METHODOLOGIES FOR MODELING AND OPTIMIZATION OF 2.5-D AND 3-D INTEGRATION ARCHITECTURES FOR COMPUTE-IN-MEMORY APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

By

Ankit Kaul

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Engineering
Department of Electrical and Computer Engineering

Georgia Institute of Technology

December 2023

METHODOLOGIES FOR MODELING AND OPTIMIZATION OF 2.5-D AND 3-D INTEGRATION ARCHITECTURES FOR COMPUTE-IN-MEMORY APPLICATIONS

Thesis committee:

Dr. Muhannad Bakir
Department of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Suman Datta
Department of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
Department of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Vanessa Smet
Department of Mechanical Engineering
Georgia Institute of Technology

Dr. Azad J. Naeemi
Department of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hemanth Dhavaleswarapu
Advanced Packaging
Advanced Micro Devices Inc.

Date approved: August 18, 2023

How beautiful is the world because it gives us time and opportunity to help others.

Anonymous

For my parents Shri. Virender and Smt. Usha Kaul, my mother Late Smt. Anjali Kaul, my grandmother Late Smt. Kaushalya Tickoo, my grandfather Shri. Rattan Lal Tickoo, and my sister Divya for their limitless love, strong support, and countless sacrifices.

ACKNOWLEDGMENTS

Versions of "go ahead with what you want to do" have repeatedly been said to me by the people who motivate me in life.

I would like to express my sincere gratitude to my advisor Dr. Muhannad S. Bakir for his support and guidance through my Ph.D journey. His professional advice and personal support during challenging times have been invaluable for me. He has been instrumental in my learnings, both technical and personal. He always encourages high quality research and emphasizes the importance of collaboration and teamwork. I have deep gratitude toward Dr. Bakir for encouraging and supporting many experiences for my learning while also being a supportive and caring figure during trying times.

I would like to thank Dr. Shimeng Yu and Dr. Arijit Raychowdhury for serving on my reading committee. Feedback form Dr. Yu and Dr. Raychowdhury on the preliminary and future work in my Ph.D. proposal was invaluable towards the directions we explored in this thesis. Dr. Yu's suggestions and advice on 3D RRAM thermal modeling were formative in my understanding and invaluable for the analyses we explored together. Dr. Raychowdhury's inputs and advice on our collaboration on multilevel signaling for chiplet-to-chiplet communication in 2.5D integration were crucial and extremely helpful for our collaboration. I am grateful and honored to Dr. Azaad Naeemi, Dr. Vanessa Smet, and Dr. Hemanth Dhavaleswarapu for serving on my doctoral thesis committee.

I want to thank the support and guidance of all of my past and current fellow members of the Integrated 3-D Systems (I3DS) group, who I owe most of my knowledge and who have made a positive impact through the fun discussions and constructive criticism: Dr. Thomas E. Sarvey, Dr. Sreejith Kochupurackal Rajan, Madison Manley, Dr. Md Obaidul Hossen, Dr. Joe L. Gonzales, Dr. William Wahby, Dr. Paul K. Jo, Dr. Yang Zhang, Dr. Congshan Wan, Dr. Reza Abbaspour, Dr. Muneeb Zia, Dr. Ting Zheng, Shengtao Yu, Jiaao Lu, Youngtak Lee, Carl Li, Shane Oh, Michael A. Nieves Calderon, Ashita Victor,

Philip M. Anschutz, Geyu Yan, Wanshu Zeng, Srujan Penta, Euichul Chung, Rohan Sahay, Zhonghao Zhang, Erik W. Masselink, and Ziyu Liu. I am particularly thankful to Dr. Sreejith Kochupurackal Rajan for his attention to detail and emphasis on pushing boundaries in research, any ounce of which if I possess, I picked up from him. I am also thankful to Sreejith for sticking through difficult times while being my best friend and my best critic. A big thank you to Madison Manley for her contributions to the device-integration chapters, open-sourcing our flows, and for all the fun conversations at work and during travel. I thank Dr. Xiaochen Peng for her contribution and her thoughts on the 3D RRAM thermal modeling flow. I would also like to thank Dr. Yandong Luo and James Read for their invaluable support on developing the device-integration PDN and thermal methodologies.

I would like to thank Jim Dodrill from Arm Inc. for providing me with an opportunity to work on an exciting summer internship and for inspiring me. I enjoyed my discussions with Jim, Saurabh Sinha, Rossana Liu, Rahul Mathur, Ashley Crawford, and other folks at Arm from whom I learnt immensely.

I am also grateful for the exciting internship opportunity with the Adanced Packaging group at AMD. I am fortunate to have worked with Dr. Hemanth Dhavaleswarapu, Dr. Rahul Agarwal, and Chandra Mandalapu. I am thankful to them for advising me, and for the valuable professional and personal feedback. I enjoyed my discussions with Dr. Dhavaleswarapu, Dr. Agarwal, Chandra, Dr. Thomas Burd, Srividhya Venkataraman, Arun Kumar Karunanithi, Dr. Brett Wilkerson, Dr. Raja Swaminathan, Chintan Buch, Dr. Arsalan Alam, and Ivor Barber and thank them for the fun conversations and advice.

I am grateful for The J.N. Tata Endowment for supporting my graduate school studies when I started my masters program at Georgia Tech. Without their support, my graduate school path would have been much harder.

I am grateful to have the limitless love and support of my parents, my grandparents, my sister, and other members of my family. My beloved father Shri Virender Kaul, beloved mother Smt. Usha Kaul, beloved grandfather Shri. Rattan Lal Tiku, beloved grandmother

Smt. Kaushalya Tiku, beloved mother late Smt. Anjli Kaul, and beloved sister Divya Kaul have been my beacons of strength and joy through this journey. I am fortunate to have such a loving family and I hope that I can live up to their hopes and dreams. This thesis was possible because of their belief in me more than that of myself. I am also fortunate to have kind and thoughtful friends as my support system. I am grateful to Satya Swaroop Panda for inspiring me to apply for a Ph.D. when I was not motivated enough to pursue one. I am grateful to Chetan Chandra, Dr. Sreejith K. Rajan, Ananth Noorithaya, Dr. Steven Schwartz, Dr. Mohit Gupta and Dr. Aarohi Shah for the pep talks and for humoring me when I was at my worst. I also can never forget the support of my friends, family, and Dr. Bakir during the time I met with a small accident.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

2-D two dimensional

2.5-D 2-D enhanced

3-D three dimensional

3-D-HI three dimensional heterogeneous integration

ADCs analog-to-digital converters

AMD Advanced Micro Devices, Inc.

AP advanced packaging

ASIC application specific integrated circuit

BEOL back end of line

BS-PDN backside power delivery network

CIM compute-in-memory

CMOS complementary metal oxide semiconductor

CPUs central processing units

D2W die to wafer

DBHi direct bonded heterogeneous integration

DNN deep neural network

DOCI dense off chip integration

DRAM dynamic random access memory

DTCs deep-trench capacitors

ECRAM electrochemical random access memory

EFB elevated fanout bridge

EMIB embedded multi-die interconnect bridge

eNVM emerging nonvolatile memory

EPB energy-per-bit

F2F face-to-face

FeFET ferroelectric field effect transistor

FPGAs field programmable gated arrays

GAA gate all around

GPGPUs general purpose graphics processing units

GPUs graphics processing units

HI heterogeneous integration

HIST heterogeneous integration stitching technology

HPC high performance computing

ICs integrated circuits

IoT internet of things

IP intellectual property

IRDS International Roadmap for Devices and Systems

MAC multiply-and-accumulate

MCM multi-chip module

MIM metal-insulator-metal

NBTI negative-bias temperature instability

PCM phase change memory

PDN power delivery network

PPAC power, performance, area, and cost

PSN power supply noise

RRAM resistive random access memory

SAR-ADC successive-approximation-register analog-to-digital converter

SoC system-on-a-chip

SOT-MRAM spin-orbit-torque magnetic random access memory

SRAM static random access memory

STT-MRAM spin-transfer-torque magnetic random access memory

TDP thermal design power

TPUs tensor processing units

TSMC Taiwan Semiconductor Manufacturing Company Limited

TSV through-silicon via

VMM vector-matrix multiplication

W2W wafer to wafer

SUMMARY

In the wake of data-intensive computing, von Neumann-based traditional architectures and conventional methods of integration such as monolithic 2-D, are facing multiple challenges with reduced performance and higher costs. Such systems suffer from higher latency of communication between the memory hierarchy and processing elements, high power consumption, and increased hardware cost. To support the rising demand for emerging applications and data-intensive workloads, such systems require higher memory bandwidth to reduce latency and more efficient devices to improve energy-per-operation. Additionally, non-linear trends in device densities and energy per operation, and the rising design costs of advanced technology nodes have made conventional feature scaling an expensive pursuit.

Due to the challenges with traditional von Neumann-based devices, new paradigms for compute, memory, communication, and integration have emerged. compute-in-memory (CIM) has been proposed as a potential paradigm for energy-efficient compute by reduced data movement and increased parallelism in image recognition and language translation computations. Further, a growing need for higher logic-memory bandwidth and lower chip-to-chip signal interconnection delay have led to a technological push towards heterogeneous integration. In this work, we propose methodologies to model physical effects and optimize design parameters in heterogeneous integration (HI) architectures for compute-in-memory hardware.

First, the design trade-offs of including a power delivery network (PDN) and metal-insulator-metal (MIM) capacitors in bridge-chip based 2.5-D heterogeneous platforms are investigated. It is demonstrated that including the PDN (and MIM capacitors) in the bridge-chip can be an effective technique to reduce both DC-IR-drop and *Ldi/dt* noise. Next, to address the power delivery challenges in three dimensional heterogeneous integration (3-D-HI), a systematic technology and design space exploration of power delivery for 3-D-HI

CIM systems is presented. A fast analysis flow facilitating early design-space exploration between power delivery design parameters and CIM performance metrics is proposed. By co-optimizing 3-D PDN and successive-approximation-register analog-to-digital converter (SAR-ADC) design parameters a balanced 3-D CIM design is demonstrated compared to a 3-D unoptimized implementation at iso-power and iso-area.

Next, the thermal impact of different 3-D-HI architectures on the reliability of 3-D-integrated binary resistive random access memory (RRAM) devices for CIM applications is quantified. A device-integration reliability evaluation methodology is proposed that can be used to quantify the direct impact of integration design parameters on CIM inference accuracy. Using this flow, heterogeneous 3-D logic-memory CIM accelerator designs are benchmarked against monolithic 2-D and balanced integration design parameters for maximized 3-D CIM inference accuracy are reported. The benchmark framework is released as an open-source tool for the research community.

A 3-D polylithic architecture is proposed that represents a densely integrated system divided into multiple device tiers where custom chiplets, such as power management IP, I/O drivers, and memory are embedded into the back-end of a base tier with extreme efficient signaling and large bandwidth density. Design optimization strategies for PDN in polylithic 3-D integration are presented. The scope includes a detailed design space exploration of the power supply noise effects in polylithic 3-D architectures. The thermal constraints for polylithic 3-D are evaluated with aggressive cooling to investigate thermal limits from transient- and steady-state perspectives.

CHAPTER 1
INTRODUCTION AND BACKGROUND

Semiconductor-based electronics have been supporting the world's ever rising computational demands for many years. Improvements in power, performance, area, and cost (PPAC) of such semiconductor-based hardware have been conventionally addressed by architectural innovation, process technology gains, larger die sizes, and higher power consumption. These PPAC improvement targets are evermore crucial with inexorable growth of demand for computing technologies (traditional high performance computing (HPC), compute intensive visualization) [1] and with the advent of data-intensive computing (emerging analytics and machine learning) [2]. Different applications have varied computational requirements that are summarised in this section along with the limitations of traditional compute systems and some emerging computational paradigms.

1.1 The Heterogeneous Compute Landscape

Electronics are used in many applications including: 1) high-performance, 2) high-efficiency/low-power computing, and 3) autonomous sensing and computing [3]. High-performance computing applications require more performance at constant power density and are usually constrained by thermal management. Low-power computing, commonly required in mobile applications, demand more performance and functionality at constant energy. These are typically constrained by a limited power source (battery) and total cost. Autonomous sensing and computing, under the wider umbrella of the internet of things (IoT), targets reduced leakage and variability.

Different applications can have disparate computational requirements. While some general purpose applications can benefit from single-threaded performance, other data-intensive applications are amenable to specialization and parallelization. For example,

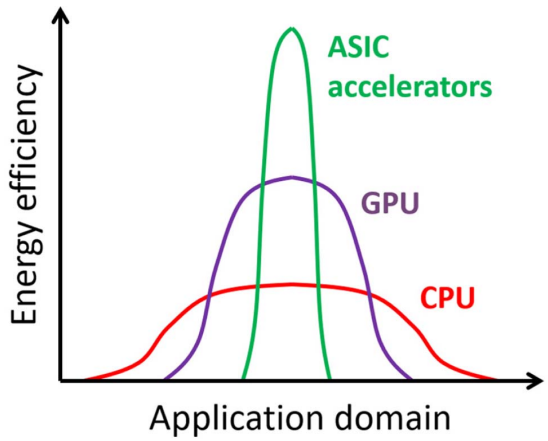


Figure 1.1: Disparate computational requirements of different applications. This leads to different energy efficiencies (performance/Watt) from general purpose to domain specific architectures (DSA). [2]

most graph algorithms can be realized by some version of matrix-vector multiplication [4] and neural networks are suited for hardware parallelization since their fundamental computation is based on multiply-and-accumulate (MAC) operations [5]. Similarly, hardware acceleration based on general purpose graphics processing units (GPGPUs) [6], field programmable gated arrays (FPGAs) [7], and custom application specific integrated circuit (ASIC) chips [8] are favourable for data intensive applications. This leads to a general trend of difference in energy efficiency for different types of compute, which increases as the hardware becomes more specialized, as shown in Figure 1.1 [1].

Some examples of high-performance compute hardware include: server central processing units (CPUs) (Advanced Micro Devices, Inc. (AMD) EPYC [9], Intel Xeon [10]), graphics processing units (GPUs) (NVIDIA RTX 4060 [11]), and GPGPUs (NVIDIA Tesla [6], AMD Instinct [12]).

High-efficiency hardware represent both efficient architectures (e.g. systolic arrays for matrix-vector multiplication in TPUs) and novel devices that help enable new architectures. Some examples of high-efficiency/low-power computing include mobile system-on-a-chip (SoC) (Apple A14, Qualcomm Snapdragon), ASIC accelerators (tensor processing units (TPUs)), and processing-in-memory (Samsung PIM).

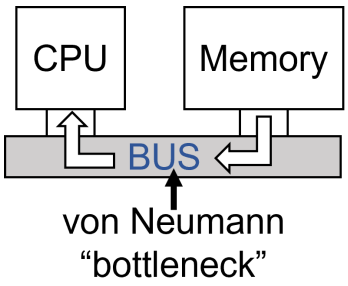


Figure 1.2: Communication bottleneck between compute and memory blocks in Von-Neumann-based architectures.

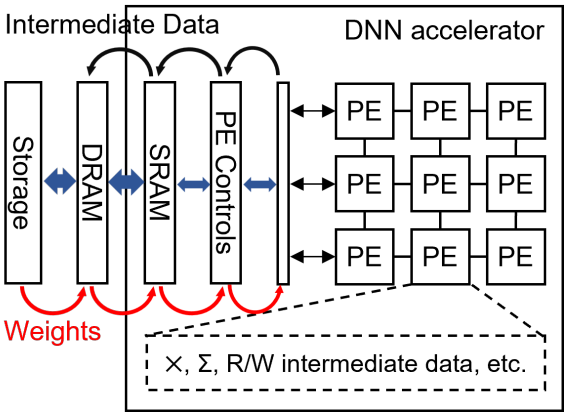


Figure 1.3: A DNN accelerator implemented using a Von-Neumann-based architecture.

1.1.1 Limitations of traditional compute systems

Traditional compute systems can be generally represented as a constitution of a compute block and a memory block interconnected with a bus for transfer of data and control logic (Figure 1.2). As an example of a system implemented using von Neumann-based arhitecture, a deep neural network (DNN) accelerator is shown in Figure 1.3. Such traditional architectures have been widely used across high-performance and high-efficiency compute hardware for decades. However, in the wake of data-intensive computing such architectures suffer from higher latency of communication between the memory hierarchy and processing elements, high power consumption, and high hardware cost. To support the rising demand for emerging applications and larger models [13], such systems require higher memory bandwidth to reduce latency and more efficient devices to improve energy-per-

operation.

1.1.2 Emerging computation paradigms

Due to the challenges with traditional von Neumann-based devices, new paradigms for compute, memory, and communication have emerged. Some examples include optical computation, near-memory computation, and in-memory computation. Optical computation [14] involves devices that use the exchange of photons as the fundamental means of representing information, and optical communication [15] uses light as the transport medium for data and logic. Optical computation and communication can be implemented using silicon photonics, and some real world examples of such devices include: optical computation [16, 17], optical communication [18, 19].

Near-memory computation [20] refers to eliminating certain levels of conventional memory hierarchy and bringing the data closer to computational cores or processing elements. This can have benefits through reduced data movement due to data residing closer to compute nodes thus saving latency and energy. Some examples of near-memory computation-based hardware include [21, 22, 23, 24]. In-memory computation, also called as compute-in-memory, refers to performing computation within memory using memory arrays or bit-cells and their interconnections to perform operations. The following subsection provides an overview of the CIM paradigm, potential benefits and disadvantages of CIM, and the various kinds of devices that can be used to implement CIM.

Compute-In-Memory

Energy consumption has been realized as the primary limiting factor in maintaining the historical rate of performance improvements in traditional and emerging computational applications [26]. Three key areas of focus where innovation is needed to continue improvements in energy efficiency are advances in: energy of compute operations (energy-per-operation), energy required to store and access data in memory (energy-per-memory

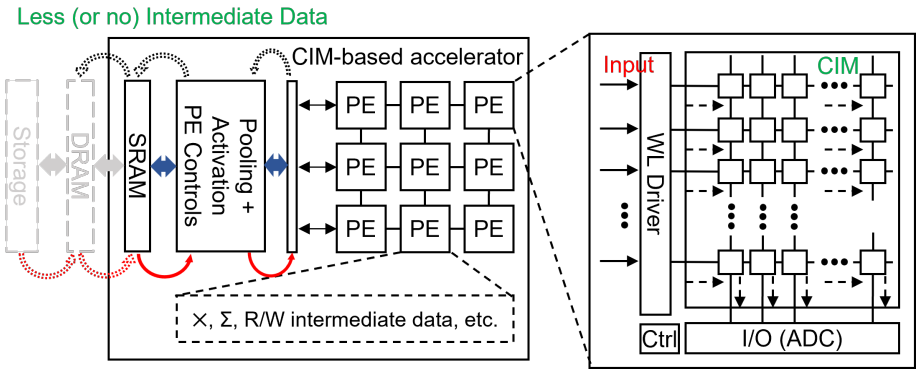
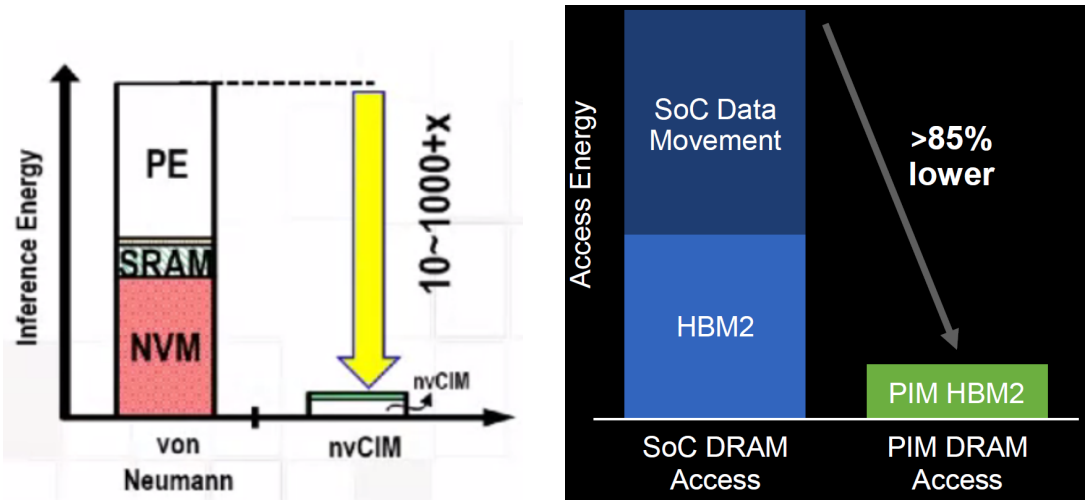


Figure 1.4: A DNN accelerator implemented using a Compute-In-Memory architecture.

bit), and energy required to communicate data externally (energy-per-communication bit) [26]. In the last few years it has been observed that energy efficiency (performance-per-watt) improvements have slowed across the heterogeneous computing landscape due to the challenges of scaling energy required across the three areas.

To improve energy efficiency (energy-per-operation and energy-per-memory bit), domain-specific architectures (such as Google's TPU [8]) or ASICs are being widely used, and could be customized for both cloud and edge applications [27]. The key bottleneck for deep learning acceleration is frequent data movement between compute units and memory units [5], that resembles the 'memory wall' challenge in traditional von-Neumann architectures. Domain specific architectures such as GPU and TPU do not solve the 'memory wall' challenge as processing elements or compute cores are at a distance from on-chip global buffers and off-chip main memory. vector-matrix multiplication (VMM) between the input vector and weight matrix, which is essentially a MAC operation, is the most energy intensive part of DNN processing. In light of this, CIM has been proposed as a promising paradigm as it realizes computation physically within the memory sub-arrays [28].

State-of-the-art image recognition model parameters have grown exponentially (upto 100's of MB [13]). CIM has been proposed as a potential paradigm for energy-efficient compute by reduced data movement and increased parallelism in image recognition and language translation computations. This can help with the reduction of energy-per-communication



(a) (b)

Figure 1.5: (a) By storing weights in the memory array and leveraging parallel MAC operations, CIM can achieve lower energy per operation through suppressed data movement across memory layers compared to von Neumann-based computation [25]. (b) Key algorithmic kernels can be executed directly in memory saving precious communication energy [26].

bit to improve energy efficiency [26]. A representative system (a DNN accelerator) implemented using CIM-based architecture is shown in Figure 1.4. By storing weights in the memory array and leveraging parallel multiply and accumulate, CIM can achieve lower latency and energy per operation through suppressed data movement across memory layers compared to von Neumann-based computation Figure 1.5. Reduced intermediate data can provide further reduction in latency.

Emerging NVM (eNVM) Devices for CIM

static random access memory (SRAM) technology has been used to implement CIM due to multiple benefits in large on-chip capacity and availability at the latest technology node. However, SRAM and dynamic random access memory (DRAM) have some drawbacks such as volatility with significant leakage power or refresh power consumption, especially in edge devices where dynamic power gating is desired. emerging nonvolatile memory (eNVM), such as RRAM, phase change memory (PCM) etc. may become more compet-

itive than SRAM/DRAM as CIM synaptic devices on power constrained platforms due to their non-volatility (turn on-off without losing stored weights), higher bit density and low leakage, enabling large embedded memory and high energy-efficiency. Since eNVM bitcells typically have a smaller layout area than SRAM bitcells and possibly offers multi-bit per cell, they can yield a higher integration density at the same technology node. eNVM devices of interest include RRAM [29], PCM [30], spin-orbit-torque magnetic random access memory (SOT-MRAM) [31], spin-transfer-torque magnetic random access memory (STT-MRAM) [32], ferroelectric field effect transistor (FeFET) [33] and electrochemical random access memory (ECRAM) [34]. In this work, we study RRAMs as the device of interest in system integration for their low-leakage and high bit-density features [35].

Some of the commercially available fabrication processes for eNVM include: Taiwan Semiconductor Manufacturing Company Limited (TSMC) 40 nm RRAM (capacity: 256K $\times$ 44) [36], TSMC 28 nm RRAM (capacity: 0.5 Mb) [37], TSMC 22 nm RRAM (density: 10.24 Mb/mm2) [38], Intel's 22 nm RRAM (density: 10.1 Mb/mm2) [39], TSMC's 40 nm PCM [40], STMicroelectronics' 28 nm PCM (capacity: 16MB) [41], TSMC's 22 nm STT-MRAM (capacity: 32Mb) [42], Intel's 22 nm STT-MRAM (density: 10.6 Mb/mm2, capacity: 7 Mb) [43], Globalfoundries' 22 nm STT-MRAM (density: 40 Mb/mm2) [44], Samsung's 28 nm STT-MRAM (density: 128 Mb/mm2, capacity: 8 Mb) [45], Globalfoundries' FeFET at 28 nm (capacity: 64 kbit) [46] and 22 nm (capacity: 32 MBit/cell) [47].

1.2 The Heterogeneous Integration Motivation

The varied demands of compute applications were traditionally met by monolithic fabrication of semiconductor devices, typically referred to as monolithic two dimensional (2-D) integrated circuits (ICs). Some examples of monolithic ICs include: NVIDIA RTX 4060 GPU [11], Intel Haswell based Core i7 CPU [50] and AMD Ryzen 4000 CPU [51], Cerebras Wafer Scale Engine [52], Apple A16 bionic SoC [53], etc. As discussed previously,

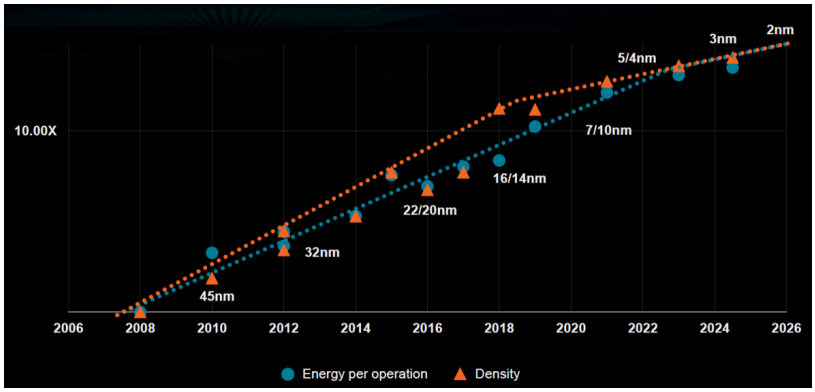


Figure 1.6: Non-linear trend for transistor density and energy per operation. [26]

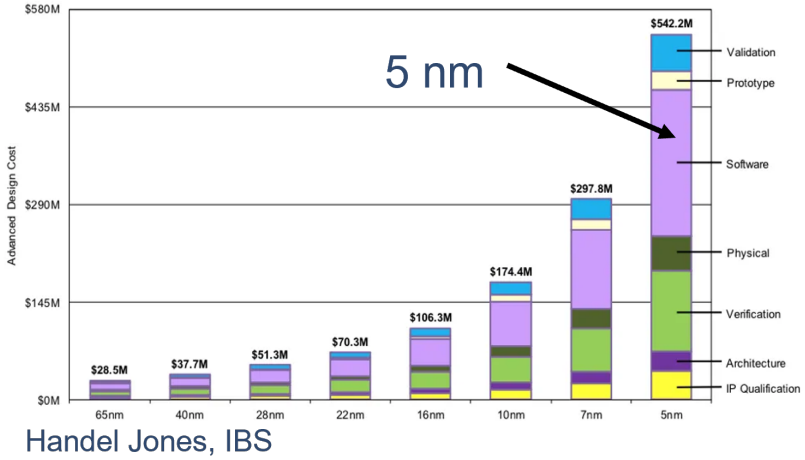


Figure 1.7: Surging average design cost of next-gen nodes. [48]

historic innovation in monolithic IC performance was through architectural innovation, process technology gains, larger die sizes, and higher power consumption. To address the growing compute and memory demands of data-intensive computing, more on-chip functionality needs to be added and this has lead to an unsustainable increase in die sizes [54]. However, non-linear trends in device densities and energy per operation (Figure 1.6 [26]), and the rising design costs of advanced technology nodes (Figure 1.7 [48]) has made conventional feature scaling, and thus working with large die sizes, an expensive pursuit [26]. Furthermore, not all parts of an SoC benefit from the leading-edge costly technology (Figure 1.8 [49]). There are additional challenges with monolithic integration from design and device/material aspects. When a new technology node is introduced, existing intellectual

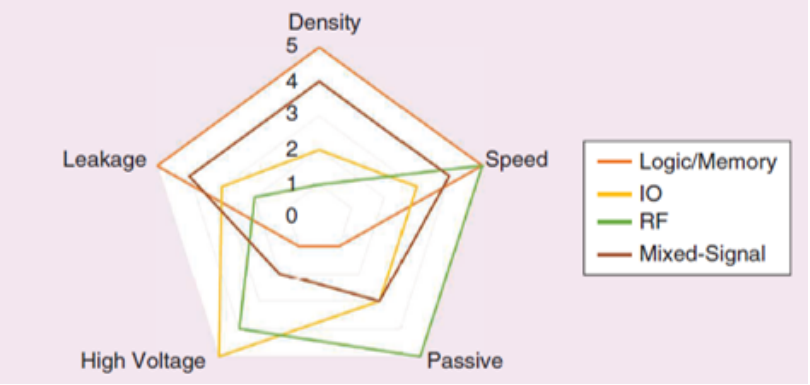


Figure 1.8: Application-based disparate technology requirements. [49]

property (IP) needs time to mature to the design rules of the new node causing potential delays in fabrication. Introducing new devices and materials to complementary metal oxide semiconductor (CMOS) processes can also be a challenge from an economic and logistical perspective. For instance, photonic devices require poly-Silicon while CMOS does not uses poly-Si, and the leading edge photonic devices are fabricated at 45nm [55] while leading-edge CMOS is at 5nm [56]. Similarly, integrating devices with new materials such as PCM [30], RRAM [29], FeFETs [33] with mature 60-mask-layer CMOS processes [57] could have challenges making them not easy to implement monolithically. A combination of these reasons pushed the packaging technology to move toward modular designs.

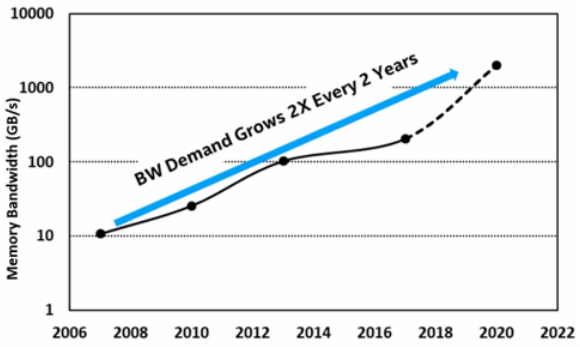


Figure 1.9: CPU–memory BW trend per memory device as a driver for Heterogeneous Integration. Memory BWs have continually increased over time to support the CPU performance. [58]

1.2.1 Opportunities for Heterogeneous Integration: A move towards modular designs

The International Roadmap for Devices and Systems (IRDS) 2020 "Package Integration" white paper [59] describes "HI" or "dense off chip integration (DOCI)" as the "approach and strategy of using advanced packaging (AP) consisting of scaling, higher feature densities of traditional package elements, materials and structures leading to tighter integration and better performance to integrate at the package rather than single chip level, for systems with conventional single-die like performance at lower cost." Some potential benefits of heterogeneous processes and integration include: multiple processes optimized for individual IPs, multi-chip integration by AP, and no reticle limit on the overall product. HI and AP is driven by three factors: 1) Faster movement of big data, which translates to need for high bandwidth (Fig. Figure 1.9), low latency, and low power interconnection, 2) need to integrate IP on different nodes and fabrication processes, and 3) yield resiliency [60].

Some examples of heterogeneous integration in the literature include Xilinx's "Everest" [61], AMD's chiplet-based "Rome" and "Matisse" SoCs [62], as well as Intel's bridge-based [63] and through-silicon via (TSV)-based 3-D [64] integration. A fundamental objective of most advanced integration schemes is to connect dis-aggregated chips to match the functionality of monolithic SoCs.

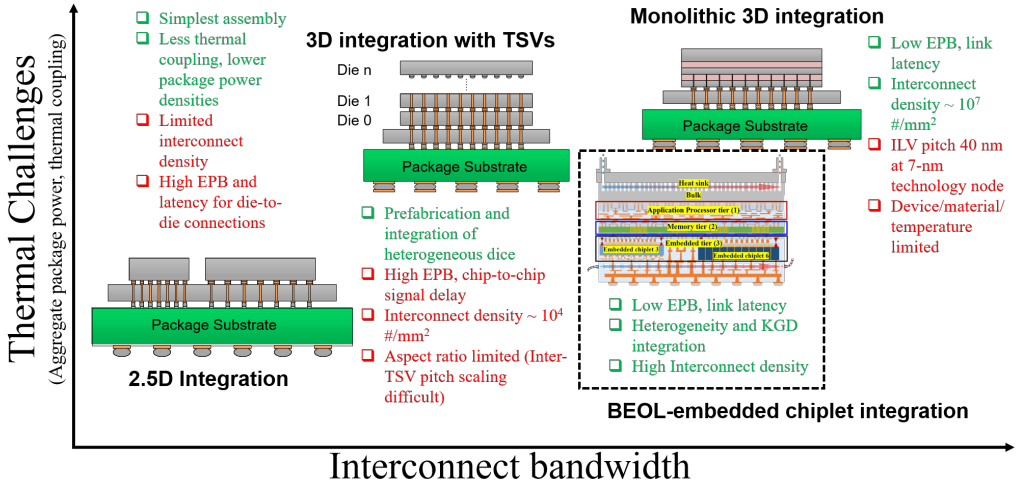


Figure 1.10: Move to Heterogeneous Integration

Benefits and Trade-offs of 3-D Heterogeneous Integration

The need for higher bandwidth and lower delay in chip-to-chip signal interconnections has led to a technological push towards modular architectures such as TSV based three dimensional (3-D) ICs [65]. Some of the benefits of TSV-based 3-D integration include lower signaling energy-per-bit (EPB), lower link latency, and higher interconnect density compared to other enhanced-2-D integration schemes such as interposers and bridge-based integration. However, relative to monolithic 3-D ICs, conventional TSV-based 3-D integration is expected to have higher EPB, higher inter-chip link latency, and lower interconnect density [66]. Owing to this performance gap (Figure 1.10), there is a significant interest in monolithic 3-D fabrication. However, limitations in devices, materials, and temperatures make monolithic 3-D integration challenging and limiting.

Monolithic 3-D ICs can have higher 3-D connectivity (with the use of nanoscale inter-layer vias) but are limited in providing heterogeneity of devices at disparate technology nodes. On the other hand, while conventional microbumps and TSV-based 3-D die stacking facilitates integration of pre-fabricated dice of varying technology nodes, such integration provides limited interconnect density relative to monolithic 3-D integration [67].

Need for HI in CIM

Among various heterogeneous integration (HI) architectures, such as multi-chip module (MCM), 2.5-D, and 3-D, 3-D-HI can provide higher compute density and signaling EPB through a reduced footprint and interconnection length, respectively, compared to MCM and 2.5-D [68]. A growing need for higher logic-memory bandwidth and lower chip-to-chip signal interconnection delay have led to a technological push towards 3-D-HI such as through-silicon via (TSV)-based 3-D integrated circuits (ICs) [68, 69, 70]. Although HI can enable dense memory-logic integration needed for state-of-the-art CIM hardware accelerators, there are power delivery and thermal challenges with HI for CIM.

	Pitch at 14 nm	Pitch at 10 nm	Scaling Factor	Aspect Ratio
M0		40		1.59
M1	62	40	0.645161	1.19
M2	61	36	0.590164	1.15
M3	55	44	0.8	1.15
M4	52	44	0.846154	1.15
M5	52	52	1	1.6

(a)

	Relative C/μm	Relative R/μm	RC Delay (per μm^2)	Scaled RC Delay
14nm	1.0	1.0	1.0	1.0
10nm	0.84	2.83	2.40	1.2

(b)

Figure 1.11: BEOL capacitance and resistance change node to node [71]

1.2.2 Interconnection Challenges

BEOL impedance scaling challenges

According to the IRDS 2020 "More Moore" roadmap [3], interconnect resistance has entered an exponential increase domain due to non-ideal scaling of barrier material for Cu, leading to less conductor volume and increased scattering at the interconnect surface and grain-boundary interfaces. Clocking frequency (f_{max}) at nominal supply voltage is forecasted to improve from 3.1 GHz in 2020 to 3.5 GHz in 2025, and 2.9 GHz in 2034. This limited scaling is due to increasing parasitics, particularly interconnect resistance, and limited gate drive ($V_{gs} - V_t$) as a result of supply voltage scaling.

A recent study [71] has highlighted the impact of change in node to node back end of line (BEOL) capacitance and resistance. RC scaling from node to node is driven by multiple factors including conventional dimensional scaling along with material and structural changes. With a case study based on Intel's Core i7 CPUs, according to the authors BEOL R,C scaling from 14 nm to 10 nm suggests a 0.84$\times$ decrease in line-to-line C/μm, 2.83$\times$ increase in R/μm, and 2.4$\times$ increase in RC delay/μm^2 (Figure 1.11). Therefore, it is important to understand and better characterize the challenges that change in BEOL parasitics pose to signaling performance of both conventional and non-von Neumann digital systems.

As 3-D integration is expected to be more widely adopted in the next decade as a form

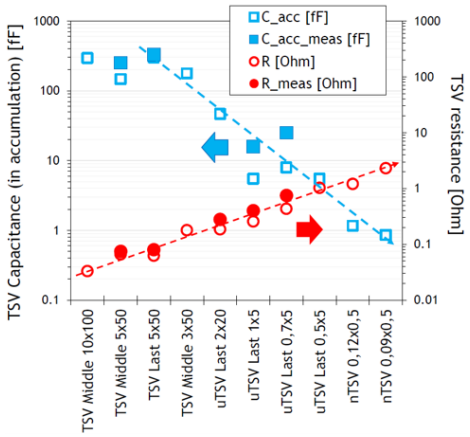


Figure 1.12: TSV capacitance and resistance change [75]

of HI [3], it is important to analyze different integration schemes for potential signaling performance bottlenecks and non-starters to provide design guidelines for a viable design space. Compared to single die system-on-chips, TSV-based 3-D [65] enables diverse heterogeneity in device integration from different technology nodes and improves overall yield through splitting larger monolithic dice into multiple smaller dice [72]. Monolithic 3-D integration is enabled through fabrication of high-density fine-pitch inter layer vias (ILVs), low-temperature active layer fabrication processes, and emerging nanotechnology techniques [73]. ILVs can enable higher inter-layer connectivity compared to both conventional 2-D and TSV-based 3-D and higher interconnect density than TSV-based 3-D [74, 67]. This implies lower EPB and link latency between devices in different active layers in a monolithic IC compared to TSV-based 3-D ICs.

Furthermore, TSVs, which are used to enable heterogeneous integration of disparate logic and memory units, have seen a steady scaling trend with reduction in pitch/diameter and thus enabling interconnect density. However, this leads to worsening of TSV resistance, which can be seen from Figure 1.12.

Power Delivery Challenges: Implications for Monolithic and HI

A factor that poses challenge in power delivery design for edge intelligent hardware is the current trend of increasing power and power density in recent CIM and hardware accel-

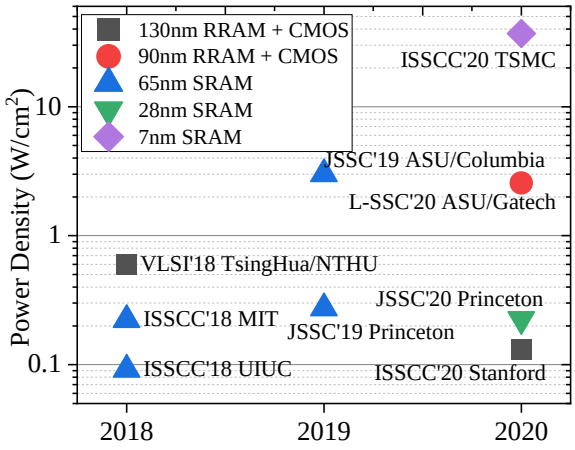


Figure 1.13: Power density trend of recent CIM hardware accelerators.

erators, as shown in Figure 4.1. Increasing DNN model sizes and workload complexity can lead to larger die sizes due to a higher demand for on-chip resources such as memory arrays, analog-to-digital converters (ADCs), etc. When it comes to low-power edge applications, the primary motivations are to improve energy efficiency (tera-operations-per-sec-per-Watt or TOPS/W) and compute efficiency (tera-operations-per-sec-per-mm^2 or TOPS/mm^2). This can be achieved through device scaling and reducing the overall hardware form-factor. As a result, the area occupied by an edge intelligent hardware and voltage regulators will need to shrink. A push for thinner devices usually corresponds to reduction in height of the die and the power delivery components such as interconnects, capacitors and inductors. Additionally, recent work has demonstrated performing vector-matrix-multiplication in parallel on multiple CIM cores, which introduces certain non-idealities such as core-to-core variation of IR-drop and supply voltage instability [76]. All these trends introduce multiple unique challenges in designing a robust PDN for CIM.

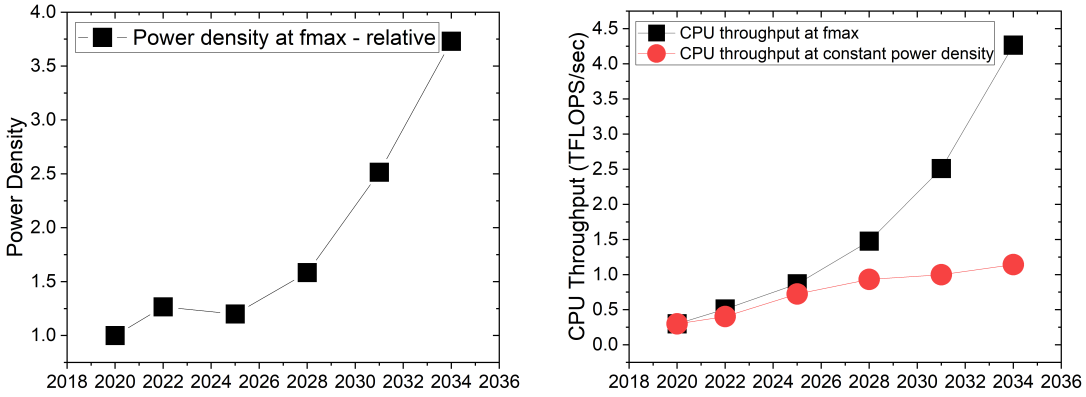
A challenge in the realization of advanced 3-D heterogeneous integration is evaluating the potential impacts and benefits of the PDN. According to [77], for the same design implemented in 2-D versus 2-stack 3-D, the 3-D design consumes potentially 50% of the 2-D area along with a similar power requirement. However, since the 3-D design has fewer number of bumps, the current per bump is increased due to a smaller footprint. This trans-

lates directly to increased power density and requires evaluation of PDN design parameters for careful floorplanning to avoid power and thermal hotspots.

Emerging 3-D integration techniques such as die to wafer (D2W) and wafer to wafer (W2W) are promising for "More than Moore" scaling and DOCI for advanced technology nodes. For these integration techniques, an optimal PDN design for meeting design specifications is non-trivial. This is because with 3-D integration there can be challenges such as inter-tier power supply noise (PSN) coupling [78], trade-off between inter-tier power and signal routing, and limited scaling of BEOL interconnect parasitics.

Related work from literature has explored some of the above highlighted challenges. Yang et. al. [79] presented a PDN analysis framework for emerging 2-D enhanced (2.5-D) and 3-D heterogeneous integration platforms and benchmarked interposer and bridge-chip-based integration technologies from a PDN perspective. Their analysis of impact of bridge-based integration suggested minimizing the overlap region between a bridge and a die, and using multiple bridge-chips instead of a single large bridge-chip to mitigate PSN. They also proposed using through-bridge vias (TSVs) to improve PSN. However, this analysis only focused on PDN challenges and opportunities for 2.5-D integration. Kahng et. al. [80] presented a a novel power delivery path-finding methodology for emerging D2W face-to-face (F2F) integration. Their study shows "scale-independence" of IR drop behavior due to regular power and ground TSVs for a D2W F2F 3-D-IC example.

Furthermore, in the wake of limited PPAC gains from conventional scaling, multiple scaling boosters have been proposed for advanced nodes [3]. These include gate all around (GAA)structures such as lateral nanosheets [81], sequential integration of vertical and lateral GAA devices [82], backside power delivery network (BS-PDN)[83], etc. In [83], the authors present a PDN modeling framework for BS-PDN configurations. They reported greater than $4\times$ reduction in PSN in the BS-PDN configurations relative to conventional BEOL PDN. Thus, it is important to evaluate the potential benefit of such advanced power delivery schemes for advanced HI techniques.



(a) (b)

Figure 1.14: Scaling projection of (a) power density and (b) computation throughput of CPU cores at the maximum clock frequency and at thermally-constrained average frequency

1.2.3 Thermal Challenges

Stalled Power Densities and Increasing Package Power

An exponential increase in transistor density and limitations in supply voltage scaling and interconnect parasitics, particularly interconnect resistance, have lead to a stall in microprocessor powers ("power wall"), and thus, clock frequency. Additionally, in the post-Dennard era, stalling of power densities due to limitations of conventional cooling techniques constituted a "thermal wall." According to the 2020 IRDS [3], power density poses a significant challenge for scaling, especially due to it's expected 2.5× increase by 2031 due to 3-D integration (Figure 1.14a). Furthermore, thermal constraints are expected to reduce the average CPU frequency to 0.8 GHz. This raises an opportunity for advanced cooling solutions to maintain an overall computation throughput scaling of ×14 over six node generations instead of ×3.8 with a thermally-constrained system (Figure 1.14b). The thermal design power of server CPUs and GPUs have also been increasing by ≈ 7% per year over the last decade Figure 1.15 [54].

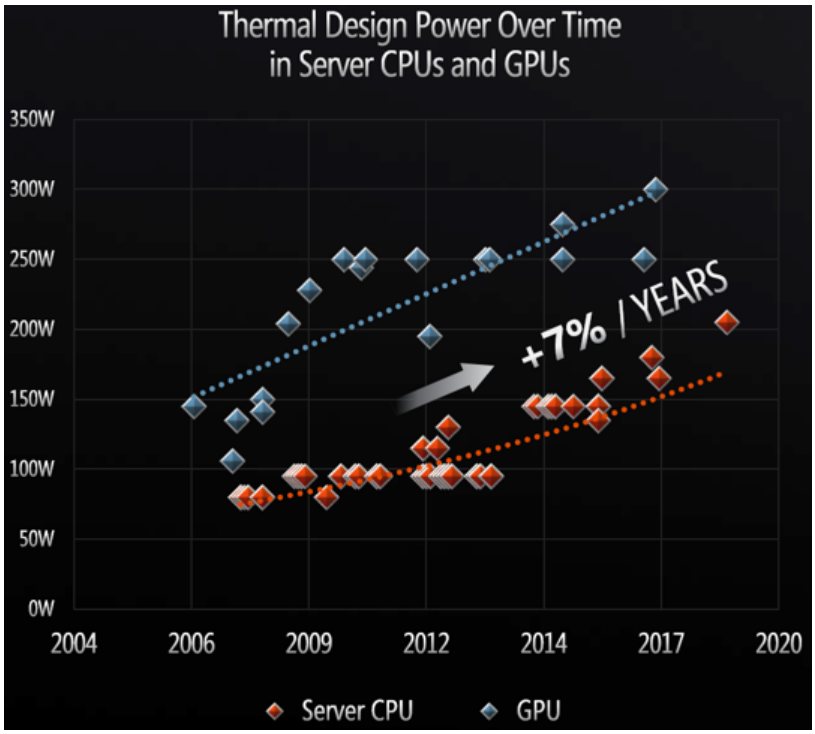
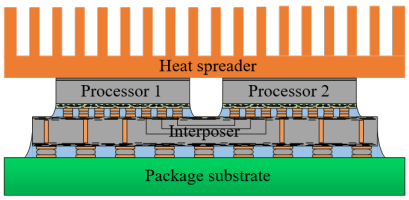


Figure 1.15: Increasing TDP of server CPUs and GPUs over the last decade. [54]

Thermal Implications for Monolithic and HI

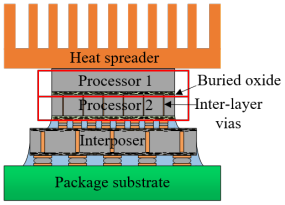
Thermal challenges described previously impact both monolithic and HI systems. For monolithic ICs, higher junction temperatures might lead to thermal throttling and other reliability related challenges such as negative-bias temperature instability (NBTI) and electromigration. For modular designs, that constitute multiple dice assembled in close proximity, in addition to the challenges with monolithic ICs, thermal coupling is an additional challenge. This subsection covers the challenges in thermal coupling with a case study to compare thermal coupling effects in representative HI systems.

Silicon interposer-based integration is capable of supporting higher interconnect densities (0.5-1.0 μm line/space) than organic substrates (2-5 μm) along with less thermal coupling and lower package power densities compared to 3-D integration [84]. However, Si-interposers are potentially more expensive compared to organic substrates, highlighting a tradeoff between cost and density. Moreover, interposer-based links can also have higher



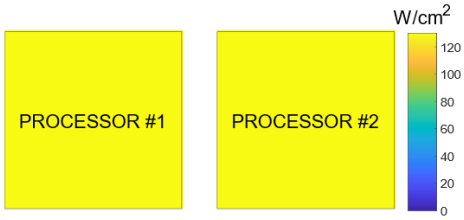
Interposer 2.5-D integration

(a) Interposer-based 2.5-D integration

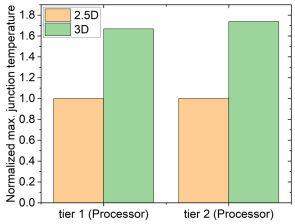


3D integration

(b) 3-D integration



(c) Power Densities



(d) normalized maximum junction temperatures

Figure 1.16: (a) Interposer-based 2.5-D and (b) 3-D integration examples. (c) Interposer-based 2.5-D and 3-D integration normalized maximum junction temperatures: Tier powers: 1. processor (150 W), 2. processor (150 W)

EPB and latency for die-to-die connections compared to 3-D integration.

Figure 1.16 (a,b) shows examples of 2.5-D and 3-D integration technologies. In the first approach (Figure 4.2a), active dice are connected to an interposer using C4 or microbumps for inter-die connectivity. Figure 4.2b shows a monolithic 3-D integration scheme where inter-layer vias are used for inter-tier connectivity of active device layers. As shown in Figure 1.16d, using conventional cooling techniques, 3-D integration of logic-on-logic tiers leads to a worst case 73% higher maximum junction temperatures ($T_{j,max}$) compared to an equivalent 2.5-D case, for a uniform per tier power of ≈150 W based on server thermal design power (TDP) estimates [85]. This can be attributed to increased volumetric power in 3-D ICs, which can lead to higher inter-tier steady state temperatures and transient thermal coupling. Moreover, this disparity in 3-D thermal performance is expected to only worsen with additional tiers and with the presence of hot-spots. However, there are significant electrical benefits from 3-D integration technologies including lower signaling EPB, lower interconnect latency, and higher interconnect density compared to other 2.5-D integration

schemes such as interposers and bridge-based integration [66, 84].

1.2.4 Device-System Integration Challenges

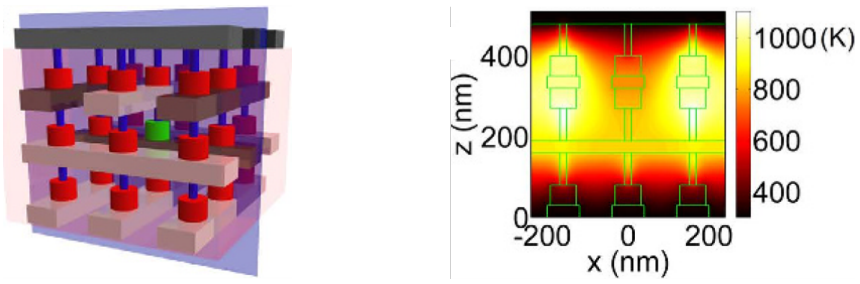
Application Performance as a Function of Physical Effects

Resistive Random Access Memory (or RRAM) [86] is an attractive form of emerging non-volatile memory in hardware implementation of neural networks. Device defects in RRAMs over time lead to temporal variation in bit-cell conductance, which contributes to loss in neural network image recognition accuracy over time. Previous retention induced conduction variation studice [87, 88] have not considered thermal effects on RRAM device tiers due to neighboring tiers in a 3-D IC form factor. Moreover, there are no prior efforts examining how 3-D integration approaches and choice of cooling architecture directly impacts RRAM's retention.

Thermal crosstalk and scaling potential under thermal effects in a 3-D RRAM crossbar array were investigated in [88]. The authors suggest that thermal crosstalk in 3-D RRAM arrays could deteriorate device retention performance and lead to data storage failure from LRS (low resistance state) to HRS (high resistance state) of the disturbed RRAM cell (Figure 1.17). They provide and verify, via numerical simulations, potential methods to alleviate thermal crosstalk. In [89], the authors studied the temperature behavior of RRAMs based on HfO_2 dielectric. They showed, via simulations, that RRAM resistance increases as temperature rises and at a certain temperature diffusion effects rise exponentially and destroy the conductive filament (CF).

Thus, it is important to evaluate the impact of cooling architectures on binary and multi-level RRAM devices in a 3-D IC form factor by quantifying image recognition accuracy over time of BEOL RRAM hardware for applications such as CIM accelerators for image recognition [90].

Power supply noise effects in 3-D-HI can bring additional challenges to computational accuracy. These include steady-state PSN due to IR-drop on additional interconnects (re-



(a) (b)

Figure 1.17: (a) A 3-D RRAM array and (b) corresponding thermal crosstalk [88]

sistive on-die PDN, TSVs, I/O bumps, etc) and inter-tier supply voltage variation. These effects lead to variations in the analog outputs of the memory arrays and the reference voltages in the ADCs contributing to sensing errors in the ADCs. These errors can significantly impact CIM inference accuracy.

1.3 Research Objectives and Contribution

The objective of this research is to investigate power delivery network, thermal management, and die-to-die signaling constraints in 3-D heterogeneous integration architectures for compute-in-memory applications.

1. The design trade-offs of including a PDN and MIM capacitors in bridge-chip based 2.5-D heterogeneous platforms are investivated. While bridge-chip-based interconnect platforms may potentially present PDN challenges, it is demonstrated that including the PDN in the bridge-chip can be an effective technique to reduce both DC-IR drop and *Ldi/dt* noise. Results from inclusion of a PDN and MIM capacitors in the bridge-chip and the corresponding impact on DC-IR drop, *Ldi/dt* noise, and high-frequency ripple are presented. The tradeoffs between bridge-chip sizing and bridge-chip PDN configurations on the maximum transient supply noise are explored. We demonstrate that MIM capacitors in the bridge-chip PDN are an effective technique to meet PSN design targets for 2.5-D architectures.

2. To address the power delivery challenges in 3-D-HI described earlier in this chapter, a systematic technology and design space exploration of power delivery for 3-D-HI CIM systems is presented. A fast analysis flow facilitating early design-space exploration between power delivery design parameters and CIM performance metrics is proposed. By co-optimizing 3-D PDN and SAR-ADC design parameters a balanced 3-D CIM design is demonstrated compared to a 3-D naive implementation at iso-power and iso-area. Trade-offs for such an approach are discussed.

3. A device-integration towards application-level reliability evaluation methodology is proposed that can be used to quantify the direct impact of integration design parameters on CIM inference accuracy. Using this flow, heterogeneous 3-D logic-memory CIM accelerator designs are benchmarked against monolithic 2-D and balanced integration design parameters for maximized 3-D CIM inference accuracy are reported. The benchmark framework is released as an open-source tool for the research community.

4. A 3-D polylithic architecture is proposed that represents a densely integrated system divided into multiple device tiers where custom chiplets, such as power management IP, I/O drivers, and memory are embedded into the back-end of a base tier with extreme efficient signaling and large bandwidth density. The thermal constraints for polylithic 3-D are evaluated with aggressive cooling to investigate thermal limits from transient- and steady-state perspectives.

5. Design optimization strategies for PDN in polylithic 3-D integration are presented. The scope of this work is a detailed design space exploration of the power supply noise effects in polylithic 3-D architectures. We propose three polylithic PDN designs and benchmark their IR drop as a function of tier power, number of embedded chiplets, hot-spot location, and TSV diameter and distribution to provide design limitations and insights.

1.4 Organization of this Thesis

The chapters of this thesis are organized as follows:

1. **Chapter 2** investigates design trade-offs in the PDN of bridge-chip based 2.5-D heterogeneous platforms. We demonstrate that including a PDN in the bridge-chip can provide significant reduction in DC-IR drop, Ldi/dt noise, and high-frequency ripple compared to the baseline.

2. **Chapter 3** presents a device-integration methodology to facilitate early design-space exploration. Trade-offs between power delivery design parameters and CIM performance metrics are quantified.

3. **Chapter 4** proposes a device-integration towards application methodology to quantify the impact of integration architectures on RRAM reliability for CIM applications. A comprehensive design-space exploration of PDN design for 3-D-HI CIM hardware is presented.

4. **Chapter 5** presents the thermal evaluation of a back-end-of-line-embedded chiplet integration scheme (polylithic 3-D), where custom chiplets are embedded into the BEOL of an application processor tier for CIM applications.

5. **Chapter 6** presents a study to evaluate the power delivery constraints for polylithic 3-D integration of BEOL-embedded chiplets for CIM applications.

6. **Chapter 7** discusses the potential impact of this research and potential future directions are summarized.

CHAPTER 2

DESIGN CONSIDERATIONS FOR POWER DELIVERY NETWORK AND METAL-INSULATOR-METAL CAPACITOR INTEGRATION IN BRIDGE-CHIPS FOR 2.5-D HETEROGENEOUS INTEGRATION

The current trend of increasing power and power densities in recent CIM and hardware accelerators, as shown in Figure 2.1 poses a challenge for efficient and low-cost power delivery design for edge-intelligent hardware. Increasing DNN model sizes and workload complexity can lead to larger die sizes due to a higher demand for on-chip resources such as memory arrays, ADCs, etc. When it comes to high-performance applications, the primary motivations are to increase overall system throughput (tera-operations-per-sec or TOPS) and energy efficiency (TOPS/W). Higher system throughput can be achieved through either higher operating frequency or added resources for parallel computing, both of which can lead to increased system power dissipation. This can lead to thermal challenges in high-performance hardware and particularly in CIM hardware that has a higher demand for on-chip resources. As opposed to Si-interposers, that have limits on the maximum interposer size for increasing on-package silicon, bridge-based 2.5-D HI is a promising approach to scale on-package silicon. However, all these trends introduce multiple unique challenges in designing a robust PDN for 2.5-D integrated CIM systems.

There can be additional challenges to CIM computational accuracy using bridge-based 2.5-D-HI. These include steady-state PSN due to IR-drop in areas of bridge and chip overlap and lack of power-sharing between dice through the bridge-chip. Increased PSN in memory arrays can lead to variations in array analog outputs and reference voltages in ADCs, leading to sensing errors in ADCs. These errors can significantly impact CIM inference accuracy. To address these challenges, we present a systematic technology and design space exploration of power delivery for 2.5-D-HI systems in this chapter.

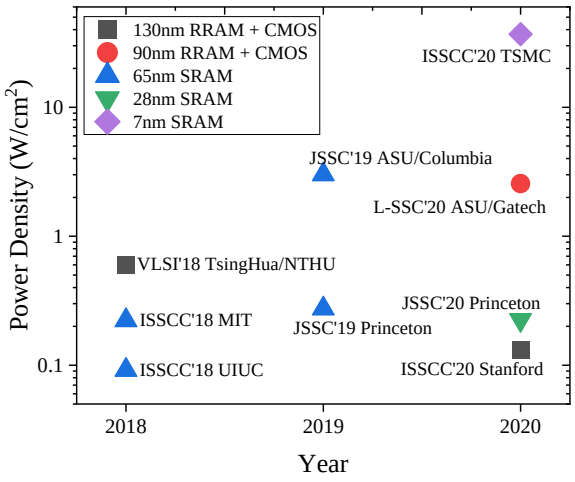


Figure 2.1: Power density trend of recent CIM hardware accelerators.

2.1 Introduction

The heterogeneous integration of chiplets fabricated using different process nodes is proving to be critical in continuing the performance and energy scaling of high-performance computing systems while also addressing economic viability. Advanced integration technologies such as silicon interposer [91], localized bridge-chip interconnects, (2.5-D integration [92, 93, 94, 95]) and 3-D die stacking ([96, 97, 98]) enable extreme interconnect densities. This has led to the rapid adoption of chiplet based architectures.

High-density interconnect routing on an organic package is challenging for several technical limitations [99]. In contrast, silicon interposer technology offers much higher interconnect density due to silicon-based processes. From a power delivery perspective, a potential advantage of using Si interposers is the possibility of integrating deep-trench capacitors (DTCs) [100] or MIM capacitors [101] in the interposer for high-frequency noise suppression. However, using Si-interposers presents some challenges as all the power from the package is routed through the TSVs resulting in added resistance and inductance. In addition, the Si-interposer needs to be large enough to accommodate all needed dice mounted, which can lead to yield and cost challenges.

To circumvent these issues, localized silicon bridges with just enough area to support

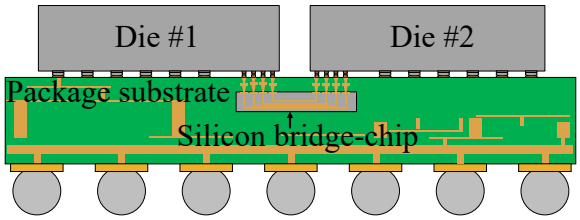


Figure 2.2: Localized silicon bridge-based 2.5-D heterogeneous integration.

die-to-die signaling interconnects can be located near die peripheries of adjacent chiplets. However, power delivery remains a challenge as the bridge-chip overlap may limit access of die peripherals to the package PDN [99]. Moreover, the number of on-chip power rails continues to scale up with the core count, which leads to a steady increase in the thermal design power [99, 102].

We investigate the design trade-offs of including a PDN and metal-insulator-metal (MIM) capacitors in bridge-chip based 2.5-D heterogeneous platforms. While bridge-chip-based interconnect platforms may potentially present PDN challenges, we demonstrate that including the PDN in the bridge-chip can be an effective technique to reduce both DC-IR-drop and *Ldi/dt* noise. We consider three scenarios: (a) inclusion of ground network in the bridge-chip, (b) inclusion of power and ground network in the bridge-chip, and (c) inclusion of metal-insulator-metal (MIM) decoupling capacitors in the bridge-chip. We model a bridge-chip based two-die PDN with die #1 emulating a high-power CPU and die #2 as an FPGA. We implement a distributed package-level PDN model to reflect the spreading effects of current in the package and the coupling between different P/G bumps. The key contributions of this work are:

1. We demonstrate that including the PDN in the bridge-chip can reduce DC-IR-drop by up to $\sim$23%, lower *Ldi/dt* noise by up to $\sim$19%, and reduce the high-frequency ripple by $>3\times$ compared to our baseline configuration.

2. We explore the tradeoffs between bridge-chip sizing and bridge-chip PDN configurations on the maximum transient supply noise. We demonstrate that MIM capacitors

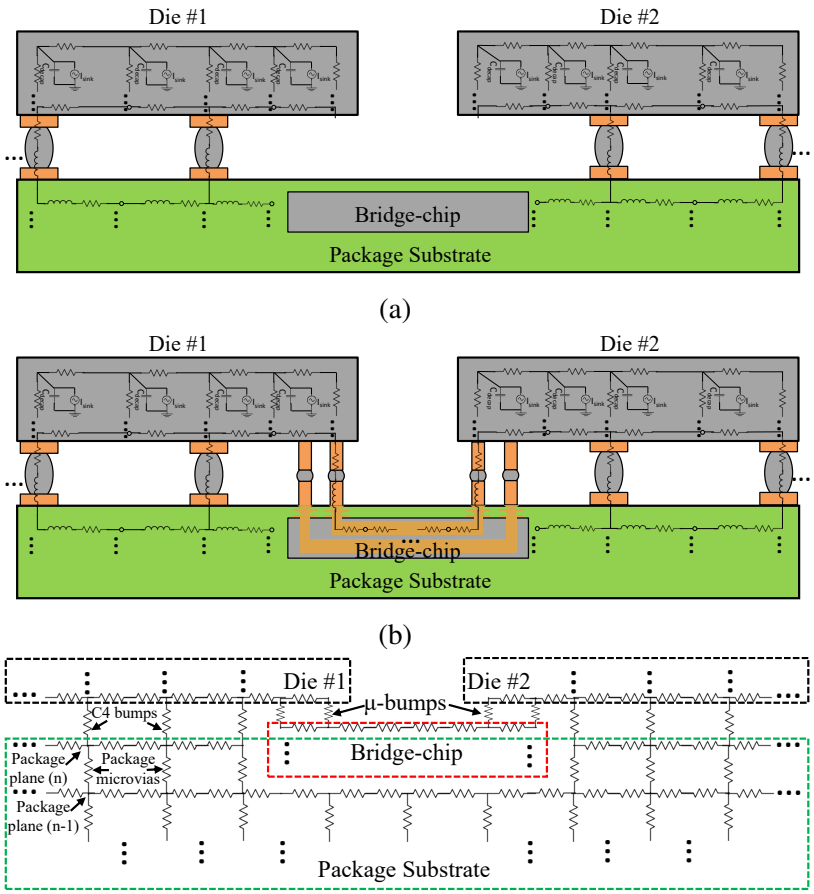


Figure 2.3: PDN schematic diagram (a) excluding bridge-chip PDN and (b) including bridge-chip PDN (c) showing the package P/G planes form a parallel resistance with the bridge-chip PDN for the on-die peripheral circuits [103].

in the bridge-chip PDN are an effective technique to meet PSN design targets for 2.5-D architectures.

2.2 Design tradeoff methodology and PDN Specifications with bridge-chip PDN

For bridge-based interconnect technologies such as embedded multi-die interconnect bridge (EMIB) [92], elevated fanout bridge (EFB) [93], direct bonded heterogeneous integration (DBHi) [95], and heterogeneous integration stitching technology (HIST) [94] the introduction of the Si-bridge may prevent the area of the die that overlaps the bridge-chip from

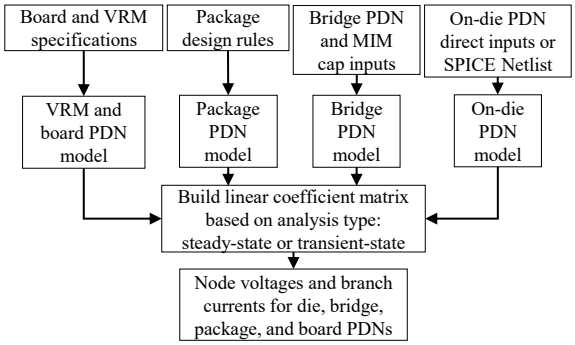


Figure 2.4: MATLAB-based PDN modeling methodology including bridge-chip PDN models with MIM capacitors.

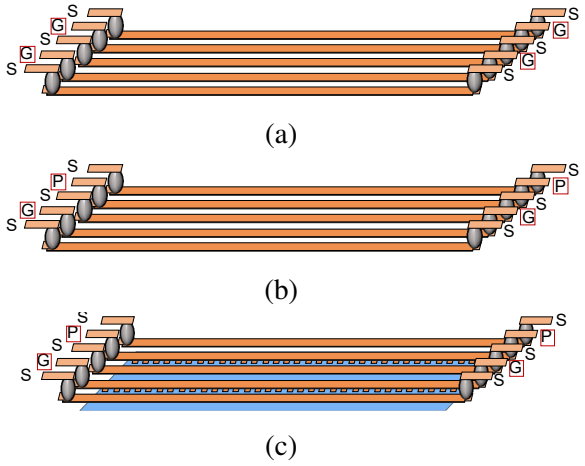


Figure 2.5: (a) Ground net in the bridge-chip, (b) power and ground nets in the bridge-chip, and (c) metal-insulator-metal capacitors in the bridge-chip [103].

having a direct power delivery path from the package [99]. Moreover, a break in the package PDN can be created in some of these technologies as a cut in the package surface is needed to make room for the bridge [95]. In other bridge-chip based package technologies the bridge-chip is mounted on the package surface, which limits direct access to the package PDN for the on-chip peripheral circuits [93, 94]. We aim to analyze a complete picture of various 2.5-D bridge-based integration. Therefore, our assumption is for the worst case, i.e., the on-die peripheral circuits do not have direct access to the package PDN in the bridge-chip and die overlap region.

Schematics for two bridge-chip power delivery scenarios considered in this work are

Table 2.1: Parameters of the PDN model [104]

Parameter	Value
On-die metal resistivity ($\Omega \cdot m$)	1.8e-8
On-die global wire Pitch/Wdith/Thickness (μm)	39.5/17.5/7
On-die intermediate wire P/W/T (nm)	560/280/506
On-die local wire P/W/T (nm)	160/80/144
on-die decap density (nF/mm^2) microbump pitch/R/L ($\mu m/m\Omega$/pH)	335 40/30.9/11.1
C4 bump pitch/R/L ($\mu m/m\Omega$/pH)	200/14.3/11.0
Package effective decap R/L/C ($m\Omega$/pH/μF)	541.5/220.7/52
Package resistivity/inductance ($m\Omega/mm$/pH/mm)	1.2/24
BGA pitch/R/L ($\mu m/m\Omega$/pH)	500/38/46
TSV R/L ($m\Omega$/pH)	54.2/77.78
PCB R/L ($\mu\Omega$/pH)	166/21
PCB Decap R/L/C ($\mu\Omega$/nH/μF)	166/19.54/240
Bridge-chip sizing (width (*mm*) $\times$ length (*mm*))	w = [1.5, 2.5, 4.5] $\times$ l = 6
MIM capacitor density (*nF/mm*^{2})	0, 5, 10, 25

shown in Figure 2.3a and Figure 2.3b. Prior work [105] considered the case described in Figure 2.3a, where the peripheral circuits on the die can not directly access the package PDN. A two die system is considered where die#1 represents a FPGA (total peak power of 44.8 W), and die#2 represents a processor die (total peak power of 74.49 W), respectively. These assumptions are hypothetical powers based on the power profiles of a 22 nm CPU die [105] and a 14 nm FPGA die [7]. The current density maps of each die and the chip-package-board PDN parameters for our model were assumed from [104]. The resistance per unit wire length for the package PDN is assumed to be 1.2 $m\Omega/mm$ [104]. The product of the number of metal layers (here two), the package size (a proxy for wire length assumed as x and y dimensions), and the square of resistance per unit length is used to estimate the total package DC resistance (DCR). This results in a package PDN DCR of $\sim$0.635 $m\Omega$, which is in close proximity of average package PDN resistance from the literature [106]. Resistance of the bridge-chip is assumed to be 20$\times$ that of the package (i.e. approximately 12.7$m\Omega$) [103]. The metal thickness for the bridge-chip is assumed to be the same as the on-die values since we model a Si-bridge, as summarized in Table 2.1 [104]. Package

mounted decoupling capacitors (decaps) are assumed to be integrated on the top side (die side) of the package with a uniform distribution outside of the die area.

In this work, microbumps are assumed to be a part of the PDN and the on-die PDN is connected to the bridge-chip PDN through a few microbumps. Next, the peripheral circuits of separate dice are assumed to share a single voltage domain. The bridge-chip metal-stack is assumed to be multi-layered to accommodate both a PDN and a signaling network. The following three cases are investigated:

- Ground (VSS) network included in bridge-chip: Availability of a common voltage domain between adjacent dice may not always be feasible. It is more common to have a shared ground between multiple voltage domains and adjacent dice (Figure 2.5a).
- Both power (VDD) and ground (VSS) network included in the bridge-chip (Figure 2.5b).
- MIM decoupling capacitors included in bridge-chip: If power and ground networks are included in the bridge-chip, MIM decoupling capacitors could be fabricated between adjacent power-ground metal layers (Figure 2.5c).

2.3 Bridge-Chip PDN Analysis for 2.5-D CPU-FPGA Integration

2.3.1 Including power and ground network in the bridge-chip

An increase in a microprocessor's load current can cause the on-die voltage to reduce temporarily (and vice versa) due to a fairly slow response time of the platform-level voltage regulation control loop (order of microseconds) [99]. These voltage fluctuations seen by the die due to voltage regulator (VR) latency can impact the system performance and latency. Thus, the total effective DC resistance (DCR) of the PDN, starting from the voltage regulator output to the on-die switching load, is an important metric to improve system efficiency and performance. Steady-state supply noise (IR-drop) due to the DCR can have implications for both analog and digital circuit performance. Lower supply voltage resulting from

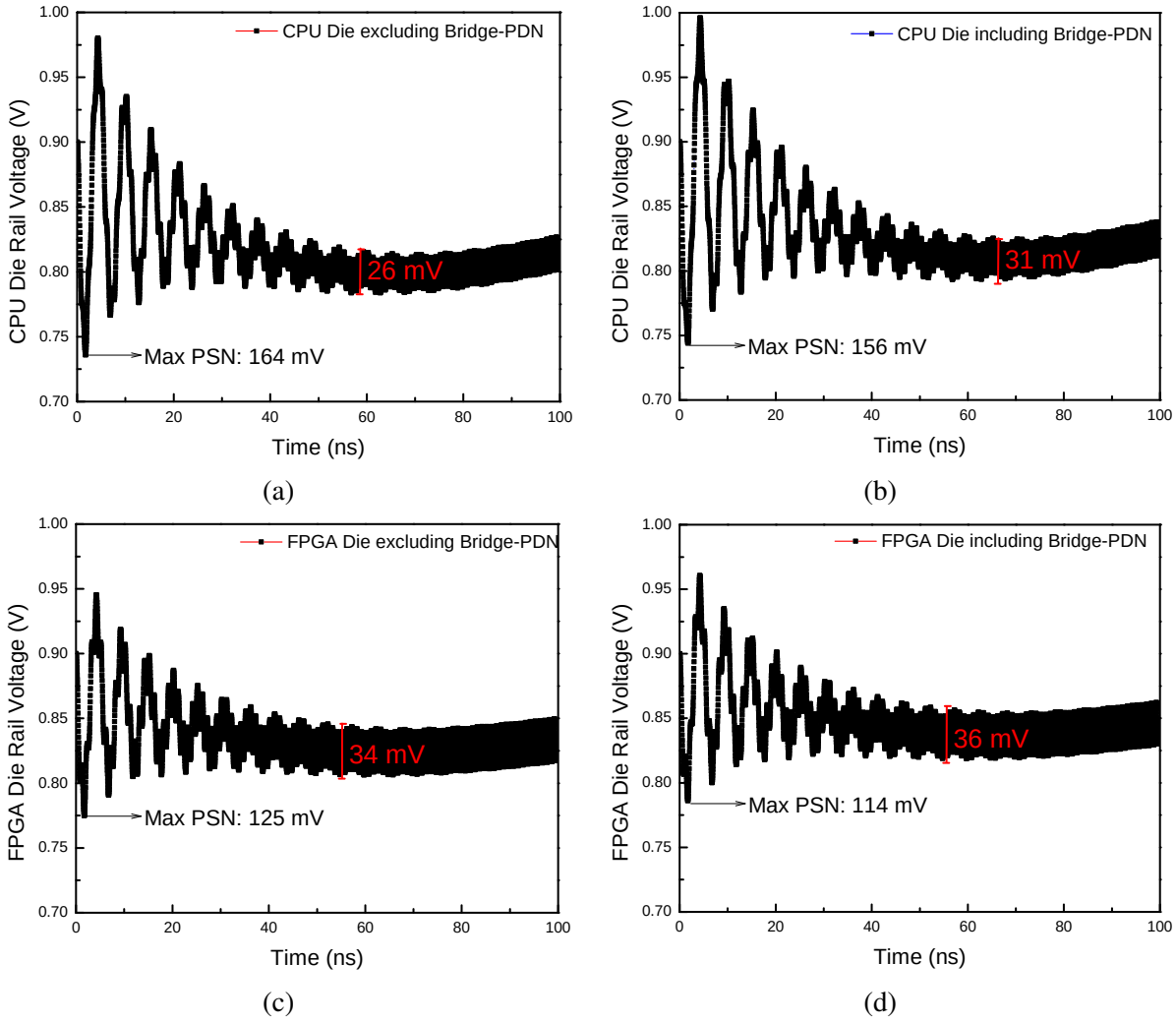


Figure 2.6: Transient analysis results for a 1 GHz pulse on-die excitation for (a) CPU die excluding bridge-chip PDN, (b) CPU die including bridge-chip PDN, (c) FPGA die excluding bridge-chip PDN, and (d) FPGA die including bridge-chip PDN [103].

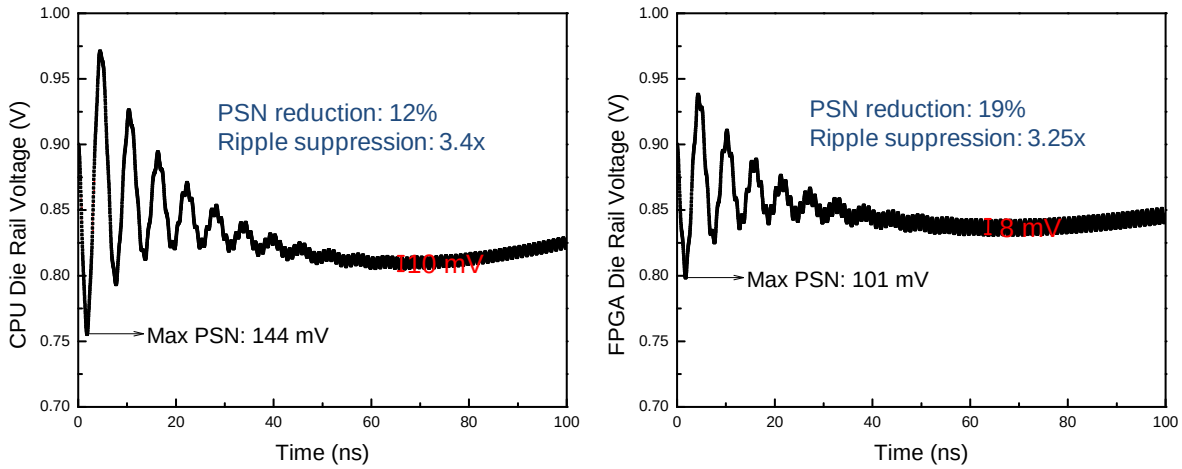


Figure 2.7: Transient analysis results including metal-insulator-metal capacitors in the bridge-chip for (a) CPU die and (b) FPGA die [103].

higher IR-drop can make transistors slower [107] which, in turn, can cause timing failures or even functional failures in critical paths. A higher IR-drop leads to a lower switching voltage in digital circuits, potentially leading to incorrect logic output levels. Moreover, power supply noise (PSN) can also introduce clock jitter in the system [108]. These effects can be potentially exacerbated with bridge-based package architectures due to the inaccessibility of on-die peripheral devices to the package PDN. For these reasons, it is critical to investigate techniques to manage the absolute IR-drop in multi-die 2.5-D packages.

For steady-state supply noise (IR-drop) analysis, we observe that without a PDN in the bridge-chip, the IR-drop in the FPGA and the CPU dice are 86 mV and 99 mV, respectively [103]. Including the ground network in the bridge-chip leads to a 10% and 8% IR-drop reduction for the FPGA die and the CPU, respectively. Including both the power and ground networks in the bridge-chip can further enhance this reduction. The IR-drop improves by 23% for the FPGA die and 17% for the CPU die, compared to the case without PDN in bridge-chip. The bridge-chip PDN appears as a parallel resistive network to the PDN for the on-die peripheral circuits, which lowers the IR-drop for each die.

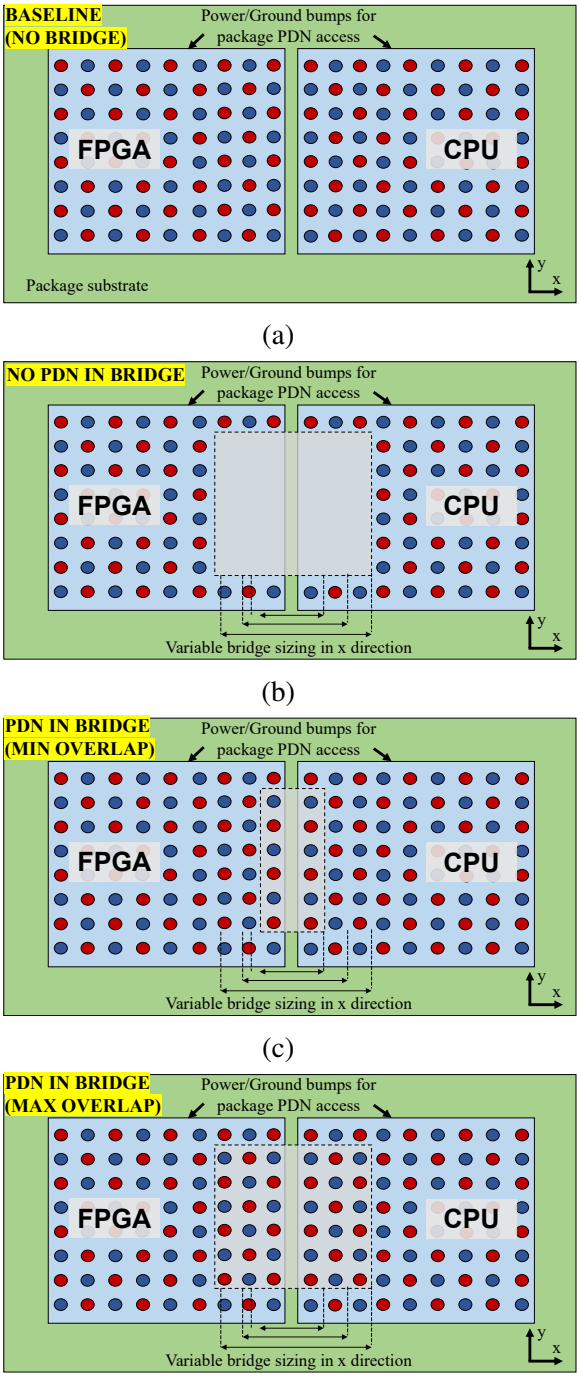
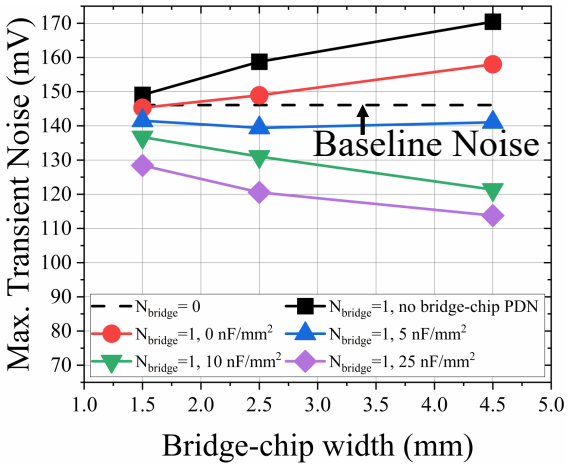
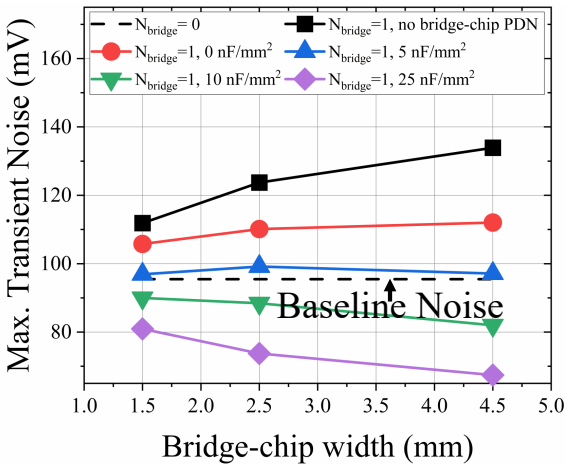


Figure 2.8: Schematic for the different configurations considered of a bridge-chip PDN with varying bridge-chip width along the x-axis as shown.



(a) CPU



(b) FPGA

Figure 2.9: Maximum transient noise for (a) CPU and (b) FPGA as a function of bridge-chip width for different bridge-chip PDN configurations. (N_{bridge} = number of bridge-chips in package)

Transient-state analyses were performed for $L\frac{di}{dt}$-based supply noise estimation for the considered configurations. A 1 GHz on-die stimulus was used for this analysis [103]. All nodes on-die are assumed to be switching simultaneously without any decoupling capacitor in the bridge, as a representative worst-case scenario. The results are shown in Figure 2.6. The first droop is a consequence of the resonance caused by the interaction between the on-die capacitance and the total package loop inductance. Including the bridge-chip PDN (without MIM capacitors) does not impact the loop inductance and the on-die capacitance since the bridge-chip PDN is in parallel to the on-die PDN. Thus, the impact on transient (first droop) noise after inclusion of both power and ground network (Figure 2.5b) in bridge-chip is marginal (5% for FPGA and 9% for CPU) compared to the case without power and ground included in the bridge-chip. This impact on the first droop noise was less significant compared to that observed in the steady-state IR-drop study since on-die resistance is impacted more significantly than loop inductance with the bridge-chip PDN inclusion. Additionally, the resistance of the on-die network and of other components in the loop dampens the voltage droop and any subsequent ringing. Since including the bridge-chip PDN reduces the effective resistance of the on-die PDN (as described previously), an increase in the high-frequency ripple (shown in red in Figure 2.6) is observed across the noise profile (19% for CPU and $\sim$6% for FPGA).

2.3.2 Decoupling (MIM) capacitors in the bridge-chip and impact of bridge-chip sizing

The first droop noise can be reduced by adding more on-die decoupling capacitors. However, area constraints can limit the amount of decoupling capacitance available on-die. MIM decoupling capacitors could potentially be fabricated between the power and ground metal layers in the bridge-chip, if a PDN is included in the bridge-chip (Figure 2.5c). For first droop noise and high-frequency ripple analysis, a modest decoupling capacitor density of 5 nF/mm^2 is assumed in the bridge-chip, with results shown in Figure 2.7. The transient (first droop) noise is reduced by 19% for the FPGA and 12% for the CPU compared to

the case without power and ground included in the bridge-chip. An improvement of 11.4% and a 7.6% for the FPGA and CPU die, respectively, is observed compared to the case of no MIM capacitors in the bridge-chip PDN. Additionally, the high-frequency ripple for both dice is 3× lower compared to the other cases presented in Figure 2.6.

The limited access to package PDN for die peripherals in the overlap region between bridge-chip and dice can lead to significant supply noise challenges. We explore the impact of bridge-chip sizing and bridge-chip PDN configuration on the maximum transient supply noise. In this study, we vary the bridge-chip width (i.e., along x-axis) from 1.5 *mm* to 4.5 *mm* while keeping a fixed length (y-axis) of 6 *mm* [104], as shown in Figure 2.8b, Figure 2.8c and Figure 2.8d. Essentially, we are changing the bridge-die overlap length, as shown in Figure 2.8b. Multiple configurations were assumed for bridge-chip PDN, including no PDN in bridge-chip, PDN in bridge-chip with no MIM caps, and PDN in bridge-chip with MIM caps. The baseline case for this study is a two-die package assembly that does not contain a bridge-chip in the package (i.e. N_{bridge} = 0) but, instead, includes microbumps across the package area to connect the dice to the package PDN (Figure 2.8a). The results are summarized in Figure 2.9a and Figure 2.9b, which show the results for the CPU and FPGA dice, respectively.

As seen in Figure 2.9a, the addition of a bridge-chip without bridge-chip PDN (square curve) leads to an increase in transient noise compared to the baseline (dashed line). The noise increases from ∼146 *mV* (baseline) to ∼149 *mV* at a bridge-chip width of 1.5mm because the access to package PDN is blocked by the bridge-chip (Figure 2.8b). As there is no bridge-chip PDN in this case, increasing the bridge-chip size (and thus the bridge-die overlap) leads to an increase in maximum noise due to a reduction in the number of bumps that offer direct package PDN access. Next, when a PDN is included in the bridge-chip without any MIM capacitors (circle curve), there is a marginal reduction in noise at 1.5 *mm* bridge-chip width compared to the 'no bridge-chip PDN' case. This noise reduces further with increasing bridge-chip width (Figure 2.8c and Figure 2.8d) because more power

and ground bumps are introduced for power delivery with a wider bridge leading to lower current per bump, and thus lower worst case noise. The next three configurations involve adding MIM capacitors to the bridge-chip PDN with densities of 5, 10 and 25 nF/mm^2 [109] distributed uniformly across the bridge-chip PDN. With a fixed MIM decoupling capacitor density across the bridge-chip PDN, increasing the bridge-chip size leads to a higher overall decap, which can reduce high-frequency noise. However, increasing bridge-chip size also leads to fewer direct access bumps between die and package PDN, thus there is a trade-off between bridge-chip size and MIM capacitor density, which is evaluated here. With a decap density of 5 nF/mm^2 (upwards triangle curve), it is observed that the reduction in noise with increasing bridge-chip width is not significant. However, increasing the MIM capacitor density to *10* nF/mm^2 and above, a greater reduction in noise is observed (downwards triangle curve) and the trend changes to lower noise with larger bridge-die overlap compared to previous curves. An MIM density of *10* nF/mm^2 and a bridge-chip width of 4.5 *mm* achieves a similar noise as that with 25 nF/mm^2 and 2.5 *mm* bridge-chip width (Figure 2.9a). For this case study, an MIM capacitor of 10 nF/mm^2 or greater can offer a significant reduction in maximum transient noise compared to 'no bridge-chip PDN' case. This noise can be reduced from $\sim$19% of VDD to $\sim$13.5% of VDD for the CPU ($\sim$15% of VDD to $\sim$9.1% of VDD for FPGA) by including a PDN in the bridge-chip with a MIM decoupling capacitor density of 10 nF/mm^2 and a bridge-chip width of 4.5 *mm*.

2.4 Related Work and Discussion

Interconnect platforms such as interposers and bridges are typically used to provide high-density, low-latency communication between two dies, such as a GPU and high-bandwidth memory (HBM). Including PDN along with signals in the bridge could create contention in the re-distribution layers (RDL) between power and signals. A potential way to get around this could be to increase the number of RDLs (EMIB uses four RDLs) to a higher value. This could incur additional complexity in fabrication of the bridge-chip and added cost.

For reference, deep trench capacitors (DTC) were integrated in a silicon interposer in [100] to improve the PSN of an HBM2E PHY by 62% using a DTC density of 300 nF/mm^2. In [101] the authors integrate MIM capacitors in a silicon interposer with a density of 17 nF/mm^2. Another work [110] has explored using wafer-on-wafer 3D stacking to closely couple DTC to an AI processor for 40% higher operating frequency and 16% less energy compared to its previous generation. In [111], a substrate-embedded bridge-chip was used with only ground nets that appear in parallel to the package surface metal layer, and not directly to the on-die PDN as was assumed in our work. It was shown in [111] that the loop inductance after including a bridge-chip was up to three times lower than without including the bridge-chip. All of the above mentioned techniques can potentially help in managing the transient first droop. A summary of some of the related work is presented in Figure 2.10.

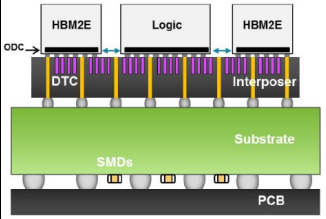
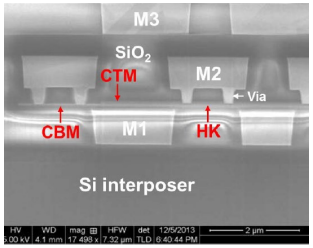
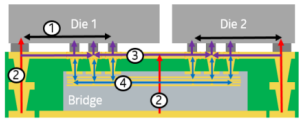
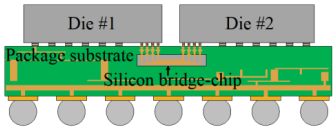
Attribute	Chen et al., 2020 [100]	Liao et al., 2014 [101]	Mahajan et al., 2019 [111]	This work
Schematic				
Interconnection method	**Si-Interposer**: µ-bump + TSVs	**Si-Interposer**: µ-bump + TSVs	**Localized Si-bridge**	**Localized bridge**
PDN in bridge/interposer	TSV and RDL in Interposer	TSV and RDL in Interposer	Only Ground Nets in bridge	2-layer P/G Network in bridge
Decoupling Capacitor in bridge/interposer/on-die	Deep Trench (**Interposer**), MiM (On-die)	High-K MiM (**interposer**)	No	MiM (**bridge**)
Decap. Density	300 nF/mm^2 (DTC)	17 nF/mm^2	N/A	Up to 25 nF/mm^2
PDN Performance	72% lower 1st droop with DTC	Not demonstrated	3 × lower loop inductance with G in bridge	23% lower IR-drop, 19% lower 1st droop, > 3× lower HF ripple
Measurement/ Simulation	Simulation	Measurement	Measurement	Simulation

Figure 2.10: A summary of the salient features of related work in literature.

2.5 Conclusion

We demonstrate that including a PDN in the bridge-chip can provide significant reduction in DC-IR-drop, *Ldi/dt* noise, and high-frequency ripple compared to the baseline case of no PDN in the bridge-chip. Key takeaways are that 2.5-D designs with both smaller-width and larger-width bridge-chips can benefit from decoupling capacitors placed closer to the on-die PDN and that there is a trade-off between the bridge-chip size and MIM capacitor density. We quantify the impact of bridge-chip size and decoupling capacitor density in the bridge-chip on the maximum transient noise. Through a bridge-chip PDN design space exploration, insights are provided which can be useful for 2.5-D design convergence.

CHAPTER 3

CO-OPTIMIZATION FOR ROBUST POWER DELIVERY NETWORK DESIGN IN 3D-HETEROGENEOUS INTEGRATION FOR COMPUTE IN-MEMORY

3.1 Introduction

To address the power delivery challenges described in Chapter 1 of this thesis, we present a systematic technology and design space exploration of power delivery for 3D-HI CIM systems. The key contributions of this chapter include:

- A fast analysis flow facilitating early design-space exploration between power delivery design parameters and CIM performance metrics.
- By co-optimizing 3D PDN and successive-approximation-register (SAR) analog-to-digital converter (ADC) design parameters, we present a balanced 3D CIM design that achieves $\approx 2\times$ inference accuracy compared to an unoptimized 3D implementation at iso-power and iso-area. Trade-offs for such an approach are discussed.

Table 3.1: Power-performance trade-offs between H3D and 2D for CIM

Metric	**2D-LL**	**3D-LL**	**3D-PP**
Area (mm^2)	115.1	3.7	56.2
Throughput (TOPS)	1.4	1.9	1226.5
TOPS/W	7.9	12.9	12.2
TOPS/mm^2	0.01	0.5	21.8
W/mm^2	0.0015	0.04	1.8

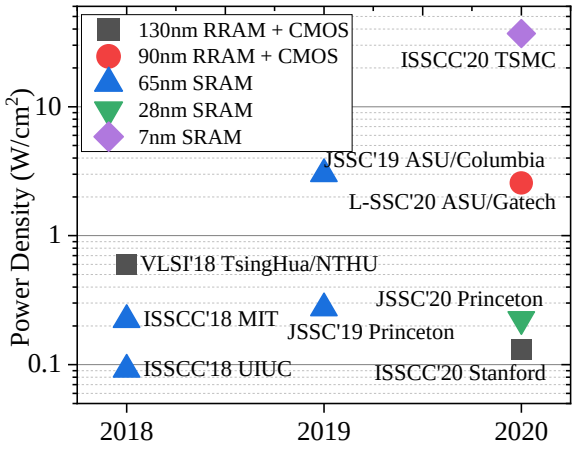


Figure 3.1: Power density trend of recent CIM hardware accelerators.

3.2 3D vs 2D trade-offs for CIM

3.2.1 3D-HI CIM Integration

To compare performance and area trade-offs between 2D and 3D for CIM, a TSV-based 3D-integrated model of an analog CIM accelerator was evaluated and the 3D design of CIM accelerators (7nm logic and 22nm RRAM memory) was applied to a VGG-8 model [13] that was trained to use 8-bit inputs and weights [112] for inference on CIFAR-10 dataset. From an architectural standpoint, two options were considered [113]:

1. A layer-by-layer (LL) system: One logic tier on the package substrate and multiple memory tiers stacked on top (as a memory cube). This design consumes low power but yields high compute latency (Figure 3.2b), and

2. A pipelined (PP) system: 3D interleaved logic and memory tiers (Figure 3.2c). This design offers high speed but consumes high power.

Figure 3.2 shows the baseline 2D (Figure 3.2a), 3D-LL (Figure 3.2b) and 3D-PP (Figure 3.2c) architecture configurations, with logic in blue and memory in green colours. A comparison of 2D baseline, 3D-LL and 3D-PP designs, with this model using 1μm diameter TSVs, is shown in Table 3.1. The performance projection is conducted with 3D

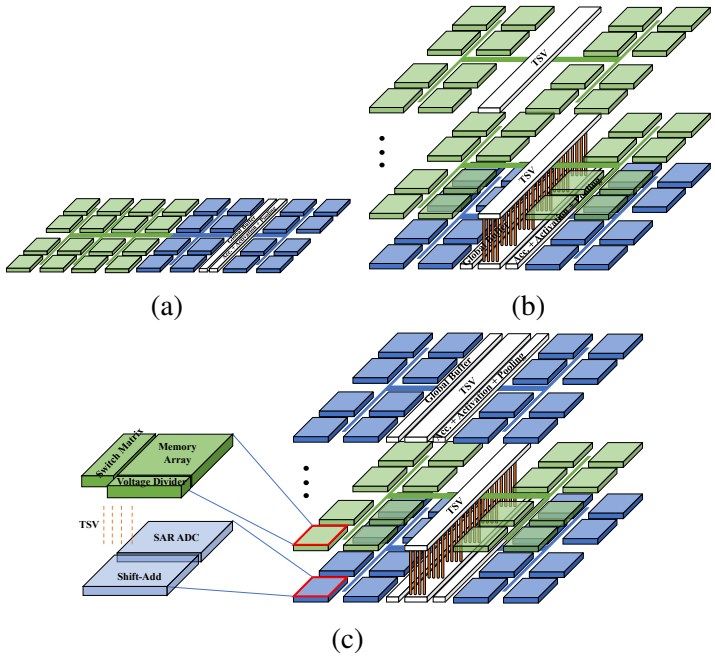


Figure 3.2: Considered (a) 2D, (b) 3D-layer by layer (3D-LL) and (c) 3D-pipelined (3D-PP) CIM architecture configurations.

NeuroSim simulator that is capable of modeling the 3D stacked CIM accelerators [113]. For 3D, memory blocks (RRAM sub-arrays and switch matrices) are assumed at 22 nm (according to the availability from industry, e.g., TSMC and Intel's latest RRAM processes) and logic (ADCs and peripherals) is assumed at 7nm to leverage the scaling benefits. For 2D, we scaled logic area from 7nm back to 22nm to consider both logic and memory at 22nm ($\approx$8$\times$ area inverse scaling). An LL architecture was assumed for the 2D design (2D-LL), as a PP design would require a large number of on-chip memory sub-arrays (for weight duplication), peripheral circuits, and buffers (to serve different DNN layers independently), which leads to a prohibitively large 2D-PP area. The total number of operations (total computations needed architecturally to complete an inference workload) were assumed to be the same for all 2D and 3D designs. The energy efficiency (TOPS/W) is observed to be 63% higher with a 3D-LL implementation vs 2D-LL. The operation-density (TOPS/mm^2) is 50$\times$ and $>$2000$\times$ higher with 3D-LL and 3D-PP, respectively, vs 2D-LL. However, with similar TOPS and lower area this leads to a higher power and power density in both 3D-LL

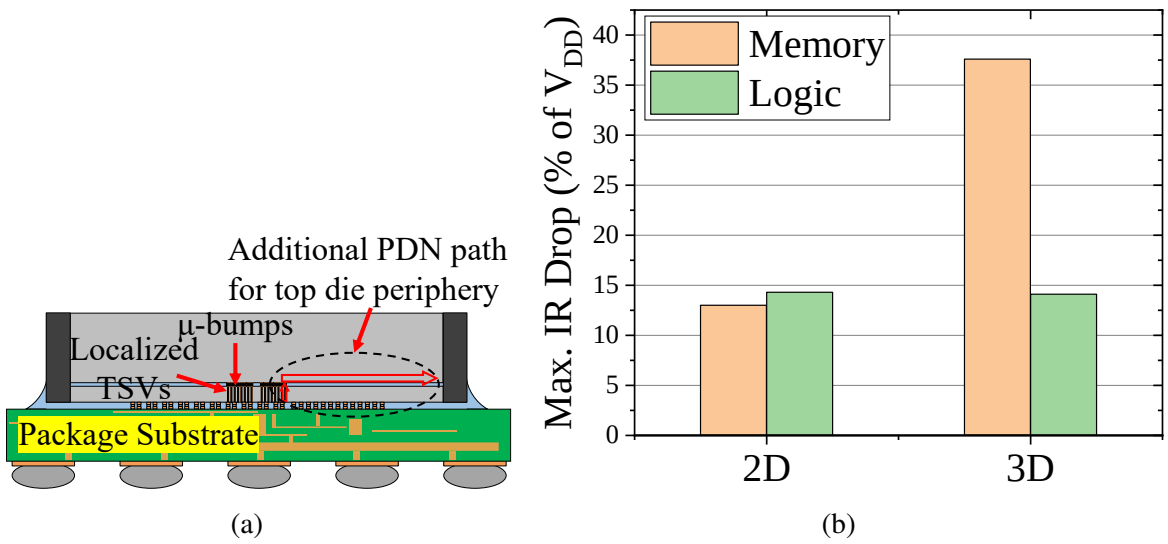


Figure 3.3: (a) Cross-section view of the power delivery network of a 2-tier 3D stack. (b) IR-drop differences between baseline 2D and 2-tier TSV-based 3D design.

and 3D-PP compared to 2D-LL.

The CIM weights need to be duplicated in a 3D-PP design to synchronize timing between different sized convolutional layers, which corresponds to more buffers and ADCs for a PP design. Due to this, the 3D-PP design has a larger total power and logic area than 3D-LL. Due to these additional resources, the 3D-PP design also requires larger on-chip interconnect length than 3D-LL leading to added energy and latency. 3D-PP also experiences higher leakage than 3D-LL due to higher power dissipation. For these reasons, 3D-PP has a marginally lower energy efficiency (TOPS/W) than 3D-LL. Nevertheless, as summarized in Table 4.1, both 3D designs offer higher throughput (TOPS), performance-per-Watt (TOPS/W), and operation density (TOPS/mm^2), and lower footprint, compared to the 2D baseline. Next we look at the power delivery challenges specific to 3D-HI compared to 2D baseline.

3.2.2 Power Delivery Challenges in 3D-HI

While the power delivery challenges mentioned in Section 1 exist for both monolithic 2D and multi-die systems, they are exacerbated in 3D ICs. Figure 3.3 presents the comparison

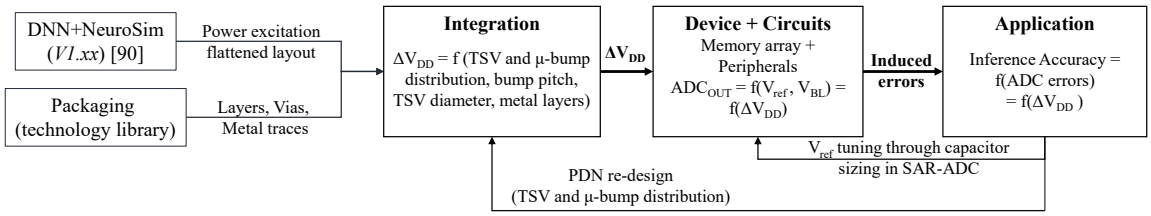
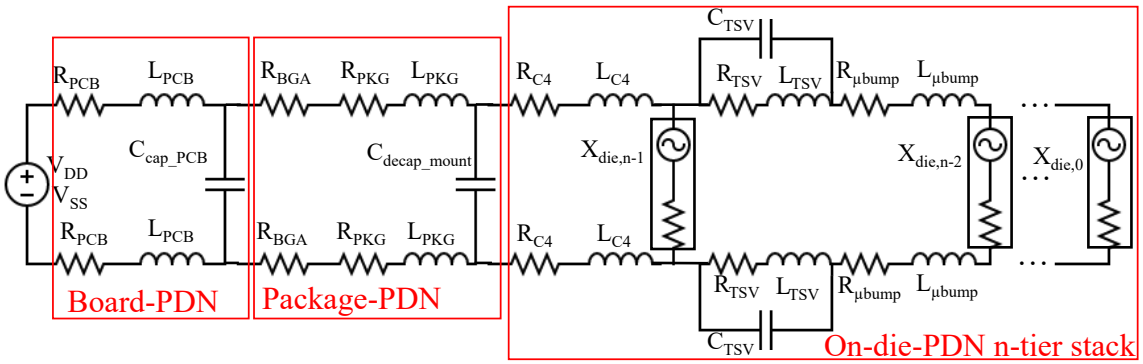


Figure 3.4: 3D CIM PSN evaluation and co-design methodology from device/integration towards application-level. (Note: ΔV_{DD} = Variation in supply voltage (mV); ADC_{OUT} = digital ADC output; V_{ref} = ADC reference voltage (V).

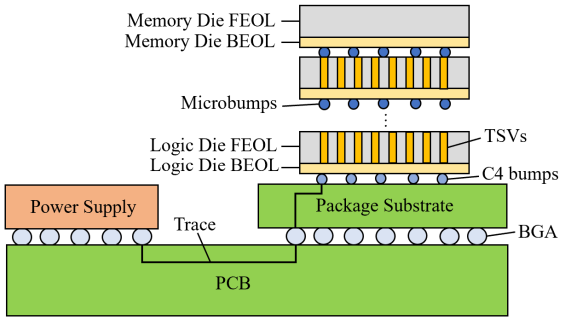
of IR-drop as a percentage of supply voltage for both 2D and 3D ICs. Both designs were implemented using 22nm logic (CMOS) and 22nm memory (RRAM) with the PDN modeling methodology described in section 3.1. The total power for both 2D and 3D was 218W (214.2W-logic, 3.8W-memory) and the only difference between the two designs is that the 3D case is a memory-on-logic face-to-back configuration that uses 1 μm diameter TSVs and 200 μm pitch microbumps for power delivery to top tier (memory) (Figure 5.3a). The maximum IR-drop in the memory tier in 3D was $>2.5\times$ compared to 2D (37% vs 13%) (Figure 5.3b). This maximum IR-drop is seen in the corner regions of the memory die and is mainly due to the additional resistive on-die PDN of the memory die. Noise for the logic tier is similar in 2D and 3D since logic is assumed to be on the bottom tier in 3D and its PSN is not affected significantly by the low-power memory tier on top. The scope of this work is to co-design PDN and ADC design parameters to reduce ADC errors and optimize CIM inference accuracy, and is reported in the following sections.

3.3 3D CIM PSN Evaluation Methodology from Device/Integration towards Application-level

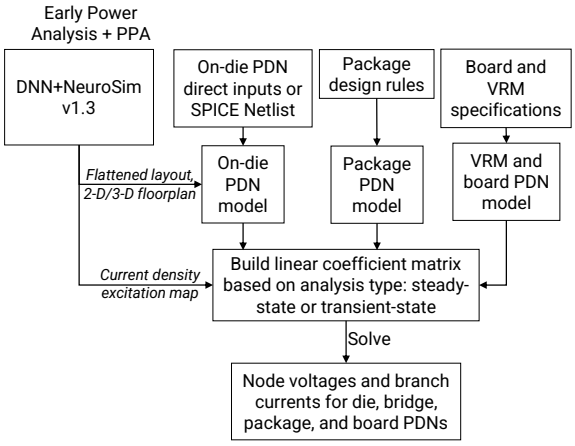
Figure 4.3 illustrates the proposed methodology to quantify the impact of 3D-HI PDN design parameters on ADC output errors and to co-design PDN and ADC parameters to maximize CIM inference accuracy. This flow combines a finite-difference method-based PSN analysis framework with a CIM inference accuracy estimation framework and the



(a)



(b)



(c)

Figure 3.5: (a) PDN modeling hierarchy. From left to right: lumped model of the board-level PDN, distributed model of the package-level PDN, and the distributed model of the on-chip PDNs in an n-tier 3D stack including the TSVs and microbumps. (b) Cross-section view of the power delivery network of a n-tier 3D stack. (c) Flow diagram of the 3D PDN analysis showing different steps of the framework.

details of each flow component are described as follows:

3.3.1 3D PDN modeling methodology

Figure 5.3a shows the PDN structure of an n-tier 3D stack with a cross-section in Figure 5.3b. We implement a distributed package-level PDN model, unlike previous works that assume a lumped package model, to reflect the spreading effects of current in the package and the coupling between different P/G bumps, especially when dice share package PDN planes. Figure 3.5c presents the flow for both steady-state and transient analyses and this is an updated version of the flow presented in [114]. The flow begins with the generation of the RLC network models of the board, package, and the on-die PDNs. These models are subsequently combined to solve for nodal voltages and branch currents. The key contributions in this updated flow include: a) support for modeling of 3-D packaging architectures, and b) support for interface with open source tools such as DNN+NeuroSimV1.xx (a pre-RTL simulator [90]).

An ideal voltage regulator module (VRM) is assumed for board-level PDN and a lumped R/L network is used to model board-level current spreading. The equivalent series resistance and equivalent series inductance of board-level decoupling capacitors are included in the model. The package power/ground planes are modeled as two layers with the bottom layer connected to the motherboard by ball grid arrays and the top layer connected to an on-die PDN through C4 bumps. Each node in the two layers is connected to six adjacent nodes through an R-L pair, representing either package traces or inter-layer vias. Die-side decaps are assumed to be connected to the top package PDN layer.

The on-die PDN consists of several metal layers where the P/G wires are parallel to each other within a layer, and adjacent layers are orthogonal to each other. To better reflect the interleaved nature of the on-die PDN and capture the effect of on-die vias, the on-die PDN is modeled as a two-layer structure. The metal wires on each on-die PDN are typically uniformly distributed, but if the actual layout is non-uniform, our flow calculates

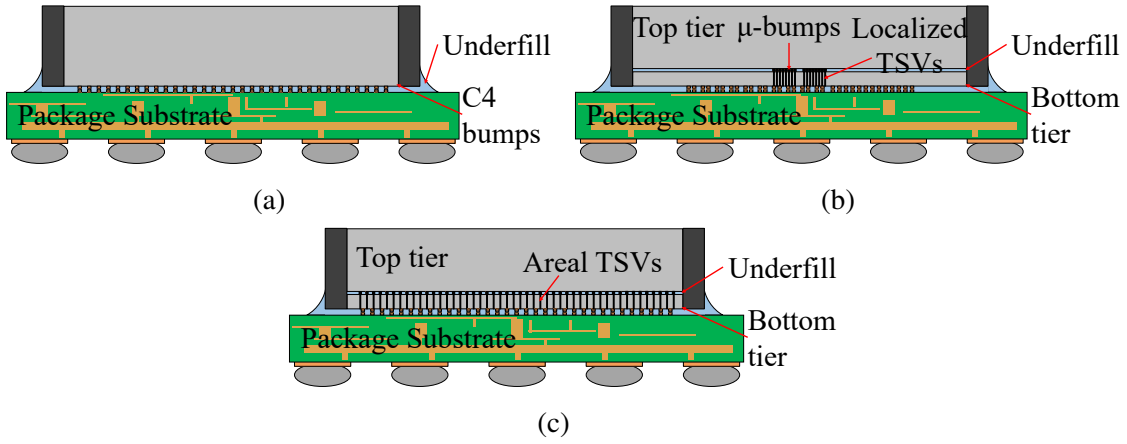


Figure 3.6: Cross-section of the considered integration architectures for PDN evaluation. (a) Monolithic 2D (baseline). TSV and microbumps based 2-tier 3D with (b) localized and (c) areal TSV distribution.

the effective wire pitch and via density to reorganize the PDN layout. For each on-die layer, we map fine-granularity PDN layout to coarse mesh grids at a C4 bump granularity. The equivalent parallel resistance is calculated, for each coarse grid containing multiple vias and metal wires, and assigned using models described in [115]. All coarse PDN layers with x-axis and y-axis metal wires are mapped onto the top and bottom layers, respectively. R_{via} are the effective resistances of vias between adjacent metal layers, and R_{TSV} is the resistance of TSVs between multiple 3D dice. Although this framework can be used to model both steady-state and transient PSN, in this work we focus on steady-state analysis.

3.3.2 Inference Accuracy Estimation

To estimate the impact of steady-state PSN on inference accuracy, we simulated the inference operation of VGG-8 for each 3D-HI design and the 2D baseline. Simulations were performed in Pytorch [117] where weights of VGG-8 were mapped to a grid of memory blocks containing the RRAM devices and necessary peripheral circuits following the method outlined in [90]. 8-bit weights are represented by a group of 8 RRAM cells, where a ”1” is represented by the low-resistance state (LRS) and a ”0” is represented by the high-resistance state (HRS). Inputs are binarized and applied to the gates of the access transistors

Table 3.2: Experimental Setup

PDN Model parameters	
Parameter	Value
On-die metal resistivity (ohm-m)	1.8e-8
On-die global wire Pitch/Width/Thickness (um)	39.5/17.5/7
On-die intermediate wire P/W/T (nm)	560/280/506
On-die local wire P/W/T (nm)	160/80/144
On-die decap density (nF/mm2)	335
microbumps pitch/R/L (um/m-ohm/pH)	40/30.9/11.1
C4 bump pitch/R/L (um/m-ohm/pH)	200/14.3/11
Package effective decap R/L/C (m-ohm/pH/uF)	541.5/220.7/52
Package resistivity/inductance (*m-ohm/mm*/ pH/mm)	1.2/24
BGA pitch/R/L (um/m-ohm/pH)	500/38/46
TSV R/L (m-ohm/pH)	54.2/77.78
PCB R/L (u-phm/pH)	166/21
PCB decap R/L/C (u-phm/nH/uF)	166/19.54/240
Power (W)	218.01 (2D) (Logic: 214.25, Mem: 3.76) 118.61 (3D) (Logic: 114.85, Mem: 3.76)
Device and Circuit parameters	
RRAM model	[116]
bits/cell	1
On/off ratio	$20.7\text{k}\Omega/100\text{k}\Omega$
RRAM array size	128x128
Total number of arrays	23,048
Total chip memory capacity	45MB
TOPS/W	8
Supply voltage to both tiers	0.9V
SAR-ADC precision	4b
Neural network	VGG-8
Precision I/W	8b/8b
Network size	12.7MB
Convolutional/dense layers	6/2
Duplication amount (layer 1 - layer 8)	64/64/16/16/4/4/1/1

in each row, connecting the RRAM to the grounded select line (SL). Resulting voltages on the bitline (BL) are calculated by the voltage division between the pull-up PMOS and the equivalent resistance of the RRAMs connected in parallel. Calculated BL voltages are then used in the simulation of the analog-to-digital (ADC) conversion in the logic tier. See Figure 3.7 for a visual depiction of a memory block and its corresponding logic block. Digital ADC outputs are then shifted and added to construct the final output of each layer. Outputs from each layer are then sent as inputs to following layers allowing for a holistic simulation of the inference operation on a CIM system.

In the RRAM model used in these simulations [116], the on-state resistance is 20.7 kΩ and the off-state resistance is 100 kΩ, resulting in an on/off ratio of 4.83. In order to achieve accuracy comparable to the software baseline ($\approx$90$\%$) without considering IR-drop, the operation of each RRAM array was limited to 3 rows at a time. This comes at a trade-off in operating latency, but it allows the effect of IR-drop to be isolated from other device non-idealities such as the limited on/off ratio. To give an estimate for how a larger scale chip will be impacted by PSN, we duplicated shallow layers of VGG-8 (see Table 3.2). This also increases the throughput of the chip and reduces the latency of inference. After duplication, the total model size is 45 MB. Other relevant parameters are listed in Table 3.2.

Once baseline accuracy was achieved, a contour plot of IR-drop was generated for each 3D-HI design following the PDN analysis framework (see Figure 3.9). The IR-drop is then used in inference accuracy simulations. Because the value of IR-drop to a compute block depends on its physical location on the chip, we injected IR-drop values to each compute block in the inference simulation according to their mapped location on the die. This way the generation of errors due to spatial IR-drop can be simulated.

3.3.3 Experimental Setup

The details of PDN model and device/circuit parameters are summarized in Table 3.2. Architecturally, a baseline 2D monolithic design is partitioned and arranged into a 2-tier

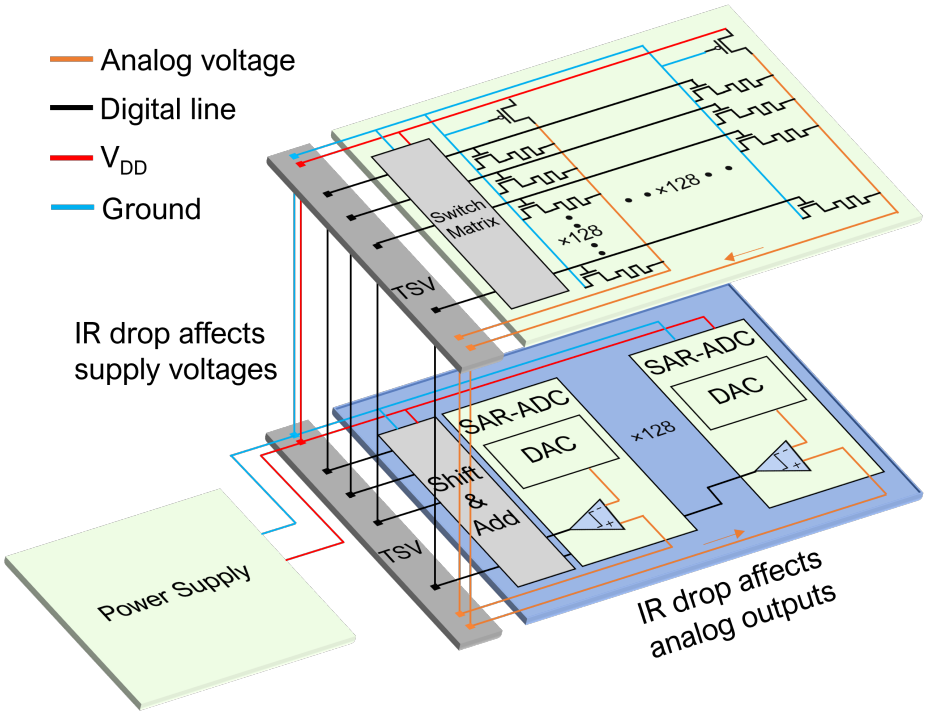


Figure 3.7: Operation of an RRAM array and corresponding ADCs in the 3D-HI design. Analog outputs were calculated in the presence of PSN, the ADC sensing process was simulated, and errors in ADC outputs were evaluated.

3D PP configuration (Figure 3.2a and Figure 3.2c). We consider two TSV and μbump-based 3D-HI architectures (Figure 3.6b, Figure 3.6c) baselined against monolithic 2D (Figure 3.6a). Figure 3.6b represents a localized distribution of TSVs and μbumps (localized-TSV 3D) while Figure 3.6c shows a uniformly spread areal TSV and μbump distribution (areal-TSV 3D). After partitioning the 2D design (i.e. both logic and memory at 22nm), we assumed a two-tier integration with a memory-on-logic approach, with memory at 22nm and logic at 7nm.

The estimated power and network/dataset assumptions are summarized in Table 3.2. The power-per-block and number of memory (switch matrix, memory arrays) and peripheral logic (shift-add, ADC, accumulation+activation+pooling, global buffer) are estimated using NeuroSim V1.3's mapper [90]. The logic and memory blocks are each mapped to separate tiers. The difference in powers between 2D and 3D designs is because 2D (logic and memory) is at 22nm while 3D logic is at 7nm and 3D memory is assumed to be at

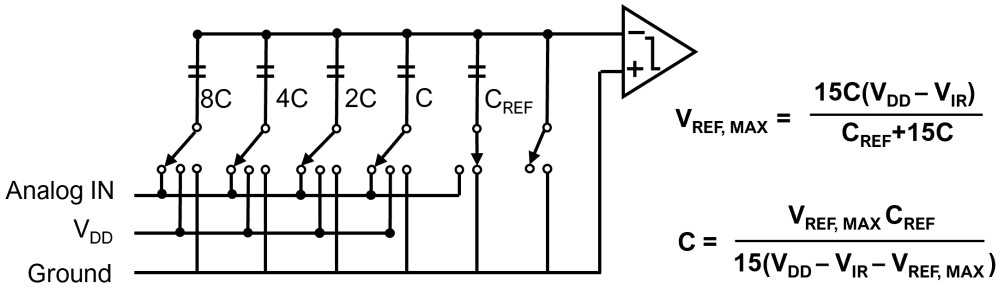


Figure 3.8: Charge phase for the maximum reference voltage in a SAR-ADC by the capacitive DAC. The reference voltages can be tuned to mitigate IR-drop by properly sizing the capacitors. C_{REF} is a reference cap of fixed size.

22nm. For this study, the assumed 3D-HI schemes are shown in Figure 3.6. For TSVs in the bottom tier we assume 1μm diameter, 2 μm pitch and an aspect ratio of 10:1 (Figure 3.6b and Figure 3.6c).

We assume 4-bit SAR-ADCs are used to sense analog outputs at the columns of each array. Because the logic tier is at a more advanced node, we can afford to include 1 SAR-ADC per column in each RRAM array, improving throughput compared to a homogeneous design where ADCs are commonly shared among multiple columns to limit their area overhead.

To account for IR-drop in the inference simulations, the BL voltages were calculated as follows:

$$V_{BL,j} = \frac{R_{eq,j} \cdot (V_{DD} - V_{IR})}{R_{eq,j} + R_{PU}} - V_{IR} \tag{3.1}$$

where j is the column in the memory array being considered, $R_{eq,j}$ is the equivalent resistance of the RRAM, R_{PU} is the resistance of the pull-up PMOS and V_{IR} is the IR-drop to the memory block. The IR-drop from the power supply to the memory block is subtracted from V_{BL} and the IR-drop from the memory block to its corresponding block in the logic tier is subtracted from the analog output voltage. Once reaching the logic block, the BL voltages are quantized through comparison to reference voltages generated by the SAR-ADCs. IR-drop to the logic tier affects the generation of these references (see Figure 3.8). Noise from IR-drop to both tiers is realized as errors in the digital outputs of the ADCs.

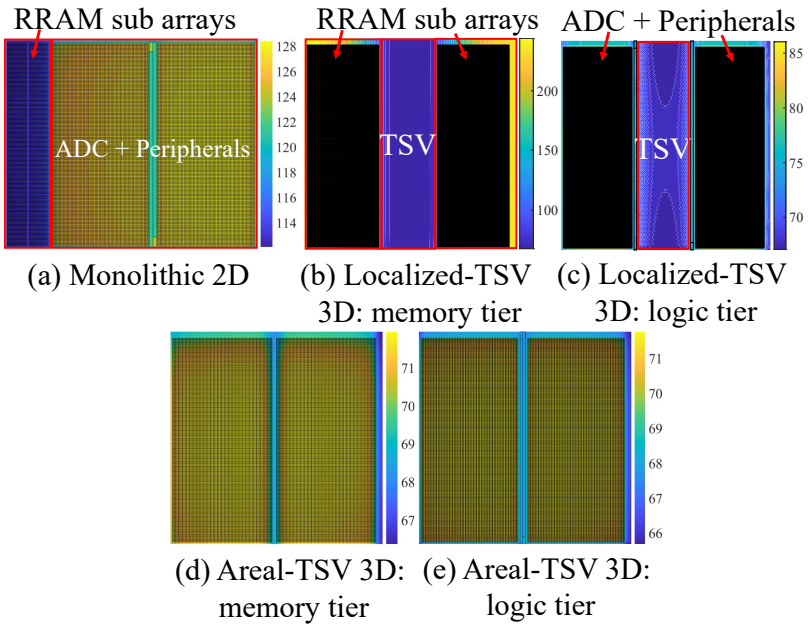


Figure 3.9: IR-drop contours for (a) baseline 2D, 2-tier localized-TSV 3D (b) memory and (c) logic tier, and areal-TSV 3D (d) memory and (e) logic tiers (die size not to scale).

We propose a strategy to mitigate the effect of IR-drop on ADC outputs by fine-tuning the capacitive DACs in the SAR-ADCs. The DAC depicted in Figure 3.8 uses a chain of parallel capacitors to generate the reference voltages used to sense analog outputs. By adjusting the size of these caps, the reference voltages can be tuned to account for the IR-drop to both the memory and logic tiers. Because IR-drop is dependant on the location on the die, each ADC can be separately tuned to offset the unique IR-drop in their respective locations. To evaluate the effectiveness of this strategy, we used the IR-drop contours in Figure 3.9 to adjust the cap sizes of each ADC in our inference simulation depending on their mapped location on the die. We note that the calculated cap sizes are smaller than the initial sizes, indicating that this strategy will result in a decrease in area overhead of the ADCs.

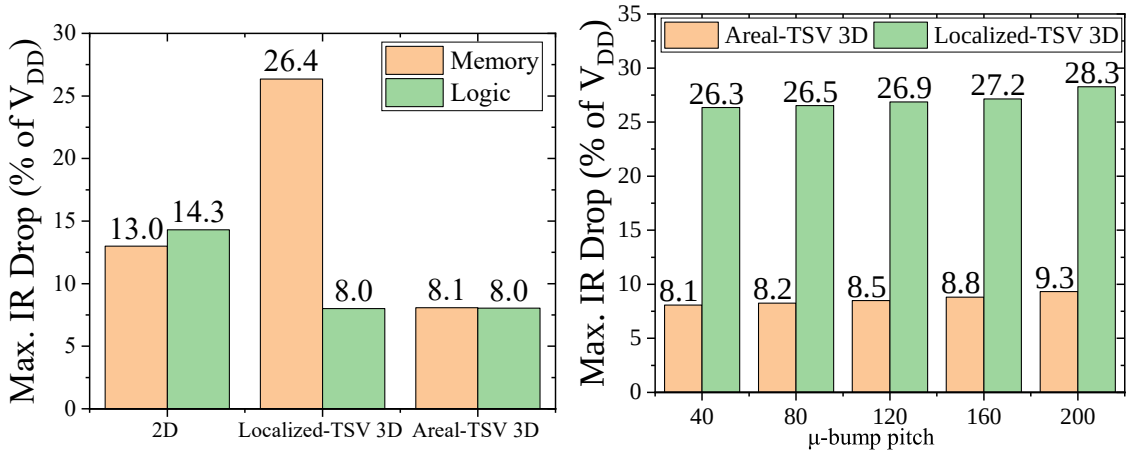


Figure 3.10: (a) Maximum IR-drop for considered configurations. (b) Steady-state IR-drop for 2-tier M-on-L configuration as a function of microbump pitch and TSV distribution.

3.4 Results

3.4.1 PDN design benchmarking: TSV and microbump analysis

Figure 3.9a depicts the IR-drop contours for 2D (logic and memory at 22nm). Figure 3.9b,c show noise contours for a 2-tier 3D with 40um microbump pitch and 1um diameter TSVs localized in the center (localized-TSV 3D). The worst-case noise in the memory tier increases from ≈13% of V_{DD} to 26% between 2D and localized-TSV 3D. To mitigate this, an areal distribution of TSVs and microbumps (Figure 3.6c) is considered (areal-TSV 3D). Figure 3.9d,e present the contours for the areal-TSV 3D. The memory tier IR-drop was reduced to ≈8% of V_{DD} at iso-power. An areal TSV distribution could also have a keep-out-zone area overhead but that is outside the scope of this work.

We observe that for localized-TSV 3D the memory tier IR-drop is minimal at the locations closest to the clustered TSVs and gets worse towards the die edges. The same trend is present in the logic tier, however the absolute noise is lower for 3D logic tier since power is directly delivered to logic through C4 bumps, similar to the 2D baseline. The worst-case IR-drop is summarized in Figure 3.10a.

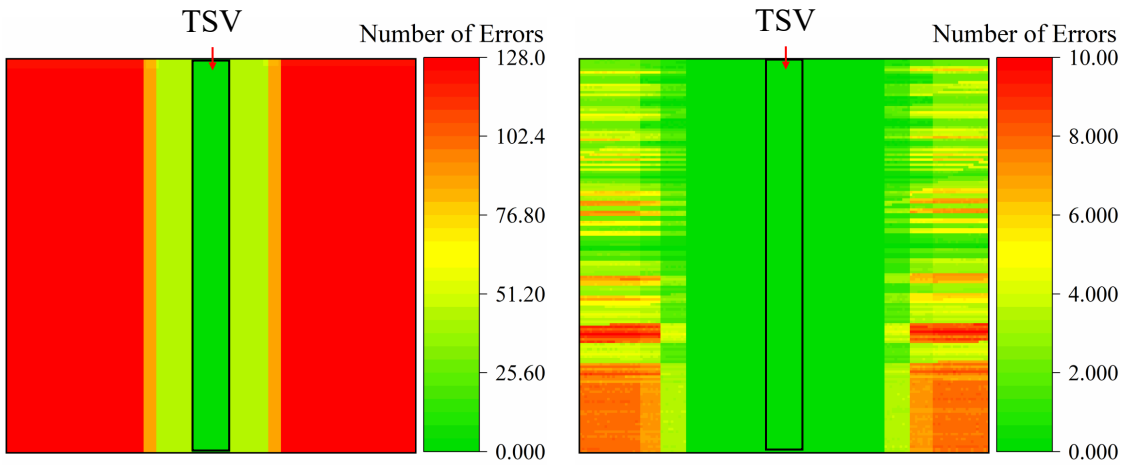
We also explored the impact of microbump pitch on the IR-drop of the considered 3D

designs. Microbump pitch plays an important role in power delivery and with dense microbumps IR-drop will decrease. We considered a range of microbump pitches between 200-40 μm (Figure 3.10b). As expected, with a reduction in microbump pitch (increase in bump density) the IR-drop reduces for both the localised-TSV 3D and areal-TSV 3D designs. However, the areal distribution of TSVs contributes significantly more than microbump pitch in reducing IR-drop. A 3-tier design (logic+memory+logic) was also explored but the IR-drop to the top logic tier was prohibitive with a supply of 0.9V and our power assumptions. We expect a full analysis for a multi-tier 3D design to be part of future work.

3.4.2 Impact of PSN on CIM errors

For each design we simulated the inference operation of VGG-8 with IR-drop to the memory tier and logic tier according to IR-drop contours from the PDN analysis framework. Each design was evaluated with and without tuning the size of the caps in the ADCs.

In Figure 3.11 we provide a map of the ADC errors distributed across the die for the localized-TSV 3D design, verifying that areas with lower IR-drop correspond to lower number of errors in ADC outputs. In this figure, each pixel represents the average number of errors in the ADC outputs. The number of errors is defined as the sum of differences between the simulated output and the ideal output. Because digital outputs range from 0-128, the maximum number of errors per digital output is 128. Noting this, we see in Figure 3.11a that large IR-drop in certain areas of the die lead to every output state being sensed incorrectly. We observe in Figure 3.11b that the cap sizing strategy can remove many of the errors present in the ADCs. However, once the IR-drop in the memory tier surpasses 170mV, or $\approx 19\%$ of V_{DD}, the strategy can no longer mitigate all of the errors. This is because IR-drop reduces the total range of the output voltages from the memory tier, reducing the sense margin of each state. When sense margins are too small, the ADCs cannot sense all outputs correctly regardless of the size the capacitors in the DAC. Error



(a) Localized-TSV 3D without sizing caps (b) Localized-TSV 3D after cap sizing

Figure 3.11: Average number of errors in ADC outputs for the localized-TSV 3D design with and without the cap sizing strategy. 23,048 total memory arrays are mapped 1:1 to blocks of ADCs in the logic tier.

maps for the 2D baseline and areal-TSV 3D case were excluded because the cap sizing strategy removes all errors present in these cases.

In Figure 3.12a we average the number of errors across the entire die for each chip design and the resulting inference accuracy is recorded in Figure 3.12b. We see that with a localized-TSV 3D design and before cap tuning, the errors increase compared to 2D baseline and they can be reduced with an areal-TSV 3D design. However this still leads to a low inference accuracies in all three cases. After cap tuning, the errors can be completely removed in both the 2D and areal-TSV 3D designs, resulting in the recovery of the baseline inference accuracy. Interestingly, despite drastic improvement in errors in the ADC outputs seen in the localized-TSV 3D case after capacitor sizing, the small number of errors that can't be mitigated have a significant impact on the overall inference accuracy, implying that co-optimizing 3D PDN and ADC design parameters can be an important methodology to achieve design robustness.

It is important to note that in all designs, IR-drop will reduce the sense margins for each output state. Therefore, any additional noise in the system will have a greater impact on the number of ADC errors than they would in the absence of IR-drop. In this study we

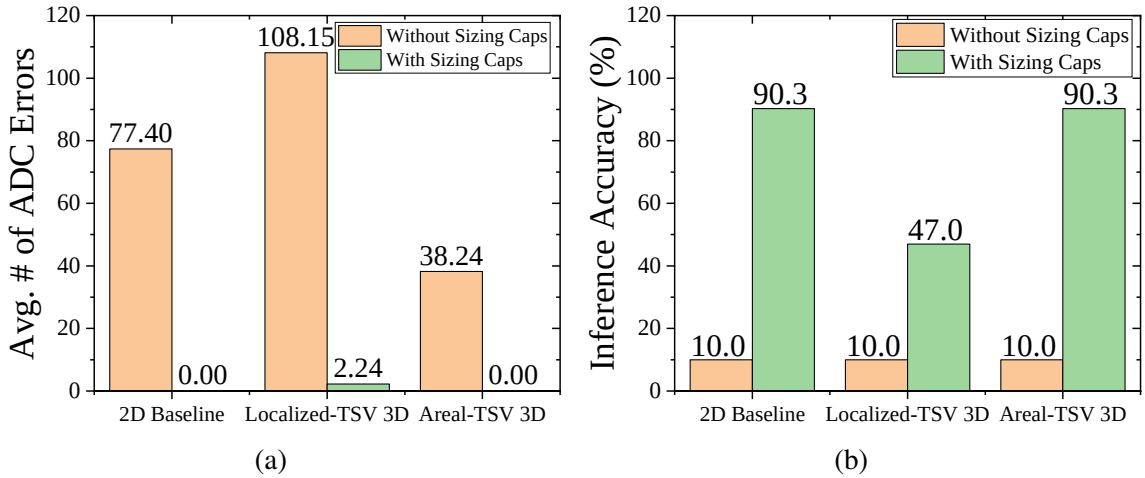


Figure 3.12: (a) Number of errors in ADC outputs averaged over the entire die and (b) corresponding inference accuracy for each design. A strategy of tuning capacitor size in the SAR-ADCs was employed to mitigate errors caused by IR-drop. Simulations were conducted in the absence of other chip non-idealities.

aimed to give a holistic analysis of the IR-drop in 3D-HI CIM systems and its effect on inference accuracy. As such, we did not consider additional non-idealities present in CIM (e.g., thermal noise caused by elevated temperature in 3D-HI). With the frameworks we have developed such an analysis is feasible but is left to be studied in future works.

3.5 Related work

[118] demonstrated a resistive random-access memory (RRAM) based CIM macro with solutions for IR-drop mitigation. They combined a hybrid analog/mixed-signal offset cancellation scheme and $I_{CELL}R_{BLSL}$ drop mitigation with a low cell bias target voltage. Their macro demonstrated robust operation (post-ECC bit error rate (BER) $< 5{\times}10^{-8}$ for 8WL CIM) while maintaining an effective cell density 1.03 – 33.1× higher than prior art and achieving 1.74 – 13.35× improved average MAC efficiency relative to the previous highest-density RRAM CIM macro. In [119], to address the challenges of increased errors in MAC operations in non-volatile memory arrays due to steady-state (IR-drop) and transient noise, the authors propose a sign-weighted 2T2R (SW-2T2R) array to reduce IR-drop by decreasing the accumulative SL current (ISL), and thus, enabling higher parallelism. They

implement a fully-integrated 784-100-10 multi-layer perceptron (MLP) model on an integrated CIM chip with 158.8kb analog ReRAMs. Their chip realizes an accuracy of 94.4% on MNIST database, an inference speed of 77 μs/image, and 78.4 TOPS / W peak energy efficiency, with CMOS circuits fabricated in a 130nm process. [120] presents characterization of the impact of IR-drop and device variation (calibrated with measured data on foundry RRAM) and evaluates different approaches to write verify. Using various voltages and pulse widths, the authors program cells to offset IR-drop and demonstrate a 136.4$\times$ reduction in BER during CIM.

3.6 Conclusion

A comprehensive design-space exploration of power delivery network design for 3D heterogeneously integrated CIM hardware is presented. A device-integration-application evaluation methodology is proposed to facilitate early design-space exploration and trade-offs between power delivery design parameters and CIM performance metrics are quantified. By co-optimizing across design hierarchies from packaging to circuits and devices, we present an areal-TSV 3D CIM design and compare it to a localized-TSV 3D implementation. For our assumed 3D CIM hardware, an areal distribution of through-silicon via (TSV) and microbumps, and a PSN-aware SAR-ADC fine-tuning achieves a 90% inference accuracy compared to 47% with a unoptimized 3D design at iso-area and iso-power. The insights provided could be useful for design convergence and performance modeling for edge intelligent 3D hardware.

CHAPTER 4

3-D HETEROGENEOUS INTEGRATION OF RRAM-BASED COMPUTE-IN-MEMORY: IMPACT OF INTEGRATION PARAMETERS ON INFERENCE ACCURACY

4.1 Introduction

Compute-in-memory (CIM) has been proposed as a potential paradigm for energy-efficient compute through reduced data movement and increased parallelism in deep neural network (DNN) computations as state-of-the-art machine learning model parameters grow exponentially (upto 100's of MB [13]). Emerging nonvolatile memory (eNVM), such as resistive RAM (RRAM), phase change memory (PCM) etc. are potential alternatives to SRAM/DRAM as CIM synaptic devices due to their higher bit density and low leakage, enabling large embedded memory and high energy-efficiency.

Among various heterogeneous integration architectures, such as MCM, 2.5-D, and 3-D, 3-D-HI can provide higher compute density and energy-efficiency through a reduced footprint and interconnection length, respectively, compared to MCM and 2.5-D. A growing need for higher logic-memory bandwidth and lower chip-to-chip signal interconnection delay have led to a technological push towards 3-D-HI such as through-silicon via (TSV)-based 3-D integrated circuits (ICs) [68, 69, 70]. Although 3-D-HI can enable dense memory-logic integration needed for state-of-the-art CIM hardware accelerators using, there are some challenges with dense 3-D integration of eNVM devices such as RRAMs.

Figure 4.1 illustrates the power densities of recent CIM and hardware accelerator monolithic 2D chip demonstrations, and an average increasing trend can be observed. To leverage the benefits of 3-D integration for CIM/accelerator hardware, with such a trend in increasing power densities, thermal effects such as inter-die thermal coupling and increased

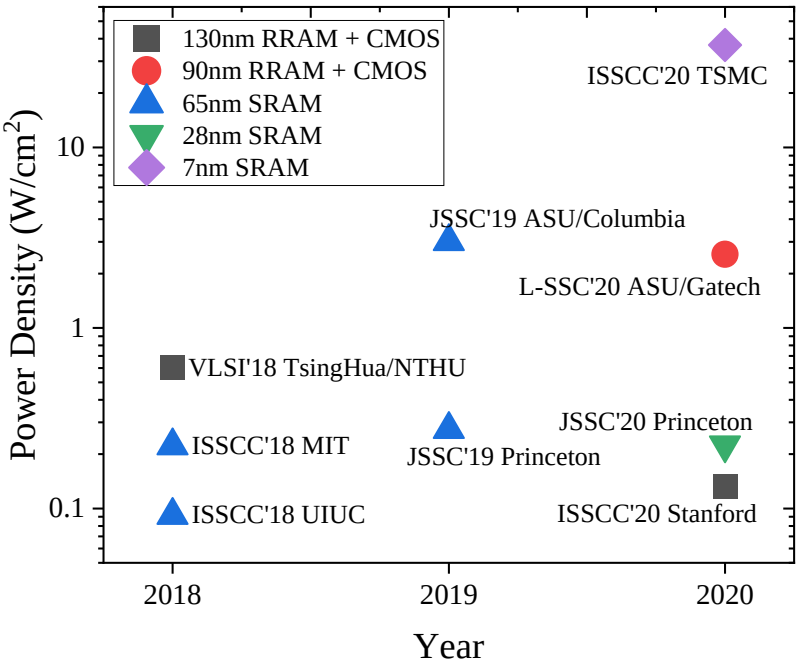


Figure 4.1: Power density trend of recent CIM and hardware accelerators.

number of hotspots could pose a significant challenge in terms of performance and reliability implications. This is because 3-D ICs can experience a significant variation in power densities compared to monolithic 2D and thermal performance may not scale linearly. Additionally, thermal-induced conductance drift remains a challenge in resistive filamentary devices such as RRAMs [121] despite their promising features as CIM synaptic devices. Due to increased volumetric power in 3-D, lower retention at higher temperatures can be more significant in dense memory-logic 3-D integration. Although prior work has investigated device retention degradation due to thermal crosstalk in high-density 3-D integration of RRAMs [122] and there has been extensive characterization of RRAM reliability mostly focused on memory applications [123], there are no previous studies that consider the impact of integration design parameters and device reliability together on system-level performance metrics for CIM applications (such as CIM inference accuracy). Shim et al. [124] performed statistical measurement and modeling of retention characteristics of multilevel RRAM-based synaptic arrays at different temperatures. They measured average conductance of a 2-bit per cell 1T1R HfO_2 based RRAM test chip, as a function of bak-

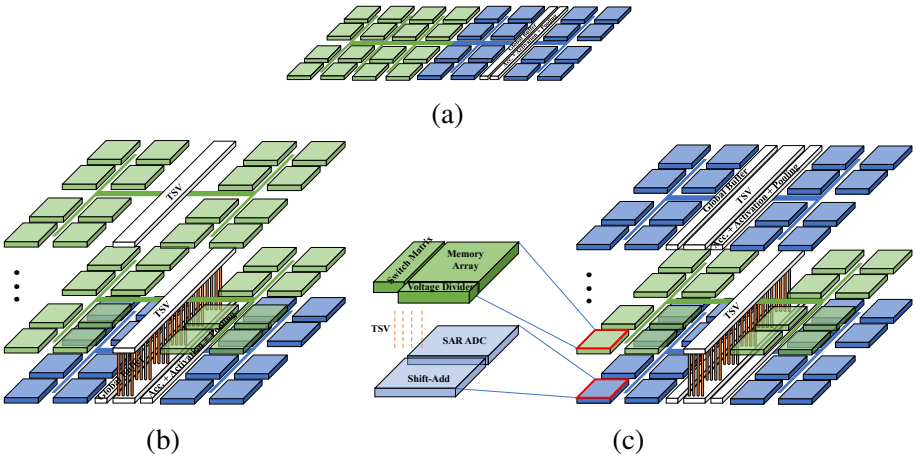


Figure 4.2: Considered (a) 2D, (b) 3-D-layer by layer (3-D-LL) and (c) 3-D-pipelined (3-D-PP) (2-tier, 3-tier and 5-tier) architecture configurations.

ing temperature. They reported a reduction in average conductance of the cells over the baking time, with a significant conductance drift rate in intermediate states. The reasoning provided for this effect was the relatively low stability of the weak conductive filament for intermediate states. Although their measurement-calibrated retention model was utilized to estimate the impact on CIM inference accuracy, their work only considered a monolithic 2D integration of CMOS logic and RRAM and thus were limited to conventional 2D evaluation.

Thus, it is important to perform early exploration to study long term reliability of such heterogeneous 3-D logic-memory CIM systems and benchmark different types of integration architectures from a system performance perspective. The main contributions of this article are:

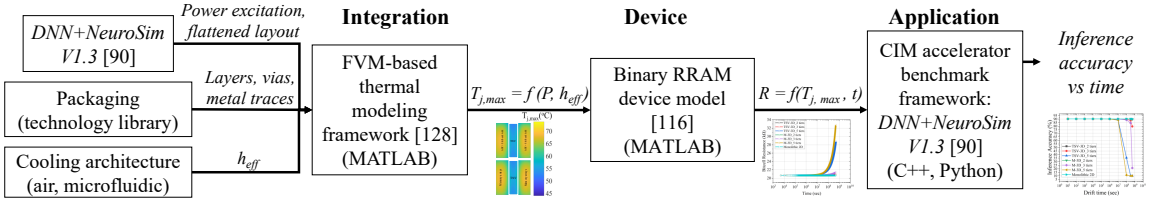


Figure 4.3: Device-integration-application-level 3-D CIM reliability evaluation flow. (Note: $T_{j,max}$ = Memory tier maximum junction temperature (^oC); P=Total package power (W); h_{eff}:=Effective heat transfer coefficient of heat sink (W/m^2.^oC); R=RRAM device resistance (Ω), t=time (sec).

Table 4.1: Power-performance trade-offs between 3-D and 2D for CIM

Metric	2D-LL	3D-LL	3D-PP
Area (mm^2)	115.1	3.7	56.2
Throughput (TOPS)	1.4	1.9	1226.5
TOPS/W	7.9	12.9	12.2
TOPS/mm^2	0.01	0.5	21.8
W/mm^2	0.0015	0.04	1.8

1. A device-integration-application-level reliability evaluation methodology is proposed that can be used to quantify the direct impact of integration design parameters on CIM inference accuracy.

2. Using this flow, heterogeneous 3-D logic-memory CIM accelerator designs - TSV-based 3-D and Monolithic 3-D-based integration of logic (7nm CMOS) and memory (22nm RRAM) tiers - are benchmarked against monolithic 2D and balanced integration design parameters for maximized 3-D CIM inference accuracy are reported.

3. We release the benchmark framework as an open-source tool (https://github.com/i3dsystems/3D_CIM_thermal_v1.0).

4.2 3-D vs 2D trade-offs for CIM

A 3-D-integrated analog CIM accelerator model was evaluated previously [113]. We applied the TSV design of CIM accelerators (7nm logic and 22nm RRAM memory) on 8-bit

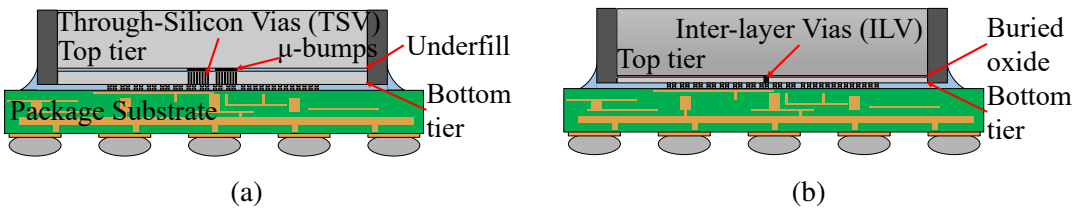


Figure 4.4: Considered (a) TSV-based 3-D and (b) Monolithic 3-D CIM configurations.

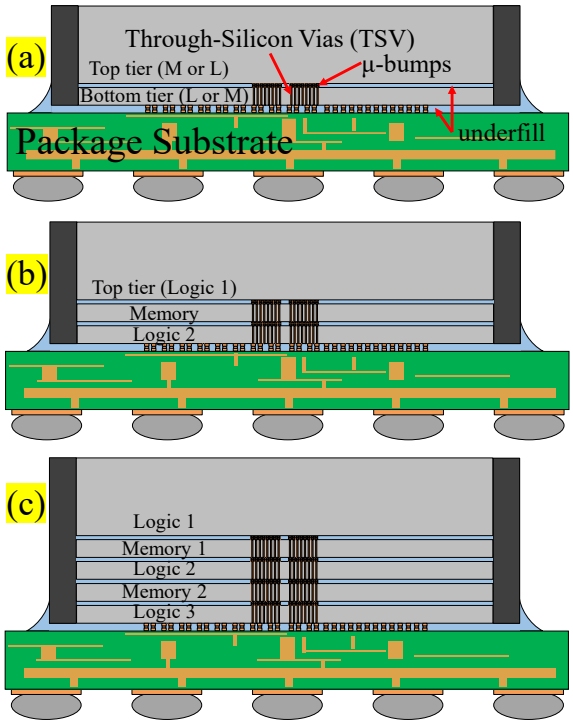


Figure 4.5: Considered (a) 2-tier, (b) 3-tier, and (c) 5-tier 3-D configurations.

ResNet-34 [125] inference for ImageNet (further details available in [113]). From an architectural perspective, we consider:

1. A layer-by-layer (LL) system (Figure 4.2b) with one logic tier on the package substrate and multiple memory tiers stacked on top (as a memory cube), which consumes low power but offers low speed, and

2. A pipelined (PP) system with 3-D interleaved logic and memory tiers (Figure 4.2c), which offers high speed but consumes high power.

Figure 4.2 shows the floorplans of baseline 2D, 3-D-LL and 3-D-PP configurations (logic in blue and memory in green). Table 4.1 provides a comparison of 2D baseline and 3-D-LL and 3-D-PP designs with 1μm TSV using this model. For 3-D, memory (RRAM) is assumed at 22 nm and logic (peripheral) is assumed at 7nm. For 2D, logic area was scaled from 7nm to 22nm to keep both logic and memory (RRAM) at 22nm ($\approx$8$\times$ area scaling). The 2D design is assumed as an LL architecture, as a PP design would require a

large number of memory arrays (for weight duplication), peripheral circuits, and buffers (to serve different DNN layers independently) on chip which leads to a prohibitively large 2D-PP area. The total number of operations (total computations needed for inference workload) for both 3-D-PP and 3-D-LL designs were fixed. The performance per watt is 63% higher with a 3-D-LL implementation vs 2D-LL. The operation-density (TOPS/mm^2) is 50$\times$ and $>$2000$\times$ higher with 3-D-LL and 3-D-PP, respectively, vs 2D-LL. However, with similar TOPS and lower area this leads to a higher power density in both 3-D-LL and 3-D-PP vs 2D-LL.

With 3-D-PP the CIM weights need to be duplicated to synchronize timing between different sized convolutional layers which corresponds to more buffers and ADCs for a PP design. Due to this 3-D-PP has a larger logic area and total power than 3-D-LL. This also means that the 3-D-PP on-chip interconnect length is larger than 3-D-LL leading to added energy and latency due to longer interconnect in 3-D-PP. Due to higher power dissipation than 3-D-LL, 3-D-PP also experiences higher leakage. Due to these two reasons 3-D-PP has a lower TOPS/W than 3-D-LL. From Table 4.1 it is clear that both 3-D designs offer higher throughput, performance-per-Watt, and operation density compared to 2D baseline. In this work, we only considered the dynamic energy of CIM arrays and peripherals. Since RRAMs exhibit negligible leakage their dynamic energy dominates. For CMOS peripherals (ADC, reference voltage circuits, etc.) the total power should be re-estimated after junction temperature simulations. Although this re-estimation was not part of this version of the methodology (section 4.3), we plan to update our flow as part of future work. The thermal performance trade-offs between 2-D and 3-D CIM are evaluated in the following sections.

4.3 Thermal-driven 3-D CIM reliability evaluation methodology

4.3.1 Simulation Flow

Figure 4.3 illustrates the proposed implementation methodology used to quantify the impact of 3-D integration design parameters on CIM inference accuracy [126, 127]. This

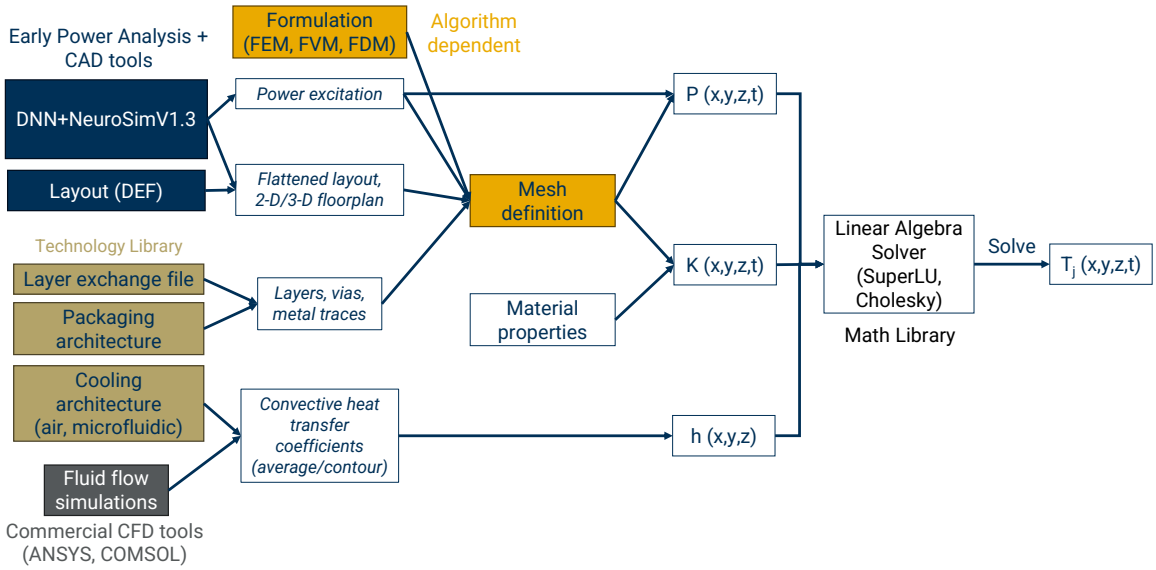


Figure 4.6: A modified version of the FVM-based thermal modeling framework described in [128] was used to model the considered integration structures.

flow combines a finite volume method (FVM)-based thermal analysis framework with a measurement-calibrated binary RRAM retention model and a CIM inference accuracy estimation framework, to perform a device-integration-application-level reliability evaluation. Details of each flow component are described as follows:

Thermal analysis

A modified version of the FVM-based thermal modeling framework described in [128] was used to model the considered integration structures, shown in Figure 4.4 and Figure 4.5, and perform steady state thermal simulations described later in this chapter. Inputs for this step include a flattened layout of the modeled CIM IP (such as memory array, ADCs, etc.), power excitation maps for each active tier/die based on the flattened layout, and a description of the die stack-up, i.e. the bulk material, interconnects, and dielectrics along with their thermal properties (thermal conductivity, specific heat capacity, etc.). The modified version of the flow (from [128]) that was used in this work is shown in Figure 4.6. The key contributions include: a) updated assumptions of the technology library such as material properties and heat transfer coefficients for accurate modeling of packaging architectures, and

b) support for interface with open source tools such as DNN+NeuroSimV1.xx (a pre-RTL simulator [90]). For validation, the memory tier maximum junction temperatures ($T_{j,max}$) from our thermal models were compared with finite-element ANSYS Mechanical APDL models (ver. 2021 R1). The maximum deviation in $T_{j,max}$ and $T_{j,min}$ between our models and ANSYS was 3.3 %. Table 4.2 summarizes the power and boundary conditions assumptions for these simulations. Although this study focuses on steady-state analysis, this framework also supports transient analysis with multi-die transient power maps including package and boundary condition definition (described in section 4.4.1).

RRAM retention

RRAMs have two switching mechanisms, non-filamentary and filamentary switching [129, 121]. This work analyzed filamentary switching because this type is commercialized in industry (e.g. IMEC, Winbond, TSMC all have used filamentary HfO_2 stacks). One concern with non-filamentary RRAM so far is that though it could support multilevel states, the retention of the intermediate states might not be stable, which could be a critical problem for weight drifting for DNN inference. We expect the study on non-filamentary type of RRAM to be a part of future work.

For the use of RRAM in CIM applications multiple circuit architectures have been proposed. Since the resistance states of RRAM can be continuously tuned, a crossbar RRAM array can be used for in-memory matrix-vector-multiplication calculations that are the core operations of different neural networks and is the intended application in this work. RRAM can also be used as a digital memory which only contains two states for binary applications [123]. In this work, we utilize arrays of 1T1R binary RRAMs for in-memory matrix-vector-multiplication calculations. While this methodology is being used for binary CIM architectures, it can also be applied to multi-level analog RRAM CIM architectures (and other devices such as ferro-electric FETs (FeFET), PCM, etc.), which will be part of future work.

Table 4.2: Experimental Setup

Parameter	**Value**
Number of chips/tiers	1, 2, 3, 5 (logic, memory)
Chip size (mm^2)	26.81 $\times$ 26.81 (2D) 9.69 $\times$ 9.69 (TSV-3D, 2 tiers) 8.36 $\times$ 8.36 (M3D, 2 tiers) 5.70 $\times$ 5.70 (TSV-3D, 5 tiers) 4.94 $\times$ 4.94 (M3D, 5 tiers)
Chip bulk thickness (μm)	Top die: 315 - 700 Bottom die: 20 (TSV-3D), 3 (M3D)
Package dimensions	15 mm $\times$ 15 mm $\times$ 1 mm
Heat Sink type	air-cooled
Heat spreader	60 mm $\times$ 60 mm $\times$ 4.5 mm
TIM thickness	TIM1: 30 μm, TIM2: 20 μm
h (W/m^2-^oC)	4.4 $\times$ 10^3 (air) [130]
Ambient Temperature	27 ^oC
Total Power (W)	118.6 Total logic: 114.85, Total memory: 3.76
Device	1-bit per cell RRAM [116], R_{on} = 20.7 kΩ, R_{off} = 100 kΩ
Network	VGG-8
Dataset	CIFAR-10

The memory tier $T_{j,max}$, as a function of integration design parameters such as architecture, number of tiers, input power, boundary conditions, etc., is an input to the second part of this flow that consists of a measurement-calibrated HfO_2-based binary RRAM device model [116]. This analytical model is used to estimate device resistance variation over time, which is converted to a resistance drift ratio. The drift ratio is calculated as (R_{10} - R_{on})/R_{on}, where R_{on} is the device low resistance state (LRS) resistance and R_{10} is the device resistance at 10 years operating at a specific junction temperature.

Ideally, an areal distribution of device junction temperatures (T_j) should be used for such an analysis where each device may experience different resistance drift leading to

Table 4.3: Tier-level Power Breakdown

Number of Tiers	**Each Logic Tier (W)**	**Each Memory Tier (W)**
2 tiers (L-M)	114.85	3.76
3 tiers (L-M-L)	57.26	3.76
5 tiers (L-M-L-M-L)	38.18	1.8

Table 4.4: Interconnect Assumptions

Interconnection		
Attribute	**TSV-3D**	**M3D**
TSV/ILV diameter (μm)	1	0.1
Number of vertical vias between two tiers	5.9×10^6	
TSV/ILV total area (mm^2)	23.60	0.24
Microbump/bonding pitch (μm)	36	0.1
Bonding layer thickness (bump height) (μm)	18	0.05

differences in retention characteristics. Therefore, using a junction temperature contour to include separate device-level (or a finer granularity than Tj,max) drift models in our inference simulations, depending on their mapped location on the die, could provide a more realistic impact on inference accuracy but can be more computationally expensive. The goal of this work was to establish the co-analysis methodology from device/integration towards application-level analysis and we assumed a conservative initial approach. We plan to include the areal distribution of drift as part of future work.

CIM inference accuracy estimation

Device retention change in RRAM-based CIM inference or training accelerators can correspond to a change in locally stored DNN weights/inputs/outpus (for weight-, input- or output-stationary dataflows [131], respectively), thus affecting the accuracy of an inference operation. The device retention drift ratio obtained from the previous step is used to adjust the drift coefficient of a retention model within DNN+NeuroSimV1.3 (a popular framework

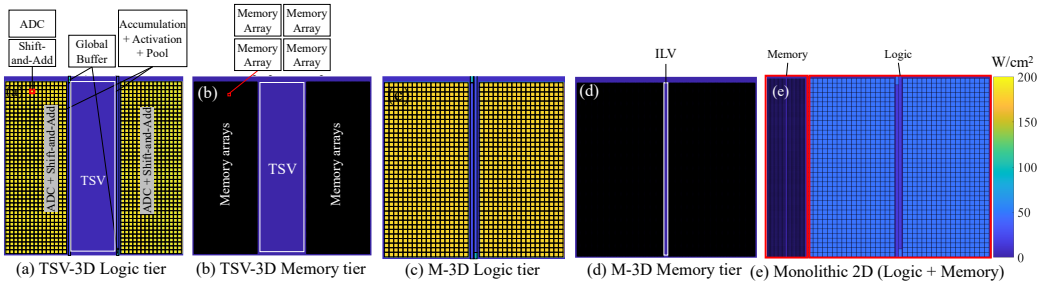


Figure 4.7: Floorplans and block-based power densities of the two-tier CIM accelerator configuration using: TSV-3D (a) logic tier, (b) memory tier and M3D (c) logic tier, (d) memory tier

to benchmark CIM accelerators [90]), to estimate the variation of CIM inference accuracy.

Using this flow, CIM inference reliability is studied to benchmark integration architectures and 3-D partitioning configurations defined in the next two sections.

4.4 Experimental Setup

4.4.1 Device, Chip, Package, Boundary Conditions, and CIM Inference Assumptions

The details of chip, package, heat spreader dimensions and boundary conditions are summarized in Table 4.2. From an architectural perspective a baseline 2D monolithic design is partitioned and arranged into multi-tier 3-D configurations as shown in Figure 4.2. Two 3-D integration architectures were considered: 1) TSV and μbump-based 3-D and 2) monolithic 3-D (Figure 4.4). After partitioning the 2D design, we assumed a two, three, and five-tier integration approach with alternating logic and memory tiers, similar to a five-tier pipelined (PP) system described in [113] and shown in Figure 4.5. The difference in chip sizes for TSV-3D and M3D arises due to the total difference in TSV/inter-layer via (ILV) area, as the considered TSVs occupy larger area compared to ILVs. We assume a state-of-the-art air cooling boundary condition [130]. For CIM-based inference, we assume an array of 1-bit per cell RRAM (R_{on} = 20.7 kΩ, R_{off} = 100 kΩ) [116], using the VGG-8 network for the CIFAR10 dataset.

4.4.2 Block-based Power Estimation

The floorplans and block-based power distribution for each logic and memory tier are shown in Figure 4.7. The power per block and number of blocks are estimated using DNN+NeuroSimV1.xx (a pre-RTL simulator [90]). The network, dataset, and device assumptions are summarized in Table 4.2. NeuroSim's mapper was used to estimate the number of memory (switch matrix, memory arrays) and peripheral logic (shift-add, ADC, accumulation + activation + pooling, global buffer) blocks for the assumed network and dataset. The energy (CV^2) for both the logic and memory blocks is estimated using NeuroSim. The logic and memory blocks are each mapped to separate tiers (top/middle/bottom) in both TSV-3D and M3D cases (Figure 4.4). In this study, we assume that the circuits are operating for a long time and that the average dynamic power for both the logic and memory tiers remain constant. We do not explicitly model switching characteristics of the devices, and assume the device switching is averaged over the power profile. As part of future work, workload dependent characteristics of the power profile could be introduced which could be used to analyze the transient noise characteristics of the assumed CIM hardware. For this study, the assumed 3-D integration schemes are shown in Figure 4.4 and Figure 4.5. The estimated total power per logic and memory tiers are mentioned in Table 4.3.

4.4.3 3-D Interconnection assumptions

The tier-to-tier interconnection (vias, I/Os) assumptions for the assumed PP architecture are summarized in Table 4.4. Each RRAM sub-array is assumed to be 128×128 and it is assumed that for each RRAM array, each row and column require an access connection. Hence, for our assumption of a 128×128 array, we need 128+128 connections per array. Multiplying this with the number of memory arrays, the total number of die-to-die interconnections required are obtained. For the TSV-3D case, a 1 μm TSV diameter was assumed for two reasons: 1) the total TSV area overhead with TSV diameter $>$ 1 μm (assuming TSV pitch = 2×diameter) becomes a significant portion of the total active area for

the number of required TSVs, and 2) for the PP architecture it was previously reported that the overall system throughput (in tera operations per second or TOPS/sec) starts to saturate for TSV diameter $\approx$ 3 μm or lower because the TSV parasitic capacitance becomes negligible compared to CMOS gate loading capacitances [113]. For the M3D case, the inter-layer via (ILV) diameter was assumed as 0.1 μm [132]. We assumed a 36 μm die-to-die microbump pitch (TSV 3D) [133], and the microbump height was assumed to be 0.5$\times$microbump pitch. While for M3D, the tier-to-tier I/O bonding pitch was assumed to be 0.1 μm with a bond height of 0.5$\times$bonding pitch.

Keep-out-zones (KOZ) help in avoiding cells being placed too close to TSVs which can cause carrier mobility variation. Two primary design considerations concerning KOZ are: hotspot mitigation and prediccarrier mobility. We find two key trade-offs in TSV and KOZ scaling from a thermal perspective: 1) smaller KOZ and smaller TSV diameter [134] can help lower thermal resistance between adjacent tiers which could be beneficial for thermal spreading, but could also increase thermal coupling between high- and low-power tiers (e.g. logic and memory) leading to higher thermal-induced conductance drift (in eNVM); 2) larger KOZ with larger TSV diameter means higher area footprint [135] but it can improve thermal spreading due to increased bulk substrate volume. For design simplicity, we do not consider KOZ effects in this work and only model the inter-TSV spacing (assuming TSV spacing = TSV diameter).

4.5 Results

4.5.1 Steady state evaluation of 3-D CIM configurations

Three partitioning configurations were considered for both TSV-3D and M3D cases: 1) two tiers: memory-on-logic (M-on-L) and logic-on-memory (L-on-M), 2) three tiers (L-M-L), and 3) five tiers (L-M-L-M-L). A summary of $\Delta T_{j,max}$ relative to the 2D baseline for each configuration is presented in Figure 4.8 and the contours for logic and memory tiers (2-tier configuration) are shown in Figure 4.9.

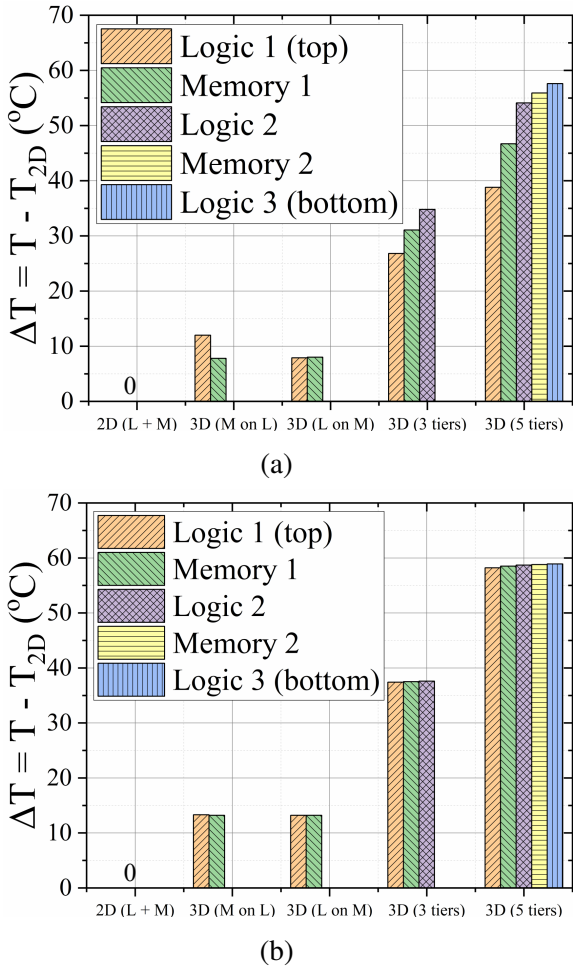


Figure 4.8: Increase in maximum temperature for different 3-D configurations in (a) TSV-3D and (b) Monolithic 3-D relative to 2D baseline.

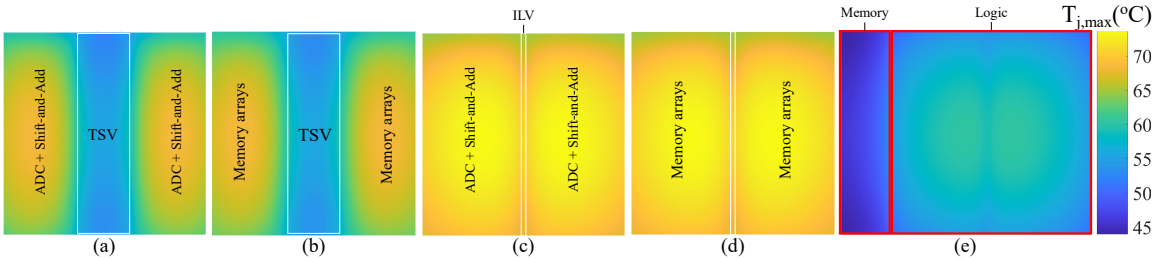


Figure 4.9: Steady state memory tier junction temperature contours for 2-tier TSV-3D (a) logic tier, (b) memory tier, 2-tier M3D (c) logic tier, (d) memory tier, and (e) monolithic 2D.

Relative temperatures ($\Delta T_{j,max}$) for each tier within all M3D configurations are similar to each other due to the low tier thickness leading to a low inter-tier thermal resistance. For both 2-tier configurations (M-on-L and L-on-M) using air-cooling, TSV-3D leads to a lower memory $\Delta T_{j,max}$ than M3D ($\approx 8^oC$ vs $\approx 13.2^oC$). This is due to the additional thermal resistance of underfill material in TSV-3D that provides better thermal isolation between logic and memory tiers. The absolute temperatures in both 2-tier configurations and in the 2D case, with the assumed boundary conditions, were not high enough to affect device retention. Redesigning the CIM system to increase the number of tiers through partitioning can provide higher throughput (TOPS) and operation density (TOPS/mm^2) using die-level pipelining (Table 4.1 and Figure 4.2c). However, the $T_{j,max}$ increases considerably (Figure 4.8) so as to affect device retention and, therefore, long-term performance i.e. accuracy of image inference operations.

4.5.2 RRAM thermal reliability in multi-tier TSV-3D and M3D

The details about considered RRAM device are noted in Table 4.2. The memory tier's retention behaviours for a binary RRAM device integrated in TSV-3D and M3D configurations are shown in Figure 4.10. We estimated the resistance values at 10 years (predicted by a measurement calibrated HfO_2 RRAM device model [116]), to find the resistance drift ratio with respect to the original LRS resistance. At higher junction temperatures the device has lower retention robustness. M3D with 5 tiers observes the highest device resistance

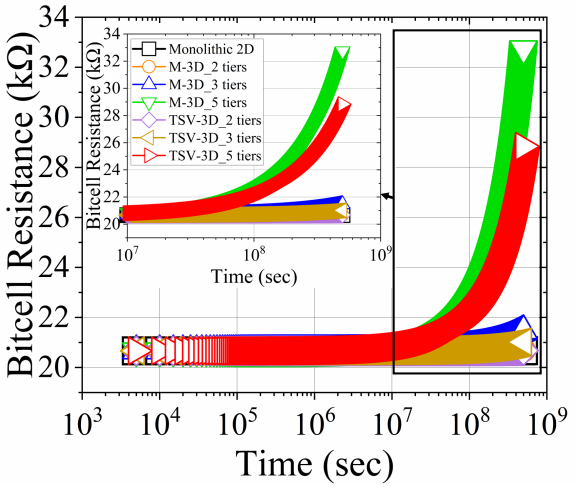


Figure 4.10: Memory tier (binary RRAM) retention for TSV-3D and M3D, both air cooling.

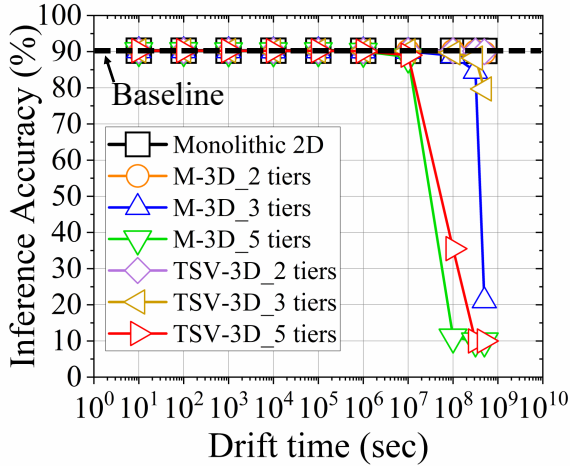


Figure 4.11: CIM inference accuracy comparison between monolithic 2D, TSV-3D and M3D (with air cooling).

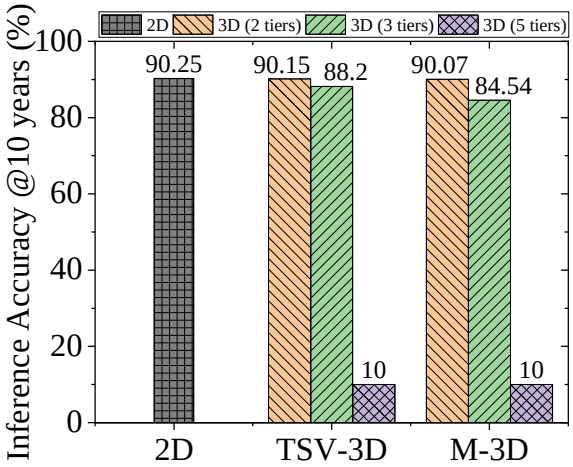


Figure 4.12: CIM inference accuracy @10 years as a function of number of tiers.

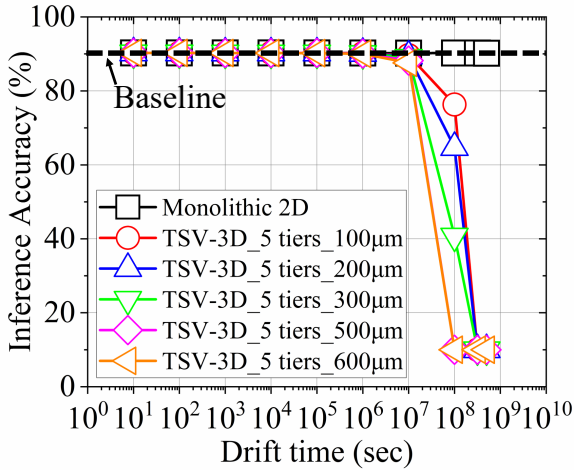


Figure 4.13: CIM inference accuracy as a function of top die bulk thickness.

drift due to the highest memory tier junction temperature, followed by the TSV-3D 5-tier configuration. The 2D and 2-tier 3-D configurations did not show any observable drift due to low vertical thermal resistance and larger die area for heat exchange leading to lower memory $T_{j,max}$.

The drop in retention with air-cooling is reasonably delayed (observed post t=10^7sec) mainly because the heat spreader dimensions (area and thickness) were increased to ensure that $T_{j,max}$ across all configurations is below 125^oC [136]. The tradeoff is that this leads to a large heat spreader introducing thermo-mechanical challenges. Such high junction temperatures can be further reduced with advanced thermal management techniques such as microfluidic cooling [137, 138, 139, 140, 141, 142]. Thermal-aware design-time partitioning [143, 144, 145] is an alternative approach, from a design standpoint, to mitigate high junction temperatures and optimize for latency and energy in 3-D ICs.

4.5.3 Multi-tier CIM Inference Accuracy

Device retention is a key factor impacting long term inference accuracy in RRAM-based CIM accelerators. The drift coefficient calculated from the retention data is used to adjust the retention speed in NeuroSim's [90] device retention model. The impact of integration architectures on CIM inference accuracy for 2-D and 3-D accelerator designs based on

VGG-8 for the CIFAR-10 dataset is shown in Figure 4.11. Due to higher memory tier $T_{j,max}$ in both 5-tier M3D and 5-tier TSV-3D compared to the 2D baseline, even after using an optimistic air-cooling configuration, the drop in inference accuracy at 10 years was ≈80%. Figure 4.12 compares inference accuracies for all configurations @ 10 years. In comparison to the 2D baseline, 2- and 3-tier configurations show minimal deviation from the baseline accuracy of 90% while redesigning 2D into 5-tiers leads to significant loss in accuracy. This significant reduction in inference accuracy can be mitigated with appropriate design-time partitioning of the 3-D stack and using more efficient thermal management architectures such as microfluidic cooling [137, 138, 139, 140, 141, 142]. For our assumed device, integration and application parameters, a 3-tier configuration provides a balanced design option between thermal and application performance.

This analysis demonstrates the potential cost of achieving higher performance through die stacking. By stacking up to five dies in a logic-memory-logic-memory-logic design (pipelined or PP design) we can achieve ≈1.5 × higher performance-per-watt compared to monolithic 2-D (Table 4.1) at the cost of ≈ 55 ^{o}C increase in junction temperature (Figure 4.8a) and an 80 % loss in inference accuracy (Figure 4.12).

4.5.4 Impact of Bulk Thickness

The top die silicon bulk thickness was varied for the TSV-3D 5-tier configuration to study the impact on image inference accuracy. The results are summarized in Figure 4.13. Increasing top die thickness increases the stack thermal resistance, which leads to a higher memory junction temperature. Compared to the 2D baseline, all 5-tier TSV-3D configurations observe ≈ 80% reduction in inference accuracy @10 years. Additionally, increasing top die bulk thickness also increases the rate of loss of accuracy, as seen in Figure 4.13 where going from 100 μm to 600 μm, accuracy drops faster with time.

4.6 Conclusion

A device-integration-application evaluation methodology is proposed that is used to quantify the impact of integration architectures on RRAM reliability for CIM applications. Two heterogeneous 3-D logic-memory CIM accelerator designs - TSV-based 3-D and Monolithic 3-D-based integration of logic (7nm CMOS) and memory (22nm RRAM) tiers - were benchmarked against monolithic 2D and balanced integration design parameters were reported for maximized 3-D CIM inference accuracy. For our assumed device, integration and application parameters, a 3-tier configuration provides a balanced design to achieve optimal system performance. The PP schemes are preferred for high-performance systems, with high operating temperature being a potential trade-off that can be improved with advanced thermal management and cooling architectures.

CHAPTER 5
BEOL-EMBEDDED 3D POLYLITHIC INTEGRATION: THERMAL CONSIDERATIONS AND IMPLICATIONS ON BEOL RRAM PERFORMANCE FOR CIM APPLICATIONS

5.1 Introduction

With increasing integration complexity and power densities in 3D integrated ICs, heat dissipation, thermo-mechanical reliability, and inter-die bonding yield are becoming challenging. In this research, we explore two such challenges of the proposed technology, which we term 3D Seamless-off-Chip-Connectivity (3D SoC+). First, we evaluate the thermal constraints for 3D SoC+ with aggressive cooling to investigate thermal limits from transient- and steady-state perspectives. Second, we present a study to evaluate the impact of cooling architectures on binary RRAM devices in a 3D IC form factor by quantifying image recognition accuracy over time of a compute in-memory accelerator based on RRAM.

The rest of the chapter is organized as follows: Section 3.2 describes the proposed 3D SoC+ scheme and reports a steady state thermal evaluation of SoC+ tier junction temperatures (T_j) as a function of (a) embedded tier power density, (b) embedded tier thickness, (c) inter-tier BEOL thickness, and (d) dielectric thermal conductivity variation. A transient thermal analysis to estimate inter-tier thermal coupling is also presented. In section 3.3, the thermal implications of 3D polylithic integration on BEOL RRAM performance for CIM applications are presented.

5.1.1 Polylithic 3D Integration

A 3D polylithic integration scheme is proposed in this work. As shown in Figure 5.1, this scheme represents a densely integrated system divided into multiple device tiers where cus-

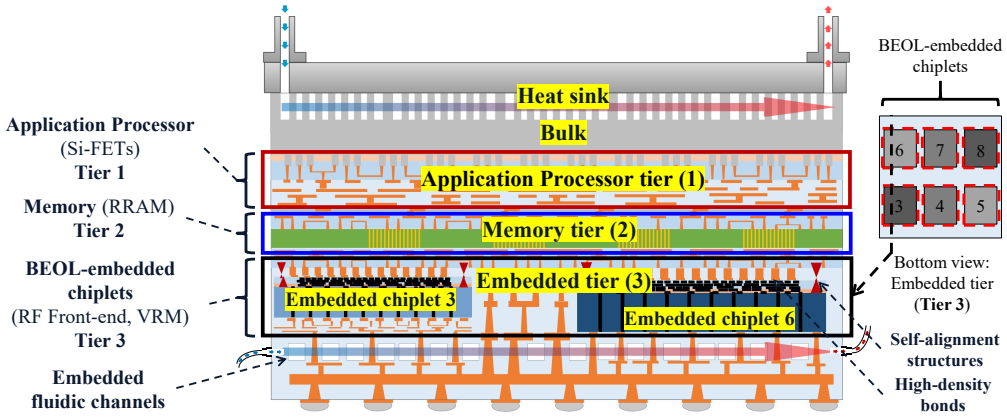


Figure 5.1: Proposed 3D Seamless off-chip Connectivity (SoC+) concept: BEOL-embedded chiplet integration

tom chiplets, such as voltage regulator modules (VRMs), I/O drivers, and RF front-ends are embedded into the back-end of an application processor (AP) tier with a monolithic memory tier, e.g. RRAM. The proposed 3D SoC+ concept aims to combine the best of both monolithic and TSV-based 3D ICs, including extreme efficient signaling and large bandwidth density (BWD). The proposed scheme can be enabled by high-density interconnections and high-accuracy self-alignment techniques.

5.2 Thermal Exploration of BEOL-Embedded Chiplet Integration

The majority of the heat generated in high-power 2D packages is typically extracted through the bulk substrate and encounters the thermal resistances associated with conduction through the bulk, thermal interface material (TIM), and the heat spreader. However, in high-power 3D packages, the heat is generated in multiple active layers along the 3D stack and thus, the heat extraction path includes thermal resistances from the bulk, BEOL, and bonding/underfill layers [146]. Therefore, in this section, we describe the proposed 3D SoC+ scheme and investigate the impact of various design parameters on thermal performance.

5.2.1 3D SoC+ Integration Scheme: Proposed Architecture

As discussed in the previous section, the goals of polylithic integration are primarily twofold: (a) to maintain integration heterogeneity by allowing one to integrate chips that are made from different materials or in different technology nodes, and (b) to increase chip-to-chip connectivity compared to TSV-based 3D integration to attain monolithic-like interconnection density. The proposed BEOL-embedded integration scheme shown in Figure 5.1 aims to achieve these goals.

The proposed 3D SoC+ architecture consists of multiple active tiers with multiple chiplets embedded within the BEOL of a primary base chip, such as an application processor (AP), as shown in Figure 5.1. The base tier (tier 1) consists of active devices in the primary bulk substrate (such as Si), and the first few metal layers of the BEOL. This AP tier is followed by a memory device tier (tier 2), which can be a monolithic memory layer such as resistive RAM (RRAM) [147]. The presence of a monolithic memory tier is application specific as it introduces stringent device, material, and temperature constraints on the fabrication process akin to monolithic 3D IC fabrication. The BEOL following tier 2 encapsulates custom chiplets, such as logic, memory, VRMs, I/O drivers, passives, RF front-end chips, etc., which are bonded to tier 2 using high-density low-temperature interconnections. Given the potential for high-power density in such an approach, it is envisioned that such a 3D stack may require single or dual-sided microfluidic cooling, as shown in Figure 5.1.

5.2.2 System Description and Specifications

We modeled the SoC+ structure using a finite-volume based thermal modeling framework described in [128]. We compared maximum junction temperatures for a steady state simulation to validate our SoC+ thermal model against ANSYS Mechanical APDL solver (ver. 19.2). The simulations were performed for high and moderate power densities use cases (defined in Table 5.1). For the high-power density case, the maximum relative error in

Table 5.1: SoC+ thermal simulations: design specifications and assumptions

Tier / Chiplet		**Power Density** (W/cm^2)	**Dimensions** ($mm \times mm \times \mu m$)	**Bulk Material**
Tier 1 (base) (AP)		100	$10 \times 10 \times 157$	Silicon
Tier 2 (RRAM)		5	$10 \times 10 \times 8$	Silicon
64emTier 3 (with embedded chiplets)	Chiplet 3	5/50	$2.5 \times 3 \times 12$	Silicon
	Chiplet 4	5/50	$2.5 \times 3 \times 12$	Silicon
	Chiplet 5	5/50	$2.5 \times 3 \times 12$	Silicon
	Chiplet 6	5/50	$2.5 \times 3 \times 12$	Silicon
	Chiplet 7	5/50/500	$2.5 \times 3 \times 12$	Silicon
	Chiplet 8	5/50	$2.5 \times 3 \times 12$	Silicon

maximum and minimum junction temperature rise between our model and ANSYS is less than 12.5 % whereas for the moderate power density case, the maximum deviation was less than 2.6%. The highest percentage error in the high-power density case was observed only for chiplet 7 (tier 3), which was designed to have the highest power density ($500 W/cm^2$), and the region in the memory tier that was right above chiplet 7. However, junction temperature percentage errors for all other chiplets and tiers were less than 2.4%. Below, we describe the modeling specifications and assumptions.

Power Dissipation Configurations

The tier power densities and dimensions are defined in Table 5.1. The power densities chosen for the AP and memory tiers are similar to high-end limits for modern designs. The power densities chosen for the embedded tier (tier 3) correspond to moderate and high power use cases of VRMs and RF front-ends. The materials mentioned for each tier in Table 5.1 correspond to plausible choices for each tier. For simplicity, Si was chosen as the bulk material for all thermal simulations in this chapter.

Table 5.2: Material specifications

Layer	**Conductivity** (W/m.K)		**Heat Capacity** (J/°C.Kg)	**Mass Density** (Kg/m^3)
	In-Plane	**Through-Plane**		
TIM	3		1000	2900
Heat-spreader	400		385	8690
Si bulk	149		705	2329
BEOL	61.173	1.6225	433	7783
Bonding (Cu+ILD)	1.6		1000	2100

Tier Dimensions and Materials

Table 5.1 lists the assumed SoC+ tier dimensions. Ideally, we seek extremely thin dice for back-end integration to reduce overall 3D IC form factor. However, choosing embedded chiplet thickness presents a trade-off between die thinning and handling, and dielectric thickness. In this chapter, we assume the total thickness of each chiplet in tier 3 as 12 μm. The total thickness assumed for tiers 1 and 2 is 157 μm (150 μm Si bulk, 5 μm BEOL, 2 μm bonding layer) and 8 μm (3 μm Si bulk, 3 μm BEOL, 2 μm bonding layer), respectively. 2 μm thick BEOL layers were considered both between tiers 1-2 and tiers 2-3.

Table 5.2 lists the material properties [148] of the layers in the 3D SoC+ stack. The bonding layer is assumed to consist of copper and inter layer dielectric (ILD; assumed as SiO_2).

Cooling configurations and boundary conditions

Boundary conditions are modeled as effective heat transfer coefficients applied uniformly over the area of a surface to be cooled. The ambient temperature is assumed to be 38 °C [148]. The impact of various design parameters on the thermal performance of the proposed 3D SoC+ scheme was evaluated in the presence of the following two types of cooling configurations:

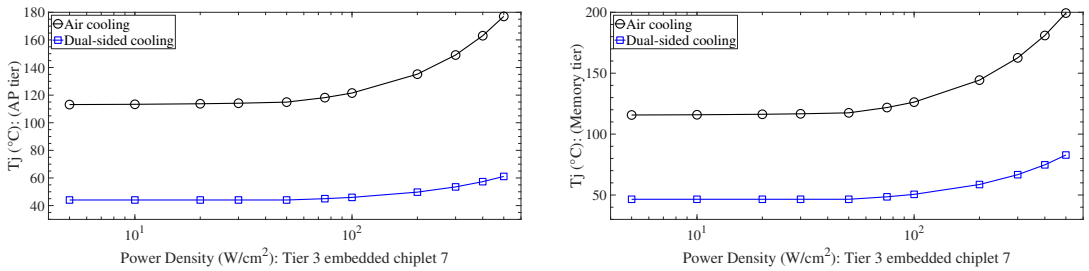
(a) Air cooling: We model an air-cooled heat-sink mounted on the top of the stack with the following layers between the heat-sink and tier 1 bulk: a top TIM, a heat-spreader, and

a bottom TIM. Sarvey $et.al.$ [149] conducted an experiment to compare the performance of an air-cooled heat-sink to that of a microfluidic heat-sink for a cooled surface area of approximately 9 cm^2 and reported the junction-to-ambient thermal resistance for a high-end air-cooled heat-sink as 0.25 °C/W. With these assumptions, an upwards effective heat transfer coefficient of 4.44×10^3 W/K·m^2 ($h_{eff,up}$) was used for our simulations, applied on the top TIM surface. The TIM and heat-spreader dimensions and material properties were assumed to be the same as specified in [148]. However, it is important to note that to effectively model a state-of-the-art air-cooled heat-sink suifor the assumed package power, a detailed design of experiments is required to determine appropriate heat-spreader dimensions for desired case-to-ambient thermal resistance. A natural cooling of 10 W/K·m^2 ($h_{eff,bot}$) is applied to the bottom surface under the stack.

(b) Dual-sided cooling (DSC): DSC refers to active cooling of top and bottom surfaces in a stack. Brunschwiler et. al. [150] have demonstrated microfluidic channels embedded in an interposer and microfluidic cooling for heat removal from the bottom and top, respectively, of a 3D stack. We use a similar approach for cooling of the 3D SoC+ stack. We model a microfluidic μ-cooler on the top substrate (above tier 1) with a heat transfer coefficient of 3.33×10^5 W/m^2-K ($h_{eff,up}$). For the bottom surface, BEOL-embedded microchannels were modeled for microfluidic cooling. For a first order estimation of the effective heat transfer coefficient under the stack ($h_{eff,bot}$), it was assumed that the heat-sink thermal resistance scales linearly with respect to the available heat-sink volume. Based on the assumed ratio of volumes of the two heat-sinks (top and bottom), the $h_{eff,bot}$ was estimated to be 3.11×10^3 W/m^2-K.

5.2.3 Impact of Design Parameters: Steady State Evaluation

In this subsection, we focus on thermal evaluation of back-end chiplet integration as a function of various design parameters. In particular, we explore embedded tier power densities, embedded tier thickness, and inter-tier BEOL thickness as they are some of the most



(a) AP tier T_j, max variation (b) Memory tier T_j, max variation

Figure 5.2: (a) Application Processor (AP) tier and (b) memory tier maximum junction temperatures as a function of chiplet 7 (embedded tier) power density

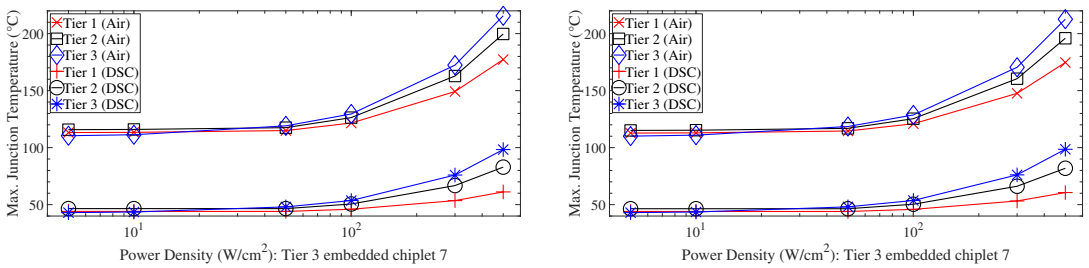
relevant factors from a thermal standpoint. We perform steady-state analyses with static uniform power dissipation on each tier and all chiplets. Additionally, to study the extent of inter-tier thermal coupling, we perform transient analysis with a processor-like simulated activity factor on tier 1. Through these analyses, the limits and challenges of dense 3D integration can be better understood.

Maximum Power Density Limits for Embedded Dice

Chip power densities can vary drastically based on function, technology node, and device technology. Heterogeneity combined with close proximity and stacking can result in high-power densities. Thus, it is important to study the limits of power densities for 3D integration as a function of cooling approach.

To investigate the limits of embedded tier power densities, we make the following assumptions. All embedded tier chips are 0.075 cm^2 in area. The AP tier, memory tier, and embedded tier chiplets are assumed to have power densities of 100 W/cm^2, 5 W/cm^2, and 50 W/cm^2, respectively. The power density of one of the embedded tier chiplets (Chiplet 7) was varied from 5 to 500 W/cm^2, and the power densities of other chiplets were kept constant. All other design parameters are listed in Table 5.1 and Table 5.2 unless otherwise explicitly specified.

The results for processor and memory tier junction temperature variation are shown in Figure 5.2 (a) and (b). The rise in maximum base silicon (tier 1) junction tempera-

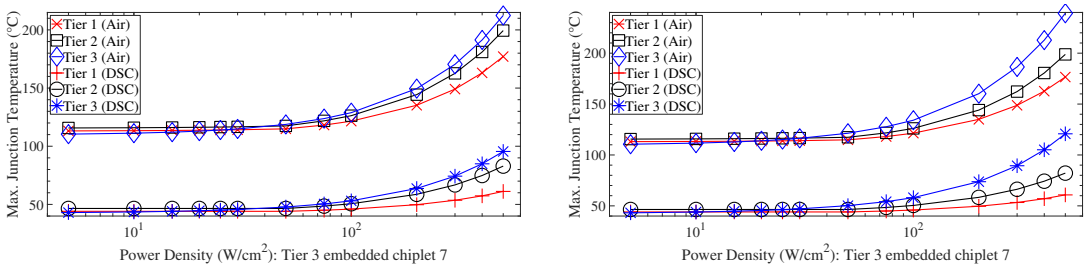


(a) Embedded tier thickness = 1μm (b) Embedded tier thickness = 50μm

Figure 5.3: Impact of embedded tier thickness scaling for thickness of (a) 1μm and (b) 50μm

ture ($T_{j,max,AP}$) (from lowest to highest embedded tier power density) for air-cooling and for DSC is 63.86 °C and 17.06 °C, respectively. The same values for tier 2 (monolithic memory) and embedded tier (chiplet 7) were 83.66 °C, 36.33 °C, and 104.96 °C, 55.37 °C, respectively. The following two observations were made: First, the rise in $T_{j,max}$ and $T_{j,min}$ were highest for the embedded tier in both air-cooling and DSC cases. This can be explained by: (a) the junction-to-heat-sink thermal resistance in the dominant heat removal path (upward) is highest for the embedded tier 3 in both cooling scenarios, and (b) the $h_{eff,bot}$ is two orders (air-cooling) and an order (DSC) of magnitudes lower than $h_{eff,up}$ (i.e. heat removal path via tier 1 (base tier) silicon). Thus, it is important to note that this result is inherent to the assumptions in the chapter and can change as a function of cooling architecture. This implies the need for a lower thermally resistive path closer to the embedded tier in the stack, since with the low thermal conductivity of the BEOL, spreading is minimal. This will be a greater concern as the number of embedded tiers increase.

Second, with the use of DSC, tier temperatures across all dice were observed to be below the critical limit of 105 °C (limit from [151]). Assuming this is the largest tolerated junction temperature, the maximum embedded chiplet power density using DSC can be higher than 500 W/cm^2 for chiplet 7 with other five chiplets at 50 W/cm^2. These observations imply that such a dense and heterogeneous scheme is thermally viable for high-power densities at the cost of cooling design and integration complexity.



(a) Inter-tier BEOL thickness = 1μm (b) Inter-tier BEOL thickness = 10μm

Figure 5.4: Impact of inter-tier BEOL thickness (BEOL between tier 2 and embedded tier) for thickness of (a) 1μm and (b) 10μm

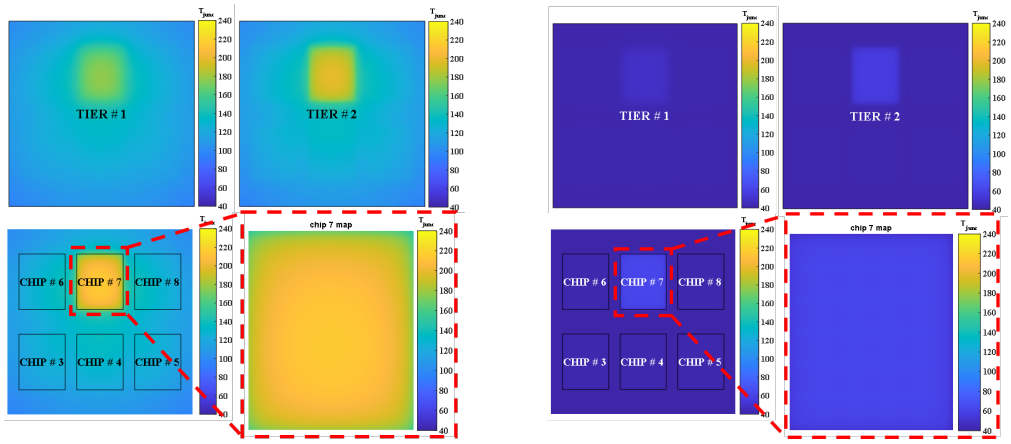
Embedded Tier Thickness

The embedded tier thickness is crucial for heat spreading and a thicker tier can reduce localized hot-spot temperatures [148], especially with high power density chips in a 3D stack. Figure 5.3 (a) and (b) show the impact of embedded tier thickness variation on tier $T_{j,max}$ as the power density of embedded chiplet 7 is varied from 5 W/cm^2 to 500 W/cm^2. We consider thickness values of (a) 1 μm and (b) 50 μm for chiplets in the embedded tier. Going from 1 μm to 50 μm at 500 W/cm^2, the $T_{j,max}$ for each tier in both the air cooled and DSC cases does not change significantly, as expected (2.61°C, 3.76°C, 2.98°C with air-cooled and 0.55°C, 0.99°C, -0.25°C with DSC for tier 1, tier 2, and tier 3 (chiplet 7), respectively). However, the within-tier temperature variation is relatively higher. For instance, the intra-tier temperature difference ($T_{j,max}$ – $T_{j,min}$) for the embedded tier in the air cooled case changes from 68.47°C to 55.6°C at 500 W/cm^2 as the embedded tier thickness is changed from 1 μm to 50 μm. The same difference changes from 9.08 °C to 8.08 °C at 5 W/cm^2. The corresponding values with the use of DSC are 41.67 °C to 35.04°C (500 W/cm^2) and 0.36 °C to 0.31 °C (5 W/cm^2). This implies that spreading in the embedded tier becomes significantly important at higher power densities. In conclusion, the intra-tier temperature spread, can be mitigated with embedded cooling, and this spread can be reduced (albeit minimally) with a thicker embedded dice. But based on our assumptions in this chapter, there does not appear to be a motivation for thick embedded chiplets.

Impact of Inter-tier BEOL Thickness

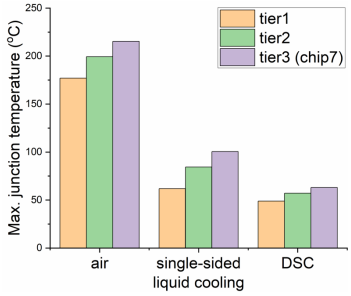
Based on the extent of required inter-tier required connectivity, for instance between the RRAM and an embedded chiplet, the number of metal layers in the BEOL may change. A higher number of metal layers would increase the BEOL thickness. An increase in inter-tier BEOL thickness effectively increases thermal resistance in the stack due to the addition of low thermal conductivity material [152, 153]. Therefore, we investigate the impact of BEOL thickness between tier 2 and the embedded tier on the $T_{j,max}$ of the three tiers.

Figure 5.4 (a) and (b) show the extent of variation in $T_{j,max}$ for AP, memory and embedded tiers as the power density of embedded chiplet 7 and the BEOL thickness between tier 2 and embedded tier were varied. The following two observations were made. First,



(a) Air-cooling

(b) Dual-sided cooling



(c) Cooling Comparison

Figure 5.5: Thermal profile ($T_{j,max}$) of AP, memory, and embedded tier with (a) air-cooling and (b) dual-sided cooling (BEOL thickness is 10 μm). (c) $T_{j,max}$ comparison for air, single-sided, and dual-sided cooling. (all results with total power = 162W).

in both cooling cases, as the BEOL thickness is increased from 1μm to 10μm, the maximum tier junction temperatures for the AP and memory tiers reduced marginally (reduced by 0.51°C and 0.61°C for air cooling and by 0.38°C and 0.76°C for DSC). However, as expected, the temperatures increased for all chiplets in the embedded tier (air cooling: increased by 26.8°C for chiplet 7 and 1.9 °C (average) for all other embedded chiplets; DSC: increased by 25.2 °C for chiplet 7 and 2.4 °C (average) for all other embedded chiplets). This can be explained by the increased BEOL thickness between tier 2 and embedded tier contributing a high thermal resistance. Figure 5.5 (a) and (b) show thermal profiles ($T_{j,max}$) of each tier for air cooling and DSC, respectively, and (c) shows ($T_{j,max}$) comparison for air, single-sided liquid, and dual-sided liquid cooling.

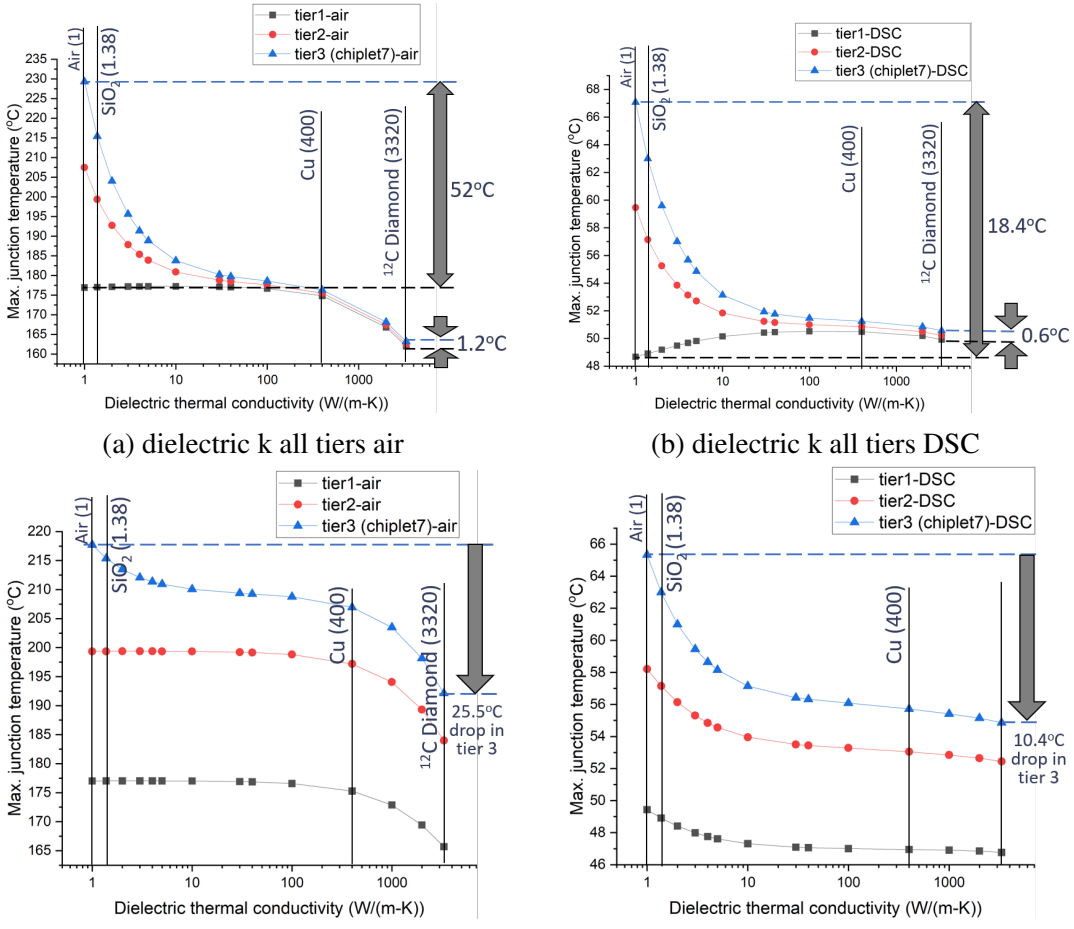
Dielectric Thermal Conductivity Variation

The thermal conductivity of dielectric material was varied: 1) for all SoC+ tiers, and 2) for just tier 3. Figure 5.6(a,b) and Figure 5.6(c,d) depict the variation in steady state ($T_{j,max}$) as the thermal conductivity of dielectric is varied from 1 W/m-K (air, low capacitance air-gap interconnects [154]) to 3320 W/m-K (diamond). With an increase in dielectric thermal conductivity for all tiers, the inter-tier temperature difference between tier 1 and tier 3 ($\Delta T_{j,max} = T_{j,max,tier1} - T_{j,max,tier3}$) decreased from 52.3^{o}C to 1.2^{o}C (air-cooling) and from 18.4^{o}C to 0.6^{o}C (DSC) (Figure 5.6(a,b)). Furthermore, with an increase in thermal conductivity of the dielectric surrounding tier 3, the $T_{j,max}$ drop for tiers 1, 2, and 3 was 11.3^{o}C, 15.3^{o}C, and 25.5^{o}C, respectively, with air-cooling, and 2.6^{o}C, 5.7^{o}C, and 10.4^{o}C for DSC (Figure 5.6(c,d)). This ≈60-70% lower reduction for DSC can be explained by the initial lower absolute temperatures compared to air-cooling. The observations are two-fold. First, dielectric materials with higher heat spreading capability might be needed when inter-tier temperature difference ($\Delta T_{j,max}$) in 3D tiers needs to be minimized. Second, thermal challenges in 3D ICs might require additional conductive and convective solutions such as heat-sinking from multiple sides and, potentially, advanced backend heat spreading

using higher thermal conductivity dielectric materials. This suggests that apart from early thermal analysis and thermal-electrical co-design, thermal challenges in 3D ICs will also require advance process solutions for heat-spreading.

5.2.4 Transient Evaluation

The transient thermal coupling between all 3D SoC+ tiers and the chiplets within the embedded tier depend on the transient activity of the active tiers and the cooling scheme used. To study the extent of thermal coupling in the 3D SoC+ configuration, in the presence of air and dual-sided cooling, the following transient-state analysis is performed. The AP tier



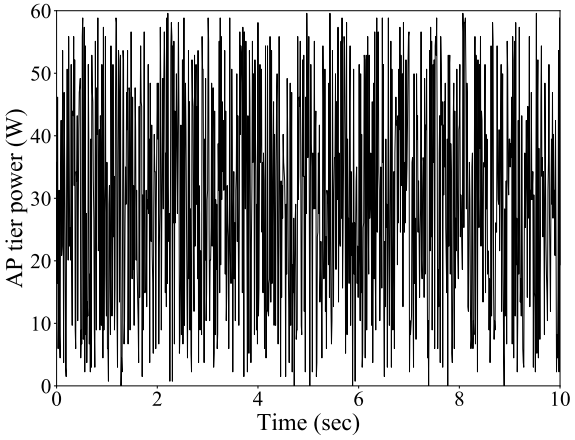
(a) dielectric k all tiers air

(b) dielectric k all tiers DSC

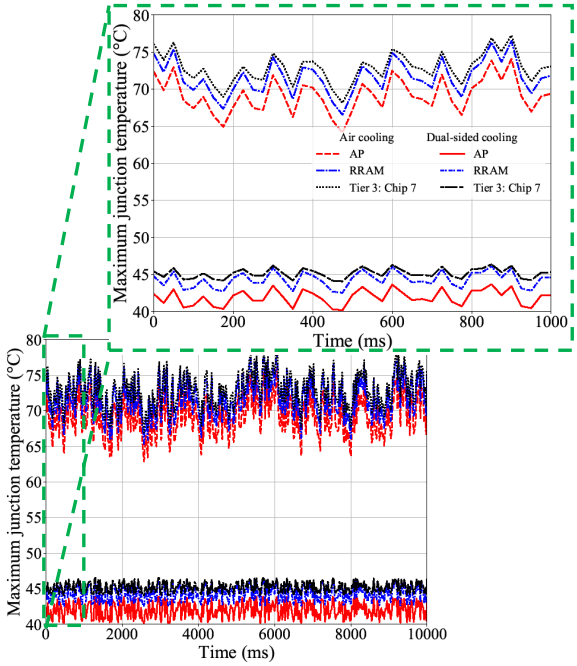
(c) dielectric k embedded tier air

(d) dielectric k embedded tier DSC

Figure 5.6: Maximum junction temperatures as a function of varying dielectric thermal conductivity in: 1) all tiers with (a) air and (b) dual-sided cooling (DSC) and 2) just embedded tier with (c) air and (d) DSC



(a) Emulated processor power variation



(b) Transient variation in maximum tier junction temperatures

Figure 5.7: (a) Emulated processor power and (b) transient variation in maximum tier junction temperatures: extent of inter-tier thermal coupling

power map emulates a processor workload with a power map shown in Figure 5.7a with an activity factor ranging between 0.01-0.80 [148]. The memory tier is assumed to have uniform power density of 5 W/cm^2, and the power densities for embedded tier chiplets are assumed to be 50 W/cm^2 (chiplets 3, 5, and 7) and 5 W/cm^2 (chiplets 4, 6, and 8).

We estimate inter-tier thermal coupling using the following method. The simulation

was run for a 10 second duration with a 10ms time step. For each two consecutive time steps (for e.g. from t=0sec to t=0.01sec and so forth), we calculate the variation in maximum junction temperatures ($T_{j,max,t+0.01s}$ - $T_{j,max,t}$) for each tier. With the presence of inter-tier coupling, a finite variation in AP tier maximum junction temperature ($\Delta T_{j,max,AP}$) should lead to a variation in a neighbouring tier's maximum junction temperature ($\Delta T_{j,max,tier}$). We quantify thermal coupling as the average (over 10sec) ratio of $\Delta T_{j,max,tier}$ and $\Delta T_{j,max,AP}$ between every two consecutive time steps, which can be represented as:

$$Coupling = avg\{\frac{(T_{j,max,t+0.01s} - T_{j,max,t})_{tier}}{(T_{j,max,t+0.01s} - T_{j,max,t})_{AP}}\} \tag{5.1}$$

Figure 5.7 (b) depicts the maximum junction temperatures of each tier. Without the presence of inter-tier thermal coupling, the junction temperatures of the monolithic memory and and embedded tiers should be constant and coupling defined by equation (Equation 5.1) would be 0. However, it can be seen from Figure 5.7 (b) that a strong inter-tier coupling exists. The thermal coupling from AP tier to the memory tier and tier 3 (chip 7), respectively, were estimated as 0.99 and 0.85 for air-cooling and 0.99 and 0.63 for DSC. This implies that with an efficient cooling solution, the inter-tier thermal coupling can be reduced. Furthermore, with the coupling estimations, it can be inferred that proximity of a heat-sink to an active tier affects both the tier's thermal coupling from neighbouring active tiers and its absolute junction temperature. The embedded tier's proximity to the bottom embedded cooling (DSC), and a higher thermal resistance between the embedded and AP tiers, enables better isolation, from the AP tier, for the embedded tier than for the memory tier.

5.3 Conclusion

This chapter presents a back-end-embedded chiplet integration scheme for heterogeneous 3D integration, which is envisioned to combine the low EPB and high BWD benefits of

monolithic 3D ICs with the integration flexibility of TSV-based 3D integration. A thermal study focusing on 3D SoC+ is presented to identify the thermal limits and challenges in such a scheme. For steady state operation of primary and embedded tiers, the impact of design parameters such as (a) embedded tier power density, (b) embedded tier thickness, and (c) BEOL thickness on maximum tier junction temperatures were evaluated with air-cooling and dual-sided cooling (DSC). Moreover, transient analysis was performed to estimate thermal coupling from the base silicon layer (tier 1) to the embedded tier. It was observed that the heat-sink proximity to an active tier affects the thermal coupling, with up to 20% reduction observed with DSC.

CHAPTER 6
DESIGN OPTIMIZATION STRATEGIES FOR POWER DELIVERY NETWORK IN POLYLITHIC 3-D INTEGRATION

6.1 Introduction

As described in the previous chapter, the technological push towards 3D heterogeneous integration such as TSV based 3D ICs is driven by the need for higher bandwidth and lower delay in chip-to-chip signal interconnections. To bridge the performance gap in connectivity and heterogeneity between monolithic 3D and TSV-based die stacking, a back-end-of-line (BEOL)-embedded integration scheme is proposed in this work where thinned dice are envisioned to be integrated close to the back-end using fine pitch interconnects (polylithic 3D integration). This can alleviate electrical signaling performance degradation due to long wire lengths and large pad sizes leading to improved EPB and lower chip-to-chip delay. Such a technology can potentially combine the benefits of current heterogeneous ICs (e.g. lower costs, technology node flexibility, higher yield, etc.) with the performance superiority of monolithic 3D ICs. However, potential challenges from a power delivery perspective can impact device performance.

We present design optimization strategies for power delivery networks (PDN) in polylithic 3-D integration. The proposed 3D polylithic architecture represents a densely integrated system divided into multiple device tiers where custom chiplets, such as power management IP, I/O drivers, and memory are embedded into the back-end of a base tier with extreme efficient signaling and large bandwidth density. The scope of this work is a detailed design space exploration of the power supply noise effects in polylithic 3-D architectures. We propose three polylithic PDN designs and benchmark their IR-drop as a function of tier power, number of embedded chiplets, hot-spot location, and TSV diameter and distribution

to provide design limitations and insights.

6.2 PDN considerations for 3D Integration

Among various heterogeneous integration (HI) architectures, such as multi-chip modules (MCM), 2.5D, and 3D, 3D-HI can provide higher compute density and signaling energy-per-bit through a reduced footprint and interconnection length, respectively, compared to MCM and 2.5D [68]. A growing need for higher logic-memory bandwidth and lower chip-to-chip signal interconnection delay have led to a technological push towards 3D-HI such as through-silicon via (TSV)-based 3D integrated circuits (ICs) [68, 69, 70]. Although 3D-HI can enable dense memory-logic integration needed for state-of-the-art CIM hardware accelerators, there are power delivery challenges with 3D-HI for CIM.

A factor that poses a challenge in power delivery design for edge intelligent hardware is the current trend of increasing power and power density in recent CIM and hardware accelerators, as shown earlier in Figure 4.1. Increasing DNN model size and workload complexity can lead to larger die sizes due to a higher demand for on-chip resources such as memory arrays, ADCs, etc. When it comes to low-power edge applications, the primary motivations are to improve energy efficiency (tera-operations-per-sec-per-Watt or TOPS/W) and compute efficiency (tera-operations-per-sec-per-square mm or TOPS/mm^2). This can be achieved through device scaling and reducing the overall hardware form-factor. As a result, the area occupied by an edge intelligent hardware and voltage regulators will need to shrink. A push for thinner devices usually corresponds to reduction in height of the die and the power delivery components such as interconnects, capacitors and inductors. Additionally, recent work has demonstrated performing vector-matrix-multiplication in parallel on multiple CIM cores, which introduces certain non-idealities such as core-to-core variation of IR-drop and supply voltage instability [76]. All these trends introduce multiple unique challenges in designing a robust power delivery network for CIM.

3D-HI brings additional challenges to computational accuracy. These include steady-

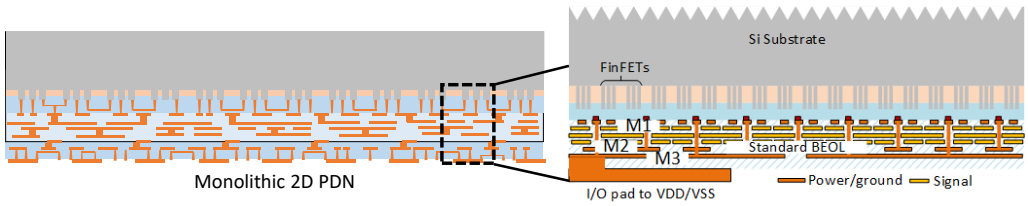


Figure 6.1: Conventional PDN cross-section for a monolithic 2D design

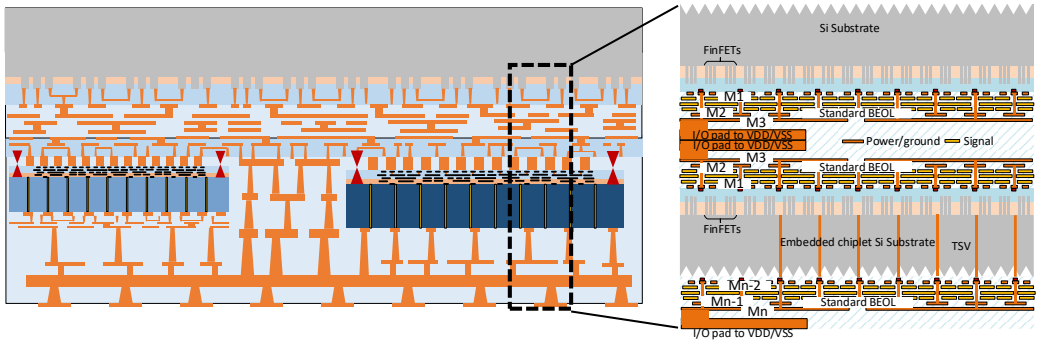


Figure 6.2: Proposed PDN cross-section for Polylithic 3D: BEOL-embedded chiplet integration

state PSN due to IR-drop on additional interconnects (resistive on-die PDN, TSVs, I/O bumps, etc) and inter-tier supply voltage variation. These effects lead to variations in the analog outputs of the memory arrays and the reference voltages in the ADCs, contributing to sensing errors in the ADCs. These errors can significantly impact CIM inference accuracy.

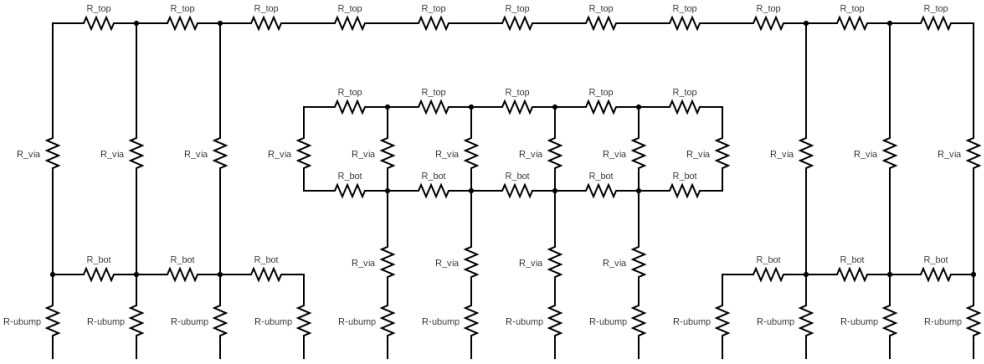
In the next few sections, we study some of these challenges and provide design optimization strategies for Polylithic 3D integration schemes.

6.3 Polylithic 3D PDN Design

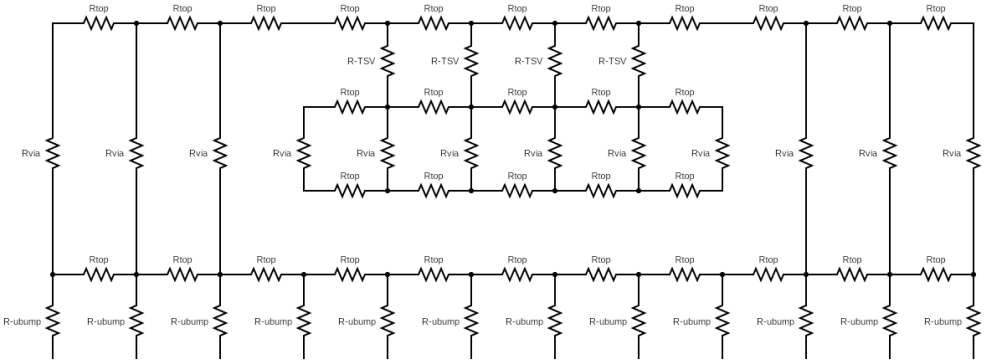
Figure 6.1 shows the PDN cross-section of a conventional monolithic 2-D IC. Figure 6.2 illustrates the cross-section representation of a polylithic 3D IC with two embedded dice under a primary top die.

We consider three cases for the polylithic 3D PDN. They are:

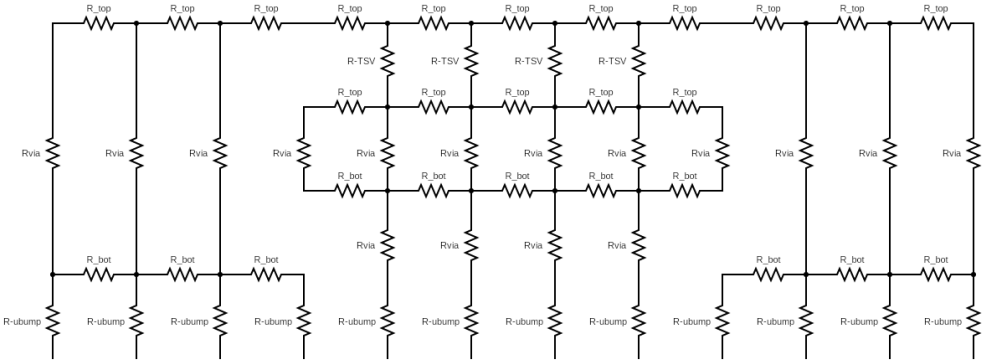
1. Case 1A: Polylithic 3D integration with BEOL vias below embedded tier connecting



(a) Case 1A: Polylithic 3D without TSV, with BEOL vias below embedded tier



(b) Case 1B: Polylithic 3D only TSVs, no BEOL vias below embedded tier



(c) Case 1C: Polylithic 3D with TSVs and BEOL vias below embedded tier

Figure 6.3: (a) Case 1A: Polylithic 3D without TSV, with BEOL vias below embedded tier, (b) Case 1B: Polylithic 3D only TSVs, no BEOL vias below embedded tier, and (c) Case 1C: Polylithic 3D with TSVs and BEOL vias below embedded tier

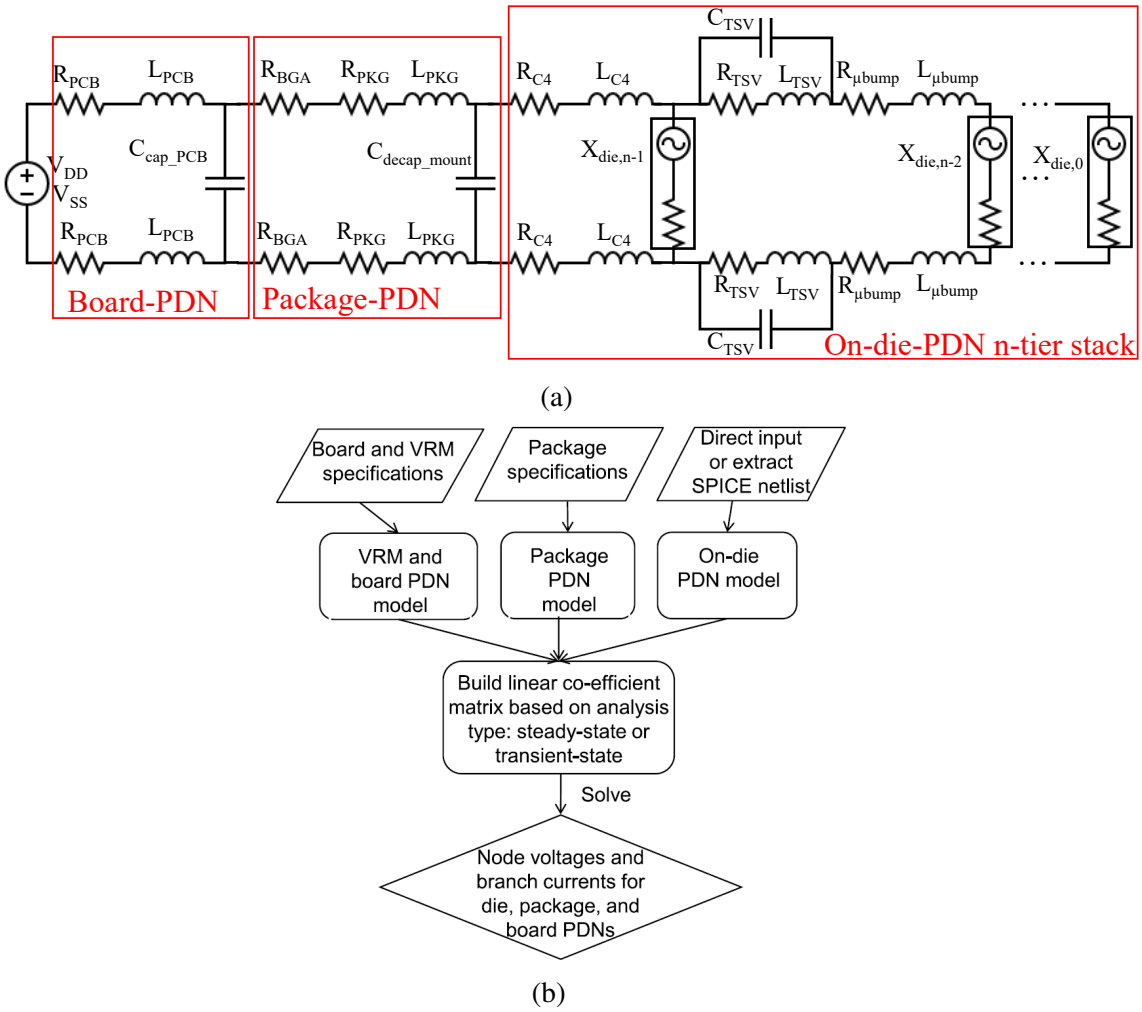


Figure 6.4: PDN modeling hierarchy: (a) Lumped model of the board-level PDN, distributed model of the package-level PDN, and the distributed model of the on-chip PDNs in an n-tier 3D stack including the TSVs and microbumps. (b) Flow diagram of the 3D PDN analysis showing different steps of the framework.

the embedded die to microbumps and no direct connections between embedded tier and top-tier. As shown in Figure 6.3a, this case includes one primary (top) tier and one embedded (bottom) tier. The top and bottom-tier PDNs are not assumed to be shared and thus are not connected with TSVs.

2. Case 1B: Polylithic 3D with TSVs connecting the top and bottom-tier and no BEOL vias connecting the embedded tier to microbumps. As seen in Figure 6.3b, this case also includes one primary (top) tier and one embedded (bottom) tier where the top

Table 6.1: Experimental Setup

PDN Model parameters	
Parameter	Value
On-die metal resistivity (ohm-m)	1.8e-8
On-die global wire Pitch/Width/Thickness (um)	39.5/17.5/7
On-die intermediate wire P/W/T (nm)	560/280/506
On-die local wire P/W/T (nm)	160/80/144
On-die decap density (nF/mm2)	335
μ-bumps pitch/R/L (um/m-ohm/pH)	40/30.9/11.1
C4 bump pitch/R/L (um/m-ohm/pH)	200/14.3/11
Package effective decap R/L/C (m-ohm/pH/uF)	541.5/220.7/52
Package resistivity/inductance (*m-ohm/mm*/ pH/mm)	1.2/24
BGA pitch/R/L (um/m-ohm/pH)	500/38/46
TSV R/L (m-ohm/pH)	54.2/77.78
PCB R/L (u-phm/pH)	166/21
PCB decap R/L/C (u-phm/nH/uF)	166/19.54/240
Die Size (mm2)	Top: 10mm $\times$ 10mm, Bottom/Embedded: 5mm $\times$ 5mm
Power (W)	Top: 1, 5, 10, 25 Bottom/Embedded: 1, 5, 25, 50
TSV Diameter (Pitch) (μm)	1 (2)
Bottom/embedded tier substrate thickness (TSV Height) (μm)	10 (TSV-stacked 3D baseline) 0.5 (embedded tier cases 1A, 1B, 1C)

and bottom-tier PDNs are assumed to be shared and are connected only with TSVs, i.e. the bottom-tier is not directly supplied power through a BEOL and microbumps underneath.

3. Case 1C: Polylithic 3D with TSVs connecting the embedded tier and top-tier and BEOL vias connecting the embedded tier to microbumps. As seen in Figure 6.3c, this case also includes one primary (top) tier and one embedded (bottom) tier where the top and bottom-tier PDNs are assumed to be shared and are connected with TSVs, and the bottom-tier is also directly supplied power through a BEOL and microbumps underneath.

6.4 Experimental Setup: 3D PDN modeling methodology

Figure 6.4a shows the PDN structure of an n-tier 3D stack. Figure 6.4b presents the flow for both steady-state and transient analyses and this is an updated version of the flow pre-

sented in [114]. The key contributions in this updated flow include support for modeling of emerging 3-D packaging architectures such as TSV and microbump-based 3-D and polylithic 3-D. We implement a distributed package-level PDN model, unlike previous efforts that assume a lumped package model, to reflect the spreading effects of current in the package and the coupling between different P/G bumps, especially when dice share package PDN planes. The flow begins with the generation of the RLC network models of the board, package, and the on-die PDNs. These models are subsequently combined to solve for nodal voltages and branch currents. Although we only present steady-state (IR-drop) analysis in this work, this flow can support transient-state analysis as well.

An ideal voltage regulator module (VRM) is assumed for board-level PDN and a lumped R/L network is used to model board-level current spreading. The equivalent series resistance and equivalent series inductance of board-level decoupling capacitors are included in the model. The package power/ground planes are modeled as two layers with the bottom layer connected to the motherboard by ball grid arrays and the top layer connected to an on-die PDN through C4 bumps. Each node in the two layers is connected to six adjacent nodes through an R-L pair, representing either package traces or inter-layer vias. Die-side decaps are assumed to be connected to the top package PDN layer.

The on-die PDN consists of several metal layers where the P/G wires are parallel to each other within a layer, and adjacent layers are orthogonal to each other. To better reflect the interleaved nature of the on-die PDN and capture the effect of on-die vias, the on-die PDN is modeled as a two-layer structure. The metal wires on each on-die PDN are typically uniformly distributed, but if the actual layout is non-uniform, our flow calculates the effective wire pitch and via density to reorganize the PDN layout [114]. For each on-die layer, we map a fine-granularity PDN layout to coarse mesh grids at a C4 bump granularity. The equivalent parallel resistance is calculated, for each coarse grid containing multiple vias and metal wires, and assigned using models described in [115]. All coarse PDN layers with x-axis and y-axis metal wires are mapped onto the top and bottom layers,

respectively. R_{via} is the effective resistance of vias between adjacent metal layers, and R_{TSV} is the resistance of TSVs between multiple 3D dice. Although this framework can be used to model both steady-state and transient PSN, in this work we focus on steady-state analyses.

6.5 Results

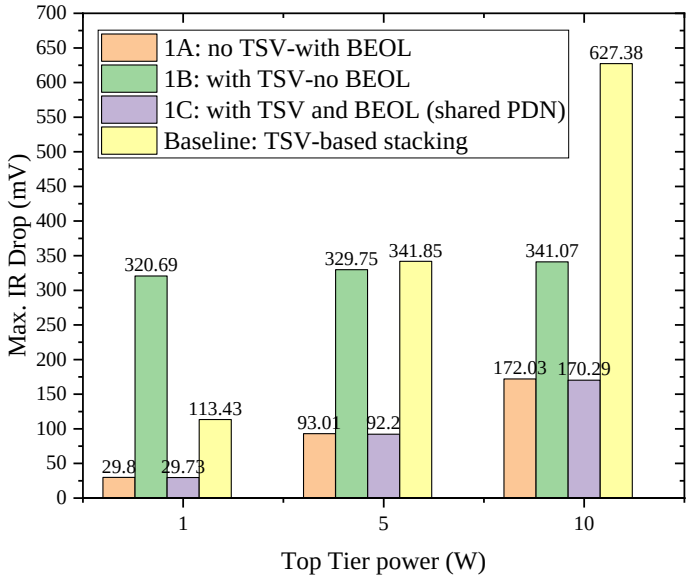


Figure 6.5: Top-tier IR-drop for polylithic 3D (cases 1A, 1B, 1C in orange, green, purple, respectively) as a function of the top-tier power. Bottom-tier power=25 W.

Benchmarking Polylithic 3D IR-drop

First we benchmark the IR-drop for polylithic 3D integration as a function of the top-tier power. The considered three cases are 1A, 1B, and 1C as described in section 6.3. Figure 6.5 illustrates the results for this study. The y-axis plots the maximum IR-drop for the top-tier, and the x-axis shows the considered top-tier powers (1 W, 5 W, 10 W). The bottom tier power was fixed at 25 W. Cases 1A and 1C show lower overall IR-drop compared to case 1B at 1 W top-tier power. This is because in case 1B, the embedded tier is only supplied power through TSVs connecting to the top-tier and not by the BEOL

and bumps below it. This leads to a longer power delivery path and overall path resistance compared to 1A and 1C where the embedded die are supplied power directly from the bumps and BEOL. Additionally, the maximum IR-drop for the top-tier increases with an increase in top-tier power.

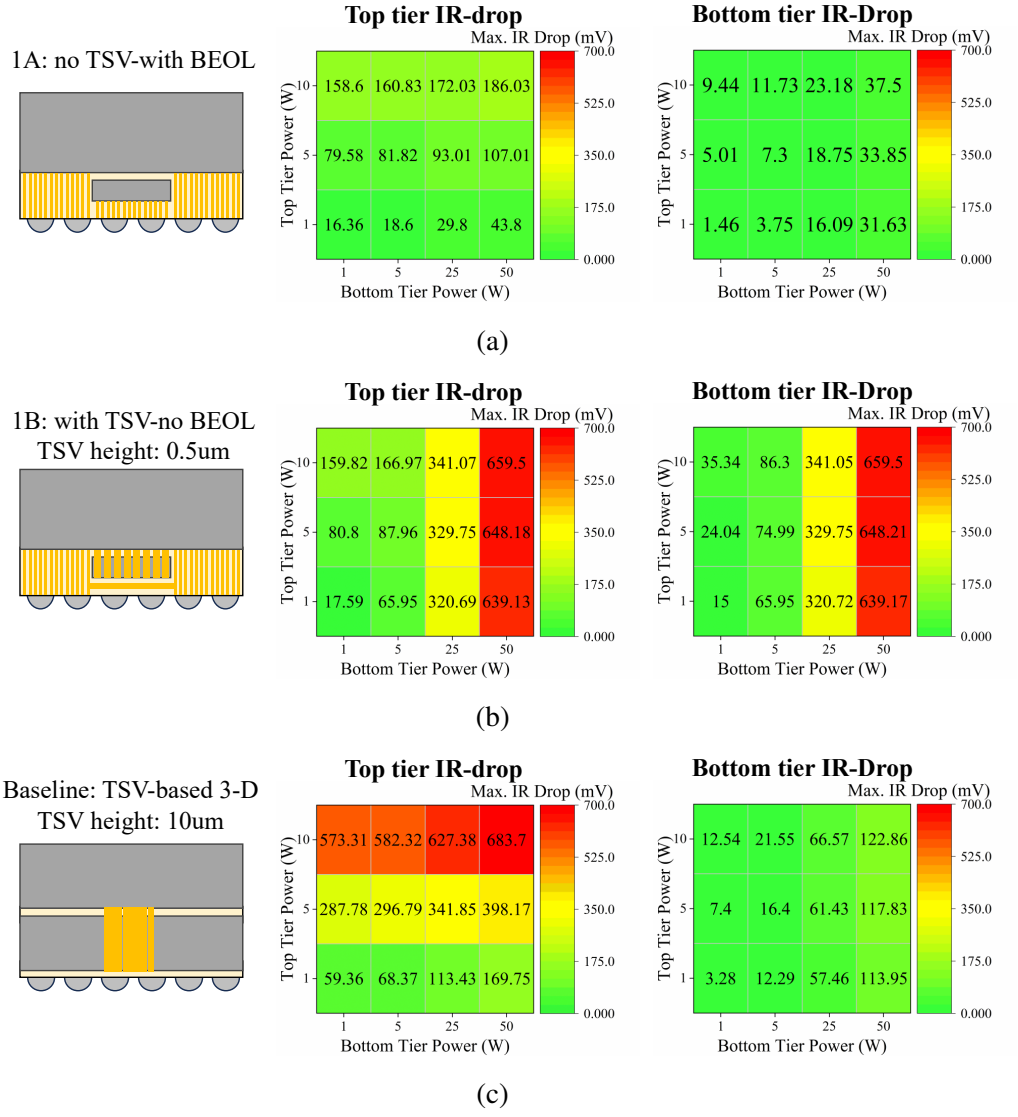


Figure 6.6: Maximum IR-drop for top and bottom-tiers in (a) Case 1A, (b) Case 1B, and (c) the baseline case.

Design Space Exploration: Tier Power

We studied IR-drop sensitivity of each tier with respect to the power dissipation of each tier to establish worst case limits and to establish optimal power combinations for each case. Figure 6.6 shows the results for this analysis. Figure 6.6a, Figure 6.6b, and Figure 6.6c show the maximum IR-drop for top and bottom-tiers in Case 1A, Case 1B, and the baseline case, respectively. Figure 6.6c (baseline) shows that the top-tier noise is sensitive to both the top and bottom-tier powers, and the sensitivity is higher towards the top-tier power. This is because power to the top-tier is delivered through TSVs that are highly resistive due to higher silicon thickness in the baseline case compared to cases 1A and 1B. Figure 6.6b illustrates that both the top and bottom-tier noise values are more sensitive to the bottom-tier power than to the top-tier power for case 1B. This is because the on-die switching loads do not have direct access to the direct PDN below the die overlap region. This leads to a significantly higher path resistance to deliver power to the bottom-tier, with tier 1 vias (shown in Figure 6.3b around the embedded tier) contributing highest to the path resistance. Figure 6.6a characterizes the power sensitivity of case 1A, and since in this case the two PDNs for top and bottom-tiers are not shared and effectively decoupled, we expect each tier's noise to be more sensitive to its own respective powers. The observations match our expectations as we learn that the top-tier noise is more sensitive to the top-tier power compared to the bottom-tier power, and vice versa. However, in the design space of powers considered in this study, the maximum noise for top-tier case 1A (top-tier power = 25W, bottom-tier power = 100W) was $\approx$23.2 lower than the baseline (top-tier power = 10W, bottom-tier power = 50W) and $\approx$20.5% lower than case 1B (top-tier power = 10W, bottom-tier power = 50W).

Chipletization of bottom-tier

Dividing the embedded tier into multiple dice and spacing the die out can help alleviate the worst case IR-drop. This is because the top-tier gets direct access to BEOL vias and

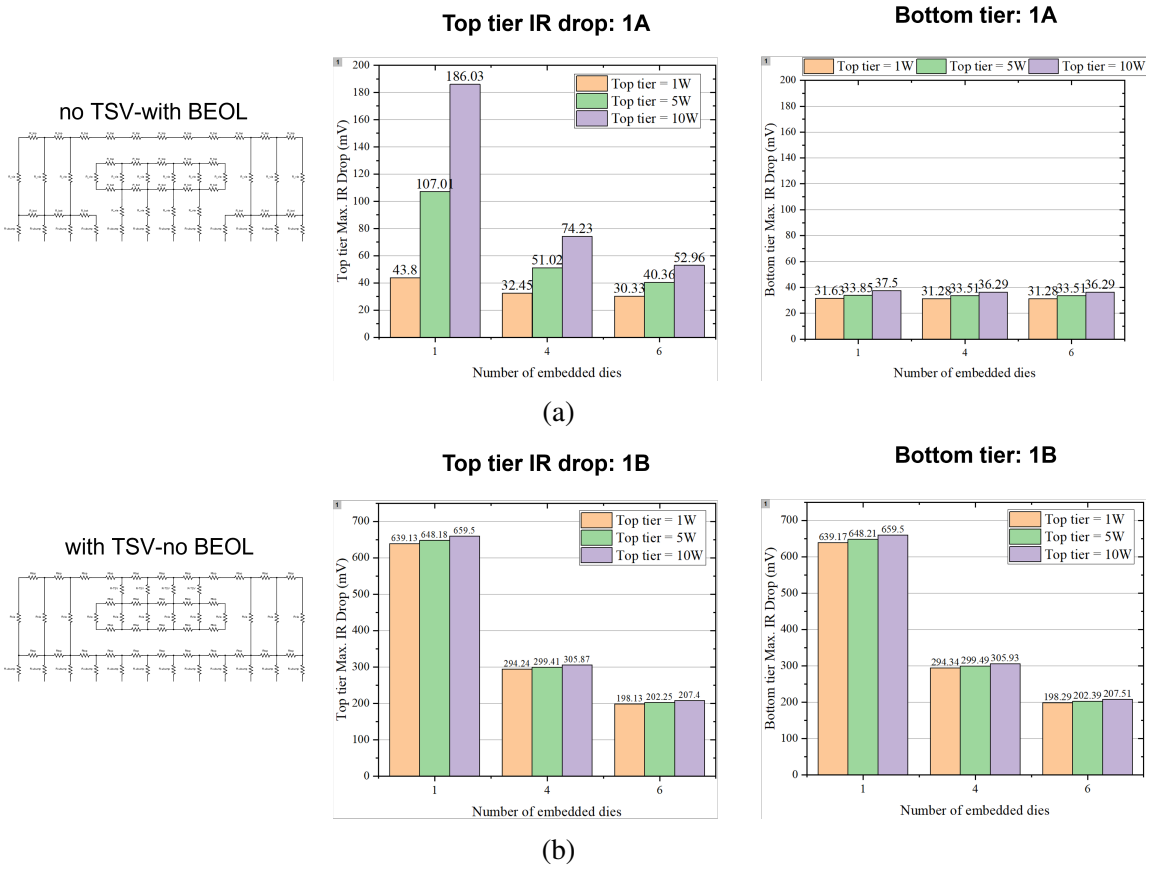


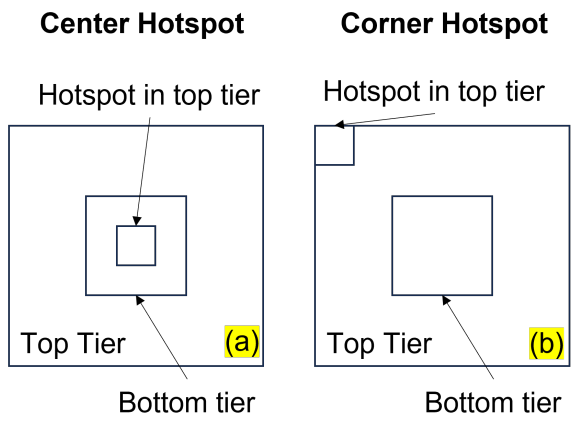
Figure 6.7: Chipletization of embedded tier for case (a) 1A and (b) 1B.

μbumps in the areas between the embedded dice. The results of this study are shown in Figure 6.7. Figure 6.7a shows the results for case 1A and Figure 6.7b for case 1B. We observe that for top-tier power = 10 W and bottom-tier power = 50W, going from 1 to 6 embedded chiplets, both top and bottom-tier noise for case 1B reduce from 73.3% of VDD to 23% of VDD (≈50% reduction). Similarly, the top-tier noise for case 1A reduces from 20.7% of VDD to 5.9% of VDD (≈14.8% reduction), and remains effectively unchanged (≈4% of VDD) for the bottom-tier. The reason for no change in case 1B bottom-tier is that the top and bottom PDNs are not shared. Thus splitting the bottom-tier and spacing out the dice does not impact the PDN path resitance for the embedded dice, however, it does reduce the path resistance for the top-tier.

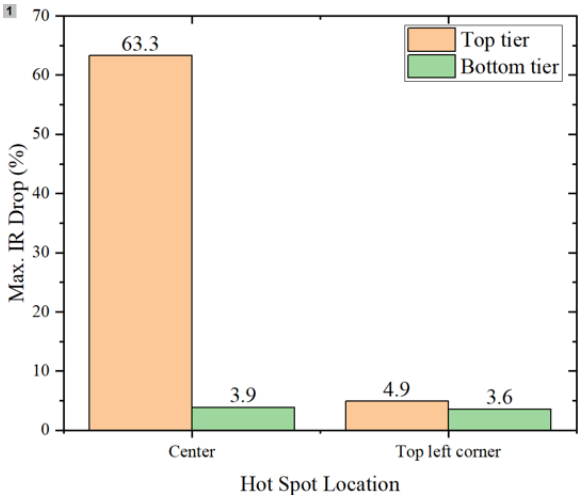
The best case top-tier IR-drop for case 1B (i.e. with 6 embedded chiplets) is higher (23%) than the worst case top-tier IR-drop for case 1A (20.7%). Additionally, there is marginal reduction in top-tier noise for case 1A going from 4 to 6 embedded chiplets. Thus, this experiment allows us to optimize the number of embedded chiplets, which were 4 for case 1A and 6 for case 1B.

Impact of hotspot

Hotspot power can have significant impact on the noise characteristics of a 3D-stacked IC. For this experiment, a Hotspot of size 1mm $\times$ 1mm was modeled on the top-tier (case 1A) with a power density of 100 W/cm^2. The bottom-tier power was set to 50 W (200 W/cm^2) while the top-tier power was set to 2 W (1 W hotspot + 1 W background). All other parameters for the top and bottom-tier dimensions were the same as those mentioned in Table 6.1. When shifting the hotspot from the center to a corner in the top-tier (as shown in Figure 6.8a), the top-tier IR-drop reduces from 63.2% to 4.9% (≈58% reduction), and the bottom-tier IR-drop remains effectively unchanged (Figure 6.8b). This demonstrates that hotspot floor-planning can be an effective strategy for PSN management.



(a)



(b)

Figure 6.8: Impact of hotspot location relative to the bottom-tier. (a) Hotspot locations considered, and (b) maximum IR-drop as a function of hotspot location.

6.6 Related Work

Related work from literature ([133], [100], [155], [156]) are summarized in Figure 6.9.

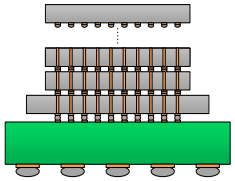
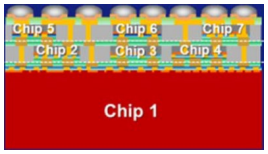
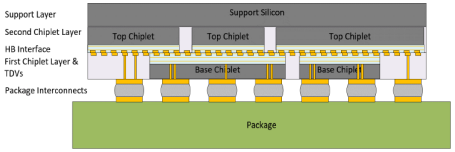
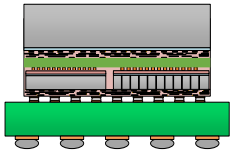
Attribute	H. Lee et al. [133]	Chen et al. [100], Cheng et al. [155]	A. Elsherbini et al. [156]	This work
Schematic				
Interconnection method	µ-bump and TSVs	Hybrid bonding D2W, W2W and TSVs	Hybrid bonding D2W, W2W and TSVs	BEOL vias and TSVs
Targeted I/O pitch (µm)	35	0.9	3 – 9	Goal: 0.5 – 5
I/O Density (count/mm^2)	1×	1512×	15× – 136×	49× – 4900×
Power Delivery Attributes	**TSV** can achieve **lower IR-drop vs BEOL via**. Stacking → extra IR-drop through TSV	TSV+TDV for lower IR-drop. IR-drop reduced from **2.3% to 1.2% using TSV array**	High aspect ratio **TDVs for lower parasitics and IR-drop** vs TSVs	Maximum **IR-drop** can be lowered **up to 3.3% vs 12.6% for baseline**
Die to Die Link Power (energy/bit) (A.U.)	1×	~ 0.05×	~ 0.05× – 0.1×	~ 0.02× – 0.67×

Figure 6.9: A summary of the salient features of related work in literature.

6.7 Conclusion

This work presents design optimization strategies for power delivery network (PDN) in polylithic 3-D integration. The proposed 3D polylithic architecture represents a densely integrated system divided into multiple device tiers where custom chiplets, such as power management IP, I/O drivers, and memory are embedded into the back-end of a base tier with extreme efficient signaling and large bandwidth density. We present a detailed design space exploration of the power supply noise effects in polylithic 3-D and reconstituted 3-D architectures. We propose three polylithic PDN designs and benchmark their IR drop as a function of tier power, number of embedded chiplets, hot-spot location, and TSV diameter and distribution to provide design limitations and insights.

CHAPTER 7
SUMMARY AND FUTURE WORK

This thesis aims to demonstrate methodologies for modeling and optimization of 2.5-D and 3-D integration architectures for compute-in-memory applications. The following sections highlight the contributions presented in this thesis and potential future directions.

7.1 Summary of the Work

The following research projects were completed and presented in the previous chapters:

- Chapter 2 demonstrated that including a PDN in the bridge-chip can provide significant reduction in DC-IR-drop, *Ldi/dt* noise, and high-frequency ripple compared to the baseline case of no PDN in the bridge-chip. Key takeaways are that 2.5-D designs with both smaller-width and larger-width bridge-chips can benefit from decoupling capacitors placed closer to the on-die PDN and that there is a trade-off between the bridge-chip size and MIM capacitor density. We quantify the impact of bridge-chip size and decoupling capacitor density in the bridge-chip on the maximum transient noise. Through a bridge-chip PDN design space exploration, insights are provided which can be useful for 2.5-D design convergence.

- Chapter 3 presented a comprehensive design-space exploration of power delivery network design for 3D heterogeneously integrated CIM hardware. A device-integration-application evaluation methodology is proposed to facilitate early design-space exploration and trade-offs between power delivery design parameters and CIM performance metrics. By co-optimizing across design hierarchies from packaging to circuits and devices, we present an areal-TSV 3D CIM design and compare it to a localized-TSV 3D implementation. For our assumed 3D CIM hardware, an areal

distribution of through-silicon vias (TSV) and microbumps, and a PSN-aware SAR-ADC fine-tuning achieves a 90% inference accuracy compared to 47% with a unoptimized 3D design at iso-area and iso-power. The insights provided could be useful for design convergence and performance modeling for edge intelligent 3D hardware.

- Chapter 4 presented a device-integration-application evaluation methodology that is used to quantify the impact of integration architectures on RRAM reliability for CIM applications. Two heterogeneous 3D logic-memory CIM accelerator designs - TSV-based 3D and Monolithic 3D-based integration of logic (7nm CMOS) and memory (22nm RRAM) tiers - were benchmarked against monolithic 2D and balanced integration design parameters were reported for maximized 3D CIM inference accuracy. For our assumed device, integration and application parameters, a 3-tier configuration provides a balanced design to achieve optimal system performance. The PP schemes are preferred for high-performance systems, with high operating temperature being a potential trade-off that can be improved with advanced thermal management and cooling architectures.

- Chapter 5 presented a back-end-embedded chiplet integration scheme for heterogeneous 3D integration, which is envisioned to combine the low EPB and high BWD benefits of monolithic 3D ICs with the integration flexibility of TSV-based 3D integration. A thermal study focusing on 3D SoC+ is presented to identify the thermal limits and challenges in such a scheme. For steady state operation of primary and embedded tiers, the impact of design parameters such as (a) embedded tier power density, (b) embedded tier thickness, and (c) BEOL thickness on maximum tier junction temperatures were evaluated with air-cooling and dual-sided cooling (DSC). Moreover, transient analysis was performed to estimate thermal coupling from the base silicon layer (tier 1) to the embedded tier. It was observed that the heat-sink proximity to an active tier affects the thermal coupling, with up to 20% reduction observed

with DSC.

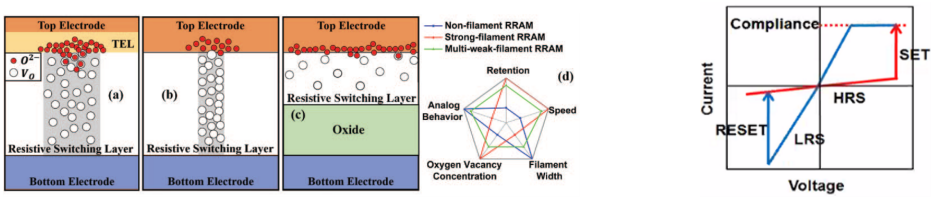
- Chapter 6 presented design optimization strategies for power delivery network (PDN) in polylithic 3-D integration. The proposed 3D polylithic architecture represents a densely integrated system divided into multiple device tiers where custom chiplets, such as power management IP, I/O drivers, and memory are embedded into the back-end of a base tier with extreme efficient signaling and large bandwidth density. We present a detailed design space exploration of the power supply noise effects in polylithic 3-D and reconstituted 3-D architectures. We propose three polylithic PDN designs and benchmark their IR drop as a function of tier power, number of embedded chiplets, hot-spot location, and TSV diameter and distribution to provide design limitations and insights.

7.2 Future Work

7.2.1 Evaluate Compute In-Memory (CIM) Inference and Training Accuracy with Multi-level RRAM Device Tier

Figure 7.1a shows three types of RRAM and their corresponding characteristics comparison [157]. Typical retention specification for non-volatile memory (NVM) devices is more than 10 years at 85 °C, and this has been met by binary RRAM devices. However, this requirement can become challenging to meet at elevated temperatures when NVM devices are integrated in 3D IC form factors with increased volumetric power. PSN management and thermal stability are of higher importance in multi-level switching to prevent the overlapping between adjacent low resistance state (LRS) levels [158] (Figure 7.1b).

Future directions could include a full system PDN and thermal design parameter analysis of polylithic 3D utilizing in-house developed and commercial frameworks. Another direction to explore could be a PDN and thermal co-simulation with multi-level RRAM BEOL device model to benchmark different 3D integration technologies and PSN manage-



(a) (b)

Figure 7.1: Three types of RRAM and the corresponding characteristics comparison. (a) Filamentary analog RRAM with multiple-weak-filaments; (b) conventional strong-filament based RRAM; (c) non-filamentary RRAM; (d) comparison of five specifications [157]. (e) I-V switching characteristics of a binary RRAM [159]

ment technologies with respect to the temporal degradation in CIM accelerator inference and training accuracy. A comprehensive reliability study of CIM accuracy change with design parameters such as increasing tier count, accurate power profile of BEOL-embedded 3D polylithic chiplets (I/O PHYs, RF-front end, etc.) is another possible vector for further work.

7.2.2 Extended PDN Benchmarking for Polylithic 3D Integration

The continued scaling of logic devices so far has ensured an increase in device density across generations of process nodes. However, on-chip Cu interconnects tend to scale poorly compared to devices due to the increased resistance of on-chip wires with reducing cross-sectional area [160]. The increase in wire and I/O resistance due to lower dimensions can increase the steady-state IR-drop in integrated circuits at advanced nodes. Moreover, with limitations in PPAC benefits from conventional device scaling, certain scaling boosters such as backside power delivery using buried power rails have been proposed for *More than Moore* integration.

Potential extensions of current work could include modeling chiplet-to-backend I/Os, proposed to be fabricated using Co atomic layer deposition (ALD), to perform static and dynamic power analyses for multi-tier polylithic 3-D and benchmark the static and dynamic power drop against that of TSV-based 3D. Analyzing the design trade-offs of backside power delivery for polylithic 3-D using buried power rails is another potential direction.

PDN-thermal co-design for polylithic 3-D integration is another interesting direction to explore.

7.2.3 Signal Channel Benchmarking for 3-D Heterogeneous Integration

As part of future work for benchmarking electrical performance of polylithic 3-D, die-to-die signal channel analysis can be explored. We present some preliminary results where repeater-based driver and receiver designs are used to model the digital signal channels for heterogeneous integration architectures. We present the results of a TSV-based 3-D design parameter simulation study and report the signaling latency, energy efficiency, and maximum areal bandwidth density of a TSV-based 3-D integration platform. In addition, we also report these metrics for polylithic 3-D and present preliminary results on the impact of process technology and temperature.

(a) (b)

Figure 7.2: Illustration of digital signal channel for (a) TSV and microbump-based and (b) polylithic and monolithic 3-D integration.

Table 7.1: Physical dimensions of each parameter of signaling models

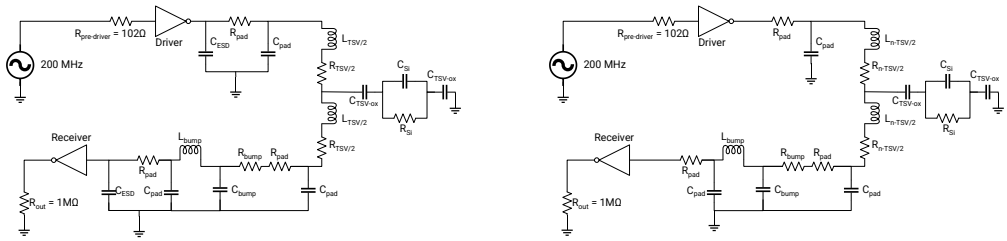
Parameter	Value
Technology Node	ASU PTM 7nm
TSV diameter (μm)	1 - 5
TSV height (μm)	50 - 300
TSV dioxide thickness (μm)	0.25
Microbump diameter (μm)	TSV diameter
Microbump height (μm)	TSV height
Microbump pitch (μm)	2*TSV diameter
Pad diameter (μm)	Microbump diameter*1.5
Pad height (μm)	Microbump height/2
Link wire length (mm)	1
Link wire pitch/thickness/width (μm)	1.6/2/0.8
ESD capacitance (fF)	50

Circuit models of digital signal channels in heterogeneous integration

The TSV-based 3-D and monolithic 3-D digital signal channels are illustrated in Figure 7.2a and Figure 7.2b, respectively. The signal channels consists of input/output (I/O) drivers and receivers, I/O pads, microbumps and chip-to-chip wires (TSVs [70, 98] or monolithic inter-layer vias [161, 162]).

The equivalent circuit models for the TSV-based and polylithic 3-D signal links are shown in Figure 7.3a and Figure 7.3b, respectively. The parasitics of the pads, microbumps and wires are included in these models. Monolithic 3-D is assumed to not require and ESD capacitances that are not included in the monolithic 3-D models. The vias models are based on the compact models described in [163, 164, 165]. For TSV-based 3-D, an ESD capacitor of 50 *fF* is added to both driver and receiver sides, and no ESD capacitance is assumed for monolithic and polylithic 3-D. A pre-driver of 102 Ω is included at the input to the driver and an output resistor of 1 *M*Ω is included as termination resistance at the receiver's output [166]. An optimal signal-to-ground (SG) TSV/microbump coupling case is considered.

The considered dimensions for I/Os and interconnects are summarized in Table 7.1. TSV specifications are assumed so as to establish the extreme case designs that are possible based on literature and fabrication constraints. The wire specifications are based on the dimensions of the top global wires from NCSU FreePDK 45 nm [167]. The wire routing configuration is assumed to use a fan-in approach as demonstrated in [92, 168], therefore



(a) (b)

Figure 7.3: Digital signal channel circuit for (a) TSV and microbump-based and (b) polylithic and monolithic 3-D integration.

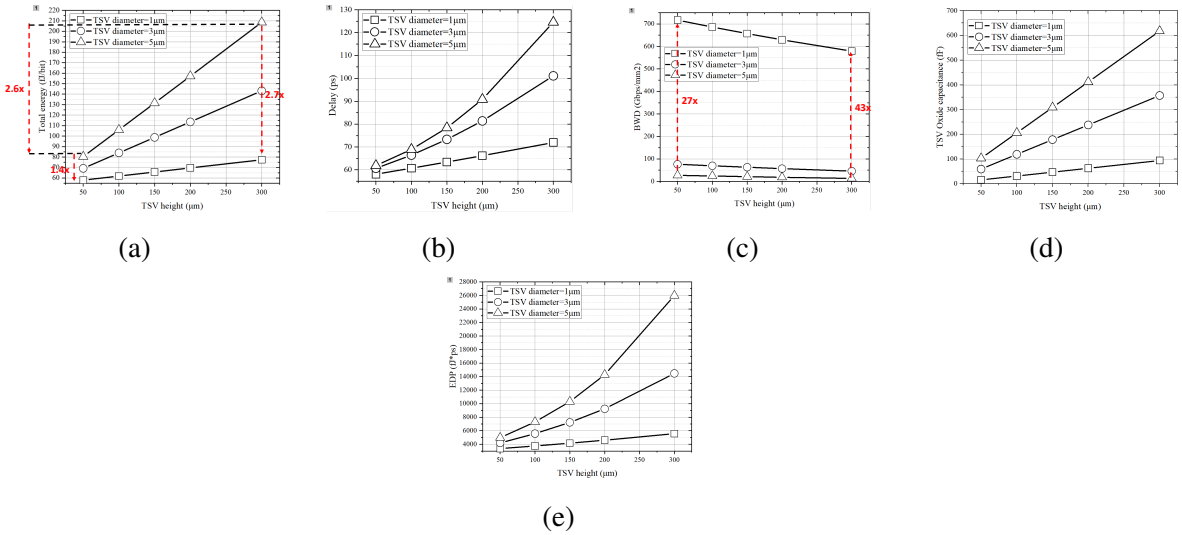


Figure 7.4: (a) Energy-per-bit, (b) Delay, (c) Bandwidth density, (d) TSV Oxide liner capacitance, and (e) Energy-delay product as a function of the TSV diameter and TSV height.

the wire pitch could be smaller than the microbump pitch. The models for parasitic estimation of each parameter are assumed from [66]. Microbumps are assumed to be cylindrical for a simplified design. Transceiver designs having multiple driver stages with a constant fan-out of 4 between stages were chosen [163]. The minimum-sized inverter that drives four identical inverters is tuned to achieve equal rise and fall times. The number of driver stages ranges from 1 to 4 stages in all simulations, and energy-delay-product (EDP) was used to optimize the number of driver stages [169]. We use a low-frequency digital signal input of 200 MHz [166] as we anticipate the signal channels are used in applications similar to Wide I/O spec [170].

TSV 3-D design parameter study

In this section, circuit models are developed in HSPICE netlists, and the 50%-to-50% propagation delay and total energy of the signal channels are simulated for TSV and microbump-based 3-D integration scenario. The device models are based on ASU PTM 7 nm HP library [84]. The version of HSPICE is PrimeSim HSPICE U-2023.03-SP1, and the BSIM model for 7 nm library is level 54 version 4.5.

Based on the design parameter specifications, the maximum data rate per link (F_{max}) was simulated using $6 \cdot \tau$ (τ is latency and $F_{max} > 1$ / $(6 \cdot \tau)$) settlement [86]. The bandwidth per unit area (Gbps/mm^2) (areal bandwidth density, BWD) is then calculated using the following equation:

$$ArealBWD = \frac{1}{P_{bump}^2} \cdot \frac{1}{(6 \cdot \tau)} (Gbps/mm^2) \tag{7.1}$$

where P_{bump} is the bump pitch with units of millimeter. We assume two rows of staggered bumps, of which half are for signals and the rest are for ground [86, 97, 93].

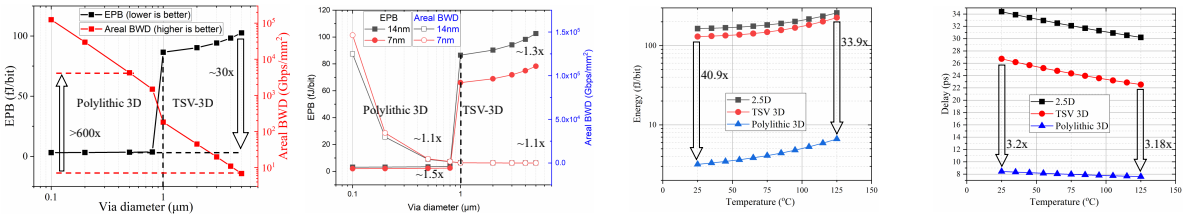
As shown in Figure 7.4a, a 1.4$\times$ reduction in energy was observed going from 5μm to 1μm TSV diameter at 50μm TSV height. A 2.7$\times$ reduction going from 5μm to 1μm TSV diameter at 300μm TSV height. A $\approx$2.6$\times$ reduction was oberved going from 300μm to 50μm TSV height @ 5μm TSV diameter.

As seen in Figure 7.4c, up to 43$\times$ improvement in bandwidth density (BWD) was observed going from 5μm to 1μm TSV diameter at 300μm TSV height. A $\approx$2$\times$ BWD improvement was observed going from 300μm to 50μm height at 5μm diameter due to lower RC parasitics.

Polylithic 3-D design parameter study

In this section, we present preliminary results from a signaling design parameter study for polylithic 3D integration as a function of via diameter, technology node, and junction temperature.

- Impact of Via Diameter: First we looked at benchmarking the energy-per-bit (EPB) and areal BWD as a function of via diameter. The results are summarized in Figure 7.5a. The figure shows that going from 0.5μm to 5μm via diameter the BWD for polylithic 3D was estimated to be 600$\times$ higher compared to TSV 3-D. The primary reasons are lower parasitic capacitance and due to the assumption that polylithic 3D



(a) (b) (c) (d)

Figure 7.5: Impact of design parameters on die-to-die signaling metrics: (a) Via Diameter, (b) Technology Scaling. (c) Energy-per-bit and (d) Delay as a function of temperature.

might not require ESD capacitance that is not included in the polylithic 3-D model (Figure 7.3b). This is similar to assumptions for monolithic 3D (without ESD), which is approximated as polylithic 3D since interconnect parasitics for polylithic 3D are expected to be in the range of BEOL-level parasitics.

- Impact of Technology Scaling: Next we look at the impact of technology scaling on the signaling metrics. In terms of node to node gains, the EPB reduced by $\approx$1.3$\times$ for TSV-3D and $\approx$1.5$\times$ for polylithic 3D, by swithcing from 14nm to 7nm. BWD in both cases increased by $\approx$1.1$\times$ in both cases.

- Impact of Temperature: Next we looked at benchmarking the signaling figures of merit for three integration architectures as a function of operating temperature. The die-to-die delay and energy-per-bit for polylithic 3D was estimated to be on an average 3.2 $\times$ lower and more than 30 $\times$ lower, respectively, compared to TSV 3D. The primary reasons are lower parasitic capacitance and with the assumption that polylithic 3D might not require ESD capacitance.

These results are preliminary and as part of future work, the parasitic models can be updated with measurement data to perform a detailed design space exploration. A summary of various proposed die-to-die signaling interfaces and a few 3-D integration hardware demonstration in literature is presented in Appendix A.

Appendices

APPENDIX A

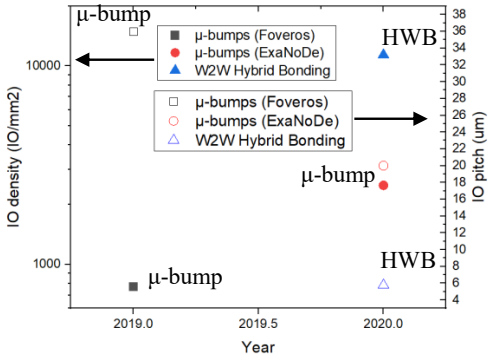
LITERATURE SURVEY

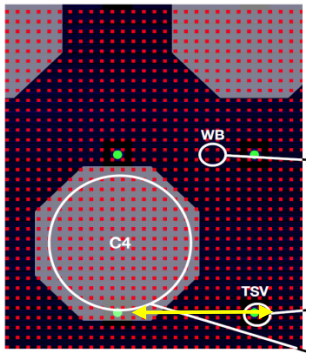
A summary of the salient features of various proposed heterogeneous integration interfaces in literature is presented in Figure A.1. A summary of 3-D integration demonstrations in literature using TSV along with hybrid bonding [68] or microbump-based [98, 171] stacking is presented in Figure A.2.

Standard	Source	Technology (2.5D/3D) (Electrical/Optical)	Bandwidth density (Linear/Areal)	Throughput / lane	Delay	PHY Energy / bit (pJ / bit)	Supply Voltage (V)
Advanced Interface Bus (AIB)	Intel	Silicon Bridge (2.5D) (E)	504 Gbps/mm (L)	Up to 2 Gbps	< 5 ns	0.85	
TeraPHY	Intel, Ayar Labs	Silicon Bridge (2.5D) (O)		Up to 5.12 Tbps			
Multi-Die I/O (MDIO)	Intel	Silicon Bridge (2.5D) (E)	1600 Gbps/mm (L)	Up to 5.4 Gbps		0.5	0.5
High Bandwidth Memory (HBM3)	JEDEC	μ-bumps and TSV (3D) (E)		4.8 Gbps		0.37	
XSR / USR	Rambus / OIF	(E)		112 Gbps			
LIPINCON	TSMC	Silicon Interposer (2.5D) (E)	536 Gbps/mm (L)	Up to 8 Gbps	< 14 ns	0.486	0.3
Bunch of Wires (BoW)	OCP / ODSA	2.5D (E)	1280 Gbps/mm (L)	Up to 16 Gbps	< 5 ns	0.7	
Infinity Fabric	AMD	MCM (E)		101.6 Gbps	< 9 ns	2	
Bandwidth Engine	Mosys	(E)		Up to 10.3125 Gbps	< 2.4 ns		
Foveros	Intel	μ-bumps and TSV (3D) (E)		0.5 Gbps		0.2	
AMBA CHI	Arm	Hybrid Bonding and TSV (3D) (E)	2276 – 3413 Gbps/mm^2 (A)	Up to 2.4 Gbps		0.013 – 0.021	0.8 – 1

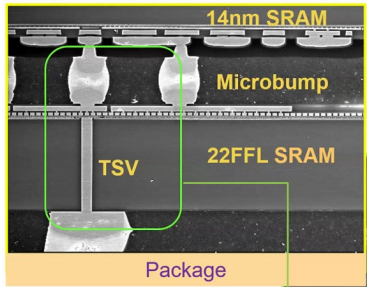
Figure A.1: A summary of the salient features of various proposed heterogeneous integration interfaces in literature.

Attribute	W2W Hybrid Bonding [68]	µ-bumps (Foveros) [98]	µ-bumps (ExaNoDe) [171]
Areal BWD (Gbps/mm^2)	27304	--	3000
EPB (fJ/bit)	21	150	--
Areal BWD/EPB (Gb/s)/(fJ/bit)	1300.19	--	--
I/O pitch (um)	5.76	36	20
IO/mm^2 (calculated from paper data)	11376.67	771.60	2500
pin speed (Gb/s)	2.4	0.5	1.2
TSV dia (um)	5	9	10
TSV pitch (um)	74.12	15	20

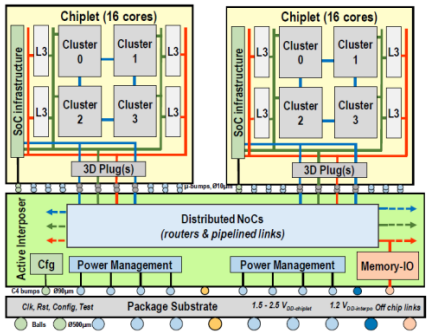




[68] S. Sinha, IEDM, 2020



[98] D. Ingerly, IEDM, 2019



[171] D. Dutoit, IEDM, 2020

Figure A.2: A summary of the die-to-die metrics of various 3-D integration demonstrations.

REFERENCES

[1] L. Su, "Delivering the future of high-performance computing," in *2019 IEEE Hot Chips 31 Symposium (HCS)*, 2019, pp. 1–43.

[2] K. Akarvardar and H.-S. P. Wong, "Technology prospects for data-intensive computing," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 92–112, 2023.

[3] *International roadmap for devices and systems: More moore*, (Accessed 16-November-2020).

[4] L. Song, Y. Zhuo, X. Qian, H. Li, and Y. Chen, "Graphr: Accelerating graph processing using reram," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, IEEE, 2018, pp. 531–543.

[5] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[6] *Nvidia tesla v100: The first tensor core gpu*, (Accessed 28-July-2023).

[7] Altera®, *Leveraging HyperFlex Architecture in Stratix 10 Devices to Achieve Maximum Power Reduction*, (Accessed 06-March-2023).

[8] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[9] S. Naffziger *et al.*, "Pioneering chiplet technology and design for the amd epyc™ and ryzen™ processor families : Industrial product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021, pp. 57–70.

[10] G. Chrysos, "Intel® xeon phi™ coprocessor-the architecture," *Intel Whitepaper*, vol. 176, no. 2014, pp. 43–50, 2014.

[11] *Geforce rtx 4060 family*, (Accessed 28-July-2023).

[12] A. Smith and N. James, "Amd instinct™ mi200 series accelerator and node architectures," in *2022 IEEE Hot Chips 34 Symposium (HCS)*, IEEE Computer Society, 2022, pp. 1–23.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, Jan. 2021, Number: 7840 Publisher: Nature Publishing Group.

[15] A. H. Atabaki *et al.*, "Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, pp. 349–354, Apr. 2018.

[16] *Pace: Photonic arithmetic computing engine*, (Accessed 27-July-2023).

[17] *Envise*, (Accessed 27-July-2023).

[18] K. Hosseini *et al.*, "5.12 tbps co-packaged fpga and silicon photonics interconnect i/o," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2022, pp. 260–261.

[19] M. Wade *et al.*, "TeraPHY: A Chiplet Technology for Low-Power, High-Bandwidth In-Package Optical I/O," *IEEE Micro*, vol. 40, no. 2, pp. 63–71, Mar. 2020, Conference Name: IEEE Micro.

[20] G. Singh *et al.*, "A review of near-memory computing architectures: Opportunities and challenges," in *2018 21st Euromicro Conference on Digital System Design (DSD)*, 2018, pp. 608–617.

[21] G. Singh *et al.*, "Nero: A near high-bandwidth memory stencil accelerator for weather prediction modeling," in *2020 30th International Conference on Field-Programmable Logic and Applications (FPL)*, 2020, pp. 9–17.

[22] K. Kara, C. Hagleitner, D. Diamantopoulos, D. Syrivelis, and G. Alonso, "High bandwidth memory on fpgas: A data analytics perspective," in *2020 30th International Conference on Field-Programmable Logic and Applications (FPL)*, 2020, pp. 1–8.

[23] R. Ben Abdelhamid and Y. Yamaguchi, "A Block-Based Systolic Array on an HBM2 FPGA for DNA Sequence Alignment," in *Applied Reconfigurable Computing. Architectures, Tools, and Applications*, F. Rincón, J. Barba, H. K. H. So, P. Diniz, and J. Caba, Eds., Cham: Springer International Publishing, 2020, pp. 298–313, ISBN: 978-3-030-44534-8.

[24] Z. Wang, H. Huang, J. Zhang, and G. Alonso, "Shuhai: Benchmarking high bandwidth memory on fpgas," in *2020 IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2020, pp. 111–119.

[25] M.-F. Chang, *Nonvolatile circuits for memory, logic, and artificial intelligence*, (Presentation 11-February-2018), 2018.

[26] L. Su and S. Naffziger, "1.1 innovation for the next decade of compute efficiency," in *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, 2023, pp. 8–12.

[27] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE Circuits and Systems Magazine*, vol. 21, no. 3, pp. 31–56, 2021.

[28] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature electronics*, vol. 1, no. 6, pp. 333–343, 2018.

[29] Y. Chen, "Reram: History, status, and future," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1420–1433, 2020.

[30] T. Kim and S. Lee, "Evolution of phase-change memory for the storage-class memory and beyond," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1394–1406, 2020.

[31] K. Garello, F. Yasin, and G. S. Kar, "Spin-orbit torque mram for ultrafast embedded memories: From fundamentals to large scale technology integration," in *2019 IEEE 11th International Memory Workshop (IMW)*, 2019, pp. 1–4.

[32] S. Ikegawa, F. B. Mancoff, J. Janesky, and S. Aggarwal, "Magnetoresistive random access memory: Present and future," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1407–1419, 2020.

[33] T. Mikolajick, U. Schroeder, and S. Slesazeck, "The past, the present, and the future of ferroelectric memories," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1434–1443, 2020.

[34] J. Tang *et al.*, "Ecram as scalable synaptic cell for high-speed, low-power neuromorphic computing," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 13.1.1–13.1.4.

[35] Y. Y. Chen *et al.*, "Balancing set/reset pulse for $> 10^{10}$ endurance in HfO_2/Hf 1t1r bipolar rram," *IEEE Trans. Electron Devices*, vol. 59, no. 12, pp. 3243–3249, 2012.

[36] C.-C. Chou *et al.*, "An n40 256k×44 embedded rram macro with sl-precharge sa and low-voltage current limiter to improve read and write performance," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 478–480.

[37] C.-F. Yang *et al.*, "Industrially applicable read disturb model and performance on mega-bit 28nm embedded rram," in *2020 IEEE Symposium on VLSI Technology*, 2020, pp. 1–2.

[38] C.-C. Chou *et al.*, "A 22nm 96kx144 rram macro with a self-tracking reference and a low ripple charge pump to achieve a configurable read window and a wide operating voltage range," in *2020 IEEE Symposium on VLSI Circuits*, 2020, pp. 1–2.

[39] P. Jain *et al.*, "13.2 a 3.6mb 10.1mb/mm2 embedded non-volatile reram macro in 22nm finfet technology with adaptive forming/set/reset schemes yielding down to 0.5v with sensing time of 5ns at 0.7v," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 212–214.

[40] J. Wu *et al.*, "A 40nm low-power logic compatible phase change memory technology," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 27.6.1–27.6.4.

[41] F. Arnaud *et al.*, "Truly innovative 28nm fdsoi technology for automotive microcontroller applications embedding 16mb phase change memory," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 18.4.1–18.4.4.

[42] Y.-D. Chih *et al.*, "13.3 a 22nm 32mb embedded stt-mram with 10ns read speed, 1m cycle write endurance, 10 years retention at 150°c and high immunity to magnetic field interference," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 222–224.

[43] L. Wei *et al.*, "13.3 a 7mb stt-mram in 22ffl finfet technology with 4ns read sensing time at 0.9v using write-verify-write scheme and offset-cancellation sensing technique," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 214–216.

[44] V. B. Naik *et al.*, "Manufacturable 22nm fd-soi embedded mram technology for industrial-grade mcu and iot applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 2.3.1–2.3.4.

[45] Y. J. Song *et al.*, "Demonstration of highly manufacturable stt-mram embedded in 28nm logic," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 18.2.1–18.2.4.

[46] M. Trentzsch *et al.*, "A 28nm hkmg super low power embedded nvm technology based on ferroelectric fets," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 11.5.1–11.5.4.

[47] S. Dünkel *et al.*, "A fefet based super-low-power ultra-fast embedded nvm technology for 22nm fdsoi and beyond," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 19.7.1–19.7.4.

[48] *Big trouble at 3nm*, (Accessed 27-July-2023).

[49] F. Sheikh, R. Nagisetty, T. Karnik, and D. Kehlet, "2.5d and 3d heterogeneous integration: Emerging applications," *IEEE Solid-State Circuits Magazine*, vol. 13, no. 4, pp. 77–87, 2021.

[50] *Intel core i7-4770k review: Haswell has landed*, (Accessed 28-July-2023).

[51] *Amd ryzen mobile 4000: Measuring renoir's die size*, (Accessed 28-July-2023).

[52] S. Lie, "Cerebras architecture deep dive: First look inside the hw/sw co-design for deep learning : Cerebras systems," in *2022 IEEE Hot Chips 34 Symposium (HCS)*, 2022, pp. 1–34.

[53] *Apple a16 die analysis*, (Accessed 28-July-2023).

[54] L. Su, *Opening plenary speaker - delivering the future of high-performance computing*, (Presentation date 15-July-2020), 2019.

[55] M. Rakowski *et al.*, "45nm cmos — silicon photonics monolithic technology (45clo) for next-generation, low power and high speed optical interconnects," in *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, 2020, pp. 1–3.

[56] G. Yeap *et al.*, "5nm cmos production technology platform featuring full-fledged euv, and high mobility channel finfets with densest 0.021μm2 sram cells for mobile soc and high performance computing applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 36.7.1–36.7.4.

[57] *The dark side of the semiconductor design renaissance – fixed costs soaring due to photomask sets, verification, and validation*, (Accessed 28-July-2023).

[58] R. Mahajan *et al.*, "Embedded multidie interconnect bridge—a localized, high-density multichip packaging interconnect," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 10, pp. 1952–1962, 2019.

[59] *International roadmap for devices and systems: Packaging integration*, (Accessed 18-November-2020).

[60] H. Azimi, *Compass virtual conference plenary: Heterogenous integration - an advanced packaging view*, (Presentation date 17-November-2020), 2020.

[61] C. Erdmann *et al.*, "A heterogeneous 3d-ic consisting of two 28 nm fpga die and 32 reconfigurable high-performance data converters," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 258–269, 2015.

[62] S. Naffziger, K. Lepak, M. Paraschou, and M. Subramony, "2.2 amd chiplet architecture for high-performance server and desktop products," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 44–45.

[63] R. Mahajan *et al.*, "Embedded multi-die interconnect bridge (emib) – a high density, high bandwidth packaging interconnect," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, 2016, pp. 557–565.

[64] W. Gomes *et al.*, "8.1 lakefield and mobility compute: A 3d stacked 10nm and 22ffl hybrid processor system in 12×12mm2, 1mm package-on-package," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 144–146.

[65] J. C. Lee *et al.*, "High bandwidth memory(HBM) with tsv technique," in *Proc. IEEE Int. SOC Conf.*, Oct. 2016, pp. 181–182.

[66] Y. Zhang, X. Zhang, and M. S. Bakir, "Benchmarking digital die-to-die channels in 2.5-D and 3-D heterogeneous integration platforms," *IEEE Trans. Electron Devices*, vol. 65, no. 12, pp. 5460–5467, Dec. 2018.

[67] *Imec magazine*, (Accessed 28-February-2020).

[68] S. Sinha *et al.*, "A high-density logic-on-logic 3dic design using face-to-face hybrid wafer-bonding on 12nm finfet process," in *Proc. IEEE Int. Elec. Devices Meeting*, 2020, pp. 15.1.1–15.1.4.

[69] J. C. Lee *et al.*, "High bandwidth memory(HBM) with tsv technique," in *Proc. IEEE Int. SOC Conf.*, Oct. 2016, pp. 181–182.

[70] W. Gomes *et al.*, "8.1 lakefield and mobility compute: A 3d stacked 10nm and 22ffl hybrid processor system in 12×12mm2, 1mm package-on-package," in *Proc. IEEE Solid State Circuits Conf.*, 2020, pp. 144–146.

[71] P. Oldiges *et al.*, "Chip power-frequency scaling in 10/7nm node," *IEEE Access*, vol. 8, pp. 154 329–154 337, 2020.

[72] L. England and I. Arsovski, "Advanced packaging saves the day! — how TSV technology will enable continued scaling," in *Proc. IEEE Int. Elec. Devices Meeting*, Dec. 2017, pp. 3.5.1–3.5.4.

[73] H. Wei, M. Shulaker, H. .-. P. Wong, and S. Mitra, "Monolithic three-dimensional integration of carbon nanotube FET complementary logic circuits," in *Proc. IEEE Int. Elec. Devices Meeting*, Dec. 2013, pp. 19.7.1–19.7.4.

[74] C. Liu and S. K. Lim, "A design tradeoff study with monolithic 3D integration," in *Proc. Int. Symp. on Quality Elec. Design*, Mar. 2012, pp. 529–536.

[75] E. Beyne, *Vlsi 2020 short course: Heterogeneous system partitioning and the 3d interconnect technology landscape*, (Presentation June-2020), 2020.

[76] W. Wan *et al.*, "A CIM chip based on resistive RAM," *Nature*, Aug. 2022.

[77] S. Sinha, X. Xu, M. Bhargava, S. Das, B. Cline, and G. Yeric, "Stack up your chips: Betting on 3d integration to augment moore's law scaling," *arXiv preprint arXiv:2005.10866*, 2020.

[78] Y. Peng, D. Petranovic, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, "Interdie coupling extraction and physical design optimization for face-to-face 3-d ics," *IEEE Transactions on Nanotechnology*, vol. 17, no. 4, pp. 634–644, 2018.

[79] Y. Zhang, M. O. Hossen, and M. S. Bakir, "Power delivery network modeling and benchmarking for emerging heterogeneous integration technologies," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 9, pp. 1825–1834, 2019.

[80] A. B. Kahng, S. Kang, S. Kim, K. Samadi, and B. Xu, "Power delivery pathfinding for emerging die-to-wafer integration technology," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2019, pp. 842–847.

[81] G. Bae *et al.*, "3nm gaa technology featuring multi-bridge-channel fet for low power and high performance applications," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 28.7.1–28.7.4.

[82] P. Batude *et al.*, "Advances in 3d cmos sequential integration," in *2009 IEEE International Electron Devices Meeting (IEDM)*, 2009, pp. 1–4.

[83] M. O. Hossen, B. Chava, G. Van der Plas, E. Beyne, and M. S. Bakir, "Power delivery network (pdn) modeling for backside-pdn configurations with buried power rails and μ tsvs," *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 11–17, 2020.

[84] H. Lee, R. Mahajan, F. Sheikh, R. Nagisetty, and M. Deo, "Multi-die integration using advanced packaging technologies," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, 2020, pp. 1–7.

[85] T. Singh *et al.*, "2.1 zen 2: The amd 7nm energy-efficient high-performance x86-64 microprocessor core," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 42–44.

[86] H. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H. .-. P. Wong, "Hfox based vertical resistive random access memory for cost-effective 3d cross-point architec-

ture without cell selector," in *2012 International Electron Devices Meeting*, 2012, pp. 20.7.1–20.7.4.

[87] P. Chen and S. Yu, "Reliability perspective of resistive synaptic devices on the neuromorphic system performance," in *2018 IEEE International Reliability Physics Symposium (IRPS)*, 2018, pp. 5C.4-1-5C.4–4.

[88] P. Sun *et al.*, "Thermal crosstalk in 3-dimensional RRAM crossbar array," *Sci Rep*, vol. 5, no. 1, pp. 1–9, Aug. 2015, Number: 1 Publisher: Nature Publishing Group.

[89] M. Villena *et al.*, "An in-depth study of thermal effects in reset transitions in hfo2 based rrams," *Solid-State Electronics*, vol. 111, pp. 47–51, 2015.

[90] X. Peng, S. Huang, Y. Luo, X. Sun, and S. Yu, "Dnn+neurosim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 32.5.1–32.5.4.

[91] P. K. Huang *et al.*, "Wafer level system integration of the fifth generation cowos®-s with high performance si interposer at 2500 mm2," in *IEEE Elec. Comp. and Tech. Conf.*, 2021, pp. 101–104.

[92] R. Mahajan *et al.*, "Embedded multi-die interconnect bridge (EMIB) – a high density, high bandwidth packaging interconnect," in *IEEE Elec. Comp. and Tech. Conf.*, May 2016, pp. 557–565.

[93] R. Swaminathan, *Advanced Packaging: Enabling Moore's Law's next frontier through heterogeneous integration*, (Accessed 06-March-2023).

[94] P. K. Jo *et al.*, "Multi-die polylithic integration enabled by heterogeneous interconnect stitching technology (hist)," in *Proc. IEEE Electrical Performance of Electronic Packaging and Systems*, 2018, pp. 11–13.

[95] K. Sikka *et al.*, "Direct bonded heterogeneous integration (dbhi) si bridge," in *IEEE Elec. Comp. and Tech. Conf.*, 2021, pp. 136–147.

[96] M.-F. Chen *et al.*, "System on integrated chips (soic(tm) for 3d heterogeneous integration," in *IEEE Elec. Comp. and Tech. Conf.*, 2019, pp. 594–599.

[97] R. Agarwal *et al.*, "3d packaging for heterogeneous integration," in *IEEE Elec. Comp. and Tech. Conf.*, 2022, pp. 1103–1107.

[98] D. B. Ingerly *et al.*, "Foveros: 3d integration and the use of face-to-face chip stacking for logic devices," in *Proc. IEEE Int. Elec. Devices Meeting*, 2019, pp. 19.6.1–19.6.4.

[99] K. Radhakrishnan *et al.*, "Power delivery for high-performance microprocessors—challenges, solutions, and future trends," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 11, no. 4, pp. 655–671, 2021.

[100] W. T. Chen *et al.*, "Design and analysis of logic-hbm2e power delivery system on cowos® platform with deep trench capacitor," in *IEEE Elec. Comp. and Tech. Conf.*, 2020, pp. 380–385.

[101] W. Liao *et al.*, "A manufacturable interposer mim decoupling capacitor with robust thin high-k dielectric for heterogeneous 3d ic cowos wafer level system integration," in *Proc. IEEE Int. Elec. Devices Meeting*, 2014, pp. 27.3.1–27.3.4.

[102] L. T. Su *et al.*, "Multi-chip technologies to unleash computing performance gains over the next decade," in *Proc. IEEE Int. Elec. Devices Meeting*, 2017, pp. 1.1.1–1.1.8.

[103] M. O. Hossen *et al.*, "Analysis of power delivery network (pdn) in bridge-chips for 2.5-d heterogeneous integration," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 12, no. 11, pp. 1824–1831, 2022.

[104] Y. Zhang, M. O. Hossen, and M. S. Bakir, "Power delivery network modeling and benchmarking for emerging heterogeneous integration technologies," *IEEE Trans. Compo. Packag. Manuf. Technol.*, pp. 1–1, 2019.

[105] Y. Zhang, M. O. Hossen, and M. S. Bakir, "Power delivery network benchmarking for interposer and bridge-chip-based 2.5-d integration," *IEEE Electron Device Letters*, vol. 39, no. 1, pp. 99–102, Jan. 2018.

[106] M. S. Gupta *et al.*, "Understanding voltage variations in chip multiprocessors using a distributed power-delivery network," in *Proc. Design, Automation and Test in Europe*, Nice, France, 2007, pp. 624–629.

[107] Yi-Min Jiang and Kwang-Ting Cheng, "Analysis of performance impact caused by power supply noise in deep submicron devices," in *Proc. ACM Design Automation Conf.*, 1999, pp. 760–765.

[108] D. Oh, "System level jitter characterization of high speed i/o systems," in *2012 IEEE International Symposium on Electromagnetic Compatibility*, Aug. 2012, pp. 173–178.

[109] H. Lin *et al.*, "84%-efficiency fully integrated voltage regulator for computing systems enabled by 2.5-d high-density mim capacitor," *IEEE Trans. VLSI Syst*, vol. 30, no. 5, pp. 661–665, 2022.

[110] Graphcore®, *Graphcore Uses TSMC 3D Chip Tech to Speed AI by 40%*, (Accessed 06-March-2023).

[111] R. Mahajan *et al.*, "Embedded multidie interconnect bridge—a localized, high-density multichip packaging interconnect," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 9, no. 10, pp. 1952–1962, 2019.

[112] S. Wu *et al.*, "Tr. and inf. with integers in deep neural networks," in *ICLR*, 2018.

[113] X. Peng, A. Kaul, M. S. Bakir, and S. Yu, "Heterogeneous 3-d integration of multitier compute-in-memory accelerators: An electrical-thermal co-design," *IEEE Trans. Electron Devices*, pp. 1–8, 2021.

[114] Y. Zhang *et al.*, "Power delivery network modeling and benchmarking for emerging heterogeneous integration technologies," *TCPMT*, vol. 9, no. 9, 2019.

[115] R. Zhang *et al.*, "Architec. implications of pads as a scarce resource," in *ISCA*, 2014.

[116] P. Chen and S. Yu, "Compact modeling of rram devices and its applications in 1t1r and 1s1r array design," *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4022–4028, 2015.

[117] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[118] S. D. Spetalnick *et al.*, "A 2.38 mcells/mm2 9.81 -350 tops/w rram compute-in-memory macro in 40nm cmos with hybrid offset/ioff cancellation and icell rblsl drop mitigation," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2023, pp. 1–2.

[119] Q. Liu *et al.*, "33.2 a fully integrated analog reram based 78.4tops/w compute-in-memory chip with fully parallel mac computing," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 500–502.

[120] B. Crafton, C. Talley, S. Spetalnick, J.-H. Yoon, and A. Raychowdhury, "Characterization and mitigation of ir-drop in rram-based compute in-memory," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 70–74.

[121] M. Lanza *et al.*, "Recommended methods to study resistive switching devices," *Advanced Electronic Materials*, vol. 5, no. 1, p. 1 800 143, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aelm.201800143.

[122] P. Sun *et al.*, "Thermal crosstalk in 3-dimensional rram crossbar array," *Scientific reports*, vol. 5, no. 1, pp. 1–9, 2015.

[123] B. Govoreanu *et al.*, "10×10nm2 hf/hfox crossbar resistive ram with excellent performance, reliability and low-energy operation," in *Proc. IEEE Int. Elec. Devices Meeting*, 2011, pp. 31.6.1–31.6.4.

[124] W. Shim, J. Meng, X. Peng, J.-s. Seo, and S. Yu, "Impact of multilevel retention characteristics on rram based dnn inference engine," in *Proc. IEEE Int. Reliability Physics Sym.*, 2021, pp. 1–4.

[125] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recog.*, 2016, pp. 770–778.

[126] A. Kaul, Y. Luo, X. Peng, S. Yu, and M. S. Bakir, "Thermal reliability considerations of resistive synaptic devices for 3d cim system performance," in *2021 IEEE International 3D Systems Integration Conference (3DIC)*, 2021, pp. 1–5.

[127] A. Kaul *et al.*, "Thermal modeling of 3d polylithic integration and implications on beol rram performance," in *Proc. IEEE Int. Elec. Devices Meeting*, 2020, pp. 13.1.1–13.1.4.

[128] Y. Zhang, Y. Zhang, and M. S. Bakir, "Thermal design and constraints for heterogeneous integrated chip stacks and isolation technology using air gap and thermal bridge," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 4, no. 12, pp. 1914–1924, 2014.

[129] H.-S. P. Wong *et al.*, "Metal–oxide rram," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[130] A. Kaul *et al.*, "Beol-embedded 3d polylithic hb integration: Thermal and interconnection considerations," in *IEEE Elec. Comp. and Tech. Conf.*, 2020, pp. 1459–1467.

[131] Y.-H. Chen *et al.*, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," in *Proc. IEEE Int. Symp. on Comp. Arch.*, ser. ISCA '16, Seoul, Republic of Korea: IEEE Press, 2016, pp. 367–379, ISBN: 9781467389471.

[132] M. M. Shulaker *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, vol. 547, no. 7661, pp. 74–78, 2017.

[133] H.-J. Lee *et al.*, "Multi-die integration using advanced packaging technologies," in *Proc. IEEE Custom Int. Circuits Conf.*, 2020, pp. 1–7.

[134] C. Okoro *et al.*, "Analysis of the induced stresses in silicon during thermcompression cu-cu bonding of cu-through-vias in 3d-sic architecture," in *2007 Proceedings 57th Electronic Components and Technology Conference*, 2007, pp. 249–255.

[135] K. Athikulwongse, A. Chakraborty, J.-S. Yang, D. Z. Pan, and S. K. Lim, "Stress-driven 3d-ic placement with tsv keep-out zone and regularity study," in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2010, pp. 669–674.

[136] X. Sun *et al.*, "Pcm-based analog compute-in-memory: Impact of device non-idealities on inference accuracy," *IEEE Trans. Electron Devices*, vol. 68, no. 11, pp. 5585–5591, 2021.

[137] D. Tuckerman and R. Pease, "High-performance heat sinking for vlsi," *IEEE Electron Device Letters*, vol. 2, no. 5, pp. 126–129, 1981.

[138] T. Brunschwiler *et al.*, "Towards cube-sized compute nodes: Advanced packaging concepts enabling extreme 3d integration," in *Proc. IEEE Int. Elec. Devices Meeting*, 2017, pp. 3.7.1–3.7.4.

[139] T. E. Sarvey *et al.*, "Microfluidic cooling of a 14-nm 2.5-d fpga with 3-d printed manifolds for high-density computing: Design considerations, fabrication, and electrical characterization," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 9, no. 12, pp. 2393–2403, 2019.

[140] T. W. Wei *et al.*, "Experimental and numerical investigation of direct liquid jet impinging cooling using 3d printed manifolds on lidded and lidless packages for 2.5d integrated systems," vol. 164, p. 114 535, 2020.

[141] R. van Erp *et al.*, "Co-designing electronics with microfluidics for more sustainable cooling," *Nature*, vol. 585, no. 7824, pp. 211–216, 2020.

[142] S. K. Rajan *et al.*, "Monolithic microfluidic cooling of a heterogeneous 2.5-d fpga with low-profile 3-d printed manifolds," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 11, no. 6, pp. 974–982, 2021.

[143] S. S. Kumar *et al.*, "Fighting dark silicon: Toward realizing efficient thermal-aware 3-d stacked multiprocessors," *IEEE Trans. VLSI Syst*, vol. 25, no. 4, pp. 1549–1562, 2017.

[144] R. Mathur *et al.*, "Thermal-aware design space exploration of 3-d systolic ml accelerators," *IEEE Jour. on Expl. Solid-State Comp. Dev. and Cir.*, vol. 7, no. 1, pp. 70–78, 2021.

[145] M. Scheuermann *et al.*, "Thermal analysis of multi-layer functional 3d logic stacks," in *Proc. IEEE Int. Conf. on 3D System Integration*, 2016, pp. 1–4.

[146] *Heterogeneous integration roadmap (HIR): Chapter 20: Thermal*, (Accessed 28-February-2020).

[147] H. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H. .-. P. Wong, "HfOx based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector," in *Proc. IEEE Int. Elec. Devices Meeting*, Dec. 2012, pp. 20.7.1–20.7.4.

[148] Y. Zhang, T. E. Sarvey, and M. S. Bakir, "Thermal evaluation of 2.5-D integration using bridge-chip technology: Challenges and opportunities," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 7, no. 7, pp. 1101–1110, 2017.

[149] T. E. Sarvey *et al.*, "Monolithic integration of a micropin-fin heat sink in a 28-nm FPGA," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 7, no. 10, pp. 1617–1624, 2017.

[150] T. Brunschwiler *et al.*, "Towards cube-sized compute nodes: Advanced packaging concepts enabling extreme 3D integration," in *Proc. IEEE Int. Elec. Devices Meeting*, 2017, pp. 3–7.

[151] R. Mahajan and B. Sankman, "3D packaging architectures and assembly process design," in *3D Microelectronic Packaging*, Springer, 2017, pp. 17–46.

[152] A. Delan, M. Rennau, S. Schulz, and T. Gessner, "Thermal conductivity of ultra low-k dielectrics," *Microelectronic Engineering*, vol. 70, no. 2-4, pp. 280–284, 2003.

[153] A. Jain, R. E. Jones, R. Chatterjee, and S. Pozder, "Analytical and numerical modeling of the thermal performance of three-dimensional integrated circuits," *IEEE Trans. Compo. Packag. Manuf. Technol.*, vol. 33, no. 1, pp. 56–63, 2009.

[154] S. Natarajan *et al.*, "A 14nm logic technology featuring 2nd-generation finfet, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm2 sram cell size," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 3.7.1–3.7.3.

[155] Y.-K. Cheng *et al.*, "Next-generation design and technology co-optimization (dtco) of system on integrated chip (soic) for mobile and hpc applications," in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 41.3.1–41.3.4.

[156] A. Elsherbini *et al.*, "Enabling next generation 3d heterogeneous integration architectures on intel process," in *2022 International Electron Devices Meeting (IEDM)*, 2022, pp. 27.3.1–27.3.4.

[157] M. Zhao *et al.*, "Investigation of statistical retention of filamentary analog rram for neuromophic computing," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 39.4.1–39.4.4.

[158] E. Pérez, C. Zambelli, M. K. Mahadevaiah, P. Olivo, and C. Wenger, "Toward reliable multi-level operation in rram arrays: Improving post-algorithm stability and assessing endurance/data retention," *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 740–747, 2019.

[159] S. Yu, *Ngc urs talk: Cognitive microsystem prototyping and benchmarking*, (Presentation date 21-October-2020), 2020.

[160] M. H. van der Veen *et al.*, "Barrier/liner stacks for scaling the cu interconnect metallization," in *2016 IEEE International Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC)*, 2016, pp. 28–30.

[161] M. D. Bishop, H.-S. P. Wong, S. Mitra, and M. M. Shulaker, "Monolithic 3-D integration," *IEEE Micro*, vol. 39, no. 6, pp. 16–27, 2019.

[162] R. Abbaspour, D. Brown, and M. Bakir, "Fabrication and electrical characterization of sub-micron diameter through-silicon via for heterogeneous three-dimensional integrated circuits," *Journal of Micromechanics and Microengineering*, vol. 27, no. 2, p. 025 011, 2017.

[163] X. Wu *et al.*, "Electrical characterization for intertier connections and timing analysis for 3-d ics," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 1, pp. 186–191, 2012.

[164] J. Kim *et al.*, "High-frequency scalable electrical model and analysis of a through silicon via (tsv)," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 1, no. 2, pp. 181–195, 2011.

[165] X. Zhang *et al.*, "Impact of on-chip interconnect on the performance of 3-d integrated circuits with through-silicon vias: Part ii," *IEEE Transactions on Electron Devices*, vol. 63, no. 6, pp. 2510–2516, 2016.

[166] M. A. Karim, P. D. Franzon, and A. Kumar, "Power comparison of 2d, 3d and 2.5d interconnect solutions and power optimization of interposer interconnects," in *2013 IEEE 63rd Electronic Components and Technology Conference*, 2013, pp. 860–866.

[167] *Nangate, nangate 45nm open cell library*, Available: https://eda.ncsu.edu/freepdk/freepdk45/, Accessed 23-August-2023.

[168] H. Braunisch, A. Aleksov, S. Lotz, and J. Swan, "High-speed performance of silicon bridge die-to-die interconnects," in *2011 IEEE 20th Conference on Electrical Performance of Electronic Packaging and Systems*, 2011, pp. 95–98.

[169] S. Abbaspour, M. Pedram, and P. Heydari, "Optimizing the energy-delay-ringing product in on-chip cmos line drivers," in *Fourth International Symposium on Quality Electronic Design, 2003. Proceedings.*, 2003, pp. 261–266.

[170] *Jedec standard*, Available: https://www.jedec.org/standards-documents/docs/jesd229, Accessed 23-August-2023.

[171] D. Dutoit *et al.*, "How 3d integration technologies enable advanced compute node for exascale-level high performance computing?" In *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 15.3.1–15.3.4.

VITA

Ankit Kaul was born in New Delhi, India, in January 1992. He received his B.E. degree in Electrical and Electronics Engineering from R.V. College of Engineering, Bangalore, India in 2013. He also received the M.S. degree in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta, USA in 2018, where he is currently a Ph.D. candidate.

From 2018 to present, he has been a graduate research assistant at the Georgia Tech. Integrated 3-D System laboratory supervised by Dr. Muhannad S. Bakir. His primary research is in the area of 2.5-D and 3-D IC design, modeling, and optimization with a concentration on thermal, power delivery, and die-to-die signaling analysis for 3-D compute-in memory hardware. His other research interests include system technology co-optimization between silicon, package, and architecture configurations, and non-volatile memory/logic for heterogeneous integration.

Outside work he enjoys biking, running, hiking, and playing badminton.