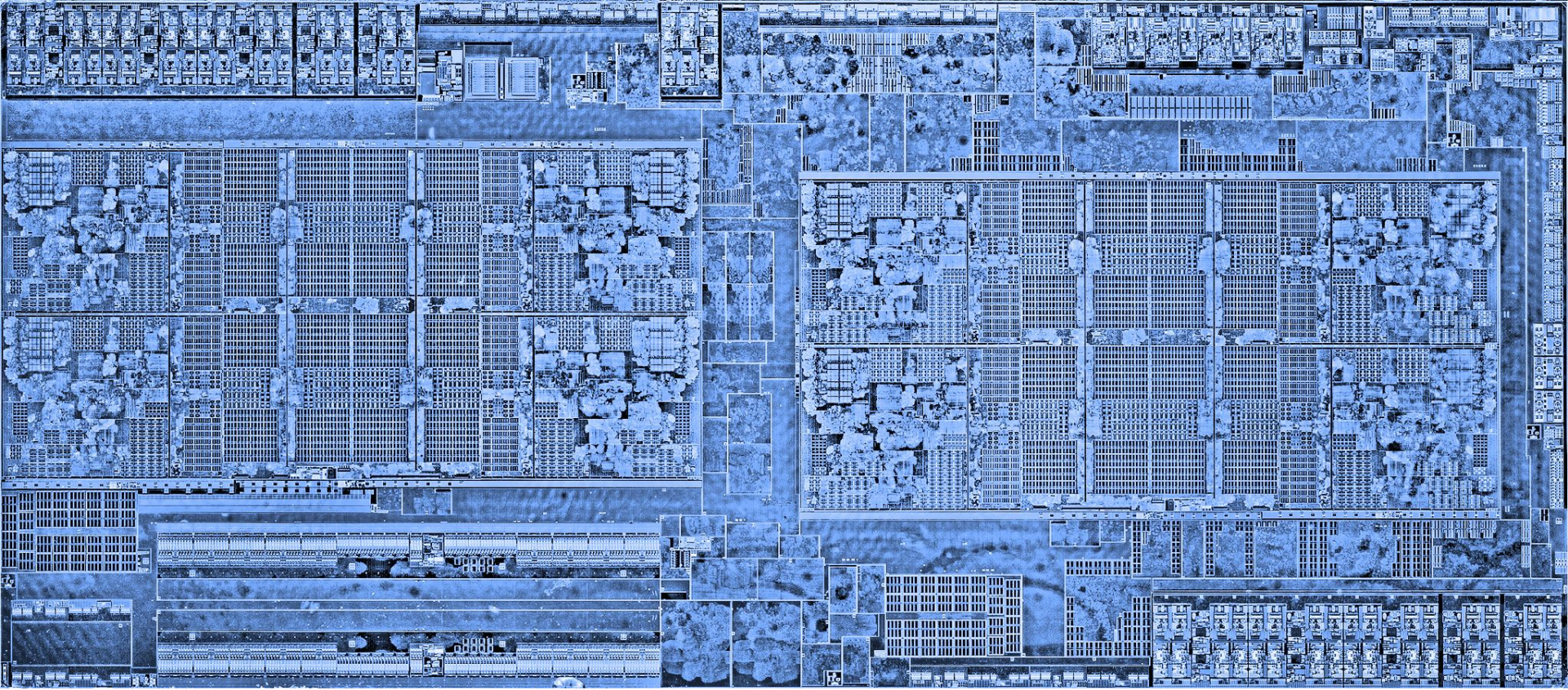


AMD Zen die shot of the Ryzen 7 octa-core die. Source: WikiChip

CTA01++: multicore and beyond

Johan Peltenburg, PhD

Postdoc @ TU Delft
j.w.peltenburg@tudelft.nl

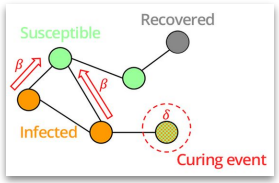
Co-founder @ Teratide
johan@teratide.io

Where I work

- Delft University of Technology
  - 26k+ students + 3k PhDs + 6k personnel
- Faculteit van Elektrotechniek, Wiskunde en Informatica
- Department of Quantum & Computer Engineering:
  - Computer Engineering
    - Accelerated Big Data Systems
  - Network Architecture & Services
  - Quantum Circuits, Architectures and Technology

TU Delft

Quantum & Computer Engineering



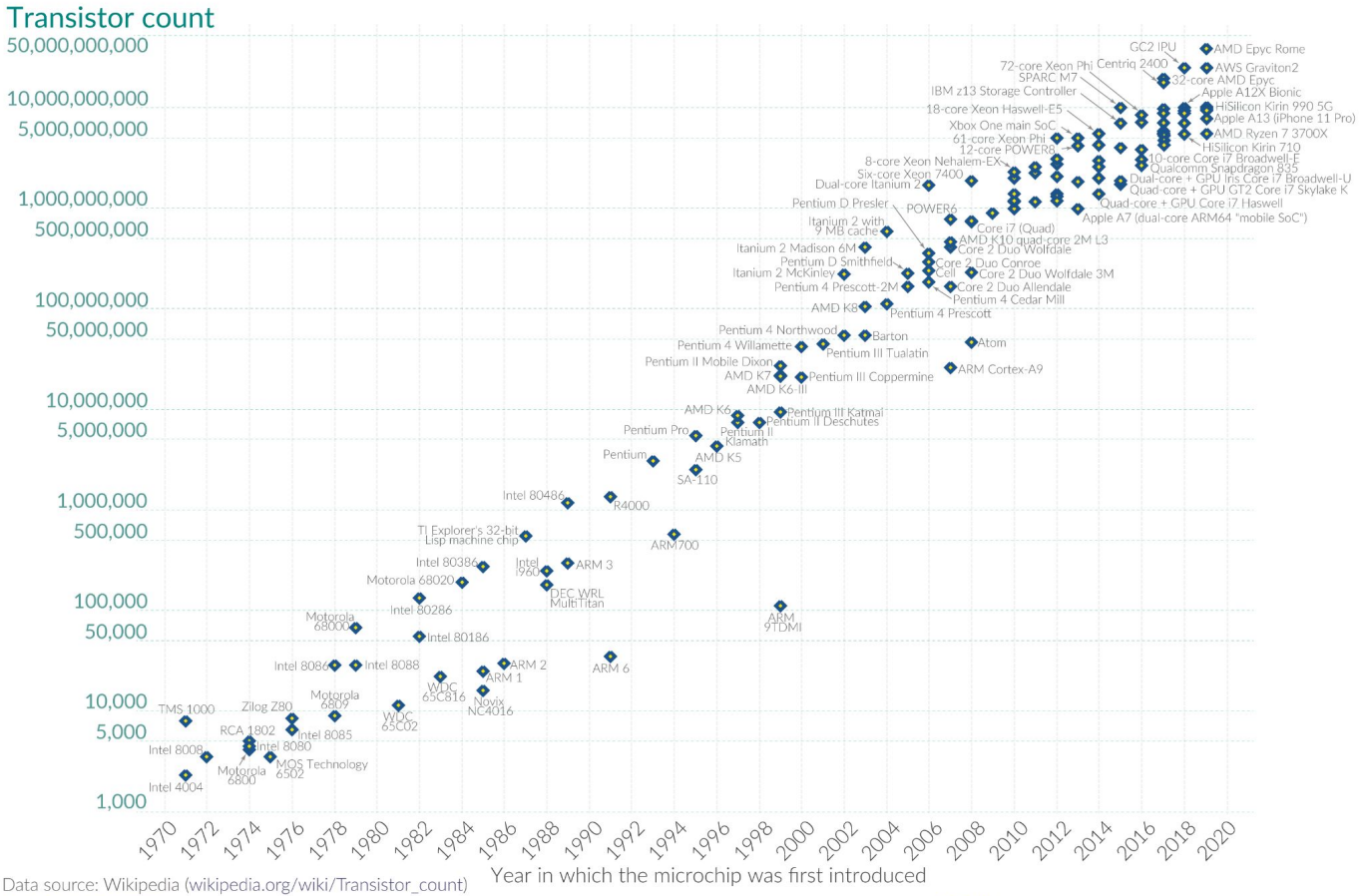
Contents

- Need for multicore processors
- Paper: Amdahl's Law in the Multicore Era
- Paper: Dark Silicon and the End of Multicore Scaling
- Accelerators

Moore's Law: The number of transistors on microchips doubles every two years

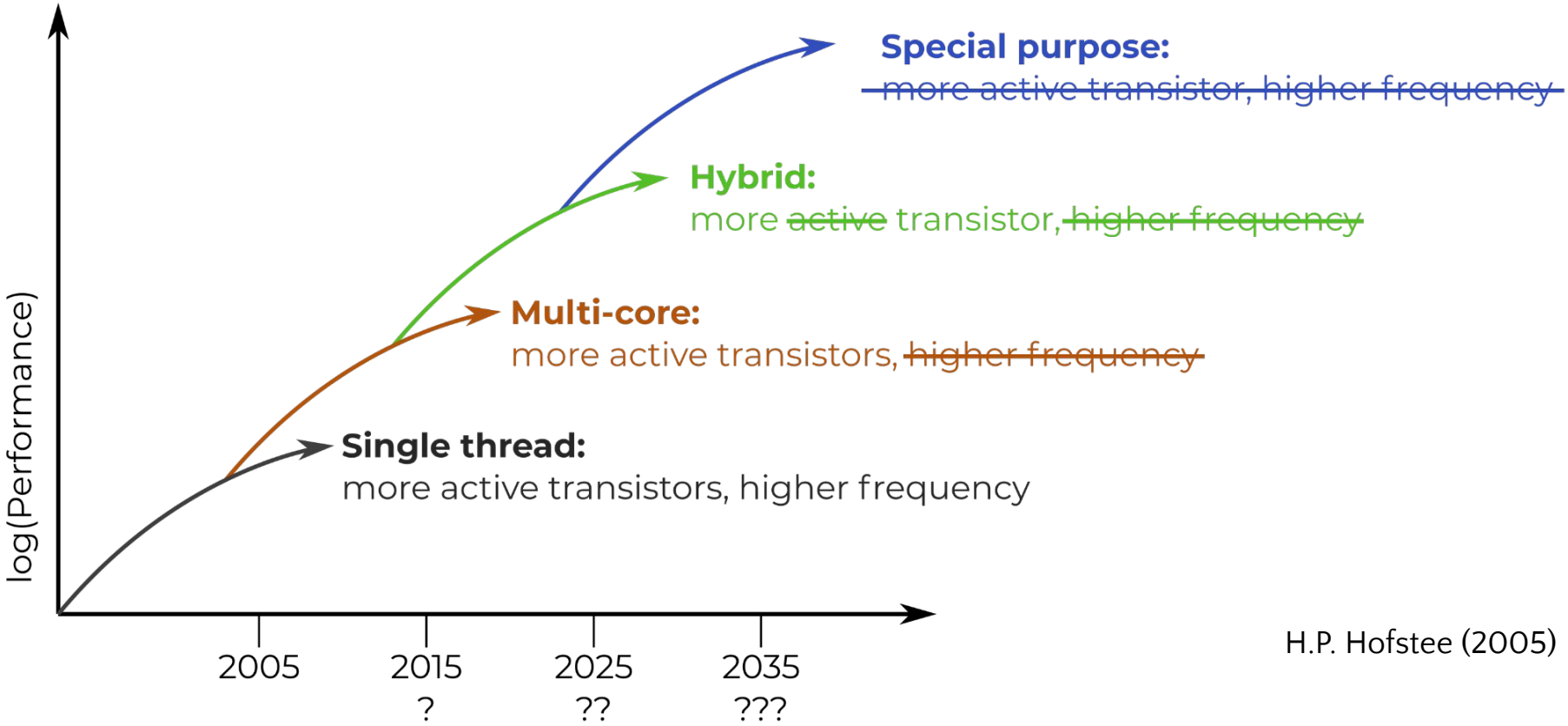
Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

What to do with these transistors?



H.P. Hofstee (2005)

What is a multicore processor?

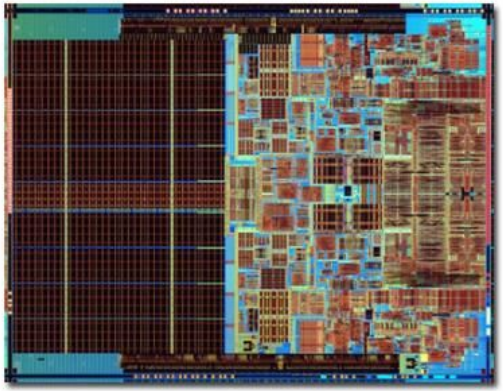
A multi-core processor is a computer processor on a single integrated circuit with two or more separate processing units, called cores, each of which reads and executes program instructions.

Characteristics:

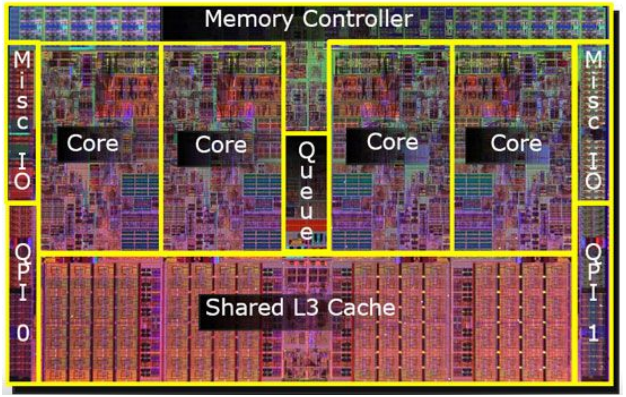
- Two or more general purpose processors
- Single component (chip or integrated circuit)
- Sharing infrastructure (sharing memory and communication resources)

Multicore examples

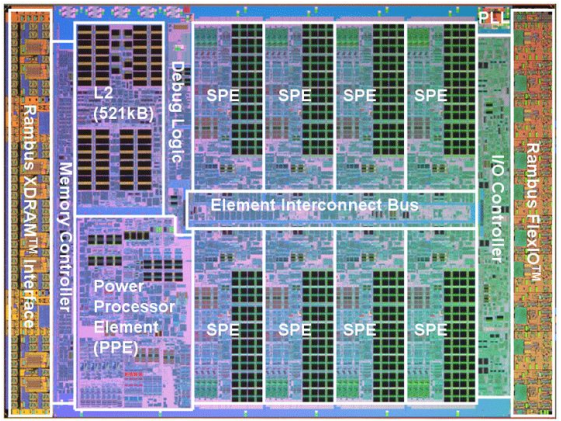
- Examples of multicore processors
  - Intel Core, Core 2, i3, i5, i7, etc.
  - AMD Athlon, Phenom, Ryzen
  - Sony CELL microprocessor
- Many other multicore examples:
  - Adapteva
  - Aeroflex
  - Ageia
  - Ambric
  - AMD
  - Analog Devices
  - ARM
  - ASOCS
  - Azul Systems .....



Intel Core 2 Duo

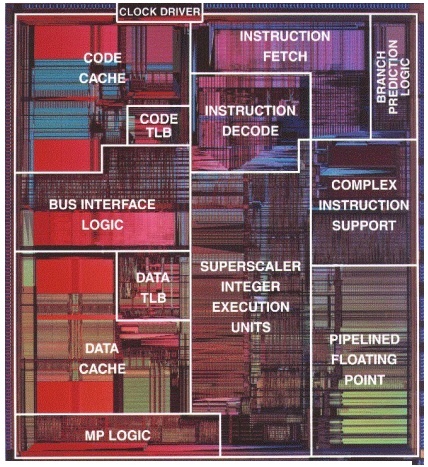


Intel i7



Sony/IBM CELL

Are these multicore processors?

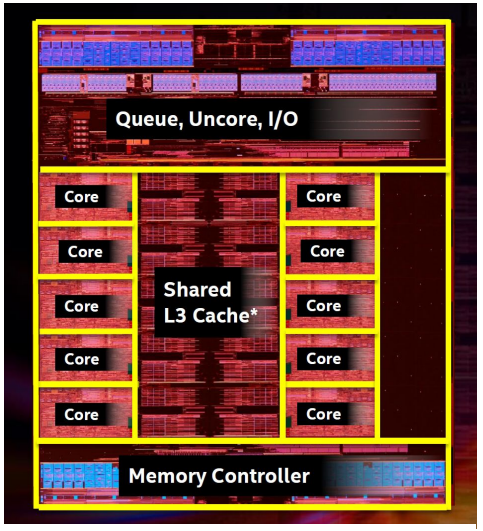


Intel P4 with SSE

Is a processor with embedded acceleration circuitry considered as a multicore?

Is a GPU considered a multicore processor?

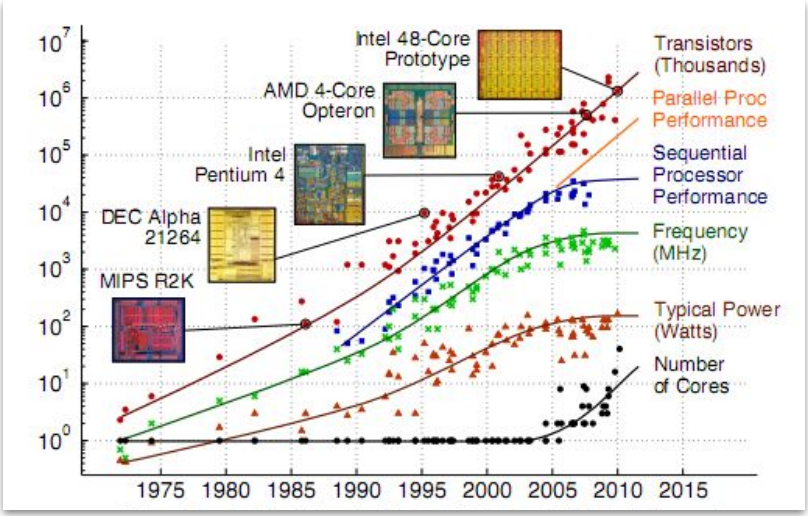
Nvidia Tesla V100



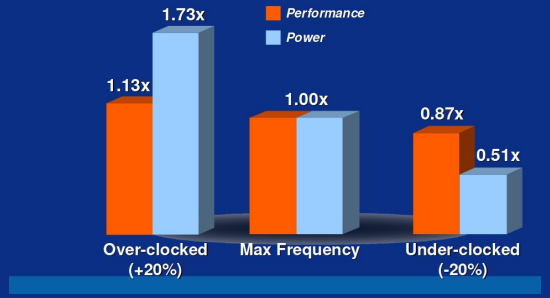
Intel i7 6950X

Are two independent processors on a single chip considered as multicore?
(Not shown)

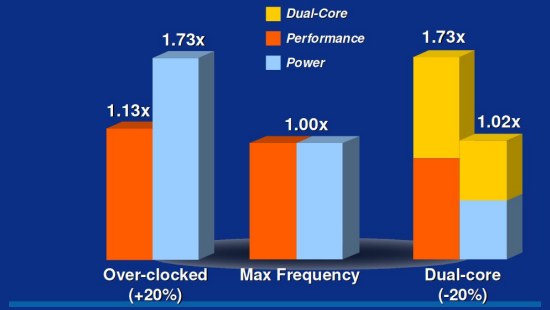
Overcoming the Power Wall



Under-clocking



Multi-Core Energy-Efficient Performance



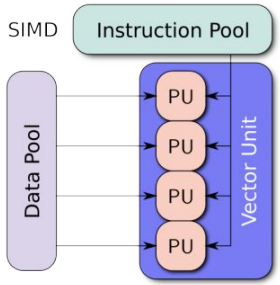
Source: Intel

ILP wall

- The ’80s: superscalar expansion
  - 50% per year improvement in performance
  - Pipeline processor
  - 10 CPI → 1 CPI
- The ’90s: the era of diminishing returns
  - Squeezing out the last bit of implicit parallelism
  - 2-way to 6-way issue, out-of-order issue, branch prediction
  - 1 CPI → 0.5 CPI
- The ’00s: the multicore era
  - The need for explicit parallelism

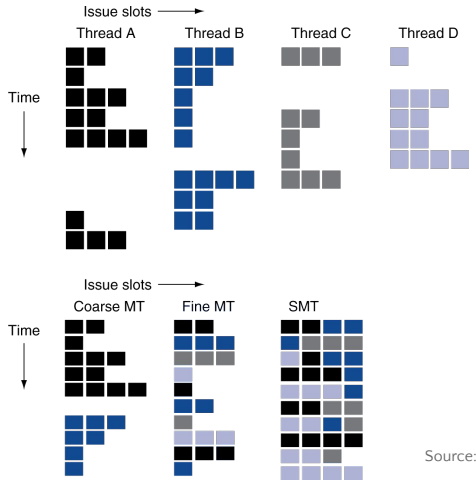
Advanced Computer Architecture topics

- Very deep pipelines
  - Intel once went up to 31!
  - Complex branch predictors
  - Speculative execution
- Advanced cache technologies
  - E.g. prefetching
- SIMD extensions
  - E.g. Intel SSE, AVX, AVX2, AVX512
  - E.g. ARM NEON



Source: https://commons.wikimedia.org

- Out-of-Order execution
- Superscalar execution
  - Can launch execution of more than one instruction.
- Simulatenous Multithreading (SMT)
  - Can run multiple threads on a single core
  - Intel calls this HyperThreading



Source: Patterson & Hennessy

What do all these improvements cost?

Pollack's Rule (of thumb)

Performance increase is roughly proportional to the square root of increase in complexity.

Exercise:

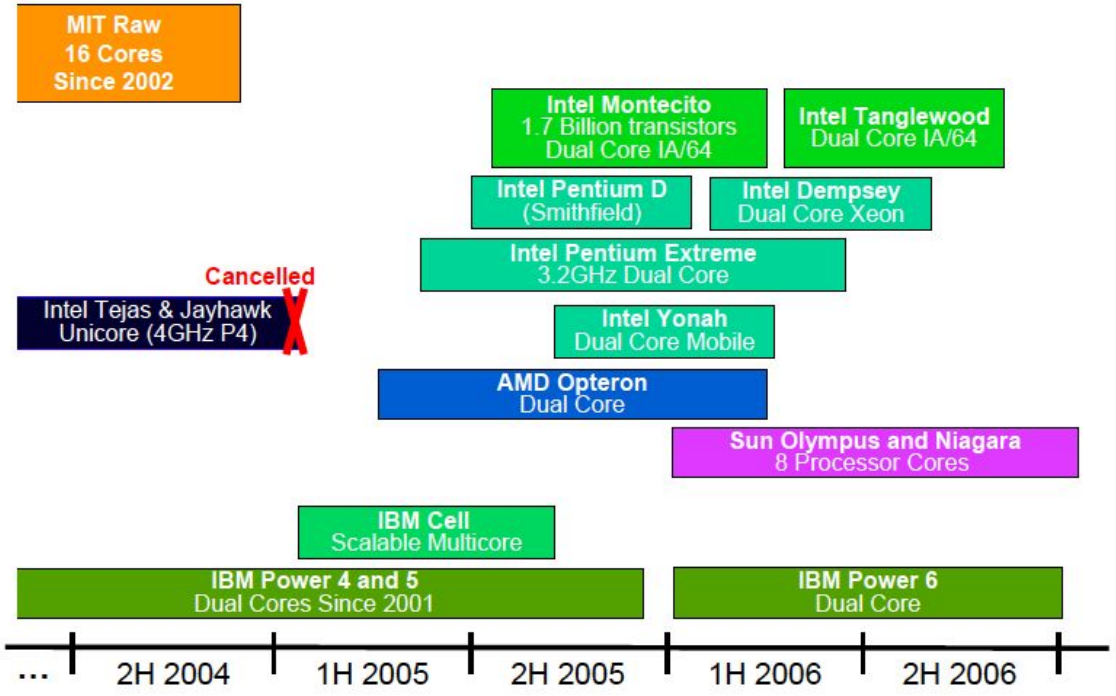
- We double the amount of transistors on our chip.
  How much performance do we get?

- We only get 1.4× (√2) improvement in performance.

Source: Borkar, S. (2007, June). Thousand core chips: a technology perspective. In Proceedings of the 44th annual Design Automation Conference (pp. 746-749). ACM.

More reasons for multicore

- Memory Wall (see book)
- Industry push



Amdahl's Law

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

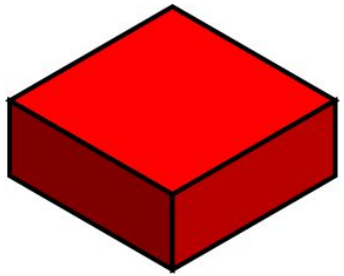
S : speedup w.r.t. single core

n : number of cores

f : parallel portion of workload

Amdahl’s Law in the Multicore Era

- Resources to build and operate a computer circuit:
  - Area, capacitive load, frequency, power, money, etc...
- Let’s forget about the specific resource and call it:
  - A “Base Core Equivalent” or BCE.
  - A “Base Core” are the resources required to implement the simplest core imaginable that can run our instruction set.



n = 1 BCE

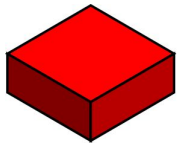
Normalized performance = 1

Source: Hill, M. D., & Marty, M. R. (2008). Amdahl's law in the multicore era. Computer, 41(7).

Exercise

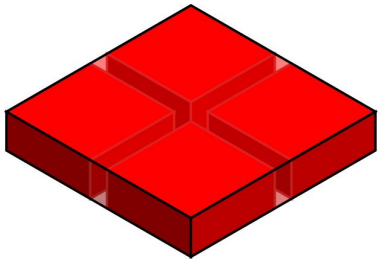
- n : the total number of BCE resources available to our design
- r : the number of BCE resources we use for our single core
- We create an architecture of r = 4 BCE
- How much performance does CPU A get compared to CPU B?

A:



n = r = 1 BCE
Performance = 1

B:

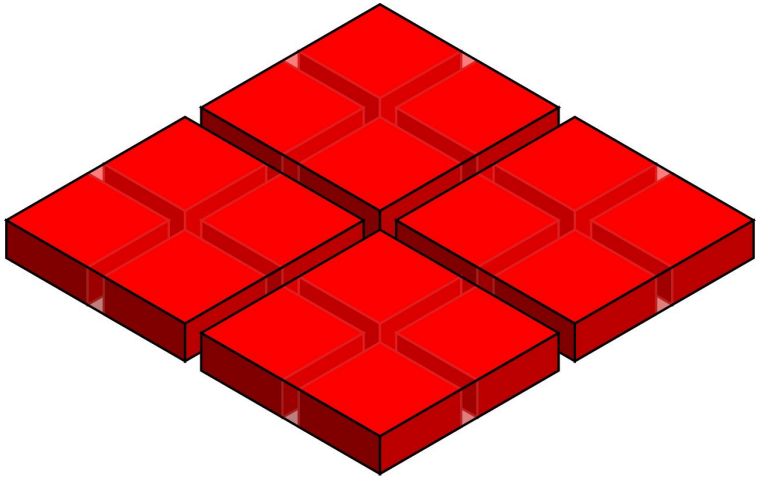


n = r = 4 BCE
Performance = ?

Performance of an r BCE core = $\sqrt{r}$
(remember Pollack's Rule)

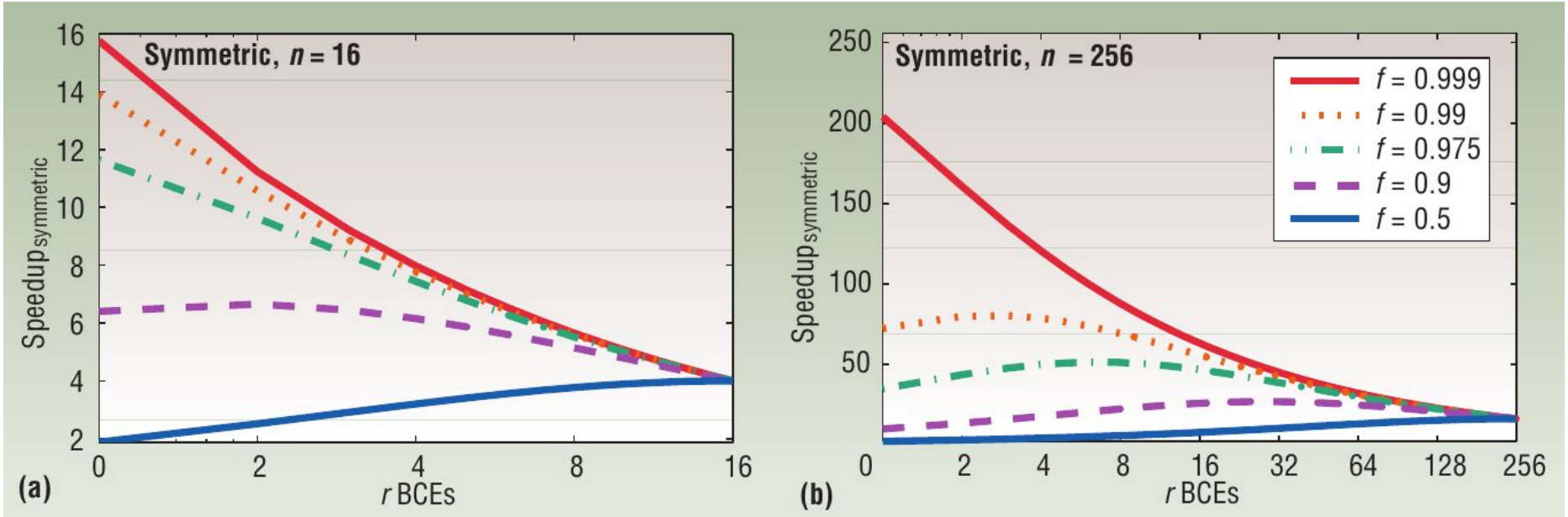
Symmetric Multicore (1/2)

- Let's build a multicore system with a resource budget of:
  - n = 16 BCE
- We give each core the same amount of resources:
  - r = 4 BCE.
- What is the speedup over 1 BCE given we have a parallel portion of f ?



$$S(n) = \frac{1}{\frac{1-f}{\sqrt{r}} + \frac{f \cdot r}{\sqrt{r} \cdot n}}$$

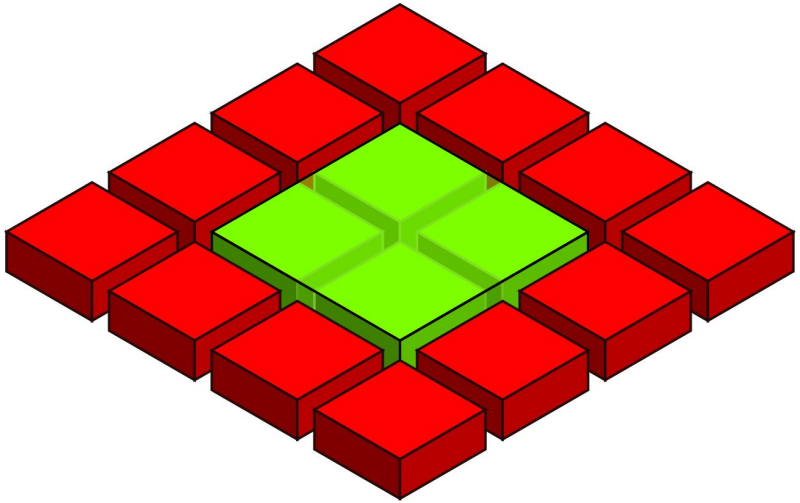
Symmetric Multicore (2/2)



Source: Hill, M. D., & Marty, M. R. (2008). Amdahl's law in the multicore era. Computer, 41(7).

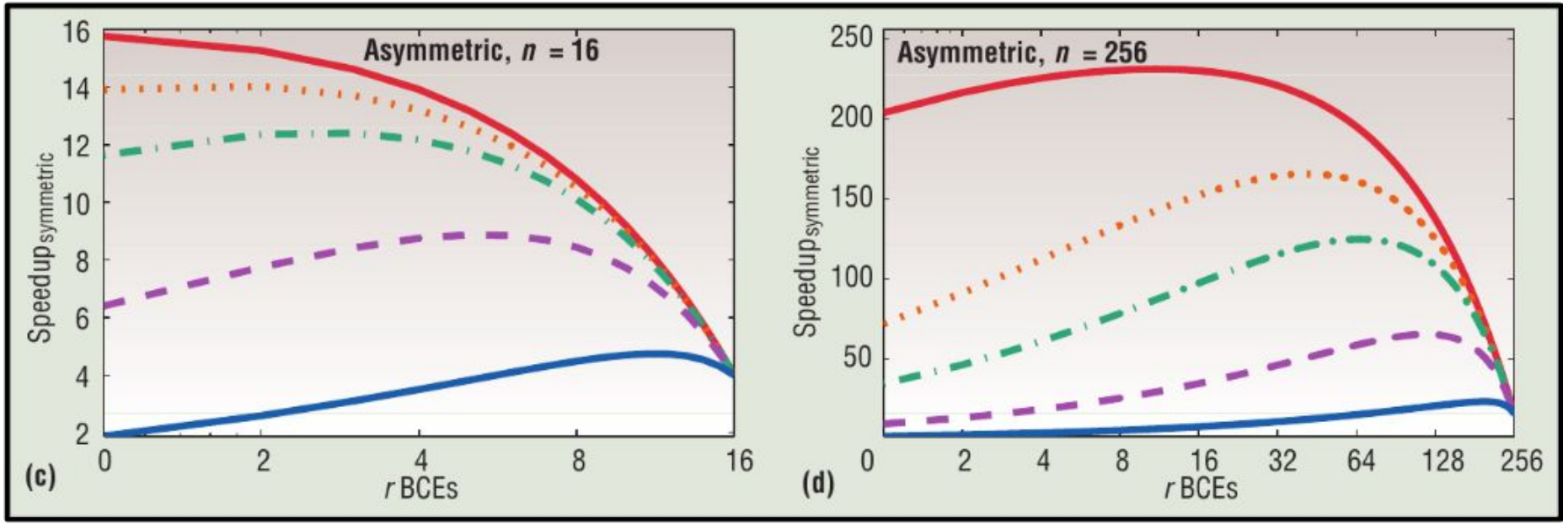
Asymmetric Multicore (1/2)

- A multicore system with an
  - n = 16 BCE budget.
- We create one big core
  - r = 4 BCE.
- We create a small, simple core out of each remaining BCE.
- What is the speedup over 1 BCE given we have a parallel portion of f ?



$$S(n)=\frac{1}{\frac{1-f}{\sqrt{r}}+\frac{f}{\sqrt{r}+n-r}}$$

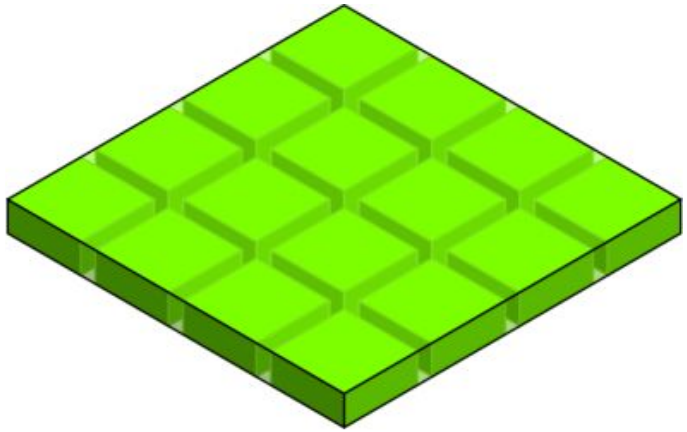
Asymmetric Multicore (2/2)



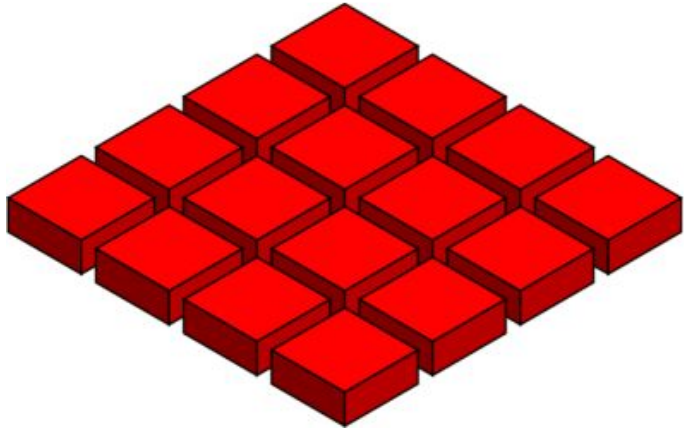
Source: Hill, M. D., & Marty, M. R. (2008). Amdahl's law in the multicore era. Computer, 41(7).

Dynamic Multicore (1/2)

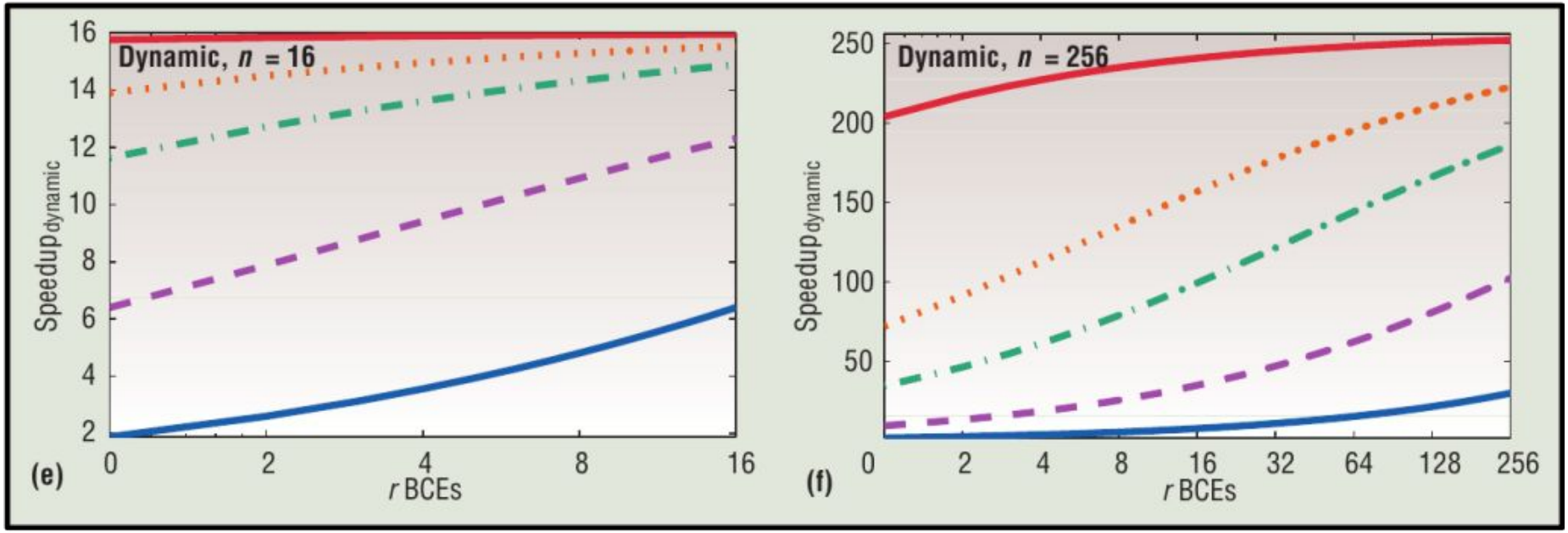
- When performing a non-parallelizable part 1-f...
  - Use all BCE to form one huge core



- When performing parallelizable part f...
  - Use all BCE to form many tiny cores



Dynamic Multicore (2/2)

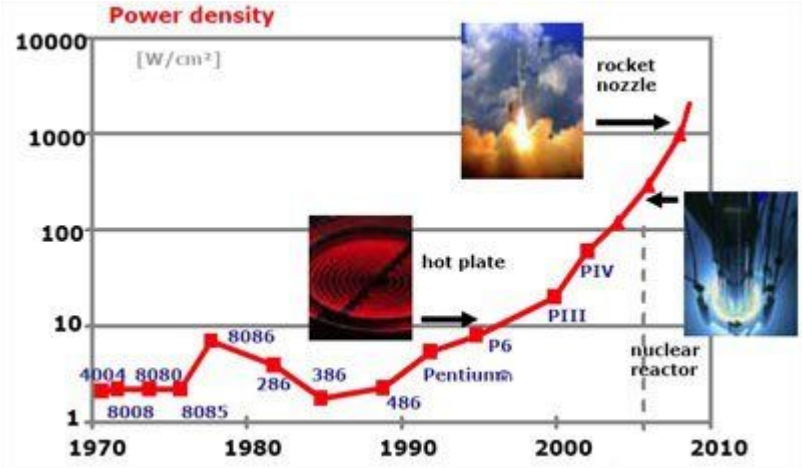


Source: Hill, M. D., & Marty, M. R. (2008). Amdahl's law in the multicore era. Computer, 41(7).

What sort of magical device would have such dynamic properties?

Dennard Scaling

- Once upon a time …
- As transistors became smaller …
- Their power density stayed constant.
- Moore's law & Dennard Scaling lived happily ever after… The end?

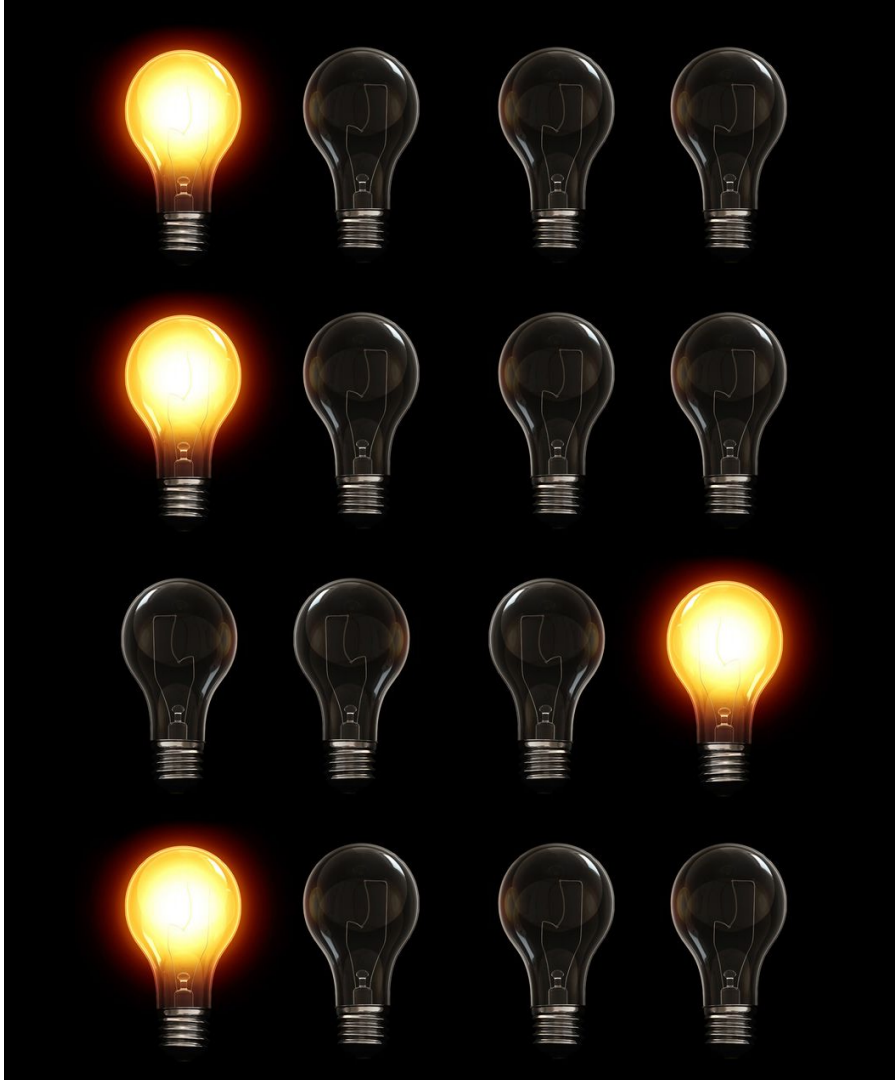


- Leakage current and threshold voltage were not taken into consideration for Dennard Scaling.
- Broke down around 2006: The Power Wall!

Source: Dennard, R. H., Gaensslen, F. H., Rideout, V. L., Bassous, E., & LeBlanc, A. R. (1974). Design of ion-implanted MOSFET's with very small physical dimensions. IEEE Journal of Solid-State Circuits, 9(5), 256-268.

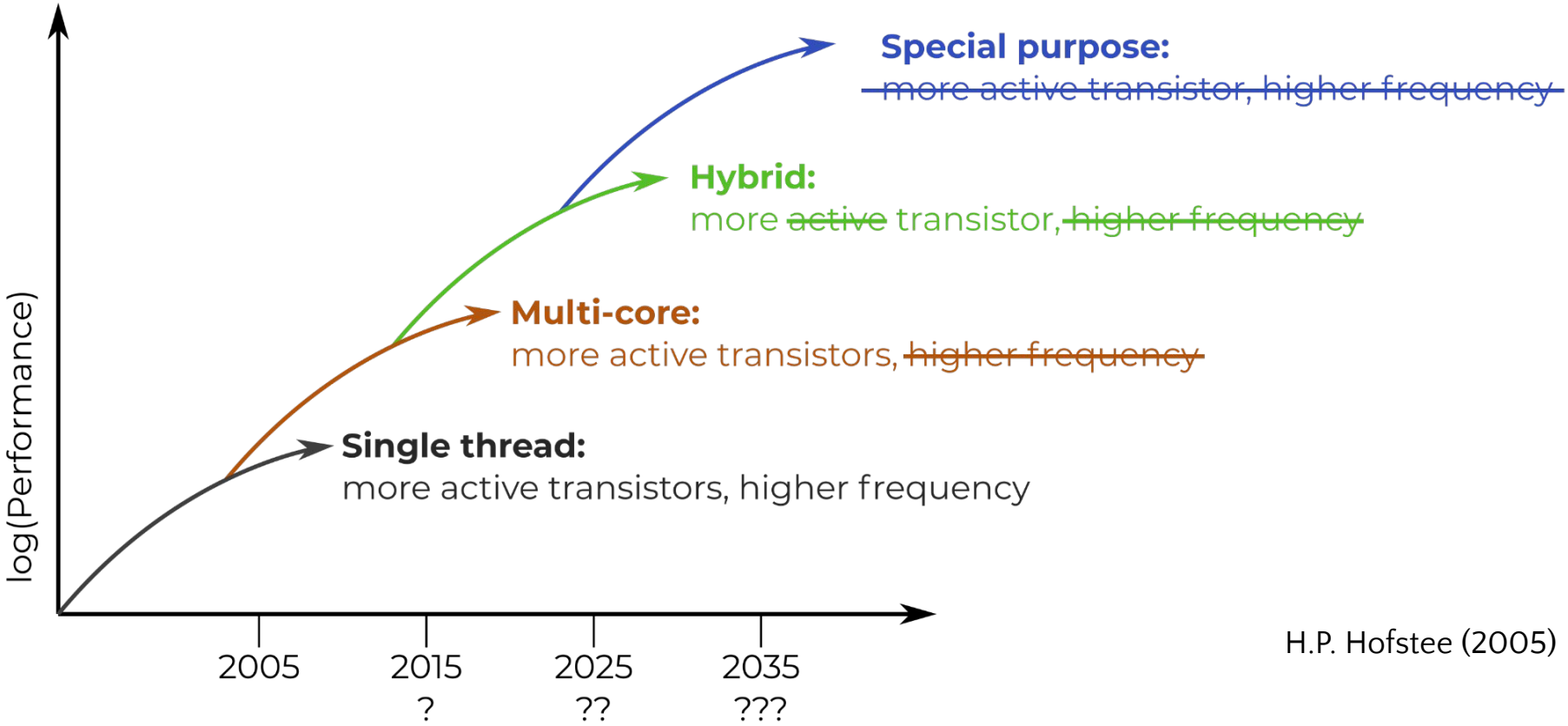
Dark Silicon

- No matter what chip topology we use (CPU-like / GPU-like)
  - We must power off parts of the chip to stay within a power budget.
  - At 8 nm, we must power off 50% of the chip continuously to stay within power budget!
- Limits to speedup in 2024:
  - Only 7.9x predicted when paper appeared in 2011!
  - Shouldn't we get ~388x according to Moore's Law?
- Don't confuse Moore's Law with performance!



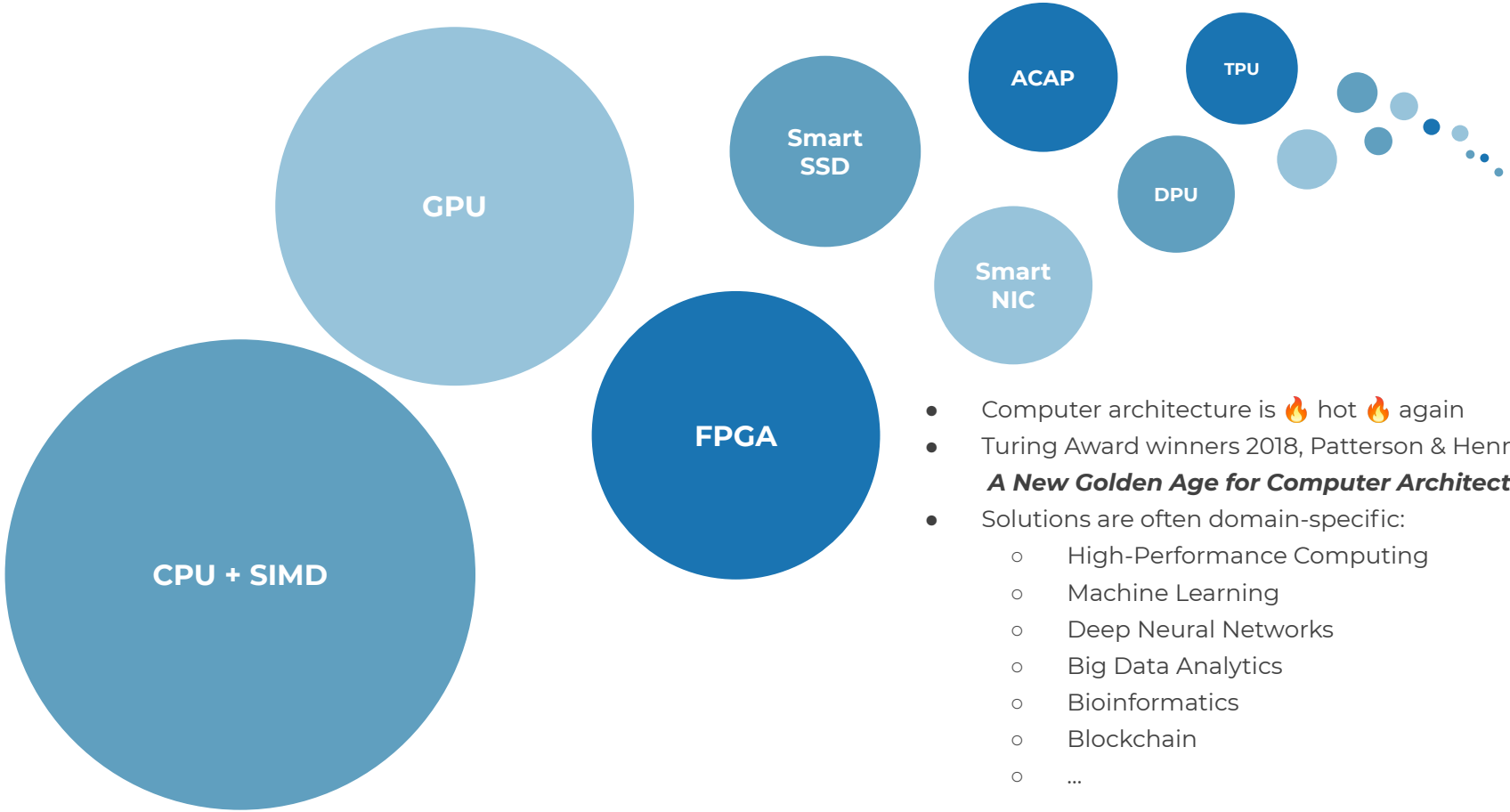
Source: Esmaeilzadeh, H., Blem, E., Amant, R. S., Sankaralingam, K., & Burger, D. (2011, June). Dark silicon and the end of multicore scaling. In Computer Architecture (ISCA), 2011 38th Annual International Symposium on (pp. 365-376). IEEE.

What to do with these transistors?



H.P. Hofstee (2005)

Accelerators

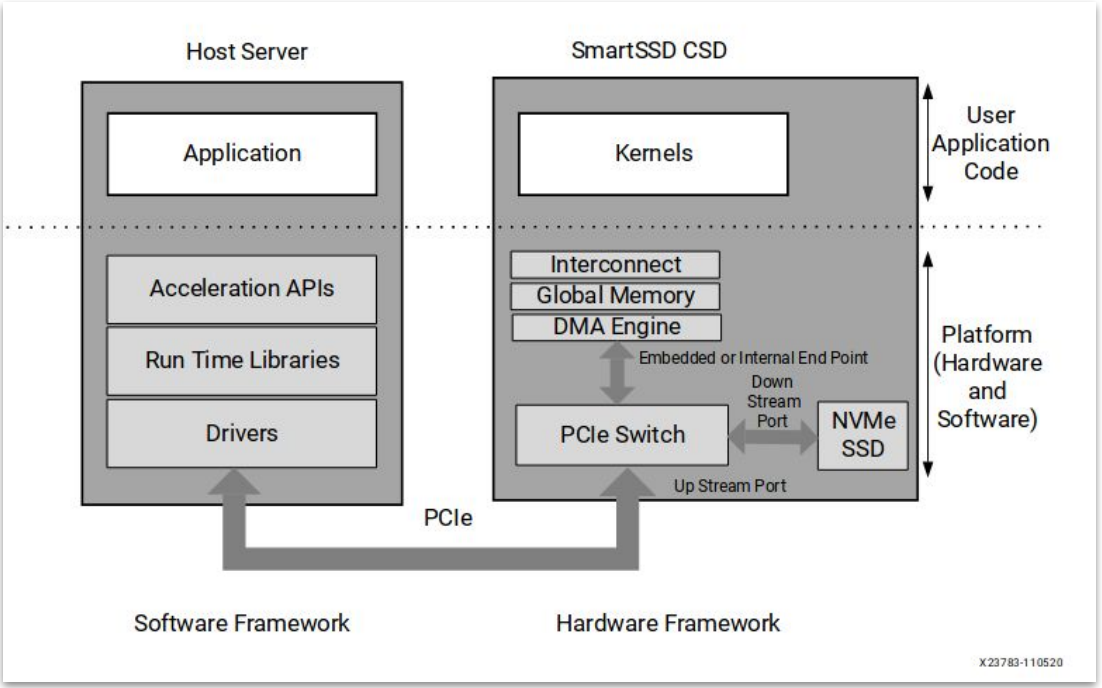


- Computer architecture is 🔥 hot 🔥 again
- Turing Award winners 2018, Patterson & Hennessy:
  A New Golden Age for Computer Architecture
- Solutions are often domain-specific:
  - High-Performance Computing
  - Machine Learning
  - Deep Neural Networks
  - Big Data Analytics
  - Bioinformatics
  - Blockchain
  - ...

Common Accelerator Characteristics

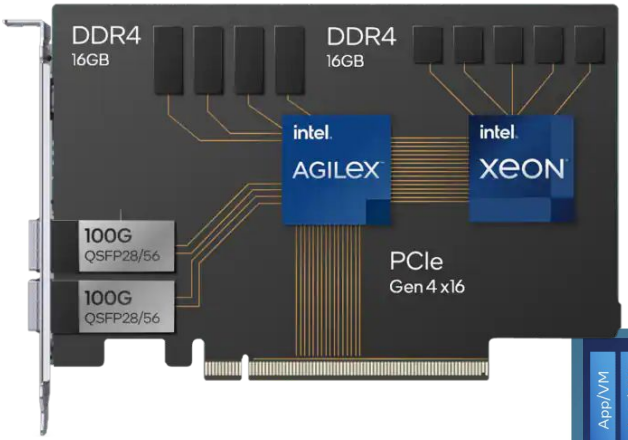
- Connected to host system with general purpose CPU
- Fast I/O to “host” main memory
  - E.g. over PCIe, OpenCAPI, CXL
- Often have local on-board memory
- Often have another form of I/O
  - E.g. SSD
  - E.g. Network
- On-board chip depends heavily on application domain
  - GPU - Embarrassingly parallel compute-intensive problems requiring many FLOPs
  - DPU - Data-intensive problems
  - FPGA - Anything faster than CPU but nobody has taped out an ASIC yet
  - ...

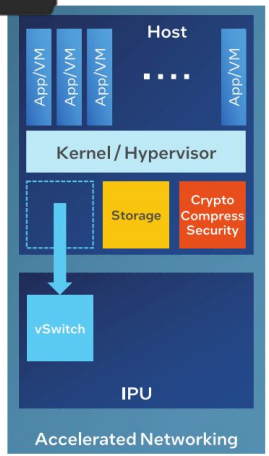
Example: Xilinx + Samsung SmartSSD

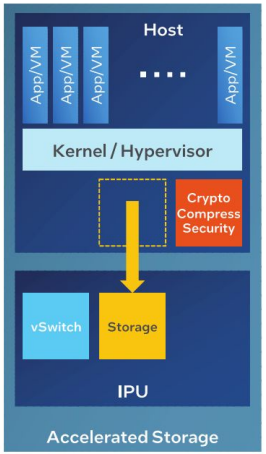


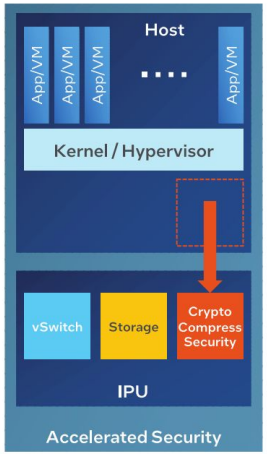
Source: Xilinx

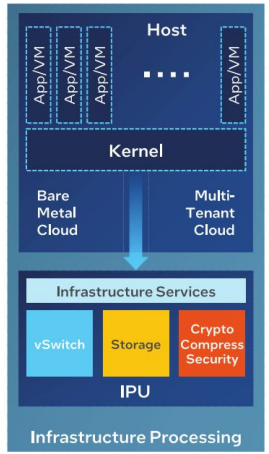
Example: Intel Infrastructure Processing Units





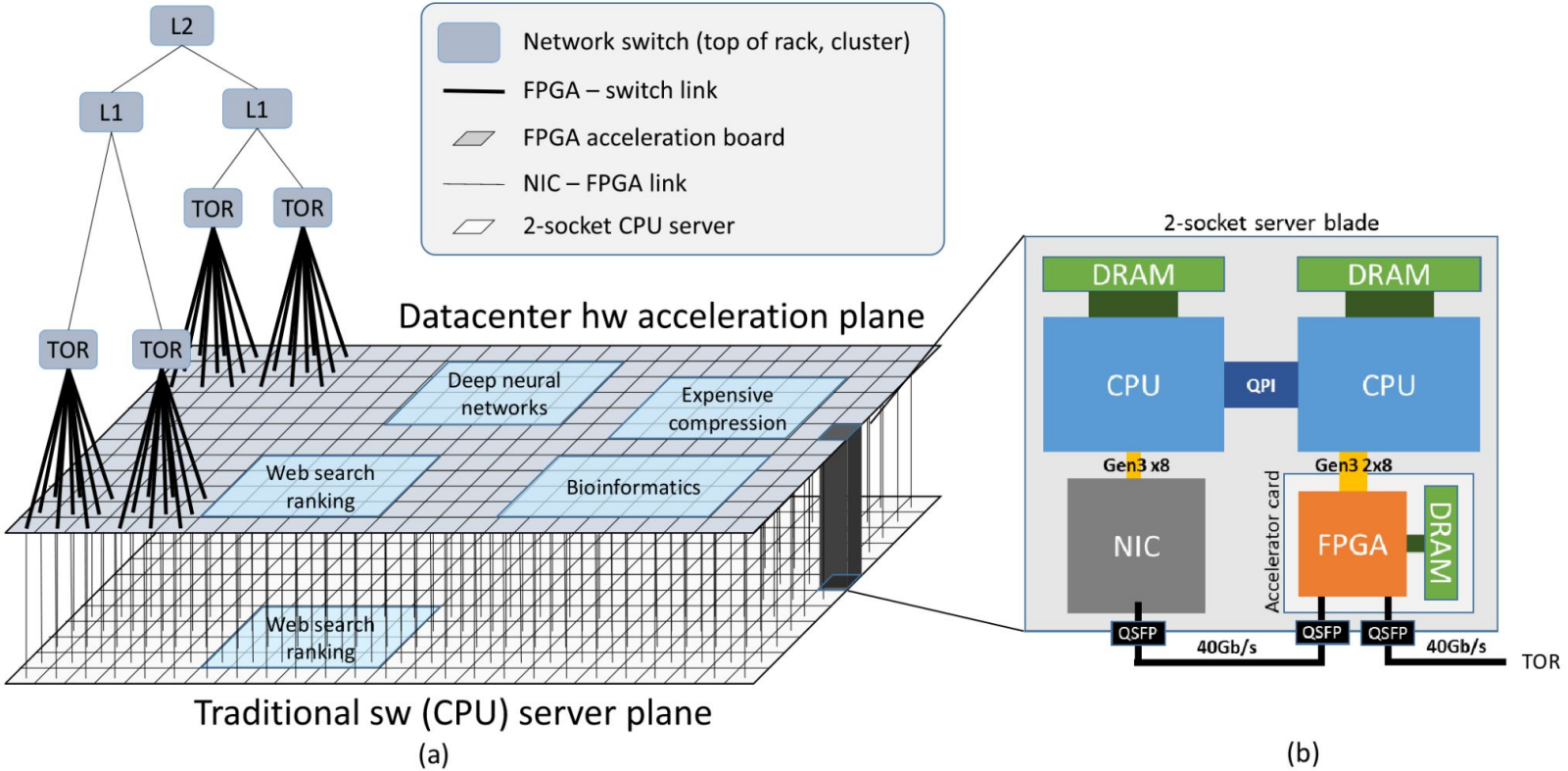






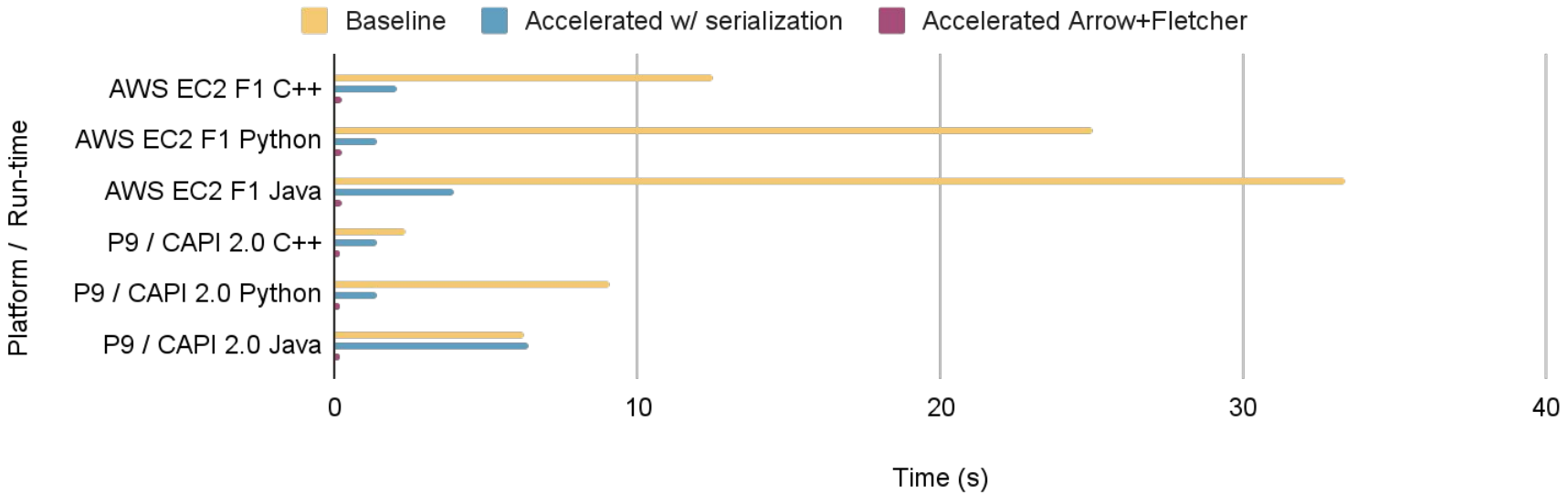
Source: Intel

Example: Microsoft Project Catapult

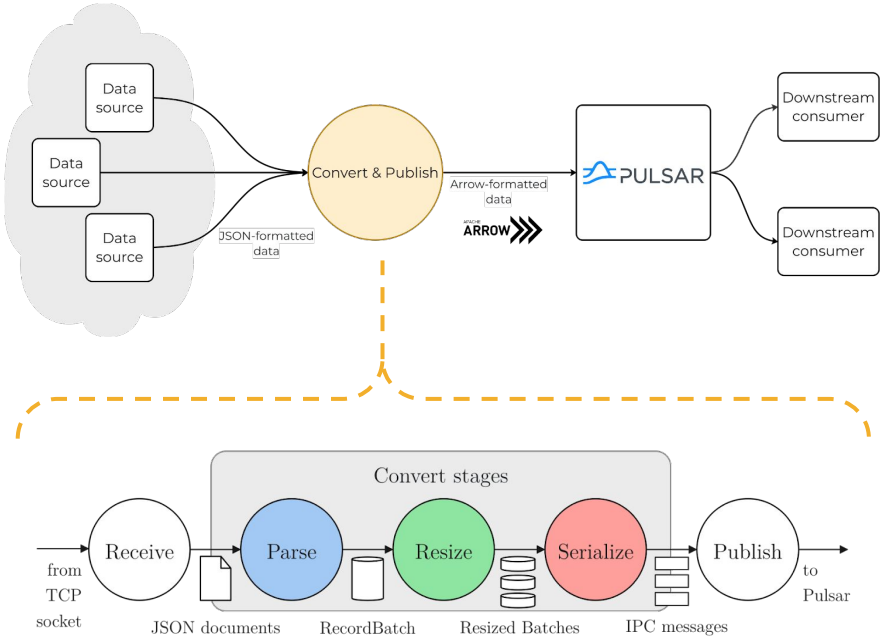


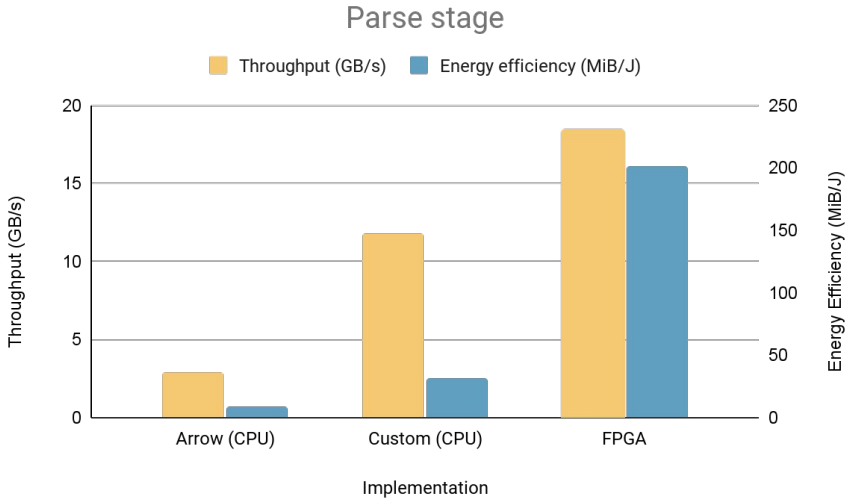
Source: Caulfield, Adrian M., et al. "A cloud-scale acceleration architecture." *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2016.

Example: Regular Expression Matching



Example: JSON parsing





System : Inspur Power Systems FP5290G2
CPU : 2x POWER9 (total 44 cores, 176 threads)
FPGA : Xilinx VU37P
Card : AlphaData ADM-PCIE-9H7
Interface: OpenCAPI 3.

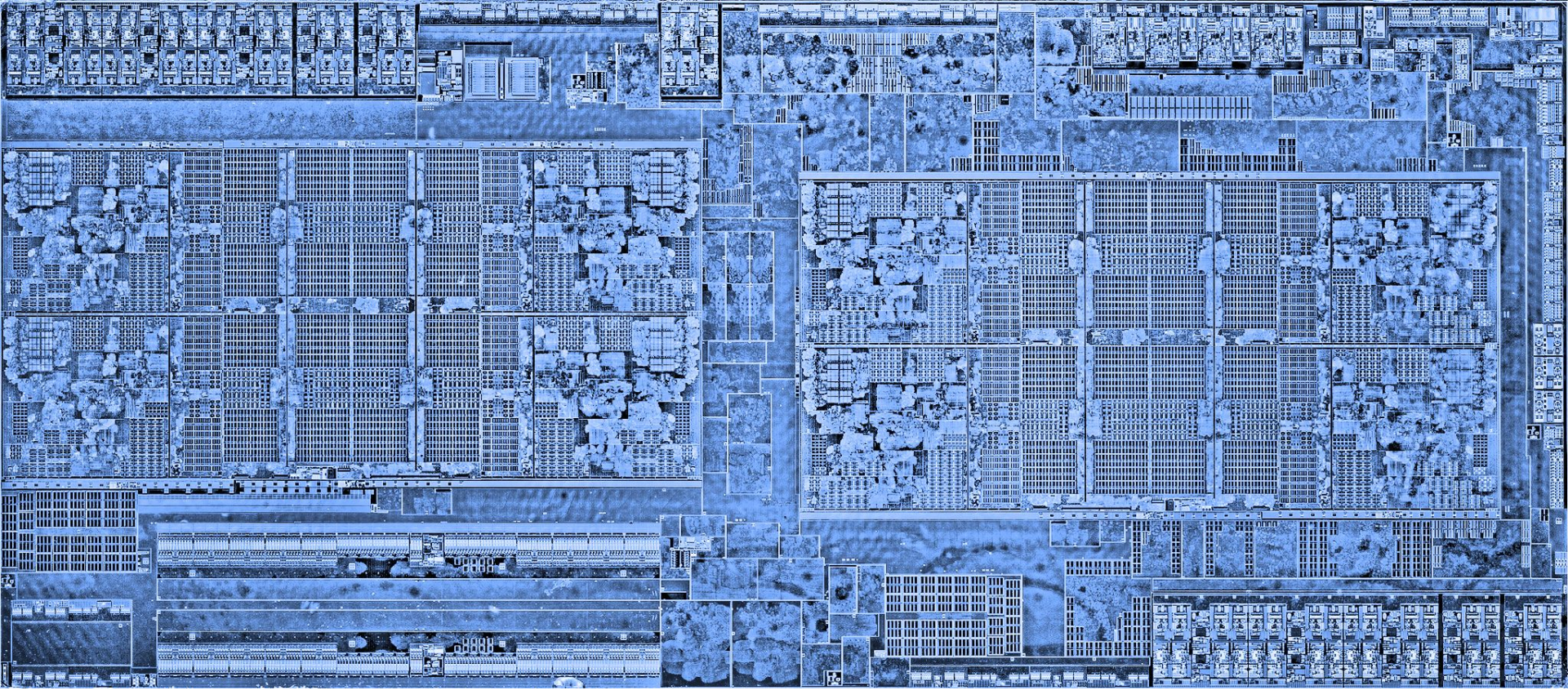
Summary

- Moore's Law continuing + various "walls"
  - Caused move from single-core to multi-core.
- Amdahl's Law in the Multicore Era
  - Depends on workload what configuration is best.
- Dark Silicon
  - Power density too high when all transistors operational
  - Need to use transistors in a smarter way
- Accelerators help on relative short-term
  - Specialize in specific application domains
  - Hot topic, especially in datacenter/cloud
  - Many kinds will appear (and disappear)
  - Lots of work to do for hardware engineers

Hardware Acceleration NL

- Group of hardware acceleration enthusiasts from Dutch research & industry
- http://hwacc.nl/

- This Friday - free online event:
  - Oct 29, 2021, 15:00 - 17:30
  - Sign up for free: https://lnkd.in/gzv_9xYS
- Talk topics:
  - Peter Hofstee, **IBM**, "Breaking the memory bottleneck in high-performance computing"
  - Rob de Jong, **Philips**, "High-performance image processing for medical applications: challenges and solutions"
  - Dirk van den Heuvel, **Topic Embedded Systems**, "Exascale high-performance computing: Infrastructural and modeling concept"
  - Matthijs Brobbel, **Teratide**, "Efficient data-centric computing using heterogeneous accelerator system"



AMD Zen die shot of the Ryzen 7 octa-core die. Source: WikiChip

CTA01++: multicore and beyond

Johan Peltenburg, PhD

Postdoc @ TU Delft
j.w.peltenburg@tudelft.nl

Co-founder @ Teratide
johan@teratide.io