A Survey of Memory Architecture for 3D Chip Multi-Processors

Yuang Zhang[1, 2], Li Li[1], Zhonghai Lu[2], Axel Jantsch[2], Minglun Gao[1], Hongbing Pan[1], Feng Han[1]

[1]Institute of VLSI Design, Nanjing University
Key Laboratory of Advanced Photonic and Electronic Materials,
Nanjing University
Nanjing, China
Email: {zhangyuang, lili, gaominglun, phb} @ nju.edu.cn

[2]Department of Electronic, Computer and Software Systems
School of Information and Communication Technology,
KTH - Royal Institute of Technology
Stockholm, Sweden
{yazhang, zhonghai, axel} @ kth.se

Abstract: 3D chip multi-processors (3D CMPs) combine the advantages of 3D integration and the parallelism of CMPs, which are emerging as active research topics in VLSI and multi-core computer architecture communities. One significant potentiality of 3D CMPs is to exploit the diversity of integration processes and high volume of vertical TSV bandwidth to mitigate the well-known “Memory Wall” problem. Meanwhile, the 3D integration techniques are under the severe thermal, manufacture yield and cost constraints. Research on 3D stacking memory hierarchy explores the high performance and power/thermal efficient memory architectures for 3D CMPs. The micro architectures of memories can be designed in the 3D integrated circuit context and integrated into 3D CMPs. This paper surveys the design of memory architectures for 3D CMPs. We summarize current research into two categories: stacking cache-only architectures and stacking main memory architectures for 3D CMPs. The representative works are reviewed and the remaining opportunities and challenges are discussed to guide the future research in this emerging area.

Keywords: 3D integrated circuit, chip multi-processor, memory architecture, non-uniform cache architecture.

1 Introduction

Recently, the performance of microprocessors has been enhanced by increasing the number of cores on a chip instead of promoting the frequency of a single core. The single chip multi-processor (CMP) [1, 2] is a trend for future microprocessor designs. Both research prototypes and industry products of CMPs have been presented. Intel has demonstrated a network based multiprocessor system with 80 cores [3]. Tilera [4] has presented a 64-core and later a 100-core multiprocessor which aim for the market of cloud computing, high performance multimedia and communication. Fujitsu has shown their new generation SPARC64 processor with 16 cores for mission-critical UNIX server in an enhanced 28 nm high-κ metal-gate (HKMG) CMOS process [5].

The increasing speed of the processor is faster than that of the memory. The processor may be starved of data which is known as the “Memory Wall” problem [6]. As the number of cores increases in CMPs, more and more data are required from the memory subsystem, which places a great deal of pressure on the I/O infrastructure. The memory bandwidth evolves as a performance bottleneck. According to the International Technology Roadmap for Semiconductors (ITRS) projection, the number of pins on a package cannot continue to grow rapidly enough to supply bandwidth for off-chip memory in the next decade. Feeding enough data to a massive number of on-die cores keeps developers facing the "Memory Wall" challenges for CMPs.

By stacking memories on top of logic layer, three-dimensional integrated circuit (3D IC) [7-10] is envisioned as a solution for future CMP design. The density of connections can be increased by using fine-pitch through-silicon vias (TSVs). TSV technologies [11, 12] promise increased system integration at low cost and reduced footprint for 3D die-stacking. Using 3D technology, we can increase the number of I/Os between both dies from 32 to thousands of bits. This massively parallel interconnection is crucial to improve memory bandwidth for CMPs. Hence, 3D integration technology is a way to mitigate the “Memory Wall” challenge in future microprocessors. Moreover, to stack memories on top of the core layers does not have the complexity of processor re-designing. Figure 1 shows a 3D multi-core network-on-chip (McNoC) in which the CMPs are connected by on chip network infrastructures [13, 14]. This 3D McNoC consists of 1 processor layer and 2 L2 cache layers.

Recently, 3D CMPs have gained great interests of both the academic community and semiconductor industry. Lots of research papers have appeared in the latest 4-5 years, and several 3D CMP prototypes have been presented. Kim et al. [15] demonstrate a 3D multi-core system with one 64-core layer and one 256 KB SRAM layer in 130nm technology. The processing foundries are Global Foundries device technology and Tezzaron TSV/bonding technology. Fick et al. [16, 17] propose a low power 64-core system that is called Centip3De. Centip3De has two stacked dies with a layer of 64 ARM M3 near-threshold cores and a cache layer. All the cores are organized to 16 4-core clusters, each connected to a 4-way 1 KB instruction cache and a 4-way 8 KB data cache. Centip3De is targeted to be a reconfigurable 7-layer 3D multi-core system which consists of 2 core layers, 2 cache layers and 3 DRAM layers [18]. Wordeman et al. [19] present a prototype of a 3D system with a memory layer and a logic layer connected by TSVs. In [20], a 2-tier design which consists of 16 processors connected by a 4×2 mesh network-on-chip (NoC) is proposed. This design is planned to be fabricated by Tezzaron technology with 130 nm low power standard library of Global Foundries.

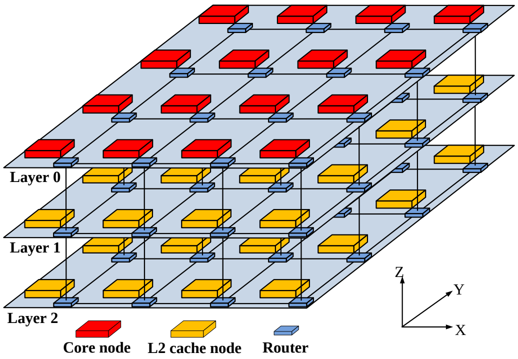

Fig.1. A 3D McNoC architecture with 1 logic layer and 2 L2 cache layers

The 3D IC technology provides opportunities for composing future systems by integrating disparate technologies as well as different technology generations in the same stack, e.g. analog circuit, CMOS, NMOS, non-volatile memories (NVMs), and cutting-edge technology nodes. To leverage this property of 3D integration, 3D CMPs can adopt disparate memory technologies like SRAM, DRAM and emerging NVM technologies, e.g. magnetic based random access memory (MRAM), phase-change memory (PCRAM), resistive RAM (RRAM), into a single tightly-coupled chip. Traditional DRAM can be integrated on top of the core layers. The density of DRAM is as high as 8 times that of SRAM. DRAM can act as the last level caches or main memories for 3D CMPs. Emerging NVMs are being explored as potential alternatives to existing SRAM/DRAM memories in future computing systems. Such emerging NVMs combine the speed of SRAM, the density of DRAM, and the non-volatility of the flash memory. Emerging NVM technologies have gained attention and are being actively investigated by the academia and industry.

The enlarged on-chip storage, integration of disparate manufacturing technologies and innovation in vertical connection give the computer designers a large space to explore. As 3D die-stacking technology is getting mature and hitting the market, it is essential to exploit the high performance memory architectures of 3D CMPs to gain the advantages of 3D technologies. Recent research has explored a variety of possibilities for die-stacking 3D CMP memory architectures, but many challenges and opportunities remain to fully realize the potential of the technology.

In this work, we intend to give an overview on memory architecture research for emerging 3D CMPs, with the concern of process availability. We look into the opportunities and challenges that the 3D CMPs bring up. The paper is organized as follows. Section 2 presents the background of 3D integration and the emerging NVM technologies. Section 3 introduces the memory architecture related terminology for 2D and 3D CMPs. Two categories of memory architectures, stacking cache-only architectures and stacking main memory architectures for 3D CMPs, are reviewed in section 4 and section 5 respectively. We discuss the design challenges and present outlooks in section 6. Finally we conclude this paper in section 7.

2 Background

2.1 2.5D and 3D IC

We briefly discuss the advantages and limitations of 2.5D and 3D integration technologies in this subsection, a full comparison among 2.5D and 3D technologies is beyond the scope of this article.

The 2.5D integration scheme [21, 22] is a revision of multi-chip module (MCM) [23, 24] with the new feature of 3D stacking ICs. 2.5D integration provides interconnections among multiple silicon dies by stacking each die on top of a common substrate (i.e. interposer). The interconnection substrate provides the necessary wiring for connecting the chips with each other and also with the pins of the whole 2.5D chip. Figure 2 shows a 2.5D IC using interposer. The interposer has TSVs connected to the metal layers on its upper and lower surfaces.

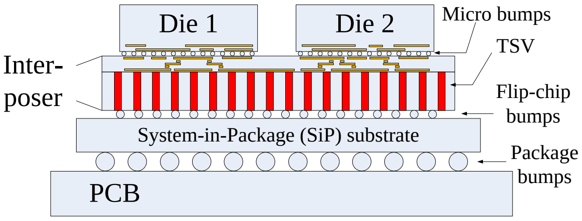

Fig.2. 2.5D integration using a silicon interposer and through-silicon vias (TSVs)

In [25], Velenis et al. compare the cost between 3D and 2.5D integration. The added interposer component of 2.5D incurs a significant cost over the necessary process cost for 3D stacking integration. In [26], Knickerbocker et al. present the actual technological achievements and key challenges of systems using 2.5D and 3D technology. 2.5D industrial products are available now. For example, Xilinx Vertex-7 2000T [27] integrates four FPGA devices to a silicon interposer, which enables more than 10,000 die-to-die connections. 2.5D integration is an evolutionary step from traditional 2D IC/System-in-Package (SiP) technology towards 3D ICs. 2.5D offers tremendous increase in capacity and performance. 2.5D also has yield advantages, because it's easier to make use of a number of small chips as opposed to a single large one. 2.5D integration may be a starting point for die integration [28].

A 3D IC [29-35] is an IC with more than one plane of devices and metals, connected by vertical interconnects. The 3D integration technology can take various approaches: (1) monolithic approach, which front-end processing is to build multiple active layers on a single wafer repeatedly and back-end processing interconnects the devices; (2) stacking approach, including wafer-to-wafer, die-to-wafer, or die-to-die stacking methods, which processes each layer separately using conventional fabrication techniques and builds up 3D IC by bonding technology [36-38]. 3D monolithic integration (3DMI) [39-41] is also known as sequential integration, in which the fine-grain integration of 3D circuits can be implemented. However, compared to the monolithic approach, the stacking approach does not require the change of conventional fabrication process. Hence, the stacking approach is much more practical, and becomes the focus of recent 3D integration research.

3D integration technology has become one of the promising approaches for future IC design which offers many more benefits than a 2D IC. Some of them are summarized as follows: (1) reduction in interconnect wire length, which results in improved wire latency and reduced power consumption; (2) improved memory bandwidth, by stacking memory on microprocessor cores with massive vertical connections between the memory layer and the core layer; (3) support for realization of heterogeneous integration, which can promote novel architecture designs; (4) smaller form factor, which results in higher packing density and smaller footprint. In [42], several methods associated with the fabrication of 3D ICs are discussed. The techniques developed by Tezzaron Semiconductor Corp. are described in detail. There exist several 3D IC products [43]. 3D memories receive a lot of attention. The 3D memories are potential future memory solutions as they feature low cost, high density and high performance. Samsung [44] presents a 3D double stacked 4 Gb multi-level cell (MLC) NAND flash memory device with shared bit line (BL) structure. The chip is composed of two stacked Si layers, each layer containing a 2 Gb MLC memory array in 45nm. The hybrid memory cube (HMC) [45] is a 3D DRAM proposed by Micron. The HMC is a heterogamous die with DRAM layers and a logic layer which contains the high-speed host interface, data distribution, address/control, refresh control and array repair functions.

2.2 TSVs

A through-silicon-via (TSV) connection is a galvanic connection between the two sides of a Si wafer that is electrically isolated from the substrate and from other TSV connections [46, 47]. A 3D CMP can have several stacked layers with different functions, e.g. processor layers, cache layers and controller layers. TSV is the enabling technology for 3D chips by stacking multiple dies together. These interconnections are formed through the dies to enable communication among dies. There are mainly three kinds of TSV processes, *via first*, *via last*, and *via middle*. When TSV is formed before the Si front-end of line (FEOL) device fabrication processing, the process is called as “via first”. In the “via middle” process, the fabrication of TSVs is after the Si FEOL device fabrication processing and before the back-end of line (BEOL) interconnect processing. In the “via last” process, TSV is performed after the BEOL processes are completed. Figure 3 shows these three kinds of TSVs.

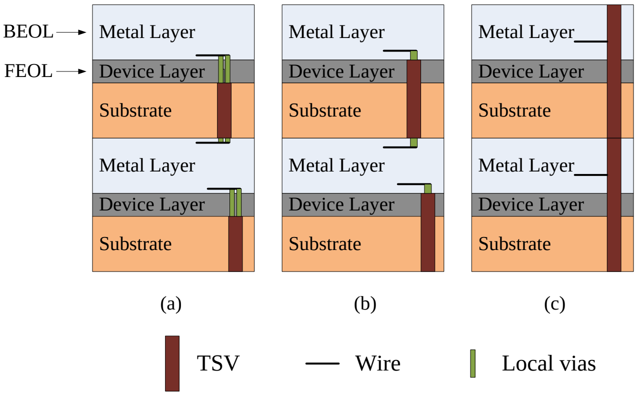

Fig.3. Schematic representation of TSV (a) via first, (b) via middle, and (c) via last process flows

TSV diameter may range from tens of microns down to nearly a micron. [48] presents that the latency for traversing a 20-layer vertical stack is 12 ps. The top view of TSVs is shown in figure 4.

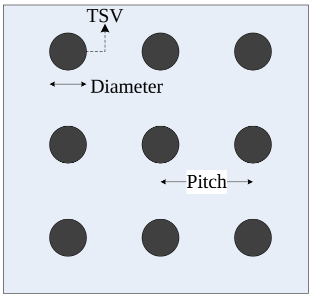

Fig.4. Top view of the inter-die vertical connections using TSV technology

According to ITRS 2011 [49], the 3D stacked IC/3D SoC TSV size parameters for the global interconnect level and the intermediate level are summarized in Table 1. Bernstein et al. [50] give an overview of the range of TSV technologies associated with applications.

Table 1. Global Interconnect Level and Intermediate Level of 3D TSV Roadmap

	Global level		Intermediate level	
	2011-2014	2015-2018	2011-2014	2014-2018
Diameter$_{\text{Min.}}$	4-8μm	2-4μm	1-2 μm	0.8-1.5 μm
Pitch$_{\text{Min.}}$	8-16 μm	4-8 μm	2-4 μm	1.6-3.0 μm
Depth$_{\text{Min.}}$	20-50 μm	20-50 μm	6-10 μm	6-10 μm
Contact pitch$_{\text{Min.}}$	20 μm	10 μm	2-3 μm	2-3 μm
No. of tiers	2-3	2-4	2-3	8-16 (DRAM)

2.3 Emerging NVM technology

Recent architecture-level studies have demonstrated that emerging non-volatile memories (NVMs), such as MRAM, PCRAM, and RRAM, have great potential to improve future memory architecture designs of 3D CMPs. These emerging memory technologies can be utilized to re-design the memory subsystems to gain the system improvement of power consumption with comparable performance to that of SRAM and DRAM memory systems [51].

2.3.1 MRAM

Magnetic-based RAM (MRAM), which is also known as STT-RAM (spin torque transfer RAM), has been considered as one of the most promising universal memory technologies due to its non-volatility, fast speed, zero standby power, and high density. The key element of MRAM cells is called magnetic tunnel junction (MTJ), which is used for binary storage. A MTJ device model is shown in figure 5.

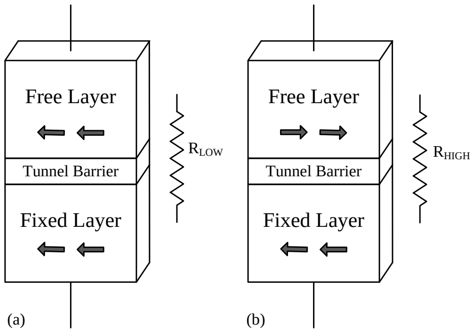

Fig.5. MTJ structure in its two states (a) parallel and (b) anti-parallel

MTJ contains a pinned layer and a free layer, which are separated by a thin insulator layer. The pinned layer has fixed magnetic direction, while the free layer can switch in between the parallel and the opposite directions. If the free layer has the same direction as the pinned layer, the electron will tunnel easily between the two layers. The MTJ resistance is low and indicates state "0". Otherwise, a bigger resistance appears and indicates state "1". For 90nm technology, the size of 1 MRAM cell is about 25% of 1 SRAM cell, and is about 1.7 times that of 1 DRAM cell [52].

Because of non-volatility, MRAM has no standby leakage power, with negligible active leakage power. For dynamic energy, compared with the same parameter of SRAM cell, an MRAM read operation consumes slightly less energy; the write operation consumes three orders of magnitude more energy per operation. Concerning speed, the read latency of MRAM is comparable with that of SRAM, and is much better than that of DRAM. However, the write latency of MRAM is much larger than that of SRAM and DRAM. The coming generation [53] of MTJ devices will reduce the energy consumption for per-switching to be values around 1 *fJ*, which is in the same order as the CMOS switching energy.

In [54], Dong et al. introduce and evaluate stacking MRAM on top of a microprocessor as a 3D IC. The simulation results show that though the performance benefit of MRAM is not obvious, the real benefit of MRAM is the power consumption. Using MRAM caches to implement L2 caches can save total power by 89% compared with SRAM caches and by 70% compared with DRAM caches. It is also attractive to build buffers in on-chip routers by scalable MTJ devices which can save the dynamic power [55].

2.3.2 PCRAM

Phase-change RAM (PCRAM) [56, 57] is a kind of non-volatile memory that exploits the property of chalcogenide glass, typically $Ge_2Sb_2Te_5$ (GST) material. PCRAM can be switched between two states, amorphous and crystalline. Figure 6 gives a basic structure of PCRAM cell.

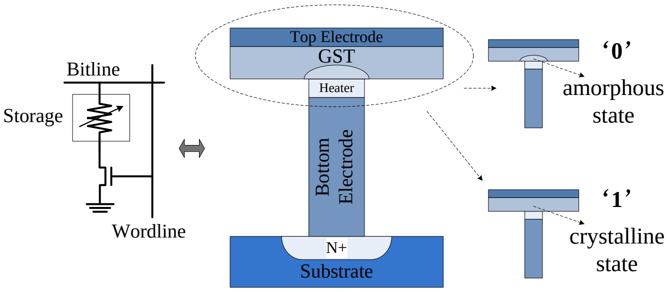

Fig.6. The basic structure of a PCRAM cell

The amorphous phase has low optical reflexivity and high electrical resistivity. The crystalline phase shows high reflectivity and low resistivity. The difference in resistance of the two states can be used to present logical states of binary data. During writes, an access transistor injects current into the storage material and thermally induces phase change, which is detected during reads. A phase-change material can be switched from one phase to another reliably, quickly, and a large number of times. For instance, 420 nm diameter $Ge_2Sb_2Te_5$ devices can be SET/RESET (recrystallized/reamorphized) in 90/1.5 ns, and 19 nm diameter devices can be switched in 2.5/0.4 ns [58]. PCRAM write endurance, which is the number of writes performed before the cell cannot be programmed reliably, is about 10^8 cycles [59] and is expected to be 10^{15} in 2015 [60]. PCRAM can act as a scalable DRAM alternative solution for increasing main memory density and capacity.

Lee et al. [61] propose a memory architecture that lays the foundation for exploiting PCRAM scalability and non-volatility as main memory. Energy of PCRAM scales down 1.6 times more quickly than that of DRAM. At 40 nm, PCRAM system energy is

61.3% that of DRAM on averaged. In [62], Qureshi et al. analyze a PCRAM-based hybrid main memory system using an architecture level model of PCRAM. They explore the trade-offs for a main memory system consisting of PCRAM storage coupled with a small DRAM buffer. Dong et al. [63] present a PCRAM model, called PCRAMsim, to bridge the gap between the device-level and system-level research on PCRAM technology. In [64], Joo et al. study a set of techniques to design an energy- and endurance-aware PCRAM cache. The timing, energy, endurance, and area of PCRAM caches are molded and integrated into a PCRAM cache simulator. Readers can refer to [65], which gives a very good overview on MRAM and PCRAM memories.

2.3.3 RRAM

Memristor-based resistive RAM (RRAM, also known as OxRAM, ReRAM) [66-69] has the property of fast access, zero standby leakage, non-volatility, multilevel capability, and relatively small write power consumption. Hence RRAM becomes one of the most promising next-generation universal memory candidates. Most of metal-oxide RRAM cell structures are composed of vertically stacked metal/metal-oxide/metal layers. The metal-oxide can be NiO [70], TiON [71], TiO_2 [72] etc. RRAM stores data based on resistance difference. The high resistance state (HRS) represents logic '0' and low resistance state (LRS) denotes '1'. Figure 7 shows a simplified schematic of the conventional structure for RRAM. A voltage larger than the set voltage is needed to trigger on the resistive switching behaviors for the subsequent cycles. The RRAM can be stacked in 3D as a crosspoint architecture. The memory cell area can be 4 F^2/n, where F is the minimum feature size and n is the number of 3D stacked memory layers [66].

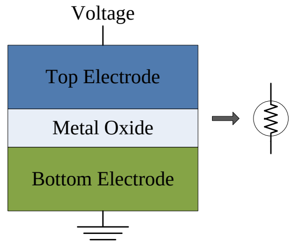

Fig.7. Structure schematic for metal-oxide RRAM

Lewis and Lee [73] give an overview of the memristors that have been produced by a number of different research teams and propose a point-by-point comparison between DRAM and RRAM, based on both existing and expected near-term memristor devices. The density of RRAM is about 10 times higher than that of DRAM. Though still suffering from a few shortcomings, RRAM shows itself to be a potential design alternative to well-established DRAM technology. Chang et al. [74] survey design challenges associated with low-voltage SRAM, memristor logic, and RRAM. The authors study integrating low-voltage RAM, RRAM and memristor logic in 3D ICs. The power consumption and thermal stress can be reduced. In [75], Chen et al. propose a RRAM based FPGA. Comparing with modern FPGA with SRAM, RRAM-based FPGA shows benefits of non-volatility, smaller area, and flexibility of configuration.

We summarize the comparison among SRAM, embedded DRAM (eDRAM), MRAM, PCRAM, and RRAM in table 2.

Table 2. Comparison of traditional and emerging NVM technologies

	SRAM	**eDRAM**	**MRAM**	**PCRAM**	**RRAM**
Density	Low	High	High	Very High	Very High
Dynamic Power	Low	Medium	Low for read, very high for write	Medium for read; high for write	Low for read, high for write
Leakage Power	High	Medium	Low	Low	Low
Speed	Very Fast	Fast	Fast read, slow write	Slow read, very slow write	Fast read, slow write
Non-Volatility	No	No	Yes	Yes	Yes
Scalability	Yes	Yes	Yes	Yes	Yes

Up to today there is no clear winner regarding which emerging NVM technology will likely be adopted in 3D CMPs, because all of them have different strength and weakness. According to current research in 3D CMPs, MRAM may act as stacking caches, and PCRAM as stacking caches and main memories due to its low static power. Thanks to its high density, RRAM is well positioned to be used as main memories in future 3D CMPs. One disadvantage of these emerging NVM technologies is that, because of the memory's non-volatility feature, it usually requires much longer latency and more energy for writes than for reads. Therefore, the 3D CMPs which intend to adopt these emerging memory technologies should take more consideration for mitigating the overheads of write operations.

3 Memory system for CMPs

In Section 3, a general memory hierarchy for CMPs is firstly revisited. Secondly, the potential memory architecture for 3D CMPs are presented. Non-uniform cache architecture (NUCA) is the main stream on-chip memory architecture for 2D CMPs, which can be migrated into 3D CMPs. Uniform memory architecture (UMA) and non-uniform memory architecture (NUMA) are two main memory architectures for large scale parallel computers. Since the stacking memory hierarchy for 3D CMPs may contain main memories, it is natural to import UMA and NUMA into the on-chip context. Thirdly, the SRAM and DRAM micro architectures are summarized.

3.1 Memory hierarchy

Because of *locality of reference*, it is possible to build a hierarchical memory system with the balance between cost and performance. The processor's memory system usually consists of different levels, where the levels (e.g. L1 caches) that are nearer from the processor give better performance while the levels (e.g. main memories) that are farther away from the processor have larger capacity and are much cheaper in price-per-bit. Hennessy and Patterson [76] give a good guide for memory hierarchy. Jacob, et al. [77] comprehensively introduce caches and DRAM memories organizations. The memory hierarchy for CMPs is shown in figure 8.

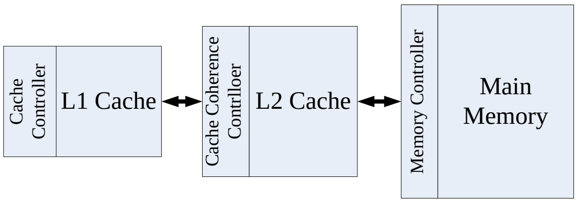

Fig.8. Memory hierarchy for CMPs

The L1 caches are usually partitioned into L1 data cache and L1 instruction cache, which are associated with the processor. The L2 caches of CMPs are usually shared, hence there is a cache coherence controller associated with each L2 cache bank. For 2D ICs, it is common that the main memories are off-chip. In 3D integration, main memories (or part of them) are possible stacked atop processor layers, resulting in larger design space of stacking memory architectures.

3.2 Memory architecture for CMPs

3.2.1 NUCA

For CMPs, the on-chip memory hierarchy is usually a cache-only system, with two or three levels of caches. The first level cache is relatively small, e.g. 16 KB, and typically private. Second level cache can be private or shared. By exploiting the variation in access time across subarrays, the on-chip last level cache (LLC) for CMPs is usually organized as NUCA [78, 79]. NUCA allows nearer cache banks to have lower access latency than further banks. Figure 9 gives a block diagram of NUCA for a McNoC system.

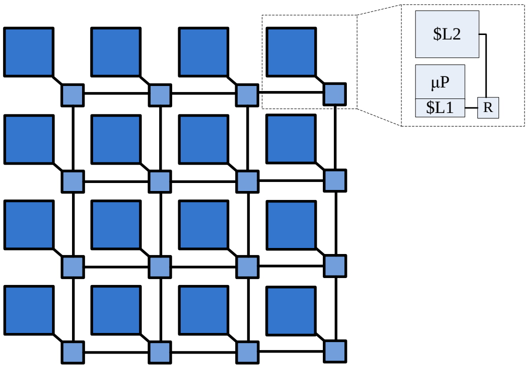

Fig.9. NUCA for McNoC

There are mainly two kinds of NUCA schemes. For *static NUCA* (S-NUCA), the mapping of lines to banks does not change. In contrast, for *dynamic NUCA* (D-NUCA), the data can be migrated to the banks that are closer to the nodes which use them recently.

3.2.2 UMA and NUMA

For 3D integrations, it is possible to stack several main memory layers on top of logic layers. The main memory architecture concept in the context of traditional large scale parallel computer system can be considered in 3D CMPs. In parallel computer systems, the shared memory system can be categorized into two classes, UMA and NUMA [80]. UMA means all the processors access the shared main memory uniformly. The access time to a main memory location is independent of which processor makes the request or which main memory location contains the requested data. Hence UMA model is suitable for symmetric multi-processor (SMP) systems. Figure 10 (a) gives a block diagram of UMA CMP. The L1 caches are associated with each processor and are not shown in figure 10.

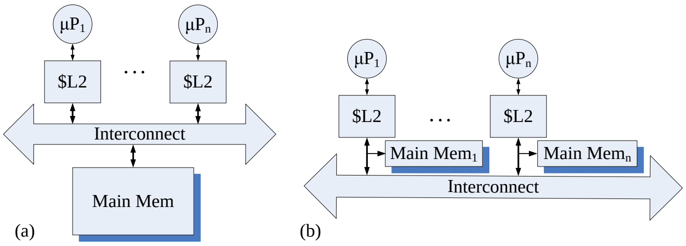

Fig.10. Main memory architecture for CMPs (a) uniform memory architecture; (b) non-uniform memory architecture

For NUMA systems, each node usually has a local main memory, as illustrated in figure 10 (b). The memory access time depends on the main memory location relative to a processor. Accessing local main memory is faster than accessing remote main memory. The local main memory controller determines whether to perform a local memory access or to send a message transaction to a remote memory controller. The main memories can be private, shared or mixed [81].

3.3 SRAM and DRAM micro architecture

SRAM is typically used to build blocks of most caches which occupy a majority portion of the processor die area. The 2D SRAM array-based components have a regular array of memory cells that can be easily spread across multiple dies. The main SRAM array can be viewed as a set of horizontal word lines and a set of vertical bit lines. If a SRAM core is organized in a page manner, a Z-decoder is needed to select the accessed page. A SRAM block structure with four pages of N×M arrays is shown in figure 11.

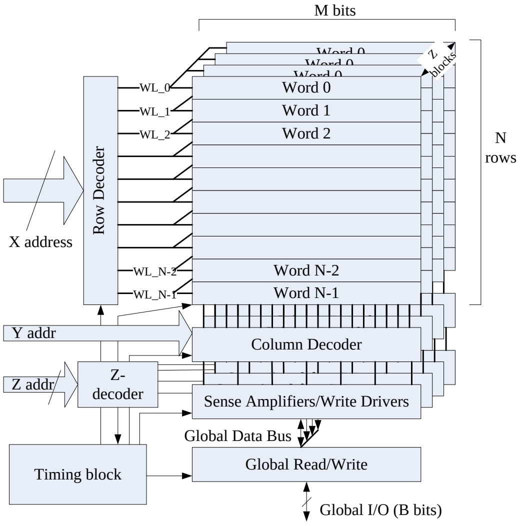

Fig.11. SRAM block diagram

A SRAM array can be divided into multiple submodules called banks. By accessing only the bank that contains the required data, significant power saving and access latency reduction can be achieved.

DRAM is typically organized into *ranks* at the highest level. Each rank is a set of DRAM devices that operate in unison. In each rank, there are multiple *banks*. Each bank can contain several *arrays*. Bank is a set of memory arrays that can be operated independently from other sets. This means that each bank can be activated, precharged, read out etc. at the same time when other banks are being operated. Hence, interleaving accesses enable high bandwidth for multiple memory banks. DRAM die contains one or more memory arrays which are grids of storage cells. The memory arrays are organized into rows and columns. By identifying the row address and column address, the memory controller can access an individual cell inside a DRAM chip. We show the DRAM architecture hierarchy in figure 12.

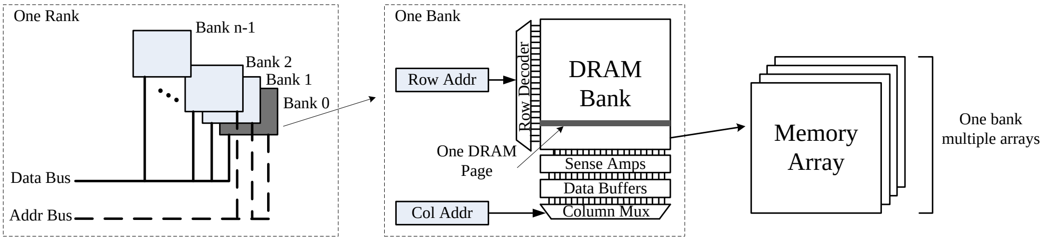

Fig.12. Hierarchy of DRAM architecture

According to the stacking memory hierarchy, we review the memory architecture research for 3D CMPs in two categories: 3D stacking cache-only architecture and 3D stacking main memory architecture in section 4 and section 5 respectively. We summarize the current research and propose the potential future efforts at the end of each section/subsection. We discuss the opportunity of NUMA for 3D CMPs in section 6.

4 Stacking cache-only architectures for 3D CMPs

Caches have been playing an important role in bridging the performance gap between high-speed processors and slow off-chip main memories. 3D CMPs can integrate much larger caches than the 2D counterparts.

4.1 Stacking caches for 3D CMPs

An intuitive design of 3D CMPs is to directly stack cache layers on top of the processor layers. The traditional cache fabrication technology is SRAM. However, DRAM has higher integration density and lower power. It is possible to stack DRAM cache layers in a 3D CMP to further enlarge the on-chip storage. New advantages in technology enable caches to be built from the emerging NVM technologies, such as MRAM and PCRAM in 3D stacked CMPs. Caches fabricated in these technologies offer dramatically different power and performance characteristics compared to SRAM based caches, particularly in the aspects of access latency, cell density, and overall power consumption. There are several works on stacking cache layers in 3D CMPs. The illustration of the stacking cache-only 3D CMP is shown in figure 13. The cache architectures discussed in this paper are LLCs. The L1 caches are the first level caches (FLCs) and usually associated with the processors at the same die. The L1 caches are not shown in figure 13 as well the following figures. The stacking LLC can be one or several banks for each storage layer.

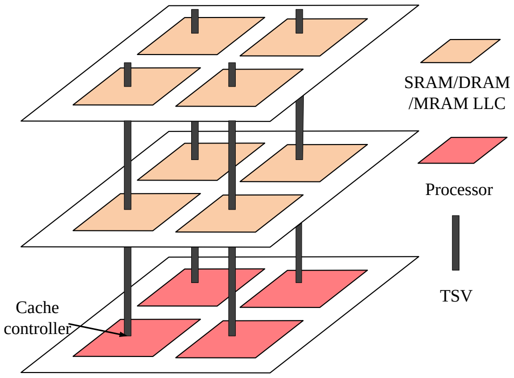

Fig.13. An illustration of stacking cache-only 3D CMP

Black et al. [82] study small scale 3D CMPs with SRAM and DRAM L2 cache layers. The baseline 2D reference design contains two cores and a 4 MB cache. The authors present three 3D stacking structures: 1) 8 MB SRAM cache on top of base line processor layer; 2) 32 MB DRAM cache on bare processor layer with no SRAMs; and 3) 64 MB DRAM atop processor layer which contains 4 MB SRAM. In architecture 3), the 4 MB SRAM is used to store the tags for DRAM caches. Experimental results show that the performance in cycles per memory access gets better with larger caches. Compared with architecture 3), architecture 2) has similar performance, with lower peak temperature and a negligible 0.08^{o}C increase over the baseline 2D reference design. Xu et al. [83] evaluates DRAM as the LLC for a medium scale 3D NoC-based CMP. The 3D CMP consists of 2-layer, one 16-core layer and one 16-bank DRAM cache layer. The experiment results show that the power consumption of 3D CMP with DRAM LLC is reduced by 25.78% compared with the SRAM counterpart. But the average cache hit latency of DRAM LLC gets 27.97% worse than the SRAM design. The performance improvements in [82] and [83] are not consistent. In [83], when the counts of processor get larger, DRAM cache may worsen the performance. The underlying reason why DRAM caches may cause performance degradation is that though the storage density of DRAM is much higher than SRAM and the large DRAM L2 caches are potential to reduce miss rates, DRAM caches make the L2 access time longer. If the negative impact of the latency issue is larger than the effects of cache miss reduction, the performance may get worse. Simply stacking DRAM caches on top of processing cores does not help to improve the performance.

Dong et al. [84] study the impact of stacking MRAM caches on top of the single core layer. For MRAM as replacement of SRAM L2 cache, the instructions-per-cycle (IPC) gets degraded by 10%, however the energy reduces by 95% on average. For memory hierarchy with stacking MRAM L3 cache, the IPC gets improved by about 57% with only 0.685W additional power consumption compared to that of the 3D processor with MRAM L2 caches. The authors also study MRAM as replacement of on-chip main memory and observe a 72% average IPC enhancement compared to the DRAM counterpart. In [85], Sun et al. investigate 3D CMPs with stacking MRAM based L2 caches and compare it against SRAM counterparts in terms of performance and energy. The experiments show that by replacing SRAM, stacking MRAM can reduce the L2 cache miss rate. However, the long write latency hurts the performance. When the write density is high, the negative impact of write operation may degrade the overall performance. Though MRAM can help to reduce the power consumption, the dynamic power consumption may increase significantly because of the high energy associated with the MRAM write operation and the amount of total power saving could be reduced. To take both the advantages of the low SRAM access latency and the high density/low power of DRAM and the emerging NVMs, one potential solution is to use hybrid cache architecture (HCA). Meanwhile, the other types of emerging storage technologies (PCRAM, RRAM, etc.) are expected in the future cache architecture exploration for 3D CMPs.

4.2 Hybrid cache architecture

By applying technologies of high density memories, the cache size can increase significantly under the same chip area constraint. Because the static power dominates for caches, leveraging emerging NVM technologies can help to lower the power consumption significantly. Hence, merging the fast SRAM and slow eDRAM/MRAM/PCRAM/RRAM etc. into the stack is attractive for 3D CMP designers. To efficiently construct and control such a hybrid scheme is an interesting topic.

The first kind of HCA is a reconfigurable structure which stacks both the fast region (SRAM) and the big region (DRAM/NVMs). The HCA switches in two working modes, when the working data set is small, the fast region is used as the LLC and when the working data get large, the big region is activated. We show the transformation of reconfigurable HCA in figure 14.

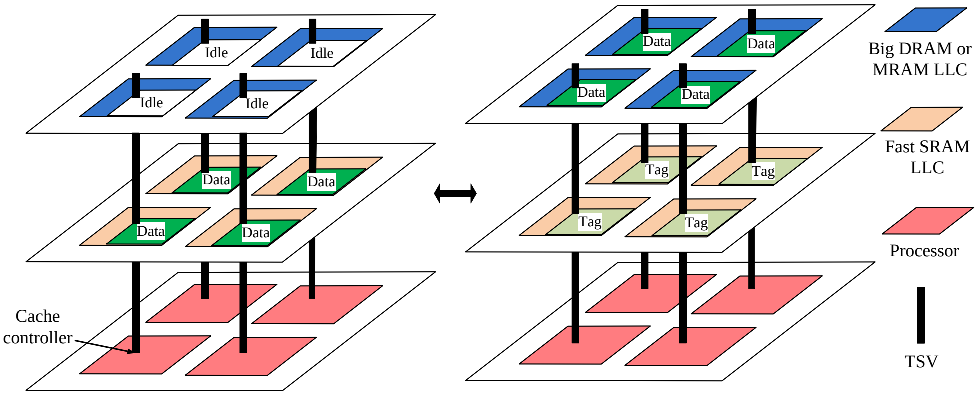

Fig.14. Transformation of reconfigurable HCA

In [86], Madan et al. propose a 3D CMP design that stacks a SRAM cache layer and a DRAM cache layer on a multi-core layer. The SRAM layer has 16 banks which is organized as an S-NUCA. The DRAM layer is on top of the SRAM layer and contains 16 banks. If the thread in a core has small data size, the data is located in the SRAM layer. If the data size gets large that overflows the contents of the SRAM cache, the DRAM cache is activated. The SRAM-DRAM HCA can be reconfigured in terms of associativity, number of sets, and block size upon phases change. The experimental results show that the IPC of the reconfigurable HCA gets the improvement up to 19% over the best baseline with a 48% network power reduction. Inoue et al. [87] introduce a 3D CMP with similar cache architecture. The proposed LLC consists of a SRAM layer and a DRAM layer. When the data size of the application is small (e.g. not bigger than 2 MB), the SRAM is used as the L2 cache to provide fast accesses. If the data size gets bigger than 2 MB, the DRAM is adopted as the L2 cache, and the SRAM stores the tags for the DRAM. On average, the hybrid approach achieves about 25% of memory performance improvement over the conventional 3D implementation.

The basic idea of the second kind of HCA is to allocate frequently accessed data in fast regions and to swap the idle data to the slow regions. We call this scheme as data migration based HCA, which is shown in figure 15.

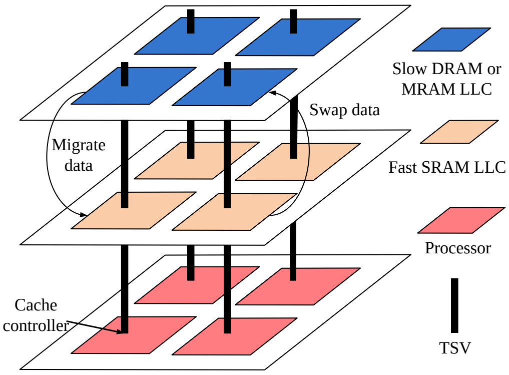

Fig.15. Data migration based HCA

To address the negative impact of long write latency and high write energy consumption of MRAM, Sun et al. [85] propose a SRAM-MRAM HCA which supports data migration to their accessing cores for 3D CMPs. The proposed hybrid architecture keeps as many write intensive data as possible in the SRAM and thereby improve both performance and power. There are 31 ways of MRAM and 1 way of SRAM in the proposed HCA. The data are migrated to SRAM from MRAM when they are written two times, which is called *intra-migration*. The read operations cause *inter-migration* which only allows data to be migrated into the SRAM bank from MRAM bank. On average, the hybrid method improves the IPC by 5.65%, and further reduces the power by 12.45% compared to the MRAM only cache architecture.

Wu et al. [88] propose two types of 3D CMP HCAs. The first one is inter-cache level HCA (LHCA), in which different levels in a cache hierarchy are made of disparate memory technologies (e.g. L2 cache by SRAM and L3 cache by MRAM/PCRAM). The second one is intra-cache level HCA (also named as RHCA in this paper), in which a single level of cache is partitioned into multiple regions with different memory technologies. For RHCA, the cache lines are considered “hot” when their accessed frequency is over the predetermined threshold. The “hot” cache line then migrates into the fast region. Simulation results show that LHCA and RHCA can provide a geometric mean 7% and 12% IPC improvement over the baseline 3-level SRAM cache design respectively. The improvement of LHCA is limited because, although using the dense stacking L3 cache can improve the performance, the performance benefit is offset by the negative long access latency of the dense L3 cache for workloads which prefer shorter latency. The power reduction reaches up to 70% across all configurations. In [89], Wu et al. propose a read-write aware HCA (RWHCA) for 3D CMPs. In RWHCA, the LLC is partitioned into one large read region (MRAM or PCRAM) and one small write region (SRAM). The data are allocated in different regions according to their load/store behavior.

Sharifi and Kandemir [90] propose a feedback control centric scheme to dynamically partition the SRAM/MRAM HCA in 3D CMPs to satisfy the application-level performance quality of service (QoS) target. The proposed scheme takes IPC targets for the applications as inputs and accordingly allocates the cache ways in the SRAM and MRAM. The experiment results indicate that the proposed scheme can satisfy the target QoS for most of the test cases.

We conclude the HCA designs for 3D CMPs in table 3.

Table 3. HCAs for 3D CMPs

	Hybrid cache architecture	
Memory type	SRAM/DRAM	SRAM/emerging NVMs (MRAM, PCRAM)

Innovation	Reconfigurable cache hierarchy [86, 87].	Cache line migration [85, 88, 89];
		Deep cache hierarchy [88];
		QoS hybrid cache partition [90].

We can find from table 3 that the current research on HCA is mainly on migrating frequently accessed data to the fast SRAM caches and swapping less used data to the slow DRAM/MRAM/PCRAM cache stacks. More sophisticated NUCA control schemes can be adopted for stacking caches and HCAs of 3D CMPs. We discuss these schemes in section 4.3.

4.3 NUCA schemes for 3D CMPs

The enlargement of caches in CMPs increases access latency. To mitigate the rising access latency, NUCA is proposed which decomposes the cache into slices. All slices are physically distributed on the die. Thus, cores can fast access the nearby cache lines.

Compared to 2D CMPs, the NUCA design for 3D CMPs is relatively new because of the much larger capacity of caches, new schemes of cache partitions, hybrid architectures, various communication infrastructures, thermal/power constraints, etc. 3D integration also provides new opportunities for cache micro architecture that is not reliable for 2D ICs. In subsection 4.3, we overview the various cache control methods in NUCA for 3D CMPs besides the research in 4.1 and 4.2, we also propose the potential design topics.

Li et al. [91] propose a 3D CMP which places CPUs on several layers and fills the remaining space with L2 cache banks. The different layers are connected by vertical buses. The average physical distance between processors and L2 cache banks is reduced in 3D CMPs compared to its 2D counterpart. The 3D D-NUCA (CMP-DNUCA-3D) and S-NUCA (CMP-SNUCA-3D) improve IPC up to 37.1% and 18.0% respectively over the 2D scheme. The data migration frequency is much reduced for CMP-DNUCA-3D and hence the power consumption is reduced.

In [92], Jung et al. propose an algorithm for set-based dynamic cache partitioning for 3D CMPs. The proposed algorithm reduces average cache access latency by assigning proper amount of cache banks to cores according to cache utility. The assignment of cache banks is done based on network hop distance. The cores are close to their frequently accessed data. Experiment results show that sum of IPC is increased by 10.4% for a 4-core CMP and by 23.3% for a 16-core CMP. Energy consumption per instruction is reduced by 10.0% and 17.9% for the 4-core CMP and the 16-core CMP respectively. Jung et al. [93] investigate the problem of partitioning shared L2 cache for multiple applications. A latency-aware utility-based cache partitioning (LUCP) method which reduces memory access time in a stacked NUCA for 3D CMPs is proposed. LUCP allocates limited cache resources to the processor cores by exploiting variations of access latency and utility of programs. To reduce the time to obtain optimal solution, the authors use a heuristic method. The experiment shows that average memory access time is reduced by up to 32.6% and on average 14.9% compared to a conventional method.

In [94], Guthmuller et al. propose a dynamically adaptive mechanism to optimize the use of the cache architecture according to the workload needs for 3D CMPs. The operating system (OS) controls the cache resource allocation. The OS allocates a large private cache to a given application or shares a given cache tile between one or several applications. The experiment result shows that the tile sharing mechanism can reduce by up to 50% both execution time of the most memory intensive application and the overall traffic to the external memory with the area overhead as 10% of the cache memory.

Loh [95] leverages the DRAM as LLC for 3D CMPs. The author proposes a novel DRAM cache architecture where each set is organized as multiple logical FIFO or queue structures. A novel pseudo-LRU based cache replacement policy is proposed. The replacement policy has the properties of providing performance isolation between cores and reducing the number of dead cache line entries. The proposed scheme provides 29.1% additional IPC on top of the benefits of simplistic 3D-stacking of DRAM.

Zhao et al. [96] investigate cache management techniques which tolerate process variation (PV) in a 3D CMP with stacked DRAM. The authors develop cache migration schemes that utilize fast banks while limiting the cost due to migration. The experimental result shows that a variation-aware management can improve the speedup over the baseline (where the slowest bank speed is assumed for all banks) by 16.5% on average. The resulting speed is only 0.8% away from a chip with no PV.

Mishra et al. [97] study the integration of STT-RAM caches in a 3D CMP and propose re-ordering schemes at the on-chip network level to mitigate the write overhead problem. The re-ordering scheme is based on that instead of sending subsequent requests to a write-busy STT-RAM bank, the network serves requests to idle cache banks to hide the write latency. Heuristics mechanisms are presented to accurately predict the busy/idle status of a cache bank a few hops away from the destination. A request prioritization mechanism is applied in the routers to effectively utilize the idle cache banks. The experiment results demonstrate that the proposed approach can lead to an average 14% improvement of IPC for a 2-layer 3D CMP with 64 nodes on each layer.

We summarize the NUCA exploration works in table 4.

Table 4. NUCA design exploration for 3D CMP

No.	Innovation
[91]	CPU/cache placement scheme; CMP-DNUCA-3D/CMP-SNUCA-3D
[92, 93]	Cache partitioning method

[94]	Dynamically adaptive 3D cache allocation scheme controlled by OS
[95]	Novel multiple logical FIFO DRAM cache architecture
[96]	Process variation aware cache management by data migration
[97]	Re-ordering network schemes to hide STT-RAM write overhead

We can find from table 4 that there are two main streams in 3D NUCA design. The first group ([91-94]) of works study the similar schemes in 3D CMPs as in 2D context. The second group are quite unique for 3D CMPs, which consider the ability of new memory structure [95], process variation [96], network re-ordering to hide STT-RAM write overhead [97] etc. For the first group, there still exists a large exploration space, e.g. victim cache planning, NUCA data block placement and replication etc. For the second group, the exploration is just a beginning, we can find more opportunities in section 6. Another important issue is cache coherence schemes for 3D CMPs. However there is almost no paper concentrated on this topic, which is expected in the future research.

5 Stacking main memory architectures for 3D CMPs

By leveraging vertical TSVs, stacking main memories atop processor layers can provide large memory bandwidth and hence mitigate the “Memory Wall” problem. Several studies show the performance benefits of 3D stacking main memories. We review the status of the current research progress of stacking main memory architectures for 3D CMPs in this section. A 3D CMP with stacking main memories is illustrated in figure 16.

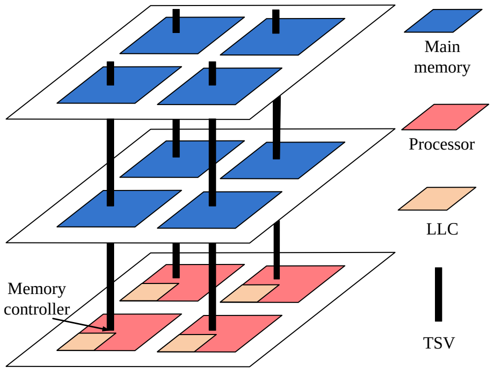

Fig.16. 3D stacking main memory architecture for CMPs

Meng et al. [98] introduce a framework to model on-chip DRAM main memory and analyze the performance, power, and temperature tradeoffs of 3D CMPs. The authors propose a runtime optimization policy to maximize performance while maintaining power and thermal constraints. The policy dynamically monitors workload behavior and selects among low-power and turbo operating modes accordingly. The experiment demonstrates that by enabling parallel access, the IPC of 3D CMP with DRAM is improved by up to 86.9% compared to single-bus access. Under the fixed voltage-frequency (V-F) settings, the IPC of runtime optimization policy improves 36.1% for a 16-core 3D CMP with stacked DRAM compared to a statically optimized 3D system. The energy-delay product (EDP) is reduced by up to 49.4% compared to a 3D system managed by a temperature triggered dynamic V-F scaling policy. In [99], Meng and Coskun study the complex interplay of performance, energy, and temperature for 3D CMPs with stacked DRAM main memories. The authors propose a memory management scheme targeting applications with spatial variations in DRAM accesses and temperature-aware mapping of memory accesses to DRAM banks. Experimental results on a 16-core 3D CMP indicate up to 88.5% improvement in EDP compared to an equivalent 2D system with off-chip memory.

Loh [100] explores an aggressive 3D DRAM organization that makes use of the wide die-to-die bandwidth for 3D CMPs. The number of ranks and memory controller interfaces are increased. The authors also propose a vector bloom filter to enable the L2 miss handling architecture (MHA) which can further utilize the increased capacity of the 3D-stacked memory system. The experiment results show that the proposed memory organization achieves a 1.75X speedup over previously proposed 3D DRAM approaches on memory intensive workloads on a quad-core processor. The scalable L2 MHA can further yield 17.8% performance improvement over the 3D stacked memory architecture.

Jacob et al. [101] evaluate shared memory architectures stacking on the high-clock-rate processors. SiGe hetero-structure bipolar transistor is the basis which builds the high-clock-rate processors. The operating frequency can be in the order of 16 or 32 GHz. The authors study various architectural design options to reduce the impact of the "Memory Wall" problem on the processor performance, e.g. the influence of multitier, multibank, multiport 3D memories for the performance of 3D CMPs. The experimental results show that 3D integrated chips can help to address the “Memory Wall” problem of CMPs.

In [102], Woo et al. declare that the L2 cache and DRAM interface need to be re-designed to take full advantage of 3D massive bandwidth. The authors study the cache line effort and find that most modern applications benefit from a larger cache line for a much larger L2 cache. However, the access latency increases almost linearly with the line size. To overcome this, the authors propose a SMART-3D memory architecture that feeds a large number of small cache lines to L2 caches once through the wide TSVs to hide the latency behind the large data transfers. The SMART-3D memory uses a vertical L2 fetch/write-back network to the 3D main memory using a large array of TSVs. The experiment results show that for single-threaded memory intensive applications, the speedups of 3D multi-core system range from 1.53 to 2.14 compared to a conventional 2D architecture and from 1.27 to 1.72 compared with prior 3D stacked memory designs on average. The energy consumption in L2 cache and 3D DRAM gets lower by the reduction of row buffer misses.

Zhang and Li [103] investigate using PCRAM to be the main memory for 3D stacking CMPs. To improve PCRAM write performance and lifetime, a hybrid PCRAM-DRAM memory architecture and an OS level paging scheme are proposed. The experimental results show that compared to die stacked planar DRAM, the proposed design can reduce the overall power consumption of the memory system by 54% with only 6% IPC degradation. The temperature of 3D chips can be alleviated up to 4.25°C and the performance has a speedup of up to 1.1X.

We summarize the main memory architecture designs for 3D CMPs in table 5.

Table 5. Stacking main memory architectures for 3D CMPs

No.	Innovation	Technology

[98, 99]	Simulation framework;
	Runtime V-F optimization policy for performance under power/thermal constraints;
	Temperature aware DRAM banks mapping;
		DRAM

[100]	Finer rank and interface; L2 MHA;	DRAM
[101]	High frequency 3D-stacked memory organization	DRAM
[102]	Redesign cache and memory using high-density TSVs	DRAM
[103]	DRAM-PCRAM hybrid main memory architecture; PCRAM-aware OS paging scheme;	DRAM/ PCRAM

We can find from table 5 that the research of 3D stacked main memory architecture is mainly on leveraging wide TSV bandwidth to enhance the performance of 3D CMPs. There also exists research on emerging NVM as main memories for 3D CMPs. We look forward to more publications on stacking main memory architectures for 3D CMPs in two aspects, explorations of stacking NVM architectures and designs of distributed shared memory (DSM) architectures.

6 Challenges and Outlook

6.1 Thermal issue

Despite the promising advantages of 3D technology, there are significant concerns for the thermal impact. The increased power density can result from placing one power hungry block over another in the multi-layered 3D stacks. Thermal hot spots in 3D chips can cause higher peak temperature than that in 2D chips. High temperature has adverse impact on circuit performance, leakage power, and reliability. It is very critical to build the thermal behavior model of 3D CMPs and investigate possible solutions to mitigate thermal problems. It is crucial to take careful thermal floor planning to avoid thermal failures [104, 105]. To mitigate the thermal impact, thermal-aware design techniques must be adopted for 3D CMP designs from circuit-level up to system architecture level.

The research presented in [48] is the first attempt to study the performance benefits of 3D technology under the impact of thermal constraints. Thermal considerations place a lower limit on the operating frequency in 3D ICs. Oprins et al. [106] present a thermal experimental and modeling characterization of a 3D packaged DRAM on logic stack. The authors study the thermal influence of logic hot spot dissipation on the temperature profile of the DRAM and the logic die using a high power and a low power experimental configuration respectively. The experiments show that the temperature increases above ambient in the logic die is up to 7X higher than the temperature increase in the DRAM die. Moreover, the authors propose a thermal finite element model to assess the thermal behavior of the packaged 3D stacks. Lee et al. [107] focus on the problem of cache data mapping for a 3D CMP with stacked L2 cache to minimize the overall system energy. An integer linear programming (ILP) based design time solution is presented. ILP generates two outputs: (1) number of cache ways to each thread, and (2) the physical location of the allocated cache ways. The simulation results show that the proposed ILP-based method yields up to 30.7% energy reduction compared to methods which consider temperature distribution only. Chatterjee et al. [108] analyze the implications of 3D die-to-die thermal coupling on power, performance and aging of SRAM. They point out that the power variations in cores may significantly affect the spatial and temporal variations in performance of SRAM blocks. The trade-off between the gains in interconnect delay and the performance degradation of SRAM caused by increased temperature should be cohesively considered for 3D CMPs. Thermal issue is a key design constraint in 3D CMPs. Hence it is important to design memory hierarchies for 3D CMPs with the concern of thermal constraint. New low-power memory technologies, memory control policies and memory architectures are expected for 3D CMPs.

6.2 TSV count

Despite the decreasing via sizes, the via pads (i.e., the via endpoints) ultimately limit the via density. Currently, via pads do not scale at the same rate as the vias themselves. Compared to a wire pitch of 0.1μm, inter-layer vias are significantly larger and cannot achieve the same wiring density as intra-layer interconnects. In a 3D IC, the die-to-die interconnect pitch must pass through the active device layer, imposing a larger area overhead than corresponding horizontal wires. Moreover, fabricating such a circuit involves several costly manufacturing steps. Each manufacturing step adds a risk for defects, which may result in potential yield reduction. According to results of the leading 3D technology owners, the yield is an exponential function of TSV defect rate and the number of TSVs. Thus, the yield exponentially decreases when the number of TSVs reaches a technology dependent limit, about from 1,000-10,000 [36]. Based on the IMEC 3D cost model [109], TSV processing cost is the dominating cost for a 3D wafer. Assuming a CMOS processing technology of 65nm with 200 mm silicon wafers, 46% to 65% costs are spent on TSV processing [110]. Besides, the thermo-mechanical stress caused by TSVs is known that severely impacts the carrier mobility of nearby devices [111, 112]. Apparently, in terms of chip area and manufacturing cost, the total number of TSVs should be limited to a practical number. The memory hierarchy designs and control schemes need to take into account the TSV numbers and costs for 3D CMPs. The research on tradeoffs between TSV counts and stacked memory performance for 3D CMPs is expected.

6.3 Wide I/O and LPDDR3

New design standards are emerging, such as wide I/O [113] and low-power DDR3 (LPDDR3) [114], to achieve high bandwidth, reduced power consumption and small form factor.

Wide I/O supports 4 128-bit channels and provides a total peak bandwidth of 12.8 GBps. Kim et al. [115] demonstrate a mobile wide I/O DRAM design with 4 channels, featuring 12.8 GBps data bandwidth. The whole chip contains four memory controllers with 128 bit wide interface. Each channel has its own input pins while sharing external power pins and internal voltage supply with other channels. Takaya et al. [116] present a test vehicle for an aggressive TSV-based wide I/O data communication in a 3-tier 3D chip stack. The width of the wide I/O in this paper is 4096 bits with capability of 100 GBps, which is superior to standard mobile memory wide I/O specifications with 512 channels at 12.8 GBps.

LPDDR3 is announced in May 2012 [114], which is designed to satisfy the memory demands of the latest generation of mobile devices such as smart phones, tablets, ultra-thin notebooks etc. LPDDR3 SDRAM is a double data rate synchronous DRAM device internally configured as an 8-bank memory. Bae et al. [117] present a 1.6 Gbps/pin×32 4 Gb LPDDR3 SDRAM design, achieving 6.4 GBps total data bandwidth in 30 nm DRAM technology. It is shown that LPDDR3 is more power efficient than wide I/O.

More works of wide I/O and LPDDR3 memory are expected in 3D CMPs. Which one is more suitable in the 3D context is still an open issue.

6.4 3D-UMA and 3D-NUMA

3D die stacking makes it possible that the off-chip main memories are integrated on-chip. The memory bandwidth can be largely expanded without the limitation of chip port counts. Meanwhile, the number of memory controllers can be enlarged and the micro-architecture of DRAM can be enhanced in the 3D context. Hence, the access latency to the stacking main memory is no longer a bottleneck. However, the access distance between the LLC and on-chip main memory may take several hops, resulting high latency and power consumption. With the counts of on-chip core increasing, the impact of main memory access latency may become more notable. Most of the existing works focus on stacking main memory to take the advantage of wide vertical bandwidth. However, the on-chip distributed memory architecture is not researched systematically. There are plenty of works for large scale parallel computer systems with distributed shared memory (DSM) [118-120]. The power and thermal constraints of 3D CMPs are different from the large scale parallel computers. The design of 3D DSMs should combine with the communication networks and take the balance between complexity and power consumption.

6.5 3D oriented memory micro-architectures

It is potential to redesign SRAM in the 3D IC context to reduce wire RC delays. 3D integration technologies can adopt existing planar processes and designs to divide the 2D SRAM to individual dies. Conventional 2D memories are designed under a variety of electrical constraints with off-chip I/O connections and packaging. However, for the 3D stacked memories, many of the basic design assumptions need to be revised. There are several works that focus on 3D SRAM and 3D DRAM memory micro organizations, which concern performance, power/thermal and 3D process variation etc. We firstly summarize the research on 3D SRAM and 3D DRAM designs respectively.

Puttaswamy and Loh [121] propose the 3D SRAM arrays from the bank level down to the circuit level. The wire length is reduced within SRAM arrays. The footprint is diminished which reduces the wires required for global routing. The diminution of wire length involves reduction of latency and energy. A 3D implementation of a 128-entry multiported SRAM array is proposed. Chen and Davis [122] study the design which converts a SRAM design from 2D to 3D by distributing the sub-arrays in a bank to several tiers. The authors present optimized sub-array partitioning for 3D SRAMs which can provide extra benefit in delay reduction. Hsu and Wu [123] also propose a 3D SRAM architecture design by splitting arrays and stacking banks. Nho et al. [124] redesign 3D SRAMs in bit level. They propose a new bit-line and decoder architecture for 3D SRAMs. The local bit-lines are extended upward, and the SRAM cells are connected vertically by inter-layer vias. The global bit-line reduces by a factor of the number of layers, which can significantly reduce bit-line capacitance. Other works [125, 126] propose novel architectures of 3D SRAM cell.

For 3D DRAM, Weis et al. [127] use commodity DRAMs to design the 3D stacked DRAM architecture with respect to performance, energy and area efficiency. Sekiguchi et al. [128] present a high data rate (1 TBps), high capacity (1 Gb) DRAM architecture with TSVs distributed in 16 memory cores. Vignon et al. [129] focus on the circuit design of DRAM architecture that improves the DRAM cache performance for low power chip design. The authors propose a DRAM architecture which uses a finer granularity matrix subdivision to reduce both the access time and the dynamic power consumption. By using a new implementation of the refresh operation, the impact on access delay and passive energy consumption can be reduced. In [130], Sun et al. firstly develop a coarse-grained partitioning strategy for 3D DRAM in intra-subarray level to reduce TSV counts. Secondly, the authors propose a heterogeneous 3D DRAM architecture to implement a large private L2 cache and the shared memory. The speed of DRAM is improved by using smaller size of individual sub-array and multiple threshold voltage. In [131], Jun et al. propose an asymmetric 3D DRAM architecture for heterogeneous CMPs and the automatic synthesis method. In the asymmetric 3D stacked DRAM architecture, the DRAM die is divided into multiple segments and the segments are optimized for different memory requirements.

Though a variety of novel micro-architectures for 3D stacked SRAM and DRAM memories are proposed, only a few of them are integrated into the 3D CMPs. A significant exploration space is left to take the advantage of the novel SRAM/DRAM designs for 3D CMPs. 3D SRAM/DRAM high level models are also expected for the system level exploration of 3D CMPs. Meanwhile, we look forward to novel micro-architecture designs on emerging NVMs for 3D CMPs.

6.6 Memory architecture for heterogeneous and embedded 3D CMPs

CMPs can be designed heterogeneously in architectures, instruction set architectures (ISAs), frequency and underlying process technology to well fit the demands of different applications [132]. Saripalli et al. [133] propose energy efficient heterogeneous CMPs in two aspects, 1) the tunnel FET-CMOS (TFET-CMOS) hybrid cores, in which TFET has a distinct energy advantage and 2) a hybrid SRAM-MRAM cache architecture with data migration, in which MRAM has little leakage power. There are works that integrate CPUs and GPUs [134, 135] in the same 2D die. The Heterogeneous System Architecture (HSA) Foundation [136] is recently formed by a number of leading companies to provide a common development approach. For 3D integration, research that explores memory architectures for 3D GPUs is emerging. Al Maashri et al. [137] study the performance of 3D stacking caches which include SRAM and MRAM on the GPU. However, there are few works that study memory architecture for 3D heterogeneous CMPs.

Another trend for 3D CMP design is the embedded system, especially mobile applications. Facchini et al. [138] examine the power and performance benefits for three different 3D stacked DRAM scenarios for mobile applications. Chang and Chen [139] discuss various noise reduction and sensing schemes for low-voltage embedded NVM of 3D ICs. Chang et al. [140] analyze the energy and performance impacts of 3D memory integration in DSP systems. Zhang et al. [141] study a 3D embedded multi-core DSP processor called 3D-iSPA, which targets multimedia applications. Hierarchical AXI crossbar is employed as the backbone interconnect. Clermidy et al. [142] investigate three promising perspectives for short-to-medium-term adoption of 3D stacking technology in high-end embedded CMPs: wide I/O memory scheme, 3D NoC and passive or active interposer. Compared to general purpose CMPs, the memory architectures for 3D embedded CMPs should be customized and adapted to embedded application's characteristics. Mobile applications put a stringent power constraint on memory architectures for embedded CMPs.

7 Conclusion

3D CMPs can exploit the high vertical bandwidth provided by TSVs and high density, low power emerging NVM technologies. 3D CMPs is a promising solution to mitigate the severe “Memory Wall” problems. In this paper, we present an overview of memory architectures for 3D CMPs. We discuss various technologies, designs and challenges. The memory architectures for 3D CMPs appear mainly in two categories: stacking cache-only architecture and stacking main memory architecture. 3D CMPs design is a promising approach for future CMP designs. However, the research on 3D CMPs is still in its infant stage. There exists a huge exploration space for 3D CMPs from the process level, circuit level up to system level. Hence, as one of the most important topics, the memory architecture of 3D CMPs needs to be explored systematically. The involved design trade-offs, particularly the thermal/power delivery and cost issues, should be studied carefully. This paper provides references to the interested readers in 3D stacking memory architecture design and guides the future research.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 61176024, 61006018, 61370040 and 61376075; Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20120091110029; IUR Cooperative Innovation Foundation-Prospective Research Project of Jiangsu Province under Grant No.BY2013072-05; A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The authors would like to thank the anonymous reviewers for their feedback and suggestions.

References

[1] Nayfeh, B. & Olukotun, K., A single-chip multiprocessor, *Computer*, volume 30, issue 9, pp. 79-85, 1997.

[2] Chiu, Jih-Ching; Yang, Kai-Ming; Chou, Yu-Liang A hyperscalar dual-core architecture for embedded systems, *Microprocessors and Microsystems*, volume 37, issue 8, pp. 929-940, 2013.

[3] Vangal, S.; Howard, J.; Ruhl, G.; Dighe, S.; Wilson, H.; Tschanz, J.; Finan, D.; Iyer, P.; Singh, A.; Jacob, T.; Jain, S.; Venkataraman, S.; Hoskote, Y. & Borkar, N., An 80-tile 1.28TFLOPS network-on-chip in 65nm CMOS, *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, Feb. 2007, pp. 98-589.

[4] Bell, S.; Edwards, B.; Amann, J.; Conlin, R.; Joyce, K.; Leung, V.; MacKay, J.; Reif, M.; Bao, L.; Brown, J.; Mattina, M.; Miao, C.-C.; Ramey, C.; Wentzlaff, D.; Anderson, W.; Berger, E.; Fairbanks, N.; Khan, D.; Montenegro, F.; Stickney, J. & Zook, J., TILE64 - Processor : A 64-core SoC with mesh interconnect, *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, Feb. 2008, pp. 88-598.

[5] Kan, R.; Tanaka, T.; Sugizaki, G.; Nishiyama, R.; Sakabayashi, S.; Koyanagi, Y.; Iwatsuki, R.; Hayasaka, K.; Uemura, T.; Ito, G.; Ozeki, Y.; Adachi, H.; Furuya, K. & Motokurumada, T., A 10th generation 16-core SPARC64 processor for mission-critical UNIX server, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2013, pp. 60-61.

[6] Wulf, W. & McKee, S., Hitting the memory wall: implications of the obvious, *ACM SIGARCH Computer Architecture News*, volume 23, issue 1, pp. 20-24, 1995.

[7] Jacob, P.; Erdogan, O.; Zia, A.; Belemjian, P.; Kraft, R. & McDonald, J., Predicting the performance of a 3D

processor-memory chip stack, *IEEE Design & Test of Computers*, volume 22 , issue 6, pp. 540-547, 2005.

[8] Sapatnekar, S. & Nowka, K, New Dimensions in 3D Integration, *IEEE Design & Test of Computers*, volume 22 , issue 6, pp. 496-497, 2005.

[9] Pamunuwa, D., Memory technology for extended large-scale integration in future electronics applications, *Design, Automation and Test in Europe*, March 2008, pp. 1126-1127.

[10] Marchal, P.; Van der Plas, G.; Eneman, G.; Moroz, V.; Badaroglu, M.; Mercha, A.; Thijs, S.; Linten, D.; Guruprasad, K.; Stucchi, M.; Vandevelde, B.; O'Prins, H.; Cherman, V.; Croes, K.; Redolfi, A.; La Manna, A.; Travaly, Y.; Beyne, E. & Cartuyvels, R., 3D technology roadmap and status, *IEEE International Interconnect Technology Conference and Materials for Advanced Metallization*, May 2011, pp. 1-3.

[11] Van der Plas, G.; Limaye, P.; Loi, I.; Mercha, A.; Oprins, H.; et al., Design issues and considerations for low-cost 3-D TSV IC technology, *IEEE Journal of Solid-State Circuits*, volume 46, issue 1, pp. 293-307, 2011.

[12] Kim, D. H.; Athikulwongse, K. & Lim, S. K., A study of through-silicon-via impact on the 3D stacked IC layout, *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers*, Nov. 2009, pp. 674-680.

[13] Kumar, S.; Jantsch, A.; Soininen, J.-P.; Forsell, M.; Millberg, M.; Oberg, J.; Tiensyrja, K. & Hemani, A., A network on chip architecture and design methodology, *IEEE Computer Society Annual Symposium on VLSI*, Apr. 2002, pp.105-112.

[14] Pavlidis, V. & Friedman, E., 3-D topologies for networks-on-chip, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, volume 15, issue 10, pp. 1081-1090, 2007.

[15] Kim, D. H.; Athikulwongse, K.; Healy, M.; Hossain, M.; Jung, M.; Khorosh, I.; Kumar, G.; Lee, Y.-J.; Lewis, D.; Lin, T.-W.; Liu, C.; Panth, S.; Pathak, M.; Ren, M.; Shen, G.; Song, T.; Woo, D. H.; Zhao, X.; Kim, J.; Choi, H.; Loh, G.; Lee, H.-H. & Lim, S. K., 3D-MAPS: 3D massively parallel processor with stacked memory, *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, Feb. 2012, pp. 188-190.

[16] Fick, D.; Dreslinski, R.; Giridhar, B.; Kim, G.; Seo, S.; Fojtik, M.; Satpathy, S.; Lee, Y.; Kim, D.; Liu, N.; Wieckowski, M.; Chen, G.; Mudge, T.; Sylvester, D. & Blaauw, D., Centip3De: A 3930 DMIPS/W configurable near-threshold 3D stacked system with 64 ARM Cortex-M3 cores, *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, Feb. 2012, pp. 190-192.

[17] Fick, D.; Dreslinski, R.; Giridhar, B.; Kim, G.; Seo, S.; Fojtik, M.; Satpathy, S.; Lee, Y.; Kim, D.; Liu, N.; Wieckowski, M.; Chen, G.; Mudge, T.; Blaauw, D. & Sylvester, D., Centip3De: a cluster-based NTC architecture with 64 ARM Cortex-M3 cores in 3D stacked 130 nm CMOS, *IEEE Journal of Solid-State Circuits*, volume 48, issue 1, pp. 104-117, 2013.

[18] Dreslinski, R. G.; Fick, D.; Giridhar, B.; Kim, G.; Seo, S.; Fojtik, M.; Satpathy, S.; Lee, Y.; Kim, D.; Liu, N.; Wieckowski, M.; Chen, G.; Sylvester, D.; Blaauw, D. & Mudge, T., Centip3De: a 64-core, 3D stacked near-threshold system, *IEEE Micro*, volume 33, issue 2, pp. 8-16, 2013.

[19] Wordeman, M.; Silberman, J.; Maier, G. & Scheuermann, M., A 3D system prototype of an eDRAM cache stacked over processor-like logic using through-silicon vias, *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, Feb. 2012, pp. 186-187.

[20] Jabbar, M. H.; Houzet, D. & Hammami, O., 3D multiprocessor with 3D NoC architecture based on Tezzaron technology, *IEEE International 3D Systems Integration Conference*, Feb. 2012, pp. 1-5.

[21] Deng, Y. & Maly, W, 2.5-dimensional VLSI system integration, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, volume 13, issue 6, pp. 668-677, 2005.

[22] Tokunaga, S., 2.5D design methodology, *18th Asia and South Pacific Design Automation Conference*, Jan. 2013, pp. 399-402.

[23] Bruce R H, Meuli W P, Ho J. Multi chip modules, *Proceedings of the 26th ACM/IEEE conference on design automation*, 1989, June 25-29, pp. 389-393.

[24] Warnock, J.; Chan, Y.; Harrer, H.; Rude, D.; Puri, R.; Carey, S.; Salem, G.; Mayer, G.; Chan, Y.-H.; Mayo, M.; Jatkowski, A.; Strevig, G.; Sigal, L.; Datta, A.; Gattiker, A.; Bansal, A.; Malone, D.; Strach, T.; Wen, H.; Mak, P.-K.; Shum, C.-L.; Plass, D. & Webb, C., 5.5GHz system z microprocessor and multi-chip module, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2013, pp. 46-47.

[25] Velenis, D.; Detalle, M.; Civale, Y.; Marinissen, E. J.; Beyer, G. & Beyne, E., Cost comparison between 3D and 2.5D integration, *4th Electronic System-Integration Technology Conference*, Sept. 2012, pp. 1-4.

[26] Knickerbocker, J.; Andry, P.; Colgan, E.; Dang, B.; Dickson, T.; Gu, X.; Haymes, C.; Jahnes, C.; Liu, Y.; Maria, J.; Polastre, R.; Tsang, C. K.; Turlapati, L.; Webb, B.; Wiggins, L. & Wright, S., 2.5D and 3D technology challenges and test vehicle demonstrations, *IEEE 62nd Electronic Components and Technology Conference*, May 2012, pp. 1068-1076.

[27] Patrick Dorsey, *Xilinx white paper: Vertex-7 FPGAs*, 2010.

[28] Deng, Y. & Maly, W., 2.5D system integration a design driven system implementation schema, *Asia and South Pacific Design Automation Conference*, Jan. 2004, pp. 450-455.

[29] Haensch, W., Why should we do 3D integration?, *45th ACM/IEEE Design Automation Conference*, June 2008, pp. 674 -675.

[30] Xie, Y.; Loh, G.; Black, B. & Bernstein, K., Design space exploration for 3D architectures, *ACM Journal on Emerging Technologies in Computing Systems*, volume 2, issue2, pp. 65-103, 2006.

[31] Loh, G. & Xie, Y., 3D stacked microprocessor: are we there yet?, *IEEE Micro*, volume 30, issue3, pp. 60-64, 2010.

[32] Chen, K. & Tan, C., Integration schemes and enabling technologies for three-dimensional integrated circuits, *IET Computers Digital Techniques*, volume 5, issue 3, pp. 160-168, 2011.

[33] Li, J.-F. & Wu, C.-W., Is 3D integration an opportunity or just a hype?, *15th Asia and South Pacific Design Automation Conference*, Jan. 2010, pp. 541-543.

[34] Loh, G. H., Computer architecture for die stacking VLSI Design, *International Symposium on Automation, and Test*, April 2012, pp. 1-2.

[35] Burns, J.; Carpenter, G.; Kursun, E.; Puri, R.; Warnock, J. & Scheuermann, M., Design, CAD and technology challenges for future processors: 3D perspectives, *48th ACM/EDAC/IEEE Design Automation Conference*, June 2011, pp. 212.

[36] Abbas Sheibanyrad, Frederic Petrot, and Axel Jantsch, editors, *3D integration for NoC-based SoC architectures*, Springer, January 2011.

[37] Yuan Xie, Jason Cong, Sachin Sapatnekar, *Three-dimensional integrated circuit design: EDA, design and microarchitectures*, Springer, 2010.

[38] Wong, S.; El-Gamal, A.; Griffin, P.; Nishi, Y.; Pease, F. & Plummer, J., Monolithic 3D integrated circuits, *International Symposium on VLSI Technology, Systems and Applications*, April 2007, pp. 1-4.

[39] Bobba, S.; Chakraborty, A.; Thomas, O.; Batude, P.; Pavlidis, V. & De Micheli, G., Performance analysis of 3-D monolithic integrated circuits, *IEEE International 3D Systems Integration Conference*, Nov. 2010, pp. 1-4.

[40] Batude, P.; Ernst, T.; Arcamone, J.; Arndt, G.; Coudrain, P. & Gaillardon, P.-E., 3-D sequential integration: A key enabling technology for heterogeneous co-integration of new function with CMOS, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, volume 2, issue 4, pp. 714-722, 2012.

[41] Vinet, M.; Batude, P.; Tabone, C.; Previtali, B.; LeRoyer, C.; Pouydebasque, A.; Clavelier, L.; Valentian, A.; Thomas, O.; Michaud, S.; Sanchez, L.; Baud, L.; Roman, A.; Carron, V.; Nemouchi, F.; Mazzocchi, V.; Grampeix, H.; Amara, A.; Deleonibus, S. & Faynot, O., 3D monolithic integration: Technological challenges and electrical results, *Microelectronic Engineering*, volume 88, issue 4, pp. 331-335, 2011.

[42] Patti, R., Three-dimensional integrated circuits and the future of system-on-chip designs, *Proceedings of the IEEE*, volume 94, issue 6, pp. 1214-1224, 2006.

[43] Tezzaron Semiconductor, http://www.tezzaron.com/.

[44] Park, K.-T.; Kim, D.; Hwang, S.; Kang, M.; Cho, H.; Jeong, Y.; ll Seo, Y.; Jang, J.; Kim, H.-S.; Jung, S.-M.; Lee, Y.-T.; Kim, C. & Lee, W.-S., A 45nm 4Gb 3-dimensional double-stacked multi-level NAND flash memory with shared bitline structure, *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, Feb. 2008, pp. 510-632.

[45] Jeddeloh, J. & Keeth, B., Hybrid memory cube new DRAM architecture increases density and performance, *Symposium on VLSI Technology*, June 2012, pp. 87-88.

[46] International technology roadmap for semiconductors, *ITRS 2009*， Semiconductor Industry Association, http://www.itrs.net/Links/2009ITRS/Home2009.htm.

[47] Kim, D. H.; Mukhopadhyay, S. & Lim, S. K., Through-silicon-via aware interconnect prediction and optimization for 3D stacked ICs, *Proceedings of the 11th international workshop on System level interconnect prediction*, July 2009, pp. 85-92.

[48] Loi, G.; Agrawal, B.; Srivastava, N.; Lin, S.; Sherwood, T. & Banerjee, K., A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy, *Proceedings of the 43rd annual Design Automation Conference*, 2006, pp.

991-996.

[49] International technology roadmap for semiconductors, *ITRS 2011* , Semiconductor Industry Association, http://www.itrs.net/Links/2011ITRS/2011Chapters/2011Interconnect.pdf.

[50] Bernstein, K.; Andry, P.; Cann, J.; Emma, P.; Greenberg, D.; Haensch, W.; Ignatowski, M.; Koester, S.; Magerlein, J.; Puri, R. & Young, A., Interconnects in the third dimension design challenges for 3D ICs, *44th ACM/IEEE Design Automation Conference*, 2007, pp. 562-567.

[51] Eleftheriou, E.; Haas, R.; Jelitto, J.; Lantz, M. & Pozidis, H., Trends in storage technologies, *Special Issue on Data Management using Modern Storage Hardware, IEEE Bulletin of the Technical Committee on Data Engineering*, volume 33, no. 4, pp. 4-13, 2010.

[52] Zhao, W.; Zhang, Y.; Lakys, Y.; Klein, J.-O.; Etiemble, D.; Revelosona, D.; Chappert, C.; Torres, L.; Cargnini, L.; Brum, R.; Guillemenet, Y. & Sassatelli, G., Embedded MRAM for high-speed computing, *IEEE/IFIP 19th International Conference on VLSI and System-on-Chip*, Oct. 2011, pp. 37-42.

[53] F. Ren, *Energy-performance characterization of CMOS/magnetic tunnel junction (MTJ) hybrid logic circuits*, Master of Science Thesis, University of California, Los Angeles, 2011.

[54] Dong, X.; Wu, X.; Sun, G.; Xie, Y.; Li, H. & Chen, Y., Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement, *Proceedings of the 45th annual Design Automation Conference*, June 2008, pp. 554-559.

[55] Schamberger, P., Zhonghai Lu; Xianyang Jiang; Meikang Qiu, Modeling and power evaluation of on-chip router components in spintronics, *Proceedings of the Sixth ACM/IEEE International Symposium on Networks-on-Chip*, May 2012, pp. 51-58.

[56] Raoux, S.; Burr, G. W.; Breitwisch, M. J.; Rettner, C. T.; Chen, Y.-C.; Shelby, R. M.; Salinga, M.; Krebs, D.; Chen, S.-H.; Lung, H.-L. & Lam, C. H., Phase-change random access memory: A scalable technology, *IBM Journal of Research and Development*, volume 52, issue 4.5, pp. 465-479, 2008.

[57] Dong, X.; Muralimanohar, N.; Jouppi, N.; Kaufmann, R. & Xie, Y., Leveraging 3D PCRAM technologies to reduce checkpoint overhead for future exascale systems, *Proceedings of the Conference on High Performance Computing Networking, Storage and Anal*ysis, Nov. 2009, pp. 1-12.

[58] Raoux, Simone, Phase change materials, *Annual Review of Materials Research*, volume 39, pp. 25-48, 2009

[59] Lee, B.; Ipek, E.; Mutlu, O. & Burger, D., Architecting phase change memory as a scalable dram alternative, *Proceedings of the 36th annual international symposium on Computer architecture*, June 2009, pp. 2-13.

[60] International technology roadmap for semiconductors, *ITRS 2007* , Semiconductor Industry Association, http://www.itrs.net/Links/2007ITRS/Home2007.htm.

[61] Lee, B.; Zhou, P.; Yang, J.; Zhang, Y.; Zhao, B.; Ipek, E.; Mutlu, O. & Burger, D., Phase-change technology and the future of main memory, *IEEE Micro*, volume 30, issue 1, pp. 143, 2010.

[62] Qureshi, M.; Srinivasan, V. & Rivers, J., Scalable high performance main memory system using phase-change memory technology, *Proceedings of the 36th annual international symposium on Computer architecture*, June 2009, pp. 24-33.

[63] Dong, X.; Jouppi, N. & Xie, Y, PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM, *Proceedings of the 2009 International Conference on Computer-Aided Design*, Nov. 2009, pp. 269-275.

[64] Joo, Y.; Niu, D.; Dong, X.; Sun, G.; Chang, N. & Xie, Y., Energy-and endurance-aware design of phase change memory caches, *Proceedings of the Conference on Design, Automation and Test in Europe*, March 2010, pp. 136-141.

[65] Xie, Y., Modeling, architecture, and applications for emerging memory technologies, *IEEE Design Test of Computers*, volume 28, issue 1, pp. 44-51, 2011.

[66] Wong, H.-S.; Lee, H.-Y.; Yu, S.; Chen, Y.-S.; Wu, Y.; Chen, P.-S.; Lee, B.; Chen, F. & Tsai, M.-J., Metal-oxide RRAM, *Proceedings of the IEEE*, volume 100, issue 6, pp. 1951-1970, 2012.

[67] Wang, C.-H.; Tsai, Y.-H.; Lin, K.-C.; Chang, M.-F.; King, Y.-C.; Lin, C. J.; Sheu, S.-S.; Chen, Y.-S.; Lee, H.-Y.; Chen, F. & Tsai, M.-J., Three-dimensional 4F^2 ReRAM with vertical BJT driver by CMOS logic compatible process, *IEEE Transactions on Electron Devices*, volume 58, issue 8, pp. 2466-2472, 2011.

[68] Xu, C.; Dong, X.; Jouppi, N. P. & Xie, Y., Design implications of memristor-based RRAM cross-point structures, *Design, Automation & Test in Europe Conference & Exhibition*, March 2011, pp. 1-6.

[69] Lee, B. & Wong, H.-S., Fabrication and characterization of nanoscale NiO resistance change memory RRAM cells with confined conduction paths, *IEEE Transactions on Electron Devices*, volume 58, issue 10, pp. 3270-3275, 2011.

[70] Russo, U.; Ielmini, D.; Cagli, C. & Lacaita, A., Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices, *IEEE Transactions on Electron Devices*, volume 56, issue 2, pp. 186-192, Feb. 2009.

[71] Hosoi, Y.; Tamai, Y.; Ohnishi, T.; Ishihara, K.; Shibuya, T.; Inoue, Y.; Yamazaki, S.; Nakano, T.; Ohnishi, S.; Awaya, N.; Inoue, H.; Shima, H.; Akinaga, H.; Takagi, H.; Akoh, H. & Tokura, Y., High speed unipolar switching resistance RAM (RRAM) technology, *International Electron Devices Meeting*, Dec. 2006, pp. 1-4.

[72] K. M. Kim, B. J. Choi, and C. S. Hwang, Localized switching mechanism in resistive switching of atomic-layer-deposited TiO_2 thin films, *Applied Physics Letters*, volume 90, issue 24, June 2007.

[73] Lewis, D. & Lee, H.-H., Architectural evaluation of 3D stacked RRAM caches, *IEEE International Conference on 3D System Integration*, Sept. 2009, pp. 1-4.

[74] Chang, M.-F.; Chiu, P.-F.; Wu, W.-C.; Chuang, C.-H. & Sheu, S.-S., Challenges and trends in low-power 3D die-stacked IC designs using RAM, memristor logic, and resistive memory (ReRAM), *IEEE 9th International Conference on ASIC*, Oct. 2011, pp. 299-302.

[75] Chen, Y.-C.; Wang, W.; Li, H. & Zhang, W., Non-volatile 3D stacking RRAM-based FPGA, *22nd International Conference on Field Programmable Logic and Applications*, Aug. 2012, pp. 367-372.

[76] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach, 5th edition*, Morgan Kaufmann, 2012.

[77] B. Jacob, S. W. Ng and D. T. Wang, *Memory systems: cache, DRAM, disk*, Elsevier, 2008.

[78] Kim, C.; Burger, D. & Keckler, S., An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches, *Proceedings of the 10th international conference on architectural support for programming languages and operating systems*, Dec. 2002, pp. 211-222.

[79] Huh, J.; Kim, C.; Shafi, H.; Zhang, L.; Burger, D. & Keckler, S, A NUCA substrate for flexible CMP cache sharing, *IEEE Transactions on Parallel and Distributed Systems*, volume 18, issue 8, pp. 1028-1040, 2007.

[80] Culler D E, Singh J P, Gupta A., *Parallel computer architecture: a hardware/software approach*, Morgan Kaufmann Pub, 1999.

[81] Xiaowen Chen, Zhonghai Lu, Axel Jantsch, and Shuming Chen, Supporting distributed shared memory on multi-core network-on-chips using a dual microcoded controller, *Proceedings of the conference for Design Automation and Test in Europe*, March 2010, pp. 39-44.

[82] Black, B.; Annavaram, M.; Brekelbaum, N.; DeVale, J.; Jiang, L.; Loh, G. H.; McCaule, D.; Morrow, P.; Nelson, D. W.; Pantuso, D.; Reed, P.; Rupley, J.; Shankar, S.; Shen, J. & Webb, C., Die stacking (3D) microarchitecture, *39th Annual IEEE/ACM International Symposium on Microarchitecture*, Dec. 2006, pp. 469-479.

[83] Xu, T.; Liljeberg, P. & Tenhunen, H., Exploring DRAM last level cache for 3D network-on-chip Architecture, *Advanced Materials Research* , volume 403, pp. 4009-4018, 2010.

[84] Dong, X.; Wu, X.; Xie, Y.; Chen, Y. & Li, H., Stacking magnetic random access memory atop microprocessors: an architecture-level evaluation, *IET Computers & Digital Techniques*, volume 5, issue 3, pp. 213-220, 2011.

[85] Sun, G.; Dong, X.; Xie, Y.; Li, J. & Chen, Y., A novel architecture of the 3D stacked MRAM L2 cache for CMPs, *IEEE 15th International Symposium on High Performance Computer Architecture*, Feb. 2009, pp. 239-249.

[86] Madan, N.; Zhao, L.; Muralimanohar, N.; Udipi, A.; Balasubramonian, R.; Iyer, R.; Makineni, S. & Newell, D., Optimizing communication and capacity in a 3D stacked reconfigurable cache hierarchy, *IEEE 15th International Symposium on High Performance Computer Architecture*, Feb. 2009, pp. 262-274.

[87] Inoue, K.; Hashiguchi, S.; Ueno, S.; Fukumoto, N. & Murakami, K., 3D implemented SRAM/DRAM hybrid cache architecture for high-performance and low power consumption, *IEEE 54th International Midwest Symposium on Circuits and Systems*, Aug. 2011, pp. 1-4.

[88] Wu, X.; Li, J.; Zhang, L.; Speight, E.; Rajamony, R. & Xie, Y, Hybrid cache architecture with disparate memory technologies, *Proceedings of the 36th annual international symposium on computer architecture*, June 2009, pp. 34-45.

[89] Wu, X.; Li, J.; Zhang, L.; Speight, E. & Xie, Y., Power and performance of read-write aware Hybrid Caches with non-volatile memories, *Design, Automation & Test in Europe Conference & Exhibition*, April 2009, pp. 737-742.

[90] Sharifi, A. & Kandemir, M., Automatic feedback control of shared hybrid caches in 3D chip multiprocessors, *19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing*, Feb. 2011, pp. 393-400.

[91] Li, F.; Nicopoulos, C.; Richardson, T.; Xie, Y.; Narayanan, V. & Kandemir, M., Design and management of 3D chip multiprocessors using network-in-memory, *Proceedings of the 33rd annual international symposium on Computer Architecture*, 2006, pp. 130-141.

[92] Jung, J.; Kang, K. & Kyung, C.-M., Design and management of 3D-stacked NUCA cache for chip multiprocessors, *Proceedings of the 21st edition of the great lakes symposium on Great lakes symposium on VLSI*, ACM, 2011, pp. 91-96.

[93] Jung, J.; Kim, S. & Kyung, C.-M., Latency-aware utility-based NUCA cache partitioning in 3D-stacked multi-processor systems, *18th IEEE/IFIP VLSI System on Chip Conference*, Sept. 2010, pp. 125-130.

[94] Guthmuller, E.; Miro-Panades, I. & Greiner, A., Adaptive stackable 3D cache architecture for manycores, *IEEE Computer Society Annual Symposium on VLSI*, Aug. 2012, pp. 39-44.

[95] Loh, G., Extending the effectiveness of 3D-stacked DRAM caches with an adaptive multi-queue policy, *42nd Annual IEEE/ACM International Symposium on Microarchitecture*, Dec. 2009, pp. 201-212.

[96] Zhao, B.; Du, Y.; Yang, J. & Zhang, Y., Process variation aware non-uniform cache management in 3D die stacked multicore processor, *IEEE Transactions on Computers*, volume PP, issue 99, pp. 1, 2012.

[97] Mishra, A. K.; Dong, X.; Sun, G.; Xie, Y.; Vijaykrishnan, N. & Das, C. R., Architecting on-chip interconnects for stacked 3D STT-RAM caches in CMPs, *Proceeding of the 38th annual international symposium on Computer architecture*, June 2011, pp. 69-80.

[98] Meng, J.; Kawakami, K. & Coskun, A., Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints, *49th ACM/EDAC/IEEE Design Automation Conference*, June 2012, pp. 648-655.

[99] Meng, J. & Coskun, A., Analysis and runtime management of 3D systems with stacked DRAM for boosting energy efficiency, *Design, Automation Test in Europe Conference Exhibition*, March 2012, pp. 611-616.

[100] Loh, G., 3d-stacked memory architectures for multi-core processors, *Proceedings of the 35th Annual International Symposium on Computer Architecture*, 2008, pp. 453-464.

[101] Jacob, P.; Zia, A.; Erdogan, O.; Belemjian, P.; Kim, J.-W.; Chu, M.; Kraft, R.; McDonald, J. & Bernstein, K., Mitigating memory wall effects in high-clock-rate and multicore CMOS 3-D processor memory stacks, *Proceedings of the IEEE*, volume 97, issue 1, pp. 108 -122, 2009.

[102] Woo, D. H.; Seong, N. H.; Lewis, D. & Lee, H.-H., An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth, *IEEE 16th International Symposium on High Performance Computer Architecture*, Jan. 2010, pp. 1-12.

[103] Zhang, W. & Li, T., Exploring phase change memory and 3D die-stacking for power/thermal friendly, fast and durable memory architectures, *18th International Conference on Parallel Architectures and Compilation Techniques*, Sept. 2009, pp. 101-112.

[104] Pavlidis, V. & Friedman, E., Interconnects in the third dimension: design challenges for 3D ICs, *44th ACM/IEEE Design Automation Conference*, June 2007, pp. 562-567.

[105] Xie, J.; Zhao, J.; Dong, X. & Xie, Y., Architectural benefits and design challenges for three-dimensional integrated circuits, *IEEE Asia Pacific Conference on Circuits and Systems*, Dec. 2010, pp. 540-543.

[106] Oprins, H.; Cherman, V.; Vandevelde, B.; Van der Plas, G.; Marchal, P. & Beyne, E., Numerical and experimental characterization of the thermal behavior of a packaged DRAM-on-logic stack, *IEEE 62nd Electronic Components and Technology Conference*, May 2012, pp. 1081-1088.

[107] Lee, S.; Kang, K. & Kyung, C.-M., Temperature- and bus traffic- aware data placement in 3D-stacked cache, *18th IEEE/IFIP VLSI System on Chip Conference*, Sept. 2010, pp. 352-357.

[108] Chatterjee, S.; Cho, M.; Rao, R. & Mukhopadhyay, S., Impact of die-to-die thermal coupling on the electrical characteristics of 3D stacked SRAM cache, *28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium*, March 2012, pp. 14-19.

[109] Velenis, D.; Stucchi, M.; Marinissen, E.; Swinnen, B. & Beyne, E., Impact of 3D design choices on manufacturing cost, *IEEE International Conference on 3D System Integration*, Sept. 2009, pp. 1-5.

[110] Xu, T.; Liljeberg, P. & Tenhunen, H., Optimal number and placement of through silicon vias in 3D network-on-chip, *IEEE 14th International Symposium on Design and Diagnostics of Electronic Circuits Systems*, April 2011, pp. 105-110.

[111] Chunbo Zhang; Leijun Li, Characterization and design of through-silicon via arrays in three-dimensional ICs based on thermomechanical modeling, *IEEE Transactions on Electron Devices*, volume 58, issue 2, pp. 279-287, Feb. 2011.

[112] Chen, Z.; Song, X. & Liu, S., Thermo-mechanical characterization of copper filled and polymer filled tsvs considering nonlinear material behaviors, *59th Electronic Components and Technology Conference*, May 2009, pp. 1374-1380.

[113] *Wide I/O single data rate (JEDEC Standard JESD229)*, JEDEC Solid State Technology Association, December 2011, http://www.jedec.org.

[114] *Low power double data rate 3 (JEDEC Standard JESD209-3)*, JEDEC Solid State Technology Association, May 2012, http://www.jedec.org.

[115] Kim, J.-S.; Oh, C. S.; Lee, H.; Lee, D.; Hwang, H.-R.; Hwang, S.; Na, B.; Moon, J.; Kim, J.-G.; Park, H.; Ryu, J.-W.; Park, K.; Kang, S.-K.; Kim, S.-Y.; Kim, H.; Bang, J.-M.; Cho, H.; Jang, M.; Han, C.; Lee, J.-B.; Kyung, K.; Choi, J.-S. & Jun, Y.-H., A 1.2V 12.8GB/s 2Gb mobile Wide-I/O DRAM with 4×128 I/Os using TSV-based stacking, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2011, pp. 496-498.

[116] Takaya, S.; Nagata, M.; Sakai, A.; Kariya, T.; Uchiyama, S.; Kobayashi, H. & Ikeda, H., A 100GB/s wide I/O with 4096b TSVs through an active silicon interposer with in-place waveform capturing, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2013, pp. 434-435.

[117] Bae, Y.-C.; Park, J.-Y.; Rhee, S. J.; Ko, S. B.; Jeong, Y.; Noh, K.-S.; Son, Y.; Youn, J.; Chu, Y.; Cho, H.; Kim, M.; Yim, D.; Kim, H.-C.; Jung, S.-H.; Choi, H.-I.; Yim, S.; Lee, J.-B.; Choi, J.-S. & Oh, K., A 1.2V 30nm 1.6Gb/s/pin 4Gb LPDDR3 SDRAM with input skew calibration and enhanced control scheme, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2012, pp. 44-46.

[118] Lenoski, D.; Laudon, J.; Gharachorloo, K.; Weber, W.-D.; Gupta, A.; Hennessy, J.; Horowitz, M. & Lam, M., The Stanford Dash multiprocessor, *Computer*, volume 25, issue 3, pp. 63-79, 1992.

[119] Waingold, E.; Taylor, M.; Srikrishna, D.; Sarkar, V.; Lee, W.; Lee, V.; Kim, J.; Frank, M.; Finch, P.; Barua, R.; Babb, J.; Amarasinghe, S. & Agarwal, A., Baring it all to software: Raw machines, *Computer*, volume 30, issue 9, pp. 86-93, 1997.

[120] Laudon, J. & Lenoski, D., The SGI Origin: a ccNUMA highly scalable server, *The 24th Annual International Symposium on Computer Architecture*, June 1997, pp. 241-251.

[121] Puttaswamy, K. & Loh, G., 3D-integrated SRAM components for high-performance microprocessors, *IEEE Transactions on Computers*, volume 58, issue 10, pp. 1369-1381, 2009.

[122] Chen, X. & Davis, W., Delay analysis and design exploration for 3D SRAM, *IEEE International Conference on 3D System Integration*, Sept. 2009, pp. 1-4.

[123] Hsu, C.-L. & Wu, C.-F., High-performance 3D-SRAM architecture design, *IEEE Asia Pacific Conference on Circuits and Systems*, Dec. 2010, pp. 907-910.

[124] Nho, H.; Horowitz, M. & Wong, S., A high-speed, low-power 3D-SRAM architecture, *IEEE Custom Integrated Circuits Conference*, Sept. 2008, pp. 201-204.

[125] Thomas, O.; Vinet, M.; Rozeau, O.; Batude, P. & Valentian, A., Compact 6T SRAM cell with robust read/write stabilizing design in 45nm monolithic 3D IC technology, *IEEE International Conference on IC Design and Technology*, May 2009, pp. 195-198.

[126] Liu, C. & Lim, S. K., Ultra-high density 3D SRAM cell designs for monolithic 3D integration, *IEEE International Interconnect Technology Conference*, June 2012, pp. 1-3.

[127] Weis, C.; Wehn, N.; Igor, L. & Benini, L., Design space exploration for 3D-stacked DRAMs, *Design, Automation Test in Europe Conference Exhibition*, March 2011, pp. 1-6.

[128] Sekiguchi, T.; Ono, K.; Kotabe, A. & Yanagawa, Y., 1-Tbyte/s 1-Gbit DRAM architecture using 3-D interconnect for high-throughput computing, *IEEE Journal of Solid-State Circuits*, volume 46, issue 4, pp. 828-837, 2011.

[129] Vignon, A.; Cosemans, S.; Dehaene, W.; Marchal, P. & Facchini, M., A novel DRAM architecture as a low leakage alternative for SRAM caches in a 3D interconnect context, *Design, Automation & Test in Europe Conference & Exhibition*, April 2009, pp. 929-933.

[130] Sun, H.; Liu, J.; Anigundi, R.; Zheng, N.; Lu, J.-Q.; Rose, K. & Zhang, T., 3D DRAM design and application to 3D multicore systems, *IEEE Design Test of Computers*, volume 26, issue 5, pp. 36-47, 2009.

[131] Jun, M.; Kim, M.-J. & Chung, E.-Y., Asymmetric DRAM synthesis for heterogeneous chip multiprocessors in 3D-stacked architecture, *IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2012, pp. 73-80.

[132] Chung, E. S.; Milder, P. A.; Hoe, J. C. & Mai, K., Single-chip heterogeneous computing: does the future include custom logic, FPGAs, and GPGPUs?, *the 43rd Annual IEEE/ACM International Symposium on Microarchitecture*, Dec. 2010, pp. 225-236.

[133] Saripalli, V.; Sun, G.; Mishra, A.; Xie, Y.; Datta, S. & Narayanan, V., Exploiting heterogeneity for energy efficiency in chip multiprocessors, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, volume 1, issue 2, pp. 109-119, 2011.

[134] Watkins, M. A. & Albonesi, D. H., ReMAP: A reconfigurable heterogeneous multicore architecture, *43rd Annual IEEE/ACM International Symposium on Microarchitecture*, Dec. 2010, pp. 497-508.

[135] Yuffe, M.; Knoll, E.; Mehalel, M.; Shor, J. & Kurts, T., A fully integrated multi-CPU, GPU and memory controller 32nm processor, *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2011, pp. 264-266.

[136] Heterogeneous System Architecture (HSA) Foundation, http://hsafoundation.com/.

[137] Al Maashri, A.; Sun, G.; Dong, X.; Narayanan, V. & Xie, Y., 3D GPU architecture using cache stacking: performance, cost, power and thermal analysis, *IEEE International Conference on Computer Design*, Oct. 2009, pp. 254-259.

[138] Facchini, M.; Carlson, T.; Vignon, A.; Palkovic, M.; Catthoor, F.; Dehaene, W.; Benini, L. & Marchal, P., System-level power/performance evaluation of 3D stacked DRAMs for mobile applications, *Proceedings of the Conference on Design, Automation and Test in Europe*, April 2009, pp. 923-928.

[139] Chang, M.-F. & Chen, P.-C., Embedded non-volatile memory circuit design technologies for mobile low-voltage SoC and 3D-IC, *10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, Nov. 2010, pp. 13-16.

[140] Chang, D.; Kim, N. & Schulte, M., Analyzing the performance and energy impact of 3D memory integration on embedded DSPs, *International Conference on Embedded Computer Systems*, July 2011, pp. 303-310.

[141] Zhang, T.; Hsu, P.-Y.; Lo, W.-H.; Tseng, S.-Y.; Wu, Y.-T.; Liu, C.-N.; Yeh, J.-C.; Hwang, T. & Xie, Y., Leveraging on-chip DRAM stacking in an embedded 3D multi-core DSP system, *IEEE 54th International Midwest Symposium on Circuits and Systems*, Aug. 2011, pp. 1-4.

[142] Clermidy, F.; Darve, F.; Dutoit, D.; Lafi, W. & Vivet, P., 3D embedded multi-core: some perspectives, *Design, Automation Test in Europe Conference Exhibition*, March 2011, pp. 1-6.

Yuang Zhang received the B.Sc. degree in communication engineering from Shandong University, Jinan, China, in 2001. He is currently a Ph.D. candidate in Micro Electronics and Solid State Electronics department, Nanjing University. His main research interests focus on the architecture of network-on-chip (NoC), memory architecture and cache coherence schemes both for 2D and 3D NoCs.

Li Li received the B.Sc. degree and Ph.D. degree from Hefei University of Technology, Hefei, China, in 1996 and 2002 respectively. She has been a professor of VLSI Design Institute, School of Electronic Science & Engineering, Nanjing University since 2010. Her current research interests include the area of VLSI design for digital communication systems, multi-processor system-on-a-chip (MPSoC) architecture design methodology and SoC-IP Design.

Zhonghai Lu received the B.Sc. degree in radio and electronics from Beijing Normal University, China, in 1989, the M.Sc. degree in system-on-chip design, and the Ph.D. degree in electronic and computer system design from KTH-Royal Institute of Technology, Stockholm, Sweden, in 2002 and 2007, respectively. From 1989 to 2000, he worked extensively as an engineer in the area of electronic and embedded systems. Dr. Zhonghai Lu is currently an Associate Professor with the Department of Electronic Systems, School for Information and Communication Technology, KTH. His research interests include interconnection network, performance analysis, and design automation. He has published over 100 peer-reviewed papers in these areas.

Axel Jantsch received the M.S. degree and Ph.D. degree from Vienna University of Technology, Vienna, Austria, in 1988 and 1992 respectively. He has been a professor of Department of Electronic Systems, School for Information and Communication Technology，Royal Institute of Technology since 2002. His current research interests include embedded systems and dependable systems, modeling concepts and languages, systems on chip architecture, design methodology and tools.

Minglun Gao received the B.Sc. degree from Tsinghua University, Beijing, China, in 1968 and the Ph.D. degree from University of Dayton, Dayton, USA, in 1991 respectively. He has been a professor of VLSI Design Institute, Nanjing University since 2002. His research interests include VLSI design, network-on-chip architecture design and SoC-IP Design.

Hongbing Pan received the B.Sc. degree and Ph.D. degree from Nanjing University, Nanjing, China, in 1994 and 2005 respectively. He has been a professor of VLSI Design Institute, School of Electronic Science & Engineering, Nanjing University since 2012. His current research interests include the area of multi-processor system-on-a-chip (MPSoC) design methodology and CMOS sensor design.

Feng Han received the B.Sc. degree and the M.Sc. degrees from Nanjing University, in 2010, and 2013 respectively. He is current pursuing the Ph.D. degree in Nanjing University. His research focuses on storage architecture of 3D NoC and multi-core DSP chip design.