Journal of Systems Architecture 97 (2019) 397–415

Contents lists available at ScienceDirect

Journal of Systems Architecture

journal homepage: www.elsevier.com/locate/sysarc

A survey of optimization techniques for thermal-aware 3D processors

Kun Cao [a], Junlong Zhou [b], Tongquan Wei [a,*], Mingsong Chen [a], Shiyan Hu [c,1], Keqin Li [d,2]

[a] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China
[c] Department of Computer Science, University of Essex, Colchester CO4 3SQ, UK
[d] Department of Computer Science, State University of New York, New York, 12561, USA

ARTICLE INFO

Keywords:
3D processors
Architecture
Floorplanning
Memory management
Task scheduling
Thermal characteristics

ABSTRACT

Interconnect scaling has become a major design challenge for traditional planar (2D) integrated circuits (ICs). Three-dimensional (3D) IC that stacks multiple device layers through 3D stacking technology is regarded as an effective solution to this dilemma. A promising 3D IC design direction is to construct 3D processors. However, 3D processors are likely to suffer from more serious thermal issues as compared to conventional 2D processors, which may hinder the employment or even offset the benefits of 3D stacking. Therefore, thermal-aware design techniques should be adopted to alleviate the thermal problems with 3D processors. In this survey, we review works on system level optimization techniques for thermal-aware 3D processor design from hierarchical perspectives of architecture, floorplanning, memory management, and task scheduling. We first survey 3D processor architectures to demonstrate how a 3D processor can be constructed by using 3D stacking technology, and present an overview of thermal characteristics of the constructed 3D processors. We then review thermal-aware floorplanning, memory management and task scheduling techniques to show how the thermal impact on 3D processor performance can be reduced. A systematic classification method is utilized throughout the survey to emphasize similarities and differences of various thermal-aware 3D processor optimization techniques. This paper shows that the thermal impact on 3D processors is manageable by adopting thermal-aware techniques, thus making 3D processors into the mainstream in the near future.

1. Introduction

With the advance of aggressive technology scaling, the density of integrated circuits (ICs) has been continually increasing. However, the rate of interconnect scaling does not keep up with that of technology scaling, and interconnects still account for a large portion of the total chip capacitance [1–6]. As a result, interconnect scaling remains a major design challenge for traditional planar (2D) ICs. Three-dimensional (3D) IC is regarded as an effective solution to overcome this dilemma [7–12]. By vertically stacking several device layers connected through 3D stacking technology, 3D ICs not only significantly improve packaging density but also drastically reduce global wirelength and semiglobal wirelength.

One promising trend in 3D IC design is to construct 3D processors. Numerous works have been conducted to explore the architecture design of 3D processors, and many novel architecture designs from various perspectives have been presented, such as stacking one core layer and several cache layers, or stacking multiple core/cache layers. 3D processors can offer significant advantages, including reduction in form-factor, improvement for memory bandwidth, realization of heterogeneous integration, and resilience towards security attacks, as detailed below.

- **Reduction in form-factor.** As an important aspect of hardware design, form-factor specifies the physical specifications (e.g., size, shape) of components. Compared to the layout of conventional 2D processors, 3D processors are with smaller form-factor due to the advantages brought by the additional third dimension. Therefore, 3D processors have higher packaging density and smaller footprint, which facilitates a low-cost chip design [13–17].
- **Improvement for memory bandwidth.** For traditional 2D processors, the memory units and logic units are resided at opposite ends, which results in the performance bottleneck induced by memory bandwidth. For 3D processors, the critical path length between logic units and memory units can be considerably shortened by using Through Silicon Vias (TSVs) instead of conventional I/O pins [18–21]. Therefore, 3D processors can dramatically improve memory bandwidth.
- **Realization of heterogeneous integration.** 3D stacking technology has the ability to realize heterogeneous integration [22–26]. For example, Chen and Jha [23] presented a 3D hybrid architecture where

* Corresponding author.
E-mail address: tqwei@cs.ecnu.edu.cn (T. Wei).
[1] Senior Member, IEEE
[2] Fellow, IEEE

https://doi.org/10.1016/j.sysarc.2019.01.003
Received 21 October 2018; Received in revised form 1 December 2018; Accepted 7 January 2019
Available online 8 January 2019

a field-programmable gate array (FPGA) layer and a DRAM layer are vertically stacked on a CPU layer. The authors showed that the proposed design significantly reduces power consumption while maintaining high computation performance.

- **Resilience towards security attacks.** 3D processors have been shown to be resilient towards security attacks [27–32]. This is because by utilizing 3D stacking technique, the layers of sensitive circuits can be separated, making the function of each layer obscured. The resultant 3D architecture makes it very difficult to reverse engineer the circuit.

Although 3D processors offer significant advantages, they suffer from more serious thermal issues compared to 2D processors. This is mainly due to limited heat dissipation paths and higher power density resulting from vertical stacking of multiple active layers [33–39]. It has been shown that thermal issues have become a major challenge in deploying 3D processors [5,24,39–45]. The main negative effects incurred by thermal issues for 3D processors are listed below.

- **Reduction in throughput.** It has been shown in the literature [38] that a core of 3D processors is likely to reach the peak temperature limit before violating its power constraint. This indicates that high temperatures will prevent 3D processors from achieving high throughput.
- **Increase in energy consumption.** In the nano-era, the leakage power is gradually dominating the overall power consumptions of ICs [46–50]. Since the leakage power is positively related to the chip operating temperature, high temperatures may incur extra energy consumption of 3D processors when executing a given task set.
- **Degradation in reliability.** Lu et al. [51] observed that DRAM transient errors are closely related to processor activities through voltage and thermal coupling in the 3D processor architecture where one DRAM layer and one processor layer are vertically stacked. High temperatures will increase the transient error rate of the DRAM and degrade the reliability of 3D processors.
- **Decay in lifetime.** It has been shown in the literature [33] that the peak temperature achieved by a two-layer 3D processor with four cores in each layer is 20 °C higher than the peak temperature achieved by a 2D processor with eight cores in one layer. In addition, Zhou et al. [52] observed that compared to the cores close to the heat sink, the cores distant from the heat sink are generally with higher temperatures for 3D processors, which results in 3D processors having a higher temperature variance than 2D processors. High peak temperatures and/or temperature variances will accelerate component aging and thus decay the lifetime of 3D processors.

Scope and Contribution: In this survey, we review research works on system level optimization techniques for thermal-aware 3D processor design. In particular, we discuss thermal-aware optimization techniques for 3D processors from perspectives of floorplanning, memory management, and task scheduling. We aim to extract the main ideas of these thermal-aware design techniques, and highlight their similarities and differences. Note that there are already some surveys on 3D processors in the literatures [24,53–57]. Compared with these existing surveys, the novel contributions of this paper are summarized below.

- Unlike the surveys [24,53–55] that mainly focus on 3D processor architecture design, or the surveys [56,57] that ignore the exploration of memory management, this paper reviews the literatures related to 3D processors from system level perspectives of architecture design, floorplanning, memory management, and task scheduling.
- We give a detailed introduction to new research works (around 70%) that are not contained in these existing surveys [24,53–57] for 3D processors. These new works reflect the recent progress in research and development of thermal-aware optimization of 3D processor design.
- All the existing surveys [24,53–57] fail to provide a systematic classification of the thermal-aware optimization research. In this paper, a systematic classification method is presented to embody most of the research directions for thermal-aware 3D processor optimization.

Organization: Fig. 1 demonstrates the organization of this paper. As shown in the figure, we first review works on 3D processor architecture design in Section 2, and present the works exploring thermal characteristics of the constructed 3D processors in Section 3. We then survey the literatures on thermal-aware floorplanning in Section 4. In Section 5 we summarize the works on thermal-aware memory management while in Section 6 we outline the works on thermal-aware task scheduling. In Sections 4–6, we organize these research works into different categories to show similarities and dissimilarities of various thermal-aware optimization techniques. Finally, Section 7 concludes this paper.

2. 3D processor architecture design

In this section, we summarize 3D processor architectures to demonstrate how to construct a 3D processor by using 3D stacking technology. These 3D processor architectures include stacking main memory architecture that concentrates on processor-main memory design (Section 2.1); stacking cache architecture that stacks one processor layer and multiple cache layers (Section 2.2); stacking cache + core architecture that stacks multiple cache/core layers (Section 2.3); and other 3D processor architecture (Section 2.4).

2.1. Stacking main memory architecture

Wang et al. [58] explored the resistance drift in a multi-level cell phase change memory (MLC PCM) and proposed a novel 3D architecture to decrease the drift-related soft errors. As shown in Fig. 2, four PCM layers, one DRAM layer, and one processor layer are vertically stacked. This architecture is based on the observation that one can reconfigure the PCM read circuit instead of inserting error correction codes or using large margins to tolerate most resistance drift errors. Based on this observation, the PCM read circuit in the proposed 3D architecture is reconfigured for the purpose of tolerating resistance drift errors. Evaluations show that compared to the benchmarking PCM designs, the proposed architecture achieves 10^6 times of error rate reduction.

Lee et al. [59] proposed a novel 3D stacked DRAM architecture. By making full use of the bandwidth offered by TSVs, the proposed architecture can concurrently access multiple DRAM layers. When the DRAM layers are simultaneously transferring data, these DRAM layers should coordinate with each other in order to prevent channel contention. The authors proposed two coordination solutions: one is Dedicated-I/O, and the other is Cascaded-I/O. By allocating each DRAM layer to multiple dedicated TSVs, Dedicated-I/O achieves the statical TSV partition. However, this scheme has two disadvantages. First, each layer should be designed nonuniformly, which will increase manufacturing overheads. Second, the energy consumption of DRAM achieved by this scheme is linearly related to the number of DRAM layers. The second solution, Cascaded-I/O, overcomes the above disadvantages by time-multiplexing all TSVs. Simulation results show that compared with a baseline 3D DRAM architecture, the presented design achieves up to 55% performance improvement and 18% energy reduction.

Liu et al. [61] presented a 3D multicore processor architecture that is specifically designed to solve operational and data path problems in biological sequence analysis. The developed processor architecture has the ability to make the array of processing elements that make up a core reconfigurable, and efficiently interconnect shared buffers for achieving congestion minimization and throughput maximization. As shown in Fig. 3, every core is associated with two DRAM controllers which are interfaced with the dedicated external DRAM channels. The integrated sequencer is utilized to route local access of a core to DRAM banks, while the circuit-switching bus is utilized to route remote access of the core to DRAM banks. During the data preparation phase, the host first performs data writing to these DRAM banks and then programs the

Paper organization

Fig. 1. Organization of the paper by section.

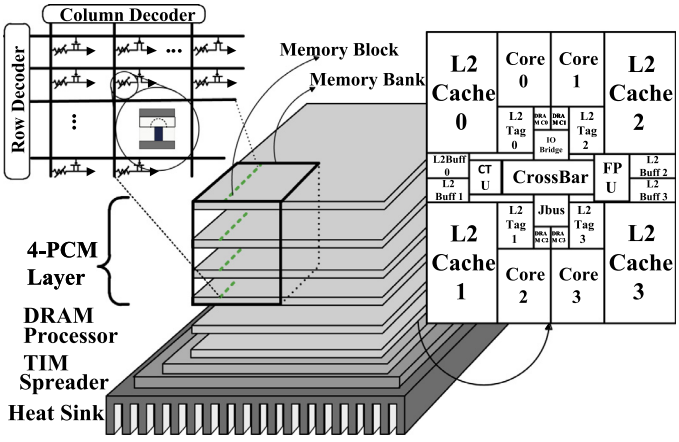


Fig. 2. A time/temperature-aware 3D processor architecture [58].

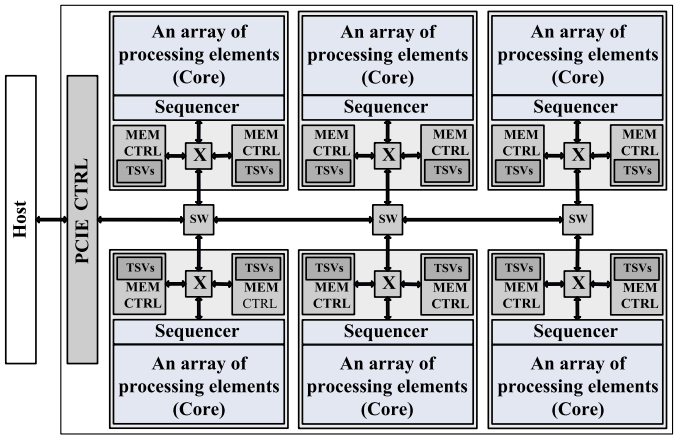


Fig. 3. A reconfigurable 3D many-core processor [60].

integrated sequencer. During the application processing phase, the integrated sequencer wisely manages the core and DRAM controllers such that this processing phase can be accelerated. After application processing is completed, the output will be stored in DRAM banks and read back by the host. Experimental results show that the proposed 3D multicore processor design can achieve up to 40 times speedup.

Tang et al. [62] proposed a realistic processing-in-memory 3D DRAM architecture. The presented design, as illustrated in Fig. 4, is based on hybrid memory cube (HMC) that is composed of numerous vaults. Every vault is equipped with a dedicated controller, and multiple vaults are independent from each other in terms of function and operation. An existing packet-transmission protocol is utilized to perform the processor-HMC communication. This protocol is interpreted by the HMC interface, and it supports read/write instructions, several data manipulation instructions and data calculation instructions. For reducing system energy consumption, the lane of the proposed architecture will be configured into quarter-width links when the target application needs to be executed. All application instructions are then fetched from the stacked memory layer, and executed on the logic layer. After application execution is finished, the lane will be recovered as full-duplex serialized links.

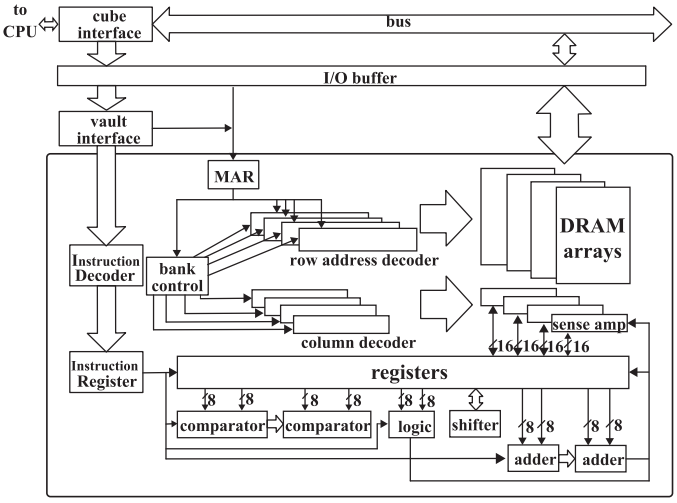


Fig. 4. An overview of the processing-in-memory architecture [62].

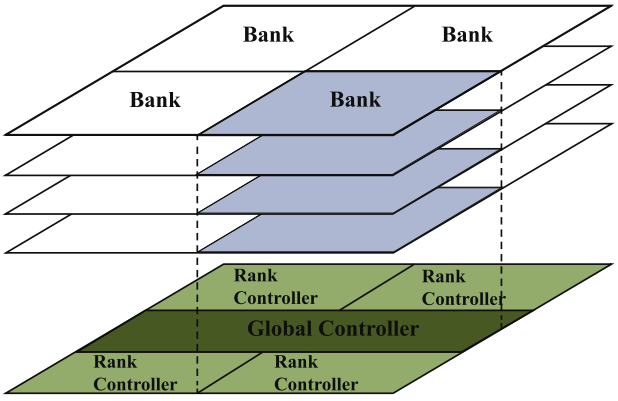


Fig. 5. The energy-efficient DRAM controller [60].

Simulation results reveal that the presented 3D DRAM architecture can gain 1.3 times speedup on average and achieve 13% energy saving.

Liu et al. [60] proposed an energy-efficient controller design for 3D stacked DRAM. As illustrated in Fig. 5, the presented controllers are the interfaces between the logic layer and the memory layers that are stacked vertically on top of the logic layer. The global controller adopts command scheduling and rank interleaving techniques to implement issuing commands in parallel. Utilizing block-in-block-out scheme instead of first-in-first-out method, the latency of read/write operation can be greatly reduced even if rank controllers are busy. A rank controller manages local rank segments and issues commands to each bank through TSVs. When a bank is idle, the rank controller will put this bank into sleep mode for power saving. Evaluation results demonstrate that the presented DRAM controller design decreases task execution time by 40.8%, improves bandwidth utilization by 66.89%, and reduces energy consumption by 27.18%.

2.2. Stacking cache architecture

Azarkhish et al. [63] focused on the cache design to reduce memory access time through exploiting the advantages of 3D stacking technology. The authors presented a scalable and synthesizable 3D-nonuniform L2 memory access architecture where several L2 memory dies with same layouts and a processor die are vertically stacked. Due to highly pipelined, this design can achieve high clock frequencies and multiple in-flight transactions. The authors implemented the proposed architecture by using the STM CMOS-28-nm technology. Experimental results demonstrate that the developed architecture can operate at 500 MHz clock frequency and obtain 34% average performance boost.

Kang et al. [64] proposed a power-efficient 3D on-chip interconnect for a multicore cluster. Fig. 6(a) shows a schematic view of the multicore cluster with 3D L2 cache. As illustrated in the figure, the multicore cluster includes multiple cores, and each core is equipped with its own data caches and private L1 instruction. The multi-banked stacked L2 cache is composed of multiple SRAM banks. All stacked SRAM banks are connected through the 3D mesh-of-tree (MoT) interconnect. Fig. 6(b) shows a geometry view of the 3D multicore cluster. MoT interconnect is located in the middle of the core layer, which well balances the memory access latency from each core. The new design of routing switch for 3D MoT interconnects can make the interconnects reconfigurable to support power gating of processing cores, cache memory banks, and the corresponding interconnect links. This reconfigurability allows to adjust power states of the interconnects to application's characteristics such as scalability for parallelism and L2 cache demand.

Kong et al. [65] utilized monolithic 3D (M3D) integration technology to design tag arrays and data arrays for last-level caches. Fig. 7 shows the proposed two architecture designs. In the first design, as shown in Fig. 7(a), the M3D technology is applied to construct tag arrays. The SRAM tag arrays are stacked vertically on top of the logic die that consists of multiple cores and the L2 cache for each core. The capacity of eDRAM cache is deemed to be 256MB, and six layers are stacked to implement 24MB SRAM tag arrays. In the second design, as illustrated in Fig. 7(b), the M3D technology is applied to construct data arrays. The capacity of cache data arrays is assumed to be 64MB, and sixteen SRAM data array layers with the capacity of each layer being 4MB are stacked above the processor core die. Experimental results show that compared to benchmarking TSV-based 3D SRAM array architectures, the proposed two designs can reduce energy consumption by up to 1.7% and 79.1%, respectively.

Nasri et al. [66] explored the use of 3D integration technology to reduce system power consumption. The authors proposed to replace conventional L2 cache designs with spin-transfer torque random access memory (STT-RAM) as on-chip L2 caches. This proposal is based on the observation that the static power consumption of STT-RAM is very close to zero due to its non-volatile nature. Therefore, STT-RAM based cache designs can implement dramatic reduction in overall system power consumption even though the technology scaling stops or the core number keeps constant. Given this, as illustrated in Fig. 8, the authors developed a 3D architecture that stacks a processor layer consisting of 16 cores and a cache layer consisting of 16 STT-RAM cache banks by using TSVs. Experiment results show that compared to the traditional L2 cache design, the proposed design achieves 51% reduction in system energy consumption as well as 37% improvement in energy delay product while only incurring 12% degradation in instruction per cycle.

2.3. Stacking cache + core architecture

Zhang et al. [67] illustrated a 3D cache resource pooling architecture. As shown in Fig. 9(a), in the presented architecture four layers where each stacked layer contains a private 1MB L2 cache and four cores are vertically placed on the stacked DRAM layers. Fig. 9(b) illustrates the cache resource pooling. The cache pooling consists of vertically adjacent caches connected by TSVs, and a shared memory controller is utilized to connect all the private L2 caches. As a result, on-chip communication is achieved by using a shared memory. Experimental results show that compared with benchmarking 3D designs, the presented cache resource pooling architecture can improve 40.4% energy-delay-area product.

Joardar et al. [68] proposed a 3D network-on-chip (NoC) architecture that enables high-performance collective communication. As shown in Fig. 10, each processor tile in the developed architecture is equipped with a primary data-transfer router and an SMART hop setup requests (SSR) router. In addition to the SSR mesh connectivity, the control network also incorporates single-bit multi-drop pre-SSR wires. Every SMART hop contains four router pipeline stages, that is, local arbitration for messages, pre-SSR transmission and arbitration for single-bit pre-cursor requests, SSR transmission for detailed bypass requests, and single-cycle multi-hop link data traversal. Experimental results show that compared with the existing 3D NoCs based on Path and Tree multicast, the proposed 3D NoC design reduces 65% and 31% network latency, respectively.

Das et al. [69] explored energy-efficient architecture design for M3D-based NoCs (as illustrated in Fig. 11). The authors focused on optimizing the placements of links and routers to guarantee optimal achievable performance. The best router and link placement is explored utilizing a machine-learning optimization solution STAGE. The basic idea of STAGE is to take advantages of the past experience in problem-solving (i.e., local search runs) to generate an evaluation function that has the ability to perform the initial state quality prediction in each iteration. The obtained evaluation function is then utilized to select several optimal initial states for the iteration. As larger design space through local search runs are explored, the accuracy of the learned evaluation function can be significantly improved. Experimental results show that compared to traditional mesh-based NoCs, the proposed M3D-based NoC lowers 32% energy-delay-product. In addition, the presented

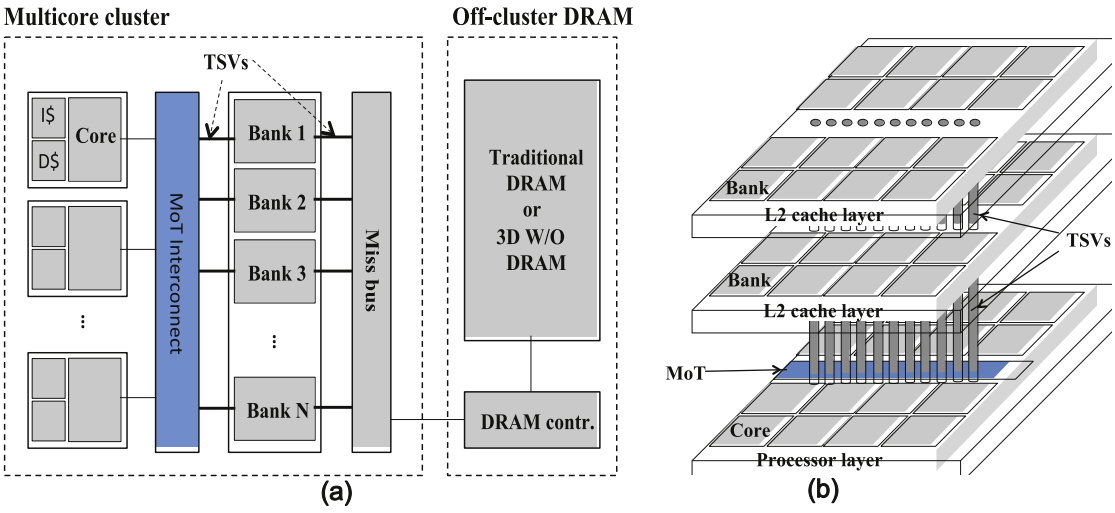


Fig. 6. A 3D multicore cluster with MoT interconnect: (a) schematic view; (b) geometric view [64].

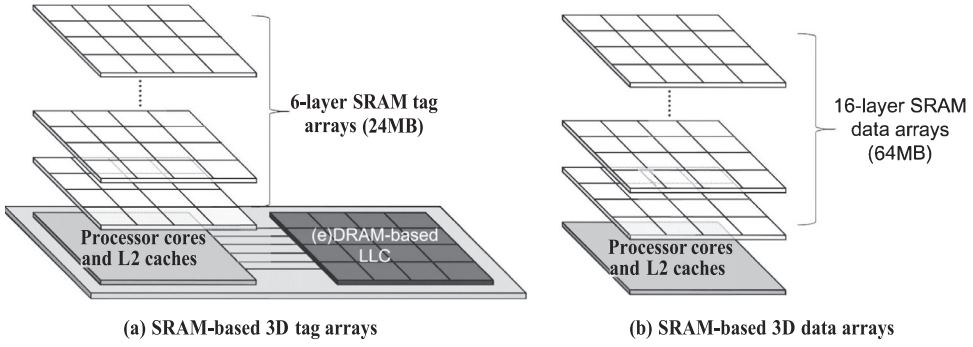


Fig. 7. A baseline 3D stacking cache architecture [65].

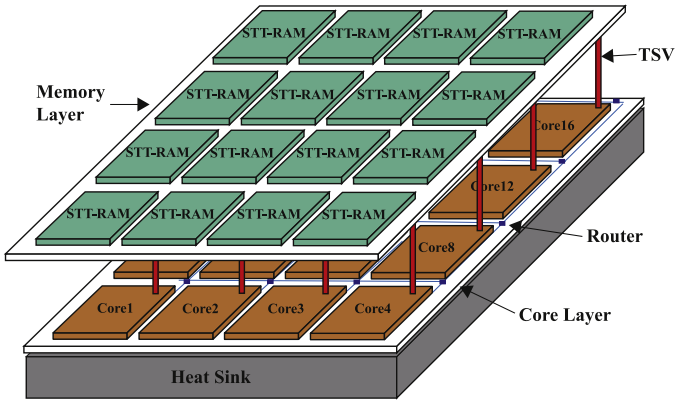


Fig. 8. An overview of the STT-RAM based 3D architecture [66].

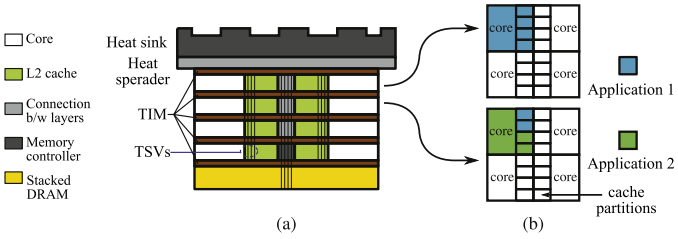


Fig. 9. A 3D cache resource pooling architecture [67].

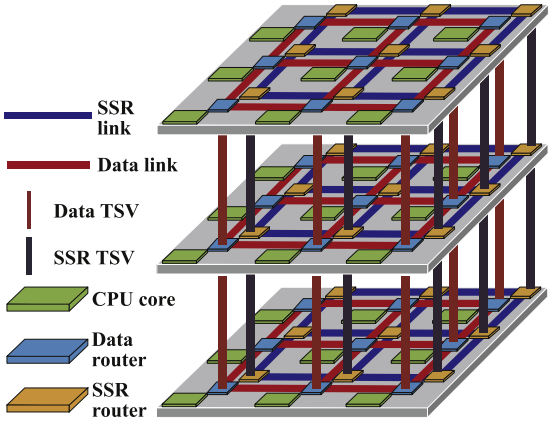


Fig. 10. The 3D NoC architecture presented in [68].

M3D-based NoC can also achieve average 28% energy-delay-product reduction compared to TSV-based counterparts.

Wang et al. [70] proposed a novel hierarchical 3D NoC architecture which enables thermal-aware task migration at runtime. The whole network hierarchy of the presented 3D NoC architecture can be divided into three levels. In the first-level network, circuit switching technique is utilized to connect nodes to form planar rings. In the second-level network, these planar rings are stacked vertically to construct cubes. In the third-level network, cubes are connected to each other to build the entire network. Routing is also achieved in a hierarchical way. To be specific, routing paths are first established within rings, and then the data that has to pass through the rings or cubes is transferred by using

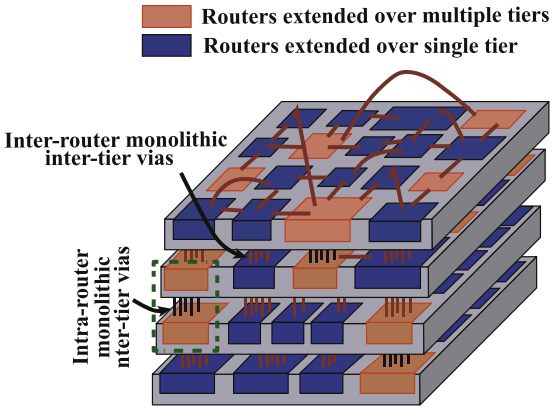


Fig. 11. An illustration of the 3D NoC architecture with M3D integration [69].

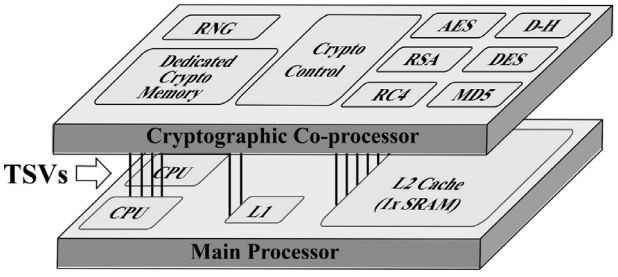


Fig. 12. A cryptographic co-processor is stacked on the main processor [28].

dimension-order routing technology. The main advantage of this routing solution is that the tasks that need to be migrated can be moved around the rings without increasing communication distance. Simulation results show that compared to benchmarking 3D NoCs built on mesh topologies, the proposed hierarchical 3D NoC architecture can reduce communication latency by up to 80%.

2.4. Other 3D processor architecture

Valamehr et al. [28] focused on the 3D architecture design from the perspective of performance and security needs. The authors proposed a novel architecture where a cryptographic co-processor is placed on the top of a commodity microprocessor by using 3D stacking technology (see Fig. 12). The authors investigated the security ramifications of the proposed 3D crypto co-processor. Various security threats are outlined, and how the 3D crypto co-processor alleviates these attacks is analyzed. Simulation results show that in addition to gaining high throughput and supporting cryptographic functions, the proposed 3D architectures can effectively alleviate some attacks, such as time-driven side channel attacks, memory remanence attacks, and access-driven cache side channel attacks.

Sepulveda et al. [71] observed that the TSV communication between stacked layers in multiprocessor system-on-chip (MPSoC) can be modified, spied and even denied by vertical communication manipulation. Given this, the authors proposed a security architecture, namely 3D-LeukoNoC. 3D-LeukoNoC consists of three main components: a) Recognizers, which are utilized to filter data, b) quality-of-service routers, which offer guaranteed quality-of-service, c) lymphocyte-B, which manages the configuration of the recognizers and quality-of-service routers and allocates TSVs. 3D-LeukoNoC prevents 3D MPSoC from being attacked by two ways. First, it can block malicious traffic by using the recognizers. Second, it can manage the inter-layer and intra-layer communication by using lymphocyte-B and quality-of-service routers at run time.

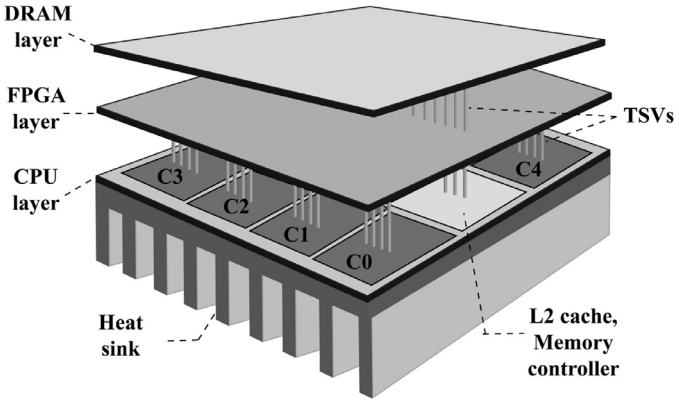


Fig. 13. A hybrid 3D processor architecture [23].

Although the proposed architecture 3D-LeukoNoC in [71] has the ability to prevent the 3D MPSoC from being attacked, it may incur performance degradation with the increase in islands on the 3D MPSoC. To guarantee both security and performance, Sepulveda et al. [72] further developed a scalable and distributed 3D-NoC-based security architecture 3D-SECTSV. 3D-SECTSV mainly consists of four key components: a) Recognizers, which are utilized to filter data, b) security and quality routers, which ensure security and quality services, c) security TSV interface, which assigns the data on the TSVs and controls the configuration of the recognizers as well as security and quality routers, and d) reconfiguration and security manager module. 3D-SECTSV can achieve the protection for 3D MPSoC by blocking malicious traffic, controlling the intra-layer communication, and isolating sensitive traffic while guaranteing system performance.

Chen and Jha [23] focused on designing a low-power 3D processor architecture. The authors found that the power consumption of general-purpose CPUs could be significantly lowered with the help of specialized accelerators. As a result, additional computational components can be powered by the saved power such that better performance is delivered. The authors proposed a 3D hybrid architecture where a DRAM layer and an FPGA layer are vertically stacked on a CPU layer (see Fig. 13). Due to high flexibility and power efficiency, FPGAs are ideal candidates to achieve low-power computation. The FPGA layer has the ability to support multiple accelerators. It can directly access the data stored in CPU caches through a well established communication mechanism. Therefore, switches between the FPGA layer and CPU layer can be extremely fast. Simulation results show that the proposed 3D hybrid architecture achieves significant reduction in power consumption.

2.5. Summary and discussion

This section reviews works related to 3D processors from an architectural design perspective. These works show that stacking main memory architecture, stacking cache architecture, and stacking cache + core architecture are the three most popular 3D processor design solutions, which offer significant advantages over traditional 2D processor design solutions. In addition to the three common designs, several novel heterogeneous integration architectures, such as the stacking FPGA-CPU design, are also proposed. However, 3D stacking technology is likely to result in high chip power density and limited heat dissipation paths, both of which may incur severe thermal issues for 3D processors. Given this, in the next section, we will outline thermal characteristics of the constructed 3D processors.

3. Thermal characteristics of 3D processors

In this section, we show the works that focus on exploring the thermal characteristics of 3D processors.

3.1. Thermal characteristic exploration

Loi et al. [73] compared the performance of 3D processors to that of conventional 2D processors under thermal constraints. The authors found that 3D processors achieve significant improvement on memory bus width and frequency, which indicates that task execution time on 3D processors can be substantially reduced. Moreover, the performance improvement of memory intensive applications executed on 3D processors is much larger than that of conventional 2D processors when both 2D processors and 3D processors increase their core operating frequencies. Further, the authors obtained thermal profiles of cores with consideration of the temperature-dependent leakage power consumption. The thermal profiles show that the maximal core operating frequency of 3D processors is lower than that of 2D processors under same peak temperature constraint.

By carefully conducting thermal simulations, Zhou et al. [52] showed that vertically adjacent layers of 3D processors are with strong temperature correlations. The authors further observed that not only the cores in the vertical direction have strong temperature correlations, but also the temperatures of cores that are far from the heat sink are usually higher than that of those cores close to the heat sink. This is mainly because the cores close to the heat sink can dissipate heat more quickly than other cores away from the heat sink.

Chatterjee et al. [74] studied the thermal coupling in 3D processors where a processor die containing several cores and a cache die containing an SRAM array are vertically stacked. The authors showed that the core power and core temperature have a significant impact on the cache temperature. Since applications executing on cores may be time-varying, power and temperatures of cores are likely to fluctuate over time. As a result, SRAM blocks in the cache of the 3D architecture have high spatial and temporal temperature variations. The high temperature and temperature variations will incur 30% performance degradation, two times increase in array leakage, and accelerated device aging. In addition, the authors found that spatial and temporal variations in SRAM block performance can be formulated a function of core power variations.

From the perspective of Amdahl's law, Yavits et al. [38] studied the thermal effects on scalability and performance of 3D processors. Unlike actually measuring temperature, the authors focused on qualitative trend analysis. By carefully choosing analytical models, the authors found that the peak temperature of 3D processors grows with core number and task parallelism. This leads to the fact that cores of 3D processors may reach their temperature limits before violating power constraints. Moreover, the authors showed that the peak temperatures of cores are likely to exceed the temperature limit of the DRAM vertically stacked on the core layers, which complicates processor-DRAM integration and reduces the scalability of 3D processors.

Tavakkoli et al. [75] investigated several key thermal attributes of 3D processors. The authors found that when more processing is done by the device layer closest to the heat sink, 3D processors exhibit best performance in reducing hotspot temperature. Within a device layer, the uneven power distribution among processors will result in the deterioration of hotspot temperature. At the same power density, increasing the ratio of processor area to the total chip area can efficiently reduce hotspot temperature for both the uniform TSV arrangement and the centralized TSV arrangement. However, when the ratio is greater than 40%, the performance achieved by the uniform TSV arrangement is superior to that of the centralized TSV arrangement, but when the ratio is less than 40%, the performance achieved by the centralized TSV arrangement is superior to that of the uniform TSV arrangement.

Knechtel et al. [76] observed that peak temperatures of 3D processors have a close relationship to the number of stacked layers. To be specific, the peak temperatures achieved by the GSRC and IBM-HB+ benchmarks performed on the three-layer 3D processor are 2.1 and 1.4 times higher than those performed on the two-layer 3D processor, respectively. The peak temperatures achieved by the GSRC and IBM-HB+ benchmarks performed on the four-layer 3D processor are 3.1 and 2.1 times higher than those performed on the two-layer 3D processor, respectively. Although there exists a thermal path toward the package, the authors found that most of heat generated by the layers away from the heat sink can only be dissipated through the heat sink. Therefore, a large amount of heat should overcome the "thermal barrier" caused by the bonding thermal resistance between layers.

3.2. Summary and discussion

This section reviews works from the perspective of thermal characteristic exploration of 3D processors. These works indicate that while 3D stacking technology offers many advantages, the benefits from this technology may be potentially offset by thermal issues that significantly affect chip performance. Therefore, it is crucial to design effective thermal management mechanisms for 3D processors. In the next sections, a system level approach from the perspectives of floorplanning, memory management and task scheduling is utilized to summarize thermal-aware optimization techniques that can be adopted to alleviate thermal issues for 3D processors.

4. Thermal-aware floorplanning

In this section, we summarize the works on thermal-aware 3D IC floorplanning methods that can be utilized to guide the early design of 3D processors. These methods includes thermal through-the-silicon via (TTSV) insertion based floorplanning (Section 4.1), force-directed technique based floorplanning (Section 4.2), meta-heuristic based floorplanning (Section 4.3), and other floorplanning approach (Section 4.4).

4.1. TTSV insertion based floorplanning

Song and Zhang [77] proposed an efficient TTSV placement method to achieve peak temperature minimization for 3D ICs. Due to the difficulty in simultaneously optimizing the heat flows in all directions, the TTSV floorplanning is separated into two parts: one is the vertical TTSV floorplanning and the other is the horizontal TTSV floorplanning. The vertical TTSV floorplanning is formulated as a convex programming problem to implement the allocation of vias to different layers. The horizontal TTSV floorplanning that is designed based on the techniques of path counting and heat propagation optimizes the process of assigning vias within one layer to different tiles. At each TTSV floorplanning level, the TTSV floorplanning scheme implements iterative alternation between vertical TTSV distribution and horizontal TTSV distribution. Evaluation results present that compared to benchmarking floorplanning schemes, the presented method can reduce the number of TTSVs by up to 68% with similar runtime for achieving the required temperature.

Li et al. [78] presented a TTSV insertion based algorithm to reallocate whitespace for 3D ICs. The proposed algorithm jointly optimizes the number of TTSVs, total wirelength, and performance in microarchitecture such that i) the required temperature can be reached by using TTSVs as few as possible, ii) the total wirelength is not significantly enlarged, and iii) the performance estimation of the design is not seriously degraded. To achieve a tradeoff among these optimization objectives, the requirements of TTSV number are first estimated based on an explicit thermal profile model. Then, TTSV requirements and other two optimization objectives are formulated as linear programming problems for simultaneously dealing with multiple optimization objectives and constraints. Simulation results reveal that the presented algorithm can reduce the number of TTSVs and total wirelength by 14% and 2%, respectively. At the same time, the results also demonstrate that the algorithm can redcue the number of TTSVs by 60% while keeping performance constant.

Wen et al. [79] presented a novel TTSV insertion technique for 3D floorplanning. In the beginning, a power density clustering approach is

utilized to cluster all modules and place every module cluster to one of the regions that are partitioned by via-channels for the purpose of handling large-scale designs. The via-channel is only introduced to simplify the complexity of TTSV insertion. After modules are placed, in order to satisfy temperature constraints, the positions of to-be-inserted TTSVs are iteratively determined by adopting a fast analytical computation technique combined with a precise heat model. The basic idea of this technique is to insert extra TTSVs near the critical position where peak temperature occurs such that the heat at that critical space can be quickly released. Experimental results demonstrate that the presented approach achieves remarkable peak temperature reduction with an acceptable amount of TTSVs.

Budhathoki et al. [80] proposed a two-stage 3D floorplanning solution. The first stage is designed to optimize packing area, total wirelength, and signal through-the-silicon vias (STSVs) using an existing 3D floorplanning tool Corblivar. The second stage is the thermal assessment, which aims to meet the thermal constraints by using as few TTSVs as possible. In this stage, an iterative TTSVs deployment algorithm is utilized to modify the thermal conductivity of design regions that exhibit local maximum temperatures. This algorithm will not terminate until the chip peak temperature is no more than the specified threshold temperature. Experimental results demonstrate that the presented floorplanning solution reduces peak temperature by 100°K at 0.5% TTSV density.

4.2. Force-directed technique based floorplanning

Zhou et al. [81] developed a force-directed technique based 3D floorplanning solution. The solution consists of two parts: one is the global placement and the other is the legalization. To be specific, an initial 3D layout is first generated by spreading blocks laterally in the 2D plane, and then the positions of these blocks are optimized in continuous 3D space. A 3D force-directed optimization with layer assignment algorithm will be invoked once the sum of overlapping areas between blocks is reduced to a specified threshold. After several iterations of the algorithm, the blocks are uniformly placed between layers and within layers. After the global placement part, a multilayer packing solution with few residual overlaps is generated. In order to produce a feasible placement, the legalization part slightly modifys the solution to obtain a non-overlapping packing while maintaining the original topological relationships between blocks unchanged. Experimental results show that compared to benchmarking floorplanning solutions, the proposed scheme can reduce chip area, total wirelength, via count and peak temperature by 6%, 16%, 22%, and 6%, respectively.

Huang et al. [82] proposed a three-stage 3D floorplanning method by utilizing the force-directed technique. The first stage is the lateral spreading where modules can only be moved laterally. When the total area of modules reaches a specified threshold, this stage will stop and output a 2D floorplanning solution. An existing thermal tool ISAC is utilized in this stage to implement the calculation of lateral thermal force. The second stage is the 3D simultaneous spreading, which spreads modules from 2D space to 3D continuous space. This stage utilizes an efficient method that is developed based on power density and 2D thermal analysis technique to derive the thermal force. The third stage is an iterative layer allocation and global spreading process, which implements the assignment of modules to chip layers. This stage generates an optimal solution with better thermal distribution and smaller chip area. Evaluation results reveal that the presented 3D floorplanning approach averagely reduces the temperature by 8% and runtime by 10.7% with no more than 3% increase in chip area and wirelength.

Kim et al. [83] proposed a three-stage 3D floorplanning solution based on force-directed technique. The first stage is the initial placement that performs the calculation of initial cell locations by using force-directed quadratic placement method. The second stage is the global placement the aims to reduce the number of cell overlaps by utilizing the move fore and the hold force. Since rapid cell movement may reduce the global placement quality, the process of removing overlaps is gradually implemented. The global placement stops when the remaining cell overlap amount is no more than the predefined overlap rate. The third stage is the detailed placement that utilizes an existing placer integrated in the tool of Cadence system-on-chip Encounter to perform detailed cell placement. Experimental results demonstrate that the constructed 3D ICs using the presented floorplanning approach can shorten the wirelength by up to 25% compared to traditional 2D ICs.

Athikulwongse et al. [84] proposed two effective heuristics that implement force-directed temperature-aware placement. The first heuristic is a through-the-silicon via (TSV) spread-alignment method which is designed based on two thermal properties of TSVs. First, TSVs occupy placement areas while consuming no power. Second, TSVs dissipate most of the heat via the polymer adhesive between layers. Based on the two thermal characteristics, the TSV spread-alignment method achieves uniform thermal conductivity by laterally spreading TSVs in each layer while increasing vertical overlaps between TSVs across multiple layers in a 3D stack by perturbing TSV positions. The second heuristic is a thermal coupling-aware placement method, which is developed based on the observation that the temperature at a certain position is related to its thermal conductivity and power density. It uses thermal conductivity-based force to place the logic cells on each layer while utilizing power density-based force to position TSVs. Experimental simulations show that compared to benchmarking methods, the presented heuristics can shorten the routed wirelength at a similar temperature, or reduce the temperature under similar routed wirelength.

4.3. Meta-heuristic based floorplanning

Cuesta et al. [85] focused on jointly minimizing the peak temperature and total wirelength of 3D ICs. Since simultaneously achieving the minimization of the two objectives is NP-hard, the authors decomposed the original multi-objective optimization problem into two subproblems: minimizing the peak temperature and minimizing the total wirelength. After an accurate analysis of 3D IC thermal behaviors, the authors formulated two mixed integer linear program (MILP) problems, and utilized an existing MILP solver to solve the formulated problems. Although MILP is a feasible technique, the MILP based algorithm will become complicated as the problem size increases. Given this, the authors developed two straightforward multiobjective evolutionary algorithms (MOEAs) in [36] to replace the MILP technique. MOEAs are stochastic optimization heuristic algorithms that adopt the population genetics to explore the solution space for a given problem. The operations of selection, crossover, and mutation are used to generate a set of solutions that are conducive to obtain the best solution. In [86], the authors further utilized MOEAs to optimize the placement of TSVs and functional units. Experimental results show that MOEAs can achieve a better tradeoff between the quality of solutions and the runtime required to derive these solutions.

Dash et al. [87] presented an evolutionary computation based floorplanning approach to achieve peak temperature minimization for 3D NoCs. The basic idea is to separate the heat sources from each other as much as possible while placing them as close as possible to the layer edges such that heat can be dispersed across the whole layers. The proposed floorplanning solution is achieved by two stages. At the first stage, an original floorplan that ignores the thermal properties of functional units is generated. At the second stage, MOEA is utilized to obtain an optimized floorplan based on the original floorplan. First, the operations of selection, crossover, and mutation in MOEA are applied to the original floorplan to obtain a floorplan set that takes thermal properties of functional units into consideration. Then, the floorplan with best fitness value is selected as the candidate floorplan, and the corresponding thermal diagrams are derived for every layer. The above two procedures at this stage will be repeated until the stopping criterion is satisfied. At the end of the second stage, an optimal floorplan with minimal on-chip peak temperature is generated.

Saha and Sur-Kolay [88] proposed a MOEA based scheme to perform the TSV placement and assignment for jointly optimizing the power density, wire congestion, TSV boundary distance and inter-layer wirelength. In the presented scheme, an initial floorplanning is first generated in which device layers with intra-layer routing interconnects are tentatively placed. Then, TSV candidate locations, inter-layer net coordinates, and the power profiles and wire congestion profiles of each device layer are collected. Based on these collected information, MOEA is utilized to generate the optimal TSV positions and the best assignment of TSVs to inter-layer nets. Experimental results show that average performance of the four objectives achieved by the presented solution is at the allowable range, and convergence time required by the proposed scheme is reasonable.

Chen and Ruan [89] proposed a 3D floorplanning method to enhance chip reliability and thermal dissipation. First, 2D modules are grouped into several clusters, and a module cluster with high power density will be allocated to the layer near the heat sink. Second, the position of each module is derived by utilizing simulated annealing for jointly minimizing total wirelength, chip area, and STSV density of every cluster. Third, STSVs are placed to suitable locations for enhancing chip reliability by using a redundant STSV insertion algorithm and a modified Ford-Fulkerson algorithm. Finally, the optimal number of TTSVs are determined for guaranteing the peak temperature constraint and reducing the cost of TTSVs. Experimental results present that compared to benchmarking floorplanning solutions, the proposed 3D floorplanning method can effectively improve chip reliability and reduce peak temperature by using a minimal number of TTSVs.

Lin et al. [90] proposed a four-stage 3D floorplanning algorithm that has the ability to handle different types of modules simultaneously. The first stage is the layer assignment that assigns all modules to different layers by utilizing simulated annealing technique. The second stage is the global distribution that adopts an analytical based method to distribute each module across its placement region with consideration of optimizing total chip wirelength. The third stage is the local legalization that utilizes integer linear programming technique to determine the exact location or the shape of each module if the module is a soft module or a folding module. The fourth stage is the monolithic inter-tier via or TSV assignment that uses network flow algorithms to perform the allocation of monolithic inter-tier vias or TSVs to proper whitespace.

Tabrizi et al. [91] proposed a force-directed simulated annealing based 3D floorplanning framework. Unlike the standard simulated annealing that selects a random neighbor solution to replace the current solution, the operation of the new solution selection in the presented framework is performed in a probabilistic way. That is, both the acceptance of a move and the selection of a new solution are based on the probability that is specified by the molecular force modeled systems. The presented framework can be further modified for thermal-aware 3D floorplanning, the basic idea of which is to move the cells in the regions with high temperatures to the regions with low temperatures by performing thermal force calculation. Experimental results show that, without sacrificing the quality of solution, the execution time of the proposed solution can be dramatically reduced since the force-directed move method accelerates the convergence speed.

Knechtel et al. [76] extended the existing 3D floorplanning tool Corblivar that adopts simulated annealing techniques to tackle the problems of multi-objective 3D IC floorplanning optimization. The developed method first determines the contiguity (i.e., spatial relationship) between modules for a given floorplanning layout. Then, the system-level timing of the layout is evaluated, and the resulting feasible voltage for each module is derived. Next, a bottom-up procedure of merging modules into all candidate voltage-volume arrangements is implemented. This procedure adopts heuristic yet efficient pruning techniques such that the proposed approach can be readily integrated into Corblivar's inner loop with no prohibitive increase in computational efforts. Finally, a top-down process is performed to choose the optimal subset of voltage volumes with consideration of design constraints (e.g., fixed outlines and critical delays).

4.4. Other floorplanning approach

Li et al. [92] proposed a thermal-aware power/ground (P/G) TSV planning method. First, an initial P/G grid without TSVs for every layer is constructed, and the analysis of initial thermal and the calculation of resistance of each metal wire are performed. Second, sensitivity estimation grids on every layer are created and the whitespace between blocks for every layer is swept. Third, the sensitivity of the midpoint of each grid is derived for the purpose of guiding the subsequent TSV insertion process. Fourth, TSVs are inserted into the P/G network, and the overall IR drop under the updated thermal profiles is calculated. The above process from the second step to the fourth step will continue until the voltage of each node is no less than a specified threshold or the overall IR drop no longer decreases. Evaluation results demonstrate that the presented P/G TSV planning approach not only decreases the number of violated nodes by 82.4%, but also improves maximum IR drop by 42.3%.

Xu and Chen [93] proposed an efficient thermal analysis method for fixed-outline 3D floorplanning. First, the thermal distributions of all blocks located at different positions are simulated before floorplanning. Second, bilinear interpolation technique is utilized to achieve quick temperature estimation during floorplanning according to the simulated thermal profiles. Third, a probability-based whitespace redistribution algorithm is performed before subsequent TSVs insertion to balance the whitespace distribution for the specified floorplan. Finally, a heuristic, combining wirelength and temperature minimization with the shortest path and min-cost-max-flow, is utilized to allocate the TSVs. Evaluation results reveal that the presented approach reduces 6.7% peak temperature on average with shorter runtime compared with the benchmarking schemes for 3D fixed-outline floorplanning.

Knechtel and Sinanoglu [94] proposed a novel floorplanning methodology to tackle the problem of thermal leakage of secret computation/activity patterns within 3D ICs. The authors found that TSV distributions and power density distributions are two main factors that determine the correlation between activity/power patterns and thermal behaviours. Given this, the basic idea of the presented 3D floorplanning methodology is to decorrelate the thermal behaviour from underlying activity/power patterns. During the iterative floorplanning, several 3D floorplans that have smaller correlations coefficients and spatial entropies are remarked as possible optimal solutions with respect to thermal leakage, while the final best solution also takes other metrics such as critical delays and wirelength into consideration. Experimental results reveal that compared to benchmarking power-aware floorplanning solutions, the proposed floorplanning methodology achieves average 13.22% reduction in peak temperature while only increasing 1.08% wirelength and 5.38% power consumption.

4.5. Summary and discussion

This section reviews works from the perspective of thermal-aware 3D IC floorplanning in order to guide the 3D processor design. These works indicate that wirelength and temperature are two key concerns in the 3D IC design phase. A closer placement of the components will result in shorter wirelength but higher temperature because the heat of each component is difficult to dissipate. Likewise, the farther apart the components, the lower the temperature, but the longer the wirelength. As a result, it is necessary to consider the joint optimization of wirelength and temperature, instead of only taking one of them into account at early design time. TTSV insertion, force-directed technique, and meta-heuristic are the three most commonly adopted floorplanning methods for achieving a better tradeoff between shortening wirelength and lowering temperature.

Table 1
A summary of references focusing on thermal-aware memory management.

Concentration	Reference	Method
Optimization for peak temperature (Section 5.1)	[95]	Three-step memory mapping
	[96]	Adaptive data placement
	[97]	Memory address management
Optimization for throughput (Section 5.2)	[98]	Memory refresh control
	[99]	Circuit design+memory refresh control
	[100]	Adaptive cache access mechanisms
	[101]	Adaptive thermal-aware routing
Optimization for energy (Section 5.3)	[102]	Memory refresh control
	[103]	Memory refresh control
	[104]	Cache compression
	[105]	Energy-aware routing
Combination optimization for peak Temperature, throughput, and energy (Section 5.4)	[106]	Bank reordering+bank swapping
	[107]	Temperature prediction based liquid cooling
	[108]	Adaptive cache refresh management
	[109]	An integrated solution
	[110]	Cool-path based routing solution
	[111]	Dynamic buffer allocation
Optimization for reliability (Section 5.5)	[112]	Variation-aware wearout management
	[113]	Error-correction-code based fault-tolerance
	[114]	Error-correction-code organization technique
	[115]	Thermal guard-band setting strategy
	[51]	Operating point tuning technique

5. Thermal-aware memory management

In this section, we review the works on thermal-aware memory management for 3D processors. As listed in Table 1, these works can be divided into five categories according to their concentrations: 1) Optimization for peak temperature (Section 5.1), 2) optimization for throughput (Section 5.2), 3) optimization for energy (Section 5.3), 4) combination optimization for peak temperature, throughput, and energy (Section 5.4), and 5) optimization for reliability (Section 5.5).

5.1. Optimization for peak temperature

Hsieh and Hwang [95] proposed a memory mapping scheme to optimize the peak temperature of 3D processors. The proposed algorithm considers both physical level and software level concerns. In physical level, a memory allocator should jointly take the power behavior and physical location of a memory block into consideration. In software level, the temperature of a memory block is in fact dynamically changing during program execution. Taking into account the above two aspects, the proposed memory mapping scheme consists of three steps. The first step aims to find candidate memory configurations for a memory system with given parameters. The second step analyzes memory requirements of an application over time, and groups memory blocks with similar behaviors in terms of access frequency into several segments. Based on the above two steps, the third step performs memory mapping by using MILP techniques for achieving peak temperature minimization. Evaluation results demonstrate that compared to the straightforward mapping solution, the presented memory mapping method lowers peak temperature by 17.1 °C and 9.9 °C for single-core systems and multicore systems, respectively.

Beigi and Memik [96] proposed an adaptive data placement algorithm to minimize the peak temperature of hybrid caches consisting of STT-RAM and SRAM. The proposed algorithm first acquires the temperature information of each bank by using on-chip thermal sensors. Based on the temperature information, banks are then classified into two categories: one is the semi-hot status and the other is the normal status. For banks in the normal status, the proposed algorithm performs intra-bank data migration; for banks in the semi-hot status, the proposed algorithm performs inter-bank data migration. As a result, the data with high access frequency will be migrated from hot banks to cool banks with low read/write latency such that the access number of hot banks is reduced. Experimental results show that compared to benchmarking schemes, the proposed approach can lower the peak temperature by up to 5.6 °C with 11.6% performance improvement and 6.5% power reduction.

Meng and Coskun [97] proposed a memory address mapping algorithm to jointly optimize the DRAM peak temperature and temperature variance of 3D processors. The proposed policy is tailored to memory-intensive applications that exhibit spatial variations in the access rates of DRAM banks. The basic idea of this strategy is to map the memory addresses that are frequently accessed to the physical banks that have low temperatures during virtual-physical address matching stage. According to the average-case analysis, this memory mapping strategy can be determined statically and thus incurs no runtime time overheads. Moreover, it can be wisely updated if the average-case workload has a dramatic change. Experimental results present that the proposed memory address management algorithm can reduce 1.42 °C DRAM peak temperature and 1.6 °C temperature variations.

5.2. Optimization for throughput

Guan and Wang [98] found that the peak temperature based all-bank DRAM refresh strategy for 3D processors incurs significant performance degradation. Given this, the authors proposed a novel temperature-aware DRAM refresh strategy with consideration of alleviating performance penalty. The proposed technique first tracks the temperature of each bank, and then adjusts the refresh rate of each bank according to the tracked bank temperature. To be specific, this technique only refreshes the memory banks reaching the peak temperature at a high rate, while refreshing the memory banks that do not reach the peak temperature at a low rate. As a result, this refresh scheme provides more memory read/write access, which indicates memory performance is improved. Experimental results show that system throughput achieved by the presented approach is up to 12.63% higher than that of the benchmarking all-bank refresh strategy.

By extending the work presented in [98], Guan et al. [99] further developed a circuit based DRAM refresh scheme. Two circuits including a temperature setting comparison circuit and a counter updating circuit are designed in order to collect the exact operation temperature of DRAM banks. The authors discussed in detail the hardware overheads of all key components in two circuits. Extensive simulations are con-

ducted to validate the effectiveness of the circuit based DRAM refresh scheme. Simulation results demonstrate that with the help of two designed circuits, the temperature-aware DRAM refresh strategy presented in [98] achieves a remarkable performance improvement over benchmarking refresh schemes.

Xiao et al. [100] proposed a novel design method that can accommodate temperature-related delay variations to optimize the average performance of last-level cache. The authors demonstrated this method with two types of thermally adaptive cache access mechanisms. The first mechanism utilizes the interface between the core and adjacent cache bank to control the cache access time that is formulated as a function of temperature. It reduces the number of cache access cycles that are proportional to temperature drops during application execution, thus effectively enhancing cache performance. The second mechanism adjusts the operating frequencies of cores and cache to match the time-varying cache hit time. It attempts to boost the core frequency when the temperature of vertically adjacent cache bank is low, and maintain a constant cache access based on the SRAM temperature-delay profile. Experimental results present that compared to benchmarking schemes utilizing worst-case design margins, both proposed mechanisms can enhance cache performance by more than 20% with extra up to 3% energy efficiency improvement.

Jiang et al. [101] designed an adaptive routing scheme for 3D NoCs. In the presented scheme, the messages that need to be transmitted are first routed in the horizontal layer using a 2D intra-layer routing method until they reach the intermediate router, and then vertically moved to the destination router in other layer. However, these messages may not reach the desired intermediate router due to the bad thermal profiles of this intermediate router. In this case, the messages will be first passed down to next layers until a feasible intermediate router is found, and then routed again. Since the authors assumed that there are no overheated routers in the bottom layer that is closest to the heat sink, messages could be successfully transmitted from the intermediate router located at the bottom layer to the destination router in worst cases. Experimental results reveal that the developed routing approach can boost throughput by up to 56.9%.

5.3. Optimization for energy

Sadri et al. [102] proposed a memory refresh control based scheme to optimize the energy consumption of 3D processors. The proposed scheme considers both peak temperatures and temperate variations of memory banks in lateral and vertical directions. The authors observed that the uniform refresh rate for all memory banks will result in a mass of hotspots and high temperature variations. Based on this observation, the basic idea of the proposed scheme is to enable each bank to intelligently select refresh rate according to its own peak temperature. For the purpose of quantifying the advantages of proposed management scheme, a virtual infrastructure integrating the whole key features of 3D processors is built. Experimental results present that the developed memory refresh approach can reduce the DRAM refresh power by up to 16%.

Shin et al. [103] presented a memory refresh scheme to improve the DRAM power consumption of 3D processors. The key idea is to map the page table to a specified physical address space such that memory refresh operations of this physical address space can be self-managed by wisely refreshing rows with valid data. However, this idea is difficult to implement because it requires extensive modifications to the operating system and system hardware. Given this, the authors divided the management structure of page table into two parts: one for the page directory and the other for the page table. As a result, the page table can point to a specified physical memory address by inspecting the allocated or deallocated the page frame numbers. The main advantage of this solution is that it only requires a slight modification of the operating system without the need to modify the underlying system hardware. Evaluation results demonstrate that the presented solution reduces the memory refresh power by up to 98% during idle time, while generating only 64 KB of DRAM register overhead and 16 KB of SRAM register overhead.

Park et al. [104] targeted at optimizing the static energy consumption of 3D processors by using cache compression technique. The authors explored the relationship between two factors. One is the whole static energy consumed by cores and caches, and the other is the cache compression rate. The authors observed that there exists a best pair of the two factors, which indicates that there exists an optimal compression rate. Based on this observation, the authors presented a cache compression policy which enables a cache to selectively compress its stored data according to data access frequency. By compressing the cache data with less access rather than compressing all cache data, the data decompression overhead can be minimized. Experimental results show that compared to benchmarking cache management approaches, the proposed cache compression scheme can reduce up to 40% energy consumption.

Yao et al. [105] developed an efficient routing solution to optimize the energy consumption of 3D NoCs. The presented solution is implemented by a thermal-aware routing unit consisting of a shortest path routing unit, a global temperature table, and a thermal-effect modeling unit. For every packet that needs to be transmitted, all the shortest routing paths from the source router to the destination router are first listed by the shortest path routing unit. The thermal-induced power losses of these shortest routing paths are then compared with each other by the thermal-effect modeling unit. Specifically, the thermal effect modeling unit acquires the information of all the shortest routing paths from the shortest path routing unit and obtains the temperature information from the global temperature table. Based on the information, the thermal-induced power loss of every shortest routing path is calculated. Finally, the routing path that incurs minimal thermal power loss will be chosen as the optimal routing path. Experimental results demonstrate that the presented routing algorithm reduces energy consumption by up to 25%.

5.4. Combination optimization for peak temperature, throughput, and energy

Unlike some works that separately optimize peak temperature or throughput, Lin and Lin [106] focused on jointly optimizing peak temperature and throughput. The authors proposed a thermal/performance-aware memory address mapping scheme that consists of two parts: one is bank reordering and the other is bank swapping. The basic idea of bank reordering is to decrease the vertical stacking of these banks that are frequently accessed in the adjacent layers. The basic idea of bank swapping is to switch banks based on the design tradeoff between thermal constraints and bandwidth constraints. In bank swapping, the banks that are far from the heat sink are swapped to these banks that are near the heat sink. As a result, channel access and bank access of different mappings can be redistributed. Simulation results show that the thermal/performance-aware memory address mapping scheme can lower up to 12.3 °C peak temperature at the expense of acceptable memory bandwidth degradation.

Beigi and Memik [107] presented a novel 3D processor-cache architecture where a liquid cooling layer is placed between a processor layer and an STT-RAM last-level cache layer. The authors proposed TESLA to optimize the energy consumption and temperature for the presented 3D architecture. TESLA consists of a monitoring component, sampler, predictor, and flow rate controller. The basic idea of TESLA is to forecast the necessary flow rate for cooling such that the energy consumption and temperature can be optimized. Based on the information obtained from the monitoring component and sampler, the predictor can forecast the future temperatures of cache banks. Based on these predicted temperatures, an optimal flow rate for the cooling layer is derived by the rate adjustment controller. Evaluation results show the presented method can achieve a throughput improvement of up to 19.1% and a power reduction of 14.6%.

Li et al. [108] proposed an adaptive cache refresh scheme that can balance processor performance with memory energy consumption of 3D

processors. The proposed policy is mainly composed of two parts: one is a system profiling strategy and the other is a scale-and-check strategy. The system profiling strategy aims to assess the sensitivity of an application to cache latency under various memory refresh rates. The scale-and-check strategy aims to monitor the margin between peak temperature and temperature threshold of the 3D processor under various memory refresh rates. Based on the above two strategies, an optimal refresh rate can be found where the benefits of improved cache access latency can balance out the lowered core frequency due to thermal control. Simulation results show that the proposed adaptive cache refresh management can achieve 105% throughput improvement with 72.5% energy saving.

Asad et al. [109] developed an integrated solution to optimize the performance and energy consumption of 3D processors. The proposed solution first randomly assigns some applications to cores, and then executes these applications for a given time. After this given time, the throughput of all cores are predicted and a two-stage (inter-region and intra-region) mapping scheme based on the predicted throughput is utilized to allocate the remaining applications to cores. After this two-stage mapping, a reconfiguration scheme is performed in each reconfiguration time interval. The proposed reconfiguration scheme consists of two phases. One is a design phase that derives the average frequency required by workload and the configuration of cache hierarchy. The other is a runtime phase that determines the cache configuration and the voltages and frequencies of cores based on the information obtained at design phase. Experimental results show that the developed approach can improve 54.3% throughput and 61% energy-delay product.

Fu et al. [110] proposed a routing solution composed of a intra-layer routing algorithm and a inter-layer routing algorithm to optimize the peak temperature and throughput of 3D NoCs. The key idea of the intra-layer routing algorithm is to derive the coolest routing path among all feasible pathways such that the probability of occurrence of thermal hotspot can be reduced. The determination of the coolest routing path is converted to a shortest path problem that can be solved by utilizing dynamic programming technique. In the inter-layer routing algorithm, the downward layer that traffic loads are redirected downward to is first determined. The packets above the downward layer are then redirected to the determined downward layer while the remaining packets are routed by using a lateral-first and vertical-last routing method. In this routing method, packets first move from the source router to an intermediate router whose 2D coordinate is the same as the destination router, and then travel vertically to the destination router. Experimental results demonstrate that the presented routing solution achieves 3.9 °C reduction in peak temperature and 78.3% improvement in throughput.

Chou et al. [111] proposed a dynamic buffer allocation technique to improve traffic congestion of 3D NoCs. The basic idea of the developed buffer control technique is to avoid traffic congestion by lengthening the length of input buffer and shortening the length of output buffer of the near-overheated router. To be specific, the temperature of each router is first predicted and compared with each other. Then, the length of input buffer of the router with a higher predicted temperature is lengthened, and the length of its output buffer is shortened. Once a hotspot is detected, the capability of switching packets in the near-overheated router will be reduced. However, this process may result in traffic congestion around the near-overheated router. Given this, an existing congestion-aware routing technique is utilized to deliver packets away from the congested region such that the packet switching frequency is reduced and the temperature of the hotspot is decreased. Finally, the buffer length of this router is recovered to normal status. Experimental results reveal that the presented technology can reduce temperature deviation by 39.4% and boost system performance by 24.8%.

5.5. Optimization for reliability

Tajik et al. [112] addressed the problem of managing negative bias temperature instability (NBTI) induced wearout in a 3D multicore architecture. A variation-aware wearout management (VAWOM) scheme is proposed to reduce the cache and core wearout induced by NBTI such that the lifetime reliability of the 3D multicore architecture is improved. The proposed scheme takes into account two different types of variations. One is core variation in frequency, and the other is cache bank variation in access time. VAWOM utilizes task migration strategy to balance temperatures of cores such that the NBTI effects on cores can be alleviated. For the purpose of improving cache lifetime, a proactive recovery for cache banks is adopted. In each run, VAWOM first puts a selected cache bank into recovery mode (named “recovery bank”), and then migrates the data stored in this recovery bank to the earlier disabled/recovered cache bank. Experimental results show that VAWOM can reduce 30% threshold voltage degradation with negligible performance cost.

Wang et al. [113] proposed a design strategy to tolerant radiation-induced soft errors for 3D processors. This strategy can provide a strong fault-tolerance for weak DRAM cells with extremely little redundancy storage overhead and delay overhead. The authors observed that fault-tolerance to unrepaired weak memory cells can be achieved by using error correction codes thanks to the logic die stacked on multiple DRAM dies. Therefore, the basic idea is to exploit the configurability of error-correction-code decoding and the detectability of weak cells. However, the actual implementation of this idea is not easy due to the constraints of silicon cost and data access latency. Given this, the authors developed an effective implementation solution that achieves minimal delay and cost while improving the detection accuracy of weak cells and the ability to tolerate soft errors. Experimental results demonstrate that the presented solution can tolerate the weak cell rate by up to 1×10^{-4} with negligible performance degradation.

Han et al. [114] present an effective error-correction-code organization technique to enhance the reliability of 3D processors. By forming a heterogeneous error-correction-code organization across different memory layers, the proposed scheme can enhance error-correction-code capability in the memory layers with low reliability. It adopts free spare columns of one DRAM layer to store extra check-bits of other DRAM layers, thus requiring no additional redundant arrays. For example, the extra check-bits of upper memory layers can be stored in the unused spare columns of lower memory layers and vice versa. Therefore, the error-correction-code organization of each memory layer is configurable for operational environments. Experimental results reveal that compared with the benchmarking method, the presented technique can achieve three times performance improvement in terms of tolerating one bit-error.

Lim et al. [115] focused on improving the DRAM data reliability for 3D processors. The authors observed that the inaccurate peak temperature prediction of DRAM cells results in an incorrect DRAM refresh rate. When the DRAM refresh rate is insufficient, the DRAM data reliability is violated. However, when the DRAM refresh rate is too high, extra refresh power overhead is incurred. To tackle this problem, the authors proposed a thermal guard-band setting strategy, which takes into consideration ont only the data read delay of temperature sensor but also the positional difference between the DRAM cell and the temperature sensor. This strategy can provide accurate peak temperature prediction of DRAM cells and thus it can be safely utilized to control the DRAM refresh rate. Simulation results reveal that the developed method can reduce up to 50% refresh power overheads with guaranteed data reliability.

Lu et al. [51] proposed an effective technique to reduce the voltage-noise induced DRAM transient faults for 3D processors. This fault-tolerance technique is based on the following two observations. First, DRAM transient errors are closely related to processor activities via thermal and voltage coupling. Second, the bit-cell leakage of DRAM is in fact an accumulative process, which suggests that the leakage is more strongly correlated with IR drop compared with transient droop. Based on the two observations, the authors proposed an operating point tuning technique that jointly optimizes the operating frequencies of cores and the refresh rates of DRAM. This technique adopts dynamic frequency

Table 2
A summary of references focusing on thermal-aware task scheduling.

Concentration	Reference	Method
Optimization for peak temperature (Section 6.1)	[116]	Tasks with high power to cores with low thermal resistance
	[117]	Rotation scheduling based scheme
	[118]	Four-phase thermal-aware scheme
	[119]	Design time optimization+runtime adjustment
	[120]	Thread migration
Optimization for throughput (Section 6.2)	[121]	Approximated task allocation
	[122]	Approximated task allocation+task migration
	[123]	Core-memory co-scheduling
	[124]	Thermal-efficient synthetic real-time scheduling
	[125]	Thermal-throttling server based technique
	[126]	Branch-and-bound method based scheme
Optimization for energy (Section 6.3)	[46]	Simulated annealing based three-stage method
	[127]	Three-stage task scheduling framework
Combination optimization for peak Temperature, throughput, and energy (Section 6.4)	[128]	Thermal characteristic extraction based method
	[129]	Voltage assignment policy
	[130]	Bottom-to-up task scheduling
	[131]	Objective decomposition method
	[132]	Offline mapping and online mapping
	[133]	Genetic algorithm based task mapping
	[134]	Task mapping and core pipeline control
Optimization for lifetime (Section 6.5)	[135]	Aging-aware runtime task mapping framework

scaling for cores to implement a flexible borrow-in mechanism that effectively improves DRAM resilience without incurring performance degradation. Experimental results demonstrate that the presented technique boosts performance by 27% while achieving fault-tolerance.

5.6. Summary and discussion

This section surveys the literatures related to 3D processors from the perspective of thermal-aware memory management. These works indicate that peak temperature, throughput, and energy are the three main concerns for memory management of 3D processors. A lot of works focus on single-objective optimization that involves only a single objective function, i.e., separately optimizing peak temperature, energy or throughput. While some works aim to perform the multi-objective optimization that involves more than one objective functions, such as simultaneously optimizing peak temperature and throughput, peak temperature and energy, or throughput and energy. In addition to aforementioned concerns, reliability optimization for 3D processors has attracted much attention in recent years. These reliability related works mainly attempt to enhance 3D processor performance in terms of energy consumption or throughput while providing required fault-tolerance levels.

6. Thermal-aware task scheduling

In this section, we review the works on thermal-aware task scheduling for 3D processors. As listed in Table 2, these works can be divided into five categories according to their concentrations: 1) Optimization for peak temperature (Section 6.1), 2) optimization for throughput (Section 6.2), 3) optimization for energy (Section 6.3), 4) combination optimization for peak temperature, throughput, and energy (Section 6.4), and 5) optimization for lifetime (Section 6.5).

6.1. Optimization for peak temperature

Tsai and Chen [116] proposed an online task scheduling strategy to achieve peak temperature minimization for 3D processors. The proposed strategy allocates the tasks with high peak power to these cores which are close to the heat sink for the purpose of balancing power consumption among cores. To satisfy temperature and real-time constraints, the operating frequencies of cores are determined by their respective stack supervisors. A stack supervisor consists of two parts: one is a thermal predictor and the other is a core operating frequency manager. The thermal predictor derives each core temperature with consideration of vertical and horizontal heat effects. The core operating frequency manager selects cores for task execution and determines the core frequencies using first-fit policy with consideration of temperature and real-time constraints. Simulation results show that the proposed online real-time task scheduling scheme can prevent cores from overheating without sacrificing system performance.

Li et al. [117] proposed rotation scheduling based algorithms to optimize the peak temperature of 3D processors. The proposed algorithms are tailored to the applications with inter-iteration data dependencies. By rotating down delays repeatedly, several flexible static direct acyclic graphs (DAGs) of an application are first generated. For each DAG, its task schedule with optimal peak temperature is then obtained and taken as a candidate task schedule by using a greedy heuristic scheme. Finally, the best task schedule with minimal peak temperature is selected among these candidate task schedules obtained previously. Simulation results present that the proposed rotation scheduling based solution reduces the peak temperature by up to 8.1 °C.

Cheng and Hsu [118] proposed a four-stage task scheduling scheme for 3D processors. In the first stage, based on the collected information about memory and task program, each task program is partitioned into several data segments that can fit the memory bank size. In the second stage, power consumption and memory access of cores are balanced by using a task allocation algorithm. In the third stage, a memory mapping algorithm is performed to prevent vertical adjacent memory banks from being simultaneously accessed. In the fourth stage, a round-robin scheduling with idle slot insertion scheme is adopted to schedule tasks under timing constraints. Simulation results demonstrate that the presented four-stage solution can effectively improve heat dissipation.

Chaturvedi et al. [119] focused on optimizing the peak temperature for throughput and periodic constrained applications. The authors proposed a two-stage approach with design time optimization and runtime adjustment. In design time, a thermal profiling algorithm based task allocation scheme first assigns tasks to suitable cores with consideration of power balance among cores under throughput constraints. Then the operating frequencies of tasks are scaled by using available slacks generated from the task assignment process. At runtime, based on the runtime information, tasks on the cores with high temperatures are migrated to the cores with low temperatures without violating the application periodic constraint. Experimental results show that compared with bench-

marking methods, the proposed two-stage approach reduces up to 14 °C peak temperature.

Zhao et al. [120] aimed to reduce the peak temperature and temperature variance of 3D processors by migrating threads among cores. The authors explored four thread migration schemes. The first one is the rotation scheme which rotates threads in a round-robin way. This scheme balances power consumption of cores, thus resulting in minimal thermal variance. The second one is the pair-wise scheme that wisely switches hot threads and cool threads. Both the rotation scheme and pair-wise scheme are static algorithms and thread migration is always carried out in the two algorithms even if the thermal variance of cores is in the acceptable range. The remaining two schemes are dynamic algorithms designed to avoid unnecessary thread migrations. Thread migrations are merely performed when the temperature variance of cores is large enough. Experiment results show that the presented thread migration algorithms can reduce up to 8 °C peak temperature with small thermal variation.

6.2. Optimization for throughput

Lung et al. [121] proposed an approximated task assignment strategy to maximize 3D processor throughput under thermal constraints. The proposed scheme is developed by rewriting temperature equations such that an incremental thermal update can be implemented. For each core, its current temperature is first obtained via using the incremental update method. Then, a new arrival task is allocated to the core with lowest temperature rise among unassigned cores. The above process will stop when all new arrival tasks are assigned to cores. However, this approximated task allocation algorithm assumes that all tasks are with the same finish time. This is a strong assumption since most tasks may not finish executions at the same time. Given this, the authors further considered the situation where both new arrival tasks and unfinished tasks exist in [122]. A task migration-aware scheme is proposed to perform task allocation. A migration-penalty function is defined to reflect task migration cost. The developed task migration-aware scheme makes decisions with consideration of the tradeoff between task migration cost and core temperatures. Experiment results show that the migration-aware method achieves 20.82 times speedup with no more than 4.39% throughput degradation compared to an exhaustive benchmarking method.

Chaparro-Baquero et al. [123] proposed a core-memory co-scheduling strategy for real-time tasks. The proposed method is designed based on a resource model which proactively and periodically suspends request services while guaranteing resource availability. This method utilizes the feasibility conditions for a periodic resource server to satisfy task timing constraints. Meantime, the method makes use of periodic behaviors of cores and memory, and achieves the deterministic guarantee for thermal constraints by formulating the temperature dynamics analytically. It wisely selects periodic server settings for memory bus arbitration and cores to meet the peak temperature limits of the processor layer and memory layers. Experimental results demonstrate that the presented method improves throughput by 19.5%.

Tsai et al. [124] proposed a synthetic task scheduling scheme for 3D processors. The developed method first allocates a suitable thermal budget to each general purpose core and special purpose core by using the technique of thermal size ratio detection. Then, based on the allocated thermal budget, the operating frequencies of both special purpose cores and general purpose cores are determined to meet thermal constraints. Next, a schedulability test is conducted for all tasks. If a task cannot pass the schedulability test, it will be refused since the timing constraint of this task is definitely violated. Otherwise, a global dispatcher will assign this task to a suitable core, and the task performs its execution on the allocated core by making use of the core's assigned thermal budget. Experimental results demonstrate that the presented scheme dramatically boosts task schedulability, indicating improved system throughput.

Tsai and Chen [125] presented an online task scheduling framework to achieve the tradeoff between temperature and schedulability. At runtime, based on current temperatures of all cores, a thermal-throttling dispatcher is utilized to assign available temperature slack to each core. When a new task has arrived, the dispatcher will assign this task to a suitable core when the task has passed the admission control. After that, a thermal-throttling server associated with the core will schedule the task. Dynamic voltage and frequency scaling (DVFS) technique is utilized to lower core temperature under task timing constraint. The current frequency assignment is fed back from each thermal-throttling server to the thermal-throttling dispatcher. Based on the feedback information, thermal-throttling dispatcher can adjust the temperature slack of each core to optimize the admission control's schedulability bound. Simulation results reveal that the presented online task scheduling method achieves a better tradeoff between temperature and schedulability.

Li et al. [126] proposed a two-stage runtime mapping scheme to reduce application running time under thermal constraints. The first stage is to select a specific shape of 3D cuboid core region for each application. A branch-and-bound method is utilized in this stage to search for the optimal tree node that corresponds to the best combination of core region shapes for applications. The second stage is to determine the exact core region locations for reducing the peak temperature and core fragmentation. To this end, round-robin method is utilized to position the core regions at one of the four corners of 3D processors, and then tasks of an application are assigned to the cores that are located at the unused core region by utilizing an existing mapping algorithm. Simulation results show that compared to benchmarking mapping approaches, the presented scheme achieves up to 48% reduction in application running time.

6.3. Optimization for energy

Cheng et al. [46] explored the tradeoff between interconnect energy and temperature when tasks are executed on 3D homogeneous multiprocessor system-on-chips (MPSoCs). The authors first formulated the thermal-constrained interconnect energy minimization problem, and then proved that the formulated problem is NP hard. Given this, a heuristic scheme is presented to address the energy minimization problem. The proposed heuristic scheme consists of three steps. In the first step, an initial solution to task-to-core allocation is generated by using a temperature-balanced method. In the second step, based on the initial solution, an intermediate solution is derived by using a developed greedy method for the purpose of achieving largest possible reduction in interconnection energy with consideration of thermal constraints. In the third step, the obtained intermediate solution is optimized by using simulated annealing techniques. The annealing procedure will stop when the intermediate solution can no longer be improved within given iteration number. Simulation results demonstrate that compared to thermal-balanced benchmarking solutions, the proposed heuristic can reduce more than 25% interconnect energy consumption on average while achieving almost same peak temperature.

Jin et al. [127] presented a three-stage task scheduling framework to minimize the energy consumption of voltage-frequency island based 3D MPSoCs. In the first stage, tasks are allocated to cores by using an energy-aware task scheduling algorithm with consideration of retaining maximum optimization space for subsequent voltage/frequency scaling to achieve computation energy minimization. In addition, the total application execution time is split into an execution timing series that facilitates the latter power balancing algorithm under timing constraints. In the second stage, cores are first allocated to core stacks for the purpose of reducing communication energy consumption. Several tasks are then re-allocated to cores for achieving stack power balance during all the timing episodes in the execution timing series. In the third stage, the process of mapping core stacks to the hardware platform and voltage-frequency islands partition is implemented. Through regarding every

core stack as a unit, the above complex process can be simplified to 2D problems. Experimental results demonstrate that the energy consumption of the presented three-stage task scheduling framework is reduced by 15.8% on average.

6.4. Combination optimization for peak temperature, throughput, and energy

Cox et al. [128] proposed a two-step approach to optimize the peak temperature and energy consumption of 3D processors. In the first step, a simple yet flexible physical chip model is utilized to extract thermal characteristics of the studied 3D processor architecture before task allocation. In the second step, a task allocation algorithm is carried out to find the optimal task-to-core mapping that matches the power distribution generated in the first step with considerations of achieving energy consumption minimization and meeting storage and timing constraints. The main advantage of this task mapping process is that it requires no iterative thermal simulations, which significantly reduce the time overhead to derive task schedules. Experimental results reveal that the presented two-step approach lowers peak temperature by 7 °C and reduces communication energy by 42%.

Liao et al. [129] proposed an online voltage assignment policy to optimize system throughput and peak temperature. The proposed policy first generates thermal profiles for all cores under various voltage levels. Based on these generated thermal profiles, the proposed policy then allocates an initial voltage level to each core for the purpose of reducing temperature increase. Tasks are assigned to suitable cores using an existing super-task-to-super-core scheme. A DVFS based voltage scaling technique is developed to quickly reduce the temperature of overheated cores. This technique not only triggers DVFS on the overheated cores, but also triggers DVFS on the cores that are vertically adjacent to the overheated cores. Experiment results show that compared to benchmarking schemes, the proposed policy can improve throughput by up to 32% and lower hotspot occurrences by up to 53%.

Cui et al. [130] proposed a bottom-to-up task scheduling scheme to optimize peak temperature and makespan for the 3D two-layer multi-core processor architecture. The proposed scheme first allocates tasks to the cores on the bottom layer near the heat sink with considerations of core power balance and task execution time. It then wisely migrates some tasks which can be in parallel executed from the cores on the bottom layer to the cores on the top layer for reducing makespan while improving thermal profiles of cores. Experimental results show that compared to benchmarking schemes, the proposed approach can reduce up to 7.95 °C peak temperature and 4% makespan.

Zhu et al. [131] proposed a two-stage task scheduling strategy to optimize makespan and peak temperature for 3D processors. Unlike some scheduling schemes that jointly optimize makespan and temperature, the proposed strategy decouples the joint optimization and optimizes makespan and temperature separately. The first stage aims to minimize makespan at design time. This stage utilizes the combination of exhaustive exploration and genetic algorithm to generate best super tasks consisting of several individual tasks for the purpose of makespan minimization. The second stage aims to optimize peak temperature at runtime. This stage develops a thermal rank model and a combined power model to quantify thermal efficiencies of all cores for best usage. Based on the two models, two heuristics are designed to allocate super tasks to available cores for achieving peak temperature minimization while ensuring that the optimized makespan derived in the first stage is not reduced. Experiment results show that compared to benchmarking methods, the proposed two-stage scheme can reduce the average peak temperature by 6.3 °C while improving performance by 6.8%.

Singh et al. [132] proposed an efficient 3D video application mapping strategy to jointly optimize peak temperature and energy consumption of 3D processors under throughput constraint. The proposed strategy consists of offline and online parts. The offline part first obtains the characteristics of the 3D video application and 3D processor architecture by using an offline analysis method. Based on these characteristics and throughput constraint, an optimal task mapping is then found by using an iterative search scheme considering power distribution of cores. For a computed task mapping during the iterative search process, a novel thermal analysis method is utilized to generate the temperature distribution of cores such that the power distribution of cores can be easily derived. The online part selects the optimal task mapping obtained from the offline part, and it is triggered immediately at the 3D video application startup. Simulation results show that compared to benchmarking mapping schemes, the proposed strategy can reduce 76% communication energy consumption and 4 °C peak temperature on average.

Shen et al. [133] proposed a two-stage task allocation method to achieve reduced peak temperature and minimal communication energy consumption for 3D NoCs. The first stage performs communication-aware task group assignment. It first divides the tasks that are communication intensive into multiple task groups with balanced power consumption, and then utilizes genetic algorithms to allocate these task groups to core stacks with consideration of communication energy consumption. The second stage performs thermal-aware allocation of the task in a task group to the core in a core stack. For achieving peak temperature minimization, the main idea of this stage is to allocate the task with maximal power consumption to the core with minimal cost until all tasks are assigned to cores. The cost of a core is jointly determined by its power consumption and cooling efficiency. Evaluation results reveal that the presented approach decreases peak temperature by up to 5.75K and reduces communication energy consumption by up to 63.34%.

Yoon et al. [134] presented a novel task scheduling scheme to jointly optimize energy consumption and throughput of 3D processors. The proposed scheme is composed of two parts: one is the thermal-aware task mapping and the other is the workload-aware core pipeline control. The basic idea of the thermal-aware task mapping is to assign hot tasks to these cores that are located at the side of processor die for quickly dissipating the heat generated by cores. The main idea of the workload-aware core pipeline control is to throttle the fetch and issue widths of cores according to the types of workloads for the purpose of reducing the dynamic power consumption of cores. If a core is to execute memory-intensive tasks, the pipeline width of the core will be dynamically reduced since long memory access latency can to some extent hide the performance loss caused by the reduction of core pipeline widths. On the contrary, if a core is to execute computation-intensive tasks, the pipeline width of the core will be conservatively adjusted since the performance of computation-intensive tasks is more sensitive to pipeline width than that of memory-intensive tasks. Experimental results reveal that the presented scheme can not only reduce 7.6% energy consumption but also improve performance by 0.4% on average.

6.5. Optimization for lifetime

Raparti et al. [135] proposed an application mapping framework to prolong the lifetime of 3D processors. The framework is implemented in two nested processes: one is the aging-aware inner-loop of application mapping and DVFS scheduling, and the other is the outer-loop of aging analysis of circuit and power delivery networks. In the inner-loop, an existing thermal- and communication-aware incremental-mapping heuristic is adopted to allocate tasks of an allocation to suitable cores. Simulated aging and power sensors are utilized to periodically feed back their monitoring values. Based on these measured values, DVFS technique is adopted in the application scheduling stage to wisely adjust the supply voltage and operating frequency of each selected core. In the outer-loop, aging analysis is performed at the end of the current epoch by utilizing the system-stats produced by the inner-loop over the last epoch. Based on the analysis, the aging parameters of 3D processors are hence updated for scheduling next application. Simulation results show that compared to benchmarking schemes, the proposed application mapping framework can prolong the chip lifetime by up to 25%.

6.6. Summary and discussion

This section surveys thermal-aware task scheduling techniques for 3D processors. Similar to the works on thermal-aware memory management presented in Section 5, many of the research efforts shown in this section are also dedicated to achieving single-objective optimization of peak temperature, energy, or throughput for 3D processors. Meanwhile, some works concentrate on the multi-objective optimization for the combination of peak temperature, throughput and energy. Several novel thermal-aware task scheduling schemes are proposed to jointly optimize peak temperature and throughput, peak temperature and energy, or throughput and energy.

7. Conclusions

In this paper, we review thermal-aware optimization techniques proposed for 3D processors by using a system level approach. We first survey the works on 3D processor architecture design and outline thermal characteristics of the constructed 3D processors. These works present that due to limited heat dissipation paths and higher power density, 3D processors are likely to suffer from more serious thermal issues compared to traditional 2D counterparts. We then summarize the works that aim to reduce thermal impact on performance improvement of 3D processors from the perspectives of floorplanning, memory management, and task scheduling. These works demonstrate that by carefully designing thermal-aware floorplanning, memory management or task scheduling schemes, the thermal impact on 3D processors is manageable. Therefore, it is expected that 3D processors will be the mainstream in the near future.

Acknowledgments

This work was supported in part by the Shanghai Municipal NSF under Grant 16ZR1409000, in part by the China HGJ Project under Grant 2017ZX01038102-002, in part by the National NFSC under Grant 61802185 and Grant 61872147, in part by Jiangsu NSF under Grant BK20180470, and in part by the Fundamental Research Funds for the Central Universities.

References

[1] ITRS, International technology roadmap for semiconductors, http://www.itrs2.net, (2018).
[2] J. Zhou, K. Cao, P. Cong, T. Wei, M. Chen, G. Zhang, J. Yan, Y. Ma, Reliability and temperature constrained task scheduling for makespan minimization on heterogeneous multi-core platforms, J. Syst. Software 133 (2017) 1–16.
[3] A. Agrawal, J. Torrellas, S. Idgunji, Xylem: enhancing vertical thermal conduction in 3D processor-memory stacks (2017) 546–559.
[4] J. Lin, J. Yang, Routability-driven TSV-aware floorplanning methodology for fixed-outline 3-D ICs, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 36 (11) (2017) 1856–1868.
[5] R. Salamat, M. Khayambashi, M. Ebrahimi, N. Bagherzadeh, LEAD: an adaptive 3D-NoC routing algorithm with queuing-theory based analytical verification, IEEE Trans. Comput. (2018) 1.
[6] A. Coelho, A. Charif, N. Zergainoh, J. Fraire, R. Velazco, A soft-error resilient route computation unit for 3D networks-on-chips, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2018, pp. 1357–1362.
[7] F. Le, S. Lee, Q. Zhang, 3D chip stacking with through silicon-vias (TSVs) for vertical interconnect and underfill dispensing, J. Micromech. Microeng. 27 (4) (2017).
[8] D. Zhang, J. Lu, 3D Integration technologies: an overview, Mater. Adv. Packag. (2017) 1–26.
[9] J. Lin, C. Huang, J. Yang, Co-synthesis of floorplanning and powerplanning in 3D ICs for multiple supply voltage designs, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2018, pp. 1339–1344.
[10] J. Lin, C. Huang, General floorplanning methodology for 3D ICs with an arbitrary bonding style, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2018, pp. 1199–1202.
[11] B. Ku, Y. Liu, Y. Jin, S. Samal, S. Lim, Design and architectural co-optimization of monolithic 3D liquid state machine-based neuromorphic processor, in: Proceedings of the ACM Design Automation Conference, 2018, pp. 1–6.
[12] Z. Wang, H. Gu, Y. Chen, Y. Yang, K. Wang, 3D Network-on-chip design for embedded ubiquitous computing systems, J. Syst. Archit. 76 (2018) 39–46.
[13] K. Kang, L. Benini, G.D. Micheli, Cost-effective design of mesh-of-tree interconnect for multicore clusters with 3-D stacked L2 scratchpad memory, IEEE Trans. Very Large Scale Integr. VLSI Syst. 23 (9) (2015) 1828–1841.
[14] W. Wang, Y. Han, H. Li, L. Zhang, Y. Cheng, X. Li, PSI conscious write scheduling: architectural support for reliable power delivery in 3-D die-stacked PCM, IEEE Trans. Very Large Scale Integr. VLSI Syst. 24 (5) (2016) 1613–1625.
[15] A. Abdallah, 3D integration technology for multicore systems on-chip, Adv. Multicore Syst. On-Chip (2017) 175–199.
[16] X. Hu, D. Stow, Y. Xie, Die stacking is happening, IEEE Micro 38 (1) (2018) 22–28.
[17] Z. Ghaderi, A. Alqahtani, N. Bagherzadeh, AROMa: aging-aware deadlock-free adaptive routing algorithm and online monitoring in 3D NoCs, IEEE Trans. Parallel Distrib. Syst. 29 (4) (2018) 772–788.
[18] T. Zhang, C. Xu, K. Chen, G. Sun, Y. Xie, 3D-SWIFT: a high-performance 3D-stacked wide IO DRAM, in: Proceedings of the ACM International Great Lakes Symposium on VLSI, 2014, pp. 51–56.
[19] D. Kim, K. Athikulwongse, M. Healy, M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y. Lee, D. Lewis, T. Lin, C. Liu, S. Panth, M. Pathak, M. Ren, G. Shen, T. Song, D. Woo, X. Zhao, J. Kim, H. Choi, G. Loh, H. Lee, S. Lim, Design and analysis of 3D-MAPS (3D massively parallel processor with stacked memory), IEEE Trans. Comput. 64 (1) (2015) 112–125.
[20] C. Chou, A. Jaleel, M. Qureshi, BATMAN: techniques for maximizing system bandwidth of memory systems with stacked-DRAM, in: Proceedings of the ACM International Symposium on Memory Systems, 2017, pp. 268–280.
[21] M. Agyeman, A. Ahmadinia, N. Bagherzadeh, Energy and performance-aware application mapping for inhomogeneous 3D networks-on-chip, J. Syst. Archit. 89 (2018) 103–117.
[22] D. Jeon, N. Ickes, P. Raina, H. Wang, D. Rus, A. Chandrakasan, 24.1 a 0.6 v 8mw 3D vision processor for a navigation device for the visually impaired, in: Proceedings of the IEEE International Solid-State Circuits Conference, 2016, pp. 416–417.
[23] X. Chen, N. Jha, A 3-D CPU-FPGA-DRAM hybrid architecture for low-power computation, IEEE Trans. Very Large Scale Integr. VLSI Syst. 24 (5) (2016) 1649–1662.
[24] J. Zhao, Q. Zou, Y. Xie, Overview of 3-D architecture design opportunities and techniques, IEEE Des. Test 34 (4) (2017) 60–68.
[25] S. Ohira, T. Matsumura, Design for three-dimensional sound processor using high-level synthesis, in: Proceedings of the IEEE International Symposium on Diagnostics of Electronic Circuits & Systems, 2017, pp. 190–193.
[26] J. Lau, Heterogeneous integration by FOWLP, Fan-Out Wafer-Level Packag. (2018) 269–303.
[27] Tezzaron Semiconductor Corporation, 3D-ICs and integrated circuit security, http://www.tezzaron.com/about/papers/3D-ICs_and_Integrated_Circuit_Security.pdf, (2018).
[28] J. Valamehr, T. Huffmire, C. Irvine, R. Dick, R. Kastner, C. Koc, T. Levin, T. Sherwood, A Qualitative Security Analysis of a New Class of 3-D Integrated Crypto Co-processors, in: Cryptography and Security: From Theory to Applications, Springer Berlin Heidelberg, 2012, pp. 364–382.
[29] P. Gu, S. Li, D. Stow, L. Liu, Y. Xie, E. Kursun, Leveraging 3D technologies for hardware security: opportunities and challenges, in: Proceedings of the IEEE International Great Lakes Symposium on VLSI, 2016, pp. 347–352.
[30] P. Gu, D. Stow, R. Barnes, E. Barnes, Y. Xie, Thermal-aware 3D design for side-channel information leakage, in: Proceedings of the IEEE International Conference on Computer Design, 2016, pp. 520–527.
[31] Y. Xie, C. Bao, C. Serafy, T. Lu, A. Srivastava, M. Tehranipoor, Security and vulnerability implications of 3D ICs, IEEE Trans. Multi-Scale Comput. Syst. 2 (2) (2016) 108–122.
[32] C. Yan, J. Dofe, S. Kontak, Q. Yu, E. Salman, Hardware-efficient logic camouflaging for monolithic 3-D ICs, IEEE Trans. Circuits Syst. II Express Briefs 65 (5) (2018) 799–803.
[33] W. Hung, G. Link, Y. Xie, N. Vijaykrishnan, M. Irwin, Interconnect and thermal-aware floorplanning for 3D microprocessors, in: Proceedings of the IEEE International Symposium on Quality Electronic Design, 2006, pp. 98–104.
[34] H. Yan, Q. Zhou, X. Hong, Thermal aware placement in 3D ICs using quadratic uniformity modeling approach, Integr. VLSI J. 42 (2) (2009) 175–180.
[35] L. Xiao, S. Sinha, J. Xu, E. Young, Fixed-outline thermal-aware 3D floorplanning, in: Proceedings of the IEEE Asia and South Pacific Design Automation Conference, 11, 2010, pp. 561–567.
[36] D. Cuesta, J. Risco-Martin, J. Ayala, J. Hidalgo, 3D thermal-aware floorplanner using a MOEA approximation, Integr. VLSI J. 46 (1) (2013) 10–21.
[37] C. Wang, J. Zhou, R. Weerasekera, B. Zhao, X. Liu, P. Royannez, M. Je, BIST Methodology, architecture and circuits for pre-bond TSV testing in 3D stacking IC systems, IEEE Trans. Circuits Syst. I Regul. Pap. 62 (1) (2015) 139–148.
[38] L. Yavits, A. Morad, R. Ginosar, The effect of temperature on amdahl law in 3D multicore era, IEEE Trans. Comput. (2016) 1–5.
[39] S. Kandlikar, A. Ganguly, Fundamentals of Heat Dissipation in 3D IC Packaging, in: 3D Microelectronic Packaging, Springer, 2017, pp. 245–260.
[40] B. Joardar, W. Choi, R. Kim, J. Doppa, P. Pande, D. Marculescu, R. Marculescu, 3D NoC-enabled heterogeneous manycore architectures for accelerating CNN training: Performance and thermal trade-offs, in: Proceedings of the IEEE/ACM International Symposium on Networks-on-Chip, 2017, pp. 1–8.
[41] S. Chatterjee, S. Roy, C. Giri, H. Rahaman, Modeling and analysis of transient heat for 3D IC, in: Proceedings of the Springer International Symposium on VLSI Design and Test, 2017, pp. 365–375.
[42] S. Kumar, A. Zjajo, R. Leuken, Fighting dark silicon: toward realizing efficient thermal-aware 3-D stacked multiprocessors, IEEE Trans. Very Large Scale Integr. VLSI Syst. 25 (4) (2017) 1549–1562.
[43] D. Lee, S. Das, P. Pande, Analyzing power-thermal-performance trade-offs in a high-performance 3D NoC architecture, Integration (2018) 1.
[44] H. Zhu, F. Hu, H. Zhou, D. Pan, D. Zhou, X. Zeng, Interlayer cooling network design for high-performance 3D ICs using channel patterning and pruning, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 37 (4) (2018) 770–781.

[45] C. Liao, C. Wen, SVM-Based dynamic voltage prediction for online thermally constrained task scheduling in 3-D multicore processors, IEEE Embed. Syst. Lett. 10 (2) (2018) 49–52.
[46] Y. Cheng, L. Zhang, Y. Han, X. Li, Thermal-constrained task allocation for interconnect energy reduction in 3-D homogeneous MPSocs, IEEE Trans. Very Large Scale Integr. VLSI Syst. 21 (2) (2013) 239–249.
[47] K. Cao, J. Zhou, P. Cong, L. Li, T. Wei, M. Chen, S. Hu, X. Hu, Affinity-driven modeling and scheduling for makespan optimization in heterogeneous multiprocessor systems, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. (2018).
[48] M. Shafique, S. Garg, J. Henkel, D. Marculescu, The eda challenges in the dark silicon era: temperature, reliability, and variability perspectives, Proc. ACM Des. Autom. Conf. (2014) 1–6.
[49] M. Scrbak, M. Islam, K. Kavi, M. Ignatowski, N. Jayasena, Exploring the processing-in-Memory design space, J. Syst. Archit. 75 (2017) 59–67.
[50] A. Andreev, A. Sridhar, M. Sabry, E. Barnes, M. Zapater, P. Ruch, B. Michel, D. Atienza, Powercool: simulation of cooling and powering of 3D MPSocs with integrated flow cell arrays, IEEE Trans. Comput. 67 (1) (2018) 73–85.
[51] T. Lu, C. Serafy, Z. Yan, A. Srivastava, Voltage noise induced DRAM soft error reduction technique for 3D-CPUs, in: Proceedings of the ACM International Symposium on Low Power Electronics and Design, 2016, pp. 82–87.
[52] X. Zhou, Y. Xu, Y. Du, Y. Zhang, J. Yang, Thermal management for 3D processors via task scheduling, in: Proceedings of the IEEE International Conference on Parallel Processing, 2008, pp. 115–122.
[53] Y. Zhan, S. Kumar, S. Sapatnekar, Thermally aware design, Found. Trends Electron. Des. Autom. 2 (3) (2008) 255–370.
[54] G. Sun, Y. Chen, X. Dong, J. Ouyang, Y. Xie, Three-dimensional integrated circuits: design, EDA, and architecture, Found. Trends Electron. Des. Autom. 5 (3) (2011) 1–151.
[55] Y. Zhang, L. Li, Z. Lu, A. Jantsch, M. Gao, H. Pan, F. Han, A survey of memory architecture for 3D chip multi-processors, Microprocess. Microsyst. 38 (5) (2014) 415–430.
[56] J. Kong, S. Chung, K. Skadron, Recent thermal management techniques for microprocessors, ACM Comput. Surv. 44 (3) (2012) 1–42.
[57] M. Sabry, D. Atienza, Temperature-aware design and management for 3D multi–core architectures, Found. Trends Electron. Des. Autom. 8 (2) (2014) 117–197.
[58] W. Wang, Y. Han, H. Li, L. Zhang, Y. Cheng, X. Li, A reliable 3D MLC PCM architecture with resistance drift predictor, in: Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks, 2014, pp. 204–215.
[59] D. Lee, S. Ghose, G. Pekhimenko, S. Khan, O. Mutlu, Simultaneous multi-layer access: improving 3D-stacked memory bandwidth at low cost, ACM Trans. Archit. Code Optim. 12 (4) (2016).
[60] Y. Liu, S. Peng, H. Hwang, Wide-I/O 3D-staked DRAM controller for near-data processing system, in: Proceedings of the IEEE International Symposium on VLSI Design, Automation and Test, 2017, pp. 1–4.
[61] P. Liu, A. Hemani, K. Paul, C. Weis, M. Jung, N. Wehn, 3D-stacked many-core architecture for biological sequence analysis problems, Int. J. Parallel Program. 45 (6) (2017) 1420–1460.
[62] Y. Tang, Y. Wang, H. Li, X. Li, ApproxPIM: exploiting realistic 3D-stacked DRAM for energy-efficient processing in-memory, in: Proceedings of the IEEE Asia and South Pacific Design Automation Conference, 2017, pp. 396–401.
[63] E. Azarkhish, D. Rossi, I. Loi, A modular shared L2 memory design for 3-D integration, IEEE Trans. Very Large Scale Integr. VLSI Syst. 23 (8) (2015) 1485–1498.
[64] K. Kang, S. Park, J. Lee, L. Benini, G. Micheli, A power-efficient 3-D on-chip interconnect for multi-core accelerators with stacked L2 cache, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2016, pp. 1465–1468.
[65] J. Kong, Y. Gong, S. Chung, Architecting large-scale SRAM arrays with monolithic 3D integration, in: Proceedings of the IEEE International Symposium on Low Power Electronics and Design, 2017, pp. 1–6.
[66] A. Nasri, M. Fathy, A. Broumandnia, An energy-efficient 3D-stacked STT-RAM cache architecture for cloud processors: the effect on emerging scale-out workloads, J. Supercomput. (2017) 1–15.
[67] T. Zhang, J. Meng, A. Coskun, Dynamic cache pooling in 3D multicore processors, ACM J. Emerg. Technol. Comput. Syst. 12 (2) (2015) 1–20.
[68] B. Joardar, K. Duraisamy, P. Pande, High performance collective communication-aware 3D network-on-chip architectures, in: Proceedings of the IEEE Design, Automation & Test in Europe Conference & Exhibition, 2017, pp. 396–401.
[69] S. Das, J. Doppa, P. Pande, K. Chakrabarty, Monolithic 3D-enabled high performance and energy efficient network-on-chip, in: Proceedings of the IEEE International Conference on Computer Design, 2017, pp. 233–240.
[70] X. Wang, Y. Jang, M. Yang, H. Li, T. Mak, HRC: A 3D NoC architecture with genuine support for runtime thermal-aware task management, IEEE Trans. Comput. 66 (10) (2017) 1676–1688.
[71] J. Sepulveda, G. Gogniat, D. Florez, J. Diguet, C. Pedraza, M. Strum, 3D-LeukoNoc: a dynamic NoC protection, in: Proceedings of the IEEE International Conference on ReConFigurable Computing and FPGAs, 2014, pp. 1–6.
[72] J. Sepulveda, G. Gogniat, D. Florez, J. Diguet, R. Pires, M. Strum, TSV Protection: towards secure 3D-MPSoc, in: Proceedings of the Latin American Symposium on Circuits and Systems, 2015, pp. 1–6.
[73] G. Loi, B. Agrawal, N. Strivastava, S. Lin, T. Sherwood, A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy, Proc. ACM Des. Autom. Conf. (2006) 991–996.
[74] S. Chatterjee, M. Cho, R. Rao, S. Mukhopadhyay, Impact of die-to-die thermal coupling on the electrical characteristics of 3D stacked SRAM cache, in: Proceedings of the IEEE International Symposium on Semiconductor Thermal Measurement and Management, 2012, pp. 14–19.
[75] F. Tavakkoli, S. Ebrahimi, S. Wang, K. Vafai, Analysis of critical thermal issues in 3D integrated circuits, Int. J. Heat Mass Transf. 97 (2016) 337–352.
[76] J. Knechtel, J. Lienig, A. Elfadel, Multi-objective 3D floorplanning with integrated voltage assignment, ACM Trans. Des. Autom. Electron. Syst. 23 (2) (2017) 1–25.
[77] J. Song, Y. Zhang, Thermal via planning for 3-D ICs, in: Proceedings of the International Conference on Computer-Aided Design, 2005, pp. 745–752.
[78] X. Li, Y. Ma, X. Hong, S. Dong, J. Cong, LP Based white space redistribution for thermal via planning and performance optimization in 3D ICs, in: Proceedings of the IEEE Asia and South Pacific Design Automation Conference, 2008, pp. 209–212.
[79] C. Wen, Y. Chen, S. Ruan, Cluster-based thermal-aware 3D-floorplanning technique with post-floorplan TTSV insertion at via-channels, in: Proceedings of the IEEE Asia Symposium on Quality Electronic Design, 2013, pp. 200–207.
[80] P. Budhathoki, J. Knechtel, A. Henschel, I. Elfadel, Thermal-driven 3d floorplanning using localized TSV placement, 3D Stacked Chips (2016) 195–209.
[81] P. Zhou, Y. Ma, Z. Li, R. Dick, L. Shang, H. Zhou, X. Hong, Q. Zhou, 3D-STAF: Scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits, in: Proceedings of the IEEE International Conference on Computer-Aided Design, 2007, pp. 590–597.
[82] Y. Huang, Q. Zhou, Y. Cai, H. Yan, A thermal-driven force-directed floorplanning algorithm for 3D ICs, in: Proceedings of the IEEE Conference on Computer-Aided Design and Computer Graphics, 2009, pp. 497–502.
[83] D. Kim, K. Athikulwongse, S. Lim, Study of through-silicon-via impact on the 3-D stacked IC layout, IEEE Trans. Very Large Scale Integr. VLSI Syst. 21 (5) (2013) 862–874.
[84] K. Athikulwongse, M. Ekpanyapong, S. Lim, Exploiting die-to-die thermal coupling in 3-D IC placement, IEEE Trans. Very Large Scale Integr. VLSI Syst. 22 (10) (2014) 2145–2155.
[85] D. Cuesta, J. Risco-Martin, J. Ayala, D. Atienza, 3D thermal-aware floorplanner for many-core single-chip systems, in: Proceedings of the IEEE Latin American Test Workshop, 2011, pp. 1–6.
[86] D. Cuesta, J. Risco-MartÅ!'Âªn, J. Ayala, J. Risco-Martin, J. Hidalgo, Thermal-aware floorplanner for 3D IC, including TSVs, liquid microchannels and thermal domains optimization, Appl. Soft Comput. 34 (2015) 164–177.
[87] R. Dash, J. Risco-Martin, A. Turuk, J. Ayala, V. Pangracious, A. Majumdar, A bio-inspired hybrid thermal management approach for three-dimensional network-on-chip systems, IEEE Trans. Nanobioscience 16 (8) (2017) 727–743.
[88] D. Saha, S. Sur-Kolay, Multi-objective optimization of placement and assignment of TSVs in 3D ICs, in: Proceedings of the IEEE International Conference on VLSI Design, 2017, pp. 372–377.
[89] Y. Chen, S. Ruan, A cluster-based reliability-and thermal-aware 3D floorplanning using redundant STSVs, in: Proceedings of the IFIP/IEEE International Conference on Very Large Scale Integration, 2015, pp. 349–354.
[90] J. Lin, P. Chiu, Y. Chang, SAINT: handling module folding and alignment in fixed-outline floorplans for 3D ICs, in: Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, 2016, pp. 1–7.
[91] A. Tabrizi, L. Behjat, W. Swartz, L. Rakai, A fast force-directed simulated annealing for 3D IC partitioning, Integr. VLSI J. 55 (2016) 202–211.
[92] Z. Li, Y. Ma, Q. Zhou, Y. Cai, Y. Wang, T. Huang, Y. Xie, Thermal-aware power network design for IR drop reduction in 3D ICs, in: Proceedings of the IEEE Asia and South Pacific Design Automation Conference, 2012, pp. 47–52.
[93] Q. Xu, S. Chen, Fast thermal analysis for fixed-outline 3D floorplanning, Integr. VLSI J. 59 (2017) 157–167.
[94] J. Knechtel, Q. Sinanoglu, On mitigation of side-channel attacks in 3D ICs: decorrelating thermal patterns from power and activity, Proc. ACM Des. Autom. Conf. (2017) 1–6.
[95] A. Hsieh, T. Hwang, Thermal-aware memory mapping in 3D designs, ACM Trans. Embedded Comput. Syst. 13 (4) (2013) 1–22.
[96] M. Beigi, G. Memik, TAPAS: temperature-aware adaptive placement for 3D stacked hybrid caches, in: Proceedings of the ACM International Conference on Memory Systems, 2016, pp. 415–426.
[97] J. Meng, A. Coskun, Analysis and runtime management of 3D systems with stacked DRAM for boosting energy efficiency, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2012, pp. 611–616.
[98] M. Guan, L. Wang, Temperature aware refresh for DRAM performance improvement in 3D ICs, in: Proceedings of the IEEE International Symposium on Quality Electronic Design, 2015, pp. 207–211.
[99] M. Guan, L. Wang, Improving DRAM performance in 3-D ICs via temperature aware refresh, IEEE Trans. Very Large Scale Integr. VLSI Syst. 25 (3) (2017) 833–843.
[100] H. Xiao, W. Yueh, S. Yalamanchili, Thermally adaptive cache access mechanisms for 3D many-core architectures, IEEE Comput. Archit. Lett. 15 (2) (2016) 129–132.
[101] X. Jiang, X. Lei, L. Zeng, T. Watanabe, High performance virtual channel based fully adaptive thermal-aware routing for 3D NoC, in: Proceedings of the IEEE International Symposium on Quality Electronic Design, 2017, pp. 289–295.
[102] M. Sadri, M. Jung, C. Weis, N. Wehn, L. Benini, Energy optimization in 3D MPSocs with wide-I/o DRAM using temperature variation aware bank-wise refresh, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2014, pp. 1–4.
[103] H. Shin, Y. Park, D. Choi, B. Kim, D. Cho, E. Chung, EXTREME: Exploiting page table for reducing refresh power of 3D-stacked DRAM memory, IEEE Trans. Comput. 67 (1) (2018) 32–44.
[104] J. Park, J. Jung, K. Yi, C. Kyung, Static energy minimization of 3D stacked L2 cache with selective cache compression, in: Proceedings of the IFIP/IEEE International Conference on Very Large Scale Integration, 2013, pp. 228–233.
[105] K. Yao, Y. Ye, S. Pasricha, J. Xu, Thermal-sensitive design and power optimization for a 3D torus-based optical NoC, in: Proceedings of the IEEE International Conference on Computer-Aided Design, 2017, pp. 827–834.

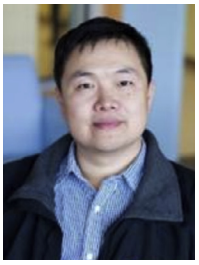
[106] S. Lin, J. Lin, Thermal-and performance-aware address mapping for the multi-channel three-dimensional DRAM systems, IEEE Access 5 (2017) 5566–5577.
[107] M. Beigi, G. Memik, TESLA: using microfluidics to thermally stabilize 3D stacked STT-RAM caches, in: Proceedings of the IEEE International Conference on Computer Design, 2016, pp. 344–347.
[108] D. Li, K. Zhang, A. Guliani, S. Ogrenci-Memik, Adaptive thermal management for 3D ICs with stacked DRAM caches, Proc. ACM Des. Autom. Conf. (2017) 101–112.
[109] A. Asad, O. Ozturk, M. Fathy, M. Jahed-Motlagh, Optimization-based power and thermal management for dark silicon aware 3D chip multiprocessors using heterogeneous cache hierarchy, Microprocess. Microsyst. 51 (2017) 76–98.
[110] Y. Fu, L. Li, K. Wang, C. Zhang, Kalman predictor-based proactive dynamic thermal management for 3-D NoC systems with noisy thermal sensors, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 36 (11) (2017) 1869–1882.
[111] C. Chou, Y. Lin, K. Chiang, K. Chen, Dynamic buffer allocation for thermal-aware 3D network-on-chip systems, in: Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan, 2017, pp. 65–66.
[112] H. Tajik, H. Homayoun, N. Dutt, VAWOM: Temperature and process variation aware wearout management in 3D multicore architecture, Proc. ACM Des. Autom. Conf. (2013) 1–8.
[113] H. Wang, K. Zhao, M. Lv, X. Zhang, H. Sun, T. Zhang, Improving 3D DRAM fault tolerance through weak cell aware error correction, IEEE Trans. Comput. 66 (5) (2017) 820–833.
[114] H. Han, J. Chung, J. Yang, READ: Reliability enhancement in 3D-Memory exploiting asymmetric ser distribution, IEEE Trans. Comput. (2018) 1.
[115] J. Lim, H. Lim, S. Kang, 3-D Stacked DRAM refresh management with guaranteed data reliability, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 34 (9) (2015) 1455–1466.
[116] T. Tsai, Y. Chen, Thermal-aware real-time task scheduling for three-dimensional multicore chip, in: Proceedings of ACM International Symposium on Applied Computing, 2012, pp. 1618–1624.
[117] J. Li, M. Qiu, J. Niu, L. Yang, Y. Zhu, Z. Ming, Thermal-aware task scheduling in 3D chip multiprocessor with real-time constrained workloads, ACM Trans. Embedded Comput. Syst. 12 (2) (2013) 1–22.
[118] W. Cheng, T. Hsu, Thermal-aware task allocation, memory mapping, and task scheduling for 3D stacked memory and processor architecture, in: Proceedings of the IEEE TENCON Spring Conference, 2013, pp. 95–98.
[119] V. Chaturvedi, A. Singh, W. Zhang, T. Srikanthan, Thermal-aware task scheduling for peak temperature minimization under periodic constraint for 3D-MPSoCs, in: Proceedings of the IEEE International Symposium on Rapid System Prototyping, 2014, pp. 107–113.
[120] D. Zhao, H. Homayoun, A. Veidenbaum, Temperature aware thread migration in 3D architecture with stacked DRAM, in: Proceedings of the IEEE International Symposium on Quality Electronic Design, 2013, pp. 80–87.
[121] C. Lung, Y. Ho, D. Kwai, S. Chang, Thermal-aware on-line task allocation for 3D multi-core processor throughput optimization, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2011, pp. 1–6.
[122] C. Yu, C. Lung, Y. Ho, R. Hsu, D. Kwai, S. Chang, Thermal-aware on-line scheduler for 3-D many-core processor throughput optimization, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 33 (5) (2014) 763–773.
[123] G. Chaparro-Baquero, S. Sha, S. Homsi, W. Wen, G. Quan, Thermal-aware joint CPU and memory scheduling for hard real-time tasks on multicore 3D platforms, in: Proceedings of the IEEE International Green and Sustainable Computing Conference, 2017, pp. 1–8.
[124] T. Tsai, Y. Chen, X. He, C. Li, STEM: a thermal-constrained real-time scheduling for 3D heterogeneous-ISA multicore processors, IEEE Trans. Comput. 67 (6) (2018) 874–889.
[125] T. Tsai, Y. Chen, Thermal-throttling server: a thermal-aware real-time task scheduling framework for three-dimensional multicore chips, J. Syst. Softw. (2016) 11–25.
[126] B. Li, X. Wang, A. Singh, T. Mak, On runtime communication-and thermal-aware application mapping in 3D NoC, in: Proceedings of the IEEE/ACM International Symposium on Networks-on-Chip, 2017, pp. 1–8.
[127] S. Jin, Y. Wang, T. Liu, On optimizing system energy of voltagecfrequency island based 3-D multi-core socs under thermal constraints, Integr. VLSI J. 48 (2015) 36–45.
[128] M. Cox, A. Singh, A. Kumar, H. Corporaal, Thermal-aware mapping of streaming applications on 3D multi-processor systems, in: Proceedings of the IEEE International Conference on Embedded Systems for Real-time Multimedia, 2013, pp. 11–20.
[129] C. Liao, C. Wen, K. Chakrabarty, An online thermal-constrained task scheduler for 3D multi-core processors, in: Proceedings of the IEEE Design, Automation and Test in Europe Conference and Exhibition, 2015, pp. 351–356.
[130] Y. Cui, W. Zhang, V. Chaturvedi, W. Liu, B. He, Thermal-aware task scheduling for 3D-network-on-chip: a bottom-to-top scheme, J. Circuits Syst. Comput. 25 (1) (2016) 1–20.
[131] Z. Zhu, V. Chaturvedi, A. Singh, W. Zhang, Y. Cui, Two-stage thermal-aware scheduling of task graphs on 3D multi-cores exploiting application and architecture characteristics, in: Proceedings of the IEEE Asia and South Pacific Design Automation Conference, 2017, pp. 324–329.
[132] A. Singh, M. Shafique, A. Kumar, J. Henkel, Analysis and mapping for thermal and energy efficiency of 3-D video processing on 3-D multicore processors, IEEE Trans. Very Large Scale Integr. VLSI Syst. 24 (8) (2016) 2745–2758.
[133] L. Shen, N. Wu, G. Yan, F. Ge, Thermal-aware task mapping for communication energy minimization on 3D NoC, IEICE Electron. Express 14 (22) (2017) 1–9.
[134] C. Yoon, J. Shimand, B. Moon, J. Kong, 3D Die-stacked DRAM thermal management via task allocation and core pipeline control, IEICE Electron. Express 15 (3) (2018) 1–12.
[135] V. Raparti, N. Kapadia, S. Pasricha, ARTEMIS: An aging-aware runtime application mapping framework for 3D NoC-based chip multiprocessors, IEEE Trans. Multi-Scale Comput. Syst. 3 (2) (2017) 72–85.

Kun Cao is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, East China Normal University, Shanghai, China. His current research interests are in the areas of high performance computing, heterogeneous multiprocessor systems, and cyber physical systems. He received the Reviewer Award from Journal of Circuits, Systems, and Computers, in 2016.

Junlong Zhou received the Ph.D. degree in Computer Science from East China Normal University, Shanghai, China, in 2017. He was a Visiting Scholar with the University of Notre Dame, Notre Dame, IN, USA, during 2014–2015. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include real-time embedded systems, cloud computing, and cyber physical systems. Dr. Zhou has been an Associate Editor for the Journal of Circuits, Systems, and Computers since 2017.

Tongquan Wei received his Ph.D. degree in Electrical Engineering from Michigan Technological University in 2009. He is currently an Associate Professor in the Department of Computer Science and Technology at the East China Normal University. His current research interests include Internet of Things, edge computing, cloud computing, and design automation of intelligent and CPS systems. He serves as a Regional Editor for Journal of Circuits, Systems, and Computers since 2012.

Mingsong Chen received the B.S. and M.E. degrees from Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2003 and 2006 respectively, and the Ph.D. degree in Computer Engineering from the University of Florida, Gainesville, in 2010. He is currently a full Professor with the Department of Embedded Software and Systems of East China Normal University. His research interests are in the area of design automation of cyber-physical systems, formal verification techniques and mobile cloud computing.



Shiyan Hu received his Ph.D. in Computer Engineering from Texas A&M University in 2008. He is the Chair and Professor of Cyber-Physical Systems at University of Essex, UK. He was an Associate Professor and Director of Center for Cyber-Physical Systems at Michigan Tech., USA. He was also a Visiting Professor at IBM Research (Austin) in 2010, and a Visiting Associate Professor at Stanford University from 2015 to 2016. His research interests include Cyber-Physical Systems (CPS), CPS Security, Smart Energy CPS, Data Analytics and Computer-Aided Design of VLSI Circuits, where he has published more than 100 refereed papers.

He is an ACM Distinguished Speaker, an IEEE Systems Council Distinguished Lecturer, an IEEE Computer Society Distinguished Visitor, a recipient of the 2017 IEEE Computer Society TCSC Middle Career Researcher Award, the 2014 National Science Foundation (NSF) CAREER Award, and the 2009 ACM SIGDA Richard Newton DAC Scholarship. His publications have received a few distinctions, which includes the 2018 IEEE Systems Journal Best Paper Award, the 2017 Keynote Paper in IEEE Transactions on Computer-Aided Design, and the Front Cover in IEEE Transactions on Nanobioscience in March 2014. He is the Chair for IEEE Technical Committee on Cyber-Physical Systems. He is the Editor-In-Chief of IET Cyber-Physical Systems: Theory&Applications. He serves as an Associate Editor for IEEE Transactions on Computer-Aided Design, IEEE Transactions on Industrial Informatics, IEEE Transactions on Circuits and Systems, ACM Transactions on Design Automation for Electronic Systems, and ACM Transactions on Cyber-Physical Systems. He has served as a Guest Editor for 8 IEEE/ACM Journals such as Proceedings of the IEEE and IEEE Transactions on Computers. He has held chair positions in numerous IEEE/ACM conferences. He is a Fellow of IET.

Keqin Li is a SUNY Distinguished Professor of computer science in the State University of New York. He is also a Distinguished Professor of Chinese National Recruitment Program of Global Experts (1000 Plan) at Hunan University, China. He was an Intellectual Ventures endowed visiting chair professor at the National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China, during 2011–2014. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, CPU-GPU hybrid and cooperative computing, multicore computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file sharing systems, mobile computing, service computing, Internet of things and cyber-physical systems. He has published over 590 journal articles, book chapters, and refereed conference papers, and has received several best paper awards. He is currently serving or has served on the editorial boards of IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Computers, IEEE Transactions on Cloud Computing, IEEE Transactions on Services Computing, and IEEE Transactions on Sustainable Computing. He is an IEEE Fellow.