Parallel Computing (CS 633)

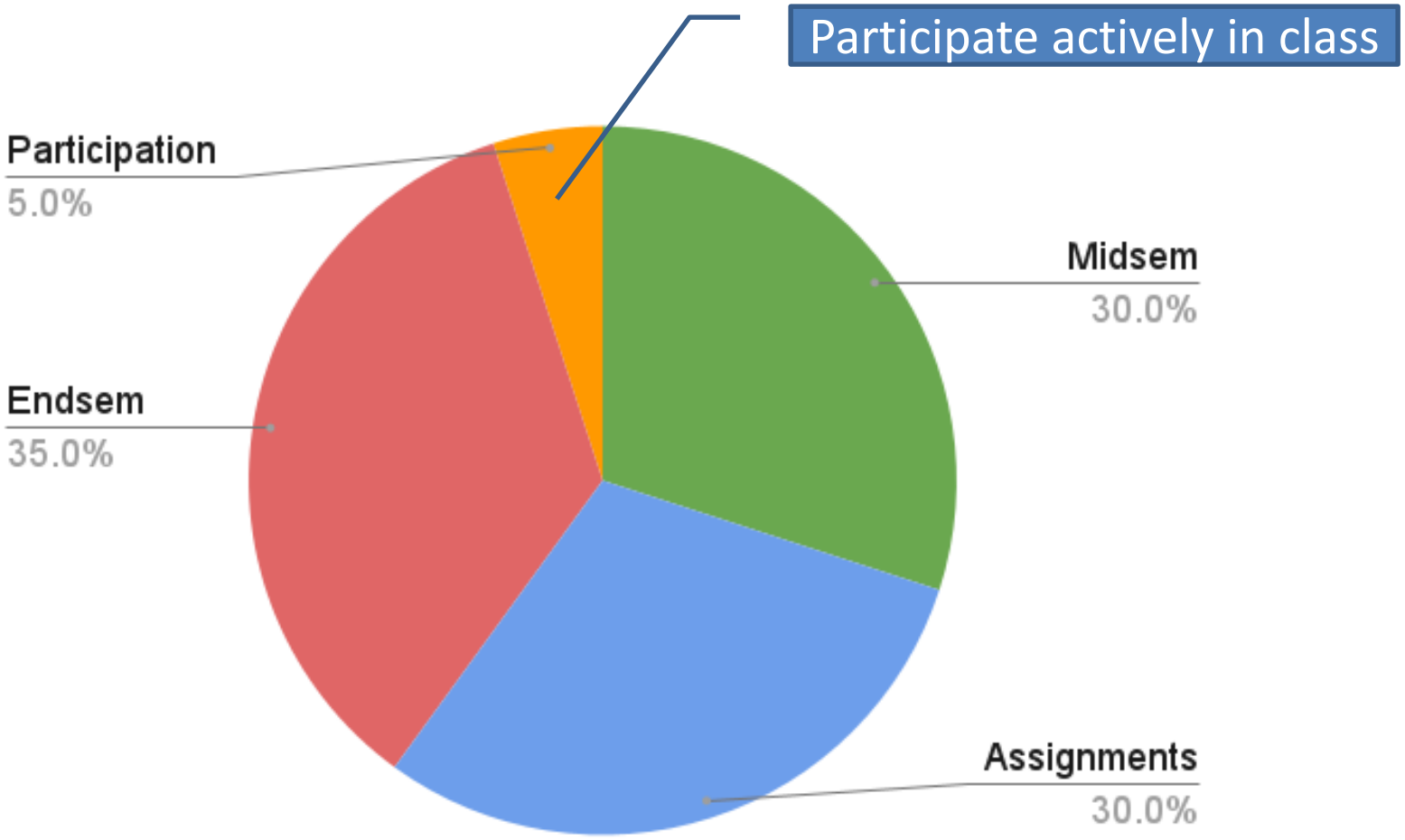
January 8, 2024

Preeti Malakar
pmalakar@cse.iitk.ac.in

Logistics

- Class hours: MW 3:30 – 4:45 PM (RM 101)
- Office hour: MTW 5:00 – 5:30 PM (KD 221)
- https://www.cse.iitk.ac.in/users/cs633/2023-24-2
  - Lectures will be uploaded after every class
- Announcements/uploads on
  - MooKIT
  - Course email alias
- Email to the instructor should always be prefixed with [CS633] in the subject

Grading Policy



Switch OFF All Devices

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

ACADEMIC SECTION

No. DOAA/20010-11/Mobile

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

Assignments

- Programming assignments in C
- In a group (group size = 3)
  - Send group member information by Jan 14 to {gsarkar,madhavm}@cse.iitk.ac.in
  - Include clearly names, roll numbers, IITK email-ids
  - Subject of email [CS633 Group]
  - Change in group formation is not allowed
- Mode of submission will be explained in due time

Assignments

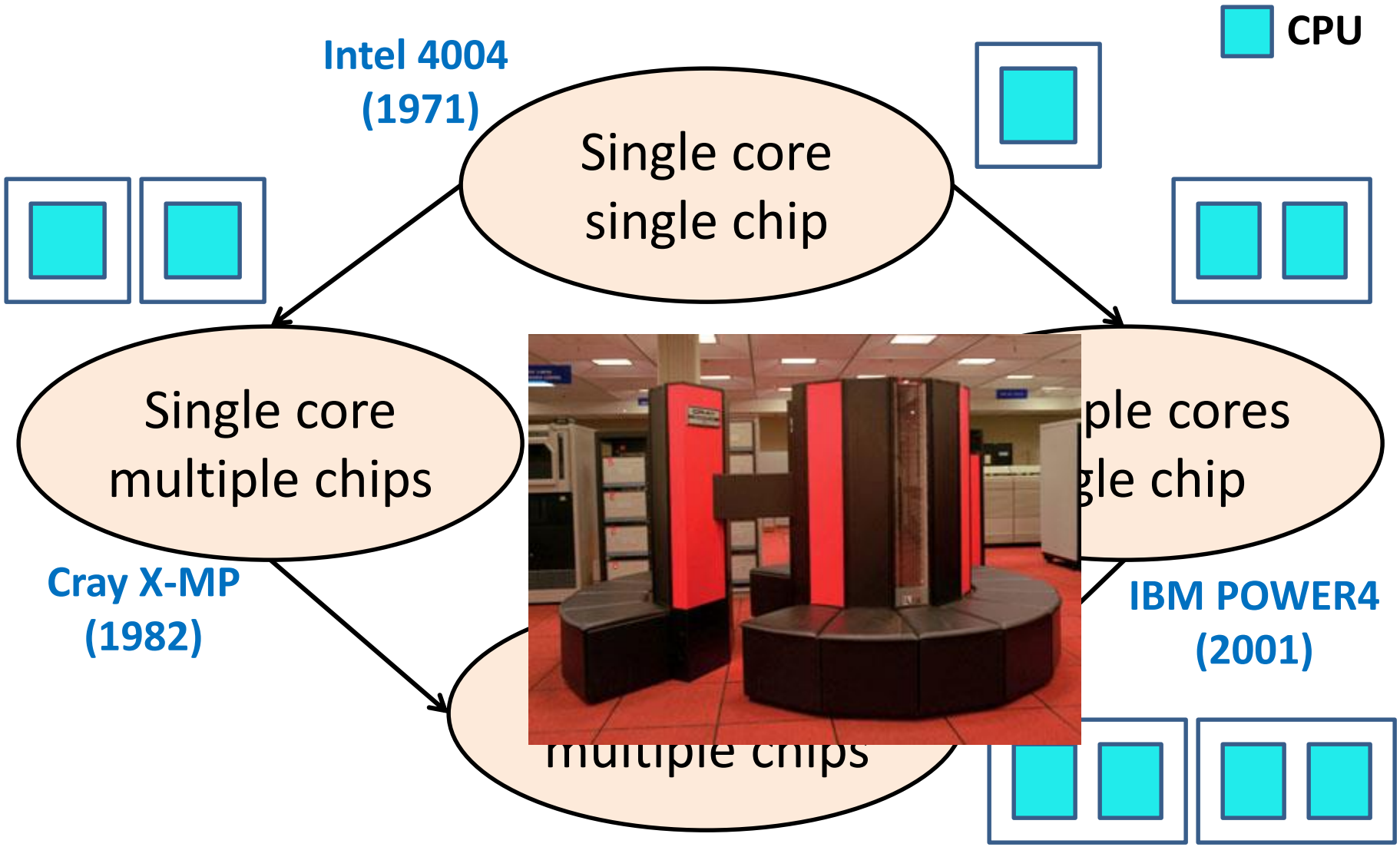
- Credit for early submissions (+5 / day)
  - Max credit: +15 / assignment
  - Last date of submission will be considered only
- Score reduction for late submissions (-3 / day)
  - Max 2 late days / assignment
- None of the assignments can be completed in a day!

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

Lecture 1

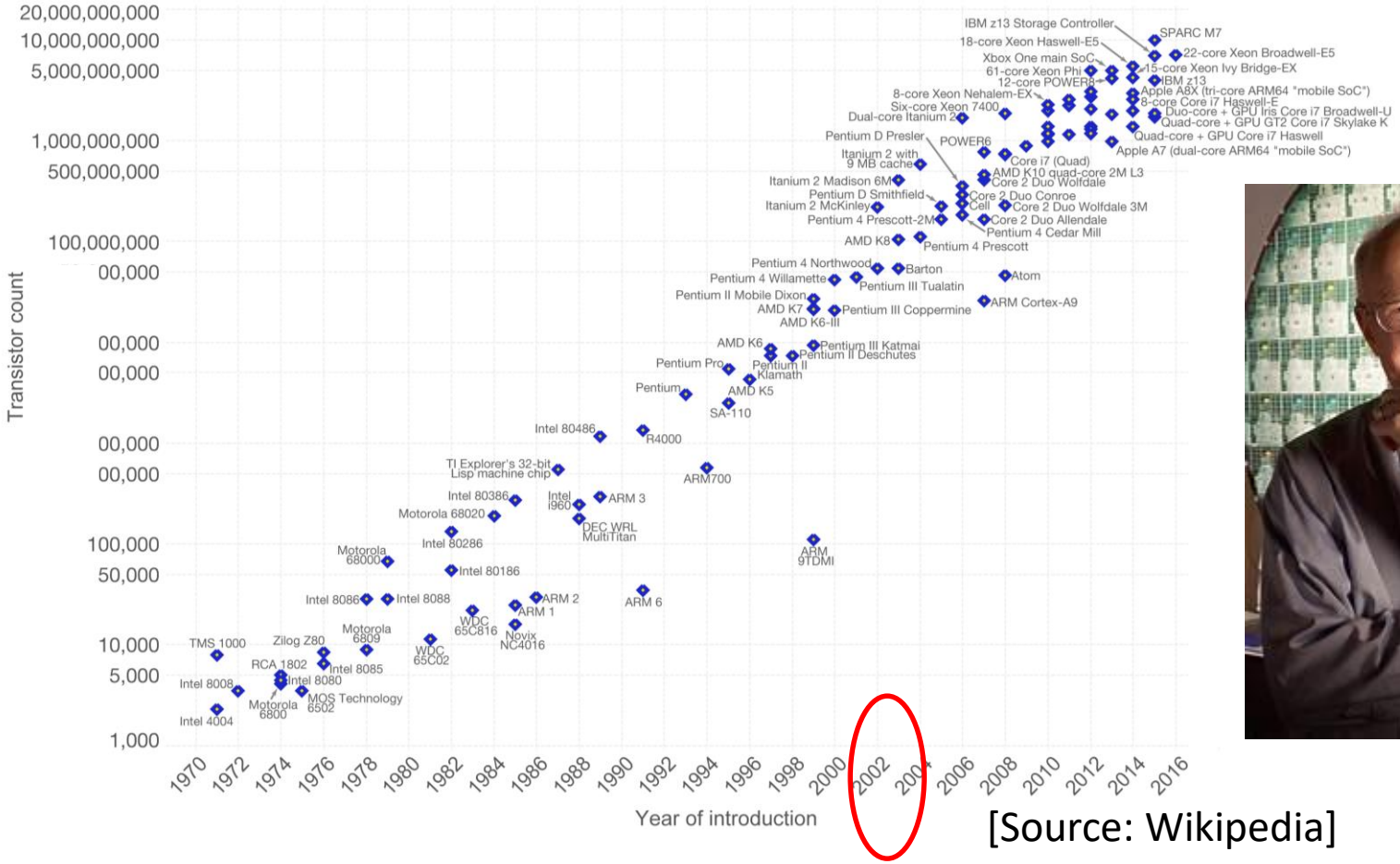
Introduction

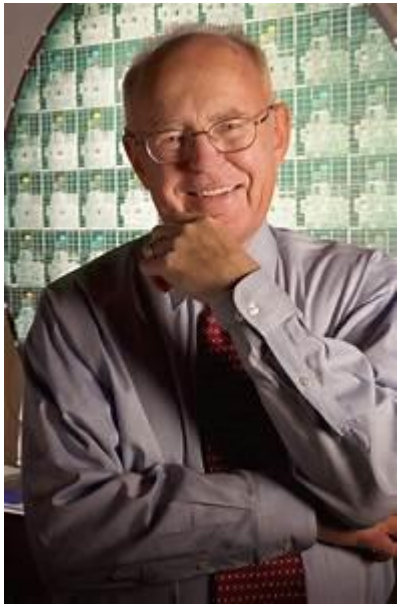
Multicore Era



Moore's Law (1965)

Number of transistors in a chip doubles every 18 months

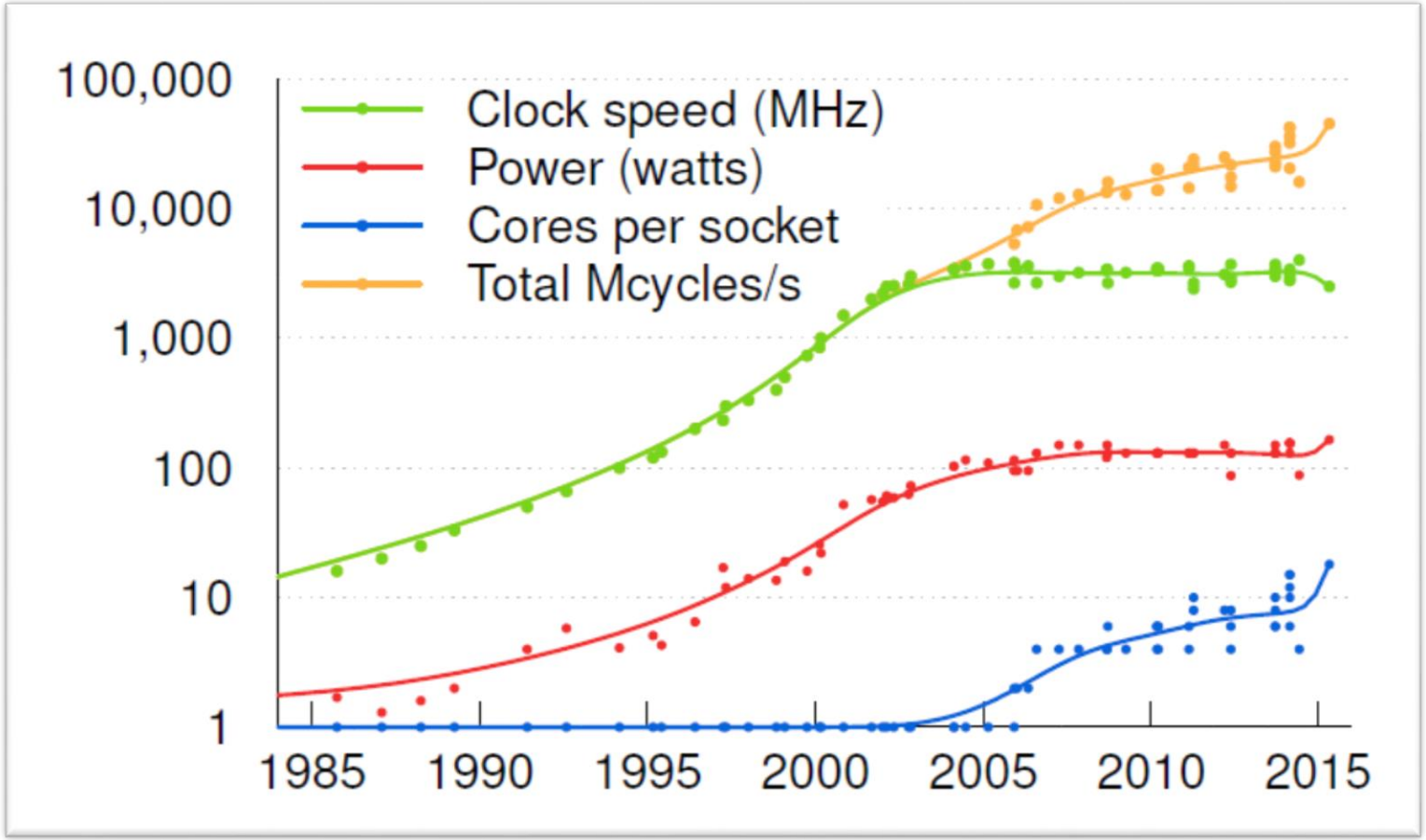




[Source: Wikipedia]

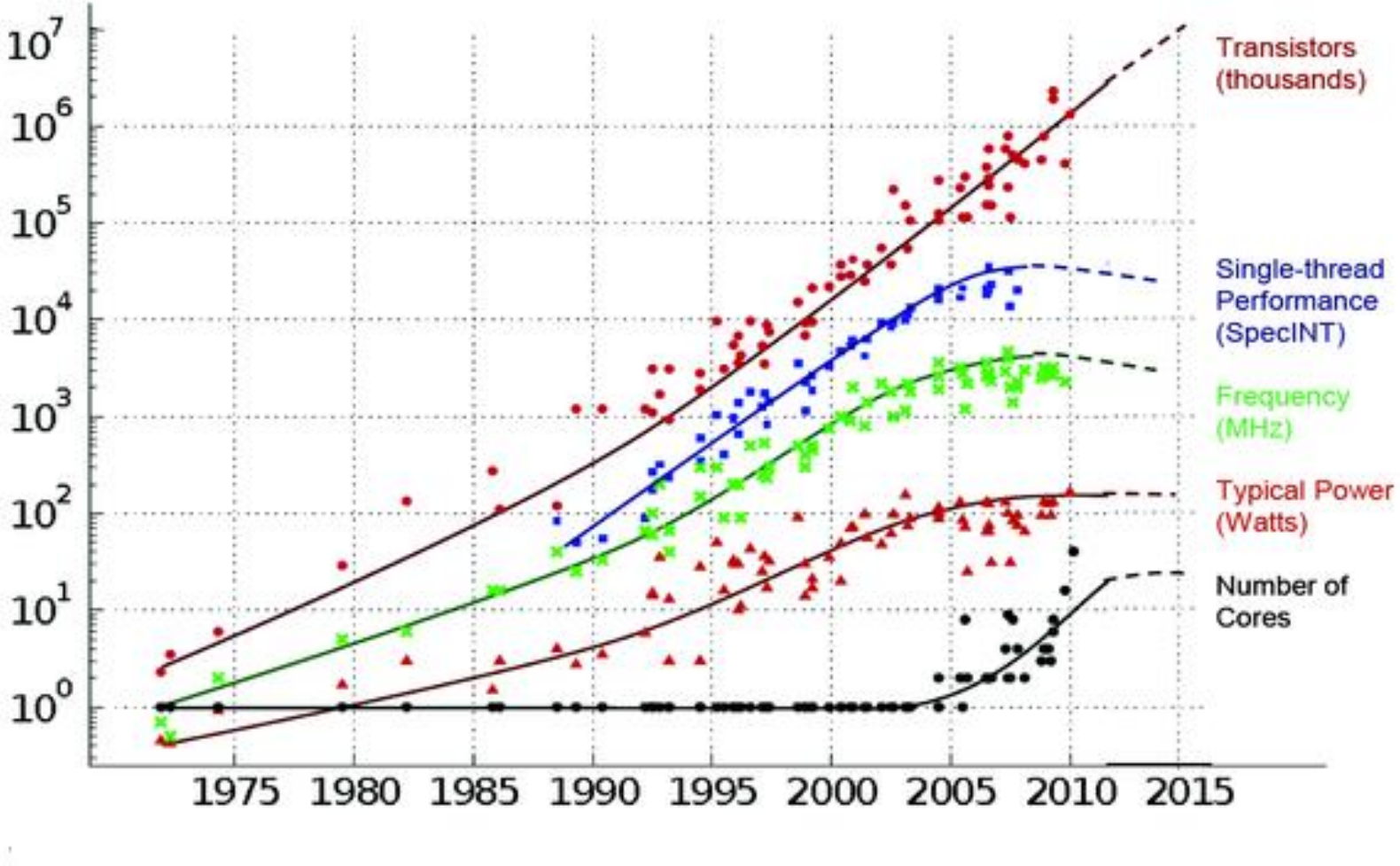
"However, it must be programmed with a more complicated parallel programming model to obtain maximum performance."

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'23)

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
			561.20		
			442.01	537.21	29,899

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

green500.org (Nov'23)

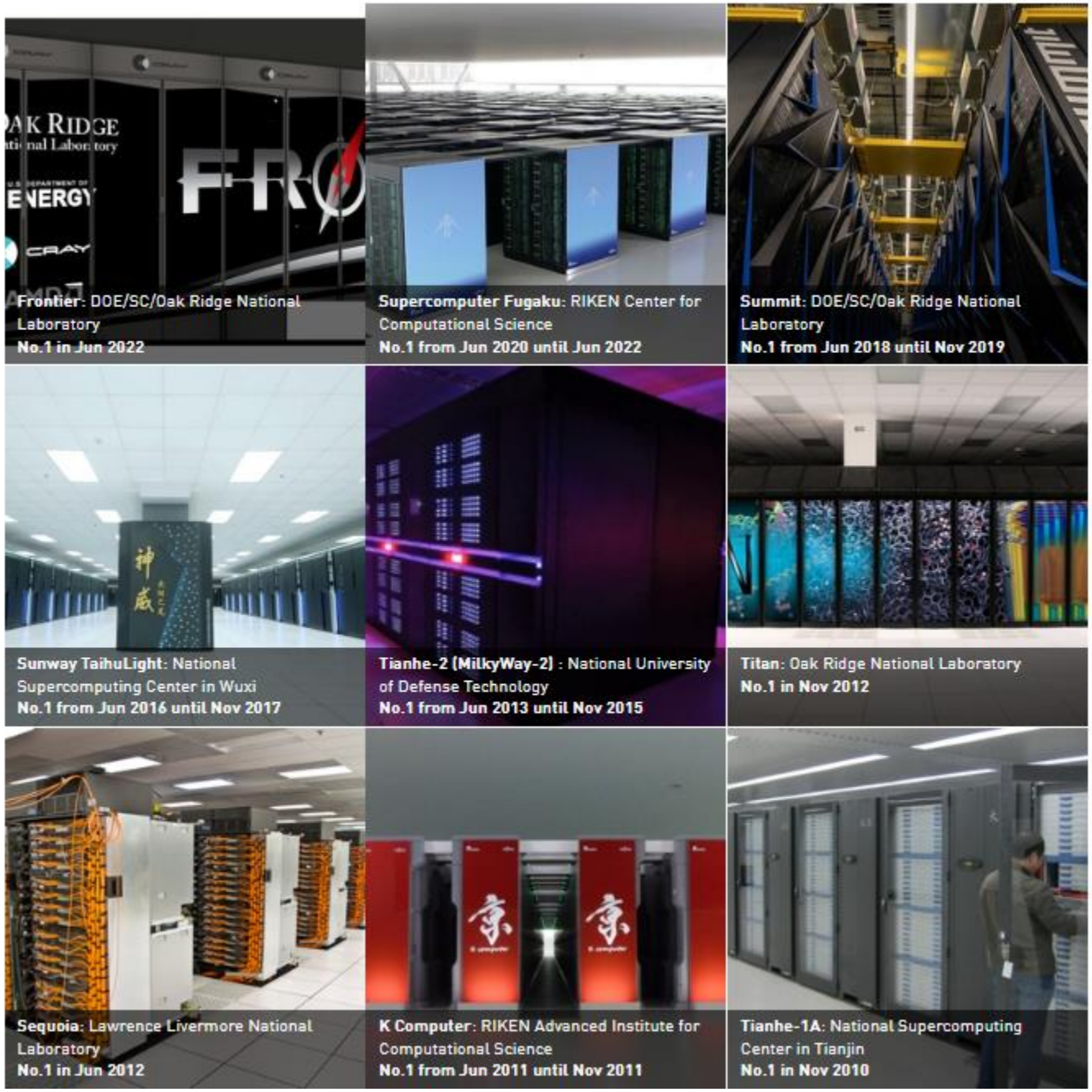
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	293	**Henri** - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States	8,288	2.88	44	65.396
2	44	**Frontier TDS** - HPE Cray EX235a, 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

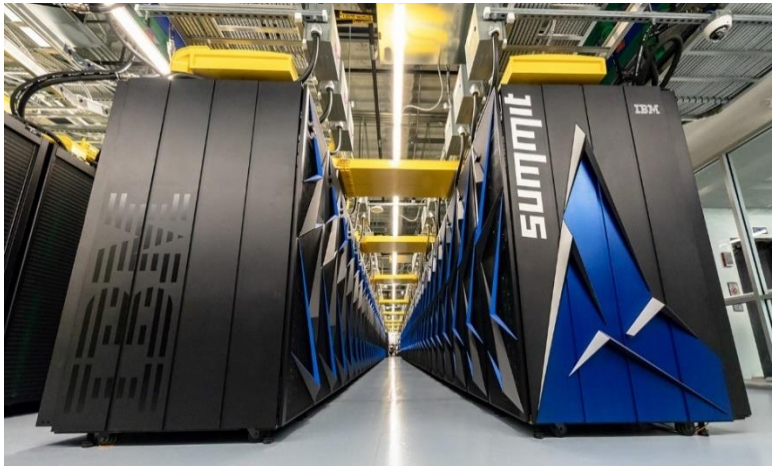
RANK	PREVIOUS RANK	MACHINE	VENDOR	TYPE	NETWORK	INSTALLATION SITE	LOCATION
1	new	Supercomputer Fugaku	Fujitsu	Fujitsu A64FX	Tofu Interconnect D	RIKEN Center for Computational Science (R-CCS)	Kobe Hyogo
2	new	Wuhan Supercomputer	HUST	Kunpeng 920+Tesla A100	Custom	Wuhan Supercomputing Center	Wuhan
3	3	Frontier	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

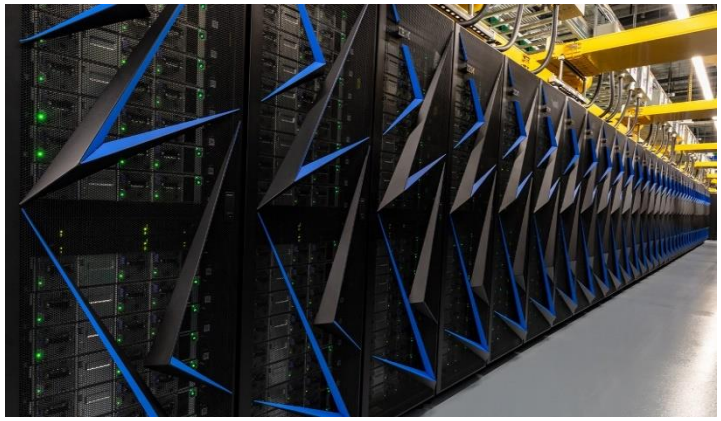
Top #1 supercomputer

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



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.

https://www.science.org/content/article/climate-change-threatens-supercomputers

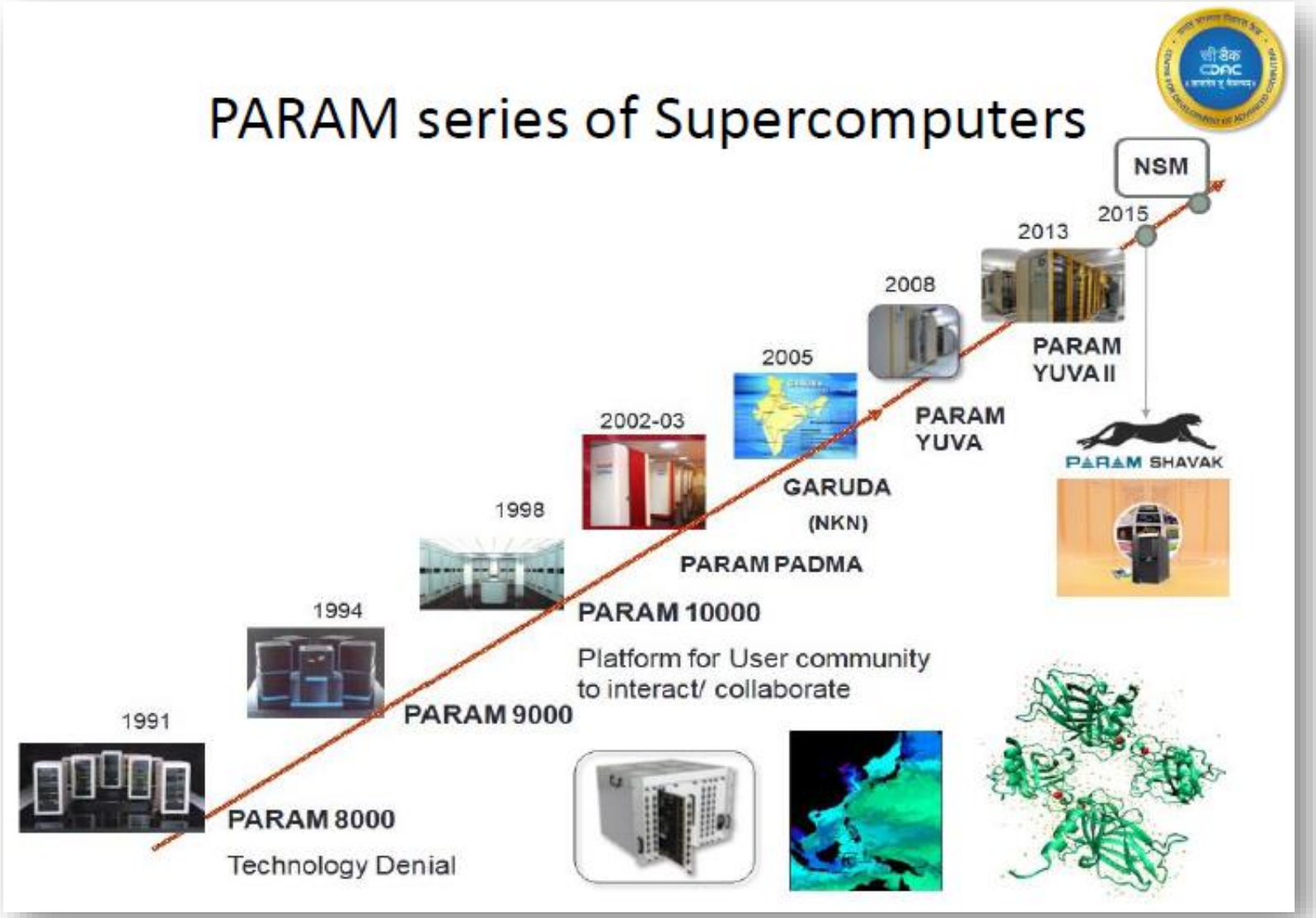
Top Supercomputers from India

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

Supercomputing in India [topsc.cdacb.in, Jul'23]

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	Centre for Development of Advanced Computing(C-DAC),Pune	PARAM SIDDHI is a NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband .consists of 42 nos of NVIDIA DGX A100 nodes, Dual socket populated with More Info OEM:NVIDIA Bidder: ATOS	41664/2/42	4619	5267.14
3	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
4	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
5	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

Source: www.iitk.ac.in



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

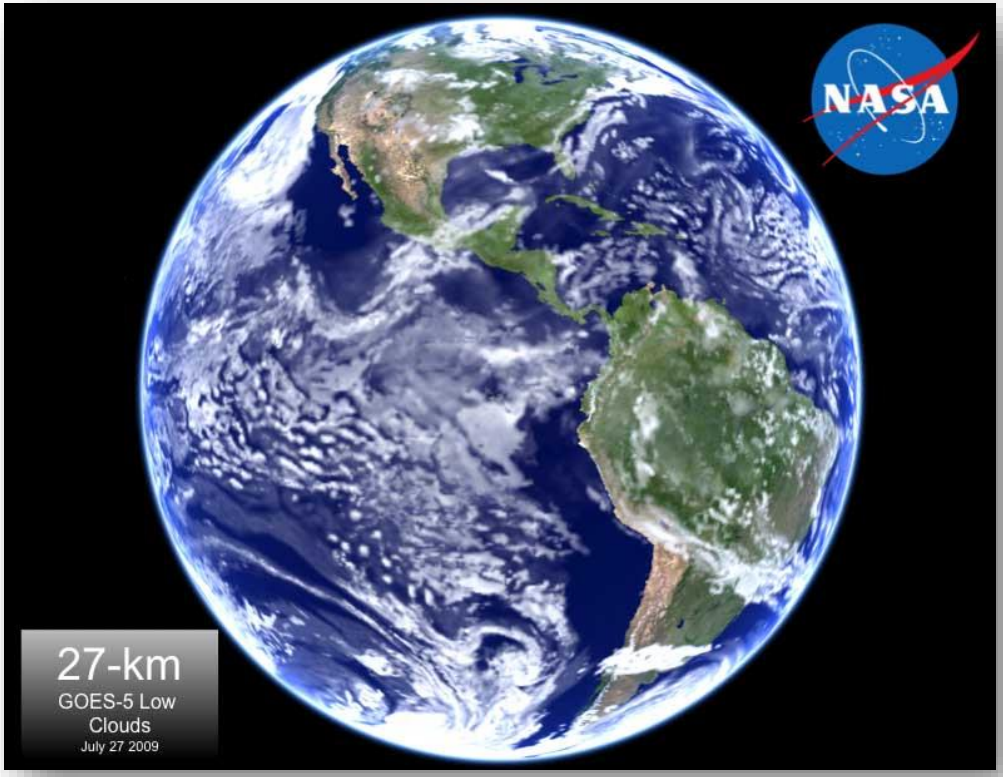
PARAM Smriti

PARAM Utkarsh

PARAM Ganga

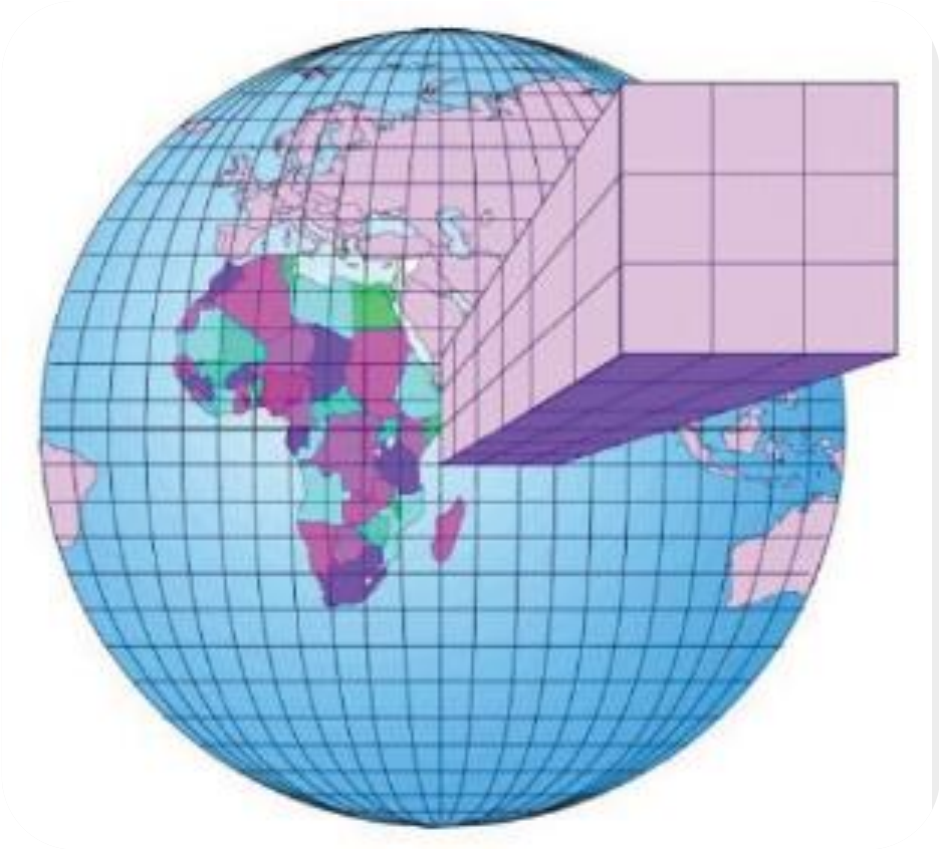
Big Compute

Massively Parallel Codes



Climate simulation of Earth [Credit: NASA]

Discretization

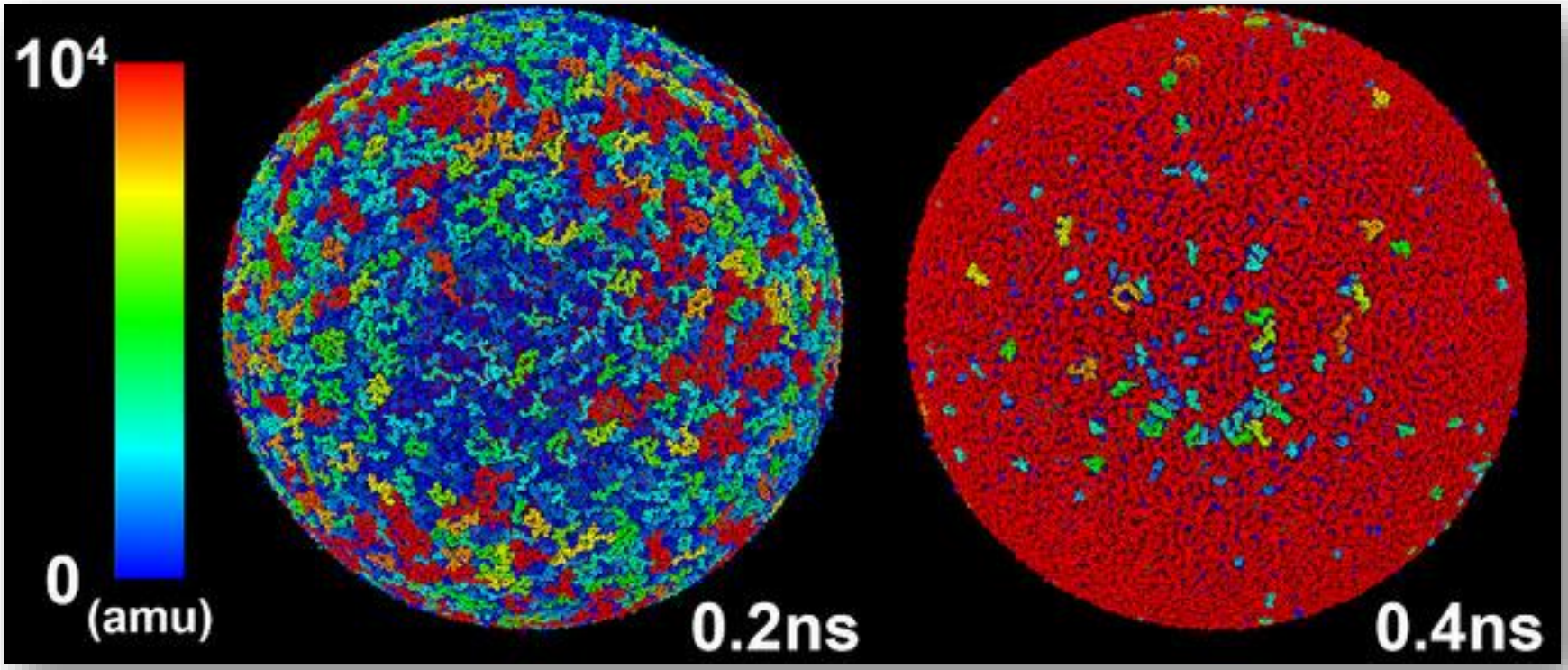


Gridded mesh for a global model [Credit: Tompkins, ICTP]

Numerical Weather Models

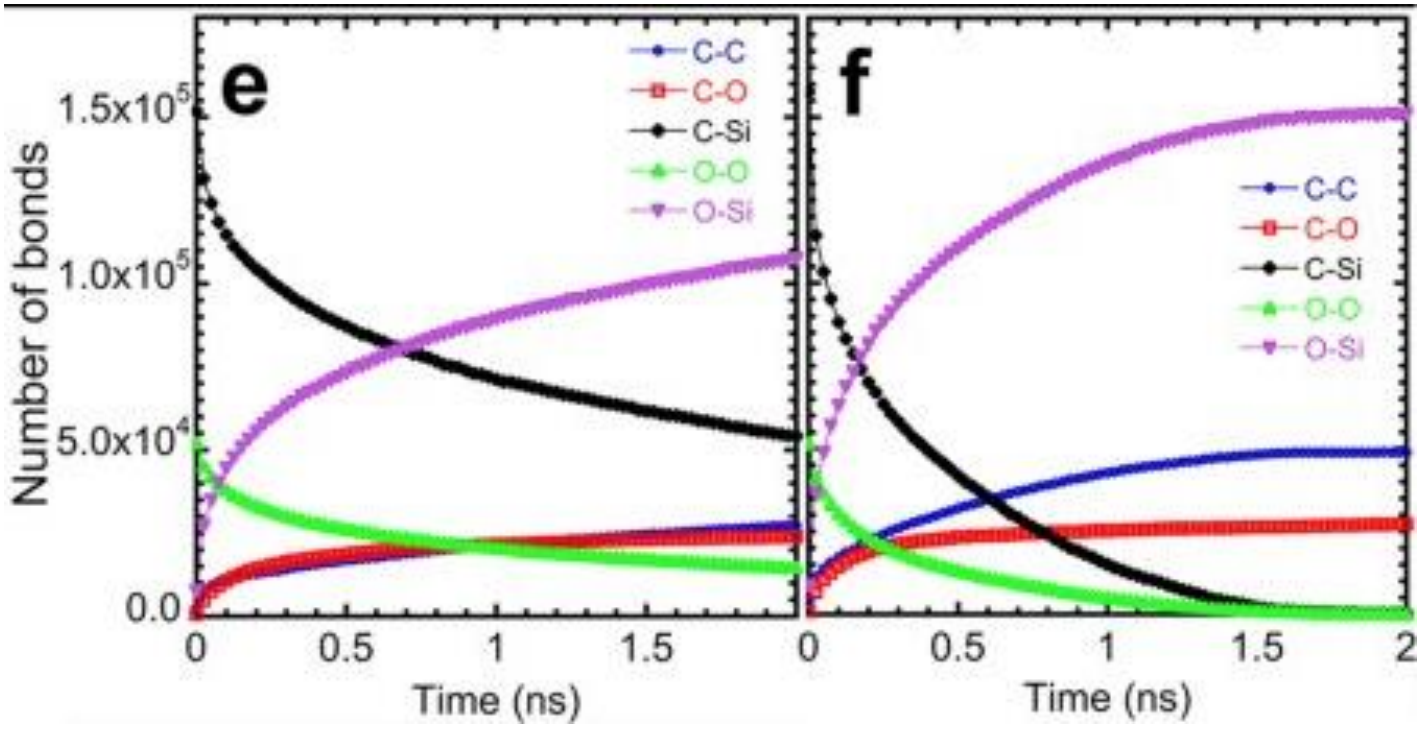
- Use numerical methods to solve equations that govern atmospheric processes
- Are based on fluid dynamics and depend on observations of meteorological variables
- Are used to obtain nowcast/forecast

Massively Parallel Simulations



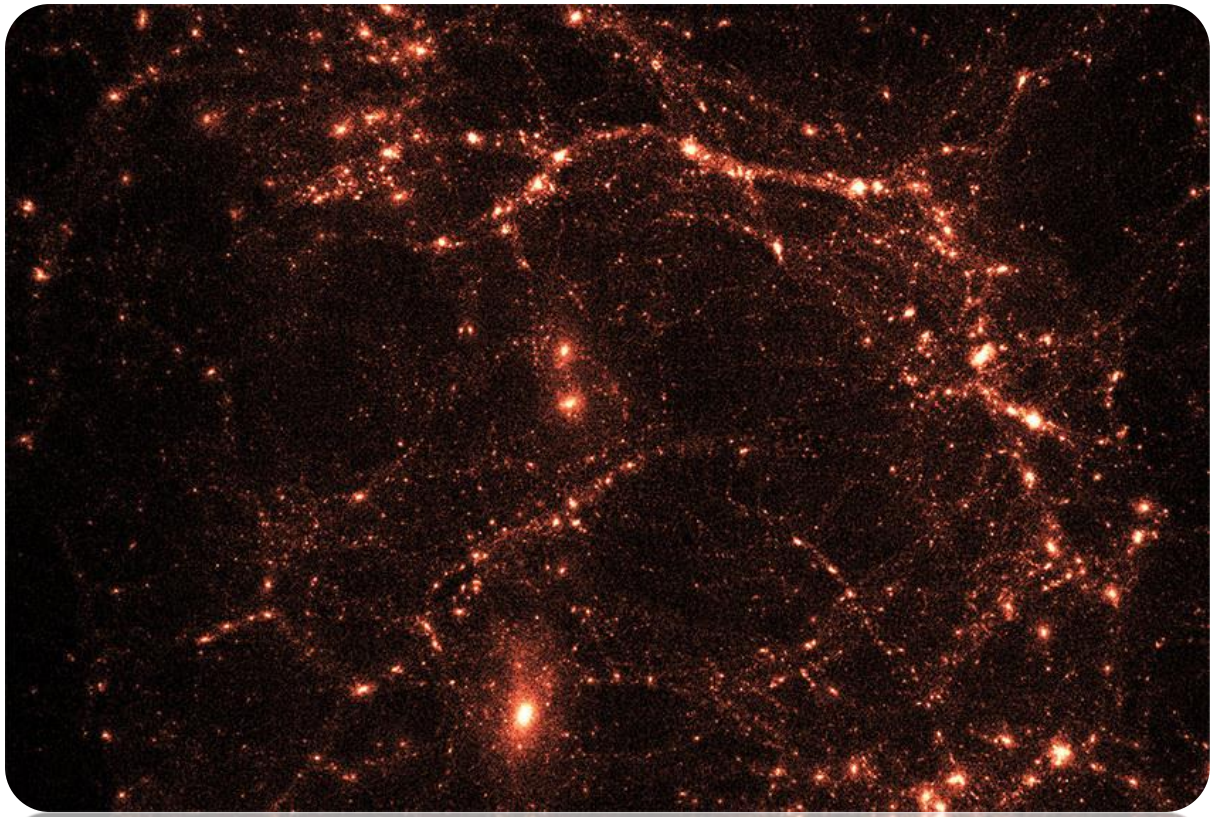
Self-healing material simulation
[Nomura et al., "Nanocarbon synthesis by high-temperature oxidation of nanoparticles", Scientific Reports, 2016]

Massively Parallel Analysis



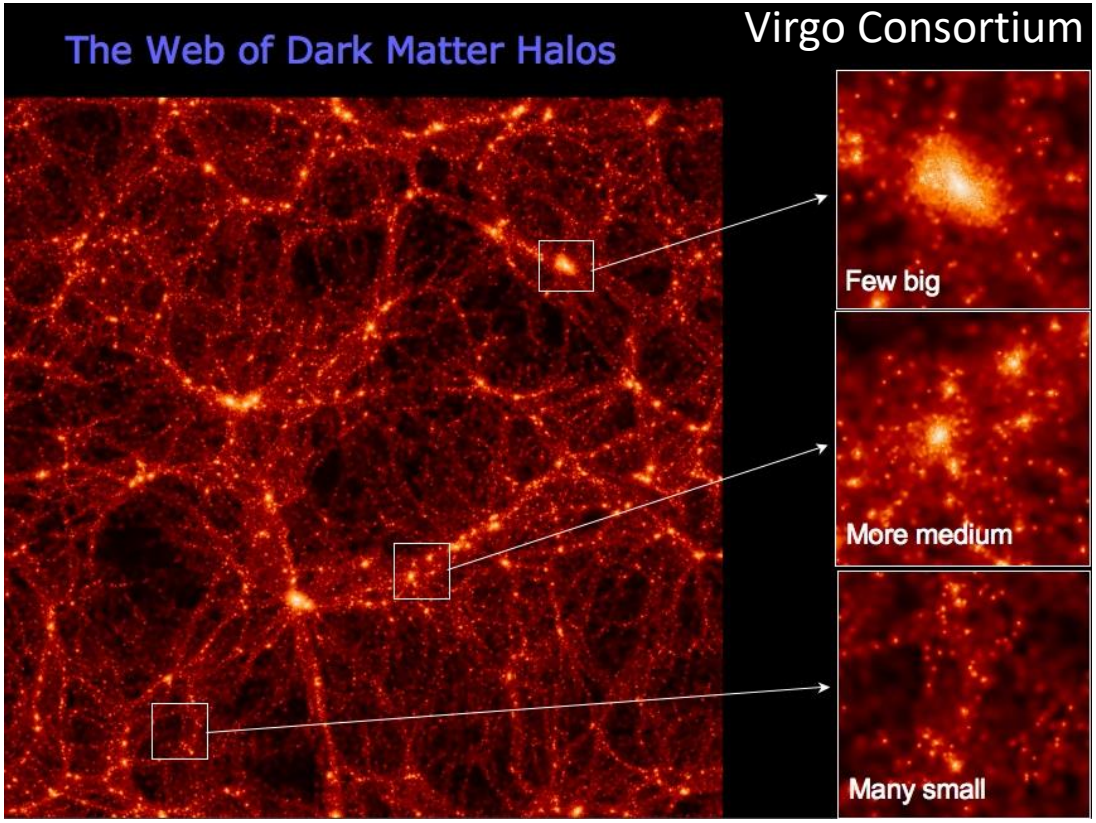
[Nomura et al., "Nanocarbon synthesis by high-temperature oxidation of nanoparticles", Scientific Reports, 2016]

Massively Parallel Codes

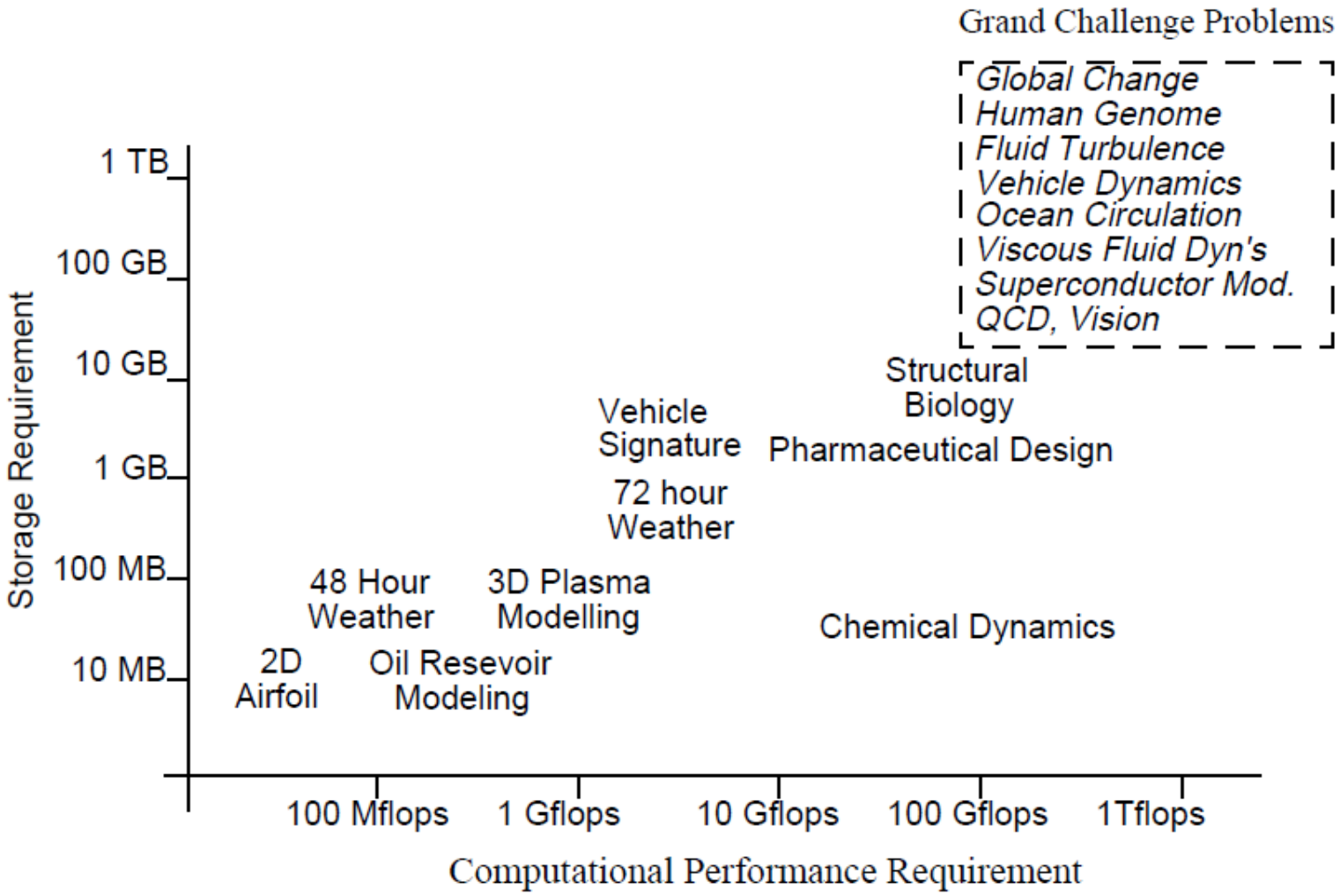


Cosmological simulation [Credit: ANL]

Massively Parallel Analysis



Computational Science

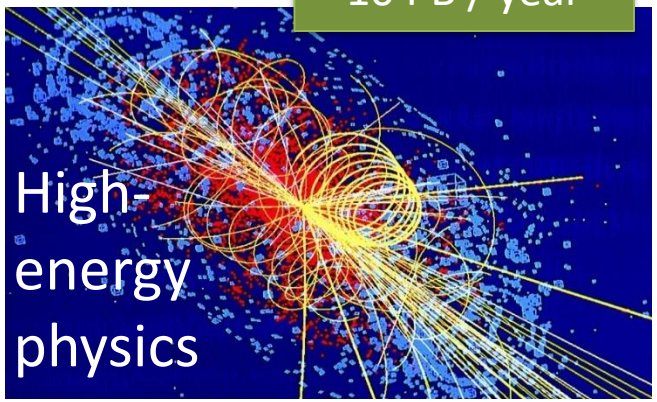


[Source: Culler, Singh and Gupta]

Big Data

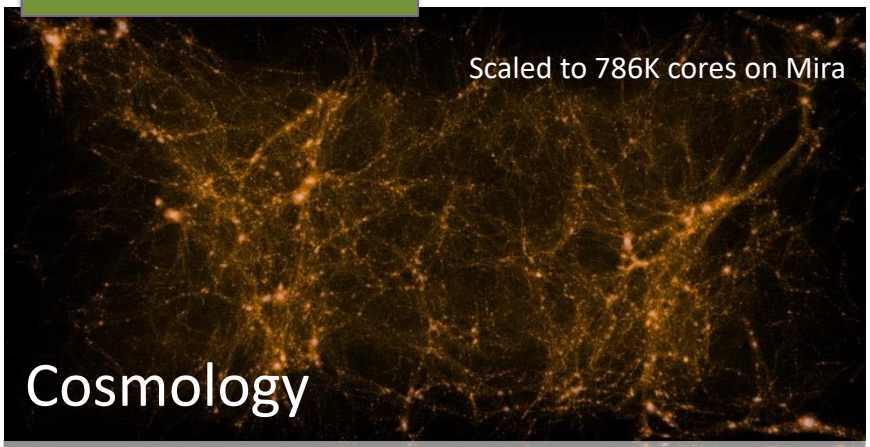
Output Data

10 PB / year



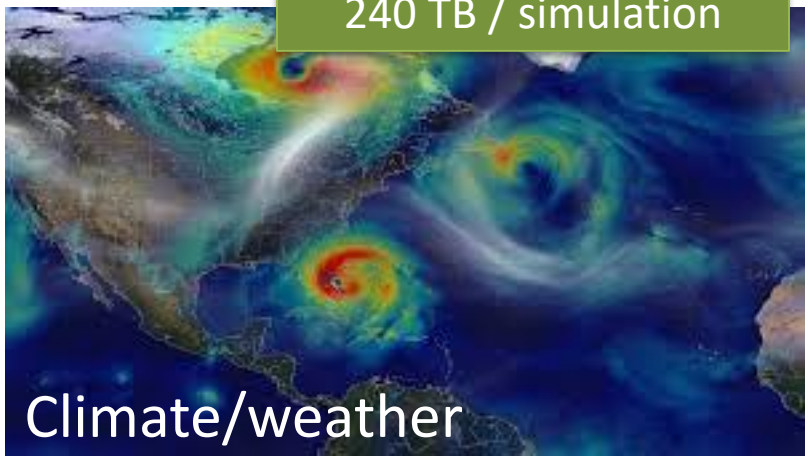
Higgs boson simulation
Source: CERN

2 PB / simulation



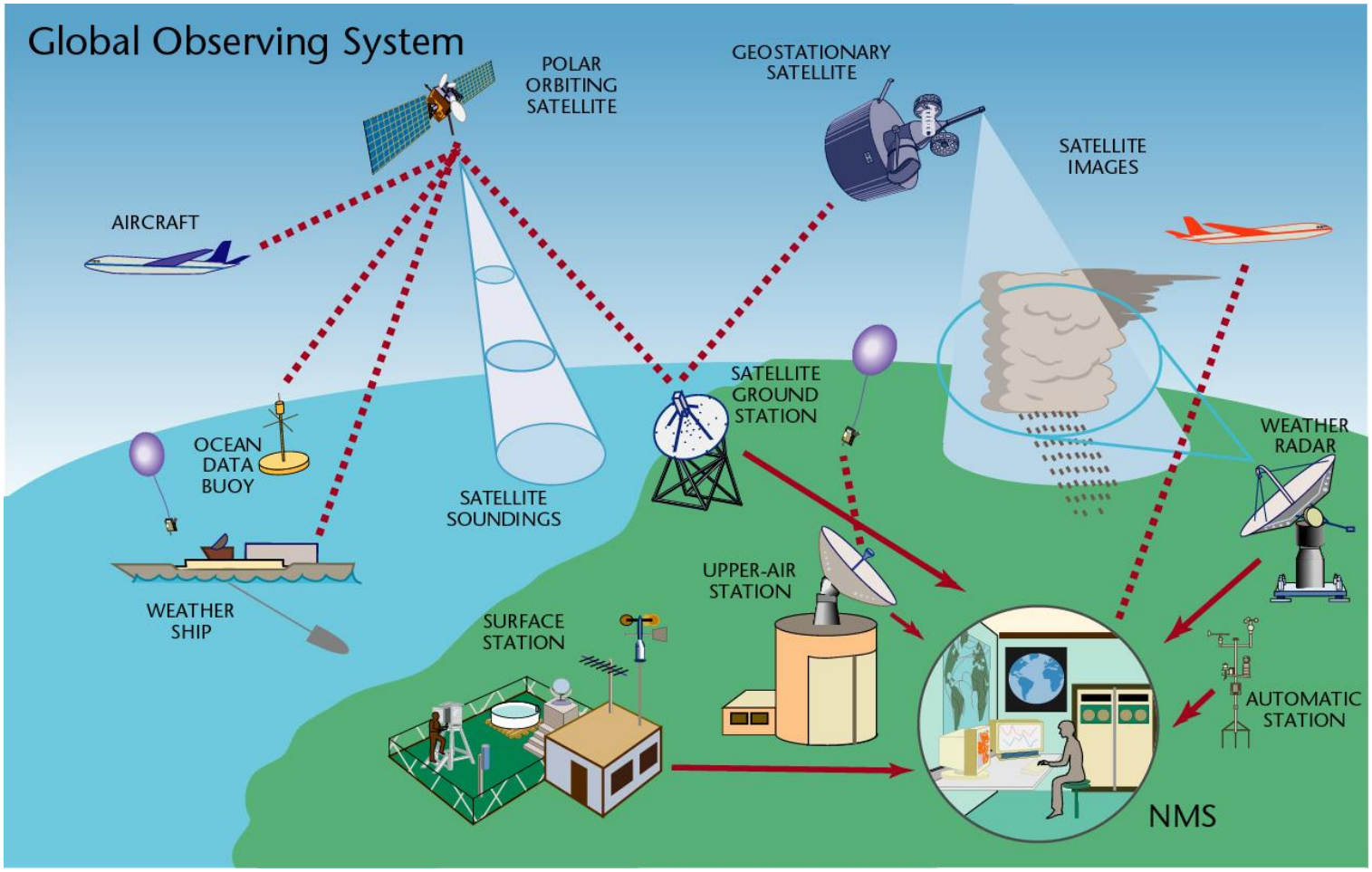
Q Continuum simulation
Source: Salman Habib et al.

240 TB / simulation



Hurricane simulation
Source: NASA

Input Data



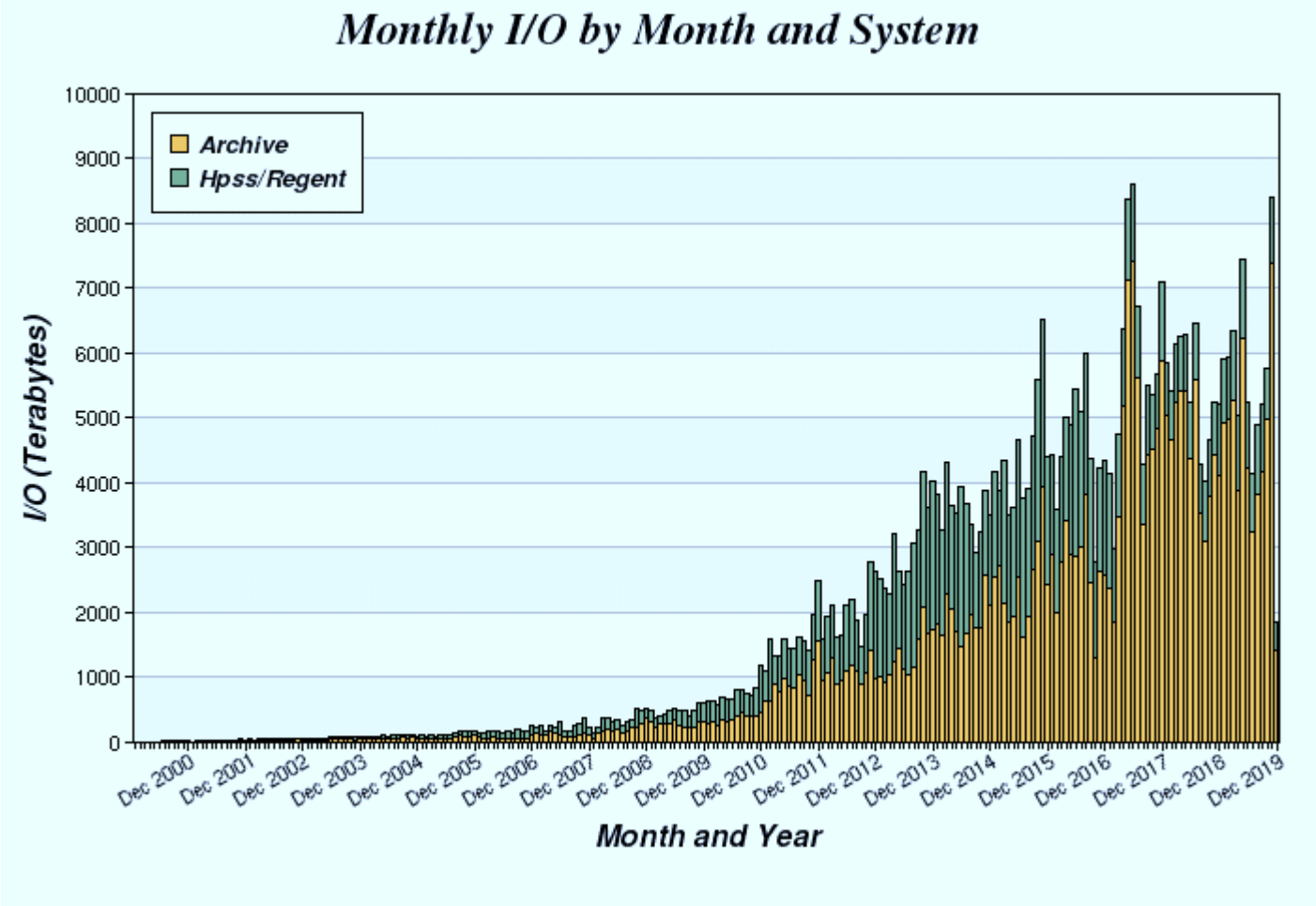
[Credit: World Meteorological Organization]

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)	

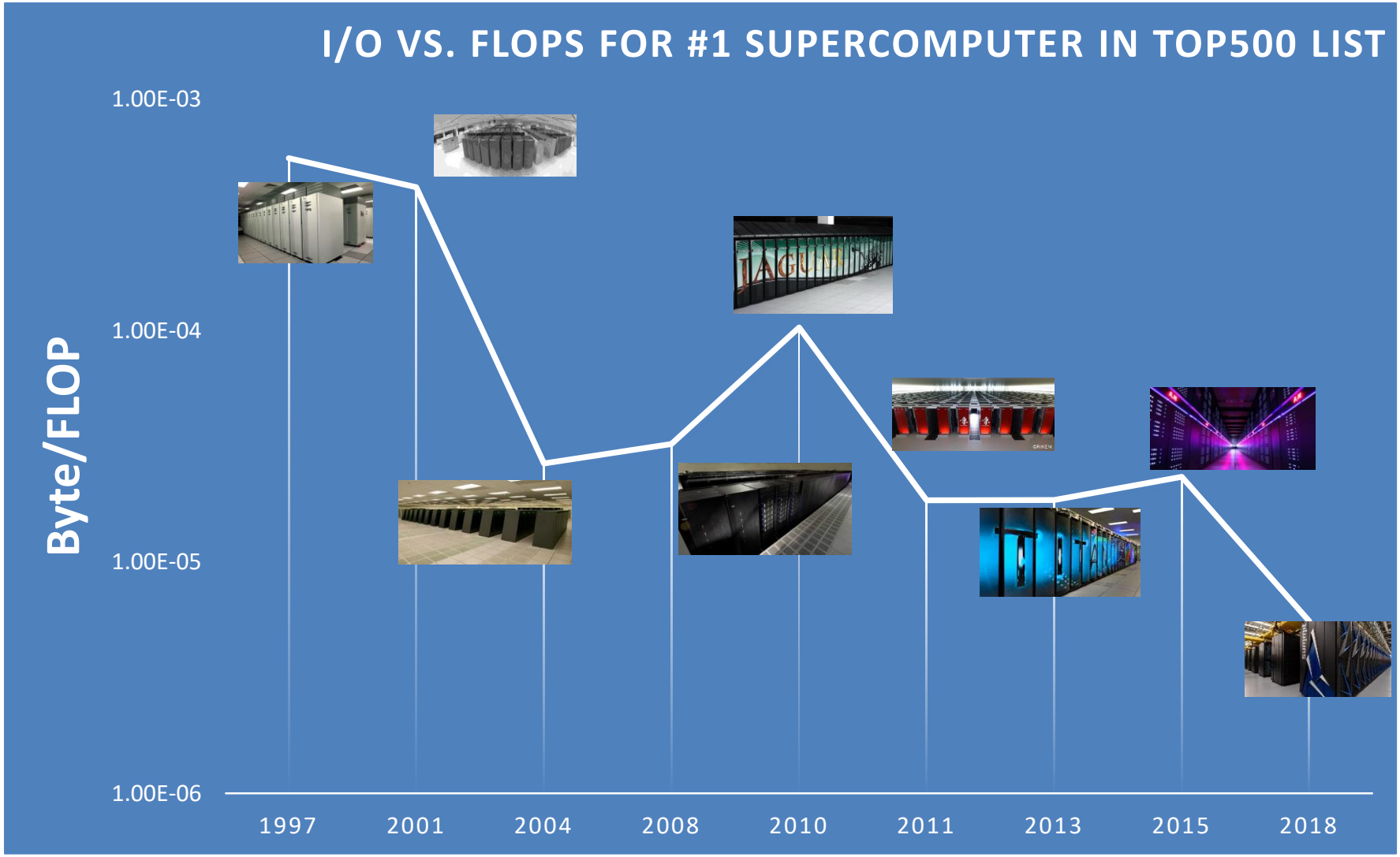
[Credit: Pavan Balaji@ATPESC'17]

I/O trends

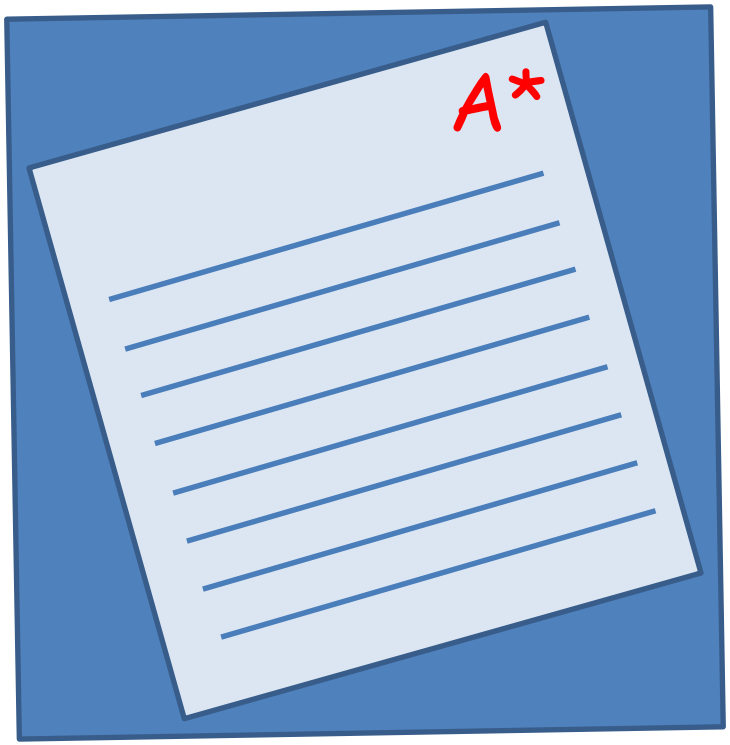


NERSC I/O trends [Credit: www.nersc.gov]

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

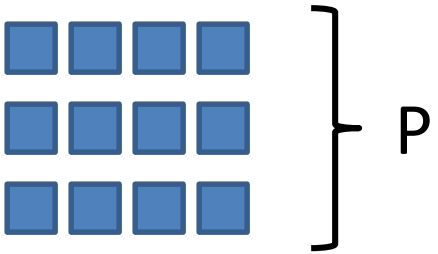
Serial

```
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

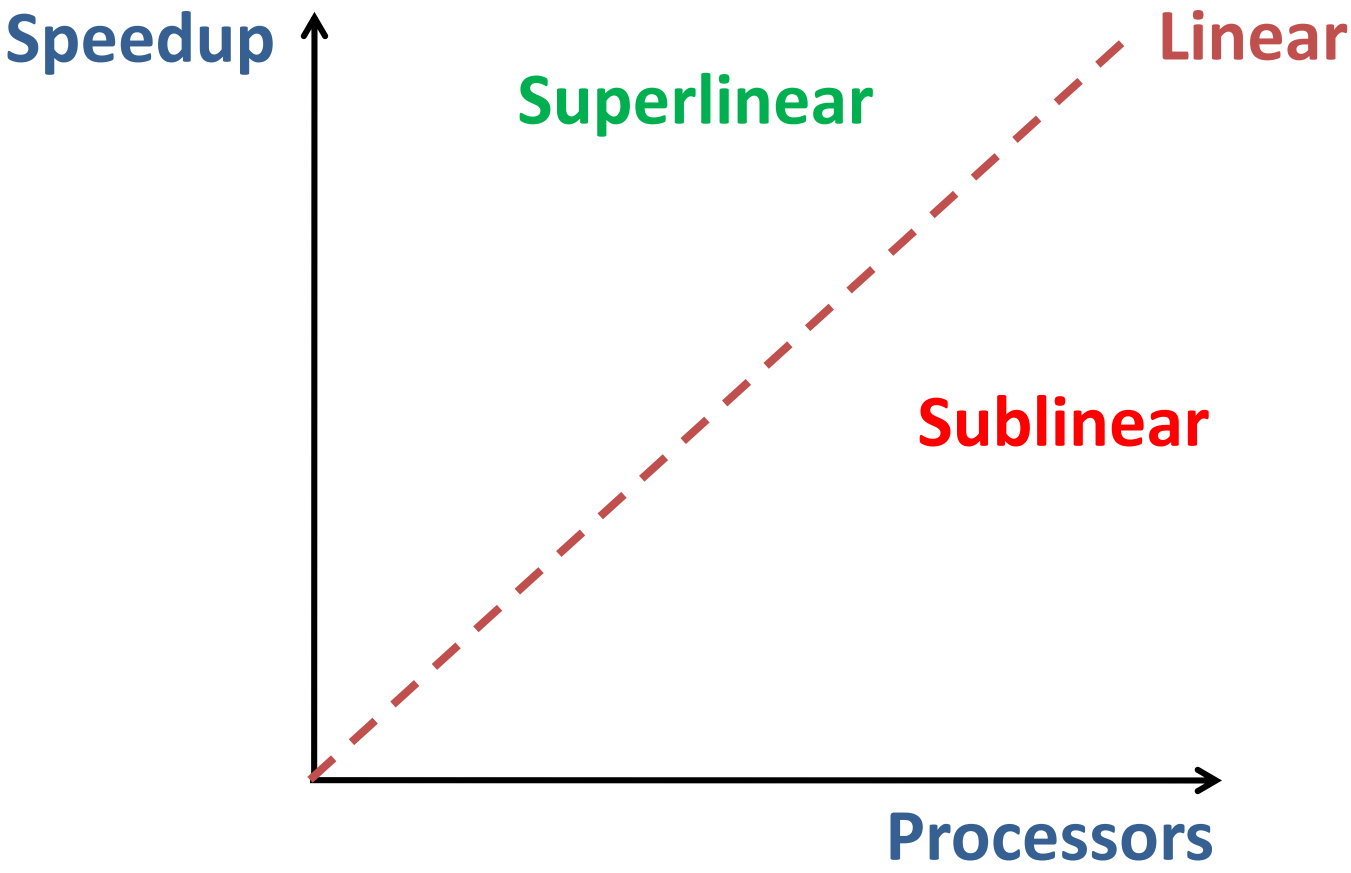
$$E_P = \frac{S_P}{P}$$

Parallel Performance (Parallel Sum)

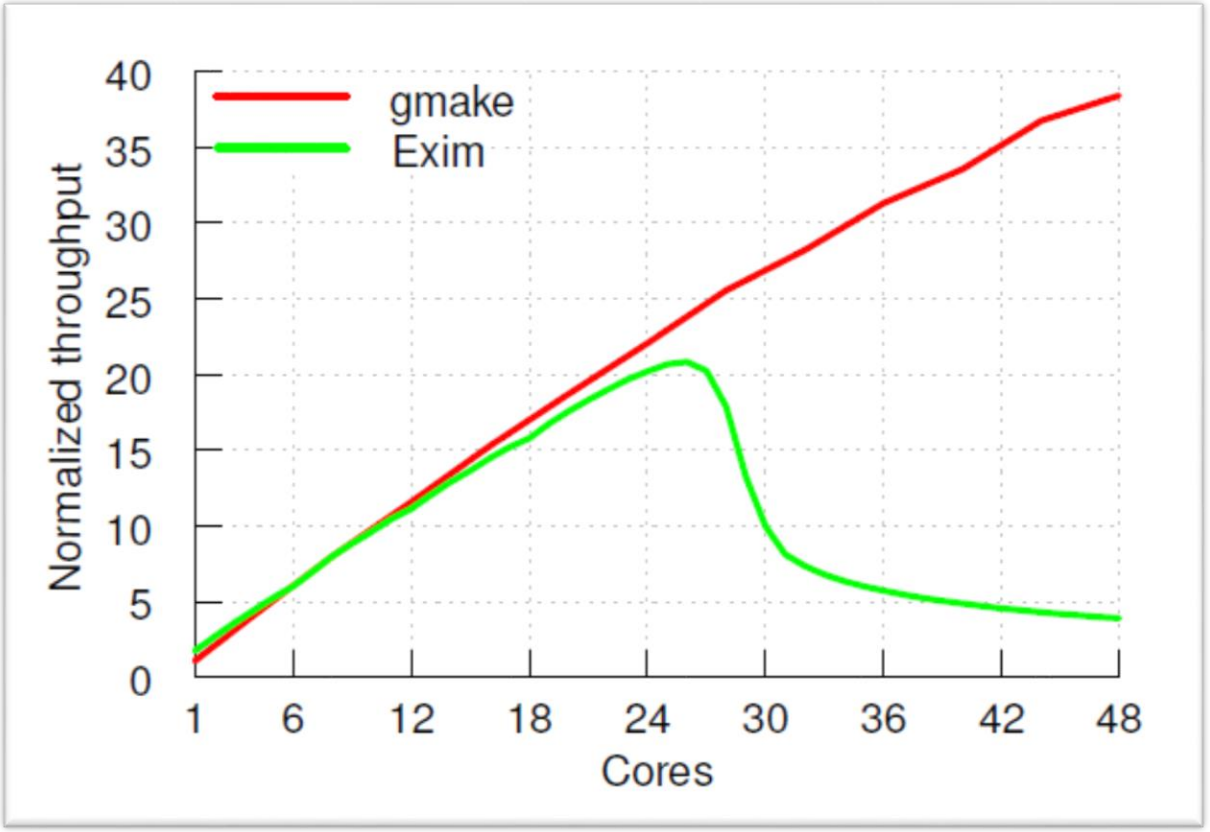
Parallel efficiency of summing 10^7 doubles

#Processes	Time (sec)	Speedup
1	0.025	1
2	0.013	1.9
4	0.010	2.5
8	0.009	2.8
12	0.007	3.6

Ideal Speedup

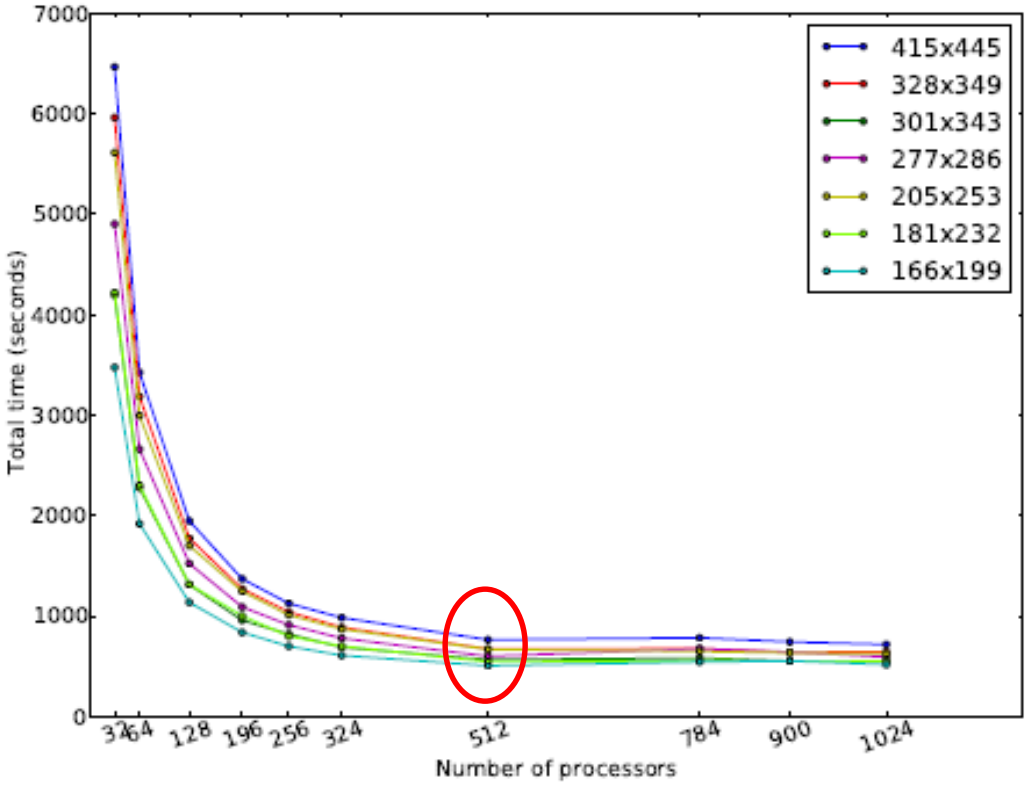


Issue – Scalability



[Source: M. Frans Kaashoek, MIT]

Scalability Bottleneck



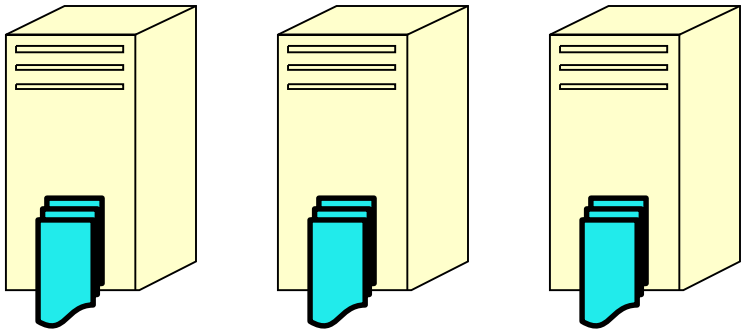
Performance of weather simulation application

Parallelism

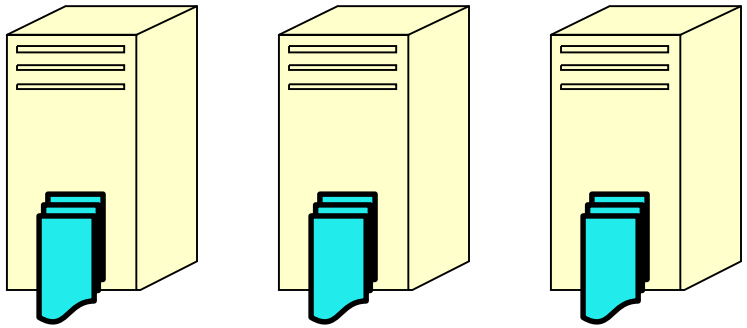
A parallel computer is a collection of processing elements that communicate and cooperate to solve large problems fast.

– 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

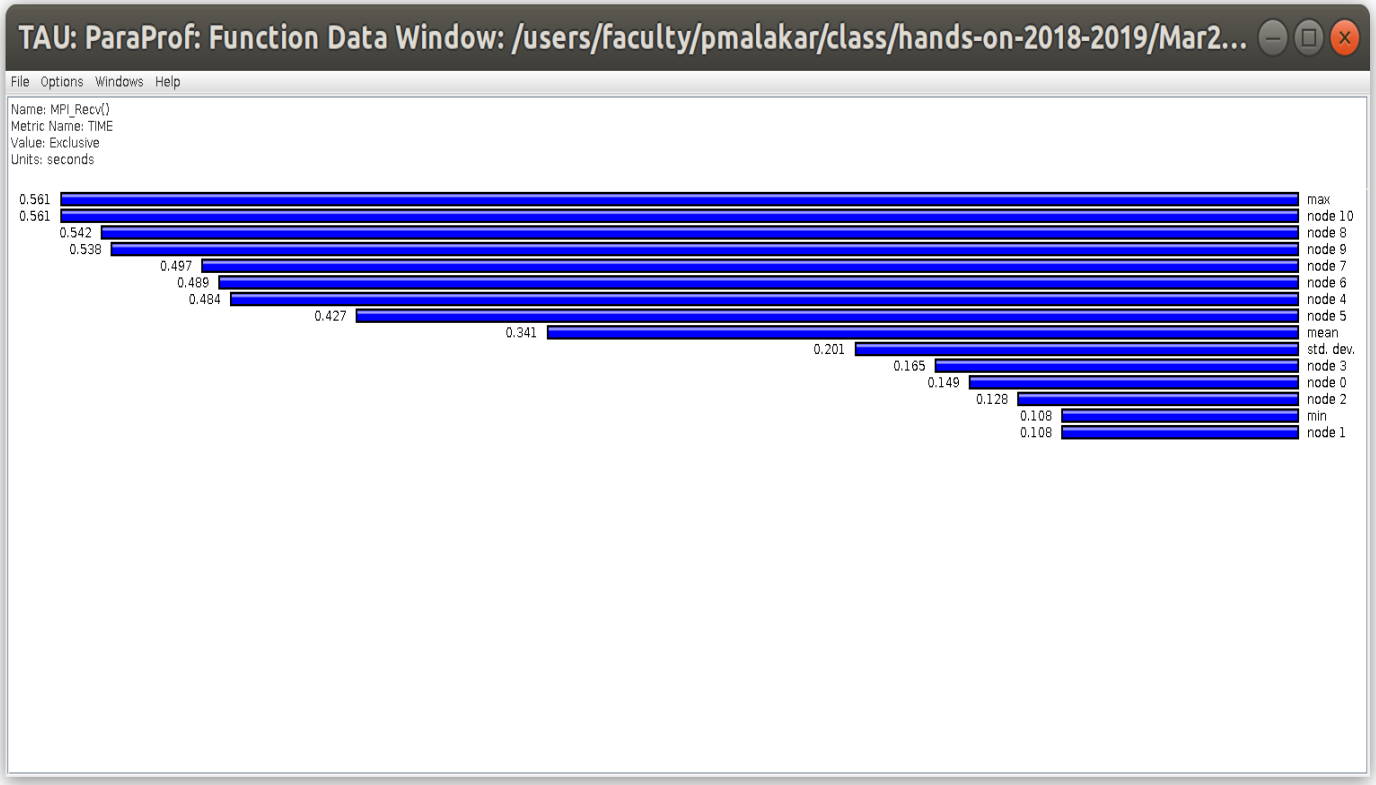
Designing parallel codes

Performance analysis

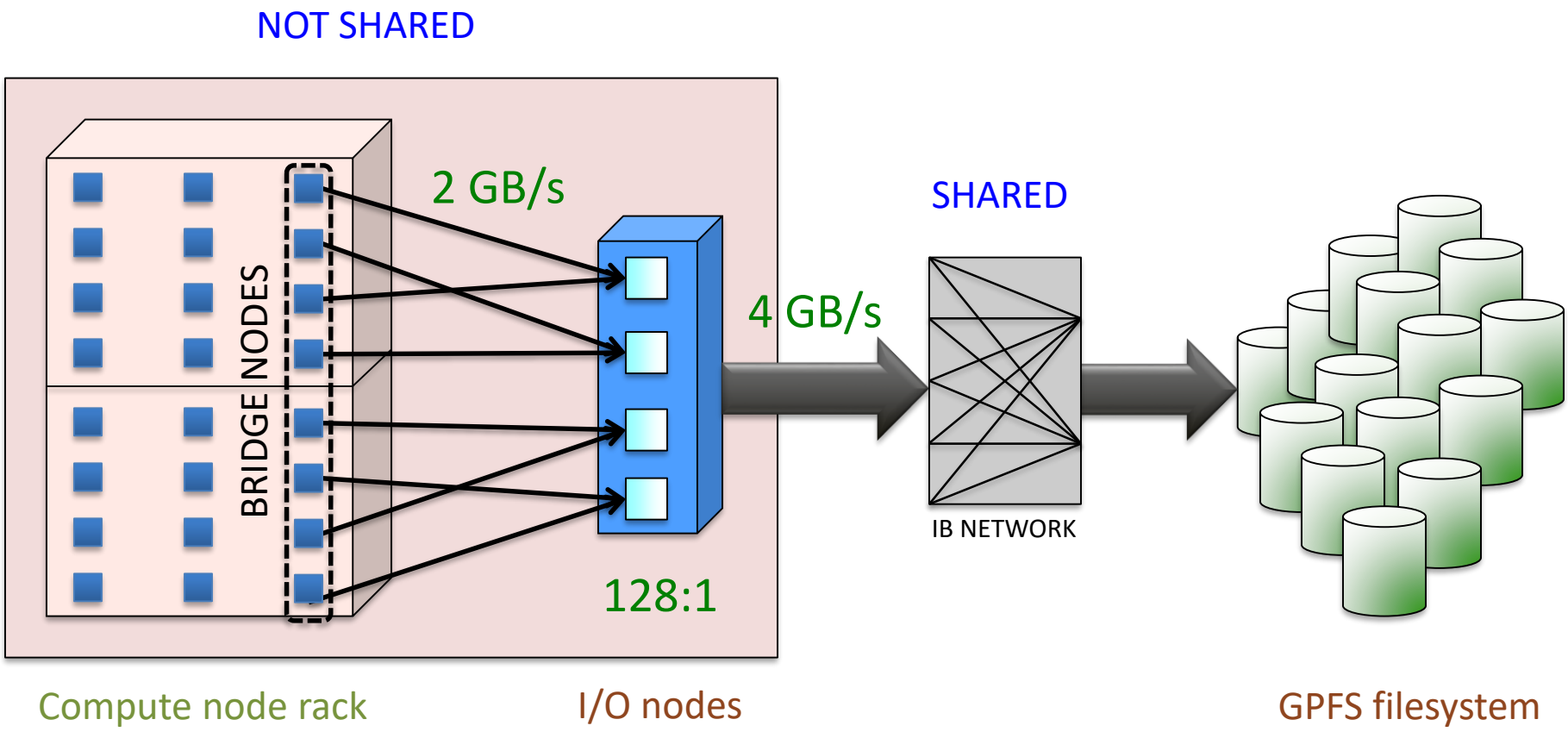
Message Passing Paradigm

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

Profiling


TAU: ParaProf: Function Data Window: /users/faculty/pmalakar/class/hands-on-2018-2019/Mar2...
File Options Windows Help
Name: MPI_Recv()
Metric Name: TIME
Value: Exclusive
Units: seconds
0.561 max
0.561 node 10
0.542 node 8
0.538 node 9
0.497 node 7
0.489 node 6
0.484 node 4
0.427 node 5
0.341 mean
0.201 std. dev.
0.165 node 3
0.149 node 0
0.128 node 2
0.108 min
0.108 node 1

Parallel I/O


NOT SHARED
BRIDGE NODES
2 GB/s
4 GB/s
SHARED
IB NETWORK
128:1
Compute node rack
I/O nodes
GPFS filesystem

Job Scheduling

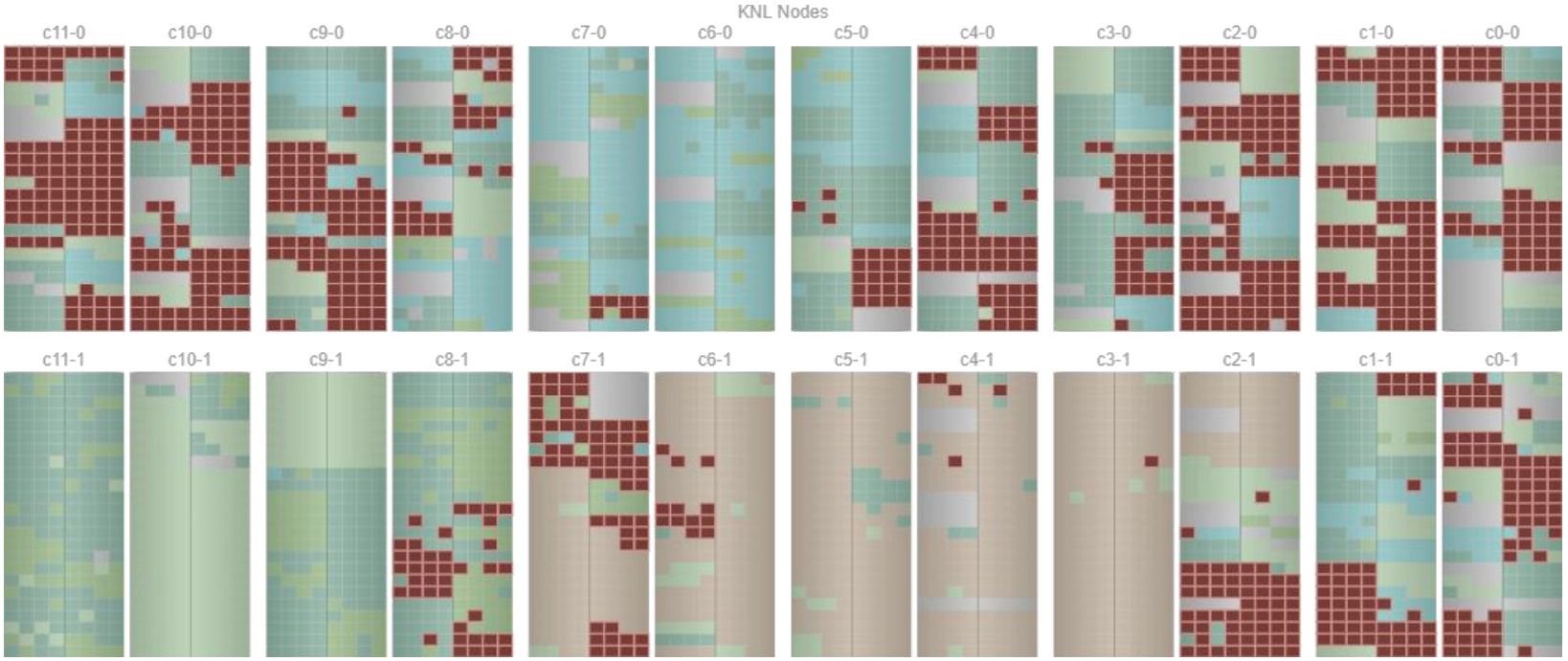
NODES

JOBS

USERS

Example of a real supercomputer activity
- Argonne National Laboratory Theta jobs

Supercomputer Activity



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