Introduction to High-Performance Computing

Giorgio Amati
Alessandro Ceci

Corso di dottorato in Ingegneria Aeronautica e Spaziale 2025

g.amati@cineca.it / g.amaticode@gmail.com
alessandro.ceci@uniroma1.it

Just for curiosity....

- ✓ Experience with HPC machine?
- ✓ Fortran, C, C++, anything else?
- ✓ Parallel paradigm: MPI, OpenMP, OpenACC, OpenMP offload, ….
- ✓ Linux, Windows, MacOS, (*NIX)
- ✓ Are you a Mathematician, a Physicist, an Engineer or a Computer Scientist?
- ✓ Do you know what is:
  - A Memory System?
  - A Cache?
  - A Floating Point Unit (FPU)?
  - A pipeline?
  - Moore Law?
  - Amdhal Law?

Agenda

- ✓ **HPC: What is it?**
- ✓ Hardware: how it works
  - Memory subsystem
  - Floating point units
- ✓ Algorithm vs. Implementation
- ✓ Compiler
- ✓ Parallel Paradigm
  - CPUs
  - GPUs
- ✓ Conclusions & Comments

Questions!

#1: Why all this complexity?
#2: Is the market big enough to survive for a HPC firm?
#3: Which skill is the more important?
#4: What is the performance range?
#5: Why GPUs?
#6: And for the next 20 years?

HPC: what it is?

From wikipedia:

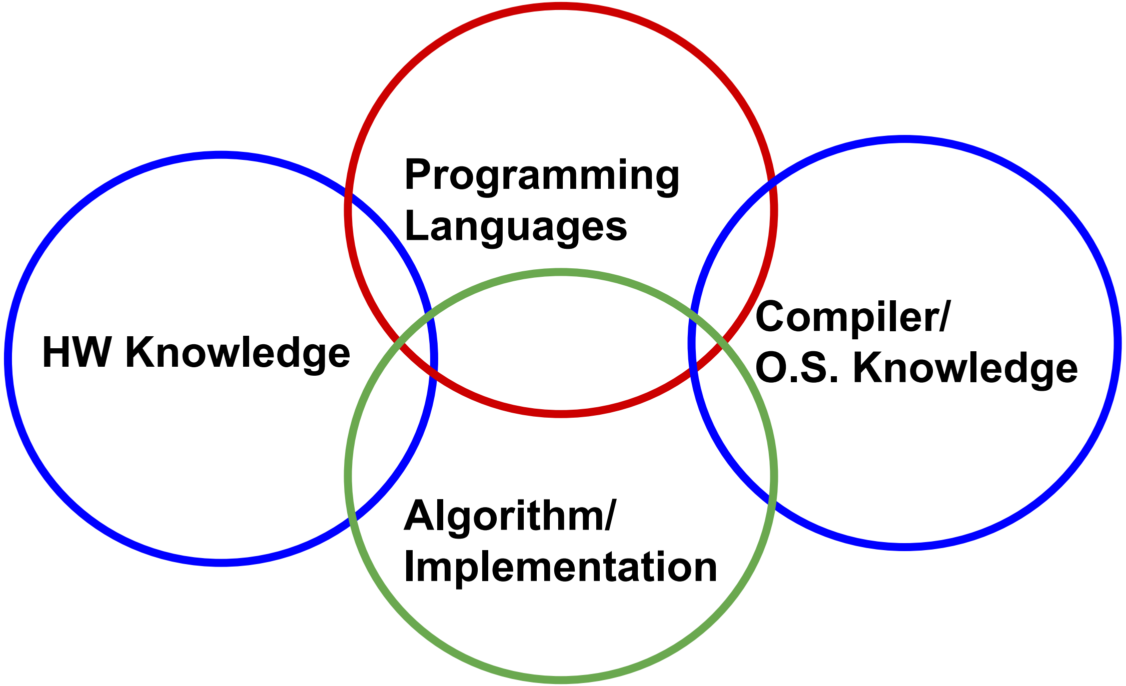
✓ “High-performance computing (HPC) uses supercomputers and computer clusters to solve advanced computation problems”

Personal definition:

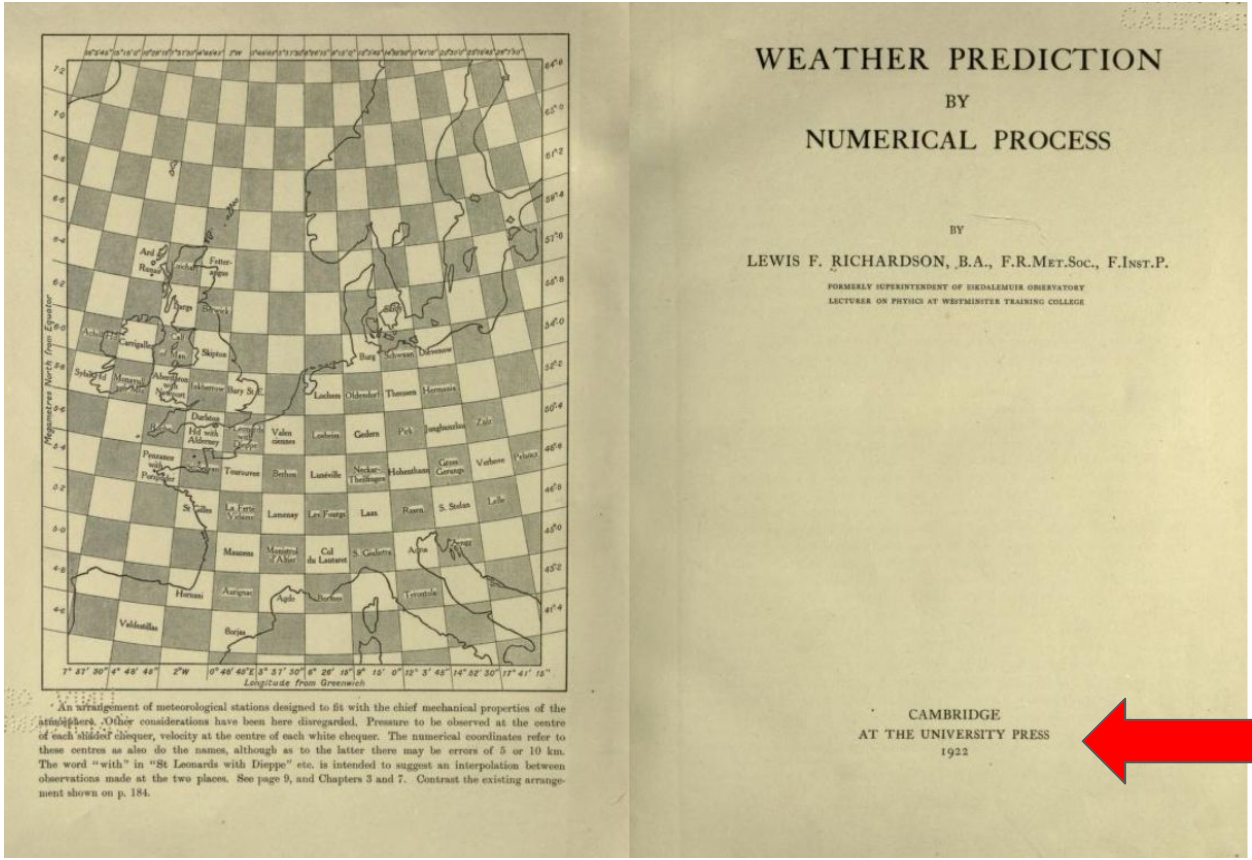
✓ It is the overlap of different skills, all devoted to exploit HW performance as much as possible (both serial and/or parallel, but not limited to supercomputers…)

HPC: what it is?

✓ These are the main skills for efficient HPC



HPC: older than computers?



An arrangement of meteorological stations designed to fit with the chief mechanical properties of the atmosphere. Other considerations have been here disregarded. Pressure to be observed at the centre of each shaded chequer, velocity at the centre of each white chequer. The numerical coordinates refer to these centres as also do the names, although as to the latter there may be errors of 5 or 10 km. The word "with" in "St Leonards with Dieppe" etc. is intended to suggest an interpolation between observations made at the two places. See page 9, and Chapters 3 and 7. Contrast the existing arrangement shown on p. 184.

WEATHER PREDICTION

BY

NUMERICAL PROCESS

BY

LEWIS F. RICHARDSON, B.A., F.R.MET.SOC., F.INST.P.

FORMERLY SUPERINTENDENT OF ESKDALEMUIR OBSERVATORY

LECTURER ON PHYSICS AT WESTMINSTER TRAINING COLLEGE

CAMBRIDGE

AT THE UNIVERSITY PRESS

1922

HPC older than computers

Meteorologist Lewis Fry Richardson, creator of the first dynamic model for weather prediction, proposes the creation of a “forecast factory” that would employ some 64,000 human computers sitting in tiers around the circumference of a giant globe. Each calculator would be responsible for solving differential equations related to the weather in his quadrant of the earth. From a pedestal in the center of the factory, a conductor would orchestrate this symphony of equations by shining a beam of light on areas of the globe where calculation was moving too fast or falling behind.

https://www.historyofinformation.com/detail.php?id=59

Example: matrix-matrix multiplication

Simple problem: for 2 n^2 matrices we have to:

- ✓ compute n^3 products and n^3 sums
- ✓ load 2*n^2 data and to store n^2 data
  - Ratio computation vs. load/store is O(n)!

```
do j = 1, n
   do k = 1, n
      do i = 1, n
         c(i,j) = c(i,j) + a(i,k)*b(k,j)
      enddo
   enddo
enddo
```

- ✓ **MM multiplication is used for supercomputing rankings (top500)**

Example: matrix-matrix multiplication

- Performance depends on many aspects:
  - Coding --> 1 vs.2 , 3 vs. 4
  - HW knowledge --> 1 vs. 2, 3 vs. 4
  - HW used --> 2 vs. 3, 4 vs 5
  - (Optimized) Libraries --> 5

#test	Size	HW	MFlops	Ratio
1-Cache unfriendly	2048	CPU	201	-
2-Cache friendly	2048	CPU	4870	24x
3-OpenACC	8192	GPU-V100	361328	1797x
4-OpenACC+unrolling	8192	GPU-V100	448923	2233x
5-Matmul	16384	GPU-A100	6721790	33441x

Few "facts" about HPC

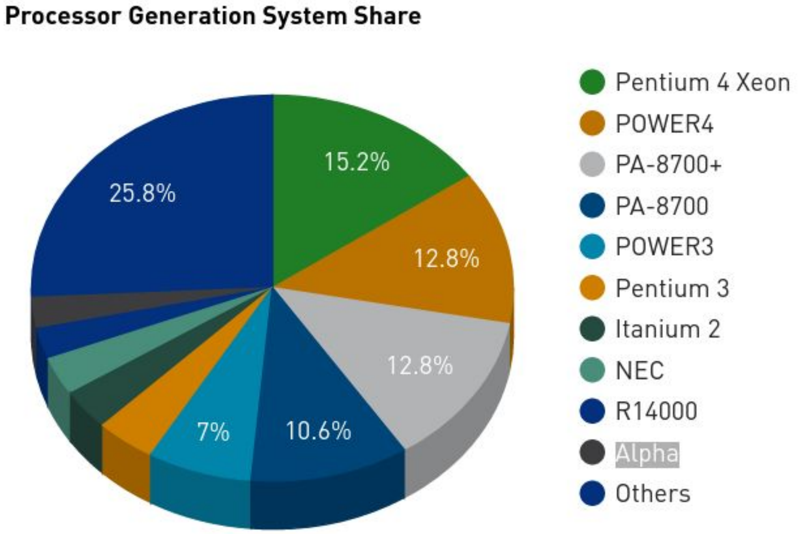
HPC market is not big enough to survive…

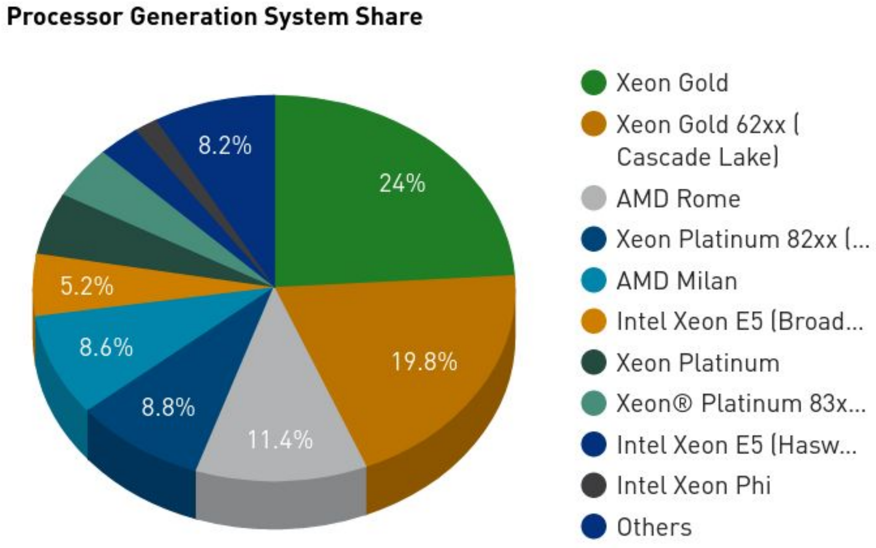
- ✓ SGI
- ✓ Compaq
- ✓ Digital
- ✓ SUN
- ✓ SiCortex
- ✓ MTA
- ✓ CRAY
- ✓ CONVEX
- ✓ CDC
- ✓ Thinking Machine
- ✓ Quadrics/APE

- ✓ IBM
  - Power3/4/.../9
- ✓ Intel
  - Itanium
  - Phi
- ✓ HP
- ✓ NVIDIA
- ✓ FUJITSU
- ✓ AMD
  - Opteron
- ✓ NEC
  - SX6

Few “facts” about HPC

Top500 list: June 2003 vs November 2022





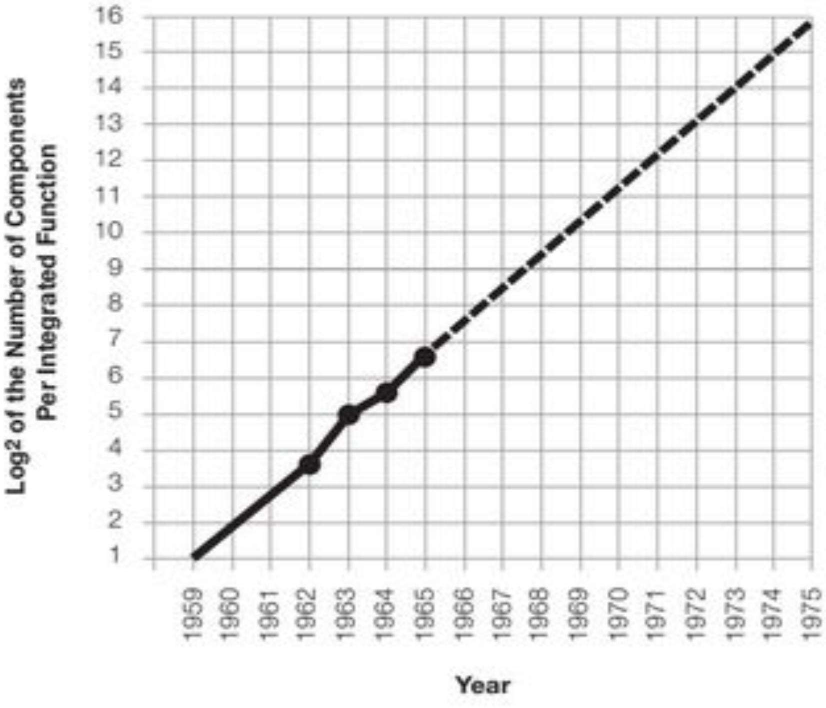
Moore's Law

In his 1965 article, Moore (Intel co-founder) planned the increase of the # of transistors up to 1975.

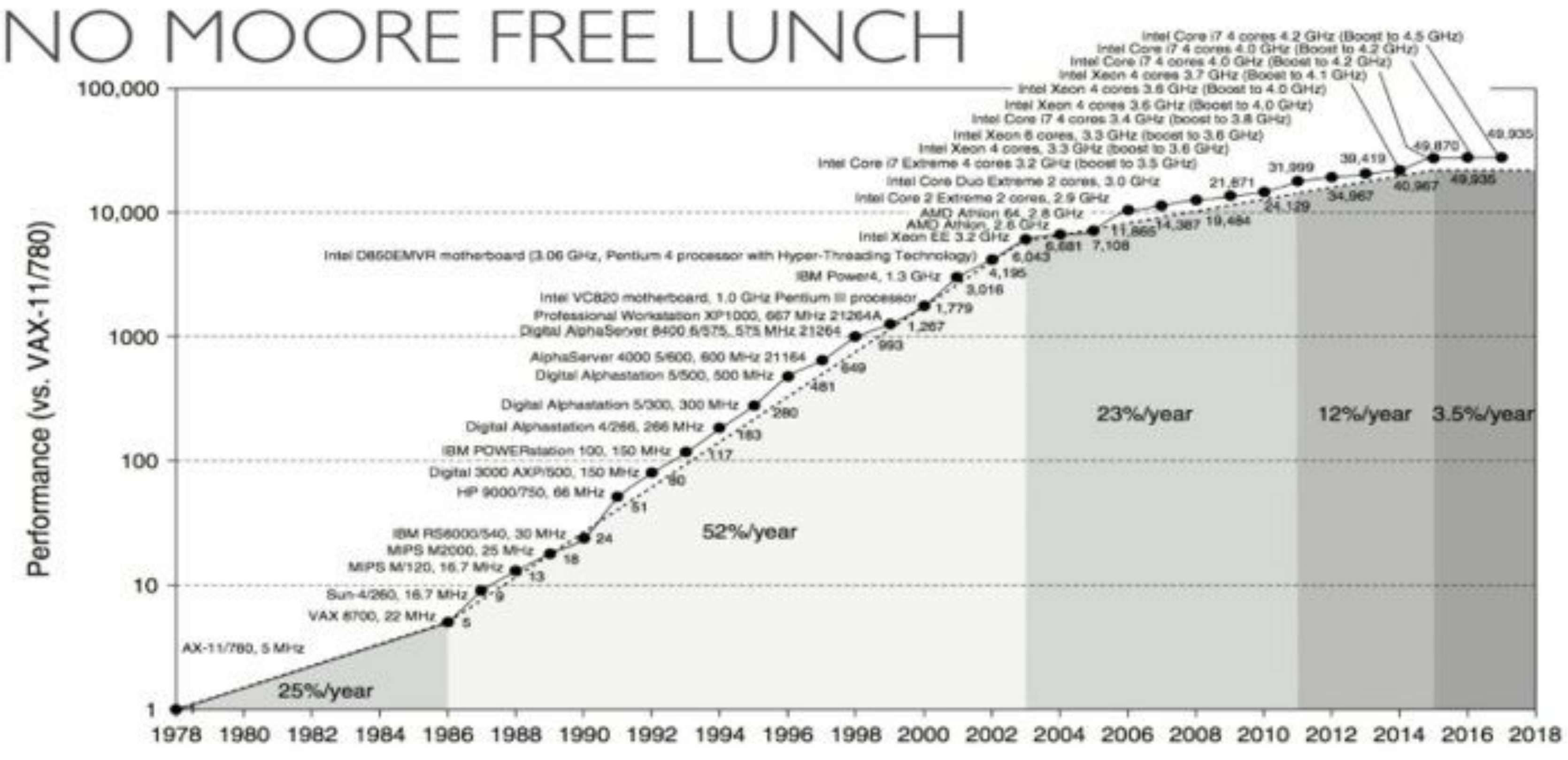
"With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip"

He stated a 2x increment of transistors every 18 Month.

- ✓ This law is still valid now (in some form): to "survive" in the market HW firms must follow this law
- ✓ Now "transistor shrinking" is much harder: we are near to quantum effects
- ✓ New "ideas" must be found

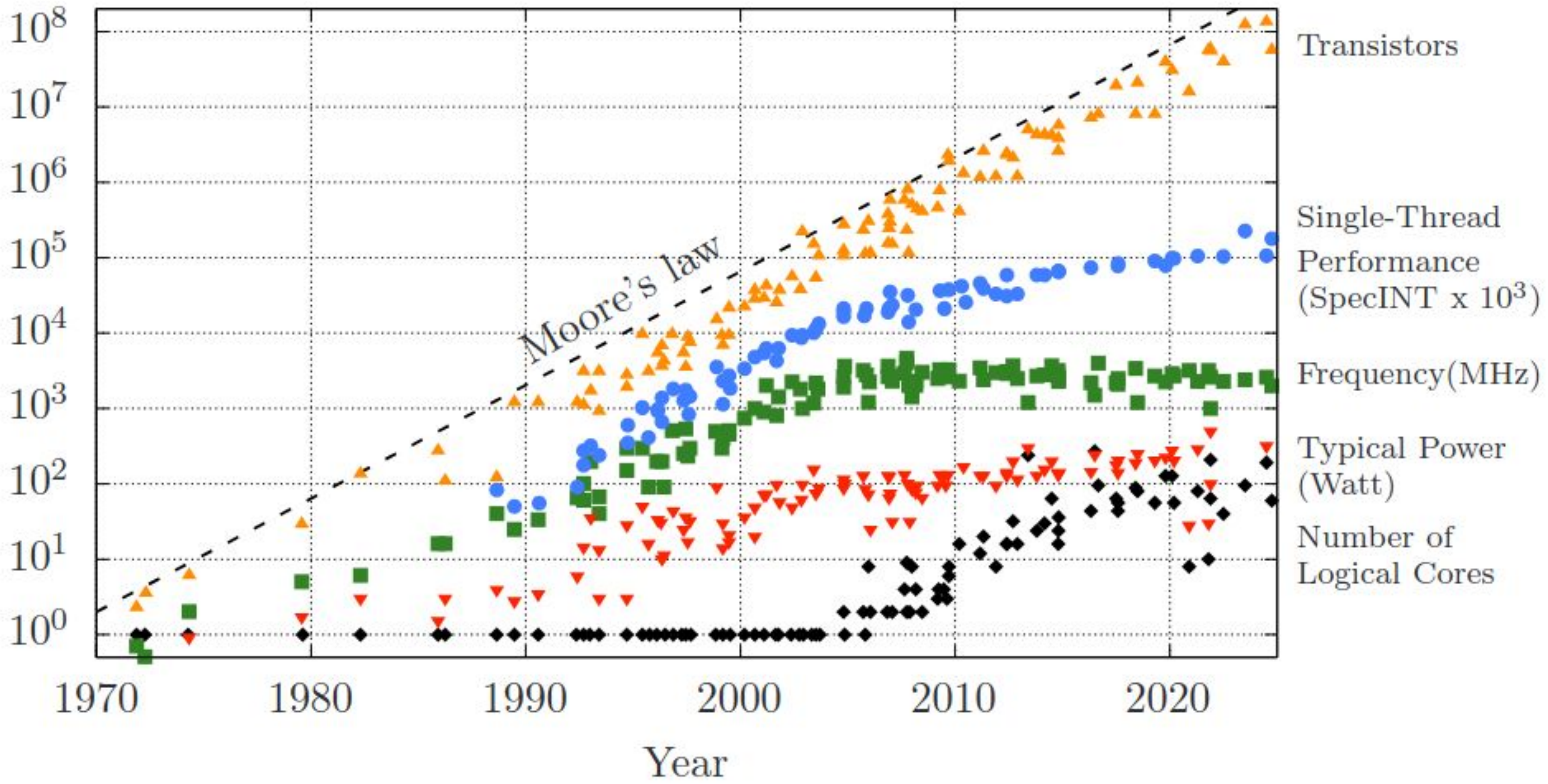


Moore's Law



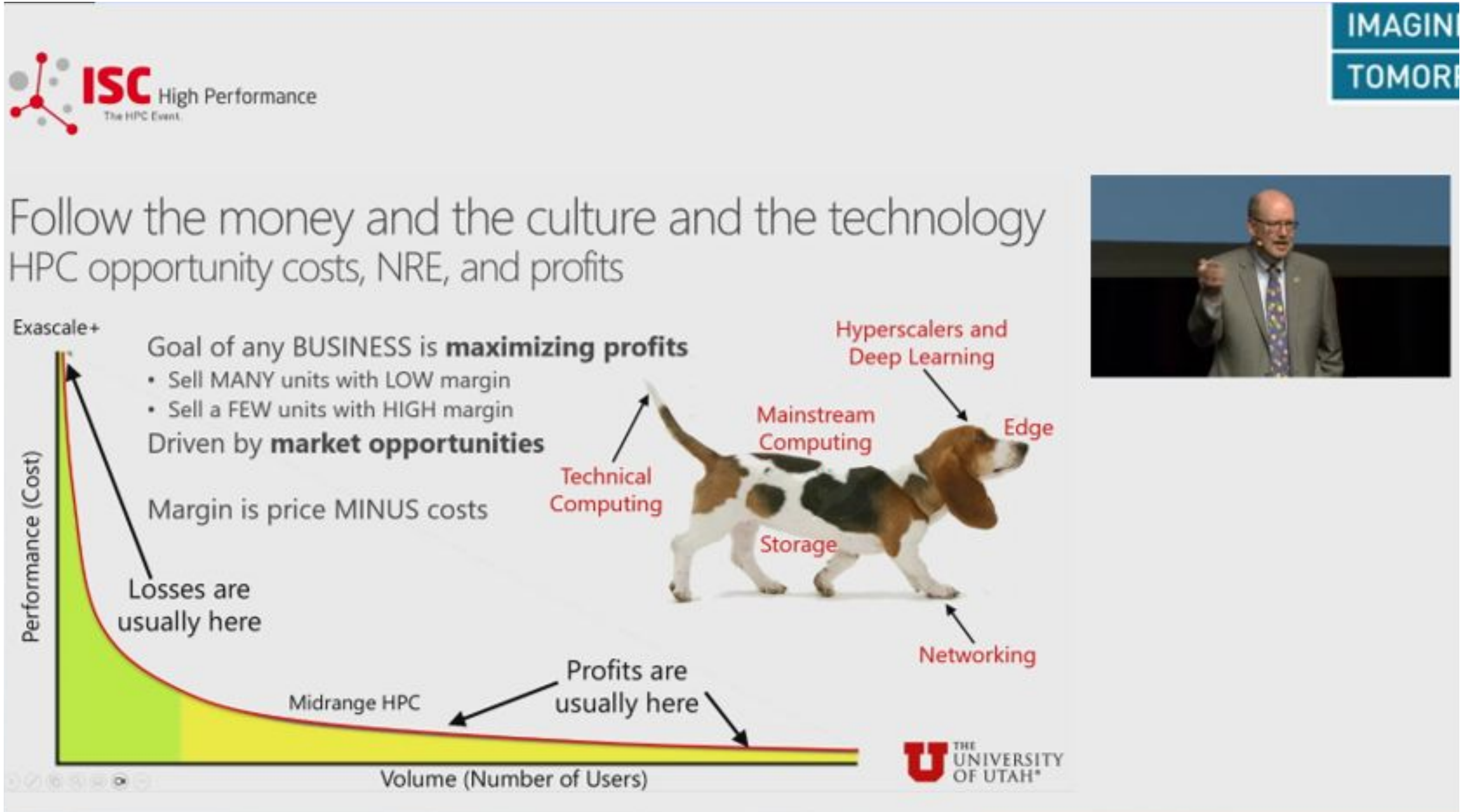
Hennessy, John L., and David A. Patterson. *Computer architecture: a quantitative approach*. 6th ed, 2017

Moore's Law

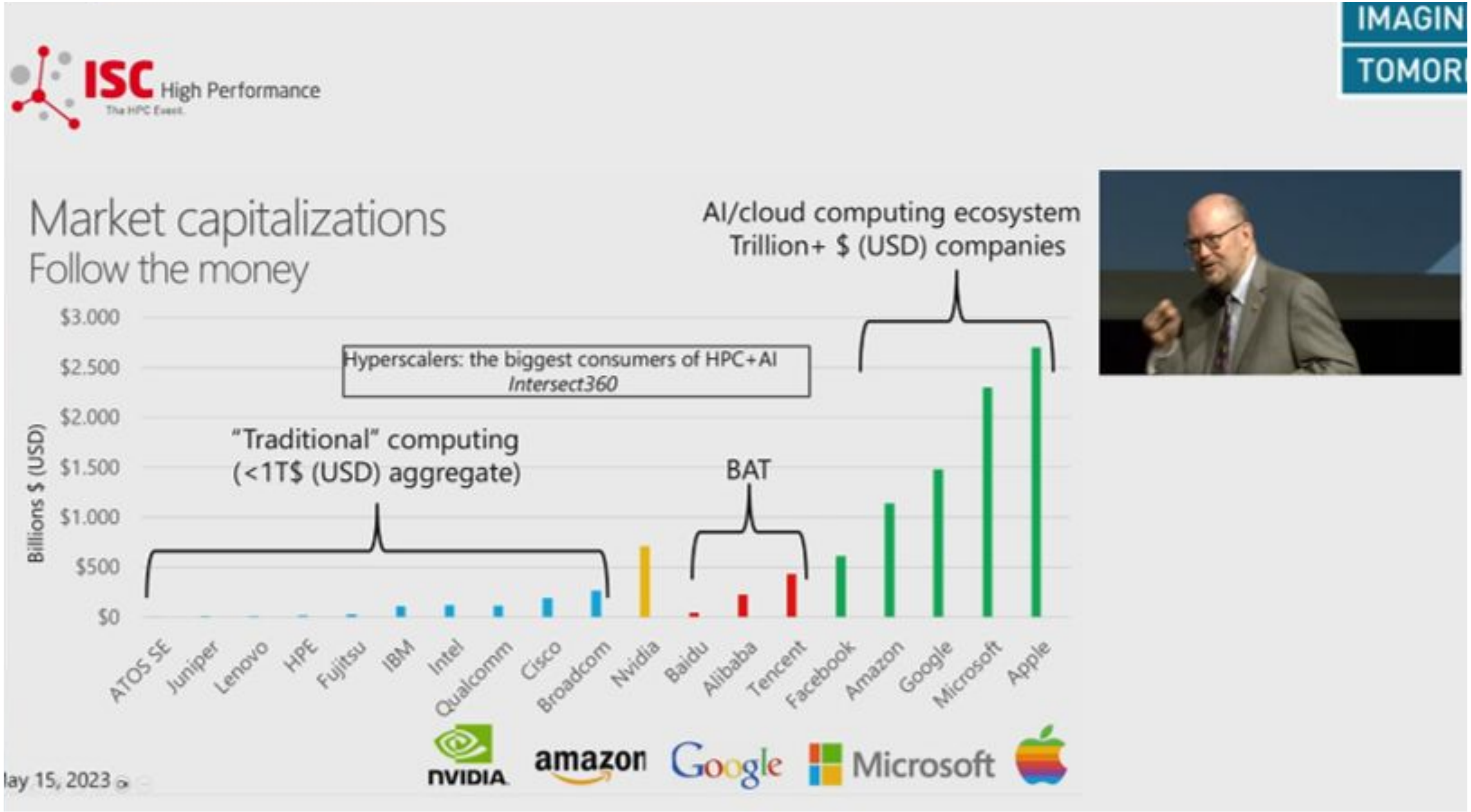


10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
Moore's law
1970
1980
1990
2000
2010
2020
Year
Transistors
Single-Thread Performance (SpecINT x 10^3)
Frequency(MHz)
Typical Power (Watt)
Number of Logical Cores

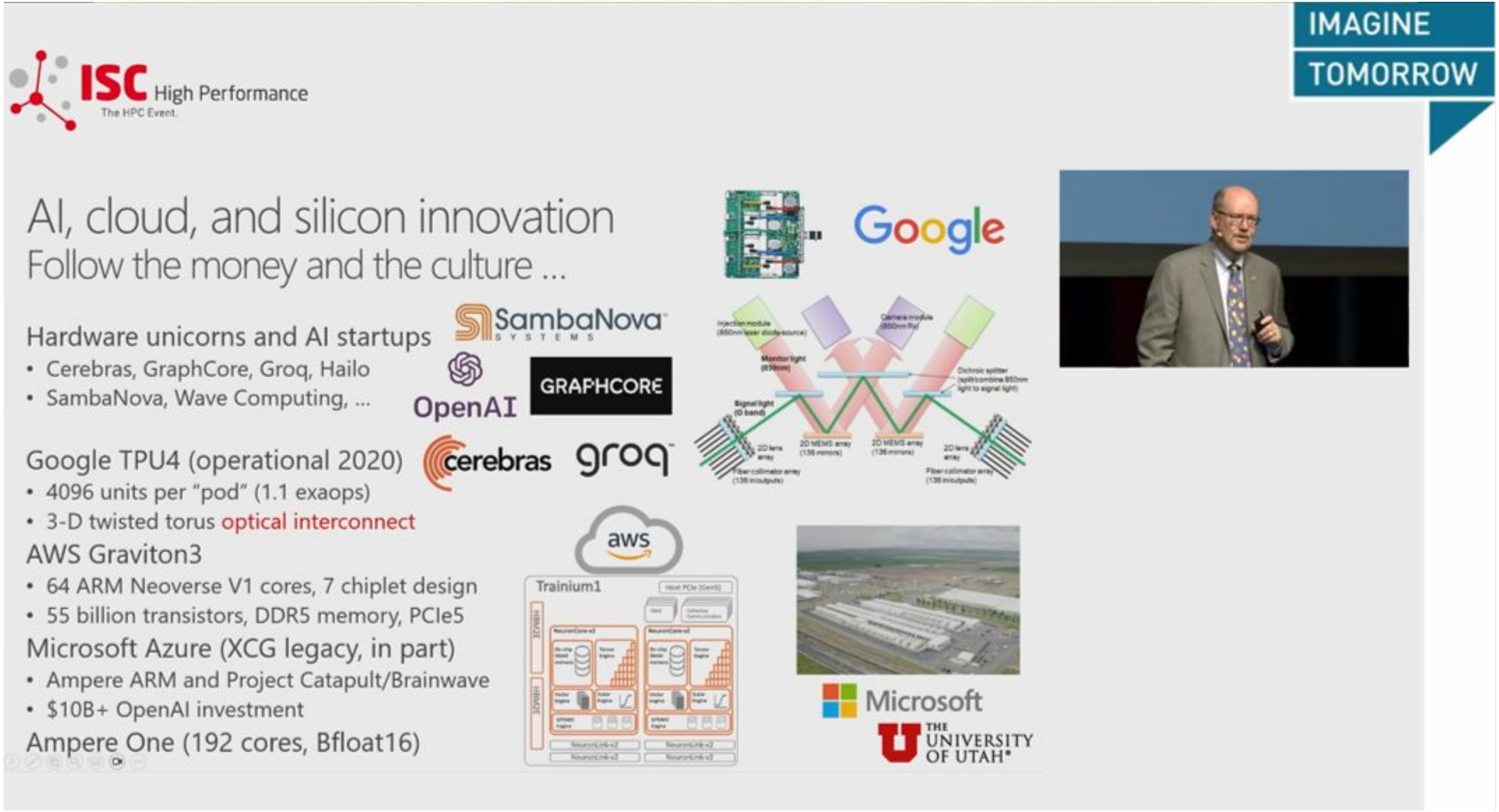
Follow the money! (from D.Reed@ISC2023)



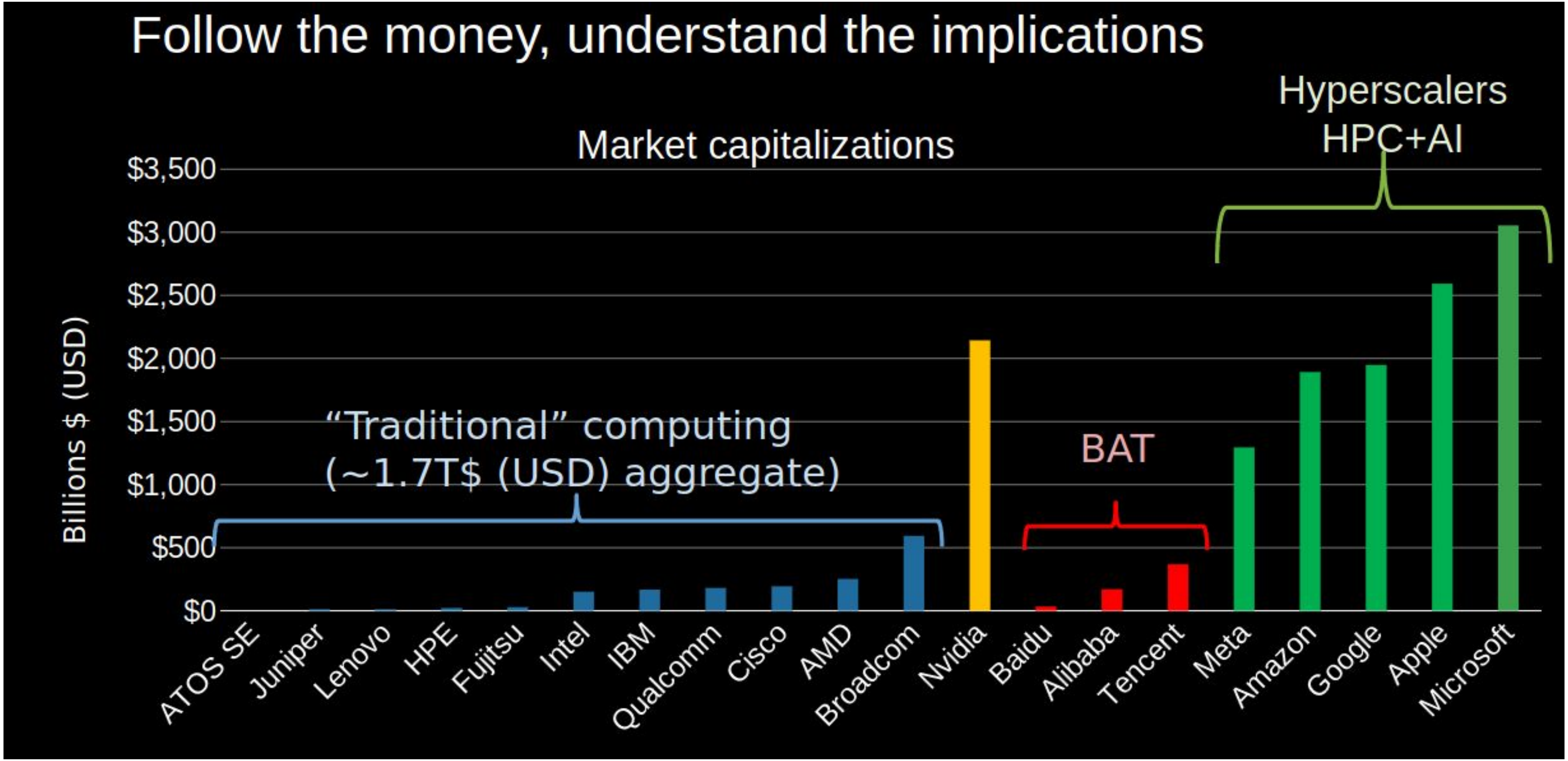
Follow the money! (from D.Reed@ISC2023)



Follow the money! (from D.Reed@ISC2023)



Follow the money!/2 (from Yelick@ISC2024)

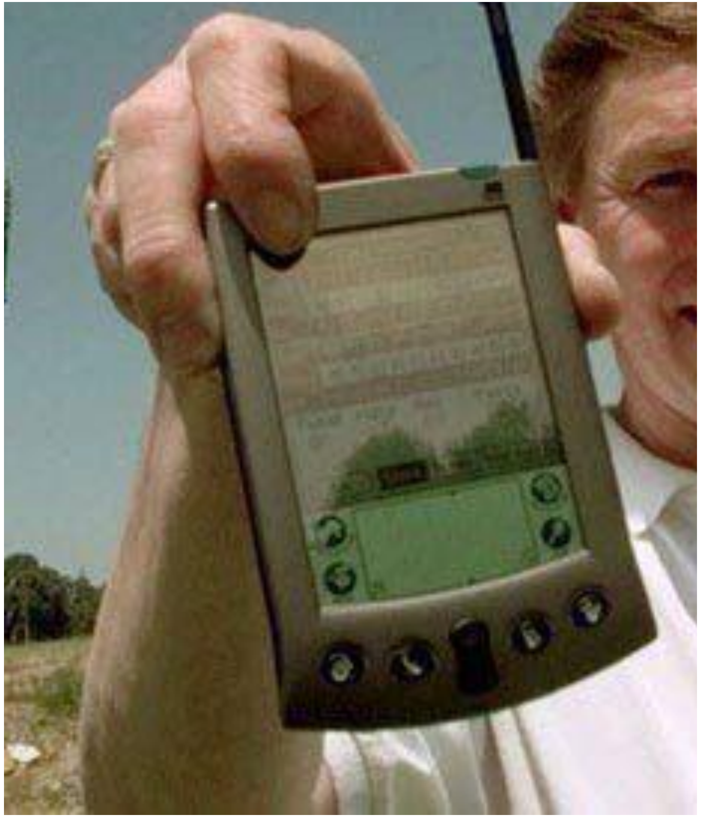


For Boomer only: Palm pilot (PDA)

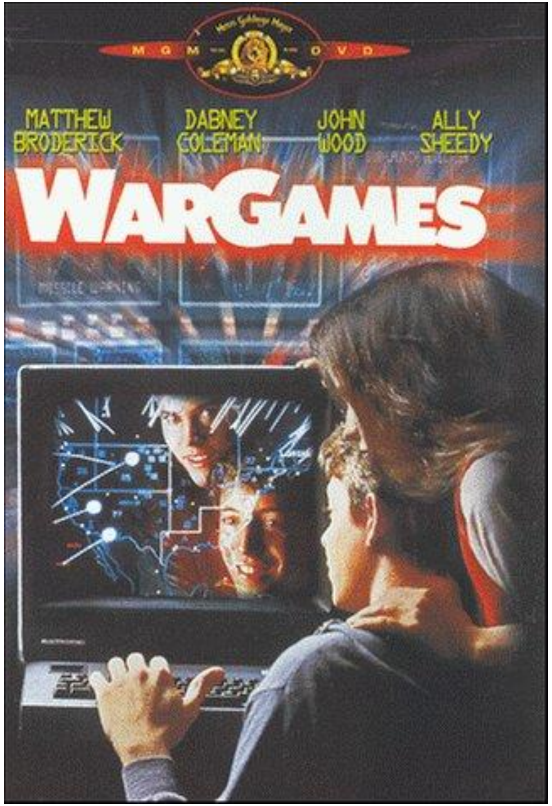
Do we remember there was a company called Palm, which used to be the pioneer in Personal Digital Assistant (PDA) industry, they reached the peak of success in 2000 with **a market cap exceeding Apple, Amazon, Google, and Nvidia combined**. 🤯

However, their attachment with physical buttons and slowness to adapt to touchscreens led to their downfall. HP acquired palm in 2009 and tried to revive the industry but failed and was discontinued in 2010.

Moral: Today's tech titan could be tomorrow's trivia answer. In this brutal tech world, any tech startup is as good as their last update.



HPC@movie


MGM DVD
MATTHEW BRODERICK
DABNEY COLEMAN
JOHN WOOD
ALLY SHEEDY
WARGAMES


fineart america

Old stuff….


Altix 4700
sgi
2023/1/20 09:13

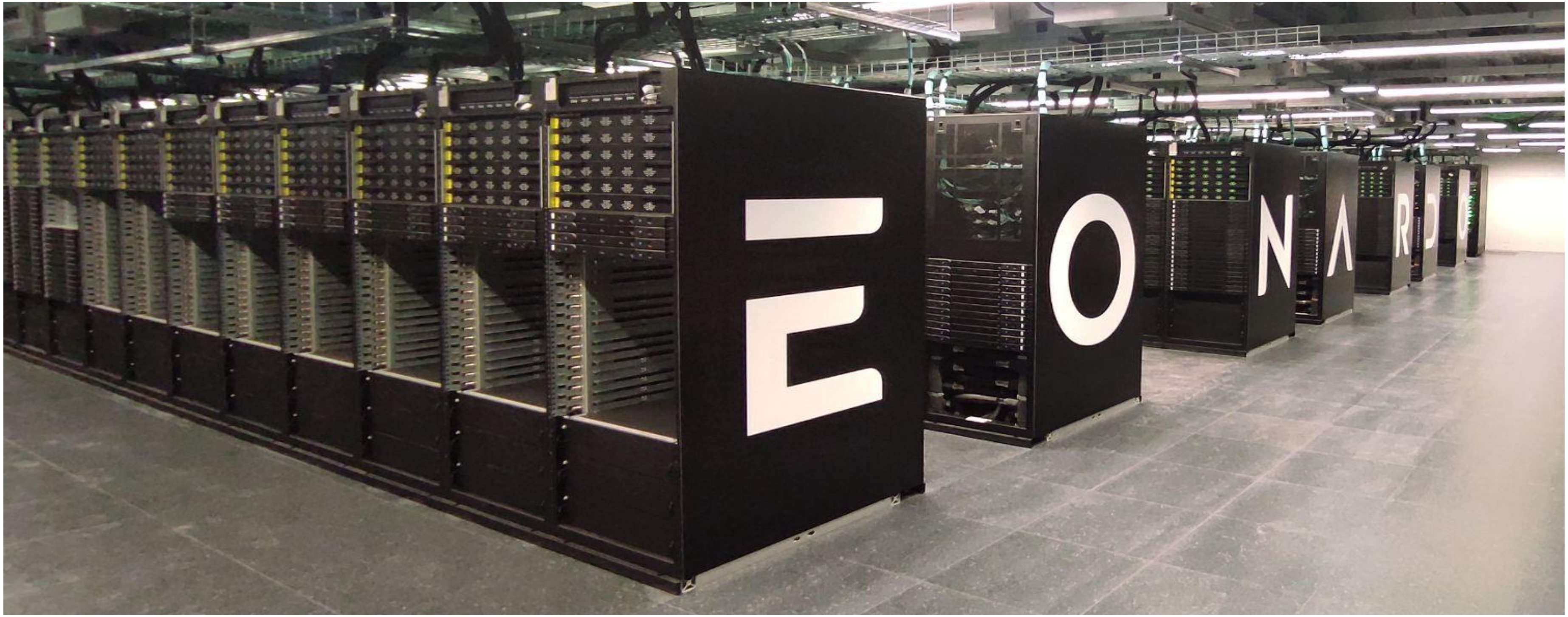

ORIGIN 2000
SiliconGraphics
ORIGIN 2000
2023/1/20 09:13


CONVEX COMPUTER CORPORATION
CONVEX


COMPAQ ALPHASERVER GS140

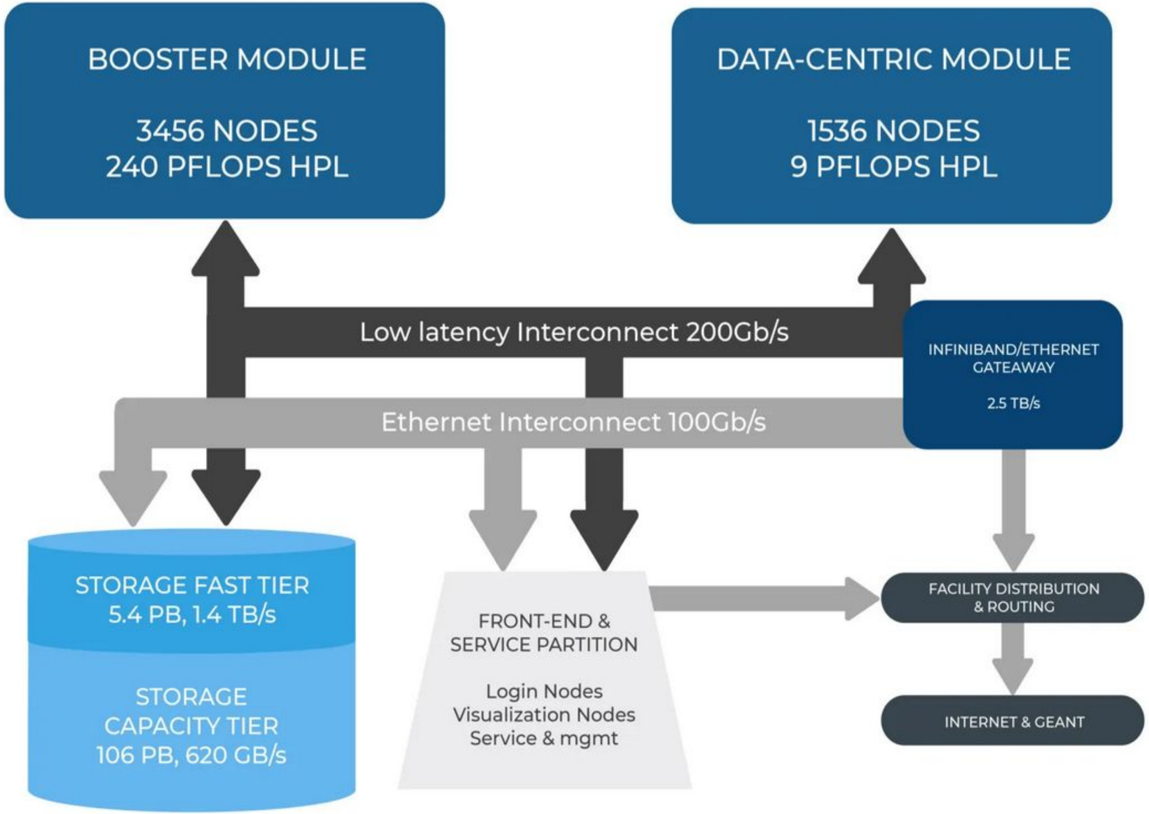

COMPAQ
2023/1/20 09:13

Leonardo



Leonardo: main figures

- ✓ 1536 CPU-based nodes
  - 172032 cores
- ✓ 3456 GPU-based nodes
  - 13824 GPU
  - 110592 cores
- ✓ 155 Racks
  - 16 CPU racks
  - 116 GPU racks
  - 12 I/O racks
  - 1 System racks
  - **About 300’000 Kg!**
- ✓ Power Requirements
  - HPL: ~ 8.0 MW
  - Operational: ~ 6.0 MW



Heterogeneous Cluster

Real word systems are now “heterogeneous”.

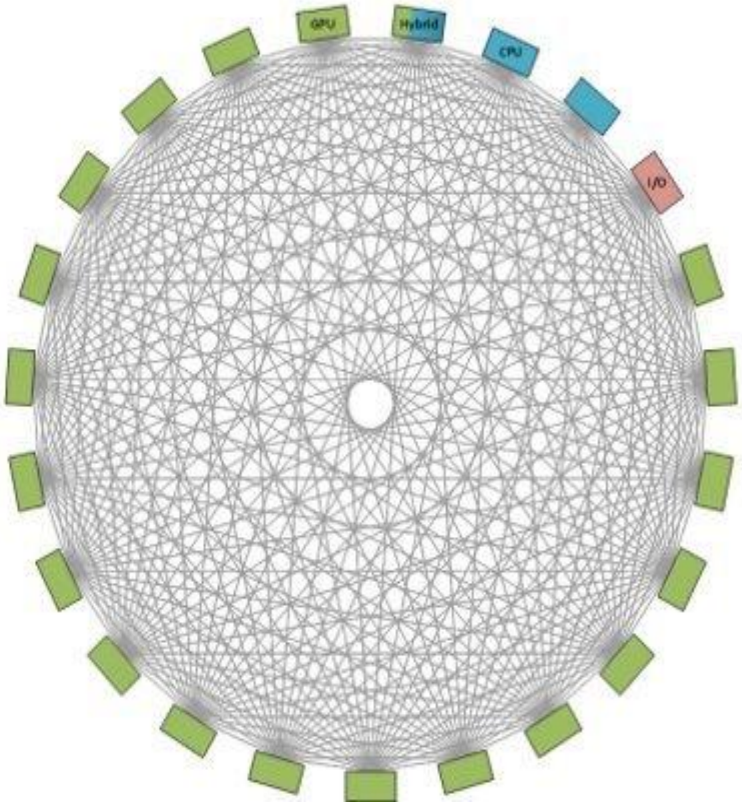
✓ Three different level of parallelization to manage
  - A cluster (i.e. distributed memory system) of nodes (up to 1000 or more)
  - Each of them is a shared memory (non uniform) system of cores (up to 128 or more)
  - with GPUs (up to 8 GPUs) connected via PCI Bus

✓ For example, Leonardo Booster has:
  - 3456 Nodes, interconnect via infiniband network at 200Gb/s
  - Each node has 1 CPU with 32 Core with 512GB RAM
  - Each node has 4 GPU, each with 64GB HBM

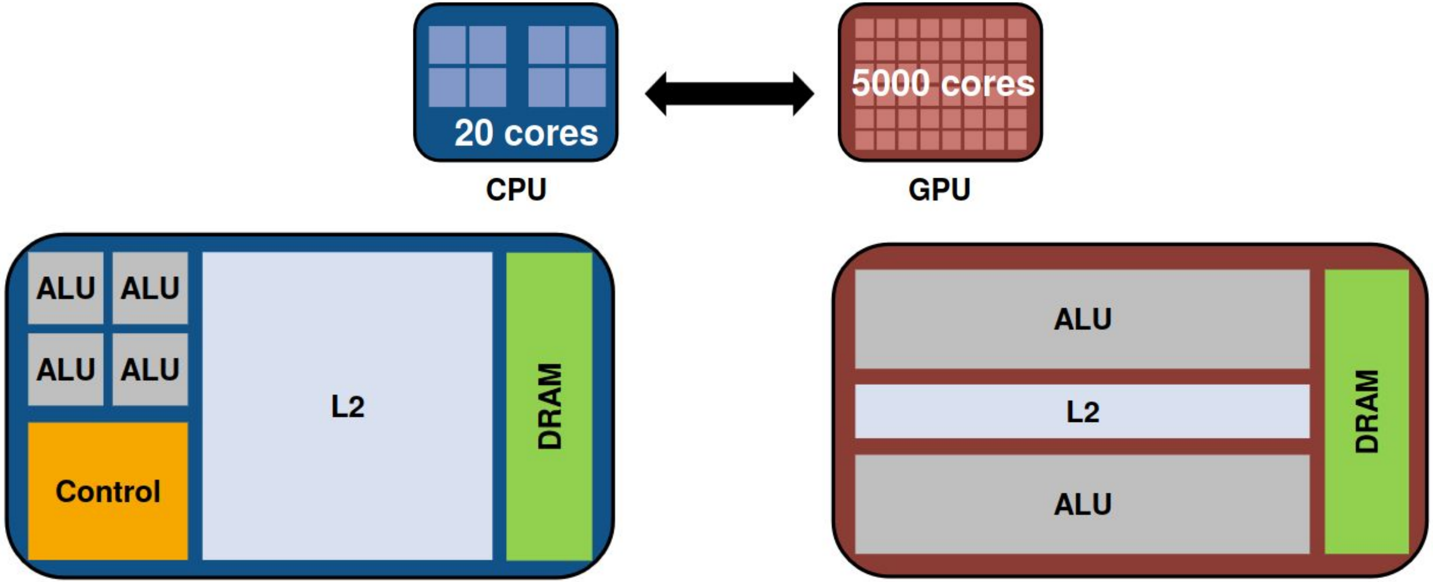
Leonardo Network (Dragonfly+)

Each Cell consists by 6 Racks

- ✓ 19 GPU-Cells 6 rack each
- ✓ 2 CPU-Cells
- ✓ 1 Hybrid-Cell (CPU/GPU)
- ✓ 1 I/O Cell



CPU vs. GPU



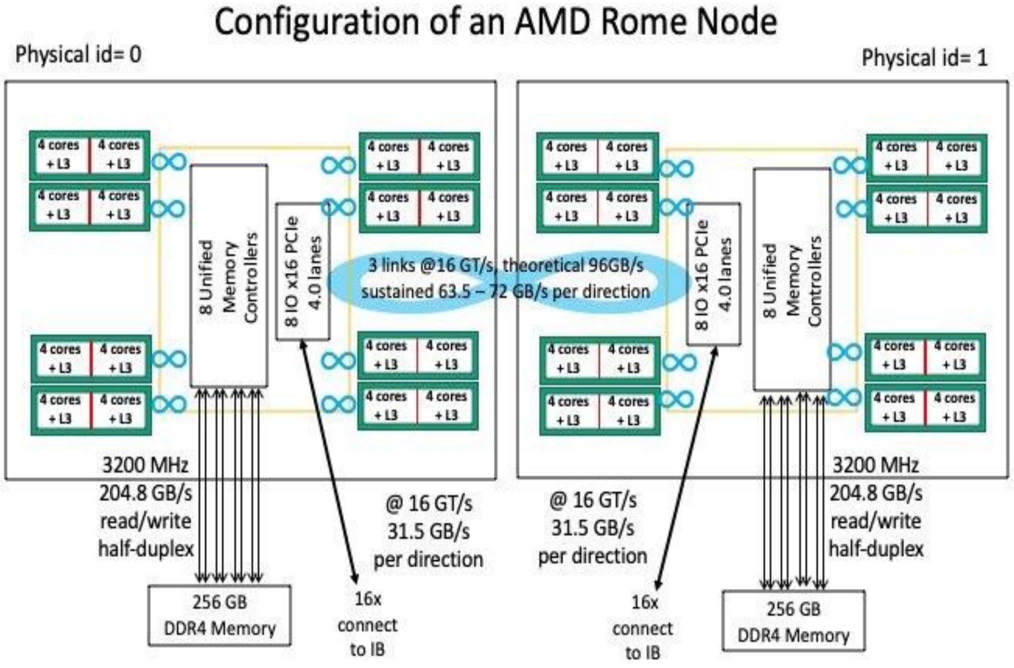
- ✓ Optimized for low latencies
- ✓ Huge caches
- ✓ Control logic for out-of-order and speculative execution
- ✓ **Targets on general-purpose applications**

- ✓ Optimized for data parallel throughput
- ✓ Memory latency tolerant
- ✓ More transistors dedicated to computations
- ✓ **Targets on special applications**

CPU structure

https://www.nas.nasa.gov/hecc/support/kb/amd-rome-processors_658.html

- ✓ Order of 40 Billion Transistors
- ✓ 64 Cores: each core
  - ✓ 2 FMA units at 256-bit
  - ✓ 4 x86 instruction per cycle
  - ✓ 4 flops per cycle



GPU structure

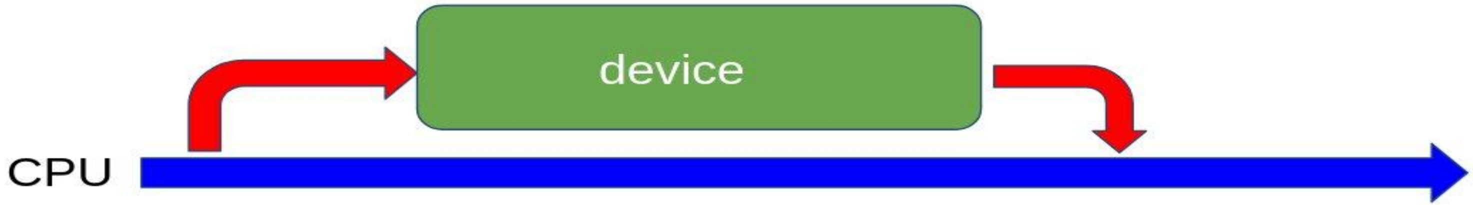
NVIDIA GA100

- ✓ up to 128 Streaming multiprocessor (SM)
- ✓ Each SM has
  - 64 FPU@32bit
  - 32 FPU@64bit
  - 64 INT@32bit

GPU offload

Offloading:

- ✓ Some work is “demanded” to an “external” device (GPU,FPGA,...)
- ✓ Explicit data movement back and from the device
- ✓ The bottleneck is the data movement
- ✓ Usually, devices has less memory than CPU
    - Leonardo: 4x64GB vs. 512 GB



Why GPUs?

- ✓ **Pro**
  - GPUs are more powerful: 1 GPU ~ 10x CPUs (Peak Mflops)
  - GPUs ask for less space: for same performance CPUs ask for ~3x racks
  - GPUs are less expensive: for same peak performance CPUs are ~2x expensive
  - GPUs ask for (relative) less power: for same peak performance CPUs ask ~4x energy
- ✓ **Cons**
  - GPUs are less flexible respect CPUs
  - Some algorithm are not GPU-friendly
  - There's no a common programming model between different vendors
  - Porting to GPU is expensive and error-prone procedure

Top500 (20 years ago)

Rank	System	Cores	Rmax (GFlop/s)	Rpeak (GFlop/s)	Power (kW)
1	Earth-Simulator, NEC Japan Agency for Marine-Earth Science and Technology Japan	5,120	35,860.00	40,960.00	3,200
2	ASCI Q - AlphaServer SC45, 1.25 GHz, HPE Los Alamos National Laboratory United States	8,192	13,880.00	20,480.00	
3	X - 1100 Dual 2.0 GHz Apple G5/Mellanox Infiniband 4X/Cisco GigE, Self-made Virginia Tech United States	2,200	10,280.00	17,600.00	
4	**Tungsten** - PowerEdge 1750, P4 Xeon 3.06 GHz, Myrinet, DELL EMC NCSA United States	2,500	9,819.00	15,300.00	
5	**Mpp2** - Cluster Platform 6000 rx2600 Itanium2 1.5 GHz, Quadrics, HPE DOE/SC/Pacific Northwest National Laboratory United States	1,936	8,633.00	11,616.00	
6	**Lightning** - Opteron 2 GHz, Myrinet, Linux Networx Los Alamos National Laboratory United States	2,816	8,051.00	11,264.00	

- 1 Self-made Supercomputer
- No GPU
- 1 Vector Machine (NEC)
- 5 different vendor/integrator
- 3.2 MW for #1
- ~10 TF/MW

- ~~Linux Network~~
- ~~NEC~~

Top500 (10 years ago)

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P, NUDT National Super Computer Center in Guangzhou China	3,120,000	33,862.70	54,902.40	17,808
2	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x, Cray/HPE DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.00	27,112.55	8,209
3	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom, IBM DOE/NNSA/LLNL United States	1,572,864	17,173.22	20,132.66	7,890
4	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect, Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.00	11,280.38	12,660
5	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom, IBM DOE/SC/Argonne National Laboratory United States	786,432	8,586.61	10,066.33	3,945
6	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x, Cray/HPE Swiss National Supercomputing Centre (CSCS) Switzerland	115,984	6,271.00	7,788.85	1,754

- 2 GPUs based
- 1 CPU based (K-Computer)
- 3 Manycore (BG/intel Phi)
- 4 different vendor/integrator
- Up to 30 MW
- ~1.9 PF/MW

- ~~NUDT~~
- ~~IBM~~

Top500 (23/11)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	4,742,808	585.34	1,059.33	24,687
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Microsoft Azure United States	1,123,200	561.20	846.84	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
6	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404

- 5 GPU-based
- 1 CPU-based (Fugaku)
- 5 different vendor/integrator
- Up to 30 MW for #1
- ~43 PF/Mw

And about precision?

FP4 Tensor Core	1,400 | 1,100[2] PFLOPS
FP8/FP6 Tensor Core	720 PFLOPS
INT8 Tensor Core	23 PFLOPS
FP16/BF16 Tensor Core	360 PFLOPS
TF32 Tensor Core	180 PFLOPS
FP32	6 PFLOPS
FP64 / FP64 Tensor Core	100 TFLOPS

NVIDIA GB300 Spec

- **FP32/FP64 = 60x**
- **FP4/FP64 = 14'000x**

Are you ready for very low precision?

It works for ML/AI but for classic science (e.g. CFD)?

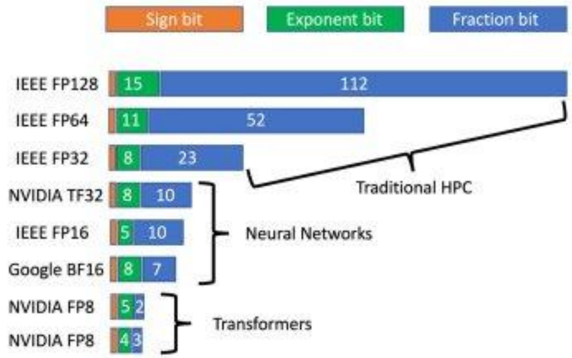


FIGURE 1. Overview of FP representations.

Questions!

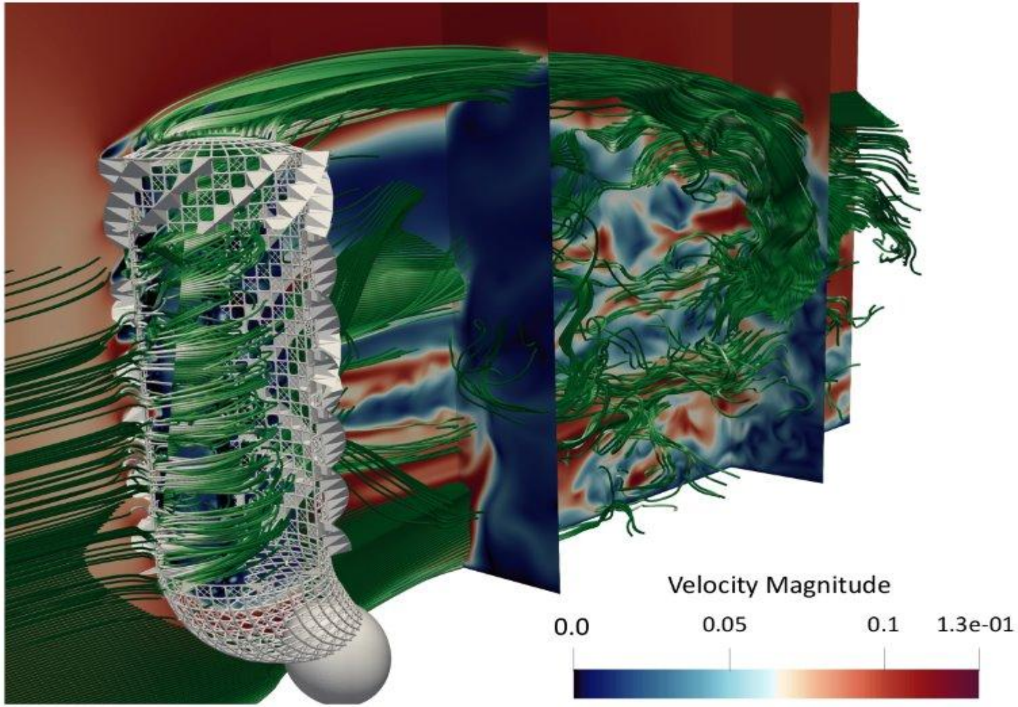
#1: Why all this complexity?
#2: Is the market big enough to survive for a HPC firm?
#3: Which skill is the more important?
#4: What is the performance range?
#5: Why GPUs?
#6: And for the next 20 years?

Some HPC works (CFD)

- ✓ LBM Performance (G. Amati et al.)
- ✓ Turbulent Pipe Flow (S. Pirozzoli et al.)
- ✓ Supersonic Boundary Leyer (M. Bernardini et al.)
- ✓ Turbulence with Polymers (P. Gualtieri et al.)

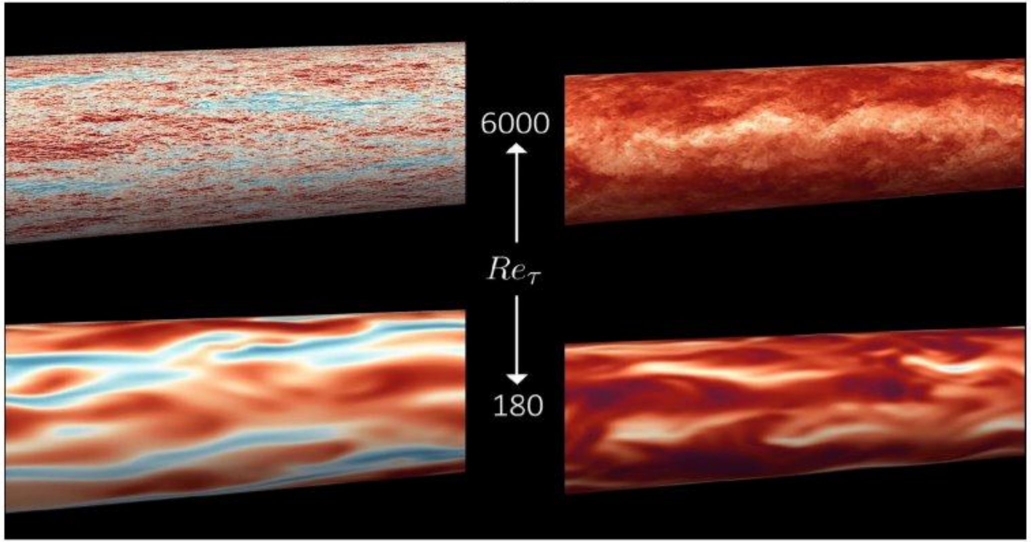
LBM

- ✓ G. Amati (CINECA) et al.
  - Flow through Silica Sponge
  - MPI+openACC
- ✓ Max Run
  - Re = ~5000
  - # GPU = 12'000
- ✓ Production run
  - Re = 5000
  - # GPU = 840
  - **Gridpoints ~ 107'000'000'000**
- ✓ https://www.researchgate.net/profile/Giorgio-Amati
- ✓ https://www.nature.com/articles/s41586-021-03658-1



Turbulent pipe flow

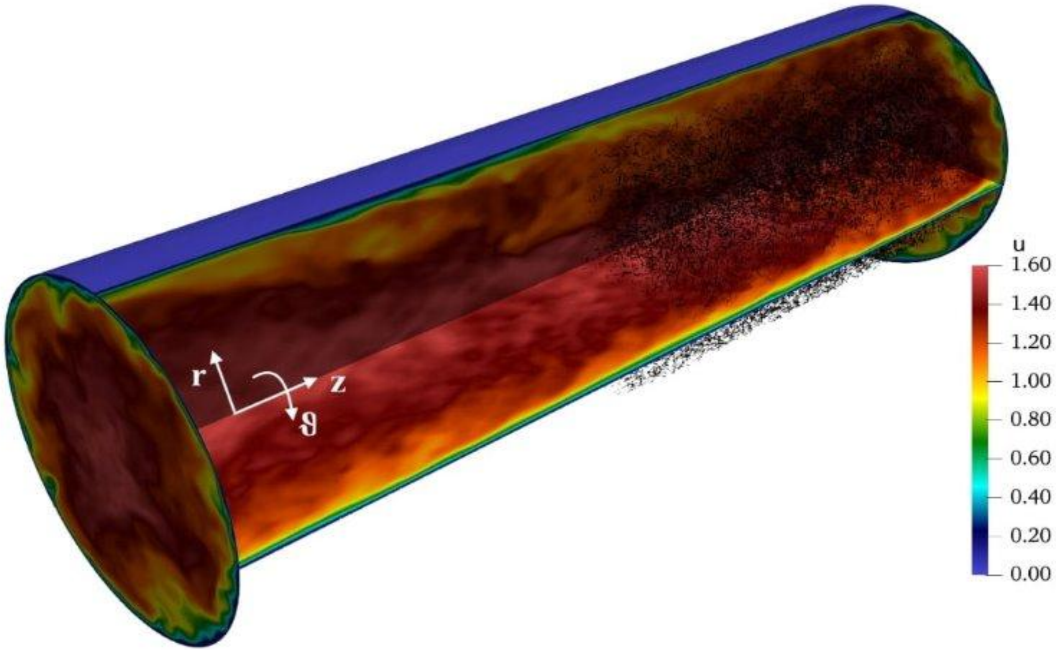
- ✓ S. Pirozzoli, A. Ceci (Univ. Rome)
  - Turbulent channel/pipe flow
  - NCCL+CUDAFortran
- ✓ Max Run (scalability)
  - Re = ~12'000
  - GPU = 2048
- ✓ Production run
  - Re = 12'000
  - GPU = 512Gridpoint ~ **70'000'000'000**
- ✓ https://www.researchgate.net/profile/Sergio-Pirozzoli
- ✓ https://youtu.be/vO0w5LUsLiM?si=bzXkFad0Tthh7xUP



Supersonic BL with microramps

- ✓ M. Bernardini(Univ. Rome)
  - MPI+CUDAFortran
  - MPI+HIP
- ✓ Max Run (scalability)
  - GPU = 12288
- ✓ Production run
  - Mach=2
  - Re_tau=2000
  - GPU = 2048
  - Gridpoint ~ **75'000'000'000**
- ✓ https://www.researchgate.net/profile/Matteo-Bernardini-2

Turbulence + polymers



- ✓ P. Gualtieri (Univ. Rome)
  - MPI+CudaFortran
- ✓ Production run
  - Re = 40000
  - Re_tau = 1000
  - GPU ~ 2048
  - Grid: ~**15'000'000'000**
  - Polymer~ **1'000'000'000**
- ✓ https://scholar.google.it/citations?user=GmSjDjMAAAAJ&hl=it

Recap

- ✓ Different skills are required to achieve “good” performance
- ✓ Performance is not only a problem of choosing the right (powerful) HW
- ✓ HW evolution is driven by mass market
- ✓ All firms devoted only to HPC have not survived to the market
- ✓ Users should (or must) be flexible enough to follow HW & SW evolution
- ✓ A Correct code could be efficient or not. With different order of magnitude!
- ✓ Today any processor is a parallel one
  - To have a parallel code doesn’t mean to have an efficient one
- ✓ To be fast is secondary respect to be correct
  - “Premature optimization is the root of all evils” (D. Knuth)
- ✓ But you’ll must face optimization issues soon
  - 1 way to go fast, 100 ways to go slow!
- ✓ Today CPUs/GPUs can have order of 100’000’000’000 transistors

Take home message

- ✓ HPC is complicate stuff, many skills are needed
  - you have to know how HW works
  - you have to know how SW works
  - you have to know about numerics
  - you have to know about the problem to solve
- ✓ Good/Bad performances depends (also) from the user
- ✓ Things change over the year
- ✓ Sorry: no “silver bullet” or “free lunch”
- ✓ No one will develop a CPU/GPU specifically for you……