Trends of Terascale Computing Chips in the Next Ten Years

Zhonghai Lu * and Axel Jantsch

Abstract – *Moore's law steadily continues though facing a number of challenges. This paper identifies ongoing and desirable trends to exploit the technology capacity and further Moore’s law for terascale on-chip computing architectures in the next ten years. Four foreseeable trends are: from single core to many cores, from bus-based to network-based interconnect, from centralized memory to distributed memory, and from 2D integration to 3D integration. We motivate these trends and show that the number of design choices for computing chips is increasing rapidly, leading to an exploding design space with uncountable opportunities for the innovative architect. Moreover, we envision that the multi-core Network-on-Chip will become an infrastructure backbone and accumulate many other infrastructural functions such as memory, power and resource management, testing and diagnostic services.*

Index Terms — Computer Architecture, Network-on-Chip, Multi-core System, Distributed Memory, 3D Integration

I. INTRODUCTION

Moore’s law has sustained over the last four decades since 1968. Today we have entered the billion transistor era, aiming for terascale performance. To exploit the technology capacity, processors, computers and System-on-Chips (SoCs) have experienced extremely exciting developments in continuously increasing performance for known applications and realizing more and more complex applications. Also, the performance cost ratio has increased steadily with more optimized power consumption. However, fundamental limits and bottlenecks have puzzled processor, computer and SoC architects. With technology and voltage scaling, we are more and more approaching physical limits in transistor size, transistor voltage and switching threshold.

The limit of transistor size: Today we have most of advanced designs fabricated in 45 nm or 32 nm. However, such transistor scaling is difficult to be sustainable in the long term as we are approaching more and more to the physical limit. The ITRS scaling targets to continue the historic 17%/year improvement in transistor delay, i.e., in $CV_{dd} / I_{d,sat}$, where C is the MOSFET capacitance including parasitics, V_{dd} the power supply voltage and $I_{d,sat}$ the drain saturation current. The shorter the delay, the better the MOSFET performance. Significant scaling difficulties have already been encountered, and are expected to worsen in the coming years as the gate length is scaled to well below 30 nm. 2007 ITRS projects the MPU high performance physical gate length to be 10 nm in 2015. However, manufacturable solutions for those below 20 nm are not known, and innovations are greatly desired [6].

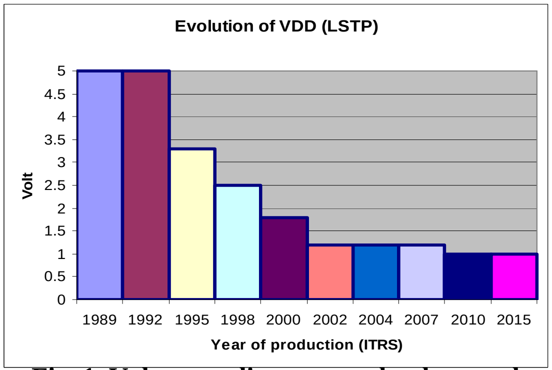

Fig. 1. Voltage scaling over technology nodes

The limit of transistor voltage and switching threshold: Low power designs have relied on reducing the supply voltage and lowering the switching threshold [19][20]. With transistor scaling, V_{dd} is also scaled down. However, the threshold voltage, V_t , cannot be scaled down significantly, since the source/drain sub-threshold leakage current, $I_{sd,leak}$, increases sharply with decreased V_t , and it is important to keep $I_{sd,leak}$ within tolerable limits [6]. As depicted in Fig. 1, in the ten years from 1992 (500 nm nodes) to 2002 (120 nm nodes), V_{dd} has reduced from 5 to 1.2 volts. Afterwards, 1.2, 1.1 volts have become another plateau. In older technologies (250 nm and above), the leakage power was marginal with respect to the switching power, and minimizing switching power had priority. In deep submicron technologies, the leakage power becomes critical. It accounts for around 5-10% of power budget at 180 nm; this grows to 20-25% at 130 nm and to 35-60% at 65 nm [6].

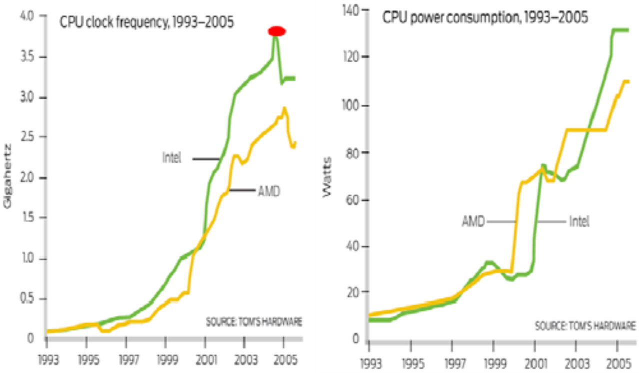

Fig. 2. Single core frequency and power

While the limits of transistor size and switching threshold

[1] This work was supported in part by the EU Commission through the FP7 projects Elite, ArtistDesign and Mosart under contract No. 215030, No. 214373 and No. 215244, respectively.

Z. Lu and A. Jantsch are with Dept. of Electronics, Computer and Software Systems, KTH – The Royal Institute of Technology, Stockholm, Sweden (e-mail: zhonghai@kth.se, axel@kth.se).

* Corresponding author.

are approached, single core frequency reaches limit. As illustrated in Fig. 2a, single core speed has greatly enhanced over ten years from about 200 MHz in 1993 to about 4 GHz in 2004 for the Intel processors and 3.0 GHz for the AMD processors [1]. Note that over-clocking processors with higher frequency is possible, but requires special cooling method and is not sustainable. Afterwards, the processor speed does not increase anymore. Historically the three factors for performance enhancement are ILP, gates per clock and process technology. However, the first two factors have essentially reached their limit [4]. The process technology has been improved but it becomes wire and power limited rather than device limited. Clocking processors with high frequency creates huge problem in power consumption and heat dissipation in addition to clocking synchronization, signal and power integrity. Processors have never been fast enough, but already too hot. As shown in Fig. 2b, we can see that the power consumption is also peaked for the processors. The power consumption for the highest speed Intel processor reaches 130 Watt. Technically it is mainly power and heat that shows stop-sign to increasing single processors' speed.

Facing these limits, how shall we further increase the performance of computing chips? Apparently, there are numerous computation intensive applications in consumer electronics, networking and communication, simulation, and ubiquitous intelligence etc., which require huge processing power. Hence further increasing performance under power and cost constraints has been the key issue for high-end computing chips both in general and specific domains. In this paper, we systematically identify and discuss foreseeable trends in *computation*, *communication*, *storage*, and *process technology* for computing chips in the next ten years. Four trends are mapped to the four aspects: from *multi-core to many cores (100-1000)*, from *bus to network*, from *centralized memory to distributed memory*, from *2D integration to 3D integration*. The four trends are discussed separately but they are by no means isolated. On the contrary, they should be viewed as an integrated solution for terascale computing. A potential platform which allows to effectively integrate these four trends is Multi-core Network-on-Chip (McNoC), which is desirable to feature a 3D architecture and distributed memory.

In the remainder of the paper, we elaborate our discussions on the four trends, analyze the underlying reasons, and discuss their pros and cons, obstacles and promises, status and challenges. As an integration platform, we also briefly discuss McNoCs. Finally we give concluding remarks.

II. Architectural Trends

We discuss the four trends one by one, starting from single- to multi- many cores, from bus to network interconnect, then from centralized to distributed memory, and finally from 2D to 3D integration.

A. From Single/Multi Core to Many/Hundred Cores

Multi- and many core processors, computers and SoCs are already the practice of today's chips. In the general-purpose domain, the number of cores is in the range of 2 to 4 cores for desktop PCs, 2 to 16 cores for servers. In the embedded domains, the number of cores ranges from 4 to 64 cores (The Kilocore of Rapport has now 256 cores but small 8-bit cores).

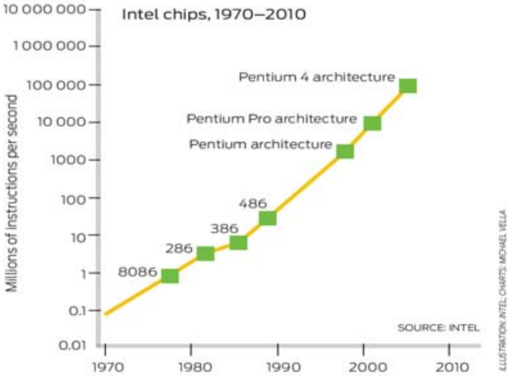

Fig. 3. The Intel processors' performance

While it is infeasible to squeeze the speed of single cores, the solution to further performance is to use more cores with reasonable speed on a single chip. After all, the ultimate application performance is measured not in frequency but workload completed per unit time, for example, millions instructions per second (MIPS) or instructions per cycle (IPC). As shown in Fig. 3, the Intel multi-core architecture processors have continually increased MIPS even though the single core speed is reduced [1]. Using more cores increases parallelism, which is fundamentally more power efficient than a sequential architecture [19]. For instance, splitting a computation in two (parallel processing) and running it as two parallel independent tasks has the potential to cut the power in half without slowing the computation. With many cores on a single chip, we can partition the entire chip capacity into many modular synchronous regions with explicit parallelism. Such a divide-and-conquer approach is beneficial to alleviate a number of concerns in synchronicity, global wire delay, reliability etc., offering high computation capability and communication bandwidth with low power.

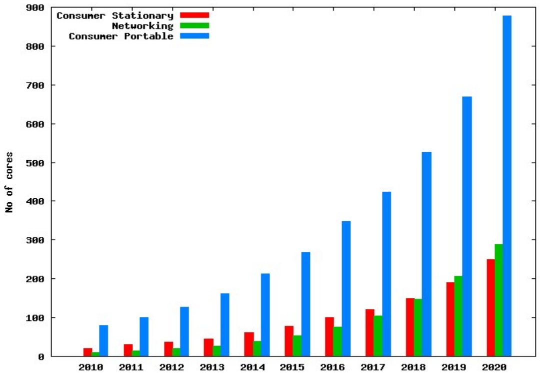

Fig. 4. Number of cores on chips

As predicted by ITRS and shown in Fig. 4 [6], the number of cores will increase from 25 to 300 for stationary computers, from 10 to 500 for networking applications, and from 64 to 900 for consumer portable devices in the next ten years. Even thousand core chips have been discussed in [5] from a technology perspective.

B. *From Bus Interconnect to Network Interconnect*

Global interconnect has become a bottleneck when the system size increases due to the following reasons.

Global wires do not scale [7][22]: Technology scaling does not treat wire delay and gate delay equally. While gate delay (transistor switching time) has been getting dramatically smaller in proportion to the gate length, wires have slowed down. Wire delay is dominating and wires across the chips take a few cycles. At and beyond 130 nm, multiple or even tens of cycles are required to transmit a signal across its diameter in a top-level metal wire depending on the clock rate assuming best transmission conditions such as very low-permittivity dielectrics, resistivity of pure copper, high aspect ratio (ratio of wire height to wire width) wires and optimally placed repeaters [7]. This means that the chip is becoming more communication-bound rather than capacity-bound.

Global clocking does not scale [8][23]: Traditionally IC designs have followed the globally synchronous design style where a global clock tree is distributed on the chip, and logic blocks function synchronously. However, this style is unlikely to survive very long for large chips. A clock tree is consuming larger portions of power and area budget, and clock skew is claiming an ever larger portion of the total cycle time [23]. Global synchronization is impossible [8].

Global buses do not scale well in terms of bandwidth, clocking frequency and power [9][24]. First, a bus system has very limited concurrent communication capability since only one device can drive a bus segment at a time. Current SoCs integrate only several processors and, rarely, more than 10 bus masters. Second, as the number of clients grows, the intrinsic resistance and capacitance of the bus also increase. This means that the bus speed is inherently difficult to scale up. Third, a bus is inefficient in energy since every data transfer is broadcast. The entire bus wire has to be switched on and off. This means that the data must reach each receiver at great energy cost. Although improvements such as multiple bridged and hierarchal buses, split-transaction protocols and advanced arbitration schemes for buses have been proposed, these incremental techniques cannot overcome the fundamental problems.

We may also justify the necessity of revolutionizing the bus to network interconnect from another angle. As discussed in Section II.A and Section II.C, respectively, there will be more cores and memories distributed on chips. Such parallel operations of cores and memories require concurrent pipelined processing of communications rather than serializing all the communications. Viewing from this angle, a communication network is an inevitable solution for the global interconnect, resulting in a so called network-on-chip (NoC) [24], which is a hot topic today.

NoC research and practices started from around year 2000, and so far have 10 years of history. State-of-the-art NoCs as general purpose platforms have a regular mesh topology with tens of nodes, for example, Tilera 64 (8x8 nodes) and Intel's teraflop research prototype (8x10 nodes). Irregular and specialized NoCs are developed for customized and application domain specific solutions, as for instance promoted by Arteris Inc [25].

C. *From Centralized Memory to Distributed Memory*

Memory is a first class citizen in a computing system. Its organization and size are crucial for performance, power and cost. As shown in Fig. 5, the memory content in SoCs has increased dramatically from 20% in 1999 to 83% in 2008 [2]. This trend will likely continue.

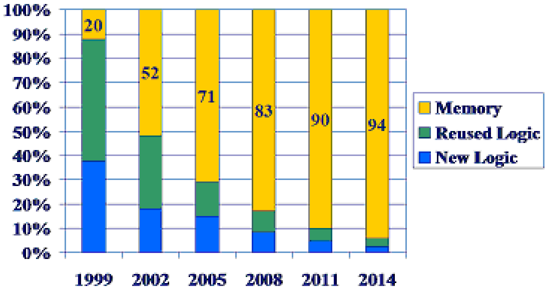

Fig. 5. Memory content in SoCs

Traditionally, memories are organized hierarchically and centrally. The hierarchical organization efficiently hides the communication latency and can greatly enhance the computation performance. However, the centralized and often monolithic organization of caches, on-chip memory and off-chip memory is not scalable. The need for large amount of storage results in the demand of extremely high bandwidth with low latency. This calls for a parallel organization of memory in order to enable concurrent accesses. For instance, as discussed in [17], caches are also becoming wire delay dominated under the submicron process technologies and cache access times to different lines are non-uniform. Given the same size of caches, performance can be improved by organizing caches into distributed, small cache banks. To de-centralize the single access point to memory, shared caches and on-chip memories are preferably organized in a distributed fashion. Due to technology advances, such as high density embedded memory (e.g. Z-RAM from Innovative Silicon) and 3D integration, high bandwidth, parallel access to memory is becoming feasible and preferable. For instance, G. Loh proposes a 3D-stacked memory architecture where each core has its own memory bank with significant performance gains [3].

Distributed memory presents a number of challenges. Providing architectural support for programming paradigms based on *shared variable* and *message passing* communication is a pressing challenge. There exists an urgent need to support distributed but shared memory (DSM), in order to re-use huge amount of shared variable legacy code. To increase productivity, reliability and reduce risk, reuse of proven legacy code is a must. From the programmer's point of view, the shared variable paradigm [18] is relatively easier, as it provides a single shared address space and transparent implicit communication and thus there is no need to worry about the destination, as required by message passing. In this regard, efficient and scalable mechanism of cache coherency and memory consistency must be developed. Nevertheless, message passing is inherently more scalable because of independent states and state management at each node.

Therefore, developing efficient distributed memory organization and mechanisms for message passing paradigm, particularly for large scale chips, will become essential.

D. From 2D Integration to 3D Integration

As the feature size of transistors shrinks to below 250 nm, wire delay becomes more and more significant with respect to gate delay [6]. Because of increasing interconnect delay with shrinking feature size, conventional two-dimensional (2D) IC integration technologies have become performance bottleneck, especially for 45 nm and beyond. Wire sizing and repeater insertion are commonly used techniques to deal with the problem, but they are inadequate for 32 nm beyond. Even with repeater insertion, the interconnect delay is more than one order of magnitude worse than gate delay [6]. Repeaters also consume much power and routing resources. Besides, as we are approaching the physical limitation, signal integrity, power integrity and dissipation, leakage power, clock distribution and yield issues are becoming increasingly intractable [21].

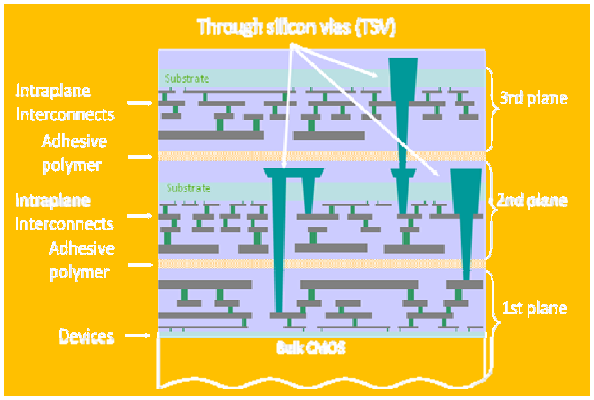

Fig. 6. TSV-based 3D integration

To tackle such problems, introducing another dimension (resulting in 3D integration) such as folding and wafer/die stacking technology becomes necessary in order to shorten topological distance, thus beneficial for both performance and power. Currently, a number of 3D integration technologies [10][11][12] emerge, such as Through-Silicon-Vias (TSVs) [11], thinned silicon and silicon-to-silicon fine pitch interconnections, wireless communication between 2D planes and 3D wafer wire-bonding technology etc. These technologies enable to stack multiple dies on a single chip, creating 3D Integrated Circuits (3D-ICs) and offering an opportunity to be the next performance growth engine [13]. Fig. 6 illustrates the cross-section of a 3D IC using TSVs to make connections between the three planes [26]. These technologies may also enable heterogeneous and new classes of complex applications with significantly improved performance, energy efficiency, product miniaturization, cost reduction, and modular design for improved time to market. Such technologies are currently available from a number of companies and labs such as IBM, IMEC, Honda, Tezzaron Semiconductor Corporation and MIT Lincoln Laboratory etc.

In making a choice between 2D and 3D technology, a key question is: should we continue to reduce the transistor feature size but keeping 2D integration, or should we go for 3D integration but keep the same process technology, or reduce the transistor size via 3D techniques? As analyzed above, shrinking feature size has complicated a number of issues in clocking, power distribution, leakage power, reliability, yield, and cost problems, let alone it is approaching to the physical limit. In the end, furthering traditional 2D integration may become technically feasible, but economically infeasible. Under such circumstance, 3D integration has attracted special attention. 3D integration technologies have been researched over the last 10 years [10][12]. The technologies deal with different levels of 3D integration, such as from 2D to 3D transistors, 2D to 3D circuits and 2D to 3D interconnects. 3D integration has improved form factor since it can provide smaller footprint and/or increased density, resulting in higher yield. It offers high performance because it has the possibility to use vertical connections to achieve higher bandwidth and lower latency. It has potentially lower power consumption because it has shorter wires and lower overall I/O count. It also allows heterogeneous integration since each plane in the 3D structure can use a different process technology for different applications such as sensors, RF, analogy planes etc.

We can imagine that cost is the deciding factor for such 3D technologies. We expect that 3D integration technologies have also cost containment potential. At the device level, they can address slow-down in the productivity gain resulting from scaling; At the die level, they can increase functionality and performance with more power reduction when compared to 2D SoCs; At the factory level, they provide optimized cost structure for each level in 3D stack; At the market level, they may enjoy short time to market for products.

III. MULTI-CORE NETWORK-ON-CHIP (MCNOC)

In the previous section, we have discussed the four trends. From the computation perspective, the trend is from single to many cores; From the interconnect perspective, the trend is from bus to network; From the storage perspective, the trend is from centralized to distributed memory organization; From the perspective of process technology, the trend is from 2D to 3D integration. Combining the four trends, we end up with a multi-core network-on-chip (McNoC) platform. Note that NoC has been used with two meanings in the literature. In the narrow sense, it refers only to the on-chip network; in the wide sense, it refers to the entire system featuring a network as the global interconnect. Here we use the wider meaning.

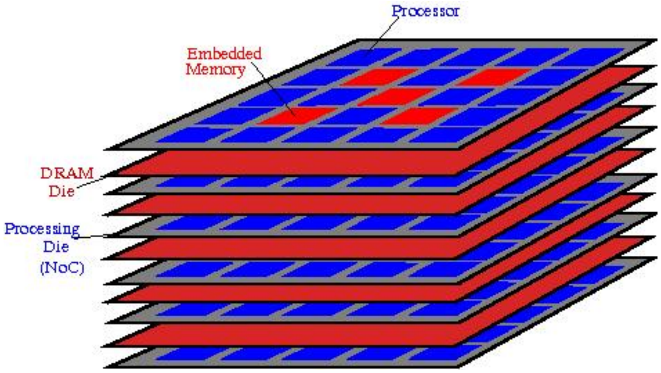

Fig. 7. A 3D McNoC

As a promising McNoC, 3D NoCs offer a great potential to integrate many computing cores, distributed memories in a 3D network topology with power and cost efficiency. Fig. 7 depicts an example of 3D NoC, where eleven planes are stacked and each plane could be a pure memory plane or a plane mixed with processors and memory blocks. 3D NoCs [14] scale 2D NoCs over the third dimension, overcoming the limited scalability of 2D NoCs by using short and fast vertical interconnects of 3D-ICs. Compared with 2D NoCs, 3D NoCs greatly reduce the network diameter and overall communication distance, thus improving communication performance and reducing power. Till today extensive results have shown that 3D networks improve 2D network performance in terms of delay and throughput [14][15][16]. Such studies provide huge promises in further continuing Moore's law, and exploiting the increased capacity for applications with exponentially growing performance characteristics.

IX. Concluding remark

On-chip system architectures have been experiencing exciting advancement towards high performance under tight cost and power constraints. In this paper, we have discussed four limits, four trends and one platform. The four limits are: transistor size limit, power supply and switching threshold limits, and single core frequency limit. In facing these limits, further enhancement of performance calls for innovative solutions. We have identified four ongoing and desirable trends, which are performance enhancement enablers. From the computation perspective, the trend is to go from single to many (100-1000) cores; From the interconnect perspective, the trend is to go from bus to network; From the storage perspective, the trend is to go from centralized to distributed memory; From the perspective of process technology, the trend is to go from 2D integration to 3D integration. Combining the four trends, 3D multicore NoCs would be a promising infrastructure backbone and accumulate many other infrastructural functions such as memory, power and resource management, testing and diagnostic services for high performance applications.

We conclude that profound changes in the architectures of processors, computers and SoCs are taking place. It is therefore important to identify and keep track of those changes. Innovations are required in all aspects of the system architecture while addressing scalability, manufacturability and reliability, IP reuse and productivity. The architectural choices will steadily increase until standard solutions for design methodology with cores as elementary elements, programming models, 3D platforms, and memory architectures have been established. Looking into the future, yesterday's supercomputer will become tomorrow's mobile device. Highly sophisticated applications, such as intelligent language translation, 3D gaming, automatic driving, cognitive radio, etc., will benefit from their unprecedented performance.

References

[1] Philip E. Ros, "Why CPU Frequency Stalled", *IEEE Spectrum*, 45(4), April 2008.

[2] E. J. Marinissen, B. Prince, D. Keltel-Schulz and Y. Zorian, "Challenges in embedded memory design and test", *Proceedings of Design, Automation and Test in Europe Conference (DATE'05)*, vol. 2, pp. 722-727, Mar. 2005.

[3] G. Loh, "3D-Stacked Memory Architectures for Multi-Core Processors", *Proceedings of the 35th ACM/IEEE International Symposium on Computer Architecture (ISCA´08)*, June 2008.

[4] M. Horowitz and W. Dally, "How Scaling will Change Processor Architecture", *IEEE International Solid-State Circuits Conference (ISSCC'04)*, Digest of Technical Papers, pp. 132-133, Feb. 2004.

[5] S. Borkar, "Thousand Core Chips: A Technology Perspective", *Proceedings of the 44th ACM/IEEE Design Automation Conference (DAC'07)*, pp. 746-749, Jun. 2007.

[6] Semiconductor Industry Association, International Technology Roadmap for Semiconductors (ITRS) documents and updates, http://www.itrs.net/Links/2007ITRS/Home2007.htm, 2007.

[7] V. Agarwal, M. S. Hrishikesh, S. W. Keckler, and D. Burger. "Clock rate versus IPC: the end of the road for conventional micro-architectures". In *Proc. of the 27th Annual International Symposium on Computer Architecture (ISCA'00)*, pages 248–259, 2000.

[8] A. Iyer and D. Marculescu. Power and performance evaluation of globally asynchronous locally synchronous processors. In *Proceedings of the 29th Annual International Symposium on Computer Architecture (ISCA'02)*, pages 158–168, 2002.

[9] T. Claasen. An industry perspective on current and future state-of-the-art in system-on-chip (SoC) technology. *Proceedings of the IEEE*, 94(6):1121–1137, June 2006.

[10] G. Philip, B. Christopher, and P. Ramm. Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits. *Wiley-VCH*, 2008.

[11] K. Snoeckx, E. Beyne, and B. Swinnen. Copper-nail TSV technology for 3D-stacked IC integration. *Solid State Technology*, 50(5), May 2007.

[12] J. U. Knickerbocker, editor. 3D Chip Technology, volume 52. *IBM Journal of Research and Development*, 2008.

[13] P. Emma and E. Kursun. Is 3D chip technology the next growth engine for performance improvement? *IBM Journal for Research and Development*, Nov. 2008.

[14] L. P. Carloni, P. Pande, and Y. Xie. *Networks-on-chip in emerging interconnect paradigms: Advantages and challenges*. In Proceedings of the 3rd ACM/IEEE International Symposium on Networks-on-Chip (NOCS'09), San Diego, CA, May 2009.

[15] B. Feero and P. Pande. Networks-on-chip in a three-dimensional environment: A performance evaluation. *IEEE Transactions on Computers*, May 2008.

[16] V. Pavlidis and E. Friedman. 3-D topologies for networks-on-chip. *IEEE Trans. on Very Large Scale Integration Systems*, 15(10), 2007.

[17] C. Kim, D. Burger and S. W. Keckler, An Adaptive, NonUniform Cache Structure for Wire Delay Dominated On Chip Caches, *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems* (ASPLOS'02), 2002.

[18] I. Park and S. W. Kim, "The distributed virtual shared-memory system based on the InfiniBand architecture", *Journal of Parallel and Distributed Computing*, vol. 65, no. 10, pp. 1271-1280, October 2005.

[19] T. Mudge. Power: A first-class architectural design constraint. *IEEE Computer*, 34(4):52–58, April 2001.

[20] V. Raghunathan, M. B. Srivastava, and R. K. Gupta. A survey of techniques for energy efficient on-chip communication. In *Proceedings of Design Automation Conference (DAC´03)*, June 2003.

[21] J. W. McPherson. Reliability challenges for 45nm and beyond. In *Proceedings of the 43rd Design Automation Conference (DAC'06)*, pages 176– 181, July 2006.

[22] R. Ho, K. Mai, and M. Horowitz. The future of wires. *Proceedings of the IEEE*, 89(4):490–504, April 2001.

[23] J. Öberg. In Networks on Chip, chapter "Clocking Strategies for Networks on Chip". *Kluwer Academic Publisher*, 2003.

[24] A. Jantsch and H. Tenhunen, editors. Networks on Chip. *Kluwer Academic Publisher*, 2003.

[25] Arteris Inc. www.arteris.com

[26] R. J. Gutmann *et al*., "Three-Dimensional (3D) ICs: A Technology Platform for Integrated Systems and Opportunities for New Polymeric Adhesives," *Proceedings of the Conference on Polymers and Adhesives in Microelectronics and Photonics*, pp. 173-180, October 2001