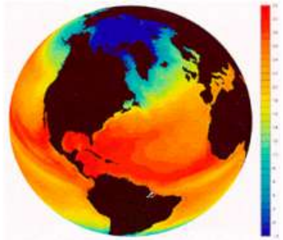
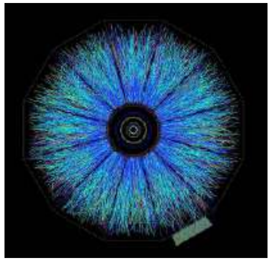
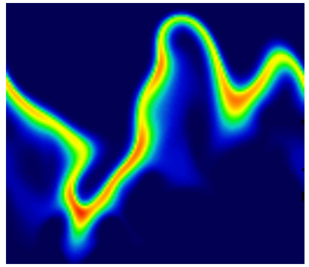
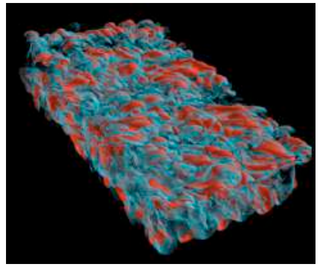
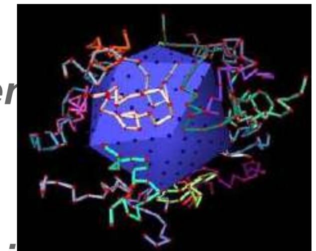
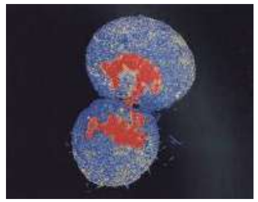
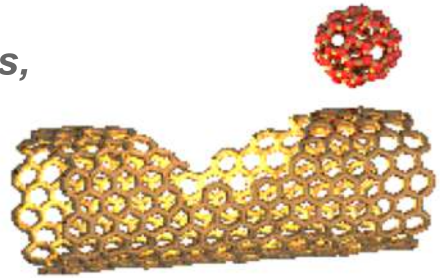
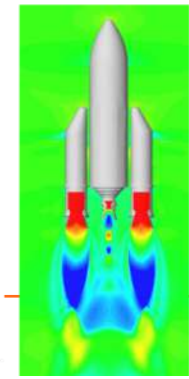
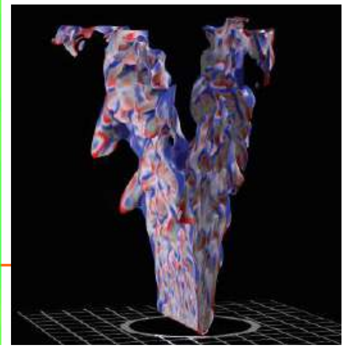
Introduction to HPC

EUDAT – PRACE Summer School on managing scientific data from analysis to long term archiving, 23-27 September 2019, Trieste, Italy

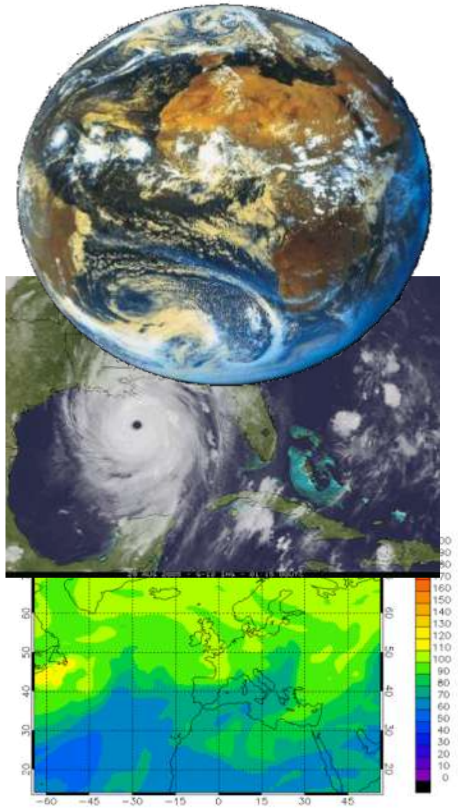
Leon Kos, University of Ljubljana

Why supercomputing?

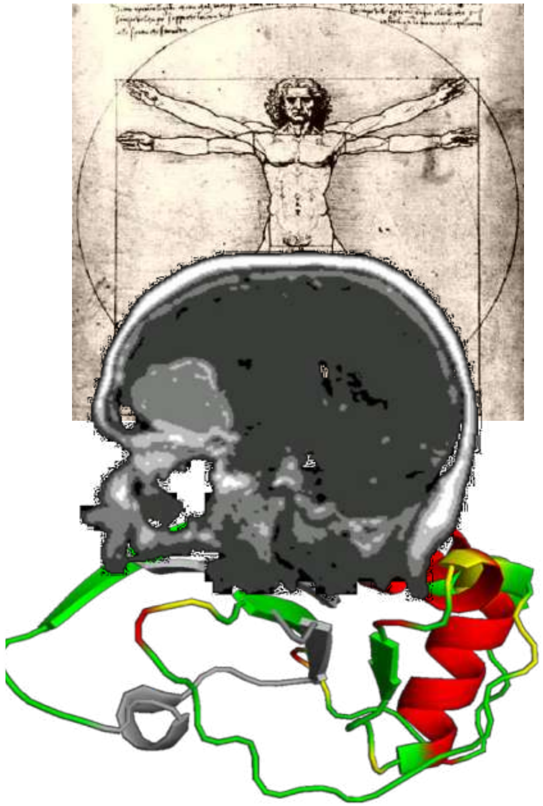
- Weather, Climatology, Earth Science
  - degree of warming, scenarios for our future climate.
  - understand and predict ocean properties and variations
  - weather and flood events
- ***Astrophysics, Elementary particle physics, Plasma physics***
  - ***systems, structures which span a large range of different length and time scales***
  - ***quantum field theories like QCD, ITER***
- ***Material Science, Chemistry, Nanoscience***
  - ***understanding complex materials, complex chemistry, nanoscien***
  - ***the determination of electronic and transport properties***
- ***Life Science***
  - ***system biology, chromatin dynamics, large scale protein dynamics, protein association and aggregation, supramolecular systems, medicine***
- ***Engineering***
  - ***complex helicopter simulation, biomedical flows, gas turbines and internal combustion engines, forest fires, green aircraft,***
  - ***virtual power plant***

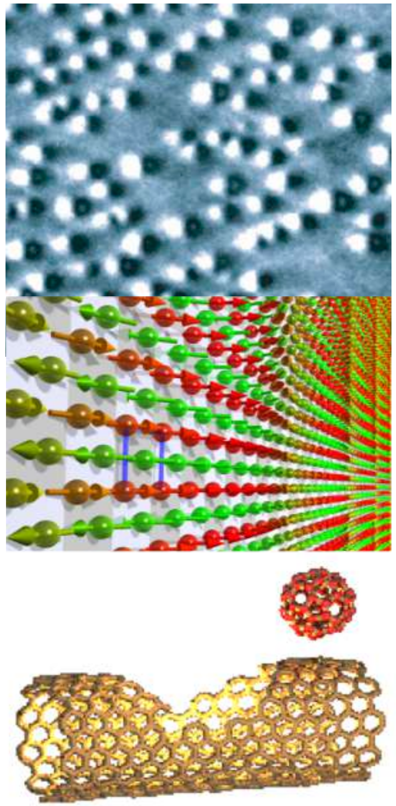
Supercomputing drives science with simulations



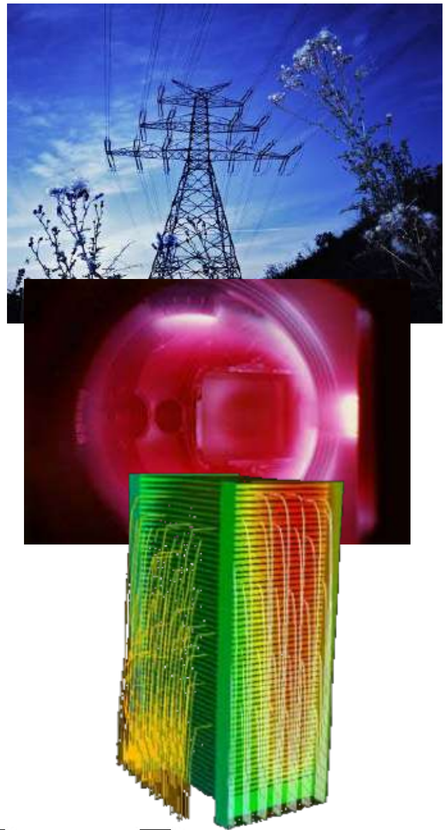
Environment
Weather/ Climatology
Pollution / Ozone Hole



Ageing Society
Medicine
Biology



Materials/ Inf. Tech
Spintronics
Nano-science



Energy
Plasma Physics
Fuel Cells

TOP 500

https://www.top500.org/lists/2019/06

- TOP 10 Sites for June 2019
- All 500 are peta-Flop/s systems

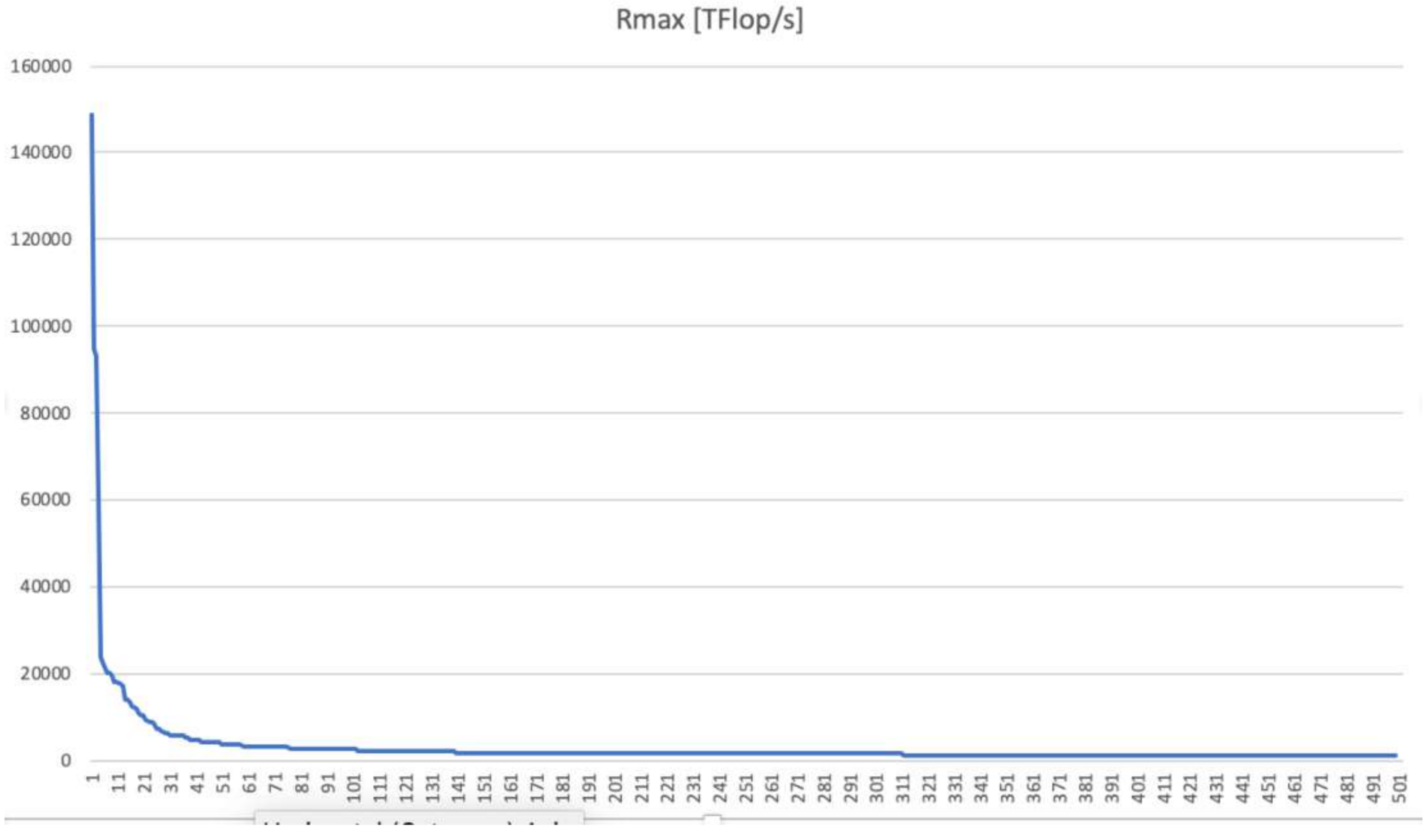
- GREEN 500
  - #469 on TOP500 is DGX SaturnV Volta system NVIDIA system installed at NVIDIA and FIRST on Green 500!

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	2,414,592	148,600.0	200,794.9	10,096
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	**Frontera** - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR , Dell EMC Texas Advanced Computing Center/Univ. of Texas United States	448,448	23,516.4	38,745.9	
6	**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
7	**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	20,158.7	41,461.2	7,578
8	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649
9	**SuperMUC-NG** - ThinkSystem SD650, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path , Lenovo Leibniz Rechenzentrum Germany	305,856	19,476.6	26,873.9	
10	**Lassen** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100 , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	288,288	18,200.0	23,047.2	

High Performance Linpack – June 2019

- Mainstream 1-5 PFlop/s
- Knee 5-20 PFlop/s
- Leaders 20-150 PFlop/s

- Exascale or ExaFlop/s

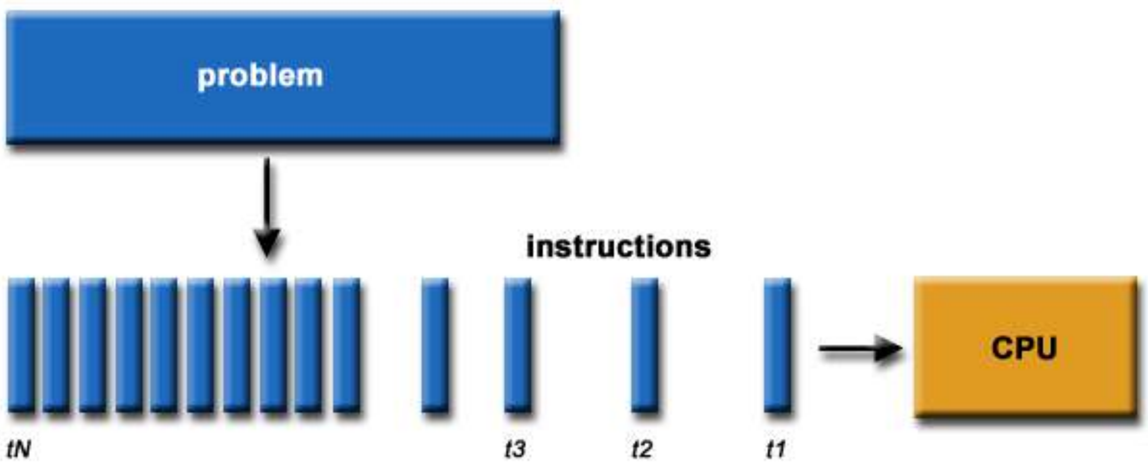


Towards Exa Scale

- GPU Accelerators everywhere?
- Summit, using only 4,608 nodes and 10PB RAM
  - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband
- Sunway TaihuLight, 40,960 nodes and 1PB RAM
  - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway

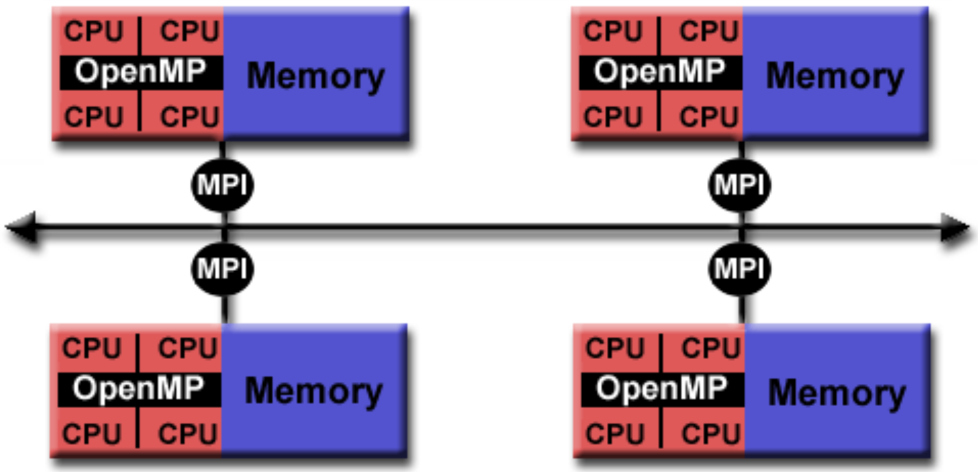
Introduction to parallel computing

- Usually is the program written for serial execution on one processor
- We divide the problem into series of commands that can be executed in paralllel
- Only one command at a time can be executed on one CPU



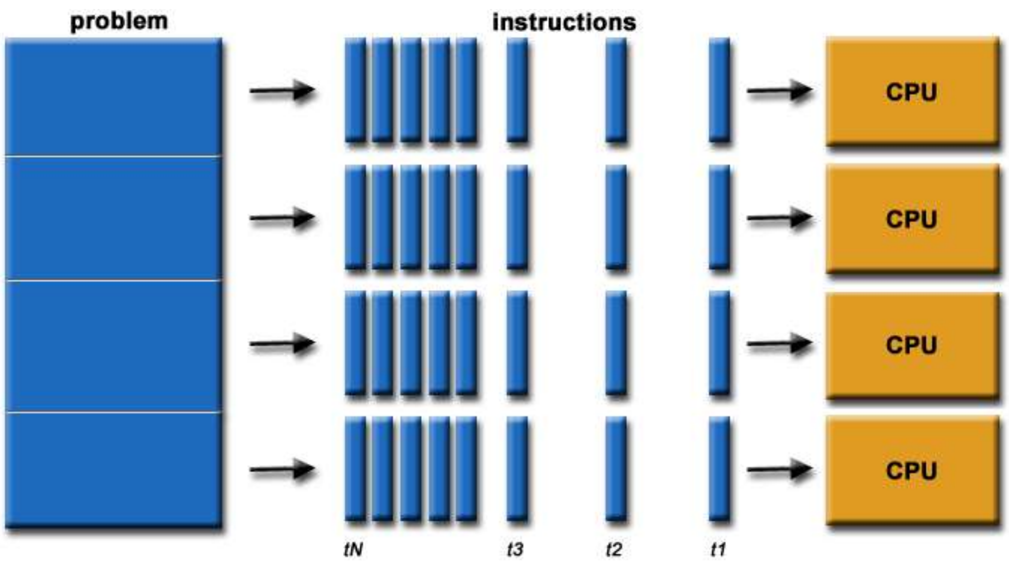
Parallel programming models

- Threading
- ***OpenMP* – *automatic parallelization***
- Distributed memory model = ***Message Passing Interface (MPI)* – *manual parallelization needed***
- ***Hybrid model OpenMP/MPI***
- ***Accelerators (GPU)***
- ***Heterogeneous computing***



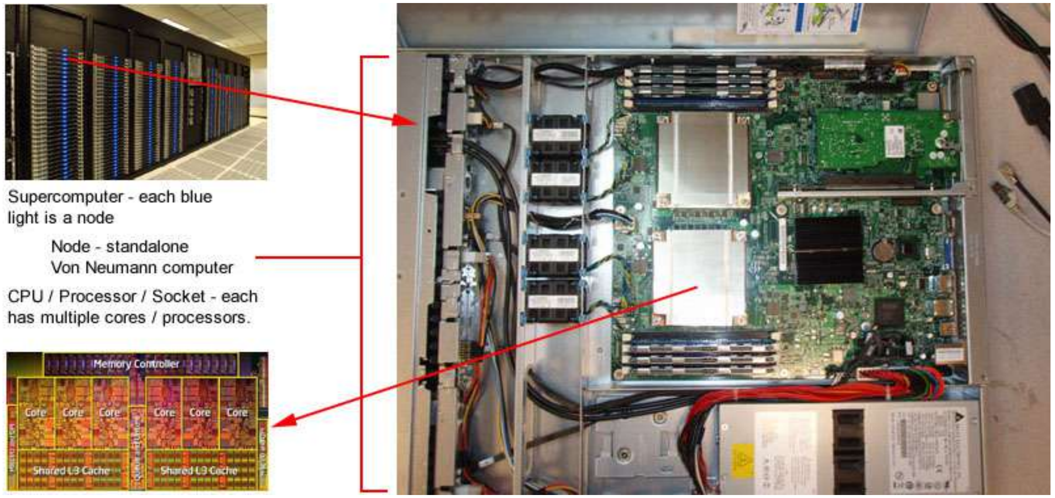
Embarrasingly simple parallel processing

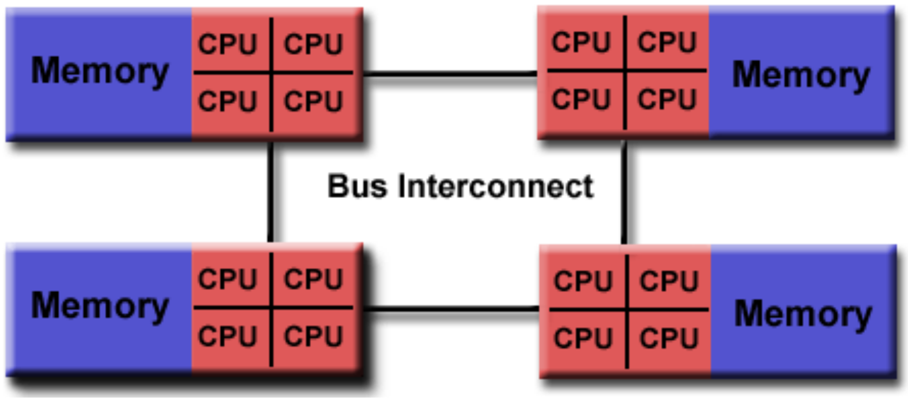
- Parallel processing of the same subproblems on multiple prooocessors
- No communication is needed between processes



Logical view of a computing node

- Need to know computer architecture
- ***Interconnect bus for sharing memory between processors (NUMA interconnect)***

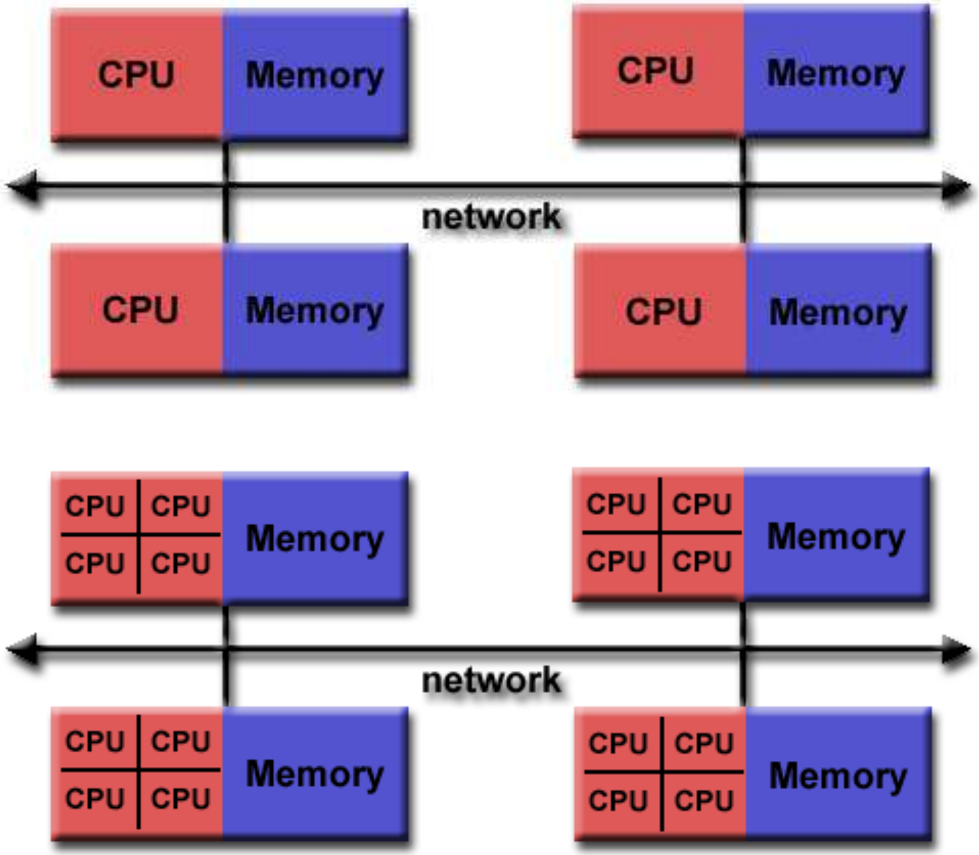




Nodes interconnect

- Distributed computing
- Many nodes exchange messages on
  - high speed,
  - low latency interconnect such as

    Infiniband



Development of parallel codes

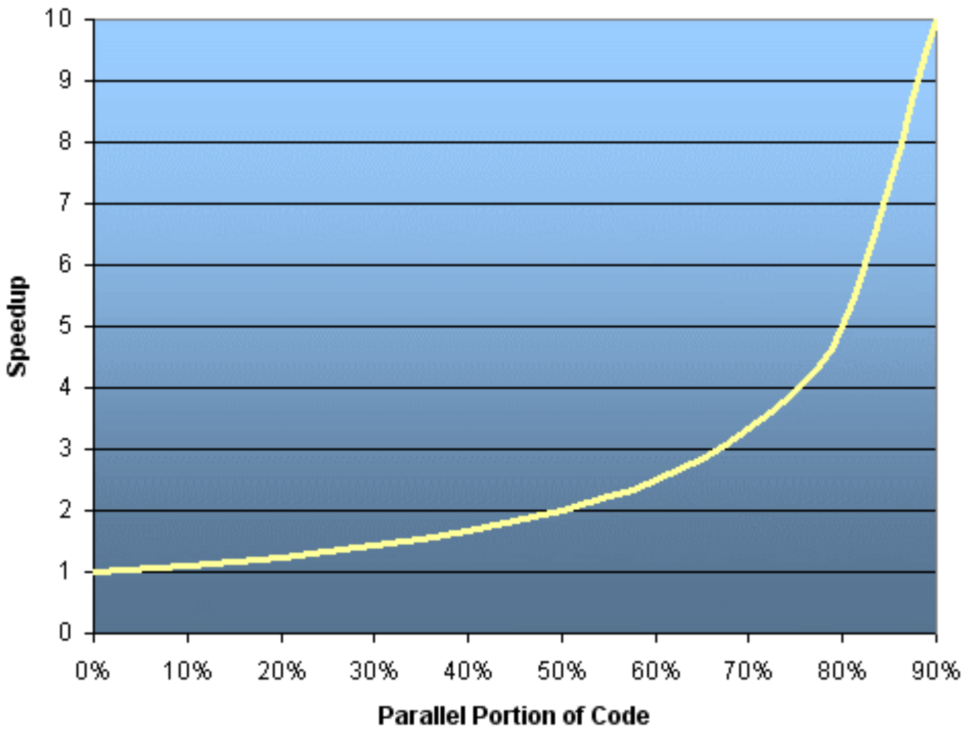
- Good understanding of the problem being solved in parallel
- How much of the problem can be run in parallel
- Bottleneck analysys and profiling gives good picture on scalability of the problem
- We optimize and parallelize parts that consume most of the computing time
- Problem needs to be disected into parts functionally and logically

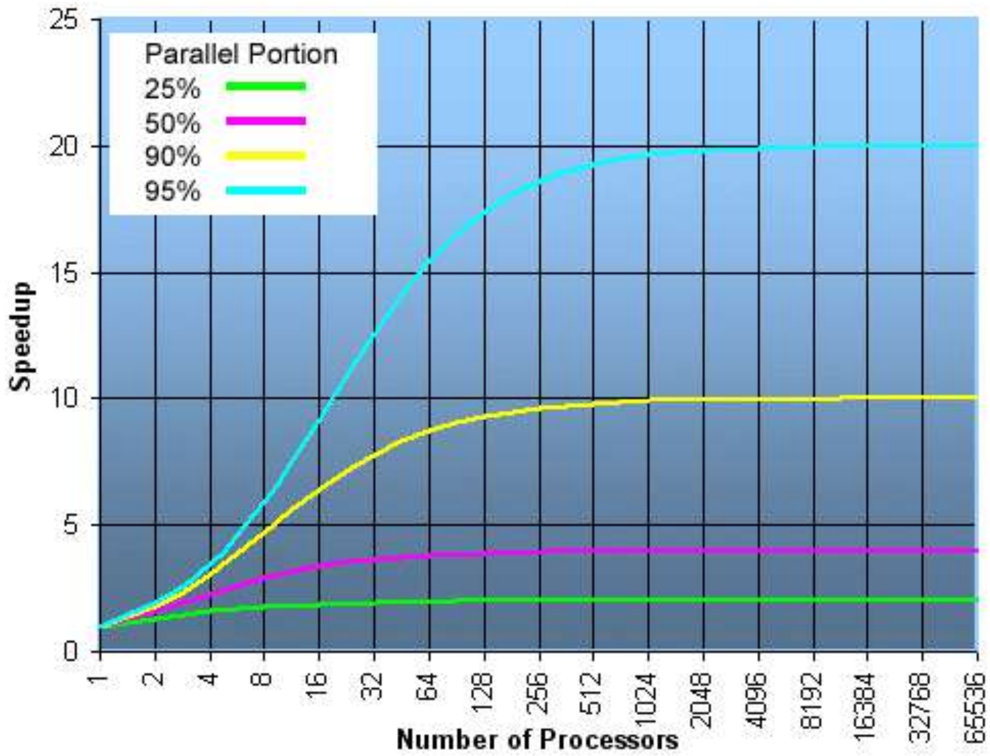
Interprocess communications

- Having little an infrequent communication between processes is the best
- Determining the largest block of code that can run in parallel and still provides scalability
- Basic properties
  - ***response time***
  - ***transfer speed - bandwidth***
  - ***interconnect capabilities***

Parallel portion of the code determines code scalability

- Amdahlov law ***Speedup = 1/(1-p)***
  - ***1% of serial code gives max speedup of 100***





Questions and practicals on the GALILEO cluster

- Demonstration of the work on the cluster by repeating
- Learning basic Linux commands
- SLURM scheduler commands
- Modules
- Development with OpenMP and OpenMPI parallel paradigms
- Excercises and extensions of basic ideas
- Instructions available at

http://www.prace-ri.eu

THANK YOU FOR YOUR ATTENTION

www.prace-ri.eu