Interaction of Scaling Trends in Processor Architecture and Cooling

Wei Huang[1,4*], Mircea R. Stan[2], Sudhanva Gurumurthi[1], Robert J. Ribando[3], and Kevin Skadron[1]

[1] Department of Computer Science, University of Virginia, Charlottesville, VA

[2] Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA

[3] Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, VA

[4] IBM Austin Research Laboratory, Austin, TX

[*] Wei Huang, whuang@virginia.edu

Abstract

It is predicted that two important trends are likely to accompany traditional CMOS semiconductor technology scaling—chip multiprocessors and 3D integration. With the ever-increasing power consumption and the consequent difficulty in heat removal, it is important to consider the limits and implications of different cooling methods for the upcoming manycore and 3D era. In this paper, we consider both technology scaling and manycore architecture scaling trends in conjunction with conventional air cooling and advanced microchannel cooling for both 2D and 3D microprocessors and identify interesting inflection design points down the road.

Keywords

Technology scaling, manycore architecture, processors, cooling solution, 3D integration.

1. Introduction

CMOS semiconductor technology has been continuously following Moore's Law for the last few decades. Moore's Law states that the number of transistor doubles about every other year [1]. This trend will continue in the foreseeable future. One consequence of technology scaling is the increased power per unit area. This is because transistor size scales faster than power per transistor. As a result, chip-level cooling solutions have been improved over the years to deal with more and more power-hungry silicon chips—from natural to forced air cooling with heat sinks and even liquid cooling. More recently, some advanced cooling techniques such as microchannel cooling have shown promising results for effective cooling of hot silicon chips [2].

Aside from silicon technology scaling and the evolution of the cooling solutions, there are two other important ongoing technical trends accompanying the roadmap of future microprocessors—manycore chip multiprocessors [3] and 3D integration [4]. Processors now integrate more and more parallel processing cores on the same chip. This is caused by the diminishing return on performance and the increasing overhead on power and area in scaling traditional single-core processors. The number of cores on a silicon chip is predicted to double with every technology generation [5]. Soon, a microprocessor can hold up to hundreds or more processor cores (i.e. "manycore" processor) [5]. In fact, current graphics processors already have hundreds of parallel processing units in a single chip [6] [7]. Another challenge caused by the continued technology scaling is the increased on-chip communication overhead, due to the fact that on-chip metal wires scale worse than transistors [8]. One solution is to stack multiple silicon dies in the third dimention, and connect them at the microscale level with through silicon vias (TSV). 3D integration greatly relieves the communication delay, and provides significantly more on-chip bandwidth that is required by faster circuits and more and more integrated processor cores [9]. 3D chips also allow the integration of silicon dies from different fabrication processes. For example, a DRAM die can be stacked with a processor die, bringing CPU and main memory into the same chip, which can drastically increase performance of memory operations [10]. However, 3D integration poses a serious thermal challenge—the power density is much higher, because multiple heat-dissipating silicon dies are stacked together.

Given the technology scaling trends, the continued increase in number of cores, and the possibility of 3D integration, it is important to investigate the thermal impact on future 2D and 3D chip multiprocessors. It is also important to identify the limits of different cooling solutions and the consequent implications on future processor design. All the existing studies focus only in one of the many aspects of the whole picture. Instead, this paper combines all aspects (technology scaling, manycore, 3D and cooling) and tries to identify important possible architecture inflection points that are caused by the thermal limitations down the roadmap.

A proper cooling solution that potentially makes future chip multiprocessors free of the thermal limit needs to meet the following criterion—it has to be scalable. In this sense, all 2D cooling methods will be limited because processor die area cannot scale up as fast for yield reasons. Good 2D cooling solution such as liquid cooling or even 2D microchannel liquid cooling can postpone the limit by a few generations. But ultimately, the exponential scaling of power and power density together with relatively fixed die area will stop it. Going to 3D integration exacerbates the 2D surface cooling problem. However, 3D integration also gives us the unique opportunity for a scalable cooling solution that is 3D. For the first time, we can possibly achieve both scalable cooling and scalable performance gain with a 3D cooling solution. So far, there is one 3D cooling technique shown promising results—3D microchannel liquid cooling. Therefore, in this paper, for the 3D chips, we will focus on 3D microchannel cooling. The arguments would also apply to any other emerging 3D cooling technique as long as it is scalable with the number of silicon layers.

This paper looks at the thermal impact of air cooling and microchannel cooling on manycore processors, based on projected

power densities for cores and peripheral parts of the processor. It also considers how the limit of air cooling and microchannel cooling would impact the 2D and 3D processor design. The results in this paper are not meant to be exactly accurate. We simply do our best to identify the overall trend with first-order models and stimulate collaborations between thermal designers and chip designers. *From our preliminary results, we found that air cooling will soon run out of steam in a few generations, and advanced 2D cooling such as 2D microchannel cooling seems to be able adequately remove heat manycore chips in the foreseeable future. However, there are several concerns that might make 2D microchannel cooling less efficient, thus maybe requiring 3D microchannel cooling, which makes 3D chip multiprocessors highly scalable in both performance and thermal design power (TDP).*

We first start with a review of related work. In Section 2, we introduce assumptions of cooling limits for air cooling and microchannel cooling, which are based on existing studies. Then, in Section 3 to Section 5, we briefly discuss the first-order models we use to account for CMOS technology scaling, manycore architecture scaling and 3D integration. In Section 6, we present our preliminary data and identify several insights regarding the potentials of different cooling solutions. In Section 7, we conclude this paper.

2. Limits of Cooling Solutions

In order to derive interesting inflection points that are caused by thermal limitations along the roadmap of future microprocessors, we have to first identify the cooling limits of different cooling solutions, beyond which another novel cooling solution is required. We mainly consider air cooling and microchannel liquid cooling in this paper, and leave other cooling solutions as future work.

2.1. Air cooling

There is no clear definition of the exact air cooling limit, as it is related to fan size, fan air volume speed, heatsink configurations and the choice of materials for heat sink, heat spreader and thermal interface material. Papers by Rodgers et al.[11] and Nakayama [12] survey air cooling limits. In particular, Zhou et al. [13] mention a limit of 150W/cm^2 in term of average chip power density. Processors usually have non-uniform power densities due to non-uniform activities at different locations on chip. Therefore, local power density can be much higher than average chip power density, so that power accommodated by a given cooling solution may be reduced. In this paper, we look at both overall chip power densities and localized core power densities. We also use the air-cooling limit of 150W/cm^2 (i.e. 1.5W/mm^2) as in [13]. Keep in mind if we consider local hot spots, the air-cooling limit can be much less than this value.

2.2. Microchannel cooling

As thermal designers realize the approaching of the air cooling limit, microchannel cooling has received considerable attention in the last few years as a plausible alternative. Experimental implementations have shown promising cooling results for microchannel liquid cooling. Tuckerman et al. [14] reported a high cooling rate of 7.9W/mm^2 for a relatively large channel size in 1981. More recent experiments by Koo et al. [2] show a moderate 1.35W/mm^2 for one layer in a modern 3D chip configuration in 2005. Brunschwiler et al. [15] at IBM reported up to 6.8W/mm^2 cooling capability for a 1cm^2 chip, which drops significantly to 4.25W/mm^2 when applied to a larger 4cm^2 chip, due to the reduced efficiency in pumping liquid into longer microchannels in larger chips. In this paper, we use the highest reported cooling limit (7.9W/mm^2).

By using high values for both air cooling and microchannel cooling limits, we are conservative in predicting the time where these cooling limits are met. In other words, the cooling limits will likely be met earlier than we predict.

3. Technology Scaling

As the feature size scales further into the sub-100nm range, conventional ideal CMOS scaling is not valid any more. It has been difficult to further scale the threshold voltage of a transistor without excessive leakage current and reliability issues. As a result, supply voltage scales very slowly to maintain an adequate overdrive voltage on the transistors. This is known as non-ideal CMOS scaling. This directly leads to the fact that transistor size scales faster than its power consumption, and hence power density (power per unit area) has been scaled up over the years. Following the non-ideal scaling analysis in [16], we can derive that the power consumption of a circuit (or processor core) with fixed architecture will be proportional to Vdd^2 across generations, i.e.

$$P_{n+1} = \left(\frac{Vdd_{n+1}}{Vdd_n}\right)^2 P_n \tag{1}$$

where Vdd is the supply voltage, n and $n+1$ denoting technology generations. Power density of a processor core with the same architecture across generations would scale as follows,

$$PD_{n+1} = \left(\frac{1}{s}\right)^2 \left(\frac{Vdd_{n+1}}{Vdd_n}\right)^2 PD_n \tag{2}$$

where PD is power density of the same processors core across technology generations. s is the scaling factor of technology feature size and is around 0.7. Power density usually scales up because feature size scales down faster than Vdd.

Furthermore, we assume the die size remains relatively constant for each processor family we investigate. This is mostly true according to past processor data, due to the fixed size of a lithographical reticule, which has been around 2cm by 2cm maximum. Further increasing the reticule size is extremely expensive because of the cost of making a large reticule and the significantly reduced yield of processed wafers.

Further detailed information about technology scaling trend can be found in [17].

4. Manycore Chip Multiprocessors

In addition to technology scaling, there are multiple possibilities regarding how processing cores in chip multiprocessors would scale in terms of number of cores and the microarchitecture of each core. For the number of cores, it is generally agreed that it would double (i.e. $2\times$) with every technology generation, at least for the next few generations. This is the assumption we use in this paper as well. As for the microarchitecture of each core, according to recent processor data from

major processor manufacturers, they tend to keep the microarchitecture fixed for a few generations by just scaling the technology and adding more cores. It is also likely that as more and more cores are added, each core may become slightly simpler across generations in order to accommodate more cores in the chip for higher parallel performance. However, core complexity is not likely to scale infinitely—too many simple cores would face the limitations posed by the "uncore" components (e.g. on-chip communication network among cores, caches and memories, I/Os) [18]. Additionally, in order to maintain performance for workloads with moderate to high sequential parts, where parallel processing does not help at all, it is important not to excessively scale down the core complexity, at least not for all the cores. This leads to heterogeneous cores in the same chip [19]. All these are very interesting research topics and are receiving great attention from the chip architecture community. Fig. 1(a) and (b) shows both actual core power densities and normalized power densities (normalized to latest technology in each family) of recent processors based on Intel Core Architecture, Intel NetBurst Architecture, and IBM POWER Architecture, respectively. Although the actual power density increases in each family across generations (Fig. 1(a)), the relatively constant and even less normalized power density scaling indicates that the core architectures in each processor family stay relatively constant, with a trend towards slightly simplified cores for better power efficiency (Fig. 1(b)).

Therefore, in this paper, we make a simple assumption that the core microarchitecture remains the same as technology scales. We look at recent processors that have a few to tens of processing cores from industry leaders. For each processor family, we scale the technology and the number of cores respectively and try to identify the generation where certain cooling limits are hit. It is important to notice that we do cover the cases where many simpler cores are integrated, by investigating the scaling trend an 80-core network-on-chip (NoC) processor from Intel, and the Sun Niagara processor family which has up to 16 cores to date.

Aside from power and power density scaling of cores, it is also important to consider the "uncore" components. This is because in the era of chip multiprocessors, the "uncore" parts, which include on-chip network, lower-level caches and I/O pads and their drivers, etc., also consume a significant amount of power. As the number of cores increases, the activities and loads on these 'uncore" components are likely to scale up rapidly, leading to high power consumption that may even comparable or higher than the power consumption of the cores. In this paper, we make the following assumptions for lower-level caches, on-chip network and I/O power scaling, respectively.

- Lower-level caches (LLC): We assume the same amount of die area is assigned to LLC. With the LLC area fixed, we can scale up the LLC power according to non-ideal scaling analysis. The power scaling factor varies from 1.3 to almost 2.0 for different processes. We use a representative value of 1.6 in this paper [17].

- On-chip network (OCN): OCN power increases with core numbers, as more cores need to communicate with each other. A first-order OCN power model for a regular 2D mesh network topology is presented in [20]. Where the authors derive that total OCN power is approximately proportional to the square root of number of cores (i.e. $\sqrt{N}$). In this paper, we adopt this model.

- I/O power: So far, I/O power has been kept around 10% of total chip power [21]. However, as the number of cores increases, significantly more off-chip signals and memory I/O accesses are required, especially for the case of 2D chip multiprocessor. It is predicted that the number of I/Os and the total I/O bandwidth will increase exponentially [22]. Therefore, the power that is needed to drive the off-chip I/Os will also increase exponentially. In this paper, we assume the I/O power also doubles every technology generation.

5. 3D Integration

One way to significantly reduce the off-chip I/O accesses is 3D integration. 3D integration allows shorter interconnects that would run across chip in the 2D case. It also allows integrating different processing technologies into the same 3D chip. For example, a silicon layer from the DRAM process can be stacked onto a silicon layer from the logic process, making it possible to integrate off-chip main memory on chip, which drastically improves memory bandwidth and reduces I/O power consumption. With 3D chips, the on-chip network can also be 3D, which can reduce OCN power consumption as well.

However, 3D integration also introduces a severe thermal problem—it sums up the power density of all silicon layers and poses a real challenge to conventional air-cooling solutions. Recently, 3D inter-layer microchannel liquid cooling has drawn attention as a promising alternative to heatsink-based air cooling [15, 2, 23]. With 3D inter-layer cooling, it is possible to scale up the thermal design power (TDP) almost linearly with the number of silicon layers. This is not possible with heatsink-based surface air cooling. Optimistically, microchannel cooling is reported to be able to cool down an average chip power density of 7.9W/mm^2 [14]. For modern chips with larger die size, in the environment of industry research labs, it is reported to be able to cool down and average chip power density of 4.25W/mm^2 or less [15].

6. Results and Discussions

We now scale from several representative data points of existing modern processors by major manufacturers (e.g. Intel and Sun) using models and assumptions presented in previous sections. Our scaling base cases are Sun Niagara T1 and T2 at 90nm and 65nm technologies [24] [25], Intel Core 2 Duo (Allendale) [26] at 65nm, and Intel 80-core Network-on-Chip (NoC) [27] processor at 65nm. We look at both the scaled power and the scaled power density of individual processor core as well as the entire chip. Total number of cores doubles every generation for each processor family. All the results are for 2D chips. The results are shown in Fig. 3 – Fig. 5.

As can be seen, both power and power density will be increasing exponentially, indicating greater challenges with both heat removal and power delivery. In addition, we can see that total chip power and power density scaled up faster than those of the cores, indicating the "uncore" components become more

and more important, if not dominating the cores, for future manycore chip multiprocessors. We also mark the air cooling limit (1.5W/mm^2) as a horizontal line in the power density charts.

In Fig. 5, we also explore the possibility where the frequency of each core is scaled up by 1.2× in addition to technology scaling for each generation. Scaling up operating frequency has been a general way to increase performance in the past decades. Here, we show that this is especially unfavorable from the thermal point of view—it significantly increase total power and chip power density, to the point that even advanced inter-layer microchannel cooling would likely to fail at the 16nm technology node (with an average chip power density of 5.35W/mm^2).

6.1. Air cooling: running out of steam

One observation from the results is that air cooling limit (1.5W/mm^2 average chip power density) will be hit in about two generations in every processor family. More aggressive low-power techniques and further simplification of microarchitectures may be able to postpone this a little bit, but the end of air cooling seems inevitable. In addition, 1.5W/mm^2 is a fairly optimistic limit for air cooling, actual air limit can happen even earlier. Going to 3D integration with air cooling even exacerbates the problem, as stacking multiple chips on top of each other increases power density linearly to the number of silicon layers, making air cooling limit be met much sooner in 3D integration. Since 3D integration is not in mass production yet, it is possible that air cooling would not even be a valid cooling option for high-performance 3D processors.

6.2. 2D microchannel cooling: a feasible solution

Among all the preliminary results, none of them hit the microchannel cooling limit (7.9W/mm^2), at least through 16nm technology node. This makes 2D microchannel cooling a feasible cooling option for future high-performance chips.

If microchannel cooling becomes practical and reaches a point of economies of scale that make it viable for mass-market systems, evolutionary scaling of the current architectural approach (basically doubling the number of general-purpose cores and keeping the microarchitecture constant) becomes feasible.

However, it's not clear we'll reach that point in 2D. In that case, the air-cooling limit forces us to extract greater efficiency out of the architecture, by using more specialized cores (for example, graphics processors and other domain-specific coprocessors can be an order of magnitude or better in energy efficiency, i.e. performance per watt), leading to a growth in heterogeneous architectures; It is also possible to embed DRAM on the same chip to reduce power wasted on off-chip I/O, or go to 3D integration for the same reason. If this is true, there will never be enough of a market for 2D microchannel to become mainstream.

Another limitation of microchannel cooling for large 2D chips is the inefficiency of pumping cooling liquid into longer microchannels. Experimental results have shown that larger chips significantly reduce the cooling efficiency of microchannel cooling [15]. This may require a shift to 3D chip and consequently 3D microchannel cooling.

6.3. 3D microchannel cooling: the ultimate solution

Ultimately, 3D microchannel cooling seems to be the solution that solves thermal problems in future high-performance chip multiprocessors [23], if this technology matures in the near future. Once we go in the direction of 3D, microchannel cooling might become more viable, because we have a much more severe cooling problem, and we have more layers in which to put the channels. Indeed, microchannel cooling may be a pre-requisite to go beyond just two layers as most existing 3D chip implementations already show. Being able to have parallel cooling paths in 3D microchannel cooling among silicon layers makes the thermal design power (TDP) of a 3D chip almost linearly scalable to the number of layers, which also means great performance scalability.

On the other hand, with 3D microchannel cooling, the microchannels may complicate the manufacturing process and compete for precious chip areas with through-silicon vias (TSV), so there is an interesting tradeoff between inter-layer cooling and inter-layer communication. Also, once we go to 3D integration, if microchannel cooling is effective enough, it is also possible that chip architectures may revert back to a general-purpose organization. All these are interesting research questions that need attention and collaborations from both computer architecture and thermal design communities.

Although we exclusively consider 3D microchannel liquid cooling for 3D chips in this paper, it is important to remember that the arguments also apply to any other emerging 3D cooling techniques as long as they are scalable with the number for silicon layers in a 3D chip.

6.4. Other cooling methods

Aside from convectional air cooling and the more recent microchannel inter-layer cooling, several other cooling solutions were also proposed. For example, phase-change cooling [28], spot cooling for local hot spots [29], and thermal-electric couple cooling [30]. However, available data and scalability of these cooling solutions are hard to find and predict. Therefore, we leave them as future extensions to the work presented in this paper.

7. Conclusions

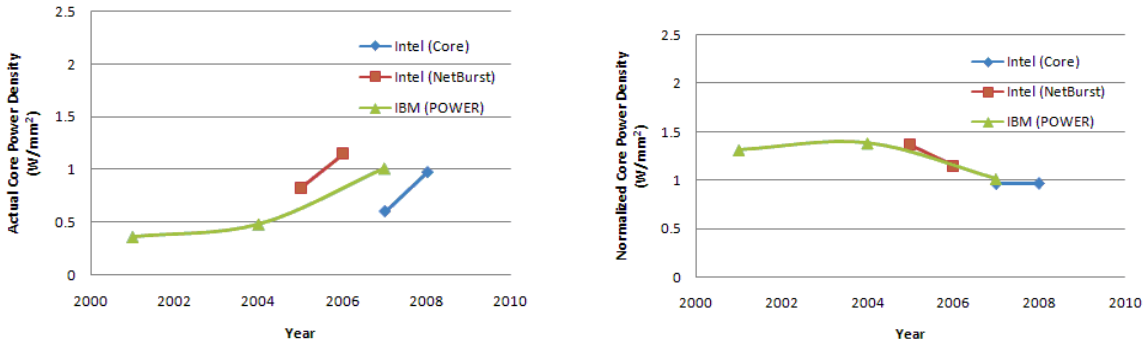
With the advent of chip multiprocessors and 3D integration, it is important to put future high-performance processors in the cooling perspective and find out the thermal impact on these new paradigm shifts, together with the continued Moore's Law and technology scaling. In this paper, we investigate the trends of power and power density scaling roadmap for chip multiprocessors, and identify important points where conventional air cooling limit will be hit. Our preliminary results show that air cooling will end in the near future, and 2D microchannel cooling is a promising alternative. Ultimately, 3D microchannel cooling is necessary for continued scalability of TDP and performance.

Acknowledgement

This work is funded by NSF grants CRI-0551630 and EHS-0509245, as well as a grant from Intel Research.

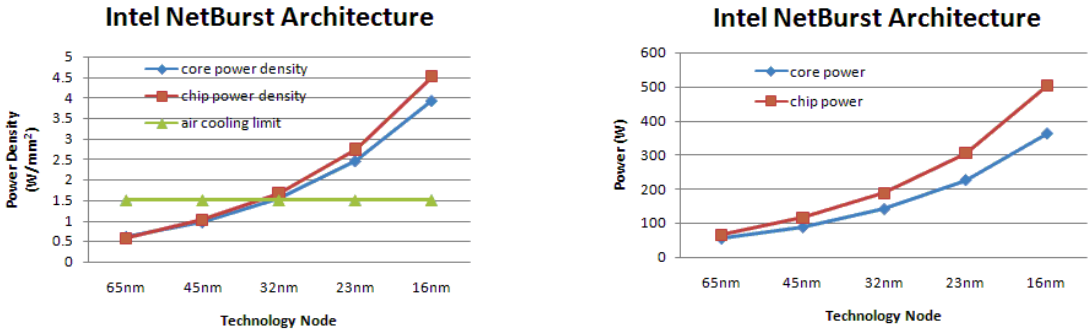
References

1. G. E. Moore. Cramming more components onto integrated circuits. *Electronics Magazine*, page 4, 1965.
2. J-M. Koo, S. Im, L. Jiang, and K. E. Goodson. Integrated microchannel cooling for three-dimensional electronic circuit architectures. *Journal of Heat Transfer*, 127:49–58, 2005.
3. B. A. Nayfeh L. Hammond and K. Olukotun. A single-chip multiprocessor. *IEEE Computer*, 30(9):79–85, 1997.
4. J. U. Knickerbocker and et al. Three-dimensional silicon integration. *IBM Journal of Research and Development*, 52(6):553, 2008.
5. S. Borkar. Thousand core chips—a technology perspective. In *Proc. of DAC*, 2007.
6. http://ati.amd.com/products/radeonhd4800/specs.html.
7. www.nvidia.com/page/geforce_8800.html.
8. R. Ho, K. Mai, and M. Horowitz. The future of wires. *Proceedings of the IEEE*, 89(4):490–504, 2001.
9. B. Black et al. Die stacking (3D) microarchitecture. In *International Symposium on Microarchitecture*, 2006.
10. G. Loh. 3d-stacked memory architectures for multi-core processors. In *International Symposium on Computer Architecture*, 2008.
11. P. Rodgers, V. Eveloy, and M. Pecht. Limits of air-cooling: Status and challenges. In *SEMITHERM*, 2005.
12. W. Nakayama. Exploring the limits of air cooling. *Electronics Cooling*, 12(3), August 2006.
13. P. Zhou, J. Hom, G. Upadhya, K. Goodson, and M. Munch. Electro-kinetic microchannel cooling system for desktop computers. In *Proc. of SEMI-THERM*, 2004.
14. D. B. Tuckerman and R. F. W. Pease. High-performance heat sinking for vlsi. *IEEE Electron Device Letters*, 2(5):126–129, 1981.
15. T. Brunschwiler et al. Forced convective interlayer cooling in vertically integrated packages. In *ITHERM*, 2008.
16. J. Rabaey, A. Chandrakasan, and B. Nikolic. *Digital Integrated Circuits: A Design Perspective*. Prentice Hall, Upper Saddle River, New Jersey, 2003.
17. The International Technology Roadmap for Semiconductors (ITRS), 2007.
18. G. Loh. The cost of uncore in throughput-oriented many-core processors. In *Workshop on Architectures and Languages for Throughput Applications (ALTA), in conjunction with International Symposium on Computer Architecture (ISCA)*, 2008.
19. M. D. Hill and M. R. Marty. Amdahl's law in the multicore era. *IEEE Computer*, 41(7):33–38, July 2008.
20. T. Konstantakopoulos, J. Eastep, J. Psota, and A. Agarwal. Energy scalability of on-chip interconnection networks in multicore architectures. Technical report, MIT CSAIL Technical Report, 2007.
21. S. Borkar. Intel corporation. personal communication. 2008.
22. R. Ross et al. High end computing revitalization task force (hecrtf), inter agency working group (heciwg) file systems and i/o researchworkshop report, 2006. http://institutes.lanl.gov/hec-fsio/docs/heciwg-fsio-fy06-workshop-document-finalfinal. pdf. 2006.
23. H. Jang, I. Yoon, C. Kim, S. Shin, and S. W. Chung. The impact of liquid cooling on 3d multi-core processors. In *International Conference on Computer Design*, 2009.
24. U. Nawathe et al. An 8-core 64-thread 64b power-efficient SPARC SoC. In *Proc. of ISSCC*, Febuary 2007.
25. G. K. Konstadinidis et al. Architecture and physical implementation of a third generation 65 nm, 16 core, 32 thread chip-multithreading sparc processor. *IEEE Journal of Solid-State Circuits*, 44(1):7–17, January 2009.
26. Intel xeon processor datasheets. http://www.intel.com/assets/pdf/datasheet/321321.pdf. 2009.
27. S. R. Vangal et al. An 80-tile sub-100-w TeraFLOPS processor in 65-nm CMOS. *IEEE Journal of Solid-State Circuits*, 43(1):29–41, January 2008.
28. S. Murthy, Y. Joshi, and W. Nakayama. Orientation independent two-phase heat spreaders for space constrained applications. *Microelectronics Journal*, 34(12):1187–1193, December 2003.
29. E. N. Wang et al. Micromachined jets for liquid impingement cooling of VLSI chips. *Journal of Microelectromechanical Systems*, 13(5):833–842, October 2004.
30. G. J. Snyder, M. Soto, R. Alley, D. Koester, and B. Conner. Hot spot cooling using embedded thermoelectric coolers. In *Proc. of Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, pages 135–143, March 2006.



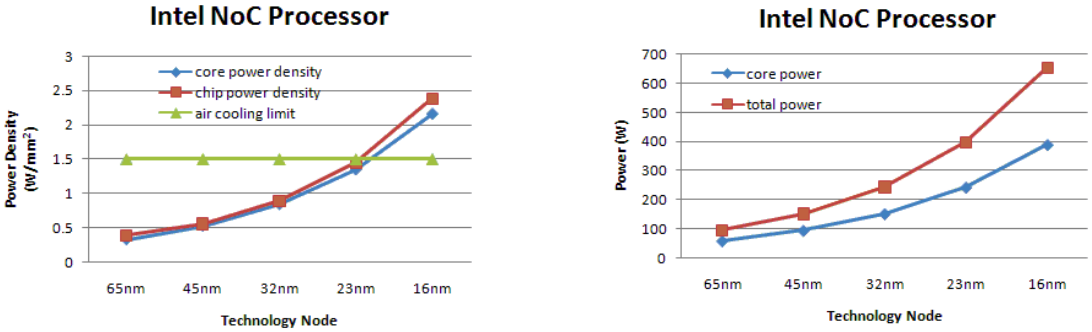
(a) (b)

Figure 1. (a) Actual core power densities and (b) normalized core power densities, for Intel Core architecture (diamond), Intel NetBurst architecture (square) and IBM POWER architecture (triangle). The normalized results in (b) show that processor core microarchitecture has stayed almost fixed or slightly simplified in the past few years for each processor family.



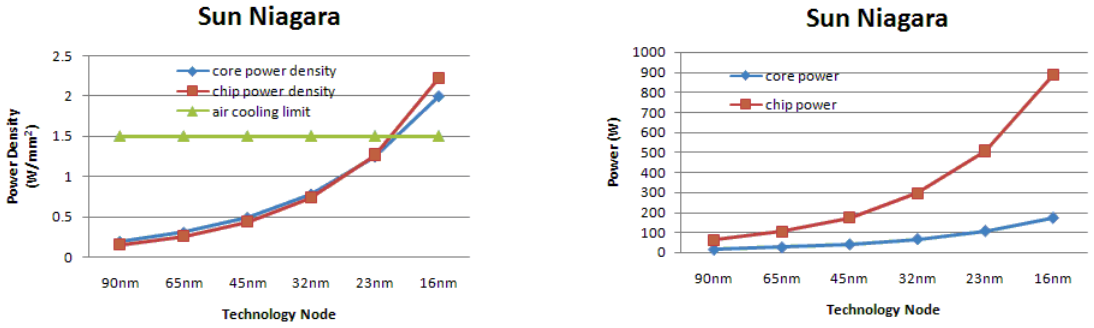
(a) (b)

Figure 2. (a) Power density scaling trend and (b) power scaling trend of Intel NetBurst architecture. Air cooling limits will be met at 32nm technology node. At 16nm, it would be even difficult for microchannel cooling to cool down such a chip if this scaling trend continues.



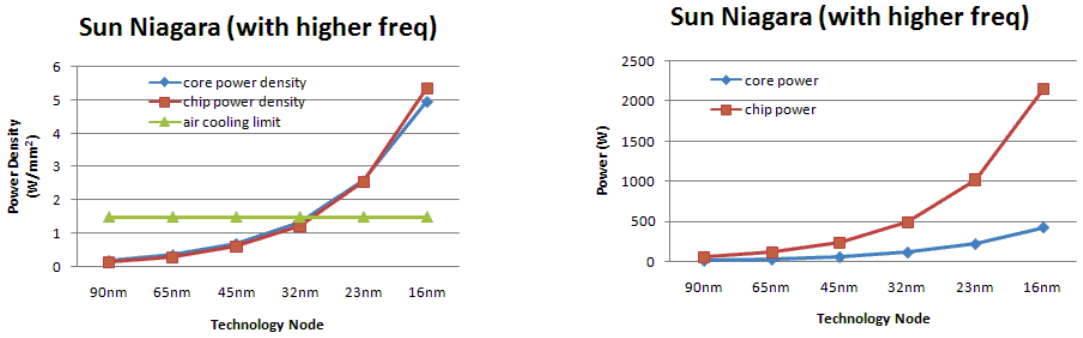
(a) (b)

Figure 3. (a) Power density scaling trend and (b) power scaling trend of Intel Network-on-Chip processor [27]. Air cooling limits will be met at 23nm technology node. 2D microchannel cooling should be able cool down such a chip at 16nm technology node.



(a) (b)

Figure 4. (a) Power density scaling trend and (b) power scaling trend of Sun Niagara chip multiprocessor [24]. Air cooling limits will be met at 23nm technology node. 2D microchannel cooling should be able cool down such a chip at 16nm technology node.



(a) (b)

Figure 5. (a) Power density scaling trend and (b) power scaling trend of Sun Niagara chip multiprocessor [24], with additional frequency scaling. Air cooling limits will be met at 32nm technology node. For 16nm and beyond, it would be even difficult for microchannel cooling to cool down such a chip if this scaling trend continues.