UC Santa Barbara
UC Santa Barbara Electronic Theses and Dissertations

Title
Cost-Driven Integration Architectures for Multi-Die Silicon Systems

Permalink
https://escholarship.org/uc/item/6h4244b8

Author
Randall, Dylan Stow

Publication Date
2020

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Cost-Driven Integration Architectures for Multi-Die Silicon Systems

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Dylan Stow Randall

Committee in charge:

Professor Yuan Xie, Chair
Professor Timothy Sherwood
Professor Li-C. Wang
Professor Zheng Zhang

December 2020

The Dissertation of Dylan Stow Randall is approved.

Professor Timothy Sherwood

Professor Li-C. Wang

Professor Zheng Zhang

Professor Yuan Xie, Committee Chair

September 2020

Cost-Driven Integration Architectures for Multi-Die Silicon Systems

Copyright © 2020

by

Dylan Stow Randall

This dissertation is dedicated to my family for all of their love and support. To my grandparents, parents, and wife: thank you for encouraging me to push myself and pursue my dreams.

Acknowledgements

I am very grateful for all of the support I received during this graduate education.

First, I want to thank my advisor, Professor Yuan Xie, for providing me with the resources and guidance I needed to complete this work. Most of all, I am thankful for the autonomy that Professor Xie entrusted in me, allowing me to pursue my research interests while also moving around the country, collaborating with industry, and starting my family. Further, he pushed me outside my comfort zone to work hard, investigate challenging topics, and share my findings around the world. This research builds upon the methods and findings of decades of his prior work, so I could not have completed it without him.

I would like to thank my industry mentors for providing thoughtful guidance into my research and for showing me the future of semiconductor technology. Dr. Gabriel H. Loh, Dr. Sudhanva Gurumurthi, Dr. Amin Farmahini-Farahani, Dr. Michael Ignatwoski, and Dr. Tanniya Siddiqua all taught me a great deal during our research collaborations, in and out of my time at AMD Research, as did Dr. Kambiz Samadi and Dr. Yang Du during my time at Qualcomm Research.

I was fortunate to collaborate with many intelligent, curious, and, hard-working researchers at UCSB Scalable Energy-efficient Architecture Lab, especially Dr. Itir Akgun, Peng Gu, Russell Barnes, Liu Liu, Dr. Shuangchen Li, Dr. Xing Hu, Wenqin Huangfu, Tianqi Tang, and Prashansa Mukim. The researchers in SEAL provided me with invaluable inspiration and insight and taught me many topics outside of my research interests that I will take with me into the rest of my career.

Finally, I would like to thank my parents and brother for encouraging me to follow my goals, my grandparents for providing an excellent education, my wife for always being by my side when I need support, and my daughters for providing infinite inspiration.

Curriculum Vitæ
Dylan Stow Randall

Education

2020	Ph.D. in Electrical and Computer Engineering (Expected), University of California, Santa Barbara.
2017	M.S. in Electrical and Computer Engineering, University of California, Santa Barbara.
2013	B.S. in Engineering, Harvey Mudd College.

Publications

[J1]. **Dylan Stow**, Amin Farmahini-Farahani, Sudhanva Gurumurthi, Michael Ignatowski, Yuan Xie. "Power Profiling of Modern Die-Stacked Memory." *IEEE Computer Architecture Letters (CAL)*, 2019.

[J2]. Itir Akgun, **Dylan Stow**, Yuan Xie. "Network-on-Chip Design Guidelines for Monolithic 3-D Integration." *IEEE Micro*, 2019.

[J3]. Xing Hu, **Dylan Stow**, Yuan Xie. "Die Stacking is Happening." *IEEE Micro*, 2018.

[C1]. **Dylan Stow**, Itir Akgun, Wenqin Huangfu, Yuan Xie, Xueqi Li, Gabriel H. Loh. "Efficient System Architecture in the Era of Monolithic 3D: Dynamic Inter-Tier Interconnect and Processing-in-Memory (Invited)." *Design Automation Conference (DAC)*, 2019.

[C2]. **Dylan Stow**, Itir Akgun, Yuan Xie. "Investigation of Cost-Optimal Network-on-Chip for Passive and Active Interposer Systems." *IEEE/ACM International Workshop on System-Level Interconnect Prediction (SLIP)*, 2019.

[C3]. Peng Gu, **Dylan Stow**, Prashana Mukim, Shuangchen Li, Yuan Xie. "Cost-Efficient 3D Integration to Hinder Reverse Engineering During and After Manufacturing." *IEEE Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*, 2018.

[C4]. **Dylan Stow**, Yuan Xie. "Navigating the Die-Integration Design Space: System Yield and Cost Analysis of 3D and 2.5D Packaging." *Government Microcircuit Applications and Cirtical Technology Conference (GOMACTech)*, 2018.

[C5]. Peng Gu, **Dylan Stow**, Yuan Xie. "Circuit Obfuscation and Thermal Side-Channel Masking using 3D Die-Stacking." *Government Microcircuit Applications and Cirtical Technology Conference (GOMACTech)*, 2018.

[C6]. **Dylan Stow**, Yuan Xie, Tanniya Siddiqua, Gabriel H. Loh. "Cost-Effective Design of Scalable High-Performance Systems Using Active and Passive Interposers." *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2017.

[C7]. Jaya Dofe, Peng Gu, **Dylan Stow**, Qiaoyan Yu, Eren Kursun, Yuan Xie. "Security Threats and Countermeasures in Three-Dimensional Integrated Circuits." *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2017.

[C8]. **Dylan Stow**, Itir Akgun, Russell Barnes, Peng Gu, Yuan Xie. "Cost Analysis and Cost-Driven IP Reuse Methodology for SoC Design Based on 2.5D/3D Integration." *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2016.

[C9]. Peng Gu, **Dylan Stow**, Russell Barnes, Eren Kursun, Yuan Xie. "Thermal-aware 3D Design for Side-Channel Information Leakage." *IEEE International Conference on Computer Design (ICCD)*, 2016.

[C10]. **Dylan Stow**, Itir Akgun, Russell Barnes, Peng Gu, Yuan Xie. "Cost and Thermal Analysis of High-Performance 2.5D and 3D Integrated Circuit Design Space." *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2016.

[C11]. Peng Gu, Shuangchen Li, **Dylan Stow**, Russell Barnes, Liu Liu, Yuan Xie, Eren Kursun. " Leveraging 3D Technologies for Hardware Security: Opportunities and Challenges." *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2016.

Abstract

Cost-Driven Integration Architectures for Multi-Die Silicon Systems

by

Dylan Stow Randall

The consistent cadence of Moore's Law has long driven improvement in compute performance by delivering increased transistor counts at equivalent cost-per-area. However, as transistor process technology advances into the single-digit nanometer nodes, it is evident that the required fabrication complexity has jeopardized the rate of transistor-size reduction as well as the cost-per-transistor. To continue increasing transistor counts without increasing manufacturing cost, an alternative solution is to utilize new multi-die packaging methods. Manufacturing yields can be improved by fabricating multiple smaller "chiplets" and validating each one before integrating them together, thus reducing the cost per functional system or increasing the number of transistors at the same cost. Additionally, multi-die chiplet systems provide opportunities for heterogeneous process integration and design reuse that can further reduce cost.

However, partitioning a monolithic chip into multiple chiplets introduces several new design considerations. First, which packaging technology, such as 2.5D interposer integration or Through-Silicon-Via (TSV) 3D stacking, is best suited to integrate a given multi-die system? Second, how can a partitioned chiplet system effectively communicate between dies without introducing bandwidth bottlenecks or excessive communication latency? Although many industry roadmaps are making clear transitions towards multi-die system design, academic research has made few strides towards answering these questions with respect to the range of modern packaging advances, and many questions about system and architecture design remain unanswered.

This dissertation seeks to address these challenges by providing methodology and analysis of the cost, power, and communication performance of several of the most promising packaging opportunities, including stacked 3D TSV integration, passive and active 2.5D interposer integration, and sequential monolithic 3D fabrication. First, manufacturing yield and cost models, with consideration of modern binning techniques, are developed and analyzed to demonstrate the cost benefit and overheads of several multi-die integration strategies. Second, these models are expanded with thermal-analysis-driven packaging and cooling costs to provide further insight into the overheads of increased circuit density. Third, a detailed investigation into the power density of modern 3D-stacked dynamic memory is performed to illustrate the challenges of high-density packaging. Fourth, based on these thermal-density concerns, this dissertation analyzes the challenges of manufacturing cost and inter-die communication for reduced-density interposer-based systems, using either passive or active silicon interposer substrates, and then demonstrates effective Network-on-Interposer topologies. Fifth, the communication requirements for future high-density Monolithic 3D systems are investigated and used to develop improved Network-on-Chip topologies that meet the unique requirements of sequential 3D fabrication.

Contents

Curriculum Vitae **vi**

Abstract **viii**

1 Introduction **1**

2 Cost Modeling of Die-Stacked Silicon Systems **6**
2.1 Die Manufacturing Costs 7
2.2 3D System Integration Costs 13
2.3 2.5D System Integration Costs 15
2.4 Die-Stacked Manufacturing Cost Comparison 17
2.5 Non-Recurring Engineering (NRE) Costs 24
2.6 Flexible Interconnect Architecture for Design Reuse 26

3 Cost Overhead of Increased Thermal Density in Die-Stacked Systems **32**
3.1 Baseline 2D Thermal Model 34
3.2 3D Thermal Model 35
3.3 2.5D Thermal Model 37
3.4 Cooling and Package Cost Estimation 38
3.5 Thermal-Aware Design Space Exploration 40
3.6 Conclusion 42

4 Power Modeling and Projection of Future Die-Stacked Dynamic Memory **43**
4.1 High Bandwidth Memory 44
4.2 HBM Power Modeling Methodology 46
4.3 Power Profiling for HBM2 48
4.4 Projection for Future Memory Power 52
4.5 Conclusion 55

5 System Integration with Interposers: Active versus Passive Technology 56
5.1 Motivation for Chiplet Partitioning . . . 57
5.2 Interposer Technology Selection . . . 59
5.3 Scaling of Interposer Link Width and Frequency . . . 65
5.4 Interposer Cost Comparison . . . 70
5.5 Performance of Network-on-Interposer . . . 79
5.6 Conclusion . . . 93

6 Network-on-Chip for Monolithic 3D Integration 95
6.1 Monolithic 3D Integration . . . 97
6.2 Necessity for M3D Network-on-Chip . . . 101
6.3 M3D Interconnect Characteristics . . . 107
6.4 M3D NoC Design Guidelines . . . 111
6.5 Conclusion . . . 116

7 Summary 117

Bibliography 122

Chapter 1

Introduction

The semiconductor industry has seen tremendous growth because of the economic and performance scaling of increasing integration complexity, with each new process node delivering more transistors-per-silicon-area with better performance and lower cost-per-transistor. By following the cadence of this scaling trend, formally recorded by Moore [1], the industry has been able to scale integrated circuits from several transistors to several billion, enabling the integration of floating point units, cache memories, multiple cores, graphics processing units, and power management units. In the maturing System-on-Chip era, a single die may additionally include modems, digital signal processors, heterogeneous cores, and application-specific accelerators to provide further system efficiency and market differentiation.

After several decades, however, these industry-defining scaling trends are struggling to continue. Power scaling has slowed significantly, leading to the current "dark silicon" era of power-efficiency-driven design. While the number of transistors-per-die continues to increase with smaller process nodes, many foundries have already failed to achieve the targeted area-scaling-per-transistor, and the cadence of new process technologies is slowing and expected to slow down further for future nodes. Perhaps most importantly,

Moore's famous observation on the cost-per-transistor may no longer hold, as the fabrication of sub-20nm FinFET transistors is sufficiently complex to require difficult and expensive fabrication technologies that translate to yield challenges and additional wafer cost. Thus, circuit designers and computer architects can no longer depend on the free availability of additional transistors and integration opportunity with each new process node. Additionally, non-recurring engineering costs have also increased quickly during the last several process nodes due to fabrication complexity issues, such as complex layout design rules and multiple masks per layer, and due to the system complexity challenges of verifying billions of logic gates.

Despite these major setbacks, alternate integration technologies may be able to function in tandem with traditional process node scaling to provide cost reductions and more transistors per circuit. Through die-level integration technologies, multiple dies can be connected electrically and physically to produce a larger integrated circuit. These technologies include 3D die stacking with connections through either face-to-face microbumps or face-to-back Through-Silicon Vias (TSVs), or through interposer-based 2.5D integration that uses a large passive or active routed die to provide interconnect between dies and the substrate, as demonstrated in Figure 1.1. By partitioning a monolithic SoC across multiple small die, yield-per-die can be greatly improved and metal layer count can be lowered, reducing the total circuit cost if integration overheads are sufficiently low. Die-level integration also allows for new integration strategies, such as heterogeneous process integration with different process technologies between dies, that can be used to further reduce costs or optimize performance. These technologies can also act as a platform for the reuse of hard intellectual property, allowing for system reconfiguration through the integration of different die combinations while amortizing non-recurring overheads of design, verification, and masks.

However, the range of multi-die systems with many advanced packaging options opens

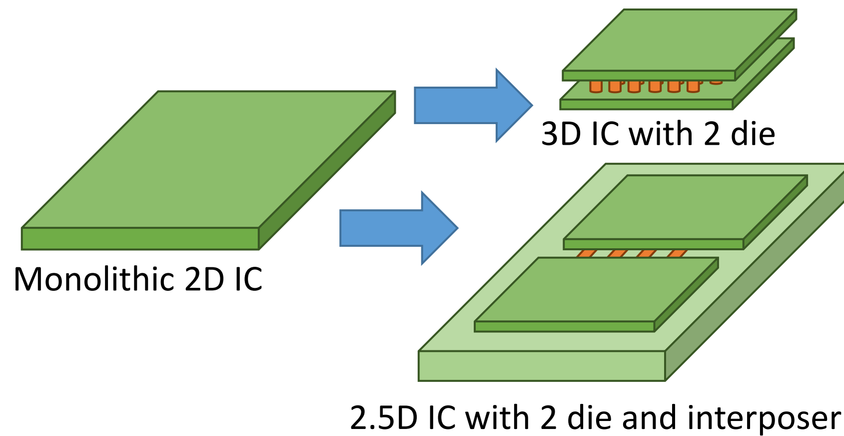


Figure 1.1: Die-level integration through TSV-based 3D and interposer-based 2.5D technologies.

up an increasingly complex design space that has not been systematically studied, especially given recent advancements in the integration technologies. Accordingly, this dissertation seeks to introduce the methodology necessary make these critical integration decisions, while providing analysis on several promising packaging technologies to determine the best options for different types of systems and their on-die communication requirements.

First, a model is developed for comparing the manufacturing yields and costs of different integration options, including 2.5D interposer substrates and Through-Silicon Via 3D stacking, versus traditional monolithic fabrication. A baseline die cost model is expanded for each multi-die alternative, with applicable manufacturing overheads, to provide answers to the critical decision of whether a packaging strategy is cost-effective for a given design. The model is further developed to better represent the functional and parametric binning strategies used in the manufacturing of modern processors, which disable faulty modules after fabrication and validation to produce partially-functional systems. Analysis of these models, using industry price and yield details, suggests that

both 2.5D interposer and 3D TSV integration can improve yield and manufacturing cost for a range of systems from mainstream to high-performance.

Second, to account for the overhead of increased thermal density, these multi-die cost models are expanded with thermal models, which account for the cost overhead of any packaging and cooling improvements that are needed to offset an increase in thermal density. After accounting for thermal management overhead, the results of the prior manufacturing cost comparison change to increasingly favor 2.5D interposer systems, with reduced thermal density, over die-stacked 3D integration, although both strategies can be cost-effective versus traditional fabrication for different niches of system area and power density.

Third, the implications of die-stacking on power density are further investigated by analyzing the power usage of modern die-stacked dynamic memories. An architectural power model is developed for the family of High Bandwidth Memory (HBM) devices, with support for memory traces generated from common architecture simulators, and the model is then validated against real HBM hardware. The resulting model is used to generate detailed power breakdowns across a range of memory behaviors, from idle to peak power, thus providing insight into the significant power contribution from modern stacked memories. Further, the power of near-future memories, with higher bandwidth and memory densities, is projected to suggest that even greater peak power densities may occur, making in-package memory an increasingly critical component in system design.

Fourth, in response to the promising cost benefits previously demonstrated by interposer-based integration, a detailed investigation is performed into the cost and performance trade-offs of different interposer technologies: passive interposers with relatively low cost but simple metal-only construction, versus active interposers with integrated transistors that can reduce link latency and provide on-interposer router networks. Through an analysis of the Network-on-Interposer design space, this section demonstrates that

both active and passive interposers can be cost-effective versus traditional fabrication while providing significant inter-chiplet network bandwidth. Results suggest that active interposers are able to reduce network latencies versus passive interposers, through a combination of lower-latency links and synchronous on-interposer router clocking, but at an increased fabrication cost, even when on-interposer fault tolerance techniques are applied to improve yield.

Fifth, a similar methodology is applied while looking forward to future monolithic 3D (M3D) integration, in which multiple tiers of transistors are fabricated sequentially into a single die with very high inter-tier communication performance. To motivate the need for efficient monolithic 3D on-die communication, a Processing-in-Memory accelerator is reviewed in the context of M3D, demonstrating significant communication bottlenecks. Next, the unique process considerations of M3D fabrication, such as temperature-sensitive materials and contested metal from increased transistor density, are mapped onto the requirements of on-die network link topologies. A performance investigation into the resulting M3D Network-on-Chip design space provides best practices for M3D partitioning decisions and network architectures.

Chapter 2

Cost Modeling of Die-Stacked Silicon Systems

Due to the increasing fabrication and design complexity with new process nodes, the cost-per-transistor trend originally identified in Moore's Law [1] is slowing when using traditional integration methods. However, emerging die-level integration technologies may be viable alternatives that can scale the number of transistors-per-integrated-circuit while reducing the cost-per-transistor through yield improvements across multiple smaller dies. Additionally, the escalating overheads of non-recurring engineering costs like masks and verification can be curtailed through die-integration-enabled reuse of intellectual property across heterogeneous process technologies. In this chapter, analytical cost models for Through-Silicon Via (TSV) 3D and interposer-based 2.5D die integration are developed and employed to demonstrate the potential cost reductions across semiconductor markets. Further, this work proposes a methodology and platform for design reuse based on these integration technologies.

2.1 Die Manufacturing Costs

In order to accurately study the relative costs of traditional 2D integrated circuits versus 3D or 2.5D die-level integrated circuits, an analytical cost model was developed to determine the approximate manufacturing cost per integrated circuit. The methodology includes flexible estimation of die area, metal layer count, and die yield to determine cost per individual die. The model is then extended to include 3D processing and bonding overhead, additional TSV area, 2.5D interposer cost, and modular binning methodology employed in modern integrated circuits. Later chapters further expand the models with packaging and cooling costs to estimate the final manufacturing cost of the packaged integrated circuit.

The cost of an individual die, whether the monolithic die in a traditional 2D circuit or one of several dies in a 3D or 2.5D circuit, can be estimated from a few parameters, allowing for a wide range of design costs to be studied and compared. The choice of process technology has a major impact on the die cost as it determines the cost per wafer, number of metal layers and cost per additional layer, the average area per transistor and gate, and the defect density. Once a process technology has been selected, the expected area can be calculated from the number of transistors or gates. With the process technology, area, and number of gates per die, it is then possible to estimate the required number of metal layers. Metal layer estimation is an important step in the cost model because the maximum number of metal layers impacts the cost per wafer. Accordingly, the number of metal layers per die can be reduced by partitioning a single large die into multiple smaller die, helping to reduce the die cost.

2.1.1 Area Estimation

Die area influences dies per wafer and die yield and is therefore a critical parameter for determining silicon cost. The die area can be estimated from the selected process node and from the number of gates in the design, or by assuming an average of four transistors per gate, with the equation:

$$A_{die} = N_g * \beta\lambda^2 \tag{2.1}$$

where N_g is the number of gates in the design, λ is the feature size, and β is an empirical scaling term where $\beta\lambda^2$ is the average area per gate. Previous work has used data from industrial designs to estimate a single value of β [2]. However, a survey of modern designs reveals considerable variation between different IC markets, resulting in area estimation of up to 3x the actual design area when using previous scaling terms. Additionally, the formula assumes ideal scaling trends between each process generation. In reality, the gate pitch and minimum bit cell sizes have both scaled slower in many process technologies since 28nm [3][4]. Adjusting λ to the effective feature size, as calculated from actual gate pitch and bit cell scaling, rather than the advertised feature size, will improve estimation accuracy. Table 2.1 presents scaling coefficients for several integrated circuit markets, illustrating the variability in gate sizing for different design types, with data surveyed from 90nm to 14nm processes. Average power densities for different markets are included as well.

Once a process technology has been selected, the expected area can be calculated from the number of transistors or gates, with previous work assuming an average of four transistors per gate. In this paper, the area is estimated with the equation $A_{die} = N_g * \beta\lambda^2$ where N_g is the number of gates, λ is the feature size, and β is an empirical scaling term such that $\beta\lambda^2$ is the average area per gate. A value for β can be determined from

a survey of previous market designs, with values ranging from 450 million for dense graphics processors, 700 million for consumer CPUs, and up to 850 million for some SoCs. As some sub-28nm foundry process nodes have scaled less than expected [3][4], the λ value should be adjusted to the true effective feature size.

Design Type	Scaling Coefficient β (M)	Power Density (W/mm^2)
CPU (desktop)	720	0.45
CPU (mobile)	610	0.24
CPU (server)	670	0.44
GPU (desktop)	440	0.47
GPU (mobile)	450	0.40
GPU (server)	440	0.33
Desktop SoC	840	0.27
Mobile SoC	710	0.19

Table 2.1: Gate sizing coefficients and average power densities for commercial products from 90nm to 14nm.

2.1.2 Metal Layer Estimation

The metal layer count is also an important parameter for determining die cost, as additional metal layers require extra fabrication steps and resources. The number of required metal layers depends on the interconnect distance that must be routed in the design and therefore depends on the design complexity. The range of available metal layers may be limited by the fabrication process, but estimation methodology can predict the required number of metal layers in new ASIC designs. First, Rent's Rule [5] can be used to estimate the average wire length in the design [6]:

$$\bar{L} = \frac{2}{9}\Big(\frac{1-4^{p-1}}{1-N_{gates}^{p-1}}\Big)\Big(\frac{7N_{gates}^{p-0.5}-1}{4^{p-0.5}-1} - \frac{1-N_{gates}^{p-1.5}}{1-4^{p-1.5}}\Big) \tag{2.2}$$

where p is Rent's exponent value that expresses the route complexity. The number of metal layers can then be approximated from the average wire length with the equation:

$$n_{metal} = \frac{f.o.N_{gates}\bar{L}\omega}{\eta A_{die}} \tag{2.3}$$

where n_{metal} is the number of required metal layers, $f.o.$ is the average fanout, ω is the wire pitch, and η is the average interconnect utilization rate with consideration of percent metalization and overheads of vias and power and clock tracks. This formula assumes a uniform metal pitch across metal layers, but if specific metal stack wire dimensions are known, layer-based assignment with variable wire pitch and utilization can be employed [7][2].

Area(mm^2)	Gate Count	1 die	2 dies	3 dies	4 dies
5	21	7	7	6	6
10	41	8	7	7	7
25	103	9	8	8	7
50	207	9	9	8	8
100	413	10	9	9	9
250	1033	11	10	10	9
500	2065	12	11	11	10

Table 2.2: Estimated metal layer counts of single and partitioned die with 14nm process, $\beta = 650M$, and $\eta = 0.3$

Metal layer estimation is also useful for anticipating the available reduction in metal count from die partitioning in a 2.5D and 3D design. As shown above in Table 2.2, partitioning a design into multiple dies can reduce the number of required metal layers per die, thus decreasing wafer cost.

2.1.3 Yield Modeling and Manufacturing Cost

The cost of an individual silicon die prior to any additional steps for 3D integration can be estimated from the process technology, area, and metal layer count. To determine the die cost, the first step is to model the cost per wafer. The cost of a wafer is dependent upon the process details, wafer diameter, and foundry vendor. This process technology choice is often determined early in the design process and is an input to this model. Within each process, the price per wafer can vary by the number of required metal layers, as extra metal layers require additional processing steps. Using methodology outlined in [8], the cost per wafer is given as:

$$C_{wafer} = C_{process} + N_{metal} * C_{metal} \tag{2.4}$$

where C_{wafer} is the total wafer cost, $C_{process}$ is the base cost per wafer, and N_{metal} is the additional cost per metal layer. The values employed in our model were calculated using *IC Knowledge LLC IC Cost Model* [9].

Each fabricated wafer contains a finite number of silicon dies within its area, as determined by the wafer diameter and die area. The number of dies per wafer is calculated by:

$$N_{die} = \frac{\pi \times (\phi_{wafer}/2)^2}{A_{die}} - \frac{\pi \times \phi_{wafer}}{\sqrt{2 \times A_{die}}} \tag{2.5}$$

where N_{die} is the number of dies per wafer, ϕ_{wafer} is the wafer diameter, and A_{die} is the die area.

For a given wafer, only a percentage of dies will properly yield after fabrication. Assuming a negative binomial yield model [10], the die yield is calculated from:

$$Y_{die} = Y_{wafer} \times \left(1 + \frac{A_{die} D_0}{\alpha}\right)^{-\alpha} \tag{2.6}$$

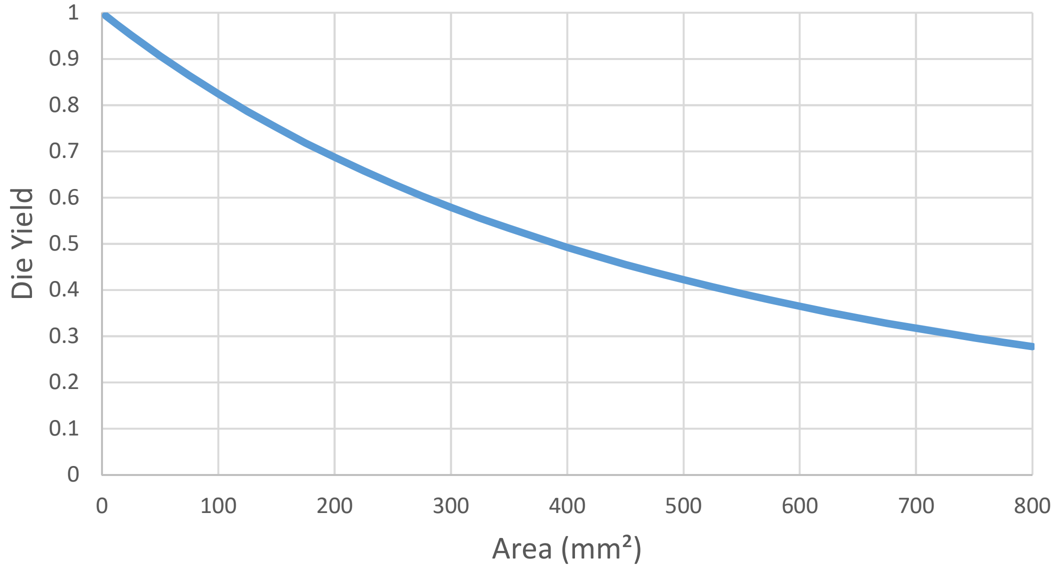


Figure 2.1: Manufacturing die yield versus area $D_0 = 0.2$ defects per cm^2

where Y_{die} is the die yield and D_0 is the defect density, which is determined by the process. For the model, $\alpha = 3$ is selected and therefore the Dingwall yield model [11] is used. An example yield trend with area is shown in Fig. 2.1. Note that as the area increases, the yield rapidly decreases, thus encouraging an SoC design with multiple small dies over a single large die.

The final die manufacturing cost C_{die} can then be calculated with:

$$C_{die} = \frac{\frac{C_{wafer}}{N_{die}} + C_{test}}{Y_{die}} \tag{2.7}$$

where C_{test} is the cost of die testing and where C_{wafer} is the process-dependent wafer cost.

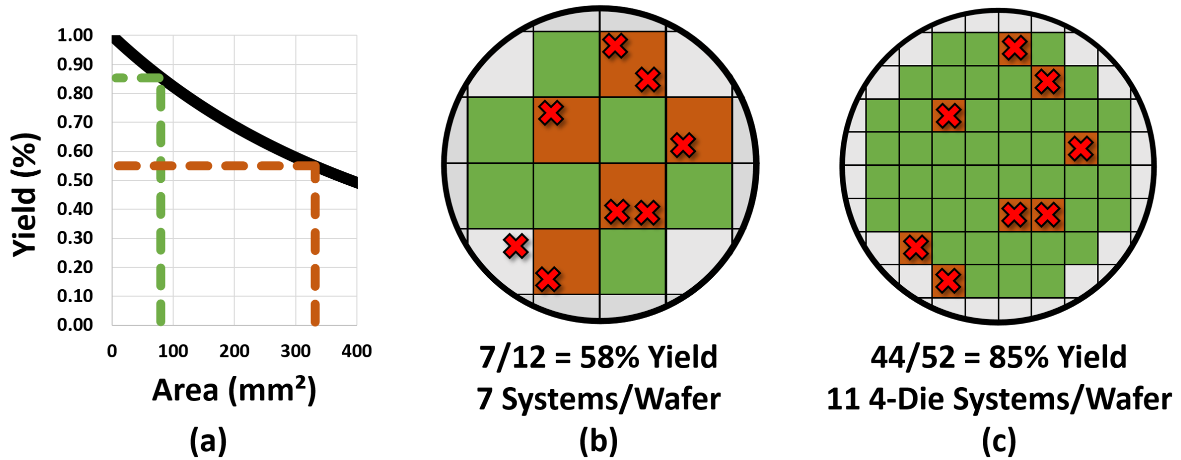


Figure 2.2: (a) Yield versus area with defect density $D_0 = 0.2/cm^2$ and $\alpha = 3$, with examples design points highlighted at $A = 336mm^2$ $(Y = 0.55)$ and $A = 84mm^2$ $(Y = 0.85)$. Scaled visual examples with eight defects per wafer are shown for (b) a monolithic system and (c) an equivalent 4-chiplet system.

2.2 3D System Integration Costs

For complex integrated circuits with high transistor count and large area, a single 2D die will be difficult to fabricate and will therefore be expensive to produce. Partitioning a large design into multiple small die will improve the yield per die, as shown above in Equation (2.6), and can reduce the overall silicon cost, but these dies must be tightly integrated to maintain performance. 3D packaging technology includes several techniques for creating vertical interconnects between stacked silicon die. The most mature of these technologies is the use of Through-Silicon Vias (TSVs) to provide high-bandwidth connections between layers with latency that meets or exceeds on-die global routes [12]. Although partitioning can improve die yield, 3D integration also introduces additional manufacturing overheads that may add to silicon costs, including wafer thinning, via production, and die bonding.

For this model, dies in a 3D stack are assumed to use face-to-back arrangement, which offers the option to stack beyond 2 layers at the expense of additional via area. TSVs,

with manufactured diameter below 1 μm [12], introduce area overheads to the die that must be considered for accurate cost modeling. The number of TSVs between two layers will depend on partitioning decisions and circuit organization. For many circuit designs, the number of TSVs will be set by global interconnect buses between layers. From the known TSV count X_{TSV}, the adjusted area for a die in the 3D stack A_{3D} is:

$$A_{3D} = A_{die} + X_{TSV} A_{TSV} \tag{2.8}$$

where A_{die} is the original area and A_{TSV} is the area per TSV, including keep-out boundary. As TSVs block routable area, only A_{die} is available for metal interconnect routing.

In an ASIC design, the via count between two layers can be estimated from the gate counts and Rent's Rule coefficients [8], as follows:

$$\begin{aligned} X_{TSV} =& \alpha k_{1,2}(N_1 + N_2)(1 - (N_1 + N_2)^{p_{1,2}-1}) - \\ & \alpha k_1 N_1 (1 - N_1^{p_1 - 1}) - \alpha k_2 N_2 (1 - N_2^{p_2 - 1}) \end{aligned} \tag{2.9}$$

where $k_{1,2}$ and $p_{1,2}$ are equivalent Rent's Rule coefficients [13].

$$\begin{aligned} p_{1,2} &= \frac{p_1 N_1 + p_2 N_2}{N_1 + N_2} \\ k_{1,2} &= \left(k_1^{N_1} k_2^{N_2} \right)^{1/(N_1+N_2)} \end{aligned} \tag{2.10}$$

To calculate the cost of a 3D die stack, the costs of the individual dies are first calculated using the die cost estimation equations. For each die in the stack, the cost is increased by extra TSV area overhead and additional process costs for wafer thinning and TSV processing. This model assumes die-to-wafer stacking and known good die testing (KGD) before bonding, which have been shown to reduce net cost when die yields are low [14]. It also assumes no stack testing between bonding steps, which has also been

shown to be cost effective when bond yields are high [15].

The net cost of the 3D stack C_{3D} is calculated from:

$$C_{3D} = \frac{\sum_{i=1}^{n} (\frac{C_i}{y_i}) + (n-1)C_{bond}}{Y_{bond}^{n-1}} \tag{2.11}$$

where Y_{bond} is the bond yield, n is the number of die, C_i and y_i are the silicon cost and yield of a given die, and C_{bond} is the cost of alignment and bonding between die. Note that die layers that require TSVs will have higher process costs $C_{process}$ during the wafer cost C_{wafer} calculation.

2.3 2.5D System Integration Costs

Recently, 2.5D packaging has emerged as an alternate integration technology to full 3D packaging, particularly for the integration of 3D-stacked dynamic memory into GPUs and FPGAs. In 2.5D packaging, an interposer substrate is used to integrate multiple die or 3D die stacks into a single package. Interposer route dimensions approach those of the on-die metal, providing high bandwidth between die. Die can be flip-chip bonded to the interposer with microbumps, currently on the scale of 25 μm and below [12], instead of with TSVs, so integration can be improved with reduced design overhead. However, the interposer itself, currently fabricated in silicon using traditional foundry processes, adds additional manufacturing cost to the design, as the interposer is a large silicon die that often includes TSVs for die-to-package power and signal connectivity. Savings from die yield improvements will be reduced by the overhead of interposer silicon cost, but interposer costs are significantly less than die costs for comparable areas due to the lack of active transistors and small number of routing layers. Figure 2.3 compares the costs of chips with different areas using a 65 nm interposer process and a 65nm CMOS logic

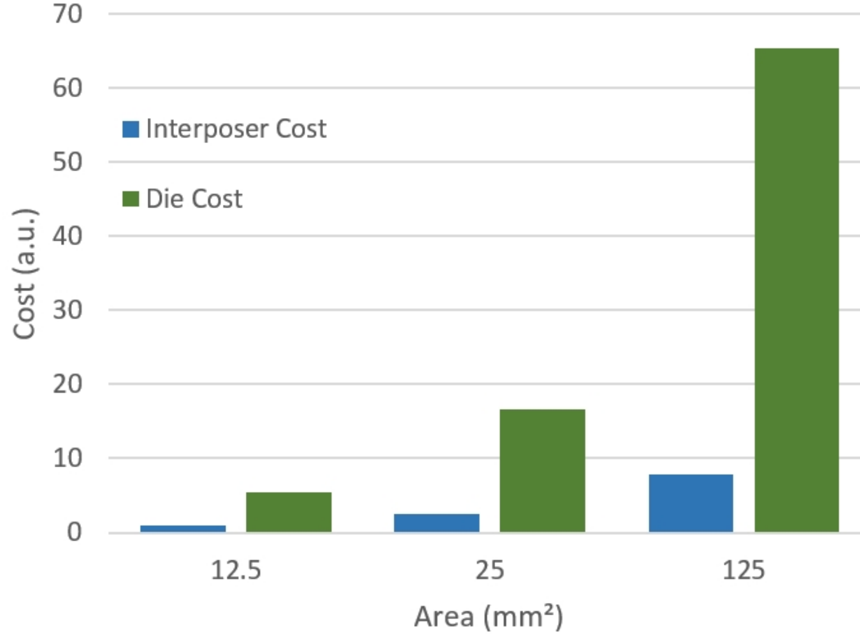


Figure 2.3: 65nm Interposer Cost and 65nm CMOS Die Cost with Area

process with 7 metal layers to illustrate the relative price difference.

The net wafer-level cost of the 2.5D silicon stack can be calculated as:

$$C_{2.5D} = \frac{\frac{C_{int}}{y_{int}} + \sum_{i=1}^{n} \left(\frac{C_i}{y_i} + C_{bond_i}\right)}{Y_{bond}^{n-1}} \tag{2.12}$$

where $C_{2.5D}$ is the net silicon cost of the 2.5D stack, Y_{bond} is the bond yield between a die or die stack and the interposer, n is the number of die/stacks being bonded to the interposer, C_{int} and y_{int} are the interposer silicon cost and yield, C_i and y_i are silicon cost and yield of a given die or stack, and C_{bond_i} is the bond cost for a given die or stack, which can vary with required accuracy and other manufacturing considerations. For the model, it is again assumed that known-good-die testing is used before bonding to the interposer, and no die testing between bonding steps. Required interposer area can be approximated as the sum of the footprint areas of the bonded die and die stacks, as interposer's fine-pitch routing and bond alignment does not require the same extra

footprint area that coarse-pitched organic substrates frequently require. This model does not consider die bonding on both sides of the interposer, which reduces required interposer area but requires vertical interconnect spacers to attach the interposer to the substrate and introduces cost and thermal complexity outside of the scope of this model.

2.4 Die-Stacked Manufacturing Cost Comparison

With cost estimation methodology for 2D, 2.5D, and 3D integrated circuits, the silicon fabrication costs at different design sizes can be compared to determine the enabling points of the different die integration technologies. Unless otherwise noted, the parameters outlined in Table 2.3 are assumed for cost estimation.

Table 2.3: Assumed values for design exploration.

Feature Size (λ)	14 nm	Y_{wafer}	98%
Area Scaling (β)	650M	Y_{bond}	99%
Rent's Coefficient (k)	4.0	D_{TSV}	1 μm
Rent's Exponent (p)	0.6	$D_{\mu bump}$	25 μm
Metal Utilization (η)	30%	Interposer Feature Size	65 nm
Gate Pitch	$4.5 \times \lambda$	Average Fan-out ($f.o.$)	4
Wire Pitch	$3.6 \times \lambda$	Defect Density (D_0)	$0.2/cm^2$

Figure 2.4 shows the relative fabrication costs for designs of various gate counts using traditional 2D fabrication, interposer-based 2.5D fabrication with the design partitioned into either 2 or 4 smaller dies, and TSV-based 3D integration with the design partitioned into either 2 or 4 layers. Fabrication cost for the 2.5D circuits includes cost and yield for the interposer and bonding steps. Fabrication costs of the 3D circuits includes process overhead for the addition of TSVs and extra thinning, TSV area overhead, and bonding costs and yields.

As gate count increases, both 2.5D and 3D circuits become more cost effective than single-die designs because of the area-dependent yield trend described in Equation (2.6).

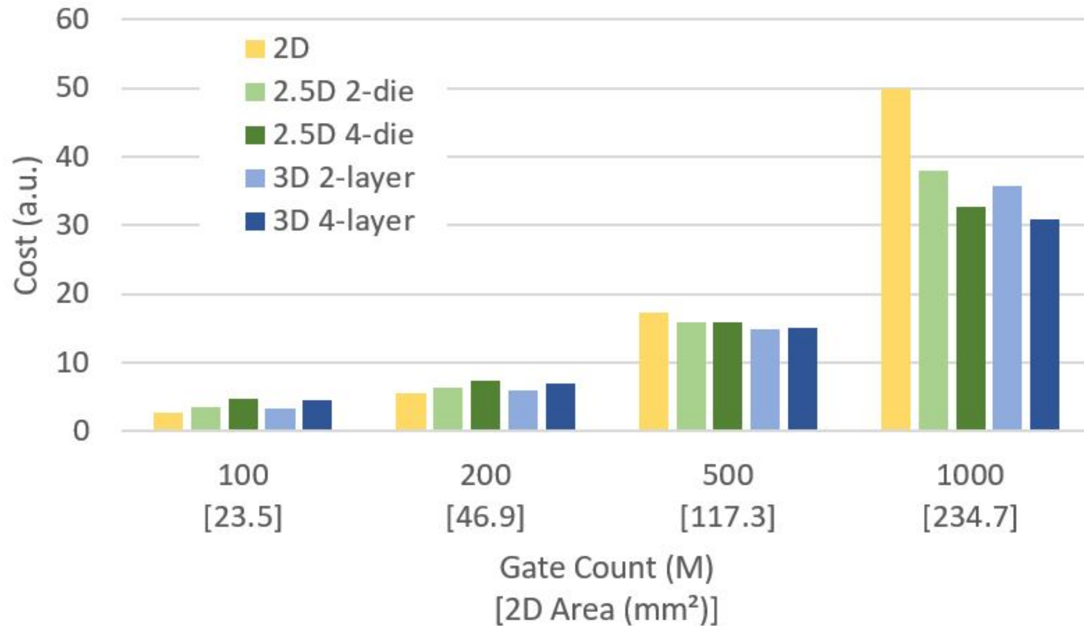


Figure 2.4: Fabrication cost versus gate count in 14nm process for 2D, 2-die 2.5D, 4-die 2.5D, 2-layer 3D, and 4-layer 3D designs

For the same number of die partitions, 3D fabrication is lower cost than 2.5D fabrication because of the interposer silicon overhead, which, although much cheaper than the active silicon, also exhibits reduced yield at large die area. Table 2.5 shows, for multiple bond yields, the number of gates at which 2.5D and 3D integration become cheaper to fabricate than single-die designs. The enabling points are also dependent upon the process technology, as shown in Table 2.4. Figures 2.5 and 2.6 show the relative cost contributions of different fabrication factors, including die cost, testing cost, interposer cost, TSV overhead, and bonding cost, at two different design sizes. The relatively high cost enabling points, in terms of gate count and area, of 2.5D and 3D integration confines their cost effective use to high-performance IC markets with greater design complexity.

Table 2.4: Enabling points, in gate counts, in 16nm, 28nm, and 40nm process technologies.

		Gate Count (M)	2D Area (mm^2)
16 nm	2.5D	262	75.1
	3D	177	50.7
28 nm	2.5D	231	117.7
	3D	133	67.8
40 nm	2.5D	107	111.3
	3D	87	90.5

Table 2.5: Enabling points, in gate counts, of 2.5D and 3D fabrication in 14 nm process.

Bond Yield (%)	2.5D 2-die	2.5D 3-die	2.5D 4-die	3D 2-layer	3D 3-layer	3D 4-layer
0.99	325 M	361 M	376 M	262 M	270 M	326 M
0.95	481 M	536 M	615 M	288 M	394 M	487 M
0.90	747 M	770 M	923 M	383 M	555 M	666 M

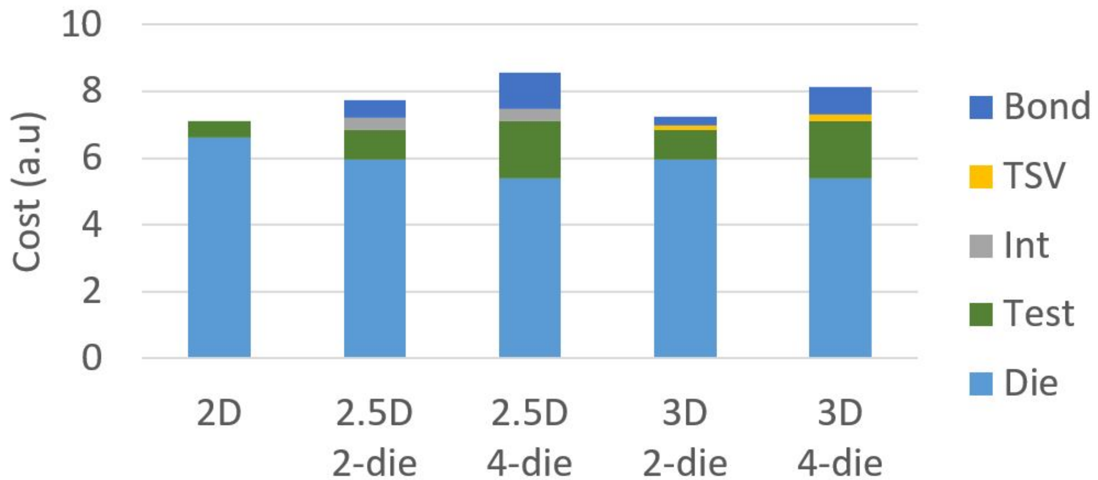


Figure 2.5: Cost contribtions in 14nm process with 250M gates

2.4.1 A Core Binning Yield Model for Modular Circuits

The yield model in Equation 2.6, although commonly used in prior work, is not representative of the fabrication of modern large area integrated circuits. With chip sizes that can approach the reticle limit, the yield for a defect-free die can be very low, even for mature process nodes. For example, according to Equation 2.6 with $\alpha = 3$, a 600 mm^2

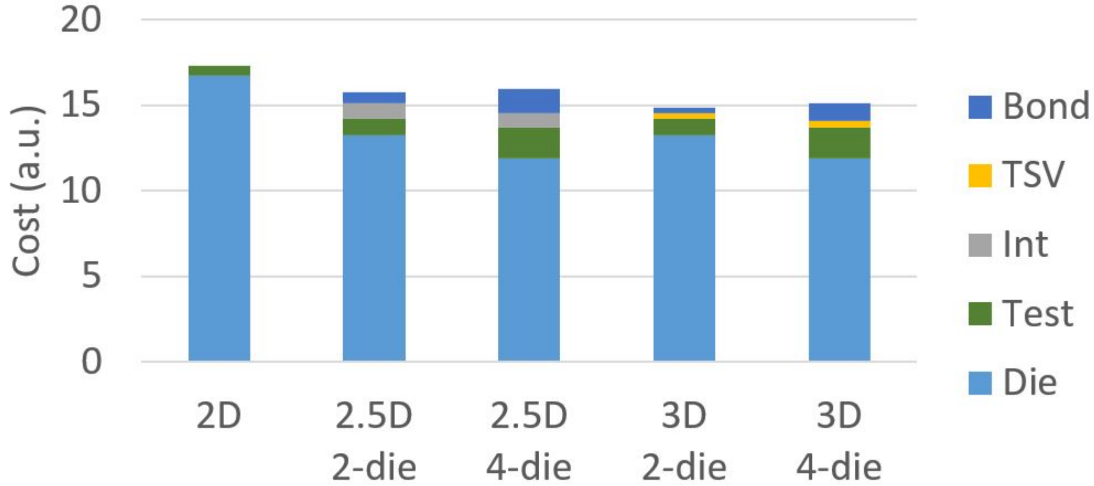


Figure 2.6: Cost contributions in 14nm process with 500M gates

GPU die [16] in a mature process node with defect density $D_0 = 0.2\ cm^{-2}$ [17] would have a die yield (before parametric variation) of only 36%. For an emerging process with $D_0 = 0.5\ cm^{-2}$, yield is only 12.5%. In order to improve revenue and produce more functional parts, leading manufacturers of CPUs, GPUs, and other high performance circuits rely on **binning** at the core unit level. If a defect is present in a modular core, the impacted segment of the die is disabled and the chip is sold with reduced functionality at a lower price.

In order to model the distribution of defects between and within the dies, utilize the derivation equation of the negative binomial yield model, shown below in Equation 2.13.

$$P_{defect} = \frac{\Gamma(d+\alpha)}{d! * \Gamma(\alpha)} * \frac{\beta^d}{(\beta+1)^{d+\alpha}} \tag{2.13}$$

The probability that a die has d defects is calculated using the gamma function $\Gamma(x)$ and constant β defined as:

$$\beta = \frac{D_0 A}{\alpha} \tag{2.14}$$

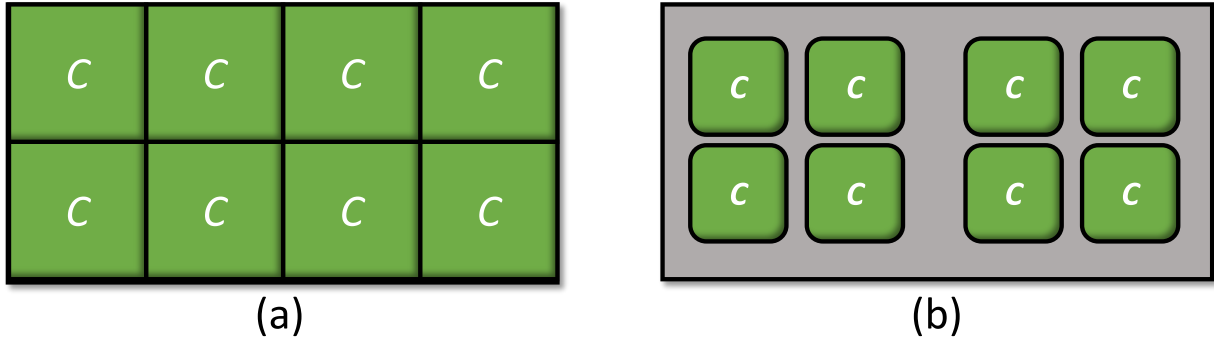


Figure 2.7: (a) An 8 core die in which 100% of the die can be flexibly disabled for binning. (b) A representative system where only the cores, which make up 50% of the area, can be disabled for binning.

Within the relatively local area of a given die, it is assumed that defects are randomly distributed (Poisson) across the cores and uncore area. Multiple defects may fall into the same core, resulting in more functional cores after binning. The probability of a die with d defects and c binnable modular cores to have g good, functional cores is:

$$P_{good} = \frac{S(d, c-g)\binom{c}{c-g}(c-g)!}{c^d} \tag{2.15}$$

where $S(d, c-g)$ is the Stirling number of the second kind. Equation 2.15 assumes that the whole die is partitionable for binning. In real designs, non-modular uncore units like interconnect fabric and system management contribute significant die area and are not easily disableable. Figure 2.7 shows an 8 core processor with (a) fully partitionable die area (b) 50% binnable core area and 50% critical uncore area, representative of modern designs. Equation 2.15 can be expanded to account for non-modular critical area percentage η:

$$P_{good_\eta} = P_{good} * (1-\eta)^d \tag{2.16}$$

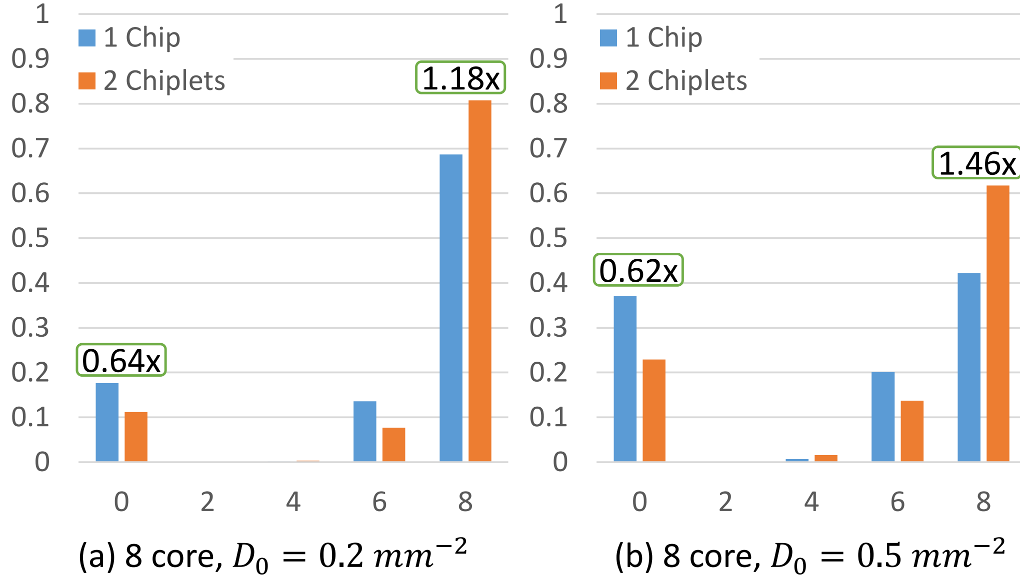


Figure 2.8: Yield distribution of binned dies after manufacturing for each functional core count bin - Eight-Core CPU.

2.4.2 Core Binning and Cost Results for Mainstream 8-Core CPU

By taking the sum of products of Equations 2.13 and 2.16 across all defect counts, the yield distribution for each number of functional cores can be determined. This section first investigates a mainstream 8 core desktop/workstation consumer processor with $A = 200mm^2$, $\alpha = 3$, and $\eta = 0.5$, as shown in Figure 2.7b. Binning is performed at multiples of 2 cores, as in modern commercially available processors. For the 2 chiplet design, a greedy matching process is used to produce as many fully-enabled processors as possible. A per-chiplet bond yield $Y_{bond} = 99\%$ [18] is included in the 2 chiplet system yield distribution to reflect pessimistic integration losses. Binned yield distribution results are shown in Figure 2.8 (a) and (b) for a mature defect density of $D_0 = 0.2\ cm^{-2}$ and for a cutting-edge defect density, or potentially a low yield future process, with $D_0 = 0.5\ cm^{-2}$. The yield improvement from chiplet partitioning and KGD testing

translate to a reduction in unsalvageable chips and an increase in the number of fully enabled, high margin chips. At the defect rates for a mature process and an emerging process, the number of fully function cores increased by 1.18x and 1.46x and the number of failing systems decreased by 0.64x and 0.62x.

Speed Bin	2 core	4 core	6 core	8 core
Target	1	1.7	2.5	5
Slow	0.8	1.5	2	3.7

Table 2.6: Normalized price per core count of existing consumer processors at two speed bins.

To measure the total utility of these improvements to yield and functionality, utilize the estimated price of equivalent commodity processors as a representative value metric. Table 2.6 lists normalized, approximate price ratios for each core count at two speed bins based on currently available consumer devices. To model parametric yield, which can also be improved through die partitioning and known good die matching [19], a Gaussian frequency distribution is assumed for each core, with any cores with frequency below one standard deviation of the mean binned to "Slow" and average and faster cores binned to "Target". Under this simple parametric model, about half of the 4 core chiplets will achieve the target speed, while only a quarter of the 8 core chips can meet the target. Through a combination of functional and parametric yield improvements, the utility value metric of the 2 chiplet system is improved by 20.8% when $D_0 = 0.2\ cm^{-2}$ and by 41.4% when $D_0 = 0.5\ cm^{-2}$.

2.4.3 Core Binning and Cost Results for Server 32-Core CPU

While modest yield improvements are seen from chiplet partitioning for the consumer processor at mature defect densities, increasingly significant gains are seen for larger area circuits like server processors that exhibit greater yield challenges. Yield distributions for an example 32 core server processor with $A = 600mm^2$ are shown in Figure 2.9 (c)

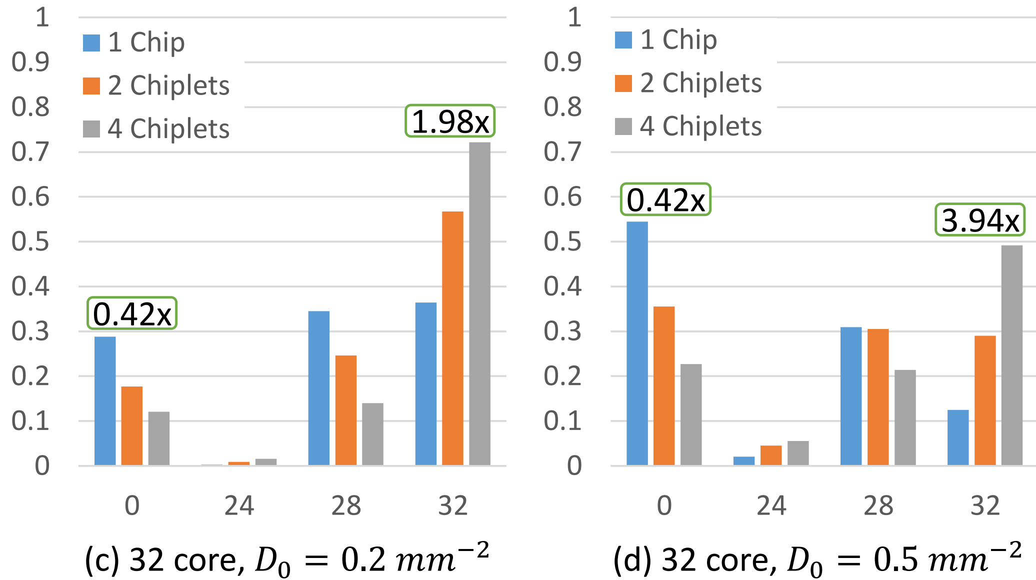


Figure 2.9: Yield distribution of binned dies after manufacturing for each functional core count bin - 32-Core CPU.

and (d) for the same $D_0 = 0.2\ cm^{-2}$ and $D_0 = 0.5\ cm^{-2}$, respectively. Die partitioning results in a 0.42x reduction in failing chips and a 1.98x improvement in the number of fully enabled chips for the mature process, and a 0.42x reduction in failures and a very sizable 3.94x improvement in full enablement for the emerging process.

2.5 Non-Recurring Engineering (NRE) Costs

he cost model and analysis presented in the previous sections cover the manufacturing cost of an integrated circuit, which can be examined without concern for specific product volume quantity. The total cost of an integrated circuit must also include the contribution of non-recurring engineering (NRE) costs that are paid once and then amortized across the number of produced circuits. For sufficiently large volumes, these non-recurring costs are distributed across all devices for a small impact on cost per device. Unfortunately, increases in design complexity with smaller process nodes and higher levels of integration

are resulting in growing NRE costs. This includes physical complexity from smaller feature size, requiring multiple masks per layer and more difficult design rules, as well as system complexity from more transistors per integrated circuit, resulting in more difficult design and verification. Industry analysts have observed a rise in the design cost of a standard SoC by 2.7x between 28nm and 14nm designs, and anticipate a further increase to 9x, over $270 million, from 28nm to 7nm [20]. Similarly, projections from ARM suggest that because of non-recurring costs, average sized designs will be prohibitively expensive with volumes less than 10 million units and even at high volumes there will be an increase in cost per constant area [21]. To demonstrate the relative impact of NRE costs, Figure 2.10 uses the constant area die cost projections from [21] to break down the total die cost between silicon manufacturing cost and amortized NRE overhead for a sample 14nm chip at different product volumes. Similar but less extreme trends are seen for older process technologies with reduced process complexity, while future technology nodes will require even larger unit volumes to amortize the expensive non-recurring costs.

As demonstrated in the previous sections, die-level integration may be a cost-driven design method towards the reduction of manufacturing costs and the improvement of integrated circuit yields, especially for larger designs like SoCs. However, non-recurring engineering costs may still dominate the overall product cost, especially at lower volumes, if a design is only partitioned across multiple dies. To further reduce costs and combat the trend of increasing NRE, die-level integration may also be employed as a platform for the reuse of intellectual property at the die level. The dies of non-critical logic can be used across SoC designs to further amortize the initial non-recurring costs of mask and design and to greatly reduce the verification effort. Additionally, heterogeneous process technologies can be employed, so any critical logic can move to the highest performing node while reusable logic may remain in an older process technology with mature yields.

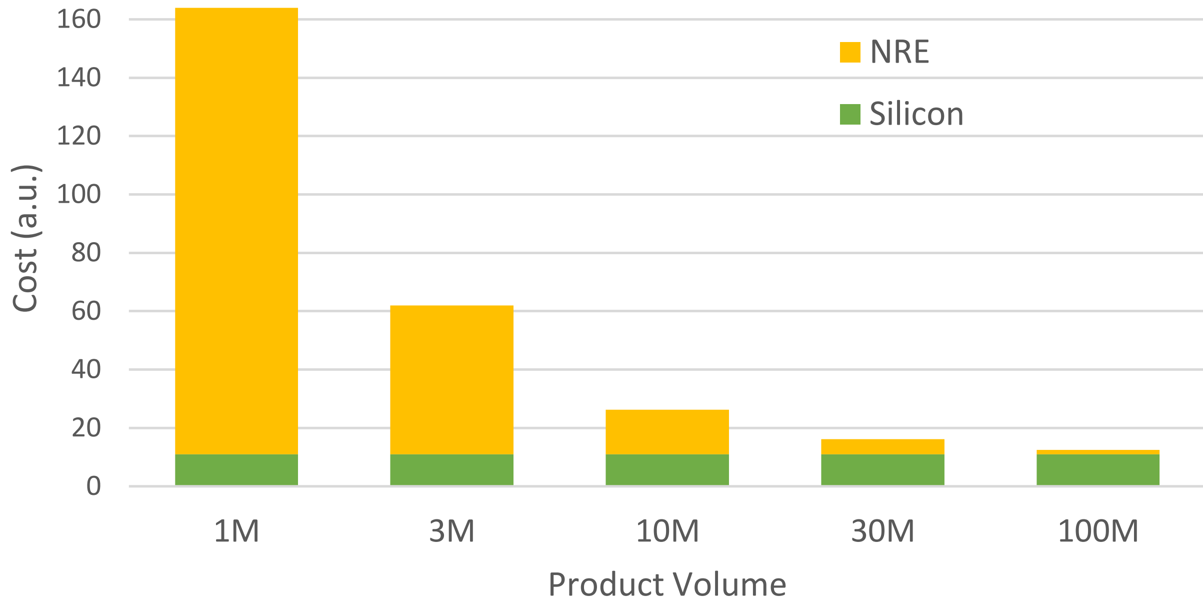


Figure 2.10: Total die cost breakdown between silicon fabrication and NRE overhead for sample 14nm chip

2.6 Flexible Interconnect Architecture for Design Reuse

In traditional SOC integration, various Intellectual Property (IP) blocks are connected with standardized bus (such as AMBA bus) or customized point-to-point interconnects. Such interconnect architecture is typically fixed for one particular design, and does not provide flexibility to be reused for future chip integration, because the interconnect fabrics and IP blocks are all fabricated on the same 2D chip. Such interconnect fabric is normally built for *providing best average case performance* across a generic set of applications. The reason being, no fixed on-chip network design efficiently supports all different types of communication requirements. Therefore it lacks flexibility and is unable to adapt to dynamically changing and special communication requirements at runtime. In addition, in 2D as well as 3D designs the interconnect fabric is tightly coupled with computing (core) and storage (cache) components. This pushes chip designers to adopt low complexity, structured and regular interconnect design to *limit pre-fabrication and*

post-fabrication verification efforts and costs.

This section proposes that 2.5D/3D integration technology be leveraged to design flexible interconnect topology that can ensure existing or future IP block can be easily swapped, so that each device technology's unique performance characteristic can be preserved while enabling fast and compatible integration.

The reuse of intellectual property at the die-integration level may be realized through both 3D and 2.5D technologies. Reuse with TSV-based 3D integration would normally be limited by the predetermined placement and connectivity of the TSVs. In our work, the key concept is to adopt the *Network-on-Chip* concept to replace the traditional bus interconnect structure, and decouple the interconnect fabric from the IP blocks, and implement the interconnect architecture in a separate silicon layer called **interconnect service layer (ISL)** [22], either on the interposer (for 2.5D) or as an independent layer in 3D stacking. Such decoupling can provide reduced manufacture cost and offer more reliable and flexible interconnect layer compared to its traditional 2D counterparts. The decoupled ISL can contain multiple on-chip networks such as mesh, ring, hierarchical bus topologies, etc. With ISL the constraints on the on-chip network router area and link bandwidth can be relaxed, and it can also support different manufacture volume for each die in 3D to reduce the overall cost. For example, the proposed ISL (either as 2.5D interposer or as a separate layer in true 3D stacking) can be manufactured with much larger volume than the IP blocks, then it can be integrated with various IP blocks on various design, such as with different number of CPU cores and various analog/mixed signal IP blocks.

The service layer is an additional die that provides a network-on-chip (NoC) with interconnect and routers to connect the TSVs between the integrated die above and below it, allowing for non-adjacent units within and across dies to communicate efficiently. An example ISL system is shown in 2.11a. Because of the relative uniformity and sparseness

of the service layer, it can be produced at high yield and volume to minimize the manufacturing cost. A catalog of several ISL designs of varying sizes could be used to address different product markets, but to minimize the number of mask sets each ISL design can allow for physical slicing of the full die to match the area of a given product, reaching a larger product volume and amortizing design and mask costs.

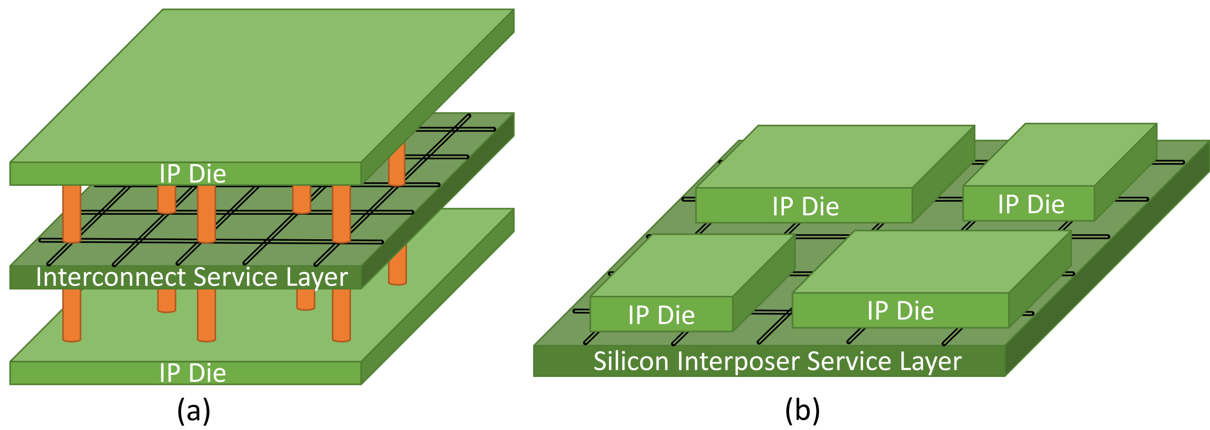


Figure 2.11: (a) Interconnect Service Layer (ISL) for TSV-based 3D (b) Silicon Interconnect Service Layer (SISL) for active interposer 2.5D

Similarly, interposers can be used as a platform to provide connection between different die-based IPs. For example, Figure 2.12 shows the 2.5D interposer design with various IP cores sitting on top of an interposer. The ISL interconnect network will be implemented on the interposer layer to provide flexible and reusable interconnect architecture for IP cores which are sitting on top of the interposer.

Note that since this work proposes to use an on-chip network with router design for the ISL, there will be active devices such as logic gates on the ISL. For an interposer-based design, depending on the interposer type, one can implement the on-chip network router on the active interposer, or one can put only the metal routing on the passive interposer but put the active devices for routers in the IP cores (thus reducing flexibility): (1) With current passive interposer technologies, which only have passive interconnect in order to

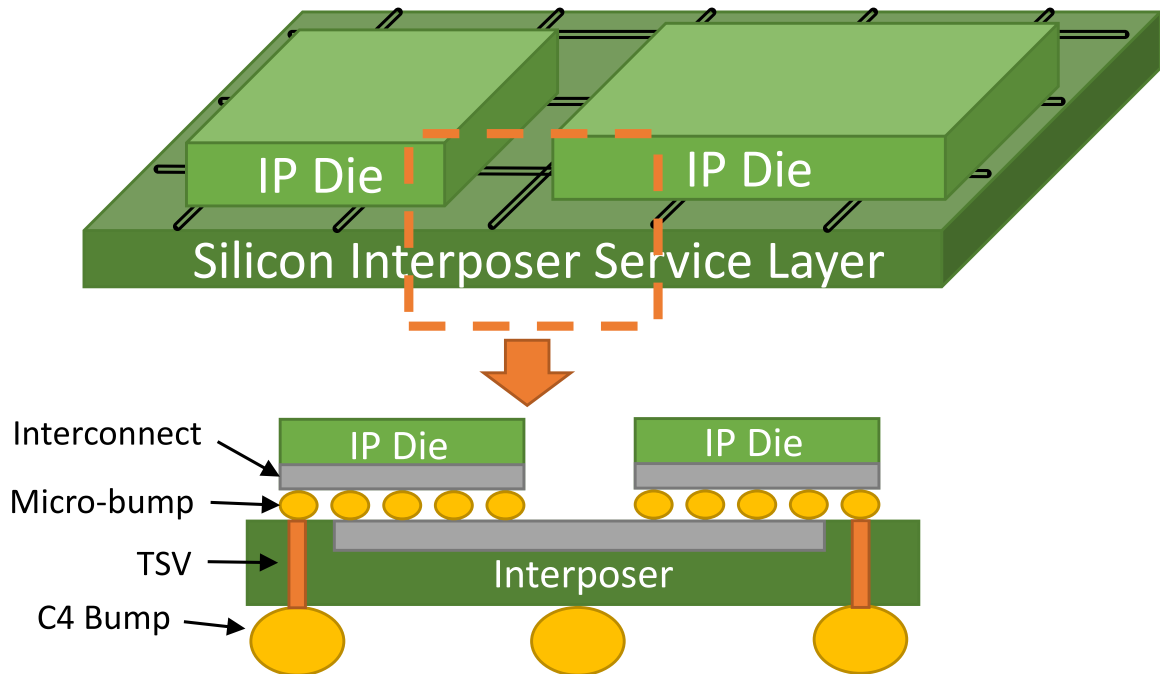


Figure 2.12: Physical connections of 2.5D interposer with ISL on-chip network

minimize process cost and maximize yield, flexible reconfiguration will require for each IP die to dedicate area for active router hardware. (2) Alternatively, active interposers could be used to provide a disaggregated interconnect network with active routing hardware.

Such Silicon Interposer Service Layer (SISL) provides a base platform with a regular network structure and dedicated routers that can be employed to connect any IP dies that are integrated onto the interposer. Dies are bonded to the interposer with micro-bumps at the top metal layer, so physical redesign of existing IP onto the SISL platform may be small and could be achieved by altering only the top metal masks. Active interposers are essentially large traditional CMOS die, and as such do not benefit from the reduced wafer cost of passive interposer processes, but the sparse nature of the active die, with only a percentage of active area, means that yields may be very high despite the large area [23]. An example Silicon Interposer Service Layer topology is shown in Figure 2.11b.

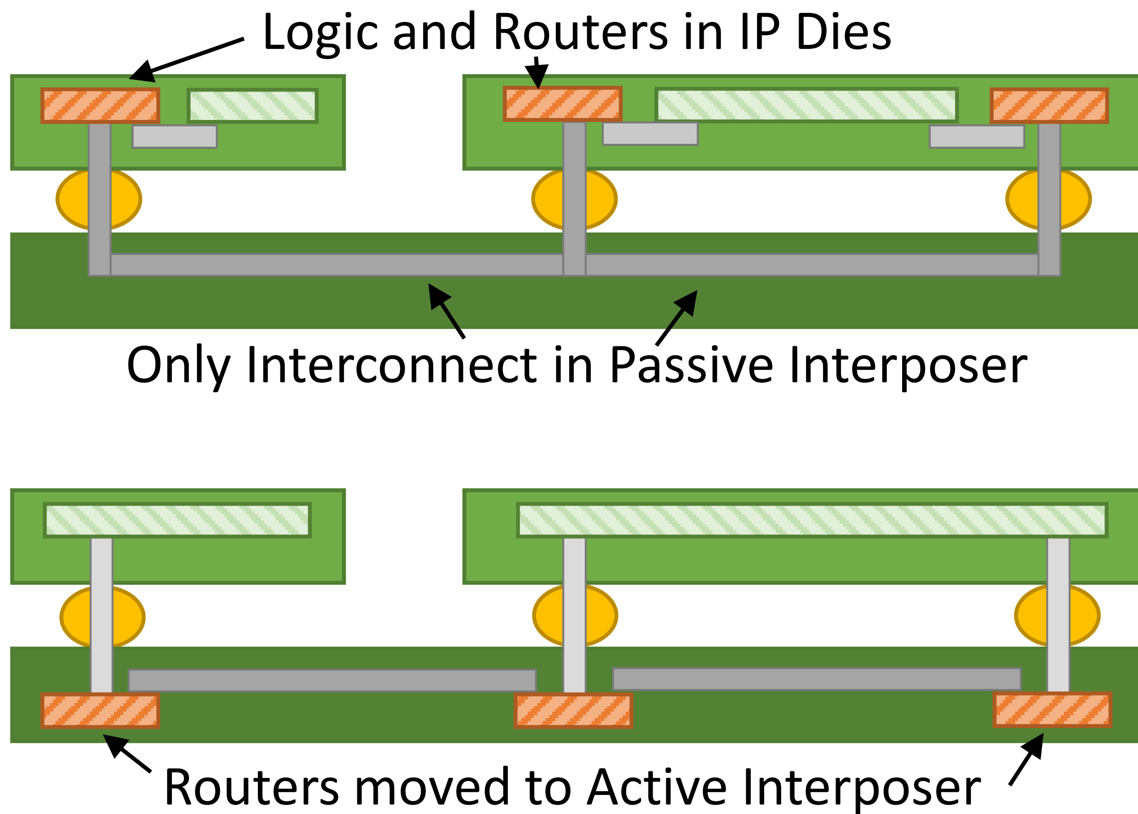


Figure 2.13: SISL implementations with passive and active interposer technologies

For both the 3D ISL and 2.5D SISL platforms, the service layer may contain multiple superimposed heterogeneous networks to provide better communication flexibility and to allow for unnecessary networks to be gated for power. An SISL example is shown in Figure 2.14, where each node represents a logical connection, realized physically by multiple micro-bumps, between the interposer network and the bonded IP die.

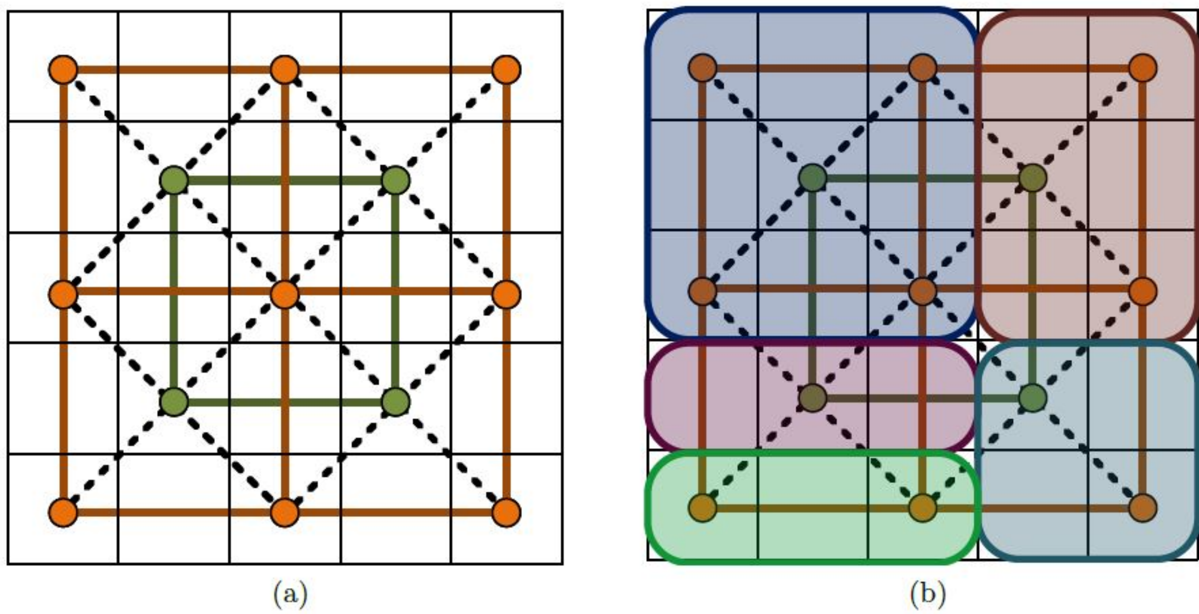


Figure 2.14: Example active network for SISL implementation: (a) 2x2 and 3x3 superimposed mesh topologies, (b) example overlay of IP blocks

Chapter 3

Cost Overhead of Increased Thermal Density in Die-Stacked Systems

As demonstrated in the previous chapter, three-dimensional integrated circuit (3D IC) packaging is a promising technology that advances Moores law, with potential improvements to performance and power and footprint by vertically integrating multiple die [24]. With 3D die-stacking, IC designs can be partitioned across layers for improvements in performance and fabrication cost. Partitioning can reduce interconnect distances between units, thus reducing interconnect delay and power. Smaller die sizes improve silicon yields and allow for higher transistor counts, helping to extend Moore's law even as transistor scaling becomes increasingly difficult.

Despite the opportunities and benefits of 3D IC design, the technology also introduces new challenges. The high transistor density of vertical stacking leads to elevated power density, raising die temperatures and thus requiring more expensive packaging and cooling solutions. Through-silicon vias (TSV) also impose additional area overhead and require design and floorplan constraints to ensure connections between layers. Furthermore, for memory-on-logic 3D stacking, it also requires a close co-design between the logic design

team (such as the CPU/GPU vendors) and the memory design team (such as the memory vendors).

Consequently, even though 3D IC design and architecture have been explored for more than a decade [24, 25], interposer-based 2.5D integration is now emerging as an alternative technology to TSV-based 3D integration. Interposer-based 2.5D integration provides the benefit of close die integration with fewer design and thermal requirements. It also decouples the processor design from the design of the memory stack, reducing the design complexity while improving flexibility. As a result, industry has adopted such 2.5D approaches in commercial products, such as Xilinx's FPGA [26] and the AMD Fury X GPU [27].

When a design strategy (either 2D or 2.5D or 3D design) has to be made, all benefits ultimately have to be justified with cost evaluation. For example, given a specific performance and power targets, which design option will result in a lower cost? Consequently, system-level cost analysis at early design stages is imperative to decide on whether 2.5D or 3-D integration should be adopted. Previous models have provided insight into cost-driven design decisions, but have not included a flexible thermal model for measuring packaging and cooling costs across the range of possible IC designs, which may be necessary given the increased thermal density of die-stacked systems.

In this chapter, the cost models from the previous chapter are expanded with a thermal model is included to determine optimal packaging and cooling costs for full IC systems. The expanded model is then utilized to explore and characterize the design space for the emerging integration options. The model suggests that for a 14nm process, 2.5D and 3D integration becomes feasible for designs larger than 100 mm^2, and power density must be below $0.4W/mm^2$ in order for 3D stacking to be more cost efficient than 2.5D and 2D. The best choice between different 2D, 2.5D and 3D partitioning schemes are then presented across the range of high-performance power densities and gate counts.

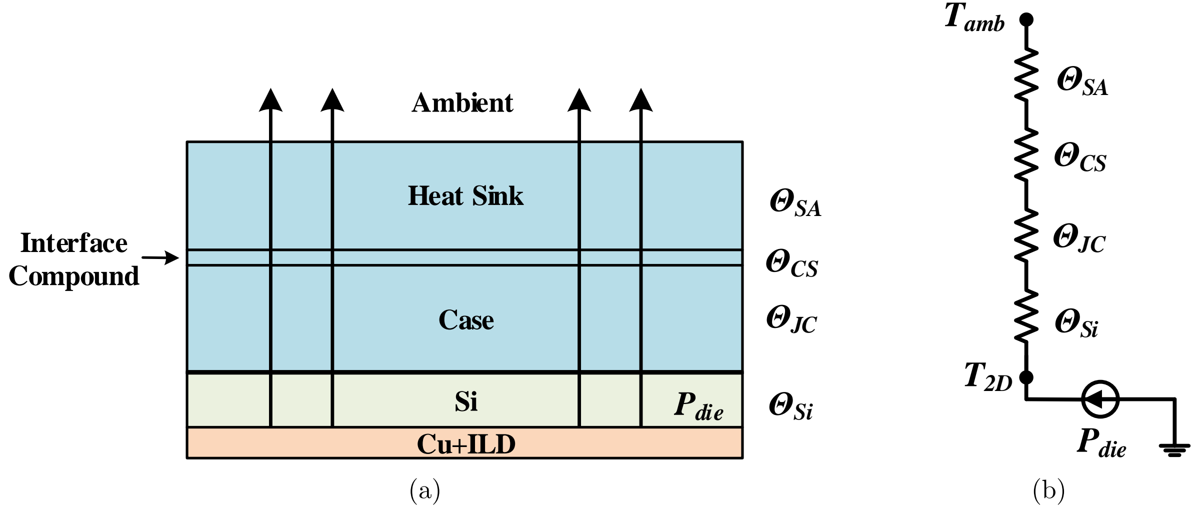


Figure 3.1: 2D thermal model representation. (a) Schematic of the 2D IC thermals; (b) Effective thermal resistance of 2D IC.

3.1 Baseline 2D Thermal Model

The one-dimensional heat equation for a 2D die is given by:

$$T_{2D} = T_{ambient} + (\Theta_{JC} + \Theta_{CS} + \Theta_{SA}) \times P + \Theta_{Si} \times P \tag{3.1}$$

where $T_{ambient}$ is the ambient temperature in ($^{\circ}C$), Θ_{JC} is the junction-to-case thermal resistance, Θ_{CS} is the thermal resistance of the interface compound between the case and heat sink, Θ_{SA} is the thermal resistance between the heat sink and ambient with units in ($^{\circ}C/W$), P is the power dissipation, and Θ_{Si} is the thermal resistance of the silicon layer, where the die is integrated face down. For face-up wire bond packaging, the thermal resistance of the silicon layer is replaced with Θ_{CuILD}, which is the thermal resistance of the metal layers with 50% metalization. For the remainder of this section, it is assumed that the dies are integrated face down unless noted.

Previous studies [28][29] assume heat removal only from the bottom surface of the die via the package and the board. However, greater than 90% of the heat in high-

performance designs may be transferred out of the heat sink [30]. Moreover, package junction-to-ambient thermal resistance Θ_{JA} values from industry are inadequate for this power range. Therefore, the thermal model assumes an external heat sink and disregards heat removal from the bottom surface. The heat escape path of conventional chips is vertical, with active cooling hardware placed on the top of the chip. Power is generated between the Si substrate and metal layers. Thermal resistances Θ_{JC}, Θ_{CS}, and Θ_{SA} contribute to junction-to-ambient temperature, while Θ_{Si} contributes to the junction temperature. Equation (3.1) describes the thermal resistances in series that make up the effective thermal resistance.

The choice of package and heat sink is vital in cooling a high-power die, as both contribute significantly to the maximum average temperature of a chip. The thermal model is integrated into the cost model in order to estimate package and cooling costs. Assuming that the chip can reach the allowed maximum temperature T_{max}, the most cost-effective package and heat sink combination can be found that satisfies this constraint.

3.2 3D Thermal Model

In order to estimate the temperature increase due to stacking multiple active layers, the 2D thermal model is expanded to include power generation at each layer and thermal resistances between stacked dies. The maximum average die temperature is observed at the layer farthest away from the heat sink. The one-dimensional heat equation of a 3D stacked die with n active layers, is therefore given by [28]:

$$T_{3D} = T_{ambient} + \sum_{i=1}^{n} (\Theta_D (\sum_{j=i}^{n} P_j)) \tag{3.2}$$

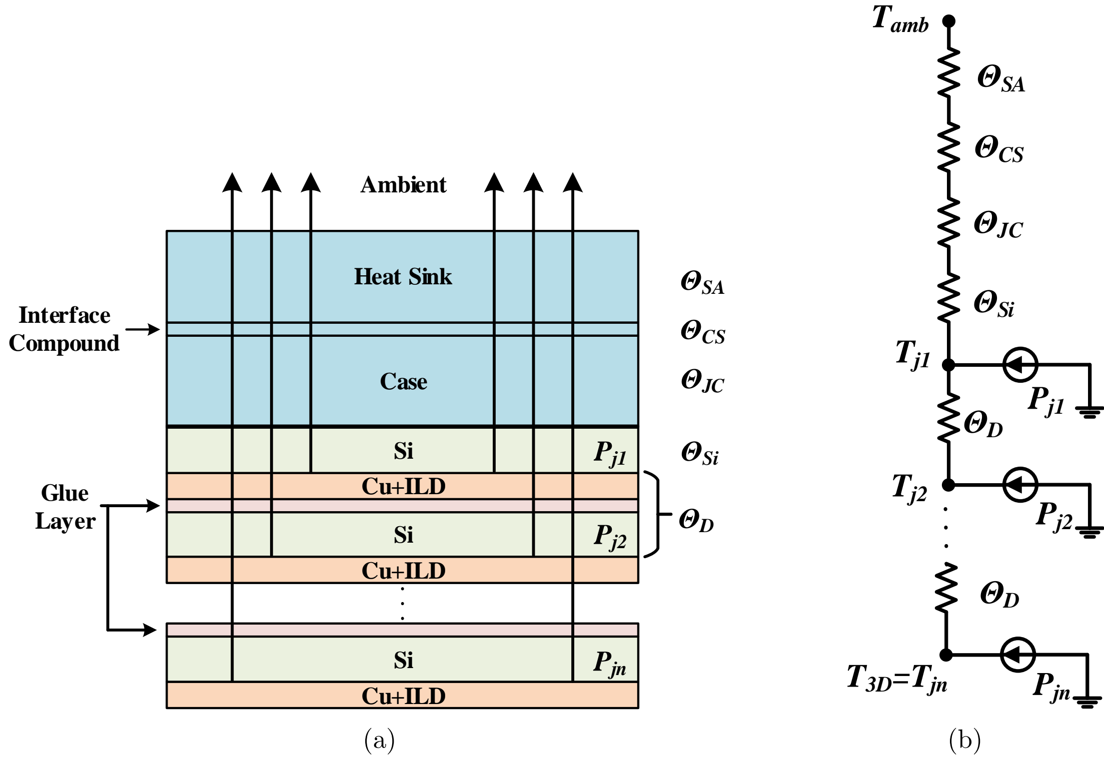


Figure 3.2: 3D thermal model representation. (a) Schematic of the 3D IC thermals; (b) Effective thermal resistance of 3D IC.

where Θ_D is the thermal resistance between the $(i-1)$ and i^{th} layers, calculated as below:

$$\Theta_D = \begin{cases} \Theta_{JC} + \Theta_{CS} + \Theta_{SA} + \Theta_{Si}, & \text{if } i = 1. \\ \Theta_{Si} + \Theta_{glue} + \Theta_{CuILD}, & \text{if } i \neq 1. \end{cases} \tag{3.3}$$

Figure 3.2 illustrates the 3D thermal model. The thermal resistance between 3D-stacked dies takes into account the resistances of silicon, glue, and metal layers. According to equation (3.3), the die temperature of lower layers are also affected by the power dissipation and the effective thermal resistance of the layers above. Compared to conventional dies, 3D integration results in higher die temperatures and therefore requires better packaging and more aggressive cooling solutions to maintain the same maximum allowed die temperature.

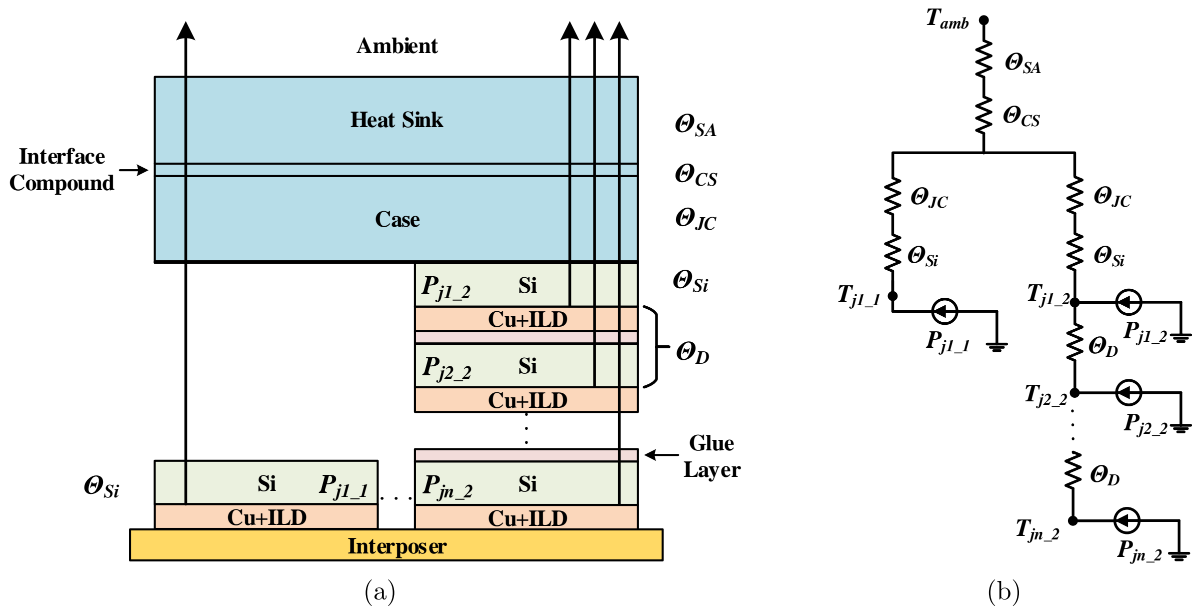


Figure 3.3: 2.5D thermal model representation. (a) Schematic of the 2.5D IC thermals; (b) Effective thermal resistance of 2.5D IC.

3.3 2.5D Thermal Model

In order to estimate the maximum temperature of a 2.5D die-on-silicon interposer, the 3D thermal model is expanded to consider multiple die stacks. Figure 3.3 describes the 2.5D thermal model in which separate stacks share the same junction-to-ambient thermal pathway but have different stack temperatures in *parallel*. The stack with the maximum die temperature determines the upper bound for cooling and package cost of 2.5D integration.

Assuming the homogeneous partitioning of a 2D design retains the same power density per partition, it is evident from our thermal model that an all-single-layered 2.5D partitioning of a design will yield similar thermals as its 2D counterpart. On the other hand, homogeneously partitioning a 2D design to form a single stack will yield the same thermals as the 3D integration. Our thermal model confirms that any 2.5D design in between these two arrangements will yield thermals greater than 2D and less than 3D

integration, as also stated in [31].

3.4 Cooling and Package Cost Estimation

When high-performance thermal management is required, the choices of packaging and cooling solution are major contributors to the final system cost of the integrated circuit. As shown in Equation (3.2), the package and heat sink thermal resistances are modeled in series, and thus sufficiently low thermal resistances for both are required for proper operation of the chip.

To determine heat sink cost and effectiveness, commercial cooling solutions were surveyed across a design range that included passive fin heat sinks, heat sinks with powered fans, complex heat pipe coolers, and high-end liquid coolers [32] [33]. A continuous cost-versus-thermal resistance curve was extracted from the cooling solution data, shown in Figure 3.4, to estimate a heat sink cost given a required thermal resistance. There is a range of options available for heat sinks with varying thermal resistances. The low-cost end includes passive fanless heat sinks and low-performance airflow-cooled heat sinks. Large-volume fan-cooled heat sinks with fins provide higher performance, and liquid coolers top the cooling solutions in performance. Note the steep cost increase as thermal resistance approaches $\Theta_{SA} = 0.07\ ^{\circ}C/W$ and the lack of commercial solutions beyond this point, suggesting a limit to the currently available heat sink capability.

Package cost is determined by multiple design factors, but the package technology type has the greatest influence on the thermal resistance of the package and on the overall package cost. A package type can be selected to meet thermal resistance requirements, which then determines the scaling of other package cost contributors. During system optimization, the model selects the most cost efficient package/heat sink combination to achieve the necessary thermal resistance. Cost-efficient package types included in the

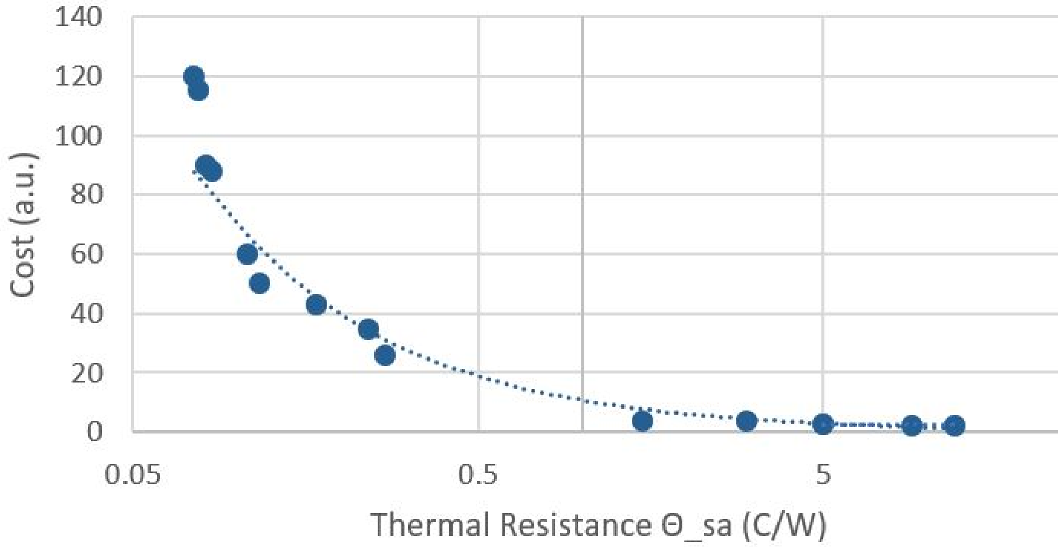


Figure 3.4: Cooling cost versus efficiency

model are pBGA, fcBGA, and cBGA, with thermal resistances of 0.44, 0.20, and 0.03 $°C/W$, respectively [34].

For a selected package type, other cost parameters include package area, pin count, substrate layer count, and package volume. Package area is determined from the chip area, pin count and substrate layer count are determined by IC electrical requirements, and package volume depends on the target market. The package cost $C_{package}$ can be calculated by:

$$C_{package} = \mu_V(\mu_L N_L)[C_{base} + \mu_A A + \mu_p * N_p] \tag{3.4}$$

where μ_V is a market volume scalar, N_l and μ_L are the substrate layer count and scaling, C_{base} is the base package cost for the selected type, A and μ_A are the chip area and scaling, and N_p and μ_p are the pin count and scaling.

Table 3.1: Package types with thermal resistances.

Package Type	Thermal Resistance ($°C/W$)
pBGA	0.44
fcBGA	0.20
cBGA	0.03

3.5 Thermal-Aware Design Space Exploration

Although our results show that the silicon cost of 3D integration is consistently less than that of 2.5D integration because of the interposer overhead, the introduction of thermal-dependent packaging and cooling costs results in a new cost-driven design space. Figure 3.5 shows the system costs across a range of gate counts and power densities for a 14nm process, pin count of 1150, max junction temperature of $100°C$, and ambient temperature of $30°C$. The highlighted values in green reflect the best design choice at a given design size and power density. The value marked in red is too hot to cool with conventional thermal management solutions.

For 14nm designs smaller than $100mm^2$, 2D design is the most cost effective because of minimal fabrication overheads and efficient cooling. 3D stacking is the most cost efficient only when power densities are at or below $0.4W/mm^2$. For reference, average mobile microprocessors have a power density of $0.2W/mm^2$ and desktop CPU and GPU parts have power densities from $0.3-1.0W/mm^2$. At all other power densities and gate counts, 2.5D integration is more cost-efficient than 2D and 3D integration because of the balance of yield improvements from die partitioning and reasonable thermal management. For closer inspection, Figure 3.6 shows the relative cost breakdown between chip fabrication and package/cooling costs at $0.4W/mm^2$ and $200mm^2$.

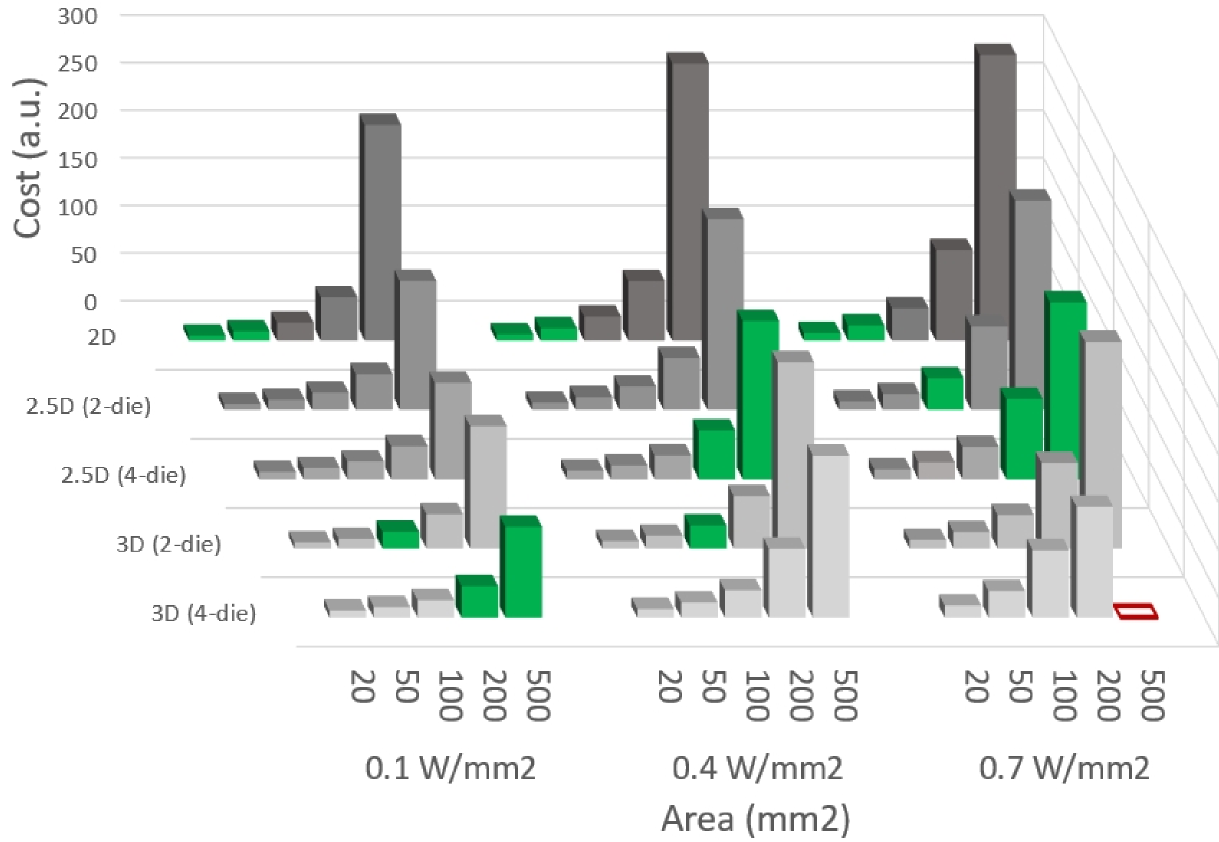


Figure 3.5: Full 14nm system costs of 2D, 2.5D, and 3D designs at different power densities

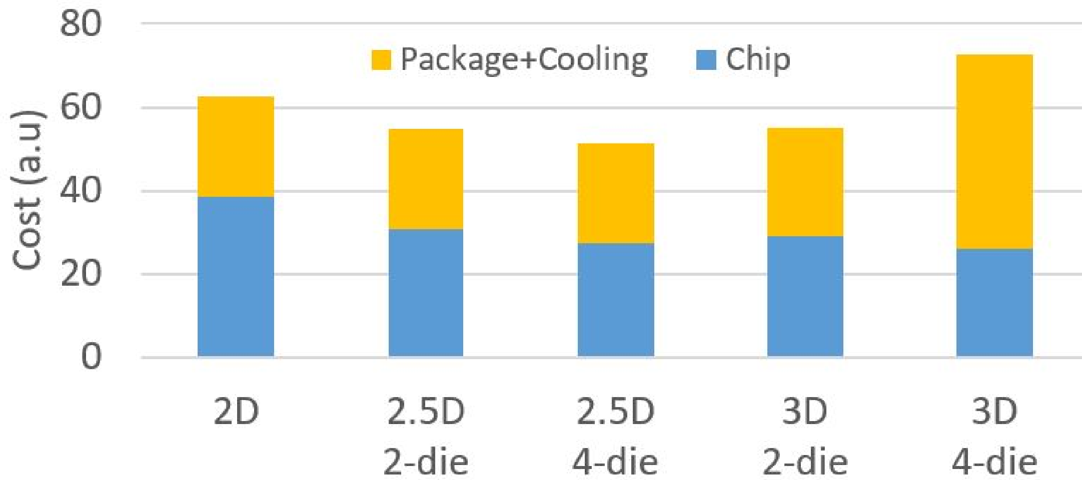


Figure 3.6: 14nm Cost Breakdown at $0.4W/mm^2$ and $200mm^2$

3.6 Conclusion

Cost analysis at the early design stage is the key to help decide the design strategy of using either 2.5D or 3D integration. This chapter presents a system-level cost model to compare the silicon fabrication, packaging, and cooling costs between 2D, 2.5D, and 3D systems. The complete cost model shows that 3D integration is still more cost-effective than 2D at lower power densities, but 2.5D integration offers the most cost-effective solution for designs with more than 100 million gates with a wide range of power densities.

Chapter 4

Power Modeling and Projection of Future Die-Stacked Dynamic Memory

DRAM memory bandwidth remains a key system bottleneck, especially for emerging memory-intensive exascale and deep learning applications. To improve memory bandwidth and interface efficiency, die-stacked memories like High Bandwidth Memory (HBM) have been developed for in-package integration with GPU, FPGA, and ASIC systems. By vertically stacking multiple DRAM dies and interfacing over fine-pitched interposer packages, the second generation of HBM technology can achieve peak bandwidths of 256 GB/s per stack with 3.5x better energy-efficiency than GDDR5 [35]. Despite these efficiency gains, the demand for memory bandwidth has not subsided. Future memory standards thus face the challenge of improving efficiency at the same time that DRAM technology approaches limits to scalability [36]. Further, this memory integration brings additional power and heat directly into the processor package, complicating the already difficult challenge of heat removal. Accordingly, the power consumption of die-stacked

memories is now a critical constraint.

Although HBM is becoming increasingly important to commercial designs, few details about HBM power are available to the research community. To provide insight into the energy expenditure of modern stacked memories, this work presents a hardware-validated power model, based on previously-validated architectural power methodology [37], for the family of High Bandwidth Memory configurations. The model is then employed to profile the power breakdowns of several HBM2 configurations across the range of real and synthetic workloads. Looking forward, the power of potential future memory configurations is also projected based on expected technology trends. After analyzing these future configurations with architecture-appropriate traces, this work suggests that direct capacity and frequency scaling of the existing HBM2 architecture could increase memory power by as much as **42%**, highlighting the need for increased efficiency in future memory architectures.

4.1 High Bandwidth Memory

To improve memory interface efficiency, 3D memory stacks with multiple DRAM dies can be placed adjacent to the compute die using 2.5D interposer packages [35], as visualized in Figure 4.1(b). This packaging allows for very wide interfaces, while simultaneously reducing the signal distance and frequency. The initial JESD235A JEDEC High Bandwidth Memory standard (HBM2) can provide 256 GB/s of bandwidth for each memory stack [38], while the JESD235B update employs higher frequencies to provide up to 307 GB/s [39].

Unlike DDR and GDDR, a single HBM2 stack contains eight 128-bit wide asynchronous data channels and a large capacity in a much smaller footprint: 4 GB for a four-high stack or 8 GB for an eight-high stack. To increase bandwidth, HBM2 splits

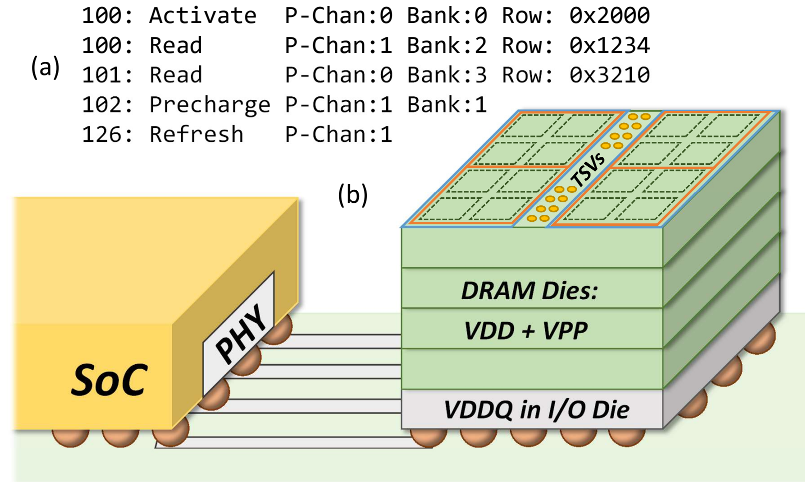


Figure 4.1: (a) Example memory command trace for one channel. (b) SoC with four-high HBM stack, with HBM voltage domains VDD, VDDQ, and VPP as well as SoC PHY. On the top die, two channels (blue), are shown with two pseudo-channels (orange), each with four banks. HBM2 may have eight to thirty-two banks per pseudo-channel.

each channel into two 64-bit pseudo-channels that share a command interface and clock signals but access discrete memory spaces. Peak memory bandwidth requires alteration between the two pseudo-channels, as demonstrated in Figure 4.1(a). To better utilize the shared interface, HBM2 also introduces auto-precharge and implicit precharge behaviors to reduce row command contention. The row (activation, precharge, refresh, power-down) and column (read, write) interfaces are also separate in HBM2, so one pseudo-channel can dispatch a row command while the other one dispatches a column command.

In other ways, HBM is similar to other SDRAM memories. To access data, a row in the pseudo-channel must first be activated and moved to the row buffer for further access. After a delay, column read and write commands can then communicate a burst of four 64-bit values with the row buffer. Before accessing a new row, the current row must be closed with a precharge command. In HBM2, a precharge can also be performed by setting the auto-precharge flag to a column access or, with implicit precharge, by activating a different row. To improve parallelism, memory is partitioned into eight to

thirty-two banks (based on channel capacity and stack height) with unique row buffers. All banks must be periodically refreshed by using either individual single bank refreshes, a bulk refresh for all banks, or by entering a low-power self-refresh state.

4.2 HBM Power Modeling Methodology

To provide detailed power breakdowns across the range of HBM configurations and workloads, a memory power model was created and validated. The power model is an *architectural* model designed to compare memory architecture configurations and memory traces, as opposed to circuit-level models that return the energy of individual memory operations [40][41]. The HBM power model is based on the previously-validated methodology developed by the DRAMPower architectural memory power models [37][42], but rewritten to provide accurate support for the family of High Bandwidth Memory architectures. As in DRAMPower, the HBM power model requires canonical memory power measurements, termed *IDD Specifications* in the memory standards [38], as inputs. The *IDD* inputs can be provided by the vendor, measured from commercial hardware, or estimated using a circuit-level model [40]. By specifying architectural parameters, timing constraints, and the *IDD* specifications, the dynamic energy for each memory event (e.g., activation, precharge, read) and the static power for each idle or active power state (e.g., bank open, closed, power-down) can be calculated for the specified HBM configuration. For example, open bank static power is calculated by normalizing the *IDD3N* benchmark by the number of banks, while read energy is calculated by subtracting *IDD3N* from *IDD4R* and normalizing by the burst length and data rate [37]. The model supports all pseudo-channel timing interactions specified in the JEDEC JESD235A standard [38], including refresh requirements, as well as HBM commands that are not supported in earlier memories and prior power models, including auto-precharge, implicit precharge,

and single bank refresh.

The model also includes the power contribution from the PHY on the SoC, as visualized in Figure 4.1(b), as this is also a significant power draw that scales with memory intensity. The PHY power contribution includes the I/O power for write commands, as write data must be transmitted to the memory. PHY power contributions during read, write, and idle behavior are measured using Synopsys PrimeTime for a tape-out quality PHY macro in a leading-edge process technology.

The model can use the architectural and timing specifications to generate synthetic traces with parameterized characteristics, including average bandwidth, average sequential accesses per row, read-write ratio, and bulk or single bank refresh. The model can also analyze a memory controller instruction trace from a simulator like gem5 [43]. Each channel trace is analyzed to record all dynamic memory events and all bank state transitions before calculating the power distributions.

The architectural HBM power model was validated by collecting power measurements from GPU test hardware. Results for a maximum activity memory power virus with closed-row read requests and standard refresh timing, representative of a peak power scenario, are shown in Figure 4.2. Error bars indicate one standard deviation of model *IDD* inputs and inter-stack measurement variation. Power error for the model was -6% for a four-high configuration and -11% for an eight-high configuration, indicating that the model slightly underestimated the peak power magnitude. However, the relative contributions of each voltage domain accurately match between modeled and measured power profiles, within less than 2% difference.

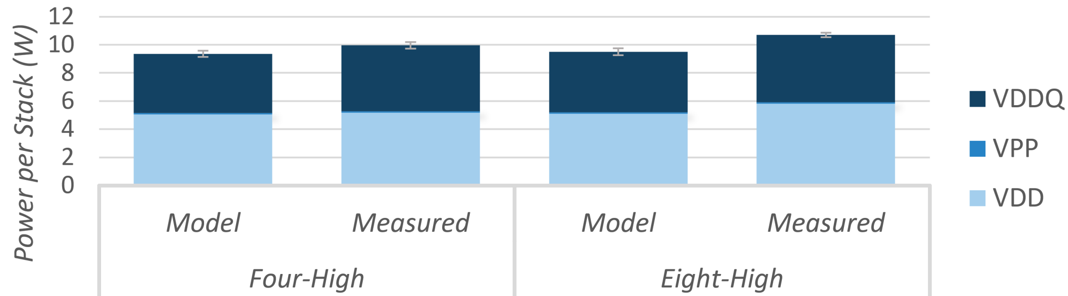


Figure 4.2: Power model validation for memory power virus benchmark for four-high and eight-high configurations, with voltage domain breakdown.

4.3 Power Profiling for HBM2

To highlight the range and breakdown of HBM2 power across benchmarks, this section provides a detailed power analysis for a 4 GB HBM2 stack composed of four 8 Gb DRAM dies, running at 1000 MHz interface frequency. The model was configured with JEDEC standard architectural specifications and voltages [38], with core voltage $V_{DD} = 1.2V$, I/O voltage $V_{DDQ} = 1.2V$, and row access high voltage domain $V_{PP} = 2.5V$. The PHY has matching I/O voltage $V_{DDQ} = 1.2V$. Power values are calculated from architecture-specific synthetic benchmarks generated with the specified total bandwidth. The notation Read_X and Write_X specify the row locality, where X is the average number of read/write requests per opened row.

Low-activity power results are shown with breakdown by command type in Figure 4.3(a), where "Static" is static leakage power, "Row" is power from row commands (activate, precharge, and refresh), "Column" is power from read or write commands, and "PHY" is the power from the SoC PHY. *Self-Refresh* is the lowest data-retaining idle power state for HBM2, with disabled clocks and internally-timed refresh. *Precharge Power-Down* is a low power state with all rows closed and disabled clocks, but the clock must be periodically restarted to execute a refresh. *Idle* is a similar case, but the clock is not disabled. As expected, static power dominates the low-activity cases, but dynamic

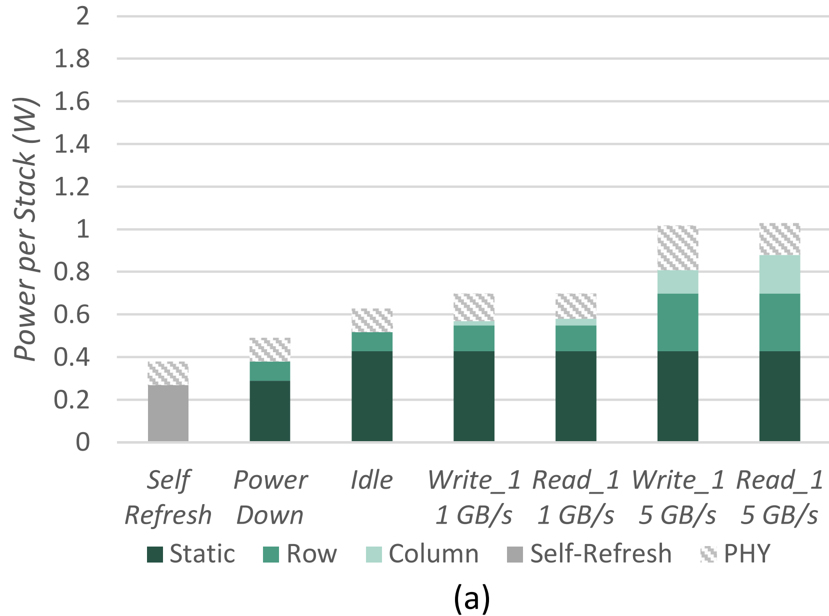


Figure 4.3: Command type power breakdowns for HBM2 four-high 4 GB Stack by for (a) low memory activity. Please note that power range (2W) differs from that of Figure 4.4 (10W, 16W). Write_X signifies row locality, where X is the average number of writes per open row.

power is equally important even at low bandwidths of 5 GB/s per stack.

Next, Figure 4.4(b) shows results for two example compute kernels: SNAP, a particle transport kernel, and DGEMM, a matrix multiplication kernel. Each kernel is run on two input sizes, with statistics listed in Table 4.1. Memory traces for the kernels were generated with gem5 [43] and an industry in-house SoC simulator. These kernels demonstrate low spatial locality, wide variability in read/write ratio, and wide variability in bandwidth and memory power even between different input sizes of the same kernel.

Finally, Figure 4.4(c) shows the power at the maximum bandwidth for row localities of one, two, and eight sequential accesses per row. Note that maximum bandwidth in each benchmark is limited from reaching the theoretical 256 GB/s peak because of refresh requirements, and the random access scenario (one access per row) is limited because each activation command requires two cycles to complete. As expected, random

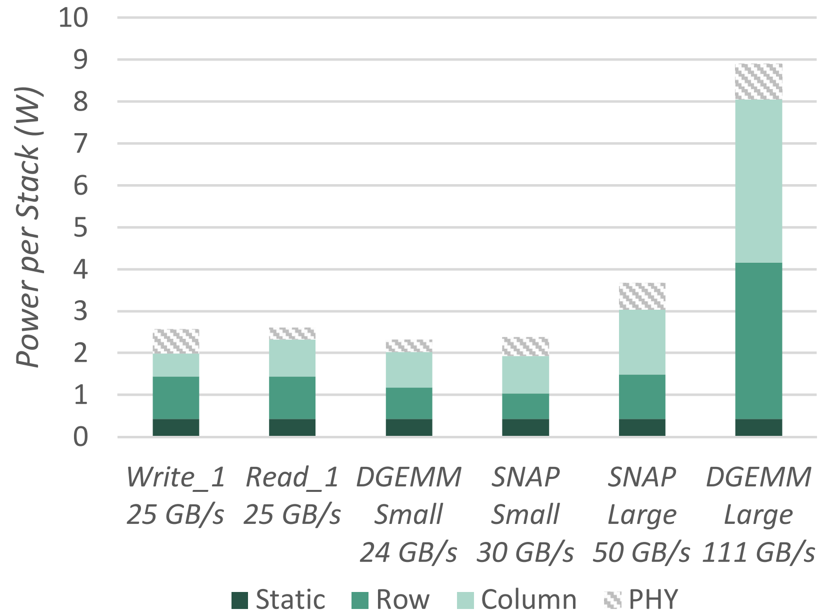
Table 4.1: Kernel memory trace behaviors

Kernel	Size	Avg. BW (GB/s)	Read/Write Ratio	Avg. Access Count Per Row
SNAP	Small	30.0	2.10	2.15
SNAP	Large	49.8	2.49	1.87
DGEMM	Small	24.3	9.87	1.34
DGEMM	Large	110.9	167.4	1.15

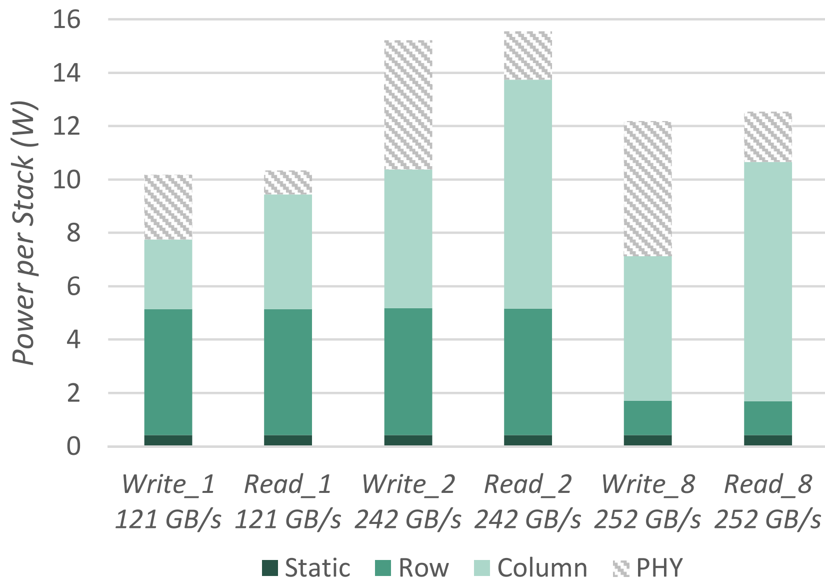
access behavior has significant Row access power, with no reuse of the row buffer, while sequential accesses are dominated by data transfer energy from Column commands. With two column accesses per row, similar to the SNAP and DGEMM kernels, the total power between the HBM2 and PHY reaches 15.6W per stack at maximum bandwidth (13.6W from memory), which is a significant power contribution when a system may have multiple memory stacks. As the number of accesses per row increases to eight, efficiency and bandwidth improve as the row buffer is better utilized and the number of activations and precharges per read/write decreases. Efficiency at the maximum bandwidth, including PHY power, is 6.1 pJ/bit.

4.3.1 Stack Height Sensitivity

Memory capacity in the same footprint can be increased by increasing the number of DRAM layers per stack. In this section, the four-high HBM2 configuration with 4 GB capacity is compared against an eight-high configuration with 8 GB capacity and the same 1000 MHz interface frequency and voltage domains. The eight-high stack doubles the capacity and number of banks, but otherwise has identical architecture and timing constraints (except when issuing consecutive reads between the top and bottom groups of four DRAM layers). Power results are shown in Figure 4.5, with annotation showing the percent increase in total power. Static power increases by about 1.4x due to the increased capacity, and there are also small increases in the row power (about 1.1x), partially due



(b)



(c)

Figure 4.4: Command type power breakdowns for HBM2 four-high 4 GB Stack by for (b) example kernels, and (c) high memory activity. Please note that power range differs (10W, 16W) between each figure and Figure 4.3 (2W). Write_X signifies row locality, where X is the average number of writes per open row.

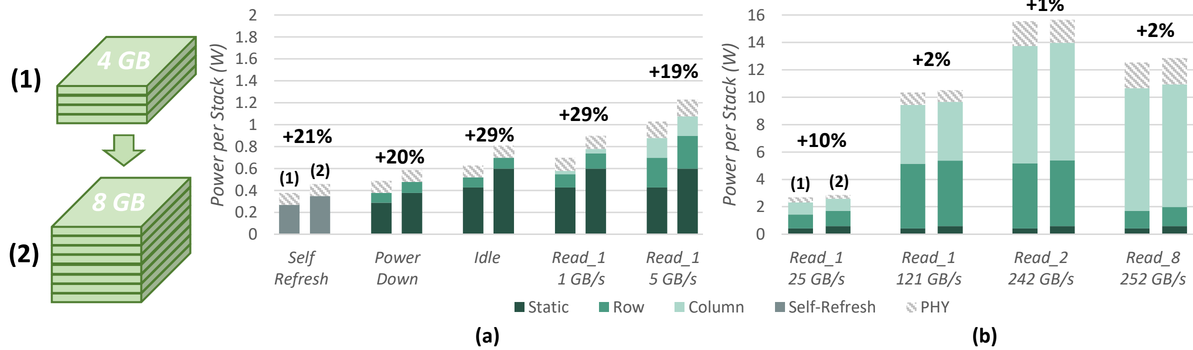


Figure 4.5: Stack height sensitivity: (1) HBM2 four-high 4 GB power versus (2) HBM2 eight-high 8 GB power, by command type. Please note that power ranges differ (2W, 16W) between (a) and (b).

to the increased TSV resistance to the highest layers. Although increasing the stack height substantially increases the power of low-activity cases, the power of high-activity cases increases much less. This is especially true for the column-dominated activity of high-locality high-bandwidth scenarios, as increasing stack height has minimal impact on read and write power. Efficiency at maximum bandwidth for the eight-high stack, including PHY power, slightly decreases to 6.2 pJ/bit.

4.4 Projection for Future Memory Power

Although the initial HBM2 specification can achieve bandwidths of 256 GB/s and DRAM densities of up to 8 GB per stack, these metrics may be insufficient to meet the demands of high-performance applications in the exascale era [44]. As evidence, the JESD235B update increased the maximum stack height to twelve DRAM layers and increased the interface frequency to 1200 MHz to achieve 307 GB/s [39]. If HBM2 can already contribute more than 15W per stack, then future memory standards must consider power as a critical constraint.

This section analyzes a future memory that continues the HBM2 scaling trends of

capacity and frequency improvement, as seen in the JESD235B update [39]. Specifically, this configuration: 1) increases the interface frequency from 1000 MHz to 1400 MHz, 2) shrinks DRAM process technology by one node, 3) increases die capacity from 8 Gb to 12 Gb, and 4) increases stack height from eight-high to twelve-high. Without any major changes to internal architecture, an increase in interface frequency by 1.4x also necessitates an equivalent increase of the internal frequency. This configuration is closely compatible with existing HBM2 memory controllers. This configuration will be referred to as *HBM358* in reference to the increased theoretical bandwidth of 358 GB/s per stack.

To project from existing HBM2 to future stacked memory configurations, each voltage domain component of each *IDD* input was scaled based on architecture and technology factors. *IDD* parameters were scaled for the stack height, interface frequency, internal frequency, technology scaling, and die capacity. The 1.4x increase in interface and core frequencies is modeled as a 1.4x linear scale on all voltage domain components of the *IDD4R* and *IDD4W* read and write measurements, reflecting the increased activity. The process technology shrink, based on recent technology projections [36] and results from circuit-level memory modeling [40], acts as a linear scale of approximately 0.9x on all V_{DD} and V_{PP} domains. Increasing die capacity by 1.5x impacts the V_{DD} and V_{PP} domains as a linear increase of 1.5x for static power and refresh (as more cells leak and must be refreshed, respectively), but as a square-root scale ($\sqrt{1.5}$) for column access and row activation due to greater interconnect distances and larger bank sizes. Finally, stack-height scaling is calculated based on the scaling delta between the HBM2 four-high and eight-high configurations: a 1.4x linear scale on static power components and a 1.1x linear scale on activation and refresh, both for the V_{DD} and V_{PP} domains. For the SoC's PHY, idle power undergoes no change, while read and write scale linearly with the frequency increase.

Figure 4.6 shows the projected *HBM358* power for low and moderate activity memory

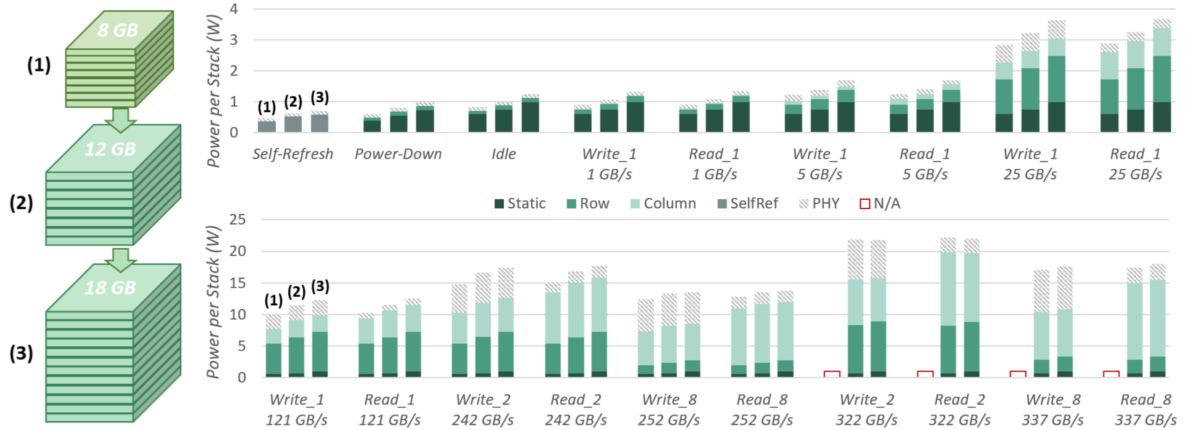


Figure 4.6: Scaled-up HBM358: Power comparison between (1) eight-high 8 GB HBM2, (2) eight-high 12 GB HBM358, and (3) twelve-high 18 GB HBM358. Note that no results can be shown above 252 GB/s for the HBM2 configuration (1) due to HBM2 bandwidth limits.

behavior (top) and for high activity behavior (bottom). For each scenario, the HBM2 eight-high configuration is compared against two *HBM358* configurations: one eight-high stack and one twelve-high stack, both with larger 12 Gb die capacity, greater frequency, and smaller DRAM technology. The increase in die capacity and stack height leads to a significant increase in static power for low-activity behavior, as well as a noticeable impact on row power for moderate-activity behavior. Power also increases for high-activity cases, and in new high-bandwidth cases (322 GB/s per stack with two reads per open row), the total HBM and PHY power reaches 22.2W, a 42% increase over maximum HBM2 power of 15.6W. Power at maximum bandwidth can be up to 18.0W (including PHY power), with efficiencies of 6.4 pJ/bit for the eight-high stack and 6.5 pJ/bit for the twelve-high stack.

The significant power at maximum bandwidth, dominated by data transmission power, could be addresses in future standards through interface voltage reduction, improved bus encoding, and other signaling techniques. However, the peak power scenario, when row locality is not as high, also has a significant power component from row acti-

vation and precharge. Solutions to this challenge include memory architecture changes to page size and row size to reduce energy per activation, as well system and software improvements to improve row locality and reduce the number of activations.

4.5 Conclusion

Emerging systems will require greater memory capacity and bandwidth, making stacked memories an increasingly critical component. Memory power has traditionally been overshadowed by processor power, but the results from this validated memory power model suggest that memory power now needs to be addressed through circuit and architecture improvements. While today's HBM2 can contribute more than 15W per stack, projections for future HBM scaling suggest peak-bandwidth power up to 18W and new power peaks of 22W during reduced-locality traffic. These projections highlight the need for continued research in efficient 3D memory architectures to improve leakage power and row access power, I/O and packaging design to improve transmission power, and memory management to improve row locality.

Chapter 5

System Integration with Interposers: Active versus Passive Technology

Cutting-edge high performance systems demand larger and denser processors, but future lithographic nodes are expected to introduce higher manufacturing costs and yield challenges. Die-level integration technologies like passive interposer-based 2.5D have demonstrated the potential for cost reduction through die partitioning and yield improvement, but system performance and scalability may be impacted. Alternatively, active interposer technology, the intersection of 3D and 2.5D methodologies, can provide higher performance interconnect networks to integrate chiplets, but the active interposer die is itself subject to cost and yield concerns. In this section, a cost and performance comparison is performed between systems manufactured with traditional monolithic 2D fabrication, 2.5D passive interposer integration, and 2.5D/3D active interposer integration to demonstrates the trade-offs between interposer types for current and future high performance systems. This section compares the relative cost and performance scaling trade-offs of passive and active interposer dies for several potential systems, demonstrating that both methodologies can indeed provide cost-effective integration for different system require-

ments. Additionally, the work demonstrate how the extra ”prepaid” silicon area of the interposers can be leveraged for fault tolerance to improve yield and cost-effectiveness. Next, the design space of Network-on-Interposer architecture is explored for both active and passive interposers to compare communication performance versus process technology and cost, while also revealing which network topologies are best suited for interposer integration.

5.1 Motivation for Chiplet Partitioning

As outlined in the ITRS 2.0 roadmap [45] [46], the datacenter and microserver markets demand increasingly performant and localized processing, with a roughly 3x increase in available memory and 4x increase in the number of processor cores per socket and rack unit, respectively, over the next 10 years. Similarly, the push for high-performance exascale supercomputing will likely require complex heterogeneous SoCs with many cores and integrated memory to provide sufficient bandwidth and data localization to meet efficiency requirements [47]. Modern manycore server processors from Intel and AMD, such as the 32 core AMD “Naples” processor, demonstrate that the industry is indeed moving in the direction to meet these datacenter and microserver demands.

Unfortunately, the ability to meet these demands with conventional process scaling is becoming increasingly difficult and expensive. The Moore’s Law cadence target has already been missed, with almost all foundries are no longer able to meet desired transistor scaling in the most recent process nodes [20] and future process roadmaps slowing for each new node [48]. Increased process complexity has led to more expensive fabrication and longer manufacturing cycle times [49], and as transistor cost reduction slows, yield and endurance challenges grow, and cost per area increases [50] [21], it becomes increasingly costly to meet the market requirements for denser, larger integrated circuits.

As Moore's Law slows and cost per semiconductor area increases [21], computer architects will need to utilize new architectures and packaging technologies to achieve the system improvements necessary to meet these demands. Recently, the concept of multi-die integration has received renewed attention as a promising solution to these challenges. Instead of the monolithic fabrication of modern Systems-on-Chips (SoC) onto a single large die, die-level integration technologies integrate multiple semiconductor dies, each fabricated individually, into a single package, as demonstrated in Figure 5.1. This concept, proposed as far back as the original Moore's Law paper [1], has long been considered as a method to improve semiconductor yields through the use of Known Good Die (KGD) validation, ensuring the functionality of each die before integration. While a single critical defect can disable the functionality of an entire monolithic system, pre-bond validation of each smaller die reduces the loss per defect and results in a greater number of fully-functional systems [51].

Historically, multi-chip module (MCM) packages have been used to integrate multiple dies onto a single substrate, and recently MCMs have again been deployed to improve yield and provide product scaling [52]. Although MCMs can provide a platform for die integration, the coarse-pitch substrate interconnect can only provide limited bandwidth, with reduced efficiency and increased latency, compared to on-chip interconnect [53]. However, these limitations can be solved by instead using fine-pitched silicon interposers, which have already come to market for several high-end devices, including the AMD Fiji GPU with High Bandwidth Memory integration for improved performance, efficiency, and footprint [18] and the Virtex-7 FPGA from Xilinx [54] with multiple FPGA slices for improved yield and with heterogeneous transceiver chiplets for configurability and performance. These interposers utilize standard semiconductor interconnect technology, such as that in the 65nm process node, and bond each die using fine pitched microbumps, with current bump pitch of 55μm and future pitches of 20μm [55]. With interposer

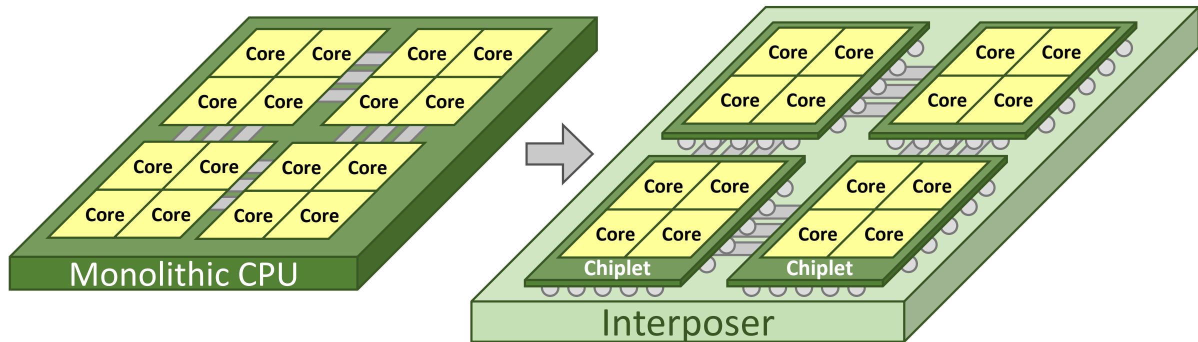


Figure 5.1: Transition from monolithic manycore CPU to interposer-based 2.5D system with multiple chiplets

integration technology, systems can realize the yield and flexibility benefits of multi-die integration while maintaining the high performance Network-on-Chip (NoC) fabric utilized to connect modules in modern SoCs. However, the usage of interposers has so far been limited to these niche cases, while the wider high-performance market could potentially benefit from Network-on-Interposer adoption. In prior analysis of cost-driven design methodology, both 2.5D and 3D designs were shown to have lower post-yield manufacturing costs than 2D SoCs for midsize and large systems, but only 2.5D designs were cost effective for high power designs, while 3D suffered from increased packaging and cooling costs when thermal management was considered [56].

5.2 Interposer Technology Selection

While recent research has demonstrated the cost and performance benefits of interposer-based system integration [57][19][58] [51][59], the details of the interposer technology and design are still an area of development and debate. Commercial interposers are today manufactured as *passive* silicon interposers [18], which contain metal interconnect but do not have active CMOS transistors. The simple nature of passive interposers greatly reduces wafer costs, but without transistors the interposer can only provide non-repeated

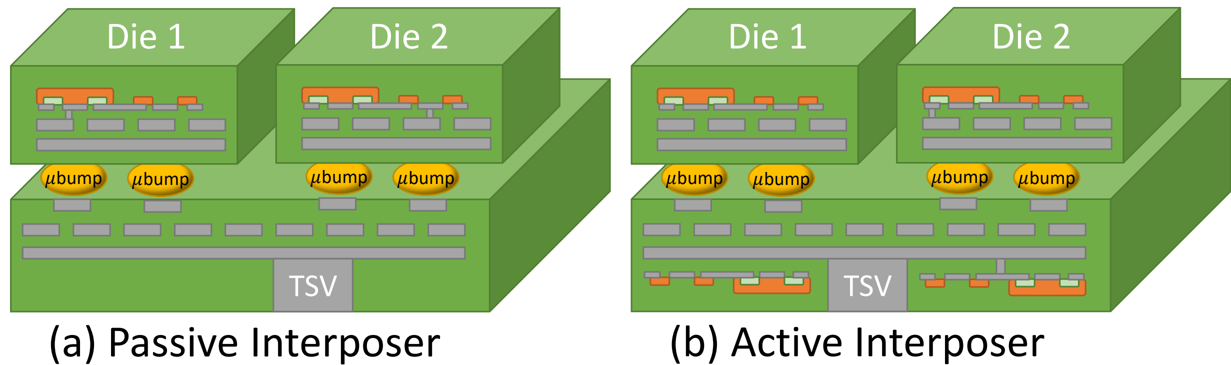


Figure 5.2: Illustrative two chiplet system, integrated with microbumps using (a) Passive interposer with only passive interconnect and TSVs (b) Active interposer with active CMOS logic.

point-to-point connections between chiplets, limiting their ability to provide sufficient bandwidth and latency for new high performance systems.

An alternative to passive interposers is *active* interposer technology, in which the interposer is instead manufactured from a standard CMOS process (with the addition of die thinning and Through-Silicon Via [TSV] insertion). Active interposers [60] are an emerging combination of 2.5D and 3D integration that balances the simplified design methodology and thermal management of passive 2.5D but leverages standard CMOS processes to integrate active transistor devices into the interposer for faster repeated interconnect and flexible Network-on-Chip (NoC) for better chiplet connectivity [61]. Active interposers have been utilized in several recent studies [61][19][59][55] to provide high-speed repeated links and to move NoC routers onto the interposers, thereby providing more network bandwidth than available on monolithic SoCs. Active interposers have been demonstrated to improve signaling and efficiency over passive interposer [62] [63], and functional samples with active NoC were recently fabricated [55]. Unfortunately, the active interposer can become a significant cost overhead, especially if it utilizes advanced process technologies. To minimize active interposer cost, previous work has limited the amount of active logic area on the interposer to improve yield [19].

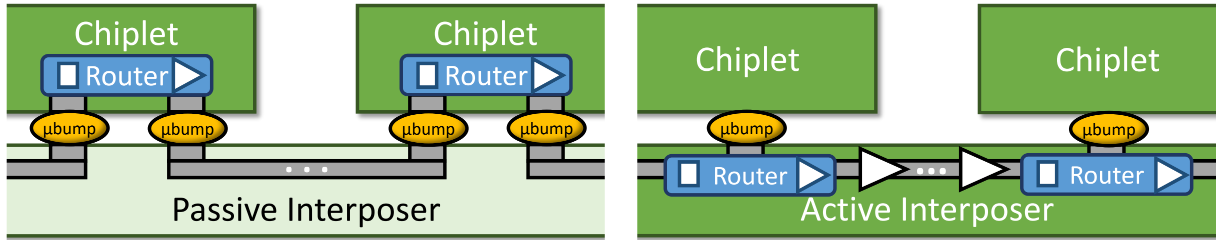


Figure 5.3: NoC integration topology for passive and active interposer.

The transition from a passive to active interposer increases the interposer cost overhead due to additional process complexity, and the active interposer itself could become a large, low-yield die that increases system cost. To date, no active interposers have been adopted in commercial designs due to these cost concerns. As such, all recent active interposer work has focused on "minimally active" interposers [19] [55] with only a small percentage of the available area utilized to minimize yield losses. Some work has gone as far as simplifying the transistors to minimize the number of extra process steps, at the expense of transistor functionality [64]. Yet in all of these minimally active designs, a large and costly active CMOS die is being produced and paid for, but little effective area is being utilized. The following sections details the specific circuit differences for general, scalable Network-on-Chip integration over active and passive interposer substrates.

5.2.1 Active Interposer Integration

To provide high-bandwidth, low-latency, scalable connectivity between modules within and between chiplets, Network-on-Chip interconnection architectures can be deployed that span the chiplets and the interposer. These interconnection networks, composed of router modules and connecting links, pass message packets between the logic module and memory controller nodes on the chiplets. While many different network topologies are possible, the selection of either active or passive interposer technologies influences many

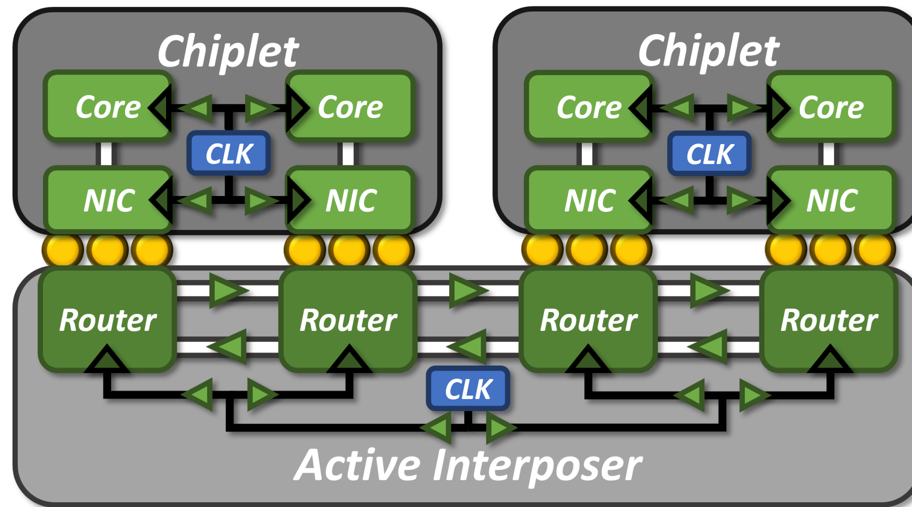


Figure 5.4: Logical diagram of Network-on-Active-Interposer, with on-interposer routers, repeated interconnect, and synchronous interposer clock.

logical and physical aspects of the resulting network.

Active interposers are manufactured to include active transistor devices within the interposer, providing several unique capabilities not possible with passive interposers:

- **Router Location:** Router modules can be moved from the chiplets to the interposer, reducing area in the chiplets and therefore improving chiplet yield and cost. Further, by moving routers into the interposer, links do not need to pass through the chiplet-interposer interface, thus saving microbump resources for the power delivery network.

- **Link Transmission:** Links on active interposers can utilize repeaters to reduce transmission delay, reducing the delay of long routes from a quadratic to linear relation with distance. Reduced link delays allow for faster NoC frequencies and further single-cycle link distances. Further, with routers in the interposer, delay is reduced by avoiding microbump capacitance [65] and the capacitance of Electrostatic Discharge (ESD) protection circuits, which are required to protect the

chiplets interface during bonding [64] (and contribute significantly greater capacitance than a microbump [59]).

- **NoC Clocking:** A primary challenge for SoC-scale synchronous NoCs is the distribution of a low-skew, low-jitter clock to all router modules [66]. Alternative clock schemes, like mesochronous or asynchronous NoC, may incur area or latency overheads and may be difficult to design using standard EDA flows. However, by utilizing the buffers and interconnect available in the active interposer, H-tree clock networks can be generated with sufficiently low jitter (less than $150ps$) to allow for GHz-frequency synchronous NoC in the active interposer [67]. The interposer can even supply a clock signal to the chiplets to avoid interposer-chiplet synchronization latency [68], but this work assumes that a synchronization latency overhead is incurred between the interposer and chiplets to enable flexible DVFS of each core.

This active interposer NoC architecture is depicted in Figure 5.4.

An active interposer can be implemented using existing CMOS technologies. Because global interconnect performance has remained relatively constant throughout recent process technologies, the link delay has low sensitivity to the interposer process selection. However, NoC routers can consume significant area when implemented with many ports and wide flits, especially in older, larger process technologies. Advanced process technologies can implement complex routers with much smaller areas to improve yield, but wafer costs are significantly higher than for older processes. Relative process technology costs are shown in Figure 5.9 and router area is addressed in Section 5.5.1.

5.2.2 Passive Interposer Integration

Unlike active interposers, passive interposers only contain metal interconnect, so they cannot include active logic like routers, repeaters, or FIFO queues in the interposer. This

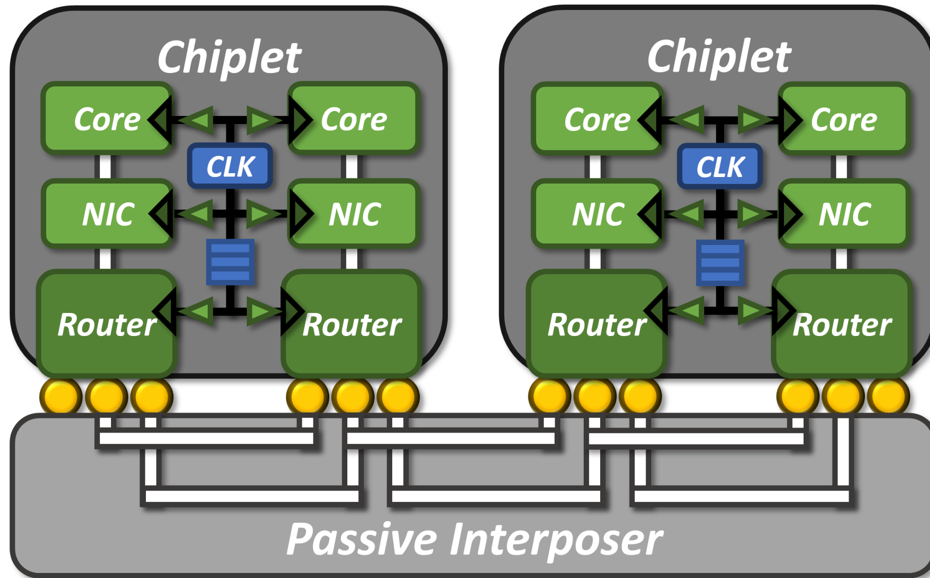


Figure 5.5: Logical diagram of Network-on-Passive-Interposer, with on-chiplet routers and non-repeated interconnect.

leads to several challenges when constructing an NoC across the interposer:

- **Router Location:** Router modules must be placed on the chiplets, contributing to chiplet area and degrading yield and cost. Additionally, links that pass through the chiplet-interposer interface consume microbump resources.

- **Link Transmission:** Links on passive interposers cannot utilize repeaters, so long routes exhibit a quadratic delay with distance, resulting in longer link latencies or slower networks. Links must additionally pass through two chiplet-interposer interfaces, which introduce additional capacitive loads from the microbumps and ESD protection circuits.

- **NoC Clocking:** The challenge of low-skew, low-jitter clock distribution to all routers is complicated by the fact that the passive interposer cannot generate a clock on the interposer and cannot include buffers to drive the low-jitter clock network. To complicate matters further, any clock network that spans and utilizes

the chiplets must contend with the potentially large inter-chiplet process variation and power noise. Accordingly, a synchronous NoC with GHz frequency is not likely to be feasible with a passive interposer. Instead, a globally asynchronous locally synchronous (GALS) paradigm is more likely. In such a scheme, routers on each chiplet are synchronous, but synchronization must occur when transmitting between chiplets, incurring a synchronization delay of several cycles [19]. To allow for flexible DVFS of each core without constraining the NoC frequency, additional synchronization occurs between the routers and Network Interface Controller (NIC) within the same chiplet.

This passive interposer NoC architecture is depicted in Figure 5.5.

5.3 Scaling of Interposer Link Width and Frequency

In the previous section, significant improvements in manufacturability are shown from the chiplet partitioning of large monolithic systems. This technique can be enabled by multiple emerging packaging technologies, but the requirements for high bandwidth, high efficiency, and low latency in performance-targeting systems are difficult to achieve with coarse-featured package-level integration techniques. The fine-featured die-level integration of passive or active interposers, however, is able to concurrently meet these performance goals. Within this interposer design space, circuit-level differences between active and passive interposers determine the feasible NoC architecture designs and resulting performance. In this section, these interposer NoC architectures are analyzed in terms of scalability, area overhead, and link frequency in order to assist designers in the proper interposer technology selection to meet system requirements.

The interconnect-only nature of passive interposers, versus the embedded routers and low latency repeated wires of active interposers, leads to major differences in NoC de-

sign between the two interposer types. For the passive interposer, all routers must be fabricated into the chiplet dies, contributing chiplet area overhead. Each network link is driven from the output channel through the microbumps into the passive interposer, where it travels along a long unbuffered interconnect link before again passing through a microbump to the receiving router input channel. With routers in the chiplets, all inter-chiplet NoC links, in all directions, must pass through these die-die connections, which often include electrostatic discharge (ESD) protection overheads. The active interposer, however, only needs to add a single high bandwidth hop from a chiplet node to an on-interposer router. Within the active interposer, the flit can be passed between routers without the overhead of die-die microbump transmissions. Additionally, repeaters along the links can reduce interconnect transmission delay and increase the achievable network frequency. The increased design flexibility of the active interposers, with reduced constraints on microbump utilization and router placement, presents a wide range of network architecture opportunities to meet performance requirements [19], which for exascale systems may be multiple Tbytes per second of memory bandwidth [47]. The network architecture differences between interposer types are demonstrated in Figure 5.3.

5.3.1 Router and Microbump Technology Scalability

One necessary design consideration for interposer-based NoC is the area scalability of the microbump arrays versus the area of the process technology-dependent routers. Modern microbump technology is standardizing on $40\mu m$ pitch, with potential reduction to 5 μm pitch in the future [69]. At current pitches, a 512-bit link spans an area of at least 0.82 mm^2 (not including any local microbump allocation for power or clocks), and a 256-bit link is half this area at 0.41 mm^2. A 5x5 router for a passive interposer will have 2 unidirectional links internal to the chiplet and 8 through-interposer links for

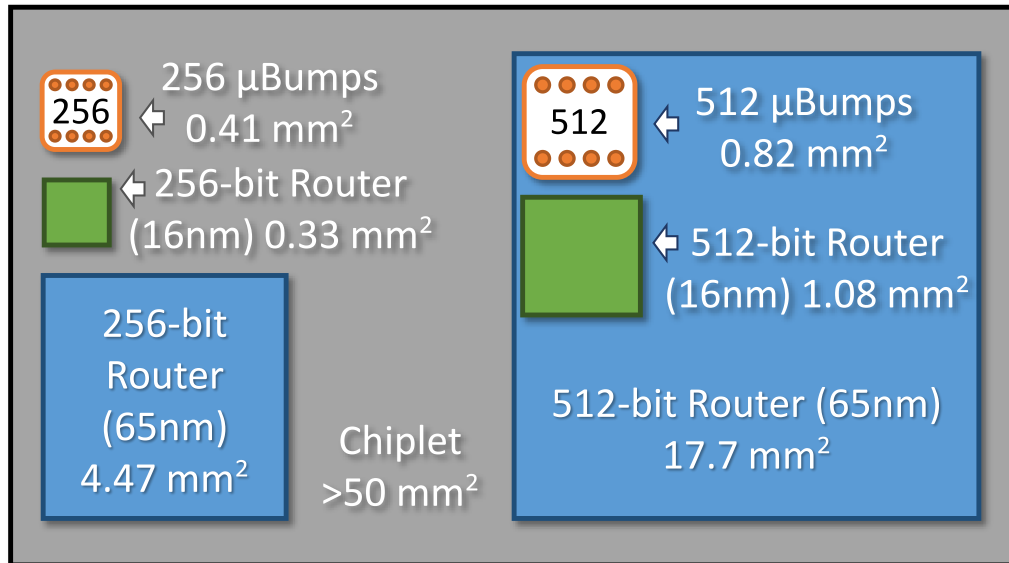


Figure 5.6: Scale comparison of $40\mu m$ pitch microbump arrays to 256-bit and 512-bit flit width routers in 16nm and 65nm technologies.

the 4 cardinal directions, thus requiring 8 microbump arrays of the link width. This is still a reasonably small percentage of the available chiplet bandwidth (only 13% of even a small 50 mm^2 chiplet), but it could limit the number of routers per chiplet. Of more significant concern is the scalability of microbump pitch with router size. Using the *McPAT* modeling framework, the areas for a 5x5 NoC router can be generated for a range of process technologies from 16nm to 65nm and beyond [70]. Figure 5.6 illustrates potential scaling issues for both active and passive interposers. For passive interposers, the area of a router in a modern 16nm process is slightly smaller than the area of a single microbump array of the same width, but because 8 unidirectional links are required between the chiplet and passive interposer, sufficient fanout wiring must be added, further consuming chiplet resources. For an active interposer in an aging technology node like 65 nm, the router area can be an order of magnitude larger than a single microbump array. This facilitates the low-overhead communication between the interposer and chiplet, but it demonstrates limits to the number of routers in the active

interposer when older processes are selected.

5.3.2 Link Frequency in Active and Passive Interposers

The lack of active devices in the passive interposer requires that routers are placed in the chiplet dies and that links must route through the die-die microbumps and across longer unrepeated interconnect.The circuit models for the different interposer types are shown in Figure 5.7 for the passive interposer link, with microbump RC, and for an active interposer link with N repeaters. To achieve high bandwidth and low latency routing, the active interposer has the advantage of lower RC (without die-die connections) and reduced interconnect delay from repeaters. Further, the die-die interconnect needs ESD protection on the bumps to protect the circuit during manufacturing, resulting in additional capacitive load for each passive interposer link. To model the difference in link delay and maximum network frequency, the circuits were simulated using HSPICE using the 65nm PTM models for transistor and interconnect [71]. For each specified link distance, the drivers and repeaters were optimized to minimize link delay. Maximum bitrate results are shown in Figure 5.8 for interconnect settings with 350 nm wire width and spacing [69], 1.2 μm thickness, starting driver width of 2x, and maximum repeater width of 64x. To demonstrate sensitivity, two curves are shown for the active interposer: one with the same capacitive load as the passive interposer with ESD protection overhead (200 fF) and one with a lower load of 50 fF. The microbumps, with self capacitance of only 15 fF [69], introduce limited overhead compared to the lengthy interconnect. The repeaters, however, provide a significant advantage to the active interposer, which is able to achieve several times less delay than the passive interposer for the same link length. For example, The active interposer can thus provide a greater range of NoC performance, with reduced latency links for higher network frequency or longer physical links at the

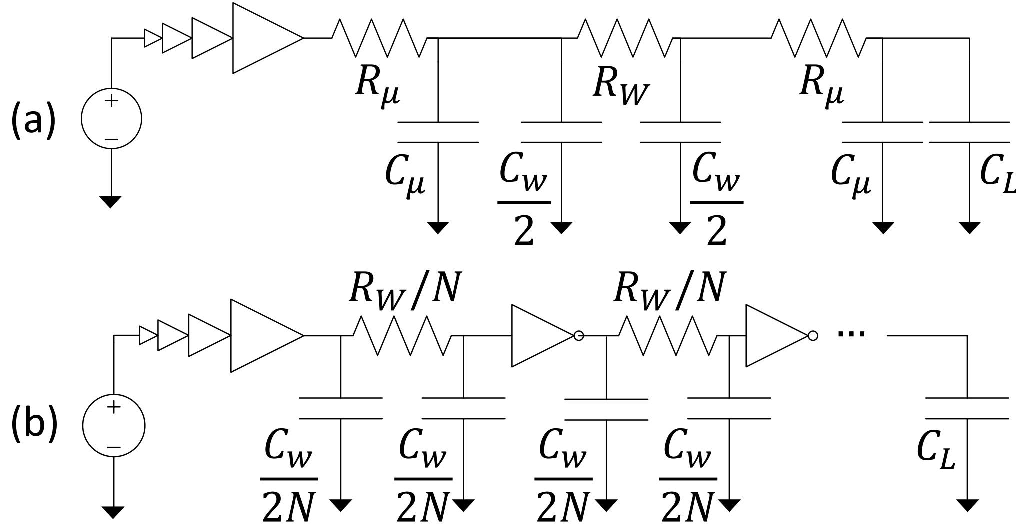


Figure 5.7: NoC circuit differences between (a) passive and (b) active interposers.

same frequency.

As demonstrated with the models developed in this work, interposer-enabled die partitioning can provide a significant improvement to the manufacturability of previously unconsidered systems as small as mainstream desktop processors. For a 32 core server processor, yield loss was reduced by 0.42x and the number of fully enabled systems was increased by 1.98x-3.94x, depending on process maturity. Contrary to prior assumptions, these yield improvements can be cost-effectively realized, while still providing high performance communication, through the use of either active or passive interposers, depending on performance and cost requirements. Active interposers can provide several times lower latency and higher throughput links than passive interposers, but the low wafer cost of passive interposers provides a cost advantage, even after including active interposer fault tolerance methods for improved yield. This work aims to provide system designers with the proper guidelines and tools for determining the best interposer solution to meet system requirements, proliferating the usage of the interposer-chiplet design approach to a broader application range.

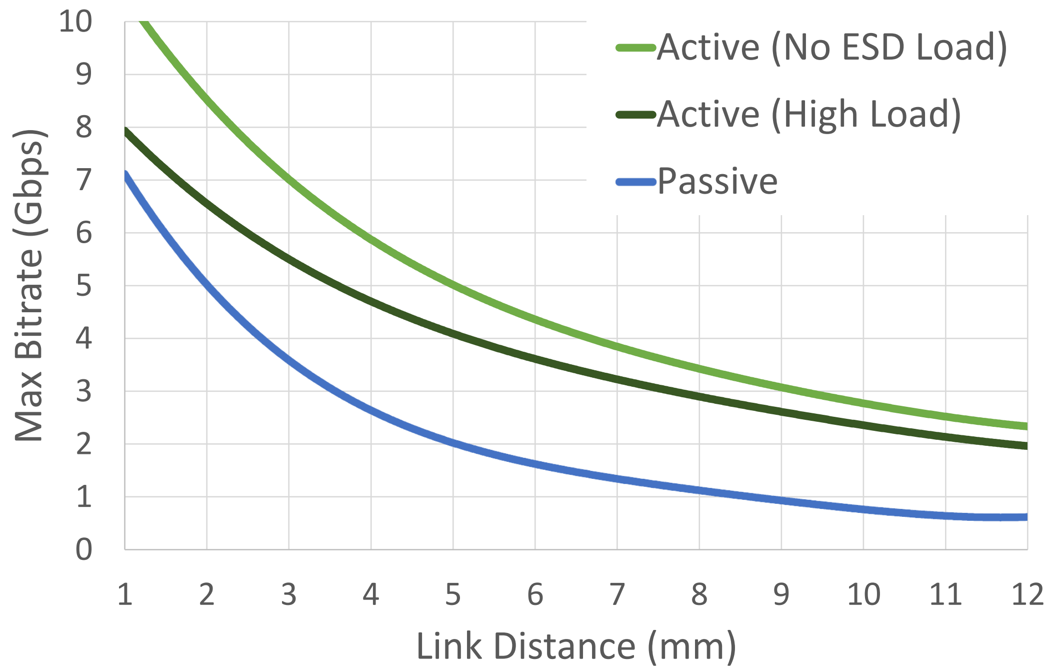


Figure 5.8: Maximum bitrate versus link distance for the passive interposer, active interposer with same load capacitance as passive interposer, and active interposer without ESD load overhead.

5.4 Interposer Cost Comparison

As previously demonstrated, the partitioning of a large monolithic SoC into multiple chiplets can result in significant improvements to yield and functionality. Active and passive interposers are able to provide high bandwidth NoCs for chiplet reintegration to meet a range of performance requirements. Unfortunately, interposer fabrication and chiplet bonding add manufacturing cost overheads that may diminish the total system cost benefits. Additionally, although active interposers demonstrate lower link latency, higher bitrates, and more flexible NoC architectures, the extra process and design complexity versus passive interposers translates to further cost and yield overheads. In this section, the relative magnitudes of these overheads are analyzed versus system cost improvements across a range of interposer technology choices. Results suggest that active interposers are indeed consistently more expensive than passive interposers, but that

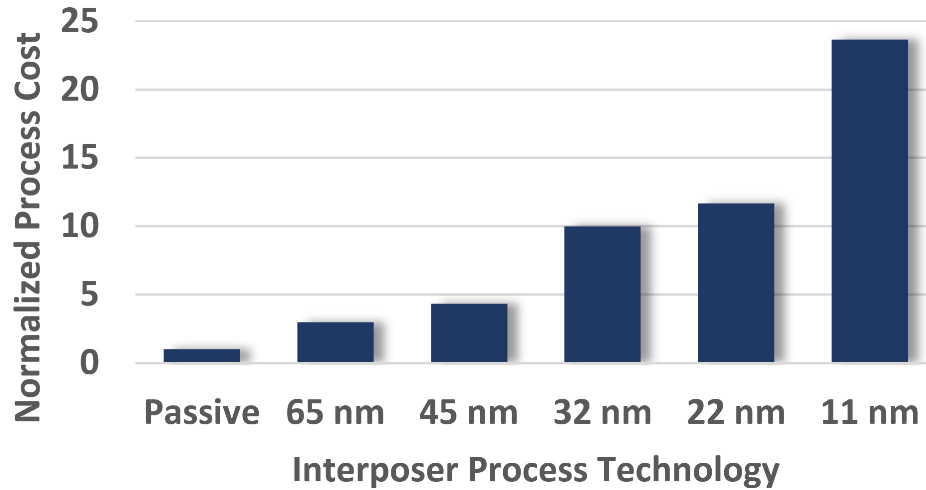


Figure 5.9: Normalized wafer costs for passive and active interposer.

with proper technology selection they are both cost effective integration solutions for high performance systems. Further, based on the presented yield and cost breakdowns, the "prepaid" vacant area of the interposer is leveraged for fault tolerance to reduce the active interposer cost overhead. In order to meet system requirements, system designers can leverage the analysis and techniques in this section to balance the cost versus performance trade-offs between active and passive interposers.

5.4.1 Interposer Yield and Cost

Employing an interposer introduces cost overheads, including the manufacturing of an additional large silicon die. To minimize this overhead, passive interposers only include interconnect layers, and no active transistors, and therefore have high yields and much lower wafer costs. Active interposers, reliant on relatively expensive transistor fabrication, may need to limit the amount of active area devoted to routers and other logic in order to minimize yield loss. Active interposer process technology selection determines the wafer cost (advanced processes are much more expensive for equal area) and the active

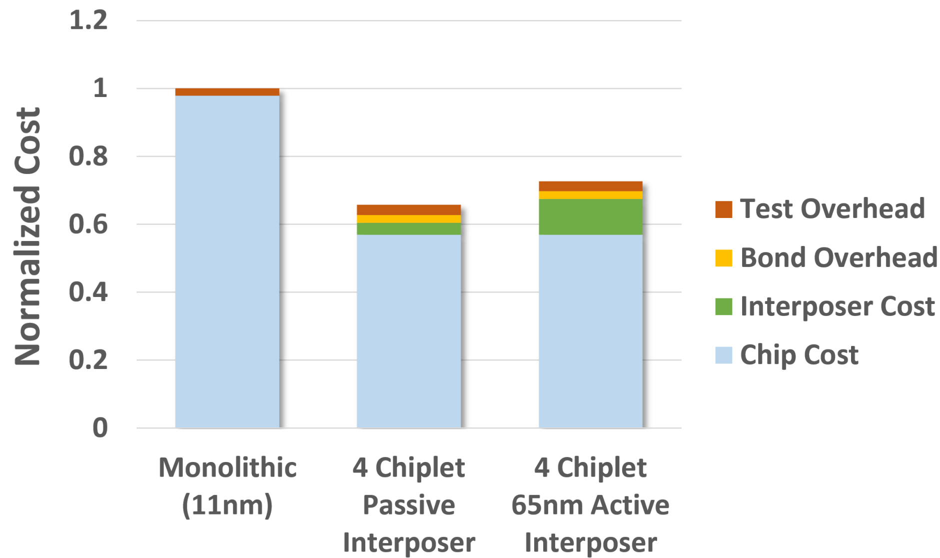


Figure 5.10: Manufacturing cost example for $336mm^2$ chip in $11nm$ process, with equivalent four-chiplet systems on $448mm^2$ passive and $65nm$ active interposers.

area (advanced processes can implement routers in less area, thus improving yield given the same defect density). Figure 5.9 demonstrates the relative process costs for different possible interposer technologies, based on TSV integration and process cost data from *IC Knowledge* [9]. An additional integration cost overhead is the cost of KGD validation of each chiplet before bonding. Although validation time (and therefore cost) primarily scale with design size and complexity, constant per die overheads result in greater validation costs for multiple small dies than for a single monolithic die [9]. Despite these overheads, prior work has demonstrated that both passive and active interposers can indeed be more cost effective than monolithic systems at standard design sizes [57][59]. An example is shown in Figure 5.10 for a $336mm^2$ monolithic system, compared against four-chiplet systems on high-yield $448mm^2$ passive and $65nm$ active interposers (assuming a conservative $1mm$ spacing between and around the chiplets), with $D_0 = 2000/m^2$, $\alpha = 3$, and 99% bond yield.

Unlike most silicon circuits, a passive interposer is primarily metal interconnect, sur-

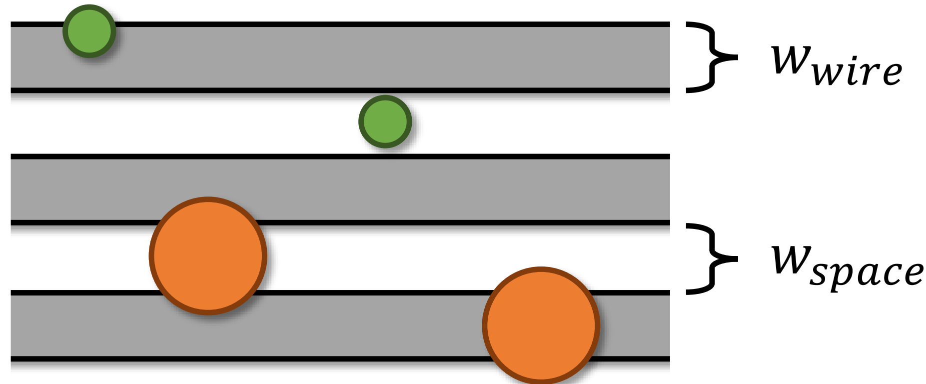


Figure 5.11: Critical defects (orange) cause shorts and cuts in the interconnect, while smaller defects (green) are non-critical.

rounded by vacant space. An active interposer is similar in design, but may also have sparse logic activity for routers and repeaters. The prior assumptions for Equation 2.6 for critical area and defect density are inaccurate for interconnect yield, since a wider route is instead more resilient to a small defects that would disrupt minimally sized features. As shown in Figure 5.11, failures occur as shorts between wires (in the same or adjacent layers) or as open cuts [72]. Large wires and spacings require larger sized defects to cause a failure, and historically densities for larger defect sizes drop quickly compared to the critical feature size [73]. Maximum density minimally sized wires will have lower yield, while wide or sparse interconnect improves yield. Based on this yield model, this increased resiliency is modeled by reducing the defect density from $D_0 = 0.2\ cm^{-2}$ to $0.05\ cm^{-2}$ for interconnect area of the specified dimensions.

5.4.2 Active and Passive Interposer Cost

The interposer-enabled die-level integration introduces cost overheads into the manufacturing of each system, but the exact amount depends on multiple design decisions, including the interposer process technology (including active or passive), the number of

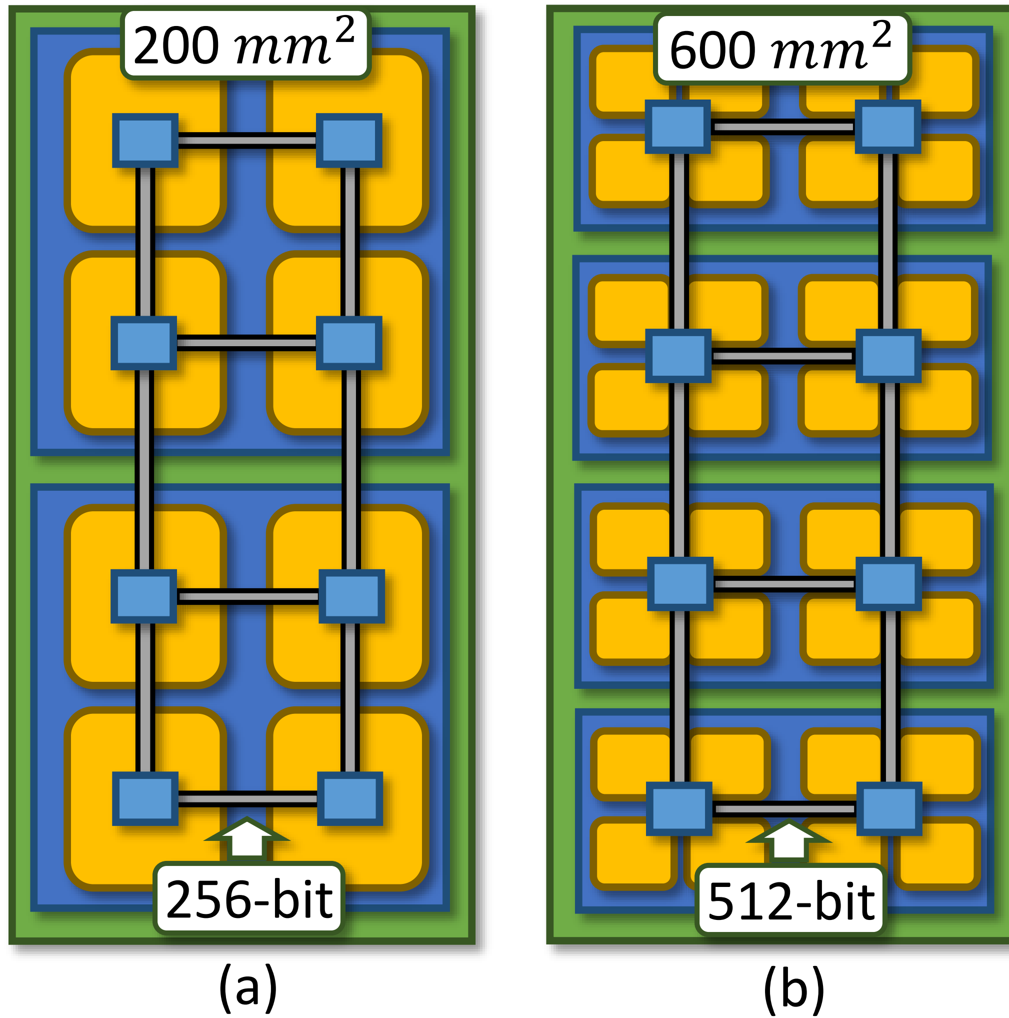


Figure 5.12: Chiplet partitions and 2x4 NoC meshes for (a) 8 core and (b) 32 core systems.

chiplets that must be bonded, and the complexity of the interposer interconnect system that will influence yield. The systems shown in Figure 5.12 are studied: an 8 core processor partitioned into 2 chiplets and a 32 core system partitioned for yield improvement into 4 chiplets. To provide low latency and high bandwidth, a NoC with link width of 256 bits with one router per core is used for the 8 core circuit, and a link width of 512 bits with 8 routers (two per chiplet, with local networks within each 4 core cluster), both in a 2x4 mesh.

Using the prior yield models, with publicly licensable industry wafer and TSV costs from *IC Knowledge, LLC* [9], this section presents the total system costs for a selection of interposer process technologies, with results shown in Figures 5.13 and 5.14. The interposer is assumed to add a 10% area overhead for space around the chiplets. For each interposer, the cost overhead include the base interposer silicon (as if it had ideal yield), the losses from router and interconnect yields, the bonding cost overhead from process complexity, and pessimistic bond yield of 99%. Additionally, the passive interposer includes a cost overhead for the NoC router area that must be added to the chiplets. The total cost of the chiplets is included for each interposer process, and the resulting systems are compared against the cost to manufacture a monolithic chip. The yields for the chiplets and monolithic die use the core-binning methodology presented earlier, with the yield defined as the percentage of dies that produce *some* level of functional binning. Thus these results represent that average manufacturing cost per functional system, but do not reflect the improvements to core count and speed binning that come from chiplet partitioning and matching.

As visible from the results, the interposer price is generally dominated not by the yield of the interconnect and routers, but by the base fabrication cost of the silicon. The most recent process nodes at 28nm and 16nm demonstrate increasing price per area, and although router area scales and yield improves, the area of the interposer is constrained by the chiplets and does not shrink. With these recent processes, the base interposer cost outweighs the yield improvements from chiplet partitioning, especially for the smaller 8 core system. However, the passive and active interposers at older processes can be cost effective even for mainstream systems, and they demonstrate significant reductions for the large area 32 core system, even before core and parametric binning improvements are considered. The passive interposer is consistently lower cost than the active interposer, but a fully active interposer at a mature process still demonstrates cost effectiveness

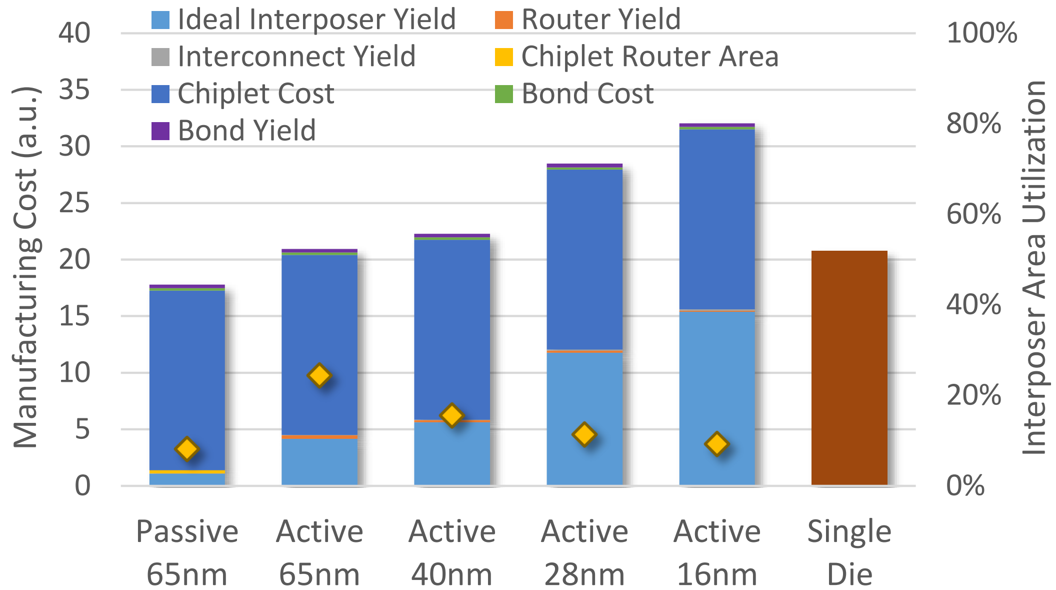


Figure 5.13: Manufacturing cost breakdown of 8 core 2-chiplet interposers systems versus 16nm monolithic die. Interposer utilization on secondary axis.

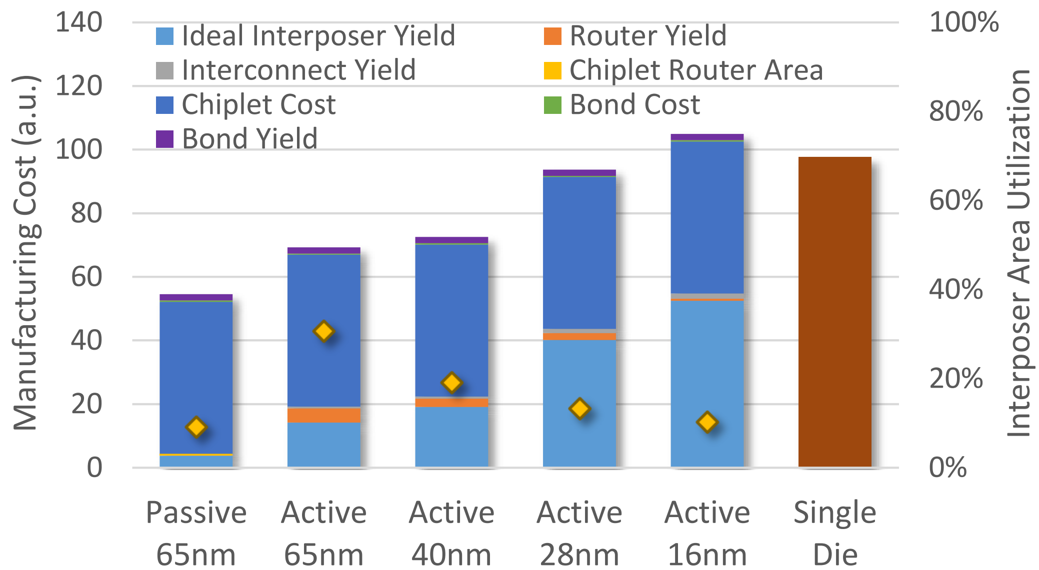


Figure 5.14: Manufacturing cost breakdown of 32 core 4-chiplet interposers systems versus 16nm monolithic die. Interposer utilization on secondary axis.

while supporting improved network performance and design flexibility. System designers can thus select the proper interposer solution for the project-specific performance and cost needs.

5.4.3 Interposer Yield Increase through Fault Tolerance

Although the active interposer suffers from increased wafer cost and has vulnerable critical logic area, it is possible to reduce the vacant "prepaid" area and active devices can be used to improve active interposer functionality. To improve the yield of the active interposer, the free area can be used for fault tolerance methods that virtually remove yield losses. With these techniques, active interposers are no longer constrained to "minimally-active" designs, allowing for cost-effective high performance NoC integration. These techniques are also applicable to passive interposers, but the area overhead is added to the routers in the chiplets, leading to a cost versus yield trade-off.

Fault Tolerant Routers

As shown in the prior sections, the active router area can be the largest yield concern for active interposers with high bandwidth networks. A range of existing literature has been previously proposed to add fault tolerance to NoC routers. In particular, the router design in [74] is a promising candidate that employs low overhead fault tolerance strategies to the routing computation, virtual channel allocation, switch allocation, and crossbar. The fault tolerance mechanisms add only 27% to the router area while allowing for functional behavior until a mean of 21 defects per router, virtually eliminating router failures in the active interposer with no die size increase. The addition of router fault tolerance is especially important for active interposers in aging process nodes like 65nm, in which the router area may consume a significant percentage of the active interposer. For example, the cost overhead of the 32 core active interposer in 65 nm is reduced

by 23% through router fault tolerance, with costs just 4% higher than if the interposer achieved ideal yield.

Redundant Interconnect

Applying fault tolerance to the active interposer routers greatly improves interposer yield, but the interconnect can still be a point of failure that reduces yield and thus increases interposer cost. Because interconnect yield is relatively high for wide routes, adding only a small number of redundant wires to each link in the NoC is sufficient for achieving close-to-ideal yield. Based on Equation 2.13 and the 32 core case study with 512-bit link width, interconnect yield is 97% before redundancy. By adding 2 redundant routes per bus, any 1 short or 2 cuts can be avoided, improving interconnect yield to ¿99.9%. If the routers already include fault detection mechanisms, the redundant interconnect overhead can be included with little additional overhead.

5.4.4 Other Benefits of Interposer Integration

The analysis in the previous section is meant to demonstrate the benefits to manufacturing and finances that can be realized through interposer-enabled chiplet integration, with moderate benefits for mainstream processors and significant improvements for large server processors. However, interposer-based integration also enables other significant design options. Benefits include 1) the on-die integration High bandwidth DRAM memory stacks or emerging resistive nonvolatile memories, 2) heterogeneous processes for analog/RF, high speed SerDes, etc., and 3) chiplet enabled IP reuse. The potential performance, efficiency, and cost benefits of these technologies is beyond the scope of this work, but the results from the previous sections suggest that the overheads of interposer integration are cost effective when combined with die partitioning, enabling these advanced design options free of charge with better performance and efficiency than

low-bandwidth integration methods like Package on Package (PoP). Alternatively, a high performance design that requires an interposer for memory integration should also explore the manufacturing benefits of die partitioning using the methodology described here.

5.5 Performance of Network-on-Interposer

The prior analysis has demonstrated the relative cost and frequency differences of passive and active interposers [59], but has not conclusively demonstrated the architecture-level impact of this interposer technology selection: should the interposer for a given system be passive or active, and if active which process node should be utilized? This section explores the complex interactions between the interposer technology, the Network-on-Chip topology, and the physical link and router implementation, which then determine the interposer active area and yield, the chiplet area and yield and cost, the network link frequency, and ultimately the system cost and the NoC bandwidth and latency. While it is not possible to make conclusions for all possible multi-die systems, traffic patterns, and performance requirements, this work aims for generalizable conclusions by sweeping a range of bisection bandwidth targets, representing different memory and coherence requirements, for a realistic multi-core system with integrated die-stacked memories and a synchronous NoC that is compatible with standard design automation flows. Further, NoC latency trends are analyzed to study the interaction between interposer technology and network topology. This work concludes that when only considering network bandwidth, passive interposer integration is almost always cost-optimal when compared to active interposers, even when considering the router area overhead impact on chiplet yield. However, due to longer link latencies and clock-crossing overheads in passive interposers, active interposers demonstrate a significant advantage for latency-sensitive

systems. Additionally, active interposers in mature process technologies can achieve high bandwidth and lower latency at system costs within 30% of passive interposer systems.

Interposer-based packaging is becoming a widespread methodology for tightly integrating multiple heterogeneous dies into a single package, with the potential to improve manufacturing yield and build larger-than-reticle-sized systems. However, interposer integration also introduces possible communication bottlenecks and cost overheads that can outweigh these benefits. To avoid these drawbacks, the abundant interposer interconnect can be leveraged as network-on-chip interconnection fabric to provide high-bandwidth, low-latency communication between chiplets and memory stacks. This section investigates this new interposer design space of passive and active interposer technologies, network-on-chip topologies, and clocking schemes to determine the cost-optimal interposer architectures for a range of performance requirements.

5.5.1 Interposer Comparison Methodology

To compare the cost and performance trade-offs of active and passive interposers, this work studies a base system floorplan based on recent commercial systems [52]. Figure 5.15 shows the base system with thirty-two generic cores implemented in $11nm$ process technology. The system also contains four stacks of in-package memory, similar to the JEDEC High Bandwidth Memory standard [75], with eight memory channels per stack. Sixteen memory controllers are distributed between the cores and placed on the chiplet periphery, with two memory channels managed by each memory controller. The entire system is integrated using either an active interposer or a passive interposer, which provides connectivity between nodes to route all inter-core message traffic. NIC terminals are located at each core and memory controller. Figure 5.15 demonstrates a configuration with four chiplets, but systems from one to eight chiplets are considered. The network is clocked

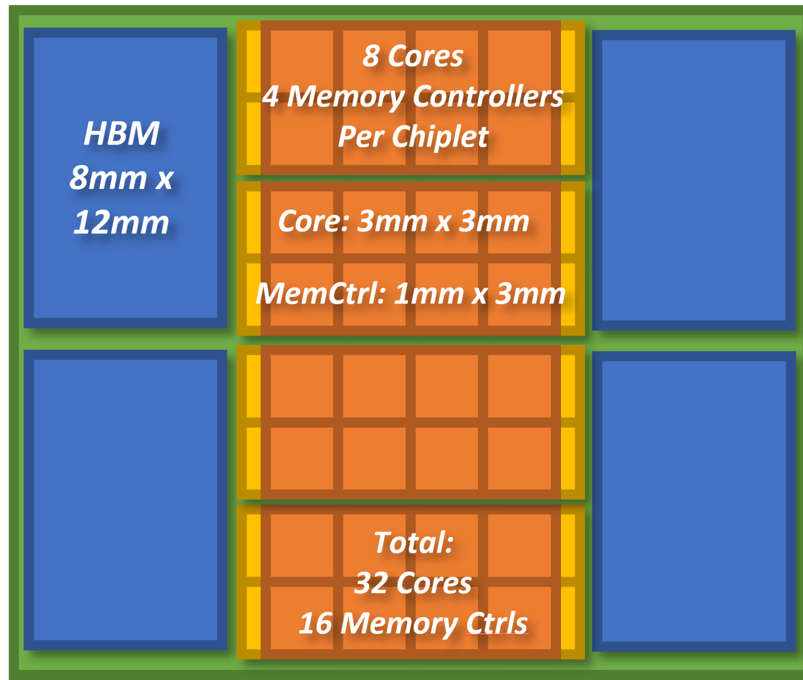


Figure 5.15: Base system under evaluation, with thirty-two $3mm \times 3mm$ cores, sixteen $1mm \times 3mm$ memory controllers, and four $8mm \times 12mm$ stacks of High Bandwidth Memory, all integrated with an interposer. The system is demonstrated with four eight-core chiplets in this example.

at 2 GHz.

The network topologies visualized in Figure 6.6 and detailed in Table 5.1 are assigned to router nodes connected to these terminals. Misaligned topology alternatives are also included [19].

The selection of active or passive interposer technology determines the relations between network performance and system cost:

- **Router Area:** With an active interposer, the routers are placed in the interposer, increasing the interposer active area used in yield calculation but reducing the chiplet area. With a passive interposer, all routers contribute to the chiplet area and are implemented in the $11nm$ process technology.

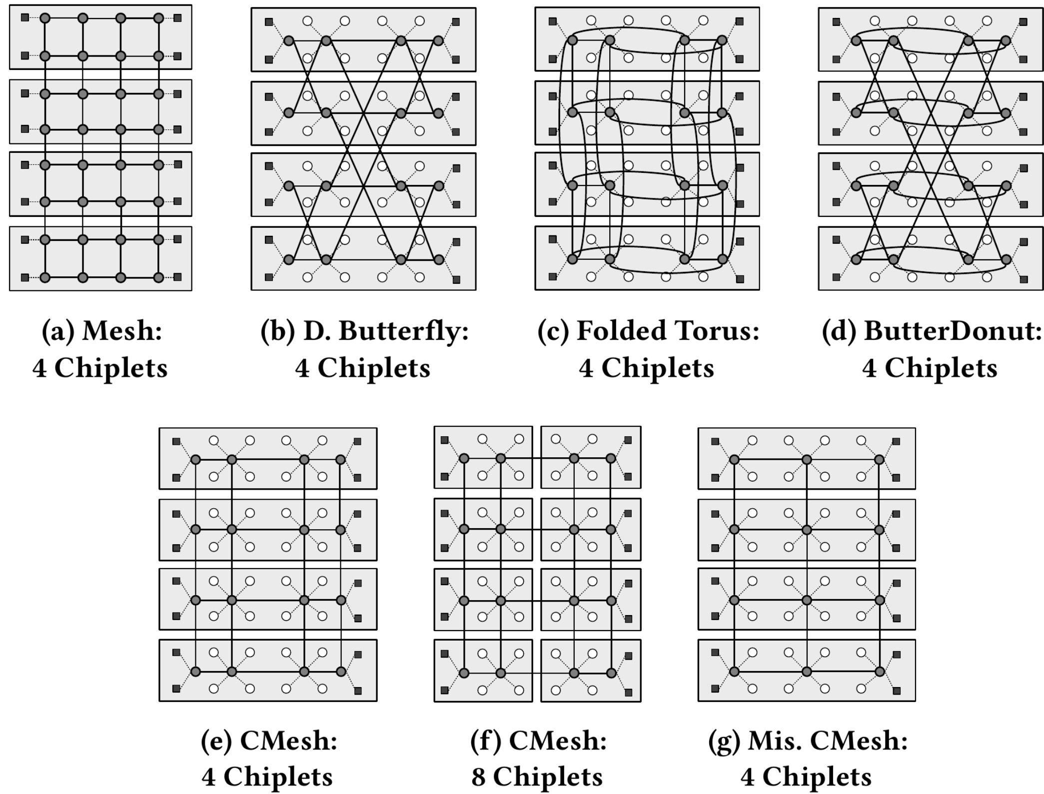


Figure 5.16: Network-on-Interposer four-chiplet topologies. Eight-chiplet and misaligned examples demonstrated for CMesh

Table 5.1: Network-on-Interposer Topologies

Topology	Nodes	Links	Diameter	Avg. Hops	Bisection Links
Mesh	32 (4×8)	52	10	4.9	8
Concentrated Mesh	16 (4×4)	18	6	3.5	5
Double Butterfly	16 (4×4)	24	3	3.1	8
Folded Torus	16 (4×4)	32	4	3.0	8
ButterDonut	16 (4×4)	28	3	3.0	12
Misaligned Concentrated Mesh	12 (3×4)	17	5	3.1	4
Misaligned Double Butterfly	12 (3×4)	16	2	3.1	8
Misaligned Folded Torus	12 (3×4)	24	3	2.7	8
Misaligned ButterDonut	12 (3×4)	20	2	2.8	12

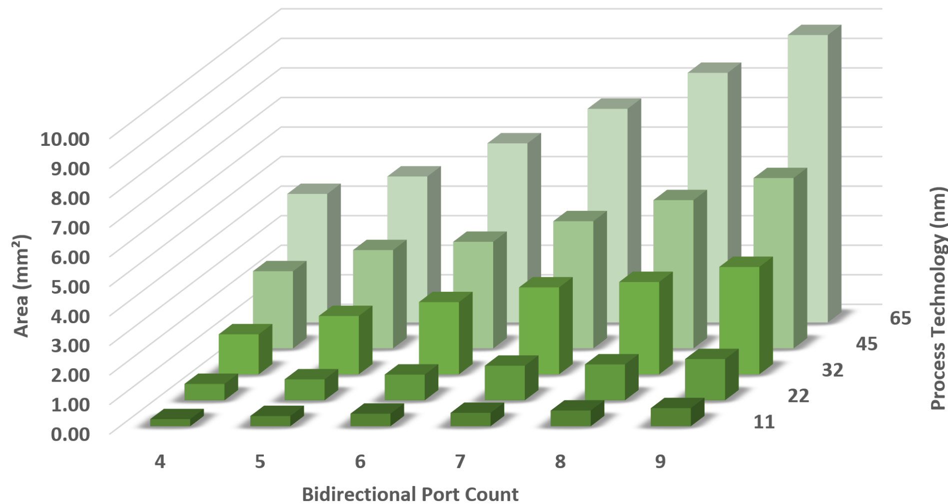


Figure 5.17: Router area with 512-bit flit width versus bidirectional port count and process technology.

- **Link Frequency:** Links on the active interposer are repeated with optimally-sized repeaters in the selected process technology to minimize link delay. For passive interposers, link delay is calculated for the non-repeated route with capacitance overhead for microbumps ($15fF$) and ESD protection ($200fF$) [59]. Delays are computed from SPICE using pi-model interconnect segments. At the 2 GHz network frequency, the active interposer is able to transmit most link distances in a single cycle. Only the longest links (the longest vertical links in the Folded Torus and the longest diagonals in the Double Butterfly and ButterDonut) require two cycles on the active interposer. The passive interposer, however, requires at least two cycles for all but the shortest links (as in the mesh or concentrating routers), and the longest diagonal links take eight cycles. Link latencies for each distance in the topologies are given in Table 5.2. Multi-cycle links, depending on EDA methodology, may require flip-flops in the active interposer or chiplets [57] to meet synchronous requirements.

- Due to clock domain division between each chiplet terminal and the interposer

Distance (mm)	Active (cycles)	Passive (cycles)
3.5	1	1
6.5	1	2
10	1	3
13	2	4
19.5	2	8

Table 5.2: Link latencies for active and passive interposers

routers, which enables core DVFS, the active interposer incurs a synchronization overhead only on the initial and final transitions from chiplet to interposer. The passive interposer must additionally synchronize when transmitting between chiplets. This work assumes a three-cycle synchronization latency at each clock crossing.

To compare each combination of interposer technology and network topology, it is necessary to first estimate the area of network routers and links and to then compute the resulting yields and costs. Router areas are estimated using the DSENT Network-on-Chip modeling tool [76], with process technologies from 11nm to 65nm. Each three-staged pipelined router has 16 virtual channels and eight buffers per virtual channel. Routers area is given for port count and process technology with 512-bit flit width in Figure 5.17. Flit width is also utilized as a parameter. Each link is composed of global-level interposer interconnect with wire width and spacing of 350nm. Based on the system floorplan, the router and link areas are calculated for each topology. Figure 5.18 shows the average and range of router utilization across topologies for each active interposer technology and flit width. Figure 5.19 shows the average and range of interposer interconnect utilization under the chiplets for flit widths from 32 to 1024 bits, assuming the availability of two X-Y routing layer pairs. Assuming a microbump pitch of 20μm, each chiplet has sufficient microbump resources to connect all topologies with at least 1024 bit flit width, even when allocating half of microbump resources to the power delivery network.

After calculating the areas of each interposer's network components, the cost of the

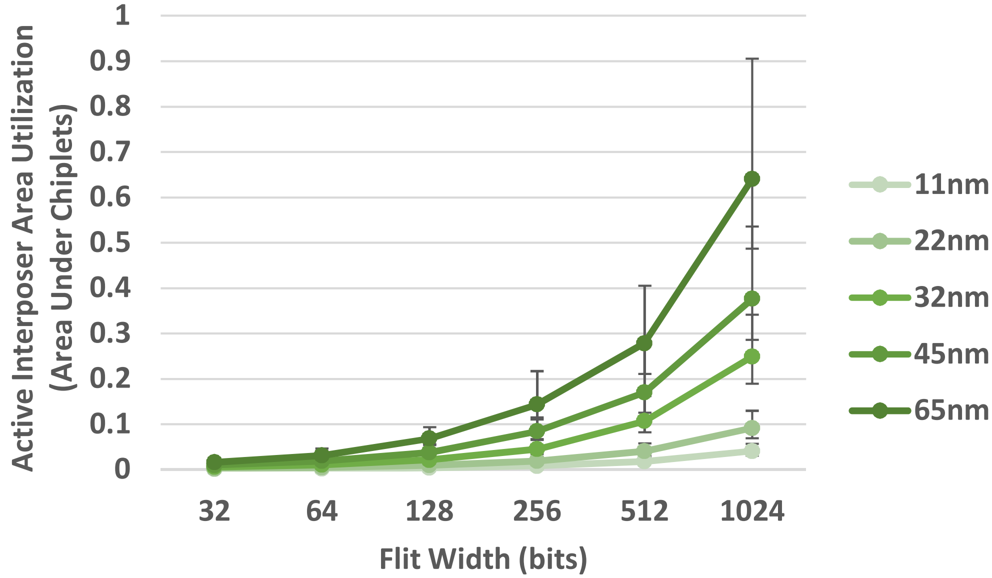


Figure 5.18: Average router utilization percentage, across network topologies, of active interposer area under the chiplets. Error bars indicate the range between topologies.

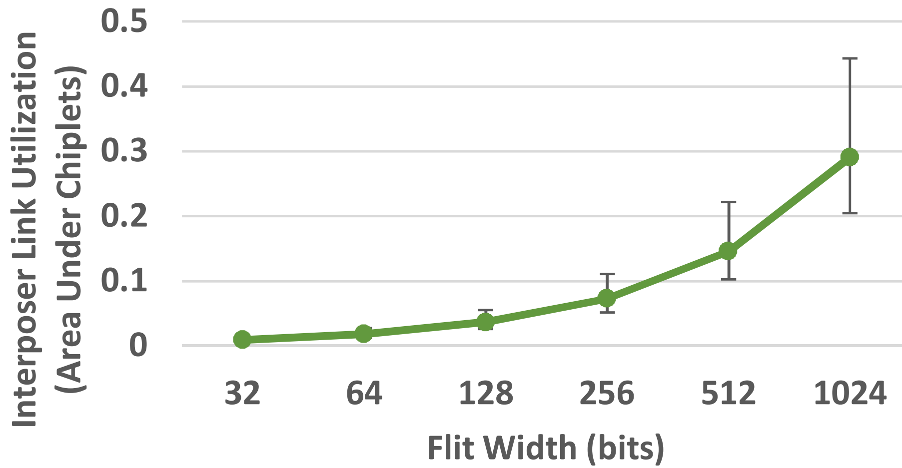


Figure 5.19: Average link utilization percentage, across network topologies, of interposer interconnect under the chiplets. Error bars indicate the range between topologies.

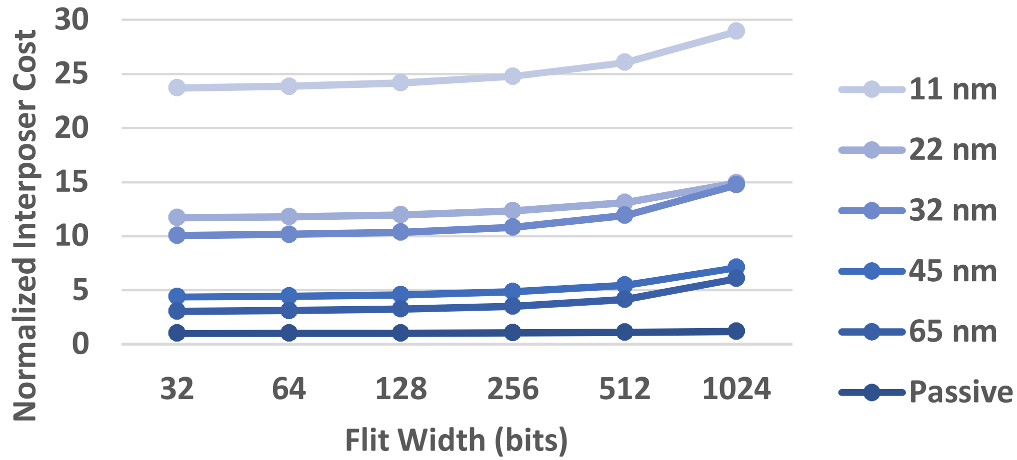


Figure 5.20: Normalized interposer cost of ButterDonut topology, for each process technology, versus flit width.

interposer and system can be computed by estimating the interposer yield. The earlier cost and yield estimation methodology is utilized to compute the interposer and chiplet active area yield, with defect density $D_0^{Logic} = 2000/m^2$, as well as the global interconnect yield with lower defect density [77] $D_0^{Link} = 500/m^2$. The manufacturing cost of each interposer and chiplet is then determined by calculating the number of dies per $300mm$ wafer and incorporating the process-dependent wafer cost, shown in Figure 5.9, and the yield loss from active area and link defects. Each chiplet is bonded to the interposer with a bond yield of 99% [18]. For all cost results, only interposer and chiplet costs (and not the memory stack costs) are included, but the interposer footprint includes the area required for the memory stacks.

The resulting normalized interposer costs for each process technology are demonstrated in Figure 5.20 for the ButterDonut topology. When only considering the interposer cost, the passive interposer is significantly less costly than any of the active interposer configurations due to a lower process cost and a lack of active area. For the active interposer technology options, the mature $65nm$ process is consistently the lowest

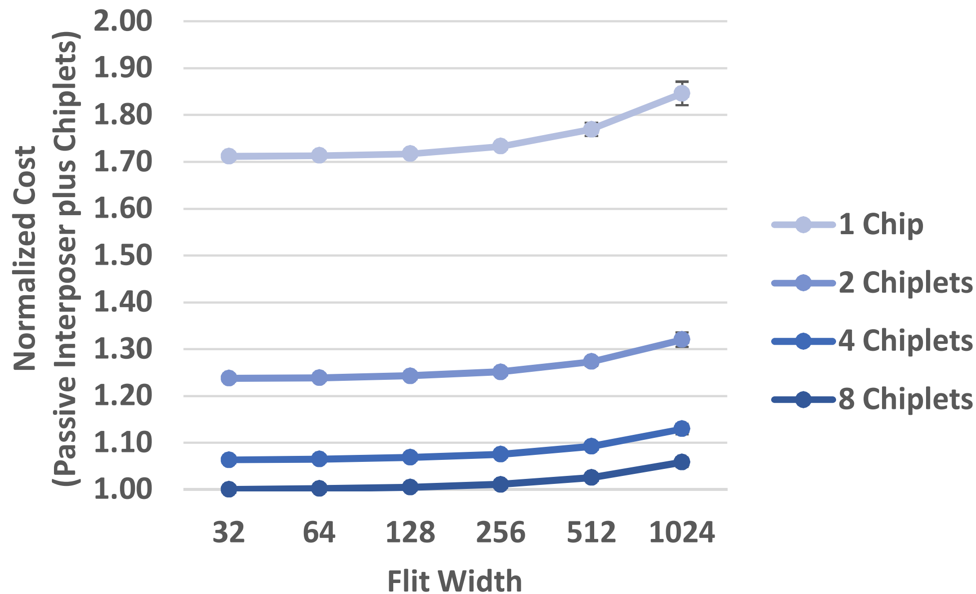


Figure 5.21: Normalized cost of the passive interposer and chiplets versus flit width, for each chiplet size, averaged over network topologies. Error bars indicate maximum and minimum variability across topologies.

cost, even despite the large active area utilization at the largest flit width of 1024 bits.

The full system costs for the ButterDonut topology with 512-bit flit width are shown in Figure 5.22. All active interposer systems have eight chiplets but vary by interposer technology. The advanced process technologies incur significantly higher interposer overheads due to higher wafer costs, while the router yield is less significant even for the 65nm process with high utilization. The passive interposer configurations are shown with a range of chiplet sizes, demonstrating the improvement in chiplet yield with increased chiplet count, as well as an increase in bond loss overhead.

5.5.2 Bisection Bandwidth of Network-on-Interposer

To compare the network performance results of a given interposer technology and network topology, the bisection bandwidth metric can be selected to measure the topology-

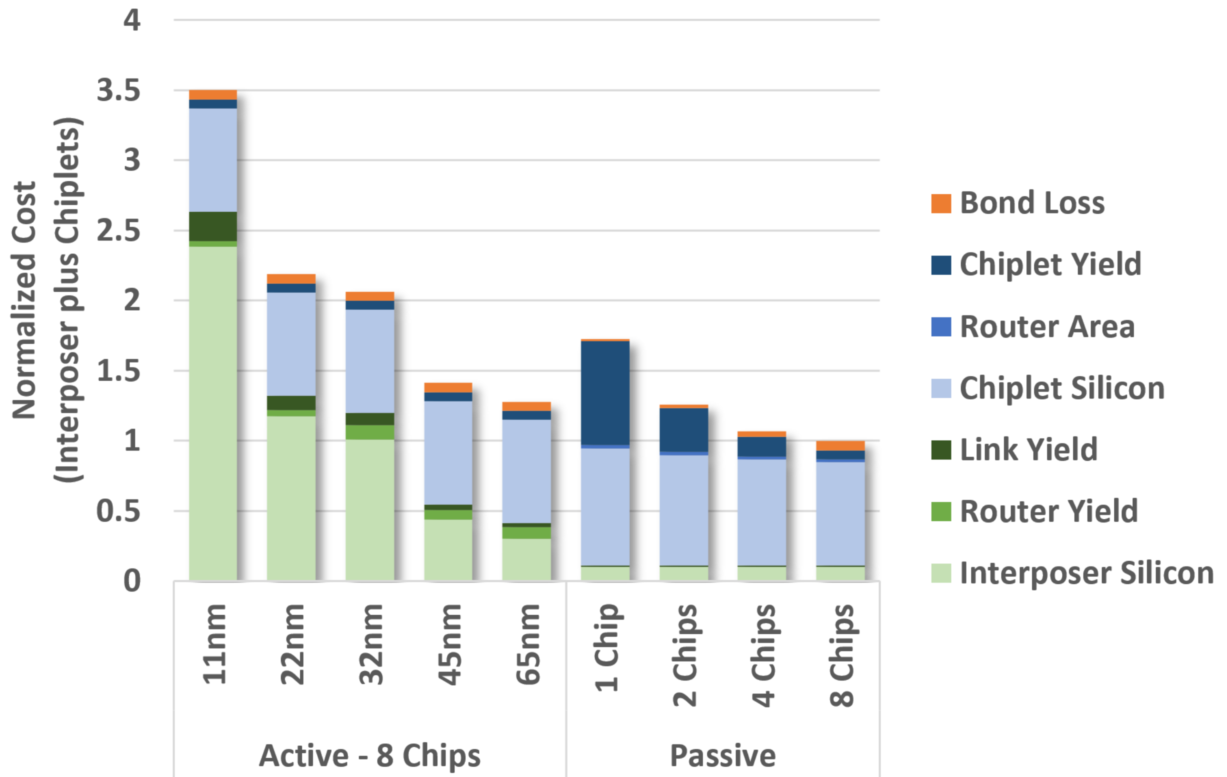


Figure 5.22: Normalized cost profile of interposer and chiplets across interposer technology options for the 512-bit flit width ButterDonut topology, with eight chiplets for active interposer configurations. Active interposer process cost dominates in advanced processes. Router yield (active) and router area (passive) are small contributions to cost.

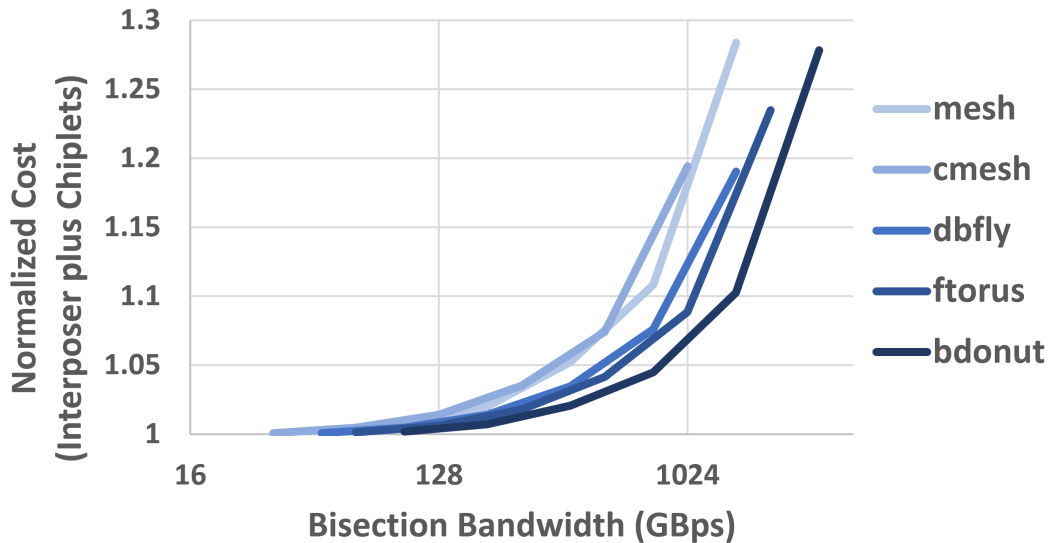


Figure 5.23: Normalized cost of interposer and chiplets versus bisection bandwidth for each topology on a $65nm$ active interposer with eight chiplets.

level peak communication between cores and memory on opposite sides of the system. The bisection bandwidth for a topology be calculated with the number of bisection links, given in Table 5.1, the flit width, and the network frequency of $2GHz$. The bisection bandwidths can then be compared against the system costs to determine the cost-optimal topology.

Figure 5.23 shows the normalized cost versus bisection bandwidth for each active interposer topology at each flit width. The bandwidth value is the average of the bisection bandwidths in the x and y directions. Each active interposer configuration utilizes the $65nm$ process and has eight chiplets. As demonstrated in the figure, the high-connectivity ButterDonut topology is consistently cost-optimal when only considering cost and bisection bandwidth, despite greater router and link area.

Figure 5.24 compares total system cost and bisection bandwidth for each interposer technology option, with eight chiplets for each active interposer system. As bisection bandwidth is sensitive only to topology, flit width, and frequency, the passive interposer achieves the same bandwidth, but at significantly lower cost. As demonstrated earlier in

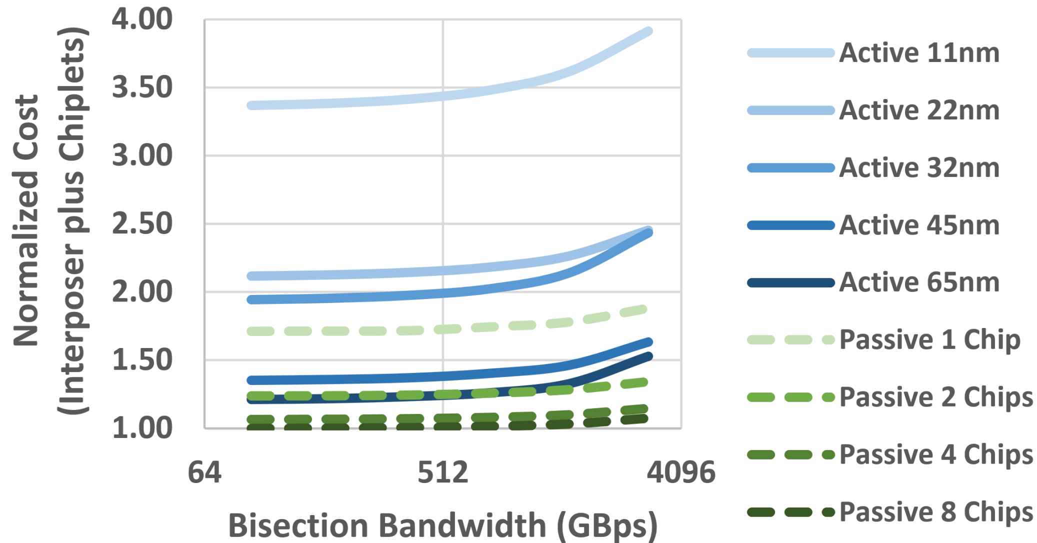


Figure 5.24: Normalized cost of interposer and chiplets versus bisection bandwidth across interposer technology options, with eight chiplets for all active interposer configurations.

Figure 5.22, the passive interposer has high yield and pays little overhead for including the router area in the chiplets. Overall, investigations of network bisection bandwidth demonstrate a system affinity for small chiplet size, to increase yield, and high-radix networks, to improve bandwidth. Router size and link width have less impact on interposer and system cost than the initial selection of interposer technology, and mature processes with lower cost are advantageous for active interposers even at high utilization and lower yield. However, bisection bandwidth is only one of multiple network metrics that should be considered.

5.5.3 Latency Evaluation of Network-on-Interposer

Although passive interposers are cost-optimal when only considering bandwidth, latency is also a critical metric for many networks-on-interposer systems. In this section, the topologies listed in Table 5.1 are mapped to passive and active interposers and compared to determine the impact on network latency.

Chiplets on passive interposers, as explained in Section 5.2.2, are clocked independently. Therefore, a three-cycle clock synchronization overhead is added to the latency of inter-chiplet links when using passive interposers. Routers within an active interposer, as explained in Section 5.2.1, are synchronous and therefore do not have this link synchronization overhead. However, the connections between the NIC terminals and routers does require a synchronization overhead for both the passive and active interposers in order to allow for independent core DVFS. Link latencies between routers for passive and active interposer implementations also depend on the distance, as previously stated in Table 5.2.

Methodology: Booksim [78] is used to evaluate the performance of network-on-interposer topologies listed in Table 5.1. For the following evaluations, the bisection bandwidth of the topologies is fixed by balancing the bisection link count with link width. The network frequency is 2 GHz, as explained in Section 5.5.1. The network is evaluated on uniform random synthetic traffic, sweeping over injection rate to observe saturation throughput as well as latency.

Active vs. Passive Interposer on Latency: Figure 5.25 shows average packet latency for networks on passive and active interposer. Two main observations can be made from this study. First, the active interposer realizes lower average latency than the passive interposer. This is mainly due to lower active interposer link latencies over the synchronous network. Second, topology decisions also impact average latency as a lower average hop count associates with lower network latency, as expected.

Aligned vs. Misaligned Topologies on Latency: Figure 5.26 shows average packet latency for aligned and misaligned topologies for active interposers. Misaligned topologies have fewer routers shared between the cores and memory controller nodes on chiplets. Therefore, misaligned topologies result in lower average latencies compared to the aligned topologies shown in Figure 5.26 due to lower average hops for both passive and

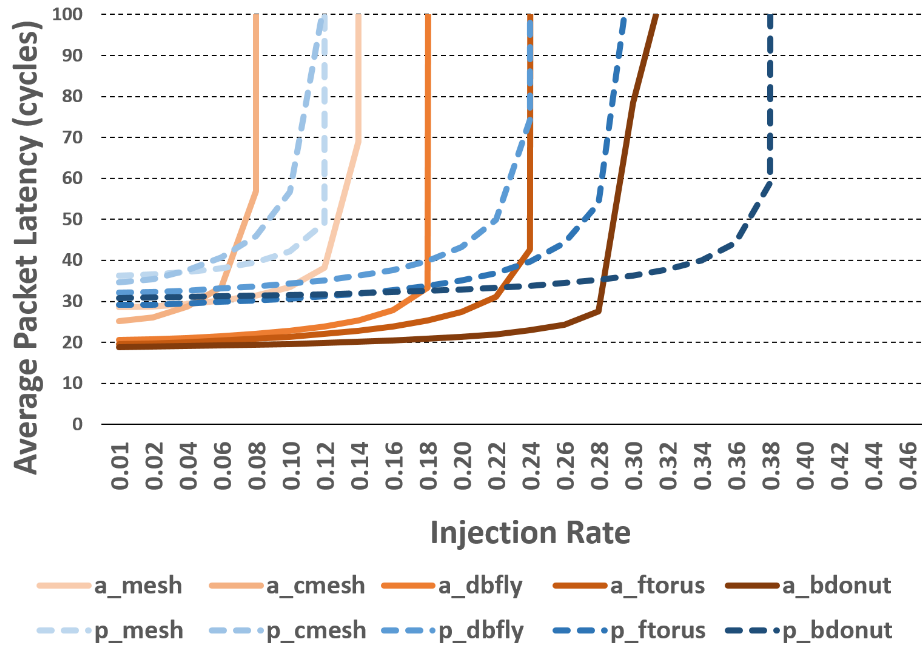


Figure 5.25: Network latency versus injection rate for active and passive interposers.

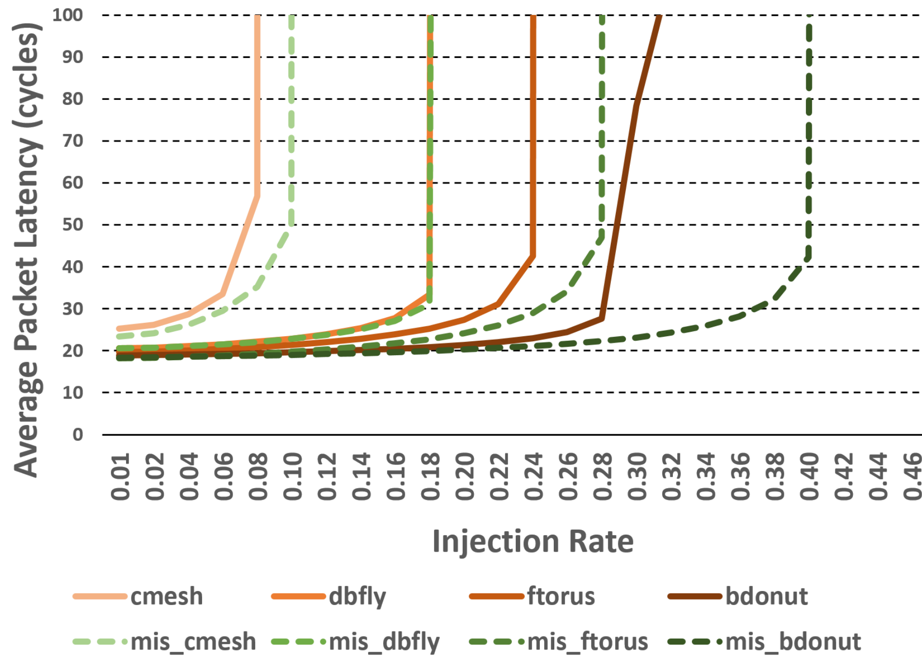


Figure 5.26: Impact of misaligned topologies on active interposer average packet latency.

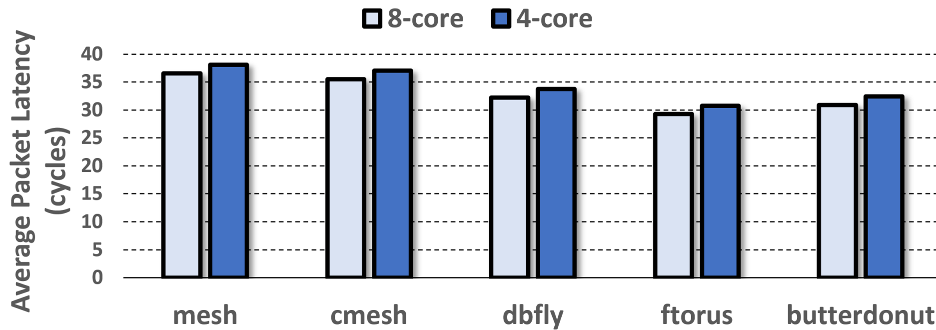


Figure 5.27: Impact of eight-core vs. four-core chiplet selection on passive interposer average packet latency.

active interposer cases. Additionally, the misaligned topologies result in higher saturation throughput and can tolerate heavier traffic loads.

Chiplet Size on Latency: Figure 5.27 shows average packet latency for eight-core and four-core chiplet sizes. This evaluation only considers the passive interposer case, since the active case does not incur an off-chiplet synchronization overhead and thus perform similarly on different chiplet sizes. As shown in the figure, smaller chiplet sizes result in higher average network latencies due to the increased frequency of synchronization overhead between chiplets.

To summarize the latency results: 1) Active interposers reduce latency versus passive interposers, 2) Misaligned topologies reduce latency and improve saturation throughput due to lower diameter, 3) Smaller chiplets can increase latency in passive interposers, but they have no effect on latency for active interposers.

5.6 Conclusion

By examining the interaction of interposer technology and network topology, this work concludes that both passive and active interposers may be cost-effective platforms for chiplet integration, depending on system requirements. From a yield and cost per-

spective, active interposers should generally be implemented using mature process technologies with lower wafer cost, as any yield benefit from smaller routers is overshadowed by the high wafer cost of advanced processes. When only considering bisection bandwidth, passive interposers achieve the same performance at lower cost than active interposers (given the same chiplet size). However, the long multi-cycle links and frequent clock-domain crossings in passive interposers introduce additional latency, and passive interposer systems may sometimes benefit from larger chiplets to reduce this latency overhead.

Chapter 6

Network-on-Chip for Monolithic 3D Integration

As classical transistor scaling becomes more challenging, alternative manufacturing technologies may be necessary to achieve continued performance and efficiency improvements. Of these emerging technologies, Monolithic 3D (M3D) integration is a promising integration methodology to increase transistor density and reduce interconnect distances. By sequentially manufacturing multiple tiers of active devices and integrating these tiers with monolithic inter-tier vias (MIV) with diameters of 100 nm or less, M3D integration can provide orders of magnitude more connectivity than current Through-Silicon Via (TSV) integration with significantly less power and delay. This technology thereby enables transistor-level and gate-level partitioning, collapsing the distance between gates within and between modules to reduce interconnect distance, average gate size, and buffer count. However, M3D also introduces new constraints and potential concerns that may have a critical impact on the system architecture and circuit design.

Prior research has characterized the benefits of M3D integration for a range of modules, architectures, and process technologies [79, 80, 81, 82, 83] using novel 3D-aware

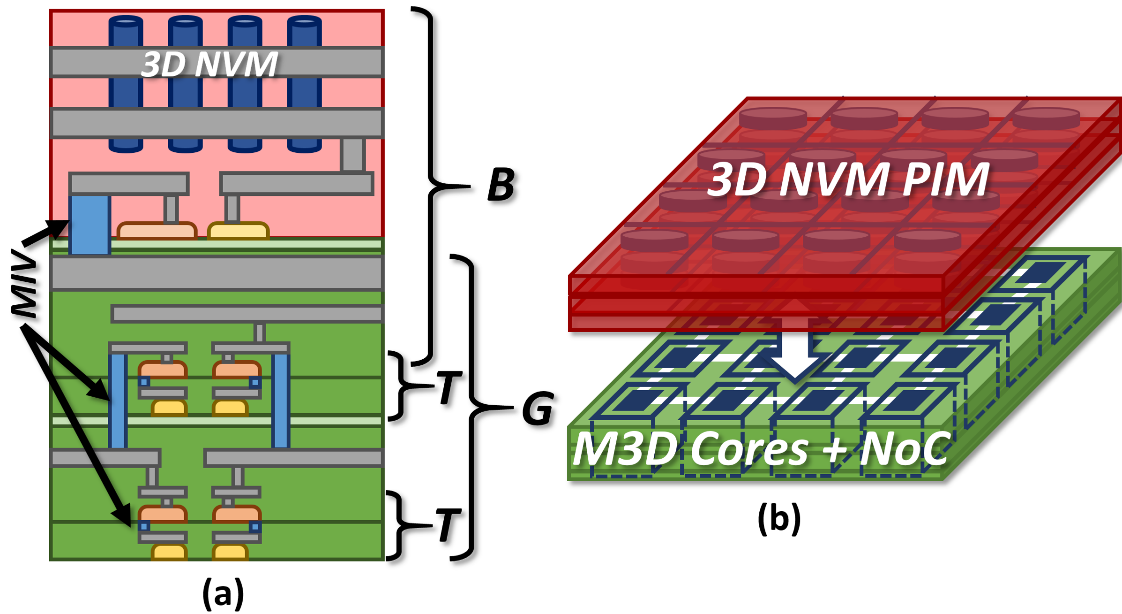


Figure 6.1: Envisioned future M3D system. (a) Process stack with T: transistor-level partitioning, G: gate-level partitioning, and B: block-level memory-on-logic partitioning with 3D NVM. (b) System with CMOS M3D cores and NoC as well as heterogeneous 3D NVM layers (separated for visibility).

design automation tools [84, 85]. Although much prior work has addressed M3D methodology and benefits, significant work remains for the optimization of system architectures to best leverage the benefits of M3D integration while handling the associated constraints and complications. Accordingly, this section investigates the interaction between M3D process technology and the proper design of system-level architectural components, especially in the context of on-chip communication. In particular, this work addressed the communication requirements of M3D systems through an analysis of Network-on-Chip (NoC) physical and logical topologies, with consideration of M3D fabrication constraints, and topology recommendations are made to improve flexibility, efficiency, and performance. This work also addresses the storage and compute requirements of M3D systems by investigating the heterogeneous integration of Non-Volatile Memory (NVM) Processing-in-Memory (PIM), which can provide efficient storage and parallel computa-

tion with little power overhead. By analyzing an NVM PIM accelerator [86], existing bottlenecks are identified that can be solved via the cooperation between NVM and CMOS M3D integration. An envisioned M3D system is depicted in Figure 6.1 with multiple tiers of M3D transistor- and gate-level integration, an M3D NoC backbone for scalable communication, and 3D NVM PIM for storage and parallel compute acceleration.

6.1 Monolithic 3D Integration

Prior work has studied the technology and design methodologies necessary for M3D, often by leveraging existing 2D design automation tools [84, 80]. These flows have been leveraged to study the transition from existing circuits into multi-tiered M3D circuits, which can be partitioned at the level of blocks, gates, or transistors.

This section provides an overview, benefits, and challenges of M3D integration, which can be partitioned at the level of transistors, gates, or blocks.

Block-Level Partitioning. In this scheme, functional blocks are partitioned into separate tiers at a coarse granularity. Block-level partitioning (*B-M3D*) has the benefit of utilizing existing layouts and macros, and fits naturally with "memory-on-logic" 3D partitioning, but does not reduce intra-block interconnect. Fine-grain blocks can reduce power by 7-16% [85], but coarse memory-on-logic partitioning achieved less than 2% power reduction [81, 84]. Inter-block distance reduction may be hindered by block imbalance when partitioning [87], and footprint reduction may also be hindered by this block imbalance [85]. Therefore, *B-M3D* achieves less efficiency improvement than *G-M3D* (and converges to *G-M3D* as block granularity becomes finer). Accordingly, this work focuses on *G-M3D* over *B-M3D* for optimal NoC-based M3D systems. However, block-level partitioning may be useful for heterogeneous M3D process integration to provide connectivity between different technologies.

Gate-Level Partitioning places gates across two or more separate layers, resulting in a coarser granularity than *TR-M3D*. Gate-level partitioning (*G-M3D*) can reduce interconnect distance, gate size, and buffer count by increasing intra-block gate density, resulting in consistent footprint reduction and power improvement (6-22%) [88, 81], especially when logic and memory can both be folded across tiers [84, 81]. Hundreds of thousands of MIVs can be used in gate-level partitioning, versus thousands in block-level partitioning. Gate-level partitioning can reduce interconnect distance, gate size, and buffer count by increasing intra-block gate density, resulting in consistent footprint reduction of 50% or more compared to 2D integration [89]. *G-M3D* can use existing 2D standard cells, however it requires 3D EDA tools [89].

Transistor-level partitioning increases circuit density by placing NMOS and PMOS transistors on separate tiers. In transistor-level partitioning (*TR-M3D*), standard cells can be created with multiple sub-100nm MIVs in each gate, resulting in size reductions of 40% for logic gates [79] and 40-45% for SRAM cells [82], but ideal footprint reduction is not achieved due to NMOS/PMOS mismatch and MIV overhead [79]. Total power can be reduced versus 2D by 3-35% [79]. One drawback of *TR-M3D* is that increased pin density results in increased route congestion. Accordingly, additional metal layers, at additional cost, are necessary to achieve optimal efficiency [90]. Despite this, digital *TR-M3D* circuits can be designed in a similar manner as a 2D process technology with standard cell place and route methodology, as all inter-tier connections are encapsulated within the standard cells.

M3D integration strategies also introduce important technology considerations that influence the circuit and system design.

MIV Overhead: The monolithic inter-tier vias introduce very little delay and area overhead, especially compared to Through-Silicon Vias. MIV diameters are similar to interconnect vias and smaller than standard cells. When driven by a 45nm 4x inverter,

MIV delay was only 40ps, more than 18x less than a 5um TSV [91].

Metal Stack: The partitioning strategy influences the optimal metal stack. Gate-level and block-level partitioning require metal stacks on each tier. Because the MIV pitch is determined by the widest metal pitch on the tier below, lower tiers are limited to intermediate-width interconnect, and global interconnect is only available to the top tier [92]. These process stacks are visualized in Figure 6.2. All partitioning strategies, especially transistor-level partitioning, increase route congestion due to increased gate or pin density [92]. More metal layers may be necessary to achieve optimal benefits.

Clock and Power Distribution: Due to the increased circuit density and interconnect contention, clock and power delivery must serve a greater load per area while using fewer interconnect resources. The clock backbone is most efficient when utilizing the top interconnect of only one tier [84]. M3D integration does improve the IR drop, skew, power, and wire overhead, for the clock and power network [93, 84], but the *relative* wire and power overheads are larger for M3D.

Process Complications: Due to the additional processing steps of sequential integration, as well as the potential increase in the number of metal layers, M3D integration is likely to increase the per wafer process costs. Analysis from prior work suggests that in 7nm technology M3D integration may only be cost effective for large, complex processors [80]. Until M3D processes develop further, temperature requirements for sequential manufacturing may require either slower tungsten interconnect on the lower tiers or weaker transistors on the upper tiers for *G-M3D* and *B-M3D* [87]. Although EDA solutions can reduce this performance degradation for some logic through selective partitioning [85], lengthy NoC links are sensitive to degradation in interconnect resistance and drive strength.

Thermal Density: Perhaps the largest challenge with M3D integration is the sudden increase in thermal density. Even with thermal-aware placement and M3D power

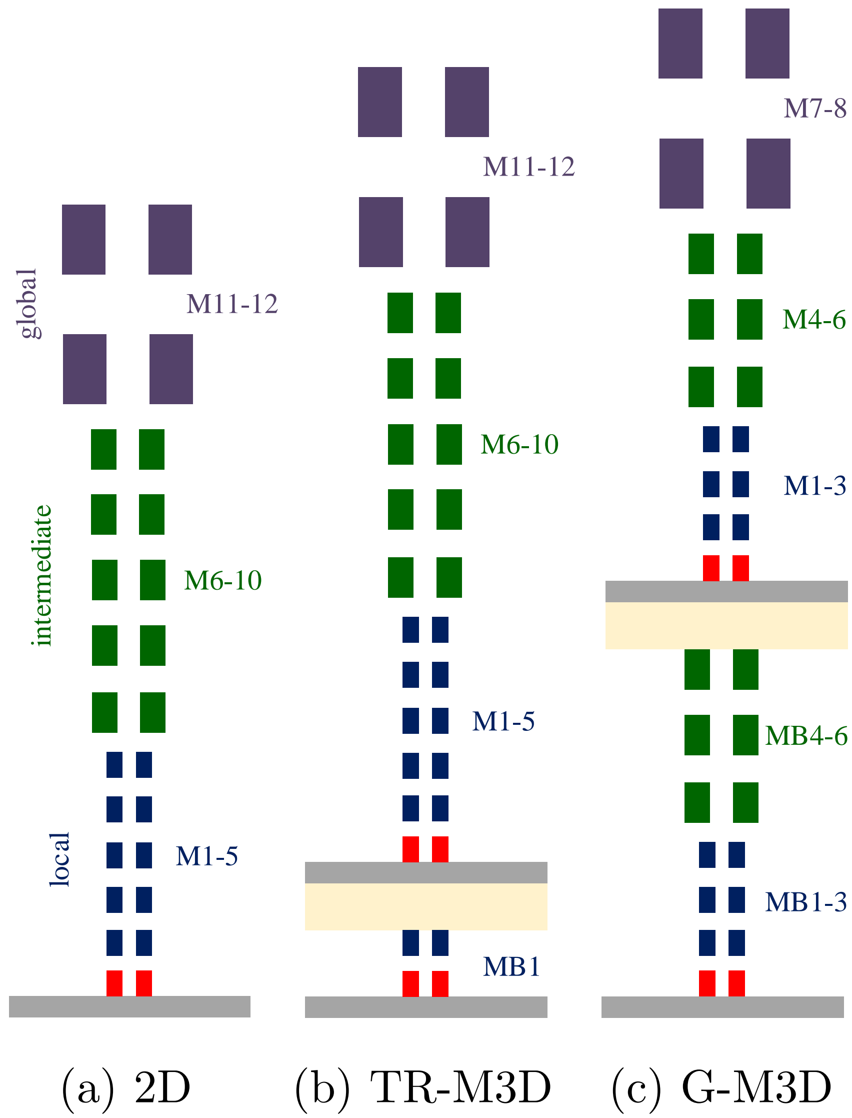


Figure 6.2: Metal layer stack diagrams.

improvement, average temperature increases almost linearly with the number of tiers [94]. Thermal management is already a critical architectural concern in the era of dark-silicon architecture, and increased thermal density may mean that M3D systems will require further throttling. Fluidic microchannels may be necessary to manage these increased thermal levels, while future device technologies like CNFETS [95] may reduce device power to acceptable levels. Unlike TSV-based 3D chips, there is little thermal resistance between tiers and MIV placement is not a concern for thermal management [94].

6.2 Necessity for M3D Network-on-Chip

Network-on-Chip can be applied to provide an efficient, scalable, and standardized communication infrastructure for future heterogeneous many-core M3D systems. NoCs leverage routing logic to improve the utilization of dedicated interconnect resources and to provide parallel, pipelined transmission. However, the high activity and additional logic can cause the NoC to constitute a significant portion of total system power [96], and poor design can result in contention and long packet latency. With proper design, NoCs can instead provide M3D systems with high bandwidth and low latency communication with reduced area and power overheads.

Prior work has studied topologies for three-dimensional NoCs, mostly for TSV-based systems where vertical links have limited bandwidth, significant delay, and considerable area overhead. These 3D NoC have introduced 1) multiple vertical hops with discrete routers on each tier, 2) shared vertical busses that can cause contention, or 3) high-radix concentration between routers and network interface units on multiple tiers [97, 98, 99, 100]. These solutions, while appropriate for TSV-based 3D integration, are less suitable for the unique properties of M3D processes.

6.2.1 Design Considerations for M3D On-Chip Communication

M3D integration introduces the following unique benefits and concerns: 1) MIV delay is a fraction of the cycle time and area overhead is minimal even for wide buses, 2) power efficiency is a primary design constraint due to thermal density, and 3) for gate-level and block-level partitioning, global interconnect is only available on the top-most tier due to MIV pitch constraints. Additionally, most M3D systems with gate-level and transistor-level partitioning will be logically two-dimensional to maximize intra-block interconnect reduction. However, heterogeneous memories or other differentiated processes will still require block-level partitioning on discrete tiers.

To best leverage the high-bandwidth, low-delay interconnect provided by Monolithic 3D integration, this work proposes a NoC without a contention-causing 3D bus or vertical logical hops. Instead, routers and links span the small physical distance between monolithic 3D tiers, compressing the 3D circuit onto a logically 2D NoC. Increasing the density of the router can reduce the intra-router distances to improve efficiency. Common NoC topologies like 2D mesh will continue to map to M3D systems.

Because of the reduced floorplan and increased density from M3D integration, an NoC mapped onto the M3D system will have reduced link distances. While it is possible to increase the router concentration factor, connecting more units to fewer routers, this increase in router port count increases per-router contention and results in poor area and efficiency scaling [98]. This work instead suggests that the NoC leverage a Single-cycle Multi-hop Asynchronous Repeated Traversal (SMART) scheme, in which flit latency can be reduced below the number of logical hops by allowing for multiple hops in a single cycle when network contention is low [101]. SMART NoCs are sensitive to the link interconnect length, so reducing the system's 2D footprint via M3D integration has been shown to increase the maximum hops per cycle while maintaining an efficient frequency,

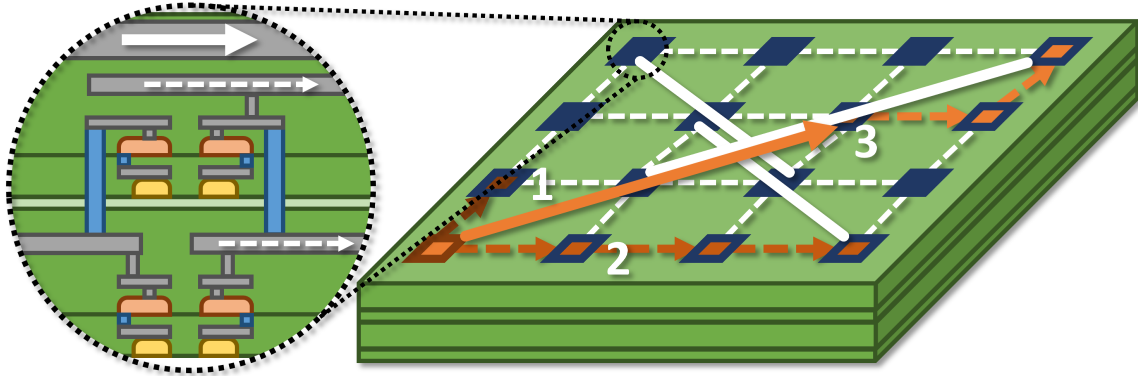


Figure 6.3: M3D NoC with inter-tier routers, SMART mesh with variable-width neighbor links (dashed) utilizing intermediate interconnect on each tier, plus additional bypass links (solid) on top level global interconnect. In this example, all nodes are within three hops and may be accessible in one cycle through: 1) neighbor links, 2) multiple neighbor links, and 3) bypass and neighbor links.

with 28% latency reduction over 2D [100].

A unique challenge to gate-level and block-level M3D integration is that global interconnect is only available on the top tier. Further, the extra demands of M3D clocking and power distribution means that much of the global interconnect will be utilized, and less is available for NoC links. Instead of utilizing global interconnect for all NoC links, this work proposes to only utilize the top tier's global interconnect for a reduced number of long-distance bypass links, while short-distance neighbor links utilize the intermediate interconnect available to each M3D tier. The SMART NoC is designed to send short-distance messages through one or more neighbor links per cycle, while long-distance messages can utilize the bypass links when available to reduce average latency without excessive global route utilization. For the example in Figure 6.3, bypass links reduce maximum hop count from six to three, enabling single-cycle SMART transmission.

To further improve NoC efficiency, the natural partitioning of the M3D tiers can be leveraged for variable-width neighbor links. In many applications, a significant portion of messages are small control messages (64 bit or less). For example, more than 50% of messages in the PARSEC 2.1 benchmark are 64 bits or less [96]. These short messages

incur poor efficiency if a wide link width is selected, but a small link width would result in significant sequentialization latency and increased likelihood of contention for wide data messages. By splitting each link and router to handle a subsection of the phit/flit width, the unused sections can be clock gated or power gated to improve efficiency. Similar ideas have been recently proposed for 2D networks to improve efficiency through clock and power gating by 25-45% [96, 102], but the discrete nature of M3D tiers, with associated clock and power network subbranches, makes this partitioning and gating a natural, low-overhead design. It also ensures a regular, balanced usage of intermediate interconnect on each tier. Figure 6.3 summarizes the proposed logically-2D M3D NoC with inter-tier routers, global bypass links, and multi-tier variable-width neighbor links.

6.2.2 Bandwidth in Processing-in-Memory Accelerators

In addition to the efficient communication provided by the NoC, efficient storage and computation are also necessary to manage M3D thermal density. These can potentially be provided by the heterogeneous integration of emerging Non-Volatile Memory (NVM), which may offer high-density storage with reduced static power and without a refresh power overhead. NVM has already been sequentially integrated in several successful demonstrations. Additionally, NVM can be utilized for very efficient Processing-in-Memory by performing massively parallel analog computations directly from the memory array. However, existing NVM PIM systems are either handicapped by the constraints of the memory process technology, or they introduce significant area overheads when adding logic to the peripheral circuits. They may also, demonstrated later, contain communication bottlenecks that cannot be managed by standard memory busses. In this section, an NVM PIM architecture is investigated to demonstrate how the high-bandwidth communication between heterogeneous M3D tiers can be used to improve PIM efficiency and

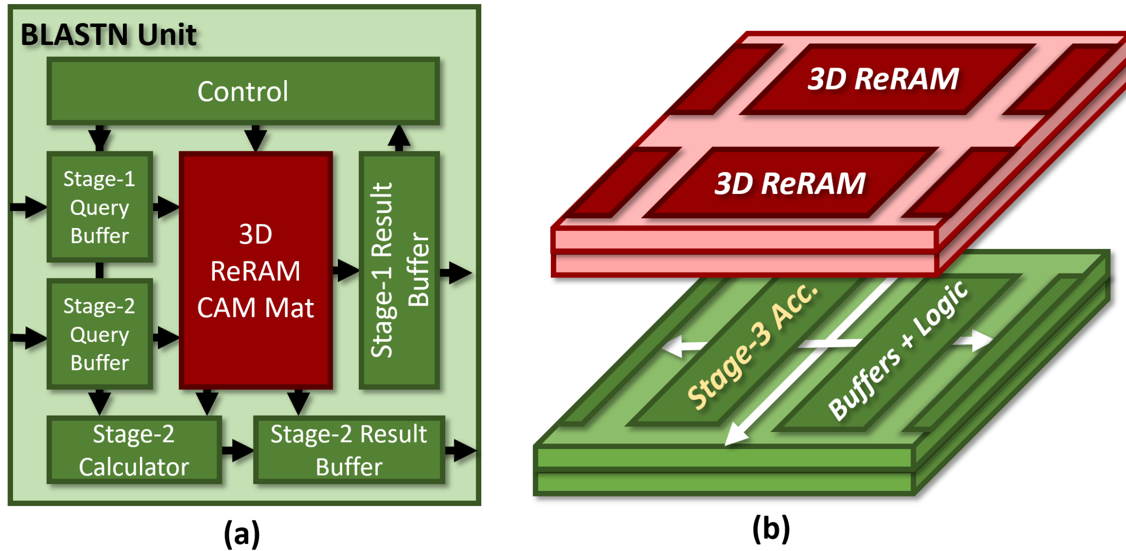


Figure 6.4: (a) 2D RADAR Unit with 3D ReRAM and CMOS buffers and logic. (b) M3D memory-on-logic system with heterogeneous process tiers, Stage-3 accelerator, and NoC.

flexibility.

As a case study, this section investigates the RADAR system: a DNA alignment accelerator that utilizes 3D-ReRAM-based CAM to perform very efficient, parallel matching operations for the BLASTN sequence alignment algorithm [86]. The BLASTN algorithm is composed of three stages: 1) word matching of fixed sized segments, 2) ungapped extension in both directions for all matching segments, and 3) gapped extension to look for dropped or swapped bases. The first two stages, word matching and ungapped extension, can be efficiently performed by exact pattern matching using ReRAM-based CAM, and they produce almost all the runtime on CPU hardware: 83.9% and 15.9%, respectively. The third stage, gapped extension, can be solved using the Smith-Waterman dynamic programming algorithm, but cannot be computed in the RADAR accelerator.

Benefits of PIM: The RADAR system is able to greatly accelerate the first two stages of the BLASTN algorithm by 1) leveraging 3D-ReRAM CAM for efficient comparison operations, 2) storing the entire DNA sequence (potentially tens of GBs) in

dense 3D NVM to minimize data movement and write energy, and 3) performing massively parallel searches across a hierarchy of multiple units with multiple ReRAM arrays with multiple rows of 3D cells. Compared to a CPU baseline, RADAR achieves a 5114x speedup with 386x energy reduction for the first two stages of BLASTN.

Limitations of PIM: 1) Although RADAR provides significant acceleration, it encounters an Amdahl's Law bottleneck with the third stage, gapped extension, of the BLASTN algorithm. Although the first two stages, which compose 99.8% of the algorithm runtime, can be accelerated by 5114 times, the remaining stage cannot be solved with the PIM hardware, and the total acceleration ratio is reduced by an order of magnitude to about 456x. 2) To support BLASTN acceleration, other support hardware, including buffers and computation logic, must be added to the ReRAM memory arrays. This extra hardware introduces more than a 25% area overhead even before considering efficiency loss from array alignment and poor memory transistor performance. 3) The internal memory bus and external interfaces are further communication bottlenecks. Each 4MB unit (with 64 ReRAM arrays) can concurrently produce over 6KB of 46-bit word match messages, which is over 18KB with the native 128-bit memory word length. For example, in a 32 GB RADAR system with 8K of these units, 155MB may be produced during each query, requiring more than 1.8 TB/s bandwidth.

Improvement with M3D: The limitations of the PIM accelerator can be removed through heterogeneous M3D integration. 1) High-performance CMOS transistors can be utilized for efficient acceleration of the gapped extension bottleneck. Prior work on FPGA-based Smith-Waterman acceleration has achieved a 330x speedup [103]; ASIC acceleration at higher frequency would be even larger. By removing the Amdahl's Law bottleneck, total speedup would improve from 456x to 4975x (only assuming a conservative 330x FPGA acceleration). 2) By moving non-ReRAM buffers and logic to the CMOS tier, ReRAM density and regularity can be improved. 3) Utilization of CMOS

for an array of distributed SW ASIC accelerators, connecteted by the NoC fabric, can provide TB/s levels of on-chip bandwidth to solve the communication bottleneck before the gapped extension stage. Variable-width NoC links are an efficient match for the short (¡64-bit) RADAR match messages.

6.3 M3D Interconnect Characteristics

In this section, the interconnection of each M3D partitioning schemes is characterized by evaluating their metal layer stacks and modeling the route delays for the NoC links.

6.3.1 M3D Partitioning Comparison

Below is a list of assumptions for the footprint and wirelength comparison between M3D partitioning techniques, based on recent M3D EDA results [90][89][84].

2D Baseline. 2D integration is selected as the baseline for footprint and wirelength comparison.

***TR-M3D* Partitioning.** Transistor-level partitioning can provide a 40% footprint reduction. The *TR-M3D* footprint is 0.60× that of 2D. This translates to wirelength of 0.77× of 2D.

***G-M3D* Partitioning.** Gate-level partitioning can provide at least a 50% footprint reduction. The *G-M3D* footprint is 0.50× that of 2D. This translates to wirelength of 0.71× of 2D.

6.3.2 Metal Layer Stack Characterization

In this section, the metal layer stack of each M3D partitioning scheme are characterized. Figure 6.2 shows the metal layer stack diagrams for 2D integration, *TR-M3D*

partitioning, and *G-M3D* partitioning schemes.

2D Baseline. The 2D baseline has 5 local metal layers, 5 intermediate metal layers, and 2 global metal layers.

***TR-M3D* Partitioning.** The top tier of the *TR-M3D* scheme identical to the 2D metal layer stack. The bottom tier only has one local metal layer used for intra-cell connectivity. Increased cell density ($1.7 - 2.0\times$ compared to 2D) results in route congestion [90]. Previous proposals add more local and intermediate metal layers [90], however this is not possible without cost overheads. Therefore, no additional metal layers are supplied for this comparison.

***G-M3D* Partitioning.** The *G-M3D* scheme requires multiple metal layers in both tiers. The bottom tier lacks global interconnects, as the MIV pitch is determined by the widest metal pitch of the bottom tier, which causes per-cell contention for the global metal resources. Due to sequential manufacturing requirements, the bottom tier may need to replace copper interconnect with slower, resistive tungsten interconnect. Both copper and tungsten interconnect are analyzed in this study.

6.3.3 Interconnect Characterization

In this section, the interconnect resources of each M3D scheme are characterized for select NoC topologies.

Methodology. HSPICE was used to compute the delay of optimally-driven, optimally-repeated links for each metal category in the ASAP7 7nm PDK. A three-segment pi-model interconnect was used to model each segment between repeaters. Worst-case parallel-neighbor parasitic capacitance is modeled to determine maximum link delay. Delay results are for copper interconnect RC values unless specified as tungsten.

Metal Layer Performance. Figure 6.5 shows link delays for each metal layer

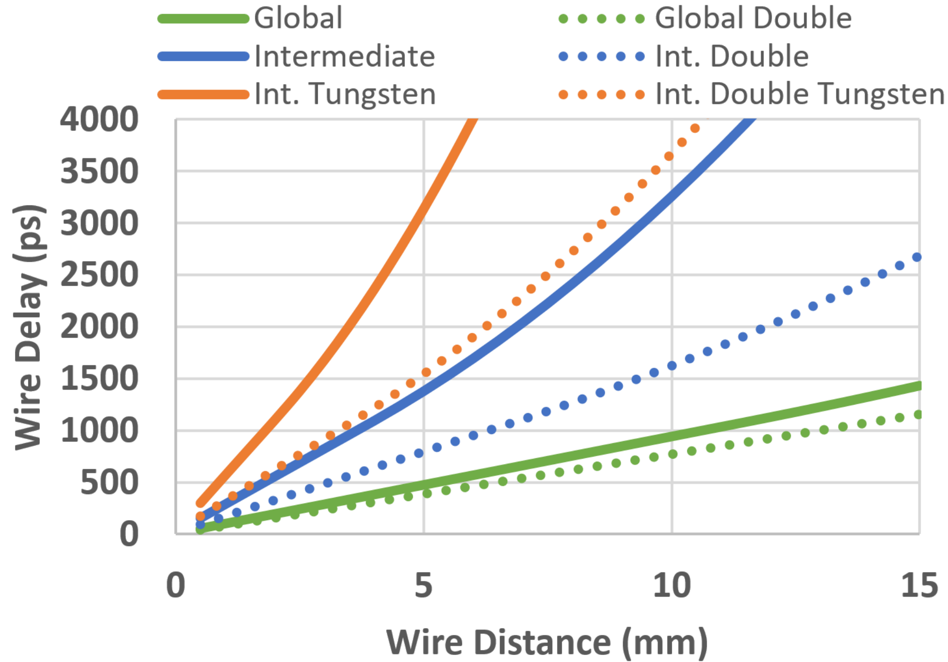


Figure 6.5: Metal layer link delay characterization versus distance. "Double" indicates double-width, double-spacing.

category plotted against link distances for the repeated links. "Double" corresponds to double-width, double-spaced interconnect, which utilizes twice the amount of available interconnect in order to reduce interconnect resistance.

NoC Baseline. As a feasible case study, a NoC-based system is selected with 16 chips that are $5mm \times 5mm$ in size (for the 2D baseline), arranged in a 4×4 layout. The design space is constrained to synchronous NoCs with single-cycle links. To maintain a fair comparison between topology resources, routers are limited to a maximum of 5 ports[1], including the local connection. Accordingly, the following topologies shown in Figure 6.6 are analyzed: mesh, torus, folded torus (FTorus), and double butterfly

[1]While not evaluated in this work, high-radix routers are starting to have a wide-spread adoption. In M3D, it is possible to implement a NoC in a true 3D approach in order to avoid hops in the vertical dimension. However, higher radix demands more routing resources. Routing resources are already limited in M3D, but narrower channels can be used if radix increases. Trade-off between higher radix and narrower channels translates to a trade-off between the number of hops and number of flits in a packet. However, maximum link length is another limiting factor, as our evaluations will show that long links often limit achievable NoC frequencies.

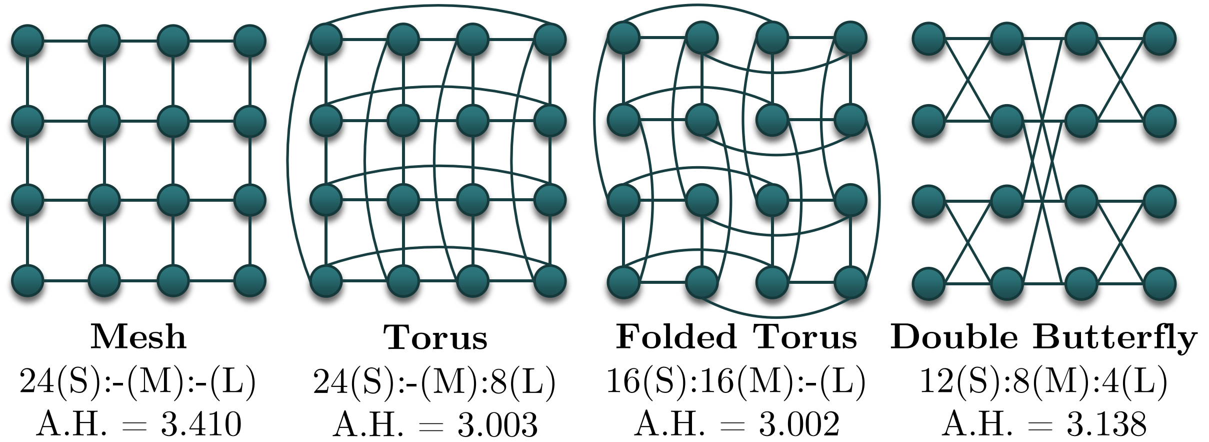


Figure 6.6: Evaluated topologies and characteristics: Number of short (S), medium (M), and long (L) links and average hops (A.H.)

M3D Type	Mesh	Torus	FTorus	DBFly
2D	1492	621	875	621
G-M3D	1857	837	1157	837
G-M3D (Tungsten)	1857	763†	763†	763†
TR-M3D	3005*	1694	1752†	1694

Table 6.1: Maximum frequency (in MHz) for M3D types and topologies. ($\star$) Maximum frequency is capped at 2000 MHz. ($\dagger$) Limited by intermediate metal layer.

(DBFly). The number of links in each topology for each relative length are listed in Figure 6.6. Note that medium links are twice as long as short links, and long links are three times as long. Diagonal distances for the double butterfly topology are Manhattan distances due to X-Y track routing.

M3D NoC Frequencies. Using the metal layer characteristics, Table 6.1 shows the maximum frequency calculations for each M3D partitioning scheme for the link distances of the four topologies that are considered. A 200ps overhead for skew and sequential capture, based on ASAP7 HSPICE timing, are added to the link delay to convert from delay to synchronous frequency. The maximum network frequency, independent of link delay, is capped at 2000 MHz based on the selected router microarchitecture. All of the longest links across topologies utilize the available global metal layers, but, to limit global

interconnect utilization, the short links utilize intermediate metal for all but the mesh topology.

The mesh topology achieves the fastest interconnect due to the short links between all router connections. The torus generally has slower links, limited by the long wrap-around links. The folded torus achieves higher frequency as the longest links are shorter. The double butterfly has equally long links as the torus, and with both, all diagonal links utilize the global interconnect.

To have a resource-equivalent comparison of the available interconnect between each manufacturing scheme, the 2D baseline has sufficient global interconnect for double-width, double-spaced links, but it suffers from the relatively long wirelength compared to M3D integration. *G-M3D* can take advantage of the multiple tiers of intermediate links to offset the limited global interconnect on the top layer. Two *G-M3D* technology options were considered: a baseline case without low-temperature process consideration and a case with tungsten interconnect on the bottom intermediate tier. For *G-M3D*, as indicated in Table 6.1, long links in global metal limit the NoC frequency, so it is safe to assign shorter links to the intermediate metal layers. *TR-M3D*, can take advantage of a reduced M3D wirelength, although the congestion from increased cell density poses a limit to interconnect resource availability.

6.4 M3D NoC Design Guidelines

Monolithic three-dimensional (M3D) integration is viewed as a promising improvement over through-silicon via based 3D integration due to its greater inter-tier connectivity, higher circuit density, and lower parasitic capacitance. With M3D integration, network-on-chip (NoC) communication fabric can benefit from reduced link distances and improved intra-router efficiency. However, the sequential fabrication methods utilized

for M3D integration impose unique interconnect requirements for each of the possible partitioning schemes at transistor, gate, and block granularities. Further, increased cell density introduces contention of available routing resources. Prior work on M3D network-on-chips has focused on the benefits of reduced distances, but has not considered these process-imposed circuit complications. In this section, NoC topology decisions are analyzed in conjunction with these M3D interconnect requirements to provide an equivalent architectural comparison between M3D partitioning schemes.

6.4.1 NoC Topologies

In this section, NoC topologies with maximum radix of 5, including the local connection, are analyzed for each M3D partition scheme. The mesh, torus, folded torus, and double butterfly, shown in Figure 6.6, are evaluated. Figure 6.6 also presents a comparison of these topologies in terms of average hops and the number of links of each distance.

Methodology. SynFull [104] integrated with BookSim [78] is used to evaluate the NoC topologies. Router microarchitecture is a 4-cycle pipeline. The default flit width (link width) is 32 bits, with 64 bits for the double-wide case. Networks are evaluated on PARSEC and SPLASH-2 benchmarks using SynFull.

Topology Comparison. Figure 6.7(a) compares the four topologies in terms of average network latency in cycles by sweeping the injection rate for uniform random synthetic traffic. As expected, the mesh has the highest zero-load network latency, due to the high average hop count. The torus and folded torus perform similarly and have lower latencies and higher saturation throughputs than mesh. Although the double butterfly topology has a lower network latency than the mesh, results suggest that it is not as scalable and has a lower saturation throughput.

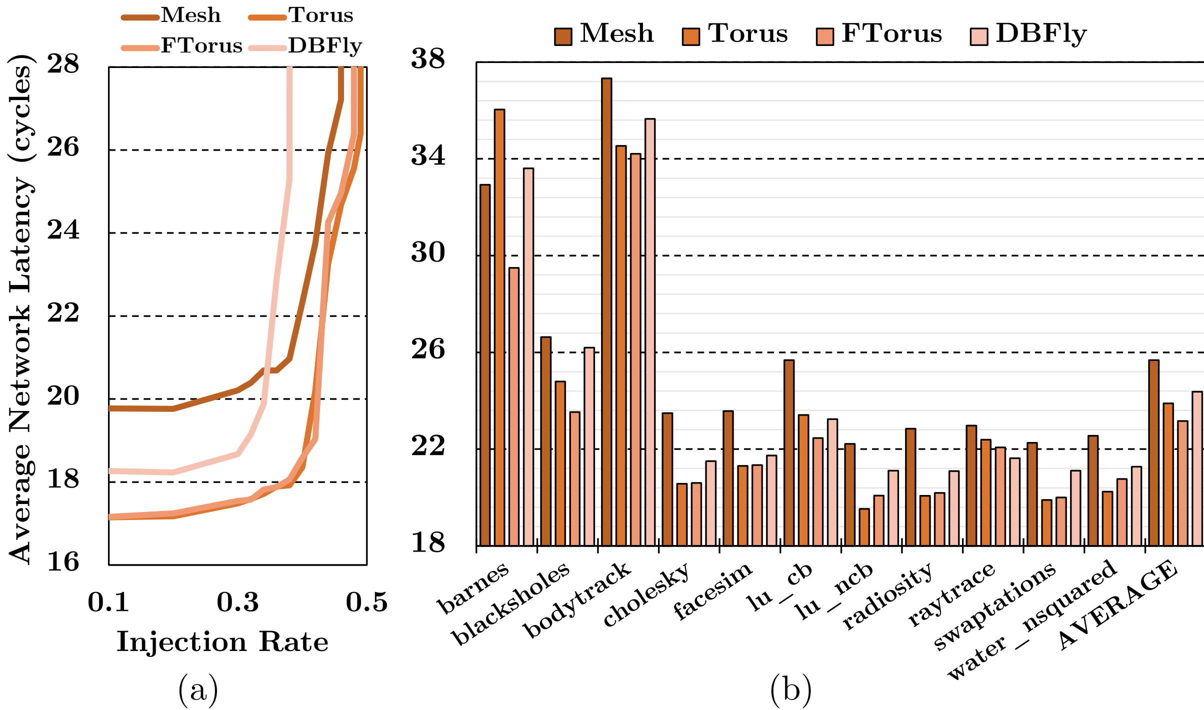


Figure 6.7: Topology comparison in terms of average network latency in cycles for: (a) uniform random synthetic traffic and (b) PARSEC and SPLASH-2 benchmarks

Figure 6.7(b) shows topology comparison results in terms of average network latency in cycles using SynFull, for the double-wide case, for PARSEC and SPLASH-2 benchmarks. The slowest topology is the mesh, while the fastest topology is the folded torus, which corresponds with their average hop counts. The average network latency changes by up to 12.5%.

These experiments were repeated for the default 32-bit link width, as well as half-wide (16-bit) links, in order to demonstrate the impact of routing congestion. Using 32-bit links increases average network latency by up to 34.2% on average over double-wide links. A pessimistic comparison is assumed with *TR-M3D* having 50% less routing resources. On average, reducing the routing resources by half increases average network latency further by up to 44.4% on average, over the default link width. For each configuration, available interconnect can either be utilized for link bit width (cycles) or physical interconnect

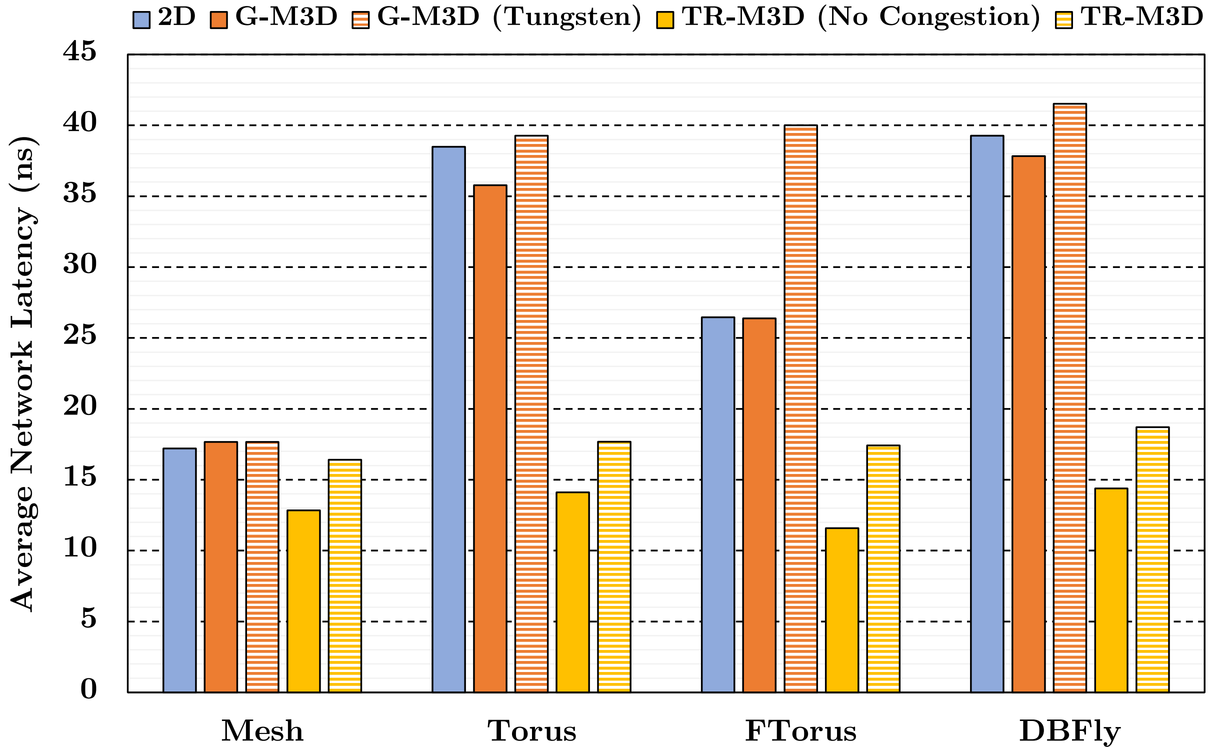


Figure 6.8: M3D NoC latency comparison

width (frequency), and the lowest-latency configuration is selected.

6.4.2 Network Latency Comparison

This section presents the network latency results for the different partitioning strategies on the four topologies by combining interconnect characterization and topology comparison for PARSEC and SPLASH-2 benchmarks. Figure 6.8 shows the average network latency of the four topologies for 2D, *G-M3D*, and *TR-M3D* partitioning schemes.

2D integration favors topologies with shorter links as the main source of latency reduction comes from higher frequency links. Due to lower cell density, 2D can utilize double-wide 64-bit links to reduce NoC latency.

G-M3D, for all but the mesh topology, uses intermediate metal across both tiers for short links and global metal in the top tier for long links. *G-M3D* latency is compara-

ble to 2D and within 10% lower latency except for mesh, without consideration of the low-temperature manufacturing requirements. When slow tungsten interconnect in the bottom tier is instead utilized, the latency significantly degrades by up to 1.51× that of 2D. The low latency of the mesh comes from using only global metal layer in the top tier, but due to global metal resource contention, *G-M3D* is constrained to default-width 32-bit links.

TR-M3D using global metal on the top tier can achieve a significant latency advantage from the reduced wirelength. Without taking routing congestion into account, and utilizing double-wide 64-bit links, *TR-M3D* latency for a given topology (FTorus) can be as low as 0.37× of 2D. When routing congestion is considered by constraining to only half the routing resources, the latency increases but still stays below that of 2D integration, improving by up to 0.46×.

6.4.3 Discussion

This section presents M3D NoC guidelines based on the results and observations above.

- Frequency plays a significant role in NoC performance and it is not possible to select a topology by only considering the average hops. Accordingly, the mesh frequently performs best, despite a greater hop count, because it can achieve faster link frequencies.

- Despite the greater distance reduction, *G-M3D* may not be better than 2D because of the limited global interconnect.

- The temperature manufacturing requirements for *G-M3D* can result in even higher NoC latency than 2D. Except for the mesh, frequency is limited by the short links

routed in intermediate tungsten metal in the bottom tier. Therefore, *G-M3D* favors topologies with short links to better utilize the limited global metal in top tier.

- *TR-M3D* can achieve significant improvement, due to shorter wirelength and the availability of global metal layers.
- With routing congestion taken into account, *TR-M3D* can still provide significant improvement over 2D for most topologies. Mesh performs the best but achieves latency similar to 2D. *TR-M3D* introduces a trade-off between extra routing resources and cost.

While thermal characteristics are not compared in this work, the NoC frequency, number of links, and link width are expected to impact the thermal characteristics of the M3D NoC.

6.5 Conclusion

Monolithic 3D integration is a promising sequential manufacturing technique that can can improve circuit density and reduce interconnect power versus TSV-based 3D integration, but system-level architectures must be optimized to improve power efficiency. The work proposes M3D-targeted NoC architectures to provide efficient communication for dense M3D systems with reduced interconnect utilization. This work also analyzes the heterogeneous M3D integration of NVM for storage and PIM with improved area and performance efficiency. Finally, this work identifies the manufacturing and interconnect challenges for various M3D partitioning schemes and analyzes M3D NoC delay and topology under a resource-equivalent comparison. Our evaluations showcase a design space exploration of M3D partitioning techniques and NoC architectures to provide design guidance on the interconnection related trade-offs for M3D-integrated NoC.

Chapter 7

Summary

With the cost and performance benefits of traditional transistor node scaling at risk, computer architects in academia and industry are beginning to transition away from monolithic System-on-Chip fabrication towards multi-die integrated systems. As demonstrated in this work, chiplet integration has the potential to improve manufacturing yields, amortize design costs through reuse, and integrate heterogeneous processes like stacked dynamic memories. Current multi-die systems leverage coarse-pitched organic substrates, or passive silicon interposers for stacked memory integration, but future performance and efficiency goals may require more advanced packaging technologies that can provide higher bandwidth and lower latency, such as Through-Silicon Via 3D integration, active interposer substrates, and monolithic 3D fabrication. At the same time, these emerging technologies introduce additional complications like increased power and thermal density and additional manufacturing cost overheads. The navigation of this increasingly complex design space of advanced packages and multi-die systems has not been thoroughly studied, especially in regards to competing packaging alternatives, yet this integration decisions impacts critical architectural features that can determine system-level performance and the construction of entire design families. This dissertation seeks

to provide the methodology necessary to navigate this new design space, while providing analysis on some of the most promising packaging technologies to suggest best practices for inter-die and inter-tier communication.

The first step towards selecting a new packaging and integration technology is to determine the cost effectiveness: can transitioning to a new technology reduce manufacturing cost, or provide more transistors at the same cost? This work starts with a cost estimation methodology for die fabrication with area and metal layer estimation and yield modeling. This baseline single-die model is then expanded for multi-die 3D integration with known-good-die validation, which introduces some fabrication and area overheads but can improve total silicon yield, with high-bandwidth low-latency inter-tier integration and reduced circuit footprint. The model is also expanded for 2.5D stacking, applicable for a number of substrates including passive and active interposers. Analysis of these models with industry price data suggests that both 3D and passive interposer 2.5D integration can be cost-effective despite their integration overheads, but only for sufficiently large-area systems that benefit from the yield improvement. Further, the multi-die cost model is modified to support modern binning techniques that are widely used in industry CPU and GPU products to improve yield by selling partially-disabled parts at lower cost. With this updated model, analysis concludes that multi-die integration is applicable for modern mainstream CPU systems, with improved full-functionality yield between 1.18x-1.46x, but best suited for large-area high-performance systems like server processors, where the number of fully-enabled dies can be increased by 2-4x through die partitioning, given current defect densities. Additionally, an investigation into the relative overhead of nonrecurring engineering costs, including mask and design costs, reveals the further benefit of chiplet-enabled die reuse, and methods for reusable integration platforms are discussed.

While the previous section demonstrated cost benefits of multi-die 3D and 2.5D inte-

gration, the models used did not account for the overhead of increased thermal density. To improve upon the prior analysis, thermal models were developed for 2D, 2.5D, and 3D systems, with cost overhead for any packaging and cooling that is needed to offset an increase in thermal density. With thermal overhead accounted for, the analysis changes to favor 2.5D passive interposer integration over TSV-based 3D integration for systems that are sufficiently large or high power, including the range of area and power density typically found in modern CPUs and GPUs. However, 3D integration, especially with only two dies per stack, may still be cost effective versus monolithic integration, even for mainstream consumer processors.

To further investigate the impact of die stacking, the next chapter presents a deeper investigation into the power density of modern die-stacked dynamic memory, which are the only fine-pitched many-die stacked circuits widely available today. This work develops an architecture-level power model for the High Bandwidth Memory family of devices, with support for memory traces from architectural simulations. Power contribution from an industry physical interface module is also included to capture the necessary power cost of data write transmission. The model is validated against real HBM hardware and then employed to provide detailed power breakdowns across the range of memory behaviors, from idle to peak random read, and for different stack heights. Then, based on industry memory roadmaps and expected scaling trends, the model is used to project the power utilization of near-future memory stacks with higher bandwidth and memory density. Given the slowdown in dynamic memory technology improvement, the model suggests that future memories could reach peak power densities of 20W per stack, which would have a significant impact on system thermal management and cost.

Next, based on the cost-benefit previously demonstrated for 2.5D interposer-based integration, this dissertation investigates the cost and performance trade-offs of active versus passive interposer technology, especially in respect to scalable high-bandwidth

Network-on-Interposer integration. While passive interposers are relatively low cost to manufacture given the simple all-metal construction, they potentially introduce communication bottlenecks that can reduce system performance. An alternative, active interposers with integrated transistors, provide several benefits for inter-die communication and system architecture, but active interposer fabrication is relatively costly. By investigating the design space of Network-on-Interposer, this work concludes that active interposer systems can still be cost effective versus large monolithic dies, while providing physical and architectural communication benefits that translate to reduced communication latencies between dies. Active interposers are most cost-effective on larger mature process technologies, despite reduced yields from an increased active area percentage, especially when fault tolerance methods are used to improve the interposer router and link yield. Passive interposers are more cost-effective than active interposers for the same bisection bandwidth and chiplet count, but networks over passive interposers have increased latency from slower repeater-less links and clock synchronization challenges, especially as chiplet count increases. The best network topologies for interposers are likely high-radix, low-hop topologies that leverage the widely-available link interconnect on the interposer.

Finally, looking forward to future monolithic 3D integration that utilizes sequential manufacturing of layers for ultra-fine inter-tier communication, a similar methodology is utilized to investigate the best communication architectures for these M3D systems. To motivate improvement in M3D on-die communication, this section investigates a promising M3D Processing-in-Memory accelerator, demonstrating that communication is a massive bottlenecks for such systems. Next, the unique process considerations of M3D fabrication and circuit technology are investigated to better allocate available resources for on-chip communication, and fabrication challenges and metal-resource contention are translated to achievable link frequencies and possible network topologies. Finally, a per-

formance investigation of these M3D Network-on-Chip topologies, utilizing the contested metal resources in the M3D stack, suggests that transistor-level M3D partitioning may be better than gate-level M3D partitioning in the context of on-chip communication.

In conclusion, this dissertation provides insight into the cost, power, and performance considerations across a range of emerging integration technologies, from interposer-based chiplet systems to monolithic 3D integration. The methodology and analysis developed in this work presents that case that multi-die and multi-tier systems are a viable, cost-effective option to increase transistor counts, but the correct architectural decisions must be made to manage manufacturing cost overheads while addressing the performance of on-die communication.

Bibliography

[1] G. E. Moore, *Cramming more components onto integrated circuits*, *Proceedings of the IEEE* **86** (Jan, 1998) 82–85.

[2] X. Dong and Y. Xie, *System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs)*, in *ASPDAC*, pp. 234–241, Jan, 2009.

[3] T. Song, W. Rim, J. Jung, *et. al.*, *13.2 a 14nm finfet 128mb 6t sram with vmin-enhancement techniques for low-power applications*, in *ISSCC*, pp. 232–233, Feb, 2014.

[4] S.-Y. Wu, C. Lin, M. Chiang, *et. al.*, *An enhanced 16nm cmos technology featuring 2nd generation finfet transistors and advanced cu/low-k interconnect for low power and high performance applications*, in *IEDM*, pp. 3.1.1–3.1.4, Dec, 2014.

[5] B. Landman and R. L. Russo, *On a pin versus block relationship for partitions of logic graphs*, *IEEE Transactions on Computers* **C-20** (Dec, 1971) 1469–1479.

[6] W. Donath, *Placement and average interconnection lengths of computer logic*, *IEEE Transactions on Circuits and Systems* **26** (Apr, 1979) 272–277.

[7] A. Kahng, S. Mantik, and D. Stroobandt, *Toward accurate models of achievable routing*, *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* **20** (May, 2001) 648–659.

[8] X. Dong, J. Zhao, and Y. Xie, *Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs*, *Computer-Aided Design of Integrated Circuits and Systems* **29** (Dec, 2010) 1959–1972.

[9] IC Knowledge LLC, *IC Cost and Price Model, 2016 Revision 05*, 2016.

[10] J. Cunningham, *The use and evaluation of yield models in integrated circuit manufacturing*, *IEEE Transactions on Semiconductor Manufacturing* **3** (May, 1990) 60–71.

[11] A. Dingwall, *High-yield-processed bipolar lsi arrays*, in *International Electron Devices Meeting*, vol. 14, pp. 82–82, 1968.

[12] J. P. Gambino, S. A. Adderly, and J. U. Knickerbocker, *An Overview of Through-silicon-via Technology and Manufacturing Challenges*, *Microelectron. Eng.* **135** (Mar., 2015) 73–106.

[13] P. Zarkesh-Ha, J. Davis, W. Loh, *et. al.*, *On a pin versus gate relationship for heterogeneous systems: Heterogeneous Rent's rule*, in *IEEE Custom Integrated Circuits Conference*, pp. 93–96, IEEE, 1998.

[14] Y. Chen, D. Niu, Y. Xie, and K. Chakrabarty, *Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis*, in *ICCAD*, pp. 471–476, Nov, 2010.

[15] M. Taouil, S. Hamdioui, E. Marinissen, and S. Bhawmik, *Using 3d-costar for 2.5d test cost optimization*, in *3D Systems Integration Conference (3DIC), 2013 IEEE International*, pp. 1–8, Oct, 2013.

[16] M. Harris, "Inside Pascal: Nvidia's newest computing platform." `https://devblogs.nvidia.com/parallelforall/inside-pascal`, June, 2016. Accessed: 2017-04-09.

[17] K. Low, "Samsung foundry's business strategy." `http://semiengineering.com/samsung-foundrys-business-strategy/`, April, 2016. Accessed: 2017-04-07.

[18] C. C. Lee, C. Hung, and C. C. et al., *An overview of the development of a GPU with integrated HBM on silicon interposer*, in *IEEE 66th Electronic Components and Technology Conference (ECTC)*, May, 2016.

[19] A. Kannan, N. E. Jerger, and G. H. Loh, *Enabling interposer-based disintegration of multi-core processors*, in *48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 546–558, Dec, 2015.

[20] M. Lapedus, "10nm versus 7nm." `http://semiengineering.com/10nm-versus-7nm/`, April, 2016. Accessed: 2017-02-20.

[21] G. Yeric, *Moore's law at 50: Are we planning for retirement?*, in *IEEE International Electron Devices Meeting (IEDM)*, Dec, 2015.

[22] X. Wu, G. Sun, X. Dong, R. Das, Y. Xie, C. Das, and J. Li, *Cost-driven 3d integration with interconnect layers*, in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, pp. 150–155, June, 2010.

[23] A. Kannan, N. E. Jerger, and G. H. Loh, *Exploiting interposer technologies to disintegrate and reintegrate multicore processors*, *IEEE Micro* **36** (May, 2016) 84–93.

[24] J. Zhao, Q. Zou, and Y. Xie, *Overview of 3d architecture design opportunities and techniques*, *Design Test, IEEE* **PP** (2015), no. 99 1–1.

[25] Y. Xie, J. Cong, and S. Sapatnekar, *Three-dimensional IC: Design, CAD, and Architecture*. Springer, 2009.

[26] K. Saban, *WhitePaper: Xilinx Stacked Silicon Interconnect Technologies*. Xilinx, 2012.

[27] B. Black, *Die-stacking is happening: AMD Fury X GPU*, in *Proceedings. of 12th Annual Conf. on 3D Architecture for Semiconductor Integration and Packaging*, Dec, 2015.

[28] S. Im and K. Banerjee, *Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs*, in *IEDM*, pp. 727–730, Dec, 2000.

[29] G. Loi, B. Agrawal, N. Srivastava, *et. al.*, *A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy*, in *DAC*, pp. 991–996, 2006.

[30] J. Galloway, S. Bhopte, and C. Nelson, *Characterizing junction-to-case thermal resistance and its impact on end-use applications*, in *ITherm*, pp. 1342–1347, May, 2012.

[31] C. Zhang and G. Sun, *Fabrication cost analysis for 2D, 2.5D, and 3D IC designs*, in *3DIC*, pp. 1–4, Jan, 2012.

[32] *Digikey*, 2015. Online. Available: http://www.digikey.com/.

[33] Anandtech, *Closed Loop AIO Liquid Coolers*, 2014. Online. Available: http://www.anandtech.com/show/7738/closed-loop-aio-liquid-coolers.

[34] E. Cooling, *Challenges in Measuring Theta jc for High Thermal Performance Packages*, 2014. Online. Available: http://www.electronics-cooling.com/2014/05/challenges-measuring-theta-jc-high-thermal-performance-packages/.

[35] M. Mantor and B. Sander, *AMD's Radeon next generation GPU Vega10*, in *HOTChips 2017*, August, 2017.

[36] *International technology roadmap for semiconductors 2.0, 2015 edition, system integration*, Report Ch 1, Semiconductor Industry Association, 2015.

[37] K. Chandrasekar, B. Akesson, and K. Goossens, *Improved power modeling of DDR SDRAMs*, in *DSD 2011*, August, 2011.

[38] *HIGH BANDWIDTH MEMORY (HBM) DRAM*, JEDEC Standards Document JESD235A, JEDEC, November, 2015. https://www.jedec.org/standards-documents/docs/jesd235a.

[39] *HIGH BANDWIDTH MEMORY (HBM) DRAM*, JEDEC Standards Document JESD235B, JEDEC, November, 2018. https://www.jedec.org/standards-documents/docs/jesd235b.

[40] T. Vogelsang, *Understanding the energy consumption of dynamic random access memories*, in *MICRO 2010*, December, 2010.

[41] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. B. Brockman, and N. P. Jouppi, *CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory*, in *DATE 2012*, March, 2012.

[42] M. Jung, D. M. Mathew, E. F. Zulian, C. Weis, and N. Wehn, *A new bank sensitive DRAMPower model for efficient design space exploration*, in *PATMOS 2016*, September, 2016.

[43] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, *The gem5 simulator*, *SIGARCH Comput. Archit. News* **39** (2011), no. 2 1–7.

[44] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, and G. Rodgers, *Achieving exascale capabilities through heterogeneous computing*, *IEEE Micro* **35** (July, 2015) 26–36.

[45] *International technology roadmap for semiconductors 2.0, 2015 edition, system integration*, Report Ch 1, Semiconductor Industry Association, 2015.

[46] J. A. Carballo, W. T. J. Chan, P. A. Gargini, A. B. Kahng, and S. Nath, *ITRS 2.0: Toward a re-framing of the semiconductor technology roadmap*, in *IEEE 32nd International Conference on Computer Design (ICCD)*, Oct, 2014.

[47] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, and G. Rodgers, *Achieving exascale capabilities through heterogeneous computing*, *IEEE Micro* **35** (July, 2015).

[48] M. Lapedus, "Uncertainty grows for 5nm, 3nm." `http://semiengineering.com/uncertainty-grows-for-5nm-3nm/`, Dec, 2016. Accessed: 2017-02-20.

[49] M. Lapedus, "Battling fab cycle times." `http://semiengineering.com/battling-fab-cycle-times/`, Feb, 2017. Accessed: 2017-02-20.

[50] S. Sutardja, *1.2 the future of IC design innovation*, in *IEEE International Solid-State Circuits Conference - (ISSCC)*, Feb, 2015.

[51] D. Stow *et. al.*, *Cost and Thermal Analysis of High-Performance 2.5D and 3D Integrated Circuit Design Space*, in *ISVLSI 2016*, pp. 637–642, July, 2016.

[52] N. Beck *et. al.*, *'zeppelin': An soc for multichip architectures*, in *ISSCC 2018*, pp. 40–42, Feb, 2018.

[53] A. Arunkumar *et. al.*, *Mcm-gpu: Multi-chip-module gpus for continued performance scalability*, in *ISCA 2017*, pp. 320–332, 2017.

[54] K. Saban, "Xilinx stacked silicon interconnect technology delivers breakthrough FPGA capacity, bandwidth and power efficiency." `https://www.xilinx.com/support/documentation/white_papers/wp380_Stacked_Silicon_Interconnect_Technology.pdf`, Dec, 2012. Accessed: 2017-02-20.

[55] P. Vivet, C. Bernard, F. Clermidy, D. Dutoit, E. Guthmuller, I. M. Panades, G. Pillonnet, Y. Thonnart, A. Garnier, D. Lattard, A. Jouve, F. Bana, T. Mourier, and S. Cheramy, *3D advanced integration technology for heterogeneous systems*, in *International 3D Systems Integration Conference (3DIC)*, Aug, 2015.

[56] D. Stow, I. Akgun, R. Barnes, P. Gu, and Y. Xie, *Cost and thermal analysis of high-performance 2.5D and 3D integrated circuit design space*, in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, July, 2016.

[57] A. Coskun *et. al.*, *A Cross-Layer Methodology for Design and Optimization of Networks in 2.5D Systems*, in *ICCAD 2018*, pp. 1–8, 2018.

[58] D. Stow *et. al.*, *Cost Analysis and Cost-driven IP Reuse Methodology for SoC Design Based on 2.5D/3D Integration*, in *ICCAD '16*, pp. 1–6, 2016.

[59] D. Stow *et. al.*, *Cost-effective Design of Scalable High-performance Systems Using Active and Passive Interposers*, in *ICCAD 2017*, pp. 728–735, 2017.

[60] J. H. Lau, *TSV manufacturing yield and hidden costs for 3D IC integration*, in *Proceedings 60th Electronic Components and Technology Conference (ECTC)*, June, 2010.

[61] N. E. Jerger, A. Kannan, Z. Li, and G. H. Loh, *NoC architectures for silicon interposer systems: Why pay for more wires when you can get them (from your interposer) for free?*, in *47th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Dec, 2014.

[62] N. Kim, D. Wu, D. Kim, A. Rahman, and P. Wu, *Interposer design optimization for high frequency signal transmission in passive and active interposer using through silicon via (TSV)*, in *IEEE 61st Electronic Components and Technology Conference (ECTC)*, May, 2011.

[63] J. Kim, *Active Si interposer for 3D IC integrations*, in *International 3D Systems Integration Conference (3DIC)*, Aug, 2015.

[64] D. Velenis, M. Detalle, and G. H. et al., *Processing active devices on Si interposer and impact on cost*, in *International 3D Systems Integration Conference (3DIC)*, Aug, 2015.

[65] P. Ehrett *et. al.*, *Analysis of microbump overheads for 2.5d disintegrated design*, report, University of Michigan Ann Arbor, 2017.

[66] S. Kumar *et. al.*, *A network on chip architecture and design methodology*, in *ISVLSI 2002*, pp. 117–124, 2002.

[67] A. Mandal *et. al.*, *An automated approach for minimum jitter buffered h-tree construction*, in *VLSI Design 2011*, pp. 76–81, Jan, 2011.

[68] A. Mandal *et. al.*, *A source-synchronous htree-based network-on-chip*, in *GLSVLSI 2013*, pp. 161–166, 2013.

[69] N. Pantano, C. R. Neve, and G. V. der Plas et al., *Technology optimization for high bandwidth density applications on 3D interposer*, in *6th Electronic System-Integration Technology Conference (ESTC)*, Sept, 2016.

[70] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, *McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures*, in *42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Dec, 2009.

[71] W. Zhao and Y. Cao, *Predictive technology model for nano-CMOS design exploration*, *J. Emerg. Technol. Comput. Syst.* (Apr., 2007).

[72] P. Christie and J. P. de Gyvez, *Prelayout interconnect yield prediction*, *IEEE Trans. Very Large Scale Integr. Syst.* **11** (Feb., 2003).

[73] J. A. Cunningham, *The use and evaluation of yield models in integrated circuit manufacturing*, *IEEE Trans. on Semiconductor Manufacturing* (May, 1990).

[74] L. Wang, S. Ma, and Z. Wang, *A high performance reliable NoC router*, in *21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan, 2016.

[75] J. Standard, *High Bandwidth Memory (HBM) DRAM*, *JESD235A* (2015).

[76] C. Sun *et. al.*, *Dsent - a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling*, in *NOCS 2012*, pp. 201–210, 2012.

[77] J. A. Cunningham, *The use and evaluation of yield models in integrated circuit manufacturing*, *IEEE Trans. Semicond. Manuf.* **3** (May, 1990) 60–71.

[78] Nan Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J. Kim, and W. J. Dally, *A Detailed and Flexible Cycle-Accurate Network-on-Chip Simulator*, in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 86–96, April, 2013.

[79] Y. Lee, D. Limbrick, and S. K. Lim, *Power benefit study for ultra-high density transistor-level monolithic 3d ics*, in *DAC '13*, May, 2013.

[80] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, and S. K. Lim, *How much cost reduction justifies the adoption of monolithic 3d ics at 7nm node?*, in *ICCAD '16*, Nov, 2016.

[81] K. Chang, D. Kadetotad, Y. Cao, J. Seo, and S. K. Lim, *Monolithic 3d ic designs for low-power deep neural networks targeting speech recognition*, in *ISLPED '17*, July, 2017.

[82] C. Liu and S. K. Lim, *Ultra-high density 3d sram cell designs for monolithic 3d integration*, in *IITC '12*, June, 2012.

[83] Y. Lee, P. Morrow, and S. K. Lim, *Ultra high density logic designs using transistor-level monolithic 3d integration*, in *ICCAD '12*, Nov, 2012.

[84] S. Panth, K. Samadi, Y. Du, and S. K. Lim, *Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **36** (Oct, 2017) 1716–1724.

[85] S. Panth, K. Samadi, Y. Du, and S. K. Lim, *Power-performance study of block-level monolithic 3d-ics considering inter-tier performance variations*, in *DAC '14*, June, 2014.

[86] W. Huangfu, S. Li, X. Hu, and Y. Xie, *Radar: A 3d-reram based dna alignment accelerator architecture*, in *DAC '18*, 2018.

[87] S. Panth, K. Samadi, Y. Du, and S. K. Lim, *Power-Performance Study of Block-Level Monolithic 3D-ICs Considering Inter-Tier Performance Variations*, in *51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, June, 2014.

[88] K. Chang, A. Koneru, K. Chakrabarty, and S. K. Lim, *Design automation and testing of monolithic 3d ics: Opportunities, challenges, and solutions*, in *ICCAD '17*, Nov, 2017.

[89] C. Liu and S. K. Lim, *A Design Tradeoff Study with Monolithic 3D Integration*, in *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, pp. 529–536, March, 2012.

[90] Y. Lee, D. Limbrick, and S. K. Lim, *Power Benefit Study for Ultra-High Density Transistor-Level Monolithic 3D ICs*, in *50th ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–10, May, 2013.

[91] C. Liu and S. K. Lim, *A design tradeoff study with monolithic 3d integration*, in *ISQED '12*, March, 2012.

[92] Y. Lee and S. K. Lim, *Ultrahigh Density Logic Designs Using Monolithic 3-D Integration*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **32** (Dec, 2013) 1892–1905.

[93] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, *Full chip impact study of power delivery network designs in monolithic 3d ics*, in *ICCAD '14*, Nov, 2014.

[94] S. K. Samal, S. Panth, K. Samadi, M. Saedi, Y. Du, and S. K. Lim, *Fast and accurate thermal modeling and optimization for monolithic 3d ics*, in *DAC '14*, June, 2014.

[95] W. Hwang, M. M. S. Aly, Y. H. Malviya, M. Gao, T. F. Wu, C. Kozyrakis, H.-S. P. Wong, and S. Mitra, *3d nanosystems enable embedded abundant-data computing: Special session paper*, in *CODES '17*, 2017.

[96] C. Li and P. Ampadu, *Energy-efficient noc with variable channel width*, in *MWSCAS '15*, Aug, 2015.

[97] V. F. Pavlidis and E. G. Friedman, *3-d topologies for networks-on-chip*, *IEEE TVLSI* **15** (Oct, 2007).

[98] J. Kimand C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das, *A novel dimensionally-decomposed router for on-chip communication in 3d architectures*, in *ISCA '07*, 2007.

[99] B. S. Feero and P. P. Pande, *Networks-on-chip in a three-dimensional environment: A performance evaluation*, *IEEE Tran. Computers* **58** (Jan, 2009).

[100] B. K. Joardar, K. Duraisamy, and P. P. Pande, *High performance collective communication-aware 3d network-on-chip architectures*, in *DATE '18*, March, 2018.

[101] T. Krishna, C. O. Chen, W. C. Kwon, and L. Peh, *Breaking the on-chip latency barrier using smart*, in *HPCA '13*, Feb, 2013.

[102] G. Michelogiannakis and J. Shalf, *Variable-width datapath for on-chip network static power reduction*, in *NoCS '14*, Sep., 2014.

[103] X. Jiang, X. Liu, L. Xu, P. Zhang, and N. Sun, *A reconfigurable accelerator for smithwaterman algorithm*, *IEEE Tran. Circuits and Systems* **54** (Dec, 2007).

[104] M. Badr and N. E. Jerger, *SynFull: Synthetic Traffic Models Capturing Cache Coherent Behaviour*, in *ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, pp. 109–120, June, 2014.