NE 155
Introduction to Numerical Simulations in Radiation Transport
Lecture 2: Computing

R. N. Slaybaugh

January 21, 2017

Outline

1. History and terminology
2. Basic computer architecture
3. Introduction to Parallelism
4. Current supercomputers

How do we measure utility?[1]

IPS (Instructions Per Second) is a measure of a computer's processor speed. IPS can be useful when comparing performance between processors made from a similar architecture, but are difficult to compare between CPU architectures

Clock rate typically refers to the frequency at which a CPU is running. It is measured in the SI unit Hertz.

FLOPS (FLoating-point Operations Per Second) is a measure of computer performance, useful in fields of scientific calculations that make heavy use of floating-point calculations.

[1] en.wikipedia.org

Computing Machines, Origins

Figure 1: Roman Abacus,
http://history-computer.com/CalculatingTools/abacus.html

Early Development of Computing

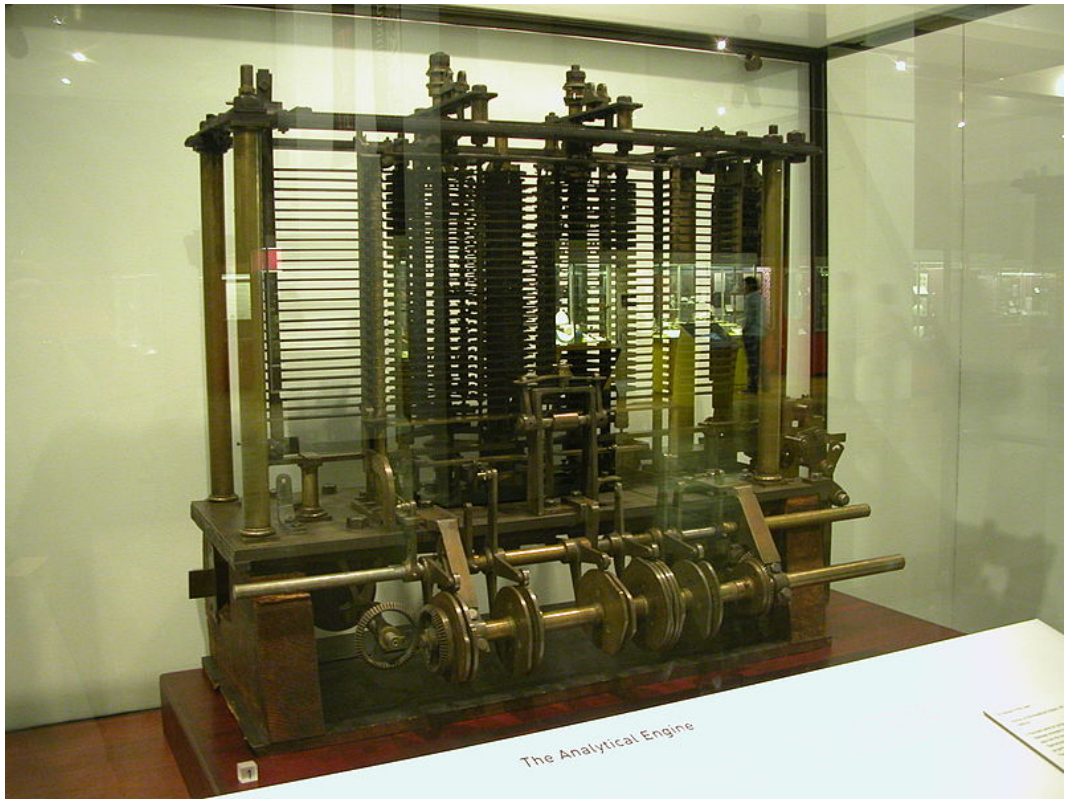


Figure 2: Reconstruction of Babbage's Analytical Engine, the first general-purpose programmable computer, http://en.wikipedia.org/wiki/History_of_computing_hardware#Punched_card_data_processing

First Electromechanical Computers

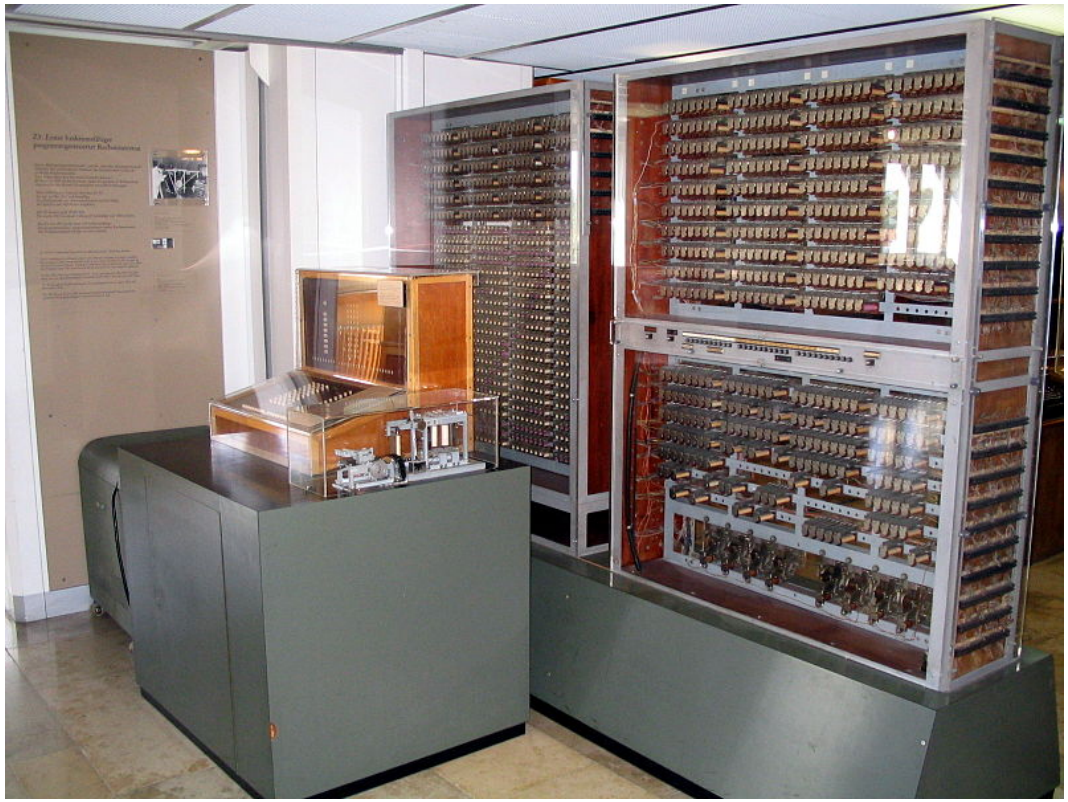


Figure 3: Zuse Z3 replica on display at Deutsches Museum in Munich, http://en.wikipedia.org/wiki/Z3_(computer)

Stored Programs

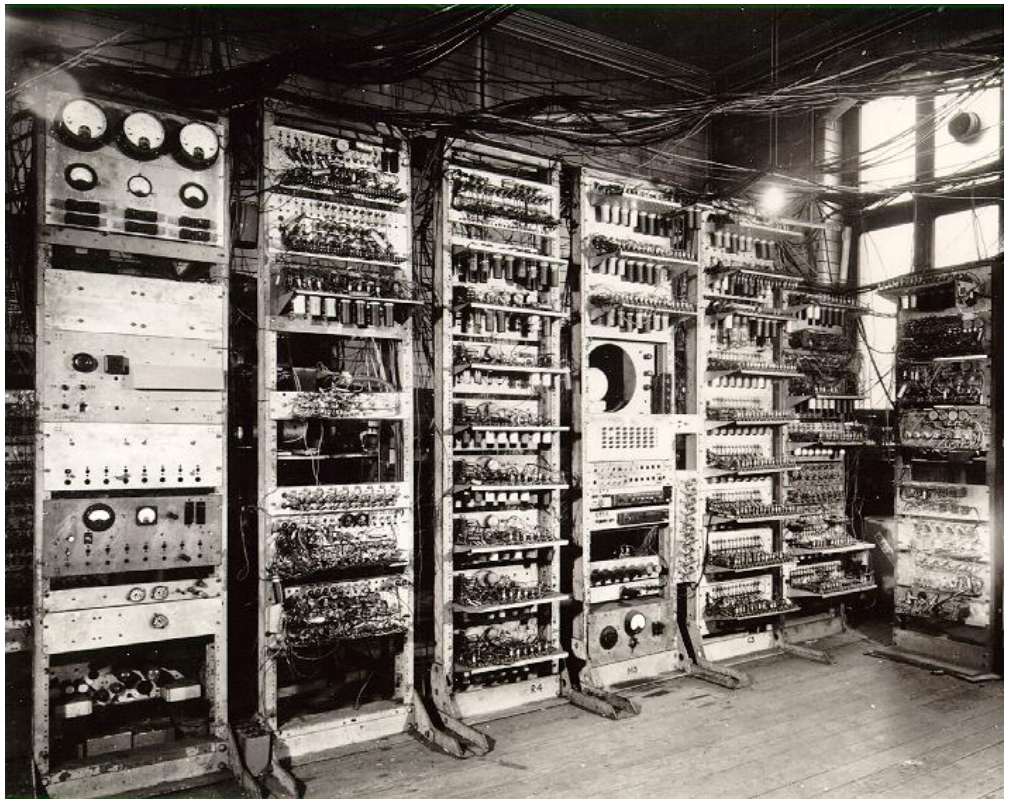


Figure 4: The Manchester Mark 1 was one of the world's first stored-program computers, http://www.computer50.org/mark1/ip-mm1.mark1.html

Microprogramming, Magnetic Storage, Transistors

- 1951: realization that CPUs can be controlled by a miniature, highly specialized computer program in high-speed ROM
- 1954: magnetic core memory was rapidly displacing most other forms of temporary storage
- 1956: IBM introduced the first disk storage unit: using 50 24-inch metal disks, it stored 5 MB of data for $10,000 per MB ($90,000 in 2014 $s)
- 1947: invention of the bipolar transistor; this replaced vacuum tubes by 1955 $\rightarrow$ "Second Generation" of computer designs

SUPERCOMPUTERS

Figure 5: The University of Manchester Atlas 1963, http://en.wikipedia.org/wiki/History_of_computing_hardware#Punched_card_data_processing

Integrated Circuit

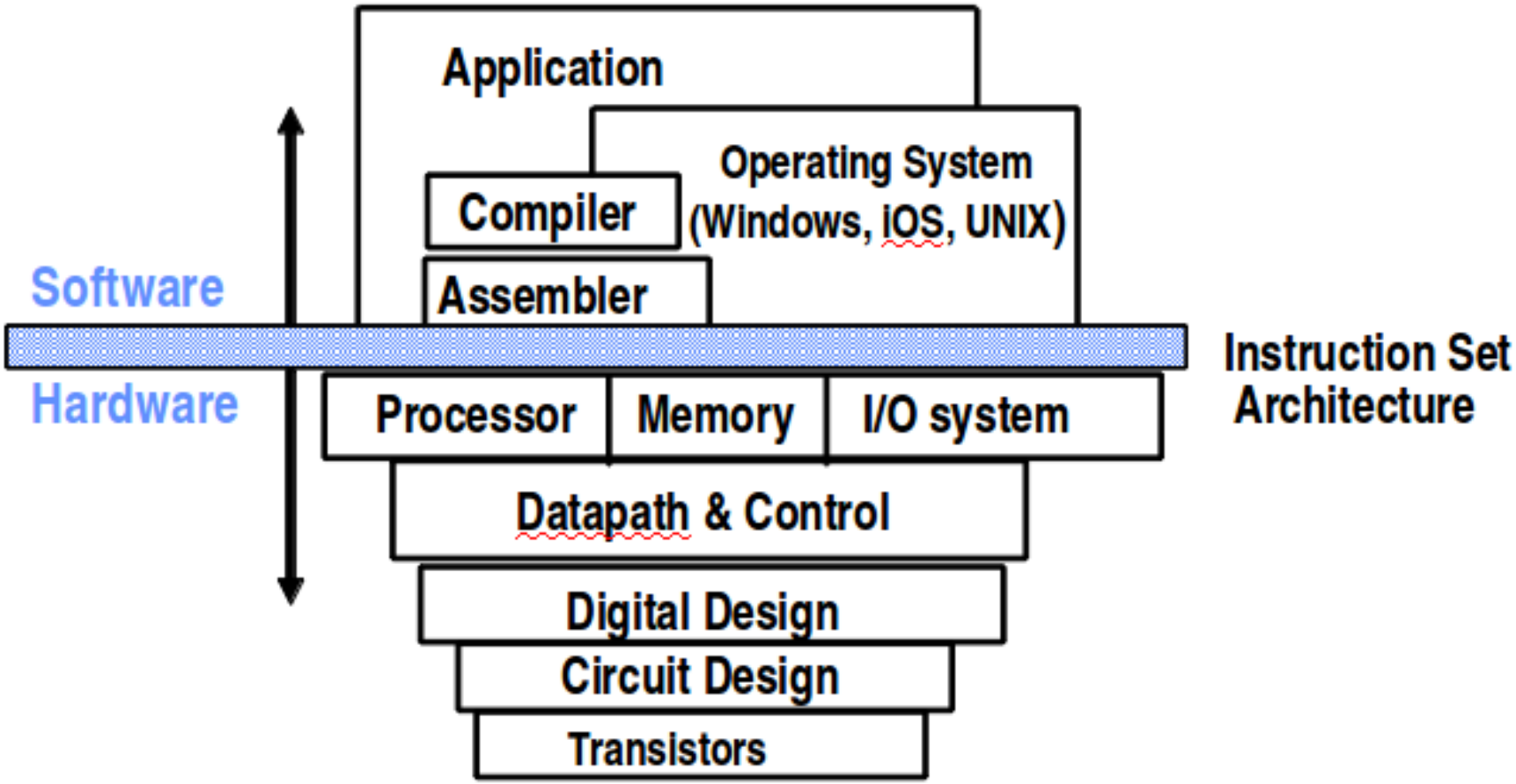
With the advent of the transistor and the work on semi-conductors generally, it now seems possible to envisage electronic equipment in a solid block with no connecting wires. The block may consist of layers of insulating, conducting, rectifying and amplifying materials, the electronic functions being connected directly by cutting out areas of the various layers.

Geoffrey W.A. Dummer, Royal Radar Establishment of the Ministry of Defence, 1952

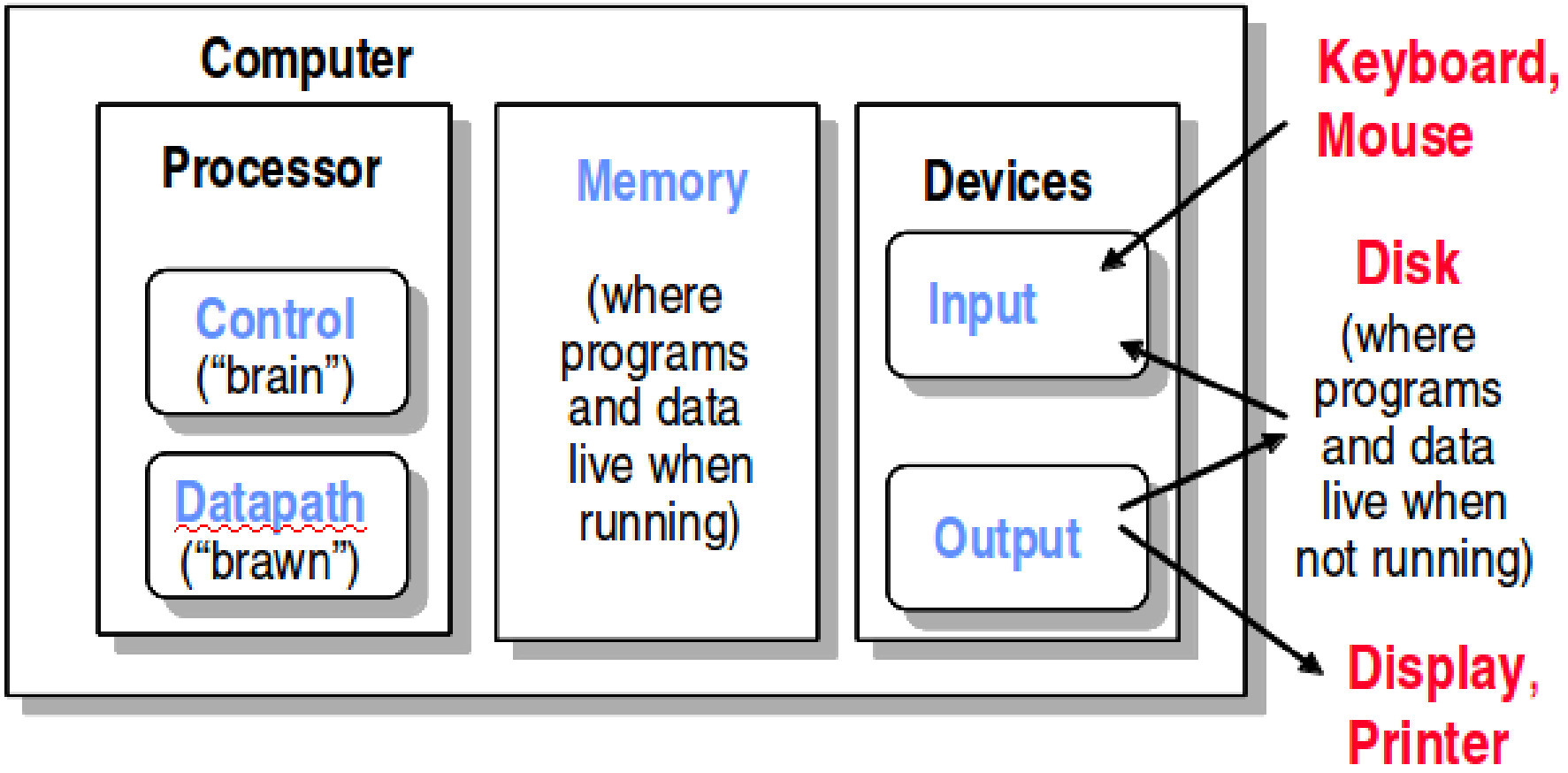
Generations 4-6

- Very large scale integration of devices on chip
- C and FORTRAN programming languages
- UNIX operating system (Bell labs, Berkeley)
- Large scale parallel processing; supercomputing centers
- Shared and distributed memory
- Parallel/vector shared/distributed memory combinations
- High speed networking

Computer Architecture



COMPONENTS


Computer
Processor
Control
("brain")
Datapath
("brawn")
Memory
(where
programs
and data
live when
running)
Devices
Input
Output
Keyboard,
Mouse
Disk
(where
programs
and data
live when
not running)
Display,
Printer

Moore's "Law"

The number of transistors on integrated circuits doubles approximately every 18 months.

- In 1965 Gordon E. Moore described this trend and predicted it to hold for at least 10 years
- He worked at Intel
- The trend has held, but is expected to change around now...

Moore's "Law"

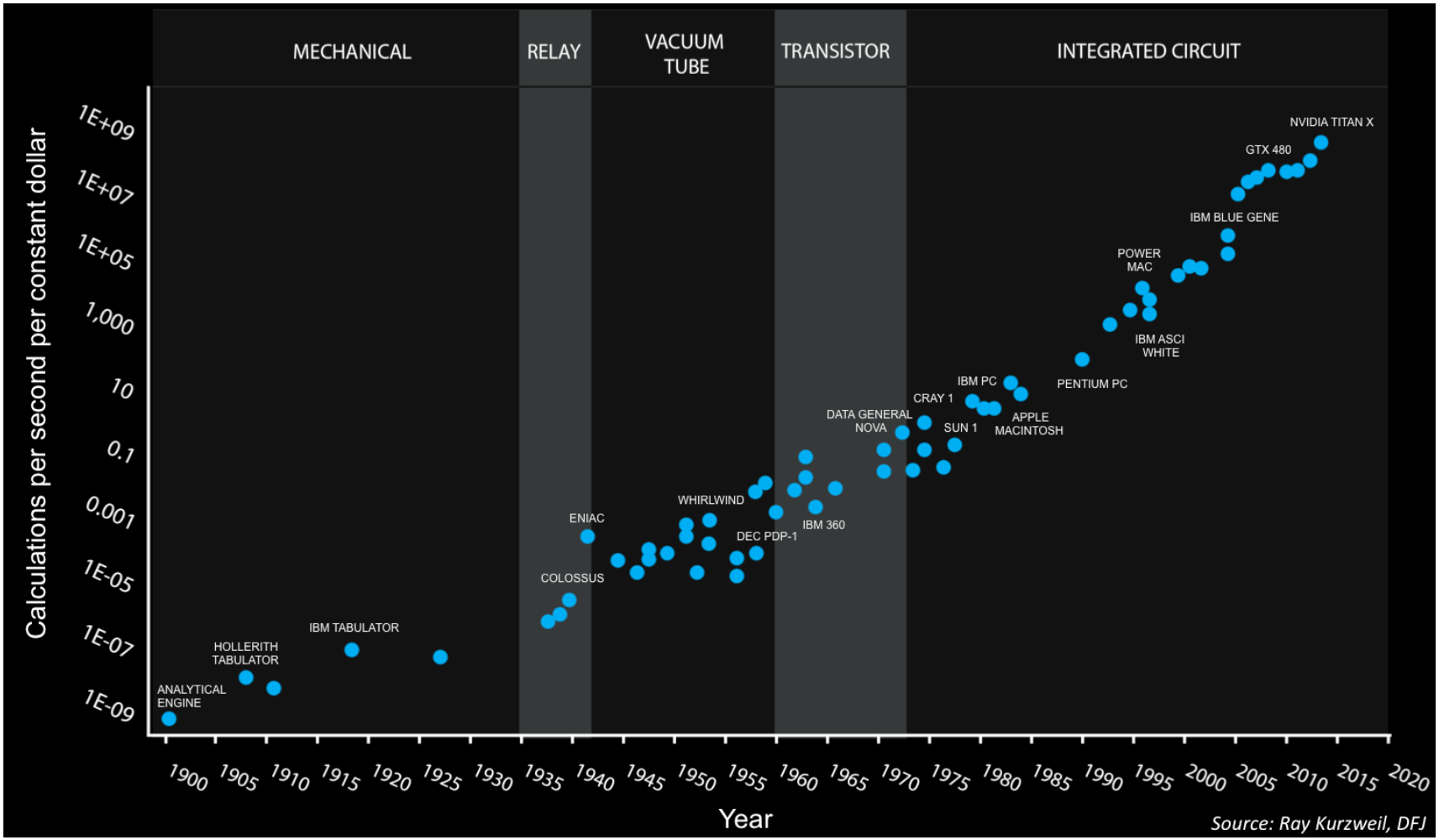


Figure 6: By Steve Jurvetson - https://www.flickr.com/photos/jurvetson/31409423572/, CC BY 2.0

What is Parallel Architecture?

A parallel computer is a collection of processing elements that cooperate to solve large problems quickly

- Resource allocation
  - How much memory?
  - How many elements?
  - How powerful are the elements?
- Data access, Communication, and Synchronization
  - How do the elements cooperate and communicate?
  - How are data transmitted between processors?
  - What are the abstractions and primitives for cooperation?
- Performance and Scalability
  - How does it all translate into performance?
  - How does it scale?

Forms of Parallelism

- **Bit level**: increases in word size reduced the # of instructions the processor needs
- **Instruction level**: hardware and/or software perform operations simultaneously when possible
- **Memory system**: overlap of memory operations with computation
- **Operating system**: multiple jobs run in parallel on commodity symmetric multiprocessors (SMPs)

There are limitations to all of these

To achieve high performance, the programmer needs to identify, schedule, and coordinate parallel tasks and data

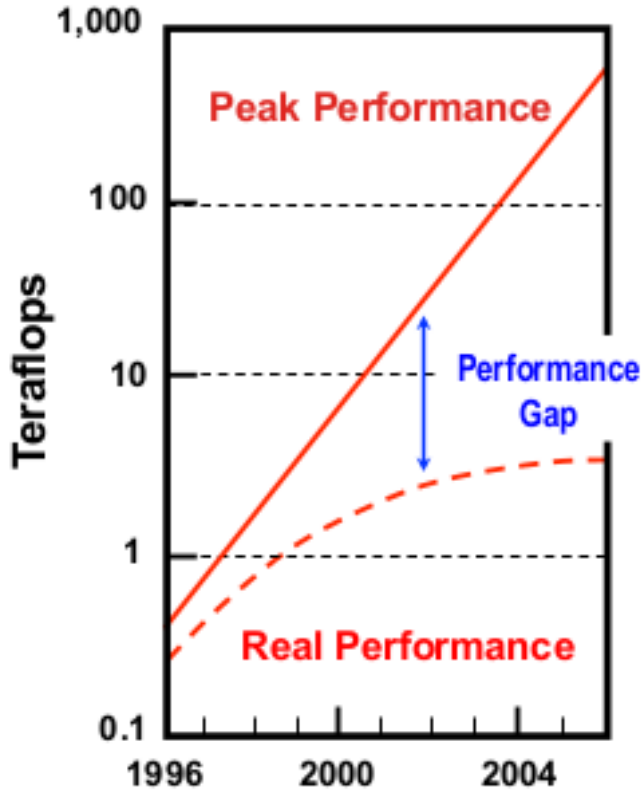
Performance

- **Strong Scaling**: increase element count with problem size fixed (solve a problem faster)

$$Speedup = \frac{\text{Time with P cores}}{\text{Time with 1 core}}$$

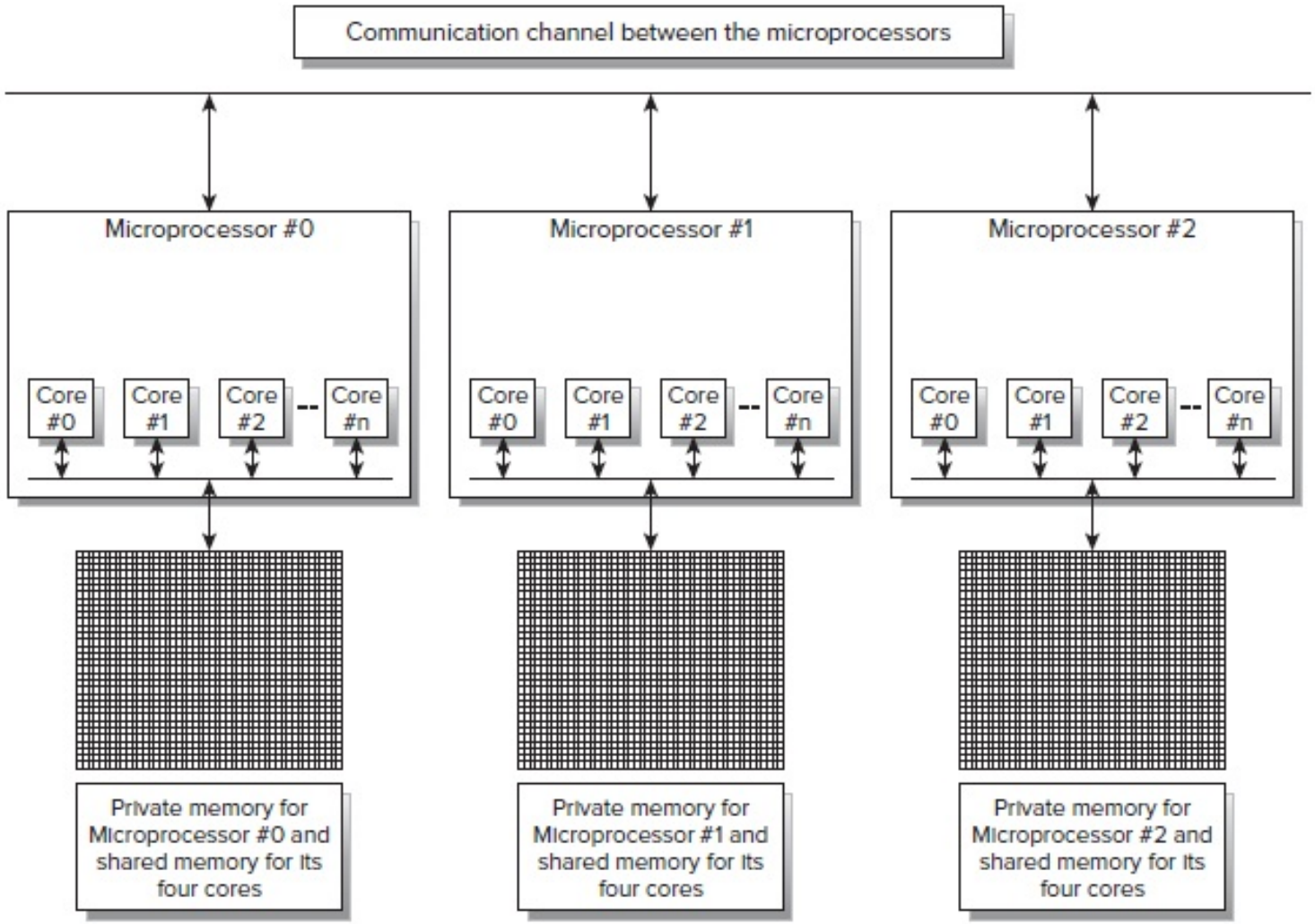
- **Weak Scaling**: increase element count and problem size to keep problem size per element fixed (solve bigger problems)

$$Speedup = \frac{\text{Time with 1 core}}{\text{Time with P cores}}$$



Types of Memory

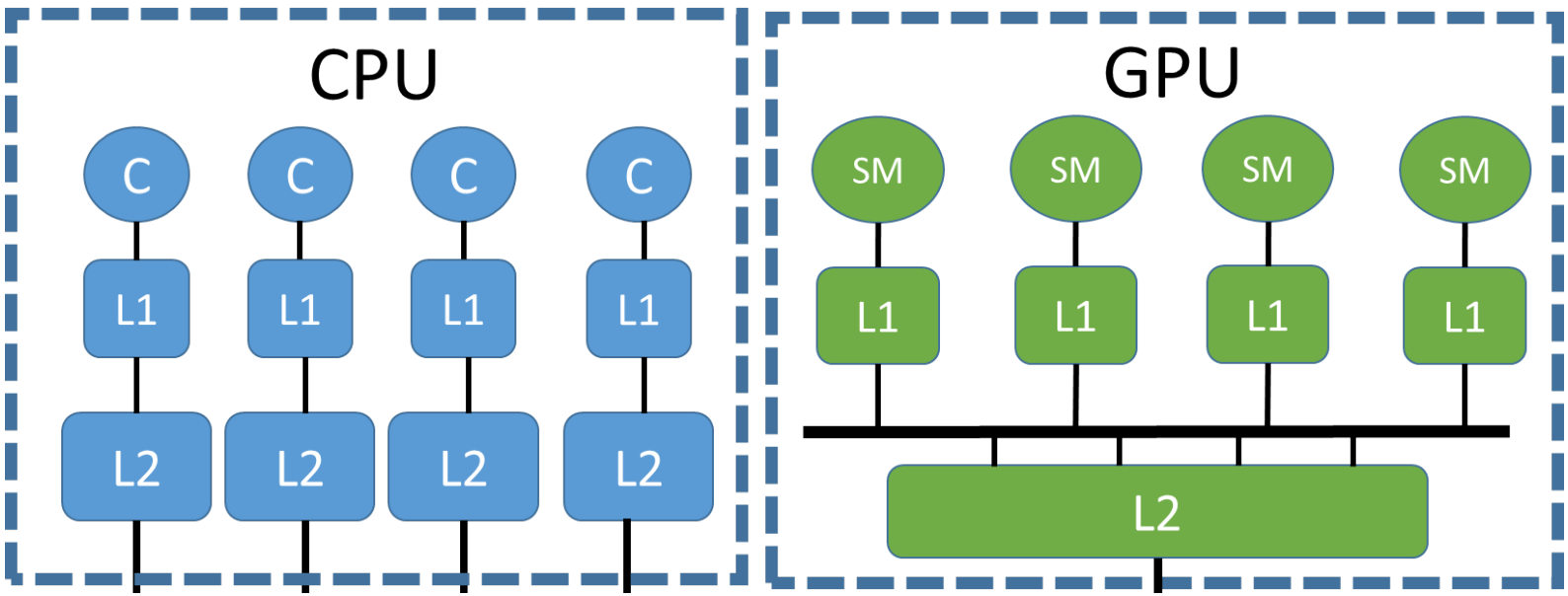
Shared Distributed Distributed Shared



GPUs

- Graphics Processing Unit (**GPU**): highly parallel, good for processing large blocks of data
- General Purpose GPU (**GPGPU**): Using a GPU to do CPU work - computational science instead of graphics

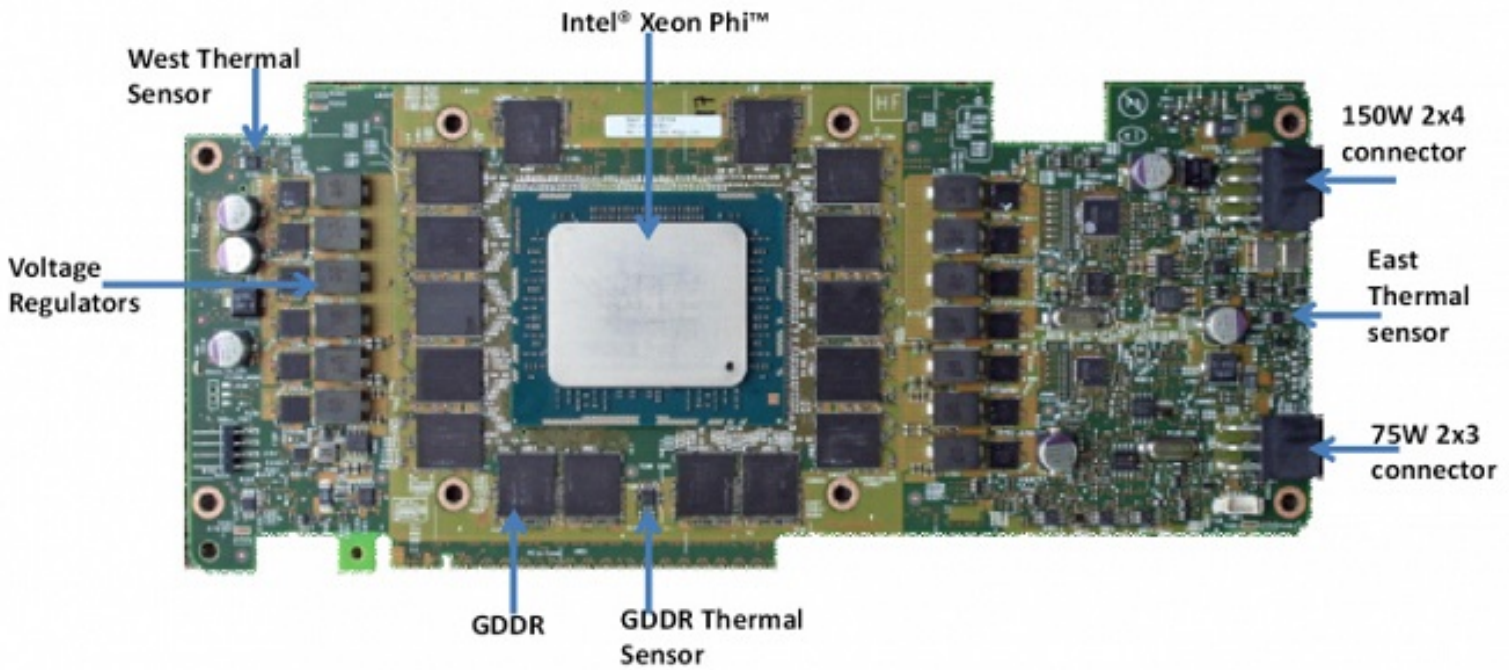
CPU vs. GPU memory structure



MICs

- Many Integrated Core (**MIC**): combines many CPU cores onto a single chip
- **Heterogeneous** architecture: GPUs + CPUs or MICs + CPUs, etc.

Intel Xeon Phi Board



Top 500 Computers, Nov 2016

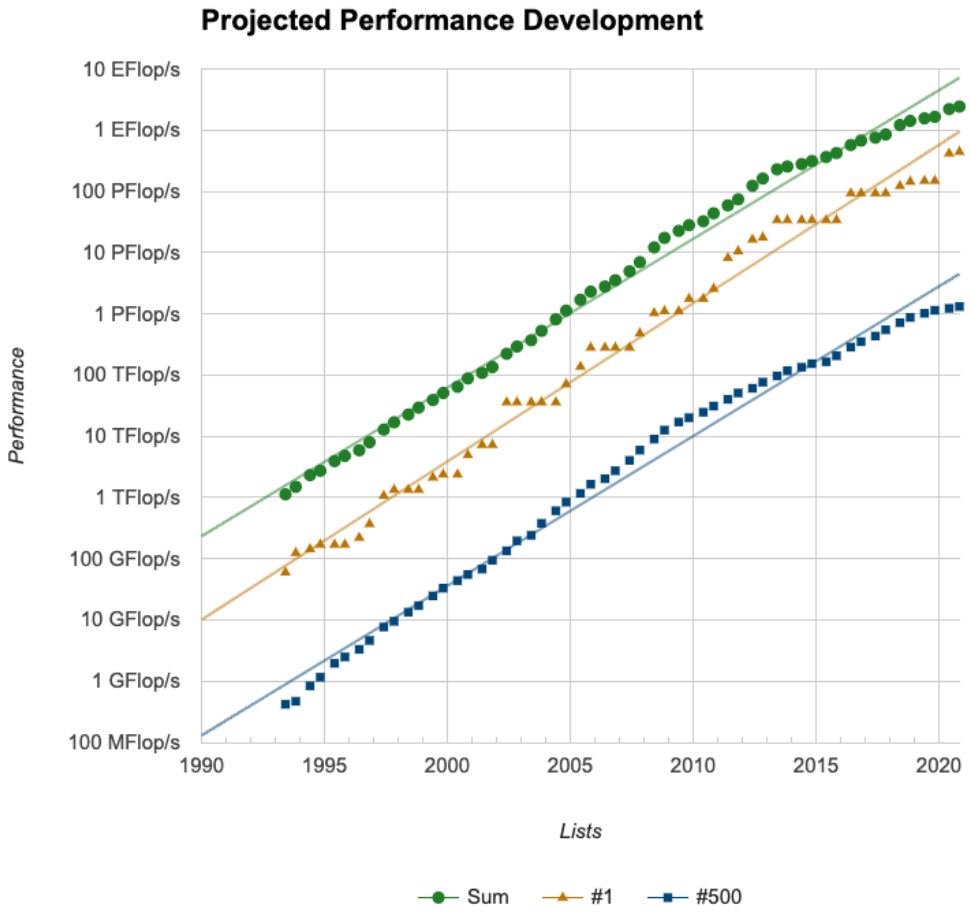


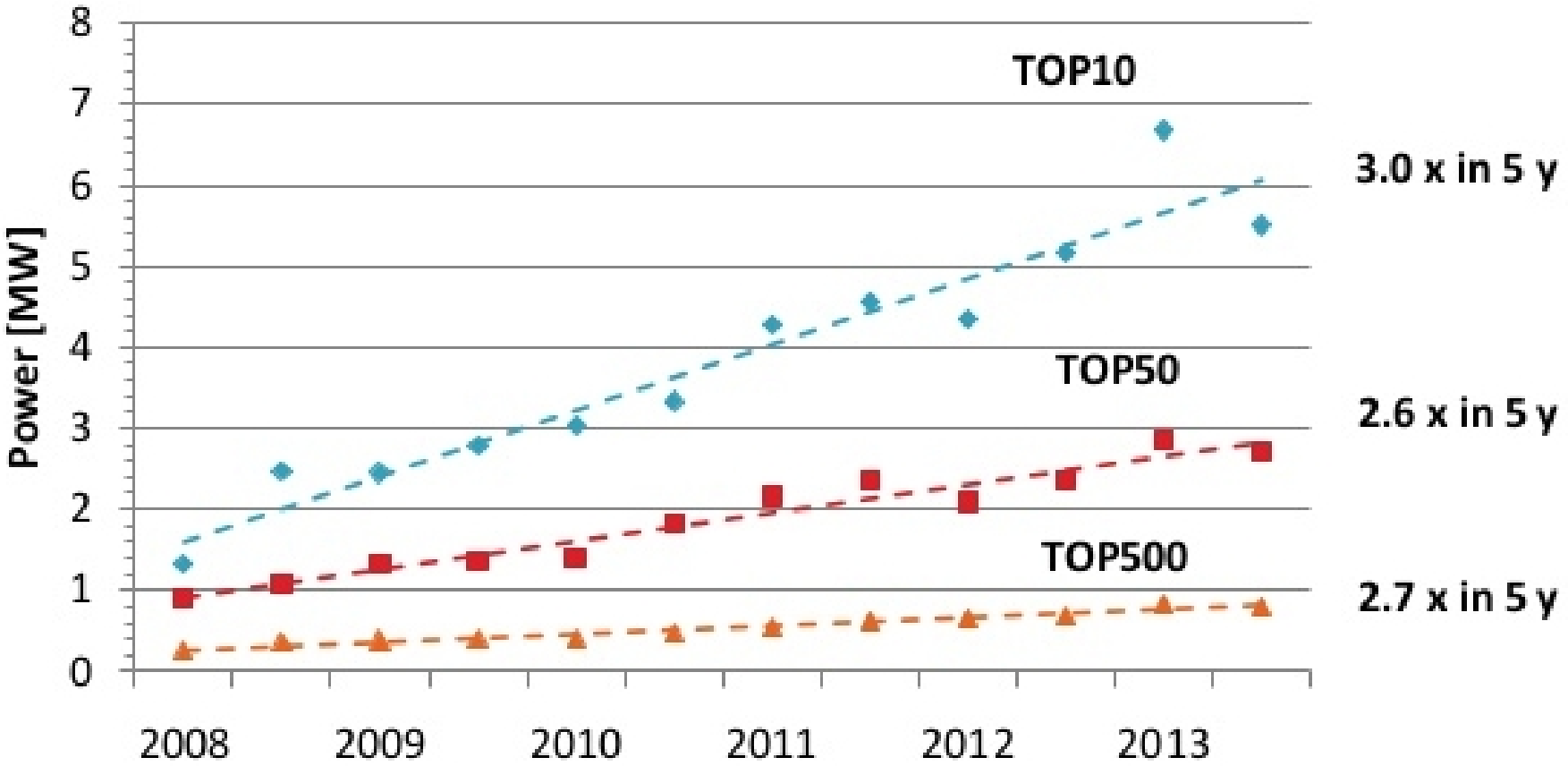
Figure 7: https://www.top500.org/statistics/perfdevel/

Top 10 Computers, Nov 2020

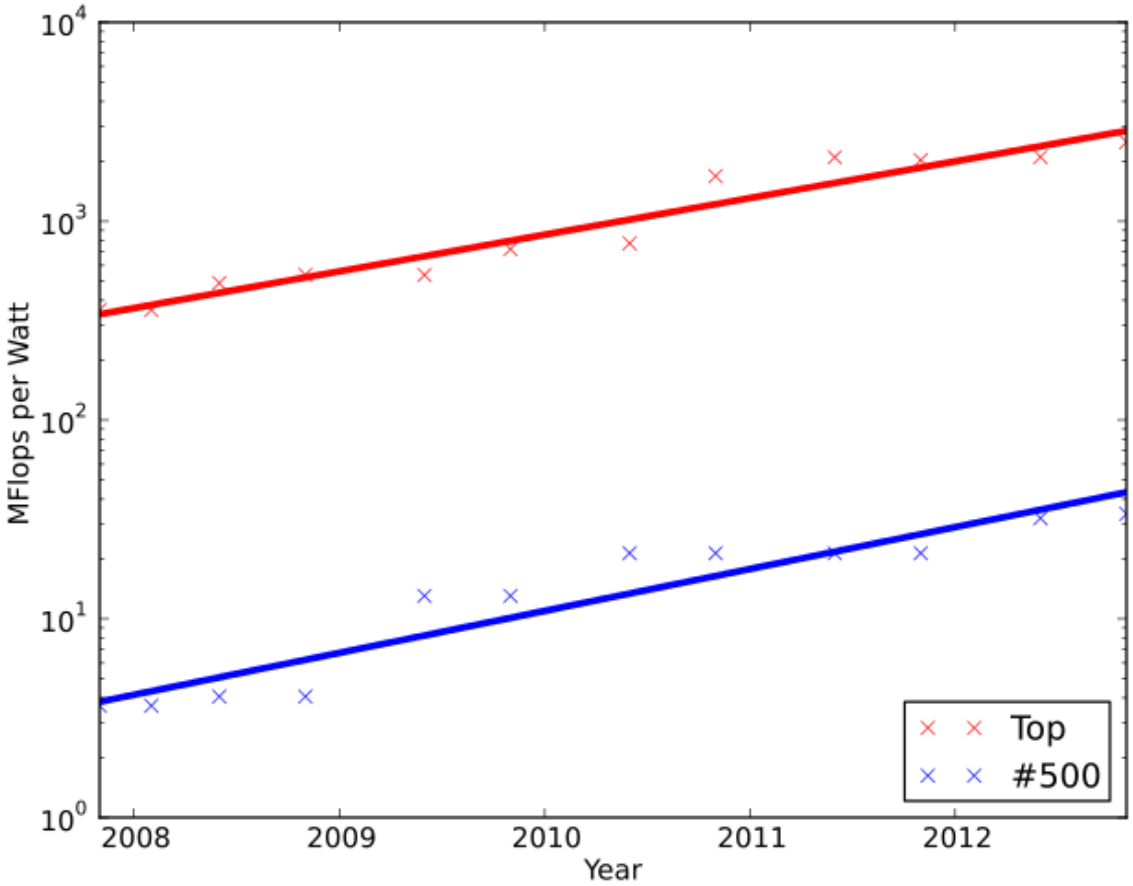
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646

Figure 8: `https://www.top500.org/lists/2020/11/`

Power Consumption, Nov 2013



Power Efficiency, Nov 2013



Power and Efficiency Trends

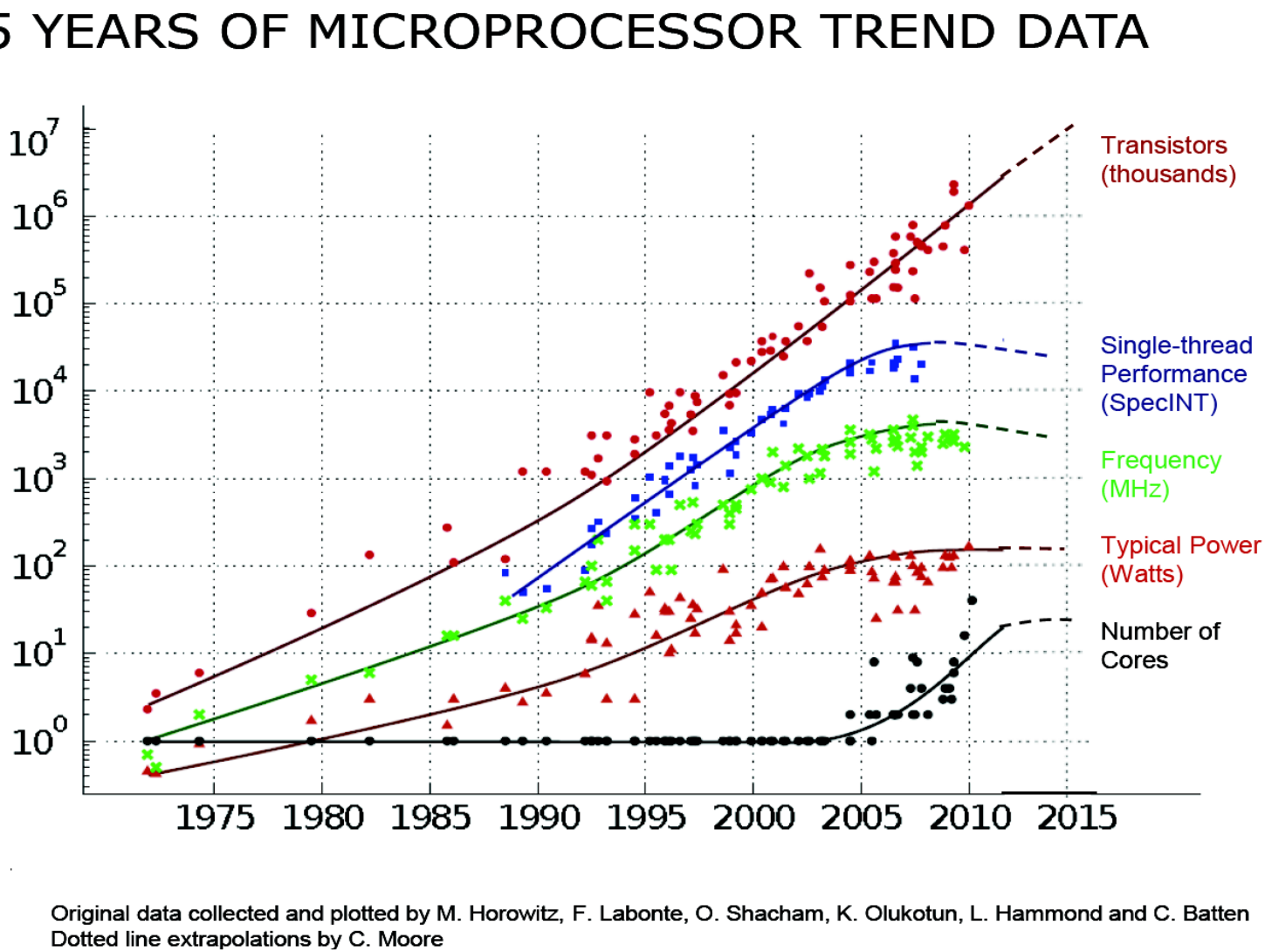


Figure 9: https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/

Where are we going?

Recap

- Humans have been working to use machines for computation for a very long time
- The 20th century saw the development of the computers we know today $\rightarrow$ development of computational science as a field
- A revolution in supercomputing began at the end of the 20th century $\rightarrow$ computational science is a major contributor to knowledge
- We are reaching the limits of "traditional" architecture growth
- What we can compute and how is tightly tied to computer architecture development