HPC-Clusters at DESY Zeuthen

Götz Waschk
DESY Zeuthen
11/22/06

Contents

- Introduction to High Performance Clustering
- Cluster Hardware at DESY Zeuthen
- Software and infrastructure
- Boring technical details

HPC Definition

- High Performance Computing
- Solve problems of a big size
- Parallel programming
- Clusters of off-the-shelf processors

Other uses for Clusters:

- High Availability
- High Throughput

MPI

- Message Passing Interface
- API for parallel programs
- single program on several machines
- one or more MPI processes per node
- synchronization with messages
  - point to point
  - broadcast
  - operations on data, e.g. addition

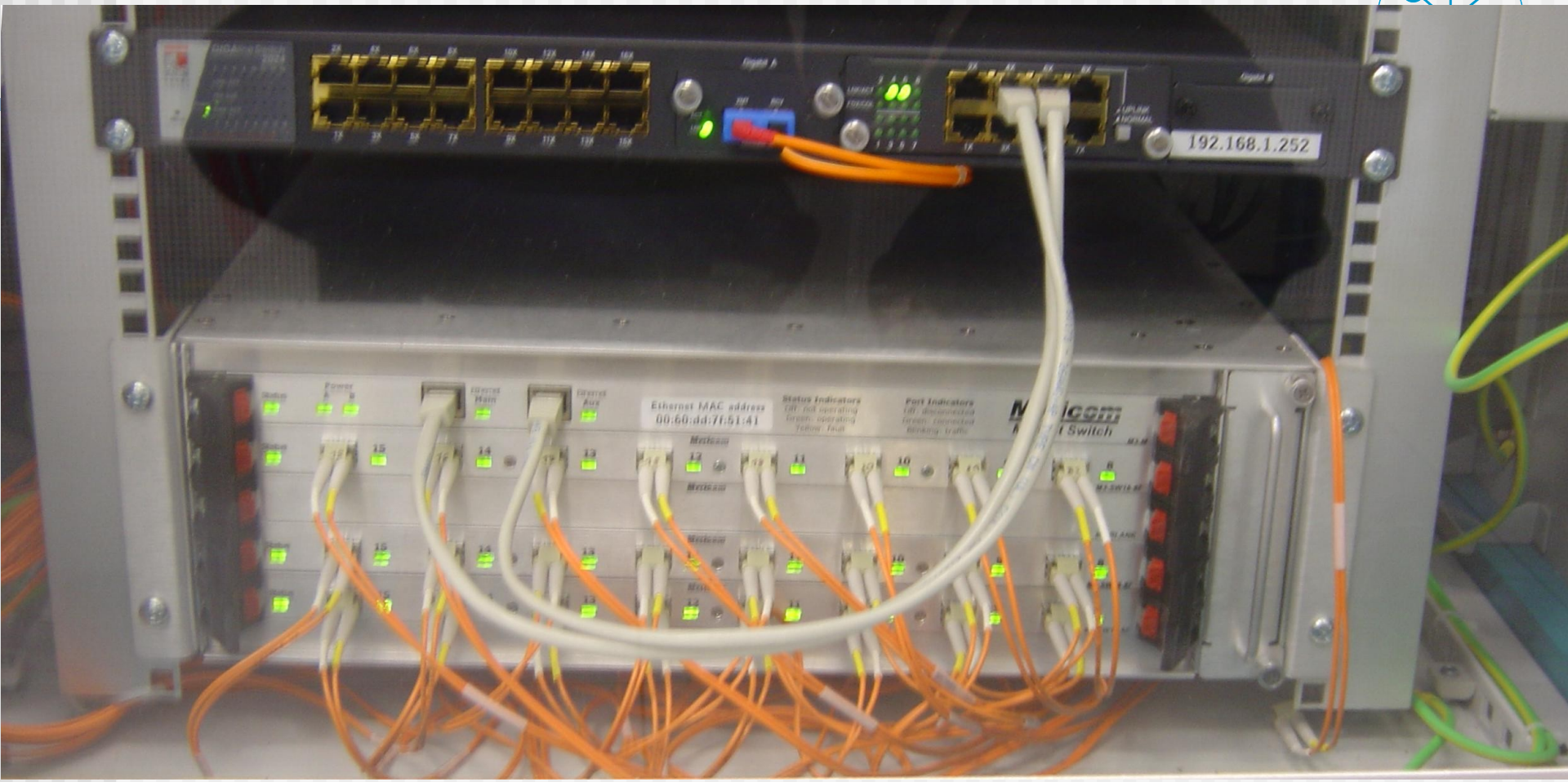
Cluster Hardware at DESY

- Common server machines connected by fast interconnect
- Optimized for high bandwidth and low latency

Geminide Cluster

- 24 nodes
- Each has two Intel Xeon 32 Bit CPUs, 1.7 or 2.0 GHz
- Myrinet 2000 network


geminide9
geminide10
geminide11
geminide12
geminide13
geminide14
geminide15


192.168.1.252
Ethernet MAC address
00:60:dd:7f:51:41

Plejade Cluster

- 16 SUN v20z servers
- 2 AMD Opteron 252 CPUs at 2.6 GHz in 64 Bit mode
- Mellanox Infiniband HA 4x network


PCI 0
PCI 1
MGMT
PCI 0
PCI 1

Previous Situation

- Private networks for the nodes
- Separate accounts at the head nodes (master, linfini)
- Ancient operating system, (S.u.S.E 7.2)
- NFS server on head nodes, but no AFS
- Batch systems (OpenPBS on master, extra installation of SGE on linfini)
- Clustware monitoring
- ➔ No longer maintainable

Now

- Nodes integrated into the DESY LAN
- Standard Scientific Linux 3 + all updates
- Integration into farm SGE batch system
- Global monitoring with Nagios
- AFS

- No dedicated head nodes
- No NFS

Software

- Same SL3 + packages as on the farm nodes
- Drivers for the Interconnect
- MPI implementations:
  - Vendor-adapted versions of MPICH 1
  - Geminide: mpichgm
  - Plejade: mvapich
  - Each in Variations for GCC, Intel and Portland Group compilers
- Special HPC libraries – at the moment only ATLAS

Sun Grid Engine

- Batch scripts like for farm jobs
  - Additional jobs script line for Parallel Environment selection:

```
#$ -pe mpich-ppn2 24
```

  - `mpich-ppn1` and `mpich-ppn2` for plejade, one or two processes per node
  - `mpichgm-ppn1` and `mpichgm-ppn2` for geminide

Grid Engine integration

Tight integration vs. loose integration

- Tight integration: all processes under SGE's control
- Loose integration: only process on the first node controlled by SGE

- We use loose integration

MPI usage

- Single binary program for all nodes in shared directory
- mpirun script started on node 0, starts all processes with ssh
- Only node 0 has AFS token
- Recommended job structure:
  1. Copy data from AFS to local file system
  2. Do calculations
  3. Collect results on node 0
  4. Write results back to AFS

Panasas file system

- Parallel file system
- File server + special software
- Linux kernel module
- Provides POSIX file system interface:

```
cd /panfs/waschk
ls
```

- Advantages: scalable with high performance
- Disadvantages: proprietary driver code, $$$


Panasas
Panasas ActiveScale
File System
2005
HPC STORAGE
PRODUCT OF TECHNOLOGY
PANASAS
2006
panasas
innovation
Los Alamos
NATIONAL LABORATORY
Fairfield

Summary

To run a parallel application you must:

1. Log into build machine, pub.ifh.de for 32 Bit, linfini for 64 Bit
2. Build application with the correct mpi compiler version
3. Write SGE job script
4. Submit it with qsub

more information at: http://dvinfo.ifh.de/Cluster