Scaling with Design Constraints – Predicting the Future of Big Chips

Wei Huang*, Karthick Rajamani*, Mircea R. Stan†, and Kevin Skadron‡

Abstract

The last few years have witnessed high-end processors with increasing number of cores and increasingly larger dies. Limited instruction-level parallelism (ILP), chip power constraints and technology scaling limitations caused designers to embrace multiple cores rather than single-core performance scaling to improve chip throughput. In this paper we try to answer whether that approach is sustainable by scaling from a state-of-the-art big chip design point using analytical models. We consider a comprehensive set of design constraints/trends including core growth rate, cache size, voltage-frequency scaling, thermal design power (TDP), hot spots and die area. We conclude that 1) Even at constatn frequency, a 2X per generation core growth will exceed TDP soon. 2) Scaling chip throughput will be difficult at constant TDP. Voltage scaling techniques, such as near-threshold operation, will be a key determinant on the extent of dark-vs-dim silicon when maximizing chip throughput. 3) Within two technology generations, the gap between technology-scaling-promised throughput and TDP-constrained throughput would need new architectural innovations to be bridged. 4) Even if relaxing strict TDP/area constraints, system power constraints might force the adoption of new packaging (3D, SiP) solutions to realize throughput growth. Then new thermal issues will be the hurdle, necessitating the adoption of better cooling solutions.

1 Introduction

Along with the continued CMOS technology scaling, frequency increase has been the dominant factor exploited for the growth of high-end computer system performance for several decades. At the same time, power has been controlled by reduction in supply voltage. However, with continued shrinking of semiconductor devices, difficulty in further scaling transistor threshold voltage limits the decreases in supply voltage. As a result, maintaining frequency growth is encountering noticeable power cost and consequently is no longer the favored path to increased performance. The prevalent approach to increased chip performance is to scale up chip throughput with more processing cores and threads.

This has caused significant growth in die area for high-end processors. For example, IBM POWER6, a dual-core (four threads) processor in IBM 65nm SOI technology, has a die size of 341mm^2 [1], and IBM POWER7, an eight-core (32 threads) processor in IBM 45nm SOI technology, has a die size of 567mm^2 [2]. Intel's Nehalem-based quad-core Xeon has a die size of 296mm^2 [3] with four cores (8 threads), whereas the 8-core Nehalem-EX has a die size of 684mm^2 [4] (16 threads).

There are three factors causing the increase in die size. First, growth in cores is exponetial (1.4-2X) in each generation while area scaling rate is about 0.6X, worse than the ideal 0.5X [5]. Second, the amount of last-level cache (LLC) per core has remained constant or increased, as seen from recent IBM, Intel and AMD processors

*Wei Huang and Karthick Rajamani are with IBM Research - Austin.
†Micea R. Stan is with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia.
‡Kevin Skadron is with the Department of Computer Science, University of Virginia

(e.g. Intel's 8-core Nehalem-EX and 10-core Westmere-EX both have 3MB L3 cache per core). Third, advance in fabrication processes have enabled reasonable yield for larger reticle sizes, lowering the hurdle to die-size growth.

While increasing parallelism rather than frequency is a more power-friendly approach to performance growth, the stricter requirement from chip power still forces processor vendors to keep the chip power at a constant *thermal design power* (TDP) in order to hold down cooling cost. For example, high-end Intel (x86-64) processors have been adhering to a 130-Watt TDP for a few technology generations.

It is not clear whether the increasing die area with multi-core growth is a sustainable approach to realize desired performance growth under power constraints, and when alternative approaches may be necessary. Additionally, what implications if any does the intersection of power constraints and technology scaling trends have for chip packaging and cooling technologies, and vice versa?

In this paper, we try to answer these questions by examining both the technology scaling predictions of the International Technology Scaling Roadmap for Semiconductors (ITRS) as well as scaling expectations for industry high-end processors [5]. The former adopts a more aggressive scaling assumptions than the latter. We employ analytical approaches to model chip power and temperature to understand the impact of scaling under different constraints/characteristics such as TDP, area, frequency, voltage scaling trends, core and cache size growth trends.

Our investigations lead us to the following conclusions. Neither reining in chip area nor the multi-core trend can rein in chip power by themselves when seeking a 2X chip performance growth per technology node. So if ever there was a time for new architectural innovations to increase chip efficiency (performance at constant power) it is now.

2 Related Work

The industry has showcased a number of state-of-the-art multi-core and many-core processors with relatively big die sizes [2, 6, 4, 7, 8, 9, 10]. There also have been abundant existing studies about many-core design considerations from both academia and industry. To list just a few examples, Borkar [11] provided an overview for the many-core design space and insights from Intel [5] for future exa-scale computing systems. Marty and Hill [12] proposed the many-core version of Amdahl's Law of performance scaling. Chung et al. [13] looked at power, performance, area tradeoffs for manycore design. Our previous work investigates the power and thermal impact on homogeneous and heterogeneous many-core processors [14, 15]. Many researchers also looked at the notion of dark silicon or dim silicon, where not all parts of silicon can operate at full speed due to power constraints. This triggers interesting architecture research ideas such as BubbleWrap [16] that takes advantage of dark silicon by allowing active cores to run faster and wear out faster, and later replace them with spare cores. Venkatesh and Sampson et al. [17, 18] looked at power-induced "utilization wall" with technology scaling work and proposed microarchitecture-level fine-grained dark silicon to improve core power efficiency. All these among many others are excellent work looking at different detailed aspects of many-core processor architecture. However, the community still needs a clearer view of the whole picture, especially where the many-core big chips are heading to, and the impact from technology scaling, together with chip- and system-level constraints.

This paper tries to clarify the future of "big chips". It provides a high-level overview for architects by considering various technology scaling scenarios together with design constraints such as power, area and temperature. It also projects possible paths and alternatives. We hope the insights from this paper would provide useful guidelines for future chip architectures and stimulate more in-depth research.

3 Scaling Methodology and Assumptions

We choose key characteristics of the Intel Nehalem-EX processor [4] to create our current-generation reference design point. The Nehalem-EX is a modern server processor with a large die. Relevant details of this reference

design point are listed in Table 1.

tech node	# of cores	frequency	TDP	die size	average power density	LLC per core
45nm	8	2.26GHz	130W	684mm^2	0.19W/mm^2	3MB

Table 1. Current-generation reference design point

Our scaling methodology projects the following from 45nm to 10nm technology nodes for this design point.

1. Power and power density. This includes active switching power (density) and leakage power (density) of cores, cache power with focus on last-level cache, and power for on-chip network. We assume power for other components such as clock spine and I/O to be relatively small, fixed percentage of total power (approximately 10% each [19, 20]) and to scale in the same fashion as core power.

2. Area. This includes area projection for cores, last-level cache as well as hot spot size within a core. Total estimated chip size includes the area for all the cores in the configuration and area for the last-level cache.

   Another factor that can potentially affect the chip size is the difficulty for fabrication and packaging technologies of scaling down I/O and power-delivery bump (C4) sizes and their pitches. It is not clear whether the need for sufficient C4 bumps will dictate the die area for future big chips. The advent of 3D integration (especially stacking of memory) and on-chip voltage regulators could potentially reduce the number of I/O and power/ground bumps. Qualitatively, the area constraint from C4 does not change the power scaling trends. But in a C4-constrained scenario, surplus die area could be used to space out hot units, reducing chip power density, which relaxes the constraint on hot spot temperature. We will leave a detailed study on this issue as a future work.

3.1 Technology and frequency scaling

For technology scaling, we adopt two representative sets of scaling parameters that are publicly available. One is from ITRS [21], with a feature size scaling factor of 0.7X, which leads to an area scaling factor of 0.5X. Combined with this area scaling assumption, we assume that chip frequency as constant to match the trend observed in the last couple of generations of high-end server processors from different vendors. Key scaling parameters from one technology node to the next are listed in Table 2. As can be seen, the ITRS scaling has almost perfect power and power density scaling (half power per generation and constant power density scaling), representing an ideal scaling trend. The same scaling factors are assumed from each technology node to the next.

feature size	area	capacitance (C)	frequency (f)	V_{dd}	power (CV_{dd}^2f)	power density
0.7X	0.5X	0.616X	1.0X	0.925X	0.527X	1.054X

Table 2. Cross-generation scaling factors for our ITRS scaling model. Constant frequency is assumed. Same set of factors used for every generation.

The other set of scaling parameters based on a recent industry projection from Intel [5] is closer to practical high-end server chip trends for area and voltage scaling, as is verified by recent Intel processor die photos and power and frequency specifications. We also observe qualitatively similar scaling trends in IBM CMOS technologies. Key scaling factors are listed in Table 3 for this *Industry* scaling model. Distinct scaling factors are used for each generation in line with the published expectation. This includes a gradually diminishing frequency increase (instead of no increase as with our ITRS model). Because of the more conservative area and voltage scaling assumptions and higher frequency target assumptions for our *Industry* model versus our *ITRS* model,

it would have a higher power and power density for same performance/area every generation. The geometric means of power and power density scaling factors for our Industry model are 0.652X and 1.141X in contrast to the 0.527X and 1.054X for our ITRS model.

tech node	feature size	area	capacitance (C)	freq (f)	V_{dd}	power ($CV_{dd}^2 f$)	power density
45-¿32nm	0.755X	0.57X	0.665X	1.10X	0.925X	0.626X	1.096X
32-¿22nm	0.755X	0.57X	0.665X	1.08X	0.95X	0.648X	1.135X
22-¿14nm	0.755X	0.57X	0.665X	1.05X	0.975X	0.664X	1.162X
14-¿10nm	0.755X	0.57X	0.665X	1.04X	0.985X	0.671X	1.175X

Table 3. Cross-generation scaling factors for our Industry scaling model, adapted from [5]

3.2 Cores

We assume a homogeneous design with identical cores for each generation. We also assume no change to the core architecture for this work - our analysis will show the need for architecture changes based on the gap between desired and estimated power, performance and area.

For the growth in number of cores across technology generations, we consider three cases: 1) double cores every generation, i.e. 8, 16, 32, 64, 128 cores from 45nm to 10nm. 2) less aggressive core scaling, i.e. 8, 12, 16, 24, 32 cores from 45nm to 10nm. In the second case, as we try to make the number of cores an even number, the scaling ratios between every two generations is 1.5X or 1.33X, with a geometric mean of 1.4. 3) For some scenarios, it is possible to scale number of cores in between the first two cases, e.g., to meet a particular TDP, a 2X scaling factor results in too high a power and a 1.4X factor leaves TDP under-utilized. We label the three core scaling cases as **2X, 1.4X, 1.4$^+$X**, respectively.

3.3 Last-level cache scaling

We assume a SRAM-based LLC as is used by most processors vendors, with the exception of high-density embedded DRAM in IBM processors [2]. SRAM cell area scaling is close to that of logic scaling. On the other hand, its supply voltage scaling is usually much slower than that of logic circuits [22]. This is required for reliable storage of bits in the presence of variations and cosmic radiation. In this work, we consider three LLC supply voltage scaling options: 1) **SV1**: aggressive scaling similar to logic (0.925X each generation), and 2) **SV2**: a slower more representative case, specifically, 0.95X, and 3) **CV**: constant SRAM supply voltage, pessimistic for now, but likely the norm after a couple of more generations.

3.4 On-chip interconnect power scaling

A recent effort on power modeling for on-chip networks, ORION [23], shows that power per hop remains relatively constant across technology generations. For a scalable network topology, the total on-chip network power is proportional to the number of cores. We use the Intel 80-core processor [24] as the reference point for per-core on-chip network power.

3.5 Leakage power scaling

In recent years, there have been significant efforts, e.g. multiple threshold voltages and body bias adjustment, to keep leakage power from dominating active power. As a result, leakage power has managed to be confined as a relatively constant portion of the total chip power. Constant leakage current per transistor width has been projected by ITRS. Intel also projects 1X to 1.43X scaling factor for chip leakage [5] power density, giving a

geometric mean of 1.19X for leakage power density growth and 0.68X for leakage power growth. As we can see, they are close to the active power and power density projection in Table 3, where the per-generation scaling factors average (geometric mean) out to 1.141X for active power density and 0.652X for active power. Therefore, in this paper, we assume leakage power scaling is the same as active power scaling.

3.6 Per-core performance scaling

For better power efficiency, one could consider lowering the operational voltage (and frequency) below the baseline projections for each technology node outlined earlier. This helps obtain a higher chip throughput under power constraints with more cores, even as it lowers per-core performance. However, as nominal voltage itself is reduced every generation, lowering operating voltage further can get increasingly difficult. We consider three scenarios for our analysis where the V_{min} for a technology can be 0.9X, 0.7X or 0.5X of its nominal voltage V_{nom}. 0.7X is representative of reductions possible today, 0.9X as we go forward and 0.5X a more aggressive possibility with specialized tuning of manufacturing and circuits (approaching near-threshold computing proposals).

For all three cases, we assume frequency reduction is roughly proportional to supply voltage reduction. For processes where frequency reduction is non-linear within a certain narrow voltage range, this assumption would represent the average per-core performance. If power reduction is required beyond V_{min} (e.g. to meet a TDP constraint), the only way to further reduce power without power gating cores is to reduce frequency, with supply voltage staying at V_{min}.

We label the three V_{min} possibilities as **0.9Vnom, 0.7Vnom, 0.5Vnom**, respectively. Note that when we scale down supply voltage, say with 0.5Vnom, we could run at 0.6Vnom if that is adequate to meet the power constraint.

3.7 Hot spot

Modern high-performance processors usually have parts of the cores heavily utilized (e.g. register files, functional units or first-level caches), which result in hot spots with much higher power density than the rest of the chip.

Previous studies [15] [25] show that as the hot spot size gets smaller, the difference between hot spot temperature and average chip temperature drops exponentially for a constant hot spot power density. So even a significant increase in hot spot power density across generations may only cause a negligible increase in hot spot temperature because of the decrease in hot spot size. Intuitively, smaller hot spot size leads to less hot spot power and more lateral heat spreading within silicon.

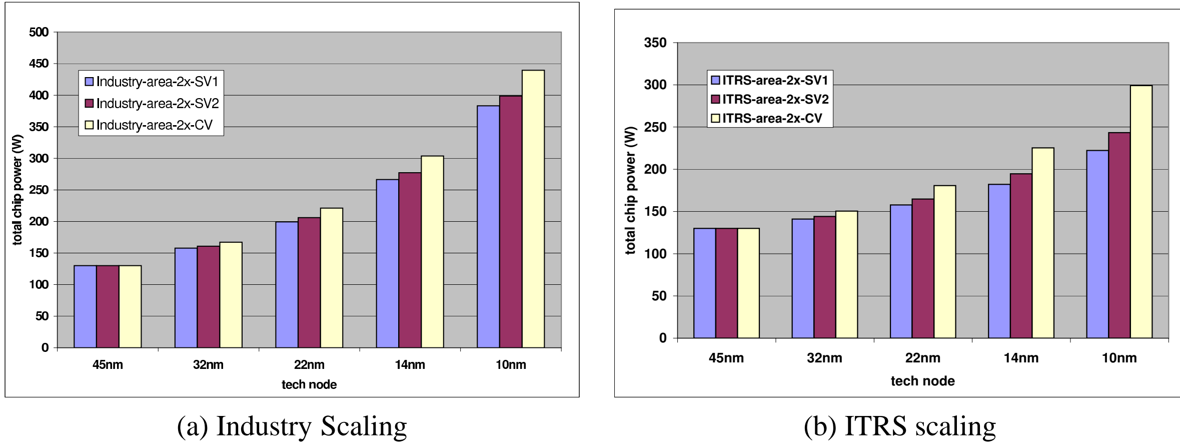
Hot spot temperature can be divided into three components [15]:

$$T_{hotspot} = T_{ambient} + T_{chip_average} + T_{hotspot_difference}$$

where $T_{ambient}$ is the ambient temperature. $T_{chip_average}$ is the chip-wise average temperature, determined by total chip power and the cooling solution. $T_{hotspot_difference}$ is the temperature difference between hot spot and average chip temperature.

We use the reported Pentium 4 hot spot size of 1mm^2 [26] as a starting point, scaling it to different technology nodes. This is reasonable since hot structures usually have minor changes from generation to generation. Additionally, we adopt a ratio of 18 between hot spot power density and average chip power density, derived from Wattch [27] scaled power estimations for Alpha EV6 processor [28]. Since both Pentium 4 and Alpha EV6 are processors with more severe hot spots, these assumptions are representative of worst-case thermal conditions. We expect future thermal scenarios to be similar to or more benign than our predictions for similar chip powers.

We assume a good, server-class air cooling solution with a 0.1K/W heatsink-to-air thermal resistance and a 35°C air temperature at the server air inlet, a typical upper limit for server operating environment.



(a) Industry Scaling

(b) ITRS scaling

Figure 1. Impact of scaled vs. constant LLC SRAM supply voltage on total chip power with constant die area.

4 Results

Before delving into the results, we first list possible design constraints.

1. Constant die area (labeled as **area**).
2. Constant thermal design power (labeled as **TDP**).
3. Constant last-level cache capacity per core (labeled as **LLC**).

These design constraints, together with the assumptions in Section 3—ITRS vs. Industry scaling, scaled vs. constant LLC supply voltage, 2X vs. 1.4X cores per generation—create multiple scaling scenarios. In the following, due to the limited space, we primarily show the cases with *Industry* scaling parameters, as this represents what is more likely to be achievable. Results with *ITRS* scaling parameters are mostly discussed qualitatively.

4.1 Constant Die Area

Figure 1 shows the total chip power for different last-level cache supply voltage scaling factors. In particular, Figure 1(a) compares Industry-area-2x-SV1 (i.e. industry scaling parameters, constant die area, 2X cores per generation, and 0.925X LLC SRAM supply voltage scaling), Industry-area-2x-SV2 (0.95X SRAM supply voltage scaling) and Industry-area-2x-CV (constant SRAM supply voltage). Overall, the impact of slow or flat SRAM supply voltage scaling on total chip power becomes more important over time (14% greater chip power for CV than SV1 at the 10nm node). With constant area, the Industry scaling parameters result in a decreasing cache capacity per core each generation. For the ITRS scaling parameters (Figure 1(b)), since core size scaling is more aggressive (0.5X, rather than 0.6X), there is more room left for LLC for a constant die area. Therefore, the impact of SRAM voltage is more significant (33% higher chip power for CV over SV1 at 10nm).

Figure 2 shows the constant area constraint's impact on LLC capacity per core. Notice that for industry scaling parameters, since core area scales less aggressively at a rate of 0.6X, double cores every generation within a constant die area would reduce LLC capacity per core. This also means that for cache-capacity dependent workloads, per-core performance could be lower for 2X than for 1.4X core growth with constant die area despite of 2X's higher total chip power.

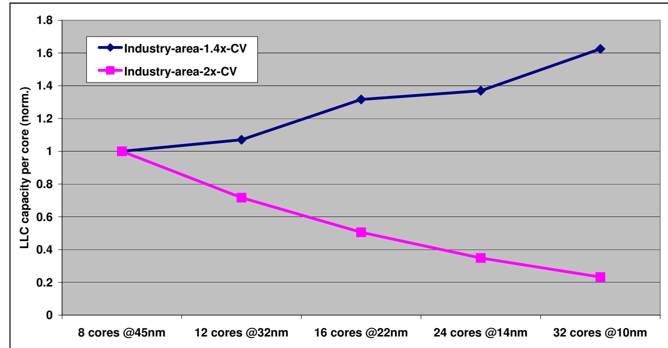


Figure 2. For constant die area and industry scaling parameters, 2x cores result in less LLC capacity per core, whereas 1.4x cores leave more room for additional LLC capacity.

On the other hand, if ITRS scaling projections are followed, since the core size scaling is the ideal 0.5X, the total core area would remain the same if we double cores every generation. This also means LLC capacity per core will remain constant, maintaining single-core performance.

4.2 Constant TDP

Power constraints at system level and power-delivery and cooling costs at chip level are keeping processor chip TDP constant. To double chip performance each generation, even maintaining single-core performance, one would need 2X cores every generation at constant LLC capacity per core. Therefore, it is natural to look at the scenario of Industry-TDP-LLC-2x-SV2 (i.e. Industry scaling, constant TDP, constant LLC per core, double cores per generation, 0.95X LLC SRAM voltage scaling).

Figure 3 shows that frequency has to drop for each core as technology scales in order to keep constant TDP with 2X cores and same LLC/core. We call this forced reduction in per-core performance "dim silicon". At the 45nm reference point, all cases start at a normalized frequency of 1.0. For each node, the 5th bar shows what the technology scaling can offer. The other bars show how much frequency has to drop as a result of the power wall (e.g. constant TDP) for different logic and SRAM voltage scaling option pairs. Since the 0.7Vnom case (1st bar for each technology node) allows more voltage reduction, it suffers less in terms of frequency reduction than the 0.9Vnom case (2nd bar). Similarly, since the SV2 case (2nd bar) has less LLC power (due to better SRAM voltage scaling), it can run at higher frequencies than the CV case (3rd bar). In combination, the CV-0.9Vnom case suffers the most frequency reduction (25% of ideal frequency at 10nm). As the feature size shrinks, the importance of running at lower voltage increases. At the 10nm node, 0.5Vnom (4th bar) can obtain a higher frequency than 0.7Vnom. If one considers 0.5Vnom as near-threshold operation, we may not need to go there till the 10nm node. It also appears that, if we can scale voltage down to 0.7Vnom at 14nm and to 0.5Vnom at 10nm, we may not need to adopt dark silicon.

Another option to maintain TDP is to scale core number slower than 2X per generation. For Industry-LLC-1.4x, we find that even for the CV (worst-case SRAM scaling) case, total chip power is still within the TDP constraint for all generations. So we can scale core number a bit more aggressively than 1.4X, but less than 2X (i.e. 1.4^{+}X), without reducing frequency, as seen in Figure 4. For example, at 22nm, 30% more cores can be supported with 1.4^{+}X than the 1.4X case.

Figure 5 shows the performance impact of dim silicon for both 2X and 1.4^{+}X core growth rates as a result of the constant TDP constraint. Moving from (a) through (d) we have worsening voltage scaling assumptions from $0.5V_{nom}$ to $0.9V_{nom}$. For each graph, the y axis is the chip throughput normalized over the single-core throughput at 45nm, where all cores are always on and core performance scales perfectly with frequency. As

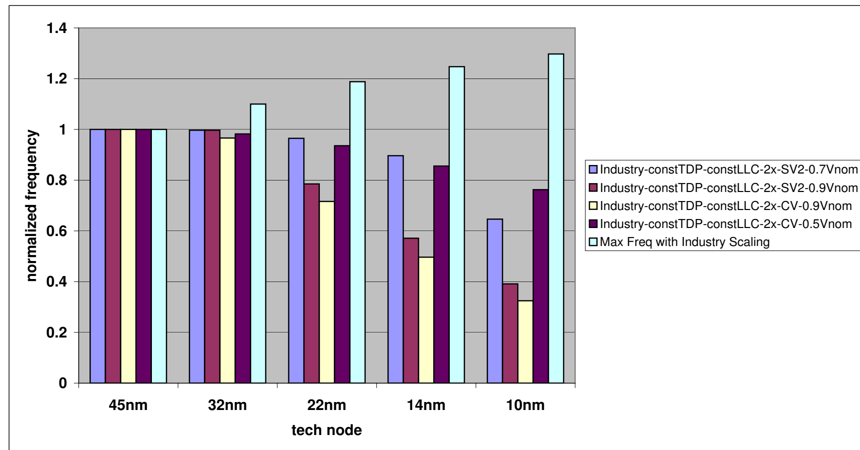


Figure 3. Frequency reduction to keep constant TDP, for different V_{min} cases.

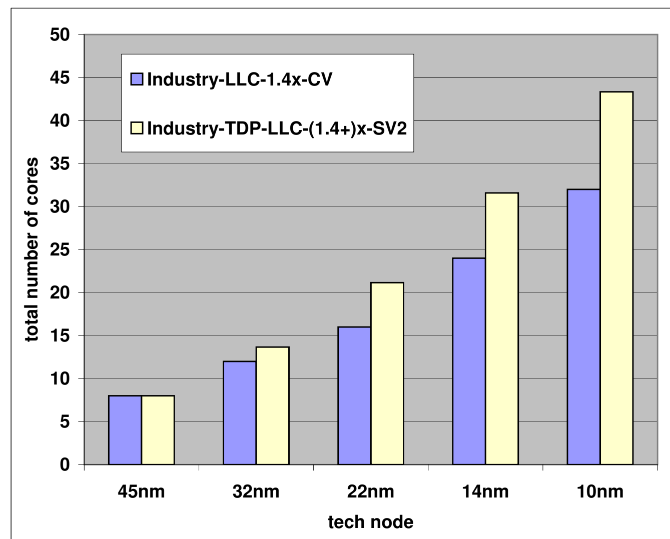
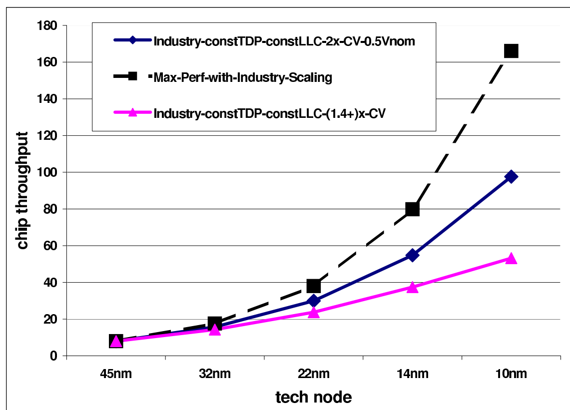
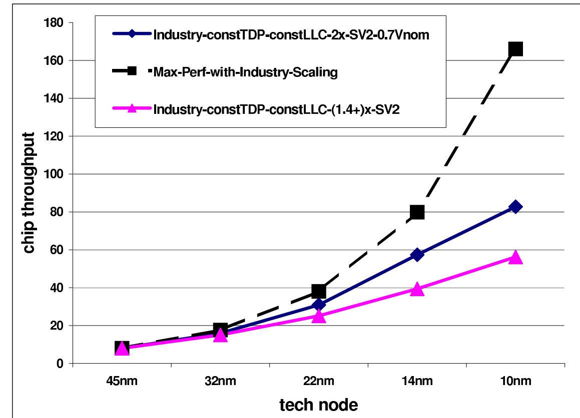


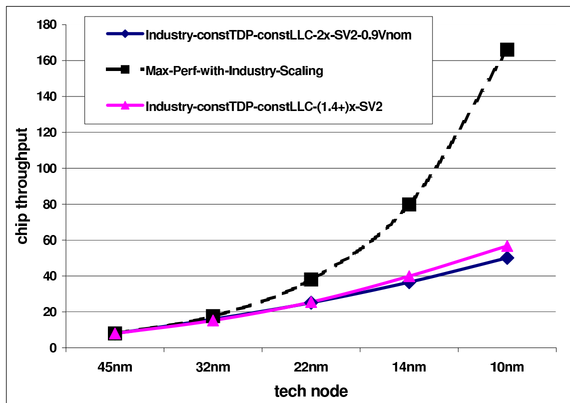
Figure 4. Number of cores to reach constant TDP without scaling down voltage.



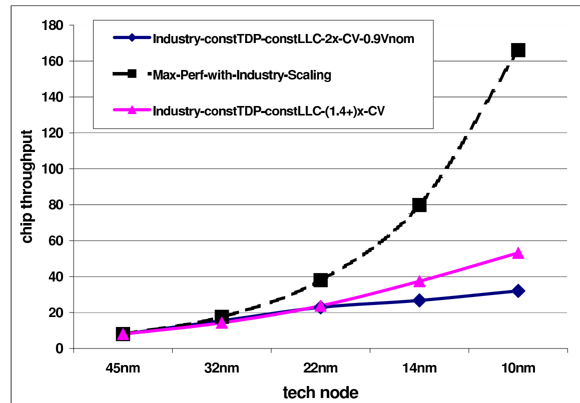
(a) CV 0.5 V_{nom}



(b) SV2 0.7 V_{nom}



(c) SV2 0.9 V_{nom}



(d) CV 0.9 V_{nom}

Figure 5. comparison of ideal throughput normalized to a single core among different Industry-TDP-LLC cases.

we can see in Figure 5(a), for Industry-TDP-LLC-CV-0.5Vnom, since more voltage reduction is allowed, the 2X cases have higher throughput than the 1.4^{+}X case, even though 1.4^{+}X cases suffer no frequency reduction. However, as voltage reduction range becomes smaller and LLC SRAM voltage scaling becomes flat, 2X cases gradually become worse than the 1.4^{+}X case ((b)-(d)). The 2X cases also use larger die area.

Maintaining TDP will keep chip performance less than expected from core growth and maximum potential frequency. Whether using fewer cores at higher frequencies or more cores with scaled down voltages and frequencies yields higher chip throughput will be determined by the extent of achievable voltage reduction both for logic as well as SRAM. From the technology point of view, the most likely case at 10nm would be SV2-$0.9V_{nom}$ (Figure 5(c)), which favors adding fewer cores and achieves about 7X throughput improvement over 45nm (55/8). Notice how far away this is from what the technology scaling and 2X cores can potentially offer (about 165/8=21X).

In order to fully exploit technology scaling and 2X cores, architecture innovations is needed to reduce chip power density. Table 4 shows the required percentage of power density reduction for Industry-TDP-LLC-2X cases to avoid sacrificing single-core performance from the *Industry* scaling scenario. Low-power semiconductor processes, but ultimately, we need architecture changes such as incorporating accelerators for higher power efficiency or stripping flexibility out of the hardware (e.g., large register files, dynamic instruction issue, caches, etc.) and relying more on the compiler.

scenarios	% of power density reduction needed at 32nm	22nm	14nm	10nm
Industry-TDP-LLC-2X-SV	11.2%	25.4%	41.0%	56.5%
Industry-TDP-LLC-2X-SV2	15.0%	30.1%	44.6%	58.3%
Industry-TDP-LLC-2X-CV	18.7%	36.2%	51.9%	65.4%

Table 4. Required chip power density percentage reduction for Industry-TDP-LLC-2X cases to fully exploit technology scaling and 2X cores

The above analysis assumes multi-program workloads that can run fully in parallel on all the cores. For a single-program workload with both sequential and parallel parts, its sequential part can run on a single core with boosted "Turbo" frequency while leaving the other cores idle. This can partially compensate for the single-core performance loss imposed by the power wall for this type of workloads. For many-core processors, the amount of single-core frequency boost in this case would likely be limited by the amount of voltage boost allowed by semiconductor process, rather than TDP and hot spot temperature.

Figure 6 shows that for 1.4^{+}X cases to meet a TDP target, the die area is less than the original area. Increased LLC for higher performance for memory-bound applications can be used to fill in the spare die area. But since more LLC would cause more power consumption, core frequency then would have to be further reduced.

For ITRS scaling parameters, as feature size scales at a rate of 0.5X, the concern of increasing die size goes away for the case of 2X cores per generation with constant LLC capacity per core. However, in order to keep constant TDP, frequency still needs to drop, but to a lesser degree than the Industry-TDP-LLC-2X case. Again, no cores need to be turned off even for the worst LLC voltage scaling and 0.9Vnom voltage reduction range. The other conclusions drawn from the Industry scaling (adding fewer cores or using more LLC) still hold.

Because of the constant TDP constraint, hot spots are not an issue in this case. Current cooling solutions already accommodate such a TDP.

4.3 Relaxed constraints

What would happen if we relax the constraints on TDP and die area?

Figure 7 shows the total chip power without TDP constraint with and without the area constraint. As we can see, adding the area constraint would slows down the increase in total chip power, but at the expense of

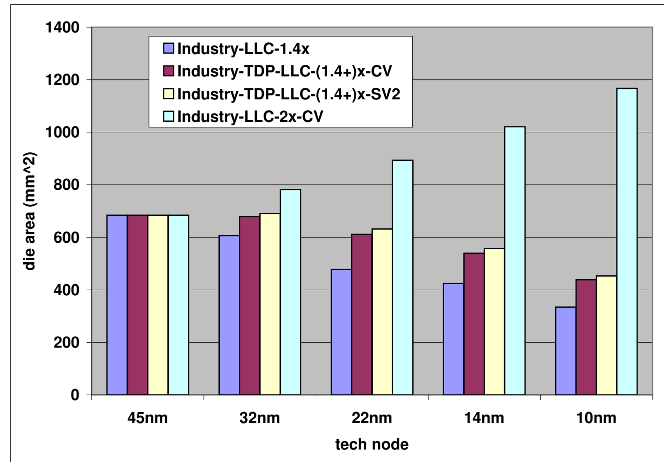


Figure 6. Die area for 1.4X and 1.4^{+}X cases. The last bar show a significant area increase when there are no TDP and area constraints.

reduced LLC capacity per core. With Industry-LLC-2x-CV, one achieves the maximum possible performance, but with significant costs: ∼2X die area (last bar in Figure 6), and ∼5X chip power at 10nm compared to 45nm (Figure 7). Such a design will hold at most for a couple generations from now before it is too costly to fabricate and too power hungry. Maybe a few niche systems might adopt such a design.

It is also worthwhile to consider how thermally limited such a non-TDP-constrained design is. Figure 8(a) shows that, with a high-end server-class cooling solution, the hot spot would reach about 110°C and be a concern only at 10nm, where a better cooling solution is required.

Figure 8(b) shows the hot spot power density at each technology node. As we can see, hot spot power density can increase by up to 3X from 45nm to 10nm. However, Figure 8(c) shows the shrinkage in hot spot size. Figure 8 (a) shows the breakdown of temperature factors that contribute to the hot spot temperature. One observation is that hot spot's temperature difference to the chip average temperature remains relatively constant, whereas the average chip temperature component (dependent on chip total power) grows most as feature size shrinks. Consequently, chip total power is the most important contributor to hot spot temperature. The reason for this is already explained in the hot spot assumptions in Section 3.

4.4 Higher level of integration

As we have seen, Industry-LLC-2X without TDP constraint achieves the highest performance with huge area and power that are hard to sustain beyond a couple of technology generations from now. More importantly, such a design poses a serious challenge at higher system levels (e.g. server, rack and data center levels).

At the server level, it is very likely that such high-power processors would take almost all of the server power budget, leaving no power for other system components such as main memory, storage and network switches. This would contradict current trends, where vendors devote an increasing fraction of system power to the off-processor components such as memory [29], to maintain system performance.

Increasing server-level power budget from today's limits would incur high cost in operational expenses and data center facilities renovations. Therefore, a power constraint at the system-level would limit the power of big chips. The solution will likely be for the chip to integrate more system components at lower system power cost. Emerging techniques such as System-in-Package (SiP) or 3D integration will be required. With higher level of integration, we might put main memory and non-volatile memory into the same chip as the processor,

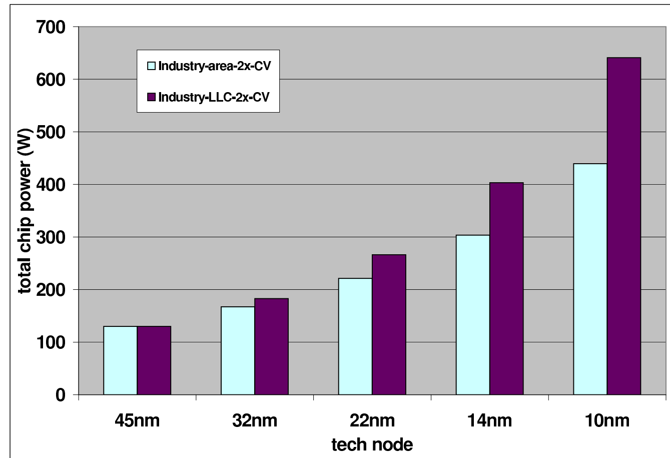
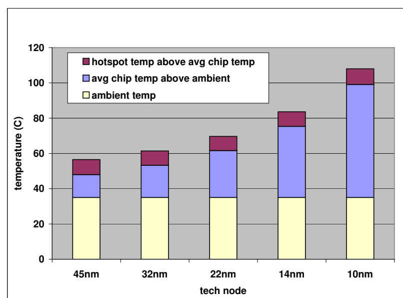
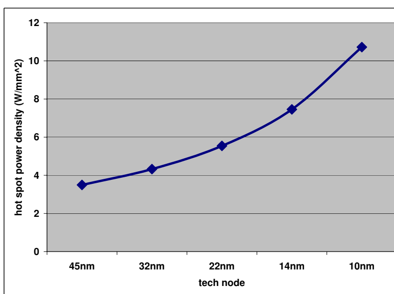


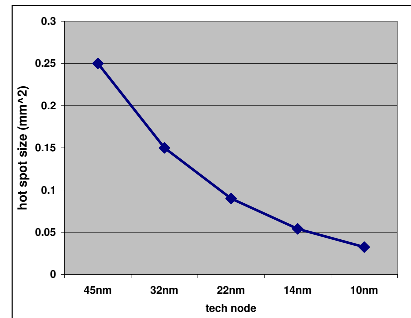
Figure 7. Chip power comparison for cases without TDP constraint.



(a)



(b)



(c)

Figure 8. (a) hot spot temperature breakdown, (b) hot spot power density scaling, and (c) hot spot size scaling

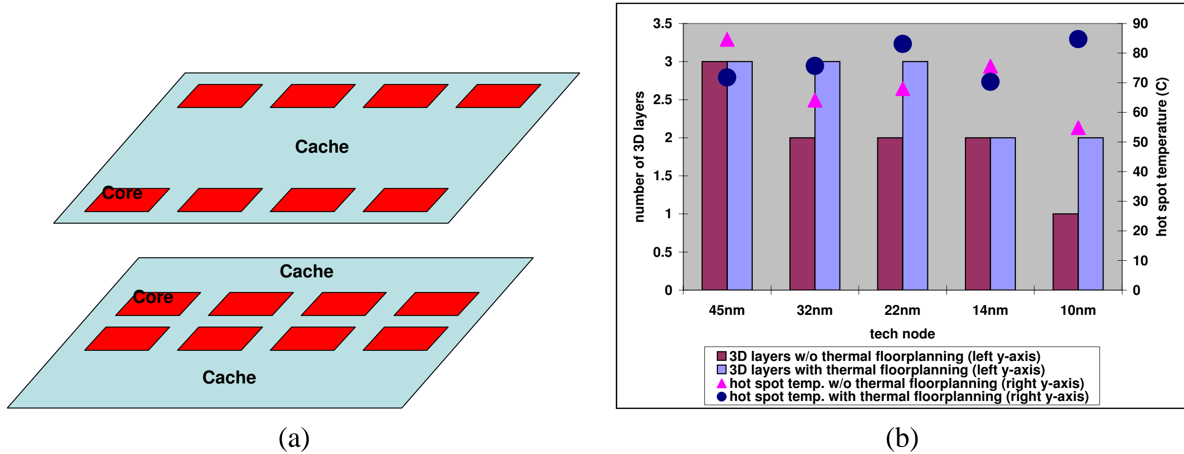


Figure 9. (a) thermal-aware 3D floorplanning (b) number of 3D layers and hottest temperature

resulting in a more power-efficient design that may meet the system power budget and reduce the need for I/O and power/groud pads.

However, 3D integration incurs higher power density and higher chip power as more layers are stacked. Therefore, temperature could become a major challenge, especially if we stack multiple processor dies together (in the case of a multi-processor server). Figure 9(a) shows a simple 3D floorplanning scheme that avoids overlapping hot core on top of each other. This helps reduce hot spot temperature and increase the number of 3D layers integrated. The bars in Figure 9(b) show the number of 3D layers that can be accomodated within an 85°C hot spot limit, for both thermal-unaware and thermal-aware 3D floorplanning schemes. The points show hot spot temperature of that number of layers. Where the number of layers is the same, the estimated hot spot temperatures for the two different schemes show the benefit of thermal-aware floorplanning.

Even with such a thermal friendly floorplan, the number of processor layers that can be stacked is still limited (three layers at most for conventional air cooling). In order to enable higher level of 3D integration, a drastic change in cooling solution is needed. Inter-layer cooling solutions such as microchannel cooling are more scalable to number of integrated 3D layers.

5 Conclusions

Non-ideal technology scaling has begun to pose a power wall on many-core big-chip designs. In this paper we examine the trend for big-chip scaling in the context of power, thermal and area constraints, together with its impacts on per-core performance and overall chip throughput, resulting in the following observations:

- Chip power will continue to grow even if die size is kept relatively constant. SRAM voltage scaling will be an important factor for power reduction with large on-chip caches. Limited SRAM voltage scaling leads to a significant power cost.

- With the trend in fixed Thermal Design Power (TDP), maintaining chip throughput growth will be a challenge. Significant drop in per-core performance will be needed, together with a judicious decrease in the degree of adding more cores and voltage reduction.

- Relaxing area and TDP constraints is necessary for the ideal 2X per generation chip throughput growth. Allowing for significantly increased power, however, will require more sophisticated cooling solutions after three generations, and will complicate system-level design as processor power will tend to eat almost

all the power budget at the system level. Power-efficient chip and system architectural innovations will become critical to maintain performance growth.

- As a result, increasing chip power might necessitate new technologies, such as 3D and SiP, to integrate more functions into the chip to operate within restricted system power limits, in three generations. Novel cooling solutions will be necessary, otherwise thermals can stop 3D integration after two generations.

Acknowledgments

The authors would like to thank Mark Sweet and Michael Floyd at IBM System and Technology Group, also Anne Gattiker, Jente B. Kuang and Peter Hofstee at IBM Austin Research Lab for insightful discussions. This work was supported in part by NSF grant MCDA-0903471.

References

[1] B. Curran et al. Power-constrained high-frequency circuits for the IBM POWER6 microprocessor. *IBM Journal of Research and Development*, 51(6):715–731, November 2007.

[2] D. Wendel et al. The implementation of POWER7: A highly parallel and scalable multi-core high-end server processor. In *Proc. of Intl. Solid-State Circuit Conf. (ISSCC)*, February 2010.

[3] http://ark.intel.com/product.aspx?id=43233.

[4] http://www.qdpma.com/CPU/CPU_Nehalem.html.

[5] S. Borkar. The Exascale Challenge - Asia Academic Forum 2010. Available at http://cache-www.intel.com/cd/00/00/46/43/464316_464316.pdf, November 2010.

[6] D. Pham et al. The design and implementation of a first-generation CELL processor. In *IEEE Solid-State Circuits Conference (ISSCC)*, February 2005.

[7] T. Fischer et al. Design solutions for the Bulldozer 32nm SOI 2-core processor module in an 8-core CPU. In *IEEE Solid-State Circuits Conference (ISSCC)*, February 2011.

[8] http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf.

[9] U. Nawathe et al. An 8-core 64-thread 64b power-efficient SPARC SoC. In *IEEE Solid-State Circuits Conference (ISSCC)*, February 2007.

[10] S. Bell et al. TILE64 processor: A 64-core SoC with mesh interconnect. In *IEEE Solid-State Circuits Conference (ISSCC)*, February 2008.

[11] S. Borkar. Thousand core chips: A technology perspective. In *Proc. of Design Automation Conf. (DAC)*, June 2007.

[12] M. Hill and M. Marty. Amdahl's law in the multicore era. *IEEE Computer*, 41(7):33–38, July 2008.

[13] E. Chung et al. Single-chip heterogeneous computing: Does the future include custom logic, FPGAs, and GPUs? In *Proc. of International Symposium on Microarchitecture (MICRO)*, November 2010.

[14] W. Huang et al. Exploring the thermal impact on manycore processor performance. In *Proc. of IEEE Semi-Therm Symposium*, February 2010.

[15] W. Huang et al. Many-core design from a thermal perspective. In *Proc. of Design Automation Conference (DAC)*, June 2008.

[16] U. Karpuzku et al. The bubblewrap many-core: Popping cores for sequential acceleration. In *Proc. of International Symposium on Microarchitecture (MICRO)*, November 2009.

[17] J. Sampson et al. Efficient complex operators for irregular codes. In *Proc. of International Symposium on High-Performance Computer Architecture (HPCA)*, February 2011.

[18] G. Venkatesh et al. Conservation cores: Reducing the energy of mature computations. In *Proc. of International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, March 2010.

[19] http://domino.research.ibm.com/comm/research.nsf/pages/r.vlsi.innovation.html.

[20] R. Ross et al. High end computing revitalization task force (hecrtf), inter agency working group (heciwg) file systems and i/o research workshop report, 2006. http://institutes.lanl.gov/hec-fsio/docs/heciwgfsio-fy06-workshop-document-finalfinal. pdf.

[21] The International Technology Roadmap for Semiconductors (ITRS), 2010.

[22] K. Zhang (Ed.). *Embedded Memories for Nano-Scale VLSIs*. Springer Inc., 2009.

[23] A. Kahng et al. Orion 2.0: A fast and accurate NoC power and area model for early-stage design space exploration. In *Proc. of Design Automation and Test in Europe (DATE)*, April 2009.

[24] 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.

[25] K. Etessam-Yazdani et al. Impact of power granularity on chip thermal modeling. In *Proc. of 10th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM)*, June 2006.

[26] A. Keane. Active micro-channel cooling. In *HotChips 16, Session 4*, August 2004.

[27] D. Brooks et al. Wattch: A framework for architectural-level power analysis and optimizations. In *Proc. Intl. Symp. of Computer Architecture (ISCA)*, pages 83–94, June 2000.

[28] K. Skadron et al. Temperature-aware microarchitecture. In *Proc. Intl. Symp. on Computer Architecture (ISCA)*, pages 2–13, June 2003.

[29] M. Ware et al. Architecting for power management: The IBM POWER7 approach. In *Proc. of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, January 2010.