Post-Dennard-Scaling und die Evolution von Multicore-Architekturen

Prof. Dr.-Ing. Christian Märtin
Computer Architecture and Intelligent
Systems
Faculty of Computer Science
Augsburg University of Applied
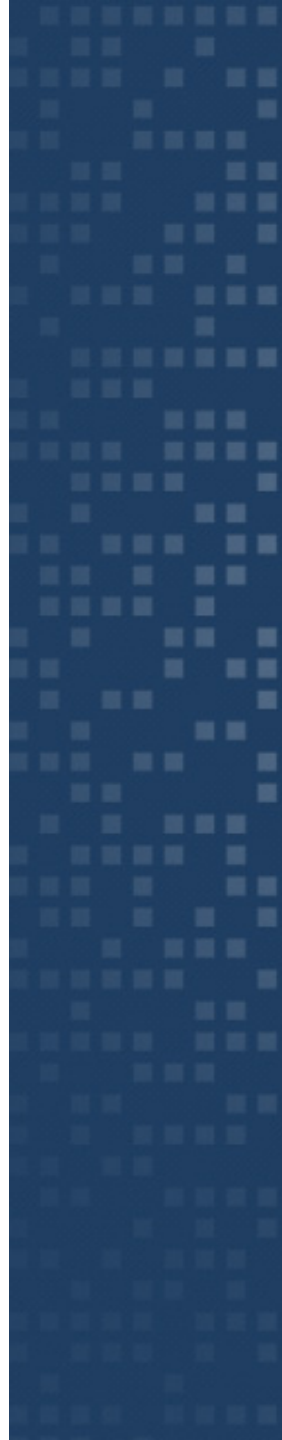**Sciences**
maertin@ieee.org

HOCHSCHULE
AUGSBURG

Content

- Introduction
- Moore's Law and Dark Silicon
- Multicore Architectures: Classes, Performance and Power Modeling
- Multicore Evolution
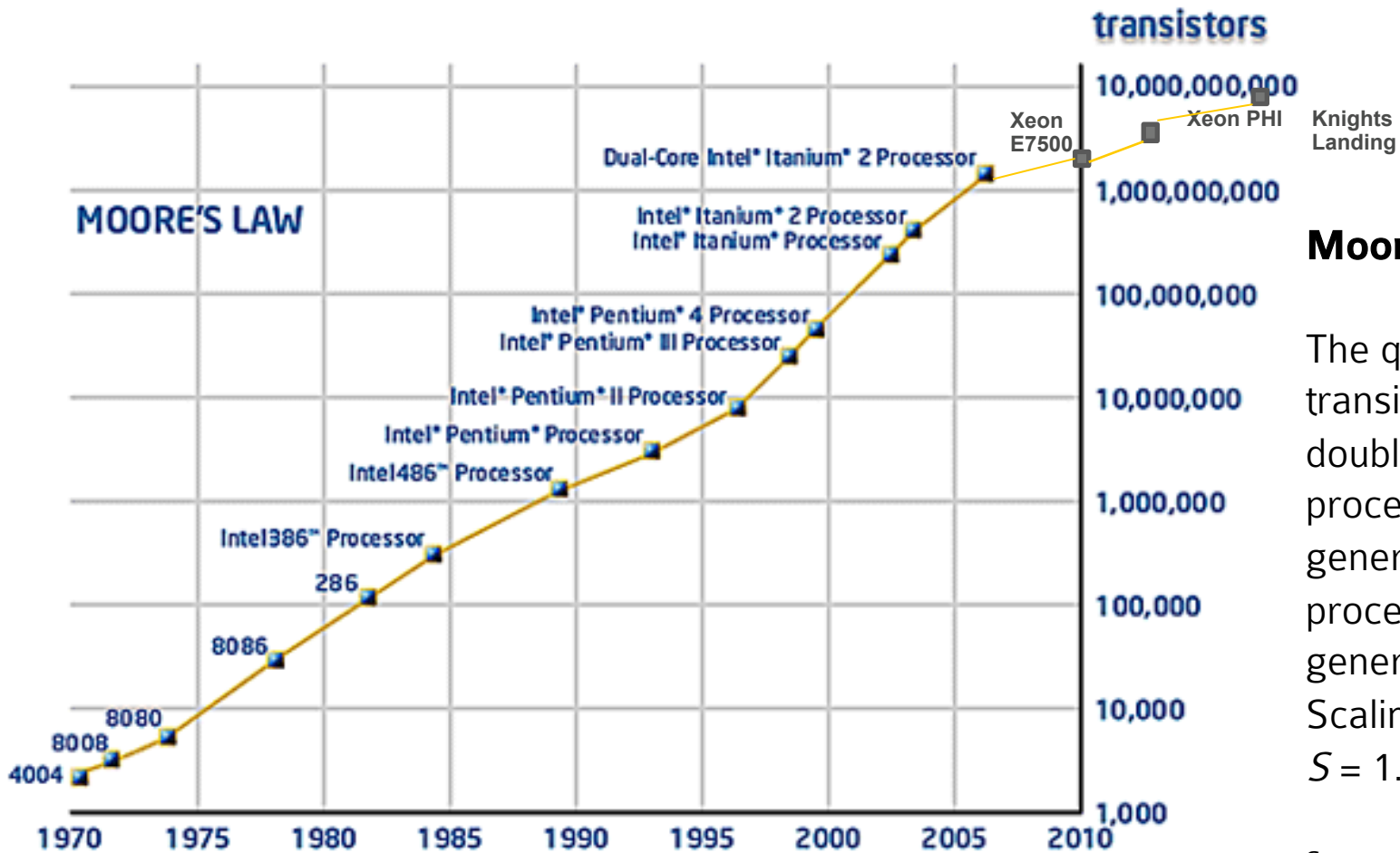- Conclusion

Introduction

Introduction

- Moore's law is still active
- Major chip makers strive hard, but successfully to solve the problems of further transistor scaling
  - strained silicon
  - high-k gate insulators
  - FinFET transistor
- Will the multicore evolution continue or will it hit the power, parellelization, and memory walls?

Moore's Law and Dark Silicon

transistors

10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000

MOORE'S LAW

Xeon E7500
Xeon PHI
Knights Landing

Dual-Core Intel® Itanium® 2 Processor

Intel® Itanium® 2 Processor
Intel® Itanium® Processor

Intel® Pentium® 4 Processor
Intel® Pentium® III Processor

Intel® Pentium® II Processor

Intel® Pentium® Processor

Intel486™ Processor

Intel386™ Processor

286

8086

8080

8008

4004

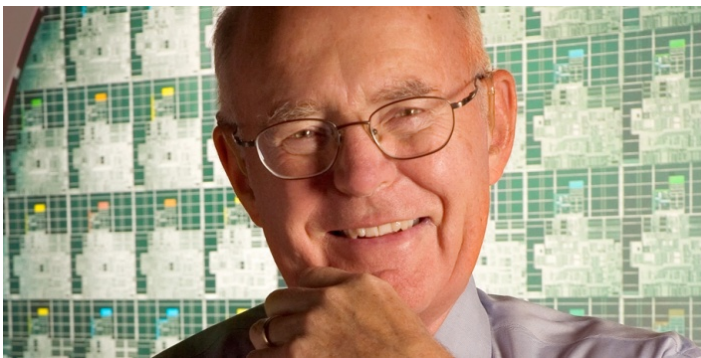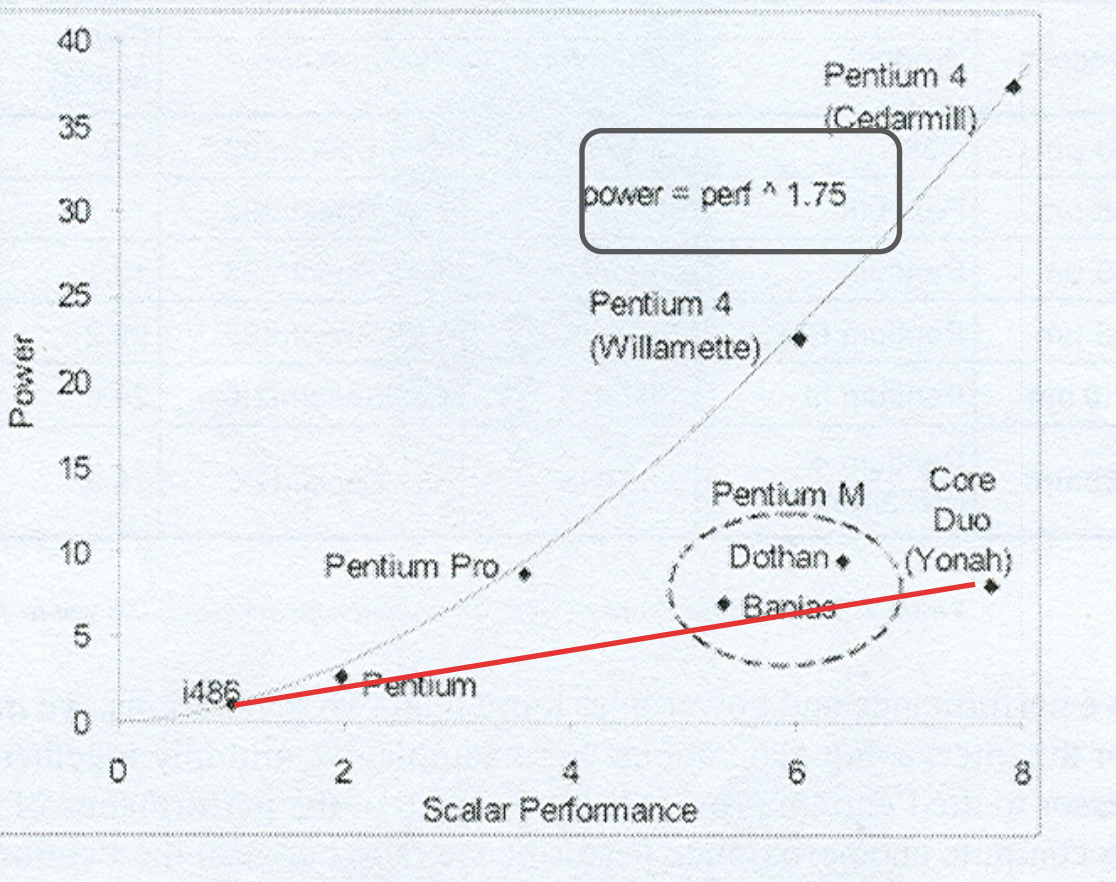1970   1975   1980   1985   1990   1995   2000   2005   2010

Moore's Law

The quantity of transistors doubles from process generation to process generation. Scaling factor $S = 1.414$

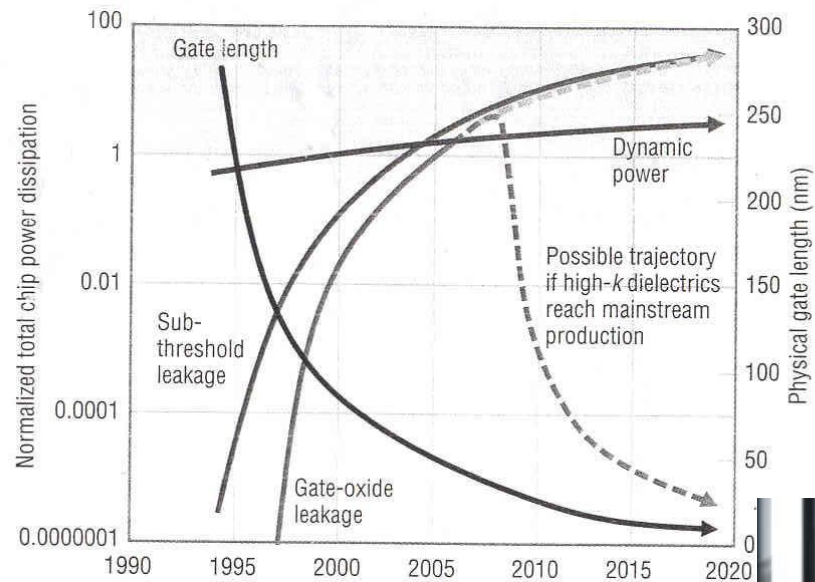Sources: Intel

Sources: Grochowski 2006
Hill & Marty 2008

Pollack's rule implies that a complex Single-Core-Processor that uses the HW resources of *r* **Basic Core Equivalents (BCE)** will only reach a performance of:

$$Perf = \sqrt{r}$$

At the same time it will consume a power of

$$P = Perf^{1.75}$$

This led to the rapid evolution of multicore processors between 2002 and today.

The leakage current in the *CMOS power equation* could be ignored until around 2004.

After that year the operating voltage could not be scaled any longer and **Dennard-Scaling** that guaranteed **2.8x** more performance from process generation to process generation for a given area and a given TDP had to be replaced by **Post-Dennard-Scaling**.

Dynamic and Static Power
Sources: IEEE 2002,
Taylor 2013

CMOS Power Equation:

$$P = QfCV^2 + VI_{leakage}$$

With **S = 1.4** power will rise with **S^2=2** per generation and chip utilization will drop with **1/S^2** per generation.

# The consequences of Post-Dennard-Scaling for Multicore Processors

What does Post-Dennard-Scaling mean for future generations of Multicore Processors?

If we start with a **22 nm** process and a quantity of 16 Basic Core Equivalents (BCEs), and if we assume an equal power envelope, an equal die-size from generation to generation, and a constant operating voltage, then with

Power = Quantity x Frequency x Capacitance x Voltage²

we receive the following scaled alternatives for the next two projected generations (i.e. **14 nm** and **11 nm**):
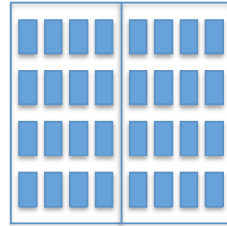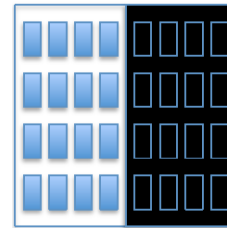
22 nm: 16 BCEs
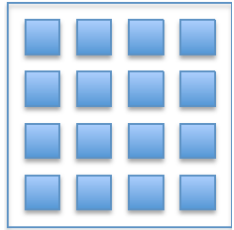
1 000 000 000 transistors

16 RISC cores @2 GHz
8 MB LLC
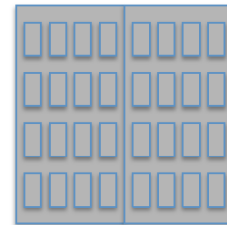Ring or mesh interconnect

14 nm 32 BCEs

2 000 000 000 transistors
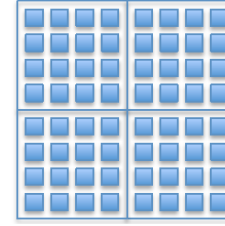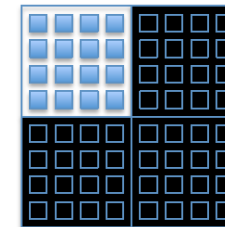
Option 1:

16 cores @2.8 GHz
16 cores dark

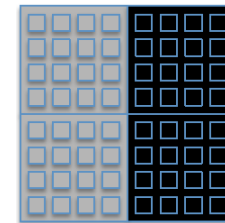**Option 2:**

32 cores @1.4 GHz

11 nm 64 BCEs

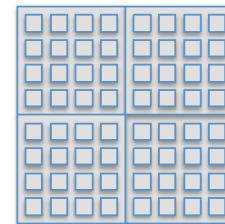4 000 000 000 transistors

Option 1:

16 cores @3.92 GHz
48 cores dark

Option 2:

32 cores @2 GHz

Option 3:

64 cores @1 GHz

Multicore Performance Projections based on Microelectronics Trends

- Due to Post-Dennard-Scaling energy efficiency only rises 40% per generation
- With a given power-envelope and with current microarchitectures performance will only increase 5.38x in 10 years compared to 30x every10 years with Dennard-Scaling (i.e. from 1974 to 2004)
- This corresponds to industry projections (e.g. S. Borkar, Intel, projected a 6x increase between 2008 and 2018)
- Hitting the power wall leads to fewer cores than projected and (too) large L3 caches
- 3D-stacking, hierarchical and/or low-energy interconnects could bring some relief in the future

Modeling Performance and Power of Multicore Architectures

Standard Performance Models

(Amdahl 1967, Hill & Marty 2008)

These speedup-oriented models build upon Amdahl's law that executes a given load with a parallelizable fraction f on an n-processor system.

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

Standard Performance Models
(Hill & Marty 2008)

Hill and Marty have extended Amdahl's law for three typical classes of multicore systems: symmetric, asymmetric, and dynamic multicore processors.

The symmetric type uses n BCEs for the parallel part, one BCE for the sequential part.
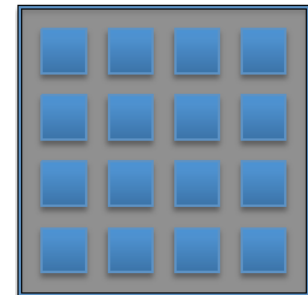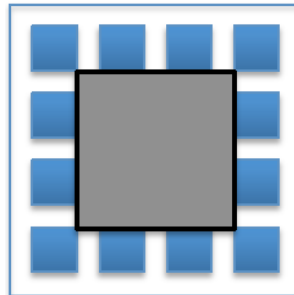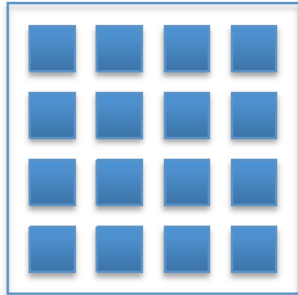
The asymmetric type uses n BCEs for the parallel part, one large core constructed from r BCEs offering a performance $perf(r)$ for the sequential part.

The dynamic type uses n BCEs for the parallel part and can be reconfigured to a large core that uses $n=r$ BCEs with $perf(r) = perf(n)$.

Standard Performance Models
(Hill & Marty 2008)

Thus, we have the following models:

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

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

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

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

For typical engineering and embedded loads the asymmetric and dynamic models deliver better speedups than the symmetric model. However, this dynamic model is not yet a realistic model for dynamic multicore operation.

Evolutionary Multicore Models

Evolution of commercial multicore processors has led to new variants of asymmetric and dynamic systems:
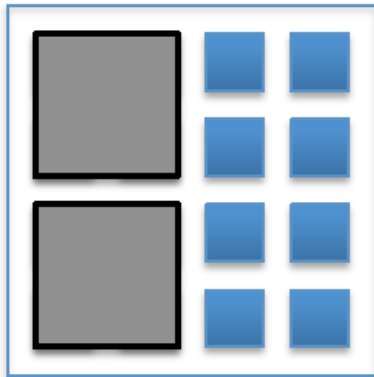
Model of an asymmetric multicore with 2 complex cores and 8 BCEs
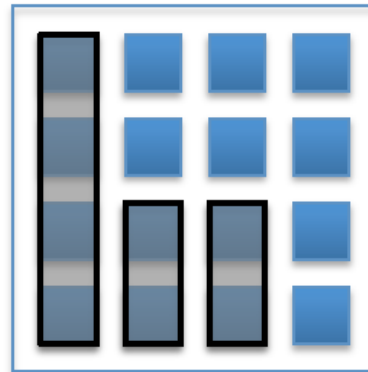
Model of a dynamic multicore with 4 cores and frequency scaling using the power budget of 8 BCEs

Heterogeneous Performance Models

Finally, two models for heterogeneous multicores are introduced. Heterogeneous multicores in addition to a complex conservative core introduce one or more unconventional (U-) cores, e.g. custom logic, FPGAs, GPU-like resources. These resources are exploited by specific parts of applications with SIMD parallelism, GPU-like multithreading, or specific parallel algorithms mapped to custom or FPGA logic:
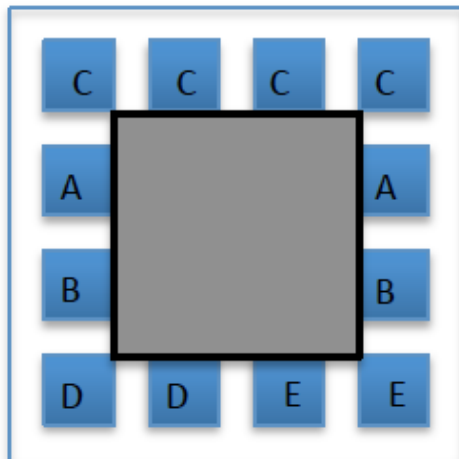
Heterogeneous multicore with one large core, 4 BCEs, two accelerators or co-processors of type A, B, D, E, each. Each co-processor/accelerator uses the same transistor budget as a BCE.

Heterogeneous Performance Models

The first model by Chung et al. (2010) assumes one large core and $n - r$ U-cores that offer a relative performance µ compared to the performance of a BCE.

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

The second model, presented here for the first time, assumes l different accelerators needing each a different amount of resources and having each a different relative performance μ_i. They each operate on different parts f_i of the application workload. If a $\mu_i=1$, then this accelerator is just an array of r_i BCEs and serves as a symmetric multicore accelerator. For power efficiency reasons, each U-core can have a relative efficiency of φ_i compared to a $\varphi_i = 1$ for a BCE.

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

Multicore Performance Scaling Study
(Esmaeilzadeh et al. 2013)

Esmaeilzadeh et al. in 2013 have presented a study that combines microelectronics scaling (ITRS/Shekar Borkar), micro-architectural, and multicore-performance scaling from 45 nm to 8 nm. The study has produced a simulator that is available online and can be executed for given parallel benchmark or abitrary loads.

The study results in a pessimistic prediction of a 3.7x to 7.9x speedup for an 8 nm multicore compared to a quadcore based on Intel's Core i7 Nehalem.

These simulation results are mainly caused by **too little exploitable parallelism** (although parallel programming benchmarks were executed) and by **dark silicon due to the power wall.**

Multicore Evolution
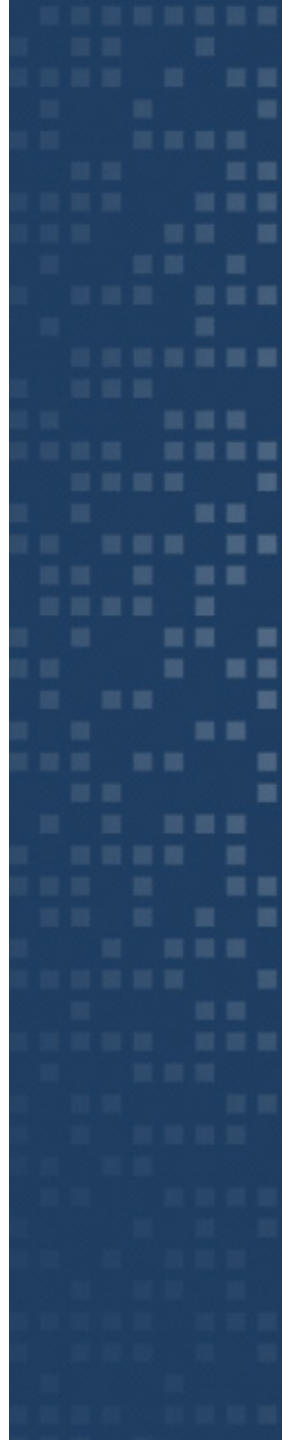
Multicore Evolution

POWER8 Processor (IBM)

Technology
- 22nm SOI, eDRAM, 15 ML 650mm2

Cores
- 12 cores (SMT8)
- 8 dispatch, 10 issue, 16 exec pipe
- 2X internal data flows/queues
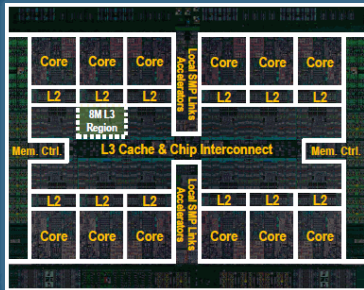- Enhanced prefetching
- 64K data cache, 32K instruction cache

Accelerators
- Crypto & memory expansion
- Transactional Memory
- VMM assist
- Data Move / VM Mobility

Caches
- 512 KB SRAM L2 / core
- 96 MB eDRAM shared L3
- Up to 128 MB eDRAM L4 (off-chip)

Memory
- Up to 230 GB/s sustained bandwidth

Bus Interfaces
- Durable open memory attach interface
- Integrated PCIe Gen3
- SMP Interconnect
- CAPI (Coherent Accelerator Processor Interface)

Energy Management
- On-chip Power Management Micro-controller
- Integrated Per-core VRM
- Critical Path Monitors

© 2013 International Business Machines Corporation

Knights Landing Processor Architecture
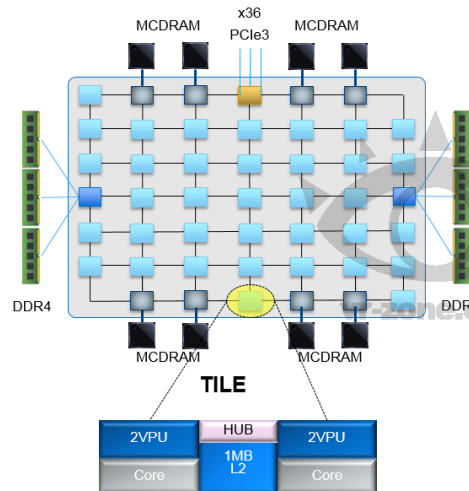
Up to 72 Intel Architecture cores based on Silvermont (Intel® Atom processor)
- Four threads/core
- Two 512b vector units/core
- Up to 3x single thread performance improvement over KNC generation

Full Intel® Xeon processor ISA compatibility through AVX-512 (except TSX)

6 channels of DDR4 2400 MHz -up to 384GB

36 lanes PCI Express* Gen 3

8/16GB of high-bandwidth on-package MCDRAM memory >500GB/sec

200W TDP

Symmetric Multicore Processors

Desktop

- **Intel Haswell (22 nm) and Broadwell (14 nm)** with 4, 6, 8 large complex cores, moderate TDP, graphics engine, AVX 256-bit wide SIMD unit, Turbo-Boost 2.0, wider internal microarchitecture (8 instruction issue-ports), larger buffers, support for transactional memory
- **AMD Kaveri,** 28 nm, 2 or 4 cores, moderate TDP, advanced Steamroller microarchitecture, huge on-chip GPU (47% of die area)

Symmetric Multicore Processors

Embedded

- **ARM Cortex-A** with advanced superscalar microarchitectures, longer pipelines, 64-bit architecture (Cortex-A53 and A57), improved branch-prediction, coherent bus interface, larger L2 caches
- **Intel Atom Z3770** with on-chip GPU and low TDP for tablets, smartphones, laptops
- **Intel Atom Silvermont**, 14 nm implementation, fast superscalar micro-architecture, low TDP

Symmetric Multicore Processors

Server

- **Intel Xeon** with Haswell (22 nm) and Broadwell microarchitecture. Haswell EP with 14 cores and 37.5 MB L3, Broadwell EP with up to 18 cores.
- **Intel Atom C2750**, Avoton, 8 core, micro-server chip, up to 2.6 GHz
- **AMD Kaveri** server chips with 8+ cores, five HT links
- **AMD 16 core ARM Cortex-A57** micro server
- **IBM Power8**, 12 core, 4 GHz, 22 nm, server chip with 410 GB/s peak memory bandwidth due to buffered Centaure memory interface chip and eight high-speed memory channels per chip, transactional memory.
- **Oracle Sparc T5**, 28 nm, 16 cores, crossbar interconnect, 8 MB L3, 8 threads per core, glue-less coherent interface between 8 chips.
- **Fujitsu Sparc64X**, 28 nm, 16 core, including SWoC accelerator for arithmetic and encryption.

Symmetric Multicore Processors

Many-Core-Chips

- **Intel Xeon PHI – 22 nm, Knights Corner** with 61 enhanced Pentium cores, ring interconnect, 512-bit vector unit, 4 HW threads, GDDR5 memory interface, 353 GB/s peak bandwidth.
- **Intel Knights Landing** (14 nm, planned for 2015) with 72 Silvermont cores, advanced vector units, new memory interface, new interconnect, 3 TFLOPS per chip, extremely energy-efficient: 10 GFLOPS/W, could be used to build first Exaflop supercomputer
- **Tilera TILE-Gx**, 40 nm, 72 RISC cores (64-bit), 5-layer mesh-interconnect to other cores, 1 GHz operating frequency, high-speed interconnect to other chips, I/O or external FPGA accelerators.

# Techniques for Asymmetric and Dynamic Multicore Processing

- **ARM big.LITTLE.** Coupling of a cluster of extremely energy-efficient cores to a cluster of high-performance cores. Dynamic frequency and voltage scaling to balance energy and optimally adapt to varying loads.
- **Intel Turbo Boost 1.0.** Balancing of the core frequency against available thread parallelism
- **Intel Turbo Boost 2.0.** Temporary boost of core performance by frequency scaling until temperature limit, after this: fall-back into low power mode.
- **Computational sprinting.** With future phase-change materials application hot spots can be accelerated.
- **Fine-grained energy management and clock gating** (e.g. Intel Haswell, IBM Power8).
- **Near-Threshold Voltage DVFS.** Rapid drop of frequency and voltage for well-known application requirements. Intel experimental x86 core.

Heterogeneous Multicore Processors

Two examples of multicore processors, where heterogeneous techniques are already applied. However, these systems can also be viewed from a different perspective.

- **Freescale QorIQ T Series.** The T4240 is a 12-core processor based on 64-bit Power architecture, with dual-threaded cores, 1.8 GHz clock, designed for heavy embedded loads. The chip can also be seen as a symmetric multicore. The heterogeneous part is its Data Path Acceleration Hardware (DPAA) with specific HW for security, pattern-matching, and data compression.
- **Altera Stratix 10.** This system can also be seen as a large, fast FPGA. What makes it a heterogeneous multicore, are the four integrated ARM Cortex-A53 cores. The chip is manufactured by Intel in the new 14 nm process with FinFET transistors.

Conclusion
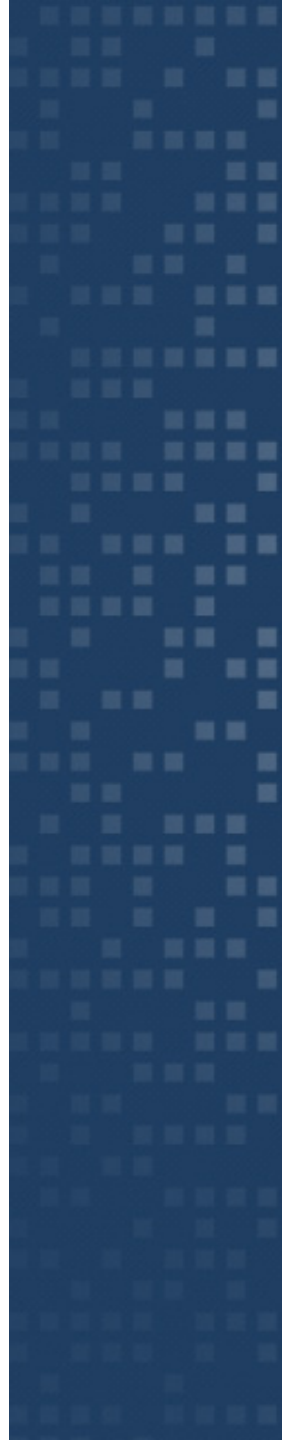
Multicore Evolution: The Next 10 Years:

Dark Silicon research has to be extended. Univ. of Southern California San Diego and Univ. of Wisconsin are the current leaders. Research topics for research and industry will be:

- What can we do with the dark silicon?
- Strategies for accelerators, heterogeneous architectures, software-specific co-processors
- Impact on embedded and engineering workloads
- Impact on real-time workloads

Multicore Evolution: The Next 10 Years:

- Intensifying the **search for new transistor materials** with the goal to lower operating voltage (R. Stevenson, IEEE Spectrum, 01/14): IBM and Imec, Leuven, Belgium, see good chances for new transistors with indium gallium arsenide and silicon germanium components: goal $\rightarrow$ lower operating voltage to 0.5 V within 5 years.
- **3D-DRAM:** vertical positioning of DRAM chips (Samsung, Micron, Hynix) could break the memory wall and change rules for memory-access speed (see R. Courtland, IEEE Spectrum, 01/14).
- Improve **software tools for parallel programming** and make exploiting the parallelism in embedded and engineering applications easier

Thank you for your interest!

Email: maertin@ieee.org

www.christianmaertin.de

Email2: Christian.Maertin@hs-augsburg.de

www.hs-augsburg.de

HOCHSCHULE
AUGSBURG