TSV-Based 3-D ICs: Design Methods and Tools

Tiantao Lu, *Student Member, IEEE*, Caleb Serafy, *Student Member, IEEE*, Zhiyuan Yang, Sandeep Kumar Samal, *Student Member, IEEE*, Sung Kyu Lim, *Senior Member, IEEE*, and Ankur Srivastava, *Senior Member, IEEE*

***Abstract*—Vertically integrated circuits (3-D ICs) may revitalize Moore's law scaling which has slowed down in recent years. 3-D stacking is an emerging technology that stacks multiple dies vertically to achieve higher transistor density independent of device scaling. They provide high-density vertical interconnects, which can reduce interconnect power and delay. Moreover, 3-D ICs can integrate disparate circuit technologies into a single chip, thereby unlocking new system-on-chip architectures that do not exist in 2-D technology. While 3-D integration could bring new architectural opportunities and significant performance enhancement, new thermal, power delivery, signal integrity and reliability challenges emerge as power consumption grows, and device density increases. Moreover, the significant expansion of CPU design space in 3-D requires new architectural models and methodologies for design space exploration (DSE). New design tools and methods are required to address these 3-D-specific challenges. This keynote paper focuses on the state of the art, ongoing advances and future challenges of 3-D IC design tools and methods. The primary focus of this paper is TSV-based 3-D ICs, although we also discuss recent advances in monolithic 3-D ICs. The objective of this paper is to provide a unified perspective on the fundamental opportunities and challenges posed by 3-D ICs especially from the context of design tools and methods. We also discuss the methodology of co-design to address more complicated and interdependent design problems in 3-D IC, and conclude with a discussion of the remaining challenges and open problems that must be overcome to make 3-D IC technology commercially viable.**

***Index Terms*—3-D integrated circuit (IC), architecture, design tools, physical design.**

I. Introduction

CMOS technology has approached a critical junction where traditional device and interconnect scaling are having trouble keeping up with Moore's law [1]. The underlying reason is that engineers are facing several dilemmas in advanced technology nodes. The first dilemma comes as chip feature size approaches the lower limits of current photolithography technology [2]. Investment in next-generation lithography solutions is possible but costly. The second dilemma is the ever-increasing leakage current. For example, thin gate oxide results in substantial gate tunneling leakage and also subthreshold leakage [3]. Employment of metal gates and high-k dielectrics [4] is an effective approach to control the leakage current, however, its compatibility with CMOS process has raised some concerns [5], [6]. The third dilemma is the increasing dominance of wire delay and power in future technology nodes [7], [8]. Limitations to DRAM bus bandwidth and speed are inherent to off-chip integration, and lead to a severe "memory wall" problem in the era of big data [9], [10]. Furthermore the power dissipated by such a coarse integration paradigm leads to high dynamic power dissipation and inhibits scaling toward envisioned exascale computing systems of the future [11].

Vertical integration provides a promising solution to reduce interconnect power and delay while increasing transistor density independent of costly device scaling [12]. Chips fabricated by existing technology can be directly bonded together to increase the number of transistors per unit area. This configuration avoids investment of new generation of devices, and enables heterogeneous technology integration on chip. Vertical integration between layers is established by through-silicon-vias (TSVs). TSVs are essentially metal pillars that penetrate the silicon substrate to engage the metal pads of the layer below. This extra connectivity in the third dimension substantially reduces the wire delay and power, and also provides high-density interconnect for data transfer between layers [13].

An illustration of a 3-D integrated circuit (IC) is shown in Fig. 1. Heterogeneous technologies such as CPUs, memories, radio frequency circuits, analog circuits and sensors can be freely integrated in the same package to avoid long and slow off-chip wires. Typically a 3-D IC is connected to PCB through C4 bumps for power delivery. A heat spreader and a heat sink are on the other side of the chip.

In summary, the expanded design space brought on by vertical integration technology reduces global interconnect length, delay, and power, and provides the potential for large improvements to latency and bandwidth. However, 3-D integration also brings many new challenges. New fabrication processes for TSV etching and wafer bonding need to be developed [14], new architectures must be developed to make use of high density interconnects [9], [10], [15]–[17], and new design methodologies must be developed to model the fundamental physical changes vertical integration brings [18]–[21].

Manuscript received March 20, 2016; revised August 14, 2016 and December 5, 2016; accepted January 19, 2017. Date of publication February 9, 2017; date of current version September 14, 2017. This work was supported in part by the National Science Foundation under Grant 0917057 and Grant CCF1302375, and in part by the DARPA IceCool Project. This paper was recommended by Associate Editor L. Behjat.

T. Lu, C. Serafy, Z. Yang, and A. Srivastava are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: ttlu@umd.edu; cserafy1@umd.edu; zyyang@umd.edu; ankurs@umd.edu).

S. K. Samal and S. K. Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sandeep.samal@gatech.edu; limsk@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2017.2666604

0278-0070 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

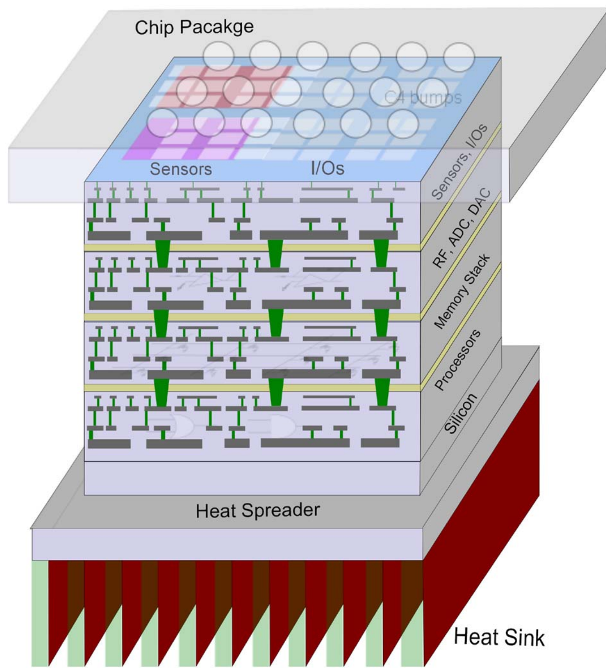


Fig. 1. 3-D IC provides a solution for heterogeneous integration. TSV establishes high-speed high-density vertical data transfer path, which is critical for modern computing and storage paradigms.

This keynote paper focuses on design method and tool challenges and state of the art solutions for TSV-based 3-D ICs, and also summarizes the current status of 3-D IC fabrication, and new architectures supported by 3-D integration. We also provide discussion about monolithic 3-D (M3-D) ICs, which recently emerges as another appealing 3-D solutions besides TSV-based 3-D ICs.

We provide overview for two primary 3-D fabrication steps in this paper: wafer stacking and TSV manufacturing. Precise alignment and bonding techniques are necessary to enable high-quality wafer stacking [22]–[24]. TSV manufacturing is another key process, and both laser drilling and plasma etching have been used to form TSVs with precise aspect ratio [25]. Details about wafer stacking as well as TSV manufacturing are discussed in Section II.

We arrange the design tool challenges in historical order in Section III. The earliest research on 3-D IC design tools were concerning its physical design. Two branches of ideas have been explored. The first idea tries to reuse 2-D physical design tools as much as possible [26], [27]. This requires certain partitioning algorithms to be applied to a flattened circuit netlist, and the gates are therefore distributed onto different 3-D layers. Conventional 2-D placer and router can then be used on each layer independently. The second idea is to develop native 3-D placer and router [18], [28]–[31]. These 3-D physical design tools need to incorporate the 3-D awareness and exploit the extra vertical connectivity to achieve better performance and lower power dissipation. Starting from early 2000, chip temperature has become an important design constraint. 3-D stacking increases power flux linear to the number of layers. Conventional air cooling is not enough to remove all the heat, and many different design approaches and novel active cooling schemes have be developed [17], [32], [33]. Meanwhile, the decrease in VDD as technology advances leads to tighter constraints on PDN impedance which are increasingly difficult to meet using small scale interconnect technology [34]–[36]. Better power delivery network (PDN) design is necessary to overcome this issue. Most recently, reliability and signal integrity metrics have added another level of design complexity. These metrics include TSV electromigration [37]–[40], thermal mechanical stress [20], [41], [42], and electrical coupling [43]–[46], and the interaction between packaging-domain and chip-domain stress [47], [48].

We also provide an in-depth discussion about advanced architectures that can potentially be unlocked by 3-D integration. We investigate the application of 3-D memories to solve the so called memory wall problem [9], [17]. 3-D memories provide a high-density, high-bandwidth, and energy-efficient solution that is favorable for many modern computing paradigms that process huge amounts of data [11], [15], [49]. We discuss the memory-on-logic design, which has also drawn attention for its low-latency and high-bandwidth communication between CPU cores and memory arrays [9], [10], [17]. Our architectural level discussion is presented in Section V.

Specifically, we arrange the 3-D IC design challenges as follows.

1) *3-D Partitioning:* Partitioning assigns placement cells to specific layers before placement and routing [26], [27]. This enables reusing 2-D physical design tools to design each layer independently. However, the quality of final 3-D IC design depends heavily on how the circuit is partitioned. Studies [50]–[54] have shown that partitioning methodologies and granularity significantly affect TSV usage, power, performance, area, and reliability (PPAR) metrics.
2) *3-D Placement:* New design constraints such as TSV placement need to be considered in 3-D ICs. TSV usage and placement significantly affect area, wire length, stress, and thermal profile [18], [28]–[31]. This 3-D problem has a high magnitude of complexity, which calls for effective algorithms and heuristics.
3) *Clock Distribution Network (CDN):* 3-D CDN design includes significant constraints on the number and position of clock TSVs [55]–[57]. Traditional clock network designs needs to be extended to incorporate TSVs' RC characteristics to satisfy performance metrics such as clock skew/slew and power.
4) *Thermally Aware Design:* The thermal problem in modern circuit is to such an extent that parts of the chip have to be shutdown to avoid thermal violations (generally called "dark silicon" [58]). This is especially severe in 3-D ICs as vertical stacking increases both power flux and thermal resistance [17], [59]. It has been more difficult for traditional air cooling schemes to remove heat inside 3-D ICs, [17], [33], [60], therefore developing active cooling schemes and corresponding design methodologies are essential.
5) *PDN:* In 3-D ICs, power is distributed layer-by-layer, from the package pins on the bottom layer to circuit components throughout the stacks. Unfortunately

significant power is lost during vertical transportation [35]. 3-D IC power delivery faces a fundamental mismatch between planar power supply and volumetric power demand [61]. Innovative 3-D power delivery solutions must provide high-quality voltage levels and deal with voltage emergencies.

6) *Signal Integrity:* Due to their size, TSVs have a large parasitic capacitance, and can easily couple to transistors, planar interconnects, and one another. TSV placement, shielding and signaling paradigms are significant contributors to TSV signal integrity [43]–[46].
7) *TSV Reliability:* TSVs introduce new failure modes such as TSV electromigration [37]–[40], thermal mechanical stress [20], [41], [42], and stress migration [62], [63]. TSV reliability depends on multiple factors such as thermal profile, current density, stress profile, TSV placement, and TSV usage. Besides, numerical methods that are conventionally used to capture TSV failures are not always computationally feasible for chip-level analysis and optimization. Therefore, efficient TSV reliability models are required.
8) *Chip/Package Interaction:* The thermo-electro-mechanical behavior of 3-D ICs is significantly affected by the package. Studies have shown that temperature, power, signal integrity, and reliability of 3-D ICs are all affected by how C4 and micro-bumps are placed and the redistribution layer (RDL) is routed [47], [48].

For each of the design challenges listed above, Section IV provides a comprehensive survey of important design solutions in the literature. We highlight how 3-D design tools and methodologies efficiently incorporate the extra design dimension, and also how they address the unique design challenges in 3-D ICs.

Section IV-I highlights currently available TSV models, including TSVs electrical, cross-coupling, stress, and electromigration models. In addition, TSV-adjacent gate timing model is also covered.

Each design challenge listed above is a complex problem on its own. Moreover, many challenges involve complex interdependencies, and improvements to one may cause degradation to another. Failure to consider these interactions could lead to an infeasible design, at which point incremental ad-hoc fixes could result in a severely suboptimal and/or costly design. Therefore, 3-D design optimization cannot be performed independently or sequentially, but requires a holistic co-design approach. Section VI shows important design tradeoff between interdependent design objectives and introduces the state-of-the-art research progress in the co-design domain.

Another interesting topic for future research is the interaction between architectural and physical design optimization. Architectural design space optimization cannot be properly performed without an understanding of the physical feasibility region of the design space. Thus the co-design paradigm must simultaneously optimize both physical and architectural design choices [17], [64], [65].

In Section VII, we provide an overview of recent research development for M3-D ICs. M3-D ICs facilitate even higher bandwidth than TSV-based 3-D ICs, but also face challenges regarding their fabrication and lack of available design tools.

Section VIII discusses open problems regarding 3-D ICs, including missing large-scale 3-D EDA and 3-D architectural modeling tools, and systematic methodologies for handling 3-D cooling and 3-D multicorner designs. Finally, Section IX concludes this paper.

II. 3-D IC FABRICATION

In this section, we provide background for two key processes during the 3-D IC fabrication flow: 1) wafer stacking and 2) TSV manufacturing.

A. Wafer Stacking

One of the enabling process technologies of 3-D ICs is vertical stacking of multiple planar wafers. Wafer stacking facilitates heterogeneous integration of chips that are designed and manufactured separately, which avoids changing existing designs and process technologies. Wafer-to-wafer (W2W) stacking and die-to-wafer (D2W) stacking are two mainstream processes. W2W is most practical for high yielding individual wafers with the same size, (for instance, memory). This is because the yield of W2W stacking decreases dramatically as more wafers are stacked. D2W stacking is more favorable for general-purpose 3-D integration, and does not require wafers/dies to have the same size.

Two general stacking paradigms exist: 1) "face-to-face" and 2) "face-to-back" bonding. The face-to-face approach directly bonds the via stubs on the front side of the chips. The front side (or the face) of a chip refers to the back-end-of-line (BEOL) layers and the back side of a chip refers to its substrate. Since the via stubs are bonded directly, the face-to-face approach requires no TSVs for interlayer power delivery and communication. The only TSVs necessary are those which connect to the I/O pins. Although this approach is simple, it only supports two-layer 3-D IC. The alternative is the face-to-back bonding. The face-to-back integration process first polishes the top chip using chemical mechanical planarization (CMP), then the substrate of the top chip is bonded to the front side of the bottom chip. An arbitrary number of chips can be stacked vertically by the face-to-back bonding, as long as the system meets its thermal and power delivery requirements.

After substrate thinning and TSV manufacturing, multiple dies are bonded vertically by precise alignment of TSVs and microbumps, in order to provide electrical connection between chips. The alignment is followed by either adhesive bond [66], oxide-oxide bond [67], or thermal compression bond [68].

B. TSV Manufacturing

The TSV manufacturing process is similar to the process for a contact hole, except that TSVs require much deeper holes that go all the way through the substrate. In general, via-last and via-first approaches are two mainstream methods. The via-last approach etches into the silicon to form the TSV hole after devices/circuits have been fully processed while the via-first approach etches the silicon before processing devices/circuits.

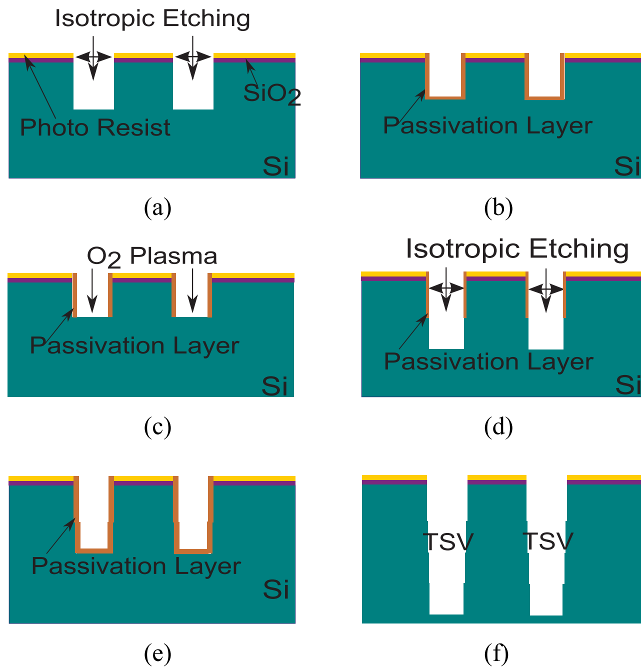


Fig. 2. Bosch process that forms TSVs using DRIE. (a) Thermal oxidation, mask patterning, and the first isotropic etching. (b) Deposit passivation layer. (c) Passivation removal by oxygen plasma. (d) Second isotropic etching step. (e) Deposit passivation leter: new isotropic etching/passivation cycle. (f) Resulting TSV cross-section with a tapering angle.

Although the two approaches differ in process details, they share similar TSV formation processes. TSV can be formed either by laser drilling or plasma etching.

The most prevailing plasma etching method is the Bosch process [69]. The Bosch process is based on deep reactive ion etching (DRIE), where isotropic plasma etch (e.g., SF_6) and deposition of etching resistant passivation layer (to protect the sidewall) alternate iteratively. Fig. 2 illustrates the overall flow of the Bosch process. The isotropic plasma etch involves removing silicon from a predefined area. Passivation material is subsequently applied to the interior of the TSV. The passivation layer at the bottom of the TSV is selectively removed by high-energy plasma to expose the bottom region. Another cycle starts from isotropic plasma etch again and at this time the sidewall of the TSV is protected by passivation material; however, bottom region can be further etched. The Bosch process can typically achieve TSVs with less than 5 μm diameter; however, the sidewall is often scalloped and tapered [70], [71].

Besides plasma etching, several publications use laser drilling to form TSVs [72]–[74]. Chen *et al.* [74] achieved TSVs with 10 μm diameter. Compared to plasma etching, laser drilling is a less-costly approach, since no photolithography or vacuum process is needed. However, laser drilling operates on a single point, and therefore is only suitable for designs containing a small number of TSVs. In addition, laser drilling damages silicon substrate around TSVs and it is often desirable to put electrical devices outside the damaged region.

The formation of TSVs is followed by the insulation step. SiO_2 layers are often used, during this step, to provide sufficient electrical isolation between metal-filled TSVs and silicon substrate. For example, Klumpp *et al.* [75] used chemical vapor deposition (CVD) to form insulation layers for its temperature stability and ideal conformality.

Finally, TSV metalization needs to be done to establish the electrical interconnect. Usually the metalization step consists of forming a diffusion barrier layer, then an adhesion layer, and finally metal filling using materials such as copper (Cu) or tungsten (W). Different deposition strategies exist, including electroplating, CVD, and physical vapor deposition (PVD). Void-free TSV metalization is very challenging for TSVs with high aspect ratio, and it is also important to develop techniques to reduce contact resistance.

TSV manufacturing involves many high-temperature processes (i.e., CVD, PVD, and annealing). Enhancement of TSV yield in the presence of repeated thermal cycling is an important manufacturing objective. Besides, when a 3-D IC is cooled from high manufacturing temperature to room temperature, the negative thermal load induces compressive stress inside the TSVs and tensile stress in the nearby silicon area, both implying severe reliability losses. Stress induced reliability challenges are discussed in detail in Section III-G.

III. Design Tool Challenges

Both academia and industry have striven to develop design tools for 3-D ICs over the past decades. Physical design tools such as partitioner, placer and clock synthesis tools are among the first to prototype 3-D ICs. One approach is to map a 2-D design into the 3-D space by reusing existing 2-D physical design tools [29], [76], [77]. The alternative is to develop native 3-D design algorithms; however, the computational complexity as well as 3-D specific design rules and restrictions must be handled effectively. As the on-chip power density continues to grow in the recent decades, sufficient cooling and high-quality power delivery have become two important design bottlenecks [35], [78]. These two bottlenecks are particularly severe in 3-D ICs since the stacking structure results in a “trapped heat” effect [17] (refer to Section III-D for more details) and significant power is lost during vertical power transportation from the power I/Os to upper-level functional units [35]. More recently, as gates and interconnects are made smaller and placed closer to each other, mitigation of parasitic influences such as coupling, thermal mechanical stress, etc. becomes increasingly important [79], [80]. TSV/package-induced thermal mechanical degradations are being investigated [47], [48].

The following sections discuss the fundamental causes of each aforementioned design tool challenge, and offer our insights into what are the fundamental directions further research must pursue.

A. Partitioning

The idea behind 3-D partitioning algorithms is to reuse 2-D placement tools once a flat 2-D circuit netlist is distributed onto 3-D layers. 3-D partitioning algorithms divide the logic blocks into different layers of a 3-D IC such that TSV usage is either minimized or less than an upper bound. Each partition needs to be assigned to a layer index. An example of gate-level partitioning is illustrated in Fig. 3. The gate-level netlist is partitioned into three layers, resulting in six TSVs. Most

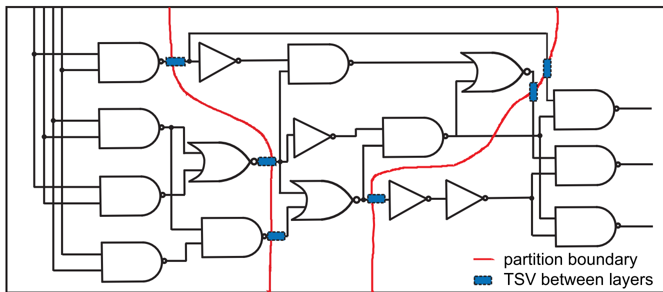


Fig. 3. Partitioning algorithm maps a gate-level circuit netlist into different layer of 3-D IC to minimize the usage of TSVs.

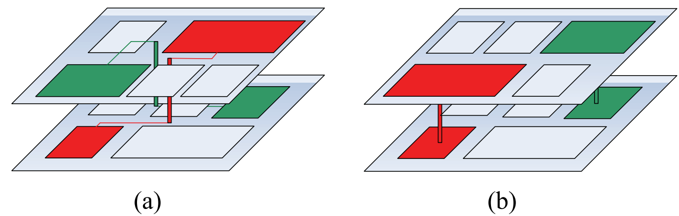


Fig. 4. Partitioning circuit netlist into layers and applying 2-D placers at each layer (a) might lead to suboptimal solutions since the interlayer connectivity is ignored. (b) Optimal 3-D placement.

of the current 3-D designs use partitioners that are originally designed for 2-D circuits. The objective of 2-D partitioning is to minimize the number of total cuts while balancing the number of members in partitions.

However, optimizing the number of TSVs is not equivalent to minimizing the number of intercluster cuts. Most partitioners designed for 2-D circuits do not account for the distribution of partitions in the vertical direction and the cuts between any pair of two partitions are treated with equal weight. In fact, cuts are cheap in adjacent 3-D layers, requiring zero or one TSV depending on bonding strategy. Cuts are expensive between two distant partitions located at nonadjacent layers as more TSVs are required. For the most common face-to-back bonding, the number of TSVs between two layers equals the product of layer index difference and number of cuts. Having the objective of minimizing the number of TSVs (or ensuring number of TSVs is below upper threshold), it is natural to penalize cuts in distant partitions.

Moreover, the netlist partitioner usually operates on a graph or a hypergraph, which is unaware of each gate's physical dimensions. Therefore naive partitioning could result in area mismatch among different layers. These challenges needs to be properly address to develop high-quality 3-D partitioners.

B. Placement

While the integration of netlist partitioning and layer-by-layer 2-D placement significantly reduces the design complexity by enabling reusing 2-D placers, the quality of the resulting placement is limited by initial netlist partition. Moreover, placing gates layer-by-layer inherently ignores the connectivity between layers, and it is possible that gates in a 3-D net are placed far away in different layers. This is illustrated in Fig. 4. In Fig. 4(a), a circuit netlist is first partitioned into layers, and 2-D placement is performed in each layer separately. However, the 3-D nets containing the red (or green) blocks have long wire length. An optimal 3-D placement is shown in Fig. 4(b), where 3-D nets effectively utilize the TSVs to reduce wire length.

The second drawback of the divide-and-conquer approach is the difficulty in controlling the TSV utilization. In the partition stage, each gate is assigned to a certain layer, therefore the number of TSVs is determined and fixed afterwards. While this scheme reduces the design complexity, it is difficult to determine what the best TSV utilization is for a given circuit graph or hypergraph, in order to minimize the total wire length.

For a given gate-level netlist, the 3-D placement problem aims to assign a position (x_i, y_i, and z_i) to each gate, so that the total wirelength and TSV number are minimized (or less than some upper limit), subject to constraints such as nonoverlapping and temperature threshold.

Two significant challenges exist for 3-D placers. The first is layer assignment for each of the gates. Unlike the 2-D placement problem which is solved in a continuous space, layer assignment is a discrete optimization problem. Solving a discrete optimization problem is usually more complicated. The second challenge is TSV placement. TSVs occupy large layout areas and their dimensions should not be ignored. TSVs can be either preplaced at certain locations and treated as placement obstacles during gate placement, or placed simultaneously with logic gates. Incorporating TSV placement into the 3-D placement flow needs to be investigated.

C. Clocking

3-D clock design has become an important and challenging research topic, striving to provide synchronization for all computations across the 3-D chip. Once the locations of the clock source [e.g., phase locked loop, delay locked loop (DLL), etc.] and clock sinks (flip-flops and latches) are known, the clock signal is delivered to each clock domain of the chip through the clock delivery network.

Like 2-D clock networks, 3-D clock networks can be implemented as either a mesh or a tree. A typical clock mesh structure is established by intersecting vertical and horizontal metal wires. Clock sinks connect to the clock mesh through short wires, often called “stubs.” The clock signal is distributed from a clock source with a high-level clock tree structure to the intersection of horizontal and vertical metal wires. A clock mesh provides low clock skew synchronization in the presence of process and environmental variations; however, it imposes wiring utilization overheads and a high power consumption.

The clock tree topology is the prevailing topology used in practice. The power consumption of a typical clock tree is much lower than clock mesh, and the 2-D synthesis flow has been heavily investigated in the past. One typical optimization goal for clock tree synthesis is to ensure zero, bounded, or useful clock skew (the maximum difference in clock arrival time between sinks). Other objectives include minimizing clock power, minimizing total wire length, etc. A naive 3-D H-tree is shown in Fig. 5. The clock signal is distributed across layers via clock TSVs. However, the H-tree structure (in blue) at the bottom layer has longer wiring, indicating late clock arrival

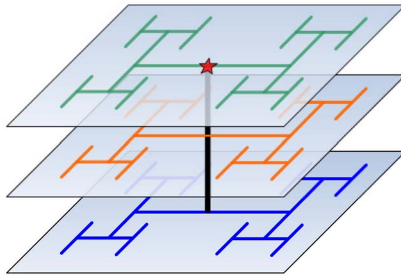


Fig. 5. Naive H-tree implementation of 3-D clock network. Clock TSVs are used to distribute clock signals across different layers, however, 3-D H-tree results in high clock skew.

time and nonzero clock skew. More sophisticated clock tree synthesis flow is obviously needed for 3-D ICs.

The 3-D clock tree synthesis flow needs to consider the usage and location of clock TSVs to optimize clock tree performance. Too few clock TSVs might not effectively exploit the potential of 3-D integration; however, too many clock TSVs might increase the clock tree's capacitive load (as TSVs have large capacitance) as well as increase the manufacturing cost and failure probability. In addition, since clock network design is generally performed after placement, 3-D clock tree algorithms need to consider the feasibility of placing clock TSVs into the layout whitespace. In cases where additional controlling signals are needed (i.e., the enable signal for clock gating), the placement of control TSV needs to be considered as well. Also, TSVs induce significant thermal mechanical stress, which affects the speed of clock buffers, potentially increasing the clock skew variation. Last but not least, 3-D clock trees can be designed in a way to have separate clock I/Os in each layer. Separate clock I/Os enable separate die testing in order to improve the yield. Prebond testability requires complete clock tree on every die; however, minimizing prebond 3-D clock tree overhead is still one of the clock tree design challenges.

Similar to 2-D ICs, 3-D clock networks can benefit from a hybrid structure that exploits the strengths of both a clock mesh and a clock tree. It deploys a loose high-level mesh, and then generates local clock trees from selective mesh stubs. This hybrid methodology has the potential to control global clock skew by tolerating process variations across different layers, and also provides opportunities for clock gating.

D. Temperature and Cooling

As very large-scale integration technology scales, electronic designs become smaller and faster, which leads to increased power density and large amounts of heat. Once a system fails to remove these heat, the temperature rises. A chip's ability to sufficiently remove the generated heat has become a dominant factor in determining performance and reliability of ICs [78]. In the era of 3-D ICs, the thermal problem is exacerbated [17], [59]. On one hand, 3-D integration enables footprint reduction by stacking several functional layers, thus causing power flux to increase linearly in the number of layers.

On the other hand, the dielectrics between functional layers have relatively low thermal conductivity, and significantly diminish heat flow from stacked layers to the heat sink in traditional air-cooling schemes. The cooling capacity on each layer of an air-cooled 3-D IC degrades as the layer moves farther away from the heatsink, therefore large thermal gradients form in the vertical direction [81]. Fig. 6 shows the steady state temperature map of a 3-D IC with two DRAM layers stacked on a 16-core multiprocessor layer. The thermal map is obtained using our in-house thermal simulator based on the work in [60]. We observe high thermal gradient both within the same layer and across vertical layers. Processor layers have highest temperature due to its high power consumption. Low-power DRAM layers have lower temperature. We also observe significant thermal coupling from the processor layer to the neighboring DRAM layer.

High temperature may cause several drawbacks in 3-D ICs.

1) Thermal runaway due to the positive feedback between leakage current and temperature [82].
2) Performance degradation (increased interconnect delay and reduced transistor performance in high temperature) [83].

Furthermore, a large vertical thermal gradient may accelerate electromigration in TSVs and increase thermo-mechanical stress due to the mismatch of thermal expansion between TSV fill material (e.g., copper) and silicon substrate [42]. This leads to significant degradation in 3-D IC reliability. Therefore, thermal-aware design is necessary and should minimize both maximum temperature and thermal gradient.

Thermal-aware design alone is not enough to efficiently mitigate the thermal problems in 3-D ICs since they often target at one single scenario (i.e., worst-case scenario) while actual thermal profile of a 3-D IC varies significantly both spatially and temporally. The variation in thermal profile implies that we can exploit the design margins to obtain the best performance using run-time thermal management methods as well [84]–[86]. Compared to 2-D ICs, the different cooling efficiency between layers and strong thermal coupling across layers make the run-time thermal management in 3-D ICs more complex [86].

E. Power Delivery

In a 3-D IC, power is delivered from off-chip package through C4 bumps and then distributed vertically through power TSVs [87]. Fig. 7 illustrates a 3-D PDN circuit model, which consists of three parts: PCB, package and on-chip circuits. The on-chip circuit is modeled as a meshed RLC network capturing the voltage distribution in both vertical and planar directions.

The vertical structure of a 3-D PDN brings several new challenges. First, as 3-D integration enables stacking multiple functional layers vertically, power scales volumetrically with the product of footprint area and the number of layers. However, the number of power delivery pins (i.e., the power delivery capacity) is a function of footprint area only. This imbalance between power supply and demand raises challenge for 3-D PDN [61]. Second, the parasitics of power/ground TSVs (P/G TSVs) affect the resonant frequency of each layer thus influencing the power noise characteristics in 3-D

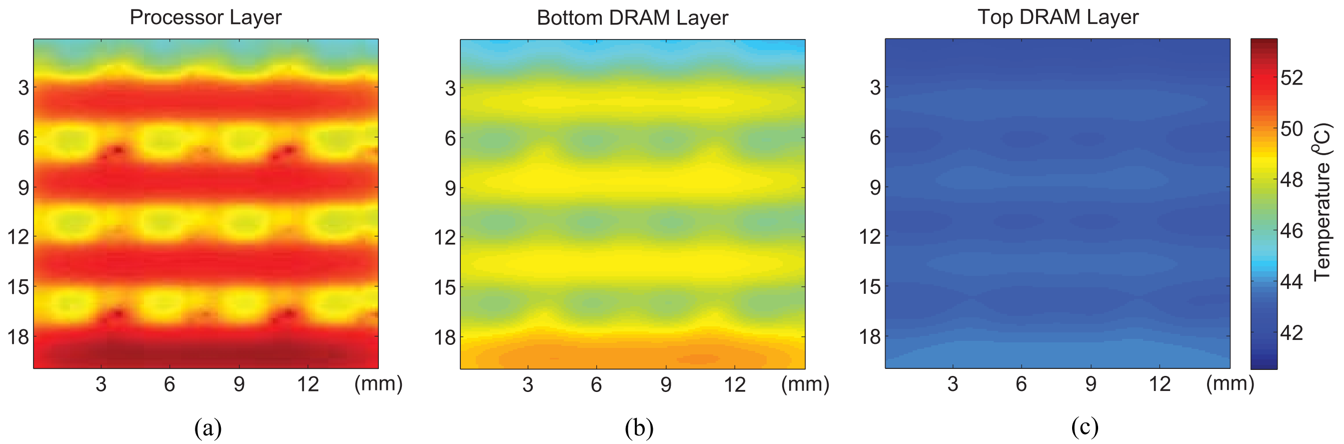


Fig. 6. Thermal map of the (a) processor layer, (b) bottom DRAM layer, and (c) top DRAM layer (close to the heat sink).

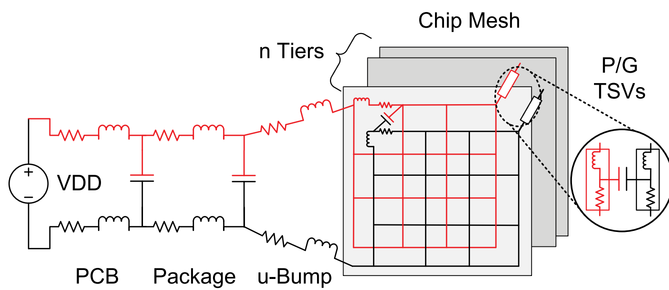


Fig. 7. Illustration of PDN model in a 3-D IC.

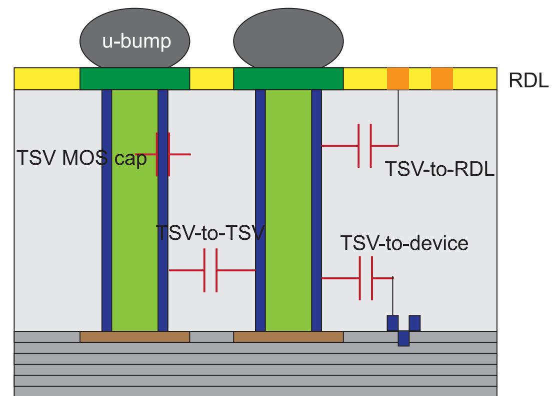


Fig. 8. New coupling parasitics in 3-D ICs.

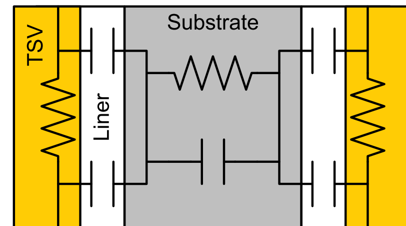


Fig. 9. TSV-to-TSV coupling circuit model.

ICs [88]. As the current draw has significant spatial variation, the PDN noise shows great variation spatially. Third, the stacking structure of 3-D ICs enables power noise from one layer to couple in neighboring layers [35]. For example, when designers fail to de-couple power noise from one layer to another, the CPU layer activity can induce significant power noise in adjacent layers [89]. Fourth, in a air-cooled 3-D IC, the heatsink and the power delivery pins are almost always on opposite ends of the chip stack [53]. This means the chip layer with the most cooling capacity (i.e., closest to the heatsink) is the one with the worst power integrity, and vice versa. This necessitates aggressive management and design methodologies considering both power delivery and temperature.

F. Signal Integrity

3-D ICs built with TSVs pose new challenges in parasitic extraction. The 3-D IC specific parasitics are mostly related to TSVs and RDL routing as illustrated in Fig. 8 [45], [90], [91]. TSVs themselves couple with each other in the substrate (TSV-to-TSV) [45], [92], [93]. In addition, TSVs affect devices nearby (TSV-to-device) [94] and RLD interconnect nearby (TSV-to-RDL) [90]. TSV itself contains MOS cap due to the copper, silicon dioxide, and silicon interface structure (TSV MOS cap) [95]. TSVs couple to wiring in metal layers above (TSV-first) and along side (TSV-last) the TSV [91]. Lastly, RDL interconnects couple each other [96]. These parasitics affect power, performance, and noise in 3-D ICs and thus need to be modeled and extracted accurately and efficiently.

One design challenge in 3-D ICs is to ensure signal voltage noise is maintained within design margins. Cross coupling between switched devices can cause increased leakage/short circuit currents and possibly result in digital glitches that affect circuit behavior or cause incorrect computations. In addition to the traditional sources of coupling noise (wires and transistors), TSVs provide a new coupling source in 3-D ICs [45], [92], [93]. TSVs have the potential to be more problematic than planar wires since they are much larger, and surrounded by a much thinner insulation layer [44], [45]. TSVs can easily couple into the conductive silicon substrate through the thin oxide liner around the TSV. From there the voltage noise can couple into other TSVs or transistors through the conductive substrate [45], [92], [95].

Fig. 9 shows a circuit model of coupling between two TSVs. TSV coupling is most strongly affected by liner capacitance which is independent of the distance between TSVs.

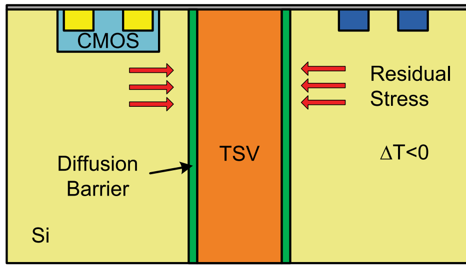


Fig. 10. TSV manufacturing process induces significant thermal mechanical stress around TSV. The stress field induces TSV stress migration and material fracture, and causes mobility deviation in nearby CMOS devices.

Thus, TSV coupling is not efficiently mitigated by increasing TSV pitch. Liu *et al.* [45] showed that increasing TSV pitch from 1 μm to 20 μm (20× increase) only reduced TSV coupling from 255 to 225 mV (12% reduction).

G. TSV Reliability

Most reliability concerns specific to 3-D ICs are related to TSVs. TSV introduces several new failure modes, many of which are derived from 3-D ICs' thermal and stress issues. The thermal issue comes from the fact that the stacked structure increases the power density without providing a sufficient heat removal path. The stress issue originates from significant difference in coefficient of thermal expansion (CTE) between TSVs (e.g., copper $1.77 \times 10^{-5}K^{-1}$) and the silicon substrate ($3.05 \times 10^{-6}K^{-1}$). When TSVs are cooled down from high manufacturing temperature to room temperature, negative thermal load is applied thus compressive and tensile stress are formed inside TSVs and neighboring substrate areas. This phenomenon is illustrated in Fig. 10.

TSV-induced reliability losses include: TSV electromigration, TSV stress migration, TSV oxide breakdown, TSV thermal cycling, and TSV stress-induced material fracture. TSV electromigration and stress-migration cause TSV metal atoms to migrate over time, gradually altering material density and resistance, and eventually causing TSV open-circuits. TSV oxide breakdown occurs when the electrical field inside TSV barrier layer exceeds its threshold, destroying the electrical isolation between TSVs and the substrate. TSV oxide breakdown occurs under high voltage stress (50 V) [97], so in general is not considered as the primary failure source in digital 3-D ICs. Thermal cycling leads to thermal fatigue and shortens TSV lifetime by introducing defects. Material fracture, initiated by manufacturing imperfections (voids inside TSVs) and accelerated in high stress environments, may lead to delaminations or cracks around the TSV structure. All the above mentioned TSV failures are exacerbated at elevated temperature.

H. Chip/Package Interaction

The mechanical stress produced in the package domain (i.e., C4 bumps, microbumps, and underfill layer) significantly impacts the stress profile in the chip domain. Due to the difference in CTEs between package material and silicon substrate, when 3-D ICs undergo thermal load, large residual stress gets coupled from the package domain to the chip domain, degrading the chip reliability. Finite-element-method (FEM) is a useful tool to analyze the interaction between package and chip domain, however, it is computationally expensive and infeasible for full-chip or full-package analysis.

IV. 3-D IC Design Tool Innovation: State of the Art

In this section, we discuss the historical as well as the state-of-the-art solutions for each of the 3-D IC design tool challenges. We also provide our insights for future research directions that are necessary to push 3-D ICs toward future commercialization.

A. Partitioning

3-D partitioning algorithms divide a gate level netlist across different layers of a 3-D IC such that TSV usage is minimized or meets design constraints. Most 3-D partitioning algorithms use 2-D partitioners to recursively partition a flattened circuit netlist into arbitrary number of 3-D layers. The objective of these 2-D partitioners is usually minimizing the number of total cuts. Circuit partitioning is an NP-hard problem, however, several efficient bi-partitioning heuristics were developed, including the Kernighan–Lin (KL) [98] algorithm and the Fiduccia–Mattheyses (FM) [99] algorithm.

In bi-partitioning algorithms such as KL and FM, a gate level netlist is represented by a graph or a hypergraph. The algorithms start with an arbitrary balanced bi-partition. During each iteration, cell swapping (or cell movement) with the largest reduction in cut size is taken. The bi-partitioning algorithms terminate if no improvement is found during a user defined number of consecutive iterations.

The partitioning problem is often solved hierarchically to provide algorithm scalability to large circuits. The idea is to progressively reduce the size of a netlist by clustering highly connected nodes into a single node. The process is repeated until the size of the clustered graph is small enough to be partitioned using well known algorithms like KL or FM. Once partitioning is done, the coarsening is undone iteratively and local improvements are made at each level of the un-coarsening loop. Academic multilevel partitioners include MLPart [100] and hMetis [101].

However, minimizing the number of TSVs is not equivalent to minimizing the number of intercluster cuts. Cuts are cheap in adjacent 3-D layers, but are expensive between nonadjacent layers since more TSVs are required. A good 3-D partitioner needs to penalize cuts between distant layers. To minimize the number of TSVs rather than to minimize the number of cuts, some preliminary 3-D partitioning work employs 2-D partitioners to partition the circuit netlist into groups, and then distribute the partitions in the vertical direction so that number of TSVs is small. Ababei *et al.* [26] mapped an initial partition result (generated by hMetis) into a matrix, and proposed a matrix manipulation heuristic to assign layer index to each of the vertices in the netlist graph. Sawicki *et al.* [27] proposed a simulated annealing-based approach to perturb the hMetis solution. Their approach allocates a node to adjacent layer or swaps nodes in different layers. The authors showed this

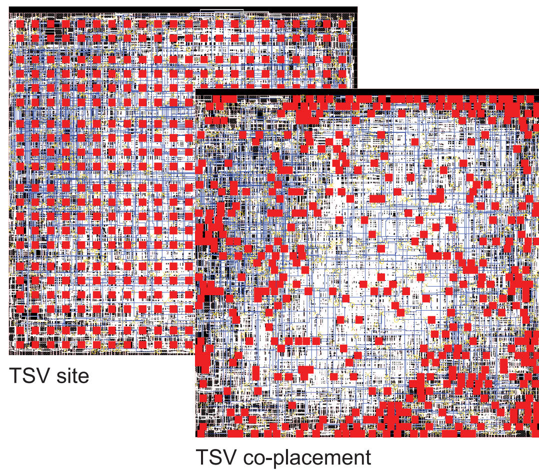


Fig. 11. Two 3-D placement styles: TSV site and TSV co-placement.

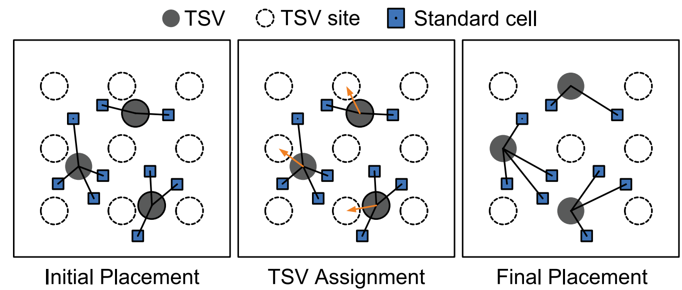


Fig. 12. TSV site placement algorithm in [28] preplaces TSVs uniformly on each layer, and performs 3-D standard cell placement. After cell placement, nearest TSV site is inserted into corresponding 3-D net.

perturbation can achieve approximately 15% reduction in TSV counts, compared to an initial solution generated by hMetis.

Pathak *et al.* [50] demonstrated the existence of *TSV usage sweet spot*: up to a certain point, the more the TSVs are used, the better the full-chip wirelength and delay become. However, if too many TSVs are used, the design quality becomes worse mainly because of the large area overhead of TSVs. This means min-cut objective may not always lead to the best design quality. Jung *et al.* [102] and Song *et al.* [103] showed that *module folding*, where a given functional module is partitioned into multiple tiers and designed under the same footprint, helps reduce power consumption. More research is needed in this area to further improve the partition quality.

B. Placement

For a given circuit netlist, the 3-D placement problem aims to assign a position (x_i, y_i, and z_i) to each cell, so that the total wirelength and TSV number are minimized, subject to constraints such as nonoverlapping gates and temperature threshold. A 3-D half-perimeter wirelength (HPWL) model that calculates cells' center-to-center distance is common for wire length estimation [44]. Quadratic total wire length function has also been employed [76]. The vertical distance can be assigned a heavier weight to penalize TSV usage [104]. Kim *et al.* [28] introduced two styles of 3-D placement, namely, TSV site and TSV co-placement as shown in Fig. 11. TSV site places all TSVs on a uniform grid and fixes their locations. TSV co-placement inserts TSV cells into the 3-D netlist and places them simultaneously with gates.

1) TSV Site: In TSV site, TSVs are preplaced on a uniform grid and are not associated with specific nets until cell placement. Cells are placed while treating TSVs as placement obstacles, and after gate placement a TSV assignment algorithm assigns each 3-D net to a TSV such that total wire length is minimized. A TSV site technique based on TSV assignment [28] is illustrated in Fig. 12.

The TSV assignment problem was investigated in [105] and [106]. Yan *et al.* [105] formulated a binary integer linear programming problem. The binary variables determined whether a TSV belonged to a specific 3-D net. Tsai *et al.* [106] used a maximum flow formulation, where TSVs and the bounding boxes of the 3-D nets that contain the TSVs were represented as nodes, and one unit flow was established when one TSV was assigned to a TSV bounding box.

2) TSV Co-Placement: The second method for TSV insertion, TSV co-placement, inserts TSV cells into the 3-D netlist and places them simultaneously with gates. TSV co-placement' has been shown to outperform TSV site by roughly 10% in wirelength while having little overhead in run-time [28].

Currently, most "true 3-D" placers [i.e., simultaneously place gates and TSVs in (x, y, and z)] have ignored TSV area (i.e., assume TSV placement is not overlap-constrained). Alternatively, the works that have considered physical placement of TSVs have used a partition-and-place methodology. Implementation of true 3-D placement technique which yields valid TSV placements is an ongoing research goal.

Some notable 3-D placement works in the literature include transformation-based placers [18], partition-based placers [107], analytical placers [29], [104], and force-directed placers [30], [31]. The current literature on transformation-based, partitioning-based, and analytical placers has modeled TSVs as zero-diameter vias which consume no placement area. Thus further legalization is required to make space for physical TSVs and produce a real 3-D placement. Force-directed placers take TSV dimensions into consideration, and thus require less post processing for legalization. Each of these placement paradigms are defined as follows.

1) *Transformation-Based Placer:* Transform a 2-D placement result into 3-D space [18]. Transformation algorithms based on stacking and folding a 2-D design are proposed. After transformation, the authors assign a layer index to each of the cells to reduce the number of TSVs and peak temperature.
2) *Partition-Based Placer:* Assign a weight to each net based on its switching activity, capacitance, and number of TSVs [107]. Then the placer applies a recursive bi-partition approach. Two cut directions are considered: a) intertier z *cut* to minimize the number of TSVs and b) intratier x/y *cut* to minimize the wire length.
3) *Analytical Placer:* Apply a density smoothing function to penalize cell overlaps in x/y directions and use

Lagrange multipliers to minimize the total wirelength, TSV usage and cell overlaps [104]. Two relaxations have been used to provide a differentiable optimization objective function: a) the 3-D HPWL model was relaxed to a log-sum-exp wirelength model and b) the vertical direction was relaxed from discrete to continuous space. After solving the Lagrangian optimization, a bell-shaped density projection function can be applied to obtain legal layer indexes for cells.

4) *Force-Directed Placer:* Minimize the quadratic total wire length function, which can be modeled as the total spring energy of an elastic spring system [30], [76]. Cell movements change the total spring energy, and the derivative of the spring energy is a force (referred as net force) which indicates which movements reduce the quadratic wire length. If all cells are movable, the trivial optimal placement is to place all cells at a single point. Another force (referred to as move force) is necessary to pull cells toward low cell density areas to remove overlap. Various other forces have been defined to improve optimization convergence time and model objectives and constraints beyond wirelength and overlap. The force-directed placer starts with an arbitrary initial 3-D placement solution. During each iteration, net force, move force, etc. are combined to determine the overall force vector, and cell positions are solved such that total force is zero on all cells. Each force is then recomputed based on the new placement and the process repeats iteratively [30], [31].

C. Clocking

Clock mesh design consists of two steps: 1) mesh construction and 2) clock sink assignment. A high-quality mesh structure minimizes total capacitance and clock skew, and the sink assignment step assigns the clock sinks such that the capacitive loading on the clock mesh structure is balanced. Load balancing is necessary to achieve good clock slew/skew.

A clock mesh provides low clock skew in the presence of process and environmental variations; however, it can impose routing overheads and is inherently high-power. Several existing approaches to reduce 2-D clock mesh overheads include: mesh size tuning [108], nonuniform grid size (more mesh wires in high cell-density regions) [109], and mesh buffer placement and sizing [108], [109]. Similar approaches can be applied to 3-D ICs as well. For 3-D clock mesh construction, TSVs can be inserted to reduce the total wire length on each die [110].

Clock tree is a more prevailing topology due to higher power efficiency and straightforward implementation. Generally the 3-D clock tree synthesis is a two-step procedure: 1) 3-D abstract clock tree generation and 2) 3-D clock tree embedding. The 3-D abstract clock tree represents the connectivity from a centralized clock source to all the sequential logic. The 3-D clock tree embedding determines the physical locations of the intermediate clock tree nodes, as well as clock TSVs.

Currently, the 3-D abstract clock tree is usually designed from a wire length minimization standpoint. For example,

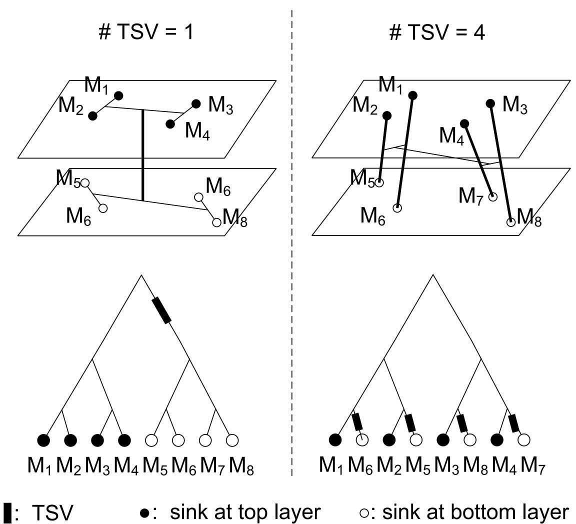


Fig. 13. Abstract clock tree generation based on 3-D method-of-median-and-mean. When TSV bound is one, clock routing is performed within each layer and one TSV connects neighboring layers. When TSV bound is more than one, clock sinks are partitioned in *X*- or *Y*-direction, and the algorithm is recursively applied on each subtree.

Zhao *et al.* [57] implemented a cutting-based approach for abstract clock tree generation, such that the clock sinks are recursively partitioned in *X*, *Y*, or *Z* directions. The partitioning direction, either *X/Y* or *Z* partition, was chosen based on the available TSV count. The location of the cutting line was determined based on the median value of the logic cells' coordinates. This method is illustrated in Fig. 13. A similar approach was reported in [111], where the partitioning direction was chosen by the HPWL of subsets. A nonpartitioning-based greedy algorithm was proposed by Kim and Kim [112]. That technique greedily picks the closest sequential pairs and recursively formed the abstract clock tree. Liu *et al.* [113] proposed a density-aware sorting algorithm to cluster clock sinks and achieve a balanced distribution of TSVs, which enhances the TSV manufacturing yield.

During 3-D clock tree embedding step, the physical locations of the intermediate clock tree nodes, as well as clock TSVs are determined such that the clock tree total wire length and clock skew are minimized. For example, Kim and Kim [112] extended the 2-D deferred-merge-embedding algorithm [114] to the 3-D space in order to decide the length of clock tree edges and the locations of clock TSVs to achieve minimum wire length and zero-skew in a 3-D clock tree.

Clock skew and power minimization are two important research topics. Lu and Srivastava [115], [116] applied clock gating techniques to reduce power consumption in 3-D ICs. Minz *et al.* [21] constructed 3-D clock trees to minimize skew while considering the underlying thermal variations. Zhao *et al.* [56] built clock trees that can be used both during prebond testing and post-bond normal operation in 3-D ICs. Kim and Kim [117] improved this paper by further minimizing clock TSV count and wirelength.

Several works have incorporated TSV stress awareness into 3-D clock tree synthesis. For example, Yang *et al.* [118]

investigated the impact of the TSV-induced stress on timing corners and optimized the location of clock buffers to minimize the stress-induced clock skew variation. Lu and Srivastava [55] incorporated TSV stress migration awareness into the 3-D clock tree synthesis flow. Lung *et al.* [119] considered the reliability issues of clock TSVs and used TSV fault-tolerant unit (TFU) to enhance clock tree robustness against TSV failures. Further tuning techniques for the TFU and associated fault-tolerant 3-D clock tree, such as better control of clock skew, TSV count, and clock slew were proposed in [120].

Moreover, it is often desirable to provide prebond testability for 3-D ICs, in which every die is tested separately before they are bonded vertically. Zhao *et al.* [56] implemented a redundant clock tree structure to enable separate testing in each layer, while minimizing the redundant clock tree's total wire length and power. Kim and Kim [117] designed 2-D clock trees on each layer for testing purposes along with the real 3-D clock tree. Transmission gates were used to disable the clock TSVs and enable the 2-D trees during testing. During normal operation the state of the transmission gates are switched, activating clock TSVs and delivering the clock signal from one centralized clock source.

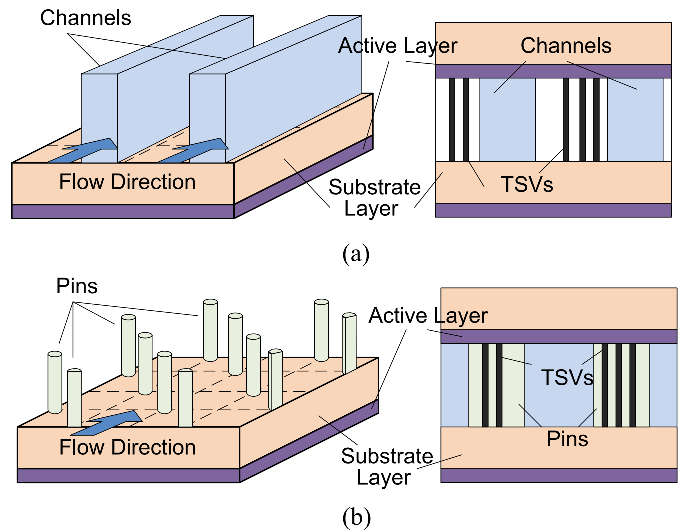


Fig. 14. Illustration of (a) microchannel cooling and (b) micropin-fin cooling.

D. Temperature and Cooling

Conventional thermal management for IC often assumes a passive cooling scheme which predominantly rejects heat through the air cooled heat sink located at the top of the chip stack. However, studies have shown that air-cooling alone does not offer sufficient cooling to realize the true potential of 3-D ICs [17], [32], [33]. Hence the application of embedded active cooling solutions, such as micro-fluidic (MF) cooling, in 3-D ICs has been investigated. Two common structures used in MF heatsinks are the microchannel and the micropin-fin, as illustrated in Fig. 14. Microchannels or micropin-fin cavities are etched directly into the substrate of each layer thus providing a conductive heat path from the heat source to the fluid coolant. Coolant (e.g., deionized water) is supplied from the inlet of microchannel or micropin-fin cavities. Heat dissipated from active layers is transferred to the coolant and pumped out of the chip. The fluid flow provides a much lower resistance path for heat flow from junction to ambient, and provides 3-D ICs with much more heat removal capacity which scales with the number of layers [32].

However, MF cooling also has several negative effects.

1) Vertical bandwidth of 3-D ICs can be restricted since TSVs cannot be co-located with the micro-cavities [121].
2) The cooling capacity of the fluid flow progressively diminishes along the direction of flow as the fluid heats up. This is referred to as the thermal wake effect, and results in systematic in-layer thermal gradients from inlet to outlet. This may have undesirable effects on chip reliability [122], [123].
3) Extra power is required to pump coolant through the cavities, thus reducing the energy efficiency of the 3-D IC [59], [60].
4) The substrate thickness is increased in order to etch microchannels or micropin-fin cavities which potentially requires TSVs with larger diameter to maintain manufacturing specifications on TSV aspect ratio [124].
5) MF cooling structures introduce new reliability issues to 3-D ICs. Cracks have the potential to form at the high-stress areas near pin-fin corners and propagate causing catastrophic chip failure [125].
6) Moreover, MF cooling capacity may degrade as corrosion/erosion occurs in the channels. Likewise, the dispatched particles as well as particles from the environment may clog the channels [125], [126].

1) Design-Time Thermal Management: Several directions have been pursued to fully exploit MF cooling: optimizing the physical structure of the MF heatsink post-floorplanning [64], [127], investigation of the co-optimization of MF cooling and 3-D architecture [121], investigation of the thermo-electrical tradeoff made by changing the structure parameters of MF heat sink [124], placement/floorplanning algorithms that are aware of MF cooling [17], [59], [122], and co-optimization of MF channel, PDN, and signal net routing [128].

Many other research directions have been developed orthogonal to MF cooling. For example, researchers have proposed thermal-aware placement to minimize the peak temperature and the thermal gradient along with traditional placement objectives like wire length and routability. Several existing thermal-aware placement algorithms use force-directed [30], transformation [18], and min-cut [129] placement techniques.

In contemporary 3-D IC placers, the position of TSVs can also be determined during the placement stage [29], [104]. Since the thermal conductivity of the TSVs (e.g., copper) is higher than that of silicon and dielectrics, the distribution of TSVs significantly influences the cooling capacity distribution. Therefore, a more advanced thermal-aware placement algorithm was proposed to optimize the position of gates and TSVs simultaneously [18]. In addition to existing signal TSVs, researchers also insert redundant TSVs, called thermal TSVs

(TTSVs), to further improve heat spreading in 3-D ICs [81]. However, adding TTSVs have several nonideal effects.

1) They increase footprint area due to large TSV dimensions.
2) They compete for space with gates, routing wires and other TSVs (i.e., signal TSVs) thus increasing the total wire length.

In order to increase the efficiency of TTSVs, a number of algorithms for determining the number and position of TTSVs during post-placement [130], [131], in-placement [132], and in-routing [19] stage have been proposed. Finally, a hybrid solution using both TTSVs and microchannels was proposed [59] which offers area savings over TTSV-only and power savings over microchannel-only.

2) Run-Time Thermal Management: Besides thermal-aware physical design, run-time thermal management approaches in 3-D ICs are also widely investigated. Under air-cooling schemes, researchers tried to perform task assignment [84] or the hybrid of dynamic voltage and frequency scaling (DVFS) and task migration [10], [85] to dynamically adjust the power of cores in 3-D multiprocessor systems-on-chips (3-D MPSoCs). Another study [86] found that the router power in 3-D MPSoCs dominated the power dissipation and proposed a dynamic routing algorithm to achieve the thermal management. Similarly, run-time thermal management techniques have been proposed for micro-fluidically cooled 3-D ICs as well, such as dynamic flow rate control [133], [134].

E. Power Delivery

In order to suppress the power noise in 3-D ICs, several categories of research are ongoing.

1) Optimizing Physical Structure of 3-D PDN: 3-D PDNs rely on power TSVs to deliver power across layers but power TSVs introduce new source of IR drop. To suppress the IR drop, researchers looked into opportunities that reduces the size and increases the density of P/G TSVs [35]. However, dense P/G TSVs reduce the flexibility of placing signal TSVs, and may cause routing congestion [135]. PDN design flows incorporating awareness of signal routing resources have been proposed [135], and several spatially nonuniform P/G TSV topologies have been investigated [136], [137]. Healy and Lim [136] investigated how the size and pitch of P/G TSVs affect the IR drop of 3-D ICs.

2) Suppressing Transient Power Noise: Transient power noise is affected by the PDN impedance. If current frequency spectrum contains significant components at anti-resonant frequency of the PDN, significant Ldi/dt droop occurs [34]. In order to suppress the transient power noise, on-chip decoupling capacitance (decap) is used [138]. 3-D integration enables innovative decap solutions such as stacking additional decap layers [139]–[141] or through-silicon-capacitors [34]. Compared to the conventional decap solution, this method saves space for placement and routing, thus further reducing the footprint of 3-D ICs.

3) Voltage Stacking: Conventionally, circuits draw current from VDD and reject it to ground in parallel and independent of one another, resulting in extremely large current demand from the power supply pins. Gu and Kim [142] stacked different supply voltages in series (i.e., GND-VDD-2VDD) to share current between stacked circuits, such that current rejected to ground from the top circuit could be reused as current drawn from VDD by the lower circuit. Although this current sharing increases the PDN impedance, power noise is efficiently suppressed by optimizing the number and value of voltage stories. However, this method requires explicit voltage regulation. Since traditional voltage regulator modules (VRMs), built on the off-chip board, are slow and inaccurate, several work investigated on-chip VRMs [143]. The distributed on-chip VRMs can also enable faster and explicit DVS management.

4) Co-Optimizing Cooling and Power Delivery: A conventional air cooled 3-D IC possesses two opposing trends in its cooling capacity and power integrity. A layer located closer to the heat sink has good cooling capability but is inherently far away from the power I/Os, resulting in a noisy PDN. To handle this dilemma, researchers have attempted to optimize the power demand across layers. For example, Panth *et al.* [53] provided a PDN aware gate-level partitioning algorithm to minimize the voltage drop while meeting the thermal and wire length constraints. Healy and Lim [136] studied three stacking topologies in a memory-over-logic system and discovered that the least power noise was achieved by interleaving DRAM and processor layers.

F. Signal Integrity

A promising solution for mitigation of TSV coupling is shielding. Shielding involves placing a grounded conductor near the switching TSV in order to cut off the propagation of electromagnetic (EM) waves through the substrate. A number of different implementations of shielding have been investigated. Khan *et al.* [144] proposed coaxial TSVs that provide a grounded conductor shell around each TSV; however, the manufacturability and cost efficiency of such a solution is doubtful. Guard ring (i.e., a ring of grounded diffusion surrounding a circuit) is a well known solution for noise isolation of transistors. Cho *et al.* [43] proposed implementing a guard ring around TSVs for noise isolation. However, that study showed that the quality of their solution was highly dependent on the depth of the doped region, and sufficient shielding could not be provided at reasonable depth levels. Because TSVs extend through the entire thickness of the substrate, coupling occurs deep within the substrate. This is in contrast with the transistors coupling problem where coupling is confined near the surface of the substrate. For these reasons guard rings are a good shielding solution for transistors, but not TSVs.

However, the principal of a guard ring can be extended to TSV coupling by using a ring of grounded TSVs [45]. The technique is successful, but consumes significant area overhead increasing the area of each shielded TSV by 9×. The area overhead can be significantly reduced by judiciously placing shields only along the coupling paths between coupled pairs of TSVs, and by using shields to simultaneously shield multiple paths where the paths intersect [92]. However, to implement such a scheme requires a chip-scale TSV coupling model capable of simultaneously modeling the pairwise coupling

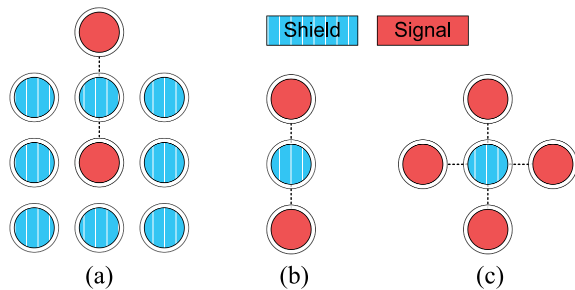


Fig. 15. (a) TSV guard ring is area inefficient for pairwise TSV coupling. (b) Judicious shield insertion only inserts shields around the coupling path. (c) Single shield can shield multiple TSV pairs.

coefficients (i.e., S-parameters) between a set of 100s–1000s of TSVs. The circuit model in Fig. 9 has been used to model coupling between a small set of TSVs, but it is computationally inefficient to use in chip-scale coupling-aware placement optimization.

Serafy *et al.* [92] proposed a chip-scale analytical model of TSV coupling (Section IV-I2), and the model was used to develop a shield insertion algorithm. The proposed algorithm identifies pairs of coupled TSVs and optimizes the shield placement to meet coupling thresholds while minimizing the total number of shields (and thus the area overhead). This method [92] matched the signal integrity achieved in [45] while reducing the total number of shields by 65%–70%. Subsequent work [44] integrated the chip-scale TSV coupling model and shield insertion algorithm into a force-directed 3-D placement flow and successfully placed a set of benchmark circuits subject to signal integrity constraints. As an alternative to TSV shielding, Peng *et al.* [46] proposed differential pairs of TSVs to pass signals between layers such that the pairs noise-cancel one another, and removed any bias noise on the pair from other sources. The approach seems promising in both area reduction and signal integrity performance.

Song *et al.* [93] extended the analytical model in [45] to consider the coupling impact of non-neighboring TSVs on a given victim TSV. The work in [95] considered various silicon effects that contribute to TSV-to-TSV coupling in 3-D ICs including MOS cap depletion region, silicon substrate effect, and the electrical field distribution around TSVs. Peng *et al.* [91] extracted TSV-to-wire coupling capacitance using a pattern matching technique. Lastly, the work in [90] conducted the full-chip extraction of metal-to-metal coupling in face-to-face bonded 3-D ICs and studied the impact on power, performance, and noise.

G. TSV Reliability

TSV reliability analysis involves complex dependencies on TSV thermal, stress, and/or current density conditions, therefore an accurate and efficient modeling approach is needed. This paper covers two broad TSV failing mechanisms, including TSV electromigration and TSV-induced thermal mechanical stress degradation. Design-time TSV lifetime estimation and tuning are useful, and run-time approaches can adjust the TSV usage according to its wear-out status.

1) TSV Reliability Modeling: TSV electromigration models have been developed by several research groups. The classic electromigration mean-time-to-failure (MTTF) model proposed in [145] is still a rule of thumb for TSV electromigration estimation. More recently, Huang *et al.* [146] proposed a physics-based electromigration modeling and assessment method for 3-D PDN. Developing efficient and accurate TSV electromigration models is still an ongoing research problem. Zhao *et al.* [147] modeled the impact of current crowding in P/G TSVs and calculated the associated IR-drop degradation. This paper is extended [148] to conduct transient modeling of PDN aging and void/hillock formation from electromigration in 3-D IC PDN. Pak *et al.* [149] studied the electromigration issues in a multiscale PDN structure that features P/G TSVs and global PDN wires versus local P/G vias and local PDN wires. Lu and Srivastava [55] developed and verified an analytical objective function to represent the intensity of TSV stress migration.

Significant modeling work has also been ongoing to capture TSV-induced thermal mechanical stress degradation. Ryu *et al.* [20] developed a semianalytical model for TSV induced thermal mechanical stress. Jung *et al.* [150] used energy release rate to model the likelihood of TSV interfacial cracks. Another important material reliability metric that has been used to study and optimize TSV-induced thermal stress is the von Mises yield criterion. The von Mises yield criterion suggests that the material yield begins when its von Mises stress exceeds certain material dependent threshold. Prior to this yield threshold, materials deform elastically and can return to original shape once external forces are removed.

2) Design-Time Approach: One example of a design-time approach to address TSV reliability concerns involves insertion of redundant TSVs into the layout to enable online reconfiguration or remapping techniques once a portion of TSVs fails. However, the number of TSVs available to designers are usually limited because TSVs need to maintain certain keep-out distance with gates and other TSVs (see Section IV-I for more details). Due to the limitation on TSV numbers, sophisticated TSV redundancy schemes are required. For example, Jiang *et al.* [151] proposed a TSV repair algorithm based on TSV redundancy network that was able to reconfigure the TSV usage when TSV failure occurs while keeping area and timing overhead of the approach small.

Besides redundant TSV insertion, physical design techniques such as floorplanning and placement can be implemented in a way such that TSV reliability is optimized. For example, Zou *et al.* [152] proposed a TSV stress-aware floorplanning method to minimize the possibility of wafer cracking and interfacial delamination. Likewise, Serafy *et al.* [65] proposed a TSV electromigration-aware floorplanner. Lu and Srivastava [153] optimized the chip thermal profile using TTSVs, in order to alleviate TSV-induced stress migration. Lu *et al.* [38] also developed an electromigration aware TSV placement heuristic to improve TSV MTTF. Electromigration is also considered during signal routing in 3-D ICs [39], [154]. Lu *et al.* [42] also developed material fracture objective functions based on von Mises stress to optimize 3-D IC von Mises stress distribution.

3) Run-Time Approach: Two branches of run-time reliability management are investigated: 1) task scheduling/migration/swapping and 2) DVFS. For example, Tajik *et al.* [155] exploited task migration in 3-D processor cores to recover 3-D cache NBTI degradation. Chantem *et al.* [156] presented a task assignment and scheduling algorithm to dynamically tackle system reliability degradation due to electromigration, stress-migration, oxide breakdown, and thermal cycling.

As an example of DVFS, Mercati *et al.* [157] dynamically tuned multicore processor operating points to mitigate long-term reliability loss without throttling performance. Zhuo *et al.* [158] applied DVFS to manage the system reliability (in terms of oxide breakdown) and performance.

In [159], thread swapping techniques along with DVFS were used to reduce the workload variance. The authors observed that reducing the variance of workload distribution can effectively improve system reliability by avoiding early failures.

H. Chip/Package Interaction

FEM for chip/package stress analysis has been used but requires long simulation time and thus is not feasible for chip-scale analysis and optimization [47]. Instead, linear superposition of stress tensors [41], [42] is a powerful method that is suitable for analyzing the interaction between package and chip and developing optimization methods to reduce the mechanical stress in 3-D ICs.

Assuming the packaging materials and the silicon substrate are linearly elastic structures, the stress value in the chip domain can be estimated using the linear superposition principle [41]: the stress coming simultaneously from several different bodies (TSVs, C4 bumps, microbumps, underfill layers, etc.) is the sum of the stresses from each source alone.

Using the linear superposition principle to estimate the von Mises stress, Jung *et al.* [48] investigated the impact of C4 bumps, microbumps and TSVs on the mechanical reliability of different layers of a 3-D IC. Different size/pitch combinations for TSVs and bumps were studied. Yang *et al.* [160] demonstrated how chip and package components affect the mobility of electrons and holes (through piezoresistive effect) as well as full-chip timing variations.

I. TSV Modeling

Accurate TSV modeling has been a primary goal of the research efforts in the field of 3-D ICs. This involves the physical and electrical properties of TSV, both of which are expected to have great impact on the cost, performance and reliability of the 3-D IC technology. Models must be both fast and accurate in order to be integrated into EDA tools for large scale 3-D IC design. Research efforts toward this goal are summarized below.

1) TSV Electrical Model: Researchers have been developing TSV resistance, inductance, and capacitance models based on physical parameters and material characteristics. TSV dc resistance is well modeled as the resistance of a metal cylinder. However, for high-frequency ac signals, TSV resistance increases due to skin effect [161]. The TSV inductance depends upon the diameter and length of the TSV and an empirical expression is given by [162]. The TSV capacitance is the series combination of the oxide and depletion capacitance, and is proportional to the length of the TSV, inversely proportional to the TSV dielectric thickness, and affected by substrate doping concentration [163], [164]. The TSV capacitance is generally considered to have the most dominant impact on TSV delay [163].

2) TSV Cross-Coupling Model: Serafy *et al.* [92] proposed an analytical model of pairwise TSV coupling as a function of global TSV placement. The TSV pair is modeled as a two-port network and the insertion loss (S_{21}) of the network is used to measure the coupling between two TSVs. During this analysis all other TSVs are modeled as faraday shields by connecting them to ground. FEM network analysis is performed to generate a library of golden data, from which a set of increasingly complex analytical models are fit.

3) TSV Coupling Model Consists of Three Parts: the contribution of the two cross coupled TSVs (ignoring all shielding) S_0, the coupling reduction attributed to each shield TSV (assuming it is the only shield) S_i and the shielding overlap between each pair of shields $M_{i,j}$. The full cross coupling magnitude would be the sum of S_0 and S_i for each shield i minus the pairwise overlap $M_{i,j}$ for each pair of shield TSVs. This model was shown to accurately model TSV-TSV S-parameters for multiple shields with arbitrary placement orientation [92].

The coupling of two TVS (S_0) was found to be well modeled by a negative exponential function of the distance between them. Likewise the contribution of each shield TSV (S_i) was modeled as a negative exponential function of the distance between the shield and the coupling path (the line connecting the coupled TSV pair). Finally, the coupling overlap between two shields ($M_{i,j}$) was found to be a linear function of both distance between the shields, and distance from each shield to the coupling path.

4) TSV Stress Model: Due to the CTE mismatch between copper (a common TSV material) and the silicon substrate, stresses are induced during the TSV manufacturing process. Ryu *et al.* [20] developed a semianalytical TSV stress model. The thermal mechanical stress value around a TSV is a function of material properties (the Young's modulus, CTE, and Poisson's ratio), the thermal load, and the distance from the measured point to the center of the TSV.

When multiple TSVs are present, researchers have assumed the TSV and silicon substrate are linearly elastic structures [41], [165]. According to the stress superposition principle, the stress coming simultaneously from several different bodies is the sum of the stress applied separately. This means that the stress value at a certain point is the accumulated stress caused by each TSV.

5) TSV Electromigration Model: Electromigration is the phenomenon whereby atoms migrate overtime, forming hillocks or voids, which eventually results in short circuit or open circuit. The classic Black's [145] equation is based on the empirical observation that the interconnect mean time to failure is inversely proportional to current density.

Besides the classic Black's equation, some recent works have shown many other forces other than current may drive electromigration, such as stress gradient, thermal gradient, and atomic concentration gradient. The physics of electromigration can be described by time-dependent multiphysics mass transportation equations [39], [40], [70], [71].

One important aspect regarding electromigration is to develop an accurate model for ac signals. Many experimental studies [166], [167] have shown that when under bidirectional current stress, the EM damage caused by electron wind of one polarity can be partially healed by the other polarity. In cases where pure ac current is applied and the average current of both polarities are of the same magnitude, the measured interconnect lifetime ranges from 30× to over 1000× comparing to dc signals with the same magnitude [166], [167]. Liew *et al.* [166] proposed a vacancy relaxation model to fit the experimental data. The average current recovery model [167] uses Black's [145] equation with the effective current density, which averages positive and negative current pulses.

6) TSV-Adjacent Gate Timing Model: TSV fabrication process involves high-temperature process (e.g., annealing and bonding), and after cooling down to room temperature, copper contracts more than silicon, and causes tensile stress in the surrounding silicon. Stress causes hole and electron mobility variations, and may cause timing violations if gates on the critical path are placed next to TSVs.

One intuitive method to decouple TSV-induced stress on carrier mobility variation is to apply keep-out-zone (KOZ) near TSVs (i.e., gates are not allowed to be placed inside KOZ). For example, Okoro *et al.* [168] reported that for a 5 μm diameter TSV, 1.5 μm keep-out distance is required for avoiding significant pMOS transistor mobility variation.

Yang *et al.* [160] further indicated that tensile stress enhances electron mobility, however, tensile stress either enhances or degrades hole mobility depending on TSV and channel direction. In other words, TSV stress has greater influence on nearby pMOS transistors on critical path and may cause timing violations. Yang *et al.* [160] considered the impact of TSV stress on the mobility variation in timing analysis. Furthermore, the TSV stress is modeled in a TSV stress-driven global placement algorithm [169]. The algorithm utilizes TSV induced stress that improves standard cell timing characteristics to minimize the total negative slack.

V. 3-D ARCHITECTURAL OPPORTUNITIES

3-D stacking offers many new opportunities for high performance CPU architectures. The memory wall [9] (Fig. 16) is a known hurdle to future performance scaling, and 3-D integration is a promising technology to overcome it. Stacked memory is already in commercial production and heterogeneous memory-on-logic CPUs are being aggressively researched and prototyped [11], [49]. Moreover, data movement overheads in both power and delay have become more and more significant as we have entered the age of big data. 3-D integration enables new solutions, such as processing-in-memory [15].

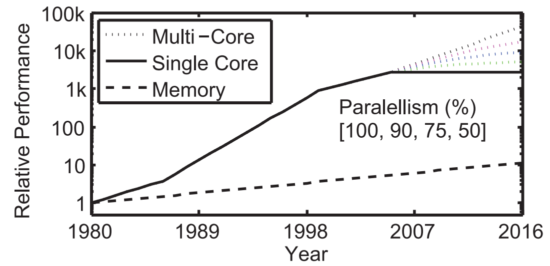


Fig. 16. Memory wall [170]: gap between processor and DRAM latency is a performance bottleneck.

A. 3-D Memory

As logic components continue to scale toward higher density and lower delay, the performance and power gaps between logic and memory continue to widen. This so called memory wall [9] is a key obstacle in the climb toward next generation computing: both mobile and exascale supercomputing. 3-D integration is an enabling technology to further the three memory design goals: higher density, higher bandwidth, and lower power. Vertical stacking inherently increases memory density within a fixed footprint area, and heterogeneous integration facilitates high speed, and/or very wide TSV memory buses which dissipate considerably less power than their off-chip counterparts.

Two main strategies have been employed toward bringing 3-D memory into the commercial market. One focuses on speed using very high speed differentially signaled serial interconnects. Although this strategy increases absolute power, the power efficiency (bandwidth per Watt) is much improved. An example of such an architecture is Micron's hybrid memory cube (HMC) [11]. Alternatively a wide parallel bus can be pursued taking advantage of the high interconnect density offered by TSV technology [11]. This strategy can significantly increase memory bandwidth without increasing power, or alternatively provide very low power operation at nominal performance. An example of such an architecture is Samsung's Wide-IO DRAM [49].

The wide-IO memory architecture consists of four independent channels each with a 128-bit data bus. Each channel contains four 64-MB arrays, for a total capacity of 1 GB per layer. The wide-IO memory can deliver peak bandwidth up to 12.8 GB s^{-1}, 4× higher than the equivalent LPDDR2 device, while increasing bandwidth per Watt of I/O power by more than 10× [49]. The wide-IO 2 specification has been released by JEDEC and makes many significant improvements [171]. The number of channels can be increased from 4 to 8, the density ranges from 8 to 32 GB and the peak bandwidth tops out at 34 (four channel) or 68 (eight channel) GB s^{-1}. Moreover, the operating voltage is reduced from 1.2 to 1.1 V, providing even lower power. Wide-IO 2 is expected to surpass the performance of LPDDR4 in 3-D stacked devices [171].

Wide-IO memory is intended to be integrated directly on top of logic using TSVs. This approach is ideal for density and power, but has thermal implications as discussed in Section III-D. Wide-IO is expected to be used in high-end smartphones, but in the absence of embedded active cooling

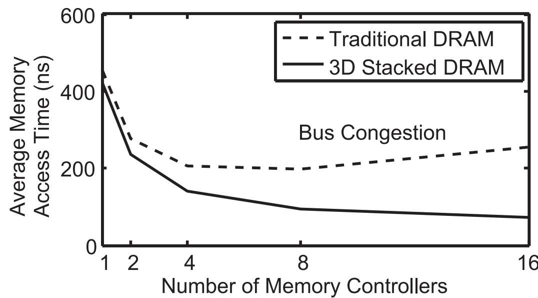


Fig. 17. Average DRAM latency versus number of MCs in a 16-core CMP [33].

schemes may not be thermally feasible in a server or supercomputer environment [172].

The HMC from Micron is connected to the CPU through a board-level high speed differential serial interface [11]. However, the cube itself is composed of stacks of DRAM on top of a layer of CMOS. This heterogeneous integration allows for optimized common logic circuits such as decoders and memory controllers (MCs) while maintaining the memory density characteristics of stacked DRAM. HMC facilitates a distributed architecture called "far" mode [11], where multiple HMCs are connected together to form a memory network for scalable high capacity memory systems. HMC moves the MC to the DRAM module itself rather than the core in order to efficiently realize such a scaled architecture.

The HMC significantly improves DRAM latency by reducing MC queuing delays and providing more memory parallelism through independent bank operation. Experimental data from first generation HMC prototype reports DRAM bandwidth of $128\,\mathrm{GB\,s^{-1}}$ while dissipating 11 W, improving bandwidth per Watt more than 3.5× over DDR4 [11].

Analysis by TSMC [172] shows that Wide-IO 2 brings the best of both worlds by providing performance parity with DDR4 while matching LPDDR4 in power dissipation. On the other hand, the HMC is a uniquely new memory architecture that pushes performance, power and price to new extremes.

B. Memory Over Logic

Heterogeneous 3-D integration can provide significant bandwidth improvements between CPU core logic and memory. Non-CMOS technologies such as DRAM, phase-change RAM (PRAM), and magnetic RAM (MRAM) [173] can be stacked directly on top of logic cores. TSVs provide much lower latency than off-chip interconnects and provide high density interconnects, facilitating wider buses and parallel memory access using multiple MCs [9], [174]. Fig. 17 shows how average memory access time can be significantly reduced by adding MCs in a 3-D stacked DRAM architecture [33]. Due to off-chip bandwidth constraints 2-D CPUs will not benefit from more than a few MCs, whereas stacked memory processors get monotonic (albeit diminishing) speedup as more MCs are added. Of course the performance benefits of adding MCs must be balanced against the associated power and area overheads. Studies have shown that the performance improvements due to main memory stacking can be up to 2× [9], [33].

The capacity of on-chip DRAM is limited to only a few GB [15], [17]. Thus most computing systems require both on and off-chip DRAM. On-chip DRAM could be leveraged as cache or a nonuniform memory access (NUMA) paradigm can be applied [175] to manage both on and off-chip DRAM as a unified main memory. Even within a stacked DRAM module nonuniform access constraints may need to be applied due to nonuniform power delivery capacity in the 3-D stack [176]. Such NUMA systems require memory swap controllers to keep hot memory pages in low-latency portions of the memory [175], [176].

Studies have shown the effectiveness of using stacked DRAM for additional cache rather than main memory. DRAM cache can offer large capacity compared to an SRAM cache of the same area [177] while maintaining higher bandwidth and lower latency compared to main memory [178]. Moreover, hot page migration into a DRAM cache can be done at the cache line granularity whereas NUMA stacked memory systems must swap memory at the page granularity, which is both inefficient and requires OS support [175].

However, there are two main limitations to DRAM cache.

1) The tag array would be unreasonably large for standard (e.g., 64 MB) cache line sizes.
2) Off-chip main memory cannot provide the necessary bandwidth to use significantly larger cache line sizes.

Jiang *et al.* [178] proposed a hot-page filtering technique to efficiently manage the DRAM bandwidth to leverage performance improvements of up to 25% from a 128 MB DRAM cache. Loh [177] leveraged the DRAM row buffer hardware to further increase DRAM cache performance by 29% by employing an adaptive multiqueue policy. On the other hand, Chou *et al.* [175] presented a low overhead technique that allows NUMA stacked memory to achieve cache-line level data migration, outperforming both DRAM cache and traditional NUMA stacked memory.

C. Processing in Memory

The age of big data has brought renewed interest to the topic of processing-in-memory. Communication had begun to dominate computation in both time and power [179] requiring new architectures to bring data closer to processing nodes. Studies have shown that big data applications are more sensitive to memory bandwidth than capacity, and in-memory storage (as opposed to disk storage) has been shown to offer lower cost and higher performance for such applications [180]. The additional memory bandwidth offered by 3-D stacked memory-on-logic can be an enabling technology for realizing the processing-in-memory paradigm.

HMCs [11] offer five tiers where the top four are DRAM, and the bottom tier contains logic circuitries for MC, I/O, and error-detection-and-correction schemes [11]. It is this logic tier that several research studies [15], [181]–[185] attempt to add computing resources for near-memory computing. The main benefit is that this logic tier is directly bonded with four tiers of DRAM dies, so memory bandwidth and access latency are both boosted significantly. The potential integration of an HMC on top of low-power processor cores is also studied

in [15] by running a set of MapReduce workloads. MapReduce workloads are a good candidate for 3-D stacked processing-in-memory because the map phase exhibits extreme levels of memory locality and execution parallelism whereas the reduce phase requires high-bandwidth random memory access [15].

Pugsley *et al.* [15] compared the performance of a traditional HMC system (i.e., off-chip shared memory bus between a multicore CPU and a cluster of HMCs) to a stacked memory on logic architecture where each core has a dedicated stacked memory slice integrated above it. The stacked memory system is able to outperform the traditional architecture because it is not constrained by HMC link bandwidth during the map phase. Moreover the total energy consumed is reduced drastically. Energy reductions range from 28% to 93% while power dissipation remains within thermally feasible limits.

VI. 3-D IC CO-DESIGN

In the previous sections, we have discussed the physical design challenges and the architectural opportunities of 3-D integration. Traditionally the physical and architectural designs are performed independently in sequence using different levels of abstraction. Moreover, even within the physical design domain, the previously discussed design problems are tackled sequentially, and cross-domain optimizations are not usually considered. A new paradigm which integrates the computational, electrical, physical, thermal, and reliability views of the system is gaining steam. This unification of diverse aspects of the overall integrated system is called co-design. Co-design enables optimizations across different layers of the design hierarchy which are not possible through a conventional top down design approach thereby unlocking new high performance configurations.

In the remainder of this paper, we present a study to exemplify the interdependence of the physical and architectural design spaces of a 3-D CPU. We use a novel simulation flow which integrates placement, temperature and reliability design challenges into a unified framework for architectural-physical optimization and analysis.

A. 3-D CPU Design Space Exploration

We perform a study involving 3-D memory-over-logic processor DSE subject to physical constraints. The design spaces considered consist of architectural parameters, floorplan topology and heatsink design. Constraints are imposed on maximum temperature and system lifetime. The optimization metric of interest is performance, measured in instructions per nanosecond. Fig. 18 illustrates the cause and effect relationships from the considered design variables to the optimization and constraint metrics of interest. The figure clearly illustrates the interdependence between the terminal and intermediate nodes, and no metric of interest can be determined without simultaneous consideration of all design variables. The gray edges and nodes in the relationship graph are not included in the study presented here, but are avenues for improvement in future work.

Furthermore, we observe that the relationship graph contains cycles, which imply nested loops within a simulation flow. An

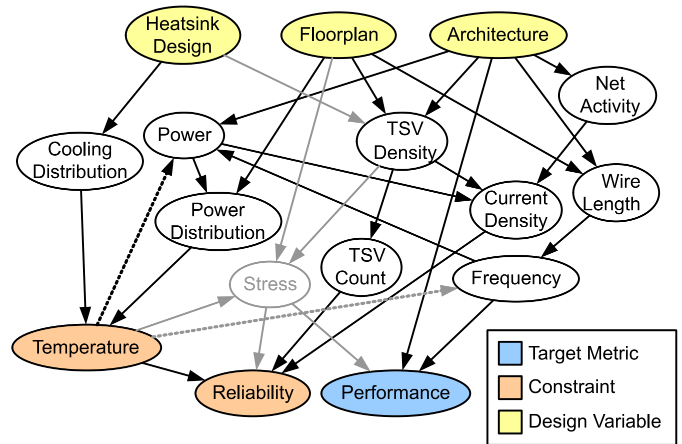


Fig. 18. Relationship graph for 3-D CPU metrics and design variables (gray edges and nodes not modeled in this paper).

example is the interdependence of temperature and leakage power. Leakage power increases as temperature elevates, and likewise temperature will rise when leakage power increases. Iterative simulations are required to accurately capture such interdependencies. Co-design DSE is a computationally intensive problem due to both optimization loops and nested simulation loops. In the following section, we present a simulation flow to capture the interdependent relationships shown in Fig. 18.

B. Simulation Approach

The simulation flow used to evaluate the 3-D CPU design space explored in the presented study is shown in Fig. 19. This flow provides an approach for design optimization considering the interdependencies shown in Fig. 18. In this paper, power, area, and timing estimates are all performed assuming the 45-nm technology node. However, other studies cited and discussed throughout this keynote paper were performed across many different technology nodes and assumptions. Modeling and optimization details of the simulation approach shown in Fig. 19 can be found in [17], [64], [65], and [186].

The architectural design space is explored by exhaustive simulation across all combinations of architectural variables of interest: number of processor cores, number of MCs, and clock frequency. Floorplan optimization is achieved using an internal optimization loop within the simulation flow to explore the possible floorplan topologies of a single core. Reliability, temperature, and performance metrics change as the floorplan design space is explored. The implications of ignoring some or all of these design metrics during optimization are reported in the results of the study. Finally, the heatsink design space is explored considering air cooling versus MF cooling and the placement of microchannels.

C. Results

The described simulation framework is applied across an architectural design space. The number of cores (16 or 32), number of MCs (2, 4, 8, or 16) and the target clock frequency (2.4, 3.0, or 3.6 GHz) are swept, and the highest performing

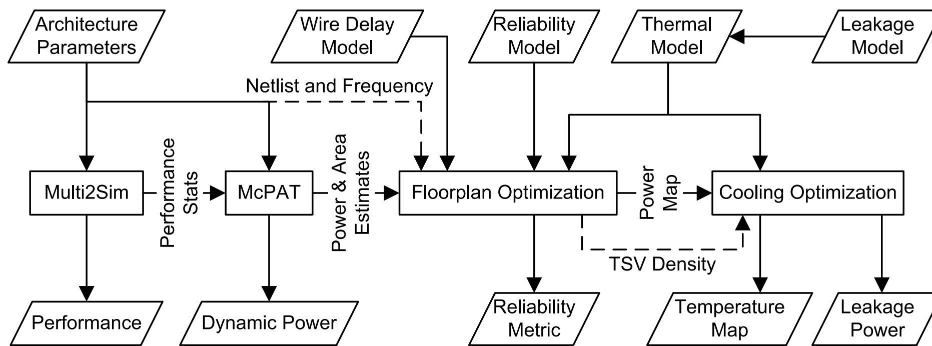


Fig. 19. Simulation flow.

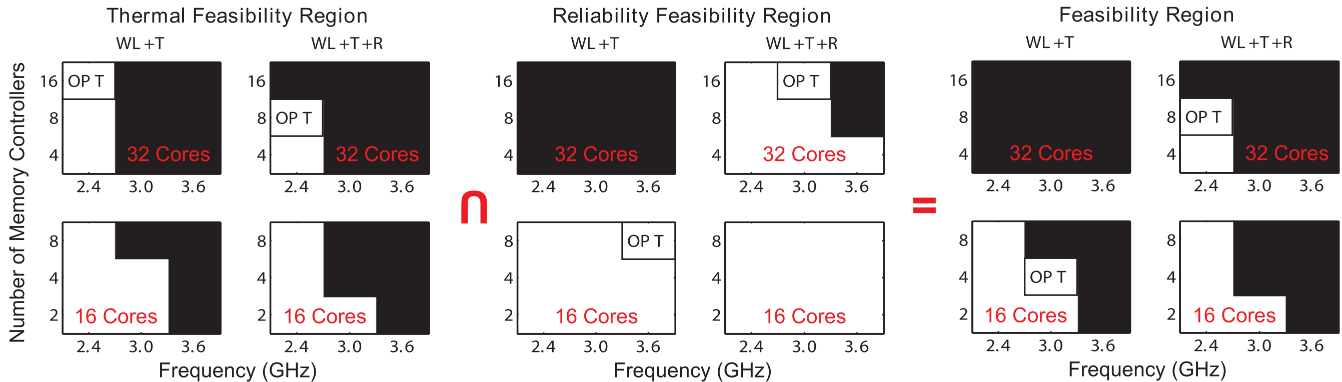


Fig. 20. Feasibility regions (shown in white) of architectural design space with highest-performance feasible architecture notated by "OPT."

thermal-reliability-feasible architecture is identified. Fig. 21 illustrates the normalized performance of the design space, evaluated over a set of parallel benchmarks from Splash-2 [187] and PARSEC [188] benchmark suites. Temperature and reliability are analyzed across the architectural design space using different floorplan objective functions and heatsink design schemes.

1) Core Stacking: In this paper, average power dissipation per core varies across the architectural design space between 5 and 7.5 W. Core power can vary significantly as memory bandwidth (i.e., number of MCs) or clock frequency is increased. This paper places 16 cores per 3-D IC layer, and considers both a 16-core (single-logic-layer) and 32-core (two-logic-layers) chip. Our chip footprint is roughly 400 mm^2 yielding average power densities of less than $30\,\text{W}\,\text{cm}^{-2}$ in the single-logic-layer chip, and roughly double that in the two-logic-layer chip. The core stacking used in our two-logic-layer chip is generally thermally infeasible with traditional air cooling [17], but our results show that MF heatsinks can potentially provide sufficient cooling to allow vertical core stacking in certain architectural configurations.

Hotspot locations can limit the feasibility of core stacking, which motivates the increased need for thermally aware physical design. Avoidance of vertical hotspot colocation is paramount when attempting to vertically stack cores [17]. The improvements to chip performance (partially due to increased core stacking) achieved by thermal aware floorplanning are visible in Fig. 22 by comparing the first two bars on the left.

2) Feasibility Region: Fig. 20 shows the feasibility region of the design space. Feasible architectures are shown in white,

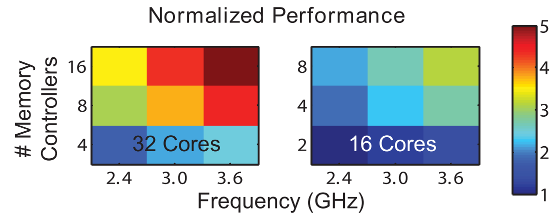


Fig. 21. 3-D CPU design space performance. Thermal-reliability feasibility of each design point depends on choice of floorplan and heatsink.

infeasible architectures are shown in black and the highest performing feasible architecture is marked with "OPT." The thermal and reliability feasibility regions are evaluated separately and their intersection defines the true thermal-reliability feasibility region. Thermal feasibility is defined as maximum on-chip temperature less than $T_{\text{violation}} = 85$ °C. Reliability feasibility was defined as $P_{\text{fail}}(t_{\text{target}}) < \alpha$, where $\alpha = 99\%$ is the reliability confidence and $t_{\text{target}} = 3$ years is the lifetime target. Two floorplan objective functions are considered. The first only includes wirelength and temperature (WL + T), whereas the second also includes reliability (WL + T + R). The results in this figure assume MF cooling with uniform microchannel placement.

Looking at the thermal feasibility region, we observe that the addition of reliability to the floorplan objective function causes the thermal feasibility region to contract, resulting in reduced optimal performance. However, considering the reliability feasibility region, the addition of reliability to the floorplan objective significantly expands the feasibility region,

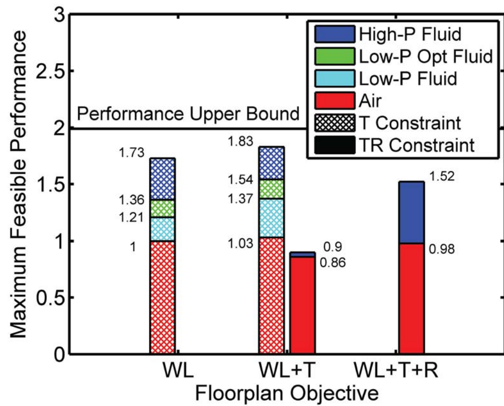


Fig. 22. CPU performance using different cooling and floorplanning co-design techniques. Floorplanning objectives include: wirelength only (WL) thermally aware (WL $+ T$) and thermal-reliability aware (WL $+ T + R$).

increasing the optimal performance significantly. This shows the potential tradeoff between temperature and reliability in 3-D CPUs. Although increased temperature does increase the probability of failure of a single TSV, it is quite possible that thermally optimized floorplans contain more 3-D nets (i.e., more cuts in the interlayer partition) in order to optimize the distribution of power. In some cases the increase in number of TSVs will outweigh the reduction in temperature when considering the net effect on system reliability. Overall, when the thermal-reliability feasibility region is considered as a whole, reliability-aware floorplanning does indeed significantly expand the total feasibility region, and result in significantly improved performance.

3) Optimal Performance: The optimal feasible performance of the investigated architectural design space is plotted in Fig. 22. Three floorplan objectives are used to generate the data, each one adding an additional term to the objective function. The data is obtained from different studies that used two different feasibility constraints. The first study [17], [64] (the two data bars on the left) used only a thermal constraint and evaluated the improvement offered by thermally aware floorplanning. The second study [65] (the two data bars on the right) further imposed a reliability constraint and studied the additional performance improvements offered by reliability aware floorplanning.

The unconstrained performance of the design space is notated as an upper bound. Likewise, four different cooling schemes are considered: 1) high-pumping-power uniform MF cooling; 2) low-pumping-power optimized MF cooling; 3) low-pumping-power uniform MF cooling; and 4) traditional air cooling. Low-pumping-power MF cooling uses 5× less pumping power, and optimized MF cooling uses the microchannel placement optimization technique described in Section VI-B.

Comparing the first (leftmost) two bars in the figure, we can see that thermal-aware floorplanning improves thermally feasible performance between 3% and 13% depending on the cooling method applied. Additionally one can observe that none of the considered cooling techniques are able to thermally unlock the entire design space, and the improvement in performance due to increasing MF cooling power 5× is less than 2×. Finally microchannel placement optimization can provide significant performance improvements while maintaining a constant pumping power, thus greatly increasing the power efficiency of the MF heatsink.

Comparing the middle two bars we observe that the expansion of the thermal feasibility region provided by MF cooling becomes a moot point when reliability feasibility is included. However, by comparing the last (rightmost) two bars we see that reliability-aware floorplanning can once again unlock the performance potential of MF cooling. Reliability feasibility does not significantly affect the potential performance of an air-cooled 3-D CPU since the architectural design points which would benefit from the expanded reliability feasibility region are still thermally infeasible. The conclusion is that aggressive cooling is required to thermally unlock 3-D CPU performance, but must also be accompanied by thermal-reliability aware physical design to realize the potential gains brought by the new cooling technology.

VII. Monolithic 3-D IC

M3-D IC is a vertical integration technology that builds up two or more tiers of devices sequentially [189], rather than bonding two parallelly fabricated dies together using bumps and/or TSVs. The key enabler is the nano-scale monolithic intertier via (MIV). The small MIV has small parasitic capacitances and overcomes the well-known shortcomings of micrometer-scale TSV including area overhead, die alignment precision, and multiphysics reliability issues. Most importantly, MIVs offer the possibility of more fine grained intertier integration compared with TSVs, which opens doors to a wide range of high-performance applications including logic/memory integration, brain-inspired computing, ultralow power mobile computing, etc.

A. Monolithic 3-D IC Manufacturing

The M3-D fabrication process flow of CEA/LETI for 2-tier is summarized in Fig. 23. Transistors on the bottom layer are fabricated with a classical thermal budget, using rapid thermal annealing at 1050 °C for dopant activation [Fig. 23(a)]. Next, a thermal oxide layer is grown on an empty wafer and hydrogen ions are implanted below the oxide surface to provide a cleave boundary. The empty wafer is attached to the existing monolithic tier using low temperature (200 °C) molecular bonding. Excess silicon is cleaved at the hydrogen boundary and the remaining thin silicon surface is polished by CMP [Fig. 23(b)]. Because tier bonding occurs before device fabrication on the top layer, the alignment issue common in TSV processes [190] is inherently nonexistent in M3-D ICs [191], [192].

Device fabrication on the stacked layers is thermally limited to certain temperature to avoid damaging previously fabricated devices and interconnects sitting below. This is a fundamental manufacturing challenge unique to M3-D ICs [193]. For the top layer, silicon substrate is grown and dopant is activated using low temperature (650 °C) solid phase epitaxial method, and gate dielectric is deposited using atomic layer

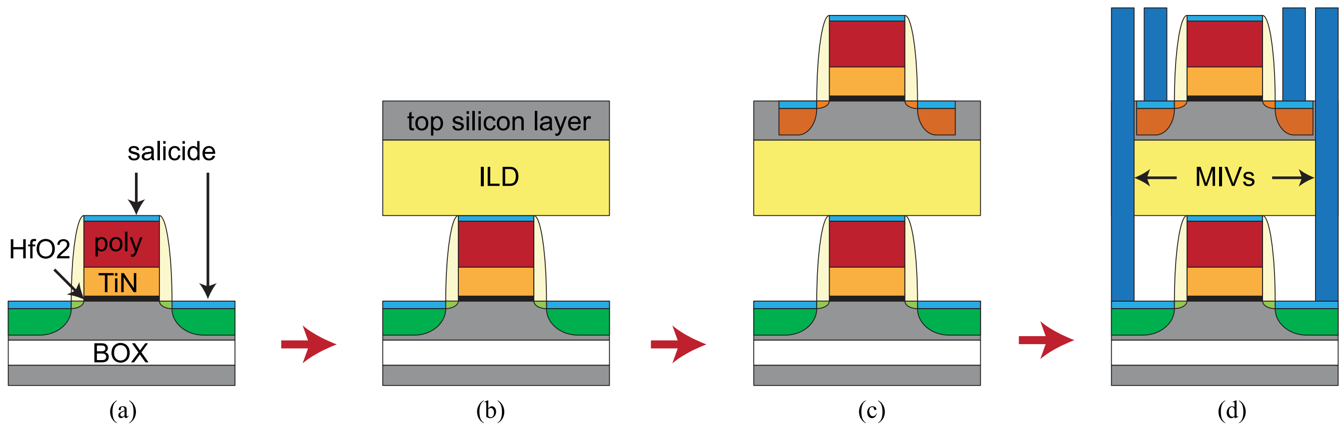


Fig. 23. (a)–(d) M3-D fabrication process flow of CEA/LETI.

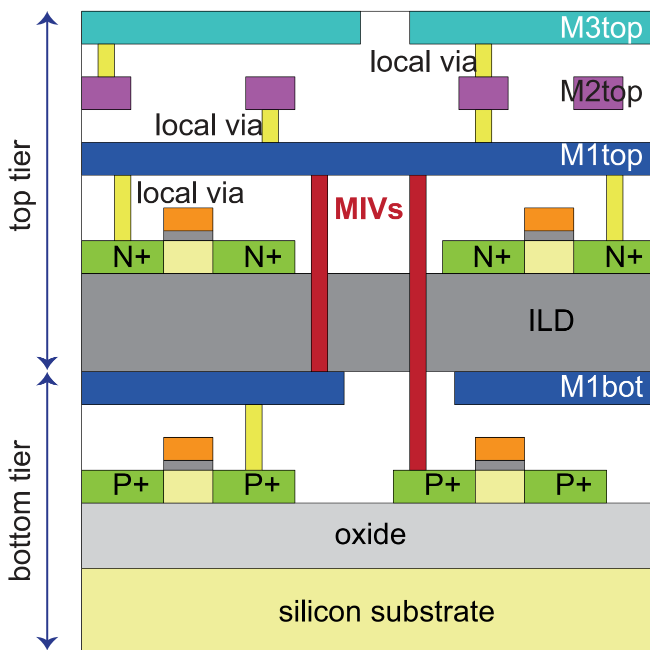


Fig. 24. Our target 2-tier M3-D stack-up.

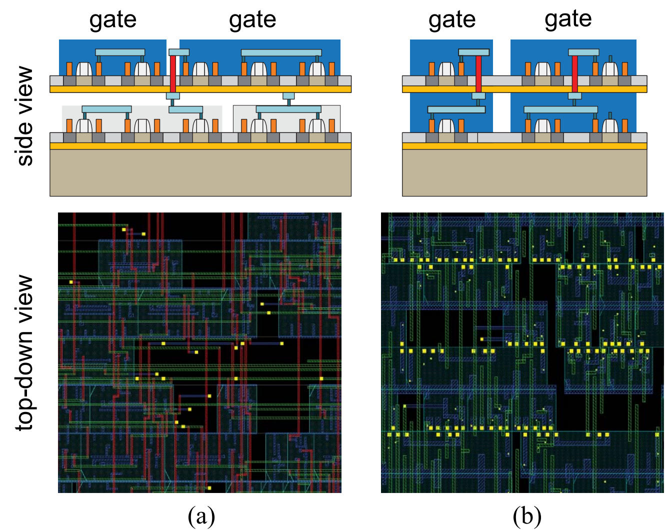


Fig. 25. M3-D IC design granularity. (a) Gate-level. (b) Transistor-level (finest). Yellow dots denote MIVs.

deposition at 350 °C, followed by thermal annealing at 515 °C [Fig. 23(c)]. Finally the contacts to the top/bottom layer transistors are fabricated. A single lithography step uses a highly selective etch to open contacts down to bottom layer transistor [Fig. 23(d)]. A two-tier M3-D stack-up is shown in Fig. 24.

In 2014, researchers from Stanford University [194] built a field-programmable gate array routing switch using a 4-tier M3-D IC, where a CMOS transistor, two RRAM devices, and a carbon nanotube transistor are vertically stacked from bottom to top. In 2015, researchers from the National Nano-Device Laboratory of Taiwan built a 6T SRAM cell in a 2-tier M3-D IC using 50-nm transistors in an ultrathin body channel (20 nm) [195]. They showed that laser annealing, combined with a relatively thick ILD between the tiers (300 nm instead of 23 nm in LETI's technology) allows comparable performance between the tiers.

B. Monolithic 3-D IC Design Granularity and Tools

There are two possible levels of design granularity for M3-D implementation: gate-level and transistor-level M3-D as shown in Fig. 25. In gate-level M3-D, individual gates occupy a single tier and their place/route are done in multiple tiers. In transistor-level M3-D, the devices in a single gate are placed in multiple tiers and connected using MIVs. For a given design, it is possible to choose between these two styles that offer trade-offs in design reuse, MIV usage, CAD tool requirements, and PPAR quality.

Bobba *et al.* [196] proposed two different strategies of stacking standard cells in M3-D designs: intracell stacking (i.e., transistor-level M3-D) and cell-on-cell (i.e., gate-level M3-D) stacking. Intracell stacking requires the modification of standard-cell design but permits a direct reuse of 2-D IC tools. In case of cell-on-cell stacking, the authors proposed a placement tool based on commercial tools and an LP formulation for tier assignment. Lee *et al.* [197] presented physical design techniques for transistor-level M3-D ICs. They first built a cell library that consists of 3-D gates, and model their timing and

power characteristics. They performed iso-performance comparisons, and demonstrated significant power benefit over 2-D. They also demonstrated that this benefit increases at future technology nodes.

Panth *et al.* [198] presented a breakthrough in physical design for gate-level M3-D. The authors showed how to use commercial 2-D IC CAD tools to build M3-D designs that outperform commercial 2-D IC designs for the first time. These techniques place the gates into half the footprint area of a 2-D IC using a commercial 2-D engine. Next, the gates are partitioned into multiple tiers to give high-quality 3-D solutions. The authors also presented techniques to utilize the commercial tool for timing optimization, clock-tree-synthesis, and intertier via insertion. Chang *et al.* [199] demonstrated the power benefits of M3-D at 28-, 16/14-, and 7-nm nodes. Their study is based on the physical design tool from [198] along with foundry PDKs and commercial designs. They showed that 28 nm offered the best power reduction because of the large gate capacitance in FinFET devices.

Ku *et al.* [200] studied M3-D cost and yield requirements so that the power benefits exceed the cost overhead of M3-D at 7-nm node. The study showed that the total BEOL metal layer usage is one of the key cost factors that must be minimized to obtain convincing power, performance, and cost improvement over 2-D ICs at 7-nm node. Samal *et al.* [201] presented design strategies to cope with BEOL degradation in M3-D.

In summary, M3-D IC technology has shown promising performance and power benefits in fine-grained design in advanced nodes, and recent studies have shown potential solutions for M3-D ICs thermal and BEOL degradation problems. More work still needs to be done at the design tool as well as fabrication levels.

VIII. Open Problems in 3-D IC Design Tools

3-D integration is a promising and exciting technology development which has already seen commercial application in the form of 3-D memory [11], [202], [203]. However, there are still significant open problems in both research and implementation. Further investment in both industry and academia is required to tackle these problems and realize the potential improvements discussed in this keynote paper.

A. Large-Scale EDA Tools

Although many EDA problems have been investigated in the literature, they are often implemented as hacks and splices of existing commercial 2-D EDA tools and custom scripts to perform 3-D analysis. These approaches suffice to show the potential of different modeling techniques and algorithms, but do not provide a realistic implementation for large-scale 3-D design automation. The development of integrated, scalable 3-D design tools is a must for 3-D stacked logic technology to become ubiquitous in the market, and will require significant commercial investment that has yet to materialize. However, as more traditional areas of investment (e.g., technology scaling) become less profitable, we expect to see a surge in 3-D IC investment in the near future.

B. Architectural Modeling

3-D integration technology brings the opportunity for new computer architectures; however, such drastic changes to the conventional computing paradigm require new architectural models of 3-D CPU performance, power, area, and timing (PPAT). The PPAT modeling challenges mainly come from significant changes in the types and organization of memory on chip, and the effects of fined-grained vertical integration on CPU subcomponents.

Stacked memory architectures have significantly different memory hierarchy topologies due to more fine grained integration with TSV technology. CPU-DRAM communication may take place over multiple independent communication channels which could be point-to-point, bus or a hybrid of both [17]. Each communication channel can be wider and/or clocked faster using high-density low-impedance on-chip interconnects. PPAT simulators must be configured to model the power and performance of such unconventional memory hierarchies. Moreover heterogeneous integration facilitates on-chip cache and/or main memory technologies such as DRAM, MRAM, and PRAM, all of which require complex MC designs [173]. Models of these technologies and their controllers must be incorporated into conventional architectural tools. Finally memory-on-chip integration could facilitate a re-emergence of large parallel interfaces as opposed to high speed serial communication for low-power designs [49]. The whole spectrum of interface implementations must have available models within a 3-D PPAT simulator for proper tradeoff analysis.

One of the main advantages of 3-D integration is the reduction to wire length due to fine grained integration. Power, delay and area estimates for circuits with regular structure (e.g., memory elements) can be estimated analytically using technology and topology parameters [204] (although 3-D implementation significantly increases the design space of the topology parameters to be considered). However, highly complex and customized circuits (e.g., ALUs) are hard to estimate analytically. For 2-D CPU analysis, empirical models that fits the real CPU data in the market have been used [205]. Since 3-D CPUs are still in the research and development stage, similar data does not exist. Developing models for 3-D function unit PPAT is a challenging and open problem.

C. 3-D IC Cooling

A significant amount of work has been done to investigate the thermal issues and cooling mechanisms in 3-D ICs, which include liquid cooling (such as MF cooling [32]), solid cooling (such as TTSV [81]), and thermoelectric cooling [206]. However, it is still an open question which cooling mechanism suits the best for 3-D ICs. Despite the fact that the embedded MF cooling has attracted more interests, there are concerns regarding the fabrication and reliability of these systems. For instance, many coolants (including water) can cause erosion and corrosion of the channel wall in microchannel cooling, which harms the circuit reliability and changes the cooling capacity [125]. More research should be done in selecting proper coolant for sustained reliability.

D. Multicorner Design for 3-D ICs

Manufacturing process corners have tremendous impacts on modern ICs' speed and power. This imposes significant challenges for 3-D IC timing closure, as different layers are manufactured separately, causing interlayer process variation. Any timing path that spans across layers will experience multiple distinct process corners. Minimizing clock skew variation across corners is so far the most prevailing way to ensure timing closure in multicorner designs. There are analysis and optimization papers for 2-D ICs' multicorner design. Nontree clock structures [207] (a top-level clock mesh connected to local clock trees), crosslink insertions [208] (in a clock tree), and buffer insertion/sizing [209], etc. have proven to be effective methods to handle clock skew variations in 2-D ICs. Since 3-D ICs benefit from existing technologies rather than relying on aggressive device scaling, researches have not fully demonstrated the effect of multiple corners on 3-D ICs' timing. Nevertheless, in the near future, modeling and optimizing multicorner design for 3-D ICs will certainly become a challenging and attractive research problem.

IX. Conclusion

3-D ICs have shown promising improvements in performance and energy efficiency independent of costly transistor scaling. However, the expanded design space brought on by 3-D integration imposes extra design complexities to the physical design domain, including partitioning, placement, clock tree synthesis, etc. Furthermore, the stacking structure inevitably increases voltage drop and on-chip temperature, which complicates thermal and power delivery design and management. Vertical vias introduce new sources of cross-coupling and reliability degradations. In this paper, we review the state-of-the-art solutions for the aforementioned challenges, and present a study showing the further performance improvement brought by the application of a co-design scheme. As design objectives in 3-D ICs are highly interdependent, co-design becomes necessary to fully exploit the true potential of 3-D integration in the future. Finally, we explore the mainstream fabrication, design tool challenges and recent research development of M3-D ICs and briefly discuss the remaining open problems which in our opinion will make significant progress toward large-scale commercial adoption of 3-D IC technology in the near future.

References

[1] A. Shah. (May 2013). *Intel: Keeping Up With Moore's Law Is Becoming a Challenge*. [Online]. Available: http://www.pcworld.com/article/2038207/intel-keeping-up-with-moores-law-becoming-a-challenge.html

[2] N. Z. Haron and S. Hamdioui, "Why is CMOS scaling coming to an END?" in *Proc. 3rd Int. Design Test Workshop*, Monastir, Tunisia, Dec. 2008, pp. 98–103.

[3] S. Tyagi, "Moore's law: A CMOS scaling perspective," in *Proc. 14th Int. Symp. Phys. Failure Anal. Integr. Circuits*, Bengaluru, India, Jul. 2007, pp. 10–15.

[4] K. Mistry *et al.*, "A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, USA, Dec. 2007, pp. 247–250.

[5] H. Gottlob *et al.*, "CMOS integration of epitaxial GD_2O_3 high-k gate dielectrics," *Solid-State Electron.*, vol. 50, no. 6, pp. 979–985, 2006.

[6] J. Lee *et al.*, "Compatibility of dual metal gate electrodes with high-k dielectrics for CMOS," in *IEEE Int. Electron Devices Meeting Tech. Dig. (IEDM)*, Washington, DC, USA, 2003, pp. 13.5.1–13.5.4.

[7] T. Osada and M. Godwin, "International technology roadmap for semiconductors," ITRS, Tech. Rep., 1999.

[8] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *Proc. Int. Workshop Syst. Level Interconnect Prediction (SLIP)*, Paris, France, 2004, pp. 7–13. [Online]. Available: http://doi.acm.org/10.1145/966747.966750

[9] G. H. Loh, "3D-stacked memory architectures for multi-core processors," in *Proc. 35th Annu. Int. Symp. Comput. Architect. (ISCA)*, Beijing, China, 2008, pp. 453–464.

[10] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints," in *Proc. 49th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Jun. 2012, pp. 648–655.

[11] J. T. Pawlowski, "Hybrid memory cube (HMC)," in *Proc. IEEE Hot Chips 23 Symp. (HCS)*, vol. 23. 2011, pp. 1–24.

[12] W. R. Davis *et al.*, "Demystifying 3D ICs: The pros and cons of going vertical," *IEEE Des. Test. Comput.*, vol. 22, no. 6, pp. 498–510, Nov./Dec. 2005.

[13] C. C. Liu, I. Ganusov, M. Burtscher, and S. Tiwari, "Bridging the processor-memory performance gap with 3D IC technology," *IEEE Des. Test. Comput.*, vol. 22, no. 6, pp. 556–564, Nov./Dec. 2005.

[14] Y. Morikawa, T. Murayama, Y. N. T. Sakuishi, A. Suzuki, and K. Suu, "Total cost effective scallop free Si etching for 2.5D & 3D TSV fabrication technologies in 300mm wafer," in *Proc. IEEE 63rd Electron. Compon. Technol. Conf.*, Las Vegas, NV, USA, 2013, pp. 605–607.

[15] S. H. Pugsley *et al.*, "NDC: Analyzing the impact of 3D-stacked memory+logic devices on mapreduce workloads," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Monterey, CA, USA, Mar. 2014, pp. 190–200.

[16] F. Li *et al.*, "Design and management of 3D chip multiprocessors using network-in-memory," in *Proc. 33rd Annu. Int. Symp. Comput. Architect. (ISCA)*, Boston, MA, USA, 2006, pp. 130–141.

[17] C. Serafy, A. Bar-Cohen, A. Srivastava, and D. Yeung, "Unlocking the true potential of 3-D CPUs with microfluidic cooling," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 4, pp. 1515–1523, Apr. 2016.

[18] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-aware 3D IC placement via transformation," in *Proc. Asia South Pac. Design Autom. Conf. (ASP-DAC)*, Yokohama, Japan, Jan. 2007, pp. 780–785.

[19] T. Zhang, Y. Zhan, and S. S. Sapatnekar, "Temperature-aware routing in 3D ICs," in *Proc. Asia South Pac. Conf. Design Autom.*, Yokohama, Japan, Jan. 2006, pp. 309–314.

[20] S.-K. Ryu *et al.*, "Impact of near-surface thermal stresses on interfacial reliability of through-silicon vias for 3-D interconnects," *IEEE Trans. Device Mater. Rel.*, vol. 11, no. 1, pp. 35–43, Mar. 2011.

[21] J. Minz, X. Zhao, and S. K. Lim, "Buffered clock tree synthesis for 3D ICs under thermal variations," in *Proc. Asia South Pac. Design Autom. Conf. (ASPDAC)*, Seoul, South Korea, Mar. 2008, pp. 504–509.

[22] Y. H. Hu *et al.*, "Process development to enable 3D IC multi-tier die bond for 20μm pitch and beyond," in *Proc. IEEE 64th Electron. Compon. Technol. Conf. (ECTC)*, Orlando, FL, USA, 2014, pp. 572–575.

[23] T. Fukushima, K.-W. Lee, T. Tanaka, and M. Koyanagi, "Advanced die-to-wafer 3D integration platform: Self-assembly technology," in *3D Integration for VLSI Systems*. Boca Raton, FL, USA: CRC Press, 2016, p. 153.

[24] Y. Xie *et al.*, "Security and vulnerability implications of 3D ICs," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 2, no. 2, pp. 108–122, Apr./Jun. 2016.

[25] K. Gopalakrishnan, A. Peddaiahgari, D. Smith, D. Zhang, and L. England, "Process development and optimization for high-aspect ratio through-silicon via (TSV) etch," in *Proc. 27th Annu. SEMI Adv. Semicond. Manuf. Conf. (ASMC)*, New York, NY, USA, 2016, pp. 460–465.

[26] C. Ababei *et al.*, "Placement and routing in 3D integrated circuits," *IEEE Des. Test. Comput.*, vol. 22, no. 6, pp. 520–531, Nov./Dec. 2005.

[27] S. Sawicki, G. Wilke, M. Johann, and R. Reis, "A cells and I/O pins partitioning refinement algorithm for 3D VLSI circuits," in *Proc. 16th IEEE Int. Conf. Electron. Circuits Syst. (ICECS)*, Hammamet, Tunisia, Dec. 2009, pp. 852–855.

[28] D. H. Kim, K. Athikulwongse, and S. K. Lim, "A study of through-silicon-via impact on the 3D stacked IC layout," in *IEEE/ACM Int. Conf. Comput.-Aided Design Dig. Tech. Papers (ICCAD)*, San Jose, CA, USA, Nov. 2009, pp. 674–680.

[29] M.-K. Hsu, Y.-W. Chang, and V. Balabanov, "TSV-aware analytical placement for 3D IC designs," in *Proc. Design Autom. Conf.*, San Diego, CA, USA, Jun. 2011, pp. 664–669.

[30] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, 2003, pp. 86–89.

[31] K. Athikulwongse, M. Pathak, and S. K. Lim, "Exploiting die-to-die thermal coupling in 3D IC placement," in *Proc. ACM Design Autom. Conf.*, San Francisco, CA, USA, Jun. 2012, pp. 741–746.

[32] J.-M. Koo, S. Im, L. Jiang, and K. E. Goodson, "Integrated microchannel cooling for three-dimensional electronic circuit architectures," *J. Heat Transf.*, vol. 127, no. 1, pp. 49–58, 2005.

[33] C. Serafy, B. Shi, A. Srivastava, and D. Yeung, "High performance 3D stacked DRAM processor architectures with micro-fluidic cooling," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, San Francisco, CA, USA, Oct. 2013, pp. 1–8.

[34] O. Guiller *et al.*, "Through silicon capacitor co-integrated with TSV as an efficient 3D decoupling capacitor solution for power management on silicon interposer," in *Proc. IEEE 64th Electron. Compon. Technol. Conf. (ECTC)*, Orlando, FL, USA, May 2014, pp. 1296–1302.

[35] N. H. Khan, S. M. Alam, and S. Hassoun, "Power delivery design for 3-D ICs using different through-silicon via (TSV) technologies," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 4, pp. 647–658, Apr. 2011.

[36] M. B. Healy and S. K. Lim, "A novel TSV topology for many-tier 3D power-delivery networks," in *Proc. Design Autom. Test Europe*, Grenoble, France, 2011, pp. 1–4.

[37] J. Pak, M. Pathak, S.-K. Lim, and D. Z. Pan, "Modeling of electromigration in through-silicon-via based 3D IC," in *Proc. Electron. Compon. Technol. Conf.*, Lake Buena Vista, FL, USA, 2011, pp. 1420–1427.

[38] T. Lu, Z. Yang, and A. Srivastava, "Electromigration-aware placement for 3D-ICs," in *Proc. Int. Symp. Qual. Electron. Design*, Santa Clara, CA, USA, 2016, pp. 35–40.

[39] M. Pathak, J. Pak, D. Z. Pan, and S.-K. Lim, "Electromigration modeling and full-chip reliability analysis for BEOL interconnect in TSV-based 3D ICs," in *Proc. ICCAD*, San Jose, CA, USA, 2011, pp. 555–562.

[40] J. Pak, S. K. Lim, and D. Z. Pan, "Electromigration study for multiscale power/ground vias in TSV-based 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 12, pp. 1873–1885, Dec. 2014.

[41] M. Jung, J. Mitra, D. Z. Pan, and S. K. Lim, "TSV stress-aware full-chip mechanical reliability analysis and optimization for 3-D IC," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 8, pp. 1194–1207, Aug. 2012.

[42] T. Lu, Z. Yang, and A. Srivastava, "Post-placement optimization for thermal-induced mechanical stress reduction," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Pittsburgh, PA, USA, 2016, pp. 158–163.

[43] J. Cho *et al.*, "Modeling and analysis of through-silicon via (TSV) noise coupling and suppression using a guard ring," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 1, no. 2, pp. 220–233, Feb. 2011.

[44] C. Serafy and A. Srivastava, "TSV replacement and shield insertion for TSV–TSV coupling reduction in 3-D global placement," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 4, pp. 554–562, Apr. 2015.

[45] C. Liu *et al.*, "Full-chip TSV-to-TSV coupling analysis and optimization in 3D IC," in *Proc. 48th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, San Diego, CA, USA, Jun. 2011, pp. 783–788.

[46] Y. Peng, T. Song, D. Petranovic, and S. K. Lim, "Silicon effect-aware full-chip extraction and mitigation of TSV-to-TSV coupling," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 12, pp. 1900–1913, Dec. 2014.

[47] S. R. Vempati *et al.*, "Development of 3-D silicon die stacked package using flip chip technology with micro bump interconnects," in *Proc. 59th Electron. Compon. Technol. Conf. (ECTC)*, San Diego, CA, USA, May 2009, pp. 980–987.

[48] M. Jung, D. Z. Pan, and S. K. Lim, "Chip/package co-analysis of thermo-mechanical stress and reliability in TSV-based 3D ICs," in *Proc. 49th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Jun. 2012, pp. 317–326.

[49] J.-S. Kim *et al.*, "A 1.2 V 12.8 GB/s 2 GB mobile wide-I/O DRAM with 4 × 128 I/Os using TSV based stacking," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 107–116, Jan. 2012.

[50] M. Pathak, Y.-J. Lee, T. Moon, and S. K. Lim, "Through-silicon-via management during 3D physical design: When to add and how many?" in *Proc. Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, 2010, pp. 387–394.

[51] M. Jung, T. Song, Y. Peng, and S. K. Lim, "Fine-grained 3-D IC partitioning study with a multicore processor," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 5, no. 10, pp. 1393–1401, Oct. 2015.

[52] W. Kim, D.-H. Kim, H. I. Hong, L. Milor, and S. K. Lim, "Impact of die partitioning on reliability and yield of 3D DRAM," in *Proc. IEEE Int. Interconnect Technol. Conf.*, San Jose, CA, USA, 2014, pp. 389–392.

[53] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Tier-partitioning for power delivery vs cooling tradeoff in 3D VLSI for mobile applications," in *Proc. 52nd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, 2015, pp. 1–6.

[54] J. Sun, J.-Q. Lu, D. Giuliano, T. P. Chow, and R. J. Gutmann, "3D power delivery for microprocessors and high-performance ASICs," in *Proc. APEC 22nd Annu. IEEE Appl. Power Electron. Conf. Expo.*, Anaheim, CA, USA, 2007, pp. 127–133.

[55] T. Lu and A. Srivastava, "Electromigration-aware clock tree synthesis for TSV-based 3D-ICs," in *Proc. 25th Edition Great Lakes Symp. VLSI (GLSVLSI)*, Pittsburgh, PA, USA, 2015, pp. 27–32.

[56] X. Zhao, D. L. Lewis, H.-H. S. Lee, and S. K. Lim, "Pre-bond testable low-power clock tree design for 3D stacked ICs," in *IEEE/ACM Int. Conf. Comput.-Aided Design Dig. Tech. Papers*, San Jose, CA, USA, Nov. 2009, pp. 184–190.

[57] X. Zhao, J. Minz, and S. K. Lim, "Low-power and reliable clock network design for through-silicon via (TSV) based 3D ICs," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 1, no. 2, pp. 247–259, Feb. 2011.

[58] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. 38th Annu. Int. Symp. Comput. Architect. (ISCA)*, San Jose, CA, USA, Jun. 2011, pp. 365–376.

[59] B. Shi, A. Srivastava, and A. Bar-Cohen, "Hybrid 3D-IC cooling system using micro-fluidic cooling and thermal TSVs," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Amherst, MA, USA, Aug. 2012, pp. 33–38.

[60] B. Shi, A. Srivastava, and P. Wang, "Non-uniform micro-channel design for stacked 3D-ICs," in *Proc. 48th Design Autom. Conf. (DAC)*, San Diego, CA, USA, 2011, pp. 658–663.

[61] P. Jain, T.-H. Kim, J. Keane, and C. H. Kim, "A multi-story power delivery technique for 3D integrated circuits," in *Proc. Int. Symp. Low Power Electron. Design*, Bengaluru, India, 2008, pp. 57–62.

[62] K. Weide-Zaage, "Exemplified calculation of stress migration in a 90nm node via structure," in *Proc. 11th Int. Conf. Thermal Mech. Multi-Phys. Simulat. Exp. Microelectron. Microsyst. (EuroSimE)*, Bordeaux, France, 2010, pp. 1–8.

[63] H. Miura and K. Suzuki, "Improvement of the long-term reliability of TSV interconnections used in three-dimensional stacked modules," in *Proc. ASME Int. Mech. Eng. Congr. Expo.*, 2014, Art. no. V010T13A077.

[64] C. Serafy, A. Srivastava, A. Bar-Cohen, and D. Yeung, "Design space exploration of 3D CPUs and micro-fluidic heatsinks with thermo-electrical-physical co-optimization," in *Proc. ASME Int. Tech. Conf. Exhibit. Packag. Integr. Electron. Photon. Microsyst.*, Brussels, Belgium, 2015.

[65] C. Serafy, T. Lu, and A. Srivastava, "Thermal-reliability physical co-optimization during architectural design space exploration of 3D-CPUs," in *Proc. GOMACTech*, Orlando, FL, USA, 2016.

[66] J. Burns *et al.*, "Three-dimensional integrated circuits for low-power, high-bandwidth systems on a chip," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2001, pp. 268–269.

[67] K. Warner *et al.*, "Low-temperature oxide-bonded three-dimensional integrated circuits," in *Proc. IEEE Int. SOI Conf.*, Williamsburg, VA, USA, Oct. 2002, pp. 123–124.

[68] R. Reif *et al.*, "3-D interconnects using Cu wafer bonding: Technology and applications," in *Proc. Adv. Metallization Conf. (AMC)*, San Diego, CA, USA, 2002, pp. 37–45.

[69] F. Laermer and A. Schilp, "Method of anisotropically etching silicon," U.S. Patent 5 501 893, Mar. 26, 1996. [Online]. Available: http://www.google.com/patents/US5501893

[70] T. Lu and A. Srivastava, "Detailed electrical and reliability study of tapered TSVs," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, San Francisco, CA, USA, Oct. 2013, pp. 1–7.

[71] T. Lu and A. Srivastava, "Modeling and layout optimization for tapered TSVs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 12, pp. 3129–3132, Dec. 2015.

[72] R. F. Toftness, A. Boyle, and D. Gillen, "Laser technology for wafer dicing and microvia drilling for next generation wafers (Invited Paper)," in *Proc. SPIE*, vol. 5713. San Jose, CA, USA, 2005, pp. 54–66.

[73] C. A. Bower *et al.*, "High density vertical interconnects for 3-D integration of silicon integrated circuits," in *Proc. 56th Electron. Compon. Technol. Conf.*, San Diego, CA, USA, 2006, pp. 399–403.

[74] Y.-H. Chen, W.-C. Lo, and T.-Y. Kuo, "Thermal effect characterization of laser-ablated silicon-through interconnect," in *Proc. 1st Electron. Systemintegr. Technol. Conf.*, vol. 1. Dresden, Germany, Sep. 2006, pp. 594–599.

[75] A. Klumpp, P. Ramm, and R. Wieland, "3D-integration of silicon devices: A key technology for sophisticated products," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Dresden, Germany, Mar. 2010, pp. 1678–1683.

[76] P. Spindler, U. Schlichtmann, and F. M. Johannes, "Kraftwerk2—A fast force-directed quadratic placement approach using an accurate net model," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, no. 8, pp. 1398–1411, Aug. 2008.

[77] (Aug. 2016). *Cadence*. [Online]. Available: https://www.cadence.com/content/cadence-www/global/en_US/home/solutions/3dic-design-solutions.html

[78] S.-C. Lin and K. Banerjee, "Cool chips: Opportunities and implications for power and thermal management," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 245–255, Jan. 2008.

[79] Y.-M. Sheu *et al.*, "Modeling mechanical stress effect on dopant diffusion in scaled MOSFETs," *IEEE Trans. Electron Devices*, vol. 52, no. 1, pp. 30–38, Jan. 2005.

[80] Y.-L. Shen, "Thermo-mechanical stresses in copper interconnects—A modeling analysis," *Microelectron. Eng.*, vol. 83, no. 3, pp. 446–459, 2006.

[81] P. Wilkerson, A. Raman, and M. Turowski, "Fast, automated thermal simulation of three-dimensional integrated circuits," in *Proc. 9th Intersoc. Conf. Thermal Thermomech. Phenom. Electron. Syst. (ITHERM)*, Las Vegas, NV, USA, 2004, pp. 706–713.

[82] N. S. Kim *et al.*, "Leakage current: Moore's law meets static power," *Computer*, vol. 36, no. 12, pp. 68–75, Dec. 2003.

[83] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[84] C. Sun, L. Shang, and R. P. Dick, "Three-dimensional multiprocessor system-on-chip thermal optimization," in *Proc. 5th IEEE/ACM/IFIP Int. Conf. Hardw./Softw. Codesign Syst. Synth. (CODES+ISSS)*, Salzburg, Austria, 2007, pp. 117–122.

[85] A. K. Coskun, J. L. Ayala, D. Atienza, T. S. Rosing, and Y. Leblebici, "Dynamic thermal management in 3D multicore architectures," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Nice, France, 2009, pp. 1410–1415.

[86] R. Al-Dujaily, N. Dahir, T. Mak, F. Xia, and A. Yakovlev, "Dynamic programming-based runtime thermal management (DPRTM): An online thermal control strategy for 3D-NOC systems," *ACM Trans. Design Autom. Electron. Syst. (TODAES)*, vol. 19, no. 1, 2013, Art. no. 2.

[87] N. H. Khan, S. M. Alam, and S. Hassoun, "System-level comparison of power delivery design for 2D and 3D ICs," in *Proc. IEEE Int. Conf. 3D Syst. Integr. (3DIC)*, San Francisco, CA, USA, 2009, pp. 1–7.

[88] J. S. Pak *et al.*, "PDN impedance modeling and analysis of 3D TSV IC by using proposed P/G TSV array model based on separated P/G TSV and chip-PDN models," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 1, no. 2, pp. 208–219, Feb. 2011.

[89] T. Lu, C. Serafy, Z. Yang, and A. Srivastava, "Voltage noise induced DRAM soft error reduction technique for 3D-CPUs," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, San Francisco, CA, USA, 2016, pp. 82–87.

[90] Y. Peng, T. Song, D. Petranovic, and S. K. Lim, "Full-chip inter-die parasitic extraction in face-to-face-bonded 3D ICs," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, Austin, TX, USA, 2015, pp. 649–655.

[91] Y. Peng, D. Petranovic, and S. K. Lim, "Fast and accurate full-chip extraction and optimization of TSV-to-wire coupling," in *Proc. 51st ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Jun. 2014, pp. 1–6.

[92] C. Serafy, B. Shi, and A. Srivastava, "A geometric approach to chip-scale TSV shield placement for the reduction of TSV coupling in 3D-ICs," *Integr. VLSI J.*, vol. 47, no. 3, pp. 307–317, 2014.

[93] T. Song, C. Liu, Y. Peng, and S. K. Lim, "Full-chip multiple TSV-to-TSV coupling extraction and optimization in 3D ICs," in *Proc. 50th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Austin, TX, USA, May 2013, pp. 1–7.

[94] H. Wang, M. H. Asgari, and E. Salman, "Compact model to efficiently characterize TSV-to-transistor noise coupling in 3D ICs," *Integr. VLSI J.*, vol. 47, no. 3, pp. 296–306, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167926013000552

[95] Y. Peng, T. Song, D. Petranovic, and S. K. Lim, "On accurate full-chip extraction and optimization of TSV-to-TSV coupling elements in 3D ICs," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, Nov. 2013, pp. 281–288.

[96] K. Yoon *et al.*, "Modeling and analysis of coupling between TSVs, metal, and RDL interconnects in TSV-based 3D IC with silicon interposer," in *Proc. 11th Electron. Packag. Technol. Conf. (EPTC)*, Singapore, Dec. 2009, pp. 702–706.

[97] H.-J. Choi *et al.*, "An experimental study on the TSV reliability: Electromigration (EM) and time dependant dielectric breakdown (TDDB)," in *Proc. IEEE Int. Interconnect Technol. Conf.*, San Jose, CA, USA, Jun. 2012, pp. 1–3.

[98] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Syst. Tech. J.*, vol. 49, no. 2, pp. 291–307, Feb. 1970.

[99] C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," in *Proc. 19th Conf. Design Autom.*, Las Vegas, NV, USA, Jun. 1982, pp. 175–181.

[100] A. E. Caldwell, A. B. Kahng, and I. L. Markov, "Improved algorithms for hypergraph bipartitioning," in *Proc. Asia South Pac. Design Autom. Conf. (ASP-DAC)*, Yokohama, Japan, Jun. 2000, pp. 661–666.

[101] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel hypergraph partitioning: Applications in VLSI domain," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 7, no. 1, pp. 69–79, Mar. 1999.

[102] M. Jung, T. Song, Y. Wan, Y. Peng, and S. K. Lim, "On enhancing power benefits in 3D ICs: Block folding and bonding styles perspective," in *Proc. 51st Annu. Design Autom. Conf. (DAC)*, San Francisco, CA, USA, 2014, pp. 1–6.

[103] T. Song, S. Panth, Y.-J. Chae, and S. K. Lim, "Three-tier 3D ICs for more power reduction: Strategies in CAD, design, and bonding selection," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, Austin, TX, USA, 2015, pp. 752–757.

[104] G. Luo, Y. Shi, and J. Cong, "An analytical placement framework for 3-D ICs and its extension on thermal awareness," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 4, pp. 510–523, Apr. 2013.

[105] H. Yan, Z. Li, Q. Zhou, and X. Hong, "Via assignment algorithm for hierarchical 3D placement," in *Proc. Int. Conf. Commun. Circuits Syst.*, vol. 2. Hong Kong, May 2005, p. 1229.

[106] M.-C. Tsai, T.-C. Wang, and T. Hwang, "Through-silicon via planning in 3-D floorplanning," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 8, pp. 1448–1457, Aug. 2011.

[107] B. Goplen and S. Sapatnekar, "Placement of 3D ICs with thermal and interlayer via considerations," in *Proc. Design Autom. Conf.*, San Diego, CA, USA, Jun. 2007, pp. 626–631.

[108] A. Rajaram and D. Z. Pan, "MeshWorks: An efficient framework for planning, synthesis and optimization of clock mesh networks," in *Proc. Asia South Pac. Design Autom. Conf. (ASPDAC)*, Seoul, South Korea, Mar. 2008, pp. 250–257.

[109] M. R. Guthaus, X. Hu, G. Wilke, G. Flach, and R. Reis, "High-performance clock mesh optimization," *ACM Trans. Design Autom. Electron. Syst.*, vol. 17, no. 3, pp. 1–17, Jul. 2012.

[110] K. Cho, C. Jang, and J.-W. Chong, "Clock mesh network design with through-silicon vias in 3D integrated circuits," *ETRI J.*, vol. 36, no. 6, pp. 931–941, Dec. 2014.

[111] T.-Y. Kim and T. Kim, "Clock tree embedding for 3D ICs," in *Proc. 15th Asia South Pac. Design Autom. Conf. (ASP-DAC)*, Taipei, Taiwan, Jan. 2010, pp. 486–491.

[112] T.-Y. Kim and T. Kim, "Clock tree synthesis for TSV-based 3D IC designs," *ACM Trans. Design Autom. Electron. Syst.*, vol. 16, no. 4, pp. 1–21, Oct. 2011.

[113] W. Liu *et al.*, "TSV-aware topology generation for 3D clock tree synthesis," in *Proc. 14th Int. Symp. Qual. Electron. Design (ISQED)*, Santa Clara, CA, USA, Mar. 2013, pp. 300–307.

[114] T.-H. Chao, Y.-C. Hsu, J.-M. Ho, and A. B. Kahng, "Zero skew clock routing with minimum wirelength," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 39, no. 11, pp. 799–814, Nov. 1992.

[115] T. Lu and A. Srivastava, "Gated low-power clock tree synthesis for 3D-ICs," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2014, pp. 319–322.

[116] T. Lu and A. Srivastava, "Low-power clock tree synthesis for 3D-ICs," *ACM Trans. Design Autom. Electron. Syst.*, vol. 1, no. 1, p. 1, 2017.

[117] T.-Y. Kim and T. Kim, "Clock tree synthesis with pre-bond testability for 3D stacked IC designs," in *Proc. 47th ACM/IEEE Design Autom. Conf. (DAC)*, Anaheim, CA, USA, Jun. 2010, pp. 723–728.

[118] J.-S. Yang, J. Pak, X. Zhao, S. K. Lim, and D. Z. Pan, "Robust clock tree synthesis with timing yield optimization for 3D-ICs," in *Proc. 16th Asia South Pac. Design Autom. Conf. (ASP-DAC)*, Yokohama, Japan, Jan. 2011, pp. 621–626.

[119] C.-L. Lung, Y.-S. Su, H.-H. Huang, Y. Shi, and S.-C. Chang, "Through-silicon via fault-tolerant clock networks for 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 7, pp. 1100–1109, Jul. 2013.

[120] H. Park and T. Kim, "Synthesis of TSV fault-tolerant 3-D clock trees," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 2, pp. 266–279, Feb. 2015.

[121] C. Serafy, Z. Yang, A. Srivastava, Y. Hu, and Y. Joshi, "Thermoelectric codesign of 3-D CPUs and embedded microfluidic pin-fin heatsinks," *IEEE Des. Test.*, vol. 33, no. 2, pp. 40–48, Apr. 2016.

[122] Z. Yang and A. Srivastava, "Co-placement for pin-fin based microfluidically cooled 3D ICs," in *Proc. ASME Int. Tech. Conf. Exhibit. Packag. Integr. Electron. Photon. Microsyst. Collocated ASME 13th Int. Conf. Nanochannels Microchannels Minichannels*, San Francisco, CA, USA, 2015, Art. no. V001T09A036.

[123] A. Ortega, S. Ramanathan, J. D. Chicci, and J. L. Prince, "Thermal wake models for forced air cooling of electronic components," in *Proc. 9th Annu. IEEE Semicond. Thermal Meas. Manag. Symp. (SEMI-THERM IX)*, Austin, TX, USA, Feb. 1993, pp. 63–74.

[124] Y. Zhang *et al.*, "Coupled electrical and thermal 3D IC centric microfluidic heat sink design and technology," in *Proc. IEEE 61st Electron. Compon. Technol. Conf. (ECTC)*, Lake Buena Vista, FL, USA, 2011, pp. 2037–2044.

[125] D. Squiller *et al.*, "Reliable integration of microchannel coolers for power electronics," in *Proc. ASME Int. Tech. Conf. Exhibit. Packag. Integr. Electron. Photon. Microsyst. Collocated ASME 13th Int. Conf. Nanochannels Microchannels Minichannels*, San Francisco, CA, USA, 2015, Art. no. V003T10A015.

[126] Z. Wan, W. Yueh, Y. Joshi, and S. Mukhopadhyay, "Enhancement in CMOS chip performance through microfluidic cooling," in *Proc. 20th Int. Workshop Thermal Investigations ICs Syst. (THERMINIC)*, London, U.K., 2014, pp. 1–5.

[127] M. M. Sabry, A. Sridhar, J. Meng, A. K. Coskun, and D. Atienza, "GreenCool: An energy-efficient liquid cooling design technique for 3-D MPSoCs via channel width modulation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 4, pp. 524–537, Apr. 2013.

[128] Y.-J. Lee, R. Goel, and S. K. Lim, "Multi-functional interconnect co-optimization for fast and reliable 3D stacked ICs," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, Nov. 2009, pp. 645–651.

[129] K. Balakrishnan, V. Nanda, S. Easwar, and S. K. Lim, "Wire congestion and thermal aware 3D global placement," in *Proc. Asia South Pac. Design Autom. Conf. (ASP-DAC)*, Shanghai, China, 2005, pp. 1131–1134.

[130] J. Cong and Y. Zhang, "Thermal via planning for 3-D ICs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, Nov. 2005, pp. 745–752.

[131] B. Goplen and S. Sapatnekar, "Thermal via placement in 3D ICs," in *Proc. Int. Symp. Phys. Design (ISPD)*, San Francisco, CA, USA, 2005, pp. 167–174.

[132] Z. Li *et al.*, "Integrating dynamic thermal via planning with 3D floorplanning algorithm," in *Proc. Int. Symp. Phys. Design (ISPD)*, San Jose, CA, USA, 2006, pp. 178–185.

[133] M. M. Sabry, D. Atienza, and A. K. Coskun, "Thermal analysis and active cooling management for 3D MPSoCs," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Rio de Janeiro, Brazil, 2011, pp. 2237–2240.

[134] M. M. Sabry, A. K. Coskun, D. Atienza, T. S. Rosing, and T. Brunschwiler, "Energy-efficient multiobjective thermal control for liquid-cooled 3-D stacked architectures," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 12, pp. 1883–1896, Dec. 2011.

[135] 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 *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, 2014, pp. 565–572.

[136] M. B. Healy and S. K. Lim, "Power delivery system architecture for many-tier 3D systems," in *Proc. 60th Electron. Compon. Technol. Conf. (ECTC)*, Las Vegas, NV, USA, 2010, pp. 1682–1688.

[137] H.-T. Chen, H.-L. Lin, Z.-C. Wang, and T. Hwang, "A new architecture for power network in 3D IC," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Grenoble, France, 2011, pp. 1–6.

[138] M. S. Gupta, J. L. Oatley, R. Joseph, G.-Y. Wei, and D. M. Brooks, "Understanding voltage variations in chip multiprocessors using a distributed power-delivery network," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Nice, France, 2007, pp. 1–6.

[139] E. Wong, J. Minz, and S. K. Lim, "Decoupling capacitor planning and sizing for noise and leakage reduction," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, Nov. 2006, pp. 395–400.

[140] G. Huang, M. Bakir, A. Naeemi, H. Chen, and J. D. Meindl, "Power delivery for 3D chip stacks: Physical modeling and design implication," in *Proc. IEEE Elect. Perform. Electron. Packag.*, Atlanta, GA, USA, 2007, pp. 205–208.

[141] E. Song, J. S. Pak, and J. Kim, "Power delivery for 3D chip stacks: Physical modeling and design implication," in *Proc. IEEE 62nd Electron. Compon. Technol. Conf. (ECTC)*, San Diego, CA, USA, 2012, pp. 2037–2044.

[142] J. Gu and C. H. Kim, "Multi-story power delivery for supply noise reduction and low voltage operation," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, San Diego, CA, USA, 2005, pp. 192–197.

[143] P. Zhou, "Interconnect design techniques for multicore and 3D integrated circuits," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Minnesoda, Minneapolis, MN, USA, 2012.

[144] N. H. Khan, S. M. Alam, and S. Hassoun, "Through-silicon via (TSV)-induced noise characterization and noise mitigation using coaxial TSVs," in *Proc. IEEE Int. Conf. 3D Syst. Integr. 3DIC*, San Francisco, CA, USA, Sep. 2009, pp. 1–7.

[145] J. R. Black, "Electromigration failure modes in aluminum metallization for semiconductor devices," *Proc. IEEE*, vol. 57, no. 9, pp. 1587–1594, Sep. 1969.

[146] X. Huang, Y. Tan, V. Sukharev, and S. X.-D. Tan, "Physics-based electromigration assessment for power grid networks," in *Proc. 51st Annu. Design Autom. Conf. (DAC)*, San Francisco, CA, USA, 2014, pp. 1–6.

[147] X. Zhao, M. Scheuermann, and S. K. Lim, "Analysis of DC current crowding in through-silicon-vias and its impact on power integrity in 3D ICs," in *Proc. ACM Design Autom. Conf.*, San Francisco, CA, USA, Jun. 2012, pp. 157–162.

[148] X. Zhao, Y. Wan, M. Scheuermann, and S. K. Lim, "Transient modeling of TSV-wire electromigration and lifetime analysis of power distribution network for 3D ICs," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, Nov. 2013, pp. 363–370.

[149] J. Pak, S. K. Lim, and D. Z. Pan, "Electromigration study for multi-scale power/ground vias in TSV-based 3D ICs," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, Nov. 2013, pp. 379–386.

[150] M. Jung, X. Liu, S. K. Sitaraman, D. Z. Pan, and S. K. Lim, "Full-chip through-silicon-via interfacial crack analysis and optimization for 3D IC," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, Nov. 2011, pp. 563–570.

[151] L. Jiang, F. Ye, Q. Xu, K. Chakrabarty, and B. Eklow, "On effective and efficient in-field TSV repair for stacked 3D ICs," in *Proc. 50th Annu. Design Autom. Conf. (DAC)*, Austin, TX, USA, 2013, pp. 1–6.

[152] Q. Zou, T. Zhang, E. Kursun, and Y. Xie, "Thermomechanical stress-aware management for 3D IC designs," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Grenoble, France, Mar. 2013, pp. 1255–1258.

[153] T. Lu and A. Srivastava, "Electrical-thermal-reliability co-design for TSV-based 3D-ICs," in *Proc. ASME Int. Tech. Conf. Exhibit. Packag. Integr. Electron. Photon. Microsyst.*, San Francisco, CA, USA, 2015, pp. 1–10.

[154] J. Pak, S. K. Lim, and D. Z. Pan, "Electromigration-aware routing for 3D ICs with stress-aware EM modeling," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, San Jose, CA, USA, Nov. 2012, pp. 325–332.

[155] H. Tajik, H. Homayoun, and N. Dutt, "VAWOM: Temperature and process variation aware wearout management in 3D multicore architecture," in *Proc. 50th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Austin, TX, USA, May 2013, pp. 1–8.

[156] T. Chantem, Y. Xiang, X. Hu, and R. P. Dick, "Enhancing multicore reliability through wear compensation in online assignment and scheduling," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Grenoble, France, Mar. 2013, pp. 1373–1378.

[157] P. Mercati, A. Bartolini, F. Paterna, T. S. Rosing, and L. Benini, "Workload and user experience-aware dynamic reliability management in multicore processors," in *Proc. 50th Annu. Design Autom. Conf. (DAC)*, Austin, TX, USA, 2013, pp. 1–6.

[158] C. Zhuo, D. Sylvester, and D. Blaauw, "Process variation and temperature-aware reliability management," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Dresden, Germany, Mar. 2010, pp. 580–585.

[159] W. Song, S. Mukhopadhyay, and S. Yalamanchili, "Architectural reliability: Lifetime reliability characterization and management of many-core processors," *Comput. Architect. Lett.*, vol. 14, no. 2, p. 1, 2014.

[160] J. S. Yang, K. Athikulwongse, Y.-J. Lee, S. K. Lim, and D. Z. Pan, "TSV stress aware timing analysis with applications to 3D-IC layout optimization," in *Proc. 47th ACM/IEEE Design Autom. Conf. (DAC)*, Anaheim, CA, USA, Jun. 2010, pp. 803–806.

[161] M. E. Goldfarb and R. A. Pucel, "Modeling via hole grounds in microstrip," *IEEE Microw. Guided Wave Lett.*, vol. 1, no. 6, pp. 135–137, Jun. 1991.

[162] R. A. Pucel, "Design considerations for monolithic microwave circuits," *IEEE Trans. Microw. Theory Tech.*, vol. 29, no. 6, pp. 513–534, Jun. 1981.

[163] G. Katti, M. Stucchi, K. D. Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ICs," *IEEE Trans. Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan. 2010.

[164] I. Savidis and E. G. Friedman, "Closed-form expressions of 3-D via resistance, inductance, and capacitance," *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 1873–1881, Sep. 2009.

[165] J. Mitra *et al.*, "A fast simulation framework for full-chip thermo-mechanical stress and reliability analysis of through-silicon-via based 3D ICs," in *Proc. ECTC*, Lake Buena Vista, FL, USA, May 2011, pp. 746–753.

[166] B.-K. Liew, N. W. Cheung, and C. Hu, "Projecting interconnect electromigration lifetime for arbitrary current waveforms," *IEEE Trans. Electron Devices*, vol. 37, no. 5, pp. 1343–1351, May 1990.

[167] L. M. Ting, J. S. May, W. R. Hunter, and J. W. McPherson, "AC electromigration characterization and modeling of multilayered interconnects," in *Proc. 31st Annu. Int. Rel. Phys. Symp.*, Atlanta, GA, USA, Mar. 1993, pp. 311–316.

[168] C. Okoro *et al.*, "Analysis of the induced stresses in silicon during thermcompression Cu–Cu bonding of Cu-through-vias in 3D-SIC architecture," in *Proc. 57th Electron. Compon. Technol. Conf.*, Sparks, NV, USA, May 2007, pp. 249–255.

[169] K. Athikulwongse, A. Chakraborty, J.-S. Yang, D. Z. Pan, and S. K. Lim, "Stress-driven 3D-IC placement with TSV keep-out zone and regularity study," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, Nov. 2010, pp. 669–674.

[170] J. L. Hennessy and D. A. Patterson, "Memory hierarchy design," in *Computer Architecture: A Quantitative Approach*. San Francisco, CA, USA: Morgan Kaufmann, Inc., 2011, pp. 390–525.

[171] JEDEC. *Wide I/O 2 (Wideio2)*. Accessed on Dec. 21, 2015. [Online]. Available: http://www.jedec.org

[172] J. Hruska. *Beyond DDR4: The Differences Between Wide I/O, HBM, and Hybrid Memory Cube*. Accessed on Dec. 21, 2015. [Online]. Available: http://www.extremetech.com

[173] X. Wu *et al.*, "Hybrid cache architecture with disparate memory technologies," in *Proc. 36th Annu. Int. Symp. Comput. Architect. (ISCA)*, Austin, TX, USA, 2009, pp. 34–45.

[174] C. Serafy, A. Srivastava, and D. Yeung, "Unlocking the true potential of 3D CPUs with micro-fluidic cooling," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2014, pp. 323–326.

[175] C. 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. 47th Annu. IEEE/ACM Int. Symp. Microarchitect. (MICRO)*, Cambridge, U.K., 2014, pp. 1–12.

[176] M. Shevgoor *et al.*, "Quantifying the relationship between the power delivery network and architectural policies in a 3D-stacked memory device," in *Proc. 46th Ann. IEEE/ACM Int. Symp. Microarchitect.*, Davis, CA, USA, 2013, pp. 198–209.

[177] G. H. Loh, "Extending the effectiveness of 3D-stacked dram caches with an adaptive multi-queue policy," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitect. (MICRO)*, New York, NY, USA, Dec. 2009, pp. 201–212.

[178] X. Jiang *et al.*, "CHOP: Adaptive filter-based DRAM caching for CMP server platforms," in *Proc. IEEE 16th Int. Symp. High Perform. Comput. Architect. (HPCA)*, Bengaluru, India, Jan. 2010, pp. 1–12.

[179] S. Jarvis, S. Wright, and S. D. Hammond, "High performance computing systems. Performance modeling, benchmarking and simulation," in *Proc. 4th Int. Workshop PMBS*, vol. 8551. Denver, CO, USA, Nov. 2014.

[180] J. Ousterhout *et al.*, "The case for RAMClouds: Scalable high-performance storage entirely in DRAM," *SIGOPS Oper. Syst. Rev.*, vol. 43, no. 4, pp. 92–105, Jan. 2010.

[181] J. Ahn, S. Hong, S. Yoo, O. Mutlu, and K. Choi, "A scalable processing-in-memory accelerator for parallel graph processing," in *Proc. IEEE Int. Symp. Comput. Architect.*, Portland, OR, USA, 2015, pp. 105–117.

[182] J. Ahn, S. Yoo, O. Mutlu, and K. Choi, "PIM-enabled instructions: A low-overhead, locality-aware processing-in-memory architecture," in *Proc. IEEE Int. Symp. Comput. Architect.*, Portland, OR, USA, 2015, pp. 336–348.

[183] A. Farmahini-Farahani, J. H. Ahn, K. Morrow, and N. S. Kim, "NDA: Near-DRAM acceleration architecture leveraging commodity DRAM devices and standard memory modules," in *Proc. IEEE Int. Symp. High Perform. Comput. Architect.*, Burlingame, CA, USA, 2015, pp. 283–295.

[184] D. Zhang *et al.*, "TOP-PIM: Throughput-oriented programmable processing in memory," in *Proc. Int. Symp. High Perform. Parallel Distrib. Comput.*, Vancouver, BC, Canada, 2014, pp. 85–98.

[185] Q. Zhu *et al.*, "A 3D-stacked logic-in-memory accelerator for application-specific data intensive computing," in *Proc. IEEE Int. 3D Syst. Integr. Conf.*, San Francisco, CA, USA, 2013, pp. 1–7.

[186] C. Serafy, "Architectural-physical co-design of 3D CPUs with micro-fluidic cooling," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Maryland, College Park, MD, USA, 2016.

[187] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The splash-2 programs: Characterization and methodological considerations," *ACM SIGARCH Comput. Architect. News*, vol. 23, no. 2, pp. 24–36, 1995.

[188] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: Characterization and architectural implications," in *Proc. PACT*, Toronto, ON, Canada, 2008, pp. 72–81.

[189] P. Batude *et al.*, "Advances in 3D CMOS sequential integration," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Baltimore, MD, USA, 2009, pp. 1–4.

[190] A. W. Topol *et al.*, "Enabling SOI-based assembly technology for three-dimensional (3D) integrated circuits (ICs)," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, USA, 2005, pp. 352–355.

[191] P. Batude *et al.*, "3D monolithic integration," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Rio de Janeiro, Brazil, 2011, pp. 2233–2236.

[192] P. Batude, M. Vinet, A. Pouydebasque, and L. Clavelier, "Enabling 3D monolithic integration," *ECS Trans.*, vol. 16, no. 8, pp. 47–54, 2008.

[193] P. Batude *et al.*, "Low temperature FDSOI devices, a key enabling technology for 3D sequential integration," in *Proc. Int. Symp. VLSI Technol. Syst. Appl. (VLSI-TSA)*, Hsinchu, Taiwan, 2013, pp. 1–4.

[194] M. Shulaker *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2014, pp. 27.4.1–27.4.4.

[195] T.-T. Wu *et al.*, "Sub-50nm monolithic 3D IC with low-power CMOS inverter and 6T SRAM," in *Proc. Int. Symp. VLSI Technol. Syst. Appl. (VLSI-TSA)*, Hsinchu, Taiwan, 2015, pp. 1–2.

[196] S. Bobba *et al.*, "CELONCEL: Effective design technique for 3-D monolithic integration targeting high performance integrated circuits," in *Proc. Asia South Pac. Design Autom. Conf.*, Yokohama, Japan, Jan. 2011, pp. 336–343.

[197] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power benefit study for ultra-high density transistor-level monolithic 3D ICs," in *Proc. 50th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Austin, TX, USA, 2013, pp. 1–10.

[198] S. A. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD methodologies for low power gate-level monolithic 3D ICs," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, 2014, pp. 171–176.

[199] K. Chang, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "Match-making for monolithic 3D IC: Finding the right technology node," in *Proc. 53rd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Austin, TX, USA, 2016, pp. 1–6.

[200] 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 *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Austin, TX, USA, 2016, pp. 1–7.

[201] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna, and S. K. Lim, "Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3D ICs," in *Proc. ICCAD*, Austin, TX, USA, 2016, p. 129.

[202] Samsung. (Aug. 2014). *Samsung Starts Mass Producing Industrys First 3D TSV Technology Based DDR4 Modules for Enterprise Servers*. [Online]. Available: http://www.samsung.com/semiconductor/about-us/news/13602

[203] Tezzaron. *Diram4^{TM} 3D Memory*. Accessed on Jul. 16, 2016. [Online]. Available: http://www.tezzaron.com/products/diram4-3d-memory/

[204] Y.-F. Tsai, F. Wang, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Design space exploration for 3-D cache," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 4, pp. 444–455, Apr. 2008.

[205] S. Li *et al.*, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitect. MICRO*, New York, NY, USA, Dec. 2009, pp. 469–480.

[206] H. N. Phan and D. Agonafer, "Experimental analysis model of an active cooling method for 3D-ICs utilizing multidimensional configured thermoelectric coolers," *J. Electron. Packag.*, vol. 132, no. 2, 2010, Art. no. 024501.

[207] H. Su and S. S. Sapatnekar, "Hybrid structured clock network construction," in *IEEE/ACM Int. Conf. Comput. Aided Design. ICCAD IEEE/ACM Dig. Tech. Papers*, San Jose, CA, USA, Nov. 2001, pp. 333–336.

[208] A. Rajaram, J. Hu, and R. Mahapatra, "Reducing clock skew variability via crosslinks," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 6, pp. 1176–1182, Jun. 2006.

[209] K. Han, A. B. Kahng, J. Lee, J. Li, and S. Nath, "A global-local optimization framework for simultaneous multi-mode multi-corner clock skew variation reduction," in *Proc. 52nd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, San Francisco, CA, USA, Jun. 2015, pp. 1–6.

[210] X. Zhao, M. R. Scheuermann, and S. K. Lim, "Analysis and modeling of DC current crowding for TSV-based 3-D connections and power integrity," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 4, no. 1, pp. 123–133, Jan. 2014.

[211] A. Husain and K.-Y. Kim, "Design optimization of manifold microchannel heat sink through evolutionary algorithm coupled with surrogate model," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 3, no. 4, pp. 617–624, Apr. 2013.

[212] M. Schultz *et al.*, "Embedded two-phase cooling of large three-dimensional compatible chips with radial channels," *J. Electron. Packag.*, vol. 138, no. 2, 2016, Art. no. 021005.

[213] (Aug. 2016). *ANSYS*. [Online]. Available: http://www.ansys.com/Products/Semiconductors/3-D-IC

[214] (Aug. 2016). *M Graphics*. [Online]. Available: https://www.mentor.com/solutions/3d-ic-design/

[215] (Aug. 2016). *Synopsys*. [Online]. Available: https://www.synopsys.com/Solutions/EndSolutions/3d-ic-solutions/Pages/default.aspx

[216] J.-M. Lin and Y.-W. Chang, "TCG: A transitive closure graph-based representation for non-slicing floorplans," in *Proc. 38th Annu. Design Autom. Conf.*, Las Vegas, NV, USA, 2001, pp. 764–769.

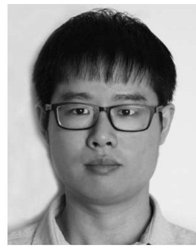
Tiantao Lu (S'12) received the B.S. degree from the Department of Electrical Engineering and Computer Science, Peking University, Beijing, China, in 2011, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD, USA, in 2016.

He is currently a Lead Software Engineer with Cadence, San Jose, CA, USA. His current research interests include physical design methodology for high-performance thermal-aware and reliable 3-D ICs.

Caleb Serafy (S'12) received the B.S. degree in computer engineering and the M.S. degree in electrical engineering from Binghamton University, Binghamton, NY, USA, in 2010 and 2011, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Maryland (UMD) at College Park, College Park, MD, USA, in 2016.

He is currently a Senior Hardware Engineer with Oracle, Santa Clara, CA, USA. His current research interests include thermal-electrical-physical co-design of 3-D ICs and power delivery network modeling and optimization.

Dr. Serafy was a recipient of the Distinguished Graduate Fellowship, the Summer Research Fellowship, and the Distinguished Dissertation Fellowship while at UMD.



Zhiyuan Yang received the B.S. degree from Zhejiang University, Hangzhou, China, in 2013. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Maryland (UMD) at College Park, College Park, MD, USA.

He has been a Research Assistant with UMD, since 2014, under the supervision of Prof. A. Srivastava. His current research interests include 3-D integrated circuits, power delivery of 3-D ICs, and thermal-architectural-physical co-design of 3-D ICs with micro-fluidic cooling.

Mr. Yang was a recipient of the Distinguished Graduate Fellowship from UMD.

Sandeep Kumar Samal (S'12) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2012, and the M.S. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2013, where he is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering.

He has authored over 20 publications in refereed journals and conferences. His current research interests include low power and reliable digital design, modeling, and analysis using through-silicon-via-based and monolithic 3-D IC technology.

Sung Kyu Lim (S'94–M'00–SM'05) received the B.S., M.S., and Ph.D. degrees from the University of California at Los Angeles, Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently the Dan Fielder Endowed Chair Professor. His current research interests include modeling, architecture, and electronic design automation for 3-D ICs.

Dr. Lim was a recipient of the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006, Distinguished Service Award in 2008, and the Best Paper Awards from the IEEE Asian Test Symposium in 2012 and the IEEE International Interconnect Technology Conference in 2014. He was on the Advisory Board of the ACM Special Interest Group on Design Automation from 2003 to 2008. He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS from 2007 to 2009. He has been an Associate Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS since 2013. He has served on the Technical Program Committee of several premier conferences in EDA.

Ankur Srivastava (M'02–SM'15) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology Delhi, New Delhi, India, in 1998, and the Ph.D. degree in computer science from University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 2002.

His current research interests include high performance, low power and secure electronic systems and applications such as computer vision, data, and storage centers and sensor networks. He has published numerous papers on the above fields at prestigious venues.

Dr. Srivastava was a recipient of the prestigious Outstanding Dissertation Award from the Computer Science Department of UCLA in 2002. His research and teaching contributions have also been recognized through various awards. He has been a part of the Technical Program and Organizing Committees of several conferences, such as ICCAD, DAC, ISPD, ICCD, GLSVLSI, and HOST. He has served as an Associate Editor for the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, and the *Integration: VLSI Journal*.