Introduction to parallel computing

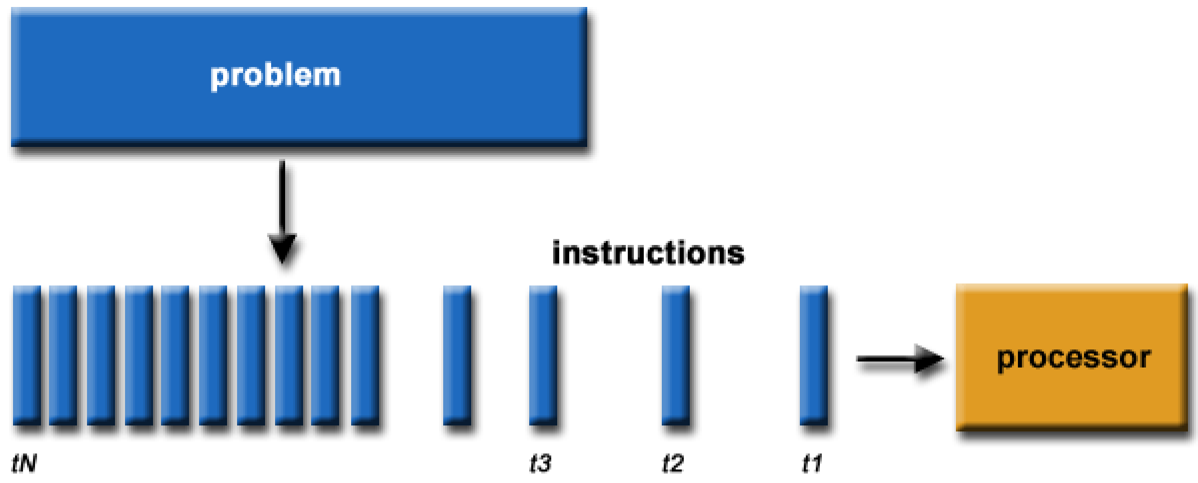
MM 2090 : Introduction to Scientific Computing

Gandham Phanikumar

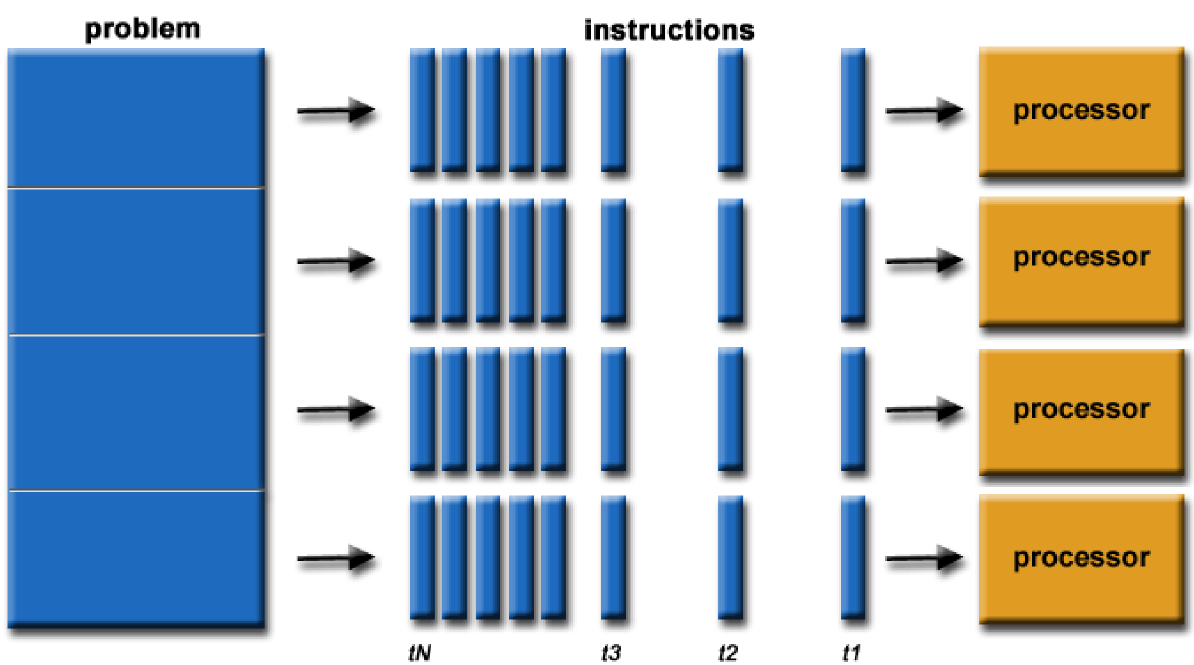
Acknowledgments:

* Blaise Barney, Lawrence Livermore National Laboratory for his article linked at https://computing.llnl.gov/tutorials/parallel_comp/
* Some images are from the internet.
* Presentation on HPC by P.G. Senapathy Centre for Computing Resources, IITM

Serial Computing



Parallel Computing

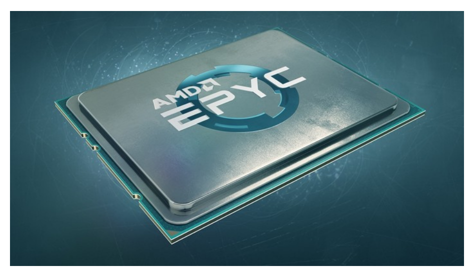


Single processors are parallel computers

Intel® Core™ i9-9920X X-series Processor

- 19.25 MB SmartCache Cache
- 12 Cores
- 24 Threads
- 4.40 GHz Max Turbo Frequency
- X - Extreme performance and mega-tasking, unlocked
- 9th Generation

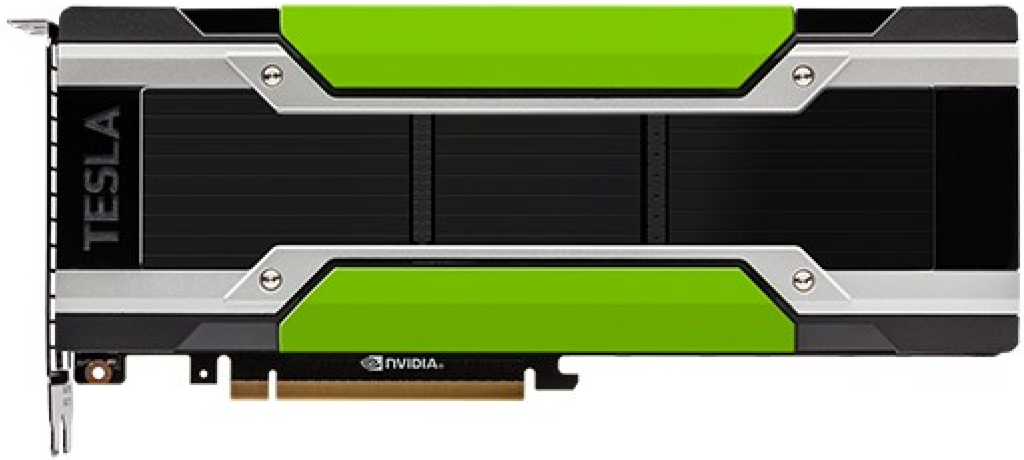
28 Cores
56 Threads



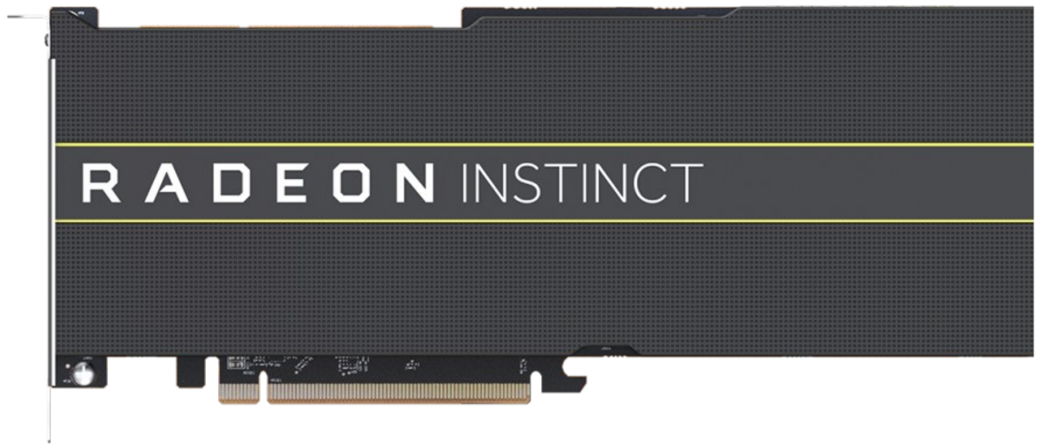
AMD EPYC 7601
32 Cores
64 Threads

AMD RYZEN
Threadripper 2990WX
32 Cores
64 Threads

GPUs are parallel computers

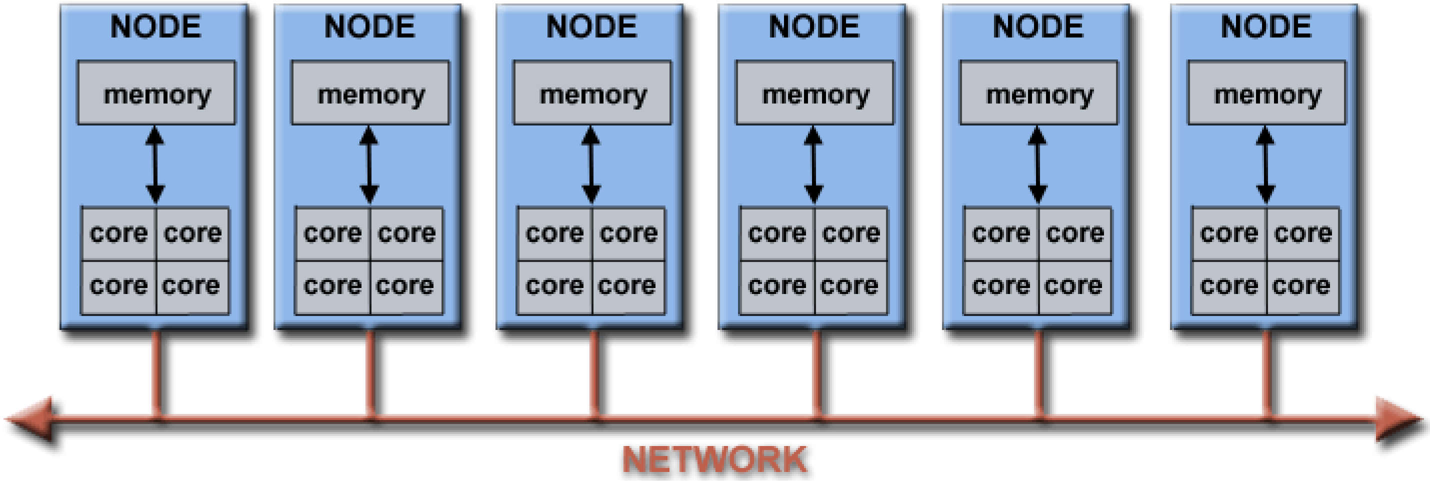


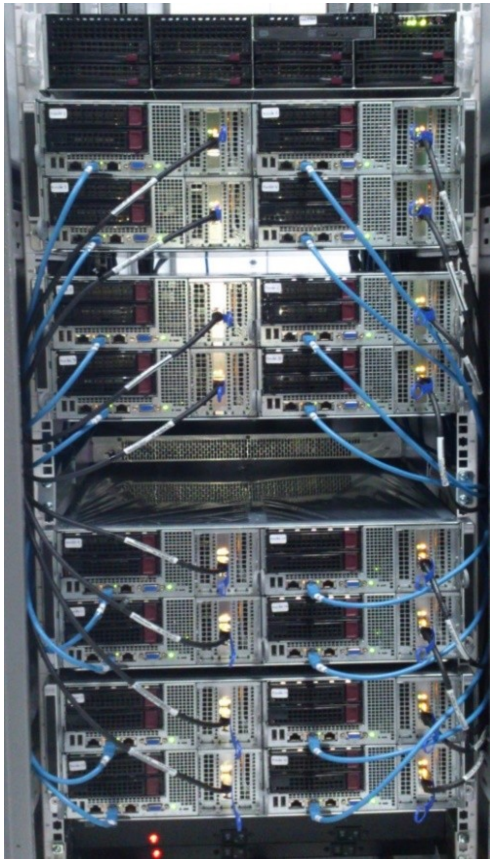
NVidia Tesla P100
3584 Cuda Cores
9.3 Teraflops in single precision
4.7 Teraflops in double precision
32 GB/s PCIe bandwidth



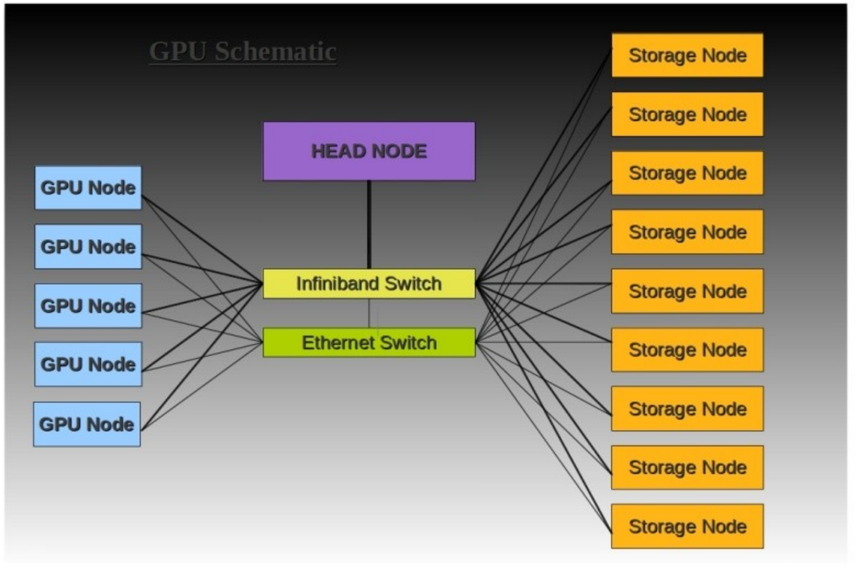
Radeon Instinct MI60
4096 Stream Processors
14.7 Teraflops in single precision
7.4 Teraflops in double precision
100 GB/s PCIe bandwidth

Networked computers are parallel computers





Libra Cluster at P.G.Senapathy Centre for Computing Resources, IIT Madras



Internet is also a Parallel Computing System

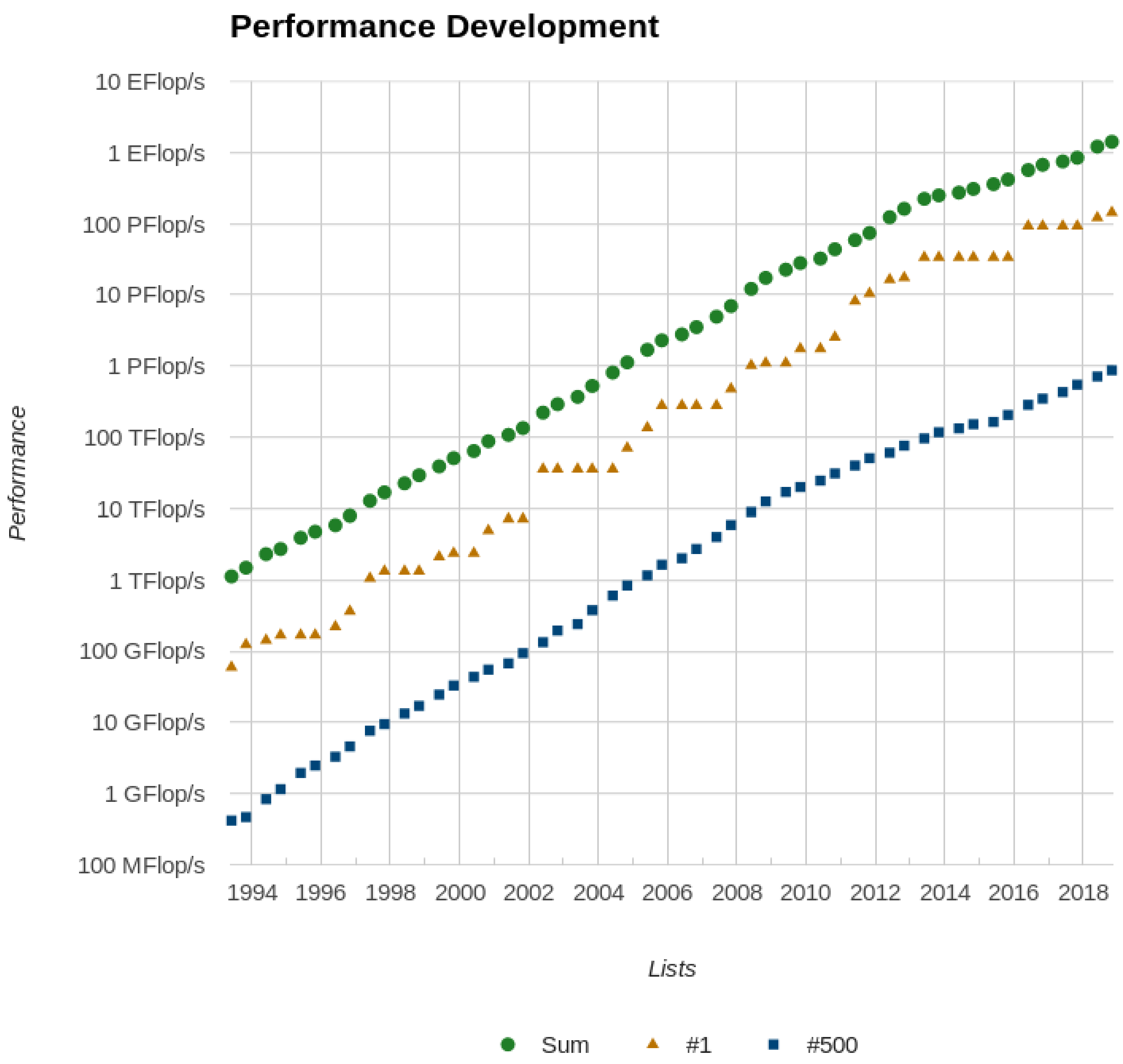
https://foldingathome.org/

Folding@home is a project focused on disease research. The problems we're solving require so many computer calculations – and we need your help to find the cures!

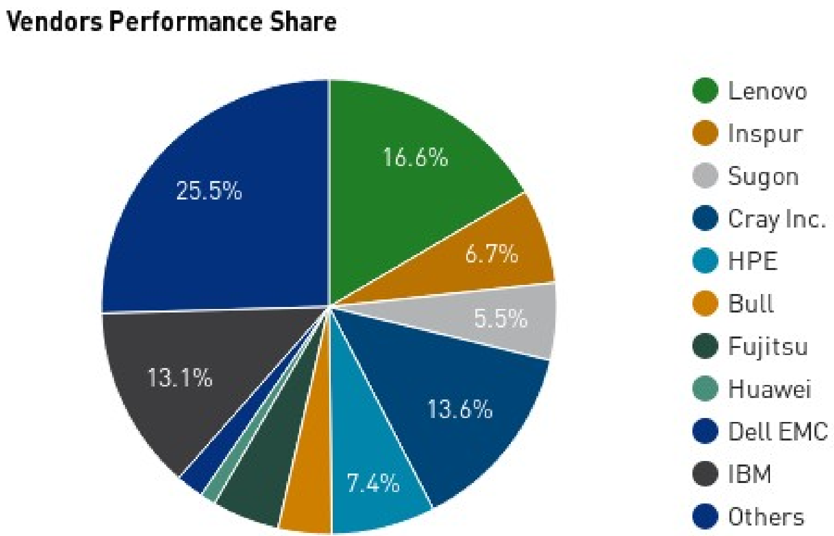
http://setiathome.berkeley.edu/

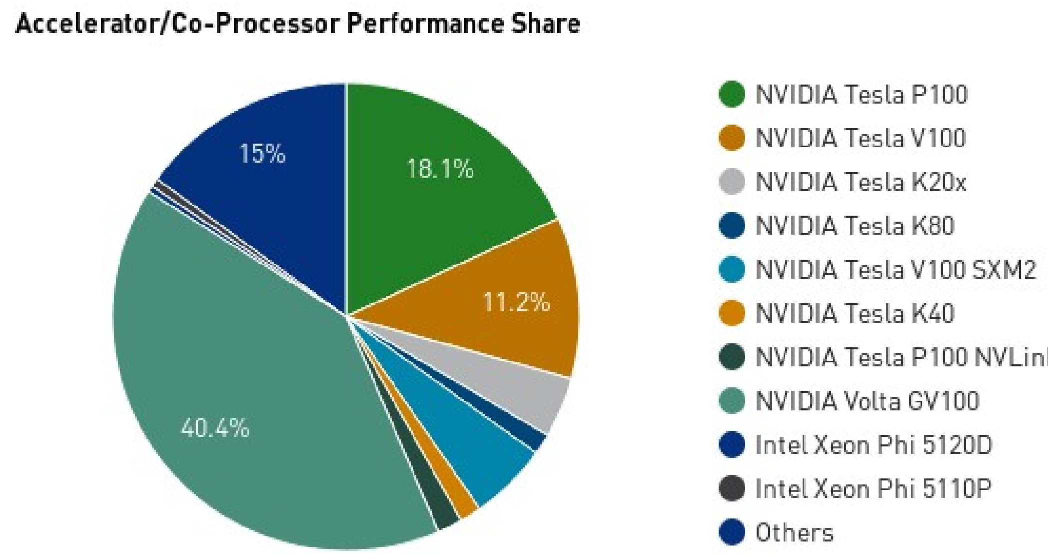
SETI@home is a scientific experiment, based at UC Berkeley, that uses Internet-connected computers in the Search for Extraterrestrial Intelligence (SETI). You can participate by running a free program that downloads and analyzes radio telescope data.

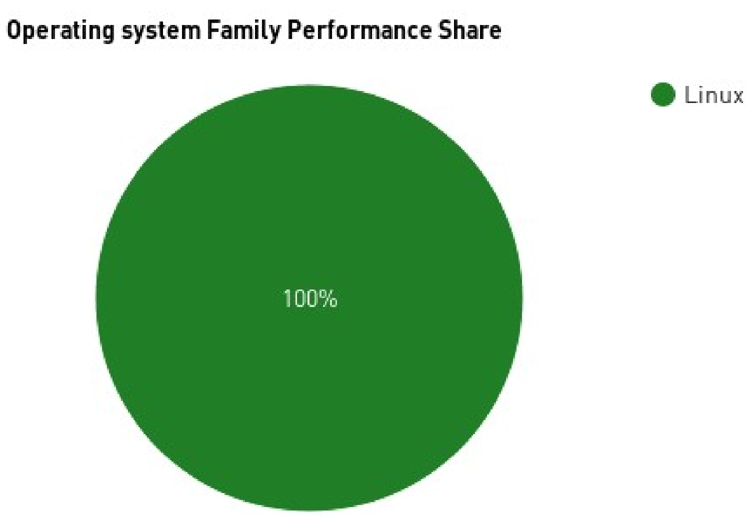
Statistics from top500.org as in Nov-2018

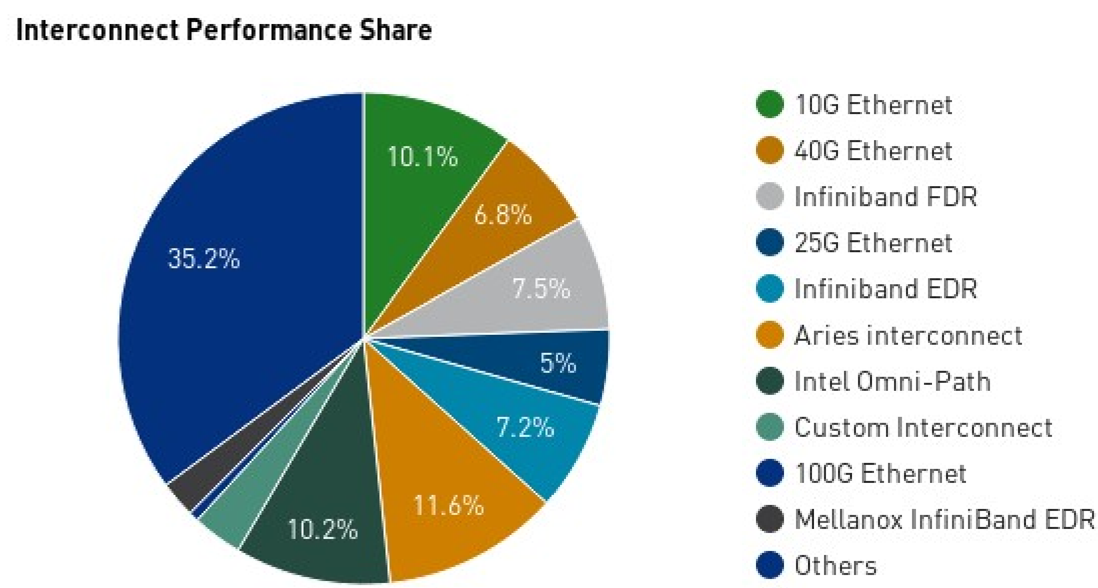


Statistics from top500.org as in Nov-2018

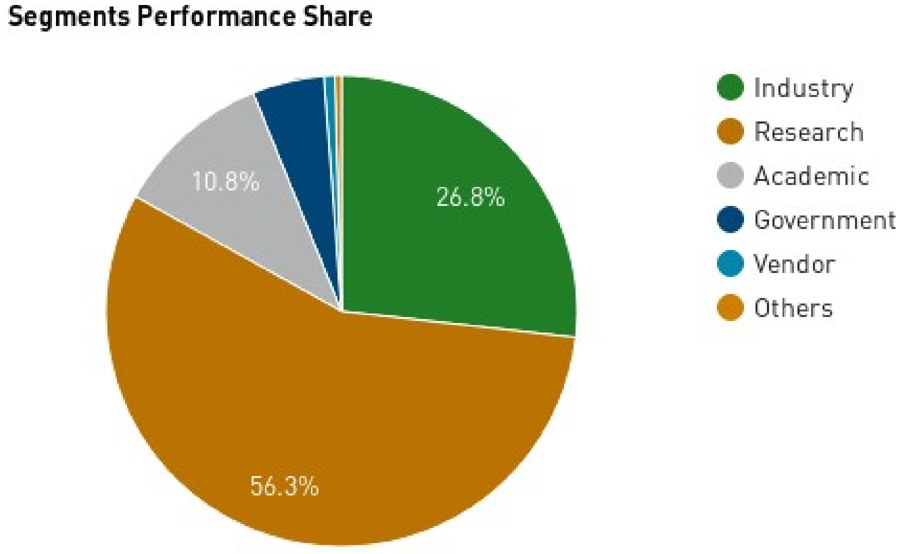


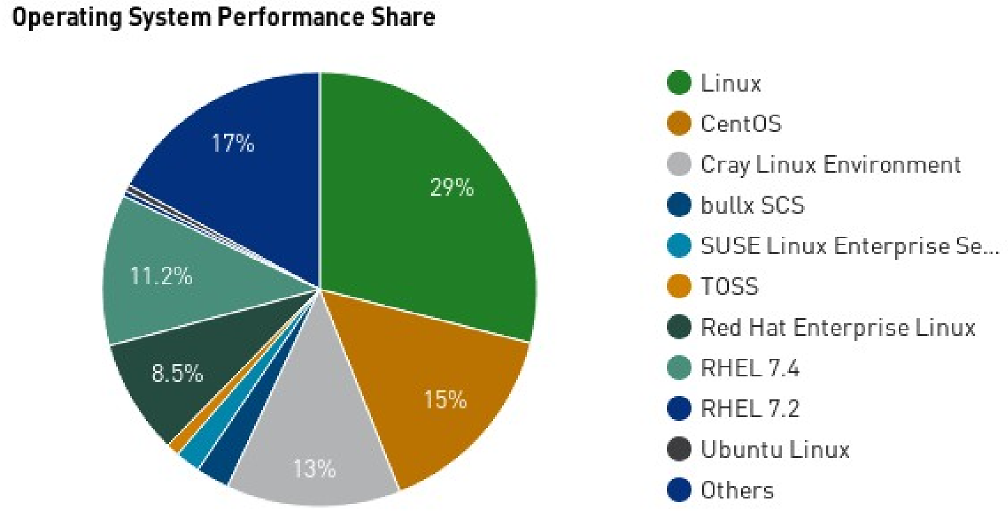






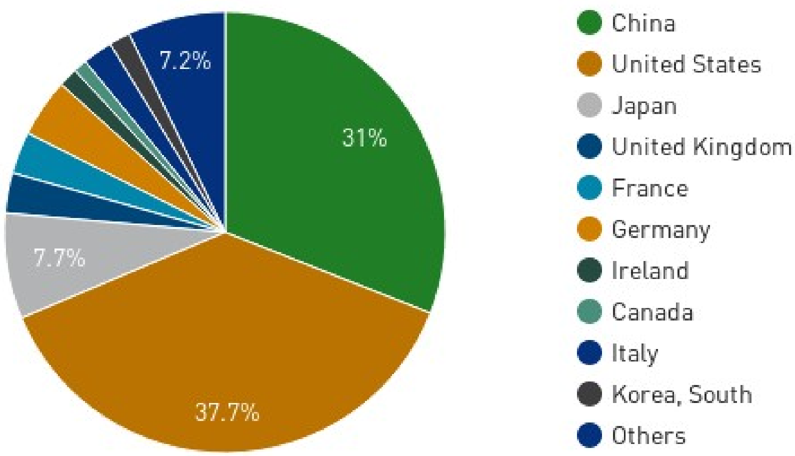
Statistics from top500.org as in Nov-2018





Programmers who use Linux in R&D units of
Industry, Academia and Government Labs

Countries Performance Share



India constitutes 0.8% of total HPC

Countries	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
China	227	45.4	438,228,339	806,368,243	26,632,672
United States	109	21.8	533,209,190	757,357,100	16,101,360
Japan	31	6.2	109,436,242	170,880,045	5,710,372
United Kingdom	20	4	41,729,303	52,509,525	1,625,892
France	18	3.6	43,580,345	66,598,837	1,792,656
Germany	17	3.4	60,502,637	86,333,952	1,575,350
Ireland	12	2.4	19,789,320	25,436,160	691,200
Canada	9	1.8	14,027,780	22,258,586	436,640
Italy	6	1.2	31,110,650	49,243,746	814,864
Korea, South	6	1.2	21,938,000	35,760,556	804,740
Netherlands	6	1.2	9,334,060	11,925,504	326,880
Australia	5	1	6,669,188	10,232,963	257,336
Poland	4	0.8	4,604,365	6,216,160	153,128
Sweden	4	0.8	4,653,054	6,565,116	139,408
India	4	0.8	8,358,996	9,472,166	272,328
Singapore	3	0.6	4,308,220	5,525,299	146,112
Russia	3	0.6	4,580,250	7,940,005	178,180
Saudi Arabia	3	0.6	10,109,130	13,858,214	325,940
Switzerland	2	0.4	23,126,750	29,347,305	453,140
South Africa	2	0.4	2,152,470	2,779,930	71,256
Spain	2	0.4	7,488,800	11,781,642	172,656
Taiwan	2	0.4	10,325,150	17,297,190	197,552
New Zealand	1	0.2	908,892	1,425,408	18,560
Norway	1	0.2	953,571	1,081,651	32,192
Brazil	1	0.2	1,123,150	1,413,120	38,400
Finland	1	0.2	1,250,000	1,689,293	40,608
Czech Republic	1	0.2	1,457,730	2,011,641	76,896

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
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	143,500.0	200,794.9	9,783
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	94,640.0	125,712.0	7,438
3	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	387,872	21,230.0	27,154.3	2,384

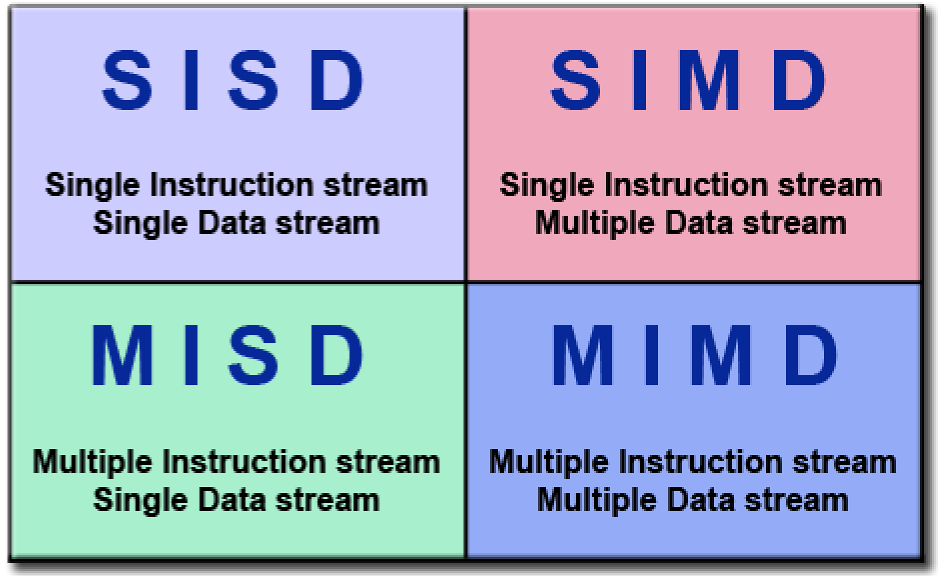
Common Feature of top ranked supercomputers:

They are hybrid systems (Multiprocessor + Accelerator) running Linux

Job Schedulers

LoadLeveler	IBM	Renamed to Tivoli Workload Scheduler LoadLeveler
Terascale open-source resource and queue manager (TORQUE)	Adaptive Computing	Since 2003
Open PBS (Portable Batch System)	Free and opensource	Since 1998
PBS Pro	Opensourced by Altair	Part of Open HPC project
Sun Grid Enginer (SGE), Computing in Distributed Networked Environment (CODINE), Global Resource Director (GRD)	Oracle Grid Engine	Since 2000 under Sun. Predates AWS on a public cloud.
Simple Linux Utility for Resource Management (SLURM)	GNU Licensed	Most popular on top500
Platform load sharing facility (LSF)	IBM Spectrum LSF Suite	Origins in University of Toronto under Utopia project → OpenLava → Platform Computing
BProc + Maui	Beowulf Clustering Project	Since 1994 at NASA to help build clusters out of regular desktop machines.

Flynn's Taxonomy

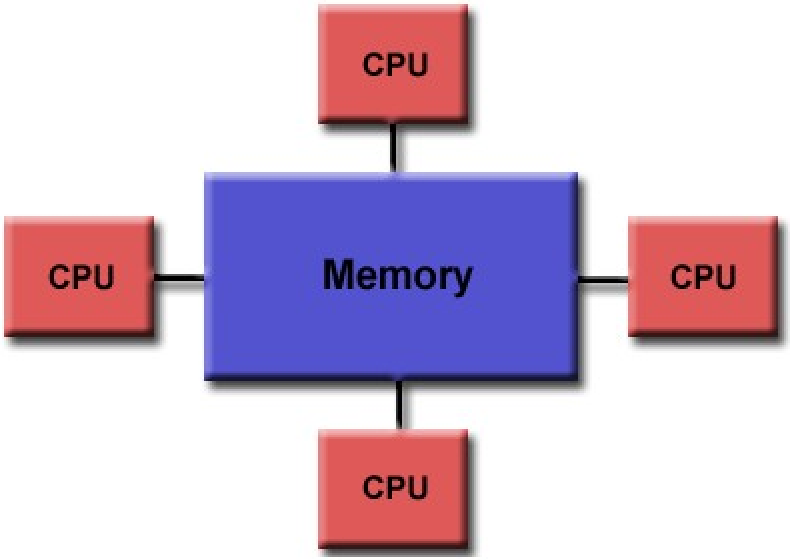


Some more Jargon:

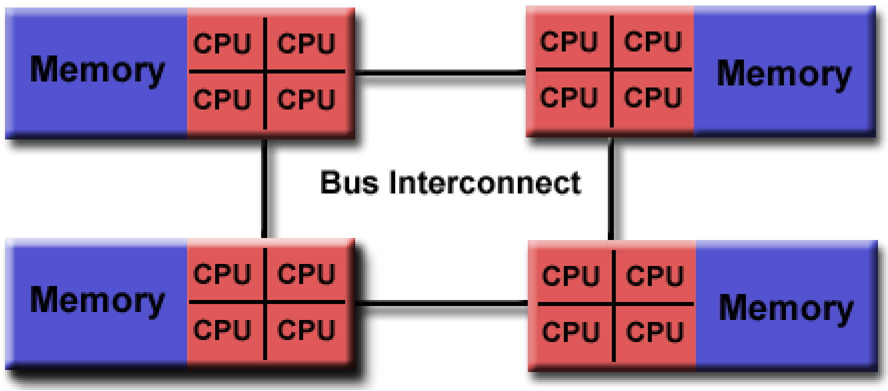
Node
CPU / Socket / Processor / Core
Task
Pipelining
Shared Memory
Distributed Memory
Symmetric Multiprocessor (SMP)
Communications
Synchronization

Latency
Bandwidth
Granularity
Observed speed-up
Parallel overhead
Massively Parallel (MP)
Embarrassingly Parallel (EP)
Scalability

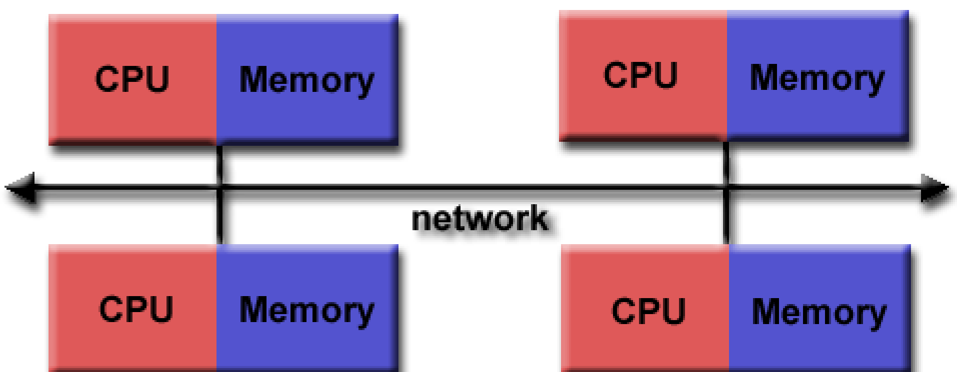
Memory Architectures



Shared Memory : Uniform Memory Access

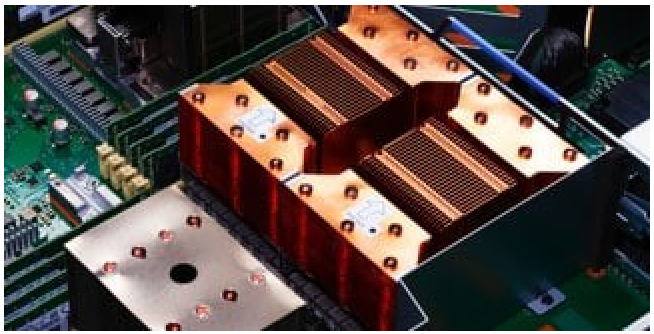


Shared Memory : Non-uniform Memory Access



Distributed Memory

Multiple sockets
Multiple CPUs on same motherboard



IBM Power9 Processor family with optimized silicon for a range of platforms

Scale out for HPC and next-gen apps

Scale up to 16 sockets to deliver the performance and capacity needed by the most-demanding enterprise workloads

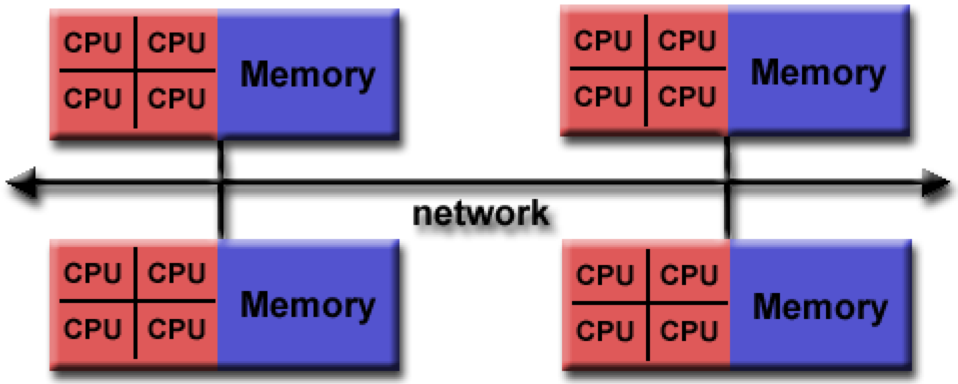
HPE Superdome Flex

A single system with 4-32 sockets and 1-48 TB of in-memory computing capacity into help you solve complex, data-intensive HPC problems at unparalleled scale.

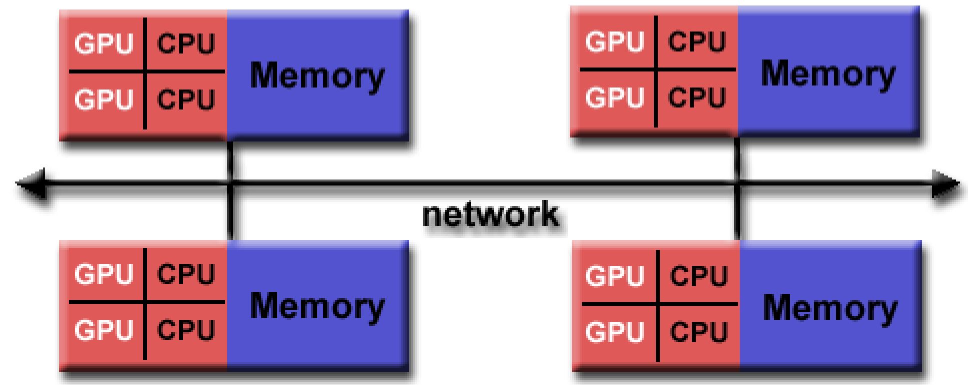
Dell PowerEdge R940xa Rack Server
4 socket

SMP systems allow for automatic parallelization by compiler.

Hybrid Architectures



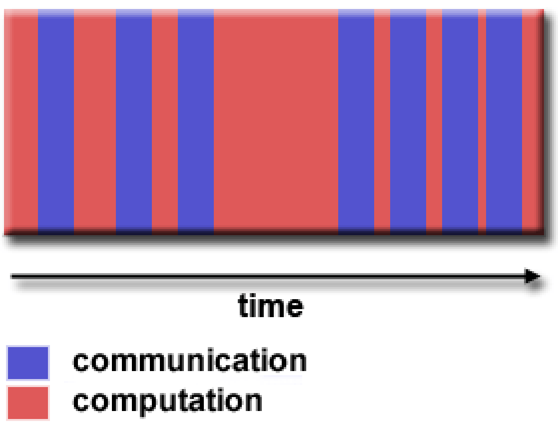
Both shared and distributed memory architectures : only CPUs



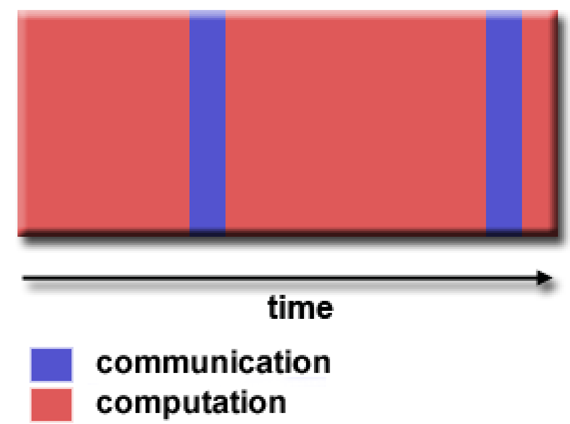
Both shared and distributed memory architectures : CPU + GPU

Granularity

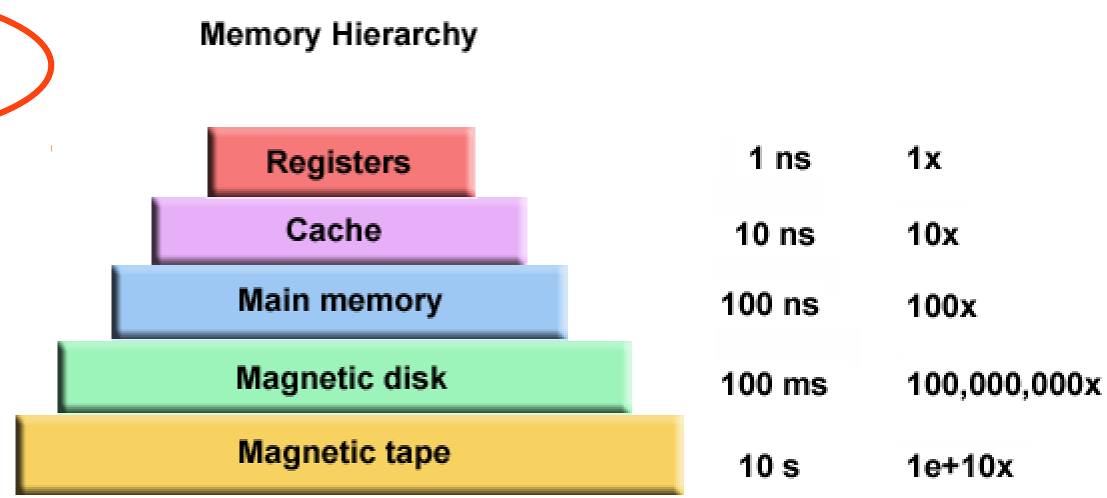
Fine grained parallelism



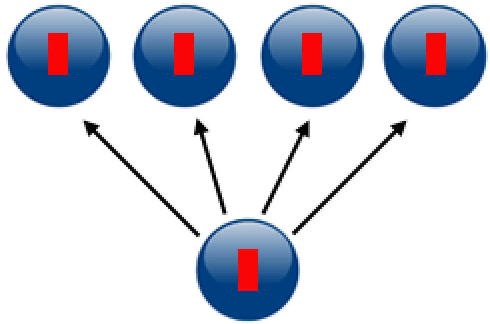
Coarse grained parallelism



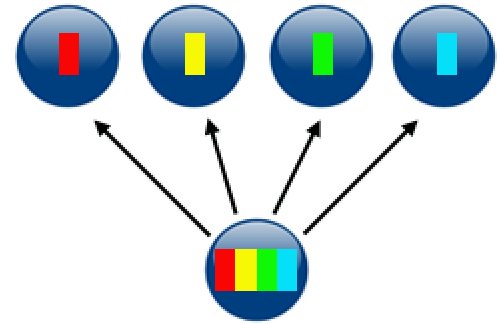
I/O could be expensive



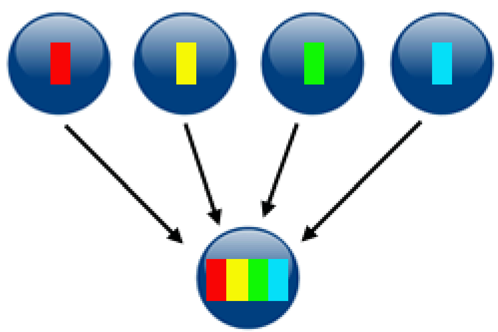
Scope of Communications



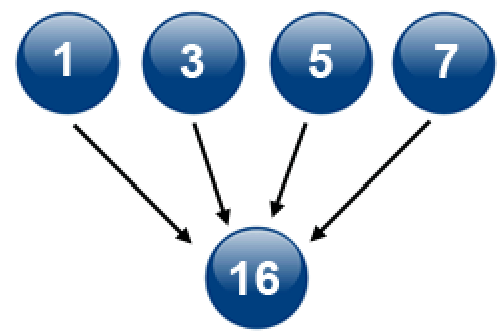
broadcast



scatter

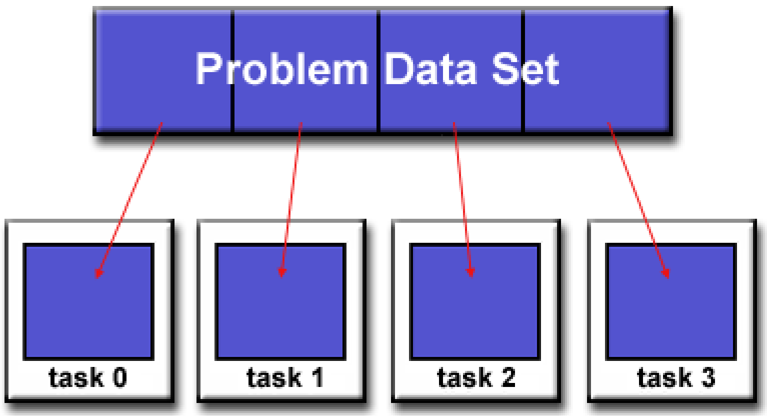


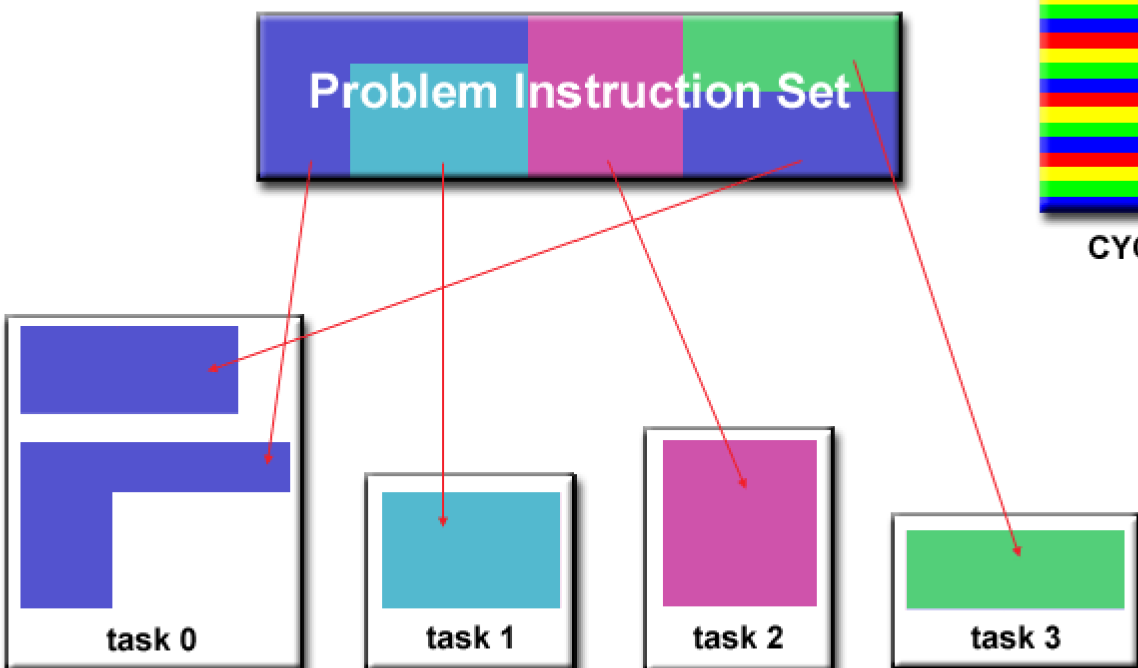
gather

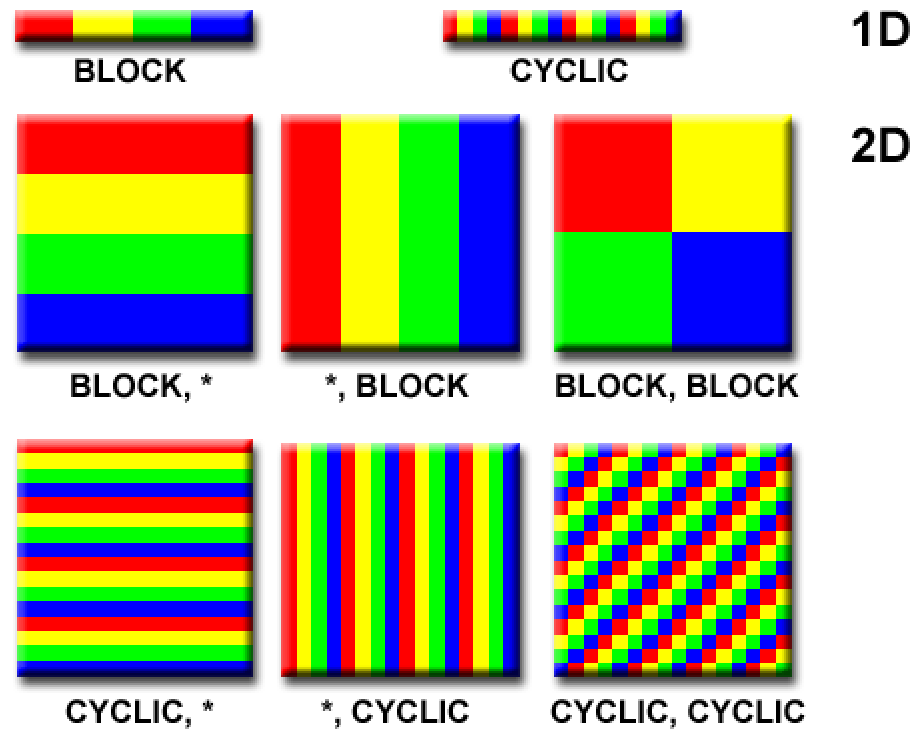


reduction

Domain Decomposition







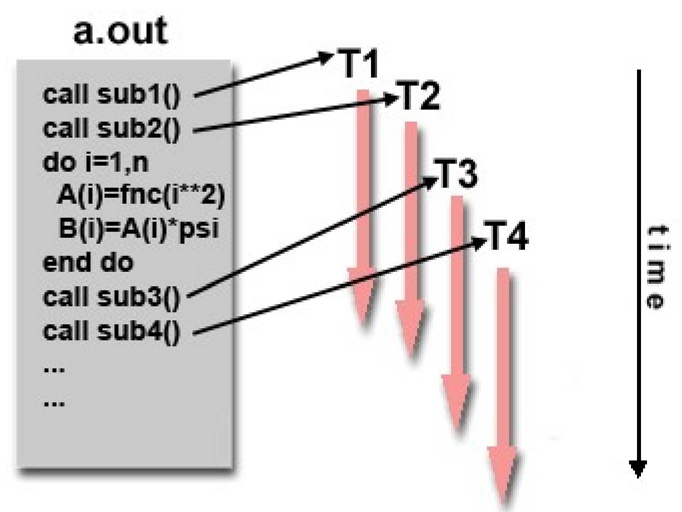
Parallel Processing : Threads Model

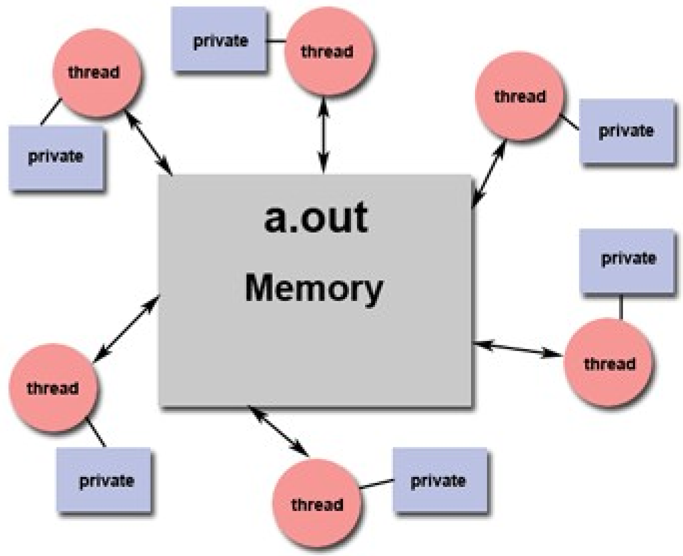
POSIX Threads

Specified by the IEEE POSIX 1003.1c standard (1995). C Language only.
Part of Unix/Linux operating systems
Library based
Commonly referred to as Pthreads.
Very explicit parallelism; requires significant programmer attention to detail.

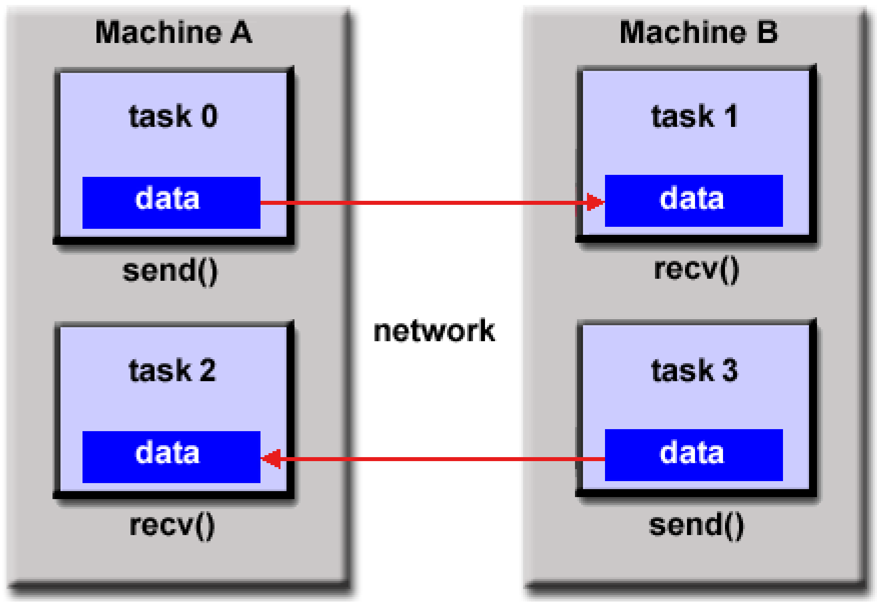
OpenMP

Industry standard, jointly defined and endorsed by a group of major computer hardware and software vendors, organizations and individuals.
Compiler directive based
Portable / multi-platform, including Unix and Windows platforms
Available in C/C++ and Fortran implementations
Can be very easy and simple to use - provides for "incremental parallelism". Can begin with serial code.





Parallel Processing : Message Passing Model



A set of tasks that use their own local memory during computation. Multiple tasks can reside on the same physical machine and/or across an arbitrary number of machines.

Tasks exchange data through communications by sending and receiving messages.

Data transfer usually requires cooperative operations to be performed by each process. For example, a send operation must have a matching receive operation.

MPI : Message Passing Interface

Profiling and Debugging

TotalView

RogueWave SOFTWARE

allinea DDT

arm Allinea Studio

Intel Inspector

Stack Trace Analysis Tool

LIVERMORE COMPUTING CENTER
HIGH PERFORMANCE COMPUTING

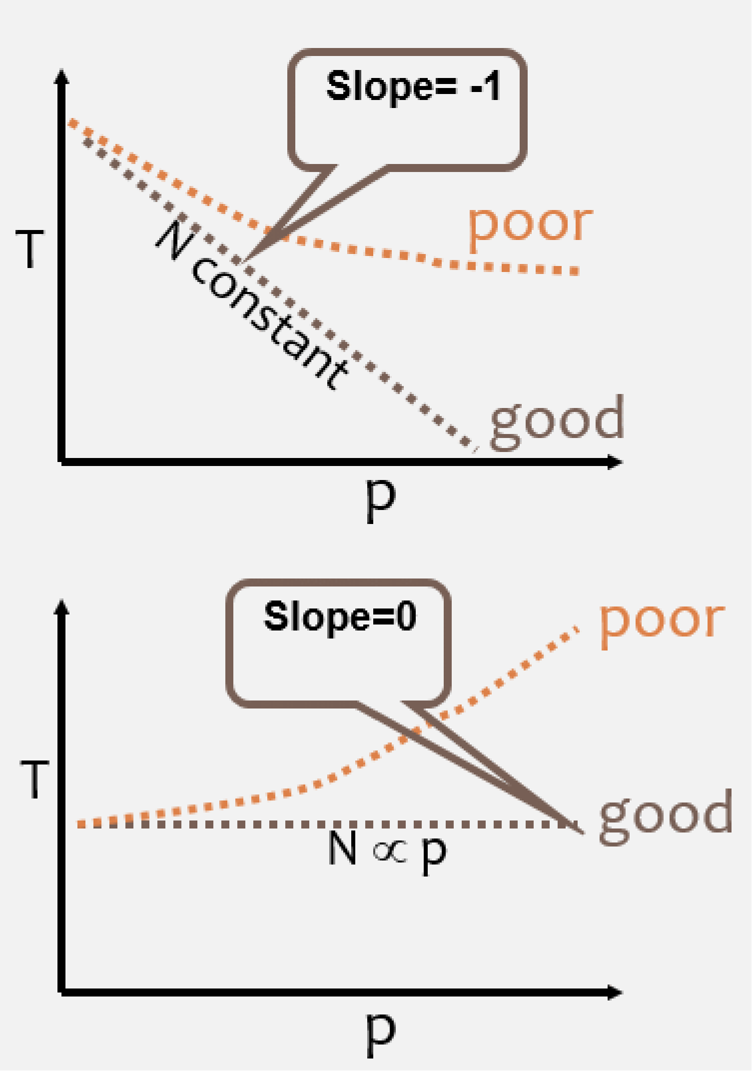
Scaling

- Strong scaling
- Weak scaling

The total problem size stays fixed as more processors are added.

Goal is to run the same problem size faster

Perfect scaling means problem is solved in 1/P time (compared to serial)



The problem size per processor stays fixed as more processors are added. The total problem size is proportional to the number of processors used.

Goal is to run larger problem in same amount of time

Perfect scaling means problem Px runs in same time as single processor run

Limits and Cost of Parallel Programming

- Amdahl's law

Speed up = 1/(1-F)
F is fraction of code parallelized

- Complexity : Design, Coding, Debugging, Tuning, Maintenance
- Portability : Vendor enhancements, non-standard APIs or libraries
- Resource Requirement : CPU time, Memory

Getting ready ...

- Profile code, identify hotspots
  (-pg option to compile & gprof command)
- Optimization of code / algorithm
- Choose a parallel paradigm
  MPI / OpenMP / OpenACC / OpenCL / CUDA / ...
- Code, code properly, code efficiently ...
- Validate with serial version, bitwise
- Benchmark speedup
- Stop after reaching scalability limit

GNR Cluster for UG students

- 1 Head Node on Super micro servers with Dual Processors, Eight-Core Intel Xeon Ivy bridge E5-2650v2 series processors with 4 X 8GB RAM and 500 GB of SATA Hard disk.
- 16 compute nodes based on super micro server with Dual processor, Eight-core Intel Xeon Ivy Bridge E5-2650v2 series Processors with 4 X 8 GB RAM and 500 GB of SATA Hard disk in each node.
- 14TB of shared storage

gnr.iitm.ac.in

- IP address: `10.200.6.5`
- Username : `mm2090`
- Keep codes in `$HOME/work` pointing to */work/oth/mm2090*
- Compile using
  /Apps/intel-compilers/impi/bin64/mpicc
- Run using
  /Apps/intel-compilers/impi/bin64/mpirun
- Use qstat and qsub commands for job status and job submission.