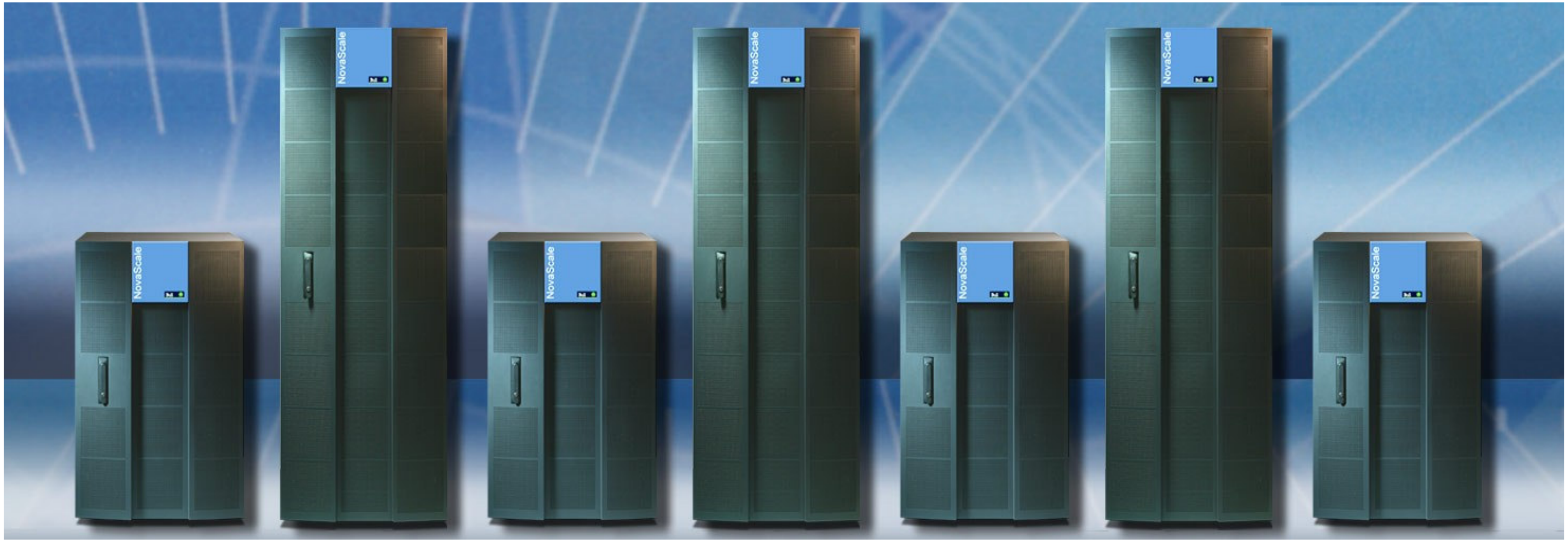


LUSTRE File System

From Clusters to Grid with 100GB/s sustainable bandwidth

jean-marc.denis@bull.net

Bull

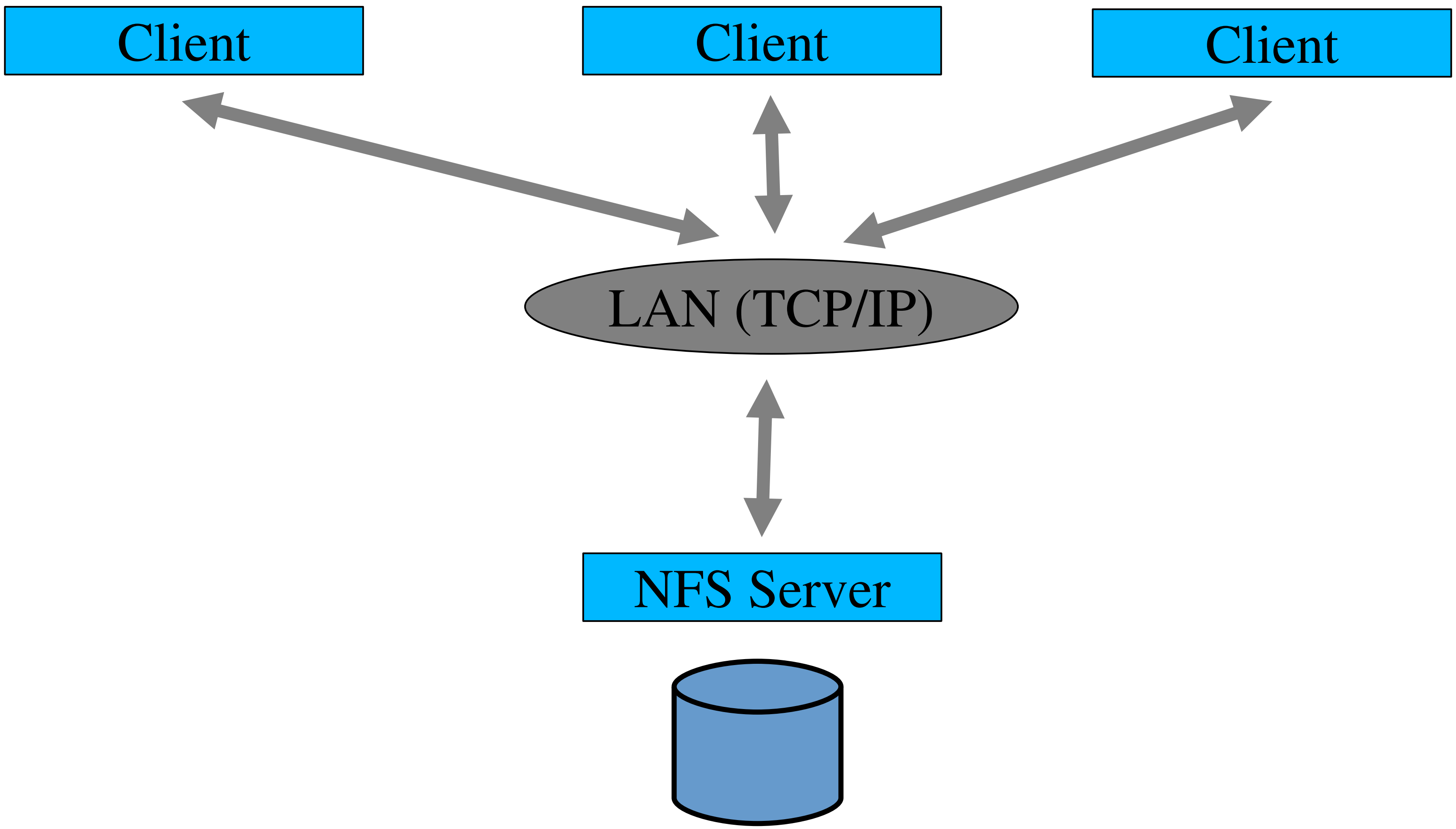
Agenda

- Concept
- Architecture
- Performances – optimizing Lustre
- TERA10 super computer: +100GB/s
- From superclusters to grid: issues
- conclusion

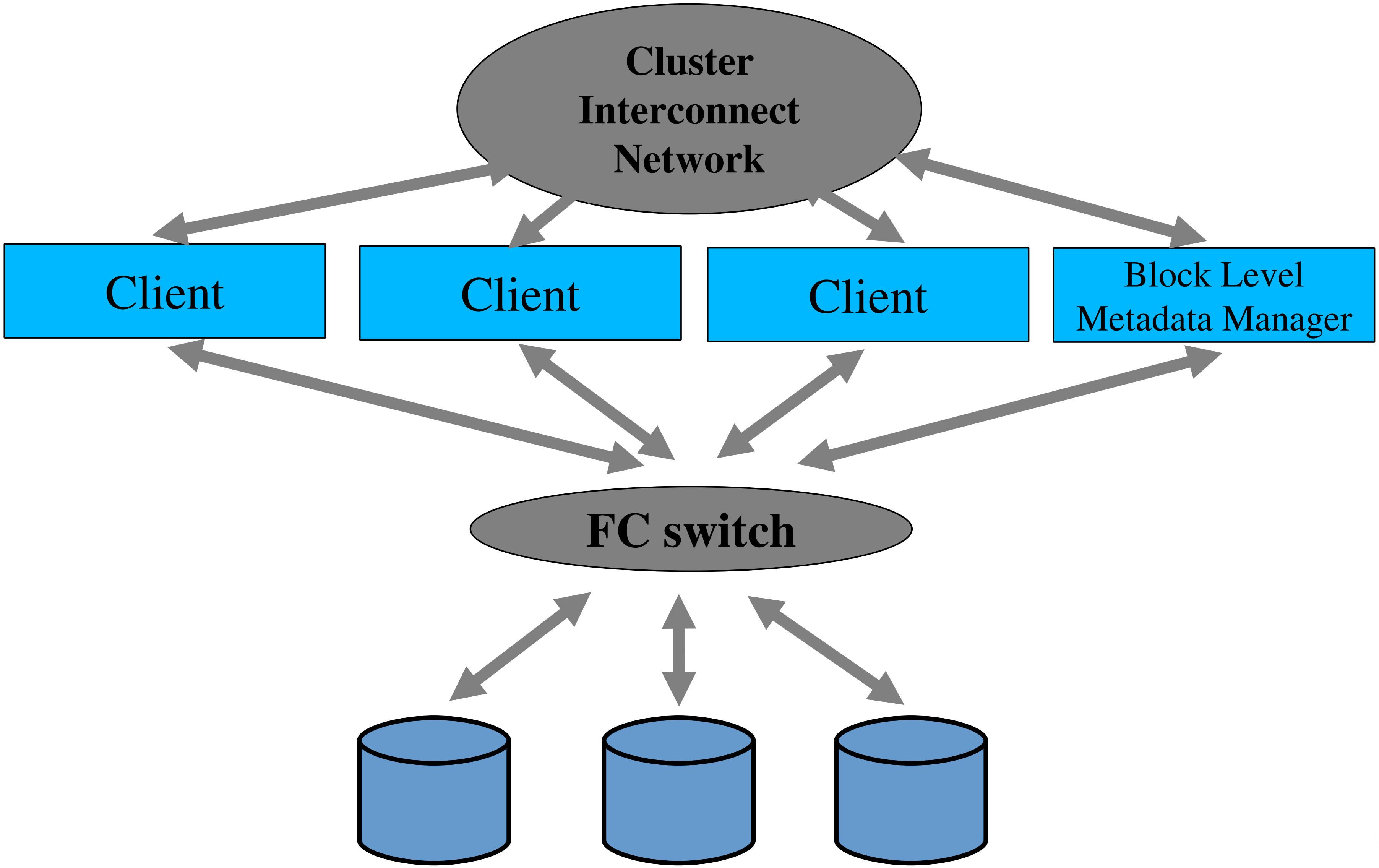
Agenda

- Concept
- Architecture
- Performances – optimizing Lustre
- TERA10 super computer: +100GB/s
- From superclusters to grid: issues
- conclusion

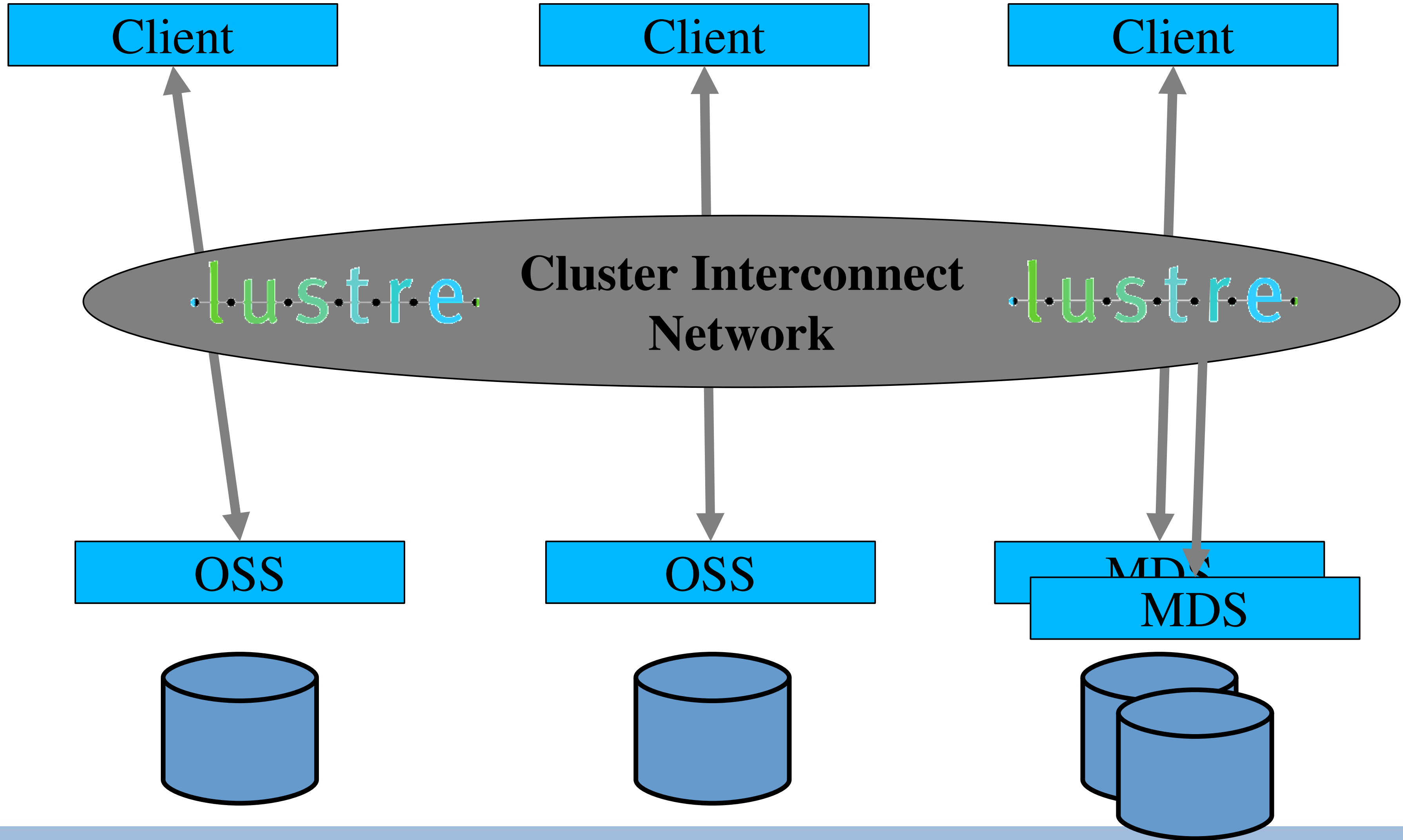
Network topology : NFS



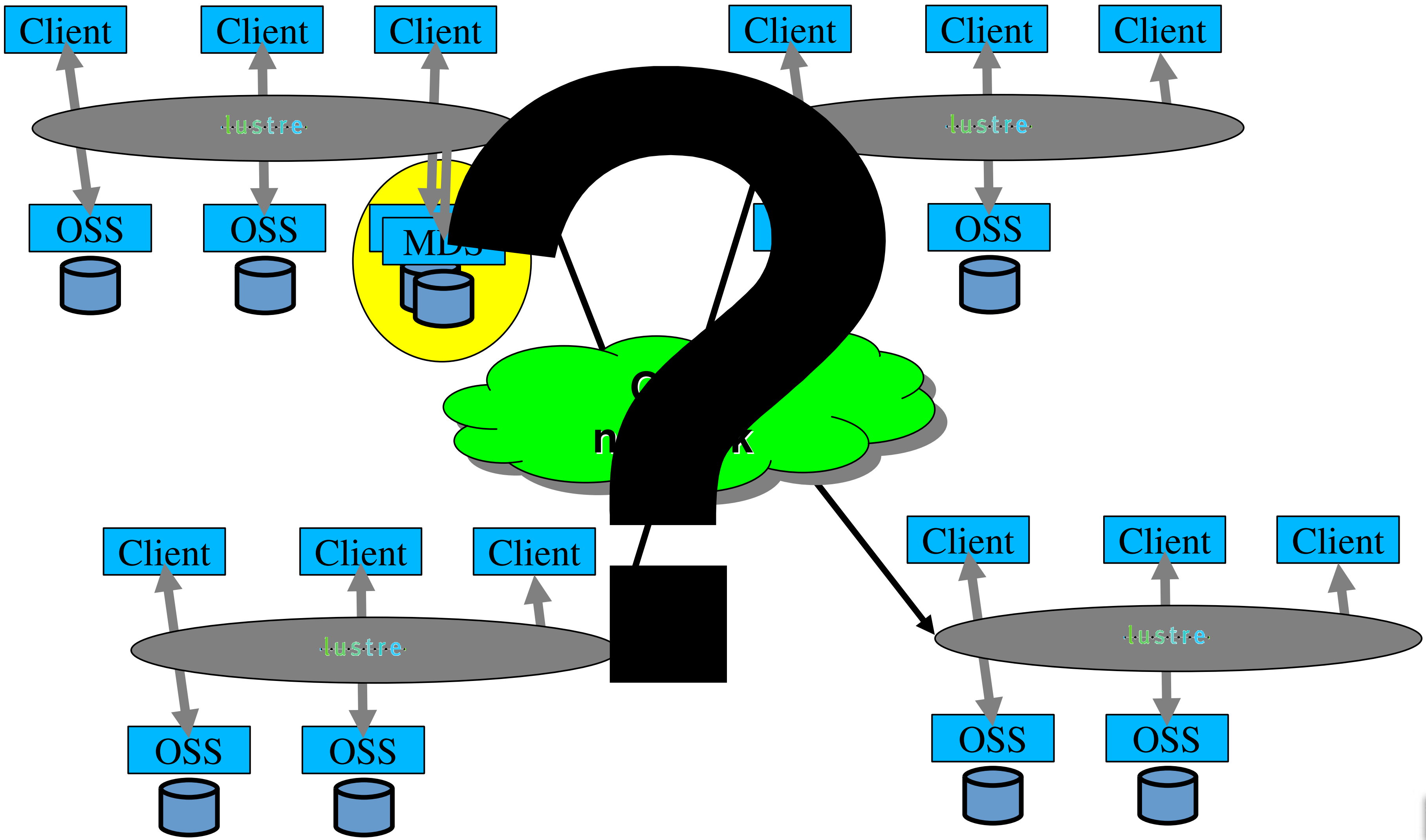
Shared Disks (GPFS, CXFS…)



Dedicated IO Servers (Lustre)



Lustre at the grid level



Agenda

- Concept
- Architecture
- Performances – optimizing Lustre
- TERA10 super computer: +100GB/s
- From superclusters to grid: issues
- conclusion

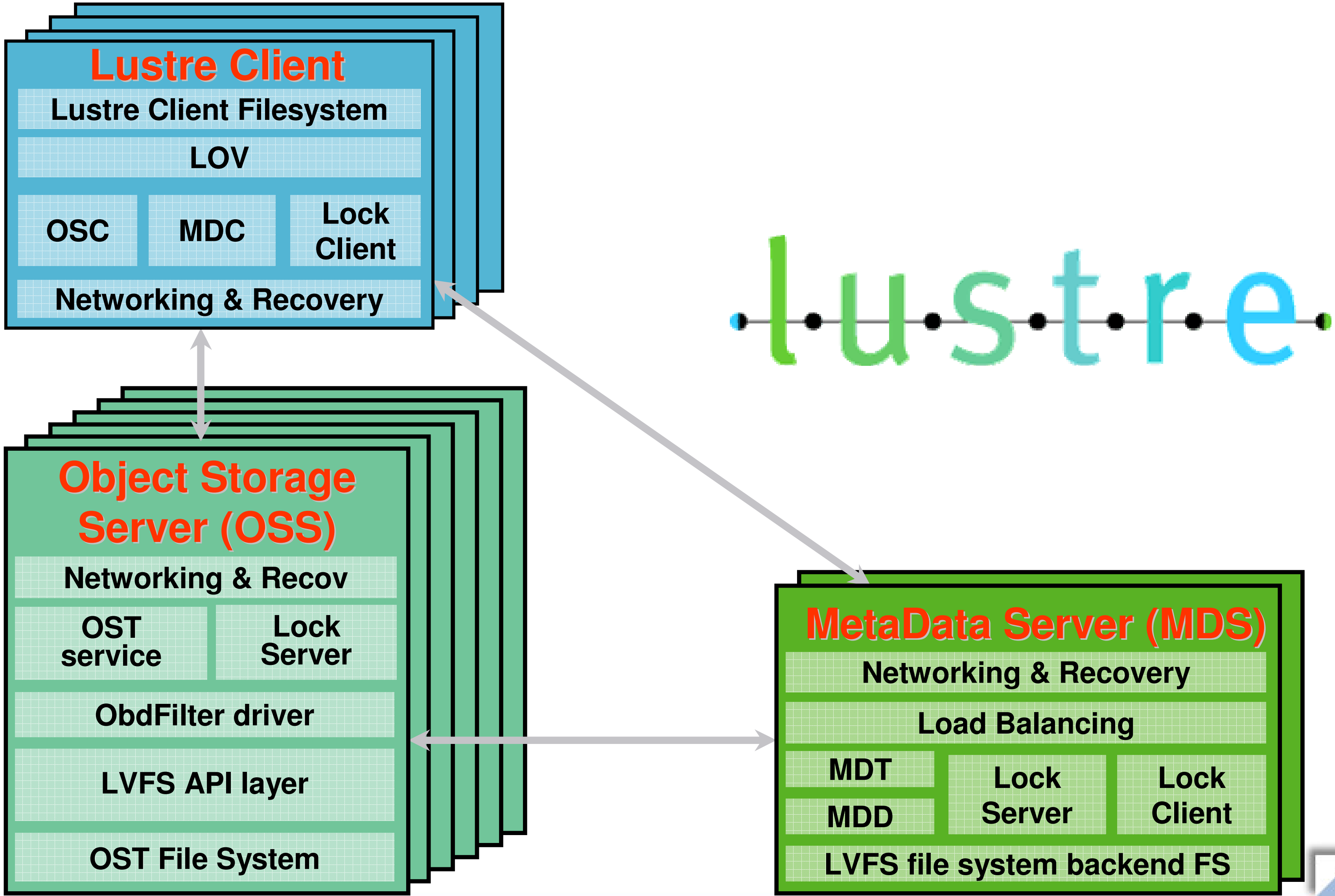
Lustre filesystem

- **Lustre is a "*NFS like*" filesystem, comparable to IBM-GPFS**
  - **Parallel**
  - **Scalable**
  - **HA (Bull development)**
  - **Opensource (GPL)**

- **Works in Linux Environment**
  - **New clients coming up in 2005 and 2006 (co-development between CFS and Bull)**

- **Wildly accepted by the largest compute centers**
  - **LANL / SANL / PNNL / LLNL**
  - **CEA**
  - **AWE**
  - **HLRS / TU Dresden**

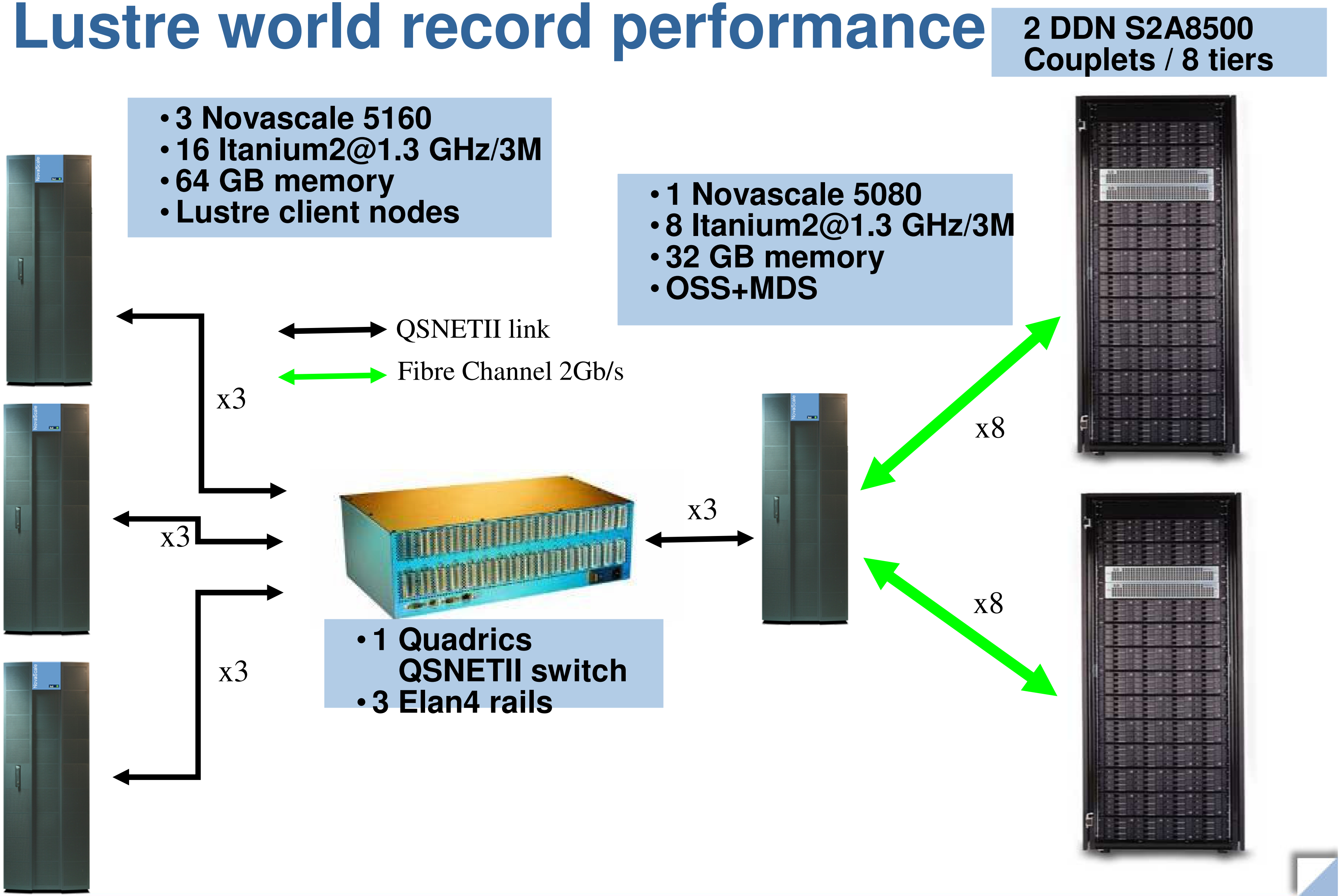
Lustre SW architecture



Agenda

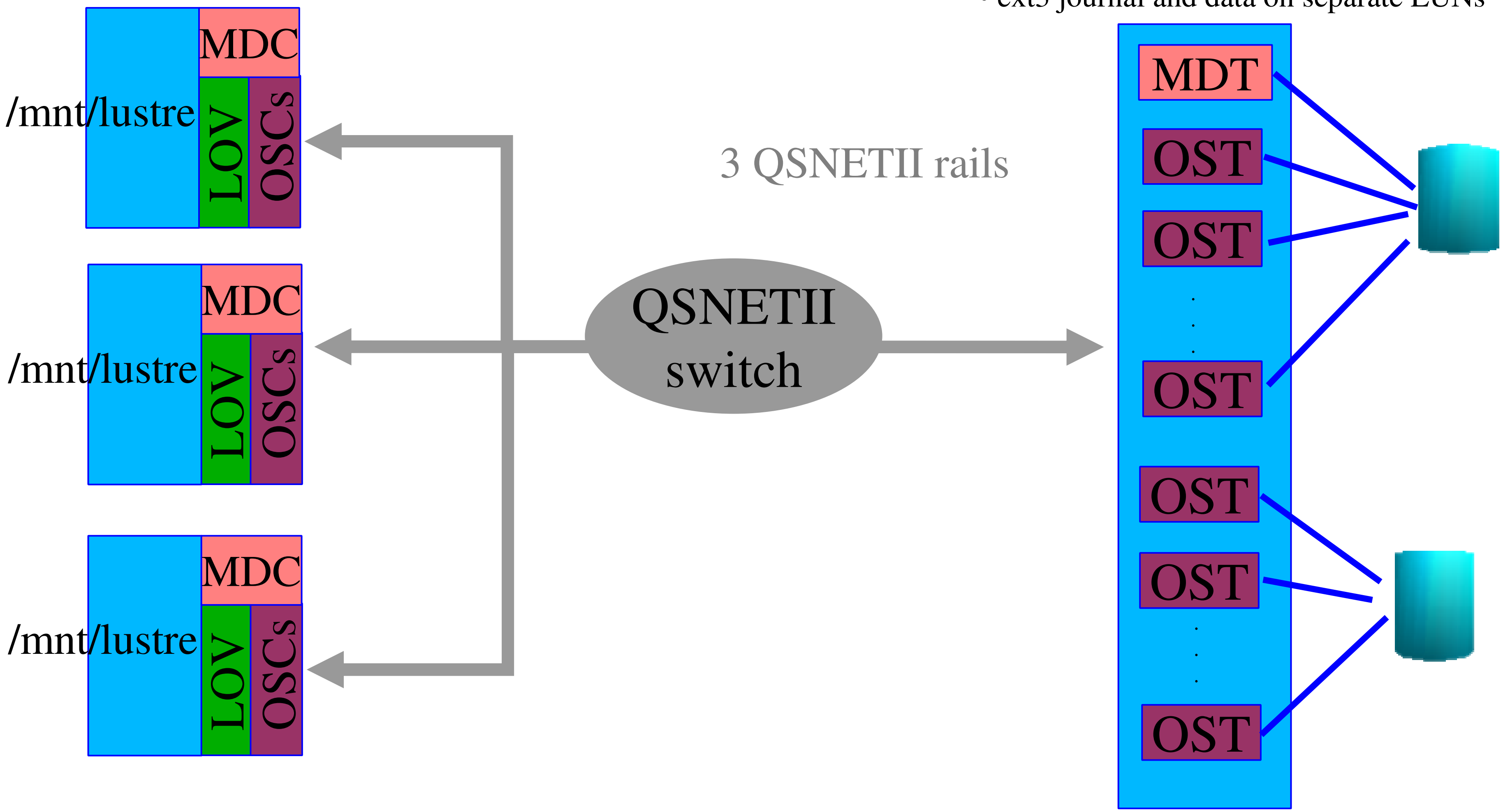
- Concept
- Architecture
- **Performances – optimizing Lustre**
- TERA10 super computer: +100GB/s
- From superclusters to grid: issues
- conclusion

Lustre world record performance



Lustre Layout

- OSS + MDS
- 16 OSTs (1 OST per FC link)
- 1 MDT
- ext3 journal and data on separate LUNs



Benchmark

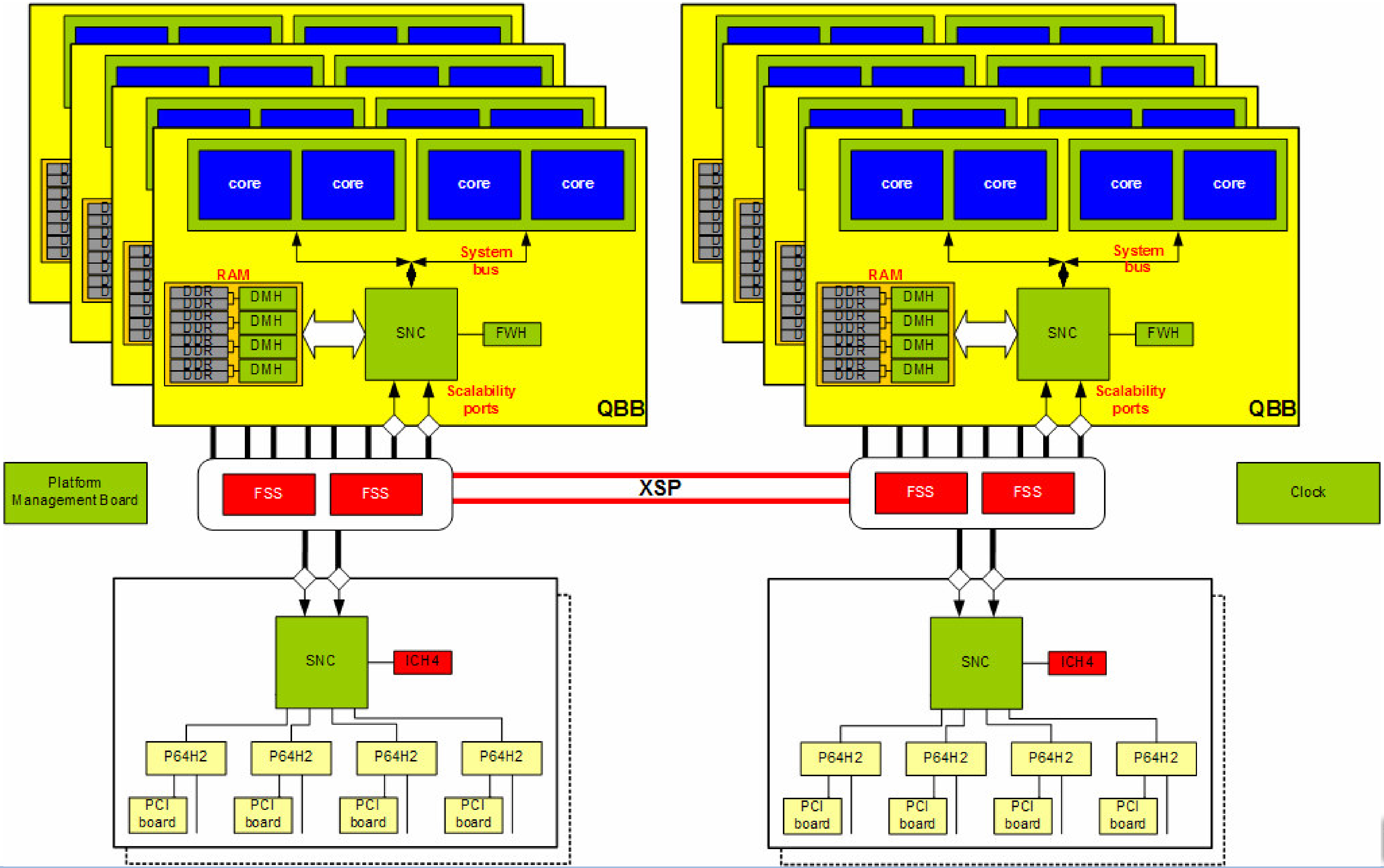
- **ES4 benchmark**
- **1 MPI task / CPU; 3 client nodes x 16 CPUs = 48 tasks**
- **Each task**
  - **allocates and initializes 1/16 of available memory with random numbers (available RAM = 58GB. 6GB allocated to system);**
  - **writes/reads the data into 1 file of 3.625 GB;**
  - **all files are in the same directory.**

Average Global Bandwidth:

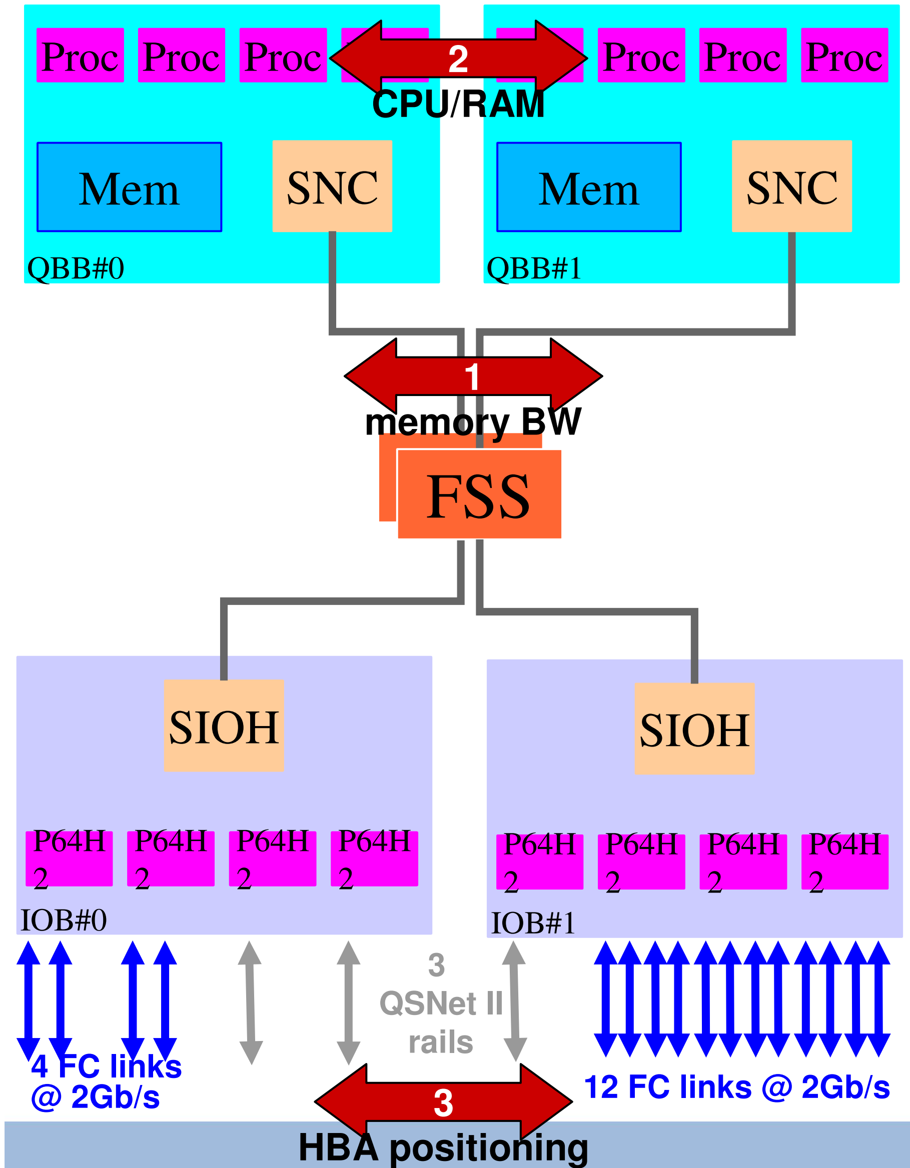
Read @ 2.1 GB/s

Write @ 2.6 GB/s

Server Architecture



OSS Internals and potential issues



- **Non Uniform Memory Architecture**
  - higher latency when accessing memory of a remote QBB
- **Independent memory controllers**
- **Linux kernel NUMA support**
  - memory allocator
    - Ex: keep memory as close as possible to the user of the memory
  - scheduler support
    - Ex: avoid migration of processes between QBBs

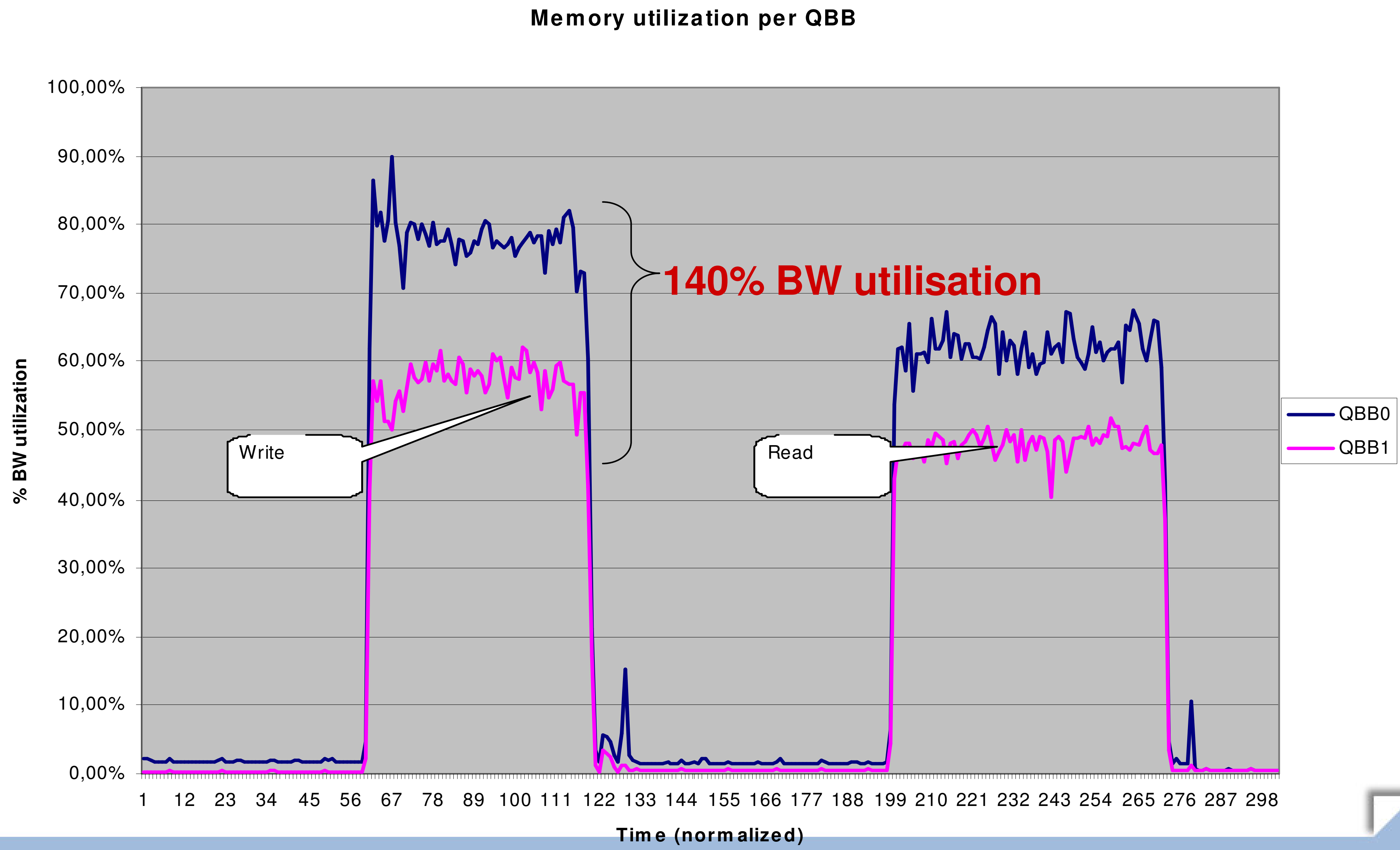
Issue #1
Balancing Lustre Memory Allocation

- The OST threads are often scheduled on the 1st NUMA node
- Lustre allocates/frees memory for each data transfer
  - lustre buffers are often located on a the same node
- ➔ Memory bandwidth usage imbalance

Solution

- What: Balance memory allocations between QBBs to reduce BW requirement per QBB
- ➔ How: bind the OST kernel threads

OSS Memory Bandwidth



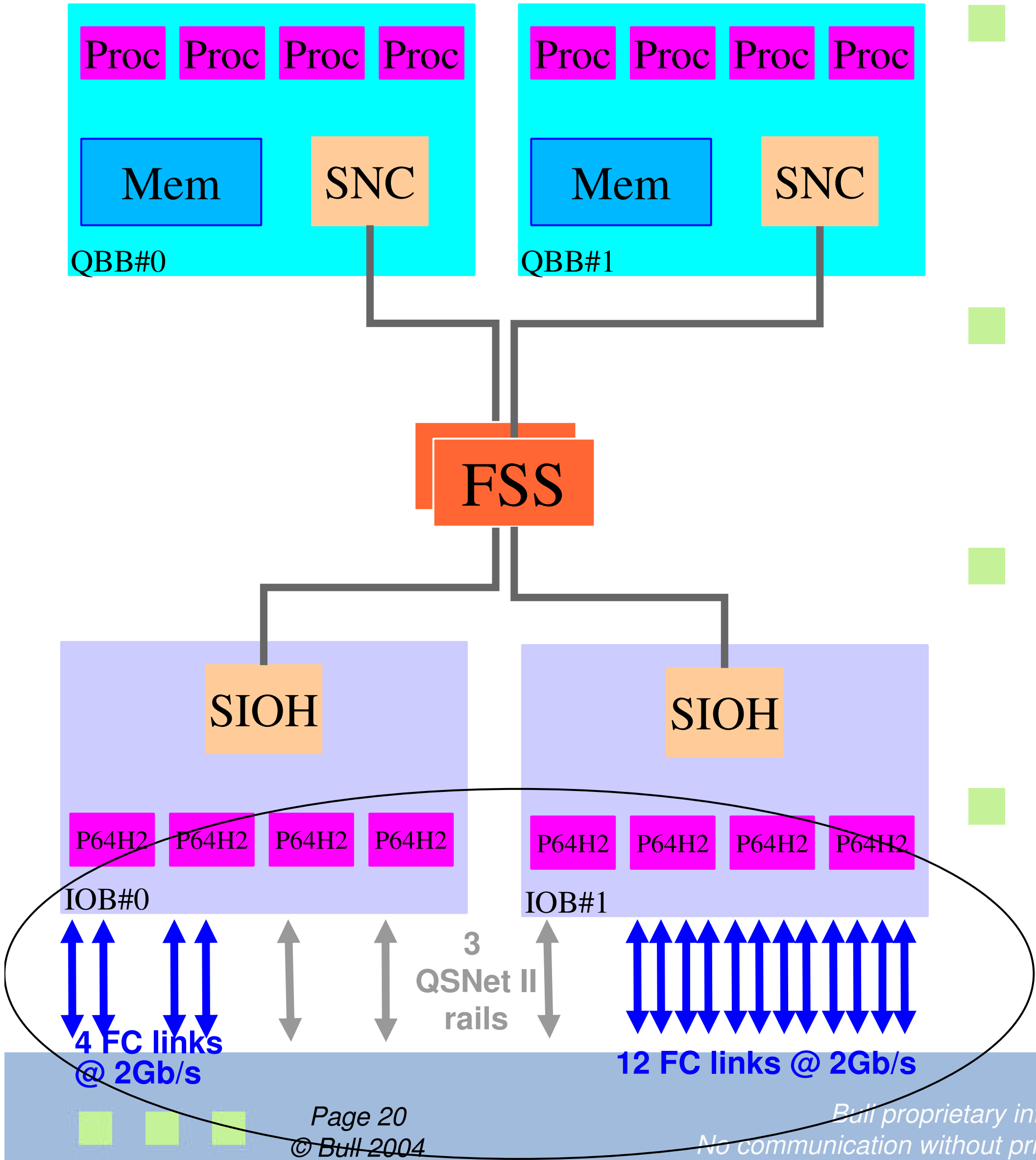
Issue #2
Memory Shortage on one NUMA Node

- **Several causes**
  - **External to Lustre (daemon, {buffer | page} cache, ...)**
  - **Lustre itself via ldiskfs journaling activity:**
    - **1 kjournald thread per OST (cpu affinity = all CPUs)**
    - **journal threads are often scheduled on the 1st NUMA node**
      - buffer cache data of journal devices are often located on the 1st NUMA node
    - **1GB journal size x 16 OSTs = 16GB = memory available on 1 QBB**

→ Memory is allocated on the 2nd NUMA node, even for OST threads scheduled on the 1st node.

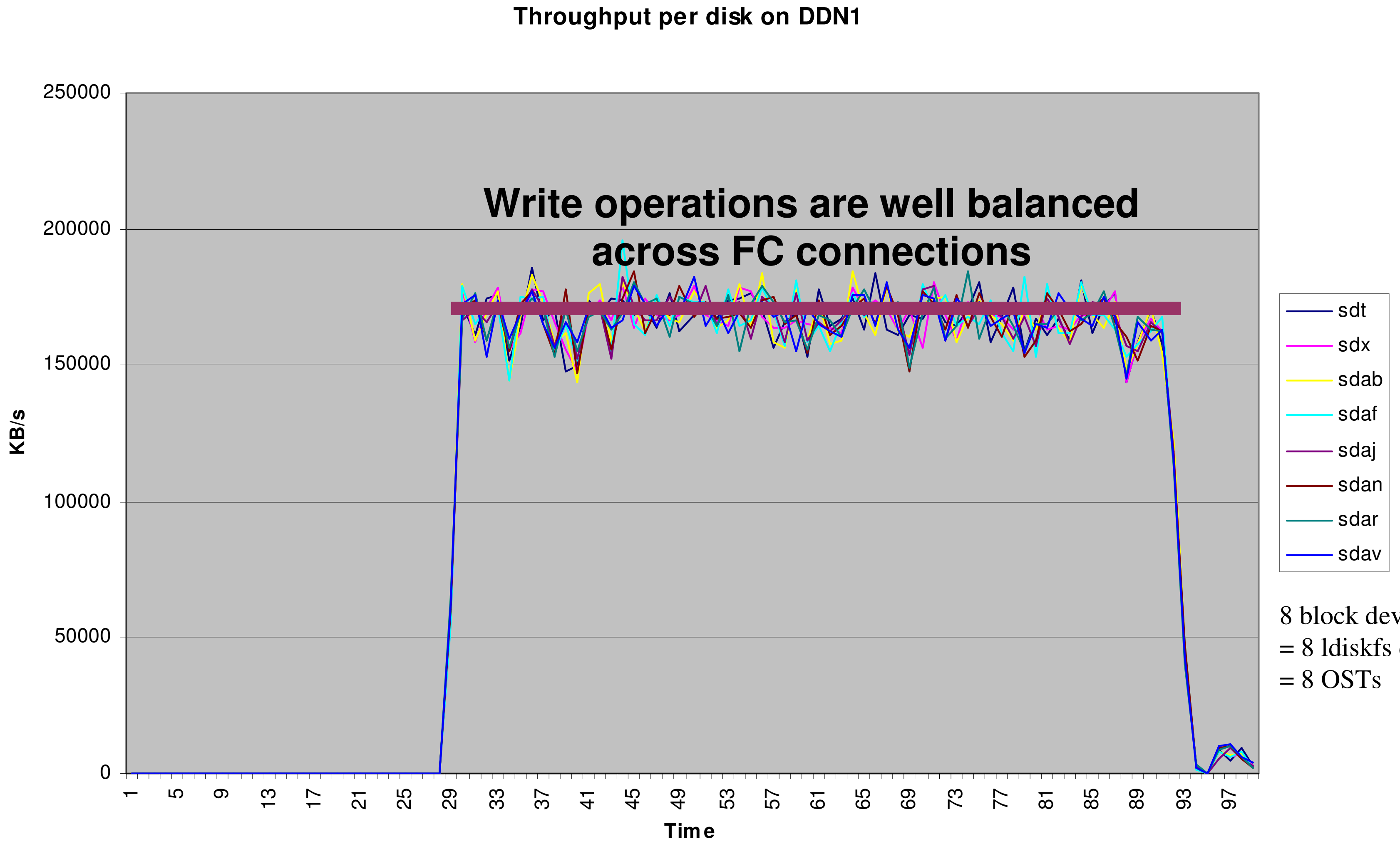
- **Current solution**
  - **bind the kjournald threads**
- **Better solution (?)**
  - **statically allocate pages for Lustre?**

Issue #3
IO board positioning



- **QsNetII**
  - **sustained BW: 900MB/s per board**
  - **3x boards**
- **FC**
  - **BW=256MB/s per board**
  - **16x board**
- **Issue: avoid bottleneck between SIOH and memory**
- **Solution: benchmark different topologies!**

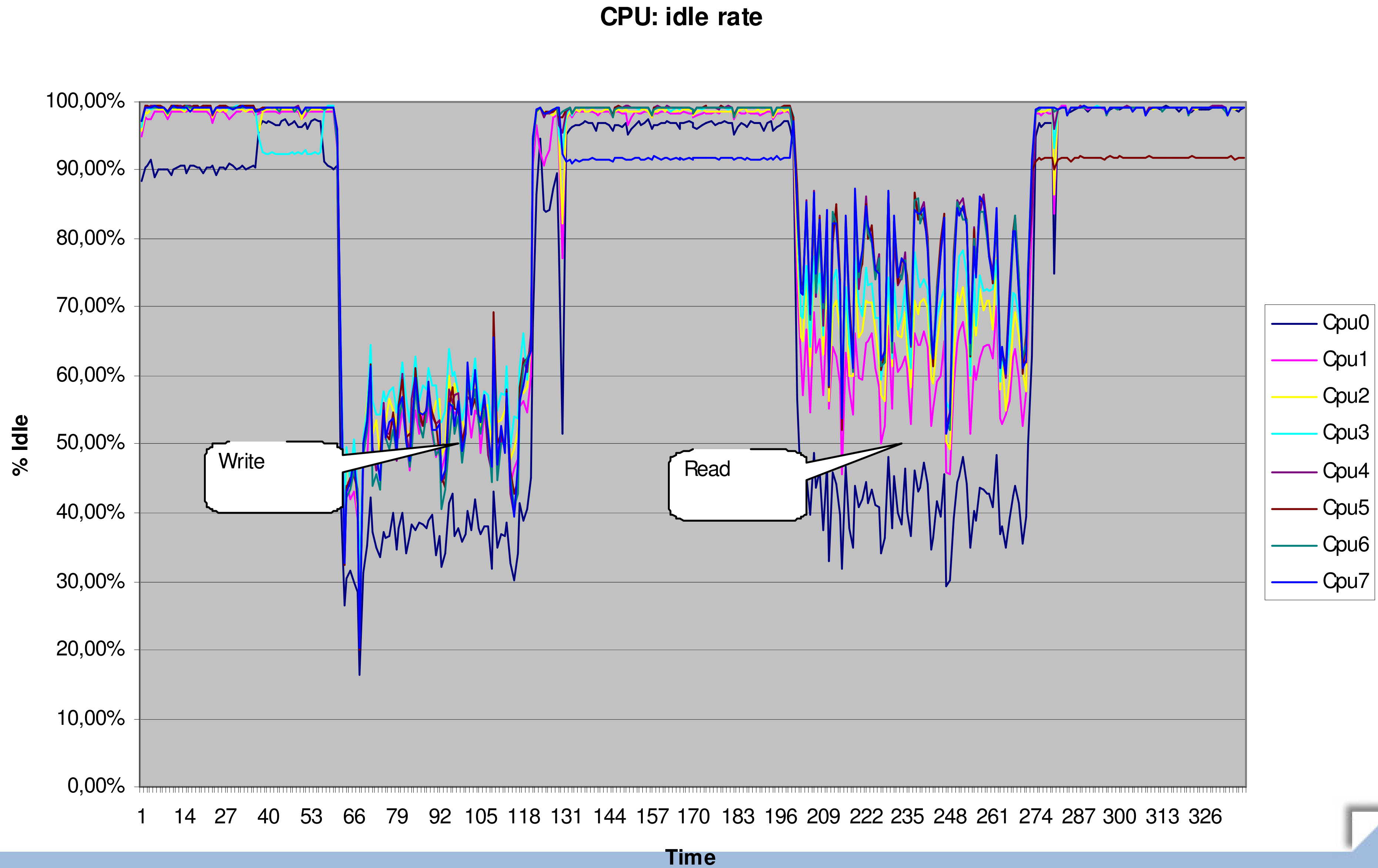
Ldiskfs Backing Storage Activity (write phase)



8 block devices
= 8 ldiskfs data LUNs
= 8 OSTs

Bull

OSS CPU Idle Rate



Agenda

- Concept
- Architecture
- Performances – optimizing Lustre
- **TERA10 super computer: +100GB/s**
- From superclusters to grid: issues
- conclusion

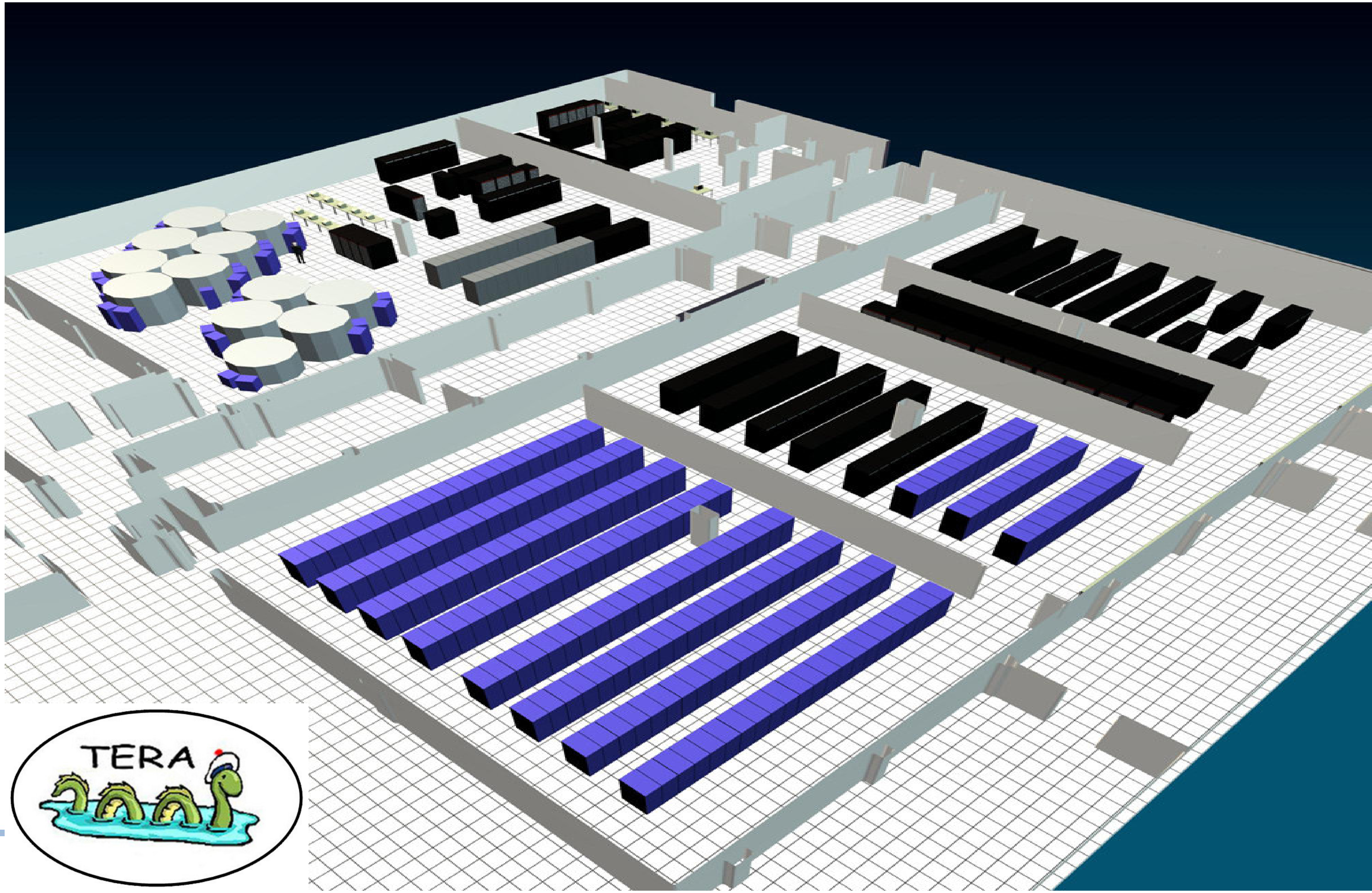
CEA TERA10 supercomputer

- **TERA10 sizing**
  - +60TFLOPS (Linpack HPC): 544x[8xMontecito]
  - 30TB RAM
  - 1PB disk space
  - 100GB/s sustained BW from RAM to disks

- **Bull approach**
  - Lustre file system
  - 2x MDS (each with 4x Intel Montecito)
  - 54x [server NS5160 with 8x Intel Montecito]
  - 27x DDN S2A9500 (FC 4Gb/s)

- **Write performance per server: 2.6GB/s sustained**
- **Total performance: 108GB/s**

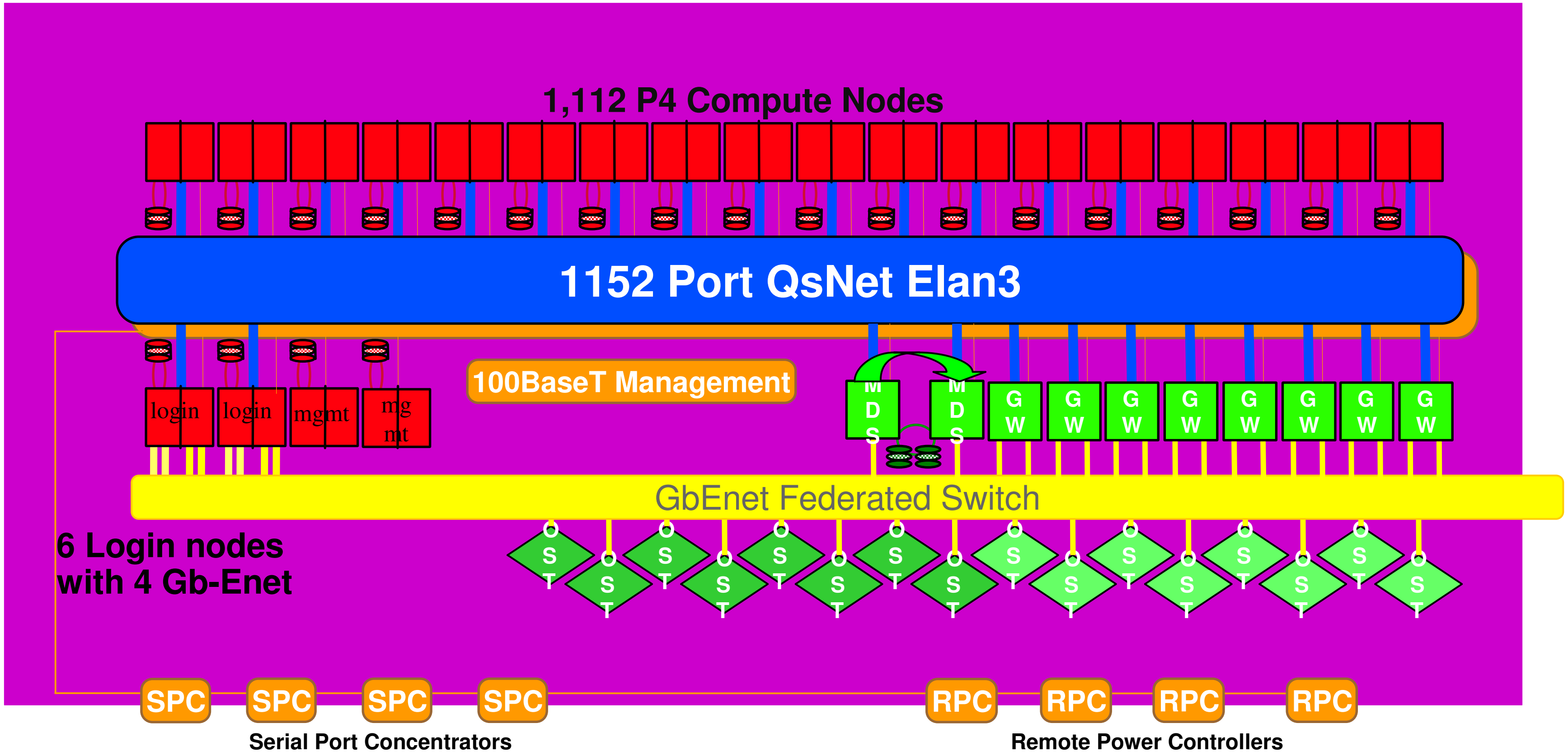
TERA10 implementation



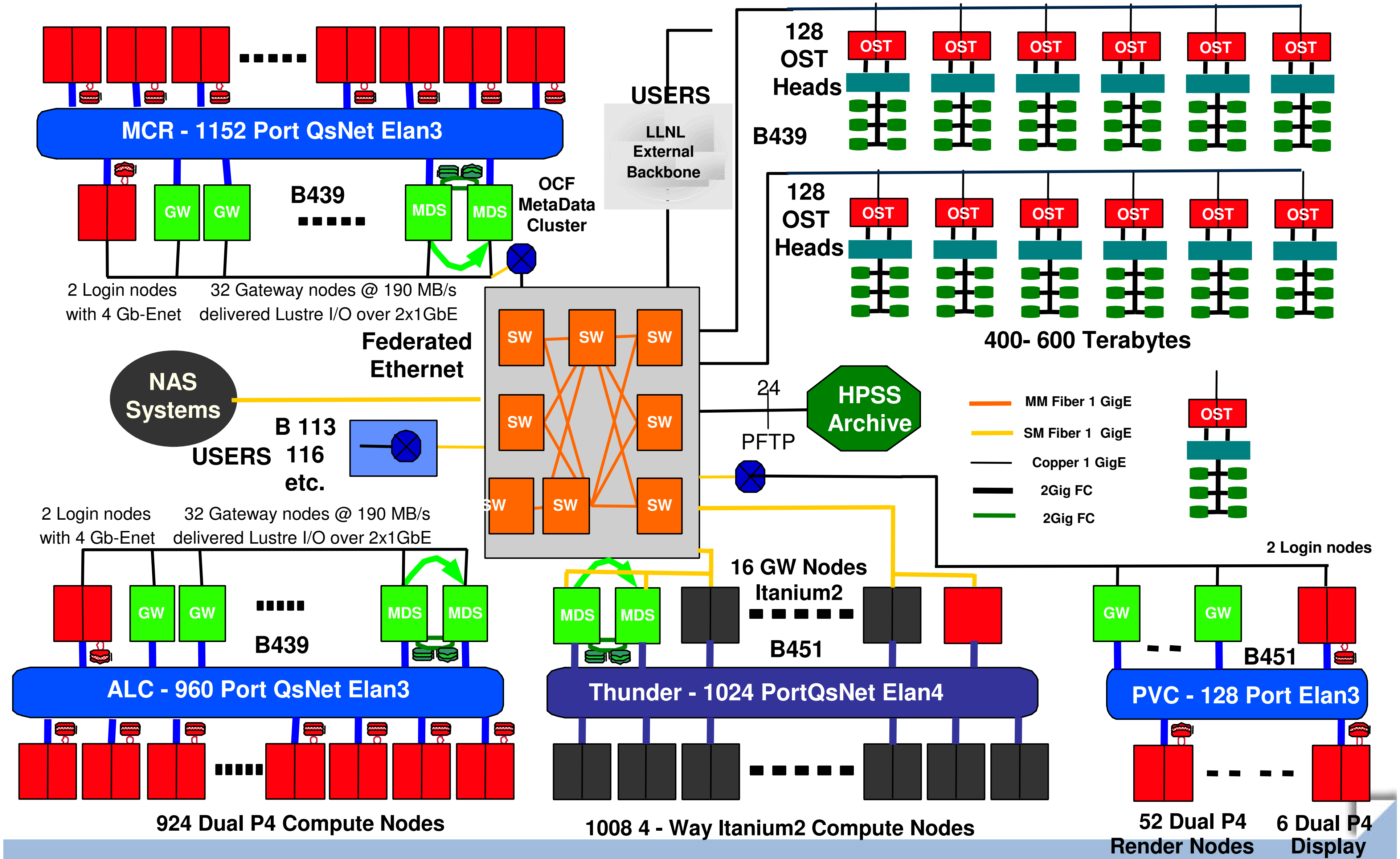
Agenda

- Concept
- Architecture
- Performances – optimizing Lustre
- TERA10 super computer: +100GB/s
- **From superclusters to grid: issues**
- conclusion

LLNL – MCR


1,112 P4 Compute Nodes
1152 Port QsNet Elan3
100BaseT Management
login
login
mgmt
mg mt
M D S
M D S
G W
GbEnet Federated Switch
6 Login nodes with 4 Gb-Enet
O S T
SPC
SPC
SPC
SPC
RPC
RPC
RPC
RPC
Serial Port Concentrators
Remote Power Controllers

LLNL – Multi Cluster



Lustre for Grid Architecture

- MDS
  - One “root” MDS
  - Several replications – one per grid site
- OSS
  - Sets of OSS – one set per grid site
- Gateways to other filesystems
  - NFS / Solaris
  - GPFS / CXFS
  - distributed

Issues

- **Reliability**
  - **MDS replica**
  - **Transactional model**
- **Security**
  - **Access rights**
  - **Encryption**
- **HW standardization**
  - **Ideal world: same storage device on all grid sites**
- **Performances**
  - **Timeout**

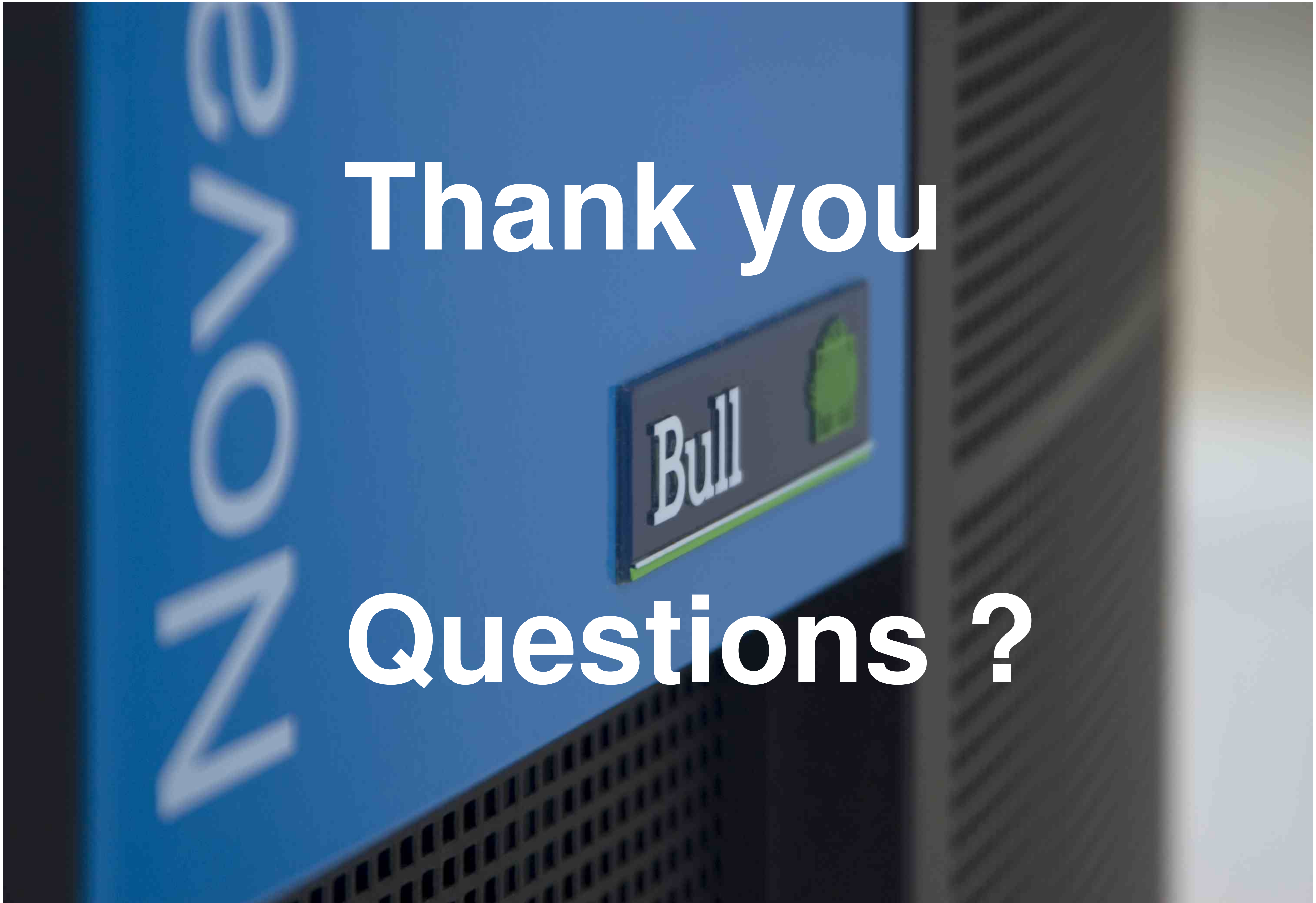
Bull

Agenda

- Concept
- Architecture
- Performances – optimizing Lustre
- TERA10 super computer: +100GB/s
- From superclusters to grid: issues
- conclusion

- **LUSTRE is mature for high-end superclusters**
  - **LLNL**
  - **CEA**
- **LUSTRE offers the most flexible architecture and the best performances**

- **“grid-ization” of LUSTRE is still at its beginning**
- **Next steps should be making the transactions securized (transactional model)**
- **Performance issues are the step further. They mostly depends on the grid network performances.**



Thank you

Questions ?