DE LA RECHERCHE À L'INDUSTRIE

Next Generation IO @ CEA Computing Centres

J-Ch Lafoucriere

ORAP Forum #39 | 2017-03-28

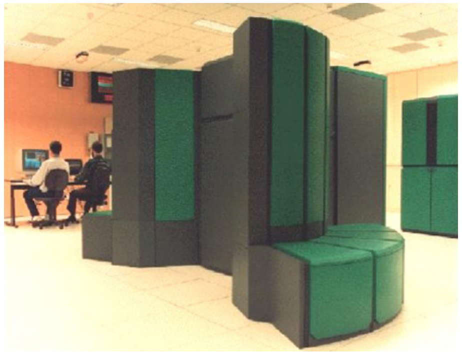
www.cea.fr

A long History of Storage Architectures

Last Century

Compute Systems

- Few Cray Supercomputers (vectors and MPP)
- Few front-end machines



YMP

Storage Systems

- Directly connected to the front-end or to the super-computer
- Data managed through HSM

T3E

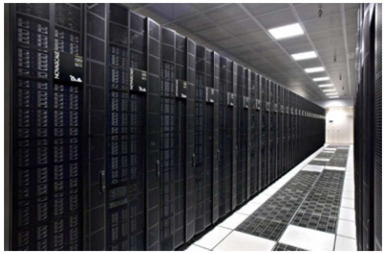
T90

Early 2000 to 2016

Homogeneous Cluster

TERA1

- Tera 1
  - Fast Local Parallel FS
  - Fast Shared Storage on striped tapes in HSM
  - Capacity Storage on large tapes



TERA10

- Tera 10
  - Fast Local Parallel FS (OpenSource)
  - Fast Shared Storage on striped disks in HSM
  - Capacity Storage on large tapes in HSM
- Curie or Tera 100
  - Fast Local Parallel FS (OpenSource)
  - Fast Shared Storage on Parallel FS (OpenSource)
  - Capacity Storage on large tapes in HSM

Curie

TERA100

Today Architecture

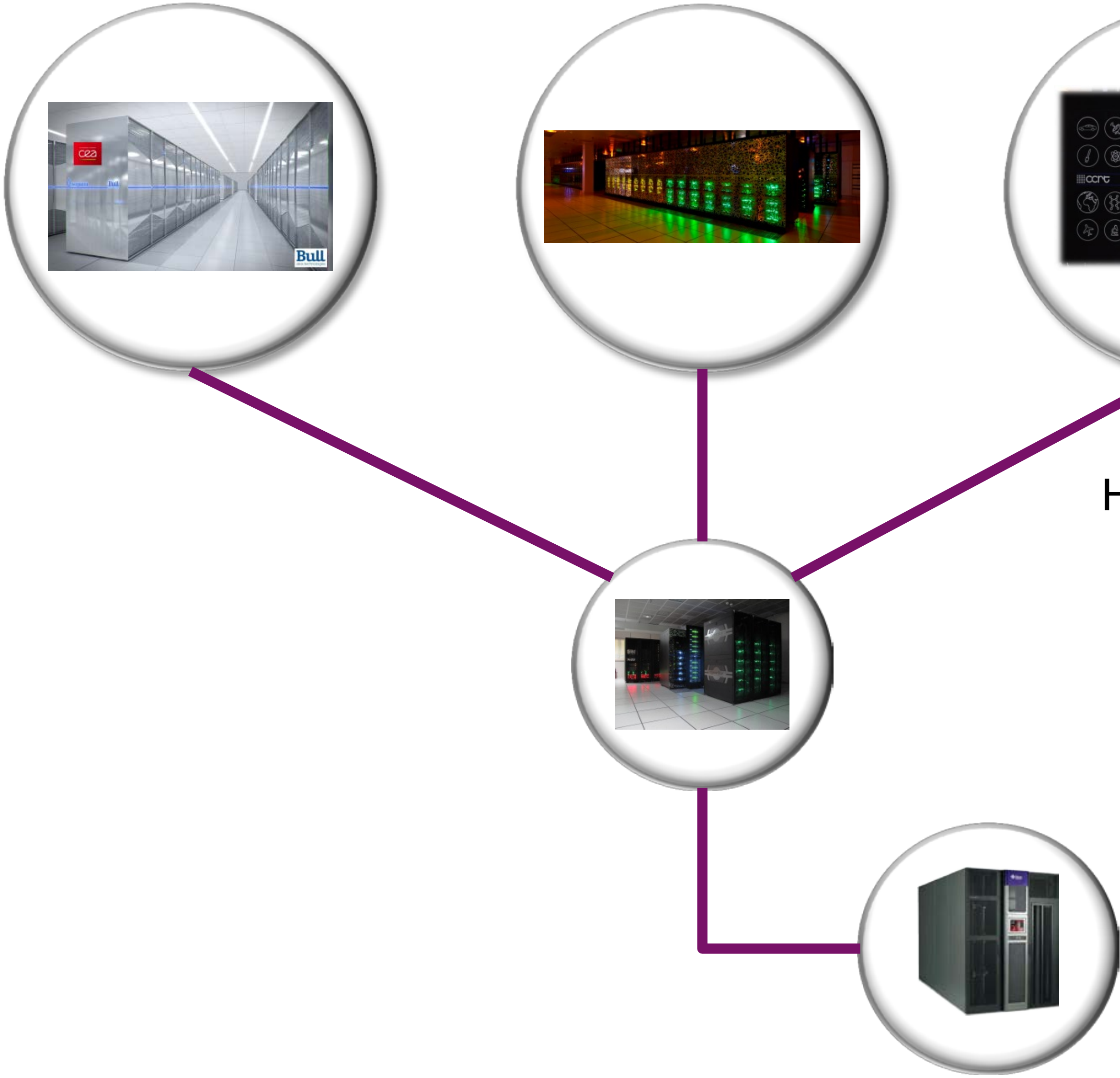
Heterogeneous Data Less Cluster: Tera 1000, Cobalt

- Multi-usage clusters: Simulation and Data Analysis
- Heterogeneous Compute resources
  - Xeon, Xeon Phi, GPU
- Data Less Clusters
  - Fast Remote Dedicated Parallel FS (OpenSource)
  - Fast Shared Storage on Parallel FS (OpenSource)
  - Capacity Storage on large tapes in HSM

TERA1000

Cobalt

Today Architecture



Heterogeneous
Compute/Analysis
Resources

High Performance Network

Heterogeneous
Storage
Resources

2020 Evolution

Exascale challenges

Exascale supercomputer will rely on new architecture

- High level of different parallelism: vectorization, threads, nodes
- High number of cores/proc => Less memory per core
- High speed network will be available

Consequences for storage

- Storage clients memory fingerprint should be reduced as much as possible
- How will behave Petascale FS?

IO Architecture: Proxy based

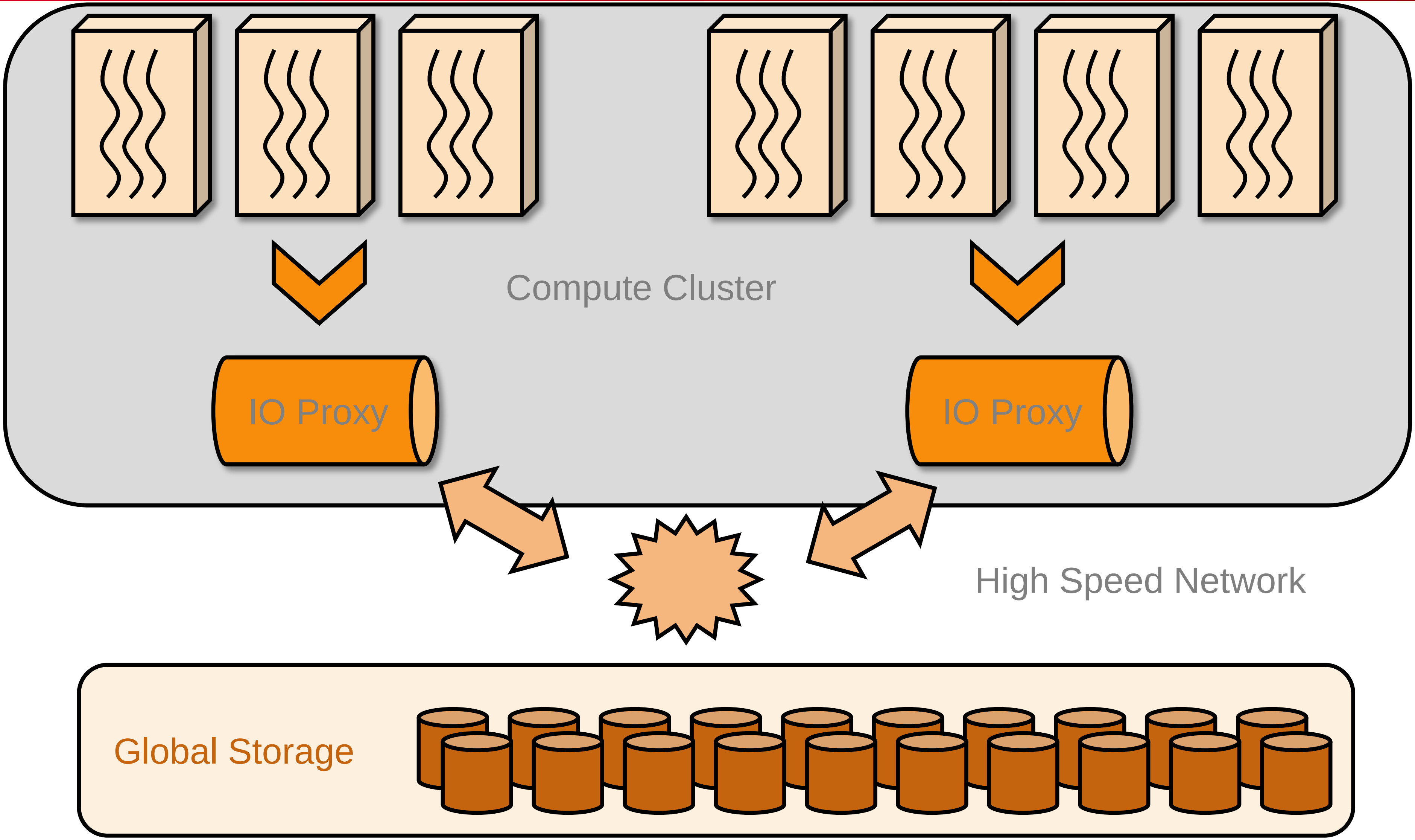
Petascale FS may not scale at millions of cores

- Client memory may be too small
- Client parallelism may be too high
- Distributed Locking may not scale

Solution: CEA IO Proxy

- Posix Compliant through client kernel access
  - Clients delegate IO calls to a job dedicated IO proxy (9P/RDMA protocol)
- Native access from User Space
  - Code or IO library calls 9P user space library
- Advantages
  - Require only a Petascale FS
  - Isolate jobs
  - Optimize IO by a shared cache effect on IO proxy node

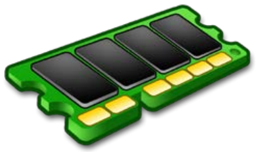
IO Proxy Architecture



Storage Evolution

Hardware Technologies

Flash memory



- Cheaper but still expensive
- Available in multiple form factors
  - DIMM
  - Disks
  - Network device (NVMe over Fabrics)
- Usage
  - As memory through pmemio API (efficient but not easy to use)
  - As block device

Object Storage Cluster

- Mero from Seagate
  - See SAGE talk
- WOS from DDN
- Good proprietary products

Object Storage Device

First Experience: Seagate Kinetic

- Idea: connect a disk with an Ethernet Interface
- Use a Key/Value store interface over tcp/ip
- Nice idea but
  - Not Open
  - KV interface too limited

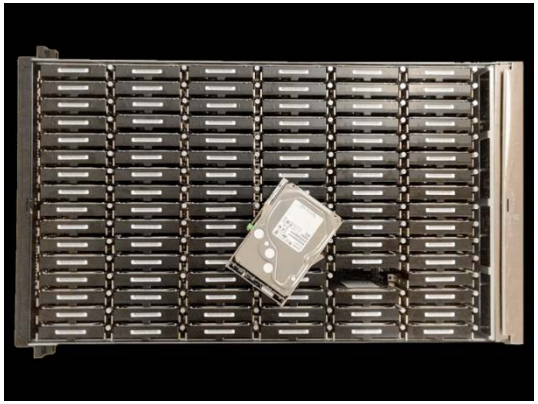
Open Approach

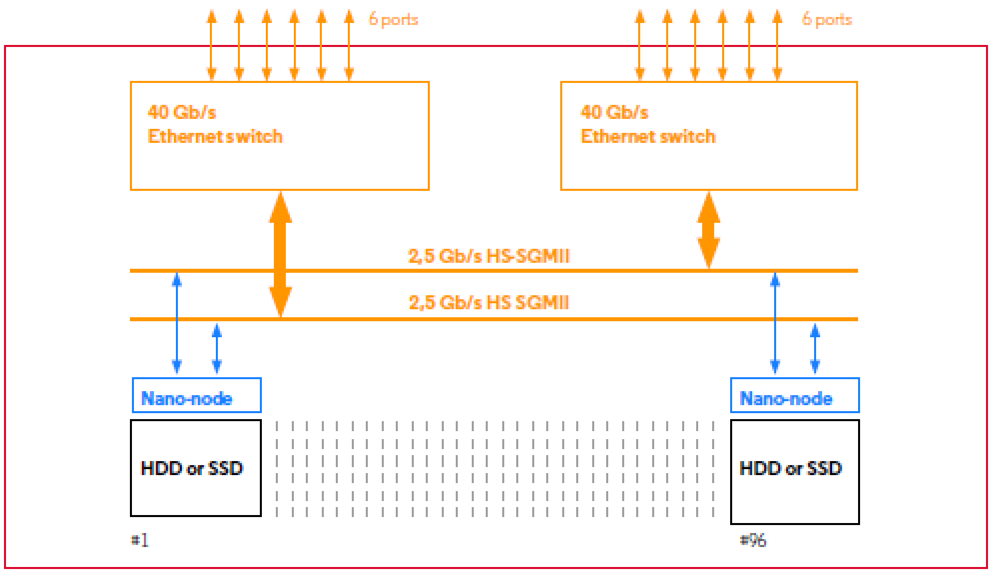
- Idea: connect an ARM interposer to a standard disk
- Product are already available: OpenIO Appliance

OpenIO Appliance (SLS-4U96)

Object Storage Appliance

- Easy to use: Ethernet interface (2.5 Gb/disk)
  - Plug and use
- Scalable Architecture
  - Nano-Node ARM interposer
- Open Source software





Object Storage Challenges

New Interfaces and no standard

- Difficult to implement from applications
- Only Mero has a RDMA based interface

Still need for legacy access

- pNFS though Ganesha-nfs project
  - Libkvns to implement tree namespace over Key/Value Store
  - Native access to object for Data

Object Storage Access

- Which API to choose?
  - Define a CEA “STD” to hide to codes?
  - Done for KVS

Our Strategy

Hide object interfaces to user

- Proxy IO will be used

Hide storage hierarchy

- Develop tools to hide storage tiers to user: phobos project
- Define interfaces to give hints from applications

Conclusions

Storage architecture evolves to a proxy based architecture

- Prototype running today
- High scale tests planned in 2017 and 2018 on a large systems @ CEA

Storage building blocks will be object based

- Software based solution (SDS)
- In network appliance for high volume deployment
- With multiple type of storage organized in distributed hierarchies

Usage Model

- Through legacy low level interface (initially)
- Through native interface for high performance
  - Opportunity to use high level interfaces: MPI-IO, HDF5, …

Thank you

Questions?

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM-Ile de France | 91297 Bruyères-le-Châtel Cedex
T. +33 (0)1 69 26 40 00 | F. +33 (0)1 69 26 70 86

Direction des applications militaires
Département sciences de la simulation et de l'information