Getting the best from Lustre in a NUMIOA and multirail InfiniBand environment

sebastien.buisson@bull.net

Architect of an Open World™

Getting the best from Lustre in a NUMIOA and multirail InfiniBand environment

sebastien.buisson@bull.net

Architect of an Open World™

Lustre with NUMIOA and multirail IB

- What is NUMIOA?
- What is multirail IB?
- How does Lustre behave on a NUMIOA server with multiple Infiniband interfaces?
- Evolutions for the multirail part
- Evolutions for the NUMIOA part
- Application to the OSS server
- Coming next...

What is NUMIOA?

Non-**U**niform **M**emory **A**ccess

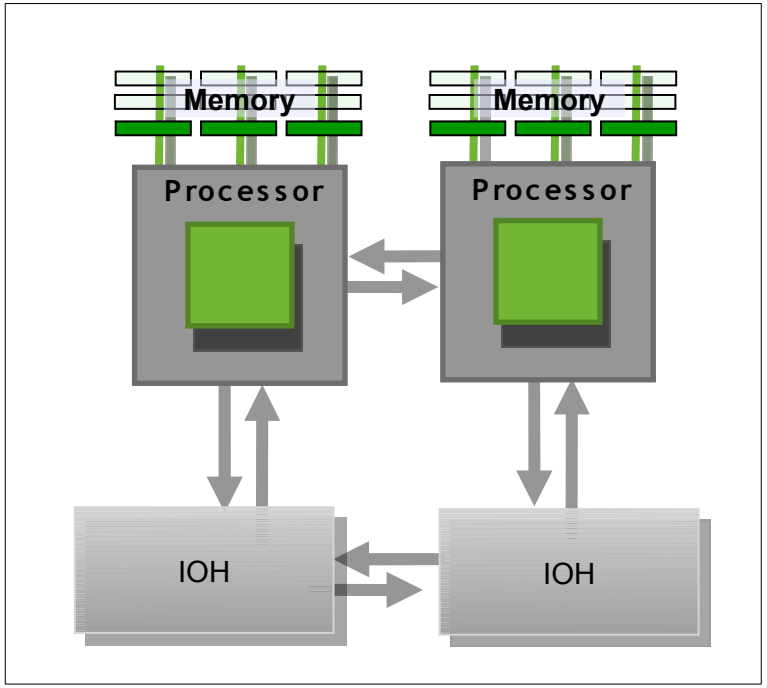
+

Non-**U**niform **IO A**ccess

=

Non-**U**niform **M**emory and **IO A**ccess

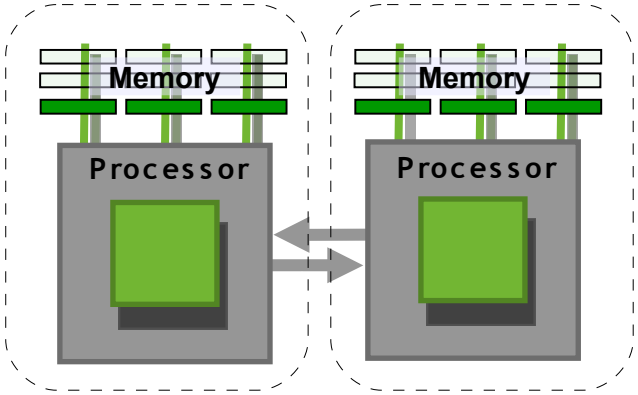
What is NUMIOA?



delimits a physical node

What is NUMIOA?

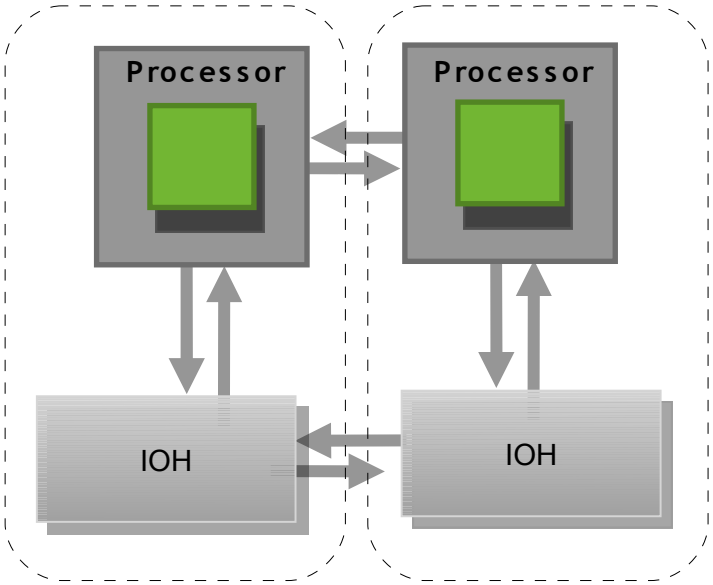
- **N**on-**U**niform **M**emory **A**ccess



- - - - delimits a NUMA node

What is NUMIOA?

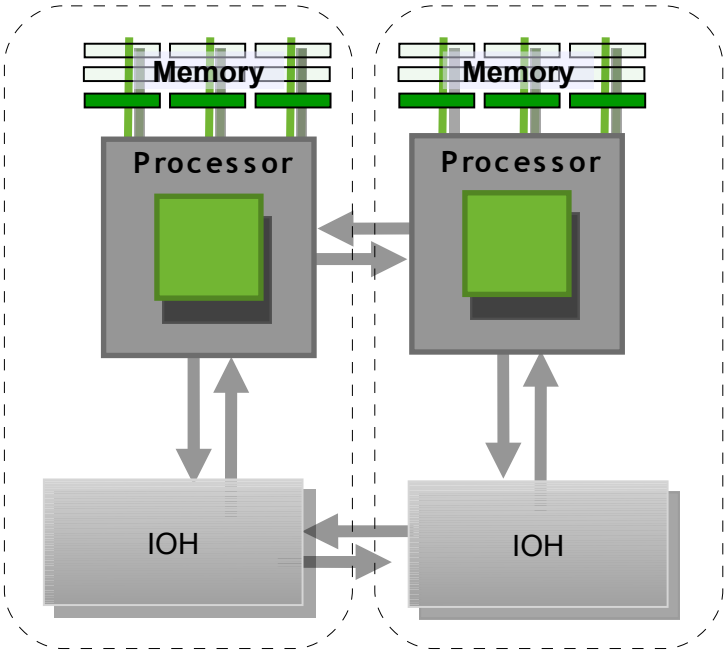
- **N**on-**U**niform **IO** **A**ccess



- - - - delimits a NUIOA node

What is NUMIOA?

- **N**on-**U**niform **M**emory and **IO** **A**ccess



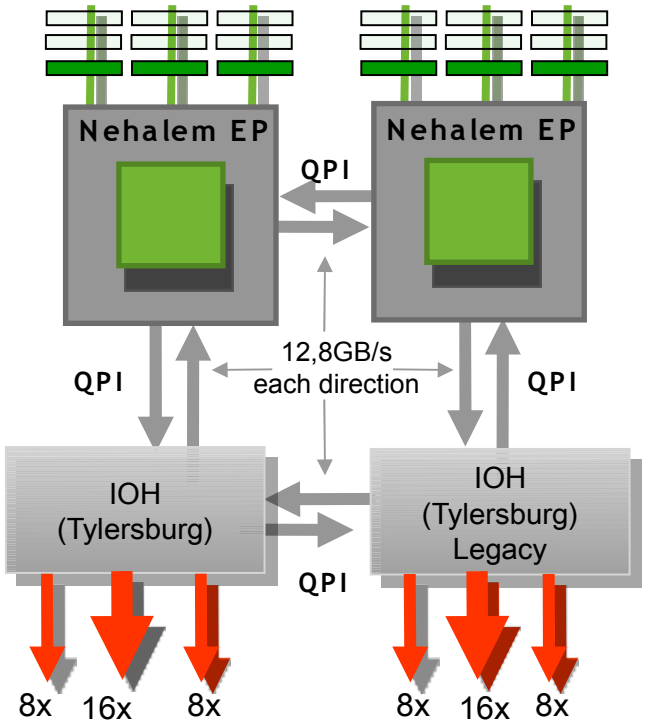
- - - - delimits a NUMIOA node

What is multirail IB?

- Taking advantage of several Infiniband interfaces
  - on clients and servers
  - for bandwidth aggregation
- Our goal:
  - Lustre network bandwidth = Σ individual link bandwidths

How does Lustre behave on a NUMIOA server?

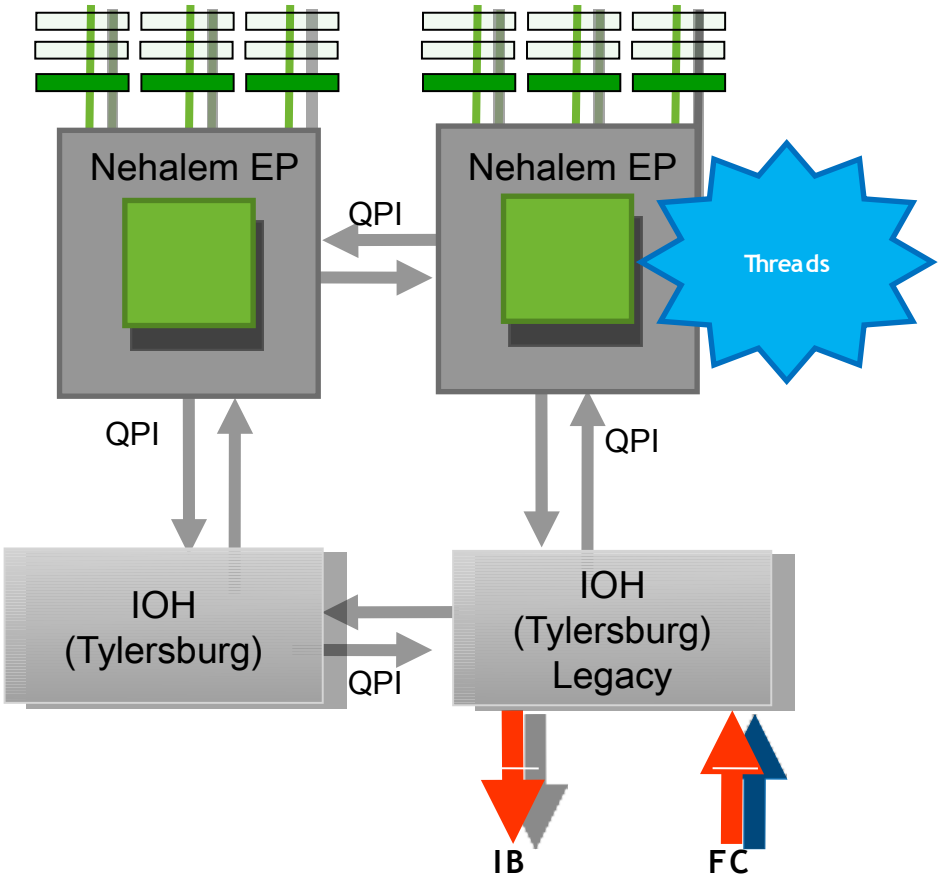
- Our testbed:



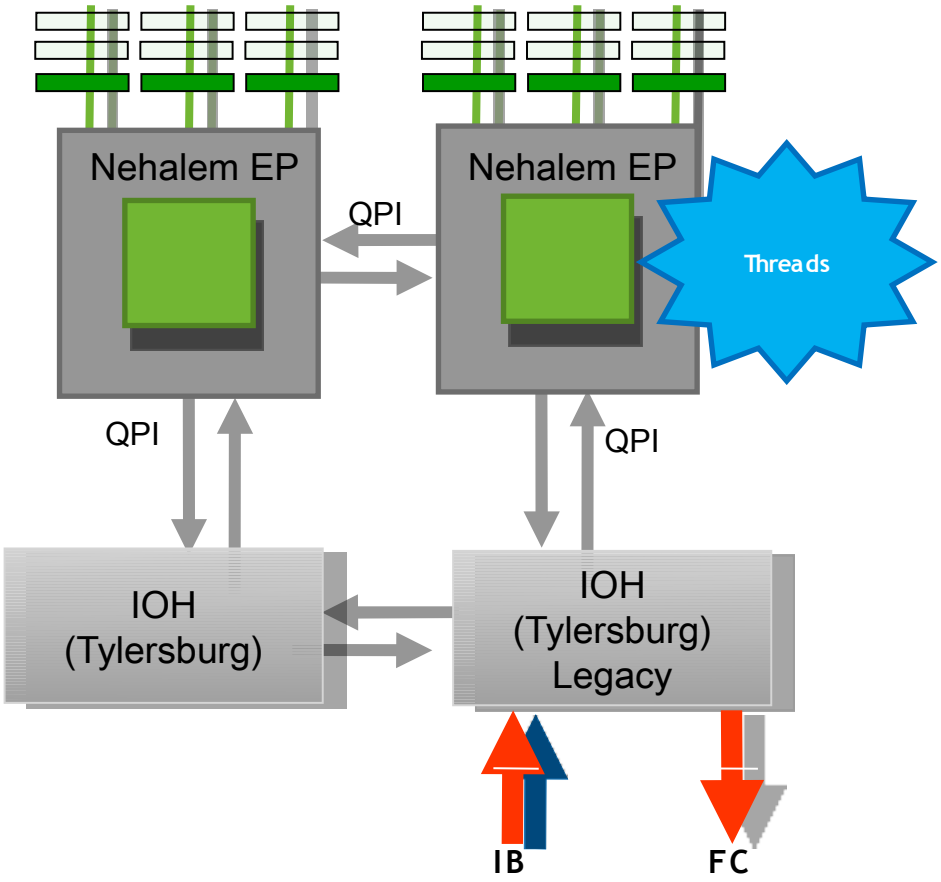
How does Lustre behave on a NUMIOA server?

- Elementary performance in R&D labs without Lustre
- Methodology
  - Disk IO bandwidth measured with xdd
  - IB bandwidth measured with qperf
  - run independently

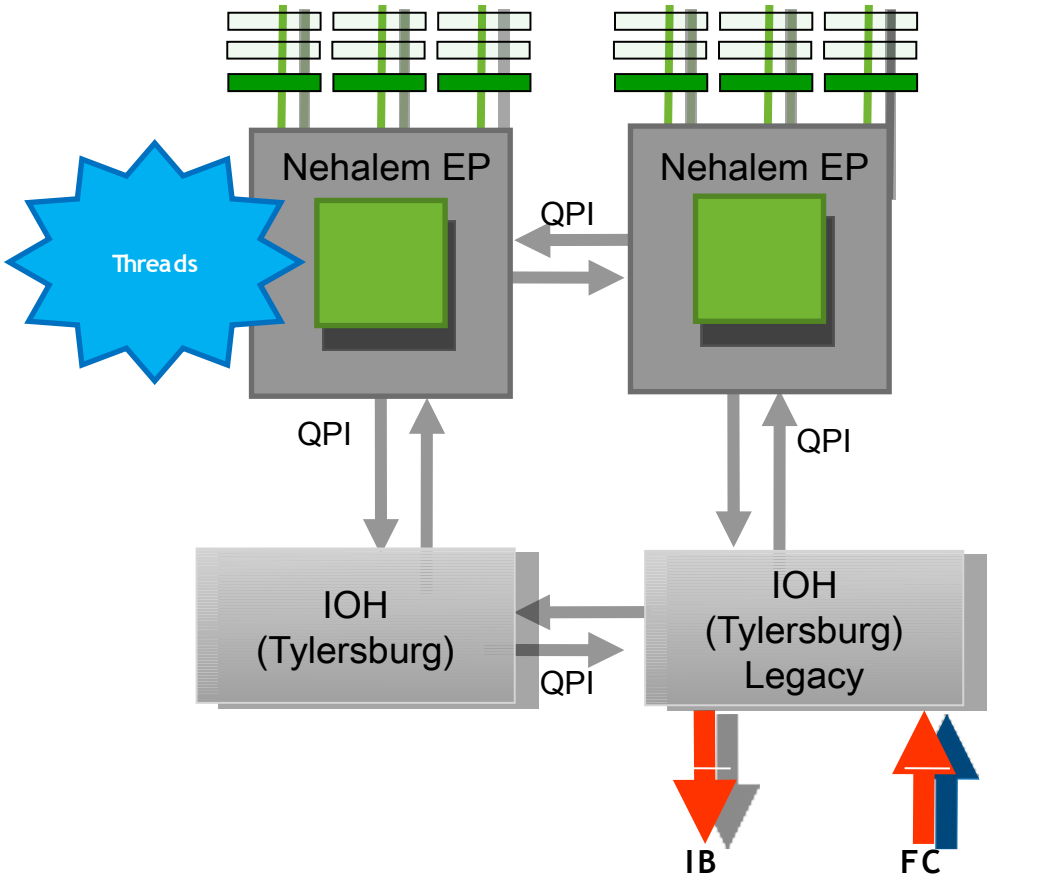
Simulated read, threads correctly localized



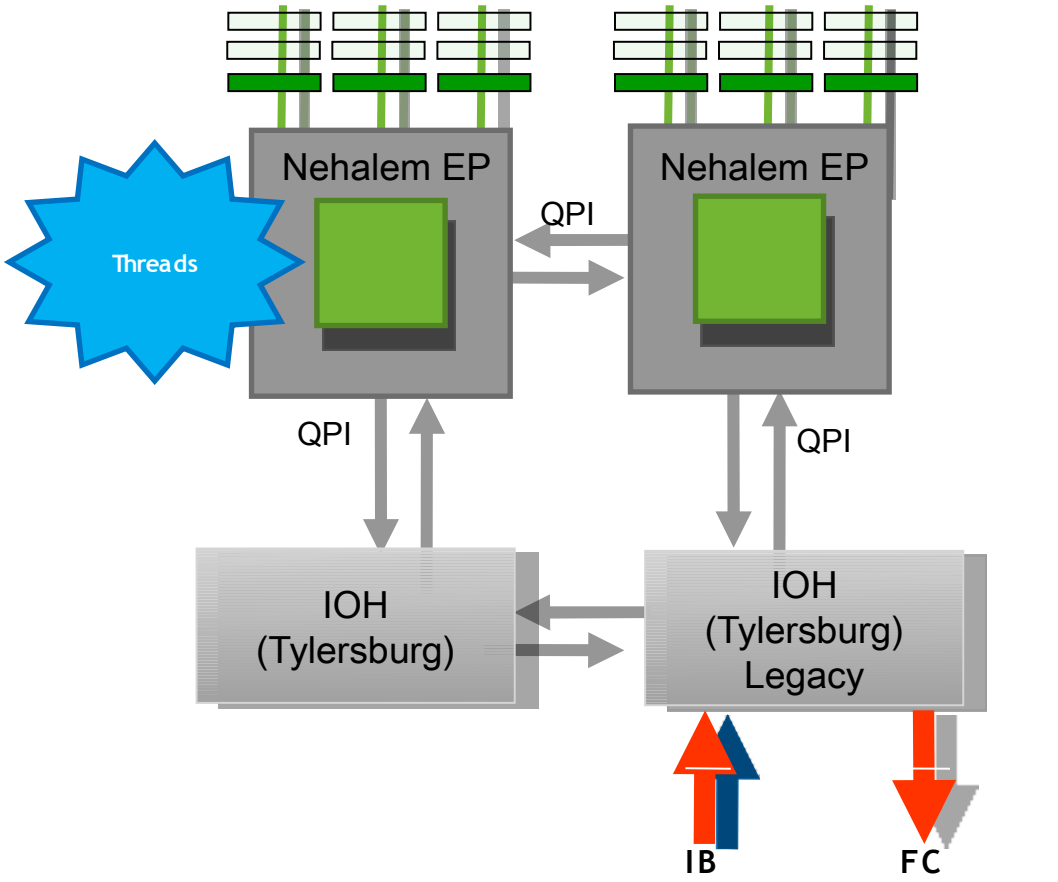
Simulated write, threads correctly localized



Simulated read, threads NOT correctly localized

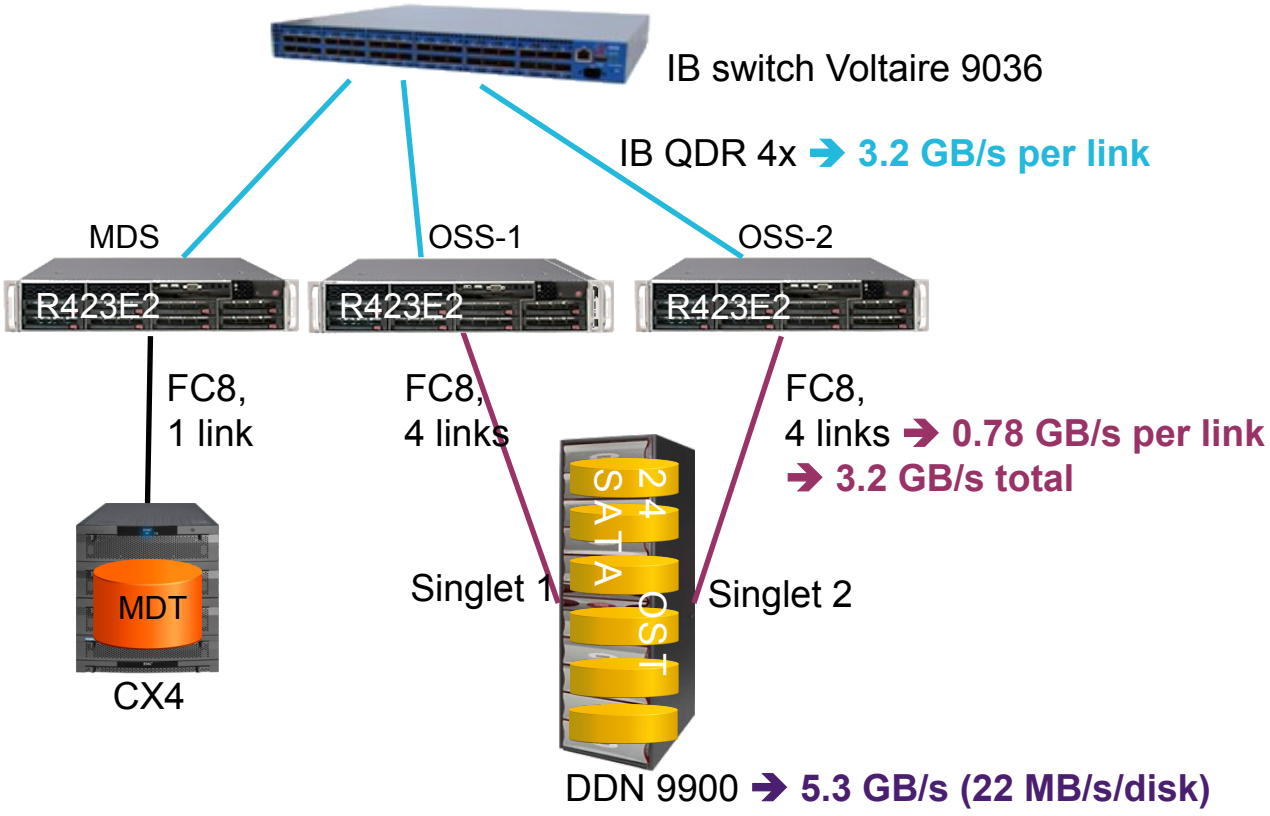


Simulated write, threads NOT correctly localized



How does Lustre behave on a NUMIOA server?

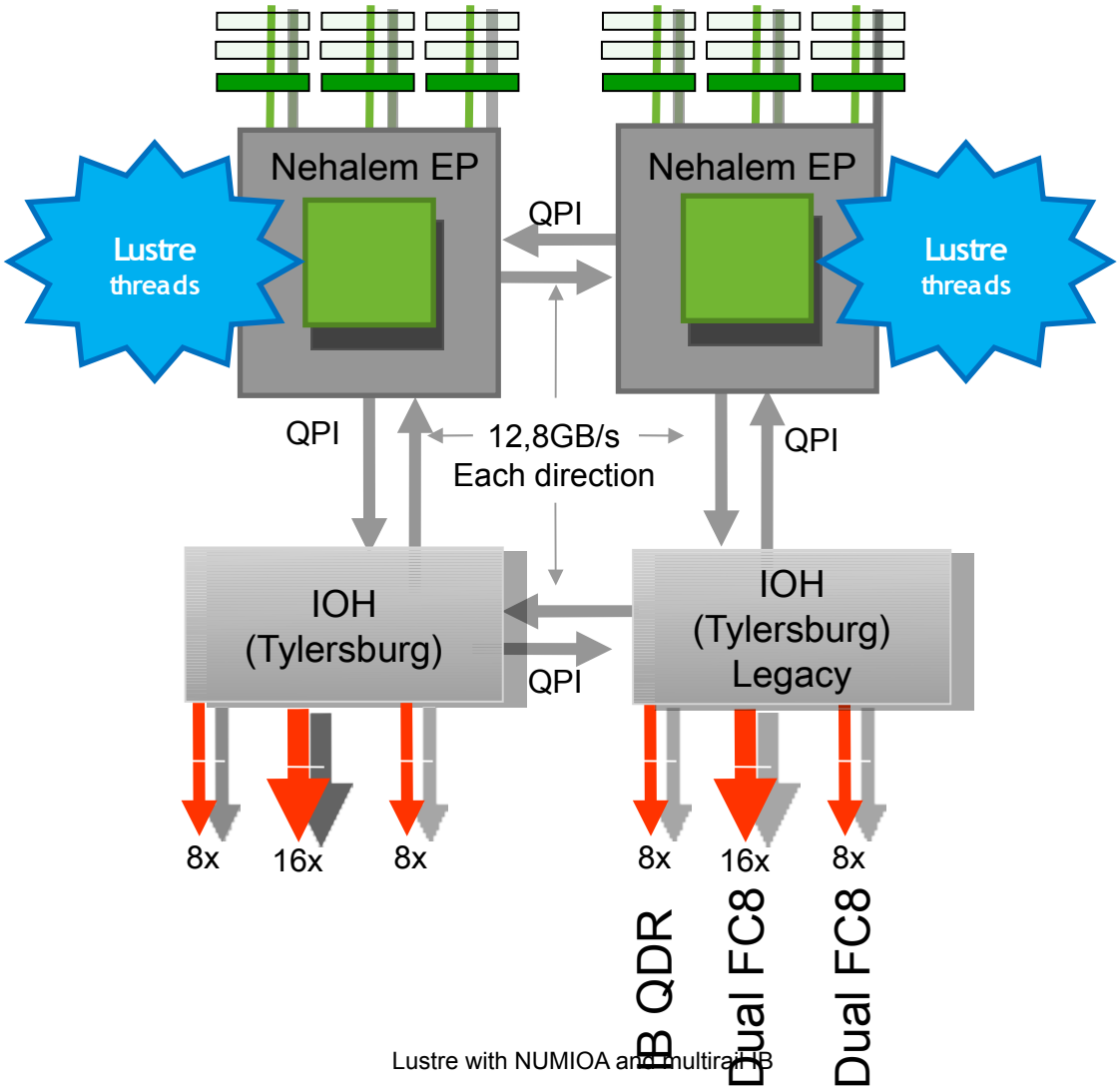
- "real life" case: customer cluster



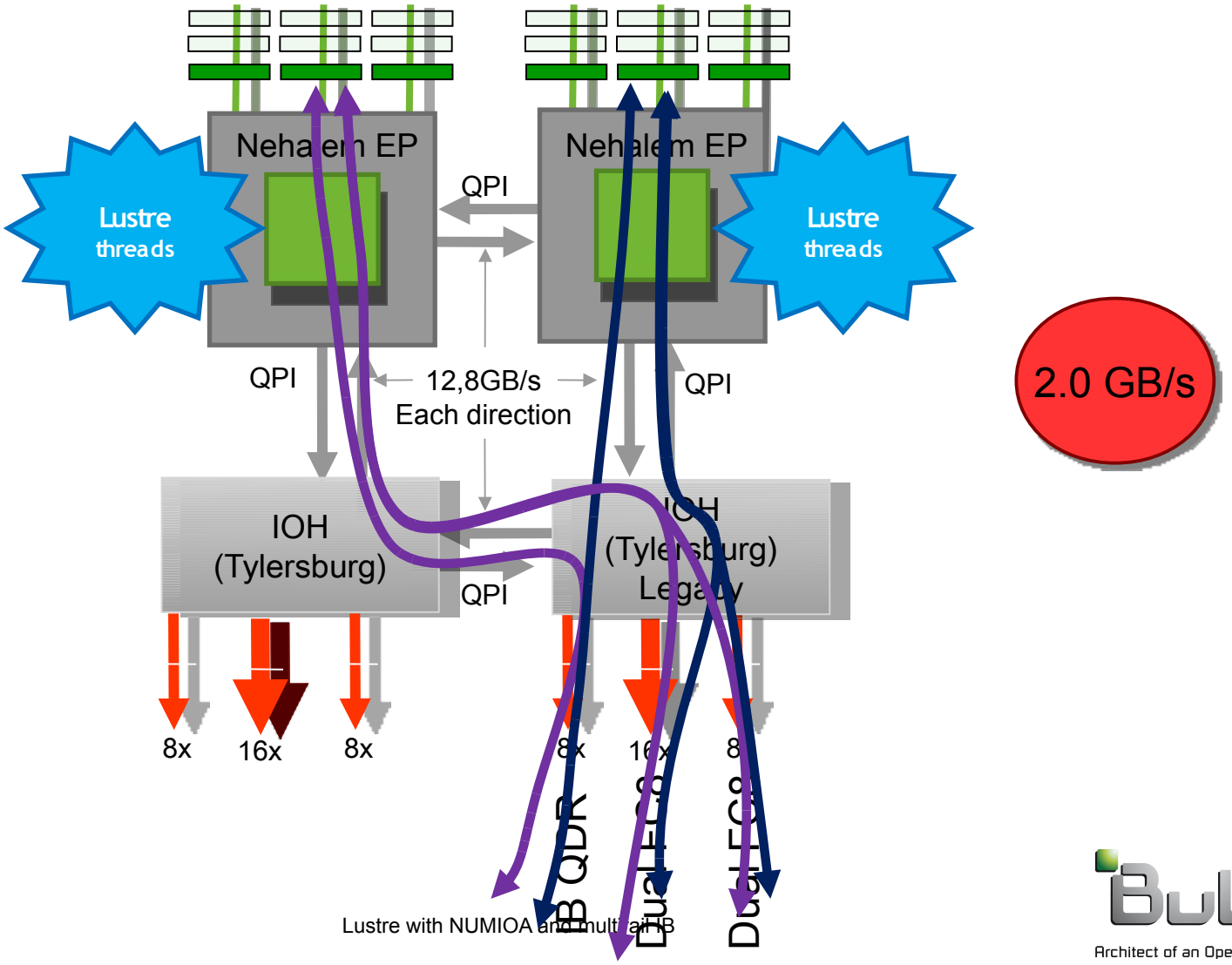
How does Lustre behave on a NUMIOA server?

- With Lustre 1.6:
  - One Compute Node gets **3 GB/s**
    - OK, network bandwidth on the client is the limiting factor
  - Two Compute Nodes get **4.0 GB/s**
    - expected 5 GB/s or more
  - What's happening in the OSS servers ?
    - **N**on **U**niform **IO A**ccess
    - **N**on **U**niform **M**emory **A**ccess

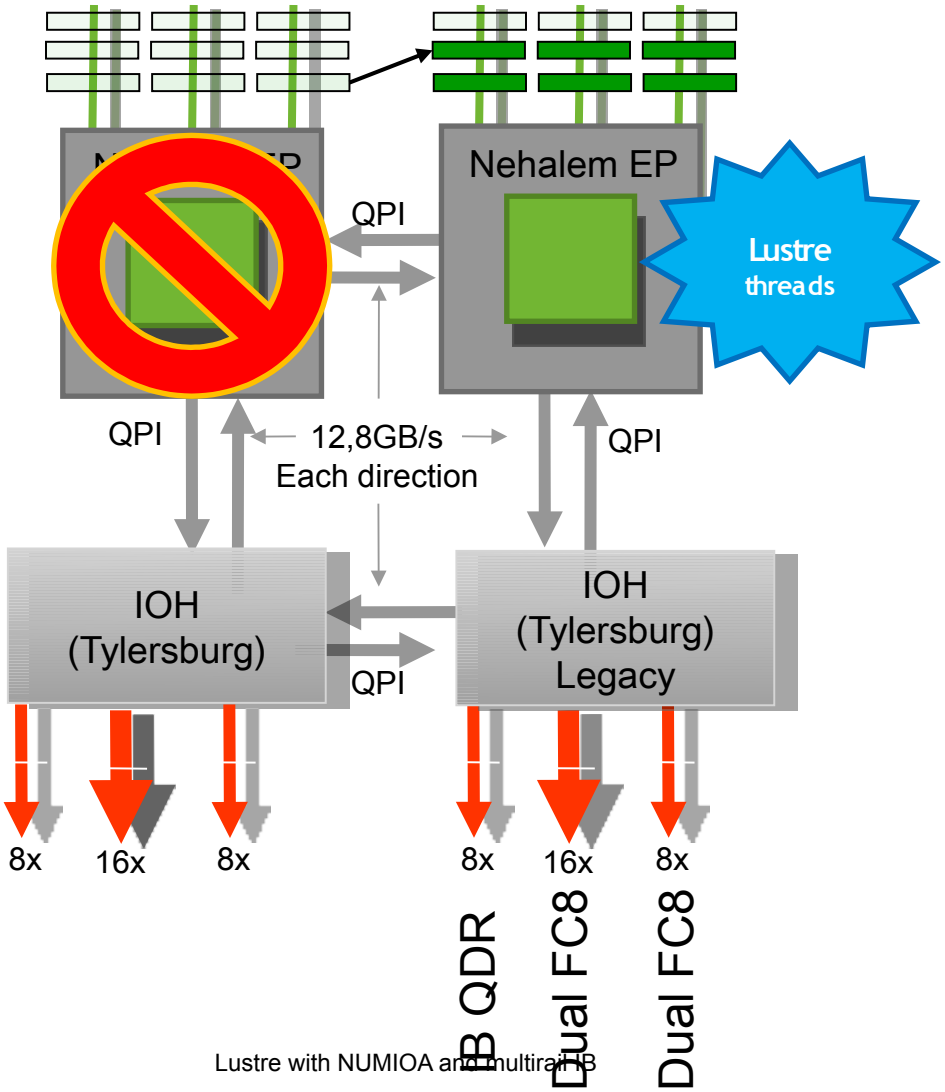
NUMIOA phenomenon



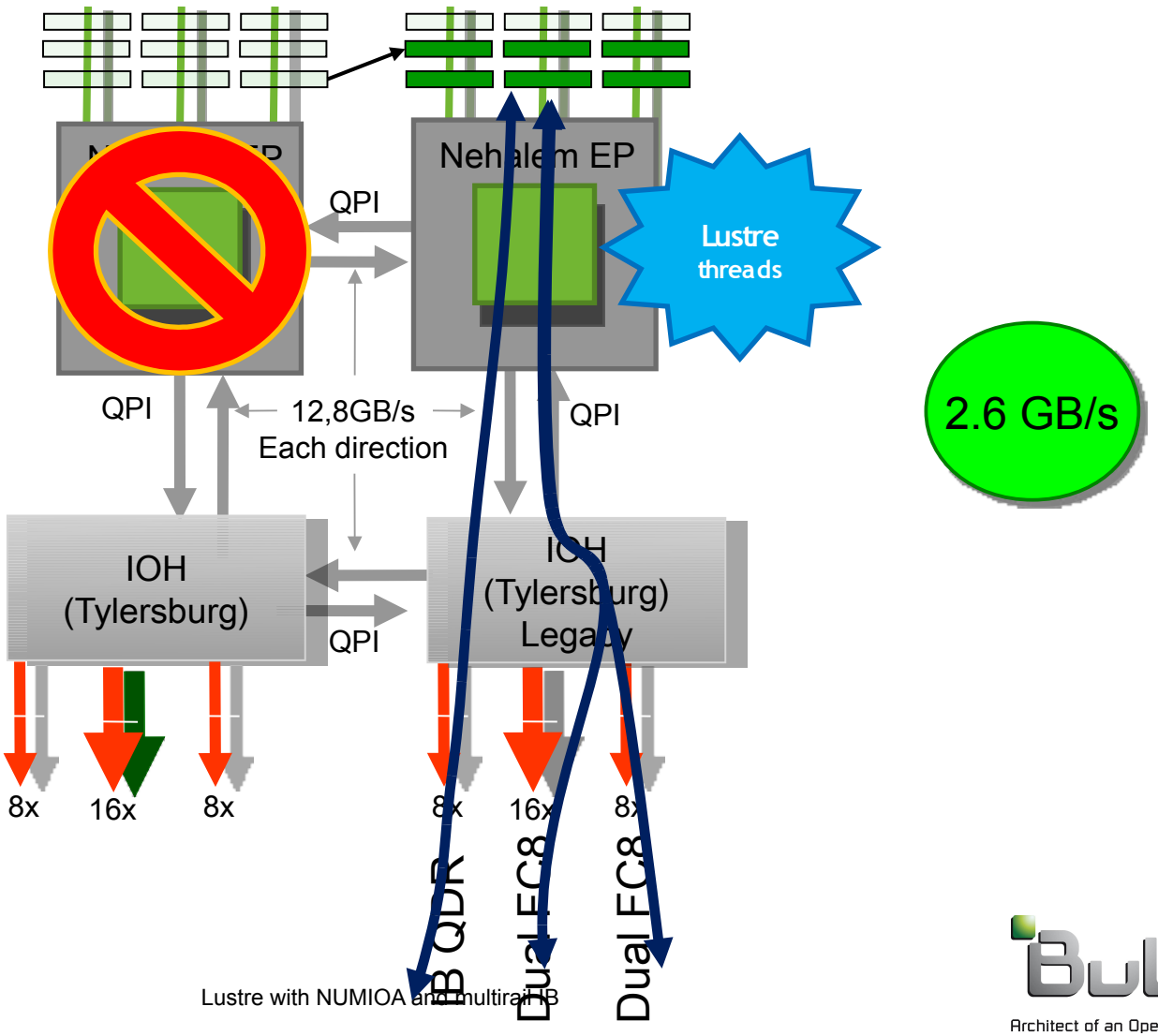
NUMIOA phenomenon



NUMIOA workaround

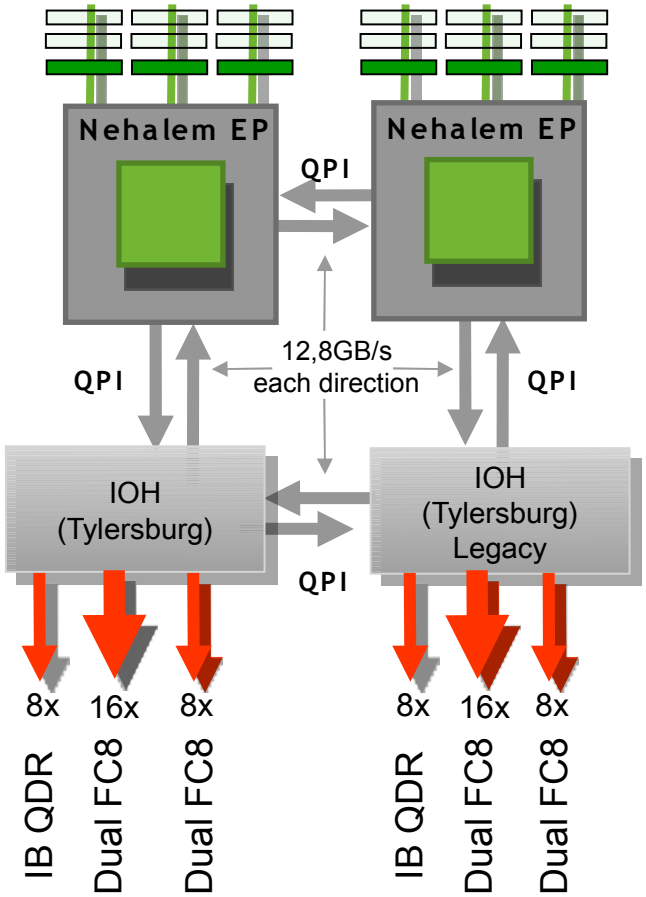


NUMIOA workaround

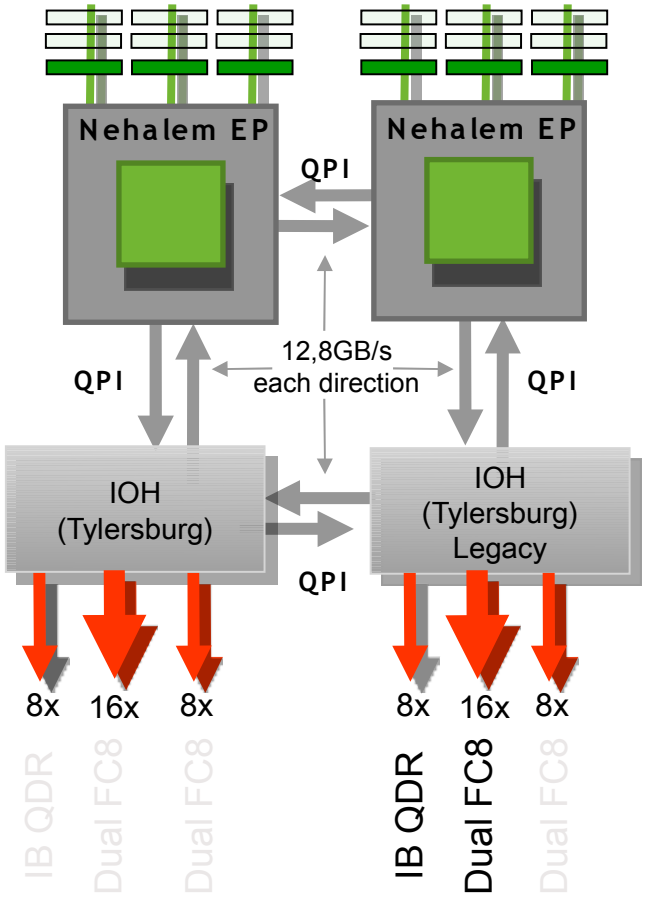


How does Lustre behave on a NUMIOA server with multiple IB interfaces?

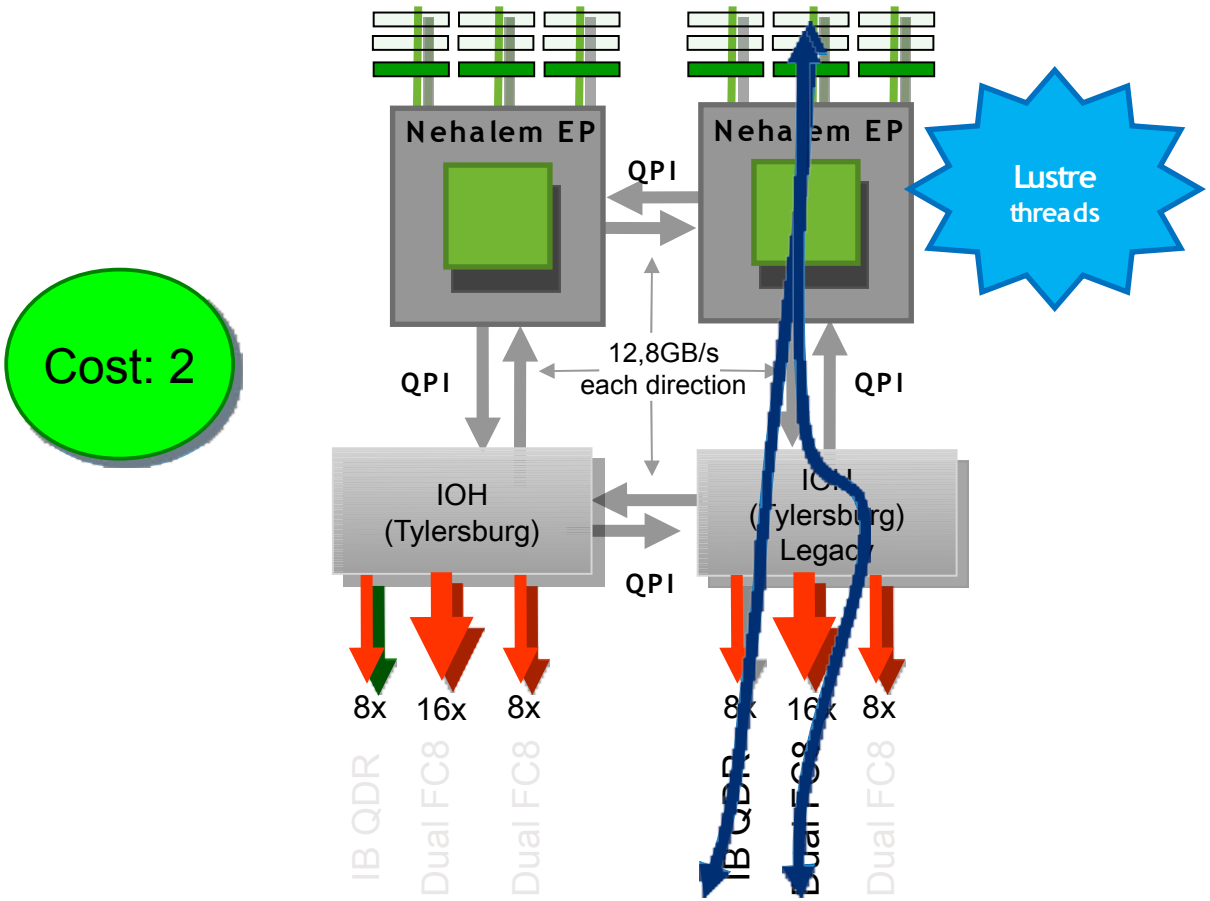
- R&D experiments with Lustre 2.0:



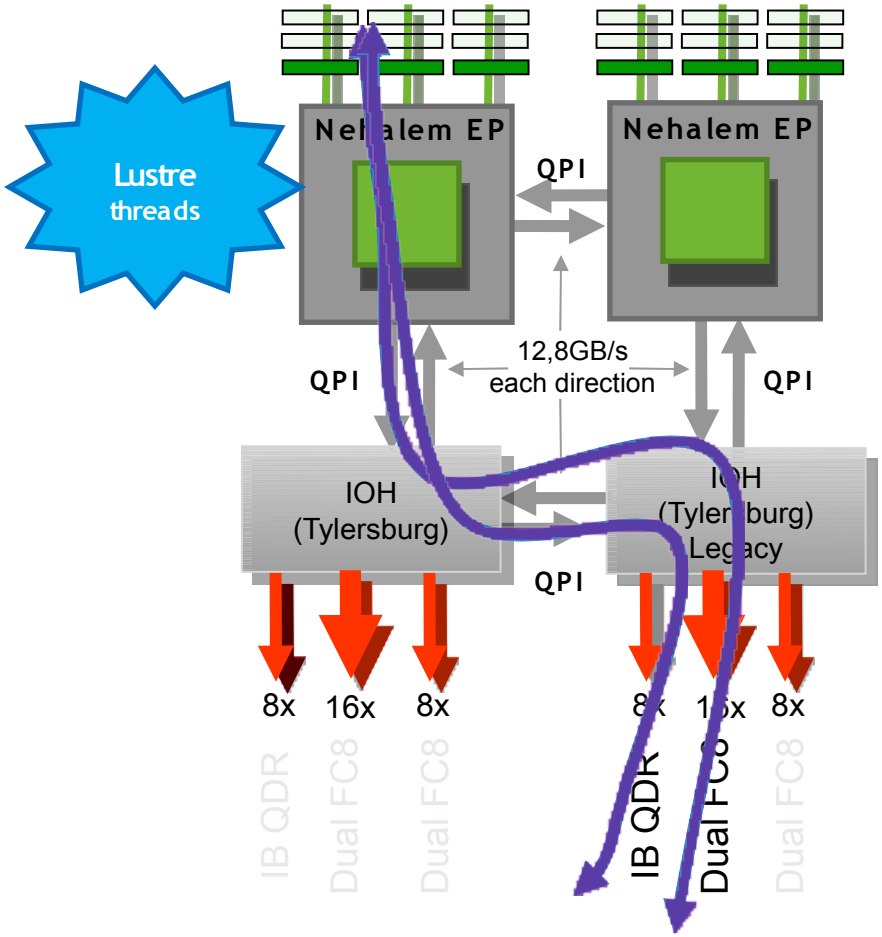
How does Lustre behave on a NUMIOA server with multiple IB interfaces?



How does Lustre behave on a NUMIOA server with multiple IB interfaces?

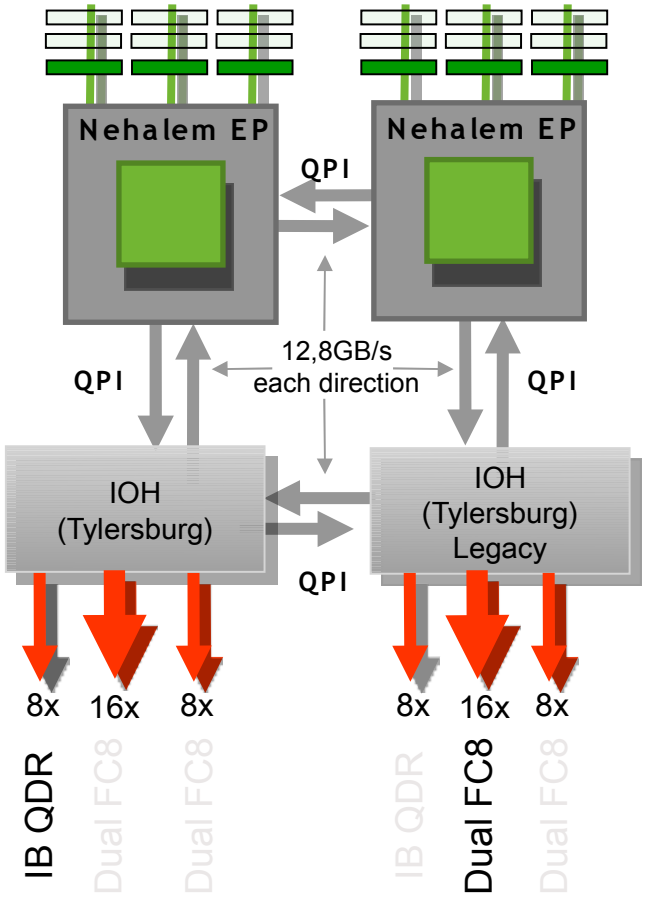


How does Lustre behave on a NUMIOA server with multiple IB interfaces?

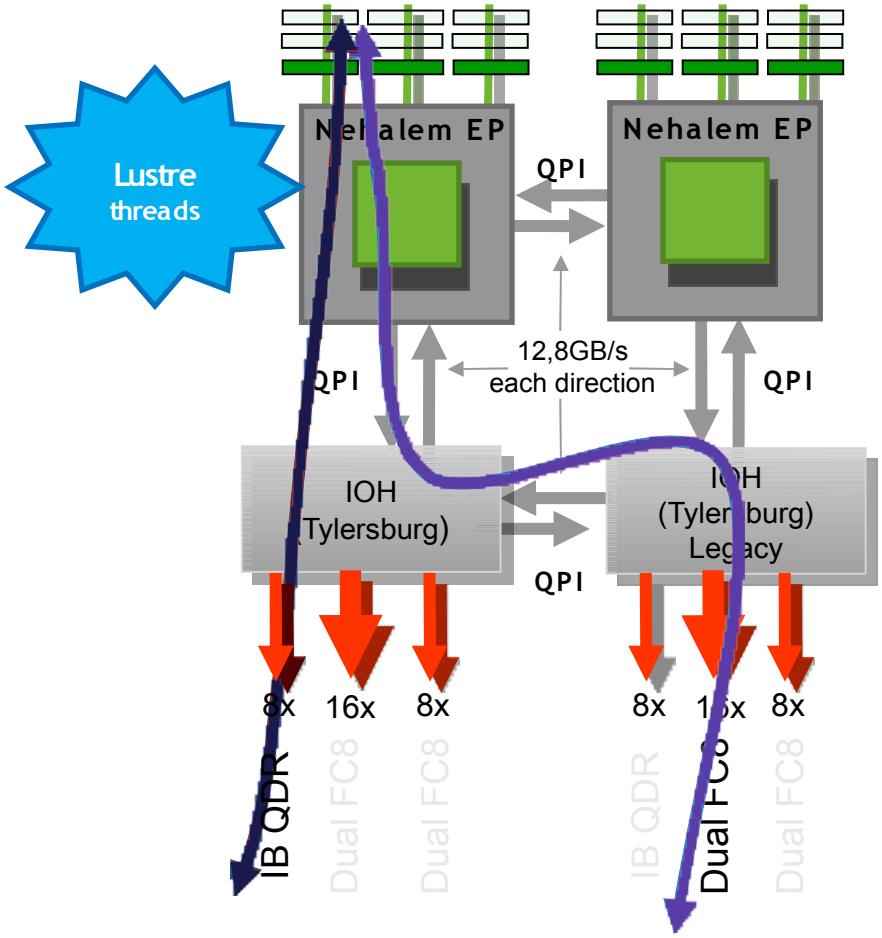


Cost: 4

How does Lustre behave on a NUMIOA server with multiple IB interfaces?

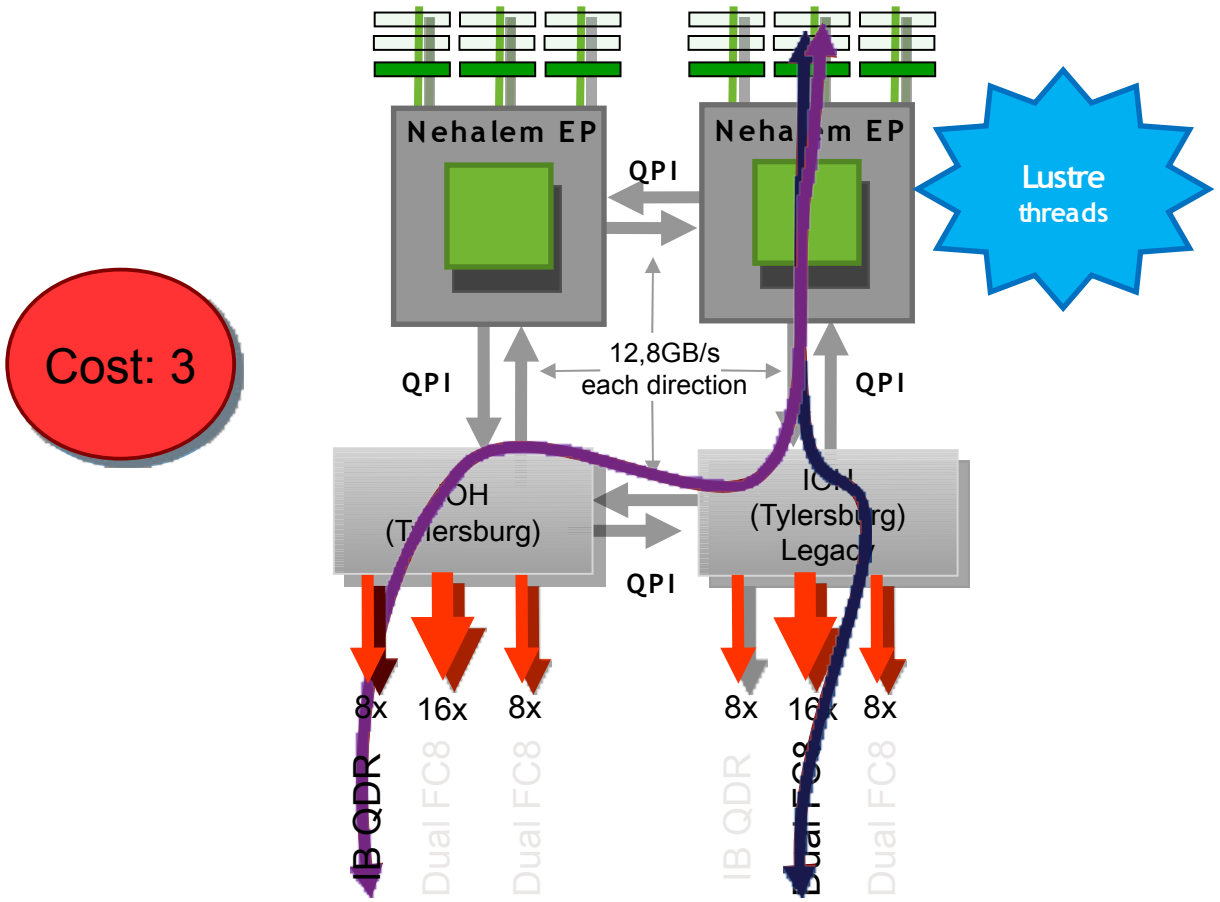


How does Lustre behave on a NUMIOA server with multiple IB interfaces?



Cost: 3

How does Lustre behave on a NUMIOA server with multiple IB interfaces?



Evolutions for the multirail part

“get rid of the 'cost 3' cases”

- An OST must be bound to a unique NID.
- With the current Lustre code
  - an OST can be reached via all the network interfaces available on the OSS
- It is a problem on NUMIOA platforms
  - Avoiding NUMIOA factor => choosing the "good" interface
    - “good”: network adapter connected to the same IOH as the FC adapter that gives access to the LUN.

Evolutions for the multirail part

Patch Lustre code: **bug 22078**

- Giving the ability to restrict the NIDs that a target (MDT or OST) registers on the MGS.
  - force client requests to go through the desired network interface
- *--network* option of mkfs.lustre and tunefs.lustre controls target binding
  - takes as values one or more LNET networks.

mkfs.lustre --ost --fsname=numafs --mgsnode=inti5@tcp0

--network=o2ib0,tcp0 /dev/sdc

- At *mount.lustre* time:
  - target registers on the MGS with only the NIDs that pertain to the specified networks.
  - If no specific network is given, all NIDs are registered.

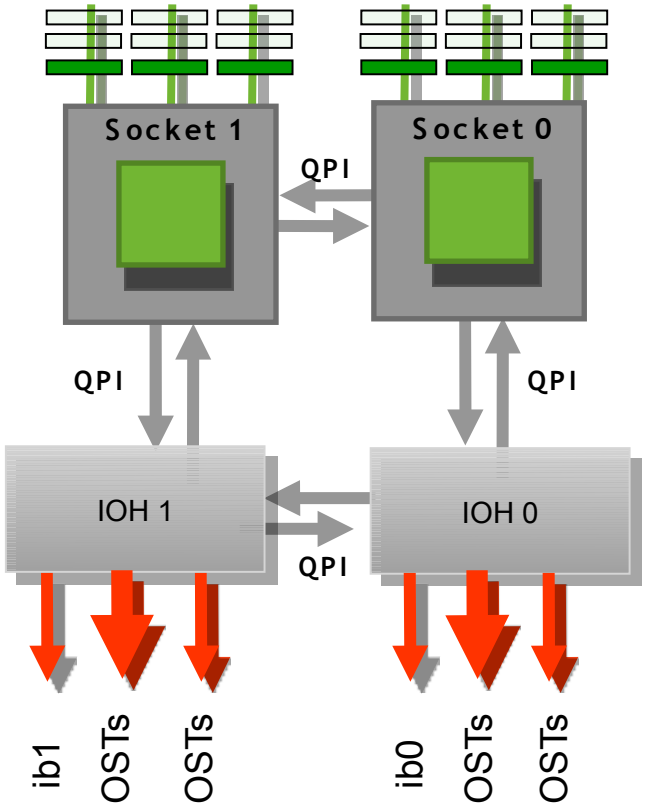
Evolutions for the NUMIOA part

“get rid of the 'cost 4' case”

- Current Lustre code is already NUMA aware.
- But this is not enough
  - Lustre code must be NUIOA aware.

- b_hd_numa branch for SMP scaling improvements (Liang Zhen's work)
  - **Bugs 19411 and 19412**
    - options libcfs cpu_numa=1
    - options lnet ni_affinity=1

Application to the OSS server



Application to the OSS server

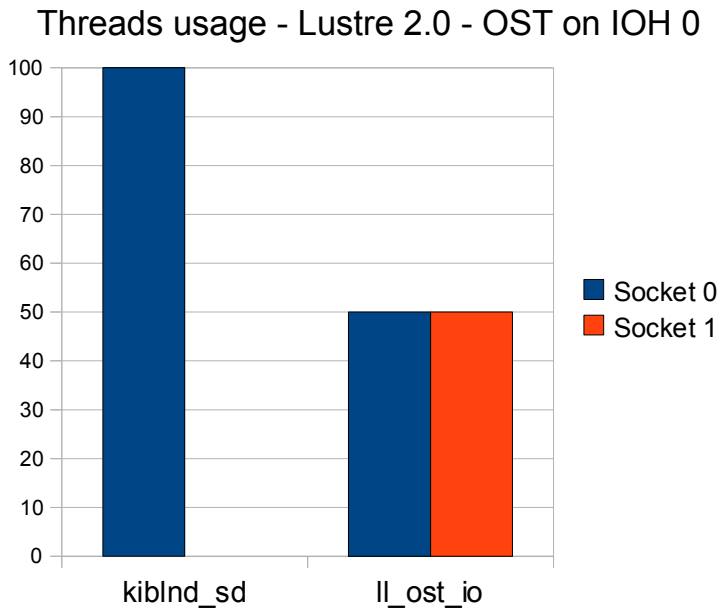
- With current Lustre 2.0 code

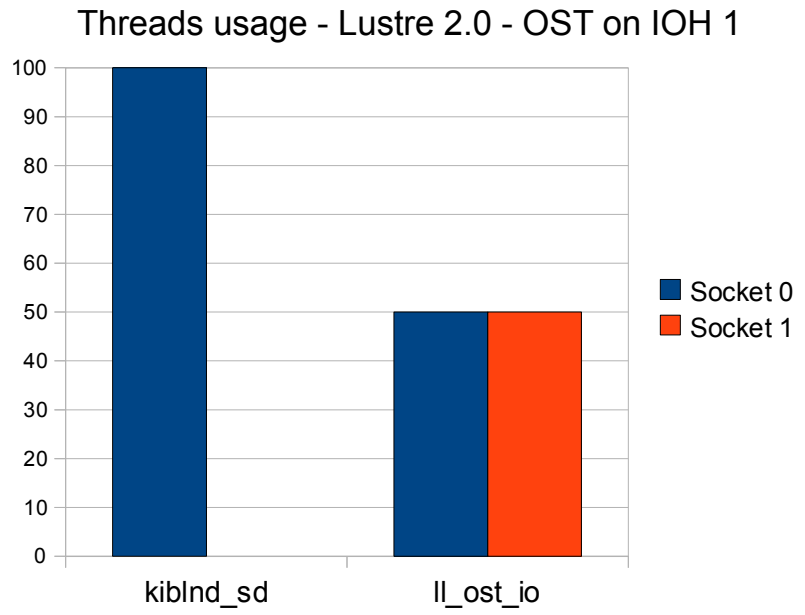
options lnet networks="o2ib0(ib1),o2ib1(ib0)”

--network=o2ib1 for OSTs on IOH 0

--network=o2ib0 for OSTs on IOH 1

2≤cost≤4





Architect of an Open World™

Application to the OSS server

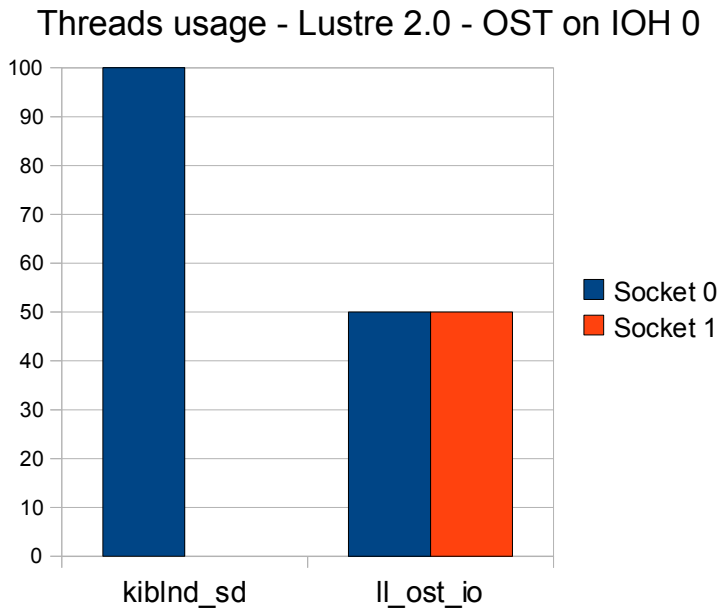
- With current Lustre 2.0 code

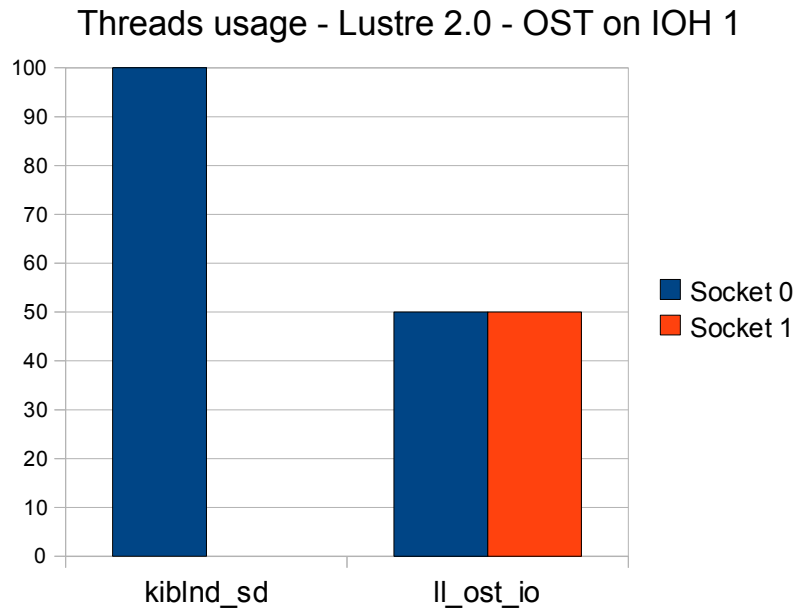
options lnet networks="o2ib0(ib1),o2ib1(ib0)”

--network=o2ib0 for OSTs on IOH 0

--network=o2ib1 for OSTs on IOH 1

Cost: 3





Architect of an Open World™

Application to the OSS server

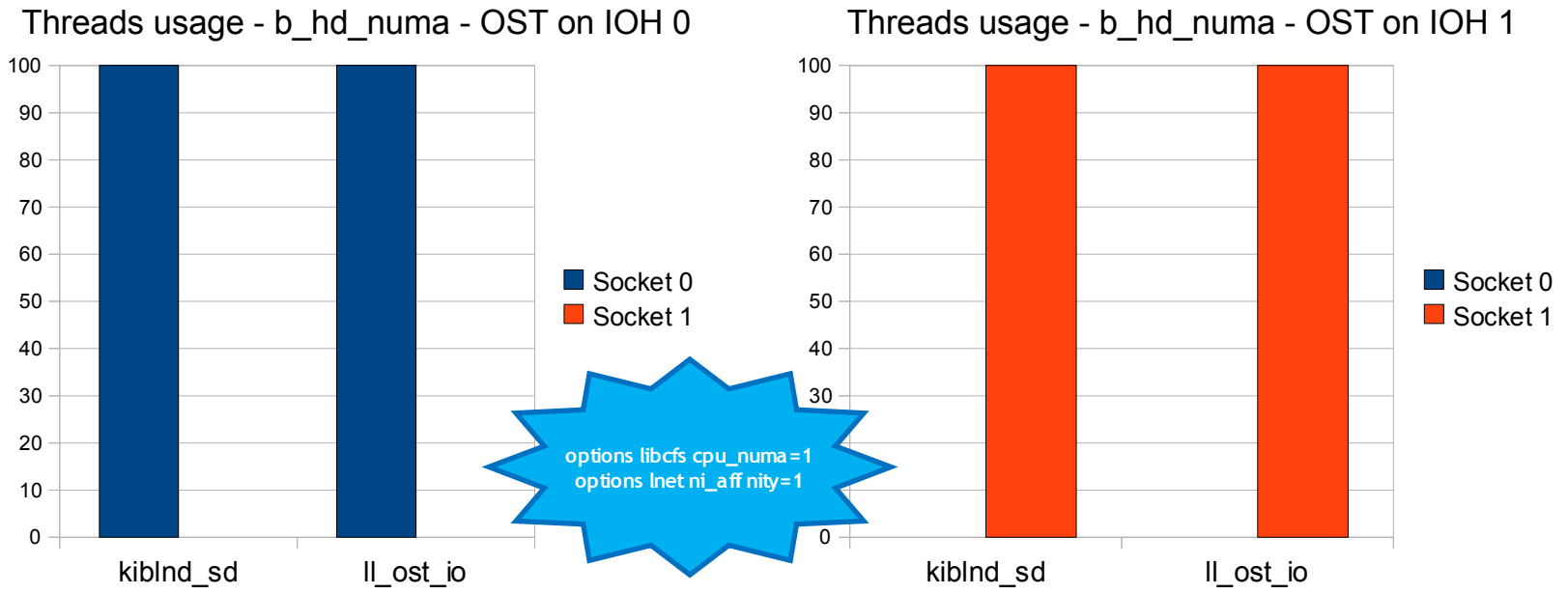
- With b_hd_numa branch

options lnet networks="o2ib0(ib1),o2ib1(ib0)”

--network=o2ib1 for OSTs on IOH 0

--network=o2ib0 for OSTs on IOH 1

Cost: 2



Application to the OSS server

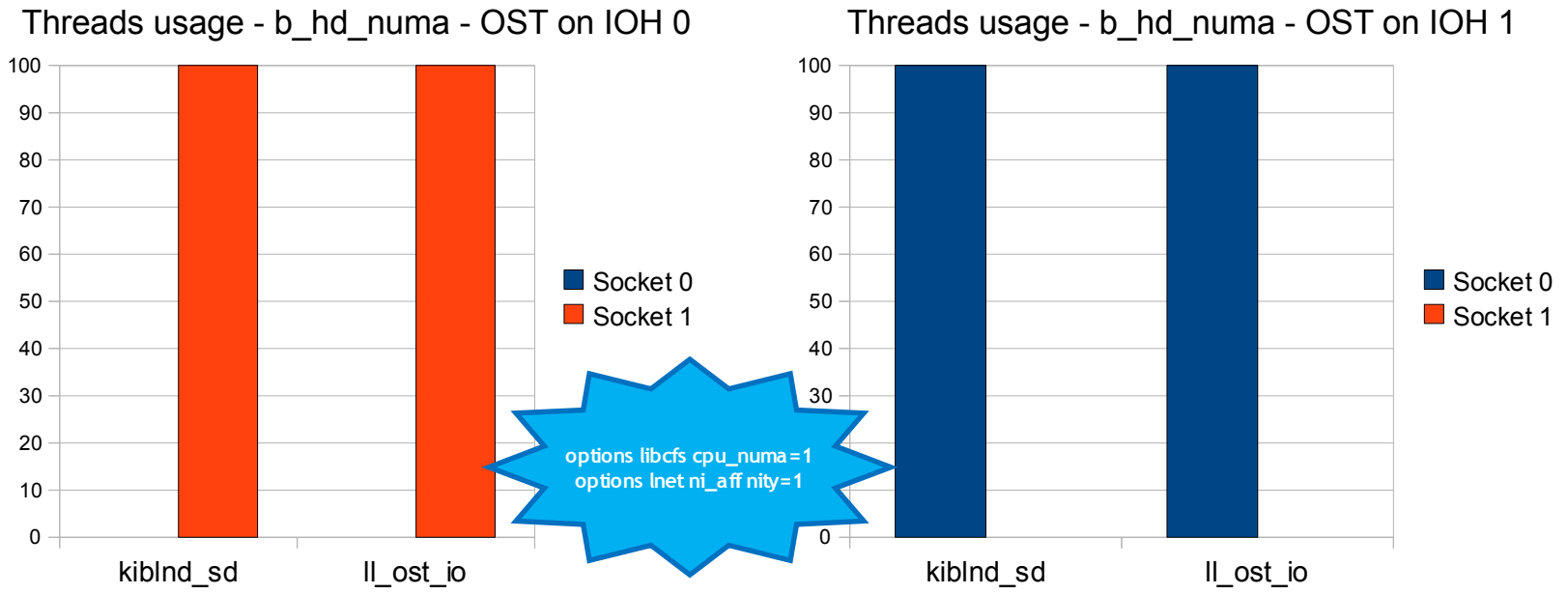
- With b_hd_numa branch

options lnet networks="o2ib0(ib0),o2ib1(ib1)”

--network=o2ib0 for OSTs on IOH 0

--network=o2ib1 for OSTs on IOH 1

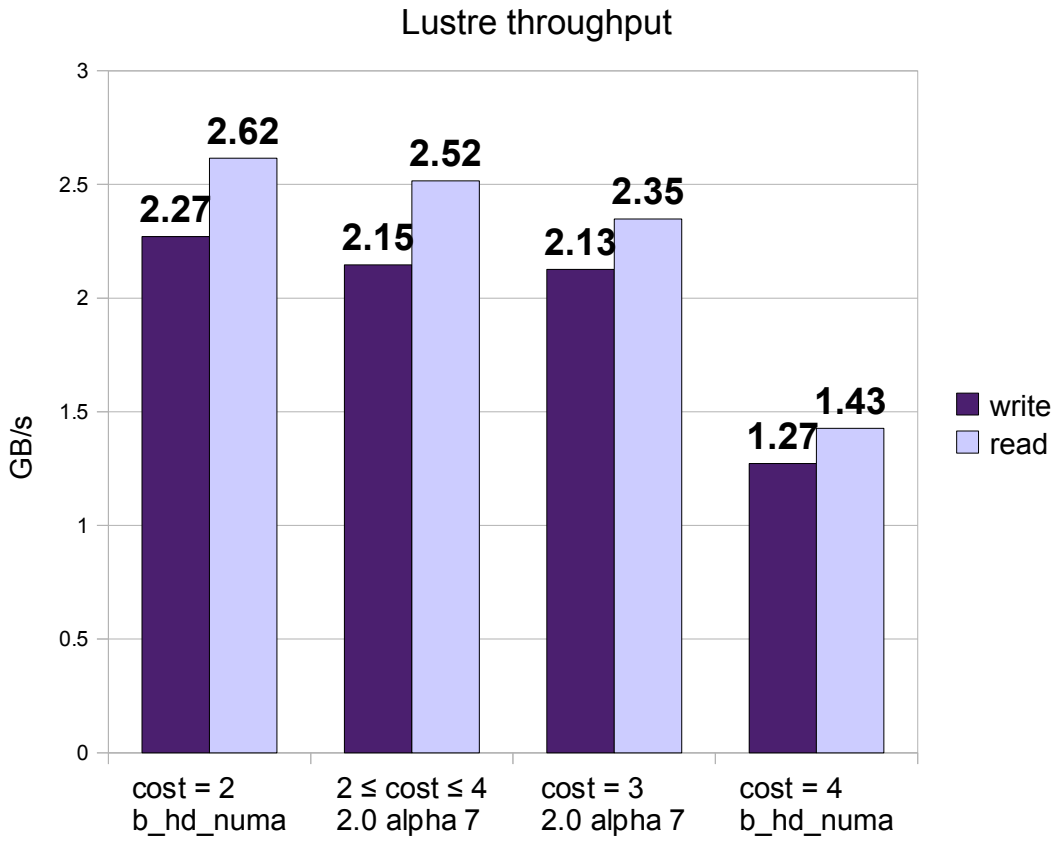
Cost: 4



Architect of an Open World™

Application to the OSS server

- Outcome:



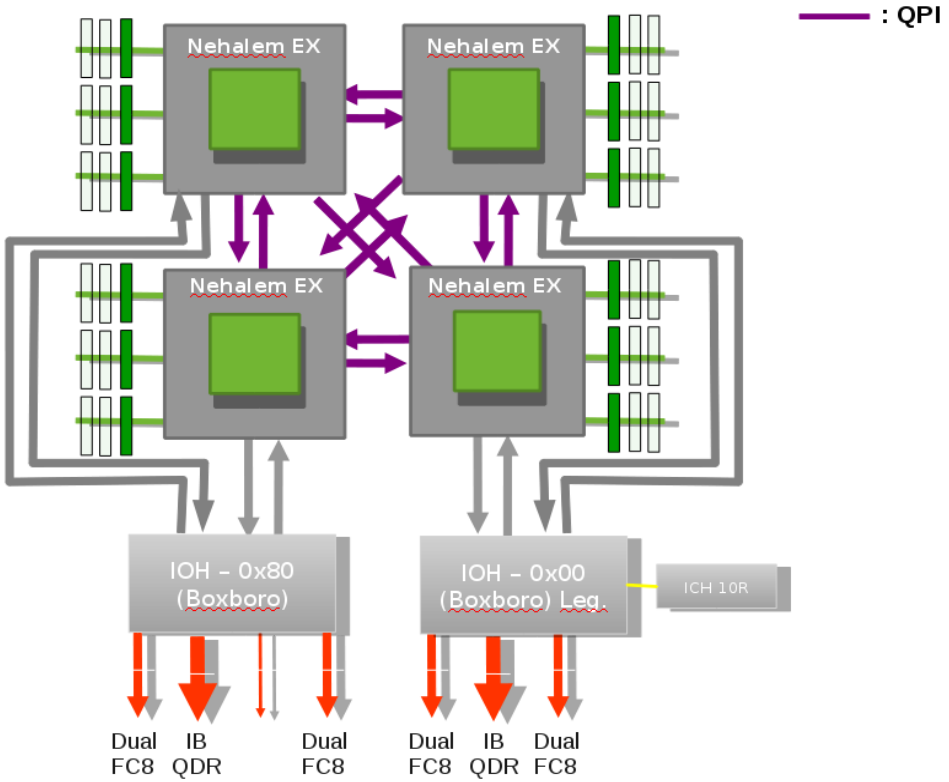
Application to the OSS server

BUT:

- With Lustre 2.0
  - We use only ib0
  - Threads are spread among socket 0 and socket 1
  - Depending on target and thread, cost can be 2, 3 or 4
  - **2.2 GB/s in write, 2.5 GB/s in read**
- With obdfilter-survey
  - **3.5 GB/s in write, 4 GB/s in read**
- Explanation for performance saturation:
  - Problem in QPI management with Nehalem EP and Tylersburg on this platform

Coming next

- Bull MESCA 4S server

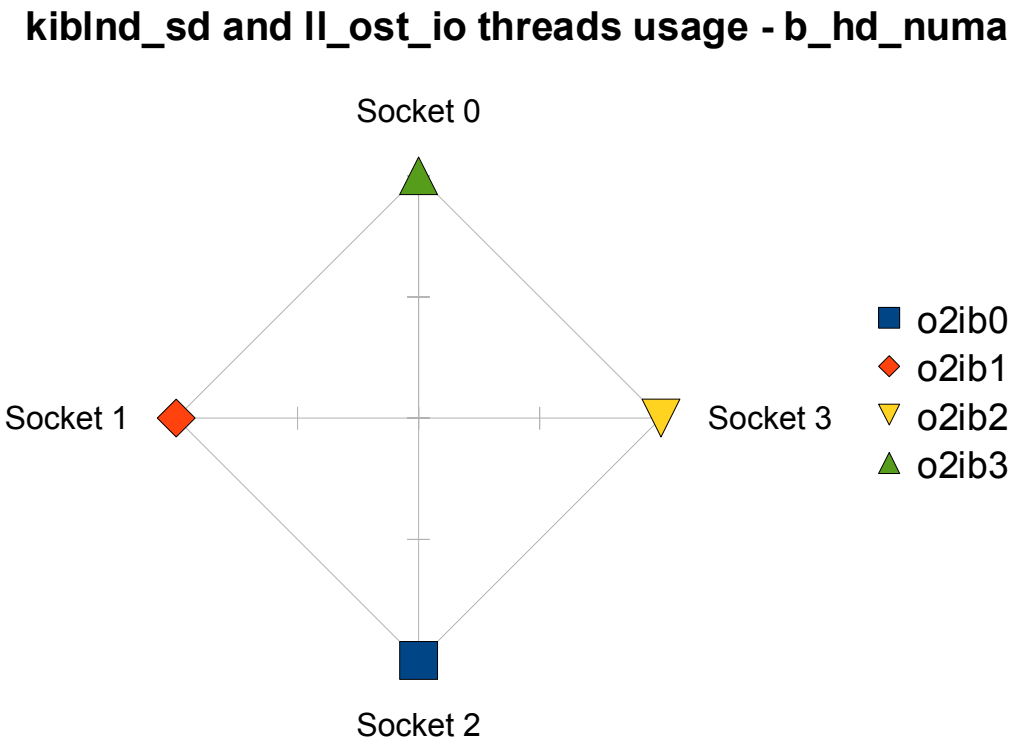


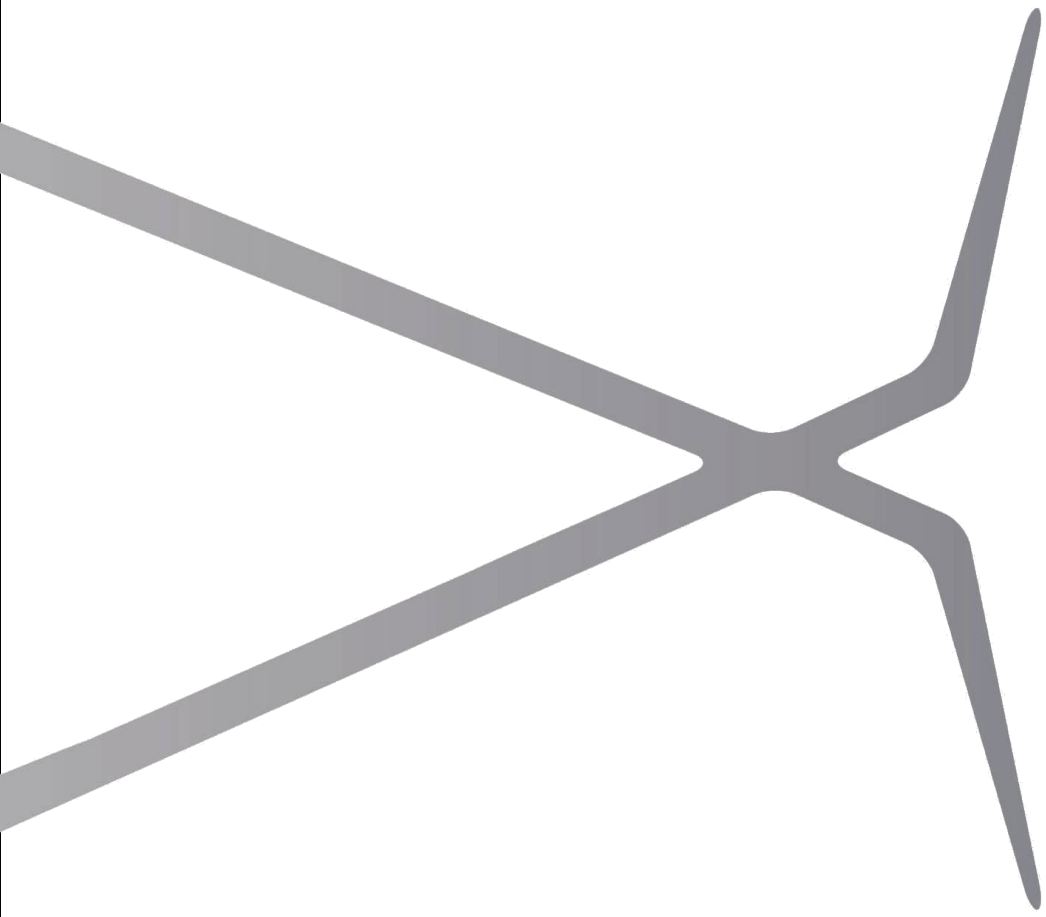
Bull MESCA 4S server

- Elementary performance tests
  - xdd for disk IOs
  - qperf for network IOs
  - Show that QPI management is better
    - we can reach QPI maximum bandwidth
- Will see with Lustre!

Lustre with NUMIOA and multirail IB

- What about a manual thread binding?





bullx

instruments for innovation