Empowered by Innovation NEC

NEC SX-8 at HLRS

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

Empowered by Innovation NEC

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

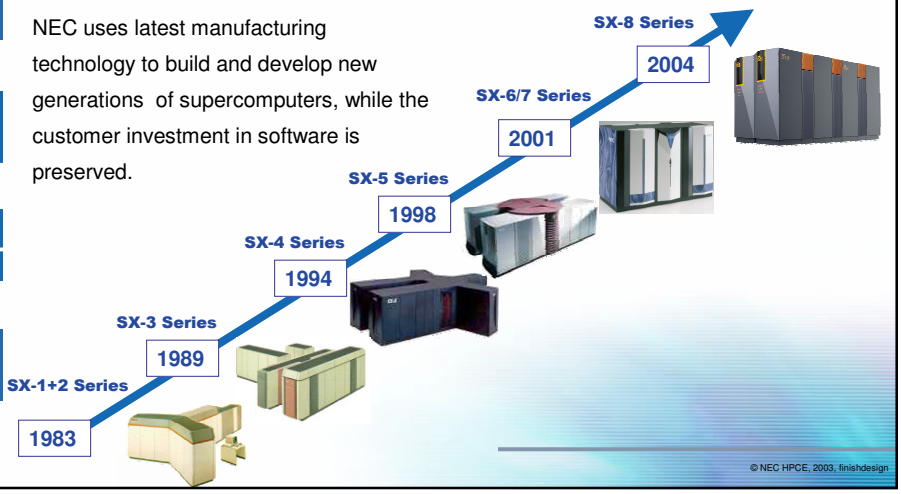
Largest dedicated HPC operation in Europe!

© NEC HPCE, 2003, finishdesign

Empowered by Innovation NEC

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.



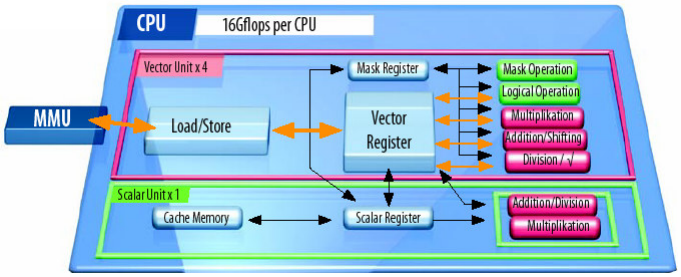
Empowered by Innovation NEC

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

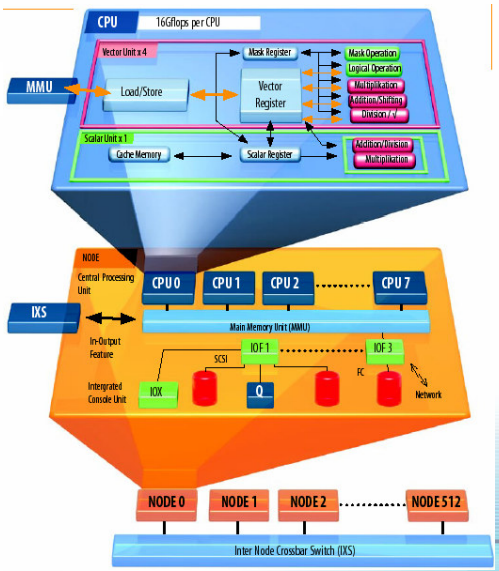
Empowered by Innovation NEC

SX-8 CPU Block Diagram



Empowered by Innovation NEC

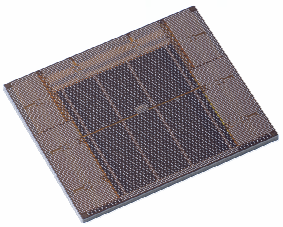
SX-8 System Architecture

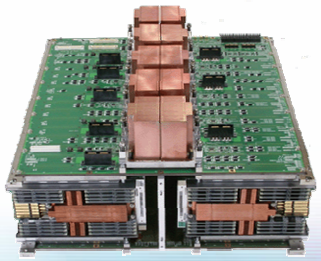


Empowered by Innovation NEC

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.

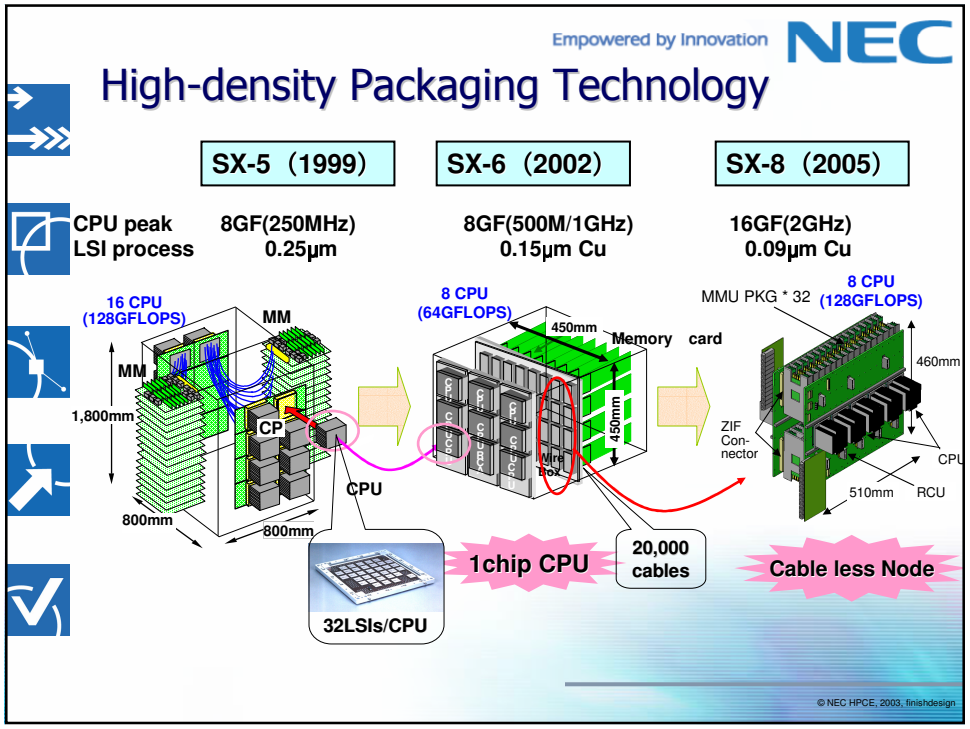


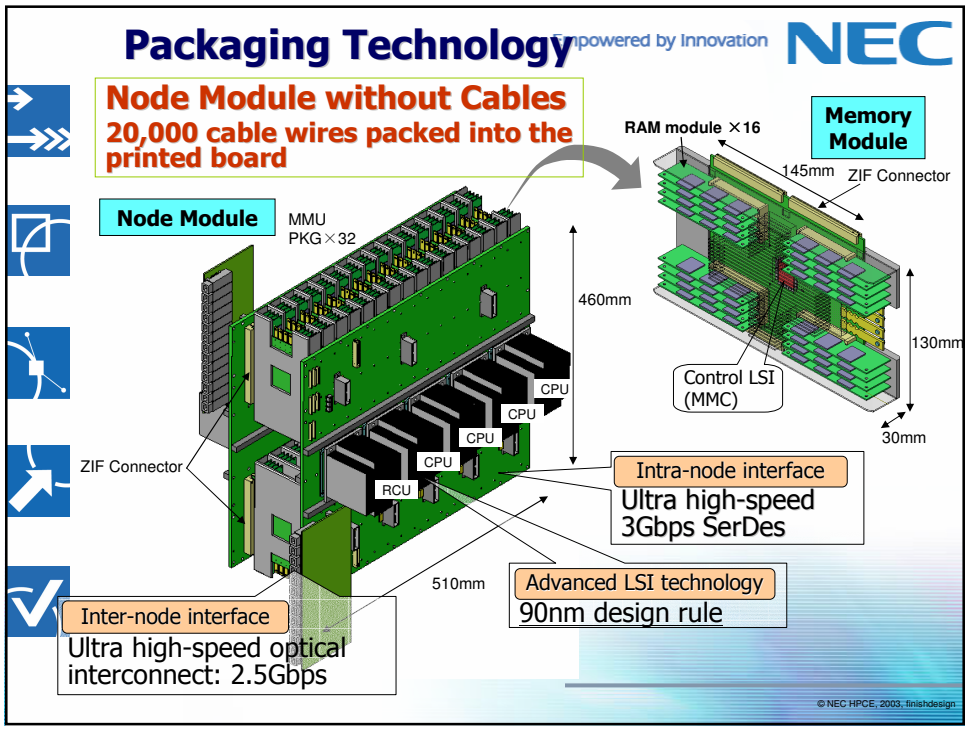


Empowered by Innovation NEC

Innovation & Customer Benefit

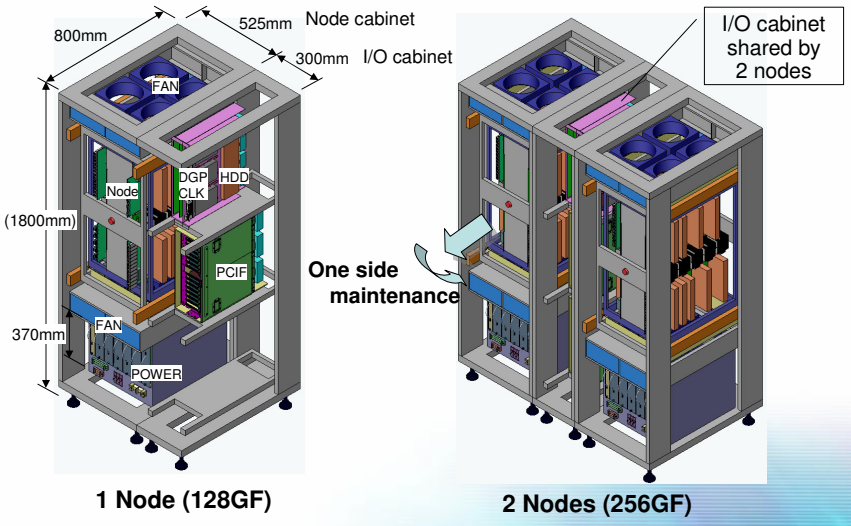
Technology	Leads to	Customer Benefit
Advanced Vector Architecture	Exploitation of fine grain parallelism	Efficient Programs, Easy Programming
Advanced LSI, 90nm CU, 8000 pins	High density packaging	Low operational and investment cost, about half the power consumption of SX-6
Optical Interconnect Cabling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


Empowered by Innovation
NEC
High-density Packaging Technology
SX-5 (1999)
SX-6 (2002)
SX-8 (2005)
CPU peak
LSI process
8GF(250MHz)
0.25μm
8GF(500M/1GHz)
0.15μm Cu
16GF(2GHz)
0.09μm Cu
16 CPU
(128GFLOPS)
MM
1,800mm
CP
800mm
800mm
CPU
8 CPU
(64GFLOPS)
450mm
Memory card
450mm
MMU PKG * 32
8 CPU
(128GFLOPS)
460mm
ZIF
Con-
nector
CPU
510mm
RCU
32LSIs/CPU
1chip CPU
20,000
cables
Cable less Node
© NEC HPCE, 2003, finishdesign


Packaging Technology
Empowered by Innovation
NEC
Node Module without Cables
20,000 cable wires packed into the printed board
Node Module
MMU
PKG×32
460mm
CPU
CPU
CPU
CPU
RCU
ZIF Connector
510mm
RAM module ×16
Memory Module
145mm
ZIF Connector
130mm
30mm
Control LSI
(MMC)
Intra-node interface
Ultra high-speed
3Gbps SerDes
Advanced LSI technology
90nm design rule
Inter-node interface
Ultra high-speed optical
interconnect: 2.5Gbps
© NEC HPCE, 2003, finishdesign

Empowered by Innovation NEC

SX-8 Node Cabinet Configuration



1 Node (128GF)

2 Nodes (256GF)

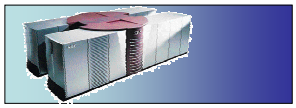
Empowered by Innovation NEC

HLRS & NEC

576 cpus @16 GF
9 TB memory

48 cpus @ 9 GF
384 GB memory

32 cpus @ 4GF
80 GB memory

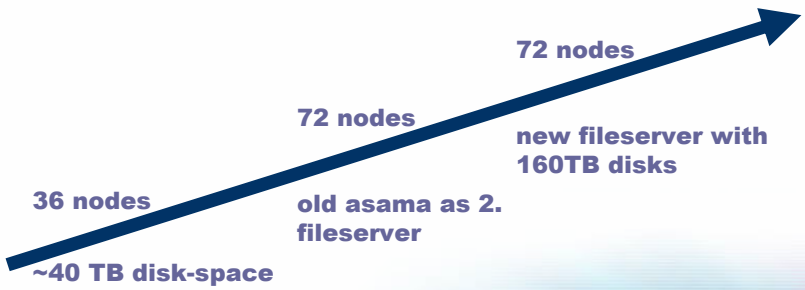


32 cpus @ 2GF
8 GB memory

|1996 |2000 |2004 | 2005

Empowered by Innovation NEC

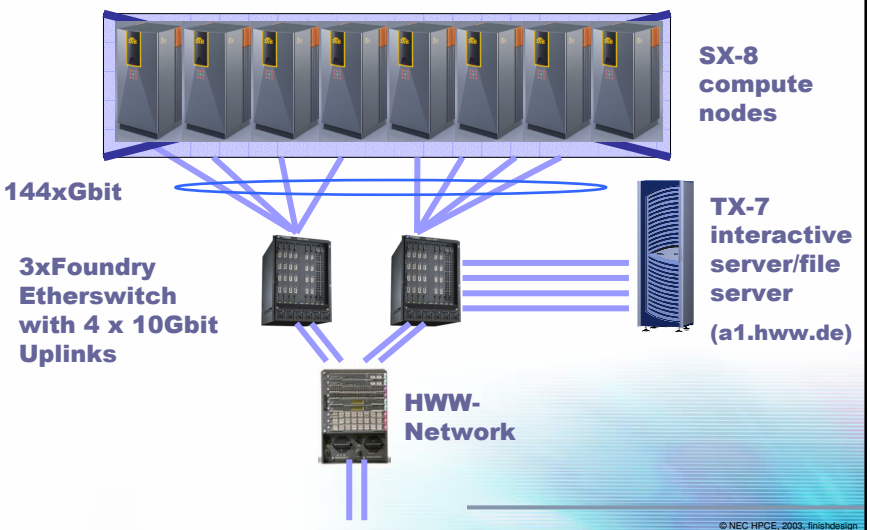
HLRS Installation Roadmap



|4/2005 |6-7/2005 |4/2006

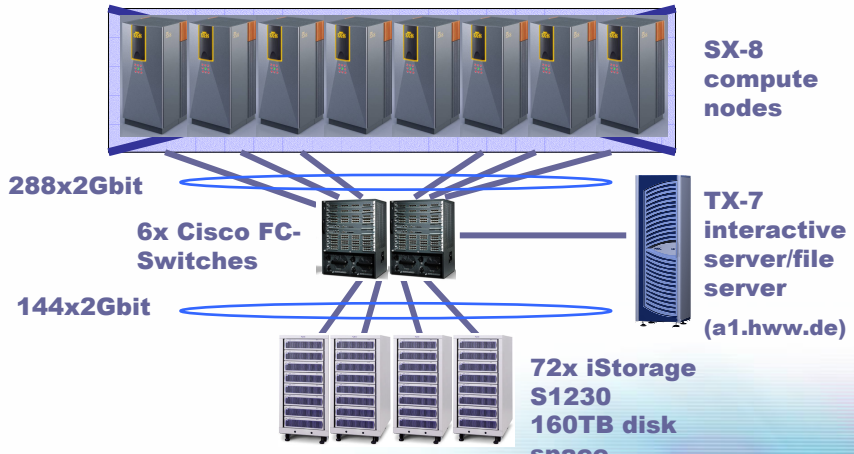
Empowered by Innovation NEC

SX-8 configuration at HLRS network



Empowered by Innovation NEC

SX-8 configuration at HLRS storage



Empowered by Innovation NEC

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:
  - -NQSII:
  - -Fortran/SX Compiler as so far
  - -C++/SX Compiler as so far
  - -HPF/SX Compiler as so far
  - -MPI/SX as so far

Empowered by Innovation NEC

TX7 interactive server a1.hww.de

- 16(32)xIntel 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

Empowered by Innovation NEC

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 tempory diskspace 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

Empowered by Innovation NEC

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

Empowered by Innovation NEC

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 $_MPINNODES -nnp 8 your_executable arg1 ...
# with 1 multi-threaded MPI-process per node:
export  OMP_NUM_THREADS=8
export  MPIEXPORT=OMP_NUM_THREADS
mpirun -nn $_MPINNODES -nnp 1 your_executable arg1 ....
```

Empowered by Innovation NEC

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 elapstim_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 MPISUSPEND=ON            # to inhbit wasting of CPU time if gang-scheduling is switched off
                                # by the operating
mpirun -nn $_MPINNODES -nnp 5 your_mpi_executable arg1 ...

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


Empowered by Innovation NEC

Current Queue Limits (April 1, 2005)

- **multi**
  - #PBS -b 32 (68) # number of nodes, on 36 (72) node hardware
  - **#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

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

Empowered by Innovation NEC

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 filesytem will increase in phase3

Empowered by Innovation NEC

Thank you.
Questions?