LUG 2011 – Lustre 2.0 and NUMIOA architectures

Diego.Moreno@bull.net
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment
- NUMIOA performance:
  - Lustre routers
  - Lustre servers
  - Lustre clients
- Improvements and future challenges
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment
- NUMIOA performance:
  - Lustre routers
  - Lustre servers
  - Lustre clients
- Improvements and future challenges
In the HPC:

- CEA's Tera100 supplier (TOP500's #6, Europe's #1)

- Our HPC strategy:
  - Based on big servers and big compute nodes:
    - Intel four-socket servers
    - Intel four-socket compute nodes
    - Bull's BCS: compute nodes with up to 16 sockets
  - Aimed to:
    - Scalability
    - 'Easy' HA management
    - Powerful compute nodes
In Lustre

- Integrating Lustre 2.0 in our releases
- Discovering and fixing bugs
- Multirail configuration:
  - Lustre servers: up to 6.4 GB/s bandwidth
  - Lustre clients: more than 10 GB/s bandwidth (Bull's BCS)
- Adapting Lustre to new NUMA architectures
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment
- NUMIOA performance:
  - Lustre routers
  - Lustre servers
  - Lustre clients
- Improvements and future challenges
What is NUMIOA?

Non-Uniform Memory Access + Non-Uniform IO Access = Non-Uniform Memory and IO Access
What is NUMIOA?

A 'single' NUMIOA machine with 2 sockets:

--- delimits a physical node
What is NUMIOA?

Non-Uniform Memory Access

--- delimits a NUMA node
What is NUMIOA?

Non-Uniform IO Access

--- delimits a NUIOA node
What is NUMIOA?

**Non-Uniform Memory and IO Access**

--- delimits a NUMIOA node
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment

NUMIOA performance:
- Lustre routers
- Lustre servers
- Lustre clients

- Improvements and future challenges
Problem to avoid

Lustre threads

Nehalem EP

QPI

IOH (Tylersburg)

12.8GB/s
each direction

IOH (Tylersburg)

Legacy

Nehalem EP

QPI

I B Q D R

D u a l F C 8

IB QDR

Dual FC8

Dual FC8

Dual FC8

8x

16x

8x

8x

16x

8x

Lustre 2.0 and NUMIOA
Lustre 2.0 and NUMIOA

*It can always get more difficult...*
NUMIOA: 4 sockets system (Intel's 7500 series)

Intel's Nehalem-EX (7500 series) 4-Socket server:
Bull's patches (bz#22078) taking advantage of several Infiniband interfaces on the same server or client:
- for bandwidth aggregation

Our goal:
- Lustre network bandwidth = \[ \sum \text{links individual bandwidths} \]

How:
- Servers are seen on the network as two different OSSs
Multirail configuration

Multirail needs a proper configuration to minimize NUMIOA:

- An OST must be bound to a unique NID
- Avoid NUMIOA factor => choose the "good" interface
  - “good”: network adapter connected to the same IOH as the FC adapter that gives access to the LUN
Multirail configuration
Bull in HPC

What is NUMIOA?

Lustre in our environment

NUMIOA performance:
- Lustre routers
- Lustre servers
- Lustre clients

Improvements and future challenges
Bad multirail configuration

Example with Lustre routers:
- Bad configuration:

Lnet routing:

- o2ib0 → o2ib2
- o2ib1 → o2ib3

Global rate: 1.4 GB/s

Crossed paths between IOHs
Bad multirail configuration

Example with Lustre routers:
- Bad configuration:

**LNet routing:**
- o2ib0 → o2ib2
- o2ib1 → o2ib3

**Global rate:**
1.4 GB/s

**Memory contention!**

Crossed paths between IOHs
Proper multirail configuration

Example with Lustre routers:
- Proper configuration:

Lnet routing:

- o2ib0 → o2ib2
- o2ib1 → o2ib3

Global rate: 5 GB/s

No crossed paths between IOHs
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment

NUMIOA performance:
- Lustre routers
- Lustre servers
- Lustre clients

- Improvements and future challenges
Obdfilter-survey is a tool used to benchmark storage systems and OSSs.

Single test not involving network (case=disk)

First tests:

![Graph showing MB/s for different numbers of active sockets (1-4) on obdfilter-survey on 4S machines.](chart.png)
Obdfilter-survey is a tool used to benchmark storage systems and OSSs.

Single test not involving network (case=disk)

First tests:

Bug with 4 and 3 sockets! (on 4S systems)
Memory contention on obdfilter-survey threads (bz#22980, jira#66)

Several patches developed on Bull's initiative and integrated in Lustre 2.1

General code also affected by these patches
obdfilter-survey bug on NUMIOA

- Memory allocation on sgppd-survey has to be optimized
Problem statement on NUMIOA servers

- If lustre threads are localized on one side then:
  - IOs are not balanced
Problem statement on NUMIOA servers

- If localized on both sides...
  - Lustre threads are not 'NUMA aware':
    - Some of them will be 'NUMA local'
    - Some of them will be 'NUMA remote'

```
Remote access
Local access
```
It should be the solution to NUMIOA problems on Lustre servers.

- Remote access
- Local access
SMP optimizations

- It should be the solution to NUMIOA problems on Lustre servers

- Code developed by Liang Zhen (Whamcloud), jira#56
  - Mainly lnet and ptlrpc patches:
    - Lnet binding on cpu is localized on the right NUMA node
  - MDD improvements

- Lustre threads will be 'NUMIOA intelligent'

- Tests of this branch has just started: still no results
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment

NUMIOA performance:
- Lustre routers
- Lustre servers
- Lustre clients

- Improvements and future challenges
Performance on NUMIOA lustre clients

- 4-sockets compute nodes:
  - Clients are also NUMIOA machines
  - Very powerful as compute nodes
  - But...
  - IO asymmetric
  - Lustre performance is impacted
Performance on NUMIOA lustre clients

Iozone tests on different kind of machines:

<table>
<thead>
<tr>
<th></th>
<th>2-sockets clients</th>
<th>4-sockets clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lustre 2.0.0.1</td>
<td>1.95 GB/s</td>
<td>[1.3 – 1.8] GB/s</td>
</tr>
</tbody>
</table>

Iozone write

Lustre performance not stable on 4-sockets platforms

Trying to run Lustre threads with numactl command doesn't improve results

SMP optimizations should solve any instability on performance
Lustre 2.0 in NUMIOA architectures

- Bull in HPC
- What is NUMIOA?
- Lustre in our environment
- NUMIOA performance:
  - Lustre routers
  - Lustre servers
  - Lustre clients
- Improvements and future challenges
Future HW challenges: BCS

Bull's Coherent Switch (BCS):
- Device designed by Bull
Bull's Coherent Switch (BCS):
- Device designed by Bull
- Interconnecting several nodes into one
Bull's Coherent Switch (BCS):
- Device designed by Bull
- Interconnecting several nodes into one
- XCSI fast link between BCS
Future HW challenges: BCS

Bull's Coherent Switch (BCS):
- Device designed by Bull
- Interconnecting several nodes into one
- XCSI fast link between BCS

Creates a super compute node:
- 128 cores in 16 sockets (4 modules model)
- Interesting for many applications needing big compute nodes

![Diagram of Bull's Coherent Switch (BCS)]
Future HW challenges: BCS

- BCS is located between IOH and sockets
- IOs become symmetric
- Now NUMA distance can be very high, so...
- Lustre threads localization is mandatory
- Lustre bandwidth could be more than 12 GB/s
Still to do...

- Test lustre 2.1
- Test and improve SMP optimizations code:
  - On servers: trying to maximize IOH isolation on Lustre
  - On clients: trying to maximize Lustre localization on 2 sockets
- Tuning recommendations for NUMIOA systems
- Improvements on Lustre benchmarking tools
- BIOS optimizations can still be done
- BCS integration