Quantum and High Performance Computing: "Accelerating" High Performance Computing with Qauntum Processing Units

April 1, 2022


Murat Manguoğlu

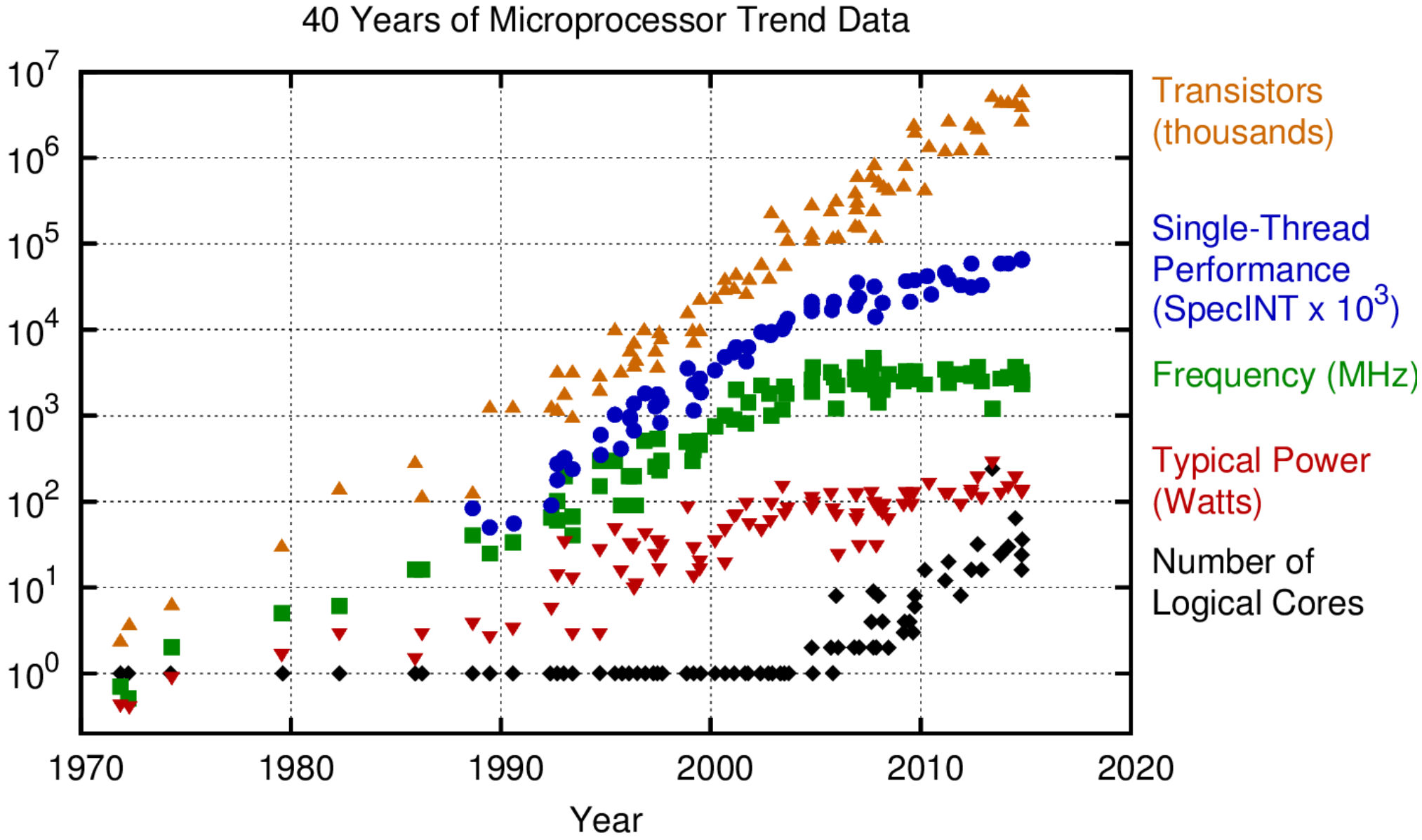
Middle East Technical University

Department of Computer Engineering

Outline

- Trends and current limitations of HPC systems
- Trends and current limitations of QC systems
- Likely merge of HPC and QC: Quantum accelerated cluster architectures (QACA)
- Performance metrics and Amdhal's law for QACA
- An example: Simon's problem and its adaptation to QACA

Trends– Moore's Law

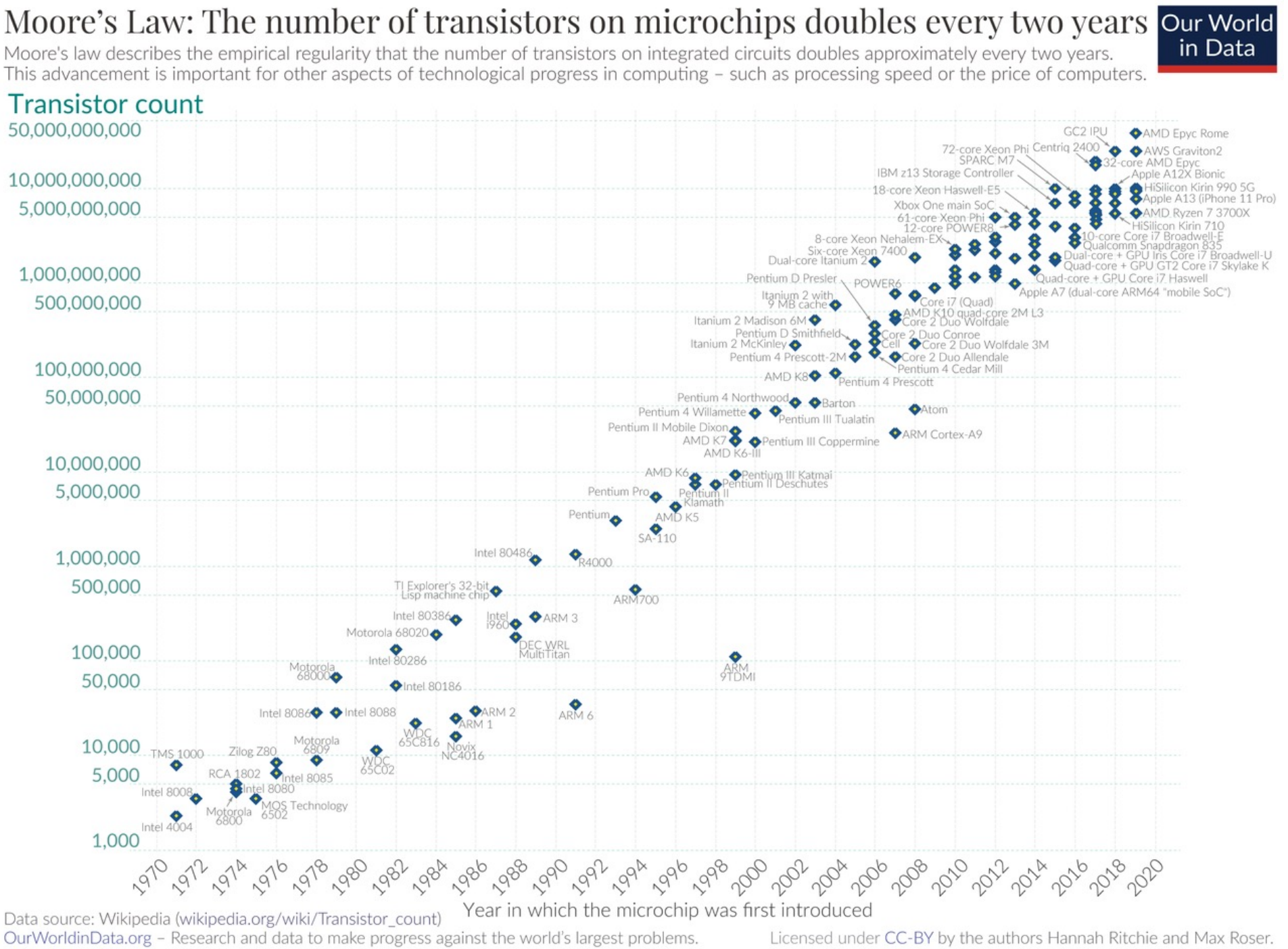


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp
Image source: https://www.karlrupp.net/wp-content/uploads/2015/06/40-years-processor-trend.png

1.8 nm chip planned by Intel in 2024

↓

Diameter of an hydrogen atom: ~0.1 nm



Most Powerful Computers

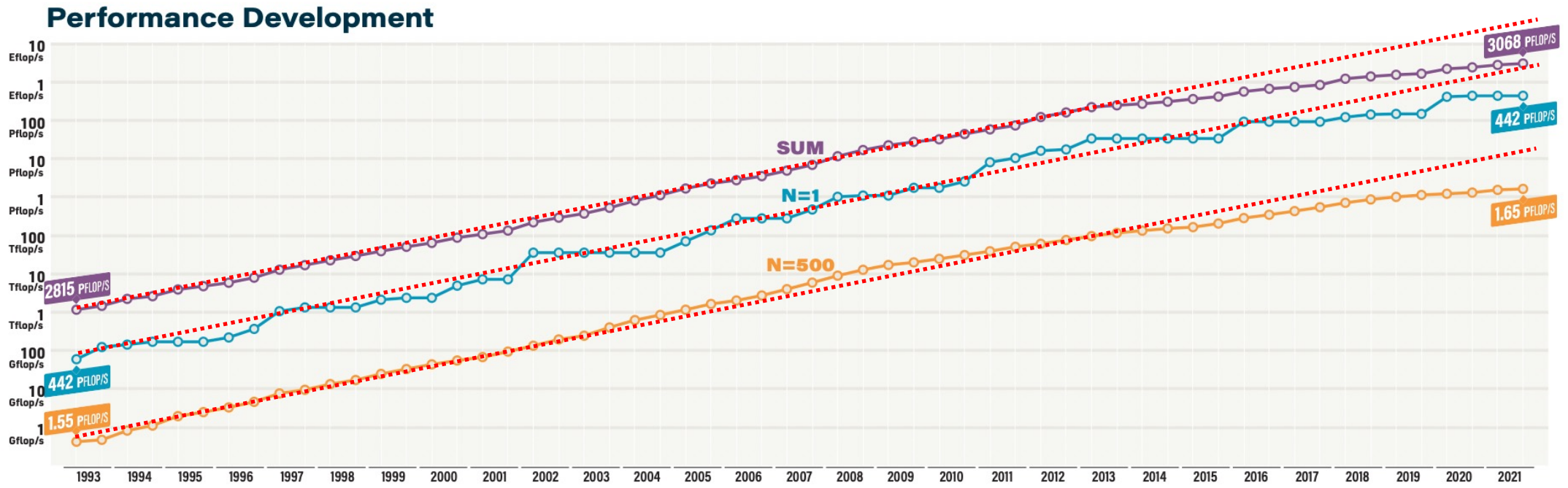
TOP 500 The List.

PRESENTED BY: ICL INNOVATIVE COMPUTING LABORATORY; BERKELEY LAB Lawrence Berkeley National Laboratory; ISC GROUP Moving Forward.

FIND OUT MORE AT: top500.org

NOVEMBER 2021

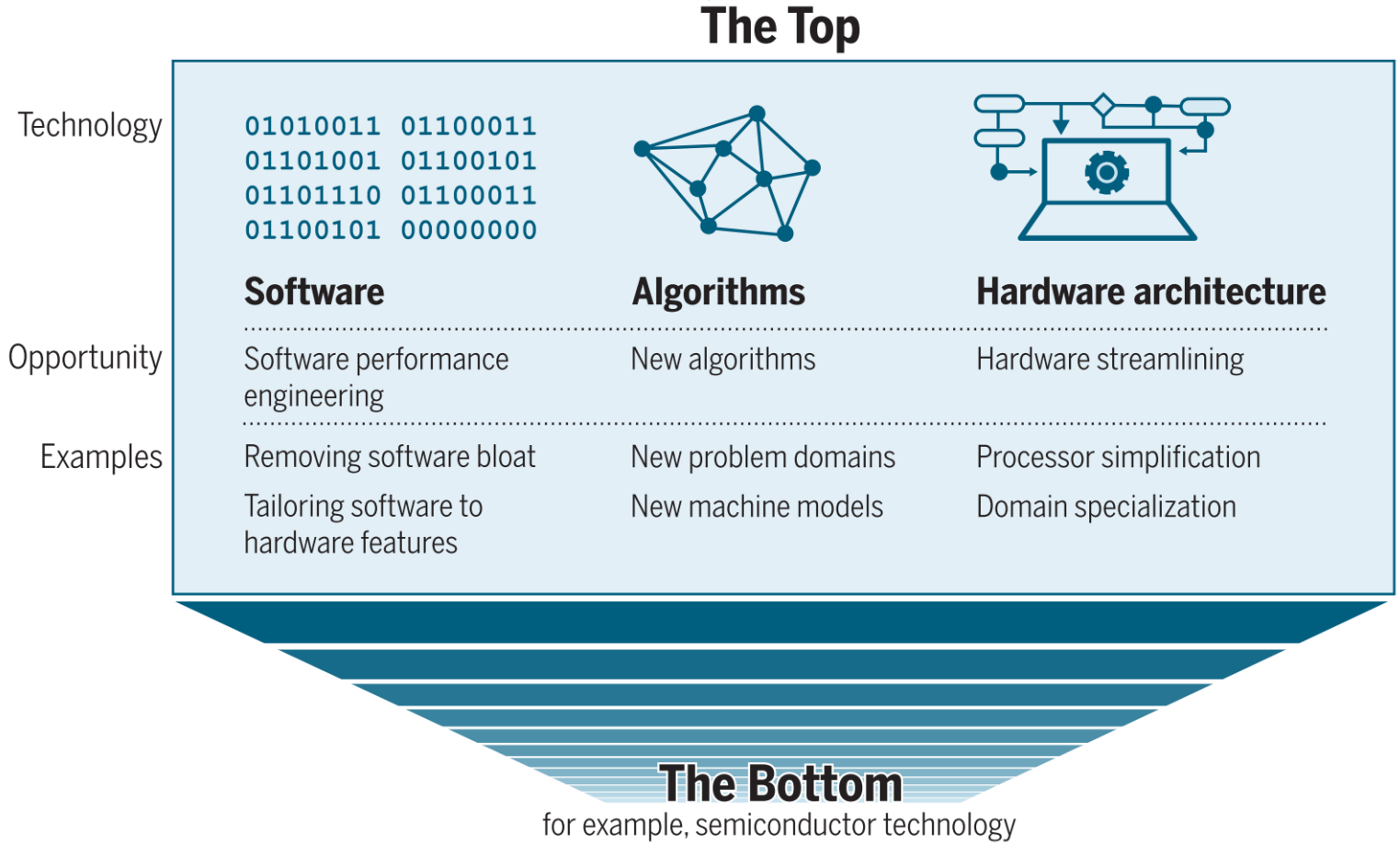
			SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	Fugaku	Fujitsu A64FX (48C, 2.2GHz), Tofu Interconnect D	RIKEN R-CCS	Japan	7,630,848	442.0	29.9
2	Summit	IBM POWER9 (22C, 3.07GHz), NVIDIA Volta GV100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/SC/ORNL	USA	2,414,592	148.6	10.1
3	Sierra	IBM POWER9 (22C, 3.1GHz), NVIDIA Tesla V100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/NNSA/LLNL	USA	1,572,480	94.6	7.44
4	Sunway TaihuLight	Shenwei SW26010 (260C, 1.45 GHz) Custom Interconnect	NSCC in Wuxi	China	10,649,600	93.0	15.4
5	Perlmutter	HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10 (274 GB)	LBNL	USA	761,856	70.9	2.58



source: www.top500.org

Moore's law might be already failing!

But there is more room for improvement in the "post-Moore era"[1]:

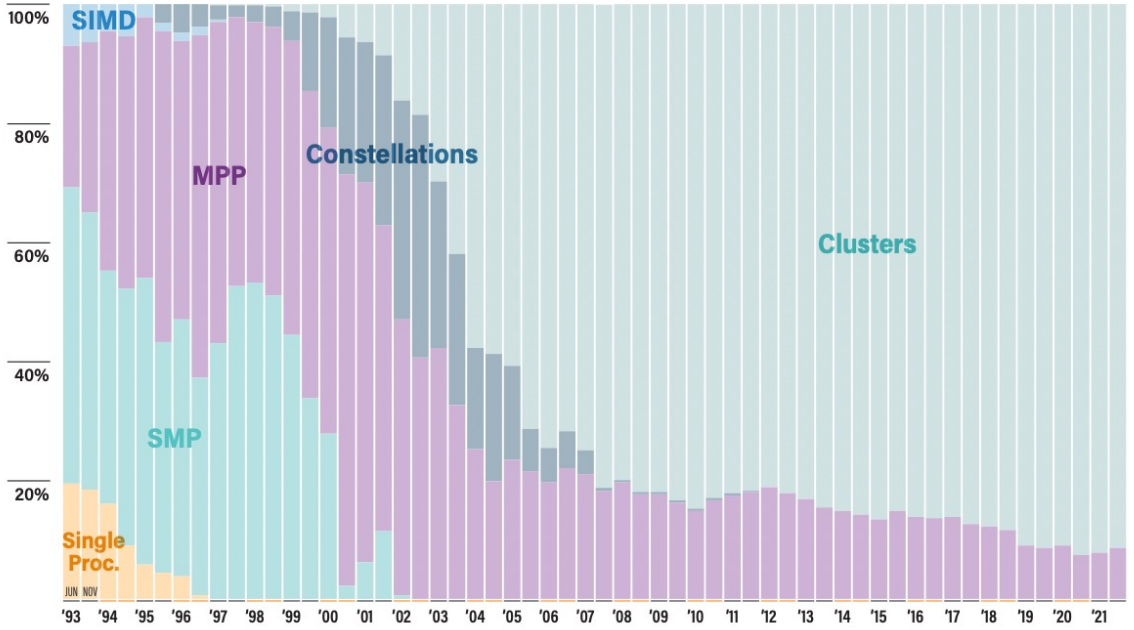


Performance gains after Moore's law ends. In the post-Moore era, improvements in computing power will increasingly come from technologies at the "Top" of the computing stack, not from those at the "Bottom", reversing the historical trend.

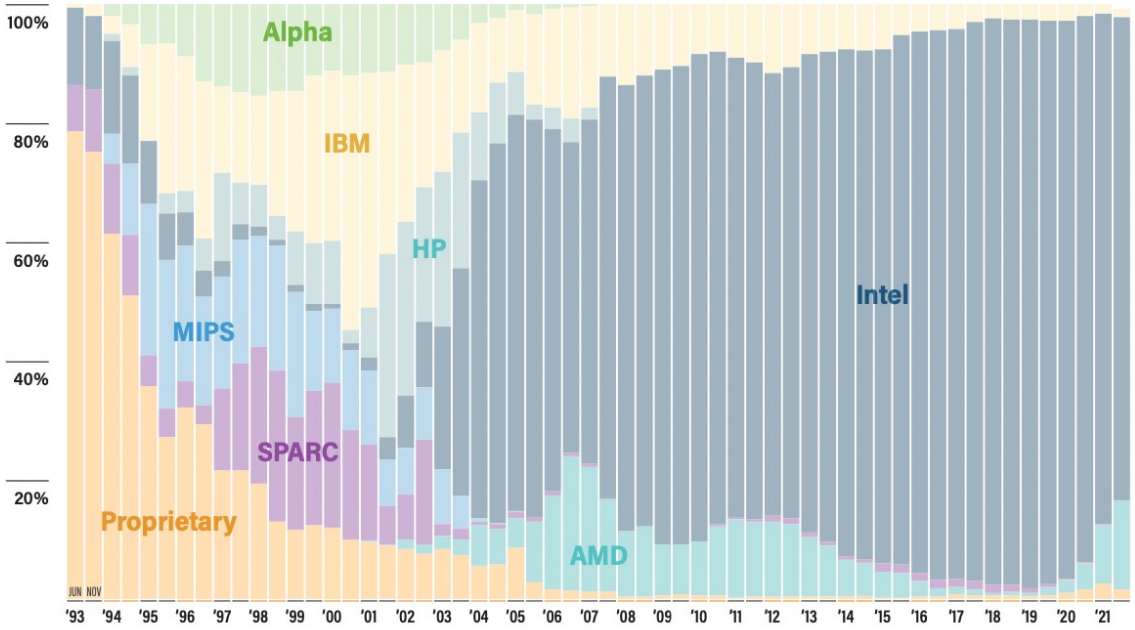
[1] Leiserson, C. E., Thompson, N. C., Emer, J. S., Kuszmaul, B. C., Lampson, B. W., Sanchez, D., & Schardl, T. B. (2020). There's plenty of room at the Top: What will drive computer performance after Moore's law?. *Science*, *368*(6495).

source: www.top500.org

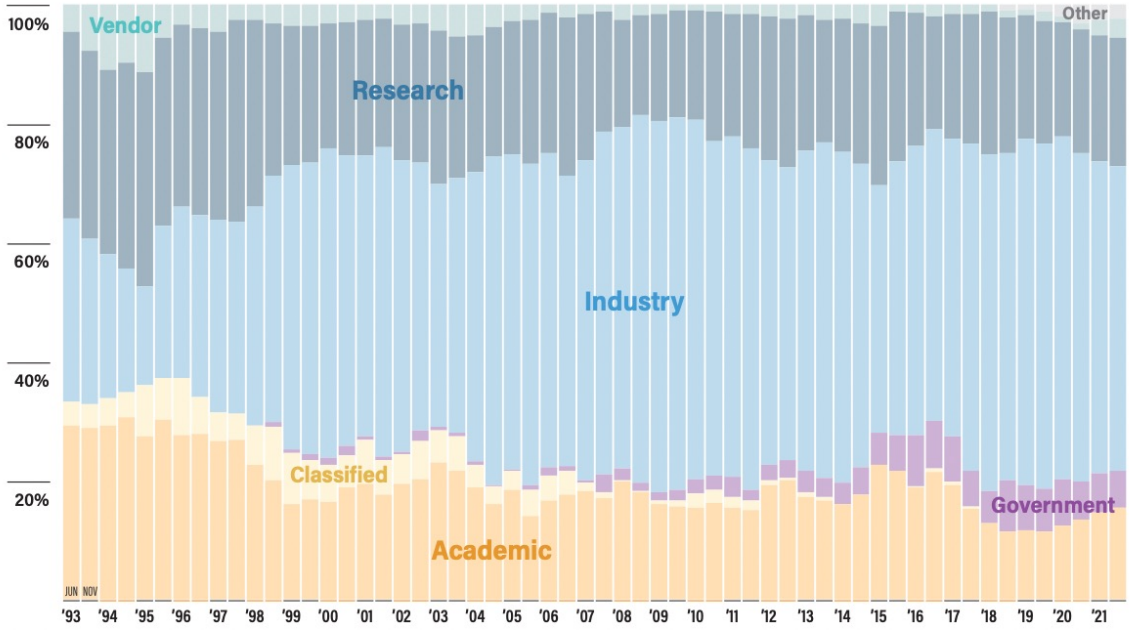
Architectures



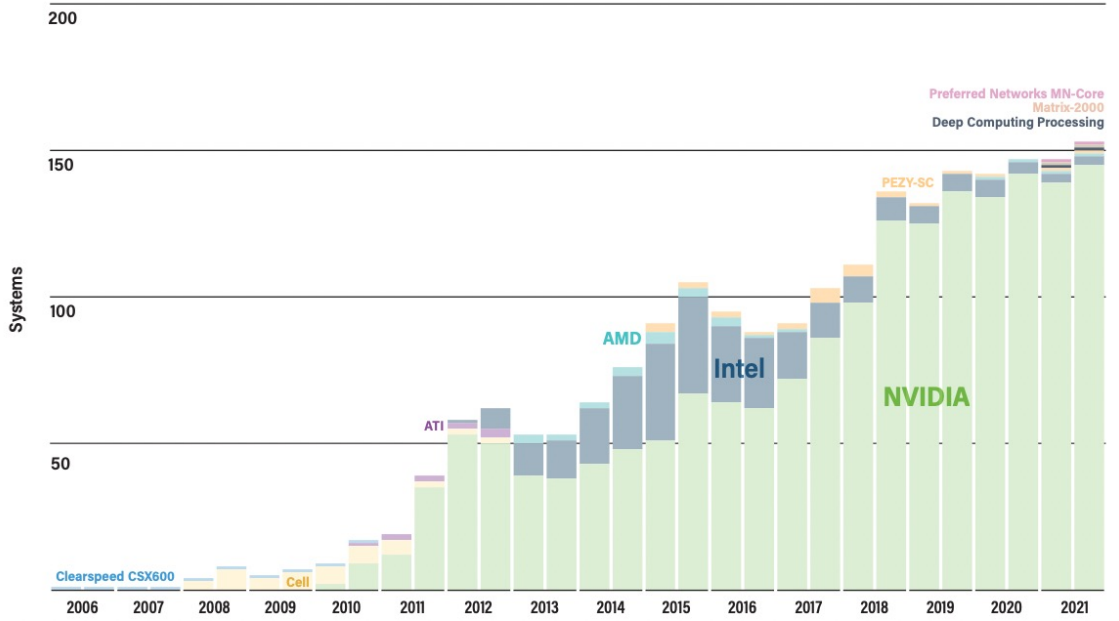
Chip Technology



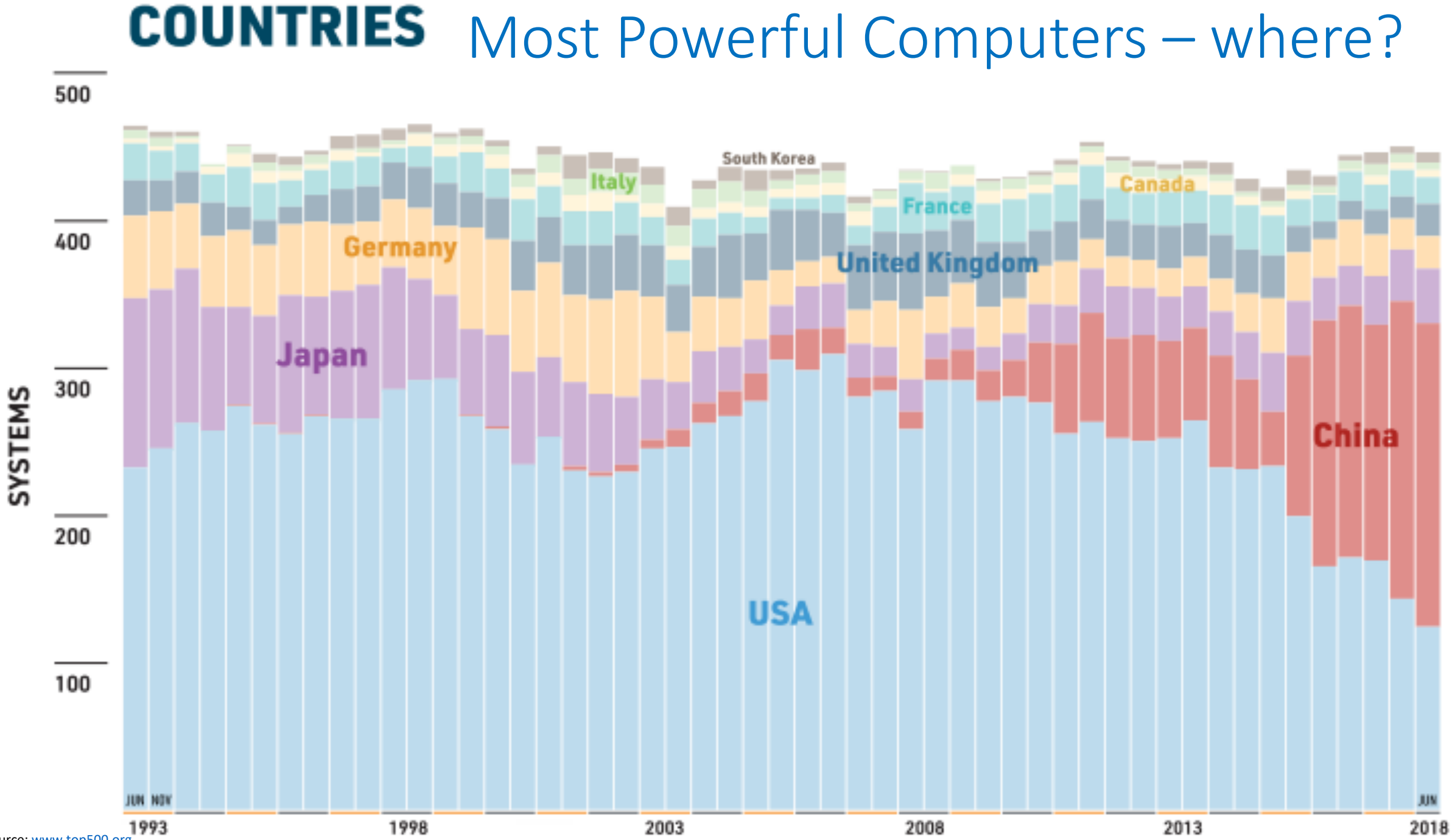
Installation Type



Accelerators/Co-processors



Most Powerful Computers – where?



source: www.top500.org

Quantum computing

Adds another layer of parallelism

- It is a generalization of classical computing but right now, they have a potential to speed up certain operations while they are not expected to be efficient in others
- It is likely that classical processors will co-exist with quantum processors
- We need to re-design algorithms for these architectures
  - Just like the best sequential algorithm is not the best one on a parallel computing platform, the best quantum algorithm may not be the best one on QAQC architectures (and the best parallel algorithm may not be the best on QAQC)

Most powerful quantum computers

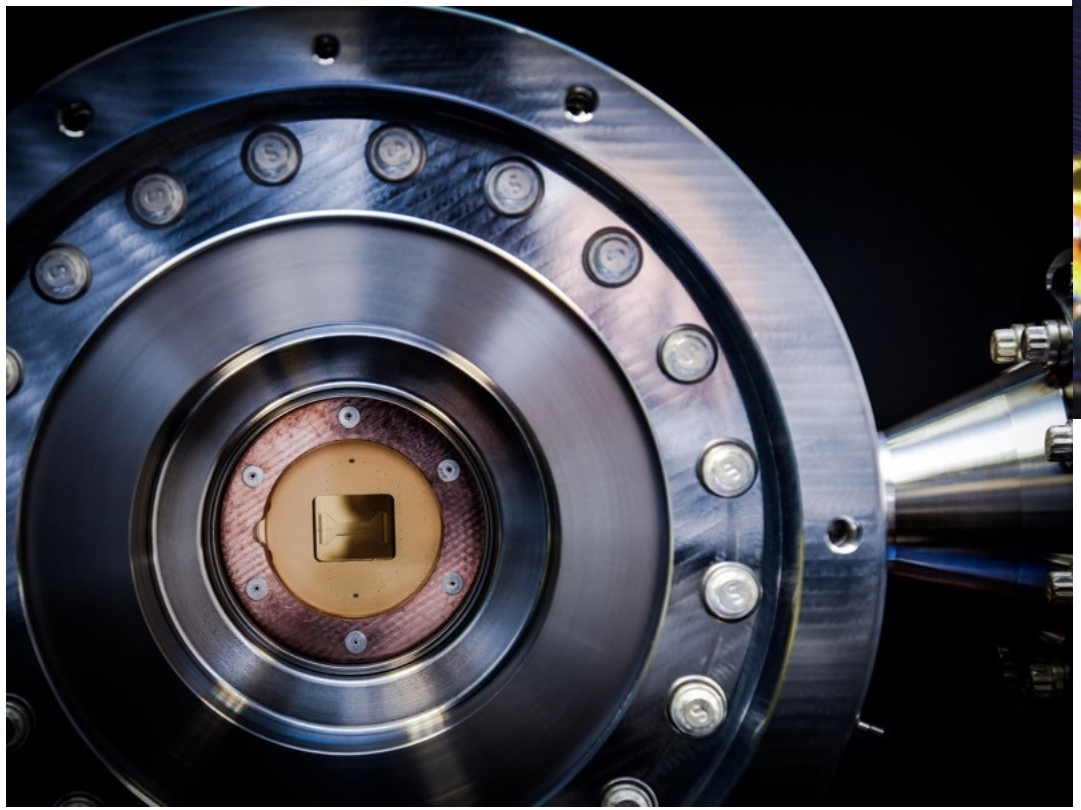
Name/Designation	Manufacturer	Architecture	Release date	Qubits
IBM Eagle	IBM	Superconducting	Late 2023	127
Jiuzhang	USTC	Photonics	2020	76
Bristlecone	Google	Superconducting	5 March, 2018	72
IBM Manhattan	IBM	Superconducting		65
Sycamore	Google	Nonlinear superconducting resonator	2019	53
IBM Q 53	IBM	Superconducting	1 October, 2019	53
IBM Q 50 prototype	IBM	Superconducting		50
N/A	Google	Superconducting	Q4 2017 (planned)	49
Tangle Lake	Intel	Superconducting	9 January, 2018	49
IBM Dublin	IBM	Superconducting		27

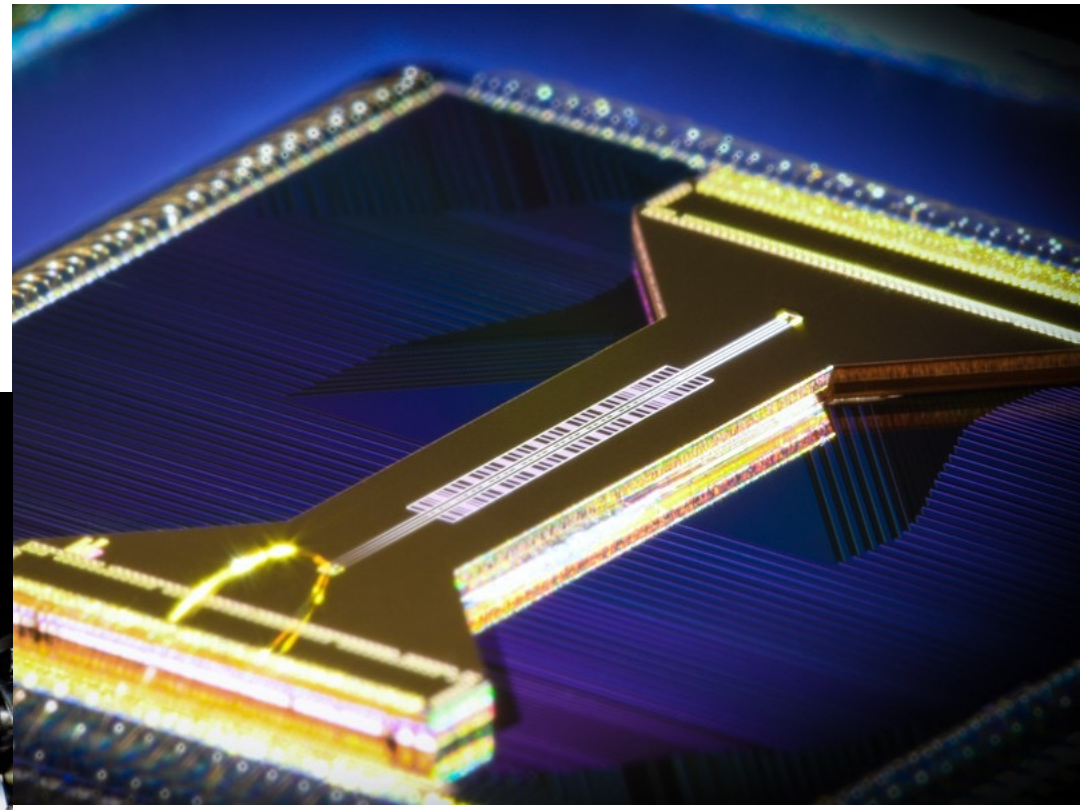
Source: IBM, Verdict, Wikipedia (sources cited: Nature, Live Science, IBM, Futurism, MIT Technology Review, IEEE Spectrum, SPIE)

Hardware

- Ion traps (atoms)
  - First quantum gate (CNOT) was build using ion traps in 1995 by C. Monreo and D. Wineland
- Superconductors (electrons)
- Optical (photons)
- Nuclear Magnetic Resonance (molecules)
- Diamond (atoms)

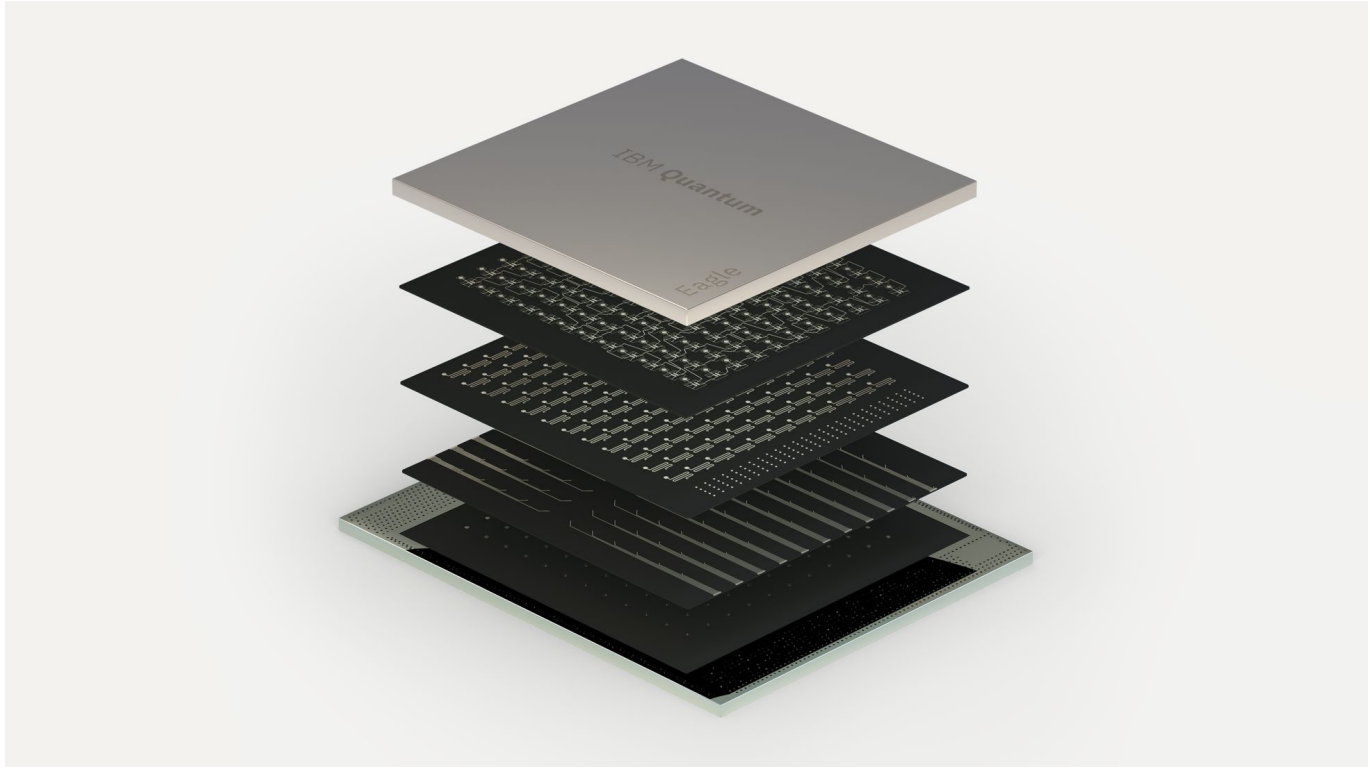
Ion traps



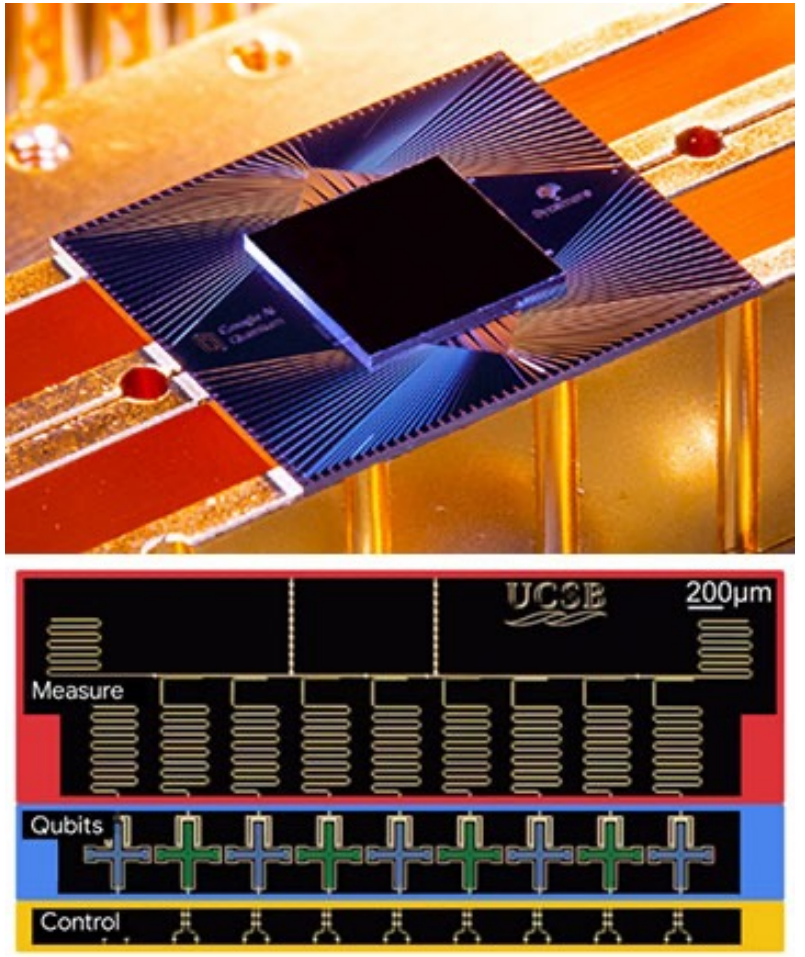


An ion trap from Honeywell's quantum computer. Credit: Honeywell Quantum Solutions

Superconductors



https://newsroom.ibm.com/2021-11-16-IBM-Unveils-Breakthrough-127-Qubit-Quantum-Processor



The Google Sycamore chip (top) involves an architecture constructed of control circuitry, superconducting qubits (in aluminum-on-silicon), and microwave resonators for measurement. [Image: Erik Lucero, Google (top); Google AI Quantum (bottom)]

Optical

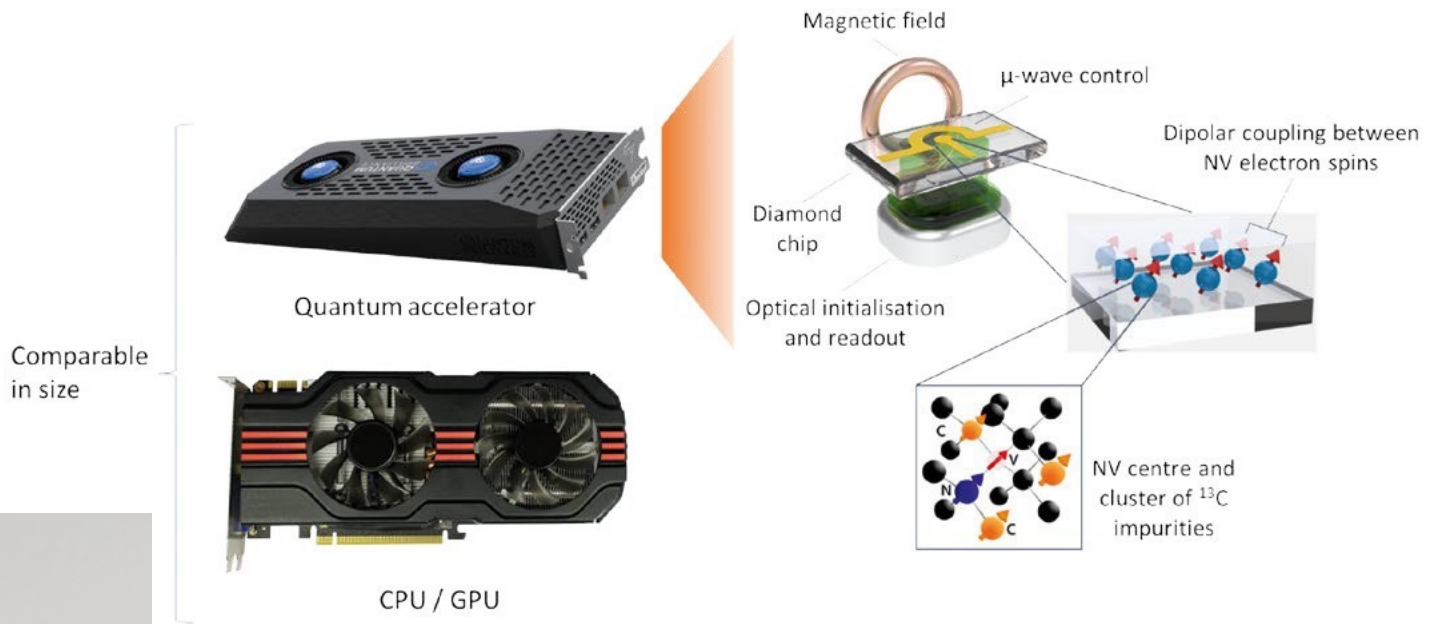
A photo of the Jiuzhang light-based quantum computer prototype
Photo: courtesy of University of Science and Technology of China

NMR

SpinQ Chief Scientist Prof. Bei Zeng from University of Guelph, announced the SpinQ Gemini, the first commercially available desktop quantum computer. Source: https://mathstat.uoguelph.ca/feature/quantum_computer

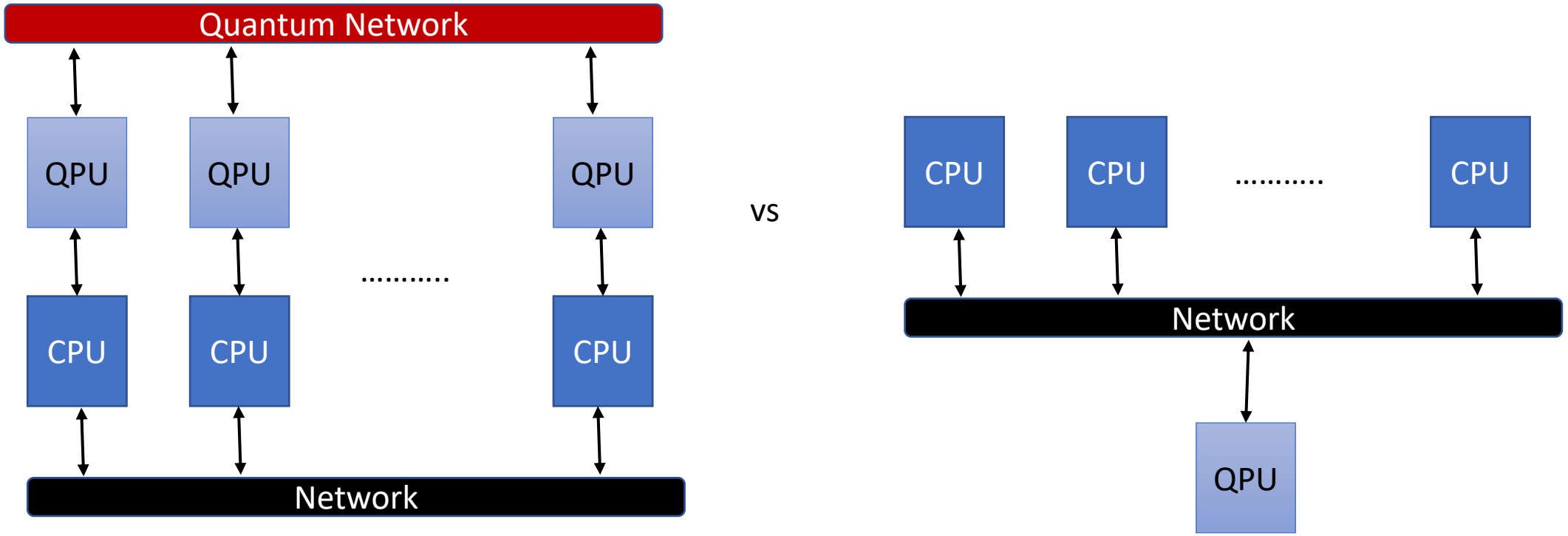
Quantum information processing experiment equipment. Image: https://ocw.mit.edu/courses/physics/8-13-14-experimental-physics-i-ii-junior-lab-fall-2016-spring-2017/experiments/quantum-information-processing/

Diamond



Room-temperature diamond Quantum Accelerators could become just another component for a PC, offering quantum capabilities when there's an advantage -**Quantum Brilliance**

Some quantum accelerated cluster architectures



Britt, Keith A., and Travis S. Humble. "High-performance computing with quantum processing units." *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 13.3 (2017): 1-13.

(Classical)Temporal Performance Metrics

- FLOPs
- Time
- Speedup
- Efficiency
- Isoefficiency
- others …

non-Temporal Performance Metrics

- Power consumption
- Accuracy
- Robustness
- Memory requirements
- others …

Parallel Running Time

- We use the Wall-Clock time (not the CPU time):
  - MPI_WTIME() - if using the MPI library
  - omp_get_wtime – if using multithreading with openmp
  - …
- Look at the wall-clock from the moment the parallel algorithm starts and at the moment last process finishes the algorithm. Time elapsed is the parallel running time: $\mathbf{T_p = t_{end} - t_{start}}$

- What is in the parallel running time?
- $\mathbf{T_p = T_{computation} + \dot{T}_{communication} + T_{idle}}$

Speed Improvement (or Speedup)

$$S \triangleq \frac{T_s^*}{T_p}$$

T_s^* : best sequential time

T_p : parallel running time

Ideal Speedup

if $S = p$,

then it is the Ideal Speedup (IS)

i.e. The applicatin is emberrasingly parallel

Parallel Efficiency

- Fraction of time in which the processing element is utilized

$$\text{eff} \triangleq \frac{S}{IS} = \frac{\frac{T_s^*}{T_p}}{p} = \frac{T_s^*}{T_p * p} = \frac{T_s^*}{\underbrace{T_p * p - T_s^*}_{T_o \triangleq \textit{Parallel Overhead}} + T_s^*}$$

$$= \frac{T_s^*}{T_o + T_s^*} = \frac{1}{1 + \frac{T_o}{T_s}}$$

Amdhal's law

T_{seq}: part of the code that can not be parallel
T_{par}: part of the code that can be parallel
p: number of processors

$T_s = T_{seq} + T_{par}$
$T_p = T_{seq} + T_{par}/p$

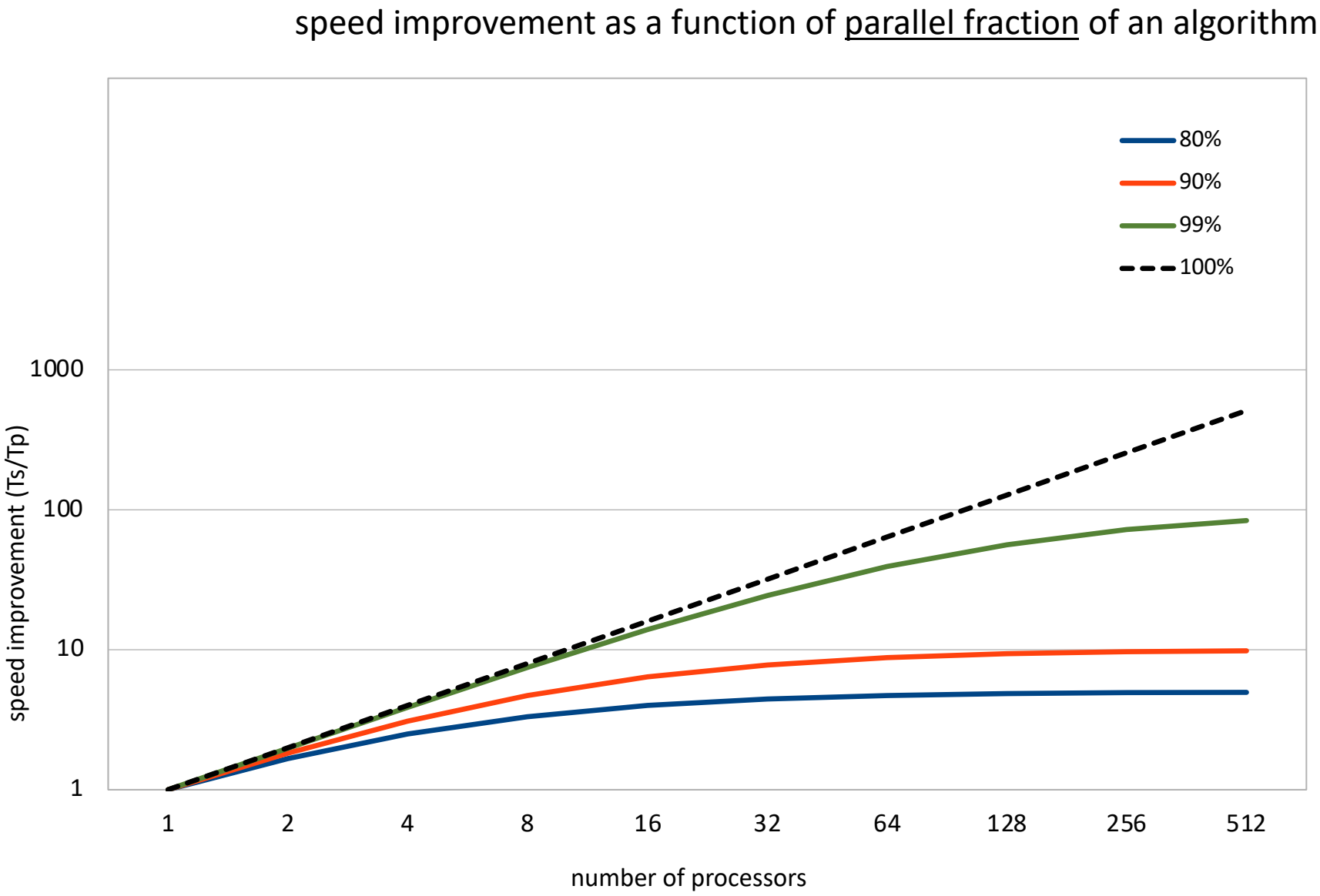
$T_s/T_p = 1/((T_{seq}+T_{par}/p)/(T_{seq}+T_{par}))$:= S (speed improvement)

$F_{seq} = T_{seq}/(T_{seq}+T_{par})$: fraction of time that can not be parallel
$F_{par} = T_{par}/(T_{seq}+T_{par})$: fraction of that that can be parallel

→ $S = 1/ (F_{seq} + F_{par}/p)$ where F_{seq} and F_{par} are fraction of serial and parallel parts of an algorithm

➔ $\lim_{p\to\infty} S \rightarrow \frac{1}{F_{seq}}$

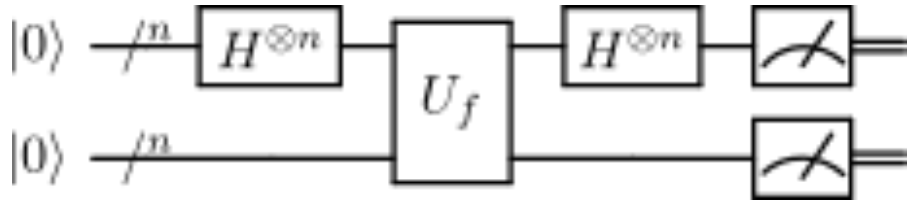
Amdhal's law


speed improvement as a function of parallel fraction of an algorithm
80%
90%
99%
100%
1000
100
10
1
speed improvement (Ts/Tp)
1
2
4
8
16
32
64
128
256
512
number of processors

An example: Simon's Periodicty Algoritm

- Finding a patterns in a function
- A combination of quantum and classical algorithms (post processing)

Step 1) (Quantum) apply the following circuit O(n) times (i.e. O(n) queries of U_f):



→ each time we obtain an equality

An example: Simon's Periodicty Algoritm

Step 2) (Classical) combined equalities define a linear system

→ solve for n unknowns (if the coefficient matrix is dense this would cost $O(n^3)$ operations for factorization and $O(n^2)$ operations for triangular solution)

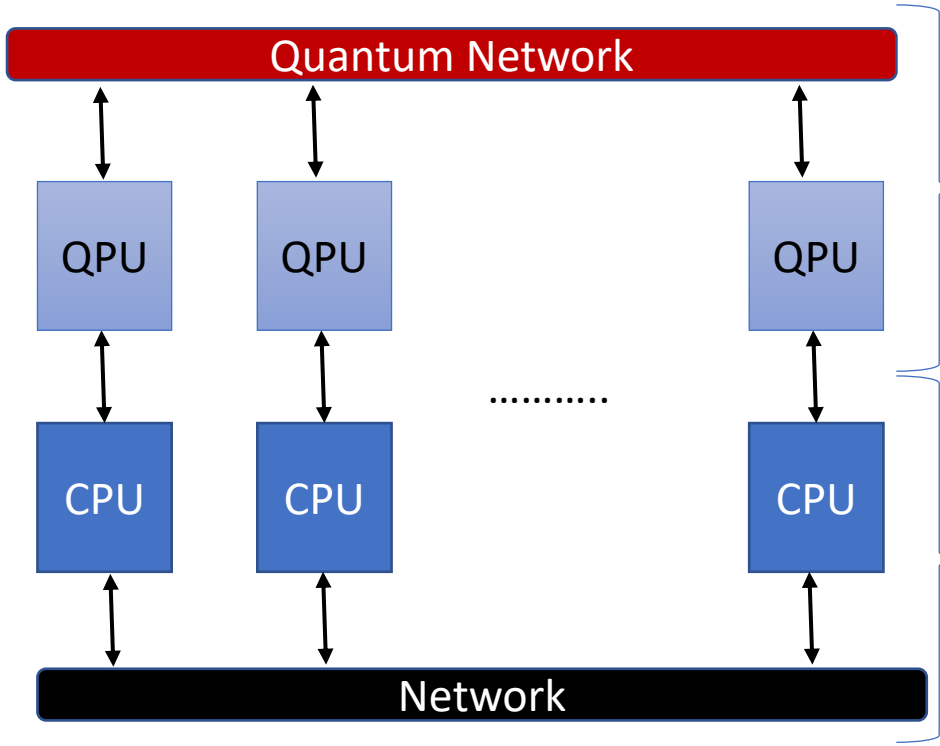
→ But the cost is likely to be lower since the coefficient matrix is sparse and this step can be "classically" parallelized

How can we adapt the algorithm for the quantum cluster architecture?

O(n) function evaluations

+

O(n^3/p) computation &
~O(n*logp) communication time [if dense, 1D block row partitioning with no pivoting]



Each node can obtain several rows of the linear system

Nodes collectively work on solving the linear system that is distributed by rows via a variant of parallel Gaussian elimination (with some communication)

Summary

- HPC and QC relationship
- Algorithms will need to be adapted to these enviroments
  - Some are more straightforward
  - Some will require a more detailed re-design of algorithms suitable for the new QACA

Thank you!