

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS

Swiss National Supercomputing Centre



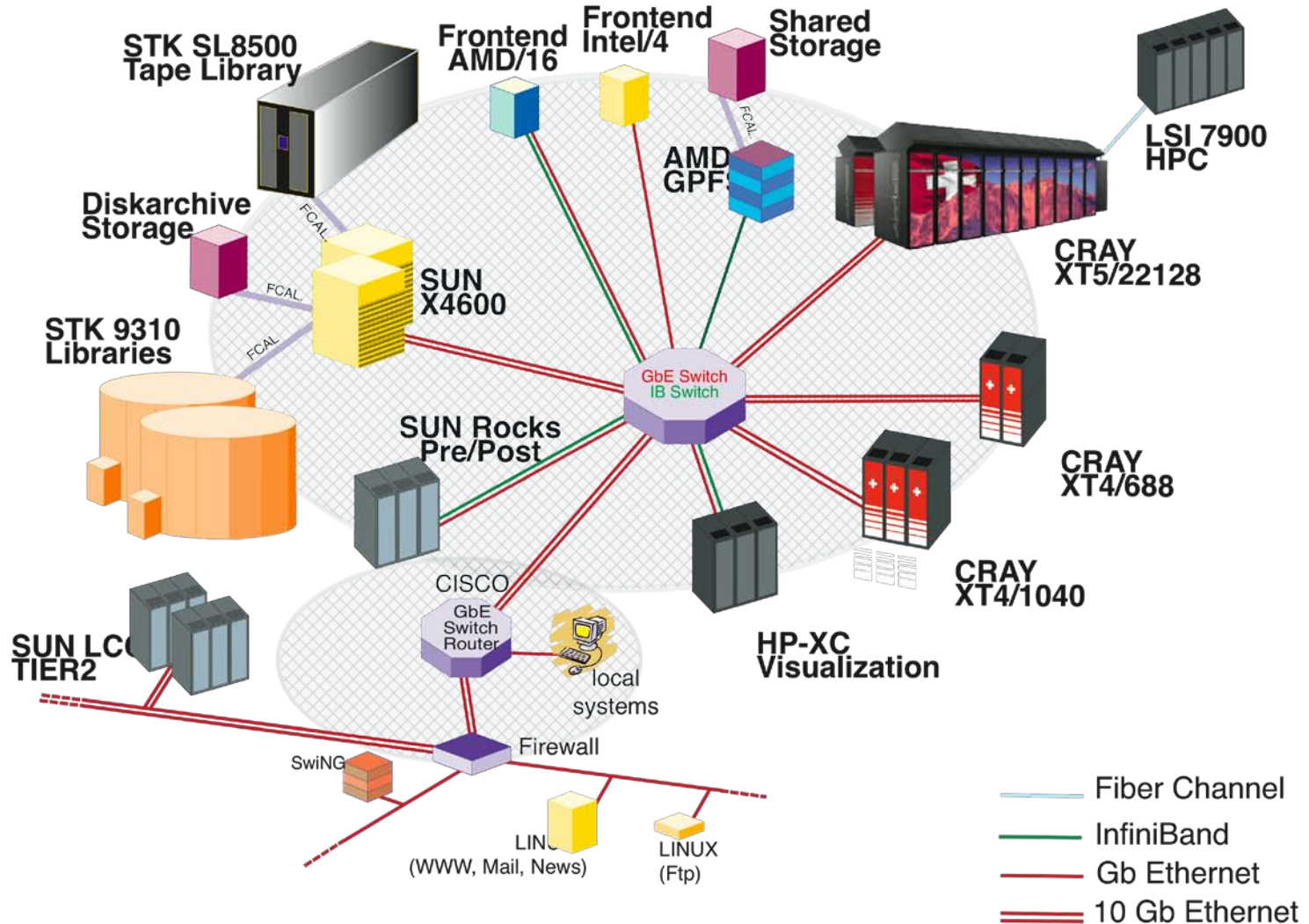
HPC Storage Systems at CSCS

Hussein N. Harake

Points to Cover

- Overview of CSCS resources
- Storage systems
- Global file-system
- IB network
- Implementation
- Case study IB congestion

Non-technical Map of CSCS



©08.03.2010 CSCS

Storage systems

- Cray XT5
 - ✓ 5 LSI HPC 7900
 - ✓ Lustre 1.6.X
 - ✓ 20 IO servers
 - ✓ 287 TB Used space
- Cray XT3/XT5
 - ✓ 1 X DDN 9550
 - ✓ Lustre 1.6.X
 - ✓ 2 IO servers
 - ✓ 83 TB Used space

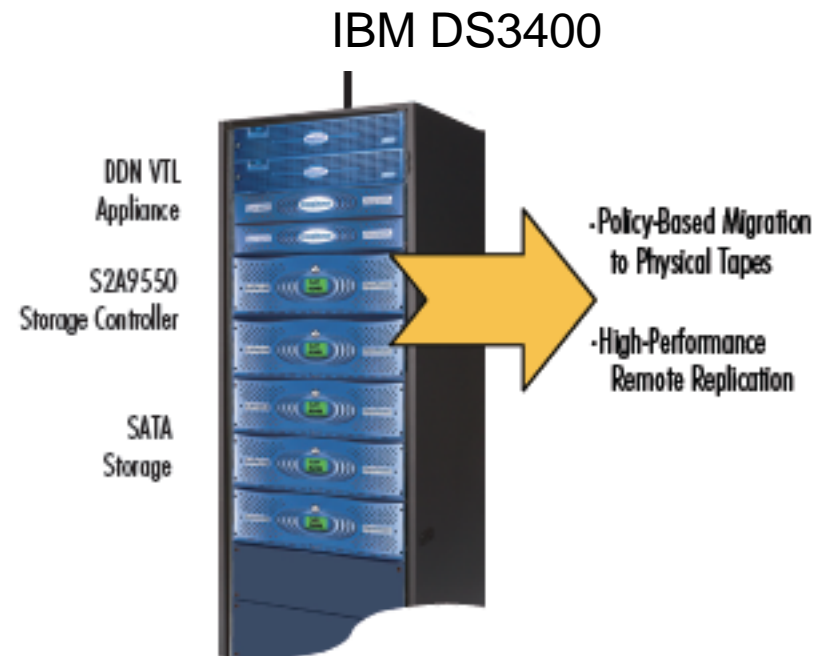


LSI HPC 7900

A scratch file-system available on each XT5 machines
Fast access, short term use and under clean policy

Storage systems

- XT4 (Meteo Swiss)
 - ✓ SUN STK 6540
 - ✓ Lustre file-system 1.6.X
 - ✓ 6 IO servers
- XT4 (Backup Meteo Swiss)
 - ✓ IBM DS3400 & Exp3000
 - ✓ Lustre file-system 1.8.X
 - ✓ 4 IO servers
 - ✓ 2 Inet routers



Each cluster usually has its own HW where scratch file-system is hosted

Storage systems

- Global File System /project (In production)
 - ✓ GPFS
 - ✓ 220 TB Used Space
 - ✓ 3 X IBM DS4800 controllers
 - ✓ 18 X EXP810 Enclosures SATA disks
 - ✓ 4 X EXP710 Enclosures FC disks
 - ✓ 4 X Data servers
 - ✓ 2 X Meta data servers
 - ✓ 12 X FC HCA's 4G/bits Host Ports



IBM DS4800

Reliable, Reasonable access time and Medium term use
Accessed from all CSCS HPC systems

Infiniband resources

- Production Systems

- ✓ IB Flextronics 144 ports 4X DDR (F- X430077)
- ✓ 4 X Switches SDR 24 ports (Flextronics, Voltaire, Cisco)
- ✓ 2 X 4036 Voltaire Switches 36 Ports QDR
- ✓ 2 X Mellanox Bridge BX4010 GW (IB – Ethernet – FC)
- ✓ SUN DS648 (M9) up to 648 Switch



Voltaire 4036

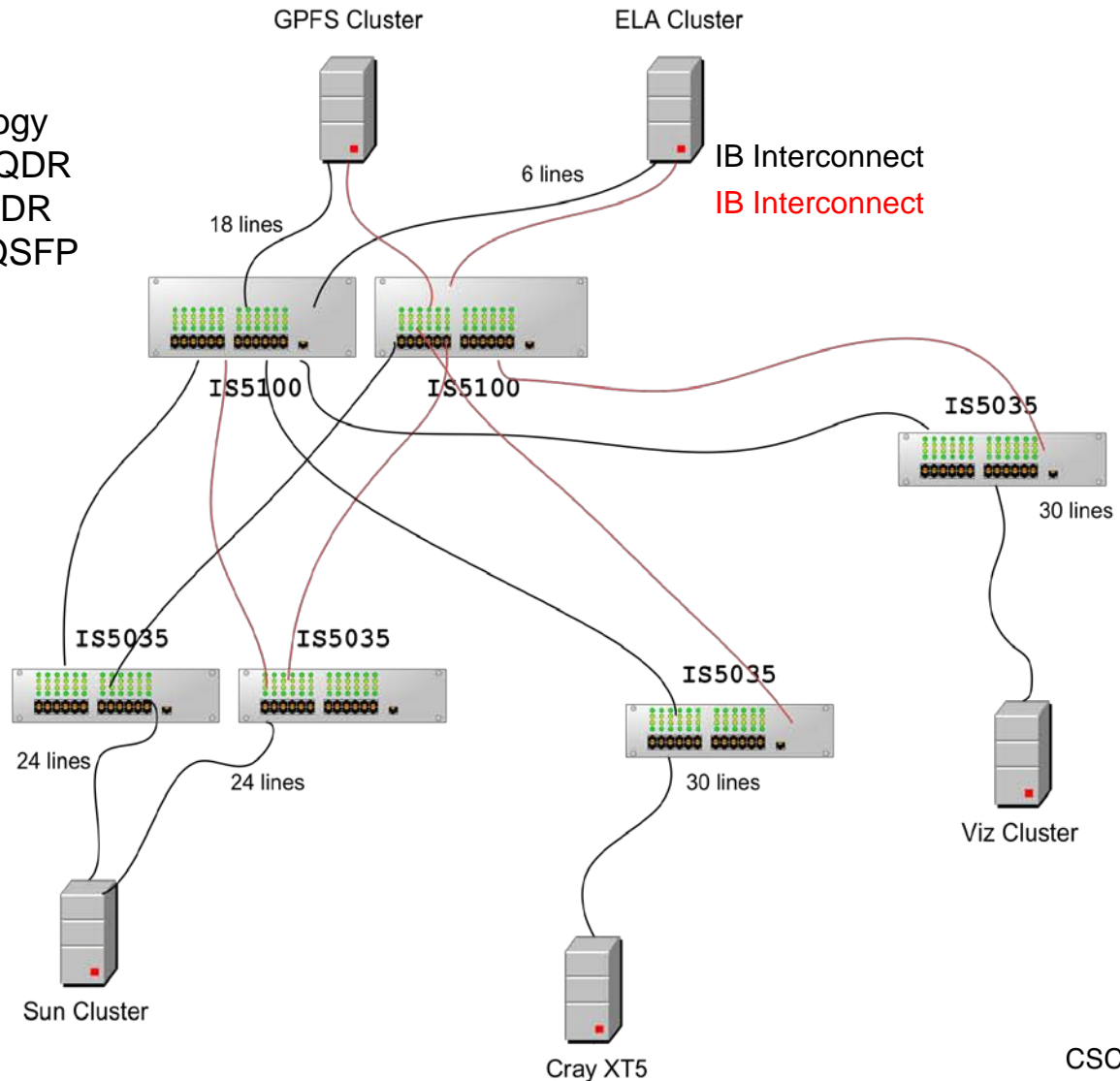
Infiniband is used for storage and computing

New Infiniband Storage Network

April / May 2010

- ✓ Based on Mellanox Technology
- ✓ 2 X IS5100 up to 108 Ports QDR
- ✓ 4 X IS5035 up to 36 Ports QDR
- ✓ 40Gb/s QDR PCI-E Gen 2 QSFP

Faster access
Congestion improvement
High availability



Implementation of Global File-System (/project)

- 1st phase:

- ✓ DDN 9550 controller
- ✓ Flextronics SDR 24 Ports Switch
- ✓ 88 TB of usable disk

- 2nd phase:

- ✓ Replacing HW with 3 DS4800
- ✓ Replacing Switch- HCA's with DDR
- ✓ Increasing capacity to 220TB
- ✓ Separating Data from Meta-Data
- ✓ Rebalancing the file-system
- ✓ No service interruption



Flex. DDR 144 ports

Implementation of /project

- 3rd phase (April / May):

- ✓ Implementing 3 X DS5100
- ✓ 8 X HD Enclosures 5060
- ✓ 4 X EXP5000 FC Disks
- ✓ Replacing Switch - HCA's with QDR
- ✓ Increasing capacity to ~500TB
- ✓ Moving Data to the new file system

- Performance improvement and adding bandwidth
- Replacing raid5 with raid6
- Adding more capacity
- Keeping a balanced system



EXP 5060

Implementation of /project

4th phase (September / October 2010):

✓ Increasing capacity to ~1.0PB



IS5100



GW BX4010

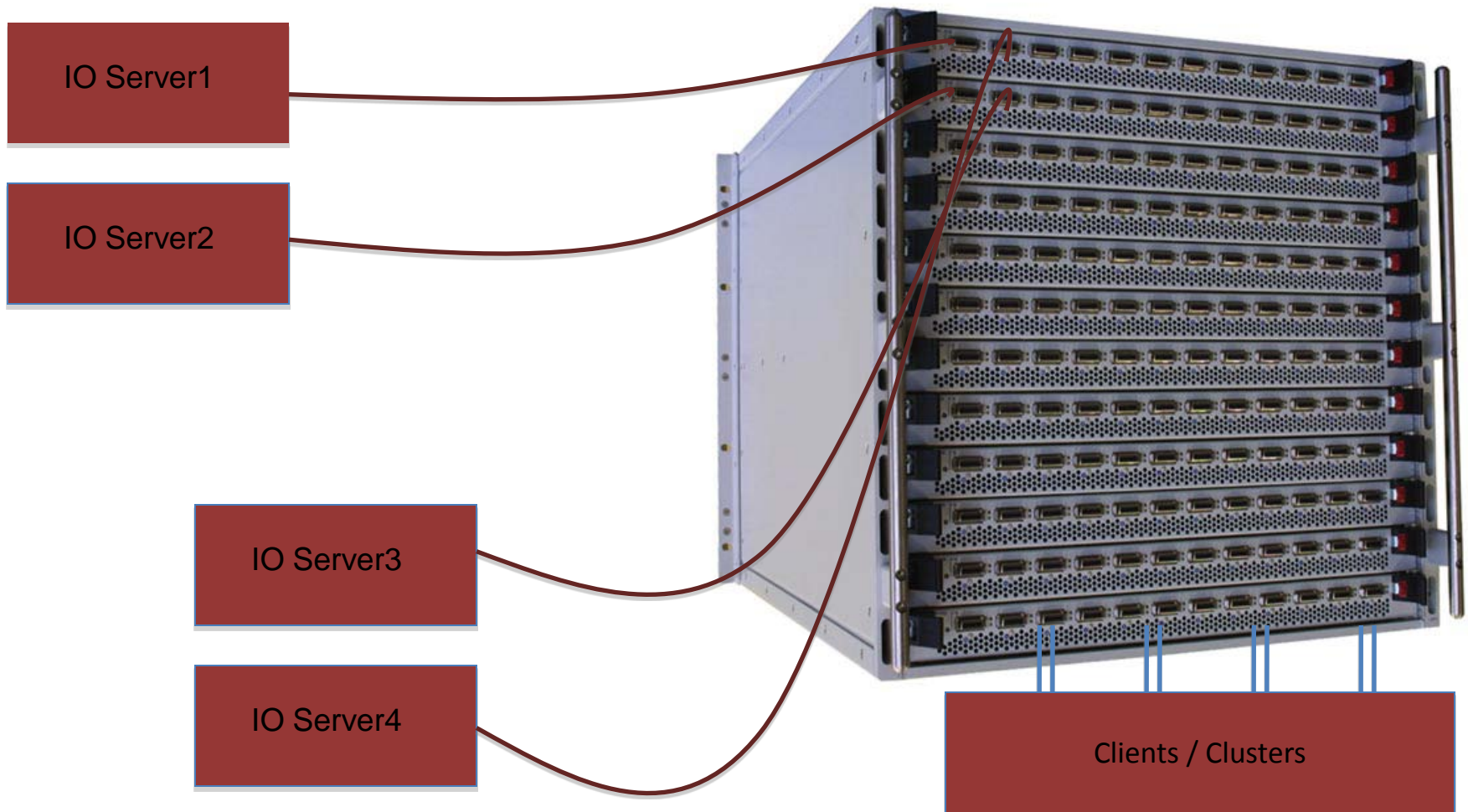
Case study Infiniband congestion

```
server:~ # ifconfig ib0
```

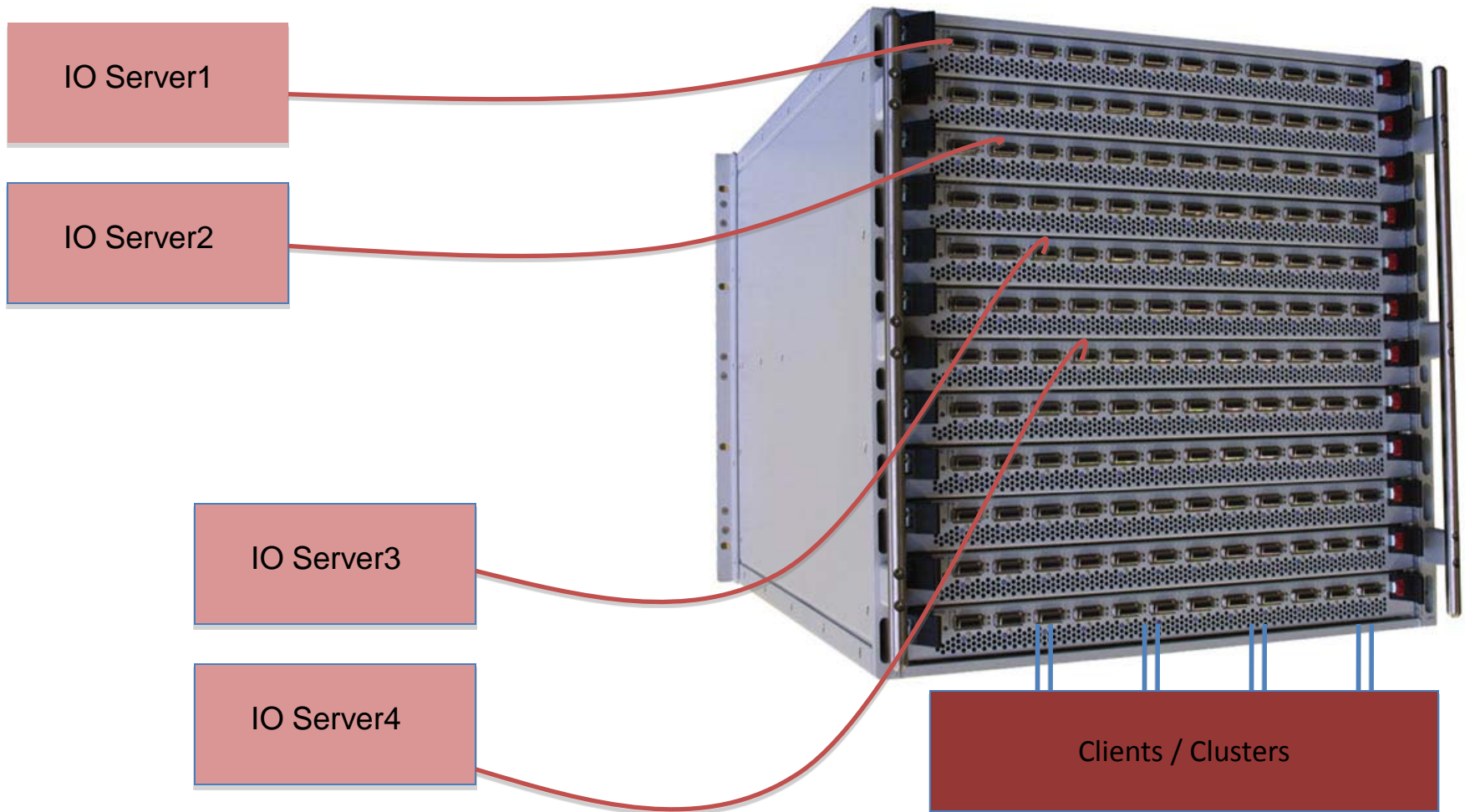
```
ib0    Link encap:UNSPEC HWaddr 80-00-04-04-FE-80-00-00-00-00-00-00-00-00-00-00
inet addr: 0.0.0.0 Bcast:0.0.0.0 Mask:255.255.255.0
inet6 addr: fe80::202:c902:26:789d/64 Scope:Link
UP BROADCAST RUNNING MULTICAST MTU:65520 Metric:1
RX packets:55376274240 errors:0 dropped:0 overruns:0 frame:0
TX packets:85269371118 errors:0 dropped:415615693 overruns:0 carrier:0
collisions:0 txqueuelen:8192
RX bytes:187729557937531 (179032857.8 Mb) TX bytes:346674676945950 (330614735.5 Mb)
```

0.5% of the transmitted packets were dropped

Case study IB congestion



Case study IB congestion



Case study IB congestion

ib0 Link encap:UNSPEC HWaddr 80-00-04-04-FE-80-00-00-00-00-00-00-00-00-00-00
inet addr:148.187.6.101 Bcast:148.187.6.255 Mask:255.255.255.0
inet6 addr: fe80::202:c902:26:789d/64 Scope:Link
UP BROADCAST RUNNING MULTICAST MTU:65520 Metric:1
RX packets:11583433402 errors:0 dropped:0 overruns:0 frame:0
TX packets:**15659669233** errors:0 dropped:**591821** overruns:0 carrier:0
collisions:0 txqueuelen:8192
RX bytes:141517594128295 (134961694.8 Mb) TX bytes:212067529766385 (202243356.4 Mb)

0.003% of the transmitted packets were dropped

