High performance computing for numerical methods and data analysis
MATH-505

Laura Grigori
EPFL and PSI

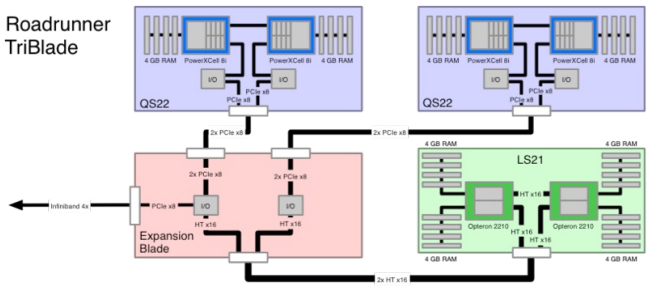
Plan

- Motivation for high performance computing
- Introduction to HPC
- Structure of the course

Evolution of high performance architectures

- Computers get faster, but their architecture gets more complex
  - From scalar (1970's), vector machines (1976 Cray 1), computers with thousands of processors (1990's, Intel paragon), distributed memory massively parallel machines (2000's)
  - To multi-core processors, accelerators, heterogeneous architectures

- First petascale system 2008, 1.33 Pflop/s
  - RoadRunner, IBM, LANL
  - A TriBlade formed by
    - Two dual-core Opterons with 16 GB of memory
    - Four PowerXCell 8i CPUs with 16 GB Cell RAM
  - A total of 13,824 Opteron cores + 116,640 Cell cores



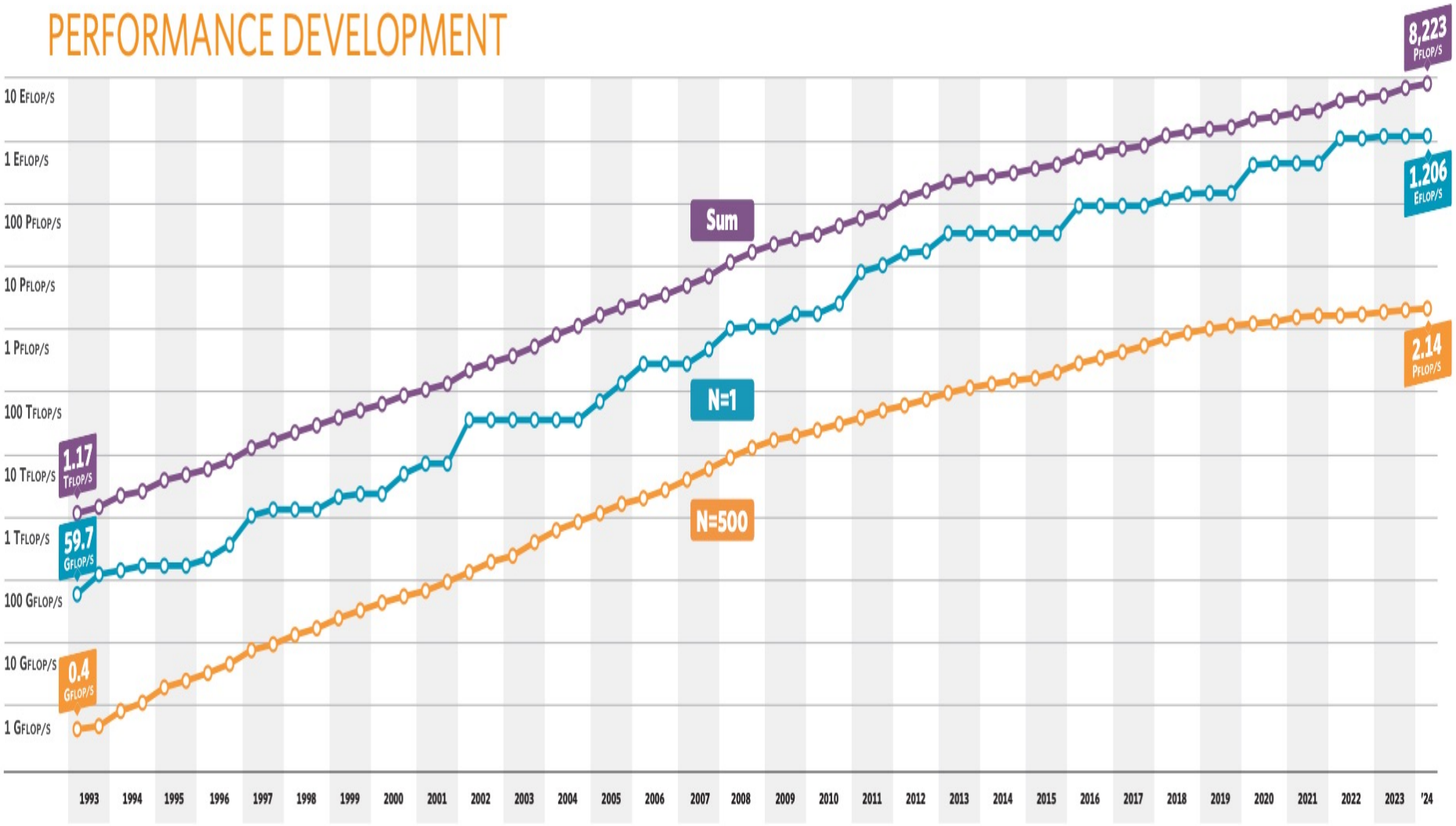
The TOP5 of the Top500, November 2023

#	NOVEMBER 2023	Manufacturer	Computer	Country	Cores	Rmax [Pflops]	Power [MW]
1	Oak Ridge National Laboratory	HPE	Frontier HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	USA	8,730,112	1,102	21.1
2	Argonne National Laboratory	HPE	Aurora* HPE Cray EX Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11	USA	4,742,808	585.3	24.6
3	Microsoft Azure	Microsoft	Eagle Microsoft NDv5 Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR	USA	1,123,200	561.2	
4	RIKEN Center for Computational Science	Fujitsu	Fugaku Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
5	EuroHPC / CSC	HPE	LUMI HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,069,760	309.1	6.0
6	EuroHPC / CINECA	Atos	Leonardo Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,463,616	174.7	5.6
7	Oak Ridge National Laboratory	IBM	Summit IBM Power System, P9 22C 3.07GHz, Mellanox EDR, NVIDIA GV100	USA	2,414,592	148.6	10.1
8	EuroHPC/BSC	EVIDEN	MareNostrum 5 ACC BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200	Spain	680,960	138.20	2.5
9	NVIDIA Corporation	NVIDIA	Eos NVIDIA DGX SuperPOD NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400	USA	485,888	121.40	
10	Lawrence Livermore National Laboratory	IBM	Sierra IBM Power System, P9 22C 3.1GHz, Mellanox EDR, NVIDIA GV100	USA	1,572,480	94.6	7.4

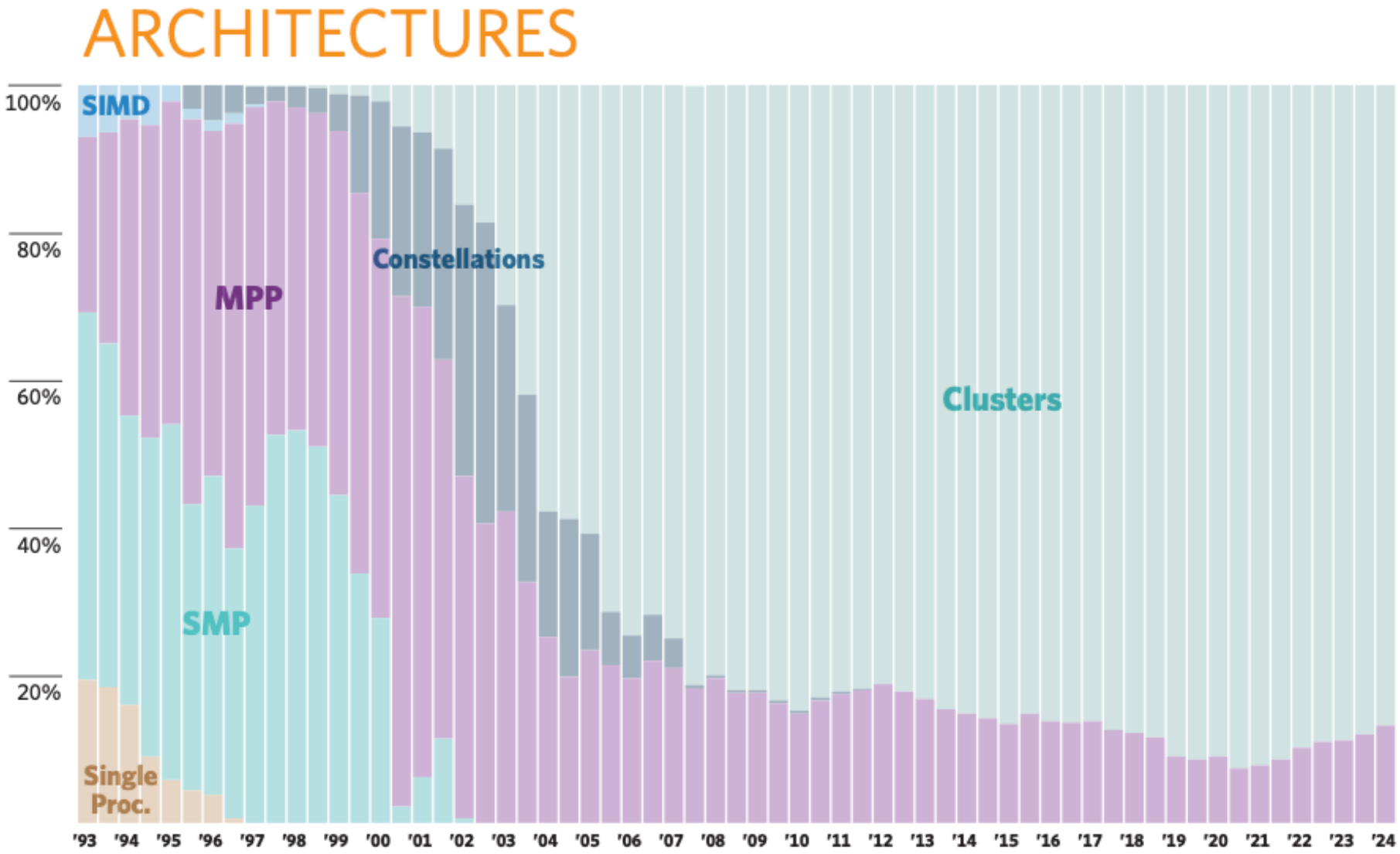
The TOP5 of the Top500, June 2024

#	NOVEMBER 2023	Manufacturer	Computer	Country	Cores	Rmax [Pflops]	Power [MW]
1	Oak Ridge National Laboratory	HPE	Frontier HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	USA	8,730,112	1,102	21.1
2	Argonne National Laboratory	HPE	Aurora* HPE Cray EX Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11	USA	4,742,808	585.3	24.6
3	Microsoft Azure	Microsoft	Eagle Microsoft NDv5 Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR	USA	1,123,200	561.2	
4	RIKEN Center for Computational Science	Fujitsu	Fugaku Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
5	EuroHPC / CSC	HPE	LUMI HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,069,760	309.1	6.0
6	CSCS	HPE	ALPS HPE Cray EX254n, NVIDIA Grace 3.1GHz, Slingshot-11	Switzerland	1,305,600	270.0	5.1
7	EuroHPC / CINECA	Atos	Leonardo Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,463,616	174.7	5.6
8	EuroHPC/BSC	EVIDEN	MareNostrum 5 ACC BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200	Spain	680,960	138.20	2.5
9	Oak Ridge National Laboratory	IBM	Summit IBM Power System, P9 22C 3.07GHz, Mellanox EDR, NVIDIA GV100	USA	2,414,592	148.6	10.1
10	NVIDIA Corporation	NVIDIA	Eos NVIDIA DGX SuperPOD NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400	USA	485,888	121.40	

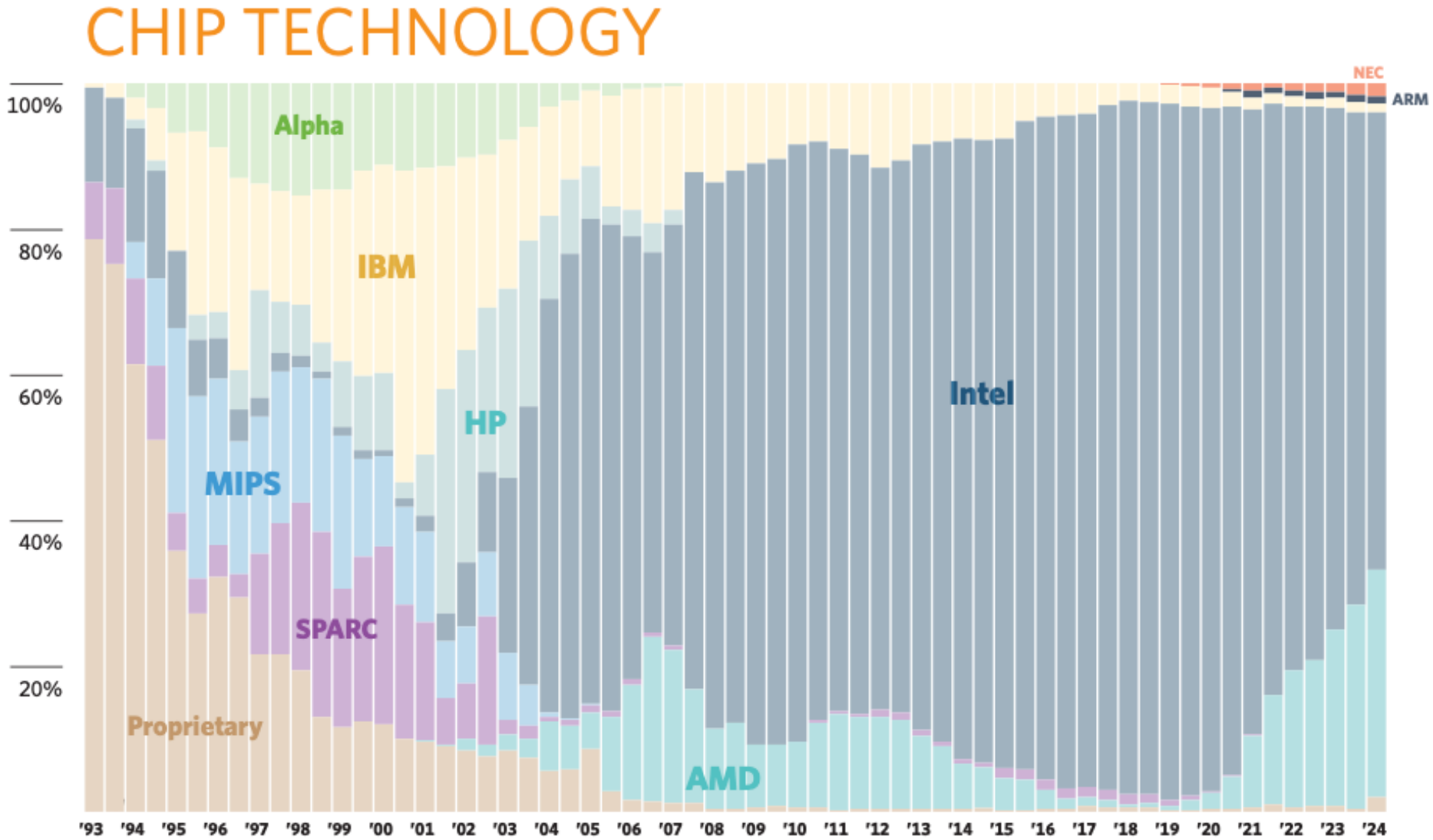
Top500, Nov 2024



Top500, Nov 2024

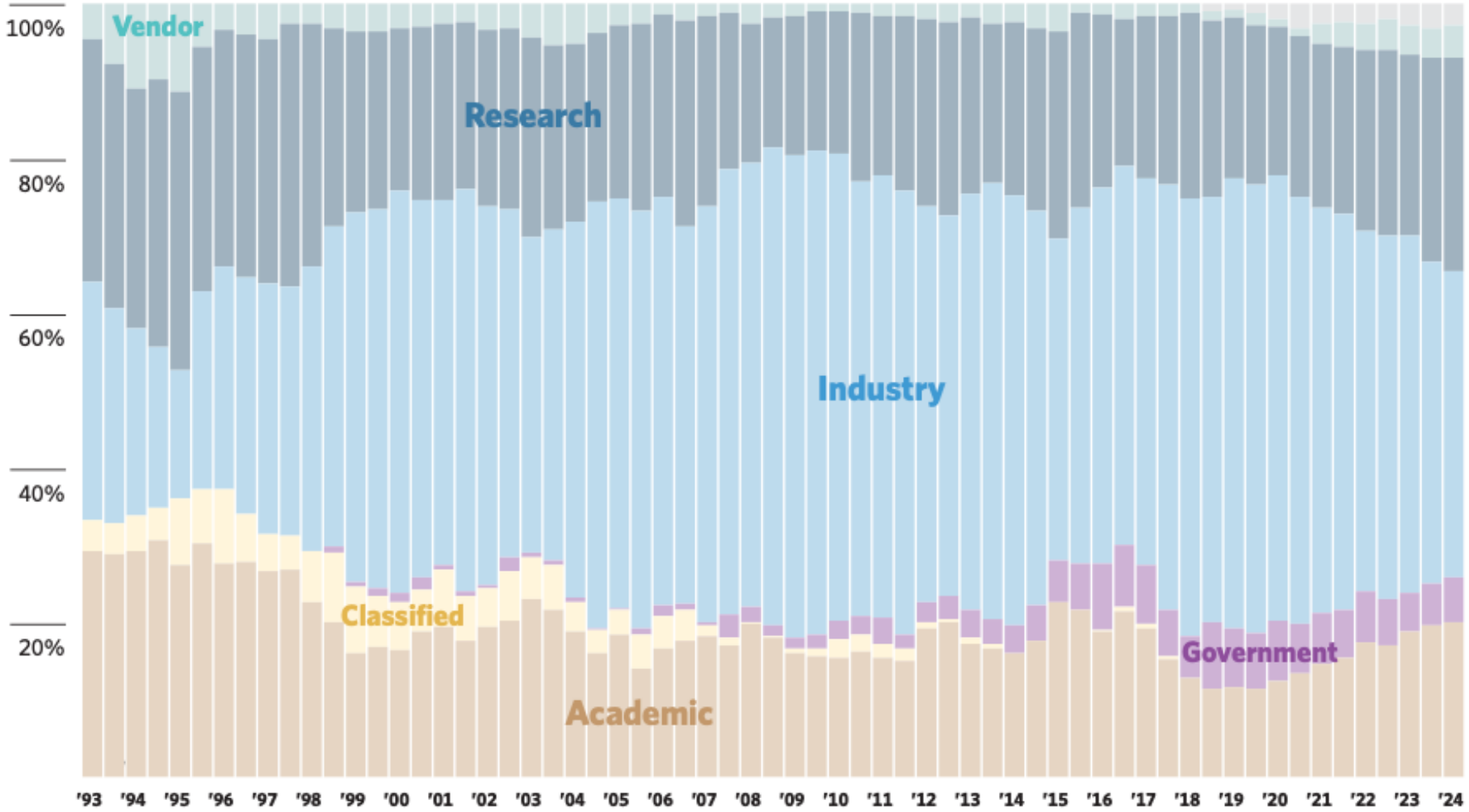


Top500, Nov 2024



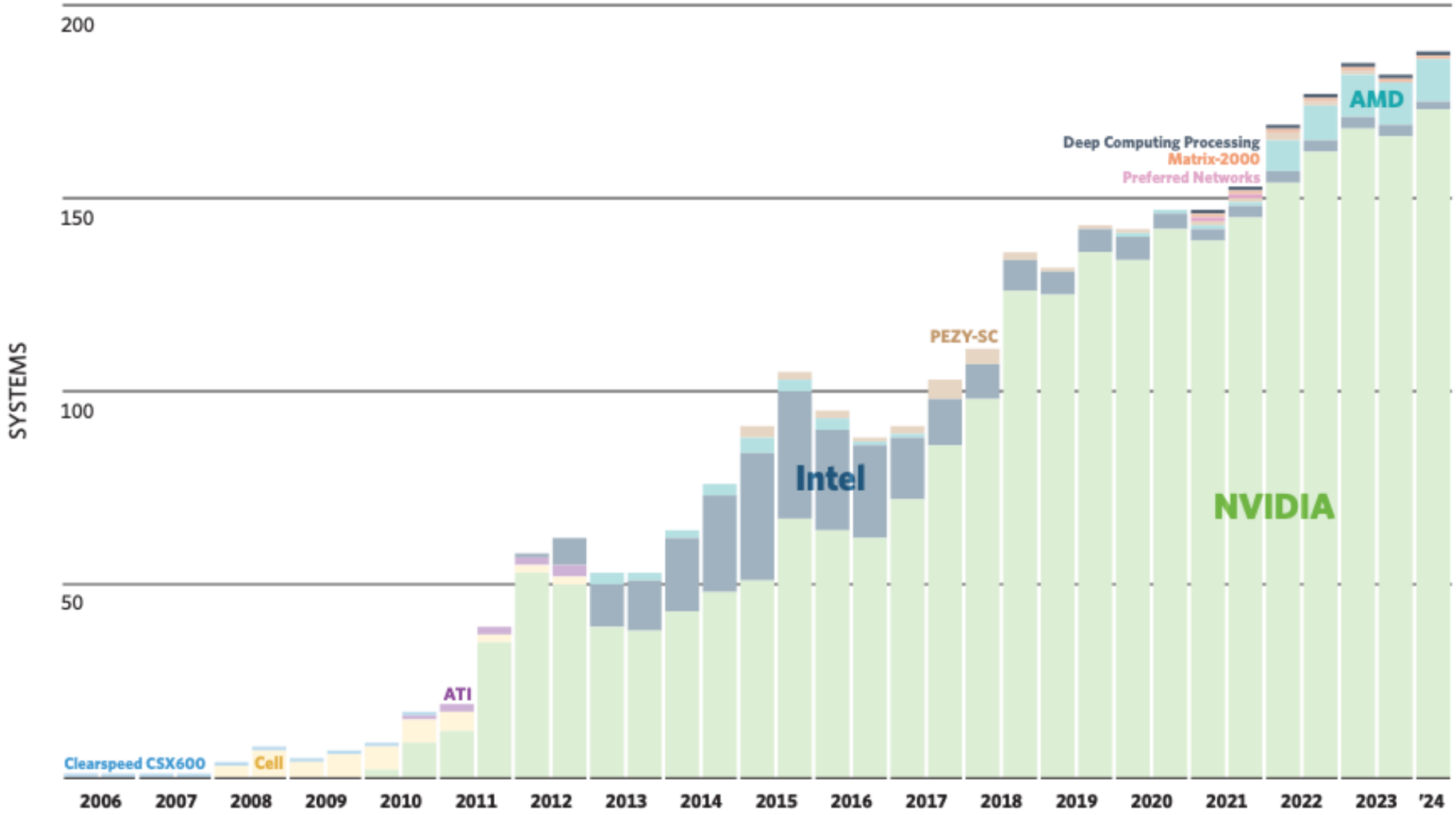
Top500, Nov 2024

INSTALLATION TYPE



Top500, Nov 2024

ACCELERATORS/CO-PROCESSORS



www.top500.org

Frontier (#1) System Overview

System Performance

- Peak performance of 1.6 double precision exaFLOPS Measured Top500 performance (Rmax) was 1.102 exaFLOPS

Each node has

- 3rd Gen AMD EPYC CPU with 64 cores
- 4 Purpose Built AMD Instinct 250X GPUs
- 4X128 GB of fast memory, 1 per GPU
- 5 terabytes of flash memory

The system includes

- 9,472 nodes
- Slingshot interconnect

Fugaku (#4) System Overview

System Performance

- Peak performance of 442 petaflops (per TOP500 Rmax),
- 2.0 EFLOPS on a different mixed-precision benchmark

Each node has

- Fujitsu A64FX CPU (48+4 cores) per node
- HBM2 32 GiB

The system includes

- 158,976 nodes
- Custom Tofu Interconnect D
- 1.6 TB NVMe SSD/16 nodes (L1)
- 150 PB Lustre Filesystem (L2)
- Cloud storage (L 3)

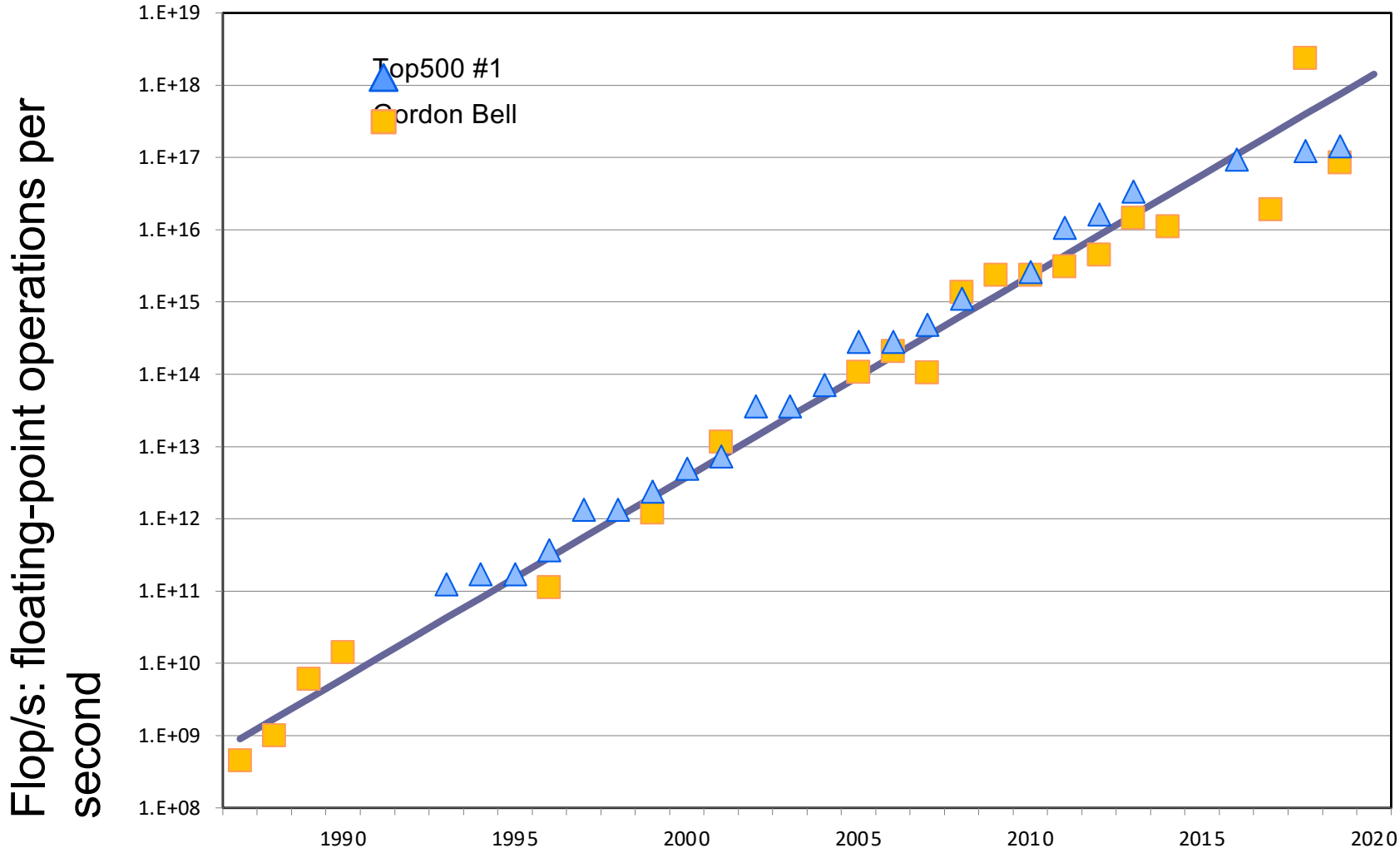
RIKEN Center for Computational Science (R-CCS)

Gordon Bell Prizes: Science at Scale

Established in 1987 with a cash award of $10,000 (since 2011), funded by Gordon Bell, a pioneer in HPC.
For innovation in applying *HPC to applications in science, engineering, and data analytics*.

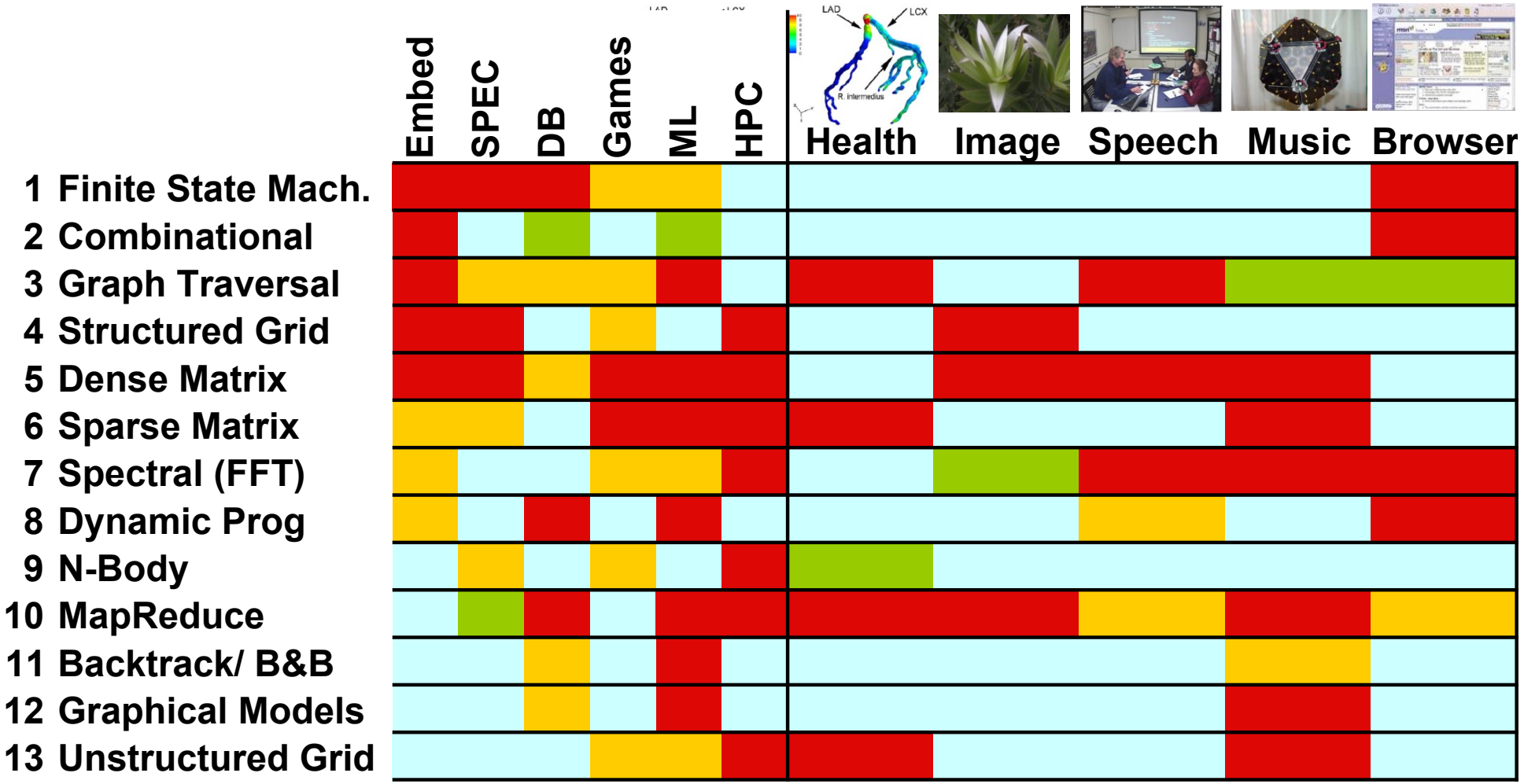
Slide from CS267 Lecture, UC Berkeley

Gordon Bell prizes vs Top 500



Slide from CS267 Lecture, UC Berkeley

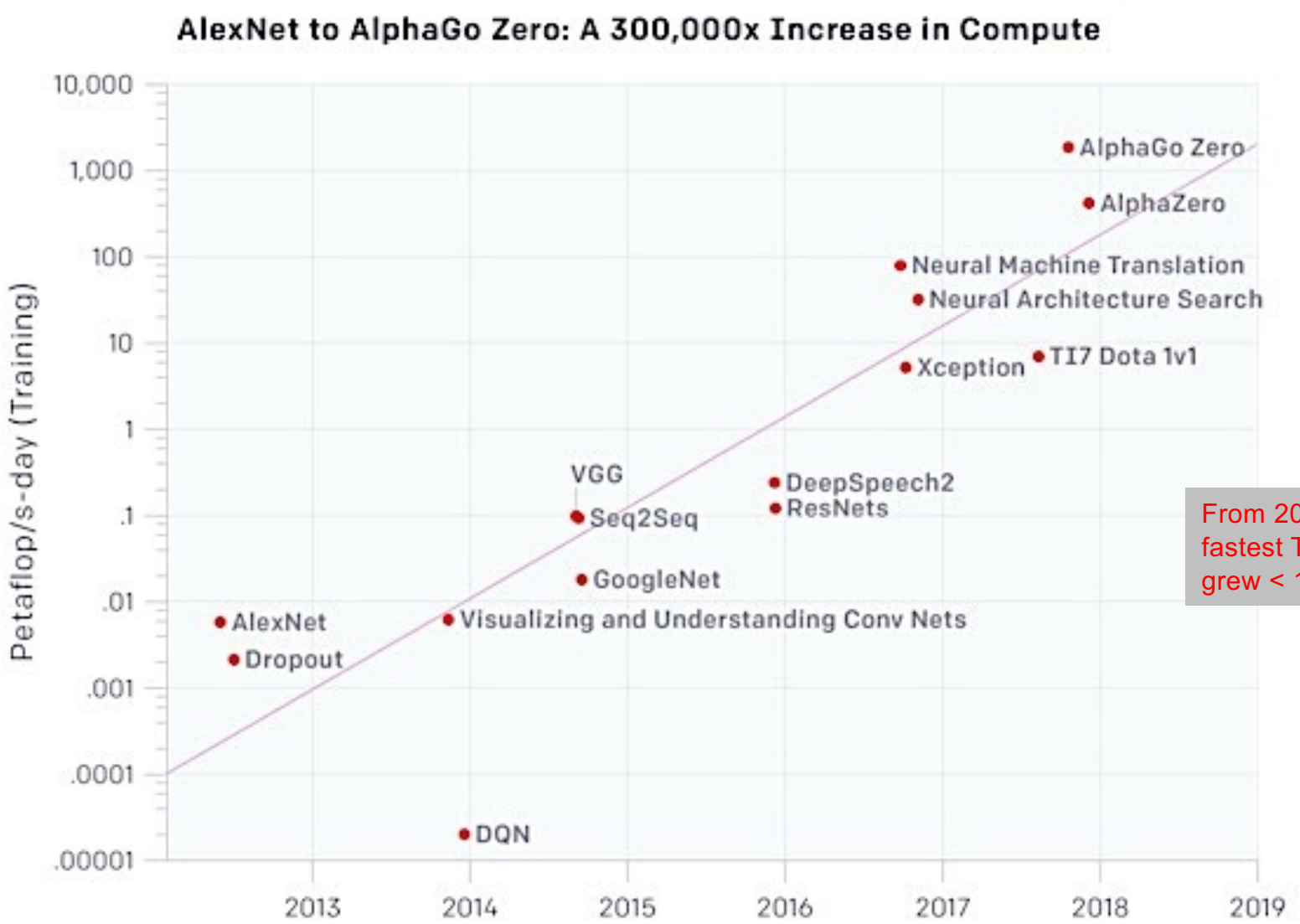
Motif/dwarf – common computational patterns



Red Hot -> Blue cool

Source slide : J. Demmel CS267 UC Berkeley

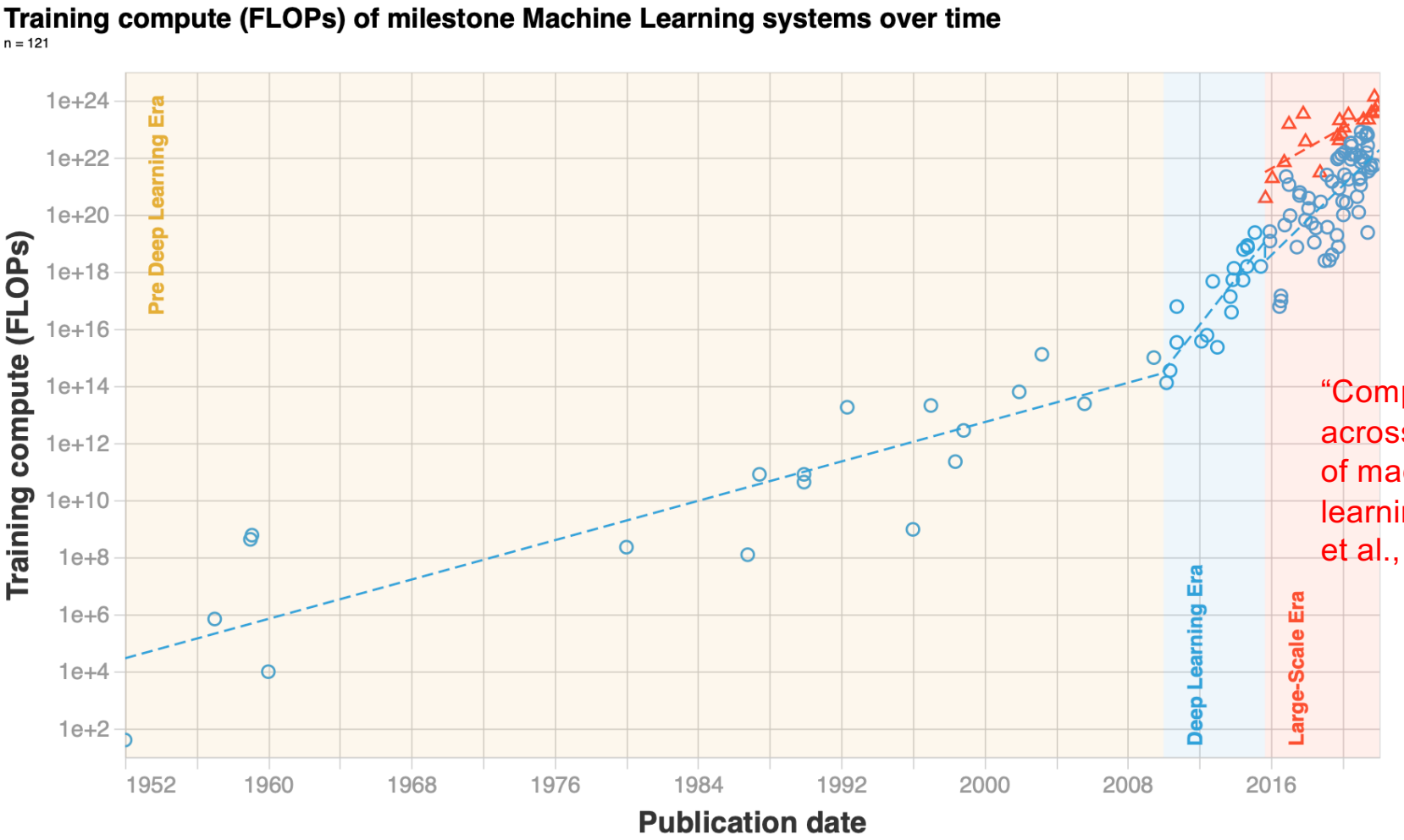
Machine learning demands



From 2011-2017 the fastest Top500 machine grew < 10x

Slide from CS267 Lecture, UC Berkeley

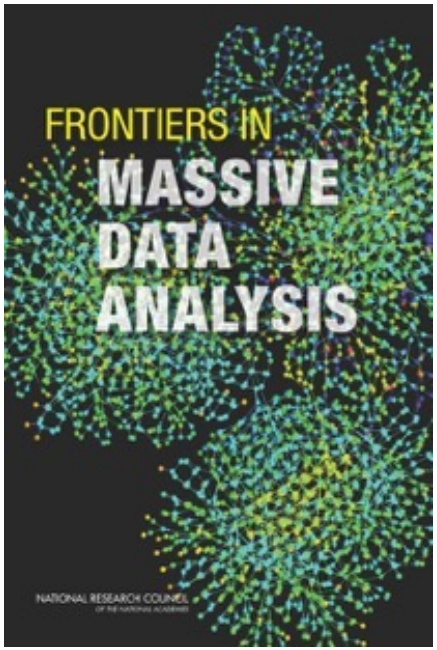
Machine learning



"Compute trends across three eras of machine learning", Sevilla et al., 2022

Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

Analytics vs. Simulation Motifs



National Academies 2013

7 Giants of Data	7 Dwarfs of Simulation
Basic statistics	Monte Carlo methods
Generalized N-Body	Particle methods
Graph-theory	Unstructured meshes
Linear algebra	Dense Linear Algebra
Optimizations	Sparse Linear Algebra
Integrations	Spectral methods
Alignment	Structured Meshes

Moore's law reinterpreted

- **Number of cores per chip can double every two years**
- **Clock speed will not increase (possibly decrease)**
- **Need to deal with systems with millions of concurrent threads**
- **Need to deal with inter-chip parallelism as well as intra-chip parallelism**

Slide source: J. Demmel, CS 267 UC Berkeley

Some Particularly Challenging Computations

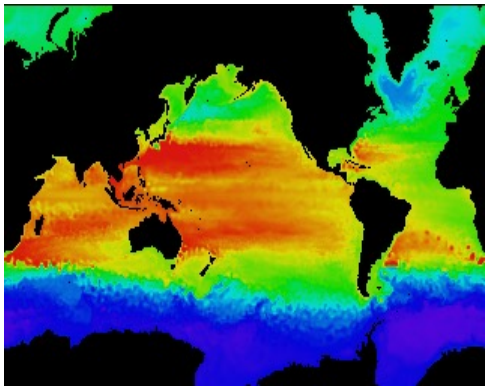
- **Science**
  - Global climate modeling
  - Biology: genomics; protein folding; drug design
  - Astrophysical modeling
  - Computational Chemistry
  - Computational Material Sciences and Nanosciences
- **Engineering**
  - Semiconductor design
  - Earthquake and structural modeling
  - Computation fluid dynamics (airplane design)
  - Combustion (engine design)
  - Crash simulation
- **Business**
  - Financial and economic modeling
  - Transaction processing, web services and search engines
- **Defense**
  - Nuclear weapons -- test by simulations
  - Cryptography

Slide source: J. Demmel, CS 267 UC Berkeley

Data driven science

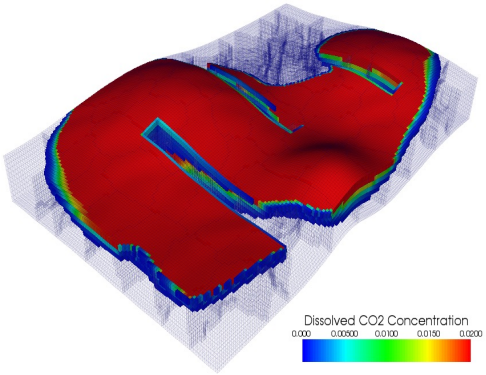
Numerical simulations require increasingly computing power as data sets grow exponentially

Climate modeling



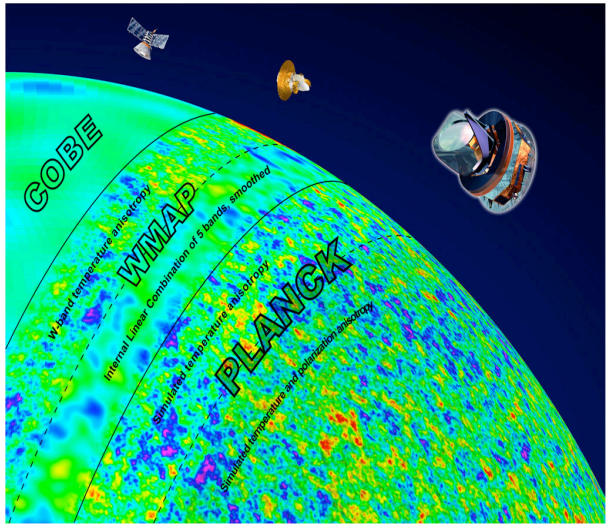
http://www.epm.ornl.gov/chammp/chammp.html

CO2 Underground storage



Source: T. Guignon, IFPEN

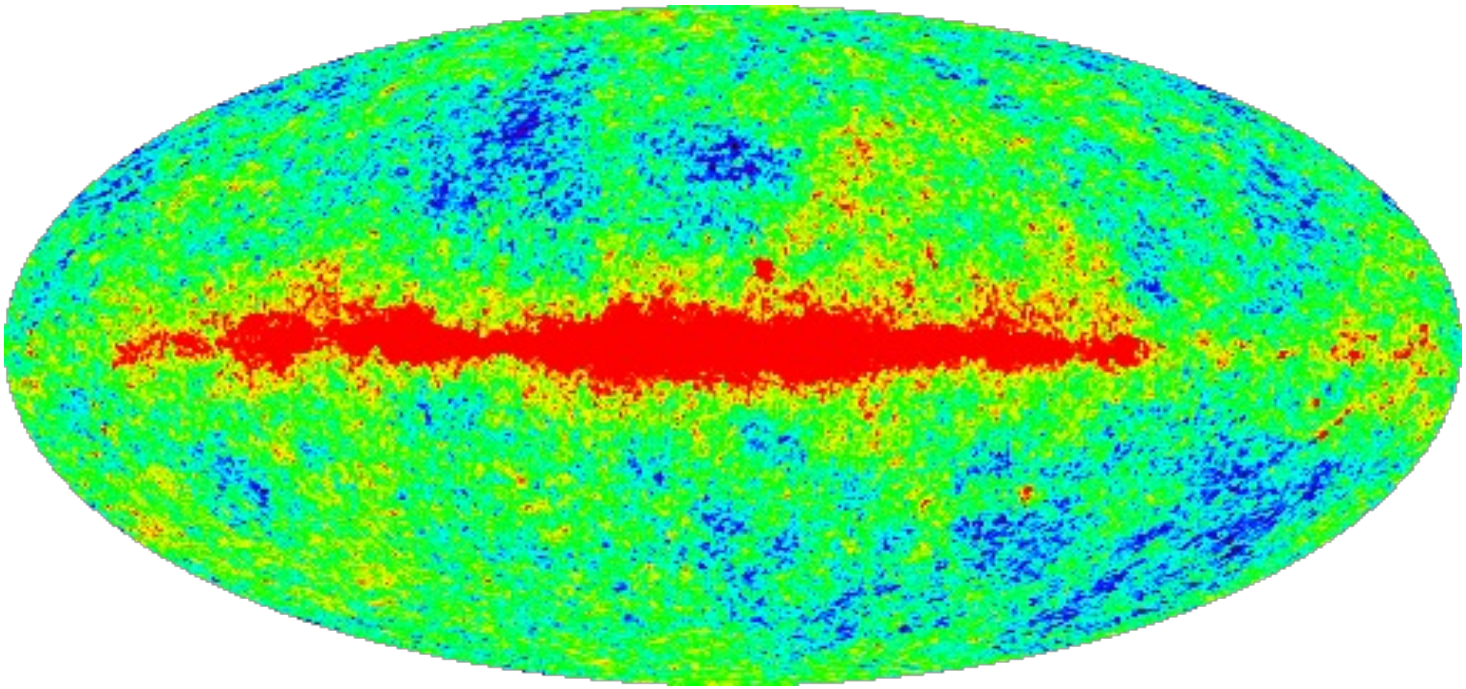
Astrophysics: CMB data analysis



http://www.scidacreview.org/0704/html/cmb.html

CMB data analysis

- Light left over after the ever mysterious «Big Bang»,
  - overall very isotropic and uniform,
  - but small - 1 part in 10^5 - anisotropies are hidden in there …
  - even smaller - 1 part in 10^6 or 10^7 - are the goal of current experiments.



- Always in need of more data
- Data sets are growing at Moore's rate

CMB data analysis in an (algebraic) nutshell

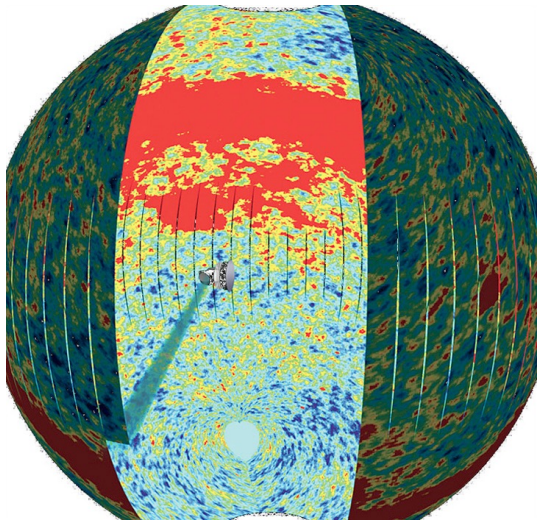
- CMB DA is a juxtaposition of the same algebraic operations
- Map-making problem
  - Find the best map x from observations d, scanning strategy A, and noise n_t

$$d = Ax + n_t$$

  - Assuming the noise properties are Gaussian and piece-wise stationary, the covariance matrix is $N = \langle n_t n_t^T \rangle$, and N^{-1} is a block diagonal symmetric Toeplitz matrix.
  - The solution of the generalized least squares problem is found by solving

$$A^T N^{-1} A x = A^T N^{-1} d$$

- Spherical harmonic transform (SHT)
  - Synthesize a sky image from its harmonic representation



- What is difficult about the CMB DA then ?

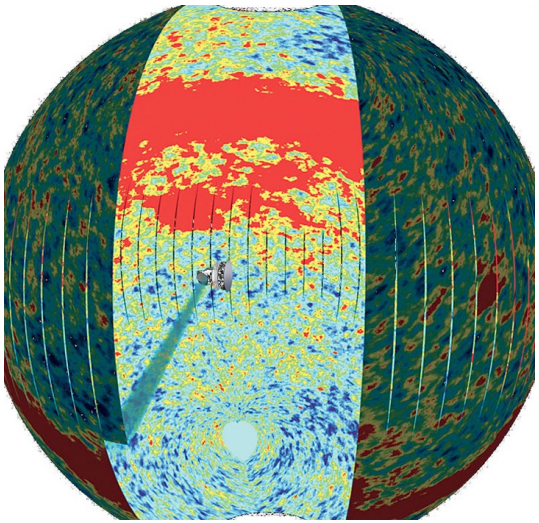
  Well, the data is BIG !

Data driven science

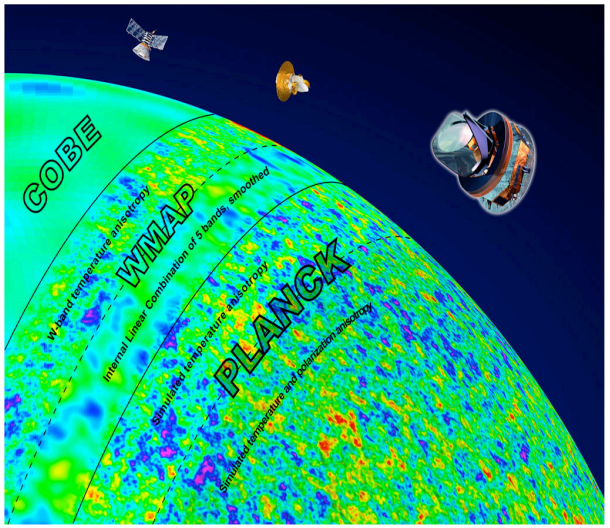
Figures from astrophysics:

- Produce and analyze multi-frequency 2D images of the universe when it was 5% of its current age.
- COBE (1989) collected 10 gigabytes of data, required 1 Teraflop per image analysis.
- PLANCK (2010) produced 1 terabyte of data, requires 100 Petaflops per image analysis.

Source: J. Borrill, LBNL, R. Stompor, Paris 7



Astrophysics: CMB data analysis



http://www.scidacreview.org/0704/html/cmb.html

CMB data analysis in an (algebraic) nutshell

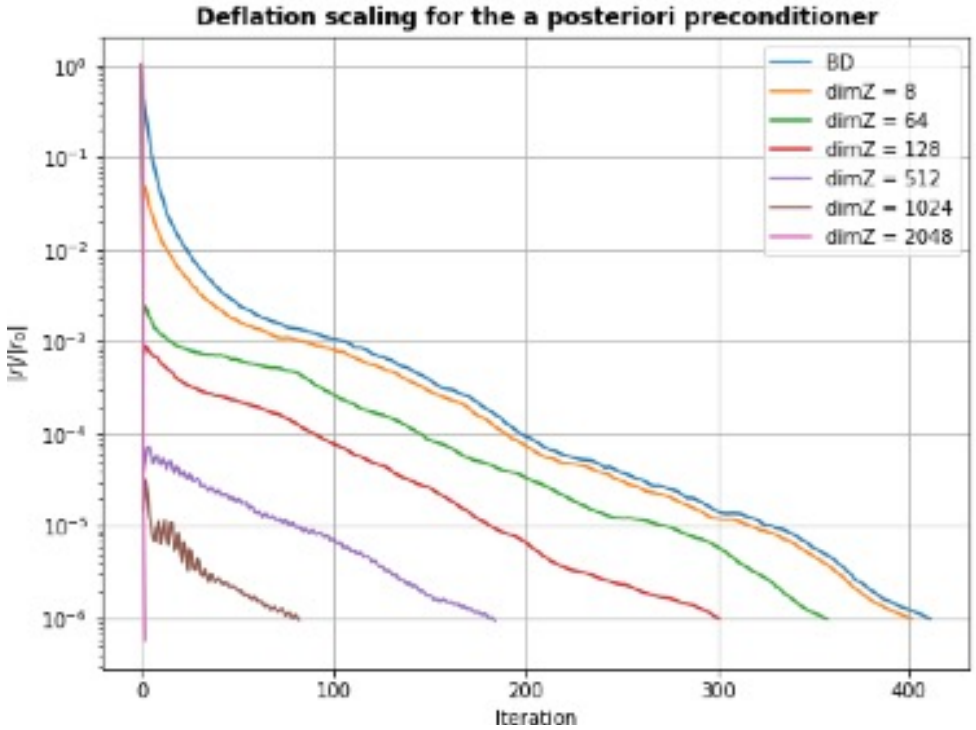
- Map-making problem
  - Find the best map x from observations d, scanning strategy A, and noise n_t

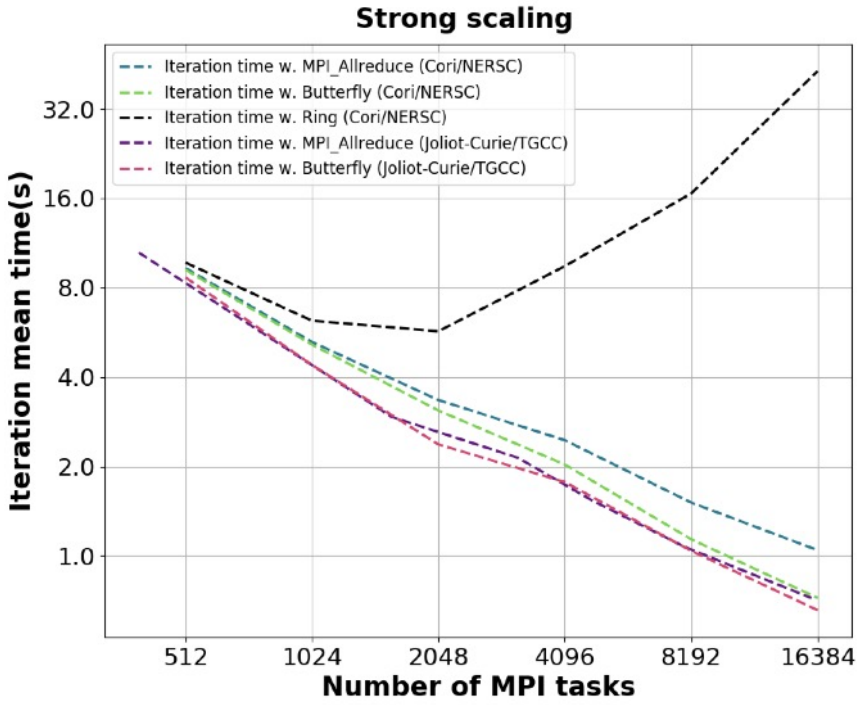
$$d = Ax + n_t$$

  - The solution of the generalized least squares problem is found by solving

$$A^T N^{-1} A x = A^T N^{-1} d$$

  - with CG, deflation techniques and optimized communication routines [Bouhargani et al., 2021].
  - $O(10^{11})$ time samples per map, $O(10^5)$ pixels per map

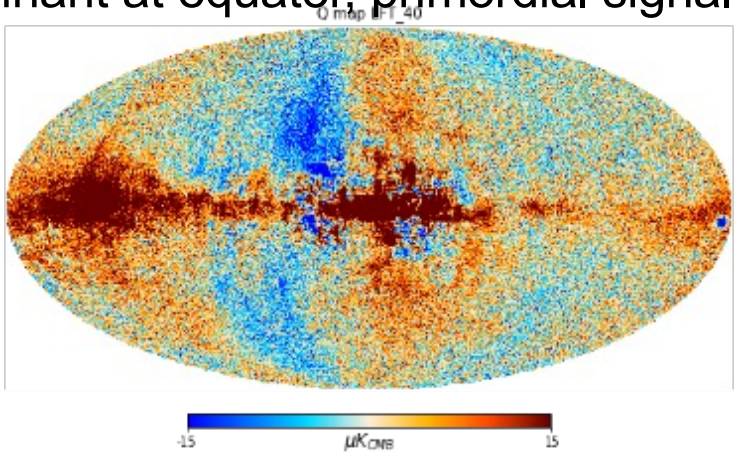


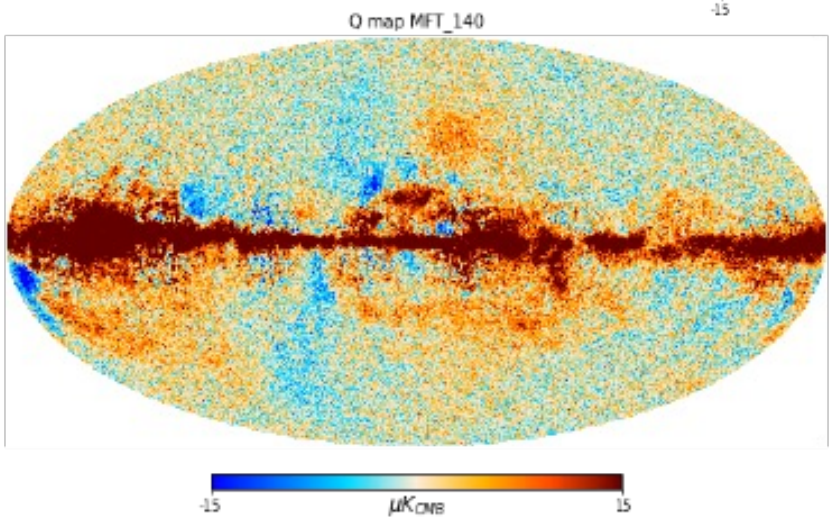


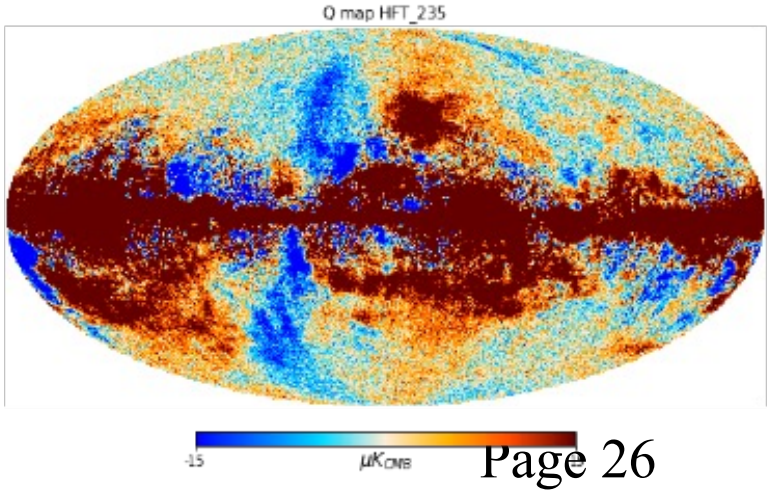
CMB data analysis – LiteBird experiment

Satellite based experiment 2029-2032 - analyzing 100 Tbytes of data

- First analysis and maps of full volume of data (all detectors and full length of the mission) as expected from LiteBIRD
- Polarization maps (Q Stokes parameter) in 40, 140, and 235 GHz frequency bands
- Galactic signal predominant at equator, primordial signal predominant at the poles







Molecular dynamics

Study time evolution of complex molecular processes with atomic-scale space-time resolution (solvated proteins, viruses, …)

- virtual microscope
- predict properties of new molecules

Consider molecular mechanics polarizable force fields

- Represent both intramolecular (chemical bonds) and intermolecular interactions (Lennard-Jones potential, Coulomb potential)

Evaluate electrostatic energy with pairwise interactions $O(N^2)$ -> $O(N \log N)$

- Separation of short range/long range interactions (relies on FFT)
- Compute the polarization energy requires solving

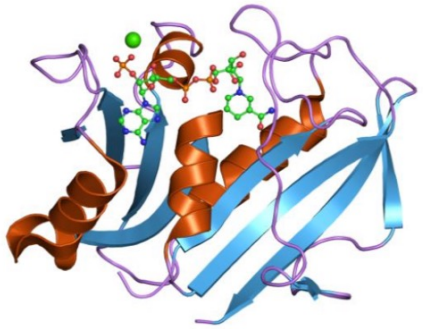
$$T\ x = E$$

T polarization matrix, E electric field produced by permanent density of charge at polarizable sites

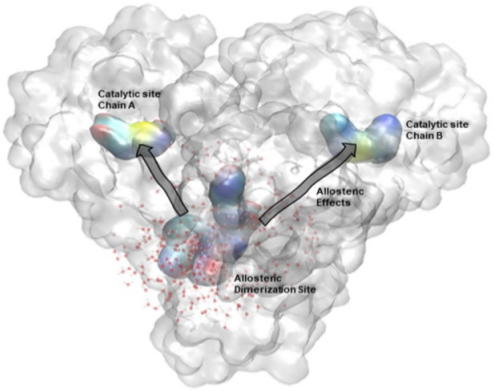
dimension of system 3N x 3N, with N number of atoms

Performance evaluation with Tinker-hp

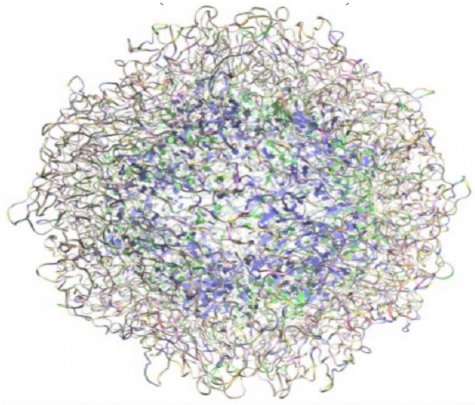
8 systems considered



DHFR (solvated), 23558 atoms

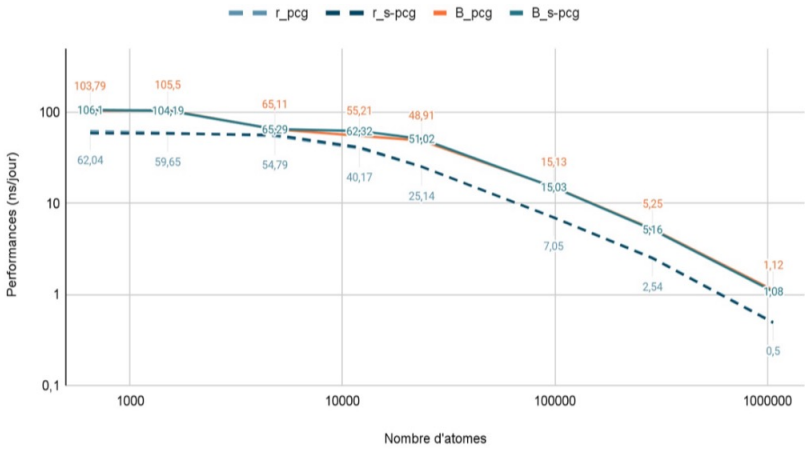


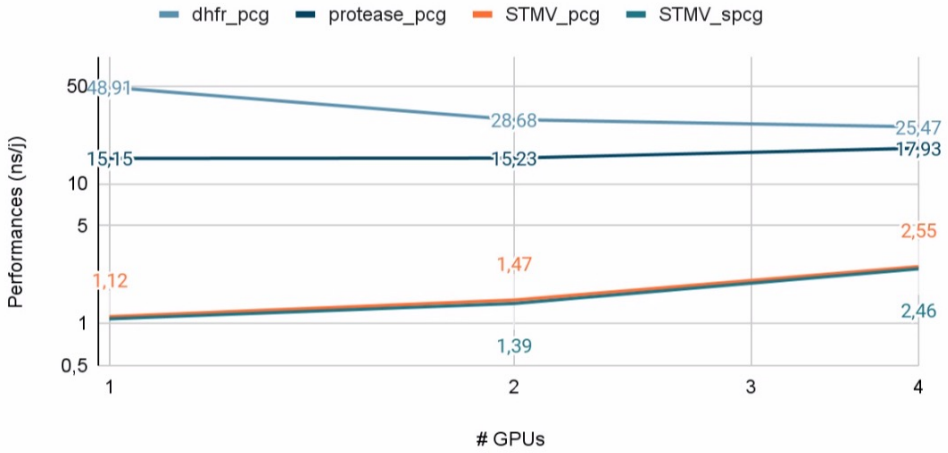
Global vue of dimer of **main protease** - protein of **SARS-CoV-2** (98695 atoms)



Satellite Tobacco Mosaic Virus STMV (solvated), 1006624 atoms (X43)

- 5000 time steps, 2 integrators: RESPA + 2fs (10 ps), BAOAB-RESPA1 + 10fs (50 ps)
- Tests on Jean Zay (HPE SGI, 1 node with 4 GPUs V100)
- Preconditioned CG versus preconditioned CA (s-step) CG

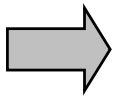




with O. Adjoua, L. Lagard ere, J. P. Piquemal (EMC2 ERC project)

Principles of Parallel Computing

- Finding enough parallelism (Amdahl's Law)
- Granularity – how big should each parallel task be
- Locality – moving data costs more than arithmetic
- Load balance – don't want 1K processors to wait for one slow one
- Coordination and synchronization – sharing data safely
- Performance modeling/debugging/tuning



All of these things makes parallel programming even harder than sequential programming.

Finding Enough Parallelism

- Suppose only part of an application seems parallel
- Amdahl's law
  - let s be the fraction of work done sequentially, so (1-s) is fraction parallelizable
  - P = number of processors

$$\text{Speedup}(P) = \text{Time}(1)/\text{Time}(P)$$

$$<= 1/(s + (1-s)/P)$$

$$<= 1/s$$

- Even if the parallel part speeds up perfectly performance is limited by the sequential part

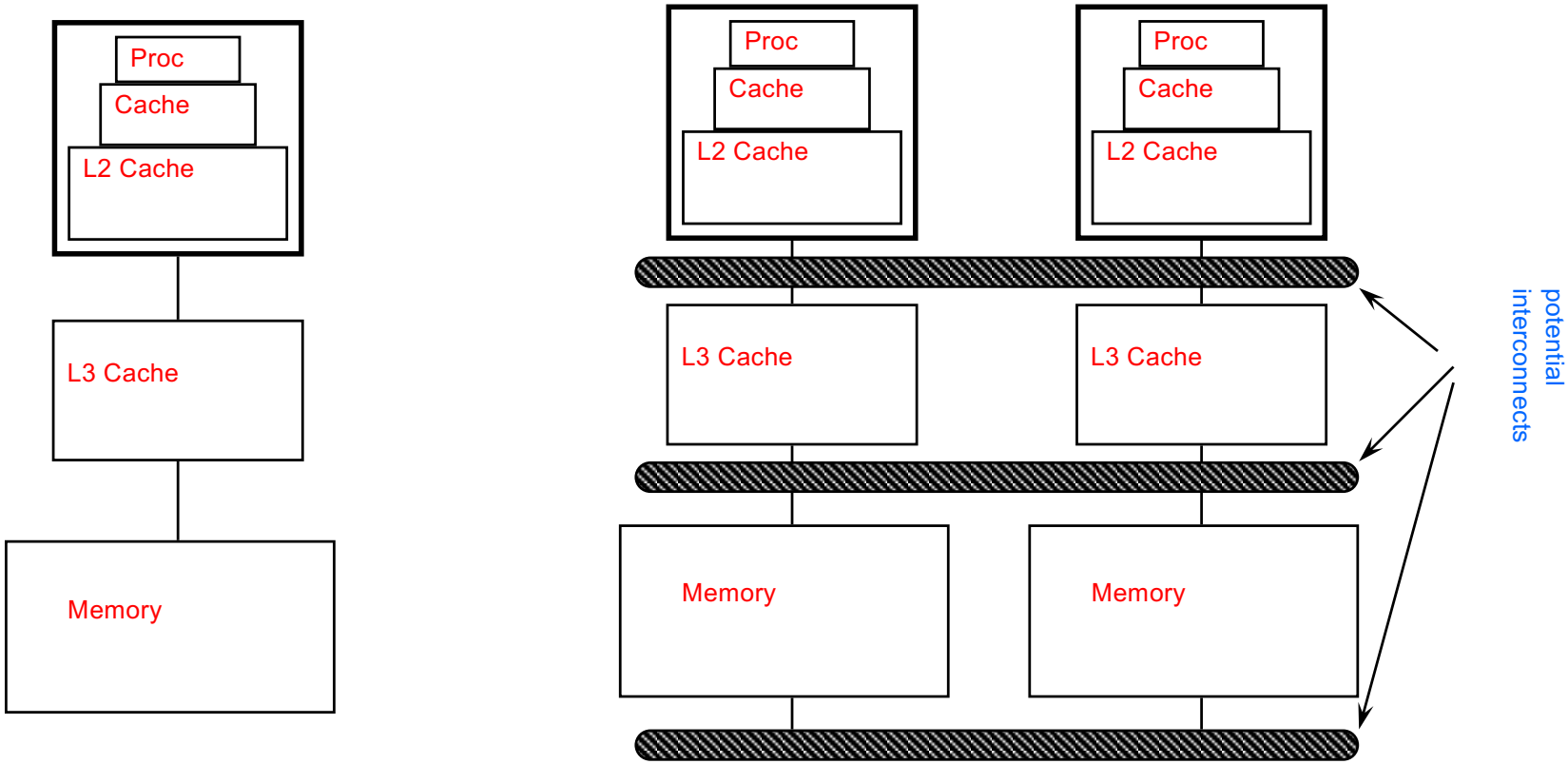
Overhead of Parallelism

- Given enough parallel work, this is the biggest barrier to getting desired speedup
- Parallelism overheads include:
  - cost of starting a thread or process
  - cost of communicating shared data
  - cost of synchronizing
  - extra (redundant) computation
- Each of these can be in the range of milliseconds (=millions of flops) on some systems
- Tradeoff: Algorithm needs sufficiently large units of work to run fast in parallel (i.e. large granularity), but not so large that there is not enough parallel work

Slide source: J. Demmel, CS 267 UC Berkeley

Locality and Parallelism

Conventional Storage Hierarchy



- Large memories are slow, fast memories are small
- Storage hierarchies are large and fast on average
- Parallel processors, collectively, have large, fast cache
  - the slow accesses to "remote" data we call "communication"
- Algorithm should do most work on local data

Slide source: J. Demmel, CS 267 UC Berkeley

Structure of the course

- Introduction to high performance computing
- Overview of state-of-the-art parallel architectures and MPI programming technique
- Factorization methods and communication avoiding algorithms
- Randomization for solving large scale problems
- Low rank matrix approximation algorithms, deterministic and randomized approaches
- Krylov subspace iterative solvers, deterministic and randomized approaches
- Applications to data science

Exercises:

- Python and MPI
- Usage of Scitas cluster (CPUs)
- Information provided next week

Sources

Moodle

- https://moodle.epfl.ch/course/MATH-505
- All relevant information will be available in the moodle
- Slides of the lecture will be available before the class

Questions:

- During the lectures (encouraged)
- In the forum
- Or directly by email

Grading

Grading on 2 projects and one quiz:

Project 1: orthogonalization techniques

- To be completed individually
- Subject provided in week 3
- Weight for the grade: 0.3
- Deadline to submit the report: week 8, November 5, 2024, 11:59PM CEST

Project 2: randomized Nystrom for low rank approximation

- To be done in groups of two
- Subject provided in week 7
- Weight for the grade - report: 0.3
- Weight for the grade - oral exam: 0.3
- Oral exam (individual, 5 mins with 3 slides to present the project + Q&A): during the January exam session
- Deadline to submit the report (one per group) + slides (individual): January 6, 2025, 11:59PM CEST

Quiz:

- Questions taken from the last lectures on Krylov subspace methods
- Weight for the grade: 0.1
- Quiz taken during week 14

Grading

Questions:

- All questions about projects/quiz/grading should be addressed to the professor

Generative AI: read the information on moodle

- Usage of Large Language Models as chatGPT, needs to be acknowledged in the report.
- Explain the usage, as improving the english and formulation, generating some code (specify which algorithms and if this concerns their initial version), debugging the code, generating figures (specify which figures), generating text (specify which parts).