embedded world 2014
Exhibition & Conference
... it´s a smarter world

Multicore Processors: Challenges, Opportunities, Emerging Trends

Christian Märtin
Faculty of Computer Science
Augsburg University of Applied Sciences
Augsburg, Germany
maertin@ieee.org

Abstract— **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 first shown, that 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. Examples of leading commercial multicore processors and architectural research trends are given to underscore the dramatic changes lying ahead in computer architecture and multicore processor design.**

Keywords—multicore processor, Moore's law, Post-Dennard scaling, multicore architectures, many-core architecture, multicore performance models, dark silicon

I. 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, software and system developers as well as end users of business, engineering and embedded applications still take it for granted, that the performance gains delivered by each new chip generation maintain a more than linear improvement over the decade ahead. 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 [1].

Very successfully and 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 [2], the ongoing multicore evolution has again hit the power wall and will undergo dramatic changes during the next several years [3]. As new analytical models and studies show [4], [5], 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 this paper 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 first 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.

The paper will therefore give an overview of the most promising current architectures and predictable trends and will finally point at some typical implementations of server, workstation, and embedded multicore chips. Multicore processor implementations in the same architectural class may vary significantly depending on the targeted application domain and the given power budget. As will be shown, the trend towards more heterogeneous and/or dynamic architectures and innovative design directions can mitigate several of the expected problems.

II. 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 [6] 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 $S = \sqrt{2}$, feature size shrinked at a rate of $1/S$ (the scaling ratio)[1], transistor count doubled (Moore´s law) and the frequency increased by 40 % [5] e very 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 isolators 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 [7]. 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.

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 [3]. But with the new voltage scaling constraints this trend cannot be maintained. Without significant architectural breakthroughs Intel's Shekar Borkar only predicts a 6 times performance increase for standard multicore based microprocessor chips in the decade between 2008 and 2018. Interestingly enough, this projection – although directly derived from the company's planned logic transistor budget, planned cache size, 2x increase in operating frequency and 3x increase by redesigned transistor devices – 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?

The rising amount of transistors used for on-chip L3 caches was the first reaction of multicore chip makers to counter this unavoidable situation. Larger caches need less than $1/10$ of the power of logic transistor blocks. But there are limits to the size of useful caches. Another reaction could be hierarchical on-chip networks, so far not used on processor dies, that would keep communication costs and power in many-core chips as low as possible [3]. Another direction could be the 3D organization of multicore processing chips with eDRAM dies, as Intel has demonstrated in [8].

III. 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" [9], [10].

[1] An error was removed. The original text was: "At the same time feature size shrinked at a rate of $S = 1/\sqrt{2}$"

Today we know, that following Pollack's rule [3], 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 $P = Perf^{\alpha}$ [2] 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 [11] 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 downscalable 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.

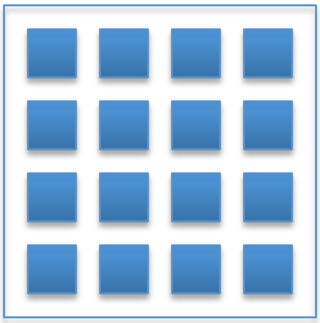

Fig. 1. Symmetric multicore processor: each block is a basic core equivalent (BCE) including L1 and L2 caches. L3 cache and onchip-network are not modeled. The computing performance of a BCE is normalized to 1.

A. 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 [12] 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.

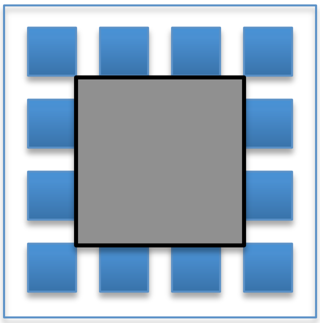

Fig. 2. Asymmetric multicore processor consisting of a complex core with $Perf = \sqrt{4}$ and 12 BCEs.

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.

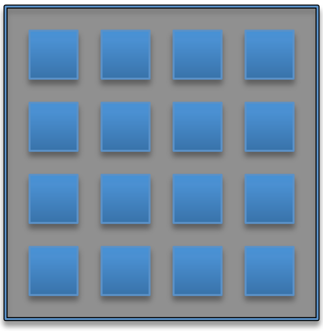

Fig. 3. Dynamic multicore processor: 16 BCEs or one large core with $Perf = \sqrt{16}$.

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 [13] 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. The equations for the symmetric case are given here:

$$W = \frac{1+(n-1)k(1-f)}{(1-f)+\frac{f}{n}} \qquad (8)$$

In this equation W is the average power consumption, k is the fraction of power that one core consumes in idle state, and the power of a core in active state is 1. For $\frac{Performance}{Power}$ we get

$$\frac{Perf}{W} = \frac{1}{1+(n-1)k(1-f)} \qquad (9)$$

However, 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).

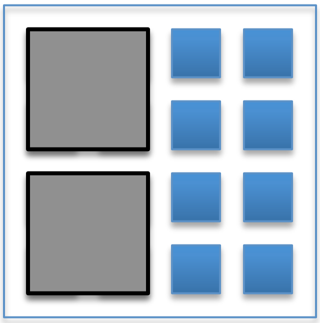

Fig. 4. Asymmetric multicore with 2 complex cores and 8 BCEs

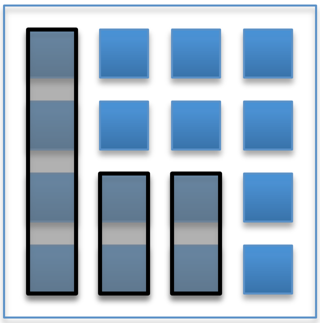

Fig. 5. Dynamic multicore with 4 cores and frequency scaling using the power budget of 8 BCEs. Currently one core is at full core TDP, two cores are at ½ core TDP. One core is switched off.

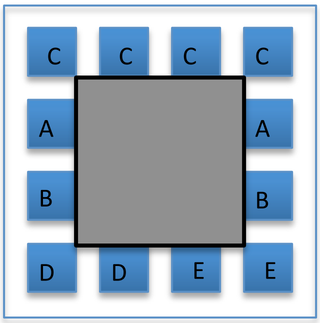

Fig. 6. Heterogeneous multicore with a large core, four BCEs, two accelerators or co-processors of type A, B, D, E, each. Each co-processor/accelerator uses the same transistor budget as a BCE.

B. 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 [11] 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):

$$S_{heterogeneous}(f,n,r,\mu) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{\mu(n-r)}} \qquad (10)$$

This equation is very similar to equation (7) that gives the speedup model for asymmetric multicore systems. In their evaluation, Woo and Lee [13] 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 [14].

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:

$$S_{flexible} = \frac{1}{\frac{1-f}{perf(r)} + \sum_{i=1}^{l} \frac{f_i}{\mu_i \cdot r_i}} \qquad (11)$$

The f_i are exploitable fractions of the application workload, $f = f_1 + f_2 + \cdots + f_l$ 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.

C. Other Studies

In [15] 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 [16]. 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. [4] 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× 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× (conservative scaling) to 7.9× 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 [5].

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 [17]. Such and other technology revolutions might affect all of the discussed models and could lead to more optimistic multicore evolution scenarios.

IV. Multicore Evolution

In order to be able to give meaningful answers to the challenging questions asked in the earlier chapters, we will now examine the current state and the ongoing evolution of multicore processor architectures.

Industry has successfully maintained Moore´s law for nearly 50 years. In the last decade commercial processor chips have arrived at a transistor count of several billion transistor devices. However, depending on the application domain and context, there exist various architectural approaches in current and planned multicore and many-core processors. In the following we will examine some important state-of-the art multicore processors as examples for the classes defined in the last chapter: symmetric, asymmetric, dynamic and heterogeneous multicore processors.

A. Symmetric Multicore Processors

During the last three technology generations (45nm to 22nm) the number of on-chip cores has not changed dramatically for mainstream and high-end server systems by Intel, IBM, Fujitsu, Oracle, and AMD. Again core microarchitecture performance and energy efficiency were improved and larger last-level caches were implemented. Much effort by all contenders is put into the memory system bandwidth optimization. Fast buffer caches are inserted between the processor cores and the memory controllers.

In the embedded sector we see a differentiation to very small and energy-efficient multicore chips on the one hand, and high-performance multicore implementations with mainstream-like micro-architectural features. With its new 22nm Atom multicore processors, in 2013, Intel has aggressively challenged the ARM-dominated market for tablet, and smartphone chips. A different picture can be seen in the area of symmetric many-core processors for embedded systems, servers, and high-performance computing. Up to 72 cores are or will be integrated in many-core products by Intel and Tilera. Implementations of research prototypes integrate even more simple cores onto the dies. Some of the research prototypes also demonstrate, how more efficient memory architectures could be constructed. In the following only a few examples of current designs can be given.

Intel Haswell and Broadwell. Intel's current microarchitecture for mainstream desktop and server processors is called Haswell [18]. It is manufactured using 22nm process technology and 3D FinFET transistors. The die size of the standard quad core Core i7 desktop chip including 8 MB of shared L3 cache, a complex graphics accelerator, fast video and audio transcoding, and 3D support is 177m^2 and integrates 1.4 billion transistors. Compared to its Sandy Bridge predecessors, the Haswell quad core chips offer a lower TDP (84 W) and an increased maximum turbo frequency of up to 3.9 GHz. Even higher frequencies are planned for 2014. The new Turbo Boost 2.0 technology uses thermal monitoring technologies and allows a higher maximum package temperature than older Core i7 chips. An eight core desktop version is planned for 2014.

The Haswell microarchitecture again uses simultaneous multithreading (Hyperthreading), allows to issue up to eight decoded instructions per cycle, and offers larger internal buffers for all stages of the superscalar pipeline compared to its predecessors. For instance, the reorder buffer (out-of-order window) now holds up to 192 instructions (Sandy Bridge: 168, Nehalem 128). The AVX2 instruction set extension offers 256-bit wide registers for enhanced SIMD parallelism. The instruction set was also extended by fast multiply-add, integer numeric, and encryption instructions. The microarchitecture has again improved its branch prediction accuracy.

The cores on the die are connected to each other and to the slices of the L3-cache and the graphics engine by a fast ring interconnect. Haswell also supports a large off-die L4 cache that is implemented in eDRAM technology and can optimize graphics and gaming performance. Most of Intel's current Xeon server chips are still manufactured using the 32nm process that allows to integrate up to ten cores. However, the Xeons already offer two memory controllers, four DRAM channels (i.e. around 100 GB/s of memory bandwidth per chip), and four Quick-Path-Interconnect-Links per chip. Haswell-based Xeon server chips in 22nm technology will be available in 2014 with up to 14 cores and cache sizes of up to 45 MB.

Multicore processors manufactured in Intel's new 14nm process are codenamed Broadwell. Early Broadwell chips will arrive in the form of desktop processors in late 2014. For 2015 Intel is planning a Broadwell-based Xeon EP chip with 18 cores.

AMD Kaveri with Steamroller microarchitecture. AMD's latest desktop processor Kaveri is manufactured in 28nm technology and packs 2.41 billion transistors on its die area of 245mm^2. At 4 GHz its TDP varies between 45W and 85W depending on the graphics load. The dual core processor is targeted at the midrange desktop and gaming market and is the first AMD chip to implement the new Steamroller microarchitecture [19]. A huge GPU (Graphics Core Next, GCN) with 512 shaders grouped in eight shader cores occupies 47% of the die area. Compute cores and GPU equally share up to 32 GB of main memory. Each dual-threaded compute core is organized in two integer modules and one FPU. Each integer module is fed by a decoder with up to four micro-instructions per cycle. The two decoders also feed the FPU. Each integer module offers four independent arithmetic pipelines working concurrently. The Steamroller microarchitecture will also be used in future eight or more core high-end server chips codenamed Berlin which will include five HyperTransport-links per chip for even better NUMA support.

IBM Power8. In September 2013 IBM has announced its new 12 core Power8 high-end server chip that will be manufactured in 22nm and enter the market in 2014 [20]. The chip will offer a 4 GHz operating frequency. The cores use a modified version of the Power7+ microarchitecture. A remarkable step forward is the support of eight simultaneous threads. Each core has a 512 KB L2 cache. All cores share a 96 MB L3 on-chip cache that is implemented in space-efficient eDRAM technology. At 4 GHz single thread performance was improved by 60% compared to the Power7+ core.

Each Power8 chip has eight high-speed memory channels, each of which is connected to a Centaur memory buffering chip that puts 16 MB L4 cache of eDRAM between the memory channel coming from the processor chip and the DDR3-1600 DRAM controller residing on the Centaur chip. Each Power8 chip can offer a sustained memory bandwidth of up to 230 GB/s with a peak bandwidth of 410 GB/s. In Power8 servers IBM will support transactional memory that accesses shared memory speculatively without locks and semaphores. If other processes have written on the same memory location the transaction has to be made undone and repeated. For typical database loads, transactional memory will lead to massive performance boosts.

Oracle Sparc T5. The Sparc T5 was announced in 2012. It is clocked at 3.6 GHz, manufactured in 28nm technology and integrates 16 cores that are interconnected by a crossbar switch to the eight banks of the shared 8MB L3 cache [21]. Core count per chip was doubled compared to the T4 predecessor. Core speed was enhanced by 20 percent. Each core can manage up to eight threads dynamically (round robin). The T5 offers instruction set extensions for data encryption. Four coherence units are positioned between the L3 cache and memory controllers to handle cache-misses and DMA to the I/O subsystem. Up to eight chips can be interconnected using a glueless one-hop interconnect that guarantees coherence between all eight sockets and up to 128 cores.

Fujitsu Sparc64 X. This chip was introduced in 2012 and contains 16 Sparc64 X processor cores. The chip is produced in 28nm technology with an enormous die size of 540mm^2 and a transistor count of 2.950 billion. The operating frequency is 3 GHz. The cores contain new hardware for a decimal and a cipher engine. Parts of these sections contain accelerator hardware for specific target functions on UNIX servers. This hardware is called SWoC and improves performance for specific decimal arithmetic and encryption functionality by a factor of up to 104 leading to a 120× better performance/Watt ratio [22].

ARM Cortex-A processors. This series of (multicore) processors offers a broad range of IP solutions from three-issue superscalar high-performance systems (e.g. ARM Cortex-A15) for mobile computing, digital home, servers and wireless infrastructures down to the extremely energy-efficient A7 and A5 cores [23]. The Cortex-A chips use advanced microarchitectures with relatively long pipelines (up to 13 stages) and efficient branch prediction hardware. For multicore support the processors are equipped with a coherent bus interface and specific hardware for multicore debug and trace purposes. The Cortex-A15 integrates 32KB L1 data and instruction caches and supports an L2-cache size of up to 4MB.

The 64-bit ARM V8 architecture was announced in 2011. Meanwhile several 64-bit implementations are available. 64-bit ARM multicore processors like the ARM Cortex-A53 are targeted at the high-end market for heterogeneous platforms including FPGA functionality and for the server market. AMD's most powerful core is the Cortex-A57 with a deep three-issue out-of-order superscalar pipeline and eight dispatch-ports (like Intel's Haswell) and an out-of-order instruction-window with size 128. AMD is planning a micro-server equipped with 16 of these ARM Cortex-A57 cores.

Intel Atom. Intel's new 64-bit Atom processors are manufactured in 22nm and are targeted both at the market for high-end smart phones and tablets [18]. The Silvermont architecture directly attacks the fastest ARM Cortex-A processors. Each core comes with a dual-issue out-of-order superscalar microarchitecture, an on-chip GPU and a very low TDP of 1.5 W. For laptops, smart phones, and tablets the Z3770D quad-core implementation with 1.5 GHz operating frequency and an on-chip GPU is already available. For embedded or large scale, low energy servers, Intel offers Atoms C2750, C2550, and C2350 codenamed Avoton with eight, four, and two cores, and a TDP of 20W, 14 W, and 6 W, respectively. Their maximum turbo frequency is 2.6 GHz. The C2750 supports memory sizes of up to 64 GB and with two memory channels reaches a maximum memory bandwidth of 25.6 GB/s.

Intel Many Integrated Cores (MIC) processors. Starting with a 32 core prototype in 2012 [24], Intel has entered the many-core market segment in 2013 with the first implementation of the Xeon PHI multicore processor, the 22nm Knight's Corner. This multicore chip operates as an accelerator co-processor in Xeon-based workstations, servers, and supercomputers. Communication to the host uses a PCI express 2.0 interface. The chip offers up to 61 64-bit x86 cores that are based on the Pentium platform. Each core comes with 512 KB of L2-cache and a completely new 512-bit vector-processing unit with a 3-address vector instruction ISA extension. Each core supports four dynamic (round robin) threads. The cores and caches are ring interconnected. The ring interface also accesses eight GDDR5 memory controllers with two memory channels each. Each chip can access 16 GB of GDDR5 memory and offers a maximum memory bandwidth of 352 GB/s. All cores are equipped with Turbo Boost 1.0 technology with a maximum turbo frequency of 1.333 GHz. The maximum TDP of the package is 300 W. In contrast to GPGPU accelerator chips, like NVIDIA's Fermi [24], which need the CUDA programming environment and cannot be seen as general purpose computers, the Intel chips can be programmed with standard languages like C++ and C# and fully support Intel and Open Source development environments like OpenMP, Intel TBB and Cilk.

For 2015 Intel has announced the 14nm MIC Knight's Landing with the same TDP of 300W, but up to 72 cores based on the 2-issue superscalar Atom Silvermont embedded platform architecture that offers many Ivy-Bridge-like microarchitecture features, 24 KB data, 32 KB instruction and 1 MB L2-caches. These new MICs will able to work as standalone systems, but will be connected with other MICs or system components over fast QPI channels. The system is targeted at a maximum performance of 3 TFLOPS, i.e. 10 $GFLOPS/W$.

Tilera TILE-Gx. Tilera is an MIT spin-off that has specialized on the development of symmetric many-core processors originally derived from the 16-core MIT RAW processor. The largest available chip, the TIL-Gx8072 [25] is implemented in 40nm technology and includes 72 efficient, 2D-mesh-connected 64 RISC cores with 32 KB L1-data and instruction caches and 256 KB L2 caches. In addition the chip offers 18 MB of coherent L3 cache. The 2D-mesh connection network consists of five independent concurrent networks. Switching occurs at a speed of one hop from core to core within one cycle. Operating frequency of the cores is 1.0 GHz. The full aggregate communication bandwidth surpasses 100 Tbps. In addition the chip includes four DDR3 memory controllers allowing up to 1 TB memory, six high-performance transaction ports for chip-to-chip or FPGA interconnect, special hardware for packet-switched communication and accelerators for encryption and security functions.

B. *Techniques for Asymmetric and Dynamic Multicore Processing*

Many of the processors discussed above, already offer techniques for dynamic processing or support asymmetric designs. Current research activities of the leading manufacturers will lead to even more efficient design solutions.

Current asymmetric multicore designs built around the ARM architecture for more flexible and even more energy-efficient embedded and fine-tuned system behavior, use a combination of one or two complex ARM cores (Cortex-A15 or Cortex-A57) and up to four extremely energy-efficient cores (Cortex-A17 or Cortex-A53). The cores can be integrated on an SoC in a so-called big.LITTLE configuration together with additional heterogeneous cores, e.g. for signal processing. They communicate via the cache-coherent CoreLink CCI-400 interconnect. To the software they appear as an homogeneous multicore processor. In addition big.LITTLE systems also show dynamic behavior, as the big cores, as well as the little cores offer different frequency (and voltage) levels and can be completely switched off, if not needed [23]. For high-end configurations system developers can also couple several Cortex-A57 and A53 cores.

The idea of dynamically changing the performance level of multicore chips as modeled by Hill and Marty was implemented in second-generation multicores in the form of dynamic voltage and frequency scaling (DVFS). Intel's Turbo Boost 1.0 allows the balancing of the core frequencies of the available cores against the available thread parallelism. Cores can completely be switched off and the frequency of the remaining threads can be raised, as long as the maximum TDP and frequency limit is respected.

Today, the operating voltage level of transistors in processors for most market segments is already very low and only 2.5× the threshold level of the transistors. DVFS downward scaling would result in a rapid drop of frequency and performance [7] and is an option only for applications with exactly known performance envelopes. Nevertheless Near-Threshold Voltage (NTV) logic is currently actively researched. Intel has developed an NTV x86 processor [26], [27] that shows a remarkable DVFS flexibility and can be seen as a prototypical core for future extremely energy-efficient multicore systems.

Currently Intel's mainstream processors use Turbo Boost 2.0 for software hot spot and single thread acceleration. This technique temporarily boosts core performance by frequency scaling for several seconds until the processor reaches its nominal temperature and as long as the thermal capacitance of the chip's heat sink is not overstrained. After this, the processor falls back to a relatively low power level. Similar techniques are applied by the four ARM Cortex-A15 cores used in big.LITTLE configurations. With future phase-change materials, an even higher temporary performance leverage, also called computational sprinting, will be possible [7].

Intel's Haswell chips offer fine-grained power management by so-called fully integrated voltage regulators on the chip that allow the separate voltage and frequency control of the cores, the integrated GPU, the L3 cache and the system interface. IBM's Power8 offers separate voltage and frequency control for each core.

C. Heterogeneous Multicore Processors

Many multicore products are offered as IP cores that can be used as building blocks for designing complex custom or FPGA-based heterogeneous multicore systems. ARM, Texas Instruments, MIPS, Freescale, Altera, Xilinx and other vendors offer solutions for various target markets that include mobile IT, automotive, manufacturing, and other areas. In the following, out of a very rich landscape, we only give two examples of typical heterogeneous designs.

Freescale QorlQ T Series. The T4240 is a 12-core processor based on Power Architecture e6500 64-bit dual-threaded cores with a clockspeed of up to 1.8 GHz [28]. The chip supports industrial embedded applications, datacenter loads, and routing applications with high-performance and low power requirements. The cores come in banks of four cores. Each bank shares 2 MB of L2 cache. The chip has three integrated DDR3 memory controllers with a 1.5 MB prefetch-cache. What makes the chip a typical heterogeneous multicore architecture is its Data Path Acceleration Hardware (DPAA) that contains an I/O manager and a frame manager that parses headers of incoming packages and assigns the packages to buffers. The queue manager assigns the packages to cores or to specific hardware that include security accelerators (SEC), a pattern matching engine (PME), and a data compression engine.

Altera Stratix 10. Altera is known for its high-end FPGA solutions [29]. The recently announced Stratix 10 SoC will be fabricated by Intel and uses the new 14nm process using FinFET transistors. Intel will integrate four ARM Cortex-A53 cores on the chip. Together with integrated FPU accelerators the multicore hardware will enable the design of very flexible applications with high-performance, as the FPGA will operate at 1 GHz, and with a low power budget.

V. 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 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 used 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.

VI. References

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

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

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

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

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

[6] 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

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

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

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

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

[11] 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

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

[13] 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

[14] 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

[15] 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

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

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

[18] http://www.intel.com

[19] http://www.amd.com

[20] Prickett Morgan, T., “Power8 Processor Packs a Twelve-Core Punch-And Then Some,” The Four Hundred, http://www.itjungle.com/tfh/thf090913-printer01.html, last access Jan. 20, 2013

[21] Feehrer, J. et al., “The Oracle Sparc T5 16-Core Processor Scales to Eight Sockets,” IEEE Micro, March/April 2013, pp. 48-57

[22] Yoshida, T. et al., “Sparc64 X: Fujitsu’s New-Generation 16-Core Processor for Unix Servers,” IEEE Micro, November/December 2013, pp. 16-24

[23] http://www.arm.com

[24] Heinecke, A., Klemm, M., Bungartz, H.-J., “From GPGPU to Many-Core: Nvidia Fermi and Intel Many Integrated Core Architecture,” IEEE Computing in Science & Engineering, March/April 2012, pp. 78-83

[25] http://www.tilera.com

[26] Ruhl, G.,”IA-32 Processor with a Wide-Voltage-Operating_Range in 32 nm CMOS,” IEEE Micro, March/April 2013, pp. 28-36

[27] Kaul, H. et al.,”Near-Threshold Voltage (NTV) Design – Opportunities and Challenges”, Design Automation Conference ’49, June 3-7, 2012, San Francisco

[28] http://www.freescale.com

[29] http://www.altera.com