

CLUSTER-COMPUTING DEVELOPMENTS IN THE UK

Computational Science and Engineering Department,

Daresbury Laboratory, Warrington, WA4 4AD

Introduction

A “cluster” is a collection of complete computers (nodes) that are physically interconnected by a high-performance “local area network” (LAN). Typically each node is a workstation or personal computer (PC). Clusters permit running parallel jobs using PVM or MPI implemented over the network as well as permitting independent use of the nodes for task farming.

The advantages of cluster computing derive from the fact that off-the-shelf commodity components are used. This offers a cost-effective solution to medium-scale computing requirements.

For a good introduction to the concepts of cluster computing and architectures available see “Scalable Parallel Computing” by K. Hwang and Z. Xu (WCB/McGraw-Hill, 1998, ISBN 0-07-031798-4).

Cluster computing solutions in the USA are becoming mainstream and may become the dominant system of the future for computational science. A 64x2-way Alpha-based system built by Alta Technology and installed at the University of New Mexico has been accepted into the USA Alliance computational meta-computing grid. A similar system, **the CPlant** <http://www.cs.sandia.gov/cplant>, is supported by Compaq under a 4-year agreement with the US DoE and is now being used at Sandia National Laboratory with a Myrinet high-performance switch for enhanced communications. This system forms part of the Accelerated Strategic Computing Initiative (ASCI) Path Forward programme.

One of the first projects, **the Beowulf**

<http://www.beowulf.org>, was started at NASA in 1994. This Web page also contains links to many related sites worldwide. Commodity cluster systems are now often known as Beowulf-class computers.

It is important for UK scientists to be able to evaluate this kind of equipment for parallel computing, as has been noted by EPSRC in recent surveys. Daresbury Laboratory therefore, as part of the Distributed Computing (DisCo) Programme, has built a 32-processor Beowulf cluster using 450 MHz Pentium III processors. Currently the processors, which are in the form of off-the-shelf PCs, each with memory and disk but no keyboard or monitor, are connected by dual fast Ethernet switches - 2x Extreme Summit48, one network for IP traffic (e.g. nfs) and the other for MPI message passing. Additional 8-port KVM switches are used to attach a keyboard and monitor to any one of the nodes for administrative purposes. The whole cluster has a single master node (with a backup spare) for compilation and resource management. All nodes are currently running RedHat Linux v6.0.

Applications, such as GAMESS-UK, DL_POLY, ANGUS, CRYSTAL, POL-ERSEM, REALC and CASTEP are being ported to the system for evaluation. Results showing their performance will be

posted on the **DisCo Web site**

<http://www.cse.clrc.ac.uk/Activity/DisCo> as they become available.

Over the coming months we also plan to evaluate a variety of networking and software options for the system. Some of the options are summarised below. Prices vary, as does performance and robustness, and it is not yet clear what will be the preferred solution for building a large-scale compute server.

Network Options

network	latency (μ s)	bandwidth (MB/s)
fast Ethernet	50	12.5
Gigabit Ethernet †	9.6	93
Myrinet	20	62
QSW QsNet ‡	5	210

† Gamma project with Packet Engines NIC.

‡ MPI short message protocol.

Figures in the table are subject to confirmation and depend on what driver hardware and software is used.

Message-passing Options

Message passing options include implementations of MPI and PVM, but there are others too.

MPICH

<http://www-unix.mcs.anl.gov/mpi/mpich/index.html> - Argonne National Laboratory's implementation of MPI

LAM/MPI

<http://www.mpi.nd.edu/lam> - Local Area Multicomputer MPI, developed at the Ohio Supercomputer Center and Univ. of Notre Dame

Globus

<http://www.globus.org/> - Metacomputing Environment

Compilers

Widely-used compilers include the Gnu family, Portland Group, KAI, Fujitsu, Absoft, NAG etc.. Compaq is about to beta test AlphaLinux compilers which are reputedly excellent. Some people already compile their applications under Digital Unix and run them on Alpha Linux, although this is not permitted under the license conditions.

Absoft Corp.

<http://www.absoft.com> - FORTRAN77 (f77) and Fortran 90 (f90)

The Portland Group

<http://www.pgroup.com> (PGI) - High Performance Fortran (pghpf), FORTRAN77 (pgf77), C and C++ (pgcc)

Numerical Algorithms Group

<http://www.nag.com> (NAG) - FORTRAN 90 (f90), Fortran 95 (f95)

Gnu CC/egcs

<http://egcs.cygnum.com> - free FORTRAN77, C, Pascal, and C++ compilers

Pentium gcc

<http://goof.com/pcg> , aka PGCC - from the Pentium Compiler Group uses Pentium-specific optimisations to produce 5%-30% speedups from regular gcc

BERT 77

<http://www.plogic.com/bert.html> - described as "an automatic and efficient Fortran paralleliser"

Lahey/Fujitsu

<http://www.lahey.com> - LF95 Linux Express fully optimising Fortran 95 compiler

Numerical Libraries

ASCI Option Red software

<http://www.cs.utk.edu/~ghenry/distrib/archive.htm> - BLAS, fast-Fourier transform, hardware performance-monitoring utilities, extended-precision and maths primitives are all available free under restricted licenses

Fast Maths library

http://www.lsc-group.phys.uwm.edu/~www/docs/beowulf/os_updates/fastMath.html and Free Fast Maths library - makes standard mathematical functions much faster

NAG

<http://www.nag.co.uk> Parallel Library - a version tuned for Beowulf systems is available commercially

Resource Management and Job Scheduling Options

LSF

<http://www.platform.com> - Load Sharing Facility from Platform Computing

LobosQ

<http://www.lobos.nih.gov> - queuing system from NIH, Bethesda USA

PBS

<http://pbs.mrj.com> - Portable Batch System developed at NASA Ames Research Center now commercially available from MRJ Inc.

Virtual Private Server

<http://www.sychron.com> - new software infrastructure for scalable internet services and enterprise services from Sychron Ltd. Oxford.

DQS

<http://www.scri.fsu.edu/~pasko/dqs.html> - Distributed Queueing System. A free batch queueing system

BVIEW

<http://w272.gsfc.nasa.gov/~udaya/Public/software/bview/bview.html> - monitoring software

bWatch

<http://www.sci.usq.edu.au/staff/jacek/bWatch> - monitoring software

BPROC

<http://www.beowulf.org/software/bproc.html> - making processes visible across nodes, allowing fork()s to happen across nodes, allowing process migration, allowing kill()s to work across nodes, currently pre-alpha release

Cluster patches for procps

<http://www.sc.cs.tu-bs.de/pare/results/procps.html> - lets you compile /proc-based programs like ps so they report on all processes on the cluster, not just the ones on the machine you're logged into

SMILE

<http://smile.cpe.ku.ac.th/software/scms/index.html> Cluster Management System - Run commands on all nodes, shut down individual nodes and sets of nodes, monitor health of nodes. Makes clusters easier to administer.

Parallel Virtual Filesystem

<http://ece.clemson.edu/parl/pvfs> - LD_PRELOAD-based filesystem modification to let you transparently stripe big files across many disks. Allows high-performance access to big datasets.

Scripts for configuring 'clone' worker nodes

<ftp://ftp.sci.usq.edu.au/pub/jacek/beowulf-utils/disk-less> - makes adding nodes to a Beowulf painless

Scripts

ftp://ftp.sci.usq.edu.au/pub/jacek/beowulf-utils/misc_scripts for doing various things on a cluster, backups, shutdowns, reboots, running a command on every node

Software

http://www.beowulf-underground.org is being added to the public domain on a daily basis, see for instance "The Beowulf Underground" URL

Other Beowulf Clusters in the UK

There are a number of other Beowulf systems being built in the UK. We list just a few of them on the DisCo Web page at **URL**

http://www.cse.clrc.ac.uk/Activity/DisCo. These include: Enterprise and Voyager (Cambridge), The Borg (Cranfield), University of Glasgow, Stac Follaidh (Lancaster), HPCI (Southampton), MadDog (UMIST).

Some Commercially available Systems

WorkstationsUK

http://www.workstationsuk.demon.co.uk

InSiliCo

http://www.insilico.co.uk

SALT

http://www.suse.de - Commercial Hardware Vendor in Germany

Sybrandt Open Systems

http://www.sybrandt.com provides a commercial Beowulf solution

Paralline

http://www.paralline.com distributes Linux clusters, high-speed networks and services

ParTec

http://www.par-tec.com supports the ParaStation project and sells clusters and services

See also the NASA Web site mentioned above.

An Example

As an example of our initial experiences using the Daresbury Beowulf we show the performance obtained from DL_POLY. The test cases were NaCl MTS Ewald 27000 ions, NaK disilicate glass 8640 ions and Gramicidin in water, SHAKE, 13390 atoms. Results from the 450 MHz Pentium III system are compared with an earlier 260 MHz Pentium II system and Cray T3Es. Clearly the single-node performance of the Pentium III is good compared to the Cray T3E-1200E, but the latter offers superior scalability for parallel programs needing a large number of processors.

Figure: DL_POLY benchmarks

