High performance data analysis

GIULIANO TAFFONI

IVOA 2017 – 23 OCT 2017 - SANTIAGO CHILE

Horizon 2020

What is HPDA


BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
SHARED
DEFINITION
CURRENT
MOVING
EXAMPLES
EVERY
PRACTITIONERS
CAPTURE
WITHIN
BUSINESS
INTERNET
DISTRIBUTED
SETS
GENOMICS
MANAGEMENT
TERABYTES
CASE
ABILITY
PETABYTES
INCLUDE
ELAPSED
SYSTEMS
TOLERABLE
SIZE
STORAGE
MPP
PARALLEL
SAN
MASSIVELY
GROW
DIFFICULTY
GARTNER
WORKING
SEARCH
RECORDS
NETWORKS
DATABASES
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
ANALYTICS
TENS
COMPLEX
CITATION
LARGER
CONTINUES
SOCIAL
RELATIONAL
SET
USE

HPDA is…

The ability of increasingly powerful HPC systems to **run data-intensive problems** at larger scale, at higher resolution, and with more elements (e.g., inclusion of the carbon cycle in climate ensemble models)

The proliferation of **larger, more complex scientific instruments** and sensor networks, from "smart" power grids to the Large Hadron Collider and Square Kilometer Array.

The growth of stochastic **modeling, parametric modeling** and other iterative problem-solving methods, whose cumulative results produce large data volumes.

The availability of newer **advanced analytics methods and tools**: MapReduce/Hadoop, graph analytics (NVIDIA IndeX), semantic analysis, knowledge discovery algorithms (IBM Watson), COMPS and pyCOMS, and more

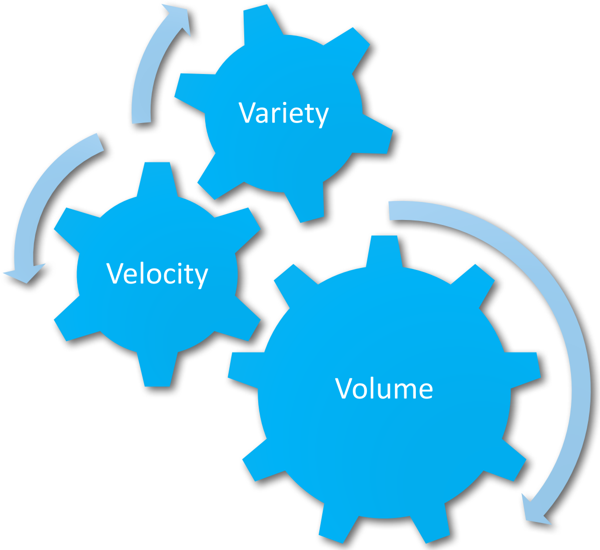
The escalating need to perform advanced analytics in **near-real time**—a need that is causing a new wave of commercial firms to adopt HPC for the first time

What drivers towards HPC

Complexity. HPC technology allows scientist to aim more complex, intelligent questions at their data infrastructures.

Time to value. Science faces ever-shortening innovation and production cycles. Analytics (including Hadoop and Spark) is moving from batch processing toward low-latency, interactive capabilities.

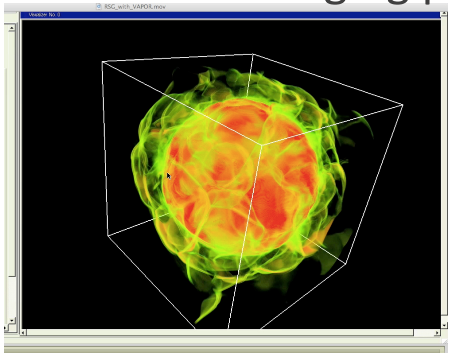
Variability. “deep” vs “Wide”
“large amount of data” vs “many variables”



What users expects from HPC

Simulations: New computing capabilities => finer results, larger parameters space, larger dynamic range

Real-time: find patterns that we do not expect and react consequently (execute new simulations or refine data reduction changing parameters)



Visualization and Interactivity:

3D visualization

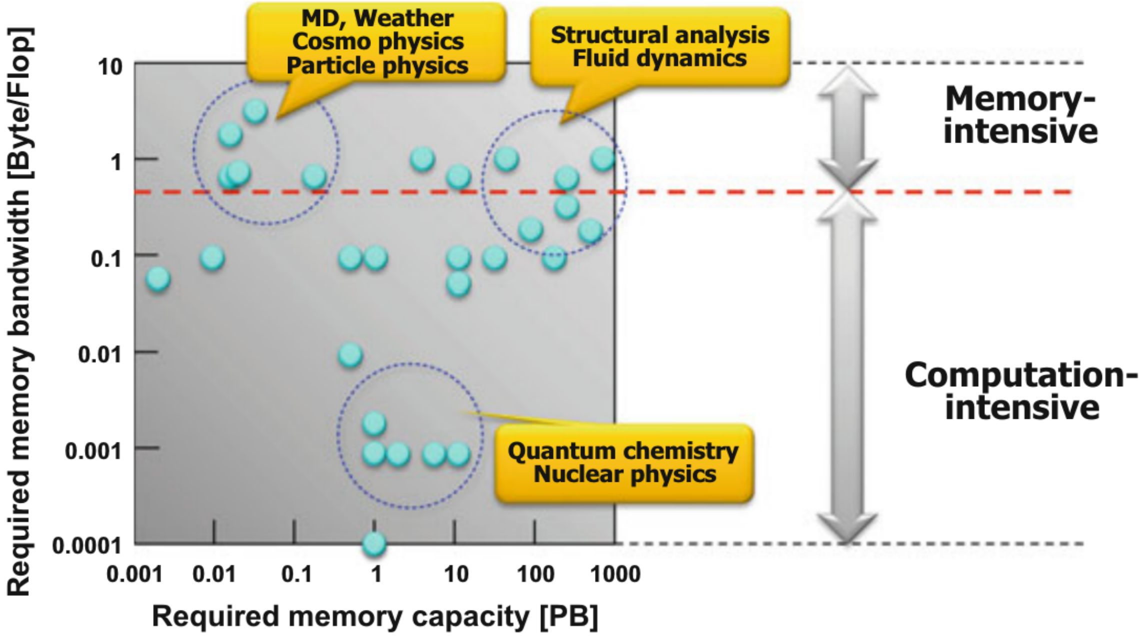
Data analytics: deep learning , machine learning..

And more

HPC and HPDA

HPC architectures today are compute-centric (FLOPS vs. IOPS)

They are not ready for I/O Intensive and memory intensive



Move code to data

Data moving is expensive, not only in time but also as energy consuption:

- Computing 1 calculation $\approx$ 1 picojoule
- Moving 1 calculation = up to 100 picojoules

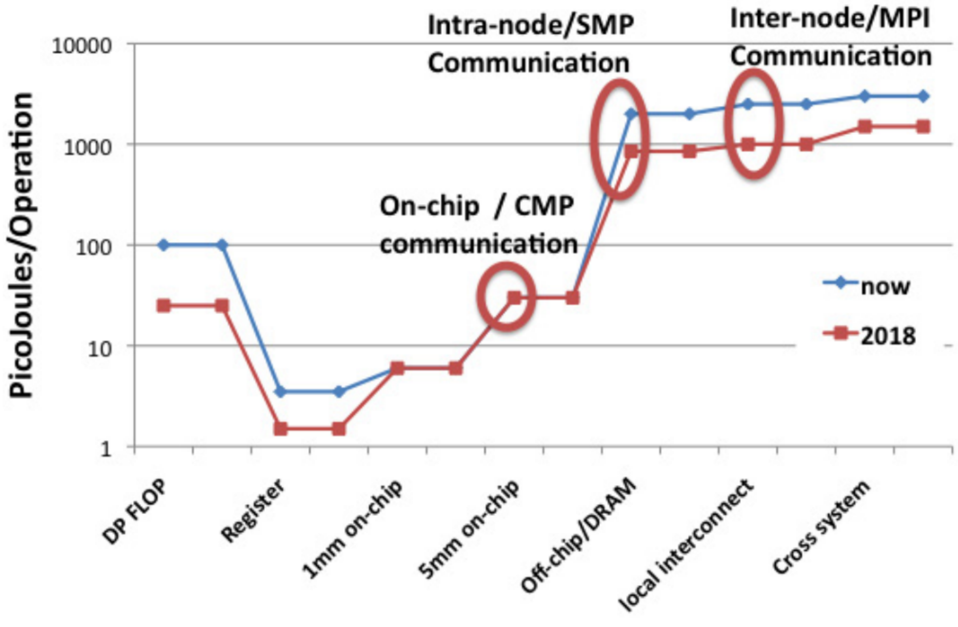
Strategy:

- Accelerate data movement at large and small scales (internet and intra-cluster): large bandwidth, photonic interconnect.
- Minimize data movements

Minimize data movement

Move your code close to the data.

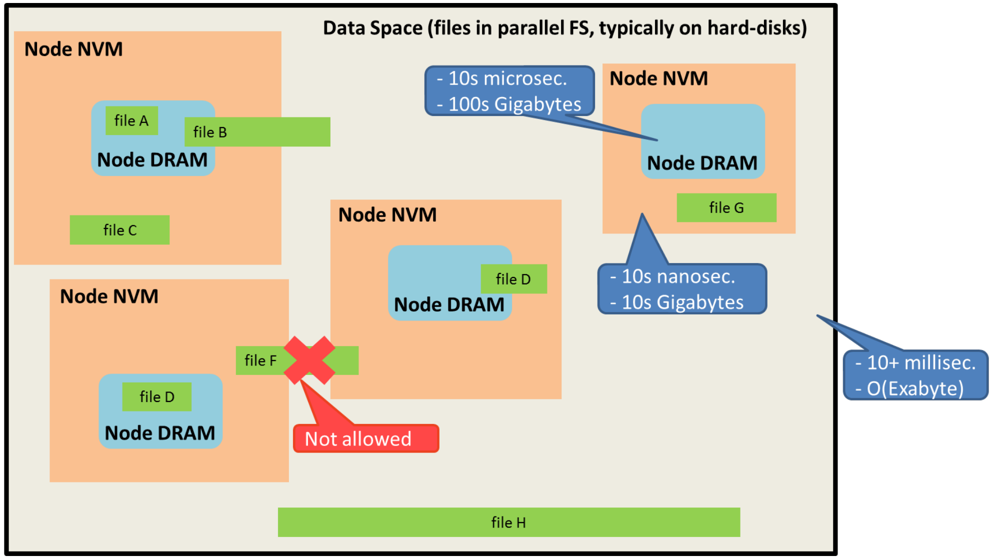
It may be not sufficient



In-memory processing

Technical Solutions

New storage systems based on Non-Volatile-Ram and tiered architectures.



Software technologies: NoSQL databases, Hadoop and MapReduce, COMPS, OMPSS, JUPITER, etc. These technologies form the core of open source software that supports the processing of large data sets across clustered systems.

How can we move code to data?

Where is my data: Peta and Exa-scale supercomputers are Tier-0 platforms.

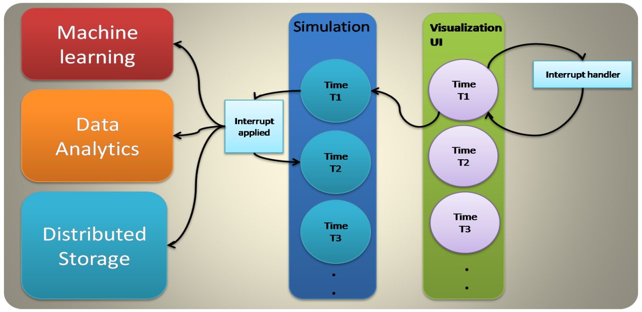
How can I run my code on it: new trends are moving towards remote interactive computing.

- e.g. Hadoop on Lustre or Beegfs
- COMP (write your app in sequential paradigm but runs in parallel) and pyCOMP+Jupiter

Real-Time in situ visualization: the use of GPUs with 3D real time visualization software. Automatic software interrupts to applications

Containers: less that 3% performance degrade

but… what happen with the I/O?



HPDA and the VO

How can we “integrate” an HPDA facility in the VO?

Data intensive computing requires a HPC PFS ==> HPC VOSpace?

Can we use standard UWS approaches?

Where is the astronomical data? (e.g. ASKAP and Lofar)