# Lecture 1: introduction to HPC (part1 and 2)

- Stefano Cozzini
- CNR-IOM and eXact lab srl

# Agenda

- Prologue: why and where HPC ?
  - What is HPC ?
  - Definitions & metrics
- What is HPC infrastructure ?
  - Parallel machines
  - Supercomputers & HPC Cluster
- Measure Performance
- HPC on the Cloud ?

# HPC: the challenge..

- Societal, scientific and economic needs are the drivers for the next generation of HPC - computing with <span style="color:red">exascale performance</span> (computers capable of performing 10 to the power of 18 floating point operations per second).

- All scientific disciplines are becoming "computational" today. Modern scientific discovery requires very high computing power and capability to deal with huge volumes of data.

- Industry and SMEs are increasingly relying on the power of supercomputers to invent innovative solutions, reduce cost and decrease time to market for products and services.

- HPC is part of a global race. Many countries (USA, Japan, Russia, China, Brazil, India) have announced ambitious plans for building the next generation of HPC with exascale performance and deploying state-of-the-art supercomputers.

From https://ec.europa.eu/programmes/horizon2020/en/h2020-section/high-performance-computing-hpc

# A few example where to apply HPC

HPC capabilities are used to solve and address scientific, industrial and societal challenges.

## Health

- Development of personalised and precision medicine to provide individual and accurate patient treatment.
- Saving time and money on the development of new drugs from the initial idea phase to the final phase of reaching the market.

| development of new drugs reduced from 17 years | HPC can help save hundreds of millions |
|---|---|
| to several months | up to billions of Euros |

- Early detection of diseases and quicker diagnosis.

## Climate change & weather forecast

- Europe paid severe weather damage costs between 1970 and 2012[1].

  150.000 lives    €270 billion in economic damages

- With HPC technology, climate scientists will be able to predict the size and paths of storms and floods with accuracy.
- This will allow implementation of measures such as alerting or evacuating people, thus saving human lives.

## Industry

- Reducing development time, minimising costs, optimising decision processes and producing higher quality goods and services.
- E.g. automotive industry will save time and money to develop new vehicle platforms, with improved environmental friendliness and passenger comfort and safety[2].

| development reduced from 60 months | HPC can help save |
|---|---|
| to 24 months | up to € 40 billions |

## Cybersecurity

- Enabling complex encryption technologies and better reactions to cyberattacks.
- Combined with Artificial Intelligence, HPC detects:
  - strange systems behaviour,
  - insider threats and electronic fraud,
  - very early cyberattack patterns (few hours instead of a few days)
  This will allow automated and immediate actions even before a cyberattack happens.

## Energy

- HPC provides critical tools for example in:
  - designing renewable energy parks
  - designing high-performance photovoltaic materials,
  - optimising turbines for electricity production.
- HPC expenditure in the energy sector is projected to grow by 5% in the next years.

# Some more examples

- See chapter one section 1.2 of reference 1

# A slogan...

Out compute = Out compete

Image from UberCloud

# HPC: A first definition

- Defining HPC (from Intersect survey)

  - High Performance Computing (HPC) is the <span style="color:red">use of servers, clusters, and supercomputers</span> – plus <span style="color:blue">associated software, tools, components, storage, and services</span> – for <span style="color:green">scientific, engineering, or analytical tasks</span> that are particularly intensive in computation, memory usage, or data management
  - HPC is used by scientists and engineers both in research and in production across <span style="color:red">industry,</span> <span style="color:blue">government</span> and <span style="color:green">academia.</span>

    [to be continued]

# Elements of the HPC...

- use of servers, clusters, and supercomputers
  - → HARDWARE
- associated software, tools, components, storage, and services
  - → SOFTWARE
- scientific, engineering, or analytical tasks
  - → PROBLEMS TO BE SOLVED..

ALL THE ABOVE DEFINES A COMPUTATIONAL INFRASTRUCTURE
aka E-INFRASTRUCTURE

# Elements of an HPC infrastructure

- High end computational  servers + <mark>accelerators</mark>
- High speed networks
- High end parallel storage
- Middleware
- Scientific/Technical/ Data analysis  software
- Research/Technical data
- Problems to be solved

IS ALL WHAT WE NEED ?

# Last but not least: people

- Human capital is by far the most important aspect
- Two important roles:
  - HPC providers
    - (plan/install/manage HPC resources)
  - HPC user :
    - use at best HPC resource

> MIXING/INTERPLAYING ROLES
> INCREASES COMPETENCE LEVELS

# From infrastructure to ecosystem

- Goal:
  - provide a computational ecosystem to satisfy all the different requirements posed by <span style="color:red">users</span>
- Which kind of requirements ?
  - All you need to to solve their computational problems !

# Which kind of users ?

- Computational scientists
- Industries with computational tasks


- Scientists with a lot of (different) data
- Industries with a lot of (different) data

# Challenges ahead HPC (I)

- HPC skilled people

# Research is changing...

- Inference Spiral of System Science
  - As models become more complex and new data bring in more information, we require ever increasing computational power
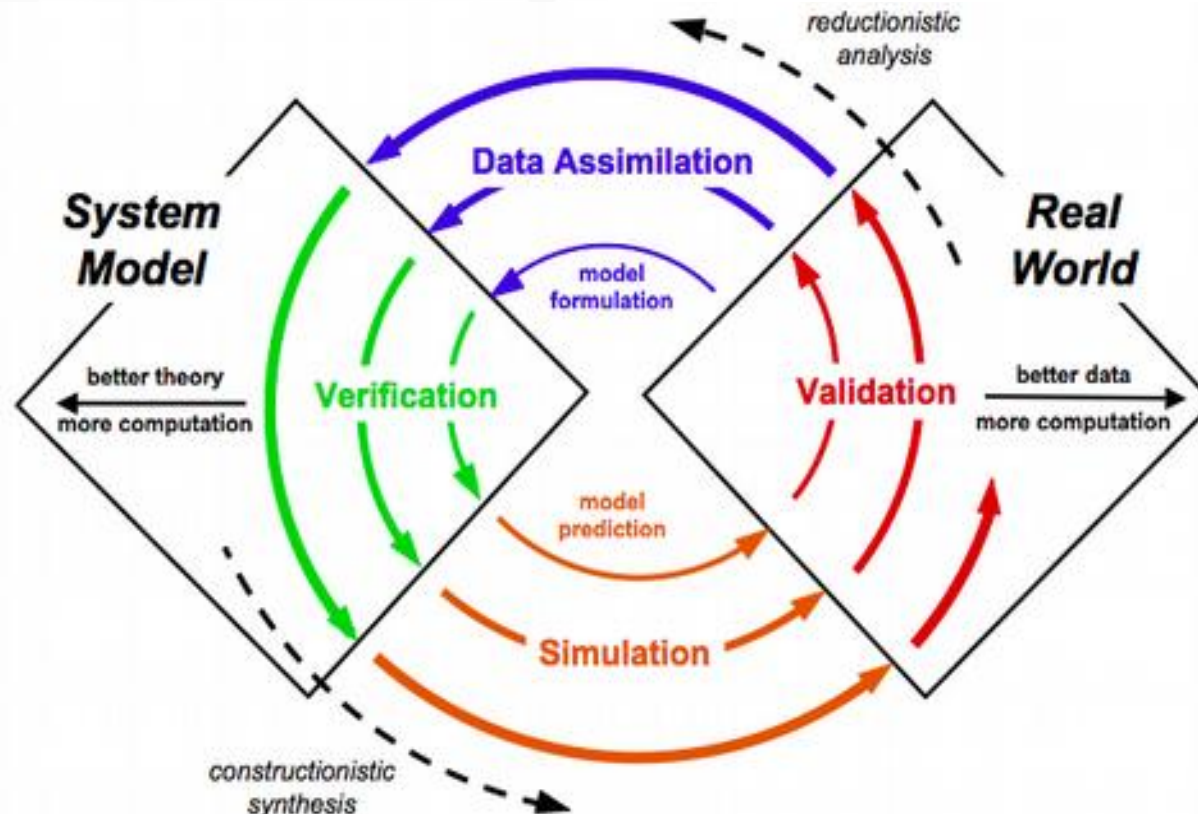
# Applications requires more data everyday..

- Simulation has become the way to research and develop new scientific and engineering solutions.

  - Used nowadays in leading science domains like aerospace industry, astrophysics, etc.

- Challenges related to the complexity, scalability and data production of the simulators arise.

  - Impact on the relaying IT infrastructure.
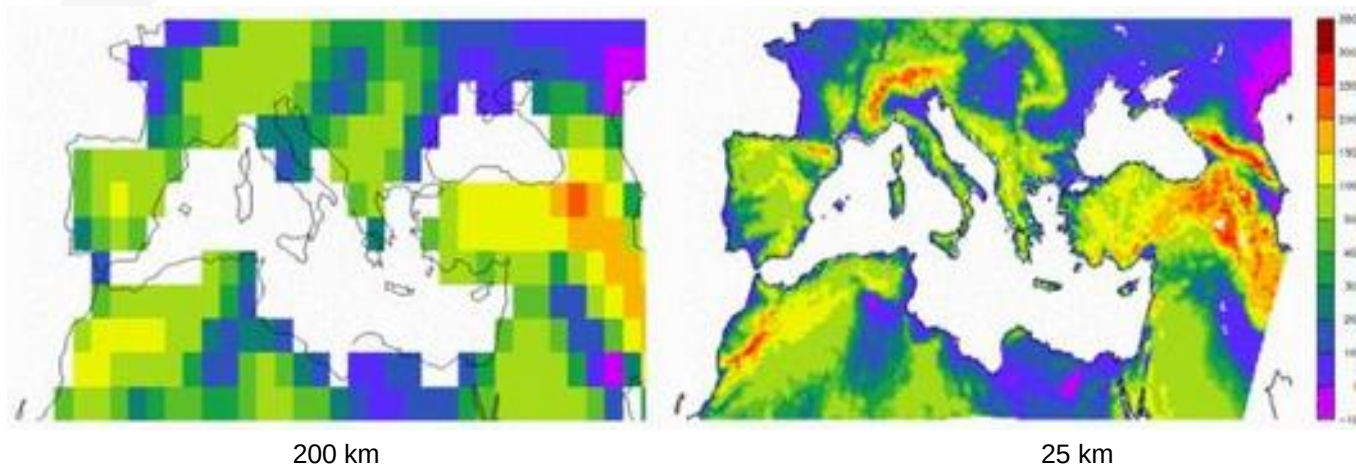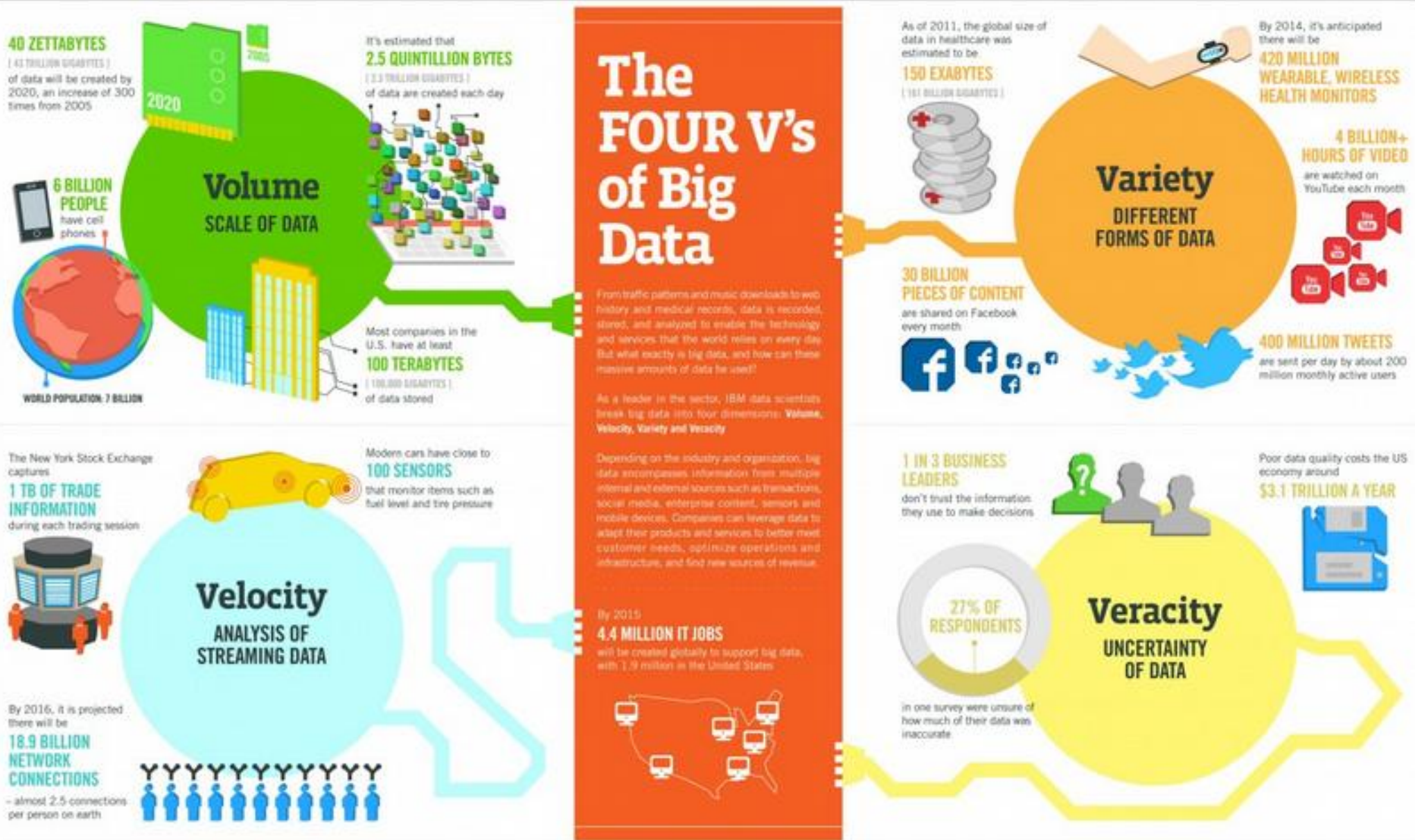


200 km                                          25 km

# BIG DATA buzzword..

# Data Intensive Science

- A fourth paradigm after experiment, theory and computation

# Big data challenge: from HPC to HPDA through AI

- Organizations are expanding their definitions of high-performance computing (HPC) to include workloads such as artificial intelligence (AI) and high-performance data analytics (HPDA) in addition to traditional HPC simulation and modeling workloads.

  From https://insidebigdata.com/2019/07/22/converged-hpc-clusters/

# Different software ecosystems for HDA and traditional computational science.
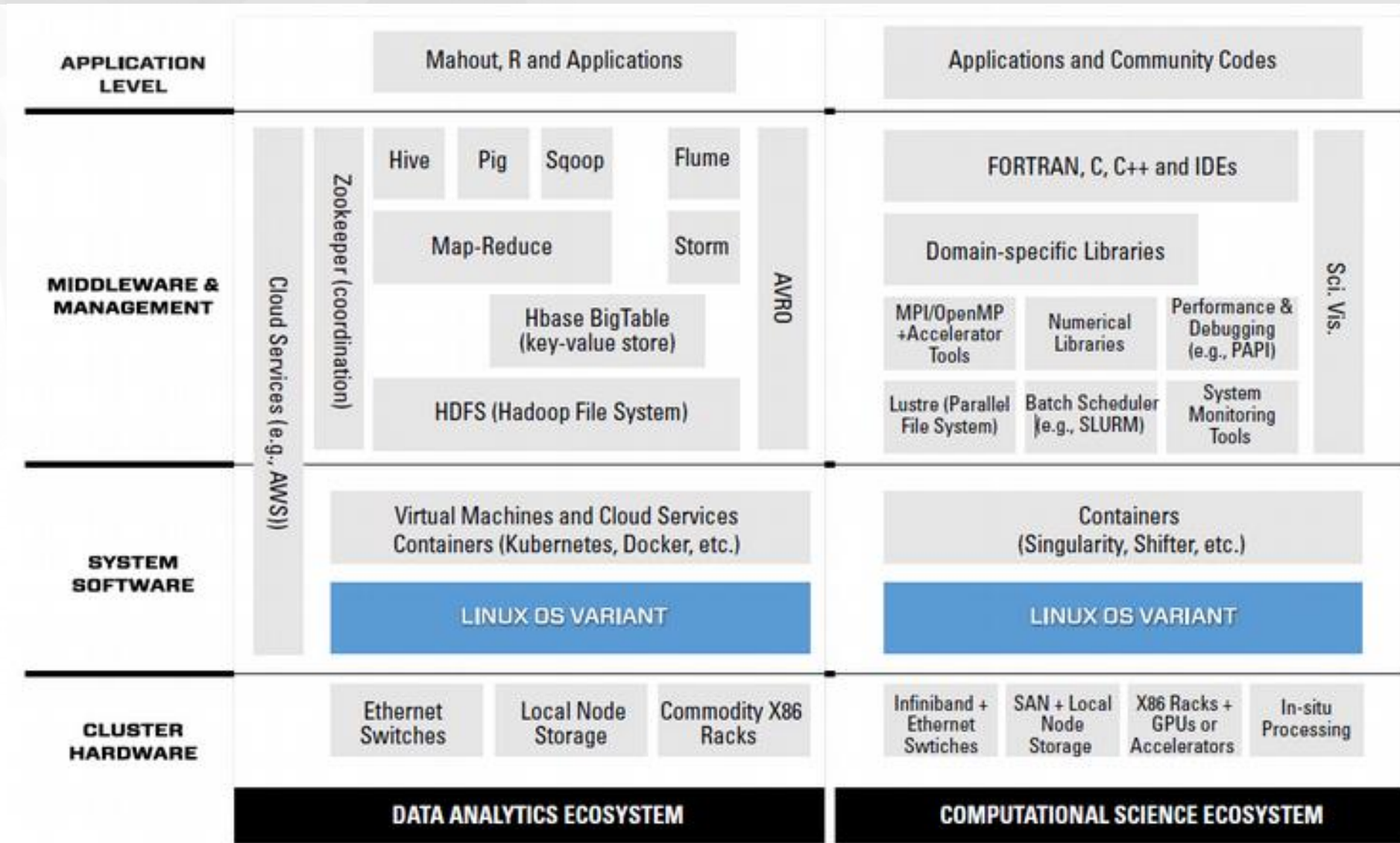


From https://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec_pathways.pdf

# Challenges ahead HPC (II)

- HPC skilled people
- BIG data: no longer HPC but HPDA/AI as well

# Is it all about Performance ?

- It is difficult to define Performance properly
- "speed" / "how fast" are vague terms
- Performance as a measure again ambiguous and not clearly defined and in its interpretation
- In any case performance it is at core to HPC as a discipline
- Let discuss it in some details

# Does  P stand just for Performance ?

- Performance is not always what matters..

    to reflect a greater focus on the productivity, rather than just the performance, of large-scale computing systems, many believe that HPC should now stand for High Productivity Computing.

    [ from wikipedia]

- P  should also stand for PROFITABILITY

# Performance vs Productivity

- A  possible definition:

  Productivity = (application performance) / (application programming effort)

- people in HPC arena have  different goals in mind  thus different expectations  and different  definitions of productivity.

  Question:  Which kind of productivity are you interested in ?

# Another HPC definition

- HPC incorporates all facets of three disciplines:
  - Technology
  - Methodology
  - Application

  The main defining property and value provided by HPC is delivering performance for end-user application
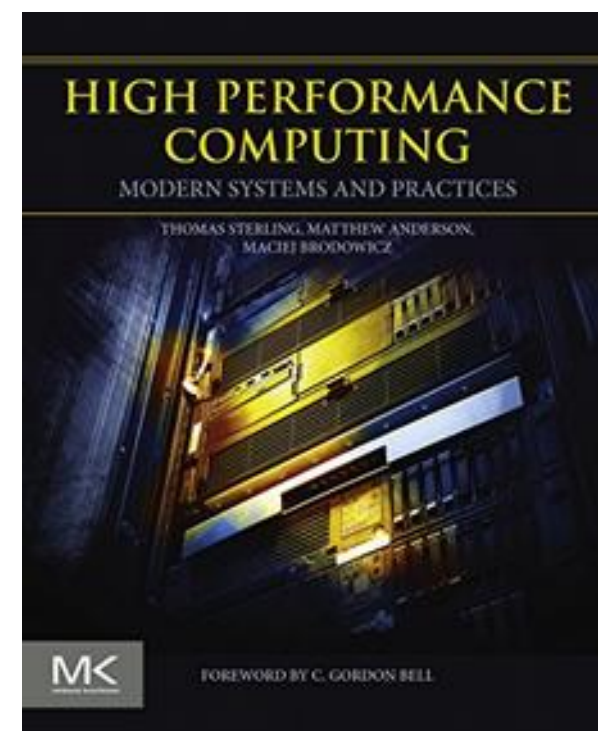


At page 39

# Let us focus on an high performance problem..

**A PARALLEL SOLUTION!**

picture from http://www.f1nutter.co.uk/tech/pitstop.php

# Analysis of the parallel solution

**FUNCTIONAL PARTITIONING**
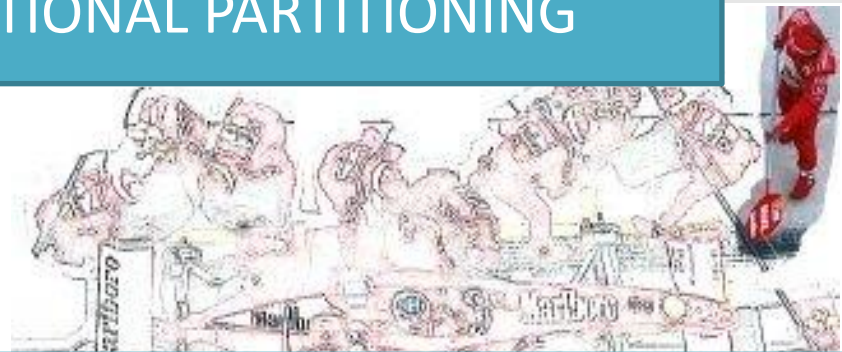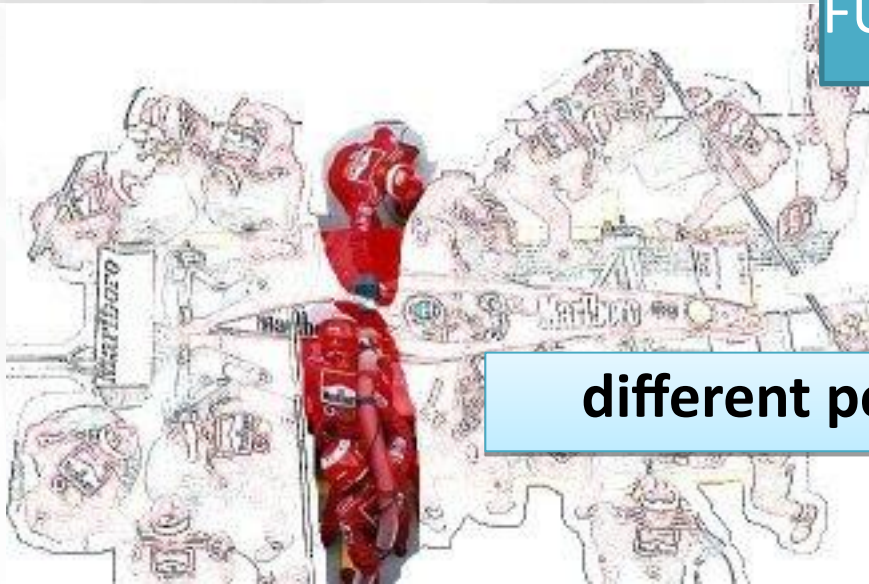
**different people are executing different tasks**

**DOMAIN DECOMPOSITION**

**different people are solving the same global task but on smaller subset**

# HPC

# =

# Parallel computing

# HPC

# =

# Parallel computers

# Parallel computers

- Tons of different machines !
- Flynn Taxonomy (1966): helps us in classifying them:
  - Data Stream
  - Instruction Stream

|  | Instruction stream | |
| --- | --- | --- |
|  | Single | Multiple |
| Data stream — Single | SISD | MISD |
| Data stream — Multiple | SIMD | MIMD |

# Flynn Taxonomy (graphical view)



Instruction

Data → **CPU** →

**SISD**
**(Single instruction stream**
**single data stream)**

Instruction

Data → **CPU** →
→ **CPU** →
→ **CPU** →

**SIMD**
**(Single instruction stream**
**multipe data stream)**

Instruction

Data → **CPU** →
→
→

**MIMD**
**(Multipe instruction stream**
**mingle data stream)**
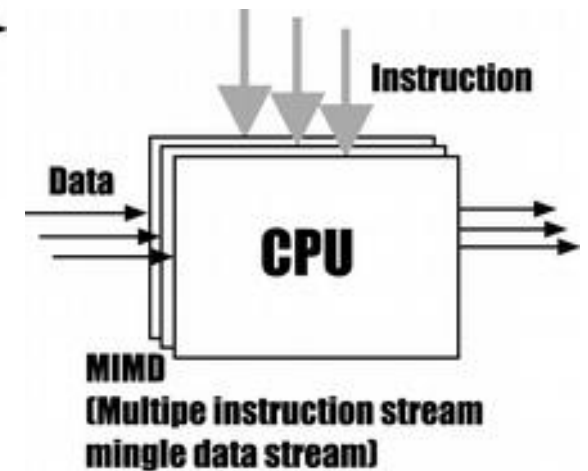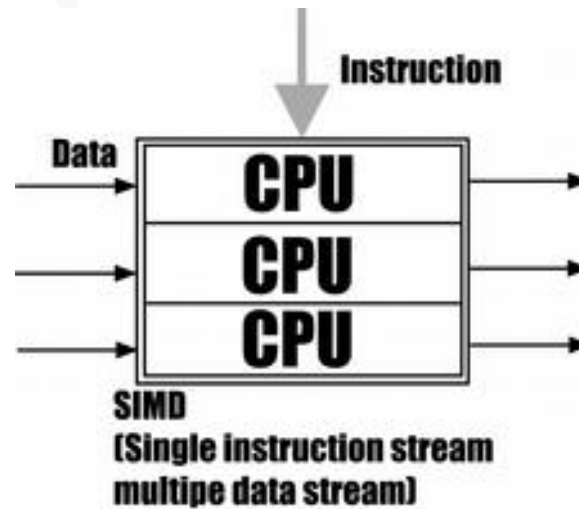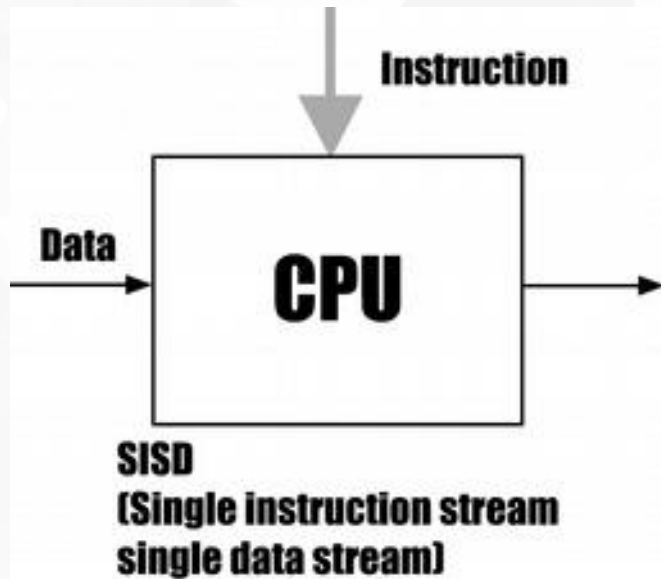
# Another important question:

- MEMORY: The simplest and most useful way to classify modern parallel computers is by their memory model:
  - SHARED MEMORY
  - DISTRIBUTED MEMORY
  - MIXED SITUATION

# Shared vs Distributed memory

- Distributed memory
  - each processor has its own local memory. Must do message passing to exchange data between processors
- Shared Memory
  - single address space. All processors have access to a pool of shared memory.
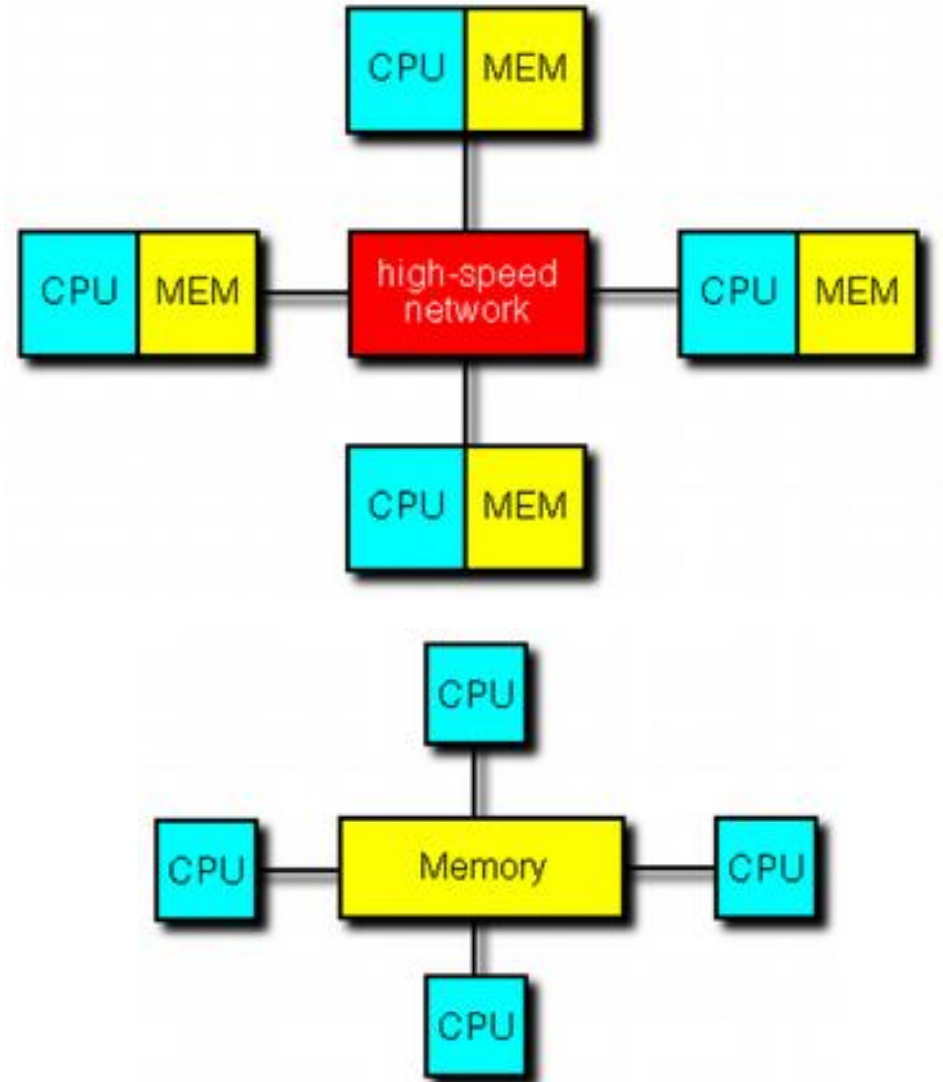
# Shared memory: UMA vs NUMA

- *Uniform memory access (UMA):* Each processor has uniform access to memory. Also known as symmetric multiprocessors (SMP)

- *Non-uniform memory access (NUMA):* Time for memory access depends on location of data. Local access is faster than non-local access.
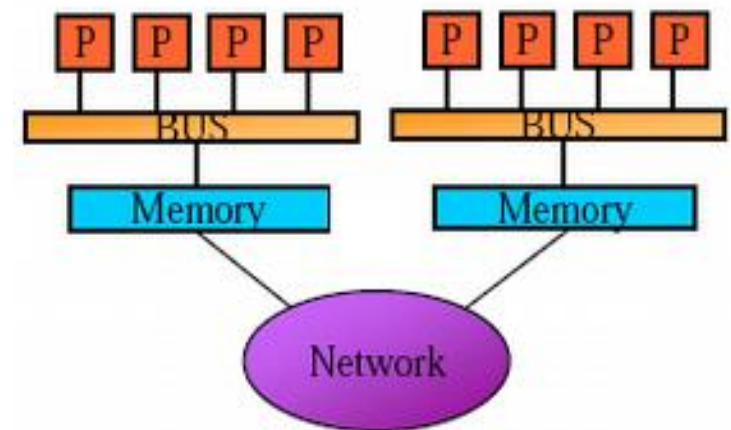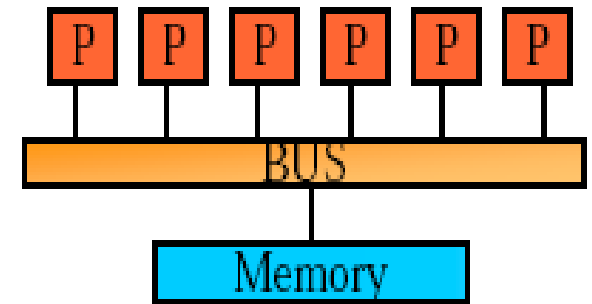
# Distributed memory machines

- The memory is physically distributed among the processors (local memory). Each processor can access directly only to its own local memory

  NO-Remote Memory Access (NORMA) model

- Communication among different processors occurs via a specific communication protocol (message passing).

# The basic distributed memory machine: clusters



User / Scheduler                    Compute Cluster                    Data Storage
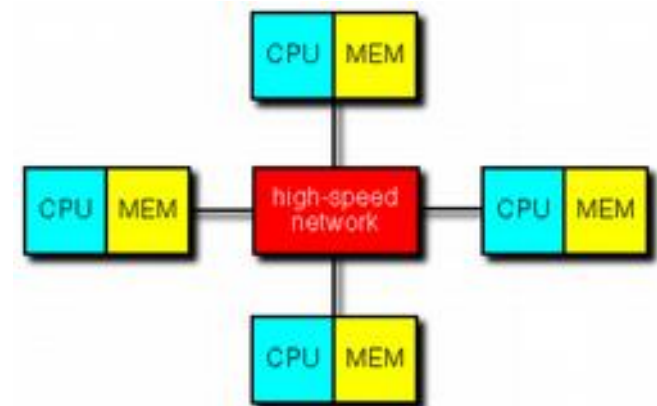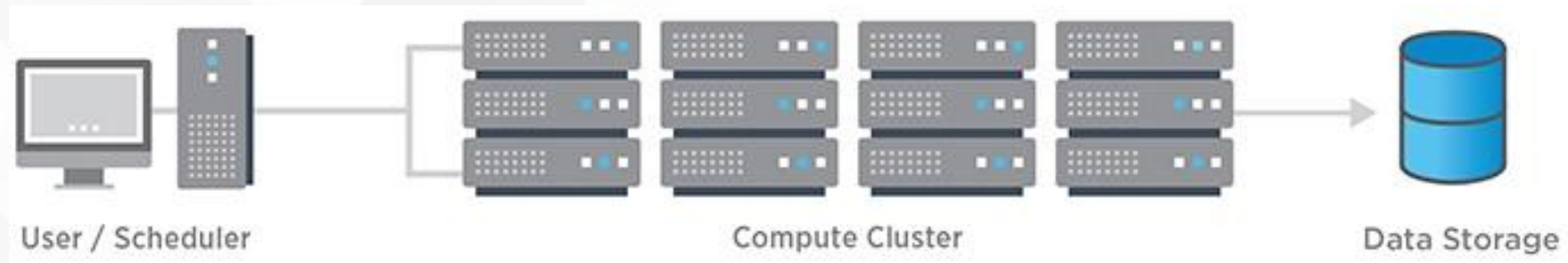
- Several computers (nodes) often in special cases for easy mounting in a rack

- One or more networks (interconnects) to hook the nodes together

- Storage facilities.

# The basic distributed memory machine: clusters



User / Scheduler          Compute Cluster          Data Storage
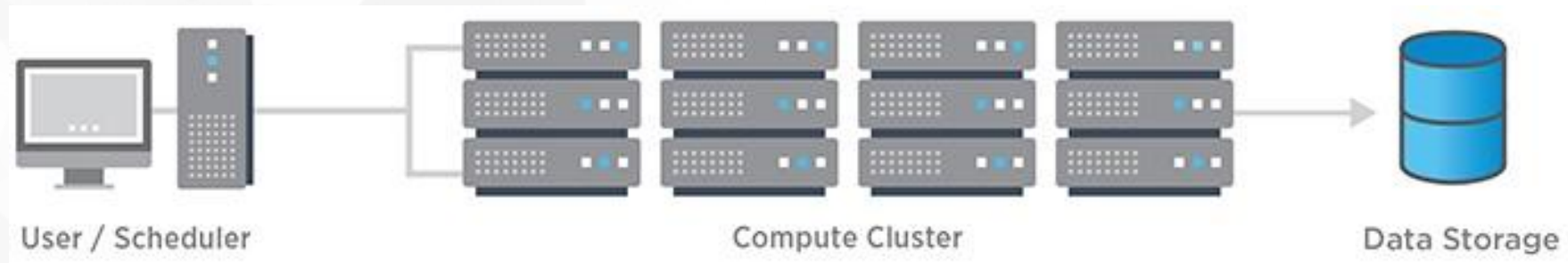
- The performance of the system are influenced by:

- Features of the node (RAM/cores/CPU frequency/ accelerator)
- Features (Topology and other) of the interconnection network

# Which architecture for HPC in 2019 ?

# The building blocks of the HPC cluster
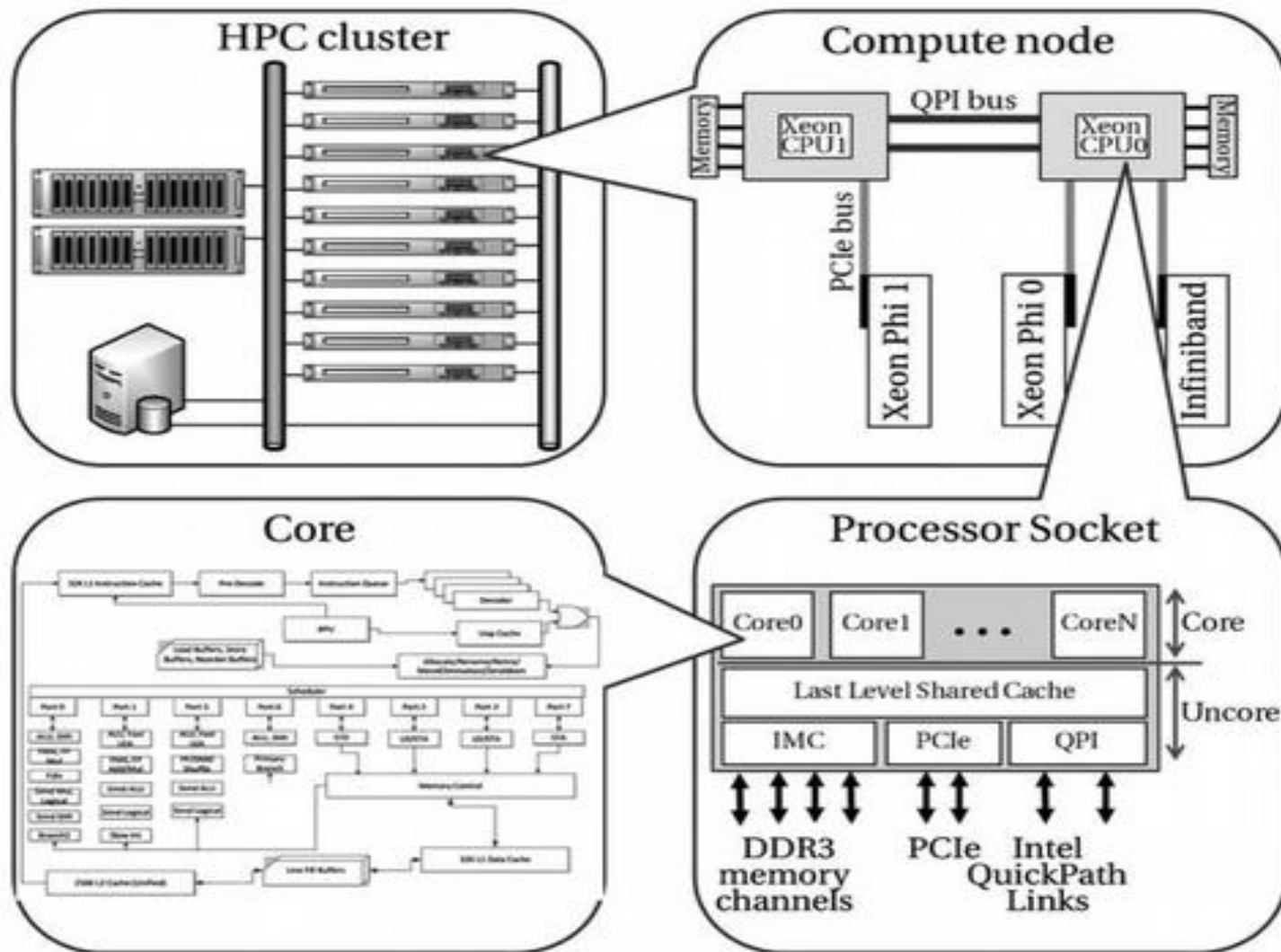
# A (not so) modern node of a cluster (Ulysses)
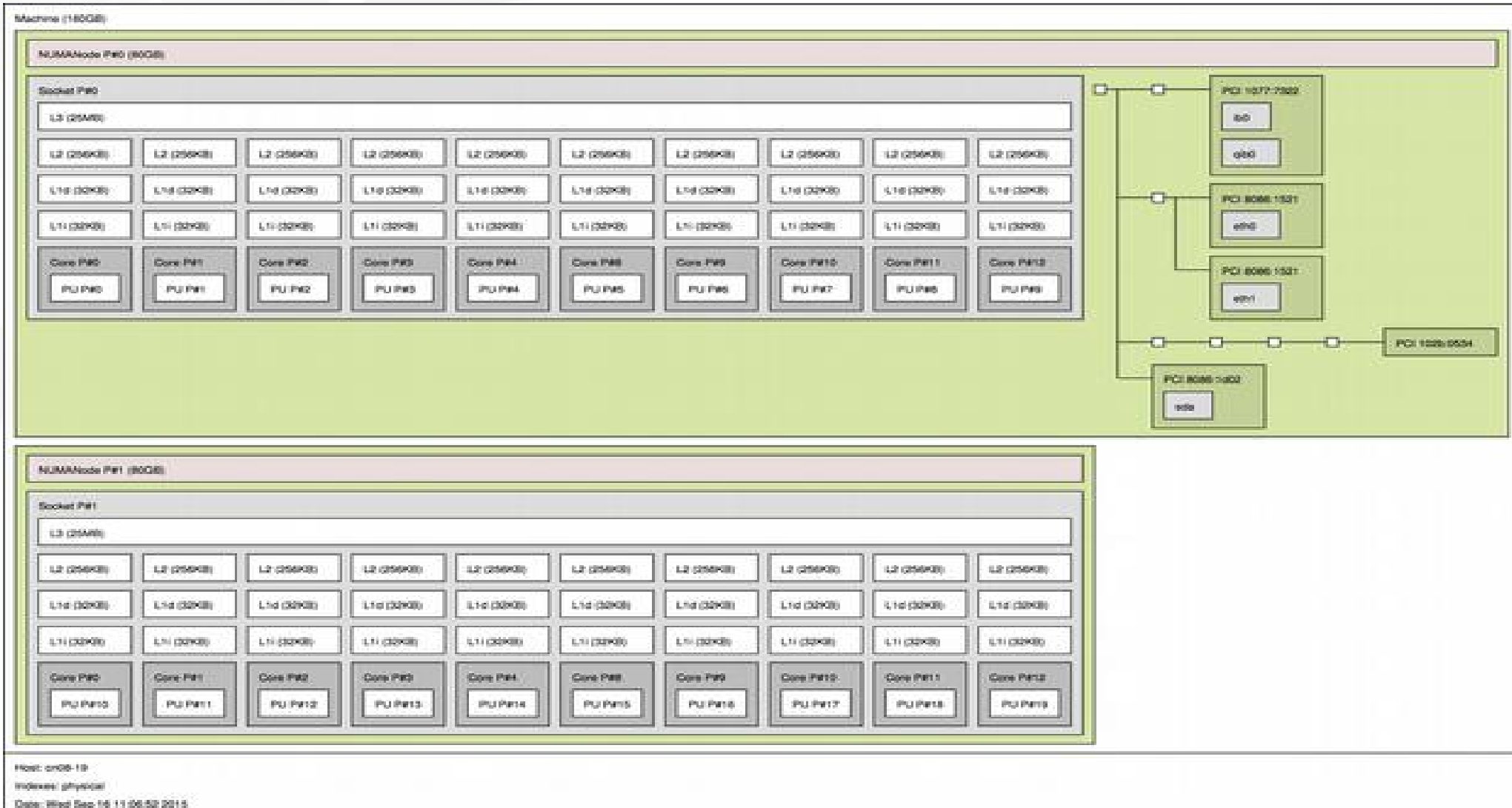
# HPC jargon

- Multiprocessor = server with more than 1 CPU
- Multicore= a CPU with more than 1 core
- Processor = CPU =socket

- BUT SOMETIME:
  - Processor= core
  - a process for each processor ( i.e. each core)

# Challenges ahead HPC (III)

- HPC skilled people
- BIG data: no longer HPC but HPDA/AI as well
- Complex and multicomponent HPC systems

# Measuring speed of HPC systems

- How fast can I crunch numbers on my CPUs ?

- How fast can I move data around ?

  - from CPUs to memory

  - from CPUs to disk

  - from CPUs on different machines

- How much data can I store ?

# Number crunching on CPU: what do we count ?

- Rate of [million/billions of] floating point operations per second ([M| G]flops) FLOPs/S

- Theoretical peak performance:

  - determined by counting the number of floating-point additions and multiplications  that can be completed during a period of time, usually the cycle time of the machine

FLOPS=Clock-rate*Number_of_FP_ operation*Number_of_cores

# Sustained (peak) performance

Real (sustained) performance:  a measure

measured by taking the time the code requires to run

$$\frac{(Number\_of\_floating\_point\_operations \text{ of the code })}{Time\ measured}$$

Number_of_floating_point_operations not easy to be defined for real application:

benchmarks are available for that..

Top500 list uses HPL Linpack:

Sustained peak performance is what's matter in TOP500

# TOP 500 List



- The TOP500 list www.top500.org
- published twice a year from 1993
  - ISC conference in Europe (June)
  - Supercomputing conference in USA (November)
- List the most powerful computers in the world
- yardstick: Linpack benchmark (LU – decomposition)

# HPL: some details

From http://icl.cs.utk.edu/hpl/index.html:

>The code solves a uniformly random system of linear equations and reports time and floating-point execution rate using a standard formula for operation count.
>
>Number_of_floating_point_operations = $2/3n^3 + 2n^2$ (n=size of the system)

```
================================================================================
T/V                N    NB     P     Q               Time                Gflops
--------------------------------------------------------------------------------
WR03R2L2        86000  1024     2     1              191.06             2.219e+03
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0043644 ...... PASSED
```

# HPL&TOP500

- For each machine the following numbers are reported using HPL:

  - Rmax: the performance in GFLOPS for the largest problem run on a machine.

  - Rpeak: the theoretical peak performance GFLOPS for the machine.

  - The measure of the power required to run the benchmark

# And the winner is...



Summit: DOE/SC/Oak Ridge National Laboratory
No.1 in Jun 2018

Sunway TaihuLight: National Supercomputing Center in Wuxi
No.1 from Jun 2016 until Nov 2017

Tianhe-2 (MilkyWay-2) : National University of Defense Technology
No.1 from Jun 2013 until Nov 2015

Titan: Oak Ridge National Laboratory
No.1 in Nov 2012

# 06/2019 Highlights (from TOP500)

The top of the list remains largely unchanged, with only two new entries in the top 10, one of which was an existing system that was upgraded with additional capacity.

- Two IBM-built supercomputers, Summit and Sierra, installed at the Department of Energy's Oak Ridge National Laboratory (ORNL) in Tennessee and Lawrence Livermore National Laboratory in California, respectively, retain the first two positions on the list. Both derive their computational power from Power 9 CPUs and NVIDIA V100 GPUs. The Summit system slightly improved its HPL result from six months ago, delivering a record 148.6 petaflops, while the number two Sierra system remains unchanged at 94.6 petaflops.

- The Sunway TaihuLight, a system developed by China's National Research Center of Parallel Computer Engineering & Technology (NRCPC) and installed at the National Supercomputing Center in Wuxi, holds the number three position with 93.0 petaflops. It's powered by more than 10 million SW26010 processor cores.

- At number four is the Tianhe-2A (Milky Way-2A) supercomputer, developed by China's National University of Defense Technology (NUDT) and deployed at the National Supercomputer Center in Guangzhou. It used a combination of Intel Xeon and Matrix-2000 processors to achieve an HPL result of 61.4 petaflops.

- Frontera, the only new supercomputer in the top 10, attained its number five ranking by delivering 23.5 petaflops on HPL. The Dell C6420 system, powered by Intel Xeon Platinum 8280 processors, is installed at the Texas Advanced Computing Center of the University of Texas.

**#1**

## Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband

| | |
|---|---|
| Site: | DOE/SC/Oak Ridge National Laboratory |
| System URL: | http://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/ |
| Manufacturer: | IBM |
| Cores: | 2,282,544 |
| Memory: | 2,801,664 GB |
| Processor: | IBM POWER9 22C 3.07GHz |
| Interconnect: | Dual-rail Mellanox EDR Infiniband |
| **Performance** | |
| Linpack Performance (Rmax) | 122,300 TFlop/s |
| Theoretical Peak (Rpeak) | 187,659 TFlop/s |
| Nmax | 13,989,888 |
| HPCG [TFlop/s] | 2,925.75 |
| **Power Consumption** | |
| Power: | 8,805.50 kW (Submitted) |

# Top 1% +1

| Rank | Site | System | Cores | Rmax (TFlop/s) | Rpeak (TFlop/s) | Power (kW) |
|---|---|---|---|---|---|---|
| 1 | DOE/SC/Oak Ridge National Laboratory United States | Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM | 2,414,592 | 148,600.0 | 200,794.9 | 10,096 |
| 2 | DOE/NNSA/LLNL United States | Sierra - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM / NVIDIA / Mellanox | 1,572,480 | 94,640.0 | 125,712.0 | 7,438 |
| 3 | National Supercomputing Center in Wuxi China | Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC | 10,649,600 | 93,014.6 | 125,435.9 | 15,371 |
| 4 | National Super Computer Center in Guangzhou China | Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT | 4,981,760 | 61,444.5 | 100,678.7 | 18,482 |
| 5 | Texas Advanced Computing Center/Univ. of Texas United States | Frontera - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR Dell EMC | 448,448 | 23,516.4 | 38,745.9 | |
| 6 | Swiss National Supercomputing Centre (CSCS) Switzerland | Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc. | 387,872 | 21,230.0 | 27,154.3 | 2,384 |

# First EU machine..

## Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100

| | |
|---|---|
| **Site:** | Swiss National Supercomputing Centre [CSCS] |
| **System URL:** | http://www.cscs.ch/computers/piz_daint_piz_dora/index.html |
| **Manufacturer:** | Cray Inc. |
| **Cores:** | 361,760 |
| **Memory:** | 340,480 GB |
| **Processor:** | Xeon E5-2690v3 12C 2.6GHz |
| **Interconnect:** | Aries interconnect |
| **Performace** | |
| **Linpack Performance (Rmax)** | 19,590 TFlop/s |
| **Theoretical Peak (Rpeak)** | 25,326.3 TFlop/s |
| **Nmax** | 3,569,664 |
| **HPCG [TFlop/s]** | 470.0 |
| **Power Consumption** | |
| **Power:** | 2,271.99 kW [Optimized: **1631.13** kW] |
| **Power Measurement Level:** | 3 |
| **Measured Cores:** | 361,760 |
| **Software** | |
| **Operating System:** | Cray Linux Environment |
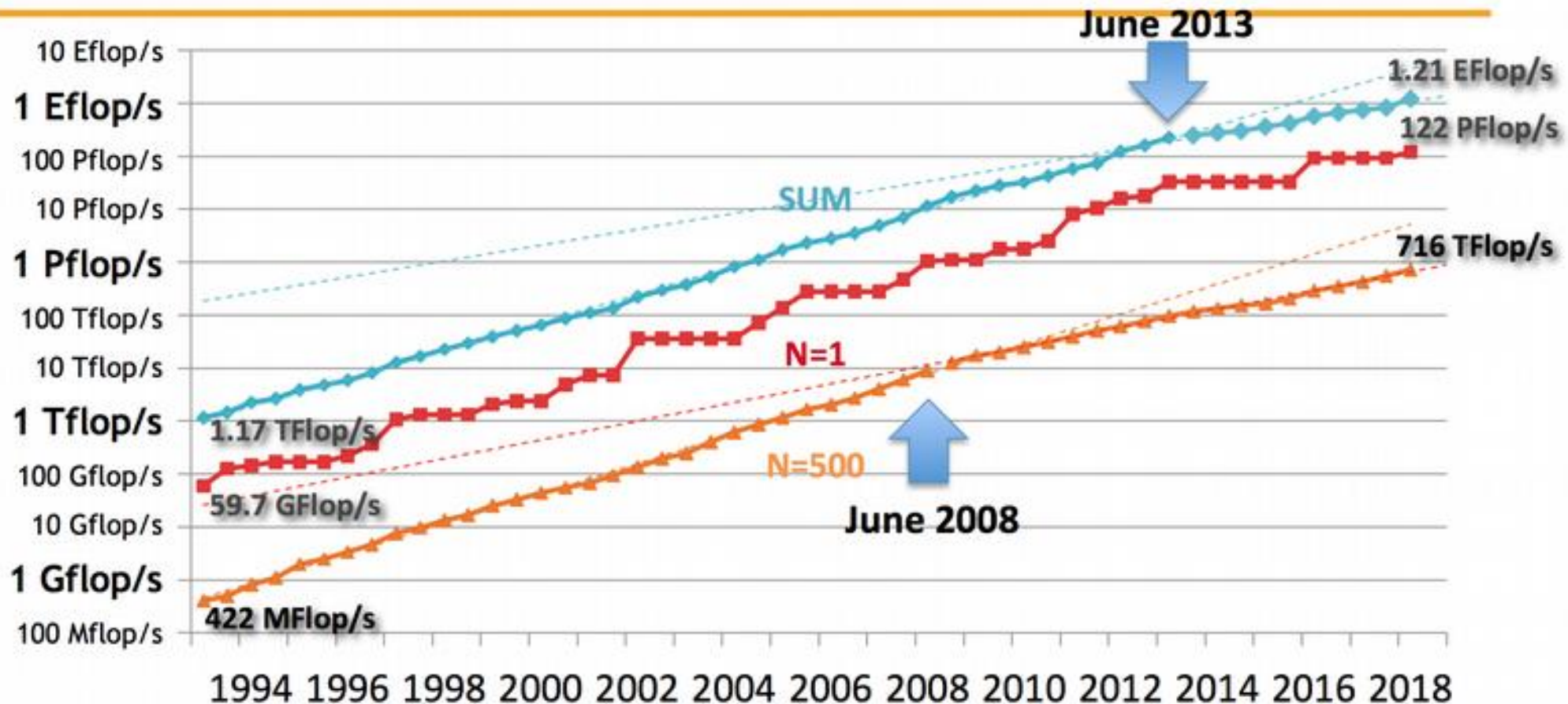
# Sustained peak performance on real scientific codes

Blue-waters at NCSA: 22,640 AMD 6276 processors
Theoretical peak performance: 13 Petaflops
Sustained performance on real scientific codes:..

| Scientific code | Number of cores | Performance achieved(PF) | runtime (hour) |
| --- | --- | --- | --- |
| VPIC | 22528 | 1.25 | 2.5 |
| PPM | 21417 | 1.23 | ~ 1 |
| QMCPACK | 22500 | 1.037 | ~1 |
| SPECF3MD | 21675 | >1 | Not reported |
| WRF | 8192 | 0,160 | <0.50 |

http://www.cray.com/sites/default/files/resources/XE6-NCSA-PFApplications-0514.pdf

# Why Performance degradation ?

- HPC system is unable to exploit all the resources all of the time
- Many different causes and many parts of the HPC are responsible all together
- At abstract level four important factors:
  - Starvation
  - Latency
  - Overhead
  - Waiting for Contention    =>SLOW

Taken from paragraph 1.4.5 of reference one

# Starvation

- Happens when sufficient work is not available at any instance in time to support issuing instructions to all functional units every cycle.

- Typical case:

  - Not enough parallel work for all processors/components

  - Parallel work not evenly distributed among all processors/components (load is not balanced)

# Latency

- Time it takes for information to move from one part of the system to the other.

- Typical cases:

  - Memory access

  - Data transfer between separate nodes

- Lot of tricks to hide latency (see next lectures)

# Overhead

- The amount of additional work needed beyond that which is actually required to perform the computation.

- Typical cases:
  - Time to spawn and synchronize parallel tasks
  - Other kind of operation not directly associated to the computation

- The above operations steals resources to the computation and should be minimized

# Waiting for contention

- Two or more request are made at the same time on the same resource (either HW or SW)..

- Typical cases:
  - Two task writing on the same disk and/or sending message to the same memory location at the same time

- Generally such events are not predictable and so difficult to avoid and to optimize.

# Challenges ahead HPC (IV)

- HPC skilled people
- BIG data: no longer HPC but HPDA/AI as well
- Complex and multicomponent HPC systems
- Sustained performance not just HPL !!!

# From twitter..

# Exercise 0..

- 0.1 Compute Theoretical Peak performance for your lap top/desktop
  - Identify the CPU
  - Identify the frequency
  - Identify the number of floating point for cycle
  - Identify how many cores
  - Put all together in one single number and tell me
- 0.2 compute sustained Peak performance for your cell-phone
  - Identify an app to run HPL
  - Run it & Tune it to get the best number you can
- 0.3 Find out in which year your cell phone/laptop could have been in top1% of Top500

# Moving data around: bits and Mb/sec

- within the node:
  - CPU-Memory: thousand of Mb/sec GByte/sec
    - 10 - 100 Gbit
  - CPU- Disks :  MByte/sec
    - 50 ~ 100 MB up  1000MB/sec
- Among node (cluster)
  - networks
    - default (commodity)
      - 1000Mbit=1Gbit
  - custom(high speed)
    - 10Gb  and now 100 Gb (and even more)

# About network for cluster

- The  performance  of the network cannot  be ignored
  - Latency:  Initialization time before data can be sent
  - Per-link Peak Bandwidth:  Maximum data transmission rate (varies with packet size)
  - Topology: how the network is done.
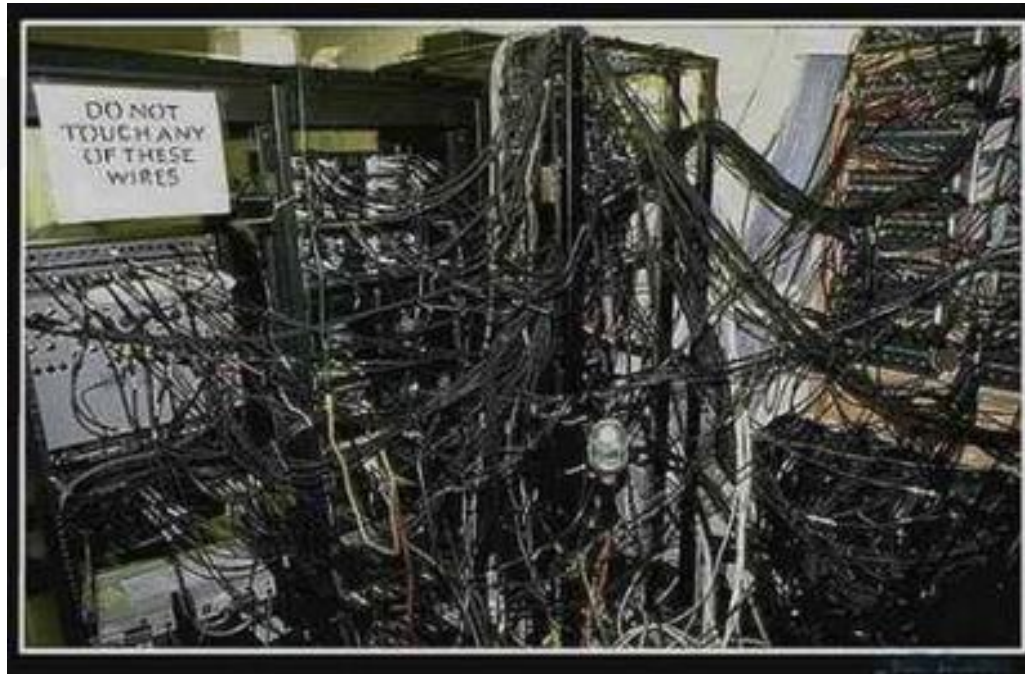
# Latency&bandwidth

| NETWORK | Latency | Bandwidth (GB/sec) |
|---|---|---|
| Gigabit | 70-40 | ~ 0.125 |
| 10G | <5 | ~1.250 |
| Infiniband 4DDR | ~1.5/1.9 | ~ 3.2 |
| Infiniband FDR | <1.0 | ~ 5 |

What is the UNIT OF MEASURE OF LATENCY ?

Microseconds: 3 order of magnitude larger than unit of measure of FP operations

# Storage size: bytes

- size of storage devices:
  - Kbyte/Mbyte -->caches/RAM
  - Gigabyte      ---> RAM/hard disks(small size)
  - Terabyte    ---> Disks/SAN
  - Petabyte   ----> SAN / Tapes devices
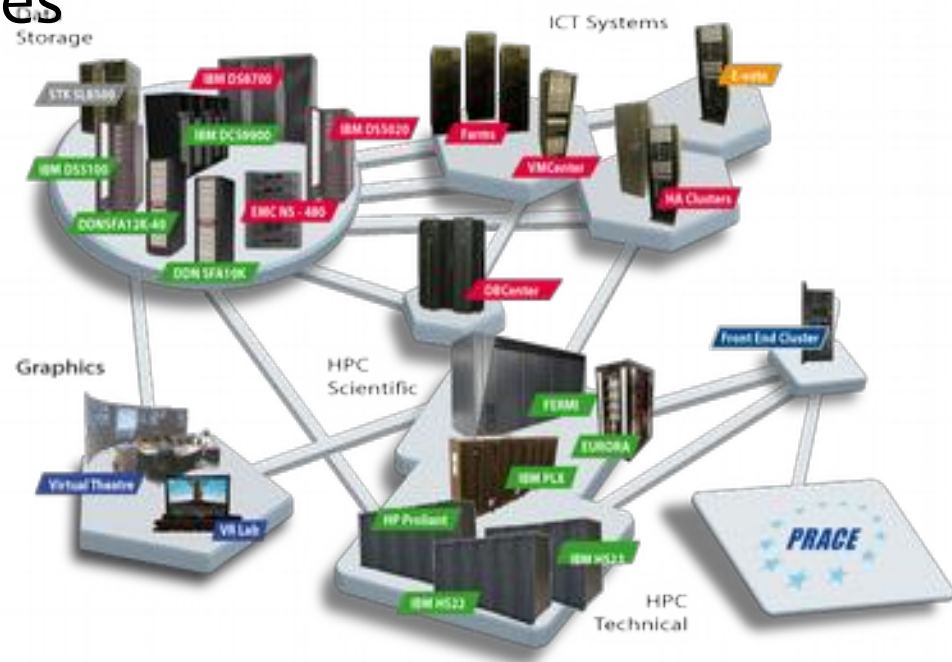
# Last but not least: Storage

- High Speed Storage is required for HPC
  - Parallel Filesystem is mandatory:
    - Lustre/GPFS/BeeGFS etc..
- Hierarchical storage is also a solution:
  - Hierarchical storage management (HSM) is a data storage technique, which automatically moves data between high-cost and low-cost storage media.
    - First layer: SSD
    - Second layer : parallel FS
    - Third layer: SAN
    - Fourth layer: Tapes

# Gigabyte or gibibyte ?

- The gibibyte is a multiple of the unit byte for digital information. The binary prefix gibi means $2^{30}$, therefore one gibibyte is equal to 1073741824bytes = 1024 mebibytes. The unit symbol for the gibibyte is GiB. Defined by the International Electrotechnical Commission (IEC) in 1998.

- The gibibyte is closely related to the gigabyte (GB), which is defined by the IEC as $10^9$ bytes = 1GiB ≈ 1.074GB.

- 1024 gibibytes are equal to one tebibyte.

- In the context of computer memory, gigabyte and GB are customarily used to mean 1024 ($2^{30}$) bytes, although not in the context of data transmission and not necessarily for hard drive size.

[Adopted from https://en.wikipedia.org/wiki/Gibibyte]

# From Wikipedia

| Multiples of bytes | | | | | v · t · e |
|---|---|---|---|---|---|
| **Decimal** | | **Binary** | | | |
| **Value** | **Metric** | **Value** | **IEC** | | **JEDEC** |
| 1000 | kB kilobyte | 1024 | KiB kibibyte | | KB kilobyte |
| $1000^2$ | MB megabyte | $1024^2$ | MiB mebibyte | | MB megabyte |
| $1000^3$ | GB gigabyte | $1024^3$ | GiB **gibibyte** | | GB gigabyte |
| $1000^4$ | TB terabyte | $1024^4$ | TiB tebibyte | | – |
| $1000^5$ | PB petabyte | $1024^5$ | PiB pebibyte | | – |
| $1000^6$ | EB exabyte | $1024^6$ | EiB exbibyte | | – |
| $1000^7$ | ZB zettabyte | $1024^7$ | ZiB zebibyte | | – |
| $1000^8$ | YB yottabyte | $1024^8$ | YiB yobibyte | | – |
| Orders of magnitude of data | | | | | |

# IOPS vs FLOPS

- HPC is too compute-centric
- Modern Scientific&technical computing requires access to data and computing

computing 1 calculation
≈ 1 picojoule

moving 1 calculation
≈ 100 picojoule
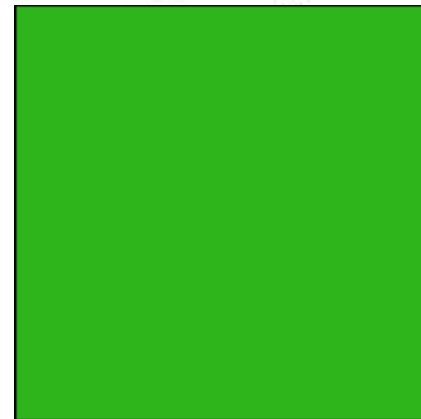
Source: IDC Direction 2013

# How much energy do we need to run an exascale machine ?

- A lot !
- Exercise: compute how much energy you need to reach 1exaflop using #1 in 2012 /2015/2018

# Top500&Green500

- Over the last years , energy efficiency increased substantially
- BUT Exaflops machine today not yet sustainable in term of energy !
- Sustainability is set to 20MW !
- Though to be reached in 2018 now postponed to 2023
- Green500 is the Top500 reorder according to energy efficiency…

# The Ultimate Goal of "The Green500"

- Raise awareness of energy efficiency in supercomputing.
  - Drive energy efficiency as a first-order design constraint (on par with FLOPS).

    Encourage fair use of the list rankings to promote energy efficiency in high-performance computing systems.

# Top Green500  June 2019

| Rank | TOP500 Rank | System | Cores | Rmax (TFlop/s) | Power (kW) | Power Efficiency (GFlops/watts) |
|---|---|---|---|---|---|---|
| 1 | 469 | **DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States | 22,440 | 1,070.0 | 97 | 15.113 |
| 2 | 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 | 10,096 | 14.719 |
| 3 | 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 | 1,649 | 14.423 |
| 4 | 393 | **MareNostrum P9 CTE** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100 , IBM Barcelona Supercomputing Center Spain | 18,360 | 1,145.0 | 81 | 14.131 |
| 5 | 25 | **TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan | 135,828 | 8,125.0 | 792 | 13.704 |

# Highlights..

- The most energy-efficient system and No. 1 on the Green500 is the DGX SaturnV Volta system, a NVIDIA system installed at NVIDIA, USA. It achieve 15.1 GFlops/Watt power efficiency. It is on position 469 in the TOP500.

- They are followed on No 3 by Summit at the Oak Ridge National Laboratory (ORNL) in Tennessee. It achieved 14.7 gigaflops/watt and is listed at number one in the TOP500.

- The Shoubu system B, a ZettaScaler-2.2 system at the Advanced Center for Computing and Communication, RIKEN, Japan was decommissioned March 2019 and was removed from the TOP500 list.

- All the top 1% have accelerators installed...

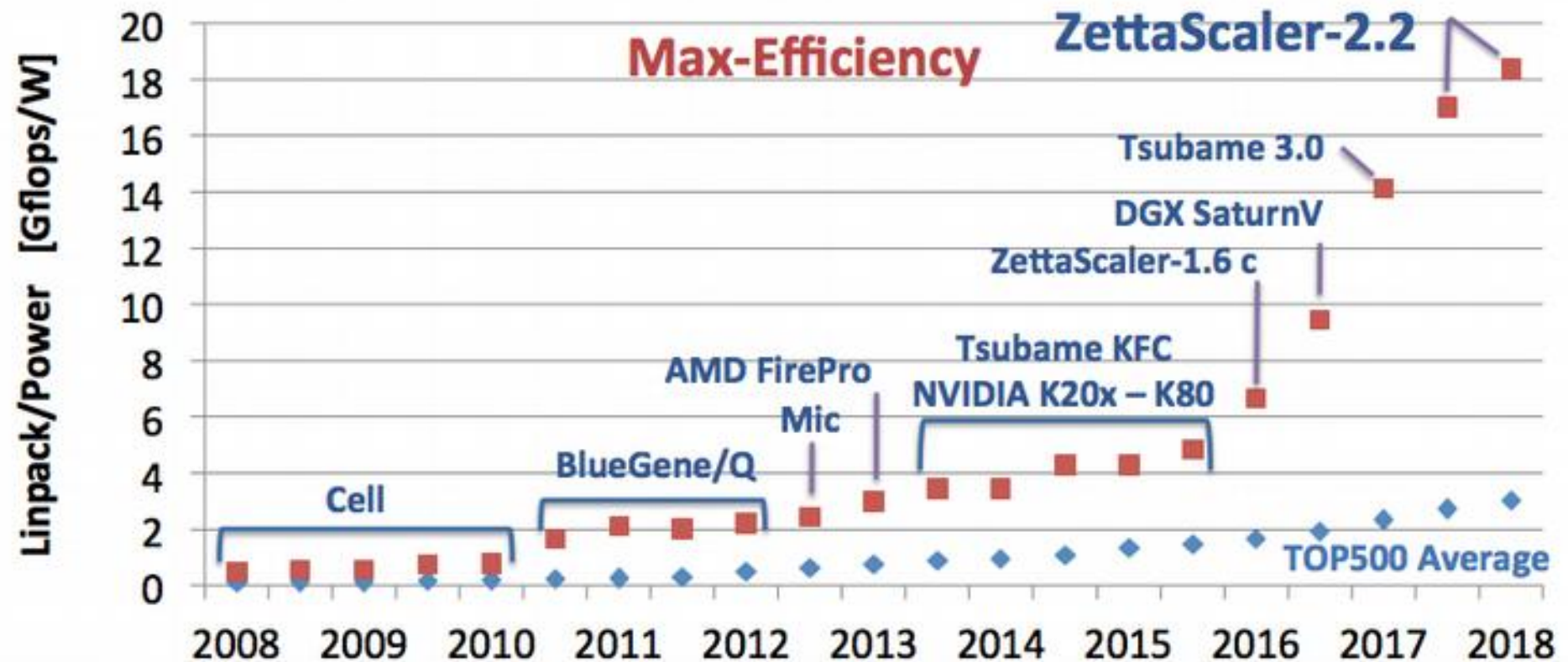# Power efficiency (1)
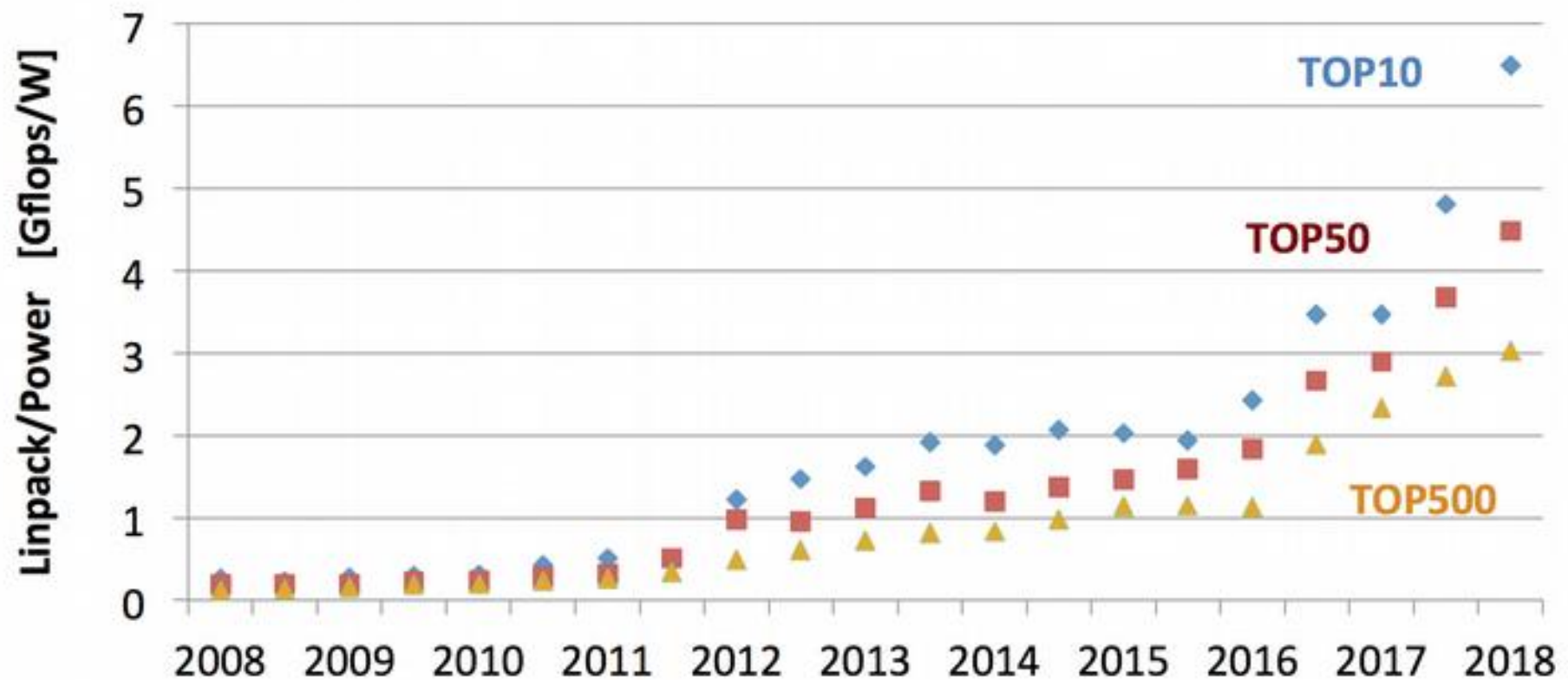
# Power efficiency (max-efficiency)

# Challenges ahead HPC (V)

- HPC skilled people

- BIG data: no longer HPC but HPDA/AI as well
  - Complex and multicomponent HPC systems

- Sustained performance not just HPL !!!

- Energy

# How much does it cost a computational infrastructure ?

- It is not just a matter of HW…
- Total Cost of Ownership is the right way to calculate the budget for an HPC infrastructure..

# Total Cost of Ownership

- It is the sum of all of the costs that a customer incurs during the lifetime of a technology solution.

- In the High Performance Computing (HPC) field, the Total Cost of Ownership is normally referred to the data center costs.

- Cost to the owner to build, operate and maintain the data center.

- Cost of Services delivered should be computed taking into account TCO.

# What should be included in the TCO for HPC ?

- Investment, operation and maintenance costs:
  - Hardware: servers, storage, networking, cabling, etc.
  - Electrical equipment: power distribution units, UPS, generators, etc.
  - Cooling systems: air conditioners, water cooling, etc.
- Infrastructure for the data center, power adaptation issues, etc.
- Energy consumption of the hardware and cooling systems
- Software licenses
- Human resources
- Maintenance

# HOW can I reduce TCO ?

# Cloud computing is the answer ?
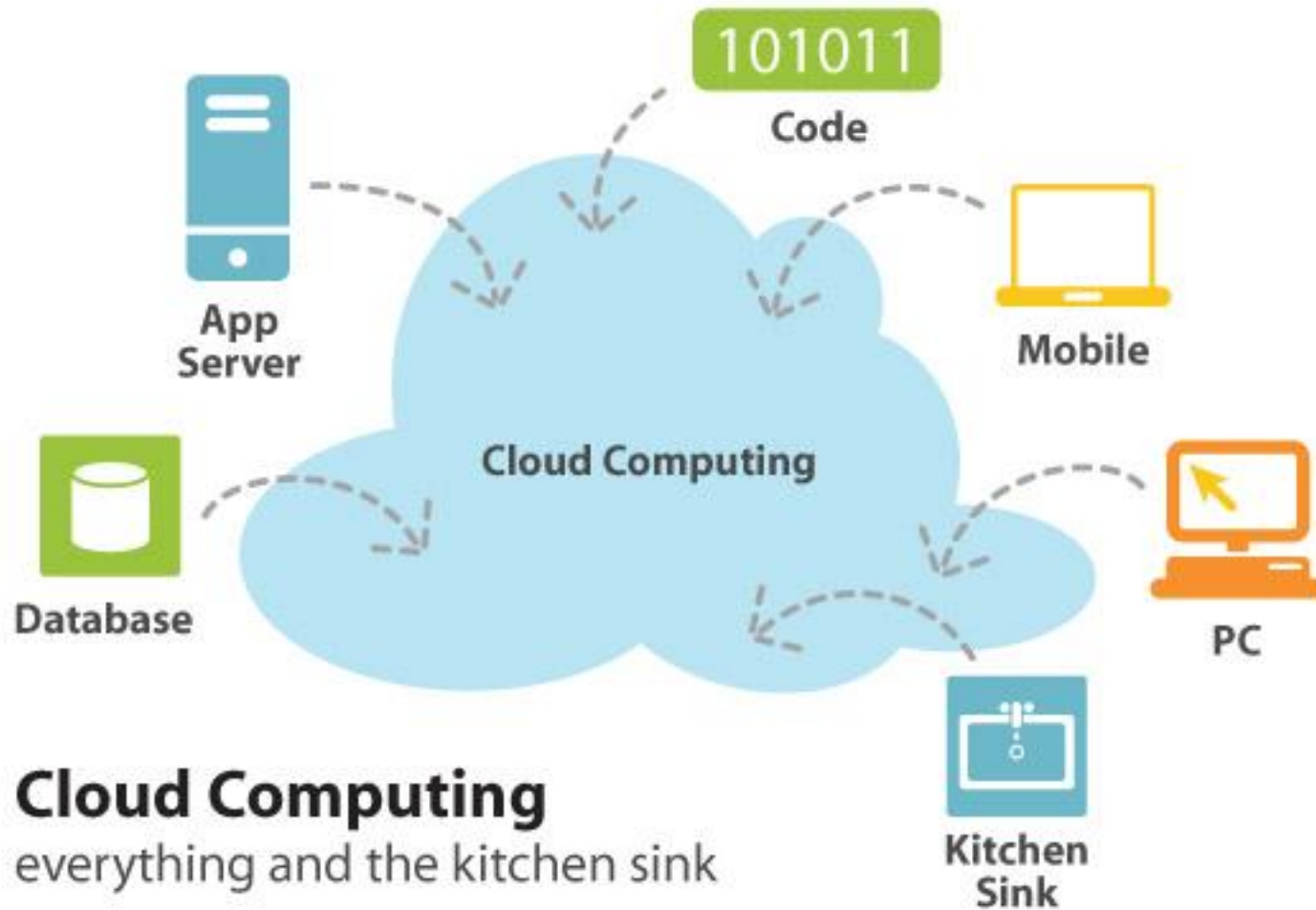
# Cloud approach
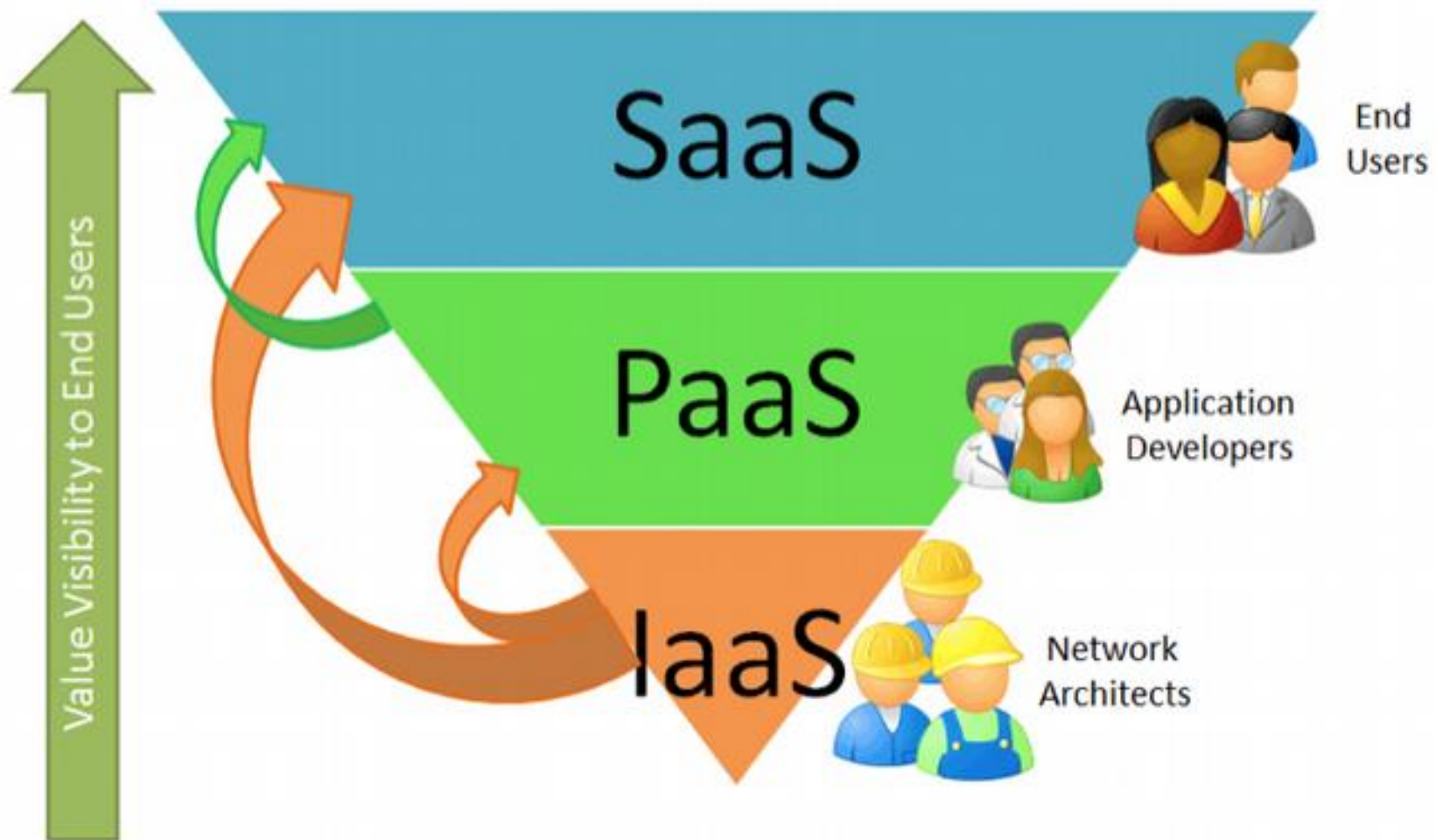
# The dream

- Cloud computing offers almost unlimited storage and instantly available and scalable computing resources,

- All the above at a reasonable metered cost.. (pay per use)

- However…
  - the use of a typical cloud needs a bit of care..
  - Remote HPC services can range from shared HPC clusters to fully virtualized cloud environments.

# Cloud computing and HPC

"The case for HPC in the cloud is growing stronger, but still has a way to go, especially for the more traditional HPC segments in the public sector"

From https://www.hpcwire.com/2018/03/15/how-the-cloud-is-falling-short-for-research-computing/

# HPC on cloud..

- cloud computing represented about 2% of the HPC market by total revenue in 2016.

- About 35% of HPC users make occasional use of public cloud resources.

- A number of vendors already exist within the industry providing HPC in the cloud solutions.

# HPC cloud providers

- AMAZON WEB SERVICES (AWS)
  - https://aws.amazon.com/it/hpc/
- MICROSOFT AZURE
  - https://azure.microsoft.com/it-it/solutions/high-performance-computing/
- IBM SPECTRUM COMPUTING
  - https://www.ibm.com/it-it/it-infrastructure/spectrum-computing
- GOOGLE CLOUD
  - https://cloud.google.com/solutions/hpc/?hl=it
- Rescale
  - ….

# Challenges ahead HPC (VI)

- HPC skilled people
- BIG data: no longer HPC but HPDA/AI as well
  - Complex and multicomponent HPC systems

- Sustained performance not just HPL !!!
- Energy
- On premises HPC or not ?

# Conclusions

- HPC is about performance but not only

- Supercomputers are clusters !

- Clusters have many different components

- Many challenges ahead to:

  - Use/Exploit a HPC system

  - Plan/ Mantain a HPC system

- There are a lot of other lectures where all what we touch just briefly in this first lecture will be analyzed in details