Empowered by Innovation NEC

SX-ACE Usage at HLRS, batch and filesystem policy

Chief System Analyst Holger Berger
holger.berger@emea.nec.com
NEC Deutschland GmbH

Access

- Access is given on request to all interested parties, not restricted to germany, decision is up to HLRS
- Usual HWW security conditions, access by SSH only
- For pricing and conditions, contact HLRS
- Login via frontend: kabuki.hww.de

Installation outline

- **64** nodes SX-ACE (one cabinet)
- 1 Node
  - 64GB RAM/256GFlops peak
  - 4 Cores 1Ghz, each 64GFlops
  - 256GB/s memory bandwidth
  - 4GB/s MPI bandwidth (single lane)
- IXS connected nodes, non blocking 2 stage fat tree network for high bisectional bandwidth
- One 10G ethernet IO connections for filesystem, no local usable disk (only OS, by iSCSI)
- One frontend, 2 socket Ivy Bridge class server for compilation and file transfers

Filesystems

- Home filesystem is shared with other machines like Cray XC-40 hornet and NEC Laki/Nehalem Cluster
- NFS home quota is small and should not be used for job data
- Scratch filesystem is NEC ScaTeFS, about 250TB
- Shared between frontend and computer nodes
- 2 servers (shared data and metadata) using NEC FC connected storage with 18 data targets with 3TB disks and 18 metadata targets using fast 200GB disks
- For usage of scratch filesystem, use the *workspace* mechanism

Workspaces

- A workspace is a directory with limited lifetime
- ws_allocate <name> <time>
  name: something you remember
  time: up to 30 days
- Will create a new directory and print it's name
- ws_find <name> returns the path to existing workspace
- ws_release <name> marks it for deleation
- ws_list shows existing workspaces

- Why: to prevent old stalled data nobody cares off anymore, allows administration to keep disk usage under control without using quotas

Batch job example, 2 nodes with MPI:

```
#!/usr/bin/bash
#PBS -b 2                          # number of nodes
#PBS -l elapstim_req=12:00:00      # max wallclock time
#PBS -j o                          # join stdout/stderr
#PBS -T mpisx                      # Job type: mpisx for MPI
#PBS -N MyJob                      # job name
#PBS -M MyMail@mydomain            # you should always specify
your email
```


Contents of job on 2 nodes with 4 threads

```
SCR=`ws_allocate MyWorkspace 2`
cd $SCR
export OMP_NUM_THREADS=4        # is default anyhow
export MPIPROGINF=YES
export MPIMULTITASKMIX=YES
export MPIEXPORT="OMP_NUM_THREADS"
mpirun -nn 2 -nnp 1 $HOME/bin/mycode
# or mpirun –nn $_MPINNODES –nnp 1 $HOME/bin/mycode
```


Interactive jobs

- qlogin –q aiq
  you get a one hour job (max) on one node

- This is a bit slowish, you do not want to edit files in that, use frontend
- Do no paste large amount of data, it will get stuck (will be fixed, understood but not yet released as of May 2015)

Batch limits

- Number of nodes: 64 at the moment (please play nicely)
- Walltime: 24 hours

- No further limits/values needed atm, number of CPU per node is defaulted to 4 (nodes are dedicated, not shared)
- No need to give CPU time limits

ScaTeFS

- ScateFS has IO servers and IO targets (iot), data and metadata is spread and stripped over the resources
- HLRS installation is 2 IO servers and 18 targets
- Default is unstriped files

```
$ scatefs_getfinfo testfile
format     : non stripe format
iot count  :                18
stripesize :         268435456
chunksize  :         268435456
filesize   :       10737418240
```


ScaTeFS

- One can specify striping for all files in directory or for single files
- 4MB striping for all files in directory:

```
$ scatefs_setdirattr -s 4m testdir
$ touch testdir/testfile
$ scatefs_getfinfo testdir/testfile
format     :     stripe format
iot count  :                18
stripesize :           4194304
chunksize  :        1073741824
filesize   :                 0
```


ScaTeFS

- 1M striping for a single file

```
$ scatefs_premap -s 1M 0 testdir/testfile1m
$ scatefs_getfinfo testdir/testfile1m
format     :     stripe format
iot count  :                18
stripesize :           1048576
chunksize  :        1073741824
filesize   :                 0
```

- The 0 in front of filename denotes the file size, if the filesize to be written is known, some information can be preallocated to speedup later processing, if not known, give 0.

ScaTeFS

- On SX-ACE, a stripesize of 4M with an IO size of 16MB gives something like 700MB/s write bandwidth into the filesystem. IO size should be multiple of stripsize, and 4M seems to be a good value for stripsize.

Empowered by Innovation

NEC