Parallel Computing at DESY Zeuthen.

Introduction to Parallel Computing at DESY Zeuthen and the new cluster machines

Götz Waschk

Technical Seminar, Zeuthen

April 27, 2010

- Introduction
- Hardware
  - Infiniband
  - M1000e blade center
  - Schematic view, blade center
  - Schematic view, storage network
- Software
  - Operating system
  - MPI versions
  - Batch System
  - Debugging support
  - Lustre
- User access
  - Batch jobs
  - Known Problems
- Summary

Parallel Computing at DESY

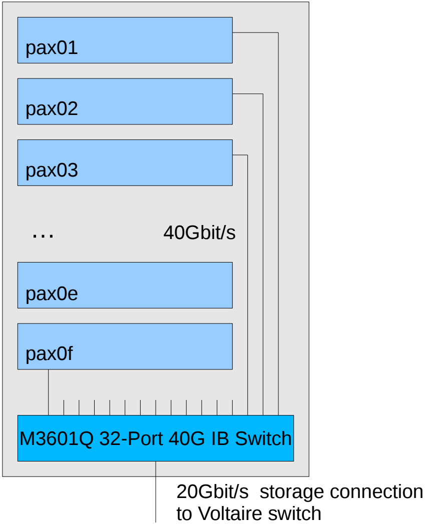
- apeNEXT Special Purpose Computer
- Local Batch Farm with slow 1G-Ethernet connections
- New Pax Clusters with Infiniband

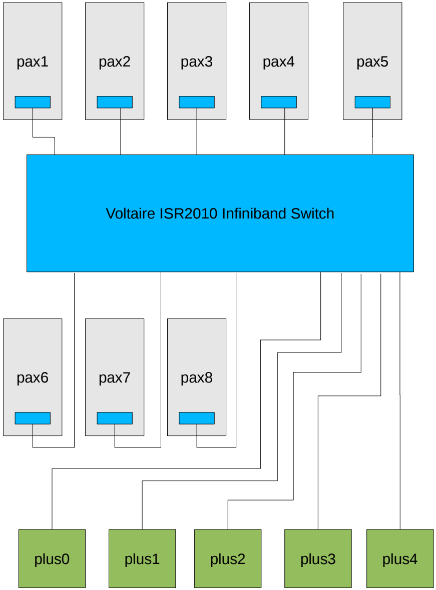
New cluster hardware

- Hardware installed in 1/2010
- 8 Dell PowerEdge M1000e Blade Centers
- M3601Q 32-Port 40G Infiniband Switches
- 16 Dell PowerEdge M610 Blade servers each
  - 2 quad-core Intel Xeon E5560 CPUs @ 2.8GHz
  - QDR 40 GBit/s Infiniband
  - 24 GB Main memory DDR3 (1.3GHz)
  - 2 × 2.5" SAS drives, 146GB, RAID0
- Total peak performance: 12 TFLOPS

Infiniband networking

- MPI communication
  - 40 GBit/s inside blade center
  - Connected to internal QDR Infiniband switch
- Storage network access to Lustre file system
  - 20 GBit/s connection per blade center
  - Connected to older Voltaire DDR Infiniband switch


pax01
pax02
pax03
...
40Gbit/s
pax0e
pax0f
M3601Q 32-Port 40G IB Switch
20Gbit/s storage connection
to Voltaire switch


pax1
pax2
pax3
pax4
pax5
Voltaire ISR2010 Infiniband Switch
pax6
pax7
pax8
plus0
plus1
plus2
plus3
plus4

Software

- Standard SL5.4 64 bit
- Same software as on desktop and farm
- Several MPI versions
  - OpenMPI
  - Mvapich/Mvapich2
  - Intel MPI test installation

Open MPI

- Open Source implementation of the MPI-1 and MPI-2 standards
- Versatile, supports many network types, batch systems
- Dynamic loading of plug-ins
- Comes with SL5.4

- Testing on your workgroup server/desktop possible
- Automatic selection of the right network transport ⇒ currently broken, use *mpirun --mca btl "ˆudapl"*
- Extra builds for Intel and PGI compilers ⇒ use *ini* to select the right version

Mvapich

- Mvapich/Mvapich2 are Infiniband ports of MPICH/MPICH2
- Needs MPD (multi purpose daemon) running on all nodes
- Not integrated with batch system
- Binaries only run on machines with Infiniband
- Comes with SL5.4 as well
- Builds for gcc and Intel compilers

Intel MPI

- Based on MPICH2
- Needs commercial or evaluation license
- Installed for testing
- No batch integration
- Supports both gcc and Intel compilers

Batch system

- SUN Grid Engine 6.2u5
- Tight integration of OpenMPI
  - OpenMPI’s mpirun uses qsub to start MPI processes
  - MPI processes are SGE tasks
  - All MPI processes have AFS token
  - All MPI processes run under SGE’s control
- Same SGE instance as used in the farm

Debugging support

- Currently, no parallel debugger installed
  - Might be purchased if demanded
  - possible choices: Intel Cluster Toolkit, Alinea DDT, Totalview
- Intel debugger 11.0 is available
- Valgrind with OpenMPI support is installed

Lustre file system

- Open Source parallel file system
- 1 Meta Data Server, 4 Object Storage Servers
- Version 1.8.2 test installation
- Advantages:
  - Scalable parallel access
  - High performance, > 500MB/s per file server
- Disadvantages:
  - Stability issues
  - Complicated administration
  - Unclear future since Oracle takeover
- Used as scratch and staging file system, *no backup!*

User access to the cluster machines

- Accessible by members of the nic, that and alpha groups
- Other users like PITZ or photon are welcome

- 2 blade centers as interactive machines: pax0 and pax1
- 6 blade centers in the batch farm

Batch job submission

Most important parameters:

- #$ -pe mpi-pax? 128
- #$ -R y
- #$ -l h_vmem=3G

Parallel jobs on the farm:

- #$ -pe multicore-mpi 8 for just 8 cores
- #$ -pe mpi 40 for larger jobs with low communication overhead

Known Problems

- Open MPI has a slow MPI_Sendrecv_replace on Infiniband [1]
- Batch system submission error: no suitable queues ⇒ reservations
- Unsolved hardware problems on some nodes ⇒ open Dell support issues
- No SGE integration of Mvapich/Intel MPI

[1] https://svn.open-mpi.org/trac/ompi/ticket/2153

Further reading

- https://dvinfo.ifh.de/Cluster/
- https://dvinfo.ifh.de/Batch_System_Usage/
- zn-cluster mailing list