Invited: Efficient System Architecture in the Era of Monolithic 3D: Dynamic Inter-tier Interconnect and Processing-in-Memory

Dylan Stow, Itir Akgun,
Wenqin Huangfu, Yuan Xie
UC Santa Barbara
Santa Barbara, California
dstow@ucsb.edu

Xueqi Li
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

Gabriel H. Loh
Advanced Micro Devices, Inc.
Bellevue, Washington

ABSTRACT

Emerging Monolithic Three-Dimensional (M3D) integration technology will not only provide improved circuit density through the high-bandwidth coupling of multiple vertically-stacked layers, but it can also provide new architectural opportunities for on-chip computation, memory, and communication that are beyond the capabilities of existing process and packaging technologies. For example, with massive parallel communication between heterogeneous memory and compute layers, existing processing-in-memory architectures can be optimized and expanded, developing into efficient and flexible near-data processors. Additionally, multiple tiers of interconnect can be dynamically leveraged to provide an efficient, scalable interconnect fabric that spans the three-dimensional system. This work explores some of the challenges and opportunities presented by M3D technology for emerging computer architectures, with focus on improving efficiency and increasing system flexibility.

1 INTRODUCTION

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 [1, 10, 12, 13, 16] using novel 3D-aware design automation tools [18, 19].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DAC '19, June 2–6, 2019, Las Vegas, NV, USA

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6725-7/19/06...\$15.00

<https://doi.org/10.1145/3316781.3323475>

Figure 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).

In Section 2 the key results, insights, and challenges revealed by these prior works are summarized to provide guidance on M3D design for the architecture community.

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 work investigates the interaction between M3D process technology and the proper design of system-level architectural components. In particular, we first address the communication requirements of M3D systems through an analysis of Network-on-Chip (NoC) physical and logical topologies, and we propose a novel SMART M3D NoC design with dynamic inter-tier management of variable-width links for improved flexibility, efficiency, and performance. We also address 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 computation with little power overhead. By analyzing an NVM PIM accelerator [4], we identify existing bottlenecks that can be solved via the cooperation between NVM and CMOS M3D integration. An envisioned M3D system is depicted in Figure 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.

2 M3D BENEFITS AND CHALLENGES

Prior work has studied the technology and design methodologies necessary for M3D, often by leveraging existing 2D design automation tools [10, 19]. 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.

Block-level partitioning 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% [18], but coarse memory-on-logic partitioning achieved less than 2% power reduction [1, 19]. Footprint reduction may also be hindered by block imbalance when partitioning [18]. However, block-level partitioning may be useful for heterogeneous M3D process integration, such as Non-Volatile Memory, to provide connectivity between different technologies.

Gate-level partitioning 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%) [1, 2], especially when logic and memory can both be folded across tiers [1, 19]. Hundreds of thousands of MIVs can be used in gate-level partitioning, versus thousands in block-level partitioning.

Transistor-level partitioning increases circuit density by placing NMOS and PMOS transistors on separate tiers. Standard cells can be created with several sub-100nm MIVs in each gate, resulting in size reductions of 40% for logic gates [12] and 40-45% for SRAM cells [16], but ideal footprint reduction is not achieved due to NMOS/PMOS mismatch and MIV overhead [12]. Total power can be reduced versus 2D by 3-35% [12]. One drawback is increased route congestion from pin density, so additional local interconnect layers (at additional cost) may be necessary [12].

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 [15].

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 [11]. All partitioning strategies, including transistor-level partitioning, increase route congestion due to increased gate or pin density [11]. More metal layers may be necessary to achieve optimal power 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 [19]. M3D integration does improve the IR drop, skew, power, and wire overhead, for the clock and power network [19, 22], 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 [10]. Temperature requirements for sequential manufacturing may require slower tungsten interconnect on lower tiers or weaker transistors on upper tiers, but process and EDA solutions can minimize the negative impact [18].

Thermal Density: Perhaps the largest challenge with M3D integration is the sudden increase in thermal density. Even with thermal-aware placement and M3D power improvement, average temperature increases almost linearly with the number of tiers [21]. 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 [5] 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 [21].

3 NETWORK-ON-CHIP IN MONOLITHIC 3D

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 [14], 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 [3, 7, 8, 20]. These solutions, while appropriate for TSV-based 3D integration, are less suitable for the unique properties of M3D processes.

3.1 Monolithic 3D NoC

As summarized in Section 2, 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, we propose a NoC without a contention-causing 3D bus or vertical logical hops. Instead, routers

Figure 2: 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.

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.

3.2 SMART M3D NoC

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 [8]. We instead suggest 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 [9]. 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, with 28% latency reduction over 2D [7].

3.3 Bypass Links to Reduce Global Interconnect Utilization

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, we propose 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 2, bypass links reduce maximum hop count from six to three, enabling single-cycle SMART transmission.

3.4 Variable-Width Links in M3D NoC

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 [14]. 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% [14, 17], 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 2 summarizes the proposed logically-2D M3D NoC with inter-tier routers, global bypass links, and multi-tier variable-width neighbor links.

4 HETEROGENEOUS M3D FOR PROCESSING-IN-MEMORY

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, as we later demonstrate, contain communication bottlenecks that cannot be managed by standard memory busses. In this section, we will investigate an NVM PIM architecture and demonstrate how the high-bandwidth communication between heterogeneous M3D tiers can be used to improve PIM efficiency and flexibility.

As a case study, we investigate 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 [4]. 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

Figure 3: (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.

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 [6]; 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, connected 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.

5 CONCLUSION

Monolithic 3D integration can improve circuit density and reduce interconnect power, but system-level architectures must be optimized to improve power efficiency. This work proposes an M3D-targeted NoC architecture 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.

ACKNOWLEDGMENTS

AMD, the AMD Arrow logo, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. © 2019 Advanced Micro Devices, Inc. All rights reserved.

REFERENCES

- [1] K. Chang, D. Kadetotad, Y. Cao, J. Seo, and S. K. Lim. 2017. Monolithic 3D IC designs for low-power deep neural networks targeting speech recognition. In *ISLPED '17*.
- [2] K. Chang, A. Koneru, K. Chakrabarty, and S. K. Lim. 2017. Design automation and testing of monolithic 3D ICs: Opportunities, challenges, and solutions. In *ICCAD '17*.
- [3] B. S. Feero and P. P. Pande. 2009. Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation. *IEEE Tran. Computers* 58, 1 (Jan 2009).
- [4] W. Huangfu, S. Li, X. Hu, and Y. Xie. 2018. RADAR: A 3D-reRAM Based DNA Alignment Accelerator Architecture. In *DAC '18*.
- [5] W. Hwang, M. M. S. Aly, Y. H. Malviya, M. Gao, T. F. Wu, C. Kozyrakis, H.-S. P. Wong, and S. Mitra. 2017. 3D Nanosystems Enable Embedded Abundant-data Computing: Special Session Paper. In *CODES '17*.
- [6] X. Jiang, X. Liu, L. Xu, P. Zhang, and N. Sun. 2007. A Reconfigurable Accelerator for Smith-Waterman Algorithm. *IEEE Tran. Circuits and Systems* 54, 12 (Dec 2007).
- [7] B. K. Joardar, K. Duraisamy, and P. P. Pande. 2018. High performance collective communication-aware 3D Network-on-Chip architectures. In *DATE '18*.
- [8] J. Kim and C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das. 2007. A Novel Dimensionally-decomposed Router for On-chip Communication in 3D Architectures. In *ISCA '07*.
- [9] T. Krishna, C. O. Chen, W. C. Kwon, and L. Peh. 2013. Breaking the on-chip latency barrier using SMART. In *HPCA '13*.
- [10] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, and S. K. Lim. 2016. How much cost reduction justifies the adoption of monolithic 3D ICs at 7nm node?. In *ICCAD '16*.
- [11] Y. Lee and S. K. Lim. 2013. Ultrahigh Density Logic Designs Using Monolithic 3-D Integration. *IEEE TCAD* 32, 12 (Dec 2013).
- [12] Y. Lee, D. Limbrick, and S. K. Lim. 2013. Power benefit study for ultra-high density transistor-level monolithic 3D ICs. In *DAC '13*.
- [13] Y. Lee, P. Morrow, and S. K. Lim. 2012. Ultra high density logic designs using transistor-level monolithic 3D integration. In *ICCAD '12*.
- [14] C. Li and P. Ampadu. 2015. Energy-efficient NoC with variable channel width. In *MWSCAS '15*.
- [15] C. Liu and S. K. Lim. 2012. A design tradeoff study with monolithic 3D integration. In *ISQED '12*.
- [16] C. Liu and S. K. Lim. 2012. Ultra-high density 3D SRAM cell designs for monolithic 3D integration. In *IITC '12*.
- [17] G. Michelogiannakis and J. Shalf. 2014. Variable-width datapath for on-chip network static power reduction. In *NoCS '14*.
- [18] S. Panth, K. Samadi, Y. Du, and S. K. Lim. 2014. Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations. In *DAC '14*.
- [19] S. Panth, K. Samadi, Y. Du, and S. K. Lim. 2017. Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs. *IEEE TCAD* 36, 10 (Oct 2017).
- [20] V. F. Pavlidis and E. G. Friedman. 2007. 3-D Topologies for Networks-on-Chip. *IEEE TVLSI* 15, 10 (Oct 2007).
- [21] S. K. Samal, S. Panth, K. Samadi, M. Saedi, Y. Du, and S. K. Lim. 2014. Fast and accurate thermal modeling and optimization for monolithic 3D ICs. In *DAC '14*.
- [22] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim. 2014. Full chip impact study of power delivery network designs in monolithic 3D ICs. In *ICCAD '14*.