High Performance Computing and Data Infrastructure

Recap on HPC concepts

2024-2025 @ Università di Trieste

Agenda

- Prologue: why and where HPC ?
- What is HPC ?
- Performance and metrics
- Supercomputers and TOP500
- Parallel computers

Before starting: HPC prefix..

Factor	Name	Symbol
10^{24}	yotta	Y
10^{21}	zetta	Z
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k

- How large is ceph storage on ORFEO ?
- How much RAM on thin nodes ?
- How large is the L1 cache of Epyc nodes ?
- What is the CPU frequency of large nodes?

HPC not easy to define..

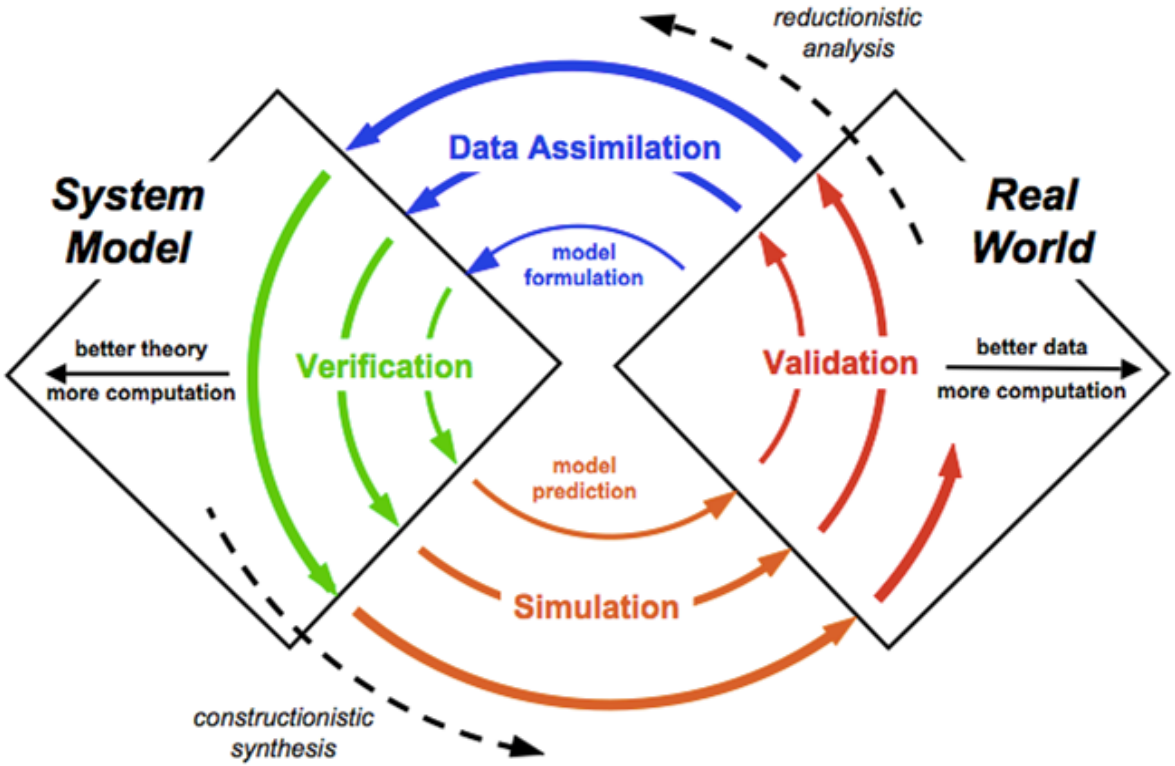
High performance computing (HPC), also known as supercomputing, refers to computing systems with extremely high computational power that are able to solve hugely complex and demanding problems.

[Taken from https://ec.europa.eu/digital-single-market/en/high-performance-computing]

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



Data are flooding us..

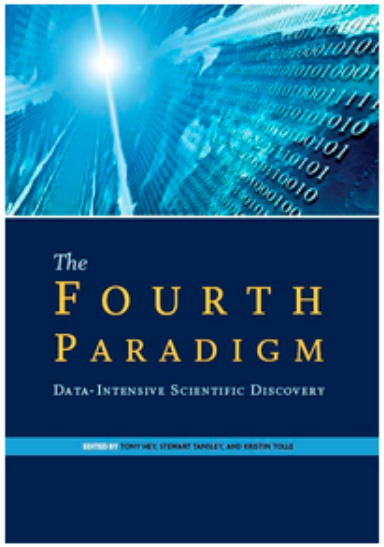
In today's world, larger and larger amounts of data are constantly being generated, from 33 zettabytes globally in 2018 to an expected 181 zettabytes in 2025). As a result, the nature of computing is changing, with an increasing number of data-intensive critical applications. HPC is key to processing and analysing this growing volume of data, and to making the most of it for the benefit of citizens, businesses, researchers and public administrations.

[Taken again from https://ec.europa.eu/digital-single-market/en/high-performance-computing]

Data intensive science

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science



Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Critical praise for *The Fourth Paradigm*

Download

- Full text, low resolution (6 MB)
- Full text, high resolution (93 MB)
- By chapter and essay

Purchase from Amazon.com

- Paperback
- Kindle version

In the news

- Sailing on an Ocean of 0s and 1s *(Science* Magazine)
- A Deluge of Data Shapes a New Era in Computing *(New York Times)*
- A Guide to the Day of Big Data *(Nature)*

Big data challenge: from HPC to HPC/AI

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

Complex problem 2: ChatGPT…

TECH

ChatGPT and generative AI are booming, but the costs can be extraordinary

PUBLISHED MON, MAR 13 2023•8:58 AM EDT | UPDATED MON, APR 17 2023•2:09 AM EDT

Jonathan Vanian
@JONATHANVANIAN

Kif Leswing
@KIFLESWING

WATCH LIVE

KEY POINTS

- The cost to develop and maintain the software can be extraordinarily high.
- Nvidia makes most of the GPUs for the AI industry, and its primary data center workhorse chip costs $10,000.
- Analysts and technologists estimate that the critical process of training a large language model such as GPT-3 could cost over $4 million.

Agenda

- ~~Prologue: why and where HPC ?~~ 
- What is HPC ?
- Performance and metrics ?
- Supercomputers and TOP500
- Parallel Computers

HPC: a ~~first~~ second definition

High Performance Computing (HPC) is the use of servers, clusters, and supercomputers – plus associated software, tools, components, storage, and services – for scientific, engineering, or analytical tasks 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 industry, government and academia.

[to be continued]

Elements of the HPC ecosystem..

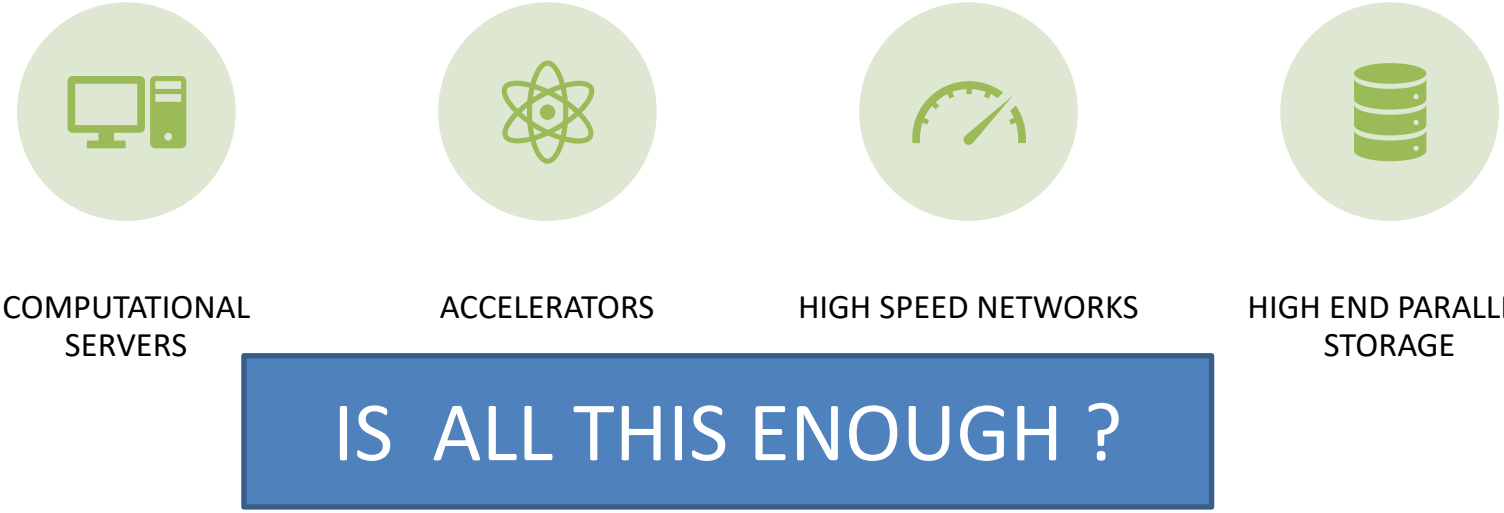
- use of servers, clusters, and supercomputers

  → HARDWARE
- associated software, tools, components, storage, and services

  → SOFTWARE
- scientific, engineering, or analytical tasks

  → PROBLEMS TO BE SOLVED..

Elements of an HPC infrastructure ?



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

Agenda

- ~~Prologue: why and where HPC ?~~ ✓
- ~~What is HPC ?~~ ✓
- Performance and metrics
- Supercomputers and TOP500
- Parallel Computers

It is 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]

Performance vs Productivity

- A possible definition:
  - Productivity = (application performance) / (application programming effort)
- Example:
  - To speed up a code by a factor of two it takes 6 months work
  - does this deserve to be done ?
- people in HPC arena have different goals in mind thus different expectations and different definitions of productivity.

How do measure (basic) performance 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

FLOPS= (total number of floating point operations done by a program) / (time the program takes to run in second)

- 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

Agenda

- ~~Prologue: why and where HPC ?~~ ✓
- ~~What is HPC ?~~ ✓
- ~~Performance and metrics~~ ✓
- Supercomputers and TOP500
- Parallel Computers

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 (HPL)

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
```

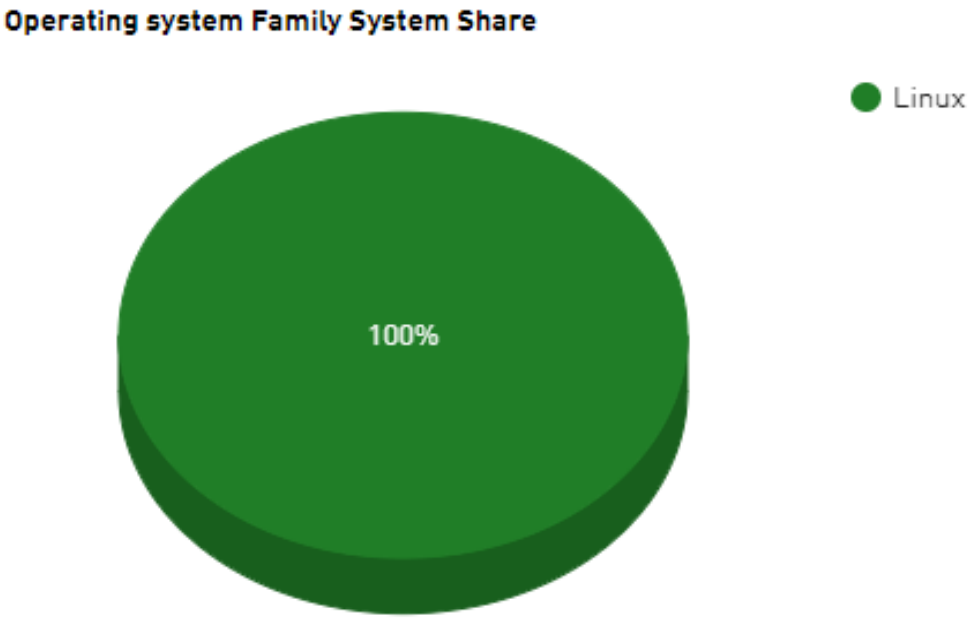
HPL&TOP 500 List

- 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

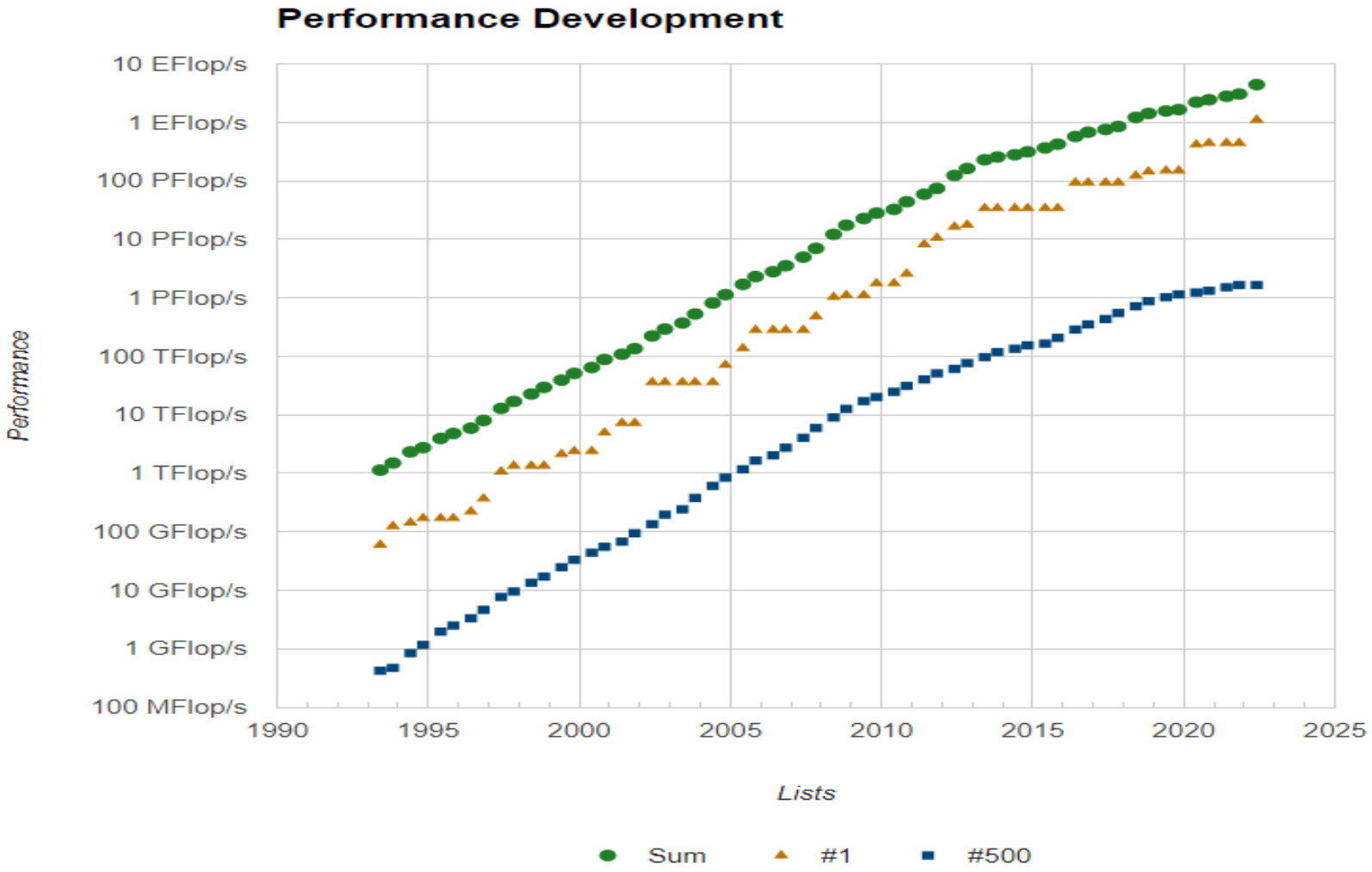
And the winner is..

Let us check this on line directly

By operating system



Performance development

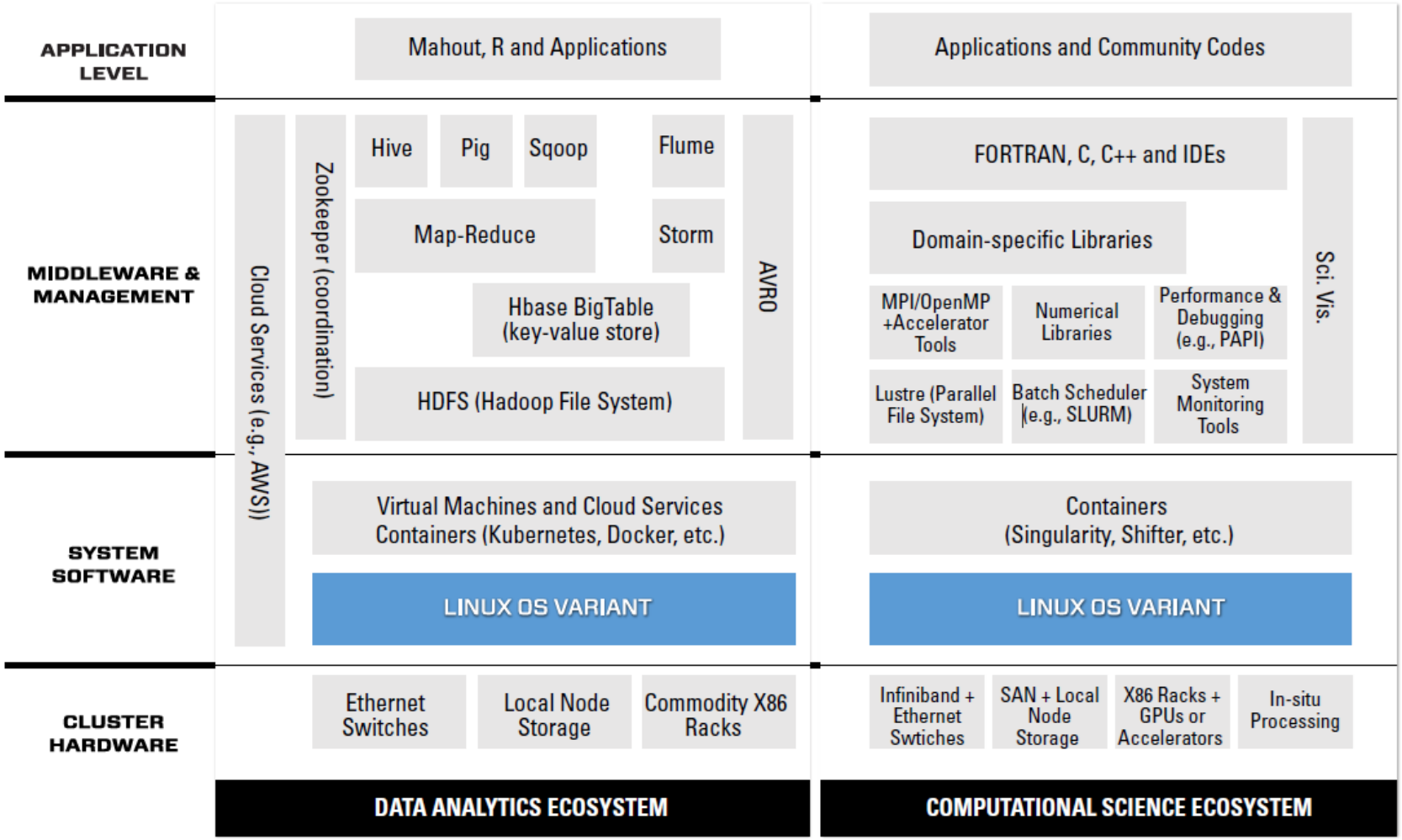



66% 06:31
Tweet
HPC Guru
@HPC_Guru
"Peak FLOPS is the best way to rank a #supercomputer or any other #HPC system"
Traduci il Tweet
Kate Hackett @HackettKate · 4g
You've been kidnapped. Your kidnappers allow you to keep tweeting to pretend everything is alright. What would you tweet that would alarm your followers without the kidnappers knowing you're asking for help?...
Mostra questa discussione
22:08 · 21 Set 19 · TweetDeck
6 Retweet 76 Mi piace
Twitta una risposta

Agenda

- ~~Prologue: why and where HPC ?~~ ✓
- ~~What is HPC ?~~ ✓
- ~~Performance and metrics~~ ✓
- ~~Supercomputers and TOP500~~ ✓
- Software for HPC and HPC/AI

Software stacks: what a difference !



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

Conclusions for today…

- HPC/AI infrastructure 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