Massive parallel programming on Graphics Processing Units and Applications (part 1)

Lokman Abbas-Turki
lokmane.abbas_turki@sorbonne-universite.fr

This project has received funding from the European High-Performance Computing Joint Undertaking under grant agreement No 101051997

October 2023

EUMaster4HPC

Plan

High Performance Computing before GPUs

- From dedicated machines to clusters of commercialized hardware
- From warfare to welfare: amortize the cost of IC
- Discussing Moore's Law and memory bottleneck

Disruption due to GPUs

- GPU vs. CPU: processors difference and latency vs. bandwidth
- When GPUs are coprocessors: Amdahl and Gustafson laws
- Scalability and the new Moore's laws

Actual and future possible evolutions

- Dedicated architectures
- CUDA becomes a standard
- From grid computing to cloud computing and quantum computing

Plan

High Performance Computing before GPUs

- From dedicated machines to clusters of commercialized hardware
- From warfare to welfare: amortize the cost of IC
- Discussing Moore's Law and memory bottleneck

Disruption due to GPUs

- GPU vs. CPU: processors difference and latency vs. bandwidth
- When GPUs are coprocessors: Amdahl and Gustafson laws
- Scalability and the new Moore's laws

Actual and future possible evolutions

- Dedicated architectures
- CUDA becomes a standard
- From grid computing to cloud computing and quantum computing

EUMaster4HPC

From parallel to serial

ENIAC 1946

- Was the first electronic general-purpose machine
- Was a parallel machine
- Later became the first Von Neumann machine
- Was used for Monte Carlo simulation
- Although developed for ballistic research, it was first used for hydrogen bomb computations

Toward transistor Machines

- Seymour Cray, "the Thomas Edison of the supercomputing industry"
- Tradic (1954): The first transistor machine, Bell Tel. Labs
- From germanium to silicon to planar process: Texas Instrument, Shockley Semiconductor Laboratory, Fairchild Semiconductor

EUMaster4HPC

From parallel to serial

Supercomputer race

- IBM machines vs. CDC machines: CDC 1604, IBM 7030 Stretch, CDC 6600, IBM System/360
- Used by specialists and dedicated essentially to military applications
- Vector processors appeared in the early 1970s
- Some well known parallel machines: ILLIAC IV (1971) and Cray 1 (1976)

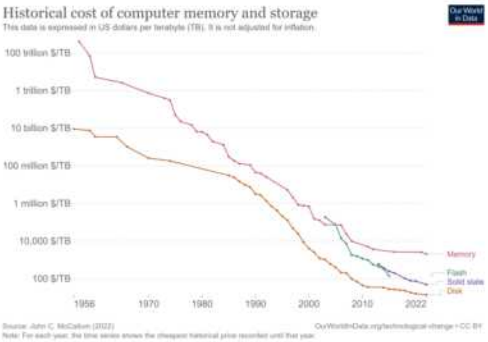
Expencive technology

- MOS transistor 1959: Bell Labs
- MOS integrated circuit used then for SRAM and DRAM
- Intel 1103 (1970): Commercialization of the first DRAM IC
- Intel 4004 (1971): Commercialization of the first microprocessor
- Amortizing the production costs by selling to the large public
- The memory hierarchy (Registers, cache, RAM, Hard Disc) is essential in computers

EUMaster4HPC

From parallel to serial

RAM became affordable



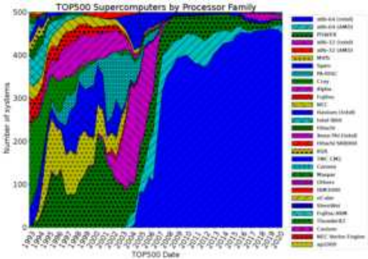
Scientific simulation

- Caltech Cosmic Cube (1981): Proposing a parallel computer at a reasonable cost
- From 1980s to 2000s: Supercomputers essentially based on serials processors RISCs and CISCs
- Difficulties due to multiplicities of platforms, inefficient inter-machine communication and insufficient documentation
- Applied mathematics expanded in the world of serial resolution of PDEs

From serial to parallel

x86 became the king

- Intel introduced the first x86 microprocessors in 1978
- IBM created PC in 1981 and wanted x86 processors
- AMD became second-source manufacturer for x86 microprocessors
- IBM continued the production of PowerPC RISC microprocessors

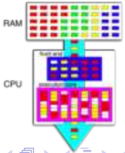


The end of Cold War

- Sandia's Paragon supercomputer, in 1993 using Intel x86
- Cray Research bankruptcy in 1995
- Roadrunner and PlayStation

Hyper-threading 2002

Logical processors sharing the same resources: cache, bus interface

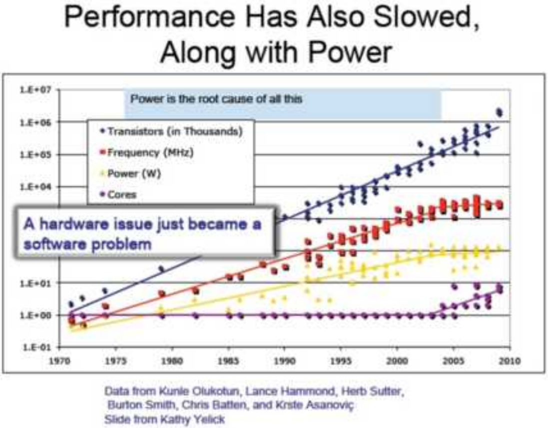


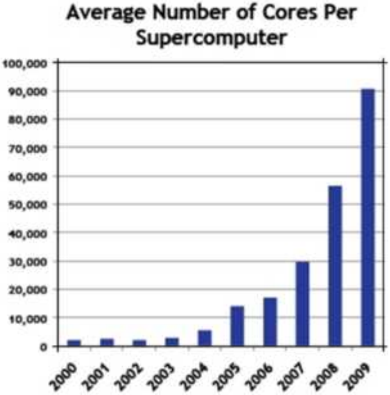
EUMaster4HPC

From serial to parallel

The best-selling author Mickeal Lewis about Silicon Valey

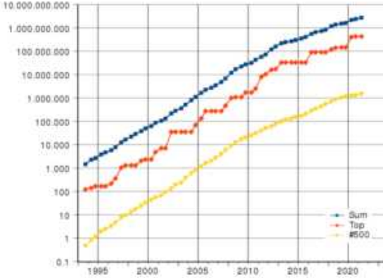
"It is a cold place, ... The people get artificially excited by technology but it does not feel warm or hot."





From serial to parallel

Maintain Moore's Law What types of processors should we add?

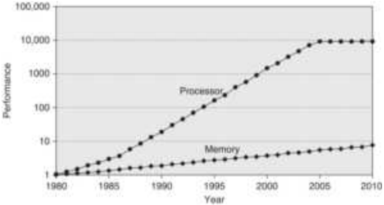


EUMaster4HPC

More statistics on https://www.top500.org/statistics/

From serial to parallel

Processor–memory performance gap



from J. L. HENNESSY and D. A. PETTERSON, *Computer Architecture*.

Larger bandwidth for almost unchanged latency

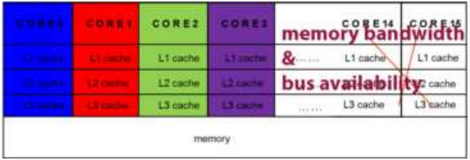
EUMaster4HPC

Plan

EUMaster4HPC

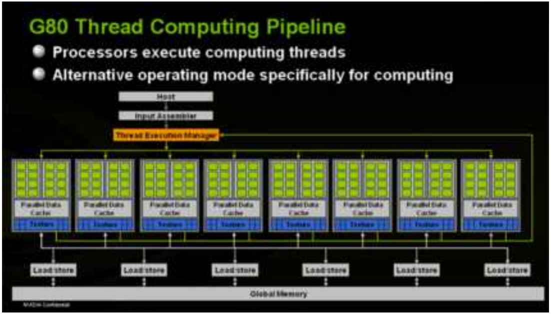
Architecture overview

Sandia National Laboratories 2009
16 cores ≈ 2 cores



The limit architecture! GPU (Graphic Processing Unit)

No branching prediction + much smaller size of cache per processor



Bill Dally, VP research at Nvidia

"Locality is efficiency, efficiency is power, power is performance, performance is king."

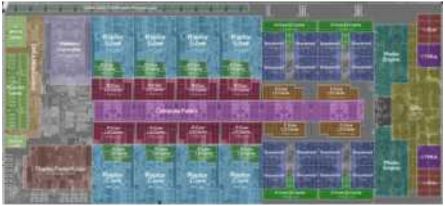
Architecture evolution

big.LITTLE + small GPU on x86

from complex to simple instruction set architectures

AI for GPUs & GPUs for clouds

more bandwidth and more dedicated cores



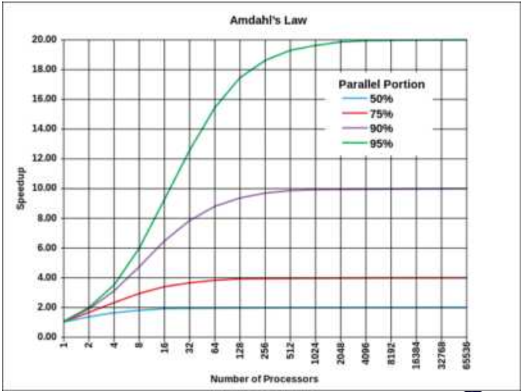
Amdahl's law

For a fixed problem

$$T(P) = T(1)\left(\alpha + \frac{1-\alpha}{P}\right), \quad S(P) = \frac{T(1)}{T(P)} = \frac{1}{\alpha + \frac{1-\alpha}{P}}, \tag{1}$$

α: the fraction of the algorithm that is purely serial

From Wikipedia



EUMaster4HPC

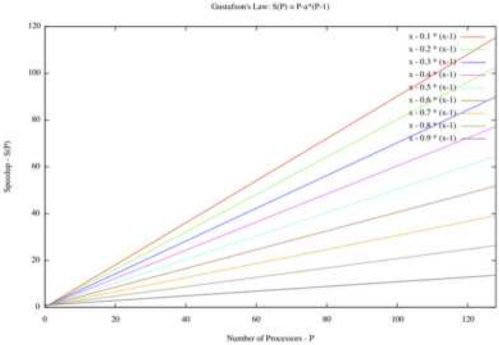
Gustafson's law

Making it bigger

$$T(1) = (\alpha + [1-\alpha]P)T(P), \quad S(P) = \frac{T(1)}{T(P)} = P - \alpha(P-1), \qquad (2)$$

α: the fraction of the algorithm that is purely serial

From Wikipedia

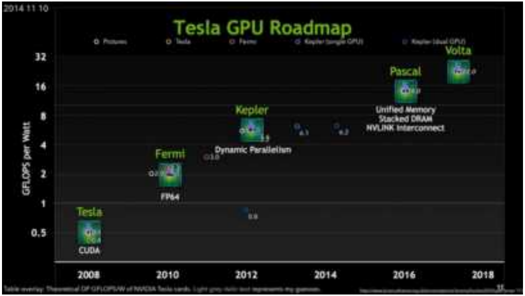


Huang's Law

Is Moore's law dead?

Maybe not dead but almost obsolete

Huang's Law is the new Moore's Law? GPUs performance more than doubles every two years



- epochai.org empirical results on GPUs:
  - the amount of FLOP/s per \$ doubles every ∼2.5 years
  - the amount of FLOP/s per \$ for machine learning workloads doubles every ∼2.07 years

Chris Miller, the author of CHIP WAR book

"So the important question isn't whether we're finally reaching the limits of Moore's Law … but whether we've reached a peak in the amount of computing power a chip can cost-effectively produce. Many thousands of engineers and many billions of dollars are betting not."

Training AI Law

IEEE Spectrum

epochai.org

EUMaster4HPC

Plan

High Performance Computing before GPUs

- From dedicated machines to clusters of commercialized hardware
- From warfare to welfare: amortize the cost of IC
- Discussing Moore's Law and memory bottleneck

Disruption due to GPUs

- GPU vs. CPU: processors difference and latency vs. bandwidth
- When GPUs are coprocessors: Amdahl and Gustafson laws
- Scalability and the new Moore's laws

Actual and future possible evolutions

- Dedicated architectures
- CUDA becomes a standard
- From grid computing to cloud computing and quantum computing

Bigger software problem

Steve Jobs "What is software?... software is something that is changing too rapidly, or you don't exactly know what you want yet, or you didn't have time to get it into hardware."

The adoption of RISC is gaining momentum

- Low power (simple) instruction set
- Intel turned down Apple proposal to build chips for mobile phones
- The rise of fabless companies: Nvidia, Qualcomm, AMD, Xilinx, IBM, Apple, Google, Amazon and others
- ARM RISC design served Apple and Qualcomm
- TSMC Grand Alliance
- Increasing use of RISC-V even by the Chinese SMIC (7nm)
- Microsoft Windows for ARM based architecture
- The rise of ARM based (AI) PCs: Qualcomm, Nvidia

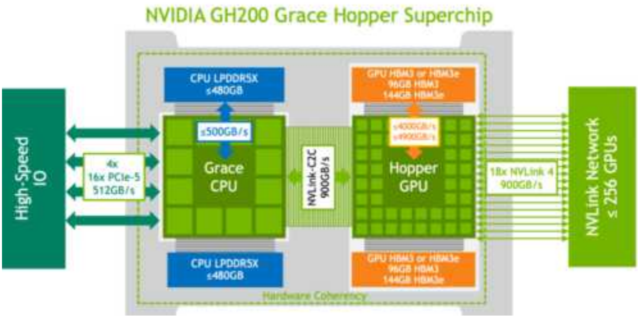
The adoption of Nvidia GPUs for AI and beyond

- Wait one year before reception when ordering H100 in september 2023
- Companies use Nvidia GPUs as collateral
- ASML CEO: "We are planning to integrate support for GPUs into all of our computational lithography software products"

EUMaster4HPC

Very expensive technology

Grace Hopper superchip



Compute-bandwidth gap Stack DRAM over GPU die
https://www.nvidia.com/en-us/on-demand/session/gtcfall22-a41187/

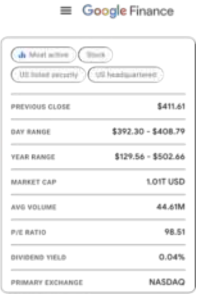
Amortization strategy
- gaming remains an important market
- clouds can include even more expensive solutions
- GPUs are champions for AI and mining crypto
- RISC architecture + GPU is very well suited to various markets: high-performance PCs, smart cars, digital twin factories, and so on

EUMaster4HPC

CUDA vs. other programming solutions

Programming options for GPUs

- OpenCL: low level language and verbose, can be implemented on all cards but less and less used
- OpenACC (came from OpenHMPP): a directives language, its use does not require to rewrite the CPU code
- CUDA: dedicated to Nvidia cards, but possible porting to
  * ROCm/HIP for AMD GPUs with a syntax that is quite similar to CUDA `https://rocm.docs.amd.com/projects/HIP/en/develop/`
  * oneAPI released recently and has a very different syntax from CUDA and HIP
- Numba JIT functions: syntax similar to CUDA



EUMaster4HPC

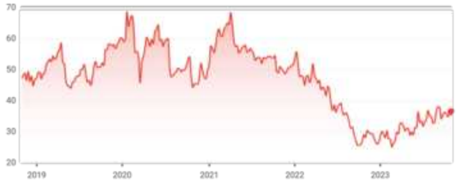
Intel Corporation

$36.50 ↓22.52% -10.61 5Y

Pre-market: $36.30 (↓0.55%) -0.20

Closed: Nov 1, 6:46:10 AM UTC-4 · USD · NASDAQ · Disclaimer

1D 5D 1M 6M YTD 1Y 5Y MAX



Google Finance

PREVIOUS CLOSE	$35.69
DAY RANGE	$35.62 - $36.57
YEAR RANGE	$24.73 - $40.07
MARKET CAP	153.88B USD
AVG VOLUME	36.70M
P/E RATIO	-
DIVIDEND YIELD	1.37%
PRIMARY EXCHANGE	NASDAQ

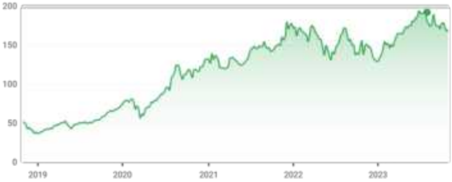
Apple Inc

$170.77 ↑229.23% +118.90 5Y

Pre-market: $170.10 (↓0.39%) -0.67

Closed: Nov 1, 7:05:37 AM UTC-4 · USD · NASDAQ · Disclaimer

1D 5D 1M 6M YTD 1Y 5Y MAX



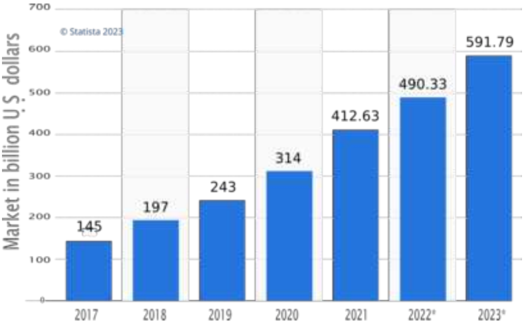
Google Finance

PREVIOUS CLOSE	$170.29
DAY RANGE	$167.90 - $170.90
YEAR RANGE	$124.17 - $198.23
MARKET CAP	2.67T USD
AVG VOLUME	54.80M
P/E RATIO	28.70
DIVIDEND YIELD	0.56%
PRIMARY EXCHANGE	NASDAQ

EUMaster4HPC

The ascent of cloud computing

doubling end-user spending for every three years



Unequal with respect to sectors

opening real future opportunities

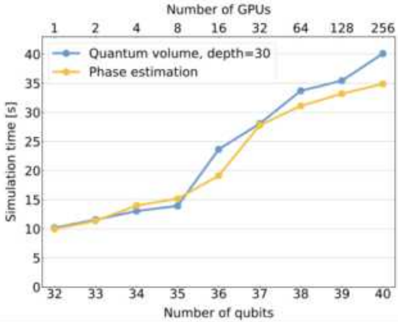
EUMaster4HPC

Disentangling Hype from Practicality: On Realistically Achieving Quantum Advantage by T. Hoefler & al.

	GPU	ASIC	Future Quantum
I/O Bandwidth	10,000 Gbit/s	10,000 G/s	1 Gbit/s
Operation throughput			
16-bit floating point	195 Top/s	550 Top/s	10.5 kop/s
32-bit integer	9.75 Top/s	215 Top/s	0.83 kop/s
binary (Boolean logical)	4,992 Top/s	77,000 Top/s	235 kop/s

Performance comparison

cuQuantum SDK: A High-Performance Library for Accelerating Quantum Science by H. Bayraktar & al.



The simulation time of the extended *Qiskit Aer* multi-node simulator on the Selene supercomputer.

EUMaster4HPC

Concluding remarks

Why programming GPUs?

- Effective amortization of very expensive technology
- Computer Graphics and gaming consume AI
- AI is general purpose
- New expensive GPUs are made up of multiple GPUs
- The multi-chip modules of GPUs favor better quantum emulation
- Expensive GPUs can be cheaply used on clouds
- After acquiring Mellanox, Nvidia is very invested in building supercomputers
- Your competitor is doing it

Ray Dalio the founder of the world's largest hedge fund

"... and then with the new technologies we are going through a time warp, we are going to be in a different world ..."

My advice

"Avoid working on anything that is not scalable with respect to an increasing size of data or not scalable with respect to greater computing power."

EUMaster4HPC