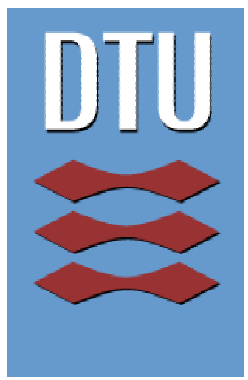
Challenges in Building LARGE Linux Clusters

Dr. Ole Holm Nielsen

Manager of computer services
Center for Atomic-scale Materials Physics (CAMP)
Department of Physics
Technical University of Denmark (DTU)
Lyngby, Denmark

E-mail: Ole.H.Nielsen@fysik.dtu.dk



Danish Center for Scientific Computing

http://www.dcsc.dk

- **DCSC** was founded by the Ministry of Research in 2001. Replaces previous supercomputing organization.
- A new, **user-driven** way to organize supercomputing.
- **Board** of 4 university and 5 external representatives.
  1 part-time **Director** with secretarial assistance.
- Scientists apply with **research projects**, asking for resources at DCSC's computers.
  Peer review of applications.
  Groups of scientists at 4 universities operate computers on behalf of DCSC.
- **Current computer installations:**
  2 Linux clusters, 1 Sun SMP, 1 IBM SMP, 2 SGI SMPs.

DCSC's Linux clusters

- www.fysik.dtu.dk/CAMP/Niflheim and www.dcsc.sdu.dk.
- Two research groups at DTU and SDU made a joint project of building two Beowulf-style Linux PC clusters.
- We asked vendors to offer maximum CPU performance within a given budget. *Compaq* (now *HP*) was selected, offering a **grand total of 1,000** EVO D510 desktop Pentium-4 PCs: 480 PCs @2.26 GHz for DTU; 520 PCs @2.0 GHz for SDU. No Microsoft Windows licenses ☺
- Peak-performance levels: **2.17 TeraFLOPS** (DTU) and **2.08 TeraFLOPS** (SDU).

Choices of technologies

- **Our primary guiding principle:**
  Maximum delivered application performance within a given budget.

Choices of technologies (2)

- **CPUs:**

  SPEC CFP-2000 floating-point benchmark (www.spec.org) is a useful performance indicator.
  Intel Pentium-4 is faster than most RISC processors.

  Pentium-4 outperforms other x86 CPUs due to high clock frequency, SSE2 64-bit vector instructions, and DDR-RAM (new Intel i845G chipset).

  Pentium-4 **price/performance** is unbeatable !

  Single-CPU systems are preferred over multi-CPU systems.

Choices of technologies (3)

- **Physical form factor:**

  Small desktop boxes are almost as compact as 1U rack nodes, but at less than half the price.

  We believe that racks may possibly be “nice-to-have” but they aren’t required, even for large clusters.

Choices of technologies (4)

- **Network:**
  100 Mbit/s Ethernet performs very well, is reliable and inexpensive.
  Gigabit Ethernet causes too many CPU-interrupts,
  and Gigabit switches are still expensive.

- **Applications:**
  Our parallel applications work well with Ethernet.
  Users have to adapt their codes to Ethernet latency/bandwidth,
  or apply to use other computer systems.

Challenges in planning a large cluster

- Sturdy storage shelves can be bought easily and at a low cost.
- Note: PCs cabinets should be mountable in a vertical position without special stands (a non-trivial point – check it out before buying).

Cooling system

- Even though a Pentium-4 PC consumes a mere 90 W, 500 PCs will dissipate a serious 45 kW.
- An adequate cooling system is a standard product, but it may take some time to do the physical installation.
  Our University Services handle all cooling systems.
- A new cooling system cost us <10% of computer budget.

Physical installation

A large truck arrives with 24 pallets of PCs.

A team of 30 students and professors are ready to do some work **in parallel** !

A detailed work plan has been written down (see our Web-page).

Unpacking 480 boxes

Registering and labeling PCs

PCs must be powered on for:

1. **BIOS setup:** Changing the factory defaults for server operation.
   Fortunately, the HP EVOs can **load BIOS parameters from diskette!**
2. **Identification:** Registering Ethernet MAC-address and attaching labels:
   Use **DHCP-booting** to get the MAC addresses onto the server.
   Use a **DYMO label printer** with Excel to generate adhesive labels.

Cabling

Constructed in a few days by electricians using standard components.
There are two 24-port switches on each top-shelf.

Network

We chose a 2-level star Ethernet topology:

Cabling is simple and can be replicated across the shelves.

HP Gigabit switch
30 ports (max. 48)

24 ports @100 Mbit/s

Alternative:

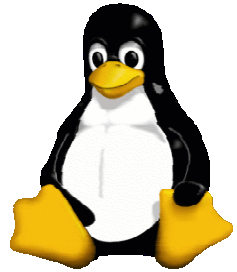
Single Ethernet switch with as many ports as you have PCs.

- However, the largest (affordable) switch has max. 240 ports. Interconnect by Gigabit would be a serious bottleneck.
- Single switch has lower hardware latency.
- Cabling many nodes to a single switch is more complex.

The cluster is ready !

We installed 480 PCs in 3.0 hours – a World Record ??

Installing 480 copies of Linux



There exists a number of toolkits for Linux clusters (OSCAR, ROCKS, …).

We selected **SystemImager** (www.systemimager.org):

- Install and customize a **Golden Client** Linux PC as usual (we use RedHat 7.3). Extract a disk-image onto the installation server.
- Client PCs do **PXE network boot** at power-on and execute a **pxelinux** boot-image that will download the Golden Client image. This process is **100% automatic** (just press power-on) !
- When the client PC reboots, the PXE-server instructs the client to boot from disk (or to do a renewed Linux installation, if desired).
- We have installed groups of 18 clients simultaneously from a HP Proliant server (2-CPU Xeon @2.4 GHz). Installation completes in 6 minutes.

Network performance looks fine

NetPIPE benchmark, www.scl.ameslab.gov/netpipe/

Network hardware	*Protocol*	*Latency (microseconds)*	*Bandwidth (Mbit/s)*	*Bandwidth (Mbytes/s)*
Same switch	TCP	43	89.7	11.2
Remote switch	TCP	55	89.7	11.2
Same switch	LAM-MPI	50	89.7	11.2
Remote switch	LAM-MPI	61	89.7	11.2
Same switch	MPICH	62	87.1	10.9
Remote switch	MPICH	73	87.1	10.9

File service

A global file space is available from a Network File System (NFS) server.

- **Server choice:**
  HP/Compaq Tru64 UNIX server with dual Alpha EV68 CPUs @833 MHz.
  Mirrored system disks. Dual power supplies. UPS.
- **Why not a Linux server ?**
  I don't trust the robustness of the Linux NFS-server implementation
  (we've been bitten a few times).
  The NFS-server must serve reliably ~500 hungry network clients.
- **Disk space:**
  A third-party RAID-5 disk with 12 times 160 GB of IDE disk
  (Ultra3 SCSI host interface) gives 1.6 TB of usable space at <10k€.
  The downside is limited monitoring capabilities.

Batch service

- De-facto standard is Portable Batch System (PBS), either Open Source from www.openpbs.org or commercial from www.pbspro.com ($1k charge/yr for educational customers).
- A contender may be the Sun Grid Engine (SGE).
- MAUI scheduler from www.supercluster.org is a fantastic policy management tool.
  – a must on any multi-user or large cluster.
  – handles priorities, fair-share and lots more.

Additional software

- Commercial compilers:
  Portland Group Fortran-90 and C++.
  Intel Fortran-90 and C++.
- ATLAS BLAS matrix-library (+ Intel MKL).
- FFTW Fourier transform library.
- MPI message-passing: LAM-MPI and MPICH.
- Python tools.

Operational experiences

- An experimental **long-distance Gigabit interconnect** (250 km) of our 2 clusters at DTU+SDU (a total of 1,000 nodes !) showed that the Linux kernel must have all MAC-addresses in a static ARP-cache (above ~500 nodes).
- **Reliability:**
  All 480 Compaq/HP EVO D510 PCs worked correctly at delivery.
  Repairs in first weeks: 2 motherboards, 2 RAM modules, 1 PSU and 1 disk.
  Ethernet network has been 100% reliable.
  Cluster availability so far: 100%.
- **Users:**
  We have about 25 users from 3 research groups.
  The batch queue was full within 24 hours of general service.
  The number of nodes requested by users is usually around 1200-1500.

Conclusions

- In Denmark we have built two Linux clusters,
  both exceeding **2 TeraFLOPS** of peak-performance.
  Minimal manpower was required, except for the physical installation.
  Funding by the new **DCSC** center.
  Detailed information available at www.fysik.dtu.dk/CAMP/Niflheim
- Desktop PCs on cheap shelves and Ethernet networking is by far the most **cost-effective** supercomputing solution available today.
  Pentium-4 CPUs have outstanding performance (cf. *SPEC CFP2000*).
- Building a 500-node cluster is straightforward, provided you plan the **installation process** carefully: There is no real need for *value-added services* from vendors. Good knowledge of Linux is a prerequisite.
- Almost all software is available as **Open Source**.
  However, we use commercial compilers and PBSPro batch system.