Processor Architecture Design Using 3D Integration Technology

Yuan Xie
Pennsylvania State University
Computer Science and Engineering Department
University Park, PA, 16802, USA
yuanxie@cse.psu.edu

Abstract

The emerging three-dimensional (3D) chip architectures, with their intrinsic capability of reducing the wire length, is one of the promising solutions to mitigate the interconnect problem in modern microprocessor designs. 3D memory stacking also enables much higher memory bandwidth for future chip-multiprocessor design, mitigating the "memory wall" problem. In addition, heterogenous integration enabled by 3D technology can also result in innovation designs for future microprocessors. This paper serves as a survey of various approaches to design future 3D microprocessors, leveraging the benefits of fast latency, higher bandwidth, and heterogeneous integration capability that are offered by 3D technology. [1]

1. Introduction

With continued technology scaling, interconnect has emerged as the dominant source of circuit delay and power consumption. The reduction of interconnect delays and power consumption are of paramount importance for deep-sub-micron designs. Three-dimensional integrated circuits (3D ICs) [3] are attractive options for overcoming the barriers in interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology.

3D integration technologies offer many benefits for future microprocessor designs. Such benefits include: (1) *The reduction in interconnect wire length*, which results in improved performance and reduced power consumption; (2)*Improved memory bandwidth*, by stacking memory on microprocessor cores with TSV connections between the memory layer and the core layer; (3) *The support for realization of heterogeneous integration*, which could result in novel architecture designs. (4)*Smaller form factor*, which results in higher packing density and smaller footprint due to the addition of a third dimension to the conventional two dimensional layout, and potentially results in a lower cost design.

This tutorial paper first presents the background on 3D integration technology, and then reviews various approaches to design future 3D microprocessors, which leverage the benefits of fast latency, higher bandwidth, and heterogeneous integration capability that are offered by 3D technology. The challenges for future 3D architecture design are also discussed in the last section.

2. 3D Integration Technology

The 3D integration technologies [25, 26] can be classified into one of the two following categories. (1) *Monolithic approach*. This approach involves sequential device process. The frontend processing (to build the device layer) is repeated on a single wafer to build multiple active device layers before the backend processing builds interconnects among devices. (2) *Stacking approach*, which could be further categorized as wafer-to-wafer, die-to-wafer, or die-to-die stacking methods. This approach processes each layer separately, using conventional fabrication techniques. These multiple layers are then assembled to build up 3D IC, using bonding technology. Since the stacking approach does not require the change of conventional fabrication process, it is much more practical compared to the monolithic approach, and become the focus of recent 3D integration research.

Several 3D stacking technologies have been explored recently, including wire bonded, microbump, contactless (capacitive or inductive), and *through-silicon vias (TSV)* vertical interconnects [3]. Among all these integration approaches, TSV-based 3D integration has the potential to offer the greatest vertical interconnect density, and therefore is the most promising one among all the vertical interconnect technologies. Figure 1 shows a conceptual 2-layer 3D integrated circuit with TSV and microbump.

[1] This work is supported in part by grants from NSF (0702617, 0903432) and SRC.

3D stacking can be carried out using two main techniques [8]: (1) *Face-to-Face (F2F)* bonding: two wafers(dies) are stacked so that the very top metal layers are connected. Note that the die-to-die interconnects in face-to-face wafer bonding does not go through a thick buried Silicon layer and can be fabricated as *microbump*. The connections to C4 I/O pads are formed as TSVs; (2) *Face-to-Back (F2B)* bonding: multiple device layers are stacked together with the top metal layer of one die is bond together with the substrate of the other die, and direct vertical interconnects (which are called *through-silicon vias (TSV)*) tunneling through the substrate. In such F2B bonding, TSVs are used for both between-layer-connections and I/O connections. Figure 1 shows a conceptual 2-layer 3D IC with F2F or F2B bonding, with both TSV connections and microbump connections between layers.

All TSV-based 3D stacking approaches share the following three common process steps [8]: (i) *TSV formation*; (ii) *Wafer thinning* and (iii) *Aligned wafer or die bonding*, which could be wafer-to-wafer(W2W) bonding or die-to-wafer(D2W) bonding. Wafer thinning is used to reduce the impact of TSVs. The thinner the wafer, the smaller (and shorter) the TSV is (with the same aspect ratio constraint) [8]. The wafer thickness could be in the range of 10 μm to 100 μm and the TSV size is in the range of 0.2 μm to 10 μm [3].

In TSV-based 3D stacking bonding, the dimension of the TSVs is not expected to scale at the same rate as feature size because alignment tolerance during bonding poses limitation on the scaling of the vias. The TSV size, length, and the pitch density, as well as the bonding method (face-to-face or face-to-back bonding, SOI-based 3D or bulk CMOS-based 3D), can have a significant impact on the 3D microprocessor design. For example, relatively large size of TSVs can hinder partitioning a design at fine granularity across multiple device layers, and make the true 3D component design less possible. On the other hand, The monolithic 3D integration provides more flexibility in vertical 3D connection because the vertical 3D via can potentially scale down with feature size due to the use of local wires for connection. Availability of such technologies makes it possible to partition the design at a very fine granularity. Furthermore, face-to-face bonding or SOI-based 3D integration may have a smaller via pitch size and higher via density than face-to-back bonding or bulk -CMOS-based integration. Such influence of the 3D technology parameters on the microprocessor design must be thoroughly studied before an appropriate partition strategy is adopted.

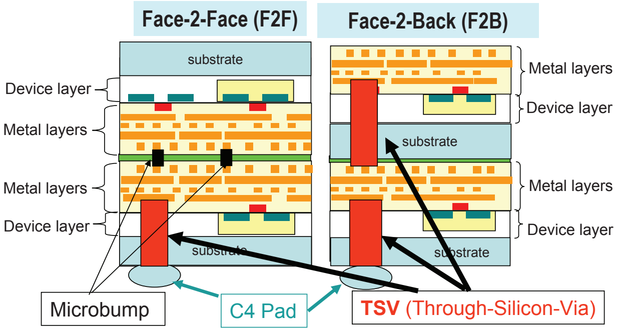

Figure 1: Illustration of F2F and F2B 3D bonding.

3. Designing 3D Processor Architecture

The following subsections will discuss various architecture design approaches that leverage different benefits that 3D integration technology can offer, namely, wire length reduction, high memory bandwidth, heterogeneous integration, and cost reduction. It will also briefly review 3D network-on-chip architecture designs.

3.1 Wire Length Reduction

Designers have resorted to technology scaling to improve microprocessor performance. Although the size and switching speed of transistors benefit as technology feature sizes continue to shrink, global interconnect wire delay does not scale accordingly with technologies. The increasing wire delays have become one major impediment for performance improvement.

Three-dimensional integrated circuits (3D ICs) are attractive options for overcoming the barriers in interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology. Compared to a traditional two dimensional chip design, one of the important benefits of a 3D chip over a traditional two-dimensional (2D) design is the reduction on global interconnects. It has been shown that three-dimensional architectures reduce wiring length by a factor of the square root of the number of layers used [9]. The reduction of wire length due to 3D integration can result in two obvious benefits: *latency improvement* and *power reduction*.

Latency Improvement. Latency improvement can be achieved due to the reduction of average interconnect length and the critical path length.

Early work on fine-granularity 3D partitioning of processor components shows that the latency of a 3D components could be reduced. For example, since interconnects dominate the delay of cache accesses which determines the critical path of a microprocessor, and the regular structure and long wires in a cache make it one of the best candidates for 3D designs, 3D cache design is one of the early design example for fine-granularity 3D partition [26]. Wordline partitioning and bitline partitioning approaches divide a cache bank into multiple layers and reduce the global interconnects, resulting in a fast cache access time. Depending on the design constraints, the 3DCacti tool [21] automatically explores the design space for a cache design, and finds out the optimal partitioning strategy, and the latency reduction can be as much as 25% for a two-layer 3D cache. 3D arithmetic-component designs also show latency benefits. For example, various designs [7, 16, 18, 22] have shown that the 3D arithmetic unit design can achieve around 6%-30% delay reduction due to the wire length reduction. Such fine-granularity 3D partitioning was also demonstrated by Intel [1], showing that by targeting the heavily pipelined wires, the pipeline modifications resulted in approximately 15% improved performance, when the Intel Pentium-4 processor was folded onto 2-layer 3D implementation.

Note that such fine-granularity design of 3D processor components increases the design complexity, and the latency improvement varies depending on the partitioning strategies and the underlying 3D process technologies. For example, for the same Kogge-Stone adder design, a partitioning based on logic level [22] demonstrates that the delay improvement diminishes as the number of 3D layers increases; while a bit-slicing partitioning [16] strategy would have better scalability as the bit-width or the number of layers increases. Furthermore, the delay improvement for such bit-slicing 3D arithmetic units is about 6% when using a bulk-CMOS-based 180nm 3D process [7], while the improvement could be as much as 20% when using a SOI-based 180nm 3D process technology [16], because the SOI-based process has much smaller and shorter TSVs (and therefore much smaller TSV delay) compared to the bulk-CMOS-based process.

Power Reduction. Interconnect power consumption becomes a large portion of the total power consumption as technology scales. The reduction of the wire length translates into the power saving in 3D IC design. For example, 7% to 46% of power reduction for 3D arithmetic units were demonstrated in [16]. In the 3D Intel Pentium-4 implementation [1], because of the reduction in long global interconnects, the number of repeaters and repeating latches in the implementation is reduced by 50%, and the 3D clock network has 50% less metal RC than the 2D design, resulting in a better skew, jitter and lower power. Such 3D stacked redesign of Intel Pentium 4 processor improves performance by 15% and reduces power by 15% with a temperature increase of 14 degrees. After using voltage scaling to lower the leak temperature to be the same as the baseline 2D design, their 3D Pentium 4 processor still showed a performance improvement of 8%.

3.2 Memory Bandwidth Improvement

It has been shown that circuit limitations and limited instruction level parallelism will diminish the benefits of modern superscalar microprocessors by increased architectural complexity, which leads to the advent of Chip Multiprocessors (CMP) as a viable alternative to the complex superscalar architecture. The integration of multi-core or many-core microarchitecture on a single die is expected to accentuate the already daunting memory-bandwidth problem. Supplying enough data to a chip with a massive number of on-die cores will become a major challenge for performance scalability. Traditional off-chip memory will not suffice due to the I/O pin limitations. Three-dimensional integration has been envisioned as a solution for future micro-architecture design (especially for multi-core and many-core architectures), to mitigate the interconnect crisis and the “memory wall” problem [14, 15, 17]. It is anticipated that memory stacking on top of logic would be one of the early commercial uses of 3D technology for future chip-multiprocessor design, by providing improved memory bandwidth for such multi-core/many-core microprocessors. In addition, such approaches of memory stacking on top of core layers do not have the design complexity problem as demonstrated by the fine-granularity design approaches, which require re-designing all processor components for wire length reduction (as discussed in 3.2).

Intel [1] explored the memory bandwidth benefits using a base-line Intel Core2 Duo processor, which contains two cores. By having memory stacking, the on-die cache capacity is increased, and the performance is improved by capturing larger working sets, reducing off-chip memory bandwidth requirements. For example, one option is to stack an additional 8MB L2 cache on top of the base-line 2D processor (which contains 4MB L2 cache), and the other option is to replace the SRAM L2 cache with a denser DRAM L2 cache stacking. Their study demonstrated that a 32MB 3D stacked DRAM cache can reduce the cycles per memory access by 13% on average and as much as 55% with negligible temperature increases.

PicoServer project [10] follows a similar approach to stack DRAM on top of multi-core processors. Instead of using stacked memory as a larger L2 cache (as shown by Intel’s work [1]), the fast on-chip 3D stacked

DRAM main memory enables wide low-latency buses to the processor cores and eliminates the need for an L2 cache, whose silicon area is allocated to accommodate more cores. Increasing the number of cores by removing the L2 cache can help improve the computation throughput, while each core can run at a much lower frequency, and therefore result in an energy-efficient many core design. For example, it can achieve a 14% performance improvement and 55% power reduction over a baseline multi-core architecture.

As the number of the cores on a single die increases, such memory stacking becomes more important to provide enough memory bandwidth for processor cores. Recently, Intel [19] demonstrated an 80-tile terascale chip with network-on-chip. Each core has a local 256KB SRAM memory (for data and instruction storage) stacked on top of it. TSVs provide a bandwidth of 12GB/second for each core, with a total about 1TB/second bandwidth for Tera Flop computation. In this chip, the thin memory die is put on top of the CPU die, and the power and I/O signals go through memory to CPU.

Since DRAM is stacked on top of the processor cores, the memory organization should also be optimized to fully take advantages of the benefits that TSVs offer [13, 15]. For example, the numbers of ranks and memory controllers are increased, in order to leverage the memory bandwidth benefits. A multiple-entry row buffer cache is implemented to further improve the performance of the 3D main memory. Comprehensive evaluation shows that a 1.75x speedup over commodity DRAM organization is achieved [15]. In addition, the design of MSHR was explored to provided a scalable L2 miss handling before accessing the 3D stacked main memory. A data structure called the Vector Bloom Filter with dynamic MSHR capacity tuning is proposed. Such structure provides an additional 17.8% performance improvement. If stacked DRAM is used as the last-level caches (LLC) in chip multiple processors (CMPs), the DRAM cache sets are organized into multiple queues [13]. A replacement policy is proposed for the queue-based cache to provide performance isolation between cores and reduce the lifetimes of dead cache lines. Approaches are also proposed to dynamically adapt the queue size and the policy of advancing data between queues.

The latency improvement due to 3D technology can also be demonstrated by such memory stacking design. For example, Li et al. [12] proposed a 3D chip multiprocessor design using network-in-memory topology. In this design, instead of partitioning each processor core or memory bank into multiple layers (as shown in [21, 26]), each core or cache bank remains to be a 2D design. Communication among cores or cache banks are via the network-on-chip(NoC) topology. The core layer and the L2 cache layer are connected with TSV-based bus. Because the short distance between layers, TSVs provide a fast access from one layer to another layer, and effectively reduce the cache access time because of the faster access to cache banks through TSVs.

3.3 Heterogenous Integration

3D integration also provides new opportunities for future architecture design, with a new dimension of design space exploration. In particular, the heterogenous integration capability enabled by 3D integration gives designers new perspective when designing future CMPs.

3D integration technologies provide feasible and cost-effective approaches for integrating architectures composed of heterogeneous technologies to realize future microprocessors targeted at the ”More than Moore” technology projected by ITRS. 3D integration supports heterogeneous stacking because different types of components can be fabricated separately, and layers can be implemented with different technologies. It is also possible to stack optical device layers or non-volatile memories (such as magnetic RAM (MRAM) or phase-change memory (PCRAM)) on top of microprocessors to enable cost-effective heterogeneous integration. The addition of new stacking layers composed of new device technology will provide greater flexibility in meeting the often conflicting design constraints (such as performance, cost, power, and reliability), and enable innovative designs in future microprocessors.

Non-volatile Memory Stacking. Stacking layers of non-volatile memory technologies such as Magnetic Random Access Memory (MRAM) [4] and Phase Change Random Access Memory (PRAM) [24] on top of processors can enable a new generation of processor architectures with unique features. There are several characteristics of MRAM and PRAM architectures that make them as promising candidates for on-chip memory. In addition to their non-volatility, they have zero standby power, low access power and are immune to radiation-induced soft errors. However, integrating these non-volatile memories along with a logic core involves additional fabrication challenges that need to be overcome (for example, MRAM process requires growing a magnetic stack between metal layers). Consequently, it may incur extra cost and additional fabrication complexity to integrate MRAM with conventional CMOS logic into a single 2D chip. The ability to integrate two differ-

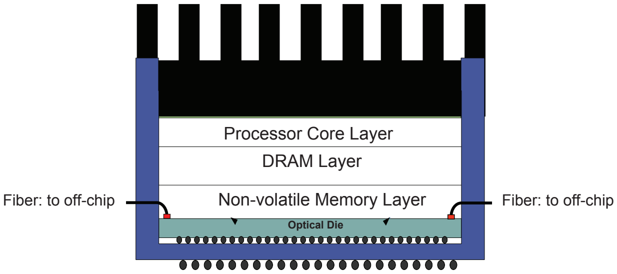

Figure 2: An illustration of 3D heterogeneous architecture with non-volatile memory stacking and optical die stacking.

ent wafers developed with different technologies using 3D stacking offers an ideal solution to overcome this fabrication challenge and exploit the benefits of PRAM and MRAM technologies. For example, Sun et al. [20] demonstrated that the optimized MRAM L2 cache on top of multi-core processor can improve performance by 4.91% and reduce power by 73.5% compared to the conventional SRAM L2 cache with the similar area.

Optical Device Layer Stacking. Even though 3D memory stacking can help mitigate the memory bandwidth problem, when it comes to off-chip communication, the pin limitations, the energy cost of electrical signaling, and the non-scalability of chip-length global wires are still significant bandwidth impediments. Recent developments in silicon nanophotonic technology have the potential to meet the off-chip communication bandwidth requirements at acceptable power levels. With the heterogeneous integration capability that 3D technology offers, one can integrate optical die together with CMOS processor dies. For example, HP Labs proposed a Corona architecture [23], which is a 3D many-core architecture that uses nanophotonic communication for both inter-core communication and off-stack communication to memory or I/O devices. A photonic crossbar fully interconnects its 256 low-power multithreaded cores at 20 terabyte per second bandwidth, with much lower power consumption.

Figure 2 illustrates such a 3D heterogenous processor architecture, which integrates non-volatile memories and optical die together through 3D integration technology.

3.4 Cost-effective Architecture

Increasing integration density has resulted in large die size for microprocessors. With a constant defect density, a larger die typically has a lower yield. Consequently, partitioning a large 2D microprocessor to be multiple smaller dies and stacking them together may result in a much higher yield for the chip, even though 3D stacking incurs extra manufacture cost due to extra steps for 3D integration and may cause a yield loss during stacking. Depending on the original 2D microprocessor die size, it may be cost-effective to implement the chip using 3D stacking [5], especially for large microprocessors. The heterogenous integration capability that 3D provides can also help reduce the cost.

In addition, as technology feature size scales to reach the physics limits, it has been predicted that moving to the next technology node is not only difficult but also prohibitively expensive. 3D stacking can potentially provide a cost-effective integration solution, compared to traditional technology scaling.

3.5 3D NoC Architecture

Network-on-chip (NoC) is a general purpose on-chip interconnection network architecture that is propsoed to replace the traditional design-specific global on-chip wiring, by using switching fabrics or routers to connect processor cores or processing elements (PEs). Typically, the PEs communicate with each other using a packet-switched protocol. Even though both 3D integrated circuits and NoCs are proposed as alternatives for the interconnect scaling demands, the challenges of combining both approaches to design three-dimensional NOCs have not been addressed until recently [6, 11, 12]. Researchers have studied various NoC router design with 3D integration technology. For example, various design options the NoC router for 3D NoC has been investigated: 1) symmetric NoC router design with a simple extension to the 2D NoC router; 2) NoC-bus hybrid router design which leverage the inherent asymmetry in the delays in a 3D architecture between the fast vertical interconnects and the horizontal interconnects that connect neighboring cores; 3) True 3D router design with major modification as dimensionally-decomposed router [11]; 4) Multi-layer 3D NoC router design which partitions a single router to multiple layers to boost the performance and reduce the power consumption [6]. 3D NoC topology design was also investigated [27]. More details can be found in [2].

4. Challenges for 3D Architecture Design

Even though 3D integrated circuits show great benefits, there are several challenges for the adoption of 3D technology for future architecture design: (1) *Thermal management.* The move from 2D to 3D design could accentuate the thermal concerns due to the increased power density. To mitigate the thermal impact, thermal-aware design techniques must be adopted for 3D architecture design [26]; (2) *Design Tools and methdologies.* 3D integration technology will not be commercially viable without the support of EDA tools and methodologies that allow architects and circuit designers to develop new architectures or circuits using this technology. To efficiently exploit the benefits of 3D technologies, design tools and methodologies to support 3D designs are imperative [25]; (3) *Testing*. One of the barriers to 3D technology adoption is insufficient understanding of 3D testing issues and the lack of design-for-testability (DFT) techniques for 3D ICs, which have remained largely unexplored in the research community.

References

[1] B. Black et al. Die stacking 3D microarchitecture. In *MICRO*, pages 469–479, 2006.

[2] L. Carloni, P. Pande, and Y. Xie. Networks-on-chip in emerging intercoonect paradigms: Advantages and challenges. In *Intl. Symp. on Networks-on-chips*, 2009.

[3] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon. Demystifying 3D ICs: the pros and cons of going vertical. *IEEE Design and Test of Computers*, 22(6):498– 510, 2005.

[4] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, and Y. Chen. Circuit and microarchitecture evaluation of 3D stacking Magnetic RAM (MRAM) as a universal memory replacement. In *Design Automation Conference*, pages 554–559, 2008.

[5] X. Dong and Y. Xie. Cost analysis and system-level design exploration for 3D ICs. In *Asia and South Pacific Design Automation Conference*, 2009.

[6] P. Dongkook, S. Eachempati, R. Das, A. K. Mishra, Y. Xie, N. Vijaykrishnan, and C. R. Das. MIRA: A multi-layered on-chip interconnect router architecture. In *International Symposium on Computer Architecture (ISCA)*, pages 251–261, 2008.

[7] R. Egawa, J. Tada, H. Kobayashi, and G. Goto. Evaluation of fine grain 3D integrated arithmetic units. In *IEEE International 3D System Integration Conference*, 2009.

[8] P. Garrou. *Handbook of 3D Integration: Technology and Applications using 3D Integrated Circuits*, chapter Introduction to 3D Integration. Wiley-CVH, 2008.

[9] J. Joyner, P. Zarkesh-Ha, and J. Meindl. A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3D-SoC). In *Proc. 14th Annual IEEE International ASIC/SOC Conference*, Sept. 2001.

[10] T. Kgil, S. D'Souza, A. Saidi, N. Binkert, R. Dreslinski, T. Mudge, S. Reinhardt, and K. Flautner. PicoServer: using 3D stacking technology to enable a compact energy efficient chip multiprocessor. In *ASPLOS*, pages 117–128, 2006.

[11] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, N. Vijaykrishnan, and C. Das. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In *Proceedings of the Annual International Symposium on Computer Architecture*, 2007.

[12] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, N. Vijaykrishnan, and M. Kandemir. Design and management of 3D chip multiprocessors using network-in-memory. In *International Symposium on Computer Architecture (ISCA'06)*, 2006.

[13] G. Loh. Extending the effectiveness of 3D-stacked dram caches with an adaptive multi-queue policy. In *International Symposium on Microarchitecture (MICRO)*, Dec. 2009.

[14] G. Loh, Y. Xie, and B. Black. Processor design in three-dimensional die-stacking technologies. *IEEE Micro*, 27(3):31–48, 2007.

[15] G. H. Loh. 3d-stacked memory architectures for multi-core processors. In *International Symposium on Computer Architecture (ISCA)*, pages 453–464, 2008.

[16] J. Ouyang, G. Sun, Y. Chen, L. Duan, T. Zhang, Y. Xie, and M. Irwin. Arithmetic unit design using 180nm TSV-based 3D stacking technology. In *IEEE International 3D System Integration Conference*, 2009.

[17] P. Jacob et al. Mitigating memory wall effects in high clock rate and multi-core cmos 3D ICs: Processor memory stacks. *Proceedings of IEEE*, 96(10), 2008.

[18] K. Puttaswamy and G. H. Loh. Scalability of 3d-integrated arithmetic units in high-performance microprocessors. In *Design Automation Conference*, pages 622–625, 2007.

[19] S. Vangal et al. An 80-tile Sub-100-W TeraFLOPS processor in 65-nm CMOS. *IEEE Journal of Solid-State Circuits*, 43(1):29–41, 2008.

[20] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen. A novel 3D stacked MRAM cache architecture for CMPs. In *International Symposium on High Performance Computer Architecture*, 2009.

[21] Y.-f. Tsai, F. Wang, Y. Xie, N. Vijaykrishnan, and M. J. Irwin. Design space exploration for three- dimensional cache. *IEEE Transactions on Very Large Scale Integration Systems*, 2008.

[22] B. Vaidyanathan, W.-L. Hung, F. Wang, Y. Xie, V. Narayanan, and M. J. Irwin. Architecting microprocessor components in 3D design space. In *Intl. Conf. on VLSI Design*, pages 103–108, 2007.

[23] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn. Corona: System implications of emerging nanophotonic technology. In *Proceedings of the 35th International Symposium on Computer Architecture*, pages 153–164, 2008.

[24] X. Wu, J. Li, L. Zhang, E. Speight, and Y. Xie. Hybrid cache architecture. In *International Symposium on Computer Architecture (ISCA)*, 2009.

[25] Y. Xie, J. Cong, and S. Sapatnekar. *Three-Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures*. Springer, 2009.

[26] Y. Xie, G. Loh, B. Black, and K. Bernstein. Design space exploration for 3D architectures. *ACM Journal of Emerging Technologies in Computing Systems*, 2006.

[27] Y. Xu et al. A low-radix and low-diameter 3D interconnection network design. In *Intl. Symp. on High Performance Computer Architecture*, 2009.