CLRC

**Technical Report**
RAL-TR-2001-025

# Wulfgar - the R&D Beowulf Cluster Project Report

**P M Oliver**



22nd June 2001

COUNCIL FOR THE CENTRAL LABORATORY OF THE RESEARCH COUNCILS

Enquiries about copyright, reproduction and requests for additional copies of this report should be addressed to:

The Central Laboratory of the Research Councils
Library and Information Services
Rutherford Appleton Laboratory
Chilton
Didcot
Oxfordshire
OX11 0QX
Tel: 01235 445384 Fax: 01235 446403
E-mail library@rl.ac.uk

ISSN 1358-6254


# CENTRAL LABORATORY OF THE RESEARCH COUNCILS

**Rutherford Appleton Laboratory Chilton, Didcot, OX11 0QX, UK**

**Information Technology Department**

**Scientific Computing Services Group**

## Wulfgar – the R&D Beowulf Cluster Project Report

Dr. Peter Oliver Email: p.m.oliver@rl.ac.uk

April 26 2001

Abstract

The report provides a detailed account of the installation, benchmarking and usage of the R&D Beowulf Cluster known as Wulfgar. The Cluster uses AMD Athlon processors connected via Myrinet, a low latency high bandwidth interconnect. The performance of the Myrinet interconnect is compared to standard 100Mbit Ethernet on a variety of applications from Computational Chemistry to Weather Modelling.

## Contents



## Tables and Figures

# 1. BEOWULF RESEARCH AND DEVELOPMENT PROJECT

This project was funded by the internal CLRC Research and Development fund. The aim was to investigate and make available to CLRC departments and other external users a test Beowulf Platform. This would then act as a technology demonstrator enabling CLRC departments to pump prime their own projects. Thus making sure that Beowulf was a good solution for their problems before making a potentially expensive mistake.

A Beowulf Cluster is a cluster of workstations connected with a fast dedicated network enabling parallel jobs to be run. There are two major considerations for a Beowulf system, processors and network. We investigated Intel and AMD processors and compared 100Mbit and Myrinet networking.

A variety of application areas were investigated including, computational chemistry, weather modelling, Genomics and CFD. Enabling access to the GRID using Globus is also considered.

# 2. PROCESSOR AND NETWORK EVALUATION

## 2.1 Introduction

The aim of the project was to investigate high performance computing using commodity components and hence maintain a good price performance ratio. Thus Intel CPU and Gigabit Ethernet were considered. However with the introduction of the new AMD Athlon K7 processor in June 1999 it was decided to obtain an Athlon system for evaluation.

## 2.2 Athlon Processor Evaluation

Two benchmarks, `FLOPS` and `STREAM`, and two computational chemistry codes `METADISE` and `STORM` were chosen for the evaluation.

### *2.2.1 Benchmarks `FLOPS` and `STREAM`*

The `FLOPS`[1] program measures the sustained MFLOPS achieved using a mixture of FADD, FSUB, FMUL, and FDIV operations based on specific 'instruction mixes'. The test MFLOPS(3) represents a good mix giving 3.4% FDIV.

The `STREAM`[2] program measures the memory bandwidth in Mbytes/s.

| Test | K7 600MHz | PII 400MHz | PIII600 est | Ratio |
|---|---|---|---|---|
| **MFLOPS(3)** | 251 | 94 | 116 | 2.12 |
| **Stream** | Copy: 469<br>Scale:440<br>Add:459<br>Triad:406 | Copy: 293<br>Scale:293<br>Add:234<br>Triad:234 | Copy: 363<br>Scale:363<br>Add:290<br>Triad:290 | Copy: 1.29<br>Scale:1.21<br>Add:1.58<br>Triad:1.4 |
| **Metadise (chem)** | 158 | 287 | 231 | 1.46 |
| **Storm (chem)** | 357 | 548 | 441 | 1.23 |

**Est is the ratio of Specfp numbers 12.8 for PII400 and 15.9 PIII600**

**Table 1 Comparison of the performance of Intel and Athlon Processors.**

From Table 1 it is clear that the Athlon processor has a clear performance advantage over the Intel processor with 2.12 times the MFLOPS and 1.4 times the memory bandwidth. For real applications the results are also impressive with METADISE and Storm being 1.46 and 1.23 times faster respectively.

The next component to be evaluated was the network.

## 2.3 Choice of Network

The network connecting the machines together is also important with the two important parameters being latency and bandwidth. Gigabit Ethernet was still too expensive to purchase components for evaluation so 100Mbit was benchmarked.

### *2.3.1 100Mbit Ethernet*

For these tests Intel 10/100 Ethernet cards were used in conjunction with a 3COM 3300XM SuperStack II switch.

A simple Ping-Pong latency and bandwidth program was used using LAM 6.3.2[3] as the MPI layer and EGCS 2.91.66 as the compiler

| TEST | RESULT |
|---|---|
| Latency (μ seconds) | 77 |
| Bandwidth (Mbytes/s) | 12 |

**Table 2 Latency and Bandwidth for 100Mbit Ethernet**

From table 2 it can be seen that good results are being obtained within the limits of the technology. For example the maximum theoretical bandwidth is 12.5Mbytes/s and 12Mbytes/s is obtained. This can be compared to the Cray T3E quoted results of 180Mbytes/s and 12μs latency. Thus the 100Mbit network is insufficient for supercomputing style applications.

After discussions with experts in the field at SuperComputing 1999 the opinion was that Gigabit Ethernet would show a small increase in bandwidth but have the same latency as 100Mbit Ethernet therefore we decided to move to Myrinet.

### *2.3.2 Myrinet*

The Myrinet card purchased were PCI64A which have the capacity to run at 66MHz and 64bits. However, they are compatible with 33MHz and 32bit PCI found in the Athlon PC chosen.

The same Ping-Pong program as for the 100Mbit test was used using MPICH v1.2.3 and GM 1.2.3.

| TEST | RESULT |
|---|---|
| Latency (μ seconds) | 15 |
| Bandwidth (Mbytes/s) | 97 |

Table 3 Latency and Bandwidth for Myrinet

From table 3 it is clear that Myrinet has a significant advantage over 100Mbit Ethernet. The latency is nearly 6 times better and the bandwidth is about 8 times better.

## 3. APPLICATIONS

In the next section applications from computational chemistry, weather modelling, CFD, Globus and genomics are presented

### 3.1 Computational Chemistry - Case Study (1) DL_POLY

For the computation chemistry package DL_POLY timings were performed on bulk ZrO2 using the Ewald sum and a 10 Angstrom cut off. Two sizes of problem were considered 6144 (medium) and 12000 (large) ions. The results for both Myrinet and 100Mbit are using 1, 2, 4, 8 and 16 processors are detailed in the graphs below.

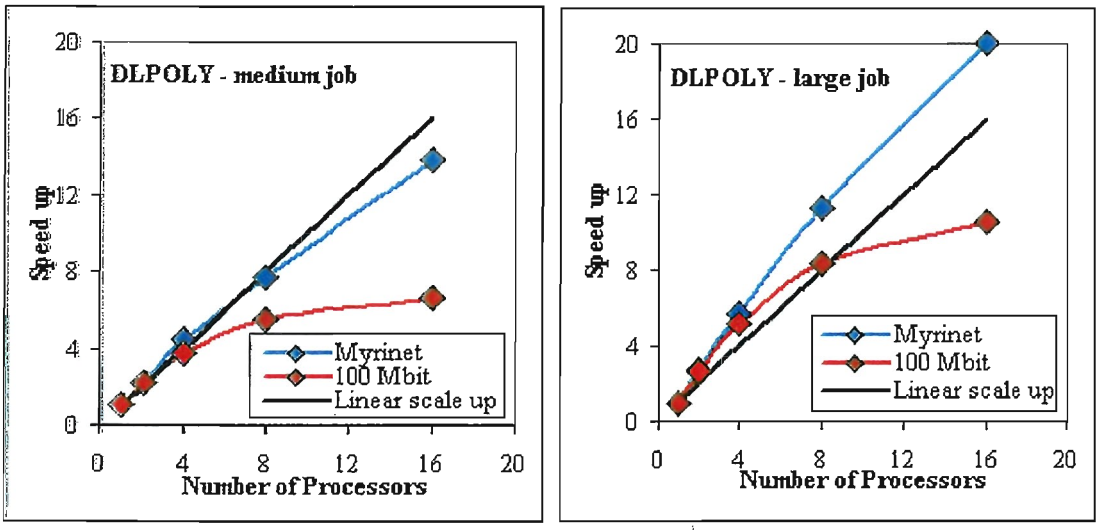


Figure 1 100Mbit vs Myrinet for DL_POLY

It can be seen from figure 1 that Myrinet is enabling more of the CPUs to be used as the speed up is close to linear. The large job however shows some interesting behaviour. The job has a super linear speedup which at first glance is difficult to understand. However, several factors could be responsible. The job could be memory bandwidth bound thus on a single CPU we only have around 500MB/s but on 16 the aggregate bandwidth is 8GB/s. The cache could also be playing a role. Each machine has 512k cache but when using 16 this aggregated to 8MB. Thus a combination of these factors gives rise to a super linear speedup.

### *3.1.1 Comparison with Cray T3E*

For the large job, 12000 ions, the timings and scalability were compared with a Cray T3E 1200E using up to 64 processors. The Cray T3E is a MPP style SuperComputer with very fast communications and processors (600MHz Alpha EV5 based).

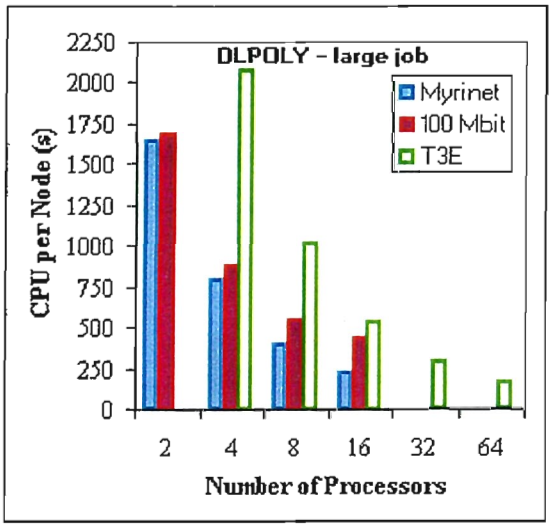


**Figure 2 100Mbit, Myrinet and Cray T3E data for the large DL_POLY job**

The Beowulf Cluster using 16 CPUs is faster than the T3E supercomputer using 32 CPUs for DL_POLY and 16 CPUs for VASP. Thus for smaller jobs the Beowulf Cluster is ideal leaving the T3E machine free for Grand Challenge Computations.

## 3.2 Computational Chemistry - Case Study (2) VASP

For the computation chemistry package VASP timings were performed on a 12 atom Pt (111) surface (medium) and a 24 atom Pt bulk cell (large) The results for both Myrinet and 100Mbit are detailed figure 3.

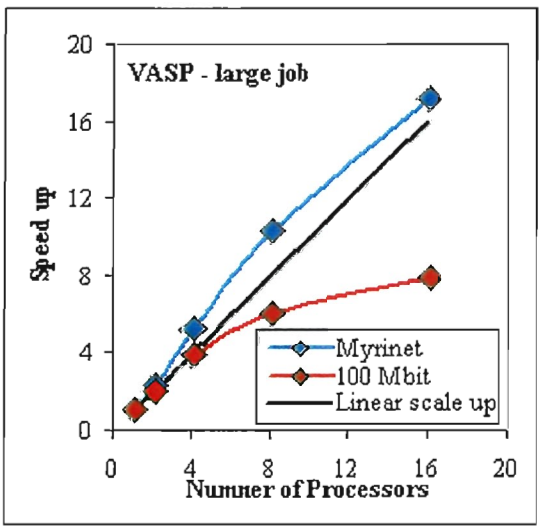


**Figure 3100Mbit vs Myrinet for VASP.**

A similar profile to DL_POLY is obtained for VASP indicating the benefit of Myrinet over 100Mbit Ethernet.

### *3.2.1 Comparison with Cray T3E*

For the large job, 24 atom Pt cell, the timings and scalability were compared with a Cray T3E using up to 16 processors.

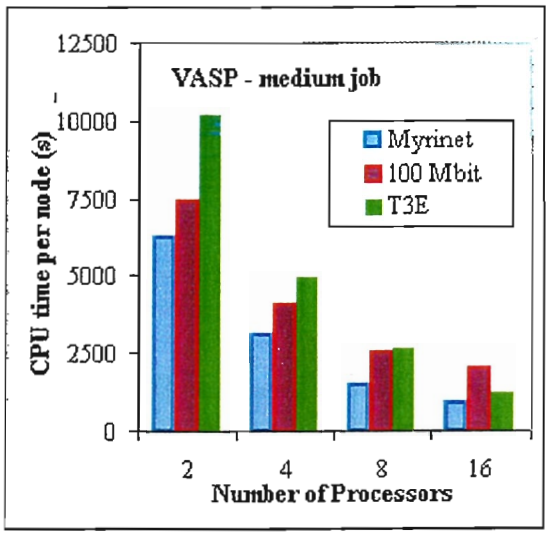


**Figure 4 100Mbit, Myrinet and Cray T3E data for the large VASP job**

VASP is a lot more demanding on the network than DL_POLY but the 16 node cluster is just faster than the same job run on 16 CPUs of a T3E.

## 3.3 Weather Modelling - Case Study (3) Unified Weather Model

The Unified weather model[45] (atmosphere only) was run on 1, 2, 9 and 16 processors using Myrinet and compared to the channel bonded 100Mbit (2x100Mbit for increased bandwidth) results obtained on a 450MHz PIII cluster. The results are show in figure 5.

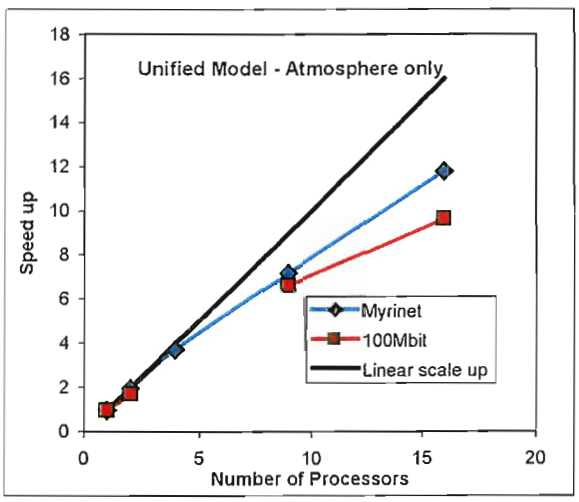


**Figure 5 Myrinet vs Channel bonded 100Mbit**

From figure 5 it is clear that the Myrinet is having a positive effect on the speed up obtained. However, the speed up is not as good as it could be with only 12 times speed up on 16 processors. Thus more work is needed on the parallelisation of the code to achieve a greater performance.

### 3.4 Globus[6] and CFD – Ron Fowler

Wulfgar was used as a Globus host in tests of the 1.1.1 and 1.1.3 releases of Globus. Various trials were made, at a fairly simple level, on the interoperability of this Linux based system with both Sun (Solaris) and IBM Power PC (AIX) systems. These showed that the hosts could work together successfully. The Globus aware versions of secure shell and ftp were also tested between these systems.

Wulfgar was also used to test a parallel computational fluid dynamics code running under Globus. The Globus aware version of MPICH (MPICH-G) was used for the communication between processes running both on Wulfgar and on a Sun system. The initial job was submitted on the Sun, using Globus to schedule the batch job on the PBS system on Wulfgar. This worked satisfactorily when the job queue was free, but otherwise has to wait until the PBS started the batch job.

No detailed measurements of the performance of the parallel CFD code were made but the code did run correctly and showed reasonable speed up for small numbers of processors. Using MPICH-G prevented use of the high speed Myrinet connection between processors, which limited the performance. The latest Globus version (1.1.4) has a new MPI implementation which should permit use of the high speed local network for nodes on Wulfgar, while using TCP connections to communicate with other systems.

The processing power of each node on Wulfgar is vastly superior to the old Sun SPARC10 system we also had access to. Hence it was only sensible to use the Sun as a master processor, sending work to Wulfgar and for visualisation of the results.

The software available on the system was very extensive, including Fortran 77 and Fortran 90 compilers, and all worked together very well.

With the commercial compilers and the up to date software tools from the RedHat distribution Wulfgar offers a very good software development environment.

### 3.5 Genomics – George Moraitakis from Birkbeck College

George Moraitakis carried out molecular dynamics simulations of proteins (lysozyme) using the GROMACS simulation package. His report follows.

All the usage so far is summarised in table 4.

a) The 1st column describes the simulations performed and its length in picoseconds (ps). The more ps performed the longer is the CPU usage.

b) The 2nd column shows the number of processors used and the third column is the compilation of GROMACS used:

default makefile options (A)
some optimisations on (B)
fortran inline loops + optimisations on (C)

c) The 4th column shows the CPU time taken, the 5th column shows how many ps of simulation are performed per CPU hour The 6th column shows how many CPU hours are required to perform 1000 ps.

d) The 7th column shows the MFlops.

e) From the results it can be seen that the three compilations do not differ much. The fortran inlined loops make the simulation slightly faster.

The same simulations were carried out on a SGI Origin 2000 using 4 processors. GROMACS was compiled with all optimisations on. Table 5 summarises the results.

From the 2 tables we see that the simulations on Origin 2000 are roughly twice faster than on Wulfgar. Scaling also seems to be better on the Origin 2000.

The results though may be biased to favour the SGI's because the creators of GROMACS have done more optimisations of the code for them and alphas (EV6) than for Linux PCs.

| SIMULATION | PROC | GMX | TIME | PS/HOUR | HOUR/NS | MFLOPS |
|---|---|---|---|---|---|---|
| warm300K 10ps | 4 | A | 0h:42:04 | 14,263 | 70.111 | 613.619 |
| 300K 1 100ps | 8 | A | 5h:10:20 | 19.334 | 51.722 | 920.751 |
| 300K 2 100ps | 4 | B | 7h:12:53 | 13.861 | 72.147 | 592.587 |
| 300K 3 200ps | 8 | B | 10h:26:36 | 19.151 | 52.217 | 911.835 |
| 300K 4 200ps | 8 | C | 10h:26:46 | 19.146 | 52.231 | 911.708 |
| 300K 5 200ps | 8 | A | 10h:27:54 | 19.111 | 52.325 | 910.009 |
| 300K 6 200ps | 2 | B | 24h:21:08 | 8.213 | 121.761 | 333.245 |
| warm500K 10ps | 2 | A | 1h:12:05 | 8.324 | 120.139 | 336.982 |
| 500K 1 200ps | 8 | C | 9h:09:53 | 21.823 | 45.824 | 879.204 |
| 500K 2 200ps | 8 | C | 11h:05:52 | 18.022 | 55.489 | 802.737 |
| 500K 3 200ps | 4 | C | 13h:43:10 | 14.578 | 68.597 | 562.213 |
| 500K 4 200ps | 8 | C | 10h:57:43 | 18.245 | 54.810 | 813.095 |
| 500K 5 200ps | 8 | B | 11h:04:44 | 18.052 | 55.394 | 808.795 |
| 500K 6 1000ps | 8 | C | 54h:04:35 | 18.492 | 54.076 | 823.189 |
| 500K 7 1000ps | 8 | A | 59h:02:49 | 16.936 | 59.047 | 812.329 |
| 500K 8 1000ps | 8 | B | 54h:19:18 | 18.409 | 54.322 | 819.349 |

**Table 4 GROMACS results for Wulfgar**

| SIMULATION | PROC | GMX | TIME | PS/HOUR | HOUR/NS | MFLOPS |
|---|---|---|---|---|---|---|
| warm300K 10ps | 4 | C | 0h:28:43 | 20.894 | 47.861 | 891.306 |
| 300K 1 1000ps | 4 | C | 47h:55:34 | 20.866 | 47.926 | 892.608 |
| warm500K 10ps | 2 | C | 0h:49:07 | 20.894 | 47.861 | 498.716 |
| 300K 2 1000ps | 2 | C | 83h:49:26 | 12.216 | 81.861 | 484.094 |
| 500K 1 1000ps | 4 | C | 44h:56:06 | 22.254 | 44.935 | 858.259 |
| 500K 2 1000ps | 4 | C | 42h:22:11 | 23.612 | 42.352 | 910.595 |

**Table 5 GROMACS result for SGI ORIGIN 2000**

## 4. USAGE

From Appendix B it can be seen that the Cluster has been very busy with usage reaching 93%. Over the latter part of 2000 and early 2001 this high usage has been sustained.

The Cluster was upgraded to 512MB/CPU in March 2001 and after the upgrade the cluster experienced a period of instability relating to memory errors. Therefore the Cluster was unavailable to users and hence the decrease in usage.

## 5. CONCLUSIONS

From all of the application areas investigated so far it is clear that Myrinet is having a large beneficial effect on the scalability of the codes. This, taken together with the superior performance of the Athlon processor over the Intel processor, makes a Myrinet cluster of Athlon processors a significant computational platform.

The GROMACS results suggest that a 4 processor Origin 2000 is as fast as 8 processors on Wulfgar for this type of work, initially making the Origin 2000 more attractive. However, when price/performance considerations are make the choice is clearly in favour of the Wulfgar Cluster. A 4 processor Origin 2000 typically costs around £50k. This is similar to the price of the whole Wulfgar Cluster of 16 processors giving the Beowulf Cluster a considerable advantage in price/performance terms.

## 6. THE FUTURE

We plan to upgrade the cluster to dual AMD Athlon CPUs in 2001 using the Wulfkit interconnect. The Wulfkit interconnect and SCALI software is more in tune with dual systems as their software is much better at shared memory MPI. The Wulfkit interconnect is also more scalable than Myrinet as it does not require switches and uses a 2D/3D torus for connectivity. The 2D torus allows up to 100 nodes to be connected with no performance degradation. For Myrinet you need larger switches which are very expensive.

## 7. PUBLICATIONS


1. R.J. Allan, S.J. Andrews, M.F. Guest, P.M. Oliver, D. Henty, L. Smith, S. Telford and S. Booth. "Design and Building ad Beowulf-Class Cluster Computers", UKHEC (2000)

2. G.W. Watson, "The Origin of the Electron distribution on SnO", Journal of Chemical Physics. **114**(2), 758 (2000). (Acknowledgement)

3. P.M. Oliver, NERC IT Manager Meeting, November 2000 http://wwwhpc.rl.ac.uk/talks/beowulf_NERC_11_00/index.shtml

4. P.M. Oliver, "Benchmarks of the new Columbus, a Compaq Alphaserver SC Cluster and the R&D Beowulf Cluster Wulfgar", Poster at HPC meeting QEII Conference Centre London, September 2000.

5. Demonstration at the E-Science Meeting at RAL, March 2001.

6. Web Page http://wwwhpc.rl.ac.uk/columbus/hw_beowulf.shtml

## 8. APPENDIX A CLUSTER CONFIGURATION

The configuration of the cluster is as follows

| |
|---|
| 16 AMD 850MHZ ATHLON PROCESSORS |
| 650MHz AMD Athlon font end with 36GB of Home filespace |
| 256MB of ECC memory per CPU (upgraded to 512MB in March 2001) |
| 10GB local /tmp space CPU |
| 100Mbit switched Ethernet |
| 16 port Myrinet switch with 16 PCI64A cards |
| Redhat 6.2 and kernel 2.2.16 |
| Portland Group Compilers v3.2 |
| OpenPBS Batch system v 2.3.11 |
| MPICH-GM v1.2.3 |
| Optimised BLAS libraries from Greg Henry and the ATLAS project |
| GLOBUS 1.1.3 Grid software |

**Table 6 Wulfgar Cluster configuration**

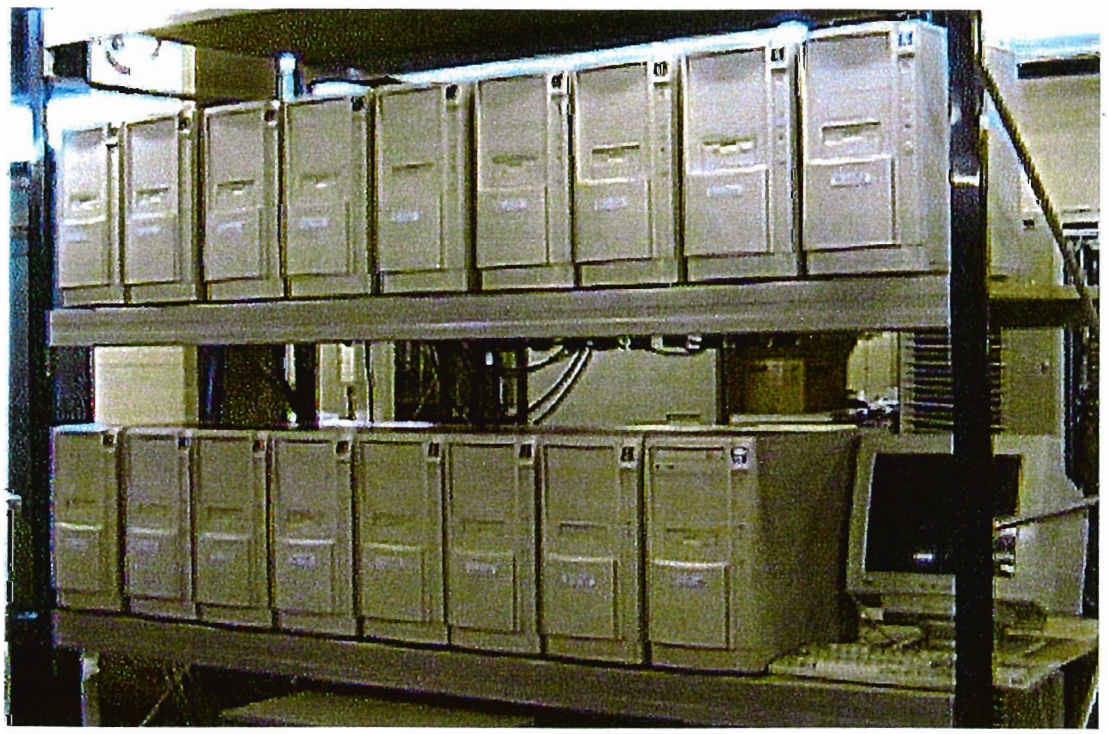

**Figure 6 Wulfgar System**

# 9. APPENDIX B USAGE

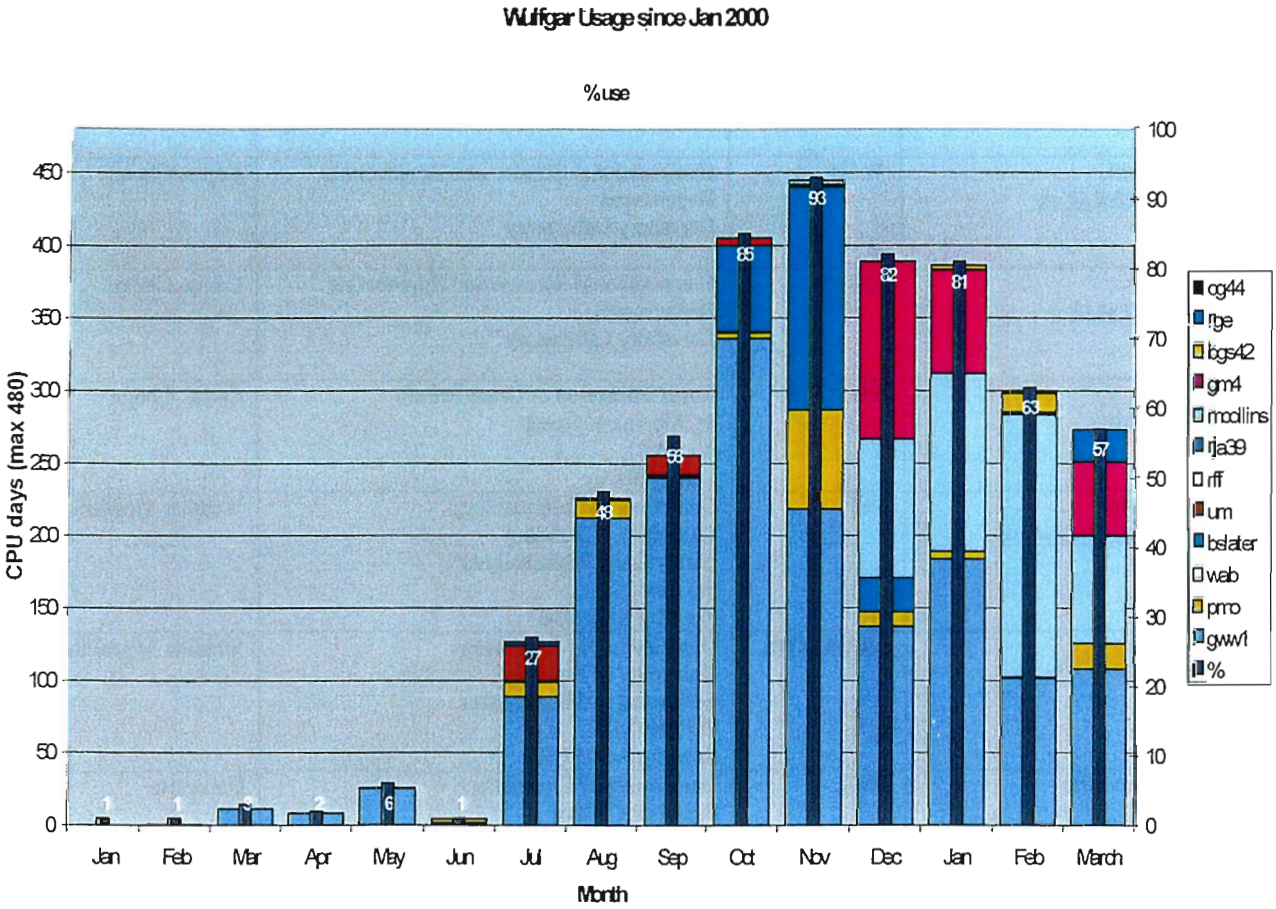


Figure 7 Wulfgar Usage since January 2000 divided by user

## 10. APPENDIX C USERS

| User | UID | Address | Field |
|---|---|---|---|
| Graeme Watson<br>watsong@tcd.ie | GWW1 | Department of Chemistry<br>Trinity College<br>Dublin 2<br>Ireland | Comp. Chem. |
| Martyn Guest<br>M.F.Guest@dl.ac.uk | WAB | Computational Science and Engineering Department<br>Daresbury Laboratory | Comp. Chem. |
| Rob Allan<br>r.j.allan@dl.ac.uk | RJA39 | Computational Science and Engineering Department<br>Daresbury Laboratory | Comp. Chem. |
| Ben Slater<br>ben@ri.ac.uk | BSLATER | Royal Institution of Great Britain<br>21 Albemarle Street<br>London<br>W1X 4BS | Comp. Chem. |
| Andrew Heaps<br>andy@met.reading.ac.uk | UM | Department of Meteorology<br>University of Reading<br>Earley Gate, Whiteknights<br>PO Box 243<br>Reading RG6 6BB | Weather Modelling |
| Mat Collins<br>mat@met.reading.ac.uk | MCOLLINS | Department of Meteorology<br>University of Reading<br>Earley Gate, Whiteknights<br>PO Box 243<br>Reading RG6 6BB | Weather Modelling |
| George Moraitakis<br>g.moraitakis@cryst.bbk.ac.uk | GM4 | Department of Chemistry<br>Birkbeck College<br>Gordon House<br>29 Gordon Square<br>London WC1H 0PP | Genomics |
| Roger Evans<br>r.g.evans@rl.ac.uk | RGE | RAL<br>Bldg R2 | |
| Chris Greenough<br>c.greenough@rl.ac.uk | CG44 | RAL<br>Bldg R27 | CFD and Globus |
| Ron Fowler<br>r.f.fowler@rl.ac.uk | RFF | RAL<br>Bldg R27 | CFD and Globus |
| Barry Searle<br>b.g.searle@dl.ac.uk | BGS42 | Computational Science and Engineering Department<br>Daresbury Laboratory | |

---

[1] Al Aburto, aburto@marlin.nosc.mil, 1992

[2] John D. McCalpin, Revision: 4.1, June 4, 1996

[3] http://www.mpi.nd.edu/lam/

[4] Unified Model from the Met Office, http://www.meto.govt.uk/

[5] Andrew Heaps, andy@met.reading.ac.uk

[6] http://www.globus.org