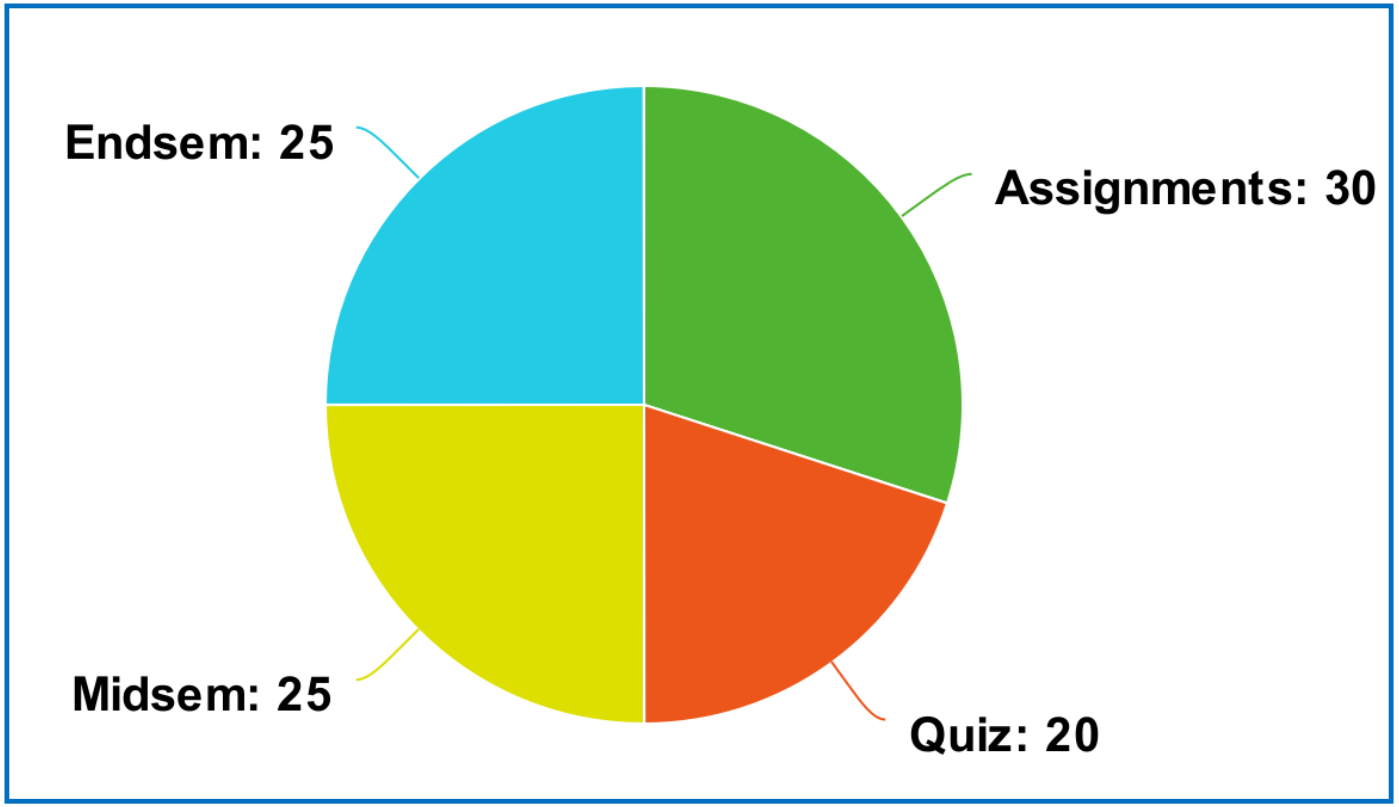
Parallel Computing (CS 633)

Preeti Malakar

Logistics

- https://web.cse.iitk.ac.in/users/pmalakar/cs633/2020-21
- Register on Piazza (https://piazza.com/iitk.ac.in/secondsemester2020/cs633)
  - Will be used for announcements, Q&A etc.
- Course-related email should always be prefixed with [CS633] in the subject
- Lectures will be uploaded every week
- Discussion hour: T at 3:30 – 4:30 PM
  - Discussion of previous week's lectures/quizzes
  - Additional meeting time (if required): F at 3:30 – 4:30 PM

Grading Policy



Assignments

- Individually or in a group (maximum group size = 2)
  - Send group member information by Jan 25
- Credit for early submissions (+5 / day)
  - Max credit: +15 / assignment
- A total of 2 extra days may be taken
- Score reduction for late submissions (-5 / day)
  - Max 3 late days / assignment
- None of the assignments can be completed in a day!

Assignments

- Programming assignments on CSE lab systems
  - C/C++
- Non-CS students – Send me an email to open a temporary CSE account
- Submission through Gitlab (git.cse.iitk.ac.in)
- Plagiarism will NOT be tolerated

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

This course …

Distributed Memory Parallelism

Parallel programming

Message passing

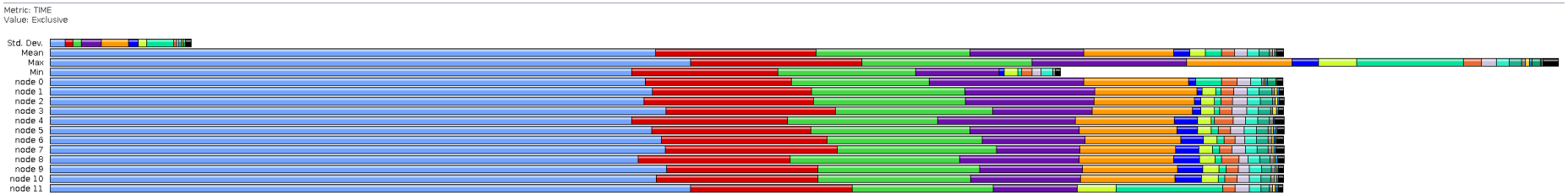
Parallel algorithms

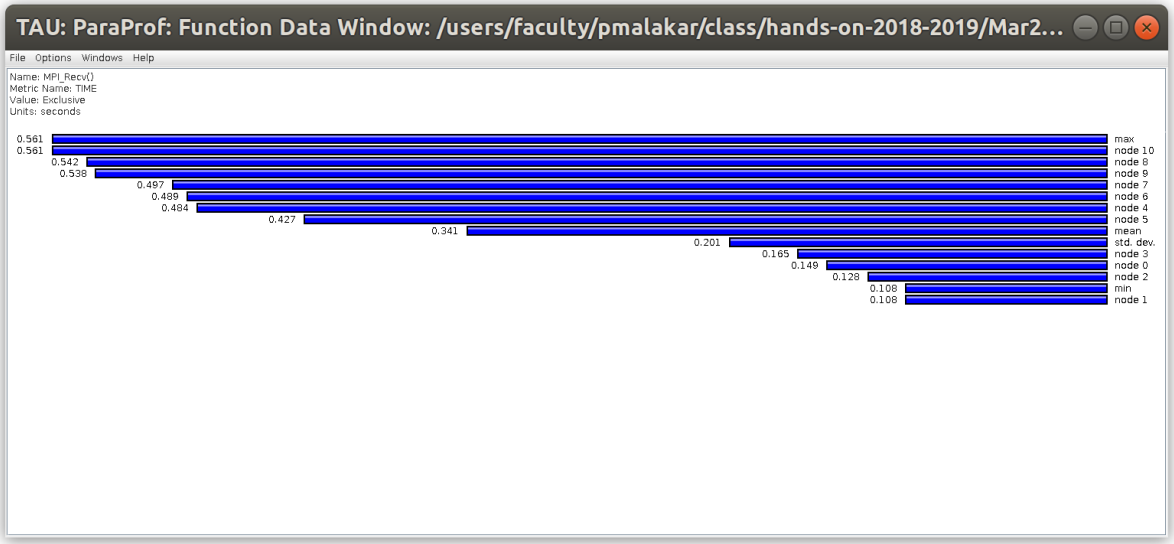
Designing parallel codes

MPI

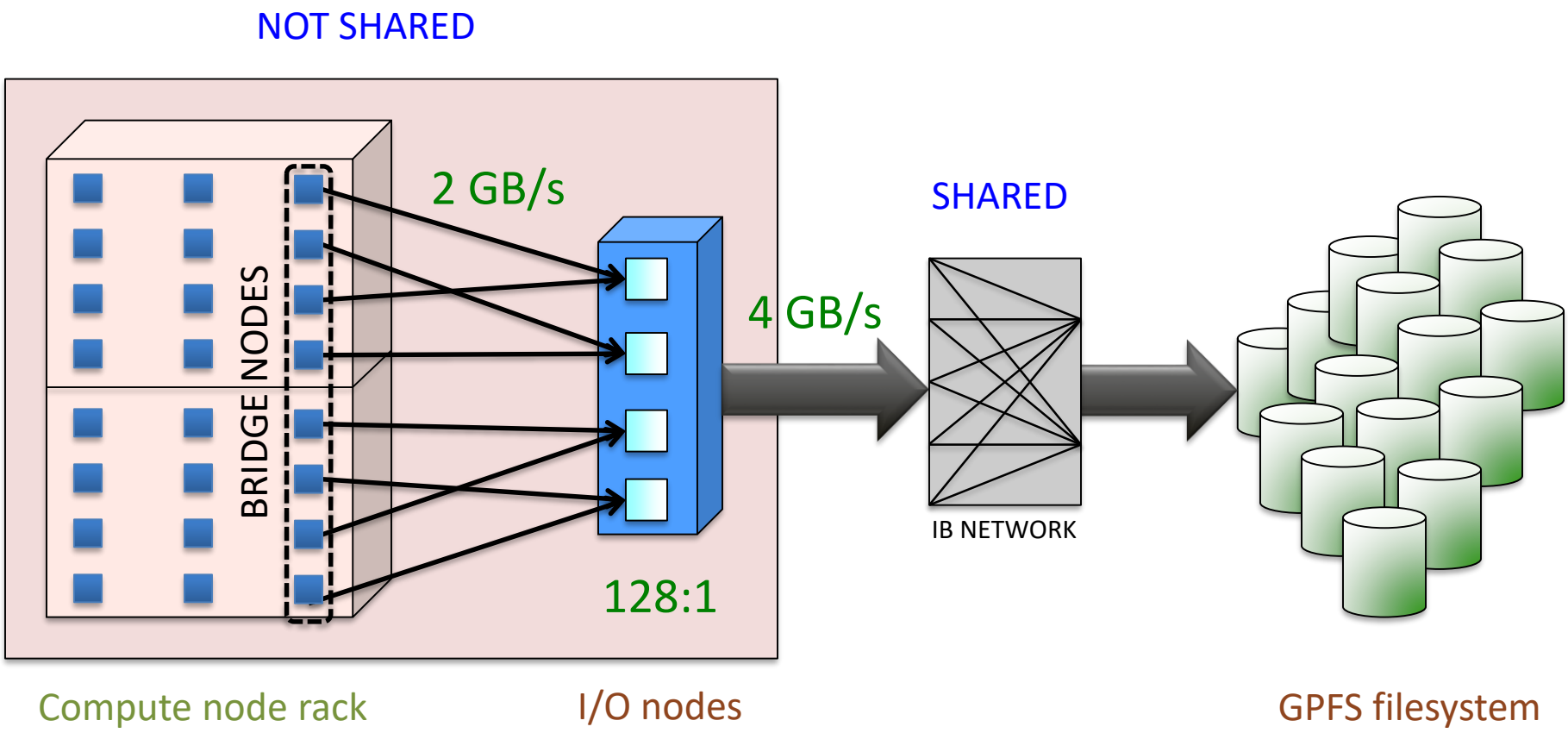
- P2P communications
- Collective communications
- Algorithms
- Performance

Profiling


Metric: TIME
Value: Exclusive
Std. Dev.
Mean
Max
Min
node 0
node 1
node 2
node 3
node 4
node 5
node 6
node 7
node 8
node 9
node 10
node 11


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

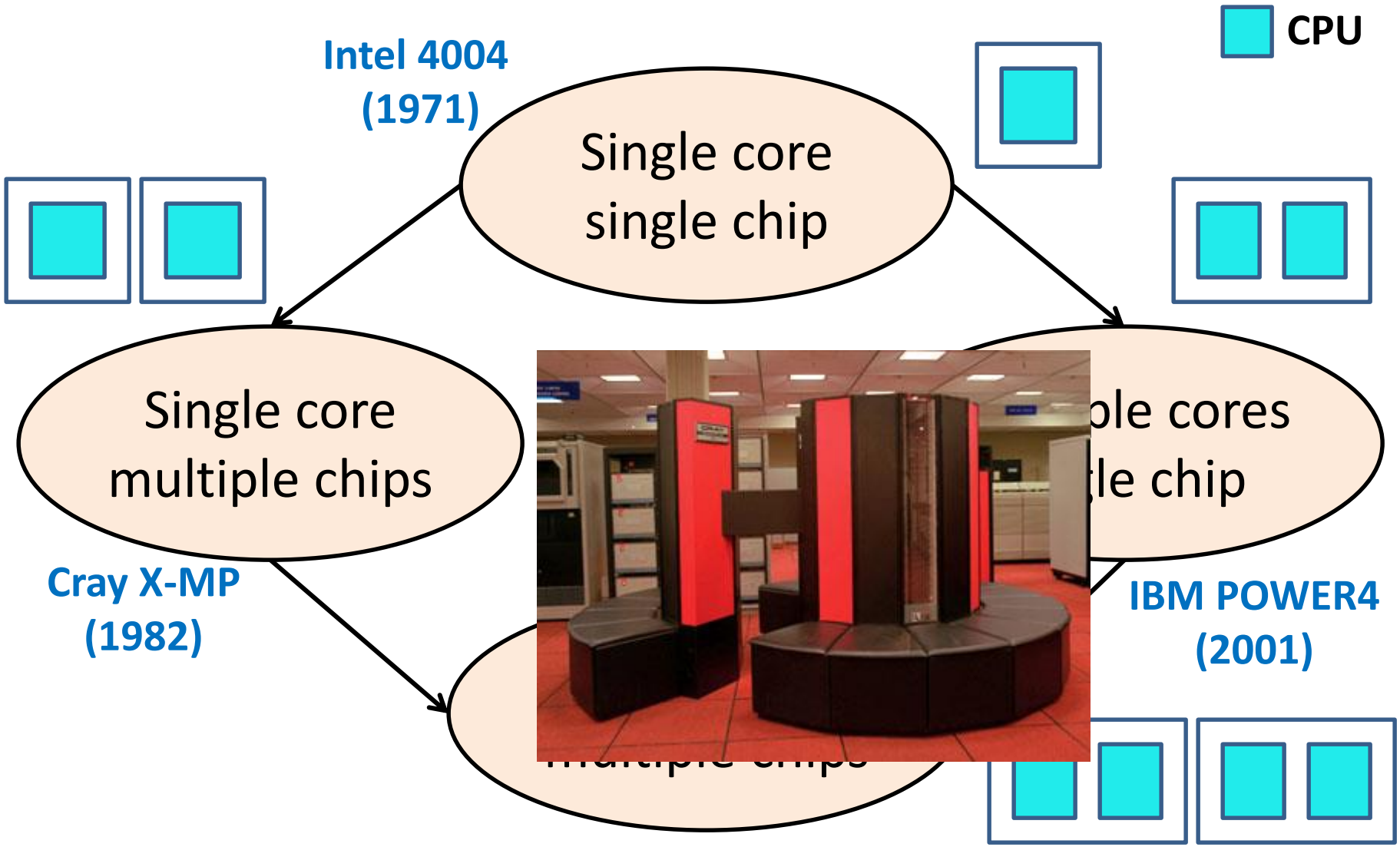
Example of a real supercomputer activity
- Argonne National Laboratory Theta jobs

Lecture 1

Introduction

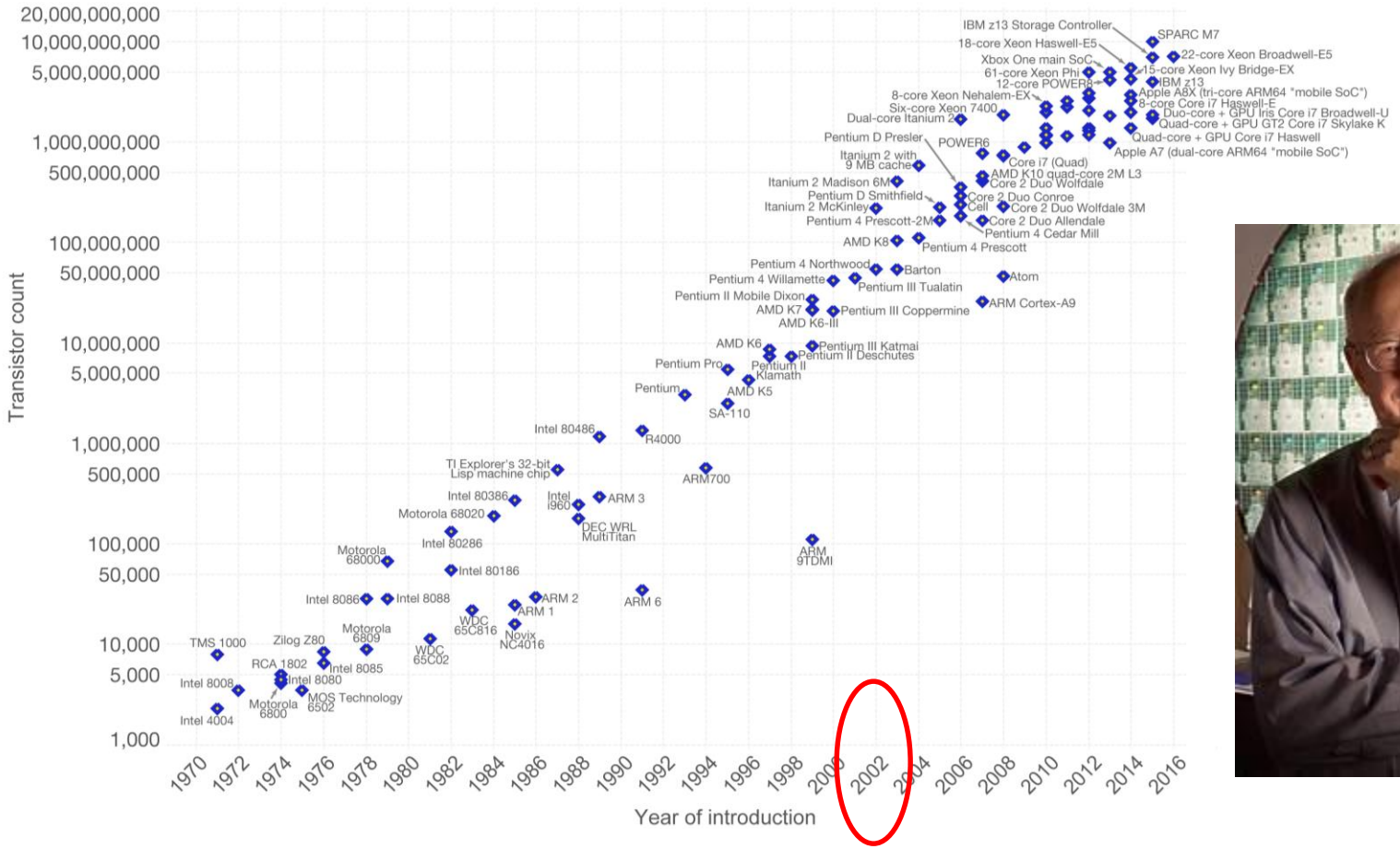
Jan 15, 2021

Multicore Era



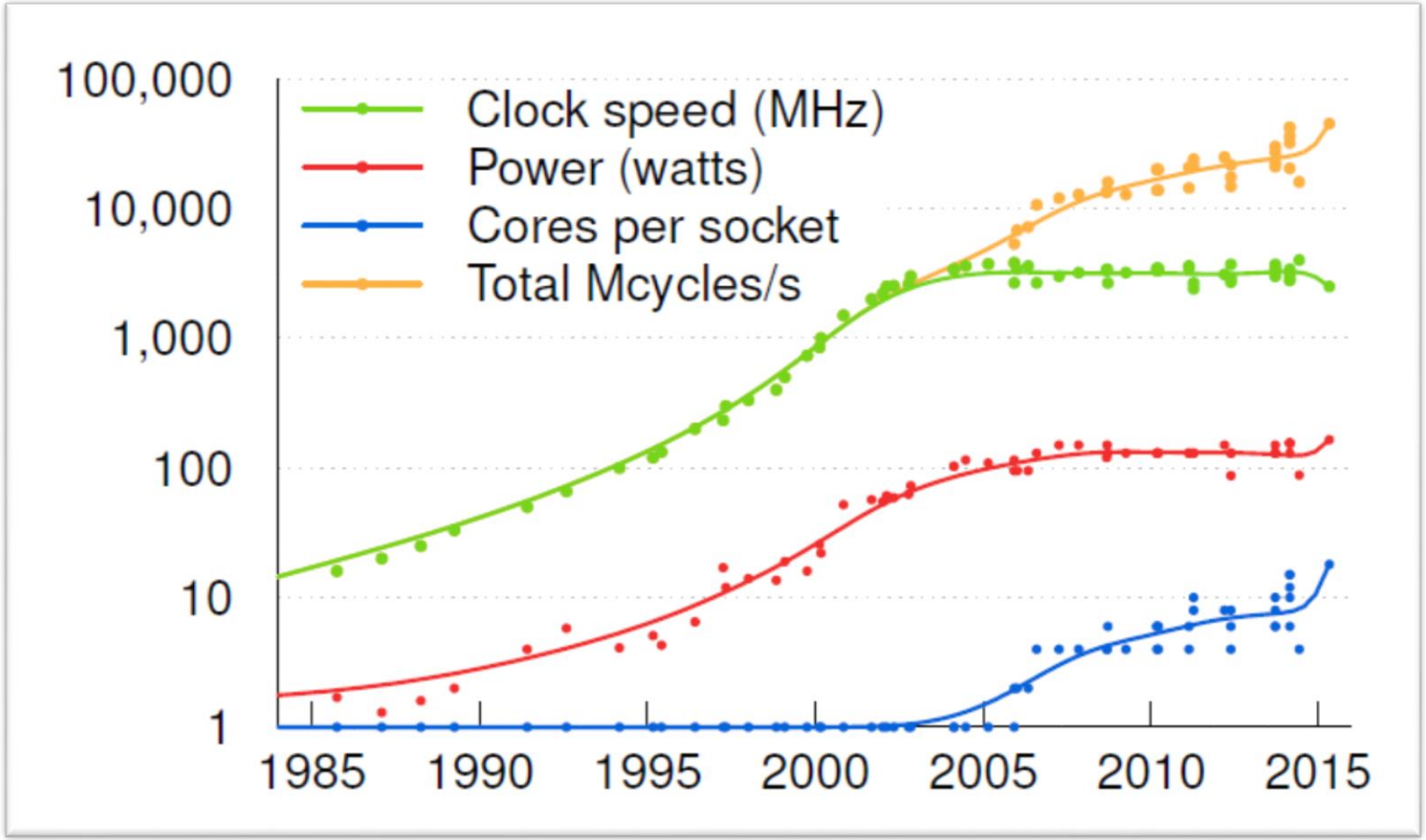
Moore's Law (1965)

Number of transistors in a chip doubles every 18 months



[Source: Wikipedia]

Trends



[Source: M. Frans Kaashoek, MIT]

top500.org (Nov'20)

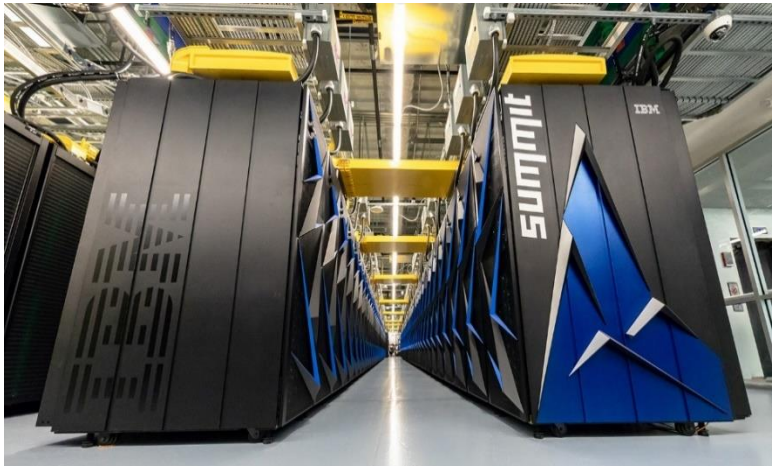
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,64		
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,4	
5	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646

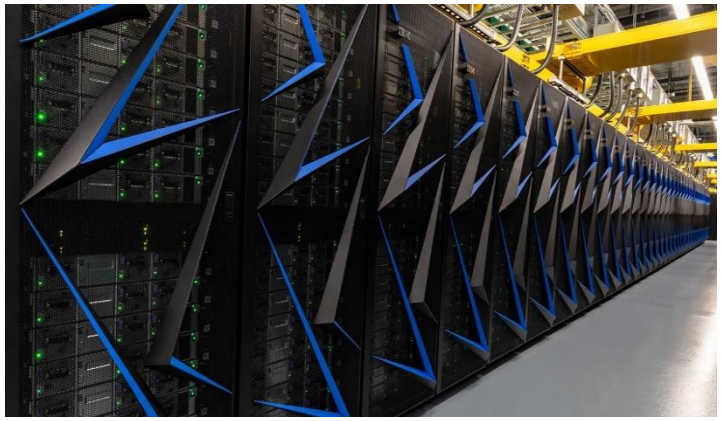
~ $200 million
~ 5600 sq. ft.
~ 13 MW power
~ 15000 l of water

green500.org (Nov'20)

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	170	**NVIDIA DGX SuperPOD** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	19,840	2,356.0	90	26.195
2	330	**MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect, Preferred Networks Preferred Networks Japan	1,664	1,652.9	65	26.039
3	7	**JUWELS Booster Module** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ) Germany	449,280	44,120.0	1,764	25.008
4	146	**Spartan2** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR Infiniband, Atos Atos France	23,040	2,566.0	106	24.262
5	5	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	2,646	23.983

Making of a Supercomputer





Source: energy.gov

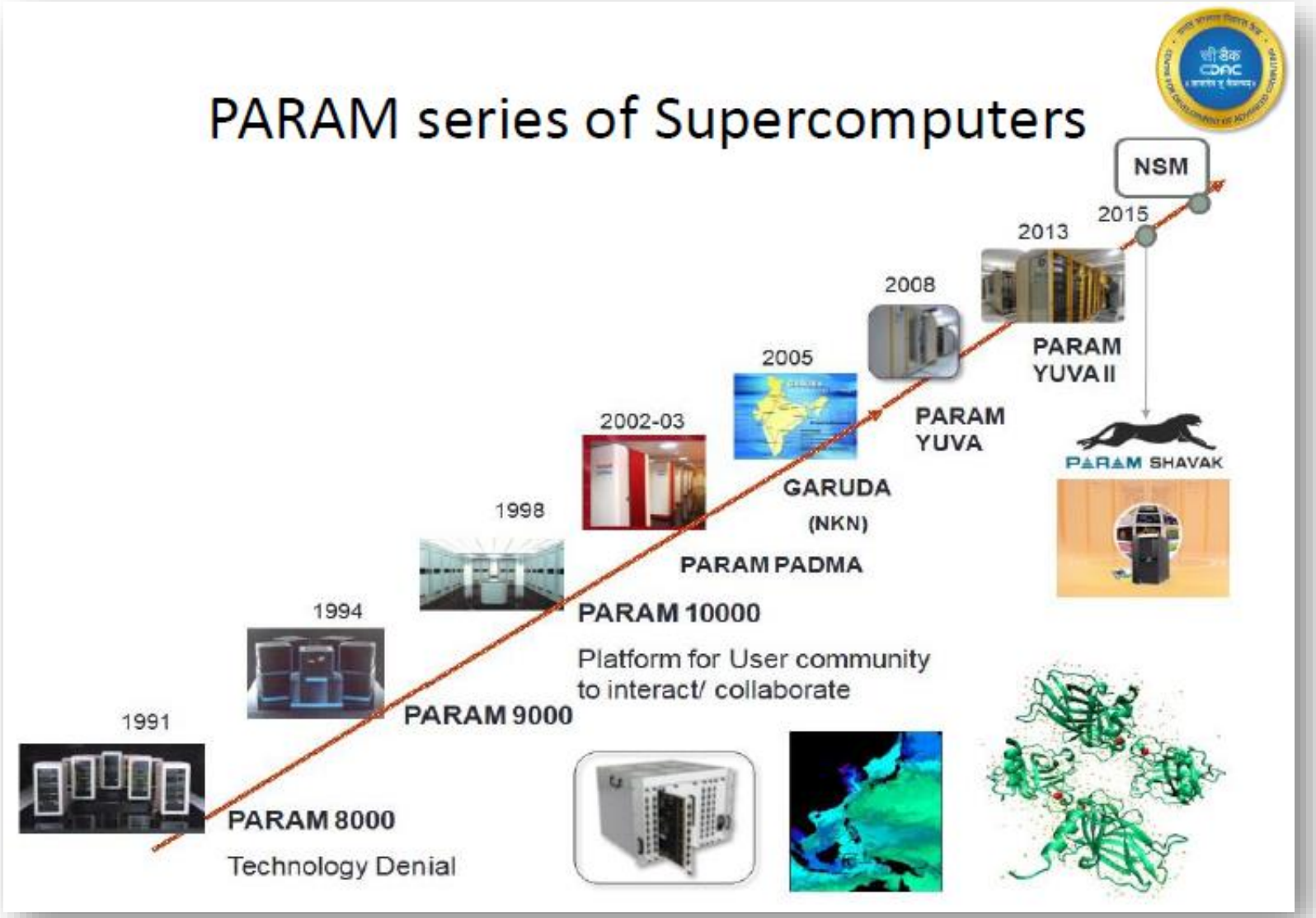
Greenest Data Centre?

Source: MIT TR 06/19

Supercomputing in India [topsc.cdacb.in, Jul'20]

Rank	Site	System	Cores/Processor Sockets/Nodes	Rmax (TFlops)	Rpeak (TFlops)
1	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. OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	119232/ /3312	3763.9	4006.19
2	National Centre for Medium Range Weather Forecasting (NCMRWF), Noida	Cray XC-40 class system with 2322 CPU-only (Intel Xeon Broadwell E5-2695 v4 CPU) nodes with Cray Linux environment as OS, and connected by Cray Aries interconnect OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	83592//2322	2570.4	2808.7
3	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 was designed and implemented by HPC Technologies team, Centre for Development of Advanced Computing (C-DAC) with total peak computing capacity of 1.66 (CPU+GPU) PFLOPS performance. The system uses the Lustre parallel file system (primary storage 1.5 PiB usable with 50 GB/Sec write throughput. Archival Storage 500TiB based on GPFS) OEM: Atos India Pvt Ltd., Bidder: Atos India Pvt Ltd.	17280/2/432	935	1290.2
4	National Atmospheric Research Laboratory(NARL),Tirupati	The Supercomputer PARAM AMBAR is based on a heterogeneous and hybrid configuration of Intel Xeon Cascade Lake processors,NVIDIA Tesla V100 with nvlink, Intel OPA 100Gbps 100% non-blocking.The system was designed and implemented by HPC Technologies team, Centre for Development of Advanced Computing (C-DAC) with C-DAC software stack OEM: Tyrone, Bidder: M/s.Netweb Technologies.	18816/2/196	919.61	1384.85
5	Supercomputer Education and Research Centre (SERC), Indian Institute of Science (IISc), Bangalore	Cray XC-40 Cluster (1468 Intel Xeon E5-2680 v3 @ 2.5 GHz dual twelve-core processor CPU-only nodes, 48 [Intel Xeon E5-2695v2 @ 2.4 Ghz single twelve-core processor+Intel Xeon Phi 5120D] Xeon-phi nodes, 44 [Intel Xeon E5-2695v2 @ 2.4 Ghz single twelve-core processor+NVIDIA K40 GPUs] GPU nodes) w/ Cray Aries Interconnect. HPL run on only 1296 CPU-only nodes. OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	36336C + 2880ICO + 126720G/ 3028C + 48ICO + 44G/ 1560C + 48ICO + 44G	901.51 (CPU-only)	1244.00 (CPU-only)
6	Indian Institute of Tropical Meteorology, Pune	IBM/Lenovo System X iDataPle DX360M4, Xeon E5-2670 8C 2.6 GHz, Infiniband FDR OEM: IBM/Lenovo, Bidder: IBM India Pvt. Ltd.	38016/ /	719.2	790.7
7	Indian Lattice Gauge Theory Initiative, Tata Institute of Fundamental Research (TIFR), Hyderabad	Cray XC-30 cluster (Intel Xeon E5-2680 v2 @ 2.8 GHz ten-core CPU and 2688-core NVIDIA Kepler K20x GPU nodes) w/Aries Interconnect OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	4760C + 1279488G/ 476C + 476G/ 476C + 476G	558.7	730.00
8	Indian Institute of Technology, Delhi	HP Proliant XL230a Gen9 and XL250a Gen9 based cluster (Intel Xeon E5-2680v3 @ 2.5 GHz dual twelve-core CPU and dual 2880-core NVIDIA Kepler K40 GPU nodes) w/Infiniband OEM: HP, Bidder: HP	10032C + 927360G/ 836C + 322G/ 418C + 161G	524.40	861.74
9	Indian Institute of Science Education And Research,Pune	The PARAM Brahma DCLC based HPC cluster has 179 Intel Xeon Platinum 8268 nodes (Bull Sequana CPU only compute blades are manufatured in India) constituting a total of 8592 CPU cores with C-DAC software stack ,total storage of 1PiB storage system and with HDR100 Infiniband Interconnect.The system was designed and implemented by HPC Technologies team, Centre for Development of Advanced Computing (C-DAC). OEM: Atos India Pvt Ltd., Bidder: Atos India Pvt Ltd.	8592/2/162	472.8	721.16
10	Indian Institute of Technology (IITBHU), Varanasi	The PARAM Shivay, first Super Computer installed by C-DAC under National Supercomputing Mission(NSM) project. It is based on a heterogeneous and hybrid configuration of BullSequana X400 series system with 192 CPU nodes, 20 High Memory nodes and 11 GPU nodes with 2xNvidia V100 accelerators. Each node having 2xIntel Xeon Skylake 6148, 20 cores processors with C-DAC Software stack. Primary Network Mellanox 100Gbps EDR Infiniband 100% fully non-blocking FatTree architecture. Lustre based primary storage 750TiB usable with 25GB/Sec write throughput. Archival Storage 250TiB based on GPFS. OEM: Atos India Pvt Ltd., Bidder: Atos India Pvt Ltd.	210/2/8400	456.9	645.12

Source: www.iitk.ac.in



Credit: Ashish Kuvelkar, CDAC

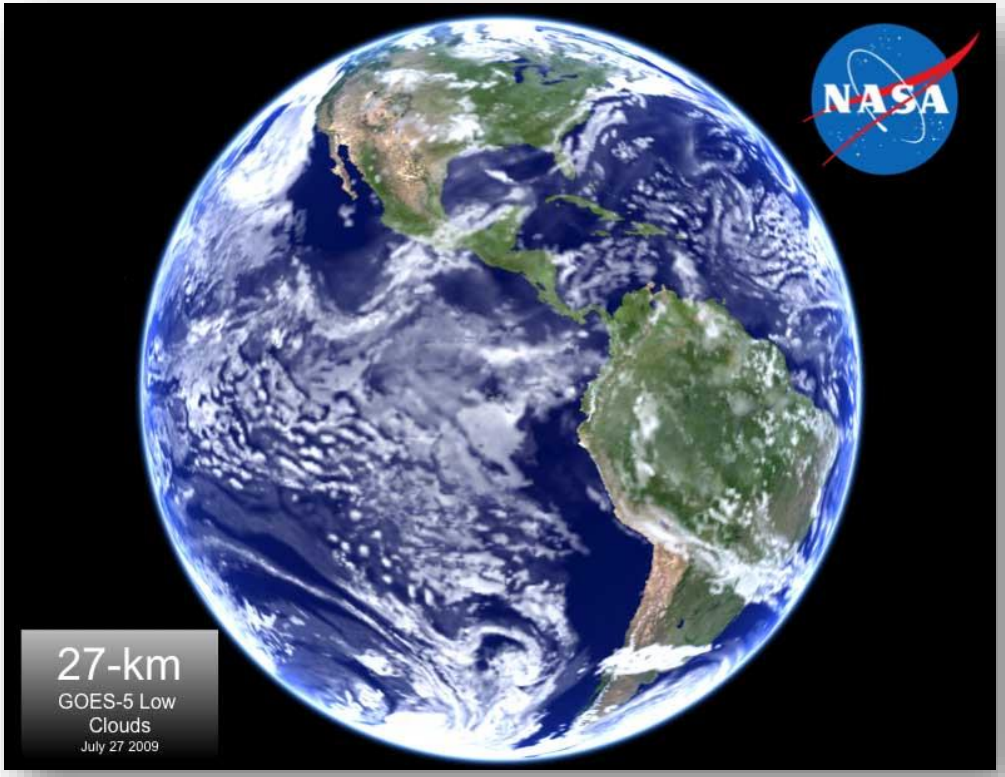
System Architecture Trends

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]

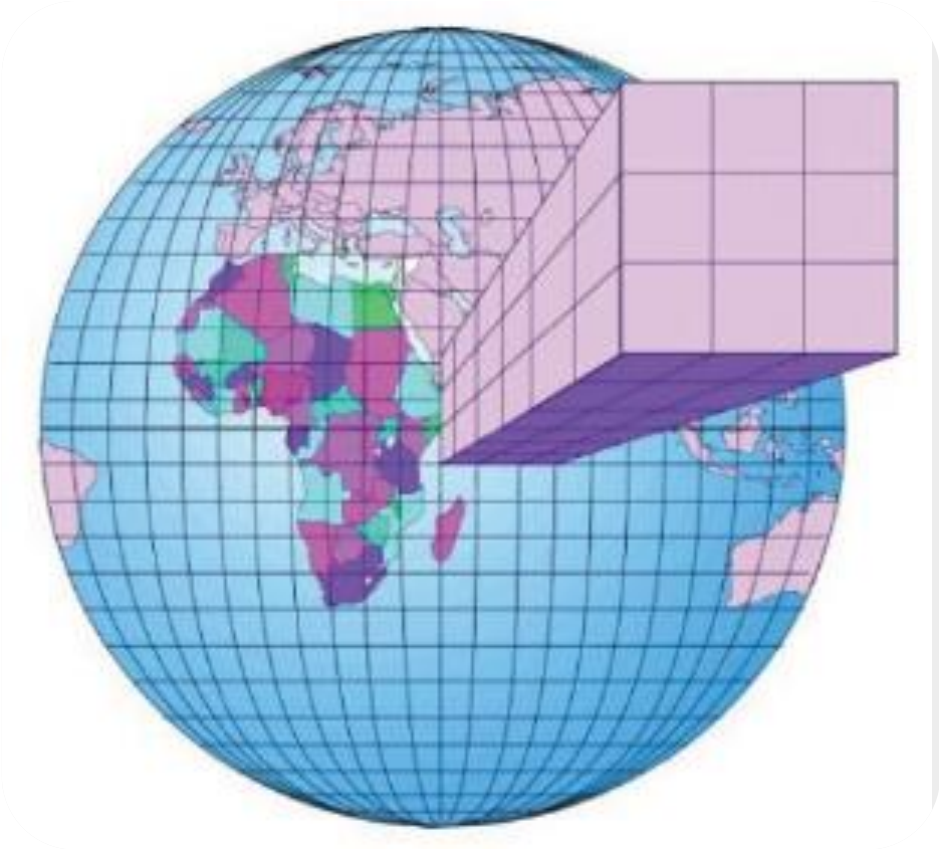
Big Compute

Massively Parallel Codes



Climate simulation of Earth [Credit: NASA]

Discretization

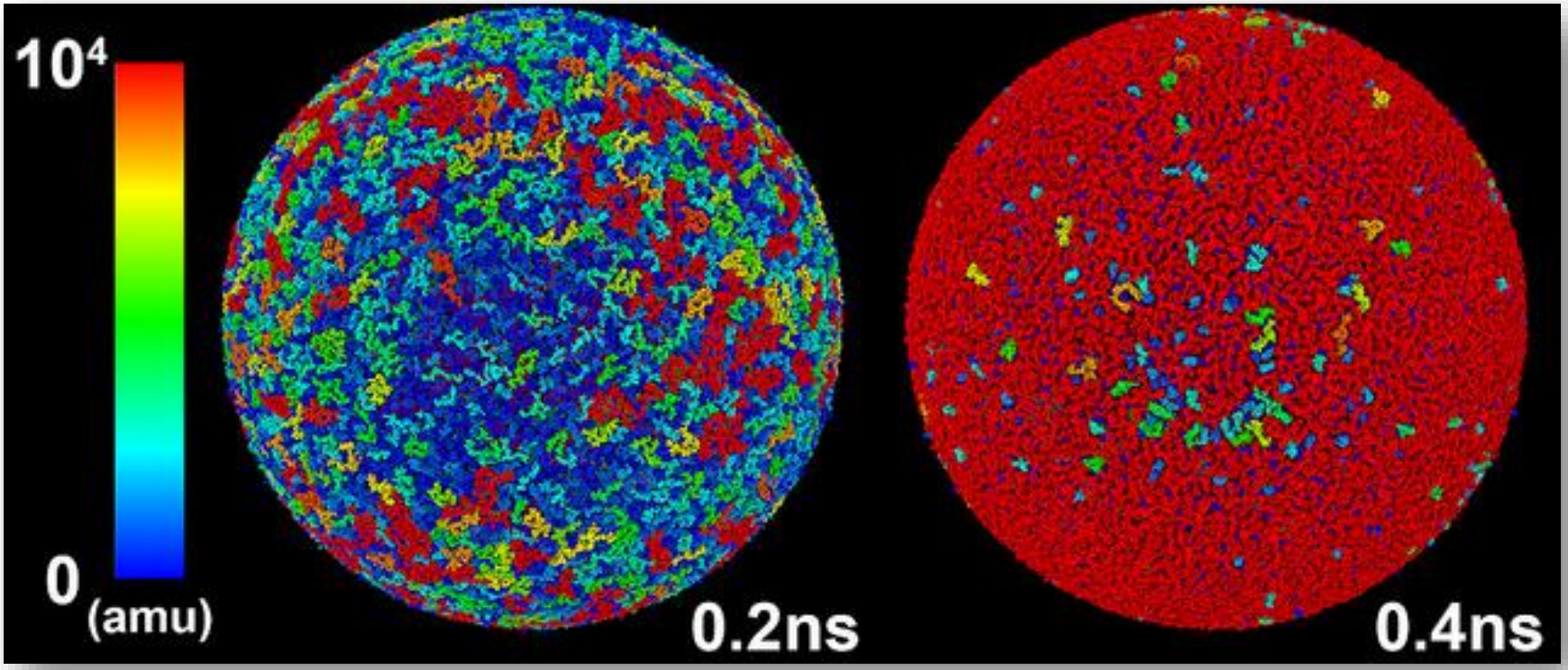


Gridded mesh for a global model [Credit: Tompkins, ICTP]

Numerical Weather Models

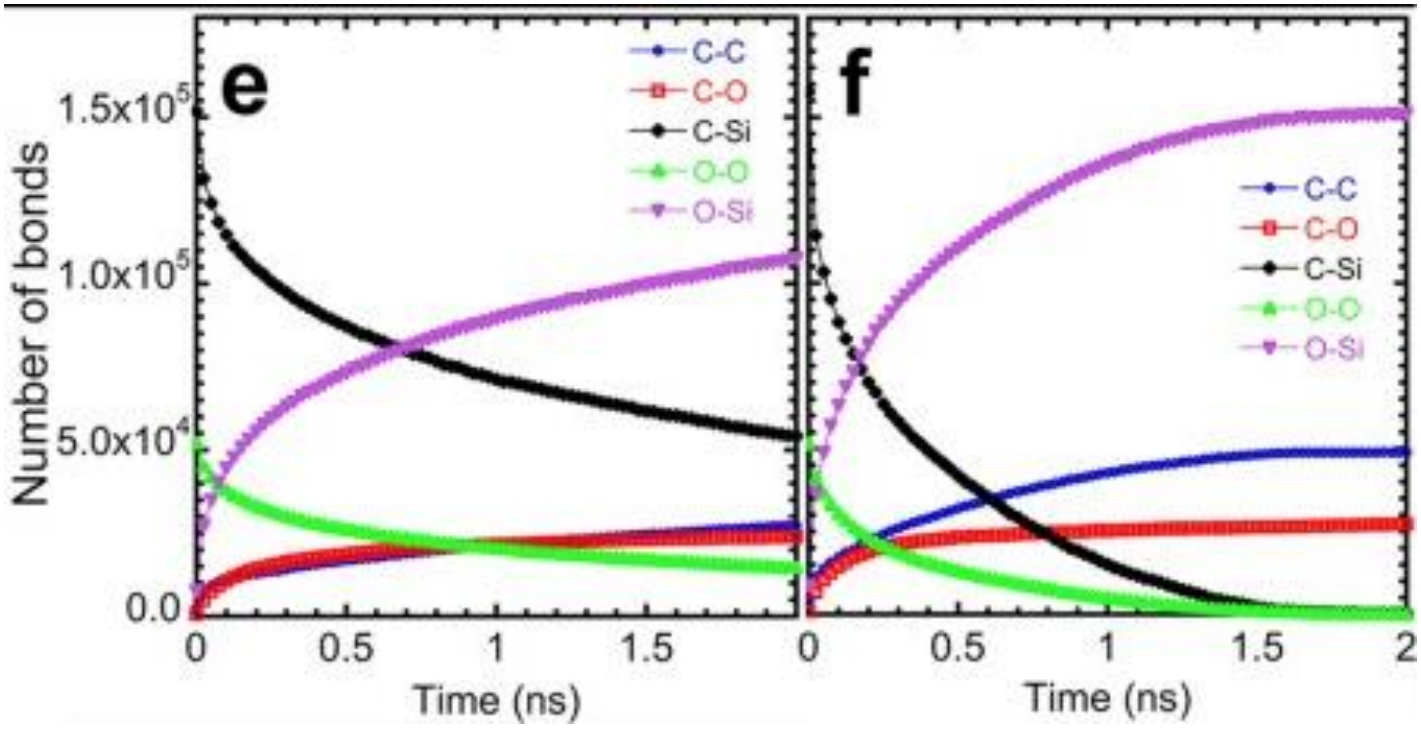
- Use numerical methods to solve equations that govern atmospheric processes
- Are based on theories and depend on observations of meteorological variables
- Are used to obtain nowcast/forecast

Massively Parallel Simulations



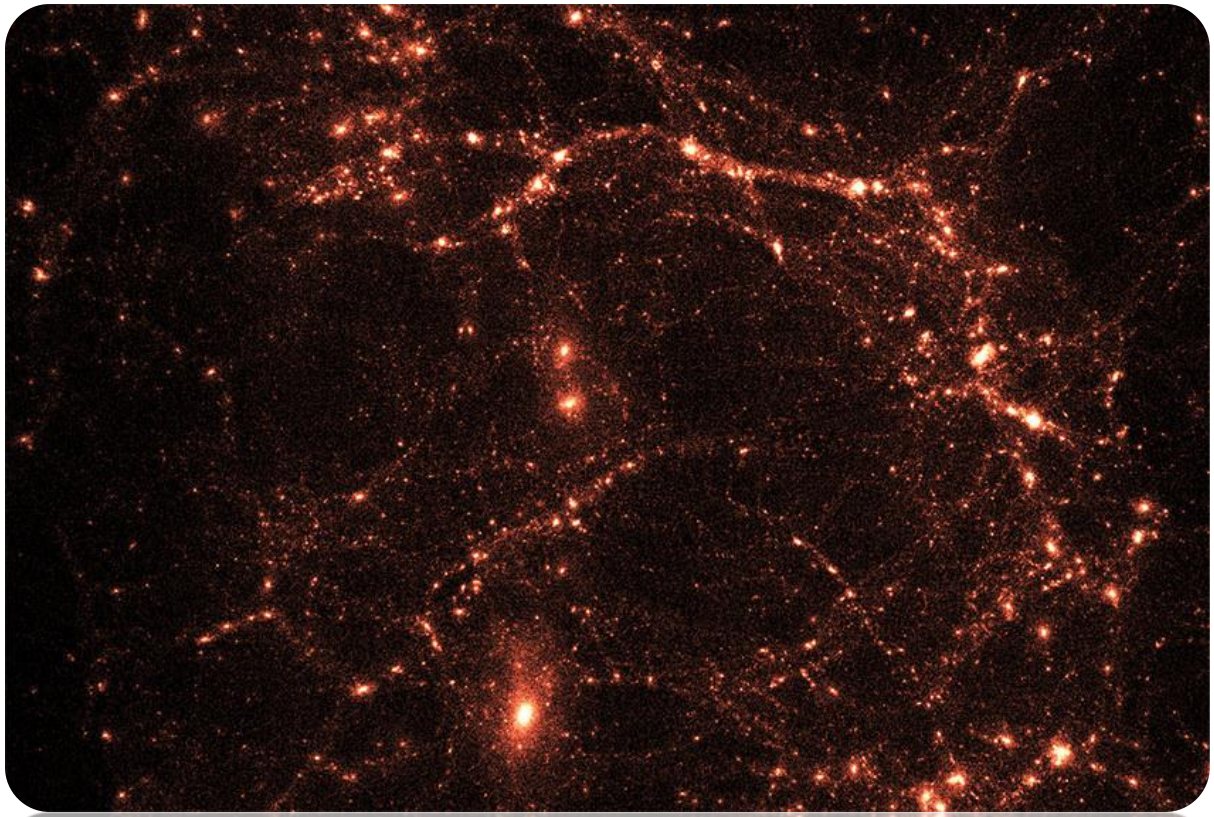
Self-healing material simulation
[Nomura et al., "Nanocarbon synthesis by high-temperature oxidation of nanoparticles", Scientific Reports, 2016]

Massively Parallel Analysis



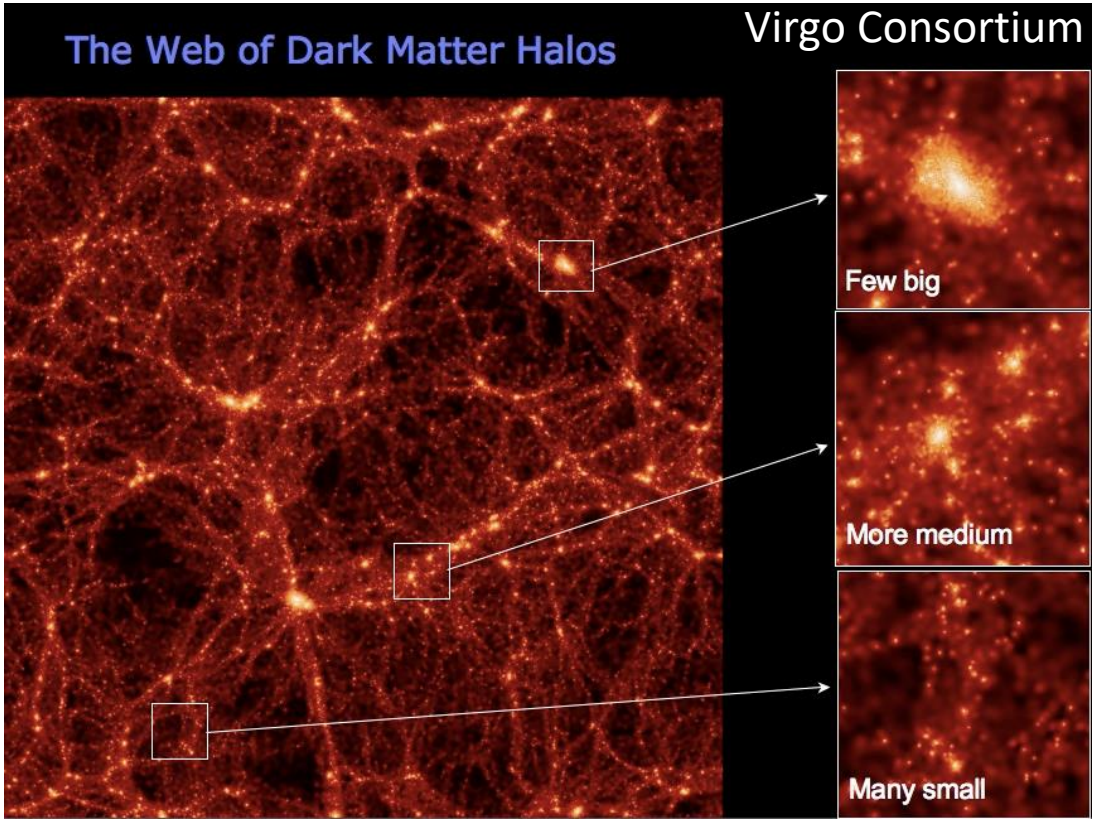
[Nomura et al., "Nanocarbon synthesis by high-temperature oxidation of nanoparticles", Scientific Reports, 2016]

Massively Parallel Codes

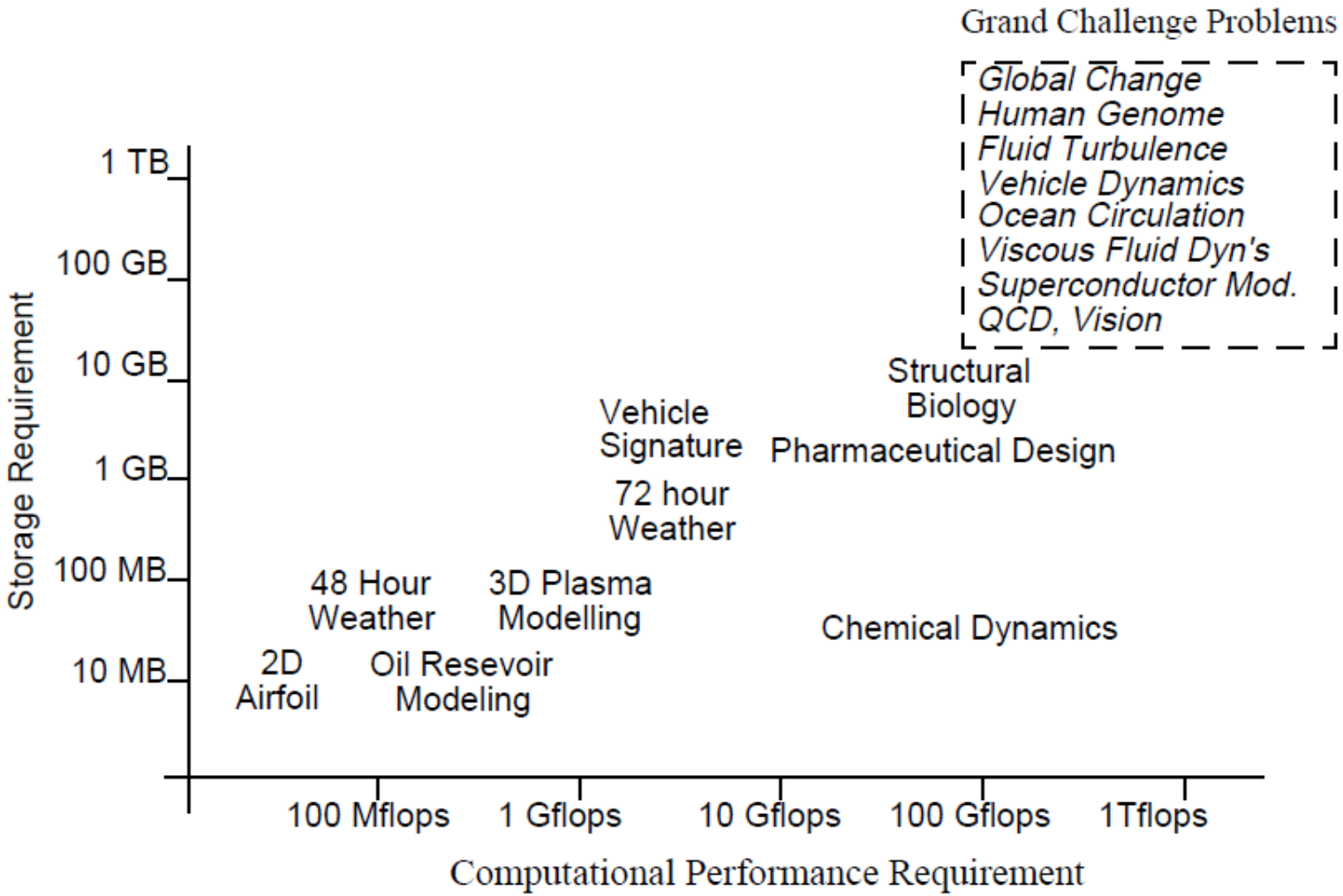


Cosmological simulation [Credit: ANL]

Massively Parallel Analysis



Computational Science

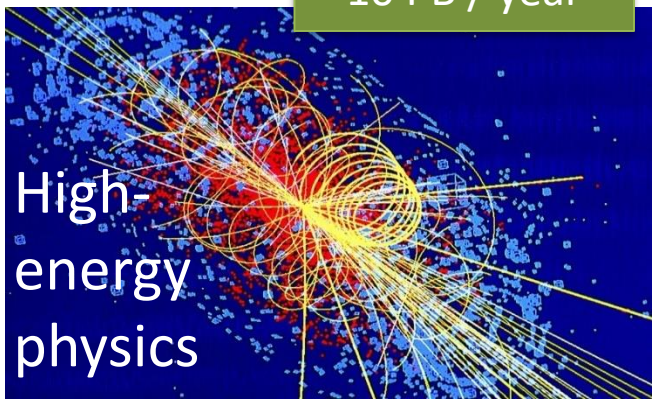


[Source: Culler, Singh and Gupta]

Big Data

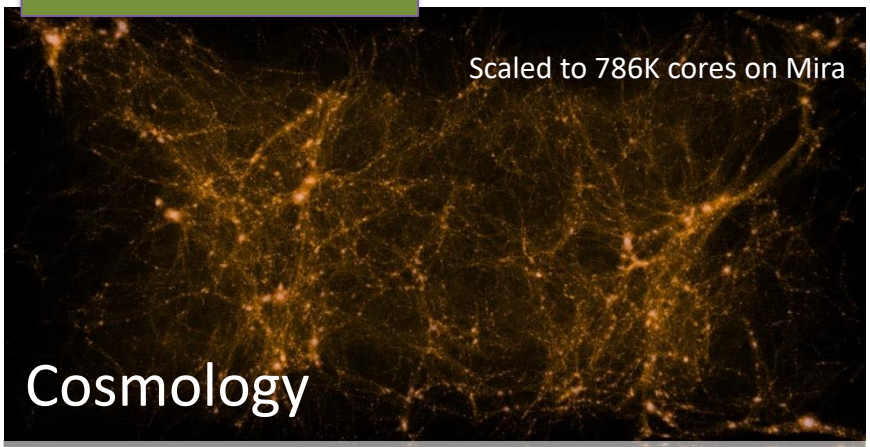
Output Data

10 PB / year



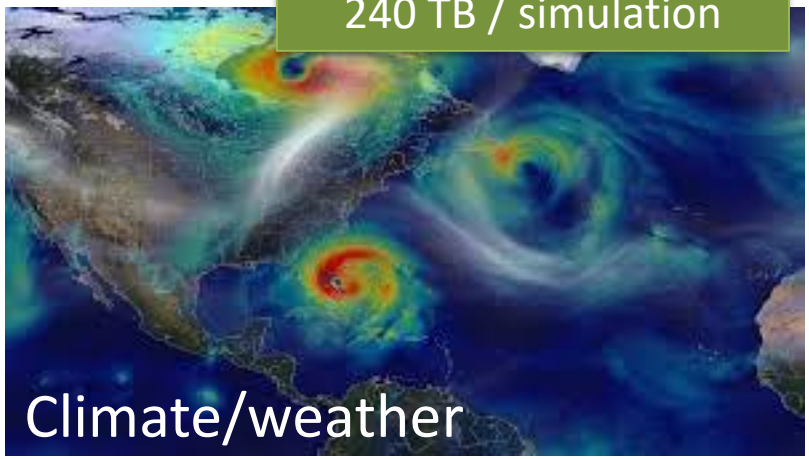
Higgs boson simulation
Source: CERN

2 PB / simulation



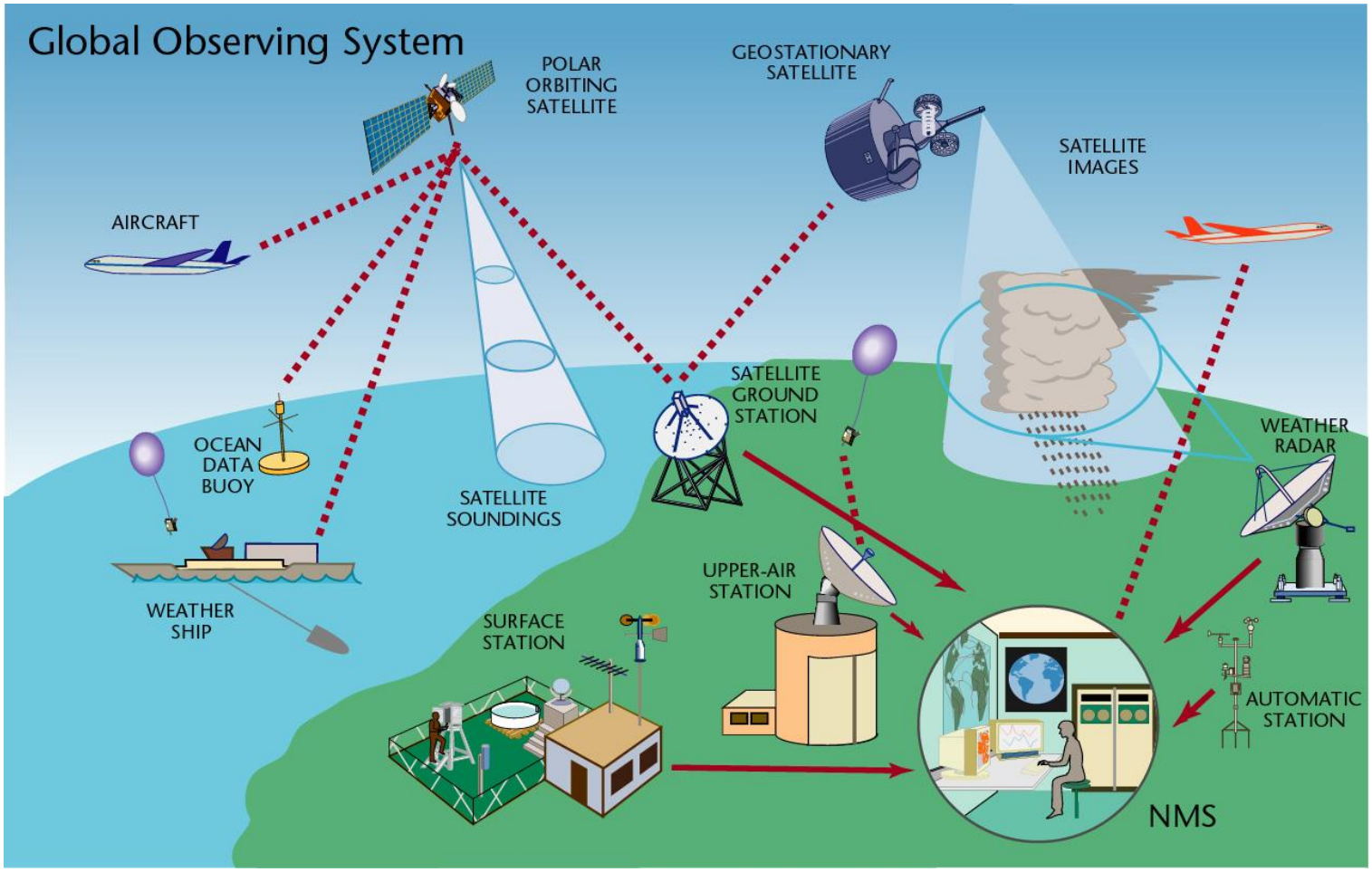
Q Continuum simulation
Source: Salman Habib et al.

240 TB / simulation



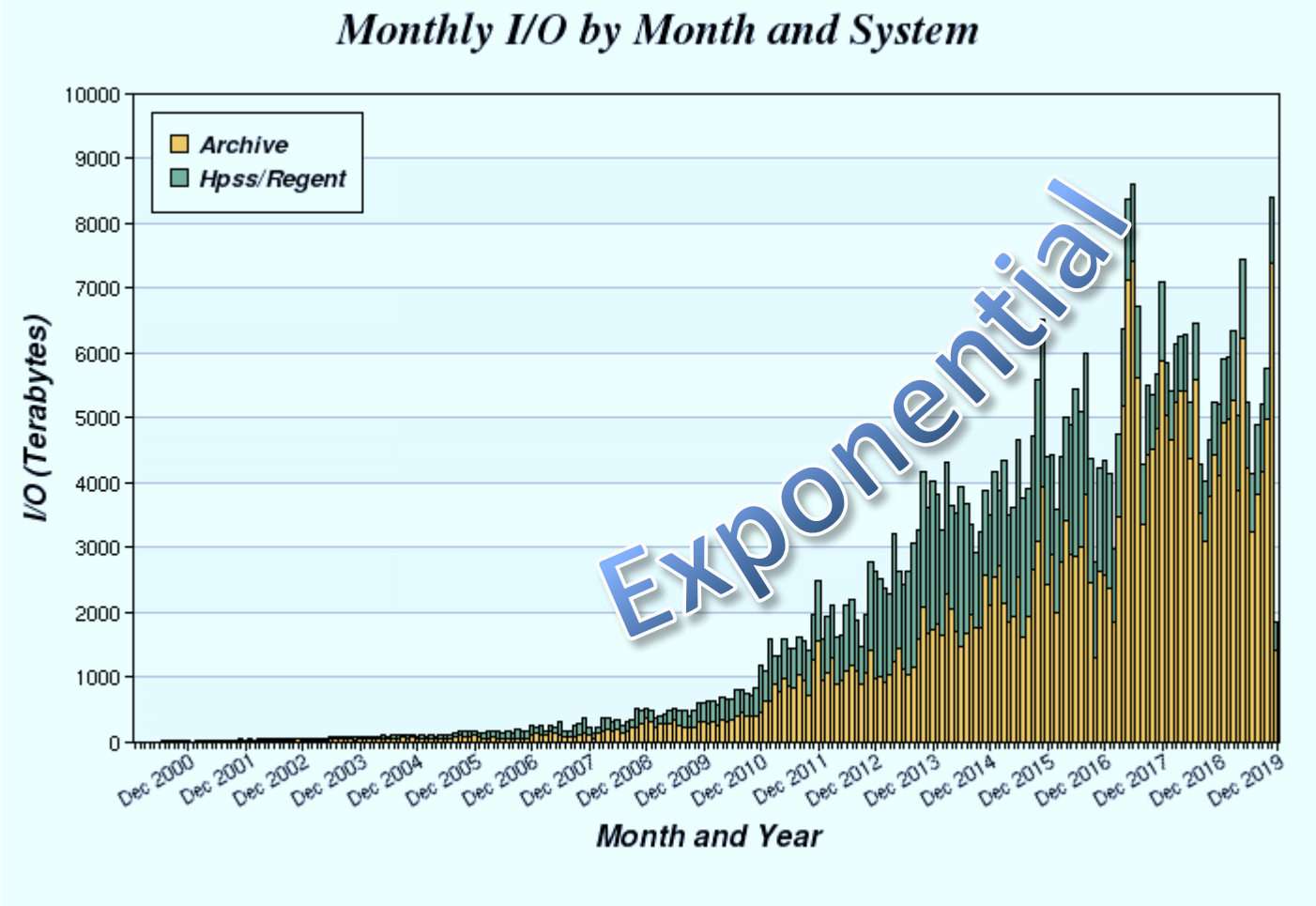
Hurricane simulation
Source: NASA

Input Data



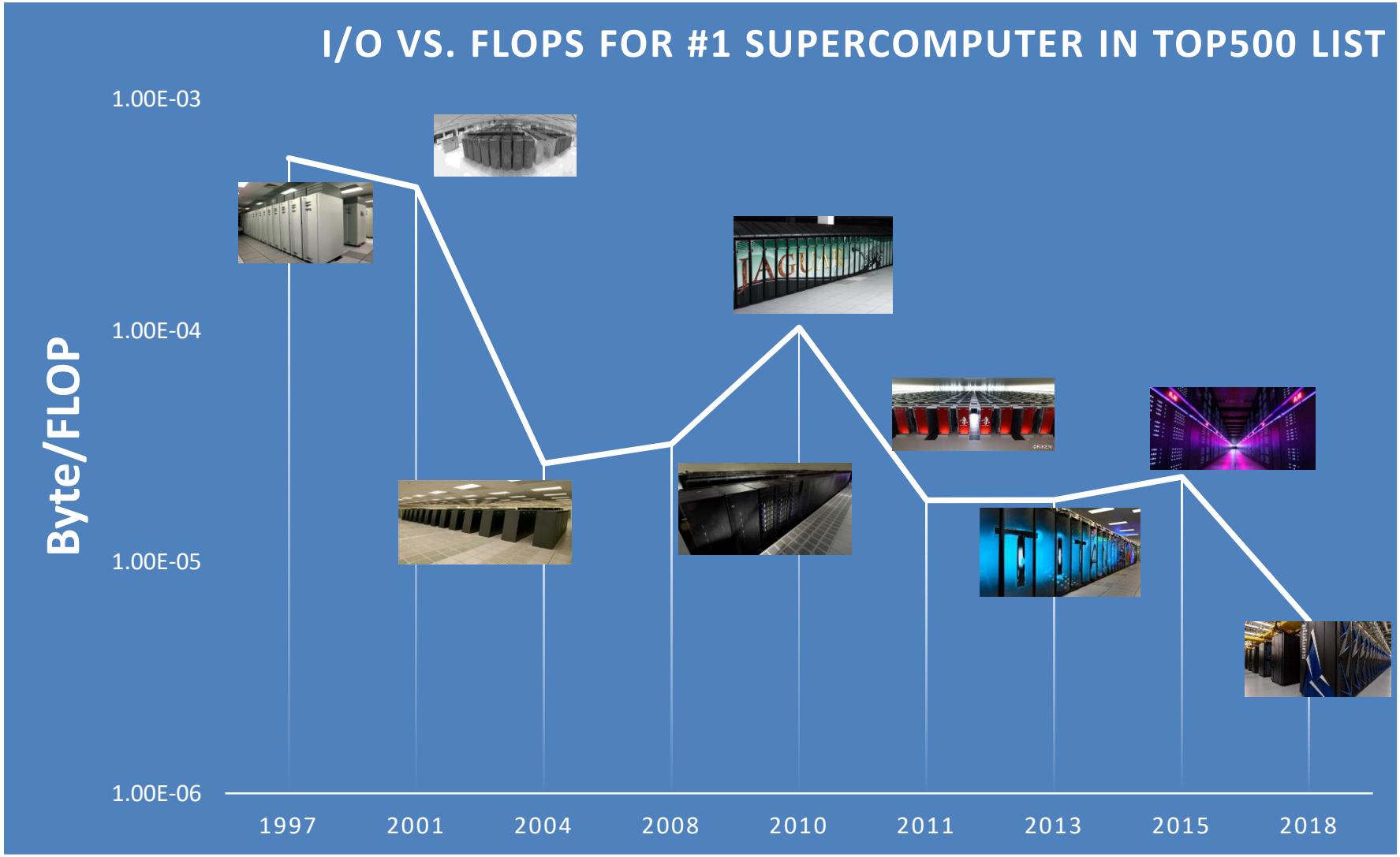
[Credit: World Meteorological Organization]

I/O trends

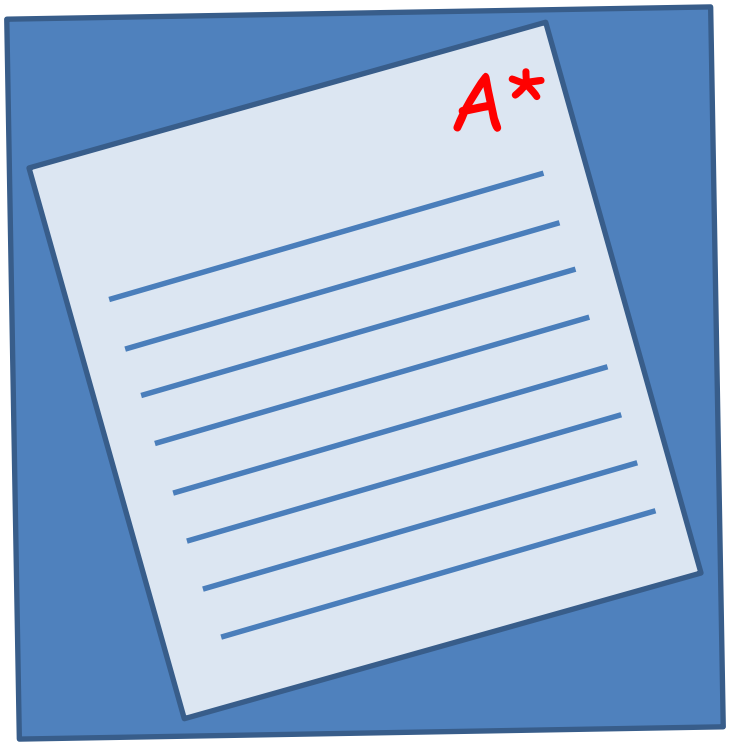


NERSC I/O trends [Credit: www.nersc.gov]

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

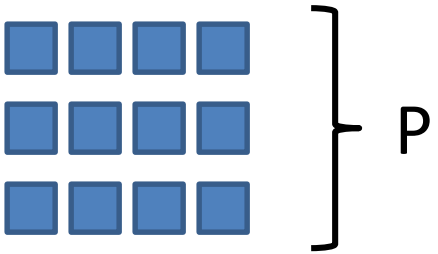
Serial

```
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

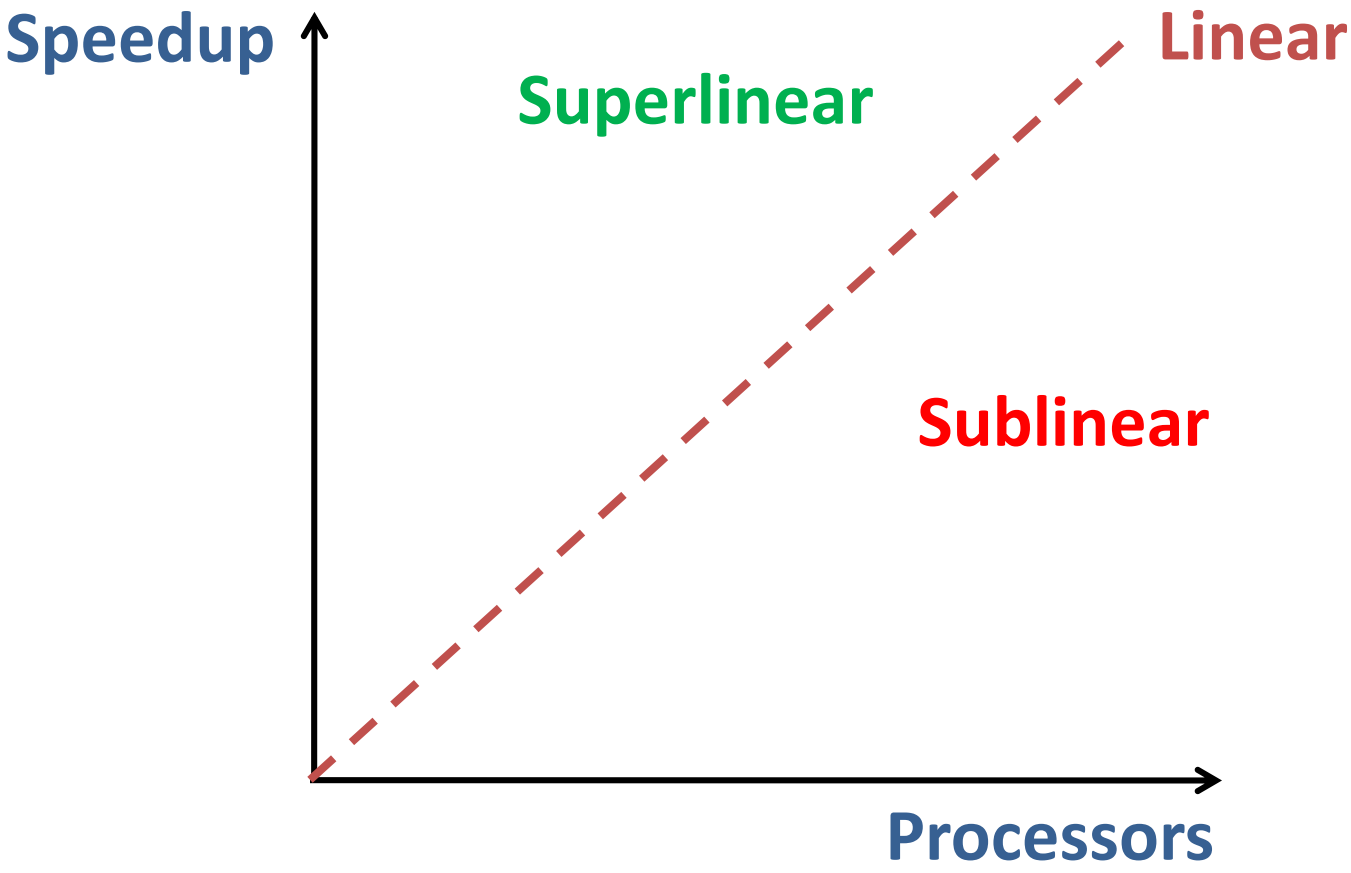
$$E_P = \frac{S_P}{P}$$

Parallel Performance (Parallel Sum)

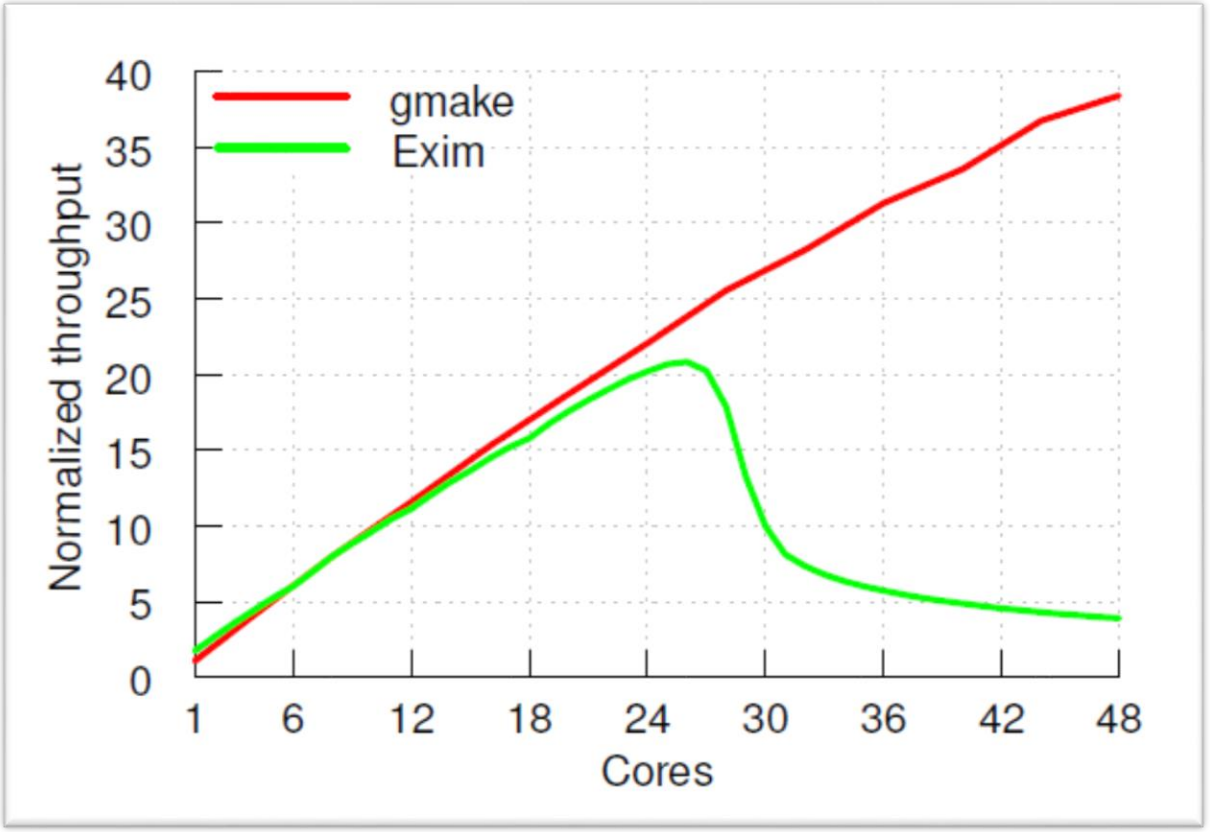
Parallel efficiency of summing 10^7 doubles

#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

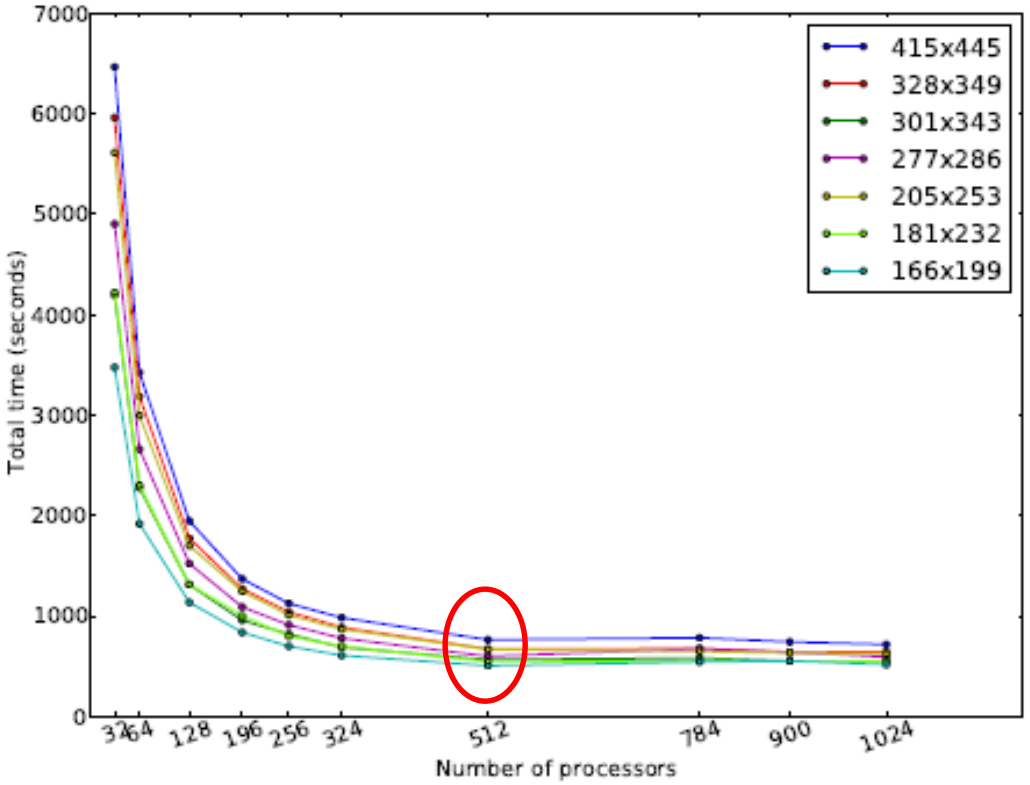


Issue – Scalability



[Source: M. Frans Kaashoek, MIT]

Scalability Bottleneck



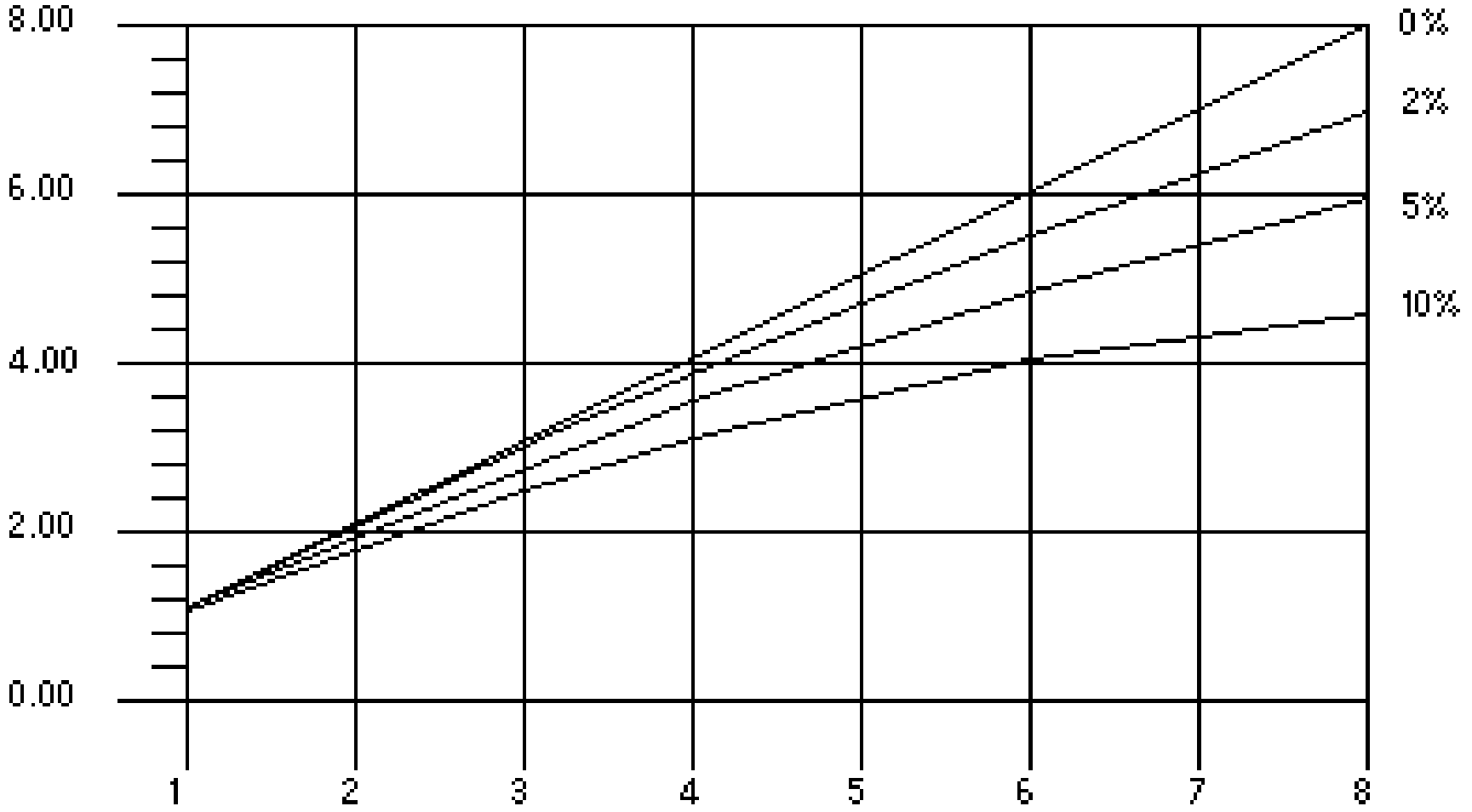
Performance of weather simulation application

A Limitation of Parallel Computing

Amdahl's Law:

$$\text{Speedup } S = \frac{1}{(1 - f) + f/P}$$

Speedup (Amdahl's Law)



Scaled Speedup

Gustafson's Law:

Speedup S' = $\dfrac{S' + N.P'}{S' + P'}$

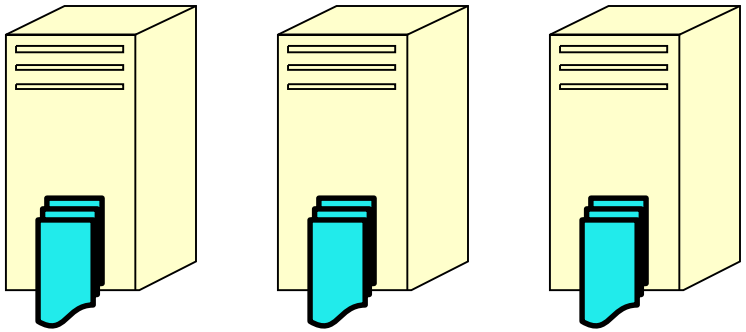
= $\dfrac{(1 - f') + N.f'}{1}$

Parallelism

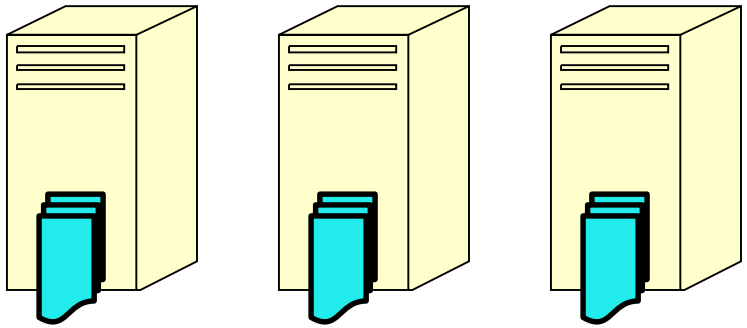
A parallel computer is a collection of processing elements that communicate and cooperate to solve large problems fast.

– 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