NEC SX8 / 576M72 @ HLRS

Operation and Usage

Thomas Beisel / Holger Berger (NEC)
beisel@hlrs.de / hberger@hpce.nec.com
Overview

• Configuration
  – System
  – Disk Subsystems

• USAGE
  – Workspace Mechanism short term Data
  – Batch Queues

• Accounting

• Problems
System Configuration

- Frontend System is single point of access
- 3 Partition types on SX-8 Nodes
System Partitions on NEC SX-8

- Interactive partition (1 Node)
  - Debugging
- Shared partition (4 Nodes)
  - Jobs < 8 CPUs
  - Unix time sharing operation
  - Charge for Resource Time (RT)
- Multi node partition (68 Nodes)
  - Jobs >= 1 Node
  - Dedicated operation
  - Charge for Wall clock * # Nodes
Frontend System

- Itanium II based server running Linux
  - Large Memory
  - High bandwidth connection to users
- Login Node
- Cross Compiling
- Performance Analyzer, Debugger
- Preprocessing
  - Grid generation
- Job preparation
- Job submission / monitoring
- Post processing
- Transfer of results to Users site
Disk Subsystem

- Local Filesystem
  - + dedicated to node
  - - jobs can’t migrate

- Global Filesystem
  - Accessible from all nodes and frontend
  - Easy monitoring of running jobs via frontend
  - 3 types of global Filesystem
  - Reconfiguration take place in 2Q 2006
  - Is not a long term storage
Global Filesystems

• NONE of the NEC Filesystems is for long term storage!

• 16 HOME each 1 TB
  – User Quota is 50 MB

• 16 scratch each 1 TB
  – This is NOT a long term storage

• ? scratch for large files
  – Only for files > 1 TB
  – Access on request only
  – Slow for small files

• Test of Next generation NEC Fileserver in progress
**Workspace Mechanism**

- Permanent disk space is ‘the´ limited resource of the NEC SX-8 installation
- Mechanism to allocate a temporary scratch-space for short term storage. Lifetime is some weeks (currently a view month but definitely less than a year)
- Will be removed automatically, but if not longer needed please release asap
- MPI-IO can stripe files on multiple Filesystems (Directories). This is supported by the workspace mechanism.
- Backup important Data into Archive
Usage of Workspace mechanism

• Allocation:
  MYSCR=`ws_allocate Temp_470K 10`
  Creates a directory which is available for 10 days

• Reference:
  WORK_DIR=`ws_allocate Temp_470K 10`
  This will not increase the lifetime of the Directory!

• List my workspaces
  ws_list
Batch Queues

- 4 Types of queues
  - **Test**  Jobs < 300s on shared Node
  - **Shared** Jobs < 8 CPUs RT-limit is 12h
  - **Multi**  Jobs >= 1 Nodes  Wall time limit 12h
  - **mighty** Jobs >= 32 Nodes Wall time limit 12h
    
    Gather Jobs for better System usage
    
    *Happy hour* take place 2 times a week

- Backfiller will rearrange Jobs in queues. This is to optimize System utilization.
  
  PLEASE  specify time limit with realistic value
Accounting

• Project based accounting / billing. Specify carefully within the Batch-Script the Project which is charged for.

• Set your default project in ~/.acct file
  
  ```
  $ cat .acct
  yyy0815
  $showacct -a
  hpc43598 yyy0815 zzz4711
  $
  ```

• Define project for Batch Jobs

  ```
  #!/usr/bin/ksh
  #PBS -q multi
  #PBS -l cputim_prc=07:55:00 # max accumulated cputime per process
  #PBS -A yyy0815 # Your Account code, see login message
  ```
Problems

In case of
- Problems
- Support request
- Special requirements
- ...

Please contact your project supervisor he/she will forward your request