Overview of 3-D Architecture Design Opportunities and Techniques

Jishen Zhao
University of California Santa Cruz

Yuan Xie
University of California Santa Barbara

Qiaosha Zou
Huawei Technologies Co. Ltd.

Editor's note:
Three-dimensional (3-D) integration, a breakthrough technology to achieve "More Moore and More Than Moore," provides numerous benefits, e.g., higher performance, lower power consumption, and higher bandwidth, by utilizing vertical interconnects and die/wafer stacking. This paper presents an overview of 3-D integration along with various design challenges and recent innovations.

—Partha Pande, Washington State University

THREE-DIMENSIONAL INTEGRATION may refer to various technologies that share a common salient feature: they integrate two or more dies of active devices in a single circuit. Compared to traditional 2-D integration technologies, 3-D integration replaces long, global wires (with massive repeaters) with short, vertical or horizontal interconnects. Looking further out, 3-D integration introduces a few benign effects that are likely to innovate computer architecture design in a nontraditional manner.

Digital Object Identifier 10.1109/MDAT.2015.2463282
Date of publication: 31 July 2015; date of current version: 30 June 2017.

First, the reduced interconnect wire length and the declined number of global signals can directly reduce the circuit delay and system power dissipation. A variety of previous studies [1] have demonstrated that the reduced wire length can lead to up to 30% delay reduction in 3-D arithmetic unit design. In addition, the reduction in wire length, repeaters, and repeating latches can be translated into power reduction due to the reduced parasitics with the shorter wire length and incorporating previously off-chip signals to be on-chip. For example, Ouyang et al. [1] demonstrated 3-D-stacked arithmetic units with up to 46% power saving compared with 2-D designs.

Second, 3-D integration enables high-bandwidth memory (HBM) due to the wide bus width between the processor and the integrated memory. It releases the constraint of off-chip pin counts, and therefore can dramatically increase memory bandwidth by an order of magnitude. For example, Microns hybrid memory cube (HMC) [2] can offer up to 160-GB/s bandwidth with a 2-GB capacity.

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

Today, memory bandwidth is a fundamental performance bottleneck. Modern applications typically adopt large memory working set and increasingly larger number of threads which can simultaneously access the memory. By offering high memory bandwidth, 3-D integration can substantially change the way memory hierarchy is designed to support high-performance and energy-efficient data movement.

Third, 3-D integration allows different circuit layers to be implemented by different and incompatible process technologies, enabling heterogeneous integration. It enables feasible and cost-effective architecture designs composed of heterogeneous technologies to achieve the More than More technology projected by International Technology Roadmap for Semiconductors (ITRS) in future microprocessors. For example, 3-D integration allows optical devices to be integrated closely with digital circuits with new architecture designs [3]. Three-dimensional integration also allows novel computer architectures to be developed such as hybrid processor cache and memory hierarchy design. Various memory technologies, such as SRAM, spin-transfer torque RAM (STT-RAM), phase-change memory (PCM), and resistive RAM (ReRAM) trade off between performance, power, density, implementation complexity, and cost. A hybrid memory hierarchy incorporated with various memory technologies can be optimized for all metrics [4]. Three-dimensional integration is a critical technology that enables such hybrid memory hierarchy designs.

Fourth, 3-D integration enables much condensed form factor compared to traditional 2-D integration technologies. Due to the addition of a third dimension to conventional 2-D layout, it leads to a higher packing density and smaller footprint. This potentially leads to processor designs with lower cost.

As such, both academic and industry communities abound with research studies that examine the rich architectural space enabled by 3-D integration, since its debut decades ago. From the academia prospective, comprehensive studies have been performed across all aspects of microprocessor architecture design by employing 3-D integration technologies, such as 3-D-stacked processor core and cache architectures, 3-D integrated memory, 3-D network-on-chip (NoC). Furthermore, a large body of research studied critical issues and opportunities raised by adopting 3-D integration technologies, such as thermal issue which is imposed by dense integration of active electronic devices, the cost issues which is incurred by extra process and increased die area, and the opportunity in designing cost-effective microprocessor architectures. From the industry prospective, 3-D integrated memory is envisioned to become pervasive in the near future. Intel's Xeon Phi processors are delivered with 3-D stacked DRAMs [5].

In this article, we overview recent novel computer architecture designs enabled by 3-D integration technologies. We will start from introducing the basis of various 3-D integration technologies. We will then review recent 3-D integrated computer architecture designs, including 3-D-stacked multicore processor design, integrated HBM and hybrid memory design, and 3-D NoC design. We will also discuss recent studies that investigate thermal and cost issued concerned with 3-D integrated architectures.

Three-dimensional integration technologies

Based on fabrication processes and media of vertical interconnects, 3-D integration technologies can be roughly categorized as through-silicon-via (TSV)-based, monolithic, and contactless 3-D integration.

TSV-based 3-D integration (Figure 1a) adopts several active device layers, which can be fabricated in parallel. These layers are then vertically stacked together to form a single integrated circuit (IC). Typically, TSVs are directly formed in each upper tier before the stacking for vertical interconnect. Therefore, performance of TSVs largely affects the yield of 3-D ICs. Steps of 3-D fabrication are different from those of 2-D counterparts mainly due to the formation of TSVs. In general, four additional steps are necessary for 3-D integration, namely etching, thinning, alignment, and bonding [6]. In terms of the bonding orientation, we can classify the bonding into face-to-back (F2B) and face-to-face (F2F) bonding methods. F2F bonding enables higher interconnect density via CuCu bonding on the top metal layer and results in no silicon area overhead for interlayer signals. In contrast, due to the alignment precision and the interconnect size, the interconnect density in F2B is relatively smaller. Three different stacking strategies can be achieved: wafer-to-wafer (W2W), die-to-wafer (D2W), and die-to-die (D2D) stacking.

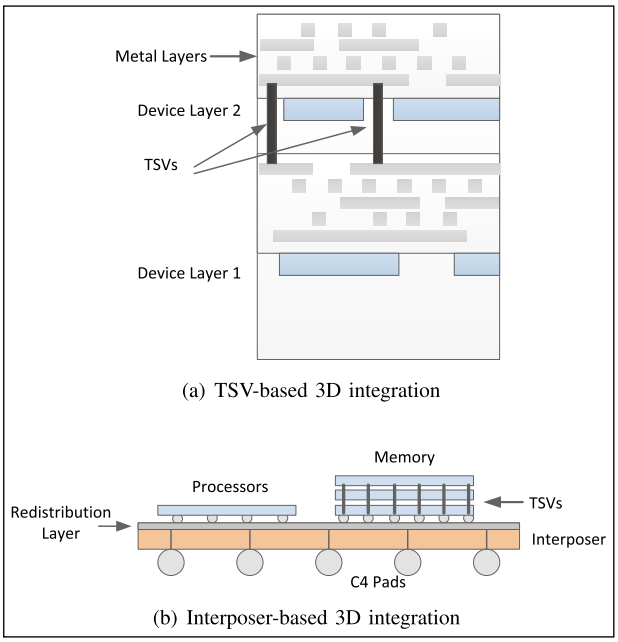

Figure 1. A conceptual view of parallel 3-D integration.

The major difference among these three is whether to slice the wafer into dice before stacking. For W2W stacking, wafers are bonded before sliced into dice. This process has the highest productivity yet the lowest compound yield. Testings are required for D2W and D2D stacking and only known-good dies (KGDs) are selected for bonding. Therefore, the compound yield is increased by preventing bad dies from stacking on good ones.

To reduce the manufacturing complexity and fabrication cost, recent 3-D IC designs employ an alternative integration technique, interposer-based integration techniques, which is also known as 2.5-D integrated circuits (Figure 1b). Interposer is a connection media containing only interconnect components (TSVs, BEOL, etc.). With 2.5-D integration, dies incorporated with active devices are integrated horizontally on the silicon interpose and connected by signals going through TSVs.

Monolithic 3-D integration [7] (Figure 2) successively grows device layers on a conventional complementary metaloxidesemiconductor (CMOS) or silicon-on-insulator (SOI) plane. Devices on the bottom layer are fabricated with traditional fabrication processes. The fabrication technique for upper layers is challenging because of two extra requirements: first, we need to guarantee device electrical characteristics in the upper layers after fabrication, such as field mobility and threshold voltage; second, building upper planes cannot compromise the electrical performance of the bottom layer. By employing sequential fabrication process, monolithic 3-D ICs allow the pitch of vertical interconnects smaller than TSVs. Therefore, monolithic 3-D integration technology enables finer grained stacking at gate level or even transistor level.

Contactless 3-D integration employs coupling of electric or magnetic fields to perform the interlayer communication with a comparable data rate as TSVs. Based on the coupling principle, contactless 3-D integration can be further categorized into capacitively and inductively coupled 3-D integration [8]. The major benefit of contactless 3-D integration is that the fabrication process is compatible with 2-D processes. However, the inductors and capacitors can dissipate certain amount of power. For example, although the inductive coupling can consume only 0.14 pJ/b, the energy consumption of the inductors is continuous, wasting significant amount of power [9], [10].

Comparison of 3-D integration technologies: Compared to TSV-based 3-D integration, monolithic integration can lead to up to 8% smaller area due to smaller size of intertier via than the TSVs, 12% lower longest path delay, and 7% lower power [7]. Nevertheless, implementing monolithic 3-D ICs requires nontrivial changes to conventional fabrication process, making it less practical compared to the TSV-based approach. One of major issues with contactless transmissions is the significant metal area consumption introduced by coils and metal plates, especially when high transmission

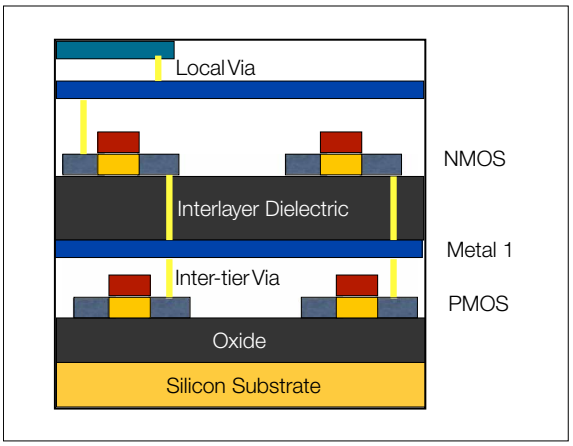

Figure 2. A conceptual sketch of sequential/monolithic 3-D integration.

intensity is required in inductively coupled designs. For example, the coil diameter is 60 μm for one signal in a 25-μm communication distance [8]. Due to the design infeasibility, fabrication complexity, and cost issues, monolithic and contactless 3-D integration is less commonly found in existing 3-D architecture designs. Most previous designs employ TSV-based 3-D integration technology.

Three-dimensional architecture design

Three-dimensional integration introduces a third dimension to circuit design, opening doors to numerous of novel computer architecture solutions toward high-performance, cost-effective, and energy-efficient computer designs. For decades, computer architects have invested substantial effort in exploring this new design space enabled by the emerging 3-D integration technologies. In this section, we review a few representative works among these research efforts.

Three-dimensional-stacked multicore processors

Three-dimensional integration technologies offer computer architects with opportunities in designing novel multicore processor architectures, with various processor function units being separated and stacked in multiple layers [11]. Compared to traditional 2-D designs, such 3-D-stacked multicore processors promise shorter interconnect delay between function units, smaller form factor, and reduced interconnect power dissipation. For example, Zhao et al. [11] explored the design space of 3-D many-core processors by partitioning processor cores and caches in two different manners: homogeneous partitioning tiles each processor core and L2 cache bank together and each layer incorporates several such tiles; heterogeneous partitioning places processor cores and the L2 cache in different layers. Their study investigated the fabrication cost and thermal effect of the two design strategies. Recent studies also explored the optimal 3-D-stacked processor caches.

High-bandwidth memories

The prominent memory wall problem limits the further performance improvement of multicore processors. Three-dimensional stacked memory with computing logic in one die has been envisioned as one of the early commercial adoption of 3-D integration. In addition to offering HBM, such 3-D integration also eliminates the redesign complexity, which is required by fine-grained partitioning involved with 3-D-integrated processor designs. The 3-D stacked memory can be implemented with two typical approaches: TSV-based 3-D stacking and interposer-based 2.5-D integration. With 3-D stacked memory, building memory layers directly on top of logic layers requires close collaboration between microprocessor vendors and memory vendors. Complexity of thermal management can pose challenges on the heat sink design, layer allocation, and floorplanning. In contrast, employing interposer-based 2.5-D integration can alleviate such design burden and decouple the memory design from more complex processor designs. Such design strategy has been demonstrated by Microns hybrid memory cube (HMC) [2] and JEDEC HBM standard [12]. HMC [2] combines high-speed logic process technology with a stack of TSV bonded memory die. It offers as much as 400 GB/s of bandwidth, while consumes 70% less energy per bit compared with state-of-the-art DRAM technologies. HBM [12] initially targets at graphic applications. Recently, AMD and SK Hynix have announced joint development of HBM stacks, which leverages TSV and wide I/O technology and conforms to JEDEC HBM standardization.

Besides traditional DRAM-based memories, 3-D integration provides another promising design opportunity in integrating the emerging nonvolatile memory (NVM) with conventional processor and memory architectures. Most NVM technologies are incompatible with the CMOS technology. They can introduce additional process steps besides the traditional CMOS process. Implementing a computer system with NVM can incur substantially increased fabrication cost. Three-dimensional integration allows heterogeneous technologies to be integrated in a single circuit. Consequently, it eases the adoption of NVMs in CMOS-based microprocessors. With separate fabrication process, NVMs and computing logics can be optimized accordingly to enable high flexibility, cost-effective, and high-performance architecture [13].

3-D/2.5-D-Integrated DRAMs as cache and memory

With 3-D integration, it is possible to integrate a large-capacity DRAM (several gigabytes) with the processor. Several recent studies explored the design

space of employing the integrated DRAM either as the last-level cache (LLC) [14] or part of the main memory [15]. The two design options trade off performance, storage overhead, and design complexity; no one-size-fits-all solution exists. As such, a natural extension is to dynamically switch the integrated DRAM between the roles of LLC and main memory. Chou et al. [16] proposed an on-chip memory design, which is managed as an LLC yet exposed to the OS as a portion of the physical address space. To this end, their design exposes the 3-D-stacked DRAM to the OS, but manages it at the cache line granularity. When a miss happens at the stacked DRAM, the design swaps the data fetched from the off-chip main memory with an existing memory line in the stacked DRAM. In this manner, the proposed design sustains the latency and the bandwidth benefits of the stacked DRAM.

3-D/2.5-D graphics processing units

GPUs require extensive memory bandwidth to support the data access of massively parallel executing threads. With the benefit of providing HBM, 3-D integrated memory is a promising candidate for graphics memory. This promising design opportunity has been explored by a couple of recent studies on energy-efficient GPU design with reconfigurable in-package graphics memory [17], [18]. Shortly after the academic studies were published, two top GPU vendors are both considering to introduce 3-D integrated graphics memory into their future products. NVIDIA adopts 3-D integrated memory in their new GPU products [19]. AMD starts to ship HBM with their GPU products in 2015 [20].

Three-dimensional NoCs

With more and more cores being integrated in a single system, the intercore connection with buses cannot fulfill communication requirements between cores. NoC is therefore emerging to tackle the communication scalability problem. Here, we consider moving NoC above cores instead of designs of router and routing algorithm to adapt the vertical connections.

Three-dimensional interconnect service layer (ISL): In 2-D designs, NoC designs are constraint by the limited routing resource and silicon area. This limitation can be relaxed with 3-D integration by building a single layer for interconnects, which decouples the interconnect component from computing and storage units. This separated interconnect layer is defined as interconnect service layer (ISL) by Wu et al. [21]. The benefits of building separate layers for NoCs are as follows: the manufacturing cost can be reduced due to the smaller die area after moving NoCs from computing layers; more reliable and flexible interconnect designs can be applied without the silicon area limitation; multiple network topologies can be supported in a single chip for different connection requirements. In their work, the core performance of using a hybrid mesh topology with coarse-grained and fine-grained meshes is improved compared to the baseline 2-D design.

Three-dimensional optical NoC: Optical NoCs are emerging to meet the continuously increased communication bandwidth problem. However, fabricating optical components on the core layer increases fabrication cost and the performance of optical components might be affected by the high core temperature. Therefore, the 3-D integration enables the heterogeneous integration of optical dies and CMOS processor dies. For example, the Corona architecture proposed by HP Labs leverages the 3-D stacking to build a nanophotonic NoC in multicore system for both intercore and off-chip communications [3]. Even though the optical NoC provides low power and high bandwidth, it faces the challenges of the thermal sensitivity and process variations, which may result in performance degradation or even malfunction if they are not properly addressed.

Three-dimensional architecture challenges

Due to technology limitations and the absence of mature design methodologies, designing 3-D architectures can be a challenging task. Among these challenges, we describe the two most critical ones in this section.

Design and fabrication cost

One major concern about adopting 3-D architectures is the cost of design and fabricating such chips: Will 3-D integrated processors and memories incur higher cost than their 2-D counterparts? A couple of recent studies comprehensively analyzed the cost issue and show that 3-D integrated architectures can in fact cost less than 2-D designs by adopting proper design strategies.

In correspondence to the fabrication process, the 3-D cost can be divided into five parts: wafer cost, bonding cost, testing cost, package cost, and cooling cost. The wafer cost captures the silicon and device formation cost of each separate die. Different from 2-D designs, the introduction of TSVs results in area overhead, which in turn affects both the silicon area and die yield. The bonding cost estimates the cost during TSV forming, wafer thinning, and bonding. The TSV forming cost varies with different TSV fabrication processes and stages. In several recent studies [11], [22], the etching and the TSV last approach are used, because the separation between die fabrication and TSV forming enables the decoupling between the wafer cost and the bonding cost. The extra testing cost in 3-D ICs can be captured in the model proposed by Chen et al. [23]. The study shows that D2W integration has higher testing cost. However, when combining with the fabrication cost, it is still more favorable because of the high stacking yield.

The package cost is determined by the package type, the package area, and the pin count. From their study [11], [22], the pin count is identified as the dominant factor in 3-D integration. In addition, the cooling cost is expected to be higher in 3-D integration because the increasing on-chip temperature needs more powerful cooling solution.

These additional costs from 3-D integration make the 3-D enabling point critical for making the design decision. Three-dimensional integration is cost efficient only when the cost saving from reduced silicon area outweighs the bonding, testing, and cooling cost. In Dongs study [22], the authors find that 3-D designs are cost efficient only when the design exceeds 50 million gates per chip with 65-nm technology node. With advanced processing technology, the enabling point moves toward a larger number of gates. Moreover, when the number of gates per chip increases, stacking more layers is considered to have cost saving.

The cost issue in 3-D ICs attracts plenty research efforts in performing test cost optimization, yield enhancement, and TSV reduction [24].

Thermal issues

Thermal is another challenging issue in 3-D architecture design due to two aspects. The first reason lies in the increasing power density of 3-D stacking. A small form factor and a large number of stacking layers deteriorate the thermal dissipation. The second reason is the lack of sufficient thermal dissipation path from devices to the ambient or the heat sink, leading to the elevated on-chip temperature. The high on-chip temperature causes many performance and reliability issues, such as thermal runaway and accelerated electromigration. Therefore, it is significantly important to perform 3-D thermal modeling/analyses and deploy efficient thermal management schemes.

Thermal modeling. Two 3-D thermal models are commonly used in thermal analyses. The first model is using partial differential equation to represent the on-chip temperature from [25]. The following equation can be used to calculate the transient thermal response with Poissons equation as steady-state and additional boundary conditions. Two methods can be applied to solve this equation: the finite difference method (FDM) and the finite element method (FEM):

$$\rho c_p \frac{\partial \mathrm{T}(r,t)}{\partial t} = k_t \nabla^2 \mathrm{T}(r,t) + g(r,t). \qquad (1)$$

The FDM and FEM computation overhead is large, which is inappropriate for thermal analysis in design space exploration. Therefore, the second compact thermal model is proposed as building the equivalent thermal circuit. This model is widely used in thermal analyses. The devices are viewed as thermal resistance and capacitance. The value of resistivity represents the capability of thermal dissipation. The higher the value is, the harder the thermal dissipation can be. Heat flows are modeled as currents while the temperature differences are modeled as voltages. The thermal capacitance captures the delay before the temperature reaching the steady state. The dominant factor to determine the circuit temperature is the heat conduction to the thermal package and to nearby circuits.

Thermal management schemes. Generally, thermal management can be performed in either design time or runtime. In design time, methodologies on thermal-aware physical level design (floorplanning, place and route, and power planning) have been extensively explored [26]. Thermal via insertion is another attractive solution [27]. Dummy vias are inserted above hot spots to help vertical thermal dissipation by leveraging the high thermal conductivity of conductive materials. In addition to aforementioned methods, powerful liquid cooling methods are attractive for 3-D designs [28]. During runtime, numerous 2-D thermal management

schemes can be applied on 3-D designs, such as dynamic voltage and frequency scaling (DVFS) and power gated.

Thermal-aware architecture design. Three-dimensional processors can employ thermal herding techniques [29] to herd or steer the switching activity to the die that is closest to the heat sink and reduce the total power (and power density). Thermal herding is proposed based on the observation that only a few of the least significant bits are needed in the data used in many integer instructions. Therefore, the datapath can be organized by assigning each 16 bits to a separate dies (four dies) in 3-D stacking; we can place the least significant bits on the top layer which is closest to the heat sink. Beyond addressing thermal issues, 3-D thermal herding techniques can also improve system performance by reducing the pipeline depth and L2 cache latency; power is also reduced due to the reduced wire length and switching activities (because of clock gating). The thermal experiments show that thermal herding techniques successfully control power density and mitigate 3-D thermal issues.

Testing challenges

Compared to traditional 2-D IC fabrication, 3-D IC fabrication incorporates more intermediate steps, such as die stacking and TSV bonding. Because of these extra steps, 3-D IC manufacturing typically performs wafer test before final assembly and packaging. Wafer test for 3-D ICs has challenges. First, existing probe technology cannot perform finer pitch and dimensions of TSV tips, hence it is limited to handling only several hundred probes at a much lower number than required TSV probes. Second, wafer test can require to create a KGD stack, which can risk damaging due to the contact of the highly thinned wafer by a wafer probe. Third, 3-D ICs can also impose intradie defects caused by thermal issues and new manufacturing steps, such as wafer thinning and bonding the top of a TSV to another wafer. Recently, Wang et al. [30] proposed a built-in self-test methodology to test TSVs in 3-D ICs prior to stacking. They also developed efficient architecture and circuit design to perform the prebond TSV scan testing. While 3-D IC testing is a crucial problem, testing-aware 3-D architecture design has remained largely unexplored in the research community and require substantial efforts in addressing the testing challenges.

Three-dimensional placement challenges

Based on the block position decided by floorplanning, the cell/gate position is determined during placement. Different from 2-D designs, the 3-D placement not only needs to consider the cell spreading on the xy plane, but also in the vertical direction. If temperature is the primary concern, then cell placements on the upper layers should be sparser than the bottom layer, which is near heat sink. Temperature and wirelength are two major design objectives during placement. Various placement methodologies are presented by previous studies. For example, Athikulwongse et al. [26] proposed two thermal-aware global placement algorithms employing the force-directed methodology to exploit the die-to-die thermal coupling in 3-D ICs. The first algorithm generates forces for TSV spreading and alignment for better heat dissipation path. The second algorithm builds forces based on thermal conductivity in cells and on power density for TSVs. Their second methods can achieve the best temperature results among state-of-the-art placers.

THREE-DIMENSIONAL INTEGRATION is one of the most promising solutions for future computing system. In this survey article, we review recent innovations in 3-D architecture design. Furthermore, we also analyzed the two major challenges in 3-D integrated processor and memory designs, including cost and thermal issues. Although the community has gained initial success on 3-D integration, yet substantial work is needed to further improvements on 3-D architecture designs and the exploration of killer applications is still needed. ■

References

[1] J. Ouyang et al., "Arithmetic unit design using 180nm TSV-based 3D stacking technology," presented at the Int. 3D Syst. Integr. Conf., 2009.

[2] "Micron hybrid memory cube," [Online]. Available: http://www. hybridmemorycube.org/

[3] D. Vantrease et al., Corona: System implications of emerging nanophotonic technology, presented at the Int. Symp. Comput. Architect., 2008.

[4] J. Zhao, C. Xu, and Y. Xie, "Bandwidth-aware reconfigurable cache design with hybrid memory technologies," in *Proc. Int. Conf. Comput.-Aided Design*, 2011, pp. 48–55.

[5] Intel, "An intro to MCDRAM (high bandwidth memory) on Knights Landing." [Online]. Available: https://software.intel.com/en-us/blogs/2016/01/20/an-intro-to-mcdram-high-bandwidth-memory-on-knights-landing

[6] V. Pavlidis and E. Friedman, *Three-Dimensional Integrated Circuit Design*, V. Pavlidis and E. Friedman, Eds. New York, NY, USA: Morgan Kaufmann, 2009.

[7] L. Chang and S. K. Lim, "A design tradeoff study with monolithic 3D integration," in *Proc. Int. Symp. Low Power Electron. Design*, 2012, pp. 16.

[8] N. Miura, K. Kasuga, M. Saito, and T. Kuroda, "An 8Tb/s 1pJ/b 0.8mm2/Tb/s QDR inductive-coupling interface between 65nm CMOS GPU and 0.1 μm DRAM," presented at the Int. Solid-State Circuits Conf., 2010.

[9] H. Zhan, H. Matsutani, M. Koibuchi, and H. Amano, "Dynamic power consumption optimization for inductive-coupling based wireless 3d nocs," *IPSJ Trans. Syst. LSI Design Methodol.*, vol. 7, pp. 27–36, Feb. 2014.

[10] T. Kagami, H. Matsutani, M. Koibuchi, Y. Take, T. Kuroda, and H. Amano, "Efficient 3-D bus architectures for inductive-coupling thruchip interfaces," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 2, pp. 493–506, 2016.

[11] J. Zhao, X. Dong, and Y. Xie, "Cost-aware three-dimensional (3D) many-core multiprocessor design," in *Proc. Design Autom. Conf.*, 2010.

[12] *Jedechbm*, [Online]. Available: http://www.jedec.org/category/technology-focus-area/3d-ics-0

[13] G. Sun, E. Kursun, J. A. Rivers, and Y. Xie, "Exploring the vulnerability of CMPs to soft errors with 3D stacked nonvolatile memory," *J. Emerging Technol. Comput. Syst.*, vol. 9, no. 3, pp. 122, 2013.

[14] G. H. Loh and M. D. Hill, "Supporting very large DRAM caches with compound-access scheduling and MissMap," *IEEE Micro*, vol. 32, no. 3, pp. 70–78, May 2012.

[15] X. Dong, Y. Xie, N. Muralimanohar, and N. P. Jouppi, "Simple but effective heterogeneous main memory with on-chip memory controller support," in *Proc. Int. Conf. High Performance Comput.*, 2010, pp. 1–11.

[16] C. Chou, A. Jaleel, and M. K. Qureshi, "CAMEO: A two-level memory organization with capacity of main memory and flexibility of hardware-managed cache," in *Proc. Int. Symp. Microarchitect.*, 2014, DOI: 10.1109/MICRO.2014.63.

[17] J. Zhao, G. Sun, G. H. Loh, and Y. Xie, "Energy-efficient GPU design with reconfigurable in-package graphics memory," in *Proc. Int. Symp. Low Power Electron. Design*, 2012, pp. 403–408.

[18] J. Zhao, G. Sun, G. Loh, and Y. Xie, "Optimizing GPU energy efficiency with 3D die-stacking graphics memory and reconfigurable memory interface," *ACM Trans. Architect. Code Optim.*, vol. 10, no. 4, pp. 24:124:25, 2013.

[19] NVIDIA, "The most advanced data center GPU ever built." [Online]. Available: http://www.nvidia.com/object/tesla-p100.html

[20] AMD, "Radeon™ R9 Series gaming graphics cards with high-bandwidth memory." [Online]. Available: http://www.amd.com/en-us/products/graphics/desktop/r9

[21] X. Wu et al., "Cost-driven 3D integration with interconnect layers," in *Proc. Design Autom. Conf.*, 2010, pp. 150–155.

[22] X. Dong, J. Zhao, and Y. Xie, "Fabrication cost analysis and cost-aware design space exploration for 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 29, pp. 1959–1972, 2010.

[23] Y. Chen, D. Niu Y. Xie, and K. Chakrabarty, "Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis," in *Proc. Int. Conf. Comput.-Aided Design*, 2010, pp. 471–476.

[24] Y. Zhao, S. Khursheed, and B. Al-Hashimi, "Cost-effective TSV grouping for yield improvement of 3D-ICs," in *Proc. Asian Test Symp.*, 2011, pp. 201–206.

[25] S. Sapatnekar, "Addressing thermal and power delivery bottlenecks in 3D circuits," in *Proc. Asia South Pacific Design Autom. Conf.*, 2009, pp. 423–428.

[26] K. Athikulwongse, M. Ekpanyapong, and S.K. Lim, "Exploiting die-to-die thermal coupling in 3-D IC placement," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, pp. 2145–2155, 2014.

[27] C.-C. Wen, Y.-J. Chen, and S.-J. Ruan, "Cluster-based thermal-aware 3D-floorplanning technique with post-floorplan TTSV insertion at via-channels," in *Proc. Asia Symp. Quality Electron. Design*, 2013, pp. 200–207.

[28] D. Kearney, T. Hilt, and P. Pham, "A liquid cooling solution for temperature redistribution in 3D IC architectures," *Microelectron. J.*, vol. 43, pp. 602–610, 2012.

[29] X. Zhou, Y. Xu, Y. Du, Y. Zhang, and J. Yang, "Thermal management for 3D processors via task scheduling," in *Proc. Parallel Process. Int. Conf.*, 2008, pp. 115–122.

[30] C. Wang et al., "BIST methodology, architecture and circuits for pre-bond TSV testing in 3D stacking IC systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 1, pp. 139–148, 2015.

Jishen Zhao is currently an Assistant Professor at the University of California at Santa Cruz, Santa Cruz, CA, USA. Her research interests are computer architecture and electronic design automation, with an emphasis on emerging technologies and high-performance computing. Zhao has a PhD from Pennsylvania State University, State College, PA, USA. She is a member of the IEEE and the Association for Computing Machinery (ACM).

Qiaosha Zou is currently a researcher at Huawei Technologies Co. Ltd., Hangzhou, China. Zou has a PhD from Pennsylvania State University, State College, PA, USA. She is a member of the IEEE and the Association for Computing Machinery (ACM).

Yuan Xie is currently a Professor at the University of California Santa Barbara, Santa Barbara, CA, USA. His research interests include 3-D ICs, memory architecture, and reliable system designs. Xie has a PhD from Princeton University, Princeton, NJ, USA. He is a Fellow of the IEEE for his contributions in design automation and architecture for 3-D ICs.

■ Direct questions and comments about this article to Jishen Zhao, University of California Santa Cruz, Santa Cruz, CA 95064 USA; e-mail: jishen.zhao@ucsc.edu.