Energy, power and other trends

Draft of project abstract and hypotheses due Friday!

Topics we've covered

- Data representations and digital computation
- Basic CPU execution
- Pipelined execution
- Caches
- Virtual memory
- Dynamic ILP
- Static ILP/Compiler considerations
- DLP
- GPUs
- ISA design
- Security

What sorts of tradeoffs (implicit or explicit) have we seen?

Ex: small/fast/expensive vs. large/slow/cheap memory

Ex: developer effort vs. program efficiency

Source

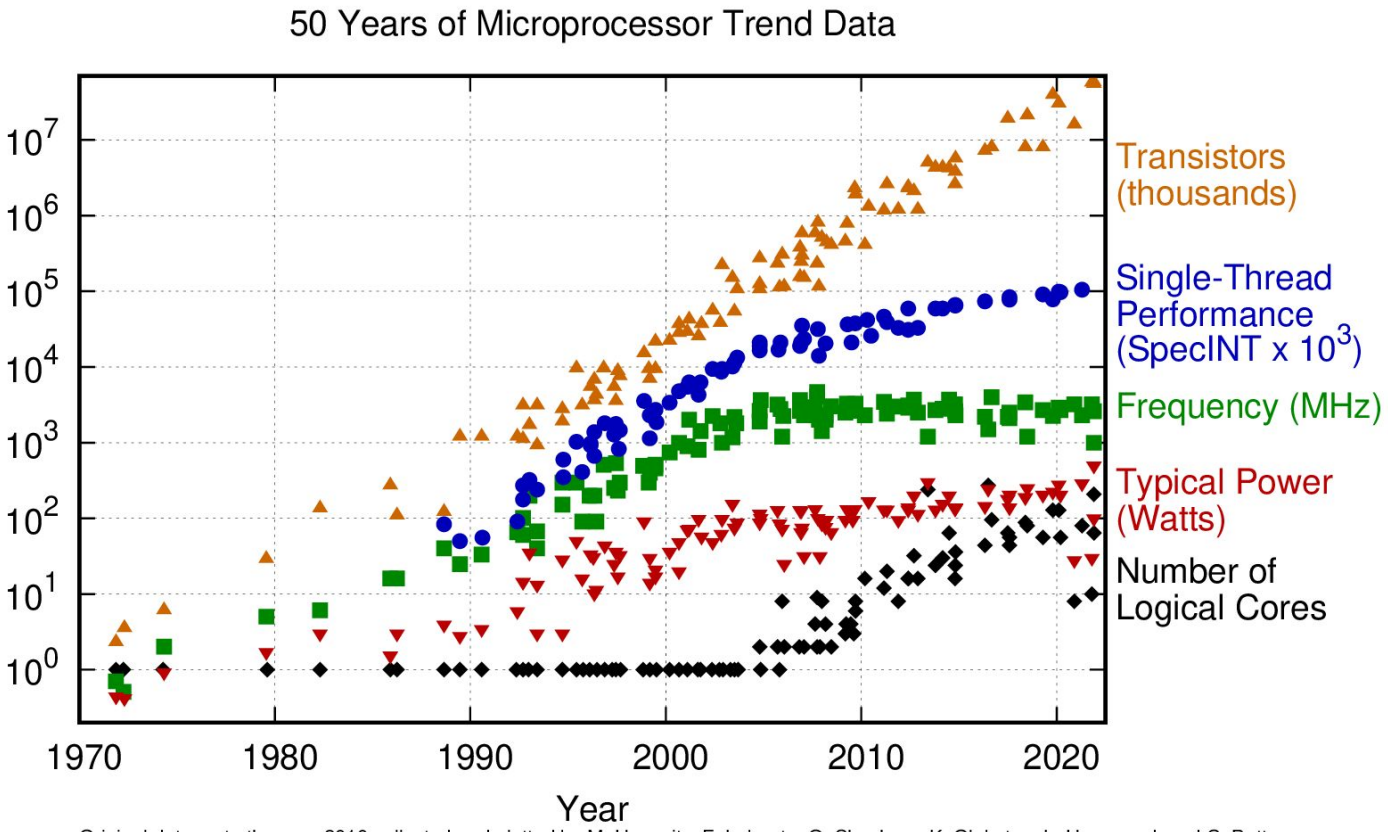

50 Years of Microprocessor Trend Data
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
1970
1980
1990
2000
2010
2020
Year
Transistors (thousands)
Single-Thread Performance (SpecINT x 10^3)
Frequency (MHz)
Typical Power (Watts)
Number of Logical Cores
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2021 by K. Rupp

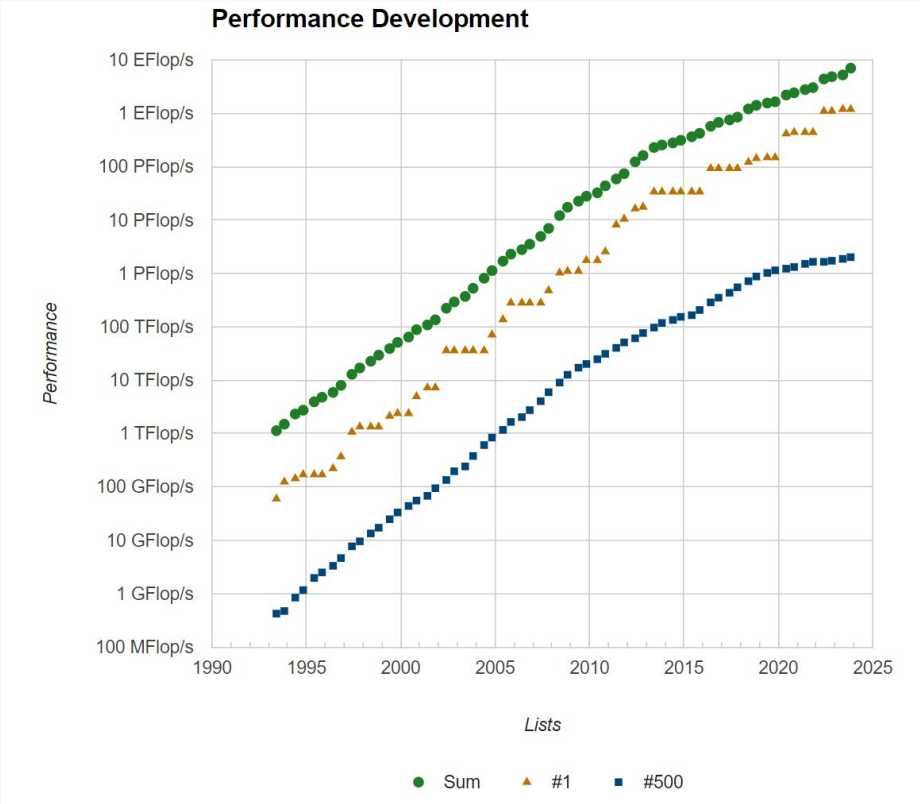
image source

The power wall

Limits to how far we can push a processor (clock frequency, transistor density) based on power limitations

Note: there are many opinions on what the limits of computer architecture are and when we'll reach them. So far, we keep seeing growth in one area or another

Top500 performance

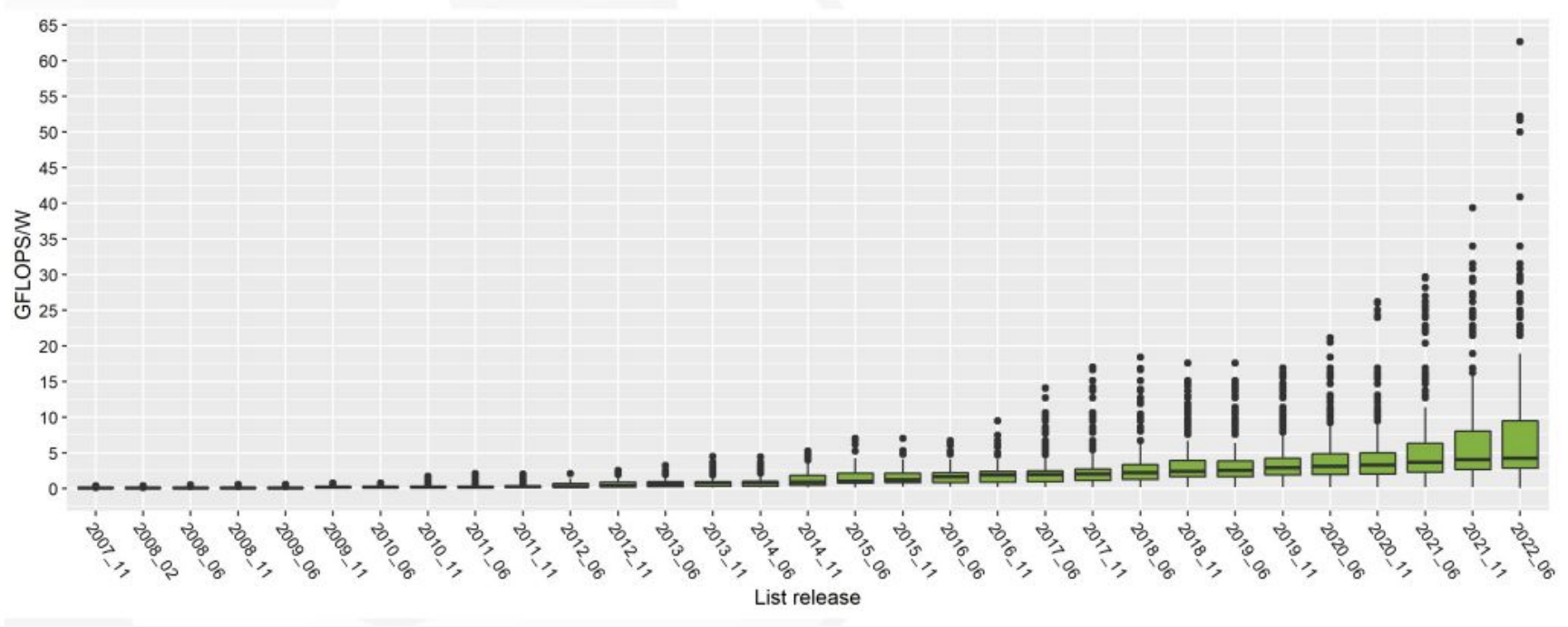


source

Green500 performance

source

Green500 data

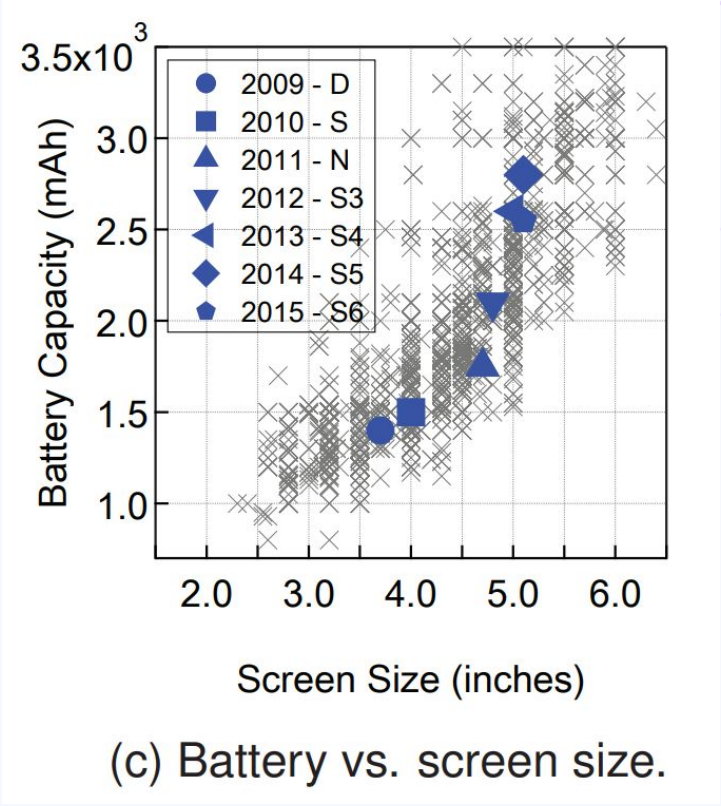


Mobile device challenges

"No Moore's Law for Batteries"

Unless battery technology sees drastic innovation, focus needs to be on energy efficiency

(Good news: mobile devices have different usage patterns than laptops/desktops)



(c) Battery vs. screen size.

M. Halpern, Y. Zhu and V. J. Reddi, "Mobile CPU's rise to power: Quantifying the impact of generational mobile CPU design trends on performance, energy, and user satisfaction," 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA), Barcelona, Spain, 2016, pp. 64-76, doi: 10.1109/HPCA.2016.7446054.
IEEE link

Power definitions

Power: energy per unit time (watt = joule/s)

Sustained power, or Thermal Design Power determines cooling requirements

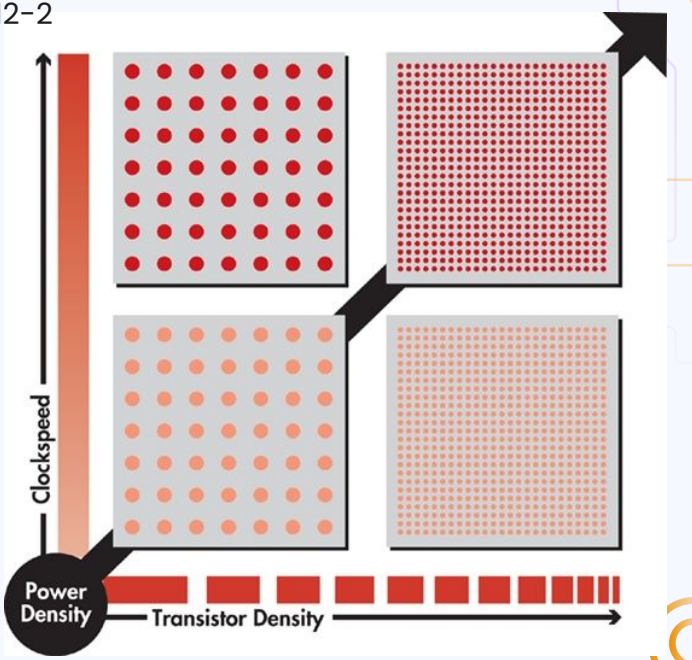
Dynamic power from switching transistors on/off

> Proportional to CV^2Af (Capacitance, Voltage, Activity Factor, frequency)

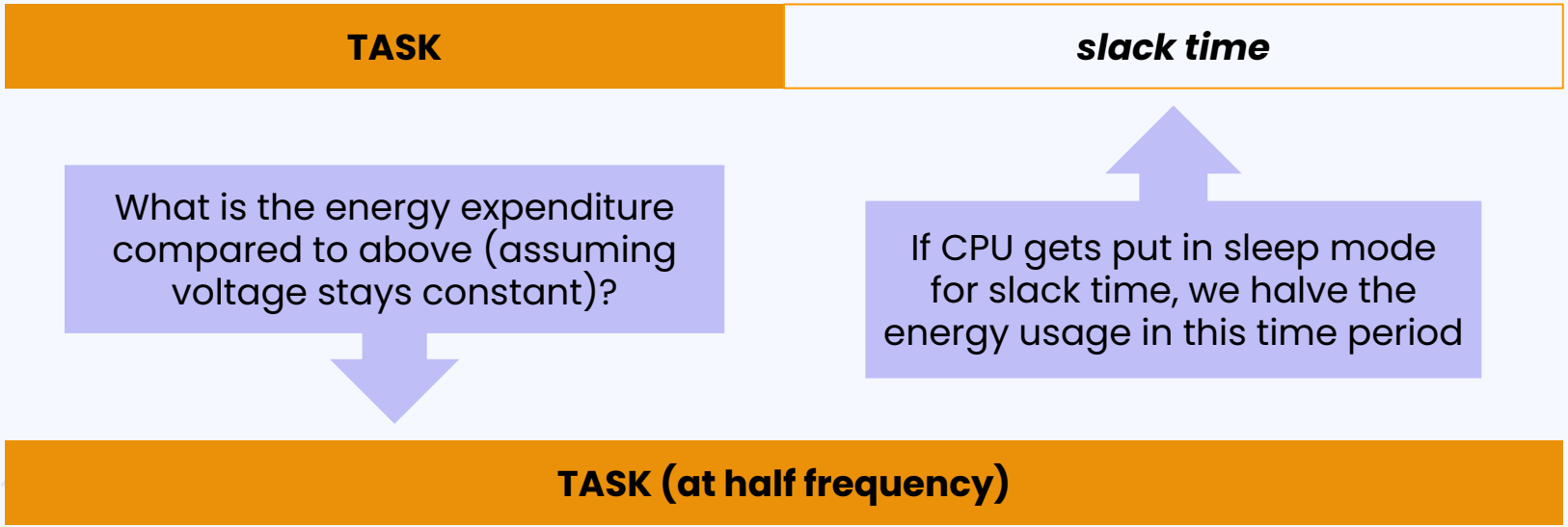
Static power from "leakage current" flowing even when transistors are off

> $I_{leak}V$ – increases as transistors shrink, increases w/ number of devices

Image source:
Stokes, Jon. *Inside the machine: an illustrated introduction to microprocessors and computer architecture*. No starch press, 2007.
Brown library link
Fig. 12-2



Adjusting clock frequency



But note that reducing frequency *also* allows us to reduce voltage! This allows us to save energy (*why?*)

DVFS

Dynamic voltage and frequency scaling

Adjusts voltage/frequency based on workload

Modern systems manage this at the OS level

- (CPUFreq governors on Linux)
- Based on coarse samples of system performance
- Require hardware interface

Other approaches to power management

Low-power modes when hardware isn't being used

- For example: I/O interrupts instead of polling
- *But* there is a latency cost to coming out of low-power mode

Clock gating

- Turn off clock to idle unit (reduces useless switching)
- IBM Power5 claim: 25% reduction in switching power w/o reduction in perf

Detecting narrow-width operands

- If buses are width 64 but using only 32 bits, disable those wires

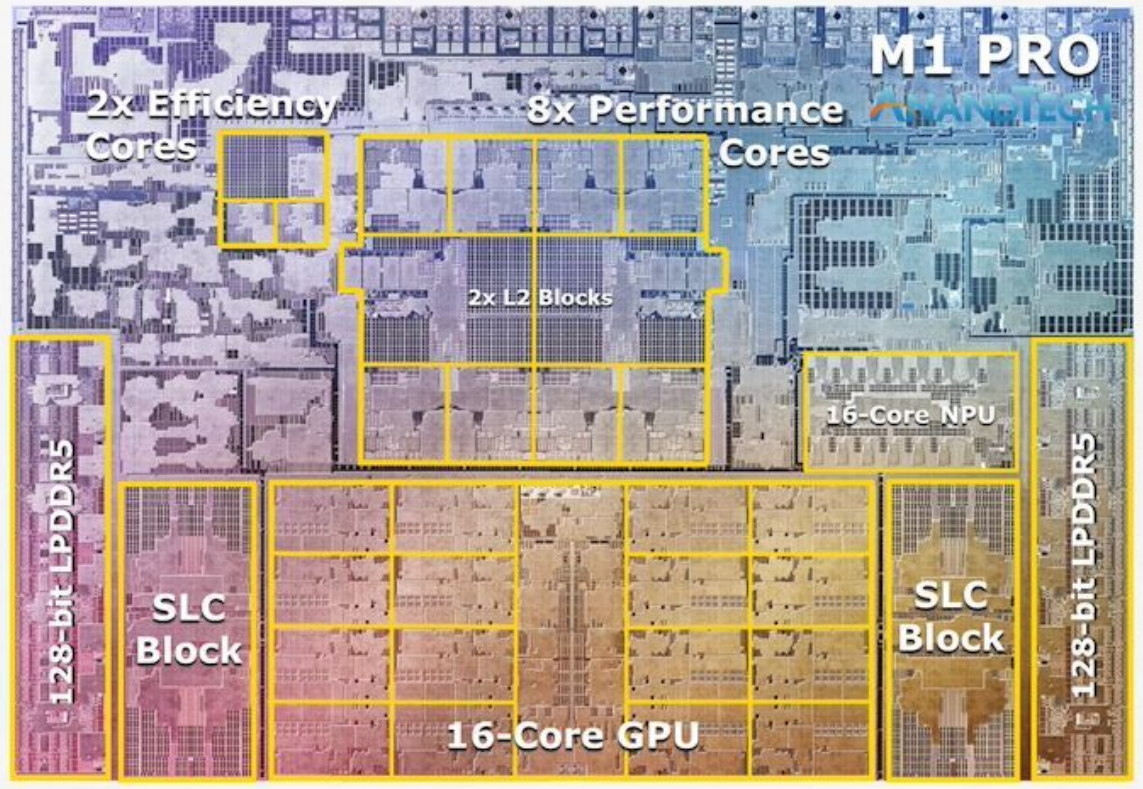
Heterogeneous architectures

Image source

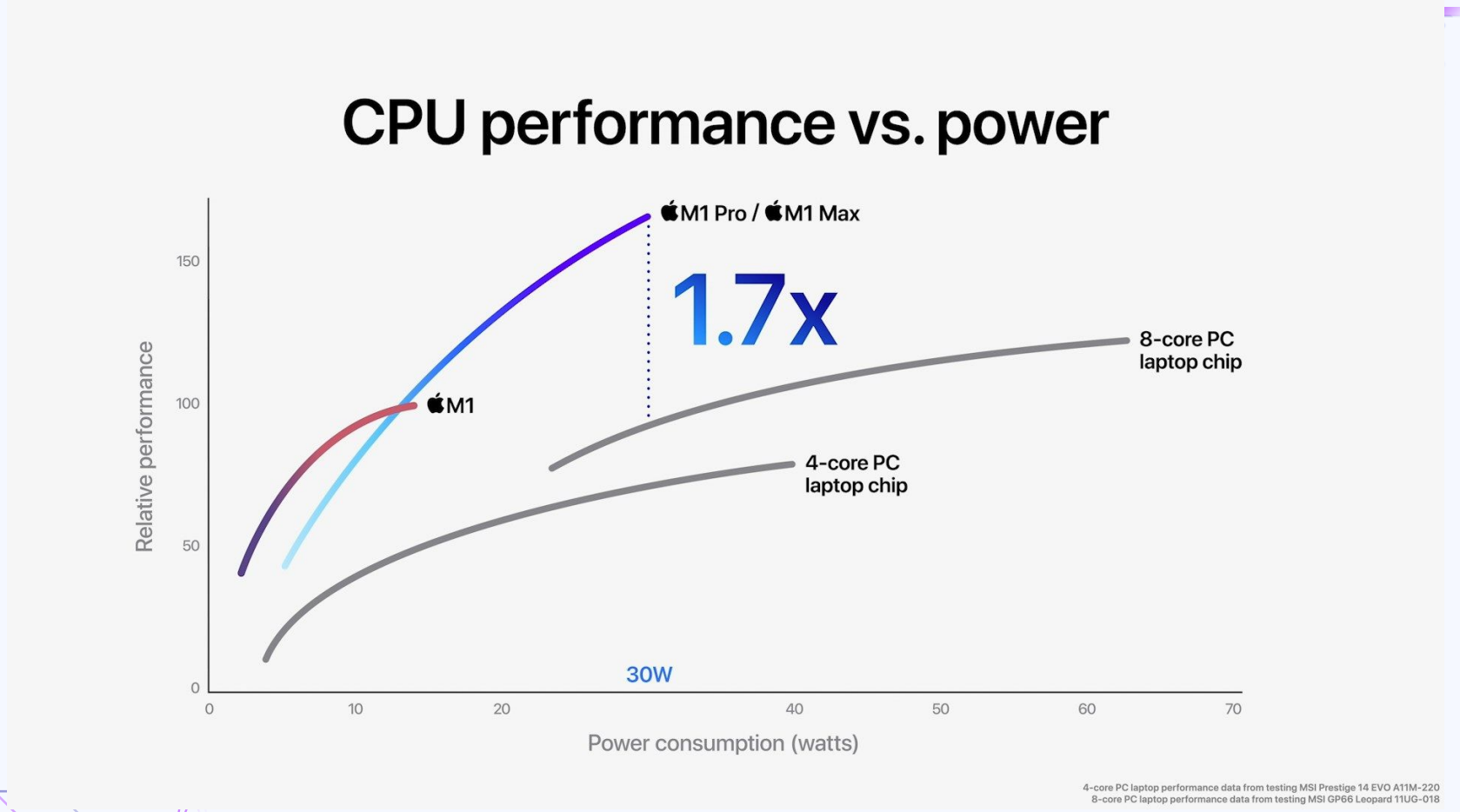
Use different types of processors (also called asymmetric) or processing units for different tasks

Allows system to use efficient hardware for specific applications

Challenges: coordination, scheduling, design



Claims by Apple, Inc (taken from AnandTech)



Domain-Specific Architecture

Specialized hardware for software domain

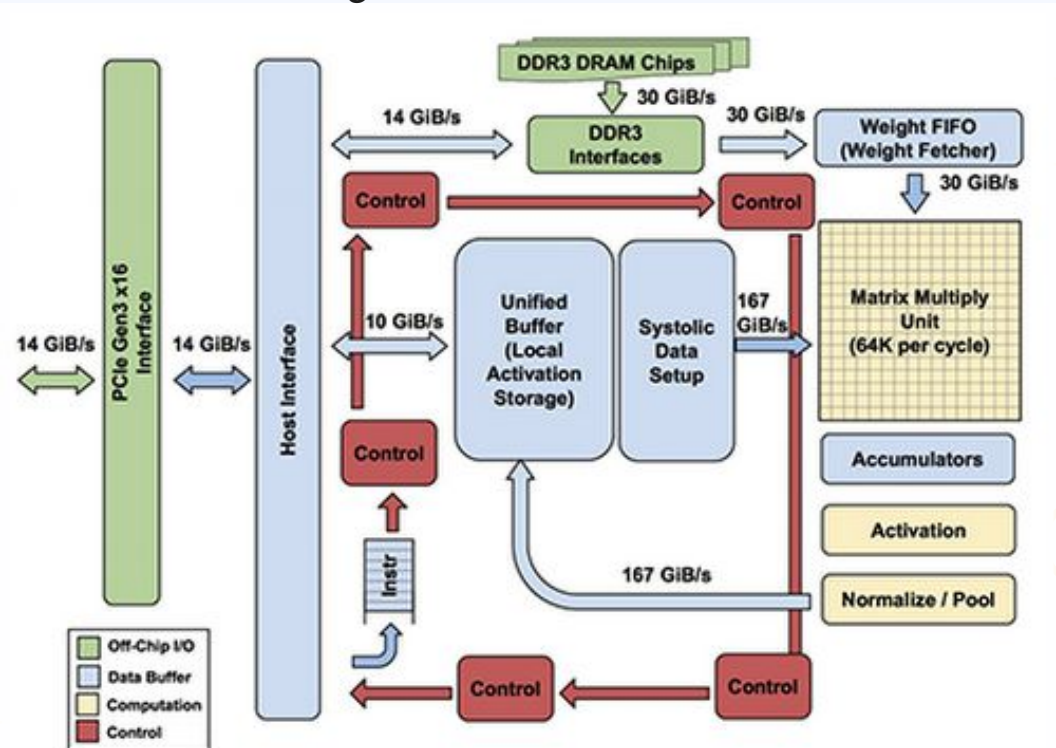
Can adapt memory, precision, parallelism to application

Increase both performance and energy efficiency!

When/how to justify design cost for specialized hardware?

What implications does this have on longevity?

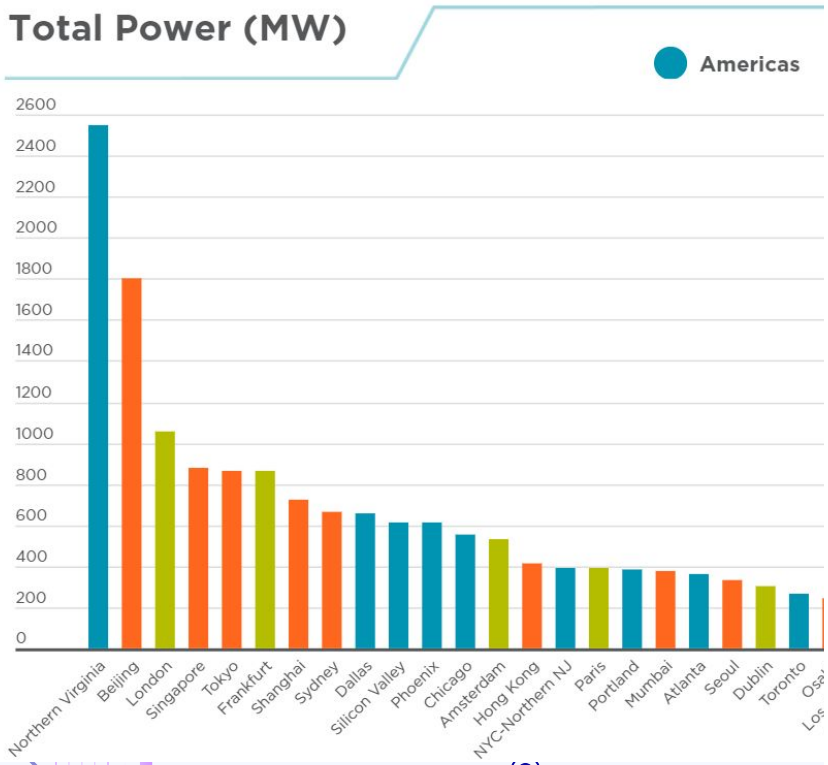
Google TPU source

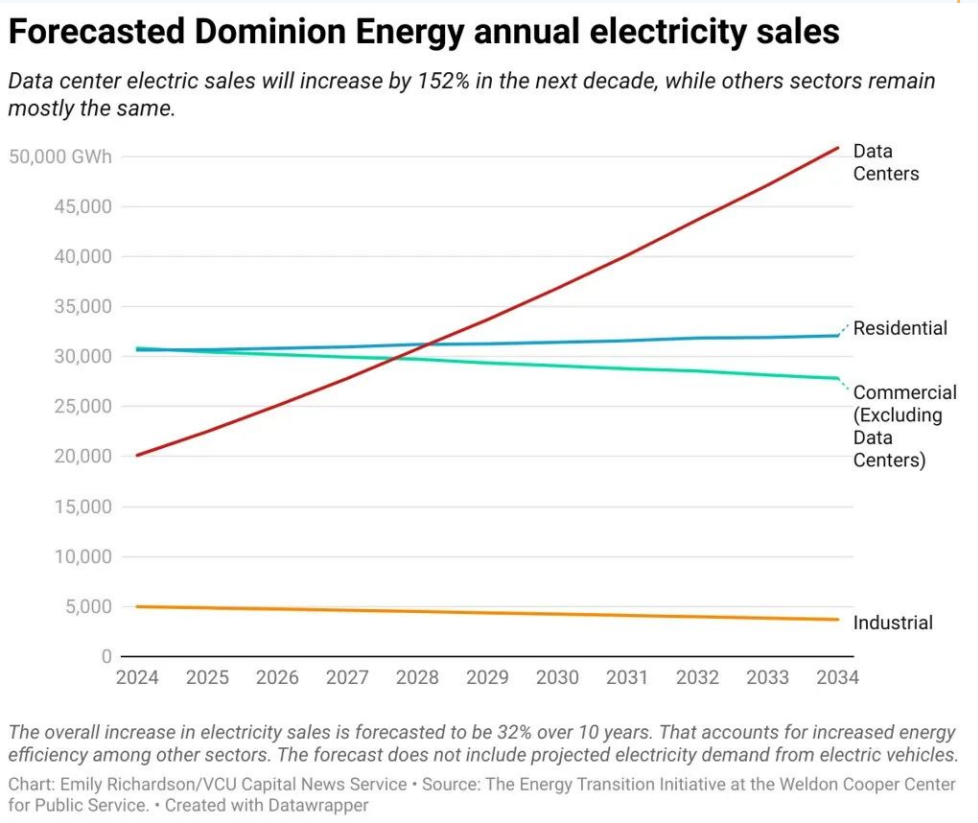


Source: Cushman & Wakefield

The other concerns with power

source





Related reading/viewing

Jim Keller: Moore's Law is Not Dead (argues that, whenever anyone says that some technology has reached its limits and cannot scale, people find ways to come up with a new technology)

Charles E. Leiserson et al. ,There's plenty of room at the Top: What will drive computer performance after Moore's law? Science 368, eaam9744(2020).

Hennessy, John L., and David A. Patterson. "A new golden age for computer architecture." Communications of the ACM 62.2 (2019): 48-60.

Kaxiras, Stefanos, and Margaret Martonosi. Computer architecture techniques for power-efficiency. Morgan & Claypool, 2008.

A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi and J. Kepner, "AI and ML Accelerator Survey and Trends," 2022 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 2022, pp. 1-10, doi: 10.1109/HPEC55821.2022.9926331.