DISSERTATION

DESIGN AND OPTIMIZATION OF EMERGING INTERCONNECTION AND MEMORY SUBSYSTEMS FOR FUTURE MANYCORE ARCHITECTURES

Submitted by

Ishan G Thakkar

Department of Electrical and Computer Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2018

Doctoral Committee:

Advisor: Sudeep Pasricha

Wim Bohm
Anura Jayasumana
Kevin Lear

ABSTRACT

DESIGN AND OPTIMIZATION OF EMERGING INTERCONNECTION AND MEMORY SUBSYSTEMS FOR FUTURE MANYCORE ARCHITECTURES

With ever-increasing core count and growing performance demand of modern data-centric applications (e.g., big data and internet-of-things (IoT) applications), energy-efficient and low-latency memory accesses and data communications (on and off the chip) are becoming essential for emerging manycore computing systems. But unfortunately, due to their poor scalability, the state-of-the-art electrical interconnects and DRAM based main memories are projected to exacerbate the latency and energy costs of memory accesses and data communications. Recent advances in silicon photonics, 3D stacking, and non-volatile memory technologies have enabled the use of cutting-edge interconnection and memory subsystems, such as photonic interconnects, 3D-stacked DRAM, and phase change memory. These innovations have the potential to enhance the performance and energy-efficiency of future manycore systems.

However, despite the benefits in performance and energy-efficiency, these emerging interconnection and memory subsystems still face many technology-specific challenges along with process, environment, and workload variabilities, which negatively impact their reliability overheads and implementation feasibility. For instance, with recent advances in silicon photonics, photonic networks-on-chip (PNoCs) and core-to-memory photonic interfaces have emerged as scalable communication fabrics to enable high-bandwidth, energy-efficient, and low-latency data communications in emerging manycore systems. However, these interconnection subsystems still face many challenges due to thermal and process variations, crosstalk noise, aging, data-snooping

Hardware Trojans (HTs), and high overheads of laser power generation, coupling, and distribution, all of which negatively impact reliability, security, and energy-efficiency. Along the same lines, with the advent of through-silicon via (TSV) technology, 3D-stacked DRAM architectures have emerged as small-footprint main memory solutions with relatively low per-access latency and energy costs. However, the full potential of the 3D-stacked DRAM technology remains untapped due to thermal- and scaling-induced data instability, high leakage, and high refresh rate problems along with other challenges related to 3D floorplanning and power integrity. Recent advances have also enabled Phase Change Memory (PCM) as a leading technology that can alleviate the leakage and scalability shortcomings of DRAM. But asymmetric write latency and low endurance of PCM are major challenges for its widespread adoption as main memory in future manycore systems.

My research has contributed several solutions that overcome multitude of these challenges and improve the performance, energy-efficiency, security, and reliability of manycore systems integrating photonic interconnects and emerging memory (3D-stacked DRAM and phase change memory) subsystems. The main contribution of my thesis is a framework for the design and optimization of emerging interconnection and memory subsystems for future manycore computing systems. The proposed framework synergistically integrates layer-specific enhancements towards the design and optimization of emerging main memory, PNoC, and inter-chip photonic interface subsystems. In addition to subsystem-specific enhancements, we also combine enhancements across subsystems to more aggressively improve the performance, energy-efficiency, and reliability for future manycore architectures.

ACKNOWLEDGEMENTS

Being grateful to the Almighty for this life, I would like to thank all the individuals whose encouragement and support has made the completion of this thesis possible.

First of all, I would like to express my gratitude to my advisor, Prof. Sudeep Pasricha, who has guided me through the entire process of doctoral study step by step. I remember my first meeting with him at CSU after I defended my MS thesis in April 2013. At that time, I was seeking to join him as a PhD student to work on new exciting research areas and prepare for a career in Engineering academia through quality research, training, and professional service. He readily agreed to advise me for my PhD research and pursuit of a career in academia. In the first year as a PhD student, the coursework and research work suggested by Dr. Pasricha helped me prepare for the basic skills needed for research. After that, it was his vision and wisdom that stimulated me to look at research problems with more critical and creative thinking, which led to several publications in well-known conferences and journals. Over countless times, I was impressed by his thoroughness and attention to detail despite his tight schedule, from which I got to know his passion and enthusiasm for research. On the other hand, he is the type of advisor that is caring enough to suggest his graduate students to slow down, get some rest, and recharge whenever he senses high pressure on them. Dr. Pasricha can also give good life advice when inquired, which helped me to overcome various difficulties and confusions in life and study during my graduate school years. I really appreciate all the help, guidance, and inspiration I received from Dr. Pasricha, who made it possible for me to survive the trial of graduate school with unforgettable memories and broadened horizons.

I would like to take this opportunity to thank the respected members of my PhD committee, Dr. Kevin Lear, Dr. Anura Jayasumana, and Dr. Wim Bohm. Their feedback helped me to rediscover my research and refine my work from different perspectives. Especially, I would like to convey my added gratefulness to Dr. Kevin Lear and Dr. Anura Jayasumana for their insightful research guidance and valuable life and career advice through different phases of my time as a PhD student. Furthermore, my special thanks to my research partner and dear friend Sai Vineel Reddy Chittamuru, whose collaboration greatly helped me to broaden my research contributions. I am also thankful to my mates in Dr. Pasricha's EPIC lab for their company and help during my Ph.D. studies: Nishit Kapadia, Srinivas Desai, Tejasi Pimpalkhute, Yaswanth Raparti, Vipin Kumar Kukkala, Daniel Dauwe, Vinay Ugave, Yi Xiang, Yong Zou, Varun Bhatt, Sai Kiran Koppu, Saideep Tiku, Shoumik Maiti, and Raja Damrela.

Moreover, I would like to thank my parents and family for their support to pursue my Ph.D. Their kindness shaped my view of this world and made me the person I am. Finally, I would like to convey a very special thanks to my wife. Without her unwavering love, care, and support through all the highs and lows of our life, I would not have achieved this feat.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

- Ishan Thakkar, Sudeep Pasricha, "3D-Wiz: A Novel High Bandwidth, Optically Interfaced 3D DRAM Architecture with Reduced Random Access Time," IEEE International Conference on Computer Design (ICCD), Oct 2014.

- Sudeep Pasricha, Ishan Thakkar, "Re-architecting DRAM memory systems with 3D Integration and Photonic Interfaces," Memory Architecture and Organization Workshop (MeAOW), Oct 2014. (Invited)

- Ishan Thakkar, Sudeep Pasricha, "A Novel 3D Graphics DRAM Architecture for High-Performance and Low-Energy Memory Accesses," IEEE International Conference on Computer Design (ICCD), Oct 2015.

- Ishan Thakkar, Sudeep Pasricha, "3D-WiRED: A Novel WIDE I/O DRAM with Energy-Efficient 3-D Bank Organization," IEEE Design & Test, vol. 32, no. 4, pp. 71-80, 2015.

- Ishan Thakkar, Sudeep Pasricha, "3D-ProWiz: An Energy-Efficient and Optically-Interfaced 3D DRAM Architecture with Reduced Data Access Overhead," IEEE Transactions on Multi-Scale Computing Systems (TMSCS), vol. 1, no. 3, pp. 168-184, Sept 2015. (Best Paper Candidate)

- Ishan Thakkar, Sudeep Pasricha, "Massed Refresh: An Energy-Efficient Technique to Reduce Refresh Overhead in Hybrid Memory Cube Architectures," IEEE International Conference on VLSI Design (VLSI), Jan 2016.

- Sai Vineel Reddy Chittamuru, Ishan Thakkar, Sudeep Pasricha, “Process Variation Aware Crosstalk Mitigation for DWDM based Photonic NoC Architectures,” IEEE International Symposium on Quality Electronic Design (ISQED), Mar 2016. (Best Paper Award Finalist)

- Sai Vineel Reddy Chittamuru, Ishan Thakkar, Sudeep Pasricha, “PICO: Mitigating Heterodyne Crosstalk Due to Process Variations and Intermodulation Effects in Photonic NoCs,” IEEE/ACM Design Automation Conference (DAC), Jun 2016.

- Ishan Thakkar, Sai Vineel Reddy Chittamuru, Sudeep Pasricha, “A Comparative Analysis of Front-End and BackEnd Compatible Silicon Photonic On-Chip Interconnects,” ACM System Level Interconnect Prediction Workshop (SLIP), Jun 2016. (Best Paper Award)

- Ishan Thakkar, Sai Vineel Reddy Chittamuru, Sudeep Pasricha, “Run-Time Laser Power Management in Photonic NoCs with On-Chip Semiconductor Optical Amplifiers,” IEEE/ACM International Symposium on Networks-on-Chip (NOCS), Aug 2016.

- Ishan Thakkar, Sai Vineel Reddy Chittamuru, Sudeep Pasricha, "Mitigation of Homodyne Crosstalk Noise in Silicon Photonic NoC Architectures with Tunable Decoupling," ACM/IEEE International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), Oct 2016.

- Ishan Thakkar, Sudeep Pasricha, “DyPhase: A Dynamic Phase Change Memory Architecture with Symmetric Write Latency,” IEEE International Conference on VLSI Design (VLSID), Jan 2017.

- Ishan Thakkar, Sai Vineel Reddy Chittamuru, Sudeep Pasricha, "HYDRA: Heterodyne Crosstalk Mitigation with Double Microring Resonators and Data Encoding for Photonic NoCs," IEEE Transactions on Very Large-Scale Integration (TVLSI), 2017.

- Ishan Thakkar, Sudeep Pasricha, "DyPhase: A Dynamic Phase Change Memory Architecture with Symmetric Write Latency and Restorable Endurance," IEEE Transactions on Computer Aided Design (TCAD), 2017.

- Sai Vineel Reddy Chittamuru, Ishan Thakkar, Sudeep Pasricha, "Analyzing Voltage Bias and Temperature Induced Aging Effects in Photonic Interconnects for Manycore Computing," ACM System Level Interconnect Prediction Workshop (SLIP), Jun 2017.

- Ishan Thakkar, Sai Vineel Reddy Chittamuru, Sudeep Pasricha, "Improving the Reliability and Energy-Efficiency of High-Bandwidth Photonic NoC Architectures with Multilevel Signaling," IEEE/ACM International Symposium on Networks-on-Chip (NOCS), Oct 2017.

- Sai Vineel Reddy Chittamuru, Ishan Thakkar, Sudeep Pasricha, "SOTERIA: Exploiting Process Variations to Enhance Hardware Security with Photonic NoC Architectures," IEEE/ACM Design Automation Conference (DAC), to appear, June 2018.

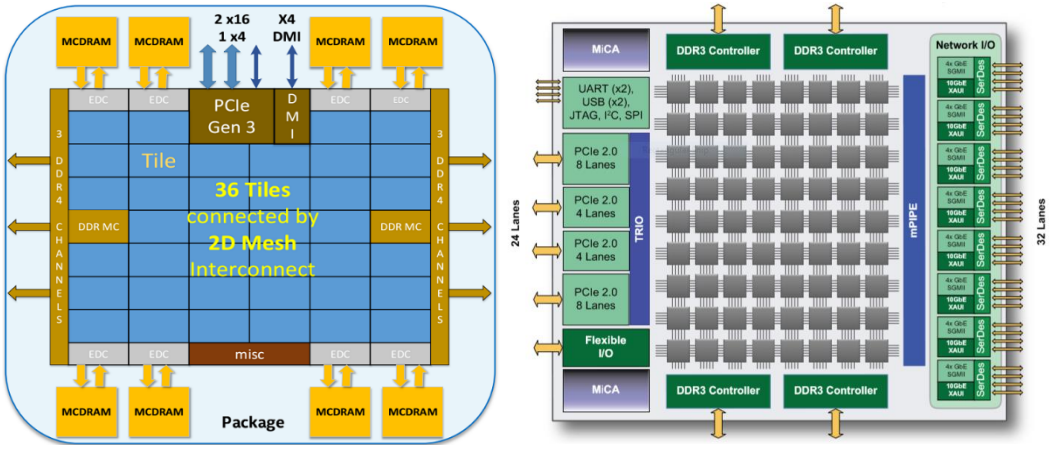
- Sudeep Pasricha, Sai Vineel Reddy Chittamuru, Ishan Thakkar, "Cross-Layer Thermal Reliability Management in Silicon Photonic Networks-on- Chip," ACM Great Lakes Symposium on VLSI (GLSVLSI), to appear, 2018.

1. INTRODUCTION

With hundreds of cores on a processor chip already becoming a reality, advanced computing has entered a manycore era! This chapter highlights the growing demand for concurrent data transfers in advanced manycore computing systems, and emphasizes the need of substituting the conventional electrical interconnects and DRAM (Dynamic Random-Access Memory) based main memory subsystems with highly scalable photonic interconnects and emerging memory subsystems (e.g., 3D-stacked DRAM and Phase Change Memory) to meet the growing data concurrency demand. Furthermore, this chapter also briefly describes the challenges faced by the emerging interconnection and memory subsystems, such as photonic interconnects, 3D-stacked DRAM, and Phase Change Memory, and presents an outline of the proposed design and optimization framework that addresses these challenges.

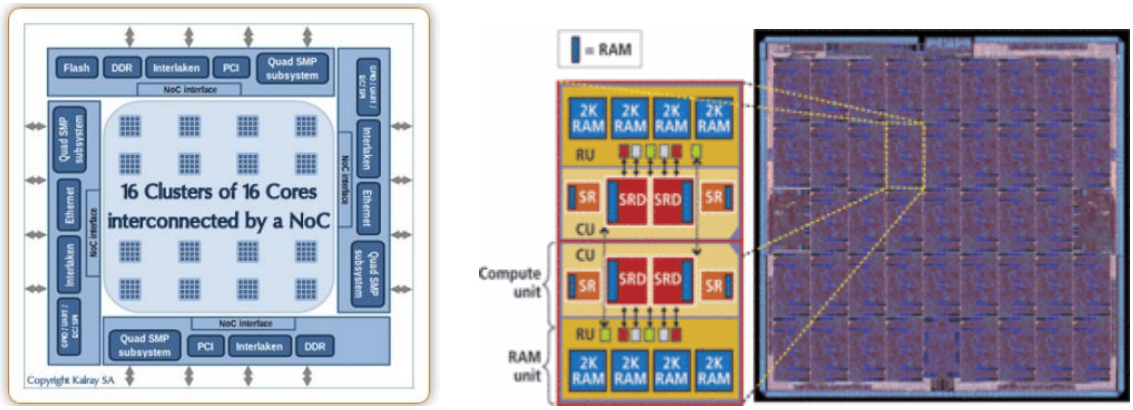
1.1. INTRODUCTION TO MANYCORE COMPUTING

According to the well-established Moore's Law [1], the transistor count on a microprocessor chip has doubled every two years or so for last 40 years due to technology scaling, although this trend is projected to end soon. As a result, integration of billions of transistors on a single chip is possible today. However, Dennard Scaling [2] ended in mid 2000s, because of which the single-thread performance, operating frequency, and power of a microprocessor chip do not grow with the growing transistor count since then. In this post Dennard-Scaling era, the available billions of transistors are used to put on more and more cores on a microprocessor chip. Consequently, microprocessors with hundreds of cores on a chip already exist today. Thus, advanced computing has entered a manycore era!



(a) **(b)**

Figure 1: (a) Intel Xeon Phi - 72 core processor [3], (b) Mellanox's 72 core processor [4], with electrical NoCs for inter-core communication, used for supercomputing applications.



(a) **(b)**

Figure 2: (a) Kalray's 288 core processor [5], (b) Ambric's 344 core Am2045 processor [6], with electrical mesh NoCs for inter-core communication, used for embedded applications.

Today, the advanced manycore microprocessors are utilized not only for supercomputing (or high-performance) applications, but also for embedded and mobile applications. For example, consider the manycore processors from Intel and Mellanox shown in Figure 1 that are used for supercomputing applications, along with the manycore processors from Kalray and Ambric shown in Figure 2 that are used for embedded applications. As evident, these manycore chips employ many processing elements along with high-capacity pools of off-chip memories. Typically, the

processing elements of a manycore processor are grouped in multiple compute clusters, which can communicate with one another and with one or more on-chip memory controllers via an electrical network-on-chip (ENoC). The utilized memory controllers connect with the employed memory modules via electrical core-to-memory interfaces. Efficient designs of interconnection fabrics (e.g., NoCs and core-to-memory interfaces) are essential to satisfy the bandwidth and latency constraints of advanced computing systems that utilize these manycore processors. It is therefore becoming evident that focus on interconnection architecture (e.g., NoCs and core-to-memory interfaces) design, customization, and exploration can provide huge performance gains in manycore processors and in advanced computing systems that utilize them.

With steadily increasing core count and growing demand of modern data-centric applications (e.g., big data, cloud computing, and IoT related applications), the demand for concurrent data transfers in advanced manycore computing systems increases by 33% every year [7]. This increasing demand for data concurrency in manycore computing systems pushes for increasingly higher bandwidth and capacity in the constituent interconnection and main memory subsystems. Unfortunately, with increasing core count, conventional electrical interconnects [3]-[6] are beginning to suffer from cripplingly high power dissipation and severely reduced performance [8]. Moreover, the susceptibility of metallic interconnects to crosstalk and electromagnetic interference has also increased with technology scaling, which has further reduced the performance and reliability of electrical interconnects [9]. On the other hand, with technology scaling, the capacity and performance of conventional planar DRAM based main memory subsystems are falling behind the capacity and performance of processing cores at 30% per year [10] and 50% per year [7], respectively. Therefore, there is a crucial need to investigate new and

more viable alternatives to conventional electrical interconnects and planar DRAM based main memory subsystems.

1.2. EMERGING INTERCONNECTION AND MEMORY SUBSYSTEMS

Emerging interconnection and memory subsystems, such as photonic interconnects (i.e., photonic NoCs and core-to-memory photonic interfaces), 3D-stacked DRAM, and Phase Change Memory, provide viable alternatives to conventional electrical interconnects and planar DRAM based main memory subsystems. This section provides a brief background on the design and functioning of these emerging subsystems and explains their benefits over the conventional interconnection and memory subsystems.

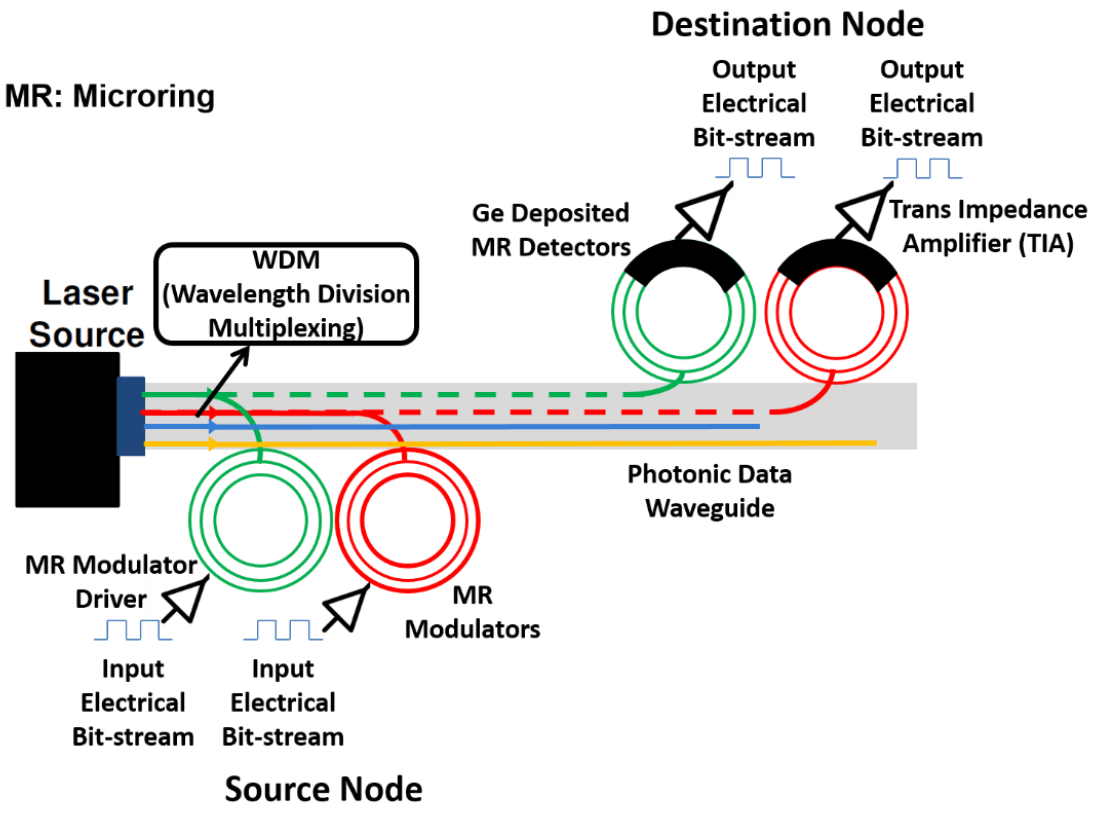


Figure 3: Overview of photonic link with wavelength division multiplexing.

1.2.1. PHOTONIC INTERCONNECTS

Recent advances in the area of silicon nanophotonics have enabled the integration of photonic devices with CMOS circuits. The resulting on-chip photonic interconnects (shown in Figure 3) have demonstrated several prolific advantages over their metallic counterparts. Photonic interconnects enable near light speed transfers as they employ photons for data communication which are 10× faster than the electrons in metallic (copper) interconnects [11]. Photonic links can also achieve distance-independent bit-rates unlike the distance dependent (crosstalk-limited) lower bit-rates in electrical wires. The photonic links are also able to employ dense wavelength division multiplexing (DWDM) [12] to achieve a bandwidth density that is 5× higher than that achieved by electrical wires. In DWDM-based photonic communication, multiple wavelengths of light can be used to simultaneously transfer multiple streams of data in a single photonic waveguide as shown in Figure 3. Additionally, because photonic links dissipate energy only at the endpoints of the communication channel [9] with low crosstalk [8] they have lower dynamic (or data-dependent) power dissipation (about 7.9 fJ/bit) than that of electronic links. Thus, silicon nanophotonics is being considered as an exciting new option for integration in future NoCs. Several photonic devices such as Microring Resonators (MRs), waveguides, and photodetectors have already been successfully fabricated and demonstrated at the chip level [13]. These devices have been used as a foundation for several PNoC architectures [14]-[16].

1.2.1.1. PHOTONIC WAVEGUIDE

In PNoC architectures, photonic waveguides are used to traverse optical signals from a source core to a destination core. Photonic waveguides, as shown in Figure 4(a), use a high refractive index silicon (Si) core (i.e., n_{si} = 3.5) and low refractive index silicon-di-oxide (SiO_2) cladding (i.e., n_{si} = 1.5) fabricated on a silicon-on-insulator (SOI) platform. These waveguides

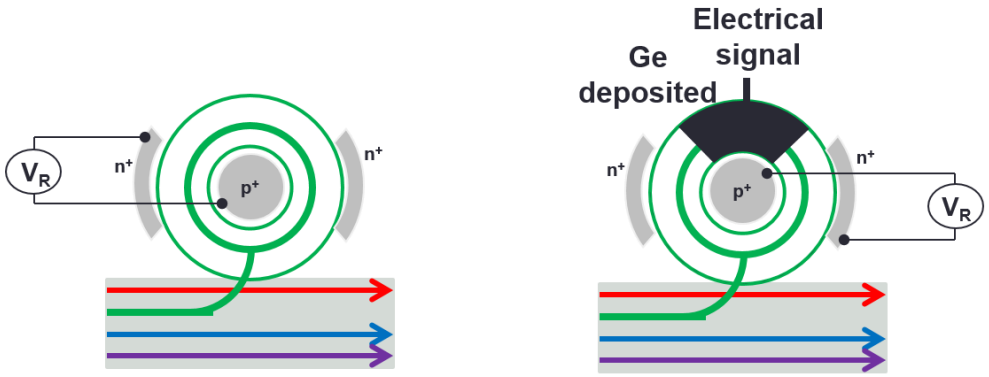
have a lower pitch and area footprint than the polymer waveguides used in [17]. Waveguides fabricated on an SOI platform have other advantages such as lower losses (on the order of 1 dB/cm) and the malleability to be curved with bend radii of ~5μm [18]. Malleability of these photonic waveguides and the SOI platform's high refractive index contrast enables fabrication of compact modulators which require lower drive voltage for high frequency operation. To support high bandwidths for future CMP applications, these photonic waveguides support dense wavelength division multiplexing (DWDM) [19], with multiple wavelengths available for concurrent data transfers in each waveguide.

Figure 4: (a) Longitudinal cross-section of photonic waveguides, (b) microring resonator, used in PNoC architectures.

1.2.1.2. MICRORING RESONATORS

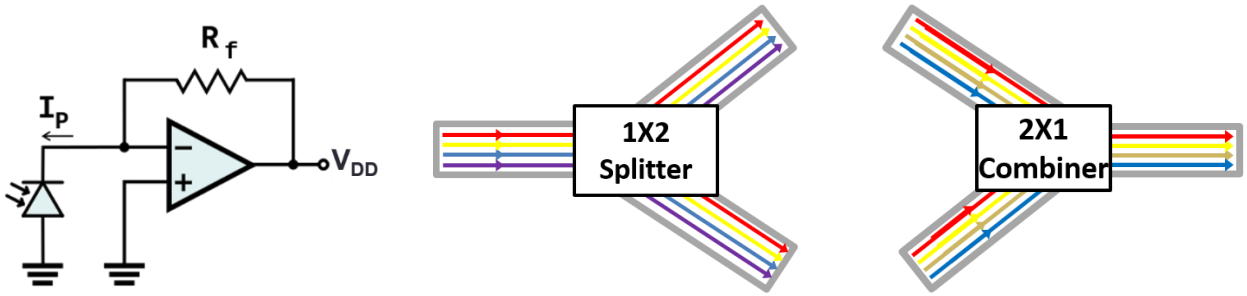
To transmit data between cores through a photonic waveguide, electrical to optical (E/O) conversion at the source and an optical to electrical (O/E) conversion at the destination is required. MRs can enable both E/O and O/E conversion in PNoCs. MRs modulate light for transmission of data at a source (data-modulation phase). MRs also detect light-modulated data from the waveguide at the destination (data-detection phase) and subsequently help with the generation of proportional electrical signals that are amplified by Trans-Impedance Amplifiers (TIAs). An MR can be functionally described as a looped photonic waveguide with a small diameter as shown in Figure 4(b).

MRs are wavelength selective and couple light when the relation $\lambda \times m = n_{eff,ring} \times 2\pi R$ is satisfied, where R is the radius of the microring resonator, $n_{eff,ring}$ is the effective refractive index, m is an integer value, and λ is the resonant wavelength [20]. As resonance wavelength is a function of R and $n_{eff,ring}$, by changing R and $n_{eff,ring}$, the resonant wavelength of the MR can be altered. It is necessary to alter resonance wavelength of an MR to remove a wavelength (in active mode to write '0'-bit) from a data waveguide, and to let a wavelength pass through (in passive mode to write '1'-bit) in a data waveguide. In general, alteration in resonance wavelength of an MR by $\Delta\boldsymbol{\lambda}$ is achieved with Δn_{eff} change in effective refractive index. There are two ways that can change the effective refractive index of an MR. Injection or removal of carriers (electrons) from the Si core of an MR alters its effective refractive index due to the Electro-Optic (EO) effect [21]. Heating of MR's also alters its effective refractive index due to the Thermo-Optic (TO) effect [22]. More details about EO and TO effects are presented in Chapter 2 and 4. However, the former method is faster and consumes lower power compared to the latter one for smaller resonance wavelength shift (i.e., <1nm) [22]. Therefore, carrier injection/removal is predominantly used to switch MRs between active and passive modes. To enable carrier injection/removal in MRs require a series of drivers. These drivers are electrical circuits which regulate carrier injection/removal rates (by altering voltage V_R shown in Figure 5) into MRs to control their resonance wavelength shifts. An MR as a modulator is shown in Figure 5(a) that removes its resonance wavelength from the data waveguide, which converts electrical signal to optical signal. Furthermore, as shown in Figure 5(b), an MR with germanium (Ge) deposited on its Si core acts as a detector to drop the corresponding resonance wavelength from the data waveguide and convert the optical signal back to an electrical signal.



(a) **(b)**

Figure 5: MR acting as (a) an active modulator to remove its resonance wavelength, (b) a detector to detect its resonance wavelength.



(a) **(b)** **(c)**

Figure 6: (a) Transimpedance amplifier (TIA), (b)1×2 splitter, (c) 2×1 combiner, used in PNoC architectures.

1.2.1.3. TRANS-IMPEDANCE AMPLIFIERS, COMBINERS, AND SPLITTERS

TIAs are used to amplify detected signals at the MR detector to digital voltage levels as shown Figure 6(a). As the signals amplified by TIAs are ultimately stored and processed on the chip, their amplitudes should match the supply voltage of logic circuits (i.e., V_{DD}). To enable amplification of signals to V_{DD}, the TIAs are typically operated at 20% higher supply voltage than V_{DD}. In addition to these TIAs, PNoCs employ splitters and combiners respectively to distribute and aggregate signal power in photonic waveguides, as shown in Figure 6(b) and (c), respectively.

1.2.2. 3D-STACKED DRAM BASED MAIN MEMORY

Today, DRAM is widely used as main memory in all types of computing systems, including servers, desktops, mobile and embedded systems, and sensors. A DRAM cell employs a capacitor to store 1-bit information in the form of electrical charge. As mentioned earlier, with technology scaling, the capacity and performance of conventional planar (2D in structure) DRAM based main memory subsystems are falling behind the capacity and performance of processing cores at 30% per year [10] and 50% per year [7], respectively. To bridge this performance gap of planar DRAMs, prefetching and high-speed interfacing techniques have been used traditionally. But using these techniques alone, meeting the high bandwidth demand beyond 70GBps per DRAM module will not be possible. On the other hand, to bridge the capacity gap of planar DRAMs, technology scaling has been used traditionally. But scaling beyond 20nm technology node makes the capacitor based DRAM cell very unstable.

As a solution to the shortcomings of planar DRAMs, some researchers have proposed 3D stacking of DRAM dies using though-silicon vias (TSVs) to achieve high-capacity and small footprint memory modules. The advances in the 3D stacking technology over the last decade have pushed the emergence of several 3D-stacked DRAM standards and architectures (e.g., [23]-[27]). For example, consider 3D-stacked DRAM standards and architectures such as Hybrid Memory Cube (HMC) [23], High-Bandwidth Memory (HBM) [25], Wide I/O DRAM [24], and Dis-Integrated RAM (DiRAM) [26]. Among these four architectures, DiRAM and HMC architectures are primarily standardized for high-performance applications, HBM standard is primarily designed for graphics applications, whereas Wide I/O DRAM standard is primarily designed for embedded applications. Use of TSVs imparts layout and packaging flexibility to these 3D-stacked DRAM

architectures, which can be leveraged to design efficient DRAM modules that can support high-bandwidth, low-latency, and energy-efficient data accesses.

1.2.3. PHASE CHANGE MEMORY

Phase Change Memory (PCM) is seen as a potential solution to the capacity and scalability related challenges of conventional planar DRAMs. Analogous to DRAM that uses a capacitor device to store a bit data, PCM utilizes a special device made of phase change material to remember information. The phase change material is one type of chalcogenide alloy, such as $Ge_2Sb_2Te_5$ (or GST in short). The GST can be switched between crystalline and amorphous states. These states have dramatically different electrical resistance. The amorphous high-resistance state (usually in the MΩ range) is used to represent a binary '0', while the crystalline low-resistance state (usually in the KΩ range) represents a '1'. The state of a GST device is preserved even after the cell is powered off, meaning that PCM is non-volatile.

Given the large resistance contrast between crystalline and amorphous states, it is possible to exploit partial crystallization states to store two or more bits per cell, forming a multi-level cell (MLC) PCM [28]. When combining MLC storage with the $4F^2$ cell size (better than the $6F^2$ cell size for DRAM), PCM can offer much better storage density than DRAM. Moreover, because the non-volatile, zero-leakage PCM cells can preserve data without power, it is possible to power down an entire memory bank or chip during idle phases to eliminate leakage power on peripheral circuits [29], which is crucial to meet the low-power requirements of future memory systems. Furthermore, the physical state based storage is immune to soft errors caused by alpha particle or cosmic radiation usually seen in voltage/charge based storages [30]. PCM also offers much better scalability as the write currents reduce with the shrinking of GST device [31]. Hence, PCM provides a truly scalable solution compared to conventional DRAM.

1.3. CHALLENGES FOR EMERGING INTERCONNECTION AND MEMORY SUBSYSTEMS

Despite the aforementioned advantages, our considered emerging interconnection and memory subsystems such as photonic interconnects (i.e., photonic NoCs and photonic interfaces), 3D-stacked DRAMs, and Phase Change Memory, face several technology-specific challenges that hinder their widespread commercial adoption. This section describes these challenges.

1.3.1. DESIGN CHALLENGES OF PHOTONIC INTERCONNECTS

We organize the design challenges faced by photonic interconnects into three categories: reliability challenges, power dissipation challenges, and security challenges. The description of each of these four challenge categories is given below.

1.3.1.1. RELIABILITY CHALLENGES

Reliability challenges in PNoC architecture design includes crosstalk noise, process variations, thermal variations, and aging of MRs. Crosstalk noise in MRs is classified into two types: heterodyne crosstalk noise and homodyne crosstalk noise. The homodyne crosstalk noise power of a particular wavelength affects the signal power of the same wavelength, whereas with heterodyne crosstalk the signal power gets affected by some noise power of one or more other (different) wavelengths. The strength of the heterodyne crosstalk noise at a detector MR depends on the following four attributes: (i) channel gap between the MR resonant wavelength and the adjacent wavelengths; (ii) Q-factors of neighboring detector MRs, (iii) the strengths of the non-resonant signals at the detector, and (iv) bit-rate or modulation rate of the photonic link. With an increase in DWDM, the channel gap between two adjacent wavelengths decreases, which in turn increases heterodyne crosstalk in detector MRs. With a decrease in Q-factors of MRs, the widths

of the resonant passbands of MRs increases, increasing passband overlap among neighboring MRs, which in turn increases heterodyne crosstalk. The strengths of the non-resonant signals depend on the losses faced by the non-resonant signals throughout their path from the laser source to the MR detector. When a data-modulated non-resonant signals passes by an MR, depending on its data bit-rate (modulation rate), a part of its signal power is dropped by the MR, which in turn affects the heterodyne crosstalk noise caused by these non-resonant signals.

Fabrication process variations (PV) induce variations in the width and thickness of MRs, which cause resonance wavelength shifts in MRs [32], [33]. PV-induced resonance shifts may reduce the channel gap between the resonances of the victim MRs and adjacent MRs, which increases crosstalk and worsens optical signal-to-noise-ratio (OSNR). The worsening of OSNR deteriorates the bit-error-rate (BER) in a waveguide. For example, a previous study shows that in a DWDM-based photonic interconnect, when PV-induced resonance shift is over 1/3 of the channel gap, BER increases from 10^{-12} to 10^{-6} [34]. Techniques to counteract the PV-induced resonance shifts in MRs involve realigning the resonant wavelengths by using localized trimming [21] or thermal tuning [22].

MR devices are highly sensitive to temperature fluctuations. With increase or decrease in temperature, the refractive index of an MR device changes, causing a change in its resonance wavelength. This wavelength is supposed to remain static, as the value assigned at design time [22]. As a result of this variation in resonance wavelength, an MR may be unable to write or read data in the waveguide. As the temperature increases or decreases from the MR's design (baseline) temperature, due to the resulting variations in refractive index, each MR now resonates with a different wavelength towards the red (i.e., red-shift) or blue (i.e., blue-shift) end of the visible

spectrum. This phenomenon reduces transmission reliability and also leads to wastage of available bandwidth.

To facilitate switching of resonance-modes of an MR with voltage biasing or trimming, a PN junction is created in the Si core of the MR surrounded by SiO_2 cladding. A positive/negative voltage bias is applied to this PN-junction to inject/remove free carriers into/out of the MR's Si core. For high frequency operation and lower power consumption, an MR's PN-junction is typically operated under a negative voltage bias (or reverse bias) [35]. The application of this voltage bias generates an electric field across the MR's Si core and SiO_2 cladding boundary. Similar to MOSFETs, this electric field generates voltage bias temperature induced (VBTI) traps at the Si-SiO_2 boundary of the MR over time (i.e., VBTI aging). Our analysis has shown that these VBTI aging induced traps alter carrier concentration in the Si core of MRs, which incur resonance wavelength shifts and increase optical scattering loss in MRs to decrease their Q-factor.

1.3.1.2. POWER CHALLENGES

Power challenges in PNoC architecture design includes high laser power dissipation and high trimming/tuning power dissipation. Data communication with photonic signals in photonic interconnects is lossy. Photonic signals traversing in waveguides incur propagation and bending losses and modulators and detectors incur through losses and modulator/detector insertion losses [12]. In addition to these losses, couplers and splitters incur coupling and splitting losses. The aforementioned losses in photonic signals demand higher laser power to ensure that all the detectors along the photonic interconnect receive sufficient signal power. This laser power dissipation needs to be controlled otherwise it will reduce the energy benefits of photonic interconnects. Another component of power dissipation in PNoC is static trimming and tuning power dissipation. As explained in subsection 1.3.1.2, an increase in process and thermal variations

increases trimming and tuning power dissipation. Further trimming/tuning power has linear dependency on the number of MRs used within a PNoC architecture. Therefore, PNoCs with a higher number of MRs (larger photonic footprint) incur more trimming/tuning power dissipation and lead to higher energy consumption.

1.3.1.3. SECURITY CHALLENGES

PNoC architectures employ shared photonic waveguides to achieve higher data rates with the minimum amount of photonic hardware [14]-[16]. Several nodes in a PNoC architecture are able to read and write data on these shared waveguides. Furthermore, several PNoC architectures [14], [15], [36] send multicast or broadcast data to multiple nodes using these shared waveguides. Despite achieving higher data rates, these shared waveguides are vulnerable to security risks. A malicious node on the shared waveguide can steal or snoop the data from the shared waveguides and transmit it to a malicious core to extract sensitive information from the data. Furthermore, malicious nodes on the shared waveguides can performs deep packet inspection and inject faults on links to develop a denial-of-service (DoS) attack. In addition, malicious nodes can corrupt data on the shared waveguides and increase bit errors in PNoCs beyond correctable limits.

1.3.2. CHALLENGES FOR 3D-STACKED DRAMS

With the advent of through-silicon via (TSV) technology, 3D-stacked DRAM architectures have emerged as small-footprint main memory solutions with relatively low per-access latency and energy costs. However, the full potential of the 3D-stacked DRAM technology remains untapped due to thermal- and scaling-induced data instability, high leakage, and high refresh rate problems along with other challenges related to 3D floor-planning and power integrity. Typically, 3D-stacked DRAMs pack more memory cells in a smaller footprint, which increases DRAM

power density and on-die temperature, causing faster capacitor leakage and data corruption. Thus, due to the decreased data retention time, 3D-stacked DRAMs require more frequent refreshes, which exacerbates the performance overhead of refresh operations in 3D-stacked DRAMs.

Despite being beneficial in terms of performance and energy-efficiency, use of TSVs in 3D-stacked DRAM architectures brings several other constraints and challenges in addition to the exacerbated refresh overhead. For instance, use of TSVs at finer granularity, i.e., at the subarray-level or tile-level granularity, allows for smaller bank sizes and larger bank counts, which typically results in greater bank-level parallelism and memory access performance. But doing do also raises 3D floor-planning related inefficiencies due to more stringent pitch-matching and wire layout constraints. Moreover, a 3D-stacked DRAM module typically employs a 3D power delivery network (PDN) to power the stacked DRAM tiers individually. Accessing the available bank-level parallelism of a 3D-stacked DRAM module through the 3D PDN raises the noise levels in the PDN across the stacked DRAM tiers in a non-uniform manner, the asperity of which depends on the under-lying PDN structure and the temperatures of the individual DRAM tiers. The increase in the PDN noise levels above acceptable limits can result in malfunctions and erroneous DRAM operation.

1.3.3. CHALLENGES FOR PHASE CHANGE MEMORY

Major challenges for the widespread adoption of phase change memory (PCM) as main memory are its asymmetric write latency and low write endurance. PCM has asymmetric write latency as writing ‘1’ in a PCM cell takes about 2-5× longer than writing ‘0’. To write ‘0’ in a PCM cell, a strong programming current pulse (called RESET pulse) of short duration is utilized. This programming pulse raises the temperature of the chalcogenide material (GST) to its melting point, after which the pulse is quickly terminated. Subsequently, the small region of melted

material cools quickly, leaving the GST material programmed in the amorphous state. As the region of the GST material that melts due to the RESET pulse is small, the required pulse time is short (tens of nanoseconds). In contrast, to write '1' in a PCM cell, a programming current pulse (called SET pulse) of a longer duration and weaker strength is applied to program the cell from the amorphous state to the crystalline state. For the SET operation, the temperature of the GST material is raised above its crystallization temperature (300 °C) but below its melting point (600 °C) for a sufficient duration of time. As the crystallization rate is a function of temperature, and given the variability of PCM cells, reliable crystallization requires a SET programming pulse of hundreds of nanoseconds. Thus, the SET latency (latency of writing '1') of a PCM cell is longer than the RESET latency (latency of writing '0'), which significantly increases average write latency of a PCM system, imposing a major challenge for PCM's widespread adoption.

Another major challenge is that a typical PCM cell can endure only 10^8-10^9 cell writes before permanent failure. In fact, faults start to appear in PCM cells long before 10^8 cell writes. As discussed in [37], the resistance of a typical PCM cell in the amorphous state (RESET state) significantly decreases as the number of cell writes increase beyond 10^4. In this state, the PCM cell is referred to as having stuck-SET failure. This trend of decreasing RESET resistance with increasing cell writes continues until near the end of the cell's lifetime, where the RESET resistance of the cell drastically increases with increase in the number of cell writes before the electrical path between the chalcogenide material GST and access device in the cell severs after about 10^8 cell writes. In this state, the PCM cell is referred to as having stuck-RESET failure. As will be explained in Chapter 13, a PCM cell that is written to every 1s would last only for about 3 years before a permanent stuck-RESET failure ends its lifetime. Thus, short lifetime caused by low write endurance presents another major challenge for PCM's widespread adoption.

1.4. DISSERTATION OUTLINE

To address the challenges presented in the previous section, in this dissertation, we propose a framework for the design and optimization of photonic interconnects and emerging memory (3D-stacked DRAM and Phase Change Memory) subsystems. A high-level preview of my contributions is given in Figure 7. As part of our proposed framework, we have proposed several cross-layer solutions that combine enhancements at the circuit level, microarchitecture level, and system level towards the design of high-bandwidth, reliable, energy-efficient, and secure photonic interconnects. Moreover, we have contributed several solutions that combine optimizations in the throughput, concurrency, access latency, energy-efficiency, and endurance of emerging memory (3D-stacked DRAM and Phase Change Memory) subsystems. The rest of this dissertation is organized as follows.

In chapter 2, we present a novel cross-layer heterodyne crosstalk mitigation framework called *HYDRA* [38] to enable reliable communication in emerging PNoC-based manycore chips. We present device-level analytical models to capture the deleterious effects of localized trimming and thermal tuning in MRs. We extend these models for system-level heterodyne crosstalk analysis. We also propose a device-level technique in this chapter for heterodyne crosstalk mitigation (DMCM) that uses double MRs to improve worst-case optical SNR (OSNR) in detectors by tailoring the MRs' passbands to have a steeper roll-off. Furthermore, a circuit-level technique for heterodyne crosstalk mitigation (EDCM) is proposed that aims to improve worst-case OSNR in detectors by encoding data to avoid undesirable data value occurrences. Lastly, we combine DMCM and EDCM into a holistic cross-layer heterodyne crosstalk mitigation framework called *HYDRA* and evaluate it on three well-known crossbar PNoC architectures as well as prior work on heterodyne crosstalk mitigation.

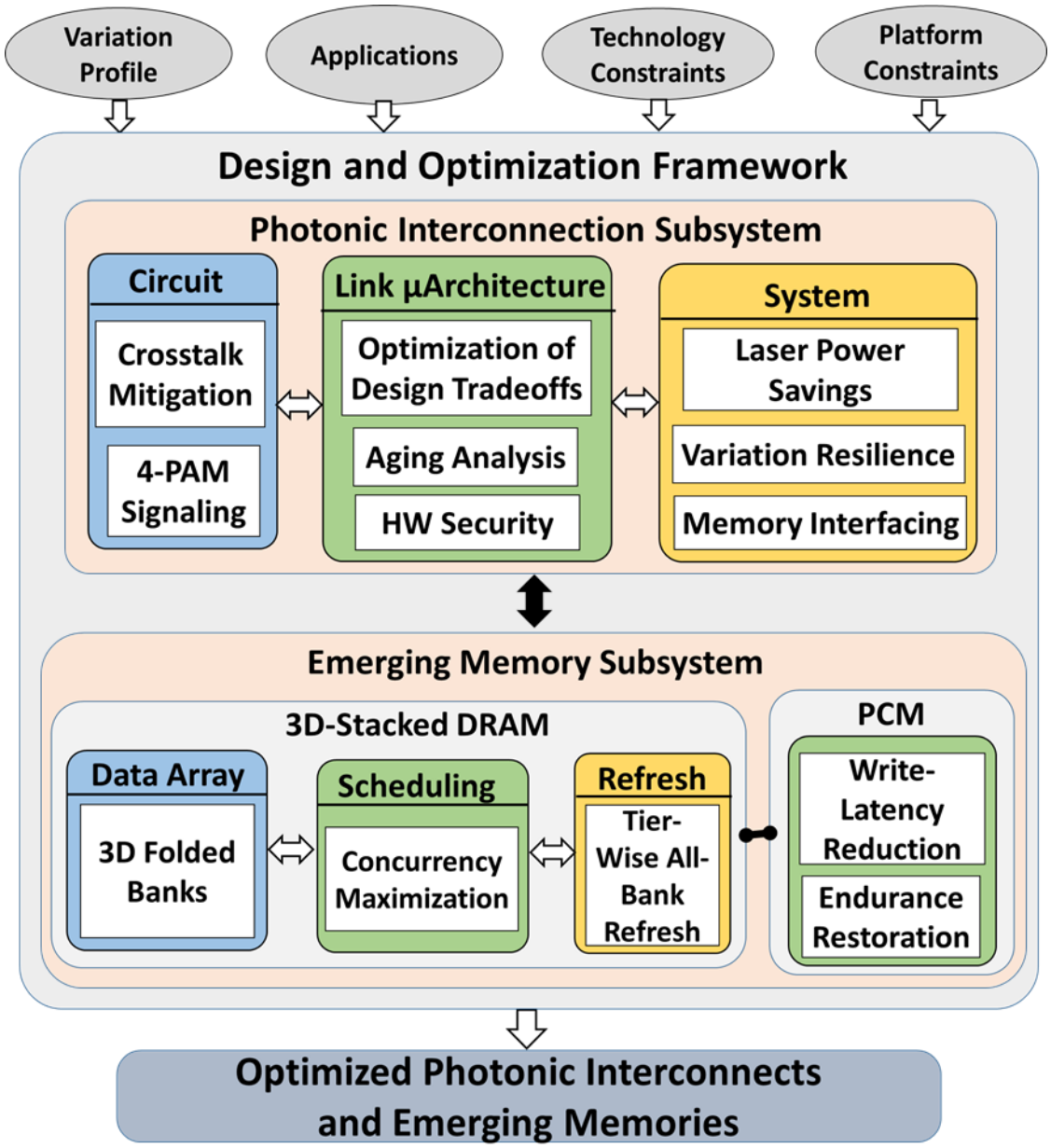


Figure 7: Outline of contributions of this dissertation.

In chapter 3, we present a novel lightweight technique to mitigate homodyne crosstalk noise in DWDM-based PNoCs [39]. Due to the resonant nature of microring resonators (MRs), the power built-up in their cavity gradually recouples back into the photonic waveguides. This recoupled power induces time-dependent unfilterable homodyne crosstalk noise, when the wavelength of the recoupled power matches with the wavelength of a signal in the waveguide. The

homodyne crosstalk in turn deteriorates the signal-to-noise ratio (SNR) and on-chip communication reliability. We evaluate the effectiveness and overhead of our homodyne crosstalk mitigating technique by implementing it for well-known PNoC architectures.

In chapter 4, we propose a thermal and process variation aware dynamic reliability management framework called *LIBRA* that integrates adaptive MR assignment at the device-level and dynamic thread migration at the system-level for PNoC-based manycore systems. The adaptive thermal and process variation aware microring assignment (TPMA) mechanism at the circuit-level tunes a set of photonic microring resonators (MRs) dynamically for reliable modulation and reception of data from a photonic waveguide in a specific temperature and process variation range. This technique aims to adapt to the changing on-chip thermal profile and maintain maximum bandwidth while minimizing trimming and tuning power in the PNoC. However, TPMA cannot control maximum on-chip temperature, whose control is critical to further minimize MR trimming and tuning power. Thus, to control maximum on-chip temperature, we devise a system-level PV-aware anti-wavelength-drift dynamic thermal management (VADTM) scheme that uses SVR based thermal prediction and dynamic thread migration, to avoid on-chip thermal threshold violations, minimize hotspots, and reduce thermal tuning power for MRs. Both TPMA and VADTM work synergistically to reduce PNoC energy consumption.

In chapter 5, we show that photonic devices fabricated with back-end compatible silicon photonic (BCSP) materials can provide independence from the complex CMOS front-end compatible silicon photonic (FCSP) process, to significantly enhance photonic network-on-chip (PNoC) architecture performance [40]. In this chapter, we present a detailed comparative analysis of a number of design tradeoffs for CMOS front-end and backend compatible devices for silicon photonic interconnects. A cross-layer optimization of multiple device-level and link-level design

parameters is performed to enable the design of energy-efficient on-chip photonic interconnects using BCSP devices. We show that the optimized design of BCSP on-chip links renders more energy-efficiency and aggregate bandwidth than FCSP on-chip links, in spite of the inferior opto-electronic properties of BCSP devices.

In chapter 6, we show that PNoC architectures require a non-trivial amount of static laser power, which can offset most of the bandwidth and energy benefits [41]. In this chapter, we present a novel low-overhead technique for run-time management of laser power in PNoCs, which makes use of on-chip semiconductor amplifiers (SOA) to achieve traffic-independent and loss-aware savings in laser power consumption.

In chapter 7, first we argue that the use of larger number of dense-wavelength-division-multiplexed (DWDM) wavelengths to achieve higher bandwidth in PNoC architectures requires sophisticated and costly laser sources along with extra photonic hardware, which adds extra noise and increases the power and area consumption of PNoCs [42]. Then we present a novel method (called *4-PAM-P*) of generating four-amplitude-level optical signals in PNoCs, which doubles the aggregate bandwidth without increasing utilized wavelengths, photonic hardware, and incurred noise, thereby reducing the bit-error-rate (BER), area, and energy consumption of PNoCs.

In chapter 8, we study the VBTI aging in MRs and its impact on PNoC architectures [43]. At the device-level, we carefully developed analytical models for trap generation with VBTI aging in MRs. We also devise analytical models in this chapter that determine variations of MR resonance wavelength shifts and Q-factor with aging-induced traps. These models are further extended to examine the impact of different operating temperatures and bias voltages, as well as process variations. From those models, we follow a mathematical bottom-up approach to analyze the system-level impact of aging on different PNoC architectures.

In chapter 9, we present a framework [44] that protects data from snooping attacks and improves hardware security in PNoCs. We analyze security risks in photonic devices and extend this analysis to the link-level, to determine the impact of these risks on PNoCs. We propose a circuit-level PV-based security enhancement scheme that uses PV-based authentication signatures to protect data from snooping attacks in photonic waveguides. We propose an architecture-level reservation-assisted security enhancement scheme to improve security in DWDM-based PNoCs.

In chapter 10, we introduce *3D-ProWiz* [27], which is a high-bandwidth, energy-efficient, optically-interfaced 3D DRAM architecture with fine grained data organization and activation. *3D-ProWiz* integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism. A novel method of routing the internal memory bus to individual subarrays using TSVs and fanout buffers enables *3D-ProWiz* to use smaller dimension subarrays without significant area overhead. The use of TSVs at subarray-level granularity eliminates the need to use slow and power hungry global lines, which in turn reduces the random-access latency and activation-precharge energy. *3D-ProWiz* yields the best latency and energy consumption values per access among other well-known 3D DRAM architectures.

In chapter 11, we present a high-bandwidth 3D graphics DRAM architecture *3D-SGDRAM* with reduced access time and energy consumption [45]. A novel 3D bank organization is employed with TSVs at subarray-level granularity to activate an optimal number of subarrays in lock-step to guarantee fast and low-energy memory access without significant area overhead. A new bitline interface enables access to only a selective group of bitlines in all active subarrays during a memory transaction, which greatly reduces row activation energy with optimal page size.

In chapter 12, we present a novel, energy-efficient DRAM refresh technique called *massed refresh* [46] that simultaneously leverages bank-level and subarray-level concurrency to reduce

the overhead of distributed refresh operations in the Hybrid Memory Cube (HMC). In *massed refresh*, a bundle of DRAM rows in a refresh operation is composed of two subgroups mapped to two different banks, with the rows of each subgroup mapped to different subarrays within the corresponding bank. Both subgroups of DRAM rows are refreshed concurrently during a refresh command, which greatly reduces the refresh cycle time and improves bandwidth and energy efficiency of the HMC.

In chapter 13, we present a novel PCM architecture called *DyPhase* [47], which uses partial-SET operations instead of the conventional SET operations to introduce a symmetry in write latency, thereby increasing write performance and throughput. However, use of partial-SET decreases data retention time. As a remedy to this problem, *DyPhase* employs novel distributed refresh operations in PCM that leverage the available power budget to periodically rewrite the stored data with minimal performance overhead. Unfortunately, the use of periodic refresh operations increases the write rate of the memory, which in turn accelerates memory degradation and decreases its lifetime. *DyPhase* overcomes this shortcoming by utilizing a proactive in-situ self-annealing (*PISA*) technique that periodically heals degraded memory cells, resulting in decelerated degradation and increased memory lifetime.

Chapter 14 concludes this dissertation. We summarize our overall body of research and make recommendations for future research.

2. HYDRA: HETERODYNE CROSSTALK MITIGATION WITH DOUBLE MICRORING RESONATORS AND DATA ENCODING FOR PHOTONIC NOCS

DWDM in photonic links increases susceptibility to intermodulation and off-resonance filtering effects, which reduces optical signal-to-noise ratio (OSNR) for photonic data transfers. Additionally, process variations induce variations in the width and thickness of MRs causing resonance wavelength shifts, which further reduces OSNR, and creates communication errors. This chapter proposes a novel cross-layer framework called HYDRA to mitigate heterodyne crosstalk due to process variations, off-resonance filtering, and intermodulation effects in PNoCs. The framework consists of two device-level mechanisms and a circuit-level mechanism to improve heterodyne crosstalk resilience in PNoCs. Simulation results on three PNoC architectures indicate that HYDRA can improve the worst-case OSNR by up to 5.3× and significantly enhance the reliability of DWDM-based PNoC architectures.

2.1. MOTIVATION AND CONTRIBUTIONS

MRs suffer from intrinsic crosstalk-noise and power-loss due to their design imperfections. Prior work [48] categorizes crosstalk noise into two types: homodyne (coherent) and heterodyne (incoherent). The homodyne crosstalk noise power of a particular wavelength affects the signal power of the same wavelength, whereas with heterodyne crosstalk the signal power gets affected by some noise power of one or more other (different) wavelengths. Heterodyne crosstalk is a major contributor of noise in DWDM-based PNoCs, and reduces OSNR and reliability in PNoCs [48].

Due to the heterodyne crosstalk phenomenon, when a data-modulated wavelength passes by an MR, depending on its data bit-rate (modulation rate), average spectral power, and its relative detuning from the resonance of the MR, part of its power is dropped by the MR [49]. All modulator,

filter, and switch MRs can drop signal power due to heterodyne crosstalk. This heterodyne crosstalk induced signal power drop creates impairments in the passing non-resonant signals. These impairments in a signal result in smoothened transition edges, lengthened rise and fall times, dampened signal amplitude, suppressed signal strength, and reduced extinction ratio, which causes data errors in the signal [50]. The overall impact of these signal impairments is manifested as a power penalty, which is defined as the amount of extra power required at the detector to overcome the data errors caused by these signal impairments.

Heterodyne crosstalk induced signal power drop has an additional effect, referred to as off-resonance filtering, at the filter MRs that are coupled with detectors. When a filter MR drops some power from the adjacent non-resonant signals on to a detector at its drop port, this dropped optical power (i.e., crosstalk noise power) produces proportional (pessimistic case) or shot-noise limited (optimistic case) noise current in the detector. This noise current increases the noise floor of the detector, increasing the minimum detectable signal power for the detector. As a result, the detector requires larger signal power to achieve a target OSNR in the presence of this crosstalk noise power. One of our goals is to reduce crosstalk noise power in detectors due to this off-resonance filtering effect.

The strength of the heterodyne crosstalk noise power at a detector depends on the following three attributes: (i) channel gap between the MR resonant wavelength and the adjacent wavelength signals; (ii) Q-factors of neighboring detector-coupled filter MRs, and (iii) the strengths of the non-resonant signals at the detector-coupled filter MR. With increase in DWDM, the channel gap between two adjacent wavelength signals decreases, which in turn increases heterodyne crosstalk noise power in detectors. With decrease in Q-factors of MRs, the widths of the resonant passbands of MRs increase, increasing passband overlap with neighboring non-resonant signals, which in

turn increases heterodyne crosstalk noise power. The strengths of the non-resonant signals depend on the losses faced by the non-resonant signals throughout their path from the laser source to the detector-coupled MR filter.

Intermodulation (IM) crosstalk has the biggest influence on the last attribute discussed above, causing suppression (or loss) of signal strength of non-resonant signals in a DWDM waveguide [51]. IM crosstalk occurs when a modulator MR induces impairments in, and as a result, suppresses the neighboring non-resonant signals. Thus the level of heterodyne crosstalk noise power and resultant OSNR at the detector depends on the amount of IM crosstalk induced signal suppression at the modulator. This motivates mitigating the effects of IM crosstalk induced signal suppression on heterodyne crosstalk by controlling the strengths of the non-resonant signals at the detector.

Additionally, fabrication process variations (PV) induce variations in the width, thickness, and doping concentration width and thickness of active MRs, which cause resonance wavelength shifts in MRs [33], [52]. PV-induced resonance shifts, when uncompensated, may reduce the gap between the resonances of the victim MRs and adjacent MRs, which increases crosstalk and worsens OSNR. For example, a previous study shows that in a DWDM-based photonic link with 1.48nm channel spacing and 4 Gbps bit-rate, when PV-induced resonance shift is over 1/3rd of the channel gap, bit-error-rate (BER) increases from 10^{-12} to 10^{-6} [34]. Techniques to counteract PV-induced resonance shifts in MRs involve realigning the resonant wavelengths by using localized trimming [21] or thermal tuning [53]. Localized trimming induces a blue shift in the resonance wavelengths (to compensate PV-induced red shifts) of MRs using carrier injection into MRs, whereas thermal tuning induces a red shift in the resonance wavelengths (to compensate PV-induced blue shifts) of MRs through heating or thermal tuning of MRs using micro-heaters. However, our analysis has shown that localized trimming and thermal tuning increase intrinsic

optical loss in MRs and signal loss in waveguides due to the free carrier absorption effect (FCA) [54] and increased optical scattering [55]. It is important to address this increase in loss, which drives the MR away from critical coupling and decreases its Q-factor, increasing heterodyne crosstalk and reducing OSNR [56].

In this chapter, we present a novel cross-layer heterodyne crosstalk mitigation framework called HYDRA to address the abovementioned challenges and enable reliable communication in emerging PNoC-based manycore chips. Our framework has low overhead and is easily implementable on any existing DWDM-based PNoC without major modifications to the architecture. Our novel contributions are:

- We present device-level analytical models to capture the deleterious effects of localized trimming and thermal tuning in MRs. We also extend these models for system-level heterodyne crosstalk analysis;
- We propose a device-level method for IM effect induced signal suppression aware heterodyne crosstalk mitigation (IMCM) that improves worst-case OSNR in detectors by controlling non-resonant signal power;
- We propose another device-level technique for heterodyne crosstalk mitigation (DMCM) that uses double MRs to improve worst-case OSNR in detectors by tailoring the MRs' passbands to have steeper roll-off;
- We propose a circuit-level technique for heterodyne crosstalk mitigation (EDCM) that improves worst-case OSNR in detectors by encoding data to avoid undesirable data value occurrences;

- We combine IMCM, DMCM, and EDCM into a holistic cross-layer heterodyne crosstalk mitigation framework called HYDRA and evaluate it on three well-known crossbar PNoC architectures as well as prior work on heterodyne crosstalk mitigation.

2.2. RELATED WORK

An important characteristic of photonic signal transmission in on-chip photonic waveguides is that it is inherently lossy, i.e., the light signal is subject to losses such as insertion losses in MR modulators and filters [57], propagation and bending loss in waveguides, and splitting loss in splitters. Such losses negatively impact signal strength in waveguides, which reduces OSNR for a given noise power. In addition to the optical signal loss, crosstalk noise of the constituent MRs also deteriorates OSNR. Crosstalk noise in PNoCs usually occurs due to imperfections in MRs used as optical modulators, filters, and switches. This crosstalk noise can be classified as homodyne or heterodyne.

For homodyne crosstalk, the noise power has the same wavelength as the signal power. As demonstrated in [16], out-of-phase homodyne crosstalk noise always degrades signal integrity. Homodyne crosstalk may either contribute to noise or cause fluctuations in signal power, which makes the analysis and mitigation of homodyne crosstalk complicated and beyond the scope of this work. On the other hand, heterodyne crosstalk occurs when an MR picks up some optical power from non-resonant signals (as explained in Section 2.1). This chapter proposes solutions to mitigate heterodyne crosstalk due to the off-resonance filtering effect. In the rest of the chapter, we use the term crosstalk to refer to heterodyne crosstalk, unless specified otherwise.

A few prior works have analyzed crosstalk in PNoCs. The effect of crosstalk noise on OSNR is shown to be negligible in the WDM system presented in [58], as this system uses only four WDM wavelengths per waveguide with 1.3nm channel spacing and 4 Gbps bit-rate. In [51], IM

crosstalk is shown to be negligible for a WDM link operating at 10 Gbps with a channel spacing of 1.6nm. However, in PNoC architectures that use DWDM (e.g., Corona [59] with 64 wavelength DWDM), significant crosstalk noise is expected. The damaging impact of crosstalk noise in the Corona PNoC is presented in [60], where worst-case OSNR is estimated to be 14dB in data waveguides, which is insufficient for reliable data transfers. To mitigate the impact of crosstalk noise in DWDM-based PNoC architectures, two encoding techniques and one wavelength spacing technique were presented in [61], [62]. However, none of these works considers the system-level impact of IM effects, off-resonance filtering, or process variations on crosstalk noise in DWDM-based PNoCs.

Fabrication-induced process variations (PV) impact the cross-section, i.e., width and height, of photonic devices such as MRs and waveguides. In MRs, PV causes resonance wavelength drifts, which can be counteracted by using device-level techniques such as localized trimming [21] and thermal tuning [53]. Trimming induces a blue shift in the resonance wavelengths of MRs using carrier injection into MRs, whereas thermal tuning induces a red shift in the resonance wavelengths of MRs through heating of MRs using ring heaters. Such device-level techniques are essential to overcome PV-induced drifts, but they incur high power overheads and may increase signal loss and crosstalk noise, thereby reducing OSNR. This motivates the use of supplementary system-level approaches to reduce the overheads of device-level techniques. A few prior works have explored the impact of PV on DWDM-based PNoCs at the system-level [34], [63]. In [34], a thermal tuning based approach is presented that adjusts chip temperature using dynamic voltage and frequency scaling (DVFS) to compensate for chip-wide PV-induced resonance shifts in MRs. In [63], a methodology to salvage network-bandwidth loss due to PV-drifts is proposed, which reorders MRs and trims them to nearby wavelengths. But the achievable benefits for all these

supplementary system-level techniques highly depend on the underlying system architecture and they also ignore the harmful effects of device-level PV remedies (i.e., trimming and tuning) on crosstalk.

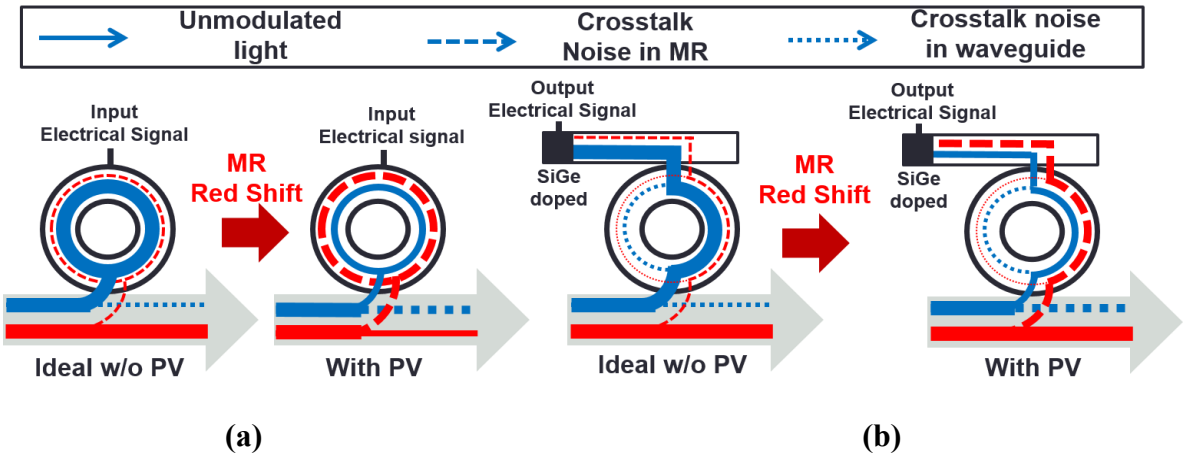


Figure 8: Impact of PV-induced resonance shifts on MR operation in DWDM waveguides (note: only PV-induced red shifts are shown): (a) MR as active modulator with PV-induced red shift, modulating in-resonance wavelength, (b) detector-coupled MR filter with PV-induced red shift, filtering its resonance wavelength and dropping it on the detector.

2.3. PV-AWARE CROSSTALK ANALYSIS

An MR can be considered to be a looped photonic waveguide with a small diameter, not to be confused with the straight photonic waveguide used for wavelength-parallel data transfers for which MRs serve as modulators and filters. Variations in an MR's dimensions due to PV cause a "shift" in its resonance wavelength. Figure 8 shows the impact of PV on crosstalk noise (dashed lines) in MRs. From Figure 8(a), PV-induced red shifts in MR modulators increase crosstalk noise in the waveguide and decrease signal strength of non-resonating wavelength signals. Figure 8(b) shows how PV-induced red shifts increase detected crosstalk noise and decrease detected signal power of resonance wavelengths in detectors, which in turn reduces OSNR and photonic data communication reliability. As discussed earlier, localized trimming and thermal tuning are essential to deal with PV-induced resonance red and blue shifts in MRs, respectively. However,

the use of these methods in an MR alters its intrinsic optical properties, which leads to increased crosstalk and degraded performance in PNoCs that use these MRs.

2.3.1. IMPACT OF LOCALIZED TRIMMING ON CROSSTALK

The localized trimming method injects extra free carriers in the circular MR waveguide to counteract the PV-induced resonance red shift. The introduction of extra free carriers reduces the refractive index of the looped MR waveguide, which in turn induces a blue shift in resonance to counteract the PV-induced red shift. However, the extra free carriers increase the absorption related optical loss in the MR due to the free carrier absorption effect (FCA) [54]. The increase in optical loss results in a decrease of MR Q-factor, which increases MR insertion loss and crosstalk. We use a PV map (described in more detail in Section 2.4) to estimate PV-induced shifts in the resonance wavelengths of all the MRs across a chip. Then, for each MR device, we calculate the amount of change in refractive index (Δn_{si}) required to counteract this PV-induced wavelength shift using the following equation [64]:

$$\Delta\lambda_r = \frac{\Delta n_{eff} * \lambda_r}{n_g} \approx \frac{\Gamma * \Delta n_{Si} * \lambda_r}{n_g}, \quad (1)$$

where, $\Delta\lambda_r$ is the PV-induced resonance shift that needs to be compensated for, λ_r is the target resonance wavelength of the MR, and n_g is the group refractive index (ratio of speed of light to group velocity of all wavelengths traversing the waveguide) of the MR waveguide. Moreover, Δn_{eff} is the change in effective index that is approximately equal to $\Gamma * \Delta n_{si}$, where Γ is the confinement factor describing the overlap of the optical mode with the MR waveguide's silicon core. The waveguides used in this study (both MRs' looped waveguides and straight bus waveguides) are rectangular channel waveguides fabricated using Si-SiO_2 material with a cross section of

450nm×220nm. We model these waveguides using a commercial eigenmode solver [65], based on which the values of Γ and n_g at 1550nm are calculated to be 0.78 and 4.16, respectively.

The change in free carrier concentration required to induce refractive index change of Δn_{si} at around 1.55μm wavelength can be quantified as follows [54]:

$$\Delta n_{si} = -8.8 \times 10^{-22} \Delta N_e - 8.5 \times 10^{-18} (\Delta N_h)^{0.8}, \quad (2)$$

where, ΔN_e and ΔN_h are the change in free electron concentration and free hole concentration respectively. The change in the absorption loss coefficient ($\Delta \alpha_{si}$) due to the change in free carrier concentration (owing to the FCA effect) can be quantified using the following equation [54]:

$$\Delta \alpha_{si} = 8.5 \times 10^{-18} \Delta N_e + 6.0 \times 10^{-18} \Delta N_h, \quad (3)$$

Quality factor (Q-factor) is a measure of the sharpness of the MR's resonance relative to its central (resonant) wavelength [64]. The Q-factor of MRs affects the magnitudes of crosstalk penalties (as explained in [51] and [49]) and determines the photon-lifetime limited allowable bitrate of signals [40]. Moreover, the Q-factor of an MR represents the number of oscillations of the field in the MR before the circulating field-energy in the MR is depleted to 1/e of the initial energy [64]. Now, from [64], the field-energy decay in the MR cavity depends on the losses in the cavity. Therefore, the Q-factor of an MR depends on the MR's loss coefficient (α) along with some other parameters. The relationship between the Q-factor and the change in absorption loss coefficient ($\Delta \alpha_{si}$) is given by the Eq. (4) and (5) [64]:

$$Q' = Q + \Delta Q = \frac{2\pi^2 R n_g \sqrt{r_1 r_2 a'}}{\lambda_r (1 - r_1 r_2 a')}, \quad (4)$$

$$a' = a + \Delta a = e^{-\pi R (\alpha + \Gamma . \Delta \alpha_{Si})}, \quad (5)$$

where, r_1 and r_2 are the self-coupling coefficients of an add-drop MR (defined in [64]); R is the MR radius; a' is the resultant round-trip field-transmission after an arbitrary change Δa in the

original round-trip field-transmission a; $\Delta\alpha_{Si}$ is the change in the MR's original loss coefficient α; and ΔQ is the change in the loaded Q-factor (Q). Eq. (4) gives the resultant loaded Q-factor Q' for an add-drop MR. Similarly, the Q' for an all-pass MR (described in [64]) can be modeled by setting r_2=1 in Eq. (4). Note that, as depicted in Figure 8, we use all-pass MRs as modulators and add-drop MRs as filters and switches.

Now, the original loss coefficient α is a sum of three components: *(i)* intrinsic loss coefficient due to material loss and sidewall roughness induced scattering loss; *(ii)* bending loss coefficient, which is a result of the curvature in the MR; and *(iii)* the absorption effect factor that depends on the original free carrier concentration in the waveguide core. Typically, the localized trimming method (when used to induce a blue-shift in the MR resonance) injects excess concentration of free carriers into the MR, which increases the absorption loss coefficient (positive $\Delta\alpha_{si}$). As evident from Eq. (5), a positive value of $\Delta\alpha_{si}$ results in a decrease in a', which in turn decreases the Q-factor Q' (from Eq. (4)). This causes a broadening of the MR passband, which results in increased insertion loss, crosstalk noise, and signal impairment/degradation related power penalty.

We model the MR transmission spectrum using a Lorentzian function [66]. In Eq. (6), this function is used to represent coupling factor φ [48] between wavelength λ_i and an MR with resonance wavelength λ_j. From [48], we use this coupling factor φ to model the heterodyne crosstalk noise power (of wavelength λ_i) that is dropped on the detector at the drop port of a filter MR with resonance wavelength λ_j. From [51], intermodulation crosstalk incurred by a modulator MR induces signal impairment, suppressing the power in the adjacent signal. As in [51], we use the same Lorentzian function to determine a loss factor γ in Eq. (7), which is the factor by which signal power of a wavelength λ_i is suppressed when it passes by a modulator MR whose resonance wavelength is λ_j. Thus, when a wavelength signal in a waveguide passes by a modulator MR, the

intermodulation-crosstalk induced bit-rate independent suppression in its power can be modeled as a through loss, which is defined as γ times the signal power before it passes by the MR.

Now from Eq. (3)-(5), Q' of an MR decreases with localized trimming based increase in carrier concentration. This in turn increases φ and crosstalk noise power (Eq. (6)). Note that we do not consider the effect of decrease in free carrier concentration, as we use only carrier injection for both modulation and trimming (to counteract PV-induced red shifts). As would be clear in Section 2.3.2, we do not need to use carrier depletion with trimming, as we would rather heat up the MRs at higher temperatures to counter the PV-induced blue shifts.

$$\Phi(\lambda_i, \lambda_j, Q') = (1 + (\frac{2Q'(\lambda_i - \lambda_j)}{\lambda_j})^2)^{-1}, \quad (6)$$

$$\gamma(\lambda_i, \lambda_j, Q') = (1 + (\frac{2Q'(\lambda_i - \lambda_j)}{\lambda_j})^{-2})^{-1}, \quad (7)$$

2.3.2. IMPACT OF THERMAL TUNING OF MR ON CROSSTALK

As mentioned earlier, localized trimming based carrier injection induces blue shifts in resonance wavelengths of MRs, which can be used to compensate PV-induced red shifts in resonance wavelengths. In contrast, thermal tuning of MRs incurs red shifts in resonance wavelengths of MRs, which can be used to compensate PV-induced blue shifts in resonance wavelengths. From Section 2.3.1, localized trimming results in increased absorption loss coefficient and subsequent decrease in Q-factor and increase in insertion loss and crosstalk power penalties. Similarly, it can be intuitively inferred that heating of MRs would also increase the absorption loss coefficient in MRs, because, the increase in temperature from the heating of MRs imparts enough energy to some valence electrons of doped silicon (constitutive semiconductor material of MRs) so that they become free carriers. However, these extra free electrons do not significantly increase the net concentration of free carriers in doped silicon. This is because, in

doped silicon, the majority of free carriers emanate from the ionization of dopant atoms and usually, all the dopant atoms are completely ionized at room temperature [67]. Thus, any increase in the MR operating temperature above room temperature does not cause ionization of any more dopant atoms. As a result, the concentration of the majority free carriers, and hence, the net free carrier concentration in doped silicon does not change with heating of MRs. Therefore, heating of MRs does not increase the absorption loss coefficient of MRs.

The scattering loss coefficient (that gives fractional loss in signal amplitude) of an MR's circular waveguide is proportional to the refractive index contrast between the core and the cladding ($n_{Si} - n_{SiO2}$) of the MR waveguide and the size of the surface roughness σ, and is given by the following equation [55], [68]:

$$\alpha_{\text{scatter}} = \frac{4(\cos\theta)^3 k_0{}^2 n_1{}^2 \sigma^2}{\sin\theta} . \left(\frac{k_0 \sqrt{n_1{}^2 (\sin\theta)^2 - n_2{}^2}}{L k_0 \sqrt{n_1{}^2 (\sin\theta)^2 - n_2{}^2} + 2} \right), \tag{8}$$

where, $\alpha_{scatter}$ is scattering loss coefficient, k_0 is the free-space wave number, $n_1 = n_{Si}$ is the MR core's refractive index, $n_2 = n_{SiO2}$ is the MR cladding's refractive index, L is the MR thickness, and θ is the propagation angle for the fundamental mode in the MR. With heating of the MR, the refractive index n_{Si} (of the MR's core) and the refractive index n_{SiO2} (of the MR's cladding) increase to their new values of $n_{Si} + \Delta n_{Si}$ and $n_{SiO2} + \Delta n_{SiO2}$ respectively, which are given by the following equations (9) and (10).

$$n_{Si}^{T+\Delta T} = n_{Si}^{T} + \Delta n_{Si} = n_{Si}^{T} + \frac{\delta n_{Si}}{\delta T} . \Delta T, \tag{9}$$

$$n_{SiO2}^{T+\Delta T} = n_{SiO2}^{T} + \frac{\delta n_{SiO2}}{\delta T} . \Delta T, \tag{10}$$

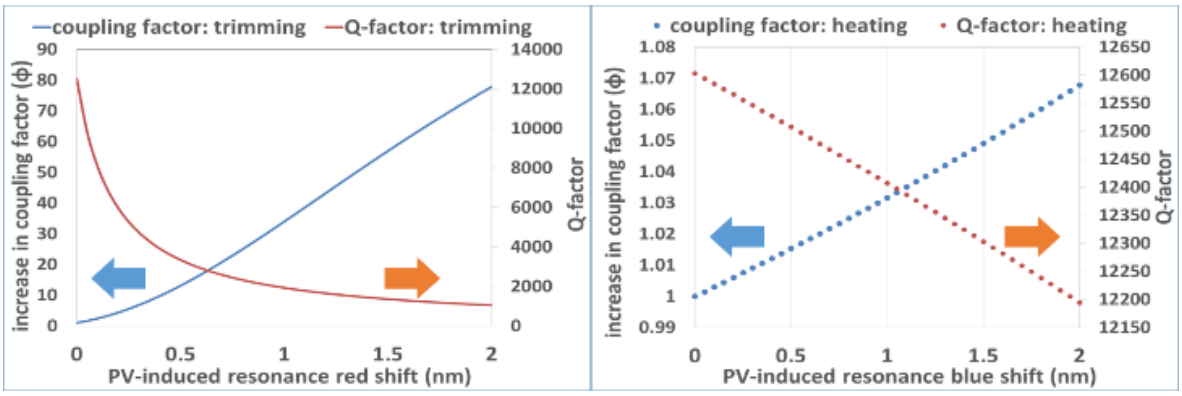
where, $\delta n_{Si}/\delta T$ and $\delta n_{SiO2}/\delta T$ are the thermo-optic coefficients of Si (MR's core) and SiO_2 (MR's cladding) materials respectively, and they assume the values of 1.86×10^{-4} K^{-1} and 1×10^{-5} K^{-1}

respectively [68]. ΔT is an increase in temperature of the MR due to heating. Due to smaller thermo-optic coefficient of SiO_2 and smaller mode field confinement in SiO_2 cladding, the effects of temperature change on $n_{SiO2}^{T+\Delta T}$ is negligible. If a blue shift in an MR's resonance wavelength of $\Delta\lambda_r$ is to be compensated by heating the MR, the required increase in MR's temperature can be computed using the following equation [69].

$$\Delta\lambda_r = \Gamma.\frac{\delta n_{Si}}{\delta T}.\frac{\lambda_r}{n_g}.\Delta T, \tag{11}$$

Now, as the thermo-optic coefficient of *Si* is greater than that of SiO_2, $n_{Si}^{T+\Delta T}$ increases faster with increase in temperature than $n_{SiO2}^{T+\Delta T}$. As a result, the difference $({n_1}^2(\sin\theta)^2 - {n_2}^2)$ in Eq. (8), which depends on the index contrast between the core and the cladding, increases with increase in temperature. This leads to an increase in $\alpha_{scatter}$ with increase in temperature (Eq. (8)). Now, similar to the case of localized trimming, this increase in scattering loss coefficient leads to decrease in MR Q-factor. Using Eq. (8), the increased value of scattering loss coefficient $\alpha_{scatter}$ can be calculated, which then can be used in place of ($\alpha+\Delta\alpha_{Si}$) in Eq. (5) to find the decreased value of Q-factor from Eq. (4).

To model and compare the effects of localized trimming and thermal tuning of MRs, we simulate an MR with a radius (R) of 1.8μm (deemed as implementable with CMOS-type processes based on projections from [70]) considering initial original Q-factor of 12500, self-coupling coefficients r_1=0.99, r_2=0.99, and field-transmission coefficient a of 0.991. Note that we use the initial Q=12500, because it gives the optimum value of total MR filter penalty for 5Gbps bitrate and 64 channels (as projected from Chapter 7). Also, note that we assume initial $\alpha_{scatter}$=0.14cm^{-1}, which corresponds to σ=1nm, n_{Si}=3.5, n_{SiO2}=1.5, L = 220nm, and θ=26.51 in Eq. (8).



(a) **(b)**

Figure 9: (a) Effect of localized trimming, (b) effect of thermal tuning, on the Q-factor and fractional increase in coupling factor of an example MR. Here, the fractional increase in coupling factor is calculated w.r.t. the original coupling factor of the MR without PV.

Using Eq. (1)-(10), we evaluate the values of Q-factor and increase in coupling factor φ for this example MR, when PV-induced red/blue shifts of different values in the resonance wavelength of this MR are compensated by using localized trimming/thermal tuning. Figure 9(a)-(b) plot these values of Q-factor and φ for localized trimming and thermal tuning respectively. From the figure, compensating 2nm PV-induced red shift in an MR's resonance wavelength with localized trimming decreases the MR's Q-factor by 91.7% and increases φ by 77.8× compared to original Q-factor and coupling factor, respectively. Furthermore, compensating 2nm of PV-induced blue shift in MR's resonance wavelength with thermal tuning decreases the MR's Q-factor by only 3.25% and increases φ by 1.07× compared to the original Q-factor and coupling factor, respectively. Thus, it can be concluded that thermal tuning of MRs has negligible impact on MRs' Q-factor and coupling factor compared to localized trimming. Therefore, compared to localized trimming, thermal tuning does not significantly increase insertion loss and crosstalk penalties for MRs.

However, note that thermal tuning cannot compensate for PV-induced red shifts in MRs' resonance wavelengths. Therefore, in a typical PNoC, where both red and blue shifts in MRs' resonance wavelengths are present, the use of localized trimming is inevitable. As a result, it is

imperative to overcome the poor efficiency of localized trimming. We propose, as part of our *HYDRA* framework, a circuit-level data encoding technique (*EDCM;* Section) that mitigates the effect of PV-remedial techniques (both localized trimming and thermal tuning) on MR crosstalk penalties. Furthermore, this chapter only analyzes the impact of PV and its remedial techniques on crosstalk noise. Evaluating the impact of thermal variations on crosstalk noise is beyond the scope of this chapter. In the next subsection, we use the derived values of Φ and γ from this and the previous section to model worst-case crosstalk and OSNR for the Corona PNoC, in the presence of process variations.

2.3.3. PV-AWARE CROSSTALK MODELS FOR CORONA PNOC

We characterize crosstalk in waveguides with DWDM for the Corona PNoC enhanced with token-slot arbitration [59]. We present equations to model the off-resonance filtering effect induced crosstalk noise power and resultant OSNR in the detectors of receiver groups. Before presenting actual equations, we show notations for parameters used in the equations, in Table 1 and Table 2.

The Corona PNoC is designed for a 256-core single-chip platform, where cores are grouped into 64 clusters, with 4 cores in each cluster. A photonic crossbar topology with 64 data waveguide groups is used for communication between clusters. Each data waveguide group consists of 4 multiple-write-single-read (MWSR) waveguides with 64-wavelength DWDM in each waveguide. As modulation occurs on both positive and negative edges of the clock in Corona, 512 bits (cache-line size) can be modulated and inserted on 4 MWSR waveguides in a single cycle by a sender. Each of the 64 data waveguide groups starts at a different cluster called ‘home-cluster’, traverses other clusters (where modulators can modulate light and receivers can filter and detect this light), and finally ends at the home-cluster again, at a set of receivers (optical termination).

Table 1: Photonic power loss, crosstalk coefficients [74], [100].

Notation	Parameter type	Parameter value (in dB)
L_P	Propagation loss	-0.274 per cm
L_B	Bending loss	-0.0085 per 90^o
L_{S12}	1×2 splitter power loss	-0.2
L_{S14}	1×4 splitter power loss	-0.2
L_{S16}	1×6 splitter power loss	-0.2

Table 2: Other model parameter notations [74].

Notation	Crosstalk Coefficient	Parameter Value
Q	Q-factor	9000
RS	Detector responsivity	0.8 A/W
L	Photonic path length in cm	
B	Number of bends in photonic path	
λ_j	Resonance wavelength of MR	
R_{S12}	Splitting factor for 1×2 splitter	
R_{S14}	Splitting factor for 1×4 splitter	
R_{S16}	Splitting factor for 1×6 splitter	

A power waveguide supplies optical power from an off-chip laser to each of the 64 data waveguide groups at its home-cluster via a series of 1×2 splitters. In each of the 64 home-clusters, optical power is distributed among 4 MWSR waveguides equally using a 1×4 splitter with splitting factor R_{S14}. As all 1×2 splitters are present before the last (64^{th}) waveguide group, this waveguide group suffers the highest signal power loss. Therefore, the worst-case signal and crosstalk noise exists in the detectors of the receiver group of the 64^{th} cluster node, and this node is called the worst-case power loss node (N_{WCPL}) in the Corona PNoC.

For this N_{WCPL} node of the Corona PNoC, the signal power ($P_{signal}(\lambda_j)$) and crosstalk noise power ($P_{noise}(\lambda_j)$) received at a receiver (i.e., detector-coupled MR filter) with resonance wavelength λ_j are expressed in Eq. (12) and (13) respectively. $K(\lambda_i)$ in Eq. (14) represents signal power loss of λ_i before the receiver group of N_{WCPL}. $\psi(\lambda_i, \lambda_j)$ in Eq. (15) represents signal power loss of λ_i before the receiver with resonance wavelength λ_j within the receiver group of N_{WCPL}.

$P_S(\lambda_i, \lambda_j)$ in Eq. (16) is the signal power of the λ_i wavelength in the waveguide that has reached the receiver with λ_j resonance wavelength in the receiver group of N_{WCPL} after passing through all the preceding receivers. Due to PV (more details about modeling of PV in PNoCs are presented in the next subsection), crosstalk coupling factor (φ, Eq. (6)) increases with decrease in loaded Q-factor (Q', which is calculated by using Eq. (4) and Eq. (5)), which in turn increases off-resonance filtering effect induced crosstalk noise in the detectors. Furthermore, $Q'_{(x\times y)+j}$ is defined as the Q-factor of j^{th} MR which is in the $x+1^{th}$ node and each node is having 'y' number of MRs. We can define OSNR(λ_j) at the detector in the receiver (with resonance wavelength λ_j) of N_{WCPL} as the ratio of $P_{signal}(\lambda_j)$ to $P_{noise}(\lambda_j)$, as shown in Eq. (17). These equations (i.e., (12)-(17)) are based on the models presented in the prior works [48] and [60].

$$P_{signal}(\lambda_j) = \Phi(\lambda_j, \lambda_j, Q'_{(63\times 64)+j})\, P_S(\lambda_j, \lambda_j), \tag{12}$$

$$P_{noise}(\lambda_j) = \sum_{i=1}^{n} \Phi(\lambda_i, \lambda_j, Q'_{(63\times 64)+j}) \left(P_S(\lambda_i, \lambda_j)\right) (i \neq j), \tag{13}$$

$$K(\lambda_i) = (R_{S14})(L_{S14})(L_P)^L (L_B)^B \prod_{n=1}^{63} \prod_{j=1}^{64} \gamma\left(\lambda_i, \lambda_j, Q'_{((n-1)\times 64)+j}\right), \tag{14}$$

$$\psi(\lambda_i, \lambda_j) = \prod_{k=1}^{(k-1)<j} \gamma\left(\lambda_i, \lambda_k, Q'_{(63\times 64)+k}\right), \tag{15}$$

$$P_S(\lambda_i, \lambda_j) = K(\lambda_i)\psi(\lambda_i, \lambda_j)\, P_{in}(i), \tag{16}$$

$$OSNR(\lambda_j) = \frac{P_{signal}(\lambda_j)}{P_{Noise}(\lambda_j)}, \tag{17}$$

2.3.4. MODELING PV OF MR DEVICES IN CORONA PNOC

We adapt the VARIUS tool [71], similar to prior work [63], to model die-to-die (D2D) as well as within-die (WID) process variations in MRs. We consider photonic devices with a silicon (Si) core and silicon-dioxide (SiO_2) cladding. VARIUS uses a normal distribution to characterize on-chip D2D and WID process variations.

The key parameters are mean (μ), variance (σ^2), and density (ω) of a variable that follows the normal distribution. As wavelength variations are approximately linear to dimension variations of MRs, we assume they follow the same distribution. The mean (μ) of wavelength variation of an MR is its nominal resonance wavelength. We consider a DWDM wavelength range in the C and L bands [72], with a starting wavelength of 1550nm and a channel spacing of 0.8nm. Hence, those wavelengths are the means for each MR modeled. The variance (σ^2) of wavelength variation is determined based on laboratory fabrication data [33] and our target die size. We consider a 256-core chip with die size 400 mm^2 at a 22nm process node. For this die size we consider a WID standard deviation (σ_{WID}) of 0.61nm [63] and D2D standard deviation (σ_{D2D}) of 1.01nm [63]. We also consider a density (ω) of 0.5 [63] for this die size, which is the parameter that determines the range of WID spatial correlation required by the VARIUS tool. With these parameters, we use VARIUS to generate 100 PV maps, these maps are used to model PV in Corona PNoC.

2.4. HYDRA FRAMEWORK: OVERVIEW

Our proposed cross-layer *HYDRA* framework enables crosstalk resilience in DWDM-based PNoC architectures by integrating device-level and circuit-level enhancements that seamlessly work together. Figure 10 gives a high-level overview of our framework. The IM effects induced signal suppression aware crosstalk mitigation (*IMCM*) scheme employs additional MRs to decrease wavelength-specific crosstalk noise at the detectors of DWDM-based photonic links. The double MR based crosstalk mitigation mechanism (*DMCM*) employs double microrings (DMRs) as signal filters to reduce the crosstalk noise at the detectors. This technique improves OSNR in DWDM-based photonic links. However, excessive usage of DMRs (or higher-order filters) increases area, PV redress power (static power required to counter PV-induced resonance drifts in the DMRs) and laser power overheads for PNoC architectures [73]. Thus, to reduce these

overheads, we also devise a circuit-level crosstalk mitigation mechanism (*EDCM*) that uses a 5-bit encoding mechanism to intelligently reduce undesirable data value occurrences in a photonic waveguide. This allows for further reduction in crosstalk noise and more effectively improves OSNR in DWDM-based PNoC architectures. The next three sections present details of the *IMCM*, *DMCM*, and *EDCM* techniques.

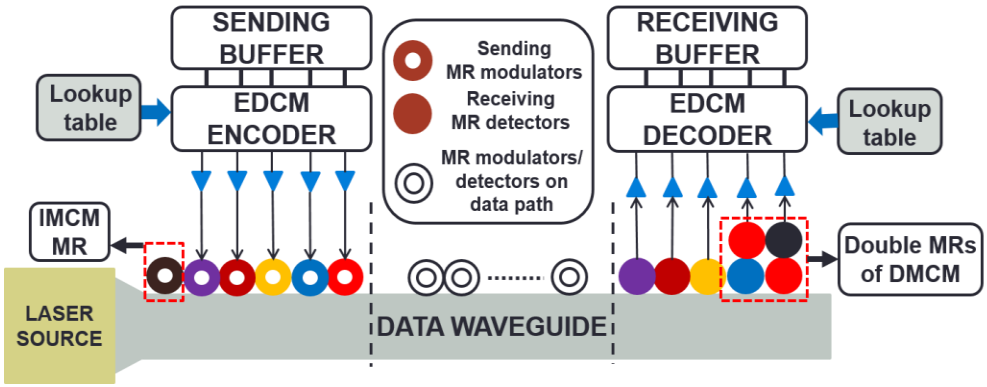
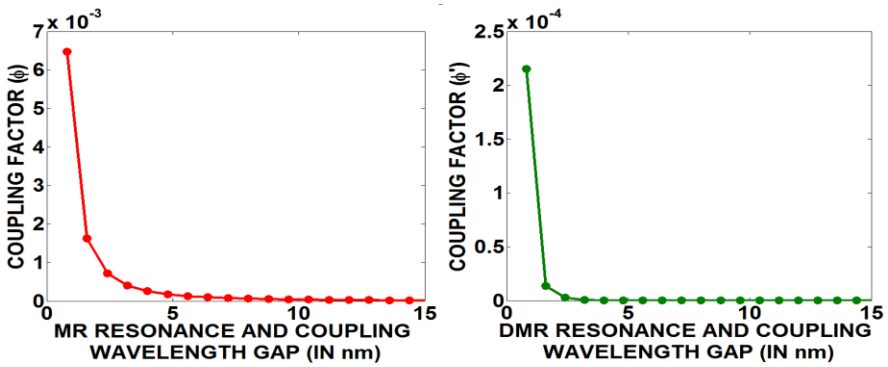


Figure 10: Overview of cross-layer *HYDRA* framework that integrates a device-level IM-aware crosstalk mitigation mechanism (IMCM) [56], a device-level double MR based crosstalk mitigation mechanism (DMCM) and a circuit-level 5-bit crosstalk mitigation mechanism (EDCM).

2.5. CROSSTALK MITIGATION WITH DMCM

Crosstalk noise in the detectors of DWDM-based PNoCs is mainly caused due to inefficient coupling of filter MRs, as filter MRs in their active mode not only couple photonic power from their resonance wavelengths but also couple a small amount of photonic power from other wavelengths in the waveguide. The coupling factor φ in Figure 11(a) represents the fraction of signal power of non-resonant wavelength coupled by an MR filter. This coupled power is then dropped on a detector at the MR's drop port. Figure 11(a) illustrates the variation of φ (using Eq. (6)) with increase in gap between the MR resonance wavelength and the non-resonant wavelength available in the waveguide. It can be seen that φ decreases abruptly with an increase in this gap. The first immediate non-resonance wavelength has almost 4× higher coupling factor than the

second immediate non-resonance wavelength considering a channel spacing of 0.8nm, Q=12500, and 5 Gbps bit-rate. We choose these values of channel spacing, Q, and bitrate, as they provide optimal value of total filter penalty for single MR filters (as projected from [49]). Thus non-resonant wavelengths closer to the MR filter's resonances create greater crosstalk noise.



(a) **(b)**

Figure 11: Coupling factor (φ/φ') variation with increase in gap between the non-resonant wavelength available in the photonic waveguide and the resonance wavelength of (a) a single MR filter, and (b) a DMR filter.

One way of reducing this crosstalk noise is to increase the Q-factor of MR filters so that φ is reduced. But doing so would increase the photon-lifetime in MR filters limiting their maximum allowable bit-rate (Chapter 5). An alternate method for reducing crosstalk is to use second-order filters with double MRs (DMRs), as used in [73] and Chapter 3, for steeper roll-off of filter response. The use of a DMR filter in place of a single MR filter is depicted in Figure 12. To further reduce crosstalk, use of filter MRs of even higher order (3rd order or higher) is possible, but as explained in Chapter 3, the use of higher-order MR filters and the choice of Q for the MR stages trade off crosstalk suppression with signal degradation due to signal side-lobe truncation. From [73], the DMRs present lower signal degradation power penalty than third order and first-order (single MR) MR filters. The optimal crosstalk performance for DMRs is achieved at 12.5Gbps bitrate or lower with 0.8nm channel spacing and the individual MRs having the Q-factor of 8000

[73]. For these reasons, in this chapter we use DMR filters with individual MR Q-factor of 8000 to reduce crosstalk noise.

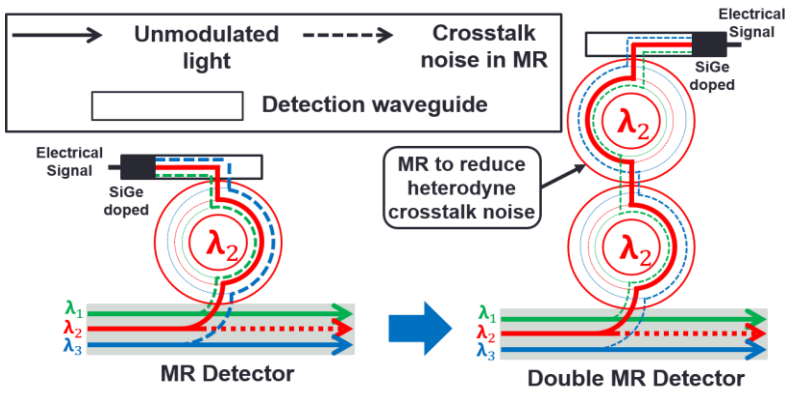


Figure 12: Crosstalk mitigation with double microring resonators: (a) MR detector operation when receiving its resonance wavelength, (b) double MR operation when receiving its resonance wavelength.

2.5.1. MODELING OF DMR FILTERS

In this section, we model the resultant coupling factor φ' and signal suppression/loss factor γ' due to the steeper roll-off of a DMR filter response. From [74], in analogy to electronic filter design, the effect of steeper roll-off of a DMR filter response can be modeled as a maximally flat Butterworth filter response. From [74], the shape, and hence the Q-factor of the Butterworth filter response does not change for higher order filters (and hence for a DMR) except that the roll-off becomes steeper. Therefore, a Butterworth type of DMR filter response can be modeled by simply setting the exponent of the term $2Q'(\lambda_i-\lambda_j)/\lambda_j$ in Eq. (6) and (7) to four instead of two. As a result, Eq. (6) and (7) can be revised for a DMR to be Eq. (18) and (19), respectively.

$$\Phi'(\lambda_i, \lambda_j, Q') = (1 + (\frac{2Q'(\lambda_i-\lambda_j)}{\lambda_j})^4)^{-1}, \quad (18)$$

$$\gamma'(\lambda_i, \lambda_j, Q') = (1 + (\frac{2Q'(\lambda_i-\lambda_j)}{\lambda_j})^{-4})^{-1}, \quad (19)$$

Here, as the Q-factor for a Butterworth DMR filter does not change from the Q-factor of a single MR filter, Q' in Eq. (18) and (19) can be modeled as the loaded Q-factor of the individual MRs using Eq. (4) and (5). We modeled a DMR with an original Q-factor of 8000 (corresponding to self-coupling coefficients r_1=0.985, r_2=0.985, and field-transmission coefficient a of 0.985 in Eq. (4) and (5)). Based on this model, we simulated φ' using Eq. (18). Figure 11(b) illustrates the variation of φ' (using Eq. (18)) with increase in gap between the DMR resonance wavelength and the non-resonant wavelength available in the waveguide. By comparing Figure 11(a) with Figure 11(b), it is evident that φ' of the MR's immediate non-resonant wavelength (with a channel gap of 0.8nm) for the DMR filter is about 30× smaller than φ for the single MR filter. Since the coupling factor is used to determine crosstalk noise power in the filter-coupled detectors, it is evident that the DMR filter reduces the crosstalk noise power by about 30×. Thus, it can be concluded that the use of DMR filters in place of signal MR filters at the receiver nodes of PNoCs results in significantly less crosstalk noise power at the detectors. Thus, our double-MR enabled crosstalk mitigation (DMCM) scheme uses DMR filters in place of single MR filters and achieves significant reduction in crosstalk noise power and improvement in OSNR at the detectors.

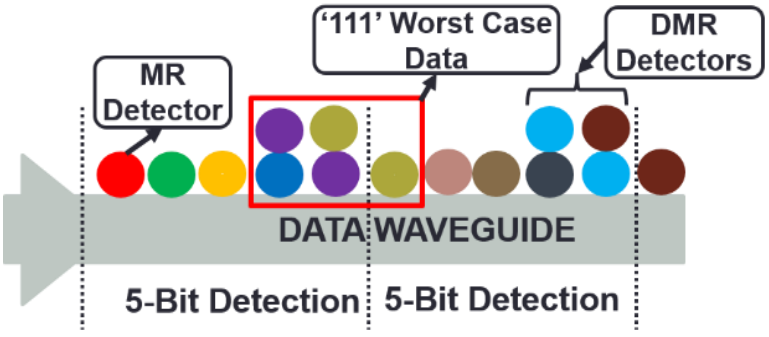


Figure 13: Organization of MR and DMR detectors in a detecting node on a photonic data waveguide with the EDCM mechanism.

2.5.2. OVERHEAD ANALYSIS FOR OUR DMCM SCHEME

In this section, we discuss the overhead of using DMR filters. From [64] and [74], as depicted in Figure 12(b), in a DMR, both the constituent MRs should be in resonance with the same wavelength (i.e., λ_2 in Figure 12(b)) to achieve a smooth filter response without any ripples or multiple peaks. However, in reality, due to the presence of PV, the constituent MRs end up having different resonance wavelengths after fabrication, which results in multiple peaks in the DMR filter's response. Therefore, the resonances of both the individual MRs of a DMR need to be aligned with trimming or tuning, which almost doubles the required trimming or tuning power for DMR filters compared to single MR filters. In addition to this, a DMR filter incurs crosstalk induced signal impairment related power penalty of 0.5dB for 0.8nm channel spacing [73] and incurs about 1.5dB insertion loss [73]. Moreover, thermal stabilization of a DMR requires 0.9mW more power [73]. Because of all these penalties, too much use of DMR filters result in a very high-power overhead. Nevertheless, we propose an intelligent method of using a few DMR filters along with a data encoding mechanism (see next section) to limit the use and overheads of DMRs and further mitigate the crosstalk noise power in the detectors.

2.6. CROSSTALK MITIGATION WITH EDCM

The crosstalk noise in a detector is also highly dependent on the strengths of the non-resonant signals at the detector. Crosstalk noise increases with increase in signal power of non-resonant wavelengths. Based on this observation, one can conjecture that crosstalk noise may be mitigated by placing one or more '0's adjacent to '1's in the data in the waveguide, to reduce photonic signal strength of non-resonant wavelengths. In this section, we present a novel technique (EDCM) at the circuit-level for mitigation of crosstalk noise in DWDM-based PNoCs.

DMRs in the DMCM technique presented in Section 2.5 increase laser power (because of higher signal loss due to higher crosstalk power penalty and insertion loss) and redress power dissipation overheads. These power overheads increase with an increase in the number of DMRs, hence there is a need to reduce the number of DMRs used with photonic waveguides. In DMCM, DMRs are beneficial when there are consecutive '1's in the parallel data word being transmitted, because consecutive '1's imply higher signal strength in the immediate non-resonant wavelengths. One way to reduce the number of DMRs while still minimizing the crosstalk noise due to consecutive '1's is by reducing the number of consecutive '1's in the parallel data word being transmitted. To do so, we propose a circuit-level scheme that employs a sophisticated encoding mechanism.

Our proposed circuit-level DMR-aware crosstalk mitigation mechanism (EDCM) places one or more '0's adjacent to '1's in the data to restrict the number of consecutive '1's in the data stream to three. EDCM employs 5-bit encoding for every 4-bit data block to restrict the number of consecutive '1's to two in the data block, which in turn limits the worst-case number of consecutive '1's in the data stream to three. Figure 13 shows the organization of MRs and DMRs in the implementation of the proposed EDCM encoding mechanism along with the location of occurrence of worst-case consecutive '1's. Table 3 shows the 5-bit codes in the EDCM scheme, to replace 4-bit data words. To implement this encoding technique on a 64-bit word, 16 additional bits are required, which in turn increases the number of MR devices by 25%. However, EDCM reduces the number of DMR detectors required by DMCM and reduces the total number of MR detectors by 12.5%. We propose to use an SRAM based lookup table with a size of 80-bits to facilitate encoding and decoding of data in each modulating and detecting node for our EDCM

mechanism. This encoding and decoding mechanism incurs a delay overhead of approximately one clock cycle, which we account for in our simulation analysis.

Table 3: Code words for EDCM technique.

Data Word	Code Word	Data Word	Code Word
0000	00000	1000	01000
0001	00001	1001	01001
0010	00010	1010	01010
0011	00011	1011	01011
0100	00100	1100	10100
0101	00101	1101	10010
0110	10011	1110	10001
0111	10101	1111	10000

2.7. HYDRA INTEGRATION WITH PNOCS

2.7.1. CORONA PNOC WITH HYDRA FRAMEWORK

In this subsection, we extend the PV-aware crosstalk models of the Corona PNoC from subsection 2.3.3 to devise PV-aware crosstalk models for Corona enhanced with the *HYDRA* framework. To integrate HYDRA with the Corona PNoC, we increase the DWDM degree in the MWSR waveguides from 64 to 65 (i.e., channel spacing is reduced from 0.8nm to 0.79nm) and increase the number of MWSR waveguides in each channel from 4 to 5 to facilitate simultaneous transfer of an entire packet (which requires 512 bits before encoding). To distribute optical power between these waveguides, there is also a need to replace 1×4 splitters with 1×5 splitters with a splitting factor of R_{S15}. Because of the increase in DWDM from 64 to 65 the number of modulators in the modulating node increases from 64 to 65. Furthermore, we need to add an additional *IMCM* MR in all modulating nodes on each MWSR waveguide, thus the total number of modulators in each modulating node on each MWSR waveguide increases to 66. In the detecting node, first we need to increase the number of detector MRs on each data waveguide from 64 to 65 and secondly as shown in Figure 13 in each group of 5 consecutive detector MRs we need to replace the last two

detector MRs with DMR detectors (replace φ, and γ with φ', and γ' respectively). Therefore, equations (12), (13), (14), and (15) for worst-case signal and crosstalk noise power are changed to equations (20), (21), (22), and (23) below respectively.

$$P_{signal}(\lambda_j) = \Phi'(\lambda_j, \lambda_j, Q'_{(63\times66)+j}) \, P_S(\lambda_j, \lambda_j), \tag{20}$$

$$P_{noise}(\lambda_j) = \sum_{i=1}^{n} \Phi'(\lambda_i, \lambda_j, Q'_{(63\times66)+j}) \left(P_S(\lambda_i, \lambda_j)\right) (i \neq j), \tag{21}$$

$$K(\lambda_i) = (\mathrm{R}_{S15})(L_{S15})(\mathrm{L}_P)^L(\mathrm{L}_B)^B \prod_{n=1}^{63} \prod_{j=1}^{66} \gamma'^{\left(\lambda_i, \lambda_j, \mathrm{Q}'_{((n-1)\times66)+j}\right)}, \tag{22}$$

$$\psi(\lambda_i, \lambda_j) = \prod_{k=1}^{(k-1)<j} \gamma'(\lambda_i, \lambda_k, \mathrm{Q}'_{(63\times66)+k}), \tag{23}$$

2.7.2. FIREFLY PNOC WITH HYDRA FRAMEWORK

To investigate the efficacy of integrating our *HYDRA* framework into other PNoC architectures, we integrated it with the Firefly [16] crossbar-based PNoC architecture. Firefly PNoC, for a 256-core system, has 8 clusters (C1-C8) with 32 cores in each cluster. Within each cluster, a group of four cores are connected to a router through a concentrator. Thus each cluster has 8 routers (R1-R8) and these routers are electrically connected using a mesh topology. Firefly uses photonic signals for inter-cluster communication. Unlike the MWSR waveguides used in the Corona crossbar, Firefly uses reservation-assisted single write multiple reader (R-SWMR) data waveguides in its crossbar. Each data channel in Firefly consists of 8 SWMR waveguides, with 64 DWDM in each waveguide. Firefly uses only 1/8th of the MRs on each data waveguide compared to Corona, as only eight nodes are capable of accessing each SWMR waveguide.

In our implementation of Firefly, we considered a power waveguide similar to that used in Corona and determined that the worst-case power loss node (N_{WCPL}) is at the detectors of C4R0, which is the router-0 (R0) of cluster-4 (C4) in this architecture. Similar to Corona, in Firefly, the

worst-case signal and noise power in the detectors of router C4R0 are calculated using Eq. (12)-(17). But as Firefly has fewer number of MRs in its data channels, this in turn changes the signal and crosstalk noise power losses before the detector group of N_{WCPL}.

To integrate *HYDRA* with the Firefly PNoC, we need to increase the DWDM degree in SWMR waveguides from 64 to 65 and increase the number of SWMR waveguides in each channel from 8 to 10 to facilitate simultaneous transfer of an entire packet (which requires 512 bits before encoding). To deal with the increase in DWDM degree, we need to increase the number of modulators and detectors from 64 to 65 on each SWMR waveguide in a modulating node and detecting node respectively. Further, we need to add an additional *IMCM* MR in all modulating and detecting nodes on each SWMR waveguide, which increases the total number of MRs in each modulating and detecting node on each SWMR waveguide to 66. Also, in each detecting node, for each group of 5 consecutive detector MRs (excluding the *IMCM* MR in that detecting node) we need to replace the last two detector MRs with DMR detectors (see Figure 13). Lastly, we determine worst-case OSNR using Eq. (20)-(23) with modified through losses.

2.7.3. FLEXISHARE PNOC WITH HYDRA FRAMEWORK

We also investigated integrating *HYDRA* with the Flexishare [16] PNoC architecture with 256 cores. We considered a 64-radix, 64 node Flexishare architecture with 4 cores in each node having 32 data channels for inter-node communication. Each data channel in Flexishare has four multiple write multiple read (MWMR) waveguides with 64 DWDM in each waveguide. Similar to the MWSR data waveguides of Corona, multiple write multiple read (MWMR) data waveguides in Flexishare also uses the models from Eq. (12)-(17) presented in subsection 2.3.3 to determine the received crosstalk noise and OSNR at detectors for the node with worst-case power loss (N_{WCPL}), which corresponds to detectors of node 63 (R_{63}).

To integrate *HYDRA* with Flexishare, we need to increase the DWDM degree in the MWMR waveguides from 64 to 65 and increase the number of MWMR waveguides in each channel from 4 to 5 to simultaneously transfer 512 bits. We also need to increase the number of modulators and detectors from 64 to 65 on each MWMR waveguide in each modulating and detecting node. Similar to the Firefly PNoC, we need to add an additional *IMCM* MR in all modulating and detecting nodes on each MWMR waveguide, which increases the total number of MRs in each modulating and detecting node on each SWMR waveguide to 66 respectively. In the detecting nodes of Flexishare, for each group of 5 consecutive detector MRs (excluding the *IMCM* MR in that detecting node), we need to replace the last two detector MRs with DMR detectors. Lastly, we can use Eq. (10)-(23) to determine worst-case OSNR.

2.8. EVALUATION

2.8.1. SIMULATION SETUP

To evaluate the efficacy of our proposed cross-layer crosstalk noise mitigation framework *HYDRA* which combines device layer (*IMCM*, *DMCM*) and circuit layer (*EDCM*) mechanisms for DWDM-based PNoCs, we integrate the framework with the Corona, Firefly, and Flexishare crossbar-based PNoCs. We modeled and performed simulation based analysis of the HYDRA-enhanced Corona, Firefly, and Flexishare PNoCs using a cycle-accurate SystemC based NoC simulator, for a 256-core single-chip architecture at 22nm. We validated the simulator in terms of power dissipation and energy consumption based on the results obtained from the DSENT tool [75]. We used real-world traffic from applications in the PARSEC benchmark suite [76]. GEM5 full-system simulation [77] of parallelized PARSEC applications was used to generate traces that were fed into our cycle-accurate NoC simulator. We set a “warmup” period of 100 million instructions and then captured traces for the subsequent 1 billion instructions. These traces are

extracted from parallel regions of execution of PARSEC benchmark applications. We performed geometric calculations for a 20mm×20mm chip size, to determine lengths of MWSR, SWMR, and MWMR waveguides in the Corona, Firefly, and Flexishare PNoCs. Based on this analysis, we estimated the time needed for light to travel from the first to the last node as 8 cycles at 5 GHz clock frequency. We use a 512-bit packet size, as advocated in the Corona, Firefly, and Flexishare PNoCs.

The static and dynamic energy consumption of electrical routers and concentrators in Corona, Firefly, and Flexishare PNoCs is based on results from the open source DSENT tool [75]. We model and consider area, power, and performance overheads for our framework implemented with the Corona, Firefly, and Flexishare PNoCs, as follows. *HYDRA* with Corona, Firefly, and Flexishare PNoCs has an electrical area overhead estimated to be 6.4 mm^2, 12.7 mm^2, and 3.4 mm^2 respectively and power overhead of 0.23 W, 0.44 W, and 0.36 W respectively, using gate-level analysis and the CACTI 6.5 [78] tool for memory and buffers. The photonic area overhead of Corona, Firefly, and Flexishare architecture is 9.63 mm^2, 19.83 mm^2, and 5.2 mm^2 respectively, based on the physical dimensions [72] of their waveguides, MRs, and splitters. For energy consumption of photonic devices, we adapt model parameters from recent work [48], [79], [80], with 0.42pJ/bit for every modulation and detection event and 0.18pJ/bit for the driver circuits of modulators and photodetectors. The MR trimming power is set to 130μW/nm [21] for current injection (blue shift) and tuning power is set to 240μW/nm [53] for heating (red shift).

2.8.2. WORST-CASE OSNR COMPARISON FOR VARIOUS PNOCS

Our first set of simulation results compares the baseline (without any crosstalk-mitigating enhancements) Corona, Firefly and Flexishare PNoCs with four variants of these architectures corresponding to three crosstalk-mitigation strategies from prior work (PCTM5B and PCTM6B

from [61], PICO from [56]) and our proposed *HYDRA* framework from this chapter. PCTM5B and PCTM6B are encoding schemes that replace each 4-bits of a data word with 5-bit and 6-bit code words respectively. These schemes aim to reduce photonic signal-strength of immediate non-resonant wavelengths (adjacent wavelengths in DWDM) to decrease crosstalk and improve OSNR in MR detectors. PICO is a process-variation aware crosstalk mitigation mechanism which also encodes data to reduce photonic signal-strength of immediate non-resonant wavelengths based on the process variation profile of the receiving MR detectors.

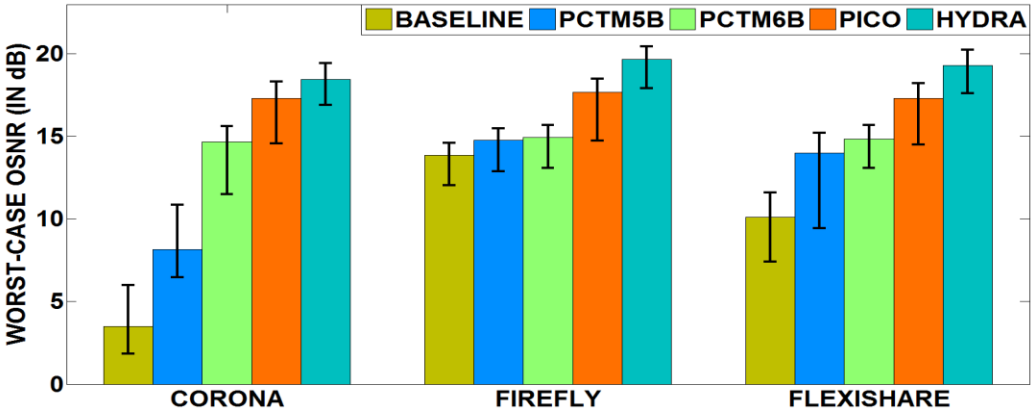


Figure 14: Worst-case OSNR comparison of *HYDRA* with PCTM5B [60], PCTM6B [60], and PICO [56] for Corona, Firefly, and Flexishare PNoCs. Bars show mean values of OSNR across 100 PV maps; confidence intervals show variation in OSNR values.

Utilizing the models presented in Sections 2.3.3 and 2.8.1, we calculate the received crosstalk noise and OSNR at detectors for the node with worst-case power loss (N_{WCPL}), which correspond to MR detectors in cluster 64 for the Corona PNoC, MR detectors of router C4R0 for the Firefly PNoC, and MR detectors of node R_{63} for the Flexishare PNoC. Figure 14 summarizes the worst-case OSNR results for the baseline, PCTM5B, PCTM6B, PICO, and *HYDRA* configurations of the three PNoC architectures considered. From the figure, it can be observed that Corona PNoC with *HYDRA* has 5.3×, 2.26×, 1.25×, and 1.06×, Firefly PNoC with *HYDRA* has 1.42×, 1.33×, 1.32×, and 1.13×, and Flexishare PNoC with *HYDRA* has 1.96×, 1.41×, 1.33×, and

1.14× worst-case OSNR improvements on average, compared to the baseline and PCTM5B, PCTM6B, and PICO enhanced variants of these architectures respectively.

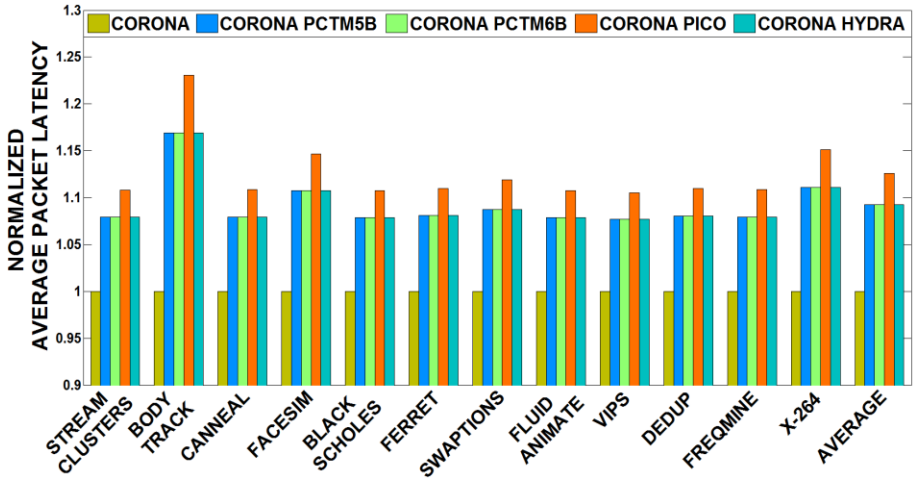
Both PCTM5B and PCTM6B eliminate occurrences of '111' in a data word and have limited occurrences of '11', which helps to reduce crosstalk noise in the detectors. But these techniques are unable to eliminate all occurrences of '11', because of which these techniques are unable to achieve higher reduction in crosstalk noise and significant improvement in OSNR. PICO considers the PV-profile of detecting nodes and performs encoding on specific wavelengths where there is high signal loss due to trimming to reduce crosstalk noise and improve OSNR in PNoCs. But even with the PICO technique, there still exist occurrences of '111' and '11', because of which OSNR gains with PICO are on the lower side. In contrast, *HYDRA* virtually eliminates all of the occurrences of '111' and '11' from the data word by combining benefits from *IMCM* and *DMCM*, and using *EDCM*'s 5-bit encoding mechanism. Although *EDCM*'s 5-bit encoding still results in limited occurrences of '111' and '11' in a data word, the DMRs of *DMCM* reduce the impact of consecutive '1's in the data word by removing crosstalk noise generated by these '1's in detector MRs. Thus *HYDRA* demonstrates higher OSNR gains compared to the best known previously proposed techniques. Furthermore, the OSNR values achieved with HYDRA (see Figure 14) are sufficient to enable reliable data transfers in PNoCs such as Corona, Firefly, and Flexishare.

2.8.3. OVERHEAD ANALYSIS OF HYDRA WITH VARIOUS PNOCS

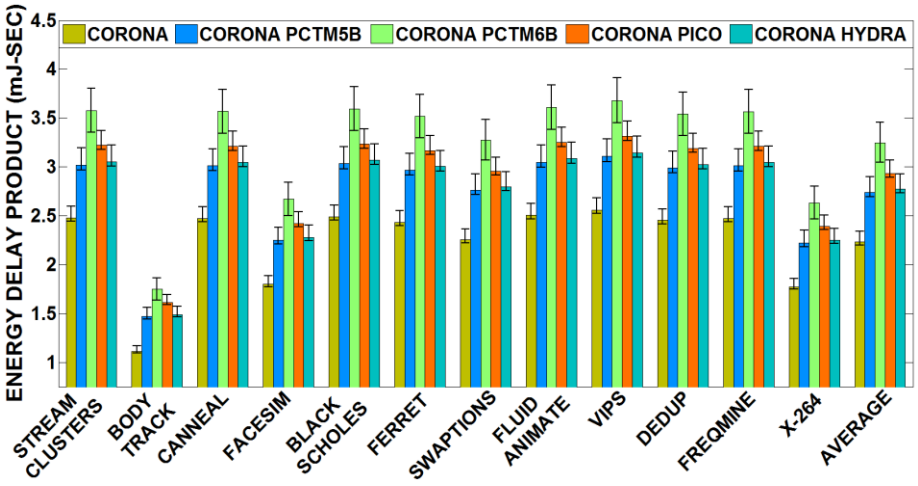
Our last set of results quantify the overhead for the proposed *HYDRA* framework and other techniques when used with the Corona, Firefly, and Flexishare PNoCs. Figure 15(a) and Figure 15(b) present detailed simulation results that quantify the average network packet latency and energy-delay product (EDP) for five Corona configurations. Results are shown for 12 multi-threaded PARSEC benchmarks. From Figure 15(a) it can be seen that on average, Corona with

HYDRA has 9.24% higher latency compared to the baseline. The additional delay due to encoding and decoding of data with *HYDRA*, PCTM5B, PCTM6B, and PICO contributes to their increase in average latency. The penalty due to encoding/decoding is approximately 1 cycle in PCTM5B, PCTM6B, and *HYDRA*. Thus *HYDRA* has a similar overhead compared to PCTM5B and PCTM6B. However, *PICO* has a 2 cycle penalty, which increases its delay compared to *HYDRA* by 3.1%. Note that for the chosen clock frequency, PV in photonic components does not change the number of clock cycles for various operations, such as encoding/decoding, modulation/detection etc., therefore Figure 15(a) does not have confidence intervals or variations in packet latency due to PV.

From the results for EDP shown in Figure 15(b), it can be seen that on average, the Corona configuration with our *HYDRA* framework has 24.3% higher EDP compared to the baseline. The increase in EDP for Corona with *HYDRA* is not only due to the increase in average latency, but also due to the addition of extra bits for encoding and decoding, which leads to an increase in the amount of photonic hardware in the architectures (more number of MRs, complex splitters). This in turn increases static power dissipation. Dynamic power also increases in these architectures, but by much less amount. However, EDP for Corona with *HYDRA* is 17.1% and 5.7% lower compared to PCTM6B and PICO respectively. The higher latency of PICO compared to *HYDRA* increases its EDP, whereas *HYDRA* has lower EDP compared to PCTM6B because *HYDRA* conserves laser and MR trimming/tuning power due to a lower photonic hardware footprint compared to PCTM6B. The EDP for Corona with *HYDRA* is 1.3% higher compared to PCTM5B. Although PCTM5B and *HYDRA* have similar average latency, the increase in number of MRs in *HYDRA* due to the presence of *IMCM* MRs and DMRs increases its laser and trimming/tuning power, which in turn increases its EDP.



(a)



(b)

Figure 15: (a) Normalized average latency, (b) energy-delay product (EDP) comparison between Corona baseline and Corona configurations with PCTM5B, PCTM6B, PICO, and *HYDRA* techniques, for PARSEC benchmarks. Latency results are normalized to the baseline Corona results. In the EDP plot, bars represent mean values of EDP across 100 PV maps; confidence intervals show variation in EDP.

Figure 16(a) and Figure 16(b) summarize the average network packet latency and EDP results for the five configurations of Firefly and Flexishare PNoCs. Results are shown for twelve multi-threaded PARSEC benchmarks and are averaged across these benchmark applications, for brevity.

From Figure 16(a) it can be observed that on average, Firefly with *HYDRA* has 5.2% and Flexishare with *HYDRA* has 10.6% higher latency compared to their respective baselines. The additional delay due to encoding and decoding of data with *HYDRA* contributes to its increase in average latency over the respective baselines of Firefly and Flexishare PNoCs. The latency overhead for Firefly with HYDRA is lower compared to Corona and Flexishare with *HYDRA*. This is because Firefly is a hybrid PNoC where some portion of data traverses through electrical links. This data over electrical links is unaffected by the extra encoding/decoding delays in *HYDRA*, whereas in Corona and Flexishare the entire traffic traverses through photonic waveguides. Much like Corona (Figure 15(a)), the Firefly and Flexishare architectures with *HYDRA* have similar latency values compared to these architectures with PCTM5B and PCTM6B (Figure 16(a)). Furthermore, Firefly with *HYDRA* has 2.7% and Flexishare with *HYDRA* has 3.2% lower latency compared to PICO. Reduction in number of encoding or decoding cycles from 2 to 1 from PICO to *HYDRA* reduces average latency of HYDRA.

From the results for EDP shown in Figure 16(b), it can be seen that on average, the Firefly and Flexishare configurations with our *HYDRA* framework have 5% and 22% higher EDP compared to their respective baselines. EDP overhead for Firefly is relatively lower compared to the Corona and Flexishare architectures because of its lower latency overheads and smaller increase in laser/trimming power due to lesser increase in the amount of photonic hardware. Firefly with *HYDRA* has 3.1% and 2.4%, and Flexishare with *HYDRA* has 5.9% and 2.4% lower EDP compared to the respective architecture configurations with PCTM6B and PICO. Additionally, compared to Firefly and Flexishare configurations with PCTM5B, the configurations of the same architectures with *HYDRA* framework have 0.6% and 1.5% higher EDP respectively.

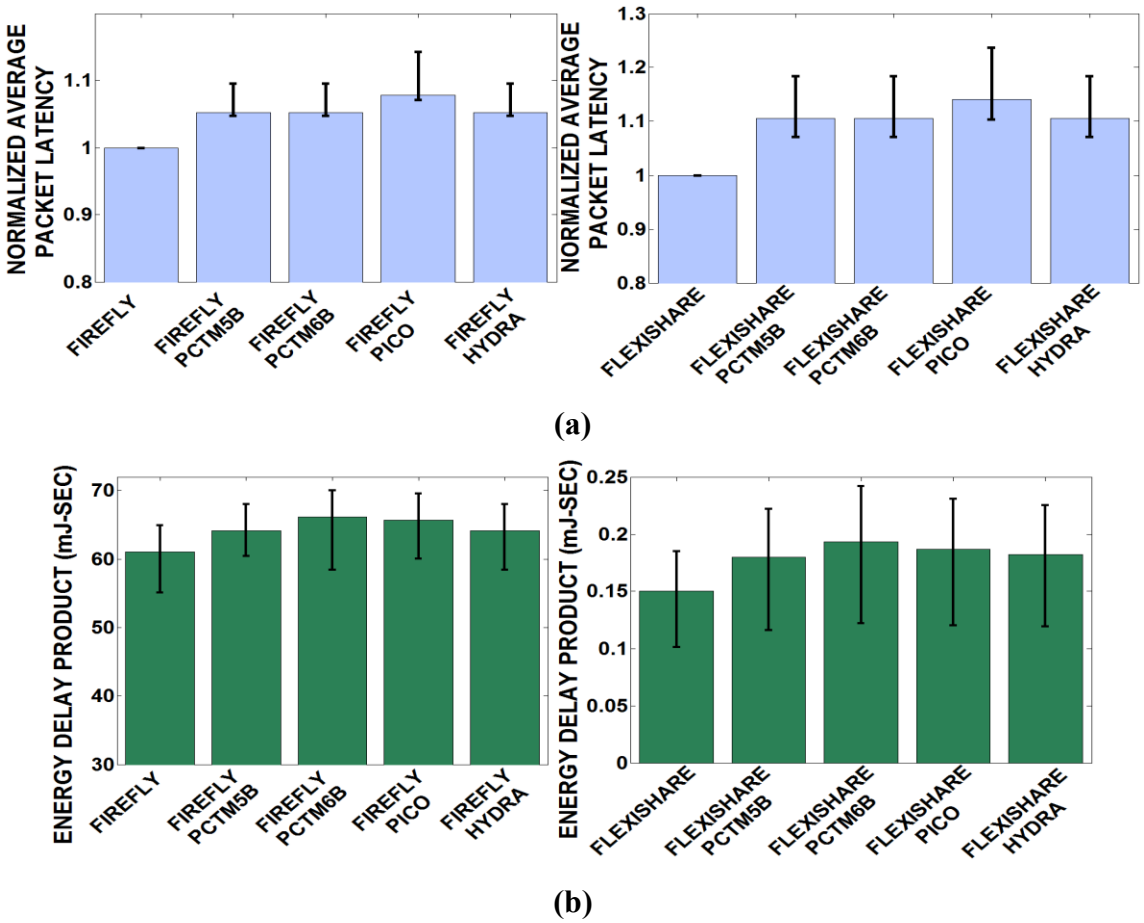


Figure 16: (a) Normalized average latency, (b) energy-delay product (EDP) comparison between variants of Firefly and Flexishare PNoCs, which include their baselines and their variants with PCTM5B, PCTM6B, PICO, and *HYDRA* techniques, for PARSEC benchmark applications. Latency results are normalized with their respective baseline architecture results. Bars represent mean values of latency and EDP for 100 PV maps; confidence intervals show variation in latency and EDP across PARSEC benchmarks.

2.9. CONCLUSIONS

In this chapter, we presented a novel cross-layer crosstalk mitigation framework for the reduction of crosstalk noise in the detectors of DWDM-based PNoC architectures. Our proposed HYDRA framework seamlessly integrates two device-layer techniques and one circuit layer technique to enable interesting trade-offs between reliability, performance, and energy overheads for the Corona, Firefly, and Flexishare crossbar-based PNoC architectures. Our simulation based analysis showed that the HYDRA framework improves worst-case OSNR by up to 5.3× compared

to the baseline architectures, and by up to 1.14× compared to the best known PNoC crosstalk mitigation scheme from prior work. Thus, HYDRA is an attractive solution to enhance reliability in emerging DWDM-based PNoCs.

3. MITIGATION OF HOMODYNE CROSSTALK NOISE IN SILICON PHOTONIC NOC ARCHITECTURES WITH TUNABLE DECOULING

Photonic network-on-chip (PNoC) architectures employ photonic waveguides with dense-wavelength-division-multiplexing (DWDM) for signal traversal and microring resonators (MRs) for signal modulation, to enable high bandwidth on-chip transfers. Unfortunately, due to the resonant nature of MRs, the power built-up in their cavity gradually recouples back into the photonic wave-guides. This recoupled power induces time-dependent unfilterable homodyne crosstalk noise, when the wavelength of the recoupled power matches with the wavelength of a signal in the waveguide. The homodyne crosstalk in turn deteriorates the signal-to-noise ratio (SNR) and on-chip communication reliability. This chapter presents a novel lightweight technique to mitigate homodyne crosstalk noise in DWDM-based PNoCs. We evaluate the effectiveness and overhead of our technique by implementing it for well-known PNoC architectures, including Corona, Firefly and Flexishare. Experimental results indicate that our approach when implemented on these PNoCs can improve the worst-case SNR by up to 37.6% compared to the baseline versions of these PNoCs, thereby significantly enhancing reliability, at the cost of up to 19.2% energy overhead and 1.7% photonic area overhead.

3.1. INTRODUCTION

Recent developments in silicon photonics have enabled the integration of photonic components with CMOS circuits on a chip. Several photonic network-on-chip (PNoC) architectures have been proposed to date, e.g., [14]-[16], and [81]. These PNoCs employ multiple on-chip photonic links that use microring resonator (MR) modulators to convert (i.e., modulate) electrical signals to photonic signals that travel through a photonic waveguide, and MR filter

detectors that detect and drop photonic signals on to a photodetector to recover an electrical signal. Each MR has a unique set of (resonance) wavelengths that it can couple to and work correctly with. Typically, photonic waveguides are designed to support dense wavelength division multiplexing (DWDM), where different wavelengths are multiplexed in the waveguide. The use of multiple MRs that are in resonance with these wave-lengths enables high-bandwidth parallel transfers in photonic waveguides (links).

One of the key challenges for the widespread adoption of DWDM-based PNoC architectures is to minimize the crosstalk noise that emanates from their MRs. In a DWDM-based PNoC, crosstalk occurs when, due to the non-idealities in the resonance of MRs, some optical noise power at an unwanted phase/wavelength mixes with signal power. Crosstalk is an intrinsic property of every MR, so both MR modulators and MR detectors are susceptible to it. The crosstalk noise of MRs negatively affects signal-to-noise ratio (SNR) in waveguides. For instance, the damaging impact of cross-talk noise in the Corona PNoC is presented in [48], where worst-case SNR is estimated to be 14dB, which is insufficient for reliable data transfers, as its corresponding bit-error-rate (BER) is very high, in the order of 10^{-3}.

Crosstalk noise can be classified broadly as homodyne or heterodyne [82]. The homodyne crosstalk noise power of a particular wavelength affects the signal power of the same wavelength, whereas in case of heterodyne crosstalk, the signal power gets affected by some noise power of one or more other different wavelengths. Both homodyne and heterodyne crosstalk noise may affect signal integrity in a waveguide. Several techniques have been proposed in prior works, e.g., [61], [82], and [83], to mitigate heterodyne crosstalk noise in PNoCs. In general, the sources and effects of heterodyne crosstalk noise in PNoCs are predictable and can be easily controlled by filtering. This fact makes mitigation of hetero-dyne crosstalk in PNoCs relatively less critical. In

contrast, homo-dyne crosstalk noise cannot be easily filtered out, as both the noise power and signal power are of the same wavelength. Moreover, the effect of homodyne crosstalk noise on signal integrity depends on the phase of the noise relative to the signal, which makes its analysis and mitigation more complex. Although, the homodyne crosstalk phenomenon is well understood in relation to WDM long-haul optical networks [84]-[86], it is relatively less explored and understood in relation to PNoCs. Only two prior works analyze the effects of homodyne crosstalk on SNR of PNoCs [48], [82]. However, these works neither explain causes of homodyne cross-talk at the device-level, nor present any technique to mitigate it.

Main Contributions: In this chapter, we first demonstrate that due to the resonant nature of MRs the circulating power in MRs builds up with time and ultimately recouples back in the waveguide to induce homodyne crosstalk noise. Then, we present a detailed analysis of the characteristics of circulating power in MRs, before presenting a simple, low overhead device-level solution to mitigate the effects of homodyne crosstalk on signal integrity. Our technique is easily implementable in any existing DWDM-based PNoC without requiring major modifications to the architecture. To the best of our knowledge, this is the first work that attempts to mitigate homodyne crosstalk noise in PNoCs. Our novel contributions in this work are summarized below:

- We demonstrate how the circulating power of an MR, built up due to the resonant nature of the MR, induces homodyne cross-talk noise in the waveguides of a PNoC;
- We perform in-depth analysis and characterization of time-dependent characteristics of circulating power in MRs to under-stand the behavior of the induced homodyne crosstalk;
- Based on this analysis, we propose a low-overhead homodyne crosstalk mitigation (HCTM) technique to control the emanation of homodyne crosstalk noise from MRs by tuning the decoupling of circulating power from the MRs;

- We evaluate the effectiveness and overhead of our HCTM technique by implementing it for well-known PNoC architectures, including Corona [14], Firefly [15], and Flexishare [16], running real-world multi-threaded PARSEC benchmarks.

3.2. BACKGROUND AND RELATED WORK

DWDM-based PNoCs utilize several photonic devices such as microring resonators (MRs) as modulators, switches/injectors, and detectors; photonic waveguides; splitters; and trans-impedance amplifiers (TIAs). The interested reader is directed to [61] for more detailed discussion on each of these components.

As described in [87], an important characteristic of photonic signal transmission in on-chip photonic waveguides is that it is inherently lossy, i.e., the light signal is subject to various types of losses such as through-loss in MR modulators and detectors, modulating losses in modulator MRs, detection loss in detector MRs, propagation and bending loss in waveguides, and splitting loss in splitters. Such losses reduce SNR in photonic waveguides. In addition to the photonic signal loss, crosstalk noise in MRs also deteriorates overall SNR. Both modulators and detectors are susceptible to crosstalk noise in DWDM-based PNoCs. The degraded SNR deteriorates bit-error-rate (BER), jeopardizing the reliability of on-chip data transfers.

As mentioned earlier, crosstalk noise can be classified as homodyne or heterodyne [82]. In case of heterodyne crosstalk, as the noise power and signal power have different wavelengths, the noise power always deteriorates the signal integrity and BER irrespective of its relative phase to the signal. The strength of the heterodyne crosstalk noise at a detector/modulator MR depends on the following four attributes: *(i)* channel gap between the MR resonant wavelength and the adjacent wavelengths, *(ii)* Q-factors of neighboring MRs, *(iii)* data-rate at which the MRs operate [49], and *(iv)* relative data occurrences in neighboring wavelength-channels [61]. With increase in

DWDM, the channel gap between two adjacent wavelengths decreases, which in turn increases heterodyne crosstalk in MRs. With decrease in Q-factors of MRs, the widths of the resonant passbands of MRs increases. Due to the wider passbands of MRs, passband overlap among neighboring MRs increases, which in turn also increases heterodyne crosstalk.

Several techniques have been proposed in prior works, e.g., [60], [83], and [88]-[92], to mitigate heterodyne crosstalk noise in PNoCs. Among them, in [88] and [89], the authors use MR based high-order filters to filter out heterodyne noise from the signal before detection. Xie et al. in [83] analyze the worst-case crosstalk noise and SNR in mesh-based PNoCs, and propose changes in the design of the photonic router to reduce worst-case crosstalk noise. In [61], Chittamuru et al. propose an encoding based approach that reduces undesirable data value occurrences in photonic waveguides to reduce the passband overlap among neighboring MRs, thereby reducing heterodyne crosstalk noise. Some other prior works also explore heterodyne crosstalk in PNoCs, e.g., [90]-[92], but each of them presents detailed analysis of heterodyne crosstalk noise for a specific PNoC architecture without aiming to mitigate the noise.

For homodyne crosstalk, as the noise power and signal power have the same wavelength, its effects on signal integrity depend on the phase of the noise power. If the homodyne crosstalk noise is in phase with the signal power, the noise power increases the signal power, increasing SNR. In contrast, if the homodyne crosstalk noise is out of phase with the signal power, it deteriorates the signal, *decreasing* SNR. Out-of-phase homodyne crosstalk noise, whether coherent (phase-correlated) or incoherent (phase-uncorrelated) to the signal, always degrades signal integrity [48].

The homodyne crosstalk phenomenon in relation to WDM long-haul optical networks is reported in several prior works [84]-[86]. Y. Shen et al. in [86] develop analytical expressions that captures the impact of homodyne crosstalk noise (coherent and incoherent) on an optical signal

passing through optical cross-connect nodes in WDM long-haul optical networks. They do not present any noise mitigation technique. Homodyne crosstalk noise has been shown to originate in multiplexer/demultiplexer units and cross-connect nodes of long-haul optical networks, and can be mitigated either by using phase scrambling at the receiver end (as discussed in [85]) or by intelligent management of the optical network (as discussed in [84]). The homodyne crosstalk mitigation techniques discussed in [84] and [85] are specific to the origins of crosstalk noise in and the structure and architecture of the underlying long-haul optical networks. Therefore, these techniques cannot be generalized for use at the chip level to mitigate the homodyne crosstalk noise in PNoC architectures.

As mentioned earlier, homodyne crosstalk is relatively less explored and understood at the chip level and in relation to PNoC architectures. Only two prior works (i.e., [48] and [82]) have shown homodyne crosstalk noise to originate from MRs of PNoCs and have analyzed the effects of homodyne crosstalk on the SNR of PNoCs. Nikdast et al. in [82] analyze the worst-case homodyne incoherent crosstalk noise and resulting SNR in arbitrary fat-tree-based PNoCs. Duong et al. [48] analyze homodyne coherent crosstalk noise in ring-based DWDM PNoCs. As demonstrated in [48], the worst-case SNR due to homodyne coherent noise in the data waveguide of Corona PNoC is estimated to be 14dB, which corresponds to BER of 10^{-3}. Now, as explained in [83], the value of BER should be 10^{-9} or less for reliable communication, which implies that the BER of 10^{-3} obtained in [48] for the Corona PNoC can severely harm the communication reliability. This implies that, in PNoCs, homodyne noise can be severe and should be significantly mitigated to ensure reliable communication. But none of the prior works [48] and [82] aims to mitigate homodyne crosstalk noise.

Now, one way of mitigating the reliability concerns related to the homodyne crosstalk noise is using error-correcting codes (ECCs). However, ECCs cannot eliminate all the faults or errors in a PNoC, when the BER is very high in the order of 10^{-3} or higher. Moreover, ECCs, whether implemented in software or hardware, consume high energy and incur significantly high-performance overhead. In addition to this high overhead, ECCs do not guarantee elimination of all errors. Thus, the effectiveness of ECCs in mitigating the homodyne crosstalk noise is questionable. On the other hand, any higher-level homodyne noise mitigation technique, other than ECCs, depends on the system's ability to differentiate between the noise power and the signal power. But as the wavelength of homodyne noise power is same as that of the signal power, it is very difficult to differentiate or filter the noise power from the signal power at any higher level. Therefore, the effectiveness of such higher-level techniques is also questionable.

In contrast to the ECCs and other high-level techniques, in this chapter, we present a low-overhead device-level technique called HCTM to mitigate homodyne crosstalk. Our HCTM technique extenuates the root cause of homodyne crosstalk noise, i.e., the circulating power in MRs. This enables HCTM to eliminate all the errors incurred due to the homodyne crosstalk noise. Thus, it can be implied that our HCTM technique is better than ECCs and any such higher-level technique in ensuring the reliability of communication in PNoCs.

3.3. HOMODYNE CROSSTALK: CAUSE AND EFFECTS

In Section 3.3.1, we present the general properties of MRs. Then we present detailed analysis of spectral and temporal characteristics of MRs in Section 3.3.2 to understand homodyne crosstalk in PNoCs. Finally, in Section 3.3.3, we discuss modeling of homodyne crosstalk in PNoCs.

3.3.1. GENERAL PROPERTIES OF MICRORING RESONATORS

Before going into the specifics of homodyne crosstalk noise, we first discuss the general properties of MRs. Optical MRs are extensively described in literature [64], [93]. As shown in Figure 17(a), an MR consists of a looped optical waveguide coupled to a straight bus waveguide (BWG). The straight BWG acts as a coupling mechanism to access the loop. The looped waveguide of the MR could be circular in shape (Figure 17(a)) or elongated (Figure 17(b)). The elongated MR with a straight section along one direction (typically along the coupling section) is called *racetrack* MR. The elongated shape of the racetrack MR renders better control of the coupled optical power [93]. Otherwise, the racetrack MR is functionally and characteristically similar to the circular MR. Therefore, the analysis presented in this section is equally applicable to both types of MRs. Moreover, the analytical models presented in this section and subsequent sections hold true for both types of MRs and can be used to capture the crosstalk noise characteristics of both types of MRs.

As shown in Figure 17(a), the BWG that is coupled to the MR has two ports termed as *input* and *pass*. The coupled optical field (E_C), the circulating optical field (E_{circ}), and the output optical field (E_{pass}) in the figure depend on the input optical field (E_{in}) through the field transmission coefficient (t) and field-coupling coefficient (κ). The relationship among these optical fields can be given by the following equations, which are derived from the transfer matrix model described in [93]:

$$E_c = j\kappa E_{in} + tE_{circ}\,, \tag{24}$$

$$E_{pass} = E_{tr} + E_N = tE_{in} + j\kappa E_{circ}, \tag{25}$$

$$E_{circ} = ae^{j\varphi}E_c, \tag{26}$$

Here, j = $\sqrt{-1}$, which represents π phase shift of the ideal MR to BWG coupler, a is the roundtrip amplitude transmission coefficient, including both propagation loss in the ring and loss in the couplers, and φ is roundtrip phase shift of the field. The coefficients t and κ are such that t^2 and κ^2 are the power splitting ratios of the coupler, and they are assumed to satisfy $t^2+\kappa^2=1$, which means there are no extra losses in the coupling section. As evident from Eq. (24)-(26), E_{circ} provides a positive feedback for E_C, which ultimately affects E_{circ} and E_{pass}. As a result, in the presence of non-zero E_{in}, E_{circ} grows with every roundtrip due to increasing E_C, resulting in an increase of circulating optical power (P_{circ}) with time. From Eq. (25), a part of E_{circ} recouples back in the BWG from the MR as a noise field ($E_N = j\kappa E_{circ}$), which adds up with the transmitted field ($E_{tr}=tE_{in}$) to give E_{pass} as the output field. In fact, E_N is the homodyne noise field, as its wavelength is the same as the wavelength of E_{tr} and E_{in}.

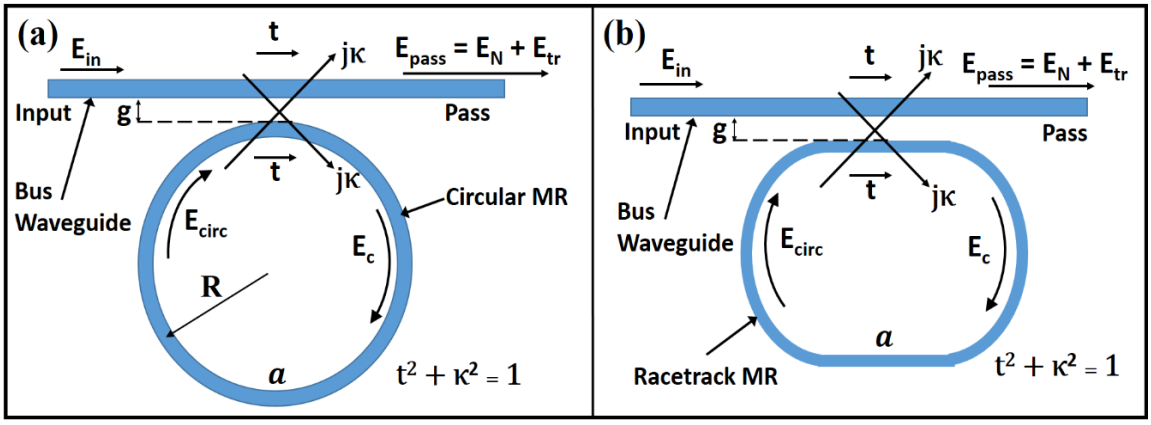


Figure 17: Illustrations of waveguide coupled photonic microring resonators (MRs): (a) circular MR, (b) racetrack MR.

For a given cross-section of the BWG and the MR's loop waveguide, the optical power travelling in these waveguides is proportional to the square of the traversing optical field [64]. This implies that the relationship between the coupled optical power (P_C), circulating optical power

(P_{circ}), and output optical power (P_{pass}) to the input optical power (P_{in}) can be derived from Eq. (1)-(3), which is given in the following equations [63]:

$$B = \frac{P_{circ}}{P_{in}} = \left|\frac{E_{circ}}{E_{in}}\right|^2 = \frac{\kappa^2 a^2}{1 - 2at\cos\varphi + a^2 t^2}, \tag{27}$$

$$P_{pass} = P_{tr} + P_N = \frac{a^2 - 2at\cos\varphi + t^2}{1 - 2at\cos\varphi + a^2 t^2} P_{in}, \tag{28}$$

$$P_N = -\left(\frac{2at\kappa^2 e^{j\varphi}}{1 - ate^{j\varphi}} + B\kappa^2\right) P_{in}, \tag{29}$$

$$P_{tr} = t^2 P_{in}, \tag{30}$$

Here, B is the buildup factor, which is the ratio of P_{circ} to P_{in} (Eq. (27)). Eq. (28) gives the value of P_{pass} in terms of t, a, and φ. This equation can be broken down in Eq. (29) and Eq. (30) to show two components of P_{pass}: P_N (noise power) and P_{tr} (transmitted signal power). Note that P_N (Eq. (29)) is homodyne noise power, as its wavelength is the same as that of P_{tr} and P_{in}.

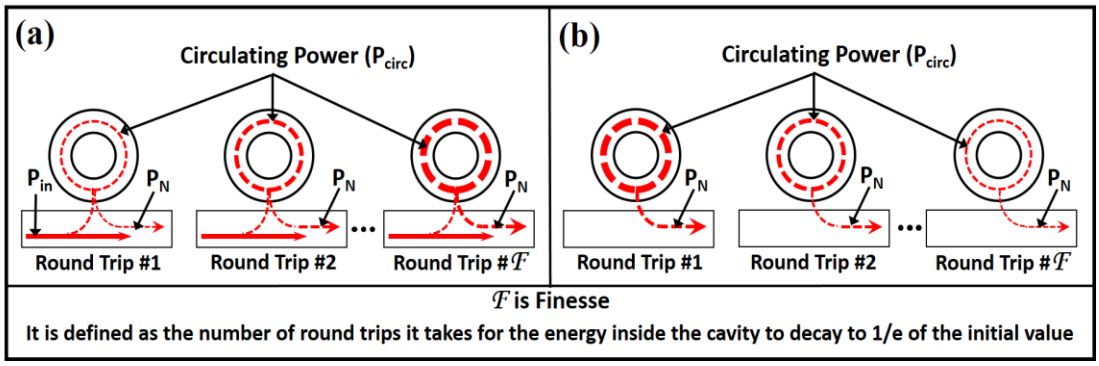


Figure 18: Illustration of gradual increase and decrease of power circulating in MR: (a) when $P_{in}\neq 0$, (b) when $P_{in}=0$.

Note that the noise characteristics of MRs that are captured by the models presented in Eq. (24)-(30) do not differ for circular and racetrack MRs. Only the amount of noise power in MRs

differs between the two types of MRs. It is because only the values of coefficients *a*, *t* and κ differ between circular and racetrack MRs.

In summary, for a non-zero input power (P_{in}), the circulating power (P_{circ}) in the MR builds up by the factor B at steady state, which ultimately contributes to the homodyne noise power (P_N) in the BWG (factor $B\kappa^2$ in Eq. (29)).

3.3.2. SPECTRAL AND TEMPORAL CHARACTERISTICS OF P_N

As evident from Eq. (29), the homodyne noise power (P_N) depends on roundtrip phase shift φ. The value of φ changes periodically with wavelength of operation, which implies that use of Eq. (28), (29) requires determination of φ that is equal to $2\pi L n_{eff}/\lambda$, where L is the roundtrip length of MR and λ is operating wavelength. This requires complicated calculation of n_{eff} that depends on MR geometry and materials. A more commonly used equation for the power transmission (and hence for the noise power) on and off resonance of an MR is the Lorentzian transfer function [94]:

$$T_N = \frac{\left(P_N^{\varphi=2\pi}/P_{in}\right)}{1+\left(2Q(\Delta\lambda/\lambda_r)\right)^2}, \tag{31}$$

where, Q is loaded quality factor of the MR, λ_r is MR's resonance wavelength, and $\Delta\lambda$ is the difference between the operating wavelength and λ_r. Note that, for $\Delta\lambda$=0, Eq. (31) gives normalized homodyne crosstalk noise (normalized to P_{in}) for on-resonance MR operations, whereas for $\Delta\lambda\neq0$, Eq. (31) gives normalized homodyne crosstalk noise for off-resonance MR operations.

Note that, at the *critical coupling condition* (when $t=a$ in Eq. (28)), whether on-resonance or off-resonance, P_{pass} is minimized, but the effect of P_N on P_{tr} and hence on P_{pass} is maximized. This implies that the critical coupling condition does not alleviate P_N, even though it minimizes P_{pass}.

In our analysis so far, we have considered only the spectral properties of P_N, which characterizes the behavior of P_N subject to various wavelengths of light. As evident from Eq. (29), in addition to the parameters φ, t, a, and κ, the noise power P_N also depends on the P_{circ} build-up factor B and input power P_{in}. Parameters B and P_{in} not only have spectral properties (such as operational wavelength, field phase φ), but also have temporal properties, i.e., B and P_{in} may vary with time. For instance, in the PNoC paradigm, the value of P_{in} typically keeps flipping between logic '1' and logic '0' with time.

As mentioned earlier, in the presence of a non-zero P_{in}, P_{circ} in an MR gradually increases with every roundtrip (and hence with time) and builds up by the factor B at steady state. As P_N depends on P_{in} and P_{circ} ($B\kappa^2 P_{in}$ - Eq. (29)), P_N also follows the same trend as P_{circ}. This fact is depicted in Fig. 2(a), which illustrates how P_{circ} and P_N grow with every roundtrip in the presence of a non-zero P_{in}. If P_{in} drops to zero while P_{circ} is growing, then P_{circ} starts decaying with time from whatever value it had grown to before P_{in} dropped to zero. This fact is depicted in Figure 18(b), which illustrates how P_{circ} and P_N decays with every roundtrip when P_{in}=0. The steady state value of P_{circ} (and hence of P_N) and the rate with which P_N grows/decays every round trip (or with time) depends on the cavity photon lifetime (τ_p) and Q-factor of the MR through a natural exponential function. The growth/decay of P_N as a function of time is expressed as [95]:

$$P_N(t) = \begin{cases} T_N\left(1 - e^{-\left(t/\tau_p\right)}\right), & P_{in} \neq 0 \\ T_N^{t=0}\left(e^{-\left(t/\tau_p\right)}\right), & P_{in} = 0 \end{cases}, \tag{32}$$

$$\tau_p = Q\lambda/2\pi c, \tag{33}$$

where, T_N can be obtained from Eq. (31). When $P_{in} \neq 0$, P_N grows exponentially until it reaches the steady state value T_N in τ_p time. When P_{in}=0, P_N decays exponentially starting from its initial

value of $T_N^{t=0}$. As the value of P_{in} typically keeps flipping between logic '1' and logic '0' with time in a typical PNoC waveguide, it implies that P_{in} consists of a sequence of falling and rising edges. Therefore, it can be inferred that the value of P_N at any given time depends on the instantaneous value of P_{in} and on the amount of time that has elapsed from the last falling or rising edge of P_{in}. Moreover, the time between two successive falling/rising edges of P_{in} depend on the bit-period and the bit-pattern of P_{in}, which implies that P_N also depends on bit-period (bit-rate) and bit-pattern of P_{in}.

In summary, the homodyne crosstalk noise (P_N) originating from MRs varies with field phase φ, which in turn depends on the operating wavelength. P_N also varies with time, and the time-dependent variation profile of P_N depends on the bit-period (and hence the bit-rate) and bit-pattern of the input signal (P_{in}) in the photonic waveguide, along with some time-independent properties of MRs such as photon lifetime (τ_p) and Q. The time-varying homodyne noise (P_N), when coupled in the BWG, gets mixed up with the time-varying transmitted signal power (P_{tr}) causing random fluctuations in the amplitude of the output signal power (P_{pass}). These random fluctuations in P_{pass} reduce the optical eye opening, which in turn increases the BER and reduces the communication reliability in PNoCs.

3.3.3. MODELING OF HOMODYNE NOISE (P_N) IN PNOCS

Our ultimate goal in this chapter is to analyze and mitigate the homodyne crosstalk noise in PNoC architectures. As mentioned earlier, PNoCs are typically designed to support DWDM, where a large number of wavelengths are multiplexed in a single BWG. The group of multiple MRs that are in resonance with these DWDM wavelengths and that are "arrayed" along the BWG can be referred to as an *MR bank*. This MR bank can be a modulator bank or detector bank. In a typical PNoC, the BWG passes through multiple nodes, and a pair of such modulator and detector

MR banks is used at every node, which enables high bandwidth parallel data transfers in the BWG. Thus, modulator banks, detector banks, and BWGs are basic building blocks of a DWDM PNoC. Before we extend our homodyne noise model (discussed in previous subsections) to the entire PNoC, it is important to first extend it to these basic building blocks. Our aim is to reduce the worst case P_N in a PNoC, thus we model only the worst-case value of P_N in MR banks, BWG, and the PNoC. Note that the worst-case P_N occurs at steady state ($t=\tau_p$ in Eq. (32)), thus, the value of worst-case P_N does not change with time.

Modeling of P_N in MR Modulator and MR Detector Banks: An MR modulator bank at a node along the BWG of a PNoC could be in either active state or inactive state. Similarly, an MR detector bank could be in either active state or inactive state. All MRs in an active modulator bank become ON and OFF resonance with their corresponding wavelengths with time to modulate a particular bit pattern on their corresponding wavelengths. All MRs in an inactive modulator bank and in an inactive detector bank are OFF resonance with their corresponding wavelengths, whereas the MRs of an active detector bank are ON resonance with their corresponding wavelengths. The worst-case homodyne crosstalk noise for a DWDM wavelength emanating from a modulator or a detector MR bank can be modeled by extending Eq. (31) as follows:

$$P_N^{Bank}(\lambda_j) = \sum_{i=0}^{n-1} \frac{P_N^{\varphi=2\pi}/P_{in}}{1 + \left(2Q\{|\lambda_i - \lambda_j|/\lambda_i\}\right)^2}, \tag{34}$$

where, n is the number of MRs in the bank that is the same as the number of DWDM wavelengths. $P_N^{\varphi=2\pi}$ is calculated by putting $\varphi=2\pi$ in Eq. (29). λ_j is the wavelength of P_N. λ_i is the resonance wavelength of the i^{th} MR of the bank. In case of an active modulator bank or detector bank, for a given value of λ_j, there exists exactly one i (and corresponding MR) for which $\lambda_j = \lambda_i$.

In contrast, in case of inactive modulator bank or detecting bank, for a given value of λ_j, there exists no *i* for which $\lambda_j = \lambda_i$. Note that λ_j gets picked up by each MR in the bank, the extent of which depends on the distance $|\lambda_i - \lambda_j|$ of λ_j from the resonance wavelength (λ_i) of the corresponding MR. Therefore, optical power of wavelength λ_j builds up in each MR of the bank, which ultimately contributes to the homodyne crosstalk noise of wavelength λ_j. Thus, Eq. (34) models the total worst-case homodyne crosstalk noise emanating from all the MRs of the bank combined.

Modeling of P_N in DWDM Bus Waveguide (BWG): When a source node sends a data packet along a BWG of a PNoC to a destination node, the modulator bank of the source node remains active, the detector bank of the destination node remains active, and all other modulator and detector banks along the BWG remain inactive. In this scenario, all wavelengths of the DWDM spectrum are modulated by the source modulator bank and then they travel along the BWG from the source node to the destination node where they are detected by the destination detector bank. For a particular wavelength from the DWDM spectrum, the homodyne crosstalk noise at that wavelength originates from all the intermediate MR banks along the BWG that add up and contribute to the total noise power $P_N^{BWG}(\lambda_j)$ at the destination detector MR. The worst-case homodyne crosstalk noise at a wavelength in a DWDM bus waveguide between the source modulator bank and the destination detector bank can be assessed using the following equation:

$$P_N^{BWG}(\lambda_j) = \sum_{l=1}^{L_D+L_M} P_N^{Bank}(j) + \sum_{m=1}^{M_D+M_M} P_N^{Bank}(j), \tag{35}$$

Here, L_D and L_M are the numbers of active MR detector banks and active MR modulator banks respectively, whereas M_D and M_M are the numbers of inactive MR detector banks and inactive MR modulator banks, respectively.

3.4. MITIGATION OF HOMODYNE CROSSTLK

As implied from the conclusions of our analysis presented in the previous section and Eq. (29), the worst-case homodyne crosstalk noise P_N originating from a particular MR depends on the MR's power build-up factor B and power-coupling coefficient κ^2. As evident from Eq. (27), B is a function of the roundtrip power-transmission coefficient a^2. Therefore, P_N can be mitigated by reducing either κ^2 or B (hence a^2). However, reducing κ^2 increases t^2 (as $\kappa^2 = 1-t^2$), which in turn increases P_{tr} (Eq. (30)). The increase in P_{tr} reduces the extinction ratio [64], which increases the susceptibility of signal transmission (P_{pass}) to noise (P_N) as evident from Eq. (28). Therefore, reducing κ^2 is not a desirable option for mitigating P_N. Alternatively, reducing a^2 can be achieved by increasing the roundtrip power loss ($1 - a^2$) in an MR. The roundtrip loss in an MR is in general attributed to propagation loss and bending loss in an MR's looped waveguide, and coupling loss at the MR-BWG coupling region [64]. The propagation loss can be increased by increasing either the sidewall roughness of or the free carrier concentration in the MR's waveguide [96]. However, increasing free carriers would alter the resonance of the MR, and increasing sidewall roughness is difficult to control.

We propose a simple and elegant method of increasing loss in an MR to reduce homodyne P_N with more predictable control. Our homodyne crosstalk mitigation (HCTM) technique has two components. *First*, we propose to use racetrack MRs (Figure 19), which adds excess bend losses in MRs due to mismatch losses at the straight-bend transition of MRs [64]. *Second*, as shown in Figure 19, we introduce a secondary coupling waveguide for an MR that can decouple most of the circulating power (P_{circ}) out of the MR before it can recouple back in the BWG, thus increasing loss ($1 - a^2$) to reduce B, and noise power P_N (as per Eq. (27), (29)). Decoupling P_{circ} out of the MR greatly reduces P_N irrespective of the temporal variation profile of P_{in}. Thus, our HCTM

technique mitigates the homodyne noise in PNoC architectures by extenuating its root cause, i.e., P_{circ} in MRs. Therefore, it can be implied that HCTM technique is agnostic to the underlying PNoC architecture, as it reduces homodyne crosstalk noise in all types of PNoC architectures in the same way, i.e., by extenuating P_{circ} in MRs.

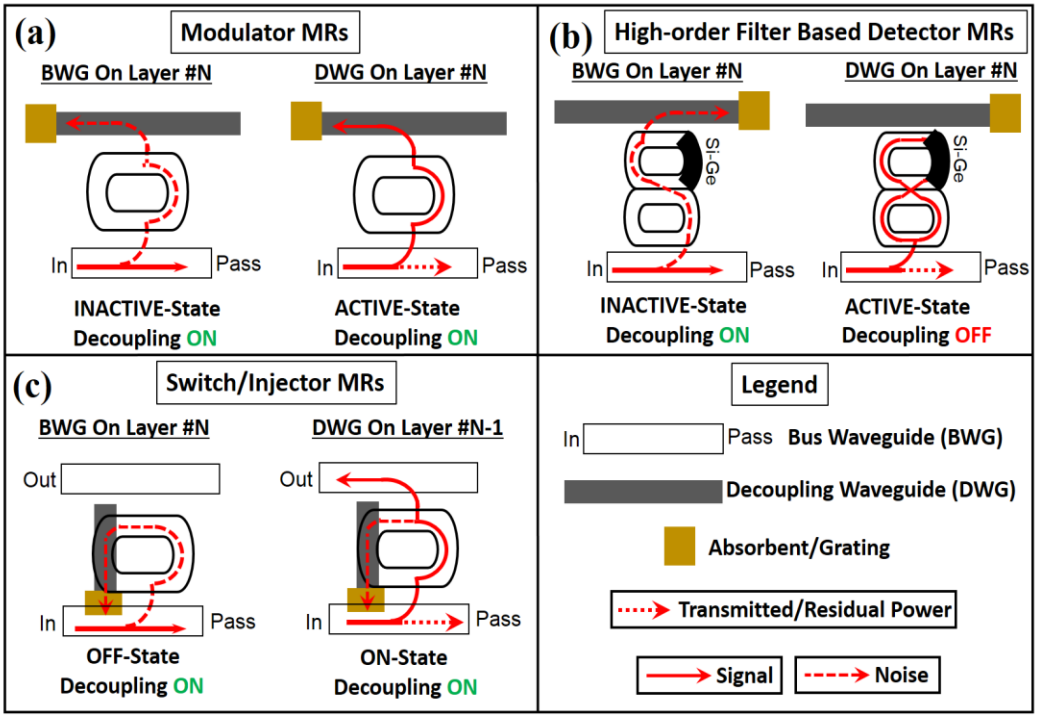


Figure 19: Illustration of decoupling waveguide (DWG) implementation for: (a) modulator MRs, (b) detector MRs, (c) switch MRs.

We refer to this secondary coupling waveguide as decoupling waveguide (DWG). The inclusion of the DWG is functionally similar to using add-drop port MRs (described in [93]) except that the DWG has an absorbent/grating structure at its drop port, which absorbs or scatters out all the power that is dropped at the drop port. The absorbent/grating structure is the same as used in [68] as part of the dithering signal based mechanism proposed for in-situ, high-speed detection of temperature and process variation in silicon photonic chips. Note that an increase in round-trip loss decreases MR's *Q* factor, which may result in increase of heterodyne crosstalk, as discussed in

Section 3.2. The decrease in Q may also increase the through loss of the MR, resulting in a decrease of signal strength and consequently a decrease in SNR. This increase in heterodyne crosstalk noise and MR's through loss can be mitigated by using higher-order MR filters with detector MRs [88] [89], in conjunction with encoding based techniques [60]. The use of higher-order filters with detector MRs is illustrated in Fig. 19(b), the details of which are presented later in this section. Moreover, note that the introduction of a DWG along the BWG incurs negligible amount of inter-waveguide crosstalk, as we assume minimum 10μm separation (equal to the diameter of the coupled racetrack MR) between the DWG and BWG.

As shown in Figure 19(c), a DWG can be used to decouple P_{circ} from a switch MR as well. Optical switch MRs are typically coupled into two waveguides. As a result, they have four ports (Figure 19(c)). As a switch MR is already coupled to two parallel waveguides, the introduction of a third waveguide as DWG in the same plane would incur extra loss due to waveguide crossings. Therefore, to avoid waveguide crossings and related losses, we introduce the DWG for a switch MR on a separate layer below or above the MR. A monolithic integration of DWG on a separate layer above or below the layer of other photonic devices is possible with the monolithic multilayer integration technology described in [69]. As discussed in [69] and Chapter 5, this type of monolithic multilayer integration of photonic devices using CMOS back-end compatible materials can be highly area-efficient, energy-efficient and cost-effective. In the absence of a DWG, the circulating power in an inactive/active (OFF-state/ON-state) switch MR can recouple back at the pass/output port, which will induce homodyne crosstalk at the pass/output port. Nevertheless, in the presence of a DWG, P_{circ} is decoupled from the MR to the DWG, which mitigates the homodyne crosstalk noise at pass/output port of the switch MR. Figure 19(b) shows two cascaded racetrack rings, which manifests a symbol representing a higher order filter for the embedded Si-

Ge detector. If this detector MR is in an inactive state, P_{circ} needs to be decoupled from it to reduce the homodyne crosstalk noise. In contrast, if P_{circ} is decoupled when the detector MR is in active state, then the signal strength (which in fact depends on P_{circ} in this case) reduces, which in turn deteriorates the SNR.

The above discussion implies that a mechanism is needed for the DWG to detune its decoupling capability when the detector MR is in active state. This motivates the design of a tunable DWG (TDWG), which is described in detail in the next section.

3.5. TUNABLE DECOUPLING WAVEGUIDE

Before we describe the design of a TDWG, it is important to understand how a TDWG works. The straight section of a racetrack MR, which is parallel to the TDWG (Figure 20), works as a directional coupler. Some part of the optical power that travels along this straight section of the MR is coupled to the TDWG depending on the design of the TDWG. Typically, a TDWG is a standard *Si*-SiO_2 waveguide, with a *Si* core and SiO_2 cladding. The coupling of power between the MR and the TDWG is given by the following equations [97]:

$$P_{TDWG}(z) = P_{MR}\frac{{K_d}^2}{{K_d}^2 + \delta^2}\left[\sin\left\{\left({K_d}^2 + \delta^2\right)^{\frac{1}{2}}z\right\}\right]^2, \tag{36}$$

$$\delta = (\beta_{MR} - \beta_{TDWG})/2, \tag{37}$$

$$P_{MR} = (a^2\kappa^2 P_{in})/2, \tag{38}$$

Here, K_d is coupling coefficient, P_{TDWG} is the amount of power decoupled from the MR into the TDWG, β_{MR} and β_{TDWG} are propagation constants of light in the MR waveguide and TDWG respectively, z is the length of the straight section of the MR, and P_{MR} is the optical power in the MR waveguide at the coupling region. From Eq. (36)-(38), P_{TDWG} depends on K_d. As discussed in

[97], K_d depends on field penetration depth (*d),* gap (*g*), and the refractive index of the MR waveguide's core and TDWG's core (n_{Si}). This implies that K_d, and hence P_{TDWG} can be controlled by tuning n_{Si}. The change in n_{Si} can be achieved by changing the TDWG's *Si* core free carrier concentration, which works on the same principles of free carrier dispersion as in the state-of-the-art MR modulators [96]. Similar to MR modulators, the TDWG can be doped like a PN-junction to function in a reverse-biased manner to render better control on free carrier concentration in the TDWG's *Si* core with faster response.

Figure 20 depicts the cross-sectional structure of the TDWG's PN-junction along with the gap *g* and penetration depth (coupling depth) *d*. Figure 20(a) and (b) illustrate how the decoupling capabilities of a TDWG can be tuned ON and OFF by changing the reverse bias condition of the constituent PN-junction.

As shown in Figure 20(a), at zero bias condition, the PN-junction has a narrower depletion region resulting in higher free carrier concentration in the core. In this case, the coupling depth (*d*) of the TDWG is smaller than the gap *g*. In contrast, as shown in Figure 20(b), an application of a non-zero reverse bias across the junction increases the depletion width resulting in the decrease of carrier concentration, which in turn increases n_{Si} causing *d* to penetrate beyond *g*. When *d* penetrates beyond *g*, a significant amount of power is decoupled from the MR into the TDWG core. Thus, n_{Si} of a TDWG can be altered by changing the free carrier concentration in the TDWG core, which ultimately tunes the TDWG decoupling capability between ON and OFF states.

It is imperative to design n_{Si} and carrier concentration values of the PN-junction appropriately to get the right amount of decoupling for ON and OFF states of the TDWG. To understand the relationship among the decoupled power P_{TDWG}, free carrier concentration (N_e=N_h), and n_{si}, we plot P_{TDWG} (in dB) and N_e=N_h versus n_{Si} in Figure 21 for a TDWG. We sweep n_{Si} from

2.4 to 4.1. As shown in Figure 21, the TDWG can be designed to decouple up to 30dB power from the MR. The smallest value of decoupled power (0.1dB) can be achieved for n_{Si}=2.4 and N_e=6×10^{20}. At this low value of decoupled power, the TDWG can be considered in OFF state, as it decouples negligible power from the MR.

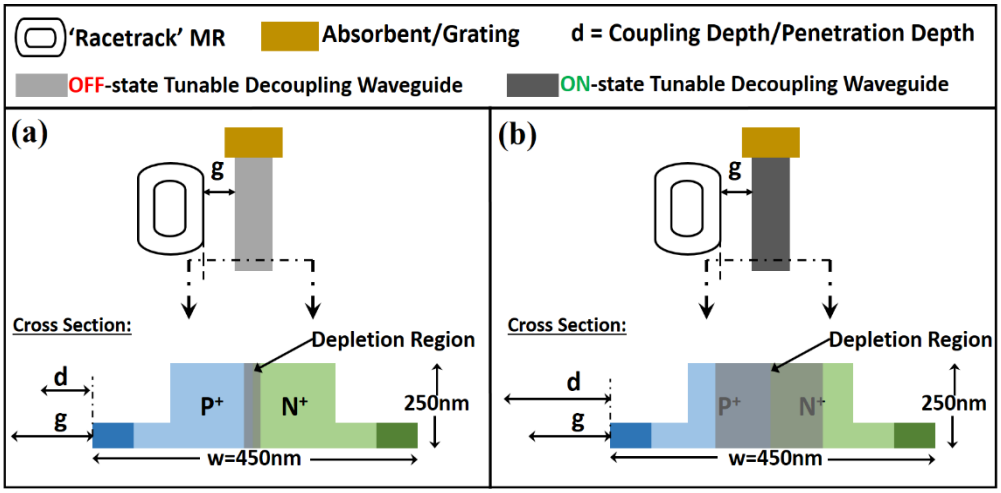


Figure 20: Cross-sectional structure of tunable decoupling waveguide (TDWG): (a) OFF-state TDWG, (b) ON-state TDWG.

We propose to dope the TDWG's PN-junction (Figure 20(a)) to achieve N_e=6×10^{20} at zero bias condition. Thus, the default state of the TDWG is OFF state. As evident from Figure 21, the TDWG can be switched ON by applying a reverse biased voltage across the junction, which would in turn decrease N_e below 6×10^{20} causing decoupled power of greater than 0.1dB. The target value of decoupled power for the ON-state TDWG should be chosen based on the underlying BWG and PNoC architectures, so that the worst-case P_N emanated from MRs is minimized by maximizing the extenuation of P_{circ} from MRs. For a given value of signal power, the worst-case (maximum) P_N results in the worst-case (minimum) SNR. Moreover, the value of decoupled power that maximizes the extenuation of P_{circ} not only minimizes the worst-case P_N, but also guarantees the minimization of the best-case and average-case P_N. Therefore, the minimization of the worst-case

P_N not only maximizes the worst-case SNR, but also guarantees the maximization of the best-case and average-case SNR values.

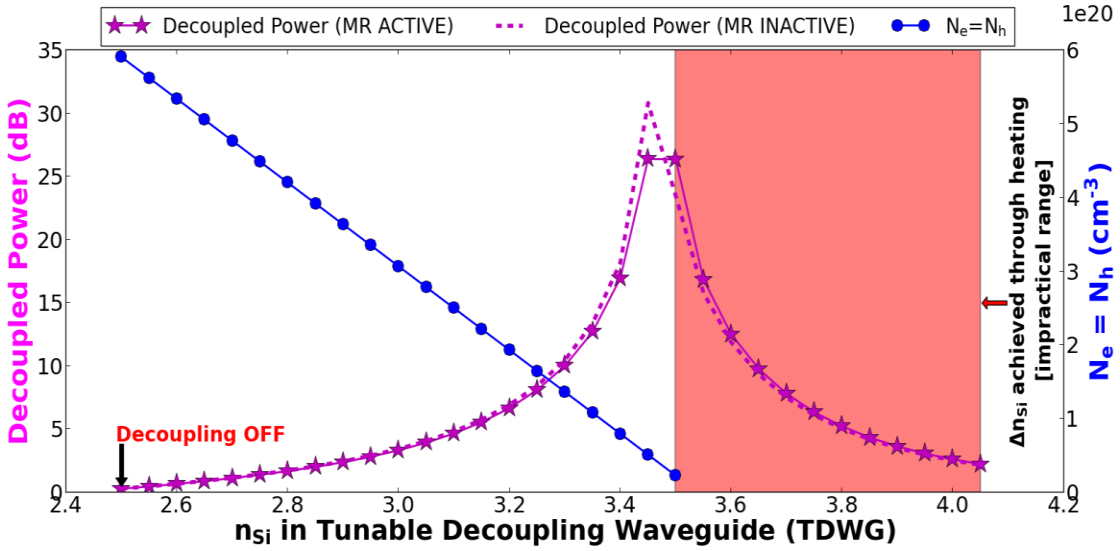


Figure 21: Decoupled power and free-carrier concentration (N_e) vs. n_{Si} of TDWG for ACTIVE, INACTIVE state MRs. (g=200nm, z=2µm).

However, note that the maximization of the worst-case SNR is more important than the maximization of the best-case or average-case SNR, as the worst-case SNR determines the BER, which in turn affects the reliability of communication in PNoCs. Therefore, the target value of decoupled power for the ON-state TDWG should be chosen based on the underlying BWG and PNoC architectures, so that the worst-case SNR is maximized. As an illustration of this fact, we present an architecture-specific implementation of TDWG for the Corona PNoC in Section 3.5.1. Note that the use of TDWG to mitigate homodyne crosstalk incurs overhead, the analysis of which is presented in Section 3.5.2.

3.5.1. IMPLEMENTATION OF TDWG FOR THE CORONA PNOC

We now demonstrate how the use of TDWG can mitigate homodyne crosstalk noise in the Corona PNoC. We direct the reader to [14] for detailed information about the Corona PNoC

architecture. Briefly, Corona consists of 256 general-purpose cores grouped into 64 four-core clusters. These clusters are connected together through three bus waveguides (BWGs) that include an optical crossbar for data communication, a broadcast bus for multi-casting, and an arbitration waveguide. The main laser source is fed into a loop and split into these BWGs. The optical crossbar for data communication is comprised of a multiple-write-single-read (MWSR) BWG, which starts at cluster 0, passes through 62 intermediate clusters, and ends at cluster 63. Each cluster has a bank of 64 MRs arrayed along the BWG. When an optical signal travels from cluster 0 to cluster 63 along the BWG, all the MRs (total 4096 MRs; 64 clusters having 64 MRs each) along the BWG incur MR through losses in the signal. Thus, a very large number of utilized MRs incur a very high value of optical signal loss in Corona data BWG, which results in very poor signal strength. A very poor signal strength in the Corona data BWG makes it more susceptible to homodyne noise.

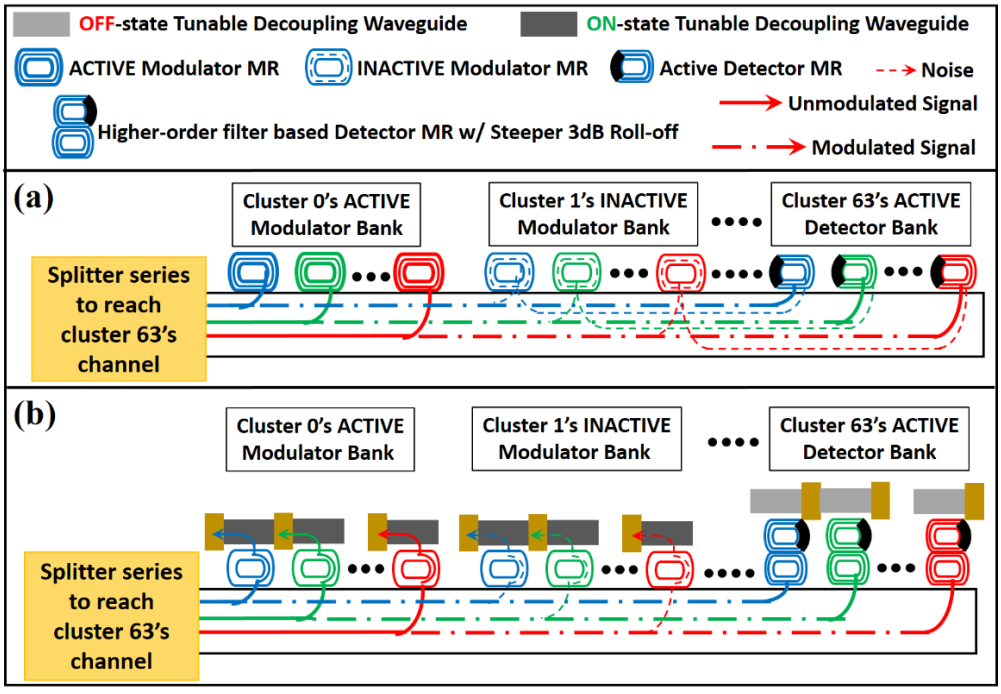


Figure 22: Demonstration of (a) worst-case homodyne noise in Corona PNoC's data bus waveguide, (b) use of tunable decoupling waveguide in Corona PNoC's data bus waveguide to mitigate homodyne noise.

Figure 22(a) shows the worst-case homodyne noise scenario in the data BWG. From the figure, the modulator bank of cluster 0 is in active state and it modulates a data packet on the DWDM wavelengths of the BWG. This data packet travels along the BWG and passes through inactive-state MR modulator banks of 62 intermediate clusters before reaching the active-state MR detector bank of cluster 63. As shown in the figure, while the data packet travels through cluster 0 to cluster 63, the modulator banks of all 62 intermediate clusters incur homodyne crosstalk noise in the data packet. This noise is picked up by the detector bank of cluster 63 along with the modulated data packet.

Figure 22(b) demonstrates the use of TDWGs to mitigate the homodyne crosstalk noise in the Corona data BWG. As shown in the figure, the TDWGs corresponding to the modulator banks of cluster 0 to cluster 63 are in ON state, whereas the TDWGs corresponding to the detector bank of cluster 63 are in OFF state. The ON-state TDWGs of intermediate modulator banks decouple the circulating power from their respective MRs, which in turn greatly reduces the homodyne crosstalk noise. As discussed in Section 3.4, we propose to use higher-order (3^{rd} order or higher) filters with detector MRs in the detector bank to reduce the heterodyne crosstalk noise, which is illustrated in the figure.

To select the most appropriate value of decoupled power for ON-state TDWGs, we evaluated the worst-case SNR in the data BWG and the control BWG (described in [14] and [48]) of the Corona PNoC as a function of decoupled power using the equations in Section 3.3 and 3.5. In our calculation of SNR, we also factored in the insertion losses incurred by the splitters, waveguide, and MRs, along with the heterodyne crosstalk noise incurred due to the passband overlap of MRs [56]. The plots of the worst-case SNR are shown in Figure 23. From the figure, the improvements in SNR for the data BWG and the control BWG saturate for values of decoupled power greater

than 15dB and 20dB respectively. Therefore, we select 15dB as the target value of decoupled power for TDWGs in the Corona data BWG and 20dB for the control BWG. The architectures of the Corona data BWG and control BWG differ from each other, which results in different values of desired decoupled power for them. Hence, it can be concluded that the target values of decoupled power for ON-state TDWGs should be chosen based on the underlying BWG and PNoC architecture.

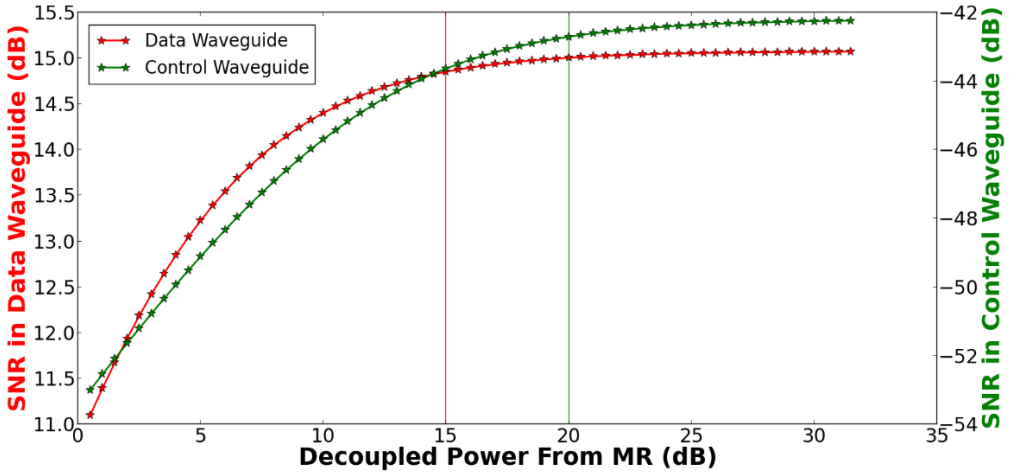


Figure 23: SNR vs. decoupled power for Corona PNoC's data and control waveguides.

From Figure 23, the worst-case SNR for Corona PNoC's data BWG increases from about 11dB to 15dB when the decoupled power is increased from about 0.1dB to 15dB. Now, as discussed in [83], the worst-case SNR, when expressed in absolute value (not in dB), is related to the BER through the following equation: $BER = 0.5 \times exp(-SNR/4)$. From this relation between the worst-case SNR and BER, the increase in the worst-case SNR from 11dB to 15dB corresponds to the improvement (decrease) in BER from 2.1×10^{-2} to 1.8×10^{-4}. From [83], the maximum allowable BER for reliable communication is 10^{-9}. This implies that even after achieving about 100× improvement in BER (from 2.1×10^{-2} to 1.8×10^{-4}), obtained by minimizing the worst-case homodyne noise, sufficiently reliable communication cannot be achieved. This is because, even

after minimizing the homodyne crosstalk noise, the heterodyne crosstalk noise still exists in PNoCs, which needs to be significantly mitigated to achieve BER in the order of 10^{-9} or less and reliable communication in PNoCs. However, note that this chapter focuses only on mitigating the homodyne crosstalk noise, and the problem of heterodyne crosstalk mitigation is out of the scope of this work.

3.5.2. DEVICE-LEVEL OVERHEAD ANALYSIS

We now evaluate the area and energy overhead incurred by our proposed TDWGs. The default state of a TDWG is the OFF state, where it does not consume static or dynamic energy. A TDWG consumes dynamic switching energy every time it is switched ON, and after switching ON it consumes static power related to the reverse saturation current in the constituent PN-junction for the entire time it is ON. As illustrated in Figure 19, only the TDWGs corresponding to detector MRs need to dynamically switch ON and OFF. The TDWGs corresponding to modulator and switch MRs always remain ON. Therefore, the TDWGs corresponding to modulator and switch MRs do not consume dynamic switching energy, but consume only static power. In contrast, as the TDWGs corresponding to detector MRs need to switch OFF while the detector MRs are detecting data bits and then switch ON immediately after the detection event, they consume dynamic energy every time a data packet is transferred.

The amount of static power and dynamic energy consumed depends on the level of decoupled power in the ON-state TDWGs. In Figure 24, we plot static power density (fW/cm^2) and dynamic energy consumption values versus decoupled power. From the figure, static power density decreases and dynamic energy increases with increase in decoupled power. Moreover, it can be observed that for the 15dB decoupled power chosen for the Corona data BWG, the values of dynamic energy and static power density are 4.3pJ and $31.76fW/cm^2$ respectively.

Furthermore, as mentioned in Figure 21, we select the length of the MR's straight section (z) that is coupled to the TDWG to be 2μm. As depicted in Figure 20, the width (w) of the TDWG is 450nm. Therefore, if we consider the length of a TDWG to be 2μm more to account for its portion that is not coupled to the MR, and if we assume the length and the width of the absorbent/grating to be 1μm each, then a single TDWG along with the grating would consume 2.8μm^2 surface area. As depicted in Figure 20, the height of a TDWG is 250nm, which results in the volume of a single TDWG along with the grating to be 0.7μm^3.

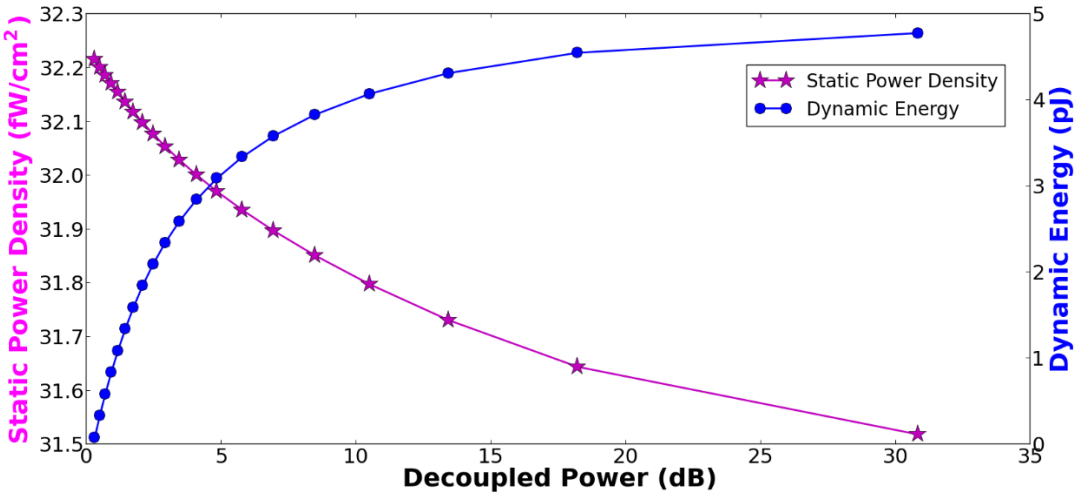


Figure 24: Static power density and dynamic energy overhead of tunable decoupling waveguide versus decoupled power.

Note that the introduction of TDWGs incur extra fabrication cost that is related to the amount of utilized silicon real estate. The amount of utilized silicon real estate is proportional to the surface area and volume overhead of TDWGs. As mentioned earlier, a single TDWG along with the grating would consume 2.8μm^2 surface area, which is very less compared to the surface area of 78.5μm^2 consumed by a typical circular MR. This implies that the introduction of TDWGs incur minimal cost overhead related to device fabrication. Moreover, TDWGs have significantly low

engineering design cost as well, as the structure of the TDWGs is simple and fairly repetitive, which reduces the complexity of mask-set designs and fabrication process.

3.6. EVALUATION

3.6.1. EVALUATION SETUP

To evaluate our proposed homodyne crosstalk noise mitigation (*HCTM*) approach in DWDM-based PNoCs, we implement and integrate it with three well-known crossbar-based PNoCs: Corona [14], Flexishare [16], and Firefly [15]. We modeled and performed simulation based analysis of these enhanced PNoCs using a cycle-accurate NoC simulator, for a 256-core system at 22nm. GEM5 full-system simulation [77] of parallelized PARSEC applications [76] was used to generate traces that were fed into our cycle-accurate NoC simulator. We set a warm-up period of 100 million instructions and then captured traces for the subsequent 1 billion instructions. We performed geometric calculations for a 20mm× 20mm chip, to determine lengths of waveguides in the Corona, Firefly and Flexishare PNoCs. Based on this analysis, we estimated the time needed for light to travel from the first to the last node as 8 cycles in all of these PNoCs at 5 GHz clock frequency. We use a 512-bit packet size, as advocated in all of these PNoCs.

The static and dynamic energy consumption of electrical routers and concentrators in Corona, Firefly, and Flexishare is based on results from the DSENT [75] tool. We evaluate area and energy overheads for our *HCTM* technique when implemented with the above mentioned PNoCs, the results of which are given in Section 6.3. For energy consumption of photonic devices, we adapt parameters from recent work [60], [79], and [80], with 0.42pJ/bit for every modulation and detection event and 0.18pJ/bit for the driver circuits of modulators and photodetectors. We used 0.02dB, 1dB/cm, 0.005dB/90°, and 0.5dB values of MR through loss, waveguide propagation loss, waveguide bending loss and coupling loss respectively, to determine the photonic laser power

budget and correspondingly the electrical laser power. The MR trimming power is set to 130μW/nm [22] considering a 40°K trimming range.

3.6.2. EVALUATION RESULTS FOR STATE-OF-THE-ART PNOCS

First, we evaluate the worst-case SNR in Corona, Flexishare and Firefly PNoCs when used with our *HCTM* technique, and compare the results with the SNR in their respective baselines. We considered two different baseline configurations for each PNoC architecture: 1) a configuration with only heterodyne crosstalk without any homodyne crosstalk (ideal case); 2) a configuration with homodyne as well as heterodyne crosstalk.

In the Corona PNoC with token ring arbitration [14], we consider multiple write single read (MWSR) data BWGs as well as control/arbitration BWGs for our evaluation of worst-case SNR. Likewise, we consider the multiple write multiple read (MWMR) data BWGs of Flexishare PNoC [16] for evaluation. The Flexishare PNoC is a 64-radix, 64-node architecture with 4 cores in each node having 32 data channels for inter-node communication. We also consider the data BWGs of the Firefly PNoC [15], which are configured as reservation-assisted single write multiple reader (R-SWMR) data BWGs.

Generally, in a PNoC, for given crosstalk noise, the worst-case SNR happens at the MR detector bank of a node for which the loss in signal is the highest, as the highest loss of signal renders the worst signal power and hence the worst SNR. We refer to such a node as the worst-case power loss node (WCPLN). Cluster #63 in Corona PNoC and node 63 (R63) in Flexishare PNoC are the WCPLNs. Similarly, in Firefly PNoC [15], the router 0 of cluster 4 (C_4R_0) is the WCPLN. We utilize the models presented in [48] (for heterodyne crosstalk) and Section 3 and 5 of this chapter (for homodyne crosstalk) to calculate the total received crosstalk noise (including both homodyne and heterodyne crosstalk noise) and SNR at the detectors of these WCPLNs. Note

that the worst-case heterodyne crosstalk noise for a data BWG of any PNoC occurs when all the 64-bits of the received data packet are 1's. In contrast, as implied from the discussion given in Section 3.5.1, the worst-case homodyne crosstalk noise in a data BWG of any PNoC occurs when the signal travels through the maximum number of intermediate nodes on its path to the WCPLN (because in this case the individual contributions of homodyne noise from all the intermediate nodes add up to incur the worst noise at the WCPLN).

Figure 25 summarizes the worst-case SNR results. It can be observed that SNR values for PNoCs with *HCTM* (brown bars) are very close to ideal SNR values (blue bars). This corroborates the excellent capabilities of our proposed *HCTM* technique in reducing homodyne crosstalk noise. From the figure, it can be observed that data BWGs of Corona, Flexishare and Firefly PNoCs with HCTM have 37.6%, 12.2%, and 4.6% SNR improvements on average compared to their respective baselines. Further, it can also be seen that the worst-case SNR of the Corona control BWG with HCTM increases by 20.1% compared to its baseline.

3.6.3. SYSTEM-LEVEL OVERHEAD ANALYSIS

In Section 3.5.2, we presented area and energy overhead of TDWGs. In this section, we present how the area and energy overheads of individual TDWGs manifest at the system-level. To know how the energy overhead of TDWGs affect the total energy consumption of the overall system, we perform simulation-based quantification of energy consumption for the aforementioned three PNoCs when used with *HCTM*. In our simulation study, we take 15dB as the target value of decoupled power for the TDWGs of the data BWGs of Firefly and Flexishare PNoCs, which is the same as chosen for the Corona data BWGs in Section 3.5.1.

The results of this simulation study are plotted in Figure 26. As evident from the figure, the Corona, Flexishare, and Firefly configurations with our *HCTM* technique have 19.2%, 18.5% and

0.5% higher energy consumption on average compared to their baseline configurations respectively. Total energy consumption for PNoCs with *HCTM* increases due to the additional dynamic energy and static energy consumption of the constituent TDWGs. As explained in section 3.5.2, the TDWGs corresponding to detector MRs consume dynamic energy every time a data packet is transferred over their corresponding BWG. The Firefly architecture, being a hybrid electro-photonic NoC, routes fewer packets through photonic BWGs compared to Corona and Flexishare PNoCs, which in turn reduces the dynamic energy consumption in its TDWGs resulting in reduced total energy overhead.

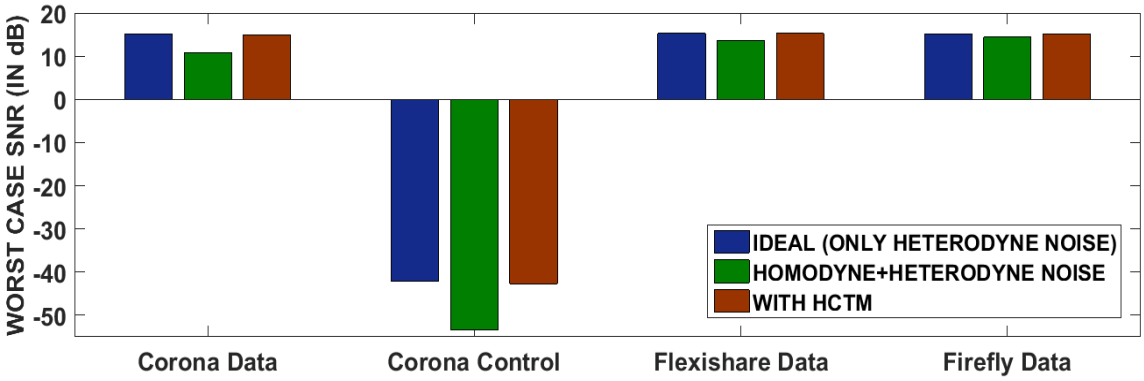


Figure 25: Worst-case SNR comparison of data and control waveguides of Corona, data waveguides of Flexishare, and data waveguides of Firefly with *HCTM* and their respective baseline configurations.

Lastly, we took the physical dimensions of MRs and splitters from [72], and used them with the dimensions of TDWGs (from Section 3.5.2) to evaluate the photonic area overhead for *HCTM*. We found the photonic area overhead for *HCTM* in the Corona, Flexishare and Firefly PNoCs to be 6.76mm^2, 2.82mm^2, and 5.63mm^2 respectively. This area overhead is small relative to the total planar area of the chip (400 mm^2).

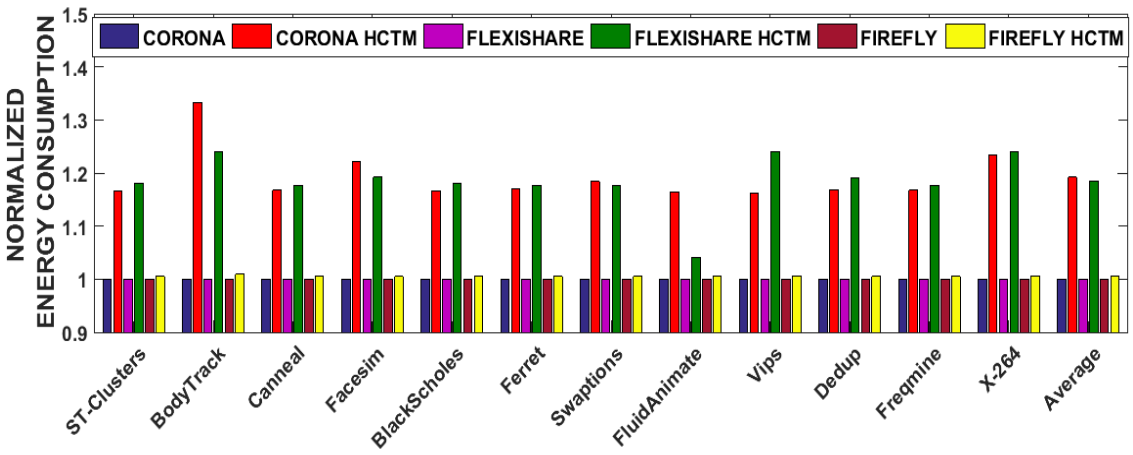


Figure 26: Comparison of normalized energy consumption in Corona, Firefly and Flexishare PNoCs with *HCTM* and their respective baselines, for 12 PARSEC benchmarks. Energy consumption values are normalized to energy values of baseline configurations.

3.7. CONCLUSIONS

In this chapter, we demonstrated that the circulating power of MRs, built up due their resonant nature, induces time-dependent homodyne crosstalk noise in DWDM PNoCs. We also presented a lightweight low overhead homodyne crosstalk mitigation (*HCTM*) technique for the reduction of homodyne crosstalk noise in DWDM PNoCs. The HCTM technique is agnostic to the time-dependent characteristics of homodyne crosstalk. Moreover, it also shows interesting trade-offs between reliability and energy overhead. We evaluated the effectiveness and overhead of our HCTM technique by implementing it for well-known PNoC architectures, including Corona, Firefly and Flexishare. Our experimental analysis showed that our approach when implemented on these PNoCs can improve the worst-case SNR by up to 37.6% compared to the baseline versions of these PNoCs, thereby significantly enhancing reliability, at the cost of up to 19.2% energy overhead and 1.7% photonic area overhead. Thus, *HCTM* represents an attractive solution to enhance reliability in emerging DWDM-based PNoCs.

4. LIBRA: THERMAL AND PROCESS VARIATION AWARE RELIABILITY MANAGEMENT IN PHOTONIC NETWORKS-ON-CHIP

PNoCs operation is very sensitive to on-chip temperature and process variations. These variations can create significant reliability issues for PNoCs. For example, a microring resonator (MR) may resonate at another wavelength instead of its designated wavelength due to thermal and/or process variations, which can lead to bandwidth wastage and data corruption in PNoCs. This chapter proposes a novel run-time framework called *LIBRA* to overcome temperature- and process variation- induced reliability issues in PNoCs. The framework consists of (i) a device-level reactive MR assignment mechanism that dynamically assigns a group of MRs to reliably modulate/receive data in a waveguide based on the chip thermal and process variation characteristics; and (ii) a system-level proactive thread migration technique to avoid on-chip thermal threshold violations and reduce MR tuning/ trimming power by dynamically migrating threads between cores. Our simulation results indicate that *LIBRA* can reliably satisfy on-chip thermal thresholds and maintain high network bandwidth while reducing total power by up to 61.3%, and thermal tuning/trimming power by up to 76.2% over state-of-the-art thermal and process variation aware solutions.

4.1. INTRODUCTION

As advocated by prior works [14]-[16], PNoCs are expected to be 3D-stacked on top of their respective manycore chips. Therefore, the MRs of PNoCs will be placed on top of, and hence in close proximity to, processing cores. Variations in core workloads lead to variations in their power dissipation, which in turn can alter the temperatures of the cores and MRs in their vicinity. For instance, the temperature on a typical manycore chip can easily vary by as much as 90°C [98].

Unfortunately, MRs are very sensitive to these on-chip thermal variations (TV): their effective refractive indices, and hence their resonance wavelengths are altered if their operating temperatures change. Therefore, in a typical PNoC, the resonance wavelengths of the utilized modulator MRs may not align with, and hence may not modulate their assigned carrier wavelengths [22]. This may result in bandwidth wastage, or worse, data corruption when detector MRs are unable to read from their assigned carrier wavelengths [99].

In addition to TV, MRs are also susceptible to fabrication process variations. Process variations (PV) induce variations in the width, thickness, and doping concentration of MRs (see Section 4.3.2), causing resonance wavelength shifts in MRs [33], [32]. PV measurements of fabricated MR devices indicate a standard deviation (σ) of 1.3 nm in width, which translates to a 0.76nm shift in an MR's resonance wavelength [34]. These PV-induced resonance wavelength shifts in MRs also cause bandwidth wastage and data corruption.

The adverse effects of PV and TV related to resonance shifts in MRs, and their performance and reliability impacts, can be redressed by realigning the resonant wavelengths of MRs with their assigned carrier wavelengths using localized trimming [21] and thermal tuning [22] mechanisms. Trimming alters the free-carrier concentration in an MR core, whereas thermal tuning uses integrated micro-heaters to alter local temperatures at MRs. But these mechanisms come with high power and performance overhead [22]. Hence, it is essential to intelligently manage thermal and process variations in PNoC-based manycore systems, to achieve reliable communication with minimal trimming and tuning costs.

In this chapter, we aim to minimize the need for (and overheads of) localized thermal tuning and trimming in PNoCs while coping with process and thermal variations, thereby easing the adoption of PNoCs in future manycore systems. We propose a novel thermal and process variation

aware dynamic reliability management framework called LIBRA that integrates adaptive MR assignment at the device-level and dynamic thread migration at the system-level for PNoC-based manycore systems. Our novel contributions as part of the LIBRA framework are summarized below:

- We design a novel thermal and process variation aware MR assignment (TPMA) mechanism at the device-level, which dynamically assigns a set of MRs to the utilized set of carrier wavelengths at run-time. TPMA enables reliable modulation and reception of data with minimal overheads, while maintaining the maximum possible bandwidth;
- We propose a novel PV-aware anti wavelength-shift dynamic thermal management (VADTM) mechanism at the system-level, which uses support vector regression (SVR) based temperature prediction and dynamic thread migration to avoid on-chip thermal threshold violations and reduce trimming/tuning power for MRs;
- We evaluate our LIBRA (TPMA+VADTM) framework on a 64-core chip, and compare it with four state-of-the-art thermal management solutions: an MR-aware thermal management (RATM) framework [100], an MR PV-aware thermal management (FATM) framework [11], a predictive dynamic thermal management (PDTM) framework [101], and an MR-aware thermal management (SPECTRA) framework [99]; and show significant reduction in maximum temperature and trimming/tuning power costs compared to these solutions.

4.2. RELATED WORK

Traditional electrical NoC communication fabrics are projected to suffer from cripplingly high power dissipation and severely reduced performance in future manycore systems [8]. The higher bandwidth density and lower power dissipation possible with silicon-photonic links,

compared to electrical wires, has made them an attractive option for manycore systems. Recent research has thus focused on exploring a wide spectrum of network topologies and protocols to enable efficient PNoC architectures [102], [81].

PV and TV in silicon-photonic links represent important challenges for the widespread adoption of PNoC architectures. Several techniques have been proposed to reduce thermal hotspots and gradients using DVFS [103]-[105], workload migration [81], [101], [106], [107], and liquid cooling [108]-[110]. A few PV-aware application mapping frameworks have also been proposed [111], [112] that optimize performance and energy in manycore systems. In [111] a run-time application-mapping strategy was presented, which considers the variation profile of a manycore processor to maximize performance and reduce leakage-power for a given fixed power budget. In [112] a framework was presented that integrates reliability and variation-awareness in a run-time variable degree-of-parallelism (DoP) application-scheduling methodology to enhance manycore performance. However, these techniques do not consider the unique challenges (e.g., MR resonance wavelength shifts) and constraints (e.g., wavelength match between sender and receiver MR pairs) that exist in PNoCs.

A few prior works have analyzed the impact of TV and PV on PNoCs at the device-level, link-level, and system-level, and proposed solutions to remedy these variations. The device-level efforts have mainly proposed various athermal photonic devices to reduce localized tuning/trimming power in MRs. These design-time solutions include using materials such as cladding to reduce thermal sensitivity [113], [114], and using heaters and temperature sensors for thermal control [115]. An electrical backend capable of bit re-shuffling was proposed in [116] to enhance photonic link robustness against TV and PV with lower MR tuning power. While these device- and link-level techniques are promising, they either possess a high-power overhead or

require costly changes in the manufacturing process (e.g., larger device areas) that would decrease bandwidth density and area efficiency.

At the system-level, the overhead associated with localized tuning of MRs was reduced in [22] using the group shift property of co-located MRs as part of a method to trim a group of rings at the same time. In [100], a ring-aware thread scheduling policy was proposed to reduce on-chip thermal gradients in a PNoC. In [117], a thread migration mechanism was proposed to minimize on-chip thermal gradients within a PNoC. In [118], an island of heater based thermal management framework was proposed to adapt groups or islands of MRs within PNoCs to on-chip thermal variations. A few prior works have also explored the impact of PV on DWDM-based photonic links at the system-level [63], [119], [120]. A reliability-aware design flow to address variation induced reliability issues is proposed in [119], which uses athermal coating at fabrication-level, voltage tuning at device-level, as well as channel hopping at the system-level. In [63], a methodology to salvage network-bandwidth loss due to PV-shifts is proposed, which reorders MRs and trims them to nearby wavelengths. In [120], power-efficient techniques are proposed, based on inter-channel hopping and variation-aware routing to compensate for PV effects at runtime. A few system-level works [11], [34], [121] also consider the impact of both TV and PV on optical links. In [34], a thermal-tuning approach is presented that adjusts chip temperature using DVFS to compensate for chip-wide thermal and process variation induced resonance shifts in MRs and improve system performance. In [11], a PV aware workload allocation policy is presented to reduce the thermal tuning power of PNoCs. In [121], a tunable laser source design is demonstrated, in which the signal power at the source is adapted to compensate for signal losses due to TV and PV across optical interconnects. None of these system-level solutions for PNoCs considers the impact of the relationship between thermal hotspots and transmission reliability.

To address these shortcomings of prior work, we proposed the SPECTRA framework in our prior work [99]. SPECTRA is a cross-layer framework that combines two dynamic thermal management mechanisms to reduce maximum on-chip temperature and conserve trimming and tuning power of MRs in DWDM-based PNoC architectures. Our proposed LIBRA framework in this chapter improves upon SPECTRA, by (i) considering the impact of PV on dynamic thermal management; (ii) utilizing a new device-level TV and PV aware ring assignment mechanism; and (iii) utilizing a new system-level PV-aware thread migration mechanism. Sections 4.4-4.6 describe our proposed framework which is then evaluated in Section 4.7 against prior work.

4.3. IMPACT OF TV AND PV ON DWDM BASED PNOCS

In this section, we explain the key impacts of PV and TV on DWDM based PNoCs. Although most silicon-based photonic devices exhibit some susceptibility to temperature and process variations, the high wavelength selectivity of MRs makes them especially susceptible to these variations. Therefore, we primarily focus on the impacts of TV and PV on MRs.

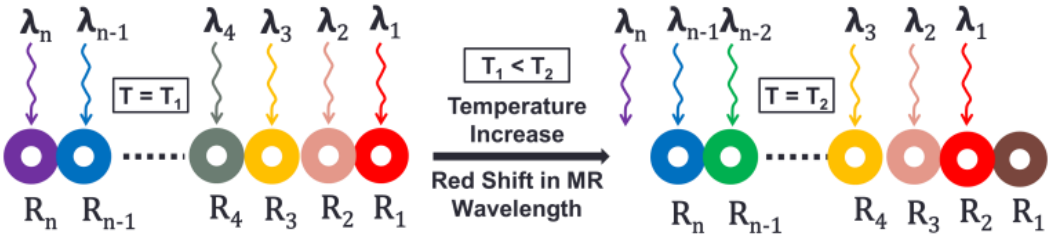


Figure 27: Impact of temperature increase on an MR bank.

4.3.1. IMPACT OF TV ON DWDM BASED PNOCS

In a DWDM PNoC, the temperatures of the individual compute nodes and their associated MR banks follow the workload-dependent temperatures of the processing cores in the nodes. As the application workload of each core in a manycore system usually differs from that of other cores

and also varies with time, the temperatures of the cores (and thus nodes) of the system differ from one-another and vary with time. As a result, the temperatures of different MR banks of the PNoC also differ from one another and vary with time.

Typically, the MR banks of each PNoC node are designed to resonate with and operate upon their assigned carrier wavelengths at a specific temperature, e.g., room temperature. But due to the time- and workload-dependent temperature variations, the resonances of different MR banks shift away from their assigned carrier wavelengths by different amounts.

For example, Figure 27 depicts an MR bank with MRs R_1-R_n that are designed to resonate with their assigned carrier wavelengths λ_1-λ_n, respectively, at temperature T_1. As the temperature increases to T_2, the resonance wavelength of each MR shifts away from its assigned carrier wavelength towards the red end of the spectrum (i.e., *red-shift*). This red-shift is shown in the figure where, at temperature T_2 (T_2>T_1), the resonance wavelength λ_i of MR R_i is in line with the carrier wavelength λ_{i-1}. Consequently, the carrier wavelength λ_n is not assigned to any of the MRs. This results in *bandwidth wastage* if the MR bank is a modulator MR bank, as λ_n cannot be modulated by any modulator MR now. This example scenario can also result in *data corruption* if the MR bank is a detector MR bank, as λ_n cannot be received by any detector MR. Similarly, if T_2 < T_1, the resonance wavelength λ_i of each R_i shifts towards the blue end of the spectrum (i.e., *blue-shift*), which may leave λ_1 unassigned, causing bandwidth wastage or data corruption. Thus, during the runtime of a PNoC, an increase in an MR bank's temperature red-shifts the resonances of all its MRs, whereas a decrease in an MR's temperature blue-shifts the resonances of all its MRs.

The amount of shift in an MR's resonance not only depends on the magnitude of temperature change, but also on the MR's structure and geometry manifested as its effective refractive index n_{eff}. Typically, an MR is a looped waveguide with a silicon (Si) core and silicon dioxide (SiO_2)

cladding, irrespective of whether it is used as a modulator or a detector. The change $\Delta\lambda_r$ in the resonance wavelength λ_r of an MR due to an arbitrary change ΔT in its local temperature is given by the following equation [69]:

$$\frac{\Delta\lambda_r}{\Delta T} = \frac{\delta n_{eff}}{\delta T}\frac{\lambda_r}{n_g} = \left(\Gamma_{Si}\frac{\delta n_{Si}}{\delta T} + \Gamma_{SiO2}\frac{\delta n_{SiO2}}{\delta T}\right)\frac{\lambda_r}{n_g}, \tag{39}$$

Here, n_g is the group refractive index (ratio of speed of light to group velocity of all wavelengths traversing the waveguide) of the MR waveguide. Γ_{Si} and Γ_{SiO2} are the modal confinement factors of the MR's core (Si) and cladding (SiO_2), respectively. $\delta n_{Si}/\delta T$ and $\delta n_{SiO2}/\delta T$ are the thermo-optic coefficients of *Si* (MR's core) and SiO_2 (MR's cladding) materials, with values of 1.86×10^{-4} K^{-1} and 1×10^{-5} K^{-1}, respectively [69]. As the thermo-optic coefficient of *Si* is an order of magnitude greater than that of SiO_2, and as Γ_{Si} is also greater than Γ_{SiO2} for a typical MR, the contributions from the MR's cladding (SiO_2) in Eq. (39) can be ignored. Consequently, Eq. (39) reduces to:

$$\Delta\lambda_r = \Gamma_{Si}.\frac{\delta n_{Si}}{\delta T}.\frac{\lambda_r}{n_g}.\Delta T, \tag{40}$$

Note that the MRs used in this study are looped channel waveguides with a cross section of 450nm×220nm. We model these MRs using a commercial-grade eigenmode solver [65], based on which the values of Γ_{Si} and n_g at 1550nm are calculated to be 0.78 and 4.16, respectively.

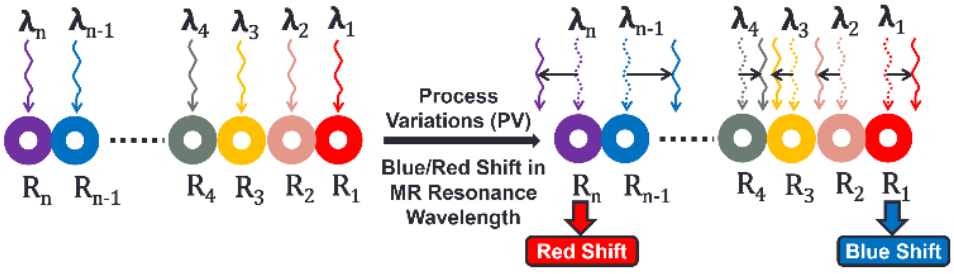


Figure 28: Impact of PV on DWDM based PNoCs.

4.3.2. IMPACTS OF PV ON DWDM BASED PNOCS

Ideally, without any fabrication-induced PV, a sender or a receiver node can modulate and detect all of the carrier wavelengths available in the waveguide without any bandwidth loss or error. But in reality, similar to deep submicron electronic devices, photonic devices such as MR modulators, MR detectors, grating couplers, splitters etc. also suffer from significant PV [122]. In this chapter, we mainly focus on the severe PV effects in MRs. The MR structure is very sensitive to PV, much like it is to TV. Due to PV effects, the widths, heights, and side wall roughness of MRs can deviate from desired values after fabrication. Consequently, the resonance wavelengths (λ_r) of the MRs also deviate from their designed values. For example, 1nm of variation in width and height of an MR can lead to 0.58~1nm and ~2nm shift in its resonance wavelength, respectively [33].

As discussed earlier, PNoCs employ DWDM-based photonic links with cascaded MRs (i.e., MR banks) in their sending and receiving nodes. Unlike TV that induces systematic red or blue shifts in all the MRs of an MR bank, PV can incur random shifts in the resonance wavelengths of the MRs of a single bank, as shown in Figure 28. From this figure, MRs R_1, R_4, …, R_{n-1} have blue shift in their resonance wavelengths and MRs R_2, R_3, …, R_n have red shift in their resonance wavelengths. Much like with TV, PV can also throw the resonances of the MRs out of alignment with their assigned carrier wavelengths, which can ultimately lead to bandwidth wastage and/or data corruption.

In summary, to enable reliable photonic communication, there is a need to mitigate the combined impact of TV and PV on PNoCs. This chapter presents a cross-layer framework that uses device-level and system-level enhancements to remedy the combined impact of TV and PV. Before discussing our proposed framework, we present our performance, power, and thermal setup

for modeling manycore systems with PNoCs in the next subsection. We also present a characterization of the impact of TV and PV on the MRs of a typical DWDM PNoC based manycore system, in this next subsection.

4.3.3. MODELING PV AND TV IN PNOC ARCHITECTURES

To model and characterize TV and PV in a manycore system with a PNoC, we developed a simulation framework, which integrates performance, power, thermal, and variation simulators, as shown in Figure 29. We considered a three-layered 3D-stacked 64-core system as advocated in existing PNoC architectures [14], [15] with a planar die area footprint of 400mm^2. The top layer is the core-cache layer, the middle layer is the conversion layer with digital and analog circuits that support electrical-to-optical (E/O) and optical-to-electrical (O/E) conversion of data, and the bottom layer is the photonic layer with photonic components and devices (e.g., MRs, waveguides, ring heaters, etc.) that comprise a PNoC.

We use Sniper [123] to simulate the performance of the manycore system while it executes multithreaded applications from SPLASH-2 [124] and PARSEC [76] benchmark suites. To factor in the varying system utilizations as a contributor to the dynamic TV in the processing cores and its impact on the associated photonic devices (e.g., MRs), we run each application on a target 64-core system (see Section 4.7) with 8, 16, 32, 48, and 64 threads. To capture runtime behavior of an application, we generate performance traces using Sniper, which are fed to MCPAT [125] to generate power traces at core-level granularity. We use published power dissipation data from Intel's Single-Chip Cloud Computer (SCC), scaled to 32 nm, to calibrate our dynamic power data. The power traces generated by McPAT are given as inputs to the 3D-ICE tool [126] for transient thermal simulations (see Figure 29). Some of the key materials used in the construction of the 3D-

stack in the 3D-ICE tool and their properties are shown in Table 4. Additionally, we consider a heat sink adjacent to the core-cache layer for dissipation of heat in the environment.

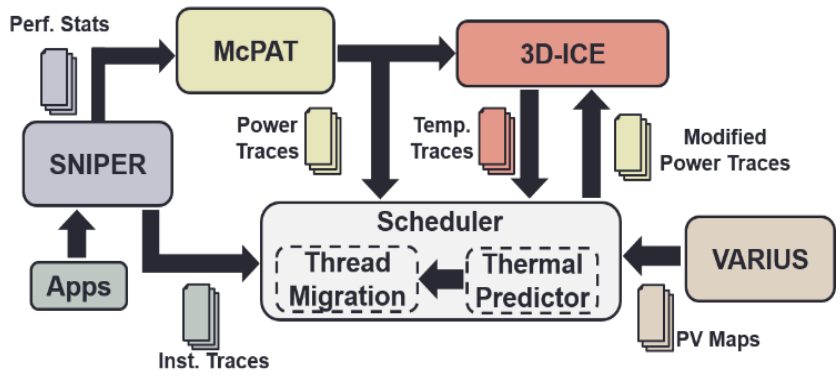


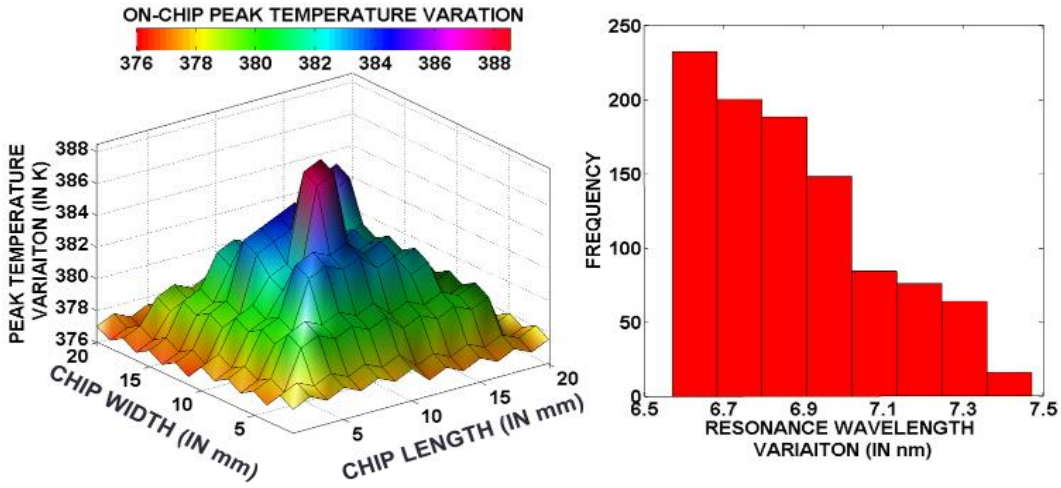
Figure 29: Simulation framework to analyze TV and PV in a manycore system with a PNoC architectures; the framework integrates performance, power, thermal, and variation simulators.

Table 4: Properties of materials used by 3D-ICE Tool [126].

Material	Thermal Conductivity	Volumetric Heat Capacity
Silicon	1.30e-4 W/µm K	1.628e-12 J/µm^3 K
Silicon di oxide	1.46e-6 W/µm K	1.628e-12 J/µm^3 K
BEOL	2.25e-6 W/µm K	2.175e-12 J/µm^3 K
Copper	5.85e-4 W/µm K	3.45e-12 J/µm^3 K

We analyzed the spatial variation in the peak temperatures of various tiles (at core-level granularity) of the photonic layer. For the 64-core system each tile has an estimated area of 6.25 mm^2 (i.e. 2.5×2.5 mm^2). We executed 64-threaded versions of the blackscholes (*BS*), bodytrack (*BT*), vips (*VI*), facesim (*FS*), fluidanimate (*FA*), swaptions (*SW*), barnes (*BA*), fft (*FFT*), radix (*RX*), radiosity (*RD*), and raytrace (*RT*) applications from the PARSEC and SPLASH2 benchmark suites on the 64-core system with one application running at a time. We monitored the peak temperature of each part of the photonic layer for every application and plotted the maximum peak temperature of each part across all the applications, as shown in Figure 30(a). From this figure, we can observe the maximum possible temperature-rise (above the room temperature) for any part of

the layer, which caps all possible dynamic TV values for that part. From Figure 30(a), higher peak temperatures are obtained at the center of the chip while relatively lower peak temperatures are achieved at the periphery of the chip. The main reason for the higher temperature at the center of the chip is the inefficiency of the heat sink to remove heat from the center of the chip. Furthermore, using Eq. (39) and (40), we determined the resonance wavelength shifts because of the peak temperature-rises, which are presented as a histogram in Figure 30(b). As evident from this figure, TV can induce up to a 7.4nm shift in MR resonances.

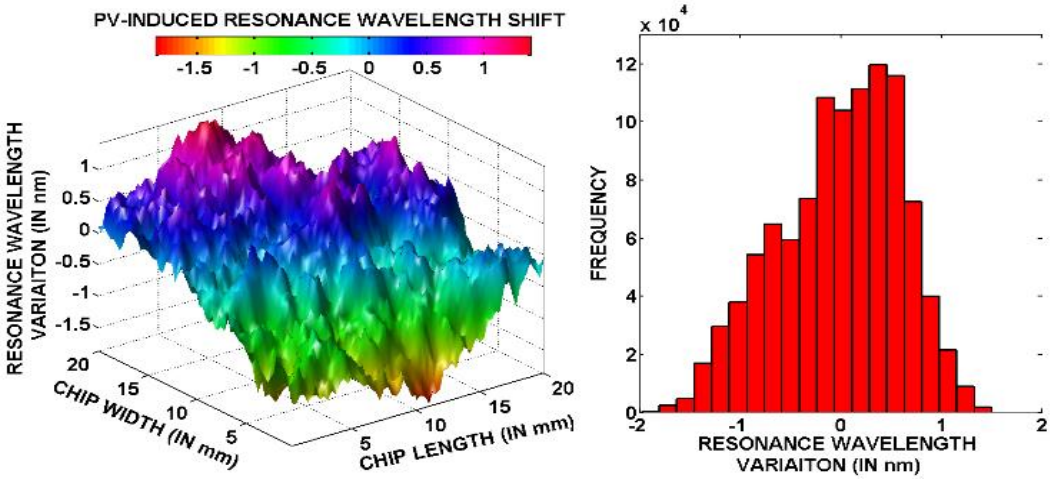


(a) **(b)**

Figure 30: (a) spatial variation in peak temperatures, (b) histogram of peak TV-induced resonance wavelength variation across a chip of size 400mm^2 using 3D ICE tool while executing 64 threaded PARSEC and SPLASH2 benchmark applications on a 64-core CMP.

In addition to TV, we also analyzed PV in PNoCs with the simulation setup presented in Figure 29. We adapted the VARIUS tool [71] to model die-to-die (D2D) as well as within-die (WID) process variations in MRs for the PNoC. VARIUS uses a normal distribution to characterize on-chip D2D and WID process variations. The key parameters are mean (μ), variance ($\sigma 2$), and density (α) of a variable that follows the normal distribution. As wavelength variations are approximately linear to the dimension variations of MRs, we assume they follow the same

distribution. The mean (μ) of wavelength variation of an MR is its nominal resonance wavelength. For PNoCs, we considered waveguides with 32 DWDM degree sharing the working band 1530–1625nm (i.e., C and L bands) with a wavelength channel spacing of 1.48nm. Hence, those wavelengths are the means for each MR modeled. The variance ($\sigma 2$) of wavelength variation is determined based on laboratory fabrication data [33] and our target die size. For a 64-core chip with 400mm^2 size at 32nm node, we consider a WID and D2D standard deviations of σWID = 0.61nm σD2D = 1.01nm, respectively [63]. We also consider a density (α) of 0.5 [63] for this die size. With these parameters, we use VARIUS to generate 100 process variation maps.



(a) **(b)**

Figure 31: (a) PV-induced resonance wavelength variation, (b) histogram of resonance wavelength variation across a chip of size 400mm^2.

We depict a PV map in Figure 31(a), which shows a spatial variation in PV-induced resonance wavelength shifts on the photonic die. Each PV map contains over one million points indicating the PV-induced shifts in MR resonances. The total number of points picked from these maps equal the number of MRs in the PNoC. We also present these points as a histogram in Figure 31(b). As evident from the histogram, PV can induce resonance wavelength shifts in the range of

-1.8nm to 1.6nm. However, we observed that this range can increase up to -3nm to 3nm for other PV maps.

4.4. OVERCOMING PV/TV INDUCED RESONANCE WAVELENGTH SHIFTS

The adverse effects of PV and TV, i.e., resonance shifts in MRs and their performance and reliability impacts, can be overcome by realigning and locking the resonance wavelengths of the individual MRs with the utilized carrier wavelengths. As PV is a static phenomenon, the PV-induced resonance shifts need to be overcome only once at system initialization. In contrast, due to the dynamic nature of TV, the TV-induced resonance shifts require runtime thermal stabilization of MRs. A stable locking of MR resonances with the utilized carrier wavelengths can be achieved using device-level (MR-level) mechanisms, such as localized trimming [21] and/or thermal tuning [22], with a dithering signal based feedback control [69]. However, the localized trimming and thermal tuning mechanisms proposed in prior work come with several challenges, which must be overcome to ease the adoption of PNoCs for future manycore systems.

First, thermal tuning and localized trimming mechanisms cannot provide sufficient tuning range to remedy PV/TV-induced resonance shifts in MRs. For instance, from Section 4.3, TV and PV together can induce shifts in MR resonance wavelengths of up to 10.4nm, i.e., 7.4nm for TV and ±3nm for PV. Therefore, compensating these TV/PV-induced resonance shifts would require a net tuning range of 10.4nm. But localized trimming can provide a tuning range of only 1.5nm at most [119]. In contrast, thermal tuning can provide a tuning range of about 6.6nm corresponding to the temperature range of up to 60K [69] at 0.11nm/K sensitivity [22]. Thus, even the thermal tuning and localized trimming together (i.e., 6.6nm+1.5nm tuning range) cannot provide the required tuning range of ~10.4nm. Another challenge for these mechanisms is their significant power overhead. A typical MR may consume 130μW of trimming power or 240μW of thermal

tuning power to remedy 1nm shift in its resonance wavelength, depending on its size, structure, and integration feasibility. To remedy a larger shift of ~10.4nm, a single MR may consume as much as ~1.35mW of trimming power or ~2.5mW of thermal tuning power. As a DWDM PNoC may have thousands of MRs, the total power overhead of PV/TV remedy can easily be in the range of a few tens of watts, which is a prohibitively high-power overhead for chip-scale systems and must be minimized to make the total power costs of large-scale DWDM PNoCs manageable.

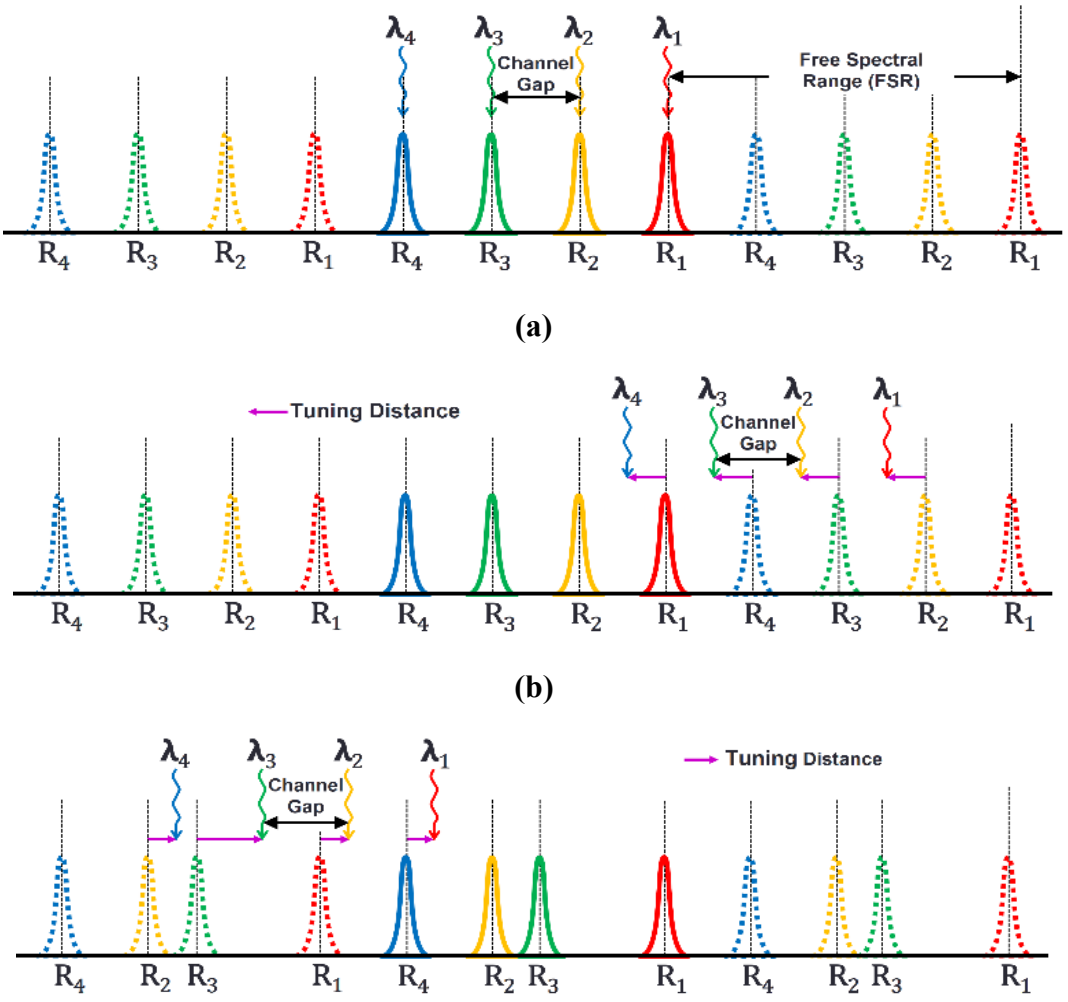


Figure 32: Periodic resonances *(R_1-R_4)* of an example bank of four MRs and their assigned carrier wavelengths *(λ_1-λ_4)* for (a) an ideal case with no resonance shifts, (b) a case with systematic blue-shifts in resonances, (c) a case with random red-shifts in resonances.

Fortunately, due to the periodicity of MR resonances, the resonance of none of the MRs in a PNoC needs to be tuned for more than a single channel gap [116]. This makes the required tuning range and the total tuning power more manageable. To understand this, consider Figure 32. The periodic resonances (R_1-R_4) of an example bank of four MRs and their assigned carrier wavelengths (λ_1-λ_4) for an ideal case with no PV or TV are shown in Figure 32(a). Due to the absence of PV/TV, the resonances of all MRs are aligned with their assigned carrier wavelengths. Figure 32(b) shows systematic blue-shifts of over two channel gaps in the resonances of all four MRs. In this case, the MR resonances can be re-aligned to their nearest carrier wavelengths followed by electrical repositioning of bits using backend barrel-shifters or pipelined shift registers [116]. In case the random PV throw the MR resonances out of order (Figure 32(c)), use of bit reordering multiplexers at the backend can still allow the MR resonances to be re-aligned to their nearest carrier wavelengths. Thus, due to the periodicity of MR resonances, and the use of bit reordering/repositioning techniques, the necessary tuning distance for the individual MRs reduces to less than one channel gap.

Our previously proposed SPECTRA framework [99] uses a different approach to reduce the required tuning distance and power overhead of PV/TV remedy. It integrates one system-level and two device-level optimizations. At the device-level, the SPECTRA framework utilizes three more MRs than the number of utilized carrier wavelengths, and thus, increases the available tuning range by three channel gaps. This mechanism reassigns the extra MRs to operate on nearby carrier wavelengths in the case when the resonances shift by less than three channel gaps. The need for remedying resonance shifts of more than three channel gaps is eliminated by reducing the range of temperature swings of the individual cores below the threshold levels that can induce resonance shifts of greater than three channel gaps. For that, an adaptive thread migration policy is used at

the system level, which also eliminates the need of bit-shifting. Moreover, SPECTRA adaptively chooses the least power-consuming method from *thermal tuning* and *localized trimming* as the preferred method for PV/TV remedy. Thus, SPECTRA conserves the total power required for PV/TV remedy with low latency overhead. However, the SPECTRA framework does not deal with PV and its benefits come with the area and power overheads of the extra MRs and bit-reordering multiplexers [116]. To address these shortcomings of the SPECTRA framework, we propose a new TV and PV aware reliability management framework called LIBRA, which is described next.

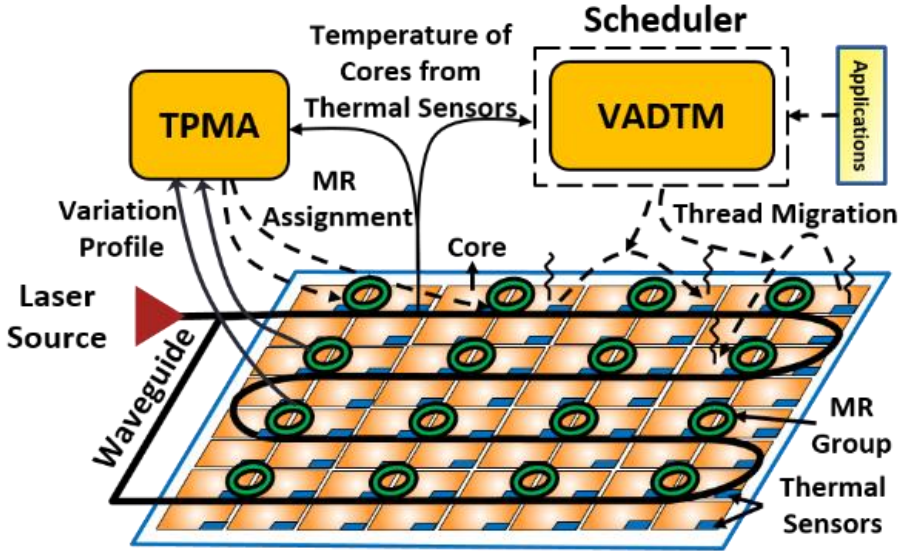


Figure 33: Overview of LIBRA framework that integrates a device-level thermal and process variation aware microring assignment mechanism *(TPMA)* and a system-level variation aware anti wavelength-shift dynamic thermal management *(VADTM)* technique.

4.5. LIBRA FRAMEWORK: OVERVIEW

Our *LIBRA* framework enables reliability-aware run-time PNoC management while rectifying TV and PV in MRs by integrating device-level and system-level enhancements. Figure 33 gives a high-level overview of our framework. The thermal and process variation aware microring assignment (*TPMA*) mechanism dynamically assigns each MR to the nearest available

carrier wavelength, which enables reliable modulation and reception of data while maintaining the maximum possible bandwidth. This device-level mechanism also adaptively chooses the least power-consuming method from thermal tuning and localized trimming as the preferred method for PV/TV remedy, and thus, reduces the total power for PV/TV remedy in the PNoC. However, limiting the peak temperature swings below threshold levels is critical to further reduce the total power for PV/TV remedy. To achieve this, we devise a PV-aware anti-wavelength-shift dynamic thermal management (*VADTM*) scheme that uses support vector regression (SVR) based temperature prediction and dynamic thread migration, to avoid on-chip thermal threshold violations, minimize on-chip thermal hotspots, and reduce thermal tuning power for MRs. The next two sections present details of the *TPMA* and *VADTM* schemes.

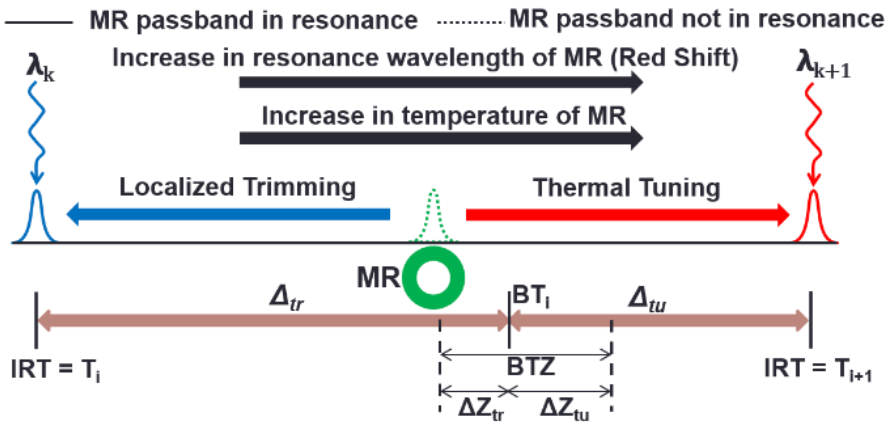


Figure 34: Red shift of MR with increase in temperature from IRTs T_i to T_{i+1} with trimming and tuning range of temperatures between these IRTs.

4.6. TV AND PV AWARE MICRORING ASSIGNMENT (TPMA)

4.6.1. TV AWARE MICRORING ASSIGNMENT (TMA)

As discussed in Section 4.3.1, TV shifts MR resonances, which can prevent MRs from reading or writing to their assigned carrier wavelengths. Fortunately, there is a linear dependency between temperature increase and resonance wavelength shift [119], which we exploit in our TV-

aware microring assignment (*TMA*) mechanism that dynamically assigns each MR to the nearest available carrier wavelength.

Figure 34 shows how at temperatures T_i and T_{i+1} ($T_{i+1} > T_i$), an MR resonance is in exact alignment with the available wavelengths λ_k and λ_{k+1}, respectively. These temperatures are called ideal resonant temperatures (IRTs). When the MR temperature is in between IRTs T_i and T_{i+1}, as shown in Figure 34, the MR needs to be either *trimmed* to resonate to λ_k (which is the resonance wavelength of an MR at temperature T_i) or thermally *tuned* to resonate to λ_{k+1} (which is the resonance wavelength of an MR at temperature T_{i+1}). To adaptively choose the least power consuming method from trimming and thermal tuning, we divide the temperature range between IRTs T_i and T_{i+1} into two parts: trimming temperature range (Δ_{tr}) and tuning temperature range (Δ_{tu}). For an MR at temperature T, if $(T_i + \Delta_{tr}) > T > T_i$ we perform trimming as it takes the least power, else if $(T_i + \Delta_{tr}) < T < T_{i+1}$ we perform tuning as it takes the least power (see Figure 34). At the boundary of the trimming and tuning temperature ranges, where $T_{i+1}-\Delta_{tu} = T_i+\Delta_{tr}$, both trimming and tuning consume equal power, and hence, an MR can be either trimmed or tuned. This temperature is called the boundary temperature (BT_i). It has been shown that for a small resonance wavelength shift (<1nm), thermal tuning power is higher compared to trimming power to mitigate the same amount of TV-induced shift [22]. Thus, our *TMA* approach considers a higher trimming temperature range compared to tuning temperature range $(\Delta_{tr}>\Delta_{tu})$, to minimize total trimming and tuning power.

In *TMA*, MRs are dynamically shifted (trimmed or tuned) to an appropriate IRT for correct operation based on their current temperature. Figure 35(a)-(d) show four different MR wavelength assignment configurations at successive IRTs T_1, T_2, T_3, and T_4, where $T_4>T_3>T_2>T_1$. If the MR group temperature T is such that $(T_1 - \Delta_{tu}) < T < (T_1+ \Delta_{tr})$ then the assignment in Figure 35(a) is

chosen, otherwise if $(T_2 - \Delta_{tu}) < T < (T_2 + \Delta_{tr})$, $(T_3 - \Delta_{tu}) < T < (T_3 + \Delta_{tr})$, or $(T_4 - \Delta_{tu}) < T < (T_4 + \Delta_{tr})$ then the assignment in Figure 35(b), Figure 35(c), or Figure 35(d) is chosen, respectively. One critical observation in the assignment shown in Figure 35(a) is that MRs R_1-R_n are in resonance with λ_1- λ_n within the same Free Spectral Range (FSR_i), whereas, in Figure 35(b) at IRT T_2, MRs R_2-R_n are in resonance with λ_1-λ_{n-1}, respectively in FSR_i and MR R_1 is in resonance with λ_n of the next FSR (i.e., FSR_{i+1}). In this assignment and the ones shown in Figure 35(c) and Figure 35(d), as explained in Section 4.4, there is a need to reposition bits in electrical domain using backend barrel-shifters or pipelined shift registers. The assignments shown in Figure 35(b), Figure 35(c), and Figure 35(d) require one, two, and three bit shifts, respectively, to retrieve the original data.

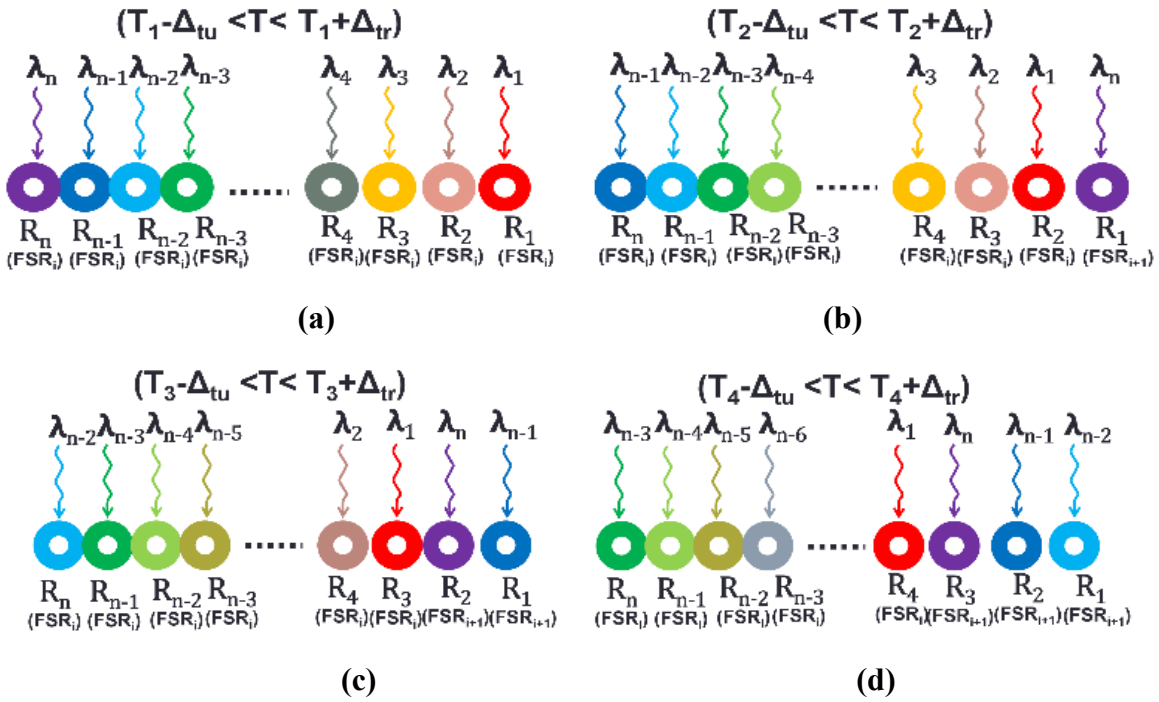


Figure 35: Thermal aware assignment of microrings *(R_{1-n})* to wavelengths *(λ_{1-n})* at four successive IRTs *T_1, T_2, T_3,* and *T_4* in *TMA* mechanism.

TMA represents a powerful reactive technique to adapt to on-die thermal variations with low overhead while ensuring reliable and high-bandwidth communication in MR based PNoCs. But

there is scope for three further enhancements. First, *TMA* does not consider the impact of PV on MRs, thus there is a need to readapt *TMA* to address the impact of PV on MRs, which is discussed in subsection 4.6.2. Second, there is a need to proactively control the peak on-chip temperature to reduce the range of on-chip temperature swings, which ultimately limits the number of required bit shifts (this work caps the number bit shifts to three as shown Figure 35(d)) and reduces the latency to retrieve the original data. Third, at the BT temperature (Figure 34), maximum trimming or tuning power is required to realign the MR resonances to their nearest carrier wavelengths. Thus, avoiding BT temperatures at MRs can reduce trimming and tuning power overhead. As shown in Figure 34, we define a boundary temperature zone (BTZ) around each BT_i. This zone includes temperatures T such that BT_i-ΔZ_{tr} < T < BT_i+ΔZ_{tu} where ΔZ_{tr} and ΔZ_{tu} are designer specified parameters. Cores with corresponding MR bank temperatures that are within BTZs are called boundary temperature cores (BTCs). As BTCs possess the highest trimming and tuning power overhead for their corresponding MR bank, a mechanism that reduces the number of BTCs can save trimming and tuning power. Section 4.7 describes such a mechanism, which also controls the range of on-chip temperature swings within allowable limits.

4.6.2. READAPTING TMA FOR PROCESS VARIATIONS (PMA)

In this subsection, we readapt the *TMA* mechanism to address the impact of PV on MR resonances. When using the *TMA* mechanism, PV-induced red or blue shift ($\Delta\lambda_{PV}$) alters the resonance wavelength (λ_{BTi}) of an MR at BT_i to λ_{BTR} or λ_{BTB}, respectively, as shown in Figure 36. This violates the actual definition of BT, which is the temperature from which either trimming to λ_k (which is the resonance wavelength of an MR at temperature T_i) or tuning to λ_{k+1} (which is the resonance wavelength of an MR at temperature T_{i+1}) dissipates equal power. For example, in case of a PV-induced blue shift, tuning λ_{BTB} to λ_{k+1} would consume more power than trimming it to λ_k,

as λ_{BTB} is shifted towards λ_k from λ_{BTi}. Therefore, in the *TMA* that is readapted for process variations (*PMA* mechanism), we propose to either increase or decrease the BTs in line with the PV-induced red or blue shifts in MR resonances, respectively.

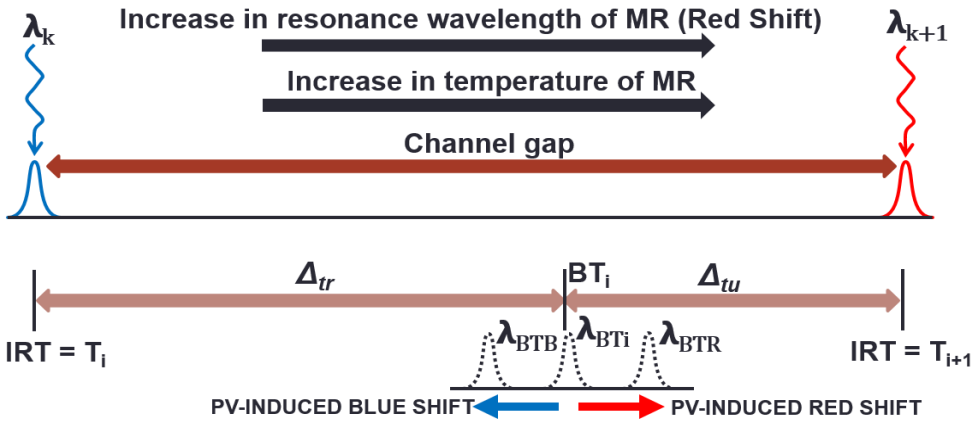


Figure 36: Impact of PV-induced red and blue shift on boundary temperature on *TMA*.

In our *PMA* mechanism, first, the PV-induced resonance shifts in MRs are gauged in situ at system initialization by using a dithering signal based control system [69]. The overhead of this in-situ PV detection technique is considered in our results section as dithering power. In our analysis, we model and estimate PV in MRs using the VARIUS tool [71], a description of which is already given in Section 4.3.3. Once PV-induced red or blue shifts of MRs are determined, we estimate the average resonance shift (in nm) across all MRs of each MR bank. We use each average shift value ($\Delta\lambda_{PV,ave}$) to determine the shift in BT (i.e., ΔBT_i) for all the MRs of the corresponding MR bank using Eq. (41), where TS is the MR thermal sensitivity obtained from Eq. (40) as $\Delta\lambda_r/\Delta T$.

$$\Delta BT_i = \frac{\Delta\lambda_{PV,ave}}{TS}, \tag{41}$$

Once the ΔBT_i values for all MR banks of the PNoC are obtained, we revise the BTs of each MR bank by either adding or subtracting the corresponding ΔBT_i value from the original BT.

Similar to the *TMA* mechanism, we then build BTZs around these updated BTs. Note that we cannot shift the original BT beyond a particular temperature range (i.e. $\Delta BT_i > \Delta_{tu}$ and $\Delta BT_i < -\Delta_{tr}$), especially when the PV-induced resonance wavelength shifts are greater than one channel gap (CG). Unfortunately, for state-of-the-art fabrication processes, the maximum PV-induced wavelength shifts are around ± 3nm (> one channel gap of 1.48nm). Shifting BT beyond a certain range to compensate for larger PV-induced shifts will also lead to higher tuning and trimming power dissipation.

Figure 37 shows an example of a larger PV-induced blue shift, which alters the resonance wavelength (λ_{BTi}) of an MR at BT to λ_{BTB}. One possible solution is to bring back the resonance wavelength to λ_{BTi}. But this is not always possible especially when the chip is operating at lower temperatures. Therefore, we propose to shift this λ_{BTB} to λ_{BTi-1} instead of λ_{BTi}, i.e., instead of decreasing BT by a larger amount here we increase BT by a smaller amount. In order to facilitate this shifting, similar to *TMA*, we perform ring assignment along with extra bit shifts. At a channel spacing of 1.48nm, to compensate for peak PV-induced resonance shift of ± 3nm, two extra bit shifts (forward and backward bit shifts to compensate positive and negative PV induced resonance shift) are needed.

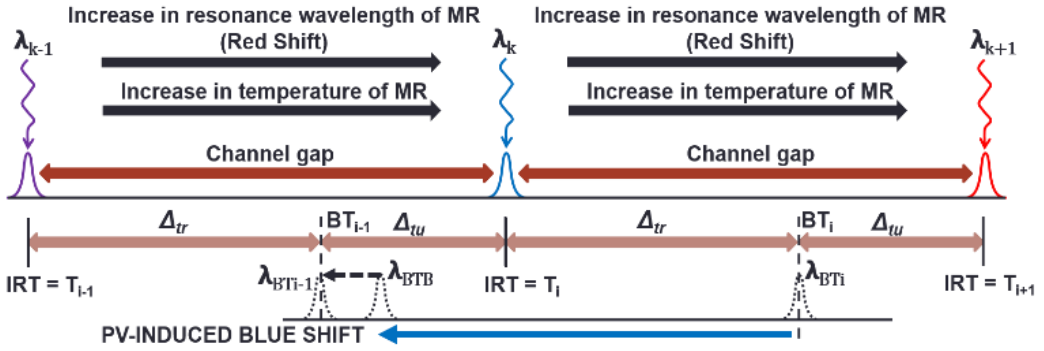


Figure 37: Boundary temperature adaptation for larger PV-induced blue shifts in *PMA*.

Overheads: Our proposed *TPMA* scheme is a combination of the two previously proposed techniques: *TMA* and *PMA* (this subsection). *TPMA* requires a maximum of five-bit shifts, which include three for *TMA* and two for *PMA*. These additional bit shifts in *TPMA* incur latency overhead. This latency overhead is quantified in more detail in Section 4.8. Furthermore, with *TPMA* each MR bank requires a Read Only Memory (ROM) to store its corresponding three BT values, which are determined using PV profiling at design time, as discussed earlier. This ROM also stores beginning and ending temperatures of three BTZs in each MR bank. We have considered 16-bits to store each temperature value. As there is a need to store nine different temperature values (three BTs, three BTZ start temperatures, three BTZ end temperatures) for each MR bank, we need a ROM that can store 144-bits. Moreover, a 16-bit comparator circuit is needed for each MR bank to determine the range of operation of MRs (i.e., trimming or tuning temperature range). This comparator is also used to determine whether an MR bank is in BTZ or not. Therefore, one input for this comparator comes from a thermal sensor (i.e., information on current temperature) and the other input is from the ROM. The area and power overhead of the ROM and comparator is quantified in detail in Section 4.8.

Table 5: List of VADTM parameters and their definitions.

Symbol	Definition
IPC_i	Instructions per cycle of i^{th} core
CT_i	Current temperature of i^{th} core
TN_i	Average temperature of immediate neighboring cores of i^{th} core; if this core is on chip periphery and missing neighbors, then we consider virtual neighbor cores at ambient temperature in lieu of the missing cores
PT_i	Predicted temperature of i^{th} core
T_t	Thermal threshold
BTCs	Boundary temperature cores
NBTCs	Non-boundary temperature cores

4.7. VARIATION AWARE ANTI WAVELENGTH-SHIFT DYNAMIC THERMAL MANAGEMENT (VADTM)

To proactively reduce thermal hotspots (which in turn will reduce instances of 'irrecoverable shift') and control on-die temperature (to reduce the number of BTCs), we propose a system-level variation aware anti wavelength-shift dynamic thermal management (*VADTM*) technique, described below.

4.7.1. OBJECTIVE

The primary goals with *VADTM* is to maintain the temperature of all of the cores on a die below a specified thermal threshold, i.e., for all cores $1 \leq i \leq N$, $T_i < T_t$ where T_i is the temperature of core i and T_t is threshold temperature. We utilize support vector based regression (SVR) to predict the future temperature of a core. This predicted temperature is compared with a thermal threshold to determine the potential for a thermal emergency. If such a potential exists, threads are migrated to available BTCs. These BTCs are determined based on the PV profile of MRs and ring blocks that are used to send and receive data from these cores. Migration to a BTC has a twofold benefit. First by moving the thread away from a core that could suffer a thermal emergency, we avoid instances of irrecoverable shift in the MR groups of that core. Second, by moving the thread to a BTC, the temperature of the BTC will increase resulting in that core no longer being a BTC (consequently the temperature of the core's MR groups will also increase, taking them outside of their BTZ and closer to IRTs, which will reduce trimming/tuning power). The parameters used to describe *VADTM* are shown in Table 5.

4.7.2. THERMAL MANAGEMENT FRAMEWORK

Figure 38 illustrates the entire *VADTM* technique. For each core, we periodically monitor the IPC value from performance counters and temperature from on-chip thermal sensors. If a

thermal emergency is predicted for a core by the SVR predictor, then *VADTM* initiates a thread migration procedure, otherwise no action is taken.

Algorithm 1: *VADTM* thread migration algorithm

Inputs: Current core temperature (CT_i), average neighboring core temperature (TN_i), current core IPC (IPC_i)

```
1:  for each core i do  // Loop that predicts future temperature
2:      PT_i = SVR_predict_future_temperature (CT_i, TN_i, IPC_i)
3:  end for
4:  for each core i do // Loop that checks for free BTCs and NBTCs
5:      if CT_i in BTZ and IPC_i == 0 then
6:          List_BTC = Push i  //add core to BTC list
7:      else if IPC_i == 0 then
8:          List_NBTC = Push i //add core to NBTC list
9:      end if
10: end for
11: for each core i do   // Loop that performs thread migration
12:     if PT_i ≥ T_t then
13:         if List_BTC ≠ {} then
14:             Migrated_core = Find_min_temperature_core(List_BTC)
15:             Do_thread_migration( core_i → Migrated_core)
16:             List_BTC = Pop i
17:         else if List_NBTC ≠ {} then
18:             Migrated_core = Find_min_temperature_core(List_NBTC)
19:             Do_thread_migration( core_i → Migrated_core)
20:             List_NBTC = Pop i
21:         end if
22:     end if
23: end for
```

Output: Thread migration to BTC or NBTC cores

Algorithm shows the pseudo-code for the *VADTM* thread migration procedure. First, the future temperature *(PT_i)* of the i^{th} core is predicted using the SVR based predictor with inputs: core temperature *(CT_i)*, core IPC *(IPC_i),* and temperature of neighboring cores *(TN_i)* in steps 1-3. The list of available BTCs (i.e., those that are not currently executing any thread) and available NBTCs is obtained in steps 4-10. In steps 11-12, a loop iterates over all cores and checks for possible thread migration conditions (i.e., thermal emergency cases where current core predicted temperature *(PT_i)* is greater than thermal threshold *(T_t)*). If a thread migration is required, then in steps 13-21, we check for free BTCs, and if they are available then we migrate the thread from the

current core to the BTC with lowest temperature, else we migrate the thread to a free NBTC with lowest temperature. This *VADTM* thread migration procedure is invoked at every epoch (1ms).

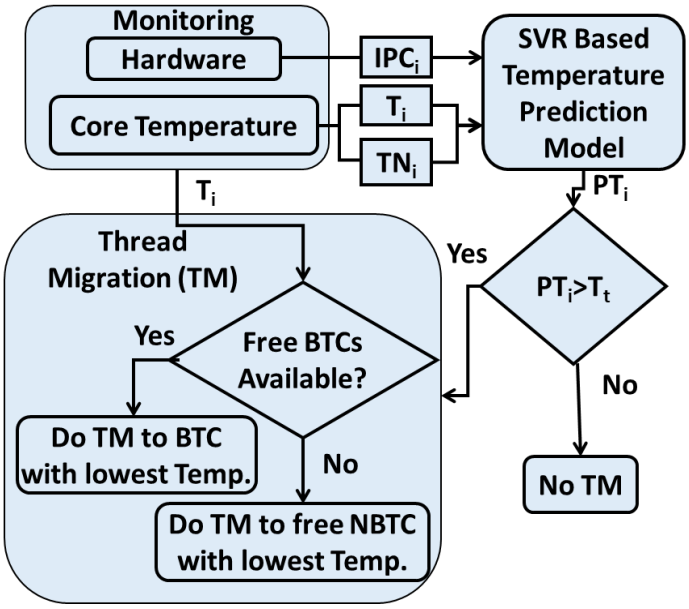


Figure 38: Overview of *VADTM* in *LIBRA* framework with support vector regression (SVR) based temperature prediction model.

4.8. EVALUATION

4.8.1. SIMULATION SETUP

We target a 64-core manycore system to evaluate our *LIBRA* (*TPMA+VADTM*) framework. Each core has a Nehalem x86 [123] micro-architecture with 32KB L1 instruction and data caches and a 256KB L2 cache, at 32nm and running at 5GHz. We evaluate *LIBRA* on two well-known PNoC architectures: Corona [14] and Flexishare [16]. Corona uses a 64×64 multiple write single read (MWSR) crossbar with token slot arbitration. Flexishare uses 32 multiple write multiple read (MWMR) waveguide groups with a 2-pass token stream arbitration. Each MWSR waveguide in Corona and MWMR waveguide in Flexishare is capable of transferring 512 bits of data from a source node to a destination node.

We modeled and simulated these architectures with the *LIBRA* framework for multi-threaded applications from the SPLASH-2 [124] and PARSEC [76] benchmark suites. Simulations were performed with a "warm-up" period of 100-million instructions and execution period of one billion cycles. Power and instruction traces for the benchmark applications were generated using the Sniper 6.0 [123] simulator and McPAT [125]. We used the 3D-ICE tool [126] for thermal analysis. The ambient temperature was set to 303K and the thermal threshold (T_t) was set to 353K.

We model and consider area, power, and performance overheads for our framework in our analysis. *LIBRA* with both Corona and Flexishare PNoCs has an electrical area overhead of 0.34 mm^2 and a power overhead of 57 mW using gate-level analysis and the CACTI 6.5 [78] tool for memory and comparators. The MR trimming power is set to 130μW/nm [21] for current injection (blue shift) and tuning power is set to 240μW/nm [22] for heating (red shift). To compute laser power, we considered detector responsivity as 0.8 A/W [80], MR through loss as 0.02 dB, waveguide propagation loss as 0.274 dB/cm, waveguide bending loss as 0.005 dB/90^0, and waveguide coupler/splitter loss as 0.5 dB [80]. We calculated photonic loss in components using these values, which sets the photonic laser power budget and correspondingly the electrical laser power. For energy consumption of photonic devices, we adapt parameters from [80], with 0.42pJ/bit for every modulation and detection event, and 0.18pJ/bit for the driver circuits of MR modulators and photodetectors.

We also considered thread migration overhead in our simulations that ranged from 500-1000 cycles to account for startup latency (extra cache misses, branch mispredictions) in the migrated core. Further, our simulations considered PNoC latency to transfer data from architectural registers from the source core to the migrated core. This latency depends on locations of the cores and traffic conditions. As presented in Section 4.5, to minimize trimming and tuning power consumption, for

a fixed channel gap of 1.48nm trimming temperature range (Δ_{tr}) and tuning temperature range (Δ_{tu}) for *TPMA* are calculated as 8.73K and 4.72K respectively. To minimize trimming and tuning power consumption further with lower performance overhead, there is a need to optimize ΔZ_{tr} and ΔZ_{tu} for *TPMA*. Therefore, we performed a sensitivity analysis to determine ΔZ_{tr} and ΔZ_{tu} values, as discussed in the next subsection.

4.8.2. SENSITIVITY ANALYSIS

Our first set of experiments involves a sensitivity analysis to explore the impact of the ΔZ_{tr} and ΔZ_{tu} parameters on *LIBRA*. We analyzed trimming and tuning power dissipation and execution time of the Flexishare PNoC with different values of these parameters. To be consistent with Δ_{tr} and Δ_{tu}, we consider the ratio of ΔZ_{tr} and ΔZ_{tu} to be equal to the ratio of Δ_{tr} and Δ_{tu}. For a fixed channel gap (i.e., 1.48nm), as presented above, the ratio of Δ_{tr} and Δ_{tu} is constant. Therefore, we determine the optimal ΔZ_{tu} with a sensitivity analysis and then we use that value to determine ΔZ_{tr}.

We considered 48-threaded *FS*, *FA*, and *BS* benchmark applications for our sensitivity analysis. Figure 39 shows the decrease in trimming and tuning power (TP) and increase in application execution time (ET) for the *LIBRA* framework while executing three benchmark applications on the Flexishare PNoC, with ΔZ_{tu} varying from 0.4K to 4K. We computed the decrease in TP and increase in ET with respect to baseline Flexishare PNoC architecture employing the FATM thread scheduling policy [11]. In this analysis, we presented results that are averaged across 100 PV maps. The three benchmarks were chosen as they resulted in high (*FS*), medium (*FA*), and low (*BS*) peak temperatures, which allowed us to explore the impact of thread migration overheads on ΔZ_{tu}. At a particular ΔZ_{tu}, this figure shows higher TP savings for high peak temperature workloads (i.e., *FS*) compared to low peak temperature workloads (i.e., *BS*), as *LIBRA* effectively controls peak temperature and thereby reducing overall TP. Also, the percentage of

increase in application execution time is higher for high peak temperature workloads (i.e., *FS*) compared to low peak temperature workloads (i.e., *BS*), as *LIBRA* incurs more number of thread migrations for these workloads, which ultimately increases ET.

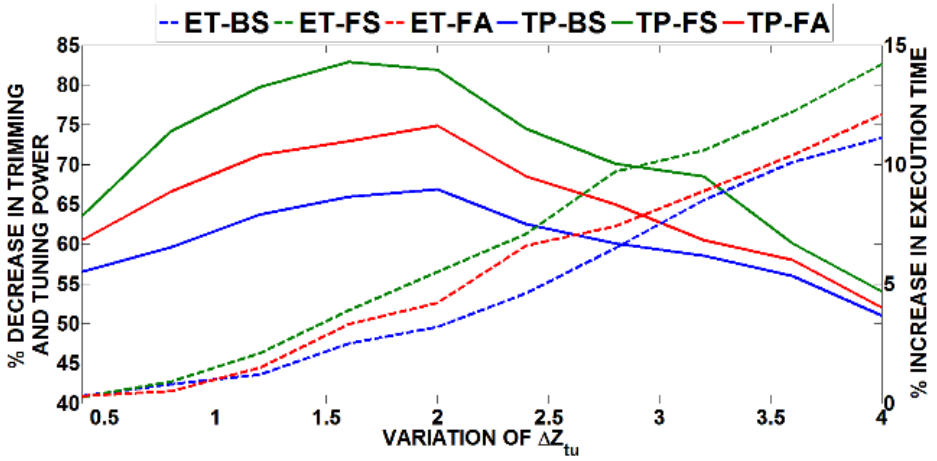


Figure 39: Percentage of decrease in trimming/tuning power (TP) and percentage of increase in execution time (ET) comparison across different ΔZ_{tu} values for *LIBRA* framework implemented on Flexishare PNoC in a 64-core CMP executing blackscholes (*BS*), Facesim (*FS*), and Fluidanimate (*FA*). Presented results are averaged across 100 PV maps. All percentage increments/decrements are calculated w.r.t baseline Flexishare PNoC employing frequency align scheduling policy (FATM).

A careful observation of Figure 39 shows that for all the benchmark applications, LIBRA's TP decreases with initial increase in ΔZ_{tu} and increases with further increase in ΔZ_{tu}. The main reason for this behavior is that at smaller values of ΔZ_{tu} *LIBRA* benefits by increasing temperature of BTCs, which ultimately reduces the number of MR groups within BTZs. Furthermore, larger values of ΔZ_{tu} increase BTZ size and the number of BTCs within it, so there is more chance that threads are migrated to cores whose temperatures are away from their BTs, which reduces the percentage of decrease in trimming and tuning power (TP; see Figure 39). Moreover, with increase in ΔZ_{tu} the number of thread migrations increase as more number of BTCs are available, which ultimately increases total execution time of the application (TP; see Figure 39). Thus, we set ΔZ_{tu}

to 2K, to achieve higher TP savings with lower ET overhead. Using ΔZ_{tu}, as explained above, we determined ΔZ_{tr} as 3.7K. We used these values of the ΔZ_{tu} and ΔZ_{tr} parameters for our *LIBRA* framework in the rest of our analysis.

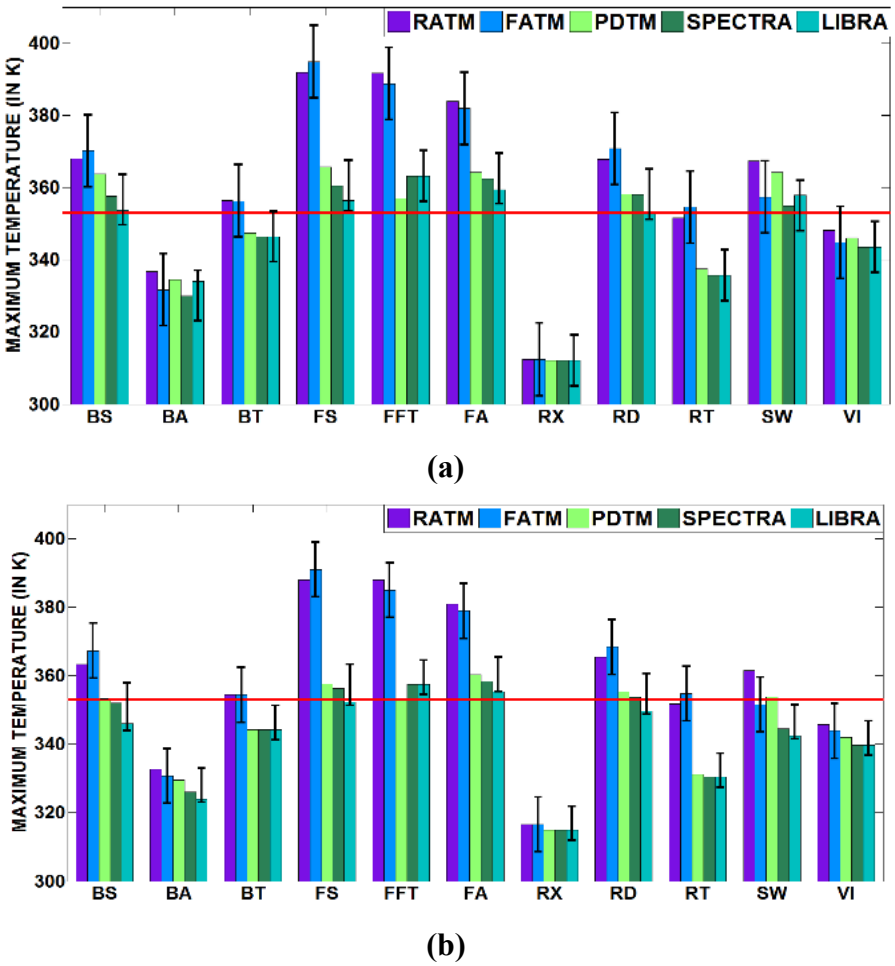


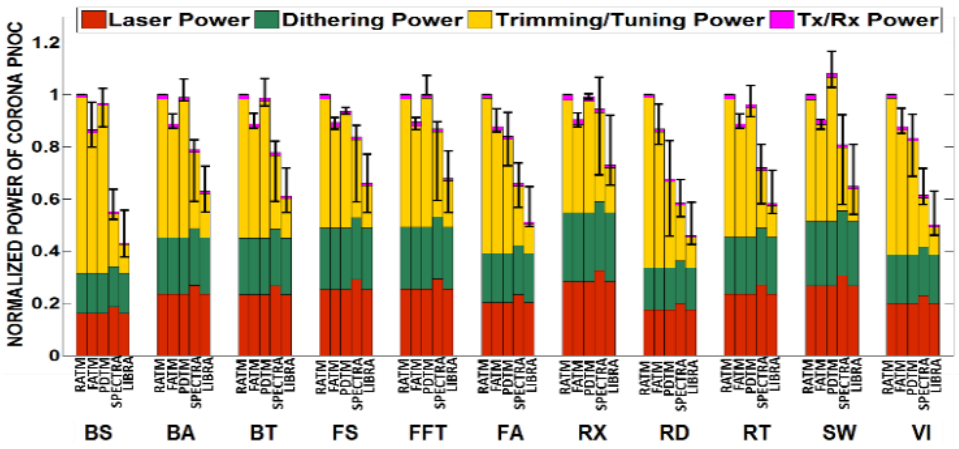
Figure 40: Maximum temperature comparison for *LIBRA* with RATM [133], FATM [145], PDTM [139] and SPECTRA [33], for (a) 48 thread, and (b) 32 thread PARSEC and SPLASH-2 benchmarks executing on 64-core manycore system with Corona PNoC. Bars show mean values of maximum temperature across 100 PV maps; confidence intervals show variation in maximum temperature.

4.8.3. COMPARISON RESULTS

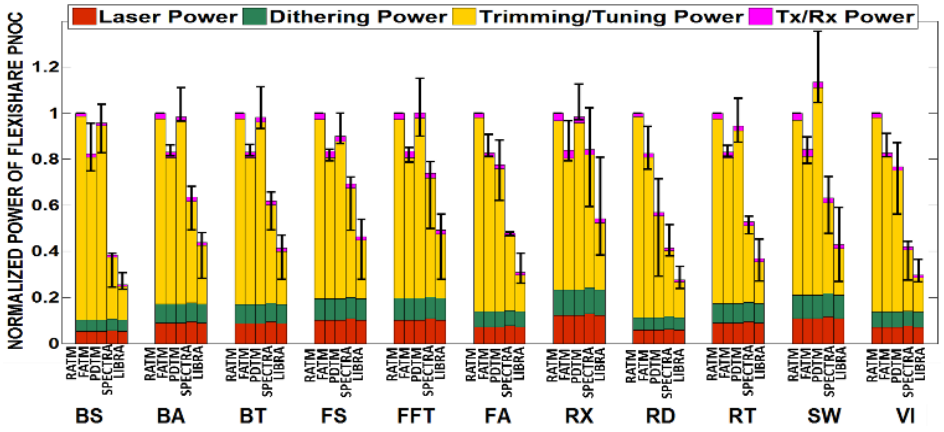
We compared the performance of our *LIBRA* framework with four prior works on manycore thermal management: a ring aware policy (RATM) [100], frequency align policy (FATM) [11], a

predictive dynamic thermal management (PDTM) framework [101], and the SPECTRA framework from our prior work [99]. RATM distributes threads uniformly across cores that are closer to PNoC nodes first and then distributes the remaining threads in a regular pattern from outer cores to inner cores. FATM distributes threads across cores based on the process variation profile of ring blocks that are in the proximity of these cores. PDTM uses a recursive least square based temperature predictor to determine if the predicted temperature of a core exceeds a thermal threshold, and if so then thread migration is performed from that core to the coolest core that is not executing any threads. SPECTRA performs ring assignment at the device-level and SVR prediction based proactive thread migration at the system-level for thermal reliability management in PNOCs.

Figure 40(a)-(b) show the maximum temperature obtained with the five frameworks across eleven applications from the PARSEC and SPLASH-2 benchmarks suites with 48 and 32 thread counts executing on a 64-core system with the Corona PNoC [14], [59] architecture. As *LIBRA* and FATM perform thread management based on the PV profile of MRs, only these frameworks have confidence intervals in Figure 40. From Figure 40(a) it can be observed that some applications (e.g., *FA* and *SW*) with 48 threads exceed the threshold (353K) for all frameworks, as there are insufficient number of free cores on the chip whose temperature is below the thermal threshold to migrate threads. However, with a more manageable number of threads, the situation improves. In Figure 40(b), for the case with 32 threads, our *LIBRA* framework avoids violating thermal thresholds for very small number of benchmark applications with 32 threads. On average, *LIBRA* has 14.6K and 17.5K lower maximum temperature compared to the RATM policy for 48 and 32 threads, respectively.



(a)



(b)

Figure 41: Normalized power dissipation (Laser Power, Dithering Power, Trimming/Tuning Power, and Modulating and Detecting (Tx/Rx) Power) comparison for *LIBRA* with RATM [133], FATM [145], PDTM [139] and SPECTRA [33] for 48 threaded applications of PARSEC and SPLASH-2 suites executed on (a) Corona, (b) Flexishare PNoC architectures for a 64-core manycore system. Results shown are normalized wrt RATM, therefore, RATM does not have confidence intervals. Bars show mean values of power dissipation across 100 PV maps; confidence intervals show variation in power dissipation.

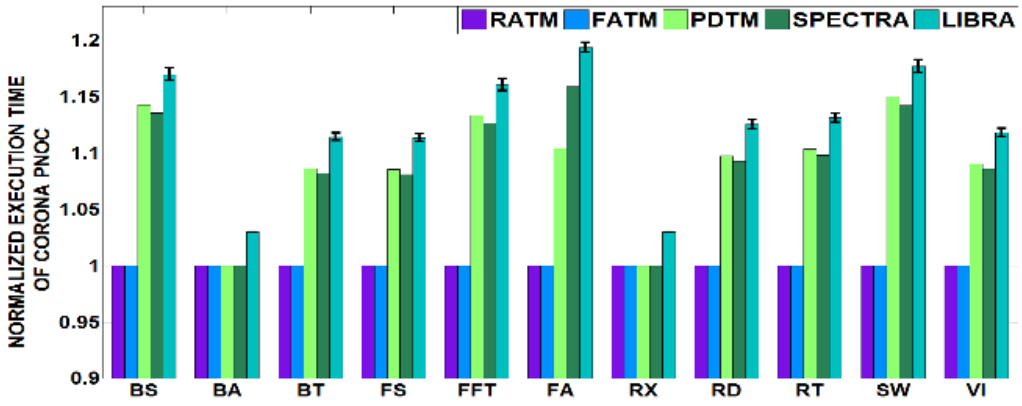
In addition, on average *LIBRA* has 13.5K and 16.9K lower maximum temperature compared to the FATM policy for 48 and 32 threads, respectively. *LIBRA* migrates threads from hotter cores to cooler cores to control maximum temperature, whereas no thread migration is performed in both RATM and FATM when the on-chip thermal threshold temperature (i.e., 353K) is reached, as these mechanisms are simple thread allocation policies without control on peak temperature. For most

of the benchmarks, maximum temperatures with PDTM, SPECTRA, and *LIBRA* are below the thermal threshold. However, on average *LIBRA* has 3.2K and 3.5K lower maximum temperature compared to PDTM for 48 and 32 threads, respectively. This is because *LIBRA* employs a more accurate SVR based prediction approach which reduces the increase in peak temperature due to mispredictions, compared to the low accuracy of the least square regression mechanism in PDTM. Lastly, *LIBRA* has a 0.8K and 1.9K lower maximum temperature compared to SPECTRA for 48 and 32 threads, respectively. Even though both *LIBRA* and SPECTRA prefer to migrate threads to BTCs, the maximum temperatures with *LIBRA* are sometimes lower compared to SPECTRA, as *LIBRA* is able to perform thread migrations more often to lower temperature BTCs compared to SPECTRA.

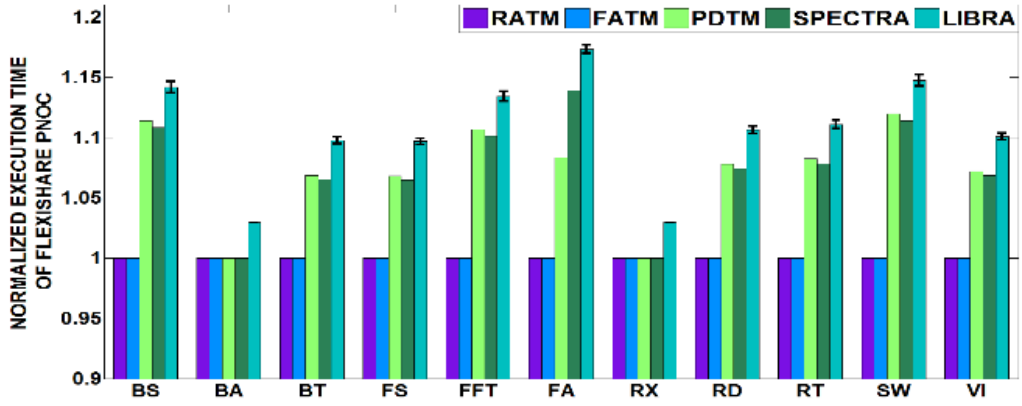
In the interest of brevity, we do not show maximum temperature results for the Flexishare PNoC architecture. We observed a similar trend in maximum temperature variations for Flexishare as we did for Corona (Figure 40).

Figure 41 shows the power dissipation comparison for the five frameworks across multiple 48-threaded applications for the Corona and Flexishare PNoC architectures, respectively. One of the main reasons why *LIBRA* has lower power dissipation than RATM, FATM, and PDTM is that it more aggressively reduces trimming and tuning power in both Corona and Flexishare PNoCs. From Figure 41(a), *LIBRA* has 74.5%, 67.4%, and 70.8% lower trimming and tuning power on average compared to RATM, FATM, and PDTM for Corona. Furthermore, from Figure 41(a), *LIBRA* also has 76.2%, 68.3%, and 72.5% lower trimming and tuning power on average compared to RATM, FATM, and PDTM for Flexishare. The *TPMA* technique in *LIBRA* intelligently conserves trimming and tuning power compared to RATM, FATM, and PDTM by performing process variation aware MR reassignment with increase in temperature, while our *VADTM* further

improves trimming and tuning power savings with its intelligent thread migration to BTCs. Lastly, the *TPMA* mechanism in *LIBRA* adapts intelligently to the PV profiles of MRs, reducing it's trimming and tuning power dissipation by 46.3% and 48.1%, compared to SPECTRA for the Corona and Flexishare architectures, respectively.



(a)



(b)

Figure 42: Normalized average execution time comparison of *LIBRA* with RATM [133], FATM [145], PDTM [139], and SPECTRA [33] for (a) Corona, (b) Flexishare PNoCs for 48 threaded applications from PARSEC and SPLASH-2 suites executed on 64-core system. Results shown are normalized wrt RATM. Bars show mean values of execution time across 100 PV maps; confidence intervals show variation in execution time.

Figure 41 also shows the laser power comparison of the five frameworks for the Corona and Flexishare architectures. It can be observed that Corona and Flexishare with *LIBRA* need similar laser power as Corona and Flexishare architectures with RATM, FATM, and PDTM. Furthermore,

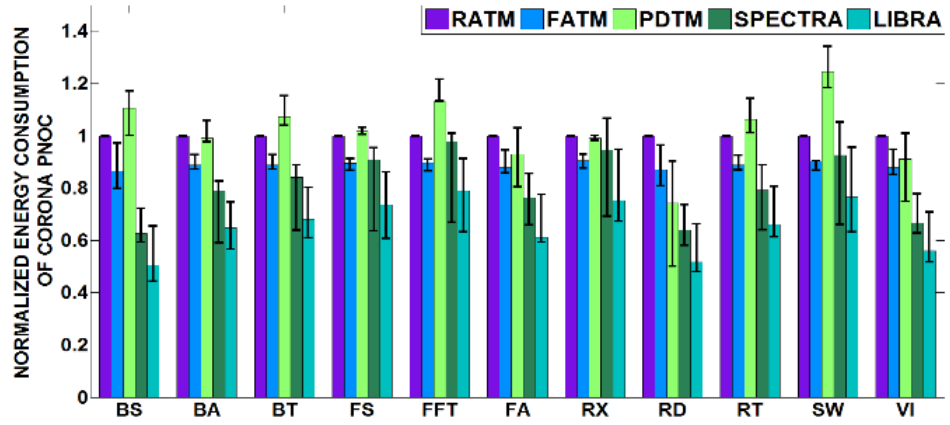
LIBRA requires 12.9% and 6.4% lesser laser power compared to SPECTRA for Corona and Flexishare. The extra MRs used in SPECTRA to compensate for TV-induced resonance shifts contribute to the increase in laser power compared to *LIBRA* for both architectures. From these results it can also be observed that the laser power saving in Corona is higher than for the better performance optimized architecture of Flexishare.

In summary, *LIBRA* saves considerable trimming/tuning power to ultimately achieve overall power reduction. From the power analysis in Figure 41(a), *LIBRA* with Corona has 40.8%, 34.1%, 37.2%, and 21.4% lower total power dissipation compared to Corona with RATM, FATM, PDTM, and SPECTRA, respectively. Further from Figure 41(b) it can be seen that Flexishare with *LIBRA* has 61.3%, 52.9%, 57.4%, and 32.8% lower power dissipation compared to Flexishare with RATM, FATM, PDTM, and SPECTRA, respectively.

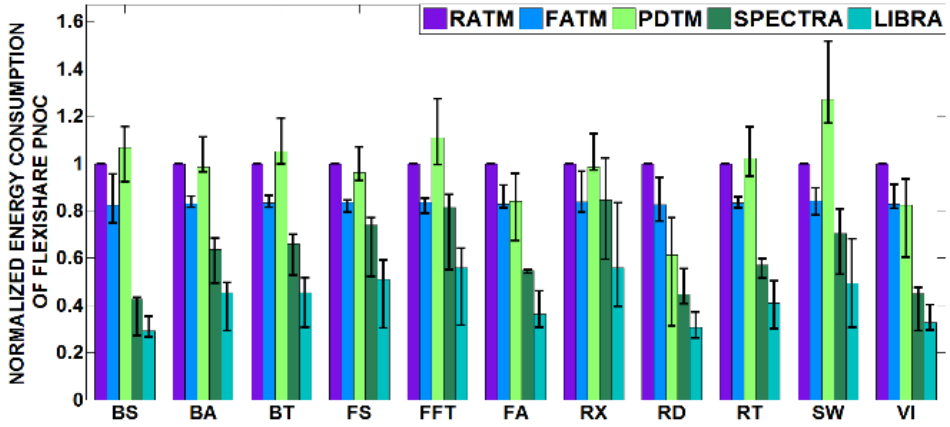
Figure 42 shows the average execution time comparison between the five frameworks across the 11 48-threaded applications from PARSEC and SPLASH-2 suites, for the Corona and Flexishare PNoC architectures, respectively. As only *LIBRA* performs thread migration based on the PV profile of MRs, therefore, this framework has confidence intervals on execution time shown in Figure 42.

From Figure 42(a), it can be seen that Corona with *LIBRA* has 12.4% higher execution time compared to Corona with RATM and FATM. Corona with *LIBRA* needs extra execution time to migrate threads between cores and to reorder bits using shift registers whereas the RATM and FATM policies simply schedule threads without any thread migration and bit reorder, and thus do not possess such overheads. Further, Corona with *LIBRA* has 3.2% higher execution time compared to PDTM. Despite *LIBRA* using a faster SVR based temperature predictor compared to a more complex recursive least square based regression predictor in PDTM, the higher number of

thread migrations (to adapt to PV profiles of MRs) and bit reordering operations in *LIBRA* contribute to an increase in execution time.



(a)



(b)

Figure 43: Normalized energy consumption comparison of *LIBRA* with RATM [133], FATM [145], PDTM [139], and SPECTRA [33] for (a) Corona, (b) Flexishare PNoCs for 48 threaded applications from PARSEC and SPLASH-2 suites executed on a 64-core system. Results shown are normalized wrt RATM, therefore, RATM does not have confidence intervals. Bars show mean values of energy consumption across 100 PV maps; confidence intervals show variation in energy consumption.

Similarly, from Figure 42(b) Flexishare architecture with *LIBRA* has 10.6% higher execution time compared to Flexishare with RATM and FATM. In addition, *LIBRA* also has 2.8% higher execution time compared to Flexishare with PDTM. The figures also indicate that the execution

time overheads for *LIBRA* are lower when utilizing the faster Flexishare architecture compared to the slower Corona architecture. Moreover, the bit-shifting overhead in *LIBRA* increases its execution time by 4.2% and 3.2% compared to the SPECTRA framework with Corona and Flexishare PNoCs, respectively. From the execution time results, it can be summarized that the significant power benefits achieved with *LIBRA* come at some cost: an increase in execution time.

Lastly, from the power dissipation and execution time results, we obtain energy consumption results for the five frameworks, as shown in Figure 43. On average, for Corona, energy consumption with *LIBRA* is 34.5%, 25%, 35.4%, and 18.7% lower compared to RATM, FATM, PDTM and SPECTRA, respectively. For the Flexishare architecture, *LIBRA* has 57.3%, 47.9%, 55.6%, and 31.1% lower energy consumption compared to RATM, FATM, PDTM, and SPECTRA respectively.

In summary, from the above results, it is apparent that our proposed PV-aware *LIBRA* framework outperforms previously proposed approaches for thermal management in manycore systems with PNoCs by combining a novel reactive device-level technique (*TPMA*) that improves waveguide channel utilization with a novel system-level proactive thread migration technique (*VADTM*). The excellent power and energy savings compared to previous approaches strongly motivate the use of our thermal management framework in future PNoC based manycore architectures.

4.9. CONCLUSIONS

In this chapter, we presented *LIBRA* framework that combines two novel dynamic thermal management mechanisms for the reduction of maximum on-chip temperature and conservation of trimming and tuning power of MRs in DWDM-based PNoC architectures. These techniques (*TPMA* at the device-level, *VADTM* at the system-level) constitute a hybrid reactive-proactive

management framework that demonstrates interesting trade-offs between performance and power/energy across two different state-of-the-art crossbar-based PNoC architectures. Our experimental analysis on the well-known Corona and Flexishare PNoC architectures showed that *LIBRA* can notably conserve total power by up to 61.3% (trimming and tuning power by up to 76.2%) and total energy by up to 57.3%.

5. A COMPARATIVE ANALYSIS OF FRONT-END AND BACK-END COMPATIBLE SILICON PHOTONIC ON-CHIP INTERCONNECTS

Photonic devices fabricated with back-end compatible silicon photonic (BCSP) materials can provide independence from the complex CMOS front-end compatible silicon photonic (FCSP) process, to significantly enhance photonic network-on-chip (PNoC) architecture performance. In this chapter, we present a detailed comparative analysis of a number of design tradeoffs for CMOS front-end and back-end compatible devices for silicon photonic interconnects. A cross-layer optimization of multiple device-level and link-level design parameters is performed to enable the design of energy-efficient on-chip photonic interconnects using BCSP devices. The optimized design of BCSP on-chip links renders more energy-efficiency and aggregate bandwidth than FCSP on-chip links, in spite of the inferior opto-electronic properties of BCSP devices. Our experimental analysis compares the use of BCSP and FCSP links at the architecture level, and shows that the optimized design of the BCSP-based Firefly PNoC achieves 1.15× greater throughput and 12.4% less energy-per-bit on average than the optimized design of FCSP-based Firefly PNoC. Similarly, the optimized design of the BCSP-based Corona PNoC achieves 3.5× greater throughput and 39.5% less energy-per-bit on average than the optimized design of FCSP-based Corona PNoC.

5.1. INTRODUCTION

Recent advances in silicon photonics (SiP) based on the silicon-on-insulator (SOI) process have produced high performance building blocks such as modulators, detectors, filters, and switches that are highly desirable for high-bandwidth and energy-efficient on-chip photonic interconnects [127]-[130]. However, the SOI platform restricts SiP circuits to a single layer, which limits the number of devices that can fit on a chip. Also, the modern SOI process offers a very thin

layer of buried oxide (BOX) (200nm thick BOX at 45nm and thinner for advanced technology nodes), which does not provide the necessary optical isolation required to guide light into SiP devices, resulting in large optical losses due to scattering [131]. To address these issues, recent efforts have proposed back-end integration of SiP devices with CMOS logic. In [132], electro-optic polymer and germanium, and in [133] III–V compounds are used as the active materials. However, fabrication of SiP devices using polymer based or III-V compound based materials requires heterogeneous integration with CMOS logic, which is very costly, requiring specialized foundries.

As a solution to these limitations, Lee et al. in [131] discussed the use of back-end compatible silicon nitride (SiN) material to produce low-loss passive optical waveguides and the use of excimer laser annealed (ELA) quasi-single-crystalline polysilicon (pSi) and polycrystalline germanium (Ge) to produce active microring modulators and detectors. Traditionally, the photonics community has largely ignored pSi due to the challenges introduced by its high optical losses and inferior electrical properties. Similarly, the stress issues complicating the deposition of SiN films thick enough for guiding in the telecom wavelength range have limited the use of low-loss SiN waveguides only for visible wavelengths [131]. However, recent advances in back-end integration technology have led to several pSi and SiN devices being demonstrated with performance and loss values comparable to front-end integrated crystalline silicon (cSi) devices [134]-[137].

In this chapter, we refer to SiP devices made of pSi and SiN materials as back-end compatible SiP (BCSP) devices, whereas we refer to SiP devices made of front-end integrated cSi material as front-end compatible SiP (FCSP) devices. BCSP devices provide independence from complex CMOS front-end processes. Moreover, the possibility of low-temperature multi-layer

deposition of pSi and SiN materials on top of CMOS metallization layers, as demonstrated in [135], enables multi-level integration for 3D photonic networks-on-chip (PNoCs) on a logic chip. Thus, BCSP has a multitude of benefits over FCSP, which favors the use of such devices in the PNoCs of the future.

The design and characteristics of active and passive SiP devices control the feasibility, reliability, and performance of the entire SiP PNoC. Therefore, the designers of PNoCs should follow a strict set of device-level design guidelines to ensure good system performance. Existing device-level design guidelines, as presented in [138] and [139], are prepared for FCSP-based devices and systems. But the optical and electrical properties of BCSP devices are different from those of FCSP devices [131], which implies that a distinct set of design guidelines are required for BCSP systems. *For the first time, in this work we analyze a number of device-level tradeoffs for BCSP devices to derive design guidelines for BCSP-based PNoC architectures.*

From our analysis of device-level tradeoffs, we observed that the design of energy-efficient, low-noise, and high-aggregate-bandwidth BCSP interconnects requires cross-layer optimization of a number of interdependent device-level and link-level parameters. In recent years, several works have discussed such cross-layer optimization of parameters for FCSP interconnects [57], [73], [140]-[142]. In [140], the impact of fabrication-induced process variations and power-induced thermal variation on FCSP devices and its impact on the reliability, power dissipation, and performance of FCSP PNoCs was studied. Mohamed et al. in [141] presented analytical models of FCSP devices and analyzed the design tradeoffs for their applications at the network level. In [73], a high-aggregate-bandwidth microring link is analyzed to determine energy-efficiency and bandwidth-density for the link using best-of-class FCSP devices. Hendry et al. in [57] present physical layer analysis and modeling of FCSP-based dense wavelength division

multiplexed (DWDM) bus architectures. In [142] and Chapter 10, optimized photonic link architectures comprised of FCSP devices are used to achieve high-bandwidth and energy-efficient data transfers between core and off-chip memory. Unlike any of these prior works, we perform a cross-layer analysis of design tradeoffs for BCSP interconnects and compare the results of this analysis with the results of a similar analysis for FCSP interconnects. Our results provide a better understanding of available design choices for realizing energy-efficient and terabyte-per-second scale PNoCs.

We summarize the key contributions in this chapter as follows:

- We present and analyze a number of device-level design tradeoffs for BCSP devices involving Q-factor, optical power loss in microring cavity, and modulator bit-rate as a function of radius;
- We characterize interdependence between various device-level and link-level design parameters of BCSP devices, and perform cross-layer optimization of these parameters, to realize energy-efficient and high-aggregate-bandwidth BCSP on-chip links;
- We perform a similar cross-layer analysis and optimization for FCSP devices and compare results with those for BCSP devices;
- We evaluate the impact of optimized designs of FCSP and BCSP links on the performance and energy-efficiency of two well-known PNoC architectures: Corona [59] and Firefly [15].

5.2. ANALYSIS OF DESIGN TRADEOFFS

A typical PNoC consists of microring resonators (MRs) that are coupled to one or more photonic DWDM bus waveguides (WGs) [81], [143], [144]. These MRs serve as modulators, filters, and switches. We direct the reader to [141] for more details on MR design and operation.

The feasibility, reliability, energy-efficiency, and performance of PNoCs depend on various device-level and link-level design parameters. Our goal in this section is two-fold: (1) to understand and analyze the tradeoffs present among various device-level and link-level design parameters of PNoCs; (2) to understand how these tradeoffs differ between BCSP and FCSP based PNoCs. As a first step towards achieving these goals, we present analytical models of BCSP and FCSP devices (Section 5.2.1). Then, using these models, we analyze the tradeoffs among various device-level (Section 5.2.2) and link-level (Section 5.2.3) design parameters for BCSP and FCSP devices.

Table 6: Definitions and typical values of some constants for MRs.

	Definition	Value	
		BCSP	FCSP
n_{eff}	Effective refractive index of MR [145]	2.49	2.45
n_g	Group refractive index of MR [145]	4.26	4.21
n_{SiO2}	Refractive index of SiO_2 cladding [134]	1.48	
n	Refractive index of an MR's looped WG core [146] [147]	pSi	cSi
		3.48	3.47
$C1$	Coefficients based on the material and geometry of MR [145]	132	126
$C2$		10	10.1
R_s	Series resistance of MR [134] [148] (in Ω)	750	250
α_i	Intrinsic optical loss due to bulk defects and surface roughness in MR [134][149] (in cm^{-1})	3.87	2
α_d	Optical absorption loss in MR (in cm^{-1})	0.23	0.23
α_b	Bending loss due to MR curvature	Eq. (4)	
-	Cross-section dimensions of MR's looped WG	450nm×250nm	

5.2.1. BCSP AND FCSP DEVICE MODELING

In a PNoC, MRs are coupled to one or more DWDM bus WGs, and serve as modulators, filters, and switches [141]. For a passive component such as a filter, the MR can be considered as a looped photonic WG with a small diameter. For an active component such as a modulator, the MR's looped WG is doped such that it may be addressed as a PN junction device. The tradeoffs

among the design parameters of a DWDM bus WG are mostly straightforward. In contrast, the resonant nature of an MR creates several complex tradeoffs among its design parameters. For this reason, in this subsection we present analytical device models for passive and active MRs. These models are equally relevant for BCSP and FCSP types of MRs, as they both have similar geometry, and work on the same principle.

Models for Passive Microring Resonators: A passive MR acts as a bandpass filter, the characteristics of which are defined by the resonant wavelength (λ_r), round-trip optical loss (a^2), and Q-factor. The Q-factor of a passive MR that is coupled to a WG is known as loaded Q-factor Q_L [64], which is inversely proportional to the full width of its passband at half the maximum (FWHM) transmission. The Q_L, a^2, and λ_r parameters, assuming a critical coupling of the MR to a WG, can be expressed as [64], [68]:

$$Q_L = \frac{2\pi^2 n_g R a}{\lambda_r(1-a^2)}, \tag{42}$$

$$\lambda_r = \left(2\pi R n_{eff}\right)/m, \tag{43}$$

$$a^2 = \exp(-2\pi R(\alpha_i + \alpha_b + \alpha_d)), \tag{44}$$

$$\alpha_b = C1 * \exp(-C2 * R), \tag{45}$$

where, R is MR radius; m is the resonant mode number; n_{eff}, n_g, $C1$, and $C2$ are constants; and α_i, α_b, and α_d are loss coefficients. The definitions and typical values of these constants are given in Table 6. From Eq. (42)-(45), the device-level parameters of a passive MR device such as round-trip optical loss (a^2), resonant wavelength (λ_r), and loaded Q-factor (Q_L) ultimately depend on the MR radius (R).

Models for Active (Doped) Microring Resonators: A doped MR acts as a modulator, a filter, or a switch, the characteristics of which are defined by the values of λ_r, Q_L, a^2, bit-rate, free-

spectral range (*FSR*), and modulation shift ($\Delta\lambda_r$). Similar to passive MRs, Eq. (42)-(45) hold for doped MRs too. So, the values of Q_L and a^2 depend on R for doped MRs as well.

Doped MRs are doped in a similar manner as PN junctions. The free carrier concentration in a PN junction based MR can be controlled by applying forward or reverse biased voltage across the junction. The change in free carrier concentration alters the optical properties of the MR owing to the free carrier dispersion (FCD) and the free carrier absorption (FCA) effects [96]. The FCD effect alters the refractive index n and the FCA effect alters the absorption related loss coefficient α_d. The change in n in turn leads to a shift in the passband of the MR. The passband shift affects the light transmission from the source to the MR output, thereby achieving modulation, filtration, or switching of the input light signal. We assume the PN-junctions of doped MRs to be reverse-biased, as the doped MRs with reverse-biased PN-junctions render faster electrical response for high bandwidth modulation [20]. We also assume the doping concentrations of $N_a = N_d = 3\times10^{18}$ cm^{-3} (N_d for electrons in N-region and N_a for holes in P-region), as assumed in prior work [134].

We also study the effect of *MR radius* on bit-rate of a doped MR. As discussed in [20], the bit-period (and hence bit-rate) of a reverse-biased PN-junction based MR is limited either by the $R_S C_J$ time constant (where C_J is junction capacitance and R_S is series resistance) or by the photon lifetime of the MR, depending on which of the two is greater. C_J depends on the junction area, which in turn depends on the *MR radius*. The photon lifetime for an MR device is given by $\tau_p = (Q_L\lambda_r/2\pi c)$, which is a function of Q_L [20]. As explained earlier, Q_L of the MR depends on the radius (R), which implies that the photon lifetime of an MR also depends on R. Moreover, the resonance of an MR cavity is cyclic in nature, and the free spectral range (*FSR*; wavelength range between two successive resonances of an MR), is defined as [64]: $FSR = \lambda^2/2\pi R n_g$.

In summary, the device-level parameters of a doped MR such as round-trip optical loss (a^2), loaded Q-factor (Q_L), bit-rate ($R_S C_J$ time or photon lifetime), and FSR ultimately depend on the MR radius (R).

5.2.2. DEVICE-LEVEL DESIGN TRADEOFFS

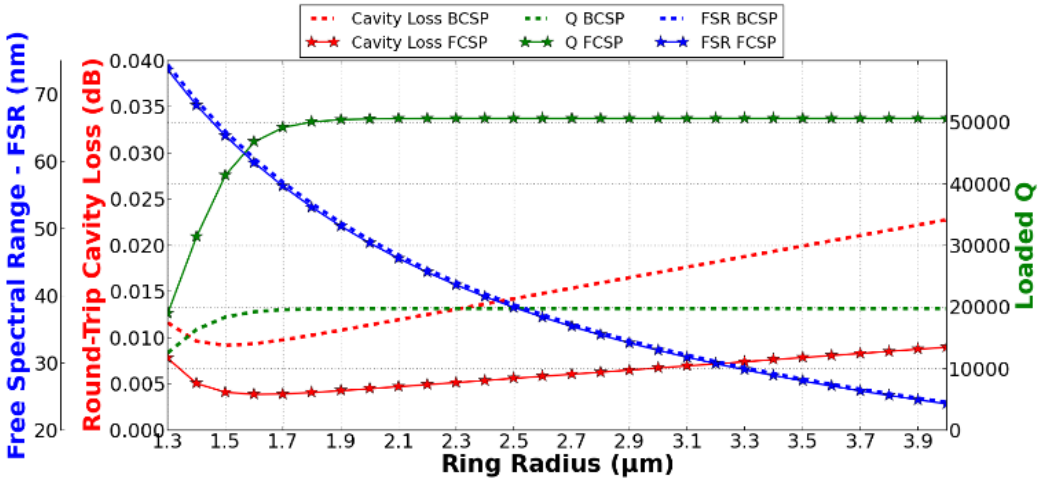
In this subsection, first, we present design tradeoffs for doped and passive MRs and then we present tradeoffs for passive WGs, for both FCSP and BCSP types of implementations.

Active/Passive Microring Resonators: As concluded in Section 5.2.1, various device-level design parameters of passive and active (doped) MRs ultimately depend on *MR radius* (*R*). This dependence of design parameters on *R* exists for both BCSP and FCSP MRs, because MRs in both cases operate on the same principle. The values of coefficients *C1*, *C2*, α_i, n_g, R_S and n_{eff} decide the degree by which various design parameters depend on *R*. The values of *C1*, *C2*, n_g and n_{eff}depend on the refractive index of MR materials and the device geometry.

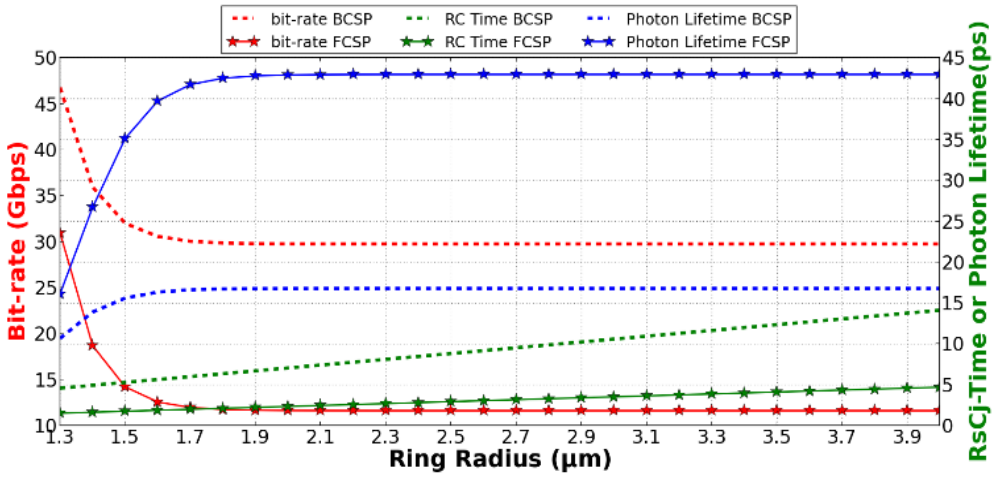
BCSP MRs are made of pSi (core)-SiO_2 (cladding), whereas FCSP MRs are made of cSi (core)-SiO_2 (cladding), with both types of MRs having the same device geometry. The optical properties of pSi and cSi are marginally different, as pSi exhibits high intrinsic optical loss due to surface roughness, grain boundaries, and dangling bonds [131]. As a result, values of *C1*, *C2*, α_i, n_g, R_S and n_{eff} differ between FCSP and BCSP MRs, causing the degree by which various device-level design parameters depend on R to differ for BCSP and FCSP MRs.

For this study, we modeled BCSP and FCSP MRs (both active/doped and passive) with the cross-sectional dimensions of 450nm×250nm, using the finite difference method [145]. For these models, we used the refractive index values *n* and n_{SiO2} from Table 6 and calculated the values of *C1*, *C2*, n_g and n_{eff} for λ = 1600nm, which are also given in Table 6. We explain the reason behind using λ = 1600nm later when we explain the design tradeoffs for WGs. Using these values

of the coefficients, we calculated the values of various design parameters using the equations presented in Section 5.2.1.



(a)



(b)

Figure 44: (a) Loaded Q factor, round-trip cavity loss, FSR, (b) R_SC_J time delay, photon lifetime, and bit-rate vs. MR radius for BCSP and cSi FCSP MRs. The curves of BCSP FSR and FCSP FSR are overlapped.

Figure 44 shows the various device-level design parameters such as R_SC_J time delay, photon lifetime, round-trip optical loss (a^2), loaded Q-factor (Q_L), and *FSR* versus the *MR radius* (*R*) for BCSP and FCSP MRs. We use the equations given in [67] to model C_J for BCSP and FCSP MRs.

From the figure, it can be observed that the degree by which the values of Q_L, a^2, R_SC_J, *FSR* and photon lifetime depend on the *MR radius* (*R*) differs between BCSP and FCSP MRs. The round-trip cavity loss (a^2, shown with red lines in Figure 44(a)) of a BCSP MR is greater than that for an FCSP MR for all values of *R*. This is due to the higher loss coefficients for BCSP MRs (Table 6). The larger value of round-trip loss in case of a BCSP MR results in a smaller value of Q_L (green lines in Figure 44(a)). The smaller Q_L of a BCSP MR results in a broader passband compared to an FCSP MR, which leads to higher insertion loss for a BCSP MR. Nevertheless, our analysis in Section 5.3.4 finds that the optimal design of BCSP links made of BCSP MRs renders more energy-efficiency than the optimal design of FCSP links made of FCSP MRs.

As described in [20], the rise-time and fall-time, and hence the bit-period of an MR is controlled by either the R_SC_J time delay or the photon lifetime, depending on which one of the two is greater. From Figure 44(b), the photon lifetime (blue lines) of FCSP and BCSP MRs is greater than their R_SC_J time delay (green lines), which implies that the bit-rate (inverse of bit-period) of BCSP and FCSP MRs is limited by the photon lifetime. In addition, the photon lifetime of FCSP MRs is greater than BCSP MRs, which leads us to the important conclusion that the bit-rate of BCSP MRs is greater than bit-rate of FCSP MRs for all values of MR radius.

Passive Waveguides (WGs): Next, we discuss the design tradeoffs of FCSP and BCSP passive WGs. Typically, FCSP WGs are fabricated using cSi core and SiO_2 cladding, whereas BCSP WGs are made of SiN core and SiO_2 cladding. The SiN-SiO_2 WGs have very high propagation loss (about 6dB/cm) in the C-band due to N-H and Si-H bond absorption harmonics, therefore, SiN-SiO_2 WG systems are typically operated in the L-band (near 1600nm) where they exhibit lower propagation loss (about 1dB/cm) [131]. Because of this reason, we analyze all the device-level parameters discussed in the preceding subsection for the 1600nm operating

wavelength. As discussed in [131], due to the ability of multilayer integration, superior coupling characteristics, and comparable propagation loss, the BCSP SiN-SiO_2 WGs outperform the FCSP cSi-SiO_2 WGs despite having higher scattering losses.

Furthermore, the maximum allowable optical power (*MAOP*) in SiN-SiO_2 and cSi-SiO_2 WGs is limited due to the emergence of nonlinearity effects at higher optical power, which incurs additional signal loss and degrades the performance of these WGs. The BCSP SiN-SiO_2 and FCSP cSi-SiO_2 WGs exhibit different types of nonlinear optical effects. The dominant nonlinear optical effects in the FCSP cSi-SiO_2 WGs are the two-photon absorption (TPA) effect and the resulting FCD and FCA effects [139]. The TPA induced FCA effect limits the *MAOP* in an FCSP cSi-SiO_2 bus WG to 100mW [57], [139]. In contrast, due to the absence of free carriers in SiN material, the TPA effect and the resulting FCA effect are not present in BCSP SiN-SiO_2 WGs [150]. However, the dominant nonlinear optical effects in the FCSP SiN-SiO_2 WGs are the second and third harmonic generation, which limits the *MAOP* in a BCSP SiN-SiO_2 bus WG to 350mW [150]. It will be evident from the discussion in Section 5.2.3 that a higher value of *MAOP* ultimately results in a larger number of DWDM channels in a SiN-SiO_2 BCSP bus WG than in FCSP cSi-SiO_2 WGs.

5.2.3. LINK-LEVEL DESIGN TRADEOFFS

In section 5.2.2, we presented the design tradeoffs among various device-level parameters such as *MR radius*, Q_L, *bit-rate*, nonlinear power limit, and *FSR*. In this subsection, we analyze how these parameters would affect design decisions at the higher link-level.

An on-chip SiP link typically comprises of a group of modulator MRs, a group of detector MRs with photodetectors, and a DWDM bus WG. The photonic signal transmission in on-chip SiP links is inherently *lossy*, i.e., the light signal is subject to losses such as insertion loss and modulation crosstalk related loss in modulator MRs, insertion loss and sideband truncation related

loss in detector MRs, and propagation and bending loss in WGs. All wavelength channels of a DWDM WG are subject to these losses. To ensure that signals of all channels propagating through the SiP link reach their destination before attenuating below the sensitivity threshold of the detector (minimum detectable power), the aggregate loss of all the channels along that link must fall within an acceptable range. This constraint is called the *optical power budget* and can be calculated in dB as the difference between the *MAOP* and the detector sensitivity. The optical power budget in dB (P_{Budget}^{dB}) determines how much loss can be present in the SiP link [139], which can be summarized as [57]:

$$P_{Budget}^{dB} \geq P_{Loss}^{dB} + 10log_{10}(N_{\lambda}), \tag{46}$$

where N_{λ} is the number of wavelength channels used in the link, and P_{Loss}^{dB} represents the sum of the loss contributions (in dB) incurred on a single channel by all the components (WG, detector and modulator MRs) present along the SiP link.

In this study, we assume the shot-noise limited sensitivity threshold of -22dBm for the FCSP photodetectors, as used in [57]. Due to the adverse effects of grain boundaries and dangling bonds, BCSP photodetectors are inherently more susceptible to noise than FCSP ones. Therefore, we assume a greater value of sensitivity threshold (-20dBm) for the BCSP photodetectors. From Section 5.2.2, the TPA-effect limited *MAOP* for an FCSP WG is 20dBm (100mW), whereas the harmonic generation effect limited *MAOP* for a BCSP WG is 25.4dBm (350mW). As a result, an FCSP link has P_{Budget}^{dB}= 42dB, whereas a BCSP link has P_{Budget}^{dB}= 45.4dB. The higher value of $\mathrm{P_{Budget}^{dB}}$ for the BCSP link allows a larger amount of aggregate loss $(P_{Loss}^{dB} + 10log_{10}(N_{\lambda}^{PB}))$ to be present in the BCSP link than in FCSP links.

For a given value of single channel loss P_{Loss}^{dB}, the N_λ in Eq. (46) should be less than a threshold value to limit the aggregate loss of the link within the *power budget* (P_{Budget}^{dB}). This threshold value (denoted as N_λ^{PB}) gives a P_{Budget}^{dB}-limited number of channels per WG. Along with the P_{Budget}^{dB}, the *FSR* of the largest MR along the WG also limits the number of channels per WG. The FSR-limited number of channels is given as N_λ^{FSR}= *FSR*/*CS*. Here, *CS* represents channel spacing, which is the distance between two adjacent wavelength channels of the SiP link. The actual feasible number of channels (N_λ^{Act}) per WG should be less than or equal to both N_λ^{PB} and N_λ^{FSR}. For a small enough value of P_{Loss}^{dB}, a given SiP link can have $N_\lambda^{PB} > N_\lambda^{FSR}$. In this case, N_λ^{Act} is the FSR-limited value N_λ^{FSR}. But, if the value of P_{Loss}^{dB} is greater than some threshold, then N_λ^{PB} becomes less than N_λ^{FSR}, and N_λ^{Act}=N_λ^{PB}. Thus, the actual number of channels (N_λ^{Act}) that are available for use per WG is $N_\lambda^{Act} = \underset{N_\lambda>0}{min}\left(N_\lambda^{FSR}, N_\lambda^{PB}\right)$.

In this study, we assume the cross-sectional dimensions of 450nm×250nm and WG propagation loss of 1dB/cm for both BCSP and FCSP WGs. We calculate the insertion loss and crosstalk related power penalty for the modulator MRs using the method described in [138], [139]. Moreover, to calculate the insertion loss and sideband truncation related power penalty for detector MRs, we use the experimentally validated analytical method described in [49]. From [138], the insertion loss and the crosstalk power penalty of modulator MRs depend on the Q_L, *channel spacing* (*CS*), and *modulation shift* (*MS*). *MS* is the amount by which the passband of a modulator shifts while modulating a signal. From [49], the insertion loss and the power penalty due to sideband truncation of MR detectors depend on the Q_L, *CS*, and *bit-rate (BR)*. Thus, the link-level design parameters P_{Loss}^{dB} and N_λ depend on some link-level design parameters such as CS, MS, and P_{Budget}^{dB}, as well as on some device-level design parameters such as Q_L and *BR* of MRs.

The observation above implies that the various device-level and link-level design parameters are interdependent. Figure 45 shows this interdependence among various design parameters of SiP links. The figure shows how the channel spacing (*CS*), modulation shift (*MS*), link-length, P_{Budget}^{dB}, and MR radius (*R*) do not depend on any other parameter in the dependence hierarchy. The combination of these five parameters in turn controls all the other parameters, which ultimately affects the aggregate bandwidth and power of the SiP link.

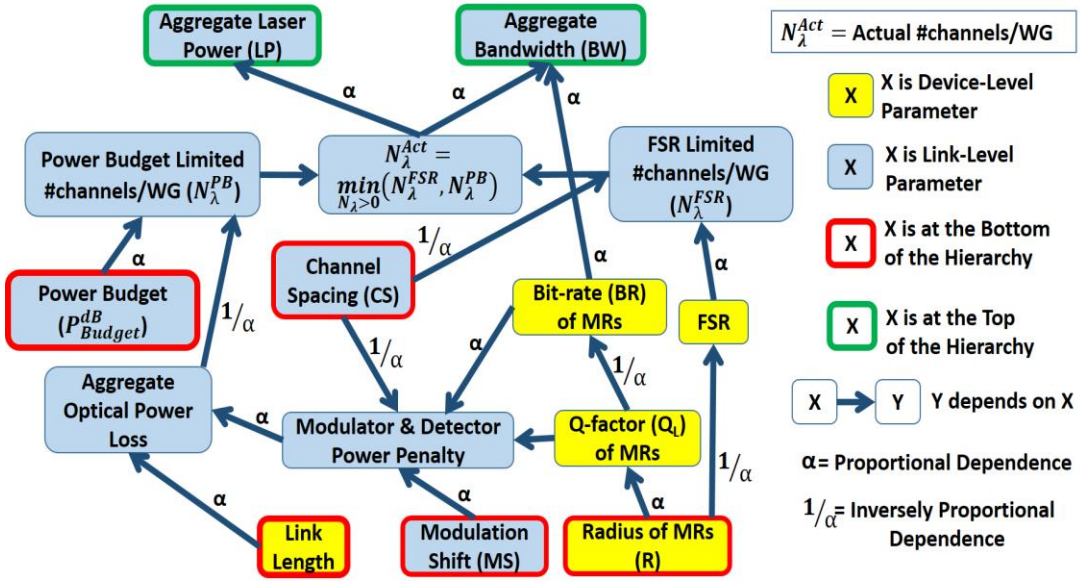


Figure 45: Interdependence among various link-level and device-level design parameters of on-chip SiP interconnects.

Consider Figure 46 to understand how the link-level design parameters such as channel loss (P_{Loss}^{dB}) and aggregate bandwidth depend on the power budget (P_{Budget}^{dB}), MR radius (*R*), link-length, *MS* and *CS*. Figure 46(a), Figure 46(c) show aggregate bandwidth versus *R* and *CS*, whereas Figure 46(b), Figure 46(d) show P_{Budget}^{dB} and P_{Loss}^{dB} values versus *R* and *CS*, for 5cm long BCSP and FCSP links, with *MS*=6pm. From Figures 46(b), 46(d), the FCSP link has P_{Budget}^{dB}= 42dB, whereas the BCSP link has P_{Budget}^{dB}= 45.4dB for all the values of *R* and *CS*.

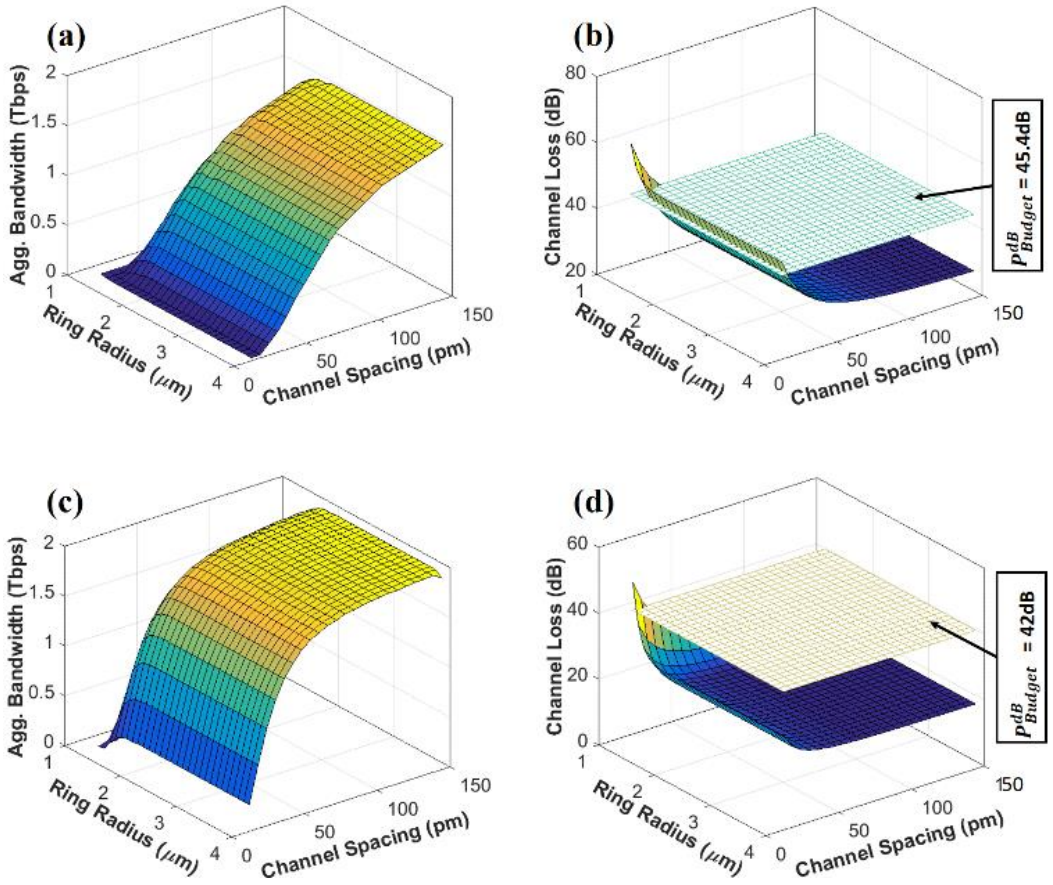


Figure 46: Aggregate bandwidth versus MR radius (*R*) and channel spacing (*CS*) for (a) a BCSP link, (c) an FCSP link. Power budget (P^{dB}_{Budget}) and channel loss (P^{dB}_{Loss}) versus *R* and *CS* for (b) a BCSP link, (d) an FCSP link. All plots are for 5cm link-length and *MS* of 6pm.

The maximum aggregate bandwidth of 1.47Tbps for the BCSP link occurs at *R*=1.9μm and *CS*=150pm (Figure 46(a)), which corresponds to P^{dB}_{Loss} of 28.3dB (Figure 46(b)) and Q-factor of 20,000. The maximum aggregate bandwidth of 1.93Tbps for the FCSP link occurs at *R*=2.1μm and *CS*=150pm (Figure 46(c)), which corresponds to P^{dB}_{Loss} of 19.7dB (Figure 46(d)) and Q-factor of 52,000. The smaller Q-factor renders higher power penalty due to MR sideband truncation for the BCSP link [49], which results in greater P^{dB}_{Loss} for the BCSP link than the FCSP link. Based on P^{dB}_{Loss} and P^{dB}_{Budget}values, the BCSP link and the FCSP link result in P^{dB}_{Budget} limited N^{Act}_{λ} values of 51 and 169 respectively. Thus, the BCSP link has less number of channels per waveguide.

Moreover, the values of Q-factor translate into channel bit-rate values of 28.7Gbps and 11.4Gbps for the BCSP link and the FCSP link respectively, which results in less maximum aggregate bandwidth of 1.47Tbps for the BCSP link than the maximum aggregate bandwidth of 1.93Tbps for the FCSP link.

Thus, for the given values of link-length=5cm, *MS*=6pm, and P_{Budget}^{dB}, the values of *R* and *CS* ultimately control P_{Loss}^{dB} and aggregate bandwidth of the BCSP and FCSP links. Similarly, for given values of *R*, P_{Budget}^{dB}, and *CS*, the link-length and *MS* can be shown to affect the ultimate values of P_{Loss}^{dB} and aggregate bandwidth. Thus, it can be concluded that the combination of the parameters *R*, *CS*, *MS*, link-length, and P_{Budget}^{dB} controls all the other parameters in the dependence hierarchy in Figure 2, which ultimately affects the aggregate bandwidth and P_{Loss}^{dB} of the SiP link. However, note that the values of link-length and P_{Budget}^{dB} cannot be varied for link optimization, as P_{Budget}^{dB} has a fixed value based on the underline device technology (FCSP or BCSP), and the link-length has a fixed value based on the layouts of and the distance between the source and destination. For this reason, the parameters *R*, *CS*, and *MS* are the only independently optimizable parameters in the dependence hierarchy in Figure 45.

Lastly, as evident from Figure 46, the decrease in *CS* results in the decrease of aggregate bandwidth but the increase of P_{Loss}^{dB}. Similarly, the increase in *R* results in the increase of aggregate bandwidth but the decrease of P_{Loss}^{dB}. Along the same lines, the increase in *MS* also affects the aggregate bandwidth and P_{Loss}^{dB} in opposite manners. Thus, it can be inferred that *the parameters R, CS and MS affect different parameters of the dependence hierarchy in different ways. Therefore, it is imperative to optimize all three of them simultaneously, to achieve energy-efficient and high-aggregate-bandwidth on-chip SiP links.* The next section discusses such an optimization step.

5.3. CROSS-LAYER OPTIMIZATION

In this section, we present a cross-layer optimization of various device-level and link-level parameters for BCSP and FCSP interconnects. These parameters depend on one another as shown in Figure 45.

5.3.1. PROBLEM FORMULATION

As the *MR radius (R)*, *CS*, and *MS* are the only independently optimizable parameters in the dependence hierarchy given in Figure 45, we use all possible values of these three variables as an input to our problem of parameter optimization. In Figure 44, the Q_L of the MRs saturates for a radius of about 3-4μm. Moreover, researchers have demonstrated in [138] that the minimal radius to obtain an intrinsic Q of 20,000, which corresponds to an optical bandwidth of 20GHz around the wavelength of 1.55μm, is 1.37μm. Furthermore, for any MR radius of greater than 4μm, the FSR becomes very small leading to an undesirably small value of N_λ^{FSR}, which results in poor aggregate bandwidth. Due to these reasons, we define the set of all possible viable values of *MR radius* $R = \{r | r \in Q^+; r\ is\ in\ \mu m; 1.3\mu m \leq r \leq 4.0\mu m; (r/0.1) \in N\}$, which has 28 elements. We aim to design SiP interconnects in ultra-dense WDM (UDWDM) regime, for which the *CS* is usually kept smaller than 25GHz or 200pm [151]. Therefore, we define the set of all possible values of CS as $\Delta = \{\delta | \delta \in N; \delta\ is\ in\ pm; 12pm \leq \delta \leq 150pm; (\delta\ mod\ 6) = 0\}$, which has 23 elements. Finally, as discussed in [138], the value of *MS* should be less than half the value of *CS* to limit worst-case insertion loss for modulator MRs. Therefore, to limit *MS* up to half of the *CS*, we define the set of all *MS* values $X = \{x | x \in N; x\ is\ in\ pm; 6pm \leq x < 75pm; (x\ mod\ 6) = 0\}$, which has 10 elements. The individual values for R, Δ and X combine to make a triplet in 28×23×10=6440 different ways. We create a set Y of these triplets, $Y =$

$\{(r_1, \delta_1, x_1), (r_1, \delta_1, x_2), \ldots, (r_{28}, \delta_{23}, x_{10})\}$ and give it as an input to our cross-layer optimization problem.

5.3.2. PROBLEM OBJECTIVE AND CONSTRAINTS

The main objective of our optimization problem is to design a single-WG SiP link of a given length with minimized *aggregate energy-per-bit (EPB)*. The *aggregate EPB* is the sum of *static EPB* (*SEPB*) and *dynamic EPB (DEPB)*. We obtain *SEPB* by dividing the aggregate laser power by aggregate bandwidth. The *DEPB* here represents *DEPB* of MRs. We calculate the *DEPB* of an MR from the required amount of charge depletion Δq to achieve corresponding *MS* using the equations given in [152]. As implied from the discussion in [138], the value of *MS* should be less than half the value of *CS* to limit the worst-case insertion loss for modulator MRs below an acceptable level, which is the constraint of the optimization problem. Out of 6440 total triplets of Y, 2268 triplets have *MS*>(*CS*/2), so they violate this constraint. Therefore, we remove these 2268 triplets from Y and define a new input set $Y'$with the remaining 4172 triplets.

5.3.3. OPTIMIZATION APPROACH

For each triplet of the constrained input set Y', first we calculate the Q_L, *FSR*, and *bit period* (2×$R_S C_J$ or 2×*photon lifetime*) using the methods and equations presented in Section 5.2.1. Using the values of Q_L and *bit period*, we then calculate the total channel loss P_{Loss}^{dB} (in dB) using the methods described in Section 5.2.2. Then, based on the optical power budget, we calculate N_λ^{Act} as described in Section 5.2.2. Next, for each triplet, we calculate the MR *bit-rate (BR)* by inverting *bit-period*. The actual feasible number of channels N_λ^{Act} is multiplied by the *BR* to obtain the *aggregate bandwidth (BW)* per WG. Using the calculated value of P_{Loss}^{dB} and P_{Budget}^{dB}, we calculate the total optical/laser power required to achieve the *BW*. We divide total laser power by the

achieved *BW* to obtain *SEPB*. We add *SEPB* and *DEPB* to obtain *aggregate EPB*. Lastly, we find an optimal triplet with minimum *aggregate EPB* out of all triplets of Y'. We use an exhaustive search approach, because it guarantees to find the optimal solution for the marginally small size of the constrained input set *Y'*.

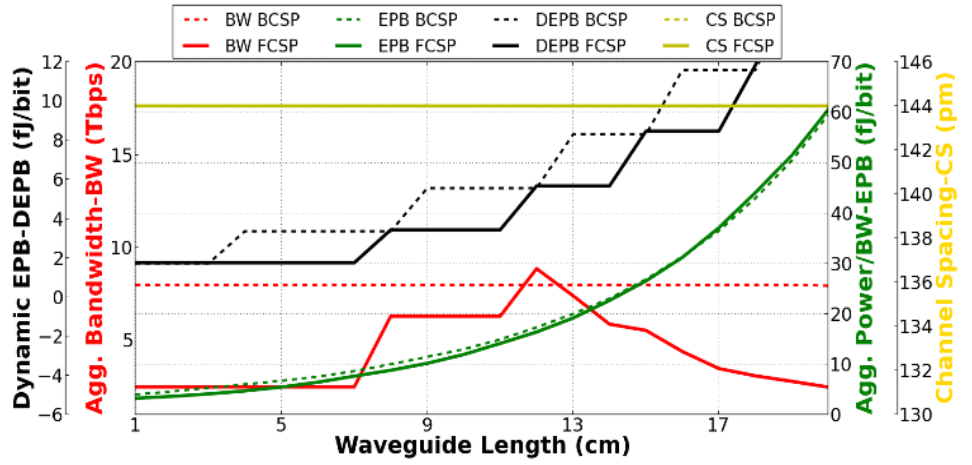
5.3.4. COMPARISON OF OPTIMIZED BCSP AND FCSP LINKS

To understand the available design choices for realizing energy-efficient and terabyte-per-second scale SiP interconnects with BCSP and FCSP, we optimize BCSP and FCSP links of 20 different lengths in the range from 1cm to 20cm using our cross-layer optimization framework. The results of this optimization are shown in Figure 47, which plots the values of various parameters obtained for the optimized BCSP and FCSP links of 20 different lengths (x-axes).

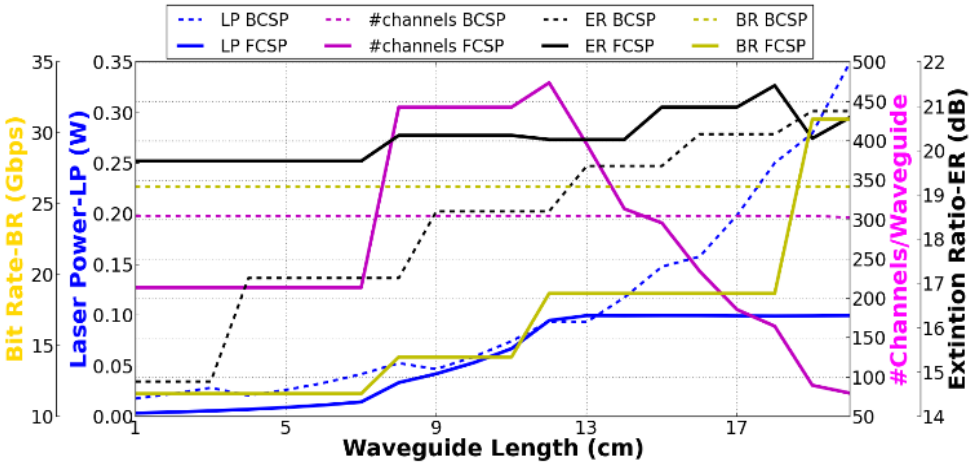
From the figure, it can be observed that the laser power (*LP)* for the BCSP link increases with increase in link-length. This is because the WG propagation loss (in dB) increases with increase in link-length, which in turn increases aggregate loss in the link, thus requiring higher *LP*. However, the BW of the BCSP link remains constant at 7.9Tbps for all link-lengths. As shown in Figure 45, the *BW* depends on only two parameters: *BR* and N_λ^{Act}. This implies that both *BR* and N_λ^{Act} should be constant for all link-lengths. As evident from Figure 47, *BR* and N_λ^{Act} actually remain constant at 26Gbps and 305 respectively for all link-lengths. Now, $N_\lambda^{Act}$is equal to the minimum of $N_\lambda^{FSR}$and N_λ^{PB}(Figure 45), which implies that either N_λ^{FSR} or N_λ^{PB} should be constant for all link-lengths. But, as the aggregate loss in the link increases, N_λ^{PB} should decrease to meet the power budget constraint in Eq. (46). This implies that N_λ^{FSR} remains constant for all link-lengths, which in turn keeps N_λ^{Act}constant. As a result, $N_\lambda^{FSR} < N_\lambda^{PB}$, and N_λ^{Act} is the FSR-limited N_λ^{FSR}. Similarly, for all the FCSP link-lengths below 8cm, the N_λ^{FSR} is less than N_λ^{PB}, as the *BW*,

BR and N_λ^{Act}(FSR-limited) are constant at 2.5Tbps, 11.5Gbps and 214 respectively. *Thus, it can be concluded that the FSR-limited value of* N_λ^{Act} *achieves constant BW for BCSP links irrespective of the link-length and link losses.*

For FCSP, at link-length of 8cm, the *BW* of the FCSP link shoots up to 6Tbps from 2.5Tbps. So, as evident from Figure 45, the increase in either *BR* or N_λ^{Act} should be the cause of it. From Figure 47, at link-length of 8cm both *BR* and N_λ^{Act} increase to 14Gbps and 440 respectively, the combined effect of which increases the *BW*. For FCSP link-lengths between 8cm and 12cm, as shown in Figure 47, both *BR* and N_λ^{Act} keep increasing with increase in link-length, which results in the increase of *BW* with increase in link-length. In addition, the *LP* also keeps increasing with link-length. However, for FCSP link-lengths beyond 12cm, *BW* decreases with increase in link-length, in spite of the increase in *BR*. This is due to decreasing N_λ^{Act}. As N_λ^{PB} becomes less than N_λ^{FSR}, N_λ^{Act} becomes power budget-limited. As the *LP* is saturated at *MAOP* of 100mW for FCSP link-lengths beyond 12cm, N_λ^{Act} keeps decreasing with increase in link-length, because of the increase in aggregate link loss with increase in link-length. This observation implies that for an FCSP link whose N_λ^{Act} is limited by the power budget, the *BW* decreases with increase in link-length and the *LP* remains constant at the *MAOP*. Decreasing *BW* at constant *LP* causes a deterioration in SEPB. These observations can be generalized to hold true for BCSP links as well, because BCSP and FCSP links operate on the same principle. Thus, it can be concluded that to design an FCSP or BCSP link to achieve high BW irrespective of the link-length and link losses, all link-level and device-level design parameters in Figure 45 should be optimized to achieve an FSR-limited value of N_λ^{Act}. For that, the channel loss P_{Loss}^{dB} (Eq. 46) should be smaller than a certain threshold value to allow N_λ^{PB} to be greater than N_λ^{FSR}.



(a)



(b)

Figure 47: (a) Aggregate bandwidth (BW), aggregate energy-per-bit (EPB), dynamic EPB (DEPB), and channel spacing (CS), (b) laser power (LP), number of channels per WG (N_{λ}^{Act}), extinction ratio (ER) and bit-rate (BR) values obtained for the optimized BCSP and FCSP links of 20 different lengths. The traces of CS BCSP and CS FCSP are overlapped.

From Figure 47, the optimal value of *CS* for all link-lengths for both the BCSP link and the FCSP link is 144pm. For all link-lengths, the BCSP link has greater *dynamic EPB* than the FCSP link. This is because the BCSP link has greater values of optimized modulation shift (not shown in the figure) than the FCSP link. Moreover, Figure 4 also plots *aggregate EPB* and extinction ratio (*ER*). Extinction ratio is defined as the ratio of the optical power in the bus WG during logic "1" state to the optical power during logic "0" state. As evident from the figure, BCSP links have

inferior *ER* compared to the FCSP links. This is because, as shown in Figure 44, the BCSP MRs have smaller Q_L and greater cavity loss than the FCSP MRs, which results in lower optical power in the bus WG for logic "1", thereby decreasing the ER. The inferior *ER* for the BCSP link decreases the signal power and increases its susceptibility to noise. Furthermore, the aggregate EPB values obtained for the BCSP links is quite comparable to those obtained for the FCSP links. *Therefore, it can be concluded from these observations that the optimized design of a BCSP link yields more aggregate bandwidth with comparable aggregate EPB, but an inferior extinction ratio than the optimized design of an FCSP link.*

5.4. EVALUATION

5.4.1. EVALUATION SETUP

We performed benchmark-driven simulation-based analysis to evaluate the impact of FCSP and BCSP devices on the performance and efficiency of two well-known crossbar PNoC architectures: Corona [14] and Firefly [15]. We modeled and simulated the Corona and Firefly PNoCs with FCSP and BCSP devices using an in-house cycle-accurate NoC simulator. We evaluated performance for a 256 core single-chip architecture at a 22nm CMOS node. We used real-world traffic from applications in the PARSEC benchmark suite [76] in our analysis. GEM5 full-system simulation [77] of parallelized PARSEC applications was used to generate traces that were fed into our cycle-accurate NoC simulator. We set a "warm-up" period of 100M instructions and then captured traces for the subsequent 1B instructions.

First, based on geometric analysis, we estimated the maximum length of the crossbar WG in both Firefly and Corona PNoCs. The maximum length of the single-write-multiple-read (SWMR) WG in Firefly PNoC is 8cm. This 8cm long SWMR WG between a source and destination node passes through 6 intermediate inactive nodes. Similarly, the maximum length of the multiple-

write-single-read (MWSR) WG in Corona PNoC is 12cm. This 12cm long WG between a source and a destination node passes through 62 intermediate inactive nodes. Each node along the crossbar WGs of both the Corona and Firefly PNoCs has arrays of modulator and detector MRs.

We model two different variants of Corona and Firefly PNoCs along with the baseline variants. One variant of Corona and Firefly each uses BCSP devices (referred to as Corona-BCSP and Firefly-BCSP), whereas the other variant uses FCSP devices (referred to as Corona-FCSP and Firefly-FCSP). The baseline variants also use the same type of front-end compatible MRs and WGs as used in the FCSP variants of the PNoCs. However, we optimize the design parameters of the FCSP variants (Firefly-FCSP and Corona-FCSP) using our cross-layer optimization framework, whereas the design parameters of the baseline variants are taken from [14] and [15] and are not optimized. We keep the number of WGs and basic floorplan of the architectures constant across all three variants. We optimized the crossbar data WG designs of all the variants of both PNoCs using the cross-layer optimization described in Section 5.3, and obtained the maximum allowed number of channels N_{λ}^{Act} for all of them. Here, N_{λ}^{Act} represents the maximum allowed DWDM degree for a given power budget. We also obtain the optical loss values and dynamic EPB values from our optimization framework. Further, we considered a fixed packet size of 512 bits across all the variants of Corona and Firefly architectures.

Table 7 summarizes the DWDM degree, optical loss, and dynamic EPB values for the different variants of the Firefly and Corona PNoCs that we consider. Our optimization framework obtains the optimal modulation shift (MS) of 18pm, 24pm, 54pm, and 72pm for the Firefly-FCSP, Firefly-BCSP, Corona-FCSP, and Corona-BCSP respectively, which results in the dynamic energy values of 3.5pJ/bit, 5.5pJ/bit, 15pJ/bit, and 20pJ/bit for the Firefly-FCSP, Firefly-BCSP, Corona-FCSP, and Corona-BCSP respectively.

Table 7: Packet size, DWDM degree, optical loss and per bit dynamic energy for different variants of Firefly and Corona PNoC architectures.

Configuration	Maximum waveguide DWDM	Selected waveguide DWDM	Optical loss data WGs (in dB)	Dynamic energy (in fJ/bit)
Firefly Baseline	64	64	-41.64	1.1
Firefly FCSP	215	128	-39	3.5
Firefly BCSP	260	256	-43	5.5
Corona Baseline	64	64	-51.4	1.1
Corona FCSP	5	4	-42	15
Corona BCSP	20	16	-44.4	20

5.4.2. EVALUATION RESULTS FOR FIREFLY PNOC

We used the reservation-assisted Firefly PNoC architecture with 64 DWDM as the baseline and compared it with two variants: Firefly-BCSP and Firefly-FCSP. As shown in Table 7, the Firefly-BCSP and Firefly-FCSP have maximum DWDM degree of 260 and 215 respectively. These values of DWDM degree are FSR-limited and we have obtained them for CS=0.15nm from our optimization framework. Prior works [153] and [154] have demonstrated 20 GHz-spaced (0.2nm-spaced), 200nm-wide comb sources, which are capable of sourcing a WG with DWDM degree of 1000 (total 1000 channels per WG). This implies that it is feasible for the Firefly-BCSP and the Firefly-FCSP to have DWDM degree of 260 and 215 respectively. However, we choose the DWDM degrees of the PNoCs to be factors of the packet-size of 512 bits. Therefore, we select the DWDM degree of the Firefly-BCSP and the Firefly-FCSP to be 256 and 128 respectively (Table 7). Moreover, to facilitate simultaneous traversal of 512 bits (entire packet) from source node to destination node in Firefly-BCSP, we have considered two SWMR WGs as a group with each WG having 256 DWDM. Further, for reasonable comparison of Firefly-BCSP with Firefly-FCSP and Firefly (baseline), we also considered two SWMR WGs as a group in these architectures as well.

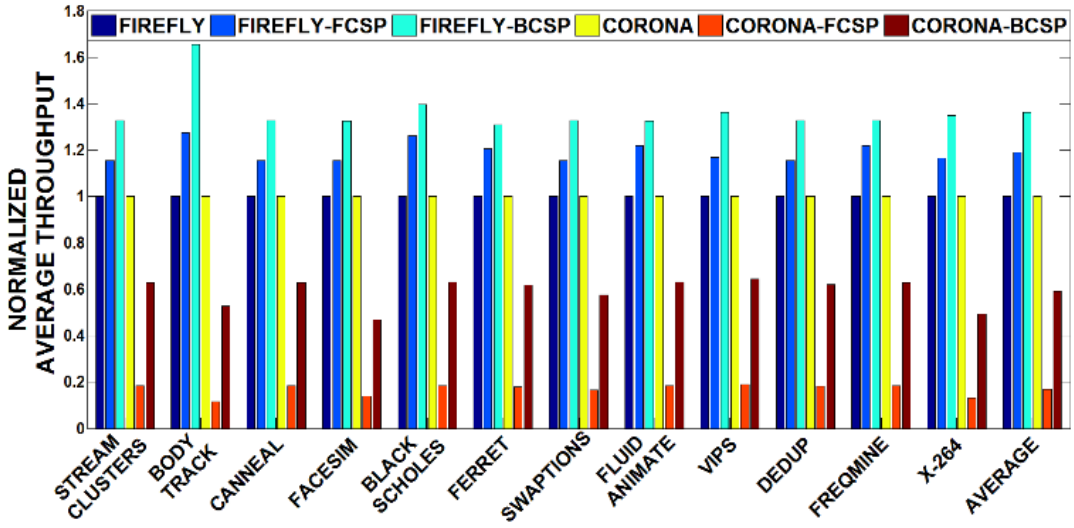


Figure 48: Throughput comparison for different variants of Firefly and Corona PNoCs. Results are shown for PARSEC applications and normalized wrt baseline architectures.

The average throughput and aggregate energy-per-bit (EPB) for all three variants of the Firefly PNoC architecture across 12 multi-threaded PARSEC benchmarks are presented in Figure 48 and Figure 49 respectively. As evident from Figure 48, Firefly-BCSP and Firefly-FCSP yield 36.4% and 19.1% higher throughputs respectively on average over the baseline Firefly. The larger value of DWDM degree for Firefly-BCSP results in greater throughput compared to Firefly-FCSP and baseline Firefly. We calculate aggregate EPB values using the same method as used in our optimization framework described in Section 5.3. From Figure 49, Firefly-BCSP and Firefly-FCSP yield 26.4% and 15.9% less aggregate EPB respectively on average over the baseline Firefly. Firefly-BCSP achieves 1.15× greater throughput and 12.4% less EPB than Firefly-FCSP. The greater throughput for Firefly-BCSP results in a lower value of aggregate EPB compared to Firefly-FCSP and baseline Firefly. The smaller value of aggregate EPB obtained for Firefly-BCSP implies that Firefly-BCSP is more energy-efficient than Firefly-FCSP.

5.4.3. EVALUATION RESULTS FOR CORONA PNOC

We performed a similar analysis for the Corona PNoC architecture with token-slot arbitration and 64 DWDM as the baseline and compared it with two variants Corona-BCSP and Corona-FCSP. As shown in Table 7, Corona-BCSP and Corona-FCSP have a power-budget limited DWDM degree of 20 and 5 respectively. As mentioned earlier, the crossbar WG of Corona is 12cm long and it passes through 62 intermediate nodes, which in turn increases the optical loss resulting in smaller values of DWDM degree compared to Firefly. Moreover, the baseline Corona has optical loss of 51.4dB (Table 6), which is significantly larger than the optical power budget of 42dB for FCSP WGs. *This implies that the DWDM degree of 64 used in the baseline Corona architecture is not feasible from a practical implementation perspective.*

The average throughput and aggregate EPB for all three variants of the Corona architecture across 12 multi-threaded PARSEC benchmarks are presented in Figure 48 and Figure 49, respectively. As the baseline Corona PNoC is not feasible, the results shown in Figure 47 and Figure 48 for the baseline Corona configuration are not practically achievable. As evident from Figure 48, Corona-BCSP and Corona-FCSP yield 40.8% and 83.1% less throughput respectively on average over the baseline Corona configuration. The baseline has a larger (but impractical to achieve) DWDM degree, which results in larger values of throughput for it compared to Corona-BCSP and Corona-FCSP. As evident from Figure 49, Corona-BCSP and Corona-FCSP yield 3.82× and 6.31× greater aggregate EPB respectively on average over the baseline. The greater DWDM degree of 64 (although impractical) results in greater throughput for the baseline, and consequently a lower value of aggregate EPB compared to Corona-FCSP and Corona-BCSP. Similarly, greater DWDM degree for Corona-BCSP yields 3.5× greater throughput for it compared to Corona-FCSP. The greater throughput results in 39.5% less EPB for Corona-BCSP compared to Corona-FCSP.

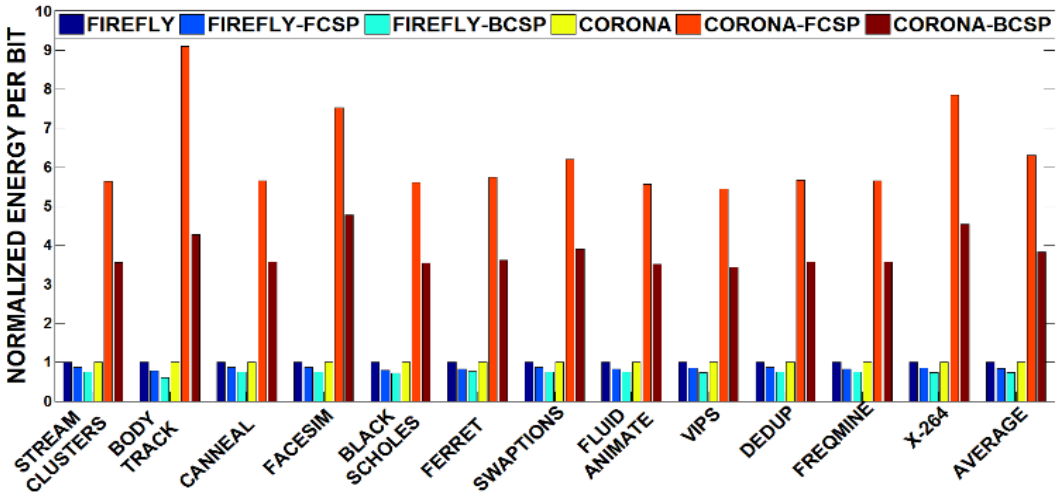


Figure 49: Energy-per-bit (EPB) comparison for variants of Firefly and Corona architectures across PARSEC applications. Results normalized wrt baseline architectures.

To summarize the major findings from our experiments, we showed that Firefly-BCSP and Corona-BCSP yields greater throughput and less aggregate EPB than Firefly-FCSP and Corona-FCSP respectively, implying that BCSP links perform better and are more energy-efficient than FCSP links. The smaller values of DWDM degree obtained for Corona-FCSP and Corona-BCSP corroborate our previous observation (Section 5.2.3) that the power budget and optical loss of BCSP and FCSP links limit the maximum allowable DWDM degree, which in turn constrains the practically achievable aggregate bandwidth and energy-efficiency in PNoCs such as Corona.

5.5. CONCLUSIONS

This chapter presented a detailed comparative analysis of a number of design tradeoffs for CMOS front-end (FCSP) and back-end (BCSP) compatible silicon photonic devices. The results of the cross-layer optimization of multiple device-level and link-level design parameters indicated that BCSP interconnects yield more throughput with comparable energy-efficiency compared to FCSP interconnects. The optimized design of BCSP-based Firefly and Corona photonic network-on-chips (PNoCs) yield 1.15× and 3.5× greater throughput with 12.4% and 39.5% more energy-

efficiency than the optimized design of FCSP-based Firefly and Corona PNoCs respectively. We showed that the greater throughput and comparable energy-efficiency obtained for BCSP links favor their use in the terabyte-per-second scale silicon photonic interconnects in future PNoCs. However, the inferior extinction ratio for BCSP links necessitates a reduction of intrinsic optical losses present in BCSP devices. Moreover, the sources of crosstalk and noise in BCSP interconnects that threaten the reliability of communication need to be thoroughly investigated and mitigated.

6. RUNTIME LASER POWER MANAGEMENT IN PHOTONIC NOCS WITH ON-CHIP SEMICONDUCTOR OPTICAL AMPLIFIERS

Photonic network-on-chip (PNoC) architectures are projected to achieve very high bandwidth with relatively small data-dependent energy consumption compared to their electrical counterparts. However, PNoC architectures require a non-trivial amount of static laser power, which can offset most of the bandwidth and energy benefits. In this chapter, we present a novel low-overhead technique for run-time management of laser power in PNoCs, which makes use of on-chip semiconductor amplifiers (SOA) to achieve traffic-independent and loss-aware savings in laser power consumption. Experimental analysis shows that our technique achieves 31.5% more laser power savings with 12.8% less latency overhead compared to another laser power management scheme from prior work.

6.1. INTRODUCTION

In the many-core era, processors with hundreds of cores on a single chip are gradually becoming a reality. The performance of these many-core processors is driven by the effectiveness of the underlying electrical network-on-chip (ENoC) fabrics that are becoming increasingly crosstalk- and energy-limited [8]. To this end, due to the recent developments in the area of silicon photonics, photonic network-on-chip (PNoC) fabrics have been projected to supersede ENoCs. PNoCs offer a multitude of benefits over ENoCs such as: 1) higher bandwidth density; 2) distance-independent bit-rate; and 3) smaller data-dependent energy. However, irrespective of network traffic and utilization, PNoCs dissipate a non-trivial amount of static power in their laser source. The high laser power overheads can offset the bandwidth and energy advantages of PNoCs.

Therefore, it is imperative to forge new techniques that can reduce the static power consumed in the laser sources of future PNoC architectures.

Several techniques have been proposed in prior works, e.g. [8], [155]-[159], that aim to reduce static power in the laser sources of PNoCs. All of these techniques leverage temporal and spatial variations in network traffic and opportunistically adjust the power in laser sources by tuning or distributing the available network bandwidth. These methods tend to notably reduce the power in laser sources during low network load conditions. However, if the losses encountered by optical signals in the network between the source and destination are high, these methods would still require excessive laser power to compensate for the high losses, even under low network load conditions. *This observation motivates the need for a technique that can provide traffic-independent and loss-aware savings in laser power.*

In this chapter, we present a novel low overhead technique for run-time management of laser power in PNoCs, which makes use of on-chip semiconductor amplifiers (SOA) to achieve traffic-independent and loss-aware savings in laser power. We refer to our SOA-enabled laser power management technique as *SOA_LPM*. Unlike the techniques proposed in prior works [155]-[159], *SOA_LPM* draws minimum power from the off-chip laser source and offloads the burden of loss-aware run-time laser power management to on-chip SOAs, which in turn enables significant savings in laser power with minimal overheads. Moreover, *SOA_LPM* is orthogonal to all the other laser power management techniques reported in prior works, and can be used in combination with any of them. Our novel contributions in this chapter are summarized below:

- We propose a low overhead, SOA-enabled, and loss-aware technique (*SOA_LPM*) to manage and optimize the laser power overhead in PNoC architectures at run-time;

- We implement our *SOA_LPM* technique on a multiple-write -multiple-read (MWMR) photonic waveguide architecture;
- We present device-level analytical models of on-chip SOAs, and based on these models, we analyze the energy, area and performance overhead of our *SOA_LPM* technique;
- We evaluate *SOA_LPM* by implementing it on a well-known PNoC architecture Flexishare [16] and compare it with another laser power management technique from prior work [156].

6.2. BACKGROUND

PNoC architectures (e.g., [12] and [81]) employ multiple high-bandwidth photonic links, each of which connects two or more nodes (e.g., cores). Typically, a large number of wavelengths are dense wavelength division multiplexed (DWDM) in a single photonic link. Each wavelength corresponds to a channel that is used for serial data transfers. Additionally, a photonic link employs microring resonator (MR) modulators (that are in resonance with the utilized wavelengths) at the source node to modulate electrical signals onto photonic signals that travel through the link, and MR detectors at the destination node to detect photonic signals and recover electrical signals. In general, the use of multiple wavelengths (or channels) in a photonic link enables high bandwidth parallel data transfers across the link.

The amount of laser power required by a source node to transfer data across the parallel wavelength channels of a photonic link to a destination node can be expressed as:

$$P_{Laser} - S \geq P_{Loss} + 10\log_{10}(N_{\lambda}), \tag{47}$$

where, P_{Laser} is the required laser power in dBm, N_{λ} is the number of wavelength channels in the link, P_{Loss} (in dB) is the total optical loss faced by a photonic signal along the link from the source to the destination, and S is the sensitivity of the detector (assumed to be -20dBm [160]). P_{Loss}

includes optical signal losses such as through loss in MR modulators and detectors, modulating losses in modulator MRs, detection loss in detector MRs, propagation and bending loss in waveguides, and splitting loss in splitters. Overall, P_{Laser} thus depends on two main factors: 1) link bandwidth in terms of N_λ, which controls the network utilization and traffic, and 2) the total optical loss P_{Loss} during photonic signal propagation [56] (Chapter 5).

As implied from Eq. (47), if a link is underutilized (due to low traffic), its laser power consumption P_{Laser} can be reduced by decreasing N_λ associated with the link. This is exactly what is done in prior works [8], [155], and [156] to reduce the total laser power consumption in PNoCs. Some other prior works also propose laser power management techniques, e.g., [157] and [158], wherein they achieve significant power savings using dynamic reconfiguration of photonic networks. In spite of requiring periodic evaluation of network traffic and expensive run-time decision-making, these methods (presented in [155]-[159]) achieve notable savings in laser power. But the savings are highly contingent on information related to network traffic and losses. In contrast to the approaches from [155]-[159], Wang et al. [159] proposed a technique that achieves loss-aware savings in laser power. However, this technique requires the network to function in the TDM communication paradigm only, which incurs architecture specific overheads, limiting the scope of this technique.

In this chapter, we present a complementary, low overhead, and SOA-enabled technique called SOA_LPM that can provide traffic-independent and loss-aware laser power savings for PNoC architectures. For transmitting a packet between source and destination nodes, *SOA_LPM* first allocates only the minimum amount of laser power to the source node that is enough for correct detection at the destination node. It then accounts for losses to be faced by the packet on its path from the source to the destination and enables the source to amplify the allocated laser

power to the necessary level by using an on-chip SOA. As will be evident in the following sections, *SOA_LPM* proves to be more energy efficient than previously proposed techniques.

Table 8: Definitions and values of parameters for our SOA model.

Parameter	Definition	Value
a_1	Constant	6.7×10^{-16} cm^2
n_0	Transparency carrier concentration	1.2×10^{18} cm^{-3}
α	Loss in SOA active region	10 cm^{-1}
Γ	Light confinement factor	0.4
L	Length of SOA active region	10μm
$\Delta\lambda$	SOA gain linewidth	95nm
I_0	Threshold input current	5μA

6.3. SEMICONDUCTOR OPTICAL AMPLIFIERS: OVERVIEW

We first give a brief explanation of the structure and behavior of SOAs, before presenting analytical models for SOA gain (Section 6.3.1) and overheads (Section 6.3.2).

A detailed description of the structure, functionality, and modeling of SOAs is given in [161]. Briefly, an SOA is an optoelectronic device, which can be heterogeneously integrated with a silicon-on-insulator (SOI) based silicon photonic chip, and under suitable operating conditions can amplify an input broadband light signal. The structure of an SOA consists of an active waveguide region made of an intrinsic narrow bandgap material (e.g., AlInGaAs and InGaAsP), which is sandwiched between n-type and p-type cladding materials (e.g., n-InP and p-InP) with wider bandgaps. Free carriers are injected into the active waveguide region from the applied bias current, which in turn causes population inversion in the active region. The population inversion of free carriers in the active region results in stimulated emission of light, which imparts "gain" to the input optical signal. The operational characteristics and the gain spectrum of an SOA depend on its structure and materials used.

6.3.1. ANALYTICAL MODEL FOR SOA GAIN

As mentioned earlier, an SOA can provide broadband optical gain in its active region through stimulated emission. This gain obtained in the active region of unity length is called material gain (g_m). The effect of g_m on SOA output power can be exponentially increased by increasing the length of the active region to provide a very high value of single-pass bulk SOA gain (G). Both g_m and G are functions of wavelength (λ) and input bias current (I), which can be expressed as [161]:

$$g_m(\lambda, I) = \left[\left(\Gamma a_1 n_0 \left\{\frac{I}{I_0} - 1\right\}\right) - \alpha\right]\left[1 - \frac{2(\lambda - 1570)^2}{\Delta\lambda^2}\right], \quad (48)$$

$$G(\lambda, I) = 10 \log_{10}\left(e^{L * g_m(\lambda, I)}\right), \quad (49)$$

Here, g_m and G are in cm^{-1} and dB, respectively; and a_1, α, n_0, Γ, L, I_0 and $\Delta\lambda$ are constants that depend on the structure and operating conditions of the SOA. We took the typical values of these constants from [161], as shown in Table 8. We modeled the SOAs used in this work using Eq. (48), (49) and the constants in Table 8.

6.3.2. OVERHEAD ANALYSIS

From the description in the previous sections, the SOA gain is proportional to SOA input current (I). As a result, use of SOA to amplify a DWDM signal comes with a non-zero power overhead corresponding to the non-zero SOA input current. The power consumed by an SOA can be calculated by multiplying the target I with the operating voltage. Based on guidelines for bulk SOA design [161], we design the length (L) of the active region and operating voltage of our SOAs to be 15μm and 1.5V, respectively. Thus, we assess the power overhead of an SOA by multiplying its target I with 1.5V. Moreover, an SOA takes about 20-50ps to adjust to the target gain [161]. In

summary, an SOA incurs power overhead equal to input current (I) × 1.5V, and latency overhead of at most 50ps (0.25 cycles at 5GHz).

Note that the broadband gain profile of an SOA is subject to fluctuations due to temperature variations and noise. To compensate for these fluctuations, our *SOA_LPM* technique operates the SOAs at such input current levels that can yield 3dB more SOA gain than the desired gain.

6.4. SOA ENABLED LASER POWER MANAGEMENT

Our proposed *SOA_LPM* technique uses SOAs in combination with comb switches [162] and lookup tables to enable loss-aware run-time laser power management in PNoC architectures. *SOA_LPM* can be easily ported to different PNoCs and its implementation depends on the type of bus waveguides (BWGs) used in PNoC architecture. To evaluate our proposed *SOA_LPM* in this work, we implement it on a crossbar based PNoC architecture Flexishare [16]. The Flexishare PNoC uses multiple write multiple read (MWMR) type of BWGs. Each node (e.g., a processing element) connected to a crossbar requires both read and write access to the other nodes. A detailed analysis of *SOA_LPM* implementation for MWMR BWGs is presented in the following sub-section.

6.4.1. IMPLEMENTATION FOR MWMR BUS WAVEGUIDE

In an MWMR BWG, multiple nodes are capable of sending and receiving data using their modulating and detecting MR banks respectively. Therefore, MWMR BWGs require arbitration among multiple sender nodes, and also receiver selection among multiple receiver nodes [16].

Figure 50 illustrates the implementation of *SOA_LPM* on a typical MWMR BWG based PNoC. As shown in the figure, the PNoC is comprised of multiple MWMR BWGs. In general, MWMR BWG based PNoCs with N nodes have N sender nodes and N receiver nodes (i.e., all N

nodes can send as well as receive), with implementation-specific *K* MWMR BWGs. Each MWMR BWG employs a comb switch [162] (a broadband MR switch that can switch the entire DWDM spectrum) and an on-chip SOA [161]. Thus, the PNoC requires *K* SOAs and *K* comb switches as there are *K* BWGs in the PNoC. The SOA of each MWSR BWG can be controlled by any of the *N* sender nodes depending on which sender node wins the arbitration.

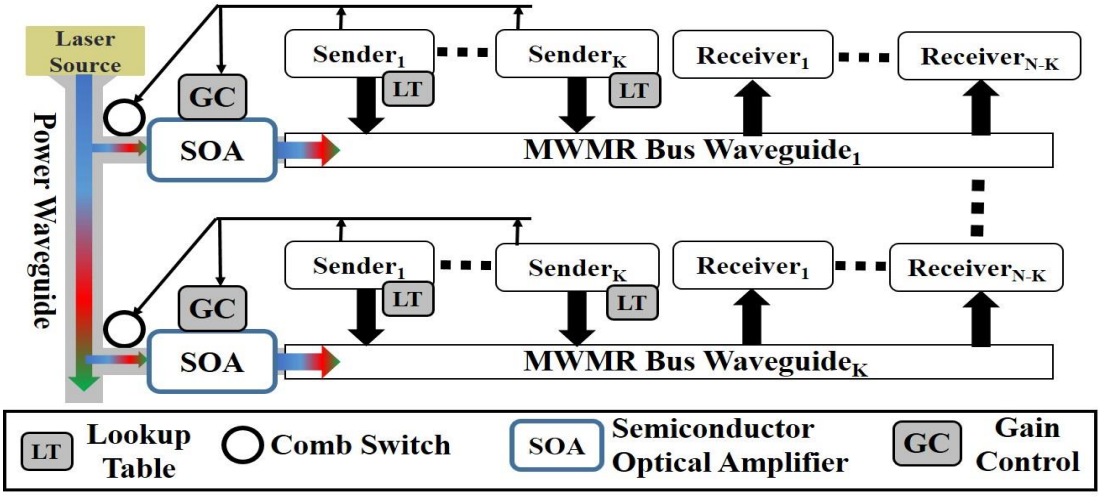


Figure 50: Implementation of SOA_LPM on MWMR BWG based PNoC.

As shown in the figure, each sender node of each MWMR BWG has an SRAM based lookup table that stores *N-1* values of loss (corresponding to *N-1* receivers). After arbitration, the authorized sender node initiates a receiver selection phase, at the end of which the sender node accesses a corresponding entry from the lookup table to determine the total loss value that the signal will face on its way to the target receiver. Then, the sender node adjusts the gain of the SOA to be equal to the loss value. Next, the sender node controls the comb switch to extract only the minimum laser power equal to S=-20dBm (detector sensitivity) from the power waveguide and provide it as an input to the SOA. The SOA amplifies the allocated laser power by a value that is equal to the accessed loss value and provides it to the sender node, which then modulates it for

communication with the target receiver. Note that we assume each entry of the lookup table to be of 8 bits, therefore, each sender has *N-1* number of 8-bit entries in its SRAM based lookup table.

6.5. EVALUATION

6.5.1. EVALUATION SETUP

We target a 256-core many-core system for evaluating our SOA enabled laser power management (*SOA_LPM*) technique. We evaluate *SOA_LPM* on a well-known crossbar-based PNoC architecture Flexishare [16]. Flexishare uses 32 groups of MWMR BWGs with a 2-pass token stream arbitration. Each MWMR BWG in Flexishare architecture is capable of transferring 512 bits of data from a source node to a destination node.

We modeled and simulated the architectures at cycle-accurate granularity with a SystemC-based NoC simulator. We used real-world traffic from applications in the PARSEC benchmark suite [76]. Full-system gem5 simulation of parallelized PARSEC benchmarks [76] was used to generate traces that were fed into our cycle-accurate NoC simulator. We set a "warm-up" period of 100M instructions and captured traces for 1B instructions. We targeted a 22nm process node and 5GHz clock frequency for the 256-core system. We considered BWGs with 64 DWDM wavelengths sharing the working band 1510nm–1590nm.

The static and dynamic energy consumption of electrical routers and concentrators in Flexishare architecture is based on results from the DSENT tool [75]. We model and consider area and performance overheads for *SOA_LPM* enabled laser power management. We estimated electrical area and power overhead using gate-level analysis and the open-source CACTI tool [78] for the SRAM-based lookup tables. The electrical area overhead, the electrical power overhead, and the photonic area overhead is estimated to be $0.8mm^2$, 0.24mW, and $237.5\mu m^2$ respectively for the Flexishare PNoC architecture.

To compute laser power, we considered detector sensitivity of -20dBm, MR through loss of 0.02 dB, waveguide propagation loss of 1dB/cm, waveguide-bending loss of 0.005dB/90^0, and waveguide coupler/splitter loss as 0.5dB [80]. We calculated photonic loss in components using these values, which sets the photonic power budget and correspondingly the electrical power for the off-chip laser source. Moreover, we take the energy/power and latency overhead values of SOAs and comb switches as discussed in Section 6.3.1.

6.5.2. COMPARISON WITH PRIOR WORK

We compared *SOA_LPM* with a dynamic laser power management technique (*BW_LPM*) from prior work [156]. *BW_LPM* performs a weighted time-division multiplexing of the photonic network bandwidth, and leverages the temporal fluctuations in network bandwidth to opportunistically save laser power. *BW_LPM* is designed to perform laser power management in MWMR BWGs [156]. Therefore, we focus on the Flexishare PNoC architecture with MWMR BWGs for our evaluation.

We analyzed power consumption and average packet latency for the *SOA_LPM* and *BW_LPM* techniques when they were integrated into the Flexishare PNoC architecture. For a fair comparison with *BW_LPM*, it is important to enable weighted time-division multiplexing of the network bandwidth in the Flexishare PNoC. Therefore, we enhanced the baseline Flexishare PNoC to enable weighted time-division multiplexing of the network bandwidth using token stream arbitration (TS) in its MWMR BWGs through the laser controller. We refer to this enhanced Flexishare PNoC as *Flexishare-TS*.

We implemented the *BW_LPM* technique on *Flexishare-TS,* and refer to the resulting PNoC configuration as *Flexishare-TS-BW_LPM*. Similar to the *BW_LPM* enhanced PNoC architecture presented in [156], the *Flexishare-TS-BW_LPM* configuration also has four laser sources along

with a laser source controller, which is capable of switching ON/OFF the laser sources based on the executing application bandwidth requirements.

We implemented *SOA_LPM* on *Flexishare-TS* to obtain the *Flexishare-TS-SOA_LPM* PNoC configuration, and compared it with the *BW_LPM* enhanced *Flexishare-TS-BW_LPM* PNoC configuration. We also implemented *BW_LPM* and *SOA_LPM* together and show the combined benefits for the resulting *Flexishare-TS-BW_LPM-SOA_LPM* PNoC configuration.

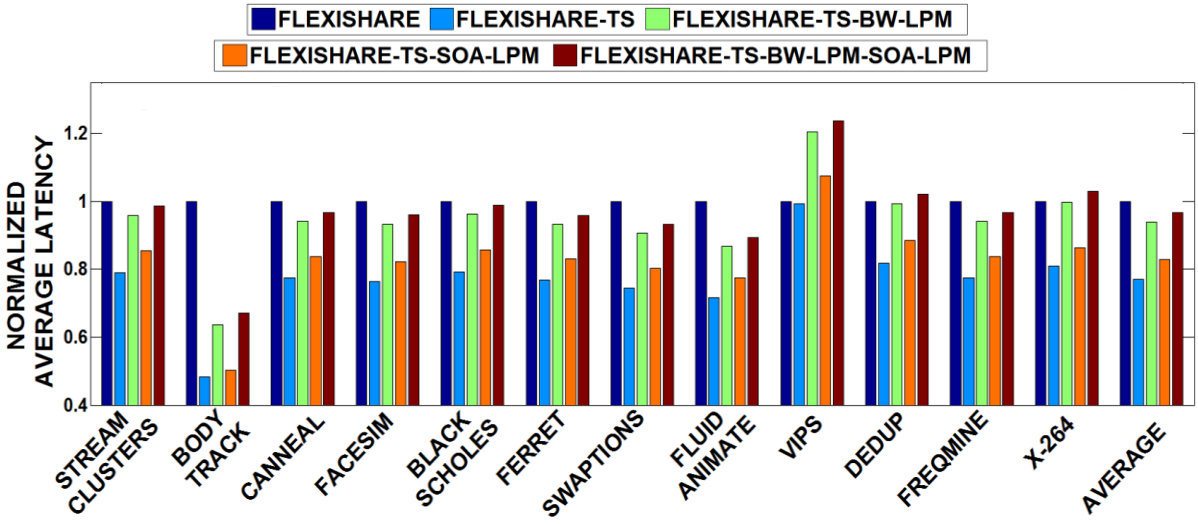


Figure 51: Average latency for different variants of the Flexishare PNoC architecture. Results are normalized wrt baseline Flexishare PNoC.

Figure 51 shows average latency results of this comparison study, with all results normalized with respect to the baseline Flexishare PNoC. As evident from the figure, it can be observed that on average, *Flexishare-TS-SOA_LPM* has 7.7% higher latency compared to *Flexishare-TS*. The inferior values of latency are due to the additional time/cycles required for switching of laser power using comb switches (0.5ns), for accessing the lookup table entries (1ns), and for amplification using SOAs (50ps). Further, *Flexishare-TS-SOA_LPM* has 12.8% lower average latency compared to *Flexishare-TS-BW_LPM*. The increase in average latency for *Flexishare-TS-BW_LPM* is because of additional cycles required for periodic computation of average packet latency in the

BW_LPM technique. *Flexishare-TS-BW_LPM-SOA_LPM* also has 25.8% higher latency compared to *Flexishare-TS*. The latencies contributed by *BW_LPM* and *SOA_LPM* cumulate to render worse average latency for *Flexishare-TS-BW_LPM-SOA_LPM*.

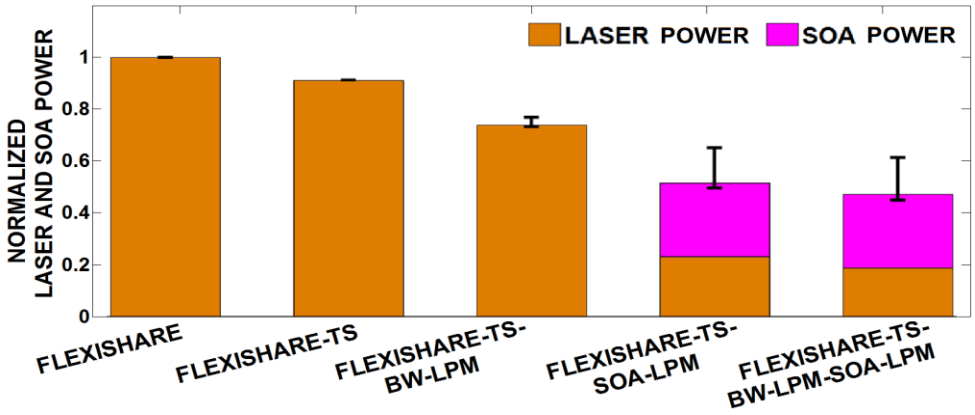


Figure 52: Average laser and SOA power consumption comparison for different configurations of the Flexishare PNoC architecture. Results are normalized wrt the baseline Flexishare PNoC architecture.

Figure 52 presents a comparison of the laser and SOA power consumption values averaged across 12 applications of the PARSEC benchmark suite for the baseline *Flexishare, Flexishare-TS, Flexishare-TS-BW_LPM, Flexishare-TS-SOA_LPM* and *Flexishare-TS-BW_LPM-SOA_LPM* PNoC configurations. The error bars in the figure represent maximum and minimum power values across the 12 applications. *Flexishare-TS-SOA_LPM* has 50.8%, 43.4%, and 30.1% lower total power consumption compared to Flexishare baseline, *Flexishare-TS*, and *Flexishare-TS-BW_LPM* respectively. For *Flexishare-TS-SOA_LPM* and *Flexishare-TS-BW_LPM-SOA_LPM*, more than 50% of the total power consumption is due to the power overhead of SOAs. Even then, these PNoC configurations possess less total power consumption than all other variants. These results corroborate the excellent capabilities of our *SOA_LPM* technique in reducing laser power and hence total power consumption. Moreover, it can be inferred that *SOA_LPM* can be combined with other laser power management techniques such as *BW_LPM* to save even more laser power.

In *summary*, our *SOA_LPM* laser power management technique saves significant power with nominal latency overheads.

6.6. CONCLUSIONS

We presented a low overhead, run-time laser power management technique called *SOA_LPM*, which makes use of on-chip semiconductor amplifiers (SOA) to achieve traffic-independent and loss-aware savings in laser power consumption. Experimental analysis showed that our technique achieves 31.5% more laser power savings with 12.8% less latency overhead compared to another laser power management scheme from prior work. Thus, we demonstrated that *SOA_LPM* represents an attractive solution to reduce laser power consumption in emerging PNoCs.

7. IMPROVING THE RELIABILITY AND ENERGY-EFFICIENCY OF HIGH-BANDWIDTH PHOTONIC NOC ARCHITECTURES WITH MULTI-LEVEL SIGNALING

Photonic network-on-chip (PNoC) architectures employ photonic waveguides with dense-wavelength-division-multiplexing (DWDM) for signal traversal and microring resonators (MRs) for on-off-keying (OOK) based signal modulation, to enable high bandwidth on-chip transfers. Unfortunately, the use of larger number of DWDM wavelengths to achieve higher bandwidth requires sophisticated and costly laser sources along with extra photonic hardware, which adds extra noise and increases the power and area consumption of PNoCs. This chapter presents a novel method (called *4-PAM-P*) of generating four-amplitude-level optical signals in PNoCs, which doubles the aggregate bandwidth without increasing utilized wavelengths, photonic hardware, and incurred noise, thereby reducing the bit-error-rate (BER), area, and energy consumption of PNoCs. Our experimental analysis shows that our *4-PAM-P* signaling method achieves equal bandwidth with 4.2× better BER, 19.5% lower power, 16.3% lower energy-per-bit, and 5.6% less photonic area compared to the best known 4-amplitude-level optical signaling method from prior work.

7.1. INTRODUCTION

In the many-core era, processors with hundreds of cores on a single chip are becoming a reality. The performance of these many-core processors is driven by the effectiveness of the underlying electrical network-on-chip (ENoC) fabrics that are becoming increasingly crosstalk- and energy-limited [8]. To this end, due to the recent developments in silicon photonics, photonic network-on-chip (PNoC) fabrics have been projected to supersede ENoCs. PNoCs offer several

benefits over ENoCs such as higher bandwidth density, distance-independent bit-rate, and smaller data-dependent energy.

Typical PNoC architectures (e.g., [8], [12], [81], and [163]) and off-chip photonic interconnects (e.g., [164] and Chapter 10) utilize several photonic devices such as multi-wavelength lasers, microring resonators (MRs), waveguides (WGs), and splitters. A broadband laser source generates light of multiple wavelengths (λs), each wavelength (λ) of which serves as a data signal carrier. Simultaneous traversal of multiple λ-signals across a single photonic WG is possible using dense wavelength division multiplexing (DWDM), which enables parallel data transfer across the photonic WG. For instance, a DWDM of 64 λs can transfer 64 data bits in parallel. At the source node, multiple MRs typically modulate multiple electrical data signals on the utilized DWDM λs (data-modulation phase). In almost all PNoCs in literature, modulator MRs utilize on-off keying (OOK) modulation, where-in the presence and absence of λs in the WG are used to represent logic '1s' and '0s'. Similarly, at the destination node, multiple MRs with photodetectors at their drop ports are used to filter and detect light-modulated data signals from the WG (data-detection phase) and generate proportional electrical signals. In general, the use of a large number of DWDM λs enables high bandwidth parallel data transfers in PNoCs.

Unfortunately, a number of challenges related to cost [165], reliability (Chapter 3), and energy-efficiency (Chapter 6) still need to be overcome for efficient implementation of PNoCs that utilize a large number of DWDM λs (typically 64 or more DWDM λs per WG). Generating a large number of DWDM λs requires a comb-generating laser source, the ineffectiveness, complexity, and cost of which increase with the number of λs generated [166]. Moreover, utilizing a larger number of DWDM λs to achieve higher-bandwidth data-transfers in a PNoC results in larger network flit size and more electrical and photonic hardware (more number of modulator and

detector MRs and their drivers). A larger network flit size also results in larger sized buffers in the network gateway interfaces, which results in significantly higher area and power overheads. Similarly, larger number of MRs and drivers also incur greater photonic area and MR heating power overheads. Furthermore, the use of a larger number of DWDM λs decreases the gap between two successive λ-channels, which in turn increases the heterodyne crosstalk noise in PNoCs, harming the reliability of communication [56], [61]. Thus, the use of larger number of DWDM λs to achieve higher bandwidth in PNoCs is not a reliable and energy-efficient option. *This motivates the need for a more reliable and energy-efficient way of achieving higher bandwidth data transfers in PNoC architectures*.

In [163] and [167], Kao et al. proposed a multilevel optical signaling format 4-PAM (4-pulse amplitude modulation) to achieve higher bandwidth and energy-efficient data communication in PNoCs. 4-PAM optical signaling format doubles the bandwidth by compressing two bits in one symbol carried out by four levels of amplitude. Kao et al. utilize superposition of two OOK-modulated optical signals of the same λ with 2:1 power ratio to create a 4-PAM λ-signal. We found that this signal superposition based 4-PAM optical signaling method (referred to as 4-PAM-SS henceforth) incurs significantly high power, photonic area, and reliability overheads (Section II). *The shortcomings of 4-PAM-SS motivate the need for a more reliable, energy- and area-efficient method of implementing 4-PAM signaling*.

In this chapter, we present a novel method (referred to as *4-PAM-P* henceforth) of generating 4-PAM optical signals, which employs only one modulator MR per λ to directly modulate the designated λ-signal in 4-PAM format. We present a search heuristic based link optimization framework that finds the optimal value of number of DWDM wavelengths (N_λ) from a constrained space of all its allowable values to achieve desired performance and/or reliability goals for the

target photonic link. We use our framework to optimize the designs of three types of photonic links, each of which uses OOK, 4-PAM-SS, or *4-PAM-P* optical signaling method. Our experimental analysis shows that PNoCs that are comprised of *4-PAM-P* based optimized links render greater reliability, energy-efficiency, and area-efficiency with equal bandwidth compared to the PNoCs that are comprised of 4-PAM-SS or OOK based optimized links. We summarize the key contributions in this chapter as follows:

- We propose a novel technique (*4-PAM-P*) of generating 4-PAM optical signals in PNoCs, which is more reliable, area- and energy-efficient than a previously proposed signal superposition based 4-PAM-SS method [167] and the conventional OOK method;
- We present a search heuristic based optimization framework that optimizes the designs of OOK, 4-PAM-SS, and *4-PAM-P* signaling based photonic links to achieve the desired performance and/or reliability goals;
- We evaluate the impact of the optimized designs of OOK, 4-PAM-SS and *4-PAM-P* based photonic links on the performance, reliability and energy-efficiency of a well-known PNoC architecture: an 8-ary 3-stage CLOS PNoC [163].

7.2. BACKGROUND AND MOTIVATION

In this section, we first present an overview of the signal superposition based 4-PAM optical signaling method (referred to as 4-PAM-SS) from [167]. In 4-PAM-SS, at first, two separate OOK-modulated signals of each of the utilized DWDM λs are generated in two separate parallel WGs. Then, these two sets of OOK-modulated DWDM signals are superposed using a combiner to generate one set of 4-PAM modulated DWDM signals. Each amplitude level of a 4-PAM λ-signal represents one of the four combinations of two bits (00, 01, 10, or 11). When the entire DWDM

spectrum of 4-PAM signals reach the destination node, each signal is filtered by its corresponding in-resonance MR and is converted back into two electrical signals by a photodetector and a back-end receiver circuit. As discussed in [167], the back-end receiver circuit consists of three sense-amplifiers and two logic gates that decode the 4-PAM modulated signal.

Ideally, in the 4-PAM-SS method, when the combiner superposes two OOK-modulated signals, a 4-PAM modulated signal is generated owing to the constructive interference between the two OOK signals. However, the constructive interference happens only if both the OOK signals have identical phases. Unfortunately, in the presence of non-idealities such as process and on-chip temperature variations, a significant phase difference exists between the two superposed OOK signals, which leads to destructive interference between them. Owing to the random nature of process and on-chip temperature variations, this incurred phase difference may fall anywhere in the range from 0 to 2π. This implies that the degree of destructive interference incurred between the OOK signals due to the phase difference (and hence the strengths of the symbols of the resultant 4-PAM signal) may fall anywhere in a very large ranges of values. This in turn makes it very hard to ensure reliability of communication with a 4-PAM-SS photonic link.

The worst-case destructive interference in 4-PAM-SS occurs when the two superposed OOK signals are completely out of phase, i.e., when the phase difference between them is an odd multiple of π. The amount of signal loss due to the superposition of two out of phase OOK signals depends on their individual signal strengths. Typically, as explained in [167], in 4-PAM-SS method, to equidistantly space the four amplitude levels of the output 4-PAM signal in the available range of optical transmission, the strengths of the individual OOK signals are kept to be two-third and one-third of the strength of the conventional OOK signal. Hence, for the best-case constructive interference between the superposed OOK signals, the strength of the resultant 4-

PAM signal becomes 2/3+1/3=1. In contrast, for the worst-case destructive interference, the strength of the resultant 4-PAM signal becomes 2/3-1/3=1/3, causing the worst-case interference-related signal loss to be -10×log(1/3) = 4.8dB.

In summary, the interference-related signal loss in 4-PAM-SS [167] reduces signal-to-noise ratio (SNR), BER, and overall communication reliability. Furthermore, as explained in [167], the 4-PAM-SS method requires additional photonic hardware, such as one asymmetric splitter, two modulator MRs per λ (for two OOK signals), and a combiner. This additional photonic hardware reduces the area benefits of the 4-PAM-SS method compared to the traditional OOK method. In this chapter, we present a novel, more reliable, and energy- and area-efficient 4-PAM signaling method (*4-PAM-P*), which overcomes the shortcomings of the 4-PAM-SS method. The next section describes our proposed *4-PAM-P* method.

7.3. PROPOSED 4-PAM-P OPTICAL SIGNALING

7.3.1. OVERVIEW

Unlike 4-PAM-SS, our proposed *4-PAM-P* method employs only one modulator MR per λ to directly modulate the designated λ-signal in 4-PAM format. This type of 4-PAM signal generating modulator MRs are demonstrated in [168]-[170]. Our proposed *4-PAM-P* method extends the use of such modulator MRs in DWDM based PNoCs. Before we discuss about how our *4-PAM-P* method works, it is important to understand how a modulator MR works in the conventional OOK method.

Ideally, a modulator MR is designed to operate in resonance with a signal-λ in its default state. But due to process and on-chip temperature variations, the MR's resonance λ often deviates from the signal-λ (black curve in Figure 53(a)). In this case, as shown in Figure 53(a), the MR's resonance λ (center/peak of the MR's passband) needs to be brought in alignment with the signal-

λ by either temperature or electrical tuning of the MR [152]. In this tuned state (purple curve), the MR remains in resonance with the signal-λ, which enables the MR to modulate logic 0 on the signal-λ, by removing the signal-λ from the WG. Thus, in this tuned state of the MR, a non-zero tuning-bias voltage (V_T) but zero signal-bias voltage ($V_0 = 0$) is applied to the MR. Hence, as shown in Figure 53(a), the total bias voltage applied to the MR in the tuned state (purple curve) is $V_B = V_T + V_0 \rightarrow V_B = V_T$. On the other hand, the MR is operated in the off-resonance state to modulate logic 1 on the signal-λ. For that, a specific non-zero signal-bias voltage V_1 is applied to the MR, which shifts the passband of the MR to be in off-resonance state (red curve). Thus, the net bias voltage applied in the off-resonance state of the MR is $V_B=V_T+V_1$.

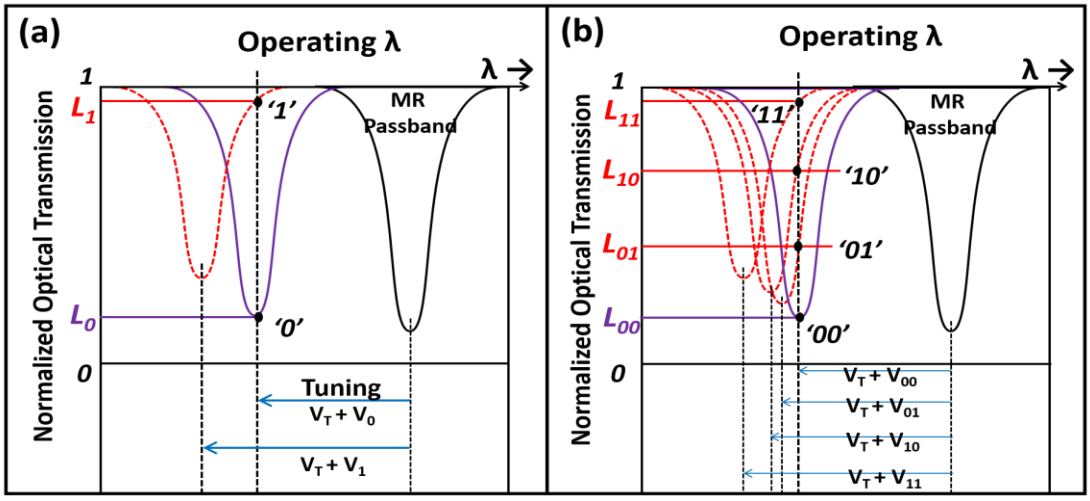


Figure 53: Illustration of optical transmission and microring resonator (MR) spectra for (a) OOK signaling, (b) proposed *4-PAM-P* signaling.

In summary, in conventional OOK modulation, to modulate a particular sequence of 1s and 0s on the signal-λ, the modulator MR is switched off and on resonance with the signal-λ by applying the net bias voltages of V_T+V_1 and V_T+V_0 to the MR, respectively. Note that the value and polarity of V_T depends on the amount and direction of the variation-induced resonance shift, which can be efficiently assessed by using the dithering signal based method demonstrated in [171].

Moreover, in the OOK method (Figure 53 (a)), two levels of optical transmissions (L_0 and L_1) are achieved that correspond to net bias voltages of V_T+V_0 and V_T+V_1. The difference between L_0 and L_1 is defined as modulation depth.

In our *4-PAM-P* method, as shown in Figure 53(b), we introduce two more intermediate levels of optical transmissions L_{01} and L_{10} between L_0 and L_1 within the modulation depth. For that, we introduce two more levels of signal-bias voltages V_{01} and V_{10} between V_0=0 and V_1. Thus, in *4-PAM-P*, signal-bias voltages/optical transmission levels V_0/L_0 (or V_{00}/L_{00}), V_1/L_1 (or V_{11}/L_{11}), V_{01}/L_{01}, and V_{10}/L_{10} correspond to bit combinations "00", "11", "01", and "10" respectively. Note that as demonstrated in [168], the resonance passbands of the carrier-depletion based optimized MRs can be shifted with a signal voltage efficiency of 2GHz/V. This allows MRs with bandwidths even as low as ~10GHz (corresponding to the quality factor of 18000) to have very fine and efficient control of optical transmission levels in the resultant 4-PAM signals [168].

In contrast to the 4-PAM-SS method, *4-PAM-P* does not use signal superposition to create 4-PAM signals, and therefore does not incur interference-related signal loss. As a result, for given noise power, *4-PAM-P* renders greater SNR with better BER. Moreover, *4-PAM-P* requires one less modulator MR per λ, and it does not require additional splitters and combiners. As a result, *4-PAM-P* consumes less photonic area and dissipates less static power related to tuning of MRs. Due to all these benefits, *4-PAM-P* is more reliable, energy- and area-efficient than the 4-PAM-SS.

However, compared to the OOK and 4-PAM-SS methods, the use of *4-PAM-P* requires some minor modifications in how electrical-to-optical (E/O) conversion of data at the sender/modulator side is implemented. At the receiver side, optical-to-electrical (O/E) conversion of data in *4-PAM-P* is carried out in the same manner as in the 4-PAM-SS method, as demonstrated in [167]. Note that from [172] and [167], a 4-PAM signal with the same baud-rate as of an OOK signal requires

4.8dB more received power to achieve the same BER as achievable by the OOK signal. Therefore, for our link- and architecture-level analysis in Section IV and V, we consider the detector sensitivity threshold for both the 4-PAM-SS and *4-PAM-P* methods to be 4.8dB more than the conventional OOK method, as per the SNR_{Target} value in Eq. (52). The next subsection describes how E/O conversion is implemented in our *4-PAM-P* method.

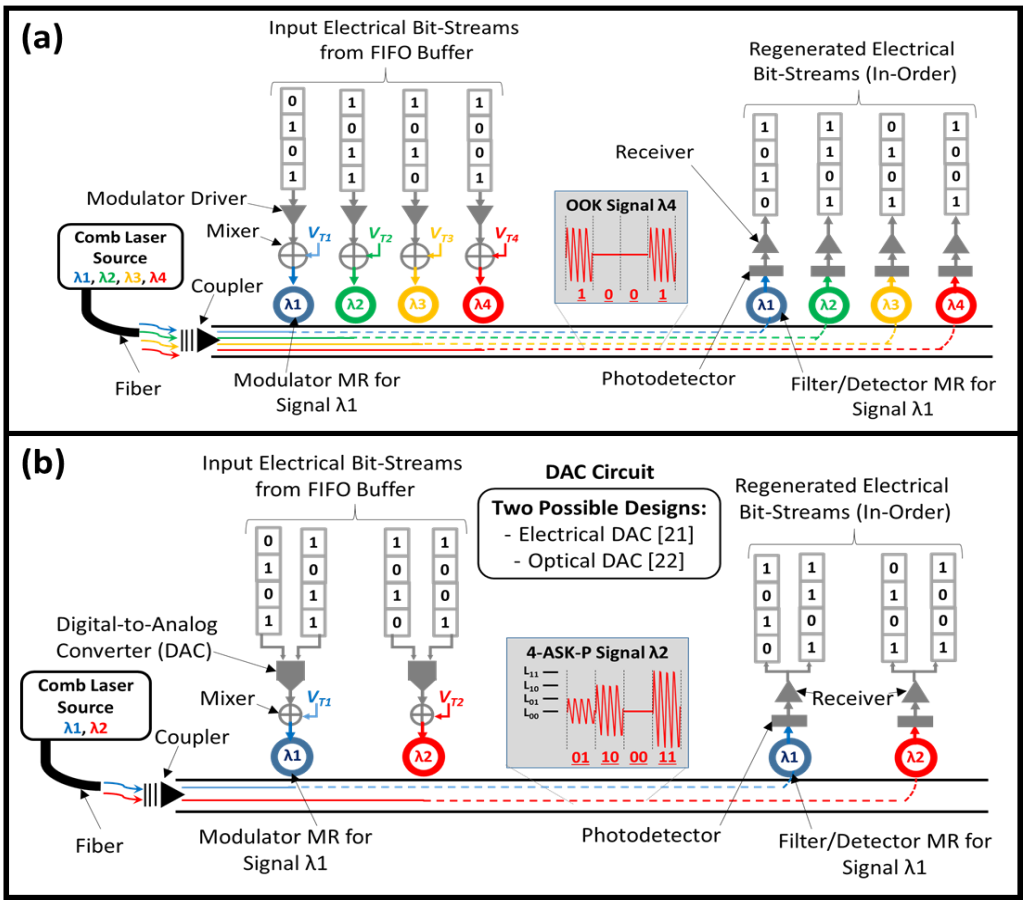


Figure 54: Schematics of (a) OOK, (b) *4-PAM-P* based photonic links.

7.3.2. E/O CONVERSION IN 4-PAM-P METHOD

In a typical PNoC, at the electrical-optical interface of a sender node, the input electrical flits are temporarily stored in first-in-first-out (FIFO) buffers before the modulator MRs convert them into the optical domain. Typically, for PNoCs with equal-sized electrical and optical flits, the size

of each entry of the FIFO buffer is equal to the size of the optical flit, which is equal to the number of bits transferred in parallel on DWDM λs [195]. N parallel electrical bit-streams can be produced when multiple N-bit entries of the FIFO buffer are evicted in sequence, triggered by consecutive clock edges (or levels). The modulator MRs at the E/O interface convert these parallel electrical bit-streams into parallel optical bit-streams.

Figure 54 (a) illustrates E/O conversion for the conventional OOK method, using an example photonic link with 4-bit flit-size. Each of the four parallel electrical bit-streams available is mapped to a designated modulator MR that is designed to operate on a particular λ. Accordingly, each of these bit-streams is applied to a driver circuit, which produces a corresponding sequence/stream of signal-bias voltages V_0 and V_1. Then, the designated modulator MR is driven by this sequence of signal-bias voltages, after the voltage sequence is offset with a constant tuning-bias voltage V_T corresponding to and adjusting for the variation-induced resonance shift in the modulator MR. As explained in the last subsection, because of this applied sequence of bias voltages, each modulator MR (corresponding to an electrical bit-stream) modulates the input electrical bit-stream onto the corresponding DWDM λ, generating an OOK-modulated optical signal.

On the other hand, as shown in Figure 54(b), in the proposed *4-PAM-P* method as well, the FIFO buffer generates four parallel electrical bit-streams. Two adjacent electrical bit-streams of these four bit-streams are applied as inputs to a digital-to-analog converter (DAC) circuit. The DAC coverts each input two-bit combination to a signal-bias voltage level out of four possible voltage levels V_{00}, V_{01}, V_{10}, and V_{11} (see Section 7.3.1). Thus, four parallel electrical bit-streams are converted in two parallel sequences of signal-bias voltages by two concurrently operated DAC units. These two parallel sequences of signal-bias voltages, after being offset by corresponding tuning-bias voltages, are applied to two designated modulator MRs that are designed to operate on

two different λs. These modulator MRs modulate the applied sequences of four-level voltages onto their corresponding λs to generate two parallel four-amplitude-level (4-PAM) optical signals.

Now, as given in the inset of Figure 54(b), two different designs of a high-speed DAC circuit are possible: electrical DAC [169] and optical DAC [170]. In [169], a 40Gbps DAC designed in 65nm CMOS is demonstrated, which utilizes a segmented pulsed-cascode output stage to achieve 4-PAM modulation on a single MR. This electrical DAC (demonstrated in [169]) consumes 3.04pJ/bit power to convert two input electrical bit-streams into a four-level electrical signal (4-PAM electrical signal) at 40Gbps. This 4-PAM electrical signal, after being properly conditioned by signal-bias voltages, is applied to an MR that generates a proportional optical 4-PAM signal. Thus, the conversion of two input electrical bit-streams into a single 4-PAM optical signal happens in two stages. In the first stage, the input bit-streams are converted into an electrical 4-PAM signal by the DAC circuit. Then in the second stage, this 4-PAM electrical signal is converted into optical domain by the modulator MR. The caveat of this electrical DAC based conversion method is that it incurs significant area and energy overhead and imposes non-linearity onto the E/O transfer functions of the driven MRs [170]. To overcome these shortcomings, Moazeni et al. in [170] utilized an optical DAC, which is basically a "spoked" MR of 5μm radius with ~10000 Q that directly converts two input electrical bit-streams into a 4-PAM optical signal for only 0.197pJ/bit power consumption at 20Gbps bit-rate. Thus, the use of these "spoked" MRs collapses the two-stage E/O conversion process into a single stage process, and thus, eliminates the need for external electrical DAC circuit. In this case, the DAC circuits shown in Figure 54(b) are eliminated and the input electrical bit-streams are directly applied to the corresponding "spoked" MRs. We utilize these optical DACs ("spoked" MRs) in our 4-PAM-P method and account for 0.197pJ/bit power

consumption (as part of Tx/Rx power) per DAC in our architecture-level evaluation presented in Section 7.5.

7.4. PHOTONIC LINK DESIGN METHODOLOGY

A naive design of photonic links can result in suboptimal values of bandwidth, power, and reliability for the associated PNoC (Chapter 5). Therefore, irrespective of the utilized optical signaling method, it becomes imperative to optimize the designs of the constituent photonic links to achieve maximum bandwidth, energy-efficiency, and reliability at the PNoC architecture level. From [56] and Chapter 5, for photonic-link design optimization, the number of DWDM λs per waveguide (N_λ) is the most important design parameter, and link power-budget ($P_{\text{Budget}}^{\text{dB}}$) is the most critical design constraint. For optimal use of the available power budget and link bandwidth (Chapter 5), parameters N_λ and $P_{\text{Budget}}^{\text{dB}}$ should meet conditions given in Eq. (50), where the expressions for the constituent terms of Eq. (50) are given in Eq. (51)-(53).

$$P_{Budget}^{dB} \geq P_{Loss}^{dB} + 10\log_{10}(N_\lambda), \tag{50}$$

$$P_{Budget}^{dB} = P_{Max} - S, \tag{51}$$

$$S = 0.5 \times SNR_{Target} + P_{Noise}^{HTC} + P_{Noise}^{Thermal}, \tag{52}$$

$$P_{Loss}^{dB} = P_{Loss}^{MMR} + P_{Loss}^{DMR} + P_{Loss}^{WGP} + P_{Loss}^{WGB} + P_{Loss}^{SpC} + P_{Loss}^{INTRF}, \tag{53}$$

Eq. (52) is derived from the equation for the required photodiode power given in [173]. In Eq. (52), *S* represents the required detector sensitivity threshold (i.e., minimum detectable power) in dBm to achieve the target SNR (SNR_{Target}). Table 9 gives the definitions and values of various parameters used in Eq. (51)-(53). As evident from Table 9, SNR_{Target} is different for OOK and 4-

PAM signals, the reason for which is explained in Section 7.4.2. We use the power drop equation given in [49] to reckon the heterodyne crosstalk noise power (P_{Noise}^{HTC}) for MR filter Q-factor of 9000 and baud-rate (or bit-rate) of 10Gbps, which defines the detector sensitivity threshold *S* in Eq. (52). From [173], the OOK and 4-PAM modulated signals have identical frequency spectra (represented by the *sinc* function). Therefore, the power drop equation given in [49] can be equally applicable to OOK and 4-PAM signals for reckoning P_{Noise}^{HTC} by simply replacing the bit-rate in the equation with the signal baud-rate. Note that, as a single symbol in a 4-PAM signal represents two bits, the bit-rate for a 4-PAM signal is twice its baud-rate (or symbol-rate).

Table 9: Definitions and values of various link design parameters.

Parameter	Definition		Value
P_{Max}	Maximum allowable optical power per WG		20dBm [202]
$P_{Noise}^{Thermal}$	Thermal noise power for detector		-22dBm [211]
SNR_{Target}	Target SNR value	OOK	21.7dB [204]
		4-PAM	31.3dB [204]
P_{Loss}^{WGB}	Waveguide bending loss (dB per 90°)		0.005 [198]
P_{Loss}^{WGP}	Waveguide propagation loss		0.27dB/cm [198]
P_{Loss}^{SpC}	Splitter + coupling loss		1.2dB [198][199]
-	Laser efficiency		30% [199]
-	Detector responsivity		1.1 A/W [199]
-	Detector bandwidth		5GHz
P_{Loss}^{INTRF}	Signal interference loss	*4-PAM*-SS	4.8dB
		4-PAM-P and OOK	0dB

We use Lorentzian function based equations given in [56] and Chapter 8 to reckon through losses of modulator and detector MRs (P_{Loss}^{MMR} and P_{Loss}^{DMR} respectively), all of which depend on N_λ. As can be implied from [174] and [49], for a given photonic link, P_{Loss}^{MMR}, P_{Loss}^{DMR}, and P_{Noise}^{HTC} increase with increase in N_λ due to corresponding decrease in channel gap. Hence, P_{Loss}^{dB}, *S*, and therefore, P_{Budget}^{dB} depend on N_λ (Eq. (51)-(53)). Now, as evident from Eq. (51), a value of N_λ can be

calculated from values of P_{Loss}^{dB} and P_{Budget}^{dB}. However, the values of P_{Loss}^{dB} and P_{Budget}^{dB} also depend on N_λ (Eq. (51)-(53)). Therefore, it can be concluded that N_λ has cyclic dependency on parameters P_{Loss}^{dB} and P_{Budget}^{dB}, i.e., N_λ depends on and also controls the parameters P_{Loss}^{dB} and P_{Budget}^{dB}. Due to this cyclic dependency, the optimal value of N_λ cannot be obtained directly from the parameters P_{Loss}^{dB} and P_{Budget}^{dB}. Therefore, we employ a search heuristic that finds the optimal value of N_λ that satisfies the constraint for given sets of input parameters, from a set of its allowable values.

7.4.1. SEARCH HEURISTIC BASED OPTIMIZATION

Our proposed search heuristic performs exhaustive search to find the optimal constraint-satisfying value of N_λ. To limit the cost and complexity of the comb-generating laser source [166], and to be consistent with the prior works on 4-PAM optical signaling [163] and [167], we limit the maximum allowable value of N_λ to 128. Moreover, as the flit-size of a PNoC is directly proportional to the value of N_λ, and as the flit-size is usually a power-of-two value, the allowable values of N_λ should also be power-of-two values. Because of these reasons, we give a set Λ = {128, 64, 32, 16, 8, 4, 2, 1}, which is a set of all allowable values of N_λ, as an input to our search heuristic. Based on the constraint in Eq. (50), we create an error function $ef(N_\lambda) = \{P_{Budget}^{dB} - P_{Loss}^{dB} - 10\log_{10}(N_\lambda)\}$. Then, for each element N_λ of the set Λ, we evaluate an error value $\epsilon = ef(N_\lambda)$ using Eq. (50)-(53) and parameter values from Table 9, and create a set E of all ϵ values. For that, we evaluate P_{Noise}^{HTC}, P_{Loss}^{MMR} and P_{Loss}^{DMR} values using equations given in [49] and [56] (as mentioned earlier). All N_λ values corresponding to the positive ϵ values in set E satisfy the constraint given in Eq. (50). But we choose the N_λ corresponding to the minimum positive value ϵ_{min} from set E as the

optimal value, because such N_λ is the maximum constraint-satisfying value of the number of DWDM λs.

Note that this search heuristic is equally applicable to OOK, 4-PAM-SS, and *4-PAM-P* methods. However, the optimal values of N_λ would differ for different methods, as the link design parameters such as signal-interference loss P_{Loss}^{INTRF} and SNR_{Target} differ between different methods (see Table 9).

7.4.2. DESIGN FOR RELIABILITY AND BANDWIDTH

We can use the search heuristic given in Section 7.4.1 to find the constraint-satisfying optimal value of N_λ that can achieve either maximum bandwidth in terms of aggregate bit transfer rate or desired reliability in terms of BER for designed photonic links. As can be implied from Chapter 5 and [57], traditionally, the designs of photonic links are optimized to achieve maximum bandwidth, and while doing so the reliability of the designed photonic links is usually disregarded. Therefore, in the link-bandwidth maximizing optimization frameworks presented in Chapter 5 and [57], only the noise-limited detector sensitivity is utilized, and the parameter SNR_{Target} is ignored. From Eq. (52), when the parameter SNR_{Target} is ignored to achieve maximum bandwidth, the detector sensitivity or the minimum detectable signal power (S) is set to be equal to the total noise power. As a result, the available power budget can accommodate greater number of λ-signals (corresponding to a larger value of N_λ) for a given P_{Loss}^{dB}, but each of these wavelength signals yields such a small value of signal power that can be hardly distinguished from the noise power. This results in poor SNR, BER, and communication reliability for the bandwidth-maximized photonic links.

In contrast, we utilize the parameter SNR_{Target} in Eq. (52) and the search heuristic to optimize the links for desired reliability. Introducing the parameter SNR_{Target} in our optimization framework

sets the minimum detectable signal power (S) to a higher level, because of which the available power budget can accommodate only a small number of λ-signals. Nevertheless, doing so ensures that all the supported λ-signals achieve the target SNR. As a result, desired BER and communication reliability can be achieved for the reliability-optimized links. As discussed in [173], BER is a more appropriate measure of reliability than SNR, and a BER of 10^{-9} is a standard target for reliable on-chip communication. Therefore, we choose appropriate values of SNR_{Target} that correspond to BER of 10^{-9} for different signaling methods. From the equations given in [173], and as shown in Table 9, to achieve a BER of 10^{-9}, OOK and 4-PAM signaling require SNR of 21.7dB and 31.3dB, respectively. F*or given noise power*, 4-PAM *signaling requires greater SNR to achieve a BER of* 10^{-9}*, as a given amount of noise power impacts a* 4-PAM *signal more than an OOK signal, due to the decreased gap between different optical transmission levels of the 4-PAM signal.*

Table 10: DWDM degree (optimal N_λ), optical loss and photonic area for different variants of CLOS PNoC.

Configuration	Waveguide DWDM (optimal N_λ)	Worst-case optical loss (in dB)	Optical loss + 10log(N_λ) (in dB)	Photonic area (in mm^2)
Reliability-optimized PNoCs				
CLOS-OOK	64	-1.7	-19.70	2.64
CLOS-4PAM-SS	4	-6.4	-12.40	2.13
CLOS-4PAM-P	16	-1.7	-13.70	2.22
Bandwidth-neutral PNoCs				
CLOS-OOK-BN	64	-1.7	-19.70	2.64
CLOS-4PAM-SS-BN	32	-6.4	-21.40	2.50
CLOS-4PAM-P-BN	32	-1.7	-16.70	2.36

As mentioned in Section 7.1, our goal in this chapter is to evaluate the impact of the optimized designs of OOK, 4-PAM-SS, and *4-PAM-P* based photonic links on the performance,

reliability and energy-efficiency of a well-known PNoC architecture: an 8-ary 3-stage CLOS PNoC [163]. To achieve this goal, we first optimize OOK, 4-PAM-SS, and *4-PAM-P* based single-waveguide photonic links for reliability to achieve BER of 10^{-9}, using our search heuristic given in Section 7.4.1. We optimize 4.5cm long single-waveguide links, as according to our geometric analysis of the CLOS PNoC, and from [163], the longest link of the CLOS PNoC is 4.5cm long. Optimizing a 4.5cm long single-waveguide link corresponds to finding the optimal value of N_λ using the search heuristic when all other parameters in Eq. (50)-(53) and Table 9 are set based on the length and geometrical parameters (e.g., number and degree of bends and number of splitters) of the link. We set the number of DWDM λs for all the links in the CLOS PNoC to be equal to the optimal N_λ obtained for the reliability-optimized 4.5cm long link.

Table 10 gives optimal N_λ and worst-case optical loss (optical loss for the longest 4.5cm link) values for OOK, 4-PAM-SS, and *4-PAM-P* based variants of the CLOS PNoC. We direct the readers to Section V for more information on the architecture of the CLOS PNoC. As our proposed *4-PAM-P* method does not require any additional photonic structures compared to the conventional OOK method and due to the absence of interference induced signal loss, the *4-PAM-P* based variants have worst-case signal loss values that are similar to the OOK based variants (Table 10). As evident from Table 10, the 4-PAM signaling based reliability-optimized variants of CLOS PNoC (CLOS-4PAM-SS and *CLOS-4PAM-P*) render smaller N_λ than the OOK signaling based CLOS-OOK PNoC. This is because the SNR_{Target} value of 31.3dB for 4-PAM signaling is greater than the SNR_{Target} of 21.7dB for OOK signaling, which reduces the available power budget for the 4-PAM signaling based links, rendering smaller N_λ for 4-PAM signaling based variants of CLOS PNoCs. Nevertheless, as will be clear in Section 7.5, despite having inferior bandwidth owing to the smaller N_λ, *CLOS-4PAM-P* PNoC achieves better energy-efficiency than CLOS-OOK.

Based on the standard dimensions and sizes of photonic devices given in Chapter 5, we evaluated the total photonic area consumption of the reliability-optimized variants of CLOS PNoC. The result of this evaluation is also given in Table 10. As evident from the optimal N_λ, worst-case loss, and photonic area results of the reliability-optimized CLOS PNoCs given in Table 10, CLOS-OOK achieves the greatest bandwidth corresponding to the largest N_λ, whereas CLOS-4PAM-SS and *CLOS-4PAM-P* achieve the best values of photonic area and worst-case loss respectively. Thus, a clear winner from the CLOS-4PAM-SS, *CLOS-4PAM-P*, and CLOS-OOK PNoCs cannot be decided by looking at the reliability-optimized results given in Table 10.

To have a fair comparison and to decide the superior method out of the three signaling methods, we evaluate bandwidth neutral designs (with equal aggregate bit transfer rates) of OOK, 4-PAM-SS, and *4-PAM-P* based variants of the CLOS PNoC, referred to as CLOS-OOK-BN, CLOS-4PAM-SS-BN, and *CLOS-4PAM-P-BN*, respectively. We evaluate the worst-case loss, optimal N_λ, and photonic area values for all three bandwidth-neutral CLOS PNoCs, which are listed in Table 10. Note that N_λ=32 for a 4-PAM signal and N_λ=64 for an OOK signal both achieve equal bandwidth (aggregate bit transfer rate), as a 4-PAM signal has 2× bit-rate than an OOK signal (Section 7.2). As evident from Table 10, among all three bandwidth-neutral variants of CLOS PNoC, *CLOS-4PAM-P-BN* achieves the best values of worst-case optical loss (which determines required laser power) and photonic area. Therefore, it can be concluded that *4-PAM-P* method is more area- and energy-efficient than 4-PAM-SS and OOK methods.

For a fairer and more comprehensive comparison of different signaling methods, it is important to evaluate the bandwidth (aggregate bit transfer rate), reliability, and energy-efficiency of all the variants of the CLOS PNoCs listed in Table 10 for real benchmark applications and in the presence of variations. Such an evaluation is presented in the next section.

7.5. EVALUATION

We considered a PNoC with an 8-ary 3-stage CLOS topology [163] for a 256-core system, with 8 clusters (C1-C8) and 32 cores in each cluster. Within each cluster, a group of four cores is connected to a concentrator. Thus, each cluster has 8 concentrators and these concentrators are connected electrically through a router for inter-concentrator communication. The CLOS PNoC uses point-to-point photonic links for inter-cluster communication, with a total of 56 waveguides being used to connect all 8 clusters of the CLOS PNoC. Each point-to-point photonic link uses either forward or backward propagating λs depending on the physical location of the source and destination clusters. This PNoC uses two laser sources to enable forward and backward communication. To power the 56 waveguides, the PNoC employs a series of 1X2, 1X7, and 1X4 splitters. Based on geometric analysis, we estimated the maximum length of a WG in the CLOS PNoC to be 4.5cm. This 4.5cm long WG act as a point-to-point link between cluster C6 and C1.

Modeling of Process Variations of MRs in CLOS PNoC: We adapt the VARIUS tool [71] to model die-to-die (D2D) as well as within-die (WID) process variations in MRs for the CLOS PNoC. We consider a 256-core chip with die size 400mm^2 at a 22nm node. For the VARIUS tool, we use the parameters and procedures given in [56] and [33] to generate 100 process variation (PV) maps, each containing over 1 million points indicating the PV-induced resonance shift of MRs. The total number of points picked from these maps equal the number of MRs in the CLOS PNoC.

Simulation Setup: We performed benchmark-driven simulation-based analysis to evaluate the impact of various signaling methods on the performance and efficiency of the CLOS PNoC architecture. We modeled and simulated all the variants (reliability-optimized as well as bandwidth-neutral) of the CLOS PNoC given in Table 10 using a cycle-accurate NoC simulator. We evaluated performance for a 256-core single-chip architecture at a 22nm CMOS node. We kept

the number of WGs and basic floorplan of the architectures constant across all the variants. We used real-world traffic from applications in the PARSEC benchmark suite [76]. GEM5 full-system simulation [77] of parallelized PARSEC applications was used to generate traces that were fed into our cycle-accurate NoC simulator. In GEM5 simulations, we set a "warm-up" period of 100M instructions and then captured traces for the subsequent 1B instructions. In our benchmark-driven simulations we evaluated total power, average latency, and energy-per-bit (EPB) values for different variants of CLOS PNoC. The results of this evaluation are given in Section 7.5.2 and Section 7.5.3.

Table 11: SNR_{Target}, detector sensitivity, channel gap (CG) between adjacent λs, and dynamic energy for different variants of CLOS PNoC.

Reliability-optimized/Bandwidth-neutral PNoCs	SNR_{Target} (in dB)	Detector sensitivity *S* - in dBm	CG (in nm)	Dynamic energy (in fJ/bit)
CLOS-OOK / CLOS-OOK-BN	21.7/0	-9.2/-22	0.83	2.2
CLOS-4PAM-SS/ CLOS-4PAM-SS-BN	31.3/0	-4.4/-22	17.33/ 1.67	4.4
				2.2
				2.2
				0
CLOS-4PAM-P/ CLOS-4PAM-P-BN	31.3/0	-4.4/-22	3.47/ 1.67	2.2
				0.5
				0.2
				0

Moreover, to evaluate dynamic energy consumption values, we extended the photonic model of the DSENT tool [75] to include all three signaling methods (OOK, 4-PAM-SS, and *4-PAM-P*). Table 11 gives SNR_{Target} values, dynamic energy values, channel gap between adjacent λs, and detector sensitivity (*S*) values (evaluated using Eq. (52)) for all the variants of the CLOS PNoC considered in our evaluation. In the table, the detector sensitivity values directly correspond to

SNR_{Target} values according to Eq. (52). The channel gap values in Table 11 correspond to a free-spectral range of 50nm and N_λ values given in Table 10.

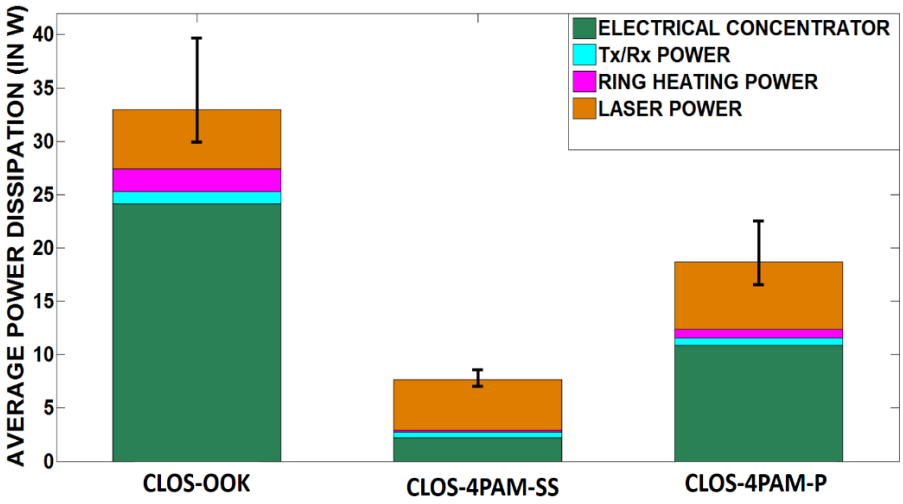
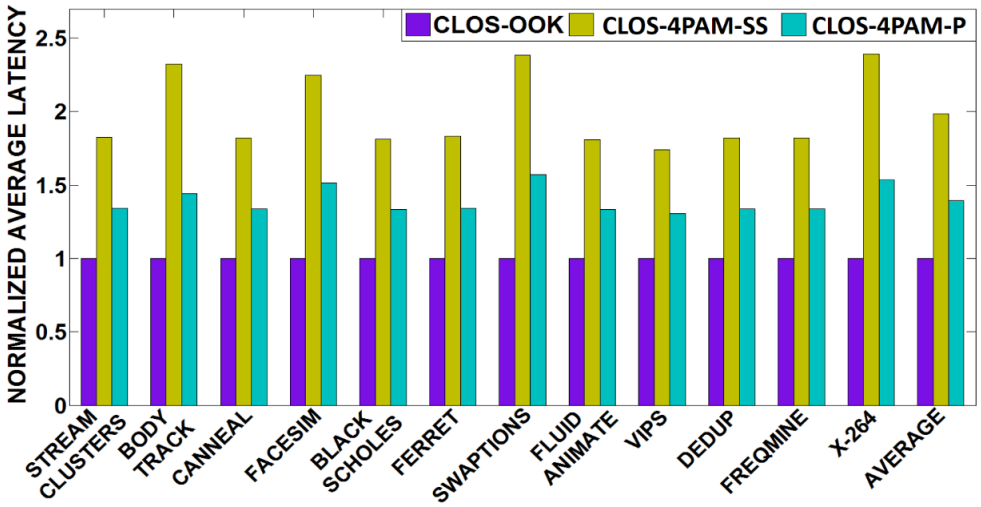


Figure 55: Average total power dissipation comparison for different reliability-optimized configurations of the CLOS PNoC architecture.

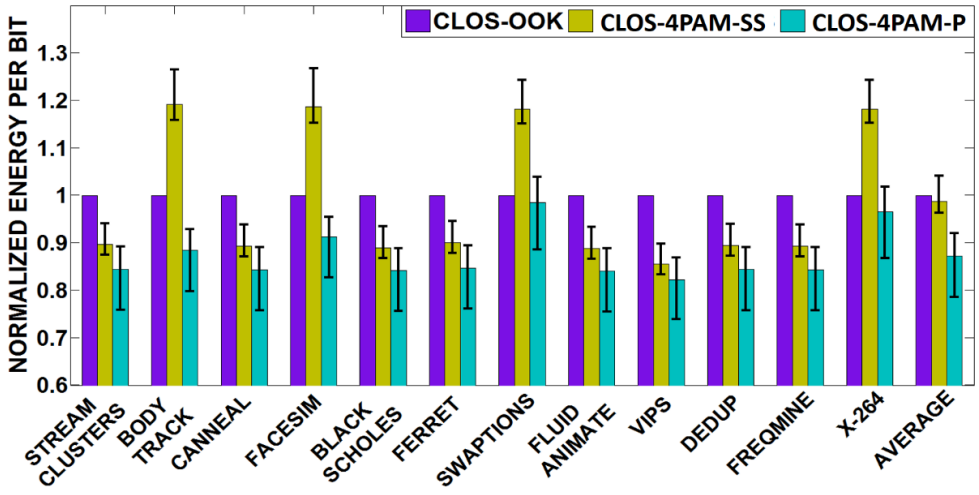
7.5.1. RESULTS FOR RELIABILITY-OPTIMIZED CLOS PNOCS

Figure 55 presents a comparison of total power dissipation values for the CLOS-OOK, CLOS-4PAM-SS, and *CLOS-4PAM-P* PNoCs. The power dissipation values in this figure are averaged across different PARSEC benchmark applications. The error bars in the figure represent maximum and minimum total power values across the 100 PV maps. As evident, compared to CLOS-OOK, *CLOS-4PAM-P* and CLOS-4PAM-SS dissipate 45.2% and 77.2% lower total power respectively. From Table 10, CLOS-OOK has the largest N_λ, whereas CLOS-4PAM-SS has the smallest N_λ. The largest value of N_λ results in the largest flit-size and hence the largest buffer-size, which in turn results in the highest power dissipation in electrical concentrators. Moreover, the largest N_λ also results in the highest number of MRs, which translates into the highest amount of MR heating power. Due to these reasons, CLOS-OOK dissipates the highest power compared to the other two variants. In contrast, the smallest value of N_λ results in the lowest power dissipation

for CLOS-4PAM-SS. Moreover, from Table 10, as the reliability-optimized CLOS-4PAM-P has greater N_λ than CLOS-4PAM-SS, CLOS-4PAM-P has higher power dissipation.



(a)



(b)

Figure 56: (a) Average packet latency, (b) energy-per-bit comparison for different reliability-optimized variants of CLOS PNoC. All results are normalized to the baseline CLOS-OOK PNoC results.

Figure 56 (a), (b) present average packet latency and aggregate energy-per-bit (EPB) for all three variants of the CLOS PNoC across 12 PARSEC benchmarks. The error bars in Figure 56(b) represent maximum and minimum EPB values across the 100 PV maps.

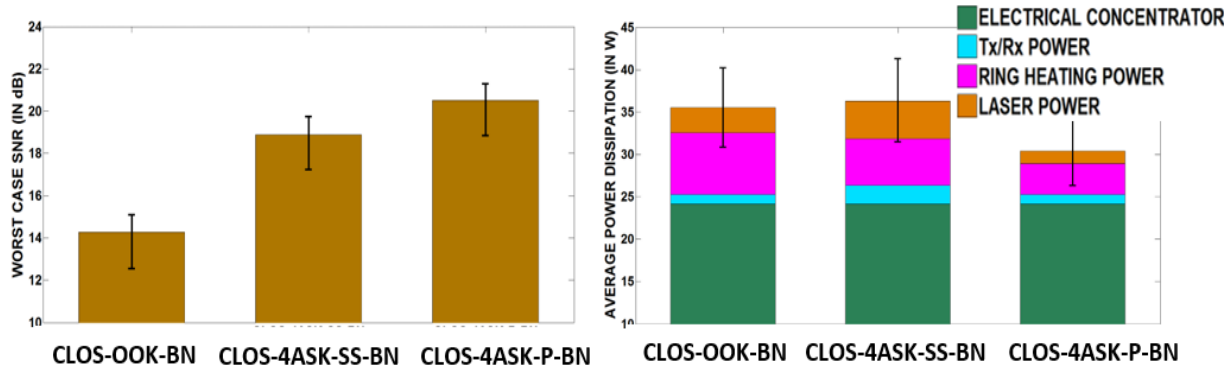
As evident from Figure 56(a), *CLOS-4PAM-P* achieves 29.8% lower average latency than CLOS-4PAM-SS, whereas CLOS-OOK achieves 50% and 28.8% lower average latency than CLOS-4PAM-SS and *CLOS-4PAM-P* respectively. The larger value of N_λ results in increased simultaneous bit transfers, which in turn renders lower average latency for *CLOS-4PAM-P* compared to CLOS-4PAM-SS. Similarly, the largest N_λ results in the lowest average latency for CLOS-OOK. From Figure 56(b), *CLOS-4PAM-P* has 12.7% and 11.5% lower EPB compared to CLOS-OOK and CLOS-4PAM-SS respectively. The 4-PAM signaling used in *CLOS-4PAM-P* makes better use of MR heating power and electrical concentrator power by modulating two bits using only one modulator MR. As a result, *CLOS-4PAM-P* renders better energy-efficiency in terms of lower EPB than CLOS-OOK. Interestingly, CLOS-4PAM-SS also uses 4-PAM signaling, but the higher value of P_{Loss}^{INTRF} for 4-PAM-SS increases the signal loss, which results in very small N_λ, and hence, worse energy-efficiency (EPB) for CLOS-4PAM-SS.

As evident from Figure 56(b), CLOS-4PAM-SS yields lower EPB than CLOS-OOK on average, as CLOS-4PAM-SS dissipates 45.2% less power, but it yields 2× average latency than CLOS-OOK on average. But for some applications such as *Bodytrack*, *Facesim*, *Swaptions*, and *X-264*, CLOS-4PAM-SS yields greater EPB than CLOS-OOK. This is because for these applications, CLOS-4PAM-SS yields greater than 2× latency compared to CLOS-OOK, the effect of which translates into greater EPB. Note that the BER for all the three reliability-optimized variants of the CLOS PNoC is 10^{-9}. *It can be observed from these results that 4-PAM-P signaling method is more energy-efficient while being equally reliable compared to OOK and 4-PAM-SS.*

7.5.2. RESULTS FOR BANDWIDTH-NEUTRAL CLOS PNOCS

As explained in Section 7.4.1, the bandwidths of all the bandwidth-neutral variants (CLOS-OOK-BN, CLOS-4PAM-SS-BN, and *CLOS-4PAM-P-BN*) are equal. Therefore, we do not present

the bandwidth (aggregate bit transfer rate) or average latency results in this section. Instead, we only present total power dissipation, EPB, and SNR/BER results.



(a) **(b)**

Figure 57: (a) Worst-case SNR, (b) average total power dissipation for different bandwidth-neutral configurations of the CLOS PNoC.

Figure 57(a) plots the worst-case SNR values for different bandwidth-neutral variants of the CLOS PNoC. CLOS-OOK-BN, CLOS-4PAM-SS-BN, and *CLOS-4PAM-P-BN* yield SNR values of 14dB, 19dB and 20.5dB respectively, which translates into BER values of 6.2×10^{-3}, 1.7×10^{-2}, and 4.1×10^{-3} respectively (using the equations given in [173]). It can be implied from these results of SNR and BER that *CLOS-4PAM-P-BN* achieves greater communication reliability (corresponding to smaller BER) than CLOS-4PAM-SS-BN and CLOS-OOK-BN. As shown in Table 10 and Table 11, among all the bandwidth-neutral PNoCs, *CLOS-4PAM-P-BN* has the least signal loss and the largest channel gap. The largest channel gap results in the least value of heterodyne crosstalk noise power P_{Noise}^{HTC} [49], [56]. Due to the combined effects of these factors, *CLOS-4PAM-P-BN* achieves the best SNR, BER, and communication reliability.

Figure 57(b) gives average power dissipation for different bandwidth-neutral variants of the CLOS PNoC. *CLOS-4PAM-P-BN* has 16.9% and 19.5% lower total power dissipation compared to CLOS-OOK-BN and CLOS-4PAM-SS-BN respectively. Decrease in laser power (due to

decrease in through losses), ring heating power (due to decrease in number of MRs) and dynamic energy (as shown in Table 11) contributes to decrease in total power dissipation of *CLOS-4PAM-P-BN* compared to CLOS-4PAM-SS-BN and CLOS-OOK-BN.

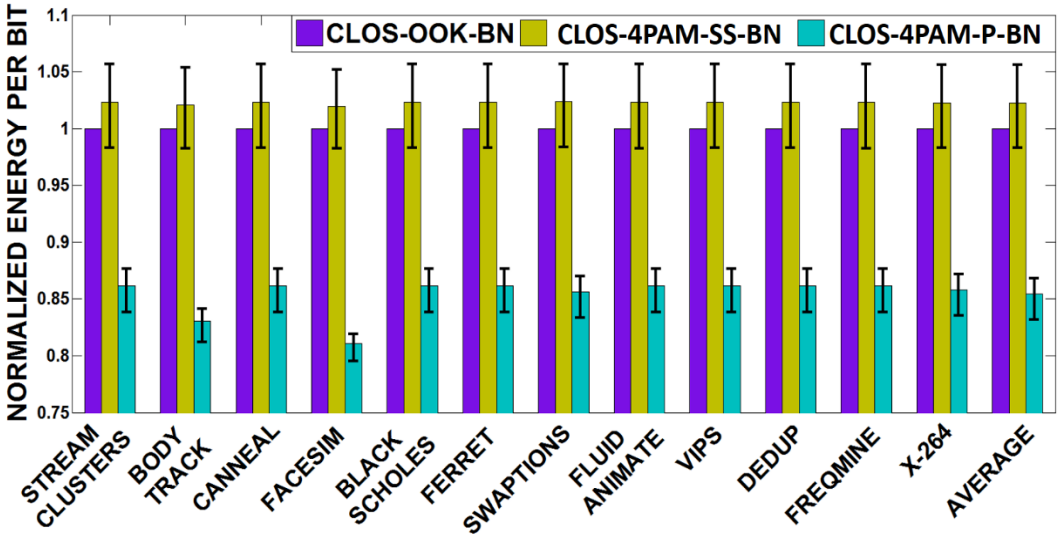


Figure 58: Energy-per-bit comparison for different bandwidth-neutral variants of CLOS PNoC across PARSEC benchmarks. All results are normalized to the baseline CLOS-OOK-BN PNoC results.

Lastly, Figure 58 presents EPB values for all three bandwidth-neutral variants of the CLOS PNoC across PARSEC benchmarks. On an average, the *CLOS-4PAM-P-BN* has 14.6% and 16.3% lower EPB compared to CLOS-OOK-BN and CLOS-4PAM-SS-BN respectively. Decrease in total energy consumption of *CLOS-4PAM-P-BN* compared to CLOS-OOK-BN and CLOS-4PAM-SS-BN reduces its EPB. Note that, in Figure 57(a), (b), and Figure 58, the error bars represent the maximum and minimum values across the 100 PV maps.

7.6. CONCLUSIONS

This chapter presented a novel method, called *4-PAM-P*, for generating 4-PAM optical signals in PNoCs, which can double the aggregate bandwidth without increasing utilized

wavelengths, photonic hardware, and incurred noise, thereby improving the bit-error-rate (BER), area-efficiency, and energy-efficiency of PNoCs. Our analysis showed that our *4-PAM-P* method achieves equal bandwidth with 4.2× better BER, 19.5% lower power, 16.3% lower energy-per-bit, and 5.6% less photonic area compared to the best known 4-PAM optical signaling method (4-PAM-SS) from prior work. Moreover, our *4-PAM-P* method achieves equal bandwidth with 1.5× better BER, 16.9% lower power, 14.6% lower EPB, and 10.6% less photonic area compared to the conventional OOK signaling method. These results corroborate the excellent capabilities of our proposed *4-PAM-P* method in achieving high-bandwidth data transfers in PNoCs with greater reliability, area- and energy-efficiency.

8. ANALYZING VOLTAGE BIAS AND TEMPERATURE INDUCED AGING EFFECTS IN PHOTONIC INTERCONNECTS FOR MANYCORE COMPUTING

To enable MRs to modulate and detect DWDM photonic signals, carrier injection in MRs through their voltage biasing is essential. But long-term operation of MRs with constant or time-varying temperature and voltage biasing causes aging. Such voltage bias temperature induced (VBTI) aging in MRs leads to resonance wavelength drifts and Q-factor degradation, which increases signal loss and energy delay product in photonic NoCs (PNoCs) that utilize photonic interconnects. This chapter explores VBTI aging in MRs and demonstrates its impacts on PNoC architectures for the first time. Our system-level experimental results on two PNoC architectures indicate that VBTI aging increases signal loss in these architectures by up to 7.6dB and increases EDP by up to 26.8% over a span of 5 years.

8.1. INTRODUCTION

MRs can be either in-resonance or out-of-resonance with respect to the utilized DWDM wavelengths. In resonance mode, an MR couples/removes light of the resonant wavelength from the waveguide, and hence, modulates logic "0" (represented by the absence of light in the waveguide) on the resonant wavelength. In contrast, in the out-of-resonance-mode, an MR does not couple any light from the waveguide, and hence, modulates logic "1" (represented by the presence of light in the waveguide) on the resonant wavelength. Thus, a particular sequence of 1s and 0s can be modulated on a wavelength by switching the corresponding MR off and on resonance with the wavelength in the same sequence. MRs can employ either voltage biasing [21] or heating [22] to switch from resonance-mode to out-of-resonance-mode or vice versa. *However, voltage*

biasing is preferred over heating [22] to switch resonance-modes of MRs, as it is faster and dissipates lower power.

To facilitate switching of resonance-modes of an MR with voltage biasing, a PN junction is created in the silicon (Si) core of the MR surrounded by silicon-di-oxide (SiO_2) cladding. A positive/negative voltage bias is applied to this PN-junction to inject/remove free carriers into/from the MR's Si core. For high frequency operation and lower power consumption, an MR's PN-junction is typically operated under a negative voltage bias or reverse bias (otherwise known as carrier depletion mode of an MR). The application of this voltage bias generates an electric field across the MR's Si (core) and SiO_2 (cladding) boundary. Similar to MOSFETs, this electric field generates voltage bias temperature induced (VBTI) traps at the Si-SiO_2 boundary of the MR over time (i.e., VBTI aging). Our analysis has shown that these VBTI aging induced traps alter carrier concentration in the Si core of MRs, which incur resonance wavelength drifts and increase optical scattering loss in MRs to decrease Q-factor of MRs.

In this chapter, for the first time, we study the VBTI aging in MRs and its impact on PNoC architectures. At the device-level, we carefully developed analytical models for trap generation with VBTI aging in MRs. We also devise analytical models that determine variations of MR resonance wavelength shifts and Q-factor with aging-induced traps. These models are further extended to examine the impact of different operating temperatures and bias voltages, as well as process variations. From those models, we follow a mathematical bottom–up approach to analyze the system-level impact of aging on different PNoC architectures. We present our aging analysis on well-known Corona [14] and Clos [102] PNoCs running real-world multi-threaded PARSEC [76] benchmarks.

8.2. RELATED WORK

Recent research on silicon photonics for manycore computing has focused on exploring a wide spectrum of network topologies and protocols to enable efficient PNoC architectures [16], [59], [102]. PNoCs utilize several photonic devices such as MRs as modulators and detectors, waveguides, splitters, and trans-impedance amplifiers (TIAs). The reader is directed to [61] and Chapter 3 for more discussion on these devices.

Fabrication-induced process variations (PV) impact the cross section, i.e., width and height, of photonic devices such as MRs and waveguides [33] and [175]. In MRs, PV causes resonance wavelength drifts, which can be counteracted by using device-level techniques such as voltage biasing (aka localized trimming) and heating (aka thermal tuning). On the other hand, thermal variations (TV) also alter the resonance wavelength of MRs, because of variations in refractive index of the core of MRs due to thermo-optic effects. Similar to PV, resonance wavelength drifts due to PV are compensated by voltage biasing and heating [22]. A few prior works have explored the impact of PV and TV on photonic links at the system-level [34], [56], [63], [121]. In [63], a methodology to salvage network-bandwidth loss due to PV-drifts is proposed, which reorders MRs and trims them to nearby wavelengths. In [34], a thermal tuning based approach is presented that adjusts chip temperature using dynamic voltage and frequency scaling (DVFS) to compensate for chip-wide PV-induced resonance shifts in MRs. In [121], a tunable laser source design is demonstrated, in which the signal power at the source is adapted to compensate for signal losses due to temperature and process variations across photonic interconnects. *All of these works ignore the harmful effects of PV and TV remedies on aging in MRs.*

Aging has become an important reliability concern for ultra-scaled semiconductor devices with significant implications for both analog and digital circuit design. The most important aging

mechanisms in CMOS devices include bias temperature instability (BTI) aging and hot carrier injection (HCI) aging. BTI causes a threshold voltage increase in MOSFETs due to trap generation at the Si-SiO_2 interface [176]. Negative BTI (NBTI) is observed in pMOSFETs, and it usually dominates the positive BTI (PBTI) observed in nMOSFETs [176]. A few prior works have analyzed the impact of NBTI aging mechanisms on MOSFET devices at the device-level. Different hydrogen diffusion models are proposed in [177] to determine trap generation at the Si-SiO_2 interface of pMOSFETs. In [178], models for trap generation in the Si-SiO_2 interface of reduced cross-section MOSFETs (e.g., narrow-width planar MOSFET, triple gate MOSFET, and surround-gate MOSFET) are presented. *However, none of these works considers the impact of aging on MRs and its implications on DWDM-based PNoCs.*

In view of the shortcomings of prior work, in this chapter we aim to analyze VBTI aging in MRs, quantify its dependence on temperature and bias voltage, and explore its impact at the PNoC architecture level.

8.3. TRIMMING (VOLTAGE BIAS) INDUCED MR AGING

8.3.1. OVERVIEW OF VOLTAGE BIAS INDUCED TRAP GENERATION IN MRS

MRs, waveguides, splitters, couplers, and TIAs are basic building blocks of PNoCs [12], [81]. MRs are essentially looped photonic waveguides with a small diameter (~a few μm), and these MRs serve as modulators to write data and detectors to read data. MRs when coupled to a waveguide in resonance-mode remove specific (resonant) wavelengths from the waveguide, whereas in the non-resonance-mode they let wavelengths simply pass through without removing them. MRs employ voltage biasing via carrier injection or removal to shift between resonance and non-resonance modes. To enable carrier injection into and removal from an MR, as shown in Figure 59, a PN junction is created in an MR's Si core surrounded by SiO_2 cladding. To switch

resonance modes at high frequency with low power dissipation using voltage biasing, an MR's PN junction needs to be reverse biased [35], which is accomplished by applying higher voltage on the n side of the PN junction (Figure 59).

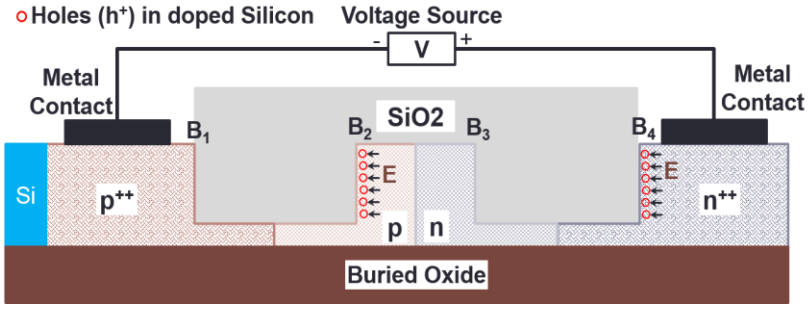
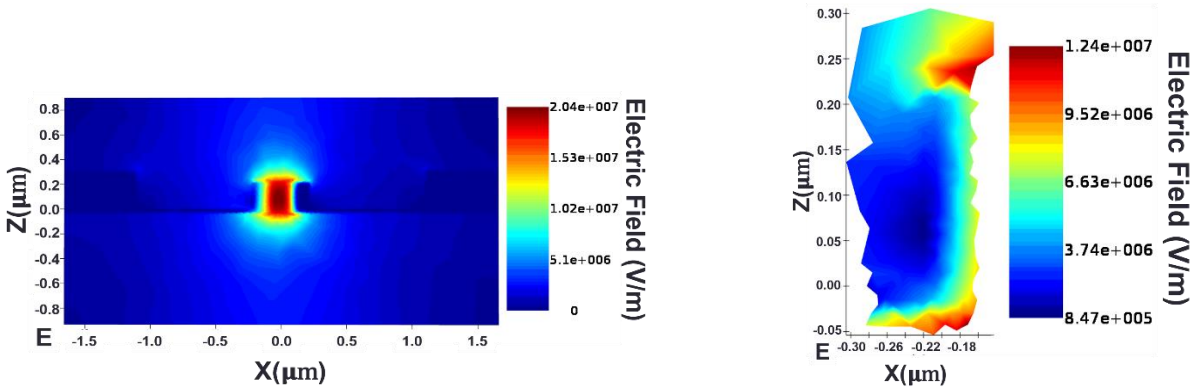


Figure 59: Cross-section of a tunable MR with PN junction in its core to facilitate carrier injection into and removal from core with voltage biasing.



(a) **(b)**

Figure 60: Distribution of electric field (E) across (a) MR waveguide, (b) Si-SiO_2 boundary B2 when -4V bias voltage is applied across PN junction.

When a negative voltage is applied across the PN junction of an MR, an electric field 'E' is generated from right to left across the Si-SiO_2 boundaries B_1, B_2, B_3, and B_4 (Figure 59). We used the Lumerical Solutions DEVICE [65] tool to construct and model the PN junction of an MR. For our preliminary analysis, we consider an MR waveguide similar to the one reported in [139] with a radius of 2µm, fabricated using standard Si-SiO_2 material with a core cross-section of 450nm ×

250nm. We then simulated the MR, using the charge transport solver in the DEVICE tool with a solver geometry of 2D y-normal, and then obtained the distribution of electric field as shown in Figure 61(a) across the MR waveguide with a bias voltage of -4V. The results from the DEVICE tool in Figure 60(a) demonstrate the presence of electric field E across all the Si-SiO_2 boundaries (i.e., B_1, B_2, B_3, and B_4). This electric field present across the Si-SiO_2 boundaries B_2 (Figure 60(b)) and B_4 attracts holes towards them (Figure 59) and generates traps across these boundaries similar to pMOSFETs [176]. However, only the traps on the B_2 boundary change the electro-optic dynamics of the MR core as it is a boundary of the MR core. Thus in this chapter we focus on analyzing trap generation on the B_2 boundary.

8.3.2. TRAP GENERATION ANALYTICAL MODEL FOR MRS

The trap generation model on the B_2 boundary of an MR is based on Si-SiO_2 boundary related hydrogen dynamics [179]. The trap generation takes place at the Si-SiO_2 boundary which is a rough surface where the highly ordered Si core and the amorphous SiO_2 cladding meet. At the junction of these dissimilar materials, some of the Si atoms from the core remain dangling without satisfied chemical bonds, thus forming boundary traps. The traps generated at the Si-SiO_2 boundary of an MR are similar to the traps generated at the Si-SiO_2 boundary of a MOSFET [179]. To improve MR performance, there is a need to reduce these boundary traps. So similar to MOSFETs, MRs are annealed in ambient hydrogen during the manufacturing process. In the presence of an electric field and thermal variations across the Si-SiO_2 boundary, the Si−H bond breaks and the hydrogen gas diffuses into the MR's SiO_2 cladding, thereby yielding passivated Si bonds (Si^*) that act as traps. Furthermore, the direction of electric field (see Figure 59) across the MR's Si-SiO_2 boundary is similar to the direction of electric field across the MOSFET's Si-SiO_2

boundary. Therefore, at a particular temperature both MRs and MOSFETs are have a similar trap generation behavior at their respective Si-SiO_2 boundaries.

Several prior works (e.g., [176]-[178]) use reaction-diffusion (RD) models to characterize boundary trap generation at the MOSFET Si-SiO_2 boundary. As boundary traps in MR's are similar to boundary traps in MOSFETs, we use the same RD model to model the boundary trap generation at the MR's Si-SiO_2 boundary. This trap generation mechanism is represented as a chemical reaction in Eq. (54), where holes (h^+) in the MR's Si core weaken a Si−H bond and hydrogen (H) is detached [177] in the presence of electric field and thermal variations:

$$Si{-}H + h^+ \leftrightarrow Si^* + H \tag{54}$$

The generated Si dangling bond (Si^*) acts as a donor-like boundary trap. The H ion released from the bond can diffuse away from the Si-SiO_2 boundary or anneal an existing trap. The boundary trap density (N_{BT}), increases with the net rate of the reaction given in Eq. (55):

$$\frac{dN_{BT}}{dt} = k_F[N_0 - N_{BT}] - k_R N_{BT} N_H \tag{55}$$

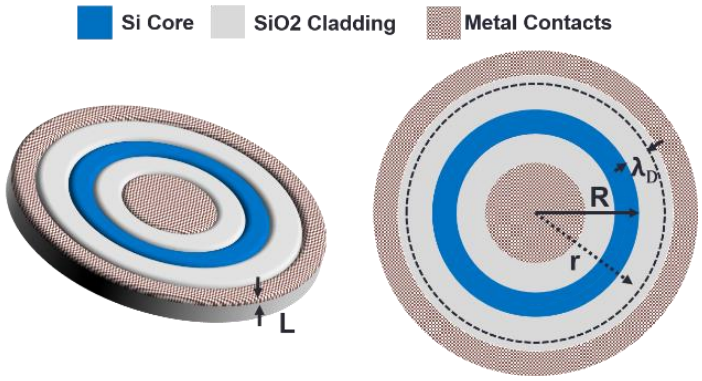
where k_F, k_R, N_0, and N_H are bond-breaking rate, bond-annealing rate, Si–H bond density available before stress, and hydrogen density at the MR's Si-SiO_2 boundary, respectively. From Eq. (55) it can be obserevd that the boundary trap generation rate increases with decrease in H ion density (N_H) at the Si-SiO_2 boundary. The diffusion of H ions away from the traps removes hydrogen from the boundary, so the boundary trap generation rate becomes limited to the diffusion rate of hydrogen. The diffusion rate of hydrogen obeys Eq. (56) [178]:

$$\frac{dN_H}{dt} = D_H \frac{d^2 N_H}{dy^2} \tag{56}$$

where D_H is the diffusion constant of hydrogen, dt is the change in time, and dy is the change in difussion distance. During the diffusion-dominated regime, the dN_{BT}/dt term is negligible compared to the other two terms in Eq. (55) and N_{BT} is significantly smaller than N_0 [178], therefore Eq. (55) can be simplified as:

$$N_{BT}N_H = \frac{k_F N_0}{k_R} \tag{57}$$

Further, the dependence of the rate of boundary trap generation on the electric field across the boundary is included in the k_F term and the temperature dependence of trap generation is incorporated via the activation energies of k_F , k_R and D_H (see Sections 0, 0).



(a) **(b)**

Figure 61: (a) MR 3D-view with Si-core, SiO_2-cladding, and metal contacts for voltage biasing, (b) top view of MR which shows hydrogen diffusion length (λ_D) across its cladding.

From the RD model presented above, the number of traps generated at the Si-SiO_2 boundary is equal to the number of hydrogen ions diffused away from the boundary. But this hydrogen diffusion depends on the geometry of the boundary. The effect of the geometry of hydrogen diffusion on the trap generation rate can be analyzed with the concept of the diffusion length λ_D, which is the distance travelled by hydrogen ion into SiO_2. As outer boundary (i.e. B_2) of an MR is similar to the surround-gate cylindrical MOSFET [177], the λ_D of an MR is similar to this

MOSFET. Therefore, based on estimations from prior works [177], [178] using Eq. (54) this diffusion length λ_D is estimated to be $(D_H*t)^{0.5}$. For the MR with outer boundary (i.e. B_2) radius R and height or thickness L depicted in Figure 61, the hydrogen diffusion is confined within the distance $R < r < R+ \lambda_D$, as shown in Figure 61(b). To determine the total hydrogen ions available within $R < r < R+\lambda_D$, there is a need to integrate all the hydrogen ions between R and $R+\lambda_D$. Thus the hydrogen profile is expressed in cylindrical coordinates and the integral becomes:

$$N_{BT}(t) = \frac{1}{2\pi RL}\int_{R}^{R+\lambda_D} N_H\left(1 - \frac{r-R}{\sqrt{D_H t}}\right) 2\pi r L\, dr \tag{58}$$

Solving Eq. (58) and substituting N_H from Eq. (57), the interface-trap density is calculated from the geometry-dependent R–D relation as:

$$N_{BT}(t) = \sqrt{\frac{k_F N_0}{k_R}}\left(\lambda_D\left(1+\frac{R}{\lambda_D}\right)(2R+\lambda_D) - \left(\frac{R^2+R(R+\lambda_D)+(R+\lambda_D)^2}{3}\right)\right)^{0.5} \tag{59}$$

From the above model it is clear that trap generation on an MR's Si-SiO_2 boundary not only depends on the operational time but also on the geometry of the boundary. *These traps are the main cause of aging in MRs. In the next subsection, we analyze how such boundary trap-induced aging impacts MR optical properties.*

8.3.3. AGING IMPACT ON MR RESONANCE WAVELENGTH AND Q-FACTOR

As discussed in the previous subsection, each trap generated on the core-cladding boundary of an MR consumes a hole from the P side of the MR core (Eq. (54)). Therefore, number of holes consumed in the silicon core is equal to number of boundary traps generated, which is otherwise $N_{BT} \approx -\Delta N_h$, where ΔN_h is the increase in free hole concentration and the negative sign represents decrease in free hole concentration. The removal of holes increases the refractive index of the core

(n_{si}) of a circular MR waveguide, which induces a red shift in an MRs' resonance. The increase in the MR's core refractive index also increases refractive index contrast between the core and cladding ($n_{Si} - n_{SiO2}$), which in turn increases the scattering related optical loss in the MR waveguide [180]. The increase in optical loss causes a decrease in MR Q-factor, which increases MR insertion loss. We quantify and model these phenomena in the rest of this section.

The change in hole concentration in an MR's core due to an MR aging induces refractive index change of Δn_{si} at around 1550nm wavelength, which can be quantified as follows [54]:

$$\Delta n_{si} = -8.8 \times 10^{-22} \Delta N_e - 8.5 \times 10^{-18} (\Delta N_h)^{0.8}, \quad (60)$$

where, ΔN_e and ΔN_h are the increase in free electron concentration and free hole concentration, respectively. Then, the increase in refractive index (positive n_{si} as a result of aging-induced negative ΔN_h) incurs resonance wavelength red shift ($\Delta \lambda_{RWRS}$) as per the following equation [54]:

$$\Delta \lambda_{RWRS} = \frac{\Delta n_{si} * \Gamma * \lambda_r}{n_g}, \quad (61)$$

where, λ_r is the initial resonance wavelength of the MR, n_g is the group refractive index (ratio of speed of light to group velocity of all wavelengths traversing the waveguide) of the MR, and Γ is the confinement factor describing the overlap of the optical mode with the MR waveguide's silicon core. The value of Γ and n_g for MR considered in our analysis are set to 0.7 and 4.2 respectively [65]. From [54], n_g accounts for refractive index dispersion and change in free carrier concentration (and hence, aging) does not significantly affect it.

An increase in the MR core's refractive index (Δn_{Si}) also increases its scattering loss coefficient. The scattering loss coefficient (that causes a fractional loss in signal amplitude) of an

MR's circular waveguide is proportional to the size of the surface roughness σ, and is given by the following equation [55], [68]:

$$\alpha_{\text{scatter}} = \frac{4(\cos\theta)^3 k_0{}^2 n_1{}^2 \sigma^2}{\sin\theta} \cdot \left(\frac{k_0\sqrt{n_1{}^2(\sin\theta)^2 - n_2{}^2}}{L k_0 \sqrt{n_1{}^2(\sin\theta)^2 - n_2{}^2} + 2} \right) \quad (62)$$

where, k_0 is the free-space wave number at 1550nm, $n_1 = n_{Si} = 3.5$ is MR core's refractive index, $n_2 = n_{SiO2} = 1.5$ is MR cladding's refractive index, $L = 250$nm is the MR thickness, and $\theta = 26.51$ is the propagation angle for the fundamental mode in the MR. $\alpha_{scatter}$ corresponding to an increase in the MR core's refractive index (Δn_{Si}) can be evaluated from Eq. (62) by putting $n_1 = n_{Si} + \Delta n_{Si}$ in it.

The Q-factor of an MR with resonance wavelength (λ_r) depends on this scattering loss coefficient. The relation between the Q-factor and $\Delta\alpha_{scatter}$, assuming critical coupling of MRs, is given by the following equation [139], where Q_A is the loaded Q-factor of the aged MR:

$$Q_A = Q + \Delta Q = \frac{\pi n_g}{\lambda_r(\alpha + \Delta\alpha_{\text{scatter}})}, \quad (63)$$

where, ΔQ is the change in Q-factor and α is the original loss coefficient, which is the sum of three components: *(i)* intrinsic loss coefficient due to material loss and sidewall roughness induced scattering loss; *(ii)* bending loss coefficient, which is a result of the curvature in the MR; and *(iii)* the absorption effect factor that depends on the original free carrier concentration in the waveguide core. As explained above, aging increases the scattering loss coefficient (positive $\Delta\alpha_{scatter}$). As evident from Eq. (63), a positive value of $\Delta\alpha_{scatter}$ results in a decrease in Q-factor. This causes a broadening of the MR passband, which results in increased insertion loss.

For our VBTI aging analysis with MRs, we have considered initial original Q-factor of 9000 and loss coefficient α of 9.5cm^{-1}. As mentioned earlier, α is the sum of the scattering loss coefficient $\alpha_{scatter}$, bending loss coefficient α_b, and absorption loss coefficient α_a, the initial values of which, in this case (for α=9.5cm^{-1}), are 3.5cm^{-1}, 3cm^{-1}, and 3cm^{-1} respectively. Note that $\alpha_{scatter}$=3.5cm^{-1} corresponds to σ=5nm in Eq. (62).

8.4. TEMPERATURE INDUCED MR AGING

Aging in MRs is also dependent on the operating temperature (T) of the devices. As temperature alters activation energy for the Si–H bond breaking and bond annealing, it alters the bond-breaking rate (k_F) and bond-annealing rate (k_R) of the reaction shown in Eq. (54). We use the Arrhenius equation [176] to determine variation in activation energies with temperature. Eq. (64) and Eq. (65) present the temperature dependence of k_F and k_R respectively:

$$k_F = k_{F0} e^{\frac{-E_F}{K_B T}} \tag{64}$$

$$k_R = k_{R0} e^{\frac{-E_R}{K_B T}} \tag{65}$$

where, E_F and E_R are activation energies of forward dissociation and reverse annealing respectively, and K_B is the Boltzmann constant. The activation energy (E_D) of diffusion of hydrogen into the cladding of MRs also depends on temperature, which in turn alters the diffusion constant of hydrogen (D_H) as per the following equation:

$$D_H = D_0 e^{\frac{-E_D}{K_B T}} \tag{66}$$

Figure 62 shows the variation of resonance wavelength red shift ($\Delta\lambda_{RWRS}$) and Q_A with aging in MRs at different temperatures. We analyze $\Delta\lambda_{RWRS}$ and Q_A across different operating

temperatures: 300K, 350K, and 400K. From the figure it can be observed that at a particular temperature, with the increase in MR aging (i.e., increase in usage time) $\Delta\lambda_{RWRS}$ increases and Q_A decreases. With MR aging, the traps on the Si-SiO_2 boundary increase, which is evident from Eq. (59). Furthermore, change in temperature also alters k_F, k_R, and D_H as per Eq. (64), (65), and (66), respectively. These rate constants ultimately change the number of traps generated at the Si-SiO_2 boundary as per Eq. (59). An increase in number of traps incurs an increase in refractive index of an MR (see Eq. (60)), which in turn increases the MR's $\Delta\lambda_{RWRS}$ (see Eq. (61)) and scattering loss ($\alpha_{scatter}$) (see Eq. (62)). Increase in $\alpha_{scatter}$ decreases an MR's Q_A as per Eq. (63). From the figure we can also observe a higher increase in $\Delta\lambda_{RWRS}$ and higher decrease in Q_A with an increase in MR's operating temperature. As the temperature increases, the activation energy (E_D) of diffusion of hydrogen (see Eq. (66)) in the cladding of an MR decreases, which increases the diffusion rate of hydrogen and further increases trap generation at the MR core-cladding boundary. This increase in number of traps ultimately leads to higher increase in RMRS and higher decrease in Q_A.

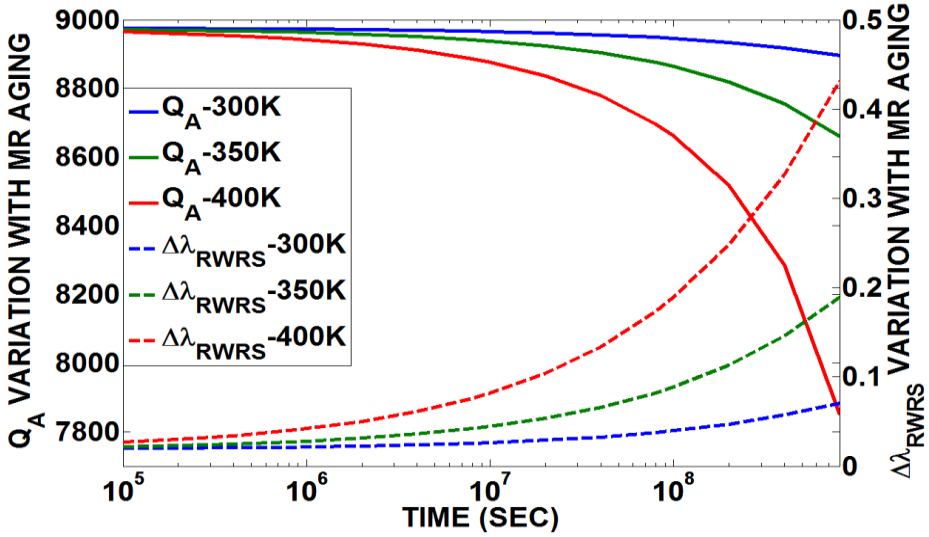


Figure 62: Variation of resonance wavelength red shift ($\Delta\lambda_{RWRS}$) and Q_A with operation time at three operating temperatures 300K, 350K, and 400K.

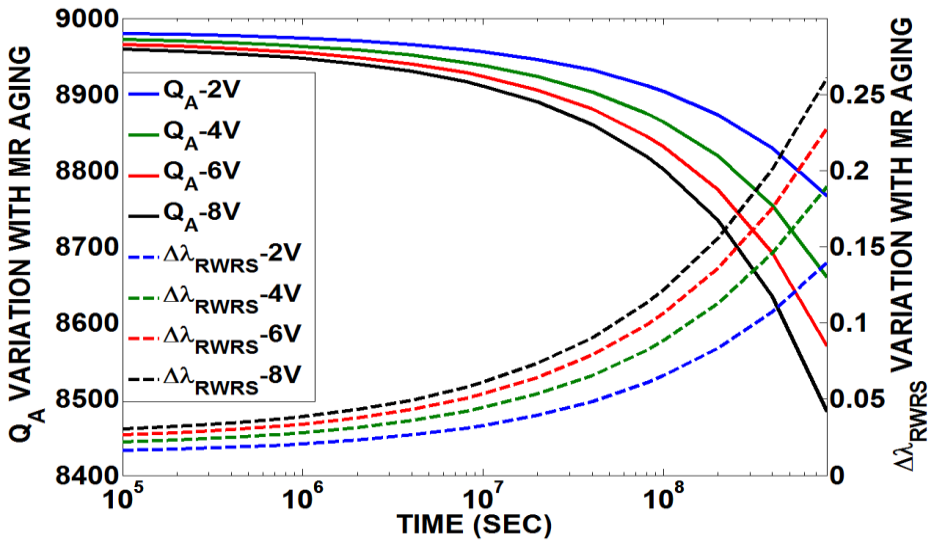


Figure 63: Variation of Q_A and resonance wavelength red shift ($\Delta\lambda_{RWRS}$) with operation time at four bias voltages -2V, -4V, -6V, and -8V.

8.5. IMPACT OF PROCESS VARIATIONS ON MR AGING

Variations in an MR's width and thickness due to process variations (PV) cause a "shift" in the resonance wavelength of the MR. As discussed earlier, voltage biasing (aka localized trimming) is essential to deal with PV-induced resonance shifts in MRs. There are other techniques such as thermal tuning that used to compensate PV-induced resonance drifts. However, thermal tuning has higher power overhead (240 μW/nm) to compensate 1nm PV-induced drift compared to localized trimming (130 μW/nm) [22]. Therefore, voltage biasing or trimming is preferred to compensate PV-induced resonance drifts over thermal tuning. Voltage biasing incurs blue shift/red shift in an MR's resonance wavelength via carrier injection/removal. To enable localized trimming in MRs to counteract PV-induced blue shifts, the negative bias voltage needs to be increased across the MR's reverse-biased PN junction. Unfortunately, this PV-induced increase in negative bias voltage results in an increase in the electric field across the MR core-cladding boundary and this electric field aggravates MR aging.

The forward dissociation constant (k_F) in Eq. (55) will depend on the electric field across core-cladding boundary (E_{OX}). Thus the equation for k_F shown in Eq. (64) is updated as per the following equation [176]:

$$k_F = B\sigma_0 \, E_{OX} \, e^{\frac{E_{OX}}{E_0}} e^{\frac{-E_F}{K_B T}} \tag{67}$$

where $\exp(E_{OX}/E_0)$ is the field dependent tunneling of holes into SiO_2 cladding, σ_0 is the capture cross-section of the Si–H bonds, and B determines field dependence of the Si–H bond dissociation.

Table 12: Notations for photonic power loss and model parameters [61].

Notation	Parameter type	Parameter value (in dB)
L_P	Propagation loss	-0.274 per cm
L_B	Bending loss	-0.005 per 90°
L_{S12}	1X2 splitter power loss	-0.2
L_{S14}	1X4 splitter power loss	-0.2
L_{S17}	1X7 splitter power loss	-0.2
L	Photonic path length in cm	
B	Number of bends in photonic path	
λ_j	Resonance wavelength of MR	
R_{S12}	Splitting factor for 1X2 splitter	
R_{S14}	Splitting factor for 1X4 splitter	
R_{S17}	Splitting factor for 1X7 splitter	

Figure 63 illustrates the impact of variation in bias voltage on $\Delta\lambda_{RWRS}$ and Q_A of MR with aging (i.e., usage time). We analyze negative voltage biases of 2V, 4V, 6V, and 8V, and the MR is assumed to be operated at 350K temperature. As explained Section 80, the charge transport solver in the DEVICE tool is used to determine electric field (E_{OX}) across the core-cladding boundary for each bias voltage across the PN junction of the MR. This tool uses MR device dimensions such as width, height and radius to determine E_{OX} at the boundary. From the figure it can be observed that with the increase in negative bias voltage, MRs incur higher $\Delta\lambda_{RWRS}$ increase (see Eq. (61)) and higher Q_A decrease (See Eq. (62)). As the negative bias voltage across the PN

junction of the MR increases, the E_{OX} across the core-cladding boundary of the MR increases. This increase in E_{OX} increases k_F as per Eq. (67), which in turn increases trap generation across the core-cladding boundary as per Eq. (59). This increase in trap generation increases $\Delta\lambda_{RWRS}$ and Q_A of an MR, as also highlighted by the Eq. (60)-(63) presented in Section 0.

8.6. IMPACT OF MR VBTI AGING ON PNOCS

8.6.1. MR AGING ANALYSIS FOR CORONA AND CLOS PNOCS

We characterize the impact of VBTI aging on two popular PNoC architectures: Corona [59] and Clos [102], both of which use DWDM-waveguides for data communication. We have considered Corona PNoC with token-slot arbitration [59] and an 8-ary 3-stage Clos PNoC [102] for our analysis. In DWDM-based waveguides, data transmission requires modulating light using a group of MR modulators equal to the number of wavelengths supported by DWDM. Similarly, data detection at the receiver requires a group of detector MRs equal to the number of DWDM wavelengths. We present analytical equations to model the impact of aging on maximum signal power loss in each architecture. Before presenting relevant equations, we provide notations for the parameters used in the equations, in Table 12.

We first model the MR transmission spectrum at a device-level and then extend these models to the system-level to determine the impact of aging on signal losses for PNoC architectures. We model the MR transmission spectrum using a Lorentzian function [66]. In Eq. (68), this function is used to represent coupling factor ϕ between wavelength λ_i and an MR with resonance wavelength λ_j. Further, using the same function, we determined loss factor γ in Eq. (69), which is the factor by which signal power of a wavelength λ_i is reduced when it passes through an MR whose resonance wavelength is λ_j. Through loss of a wavelength in a waveguide, when it passes through an MR, is defined as γ times the signal power of the wavelength before it passes through

the MR. From Eq. (62) and (63), it can be inferred that an MR's loaded Q-factor (Q_A) decreases with aging in MRs. This in turn decreases Φ and increases γ as per Eq. (68) and (69), respectively. Furthermore, as per Eq. (68) and (69) increase in $\Delta\lambda_{RWRS}$ with aging (i.e., $\Delta\lambda_{RWRSAi}$) further decreases Φ and increases γ, respectively.

$$\Phi\left(\lambda_i, \Delta\lambda_{RWRSAi}, \lambda_j, Q_A\right) = \left(1 + \left(\frac{2Q_A(\lambda_i + \Delta\lambda_{RWRSAi} - \lambda_j)}{\lambda_j}\right)^2\right)^{-1}, \quad (68)$$

$$\gamma\left(\lambda_i, \Delta\lambda_{RWRSAi}, \lambda_j, Q_A\right) = \left(1 + \left(\frac{2Q_A(\lambda_i + \Delta\lambda_{RWRSAi} - \lambda_j)}{\lambda_j}\right)^{-2}\right)^{-1}, \quad (69)$$

Corona PNoC: This PNoC is designed for a 256 core single-chip platform, where cores are grouped into 64 clusters, with 4 cores in each cluster. A photonic crossbar topology with 64 data channels is used for communication between clusters. Each channel consists of 4 multiple-write-single-read (MWSR) waveguides with 64-wavelength DWDM in each waveguide. As modulation occurs on both positive and negative edges of the clock in Corona, 512 bits (cache-line size) can be modulated and inserted on 4 MWSR waveguides in a single cycle by a sender. A data channel starts at a cluster called 'home-cluster', traverses other clusters (where modulators can modulate light and detectors can detect this light), and finally ends at the home-cluster again, at a set of detectors (optical termination). A power waveguide supplies optical power from an off-chip laser to each of the 64 data channels at its home-cluster, through a series of 1X2 splitters. In each of the 64 home-clusters, optical power is distributed among 4 MWSR waveguides equally using a 1X4 splitter with splitting factor R_{S14}. As all 1X2 splitters are present before the last (64^{th}) channel, this channel suffers the highest signal power loss. Thus, the worst-case signal loss exists in the detector group of the 64^{th} cluster node, and this node is defined as the worst-case power loss node (N_{WCPL}) in the Corona PNoC. For this N_{WCPL} node, signal power ($P_{signal}(\lambda_j)$) on each detector with resonance wavelength λ_j is shown in Eq. (70). $K(\lambda_i)$ in Eq. (72) represents signal power loss of λ_i before the

detector group of N_{WCPL} (see Table 12 for notations of different parameters). $\psi(\lambda_i, \lambda_j)$ in Eq. (71) represents signal power loss of λ_i before the detector with resonance wavelength λ_j in the detector group of N_{WCPL}.

$$P_{signal}(\lambda_j) = K(\lambda_i)\psi(\lambda_i, \lambda_j)\, \Phi(\lambda_j, \Delta\lambda_{RWRSAj}, \lambda_j, Q_{A(63\times64)+j})\, P_{in}(i), \tag{70}$$

$$\psi(\lambda_i, \lambda_j) = \prod_{k=1}^{(k-1)<j} \gamma(\lambda_i, \Delta\lambda_{RWRSAi}, \lambda_k, Q_{A(63\times64)+k}), \tag{71}$$

$$K(\lambda_i) = (R_{S14})(L_{S14})(L_P)^L(L_B)^B \prod_{n=1}^{63}\prod_{j=1}^{64} \gamma\left(\lambda_i, \Delta\lambda_{RWRSAi}, \lambda_j, Q_{A((n-1)\times64)+j}\right) \tag{72}$$

Clos PNoC: An 8-ary 3-stage Clos topology is considered for a 256-core system, with 8 clusters (C1-C8) and 32 cores in each cluster. Within each cluster, a group of four cores are connected to a concentrator. Thus each cluster has 8 concentrators and the concentrators are connected electrically through a router for inter-concentrator communication. The Clos PNoC uses photonic signals for inter-cluster communication. Unlike the MWSR waveguides used in the Corona crossbar, the Clos uses point-to-point photonic links for data communication. Each point-to-point photonic link uses either forward or backward propagating wavelengths depending on the physical location of the source and destination clusters. Each photonic link in the Clos PNoC use 128 DWDM, with 64 wavelengths for forward communication and the remaining 64 wavelengths for backward communication. Thus the Clos PNoC uses only 56 waveguides with 256 MRs on each waveguide. This PNoC uses 2 laser sources to enable forward and backward communication. To power the 56 waveguides, it is assumed that the PNoC employs a series of 1X2, 1X7, and 1X4 splitters. In our implementation of the Clos PNoC, the worst-case power loss occurs when C1 sends data to C8, as this involves the longest photonic path for data traversal. Thus the node C8 is the worst-case power loss node (N_{WCPL}) in the Clos PNoC. We use Eq. (70) to determine worst-case power loss in the Clos PNoC. But as the Clos network has lower number of waveguides and

fewer number of MRs on each waveguide, this in turn changes the signal power losses. Thus we modify Eq. (72) for the Clos PNoC as:

$$K(\lambda_i) = (\mathrm{R}_{S14})(L_{S14})(\mathrm{L}_P)^L(\mathrm{L}_B)^B \prod_{n=1}^{3}\prod_{j=1}^{64} \gamma\left(\lambda_i, \Delta\lambda_{RWRSAi}, \lambda_j, \mathrm{Q}_{A((n-1)\times 64)+j}\right) \quad (73)$$

8.6.2. MODELING PV OF MR DEVICES IN CORONA AND CLOS PNOCS

We adapt the VARIUS tool [71] to model die-to-die (D2D) as well as within-die (WID) process variations in MRs for the Corona and Clos PNoCs. VARIUS uses a normal distribution to characterize on-chip D2D and WID process variations. The key parameters are mean (μ), variance (σ^2), and density (α) of a variable that follows the normal distribution. As wavelength variations are approximately linear to dimension variations of MRs, we assume they follow the same distribution. The mean (μ) of wavelength variation of an MR is its nominal resonance wavelength. We consider a DWDM wavelength range in the C and L bands [72], with a starting wavelength of 1550nm and a channel spacing of 0.8nm. Hence, those wavelengths are the means for each MR modeled. The variance (σ^2) of wavelength variation is determined based on laboratory fabrication data [33] and our target die size. We consider a 256-core chip with die size 400 mm^2 at a 22nm process node. For this die size we consider a WID standard deviation (σ_{WID}) of 0.61nm [63] and D2D standard deviation (σ_{D2D}) of 1.01 nm [63]. We also consider a density (α) of 0.5 [63] for this die size. With these parameters, we use VARIUS to generate 100 PV maps, each containing over 1 million points indicating the PV-induced resonance shift of MRs. The total number of points picked from these maps equal the number of MRs in the Corona and Clos PNoCs.

8.7. EXPERIMENTS

8.7.1. EXPERIMENT SETUP

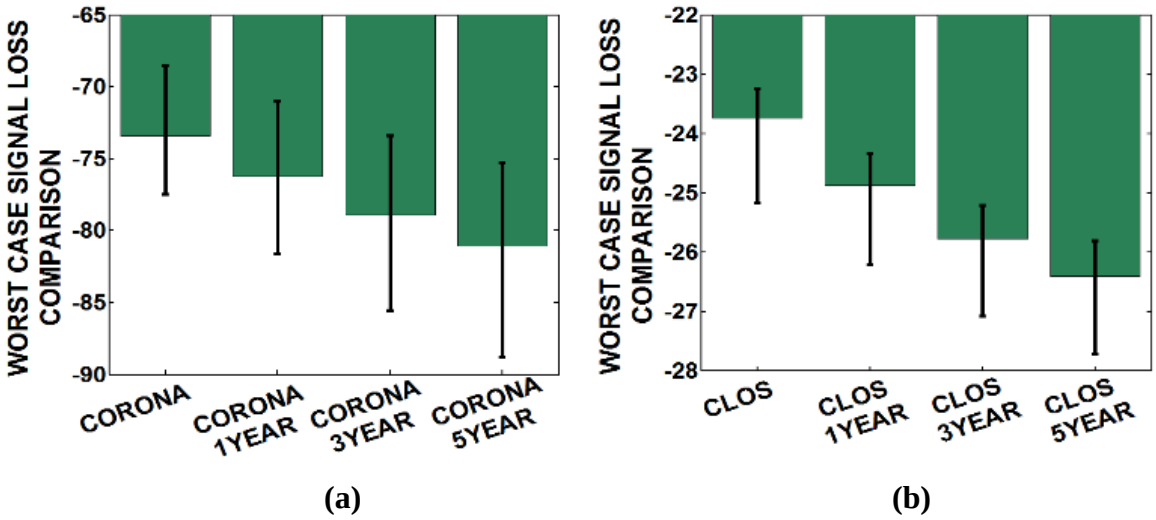
We evaluate the impact of VBTI aging on PNoCs on the Corona and Clos PNoC architectures. We modeled and performed simulation based analysis of the Corona and Clos PNoCs using a cycle-accurate NoC simulator, for a 256 core single-chip architecture at 22nm. As explained in Section 0, we generated 100 PV maps to evaluate MR aging impact on these PNoCs for different PV profiles. We used real-world traffic from applications in the PARSEC benchmark suite [76]. GEM5 full-system simulation [77] of parallelized PARSEC applications was used to generate traces that were fed into our cycle-accurate NoC simulator. We set a "warm-up" period of 100 million instructions and then captured traces for the subsequent 1 billion instructions. We performed geometric calculations for a 20mm×20mm chip size, to determine lengths of MWSR waveguides in the Corona PNoC and photonic links in the Clos PNoC. We consider a 5 GHz clock frequency of operation for the cores. A 512-bit packet size is utilized for both Corona and Clos PNoCs.

The static and dynamic energy consumption of electrical routers and concentrators in the Corona and Clos PNoCs is based on results from the open source DSENT tool [75]. For energy consumption of photonic devices, we adapt model parameters from recent work [79], [80] with 0.42pJ/bit for every modulation and detection event and 0.18pJ/bit for the driver circuits of modulators and photodetectors. We used optical loss in photonic components (Table 12) to estimate photonic laser power budget and correspondingly the electrical laser power Chapter 6.

8.7.2. EXPERIMENT RESULTS

Our first set of experiments compares the worst-case signal losses of the baseline Corona and Clos PNoCs with their variants with 1 Year, 3 Years, and 5 Years of VBTI aging. We have

performed this aging analysis across 100 PV maps as explained in Section 0. The presented results are averaged across the PV maps. Furthermore, as we are determining worst-case signal loss for Corona and Clos PNoCs with VBTI aging, therefore we performed this analysis at the peak on-chip temperature, which is estimated to be 357 K [181].

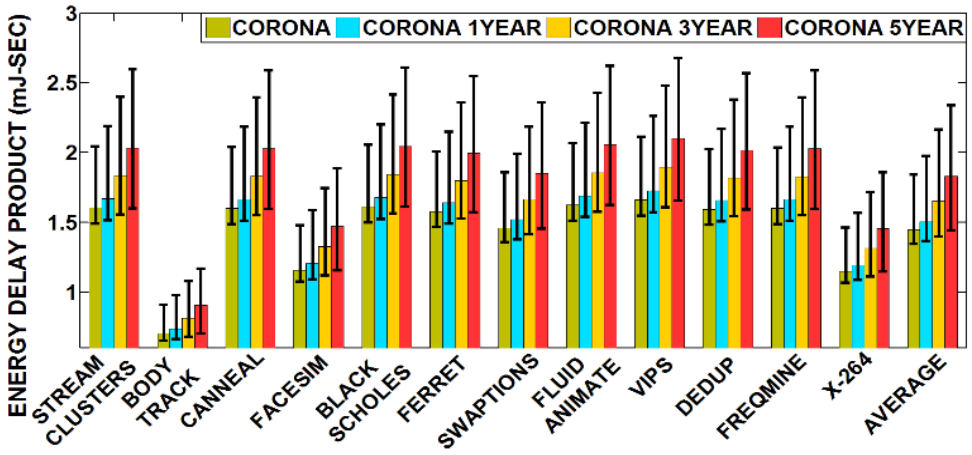


(a) **(b)**

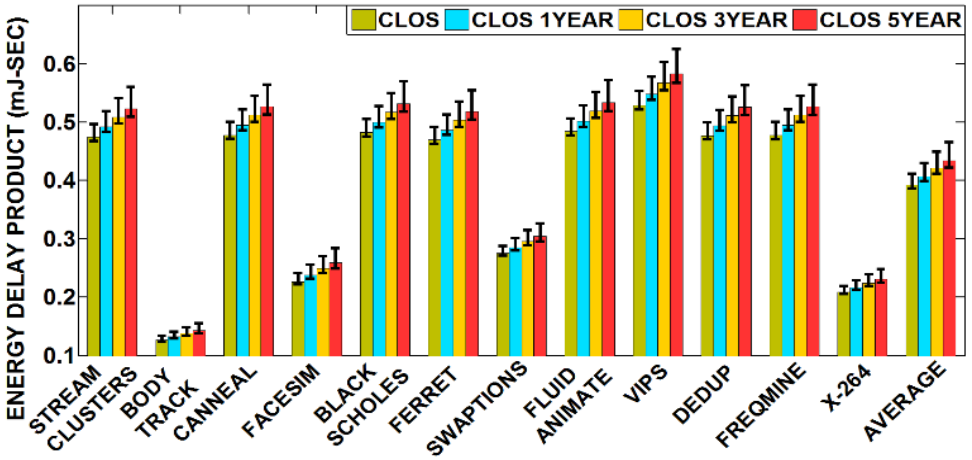
Figure 64: Worst-case signal power loss analysis of (a) Corona PNoC, (b) Clos PNoC, with 1 Year, 3 Years, and 5 Years of aging across 100 PV maps.

Utilizing the models presented in Section 0, we calculate the signal power loss at the last detector of the N_{WCPL} nodes of Corona and Clos PNoCs, which corresponds to the last MR detector of the cluster 64 and cluster 8 for the Corona and Clos PNoCs respectively. Figure 65(a) and (b) compare the worst-case signal loss of baseline Corona and Clos PNoCs with three variants of these PNoCs that undergo 1 Year, 3 Years and 5 Years of VBTI aging. The confidence intervals represent the variation in signal loss across the 100 PV maps considered. From Figure 65(a), it can be observed that compared to their respective baselines, the Corona PNoC with 1 Year, 3 Year, and 5 years of VBTI aging has 2.8dB, 5.5dB, and 7.6dB higher signal losses, and the Clos PNoC has 1.1 dB, 2.1dB, and 2.6dB higher signal losses. The increase in resonance wavelength red shift ($\Delta\lambda_{RWRS}$) and degradation in Q-factor with VBTI aging in MRs leads to increase in MR loss factor

(γ) (see Eq. (69)) and decrease in MR coupling factor (Φ) (see Eq. (68)), which ultimately increases signal losses in these PNoCs. Also, the increase in signal loss in the Corona PNoC with VBTI aging is on the higher side compared to the Clos PNoC. Corona has 16× higher number of MRs on its waveguides compared to the Clos PNoC, which in turn incurs higher signal losses on Corona's waveguides.



(a)



(b)

Figure 65: EDP comparison of (a) Corona, (b) Clos PNoCs with 1 Year, 3 Years, and 5 Years of aging considering 100 process variation maps.

Figure 65(a) and (b) present detailed simulation results that quantify the energy-delay product (EDP) for the four configurations of Corona and Clos PNoCs respectively. Results are

shown for twelve multi-threaded PARSEC benchmarks. From Figure 65(a) it can be seen that on average, Corona PNoC with 1 Year, 3 Year, and 5 years of VBTI aging has 4.1%, 14.3%, and 26.8% and Clos PNoC has 3.7%, 7.5%, and 10.6% higher EDP compare to their respective baselines. Increase in worst-case signal loss with increase in VBTI aging (see Figure 64) contributes to an increase in the PNoCs laser power, which increases total laser energy consumption in these PNoCs. Additionally, VBTI aging in MRs has positive effects on MR trimming energy consumption, as MR aging incurs red shift in resonance wavelength which naturally reduces PV-induced blue shifts in MRs and reduces total trimming energy consumption in the PNoCs. However, these trimming energy savings are relatively on the lower side compared to the increase in laser energy, which ultimately increase total energy and hence the EDP.

From the results presented in this section, we can summarize that in Corona and Clos PNoCs, VBTI aging in MRs increases signal losses by up to 7.6dB. Despite the decrease in tuning energy consumption of the Corona and Clos PNoCs with VBTI aging, the increase in their laser energy consumption increases EDP in these architectures by up to 26.8%. The signal loss and EDP increase due to VBTI aging are much lower in architectures optimized for physical-layouts such as the Clos PNoC, than in non-optimized architectures such as Corona. PNoC architectures with more MRs per waveguide (e.g., Corona) have higher VBTI aging degradation compared to PNoC architectures with less MRs per waveguide (e.g., Clos). Thus, to reduce aging effects in a PNoC, designers should reduce the number of MRs per waveguide and increase the number of these waveguides to maintain high bandwidth.

8.8. CONCLUSIONS

In this chapter, we analyzed VBTI aging in MRs used in photonic interconnects, and the dependence of this aging on voltage bias and temperature. We presented an analytical model for

trap generation on the MR core-cladding boundary with VBTI aging in MRs. We also considered the impact of process variations on aging. Our device-level results indicated that MR aging causes significant degradation in MR Q-factor and incurs notable resonance wavelength red shift. We extended our MR aging analysis to the system-level for the Corona and Clos PNoCs. The system-level analysis on these PNoCs clearly showed the damaging effects of MR aging, with worst signal loss increase by up to 7.6dB and EDP increase by up to 26.8%.

9. SOTERIA: EXPLOITING PROCESS VARIATIONS TO ENHANCE HARDWARE SECURITY WITH PHOTONIC NOC ARCHITECTURES

A Hardware Trojan in a PNoC can manipulate the electrical driving circuit of its MRs to cause the MRs to snoop data from the neighboring wavelength channels in a shared photonic waveguide. This introduces a serious security threat. This chapter presents a novel framework called *SOTERIA* that utilizes process variation based authentication signatures along with architecture-level enhancements to protect data in PNoC architectures from snooping attacks. Evaluation results indicate that our approach can significantly enhance the hardware security in DWDM-based PNoCs with minimal overheads of up to 10.6% in average latency and of up to 13.3% in energy-delay-product (EDP).

9.1. INTRODUCTION

To cope with the growing performance demands of modern Big Data and cloud computing applications, the complexity of hardware in modern chip-multiprocessors (CMPs) has increased. To reduce the hardware design time of these complex CMPs, third-party hardware IPs are frequently used. But these third party IPs can introduce security risks [182], [183]. For instance, the presence of Hardware Trojans (HTs) in the third-party IPs can lead to leakage of critical and sensitive information from modern CMPs [184]. Thus, security researchers that have traditionally focused on software-level security are now increasingly interested in overcoming hardware-level security risks.

Many CMPs today use electrical networks-on-chip (ENoCs) for inter-core communication. ENoCs use packet-switched network fabrics and routers to transfer data between on-chip components [185]. Recent developments in silicon photonics have enabled the integration of

photonic components and interconnects with CMOS circuits on a chip. Photonic NoCs (PNoCs) provide several prolific advantages over their metallic counterparts (i.e., ENoCs), including the ability to communicate at near light speed, larger bandwidth density, and lower dynamic power dissipation [9]. These advantages motivate the use of PNoCs for inter-core communication in modern CMPs [32].

Several PNoC architectures have been proposed to date (e.g., [15] and [16]). These architectures employ on-chip photonic links, each of which connects two or more gateway interfaces. A gateway interface (GI) connects the PNoC to a cluster of processing cores. Each photonic link comprises one or more photonic waveguides and each waveguide can support a large number of dense-wavelength-division-multiplexed (DWDM) wavelengths. A wavelength serves as a data signal carrier. Typically, multiple data signals are generated at a source GI in the electrical domain (as sequences of logical 1 and 0 voltage levels) which are modulated onto the multiple DWDM carrier wavelengths simultaneously, using a bank of modulator MRs at the source GI [99]. The data-modulated carrier wavelengths traverse a link to a destination GI, where an array of detector MRs filter them and drop them on photodetectors to regenerate electrical data signals.

In general, each GI in a PNoC is able to send and receive data in the optical domain on all of the utilized carrier wavelengths. Therefore, each GI has a bank of modulator MRs (i.e., modulator bank) and a bank of detector MRs (i.e., detector bank). Each MR in a bank resonates with and operates on a specific carrier wavelength. Thus, the excellent wavelength selectivity of MRs and DWDM capability of waveguides enable high bandwidth parallel data transfers in PNoCs.

Similar to CMPs with ENoCs, the CMPs with PNoCs are expected to use several third party IPs, and therefore, are vulnerable to security risks [186]. For instance, if the entire PNoC used within a CMP is a third-party IP, then this PNoC with HTs within the control units of its GIs can

snoop on packets in the network. These packets can be transferred to a malicious core (a core running a malicious program) in the CMP to determine sensitive information.

Unfortunately, MRs are especially susceptible to security threatening manipulations from HTs. In particular, *the MR tuning circuits that are essential for supporting data broadcasts and to counteract MR resonance shifts due to process variations (PV) make it easy for HTs to retune MRs and initiate snooping attacks.* To enable data broadcast in PNoCs, the tuning circuits of detector MRs partially detune them from their resonance wavelengths [15], [187], such that a significant portion of the photonic signal energy in the data-carrying wavelengths continues to propagate in the waveguide to be absorbed in the subsequent detector MRs. On the other hand, process variations (PV) cause resonance wavelength shifts in MRs [33]. Techniques to counteract PV-induced resonance shifts in MRs involve retuning the resonance wavelengths by using carrier injection/depletion or thermal tuning [32], implemented through MR tuning circuits. An HT in the GI can manipulate these tuning circuits of detector MRs to partially tune the detector MR to a passing wavelength in the waveguide, which enables snooping of the data that is modulated on the passing wavelength. *Such covert data snooping is a serious security risk in PNoCs.*

In this work, we present a framework that protects data from snooping attacks and improves hardware security in PNoCs. Our framework has low overhead and is easily implementable in any existing DWDM-based PNoC without major changes to the architecture. To the best of our knowledge, this is the first work that attempts to improve hardware security for PNoCs. Our novel contributions are:

- We analyze security risks in photonic devices and extend this analysis to link-level, to determine the impact of these risks on PNoCs;

- We propose a circuit-level PV-based security enhancement scheme that uses PV-based authentication signatures to protect data from snooping attacks in photonic waveguides;
- We propose an architecture-level reservation-assisted security enhancement scheme to improve security in DWDM-based PNoCs;
- We combine the circuit- and architecture-level schemes into a holistic framework called *SOTERIA*; and analyze it on the Firefly [15] and Flexishare [16] crossbar-based PNoC architectures.

9.2. RELATED WORK

Several prior works (e.g., [186], [188], and [189]) discuss the presence of security threats in ENoCs and have proposed solutions to mitigate them. In [186], a three-layer security system approach was presented by using data scrambling, packet certification, and node obfuscation to enable protection against data snooping attacks. A symmetric-key based cryptography design was presented in [188] for securing the NoC. In [189], a framework was presented to use permanent keys and temporary session keys for NoC transfers between secure and non-secure cores. *However, no prior work has analyzed security risks in photonic devices and links; or considered the impact of these risks on PNoCs.*

Fabrication-induced PV impact the cross-section, i.e., width and height, of photonic devices, such as MRs and waveguides. In MRs, PV causes resonance wavelength drifts, which can be counteracted by using device-level techniques such as thermal tuning or localized trimming [32]. Trimming can induce blue shifts in the resonance wavelengths of MRs using carrier injection into MRs, whereas thermal tuning can induce red shifts in MR resonances through heating of MRs using integrated heaters. *To remedy PV, the use of device-level trimming/tuning techniques is inevitable; but their use also enables partial detuning of MRs that can be used to snoop data from*

a shared photonic waveguide. In addition, prior works [56] discuss the impact of PV-remedial techniques on crosstalk noise and proposed techniques to mitigate it. *None of the prior works analyze the impact of PV-remedial techniques on hardware security in PNoCs.*

Our proposed framework in this chapter is novel as it enables security against snooping attacks in PNoCs for the first time. Our framework is network agnostic, mitigates PV, and has minimal overhead, while improving security for any DWDM-based PNoC architecture.

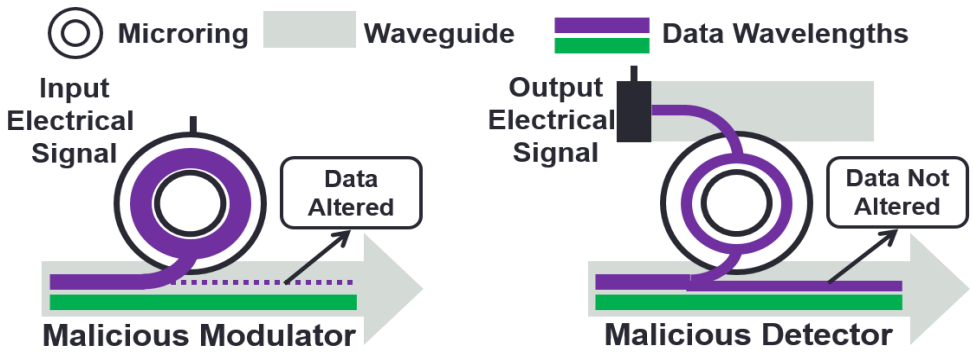


Figure 66: Impact of (a) malicious modulator MR, (b) malicious detector MR on data in DWDM-based photonic waveguides.

9.3. HARDWARE SECURITY CONCERNS IN PNOCS

9.3.1. DEVICE-LEVEL SECURITY CONCERNS

Process variation (PV) induced undesirable changes in MR widths and heights cause "shifts" in MR resonance wavelengths, which can be remedied using localized trimming and thermal tuning methods. The localized trimming method injects (or depletes) free carriers into (or from) the Si core of an MR using an electrical tuning circuit, which reduces (or increases) the MR's refractive index owing to the electro-optic effect, thereby remedying the PV-induced red (or blue) shift in the MR's resonance wavelength. In contrast, thermal tuning employs an integrated micro-heater to adjust the temperature and refractive index of an MR (owing to the thermo-optic effect)

for PV remedy. Typically, the modulator MRs and detectors use the same electro-optic effect (i.e., carrier injection/depletion) implemented through the same electrical tuning circuit as used for localized trimming, to move in and out of resonance (i.e., switch ON/OFF) with a wavelength [56]. *A Hardware Trojan can manipulate this electrical tuning circuit, which may lead to malicious operation of modulator and detector MRs, as discussed next.*

Figure 66(a) shows the malicious operation of a modulator MR. A malicious modulator MR is partially tuned to a data-carrying wavelength (shown in purple) that is passing by in the waveguide. The malicious modulator MR draws some power from the data-carrying wavelength, which can ultimately lead to data corruption as optical '1's in the data can lose significant power to be altered into '0's. Alternatively, a malicious detector (Figure 66(b)) can be *partially* tuned to a passing data-carrying wavelength, to filter only a small amount of its power and drop it on a photodetector for data duplication. This small amount of filtered power does not alter the data in the waveguide so that it continues to travel to its target detector for legitimate communication [187]. Thus, malicious detector MRs can snoop data from the waveguide without altering it, which is a major security threat in photonic links. Note that malicious modulator MRs only corrupt data (which can be detected) and do not covertly duplicate it, and are thus not a major security risk. Our analysis in Section 0 presents the impact of malicious modulator and detector MRs on photonic links.

9.3.2. LINK-LEVEL SECURITY CONCERNS

Typically, a photonic link is comprised of one or more DWDM-based photonic waveguides. A DWDM-based photonic waveguide uses a modulator bank (a series of modulator MRs) at the source GI and a detector bank (a series of detector MRs) at the destination GI. DWDM-based waveguides can be broadly classified into four types: single-writer-single-reader (SWSR), single-

writer-multiple-reader (SWMR), multiple-writer-single-reader (MWSR), and multiple-writer-multiple-reader (MWMR). As SWSR, SWMR, and MWSR waveguides are subsets of an MWMR waveguide, and due to limited space, we restrict our link-level analysis to MWMR waveguides only.

An MWMR waveguide typically passes through multiple GIs, connecting the modulator banks of some GIs to the detector banks of the remaining GIs. Thus, in an MWMR waveguide, multiple GIs (referred to as source GIs) can send data using their modulator banks and multiple GIs (referred to as destination GIs) can receive (read) data using their detector banks. Figure 67 presents an example MWMR waveguide with two source GIs and two destination GIs. Figure 67(a) and Figure 67(b), respectively, present the impact of malicious source and destination GIs on this MWMR waveguide. In Figure 67(a), the modulator bank of source GI S_1 is sending data to the detector bank of destination GI D_2. When source GI S_2, which is in the communication path, becomes malicious with an HT in its control logic, it can manipulate its modular bank to modify the existing '1's in the data to '0's. This ultimately leads to data corruption. For example, in Figure 67(a), S_1 is supposed to send '0110' to D_2, but because of data corruption by malicious GI S_2, '0010' is received by D_2. Nevertheless, this type of data corruption can be detected or even corrected using parity or error correction code (ECC) bits in the data. Thus, malicious source GIs do not cause major security risks in DWDM-based MWMR waveguides.

Let us consider another scenario for the same data communication path (i.e., from S_1 to D_2). When destination GI D_1, which is in the communication path, becomes malicious with an HT in its control logic, the detector bank of D_1 can be partially tuned to the utilized wavelength channels to snoop data. In the example shown in Figure 67(b), D_1 snoops '0110' from the wavelength channels that are destined to D_2. The snooped data from D_1 can be transferred to a malicious core

within the CMP to determine sensitive information. This type of snooping attack from malicious destination GIs is hard to detect, as it does not disrupt the intended communication among CMP cores. Therefore, there is a pressing need to address the security risks imposed by snooping GIs in DWDM-based PNoC architectures. To address this need, we propose a novel framework *SOTERIA* that improves hardware security in DWDM-based PNoC architectures.

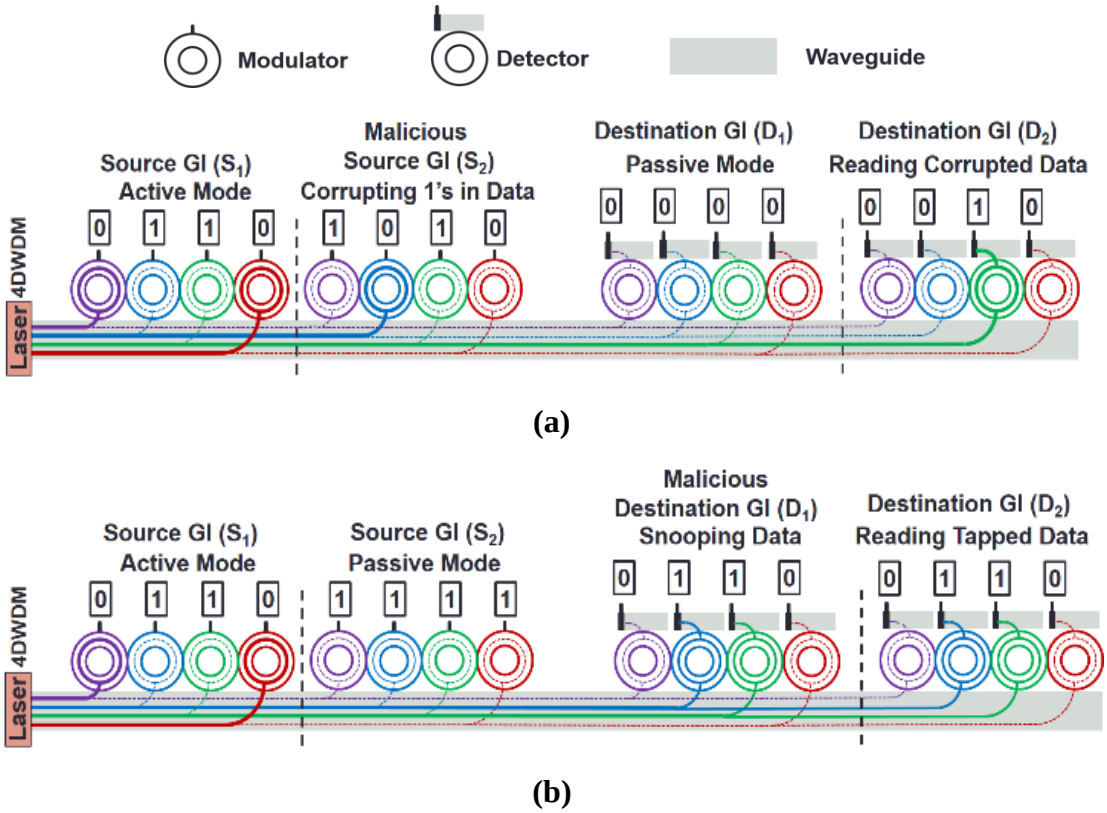


Figure 67: Impact of (a) malicious modulator (source) bank, (b) malicious detector bank on data in DWDM-based photonic waveguides.

9.4. SOTERIA FRAMEWORK: OVERVIEW

Our proposed multi-layer *SOTERIA* framework enables secure communication in DWDM-based PNoC architectures by integrating circuit-level and architecture-level enhancements. Figure 68 gives a high-level overview of this framework. The PV-based security enhancement (*PVSC*) scheme uses the PV profile of the destination GIs' detector MRs to encrypt data before it is

transmitted via the photonic waveguide. This scheme is sufficient to protect data from snooping GIs, if they do not know about the target destination GI. With target destination GI information, however, a snooping GI can decipher the encrypted data. Many PNoC architectures (e.g., [156] and [186]) use the same waveguide to transmit both the destination GI information and actual data, making them vulnerable to data snooping attacks despite using *PVSC*. To further enhance security for these PNoCs, we devise an architecture-level reservation-assisted security enhancement (*RVSC*) scheme that uses a secure reservation waveguide to avoid the stealing of destination GI information by snooping GIs. The next two sections present details of our *PVSC* and *RVSC* schemes.

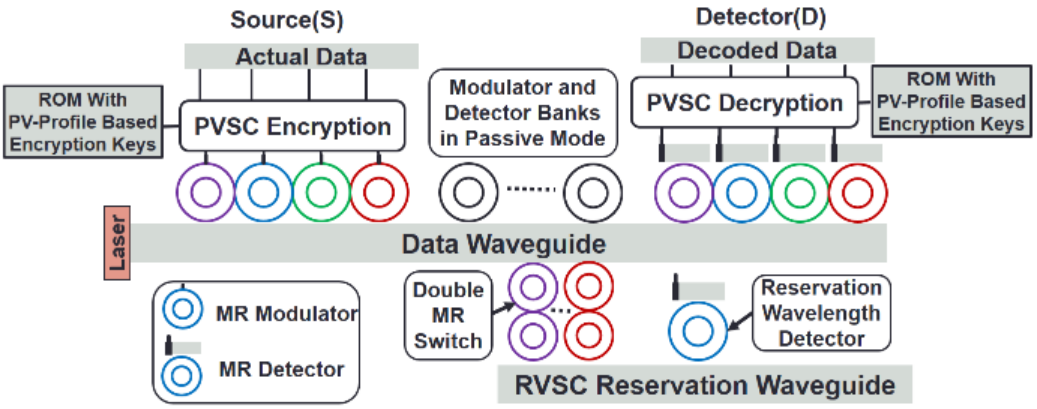


Figure 68: Overview of proposed *SOTERIA* framework that integrates a circuit-level PV-based security enhancement (*PVSC*) scheme and an architecture-level reservation-assisted security enhancement (*RVSC*) scheme.

9.5. PV-BASED SECURITY ENHANCEMENT

As discussed earlier (Section 0), malicious destination GIs can snoop data from a shared waveguide. One way of addressing this security concern is to use data encryption so that the malicious destination GIs cannot decipher the snooped data. For the encrypted data to be truly undecipherable, the encryption key used for data encryption should be kept secret from the snooping GIs, which can be challenging as the identity of the snooping GIs in a PNoC is not known.

Therefore, it becomes very difficult to decide whether or not to share the encryption key with a destination GI (that can be malicious) for data decryption. This conundrum can be resolved using a different key for every destination GI so that a key that is specific to a secure destination GI does not need to be shared with a malicious destination GI for decryption purpose. Moreover, to keep these destination specific keys secure, the malicious GIs in a PNoC must not be able to clone the algorithm (or method) used to generate these keys.

To generate unclonable encryption keys, our PV-based security (*PVSC*) scheme uses the PV profiles of the destination GIs' detector MRs. As discussed in [33], PV induces random shifts in the resonance wavelengths of the MRs used in a PNoC. These resonance shifts can be in the range from -3nm to 3nm [33]. The MRs that belong to different GIs in a PNoC have different PV profiles. In fact, the MRs that belong to different MR banks of the same GI also have different PV profiles. Due to their random nature, these MR PV profiles cannot be cloned by the malicious GIs, which makes the encryption keys generated using these PV profiles truly unclonable. Using the PV profiles of detector MRs, *PVSC* can generate a unique encryption key for each detector bank of every MWMR waveguide in a PNoC.

Our *PVSC* scheme generates encryption keys during the testing phase of the CMP chip, by using a dithering signal based in-situ method [171] to generate an anti-symmetric analog error signal for each detector MR of every detector bank that is proportional to the PV-induced resonance shift in the detector MR. Then, it converts the analog error signal into a 64-bit digital signal. Thus, a 64-bit digital error signal is generated for every detector MR of each detector bank. We consider 64 DWDM wavelengths per waveguide, and hence, we have 64 detector MRs in every detector bank and 64 modulator MRs in every modulator bank. For each detector bank, our *PVSC* scheme XORs the 64 digital error signals (of 64 bits each) from each of the 64 detector MRs

to create a unique 64-bit encryption key. Note that our PVSC scheme also uses the same anti-symmetric error signals to control the carrier injection and heating of the MRs to remedy the PV-induced shifts in their resonances.

To understand how the 64-bit encryption key is utilized to encrypt data in photonic links, consider Figure 69 which depicts an example photonic link that has one MWMR waveguide and connects the modulator banks of two source GIs (S_1 and S_2) with the detector banks of two destination GIs (D_1 and D_2). As there are two destination GIs on this link, *PVSC* creates two 64-bit encryption keys corresponding to them, and stores them at the source GIs. When data is to be transmitted by a source GI, the key for the appropriate destination is used to encrypt data at the flit-level granularity, by performing an XOR between the key and the data flit. This requires that the size of an encryption key match the data flit size. We consider the size of data flits to be 512 bits. Therefore, the 64-bit encryption key is appended eight times to generate a 512-bit encryption key. In Figure 69, every source GI stores two 512-bit encryption keys (for destination GIs D_1 and D_2) in its local ROM, whereas every destination GI stores only its corresponding 512-bit key in its ROM. Note that we store the 512-bit keys instead of the 64-bit keys as this eliminates the latency overhead of affixing 64-bit keys to generate 512-bit keys, at the cost of a reasonable area/energy overhead in the ROM. As an example, if S_1 wants to send a data flit to D_2, then S_1 first accesses the 512-bit encryption key corresponding to D_2 from its local ROM and XORs the data flit with this key in one cycle, and then transmits the encrypted data flit over the link. As the link employs only one waveguide with 64 DWDM wavelengths, therefore, the encrypted 512-bit data flit is transferred on the link to D_2 in eight cycles. At D_2, the data flit is decrypted by XORing it with the 512-bit key corresponding to D_2 from the local ROM. In this scheme, even if D_1 snoops the data intended for D_2, it cannot decipher the data as it does not have access to the correct key

(corresponding to D_2) for decryption. Thus, our *PVSC* encryption scheme protects data against snooping attacks in DWDM-based PNoCs.

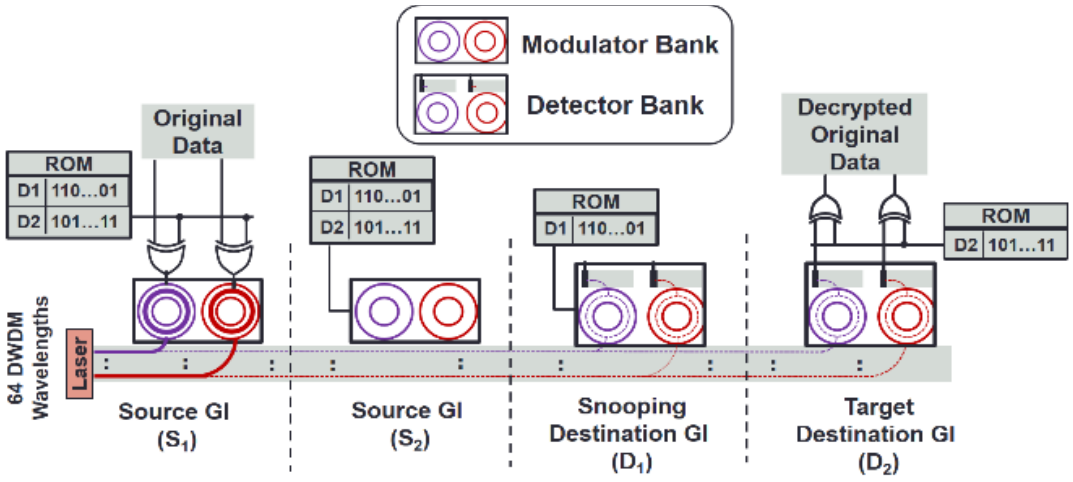


Figure 69: Overview of proposed PV-based security enhancement scheme.

Limitations of PVSC: The *PVSC* scheme can protect data from being deciphered by a snooping GI, if the following two conditions about the underlying PNoC architecture hold true: *(i)* the snooping GI does not know the target destination GI for the snooped data, *(ii)* the snooping GI cannot access the encryption key corresponding to the target destination GI. As discussed earlier, an encryption key is stored only at all source GIs and at the corresponding destination GI, which makes it physically inaccessible to a snooping destination GI. However, if more than one GIs in a PNoC are compromised due to HTs in their control units and if these HTs launch a coordinated snooping attack, then it may be possible for the snooping GI to access the encryption key corresponding to the target destination GI.

For instance, consider the photonic link in Figure 69. If both S_1 and D_1 are compromised, then the HT in S_1's control unit can access the encryption keys corresponding to both D_1 and D_2 from its ROM and transfer them to a malicious core (a core running a malicious program). Moreover, the HT in D_1's control unit can snoop the data intended for D_2 and transfer it to the

malicious core. Thus, the malicious core may have access to the snooped data as well as the encryption keys stored at the source GIs. Nevertheless, accessing the encryption keys stored at the source GIs is not sufficient for the malicious GI (or core) to decipher the snooped data. This is because the compromised ROM typically has multiple encryption keys corresponding to multiple destination GIs, and choosing a correct key that can decipher data requires the knowledge of the target destination GI. Thus, our *PVSC* encryption scheme can secure data communication in PNoCs as long as the malicious GIs do not know the target destinations of the snooped data.

Unfortunately, many PNoC architectures, e.g., [156] and [186], that employ photonic links with multiple destination GIs utilize the same waveguide to transmit both the target destination information and actual data. In these PNoCs, if a malicious GI manages to tap the target destination information from the shared waveguide, then it can access the correct encryption key from the compromised ROM to decipher the snooped data. Thus, there is a need to conceal the target destination information from malicious GIs (cores). This motivates us to propose an architecture-level solution, as discussed next.

9.6. RESERVATION-ASSISTED SECURITY ENHANCEMENT

In PNoCs that use photonic links with multiple destination GIs, data is typically transferred in two time-division-multiplexed (TDM) slots called reservation slot and data slot [156], [186]. To minimize photonic hardware, PNoCs use the same waveguide to transfer both slots, as shown in Figure 70(a). To enable reservation of the waveguide, each destination is assigned a reservation selection wavelength. In Figure 70(a), λ_1 and λ_2 are the reservation selection wavelengths corresponding to destination GIs D_1 and D_2, respectively. Ideally, when a destination GI detects its reservation selection wavelength in the reservation slot, it switches ON its detector bank to receive data in the next data slot. But in the presence of an HT, a malicious GI can snoop signals

from the reservation slot using the same detector bank that is used for data reception. For example, in Figure 70(a), malicious GI D_1 is using one of its detectors to snoop λ_2 from the reservation slot. By snooping λ_2, D_1 can identify that the data it will snoop in the subsequent data slot will be intended for destination D_2. Thus, D_1 can now choose the correct encryption key from the compromised ROM to decipher its snooped data.

To address this security risk, we propose an architecture-level reservation-assisted security enhancement (*RVSC*) scheme. In *RVSC*, we add a reservation waveguide, whose main function is to carry reservation slots, whereas the data waveguide carries data slots. We use double MRs to switch the signals of reservation slots from the data waveguide to the reservation waveguide, as shown in Figure 70(b). Double MRs are used instead of single MRs for switching to ensure that the switched signals do not reverse their propagation direction after switching (Chapter 2). Compared to single MRs, double MRs also have lower signal loss due to steeper roll-off of their filter response. The double MRs are switched ON only when the photonic link is in a reservation slot, otherwise they are switched OFF to let the signals of the data slot pass by in the data waveguide. Furthermore, in *RVSC*, each destination GI has only one detector on the reservation waveguide, which corresponds to its receiver selection wavelength. For example, in Figure 70(b), D_1 and D_2 will have detectors corresponding to their reservation selection wavelengths λ_1 and λ_2, respectively, on the reservation waveguide. This makes it difficult for the malicious GI D_1 to snoop λ_2 from the reservation slot as shown in Figure 70(b), as D_1 does not have a detector corresponding to λ_2 on the reservation waveguide. However, the HT in D_1's control unit may still attempt to snoop other reservation wavelengths (e.g., λ_2) in the reservation slot by retuning D_1's λ_1 detector. But succeeding in these attempts would require the HT to perfect the timing and target wavelength of

its snooping attack, which is very difficult due to the large number of utilized reservation wavelengths. Thus, D_1 cannot identify the correct encryption key to decipher the snooped data.

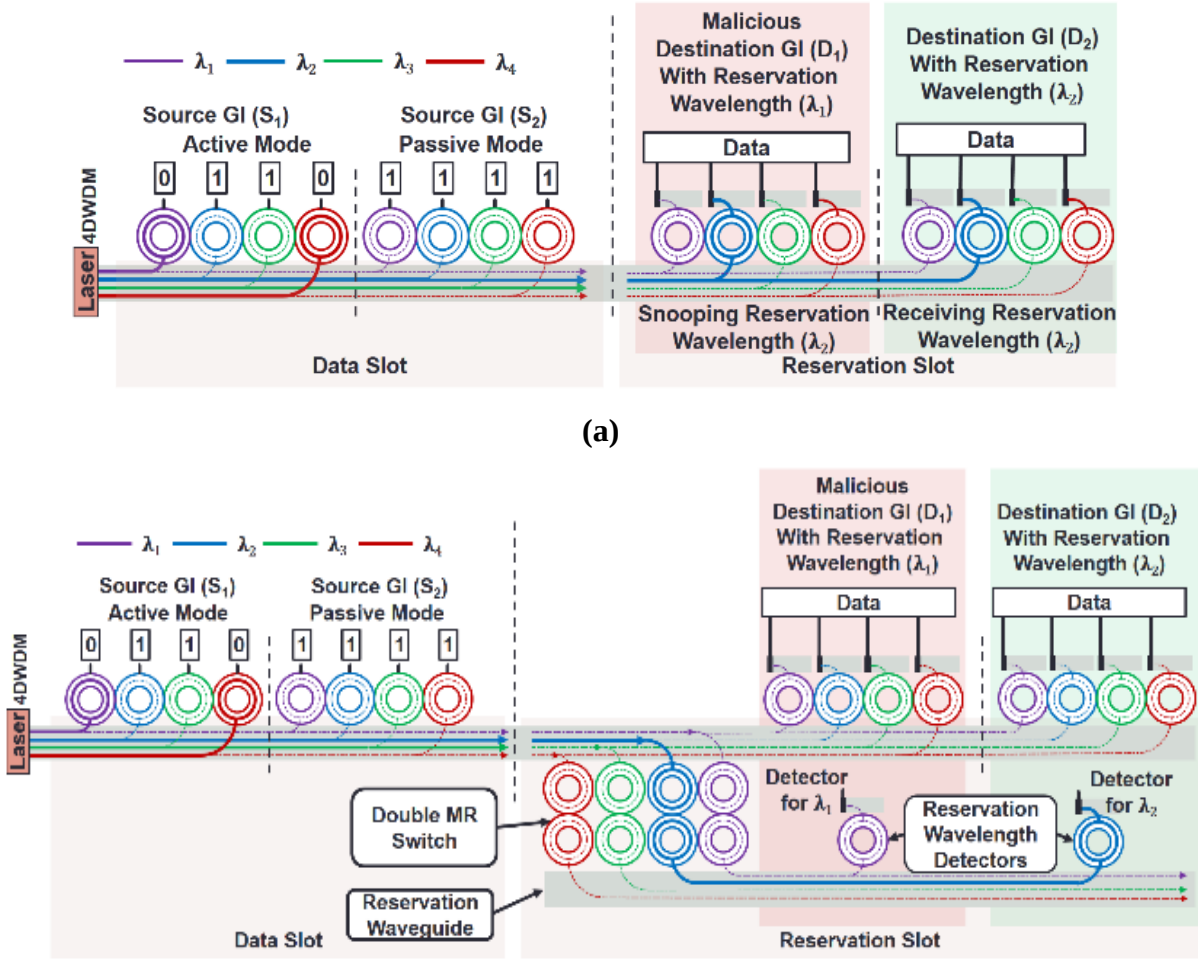


Figure 70: Reservation-assisted data transmission in DWDM-based photonic waveguides (a) without *RVSC*, (b) with *RVSC*.

In summary, RVSC enhances security in PNoCs by protecting data from snooping attacks, even if the encryption keys used for data encryption are compromised. To implement *RVSC* on a data waveguide with multiple destination GIs, we need to add a reservation waveguide with multiple detector MRs, where each detector MR corresponds to a destination GI. A group of double MRs, each of which corresponds to a reservation selection wavelength available in the waveguide,

is also needed to switch the wavelength signals of reservation slots from the data waveguide to the reservation waveguide. The introduction of the additional reservation waveguide and the group of double MRs increases signal loss and laser power. We account for this overhead in our PNoC architecture level analysis.

9.7. IMPLEMENTING *SOTERIA* FRAMEWORK ON PNOCS

We characterize the impact of *SOTERIA* on two popular PNoC architectures: Firefly [15] and Flexishare [16], both of which use DWDM-based photonic waveguides for data communication. We consider Firefly PNoC with 8×8 SWMR crossbar and a Flexishare PNoC with 32×32 MWMR crossbar with 2-pass token stream arbitration. We adapt the analytical equations from Chapter 2 to model the signal power loss and required laser power in the *SOTERIA*-enhanced Firefly and Flexishare PNoCs. At each source and destination GI of the *SOTERIA*-enhanced Firefly and Flexishare PNoCs, XOR gates are required to enable parallel encryption and decryption of 512-bit data flits. We consider a 1 cycle delay overhead for encryption and decryption of every data flit. The overall laser power and delay overheads for both PNoCs are quantified in the results section.

Firefly PNoC: Firefly PNoC [15], for a 256-core system, has 8 clusters (C1-C8) with 32 cores in each cluster. Firefly uses reservation-assisted SWMR data channels in its 8x8 crossbar for inter-cluster communication. Each data channel consists of 8 SWMR waveguides, with 64 DWDM wavelengths in each waveguide. To integrate *SOTERIA* with Firefly PNoC, we added a reservation waveguide to every SWMR channel. This reservation waveguide has 7 detector MRs to detect reservation selection wavelengths corresponding to 7 destination GIs. Furthermore, 64 double MRs (corresponding to 64 DWDM wavelengths) are used at each reservation waveguide to implement *RVSC*. To enable *PVSC*, each source GI has a ROM with seven entries of 512 bits each

to store seven 512-bit encryption keys corresponding to seven destination GIs. In addition, each destination GI requires a 512-bit ROM to store its own encryption key.

Flexishare PNoC: We also integrate *SOTERIA* with the Flexishare PNoC architecture [16] with 256 cores. We considered a 64-radix 64-cluster Flexishare PNoC with four cores in each cluster and 32 data channels for inter-cluster communication. Each data channel has four MWMR waveguides with each having 64 DWDM wavelengths. In *SOTERIA*-enhanced Flexishare, we added a reservation waveguide to each MWMR channel. Each reservation waveguide has 16 detector MRs to detect reservation selection wavelengths corresponding to 16 destination GIs. To enable *PVSC*, each source GI requires a ROM with 16 entries of 512 bits each to store the encryption keys, whereas each destination GI requires a 512-bit ROM.

9.8. EXPERIMENTS

9.8.1. EXPERIMENT SETUP

We To evaluate our proposed *SOTERIA* (*PVSC+RVSC*) security enhancement framework for DWDM-based PNoCs, we integrate it with the Firefly [15] and Flexishare [16] PNoCs, as explained in Section 0. We modeled and performed simulation based analysis of the *SOTERIA*-enhanced Firefly and Flexishare PNoCs using a cycle-accurate SystemC based NoC simulator, for a 256-core single-chip architecture at 22nm. We validated the simulator in terms of power dissipation and energy consumption based on results obtained from the DSENT tool [75]. We used real-world traffic from the PARSEC benchmark suite [76]. GEM5 full-system simulation [77] of parallelized PARSEC applications was used to generate traces that were fed into our NoC simulator. We set a “warmup” period of 100 million instructions and then captured traces for the subsequent 1 billion instructions. These traces are extracted from parallel regions of execution of PARSEC applications. We performed geometric calculations for a 20mm×20mm chip size, to

determine lengths of SWMR and MWMR waveguides in Firefly and Flexishare. Based on this analysis, we estimated the time needed for light to travel from the first to the last node as 8 cycles at 5 GHz clock frequency. We use a 512-bit packet size, as advocated in the Firefly and Flexishare PNoCs.

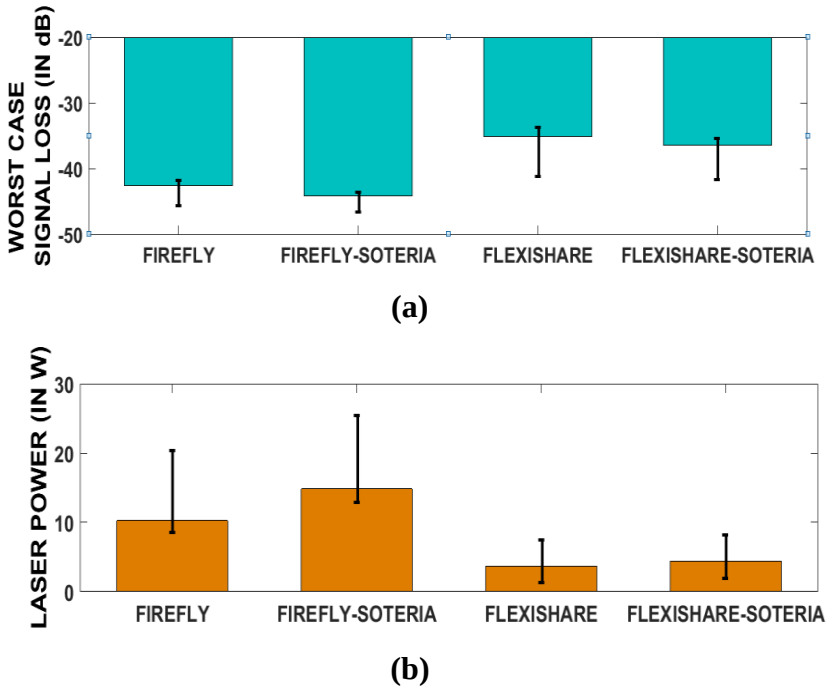


Figure 71: Comparison of (a) worst-case signal loss, (b) laser power dissipation of *SOTERIA* framework on Firefly and Flexishare PNoCs with their respective baselines considering 100 process variation maps.

The static and dynamic energy consumption values for electrical routers and concentrators in Firefly and Flexishare PNoCs are based on results from DSENT [75]. We model and consider the area, power, and performance overheads for our framework implemented with the Firefly and Flexishare PNoCs as follows. *SOTERIA* with Firefly and Flexishare PNoCs has an electrical area overhead of 12.7mm^2 and 3.4mm^2, respectively, and power overhead of 0.44W and 0.36W, respectively, using gate-level analysis and CACTI 6.5 [78] tool for memory and buffers. The photonic area of Firefly and Flexishare PNoCs is 19.83mm^2 and 5.2mm^2, respectively, based on

the physical dimensions [72] of their waveguides, MRs, and splitters. For energy consumption of photonic devices, we adapt model parameters from Chapter 3 with 0.42pJ/bit for every modulation and detection event and 0.18pJ/bit for the tuning circuits of modulators and photodetectors. The MR trimming power is 130µW/nm (Chapter 2) for current injection (to remedy PV-induced red shifts) and tuning power is 240µW/nm (Chapter 2) for heating (to remedy PV-induced blue shifts).

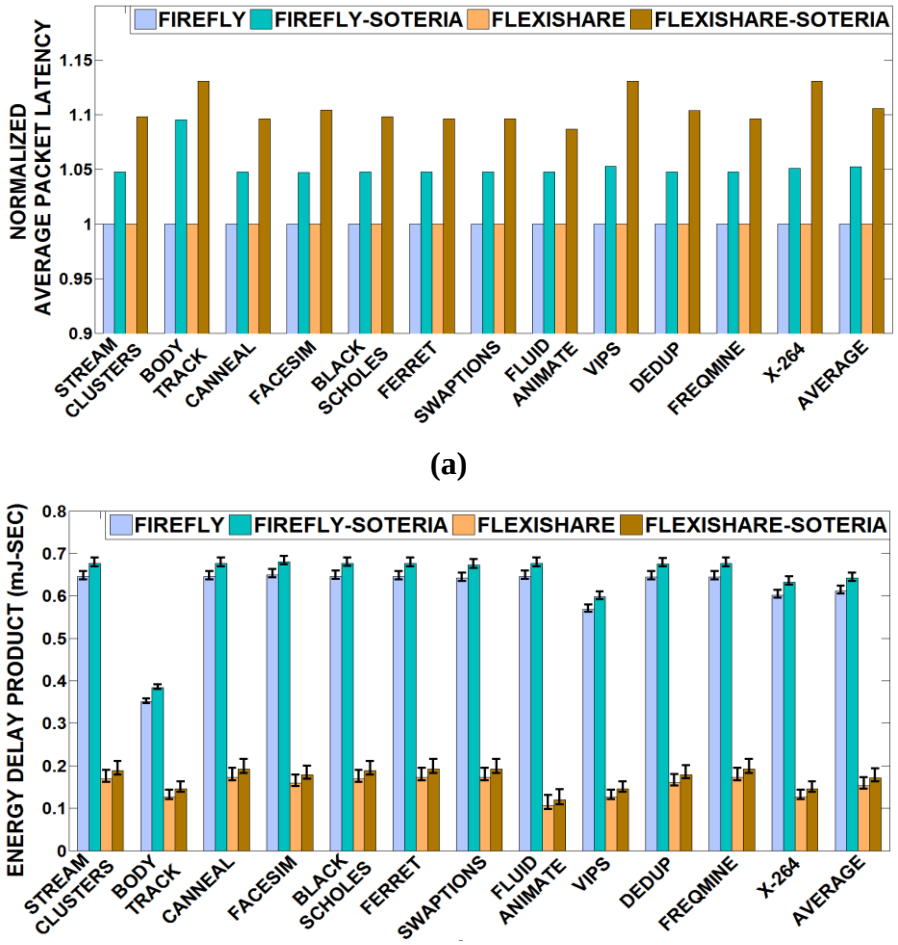


Figure 72: (a) normalized average latency, (b) energy-delay product (EDP) comparison between different variants of Firefly and Flexishare PNoCs that include their baselines and their variant with *SOTERIA* framework, for PARSEC benchmarks. Latency results are normalized with their respective baseline architecture results. Bars represent mean values of average latency and EDP for 100 PV maps; confidence intervals show variation in average latency and EDP across PARSEC benchmarks.

9.8.2. OVERHEAD ANALYSIS OF SOTERIA ON PNOCS

Our first set of experiments compare the baseline (without any security enhancements) Firefly and Flexishare PNoCs with their *SOTERIA* enhanced variants. From Section 0, all 8 SWMR waveguide groups of the Firefly PNoC and all 32 MWMR waveguide groups of the Flexishare PNoC are equipped with *PVSC* encryption/decryption and reservation waveguides for the *RVSC* scheme.

We adapt the analytical models from Chapter 2 to calculate the total signal loss at the detectors of the worst-case power loss node (N_{WCPL}), which corresponds to router C4R0 for the Firefly PNoC [15] and node R_{63} for the Flexishare PNoC [16]. Figure 71(a) summarizes the worst-case signal loss results for the baseline and *SOTERIA* configurations for the two PNoC architectures. From the figure, Firefly PNoC with *SOTERIA* increases loss by 1.6dB and Flexishare PNoC with *SOTERIA* increases loss by 1.2dB on average, compared to their respective baselines. Compared to the baseline PNoCs that have no single or double MRs to switch the signals of the reservation slots, the double MRs used in the *SOTERIA*-enhanced PNoCs to switch the wavelength signals of the reservation slots increase through losses in the waveguides, which ultimately increases the worst-case signal losses in the *SOTERIA*-enhanced PNoCs. Using the worst-case signal losses shown in Figure 71(a), we determine the total photonic laser power and corresponding electrical laser power (using laser wall-plug efficiency of 3%) for the baseline and *SOTERIA*-enhanced variants of Firefly and Flexishare PNoCs, shown in Figure 71(b). From this figure, the Firefly and Flexishare PNoCs with *SOTERIA* have laser power overheads of 44.7% and 31.40% on average, compared to their baselines.

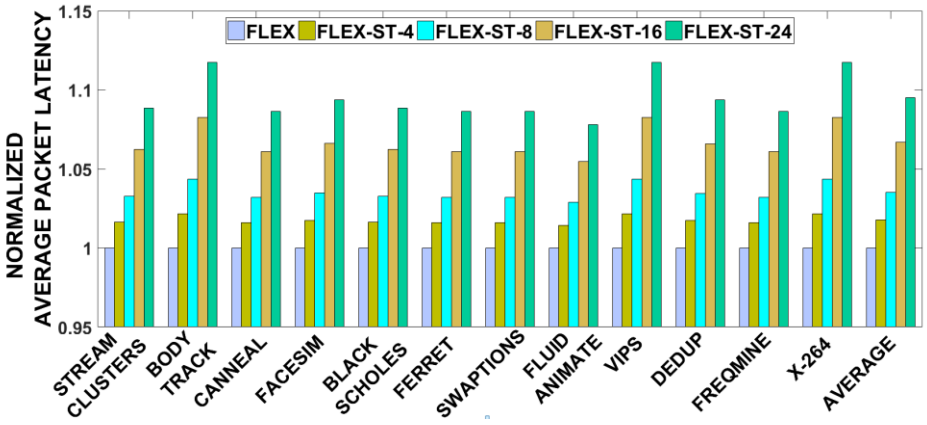
Figure 72 presents detailed simulation results that quantify the average packet latency and energy-delay product (EDP) for the two configurations of the Firefly and Flexishare PNoCs.

Results are shown for twelve multi-threaded PARSEC benchmarks. From Figure 72(a), Firefly with *SOTERIA* has 5.2% and Flexishare with *SOTERIA* has 10.6% higher latency on average compared to their respective baselines. The additional delay due to encryption and decryption of data (Section 0) with *PVSC* contributes to the increase in average latency.

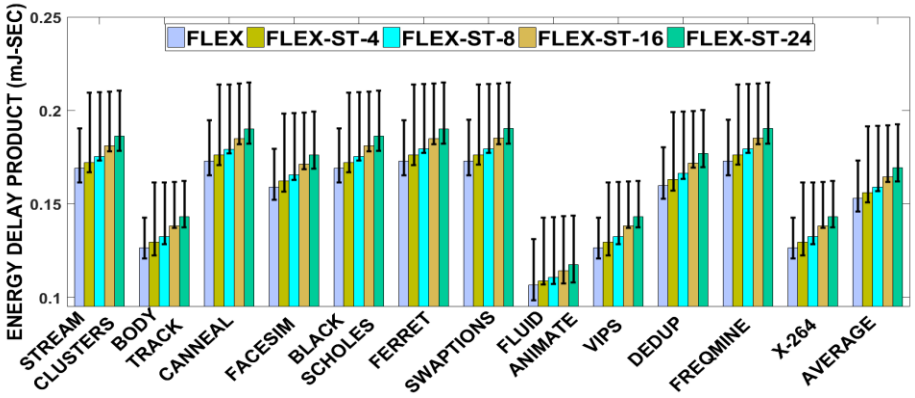
From the results for EDP shown in Figure 72(b), Firefly with *SOTERIA* has 4.9% and Flexishare with *SOTERIA* has 13.3% higher EDP on average compared to their respective baselines. Increase in EDP for the *SOTERIA*-enhanced PNoCs is not only due to the increase in their average packet latency, but also due to the presence of additional *RVSC* reservation waveguides, which increases the required photonic hardware (e.g., more number of MRs) in the *SOTERIA*-enhanced PNoCs. This in turn increases static energy consumption (i.e., laser energy and trimming/tuning energy), ultimately increasing the EDP. From the results presented in this section, we can conclude that our *SOTERIA* framework improves hardware security in PNoCs at the cost of additional laser power, average latency, and EDP overheads.

9.8.3. ANALYSIS OF OVERHEAD SENSITIVITY

Our last set of evaluations explore how the overhead of *SOTERIA* changes with varying levels of security in the network. Typically, in a manycore system, only a certain portion of the data that contains sensitive information (i.e., keys) and only a certain number of communication links need to be secure. Therefore, for our analysis in this section, instead of securing all data channels of the Flexishare PNoC, we secure only a certain number channels using *SOTERIA*. Out of the total 32 MWMR channels in the Flexishare PNoC, we secure 4 (FLEX-ST-4), 8 (FLEX-ST-8), 16 (FLEX-ST-16), and 24 (FLEX-ST-24) channels, and evaluate the average packet latency and EDP for these variants of the *SOTERIA*-enhanced Flexishare PNoC.



(a)



(b)

Figure 73: (a) normalized latency, (b) energy-delay product (EDP) comparison between Flexishare baseline and Flexishare with 4, 8, 16, and 24 *SOTERIA* enhanced MWMR waveguide groups, for PARSEC benchmarks. Latency results are normalized to the baseline Flexishare results.

In Figure 73, we present average packet latency and EDP values for the five *SOTERIA*-enhanced configurations of the Flexishare PNoC. From Figure 73(a), FLEX-ST-4, FLEX-ST-8, FLEX-ST-16, and FLEX-ST-24 have 1.8%, 3.5%, 6.7%, and 9.5% higher latency on average compared to the baseline Flexishare. Increase in number of *SOTERIA* enhanced MWMR waveguides increases number of packets that are transferred through the *PVSC* encryption scheme, which contributes to the increase in average packet latency across these variants. From the results for EDP shown in Figure 73(b), FLEX-ST-4, FLEX-ST-8, FLEX-ST-16, and FLEX-ST-24 have

2%, 4%, 7.6%, and 10.8% higher EDP on average compared to the baseline Flexishare. EDP in Flexishare PNoC increases with increase in number of *SOTERIA* enhanced MWMR waveguides. Increase in average packet latency and signal loss due to the higher number of reservation waveguides and double MRs increase overall EDP across these variants.

9.8.4. SUMMARY OF RESULTS AND OBSERVATIONS

From the results in the previous subsections, it can be concluded that our proposed *SOTERIA* framework secures data during unicast communications in PNoC architectures from snooping attacks by leveraging the benefits of our circuit-level *PVSC* and architecture-level *RVSC* techniques. *SOTERIA*-enhanced PNoCs incur minimal overheads of up to 10.6% and as low as 1.8% in average packet latency and up to 13.3% and as low as 2% in EDP compared to the baseline insecure PNoCs.

9.9. CONCLUSIONS

We presented a novel security enhancement framework called *SOTERIA* that secures data during unicast communications in DWDM-based PNoC architectures from snooping attacks. Our proposed *SOTERIA* framework shows interesting trade-offs between security, performance, and energy overhead for the Firefly and Flexishare PNoC architectures. Our analysis showed that *SOTERIA* enables hardware security in crossbar based PNoCs with minimal overheads of up to 10.6% in average latency and of up to 13.3% in EDP compared to the baseline PNoCs. Thus, we represented *SOTERIA* as an attractive solution to enhance hardware security in emerging DWDM-based PNoCs. In the future, we plan to extend our *SOTERIA* framework to enhance data security during broadcast and multicast communications in PNoCs.

10. 3D-PROWIZ: AN ENERGY-EFFICIENT AND OPTICALLY-INTERFACED 3D DRAM ARCHITECTURE WITH REDUCED DATA ACCESS OVERHEAD

This chapter introduces *3D-ProWiz*, which is a high-bandwidth, energy-efficient, optically-interfaced 3D DRAM architecture with fine grained data organization and activation. *3D-ProWiz* integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism. A novel method of routing the internal memory bus to individual subarrays using TSVs and fanout buffers enables *3D-ProWiz* to use smaller dimension subarrays without significant area overhead. The use of TSVs at subarray-level granularity eliminates the need to use slow and power hungry global lines, which in turn reduces the random-access latency and activation-precharge energy. *3D-ProWiz* yields the best latency and energy consumption values per access among other well-known 3D DRAM architectures. Experimental results with PARSEC benchmarks indicate that *3D-ProWiz* achieves 41.9% reduction in average latency, 52% reduction in average power, and 80.6% reduction in energy-delay product (EDP) on average over DRAM architectures from prior work.

10.1. BACKGROUND AND CONTRIBUTIONS

In recent years, DRAM latency has not improved as rapidly as DRAM capacity and bandwidth with shrinking technology, owing to the well-known "memory-wall" problem [190]. Continued process technology scaling has enabled commodity memory system solutions to exploit smaller and faster transistors to improve memory capacity and bandwidth. However, the traditional means of improving memory access performance by increasing clock frequency is no longer practical due to increasingly stringent power constraints that limit further frequency scaling. Thus, performance improvements in memory systems today are primarily reliant on latency tolerance

techniques such as multi-level caches, row-prefetching, burst-mode access, memory scheduling [191], and memory parallelism [192], [193]. But the performance improvements obtained through these techniques are not expected to scale well for high performance computing systems of the future [194]. Moreover, preserving the minimum standard capacitance of a DRAM cell is becoming increasingly challenging with shrinking feature size [195]. These trends are forcing designers to reinvent DRAM architectures so as to overcome the hurdles in DRAM performance scaling.

Since the emergence of 3D integration technology, 3D-stacked DRAM has been a promising option to alleviate many of the physical limitations of commodity DRAMs. In recent years, several 3D DRAM architectures have been proposed [23], [164], [196]-[201], that have exploited high bandwidth through silicon vias (TSVs) to improve memory latency and throughput. Such 3D-stacked DRAM designs require new methods for efficient address and data path routing and 3D cell organization, to realize their full potential and achieve improved memory parallelism and energy-efficiency.

In this chapter, first of all we identify the critical elements of a 3D-stacked DRAM architecture that contribute significantly to the overall latency, energy consumption, and die area. Then, we propose 3D-ProWiz, a new 3D DRAM architecture with a photonic interface that tones down the unfavorable effects of the critical DRAM elements to improve access latency and energy consumption characteristics of DRAM subsystem over prior efforts. Our 3D-ProWiz DRAM architecture is an extension of the 3D-Wiz DRAM architecture [164] with notable improvements in the bank floorplan that adhere to pitch-matching rules and awareness of inter-component connections at the subarray level, which makes the manufacturing and implementation of 3D-ProWiz DRAM more feasible. Our key contributions in this chapter can be summarized as follows:

- *Fine-grained 3D organization of DRAM data array:* Bank-level parallelism in a commodity DRAM data array is limited by the total number of banks the data array is partitioned into. A typical DRAM data array is partitioned into 4 to 16 banks. These banks are very large in capacity, ranging from 32Mb to 256Mb, and hence they are few in number. To enable higher memory parallelism in 3D-ProWiz, its data array (DRAM rank) is divided into 512 smaller banks. A very large number of banks greatly increase bank-level parallelism in 3D-ProWiz. In addition, we conservatively estimate a reasonable non-zero value of the tTAW (two banks activation window) constraint, which limits the power rail noise to a safe level while exploiting increased bank-level parallelism. Moreover, each bank in 3D-ProWiz is many-fold smaller than a commodity bank and acts independently, which greatly reduces data access energy.
- *TSV-based internal memory bus:* In 3D-stacked DRAM, the address and data buses are routed to each die layer in the vertical direction using TSVs. After reaching a die they are routed to the edges of individual banks along 2D routing paths. These 2D routing paths (or global lines) and global peripheral circuits (decoders, repeaters and drivers of global lines) incur latency, energy, and area overhead at each die. Contrarily, in 3D-ProWiz DRAM, we make intelligent use of TSVs in conjunction with fanout buffers at finer granularity to route data and address lines to individual subarrays, which eliminates the need to use 2D routing paths and global peripheral circuits at each die. This in turn enables the use of smaller sized banks and subarrays to reduce the overall area, latency, and energy consumption.
- *Reduced access time:* The DRAM access time is a function of the sum of row to column command delay (tRCD) and column command to data out delay (tCAS). For a subarray of interest, the timing constraints tRCD and tCAS are proportional to the capacitive loading of

wordlines and bitlines respectively, which are increasing functions of the number of columns and number of rows respectively [193]. We reduced the access time of 3D-ProWiz by carefully reducing the number of rows and columns in a subarray so as to not harm the DRAM cell area efficiency compared to the area efficiency of other 3D-stacked DRAM architectures. Moreover, fanout buffers and the TSV-based internal memory bus help further reduce access time in 3D-ProWiz.

- *High-bandwidth photonic interface:* We observed that the conventional electrical interfaces of the DDRx family [202], [203] and differential lane based interfaces [204] cannot support very high bandwidths. So in 3D-ProWiz, we propose using a higher bandwidth, dense wavelength division multiplexing (DWDM) based photonic interface. We also investigated the energy-efficiency of conventional electrical interfaces such as DDR3 [202], LPDDR3 [203], Wide-I/O [24] and differential serial interface [204] in terms of energy-per-byte values and compared it with the energy-efficiency of the proposed photonic interface when used with 3D-ProWiz. The results of this comparison showed that use of a photonic interface with 3D-ProWiz greatly improves throughput and energy-efficiency of the DRAM subsystem.
- *Sensitivity analysis:* We analyzed the sensitivity of our 3D-ProWiz architecture to different address mapping policies and memory scheduling policies for PARSEC benchmarks to determine a combination of policies that achieve the best energy-efficiency in terms of energy-delay product. We also analyzed the sensitivity of average latency of 3D-ProWiz and other DRAM architectures to different values of the tTAW constraint in an effort to understand the relationship between the tTAW constraint, bank-level memory parallelism, and average latency.

10.2. BACKGROUND AND MOTIVATION

In this section, we briefly discuss the state-of-the-art data array organizations for 3D-stacked DRAMs, and identify the fundamental elements that contribute significantly to the overall DRAM access latency, energy consumption, and die area.

Different organizations of data array differ in the way banks and ranks are stacked and partitioned across the 3D die-stack. Accordingly, the 3D organizations of data array are referred to as coarse-grained or fine-grained rank-level partitioning, or bank-level partitioning [205]. Loh [196] was the first to highlight the potential performance benefits achievable by altering the data array organization for conventional 3D-stacked DRAMs. Previous approaches (prior to [196]) did not fully exploit 3D-stacking technology, as the individual structures inside DRAMs were still 2D. Loh [196] proposed splitting a rank across multiple layers instead of laying out a rank on a single layer, and showed up to 1.75× DRAM performance improvement. Woo et al. [197] proposed the SMART-3D DRAM architecture that used a vertical L2-fetch and write-back network made up of a large array of TSVs to hide the latency behind large data transfers. Kang et al. [198] proposed extending the commodity DDR3 architecture to 3D-DDR3 and showed benefits in energy consumption due to the reduced length of the TSV-based memory bus. The hybrid memory cube (HMC) exploits fine-grained rank-level partitioning to further reduce the access latency and increase memory parallelism [23]. All of these approaches [23], [196]-[198] exploit the bandwidth from low-energy vertical TSVs to reduce access latency and/or energy. But, they miss out on the potential of 3D data array organization and TSVs to more aggressively improve performance, reduce access energy, and improve energy-efficiency in DRAM data arrays.

Chen et al. [205] discussed the pros and cons of coarse-grained and fine-grained rank-level partitioning for 3D DRAMs with respect to latency, energy consumption, and area efficiency. For

the coarse-grained rank-level cell organization, the inter-bank communication within one rank occurs via 2D routing paths. It is shown that activation energy and latency of the fine-grained design are reduced by 48.5% and 46.9% respectively compared to the coarse-grained design, because of the reduced bank size and optimized data path routing in the fine-grained design. The total die area for the fine-grained design reduces by 35.9% in spite of a 3.7% TSV area overhead. Also, the latency of the internal memory bus is reduced by 62.8% for the fine-grained design. The reason behind the improved results for the fine-grained model is that the model utilizes the potential bandwidth of TSVs for inter-bank transfers, which relaxes the need for 2D routing paths for such transfers and alleviates related overheads. From these results, it can be concluded that many of the limitations that arise with DRAM scaling can be overcome by intelligently organizing the data array of 3D-stacked DRAMs.

Recent advancements in TSV-based 3D-stacked DRAM technology have enabled the use of high density TSVs at subarray-level granularity [164], [199], [201], which has enabled elimination of 2D routing paths (global lines) and provided new opportunities in designing low-power and high band-width data organizations. Thakkar et al. [164] proposed the 3D-Wiz DRAM architecture for high-performance systems, which eliminates global lines by employing aggressive vertical routing of intra-bank buses using TSVs and fanout buffers. The same authors also proposed the 3D-WiRED DRAM architecture in [199] for use in energy-constrained embedded systems. 3D-WiRED renders high energy-efficiency due to its 3D folded bank organization. To ease implementation of 3D-DRAM architectures that use TSVs at subarray level granularity (such as 3D-Wiz [164], 3D-WiRED [199], and [201]), the layout of a DRAM bank should be designed at subarray level abstraction and it should be aware of inter-component connections. But the layouts of the 3D-Wiz bank [164] and 3D-WiRED bank [199] are unaware of inter-component connections

at the subarray level. The 3D-ProWiz DRAM architecture presented in this chapter notably extends the 3D-Wiz DRAM architecture [164] with an improved bank floorplan which is aware of inter-component connections at the subarray level. Moreover, we show that by carefully choosing the bank size and an appropriate scheme for TSV based routing of address and data paths, the performance and energy consumption of 3D-ProWiz DRAM can be improved even further.

The organization of the DRAM data array is only one of the limiting factors that affect DRAM performance and energy. The other more fundamental limiting factor is RC loading of the memory access path, which is a function of the number of rows and columns in the accessed subarray. This fact encourages designers to reduce the number of rows and columns in a subarray. But, such a reduction can significantly increase the area overhead, harming overall area efficiency, if it is done without due deliberation. 3D-ProWiz exploits the benefits of intelligent array organization and aggressive vertical routing of address and data paths to counteract this overhead, which enables the use of smaller dimension subarrays.

Additional performance benefits can also be achieved by increasing memory access concurrency inside DRAMs. Zhang et al. proposed 3D-SWIFT [200], which increases bank-level parallelism by employing a large number of small banks. They divide each bank of 3D-SWIFT into 16 smaller banks and operate them independently. However, 3D-SWIFT [200] does not utilize the full potential of high bandwidth TSVs, as it employs the conventional 2D structure of banks. Moreover, aggressive changes made in DRAM organization to achieve higher access concurrency and excessive use of available concurrency may result in suboptimal power consumption, which leads to excessive increase in operating temperature and increased noise in the power delivery network (PDN) of 3D-stacked DRAMs. To keep the power consumption and PDN noise within allowable limits, the bank-level parallelism in 3D-stacked DRAMs is curtailed by the tTAW

constraint. In general, the tTAW constraint controls bank-level parallelism by allowing only two bank activates in a rolling window of tTAW time. In 3D-SWIFT [200], the tTAW constraint is over-optimistically estimated to be zero. But, the definition of tTAW constraint implies that increasing the bank-level concurrency of the memory module within power noise limits requires careful design of non-zero tTAW constraint. Therefore, in our 3D-ProWiz architecture we consider a reasonable non-zero value of the tTAW constraint. As the performance and power benefits of memory parallelism also depend on ad-dress mapping and scheduling policies, we investigate the sensitivity of 3D-ProWiz to a combination of different ad-dress mapping and scheduling policies, as well as different values of tTAW constraint to determine a configuration that achieves the best performance and energy-efficiency.

Lastly, state-of-the-art DRAM modules mainly focus on exploiting concurrency inside the banks (as discussed above) due to limited concurrency outside the banks as a result of bus contention. The contention arises because the internal memory bus and interface bus are shared among all the banks due to pin-bandwidth limitation of the traditional DDR interface [202]. The Wide I/O standard [24], a new JEDEC standard for direct chip-to-chip interfacing of DRAM dies with the processor/controller, has the potential to alleviate these shortcomings for on-chip embedded DRAMs. A Wide I/O DRAM employs a wider interface (typically made of 128 TSVs) to increase the peak band-width of memory-to-core interconnect. But the TSV based core-to-memory interface can be realized only for on-chip embedded DRAMs. Its use for off-chip main memory is limited due to pin-bandwidth limitations of off-chip PCB interconnects. In the through-silicon interposer (TSI) inter-connect based memory-logic integration, the pin-bandwidth limitation can be alleviated using space-time multiplexed 2.5D I/O channels as done in [206]. The energy-efficiency of such 2.5D I/Os can be improved by adaptively adjusting the I/O output-

voltage swing under constraints of both communication power and bit error rate [207]. In contrast, our proposed 3D-ProWiz DRAM targets off-chip interconnect based memory-logic integration and uses intelligent layout of TSV buses and a high speed photonic interface to address the pin-bandwidth limitation.

10.3. 3D-PROWIZ ARCHITECTURE: OVERVIEW

In this section, an example DRAM design is described to demonstrate the 3D-ProWiz architecture. As implied from the discussion in Section 10.2, we aim to use TSVs at subarray-level granularity along with smaller subarrays and greater bank-level parallelism in 3D-ProWiz to achieve greater performance and energy-efficiency without significantly impacting area and cost. As we focus on reducing the capacitive loading of bitlines and wordlines by using smaller subarrays, we avoid time consuming exploration of the subarray design space and assume the organization of bitcells in a smaller subarray of 3D-ProWiz to be based on DRAM solutions found in [201] and in Tezzaron's DiRAM [26]. Consequently, each subarray within 3D-ProWiz DRAM is a 256×256 arrangement of bitcells and is of size 28μm × 42μm. This subarray size is smaller than the subarray size (512×512) found in all well-known 3D DRAMs. The following subsections describe the 3D-ProWiz architecture in more detail.

10.3.1. 3D-PROWIZ MODULE

In this section, we describe an example design of a 3D-ProWiz module. As shown in Figure 74(a), 8Gb modules (at 45nm) of both 3D-Wiz [164] and 3D-ProWiz DRAMs consist of a stack of 4 dies, which is divided into 4 identical ranks. Each rank has eight bankgroups. As shown in Figure 74(b), each bankgroup of a 3D-ProWiz module consists of 64 identical banks. In contrast, each bankgroup of a 3D-Wiz module consists of 128 identical banks, as shown in Figure 74(c).

The four DRAM dies of both 3D-Wiz and 3D-ProWiz modules are stacked onto one logic die. For both 3D-Wiz and 3D-ProWiz modules, all of the global control logic (except subarray-level control) of the DRAM is integrated on the logic die. The logic die also contains all of the opto-electrical circuits required to support a bidirectional, 256-bit wide, DWDM photonic data bus between the processor side memory controller and the memory module.

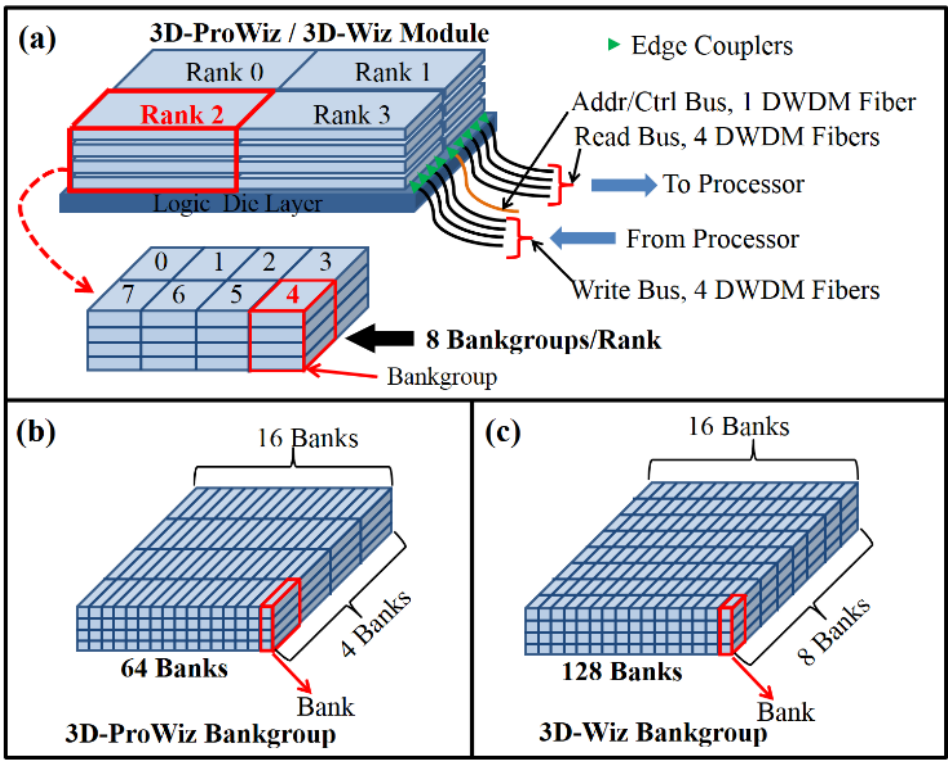


Figure 74: (a) Schematic of 3D-Wiz [11] and 3D-ProWiz modules and constituent elements, (b) schematic of 3D-ProWiz bankgroup, (c) schematic of 3D-Wiz bankgroup.

In state-of-the-art 3D-stacked DRAMs, the tTAW power constraint significantly reduces bank-level parallelism in a rank [208]. However, the relatively smaller size of banks in 3D-ProWiz relaxes the tTAW constraint. The internal data bus of 3D-ProWiz is also not shared among banks, but rather each bank has its own data bus made of TSVs. There-fore, bank-level parallelism, which is generally limited by the tTAW constraint and the shared internal memory bus together, is limited

by only the tTAW constraint in our 3D-ProWiz module. This improves bank-level parallelism for 3D-ProWiz. The analysis of the tTAW constraint for a 3D-ProWiz module is given in Section 10.5.

To effectively exploit the high-level of parallelism in a 3D-ProWiz module, we use a high bandwidth photonic interface so that a larger number of requests to different banks can be pipelined through the photonic bus. We have found in our studies that the use of a high-bandwidth DWDM based photonic interface significantly improves the throughput and energy-efficiency of the 3D-ProWiz subsystem over the DDR3 [202], LPDDR3 [203] and differential serial interface [204]. The methodology used to perform this analysis along with the detailed results of the analysis and the structure of the proposed photonic interface are described in Section 10.6.

10.3.2. FLOORPLAN OF 3D-WIZ BANK

To understand the difference between 3D-ProWiz and 3D-Wiz floorplans, firstly the floorplan of a 3D-Wiz bank is briefly revisited here. As described in [164], a 3D-Wiz bank is of 2Mb size and is folded across four DRAM layers. Each bank consists of 32 subarrays, 8 subarrays of which are on one layer. Figure 75 shows a schematic layout of the partition of 3D-Wiz bank on one die. Two adjacent banks in 3D-Wiz end up sharing a section of a TSV bus. Each bank is 8 subarrays long, one subarray wide, and 4 die layers high. All 32 constituent subarrays of a bank are addressed in parallel, and they work in lockstep to serve a cache line. As all the subarrays in a bank are addressed in parallel, only one set of row address and column address TSVs are used to address all 32 in-unison subarrays of a bank. The pre-decoder and decoder circuits for row and column ad-dress lines are located on the logic die. The decoded address lines are routed vertically via the TSV bus section, which feeds the local word lines and column select lines of subarrays in a bank. In this case, each address TSV, when it reaches a die layer, is used to feed in one 1:8 fanout

buffer after a repeater stage. The output of the 1:8 fanout buffer feeds decoded address signals to all 8 on-die in-union subarrays in parallel. Using one 1:8 fanout buffer per ad-dress TSV on each die layer enables 3D-Wiz to drive all 32 subarrays of a bank using just one set of address TSVs. The reader is directed to [164] for more details on 3D-Wiz floor-plan.

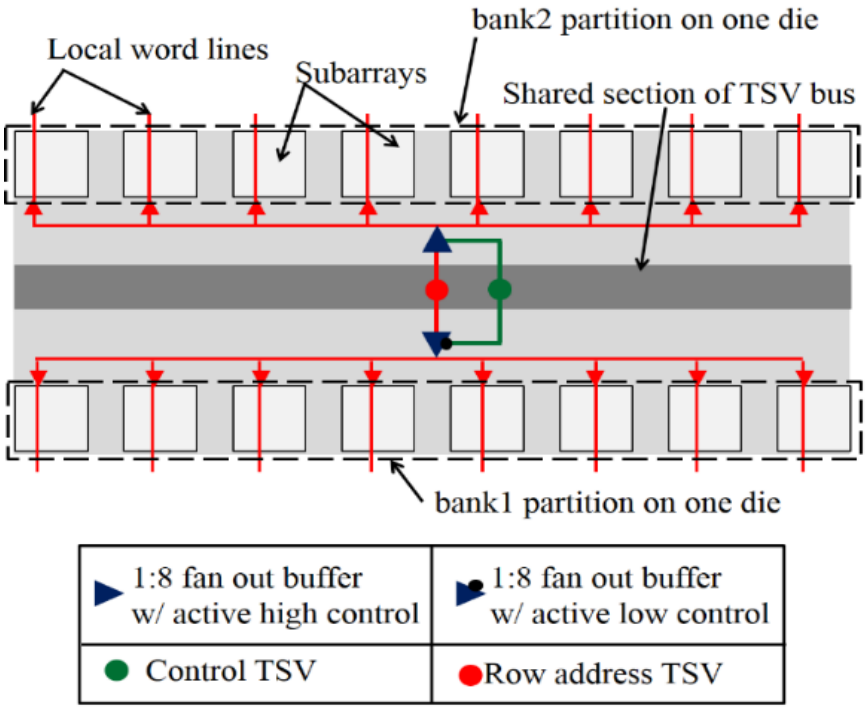


Figure 75: Schematic layout of the partition of two adjacent banks and shared TSV bus section on one die of 3D-Wiz [11] architecture.

10.3.2.1. ADHERENCE TO PITCH-MATCHING RULES

In this subsection, we explain how the use of fanout buffers for connecting TSVs to individual subarrays facilitates the floorplan of a 3D-Wiz bank to adhere to the pitch-matching rules. According to the physical design rules of 3D-stacked integrated circuits (3D-SIC), the pitch of adjacent circuit blocks on a die should match. As 3D-Wiz DRAM uses on-die metal wires to connect the output of fanout buffers to local wordline drivers, the pitch of on-die metal wires should match with the pitch of local wordline drivers. An intelligent use of shared bitline contacts

and diagonal polysilicon in the folded bitline structure of mod-ern subarrays have reduced the pitch of local wordline drivers to 4-5L [209], where L refers to minimum transistor length or feature size. Typically, on-die metal wires have a pitch of 2L [210], which matches with the pitch of local wordline drivers. On the other hand, as the fanout buffers of 3D-Wiz DRAM are fed in by TSVs, the pitch of fanout buffers should match TSV pitch. 3D-Wiz DRAM uses intermediate interconnect type of TSVs, the pitch of which is about 2.4µm [31]. A 1:8 fanout buffer realized using the design rules of 45nm technology would have 0.6µm × 1.6µm dimensions [209], which matches with the TSV pitch of 2.4µm. Thus, the pitch of all the adjacent circuit blocks in the floorplan of a 3D-Wiz bank match with each other.

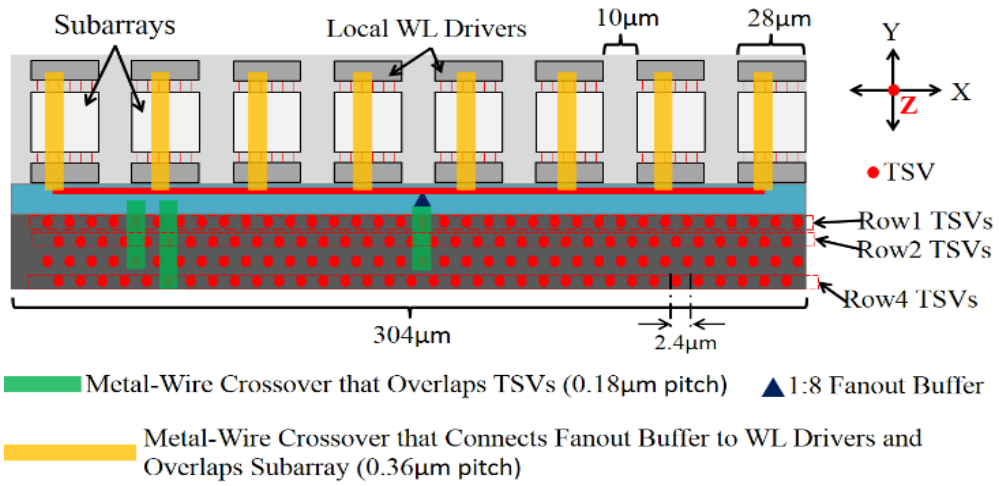


Figure 76: Schematic floorplan of a 3D-Wiz bank with dimensions and metal-wire crossovers. (The figure is not drawn to the scale)

10.3.2.2. INEFFICIENT LAYOUT OF ON-DIE METAL WIRES

In this subsection, we explain why the floorplan of a 3D-Wiz bank is not feasible even when the pitch of the adjacent blocks of the bank match with each other. In a 3D-Wiz bank, the on-die metal wires responsible for inter-component interconnection at the subarray level are inefficiently laid out. This makes the manufacturing of 3D-Wiz DRAM infeasible.

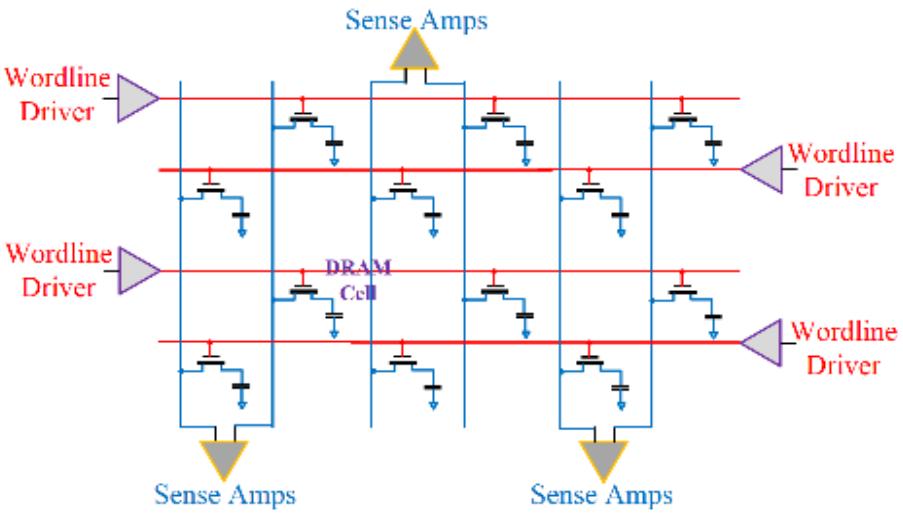


Figure 77: Schematic subarray structure with folded bitlines.

Consider Figure 76, which shows a schematic floorplan of a 3D-Wiz bank with dimensions and metal-wire crossovers. The length of a 3D-Wiz subarray along the X-direction is 38µm (including 10µm for peripherals), which makes the length of the 3D-Wiz bank to be 304µm as there are 8 subarrays in a bank along the X-direction. This implies that only 126 TSVs can fit in one row along the X-direction. A 3D-Wiz bank has total 448 TSVs [265] arranged in 4 rows, in which there are 256 row address TSVs, 64 column address TSVs and 128 data TSVs. As shown in the figure, the TSVs in row2, row3 and row4 have to connect to the fanout buffers through metal-wire crossovers (shown in green color in the figure) partially overlapping the launch pads of at least one row of TSVs. For instance, consider the metal-wire crossover (in green color) that connects a TSV of row3 to a fanout buffer (shown as dark blue triangle). This particular metal-wire crossover overlaps the adjacent TSV of row1 to connect to the fanout buffer, which interrupts the inter-die connection of the overlapped TSV (in the vertical Z-direction) across the die-stack. This results in a very inefficient arrangement that complicates the realization of the 3D-Wiz stack.

Moreover, the 3D-Wiz floorplan shown in Figure 76 does not account for how the subarray is laid out, which makes the floorplan even more inefficient. In modern subarray structures with folded bitlines, as shown in Figure 77, the driver of every alternate wordline is on the opposite side of the subarray. So, as shown in Figure 76, some of the fanout buffers have to connect to the wordline drivers on the other side of the subarray through metal-wire crossovers (shown in yellow color in Figure 76) that act as global wires overlap-ping the subarray. This arrangement acts against the original idea of eliminating global wires. In summary, from the preceding discussion it is apparent that the original floor-plan of 3D-Wiz is infeasible due to the inefficient use of metal-wire crossovers that connect different components of the bank overlapping TSVs and subarrays. Alternatively, in a 3D-Wiz bank, TSVs can be arranged on both sides of subarrays to efficiently arrange the connections of fanout buffers to respective wordline drivers without using metal-wire crossings that overlap subarrays. However, this provision would still require TSVs to be arranged in two rows along the X-direction on each side of the subarrays. So, this arrangement does not solve the infeasibility problem, as it would still result in metal-wire crossovers that overlap TSVs hindering their connection to upper layers. Therefore, we modify the floorplan of 3D-Wiz DRAM and derive a more efficient and feasible floorplan as part of our 3D-ProWiz DRAM architecture.

10.3.2.3. FLOORPLAN OF 3D-PROWIZ RANK AND BANK

In this subsection, we describe the efficient and feasible floorplan of 3D-ProWiz DRAM. Figure 79 shows the layout of all the 3D-ProWiz banks and constituent subarrays of a rank on one die layer, along with the TSV bus layout.

As shown in the figure, a rank is partitioned into 512 identical banks (64 banks along Y-axis and 8 banks along X-axis). Each bank is folded across 4 die layers (along Z-axis) and consists of a total of 64 subarrays, 16 subarrays of which are on one layer. All 64 constituent subarrays of a

bank are addressed in parallel, and they work in lockstep to serve a cache line. Each subarray has 2 data lines, which allows a bank to serve a total of 128 data bits in one data burst. Therefore, 3D-ProWiz requires a burst length of 2 to serve a 32B cache line.

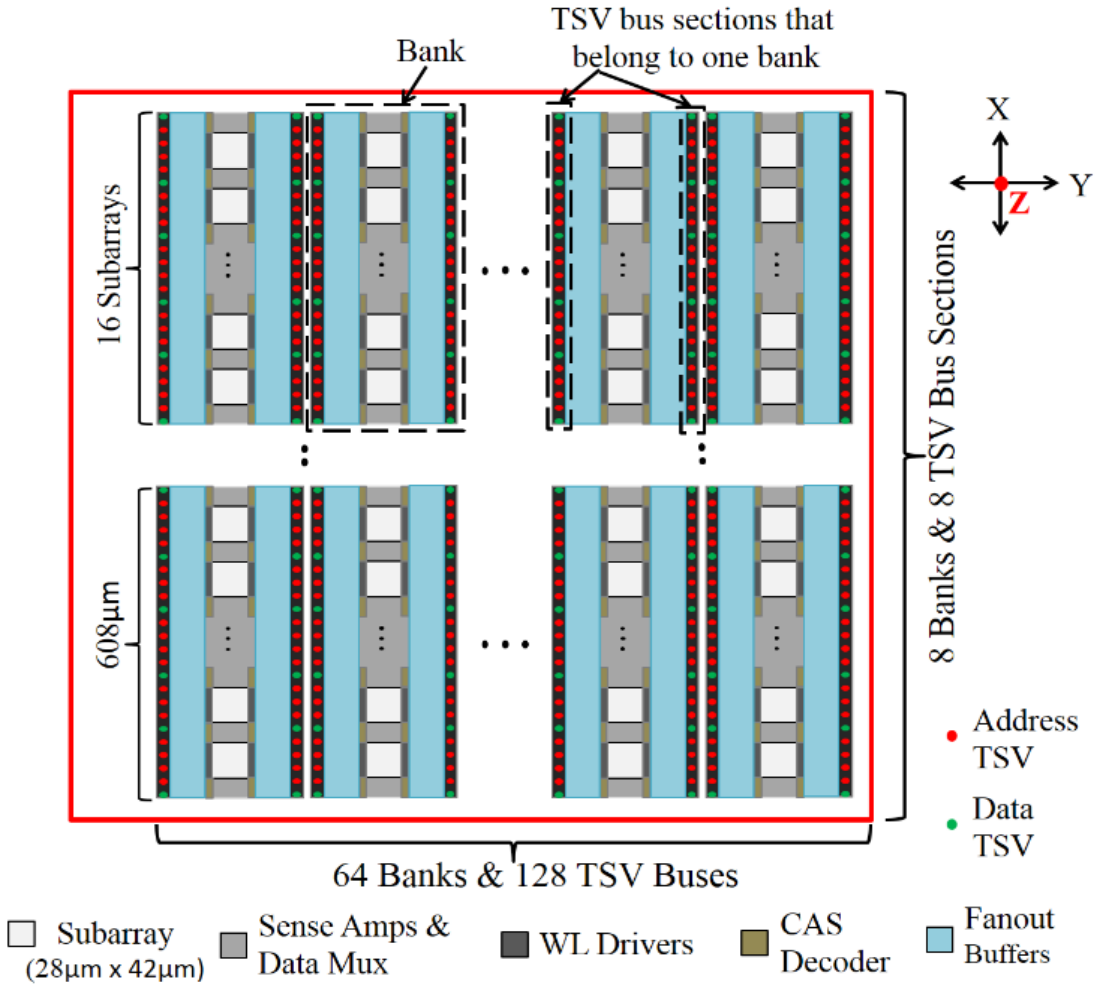


Figure 78: Schematic floorplan of on-die partition of a *3D-ProWiz* rank.

All 16 subarrays belonging to the on-die part of a bank are lined up along the X-direction (Figure 79), with number of subarrays along the Y-direction being one. Thus, a bank is 16 subarrays long, one subarray wide and 4 die layers high. In a rank, a total of 8 banks are arranged in the X-direction to form one column of banks. A total of 64 bank-columns are arranged in the

Y-direction, which is also the number of banks in a bank-row. Each bank-column has two TSV buses on two opposite sides of the bank. All TSV buses are laid out in the X-direction along the length of a bank-column. All TSVs in a TSV bus are arranged in a single row along the X-direction, thus each TSV bus is 8 banks long and one TSV wide. As shown in Figure 79, each TSV bus is logically partitioned into 8 sections. Each bank owns two TSV bus sections, which are located on two opposite sides along the X-direction and are responsible for routing address and data lines to the bank.

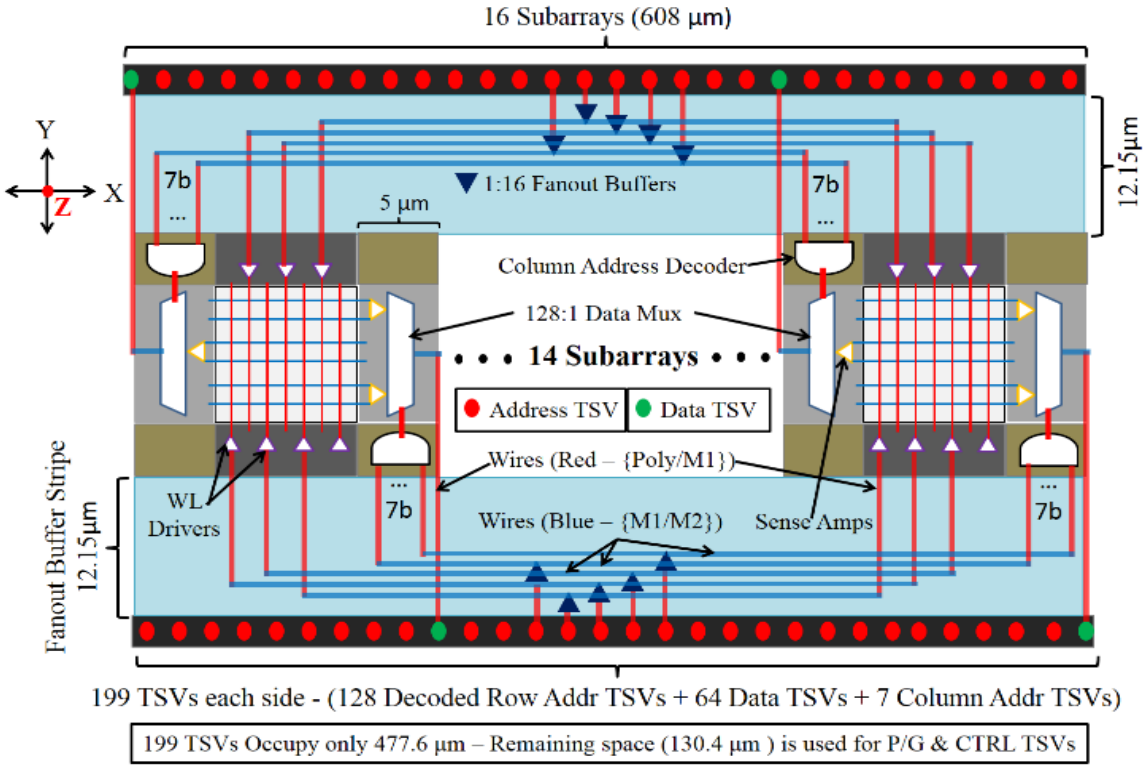


Figure 79: Schematic floorplan of a 3D-ProWiz bank and its TSV bus sections.

Figure 80 shows a detailed schematic floorplan of the on-die portion of a 3D-ProWiz bank and its TSV bus sections. The 3D-ProWiz floorplan differs from the 3D-Wiz floorplan in several aspects, which makes it more efficient and feasible. A 3D-ProWiz bank is of 4Mb size and is partitioned across all four DRAM layers. A 3D-Wiz bank consists of total 32 in-unison subarrays,

8 of which are on the same die, whereas a 3D-ProWiz bank consists of total 64 subarrays, 16 of which are on the same die. Therefore, we use 1:16 fanout buffers in a 3D-ProWiz bank instead of 1:8 fanout buffers used in the 3D-Wiz bank. Unlike the arrangement of TSVs in the 3D-Wiz bank, the TSVs in the 3D-ProWiz bank are arranged on both sides of the subarrays. The 3D-Wiz DRAM puts the decoders for column address on the logic die, whereas the 3D-ProWiz DRAM puts the decoders for column address on memory dies.

To facilitate efficient connections of TSVs to local word-line drivers on both sides of subarrays without using any metal-wire crossovers that overlap the subarrays, we arranged TSVs to be on both sides of a 3D-ProWiz bank. Also, to avoid metal-wire crossovers that would hinder connections of TSVs to upper layers, TSVs are arranged in a single row on each side of the bank. As discussed earlier, each in-unison subarray of a 3D-ProWiz bank serves 2 data bits on a read request, which sets the number of required decoded column address signals to 128. If we assume that the decoders for row ad-dress and column address are located on the logic die and decoded address signals are routed to individual subarrays through TSVs, then a 3D-ProWiz bank would have 256 row address TSVs and 128 column address TSVs along with 128 data TSVs (total 512 TSVs). As mentioned earlier and as shown in Figure 80, the length of a 3D-ProWiz subarray along the X-direction is 38μm (including 5μm for peripherals on each side), which makes the length of the 3D-Wiz bank to be 608μm as there are 16 subarrays in the bank along the X-direction. TSVs are arranged on both sides of a 3D-ProWiz bank, which implies that about half of 512 total TSVs (256 TSVs) have to fit along the 608μm long TSV bus on each side of the bank. A 608μm long TSV bus can fit 256 TSVs in a single row at TSV pitch of 2.375μm, which is 0.025μm tighter pitch than the standard TSV pitch of 2.4μm. But, this arrangement does not leave any room for control TSVs, P/G TSVs and any redundant TSVs to cope with TSV failures.

As mentioned earlier, a 3D-ProWiz bank has 128 decoded signals for column address. So, putting the column ad-dress decoders on the logic die would require 128 decoded column address lines to be routed to the small area of sense amplifiers and data mux on both sides of each subarray (as shown in Figure 80) through TSVs using 1:16 fanout buffers and metal wires. Semi-global metal wires have 0.18μm pitch at 45 nm following a conservative method of area projection [210], which implies that 128 column address wires would occupy 23μm width in the area of sense amplifiers and data mux. But, the width of the area of sense amplifiers and data mux along the X-axis is just 5μm (Figure 80). Thus, accommodating 128 column address wires in the small area of sense amplifiers and data mux is not possible.

Alternatively, we propose to locate the decoders for column address on memory dies along with sense amplifiers and data mux, as shown in Figure 80. Doing so reduces the number of TSVs required to route the column address to 7. We propose to use 7 column address TSVs on each side of the bank. As shown in Figure 80, a 7-bit address is fed to the column address decoders near each of the 16 subarrays through TSVs using 1:16 fanout buffers and metal wires. The 7-bit address is decoded by the decoder and used to select 1-bit data path through each data mux. Each subarray has one data mux on each side along the X-direction, therefore each subarray has a data I/O width of 2-bits.

An efficient floorplan for the 3D-ProWiz bank, as shown in Figure 80, renders many benefits. First of all, use of just 14 column address TSVs per bank (7 TSVs on each side) now means that only 199 TSVs (128 row address TSVs, 64 data TSVs and 7 column address TSVs), instead of 256 TSVs, have to be fit in 608μm long TSV bus on each side of the bank. Arranging 199 TSVs in one row occupies only 477.6μm of the total length at TSV pitch of 2.4μm. The remaining 130.4μm on each side can be used for P/G TSVs, control TSVs and redundant TSVs.

Moreover, the efficient arrangement of fanout buffers and their connections to TSVs and wordline drivers does not leave behind any met-al-wire crossovers that would hinder connections of TSVs to upper layers. Thus, our proposed floorplan of 3D-ProWiz bank adheres to physical design rules and renders efficient design, the implementation of which is practically possible. Note that the fanout buffer stripe (fanout buffers and their metal-wire connections to wordline drivers) occupies 24.3μm×608μm area per 3D-ProWiz bank, which is not insignificant. But, the greater benefits in performance, power and energy-efficiency obtained with the new floorplan of a 3D-ProWiz bank outweighs this area overhead. A detailed analysis of area, timings and energy for 3D-ProWiz DRAM is given in Section 10.4.

10.4. 3D-PROWIZ AREA, TIMING, AND ENERGY ANALYSIS

The area, timing, and energy analysis for the 3D-ProWiz architecture was performed by modeling the architecture using CACTI-3DD [205]. A similar analysis was conducted for other well-known 3D DRAM architectures such as the 3D stacked photonic DRAM (3DSPDRAM) [142], 3D DRAM from Samsung (3DSams) [198], and the hybrid memory cube (HMC) from Micron [23]. The results of the analysis for these architectures were compared with results for 3D-ProWiz and commodity DDR3 DRAM architectures. The models of the aforementioned 3D DRAM architectures were implemented in CACTI-3DD using the technology parameters for the 45nm node. The 3D DRAM architectures from prior work were chosen to provide a broad coverage of the full spectrum of 3D DRAM designs. For instance, the 3D DRAM architecture from Samsung realizes coarse-grained rank level stacking [198], [205], whereas 3DSPDRAM implements single subarray access (SSA) for a coarse-grained rank-level 3D data array organization [142], [205]. On the other hand, HMC from Micron is designed to exploit fine-grained rank-level partitioning [23], [205]. Even though the 3D-Wiz architecture [164] is infeasible to implement due to its

unawareness to the inter-component inter-connects at the subarray level, we compare our proposed 3D-ProWiz architecture with 3D-Wiz to highlight the micro-architectural differences between the two architectures in terms of the area, delay and energy values of the corresponding DRAM subsystems.

Table 13 lists the architectural parameters used in CACTI-3DD to model the DRAM designs. For a fair comparison between different architectures, we kept the memory capacity and page size of our memory models constant across all architectures. Table 13 also lists different timing parameters and per access energy/power values obtained from CACTI-3DD based models for the DRAM designs considered in our analysis. 3D-ProWiz demonstrates access latency of 17.3ns and row cycle time of 23ns, which are better than other architectures on average by about 49% and 57.2% respectively. 3D-ProWiz also yields activation precharge energy of 2nJ per access, which is better than other architectures on average by 23.6%.

We modeled DDR3 [202], LPDDR3 [203], and differential serial interfaces [204], which are used for DDR3, 3DSams and HMC respectively. These interfaces were modeled using the CACTI-IO tool [211]. We use a photonic interface for 3DSPDRAM [142]. We modeled the photonic interface of 3DSPDRAM and 3D-Wiz as well as our proposed photonic interface for 3D-ProWiz using the DSENT tool [75] (see Section 10.6 for details). As explained in [198], 3DSams uses a low power DDR3 interface where power-hungry redundant circuits including delay-locked-loop (DLL), input buffers, and clock circuitry are eliminated in the slave chips of the 3DSams chip-stack. This design very closely resembles the LPDDR3 interface, which is a low power DDR3 interface without DLLs and input buffers. So, we use the LPDDR3 interface for 3DSams in our studies. The refresh cycle time (tRFC) values given in Table 13 are calculated as eight times row cycle time (tRC) and additional recovery time (tREC) of 10ns. The values of tTAW constraint

used in our study are given in Table 13 as well. A detailed analysis of the tTAW constraint is presented in Section 10.5.

Table 13: Modeling parameters and timing, energy and power values for various DRAM architectures.

	DDR3	3D-Sams	3D-SPDRAM	HMC	3D-ProWiz	3D-Wiz
#ranks	4	4	4	4	4	4
#banks/rank	8	8	8	16	512	1024
#bitlines	1024	1024	1024	512	256	256
#wordlines	512	512	512	512	256	256
Page size	16Kb	16Kb	16Kb	16Kb	16Kb	8Kb
Bus width	64b	32b	8b	32b	128b	128b
Timing parameters (ns) [DoI: Delay of Interface]						
tRAS	36.7	33.5	33.5	32.1	17.3	20.2
tRC	54.2	51	63	48	23	25.1
tTAW	33	51	4	49	16	15
tRFC	443.4	418	513.6	382.6	193.8	210.8
DoI	4	4	16	4	1	2
Per access energy values (nJ) [EPA: Energy per Access]						
ActPre E	3.5	3.5	0.3	3.5	2	1.4
Read E	1.9	1.4	0.4	1.1	1.8	2.1
Interface E	12.8	7.7	0.25	7.4	0.25	0.25
EPA	18.1	12.5	0.88	11.9	4.05	3.75
Power values (mW) [BG: Background]						
BG Power	2197.8	117.4	455	1226.8	455	455
Refresh	252.6	267.9	190.8	292.7	330.2	212.5

We present detailed breakdowns of delay per access values (obtained from CACTI-3DD, DSENT and CACTI-IO based models) for various DRAMs in Figure 80. Our goal is to understand how the contribution of some critical architectural components differ across various DRAMs in defining delay per access and energy per access values for these DRAMs. The use of TSV-based internal memory buses reduces the lengths of address and data paths for the 3D-stacked DRAMs such as 3DSams, 3DSPDRAM and HMC compared to DDR3. This results in smaller values of I/O delay and sensing delay for the 3D-stacked DRAMs than DDR3. As shown in Table 13, HMC has

512 bitlines per subarray compared to 1024 bitlines per subarray in DDR3, 3DSams, and 3DSPDRAM. This results in smaller values of sensing and precharge delays for HMC compared to DDR3, 3DSams and 3DSPDRAM. In the single subarray access (SSA) architecture of 3DSPDRAM, the entire cache line is serially accessed from only one subarray through an 8-bit wide bank bus, which increases the serialization latency of the 3DSPDRAM interface. This results in larger interface delay for 3DSPDRAM than 3DSams. The use of TSVs at subarray-level granularity further reduces the lengths of address and data paths for 3D-ProWiz compared to other 3D-stacked DRAMs. The effect of this, combined with its smaller subarrays (256 × 256), leads to even smaller values of sensing delay, precharge delay, and restore delay for 3D-ProWiz. Also, the wider data bus (128-bit) in 3D-ProWiz reduces burst length to 2 cycles, lowering interface delay compared to all other DRAMs.

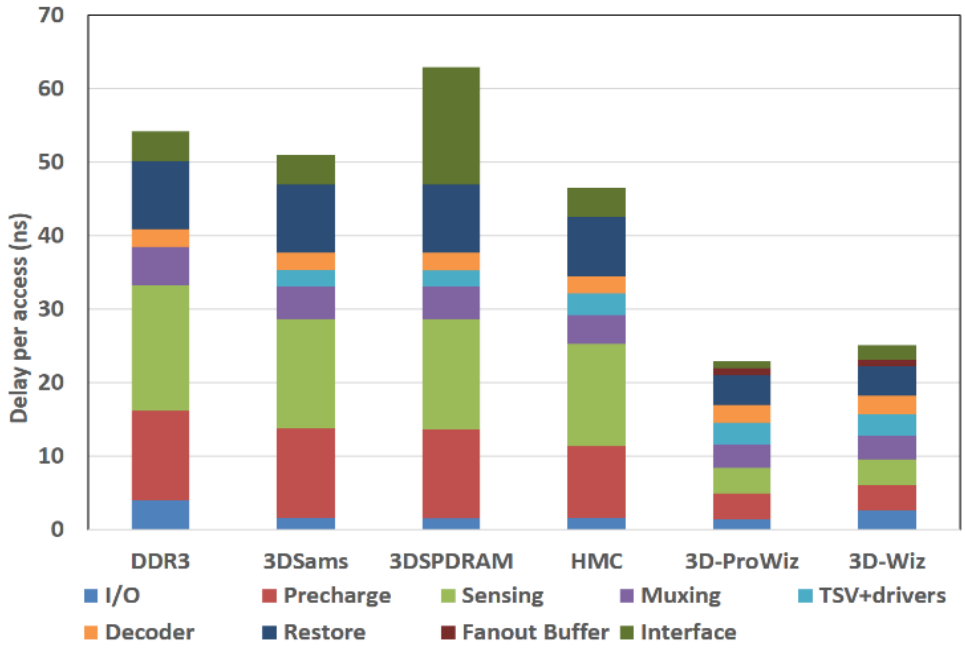


Figure 80: Breakdown of delay per access for various DRAMs.

Figure 81 presents detailed breakdowns of energy per access (obtained from CACTI-3DD, DSENT and CACTI-IO based models) for the various DRAMs. The use of a pho-tonic interface reduces dynamic interface energy for 3DSPDRAM and 3D-ProWiz compared to all other DRAMs. In spite of a similar bank size and subarray organization, 3DSPDRAM has lower energy than 3DSams, because 3DSPDRAM implements a single subarray access (SSA) [142]. In SSA, an entire cache line is served by a single subarray [142], which reduces the granularity of activation for 3DSPDRAM compared to 3DSams, reducing the activation-precharge energy for 3DSPDRAM. The combined effect of smaller subarrays and shorter address and data paths yields smaller values of sensing energy, precharge energy, and restore energy for 3D-ProWiz than DDR3, HMC and 3DSams.

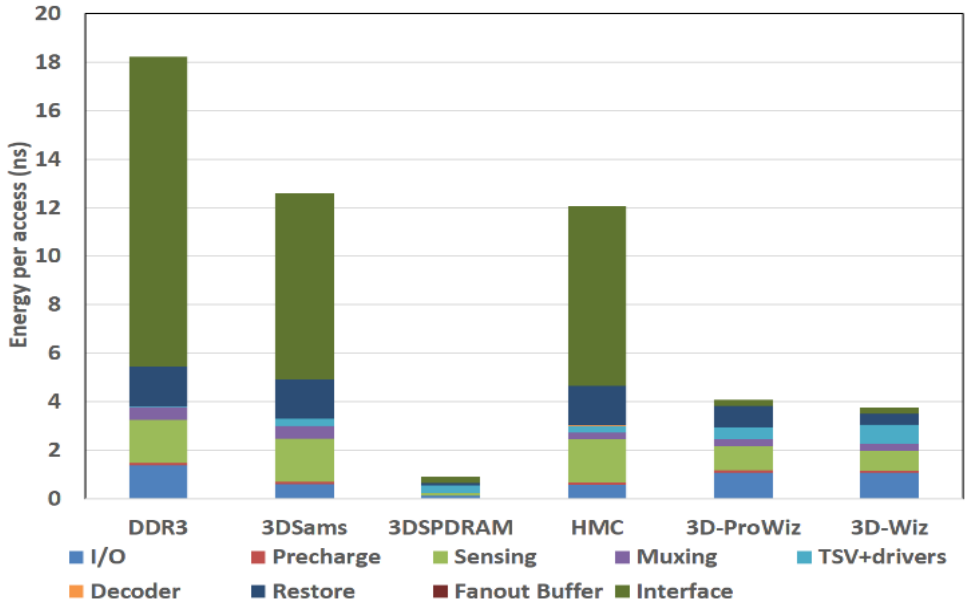


Figure 81: Breakdown of energy per access for various DRAMs.

The cost of a DRAM module depends on the area efficiency of the DRAM die. Therefore, area efficiency is an important parameter to consider while designing a new DRAM architecture. Area efficiency of a DRAM die is de-fined as the percentage of the total die area that corresponds

to the DRAM cell area. The total DRAM die area includes the area covered by peripherals, memory bus, and TSVs in addition to area covered by DRAM cells. A DRAM architecture with a high area efficiency is thus able to store more useful information compared to a lower area efficiency DRAM architecture, within the same die area.

Table 14: Breakdown of total die area for various DRAM architectures. All area values are in mm2. (AE= Area Efficiency, FoB= Fanout Buffer, M bus= Memory data bus, WL=Wordline, C. Dec=on-die column address decoders).

	DDR3	3D-Sams	3D-SPDRAM	HMC	3D-Wiz	3D-ProWiz	
Decoders	0.03	0.03	0.03	0.03	0.55	0.17	**Logic Die**
M bus	0.52	1.41	9.1	1.44	8.7	4.67	
TSV driver	0	0.02	0.01	0.06	1.5	1.45	
Interface	6.8	3.8	2.3	4.5	2.3	2.3	
Bitcells	9.39	9.39	9.39	9.39	9.39	9.39	
WL drivers	2.1	2.1	2.1	4.2	8.4	8.4	
Sense Amp	2.5	2.5	2.5	2.02	3.5	3.5	
I/O Mux	0.01	0.02	0.08	0.01	0.25	0.52	
TSVs	0	0.009	0.006	0.019	1.5	0.925	
FoB + C. Dec	0	0	0	0	6	7.56+0.85	
Total Area	21.38	19.3	25.54	21.66	29.04	31.12	
AE %	43.9	48.7	36.8	43.3	32.3	30.1	

Table 14 shows detailed breakdowns of DRAM die area for various DRAMs. All TSVs in this study were modeled based on ITRS projections for intermediate interconnect level TSVs [31]. We assume the pitch of semi-global metal wires to be 0.09μm following a more aggressive method of area projection than in [210]. Accordingly, fanout buffer stripes occupy 24.3μm×608μm area per 3D-ProWiz bank, which translates to a 7.56mm^2 of fanout buffer area per 3D-ProWiz die. The area covered by the on-die column address decoders was found to be 0.85mm^2. The area covered by electrical interfaces of DDR3, 3DSams and HMC and the area of related global I/O circuits was calculated using CACTI-IO [211]. This area is listed in Table 14 as interface area. The area covered

by the photonic interfaces of 3DSPDRAM and 3D-ProWiz were estimated using DSENT [75]. The photonic interface configuration is explained in Section 10.6.

The smaller subarray size (256 ×256) in 3D-ProWiz results in larger areas of wordline drivers and sense amplifiers per DRAM die, compared to all other DRAMs, as smaller size of subarray implies more number of subarrays for a given capacity. Similarly, for HMC, the smaller subarray size (512 ×512) results in larger area of wordline drivers and sense amplifiers than DDR3, 3DSams, and 3DSPDRAM. Due to the combined effect of larger wordline driver area, larger sense amplifier area, and fanout buffer area, 3D-ProWiz DRAM has 9.16mm^2 (41.7%) more die area than the average die area of all other DRAMs. So, the 3D-ProWiz architecture may prove to be relatively costly due to its larger die area. Nonetheless, the significant improvements in access time and energy consumption make 3D-ProWiz a promising architecture candidate for future 3D DRAMs. The higher cost of silicon real estate increases the unit cost of a 3D-ProWiz device. The unit cost of the memory device can be reduced either by reducing the required die area or by reducing the engineering design cost. The layout of a 3D-ProWiz bank (as shown in Figure 79) is designed using conservative and scalable design rules, which often lead to an oversimplified layout design resulting in larger than required die area as discussed in [209]. The required die area and unit cost of 3D-ProWiz DRAM can be reduced by using more aggressive and process-specific design rules to design the layout of 3D-ProWiz DRAM die. Moreover, the structure of 3D-ProWiz is fairly repetitive at the bank level, which re-duces the complexity of the mask-set designs and fabrication process. This in turn reduces the engineering design cost and consequently unit cost of 3D-ProWiz DRAM even further.

The scalability of the 3D-ProWiz bank-size is limited only by the non-scalable size of TSVs, as the size of all other components of the bank except TSVs scale proportionally to the scaled

technology node. As implied from Figure 79, the length covered by the TSVs controls the length of the bank, which suggests that the non-scalability of TSVs affects the bank-length more than the bank-width. This results in inefficient use of available die area, which harms the area efficiency and cost of the DRAM. To cope with this limitation, the number of subarrays along the X-direction in a bank and the fanout strength of the fanout buffers can be changed with the scaling of technology node so that the subarrays and fanout buffers cover the same length as covered by the TSVs.

Table 15: Sixteen different combinations of memory controller policies. (OP=Open Page, CP=Close Page, RBRR=Rank then Bank Round Robin, FCFS=First Come First Served).

	Address Mapping Scheme	Scheduling Policy	Page Mode Policy
RBRR_CP_AMS1	ch:rank:row:col:bank	RBRR	CP
RBRR_CP_AMS3	ch:rank:bank:col:row	RBRR	CP
RBRR_CP_AMS4	ch:rank:bank:row:col	RBRR	CP
RBRR_CP_AMS5	ch:row:col:rank:bank	RBRR	CP
FCFS_CP_AMS1	ch:rank:row:col:bank	FCFS	CP
FCFS _CP_AMS3	ch:rank:bank:col:row	FCFS	CP
FCFS _CP_AMS4	ch:rank:bank:row:col	FCFS	CP
FCFS _CP_AMS5	ch:row:col:rank:bank	FCFS	CP
RBRR_OP_AMS1	ch:rank:row:col:bank	RBRR	OP
RBRR_OP_AMS3	ch:rank:bank:col:row	RBRR	OP
RBRR_OP_AMS4	ch:rank:bank:row:col	RBRR	OP
RBRR_OP_AMS5	ch:row:col:rank:bank	RBRR	OP
FCFS _OP_AMS1	ch:rank:row:col:bank	FCFS	OP
FCFS _OP_AMS3	ch:rank:bank:col:row	FCFS	OP
FCFS _OP_AMS4	ch:rank:bank:row:col	FCFS	OP
FCFS _OP_AMS5	ch:row:col:rank:bank	FCFS	OP

10.5. SENSITIVITY ANALYSIS

In this section, we analyze the sensitivity of 3D-ProWiz DRAM to different memory controller policies in order to determine a combination of policies that yield the best performance and energy-efficiency for 3D-ProWiz. We also analyze the sensitivity of the average latency for

3D-ProWiz and other DRAM architectures to different values of the tTAW constraint to better understand the relation-ship between tTAW constraint, bank-level memory parallel-ism, and average latency.

10.5.1. SENSITIVITY TO MEMORY CONTROLLER POLICIES

The performance and energy-efficiency of a DRAM sub-system not only depends on the DRAM data organization, but also on a number of memory controller policies such as address mapping policy, page mode policy, and scheduling policy.

In modern DRAM devices, the arrays of sense amplifiers can also act as buffers (known as row buffer) that provide temporary data storage [208]. In a DRAM controller that implements the open-page policy, once a row of data is activated in a bank, it is temporarily stored in the row buffer. A row of data, which is temporarily stored in the row buffer, is referred to as a page. In case of open-page policy, spatially and temporally adjacent memory accesses to different columns of the same row/page can be made again with minimal latency, because the row/page is already active in the row buffer. In contrast to the open-page policy, the close-page policy flushes the row buffer after an access finishes. This policy is designed to favor accesses to random locations in memory and optimally supports memory request patterns with low degrees of access locality.

The task of an address mapping scheme is to minimize the probability of bank conflicts in temporally adjacent requests and maximize the parallelism in the memory system. To obtain the best performance, the choice of the ad-dress mapping scheme is often coupled to the page mode policy of the memory controller. To facilitate the pipelined execution of DRAM commands, the DRAM memory controller uses a scheduling policy that prioritizes DRAM commands based on many different factors, including, but not limited to, the age of the request, the priority of the re-

quest, the bank address of the request, the availability of resources to a given request. The reader is directed to [208] for more information on various memory controller policies.

A modern memory controller uses one of the following six address mapping schemes (AMS) [208], [212]: (1) AMS1 – ch:rank:row:col:bank, (2) AMS2 – ch:row:col:bank:rank, (3) AMS3 – ch:rank:bank:col:row, (4) AMS4 – ch:rank:bank:row:col, (5) AMS5 – ch:row:col:rank:bank, and (6) AMS6 – row:col:rank:bank:ch. Depending on the address mapping policy, the physical address is resolved into indices in terms of channel ID (ch), rank ID (rank), bank ID (bank), row ID (row) and column ID (col). The address mapping scheme AMS2 is used to maximize the rank level parallelism in the memory system, whereas the scheme AMS6 is used to maximize the channel level parallelism. All the memory systems investigated in this chapter are single channel systems with a limited number of ranks (4 ranks/channel). Therefore, the rank level and channel level parallelism that can be leveraged by these systems would be insignificant. Because of this reason, we disregard the address mapping schemes AMS2 and AMS6, and select the remaining four types of address mapping schemes (AMS1, AMS3, AMS4, and AMS5) for our study.

We chose the following two kinds of scheduling policies: (i) Rank-then-Bank-Round-Robin (RBRR) and (ii) First-Come-First-Serve (FCFS). We also analyze the following two-page mode policies: (i) Open page (OP) and (ii) close page (CP). We analyzed the sensitivity of 3D-ProWiz DRAM to 16 different combinations of these policies which are listed in Table 15. We performed trace-driven simulation analysis for PARSEC benchmarks using a cycle-accurate DRAM simulator DRAMSim2 [212]. We evaluated energy-delay product (EDP) values averaged over twelve PARSEC benchmarks [76] for all sixteen combinations of policies. The simulation method and environment are described in Section 10.7 in more detail.

Figure 82 plots EDP values averaged over twelve PARSEC benchmarks across the 16 combinations of policies. Each column in the figure gives the EDP value averaged over twelve PARSEC benchmarks for the respective combination of policies. The error bars on each column present standard deviation in the EDP value across the bench-marks. EDP values were calculated by multiplying the average-energy with average latency for memory accesses. Average-energy in each benchmark was calculated by dividing the total energy by total number of transactions. Average-latency was calculated by dividing total latency by the total number of transactions. As evident from the figure, the policy combination RBRR_OP_AMS1 yields the least value of EDP among all policy combinations.

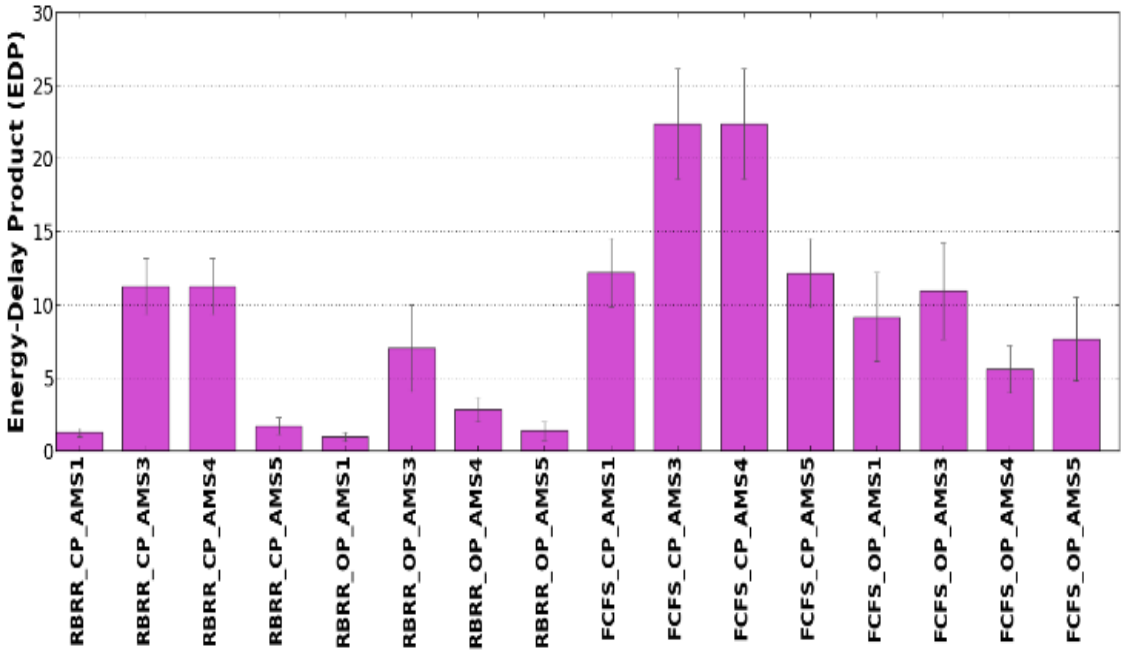


Figure 82: Energy-delay product (EDP) values averaged over the PARSEC benchmarks for sixteen combinations of policies.

The policy combinations with AMS1 and AMS5 yield less EDP than policy combinations with AMS3 and AMS4. This is because address mapping schemes AMS1 and AMS5 present more opportunity to exploit bank-level parallelism, as in these schemes consecutive memory addresses

are mapped to different banks. Increased exploitation of bank-level parallelism results in smaller value of average latency and higher throughput, which in turn results in smaller EDP. Moreover, the policy combinations with RBRR scheduling in general outperform the policy combinations with FCFS scheduling. This is because memory re-quests face more queuing delay in the case of FCFS scheduling compared to RBRR scheduling. The policy combinations with AMS4 and OP schemes yield smaller EDP in general over the policy combinations with AMS5 and OP schemes. This is because the open page (OP) scheme pro-vides higher row buffer hit rate in the case of the AMS4 address scheme than in the case of the AMS5 address scheme.

In summary, as the policy combination RBRR_OP_AMS1 yields the least EDP among all policy combinations, we propose to use this policy combination for 3D-ProWiz DRAM. We did a similar analysis for DDR3, 3DSams, 3DSPDRAM, and HMC as well, and found that policy combination RBRR_OP_AMS1 provides the best EDP values for all of these DRAMs as well. Therefore, we use the policy combination RBRR_OP_AMS1 for all the analysis presented in this chapter.

10.5.2. SENSITIVITY TO TTAW CONSTRAINT

In this subsection, we first estimate reasonable values of the tTAW constraint for all DRAM architectures being studied, before analyzing the sensitivity of average latency for 3D-ProWiz and other DRAM architectures to different values of the tTAW constraint. To prevent the adverse effects of power delivery network (PDN) noise in commodity 2D DRAMs, the number of activates to different banks in a rolling window of time are limited by the tFAW (four-bank activation window) constraint. In general, the tFAW constraint allows only four bank activates in a rolling window of tFAW time. However, the PDN noise issue is more challenging in 3D DRAMs [198], [213], which makes the tFAW constraint more critical. Therefore, the designers of 3D-stacked

DRAMs have conservatively allowed only two bank activates in a rolling window of tTAW time [24], [200]. This new constraint is called two-bank activation window (tTAW).

The tTAW constraint for a given 3D-DRAM module can be estimated if we know the maximum current that can be drawn from the PDN without violating the PDN noise limit. This is because the noise level of a PDN depends on the load current drawn from the PDN and physical locality of consecutive bank activates [198]. The closer the physical locations of banks that are consecutively activated, the more is the PDN noise level. An 8GB DIMM of DDR4x2400 [290] can draw a maximum of 1.8A current from its PDN without violating the noise limits. To reduce the noise level in the PDN for 3D-stacked DRAMs, we propose to use V_{DD}/V_{SS} edge TSV pads, as used in [198] and [213]. Accordingly, we allocate 10 V_{SS} and 5 V_{DD} pads per bankgroup of the 3D-ProWiz DRAM.

In 3D-stacked DRAMs, the PDN noise level also gets affected by the degree of bank-level concurrency, which is controlled by tTAW time. This implies that the PDN noise is actually controlled by tTAW time. Here, we present a brief quantitative analysis to show how the PDN noise level is affected by the degree of concurrency. It is implied from the discussion given in [213] that the maximum tolerable IR-drop noise in the PDN network is 75mV per V_{DD} pad. The resistance of an intermediate level TSV is 730mOhms as reported in [199], which sets the maximum allowable current draw of 100mA per V_{DD} pad to limit the IR-drop noise to 75mV per V_{DD} pad. Considering 5 V_{DD} pads per bankgroup, the assumed design of the PDN can deliver about 500mA peak current per bankgroup without violating the noise limit. In this case, the 2nJ activation-precharge energy and 23ns tRC obtained for a 3D-ProWiz bank (as shown in Table 13) indicates that concurrent refresh operations on 6 banks of the same bankgroup would draw about 522mA current ({(2nJ×6)/23ns}/1V = 522mA) from the V_{DD} pads, which would clearly violate

the IR-drop noise limit. Thus, it can be concluded from this discussion that the degree of concurrency, and hence the tTAW time should be optimized to keep the PDN noise to within allowable limits and ensure error-free memory operation. We have assumed nominal operating temperature for the preceding analysis. A detailed analysis of the effect of concurrency on temperature and related analysis of tTAW constraint are beyond the scope of this work.

Now, the V_{DD} at 45nm technology node is 1V, thus the peak current of 500mA corresponds to peak power of 500mW. Based on this information, we estimate reasonable values of the tTAW constraint for all 3D DRAMs considered in this study, for which the peak power consumption does not exceed 500mW. For that, we calculate total energy required to access two banks in parallel (2*energy per access) and divide it by 500mW. The resultant value gives the tTAW constraint in ns. The values of tTAW constraint for all 3D DRAMs given in Table 13 were estimated using this method. We take the tFAW constraint of 33ns from DDR3-1866 datasheet and use it as tTAW in our model of DDR3 memory for fair comparison of our DDR3 model with other 3D DRAMs that use tTAW instead of tFAW. As shown in Table 13, 3DSPDRAM yields tTAW of 3.6ns. Such a short tTAW time is justifiable for 3DSPDRAM, as 3DSPDRAM supports SSA architecture in which only one subarray of the target bank is activated while serving a memory request. So, owing to the reasonably large bank size for 3DSGDRAM (implied from the modeling parameters given in Table 13), consecutive activation of two different banks results in activation of two different subarrays that are physically quite far from each other, which in turn results in more stable operation for the 3DSPDRAM data array.

More accurate estimates of the tTAW constraint require detailed circuit-level simulation of the DRAM data array and power delivery network, which is beyond the scope of this chapter. But, we believe that it is important to have some insights about the relation of the tTAW constraint with

bank-level parallelism and performance. To investigate this relationship, we performed sensitivity analysis for the tTAW constraint by varying the tTAW constraint of all DRAMs in a range from 10ns to 120ns with step increase of 10ns. Thus, the tTAW constraint took 12 different values in the range. Then, we performed trace-driven simulation analysis for PARSEC benchmarks using a cycle accurate DRAM simulator DRAMSim2 [212]. We evaluated access latency values of all DRAM architectures averaged over twelve PARSEC benchmarks for all twelve values of tTAW constraint. The policy combination RBRR_OP_AMS1 was used for all simulations in this study. The simulation method and environment are described in Section 10.7 with more details.

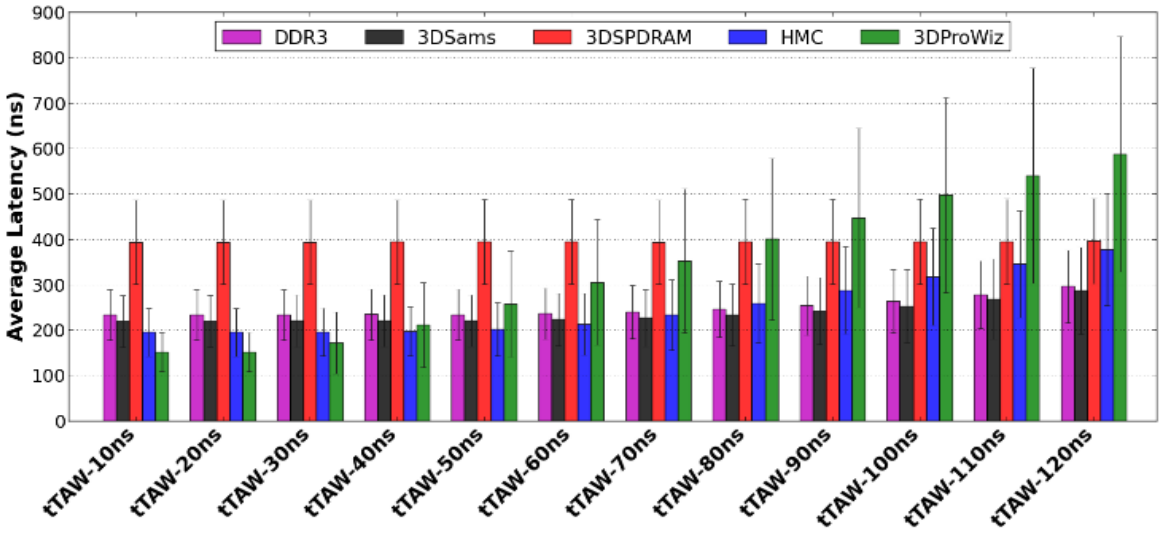


Figure 83: Access latency values averaged over the PARSEC benchmarks for twelve different tTAW values.

Figure 83 plots the average access latency values averaged over twelve PARSEC benchmarks across the twelve tTAW values. Each column in the figure shows access latency value averaged over twelve PARSEC benchmarks for the respective tTAW value of the corresponding DRAM. The error bars on each column present standard deviation in access latency across benchmarks. The access latency value for each benchmark was calculated by dividing total latency

by number of transactions. As evident from the figure, the average latency of 3D-ProWiz increases by 435ns (2.9× - to 585ns from 150ns) over 110ns increase in tTAW, which yields the sensitivity of 2.6%/ns. Based on similar interpretation of the figure for other DRAMs, the access latencies of 3D-ProWiz, HMC, 3DSPDRAM, 3DSams and DDR3 yield tTAW sensitivity values of 2.6%/ns, 0.8%/ns, 0.01%/ns, 0.2%/ns and 0.2%/ns respectively. 3DSams and DDR3 yield about the same amount of sensitivity to tTAW. From Table 13, it is observed that 3DSams and DDR3 have row cycle times (tRC) of 51ns and 54ns respectively. The *tRC* of DDR3 is about 5% larger than 3DSams. On the other hand, the tRC of 3DSPDRAM is 63ns, which is about 24% greater than the tRC of 3DSams. From Figure 83, it can be observed that 3DSPDRAM has low access latency sensitivity to tTAW values. Intuitively, these results imply that, for a DRAM subsystem, the sensitivity of access latency to tTAW is inversely proportional to the row cycle time (tRC) of the DRAM.

In contrast, the sensitivity to tTAW for HMC is much larger than 3DSams in spite of HMC having only 5% smaller value of row cycle time (48ns). However, HMC has more number of banks than 3DSams (see Table 13), which naturally increases the bank-level parallelism of HMC compared to 3DSams regardless of the tTAW constraint. Due to this reason, HMC access latency is much more sensitive to tTAW values than 3DSams. This implies that the sensitivity of access latency to tTAW is directly proportional to the number of banks in a DRAM. Moreover, among all the DRAM architectures under study, 3D-ProWiz has the smallest tRC of 23ns and the largest number of banks (512banks/rank), which results in the highest value of tTAW sensitivity for 3D-ProWiz. For a system that is relatively more sensitive to tTAW, even a small deviation from the optimal tTAW constraint (minimal value of tTAW) may result in much larger degradation of

performance. For such systems, it is very important to have as small tTAW value as possible, by designing a robust PDN that is less prone to noise.

In conclusion, we have shown that for a given DRAM subsystem, the sensitivity of access latency to tTAW constraint is directly proportional to the number of banks and inversely proportional to the row cycle time. Therefore, it is very important to design a robust PDN, which is less prone to noise and has more relaxed tTAW constraint, for a 3D DRAM subsystem with relatively small tRC and large bank count.

10.6. MODELING AND ANALYSIS OF HIGH BANDWIDTH PHOTONIC INTERFACE

In this section, we first discuss the maximum theoretical bandwidth achievable with 3D-ProWiz and justify the use of a high bandwidth photonic interface. Then we describe the functionality of the logic die that supports this interface. Lastly, we investigate the energy-efficiency of the proposed photonic interface in terms of energy-per-byte values for PARSEC benchmarks and compare it with several conventional electrical interfaces such as DDR3 [202], LPDDR3 [203], Wide-I/O [24] and differential serial interface [204].

10.6.1. BANDWIDTH ANALYSIS

As discussed earlier, bank-level parallelism, which is generally limited by the tTAW constraint and the shared internal memory bus together, is limited by only the tTAW constraint in 3D-ProWiz DRAM. This allows 3D-ProWiz to support significantly higher bandwidth than other architectures. To utilize this bandwidth, the memory-to-core inter-face should support at least a two-stage deep pipeline for memory requests with average issue rate of 2 requests per 16ns (tTAW for 3D-ProWiz is 16ns).

This organization enables a 3D-ProWiz module to achieve a peak data rate of 512×4 bits/16ns (2 cachelines/rank/16ns) assuming a 256-bit cacheline, if the memory controller is designed to send requests to all four ranks in parallel. This is equal to a 128Gbps peak bandwidth at 1GHz clock rate. This bandwidth analysis assumes close page policy. But, for an open page policy and relatively high row-buffer hit rate, a 3D-ProWiz module can be reasonably assumed to deliver a peak data rate of 512×4 bits/2ns (Burst Length=2ns), which is equal to 1024Gbps peak bandwidth. A conventional electrical bus based inter-face of the DDRx family [202], [203] cannot support such a high bandwidth due to pin-bandwidth limitations. To address this issue, the Hybrid Memory Cube (HMC) from Micron proposes the use of high speed serial links at the interface [23]. The high-speed link in HMC consists of several differential lanes as the fundamental building blocks. Each differential lane is claimed to achieve the maximum data transfer rate of 10Gbps. Thus, 3D-ProWiz would require about 103 such differential lanes to achieve a 1024Gbps data transfer rate across the interface. This is a prohibitively large number of lanes that would cause serious packaging issues due to pin-limitations.

Alternatively, several recent works on DRAM architectures have proposed dense wavelength division multiplexed (DWDM) photonic interfaces to achieve higher bandwidths [142], [214]. A DWDM optical fiber can carry approximately 64 wavelengths, which creates 64 channels on a single fiber [81]. A typical DWDM fiber link consists of several components such as ring modulators, photonic waveguides, Ser-Des (serialization/deserialization) components, and photo detectors, which work together in sync to achieve high speed and high bandwidth data transfers [81], [87]. The resulting bandwidth for a DWDM fiber link depends on the bandwidth of the slowest individual component. In recent years, innovations in Si-photonics technology have enabled on-chip Si waveguides, ring modulators, and ring detectors to function at 10-20Gbps data

rate [73]. Also, advancements in CMOS technology have enabled a single lane SerDes to operate at 25Gbps data rate [215]. Thus, single wavelength channels within fiber links can today operate at 10Gbps data rate. In this study, we conservatively operate single wavelength channels within fiber links at 5Gbps, which corresponds to a 320Gbps bandwidth per DWDM fiber. Consequently, the bandwidth requirement of 3D-ProWiz can be fulfilled by just 4 such DWDM fibers (~1280Gbps), which is easily achievable well within the pin-constraints. Therefore, we propose utilizing such a high bandwidth photonic interface for 3D-ProWiz.

The photonic interface in 3D-ProWiz is comprised of two links (as shown in Figure 74): one for reads and the other for writes. Each link consists of 4 unidirectional DWDM fibers. Each DWDM fiber in read and write links supports 64 wavelengths, making the read and write bus to be 256-bits wide each. The photonic interface also uses one additional link consisting of a single DWDM fiber for transmitting address and control signals. The address/control fiber supports a total of 40 wavelengths, with 32 wavelengths for addresses and 8 wavelengths for control signaling.

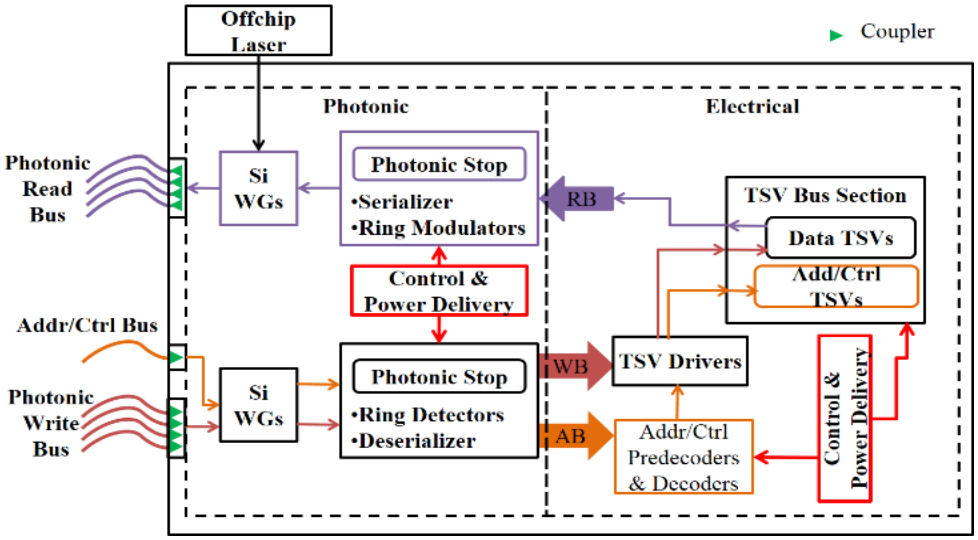


Figure 84: Functional block diagram of the logic die in 3D-ProWiz.

10.6.2. DESIGN AND FUNCTIONING OF LOGIC DIE

Figure 84 shows the functional block diagram of the logic die for 3D-ProWiz. The logic die is functionally divided into two parts: photonic and electrical. The photonic part of the logic die performs all of the optical to electrical (O-to-E) and electrical to optical (E-to-O) conversions; and optically interfaces with the memory controller on the processor chip. As shown in the figure, the photonic portion of the logic die consists of on-chip Si waveguides (Si WGs), photonic couplers, and photonic stops. The electrical part of the logic die consists of address/control decoder logic, TSV drivers, read/write buffers, and the control and power delivery net-work. It interfaces with the DRAM cell array through TSV buses. The functionality and role of all of these components on the logic die is discussed next.

Table 16: Dynamic energy, static power, losses and delay for photonic interface components [142] [75].

	Dynamic Energy	Static Power	Losses (dB)	Delay (ns)
Modulator	47 fJ/bit	250μW/ring	1	0.5
Detector	44 fJ/bit	250μW/ring	1.2	0.5
Serializer	950 fJ	1.2 mW	-	0.6
De-serializer	870 fJ	1 mW	-	0.6
Average Laser Power (mW)				
Read/Write Laser Power	131mW			
Address/Ctrl Laser Power	41mW			

For each transaction, firstly the incoming bit stream (originating at the processor side memory controller) from the photonic address/control bus is coupled to the on-chip Si waveguides by couplers. These photonic couplers are primarily responsible for realizing a low loss coupling between on-chip Si waveguides and off-chip DWDM fibers. Next, the bit stream is passed through a photonic stop, where the constituent ring detectors convert the photonic bit stream into an electrical bit stream. The serialized electrical bit stream is then de-serialized before it electrically drives the address/control bus (AB). The signals of this AB are decoded before they are used to

drive an appropriate section of one of the TSV buses. The decoded address/control signals activate an appropriate bank to serve the subsequent read or write request. For a write request, the data is written to the requested bank through the TSV bus section.

Table 17: Modeling parameters for interfaces. (AF=Activity Factor, λ=wavelength).

	DDR3	LP-DDR3	Differential	Wide-IO	Photonic
#DQ pins/λs	64	32	32	128	256
#DQS pins/λs	8	4	4	16	-
#CA pins/λs	32	32	32	32	32
Frequency	1GHz	1GHz	1GHz	800MHz	1GHz
AF DQ	1	1	1	1	1
AF CA	0.5	0.5	0.5	1	1
Energy/256b (nJ)	12.78	7.67	7.43	2.12	0.25
Static Power (mW)					
Clock	60.75	47.25	47.25	10	20
PHY	30	30	10	1	-
Termination & Bias	2107.8	1149.6	48.8	2.3	-
Total static power of photonic interface					455

10.6.3. MODELING OF INTERFACES AND ENERGY-EFFICIENCY ANALYSIS

We modeled our proposed photonic interface using the DSENT [75] tool. The timing, energy, and power values for the interface were extracted from DSENT [75], and are given in Table 16 and Table 17. In Table 16, the static power for modulators and detectors represents thermal trimming power, whereas it represents leakage power for the rest of the components. We also modeled conventional electrical interfaces such as DDR3 [202], LPDDR3 [203], Wide-I/O [24] and differential interface [204] using CACTI-IO [211]. The configuration parameters used to model the electrical inter-faces are given in Table 17. We obtained dynamic energy and static power values of all electrical interfaces from CACTI-IO based simulations, which are given in Table 17. We used the timing, energy and leakage power values given in Table 16 and Table 17 to model DDR3, LPDDR3, differential, Wide-I/O and photonic interfaces in DRAMSim2 [212].

We used the timing, energy and leakage power values given in Table 13 to model the 3D-ProWiz core in DRAMSim2.

We integrated the 3D-ProWiz core model with interface models and simulated the following five 3D-ProWiz DRAM subsystems: (1) 3D-ProWiz core with DDR3 interface, (2) 3D-ProWiz core with LPDDR3 interface, (3) 3D-ProWiz core with differential interface, (4) 3D-ProWiz core with Wide-I/O interface, and (5) 3D-ProWiz core with photonic interface. We performed trace-driven simulation analysis for PARSEC benchmarks using DRAMSim2 models of all five aforementioned DRAM subsystems. From DRAMSim2, we obtained energy-per-byte values consumed by the interfaces. In this subsection, we concentrate on investigating only the interfaces, so we do not report the energy-per-byte values consumed by the 3D-ProWiz core.

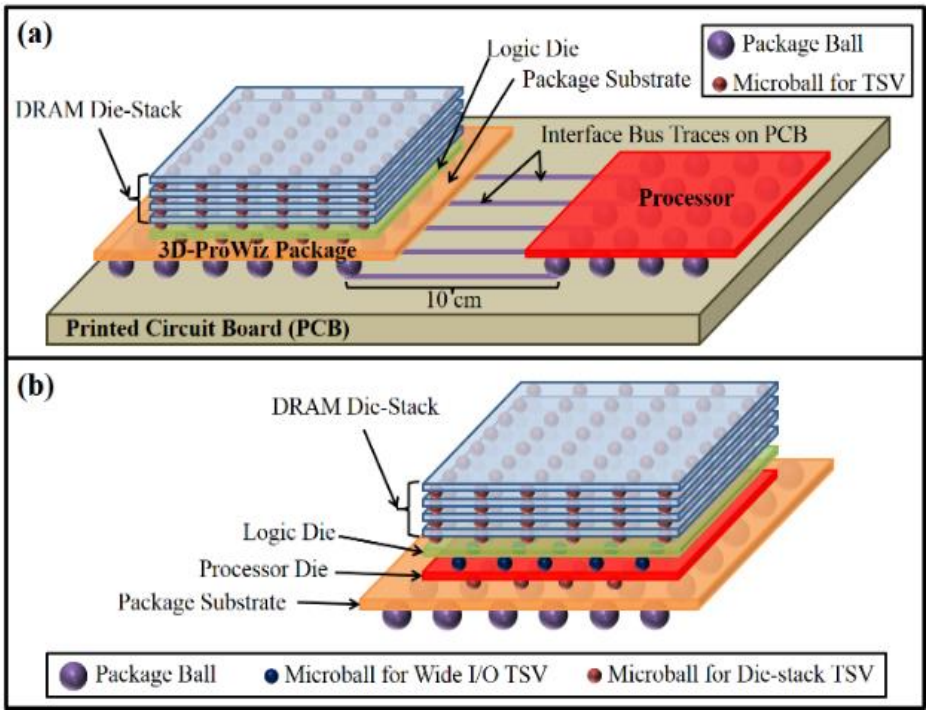


Figure 85: (a) Schematic of how DDR3, LPDDR3 and differential interfaces can be used to realize an off-chip memory-to-processor interconnect, (b) schematic of how Wide-I/O interface can be used to realize an memory-to-processor interconnect through 3D-stacking.

For better understanding of the results of these simulations, first consider Figure 85, which shows a schematic of how a 3D-ProWiz DRAM die-stack can use different type of electrical interfaces to connect to a processor. Figure 85(a) shows how the processor can be connected to an off-chip package of 3D-ProWiz die-stack using DDR3, LPDDR3, or differential interfaces. The individual dies of the DRAM die-stack are connected to the logic die using TSVs. The logic die in turn is connected to the package substrate using TSVs. The TSVs used for inter-die connections across the die-stack are global or intermediate level TSVs [31]. The 3D-ProWiz package, as shown in Figure 85, is bonded to the PCB using package level TSVs and package ballouts. Similarly, the processor package is bonded to the PCB using package level TSVs and package ballouts.

The interface bus traces are printed on the PCB, which are used to connect the memory package with the processor package through package ballouts, as shown in the figure. Typically, the distance between the memory package and processor package on today's motherboards/PCBs can be up to 10cm. The interface bus traces printed on the PCB are similar to standard transmission lines in case of DDR3 and LPDDR3 interfaces, whereas, in case of differential interface, they are similar to low-swing differentially coupled transmission lines. In contrast, the JEDEC standardized Wide-I/O interface is made of TSVs and is used to bond connect two or more devices together to be stacked upon one another. As shown in Figure 85(b), the 3D-ProWiz die-stack can be stacked on the processor die through Wide-I/O TSVs. Due to the TSV-based low-power design of Wide-I/O interface, it is not suitable for connecting the processor to an off-chip memory. The photonic interface in the 3D-ProWiz die-stack can connect to a processor using DWDM fibers through on-chip edge couplers, as shown in Figure 74.

Figure 86 shows energy-per-byte values for various inter-faces used with the 3D-ProWiz core across the PARSEC benchmarks. It can be observed that the photonic interface consumes

about 82% less energy-per-byte over all other off-chip interfaces. More specifically, the photonic inter-face consumes about 88.6%, 79.5%, and 66.6% less energy-per-byte on average over DDR3, differential, and LPDDR3 interfaces respectively. The photonic interface has the shortest burst length (2 cycles) and largest bus width, which results in the highest throughput for the pho-tonic bus. Moreover, as shown in Table 17, the photonic interface has the smallest values of per access dynamic energy and static power, which results in the smallest value of average power. The combined effect of these energy and throughput benefits renders the least energy-per-byte value for the photonic interface among the off-chip interfaces.

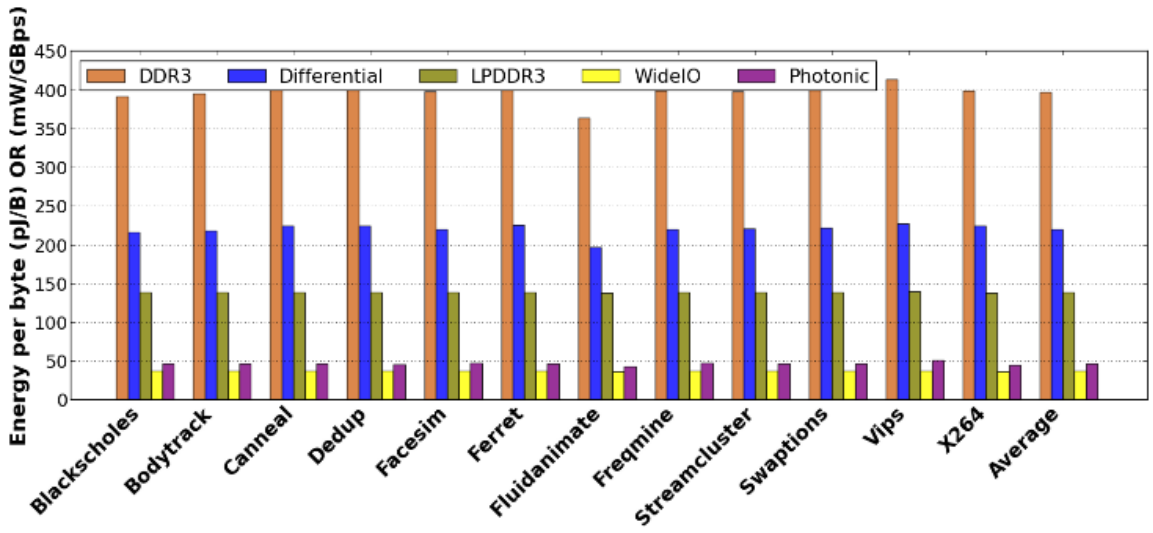


Figure 86: Energy-per-byte values for various interfaces across the PARSEC benchmarks.

However, as shown in Table 17, the Wide-I/O interface has the least static power consumption among all the inter-faces. Also, the width of the Wide-I/O interface is comparable to the width of the photonic interface. So, for the Wide-I/O interface, the combined effect of low static power and comparable interface width results in 16.7% less energy-per-byte value on average over the photonic interface.

It can be concluded from these results that in case of an off-chip memory, the photonic interface results in the best energy-efficiency over all other interfaces, whereas, if the DRAM is stacked on the processor chip in a single package, Wide-I/O interface proves to be a more promising (energy-efficient) option.

10.7. SIMULATION RESULTS

10.7.1. SIMULATION SETUP

We performed trace-driven simulation analysis to com-pare 3D-ProWiz with other DRAM architectures. Memory access traces for the PARSEC benchmark suite [76] were extracted from detailed cycle-accurate simulations using gem5 [77]. We considered twelve different applications from the PARSEC suite: Blackscholes, Bodytrack, Canneal, Dedup, Facesim, Ferret, Fluidanimate, Freqmine, Streamcluster, Swaptions, Vips, and x264. We ran each PARSEC benchmark for a "warm-up" period of 1 billion instructions and captured memory access traces from the subsequent 1 billion instructions extracted. These memory traces were then provided as inputs to the DRAM simulator DRAMSim2 [212], which we used to model 3D-ProWiz and other DRAM architectures. Table 18 gives the configuration of Gem5 that was used for this study. We chose direct mapped cache associativity because of its simplicity and low-cost implementation. The use of N-way set associative cache would change the memory access pattern, and hence it would change the behavior of the DRAM subsystem. But, the effect of different memory access patterns on the system behavior is captured by the different kinds of applications from the PARSEC benchmark suite used in this study. For this reason and for the sake of brevity we do not include the evaluation results for different cache associativity in this chapter.

We used the timing, energy and static power values given in Table 13 to characterize the various DRAM architecture in DRAMSim2. We modeled memory interfaces using the method

discussed in Section 10.6.3. Table 19 shows the memory configurations used in DRAMSim2 for the comparison across different DRAM architectures. As discussed in Section 10.4, the interface used with 3DSams in [198] closely resembles the low power DDR3 interface LPDDR3. Therefore, for this study, we use LPDDR3 interface with the 3DSams DRAM. As shown in Table 19, we use DDR3, differential and photonic interfaces with the DDR3, HMC and 3D-ProWiz DRAMs respectively. An RBRR scheduling scheme, an open page policy and rank:row:col:bank ad-dress mapping scheme were used for all simulations. Average latency, total power consumption, and energy-delay product values for the memory subsystem were obtained from DRAMSim2. The results of simulations with PARSEC benchmarks are discussed in the following subsection.

Table 18: Gem5 simulation configuration.

#Cores	4 ARM	L2 Coherence	MOESI
L1 I Cache	16KB	Frequency	5 GHz
L1 D Cache	16KB	Issue Policy of cores	In-order
L2 Cache	128KB	# Memory Controllers	1
-		Cache Associativity	Direct Mapped

Table 19: DRAMSim2 simulation configurations.

3D-ProWiz	8Gb module, 4 ranks, 512 banks/rank, Stacked die count: 4 Burst length: 2, 1GHz, Interface: photonic
HMC	8Gb module, 4 ranks, 16 banks/rank (vault), Stacked die count: 4 Burst length: 8, 1GHz, Interface: differential
3DSPDRAM	8Gb module, 8 banks/rank, 4 ranks Stacked die count: 4 Burst length: 32, 1GHz Single subarray access (SSA), Interface: photonic
3DSams	8Gb module, 4 ranks, 8 banks/rank, Stacked die count: 4 Burst length: 8, 1GHz, Interface: LPDDR3
DDR3	8Gb module, 4 ranks, 1 chip/rank, 8 banks/rank, Burst length: 8, 1GHz, Interface: DDR3

10.7.2. SIMULATION RESULTS FOR PARSEC BENCHMARKS

This subsection presents the average latency, power, and energy-delay product values for the 3D DRAM designs shown in Table 19 obtained for PARSEC benchmarks. Figure 87 shows power consumption values for the various DRAM architectures across the PARSEC benchmarks. It can be observed that 3D-ProWiz consumes about 52% less power on average over all the other DRAM architectures. More specifically, 3D-ProWiz consumes about 75.3%, 23.4% and 58.3% less power on average over DDR3, 3DSAMS and HMC respectively. 3DSPDRAM consumes about 72.3% less power on average over 3D-ProWiz. In fact, 3DSPDRAM consumes the least amount of power across all the DRAMs. The reason behind it is the single-subarray-architecture (SSA) architecture in 3DSPDRAM, which yields the smallest value of activation-precharge energy resulting in the lowest power consumption. The reason for the lower power consumption in 3D-ProWiz com-pared to other DRAMs is its smaller values of per-access energy, the effect of which cumulates over time to minimize average power consumption.

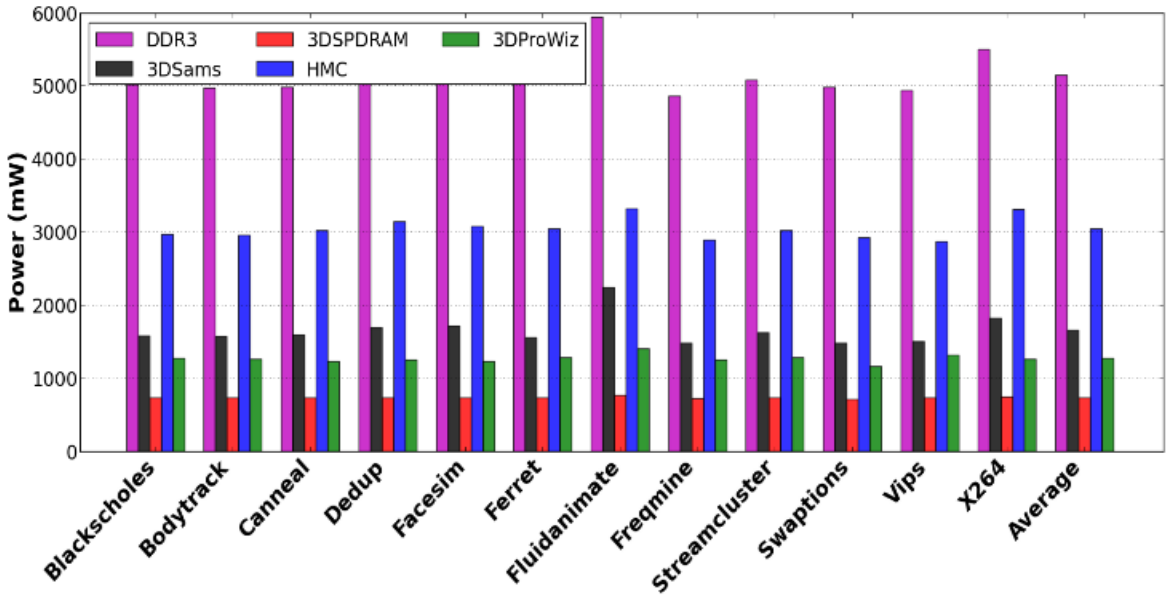


Figure 87: Power values for various DRAM architectures across PARSEC benchmarks.

Figure 88 shows the average latency values for different DRAM designs across the PARSEC benchmarks. The aver-age latency was calculated by dividing the total latency by total number of transactions. 3D-ProWiz demonstrates about 41.9% less average latency over all the other 3D DRAM architectures. More specifically, 3D-ProWiz demonstrates 35.1%, 31.3%, 61.5% and 22.9% lower average latency values over DDR3, 3DSAMS, 3DSPDRAM and HMC respectively. As discussed in Section 10.2 and in [205], the fine-grained rank-level 3D partitioning of the data array in HMC better utilizes potential TSV bandwidth compared to the coarse-grained rank-level partitioning used in 3DSPDRAM and 3DSAMS. Due to this reason, HMC has an edge over 3DSPDRAM and 3DSAMS, which translates into a performance edge for HMC over 3DSPDRAM and 3DSAMS. The reason behind the better performance of 3D-ProWiz over other DRAM designs is the reduced RC loading of the access path in 3D-ProWiz, which is a result of the smaller subarrays and elimination of global lines in the architecture. As implied from the discussion given in Section 10.4, the increased serialization delay of the SSA architecture of 3DSPDRAM results in larger average latency for 3DSPDRAM compared to other DRAMs with photonic interfaces.

Figure 89 shows the energy-delay product (EDP) values for different DRAM designs across the PARSEC bench-marks. EDP values were calculated by multiplying the average-energy with average latency for memory accesses. Average-energy in each benchmark was calculated by di-viding the total energy by total number of transactions. 3D-ProWiz yields an 80.6% lower EDP value on average over all the other 3D DRAM architectures. More specifically, 3D-ProWiz yields 89.7%, 64.4%, 75.1% and 75.6% less EDP values on average over DDR3, 3DSAMS, 3DSPDRAM and HMC respectively. These improvements in EDP for 3D-ProWiz follow directly from the power and latency improvements that were discussed earlier.

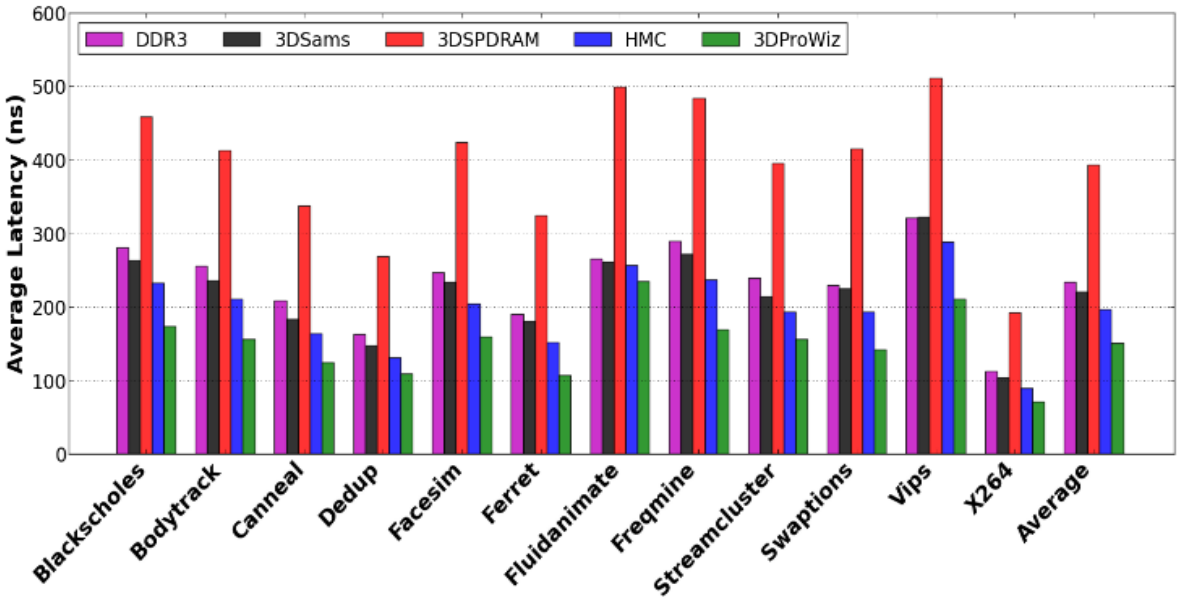


Figure 88: Average latency for various DRAM architectures across PARSEC benchmarks.

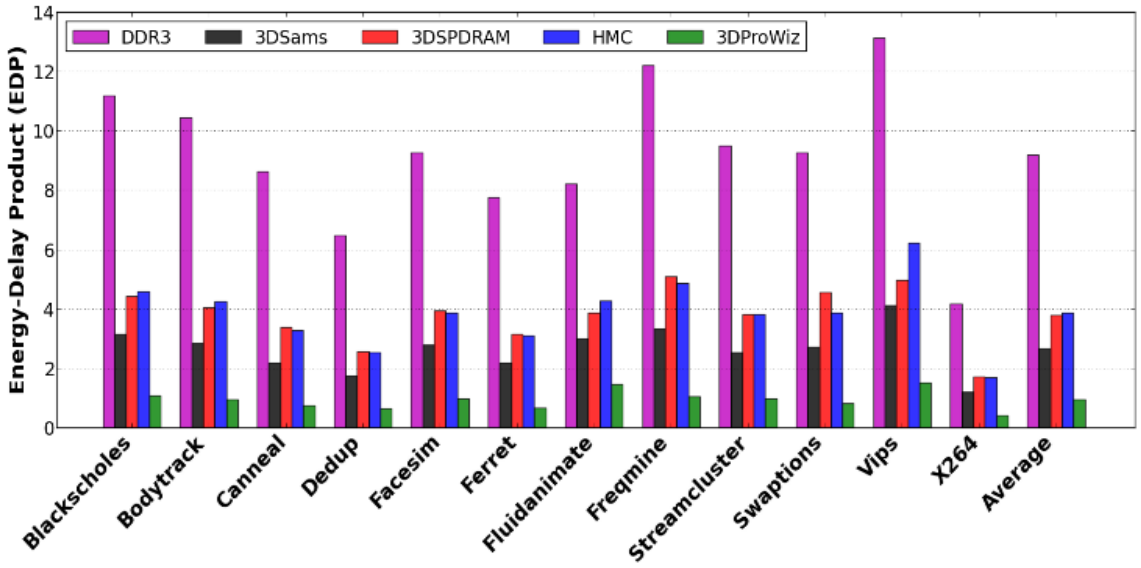


Figure 89: Energy-delay product for DRAM architectures across PARSEC benchmarks.

In summary, the 3D vertical routing of the internal memory bus using TSVs at subarray-level granularity and fanout buffers enable 3D-ProWiz to use smaller dimension subarrays. This in turn reduces the random-access latency and activation-precharge energy of 3D-ProWiz DRAM over other DRAM designs. Consequently, 3D-ProWiz yields on average 52%, 41.9% and 80.6%

improvements in power consumption, latency and energy-delay product (EDP) respectively over other DRAM architectures. 3D-ProWiz DRAM has about 9.16mm^2 (41.7%) more die area than the average die area of other DRAMs, which increases its relative cost. Nonetheless, the significant improvements in access latency, power, and energy-efficiency make 3D-ProWiz a promising architecture candidate for future 3D DRAMs.

10.8. CONCLUSIONS

This chapter introduced 3D-ProWiz, a novel high band-width and low-latency 3D DRAM architecture. 3D-ProWiz integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism than other 3D DRAM architectures. The 3D vertical routing of the internal memory bus using TSVs at subarray-level granularity and fanout buffers enable 3D-ProWiz to use smaller dimension subarrays without significant area overhead. This in turn reduces the random-access latency and activation-precharge energy. Consequently, 3D-ProWiz yields on average 52%, 41.9% and 80.6% improvements in power consumption, latency and energy-delay product (EDP) respectively over other DRAM architectures.

Detailed sensitivity analysis of the tTAW constraint in this work established that it is very important to design a robust PDN, which is less prone to noise and which has more relaxed tTAW constraint, for a 3D DRAM subsystem with relatively small row cycle time and large bank count. We also showed in this work that, in case of an off-chip memory, the photonic interface renders the best energy-efficiency over all other interfaces, whereas, if the memory is embedded with the processor chip in a single package, the Wide-I/O interface proves to be more promising and energy-efficient option.

The significant improvements demonstrated by 3D-ProWiz position it as a promising architecture for future DRAMs. The performance of the 3D-ProWiz memory system can be further

improved by using intelligent scheduling schemes and novel memory controller designs so that the greater parallelism of 3D-ProWiz architecture can be better exploited. Moreover, the capacity of the 3D-ProWiz DRAM module can be greatly scaled by using intelligent arbitration techniques for the photonic bus. Thus, with potential opportunities for further improvements, the 3D-ProWiz architecture can become an even more promising solution for future DRAM implementations.

11. A NOVEL 3D GRAPHICS DRAM ARCHITECTURE FOR HIGH-PERFORMANCE AND LOW-ENERGY MEMORY ACCESSES

This chapter presents a high-bandwidth 3D graphics DRAM architecture (*3D-SGDRAM*) with reduced access time and energy consumption. A novel 3D bank organization is employed with TSVs at subarray-level granularity to activate an optimal number of subarrays in lock-step to guarantee fast and low-energy memory access without significant area overhead. A new bitline interface enables access to only a selective group of bitlines in all active subarrays during a memory transaction, which greatly reduces row activation energy with optimal page size. Experimental results with CUDA benchmarks indicate that 3D-SGDRAM yields 57.5%, 77.7%, and 45.2% improvements in power, latency, and energy-delay product (EDP) on average over state-of-the-art GDDR5 and GDDR5M solutions.

11.1. INTRODUCTION

Recent advances in GPGPU computing have greatly increased bandwidth requirements for graphics DRAMs [216], leading to the introduction of GDDR5 memory. Today, the data rate of the GDDR5 memory interface has reached 7Gbps/pin [217]. But the high-speed interface of GDDR5 consumes high power, making it energy-inefficient for the green computing solutions in demand today. Moreover, the throughput of GDDR5 is not expected to meet the bandwidth requirements of future high-performance computing systems within stringent cost constraints. These trends highlight the need for innovation in graphics DRAM architectures to improve bandwidth, cost, and energy consumption.

Since the emergence of through silicon via (TSV) based 3D integration technology, 3D stacking has been a promising option for designing faster and more energy-efficient DRAM cores.

The main advantage of TSV-based stacking is that it reduces the wire-length between modules located on different tiers, which in turn reduces delay and energy of inter-module interconnects. Several 3D DRAM architectures (e.g., [142], [164], [197], [198], and [218]) have been reported in recent years. Such 3D DRAMs require new methods for efficient address and data path routing, and 3D cell organization, to realize their performance potential while limiting area and cost overheads.

In this chapter, we present *3D-SGDRAM*, a new 3D-stacked graphics DRAM architecture. *3D-SGDRAM* employs a new bitline interface and a bank organization based on detailed parameter characterization and optimization to achieve simultaneous improvements in performance, throughput, power, and area of the DRAM core. Our key contributions with *3D-SGDRAM* are:

- We modify the bitline interface of the DRAM core to enable access to only a selective group of bitlines in all active subarrays during a memory transaction, which helps optimize page size and related architectural parameters;
- We characterize the interdependence between various architectural parameters of the *3D-SGDRAM* bank organization and optimize these parameters, to achieve benefits in performance, power, throughput and area;
- We experimentally compare and contrast our *3D-SGDRAM* architecture with two state-of-the-art graphics DRAM architectures: GDDR5 [219] and GDDR5M [220].

11.2. RELATED WORK

Traditional GDDR5 interfaces use improved synchronization and clocking mechanisms to enable higher speed data transfers than conventional DDR DRAM architectures, but at the cost of notable complexity, noise, and larger power consumption in the DRAM cores as well as the

memory controller. Lee et al. [220] modified the conventional GDDR5 interface in their proposed GDDR5M architecture, to reduce noise, dynamic power, and standby power. But the GDDR5M memory core suffers from low-throughput and low energy-efficiency. The Wide I/O DRAM, a new JEDEC standard for on-chip embedded DRAMs [24], has the potential to overcome these shortcomings, but its use as off-chip main memory for high-performance graphics subsystems is limited due to pin-bandwidth limitations and high capacitance of off-chip PCB interconnects.

Several 3D DRAM architectures have been proposed in recent years (e.g., [142], [164], [199], [200], and [205]) that aim to improve throughput and energy consumption of main memory. Some works propose smaller subarrays that reduces the RC loading of local wordlines and bitlines (e.g., [200]), while others propose true 3D organizations of data arrays that reduce the length of global lines (e.g., [164] and [199]). Thakkar et al. [164] propose the 3D-Wiz 3D DRAM architecture that eliminates global lines by employing aggressive vertical routing of intra-bank buses using TSVs and fan-out buffers. However, using TSVs at such granularity requires optimization of a number of interdependent architectural parameters to achieve improved memory core throughput and power without significant area and cost overhead, which 3D-Wiz fails to do. Unlike 3D-Wiz, our proposed *3D-SGDRAM* architecture employs TSVs at subarray level granularity after optimizing various architectural parameters of the 3D bank organization to achieve simultaneous benefits in performance, power, throughput and area. To the best of our knowledge, *3D-SGDRAM is the first 3D DRAM architecture optimized for graphics interfaces.*

11.3. 3D-SGDRAM ARCHITECTURE OVERVIEW

Like the 3D-Wiz module depicted and described in [164], a *3D-SGDRAM* module is also a stack of 5 die layers. The top-most four layers are DRAM data layers, which are stacked on a logic layer. The DRAM data layers are divided into four equal bank groups, and each bank group has

16 banks with each bank being 16Mb in size and partitioned across all four data layers. We chose these counts for banks and bank groups for *3D-SGDRAM* to make them consistent with the respective counts in GDDR5 modules. The logic layer contains all of the logic for global control, including a bank-state controller (BSC), address decoder, and an I/O buffer for each bank to support a packet-based interface protocol, as proposed in [200]. The following subsections present more details of our *3D-SGDRAM* architecture.

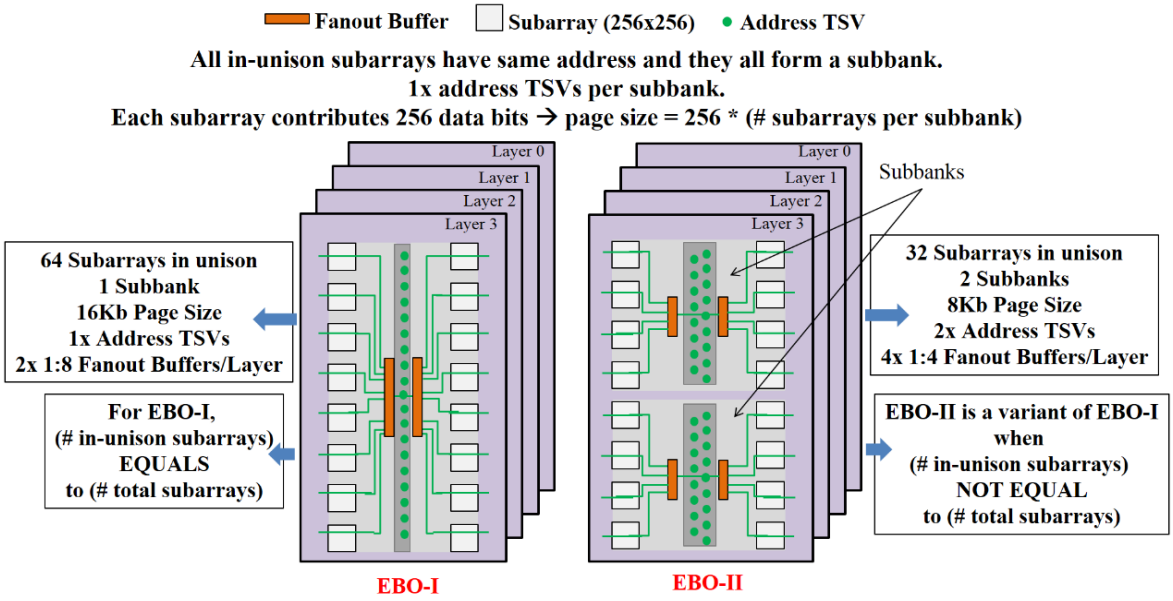


Figure 90: Two example bank organizations (EBO-1, EBO-II).

11.3.1. ARCHITECTURAL PARAMETER INTERDEPENDENCE

In this section, we identify the interdependence among various bank organization parameters in *3D-SGDRAM* and motivate the need for a novel bitline interface.

Similar to 3D-Wiz [164], *3D-SGDRAM* employs TSVs at subarray level granularity, replacing the global lines of intra-bank buses with TSVs. The TSVs in turn connect the inter-bank buses to the local lines of individual subarrays of a bank. But using TSVs at such fine granularity requires a very large number of address and data TSVs. To reduce the number of address TSVs,

3D-Wiz [164] uses 1:8 fanout buffers. However, using TSVs with fanout buffers at such fine granularity requires optimization of several interdependent architectural parameters to achieve improved memory core throughput and power without high area and cost overhead, which 3D-Wiz fails to do. Unlike 3D-Wiz, our proposed *3D-SGDRAM* architecture uses TSVs with fanout buffers obtained after an architectural parameter optimization step.

To understand the interdependence among various bank organization parameters in *3D-SGDRAM*, consider Figure 90, which shows two alternatives for employing TSVs in an example bank organization (EBO) of a 4Mb bank within *3D-SGDRAM*. Reducing the number of in-unison subarrays in a 4Mb bank with total 64 subarrays by a factor 2 (from 64 in EBO-I to 32 in EBO-II) decreases the fanout strength of fanout buffers and page size by a factor of 2 (from 1:8 and 16Kb to 1:4 and 8Kb respectively), but it also increases the number of address TSVs per bank by the same factor. Similarly, reducing the number of in-unison subarrays in a bank by a factor larger than 2 would also proportionally affect page size and number of address TSVs. Thus, reducing the number of in-unison subarrays would reduce the delay and energy of fanout buffers (which depends on the fanout strength), but it would also increase the area and energy overhead related to the address TSVs. *Such contrasting interdependence necessitates co-optimization of various architectural parameters to design the best possible bank organization in any DRAM architecture.*

Figure 91 illustrates this interdependence among various DRAM design parameters. The parameter *number of in-unison subarrays* is at the top of the dependence hierarchy and other parameters such as *page size*, *path effort*, and *number of TSVs per bank* can be observed to depend on it. These dependent parameters in turn control the area, energy, and delay of the entire memory subsystem. Among all the parameters in Figure 91 that are enclosed in rectangular boxes, the

number of in-unison subarrays is the only independent control variable, and controls dependent parameters in different ways (e.g., some are proportional, others are inversely proportional).

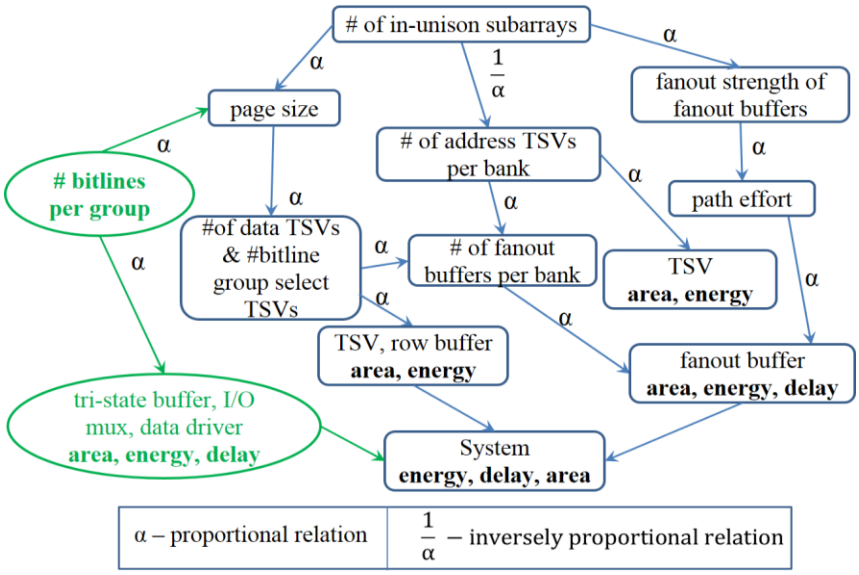


Figure 91: Interdependence among various architectural parameters.

Clearly, focusing on the co-optimization of the dependent parameters in isolation will yield suboptimal solutions that possess high energy, area, and latency for the DRAM core. Therefore, we introduce a new independent parameter *number of bitlines per group* in the dependence hierarchy that assists the parameter *number of in-unison subarrays* during an optimization phase to find better solutions with greater benefits in area, energy and latency for the DRAM core. This new parameter controls *page size* and other dependent parameters as shown in Figure 91. To realize this parameter, we propose a novel interface between bitlines and sense amplifiers to select a group of bitlines to be activated during a memory transaction, as discussed next. The reason behind the proportional dependence of *page size* on the parameter *number of bitlines per group* is explained in subsection 11.3.2.

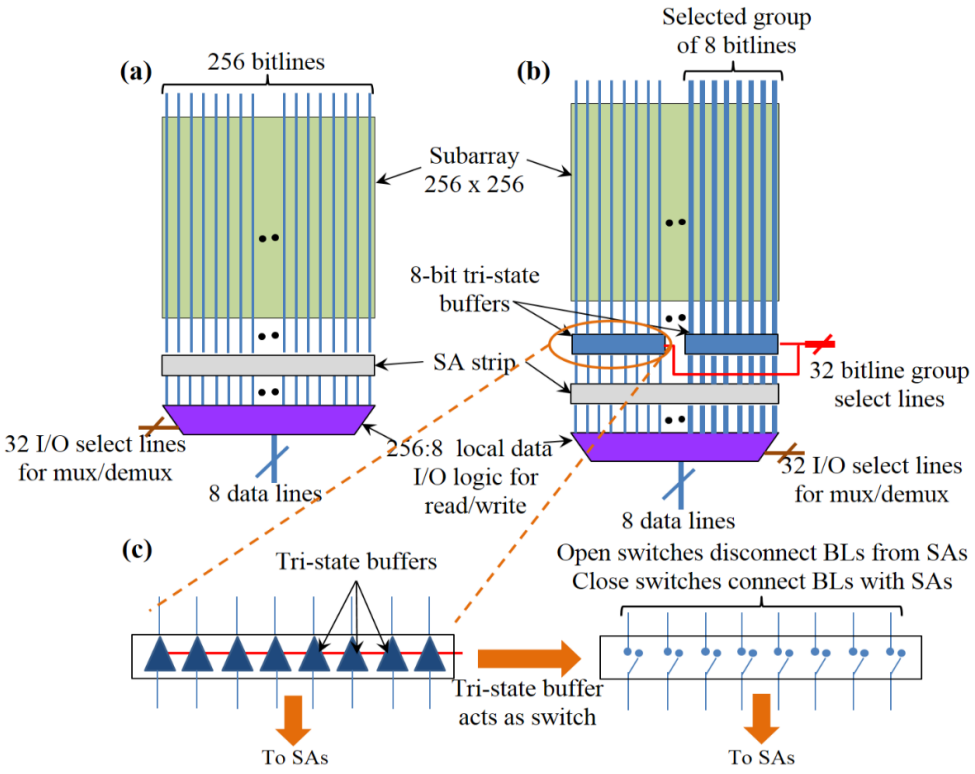


Figure 92: (a) Conventional bitline interface, (b) New bitline interface of 3D-SGDRAM, (c) Tri-state buffers of the new bitline interface.

11.3.2. NEW BITLINE INTERFACE IN 3D-SGDRAM

The bitline interface of a conventional DRAM subarray of size 256×256 is shown in Figure 92(a). All 256 bitlines of a subarray are connected to sense amplifiers (SAs). The output of SAs is connected to local data I/O logic that drives 8 bidirectional local data lines for input or output during a write or read, respectively. Our modified bitline interface is shown in Figure 92(b). Here the bitlines of a 256×256 subarray are arranged in 32 groups with each group having 8 bitlines. The grouping of bitlines is done at design time, thus the number of bitlines per group does not change dynamically. Each group of bitlines is connected to SAs through 8-bit wide tri-state buffers. So, there are a total of 32 8-bit wide tri-state buffers associated with a total of 32 groups of bitlines.

As shown in Figure 92(c), the 8-bit wide tri-state buffer acts as an array of 8 switches that connects or disconnects bitlines to or from SAs. When the tri-state buffer is in high-impedance state, it acts as an array of open switches and disconnects corresponding bitlines from SAs. When the tri-state buffer is in active state, it acts as an array of closed switches and connects corresponding bitlines to SAs. The states of these 32 tri-state buffers (32 sets of switches) are controlled by 32 bitline group (BL) group select lines, as shown in Figure 92(b).

During a memory transaction, the 32 BL group select lines are driven to set up a particular address pattern on them so that only one of 32 tri-state buffers is activated depending on the address. Thus, a particular group of 8 bitlines corresponding to the activated tri-state buffer is selected per subarray to be connected to SAs. All the other bitlines of the subarray are disconnected. The same address pattern provided to the BL group select lines is also fed to the 32 I/O select lines, which connects outputs of the 8 selected SAs to 8 local data lines.

A particular group of bitlines is pre-selected by the address pattern on the BL group select lines before serving any memory request. Therefore, a row access command results in *SAs* detecting only as many data bits per subarray as equal to the number of bitlines in the pre-selected group of BLs, instead of detecting all 256 data bits of the accessed row. As all the other bitlines are disconnected, they act as open circuit wires and do not get charged or discharged by their corresponding bitcell capacitors during row access and precharge. Moreover, only the pre-selected bitlines of all in-unison subarrays of the target subbank contribute to page size. This in turn *reduces the page size and related energy overhead without altering the number of in-unison subarrays.*

The best possible value for the *number of bitlines per group* parameter should be chosen to optimize page size at design time, and for that, the width of the tri-state buffers can be adjusted as needed. The next subsection (Section 10.3.3) describes our optimization phase to accomplish this

parameter selection. Note that introducing tri-state buffers in the bitline interface incurs extra energy, delay, and area overhead, which we analyze using CACTI-3DD [205] and account for in Table 20.

Table 20: Results of *3D-SGDRAM* design overhead analysis.

	Dynamic Energy (pJ)	Leakage Power (µW)	Area (μm^2)
TSV	0.167	-	5
32-bit Tri-state buffer	0.033	7.2	22.4
1:32 Fanout buffer	0.037	16.7	45.5

11.3.3. BANK ORGANIZATION PARAMETER OPTIMIZATION

Problem Formulation: The *number of in-unison subarrays* and the *number of bitlines per group* are the two main independent variables of the example bank organization (EBO) that control all other parameters in the dependence hierarchy shown in Figure 91. Therefore, we use all possible values of these two variables as an input to our problem of parameter optimization. We assume a fixed bank size of 16Mb to maintain consistency with GDDR5, and set the subarray size to 256×256, as found in some state-of-the-art 3D DRAMs, which in turn sets the total number of subarrays in a bank to be 256 for the bank size of 16Mb. Owing to these assumptions, the variables *number of in-unison subarrays* (Γ) and *number of bitlines per group* (X) can attain only a finite number of discrete values that satisfy the following conditions: $\Gamma, X = 2^n;\ where\ n \in N;\ and\ \Gamma, X \leq 256$. Thus, the set of all possible values for X and Γ is {1, 2, 4, 8, 16, 32, 64, 128, 256}. The individual values for Γ and X combine to make a duplet in 81 different ways. We create a set Y of these duplets, $Y = \{(\Gamma_1, X_1),\ (\Gamma_1, X_2), \ldots, (\Gamma_9, X_9)\}$ and give it as an input to our problem.

Problem Objective: The main objective of our optimization framework is to design a bank organization for 3D-SGDRAM that minimizes energy and latency, and maximizes area efficiency for the given constraints on page size.

Area efficiency directly impacts DRAM subsystem cost, and is defined as the percentage of total die area which corresponds to the DRAM cell area. We combine these objectives into a single minimization objective: *Minimize(EDP/AE)* for the DRAM core to simultaneously minimize EDP and maximize AE, where EDP is the energy-delay product and AE is the area efficiency. Out of all possible solutions, we only consider solutions that yield page size (PS) of either 8Kb or 16Kb. We chose these values of page size, because commodity GDDR3 and GDRR5 DRAMs of 1Gb density support page sizes of either 8Kb or 16Kb [219].

Brute-force search-based optimization framework: We enhance CACTI-3DD [205] with Tezzaron's high density TSV model [26] and use it in our proposed optimization framework. For each duplet of the input set *Y*, we use CACTI-3DD to arrange all 256 subarrays of a *3D-SGDRAM* bank into 256/Γ number of subbanks (as per EBO-II) with each subbank having Γ subarrays working in lockstep, and then calculate *EDP/AE* for the resulting bank organization. EDP values are obtained by multiplying energy per access (EPA) values with row cycle time (*tRC*). Then, the brute-force search algorithm finds an optimal duplet with minimum *EDP/AE* out of all duplets that satisfy given page size constraints.

Table 21: *EDP/AE* values for seven duplets that satisfy PS constraints.

Γ	32	64	64	128	128	**256**	256
X	256	128	256	64	128	**32**	64
EDP/AE	2.73	2.14	5.06	2.13	3.97	**1.74**	3.61

Table 21 shows *EDP/AE* values for seven duplets that satisfy page size (PS) constraints. From the table, duplet (256, 32) has minimum EDP/AE, which makes it the optimal solution.

Therefore, we use duplet (256, 32) to derive the 3D-SGDRAM bank organization, which has 256 subarrays working in unison and 32 bitlines per BL group. This 3D-SGDRAM configuration yields a value of 3.0nJ energy per (read) access for the DRAM core, a 20.5ns row cycle time, 8Kb page size, and 35.3% area efficiency.

11.4. SIMULATION RESULTS

We performed a benchmark-driven, simulation-based analysis to compare *3D-SGDRAM* with existing graphics DRAM architectures such as GDDR5 [219] and GDDR5M [220]. We do not show results for 3D-Wiz [164] as it is not a graphics DRAM architecture and thus not designed to support high graphics data throughputs.

Memory access traces for the CUDA benchmark suite [221] were extracted from detailed cycle-accurate simulations using GPGPUSim [222]. We configured GPGPUSim for the QuadroFX5800 GPGPU. We considered eight different applications from the CUDA benchmark suite: Breadth-First Search (BFS), Coulombic Potential (CP), LIBOR Monte Carlo (LIB), 3D Laplace Solver (LPS), MUMmerGPU (MUM), Neural Network (NN), Ray Tracing (RAY), and StoreGPU (STO).

We ran each CUDA benchmark in a continuous loop and captured memory access traces from the L2Cache-to-DRAM interface over two billion instructions. These memory traces were then provided as inputs to the DRAM simulator DRAMSim2 [212], which we modified to model *3D-SGDRAM* and the other graphics DRAM architectures. The energy and timing parameters given in Table 22 were analyzed using CACTI-3DD based models of the DRAM architectures. These parameters were provided as inputs to the DRAMSim2 based models. An FCFS scheduling scheme and a closed-page policy were used for all simulations.

Table 22: Energy and timing parameters for graphics DRAMs.

	3DSGDRAM-P	GDDR5	GDDR5M	3D-Wiz
Read E (nJ/access)	0.66	2.5	1.5	1.98
Write E (nJ/access)	0.5	2.2	1.2	1.98
ActPre E (nJ/access)	2.34	3.1	3.1	1.4
Background (mW)	115.3	810	470	105.3
tRAS (ns)	14.7	23.2	23.2	19.5
tRC (ns)	20.5	40.3	40.3	25

Figure 93 shows the power consumption values for the various DRAM architectures across the CUDA benchmarks. The figure shows the total power, which is the sum of *Background* power, read/write (*Burst*) power, and activation-precharge power (*ActPre*). It can be observed that *3D-SGDRAM* consumes about 65.8% and 44% less power on average than GDDR5 and GDDR5M respectively for the CUDA benchmarks. The smaller values of read/write and activation-precharge energies translate in lower power consumption in 3D-SGDRAM compared to GDDR5 and GDDR5M.

Figure 94 shows the energy-delay product (EDP) values for the various DRAM architectures across the CUDA benchmarks. It can be observed that *3D-SGDRAM* demonstrates about 81.8% and 71.2% less EDP on average than GDDR5 and GDDR5M, respectively. Moreover, the average latency values over all CUDA benchmarks for *3D-SGDRAM*, GDDR5 and GDDR5M memory subsystems were found to be 9.8ns, 18.1ns, and 19.3ns, respectively. The average latency for each CUDA benchmark is calculated by dividing the total latency with total number of transactions. *3D-SGDRAM* demonstrates about 45.9% and 49.2% less latency on average than GDDR5 and GDDR5M respectively. The improvements in EDP for *3D-SGDRAM* follow directly from these improvements in power and latency.

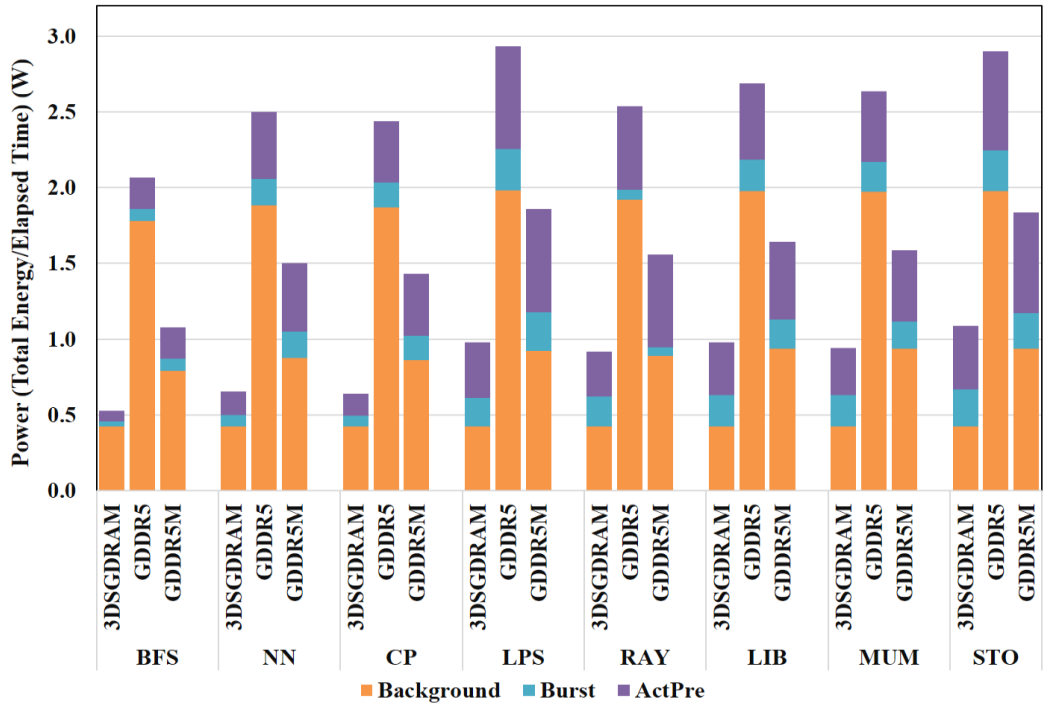


Figure 93: Power consumption for various graphics DRAMs across CUDA benchmarks.

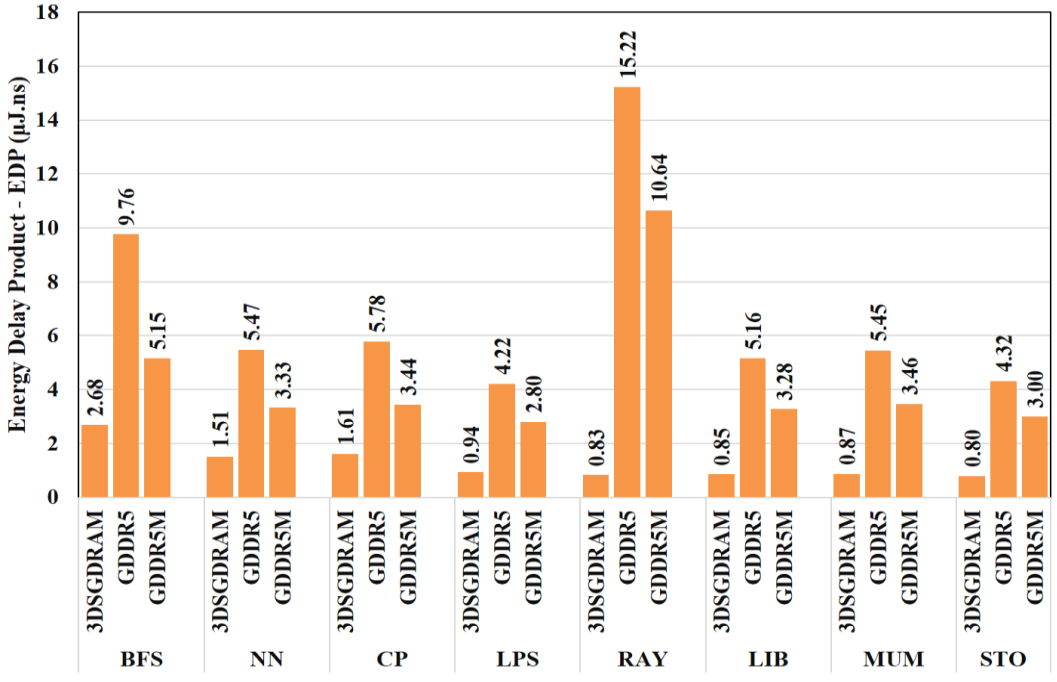


Figure 94: Energy delay product for various graphics DRAMs across CUDA benchmarks.

11.5. CONCLUSIONS

This chapter introduced a new high-bandwidth 3D graphics DRAM architecture (*3D-SGDRAM*) with reduced access time and energy consumption. A novel 3D bank organization with a new bitline interface guarantees fast and low-energy memory access with high area efficiency. *3D-SGDRAM* yields on average 57.5%, 77.7%, and 45.2% improvements in power consumption, average latency, and energy-delay product (EDP) over state-of-the-art GDDR5 and GDDR5M architectures. The significant improvements demonstrated by *3D-SGDRAM* position it as a promising solution for future graphics DRAMs.

12. MASSED REFRESH: AN ENERGY-EFFICIENT TECHNIQUE TO REDUCE REFRESH OVERHEAD IN HYBRID MEMORY CUBE ARCHITECTURES

This chapter presents a novel, energy-efficient DRAM refresh technique called *massed refresh* that simultaneously leverages bank-level and subarray-level concurrency to reduce the overhead of distributed refresh operations in the Hybrid Memory Cube (HMC). In *massed refresh*, a bundle of DRAM rows in a refresh operation is composed of two subgroups mapped to two different banks, with the rows of each subgroup mapped to different subarrays within the corresponding bank. Both subgroups of DRAM rows are refreshed concurrently during a refresh command, which greatly reduces the refresh cycle time and improves bandwidth and energy efficiency of the HMC. Our experimental analysis shows that the proposed *massed refresh* technique achieves up to 6.3% and 5.8% improvements in throughput and energy-delay product on average over JEDEC standardized distributed per-bank refresh and state-of-the-art scattered refresh techniques.

12.1. INTRODUCTION

Since the emergence of DRAM technology, reducing the overhead of refresh operations has been one of the major challenges for DRAM designers. Due to the volatile nature of DRAM cells, it is necessary to perform periodic refresh operations on them to retain the stored data bits over time. One of the downsides of periodic refresh operations is that they contribute significantly to the memory power, e.g., up to 30% of total power consumption [223], [224]. Refresh operations also require stalling waiting memory requests until the refresh finishes, which reduces DRAM performance, e.g., as explained in [225], the overall performance for a DDR4 system degrades by 18.8% for a variety of workloads due to periodic refresh.

In recent years, many techniques have been proposed to minimize the performance overhead of refresh operations in modern DRAMs [195], [223]-[230]. Although these techniques are mainly focused on 2D DRAMs, they can also be applied to emerging 3D-stacked DRAMs. However, the smaller bank-size and increased bank-level parallelism of 3D-stacked DRAMs results in higher power density and die temperature [23], [164], [199], which reduces data retention time, requiring more frequent refreshes (thus increasing refresh overhead) in 3D-stacked DRAMs. The greater power density of 3D DRAMs also increases the amount of current drawn from the power delivery networks (PDNs) of memory modules [198], which raises the noise level in PDNs resulting in degraded memory reliability and performance [198], [213]. Recent work has shown that the power density of 3D-stacked DRAMs can be decreased by decreasing refresh power. *This provides a strong motivation for 3D-stacked DRAM designers to explore new techniques that can reduce refresh overheads, thereby ensuring reliability and high performance.*

In this chapter, we propose *massed refresh*, a novel energy-efficient DRAM refresh technique that intelligently leverages bank-level as well as subarray-level parallelism to reduce refresh cycle time overhead in the 3D-stacked Hybrid Memory Cube (HMC) DRAM architecture. We have observed that due to the increased power density and resulting high operating temperatures, the effects of periodic refresh commands on refresh cycle time and energy-efficiency of 3D-stacked DRAMs, such as the emerging HMC [23], are significantly exacerbated. Therefore, we select the HMC for our study, even though the concept of *massed refresh* can be applied to any other 2D or 3D-stacked DRAM architecture as well. Our novel contributions in this chapter are summarized below:

- We demonstrate how minor changes in the DRAM bank access control logic of the HMC can be leveraged to increase bank-level parallelism during a distributed refresh operation, as part of our proposed *massed refresh* technique. We also quantify the extra overhead of area and power incurred due to these changes;
- Unlike the JEDEC standardized *all-bank refresh* technique [202] that refreshes all the banks of a rank concurrently, we restrict our proposed *massed refresh* technique to refresh only a few selected banks in parallel. This enhancement limits the PDN noise and peak refresh current to acceptable levels resulting in increased operation efficiency and performance;
- We investigate, for the first time, the effect of state-of-the-art *distributed per-bank refresh* [25] [203] and *scattered refresh* [225] techniques on DRAM energy-efficiency and compare these techniques with our *massed refresh* technique in terms of memory throughput and energy delay product (EDP) for the HMC DRAM architecture.

12.2. RELATED WORK

During the initial stages of DRAM development, the *burst refresh* scheme was used to refresh the whole DRAM device with a single refresh command in every retention cycle of 64ms under normal temperature conditions ($<85^0$C) or every 32ms under high temperature conditions ($>85^0$C). During a burst refresh operation, the entire DRAM device, including all the banks in all the ranks, becomes unavailable to non-refreshing memory tasks such as reading/writing data from/into memory for the amount of time it takes to perform the refresh, during which no productive tasks are performed.

To reduce the length of memory pauses and to increase the productivity of DRAM systems while still supporting refresh operations, the Joint Electron Device Engineering Council (JEDEC) standardized a new scheme called *distributed/ interleaved refresh* [202], wherein a memory

controller sends out multiple refresh commands in one retention cycle. In this scheme, a subset of DRAM rows, called "refresh bundle" are refreshed for every distributed refresh command issued. The time taken in refreshing a refresh bundle is termed as refresh cycle time (tRFC). The distributed refresh operation can be implemented in a DRAM rank at one of two granularities – (1) per-bank or (2) all-bank. In the *all-bank refresh* scheme that is used by all general purpose DDRx memory standards including the DDR4 standard, the refresh bundle is distributed among all banks of a rank and refresh operations of all the banks are completely overlapped in time. In the *per-bank refresh* scheme that is supported by some new memory standards such as LPDDR4 and high bandwidth memory (HBM) [25], all rows of the refresh bundle map to a single bank and their refresh operations are performed sequentially. In state-of-the-art 2D-DRAMs, depending on the workload characteristics, either *per-bank refresh* or *all-bank refresh* is favored to minimize the refresh overhead. But, as discussed in [231], the selection of an appropriate refresh method that can reduce the refresh overhead for 3D-stacked DRAMs requires much more complex deliberation and analysis of a number of architecture-level and design-level tradeoffs. We provide a brief discussion of these tradeoffs in Section 12.4.

In recent years, many research efforts have focused on reducing DRAM refresh overhead. Based on their method to minimize refresh overhead, these techniques can be classified into four different categories: *(i)* cell retention time aware methods [226], [227], *(ii)* methods that eliminate unnecessary refreshes [224], [228], *(iii)* refresh-aware scheduling methods [229], [230], and *(iv)* refresh cycle time reduction methods [195], [225].

Cell retention time aware methods exploit DRAM inter-cell variation in retention time either to minimize refresh operations [226] or to adaptively select a refresh period for a particular refresh operation [227]. But, these methods make DRAM cells more prone to errors which harms data

integrity, because they operate DRAM cells at a refresh interval beyond the specification range that DRAM manufacturer's guarantee. The methods that eliminate unnecessary refreshes [224], [228] require extensive data profiling at design time and power-hungry run-time decision making. The refresh-aware scheduling methods [229], [230] schedule refresh commands so that they do not collide with read or write commands, which improves utilization of resources and DRAM performance. But, these methods complicate memory controller design and also heavily depend on data access patterns of the workloads, providing benefits only if a large number of rows are activated. Refresh cycle time reduction methods [195], [225] identify refresh command cycle time (tRFC) as the main factor that limits DRAM performance. Among these, methods such as [195] intelligently choose refresh granularity to selectively reduce tRFC, which in turn reduces the queuing delay of memory access requests for a given application. On the other hand, *scattered refresh* [225] reduces refresh cycle time by exploiting subarray level parallelism, wherein the rows of a refresh bundle are scattered to different subarrays and their refresh operations are overlapped in time. Such refresh cycle time reducing methods have been shown to have more potential to overcome the DRAM refresh problem without the drawbacks of other methods. Besides, refresh cycle time reducing methods are orthogonal to other methods and can be applied in any combinations of the other methods.

In this chapter, we propose *massed refresh*, a novel energy-efficient distributed refresh technique that reduces refresh cycle time overhead by leveraging bank-level as well as subarray-level parallelism in the hybrid memory cube (HMC). The results of our analysis indicate that *massed refresh* can significantly reduce the overhead of distributed refresh operations in the HMC, thus improving memory throughput and energy-efficiency while keeping the PDN noise level to within acceptable limits for reliable operation.

12.3. BACKGROUND: HYBRID MEMORY CUBE (HMC)

This section provides a brief overview of the HMC architecture. The reader is directed to [23], [232] for more details on the HMC specification and operating mode configurations.

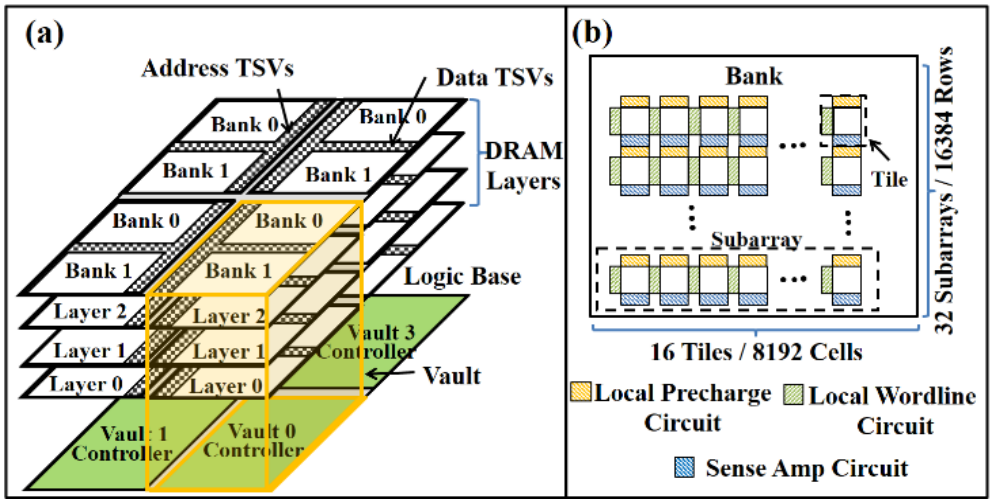


Figure 95: (a) Schematic of 4Gb hybrid memory cube (HMC) quad unit, (b) schematic of an HMC bank.

An HMC architecture consists of a single power efficient package containing multiple memory dies and one base die, stacked together using through-silicon-via (TSV) technology. The base die layer is generally configured as the logic base, or LoB. The logic layer consists of multiple components that provide both external link access to the HMC device as well as internal routing and transaction logic. The data storage in an HMC module is organized in a three-dimensional manner into multiple quad units. Figure 95(a) shows a schematic of a 4Gb HMC quad unit. As shown in the figure, one quad unit consists of four vault units. A vault is a major organization unit of data storage in an HMC that vertically spans each of the memory layers (a total of four memory layers in the example shown) using TSVs. Each vault has two 128Mb sized banks on each memory layer. Each bank has 16384 rows with each row of size 8Kb (Figure 95(b)). The rows in a bank are grouped into 32 subarrays. To reduce the capacitance of the memory access path. The subarrays

are further divided into 16 tiles with each tile having a 512×512 array of cells. The physical layout of subarrays and tiles are similar to that found in commodity 2D DRAMs [225]. As shown in Figure 95(a), each vault has a vault controller on the logic base layer that manages all memory reference operations within that vault. A vault controller is analogous to a DIMM controller. The refresh operations of each vault are managed by its corresponding refresh controller.

12.4. MINIMIZING REFRESH OVERHEAD IN 3D DRAMS

The main idea behind minimizing DRAM refresh overhead is to increase DRAM availability for non-refresh operations, so that memory performance can be improved. All existing DRAM architectures support *all-bank refresh* [202] where the corresponding rows in all banks within a rank are refreshed concurrently. This technique increases DRAM availability by reducing the refresh cycle time, and may increase DRAM performance, depending on the workload characteristics. But it is inefficient and less reliable for 3D-stacked DRAMs, because as discussed in [198] and [213], the smaller size and increased physical density of banks increases the noise level of the power delivery network (PDN) when concurrently refreshing the corresponding rows in all the banks. The increased noise level in the PDN often leads to malfunctions that result in erroneous DRAM operation, making the *all-bank refresh* method highly undesirable for 3D-stacked DRAMs.

In contrast, in *per-bank refresh* only one bank per rank is refreshed during a refresh command. All other non-refreshing banks are available during per-bank refresh. Therefore, this technique allows the non-refresh (i.e., read/write) operations targeted at non-refreshing banks to be overlapped with refresh operation of the refreshing bank. In 3D-stacked DRAMs such as the HMC, there exists greater bank-level parallelism than 2D DRAMs because of the fine-grained structure of their data array, creating even more options for overlapping refresh and non-refresh

operations. For example, in HMC, the data array is divided into multiple vaults, with each vault having multiple ranks and partitions; and it is possible to overlap refresh and non-refresh operations across multiple vaults, across ranks within a vault, and across banks within a rank. However, in traditional *per-bank refresh*, all rows in the refresh bundle are refreshed sequentially, which increases refresh cycle time by a factor equal to the refresh bundle size, and leads to inefficiencies. *Moreover, reducing the energy overhead of refresh in HMC still remains a critical problem to improve the energy-efficiency of the memory subsystem*, due to the greater power density of HMC, as discussed in Section 12.1. The next section describes our approach to overcome these challenges.

12.5. MASSED REFRESH: OVERVIEW

12.5.1. CONCEPT AND IMPLEMENTATION

In this subsection, our proposed *massed refresh* technique is explained, including a description of how the control logic and vault peripherals required to enable this technique are implemented. The following descriptions of the concept, implementation, and experimental results are presented for a 1Gb vault of the example HMC design explained in Section 12.3. For clarity, it is imperative to first briefly discuss the function of the refresh controller for *distributed per-bank refresh* [25], [203], *distributed all-bank refresh* [203], and *scattered refresh* [225], before going into the details of how it functions for our proposed *massed refresh* technique.

An HMC has a low data cell retention period of 32ms, due to the increased power density and operating temperature of its 3D-stacked memory dies. A total of 8192 refresh commands must be issued by a refresh controller in 32ms under a distributed refresh scheme, which sets the refresh command interval (tREFI) for HMC to be 3.9µs. Also, for a vault capacity of 1Gb and an 8Kb row size, refresh bundle size is 16. Typically, a refresh controller consumes the row address supplied

by an internal address counter and identifies a refresh bundle based on the implemented refresh scheme. For every refresh command, the starting value of the internal address counter is the number of refresh commands sent so far in the current refresh/retention period. In the *distributed per-bank refresh* scheme, for each refresh command, the refresh controller determines the address of the first row to be refreshed from the starting value of the internal address counter. Then, the address calculator increments the row address by one every time to determine the address of the next row to be refreshed. As an example, suppose that the first row to be refreshed in a distributed refresh command is row R-5632 in subarray SA-12 of bank B-0 on layer L-0. Therefore, the next row to refresh would be row R-5633 in the same subarray SA-12. Hence, in *distributed per-bank refresh*, as shown in Figure 96(a), all rows of a refresh bundle map to a single subarray. For brevity, Figure 96 considers a smaller refresh bundle of 4 rows for a DRAM system with 4 banks, but a similar trend holds for a refresh bundle size of 16 and a memory system with more than 4 banks as well.

In the *distributed all-bank refresh* scheme, as shown in Figure 96(e), the corresponding rows (R-5632) in all the banks (B-0 to B-3) are refreshed simultaneously, which reduces the refresh cycle time (tRFC) by several cycles. For this scheme, if the refresh bundle size (k) is equal to the number of banks (n), each bank will have only one row to refresh. Similarly, if k is greater than n, each bank will have k/n rows to refresh. Therefore, in the case of *distributed all-bank refresh*, the row address is incremented only when $k>n$.

On the other hand, in the case of *scattered refresh* [225], the address calculator increments the address counter by 8192 so that every other row of the refresh bundle is mapped to a different subarray [225]. This arrangement in conjunction with the provision of a dedicated address latch to each individual subarray ends up improving subarray-level parallelism. Hence, as shown in Figure

96(b), for *scattered refresh* the refresh operations of rows that are mapped to different subarrays are overlapped in time, which reduces tRFC by several cycles. In general, for *distributed per-bank refresh* and *scattered refresh*, the address calculator increments the row address counter by 1 and by 8192 respectively for 16 times during a refresh command before moving on to the next refresh command after the tREFI interval. More information on these refresh techniques can be found in [25], [202], [225].

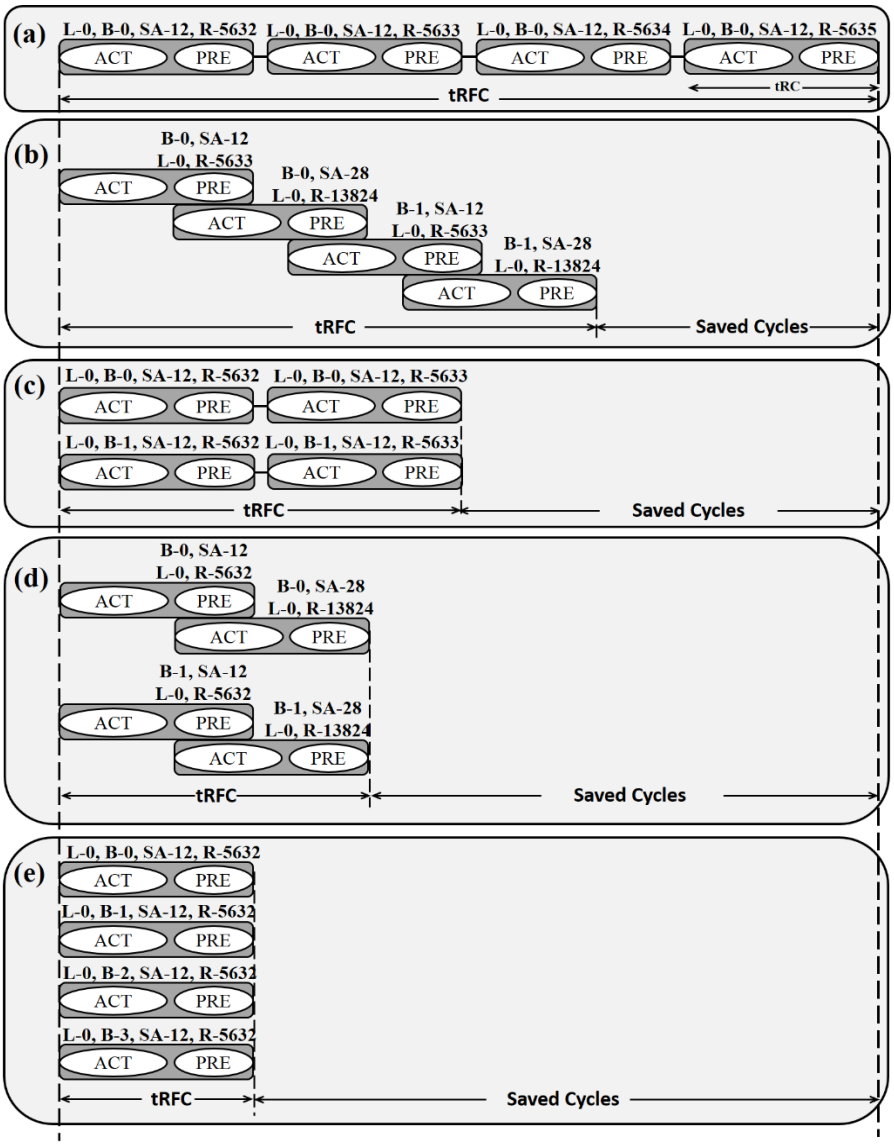


Figure 96: Refresh cycle for (a) distributed per-bank refresh, (b) scattered refresh, (c) crammed refresh, (d) massed refresh, (e) distributed all-bank refresh.

Our proposed *massed refresh* technique builds on the concept of subarray-level parallelism in *scattered refresh* by leveraging additional bank-level parallelism during a refresh command in the DRAM device. We note that as the global row address decoder is shared among all the subarrays of a bank, the refresh operations of rows that are mapped to different subarrays of a single bank are not completely overlapped in time in *scattered refresh*, which reduces efficiency. Figure 97 illustrates how bank-level parallelism is exploited in our proposed scheme to improve upon *scattered refresh*. As shown in the physical address latch of the figure, the physical row address is divided into three parts: a 14-bit row address, a 2-bit layer address, and a 1-bit bank address. Before routing the address bits to individual memory layers using TSVs, the layer address and bank address are decoded using a physical address decoder to obtain layerID (LID) and bankID (BID) respectively. The figure also shows how the physical address latch and the address decoder are shared among the read/write command scheduler and the refresh controller.

During a regular read/write command operation, the LID and BID values direct the row address bits to a particular bank on a particular memory layer. But, during a refresh command operation, the BID is masked by the refresh controller so that the row address is directed to both the banks on the target memory layer. *This arrangement simultaneously refreshes two rows in two banks on the selected layer for each physical address generated by the address calculator*. In other words, a refresh bundle is divided into two equal subgroups and both subgroups are refreshed concurrently. The refresh operations of these two subgroups are completely overlapped in time. Due to this bank-level parallelism exploited in our approach, the address counter value needs to be incremented only for 8 times to refresh a bundle of 16 rows.

The bank-level parallelism obtained by masking the BID bits can be used in two ways: one with the added subarray-level parallelism and the other without it. For clarity, we refer to our

refresh technique with the added subarray-level parallelism as the *massed refresh* technique and our technique without the added subarray-level parallelism is referred to as *crammed refresh*. As shown in Figure 96(c), in *crammed refresh*, the example refresh bundle of 4 rows is divided in two subgroups with two rows (R-5632 and R-5633) in each subgroup. These two subgroups are mapped on bank B-0 and bank B-1 on the memory layer L-0, and their refresh is completely overlapped in time, which reduces the refresh cycle time (tRFC) by a large number of cycles. However, rows R-5632 and R-5633 are refreshed sequentially in both subgroups, because they are mapped on the same subarray SA-12.

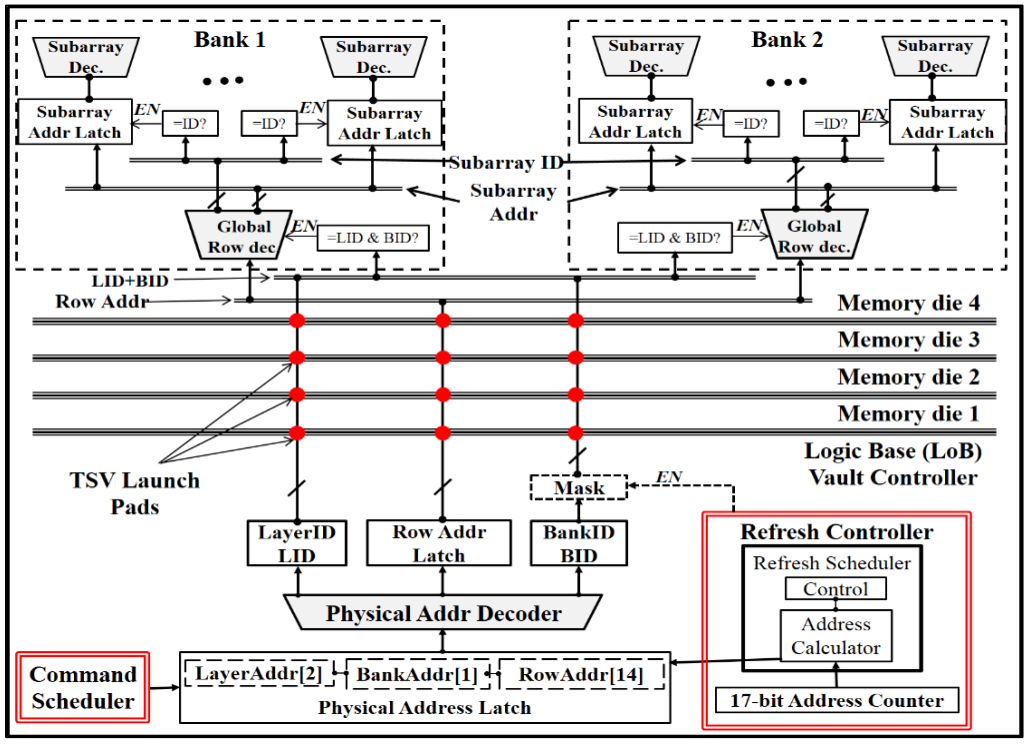


Figure 97: Schematic implementation of control logic and peripheral circuits for the bank-level and subarray-level parallelism of our proposed *massed refresh* technique.

In contrast, as shown in Figure 96(d), in *massed refresh*, the constituent rows R-5632 and R-13824 of both the subgroups (which are mapped on bank B-0 and bank B-1 on layer L-0) belong to two different subarrays, subarray SA-12 and SA-28 respectively. Therefore, refresh operations

of the constituent rows of both subgroups are further overlapped in time. Hence, *massed refresh* exploits the subarray-level parallelism in addition to the bank-level parallelism of *crammed refresh* to reduce tRFC even further. The main difference between *all-bank refresh* and *massed/crammed refresh* is that the *massed*/ *crammed refresh* methods are restricted to refresh only two banks in parallel. The smaller degree of bank-level parallelism in *crammed refresh* and *massed refresh* sets the PDN noise levels to more acceptable level, which makes them more efficient and desirable than the *all-bank refresh* method.

12.5.2. REFRESH CYCLE TIME AND OVERHEAD ANALYSIS

In this subsection, we present refresh cycle time calculations for the *crammed refresh* and *massed refresh* techniques along with those for the state-of-the-art *distributed per-bank refresh, distributed all-bank refresh,* and *scattered refresh* schemes, before analyzing their overheads.

In general, refresh cycle time (tRFC) is equal to the product of refresh bundle size (*k*) and row cycle time (tRC) with some additional recovery time (tREC) [195]. Consequently, tRFC for *distributed per-bank refresh* is (tRC*16) + tREC (as *k*=16). Typically, tRC is a sum of row access cycle time (tRAS) and precharge delay (tRP). As explained in [225], the subarray-level parallelism of the *scattered refresh* scheme hides tRP time from the tRC time of each row whose refresh operation is overlapped with another row's refresh during a refresh command. For *scattered refresh*, in a refresh bundle of 16 rows, refresh operations of only 15 rows are overlapped. Hence, tRFC for *scattered refresh* is (tRAS*15) + tRC + tREC. For *all-bank refresh*, each of the total 8 banks (*n*=8) has two rows to refresh (as *k*=16). So, tRFC for *all-bank refresh* is (tRC*2) + tREC = 80ns. In a similar manner, we can calculate tRFC for our *crammed refresh* and *massed refresh* techniques.

The calculated tRFC values for all the aforementioned refresh techniques are shown in Table 23. We consider a row cycle time (tRC) of 35ns for the example 1Gb vault of HMC (shown in Figure 95) by modeling the vault using CACTI-3DD [205] with 50nm technology parameters. Our choice of this technology node is motivated by Micron demonstrating a 4Gb HMC quad fabricated in 50nm technology [23]. We assume a recovery time (tREC) of 10ns [195] in our calculations of tRFC. It should be noted that performing refresh operations on more than one row simultaneously draws extra current from the PDN. So, it is important to ensure that the proposed *crammed refresh* and *massed refresh* techniques do not overshoot the current delivering capacity of the PDN. To investigate this aspect, we also calculated the required peak refresh current for the various refresh techniques using their 50nm technology parameter based architectural models developed in CACTI-3DD [205]. These values are shown in Table 23.

Table 23: Refresh cycle time (tRFC) and peak refresh current for state-of-the-art and proposed DRAM refresh techniques.

	Per-bank	**Scattered**	**Crammed**	**Massed**	**All-bank**
tRFC (ns)	570	420	290	220	80
Refresh current (mA)	130.4	178.1	260.7	347.6	1290.5

The peak current capacity of the 4Gb HMC quad design demonstrated by Micron was shown to be 9.2A [23], which implies that the peak current capacity of a 1Gb HMC vault can be reasonably estimated to be 2.3A. Table 23 shows that *all-bank refresh* yields the smallest tRFC. However it also has a prohibitively high peak refresh current of 1290.5mA, along with a propensity to increase PDN noise when concurrently refreshing the corresponding rows in all the banks [198] [213]. Due to these practical limitations, *all-bank refresh* is not well suited for 3D-stacked DRAMs such as HMC.

The peak refresh current for the remaining refresh techniques is significantly lower than for *all-bank refresh*. Our *massed refresh* technique has a peak refresh current of 347.6mA, which is well below the current delivering capacity of an HMC vault. Therefore, the proposed *crammed refresh* and *massed refresh* techniques can be easily implemented without overshooting the current delivery capacity in HMC or creating PDN noise issues. Moreover, our proposed crammed refresh and massed refresh techniques reduce refresh cycle time much more effectively compared to the other refresh techniques, as shown in Table 23. Also, the HMC specification defines separate pins to externally supply the DRAM wordline boost voltage V_{PP}, which relaxes the need for on-chip charge pumps for the V_{PP} supply. Dedicated pins for V_{PP} supply increase charge supplying capacity of an HMC vault, easily allowing parallel activation of two or more rows in a vault, as required in our proposed refresh techniques.

Similar to *scattered refresh*, our proposed *crammed refresh* and *massed refresh* techniques would need a 13-bit counter to keep track of the number of refresh commands in a refresh period for the example 1Gb HMC vault, along with a 17-bit internal address counter. These counters, when implemented at the 50nm technology node, consume 170μm^2 and 230μm^2 area respectively. Thus, the area overhead of the changes in the peripherals required for subarray-level and bank-level parallelism is negligible, similar to [225].

12.6. EXPERIMENTAL RESULTS

We performed trace-driven simulation analysis to compare our proposed *crammed refresh* and *massed refresh* techniques with the state-of-the-art *distributed per-bank refresh* [25] and *scattered refresh* [225] techniques. We do not compare our proposed refresh techniques against the *all-bank refresh* technique because it is impractical to implement for 3D DRAMs as discussed in Section 12.4 and 12.5.2.

Memory access traces for the PARSEC benchmark suite [76] were extracted from detailed cycle-accurate simulations using the gem5 full-system simulator [77]. Table 24 gives the configuration of the gem5 simulator that was used for this study. We considered twelve different applications from the PARSEC benchmark suite: *Blackscholes (BS), Bodytrack (BT), Canneal (CN), Dedup (DD), Facesim (FS), Ferret (FR), Fluidanimate (FA), Freqmine (FM), Streamcluster (SC), Swaptions (SW), Vips (VP),* and *x264 (X2).* We ran each PARSEC benchmark for a "warm-up" period of 1 billion instructions and captured memory access traces from the subsequent 1 billion instructions executed. These memory traces were then provided as inputs to the DRAM simulator DRAMSim2. We heavily modified DRAMSim2 to model *crammed refresh*, *massed refresh*, and the refresh techniques from prior work for comparison. Throughput and energy-delay product (EDP) values for the memory subsystem for all of these refresh techniques considered were obtained from this DRAMSim2 simulator. We performed timing and energy analysis by adapting the code for CACTI-3DD [205] to model the 1Gb HMC vault at the 50nm technology node. These parameters for the HMC architecture are summarized in Table 25. All TSVs in this study were modeled based on ITRS projections for intermediate interconnect level TSVs.

Table 24: Gem5 simulation configuration.

#Cores	4 ARM	L2 Coherence	MOESI
L1 I Cache	16KB	Frequency	5 GHz
L1 D Cache	16KB	Issue Policy	In-order
L2 Cache	128KB	# Memory Controllers	1

Table 25: Row access time (tRAS), row cycle time (tRC), activation-pre-charge energy (ActPreE), read energy (ReadE), and background power (BGP) parameter values for HMC.

tRAS (ns)	tRC (ns)	ActPreE (nJ)	ReadE (nJ)	BGP (mW)	Data bus width
25	35	1.8	2.7	11	128b

Figure 98 shows throughput values for the various refresh techniques across the PARSEC benchmarks. A rank-based round-robin scheduling scheme, rank:row:col:bank address mapping scheme, and open page policy were used for all simulations. It can be observed that *massed refresh* achieves 6.3% more memory throughput on average over *per-bank refresh* and *scattered refresh*. More specifically, the proposed *massed refresh* technique achieves 8.4%, 4.3%, and 1.4% more memory throughput on average over *per-bank refresh*, *scattered refresh* and *crammed refresh* respectively. *Massed refresh* has 62.5%, 47.6% and 24.1% less refresh cycle time over *per-bank refresh*, *scattered refresh* and *crammed refresh* respectively, which translated into the highest throughput for *massed refresh* over other refresh schemes.

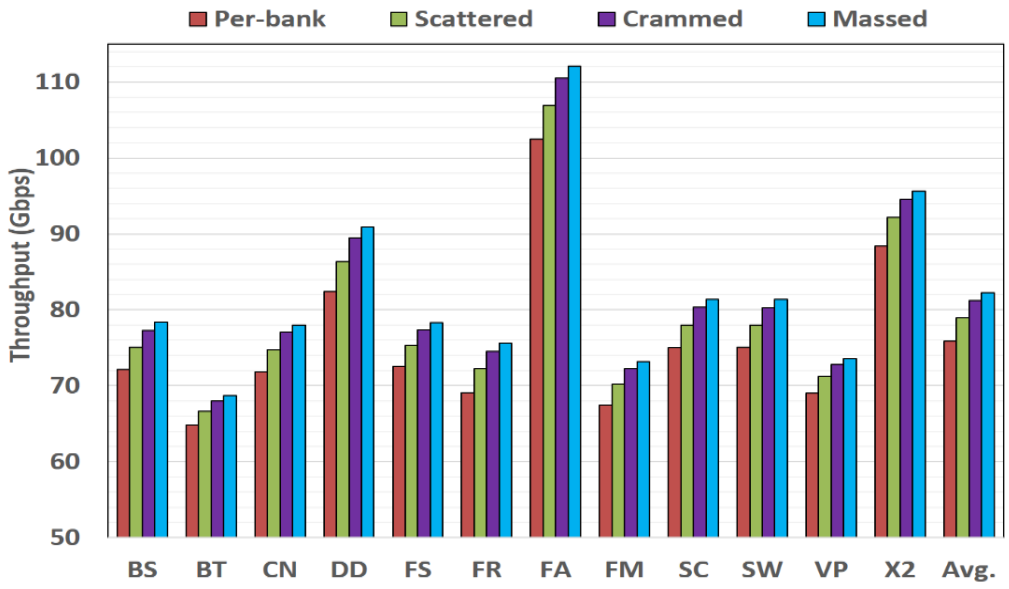


Figure 98: Memory throughput for various refresh schemes across PARSEC benchmarks.

Figure 99 shows the energy-delay product (EDP) values for the various DRAM refresh techniques across the PARSEC benchmarks. EDP values were calculated by multiplying the average-energy with average-latency for memory accesses. Average-energy in each case was calculated by dividing the total energy by the total number of memory transactions, and the

average-latency was calculated by dividing the total latency by the total number of memory transactions. It can be seen that *massed refresh* achieves 5.8% less EDP on average over *per-bank refresh* and *scattered refresh*. More specifically, our proposed *massed refresh* technique achieves 7.5%, 3.9% and 1.2% less EDP on average over *per-bank refresh*, *scattered refresh*, and *crammed refresh*, respectively. As was shown in Table 23 earlier, even though the peak refresh current (power) required for *massed refresh* is higher than all the other refresh techniques (except *all-bank refresh*), the total energy spent in refresh operations during a given amount of time remains the same for all the refresh techniques (except *all-bank refresh*). Moreover, the increased throughput of *massed refresh* translates into the least average-energy. The least average-energy and the highest throughput result in the best improvements in EDP for *massed refresh* among all of the DRAM refresh techniques, making it the most energy-efficient refresh technique compared to the other techniques.

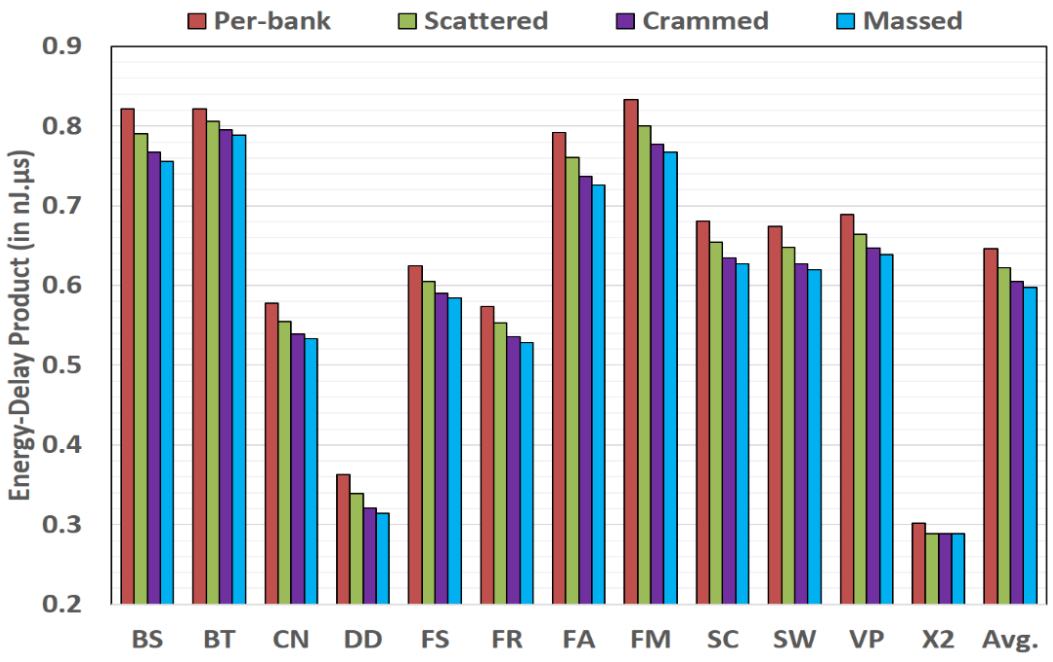


Figure 99: Energy-delay product for various refresh methods.

12.7. CONCLUSIONS

In this chapter, we proposed a new refresh technique for the 3D-stacked Hybrid Memory Cube (HMC) DRAM called *massed refresh* that yields 6.3% and 5.8% improvements in throughput and EDP on average over the JEDEC standardized *distributed per-bank refresh* and the state-of-the-art *scattered refresh* schemes. These promising results indicate that our *massed refresh* technique can significantly reduce the overhead of distributed refresh operations in the HMC (and other DRAMs) by improving their throughput and energy-efficiency.

13. DYPHASE: A DYNAMIC PHASE CHANGE MEMORY ARCHITECTURE WITH SYMMETRIC WRITE LATENCY AND RESTORABLE ENDURANCE

A major challenge for the widespread adoption of phase change memory (PCM) as main memory is its asymmetric write latency. Generally, for a PCM, the latency of a SET operation (i.e., an operation that writes '1') is 2-5 times longer than the latency of a RESET operation (i.e., an operation that writes '0'). For this reason, the average write latency of a PCM system is limited by the high-latency SET operations. This chapter presents a novel PCM architecture called *DyPhase*, which uses partial-SET operations instead of the conventional SET operations to introduce a symmetry in write latency, thereby increasing write performance and throughput. However, use of partial-SET decreases data retention time. As a remedy to this problem, *DyPhase* employs novel distributed refresh operations in PCM that leverage the available power budget to periodically rewrite the stored data with minimal performance overhead. Unfortunately, the use of periodic refresh operations increases the write rate of the memory, which in turn accelerates memory degradation and decreases its lifetime. *DyPhase* overcomes this shortcoming by utilizing a proactive in-situ self-annealing (*PISA*) technique that periodically heals degraded memory cells, resulting in decelerated degradation and increased memory lifetime. Experiments with PARSEC benchmarks indicate that our *DyPhase* architecture based hybrid DRAM-PCM memory system, when enabled with *PISA*, yields orders of magnitude higher lifetime, 8.3% less CPI, and 44.3% less EDP on average over other hybrid DRAM-PCM memory systems that utilize PCM architectures from prior works.

13.1. INTRODUCTION

For decades since its emergence, dynamic random-access memory (DRAM) has supported the demands on main memory capacity and performance. In today's Big Data era, next generation memory systems must be capable of offering the memory capacity required by large data structures [190]. However, scaling DRAM below 22nm to increase capacity is currently a major challenge [31], and at 22nm, DRAM dissipates a large amount of leakage power [233]. These limitations make DRAM less suitable for next generation main memory, especially with the proliferation of Big Data applications.

Recent advances have enabled Phase Change Memory (PCM) as a leading technology that can alleviate the leakage and scalability problems of traditional DRAM [234]. Compared to DRAM, PCM has non-volatility, superior scalability, lower standby leakage power, and comparable read latency. Single-Level-Cell (SLC) PCM differentiates between two resistance levels of a phase change material to store a logic bit (logic '0' or '1'). An SLC PCM cell also requires different durations and strengths of programming current to write '0' and '1'. The RESET operation, which writes '0' to a PCM cell, uses a short but high-amplitude current pulse to program the phase change material to the amorphous state that has high resistance. In contrast, the SET operation, which writes '1' to a PCM cell, utilizes 2–5× longer and 2–4× lower-amplitude current pulses to program the phase change material to the polycrystalline state that has lower resistance. When a page (or line) of data is written to PCM, the longer SET operation limits the write latency. The resultant longer average write latency in PCM may incur as much as 60% performance degradation [235]. Therefore, it is critical to reduce the inherently long write latency of PCM, if PCM is to become a viable replacement for DRAM in the next generation of memory systems.

In recent years, many techniques have been proposed to minimize the effect of longer write latency on PCM performance. These methods either hide longer write latency by scheduling writes among idle bank cycles [235]-[238] or provide architecture-level solutions for reducing write latency in multi-level cell (MLC) PCMs [239]-[243]. The write scheduling methods are not suitable for high-performance systems where there are few-to-no idle cycles between memory accesses, whereas the architecture-level techniques presented in [239]-[243] are specific to MLC PCMs, which is less preferred over SLC PCMs due to reliability and variation susceptibility issues. Thus, neither of these approaches is general enough to address the long write latency in PCMs. Some other techniques (e.g., [244]-[246]) have been proposed that utilize latency-aware coding schemes to encode data words, which relax the need to write '1' bits in some write operations, resulting in a reduced average write latency. However, these methods cannot eliminate the need to write '1' bits in every write operation, thus, limiting the improvement in average write latency.

This chapter presents a novel PCM architecture called *DyPhase*, which uses partial-SET operations instead of the conventional SET operations to introduce a symmetry in write latency. *DyPhase* builds upon the partial-SET PCM architecture presented in [247] and overcomes its shortcomings. To remedy the problem of shorter data retention time of partial-SET PCM bits, *DyPhase* employs novel distributed refresh operations in PCM that leverage the available power budget to periodically rewrite the stored data with minimal performance overhead. Unfortunately, the use of periodic refresh operations increases the write rate of the memory, which in turn accelerates memory degradation and decreases its lifetime. *DyPhase* overcomes this shortcoming by utilizing a proactive in-situ self-annealing (*PISA*) technique that periodically restores degraded memory cells, resulting in decelerated degradation and increased memory lifetime. In summary, the novel contributions of this chapter are:

- A dynamic phase change memory (PCM) architecture called *DyPhase* that introduces a symmetry in PCM write latency by using partial-SET operations to write '1's;
- A distributed refresh method called *Reset-pSet Refresh* as a part of *DyPhase* architecture, that periodically restores the data bits in *DyPhase* PCM to ensure reliable data retention;
- An improved refresh method called *O-pSet Refresh* that refreshes only the partial-SET bits in *DyPhase* PCM to achieve better performance than the *Reset-pSet Refresh;*
- A proactive in-situ self-annealing (*PISA*) technique that decelerates the degradation of *DyPhase* PCM cells and improves their lifetime.

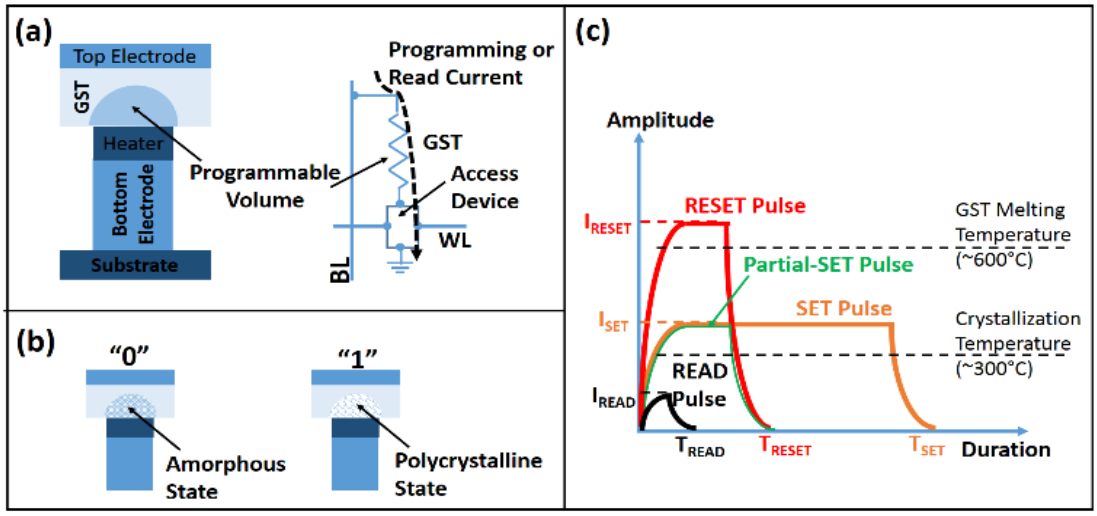


Figure 100: (a) The basic structure of PCM cell, (b) different states of PCM cell, (c) programming (write) and read pulses for PCM cell.

13.2. BACKGROUND ON PHASE CHANGE MEMORY

As shown in Figure 100(a), a PCM storage cell is comprised of two electrodes separated by a resistive heater element and phase change material, which is typically a chalcogenide material $Ge_2Sb_2Te_5$ (GST) [248]. The bottom electrode is connected to the heater element at its one end, whereas it is connected to an access device at the other end. The access device is typically one of the following three devices: a standard NMOS transistor, a bipolar junction transistor (BJT), or a

PN-junction diode. As shown in Figure 100(b), the GST can be switched between two states (polycrystalline or amorphous) with dramatically different electrical resistance. The amorphous high-resistance (usually in the MΩ range) state is used to represent a binary '0', while the crystalline low-resistance (usually in the KΩ range) state represents a '1'.

A PCM memory system has three basic operations: read, SET, and RESET. A PCM cell can be read by simply sensing the current flow through it. Due to the large gap between the two resistance levels ('0' and '1') of the GST material, the sensed currents of these two states differ by a large magnitude. The read operation latency is typically tens of nanoseconds.

As shown in Figure 100(c), in the write operations, different heat-time profiles are applied to switch cells from one state to the other. To RESET a PCM cell, a strong programming current pulse (I_{RESET}) of short duration (T_{RESET}) is required. This programming pulse raises the temperature of the GST material to its melting point, after which the pulse is quickly terminated. Subsequently, the small region of melted material cools quickly, leaving the GST material programmed in the amorphous state. As the region of the GST material that melts due to the I_{RESET} pulse is small, the required T_{RESET} is short (tens of nanoseconds). In contrast, to SET a PCM cell, a programming current pulse of a longer duration T_{SET} and weaker strength I_{SET} is applied to program the cell from the amorphous state to the polycrystalline state. For the SET operation, the temperature of the GST material is raised above its crystallization temperature (300 °C) but below its melting point (600 °C) for a sufficient duration of time. As the crystallization rate is a function of temperature, and given the variability of PCM cells, reliable crystallization requires a programming pulse of hundreds of nanoseconds [234]. Thus, the SET latency of a PCM cell is longer than both the RESET latency and the read latency.

From a system perspective, it is highly likely that both SET and RESET operations occur in a typical PCM write, where hundreds of bits are programmed in PCM cells. For this reason, the average write latency of a PCM system is limited by the slower SET operations. Therefore, the write latency in PCM memory is longer than the read latency.

13.3. RELATED WORK

In recent years, a compelling body of research has been conducted that aims to minimize the effect of longer write latency on PCM performance. The PCM latency improving methods presented in the literature can be broadly classified into the following four categories: *(i)* methods that optimize DRAM-PCM hybrid memory architectures to reduce the number of write operations to PCM [249]-[252]; *(ii)* methods that hide longer write latency by scheduling PCM writes among idle bank cycles [235]-[238]; (iii) architecture-level solutions for reducing write latency in MLC PCMs [239]-[243]; and *(iv)* methods that utilize latency-aware data coding schemes to relax the need for writing logic '1' bits in some write operations, thereby reducing the average write latency [244]-[246].

The prior works (e.g., [249]-[252]) that utilize DRAM-PCM hybrid memory systems, in general, optimize the memory space and page allocation between the DRAM and PCM parts of the main memory to reduce the number of write operations to the high-latency PCM part. In [251], Lee et al. present one of the earliest works on DRAM-PCM hybrid memory, which provides a comprehensive design space exploration of DRAM-PCM hybrid memory systems from the perspective of energy-delay efficiency. In [250], Khouzani et al. utilize DRAM as a cache to PCM and propose a DRAM page replacement algorithm along with a conflict-aware page remapping strategy to reduce the number of DRAM misses and the resultant write backs to PCM. Lee et al. [251] propose a write-history aware page replacement algorithm for hybrid DRAM-PCM

architectures that estimates future write references based on write history, and then absorbs frequent writes into DRAM. Zhang et al., in [252], propose a write-back aware last-level cache management scheme for the hybrid DRAM-PCM main memory, which improves the cache hit ratio of PCM blocks and minimizes write-backs to PCM.

A higher write latency can be masked using a DRAM-PCM hybrid memory system with intelligent page allocation and scheduling as long as there is sufficient write bandwidth [235]. Thus, a DRAM-PCM hybrid memory system draws forth an untrue masked behavior of the PCM sub-system, as the DRAM part of the hybrid system hides the longer write latency of the constituent PCM part. However, as explained in [235], the inherently longer latency of PCM write-back accesses, due to DRAM misses, may stall subsequent read accesses, significantly increasing average read latency of the hybrid system. As read accesses are latency critical, increasing read latency has significant performance impact [235]. Therefore, improving the true unmasked write latency of PCM accesses is imperative for improving the overall memory performance. For this reason, we focus on minimizing the unmasked write latency of PCM accesses in this chapter.

To reduce the unmasked write latency of PCM, some prior works (e.g., [235]-[238]) tend to schedule write requests during idle bank cycles when the target banks are not serving any other requests. In [235], Qureshi et al. present write preemption that preempts the on-going write operation to serve a newly arrived read request to the same bank, thereby reducing the effect of longer write latencies on average read latency of the PCM system. Kim et al., in [237], propose to overlap the resistance drift latency of some write operations with concurrent read operations, thereby achieving significant benefits in overall system performance. Qureshi et al., in [236], exploit the property of PCM devices where PCM write latency is longer than read latency only because of high-latency SET operations. They propose an architectural technique called PreSET

that proactively SETs all the bits in a given memory line during idle bank cycles as soon as the line becomes dirty in the cache. Thus, subsequent write operations to the line require only RESET operations, which incur much lower latency. These write scheduling methods are efficient in hiding longer write latency, but cannot reduce the fundamental latency of every write access. Therefore, in this work, we aim to reduce the fundamental latency of every PCM write access.

Some other architecture-level solutions have been proposed for reducing write latency in multi-level cell (MLC) PCMs. These techniques are specific to MLC PCMs and they are not general enough to be applicable for single-level cell (SLC) PCMs. An MLC PCM, which stores multiple bits (represented by multi-level resistance) in a single cell, offers high density with low per-byte fabrication cost [239]. However, due to cell process variations and the relatively small differences among resistance levels, MLC PCM typically employs an iterative write scheme to achieve precise control, which suffers from large write access latency [239]. Moreover, the susceptibility to variations renders MLC PCM less reliable than SLC PCM, making it less preferable over SLC PCM. Therefore, we focus on optimizing SLC PCM in this chapter.

Some other techniques (e.g., [353]-[355]) have been proposed that utilize latency-aware coding schemes to encode data words, which relax the need to write '1' bits in some write operations, resulting in a reduced average write latency. Cho et al., in [244], propose Flip-N-Write, which on every write request, updates only those bits of the new data word that differ from the original data word. Flip-N-Write also limits the required number of bit updates to half of the data word size by "flipping" (inverting) the bit values of the new data word if the number of to-be-updated bits is over half the data word size. As a result, Flip-N-Write can achieve 2× write bandwidth by doubling the write unit size without increasing the instantaneous write current. In [245], Yue et al. exploit a property of PCM cells that writing a '1' (SET operation) takes longer

time but a smaller amplitude current than writing a ‘0’ (RESET operation). They propose Two-Stage Write, wherein a write is divided into two stages: in the write-0 stage, all zeros are written at an accelerated speed, and in the write-l stage, all ones are written with increased parallelism, without violating power constraints. Two-Stage Write achieves better resource utilization and reduces the service time of writing a cache line. Li et al., in [246], propose write-once-memory (WOM) code PCM architecture (referred to as *WOMC_PCM* henceforth), wherein they encode the PCM data words using a $<2^2>^2/3$ WOM-code. A “$<v>^t/n$ WOM-code” is a coding scheme that uses n “write-once bits” to represent one of v values so that the WOM can be written a total of t times by using only RESET operations. Therefore, $<2^2>^2/3$ WOM-code used in WOMC_PCM [246] utilizes a 3-bit code to represent one of four two-bit values that can be written a total of 2 times by using only RESET operations. Thus, WOMC_PCM architecture reduces the necessity of using SET operations (to write ‘1’s) during some writes, thereby reducing the latency for those writes resulting in significantly reduced average write latency. However, these methods (Flip-N-Write [244], Two-Stage Write [245], and WOMC_PCM [246]) cannot eliminate the need to write ‘1’ bits (the need to use SET operations) in every write operation, thus, limiting the achievable improvement in average write latency.

Different from all these works, this chapter presents a novel PCM architecture called *DyPhase*, which uses partial-SET operations instead of the conventional SET operations to introduce a symmetry in write latency. *DyPhase* improves upon the partial-SET PCM architecture presented in [356], a detailed description of which is given in the next section.

13.4. BACKGROUND: PARTIAL-SET OPERATIONS

In a typical PCM, SET bits have about 1000× less resistance than RESET bits. However, the resistance contrast of only 10×, which corresponds to 10× contrast in read sense current, is

sufficient to detect a logic '1' bit as separate from a logic '0' bit. As described in [247] and [253], the resistance of a RESET bit (logic '0') can be decreased by 8-10× by applying a SET current pulse of I_{SET} amplitude but only of T_{RESET} duration. In other words, a SET pulse of T_{RESET} duration can decrease the resistance of a logic '0' bit by 8-10× to program it as a logic '1' bit. We refer to the SET pulse of I_{SET} magnitude and T_{RESET} duration as a *partial-SET* pulse and the corresponding cell-programming event as a partial-SET operation. Similarly, the conventional SET operation with current pulse of I_{SET} amplitude and T_{SET} duration is referred to as a *full-SET* operation henceforth. Thus, the latency of SET operations can be reduced to be equal to the latency of RESET operations by using partial-SET pulses instead of full-SET pulses, which can achieve the same performance from a system perspective as can be achieved by the ideal write operations with symmetric latency. However, as described in [247], there exists a reliability challenge with partial-SET operations.

Due to the resistance drift caused by the thermally activated atomic rearrangement of the amorphous structure [247], [254], the resistance of a partial-SET PCM cell increases with time. The phenomenon of resistance drift exhibits a power-law model, $R_t = R_0 \times t^v$, where R_0 is the initial resistance of cells after a write, t is the elapsed time (in seconds), and v is the drift exponent. As a RESET cell has greater R_0 than a SET/partial-SET cell, a RESET cell exhibits greater drift rate. Nevertheless, resistance drift in a RESET cell does not cause readout error, as all cell resistance values above the specified boundary threshold (reference resistance that distinguishes the RESET state from the SET/partial-SET state) represent the RESET state. In contrast, when R_t of a partial-SET cell crosses the reference resistance due to drift, it erroneously represents the RESET state. As described in [247], the partial-SET cells give readout errors due to resistance drift for elapsed time of about 4.9 seconds (the worst case). Thus, the retention time (elapsed time after which

resistance drift causes readout errors) of partial-SET cells is about 4.9 seconds. However, from [255], the drift exponent v and retention time of cells depend on temperature, type and thickness of chalcogenide material, and the distance of programmed cell-resistance R_0 from the reference resistance value. The exact value of R_0 after a cell write depends on variations in the manufacturing, programming, and crystallization processes [255]. Therefore, to account for the impact of these variations on the drift rate of partial-SET cells, we conservatively choose a smaller value of 4s for the retention time of partial-SET cells.

To improve retention related reliability of partial-SET cells, Li et al. in [247] proactively schedule full-SET write operations on partial-SET lines within their retention window. To do that, they propose maintaining a separate partial-SET queue (in addition to the regular write request queue), each entry of which stores the address and elapsed time for a partial-SET line. In their proposed PCM architecture (referred to as *pSET_PCM* henceforth), when the next regular write request becomes issuable and if the partial-SET queue has a spare entry, the line of request is written with partial-SET pulses in the PCM array and a corresponding entry is made in the partial-SET queue with its elapsed time value starting from zero. Eventually, an entry from the partial-SET queue can be released or evicted under three different scenarios. First, an entry can be released from the partial-SET queue if the target bank of the corresponding line in memory is idle and can accept a full-SET write without violating any timing or power constraints. In the second scenario, if the partial-SET queue is full and the newly issued write request arrives, the scheduler first evicts the entry with the largest elapsed time from the partial-SET queue, and issues a full-SET request to the corresponding line in the PCM array. After the full-SET request due to the evicted entry, the newly requested line is written with partial-SET pulses in the PCM array and the corresponding entry is queued in the newly vacated location in the partial-SET queue. In the third scenario, if the

partial-SET queue is full and no new write requests arrive for a long time, the elapsed time for one or more entries in the queue may reach the retention time of 4 seconds. In this scenario, all the entries whose retention times have expired are rewritten in the PCM array with full-SET operations and the corresponding entries from the partial-SET queue are released.

In any scenario, due to reliability concerns associated with the partial-SET lines, full-SET requests due to the evicted/released entries are prioritized over regular read/write requests [247]. Therefore, regular read/write requests are stalled while full-SET requests are being serviced, which increases queuing latency of the stalled requests, resulting in an increased average latency for the system. Thus, the use of the partial-SET queue and having to rewrite the evicted entries with full-SET operations significantly reduces the write latency improvements achieved by the *pSET_PCM* architecture.

One way of dealing with drift-induced errors in partial-SET cells is to reactively correct these errors using a multi-bit error correction support and a lightweight detect-and-scrub mechanism as proposed in [255]. In spite of being more effective than device-level solutions at handling drift-induced errors, the techniques in [255] incur significant overhead of error detection. In contrast, our proposed *DyPhase* architecture (discussed in the next section) proactively scrubs drift-related errors and eliminates the overhead of error detection. Moreover, our *DyPhase* architecture utilizes the concept of partial-SET operations more effectively to achieve better improvements in PCM latency and performance.

13.5. DYPHASE PCM ARCHITECTURE: OVERVIEW

DyPhase is a dynamic PCM architecture with symmetric write latency and restorable endurance. From a data-organization and structural perspective, *DyPhase* does not differ from other PCM architectures, and hence, it can be implemented on any PCM system with arbitrary

size, structure, and data-organization. The key novel attributes of *DyPhase* architecture are at the micro-architecture level, and are summarized as follows: (i) a distributed refresh method called *Reset-pSet Refresh* (Section V.B), (ii) an improved refresh method called *O-pSet Refresh* with better performance (Section V.C), and (iii) a proactive in-situ self-annealing (*PISA*) technique that improves PCM lifetime (Section VII).

13.5.1. BASELINE PCM ARCHITECTURE

Before diving into the specifics of our *DyPhase* architecture, we present the general structure and data-organization of the baseline PCM system used in this work. *DyPhase* and other PCM architectures in this chapter used for comparison purposes are all implemented on top of this baseline PCM system with their specific microarchitecture-level attributes. Figure 101 depicts the structure and data-organization of this baseline PCM. As shown in Figure 101(a), a 4GB SLC PCM DIMM rank has 8 PCM chips of 512MB size each. The total PCM capacity is divided across 8 logical banks, each of which has 32768 logical rows. Each of these rows is of 16KB in size and is striped across 8 PCM chips. Thus each chip stores 2KB of the row. As shown in Figure 101(b), in each individual chip of a logical bank, the PCM cells are hierarchically organized into blocks and sub-blocks. Peripheral circuitry, such as global decoders, buffers, sense amplifiers (S/As), and write drivers (W/Ds) are shared among blocks and are multiplexed across sub-blocks. For read and write operations, global bitline decoders (GBL_DEC) and local bitline decoders (LBL_DEC) select some bit-lines and connect them to S/As for reading or W/Ds for writing. In the case of a PCM write operation, the write current flows from W/D to the accessed cell's ground line through the local and global bitlines, GST material of the cell, and access device. For a read operation, the read current follows the same path except that it originates from the S/A.

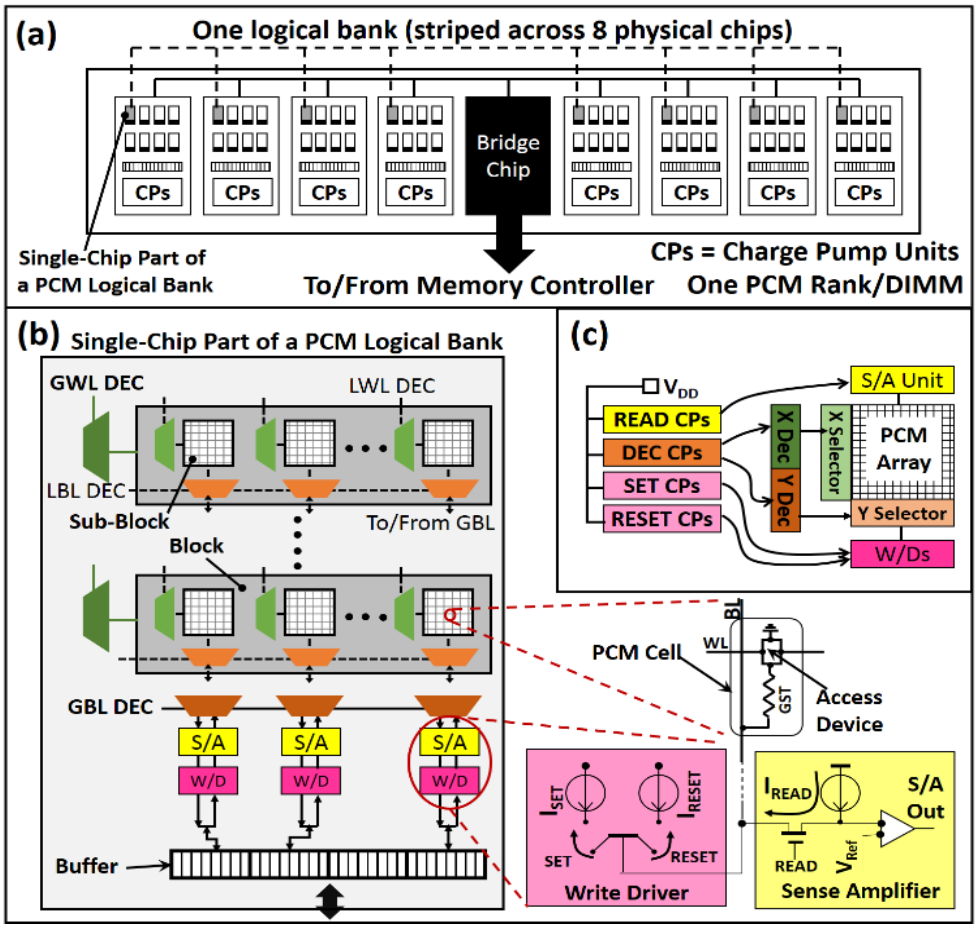


Figure 101: (a) Architecture of the baseline SLC PCM DIMM rank and a logical bank, (b) hierarchical organization of a single-chip part of a PCM logical bank in the baseline PCM rank, (c) the charge pump (CP) system for the baseline PCM array.

Table 26: Various parameters for 11nm diode-switched SLC PCM system of 4GB capacity with eight chips per rank.

	RESET	SET	READ	Partial-SET
Current pulse amplitude (µA)	40	20	4.2	20
Voltage	2.5	1.5	1.5	1.5
Current pulse duration (ns)	50	150	40	50
Energy per bit (pJ)	5	4.5	0.25	1.5
#CPs per chip	4096	8192	16384	8192
Total CP area per chip (mm^2)	3.7	1.2	0.5	1.2
Total power wastage in CPs (mW)	1250	460	190	460

As shown in Figure 101(a), each PCM chip has dedicated charge pumps (CPs), which are responsible for providing large amplitude voltage and currents for READ, RESET, and SET operations. For the baseline PCM system, we assume the use of multi-unit modular CPs as described in [243]. As the amplitudes of currents required for READ, RESET, and SET operations are different, these operations require dedicated multi-unit CPs. Figure 101(c) shows a multi-unit CP system for a PCM array with a dedicated multi-unit CP per PCM operation.

Table 26 gives relevant parameter values (voltage, current, energy, delay etc.) for different operations of the basic diode-switched 11nm 4GB SLC PCM system with one PCM DIMM rank. These values are consistent across all the PCM architectures used in this chapter. We first used NVsim [256], a CACTI-based non-volatile memory modeling tool, to calculate 22nm diode-switch PCM chip parameters. Then we scaled the calculated parameter values to 11nm PCM technology values following the PCM scaling guidelines given in [234] and [257]. According to the scaling guidelines in [257], the SET/RESET resistance increases linearly (by k) and programming current reduces linearly (by $1/k$) as the PCM feature size scales down by k. Note that the scaled values of programming currents (SET/partial-SET/RESET) given in Table 26 are greater than ITRS-projected values for the 11nm PCM process [31]. The use of higher than ITRS projected programming currents in our simulations provides significant tolerance against possible discrepancies between the ITRS projections and actual manufacturable solutions. As both SET/partial-SET and RESET resistances increase with process scaling, the dynamic range of resistance (difference in resistance between SET/partial-SET and RESET states) is preserved with process scaling. Moreover, process scaling does not affect read and write latencies, as the read/write latencies are primarily determined by the phase change material (GST in this chapter) [234].

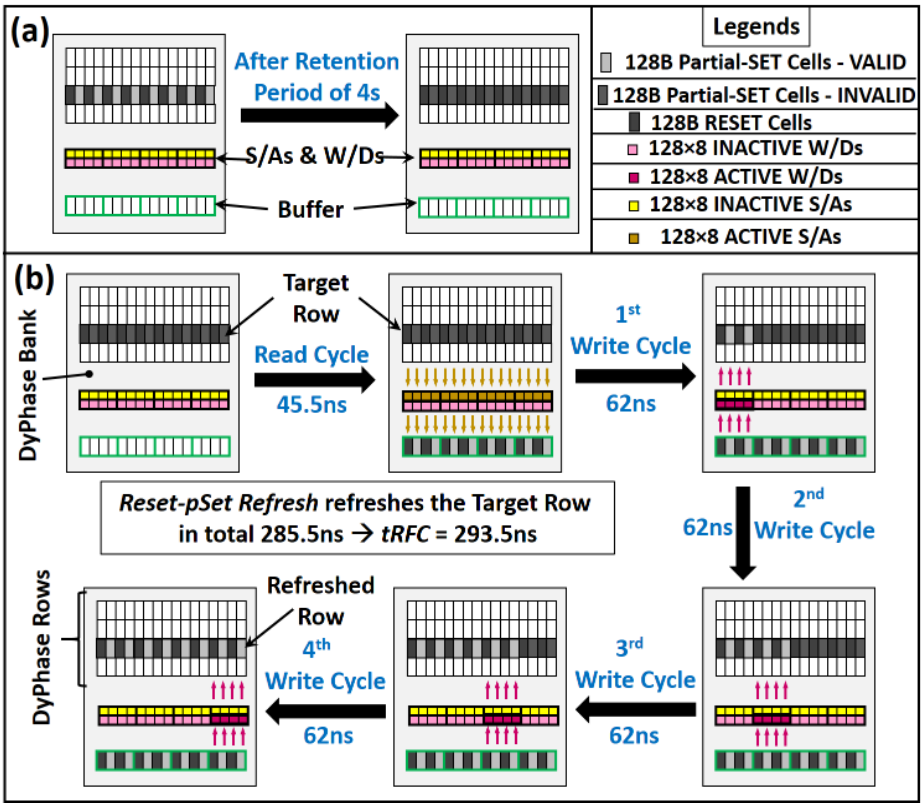


Figure 102: (a) Schematic of drift in resistance of partial-SET cells of a single-chip part of a DyPhase bank over the retention period of 4s, (b) schematic of a *Reset-pSet Refresh* cycle.

As implied in [243], the underlying design of CPs and total wasted power (leakage and parasitic) decides the peak power budget of a memory chip. Moreover, up to 75% of the total power consumed by a multi-unit CP system of a typical PCM chip is wasted as parasitic and leakage power [243]. As we adopt a modular multi-unit CP system in our baseline PCM, the total number of modular CP units decide the peak power provisioning to the PCM chip, which in turn decides how many 1-bit RESET operations can occur concurrently per chip. Hay et al. calculated in [258] that the peak cell-write current of 168mA provided by a typical DDR3-1066×16 DRAM memory can allow up to 560 simultaneous single-bit RESETs in a typical PCM with I_{RESET} of 300µA. But our assumed baseline PCM system has I_{RESET} of 40µA (Table 26), and modern DRAMs (e.g., [202] and [219]) can easily provide peak cell-write current of 200mA (or more) per

chip. This implies that our assumed baseline PCM system can support up to 5000 (200mA/40μA) 1-bit RESETs per chip. To be conservative, we assume that our baseline PCM can support up to 512B (equal to eight 64B cache lines) RESET operations per chip. The number of CP units to support this peak power provisioning and their area and power overheads are given in Table 26.

We implement the key attributes of our proposed *DyPhase* architecture on the above-described baseline PCM. Similar to the *pSET_PCM* architecture described in [247], the *DyPhase* architecture also uses partial-SET operations instead of full-SET operations to write '1's to PCM cells. *Reset-pSet Refresh* and *O-pSet Refresh*, the microarchitecture-level key attributes of the *DyPhase* architecture, are described next.

13.5.2. RESET PSET REFRESH

To deal with the retention related reliability issues of partial-SET operations, *DyPhase* PCM employs *Reset-pSet Refresh*, which is a distributed refresh technique that periodically restores data bits in *DyPhase* PCM cells every 4 seconds (the retention window of partial-SET cells from [247]) and ensures reliable data retention.

Reset-pSet Refresh and *O-pSet Refresh* are different from the distributed/interleaved refresh technique used in state-of-the-art DRAM systems [164], [199], [202] in the following ways: *(i)* the distributed refresh in DRAM restores the charge in DRAM storage cells, whereas *Reset-pSet Refresh* and *O-pSet Refresh* restore the difference in resistance between the partial-SET and RESET states of *DyPhase* PCM cells; *(ii) Reset-pSet Refresh* and *O-pSet Refresh* have longer refresh interval due to the longer data retention time for PCMs than DRAMs; *(iii) Reset-pSet Refresh* and *O-pSet Refresh* methods refresh smaller number of cells in parallel per refresh command, because refreshing a PCM cell requires larger amount of current compared to refreshing

a DRAM cell, which in turn limits the number of PCM cells that can be refreshed in parallel to a smaller number for the given peak current provisioning during a refresh command.

Typically, in the popular interleaved refresh scheme, each refresh command refreshes a certain number of rows depending on the size of the row and memory capacity. One refresh command is scheduled after every refresh interval (*tREFI*), defined as the *Retention Time/Total Number of Rows*. The number of rows that need to be refreshed during every refresh command is calculated as *Memory Capacity / (Row Size × Total Number of Rows)* (Chapter 12). The time taken by every refresh command to refresh the required number of rows is defined as refresh cycle time (*tRFC*), which depends on the granularity at which the required rows are refreshed.

As shown in Figure 102(a), in *DyPhase* PCM, due to the drift in resistance of partial-SET cells, the difference in resistance between the RESET and partial-SET cells becomes very small after 4s (retention period). As a result, logic '1' bits become very marginally distinguishable from logic '0's, potentially harming the validity of stored data. *Reset-pSet Refresh* restores the difference in resistance between partial-SET and RESET cells to be 8-10×, which ensures that logic '1's are clearly distinguishable from '0's. For that, as shown in Figure 102(b), *Reset-pSet Refresh* first reads the cells that are being refreshed and stores their data in a buffer, and then rewrites the buffered data in the cells in multiple write cycles (four write cycles in this case) depending on the granularity of refresh operation. To rewrite the buffered data, *Reset-pSet Refresh* uses RESET (for '0's) and partial-SET (for '1's) operations, which in turn ensures restoration of the difference in resistance between partial-SET and RESET cells. In a PCM system, the granularity of refresh operation is decided from the peak power provisioning to the constituent PCM chips.

Each row of a *DyPhase* logical bank is 16KB in size (stripped across 8 *DyPhase* chips; Figure 101(a)) and there are total of 32768×8=262144 rows in a 4GB *DyPhase* rank, which implies

that exactly one row should be refreshed during every refresh command. Moreover, as the retention time for partial-SET cells in *DyPhase* PCM is 4 seconds, the *tREFI* for the *DyPhase* PCM becomes 15.26μs (4s/262144). As mentioned earlier, our baseline PCM, and hence the *DyPhase* PCM, can support only up to 512B 1-bit writes per chip. But all eight chips of a *DyPhase* rank can function in parallel, which allows up to 512B×8=4KB cells to be written in parallel per *DyPhase* rank. As a result, refreshing/rewriting a 16KB row requires four write cycles where one write cycle rewrites only 4KB cells in parallel across all eight chips. The required number of write cycles per refresh command defines the granularity of the refresh operation, which in turn defines *tRFC*.

To evaluate *tRFC* for *Reset-pSet Refresh*, reconsider Figure 102(b). A *Reset-pSet Refresh* cycle is comprised of one read cycle followed by four write cycles. During the read cycle, first, the global and local wordline decoders (GWL_DEC, LWL _DEC) decode the target row address in 1.5ns (scaled down value from [234] to 11nm). Then, the S/As read data from the target row in 40ns (Table 26) and store it in the buffer in 2ns (scaled down value from [234]). As this buffer is used for refresh operations, we refer to it as refresh-buffer henceforth. After the read cycle, the first write cycle starts after providing 2ns time for data stability in the refresh-buffer. Every write cycle starts with bitline decode phase, which takes about 2ns. Then, the W/Ds write the buffered data back in the row cells in 50ns (defined by partial-SET operations; Table 26). Every write cycle of a *Reset-pSet Refresh* command refreshes 4KB cells in parallel (512B per chip) and is followed by an idle period of 10ns to limit the average power within the budget. From this analysis, the *tRFC* time for *Reset-pSet Refresh* is calculated to be 293.5ns (45.5ns (40ns+1.5ns+2ns+2ns) for the read cycle and 248ns (4× (50ns+10ns+2ns)) for total four write cycles).

In summary, the *Reset-pSet Refresh* method schedules one refresh command to a *DyPhase* rank every 15.26μs, which refreshes one logical row of 16KB size in 293.5ns. As the *Reset-pSet*

Refresh fully utilizes the available power budget, the entire *DyPhase* rank is stalled for the *tRFC* time of 293.5ns during which it does not serve any other memory requests. Thus, *Reset-pSet Refresh* enables *DyPhase* PCM to use partial-SET operations without any retention related reliability issues, and stalls only for 1.9% of the total run time of memory.

Note that, as only one row is refreshed at a time in the entire PCM system, the refresh-buffer can be shared among all the banks. Therefore, only one refresh-buffer is required per *DyPhase* rank, which is striped across 8 PCM chips and its size is exactly equal to the size of a logical row (i.e., 16KB). According to our area and power analysis, using NVSim [256] based standard cell library that is scaled for 11nm technology node, the refresh-buffer consumes 0.02mm^2 area and dissipates 10.3pJ dynamic energy and 2mW leakage power.

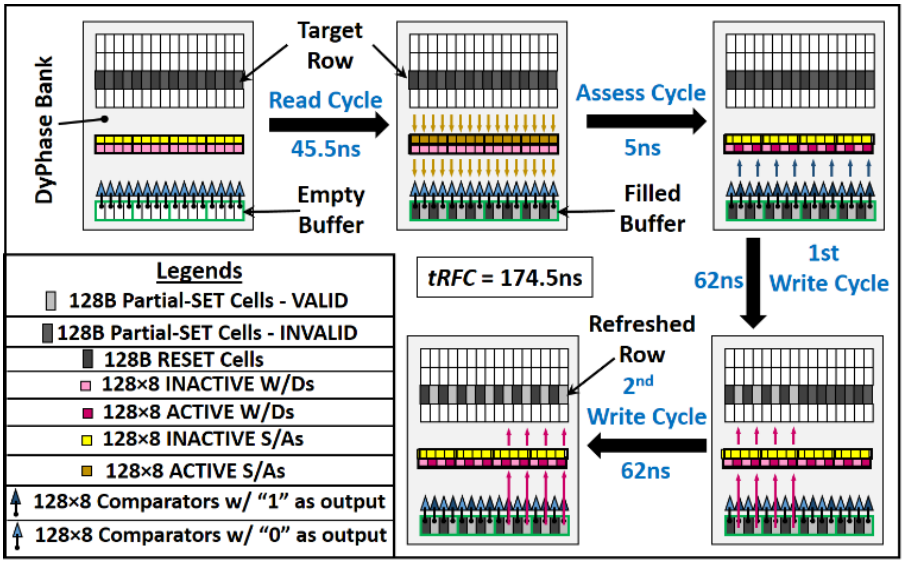


Figure 103: Schematic of an *O-pSet Refresh* cycle.

13.5.3. O-PSET REFRESH

O-pSet Refresh builds on *Reset-pSet Refresh* to reduce the *tRFC* time and energy overhead of refresh operations. The main idea is based on the fact that the resistance drift in partial-SET cells of the *DyPhase* PCM is the main contributor to the retention related reliability issues. For this

reason, it is sufficient to refresh/rewrite only the partial-SET cells of the target row in order to restore the difference in resistance between the partial-SET and RESET cells. There is no need to rewrite the RESET cells. Therefore, *O-pSet Refresh* rewrites/refreshes only the partial-SET cells of the target row. As evident from I_{SET} and I_{RESET} values given in Table 26, partial-SET operations require 2× less current than RESET operations. Because of this relaxed current requirements, *DyPhase* PCM can support total 5000×2=10000 1-bit partial-SET operations in parallel per chip, which are twice as many RESET operations per chip. However, we conservatively limit the maximum number of partial-SET operations to 8192 bits per chip. As a result, refreshing the target row of 16KB with *O-pSet Refresh* requires only two constituent write cycles, as one write cycle can rewrite 8KB (8Kb×8 chips) partial-SET cells per *DyPhase* rank.

To evaluate *tRFC* for *O-pSet Refresh*, consider Figure 103. An *O-pSet Refresh* cycle is comprised of one read cycle followed by an assess cycle and two consecutive write cycles. Similar to the read cycle of *Reset-pSet Refresh*, the read cycle of *O-pSet Refresh* also accesses the target row, reads its data and stores it in the refresh-buffer, in 45.5ns. In the assess cycle, a multi-bit voltage comparator assesses the buffered data (‘1’s and ‘0’s, i.e., digital voltage levels) in 3ns to find out exactly how many and which bits in the target row (stored in the buffer) are partial-SET bits (or ‘1’s). After knowing exactly which bits are partial-SET, the output of the comparator enables the W/Ds corresponding to the partial-SET bits in 2ns. In each of the following two write cycles, the enabled W/Ds rewrite the partial-SET bits of half (8KB) of the target row in 62ns (the same time-duration as the write cycle of *Reset-pSet Refresh*), including 2ns for bitline decode, 50ns for partial-SET operations and 10ns of idle period. Thus, an *O-pSet Refresh* has *tRFC* time of 174.5ns (45.5ns+5ns+62ns×2). Note that, similar to the *Reset-pSet Refresh*, *O-pSet Refresh* also refreshes one row every refresh command and has a *tREFI* of 15.26μs.

In summary, *O-pSet Refresh* schedules one refresh command to a *DyPhase* PCM rank every 15.26µs, which refreshes one logical row of 16KB size in 174.5ns. Thus, *O-pSet Refresh* reduces the *tRFC* by 40% compared to *Reset-pSet Refresh*. Similar to *Reset-pSet Refresh*, *O-pSet Refresh* also needs only one refresh-buffer of 16KB size per rank. In addition, it also requires one multi-bit voltage comparator of 16KB size. According to our area and power analysis, using NVSim [256] based standard cell library scaled for the 11nm node, the multi-bit comparator of 16KB size consumes 0.05mm^2 area and dissipates 18pJ dynamic energy and 3.5mW leakage power.

Lastly, note that, we refer to a *DyPhase* PCM system that employs *Reset-pSet Refresh* as *Reset-pSet DyPhase* henceforth. Similarly, we refer to a *DyPhase* PCM system that employs *O-pSet Refresh* as *O-pSet DyPhase*.

13.5.4. ANALYSIS OF WRITE ENDURANCE

Our full-system simulation analysis (Section 13.7) shows that due to the achieved improvements in write latency both the variants of our *DyPhase* architecture (*Reset-pSet DyPhase* and *O-pSet DyPhase*) yield greater performance than the baseline PCM (Section 13.5.1) and the other PCM architectures from prior works such as Partial-SET PCM (*pSET_PCM_QD8*) [247] and Write-Once-Memory-Code PCM (*WOMC_PCM*) [246]. In addition to poor write latency, poor write endurance is also a major challenge for the widespread adoption of PCM as main memory. A common measure of the PCM's write endurance is its lifetime, which mainly depends on its write rate. Assuming that PCM writes can be made uniform for the entire PCM capacity due to the use of wear-leveling techniques (e.g., [259], [260], and [261]), the PCM lifetime is given by the equation [262]:

$$Y = \frac{S.W_m}{B.F.2^{25}}, \tag{74}$$

where, Y is PCM lifetime in years, W_m is maximum allowable number of writes per cell, B is write rate in bytes/cycle, F is frequency of processor in Hz, and S is PCM capacity in bytes.

From the equation, the lifetime of a PCM system depends on its write rate (B) in terms of bytes/cycle. The number of bytes per cycle written to a PCM system depends on the application workload of the system and underlying PCM system architecture. In general, the write rate of a PCM system corresponds to the write throughput for a given application workload. However, for the *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures, the write rate is somewhat different from the write throughput due to the additional writes induced by periodic refresh operations. During the distributed refresh operations of *O-pSet* and *Reset-pSet DyPhase* architectures, an entire 4GB *DyPhase* rank is refreshed in 4s, which causes the refresh-induced write rate for the *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures to be 0.54 bytes/cycle assuming a 0.5ns cycle period. As these refresh-induced periodic writes do not stem from application workload, the net write rate for the *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures is 0.54 bytes/cycle plus the workload-induced write throughput.

Similarly, the *pSET_PCM_QD8* requires rewriting of the partial-SET-queue-evicted entries with full-SET operations. As a result, in *pSET_PCM_QD8* architecture, every workload-induced memory write operation is performed two times. This causes the net write rate of the *pSET_PCM_QD8* architecture to be twice the workload-induced write throughput.

Figure 104(a) gives the net write rate values for various PCM architectures for PARSEC benchmarks. Figure 104(b) gives lifetime values (evaluated using Eq. (74) for the net write rate values given in Figure 104(a)) for various PCM architectures. The details of the simulation setup and system configuration utilized to obtain these results are given in Section 13.7. For our evaluation of lifetime values, we took the values of W_m, F, and S to be 10^8 ([31] and [234]), 2GHz,

and 4GB (see Section 13.5) respectively. Note that the utilized hybrid PCM-DRAM configuration for main memory along with the biased coin based migration policy (see Section 13.7) favor some applications (i.e., *Blackscholes*, *Bodytrack*, *Facesim*, and *Fluidanimate*) such that majority of the writes induced by these applications are consumed by the DRAM part of the main memory. As a result, the PCM parts of all architectures in Figure 104(a) yield very small values of workload-dependent write-rates for these applications. Therefore, baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* achieve significantly long lifetimes for these applications. However, in spite of having very small values of workload-dependent write rate for the favored applications, *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures yield significantly short lifetimes compared to the other PCM architectures due to the large values of refresh-induced write rate. In fact, O-pSet DyPhase and Reset-pSet DyPhase have the highest write rate values for all the applications (Figure 104(a)), resulting in the shortest lifetimes. For instance, for *Canneal* application, *O-pSet DyPhase* has 4.6, 0.7, 3.3 years shorter lifetime than baseline, *pSET_PCM_QD8*, and *WOMC_PCM* respectively. Similarly, *Reset-pSet DyPhase* has 4.7, 0.8, 3.4 years shorter lifetime than baseline, *pSET_PCM_QD8*, and *WOMC_PCM* respectively. Thus, in spite of having superior performance and energy-efficiency, both *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures have shorter lifetimes. In the next section, we present an efficient solution for these lifetime-related shortcomings of the *DyPhase* architectures.

13.6. RESTORATION OF WRITE ENDURANCE

13.6.1. PROACTIVE IN-SITU SELF-ANNEALING (PISA)

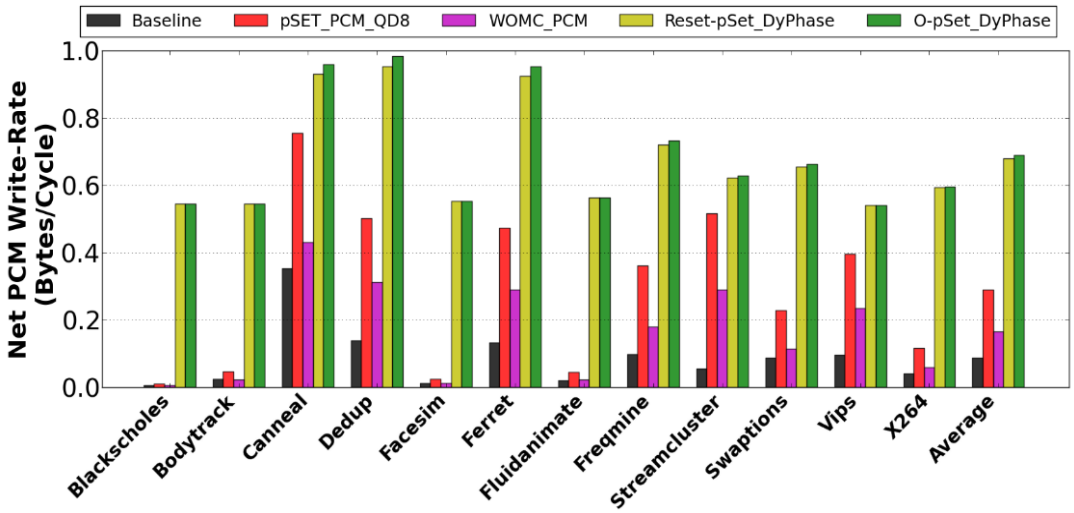
In this section, we present a system-level implementation of an improved version of a previously proposed technique called In-situ Self-Annealing (ISA) [37], which remedies write-cycle-induced degradation in PCM cells and restores their lifetime. As discussed in [37], the

resistance of a typical PCM cell in the RESET state significantly decreases as the number of cell-writes increase beyond 10^4. In this state, the PCM cell is referred to as having stuck-SET failure. This trend of decreasing RESET resistance with increasing cell-writes continues until near the end of the cell's lifetime, where the RESET resistance of the cell drastically increases with increase in the number of cell-writes before the electrical path between the chalcogenide material GST and access device in the cell severs after about 10^8 cell-writes. In this state, the PCM cell is referred to as having stuck-RESET failure. Intuitively, the number of cell-writes during a given amount of time increases with increase in write rate. Hence, the lifetime of a cell decreases with increase in write rate, which is also evident from Eq. (74).

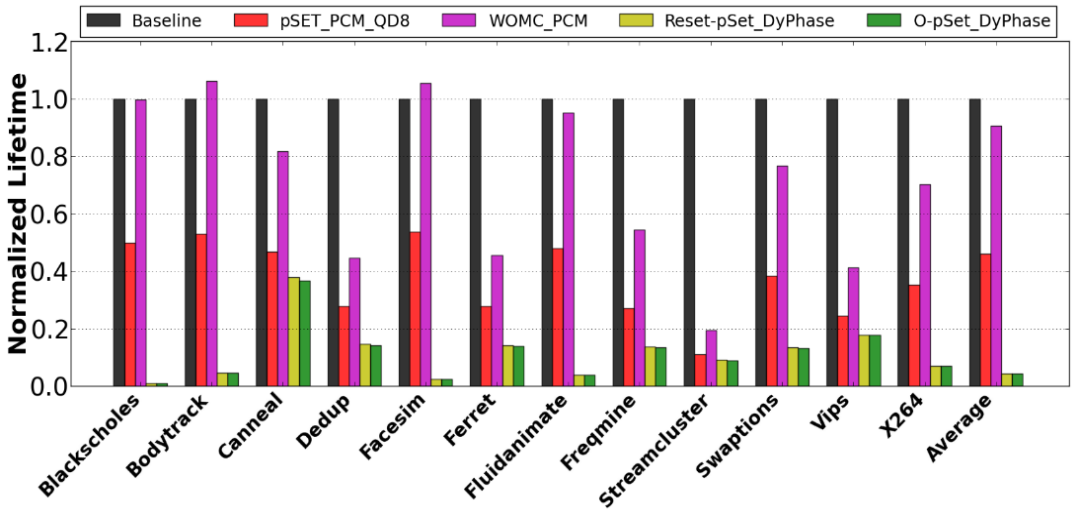
The PCM-endurance-restoring ISA technique presented in [37] monitors the resistance of each PCM cell after every RESET operation, and reactively triggers a healing procedure when the cell-resistance drops below a certain threshold value (typically corresponding to cell-writes greater than 10^4). To heal a degraded cell, the ISA technique applies a self-annealing current pulse of certain amplitude (I_{ISA}) and duration (T_{ISA}) to it. This current pulse results in localized joule heating, which in turn anneals the GST material of the cell and restores its write endurance. As discussed in [37], a healing current pulse that has 15-20% more amplitude than I_{RESET} and T_{ISA} of 10μs can completely restore the cell's endurance and lifetime, if the cell was not already written to for more than 5×10^4 times before issuing the healing current pulse. A cell can be partially restored (only by 90-95%) if the number of cell-writes is between 5×10^4 and 10^8 at the time of healing-pulse application; if a cell is written for more than 10^8 times, it cannot be restored at all with the ISA technique [37].

From the value of I_{RESET} given in Table 26, it can be implied that self-annealing current pulses with I_{ISA}=48μA (20% more than I_{RESET}) and T_{ISA}=10μs can restore the lifetime of our

DyPhase PCMs. However, the ISA technique presented in [37] reactively triggers the issuance of 10µs long self-annealing current pulses, which is impractical as it requires frequent assessment of cell resistance (typically after each write) and incurs excessive performance penalty. Therefore, an efficient workaround to this shortcoming is required to be able to effectively use the ISA technique for our *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures.



(a)



(b)

Figure 104: (a) Net PCM write rate, (b) normalized PCM lifetime values for various hybrid PCM main memory systems. Lifetime values are only for the PCM parts of the hybrid systems and are normalized wrt the PCM part of the baseline hybrid system.

As a solution, we propose a proactive technique *PISA* that has minimal performance penalty. As shown in Figure 105, *PISA* periodically schedules interleaved healing cycles, issuing a healing-burst (a burst of healing operations) of time-duration T_{HB} at the end of every restoration period (T_{RSP}). During a healing-burst, multiple healing cycles (of T_{HLC} time duration each) are interleaved with a healing interval of T_{HEALI} time-duration between two successive healing cycles. Periodically issued healing bursts ensure that an assessment of PCM cell resistance after each cell-write is not required, and hence, the overhead of healing operations is reduced.

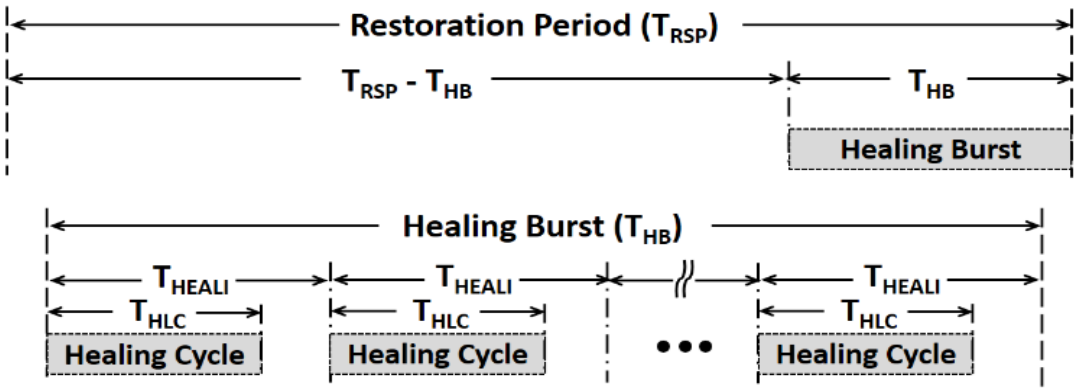


Figure 105: Periodically scheduled interleaved healing cycles during a restoration period of the *PISA* technique.

However, due to the presence of endurance variations [263], unpredictable wear-rate patterns are inflicted upon individual PCM cells. This results in an uncertain amount of degradation of individual PCM cells during every T_{RSP} period, which in turn results in uncertain amount of endurance for individual PCM cells after every T_{HB} healing burst. To guard against this uncertainty and resultant reliability issues, we propose to periodically assess the PCM cell resistance during the read cycles of periodic *O-pSet Refresh* and *Reset-pSet Refresh* commands. During the read cycle of any refresh command, if the resistance of any of the RESET cells of the accessed row is found to have reduced by more than 10% of the nominal RESET cell resistance, it serves as an

indication (as implied from [37]) that one or more cells of the accessed row may have stuck-SET failure. Upon detection of a stuck-SET failure, the *DyPhase* memory controller schedules a healing burst in coalescence with the following four successive refresh periods. The implementation details of the healing-burst schedule and *DyPhase* controller are given in the next subsection.

Moreover, we carefully select the values of T_{HB}, T_{RSP}, and T_{HEALI} to meet the following criteria, so that significant improvement in *DyPhase* PCM lifetime with minimal performance overhead can be achieved: *(i)* T_{RSP} and T_{HB} are chosen such that the number of cell-writes do not exceed 5×10^4 for any *DyPhase* cell during the restoration period. This helps in avoiding untimely cell failures and ensures complete healing of PCM cells during each restoration cycle; *(ii)* T_{HB} and T_{HEALI} are chosen such that all the rows of a *DyPhase* rank are healed during T_{HB} time without violating the instantaneous power budget; and *(iii)* the introduction of T_{HB}, T_{RSP}, and T_{HEALI} timing parameters does not add to the complexity of the design and operation of *DyPhase* memory controller.

We select T_{RSP} based on the maximum achievable write rate of the Intel Core i7-6700K processor, which is one of the fastest general purpose processors today [264]. Deciding T_{RSP} in this way ensures that the chosen value of T_{RSP} meets the above criteria for all the commodity general purpose processors available today. As described in [264], the Intel Core i7-6700K processor can achieve peak double precision floating-point performance of 81.28GFlops. If one Flop on average writes 1B in the memory (as explained in [201]) and if we assume that PCM writes can be made uniform for the entire PCM capacity, it takes 2642s (derived from Eq. (74)) for the Intel Core i7-6700K processor (that has F=4GHz) to achieve 5×10^4 cell-writes for our *DyPhase* PCM architectures (S=4GB). Now, if we choose a value of T_{RSP} that is an integer multiple of *DyPhase* PCM's retention period of 4s, we can significantly reduce the complexity of *DyPhase* controller

used for managing the new PCM timing parameter T_{RSP}, as the controller can calculate T_{RSP} by simply counting the number of elapsed retention periods. Therefore, we choose T_{RSP} to be 2640s, which is the 660th multiple of the retention period 4s.

Similarly, we make it easier for the *DyPhase* controller to manage T_{HB} and T_{HEALI} parameters by designing the *DyPhase* controller to schedule the interleaved healing operations during the T_{HB} time-duration at the same granularity at which the distributed *O-pSet Refresh* and *Reset-pSet Refresh* operations are scheduled. For that reason, and to ensure that the instantaneous power constraints are not violated, we choose T_{HB}=16s and T_{HEALI}=15.26µs. The implementation of our *PISA* technique with these values of parameters T_{HEALI}, T_{HB}, and T_{RSP} is described in the next subsection.

Note that, as I_{ISA} of 48µA is greater than I_{RESET}, and as the healing operations are different from the RESET operations, a *DyPhase* chip requires separate on-chip CP units for sourcing I_{ISA} during healing operations. The addition of extra CP units incurs area and power overheads. Nevertheless, as it will be clear from the discussion in Section 13.7.3, in spite of having high overhead, the introduction of on-chip CP units for healing operations ultimately results in better lifetime and energy-efficiency for *O-pSet DyPhase* and *Reset-pSet DyPhase*.

13.6.2. IMPLEMENTATION OF PISA TECHNIQUE ON DYPHASE PCM

This section presents the implementation of *PISA* on the *O-pSet DyPhase* and *Reset-pSet DyPhase* architectures. As mentioned in the previous section, we choose T_{HB} and T_{HEALI} to be 16s and 15.26µs respectively, which are equal to the values of 4×retention period (or 4×refresh period) and refresh interval (*tREFI*), respectively. Moreover, the restoration period of *PISA* technique (i.e., T_{RSP}) is an integer multiple (660th) of the refresh period, i.e., one restoration period of 2640s is equal to 660 refresh periods (of 4s each). To leverage this multiplicity of parameters T_{RSP}, T_{HB}, and

T_{HEALI} with *tREFI* and refresh period, the *DyPhase* controller treats every restoration period as a scheduled sequence of 660 consecutive refresh periods (of 4s each). During each of the first 656 scheduled refresh periods, the *DyPhase* controller issues a total of 256K refresh commands that refresh 256K rows (i.e., one refresh command refreshes one row) with 15.26µs refresh interval between two successive refresh commands. Every refresh command takes *tRFC* time to refresh the target row, and the value of *tRFC* depends on which refresh technique (i.e., *O-pSet Refresh* or *Reset-pSet Refresh*) is used.

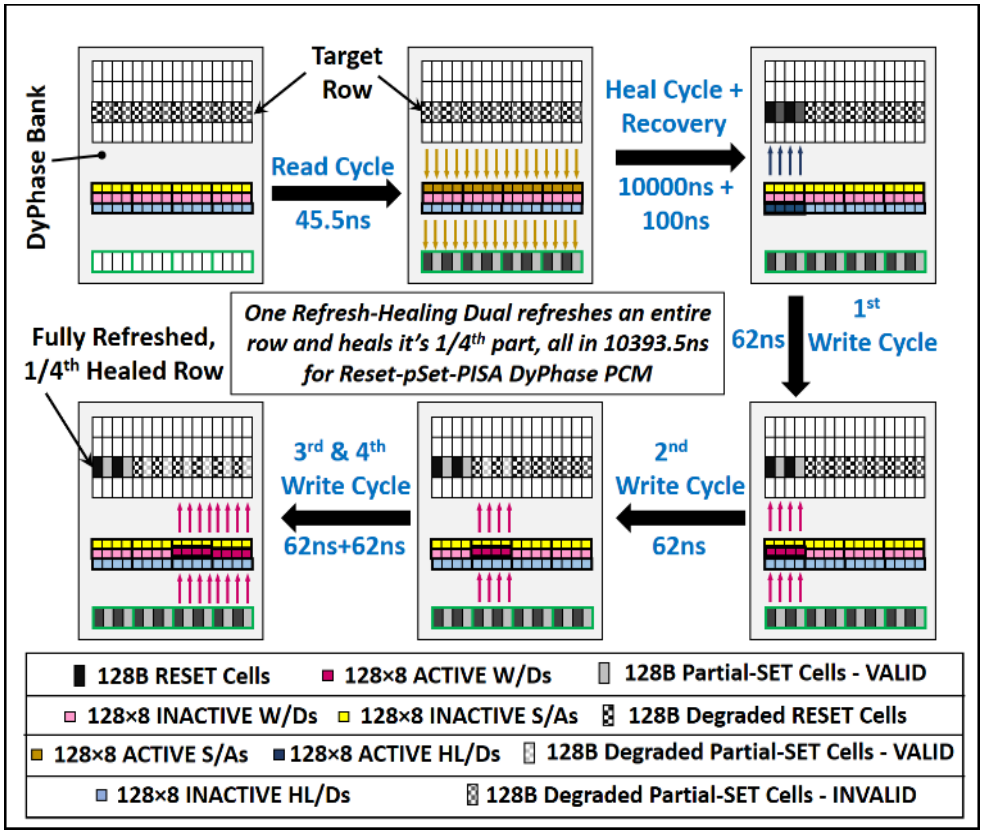


Figure 106: Schematic of a refresh-healing dual operation for the *PISA*-enabled *Reset-pSet DyPhase* PCM.

The controller proactively schedules a healing burst of T_{HB}=16s along with the four refresh periods from 657th refresh period to 660th. Thus, each healing burst has four constituent refresh periods. Now, during each of the four healing-burst-constituent refresh periods, a refresh command

and a healing operation both are scheduled together every 15.26µs. The combination of a refresh command and a healing operation is referred to as a refresh-healing dual henceforth. A total of 256K refresh-healing duals are issued during every healing-burst-constituent refresh period. Each of these refresh-healing duals requires T_{HLC} time-duration and refreshes one entire row but heals only 1/4th of the row. As a result, at the end of the first healing-burst-constituent refresh period, all the rows of a 4GB *DyPhase* rank are refreshed once, but only the 1/4th parts of all the rows are healed. Similarly, at the end of the second healing-burst-constituent refresh period, all the rows of the *DyPhase* rank are refreshed twice, but only the 2/4th parts of all the rows are healed. Consequently, at the end of the fourth and final healing-burst-constituent refresh period (i.e., at the end of the healing burst), all the rows of the *DyPhase* rank are refreshed four times, but they are all healed once.

Figure 106 illustrates a refresh-healing dual operation for the *PISA*-enabled *Reset-pSet DyPhase*. The operation consists of one read cycle, one healing cycle and 2 or 4 write cycles (2 for *O-pSet DyPhase* and 4 for *Reset-pSet DyPhase*). Similar to the read cycles of *O-pSet Refresh* and *Reset-pSet Refresh* commands, the read cycle of a refresh-healing dual also accesses the target row, reads its data and stores it in the buffer, all in a total of 45.5ns. Then follows a healing cycle, during which only 1/4th of all the cells (4KB per rank and 4Kb per chip) of the target row are healed in parallel by applying I_{ISA}=48µA to each cell for T_{ISA}=10µs duration (corresponds to 1.2nJ energy per cell). We discuss the design of healing current provisioning later in this section. The healing cycle precedes an idle recovery time of 100ns that recovers the current delivery network from the peak surge of healing current. Then follows the write cycles, each of which takes 62ns to rewrite the buffered data in the target row. Thus, if the *PISA* technique is implemented on the *O-pSet DyPhase* architecture (the resulting architecture is called *O-pSet-PISA-DyPhase* henceforth),

the refresh-healing dual requires T_{HLC} of 10274.5ns, which stalls the PCM for 1.5% of the total run time of memory. On the other hand, if the *PISA* technique is implemented on *Reset-pSet DyPhase* (the resulting architecture is called *Reset-pSet-PISA-DyPhase* henceforth), the refresh-healing dual requires T_{HLC} of 10393.5ns, which stalls PCM for 2.3% of the total run time of memory.

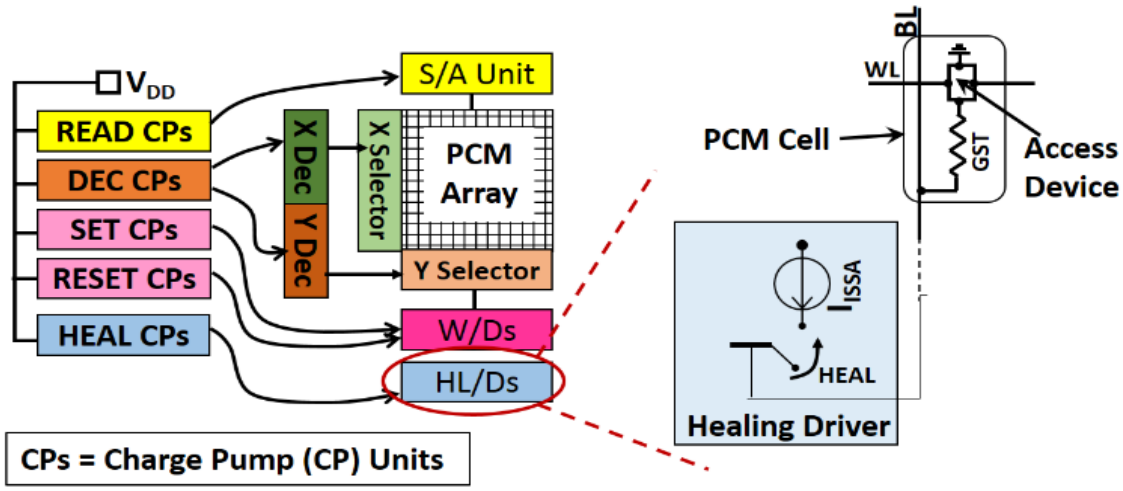


Figure 107: Schematic of a charge pump (CP) system with healing CPs and healing drives (HL/Ds).

As only a 1/4th part of the target row is healed during every healing cycle, the healing cycle requires 196.6mA (48µA×4096) of current per chip. As described in the previous subsection, we use separate on-chip CP units to provide this healing current. The introduction of separate on-chip heal CP units requires restructuring of the CP system and PCM array peripherals, as shown in Figure 107. A separate set of drivers called heal drivers (HL/Ds) are introduced. During a healing operation of a PCM cell, healing current of I_{ISA} amplitude flows from HL/D to the cell ground line through bit-line decoder, bit-line, GST, and access device. If we use the same design of on-chip CPs as used in [243], the on-chip CP units that can deliver 196.6mA current occupy 4.4mm^2 area and dissipate 1.5W of leakage power due to their very low power conversion efficiency of 25% [243]. Nevertheless, as will be clear in the next subsection, in spite of area and power overheads,

our *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* architectures achieve greater lifetime and energy-efficiency compared to the other PCM architectures from prior works.

13.7. EVALUATION

13.7.1. EVALUATION SETUP

We performed full-system simulation analysis to compare the four variants of our *DyPhase* architecture (*Reset-pSet DyPhase* and *O-pSet DyPhase* with and without *PISA*) with the baseline PCM (Section 13.5.1). We also compare with other PCM architectures from prior works: Partial-SET PCM (*pSET_PCM_QD8: the pSET_PCM* architecture discussed in Section 13.5 with partial-SET queue depth of eight entries) [247] and Write-Once-Memory-Code PCM (*WOMC_PCM*) [246]. We implement an energy- and cycle-accurate model of the baseline PCM system described in Section 13.5.1 in NVMain 2.0 [265]. We model the *DyPhase, pSET_PCM_QD8,* and *WOMC_PCM* architectures by implementing their unique characteristics on the model of the baseline PCM system. We use the resulting energy- and cycle-accurate models of PCM architectures in a DRAM-PCM hybrid main memory system.

We integrated our NVMain 2.0 based main memory system models with gem5 [77] for full-system simulations. We use the PARSEC benchmark suite for our simulation analysis because it covers a wide spectrum of working sets, locality, data sharing, synchronization and off-chip traffic [76]. We ran each PARSEC benchmark for a specific "warm-up" period and then captured full-system CPI count and memory system behavior for the subsequent region of interest (ROI) [266]. Table 27 gives the configurations of the DRAM-PCM main memory systems. Table 28 gives the configuration of gem5 that was used for this study. The first-ready first-come-first-serve (FR-FCFS) scheduling scheme was used for the DRAM parts of all main memory subsystems, whereas the FR-FCFS scheme with write buffering was used for the PCM parts. An open page policy and

row:rank:bank:channel:col address mapping scheme were used for all simulations. Average latency, CPI counts, net write throughput, and energy-delay product values for the memory subsystem were obtained from gem5 integrated NVMain 2.0.

Table 27: Configuration of hybrid DRAM-PCM memory system.

Main memory	2 channels; 1 channel for DRAM part and the other for PCM part; 4GB capacity per channel; 64-bit channel width; 1 rank per channel; 8 chips per rank; 8 banks per rank; 8 banks interleaved on 8 chips; LPDDR2-NVM interface; 11nm process for PCM;
PCM chip	$4F^2$ diode-switched cell; 8-bit width; 512MB capacity; 1V V_{DD}; 400MHz; 200mA peak current provisioning
DRAM chip	8-bit width; 512MB capacity; 667MHz; DDR3-1333_4Gb_8B_x8 chip [267]
Memory controller	FR-FCFS scheduling for DRAM; FR-FCFS with write buffering for PCM; open page policy; row:rank:bank:channel:col address mapping; 32-entry R/W queues; in-order R/W issue; biased coin based page migration policy [265]

Table 28: Gem5 simulation configuration.

Number of Cores	4	L2 Coherence	MOESI
L1 I Cache	32KB	Frequency	2 GHz
L1 D Cache	32KB	Issue policy of cores	OoO (4 issue)
Shared L2 Cache	2MB	# Memory Controllers	1
ISA/OS	ALPHA	Cache Associativity	4-way (L1); 8-way (L2)

13.7.2. EVALUATION RESULTS FOR DYPHASE WITHOUT PISA

Figure 108 shows absolute values of full-system CPI counts for various hybrid PCM main memory systems across twelve PARSEC benchmarks. *O-pSet DyPhase* yields 6.9% smaller CPI count on average over all the other hybrid PCM systems. More specifically, *O-pSet DyPhase* yields about 0.3%, 3.7%, 6.7%, and 16.8% less CPI over *Reset-pSet DyPhase*, *WOMC_PCM*, *pSET_PCM_QD8*, and baseline PCM respectively. Moreover, *Reset-pSet DyPhase* yields 3.5%, 6.5%, and 16.4% less average latency over *WOMC_PCM*, *pSET_PCM_QD8*, and baseline PCM

systems respectively. *O-pSet DyPhase* has a 40% smaller *tRFC* than *Reset-pSet DyPhase*, which results in 0.3% less CPI for *O-pSet DyPhase* compared to *Reset-pSet DyPhase*.

In spite of using partial-SET operations instead of full-SET operations to write '1's, the use of partial-SET queue and having to rewrite the evicted entries with full-SET operations reduces the CPI improvements achieved by the *pSET_PCM_QD8* system compared to *WOMC_PCM* and *DyPhase* systems. Along the same lines, as *WOMC_PCM* needs to use full-SET operations for every alternate write to the same memory location, it cannot completely eliminate the need to use full-SET operations. As a result, it has greater CPI value compared to the *DyPhase* architectures. Figure 109 gives net write throughput values for various hybrid PCM systems across the PARSEC benchmarks. *O-pSet DyPhase* achieves about 1.4%, 15.9%, 28.1%, and 69.8% more net write throughput than *Reset-pSet DyPhase*, *WOMC_PCM*, *pSET_PCM_QD8*, and baseline PCM respectively. Moreover, *Reset-pSet DyPhase* yields 14.3%, 26.3%, and 67.4% more write throughput than *WOMC_PCM*, *pSET_PCM_QD8*, and baseline PCM respectively.

From the results in Figure 109, *pSET_PCM_QD8*, *WOMC_PCM*, *O-pSet DyPhase*, and *Reset-pSet DyPhase* have less write throughput for the *Fluidanimate* application than all other applications. From [76], *Fluidanimate* shows seamless streaming behavior, which results in consecutive memory-writes being mapped to physically adjacent memory locations. This type of write behavior provides more opportunities for the utilized migration policy [265] to migrate the majority of write-intensive pages to the DRAM parts of the hybrid PCM systems. As a result, the PCM parts of these hybrid PCM systems get significantly less write requests to serve compared to their DRAM counterparts. This results in all hybrid PCM main memory systems shown in Figure 109 yielding less write throughput for *Fluidaminate* than all other applications.

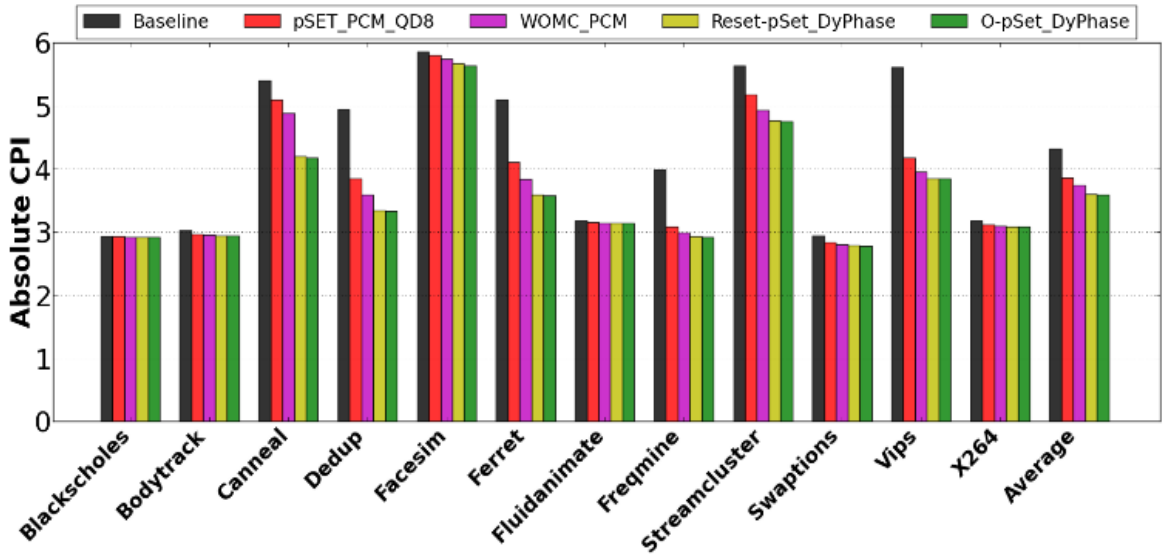


Figure 108: Absolute values of full-system CPI counts for various hybrid PCM systems across PARSEC benchmarks.

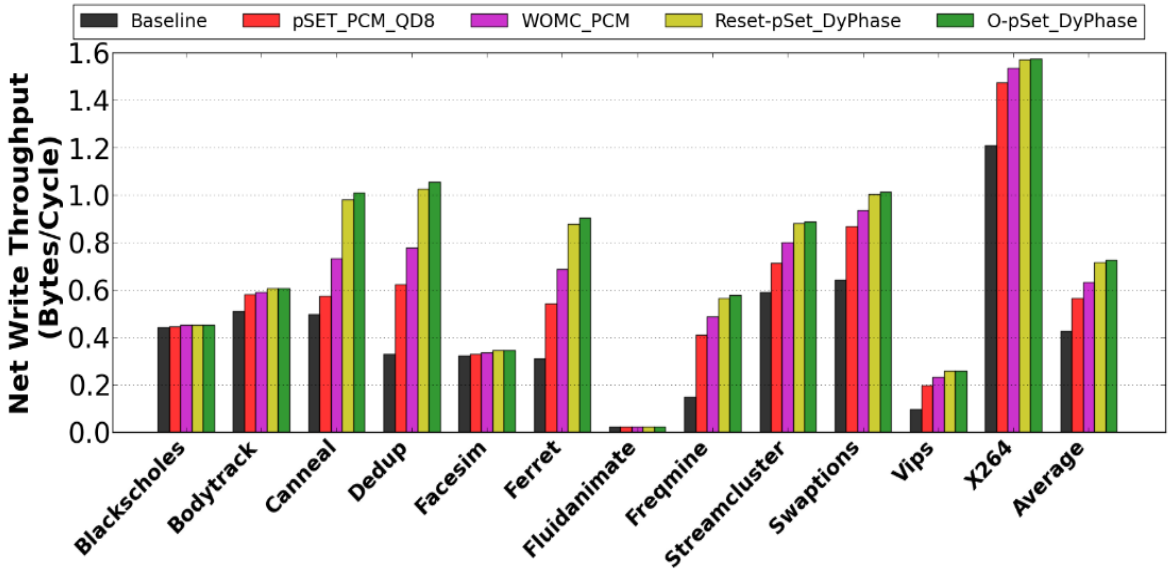


Figure 109: Absolute net write throughput values for various hybrid PCM systems across PARSEC benchmarks.

Figure 110(a) gives net throughput (read throughput + write throughput) values for various hybrid PCM systems averaged over the twelve PARSEC benchmarks. The error bars in the figure represent standard deviation of net throughput values across the PARSEC benchmarks. *O-pSet*

DyPhase achieves about 60.8%, 13.2%, and 23.7% more net throughput on average over baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* respectively. Similarly, *Reset-pSet DyPhase* achieves about 59.8%, 12.5%, and 23% more net throughput on average over baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* respectively. The reduction in the latency of write operations results in more read and write requests being served in unit time, which renders greater net throughput for *Reset-pSet DyPhase* and *O-pSet DyPhase*. The necessity of rewriting the partial-SET-queue-evicted entries with full-SET operations results in less throughput for *pSET_PCM_QD8* than *WOMC_PCM*.

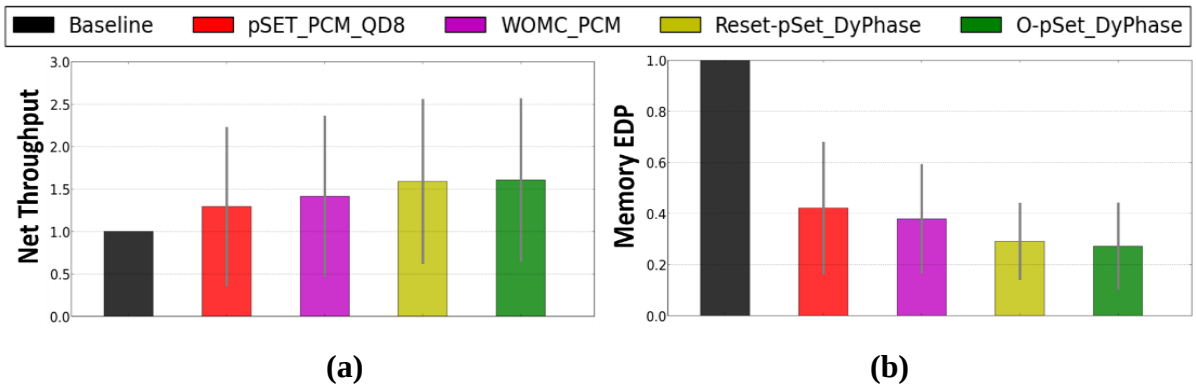


Figure 110: (a) Net throughput, (b) energy delay product (EDP) values for various hybrid PCM systems averaged across PARSEC benchmarks. The values are normalized wrt the baseline hybrid PCM system. The error bars represent standard deviation of values across the PARSEC benchmarks.

Figure 110(b) gives energy-delay product (EDP) values for various hybrid PCM systems averaged over the twelve PARSEC benchmarks. The error bars in the figure represent standard deviation of EDP values across the PARSEC benchmarks. We obtain EDP values by multiplying energy-per-bit values with average latency values. We obtain energy-per-bit values by dividing net power values with net throughput values. *O-pSet DyPhase* yields about 72.7%, 28.1%, and 35.3% less EDP on average over baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* respectively. Similarly, *Reset-pSet DyPhase* also yields about 70.8%, 23.1%, and 31% less EDP

over baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* respectively. It can be concluded that due to the achieved greater net throughput, *O-pSet DyPhase* and *Reset-pSet DyPhase* have less energy-per-bit than the other hybrid PCM systems in spite of having more net power due to the periodic refresh operations. This energy-per-bit benefit in addition to less write latency render better energy-efficiency for *O-pSet DyPhase* and *Reset-pSet DyPhase* in terms of EDP compared to other hybrid PCM systems.

In summary, both *O-pSet DyPhase* and *Reset-pSet DyPhase* based hybrid systems yield greater performance and energy-efficiency compared to the other hybrid systems that utilize PCM architectures from prior work.

13.7.3. EVALUATION OF DYPHASE WITH PISA

In this section, we evaluate the lifetime, CPI, and EDP for our *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* based hybrid PCM systems across PARSEC benchmarks, and compare them with the baseline PCM, *WOMC_PCM* and *pSET_PCM_QD8* based hybrid PCM systems.

As described in [253], due to process variations, PCM cells within a typical PCM array exhibit different rates of write degradation and healing. As a result, as can be implied from the discussions given in [37] and [253], even if we ensure that no PCM cell is written for more than 5×10^4 times during every restoration period (T_{RSP}) of 2640s, the PCM cells cannot be fully healed using the *PISA* technique. Therefore, we assume that our *PISA* technique can heal a PCM cell by up to 98%. With this assumption, we use Eq. (74) to evaluate the lifetime of the *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* hybrid systems for the PARSEC benchmarks, and compare it with the lifetime of other PCM hybrid systems.

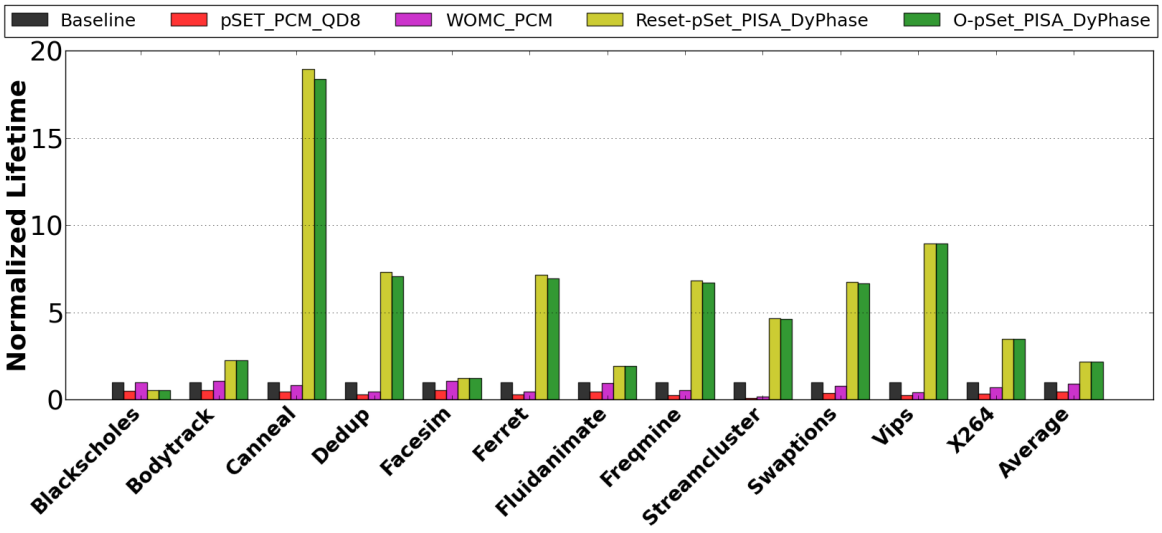
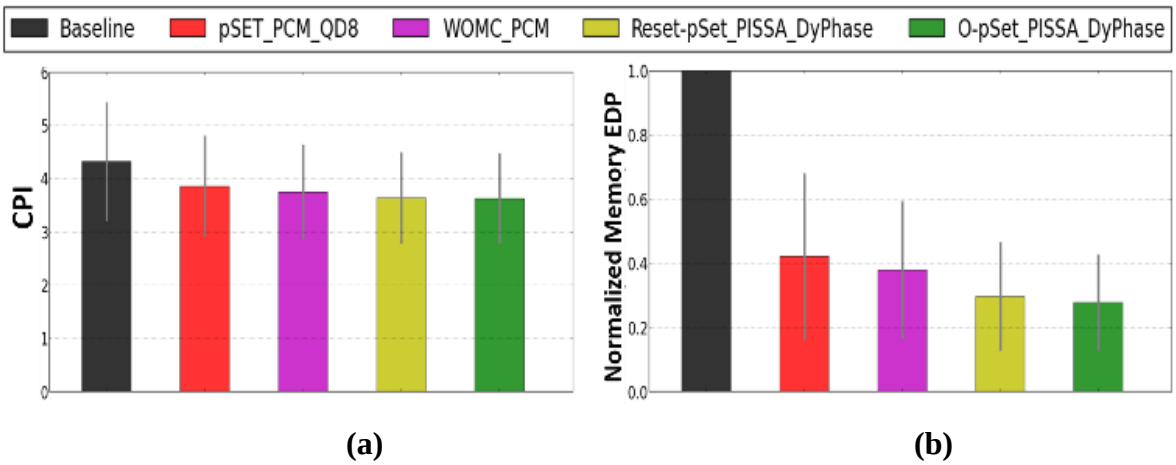


Figure 111: Normalized lifetime for various hybrid PCM systems across PARSEC benchmarks.



(a) **(b)**

Figure 112: (a) CPI, (b) EDP values for various hybrid PCM systems averaged across the PARSEC benchmarks. EDP values are normalized w.r.t. the baseline hybrid PCM system. The error bars represent standard deviation across the PARSEC benchmarks.

Figure 111 gives lifetime values for PCM parts of various PCM hybrid systems for various PARSEC benchmarks. The PCM parts of *O-pSet-PISA-DyPhase* hybrid systems achieve 113, 154, and 105 years more lifetime on average over the PCM parts of *WOMC_PCM*, *pSET_PCM_QD8* and the baseline hybrid systems respectively. *Reset-pSet-PISA-DyPhase* achieves 2, 115, 156, and 107 years more lifetime on average over *O-pSet-PISA-DyPhase*, *WOMC_PCM*,

pSET_PCM_QD8, and the baseline PCM respectively. Similar to *O-pSet DyPhase* (Figure 104(a) in Section 13.5.4), *O-pSet-PISA-DyPhase* also achieves greater write throughput, which renders greater net write rate for *O-pSet-PISA-DyPhase* compared to *Reset-pSet-PISA-DyPhase.* As a result, *O-pSet-PISA-DyPhase* achieves less lifetime than *Reset-pSet-PISA-DyPhase.* From the comparison of these lifetime values with the lifetime values given in Figure 104(b), it is evident that our *PISA* technique improves the lifetime of both the *O-pSet DyPhase* and *Reset-pSet DyPhase* by about 50 times. Note that if T_{RSP} and T_{HB} for our *PISA* technique are chosen such that the number of cell-writes do not exceed 5×10^4 for any *DyPhase* cell, then the *PISA* technique can completely heal all *DyPhase* cells, theoretically rendering infinite lifetime for *DyPhase* PCM. However, we account for possible fabrication process variations in our lifetime analysis, which results in only 50× improvement in lifetimes of *O-pSet DyPhase* and *Reset-pSet DyPhase* when integrated with *PISA.*

Figure 112(a) gives CPI values for various hybrid PCM systems averaged over the PARSEC benchmarks. The error bars in the figure represent standard deviation of CPI values across the PARSEC benchmarks. *O-pSet-PISA-DyPhase* yields about 16%, 3%, 6%, and 0.3% less CPI over the baseline PCM, *WOMC_PCM*, *pSET_PCM_QD8*, and *Reset-pSet-PISA-DyPhase* respectively. Similarly, *Reset-pSet-PISA-DyPhase* yields about 15.7%, 2.5%, and 5.7% less CPI over the baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* respectively.

Figure 112(b) gives EDP values for various PCM hybrid systems averaged over the PARSEC benchmarks. The error bars in the figure represent standard deviation of EDP values across the PARSEC benchmarks. *O-pSet-PISA-DyPhase* yields about 69.2%, 23.6%, 31.1%, and 5.4% less EDP over the baseline PCM, *WOMC_PCM*, *pSET_PCM_QD8*, and *Reset-pSet-PISA-*

DyPhase respectively. *Reset-pSet-PISA-DyPhase* yields about 67.3%, 18.6%, and 26.5% less EDP over the baseline PCM, *WOMC_PCM*, and *pSET_PCM_QD8* respectively.

13.7.4. COMPARISON OF PISA WITH ISA

As our PISA technique is an improved version of *ISA* technique from prior work [37], we compare *PISA* enabled hybrid *DyPhase* systems (*O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase*) with *ISA* enabled hybrid *DyPhase* systems (*O-pSet-ISA-DyPhase* and *Reset-pSet-ISA-DyPhase*) in terms of CPI and EDP values obtained for PARSEC benchmarks.

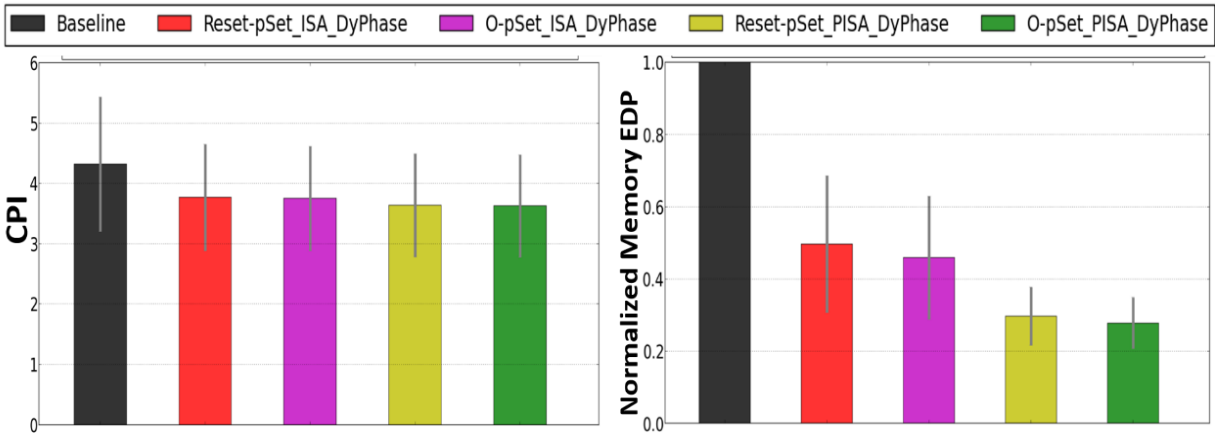


Figure 113: (a) CPI, (b) EDP values for various hybrid PCM systems averaged across the PARSEC benchmarks. EDP values are normalized wrt the baseline hybrid PCM system.

Figure 113(a) and Figure 113(b) give CPI values and EDP values, respectively, for various hybrid PCM systems averaged over the PARSEC benchmarks. The error bars in the figure represent standard deviation of CPI values across the PARSEC benchmarks. *O-pSet-PISA-DyPhase* yields about 3.7%, 3.2%, and 0.3% less CPI over *Reset-pSet-ISA-DyPhase, O-pSet-ISA-DyPhase, and Reset-pSet-PISA-DyPhase*, respectively. Moreover, *O-pSet-PISA-DyPhase* yields about 44%, 39.4%, and 6.4% less EDP over *Reset-pSet-ISA-DyPhase, O-pSet-ISA-DyPhase, and Reset-pSet-PISA-DyPhase*, respectively.

As described earlier in Section 13.6, *ISA* technique monitors the resistance of each PCM cell after every write operation, and reactively triggers a healing procedure when the cell-resistance drops below a certain threshold value. To monitor PCM cell-resistance after every cache-line write, *ISA* schedules a dummy read of the page that contains the recently written cache-line. In contrast, *PISA* proactively heals degraded memory cells without having to monitor their resistance. This prevents *PISA* from spending extra energy and time on monitoring cell-resistance, which results in superior CPI and EDP values for *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* compared to *O-pSet-ISA-DyPhase* and *Reset-pSet-ISA-DyPhase*. In summary, *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* based hybrid memory systems yield 124 and 126 years more lifetime, 8.3% and 8% less CPI, and 44.3% and 40.4% less EDP on average over the *WOMC_PCM*, *pSET_PCM_QD8* and the baseline PCM based hybrid memory systems. Furthermore, *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* based hybrid memory systems yield 41.7% and 35.2% less EDP, and 3.5% and 3.2% less CPI on average over the *O-pSet-ISA-DyPhase* and *Reset-pSet-ISA-DyPhase* memory systems. Moreover, our *PISA* technique improves the lifetime of *O-pSet DyPhase* and *Reset-pSet DyPhase* PCMs by about 50 times. Thus, it can be concluded that our *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* architectures provide very promising low-latency, energy-efficient, and highly durable solutions for future PCM implementations.

13.8. CONCLUSIONS

This chapter presented a novel PCM architecture named *DyPhase*, which uses partial-SET operations to reduce the average write latency. To remedy retention-related reliability issues associated with partial-SET operations, we presented *O-pSet Refresh* and *Reset-pSet Refresh* techniques to periodically refresh the stored data in the *DyPhase* PCM architecture. Unfortunately,

the use of periodic refresh operations increases the write rate of the memory, which in turn accelerates memory degradation and decreases its lifetime. *DyPhase* overcomes this shortcoming by utilizing a proactive in-situ self-annealing (*PISA*) technique that periodically heals degraded memory cells, resulting in decelerated degradation and increased memory lifetime. Our evaluation with PARSEC benchmarks indicate that PISA-enabled *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* based hybrid DRAM-PCM systems yield about 124 and 126 years more lifetime, 8.3% and 8% less CPI, and 44.3% and 40.4% less EDP on average over the *WOMC_PCM*, *pSET_PCM_QD8* and the baseline PCM based hybrid memory systems.

Significant improvements in latency, energy-efficiency, and durability demonstrated by our *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* architectures position them as promising solutions for future PCMs. Our proactive scrubbing techniques (i.e., *O-pSet/Reset-pSet Refresh* and *PISA*) make SLC *DyPhase* PCMs more effective in tolerating drift-induced soft errors and wear-induced hard errors. These techniques can be easily extended with low overhead and be used to scrub (refresh and heal) MLC PCMs as well. Thus, with potential extensions to MLC PCMs, our scrubbing techniques enabled *DyPhase* architectures can become even more promising solutions for future PCM implementations.

14. CONCLUSIONS AND FUTURE WORK

14.1. CONCLUSIONS

We addressed several design challenges faced by photonic interconnects and emerging memory (3D-stacked DRAM and Phase Change Memory) subsystems by proposing a design and optimization framework. Our framework contributes several cross-layer solutions that combine enhancements at the circuit level, microarchitecture level, and system level to improve the bandwidth, reliability, energy-efficiency, and security of photonic interconnects. Moreover, it also contributes solutions that combine optimizations in the throughput, concurrency, access latency, energy-efficiency, and endurance of emerging memory (3D-stacked DRAM and Phase Change Memory) subsystems. Experimental results for our proposed framework corroborate the capability of our proposed solutions in addressing key design challenges of photonic interconnects and emerging memory subsystems. Therefore, our proposed framework has the potential to be applied as a general strategy for performance, reliability, power, and security management of emerging interconnection and memory subsystems.

As our first contribution, we presented a novel cross-layer crosstalk-mitigation framework *HYDRA* that reduces crosstalk noise in the detectors of DWDM-based PNoC architectures. Our proposed *HYDRA* framework seamlessly integrates two device-layer and a circuit layer technique to enable interesting tradeoffs between reliability, performance, and energy overheads for the Corona, Firefly, and Flexishare crossbar-based PNoC architectures. Our simulation-based analysis shows that the HYDRA framework improves worst case OSNR by up to 5.3× compared to the baseline architectures, and by up to 1.14× compared to the best known PNoC crosstalk-mitigation

scheme from prior work. Thus, HYDRA is an attractive solution to enhance reliability in emerging DWDM-based PNoCs.

Next, we contributed a lightweight low overhead homodyne crosstalk mitigation (HCTM) technique for DWDM PNoCs. The HCTM technique is agnostic to the time-dependent characteristics of the homodyne crosstalk. Moreover, it also shows interesting trade-offs between reliability and energy overhead. We evaluate the effectiveness and overhead of our HCTM technique by implementing it for well-known PNoC architectures, including Corona, Firefly and Flexishare. Our experimental analysis shows that our approach when implemented on these PNoCs can improve the worst-case SNR by up to 37.6% compared to the baseline versions of these PNoCs, thereby significantly enhancing reliability, at the cost of up to 19.2% energy overhead and 1.7% photonic area overhead. Thus, HCTM represents an attractive solution to enhance reliability in emerging DWDM-based PNoCs.

In this dissertation, we have also presented LIBRA framework that combines two novel dynamic thermal management mechanisms for the reduction of maximum on-chip temperature and conservation of trimming and tuning power of MRs in DWDM-based PNoC architectures. These techniques (TPMA at the device-level, VADTM at the system-level) constitute a hybrid reactive-proactive management framework that demonstrated interesting trade-offs between performance and power/energy across two different state-of-the-art crossbar-based PNoC architectures. Our experimental analysis on the well-known Corona and Flexishare PNoC architectures has shown that LIBRA can notably conserve total power by up to 61.3% (trimming and tuning power by up to 76.2%) and total energy by up to 57.3%.

Moreover, we presented a detailed comparative analysis of a number of design tradeoffs for CMOS front-end (FCSP) and back-end (BCSP) compatible silicon photonic devices. The results

of the cross-layer optimization of multiple device-level and link-level design parameters indicate that BCSP interconnects yield more throughput with comparable energy-efficiency compared to FCSP interconnects. The optimized design of BCSP-based Firefly and Corona photonic network-on-chips (PNoCs) yield 1.15× and 3.5× greater throughput with 12.4% and 39.5% more energy-efficiency than the optimized design of FCSP-based Firefly and Corona PNoCs respectively. The greater throughput and comparable energy-efficiency obtained for BCSP links favor their use in the terabyte-per-second scale silicon photonic interconnects in future PNoCs.

Furthermore, we presented a low overhead, run-time laser power management technique called SOA_LPM, which made use of on-chip semiconductor amplifiers (SOA) to achieve traffic-independent and loss-aware savings in laser power consumption. Experimental analysis shows that our technique achieves 31.5% more laser power savings with 12.8% less latency overhead compared to another laser power management scheme from prior work. Thus, SOA_LPM represents an attractive solution to reduce laser power consumption in emerging PNoCs.

We presented a novel method, called 4-PAM-P, for generating 4-PAM optical signals in PNoCs, which can double the aggregate bandwidth without increasing utilized wavelengths, photonic hardware, and incurred noise, thereby improving the bit-error-rate (BER), area-efficiency, and energy-efficiency of PNoCs. Our analysis shows that our 4-PAM-P method achieves equal bandwidth with 4.2× better BER, 19.5% lower power, 16.3% lower energy-per-bit, and 5.6% less photonic area compared to the best known 4-PAM optical signaling method (4-PAM-SS) from prior work. Moreover, our 4-PAM-P method achieves equal bandwidth with 1.5× better BER, 16.9% lower power, 14.6% lower EPB, and 10.6% less photonic area compared to the conventional OOK signaling method. These results corroborate the excellent capabilities of our

proposed 4-PAM-P method in achieving high-bandwidth data transfers in PNoCs with greater reliability, area- and energy-efficiency.

VBTI aging in the MRs used in photonic interconnects, and the dependence of this aging on voltage bias and temperature was also analyzed in this dissertation. We presented an analytical model for trap generation on the MR core-cladding boundary with VBTI aging in MRs. We also considered the impact of process variations on aging. Our device-level results indicate that MR aging causes significant degradation in MR Q-factor and incurs notable resonance wavelength red shift. We extended our MR aging analysis to the system-level for two crossbar-based PNoCs. The system-level analysis on these PNoCs clearly shows the damaging effects of MR aging with a worst-case signal loss increase of up to 7.6dB and EDP increase of up to 26.8%.

A novel security enhancement framework called SOTERIA that secures data during unicast communications in DWDM-based PNoC architectures from snooping attacks was also presented in this dissertation. The SOTERIA framework shows interesting trade-offs between security, performance, and energy overheads for DWDM-based PNoC architectures. Our analysis shows that SOTERIA enables hardware security in crossbar based PNoCs with minimal overheads of up to 10.6% in average latency and up to 13.3% in EDP compared to the baseline PNoCs. Thus, SOTERIA represents an attractive solution to enhance hardware security in emerging DWDM-based PNoCs.

Moreover, we introduced 3D-ProWiz, a novel high bandwidth and low-latency 3D DRAM architecture. 3D-ProWiz integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism than other 3D DRAM architectures. The 3D vertical routing of the internal memory bus using TSVs at subarray-level granularity and fanout buffers enable 3D-ProWiz to use smaller dimension subarrays without significant area overhead.

This in turn reduces the random-access latency and activation-precharge energy. Consequently, 3D-ProWiz yields on average 52%, 41.9% and 80.6% improvements in power consumption, latency and energy-delay product (EDP) respectively over other DRAM architectures. The significant improvements demonstrated by 3D-ProWiz position it as a promising architecture for future DRAMs.

We also proposed a new refresh technique for the 3Dstacked Hybrid Memory Cube (HMC) DRAM called massed refresh that yields 6.3% and 5.8% improvements in throughput and EDP on average over the JEDEC standardized distributed per-bank refresh and the state-of-the-art scattered refresh schemes. These promising results indicate that our massed refresh technique can significantly reduce the overhead of distributed refresh operations in the HMC (and other DRAMs) by improving their throughput and energy-efficiency.

Lastly, we presented a novel PCM architecture named *DyPhase*, which uses partial-SET operations to reduce the average write latency. To remedy retention-related reliability issues associated with partial-SET operations, we presented *OpSet Refresh* and *Reset-pSet Refresh* techniques to periodically refresh the stored data in the *DyPhase* PCM architecture. Unfortunately, the use of periodic refresh operations increases the write rate of the memory, which in turn accelerates memory degradation and decreases its lifetime. *DyPhase* overcomes this shortcoming by utilizing a proactive in-situ self-annealing (PISA) technique that periodically heals degraded memory cells, resulting in decelerated degradation and increased memory lifetime. Our evaluation with PARSEC benchmarks indicate that PISA-enabled *O-pSet-PISA-DyPhase* and *ResetpSet-PISA-DyPhase* based hybrid DRAM-PCM systems yield about 124 and 126 years more lifetime, 8.3% and 8% less CPI, and 44.3% and 40.4% less EDP on average over the *WOMC_PCM*, *pSET_PCM_QD8* and the baseline PCM based hybrid memory systems. Significant improvements

in latency, energy-efficiency, and durability demonstrated by our *O-pSet-PISA-DyPhase* and *Reset-pSet-PISA-DyPhase* architectures position them as promising solutions for future PCMs.

14.2. SUGGESTIONS FOR FUTURE RESEARCH

The design of photonic interconnects and emerging memory subsystems will continue to face new challenges as the advanced manycore computing architectures are expected to scale significantly in the near future. Considering this fact, we envision the following likely directions for future work.

- *Fault Tolerant PNoCs:* Faults are inevitable not only in electrical interconnects, but also in photonic interconnects. Faults in PNoCs include, photonic waveguide faults, MR faults, and splitter faults due to aging and/or other mechanisms. There is a need to explore these faults in PNoCs and novel strategies are needed to reduce fault-induced performance penalties. A cross-layer approach combining architectural-level enhancements and system-level application scheduling will likely also be beneficial.
- *Aging in MRs:* In this dissertation, we already explored VBTI aging in MR's PN junctions (see chapter 8) and analyzed its impact on PNoC architectures. In addition to VBTI, MRs are prone to other aging mechanisms such as hot carrier injection (possible in PN junction of an MR) however these aging scenarios have not been explored yet.
- *Hardware Security in PNoCs:* Security is expected to be a critical concern in CMPs that use DWDM-based PNoCs for inter-core communication. Mechanisms to mitigate snoop-based attacks on PNoCs are already discussed in this dissertation (see chapter 9). Furthermore, novel strategies are needed to mitigate snoop-based attacks in multicast- and broadcast-enabled PNoC architectures. However, PNoCs are also vulnerable to Denial-of-service (DoS) based attacks and data corruption based attacks. Therefore, solutions are

needed to reduce the aforementioned security risks in PNoCs to further enhance their hardware security.

- *Variation Resilient PNoCs*: We envision finding new efficient solutions to the variation resilience and high signal-power loss problems in PNoCs through the following objectives. First, novel designs of CMOS-compatible photonic devices with significantly improved variation resilience and loss characteristics compared to the state-of-the-art SOI platform based devices can be invented. The research challenge in designing such devices is tailoring their physical size, performance, loss characteristics, and variation susceptibility at the device level, while minimizing their implementation adversities and related costs at the system level. To overcome these challenges, novel device structures, dimensions and material platforms that are inherently less lossy and susceptible to variations can be considered. These new photonic devices can be fabricated and characterized to verify their feasibility and efficiency. These new photonic devices will provide wider design space on the one hand. But on the other hand, these devices will impose new sets of design tradeoffs. The new design space can be carefully explored and the new sets of design tradeoffs can be examined to identify the best combination of design choices (i.e., best-suitable devices with best design parameters) that can provide the required variation resilience and signal loss values with minimal overheads. Moreover, leveraging the knowledge obtained from these objectives, novel architectures of data-parallel photonic interconnects with optimized energy-efficiency and reliability can be designed.
- *Optimizing Emerging Memories*: We envision improving the endurance, throughput, latency, and energy-efficiency of emerging phase change memory and magnetoresistive memory subsystems by orders of magnitude through the following objectives. To improve

the fundamental endurance, energy, and latency of phase change memory and magnetoresistive memory, novel memory cell structures can be invented that inherently attain high endurance, low access energy, and low write latency. Apart from the fundamental memory cell structure, another main contributor to the per-access latency and energy of a memory subsystem is the capacitive loading of the data access path. To significantly reduce the capacitive loading of the data access path, I will invent novel memory architectures can be invented that make use of emerging integration technologies, such as the tight-pitched metal inter-layer via based monolithic 3D integration technology and the near-field coupling based through-chip interface technology. Reducing the per-access latency will also improve the throughput of our proposed memory architectures. Another way of improving memory throughput is to increase data access parallelism by enabling fine-grained activation and access of data arrays. However, increasing the access parallelism of a memory module can increase its power dissipation. Therefore, to maximize the power capacity and access parallelism of the memory modules, their fine-grained data-array organizations and power delivery networks should be co-designed.

- *Memory-Centric Computing:* With advancements in high-speed photonic interconnects and non-volatile memory subsystems, it will be possible to collapse the traditional memory hierarchy and access various memory levels with near-uniform latency in future manycore computing systems. This will open new opportunities for long-term research in the field of memory-centric computing. The idea of memory-centric computing is to employ a massive pool of non-volatile memory at the center, which makes the big data it stores accessible to multiple compute resources with low energy and near-uniform low latency using high-speed interconnects. Memory-centric computing can eliminate the inefficiencies of the

traditional processor-centric architecture to provide massive benefits in the latency and energy costs of concurrent data communications in future computing systems.

BIBLIOGRAPHY

[1] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, vol. 38, no. 8, pp.114, 1965," *IEEE Solid-State Circuits Soc. Newsl.*, vol. 11, no. 3, pp. 33–35, Sep. 2006.

[2] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideovt, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Solid-State Circuits Soc. Newsl.*, vol. 12, no. 1, pp. 38–50, Winter 2007.

[3] Intel Corporation, "Intel® Xeon Phi™ Processor 7290F."

[4] Mellanox Technologies, "TILE-Gx72 Processor."

[5] Kalray, "KALRAY COMPUTE CORE."

[6] Ambric, "Am2045 Processor."

[7] J. C. McCalpin, "Memory Bandwidth and System Balance in HPC Systems," *Keynote at ACM/IEEE Supercomputing Conference (SC)*, 2006.

[8] L. Zhou and A. K. Kodi, "PROBE: Prediction-based optical bandwidth scaling for energy-efficient NoCs," *IEEE/ACM International Symposium on Networks on Chip (NOCS)*, 2013, pp. 1-8.

[9] D. A. B. Miller, "Device Requirements for Optical Interconnects to Silicon Chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[10] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch, "Disaggregated Memory for Expansion and Sharing in Blade Servers," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2009, pp. 267–278.

[11] J. L. Abellán, A. K. Coskun, A. Gu, W. Jin, A. Joshi, A. B. Kahng, J. Klamkin, C. Morales, J. Recchio, V. Srinivas, and T. Zhang, "Adaptive Tuning of Photonic Devices in a Photonic NoC Through Dynamic Workload Allocation," *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.*, vol. 36, no. 5, pp. 801–814, May 2017.

[12] S. Bahirat and S. Pasricha, "METEOR: Hybrid Photonic Ring-mesh Network-on-chip for Multicore Architectures," *ACM Trans Embed Comput Syst*, vol. 13, no. 3s, pp. 116:1–116:33, Mar. 2014.

[13] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson, "12.5 Gbit/s carrier-injection-based silicon micro-ring silicon modulators," *Opt. Express*, vol. 15, no. 2, pp. 430–436, Jan. 2007.

[14] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, "Corona: System Implications of Emerging Nanophotonic Technology," *IEEE/ACM International Symposium on Computer Architecture (ISCA), 2008*, pp. 153–164.

[15] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary, "Firefly: Illuminating Future Network-on-chip with Nanophotonics," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2009, pp. 429–440.

[16] Y. Pan, J. Kim, and G. Memik, "FlexiShare: Channel sharing for an energy-efficient nanophotonic crossbar," *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2010, pp. 1–12.

[17] S. Pasricha and N. Dutt, "On-Chip Communication Architectures, Volume - - 1st Edition." [Online]. Available: https://www.elsevier.com/books/on-chip-communication-architectures/pasricha/978-0-12-373892-9. [Accessed: 02-May-2018].

[18] Y. Vlasov, W. M. J. Green, and F. Xia, “High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks,” *Nat. Photonics*, vol. 2, no. 4, pp. 242–246, Apr. 2008.

[19] G. Chen, H. Chen, M. Haurylau, N. A. Nelson, D. H. Albonesi, P. M. Fauchet, and E. G. Friedman, “Predictions of CMOS Compatible On-chip Optical Interconnect,” *ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, 2005, pp. 13–20.

[20] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, “Micrometre-scale silicon electro-optic modulator,” *nature*, vol. 435, no. 7040, pp. 325–327, Jan. 2005.

[21] J. Ahn, M. Fiorentino, R. G. Beausoleil, N. Binkert, A. Davis, D. Fattal, N. P. Jouppi, M. McLaren, C. M. Santori, R. S. Schreiber, S. M. Spillane, D. Vantrease, and Q. Xu, “Devices and architectures for photonic chip-scale integration,” *Appl. Phys. A*, vol. 95, no. 4, pp. 989–997, Jun. 2009.

[22] C. Nitta, M. Farrens, and V. Akella, “Addressing system-level trimming issues in on-chip nanophotonic networks,” *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2011, pp. 122–131.

[23] J. Pawlowski, “Hybrid Memory Cube (HMC),” *Hot Chips*, 2011.

[24] “JEDEC STANDARD (JESD229): Wide I/O Single Data Rate.” JEDEC SOLID STATE TECHNOLOGY ASSOCIATION, Dec-2011.

[25] AMD, “High-Bandwidth Memory: Reinventing Memory Technology.”

[26] D. Chapman, “DiRAM Architecture Overview.” Tezzaron Semiconductors.

[27] I. G. Thakkar and S. Pasricha, “3D-ProWiz: An Energy-Efficient and Optically-Interfaced 3D DRAM Architecture with Reduced Data Access Overhead,” *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 1, no. 3, pp. 168–184, Jul. 2015.

[28] F. Bedeschi, R. Fackenthal, C. Resta, E. M. Donze, M. Jagasivamani, E. C. Buda, F. Pellizzer, D. W. Chow, A. Cabrini, and G. M. A. Calvi, "A Bipolar-Selected Phase Change Memory Featuring Multi-Level Cell Storage," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 217–227, Jan. 2009.

[29] B. C. Lee, P. Zhou, J. Yang, Y. Zhang, B. Zhao, E. Ipek, O. Mutlu, and D. Burger, "Phase-Change Technology and the Future of Main Memory," *IEEE IEEE/ACM International Symposium on Microarchitecture (MICRO)*, vol. 30, no. 1, pp. 143–143, Jan. 2010.

[30] Numonyx, "Phase Change Memory and Its Impacts on Memory Hierarchy," *Seminar at Carnegie Mellon University*, Sep-2009.

[31] "ITRS Report, 2013 Edition," International Technology Roadmap for Semiconductors (ITRS).

[32] C. Batten, A. Joshi, J. Orcutt, C. Holzwarth, M. Popovic, J. Hoyt, F. Kartner, R. Ram, V. Stojanovic, and K. Asanovic, "Building Many-Core Processor-to-DRAM Networks with Monolithic CMOS Silicon Photonics," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, vol. 29, no. 4, pp. 8–21, Jul. 2009.

[33] S. K. Selvaraja, "Wafer-Scale Fabrication Technology for Silicon Photonic Integrated Circuits," PhD Thesis, Ghent University, 2011.

[34] Z. Li, M. Mohamed, X. Chen, E. Dudley, K. Meng, L. Shang, A. R. Mickelson, R. Joseph, M. Vachharajani, B. Schwartz, and Y. Sun, "Reliability Modeling and Management of Nanophotonic On-Chip Networks," *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 20, no. 1, pp. 98–111, Jan. 2012.

[35] P. Dong, S. Liao, D. Feng, H. Liang, D. Zheng, R. Shafliha, C.-C. Kung, W. Qian, G. Li, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, "Low V_{pp}, ultralow-energy, compact, high-

speed silicon electro-optic modulator," *Opt. Express*, vol. 17, no. 25, pp. 22484–22490, Dec. 2009.

[36] S. V. R. Chittamuru, S. Desai, and S. Pasricha, "SWIFTNoC: A Reconfigurable Silicon-Photonic Network with Multicast-Enabled Channel Sharing for Multicore Architectures," *J Emerg Technol Comput Syst*, vol. 13, no. 4, pp.58:1-58:27, 2017.

[37] W. S. Khwa, J. Y. Wu, T. H. Su, H. P. Li, M. BrightSky, T. Y. Wang, T. H. Hsu, P. Y. Du, S. Kim, W. C. Chien, H. Y. Cheng, R. Cheek, E. K. Lai, Y. Zhu, M. H. Lee, M. F. Chang, H. L. Lung, and C. Lam, "A novel inspection and annealing procedure to rejuvenate phase change memory from cycling-induced degradations for storage class memory applications," *IEEE International Electron Devices Meeting (IEDM)*, 2014, pp. 29.8.1-29.8.4.

[38] S. V. R. Chittamuru, I. G. Thakkar, and S. Pasricha, "HYDRA: Heterodyne Crosstalk Mitigation With Double Microring Resonators and Data Encoding for Photonic NoCs," *IEEE Trans. Very Large Scale Integr*, vol. 26, no. 1, pp. 168–181, 2018.

[39] I. G. Thakkar, S. V. R. Chittamuru, and S. Pasricha, "Mitigation of homodyne crosstalk noise in silicon photonic NoC architectures with tunable decoupling," in *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2016, pp. 1–10.

[40] I. G. Thakkar, S. V. R. Chittamuru, and S. Pasricha, "A comparative analysis of front-end and back-end compatible silicon photonic on-chip interconnects," *ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, 2016, pp. 1–8.

[41] I. G. Thakkar, S. V. R. Chittamuru, and S. Pasricha, "Run-time laser power management in photonic NoCs with on-chip semiconductor optical amplifiers," *IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, 2016, pp. 1–4.

[42] I. G. Thakkar, S. V. R. Chittamuru, and S. Pasricha, “Improving the Reliability and Energy-Efficiency of High-Bandwidth Photonic NoC Architectures with Multilevel Signaling,” *IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, 2017, p. 4:1–4:8.

[43] S. V. R. Chittamuru, I. Thakkar, and S. Pasricha, “Analyzing voltage bias and temperature induced aging effects in photonic interconnects for manycore computing,” *ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, 2017.

[44] S. V. R. Chittamuru, I. Thakkar, V. Bhat, and S. Pasricha, “SOTERIA: Exploiting Process Variations to Enhance Hardware Security With Photonic NoC Architectures,” *IEEE/ACM Design Automation Conference (DAC)*, 2018.

[45] I. G. Thakkar and S. Pasricha, “A novel 3D graphics DRAM architecture for high-performance and low-energy memory accesses,” *IEEE International Conference on Computer Design (ICCD)*, 2015, pp. 467–470.

[46] I. G. Thakkar and S. Pasricha, “Massed Refresh: An Energy-Efficient Technique to Reduce Refresh Overhead in Hybrid Memory Cube Architectures,” *IEEE International Conference on VLSI Design and International Conference on Embedded Systems (VLSID)*, 2016, pp. 104–109.

[47] I. G. Thakkar and S. Pasricha, “DyPhase: A Dynamic Phase Change Memory Architecture with Symmetric Write Latency and Restorable Endurance,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, pp. 1–1, 2017.

[48] L. H. K. Duong, M. Nikdast, J. Xu, Z. Wang, Y. Thonnart, S. L. Beux, P. Yang, X. Wu, and Z. Wang, “Coherent crosstalk noise analyses in ring-based optical interconnects,” *IEEE/ACM Design, Automation Test in Europe Conference Exhibition (DATE)*, 2015, pp. 501–506.

[49] M. Bahadori, D. Nikolova, S. Rumley, C. P. Chen, and K. Bergman, “Optimization of Microring-based Filters for Dense WDM Silicon Photonic Interconnects,” *Measurement*, vol. 4, p. 5.

[50] B. G. Lee, B. A. Small, K. Bergman, Q. Xu, and M. Lipson, “Transmission of high-data-rate optical signals through a micrometer-scale silicon ring resonator,” *Opt. Lett.*, vol. 31, no. 18, pp. 2701–2703, Sep. 2006.

[51] K. Padmaraju, X. Zhu, L. Chen, M. Lipson, and K. Bergman, “Intermodulation crosstalk characteristics of WDM silicon microring modulators,” *IEEE Photonics Technol. Lett.*, vol. 26, no. 14, pp. 1478–1481, 2014.

[52] S. V. R. Chittamuru, I. G. Thakkar, and S. Pasricha, “Process variation aware crosstalk mitigation for DWDM based photonic NoC architectures,” *IEEE International Symposium on Quality Electronic Design (ISQED)*, 2016, pp. 57–62.

[53] C. Sun, M. Wade, Y. Lee, J. Orcutt, L. Alloatti, M. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin, B. R. Moss, R. Kumar, F. Pavanello, A. H. Atabaki, H. M. Cook, A. J. Ou, J. C. Leu, Y.-H. Chen, K. Asanovic, R. J. Ram, M. A. Popovic, and V. M. Stojanovic, “Single-chip microprocessor that communicates directly using light,” *Nature*, vol. 528, no. 7583, pp. 534–538, Dec. 2015.

[54] R. G. Beausoleil, “Large-scale Integrated Photonics for High-performance Interconnects,” *J Emerg Technol Comput Syst*, vol. 7, no. 2, pp. 6:1–6:54, Jul. 2011.

[55] P. K. Tien, “Light Waves in Thin Films and Integrated Optics,” *Appl. Opt.*, vol. 10, no. 11, pp. 2395–2413, Nov. 1971.

[56] S. V. R. Chittamuru, I. G. Thakkar, and S. Pasricha, "PICO: Mitigating heterodyne crosstalk due to process variations and intermodulation effects in photonic NoCs," *ACM/EEE Design Automation Conference (DAC)*, 2016, pp. 1–6.

[57] R. Hendry, D. Nikolova, S. Rumley, N. Ophir, and K. Bergman, "Physical layer analysis and modeling of silicon photonic WDM bus architectures," *HiPEAC Workshop*, 2014, pp. 20–22.

[58] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, "Cascaded silicon micro-ring modulators for WDM optical interconnection," *Opt. Express*, vol. 14, no. 20, pp. 9431–9436, Oct. 2006.

[59] D. Vantrease, N. Binkert, R. Schreiber, and M. H. Lipasti, "Light speed arbitration and flow control for nanophotonic interconnects," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 304–315.

[60] L. H. K. Duong, M. Nikdast, S. L. Beux, J. Xu, X. Wu, Z. Wang, and P. Yang, "A Case Study of Signal-to-Noise Ratio in Ring-Based Optical Networks-on-Chip," *IEEE Des. Test*, vol. 31, no. 5, pp. 55–65, Oct. 2014.

[61] S. V. R. Chittamuru and S. Pasricha, "Crosstalk Mitigation for High-Radix and Low-Diameter Photonic NoC Architectures," *IEEE Des. Test*, vol. 32, no. 3, pp. 29–39, Jun. 2015.

[62] S. V. R. Chittamuru and S. Pasricha, "Improving crosstalk resilience with wavelength spacing in photonic crossbar-based network-on-chip architectures," *IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2015, pp. 1–4.

[63] Y. Xu, J. Yang, and R. Melhem, "Tolerating process variations in nanophotonic on-chip networks," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2012, pp. 142–152.

[64] W. Bogaerts, P. D. Heyn, T. V. Vaerenbergh, K. D. Vos, S. K. Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. V. Thourhout, and R. Baets, "Silicon microring resonators," *Laser Photonics Rev.*, vol. 6, no. 1, pp. 47–73, 2012.

[65] "Lumerical Inc." [Online]. Available: http://www.lumerical.com/tcad-products/mode/.

[66] J. Heebner, "Nonlinear optical whispering gallery microresonators for photonics," 2003.

[67] D. A. Neamen;, *Semiconductor Physics And Devices: Basic Principles*, 3 edition. McGraw-Hill Higher Education, 2002.

[68] G. T. Reed and A. P. Knights, *Silicon Photonics: An Introduction*. John Wiley & Sons, 2004.

[69] K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," *Nanophotonics*, vol. 3, no. 4–5, pp. 269–281, 2014.

[70] C. Li, R. Bai, A. Shafik, E. Z. Tabasy, B. Wang, G. Tang, C. Ma, C.-H. Chen, Z. Peng, M. Fiorentino, R. G. Beausoleil, P. Chiang, and S. Palermo, "Silicon Photonic Transceiver Circuits With Microring Resonator Bias-Based Wavelength Stabilization in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 49, no. 6, pp. 1419–1436, Jun. 2014.

[71] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "VARIUS: A Model of Process Variation and Resulting Timing Errors for Microarchitects," *IEEE Trans. Semicond. Manuf.*, vol. 21, no. 1, pp. 3–13, Feb. 2008.

[72] S. Xiao, M. H. Khan, H. Shen, and M. Qi, "Modeling and measurement of losses in silicon-on-insulator resonators and bends," *Opt. Express*, vol. 15, no. 17, p. 10553, 2007.

[73] N. Ophir, C. Mineo, D. Mountain, and K. Bergman, "Silicon Photonic Microring Links for High-Bandwidth-Density, Low-Power Chip I/O," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, vol. 33, no. 1, pp. 54–67, Jan. 2013.

[74] D. G. Rabus, *Integrated Ring Resonators: The Compendium*. Berlin Heidelberg: Springer-Verlag, 2007.

[75] C. Sun, C. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanoic, "DSENT - A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling," *IEEE/ACM International Symposium on Networks on Chip (NOCS)*, 2012, pp. 201–210.

[76] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC Benchmark Suite: Characterization and Architectural Implications," *International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2008, pp. 72–81.

[77] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The Gem5 Simulator," *SIGARCH Comput Arch. News*, vol. 39, no. 2, pp. 1–7, Aug. 2011.

[78] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "CACTI 6.0: A Tool to Understand Large Caches," 2007.

[79] X. Zheng, D. Patil, J. Lexau, F. Liu, G. Li, H. Thacker, and Y. Luo, "Ultra-efficient 10Gb/s hybrid integrated silicon photonic transmitter and receiver," *Opt. Express*, vol. 19, no. 6, p. 5172, Mar. 2011.

[80] P. Grani and S. Bartolini, "Design Options for Optical Ring Interconnect in Future Client Devices," *J Emerg Technol Comput Syst*, vol. 10, no. 4, p. 30:1–30:25, Jun. 2014.

[81] S. Pasricha and S. Bahirat, "OPAL: A multi-layer hybrid photonic NoC for 3D ICs," *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2011, pp. 345–350.

[82] M. Nikdast, J. Xu, L. H. K. Duong, X. Wu, Z. Wang, X. Wang, and Z. Wang, "Fat-Tree-Based Optical Interconnection Networks Under Crosstalk Noise Constraint," *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 23, no. 1, pp. 156–169, Jan. 2015.

[83] Y. Xie, M. Nikdast, J. Xu, W. Zhang, Q. Li, X. Wu, Y. Ye, X. Wang, and W. Liu, "Crosstalk noise and bit error rate analysis for optical network-on-chip," *ACM/IEEE Design Automation Conference (DAC)*, 2010, pp. 657–660.

[84] S. D. Dods, J. P. R. Lacey, and R. Tucker, "Homodyne crosstalk in WDM ring and bus networks," *IEEE Photonics Technol. Lett.*, vol. 9, no. 9, pp. 1285–1287, Sep. 1997.

[85] E. Tangdiongga, I. T. Monroy, R. Jonker, and H. De Waardt, "Experimental evaluation of optical crosstalk mitigation using phase scrambling," *IEEE Photonics Technol. Lett.*, vol. 12, no. 5, pp. 567–569, May 2000.

[86] Y. Shen, K. Lu, and W. Gu, "Coherent and incoherent crosstalk in WDM optical networks," *J. Light. Technol.*, vol. 17, no. 5, pp. 759–764, May 1999.

[87] N. Dutt, S. Pasricha, "Trends in Emerging On-Chip Interconnect Technologies," *Ipsj Trans. Syst. Lsi Des. Methodol.*, vol. 1, pp. 2–17, 2008.

[88] F. Xia, M. Rooks, L. Sekaric, and Y. Vlasov, "Ultra-compact high order ring resonator filters using submicron silicon photonic wires for on-chip optical interconnects," *Opt. Express*, vol. 15, no. 19, p. 11934, 2007.

[89] Q. Li, M. Soltani, S. Yegnanarayanan, and A. Adibi, "Design and demonstration of compact, wide bandwidth coupled-resonator filters on a silicon-on- insulator platform," *Opt. Express*, vol. 17, no. 4, p. 2247, Feb. 2009.

[90] M. Nikdast, J. Xu, X. Wu, W. Zhang, Y. Ye, X. Wang, Z. Wang, and Z. Wang, "Systematic Analysis of Crosstalk Noise in Folded-Torus-Based Optical Networks-on-Chip," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 33, no. 3, pp. 437–450, Mar. 2014.

[91] Y. Xie, M. Nikdast, J. Xu, X. Wang, Z. Wang, and W. Liu, "Formal Worst-Case Analysis of Crosstalk Noise in Mesh-Based Optical Networks-on-Chip," *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 21, no. 10, pp. 1823–1836, Oct. 2013.

[92] B.-C. Lin and C.-T. Lea, "Crosstalk Analysis for Microring Based Optical Interconnection Networks," *J. Light. Technol.*, vol. 30, no. 15, pp. 2415–2420, Aug. 2012.

[93] J. Heebner, R. Grover, and T. Ibrahim, *Optical Microresonators: Theory, Fabrication, and Applications*. New York: Springer-Verlag, 2008.

[94] R. Wu, C.-H. Chen, J.-M. Fedeli, M. Fournier, K.-T. Cheng, and R. G. Beausoleil, "Compact models for carrier-injection silicon microring modulators," *Opt. Express*, vol. 23, no. 12, p. 15545, Jun. 2015.

[95] K. J. Vahala, "Optical microcavities," *Nature*, vol. 424, no. 6950, pp. 839–846, Aug. 2003.

[96] R. A. Soref and B. R. Bennett, "Electrooptical effects in silicon," *IEEE J. Of Quantum Electron.*, vol. 23, no. 1, pp. 123–129, 1987.

[97] Y. G. Boucher, "Analytical model for the coupling constant of a directional coupler in terms of slab waveguides," *Opt. Eng.*, vol. 53, no. 7, p. 71810, 2014.

[98] A. Kumar, L. Shang, L.-S. Peh, and N. K. Jha, "HybDTM: a coordinated hardware-software approach for dynamic thermal management," *ACM/IEEE Design Automation Conference (DAC)*, 2006, pp. 548–553.

[99] S. V. R. Chittamuru and S. Pasricha, “SPECTRA: A Framework for Thermal Reliability Management in Silicon-Photonic Networks-on-Chip,” *International Conference on VLSI Design and International Conference on Embedded Systems (VLSID)*, 2016, pp. 86–91.

[100] T. Zhang, J. L. Abellán, A. Joshi, and A. K. Coskun, “Thermal management of manycore systems with silicon-photonic networks,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2014, pp. 1–6.

[101] I. Yeo, C. C. Liu, and E. J. Kim, “Predictive Dynamic Thermal Management for Multicore Systems,” *IEEE/ACM Design Automation Conference (DAC)*, 2008, pp. 734–739.

[102] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanovic, and V. Stojanovic, “Silicon-photonic clos networks for global on-chip communication,” *ACM/IEEE International Symposium on Networks-on-Chip (NOCS)*, 2009, pp. 124–133.

[103] H. Hanson, S. W. Keckler, S. Ghiasi, K. Rajamani, F. Rawson, and J. Rubio, “Thermal response to DVFS: analysis with an Intel Pentium M,” *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2007, pp. 219–224.

[104] V. Hanumaiah, S. Vrudhula, and K. S. Chatha, “Maximizing performance of thermally constrained multi-core processors by dynamic voltage and frequency control,” *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2009, pp. 310–313.

[105] S. Herbert and D. Marculescu, “Analysis of dynamic voltage/frequency scaling in chip-multiprocessors,” *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2007, pp. 38–43.

[106] K. K. Rangan, G.-Y. Wei, and D. Brooks, “Thread Motion: Fine-grained Power Management for Multi-core Systems,” *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2009, pp. 302–313.

[107] E. Ipek, M. Kirman, N. Kirman, and J. F. Martinez, “Core Fusion: Accommodating Software Diversity in Chip Multiprocessors,” *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, New York, NY, USA, 2007, pp. 186–197.

[108] A. Fourmigue, G. Beltrame, and G. Nicolescu, “Efficient transient thermal simulation of 3D ICs with liquid-cooling and through silicon vias,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2014, pp. 1–6.

[109] M. M. Sabry, A. Sridhar, and D. Atienza, “Thermal balancing of liquid-cooled 3D-MPSoCs using channel modulation,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2012, pp. 599–604.

[110] A. K. Coskun, D. Atienza, T. S. Rosing, T. Brunschwiler, and B. Michel, “Energy-efficient variable-flow liquid cooling in 3D stacked architectures,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2010, pp. 111–116.

[111] B. Raghunathan, Y. Turakhia, S. Garg, and D. Marculescu, “Cherry-picking: Exploiting process variations in dark-silicon homogeneous chip multi-processors,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2013, pp. 39–44.

[112] N. Kapadia and S. Pasricha, “VARSHA: Variation and reliability-aware application scheduling with adaptive parallelism in the dark-silicon era,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2015, pp. 1060–1065.

[113] B. Guha, B. B. C. Kyotoku, and M. Lipson, “CMOS-compatible athermal silicon microring resonators,” *Opt. Express*, vol. 18, no. 4, pp. 3487–3493, Feb. 2010.

[114] S. S. Djordjevic, K. Shang, B. Guan, S. T. S. Cheung, L. Liao, J. Basak, H.-F. Liu, and S. J. B. Yoo, “CMOS-compatible, athermal silicon ring modulators clad with titanium dioxide,” *Opt. Express*, vol. 21, no. 12, pp. 13958–13968, Jun. 2013.

[115] C. T. DeRose, M. R. Watts, D. C. Trotter, D. L. Luck, G. N. Nielson, and R. W. Young, "Silicon microring modulator with integrated heater and temperature sensor for thermal control," in *CLEO/QELS Laser Science to Photonic Applications*, 2010, pp. 1–2.

[116] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanović, "Addressing link-level design tradeoffs for integrated photonic interconnects," *IEEE Custom Integrated Circuits Conference (CICC)*, 2011, pp. 1–8.

[117] M. V. Beigi and G. Memik, "Therma: Thermal-aware Run-time Thread Migration for Nanophotonic Interconnects," *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2016, pp. 230–235.

[118] D. Dang, S. V. R. Chittamuru, R. Mahapatra, and S. Pasricha, "Islands of heaters: A novel thermal management framework for photonic NoCs," *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2017, pp. 306–311.

[119] M. Mohamed, Z. Li, X. Chen, L. Shang, and A. R. Mickelson, "Reliability-Aware Design Flow for Silicon Photonics On-Chip Interconnect," *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 22, no. 8, pp. 1763–1776, Aug. 2014.

[120] M. Mohamed, Z. Li, X. Chen, L. Shang, A. R. Mickelson, M. Vachharajani, and Y. Sun, "Power-efficient variation-aware photonic on-chip network management," *ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*, 2010, pp. 31–36.

[121] R. Wu, C.-H. Chen, C. Li, T.-C. Huang, F. Lan, C. Zhang, Y. Pan, J. E. Bowers, R. Beausoleil, and K.-T. Cheng, "Variation-Aware Adaptive Tuning for Nanophotonic Interconnects," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2015, pp. 487–493.

[122] P. P. Absil, P. Verheyen, P. D. Heyn, M. Pantouvaki, G. Lepage, J. D. Coster, and J. V. Campenhout, "Silicon photonics integrated circuits: a manufacturing platform for high density, low power optical I/O's," *Opt. Express*, vol. 23, no. 7, pp. 9369–9378, Apr. 2015.

[123] T. E. Carlson, W. Heirmant, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2011, pp. 1–12.

[124] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The SPLASH-2 programs: characterization and methodological considerations," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 1995, pp. 24–36.

[125] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 469–480.

[126] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2010, pp. 463–470.

[127] S. Feng, T. Lei, H. Chen, H. Cai, X. Luo, and A. W. Poon, "Silicon photonics: from a microresonator perspective," *Laser Photonics Rev.*, vol. 6, no. 2, pp. 145–177, Apr. 2012.

[128] D. J. Thomson, F. Y. Gardes, Y. Hu, G. Mashanoich, M. Fournier, P. Grosse, J-M. Fedeli, and G. T. Reed, "High contrast 40Gbit/s optical modulation in silicon," *Opt. Express*, vol. 19, no. 12, p. 11507, Jun. 2011.

[129] S. Liao, N-N. Feng, D. Feng, P. Dong, J. E. Cunningham, Y. Luo, and M. Asghari, "36 GHz submicron silicon waveguide germanium photodetector," *Opt. Express*, vol. 19, no. 11, p. 10967, May 2011.

[130] L. Vivien, L. Virot, D. Marris-Morini, J-M. Hartmann, C. Baudot, F. Boeuf, and J-M. Fedeli, "40Gbit/s germanium waveguide photodiode," *Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, 2013, pp. 1–3.

[131] Y. H. D. Lee and M. Lipson, "Back-End Deposited Silicon Photonics for Monolithic Integration on CMOS," *IEEE J. Sel. Top. Quantum Electron.*, vol. 19, no. 2, pp. 8200207–8200207, Mar. 2013.

[132] I. A. Young, E. Mohammed, J. T. S. Liao, A. M. Kern, S. Palermo, B. A. Block, P. L. Chang, "Optical I/O Technology for Tera-Scale Computing," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 235–248, Jan. 2010.

[133] J. M. Fedeli, R. Orobtchouk, C. Seassal, and L. Vivien, "Integration issues of a photonic layer on top of a CMOS circuit," 2006, p. 61250H–61250H–15.

[134] K. Preston, B. Schmidt, and M. Lipson, "Polysilicon photonic resonators for large-scale 3D integration of optical networks," *Opt. Express*, vol. 15, no. 25, p. 17283, 2007.

[135] N. Sherwood-Droz and M. Lipson, "Scalable 3D dense integration of photonics on bulk silicon," *Opt. Express*, vol. 19, no. 18, p. 17758, Aug. 2011.

[136] A. Gondarenko, J. S. Levy, and M. Lipson, "High confinement micron-scale silicon nitride high Q ring resonator," *Opt. Express*, vol. 17, no. 14, p. 11366, Jul. 2009.

[137] K. Preston, S. Manipatruni, A. Gondarenko, C. B. Poitras, and M. Lipson, "Deposited silicon high-speed integrated electro-optic modulator," *Opt. Express*, vol. 17, no. 7, p. 5118, Mar. 2009.

[138] N. Sherwood-Droz, K. Preston, J. S. Levy, and M. Lipson, "Device guidelines for wdm interconnects using silicon microring resonators," *Workshop on the Interaction between Nanophotonic Devices and Systems (WINDS)*, 2010, vol. 43, pp. 15–18.

[139] K. Preston, N. Sherwood-Droz, J. S. Levy, and M. Lipson, "Performance guidelines for WDM interconnects based on silicon microring resonators," *Conference on Lasers and Electro-Optics (CLEO)*, 2011, pp. 1–2.

[140] Z. Li, M. Mohamed, X. Chen, A. Mickelson, and L. Shang, "Device modeling and system simulation of nanophotonic on-chip networks for reliability, power and performance," *ACM/IEEE Design Automation Conference (DAC)*, 2011, pp. 735–740.

[141] M. Mohamed, Z. Li, X. Chen, A. Mickelson, and L. Shang, "Modeling and analysis of micro-ring based silicon photonic interconnect for embedded systems," *IEEE/ACM International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2011, pp. 227–236.

[142] A. N. Udipi, N. Muralimanohar, R. Balsubramonian, A. Davis, and N. P. Jouppi, "Combining memory and a controller with photonics through 3D-stacking to enable scalable and energy-efficient systems," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2011, pp. 425–436.

[143] S. Bahirat and S. Pasricha, "Exploring Hybrid Photonic Networks-on-chip Foremerging Chip Multiprocessors," *IEEE/ACM International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2009, pp. 129–136.

[144] S. Bahirat and S. Pasricha, “A Particle Swarm Optimization approach for synthesizing application-specific hybrid photonic networks-on-chip," *IEEE International Symposium on Quality Electronic Design (ISQED)*, 2012.

[145] A. B. Fallahkhair, K. S. Li, and T. E. Murphy, “Vector Finite Difference Modesolver for Anisotropic Dielectric Waveguides,” *J. Light. Technol.*, vol. 26, no. 11, pp. 1423–1431, Jun. 2008.

[146] D. T. Pierce, “Electronic Structure of Amorphous Si from Photoemission and Optical Studies,” *Phys Rev B*, vol. 5, Apr. 1972.

[147] C. D. Salzberg and J. J. Villa, “Infrared Refractive Indexes of Silicon Germanium and Modified Selenium Glass,” *JOSA*, vol. 47, no. 3, pp. 244–246, Mar. 1957.

[148] G. Li, X. Zheng, J. Yao, H. Thacker, J. E. Cunningham, and A. V, Krishnamoorthy, “25Gb/s 1V-driving CMOS ring modulator with integrated thermal tuning,” *Opt. Express*, vol. 19, no. 21, p. 20435, Oct. 2011.

[149] Q. Xu, D. Fattal, and R. G. Beausoleil, “Silicon microring resonators with 1.5-μm radius,” *Opt. Express*, vol. 16, no. 6, p. 4309, Mar. 2008.

[150] J. Levy, “Integrated nonlinear optics in silicon nitride waveguides and resonators,” 2011.

[151] J. D. Reis and A. L. Teixeira, “Architectural optimization of coherent ultra-dense WDM based optical access networks,” *Optical Fiber Communication Conference (OFCC)*, 2011, p. OTuB7.

[152] R. Wu, C.-H. Chen, J.-M. Fedeli, M. Fournier, R. G. Beausoleil, and K.-T. Cheng, “Compact modeling and system implications of microring modulators in nanophotonic interconnects,” *ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, 2015, pp. 1–6.

[153] A. Johnson, Y. Okawachi, J. S. Levy, J. Cardenas, K. Saha, M. Lipson, and A. L. Gaeta, "Chip-based frequency combs with sub-100 GHz repetition rates," *Optics Letters*, vol. 37, no. 5, pp. 875-877, 2012.

[154] D. J. Moss, R. Morandotti, A. L. Gaeta, and M. Lipson, "New CMOS-compatible platforms based on silicon nitride and Hydex for nonlinear optics," *Nat. Photonics*, vol. 7, no. 8, pp. 597–607, Aug. 2013.

[155] B. Neel, M. Kennedy, and A. Kodi, "Dynamic Power Reduction Techniques in On-Chip Photonic Interconnects," *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2015, pp. 249–252.

[156] C. Chen and A. Joshi, "Runtime Management of Laser Power in Silicon-Photonic Multibus NoC Architecture," *IEEE J. Sel. Top. Quantum Electron.*, vol. 19, no. 2, pp. 3700713–3700713, Mar. 2013.

[157] C. Chen, J. L. Abellan, and A. Joshi, "Managing Laser Power in Silicon-Photonic NoC Through Cache and NoC Reconfiguration," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 6, pp. 972–985, Jun. 2015.

[158] Y.-H. Kao and H. J. Chao, "Design of a Bufferless Photonic Clos Network-on-Chip Architecture," *IEEE Trans. Comput.*, vol. 63, no. 3, pp. 764–776, Mar. 2014.

[159] Z. Wang, H. Gu, Y. Yang, and B. Zhang, "Power Allocation Method for TDM-Based Optical Network on Chip," *IEEE Photonics Technol. Lett.*, vol. 25, no. 10, pp. 973–976, May 2013.

[160] A. Biberman, K. Preston, G. Hendry, N. Sherwood-Droz, J. Chan, J. S. Levy, M. Lipson, and K. Bergman, "Photonic Network-on-chip Architectures Using Multilayer Deposited Silicon Materials for High-performance Chip Multiprocessors," *J Emerg Technol Comput Syst*, vol. 7, no. 2, p. 7:1–7:25, Jul. 2011.

[161] *Semiconductor Optical Amplifiers*, 2002 edition. Boston; London: Springer, 2002.

[162] A. Biberman, H. L. R. Lira, K. Padmaraju, N. Ophir, J. Chan, M. Lipson, and K. Bergman, "Broadband Silicon Photonic Electrooptic Switch for Photonic Interconnection Networks," *IEEE Photonics Technol. Lett.*, vol. 23, no. 8, pp. 504–506, Apr. 2011.

[163] T. J. Kao and A. Louri, "Design of high bandwidth photonic NoC architectures using optical multilevel signaling," *IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, 2016, pp. 1–4.

[164] I. Thakkar and S. Pasricha, "3D-Wiz: A Novel High Bandwidth, Optically Interfaced 3D DRAM Architecture with Reduced Random-Access Time," *IEEE International Conference Computing Design (ICCD)*, 2014.

[165] Y. Xu and S. Pasricha, "Silicon Nanophotonics for Future Multicore Architectures: Opportunities and Challenges," *IEEE Des. Test*, vol. 31, no. 5, pp. 9–17, Oct. 2014.

[166] C.-H. Chen, M. A. Seyedi, M. Fiorentino, D. Livshits A. Gubenko, S. Mikhrin, V. Mikhrin, and R. G. Beausoleil, "A comb laser-driven DWDM silicon photonic transmitter based on microring modulators," *Opt. Express*, vol. 23, no. 16, pp. 21541–21548, Aug. 2015.

[167] T. J. Kao and A. Louri, "Optical Multilevel Signaling for High Bandwidth and Power-Efficient On-Chip Interconnects," *IEEE Photonics Technol. Lett.*, vol. 27, no. 19, pp. 2051–2054, Oct. 2015.

[168] R. Dubé-Demers, S. LaRochelle, and W. Shi, "Ultrafast pulse-amplitude modulation with a femtojoule silicon photonic modulator," *Optica*, vol. 3, no. 6, pp. 622–627, Jun. 2016.

[169] A. Roshan-Zamir, B. Wang, S. Telaprolu, K. Yu, C. Li, R. Beausoleil, and S. Palermo, "A 40 Gb/s PAM4 silicon microring resonator modulator transmitter in 65nm CMOS," *IEEE Optical Interconnects Conference (OI)*, 2016, pp. 8–9.

[170] S. Moazeni, S. Lin, M. Wade, L. Alloatti, R. J. Ram, M. Popovic, and V. Stojanovic, “A 40-Gb/s PAM-4 Transmitter Based on a Ring-Resonator Optical DAC in 45-nm SOI CMOS,” *IEEE J. Solid-State Circuits*, vol. 52, no. 12, pp. 3503–3516, Dec. 2017.

[171] K. Padmaraju, D. F. Logan, T. Shiraishi, J. J. Ackert, A. P. Knights, and K. Bergman, “Wavelength Locking and Thermally Stabilizing Microring Resonators Using Dithering Signals,” *J. Light. Technol.*, vol. 32, no. 3, pp. 505–512, Feb. 2014.

[172] K. Szczerba, P. Westbergh, J. Karout, J. S. Gustavsson, A. Haglund, M. Karlsson, E. Agrell, A. Larsson, “4-PAM for high-speed short-range optical communications,” *IEEE J. Opt. Commun. Netw.*, vol. 4, no. 11, pp. 885–894, Nov. 2012.

[173] T. Y. Elganimi, "Performance Comparison between OOK, PPM and PAM Modulation Schemes for Free Space Optical (FSO) Communication Systems: Analytical Study," *Intl. J. Comp. App.*, vol. 79, no. 11, pp. 22-27, Oct 2013.

[174] M. Bahadori, S. Rumley, H. Jayatilleka, K. Murray, L. Chrostowski, S. Shekhar, and K. Bergman, “Crosstalk Penalty in Microring-Based Silicon Photonic Interconnect Systems,” *J. Light. Technol.*, vol. 34, no. 17, pp. 4043–4052, Sep. 2016.

[175] M. Nikdast, G. Nicolescu, J. Trajkovic, and O. Liboiron-Ladouceur, “Modeling fabrication non-uniformity in chip-scale silicon photonic interconnects,” *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2016, pp. 115–120.

[176] M. A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, “A comprehensive model for PMOS NBTI degradation: Recent progress,” *Microelectron. Reliab.*, vol. 47, no. 6, pp. 853–862, Jun. 2007.

[177] H. Kufluoglu, “MOSFET degradation due to negative bias temperature instability (NBTI) and hot carrier injection (HCI), and its implications for reliability-aware VLSI design,” *Theses Diss. Available ProQuest*, pp. 1–200, Jan. 2007.

[178] H. Kufluoglu and M. A. Alam, “Theory of interface-trap-induced NBTI degradation for reduced cross section MOSFETs,” *IEEE Trans. Electron Devices*, vol. 53, no. 5, pp. 1120–1130, May 2006.

[179] S. Ogawa and N. Shiono, “Generalized diffusion-reaction model for the low-field charge-buildup instability at the Si-SiO2 interface,” *Phys. Rev. B Condens. Matter*, vol. 51, no. 7, pp. 4218–4230, Feb. 1995.

[180] M. Lipson, “Guiding, modulating, and emitting light on Silicon-challenges and opportunities,” *J. Light. Technol.*, vol. 23, no. 12, pp. 4222–4238, Dec. 2005.

[181] M. Cho, C. Kersey, M. P. Gupta, N. Sathe, S. Kumar, S. Yalamanchili, and S. Mukhopadhyay, “Power Multiplexing for Thermal Field Management in Many-Core Processors,” *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 3, no. 1, pp. 94–104, Jan. 2013.

[182] R. S. Chakraborty, S. Narasimhan, and S. Bhunia, “Hardware Trojan: Threats and emerging solutions,” *IEEE International High Level Design Validation and Test Workshop*, 2009, pp. 166–171.

[183] M. Tehranipoor and F. Koushanfar, “A Survey of Hardware Trojan Taxonomy and Detection,” *IEEE Des. Test Comput.*, vol. 27, no. 1, pp. 10–25, Jan. 2010.

[184] S. Skorobogatov and C. Woods, “Breakthrough Silicon Scanning Discovers Backdoor in Military Chip,” *International Conference on Cryptographic Hardware and Embedded Systems*, 2012, pp. 23–40.

[185] W. J. Dally and B. Towles, “Route packets, not wires: on-chip interconnection networks,” *IEEE/ACM Design Automation Conference (DAC)*, 2001, pp. 684–689.

[186] D. M. Ancajas, K. Chakraborty, and S. Roy, “Fort-NoCs: Mitigating the threat of a compromised NoC,” *IEEE/ACM Design Automation Conference (DAC)*, 2014, pp. 1–6.

[187] C. Li, M. Browning, P. V. Gratz, and S. Palermo, “Energy-efficient optical broadcast for nanophotonic networks-on-chip,” *IEEE Optical Interconnects Conference (OIC)*, 2012, pp. 64–65.

[188] C. H. Gebotys and R. J. Gebotys, “A framework for security on NoC technologies,” *IEEE Computer Society Annual Symposium on VLSI*, 2003, pp. 113–117.

[189] H. K. Kapoor, G. B. Rao, S. Arshi, and G. Trivedi, “A Security Framework for NoC Using Authenticated Encryption and Session Keys,” *Circuits Syst. Signal Process.*, vol. 32, no. 6, pp. 2605–2622, Dec. 2013.

[190] T. Agerwala, “Exascale computing: The challenges and opportunities in the next decade,” *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2010, pp. 1–1.

[191] T. Pimpalkhute and S. Pasricha, “NoC Scheduling for Improved Application-Aware and Memory-Aware Transfers in Multi-core Systems,” *IEEE International Conference on VLSI Design and International Conference on Embedded Systems (VLSID)*, 2014, pp. 234–239.

[192] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, “A case for intelligent RAM,” *IEEE International Symposium on Microarchitecture (MICRO)*, vol. 17, no. 2, pp. 34–44, Mar. 1997.

[193] Y. H. Son, O. Seongil, Y. Ro, J. W. Lee, and J. H. Ahn, "Reducing Memory Access Latency with Asymmetric DRAM Bank Organizations," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2013, pp. 380–391.

[194] S. Beamer, C. Sun, Y-J. Kwon, A. Joshi, C. Batten, V. Stojanovic, and K. Asanovic, "Re-architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2010, pp. 129–140.

[195] J. Mukundan, H. Hunter, K. Kim, J. Stuecheli, and J. F. Martínez, "Understanding and Mitigating Refresh Overheads in High-density DDR4 DRAM Systems," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2013, pp. 48–59.

[196] G. H. Loh, "3D-Stacked Memory Architectures for Multi-core Processors," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2008, pp. 453–464.

[197] D. H. Woo, N. H. Seong, D. L. Lewis, and H.-H. S. Lee, "An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth," *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2010, pp. 1–12.

[198] U. Kang, H-J Chung, S. Heo, D-H. Park, D. Kwon, J-W. Lee, H-S. Joo, W-S. Kim, J. S. Choi, C. Kim, and Y-H. Jun, "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 111–119, Jan. 2010.

[199] I. Thakkar and S. Pasricha, "3D-WiRED: A Novel Wide I/O DRAM with Energy-Efficient 3D Bank Organization," *IEEE Des. Test*, vol. PP, no. 99, pp. 1–1, 2015.

[200] T. Zhang, C. Xu, K. Chen, G. Sun, and Y. Xie, "3D-SWIFT: A High-performance 3D-stacked Wide IO DRAM," *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2014, pp. 51–56.

[201] B. Giridhar, M. Cieslak, D. Duggal, R. Dreslinski, H-M. Chen, R. Patti, B. Hold, C. Chakrabarti, T. Mudge, and D. Blaauw, "Exploring DRAM organizations for energy-efficient and resilient exascale memories," *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2013, pp. 1–12.

[202] "JEDEC STANDARD (JESD79-3E): DDR3 SDRAM." JEDEC SOLID STATE TECHNOLOGY ASSOCIATION, Jul-2010.

[203] "JEDEC STANDARD (JESD209-3): Low Power Double Data Rate 3." JEDEC SOLID STATE TECHNOLOGY ASSOCIATION, May-2012.

[204] H. Y. To and J. A. McCall, "Differential memory interface system," US6747483 B2, 08-Jun-2004.

[205] K. Chen, S. Li, N. Muralimanohar, J.-H. Ahn, J. B. Brockman, and N. P. Jouppi, "CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory," *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2012, pp. 33–38.

[206] D. Xu, N. Yu, P. D. S. Manoj, K. Wang, H. Yu, and M. Yu, "A 2.5-D Memory-Logic Integration With Data-Pattern-Aware Memory Controller," *IEEE Des. Test*, vol. 32, no. 4, pp. 1–10, Aug. 2015.

[207] S. M. P. D, H. Yu, H. Huang, and D. Xu, "A Q-Learning Based Self-Adaptive I/O Communication for 2.5D Integrated Many-Core Microprocessor and Memory," *IEEE Trans. Comput.*, vol. 65, no. 4, pp. 1185–1196, Apr. 2016.

[208] B. Jacob, S. Ng, and D. Wang, *Memory Systems: Cache, DRAM, Disk*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007.

[209] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. Addison Wesley, 2011.

[210] R. Ho, K. W. Mai, and M. A. Horowitz, “The future of wires,” *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.

[211] N. P. Jouppi, A. B. Kahng, N. Muralimanohar, and V. Srinivas, “CACTI-IO: CACTI With OFF-chip Power-Area-Timing Models,” *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. PP, no. 99, pp. 1–1, 2014.

[212] P. Rosenfeld, E. Cooper-Balis, and B. Jacob, “DRAMSim2: A Cycle Accurate Memory System Simulator,” *Comput. Archit. Lett.*, vol. 10, no. 1, pp. 16–19, Jan. 2011.

[213] M. Shevgoor, J. S. Kim, N. Chatterjee, R. Balasubramonian, A. Davis, and A. N. Udipi, “Quantifying the relationship between the power delivery network and architectural policies in a 3D-stacked memory device,” *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2013, pp. 198–209.

[214] A. Hadke, T. Benavides, S. J. B. Yoo, R. Amirtharajah, and V. Akella, “OCDIMM: Scaling the DRAM Memory Wall Using WDM Based Optical Interconnects,” *IEEE Symposium on High Performance Interconnects (HOTI)*, 2008, pp. 57–63.

[215] C. Zhong, “25Gbps SerDes,” *IEEE Higher Speed Study Group (HSSG)*, Mar-2007.

[216] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, “GPU Computing,” *Proc. IEEE*, vol. 96, no. 5, pp. 879–899, May 2008.

[217] R. Kho, D. Boursin, M. Brox, P. Gregorius, H. Hoenigschmid, B. Kho, S. Kieser, D. Kehrer, M. Kuzmenka, M. Gjukic, W. Spirkl, H. Steffens, J. Weller, and T. Hein, “A 75 nm 7 Gb/s/pin 1 Gb GDDR5 Graphics Memory Device With Bandwidth Improvement Techniques,” *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 120–133, Jan. 2010.

[218] C. Weis, I. Loi, L. Benini, and N. Wehn, "Exploration and Optimization of 3-D Integrated DRAM Subsystems," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 32, no. 4, pp. 597–610, Apr. 2013.

[219] "1Gb (32Mx32) GDDR5 SGRAM H5GQ1H24AFR." Datasheet by Hynix, Nov-2009.

[220] H.-W. Lee, J. Song, S-A. Hyun, S. Baek, Y. Lim, J. Lee, M. Park, H. Choi, C. Choi, J. Cha, J. Kim, H. Choi, S. Kwack, Y. Kang, J. Kim, J. Park, J. Kim, J. Cho, C. Kim, Y. Kim, J. Lee, B. Chung, and S. Hong, "25.3 A 1.35V 5.0Gb/s/pin GDDR5M with 5.4mW standby power and an error-adaptive duty-cycle corrector," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2014, pp. 434–435.

[221] A. Bakhoda, G. L. Yuan, W. W. L. Fung, H. Wong, and T. M. Aamodt, "Analyzing CUDA workloads using a detailed GPU simulator," *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2009, pp. 163–174.

[222] T. Aamodt, W. Fung, and T. Hetherington, "GPGPU-Sim 3.x ,A Performance Simulator for Many-Core Accelerator Research," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2012.

[223] Y. Han, Y. Wang, H. Li, and X. Li, "Data-aware DRAM refresh to squeeze the margin of retention time in hybrid memory cube," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2014, pp. 295–300.

[224] M. Ghosh and H.-H. S. Lee, "Smart Refresh: An Enhanced Memory Controller Design for Reducing Energy in Conventional and 3D Die-Stacked DRAMs," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2007, pp. 134–145.

[225] T. V. Kalyan, K. Ravi, and M. Mutyam, “Scattered refresh: An alternative refresh mechanism to reduce refresh cycle time,” *IEEE/ACM Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014, pp. 598–603.

[226] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, “RAIDR: Retention-aware intelligent DRAM refresh,” *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2012, pp. 1–12.

[227] R. K. Venkatesan, S. Herr, and E. Rotenberg, “Retention-aware placement in DRAM (RAPID): software methods for quasi-non-volatile DRAM,” *IEEE/ACM International Symposium on High-Performance Computer Architecture (HPCA)*, 2006, pp. 155–165.

[228] S. Liu, K. Pattabiraman, T. Moscibroda, and B. G. Zorn, “Flikker: Saving DRAM Refresh-power Through Critical Data Partitioning,” *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2011, pp. 213–224.

[229] J. Stuecheli, D. Kaseridis, H. C. Hunter, and L. K. John, “Elastic Refresh: Techniques to Mitigate Refresh Penalties in High Density Memory,” *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2010, pp. 375–384.

[230] P. Nair, C.-C. Chou, and M. K. Qureshi, “A case for Refresh Pausing in DRAM memory systems,” *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2013, pp. 627–638.

[231] I. S. Bhati, “Scalable and energy efficient DRAM refresh techniques,” Ph.D. Thesis, University of Maryland, College Park, United States -- Maryland, 2014.

[232] J. D. Leidel and Y. Chen, “HMC-Sim: A Simulation Framework for Hybrid Memory Cube Devices,” *IEEE International Parallel Distributed Processing Symposium Workshops (IPDPSW)*, 2014, pp. 1465–1474.

[233] M. T. Chang, P. Rosenfeld, S. L. Lu, and B. Jacob, “Technology comparison for large last-level caches (L3Cs): Low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM,” *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2013, pp. 143–154.

[234] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, “Architecting Phase Change Memory As a Scalable Dram Alternative,” *IEEE International Symposium on Computer Architecture (ISCA)*, 2009, pp. 2–13.

[235] M. K. Qureshi, M. M. Franceschini, and L. A. Lastras-Montaño, “Improving read performance of Phase Change Memories via Write Cancellation and Write Pausing,” *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2010, pp. 1–11.

[236] M. K. Qureshi, M. M. Franceschini, A. Jagmohan, and L. A. Lastras, “PreSET: Improving performance of phase change memories by exploiting asymmetry in write times,” *IEEE International Symposium on Computer Architecture (ISCA)*, 2012, pp. 380–391.

[237] Y. Kim, S. Yoo, and S. Lee, “Write performance improvement by hiding R drift latency in phase-change RAM,” *ACM/IEEE Design Automation Conference (DAC)*, 2012, pp. 897–906.

[238] S. Kwon, S. Yoo, S. Lee, and J. Park, “Optimizing Video Application Design for Phase-Change RAM-Based Main Memory,” *IEEE Trans. Very Large Scale Integr.*, vol. 20, no. 11, pp. 2011–2019, Nov. 2012.

[239] L. Jiang, B. Zhao, Y. Zhang, J. Yang, and B. R. Childers, "Improving write operations in MLC phase change memory," *IEEE International Symposium on High-Performance Comp Architecture (HPCA)*, 2012, pp. 1–10.

[240] J. Yue and Y. Zhu, "Making Write Less Blocking for Read Accesses in Phase Change Memory," *IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, 2012, pp. 269–277.

[241] C. Pan, M. Xie, J. Hu, Y. Chen, and C. Yang, "3M-PCM: Exploiting multiple write modes MLC phase change main memory in embedded systems," *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2014, pp. 1–10.

[242] L. Jiang, Y. Zhang, B. R. Childers, and J. Yang, "FPB: Fine-grained Power Budgeting to Improve Write Throughput of Multi-level Cell Phase Change Memory," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2012, pp. 1–12.

[243] L. Jiang, B. Zhao, J. Yang, and Y. Zhang, "A low power and reliable charge pump design for Phase Change Memories," *ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, 2014, pp. 397–408.

[244] S. Cho and H. Lee, "Flip-N-Write: A simple deterministic technique to improve PRAM write performance, energy and endurance," *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 347–357.

[245] J. Yue and Y. Zhu, "Accelerating write by exploiting PCM asymmetries," *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2013, pp. 282–293.

[246] J. Li and K. Mohanram, "Write-once-memory-code phase change memory," *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2014, pp. 1–6.

[247] B. Li, S. Shan, Y. Hu, and X. Li, "Partial-SET: Write speedup of PCM main memory," *IEEE/ACM Design, Automation Test in Europe Conference (DATE)*, 2014, pp. 1–4.

[248] H. Horii, J. H. Yi, J. H. Park, Y. H. Ha, I. G. Baek. S. O. Park, Y. N. Hwang, S. H. Lee, Y. T. Kim, K. H. Lee, U-I. Chung, and J. T. Moon, "A novel cell technology using N-doped GeSbTe films for phase change RAM," *IEEE Symposium on VLSI Technology (VLSIT)*, 2003, pp. 177–178.

[249] H. G. Lee, S. Baek, C. Nicopoulos, and J. Kim, "An energy- and performance-aware DRAM cache architecture for hybrid DRAM/PCM main memory systems," *IEEE International Conference on Computer Design (ICCD)*, 2011, pp. 381–387.

[250] H. A. Khouzani, F. S. Hosseini, and C. Yang, "Segment and Conflict Aware Page Allocation and Migration in DRAM-PCM Hybrid Main Memory," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 36, no. 9, pp. 1458–1470, Sep. 2017.

[251] S. Lee, H. Bahn, and S. H. Noh, "CLOCK-DWF: A Write-History-Aware Page Replacement Algorithm for Hybrid PCM and DRAM Memory Architectures," *IEEE Trans. Comput.*, vol. 63, no. 9, pp. 2187–2200, Sep. 2014.

[252] D. Zhang, L. Ju, M. Zhao, X. Gao, and Z. Jia, "Write-back aware shared last-level cache management for hybrid main memory," *ACM/IEEE Design Automation Conference (DAC)*, 2016, pp. 1–6.

[253] G. W. Burr, A. Padilla, M. Franceschini, B. Jackson, D. G. Dupouy, C. T. Rettner, K. Gopalakrishnan, R. Shenoy, and Karidis†, "The inner workings of phase change memory: Lessons from prototype PCM devices," *IEEE GLOBECOM Workshop*, 2010, pp. 1890–1894.

[254] W. Zhang and T. Li, “Helmet: A resistance drift resilient architecture for multi-level cell phase change memory system,” *IEEE/IFIP 41st International Conference on Dependable Systems Networks (DSN)*, 2011, pp. 197–208.

[255] M. Awasthi, M. Shevgoor, K. Sudan, B. Rajendran, R. Balasubramonian, and V. Srinivasan, “Efficient scrub mechanisms for error-prone emerging memories," *IEEE International Symposium on High-Performance Comp Architecture (HPCA)*, 2012, pp. 1–12.

[256] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, “NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.

[257] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, and R. Bez, “Scaling analysis of phase-change memory technology,” *IEEE International Electron Devices Meeting (IEDM)*, 2003, p. 29.6.1-29.6.4.

[258] A. Hay, K. Strauss, T. Sherwood, G. H. Loh, and D. Burger, “Preventing PCM Banks from Seizing Too Much Power,” *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2011, pp. 186–195.

[259] J. Yun, S. Lee, and S. Yoo, “Dynamic Wear Leveling for Phase-Change Memories With Endurance Variations,” *IEEE Trans. Very Large Scale Integr.*, vol. 23, no. 9, pp. 1604–1615, Sep. 2015.

[260] M. K. Qureshi, J. Karidis, M. Franceschini, V. Srinivasan, L. Lastras, and B. Abali, “Enhancing lifetime and security of PCM-based Main Memory with Start-Gap Wear Leveling,” *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 14–23.

[261] P. M. Palangappa, J. Li, and K. Mohanram, “WOM-Code Solutions for Low Latency and High Endurance in Phase Change Memory,” *IEEE Trans. Comput.*, vol. 65, no. 4, pp. 1025–1040, 2016.

[262] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, “Scalable High Performance Main Memory System Using Phase-change Memory Technology,” *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2009, pp. 24–33.

[263] W. Zhang and T. Li, “Characterizing and Mitigating the Impact of Process Variations on Phase Change Based Memory Systems,” *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 2–13.

[264] “Intel’s ‘Skylake’ Core i7-6700K: A Performance Look,” *Techgate*, Aug-2015. .

[265] M. Poremba, T. Zhang, and Y. Xie, “NVMain 2.0: A User-Friendly Memory Simulator to Model (Non-)Volatile Memory Systems,” *IEEE Comput. Archit. Lett.*, vol. 14, no. 2, pp. 140–143, Jul. 2015.

[266] “PARSEC v2.1 for M5.” [Online]. Available: http://www.cs.utexas.edu/~cart/parsec_m5/. [Accessed: 23-May-2018].

[267] “DDR3 SDRAM MT41J256M4.” Datasheet by Micron.