

Parallel Computing (CS 633)

January 6, 2025

Preeti Malakar
pmalakar@cse.iitk.ac.in

Logistics

- Class hours: MW 3:30 – 5:00 PM (L16)
- Office hour: W 5:00 – 6:00 PM (KD 221)
- <https://www.cse.iitk.ac.in/users/cs633/2024-25-2>
 - Lectures will be uploaded after every class
- Extra class/quiz/doubts: Saturday 11 AM – 12 PM
- Announcements/uploads on
 - MookIT
 - Course email alias
- Email to the instructor should always be prefixed with **[CS633]** in the subject

Switch OFF All Devices



October 4, 2010

Use of Mobile Phones in Academic Area

The following policy will be followed regarding the usage of Mobile Phones in the Academic Area:

Examinations: Students are not permitted to carry Mobile phones inside the examination hall. The faculty members/invigilators must keep the mobile phones switched off during the conduct of the examination.

Classrooms: Mobile phones are to be switched off in class-rooms both, by students as well as Instructors/Tutors.

Laboratory/Library/Auditorium: Mobile phones are to be kept in silent mode in laboratories/library /auditorium. In case the individual would like to receive/make a call, they must do so from outside the premises.

The implementation of the above will be overseen by the following:

Examinations: Instructors/Invigilators

Classrooms: Instructors/Tutors

Laboratory: Instructors/Tutors/Officer-in-charge of the Laboratory

Library: Librarian

Auditorium: Facility-in-charge.

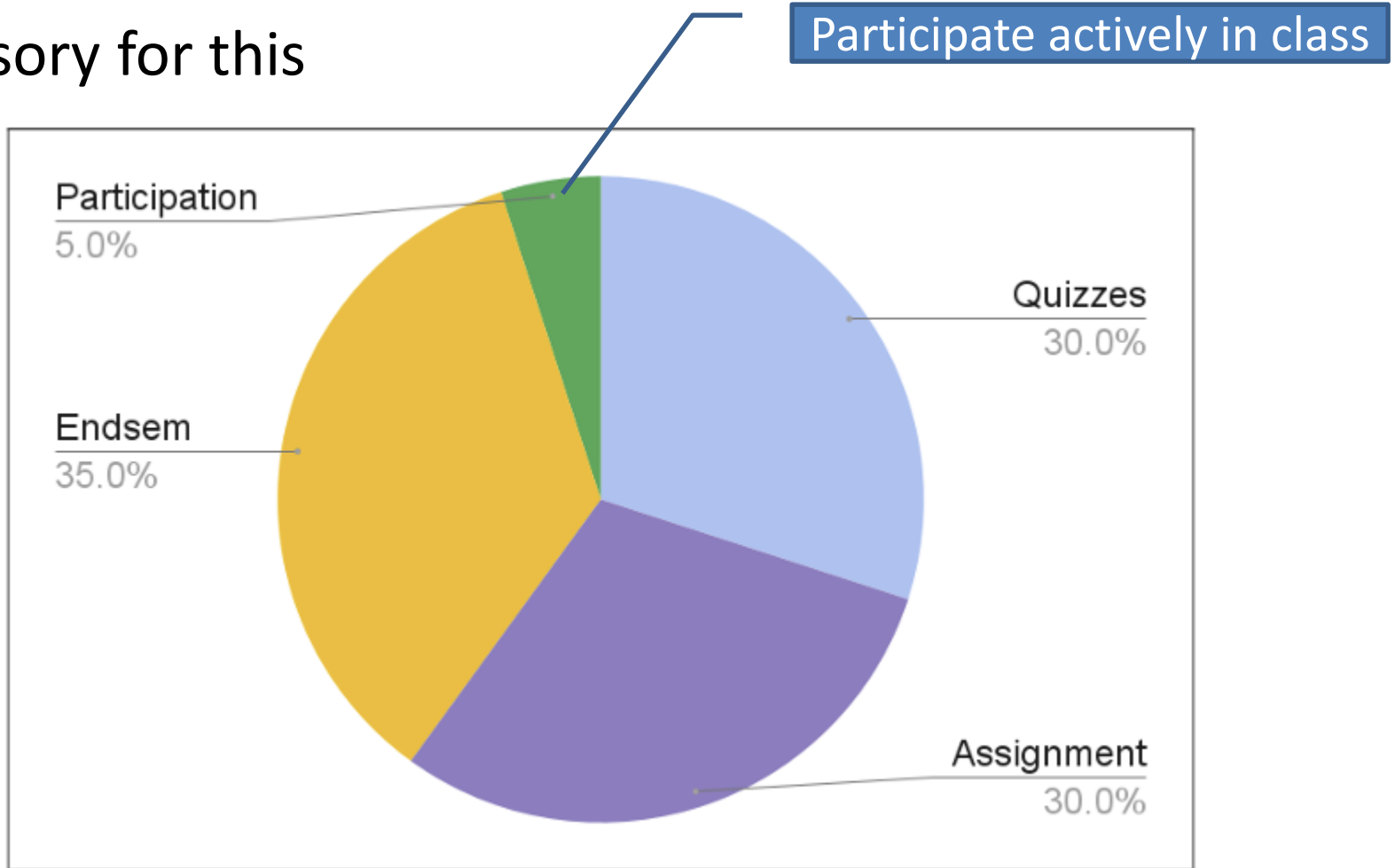


Sanjay Mittal

Dean of Academic Affairs

Grading Policy

75% attendance is compulsory for this course



Lectures

- Lecture slides are pointers for the topic
 - They won't be as verbose as a book!
- In case you miss a class, please ensure you are up to date with the lecture content
 - Either ask your friend
 - Or, ask the instructor (Saturday class)

Assignment

- One programming assignment in C
- In a group (group size = 4 or 5)
 - Send group member information by Jan 14 via Google forms (link will be shared on Jan 8)
 - Include clearly names, roll numbers, IITK email-ids
- Mode of submission will be explained in due time

Assignment

- Timeline: Early February to Early March
- Credit for early submission
- Penalty for late submission
- Cannot be completed in a day!
- Discussion is NOT allowed outside your group
 - You are responsible to maintain your code and report within your group only

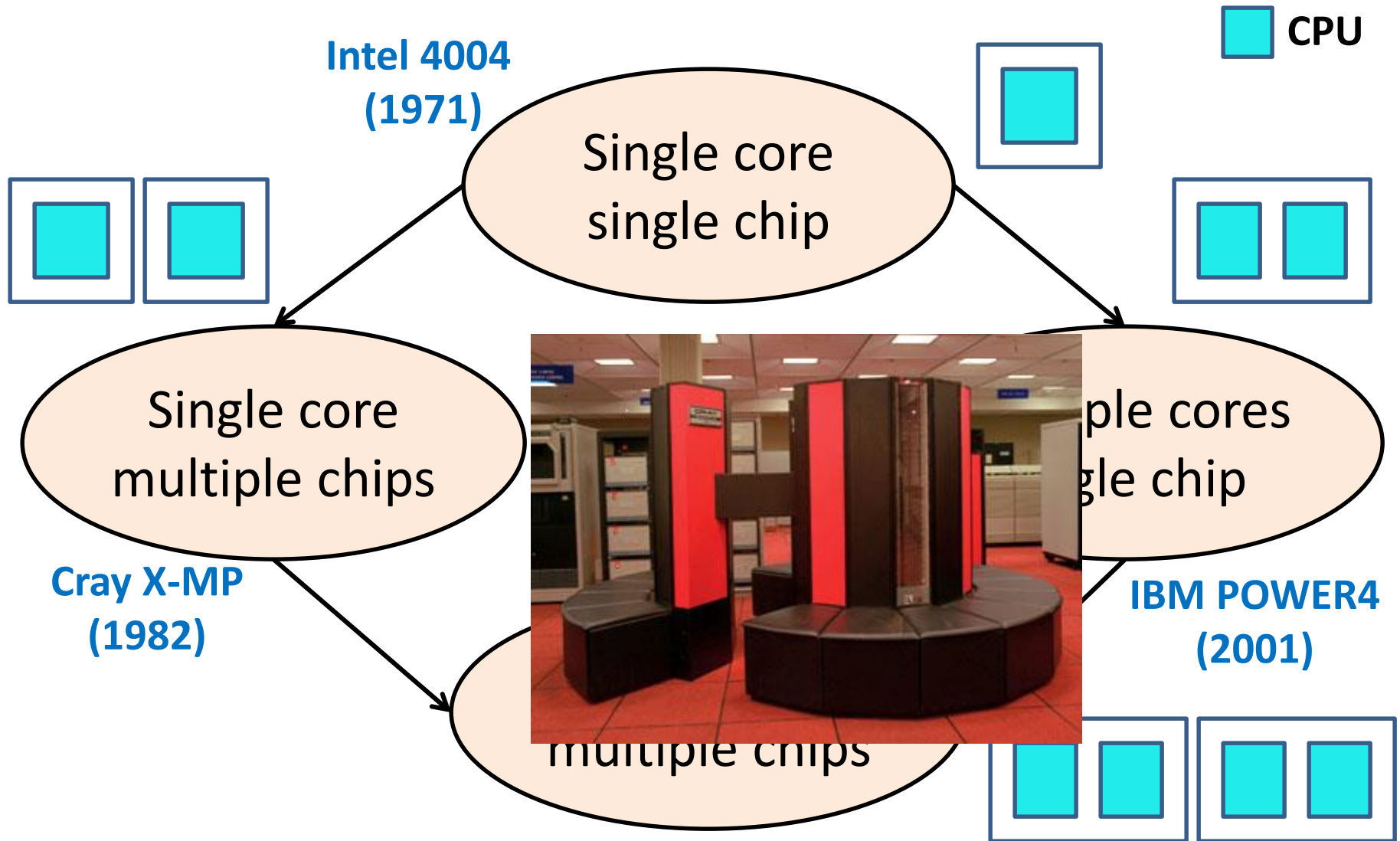
Plagiarism

Plagiarism will NOT be tolerated
Use of AI tools is NOT allowed

Lecture 1

Introduction

Multicore Era

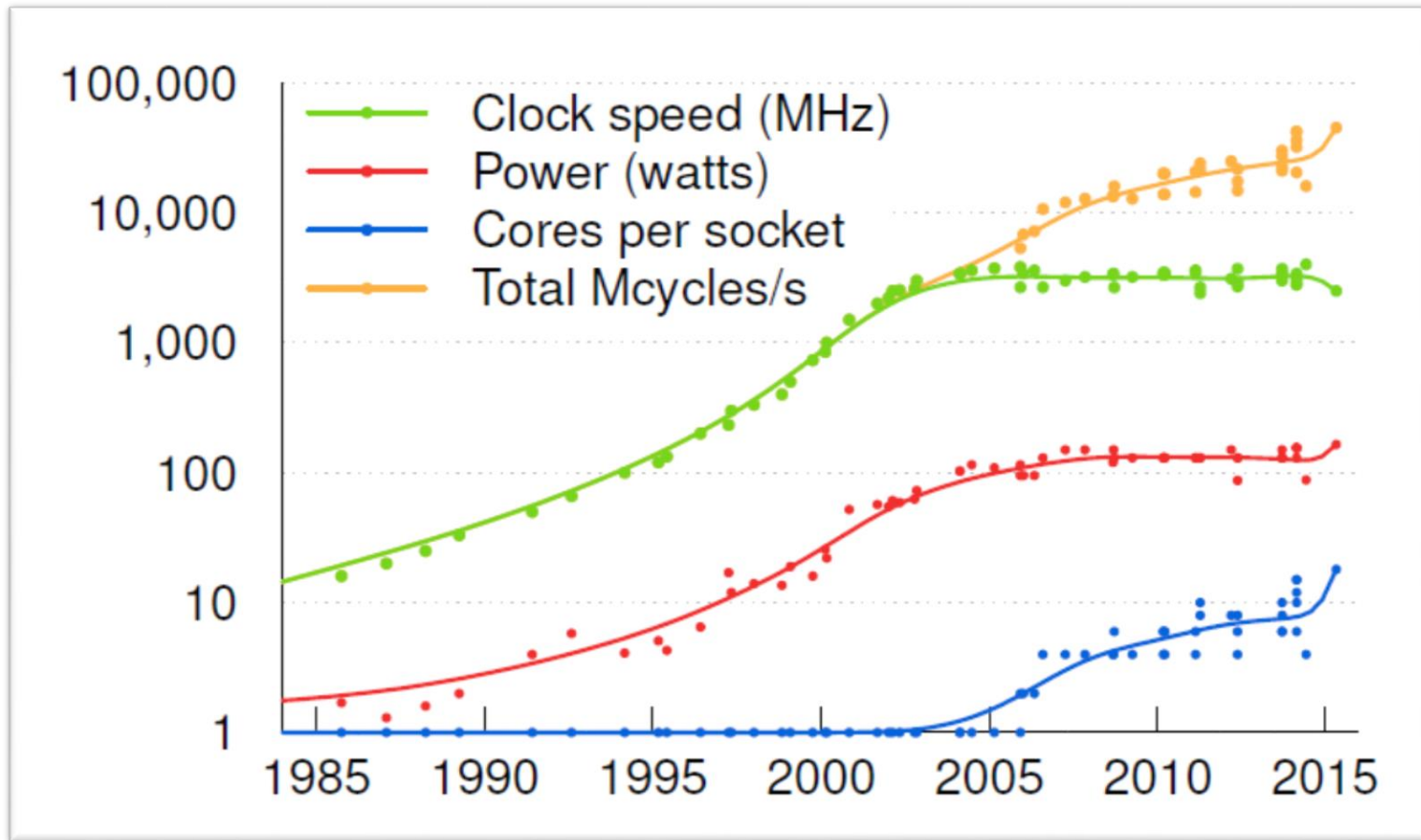


Number of transistors in a chip doubles every 18 months



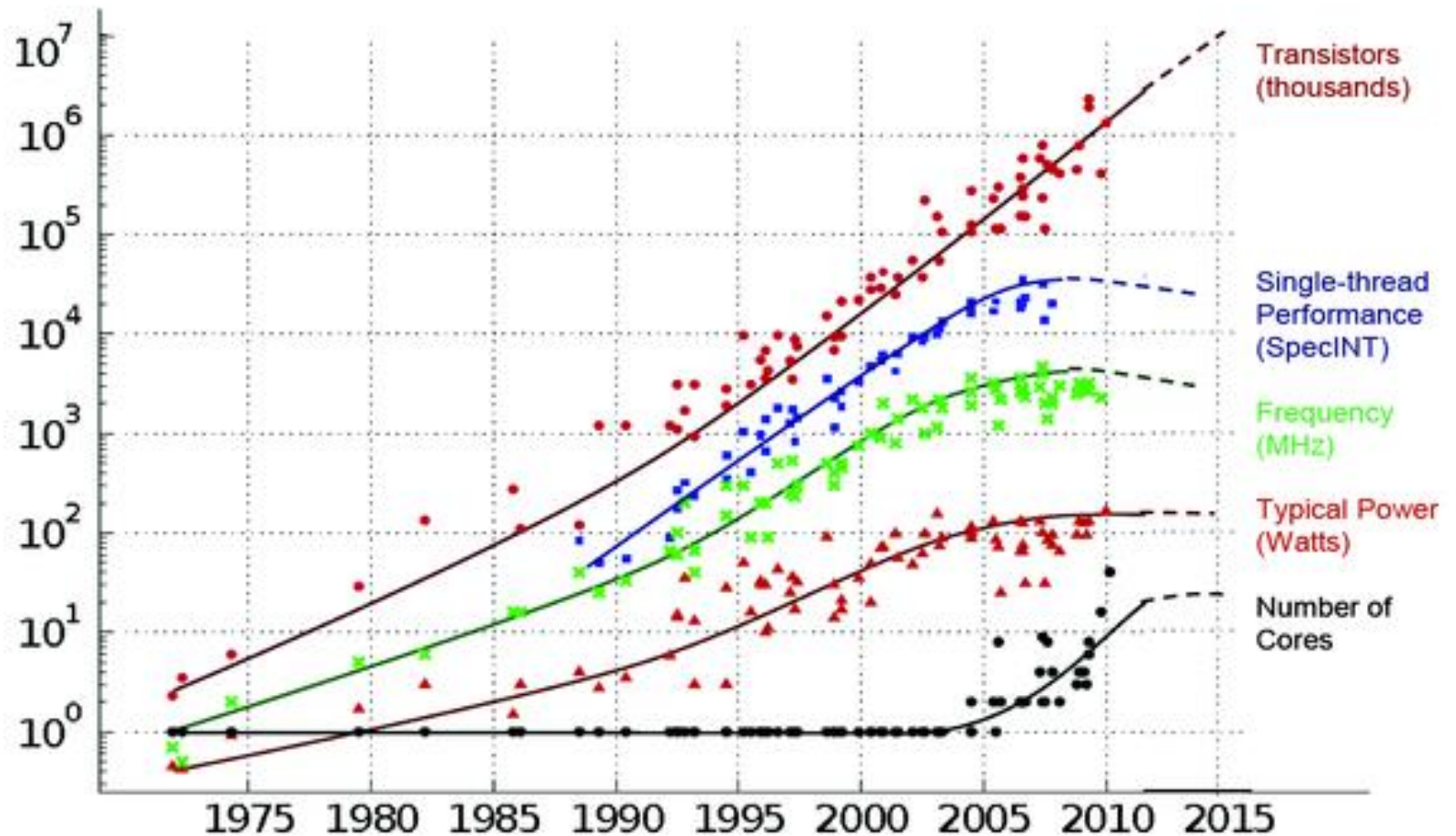
12

Trends



[Source: M. Frans Kaashoek, MIT]

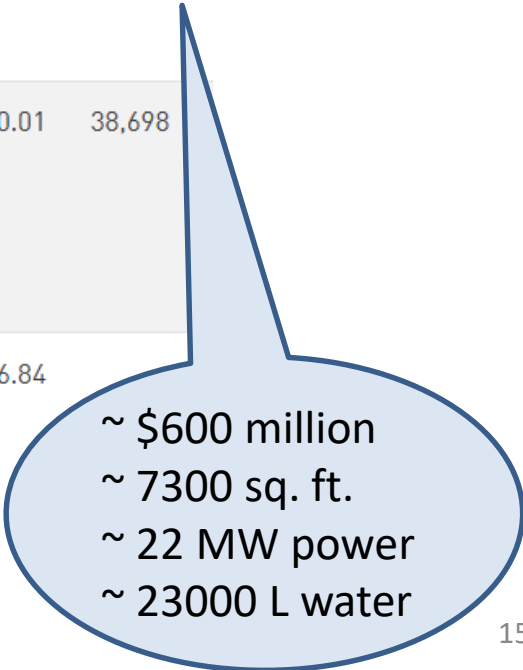
35 YEARS OF MICROPROCESSOR TREND DATA



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

top500.org (Nov'24)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	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	9,264,128	1,012.00	1,980.01	38,698
4	Eagle - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	



~ \$600 million
~ 7300 sq. ft.
~ 22 MW power
~ 23000 L water

Top #1 Supercomputer

<https://www.top500.org/resources/top-systems/>



El Capitan: DOE/NNSA/LLNL
No.1 in Nov 2024



Frontier: DOE/SC/Oak Ridge National
Laboratory
No.1 in Jun 2022



Supercomputer Fugaku: RIKEN Center for
Computational Science
No.1 from Jun 2020 until Jun 2022



Summit: DOE/SC/Oak Ridge National
Laboratory
No.1 from Jun 2018 until Nov 2019



Sunway TaihuLight: National
Supercomputing Center in Wuxi
No.1 from Jun 2016 until Nov 2017



Tianhe-2 (MilkyWay-2) : National University
of Defense Technology
No.1 from Jun 2013 until Nov 2015

green500.org (Nov'23)

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	293	Henri - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	44	Frontier TDS - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
3	17	Adastra - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	58.021

Metric of interest: Performance per Watt


[HOME](#)
[COMPLETE RESULTS](#)
[GREEN GRAPH500](#)
[SUBMISSIONS](#)
[BE](#)

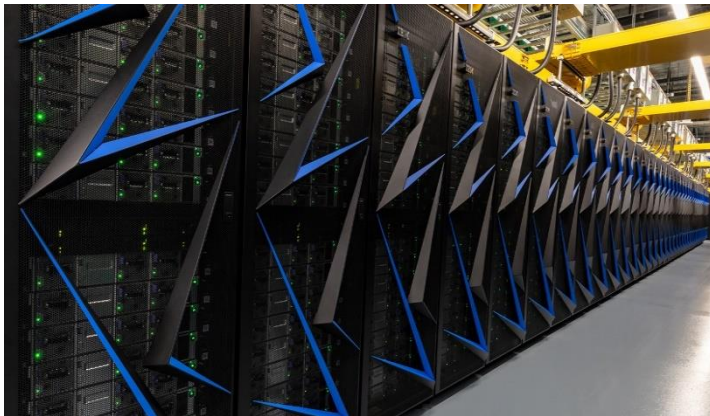
RANK ↕	PREVIOUS RANK ↕	MACHINE ↕	VENDOR ↕	TYPE ↕	NETWORK ↕	INSTALLATION SITE ↕	LOCATION ↕
1	new	Supercomputer Fugaku	Fujitsu	Fujitsu A64FX	Tofu Interconnect D	RIKEN Center for Computational Science (R-CCS)	Kobe Hyogo
2	new	Wuhan Supercomputer	HUST	Kunpeng 920+Tesla A100	Custom	Wuhan Supercomputing Center	Wuhan
3	3	Frontier	HPE	HPE Cray EX235a	Slingshot-11	DOE/SC/Oak Ridge National Laboratory	Oak Ridge TN
4	new	Pengcheng Cloudbrain-II	HUST-Pengcheng Lab-HUAWEI	Kunpeng 920+Ascend 910	Custom	Pengcheng Lab	ShenZhen

HPL-MXP MIXED-PRECISION BENCHMARK

The HPL-MxP benchmark seeks to highlight the emerging convergence of high-performance computing (HPC) and artificial intelligence (AI) workloads. While traditional HPC focused on simulation runs for modeling phenomena in physics, chemistry, biology, and so on, the mathematical models that drive these computations require, for the most part, 64-bit accuracy. On the other hand, the machine learning methods that fuel advances in AI achieve desired results at 32-bit and even lower floating-point precision formats. This lesser demand for accuracy fueled a resurgence of interest in new hardware platforms that deliver a mix of unprecedented performance levels and energy savings to achieve the classification and recognition fidelity afforded by higher-accuracy formats.

<https://hpl-mxp.org/>

Making of a Supercomputer



Source: energy.gov

Greenest Data Centre?



Source: MIT TR 06/19

“The 149,000 square foot facility built on a hillside overlooking the UC Berkeley campus and San Francisco Bay will house one of the most energy-efficient computing centers anywhere, tapping into the region’s mild climate to cool the supercomputers at the National Energy Research Scientific Computing Center (NERSC) and eliminating the need for mechanical cooling.”

BERKELEY LAB OPENS STATE-OF-THE-ART FACILITY FOR COMPUTATIONAL SCIENCE

Wang Hall takes advantage of Lab’s hillside location for advanced energy efficiency

NOVEMBER 12, 2015

Contact: Jon Bashor, jbashor@lbl.gov, 510-486-5849

A new center for advancing computational science and networking at research institutions and universities across the country opened today at the Department of Energy’s (DOE) Lawrence Berkeley National Laboratory (Berkeley Lab).

Named Shyh Wang Hall, the facility will house the National Energy Research Scientific Computing Center, or NERSC, one of the world’s leading supercomputing centers for open science which serves nearly 6,000 researchers in the U.S. and abroad. Wang Hall will also be the center of operations for DOE’s Energy Sciences Network, or ESnet, the fastest network dedicated to science, connecting tens of thousands of scientists as they collaborate on solving some of the world’s biggest scientific challenges.

Complementing NERSC and ESnet in the facility will be research programs in applied mathematics and computer science that develop new methods for advancing scientific discovery. Researchers from UC Berkeley will also share space in Wang Hall as they collaborate with Berkeley Lab staff on computer science programs.



Top Supercomputers from India (Nov'23)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
90	AIRAWAT - PSAI - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR, Netweb Technologies Center for Development of Advanced Computing (C-DAC) India	81,344	8.50	13.17	
163	PARAM Siddhi-AI - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, EVIDEN Center for Development of Advanced Computing (C-DAC) India	41,664	4.62	5.27	
201	Pratyush - Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE Indian Institute of Tropical Meteorology India	119,232	3.76	4.01	1,353
354	Mihir - Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE National Centre for Medium Range Weather Forecasting India	83,592	2.57	2.81	955

2024...

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
136	AIRAWAT - PSAI - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR, Netweb Technologies Center for Development of Advanced Computing (C-DAC) India	81,344	8.50	13.17	
188	Arka - BullSequana XH2000, AMD EPYC 7643 48C 2.3GHz, Infiniband HDR, Red Hat Enterprise Linux, EVIDEN Indian Institute of Tropical Meteorology India	290,016	5.94	7.40	1,236
189	Arunika - BullSequana XH2000, AMD EPYC 7643 48C 2.3GHz, Infiniband HDR, Red Hat Enterprise Linux, EVIDEN National Centre for Medium Range Weather Forecasting India	203,040	5.94	7.40	1,236
268	Pratyush - Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE Indian Institute of Tropical Meteorology India	119,232	3.76	4.01	1,353
400	Arka AI/ML - NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GHz, NVIDIA H100 SXM5 80GB, Octo-rail NVIDIA HDR100 Infiniband, Red Hat Enterprise Linux, EVIDEN Indian Institute of Tropical Meteorology India	8,176	2.70	3.75	
431	Mihir - Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE National Centre for Medium Range Weather Forecasting India	83,592	2.57	2.81	955

Supercomputing in India [topsc.cdacb.in, Jul'24]

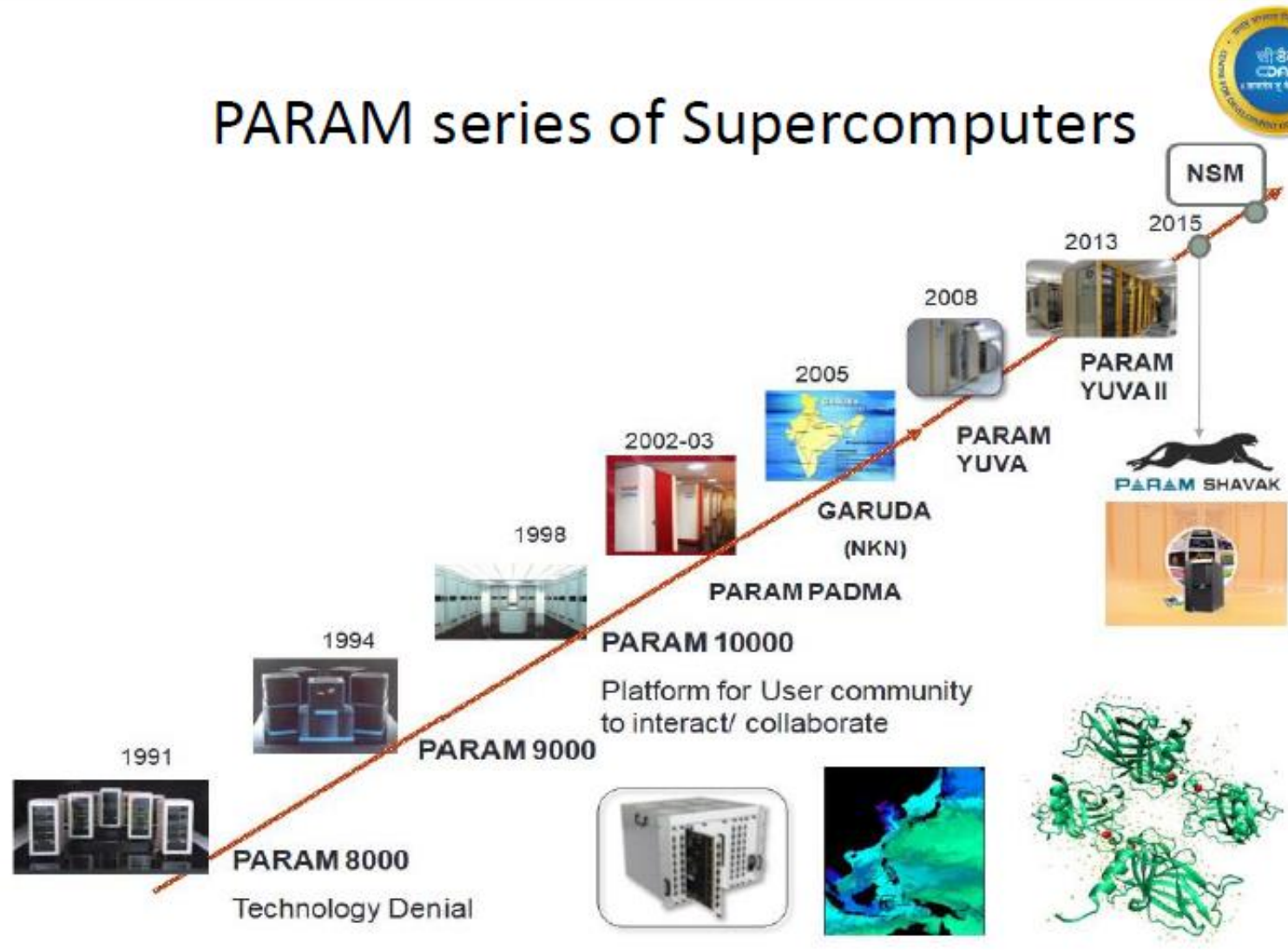
Rank	Site	System	Core/Processor/Socket/Nodes	Rmax (TFlops)	Rpeak (TFlops)
1	Center for Development of Advanced Computing (C-DAC),PUNE	AIRAWAT - PSAI is a NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHZ, NVIDIA A100, INFINIBAND HDR2 More Info OEM:NVIDIA, Bidder:Netweb	81344/2/82	8500	13176
2	Indian Institute of Tropical Meteorology(IITM),Pune	Cray XC-40 class system with 3315 CPU-only (Intel Xeon Broadwell E5-2695 v4 CPU) nodes with Cray Linux environment as OS, and connected by Cray Aries interconnect. Total storage More Info OEM:Cray, Bidder:Cray	119232//3315	3763.9	4006.19
3	National Centre for Medium Range Weather Forecasting (NCMRWF),Noida	Intel Xeon Broadwell E5-2695 v4 CPU) nodes with Cray Linux environment as OS, and connected by Cray Aries interconnect More Info OEM:Cray, Bidder:Cray	83592/2/2612	2570.4	2808.7
4	PARAM Pravega,IISc,Bangalore	The PARAM Pravega is a heterogeneous and hybrid configuration of Intel Xeon Cascade Lake processors,NVIDIA Tesla V100 with NVLink, Mellanox HDR More Info OEM:ATOS, Bidder:ATOS	29952/2/624	1702	2565
5	Indian Institute of Technology (IITK),Kharagpur	The supercomputer PARAM Shakti is based on a heterogeneous and hybrid configuration of Intel Xeon Skylake(6148, 20C, 2.4Ghz) processors, and NVIDIA Tesla V100. The system More Info OEM:Atos, Bidder:Atos	17280/2/442	935	1290.2



Param Sanganak
New HPC facility being installed @IITK

Source: www.iitk.ac.in

PARAM series of Supercomputers



Credit: Ashish Kuvelkar, CDAC

National Supercomputing Mission Sites

Completed NSM Sites

PARAM Shivay



PARAM Shakti



PARAM Brahma



PARAM Yukti



PARAM Sanganak



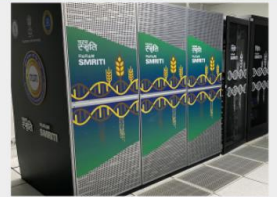
PARAM Pravega



PARAM Seva



PARAM Smriti



PARAM Utkarsh

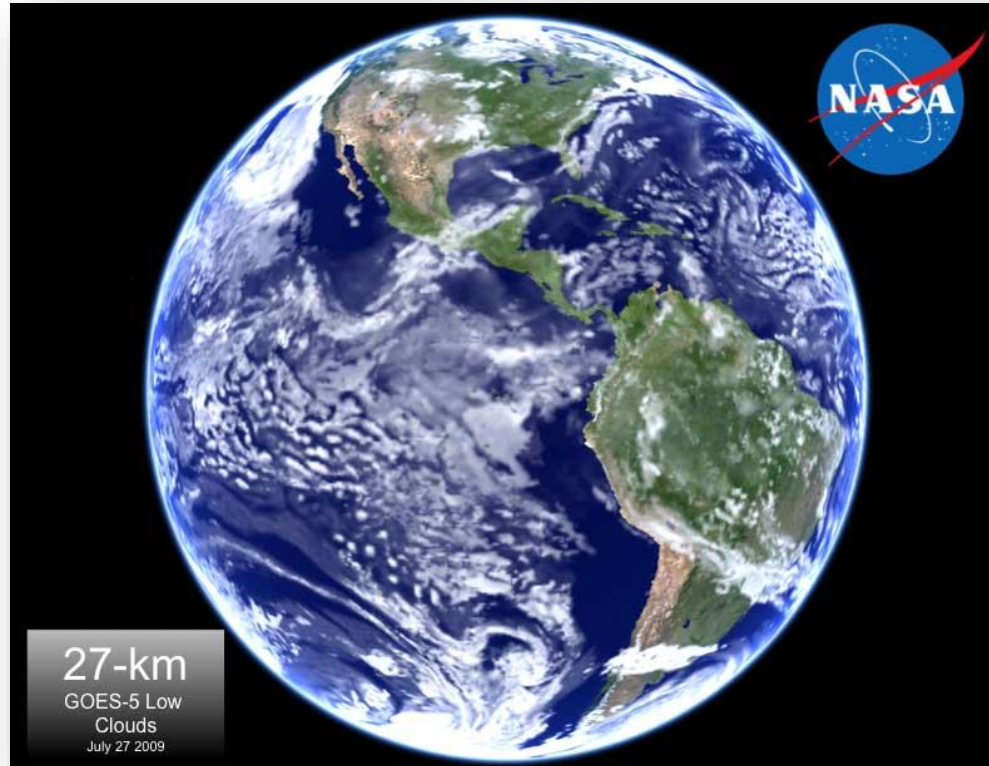


PARAM Ganga



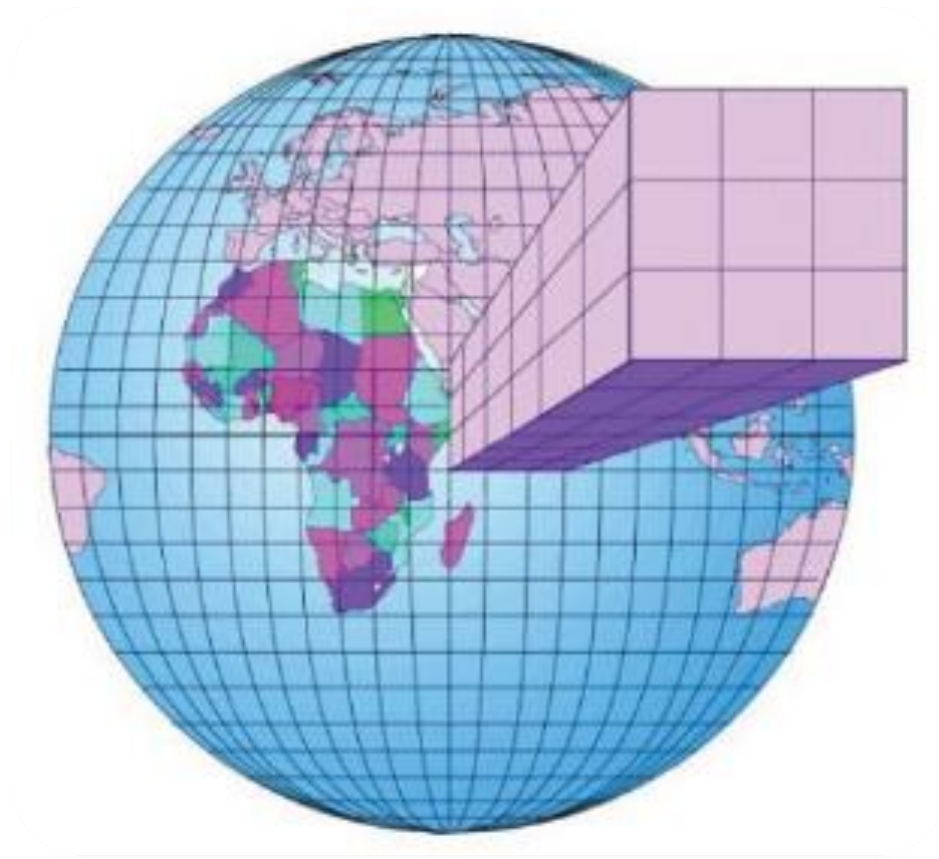
Big Compute

Massively Parallel Codes



Climate simulation of Earth [Credit: NASA]

Discretization

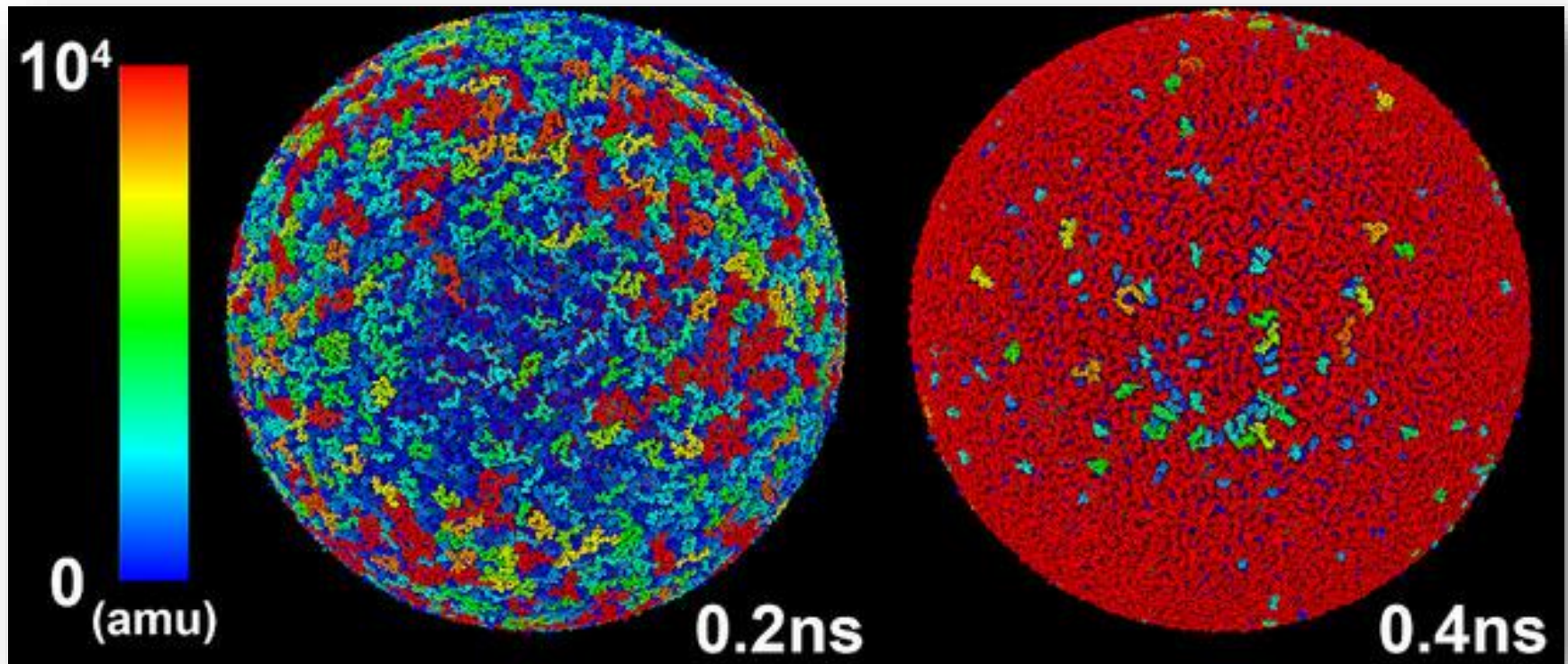


Gridded mesh for a global model [Credit: Tompkins, ICTP]

Numerical Weather Models

- Use numerical methods to solve equations that govern atmospheric processes
- Are based on fluid dynamics and depend on observations of meteorological variables
- Are used to obtain nowcast/forecast

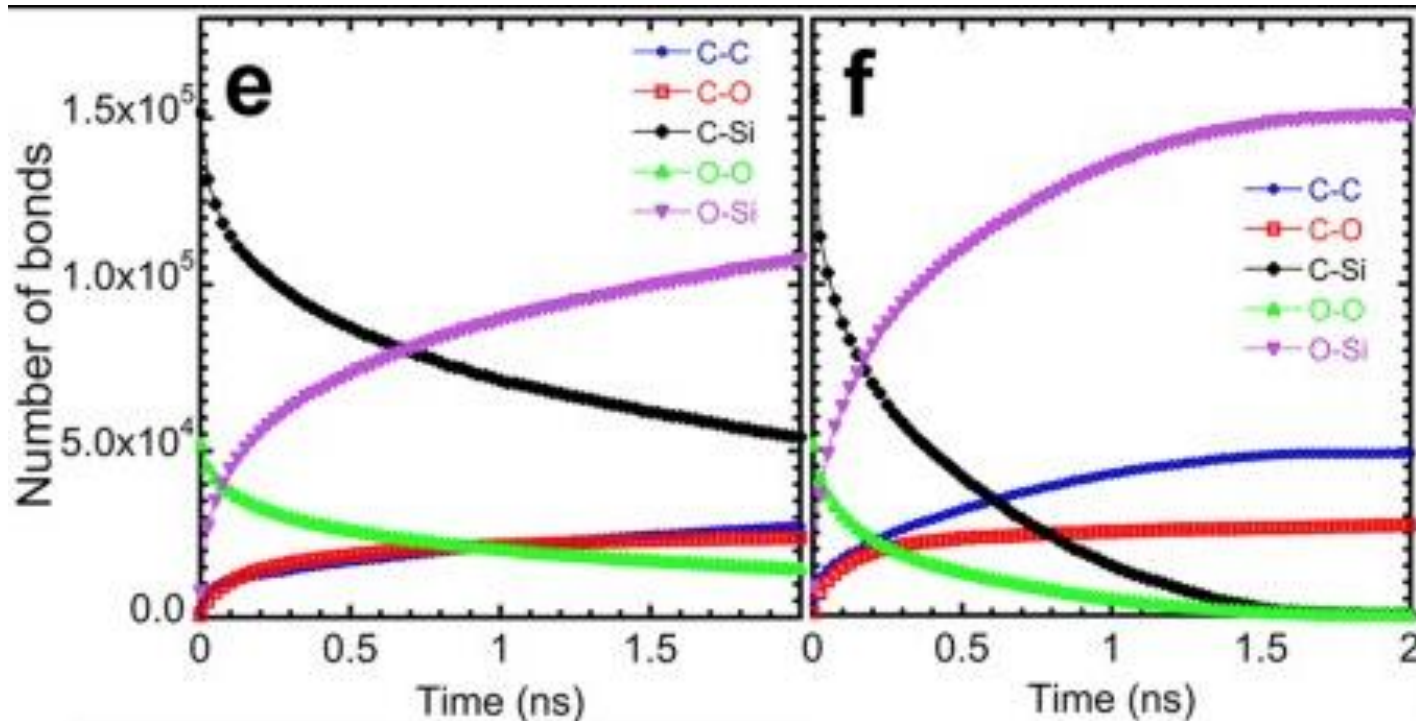
Massively Parallel Simulations



Self-healing material simulation

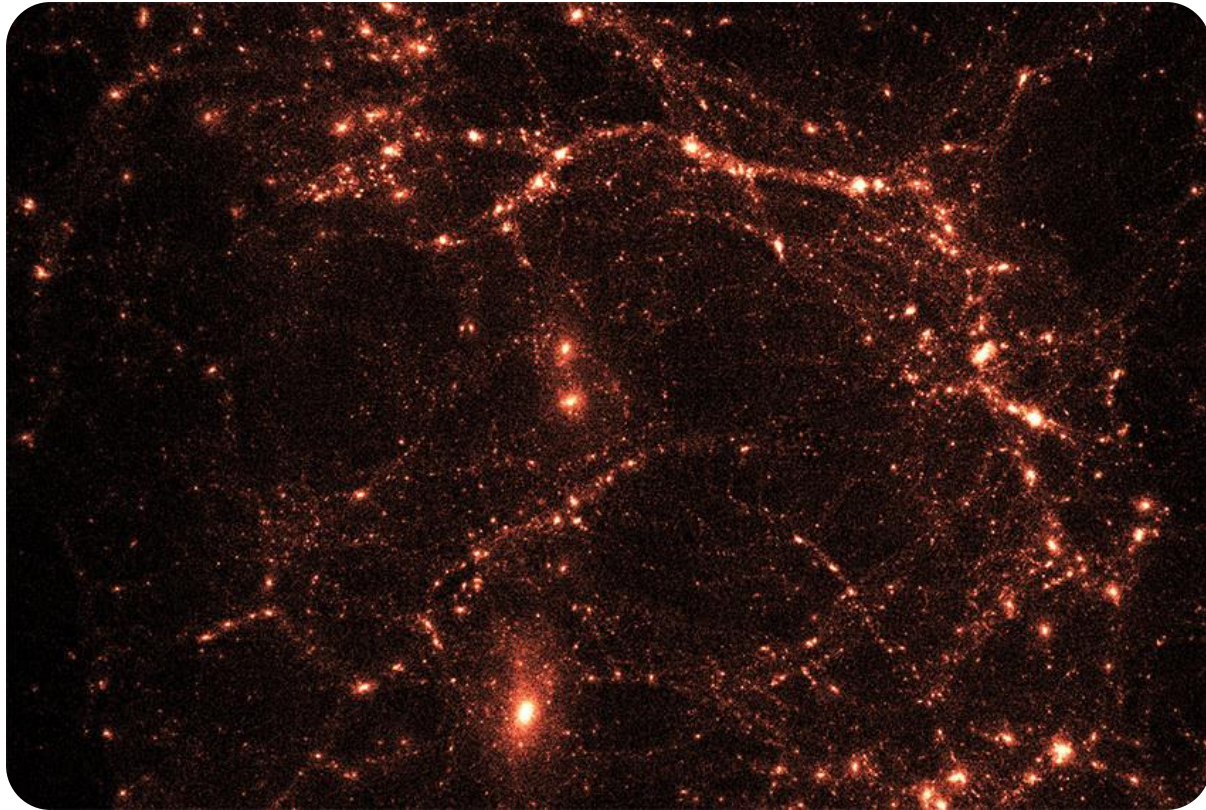
[Nomura et al., “[Nanocarbon synthesis by high-temperature oxidation of nanoparticles](#)”, Scientific Reports, 2016]

Massively Parallel Analysis



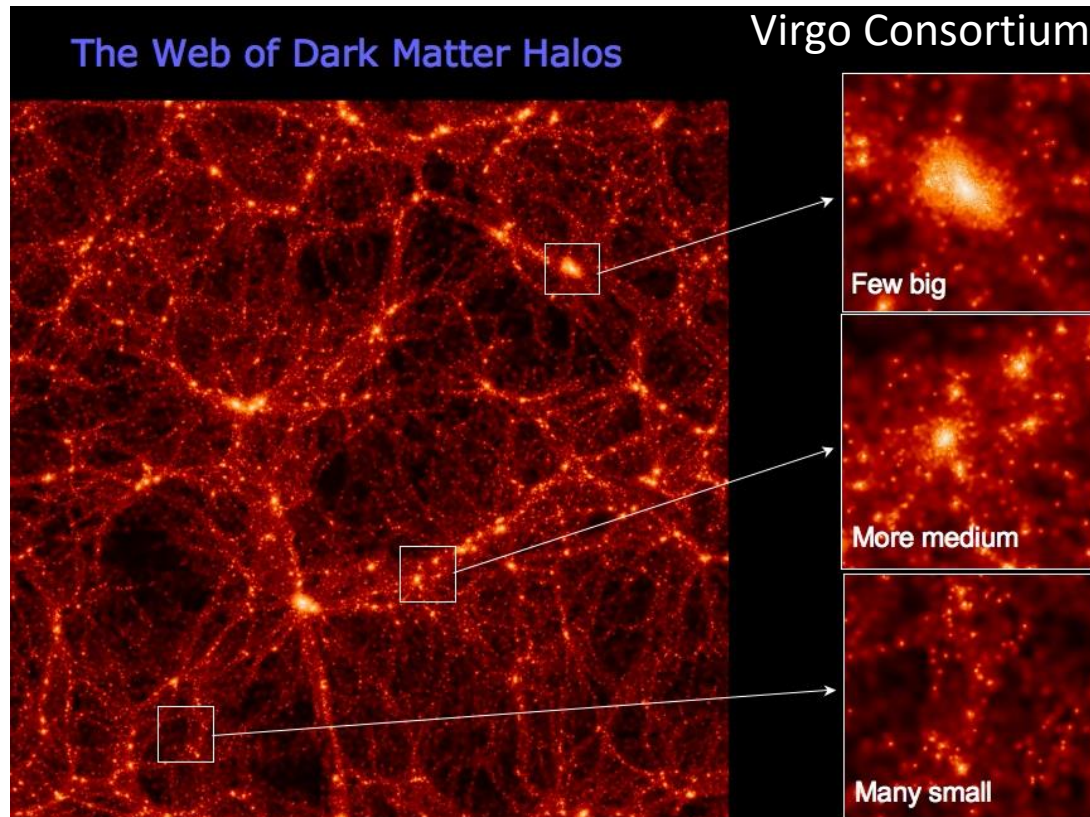
[Nomura et al., “[Nanocarbon synthesis by high-temperature oxidation of nanoparticles](#)”, Scientific Reports, 2016]

Massively Parallel Codes

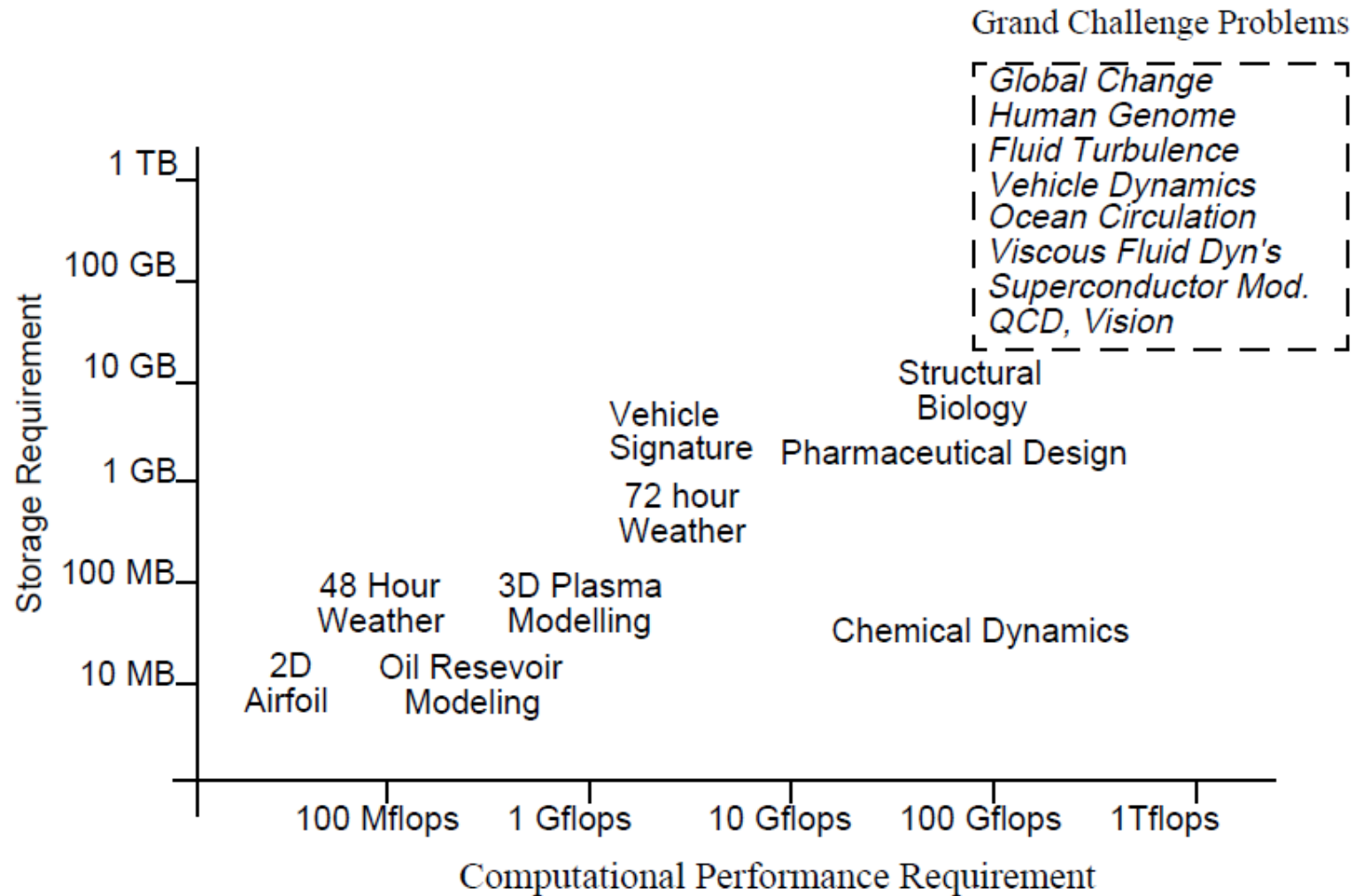


Cosmological simulation [Credit: ANL]

Massively Parallel Analysis



Computational Science

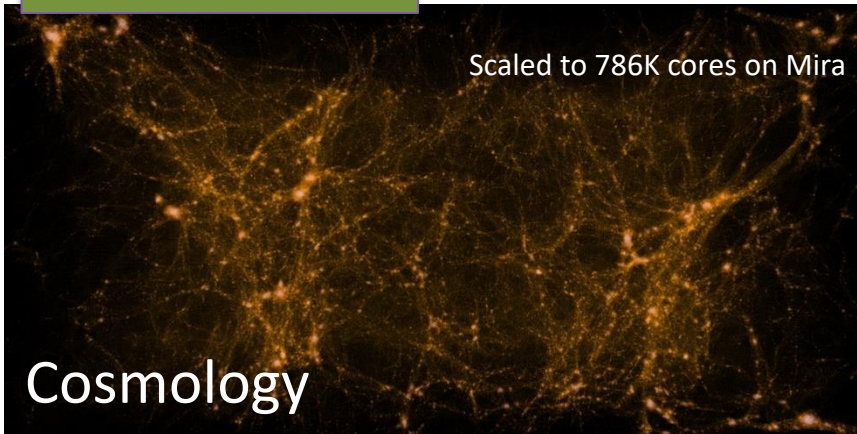


[Source: Culler, Singh and Gupta]

Big Data

Output Data

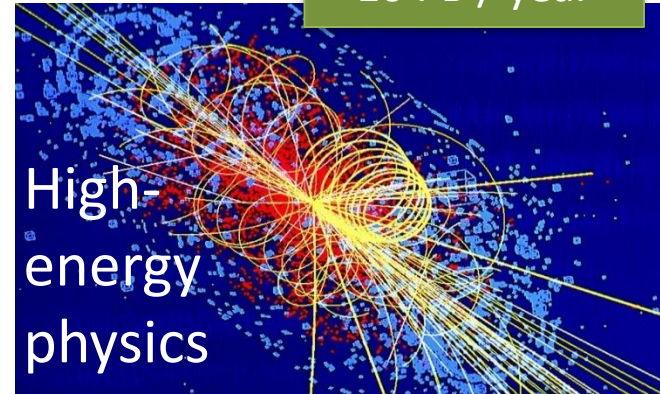
2 PB / simulation



Scaled to 786K cores on Mira

Q Continuum simulation
Source: Salman Habib et al.

10 PB / year



Higgs boson simulation

Source: CERN

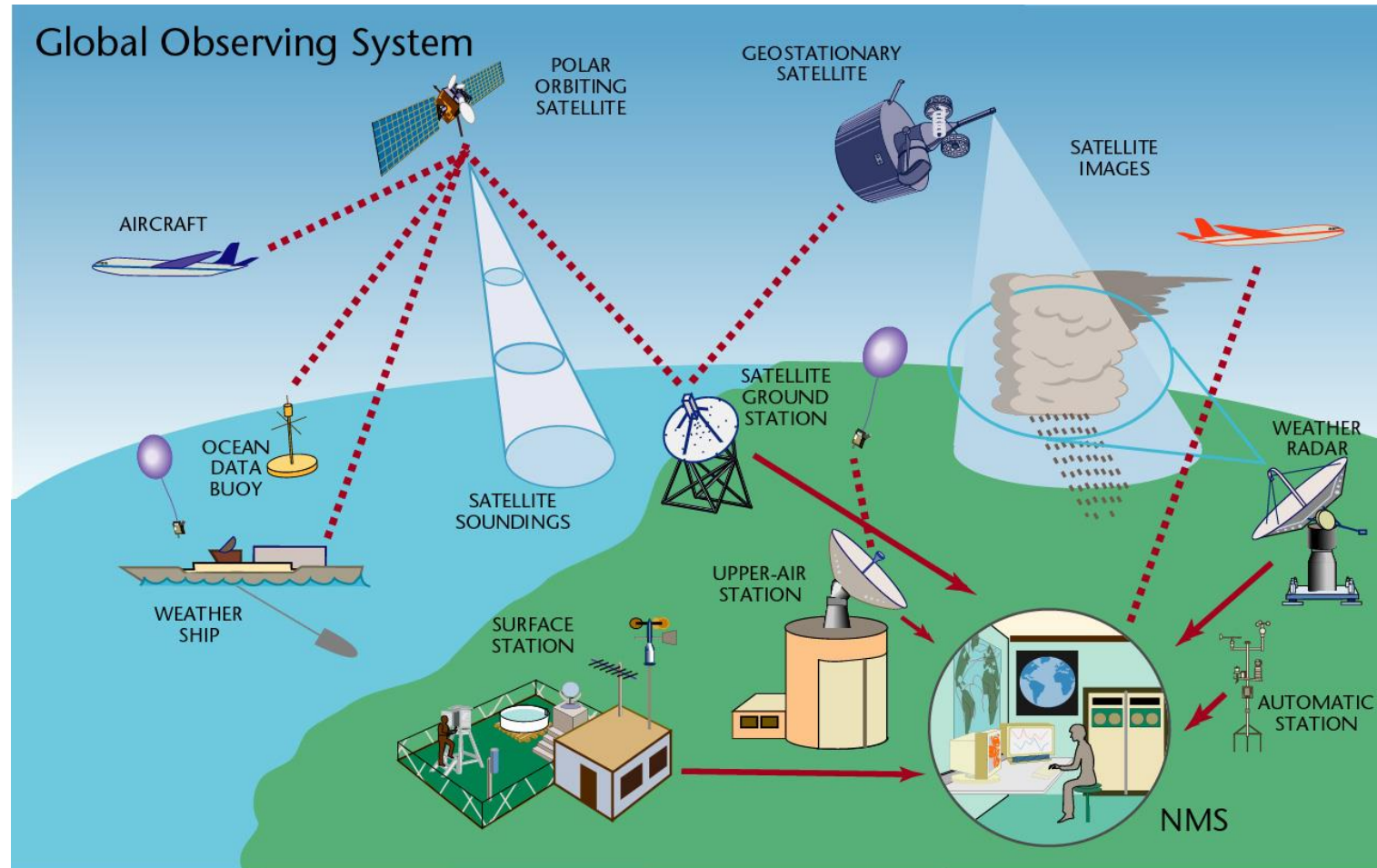
240 TB / simulation



Hurricane simulation

Source: NASA

Input Data



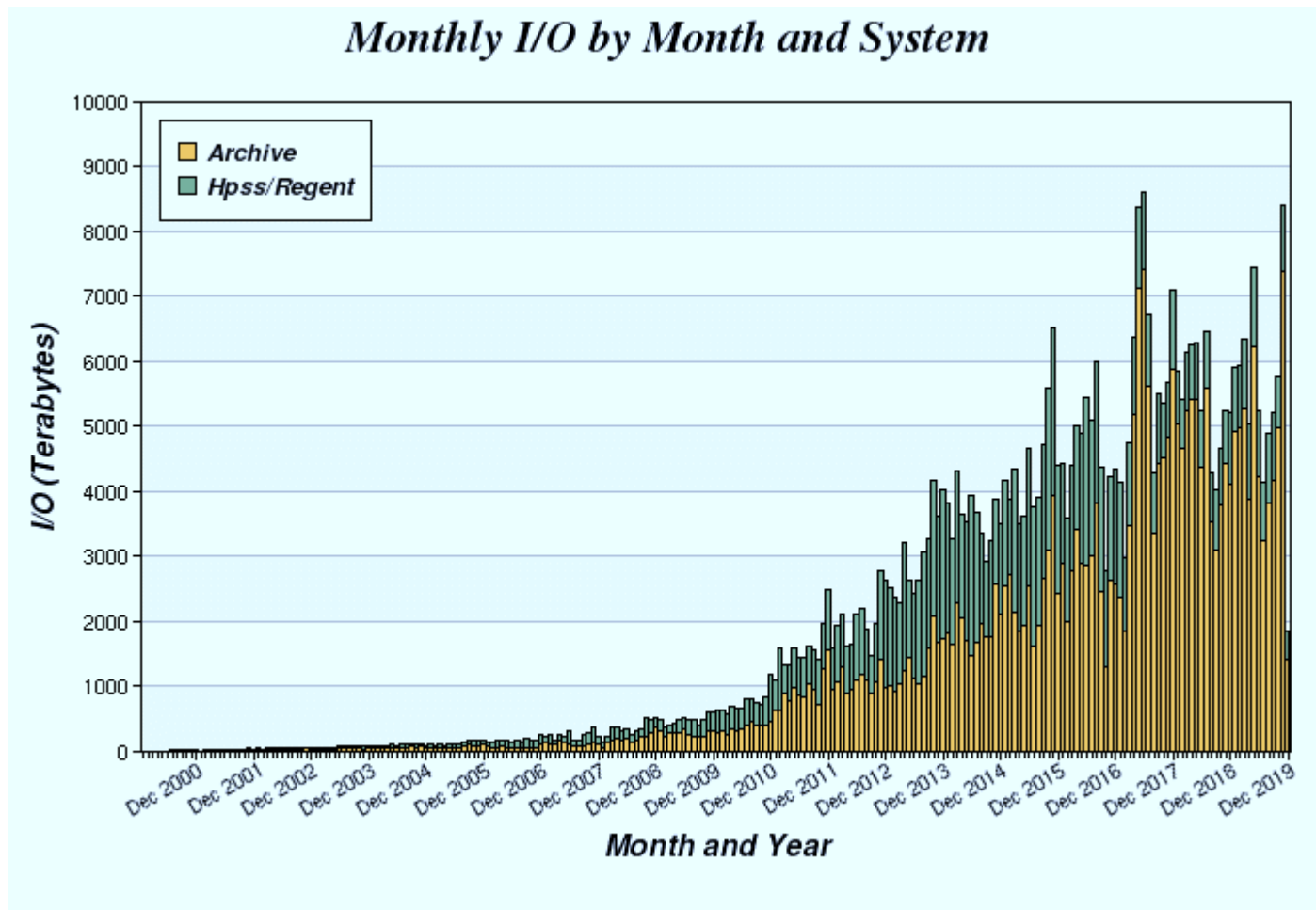
[Credit: World Meteorological Organization]

System Architecture Trends

System attributes	2010	2017-2018		2021-2022	
System peak	2 Peta	150-200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	3 TF	30 TF	10 TF	100 TF
Node memory BW	25 GB/s	0.1TB/sec	1 TB/sec	0.4TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	100,000	10,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200GB/sec	
MTTI	days	O(1day)		O(1 day)	

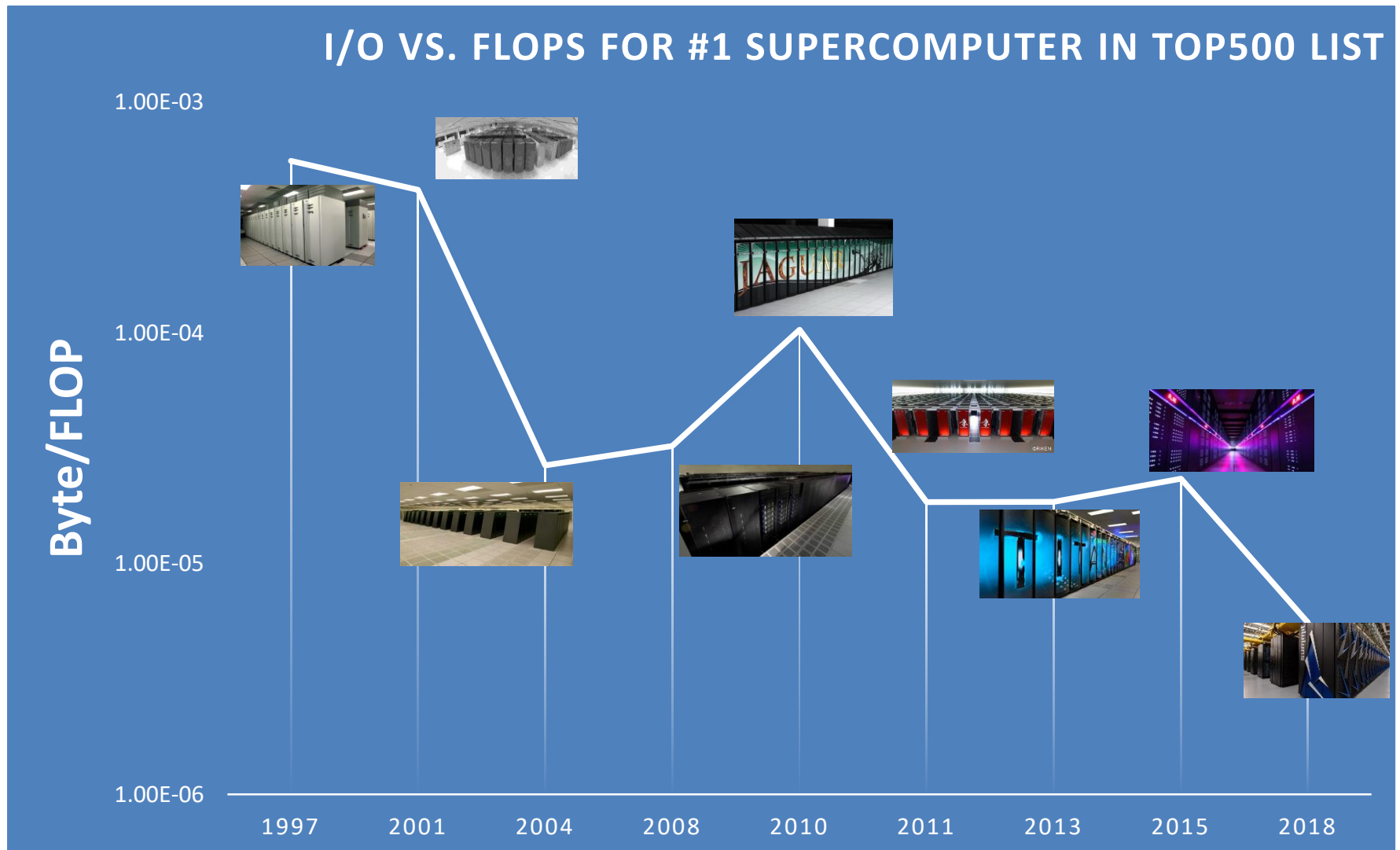
[Credit: Pavan Balaji@ATPESC'17]

I/O trends

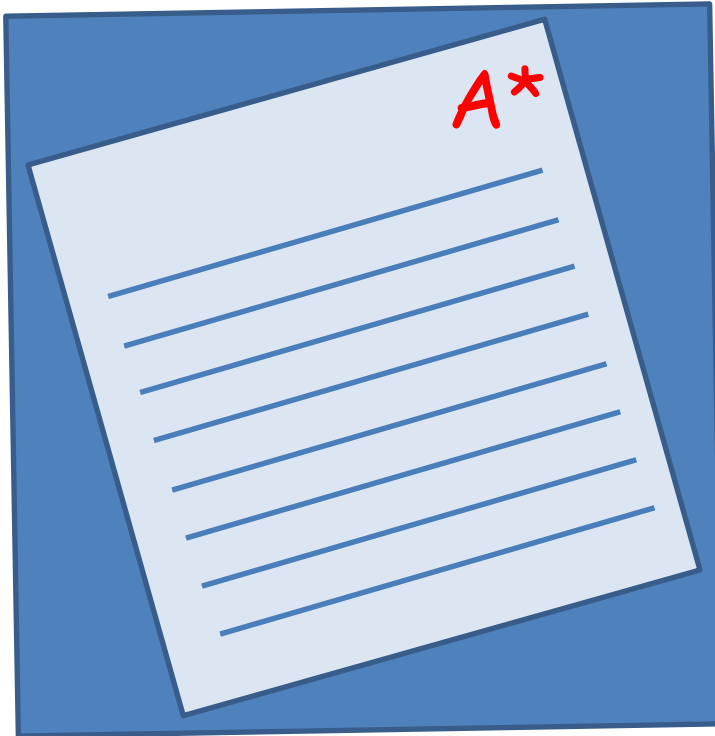


NERSC I/O trends [Credit: www.nersc.gov]

Compute vs. I/O trends



Why Parallel?



20 hours



2 hours

Not really

Parallelism

A parallel computer is a collection of processing elements that communicate and **cooperate** to solve large problems **fast**.

– Almasi and Gottlieb (1989)

Speedup

Example – Sum of squares of N numbers

Serial

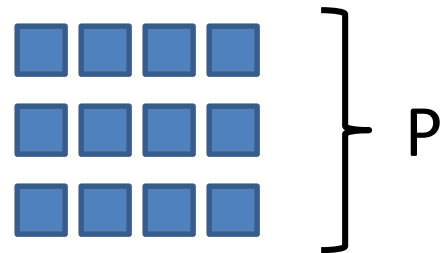
```
for i = 1 to N  
  sum += a[i] * a[i]
```



$O(N)$

Parallel

```
for i = 1 to N/P  
  sum += a[i] * a[i]  
collate result
```



$O(N/P) +$

Communication time

Performance Measure

- Speedup

$$S_p = \frac{\text{Time (1 processor)}}{\text{Time (P processors)}}$$

- Efficiency

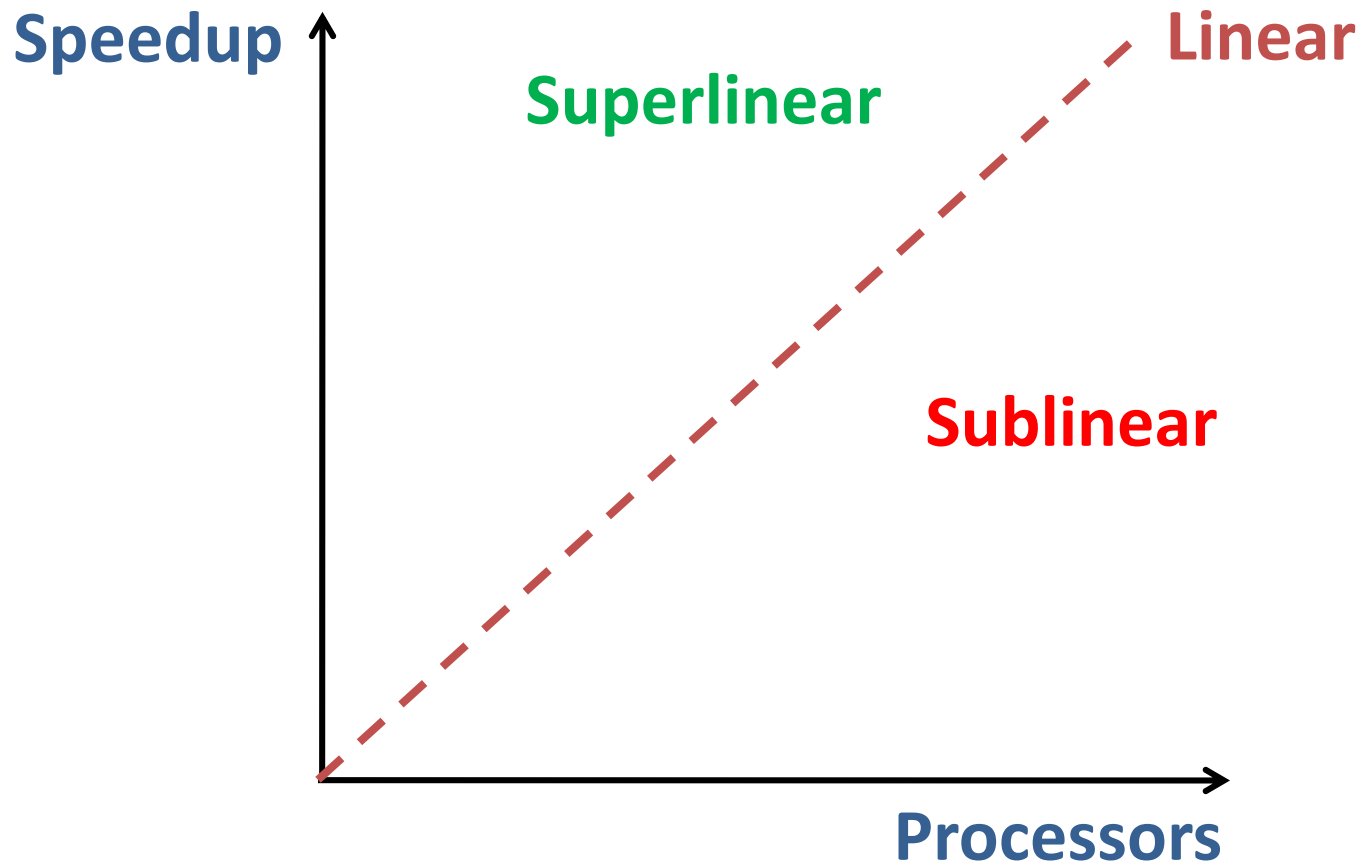
$$E_p = \frac{S_p}{P}$$

Parallel Performance (Parallel Sum)

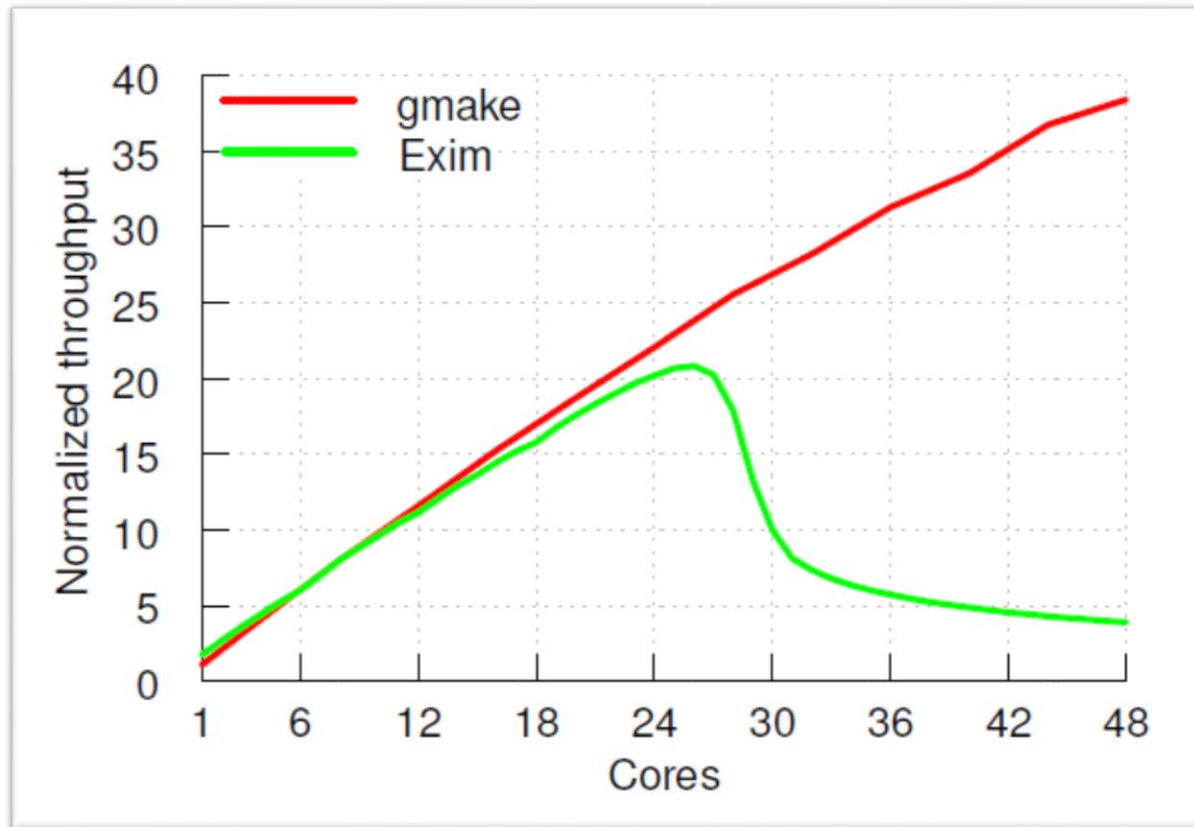
Parallel efficiency of summing 10^7 doubles

#Processes	Time (sec)	Speedup	Efficiency
1	0.025	1	1.00
2	0.013	1.9	0.95
4	0.010	2.5	0.63
8	0.009	2.8	0.35
12	0.007	3.6	0.30

Ideal Speedup

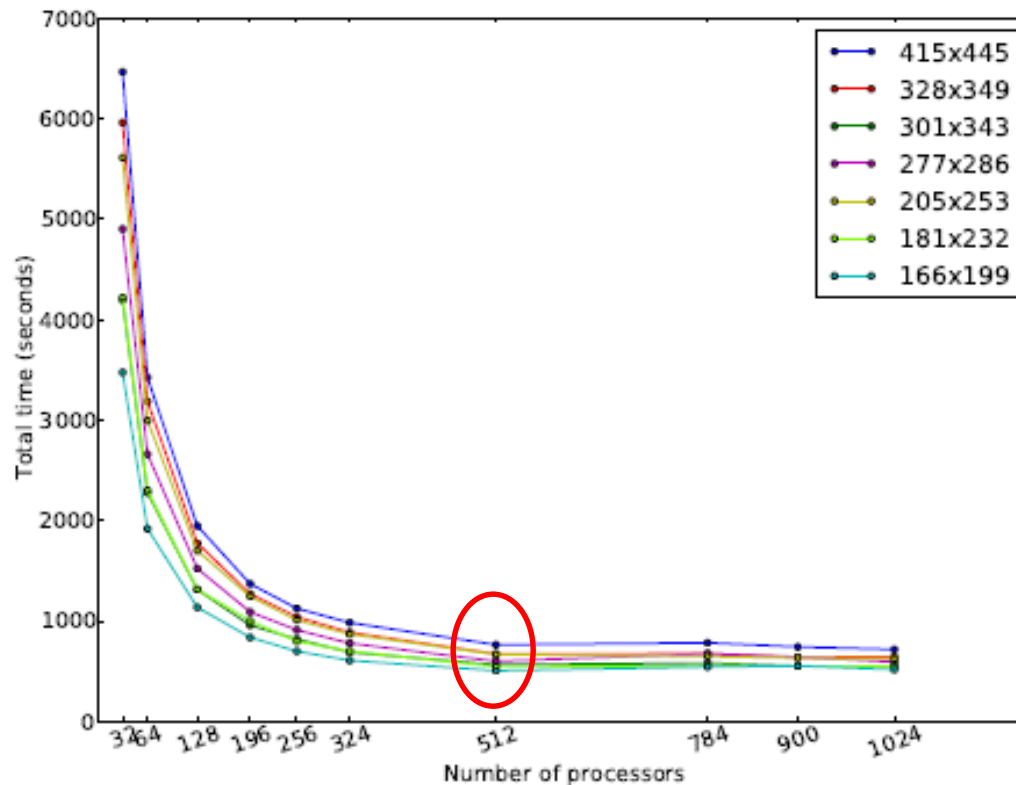


Issue – Scalability



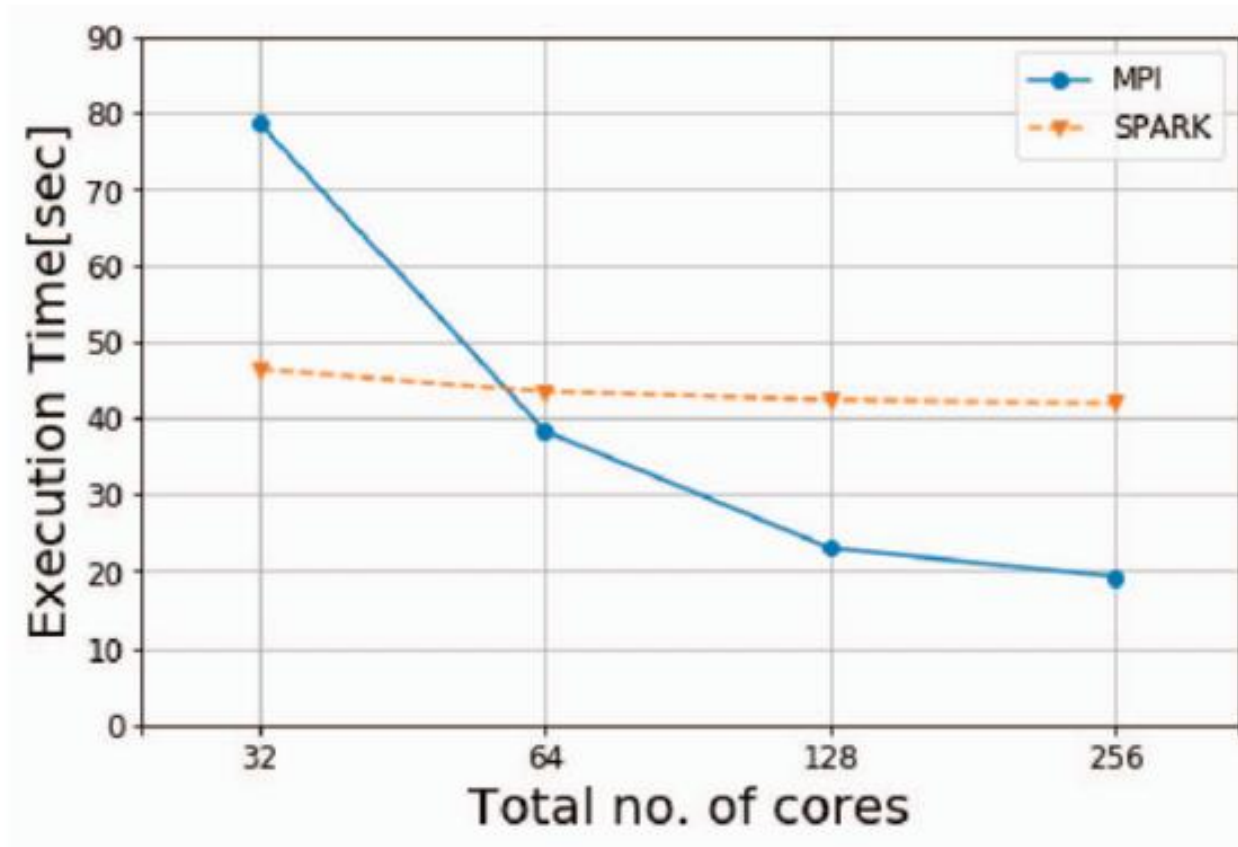
[Source: M. Frans Kaashoek, MIT]

Scalability Bottleneck



Performance of weather simulation application

Scalability and Performance



C vs. Python Parallel Performance

TABLE VI. AVERAGE EXECUTION TIME FOR 1K PARTICLES

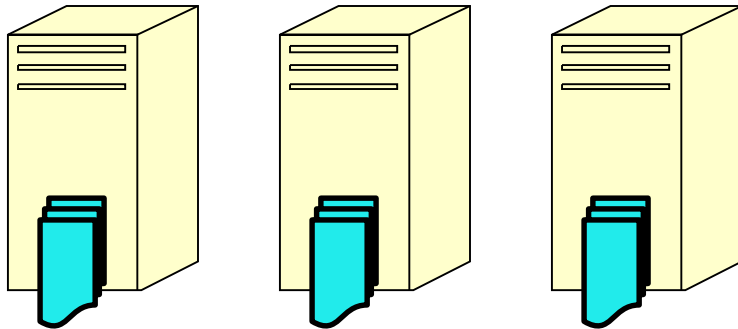
Number of Processes	Time due to C code (Serial time: 0.100565)		Time due to Python code (Serial time: 1.709605)	
	<i>OpenMP</i>	<i>MPI</i>	<i>Threading</i>	<i>MPI</i>
1	0.123928	1.718326	32.771536	31.273835
2	0.100983	1.285032	28.003258	32.362905
3	0.083024	1.167345	24.952718	38.071583
6	0.048501	1.005627	29.082573	34.398242
9	0.036877	0.947252	31.782461	39.992538
12	0.035725	0.931694	32.072825	46.381723
15	0.041571	0.995127	32.871903	65.703518

Parallelism

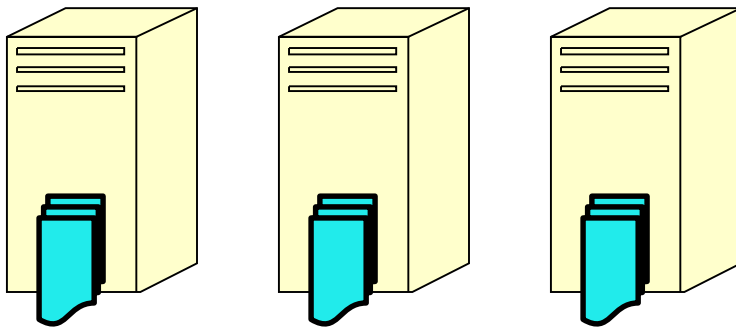
A parallel computer is a collection of processing elements that **communicate** and cooperate to solve large problems fast.

– Almasi and Gottlieb (1989)

Distributed Memory Systems



Node



Cluster

- Networked systems
- Distributed memory
 - Local memory
 - Remote memory
- Parallel file system

Parallel Programming Models

Libraries	MPI, TBB, Pthread, OpenMP, ...
New languages	Haskell, X10, Chapel, ...
Extensions	Coarray Fortran, UPC, Cilk, OpenCL, ...

- Shared memory
 - OpenMP, Pthreads, ...
- Distributed memory
 - MPI, UPC, ...
- Hybrid
 - MPI + OpenMP

This course ...

Large-scale Parallel Computing

Message
passing

Parallel
algorithms

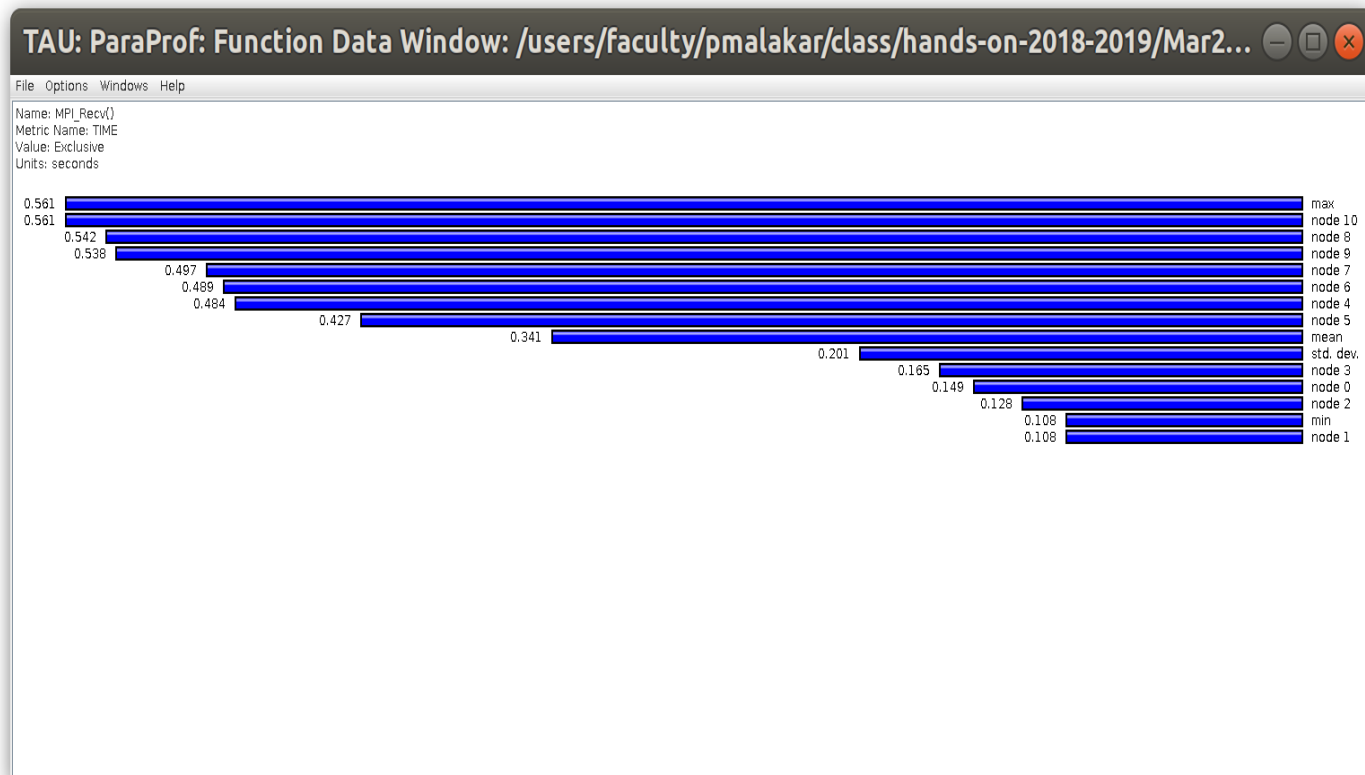
Designing
parallel codes

Performance
analysis

Message Passing Paradigm

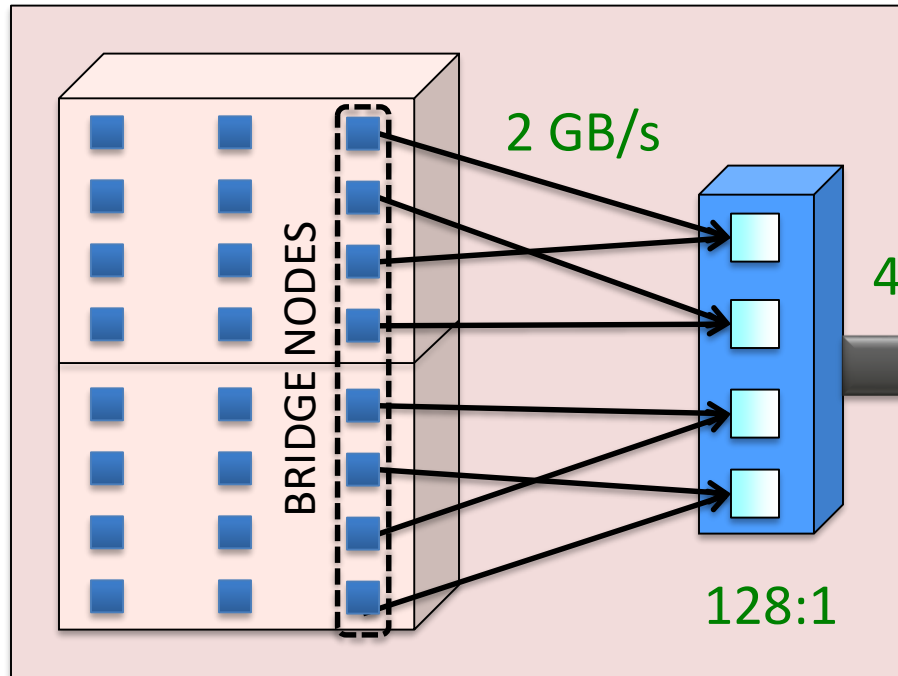
- Point-to-point (P2P) communications
- Collective communications
- Algorithms
- Performance

Profiling



Parallel I/O

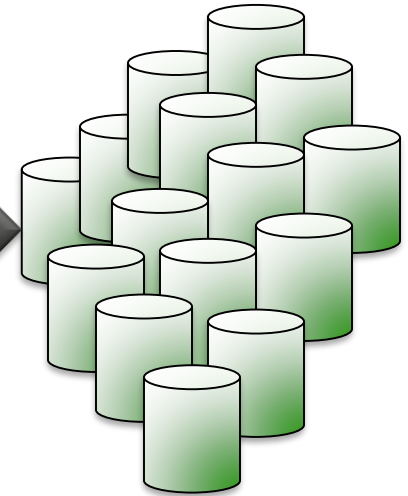
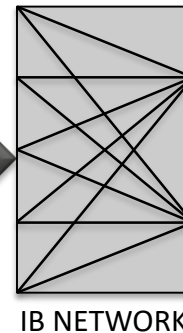
NOT SHARED



Compute node rack

I/O nodes

SHARED



GPFS filesystem

Job Scheduling



NODES



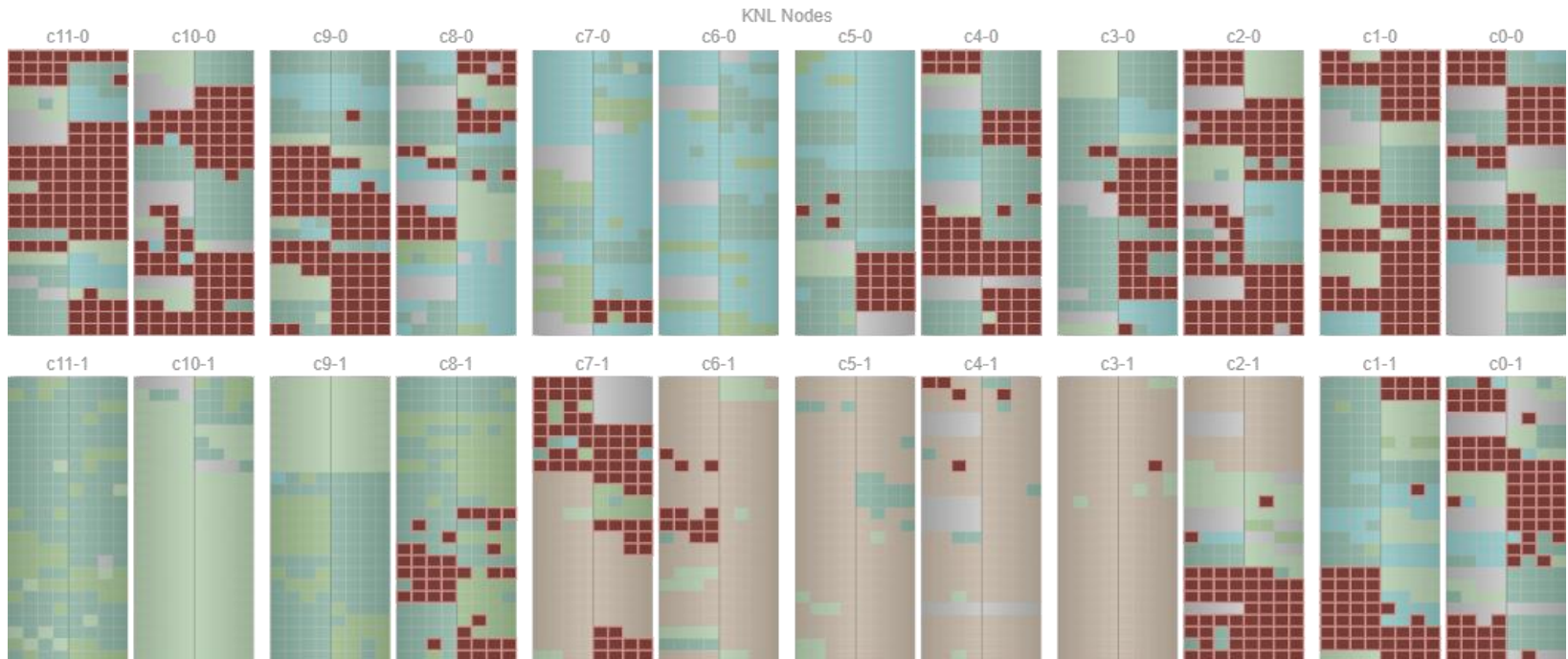
JOBs



USERS

Example of a real supercomputer activity
- [Argonne National Laboratory Theta jobs](#)

Supercomputer Activity

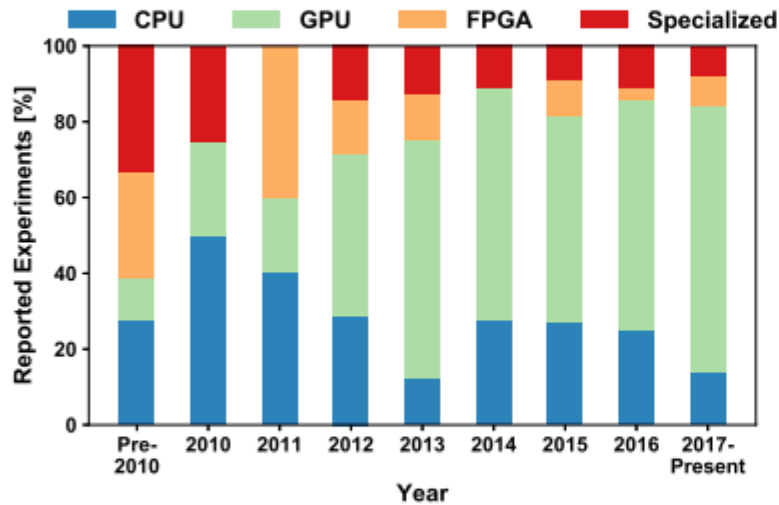


A graphical representation of all jobs running on the supercomputer

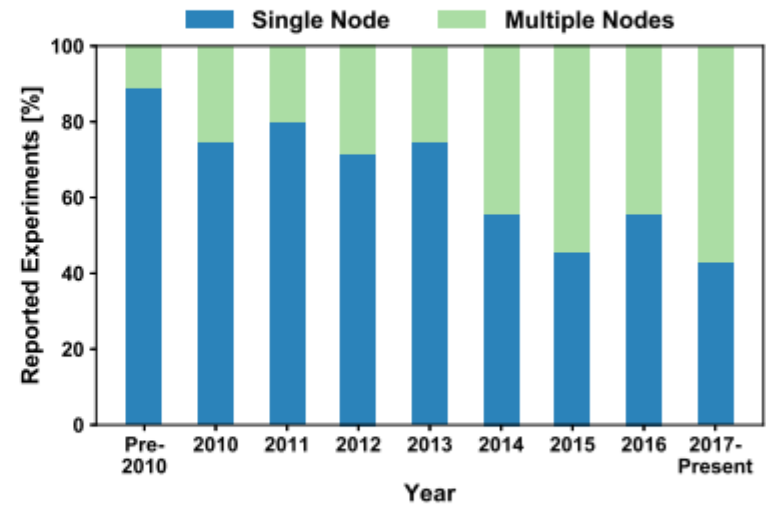
Parallel Deep Learning

65:6

T. Ben-Nun and T. Hoefler



(a) Hardware Architectures



(b) Training with Single vs. Multiple Nodes

Fig. 3. Parallel architectures in deep learning.

Reference Material

- DE Culler, A Gupta and JP Singh, Parallel Computer Architecture: A Hardware/Software Approach Morgan-Kaufmann, 1998.
- A Grama, A Gupta, G Karypis, and V Kumar, Introduction to Parallel Computing. 2nd Ed., Addison-Wesley, 2003.
- Marc Snir, Steve W. Otto, Steven Huss-Lederman, David W. Walker and Jack Dongarra, MPI - The Complete Reference, Second Edition, Volume 1, The MPI Core.
- Bill Gropp, Using MPI, Third Edition, The MIT Press, 2014.
- Research papers