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

R. N. Slaybaugh

January 20, 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.

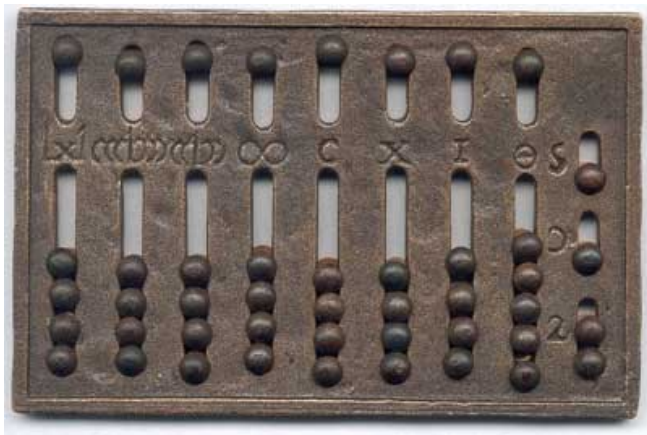[1]en.wikipedia.org

COMPUTING MACHINES, ORIGINS

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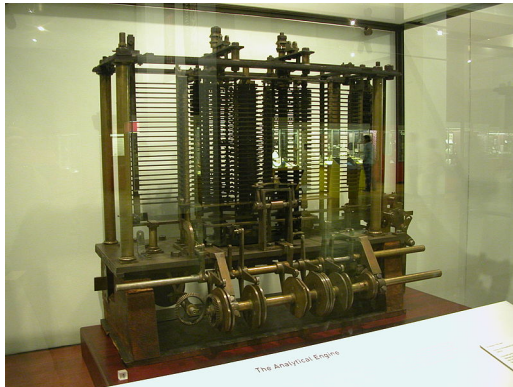
EARLY DEVELOPMENT OF COMPUTING

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
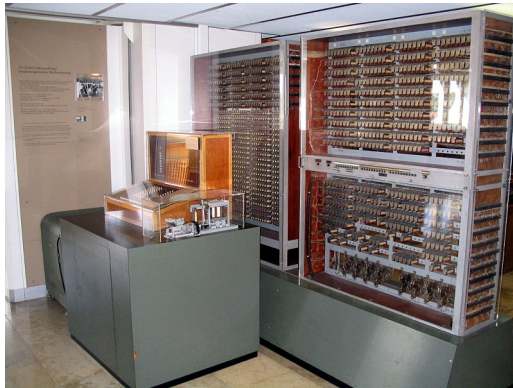
First Electromechanical Computers

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

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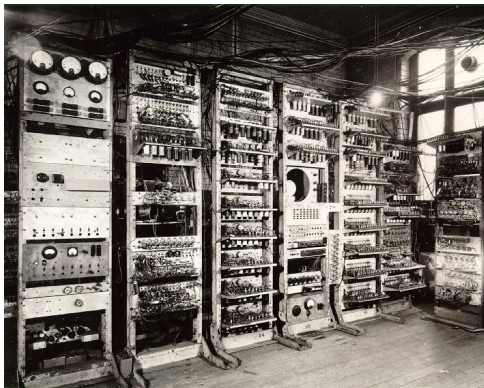
Figure 4 : The Manchester Mark 1 was one of the world's first stored-program computers,
http://www.computer50.org/mark1/ip-mm1.mark1.html

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 → "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
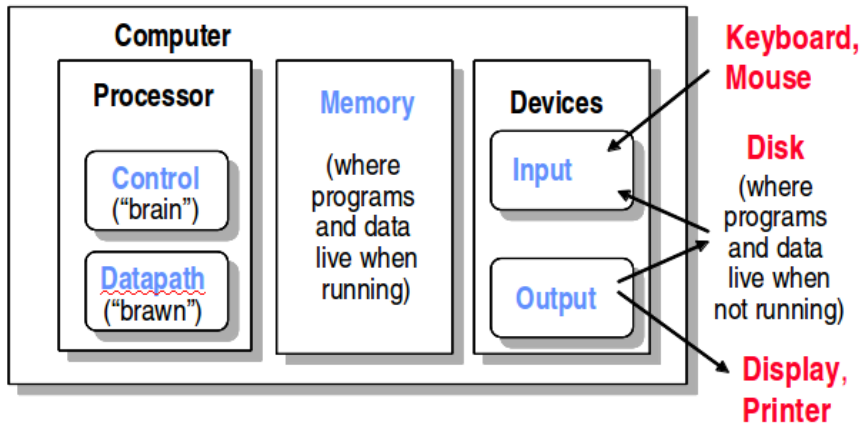
GENERATIONS 4-6

- 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
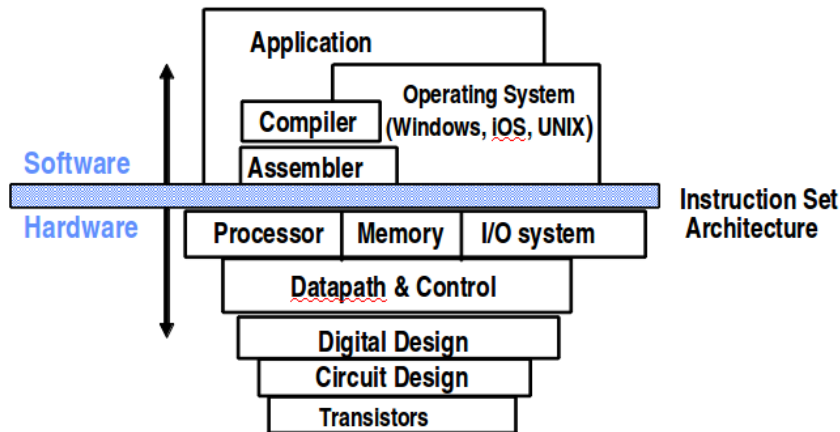
COMPONENTS

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
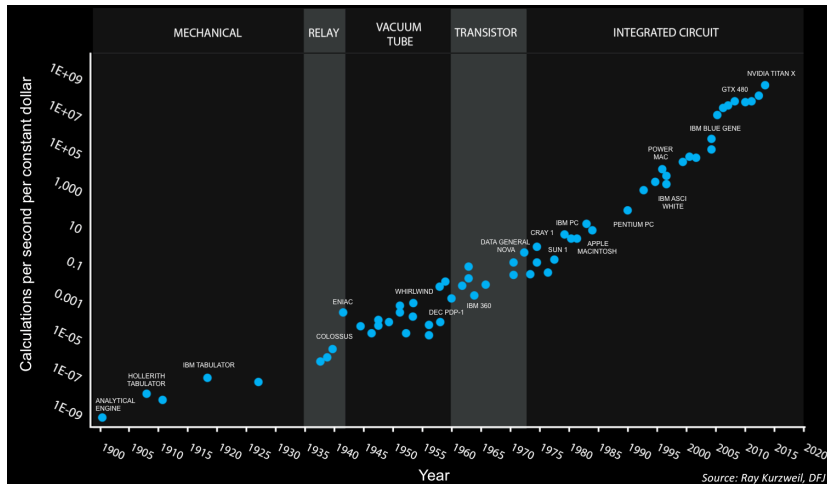 around now...

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)

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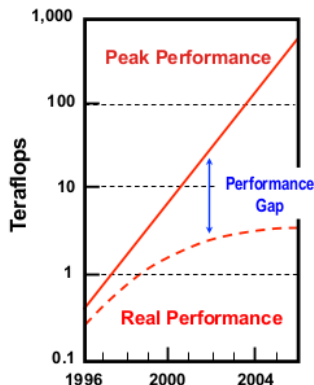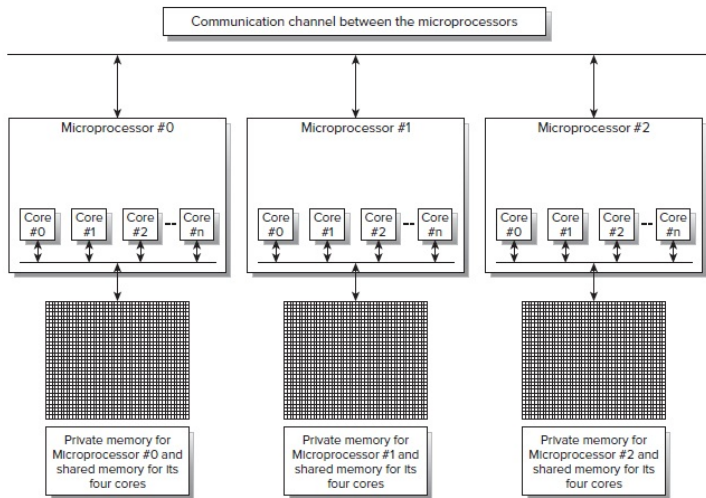
TYPES OF MEMORY

Shared   Distributed   Distributed Shared

GPUs

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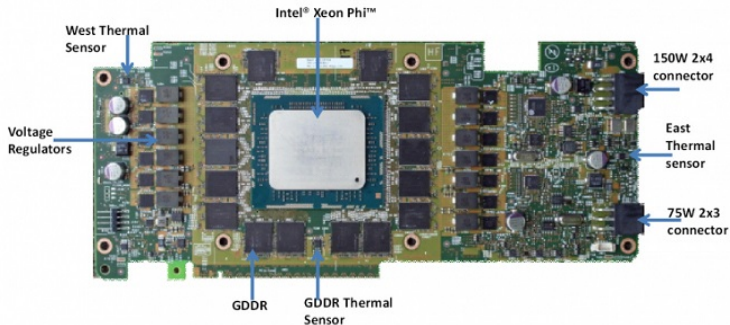- General Purpose GPU (**GPGPU**): Using a GPU to do CPU work - computational science instead of graphics
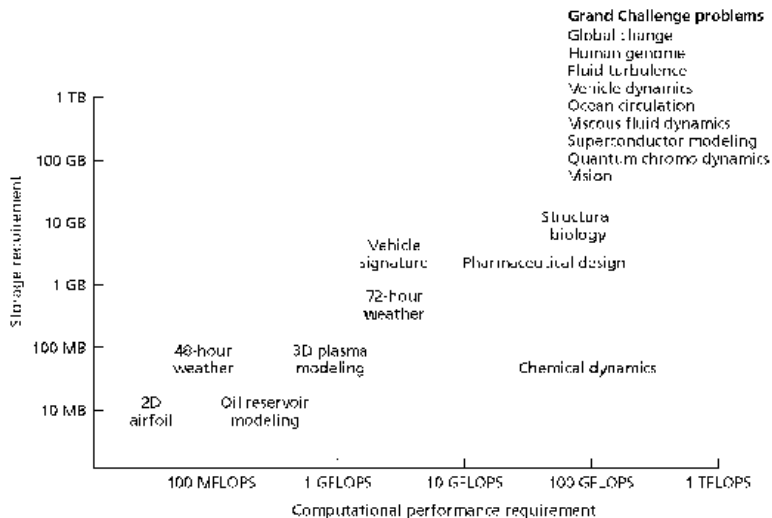
NVIDIA GPU

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

SCIENTIFIC COMPUTING DEMAND
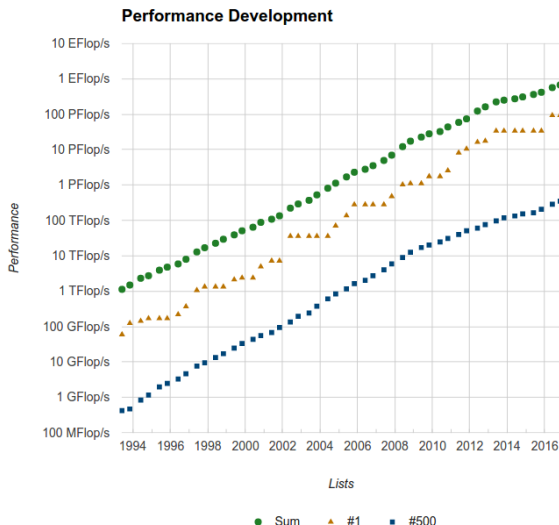
TOP 500 COMPUTERS, NOV 2016

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

TOP 10 COMPUTERS, NOV 2016

| Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
5	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939
6	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Fujitsu	556,104	13,554.6	24,913.5	2,719
7	RIKEN Advanced Institute for Computational Science (AICS) Japan	**K computer**, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
8	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	206,720	9,779.0	15,988.0	1,312
9	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
10	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	301,056	8,100.9	11,078.9	4,233

Figure 8: https://www.top500.org/lists/2016/11/
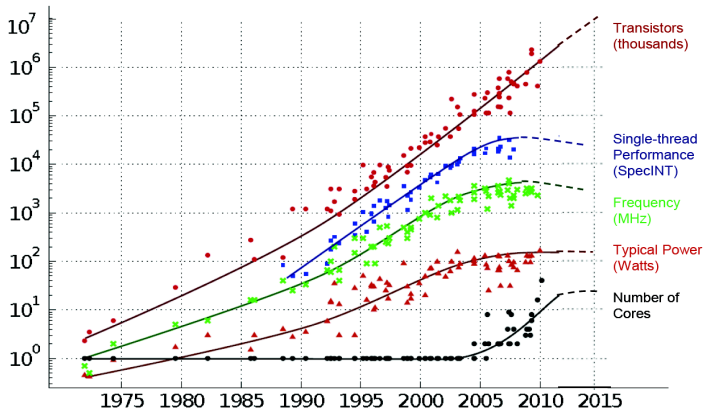
POWER CONSUMPTION, NOV 2013
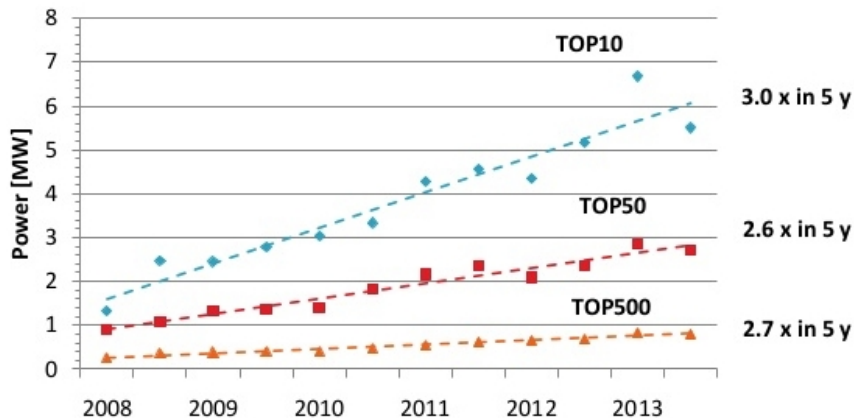
POWER EFFICIENCY, NOV 2013
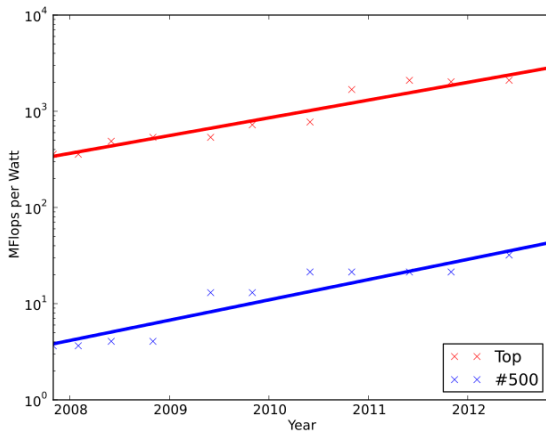
POWER AND EFFICIENCY TRENDS

35 YEARS OF MICROPROCESSOR TREND DATA

Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

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 → development of computational science as a field

- A revolution in supercomputing began at the end of the 20th century → 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