Bridging the gap between cold and hot storage at scale

Valentin HONORÉ

ensIIE & Samovar BENAGIL

May 23, 2025

PER3S Workshop

Table of Contents

Numerical simulations in HPC...

Scientific applications based on numerical simulation

- Replace real experiments:

  costly, dangerous, impossible...

- Weather forecast, physics, avionics etc.
- **Industry & Academic**
- Intense simulation programs
  - **compute-intensive :** long executions (days, weeks)
  - **memory-intensive :** compute & storage

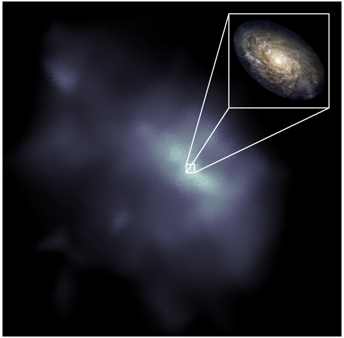


Figure: Dark Matter Halo

Source : CPAC group, ANL

… that are executed on supercomputers

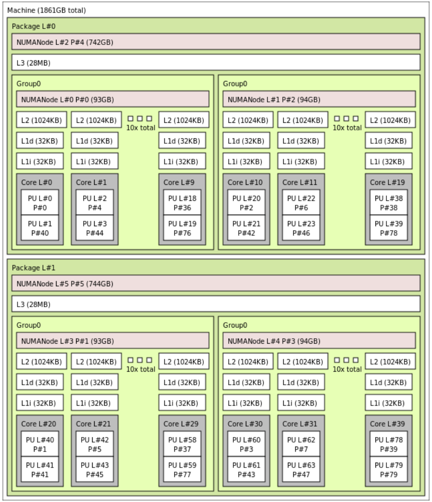


Figure: 2x Xeon CascadeLake 6230 with NVDIMMs (hwloc v2.1) Source : hwloc

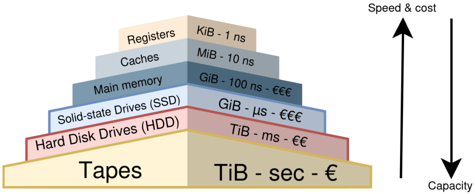


Figure: Memory hierarchy

Complex hardware and software

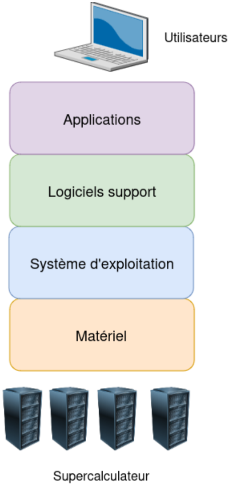


Figure: Software stack of a supercomputer

- Linux = 25 000 000 code lines
- **Overall :**
  - Hardware heterogeneity
  - Very complex software stacks

- On-demand computing resources through network
- Flexibility, elasticity, disaggregation
- Computing & storage

GPU Instances

Our comprehensive lineup of NVIDIA GPUs, including P100 and H100, covers a wide range of computing needs.

Name	vCPUs	GPU	RAM	Disks	Bandwidth	Price	Approx. per month
GPU - 3070	8 vCPUs	1 GPU	16 GB	Block Storage	2.5 Gbps	€0.98 /HOUR	€715 /MONTH
RENDER-S	10 vCPUs	1 GPU	42 GB	NVMe Local Storage or Block Storage on demand	1 Gbps	€1.24 /HOUR	€891 /MONTH
L4-1-24G	8 vCPUs	1 GPU	48 GB	Block Storage	2.5 Gbps	€0.75 /HOUR	€547.5 /MONTH
L4-2-24G	16 vCPUs	2 GPUs	96 GB	Block Storage	5 Gbps	€1.5 /HOUR	€1,095 /MONTH
L4-4-24G	32 vCPUs	4 GPUs	192 GB	Block Storage	10 Gbps	€3 /HOUR	~€2,190 /MONTH
L4-8-24G	64 vCPUs	8 GPUs	384 GB	Block Storage	20 Gbps	€6 /HOUR	~€4,380 /MONTH
L40S-1-48G	8 vCPUs	1 GPU	96 GB	Block Storage	2.5 Gbps	€1.4 /HOUR	~€1,022 /MONTH
L40S-2-48G	16 vCPUs	2 GPUs	192 GB	Block Storage	5 Gbps	€2.8 /HOUR	~€2,044 /MONTH

(Source : Scaleway)

Table of Contents

Data are getting more and more important

- LHC experiment at CERN (Source : CERN)
  - 30 PB of data per year
    - ≡ 1.2 million Blu-ray
    - ≡ 250 years of HD video
  - 100 PB permanently archived

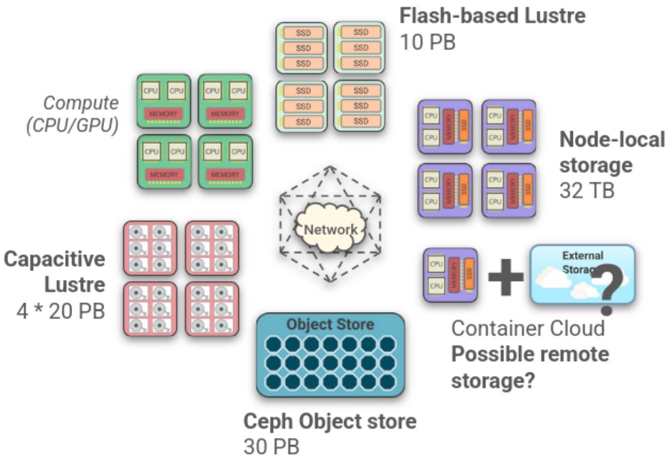
- Cloud services, with free offer (ex: gmail)
  - 15 GB of free storage
  - 1.8 billion users
  - **Free offer:** 27 EB
    - ≡ $>$ 1 billion Blu-ray
    - ≡ 225 000 years of HD video
  - *Youtube* : +30 000 years of HD video per year

Bilan : tsunami of data, how to store under **limited budget**?

Distributed storage in HPC

Two main domains

- **Hot storage** : close to computing resources
- **Cold storage** : distant (network)



Distributed storage in HPC

Two main domains

- **Hot storage** : close to computing resources
- **Cold storage** : distant (network)

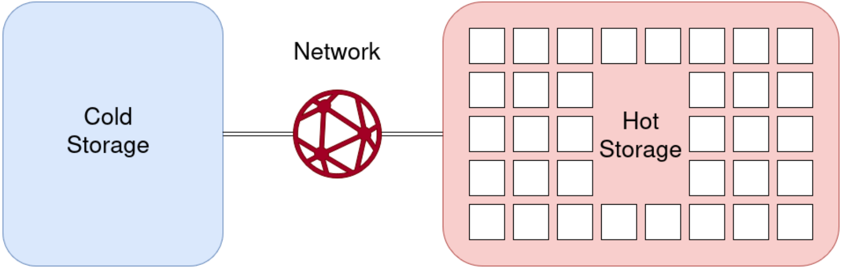


Figure: Simplified view of HPC storage

Two main domains

- **Hot storage** : close to computing resources
- **Cold storage** : distant (network)

No integration of cold storage with the operating system ☹

- Need of a **continuum**: computing $\leftrightarrow$ cold storage
- Hot storage widely studied, not the case for cold storage
- Cold storage integration = industrial priority![1]

1. "Electricity demand for data centers set to more than double by 2030"
   Source : International Energy Agency

Cold storage reference: magnetic tapes

$\approx$ 20+ TB on 1000×1km, read at 10m/sec ; 100s MB/sec

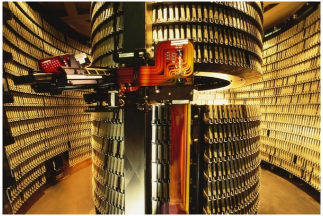
Source: Wikipedia

Inevitable for mass storage
ex : CC-IN2P3, CERN, ECMWF, Scaleway, CEA $\cdots$

☺ **Technology with many fors** (see next slide)

☹ ... and some cons : data access ($\sim$ 10s of seconds)

Adapted pour "*Write Once Read Many*"

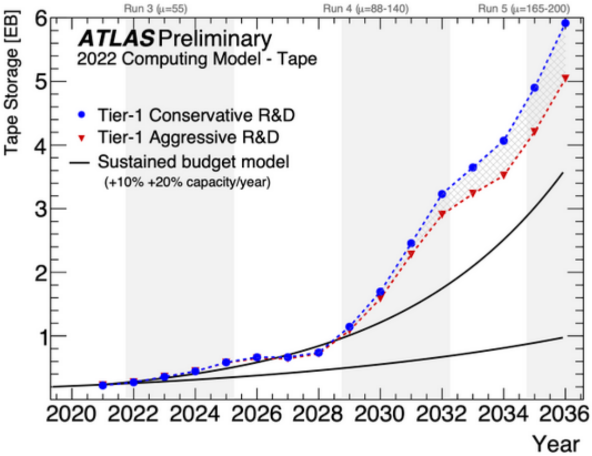


Source: Peter GINTER

Why to use tapes? (Source : CERN)

- Writing error rate per bit
  - LTO-9 (10^{-20}) is 10 000× more reliable than a 18 TB HDD (10^{-15})
- Hardware failures at CERN
  - 1% disks vs. 0.005% tapes
- Separation between data and driver
  - No data loss if the driver fails
- **Lifetime** (30+ years)
- **Energy**
  - 10× less energy than disks at equal storage amount
  - ↗ more capacity **without** extra energy

Example: tape storage for ATLAS (CERN)



Projected storage needs for ATLAS Tier-1 sites

(Source : CERN)

Table of Contents

Bring the tapes closer to the operating system

Positioning :

- Strong operational advantages ☺ but complex & unknown mechanical processes ☹
- **Objectives:**
  - continuum hot ↔ cold
  - Rethink the way data are handled and stored (hot + cold)

Proposal: bridge the gap between tapes and computing resources

1. Performance analysis of magnetic tapes
2. Integration of tapes in the OS

Related work

- (Very) few performance studies on tape systems
- Theoretical aspects already treated [ICAPS'22]
- Tapes = brick of the I/O stack for long-term storage

Bring the tapes closer to the operating system

Positioning :

- Strong operational advantages ☺ but complex & unknown mechanical processes ☹
- **Objectives:**
  - continuum hot ↔ cold
  - Rethink the way data are handled and stored (hot + cold)

Proposal: bridge the gap between tapes and computing resources

1. Performance analysis of magnetic tapes
2. Integration of tapes in the OS

Related work

- (Very) few performance studies on tape systems
- Theoretical aspects already treated [ICAPS'22]
- Tapes = brick of the I/O stack for long-term storage

Bring the tapes closer to the operating system

Positioning :

- Strong operational advantages ☺ but complex & unknown mechanical processes ☹
- **Objectives:**
  - continuum hot ↔ cold
  - Rethink the way data are handled and stored (hot + cold)

Proposal: bridge the gap between tapes and computing resources

1. Performance analysis of magnetic tapes
2. Integration of tapes in the OS

Related work

- (Very) few performance studies on tape systems
- Theoretical aspects already treated [ICAPS'22]
- Tapes = brick of the I/O stack for long-term storage

1. Performance analysis of tape storage systems

Goal : Better performance on tape storage

- Metric extraction
- Properties of internal processes (reading head acceleration, throughput etc.)
- Propose new firmware and/or sotfware APIs

Requirements:

- Hardware setup for performance evaluation **[MISSING]**

Expected results

- Performance evaluation datasets & data analysis
- APIs for efficient tape operations

1. Performance analysis of tape storage systems

Goal : Better performance on tape storage

- Metric extraction
- Properties of internal processes (reading head acceleration, throughput etc.)
- Propose new firmware and/or sotfware APIs

Requirements:

- Hardware setup for performance evaluation **[MISSING]**

Expected results

- Performance evaluation datasets & data analysis
- APIs for efficient tape operations

2. Integration of tapes in the OS

Objective : Software support of tapes in the OS

- Use the previous APIs
- Proactively optimize and reorder the requests to the tape device
  - data-packing on an application-based, pre-fetching, data-buffering, etc.
- New data structures to optimize the storage of I/O data on tapes

Requirements:

- Hardware setup for performance evaluation **[MISSING]**
- Efficient tape operation APIS **[MISSING]**

Expected results

- Piece of software smartly connecting tape drive's firmware and OS
- Evaluation setup

2. Integration of tapes in the OS

Objective : Software support of tapes in the OS

- Use the previous APIs
- Proactively optimize and reorder the requests to the tape device
  - data-packing on an application-based, pre-fetching, data-buffering, etc.
- New data structures to optimize the storage of I/O data on tapes

Requirements:

- Hardware setup for performance evaluation **[MISSING]**
- Efficient tape operation APIS **[MISSING]**

Expected results

- Piece of software smartly connecting tape drive's firmware and OS
- Evaluation setup

A very collaborative roadmap

Strong potential for industrial collaborations

- Need of real use-cases & guidance
- Joint project BPI/Inria "HyperScaleway"

Ongoing collaboration with CEA

- ANR project submission with P. Deniel
- Objective: hardware acquisition & initiate the work on the two axis

How to scale to production challenges?

- By working with an industrial
  - that is willing to share, study & modify current strategies
- Focus on a demonstrator (scientific app, user profile etc)
- ECMWF? Scaleway? Atempo? Come discuss with us!

Thank you for your attention!