NEC SX-8 at HLRS

Holger Berger, NEC HPCE,
Service&Delivery
hberger@hpce.nec.com

Company Overview

• Newly (in 2003) created NEC subsidiary
• Dedicated to HPC business
• Headquarters in Düsseldorf, Germany
• Serving the European Market
• Branch offices in France, Italy, The Netherlands, Switzerland and United Kingdom
• Application tuning & support centre in Stuttgart
• About 95 employees

Largest dedicated HPC operation in Europe!
NEC’s SX-Series is a consistent innovation driver and today’s leading high performance platform

NEC uses latest manufacturing technology to build and develop new generations of supercomputers, while the customer investment in software is preserved.

SX-8 specifications

• 16 GF / CPU (vector)
• 64GB/s memory bandwidth per CPU
• 8 CPUs / node
• 512 GB/s memory bandwidth per node
• Maximum 512 nodes
• Maximum 4096 CPUs, max 65 TFLOPS
• Internode crossbar Switch
• 16 GB/s (bi-directional) interconnect bandwidth per node
• Maximum size SX-8 is among the most powerful computers in the world
SX-8 CPU Block Diagram

SX-8 System Architecture
SX-8 Technology

- Hardware dedicated to scientific and engineering applications.
- CPU: 2 GHz frequency, 90nm-CU technology
- 8000 I/O per CPU chip
- Hardware vector square root
- Serial signalling technology to memory, about 2000 transmitters work in parallel
- 64 GB/s memory bandwidth per CPU
- Multilayer, low-loss PCB board, replaces 20000 cables
- Optical cabling used for internode connections
- Very compact packaging.

Innovation & Customer Benefit

<table>
<thead>
<tr>
<th>Technology</th>
<th>Leads to</th>
<th>Customer Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Vector Architecture</td>
<td>Exploitation of fine grain parallelism</td>
<td>Efficient Programs, Easy Programming</td>
</tr>
<tr>
<td>Advanced LSI, 90nm CU, 8000 pins</td>
<td>High density packaging</td>
<td>Low operational and investment cost, about half the power consumption of SX-6</td>
</tr>
<tr>
<td>Optical Interconnect Cabling</td>
<td>Easy Installation and Maintenance</td>
<td>Low operational and investment cost, six times less parts than SX-6</td>
</tr>
<tr>
<td>Low Loss PCB technology, serial signalling to memory</td>
<td>High packing density, easy manufacturing</td>
<td>Low operational and investment cost, 4 times less space than SX-6</td>
</tr>
</tbody>
</table>
SX-8 Node Cabinet Configuration

1 Node (128GF)  2 Nodes (256GF)

Node cabinet 300mm I/O cabinet

I/O cabinet shared by 2 nodes

One side maintenance

Node
DGP
CLK
PCIF
HDD
FAN
FAN

300mm 370mm

POWER

(1800mm)

800mm 525mm

Node cabinet

370mm 300mm

POWER

1996 2000 2004 2005

HLRS & NEC

576 cpus @16 GF
9 TB memory

48 cpus @ 9 GF
384 GB memory

32 cpus @ 4GF
80 GB memory

32 cpus @ 2GF
8 GB memory

© NEC HPCE, 2003, Finishdesign
HLRS Installation Roadmap

- 72 nodes
- 72 nodes
- 36 nodes
- ~40 TB disk-space
- old asama as 2. fileserver
- new fileserver with 160TB disks


sx-8 configuration at hlrs

- 144xGbit
- 3xFoundry Etherswitch with 4 x 10Gbit Uplinks
- SX-8 compute nodes
- TX-7 interactive server/file server (a1.hww.de)
- HWW-Network
SX-8 configuration at HLRS storage

- 288x2Gbit
- 6x Cisco FC-Switches
- 144x2Gbit
- 72x iStorage S1230
- 160TB disk space
- TX-7 interactive server/file server (a1.hww.de)

SX-8 vector compute nodes v00-v71

- 72x8 CPU nodes, 16/22 Gflops performance, 128 GB shared Memory
- 64 GB/s memory bandwidth
- connected by 16 GB/s IXS crossbar
- 6 2-Gbit FC-HBAs, 4 for global filesystems
- Software:
  - NQSI:
    - Fortran/SX Compiler as so far
    - C++/SX Compiler as so far
    - HPF/SX Compiler as so far
    - MPI/SX as so far
TX7 interactive server a1.hww.de

- 16(32)x Intel Itanium2 1.5Ghz/6M
- 256(512) GB Memory
- Partitioned, so not all CPUs will be visible to users
- 14 2-gbit FC-HBAs/12 Gbit ethernet interfaces
- Running NEC Linux based on RH Advanced Server
- SX cross compilers are installed
- Filesystem are shared with SX
- Integrated in NQSII
- Usage: Pre- and Postprocessing, compiling for SX

Changes for SX-6 Users

- New mandatory NQSII limit elapstim_req for wallclock time
- New versions of compilers supporting HW SQRT
- New mechanism to allocate disk space: Workspaces
- Idea: Users shall have large temporary disk space which can be used for job-chains without copying data to home and back
  - Call SCRDIR=`ws_allocate <name> <duration/days>`
  - 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
NQSII Configuration

- mainly as on SX-6
- 4 nodes for small jobs in shared mode, serial and < 8 CPUs
- 64 nodes for large jobs, 8 CPUs and more
- new elapse time limit of 24 hours, to increase throughput
- less restart overhead, as data can be kept in workspace
- no SCRDIR anymore, use workspace mechanism to get one
- in regular production, no checkpointing or suspending
- scheduling with fair share, so users priority is lowered with increasing usage of the resources
- it is possible to login with rsh to nodes where a owned dedicated job is running

Multi-node job example on dedicated nodes

```
#!/usr/bin/ksh
#PBS -q multi # routing queue for multi-node jobs
#PBS -l cpunum_prc=8 # cpus per Node (don't modify!!!)
#PBS -b 4 # number of nodes, e.g., 4 nodes x 8 CPUs
#PBS -l elapstim_req=24:00:00 # max wall-clock time
#PBS -l cputim_job=192:00:00 # max accumulated cputime per node
#PBS -l memsz_job=120gb # memory per node
#PBS -A yyynnnnn # Your Account code, see login message
#PBS -j o # join stdout/stderr
#PBS -T mpisx # Job type: mpisx for multi-node MPI
#PBS -N your_job_name # job name
#PBS -M your_mail@address # you should always specify your email

# with 8 MPI-processes per node:
mpirun -nn $_MPINODES -nnp 8 your_executable arg1 ...

# with 1 multi-threaded MPI-process per node:
export OMP_NUM_THREADS=8
echo MPIEXPORT=OMP_NUM_THREADS
mpirun -nn $_MPINODES -nnp 1 your_executable arg1 ....
```
Single-node job example
in time-sharing

#!/usr/bin/ksh
#PBS -q dq # routing queue for single-node jobs
#PBS -l cpunum_prc=5 # cpus per Node, e.g., 5 CPUs
#PBS -b 1 # number of nodes (don't modify!!!)
#PBS -l elapsed_req=100:00:00 # max wall-clock time
#PBS -l cputim_job=100:00:00 # max accumulated cputime
#PBS -l memsz_job=120gb # memory per node
#PBS -A yyynnnnn # Your Account code, see login message
#PBS -j o # join stdout/stderr
#PBS -N your_job_name # job name
#PBS -M your_mail@address # you should always specify your email

# with, e.g., 5 MPI-processes:
export MpSUSPEND=ON # to inhibit wasting of CPU time if gang-scheduling is switched off
mpirun -nn $_MPINNODES -np 5 your_mpi_executable arg1 ...

# with, e.g., 5 OpenMP threads
export OMP_NUM_THREADS=5
your_openmp_executable arg1 ...

Current Queue Limits (April 1, 2005)

• multi
  - #PBS -b 32 (68) # number of nodes, on 36 (72) node hardware
  - #PBS -l elapsed_req=24:00:00 # max wall-clock time
  - #PBS -l cputim_job=192:00:00 # max accumulated cputime per node
  - #PBS -l memsz_job=120gb # memory per node

• dq (single-node)
  - para: <= 8 CPUs, elapsed <= 100h, cpu <= 100h, mem <= 124 GB
  - test: <= 4 CPUs, elapsed <= 600sec, cpu <= 600sec, mem <= 10 GB
  - seriell: = 1 CPU, elapsed <= 100h, cpu <= 12h, mem <= 64 GB
Filesystem Configuration (phase 1)

- users homes are distributed over 16 filesystems /nfs/homeXX
- there are 16 scratch filesystems /nfs/scrXX
- each FS is 1 TB in size
- speed of a filesystem is around 600 MB/s from one node (4 fold striping)
- with phase3 in 2006, number of FS will reduce, and size of single FS will increase
- aggregated speed of single filesystem will increase in phase3

---

Thank you. Questions?