Hochschule
Augsburg University of
Applied Sciences

Post-Dennard Scaling and the final Years of Moore’s Law

Consequences for the Evolution of Multicore-Architectures

Prof. Dr.-Ing. Christian Märtin

Christian.Maertin@hs-augsburg.de

Technical Report
Hochschule Augsburg University of Applied Sciences
Faculty of Computer Science
September 2014

Post-Dennard Scaling and the final Years of Moore's Law

Consequences for the Evolution of Multicore-Architectures

Prof. Dr.-Ing. Christian Märtin

Hochschule Augsburg
Fakultät für Informatik
Telefon +49(0)821 5586-3454
christian.maertin@hs-augsburg.de

Fachgebiete

- Rechnerarchitektur
- Intelligente Systeme
- Mensch-Maschine-Interaktion
- Software-Technik

This paper undertakes a critical review of the current challenges in multicore processor evolution, underlying trends and design decisions for future multicore processor implementations. It is a short version of a paper presented at the Embedded World 2014 conference [1]. For keeping up with Moore´s law during the last decade, the VLSI scaling rules for processor design had to be dramatically changed. In future multicore designs large quantities of dark silicon will be unavoidable and chip architects will have to find new ways for balancing further performance gains, energy efficiency and software complexity. The paper compares the various architectural alternatives on the basis of specific analytical models for multicore systems.

Introduction

More than 12 years after IBM started into the age of multicore processors with the IBM Power4, the first commercial dual core processor chip, Moore´s law appears still to be valid as demonstrated by Intel´s fast track from 32 to 22nm mass production and towards its new 14nm CMOS process with even smaller and at the same time more energy efficient structures every two years [2]. At the extreme end of the performance spectrum, Moore´s law is also expressed by the industry´s multi-billion transistor multicore and many-core server chips and GPUs. Obviously the transistor raw material needed for integrating even more processor cores and larger caches onto future chips for all application areas and performance levels is still available.

However, in a similar way as the necessary transition from complex single core architectures with high operating frequencies to multicore processors with moderate frequencies was caused by the exponentially growing thermal design power (TDP) of the complex single core processors for reaching linear performance improvements [3], the ongoing multicore evolution has again hit the power wall and will undergo dramatic changes during the next several years [4]. As new analytical models and studies show [5], [6], power problems and the limited degree of inherent application parallelism will lead to rising percentages of dark or dim silicon in future multicore processors. This means that large parts of the chip have to be switched off or operated at low frequencies all the time. It has to be studied, whether the effects of such pessimistic forecasts will affect embedded applications and system environments in a milder way than the software in more conservative standard and high-performance computing environments.

In the following we discuss the reasons for these developments together with other future challenges for multicore processors. We also examine possible solution approaches to some of the topics. When discussing the performance of multicore systems, we must have a look on adequate multicore performance models that both consider the effects of Amdahl's law on different multicore architectures and workloads, and on the consequences of these models with regard to multicore power and energy requirements. We use the models also to introduce the different architectural classes for multicore processors. As will be shown, the trend towards more heterogeneous and/or dynamic architectures and innovative design directions can mitigate several of the expected problems.

Moore´s law and dark silicon

The major reason for the current situation and the upcoming trend towards large areas of dark silicon are the new scaling rules for VLSI design. Dennard's scaling rules [7] were perceived in 1974 and have held for more than 30 years until around 2005. As is well known, power in CMOS chips can be modeled as:

$$P = QfCV^2 + VI_{leakage} \qquad (1)$$

Q is the number of transistor devices, f the operating frequency of the chip, C the capacitance and V the operating voltage. The leakage current $I_{leakage}$ could be neglected until 2005 with device structures larger than 65nm.

With Dennard's scaling rules the total chip power for a given area size stayed the same from process generation to process generation. At the same time, with a scaling factor S=√2, feature size shrinked at a rate of *1/S* (the scaling ratio), transistor count doubled (Moore´s law) and the frequency increased by 40 % [5] every two years. With feature sizes below 65nm, these rules could no longer be sustained, because of the exponential growth of the leakage current. To lessen the leakage current, Intel, when moving to 45nm, introduced extremely efficient new Hafnium based gate insulators for the Penryn processor. When moving to 22nm, Intel optimized the switching process by using new 3D FinFET transistors that are currently used in the Haswell processors and will also be scaled down to 14nm.

However, even these remarkable breakthroughs could not revive the scaling of the operating voltage, because no further scaling of the threshold voltage is possible as long as the operating frequency is kept at the current already very low level. Therefore operating voltage has remained at a constant value of around 1 V for several processor chip generations.

With Post-Dennard scaling, like with Dennard scaling the number of transistors grows with S^2 and the frequency with S from generation to generation, i.e. the potential computing performance increases by S^3 or 2.8 between two process generations. Transistor capacitance also scales down to $\frac{1}{S}$ under both scaling regimes. However, as threshold and thus operating voltage cannot be scaled any longer, it is no longer possible to keep the power envelope constant from generation to generation and simultaneously reach the potential performance improvements. Whereas with Dennard scaling power remains constant between generations, Post-Dennard Scaling leads to a power increase of $S^2 = 2$ per generation for the same die area [8]. At the same time utilization of a chip's computing resources decreases with a rate of $\frac{1}{S^2}$ per generation.

This means that at runtime large quantities of the transistors on the chip have to be switched off completely, operated at lower frequencies or organized in completely different and more energy efficient ways. For a given chip area energy efficiency can only be improved by 40 % per generation. This dramatic effect, called dark silicon, already can be seen in current multicore processor generations and will heavily affect future multicore and many-core processors. Appendix A shows the amount of dark and dim (i.e. lower frequency) silicon for the next two technology generations.

These considerations are mainly based on physical laws applied to MOSFET transistor scaling and CMOS technology. However, the scaling effects on performance evolution are confirmed by researchers from industry. With the advances in transistor and process technology, new architectural approaches, larger chip sizes, introduction of multicore processors and the partial use of the die area for huge caches microprocessor performance increased about 1000 times in the last two decades and 30 times every ten years [4]. But with the new voltage scaling constraints this trend cannot be maintained. Without significant architectural breakthroughs Shekar Borkar based on Intel's internal data only predicts a six times performance increase for standard multicore based microprocessor chips in the decade between 2008 and 2018. This projection corresponds to the theoretical 40% increase per generation and over five generations in Post-Dennard scaling that amounts to S^5=5.38.

Therefore, computer architects and system designers have to find effective strategic solutions for handling these major technological challenges. Formulated as a question, we have to ask: Are there ways to increase performance by substantially more than 40% per generation, when novel architectures or heterogeneous systems are applied that are extremely energy-efficient and/or use knowledge about the software structure of the application load to make productive use of the dark silicon?

Modeling Performance and Power of Multicore Architectures

When Gene Amdahl wrote his famous article on how the amount of the serial software code to be executed influences the overall performance of parallel computing systems, he could not have envisioned that a very fast evolution in the fields of computer architecture and microelectronics would lead to single chip multiprocessors with dozens and possibly hundreds of processor cores on one chip. However, in his paper, he made a visionary statement, valid until today, that pinpoints the current situation: "the effort expended on achieving high parallel processing rates is wasted unless it is accompanied by achievements in sequential processing rates of very nearly the same magnitude" [10], [11].

Today we know, that following Pollack's rule [4], single core performance can only be further improved with great effort. When using an amount of *r* transistor resources, scalar performance rises to $Perf\ (r) = \sqrt{r}$. At the same time single core power *P* (or the TDP) rises with $P = Perf^{\alpha}$ [3] with $\alpha = 1.75$ or nearly the square of the targeted scalar performance. For the rest of this paper we assume that *P* includes dynamic as well as leakage power. Thus, there is a direct relation between the TDP of a sequential core, the transistor resources and the performance. Power can be expressed, as shown in [12] by:

$$P = Perf^{\alpha} = (\sqrt{r})^{\propto} = r^{\propto/2} \qquad (2)$$

When Grochowski derived his equation from comparing Intel processor generations from the i486 to the Pentium 4 Cedarmill (all of them normalized to 65nm technology), multicore evolution was slowly gathering momentum and the Post-Dennard scaling effects only were in their initial phase. With current technology, leakage currents and not further down-scalable supply voltages would forbid the use of the entire die area of mainstream microprocessor chips for a single core processor. Therefore analytical multicore models for existing or envisioned architectural alternatives have to consider the implications of different workload characteristics, both, to the performance, and the power requirements of these architectures.

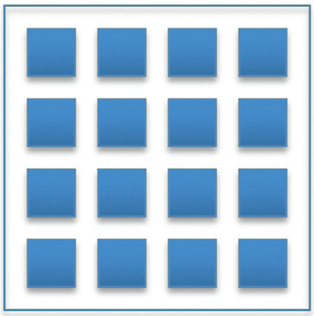


Fig. 1. Symmetric multicore processor: each block is a basic core equivalent (BCE) including L1 and L2 caches. L3 cache and onchip-network are not modeled. The computing performance of a BCE is normalized to 1.

Standard Models

In 2008 Hill and Marty started this discussion with performance models for three types of multicore architectures: symmetric (fig. 1), asymmetric (fig. 2), and dynamic (fig. 3) multicore chips [13] and evaluated their speedup models for different workloads, characterized by the respective degree *f* of code that can be parallelized as in Amdahl's law, where the Speedup *S* achieved by parallel processing is defined as:

$$S_{Amdahl}(f) = \frac{1}{(1-f)+\frac{f}{n}} \qquad (3)$$

In this equation *f* is the fraction of possible parallel work and n is the number of available cores. As Amdahl's law uses fixed application sizes, the achievable maximum speedup, whether multiprocessors or multicores are used, is always limited to $S_{max} = \frac{1}{1-f}$. 'With a fixed application size, the models presented by Hill and Marty also reach their performance limits relatively soon, and for workloads with $f < 0.99$ a large number of cores does not seem adequate.

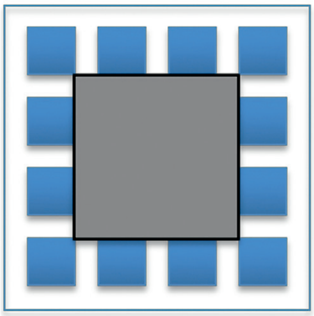


Fig. 2. Asymmetric multicore processor consisting of a complex core with *Perf* =√4 and 12 BCEs.

In the engineering and embedded system domains there are many standard and real time workloads with moderate data size, irregular control and data flow structures and with a limited degree of exploitable parallelism. For these workload characteristics, the easily understandable study of Hill and Marty with some extensions can lead to valuable insights of how different multicore architectures may influence the achievable performance.

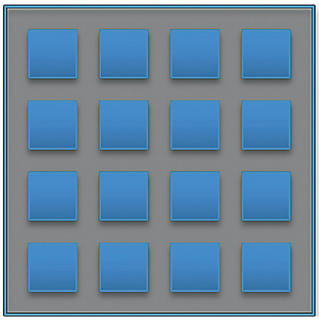


Fig. 3. Dynamic multicore processor: 16 BCEs or one large core with *Perf* =√16.

The idea behind the different models is that the given chip area allows for the implementation of n simple cores or so called basic core equivalents (BCEs) with a given performance of 1. To construct the architectural alternatives, more complex cores are modeled by consuming r of the n BCEs, leading to a performance *perf(r)*= $\sqrt{r}$ per core, if we go along with Pollack's performance rule for complex cores.

These considerations lead to the following speedup equations for the three alternatives:

$$S_{sym}(f,n,r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f \cdot r}{perf(r) \cdot n}} \qquad (4)$$

$$S_{asym}(f,n,r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{perf(r)+n-r}} \qquad (5)$$

$$S_{dyn}(f,n,r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{n}} \qquad (6)$$

Note, that for the asymmetric case in equation (5), the parallel work is done together by the large core and the n - r small cores. If either the large core, or the small cores would be running, to keep the available power budget in balance, the equation would change to:

$$S_{asym}(f,n,r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{n-r}} \qquad (7)$$

In addition to achievable performance, power (TDP) and energy consumption are important indicators for the feasibility and appropriateness of multi- and many-core-architectures with symmetric and asymmetric structures. Woo and Lee have extended the work of Hill and Marty towards modeling the average power envelope, when workloads that can be modeled with Amdahl's law are executed on multicore processors [14].

If we have a look at the evolution of commercial multicore processors, in addition to the already discussed architectural alternatives, we meanwhile can see new architectural variants for asymmetric systems (fig. 4), dynamic systems (fig. 5) and heterogeneous multicore systems (fig. 6).

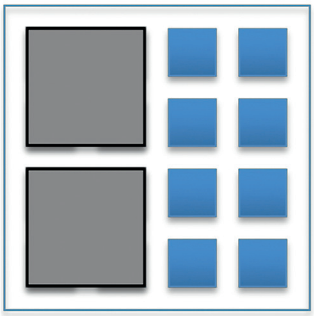


Fig. 4. Asymmetric multicore with 2 complex cores and 8 BCEs

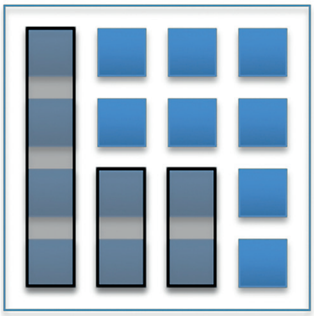


Fig. 5. Dynamic multicore with 4 cores and frequency scaling using the power budget of 8 BCEs. Currently one core is at full core TDP, two cores are at ½ core TDP. One core is switched off.

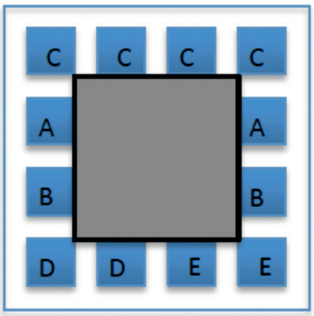


Fig. 6. Heterogeneous multicore with a large core, four BCEs, two accelerators or co-processors of type A, B, D, E, each. Each co-processor/accelerator uses the same transistor budget as a BCE.

Models for Heterogeneous Multicores

Heterogeneous architectures on first glance look similar to asymmetric multicores. However, in addition to a conventional complex core, they introduce unconventional or U-cores [12] that represent custom logic, GPU resources, or FPGAs. Such cores can be interesting for specific application types with SIMD parallelism, GPU-like multithreaded parallelism, or specific parallel algorithms that have been mapped to custom or FPGA logic. To model such an unconventional core, Chung et al. suggest to take the same transistor budget for a U-core as for a simple BCE core. The U-core then executes a specific parallel application section with a relative performance of μ, while consuming a relative power of φ compared to a BCE with $\mu = \varphi = 1$. If $\mu > 1$ for a specific workload, the U-core works as an accelerator. With $\varphi < 1$ the U-core consumes less power and can help to make better use of the dark silicon. The resulting speedup equation by Chung et al. can directly be derived from equation (7):

$$S_{heterogeneous}(f, n, r, \mu) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{\mu(n-r)}} \quad (8)$$

This equation is very similar to equation (7) that gives the speedup model for asymmetric multicore systems. In their evaluation, Woo and Lee [14] show that such asymmetric multicores are much more appropriate for power scaling than symmetric architectures, either with complex or simple cores. Going a step further, it follows that heterogeneous multicores will even scale better, if $\mu > 1$ and $\varphi < 1$ or at least $\varphi = 1$. As prototypical results in current research towards a better utilization of the dark silicon in logic chips demonstrate, future architectures might profit mostly in performance and energy consumption, if not only one, but many additional coprocessors or accelerator cores are added to one or a small number of large cores [15].

If, in addition to a conservative complex core, there are different accelerator/co-processor cores on an heterogeneous chip, they will typically be active on different parts of the entire application workload. Let us assume that the entire die area can hold resources for $n = r + r_1 + r_2 + \cdots + r_l$ BCEs with r denoting the resources for the complex core and r_i giving the resources for each specific coprocessor core. We can now derive a more flexible speedup equation that mirrors the real application structure and that can easily be used to model application-specific power and energy consumption:

$$S_{flexible} = \frac{1}{\frac{1-f}{perf(r)} + \sum_{i=1}^{l} \frac{f_i}{\mu_i \cdot r_i}} \quad (9)$$

The f_i are exploitable fractions of the application workload, , $f = f_1 + f_2 + \cdots + f_1$ with $f \leq 1$ and μ_i is the relative performance of a coprocessor, when using r_i BCE resources. If $\mu_i = 1$, the coprocessor can be interpreted as a cluster of r_i simple cores working in parallel.

Other Studies

In [16] it is shown, that there will be no limits to multicore performance scaling, if the application size grows with the number of available cores, as was stated first for classic multiprocessors by Gustafson [17]. Sun and Chen show that Gustafson's law also holds for multicore architectures and that the multicore speedup scales as a linear function of n, if the workload is also growing with the system size. They show, that even, if continuously growing parts of the workload will reside in slow DRAM memory, unlimited speedup scaling is still possible, as long as the DRAM access time is a constant. Note, that without scaling up the workload of the LINPACK benchmark, the multi-million-core supercomputers presented in the Top 500 list twice every year, would not make sense at all.

A recent and much more pessimistic study by Esmaeilzadeh et al. [5] constructs a very fine grained performance model for future multicore processors. The model includes separate partial models for Post-Dennard device scaling, microarchitecture core scaling, and the complete multicore. The model's starting point is the 45nm i960 Intel Nehalem quadcore processor that is mapped to all future process generations until a fictitious 8nm process. Following optimistic ITRS scaling, the downscaling of a core will result in a 3.9 x core performance improvement and an 88% reduction in core power consumption. However, if the more conservative projections by Shekar Borkar are applied, the performance will only increase by 34% with a power reduction of 74%. The model is evaluated using the 12 real benchmarks of the PARSEC suite for parallel performance evaluation.

The results of the study show that the entire future 8nm multicore chip will only reach a 3.7 x (conservative scaling) to 7.9 x speedup (ITRS scaling). These numbers are the geometric mean of the 12 PARSEC benchmarks executed on the model. The study also shows that 32 to 35 cores at 8nm will be enough to reach 90% of the ideal speedup. Even though the benchmarks contain relatively much exploitable parallelism, there is not enough parallelism to utilize more cores, even with an unlimited power budget. For more realistic standard applications, the speedup would be limited by the available power envelope, rather than by exploitable parallelism. The power gap would lead to large fractions of dark silicon on the multicore chip. The executable model by Esmaeilzadeh et al. can be executed for different load specifications and hardware settings. The authors have made it available at a URL at the University of Wisconsin [6].

Although this model was devised very carefully and on the basis of available facts and plausible future trends, it cannot predict future breakthroughs that might influence such hardware parameters that are currently treated like immovable objects. An example is the slow DRAM access time, which is responsible for the memory wall and has existed for decades. In Esmaeilzadeh's model it is kept constant through to the 8nm process, whereas off-chip memory bandwidth is believed to scale linearly. Currently memory makers like Samsung and HMC are working on disruptive 3D-DRAM technologies that will organize DRAM chips vertically and speedup access time and bandwidth per chip by a factor of ten, without needing new DRAM cells [18]. Such and other technology revolutions might affect all of the discussed models and could lead to more optimistic multicore performance scenarios. Read the full paper for a discussion of the ongoing evolution of commercial multicore processors, current research trends, and the consequences for software developers [1].

Conclusion

As VLSI scaling puts new challenges on the agenda of chip makers, computer architects and processor designers have to find innovative solutions for future multicore processors that make the best and most efficient use of the abundant transistor budget offered by Moore's law that will hopefully be valid for another decade. In this paper, we have clarified the reasons of the Post-Dennard scaling regime and discussed the consequences for future multicore designs.

For the four typical architectural classes: symmetric, asymmetric, dynamic, and heterogeneous multicore processors, we have compared and extended some straightforward performance and power models. Although various results from research and industrial studies predict that huge parts of future multicore chips will be dark or dim all the time, it can already be seen in contemporary commercial designs, how these areas can be dealt with for making intelligent use of the limited power budget with heterogeneous coprocessors, accelerators, larger caches, faster memory buffers, and improved communication networks.

In the future, the success of many-core architectures will first and foremost depend on the exact knowledge of specific workload properties as well as easy-to-use software environments and software developing tools that will support programmers to exploit the explicit and implicit regular and heterogeneous parallelism of the application workloads.

References

[1] Märtin, C.,"Multicore Processors: Challenges, Opportunities, Emerging Trends", Proceedings Embedded World Conference 2014, 25-27 February, 2014, Nuremberg, Germany, Design & Elektronik, 2014

[2] Bohr, M., Mistry, K., "Intel's revolutionary 22 nm Transistor Technology," Intel Corporation, May 2011

[3] Grochowski, E., "Energy per Instruction Trends in Intel Microprocessors," Technology@Intel Magazine, March 2006

[4] Borkar, S., Chien, A.A.,"The Future of Microprocessors," Communications of the ACM, May 2011, Vol. 54, No. 5, pp. 67-77

[5] Esmaeilzadeh, H. et al., "Power Challenges May End the Multicore Era," Communications of the ACM, February 2013, Vol. 56, No. 2, pp. 93-102

[6] Esmaeilzadeh et al., "Executable Dark Silicon Performance Model": http://research.cs.wisc.edu/vertical/DarkSilicon, last access Jan. 20, 2014

[7] Dennard, R.H. et al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," IEEE J. Solid-State Circuits, vol. SC-9, 1974, pp. 256-268

[8] Taylor, M.B., "A Landscape of the New Dark Silicon Design Regime," IEEE Micro, September/October 2013, pp. 8-19

[9] Borkar, S., “3D Integration for Energy Efficient System Design,” DAC’11, June 5-10, San Diego, California, USA

[10] Amdahl, G.M., “Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities,” Proc. AFIPS Conf., AFIPS Press, 1967, pp. 483-485

[11] Getov, V., “A Few Notes on Amdahl’s Law,” IEEE Computer, December 2013, p. 45

[12] Chung, E.S., et al., “Single-Chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?” Proc. 43rd Annual IEEE/ACM International Symposium on Microarchitecture, 2010, pp. 225-236

[13] Hill, M.D., Marty, M., “Amdahl’s Law in the Multicore Era,” IEEE Computer, July 2008, pp. 33-38

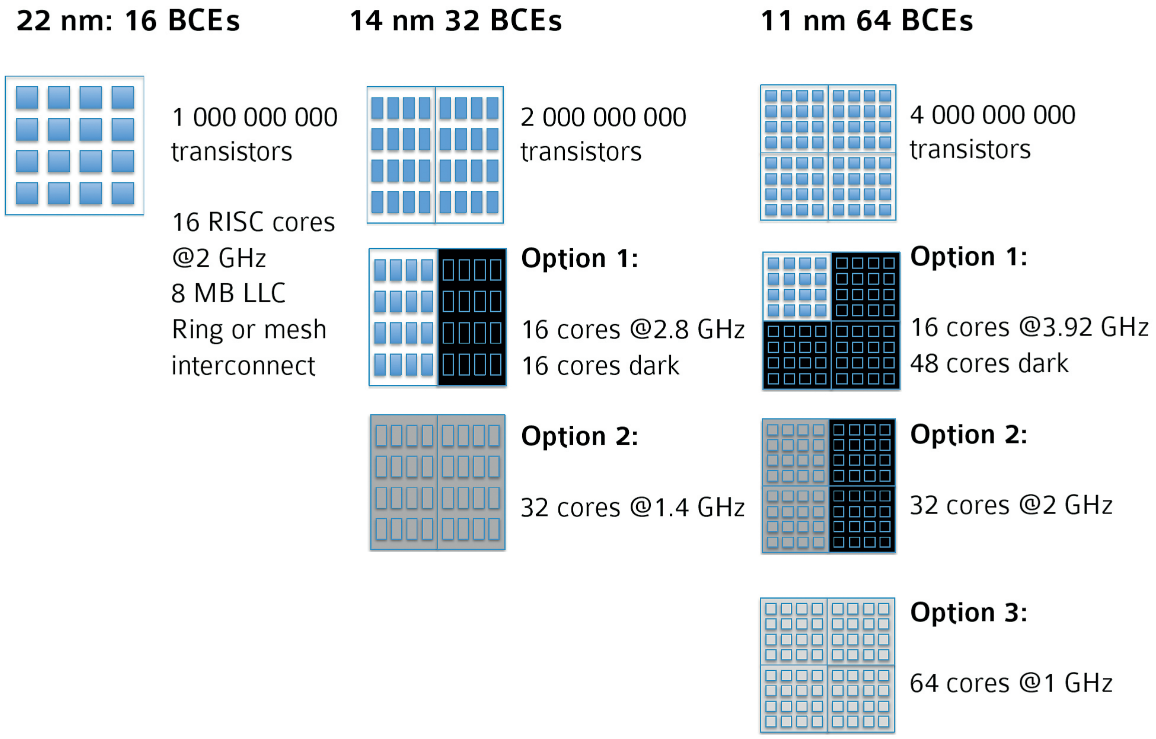
[14] Woo, D.H., Lee, H.H., “Extending Amdahl’s Law for Energy-Efficient Computing in the Many-Core Era,” IEEE Computer, December 2008, pp. 24-31

[15] Venkatesh, G. et al., “Conservation Cores: Reducing the Energy of Mature Computations,” Proc. 15th Architectural Support for Programming Languages and Operating Systems Conf., ACM, 2010, pp. 205-218

[16] Sun, X.-H., Chen, Y., ”Reevaluating Amdahl’s law in the multicore era,” J. Parallel Distrib. Comput. (2009), doi: 10.1016/j.jpdc2009.05.002

[17] Gustafson, J.L, “Reevaluating Amdahl’s Law,” Communications of the ACM, Vol. 31, No. 5, 1988, pp. 532-533

[18] Courtland, R., “Chipmakers Push Memory Into the Third Dimension,” http://spectrum.ieee.org, last access: January 20, 2014



Appendix A: Dark and dim silicon in symmetric multicore processor generations from 22nm to 11nm. Each BE represents a RISC core with 8 MB last-level cache. Communication infrastructure is included in the transistor count.