

# HPCx Quarterly Report

## Apr-Jun 2003

### 1 Introduction

This report covers the period from 1 April 2003 at 0800 to 1 July 2003 at 0800.

The next section summarises the main points of the service for this quarter. Section 3 gives details of the usage of the service, including failures, serviceability, CPU usage, helpdesk statistics and service quality tokens. Section 4 includes reports on progress against the Annual Plan from each of the functional teams. A summary table of the key performance metrics is in Section 5. The Appendices define the incident severity levels and list the current HPCx projects.

### 2 Executive Summary

- The numbers of failures and incidents are much lower than in the previous quarter and this is reflected in better MTBF and serviceability figures.
- A 2-frame software development facility was installed towards the end of this quarter. This should allow us to provide a flexible and responsive service while maintaining high MTBF and serviceability.
- During this quarter, utilisation of the service has averaged 75% on the capability region.
- However, looking ahead, it appears that demand is likely to outstrip supply throughout the next year.
- The modal job size remains 128 CPUs and more than a quarter of the utilisation was from jobs of at least 256 CPUs.
- Interactions with the users have been good:
  - the helpdesk is exceeding all its targets;
  - the first user group meeting was held and the first newsletter was distributed;
  - the HPCx annual seminar is being arranged at Daresbury in December;
  - there are now four technical reports available on the web site;

- the outreach to lifesciences has identified a number of projects due to start within the next quarter.
- We have made good progress in investigating the performance characteristics of the system:
  - in-depth studies of the HPCx IO system;
  - identifying that the LAPI performance problems appear to be associated with the LAPI\_Get function.
- The Terascaling team has begun work on a good range of applications codes and has had a number of successes:
  - significant speedups have been recorded with the most recent implementations of the NAMD and CASTEP codes, the latter through enhanced treatment of all-to-all collective operations;
  - an increasing range of tools have been made available to users of the service, including Vampir (512 CPUs), Total View and libhmd.

### 3 Usage Statistics

#### 3.1 Availability

##### 3.1.1 Failures

The monthly numbers of incidents and failures (SEV 1 incidents) are shown in the table below:

	April	May	June
Incidents	5	8	6
Failures	2	2	2

The number of incidents and failures has been significantly lower than in the previous quarter.

The following tables give more details on the attribution of the failures:

##### *April*

<i>Failure</i>	<i>Site</i>	<i>IBM</i>	<i>External</i>	<i>Reason</i>
03.083		100%		GPFS inaccessible from I1f01
03.086			100%	Loss of JANET external to the site

##### *May*

<i>Failure</i>	<i>Site</i>	<i>IBM</i>	<i>External</i>	<i>Reason</i>
03.088	100%			Uninterrupted access to the external network is site responsibility
03.092		100%		Ability to submit LL jobs is technology supplier responsibility

##### *June*

<i>Failure</i>	<i>Site</i>	<i>IBM</i>	<i>External</i>	<i>Reason</i>
03.099	100%			DL firewall failure
03.100	100%			Power incident during electrical upgrade work for development machine

#### 3.1.2 Performance Statistics

This section uses the definitions agreed in Schedule 7, ie,

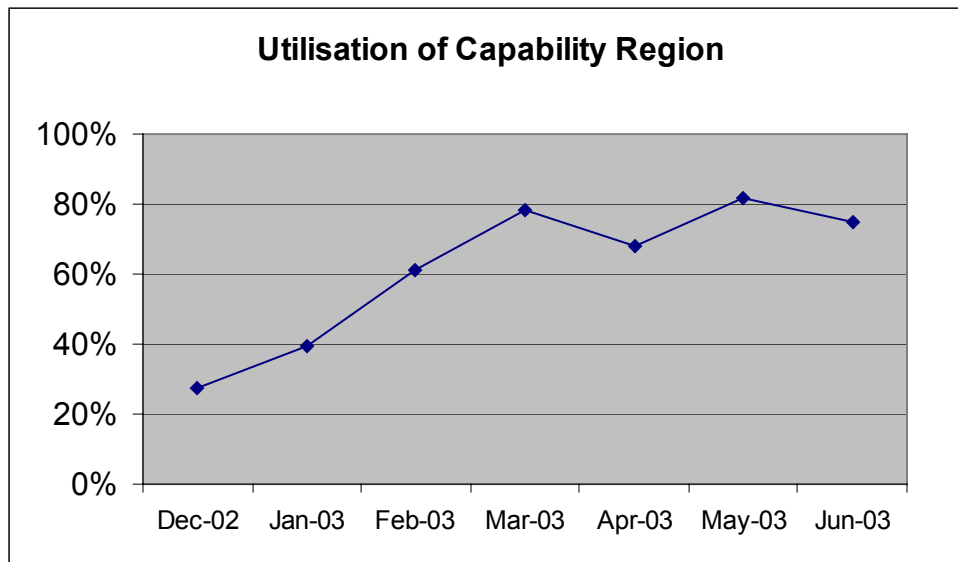
- $MTBF = (24 \times 30.5) / (\text{number of failures in month})$

- Serviceability (%) = 100 x (WCT – SDT – UDT) / (WCT– SDT)

<i>Attribution</i>	<i>Metric</i>	<i>April</i>	<i>May</i>	<i>June</i>	<i>Quarterly</i>
IBM	Failures	1	1	0	2
	MTBF	732	732	∞	1098
	Serviceability	99.9	99.8	100.0	99.9
Site	Failures	0	1	2	3
	MTBF	∞	732	366	732
	Serviceability	100.0	99.9	99.5	99.8
External	Failures	1	0	0	1
	MTBF	732	∞	∞	2196
	Serviceability	99.8	100.0	100.0	99.9
Total	Failures	2	2	2	6
	MTBF	366	366	366	366
	Serviceability	99.8	99.8	99.5	99.7

### 3.2 Capability Utilisation

The monthly utilisation for the 1024-processor capability region is shown in the graph below. This has averaged more than 75% for the last 4 months.

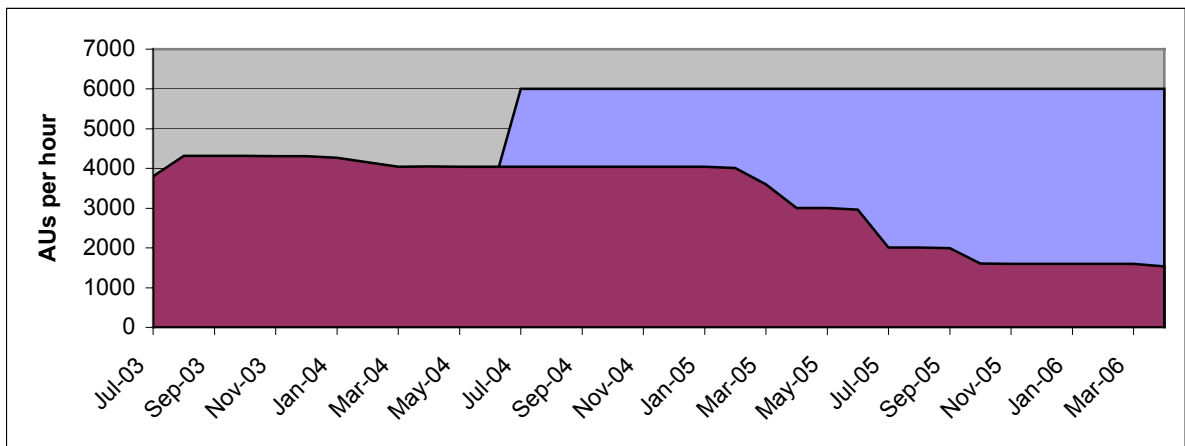


### 3.3 Capacity Planning

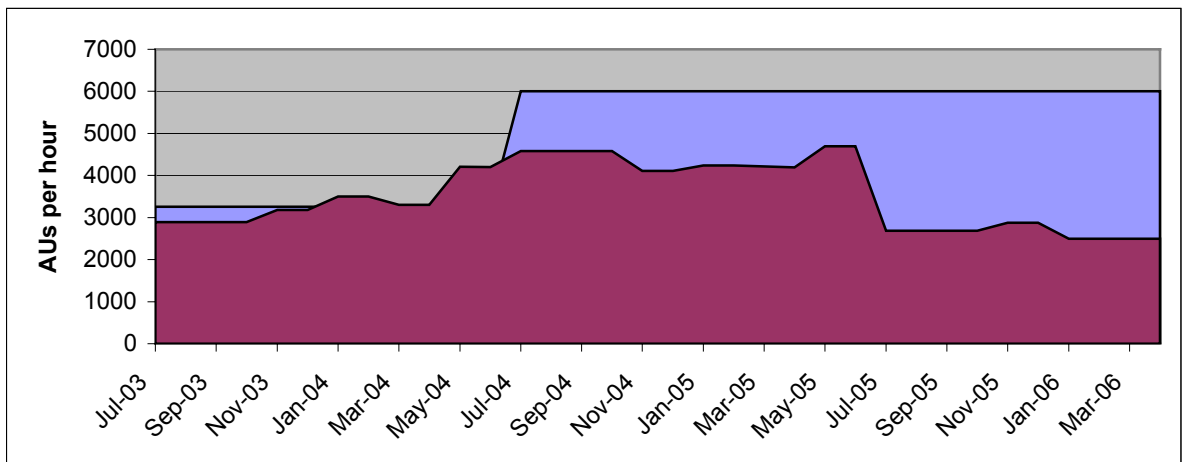
#### *Predicted Utilisation*

The following graphs show predicted utilisation against time until April 2006. The scale on the y-axis is AUs per hour, where the peak that HPCx Phase 1 could currently deliver is around 3250 AUs per hour (the light blue region in the graphs below). However, the practical maximum is probably around 80% of this, i.e., 2600 AUs per hour.

This first graph naively assumes that each project will use its remaining allocation evenly over the term of their grant. The steps in the graph show where grants come to an end but, in many cases, there are likely to be follow-on grants.



This naive analysis suggests that demand over the next year appears to exceed the practical maximum by 60%. We have repeated this analysis using the profile information in the original research grants.

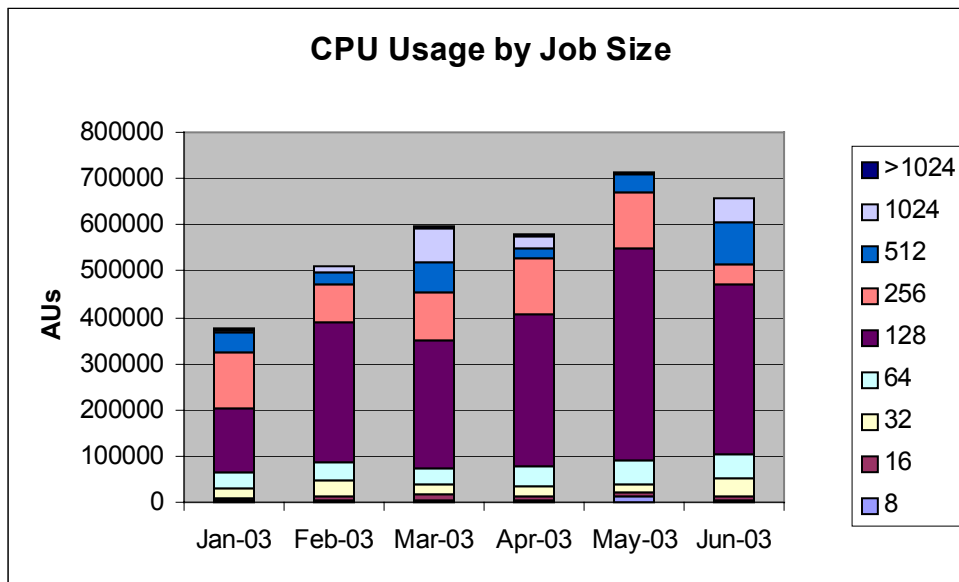


This graph suggests that, during the next year, the likely demand is 25% higher than the practical maximum. This portrays a more encouraging picture than the previous one but, nevertheless, it appears likely that, even without new grants coming on to the system, demand on the Phase 1 system will exceed supply.

### Numbers of Research Consortia

There are currently 17 research consortia using the HPCx system. The HPCx support activity is sized on a maximum of 25 concurrent research consortia (1.9.3 in Schedule 3). With the Life Sciences projects due to commence, we will soon be very close to this limit.

### 3.4 CPU Usage by Job Size



The above graph shows that the modal job size is still 128 CPUs. However, during this quarter, more than a quarter of the utilisation was from jobs of at least 256 CPUs.

### 3.5 CPU Usage by Consortium

The PIs and titles for the various consortia are listed in Appendix B.

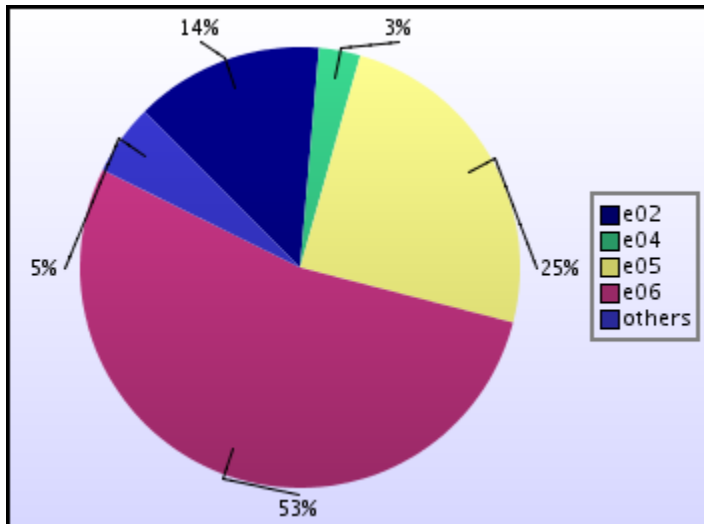
Consortium	April	May	June	Quarterly	%age
e01	1753	19313	9284	30350	0.61%
e02	177451	166316	334711	678478	13.68%
e03	5535	37189	57920	100644	2.03%

e04	18258	34830	100347	153435	3.09%
e05	637672	341919	235975	1215566	24.52%
e06	599309	1171842	879497	2650648	53.46%
e07		61	1	62	0.00%
e08		1	134	135	0.00%
e09			4957	4957	0.10%
<b>EPSRC Total</b>	<b>1439978</b>	<b>1771471</b>	<b>1622826</b>	<b>4834275</b>	<b>97.50%</b>

n01		12234		12234	0.25%
n02	290	298	32	620	0.01%
n03	9749	21785	16528	48062	0.97%
n04			4971	4971	0.10%
<b>NERC Total</b>	<b>10039</b>	<b>34317</b>	<b>21531</b>	<b>65887</b>	<b>1.33%</b>

p01			1302	1302	0.03%
<b>PPARC Total</b>			<b>1302</b>	<b>1302</b>	<b>0.03%</b>

z001	9787	15412	22500	47699	0.96%
z002	71	15	192	278	0.01%
z004	13	55	1014	1082	0.02%
z05	7005			7005	0.14%
z06	1492	220	374	2086	0.04%
<b>HPCx Total</b>	<b>18368</b>	<b>15702</b>	<b>24080</b>	<b>58150</b>	<b>1.17%</b>



## 3.6 Helpdesk

### 3.6.1 Classifications

<i>Category</i>	<i>Number</i>	<i>% of all</i>
Administrative	106	45.9
Technical	108	46.8
In-depth	16	6.9
PMR	1	0.4
TOTAL	231	100.0

<i>Service Area</i>	<i>Number</i>	<i>% of all</i>
Phase 1 platform	175	75.8
Website	24	10.4
Early User Service	1	0.4
Other/general	31	13.4
TOTAL	231	100.0

### 3.6.2 Performance

<i>All non-indepth queries</i>	<i>Number</i>	<i>%</i>	<i>Target</i>
Finished within 24 Hours	174	81.3	75%
Finished within 72 Hours	214	100.0	97%
Finished after 72 Hours	0	0	

<i>Administrative queries</i>	<i>Number</i>	<i>%</i>	<i>Target</i>
Finished within 48 Hours	104	98.1	97%
Finished after 48 Hours	2	1.9	

### 3.6.3 Experts Handling Queries

<i>Expert</i>	<i>Admin</i>	<i>Technical</i>	<i>In-Depth</i>	<i>PMR</i>
epcc.ed.ac.uk	59	47	6	0
dl.ac.uk	9	21	1	0
Sysadm	30	37	8	1
Other people	8	3	1	0

### 3.7 Service Quality Tokens

<i>Date</i>	<i>Person</i>	<i>Value</i>	<i>Comment</i>	<i>Status</i>
04-Jun-2003 09:16:33	<a href="#">Dr Dario Alfe</a>	-	the machine is now almost completely idle, waiting for space for one job on a 128_1 queue. something should be done about this, as it is not an efficient way of running the machine	Significant effort has been put into tuning the batch configuration as can be seen from the slowdown figures in the monthly reports.
04-Jun-2003 09:15:05	<a href="#">Dr Dario Alfe</a>	*	The policy of dividing the machines in pools of 8 processors each is very bad for codes which requires heavy communications, and it undermines the true capability of the machine, which may have nodes of 32 processors, and not 8, with fast communications	This restriction is currently essential to maximise the bandwidth of the Colony switch but will be removed at Phase 2.

## 4 Support

### 4.1 Applications Support (*Dr David Henty*)

The HPCx Applications Support effort is progressing well. The majority of everyday user issues are dealt with by the online documentation and user admin pages, which are continually updated. This quarter we completed four in-depth technical reports, and ran seven courses, a workshop and the first User Group meeting.

#### 4.1.1 Helpdesk

The helpdesk has continued to operate extremely smoothly, and the statistics above show that we are meeting our targets for answering user queries.

#### 4.1.2 Documentation

The HPCx User Guide is now at version 1.2. Major updates this quarter have included information on using the new LoadLeveler interactive queues, and the revised memory allocation policies for mixed MPI/OpenMP jobs (changes requested by the User Group). The document is now 50 pages in length, so to keep it at a manageable size we plan to put any more advanced documentation on the WWW within an extended FAQ.

#### 4.1.3 Technical Reports

Of the eight technical reports identified in the Annual Plan, four were due by the end of Q2. Three are already available on the WWW, but complications with the LAPI installation have slightly delayed the single-sided report (now due in July). However, an additional report on the performance of a turbulence code means that the following four completed reports are already available at <http://www.hpcx.ac.uk/research/hpc/>. A version of the report on parallel scaling (HPCxTR0301) was submitted to Supercomputing 2003 but, unfortunately, was not accepted.

- HPCxTR0301 Capability Computing, Achieving Scalability on over 1000 Processors, Joachim Hein.
- HPCxTR0302 Performance of the Turbulence Code, Joachim Hein.
- HPCxTR0303 3D FFTs on HPCx (IBM vs FFTW), Adrian Jackson, Gavin J. Pringle.
- HPCxTR0304 Investigating MPI-IO on HPCx, Elena Breitmoser.

#### 4.1.4 Training

We ran a full set of core and advanced HPC courses in Q2, together with another run of the introductory course “Using the HPCx Service”. In order to expand the audience of HPCx courses we plan to run a number of future courses at different locations and in conjunction with other events. The introductory course has been submitted as a tutorial session to the All Hands meeting, and the Optimisation course will be run immediately prior to the HPCx Annual Seminar. We are also trying to arrange a run of the MPI course in Cambridge.

The quarterly training statistics were as follows:

Course days	16
Number of courses	7
Different courses	7
Student-days for HPCx users	54
Student-days for HPCx staff	11
Student-days available for HPCx	180

#### 4.1.5 Workshops / Conferences

A workshop on “Materials Modelling on HPCx” was held at Daresbury in early April and attracted 18 delegates. Preparations are underway for the workshop on “Measuring and characterising parallel scaling on HPCx” (to be held at EPCC in late September) and for the HPCx Annual Seminar at Daresbury on 10th December. This covers all the planned events for this year.

#### 4.1.6 User Group

The first meeting was successfully held over Access Grid in April. It provided useful feedback from the users, and a number of the suggestions were taken on-board immediately (e.g., changing the day on which maintenance sessions take place). Minutes of the meeting are available on the WWW.

#### 4.1.7 Newsletter

The first newsletter was published in early June and mailed to all HPCx users. We also took the opportunity to include a copy with the final issue of UKHEC News. In total we mailed in excess of 2,200 copies, and we will also distribute newsletters at relevant events such as the upcoming all-hands meeting. The second issue is still planned to be available in time for Supercomputing in November.

### 4.1.8 Packages

Support of all packages has been formalised under a package management project, each with their own user account administered by a nominated HPCx member of staff. A total of 17 packages are now supported in this way. We have recently received the source for Gaussian which is in the process of being installed.

## 4.2 Outreach (*Dr Richard Blake*)

Over the last quarter there has been significant progress in a number of the outreach activities:

- The HPCx service has actively progressed the IBM Lifesciences programme with proposals for time on the system currently being peer reviewed by BBSRC. Unfortunately, the contractual negotiations with IBM have been protracted because of potential VAT implications that were only recognised by them late in the day. These are hopefully close to resolution. We will then develop workplans for the HPCx Added Value staff to support the projects. This will allow us to identify the residual resource available for supporting other life-science and medical simulation applications.
- HPCx is planning to follow up its initial programme of supporting a number of BBSRC projects through a meeting to develop a 'Proposal to form a Biological Consortium on the HPCx Facility' during September. The draft of the current proposal will be circulated to BBSRC and EPSRC separately.
- The UK – ETF proposal has been developed further with an emphasis on seeking to exploit the integrated capability of the systems in high-profile scientific demonstrations. The plan is to develop a more strategic framework for interacting with the ETF sites supporting scientific collaborations which demonstrate the types of investigations that can be undertaken on next-generation services. A number of target applications have been invited to develop proposals for projects with the first, a RealityGrid Lattice Boltzman simulation of a complex fluid, set for demonstration at SC 2003.
- Plans are still being developed for a "HPCx Industry Day" but the date will probably shift into December to be nearer to the HPCx Annual Seminar. Support from IBM and the various Regional Development Agencies is being sought.

- Work has progressed on reviewing the international portfolio of HPC applications in particular those being considered as part of the Blue Planet project.

### 4.3 Terascaling Applications (*Dr Martyn Guest*)

The work described below covers the period April-June 2003, and details evaluation and development terascaling activities around application codes, libraries and tools, plus details of staff training, and attendance at Consortium meetings and associated events, including presentations by members of the Terascaling Team.

#### 4.3.1 Computational Materials

##### *AIMPRO*

- Performance issues with AIMPRO are centered around the diagonalisation. Current work is focussed on profiling the current version of the code and using typical matrices from AIMPRO as testbeds for looking at diagonalisers in general.

##### *Castep*

- A customised version of MPI\_AllToAllV has been written specifically for Castep. The processors are subdivided into intra-LPAR and inter-LPAR communicators and MPI\_AllToAllV is changed to use MPI\_GatherV (intra-LPAR), local re-arrange on LPAR master processor, MPI\_ALLTO\_ALLV between master processors (inter-LPAR), local re-arrange, MPI\_ScatterV (intra-LPAR). Results show up to 33% speed-up for test cases with large numbers of processors in the main communicator (i.e. Newtep 'G' parallelisation). This is important for physically realistic jobs approaching capability computing.
- This enhancement has been incorporated in the released version of Castep.
- Work continues on Newtep. It is organised in several large modules broadly reflecting different functional aspects of the program. At least one of these modules is now too large to compile at O3 level and has been systematically split into smaller modules, without affecting the referencing statements in other modules. Some memory leaks have been captured and removed. The latest version has one MPI-bug fixed with another under investigation.

### *Crystal*

- Improved linear dependence checking in CRYSTAL, which not only indicates when a problem has occurred, but also provides a diagnosis.
- Implemented a better Fock matrix mixing scheme
- Investigations have started into methods for exploiting symmetry in the diagonalisation stage of CRYSTAL.
- Implemented first version of improved restarts for CRYSTAL. This allows restarts of the eigenvectors, which in some cases is more flexible than the normal methods.
- Work on exploitation of k and spin parallelism in CRYSTAL. The form of the basic data structures has now been decided and written, as have some of the simpler compute routines.
- Continued to reduce the amount of replicated memory, especially in the DFT portion of the code.
- Continued preparing the Rusticyanin demonstrator calculations. This has involve checking that the basis set is suitable for this case and checking that the structure derived from the PDB is chemically sensible (e.g. in PDB files hydrogen atoms are often missing).

### **4.3.2 Molecular Simulation**

#### *Plato*

- Following an investigation of the performance and scaling of this code, it is clear that the diagonalisation routines are the limiting factor. Current work is focussing on implementing a range of different diagonalisers within the code in order to investigate the relative performance.

#### *NAMD*

- At the request of Peter Coveney (e10 consortium), the performance and scaling of NAMD 2.4 and 2.5 has been investigated, using one of his data sets. This scaling study has been used to provide a bronze seal of approval for the code, providing the consortium with a 5% discount on 256 processors. The results are of general interest to all NAMD users, as they demonstrate that 2.5 shows a considerable improvement in performance. Current work is focussed on examining the performance on a larger data set.

#### *AMBER*

- Extensive profiling has been carried out on this code. Current work is focussing on the performance of the FFT routines and of the collective communications. HPCx are also in contact with Robert Duke, who has been working on improving the performance of Sander 6.

### 4.3.3 Atomic and Molecular

#### *PRMAT*

- ScaLAPACK divide-and-conquer diagonalisation routines have been incorporated into the PRMAT codes and results are being validated. This involved restructuring the data distribution to 2D block-cyclic. Timing comparisons with PeIGS are currently being undertaken. ScaLAPACK is expected to be significantly faster.
- New internal region codes have been ported to HPCx, which allow significantly larger problem sizes to be set up and run.

### 4.3.4 Molecular Electronic Structure

#### *Global Arrays and LAPI*

- The GA/LAPI and ScaLAPACK/MPI matrix multiply benchmarks that illustrate existing performance problems have been lodged with IBM for performance analysis / Federation roll-out. We await feedback.
- The matrix-matrix multiply algorithm within the GAs is being re-written adding algorithms that exploit non-blocking algorithms and LAPI vector functionality (e.g. LAPI-GETV). This should be completed shortly and will shed further light on the benchmarks above.

#### *GAMESS-UK*

- A workplan has been developed that brings some of the techniques successfully used in CRYSTAL to bear on the GAMESS-UK code. This involves the definition of data structures that will be used to hold the distributed objects and the subroutines that will be used by both GAMESS-UK and CRYSTAL to handle them. This work may be done in collaboration with T.Müller at Jülich.
- Exploratory studies on the replicated data version of the code have reduced the number of triangular arrays (dimension  $N*(N+1)/2$ ,  $N$  = number of GTOs) from six to four, using the Global arrays to house the other triangles. This permits the DFT calculation for systems with ca. 6,200 basis functions.

#### *Performance Evaluation of QC Codes*

- To shed further light on the performance of QC software, the GAMESS-US code has been ported to HPCx and an initial set of SCF benchmarks undertaken. These point to major issues surrounding the current implementation of the  $N^3$  matrix algorithms.
- A number of benchmarks illustrating the performance of key modules within the NWChem software (DFT analytic 2nd derivatives, CCSD(T) energy evaluation and MP2 gradient optimisations) have been undertaken

together with the team at PNNL. These point to latency-sensitive bottlenecks within the software that are to be targeted with a view to enhancing performance on HPCx. A Technical Report is in preparation.

### 4.3.5 Computational Engineering

#### *CFX*

- Following the resolution of CFX license server problems on HPCx, the software is now fully available and upgraded to the latest version (CFX 5.6).

#### *PNEWT*

- Analysed memory usage of PNEWT code on HPCx and forwarded information to developers.

#### *Turbulence Code*

- HPCx were asked to investigate the performance and scaling of a turbulence code, for Professor D. McComb of Edinburgh University, as the group are considering submitting an application for time on HPCx.
- The scaling study is of general interest to HPCx users and a copy of the report has been made available on the HPCx web site at: [http://www.hpcx.ac.uk/research/hpc/technical\\_reports/HPCxTR0302.pdf](http://www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0302.pdf)

### 4.3.6 Environmental Science

#### *POLCOMS*

- The coupled parallel POLCOMS + Wave Model (WAM) is now running and undergoing testing. The WAM code adds significantly to the computational demands of the code, such that in the coupled runs the WAM component takes about 90% of the memory and 90% of the time. Early results indicate that performance scales well and is about 5x faster than the Cray T3E.

#### *Ensemble Modelling*

- Following a request from Lois Steenman-Clark, the Terascaling team have been investigating and developing an ensemble modelling capability on HPCx, which will enable climate models to be run as capability jobs. The work has two components:
  - installation and testing of MPH, a program which allows multiple climate models to be task-farmed;
  - investigation of the I/O performance when several tasks are accessing the disks at the same time, as we would expect from the

task-farming scenario. This work is also highly relevant to other users, and is currently being written up as a Technical Report (as part of the work of the Software Engineering Team).

#### 4.3.7 Libraries

- The performance of ScaLAPACK and PeIGS has been tested on very large (dimensions=7500,12500) matrices output from Crystal (the 6-31G\*\* Crambin case). PeIGS results have been broken down into constituent parts and non-scaling stages of the calculation have been identified. A LAPI-based version of PeIGS has been compared with the MPI version.
- We have continued to investigate the use of a Householder diagonaliser for the first few iterations in Crystal. The results suggest that while some benefit can be gained for small and medium sized cases, the workspace memory requirement for large cases is unacceptable, e.g. while the BFG diagonaliser can cope with 6-31G Rusticyanin (~25,500 basis functions) the ScaLAPACK routines can not.
- PeIGS and ScaLAPACK have been tested on HPCx using Stephen Booth's fast MPI library.
- A technical report on 3D Fast Fourier Transforms has now been written and is available on the HPCx web site at: [http://www.hpcx.ac.uk/research/hpc/technical\\_reports/HPCxTR0303.pdf](http://www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0303.pdf) This report compares the performance of the two main FFT libraries on HPCx: IBM's ESSL/PESSL and FFTW. Both serial and parallel (distributed-memory) 3D complex-to-complex FFT routines are considered.

#### 4.3.8 Tools

##### *TotalView*

- TotalView has been installed and tested under the new interactive service. After a period of internal testing, TotalView has now been released and documented for general use.

##### *DDT*

- Streamline have implemented the DDT debugger on HPCx. Mark O'Connor visited DL and demonstrated the debugger using some simple examples. Attempts with a full-scale application have been unsuccessful and work continues.

##### *libhmd*

- A Perl script is now available to allow users to organise output from Memory-Debugging software (libhmd). Users can check where dynamic memory is being allocated in their programs and automatically view the associated line of source code.

### *Sigma*

- We are testing the Sigma Memory Analyser Tool from IBM Alphaworks, which measures memory access and cache usage, and expect to make it available to general users (license permitting).

### *Vampir*

- The Vampir performance analysis tool from Pallas GmbH is now fully installed and is available for runs up to 512 processors.

## **4.3.9 Consortium Meetings and Presentations**

- M.F. Guest, P. Sherwood, I.J. Bush, W. Smith and H. Van Dam met with T.Müller and K. Wolkersdorfer of KFA Jülich to discuss collaborative approaches to running Quantum Chemistry applications on POWER 4 systems.
- Alessandro Curioni (IBM Zurich Research Laboratory) gave a presentation to HPCx staff on the CPMD code entitled "Reaching Teraflop Performance in *Ab Initio* Molecular Dynamics: the CPMD Code".

## **4.4 Software Engineering (*Dr Stephen Booth*)**

### **4.4.1 Low Level Communications**

Investigation of the low-level communication protocols of HPCx, in particular, MPI, MPI-2 single sided, LAPI and GA-tools.

- *LAPI*: We are continuing to investigate low-level LAPI performance. We believe we have identified a particular problem with the performance of the LAPI\_Get function. We will continue to investigate this and report the results to IBM. This is likely to be one of the causes of the disappointing performance of GA-tools as this uses LAPI\_Get extensively.
- *Optimised Collectives*: The optimised collective routines developed previously have been extended and are now callable from Fortran.
- *LAPI MPI*: One of the current EPCC MSc students is working on a project to port the T3D/T3E MPI library (originally developed in EPCC) to HPCx using the LAPI library. This is progressing well and will be disseminated to HPCx users at the end of the project. This work is proving to be very useful in understanding the trade-offs in building communication libraries on IBM systems.

#### 4.4.2 Grid Integration

- Globus-2 is now working on HPCx with access permitted from machines on the ETF test grid. We are planning to extend this access generally.
- So far only one user has requested a Grid certificate for use on HPCx.

#### 4.4.3 MPI and Mixed Mode Programming

- *Single sided Communications:* A report comparing the performance characteristics of MPI-2 single sided and LAPI communications is under development.
- *MPI-IO performance:* A report on MPI-IO on HPCx is now available at <http://www.hpcx.ac.uk/research/hpc> .
- *MPI performance:* We have been investigating the most efficient way of performing boundary exchanges using MPI. Boundary exchanges are an important part of many applications. MPI has a large number of different communication calls so there are many different ways of implementing boundary exchanges. We have been investigating these to determine which method is most effective on HPCx. This work is being written up as a technical report.

#### 4.4.4 General Terascaling Techniques

We have performed a number of investigations into general performance issues on HPCx.

- *Scalability report:* A technical report on issues affecting the scalability of programs to thousands of processors on HPCx is available at <http://www.hpcx.ac.uk/research/hpc>; we have submitted an abstract covering this work to the All-Hands meeting in Nottingham.
- *I/O performance on HPCx:* We have performed an in-depth study of the I/O characteristics of the HPCx looking at the performance of the GPFS file system and the impact of AIX disk caches and the switch network on I/O performance. This work was originally undertaken as part of an investigation into ensemble modelling but the results are applicable to most users of HPCx. The work is currently being written up as a technical report.

#### **4.4.5 System Administration Functions (SAF)**

- Development of the SAF has continued during the last 3 months. As expected the overall level of effort on this activity has been reduced. Developments have included:
  - Improved error reporting.
  - Requesting new passwords through the Web interface.
  - Additional disk use reporting in project reports.
  - Improved use profile handling and capacity planning.
  - Improved PDF format output for project reports.
  - Security improvements.
  - Usability improvements.

### **4.5 Systems and Networking (*Mr Mike Brown*)**

#### **4.5.1 Staffing**

No change. Currently, staff are available on a two-shift basis (with one at weekends) with out-of-hours on call. This level of cover is being reviewed as the increased stability of the system in recent months does not require such intense on-site support.

#### **4.5.2 System Configuration**

4 x LPARs are now configured as an interactive parallel region to support parallel test and development, particularly using Totalview.

#### **4.5.3 Hardware Configuration**

No changes.

#### **4.5.4 Software Test & Development System**

A 2 x p690 test and development system was delivered towards the end of June. This system will be used principally to evaluate and validate efixes, PTF sets and AIX ML updates so as to avoid rolling out untried new software onto the production system.

#### 4.5.5 Reliability/Stability

There has been a substantial increase in reliability since the beginning of March, due principally to the application of PSSP PTF set 20 and changes to where low-level LoadLeveler functions are run.

#### 4.5.6 IBM Support

Good on-site support remains available. IBM recently proposed changes to the support structure in the UK as a result of input from HPCx, AWE and ECMWF. These will enable access to a more-focussed HPC support group and are now starting to come into operation.

#### 4.5.7 HSM under TSM

Full implementation and roll-out to users requires version 5.1 of HSM client software which has just gone on GA as from last week. Planning is on hand to install, test and validate this (possibly on the software test and development system).

#### 4.5.8 Phase 2 Migration

Discussions are ongoing with IBM as to how best plan and implement the eventual migration to a Phase 2 service platform by Q3/04.

### 4.6 Staffing

<i>AV</i>	<i>April</i>	<i>May</i>	<i>June</i>
DL	3.9	3.9	4.3
EPCC	9.6	7.7	8.6
Total	13.5	11.6	12.9

<i>Systems</i>	5.4	5.4	5.6
----------------	-----	-----	-----

## 5 Summary of Performance Metrics

<i>Metric</i>	<i>TSL</i>	<i>FSL</i>	<i>April</i>	<i>May</i>	<i>June</i>
Technology serviceability	80%	99.2%	99.9%	99.8%	100.0%
Technology MTBF (hours)	200	300	732	732	∞
Number of AV FTEs	7.5	10	13.5	11.6	12.9
Number of training days per month	30/12	40/12	7/1	16/2	16/3
Non in-depth queries resolved within 3 days	85%	97%	100.0%	100.0%	100.0%
Number of A&M FTEs	3.75	5.75	5.4	5.4	5.6
A&M serviceability	80%	100%	100.0%	99.9%	99.5%

<i>Colour</i>	<i>Meaning</i>
	Exceeds FSL
	Between TSL and FSL
	Below TSL

## Appendix A: Incident Severity Levels

**SEV 1** --- anything that comprises a FAILURE as defined in the contract with EPSRC.

**SEV 2** --- NON-FATAL incidents that typically cause immediate termination of a user application, but not the entire user service.

The service may be so degraded (or liable to collapse completely) that a controlled, but unplanned (and often very short-notice) shutdown is required or unplanned downtime subsequent to the next planned reload is necessary.

This category includes unrecovered disc errors where damage to filesystems may occur if the service was allowed to continue in operation; incidents when although the service can continue in operation in a degraded state until the next reload, downtime at less than 24 hours notice is required to fix or investigate the problem; and incidents whereby the throughput of user work is affected (typically by the unrecovered disabling of a portion of the system) even though no subsequent unplanned downtime results.

**SEV 3** --- NON-FATAL incidents that typically cause immediate termination of a user application, but the service is able to continue in operation until the next planned reload or re-configuration.

**SEV 4** --- NON-FATAL recoverable incidents that typically include the loss of a storage device, or a peripheral component, but the service is able to continue in operation largely unaffected, and typically the component may be replaced without any future loss of service.

## Appendix B: Current Projects

### EPSRC Projects

<i>Code</i>	<i>Class</i>	<i>Title</i>	<i>PI</i>
e01	1	UK Turbulence Consortium	Prof Neil Sandham
e02	1	Ab-initio simulation of covalently bonded materials	Dr Patrick Briddon
e03	1	Multi-photon, electron collisions and BEC HPC consortium	Prof Ken Taylor
e04	1	Chemreact Computing Consortium	Prof Jonathon Tennyson
e05	1	Materials Chemistry using Terascaling Computing	Prof Richard Catlow
e06	1	UK Car-Parrinello Consortium	Prof Paul Madden
e07	2	Turbulent Plasma Transport in Tokamaks	Dr Colin M Roach
e08	2	Organic Solid State	Prof Sarah Price
e09	2	Molecular Properties and their Geometry	Prof Peter Taylor
e10	1	RealityGrid	Prof Peter Coveney

### NERC Projects

<i>Code</i>	<i>Class</i>	<i>Title</i>	<i>PI</i>
n01	1	Large-Scale Long-Term Ocean Circulation	Dr David Webb
n02	1	NCAS	Prof Alan J Thorpe
n03	1	Computational Mineral Physics Consortium	Dr John Brodholt
n04	1	Shelf Seas Consortium	Dr Roger Proctor
n05	2	Non-linear Wave-particle Instabilities in Plasmas	Dr Mervyn Freeman

### PPARC Projects

<i>Code</i>	<i>Class</i>	<i>Title</i>	<i>PI</i>
p01	1	Atomic Physics and Astrophysics	Prof Alan Hibbert

## BBSRC Projects

<i>Code</i>	<i>Class</i>	<i>Title</i>	<i>PI</i>
b01	2	Quantum Chemistry Studies of the Rusticyanin Protein Crystal	Prof Samar Hasnain

## Early User Projects

<i>Code</i>	<i>Title</i>	<i>PI</i>
y001	Materials	Dr Patrick Briddon
y002	DNS of Turbulent Flow	Prof Neil Sandham
y003	Multi-photon and Electron Collision Processes	Prof Ken Taylor
y004	Materials	Prof Jonathon Tennyson
y005	UKAEA	Dr Tim Hender
y006	UK Car-Parrinello Consortium	Prof David Price
y007	Climate Modelling	Dr Lois Steenman-Clark

## HPCx Projects

<i>Code</i>	<i>Title</i>	<i>PI</i>
z001	HPCx Support	Dr Alan Simpson
z002	Systems and Operations	Mr Mike Brown
z003	Test Project	Dr Denis Nicole
z004	HPCx Training	Dr David Henty
z05	Outreach Projects	Dr Richard Blake
z06	Application Porting	Dr David Henty
z07	Package Installation	Dr Mike Ashworth