# Filesystems, File Import/Export, Parallel I/O on NEC SX-8 at HLRS

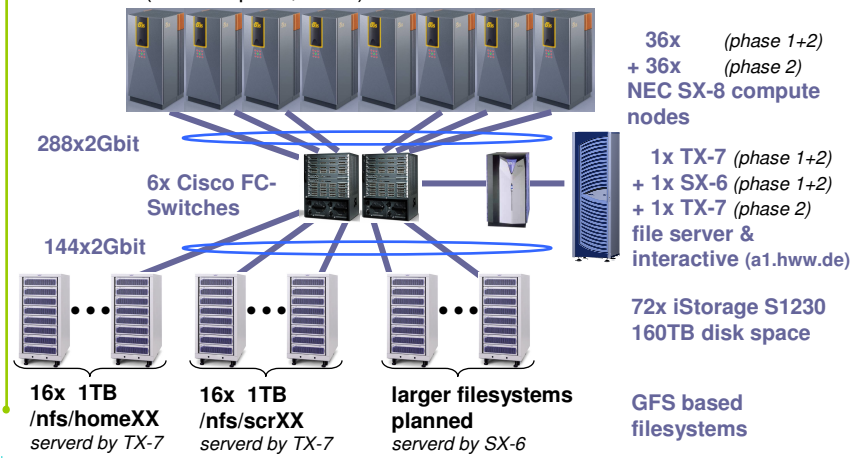
Holger Berger, Rolf Rabenseifner
hberger@hpce.nec.com, rabenseifner@hlrs.de

University of Stuttgart
High-Performance Computing-Center Stuttgart (HLRS)
www.hlrs.de



## Filesystems on NEC SX-8 at HLRS

Phase 1/2 (status April 1, 2005)




36x *(phase 1+2)*
+ 36x *(phase 2)*
NEC SX-8 compute nodes

1x TX-7 *(phase 1+2)*
+ 1x SX-6 *(phase 1+2)*
+ 1x TX-7 *(phase 2)*
file server & interactive (a1.hww.de)

72x iStorage S1230
160TB disk space

GFS based filesystems

## Workspace directories

- Idea: Users shall have large temporary diskspace which can be used for job-chains *without* copying data to home and back
  - SCRDIR=`ws_allocate *<name>* *<duration/days>*`
    - ***<name>*** **choose a name of your workspace**
    - ***<duration/days>*** **how long do you need this directory**
  - Monitor with ws_list
  - Reuse with ws_find <name> or ws_allocate
- Workspace can be accessed from frontend a1.hww.de during and after jobs, until it is wiped when reservation time is over
- All commands (ws_allocate, ws_list, ws_find, ws_release) can be used on a1.hww.de and inside of SX-8 batch jobs
- The allocation goes *round robin* through all scratch filesystems



## Workspace directories (continued)

- Example (user is hpcxxxx):
  - e.g., interactively:
    **SCRDIR=`ws_allocate TEST 7`**
    **echo $SCRDIR**
    `/nfs/scr7/ws/hpcxxxx-TEST-0`
  - **ws_list**

```
Workspace ID     Workspace location           Creation date   Remaining time
--------------------------------------------------------------------------------
rusrabe_A                                     Apr 4 14:05    7 days 23 hours
                 /nfs/scr7/ws/hpcxxxx-TEST-0
```

  - e.g., in your batchjob
    **SCRDIR=`ws_find TEST`**
    **echo $SCRDIR**
    `/nfs/scr7/ws/hpcxxxx-TEST-0`
  - e.g., if you have finished your project earlier
    **ws_release TEST**

## Home directories

- Based on the uid,
  the user will find his/her home-directory on one of the 16 filesystems
  /nfs/home1/ … /nfs/home16/



## Limits for home and scratch

- They are served by the TX-7.
- Each filesystem has only 1 TB.
- Maximal bandwidth from SX-8 is 600 MB/s.
- This bandwidth may be achievable with one CPU.
- 4 fold disk-striping into 4 RAID systems on each filesystem.
- Higher bandwidth can be achieved by using several filesystems in parallel, or the SX-6-based filesystems
- Optimization of Fortran or C I/O → next talk by Uwe Küster

## SX-6-based large filesystems (planned)

- Large filesystems for files >= 1TB
  must be currently served by the SX-6 file-server node.
- High accumulated bandwidth by
  - writing individual files by each MPI process
    with normal binary Fortran or C I/O, or
  - using MPI-I/O to write one file collectively → talk by Rolf Rabenseifner
- Not yet tested.



## Data migration

- Any data migration only on the TX-7 (a1.hww.de),
  i.e., **never** inside of SX-8 batch jobs.
- rcp – between a1.hww.de (SX-8 front-end)
  and asama.hww.de (SX-6 front-end)
  or other hww platforms
  (faster than scp)
- scp – between a1.hww.de (SX-8 front-end)
  and your institute

## Phase 3 (in 2006)

- TX-7 and SX-6 file serving is substituted by special GFS servers.
- 1 TB limit (currently on TX-7) is then obsolete.
- Number of FS will reduce, and size of single FS will increase.
- Aggregated bandwidth of single filesystem will increase.



## Further reading

| | |
|---|---|
| • HLRS | www.hlrs.de |
| • SX-8 | www.hlrs.de/hw-access/platforms/sx8/ |
| • Online courses | www.hlrs.de/organization/par/par_prog_ws/ |
| • Teraflop Workbench | www.teraflop-workbench.de/ |
| • On a1.hww.de | Login-message (more /etc/motd) |