Parallel Computing (CS 633)

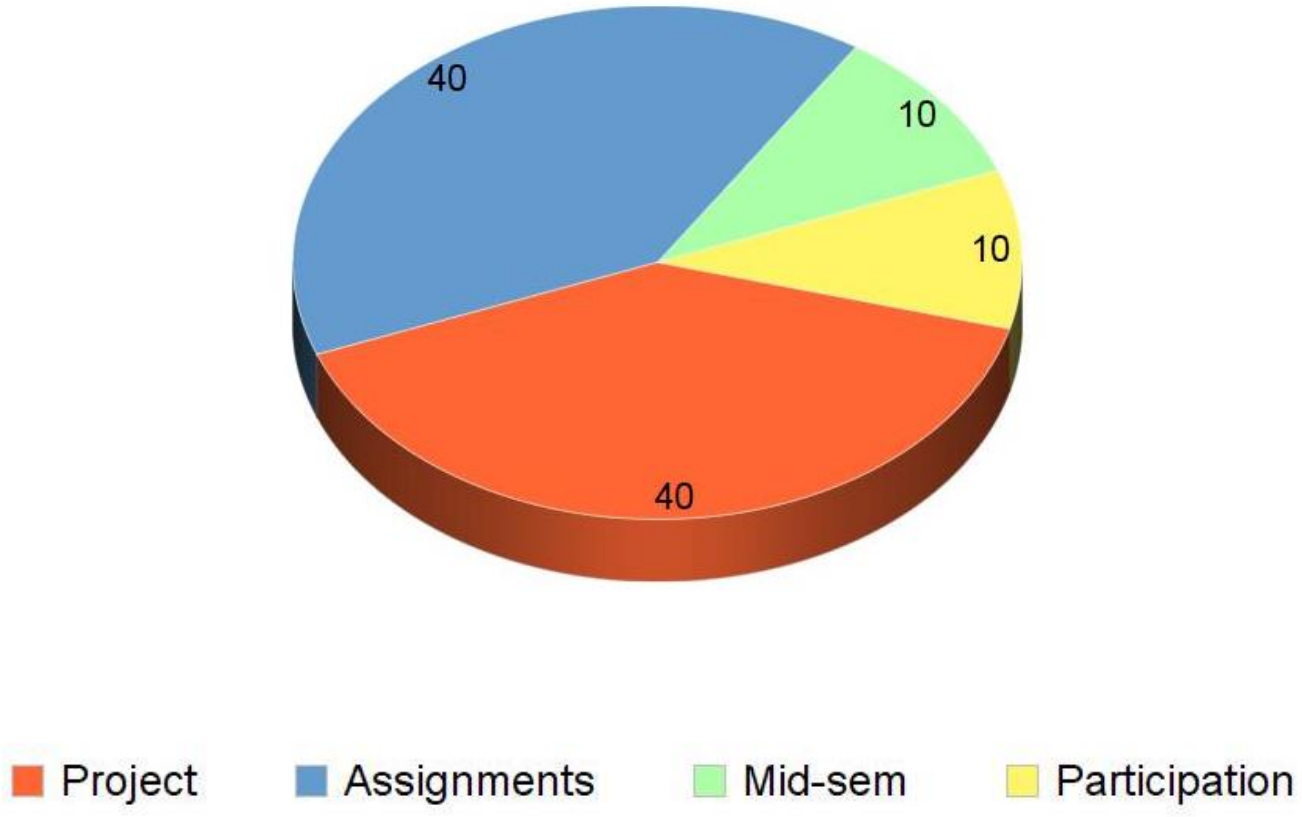
End of La-Z-Boy Programming Era – Dave Patterson

Preeti Malakar

Logistics

- T,F: 2 – 3:15 PM
- Register on Piazza
  - Will be used for announcements, Q&A etc.
- Office hours: W 10 AM – 12 PM (Appointment by email to *pmalakar@*)
- Email with prefix [CS633] in the subject
- https://web.cse.iitk.ac.in/users/pmalakar/cs633/2019

Grading



Programming

- Assignments
  - A total of 4 extra days may be taken
  - Credit for early submissions
  - Score reduction for late submissions
  - Through Gitlab
- Project
  - May select from the list
  - Discuss with me and finalize project topic by Jan 22
  - Weekly discussions on progress (Feb and Mar)
  - Report due on April 20
  - Final presentation in April
  - Credit on progress and novelty
- 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.
- Research papers

Lecture 1

Introduction

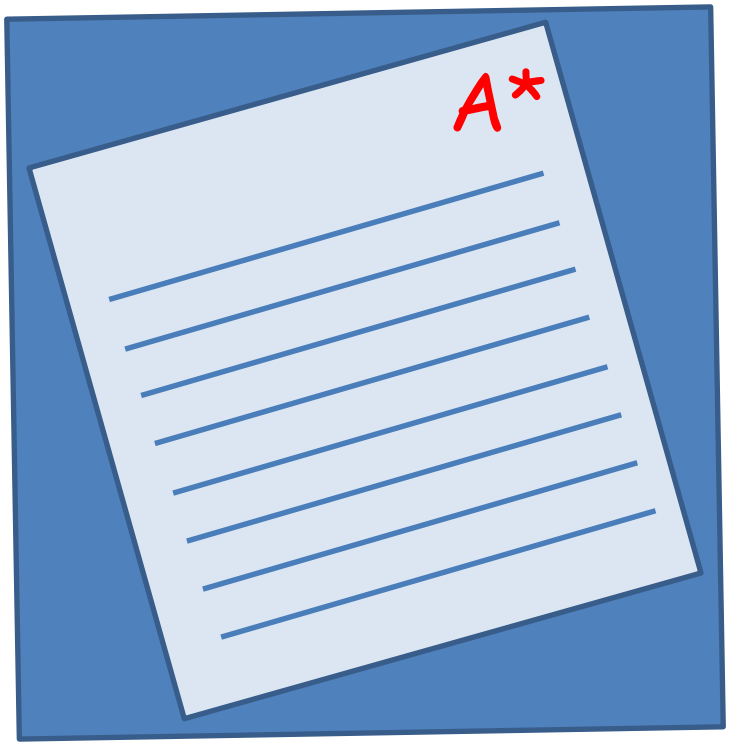
Jan 4, 2019

Parallelism

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

– Almasi and Gottlieb (1989)

Layman's Parallelism

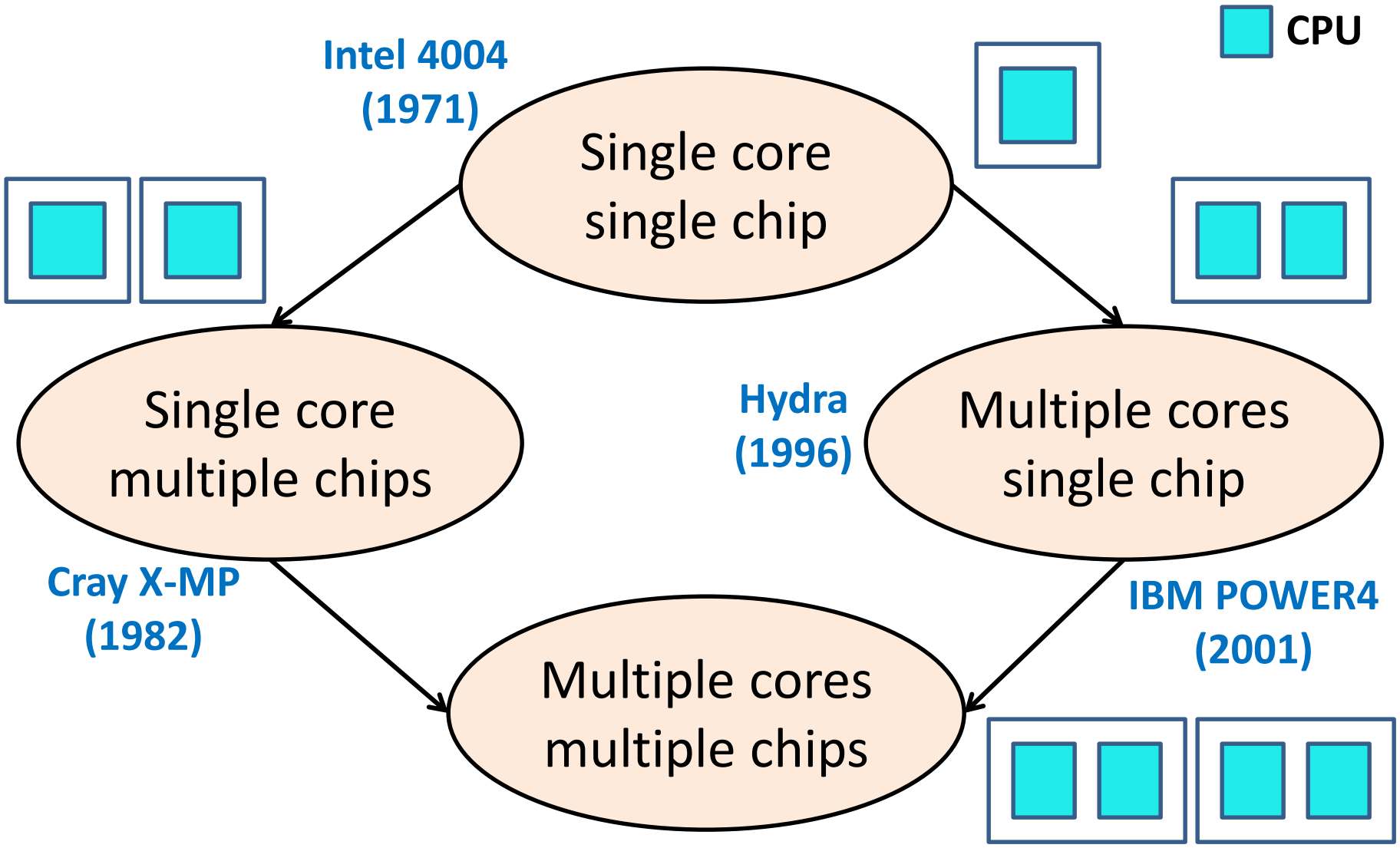


20 hours

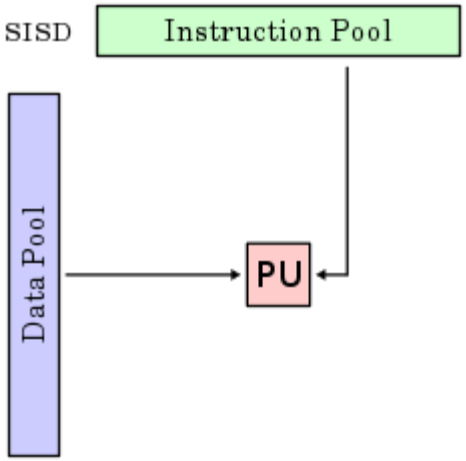
2 hours

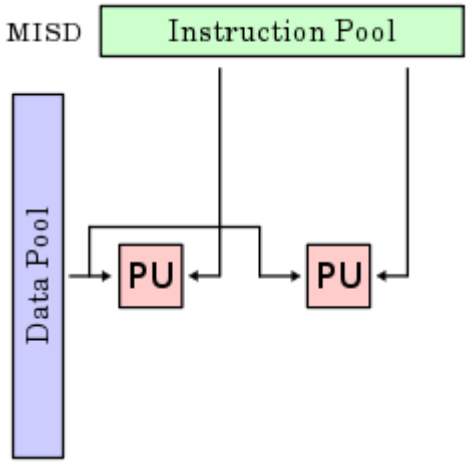
Really?

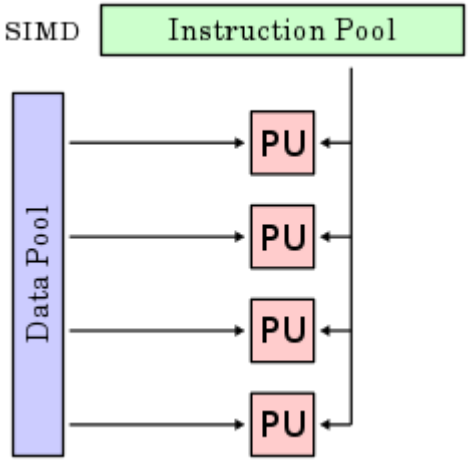
Multicore Era

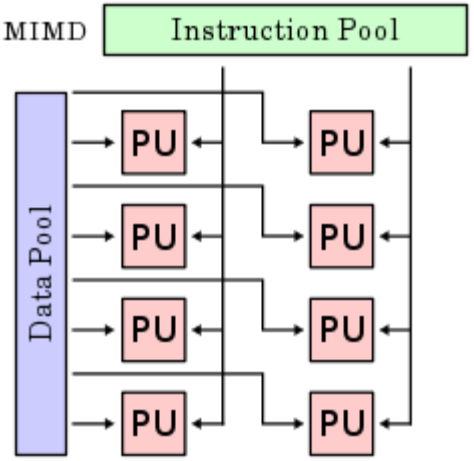


Flynn's Classification





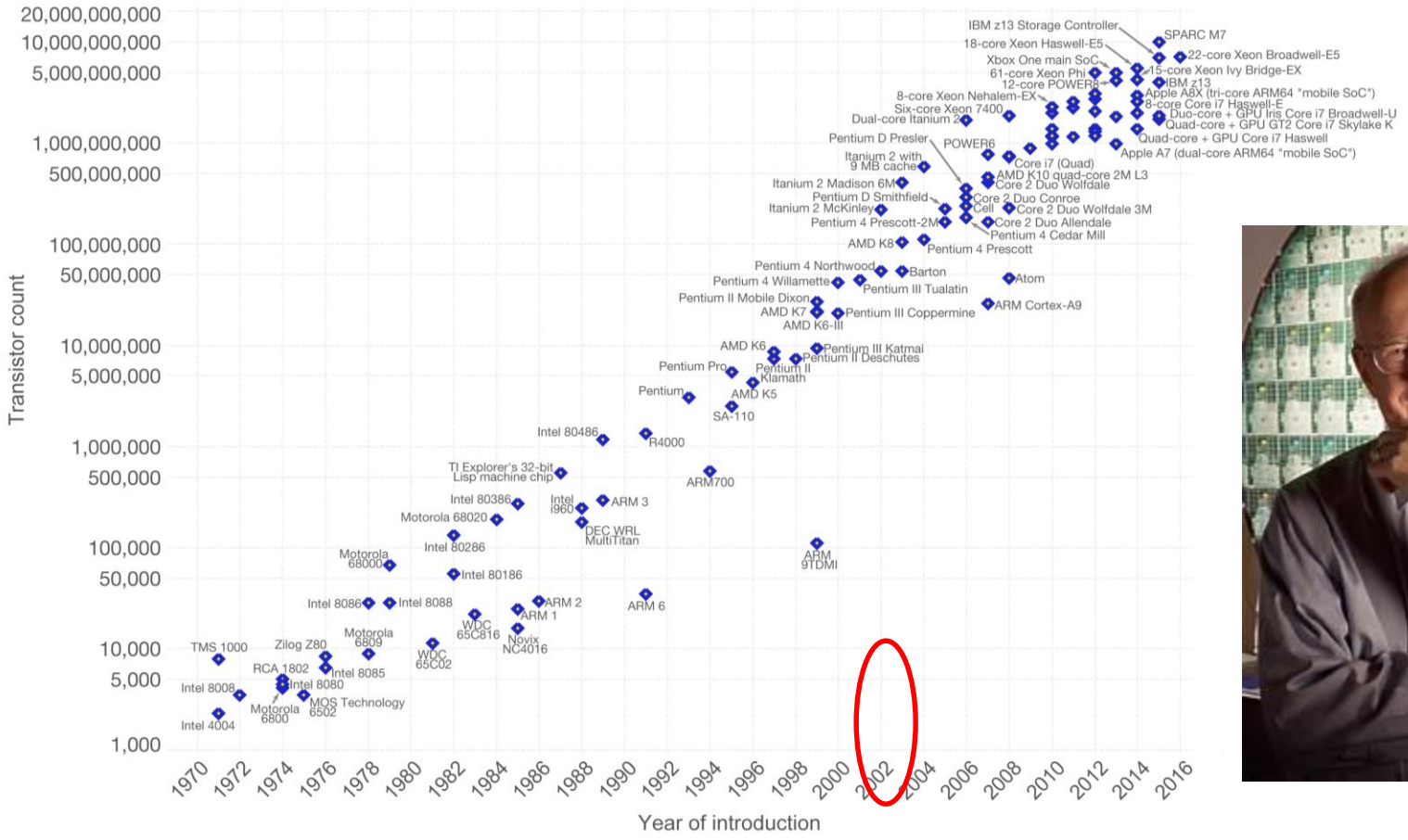




[Source: Wikipedia]

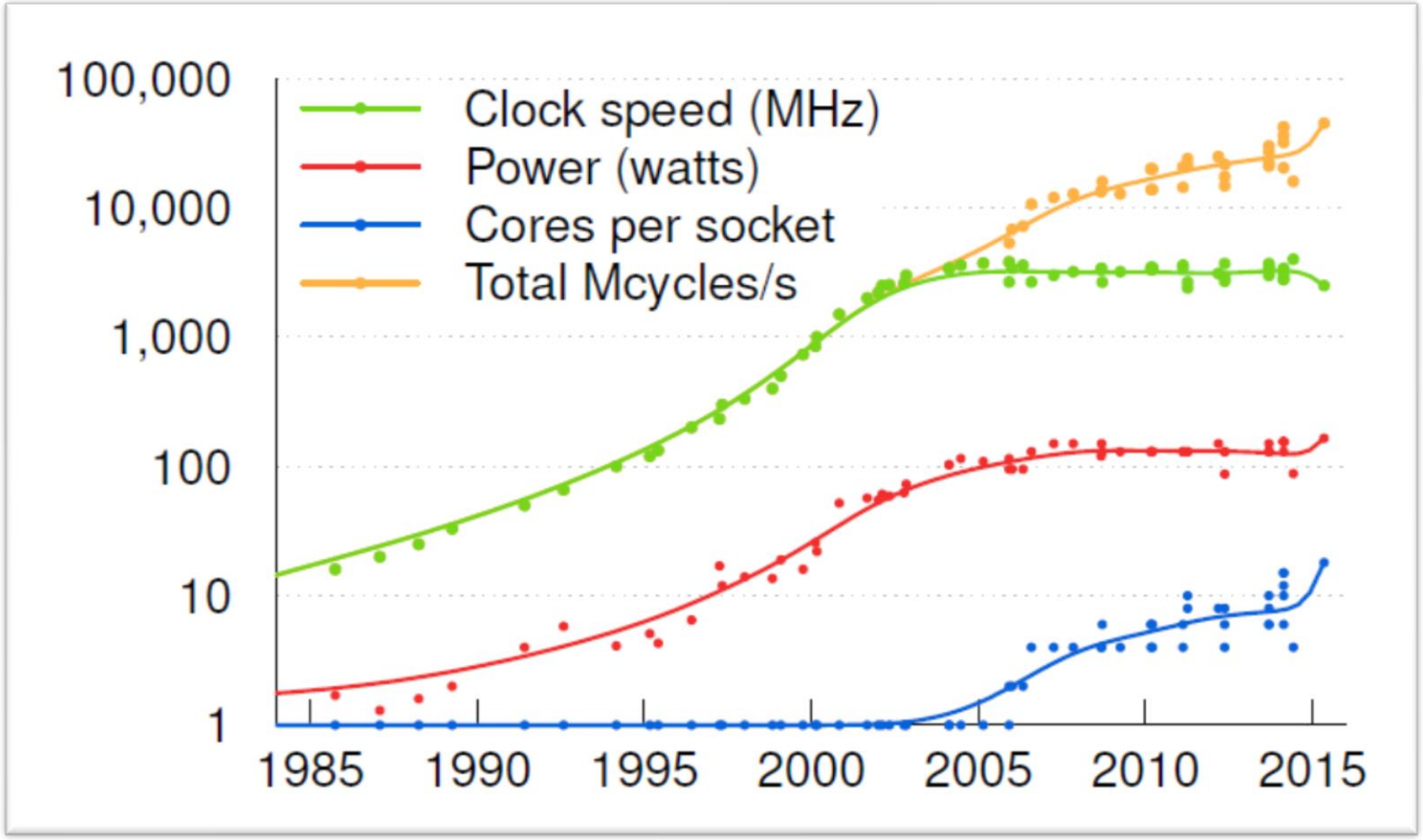
Moore's Law (1965)

Number of transistors in a chip doubles every 18 months



[Source: Wikipedia]

Trends



[Source: M. Frans Kaashoek, MIT]

Supercomputing [www.top500.org]

Rank	System	Cores
1	**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	2,397,824
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480
3	**Sunway TaihuLight** - Sunway MPP, Sunway , NRCPC National Supercomputing Center in Wuxi China	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Express-2, Matrix-2000 , NUDT National Super Computer Center in Gu China	
5	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6 NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	
6	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,072

~ $300 million
~ 9000 sq. ft.
~ 15 MW power
~ 15000 l of water

Supercomputing in India [topsc.cdacb.in]

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	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)
4	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
5	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
6	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
7	Center for Development of Advanced Computing (C-DAC), Pune	Param Yuva2 System (Intel Xeon E5-2670 (Sandy Bridge) @ 2.6 GHz dual octo-core CPU and Intel Xeon Phi 5110P dual 60-core co-processor nodes) w/Infiniband FDR OEM: Intel, Bidder: Netweb Technologies	3536C + 26520 ICO/ 442C + 442 ICO/ 221C + 221 ICO	388.44	520.40
8	CSIR Fourth Paradigm Institute (CSIR-4PI), Bangalore	HP Cluster Platform 3000 BL460c (Dual Intel Xeon 2.6 GHz eight core E5-2670 w/Infiniband FDR) OEM: HP, Bidder: HCL Infosystems Ltd.	17408/2176/1088	334.38	362.09
9	National Centre For Medium Range Weather Forecasting, Noida	IBM/Lenovo System X iDataPlex DX360M4, Xeon E5-2670 8C 2.6 GHz, Infiniband FDR OEM: IBM/Lenovo, Bidder: IBM India Pvt. Ltd.	16832/ /	318.4	350.1
10	Indian Institute of Technology, Kanpur	Cluster Platform SL230s Gen8, Intel Xeon E5-2670v2 10C 2.5 GHz, Infiniband FDR. OEM: HP, Bidder: HP	15360/1536/768	295.25	307.2

Intel Processors

- Intel 4004 (0.5 MHz, 1 core)
- …
- Sandybridge
- Ivybridge
- Haswell
- Broadwell
- Skylake (4 GHz, upto 22 cores)

Xeon Phi series

- KNC
- KNL
- KNM

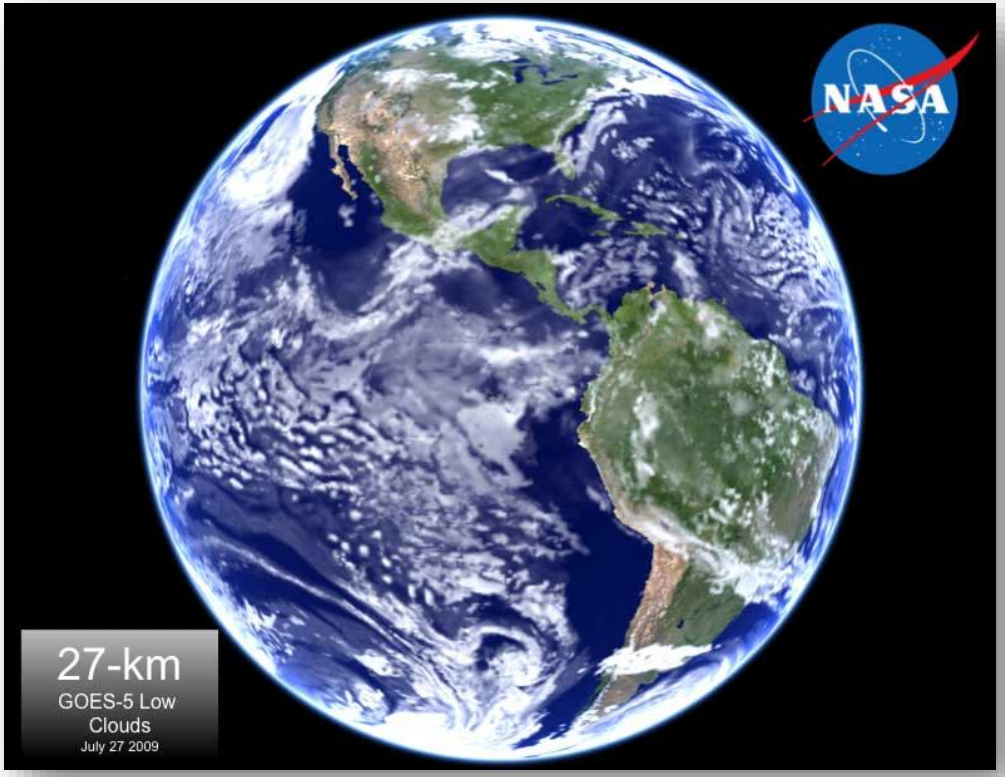
System Architecture Targets

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]

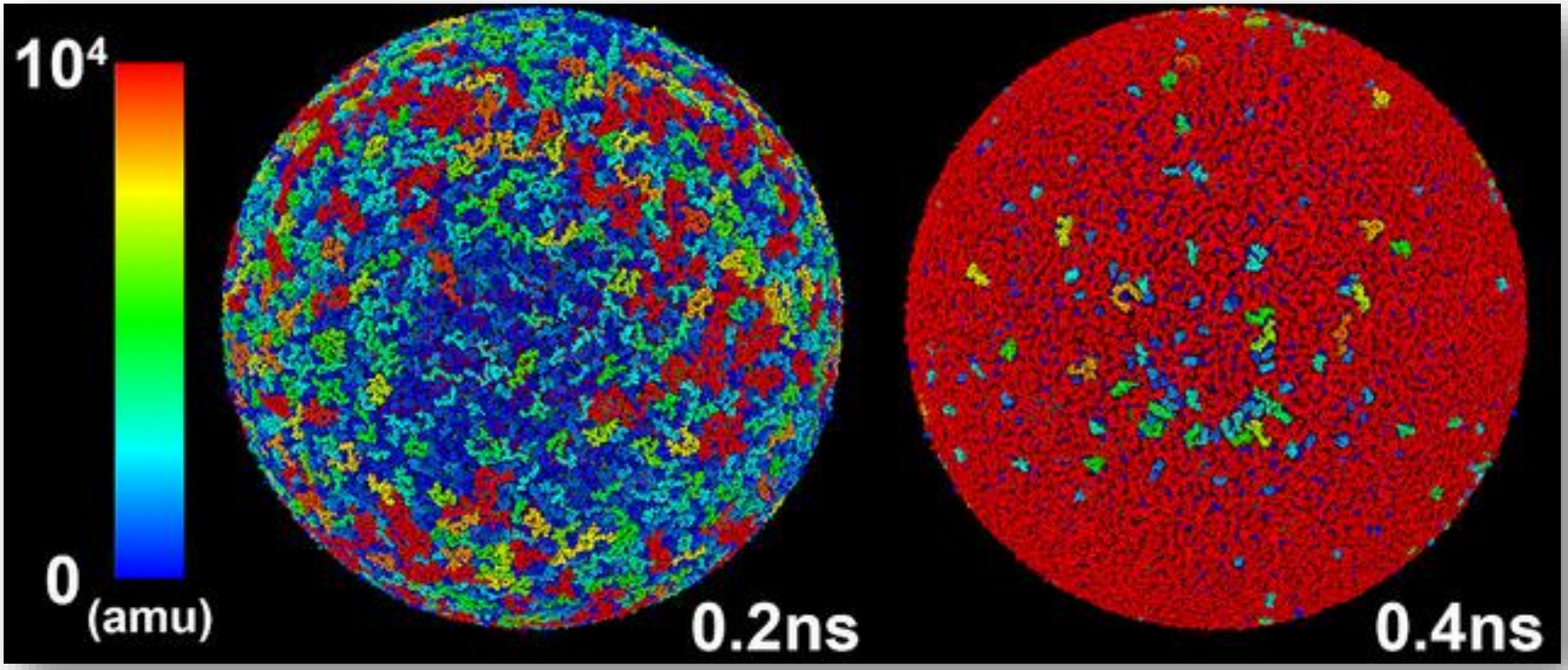
Big Compute

Massively Parallel Codes



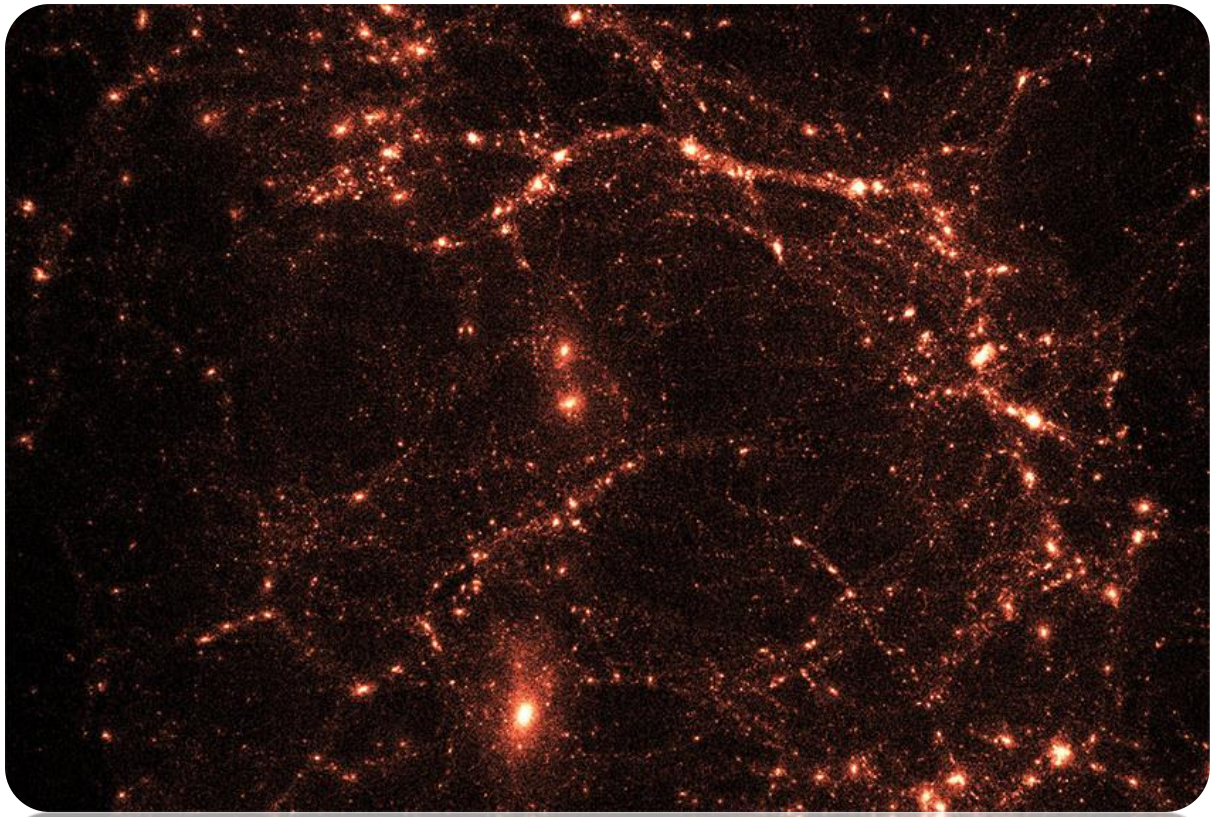
Climate simulation of Earth [Credit: NASA]

Massively Parallel Simulations



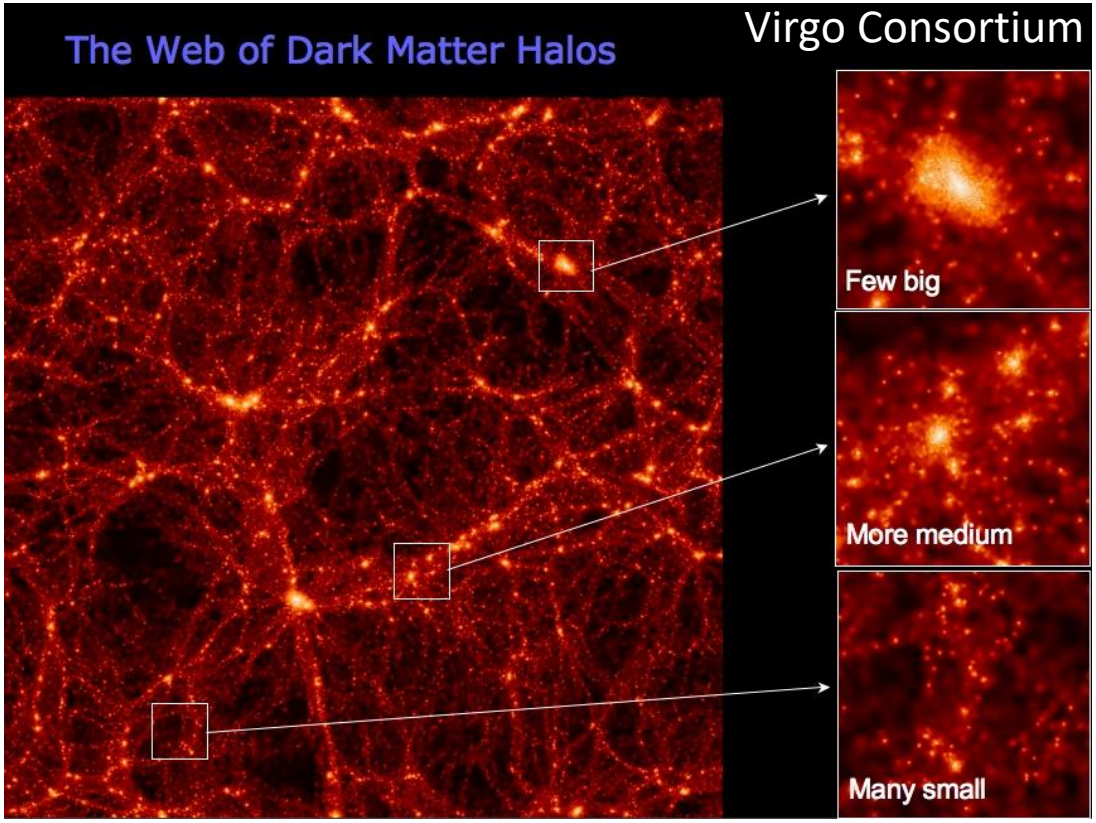
Self-healing material simulation [Nature, 2016]

Massively Parallel Codes

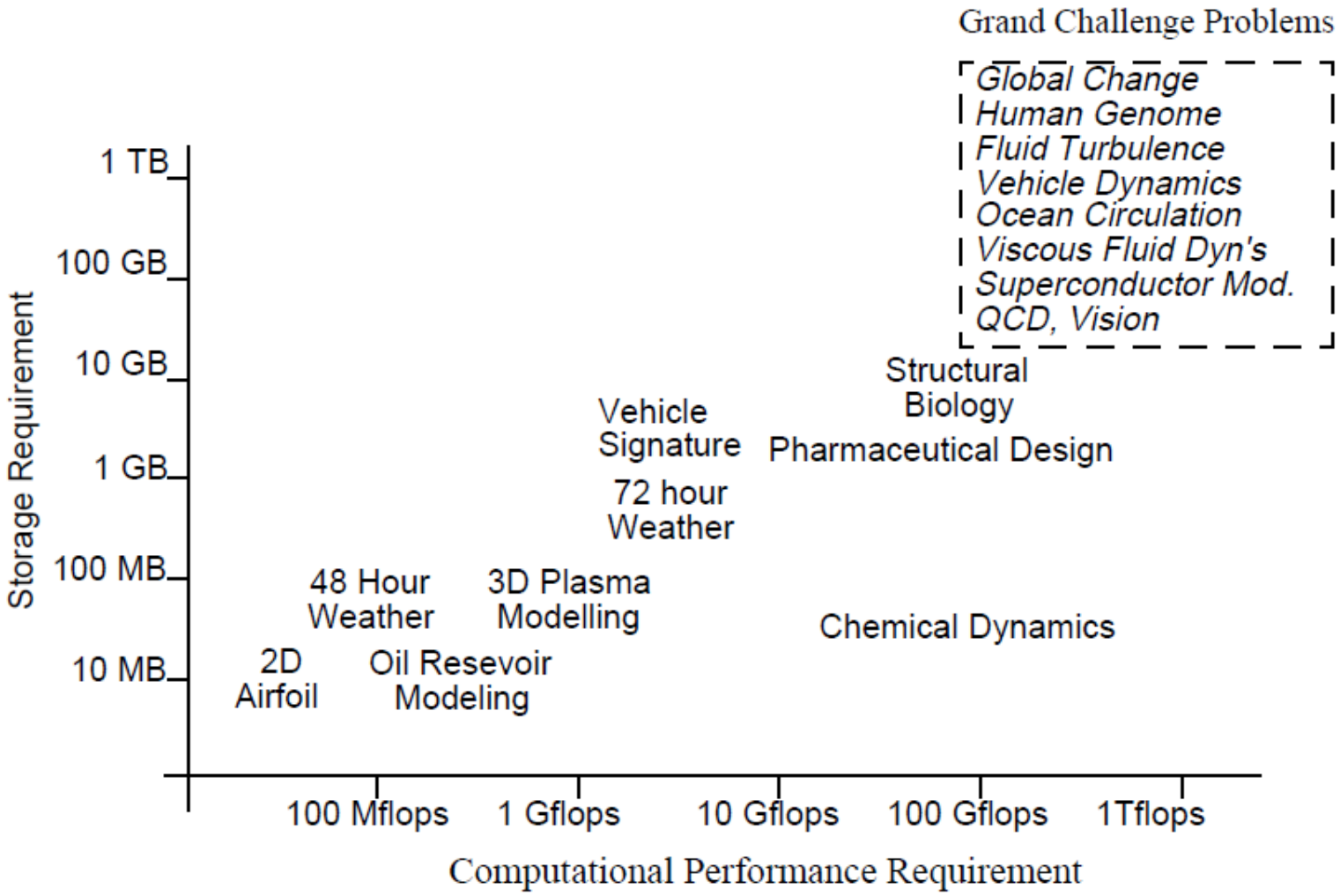


Cosmological simulation [Credit: ANL]

Massively Parallel Analysis



Computational Science

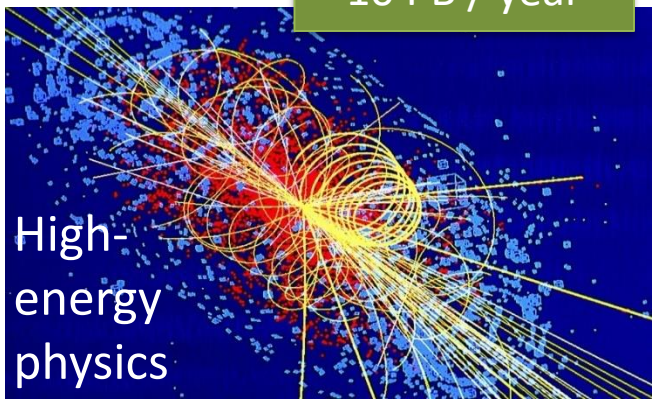


[Source: Culler, Singh and Gupta]

Big Data

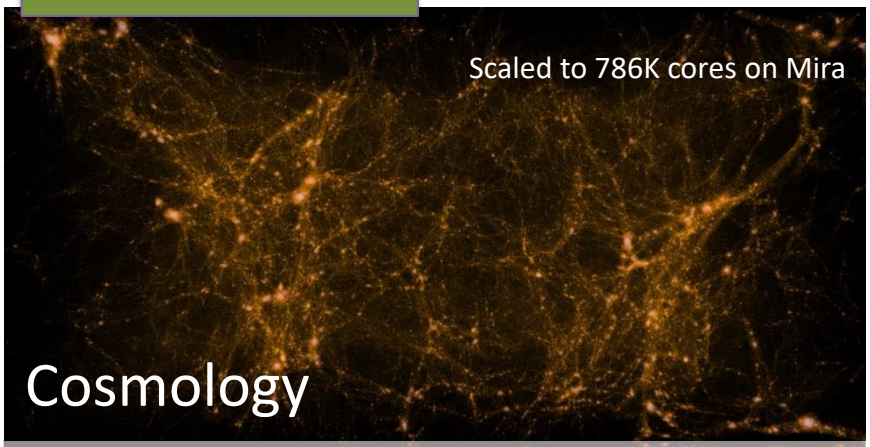
Output Data

10 PB / year



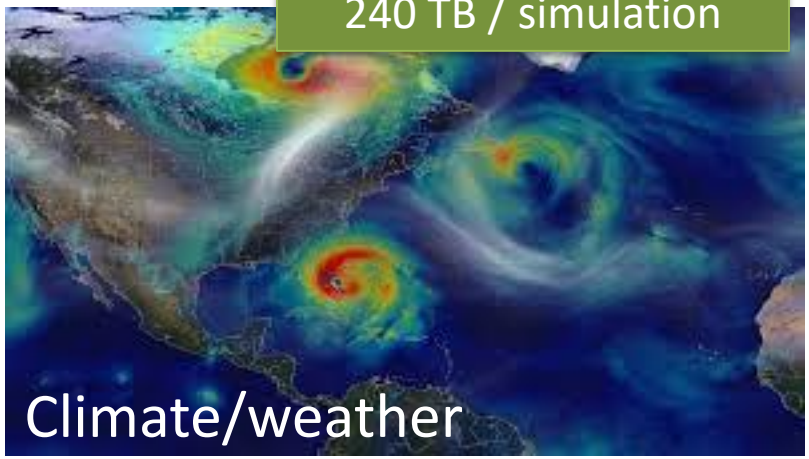
Higgs boson simulation
Source: CERN

2 PB / simulation



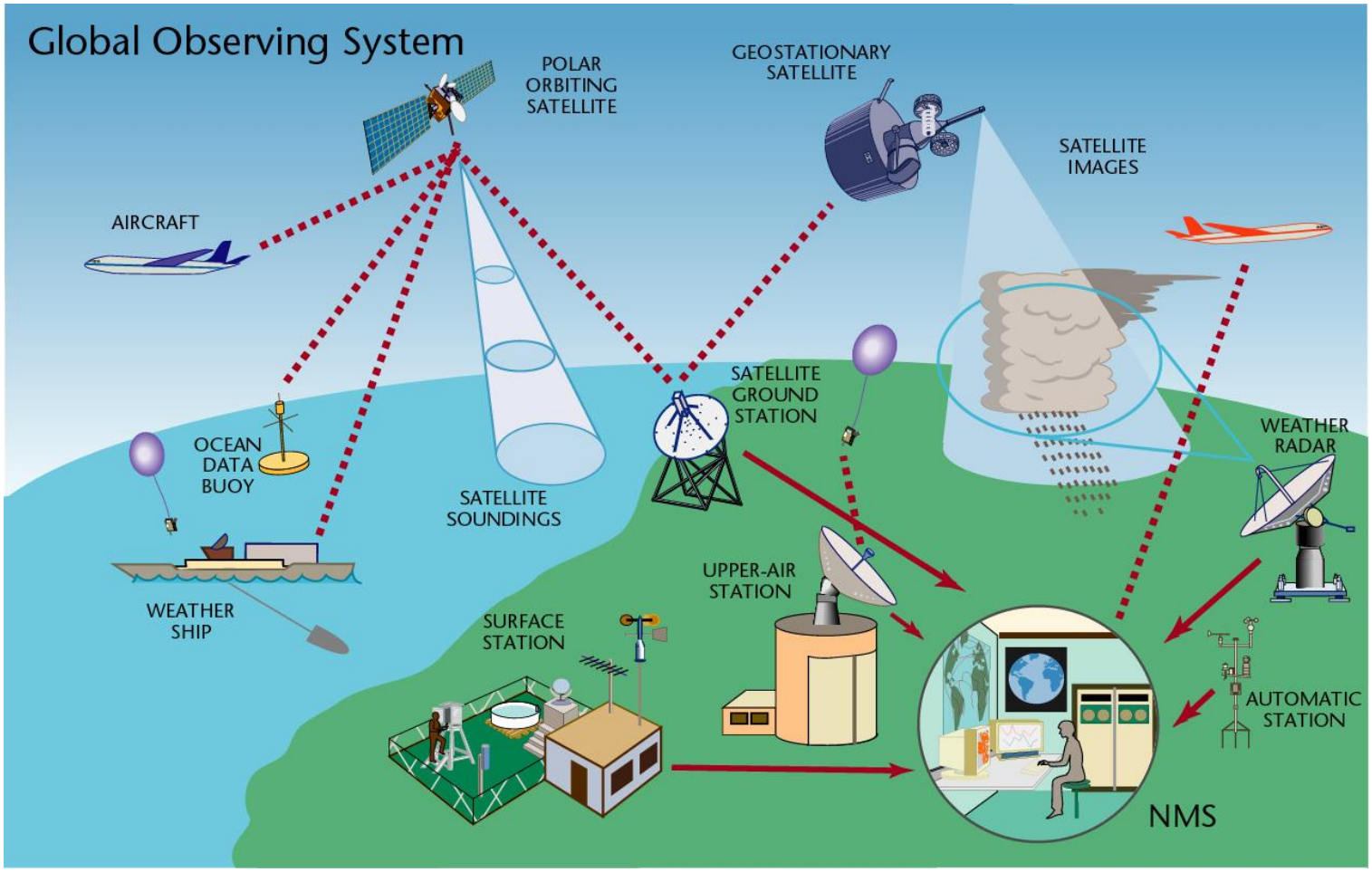
Q Continuum simulation
Source: Salman Habib et al.

240 TB / simulation



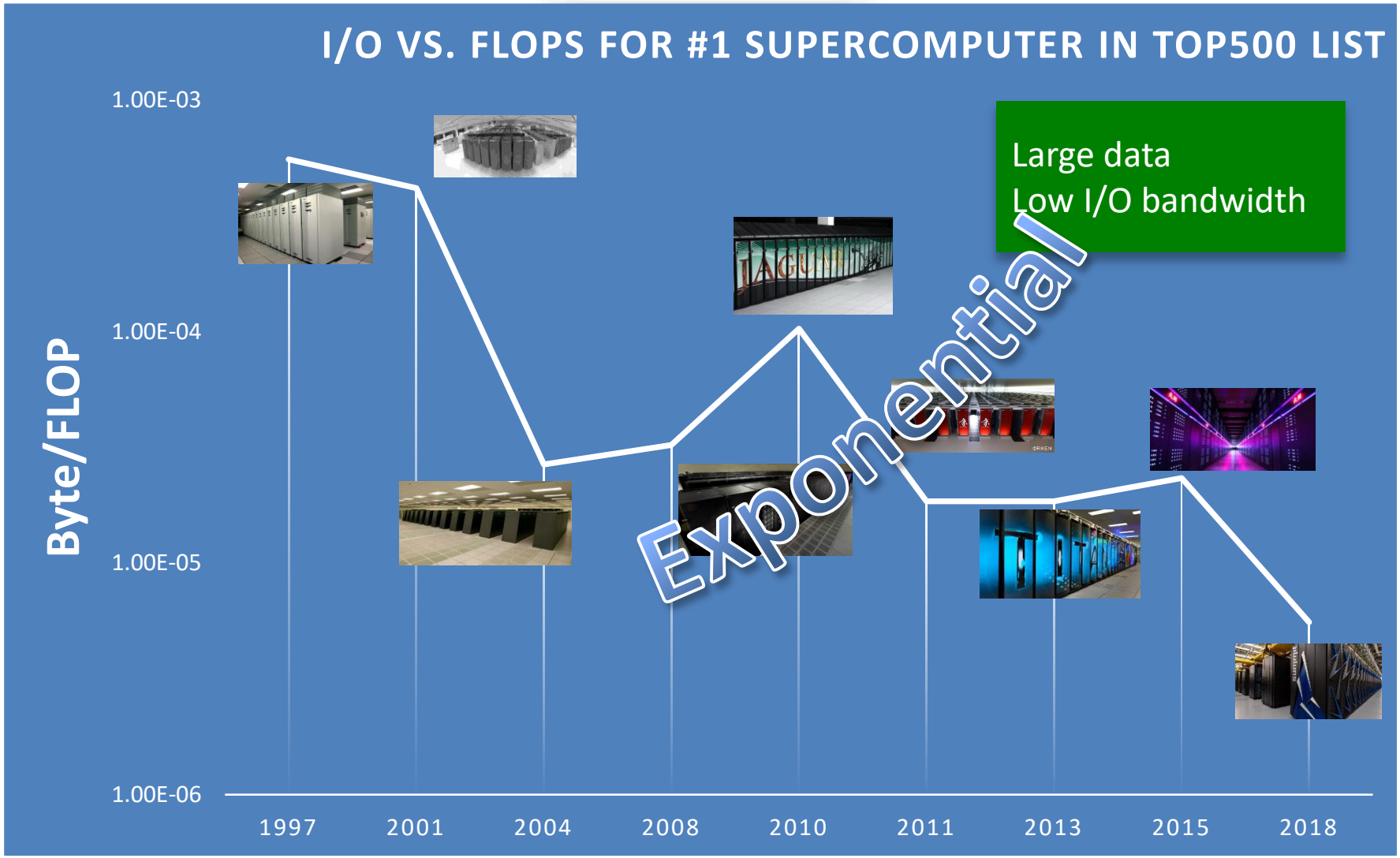
Hurricane simulation
Source: NASA

Input Data



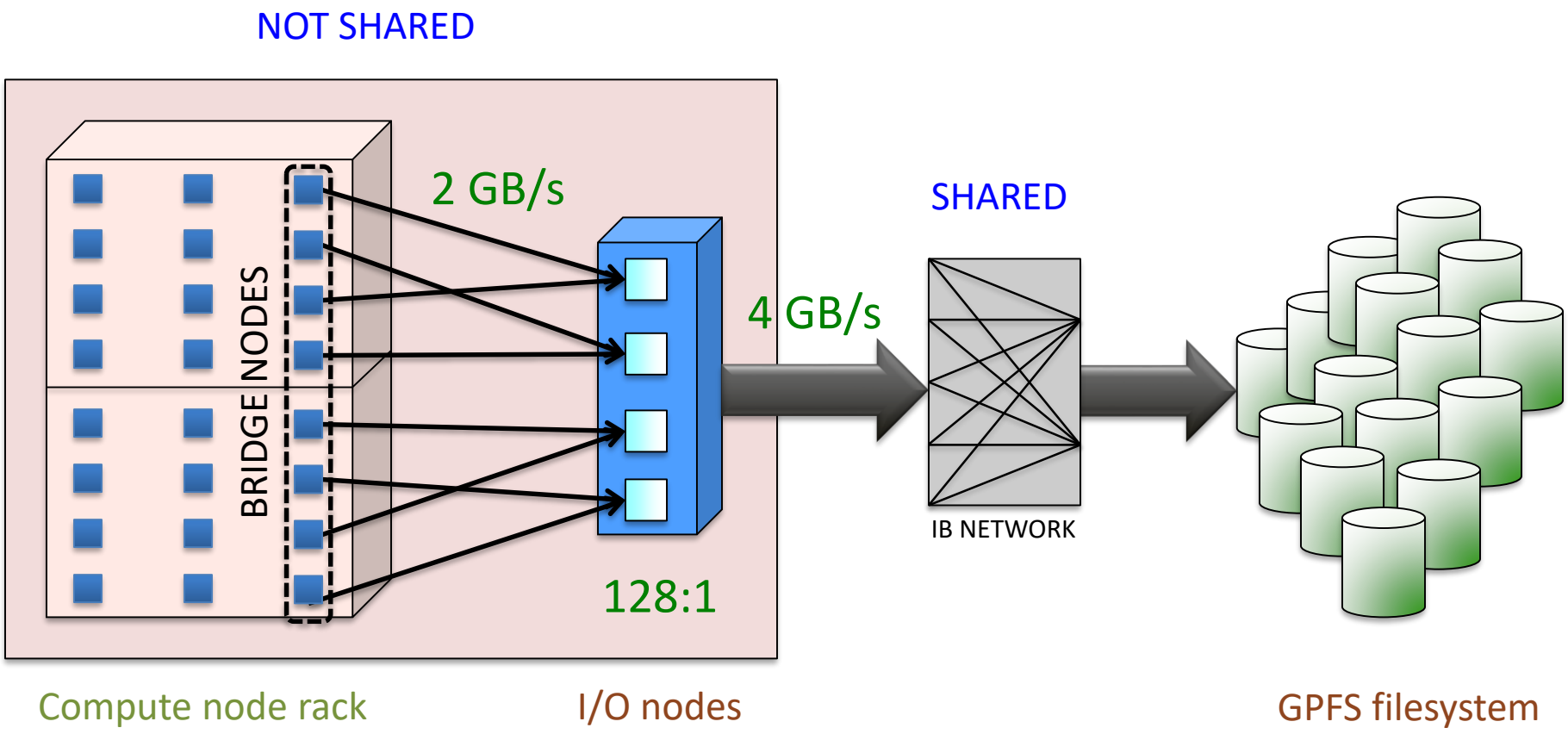
[Credit: World Meteorological Organization]

Compute vs. I/O trends



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

I/O node architecture on IBM Blue Gene/Q



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

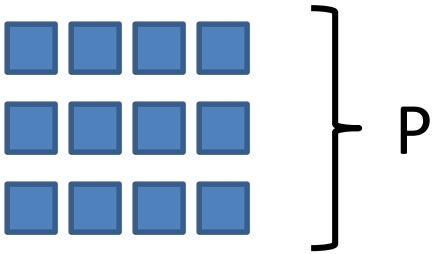
Serial

```
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

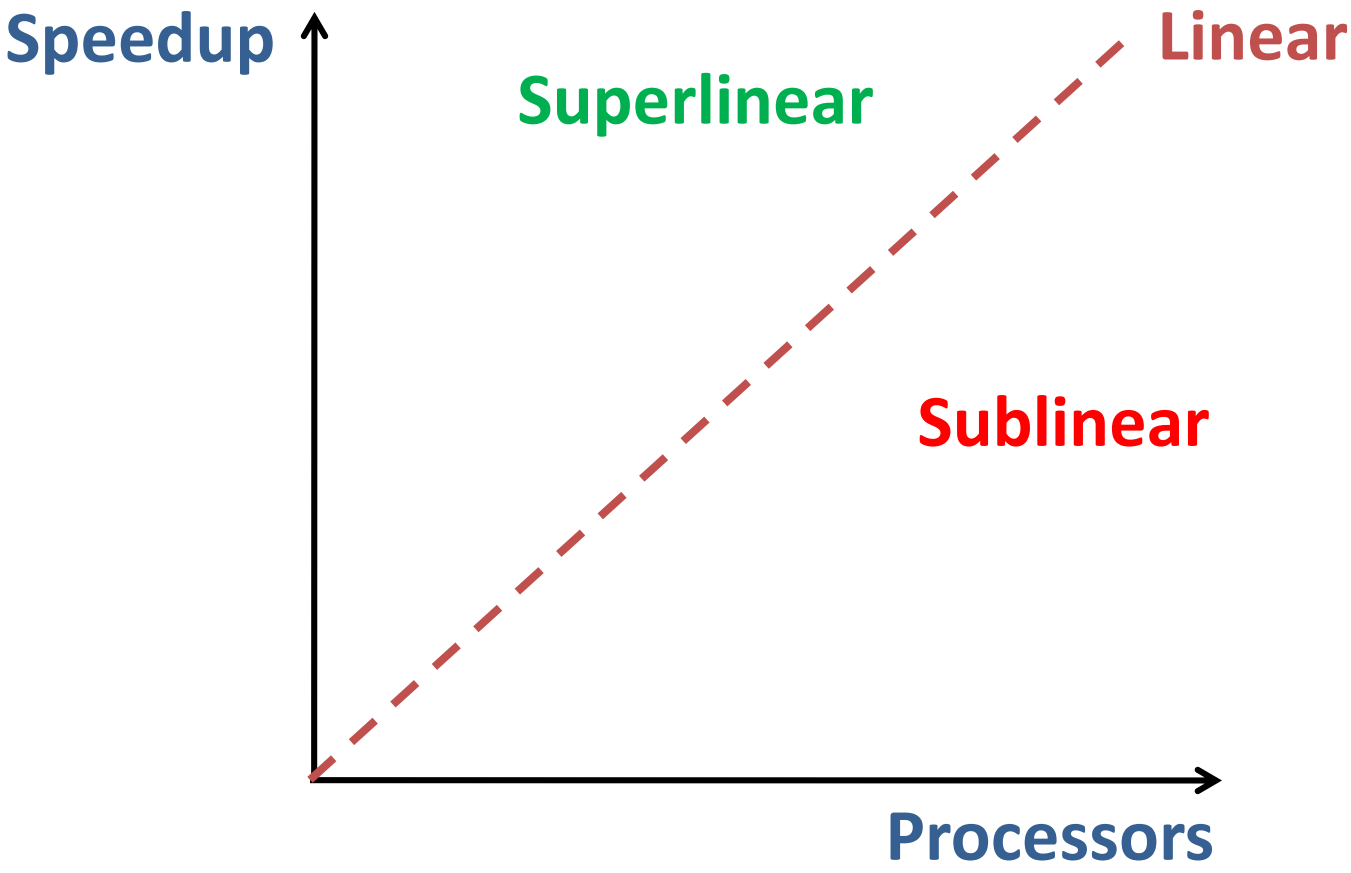
- Speedup

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

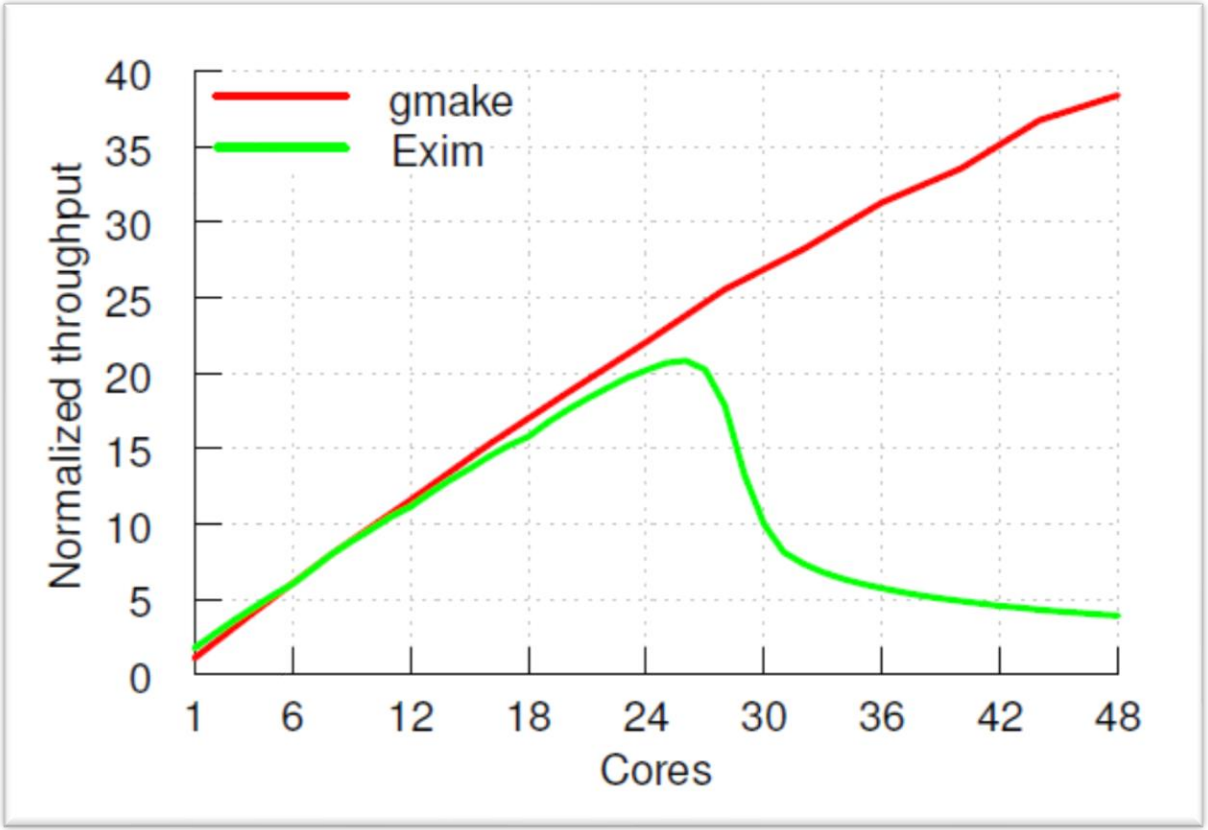
- Efficiency

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

Ideal Speedup

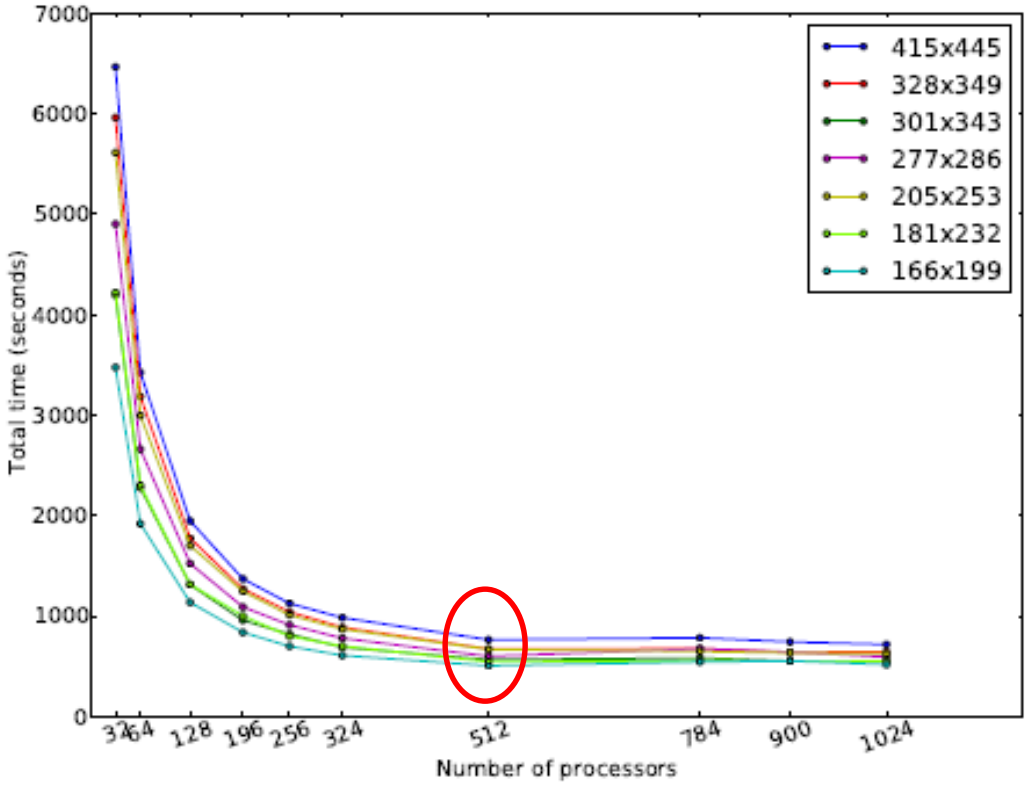


Issue – Scalability



[Source: M. Frans Kaashoek, MIT]

Scalability Bottleneck



Performance of weather simulation application

A Limitation of Parallel Computing

Amdahl's Law:

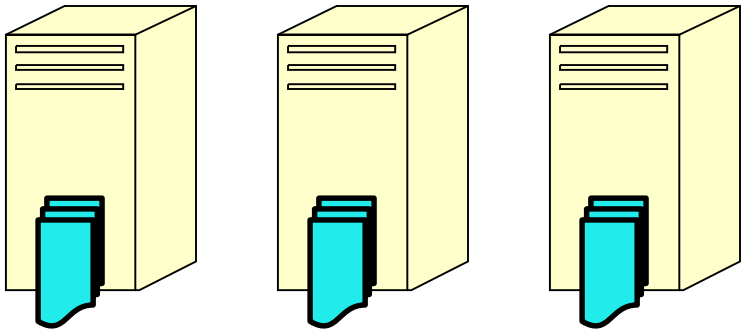
$$\text{Speedup}(f, S) = \frac{1}{(1-f) + \frac{f}{S}}$$

Parallelism

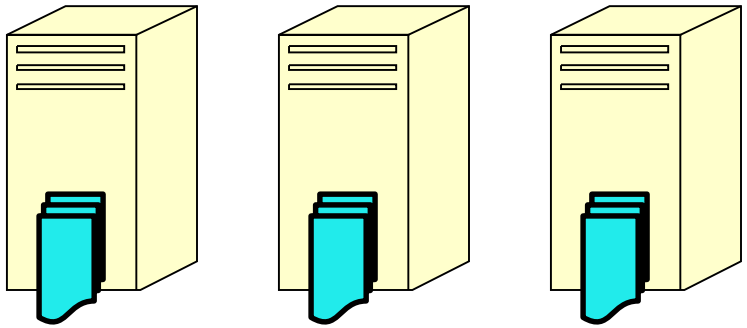
A parallel computer is a collection of processing elements that communicate and cooperate to solve large problems fast.

– Almasi and Gottlieb (1989)

Distributed Memory Systems



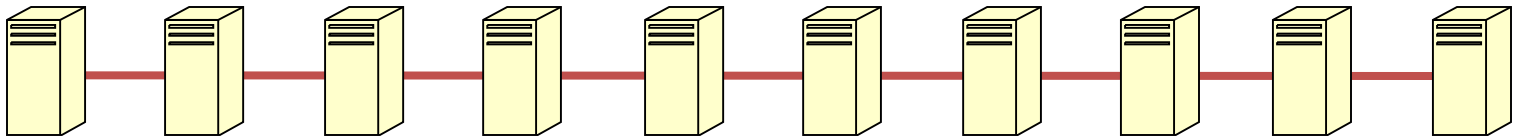
Node



Cluster

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

Interconnects

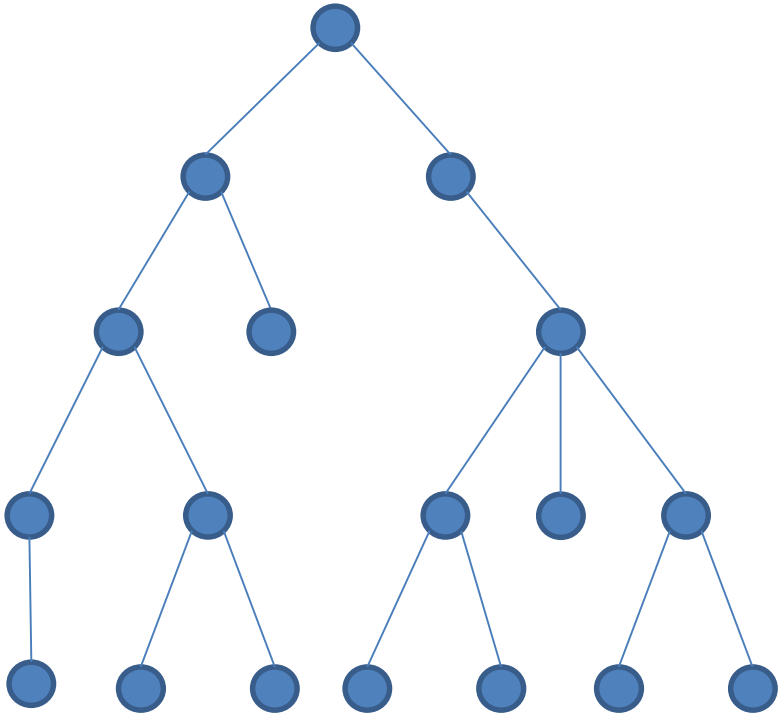


Linear Array

Attributes / Parameters

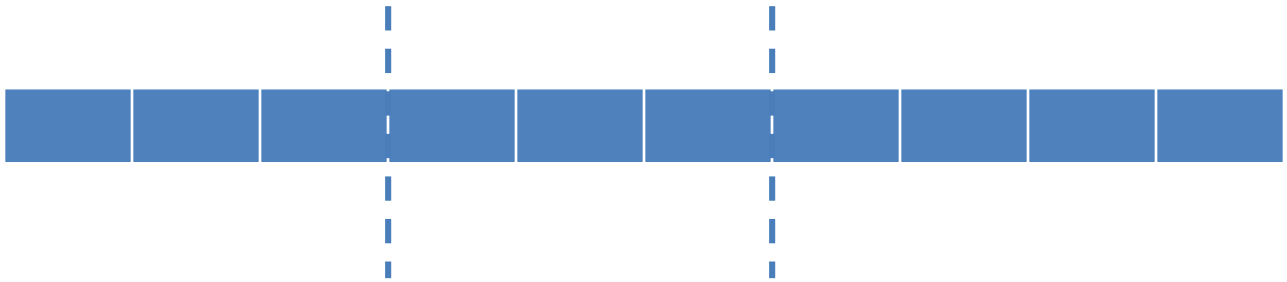
- Topology
- Diameter
- Bisection bandwidth

Task Parallel



- Group of tasks executed parallelly
- Optimal grouping
- Communications (inter and intra)

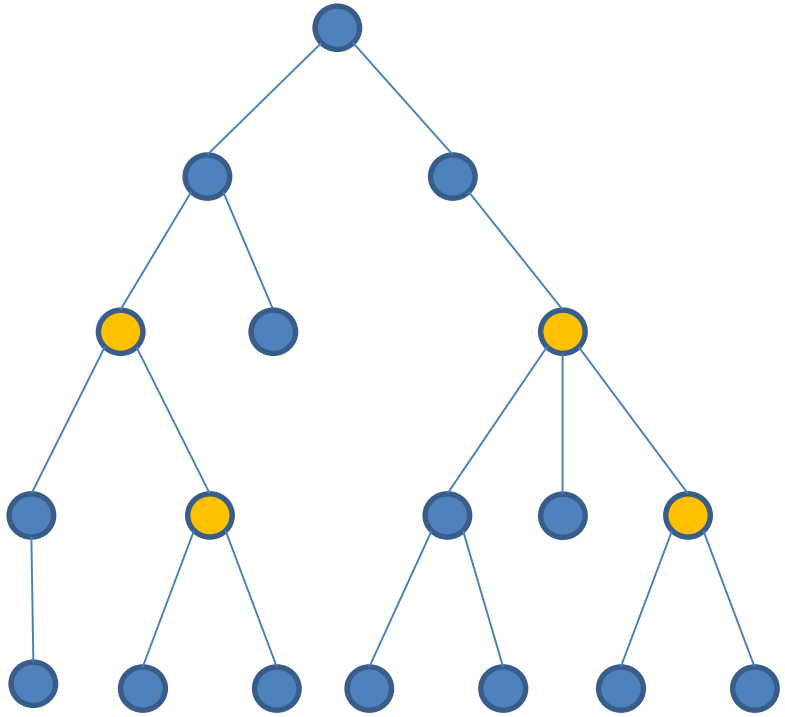
Data Parallel



- How many processes to use?
- Computation + communication cost

Load Balancing

- average computation time / maximum computation time



Job Scheduling

NODES

JOBS

USERS

Example of a real supercomputer activity

- Argonne National Laboratory Mira jobs

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

Message Passing Interface (MPI)

- Efforts began in 1991 by Jack Dongarra, Tony Hey, and David W. Walker.
- Standard for message passing in a distributed memory environment
- MPI Forum in 1993
  - Version 1.0: 1994
  - Version 2.0: 1997
  - Version 3.0: 2012

MPI Flavors

- MPICH (ANL)
- MVAPICH (OSU)
- Intel MPI
- OpenMPI

How to run MPI program?

MPI process MPI process MPI process MPI process
MPI process MPI process MPI process MPI process

4 nodes, ppn=2

mpiexec –n <number of processes> -f <hostfile> ./exe

<hostfile>

Host1:2
Host2:2
Host3:2
…

Assignment 0

0.1: Install MPICH v3.2.1.
0.2: Run examples/cpi and plot execution times for 1, 2, 4, 8 nodes with 4 ppn.

Submission:

(1) Installation steps. Summary of any hurdles you might have encountered while installation.

(2) Execution steps and speedup plot.

Due date: 11-01-2019

Quick start

- https://www.mpich.org/static/downloads/3.3/mpich-3.3-userguide.pdf

CSE Lab cluster

- ~ 30 nodes connected via Ethernet
- Each node has 12 cores
- Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz
- NFS filesystem
- Enable passwordless ssh (ssh-keygen)

General Instructions

- https://git.cse.iitk.ac.in
- Create private project named ‘CS633-2018-19-2’ and add pmalakar and TAs
- Create subdirectories for each assignment
- Include hostfile and job script in your submissions
- Run your code on CSE lab cluster
  - IP addresses: 172.27.19.2-8,10-13,15-18,20-24,27,30
  - Email if you find most of these are unreachable