Parallel I/O on JUQUEEN and JURECA

Hardware Overview

14.03.2016, Parallel I/O Workshop

Michael Stephan
JSC Supercomputer I/O Infrastructure

- JUQUEEN
- Ethernet Network
- JURECA
- JUST
Hardware View
JUQUEEN: Jülich’s Scalable Petaflop System

- IBM Blue Gene/Q JUQUEEN
- IBM PowerPC® A2 1.6 GHz,
  - 16 cores 4way SMT per node
  - 28 racks (7 rows à 4 racks)
  - 28,672 nodes (458,752 cores)
- 5D torus network
- 5.9 Pflop/s peak
  5.0 Pflop/s Linpack
- Main memory: 448 TB
- I/O Nodes: 248 (27x8 + 1x32)
  - 496 x 10GigE
JURECA: Jülich’s Multi-Purpose System

- Jülich Research on Exascale Cluster Architectures
  - Vendor: T-Platforms
- 1884 compute nodes in 34 racks
- Dual-socket Intel Xeon E5-2680 v3 Haswell CPUs
  - 24 cores @ 2.5 GHz (up to 3.3 GHz boost)
  - 480 GFlop/s peak
- RDIMM DDR4 2133 MHz
  - 128 GiB in 1605 thin nodes
  - 256/512 GiB in 196 fat nodes
- Mellanox EDR InfiniBand (100 Gb/s)
  - Two-level non-blocking fat tree
  - Two Gateway switches for storage connectivity
    - 2x 18 port FDR/40GigE gateway ⇒ 100 Gb/s to JUST
JUST: GPFS Storage Cluster

- JuElIch Storage Cluster
- GPFS-Filesystems
  - $\text{WORK}$
    - Capacity: $5.3 \text{ Pbyte, up to 200 GB/sec}$
  - $\text{HOME (/home[a-c])}$
    - Capacity: $3 \times 600 \text{ Tbyte, 12 GB/sec}$
  - $\text{ARCH (/arch, /arch2)}$
    - Capacity: $2 \times 600 \text{ Tbyte, 12 GB/sec}$
- 33 Building blocks:
  - Lenovo GPFS Storage Server solution
  - $2 \times \text{x3650 M4 server}$
  - $232 \text{ NL-SAS disks (2TB)}$
Network Infrastructure

- Cisco Nexus 7018 Switch
  - Number of slots: 18
  - Bandwidth per slot: 550 Gb/s
  - Maximum switching capacity: 17.6 Tb/s
  - Port density @ line-rate
    - 10GigE: 768 ports
    - 40 GigE: 192 ports
  - Redundant supervisors
Network Infrastructure - JUST

Diagram showing network infrastructure with GSS Bld. Block and Nexus 7k connected by discs.
Network Infrastructure - JURECA

[Diagram showing network connections and infrastructure details, including Nexus 7k, GSS Bld. Block, and various JURECA Infiniband Fabric nodes and gateways.]
Network Infrastructure – Final picture

- Storage-Cluster JUST
  - 64 Storage-Servers IBM x3650M4 a 6 x 10GE
  - 198 10GE-LAGs a 2 x 10GE, 7.6 TB/sec aggregated
  - 2 10GE-LAGs a 2 x 10GE

- Internet
  - Juniper-Cluster
  - 68 x 10GBaseT
  - Mellanox SX5036G
  - 3 x Cisco Nexus 7018 36 TB/sec actual deployed forwarding bandwidth

- IP 1
  - 2 x Mellanox SX6036G gateway switches with 3 proxy-arp instances per box
  - 6 x 40GE-LAGs a 6 x 40GE, 2.8 TB/sec aggregated

- IP 2
  - 2 x Mellanox SX6036G

- IP 3
  - 20GE-LAGs a 2 x 10GE

- IP 4
  - 24 x BlueGene Q nodes a 2 x 10GE LAGs, 9.9 TB/sec aggregated

- IP 5
  - BlueGene Q-SC JUQUEEN

- IP 6
  - Haswell-Cluster JURECA
  - about 1869 nodes with EDR-HCA alternated belonging to one of the three VLANS/pkeys
  - DataPlex GPU-Cluster JUDGE

- Campus Network
  - JuNet

- DEEP-Cluster
Blue Gene/Q: I/O-node cabling (8 ION/Rack)
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture.

© IBM 2012
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture (ioBridge is established at CN block boot time)
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture (ioBridge is established at CN block boot time)
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Rxx-M1

M1-N15
M1-N14
M1-N13
M1-N12
N12-J11
N12-J06

M1-N11
M1-N10
M1-N09
M1-N08
N08-J11
N08-J06

M1-N07
M1-N06
M1-N05
M1-N04
N04-J06
N04-J11

M1-N03
M1-N02
M1-N01
M1-N00
N00-J06
N00-J11

Rxx-ID

ID-J07
ID-J06
ID-J05
ID-J04

Rxx-M0

M0-N15
M0-N14
M0-N13
M0-N12

M0-N11
M0-N10
M0-N09
M0-N08
N08-J11
N08-J06

M0-N07
M0-N06
M0-N05
M0-N04
N04-J06
N04-J11

M0-N03
M0-N02
M0-N01
M0-N00
N00-J06
N00-J11

Note: This assumes an error-free machine; errors like missing IONs may change this picture (ioBridge is established at CN block boot time)

© IBM 2012
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture (ioBridge is established at CN block boot time)
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture (ioBridge is established at CN block boot time).
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture (ioBridge is established at CN block boot time)
Blue Gene/Q: I/O-node cabling (8 ION/Rack)

Note: This assumes an error-free machine; errors like missing IONs may change this picture.
BGQ: Tasks connected to same I/O-node

- MPIX_Calls available on BG/Q
  (see http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/UserInfo/MPIextensions.html)

- Communicator: All tasks belonging to same I/O Bridge Node

  **FORTRAN:**  MPIX_PSET_SAME_COMM_CREATE(INTEGER pset_comm_same,
                               INTEGER ierr)
  
  **C:**  #include <mpix.h>
          int MPIX_Pset_same_comm_create( MPI_Comm *pset_comm_same )

- Usage: implementation of own I/O strategy
  (One file per I/O-bridge)

- Passing new communicator to SIONlib
  paropen-Call (as local communicator)

  ```c
  ...  
  sid=sion_paropen_mpi( filename , "bw“,  
                      &numfiles, &chunksize,  
                      gcom, &lcom, &fileptr, ...);
  ...
  ```
Application View to Parallel I/O

Parallel application

HDF5

NETCDF

MPI-I/O

SIONlib

shared

POSIX I/O

Magic

Parallel file system

local
File I/O to GPFS on JURECA

Typical GPFS Architecture for HPC Clusters

Network (TCP/IP or IB)

Comp. node
Application
NSD client
NSD
GPFS server BB1
SAN
NSD

Comp. node
Application
NSD client
NSD
GPFS server BB2
SAN
NSD

Comp. node
Application
NSD client
NSD
GPFS server BBn
SAN
NSD

GPFS NSD Client
pagepool (streams)
prefetch threads

Network
transfer size

GPFS NSD Server
pagepool (disks)
NSD workers

OS Adapter / Disk Device Driver
hdisk dd
adapter dd

Storage Subsystem
SAN
Ctrl A
Ctrl B

© IBM 2012
File I/O to GPFS on JUQUEEN(I)

I/O-Forwarding

Application

BG/Q CN : CNK
BG/Q Network (dual 2GByte/s raw)
BG/Q ION : SYSIOD

2x BG/Q CN->ION links in (2GByte/s raw, 1.6GByte/s dir)

10 GbE Network

GPFS NSD
Server

OS Adapter / Disk
Device Driver

Storage Subsystem

© IBM 2012
File I/O to GPFS on JUQUEEN(II)

Parallel application

POSIX I/O

POSIX I/O

Parallel file system

I/O-Forwarding

Application

IO size

Parallelism

BG/Q CN : CNK

BG/Q Network (dual 2GByte/s raw)

BG/Q ION : SYSIOD

large_region_size

BG/Q ION: GPFS

NSD Client (64bit)

pagepool (streams)
prefetch threads

10 GbE Network

transfer size

GPFS NSD Server

pagepool (disks)

NSD workers

OS Adapter / Disk Device Driver

hdisk dd

adapter dd

Storage Subsystem

SAN

Ctrl A

Ctrl B
File I/O to GPFS on JUQUEEN(III)
GPFS Storage Server
GPFS Storage Server – Building Block

x3650 M4 Server

JBOD Disk Enclosure

**GSS 24: Light and Fast**
2 x3650 servers + 4 JBOD 20U rack

**GSS 26: HPC Workhorse**
2 x3650 servers + 6 JBOD Enclosures, 28U
GPFS Storage Server – Building Block

<table>
<thead>
<tr>
<th>GSS Model</th>
<th>JBODs (60 slots)</th>
<th>Disk Size</th>
<th>Gross Capacity</th>
<th>Spare Capacity</th>
<th>8+2P Net Capacity</th>
<th>8+3P Net Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSS-24</td>
<td>4</td>
<td>2 TB</td>
<td>464 TB</td>
<td>2 disks per DA</td>
<td>358 TB (4<em>56</em>2TB*(8/10))</td>
<td>326 TB (4<em>56</em>2TB*(8/11))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 TB</td>
<td>696 TB</td>
<td>(one DA has 58 disks)</td>
<td>538 TB (4<em>56</em>3TB*(8/10))</td>
<td>489 TB (4<em>56</em>3TB*(8/11))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 TB</td>
<td>928 TB</td>
<td></td>
<td>717 TB (4<em>56</em>4TB*(8/10))</td>
<td>652 TB (4<em>56</em>4TB*(8/11))</td>
</tr>
<tr>
<td></td>
<td>(GSS v1.5)</td>
<td></td>
<td></td>
<td></td>
<td>538 TB (6<em>56</em>2TB*(8/10))</td>
<td>489 TB (6<em>56</em>2TB*(8/11))</td>
</tr>
<tr>
<td>GSS-26</td>
<td>6</td>
<td>2 TB</td>
<td>696 TB</td>
<td></td>
<td>806 TB (6<em>56</em>3TB*(8/10))</td>
<td>733 TB (6<em>56</em>3TB*(8/11))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 TB</td>
<td>1044 TB</td>
<td></td>
<td>1075 TB (6<em>56</em>4TB*(8/10))</td>
<td>977 TB (6<em>56</em>4TB*(8/11))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 TB</td>
<td>1392 TB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(GSS v1.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GPFS Native RAID - Motivation

**Classical**

- Application
- GPFS File System
- External RAID Controller
- Physical Disk

**New approach**

- Application
- GPFS File System
- GPFS Native RAID
- Physical Disk

- Performance degradation on disk rebuild
- Silent data corruption

- Fast disk rebuild using Declustered RAID
- End-to-End data integrity (checksums, version no)
GPFS Native RAID (GNR) - Features

- End-to-end checksums against silent data corruption
- Declustered RAID
- RAID codes:
  - 3-way/4-way replication
  - 8+2p/8+3p Reed Solomon
- Recovery Groups (failover)
- Disk hospital
  - Diagnoses errors/faults in storage subsystem
Declustered RAID

- 3 groups with 6 disks each
- 1 spare disk
- 7 spare strips
- 21 stripes (42 strips)
- 7 disks
- 7 stripes per group (2 strips per stripe)
- 3 1-fault-tolerant groups
Declustered RAID Rebuild Example - Single Fault

Disk failure causes disk rebuild
- Volume degraded for a long time
- Performance impact for file system

Disk failure causes strips rebuild
- All discs involved
- Volume degraded for a short time
- Minimized performance impact