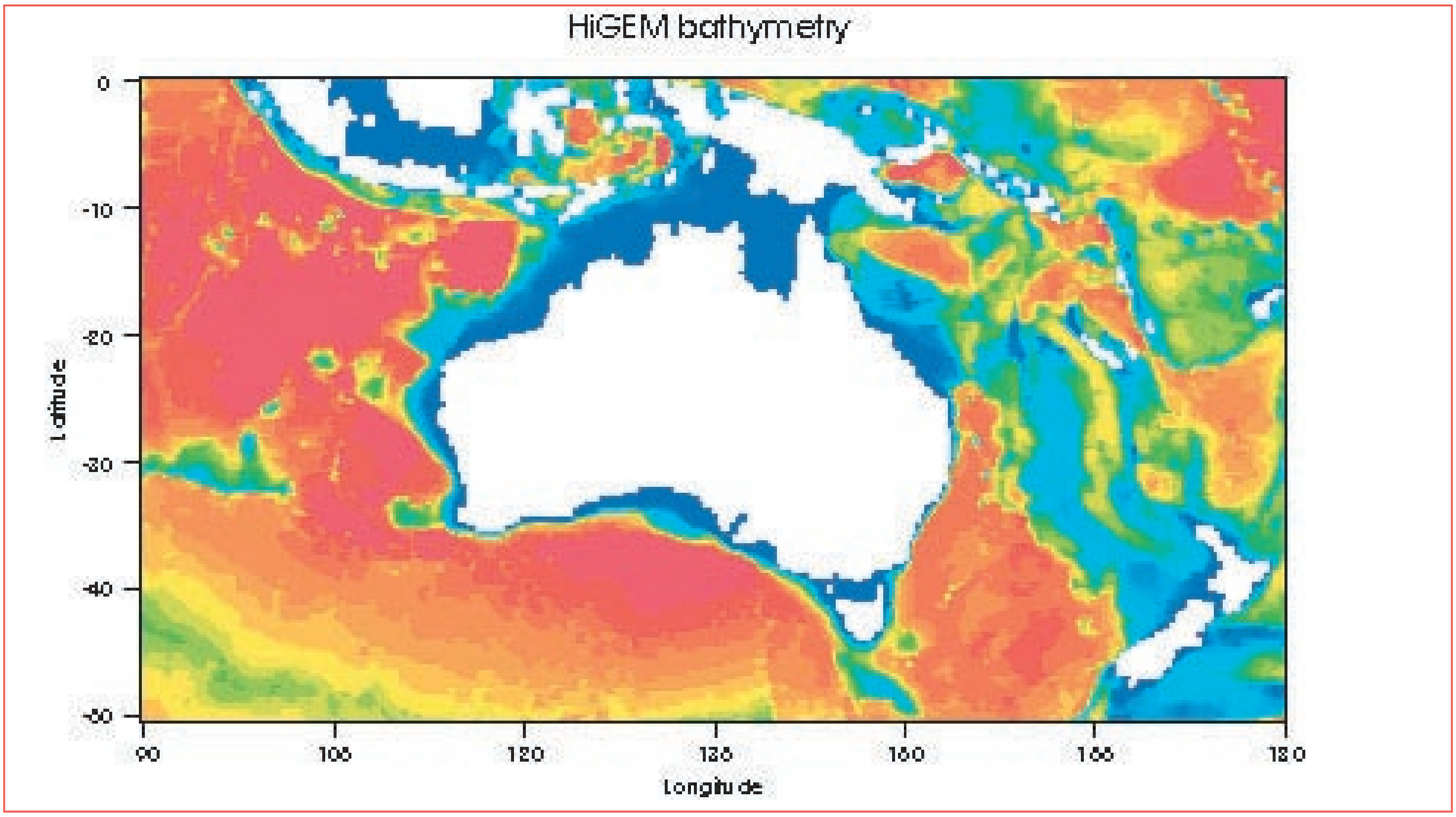
Capability Computing

The newsletter of the HPCx community

[ISSUE 3, SUMMER 2004]



Modelling the global environmental on HPCx

CONTENTS

Editorial

Andrew Sunderland, *HPCx Terascaling Team*

It is a pleasure for me to introduce the third edition of the *Capability Computing* newsletter. The HPCx user-base has continued to grow apace, and now has almost 400 registered users from 34 different projects. I would like to take this opportunity to extend a warm welcome to the nearly 100 new users who have joined since our last newsletter.

The HPCx service is now over 18 months old, and we feel it would be useful to report on some of the latest scientific research being undertaken on HPCx. Hence the theme for this edition is 'Delivering Science'. We present three articles from various HPCx consortia members, who report on their latest findings from a wide range of applications, including Computational Fluid Dynamics, Atomic and Molecular Physics and Environmental Modelling.

Elsewhere in this edition we report on the recent Phase2 upgrade to the service, with an insight into the 'behind the scenes' practical challenges our system administration team faced in ensuring a smooth transition for our users.

Finally, we also include articles detailing recent and upcoming events, including the HPCs Industry Day event held at Daresbury. This event in particular has helped raise awareness of the HPCx Service amongst potential customers from commercial and industrial organizations.

How to contact us

HPCx website: www.hpcx.ac.uk
Helpdesk: support@hpcx.ac.uk
Telephone: 0131 650 5029
Fax: 0131 650 5029

UoE HPCx Ltd
c/o EPCC, University of
Edinburgh, JCMB, Mayfield Rd,
Edinburgh, EH9 3JZ

Come and see us at SC2004…
SC2004 'Bridging Communities'

6th–12th November, Pittsburgh, USA

Damian Jones, *HPCx Administration Team*

SC2004, the world's leading conference on high performance computing, networking and storage, will be held in the brand new David L. Lawrence Convention Center in Pittsburgh on November 6-12, 2004. The Conference brings together representatives from many technical communities to exchange ideas, celebrate past successes and plan for the future.

The theme of this year's event is 'Bridging Communities' and the Conference will utilise state of the art technology, such as Access Grid, to allow participants from around the world to join. At the convention center itself, the various educational and technical programmes will all aim to create bridges to new communities, following the theme of the Conference.

As with previous Supercomputing Conferences there will be a wide variety of industrial and research exhibits, showing the latest technology available and also developments in University, Government and non-profit organisations.

Both CCLRC Daresbury Laboratory and EPCC, University of Edinburgh will be exhibiting at SC2004, and it is hoped that our booths will be co-located – as at previous events – to allow us to also provide a shared HPCx exhibit.

One of the highlights of SC2004 will be StorCloud, a multi-vendor resource available to conference participants that could reach a petabyte of random accessible storage. The goals of the StorCloud initiative are to…

- showcase evolutionary and revolutionary HPC storage technologies in a heterogeneous environment.

- provide 1 PetaByte (10^{15}) of randomly accessible storage to StorCloud participants.

- approach a 1 TeraByte (10^{12}) per second infrastructure bandwidth.

- provide a 1 GigaByte (10^{9}) per second backup bandwidth.

- incorporate/leverage SCinet infrastructure.

- manage and allocate resources to SC2004 participants.

Another highlight will be the InfoStar initiative whose goals are to:
(1) provide real-time information about multiple aspects of the conference to all participants
(2) create a searchable knowledge base about conference events and attendance for the benefit of future SC conference planners.
InfoStar will allow the Conference organisers the opportunity to inform participants of any late changes to the Conference schedule and provide updated information on speakers and presentations, as well as information about the exhibition and exhibitors booths. InfoStar will also help build up a record of information about SC2004. Participants will be able to access the information over the wireless network in the center and also at dedicated InfoStar kiosks located around the Conference hall.

Previous SuperComputing events have provided an invaluable source of information and an excellent opportunity to make contacts with various vendors, resellers and research institutions. Hopefully SC2004 will be as good, if not better.

All information, and logo, from the official SC2004 web site: http://www.sc-conference.org/sc2004.

Picture of David L Lawrence Convention Center from http://www.pittsburghcc.com.

Events

July 2004

8 Course: Improved Perfomance Scaling on HPCx (EPCC, Edinburgh)

9 New Science from Capability Computing: The Second HPCx Annual Seminar (National e-Science Centre, South College Street, Edinburgh)

13-14 Workshop: Networks for Non-Networkers (NFNN) University College London
http://grid.ucl.ac.uk/NFNN.html

15-16 2nd Annual RealityGrid Workshop (Royal Society of London)
http://www.realitygrid.org/workshop-2004/index.html

August 2004

9-12 ScicomP 10: IBM System Scientific User Group (University of Texas, Austin, TX)
http://www.spscicomp.org/

September

2-4 CCP5 Conference: Tapton Hall, University of Sheffield CCP5 is the collaborative computational project for computer simulation of condensed phases.
www.ccp5.ac.uk/meetings/ann2004/ann2004.html

20-24 Ab Initio Modeling in Solid State Chemistry (MSSC2004) Department of Chemistry, Imperial College London
http://www.cse.clrc.ac.uk/events/MSSC2004/

October

27 Course: Optimisation Techniques for the POWER4 Processor (EPCC, Edinburgh)

28 Course: Improved Performance Scaling HPCx (EPCC, Edinburgh)

November

6-12 SC'04: Pittsburgh, PA, The Supercomputing Conference, 2004
http://www.sc-conference.org/sc2004/

29-30 CCP9 Conference, Daresbury Laboratory CCP9 is the Collaborative Computational Project for the Study of the Electronic Structure of Condensed Matter
http://www.ccp9.ac.uk/index.html

MSSC2004: Modelling in Solid State Chemistry School

20-24th September 2004

Ian Bush
HPCx Terascaling Team

Our previous edition of *Capability Computing* carried a report on ab initio electronic structure calculations using the CRYSTAL03 code on HPCx. Some readers may be interested in attending a Summer School on the ab initio simulation of crystalline solids organised by the Computational Science Department of the Daresbury Laboratory, the Department of Chemistry of Imperial College, and the Davy Faraday Laboratory of the Royal Institution. The school will last five days between 20-24th September 2004.

Ab initio modelling has become an ever-increasing area of interest in solid state chemistry and materials science. The software currently available for the quantum mechanical study of electronic properties of crystalline systems features the following:

1. The programs are reliable, their acquisition and use is easy and comparatively inexpensive;
2. The range of potential applications is huge;
3. A variety of properties of great importance can be calculated;
4. Widespread experience and general consensus exist about the practical limitations of the programs and quality of results.

These properties lead to widespread use by a rapidly growing community of non-specialized users including material scientists, crystallographers and geologists.

To fully exploit the potential of these powerful tools users may benefit from an assisted introduction to their use from experts in the field.

This School is the seventh in a series. The first two editions took place back in 1994 and 1995, aiming to offer an overview on the state of the art in the quantum mechanical ab initio modeling of crystalline materials. Since 2000, the MSSC School has taken place annually.

The school is addressed to PhD students, Post-Docs and researchers with interests in solid state chemistry, physics, materials science, surface science and catalysis, and will provide an overview of the possibilities offered by current ab initio quantum mechanical techniques in characterising crystalline solids.

The capabilities of CRYSTAL03 will be illustrated, with hands-on tutorials organised in the afternoon sessions.

Further details can be found at http://www.cse.clrc.ac.uk/events/MSSC2004.

Exploit 70 Tbyte of offline tape storage

Elena Breitmoser
HPCx Terascaling Team

HPCx tape library. Photo courtesy of Tim Franks.

The Phase 2 HPCx system offers a total of 70 Tbyte of offline tape storage. Users who create a large amount of data can benefit significantly from archiving data to tape to free up their disk space.

The HPCx archiving system is build upon the Tivoli Storage Manager (TSM)[1], version 5, release 2. TSM offers both command line and GUI-based modes for archival and retrieval of data, which are described in detail in the 'HPCx Archiving User Guide, V 1.2'[2]. Unlike other tape storage devices, e.g. the T3E that EPCC hosted, the HPCx device does not implicitly migrate data automatically. The migration of data from file to tape is done at the time the user makes the request.

To use the HPCx tape archive, a user's project must first be set up for tape access. This process is explained in [2].

1. IBM Tivoli Storage Manager: www-306.ibm.com/software/tivoli/products/storage-mgr

2. HPCx Archiving User Guide V1.2: www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0405.pdf

HPCx Industry Day

Damian Jones, *HPCx Administration Team*
Richard Blake, *HPCx Applications Outreach Team Leader*

CCLRC Daresbury Laboratory hosted the HPCx Industry Day on 5 April 2004. The main objectives of the meeting were to:

- introduce, raise awareness of, and demonstrate how Terascale class High Performance Computing systems such as HPCx and successive generation of facilities can meet the challenges of industrial R&D
- promote the skill-base available in HPCx for efficiently and effectively exploiting high performance computing systems, developing new scientific functionality and simulation technologies
- explore the scale and scope of potential commercial interest in the HPCx service and subsequent generations of facilities
- explore the quality of service required by industrial users of academic research High Performance Computing services.

The audience was made up of potential users of the HPCx service from the industrial sector, a selection of software vendors and academic researchers in addition to Research Council officials with high-performance computing interests. Talks covered the areas of computational engineering, life sciences, environment, materials and chemistry simulations. We were particularly interested in over-viewing the impact that systems with sustained performances of 1 Teraflop, 10 Teraflops and 100 Teraflops would have on industrial R&D applications.

The day began with an official welcome from Prof. Paul Durham, Director of the Computational Science & Engineering Department at Daresbury Laboratory. Alan Simpson (HPCx) provided an 'Overview of the HPCx Service' and Martyn Guest (HPCx) described some of the challenges of 'HPC - Scaling to 1000's of Processors'. Adrian Mulholland (University of Bristol) overviewed 'Computational enzymology: modelling enzyme catalysis with HPC'. Mike Payne (University of Cambridge) described 'Next Generation Technologies for First Principles Atomistic Simulation' and Mark Sansom (University of Oxford) reported on 'Large Scale Simulations of Biological Membranes'. Ben Slater (The Royal Institution) discussed 'HPC in Atomic Simulations of Materials', Jonathon Chin (University College London) overviewed 'HPC in Modelling Complex Fluids' and Phil Tattersall (Qinetiq) described 'Uses of High Performance Computing in Aerodynamics and Aeroacoustics'. The day ended with a discussion of the quality of service and software environment that would need to be provided to make the HPCx service attractive to industry.

A total of sixty people attended the Industry Day - twenty from software and hardware vendors and resellers, eight from Universities, twenty-seven from research organisations and five from industrial organisations. Ideally, we would have liked to attract more people from different industrial organisations to the event but some excellent presentations made the exercise very worthwhile. Most of the talks are now available on the web-site. The HPCx service will progress its engagement with industry on a number of fronts, in particular:

- holding one-on-one meetings with the industrial groups that attended the meeting, and those that expressed interest, on the quality of service that they would require
- discussions with commercial software vendors on how to tempt their high-end users onto the facility, one key issue here is licensing costs
- discussions with key academics on how to support them bringing their industrial collaborators onto the service

To discuss commercial/ industrial access to the HPCx service please contact:

Mark Parsons, EPCC
0131 650 5022/m.parsons@epcc.ed.ac.uk

R J Blake,CCLRC Daresbury Laboratory
01925 603372/r.j.blake@dl.ac.uk

Using a deforming domain to study fluid flow

Chris Yorke
The University of Southampton

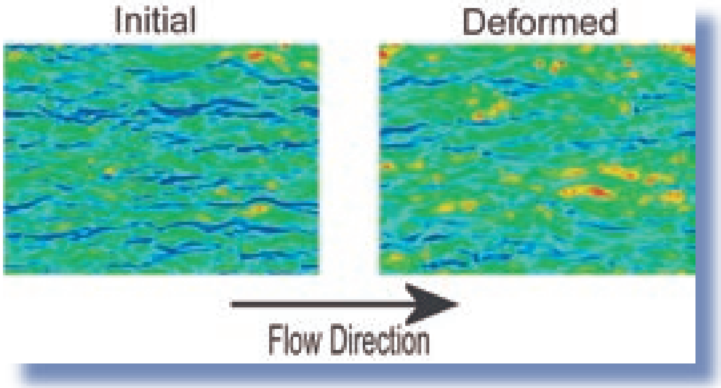


Fig. 1: Initial and deformed domain of time-developing strained-channel flow.

For many years now engineers have been using computers in order to investigate the behaviour of turbulent fluid flows. Originally, empirical models were used as the exact calculations were far beyond the computer power available at the time. As the power of computers has increased a near exact solution to flows has become possible though the use of Direct Numerical Simulation (DNS). This method solves all the scales of the turbulence rather than modelling them (see Fig. 2). Even with today's super computers such as HPCX, the range of situations to which DNS can be applied is limited. This fact leads engineers to design novel geometries which maintain simplicity, yet hold useful analogies to actual flows. The deforming channel is one such geometry.

Fig. 1 outlines the geometry of the flow which is currently being studied. The domain is deformed in time with a constant strain rate while the channel walls are gradually accelerated to reduce the relative bulk velocity. This strategy was designed to have many of the properties which are exhibited in a turbulent boundary layer which is exposed to an adverse pressure gradient (a flow travelling from a low pressure to a higher pressure).

Previous work has been carried out using this geometry using an initial bulk Reynolds number of 13,750. The current work sets out to increase the Reynolds number by 50% and to look at the flow's response to the removal of the deformation. This recovery process is well known to be complicated and long, yet there is little understanding of the processes which take place. Study of the information resulting from these DNS calculations will, it is hoped, lead to a better understanding of these processes and help the development of improved models for such flows.

To illustrate the size of the problem being solved here it should be noted that in order to gain useful statistics from the high Reynolds number problem 60,000 HPCX CPU hours were required. This generated a five sample ensemble over which the statistical quantities were averaged.

Fig. 3 illustrates the quality of data which can be obtained from the DNS. It highlights the areas which play an important part in the flow's behaviour. When data from different locations in time are compared responses to the deformation and recovery can be noted and compared to the behaviour of existing turbulence models. It is hoped this detailed knowledge will lead to an improvement in the performance of these simple turbulence models.

This project is funded by EPSRC and managed through the UK Turbulence Consortium. Further information on this project can be found in the following publications:

G.N. Coleman, J. Kim, P.R. Spalart, Direct numerical simulation of a decelerated wall-bounded turbulent shear flow, J. Fluid Mech. (2003) 495 pp. 1-18

C.P. Yorke, & G.N. Coleman, Assessment of common turbulence models for an idealized adverse pressure gradient flow, Eur. J. Mech. B/Fluids (2004) 23 pp. 319-337

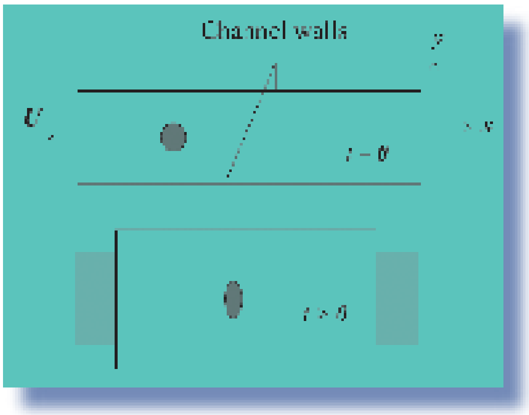


Fig. 2 (above): Slices of the flow a small distance from the wall showing the changes in the flow structure as the domain is deformed. Contours show the variation in the streamwise velocity.

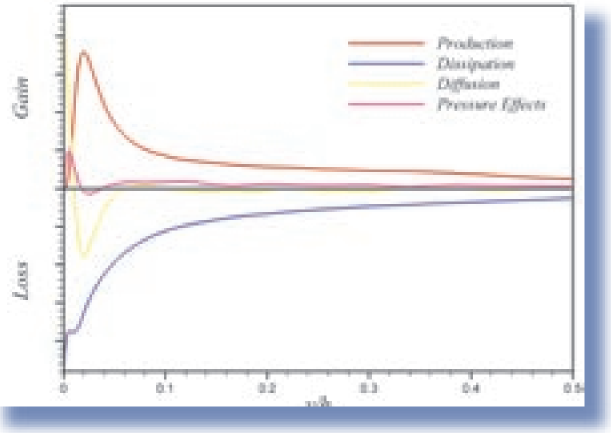
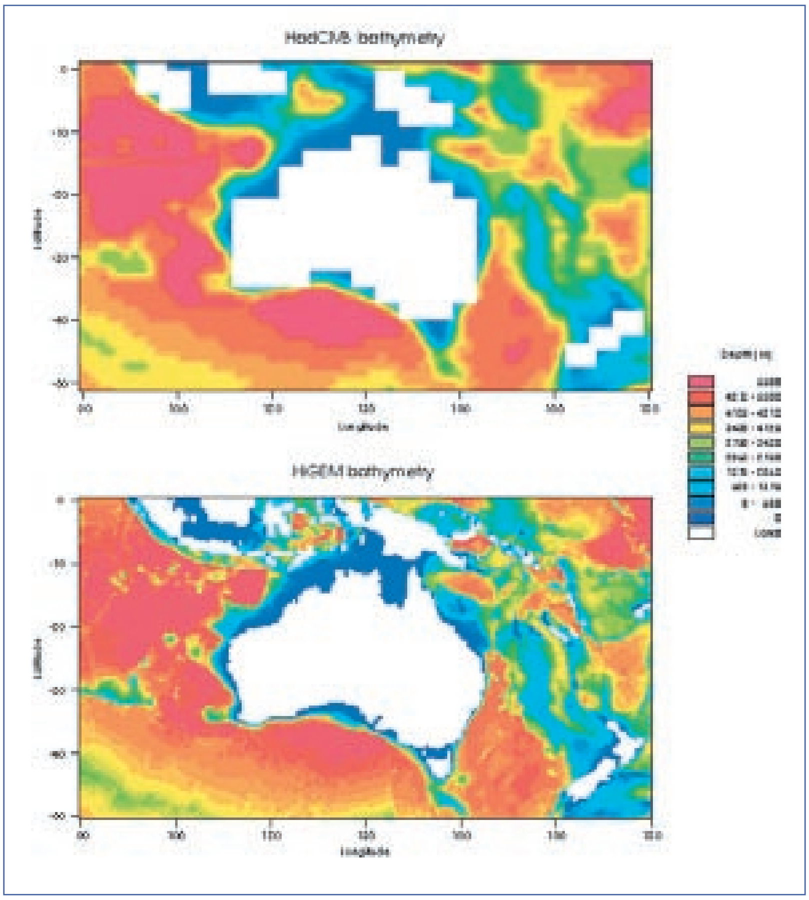


Fig. 3 (right): Turbulent Kinetic Energy Budget for strained flow. Loss/Gain against distance from wall/channel half width.

HiGEM1 on HPCx: High Resolution Global Environmental Modelling

Warwick Norton
NCAS Centre for Global Atmospheric Modelling,
University of Reading



HiGEM is a national UK programme in ‘Grand Challenge’ high resolution modelling of the global environment between NERC and the Hadley Centre of the Met Office. In a three year programme (2004-2006), the new Met Office climate model will be taken to unprecedented resolutions. This will achieve a major advance in the fidelity of simulations of the global environment, and improve our understanding of mechanisms of climate variability and change on time scales of days to centuries.

To model the global environment requires representation of the atmosphere, ocean, ice, and land surface. To do this, the earth is divided into horizontal grid boxes, these are typically 300 km square for the atmosphere and 100 km square for the ocean. However at this resolution, many processes are not properly represented, or represented at all. For example: the impact of mountains on the atmosphere; small-scale eddies in the ocean that form in regions like the gulf stream; polynyas (large cracks in sea-ice) which are sites of intense exchange between the sea and the overlying atmosphere.

The goal of HIGEM is to simulate these processes much more accurately than has hitherto been possible. Initially the HiGEM model will have a resolution of 100 km in the atmosphere and 30 km in the ocean. Subsequently the resolution of the HiGEM model will be increased to 60 km in the atmosphere, and 15 km in the ocean. 100 years of model integration will be run so that important climate processes can be studied in detail. HPCx is the only academic computer in the UK that can provide the necessary computer power. Even so, this will be a very large capacity job on HPCx requiring some 5000 hours using 256-512 processors. The HiGEM model will also be run on the Earth Simulator in Japan as part of a closely related project.

The model code itself is based on the weather forecasting model of the Met Office. It solves forms of the Navier Stokes equations and simulates the many complex processes in the atmosphere, ocean, ice and land surface. The HiGEM model integrations will produce 10s of terabytes of output and an important aspect of project is to verify this output against the latest observational measurements. The complexity of the model and huge task in analysing the output requires a multidisciplinary consortium. The HiGEM consortium (listed below) brings together a wide range of expertise including:

• Hadley Centre: Provision of new Met Office Climate Model, expert advice on climate modelling.

• Centre for Global Atmospheric Modelling: Atmospheric processes, HPC expertise.

• British Antarctic Survey: Polar processes, modelling the cryosphere.

• Centre for Ecology and Hydrology: Land surface processes and modelling.

• Environmental Systems Science Centre : Clouds and radiation processes, model evaluation against satellite data.

• Southampton Oceanography Centre: Ocean processes and modelling, remote sensing.

• University of East Anglia: Ocean processes and modelling.

• British Atmospheric Data Centre: Data management.

During the start up phase of HiGEM, which is currently underway,

Continued opposite.

Calculating electron impact excitation of iron peak elements on HPCx using parallel R-matrix codes

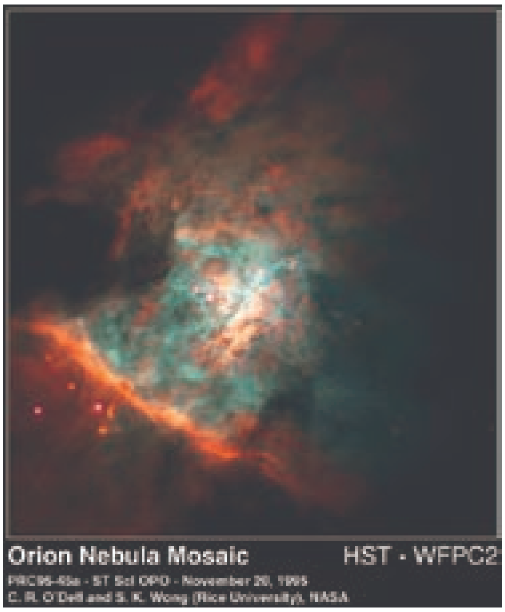
P G Burke, A Hibbert, B M McLaughlin, C A Ramsbottom, M P Scott, *Queen's University, Belfast*; V M Burke, C J Noble and A G Sunderland, *Daresbury Laboratory.*

One of the major outstanding problems in atomic physics is the accurate calculation of collision data for low ionization stages of iron peak elements such as iron, cobalt and nickel. This data is urgently required in the analysis of observations by the Hubble Space Telescope of gaseous nebulae (Figure 1) and in the analysis of laboratory spectra from, for example, laser-plasma interactions and tokamaks.

There are two main difficulties which arise from open d-shells in the target states of these ions. Firstly, a large Configuration Interaction (CI) expansion is required to adequately represent electron correlation effects within the target ion, and secondly, the open d shells give rise to a large number of target states, and in turn to hundreds, or even thousands of closely coupled channels. In addition, in order to resolve low-lying Rydberg resonances, calculations must be carried out over a very fine energy mesh.

These difficulties have necessitated a major redevelopment of the standard scalar R-matrix codes to produce the parallel PRMAT codes that are currently being used for calculations on HPCx. The latest optimised eigensolvers from the Scalapack library have been incorporated in order to address the computational bottleneck associated with diagonalizing large Hamiltonian matrices (of order 10000) in parallel. The PRMAT codes scale well and are able to exploit the high-end computing resources available on HPCx.

These codes have been applied sucessfully in LS coupling to study electron collisions with FeII, FeIII and FeIV, where comparison with previous work demonstrates the importance of including additional correlation effects and coupled channels and the need for using a sufficiently fine energy mesh. Typical results are shown in figure 2, where complex resonance structure can be seen.


Orion Nebula Mosaic
HST · WFPC2
PRC95-45a · ST ScI OPO · November 20, 1995
C. R. O'Dell and S. K. Wong (Rice University), NASA

Other computational innovations based on the R-matrix method and using HPCx facilities include the development of computer codes to study multiphoton ionisation in super-intense laser fields and the development of a 2-dimensional R-matrix propagator to study electron impact excitation and ionisation at intermediate energies.

Further information:
www.hpcx.ac.uk/research/atomic/prmat.html

Figure 1: Orion Nebula, NGC 1976, photographed from the Hubble telescope. Forbidden transitions of Fe ions are observed, for example in FeII.

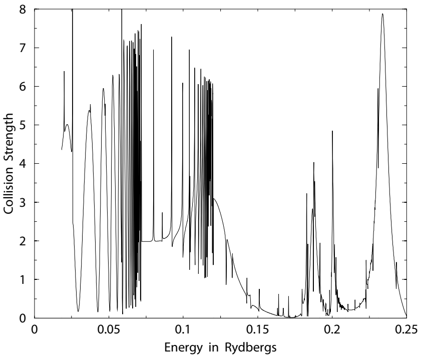


Figure 2: Typical collision strength results for electron impact excitation of FeII, illustrating the complex resonance structure that has to be resolved.

there is significant work developing new grids, boundary datasets, and testing of numerical solvers. It is envisaged production runs of the new high resolution model will start later in the year.

The figure shows how the representation of the ocean will change with HiGEM model compared to our current climate model (HadCM3). Shown are the ocean depths around Australia, red and oranges indicate deep ocean, blue and green shallow ocean, with white regions land. The increased horizontal resolution of HiGEM gives a much better representation of the coastlines, allows more islands, and a more accurate representation of the width of straits and channels. In this region this will allow a better representation of the passage of water through Indonesia which was poor in HadCM3 but known to be important for the global circulation of the ocean.

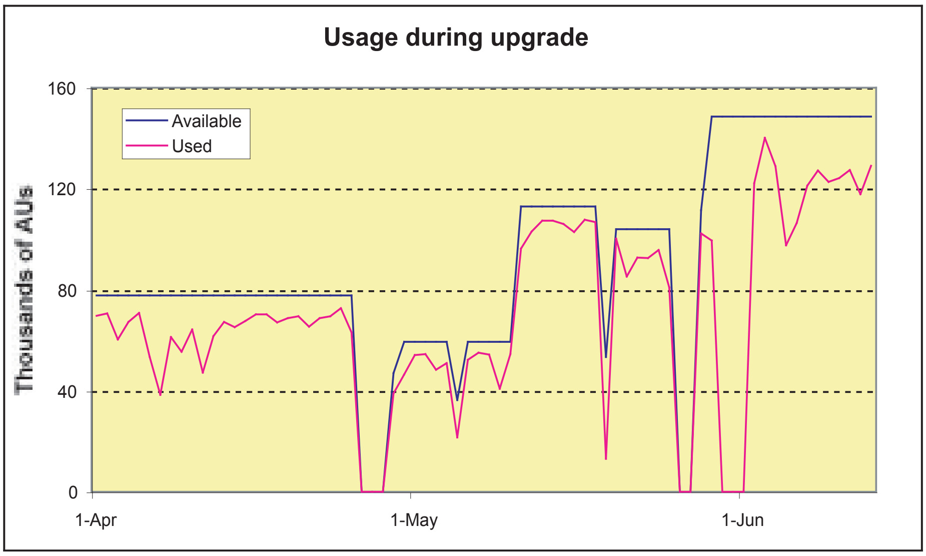
Other benefits of increased resolution in the ocean will include dramatic improvement in the representation of the gulf stream and the North Atlantic current. Increased resolution in the atmosphere will result in much more realistic low pressure systems and storm tracks across the North Atlantic. It is known that both the gulf stream and the storm tracks play an important role in the ocean's thermohaline circulation which releases a huge amount of heat into the North Atlantic and so warms Northern Europe. The fidelity of the simulations produced by HiGEM will provide better assessments of the likelihood of a collapse of the thermohaline circulation, a possibility that is currently receiving much public attention in the film 'The day after tomorrow'

Further information:
www.higem.nerc.ac.uk

A time of change

Alan Simpson
HPCx Project Director

John Fisher
HPCx User Support



HPCx has now moved to the second phase of its development, doubling its performance to more than 6 Tflops. This makes it once again the most powerful academic computing system in Europe, and opens up a new HPC landscape for researchers in the UK

The word upgrade is not really an adequate description of this move. In effect, the Phase 2 platform is an entirely new system. The 40 IBM Regatta frames have been replaced with 50 Regatta H+ frames, giving us 1600 1.7 MHz Power 4+ processors. The communication interconnect between the frames has been replaced by IBM's new High Performance Switch technology and disk space has doubled. The familiar division of each 32-processor frame into 4 8-processor LPARs has gone.

The technical challenges this exercise presented were enormous; this was probably the most testing period in the whole of the HPCx project. The plan that we followed was agreed between HPCx and EPSRC, and required us to keep downtime and disruption to an absolute minimum. As can be seen from the graph, we achieved this objective. For example, in April, when the first part of the change was carried out, the processor time available to users was reduced by less than 10%. In May, although for a time the number of frames in use had to be reduced to 20, more AUs were provided for users than in any previous month.

We can follow the stages in the changeover on the graph. The first frames were installed in the computer room while the old service was still running, and by 31 March there were 20 of them, containing 640 processors. On that date we started a Phase 2 early-use service, during which a number of users tried out the system and provided useful feedback, especially about the new documentation. On 26 April, the Phase 1 service was closed. During the next three days, there was a break while the huge exercise of transferring the service to the new platform was carried out – all the projects and user accounts and the entire file store, nearly 2.5 million files – and the system was configured for general use. Corresponding changes were made to the HPCx database and its software. The Phase 2 service opened to all users on 29 April with 20 frames. During May extra frames were added using normal maintenance sessions. There was another break in service on May 26 and 27 for testing and benchmarking, and the complete platform opened with 50 frames on 28 May.

As can be seen from the diagram, the breaks in service were quite short. It is also interesting to note how closely utilisation follows the changes in the available resource. (The break in service at the end of May resulted from a problem in the external network.)

All this was achieved without any significant problems by close cooperation between Mike Brown, the HPCx systems group and the IBM technical team. You can read a description of how they

Continued opposite

TOP500 – Moore's law in action

The new platform has now been accepted as fully operational by EPSRC. It is rated at 6,188 Gflops on the Rmax figure of the Linpack benchmark. This means that it produces 6,188 AUs an hour, or 54,244,000 AUs a year.

This figure is also used by the TOP500 list to rate the most powerful computing systems in the world. HPCx's figures have been included in the latest edition of this: we are rated at number 18 in the world. Despite doubling our size we have dropped two places!

We aren't disappointed, however. If we hadn't upgraded we would now be around number 40. We are the second system overall in Europe (the first being ECMWF), and the highest-rated academic system.

Phase 3 will have a rating of at least 12,000; if we had it now, we'd be at number 4 in the world, and ahead of every other IBM system.

TOP500: http://www.top500.org
Moore's Law: http://tinyurl.com/38g3p

Doubling the HPCx service capability: the Phase Transition

Steve Andrews
HPCx Senior System Administrator

At 13:00 on Thursday 29 April the interim Phase 2 HPCx service was released to users, one day earlier than the date that had been anticipated some 6 months before. The first user logged on 2 minutes later, and the first LoadLeveler job was running by 13:04.

This was a major step towards the full service that will officially arrive on 1 July 2004. Success in this project was the result of the meticulous planning and hard work by HPCx and IBM which began in October last year.

Phase2 vs Phase1 hardware comparison

	PHASE1	PHASE2
NODES	40 p690 Frames 1.3 GHz Power4 Processors 4 × 8 Way LPARS per Frame 1280 processors in total	50 p690+ Frames 1.7 GHz Power4+ Processors 1 × 32 Way LPAR per Frame 1600 processors in total
INTERCONNECT	SP Switch 2 with Colony PCI adaptors	HPS Switch
DISK	18 Tbytes GPFS	36 Tbytes GPFS
RMAX LINPACK	3.41 Tflop/s	6.19 Tflop/s

What made this transition unusual was the number of constraints that were imposed on the process:

• This was to be a major reconfiguration of both the hardware and system software. Thus it was not possible to operate the Phase 1 and 2 machines as a single system, a situation that would have allowed a seamless migration between the two.

• Since GPFS was structurally incompatible on the two systems all user data had to be migrated quickly and flawlessly. It was this step that determined how long the service would be closed to users.

•All accounts and environments had to be identical on the two machines so that it was not obvious to the users that anything had changed.

• All the necessary codes had to be ported to new versions of:

- AIX
- POE and associated libraries
- the interconnect (the completely new High Performance Switch)
- LPAR layout (32 vs 8 cpus per LPAR)

• Most of the Phase 1 event monitoring and operating procedures were now obsolete and had to be created or revised.

• EPSRC insisted on minimal disruption to a continuous service. In the event only three days of downtime were required to achieve the transition.

While much of this work was done in advance, the physical replacement and data migration had to be initiated and completed with a break in service. In the event all user data was transferred through the backup tape system (TSM) with a total of 2,466,867 user files (4.5TB) migrated at a top rate of 0.25TB/hr. A rigorous set of checks was employed to ensure that all the data had transferred safely.

The original 20 p690 frames of the new system are now at the full strength of 50 compute machines, a total of 1600 1.7GHz Power4+ processors. We are delighted to have doubled both the peak performance (to 6 Tflop/s) and the available disk space (to 36 TB) on time with minimum interruption to the user service.

---

carried the plan through in the article above by Steve Andrews. We also need to thank all our users for their cooperation and forbearance.

Benchmark trials on the new platform indicate that the performance of real jobs has more than doubled, with a particular improvement for those jobs which did not scale very well on the first phase interconnect. We shall now be able to routinely run jobs using 1280 processors. It is clear that this event represents a really significant increase in the capability resource available to UK researchers, and an opportunity to attack problems which up to now have been out of reach.

The upgrade to Phase 3 will double the service's capacity yet again. This will be complete by the end of 2006, and planning is already well under way.

Phase 2 hardware: www.hpcx.ac.uk/services/hardware/index.html
Phase 2 documentation:
www.hpcx.ac.uk/support/documentation/index.html

Figure 1. Scicomp9 was held in San Giovanni in Monte, Bologna. Originally a monastery, it has also been a prison and a special criminal court, before becoming part of the University of Bologna.

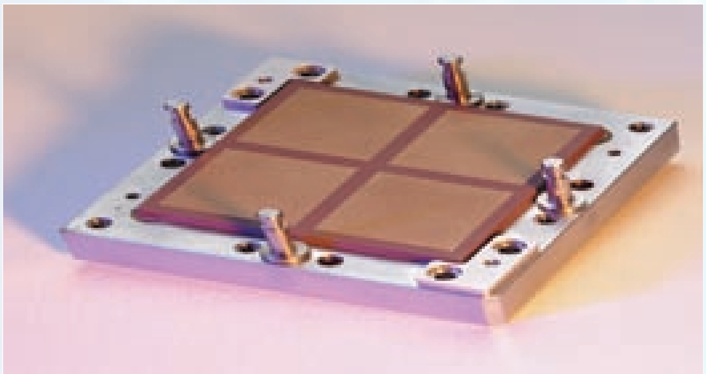


Figure 2. A Multi-Chip Module of HPCx.

Scicomp 9

CINECA, Bologna

Lorna Smith
Applications Support HPCx

The ninth IBM System Scientific Computing User Group (ScicomP9) was hosted by CINECA in Bologna in late March. In addition to a beautiful location (note the fresco in the conference hall!) the meeting was well attended by both IBM developers and HPC Users, and I thought it would be useful to summarise a few of the more interesting points raised at the meeting.

Firstly, the latest version of the (64 bit) MPI library includes some 'cluster aware' collective communication routines. Our own cluster aware routines suggest some users will see significant improvements in their code performance, and we would be keen to hear from anyone who has seen (or not seen!) performance improvement with a collective heavy code.

Also on the MPI front, striping will be including in future releases of MPI. This will allow users to make full use of all four links to the switch without needing at least four separate MPI messages. This will be particularly helpful for users with hybrid MPI/OpenMP codes, allowing full use of the links with only one MPI task per node.

Our move from 8-way to 32-way logical partitions will obviously help those with shared memory codes. Some of the discussion at this meeting however highlighted the increased complexity in the memory hierarchy, which can influence the performance of some codes. For example 32-way nodes consist of four 'multi-chip modules' (MCMs), each of which contain 8 processors. A processor can access memory located on the same MCM faster than memory located on different MCMs.

This highlights the importance of the environment variable MEMORY_AFFINITY, which causes AIX to try and allocate memory on the same MCM as the processor. In some situations the issue is complicated by the operating system migrating processes off an MCM, which can be addressed through processor binding. If you are interested in this for your own code, please contact the helpdesk (helpdesk@hpcx.ac.uk) for further details.

Myself, Joachim Hein and Andy Sunderland all gave talks at this meeting on mixed mode programming, NAMD and parallel eigensolvers respectively.

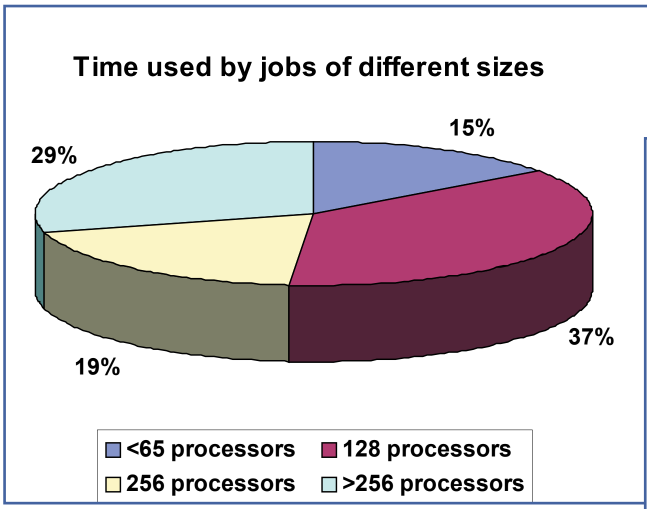
If you are interested in these, or any other talk you can access them at the ScicomP9 website:
http://www.spscicomp.org/ScicomP9/

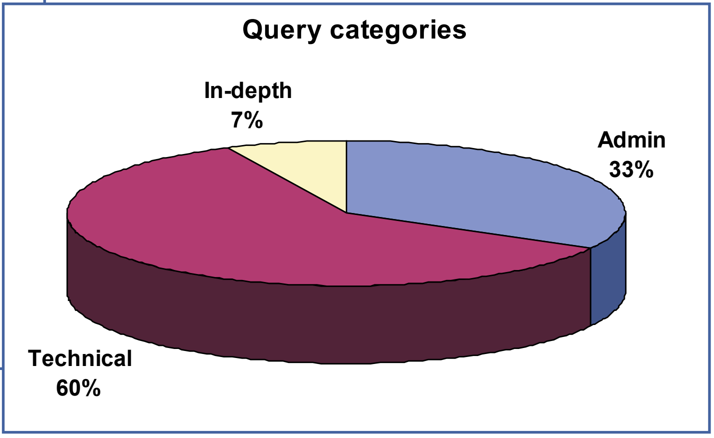
Lastly, we will be hosting ScicomP11 in Edinburgh next year. We hope to see you there.

Scicomp10

Scicomp10 will be hosted by the Texas Advanced Computing Center in Austin between August 9th–13th. The meeting will consist of presentations by both IBM staff and SP users. This year it will be co-located and co-scheduled with SP-XXL, a related group where the focus is on SP system management.

The early registration deadline is July 23rd. For those wishing to present, the abstract submission deadline is July 16th. More details of Scicomp10 snd SP-XXL can be found at:
http://www.spscicomp.org/ScicomP10/agenda.html
http://www.spxxl.org





In the last eight months...

John Fisher,
HPCx User Support

Usage

This report covers the eight months since my last report in Capability Computing, that is, October 2003 – May 2004. Over this period, take-up of HPCx resources has continued to be extremely high, well above our 75% target in most months. In May for the first time it exceeded 2 million AUs per month.

The dip in April resulted principally from time lost during the move to Phase 2. In May we provided more AUs than in any earlier month, despite the reduced number of processors for most of the month, because of very high demand and the higher rating of the processors.

An AU is equivalent to a 1 Gflops processor running for an hour, as measured by the Linpack benchmark (Rmax). The complete Phase 2 platform delivers 6,188 AUs per hour.

The pattern of use by jobs of different sizes has changed radically. Jobs of more than 256 processors, which are classed as capability work, used 29% of the time as compared to 17% in my last report. Large jobs (256 processors) used nearly half the time (48%). The proportion of 128-processor jobs has fallen from 53% to 37%.

Users

On June 8, there were 397 active users in the HPCx database, as compared to 301 in October last year. There were 34 user projects, funded as shown below.

Research council	*Projects*
EPSRC	19
NERC	5
PPARC	1
BBSRC	7
CCLRC	1
Externally funded	1

As a point of interest, we have users from 32 nationalities, including 138 (34%) from outwith the UK.

Helpdesk

We received 728 queries over the eight months of this report. The proportion of administrative queries has dropped from 40% to 33%; we think this reflects the growing maturity of the administrative website. In-depth queries are those which require several person-days of work, or which have to wait for additional input from the user; they include those awaiting action by IBM.

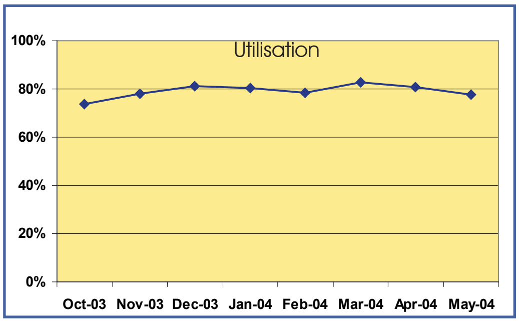
Clear-up time	*%*	*Target*
<24 hours	86.1	75%
<72 hours	99.9	97%
>72 hours	0.1	

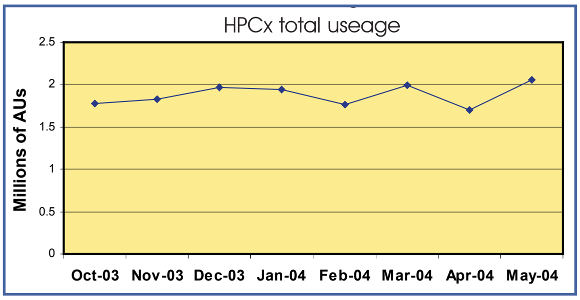
The table shows that we are still fairly easily meeting the prescribed targets for clear-up times of queries.

Training

36 days of training courses have been taught during the eight months; this corresponds to 230 student-days, excluding attendance by HPCx personnel.

Monthly and quarterly reports: www.hpcx.ac.uk/projects/reports





Time and LoadLeveler wait for no man...

In this issue, I outline how you can arrange for your program to be warned some time in advance that it is about to run out of time in the batch queue.

Q: I try always to specify the wall_clock_limit in my LoadLeveler script to be as close as possible to the actual runtime. However, if I set it even slightly too small then my job is killed before it completes and I get no output. What should I do?

A: A brief answer is given below. For full details, see ‘How can I get some warning that my batch job is going to be killed?’ at http://www.hpcx.ac.uk/support/FAQ/.

You’re right that it’s good to specify as small a wall_clock_limit as possible. This is the only information that Loadleveler has about the runtime when it schedules jobs, and there will be times when it is specifically looking for short jobs to fill up temporary gaps in the queues. However, LoadLeveler also takes the limits very seriously and, as you have noticed, will kill jobs that exceed the limit by even a second.

Below I describe a way of setting an alarm call so that your job is notified some specified time (eg several minutes) in advance of being killed so you can save data to disk and exit gracefully.

First a note on how this should be used. It’s very unlikely that your program will be able to checkpoint itself at arbitrary places in the code. However, it is very possible that you have some large outer loop (eg over timesteps or iterations) and that you can, in principle, checkpoint at the end of any loop. In such a case you would test for the alarm at the end of each loop and exit if it has gone off. If your main loop takes about five minutes, and it takes two minutes to save to disk, then setting an alarm value of eight minutes should be safe. Even in the worst case when the alarm goes off immediately after you have tested for it, you will have time to execute one more complete loop and still checkpoint safely.

The procedure relies on the fact that you can specify two wall_clock_limit’s to LoadLeveler: a hard one and a soft one. If you only set a single limit it is taken to be a hard limit, at which point the system terminates your program. However, you can also specify an additional soft limit (less than the hard one) at which point the system issues a signal, SIGXCPU, indicating that this soft limit has been reached. By default this signal is ignored, but a user can choose to trap it and trigger a bespoke signal handler. The simplest approach is to set some special alarm variable whose value can be tested from user code.

The normal LoadLeveler syntax is

```
 #@ wall_clock_limit = hardlimit
```

with the limit specified as hours:minutes:seconds. For example, with

```
 #@ wall_clock_limit = 01:30:00
```

your program will be terminated after 90 minutes. The full syntax is actually:

```
 #@ wall_clock_limit = hardlimit [, softlimit]
```

eg with

```
 #@ wall_clock_limit = 01:30:00, 01:25:00
```

your program would be sent notified (via SIGXCPU) after 85 minutes, five minutes before it is terminated.

I have implemented a couple of simple routines, HPCxAlarmSet and HPCxAlarm, so that you can use this facility without having to bother about the details of signal handlers under AIX.

There is one slight subtlety regarding which program is actually sent the signal. In the normal situation your job comprises a script which LoadLeveler executes. Unfortunately, in this case the system sends the SIGXCPU signal to the script and NOT the user program, and this can cause problems.

The trick is to run a job without any associated script. The Loadleveler parameters ‘executable’ and ‘arguments’ allow you to do this.

If the last lines of your current script are:

```
  #@ queue
  poe ./a.out
```

then you should replace them by

```
  #@ executable = /usr/bin/poe
  #@ arguments  = ./a.out
  #@ queue
```

Note that if your current script doesn’t contain the explicit call to ‘poe’ (eg you simply have ‘./a.out’) you must still specify ‘/usr/bin/poe’ as the executable: for parallel jobs, the system is actually using poe automatically to launch your program at runtime.

To use the alarm system in C programs, you need to include the header ‘hpcxalarm.h’ and link against ‘libhpcxalarm.a’. These files are located in /usr/local/packages/include/ and /usr/local/packages/lib/ respectively.

I am currently working on a Fortran interface: keep an eye on the FAQ entry noted above for news of any progress!

A typical C code would look something like:

```
#include ‘hpcxalarm.h’
void main(void)
{
  ...
  HPCxAlarmSet();       /* Set up the alarm */
  for (loop=0; loop < MAXLOOP; loop++)
  {
    ...                /* Do all the complicated work for this loop */
    if ( HPCxAlarm() )  /* Has the alarm gone off yet? */
      {
        printf(‘WARNING: Alarm Call Received!\n’);
        loop++;
        break;
      }
}
  printf(‘Checkpointing after completing %d iterations\n’, loop);
  checkpoint();
  printf(‘Finished\n’);
}
```

A working test program (including a Makefile and LoadLeveler script) is available under the FAQ entry http://www.hpcx.ac.uk/support/FAQ/alarm/.*