DE LA RECHERCHE À L'INDUSTRIE

Exascale I/O challenges

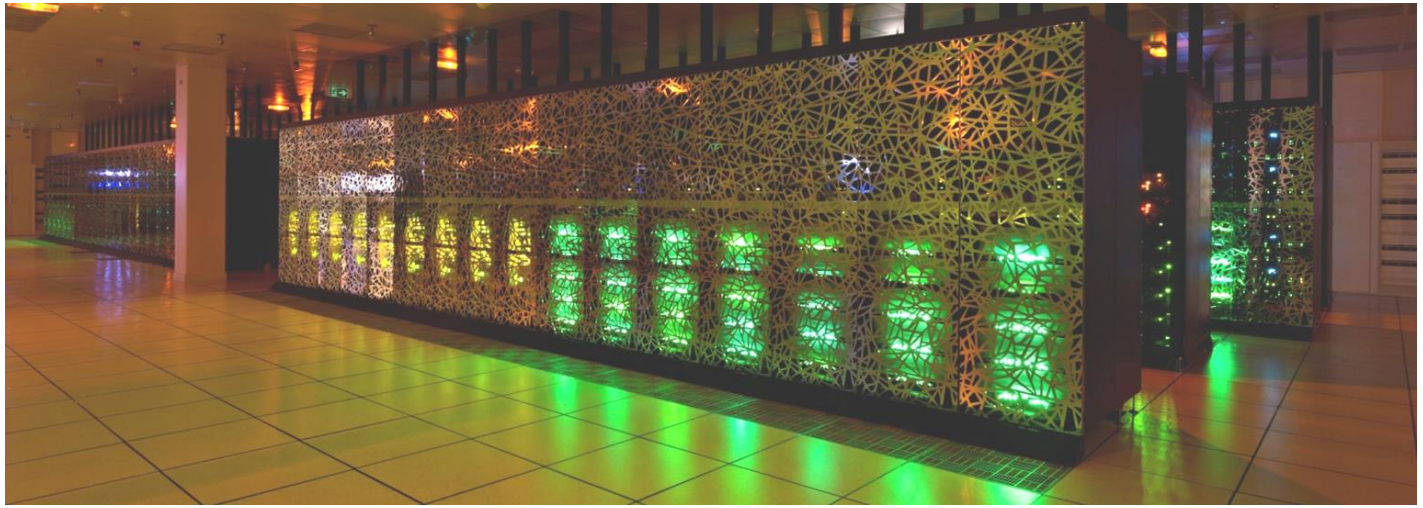
Jacques-Charles Lafoucriere
CEA

www.cea.fr

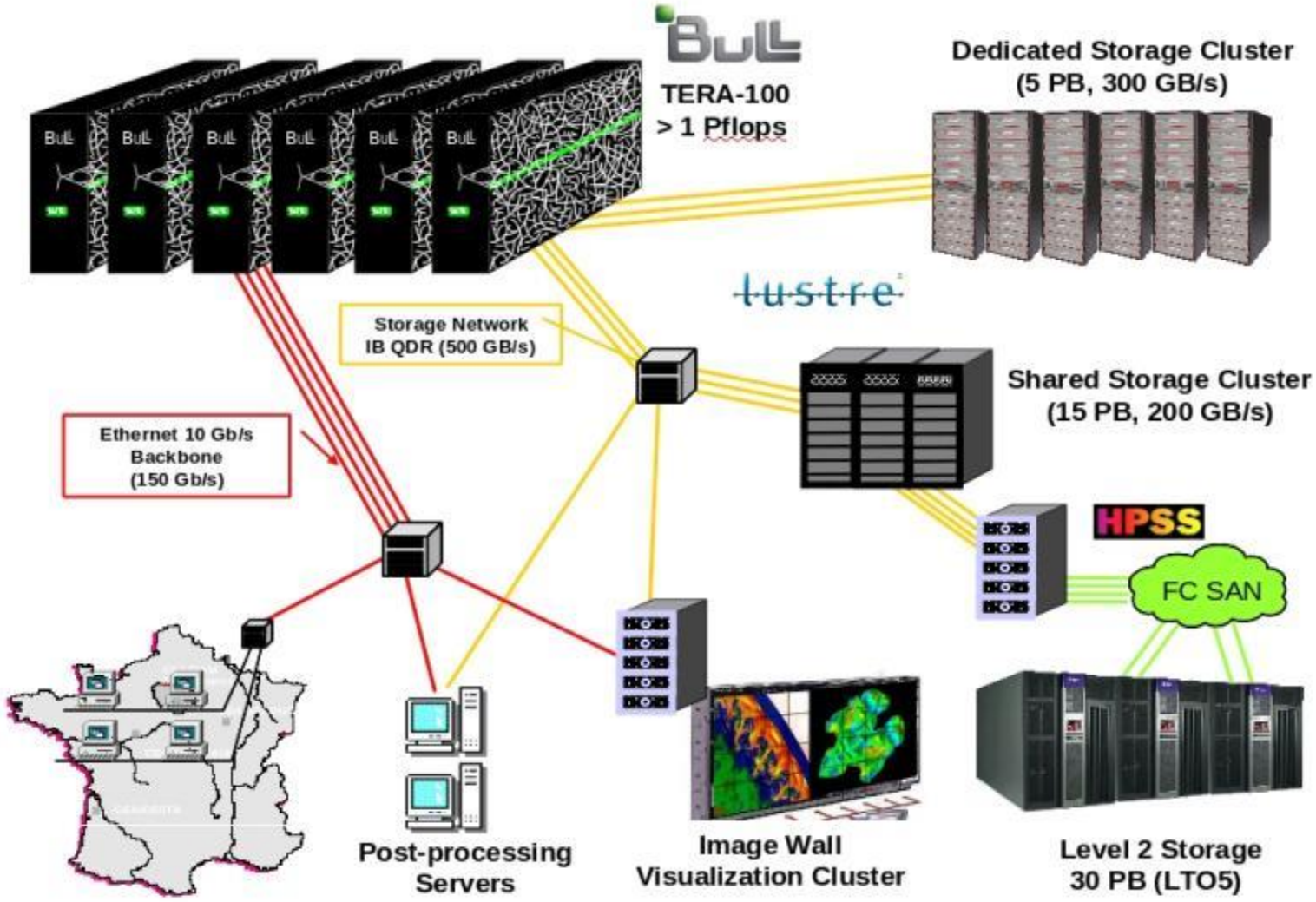
Petaflopic Computing Centers

Tera
1.25 Pflop/s
Mem: 290 TB
FS: 500 GB/s

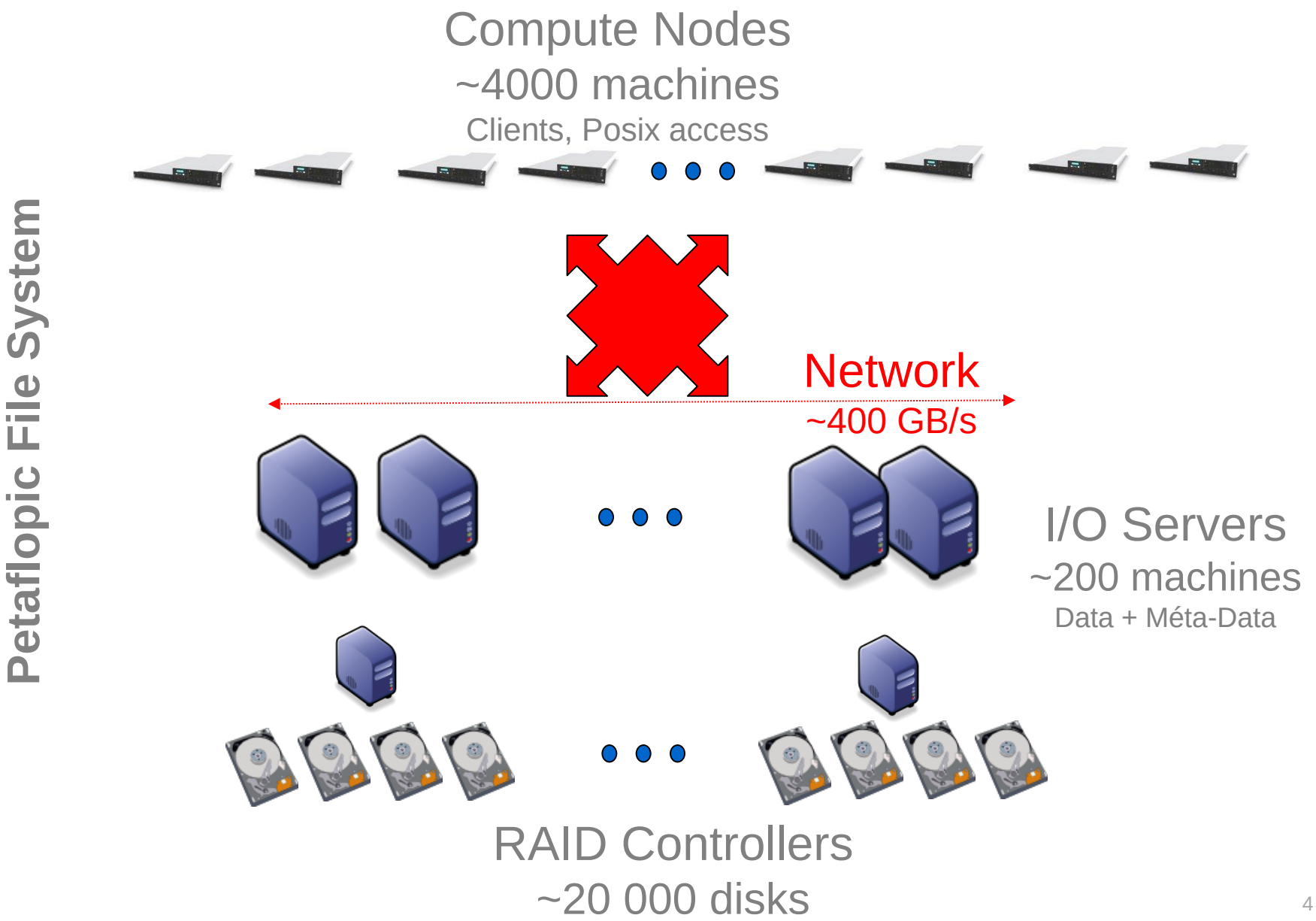
TGCC/Curie
2 Pflop/s
Mem: 340 TB
FS: 250 GB/s



Global Architecture



Storage for a Computing Center ?



Exaflop Context (2020)

HPC in 2020

Power consumption constraints drive to technological breakthrough

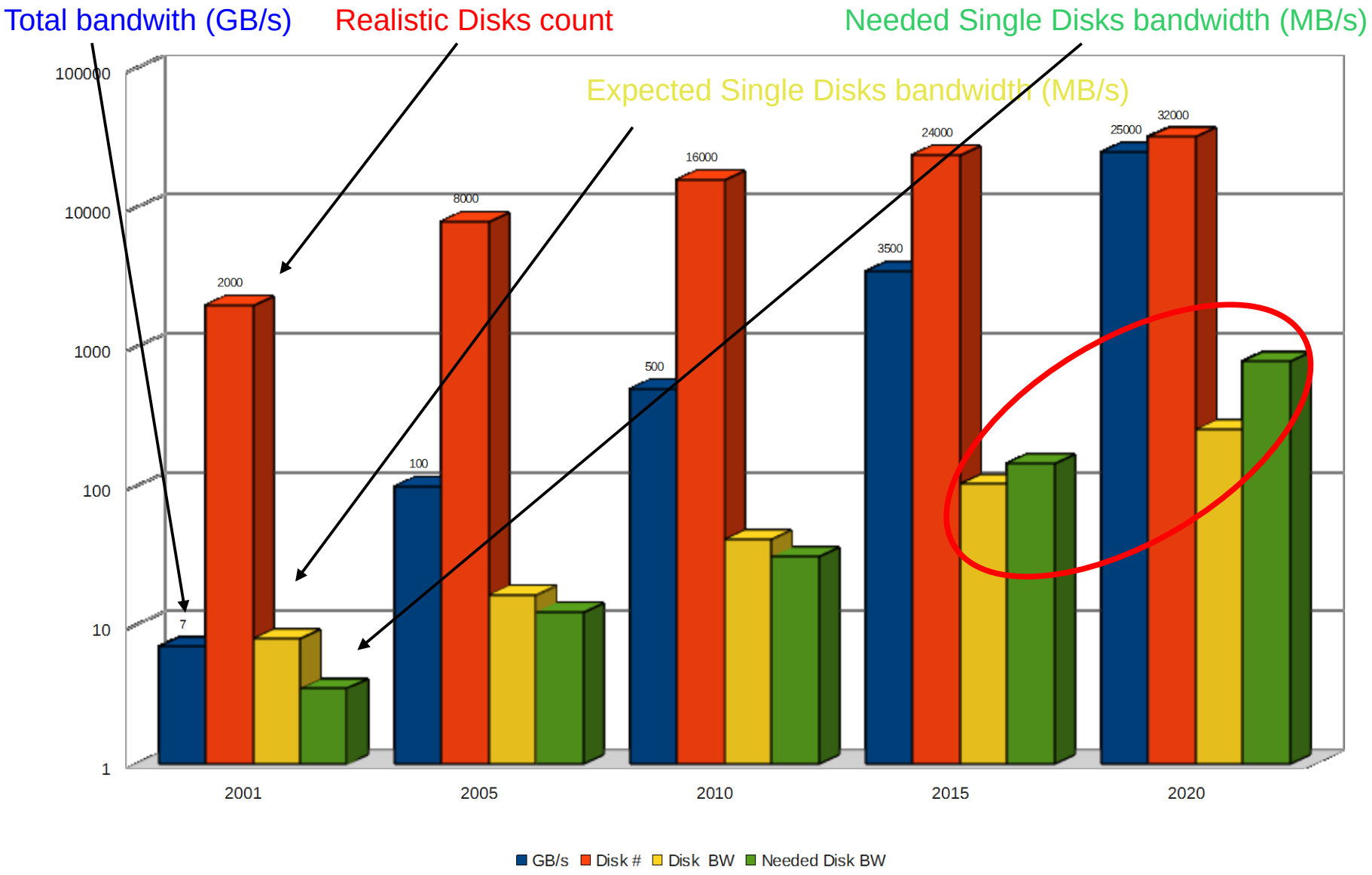
2010	2020 (STD)	2020
2 PF	1 EF	1 EF
5 MW	50 MW	20 MW
0.3 GF/W	20 GF/W	500 GF/W

New processors architectures are needed to get better Flop/Watt

- Increase parallelism
  - Vector computing
  - High number of compute threads
  - High number of compute nodes

- New programming model for applications
- Less memory available for I/O on compute nodes
- More concurrency when accessing files

Disks Trends



New Usage in Computing Center

HPC and Big Data Convergence

- Big Data applications need HPC class computer
  - Data from connected objects
  - Data from experiments
  - Data from human usage (and user them self)

- HPC centers need to evolve to support this new I/O pattern
  - Data come from external sources
  - Data cannot be recreated
  - Data are not really structured
  - “Small” random I/O
  - Intense meta-data use
  - New ratio Reads# vs Write#

2020 HPC needs

New storage building block

New storage architecture

New storage software stack

Which Solutions for HPC Storage in 2020?

Hardware

Rotating disk performance will be too slow

- Need to use a faster solution: flash based storage
  - PCIe or DIMM based to get best efficiency

Full path from client to storage has too many frontiers/hops

- Need to reduce number of interfaces between client and storage build block
  - Embedded model: run servers/application in RAID controllers

Block based architecture is too simple/low level

- Need a new architecture with a larger global view
  - Network object mode: file server becomes an object server

Clients Software

Compute Node

- Highly multi-threaded
- Few memory/thread

File system client

- Need memory for IO buffers

We have to introduce a mechanism to transfer IO to storage from compute nodes to an IO gateway

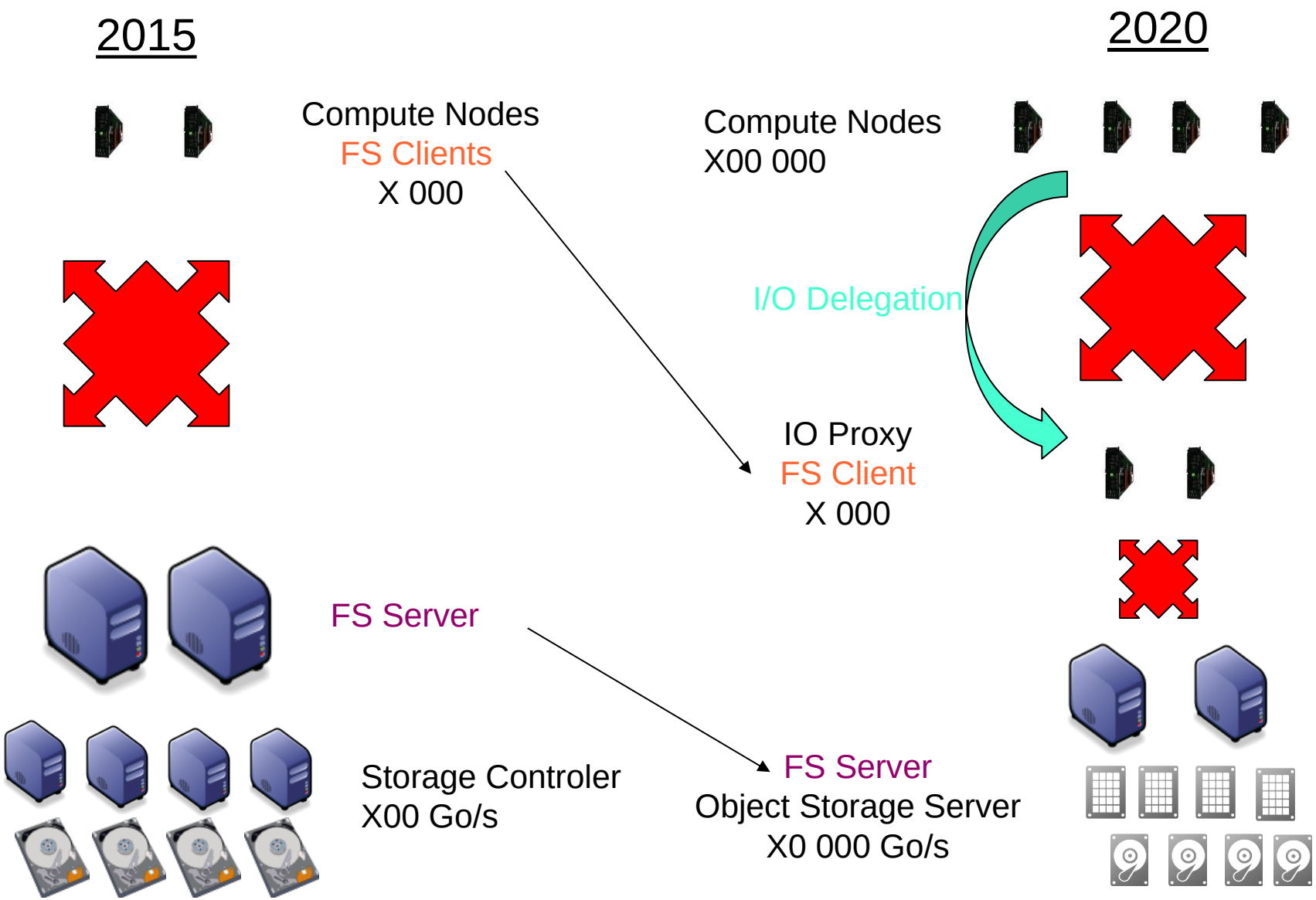
Dynamic allocation of IO gateways

Remote Direct Memory copy from compute node to limit memory use

2 tracks

- System: IO Proxy
- Applications: IO delegation

Future Architecture



Applications

New IO paradigm

- Constraints from Posix interface need to be removed
  - No more possible to offer a free/fast global coherency to applications
- Applications and resource manager need to provide help to storage (hints)
  - Topology, access mode, ...
  - Real data use knowledge is within applications
- Working groups have started discussions on new IO API for applications (expansion phase)
- **Applications will have to change their IO interfaces to get all the performances**

Conclusion

IO challenges for future storage systems

- Data and Meta-Data management
- Hints from “those who know”
- New IO servers architecture (Object storage and Embedded model)
- I/O delegations
- HPC and BigData convergence

Multiple ways already exists

Storage communities have started to work on solutions


Seagate Ember Chassis
(Currently Not for Sale)
Seagate

Questions?


Kinetic HDD
4TB


ONE STOP SYSTEMS
ONE STOP SYSTEMS
ONE STOP SYSTEMS

CEA | October 2015

Commissariat à l'énergie atomique et aux énergies alternatives
Direction des applications militaires
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

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019