Operation of the Cray T3D as a National Facility

Michael W. Brown, Edinburgh Parallel Computing Centre
The University of Edinburgh, King's Buildings, Mayfield Road,
Edinburgh, EH9 3JZ, UK

Abstract

In 1994 a 256-processor Cray T3D system was installed at the University of Edinburgh on behalf of the UK Engineering and Physical Sciences Research Council (EPSRC) after a lengthy procurement exercise. The University is under contract to the EPSRC to run a peer-reviewed national service for Grand Challenge science to the UK academic and research community. This paper describes the procurement exercise, the installation of the system at Edinburgh and the operational regime necessary to maximise utilisation, throughput and performance for production-quality codes, whilst still enabling code-development access for emerging projects.

Academic Supercomputing Provision in the UK

In recent years the provision of academic supercomputing in the United Kingdom has been within the remit the government-funded Science and Engineering Research Council (SERC), now the Engineering and Physical Sciences Research Council (EPSRC), which has been responsible for this provision on behalf of all of the research councils.

Three National Centres, analogous to the NSF-funded centres in the US, were funded and maintained, these being the Rutherford-Appleton Laboratory near Oxford, a SERC-operated laboratory which has a Cray Y-MP8I/8128, the University of London Computing Centre which has a Convex C3840 system, and the Manchester Computing Centre, based at the University of Manchester, which has a Fujitsu VPX240/10.

Access to these facilities is granted to UK academics and researchers largely through a peer-reviewed process, although small "pump-priming" allocations of time are also available.

In addition to this established vector-supercomputing provision, parallel computing facilities were made available at the University of Edinburgh (Meiko i860 and Connection Machine CM-200), at the University of Manchester (Kendall Square KSR-1) and at the SERC's Daresbury Laboratory (Intel i860) and each site was granted partial funding support for these services.

Access to these facilities was also granted under a peer-reviewed process and quite substantial amounts of access were given to "pump-priming" applications.

The Edinburgh Parallel Computing Centre

The Edinburgh Parallel Computing Centre (EPCC) is an interdisciplinary research centre within the Faculty of Science and Engineering at the University of Edinburgh and was formed in 1990 **[1]**.

Its mission is:

To accelerate the effective exploitation of parallel high performance computing systems throughout academia, industry and commerce and to enhance the international reputation of the University of Edinburgh.

EPCC is responsible for the operation and support of the external user services available on a 64 node Meiko i860 system, a 16K processor CM-200 and the Cray T3D as well as a range of smaller systems used internally for applications development.

EPCC currently employs around 50 staff and is divided into the Consultancy and Support Group which supports academic use of the computing systems and operates a European visitor programme and a Summer Scholarship Programme, the Systems Group which is responsible for the operation and administration of all of the computing systems under the control of EPCC and the Projects Group which is involved principally in undertaking contracts for industrial and commercial customers.

Over the years, EPCC has become a prime focus for parallel HPC activity in Europe, and this allied with the strong industrial and commercial activity greatly aided the University's bid to house the national parallel HPC facility based round the Cray T3D. The commercial work undertaken by EPCC, which includes that on the Cray T3D, is directly in line with the current government policy on wealth creation that was enshrined in the 1993 White Paper under which the Research Council structure was re-organised **[2]**.

1994 HPC Procurement Exercise

During 1992, an evaluation **[3]** by the Scientific Working Group to the Supercomputing Management Group of the SERC (the body that maintained overall responsibility for the provision of HPC to UK academic and research users) was undertaken, and with the expected release of significant capital funds for 1993-1994, it was decided to augment the existing national vector supercomputing facilities with what was expected to be a large parallel system.

Unlike the vector systems, which are used by very large numbers of users from all branches of science, it was expected that the use for the new facility would be tightly focussed on a number of areas of science that could make best use of very high performance computing. It was expected that the new system would complement, rather than replace, the various vector facilities and six areas of so-called "Grand Challenge" science were initially identified as being able to make the most appropriate use of such a facility.

These were:

1. Modelling of Climatic Change
2. Computer Modelling of Biological Molecules
3. Quantum Chromodynamics
4. Computational Fluid Dynamics
5. Simulations of Materials
6. Genome Data Management

The intention to focus on a small number of very large scientific applications was a complete change from the existing policy of access to the vector facilities and in many ways reflected the successful operation of the Edinburgh Meiko i860 facilities which, although centrally funded, were only available to users of two very large consortia.

Aims of the procurement

The aim of the procurement was to provide world-class high performance computing facilities to UK researchers for use in areas where computational work plays a central role by 1994, and that the peak performance of such facilities should be in excess of 10 times over the existing vector facilities.

It was intended that the system be used for running production applications, and not for research into the uses of parallel processing, and thus high reliability to maximise the productive use of the system was required.

Procurement strategy

A technical working group, called the Procurement Exercise Group, was set up late in 1992, this group being tasked with formulating the RFI (Request For Information) and OR (Operational Requirement) documents, evaluating the responses from manufacturers, drawing up short-lists and eventually making a full technical appraisal of the final responses to the tender and recommending a system for purchase.

Six vendors were invited to respond to the OR document, and five were chosen as suitable for benchmarking and subsequently were invited to tender in October 1993.

The responses from the vendors to the requirements within the OR were scored, and the results entered into an evaluation model which weighted certain aspects (for example the programming environment) against others (for example upgradeability).

The system tendered by Cray Research scored highest in the final technical evaluation, and the recommendation of the group was that the T3D system be procured.

Benchmarking

In parallel with the technical evaluation undertaken by the Procurement Exercise Group, a Benchmarking Working Group was set up whose remit was to construct a set of benchmark codes that would best reflect the intended workload of the system. This benchmark consisted principally of adapted existing production codes, but also included a synthetic benchmark suite, and it was a major task to formulate a benchmark that would be suitable for running against a number of different hardware platforms.

The eventual suite consisted of around 12 adapted user codes, plus the synthetic suite, and the vendors who were short-listed for benchmarking were required to run the complete benchmark both in an "as-is" form (where only the changes necessary to make the applications actually run were permitted) and in a fully optimised form for various sizes of system.

A benchmark witness team travelled to the US in August 1993 to supervise the running of the benchmarks and to engage in technical discussions with the vendors.

In response to the final tender, each vendor had to submit a Best and Final Offer and this contained projected benchmark performance figures which would become part of the contractual obligations made by the successful vendor.

The results from the benchmark exercise were subsequently analysed, and although none of the systems benchmarked came out top in every application, the overall performance of the Cray T3D system justified it being recommended by the Benchmark Working Group.

System selection

On the recommendations of both the Procurement Exercise Group and the Benchmark Working Group, the Cray T3D system was selected for procurement, and the contract was let to Cray Research early in January 1994 with a projected date for installation in April.

Site selection

In parallel with the hardware procurement, a separate exercise was underway to choose the site for the system to be purchased. Under the Services Directive issued by the European Economic Community, it was necessary that a full open tender under EEC rules was undertaken for this exercise, and this lengthy process was finally completed in January 1994 when the University of Edinburgh was chosen as the site for the system.

The Edinburgh Cray T3D

Hardware configuration

The system that Cray tendered was a 256-PE Cray T3D, each processor having 64 Mbytes of local memory.

As the entire system was to be stand-alone, a separate Cray Y-MP4E system was also to be provided.

In excess of 200 Gbytes of disc capacity, plus tape devices and network interfaces, was specified, and in addition the installation site was expected to provide around 1 Tbyte of automated mass-storage.

sn6007:
T3D/MC256-8 (256 PE, 8 MWord/PE)
2 x IOG's

sn1904:
Y-MP4E/264
32 MWord SSD
2 x IOC's
36 x DA301 (197 GB) on 8 x DCA-3
8 x DD60 (15 GB) on 8 x DCA-2
2 x STK 4480 on 1 x TCA-1
1 x IBM 3494 ATL system on 3 x TCA-1
2 x FDDI on 2 x FCA-1

Trialling, acceptance and installation

Factory trials were undertaken in Chippewa Falls and were completed on the 15th April 1994.

The trials, witnessed by representatives from EPSRC and the University of Edinburgh, consisted of three phases:

1. Performance trial. The benchmark codes were run and their performances (typically elapsed times) were checked against the figures that Cray had contractually committed to in their Best and Final Offer.
2. Large jobs reliability trial. The system ran for 120 hours in two configurations, the day configuration running a mix of 18 different 128-PE jobs and 16 different 256-PE jobs, and a night configuration running only the 256-PE jobs. Around 4000 jobs were executed during the 120 hours without incident. This test was intended to exercise the machine in a production job environment.
3. Smaller jobs reliability trial. The system ran for 8 hours with a mix of 4 different 32-PE jobs, 2 different 64-PE jobs and 18 different 128-PE jobs. Around 900 jobs were executed during the 8 hours without incident. This test was intended to exercise the machine in a dual development/smaller-scale production job environment.

The system was shipped to Edinburgh late in April, and installation began on the 27th April. Power was applied to the Y-MP on the 28th, and to the T3D itself 24 hours later. The software installation on the Y-MP began on the 30th, and diagnostics started running on the T3D itself on the 31st. On the 3rd May strain-testing of the T3D began with selected benchmark codes, and after the disc and netwoork configuration had been finalised the T3D began to run the full benchmark suite under test on the 7th May, entering the observed acceptance trials on the 10th May and completing the trials on 13th May.

The acceptance trials, witnessed by representatives from EPSRC and the University of Edinburgh, consisted of three phases:

1. Performance trial. This was essentially a repeat of the performance trial undertaken in Chippewa Falls.
2. Large jobs reliability trial. This was a essentially a repeat of the reliability trial undertaken in the factory, except that it was only required to run for 24 hours in which time 1127 jobs were executed.
3. Smaller jobs reliability trial. This was essentially a repeat of the smaller jobs reliability trial undertaken in the factory.

On conclusion of the acceptance trials, the system was handed over to the University in preparation for service.

The entire installation and commissioning process

was completed according to schedule and with the minimum of problems considering the extent of the works necessary to receive the Cray system, and the professional and timely approach to the installation by Cray Research personnel was much commented upon.

An initial service to selected users was started on 1st June, with full user service to the peer-reviewed groups commencing on 1st July.

Mass-storage provision

The University agreed to provide a mass-storage system for attachment to the Y-MP with an initial capacity of around 1 Tbyte. The system chosen was a new product from IBM, the 3494 Automatic Tape Library, which was based upon 3490E 800 Mbyte cartridges. The machine installed initially contained 1200 cartridge slots and three 3490E drives which were directly attached to the Y-MP by TCA-1 block-mux parallel channels. The control path to the 3494 was via a private LAN to a RS/6000 system.

The Edinburgh installation was the first attachment of a 3494 library to a Cray system on a customer site, and IBM and Cray developers worked closely on the integration of the software.

Initial Operation of the T3D

Operational requirements

Despite the original intention that there would only be a small number of very large user groups granted access to the machine, the reality was that some 20-30 groups were to be granted access to the system for the initial time allocation period that commenced in July 1994.

It was viewed as essential that the small numbers of very large groups that would be able to make effective use of the system from the very start of the service due to their experience on other large parallel platforms should be able to get access to large amounts of production time, whilst at the same time making useful portions of the machine available to users new to the technology as a development resource. The requirement to make extensive training available to the user community was viewed as essential, with significant proportions of the resource being directed towards on-line training needs.

Time allocation

Time was allocated, in processor-hours, by an EPSRC committee set up to peer-review the applications called TRAP (T3D Resources Allocation Panel). The initial time allocation period was for the four months starting 1st July, with a second period of five months to commence on 1st November, and with subsequent time allocation periods to be of six-monthly duration starting from 1st April 1995.

Time was allocated to groups (or consortia) of users who were under the control of a named principal investigator. Some of the consortia were large with some tens of members, others consisted of only a few individuals. The membership of most of the consortia was distributed geographically throughout the United Kingdom, and members ranged from senior professors to graduate students. Time was not allocated to individuals.

It was essential that the usage of resources by both individuals and user groups could be effectively monitored, and a strategy needed to be found to implement the time-allocation strategies directed by the allocation panel.

Regrettably, Cray had not provided any means of easily regulating access to the T3D resource by users on a processor-use basis, nor any means that users (or, more importantly, the principal investigators of each user consortium) could employ to interrogate their processor usage to date. In particular, it was necessary to be able to automatically deny access to the MPP for users who had exceeded their allocation of time.

Allocation using ACIDs

It was the responsibility of EPCC staff to implement a scheme to enforce the time allocation policies of TRAP, and it was decided to base this on UNICOS ACIDs.

Time was allocated (in processor hours) to ACIDs, and users were granted access to appropriate ACIDs by modifying their entries in the UDB. The actual time was converted to processor seconds, and the current allocation assigned to an ACID maintained in a special file within the system.

Users could have access to a number of ACIDs reflecting perhaps their membership of a number of user groups, or if the time given to a particular group was divided up into development and production time, and could use the 'newacct' command (or the #QSUB -A directive within a NQS job) to assign the ACID to be used for any subsequent MPP access during that session or job.

It was up to the principal investigator of each user group to decide how the time was to be assigned to users within that consortium, and while some consortia opted to leave all of their allocated time in a single ACID that was accessible to all members of that group, others chose to divide it up on an an individual basis, or to reflect the distributed nature of some of the consortia on a geographical basis. Others still divided their time into two, with the bulk of the time going to a production ACID accessible only to trusted "production" members of the group, and the residue going to a development ACID with access given perhaps to research students.

The principal investigator would be granted access to all of the ACIDs associated with the group, so that an effective check could be made on utilisation by individuals or sub-groups. With each ACID is associated an 'owner' (normally the principal investigator), who receives, on a monthly basis, an automatically generated report on the usage made by each user with access to that ACID.

The 'mppacct' command is used by users to view their own allocation, and by the systems administrators to manually adjust allocations as requested.

Example:

A consortium named "v3" has 150000-PE hours allocated to it. The principal investigator (username piv3) has requested that 35000 hours be allocated to his Edinburgh, Oxford and Manchester sub-groups, 20000 hours for development and the remainder held in reserve.

User piv3 views the initial state:

```
piv3(sn1904): mppacct
        v3:          25000:00:00
        v3-ed:       35000:00:00
        v3-oxf:      35000:00:00
        v3-man:      35000:00:00
        v3-dev:      20000:00:00
```

After some weeks of use, the following is the state:

```
piv3(sn1904): mppacct
        v3:          25000:00:00
        v3-ed:        2350:31:02
        v3-oxf:      10219:53:36
        v3-man:      14118:57:44
        v3-dev:      20000:00:00
```

The principal investigator requests EPCC to remove 20000 hours from the reserve, and give it to the v3-ed sub-group:

```
# mppacct -20000 v3
        v3:           5000:00:00

# mppacct +20000 v3-ed
        v3-ed:       22350:31:02
```

Control of access

A special local 'mppexec' command was implemented which made various checks before calling the real 'mppexec' which was not accessible to users.

These checks are:

1. Ensure that the current ACID has time allocated to it.
2. Ensure that there is sufficient time left within the ACID to enable this program to terminate (batch programs only where the #QSUB

-l mpp_t directive is enforced). This was implemented only after a very new user set only the Y-MP time limit within a 256-PE batch job, which then ran for more than 40 hours over a weekend totally destroying the entire 10000-PE hours initial time allocation for a complete multi-user consortium.

Users who have been set up to access the Y-MP, but have not yet been allocated to any user group are given a default ACID without any MPP time allocated, thus ensuring that only users who have been positively accredited to use the MPP may do so.

At the end of each T3D accounting period (currently twice per day) a local accounting program analyses the NQS logs, the /usr/spool/mpp/mppsyslog file and a local 'mppexec' logfile, and each ACID has the amount of time allocated to it decremented by the usage during that accounting period.

If a user attempts to run a program when there is no time left within the current ACID, the request is denied:

```
expc01(sn1904): mppacct
        k1-test:        0:00:00

expc01(sn1904): mppexec testjob
Permission denied
 - quota exceeded on account k1-test
```

Modus operandi

With the dual requirements to provide a production service and a development one upon the same hardware, it was decided to configure the system with a DAY and NIGHT configuration.

DAY and NIGHT NQS pipe queues were set up, with 64, 128 and 256 PE batch queues being the destinations. A separate COMPILE queue was set up for non-MPP-execution work.

The MPP system was set to deadstart twice per day, thus any applications running at the changeover time were lost. This was done to ensure that the changeover was completely clean, and that the internal usage accounting was consistent by being done on complete and closed accounting and log files. The changeover normally only took under one minute and was done under the control of 'cron'.

There was no intention to operate a vector processing service upon the Y-MP in addition to the MPP service upon the MPP. Users were encouraged to use the Y-MP only for work directly related to their MPP work, although some degree of post-processing was run by members of one consortium overnight when the interactive load on the Y-MP diminished.

DAY configuration

The DAY configuration, originally ran between 0900 and 1700 on Monday - Friday, but later revised to 1000 - 1800, was to provide limited production access but significant development access, while the NIGHT configuration (all other times including the complete weekend) was to provide solely production access for the largest codes.

Initially, the system was divided up into three pools during the DAY configuration, one containing 128-PEs that was dedicated to batch access, one containing 64-PEs for shared batch/interactive access, and one containing 64-PEs that was for interactive access only.

Batch jobs were limited to a total duration of one hour, with interactive access limited to 30 minutes. 64-PEs was the largest interactive job that could be run due to the nature of the pool structure.

This scheme was soon found to be non-optimal, with the 128-PE pool often partially (or wholly) idle due to an intermittent batch load so early in the life of the system, and the shared batch/interactive pool could become badly fragmented with small interactive jobs.

After about a month it was decided to use only a single 256-PE pool during the day, and regulate access by having a 192-PE global NQS limit. This enabled better utilisation to be achieved, with the interactive portion of the machine being able to

expand dynamically into that previously reserved for batch use if necessary. The 64-PE limit for interactive use was retained by setting it in the UDB.

Problems still ensued, the original "first-fit" PE placement algorithm often causing subsequent jobs to block due to there not being the correct shape left within the torus for the application to be placed, although such problems were largely overcome when the "best-fit" algorithm was available in UNICOS-MAX 1.1.0.4 on October 1994.

The one hour limit for daytime batch jobs was considered by some users to be too long, users were observed submitting streams of 3599 second jobs so that a single user could easily tie up the entire DAY batch programme with what could easily have been a single larger overnight job. To enhance the turnround for users requiring access to 128-PEs during the day, but only for short periods, a special short 128-PE queue was introduced with an upper limit of 20 minutes elapsed time. This queue had a higher priority than the standard 128-PE queue.

For use during the regular series of training courses, and return was made to a 3-pool configuration so that 64-PEs could be dedicated to users under training by the group access facility within the MPP configuration file. The rest of the machine was configured as a 128-PE batch pool and a 64-PE interactive pool during these periods.

If there were not enough resources available for an interactive job, it would go into WAIT state, thereby denying access to any resources for any subsequent job until that request was satisfied. This was a frequent cause of problems when the system was busy, and the local 'mppexec' was changed so as to force the "nosleep" option on interactive jobs.

NIGHT configuration

The NIGHT configuration (which also ran all day at weekends) was in place essentially to provide an exclusive batch-based service for users using significant portions of the MPP for substantial periods. 256-PE jobs were given priority, with 128-PE jobs and 64-PE jobs being run generally only when there was no more 256-PE work, and users were not encouraged to use the NIGHT configuration for jobs of very short duration or for those requiring only modest amounts of processors.

The only significant subsequent change made to this configuration was because users trying to run production jobs with less that 256-PEs found that they were being virtually locked out. By the end of the first allocation period, the utilisation on the system was such that the amount of time in the week dedicated to the NIGHT configuration was barely able to satisfy the requirements of the 256-PE jobs, much less give access to the resources for long-running 128-PE (or smaller) jobs. The decision was made to change the priority in the NIGHT queue-complex on Monday nights, this allowing any backlog of 128 or 64-PE jobs that may have built up over the weekend to drain away, before starting any further 256-PE jobs. This move was considered a success, especially as not all serious production work was 256-PE based, and users requiring large amounts of legitimate 128-PE time were not getting their fair share of the resources.

Disc allocation strategy

It was accepted that many of the applications running on the T3D would be extremely demanding on disc space, thus a sensible disc allocation strategy was required so as to ensure that the processing resource was not stalled by lack of space to store the data generated.

Tertiary level storage, in the shape of the IBM 3494 automatic tape library system based on 3490E technology was provided by the University of Edinburgh, but although of an initial 1 Tbyte capacity it was viewed that this system would be full within 6-9 months of full operation on the T3D.

The effective allocation and management of the data-storage areas within the system have been amongst the most challenging tasks associated with the operation of the T3D due to what is now perceived to be modest amounts of disc space, inadequate amounts of on-line mass-storage capacity

and the ability of the production user groups to generate data at rates greatly in excess of the original expectations.

Individual users were allocated modest (100 Mbytes each) space allocations on the various 'home' filesystems, it being imagined that users would use such space for storage of code, and not for the results generated by their applications.

Three large filesystems were created for workspace, and these are considered in turn:

1. /work. This filesystem, of 65 Gbytes in size, was created to form the workspace for the groups of users with the largest compute requirements on the T3D. Typically, these groups had > 5% of the allocated cycles on the machine, and initially there were six such user groups. Each was allocated, under group quota, a proportion of the space on /work and this guaranteed space was intended to be the repository for the input data required for a large-scale run, and for the output data generated by that run. It was not intended that the filesystem would be used for the long-term storage of data, although data-restart sets might be held within the filesystem between application runs.

2. /scratch. This filesystem, of 27 Gbytes in size, was created to form the workspace for the groups of users with more modest compute requirements. Typically, these groups had between 1% and 5% of the allocated cycles on the machine. Space was not allocated under quota, but users had to apply to have a directory created under their name. As with the space allocated for the largest user groups, this space was intended to be the repository for the input data required for a program run, and for the output data generated by that run.

3. /arch. This filesystem, of 48 Gbytes in size, was created to form the interface to the IBM 3494 automatic tape library system. Access to the library was controlled under DMF, and only those users who had a proven need to access the mass-storage system were able to use the filesystem. DMF was found to be extremely effective and was popular with the users due to its simplicity, although a number of operational problems were encountered which required a usage policy to be determined.

DMF access to the IBM

There were two main problems associated with access to the IBM through DMF:

1. Some users were storing miriads of very small files, and then requesting that they be restored in their hundreds. The mount/search/read/rewind/search/read cycle was found to take inordinate amounts of time, and some users found that it was faster to recompute the data than to get it brought back from archive.

2. There was no easy means to apply any form of off-line quotas through DMF, and as space on the IBM was extremely limited, it was expected that the initial users groups in production would be able to swamp the archiver before other users could get into production. This is precisely what happened, and when the IBM became 100% full in December 1994, over 85% of the data belonged to one user group.

The solution to the first problem was found by stating a policy that users should not consider archiving small data files, and a minimum size of 100 Mbytes was suggested as the norm. Users were instructed not to actually work within /arch, but instead to use their allocated directories in /work or /scratch as appropriate, and if they had large numbers of small data sets to archive, then these be packed into single datasets and subsequently moved to /arch for archive.

The solution to the second problem has been harder to find, and despite an upgrade to the the IBM system in January (400 more cartridge slots being provided) and the identification of 200 Gbytes worth of data that could be physically removed for on-shelf storage, the problem persists. The current plan is to simulate off-line quotas by creating a range of MSP's, each with their own subset of the

tape-space, and enabling access for user groups to a single MSP by use of the 'archmed' setting within the UDB. It is expected that this policy will delay the off-line storage problem until the expected new tape technology that will greatly expand the total capacity of the IBM system.

Upgrade of the T3D

One of the reasons behind the choice of the University of Edinburgh as the site for the Cray T3D system, was the agreement that the T3D would be enhanced by a further 64-processors before the end of 1994 at no cost to EPSRC. This came about due to the commercial arrangement between EPCC and Cray, and the success of EPCC's industrial partnership scheme which enabled the additional resources to be purchased on the back of successful commercial contracts between EPCC and a number of large companies and government agencies.

The upgrade was successfully completed during December 1994 and within the planned schedule.

Installation, trialling and acceptance

The original design of the Cray T3D allowed for upgrades only by doubling the size of the machine. This rather inflexible arrangement, appealing as it may be to those responsible for ensuring Cray's income, would have made any upgrades to the Edinburgh installation extremely expensive.

Subsequently, the system design was so changed so as to allow for non-powers-of-2 systems, and this modification was made to all new systems when still on the production line. Regrettably, the decision to make this change was made too late for the Edinburgh machine to be altered when still being constructed, so the 64-PE upgrade in fact required complete on-site frame replacement with one containing a suitably modified wire-mat.

During mid-December 1994, the Edinburgh T3D was replaced with system number 6001, the original T3D, which had been re-built into a non-powers-of-2 system. The entire swap-out and swap-in process plus integration testing and an acceptance trial took less than the 80 hours estimated, and the machine was returned to production some hours in advance of the advertised time.

The only potentially serious problem encountered during the upgrade was the initial inability to run a 256-PE job and one of 64-PEs concurrently in the upgraded machine. The reason was that the 8 x 4 x 5 'shape' of the new machine (in nodes, not PE's) precluded a 256-PE job (8 x 4 x 4) and a 64-PE job (4 x 4 x 2) from being placed upon the torus, a change to the default shape of a 64-PE job (to 8 x 4 x 1) being necessary before the acceptance trial could be undertaken.

Subsequently a limitation within the barrier network has shown that it is not possible to place 5 different 64-PE jobs within the torus at the same time, although that is the only combination that has been found not to be possible on the upgraded system.

The acceptance trial consisted of a series of three 64-PE, three 128-PE and three 256-PE jobs of different lengths, which were run within a single 320-PE pool for 8 hours. During that period, 58 jobs executed without incident.

Revised hardware configuration

sn6001:
T3D/MCN320-8 (320 PE, 8 MWord/PE)
2 x IOG's

sn1904:
Y-MP4E/264
32 MWord SSD
2 x IOC's
36 x DA301 (197 GB) on 8 x DCA-3
8 x DD60 (15 GB) on 8 x DCA-2
2 x STK 4480 on 1 x TCA-1
1 x IBM 3494 ATL system on 3 x TCA-1
2 x FDDI on 2 x FCA-1

Subsequent operation of the T3D

Operational requirements

The operational requirements for the expanded machine remained as before, to be able to give an effective mix of production and development access upon the same platform.

It was considered that it would be easier to satisfy these two sometimes opposing requirements with the expanded system.

Modus operandi

The previous DAY and NIGHT configurations were carried over, with only the modifications made to reflect the enhanced system.

Users submit jobs to the two pipe queues DAY and NIGHT and the job is routed to an appropriate batch queue depending on the resources requested. The current NQS queue complex is appended.

DAY configuration

Initially a single 320-PE pool was set up, with shared batch and interactive access. No more than 256-PEs could be used in total by batch jobs, with a maximum of 128-PEs for any particular job. The maximum duration of a batch job was retained at one hour.

This configuration was later changed to one containing 2 pools, a 64-PE pool dedicated for interactive access, and a 256-PE pool for shared interactive and batch access.

The change was made because under certain circumstances the 320-PE pool could become badly fragmented. A typical scenario was as follows. As there were less than 256-PEs in use for batch work, an NQS job would start, but it was not possible to place the required shape (say, for a 128-PE job) upon the torus due to a an ill-placed small interactive job that was in the middle of the torus. The batch job would then go into WAIT state, which would prevent any other job (batch or interactive) from starting. This caused frequent wastage of resources, with multiple complaints from users who were unable to run even the smallest job despite there being PEs available (or seemingly available). Largely separating the batch and interactive work greatly reduced the likelyhood of this happening, and even if a batch job did go into WAIT state in the 256-PE pool it would not prevent interactive jobs starting in the 64-PE pool if there were resources available.

The most recent change to the DAY configuration has been the reduction in the maximum number of PEs available to a single interactive job from 64 to 32. It was observed that some users were stacking up 64-PE interactive jobs in sequence and thus not allowing any others users to access a major part of the interactive resources.

To compensate those users who required access to 64-PEs for development purposes, a short 64-PE queue (with a maximum time limit of 20 minutes) was introduced. This queue runs with a higher priority than the standard 64-PE queue and is thus analogous to the short high priority 128-PE queue already in use.

NIGHT configuration

The NIGHT configuration was divided into two pools from the start, the 256-PE pool being used exclusively for production 256-PE jobs, and the 64-PE pool used to satisfy the requirements of those users who required interactive access during the evenings, early morning and over the weekend.

The present configuration allows interactive use up until 2000 hours and from 0700 hours each day to the 64-PE pool, and exclusive batch access between 2000 and 0700.

The most recent change has been to apply total time limits to 256-PE jobs, when previously such jobs were unbounded. With the enormous increase in the workload observed during the last two months, the backlog of work has been growing at an alarming rate, and some users were finding that they could wait well over a week before their job might be scheduled to run. Some users were submitting

jobs with a time limit of 24-hours, but these jobs always consisted of a number of more modest jobs back-to-back within the same script. The present limits are 6 hours for an overnight job, and 12 hours for one that will be scheduled to run at the weekend. Jobs requesting more than 12 hours (and those requesting no time at all) are held in a special queue that is never scheduled to run, but which is visible to the systems staff who can advise the user of his oversight.

All the limits currently applied are flexible, and the configuration is constantly being adjusted to reflect both the increasing workload and any changes in the balance of the user profile between those requiring production those requiring development resources.

The very high levels of utilisation now achieved (in excess of 86% at the time of writing) would indicate that the current configuration matches well with the present workload.

Current problems

There are a number of problems, which are affecting the usability of the system:

1. The ability for NQS to start a job, that may have taken days to climb to the head of the queue, one minute before a configuration changeover causes severe and justified irritation. It is planned to investigate the implementation of a local user exit within NQS to prevent the selection of a such a job just before a planned configuration changeover.
2. The ability for NQS to start a job that requests a number of PEs, only for the placement to fail due to other jobs being ill-placed within the torus, thus causing the job to enter WAIT state. The functionality for NQS to be able to interrogate the MPP job start and placement system to see if a job can be placed would reduce resource wastage forced by this problem.
3. The inability to control total access to resources on a UID basis. Users easily get round the UDB limit on the number of processors they can use interactively, by starting multiple logins, and by running interactive work concurrently with their own batch jobs.
4. There is an outstanding problem when communication down one of the IOG's is lost thus causing an interruption to service.
5. There is an outstanding problem when a program will enter WAIT state because the system is claiming that the program has requested a particular base PE. In these cases, the program has not specified any base PE, and in fact the reported base PE has an illegal identifier.

Utilisation

Tables summarising the utilisation on the T3D since the opening of the full user service at the start of July 1994 are appended.

MTBSI (Mean Time Between Service Incidents) is a measure of the times between unplanned downtime, an 'incident' being defined as anything that causes the T3D to be out of service at a time that it was planned to be available for users. Thus for the sake of the figures, an MPP crash, a total power outage to the site or a 12 minute delay in handing back the system after a preventative maintenance session are treated equally. A table breaking down the source of the incidents between MPP hardware, MPP software and other causes is appended, and it is clear that while the MPP hardware is broadly very reliable, that there have been a number of problems encountered with the software. With other vendors in the past, Edinburgh has been able to stretch systems to their limits and loadtest hardware and software seemingly to destruction, and the experience with the Cray T3D was no different in this respect.

Availability is defined as the fraction of the total time within the period in question that the system was available to users, thus it excludes both planned and unplanned downtime.

Servicability is defined as the fraction of the total time within the period in question less the planned

downtime, that the system was available to users.

It will be seen from the figures that the utilisation of the MPP has varied quite considerably since July 1994. Peaks were reached in September/October and again in February/March and these were due to the current policy of having fixed allocation periods for all users. Many users chose not to use their assigned cycles early in the allocation periods for a number of reasons — (1) they thought that there was plenty of time left, (2) that they would prefer a longer time to develop their applications and (3) that they would wait until more efficient compilers were available from Cray. After the end of the first allocation period, a number of user groups then embarked on extensive data-analysis of their previous results before generating new data, and this too caused a drop in utilisation prior to the running-up to the end of the second allocation period.

The fixed periods caused problems both for the EPCC systems and support staff as well as imposing an artificial pattern upon the users with the effect that the machine becomes saturated at the end of each allocation period as users try to use up their unused time.

Both EPCC and the users are lobbying for some change in the allocation period strategy now that the T3D has stabilised into a production platform.

By the beginning of March, the backlog building up for 256-PE work was in excess of 200 hours (when only 128 hours per week are available for 256-PE work), and this backlog was continuing to grow faster than it could be satisfied. Although the imminence of the end of the second allocation period was likely to be the major reason for the increase in user load, the user population was still growing and more major consortia were expected to enter full production status within a short period of time.

The load on the dual CPU Y-MP front-end has been considerable, although it was intended originally that one CPU would be removed when the additional 64-PEs were installed as part of a cost-saving exercise. Even though no heterogeneous applications are being run and users are strongly discouraged against using the Y-MP for post-processing, the average load on the Y-MP is in excess of 100% of a single CPU and very often during the DAY configuration it reaches 100% of both CPUs. Measurement has shown that a mean of about 14% of the Y-MP cycles are used in direct support of running jobs upon the MPP, and the bulk of the cycles used (in excess of 60%) are related to compilation of MPP jobs. As the system has an increasing user population this is to be expected, but an early assertion that the front-end capacity could be reduced to a single CPU without detriment to the service because the established applications in production would no longer use the compilers has proved to be a fallacy. The major groups are constantly refining and tuning their applications and new releases of the compilers are eagerly sought as the users do all possible to exploit the maximum performance out of their code.

At the time of writing there are 406 user accounts on the machine (an increase of 101 since the start of 1995), and it is interesting to reflect that 400 accounts was expected to be the maximum number that would be reached only by the end of the anticipated lifetime of the system.

Outputs from 'mppmon' are appended, the first is from a typical DAY session, the second and third display the effect of jobs blocking others on a busy machine, the fourth is from a typical NIGHT session and the fifth from a Monday evening session when the queue priority is changed to allow production jobs smaller than 256-PEs to run.

Conclusions

The Cray T3D has proved to be a successful platform upon which to operate a stable national service for a large number of disparate user groups. The dual requirements to operate a production service and a development service upon the same platform at the same time has proved a challenge, but the popularity of the service and the amount of science that has been done has proved that this has been reasonably successful.

There are a number of rough edges within the software environment offered by Cray, in particular a poor interface between NQS and the MPP resource

manager, severe difficulties in enabling DMF to effectively manage very large amounts of data when total mass-storage capacity is limited, and a very real immaturity within the MPP operational environment that forced EPCC to write its own time allocation and accounting procedures, inflexible processor placement strategies and fairly regular loss of service due to PE errors or software problems (typically timeouts).

That being said, the turnround on critical SPR's has been good and the good interface between on-site staff and developers in the US has enabled such problems to be largely got-round in a timely manner.

The University of Edinburgh has been involved in the operation of multi-user services on large-scale platforms of novel architecture since 1982, and in no case has an installed system so rapidly reached a position of stability and reliability that is comparable to machines of conventional architecture. The decision made by Cray to attach the MPP to an existing stable platform like the Y-MP under UNICOS has greatly smoothed the transition for many users and there is no doubt that this has contributed to the early stability of the T3D.

Acknowledgements

This work was undertaken under the terms of the grant from the Engineering and Physical Sciences Research Council to the University of Edinburgh.

The success of the operation of the Edinburgh Cray T3D installation has been due to the support and commitment from very many people, but the author should in particular like to give much credit to colleagues within the University Computing Services and the EPCC Consultancy and Support Group for their support and encouragement, and to extend this to the on-site personnel from Cray Research.

The author is on full-time secondment from the Edinburgh University Computing Services to EPCC and gratefully acknowledges the support of the Director of Computing and Information Technology Services and the Vice-Principal for Academic and Information Services.

References

[1] EPCC Annual Reports, *Edinburgh Parallel Computing Centre, 1990, 1991, 1992, 1993, 1994*

[2] Realising Our Potential, A Strategy for Science, Engineering and Technology, *Her Majesty's Government, 1993*

[3] Research Requirements for High Performance Computing, *SERC Scientific Working Group (Chairman Professor C.R.A.Catlow) 1992*

Utilization Figures, July 1994 - March 1995

1994	Planned (hh:mm)	Downtime Unplanned (hh:mm)	Total (hh:mm)	Incidents No.	MTBSI (hours)	Avail ability	Service ability	Proc. use (hh:mm:ss)	Proc. utilis.	MPP users
Jul	58:05	14:06	72:11	9	83	90.30%	97.95%	93335:12:08	54.24%	59
Aug	07:41	29:36	37:17	11	68	94.98%	95.98%	83038:21:10	45.90%	80
Sep	06:17	06:26	12:43	6	126	98.30%	99.16%	153230:55:48	79.14%	83
Oct	08:50	00:52	09:42	5	137	98.60%	99.87%	127892:19:30	72.90%	90
Nov	19:23	37:03	56:26	9	80	92.17%	94.72%	110780:47:14	65.12%	89
Dec	76:36	19:55	96:31	5	139	87.81%	97.22%	110380:41:16	55.08%	92
Tot:	176:52	106:58	283:50	45	99	93.64%	97.50%	678657:17:06		169

Major planned outages in July and December were for the six-monthly plant maintenance period and the 64-processor upgrade respectively.

The processor utilisation percentages for December reflect the change in the size of the machine from 256 to 320 processors during the month.

1995	Planned (hh:mm)	Downtime Unplanned (hh:mm)	Total (hh:mm)	Incidents No.	MTBSI (hours)	Avail ability	Service ability	Proc. use (hh:mm:ss)	Proc. utilis.	MPP users
Jan	55:47	01:23	57:10	3	212	91.79%	99.78%	103511:46:30	50.64%	122
Feb	08:32	08:02	16:34	9	72	97.53%	98.79%	157546:05:28	75.12%	135
Mar	04:11	02:04	06:15	1	-	97.11%	99.02%	57795:42:38	86:11%	95
Tot:	68:30	11:29	79:59	12	125	95.33%	99.24%	318853:44:36		163

Major planned outage in January was for the six-monthly plant maintenance period.

Figures for March are for the 216 hours from 1000 hours on the 1st until 1000 hours on the 10th.

All times refer to the T3D service and not to the Y-MP.

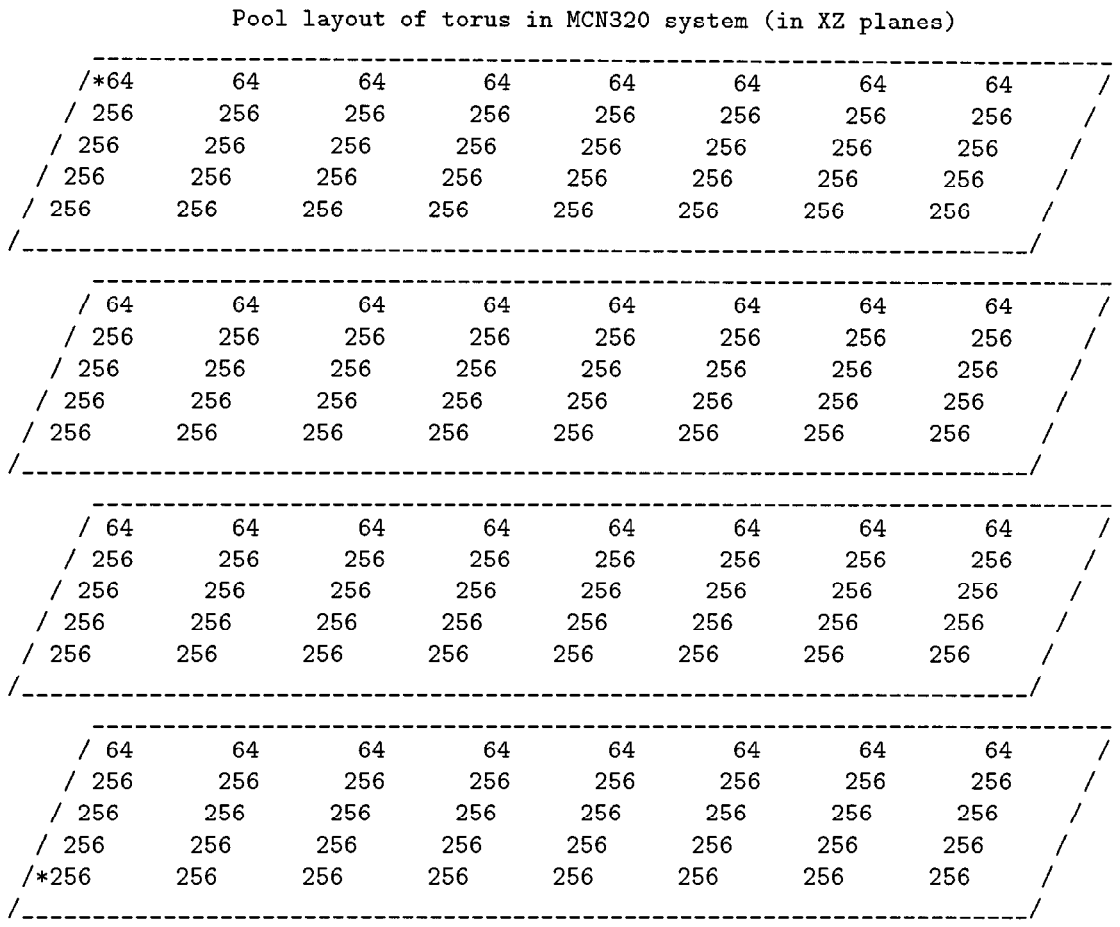
Breakdown of Incident Causes, July 1994 - March 1995

Month	MPP Hardware	MPP Software	Y-MP	Others
Jul	2	2	1	4
Aug	0	8	0	3
Sep	0	4	2	0
Oct	0	4	0	1
Nov	1	6	0	2
Dec	1	3	0	1
Jan	1	0	0	2
Feb	2	3	2	2
Mar	0	0	1	0
Tot:	7	30	6	15

NQS Queue Complexes

```
NQS 80.34 BATCH QUEUE SUMMARY
-----------------------------
QUEUE NAME              LIM TOT ENA STS QUE RUN  WAI HLD ARR EXI
----------------------- --- --- --- --- --- ---  --- --- --- ---
N_MPP_64                  2   8 yes off   8   0    0   0   0   0
MPP_256                   2  39 yes off  39   0    0   0   0   0
MPP_64                    3  13 yes  on  11   2    0   0   0   0
MPP_128                   1  18 yes  on  17   1    0   0   0   0
COMPILE                   4   1 yes  on   0   1    0   0   0   0
COMPILE_NIGHT             1   0 yes  on   0   0    0   0   0   0
N_MPP_128                 2   4 yes off   4   0    0   0   0   0
MPP_S128                  1   0 yes  on   0   0    0   0   0   0
W_MPP_256                 1   6 yes off   6   0    0   0   0   0
W_MPP_128                 2   1 yes off   1   0    0   0   0   0
W_MPP_64                  4   1 yes off   1   0    0   0   0   0
NORUN                     1   0 yes off   0   0    0   0   0   0
MPP_S64                   1   1 yes off   1   0    0   0   0   0
----------------------- --- --- --- --- --- ---  --- --- --- ---
darwin                   20  92           88   4    0   0   0   0
-------- ---------- --- --- --- --- --- ---  --- --- --- ---
-----------------------------
NQS 80.34 PIPE QUEUE SUMMARY
-----------------------------
QUEUE NAME              LIM TOT ENA STS QUE ROU  WAI HLD ARR DEP  DESTINATIONS
----------------------- --- --- --- --- --- ---  --- --- --- ---  -------------
day                       1   0 yes  on   0   0    0   0   0   0  COMPILE
                                                                   MPP_S64
                                                                   MPP_64
                                                                   MPP_S128
                                                                   MPP_128
night                     1   0 yes off   0   0    0   0   0   0  COMPILE
                                                                   COMPILE_NIGHT
                                                                   N_MPP_64
                                                                   N_MPP_128
                                                                   MPP_256
                                                                   W_MPP_64
                                                                   W_MPP_128
                                                                   W_MPP_256
                                                                   NORUN
----------------------- --- --- --- --- --- ---  --- --- --- ---  -------------
```

COMPILE	Non MPP work, typically for compilation
MPP_S64	Maximum of 64-PEs, maximum of 20 minutes MPP time
MPP_64	Maximum of 64-PEs, maximum of 1 hour MPP time
MPP_S128	Maximum of 128-PEs, maximum of 20 minutes MPP time
MPP_128	Maximum of 128-PEs, maximum of 1 hour MPP time
COMPILE_NIGHT	Non MPP work, typically for compilation
N_MPP_64	Maximum of 64-PEs, maximum of 6 hours MPP time
N_MPP_128	Maximum of 128-PEs, maximum of 6 hours MPP time
MPP_256	Maximum of 256-PEs, maximum of 6 hours MPP time
W_MPP_64	Maximum of 64-PEs, maximum of 12 hours MPP time (weekends only)
W_MPP_128	Maximum of 128-PEs, maximum of 12 hours MPP time (weekends only)
W_MPP_256	Maximum of 128-PEs, maximum of 12 hours MPP time (weekends only)
NORUN	MPP jobs requesting > 12 hours MPP time (Special scheduling applies)



Schematic representation of the 8 x 4 x 5 torus of logical nodes within the Edinburgh MCN320 system, showing thc division into 256 and 64-PE pools.

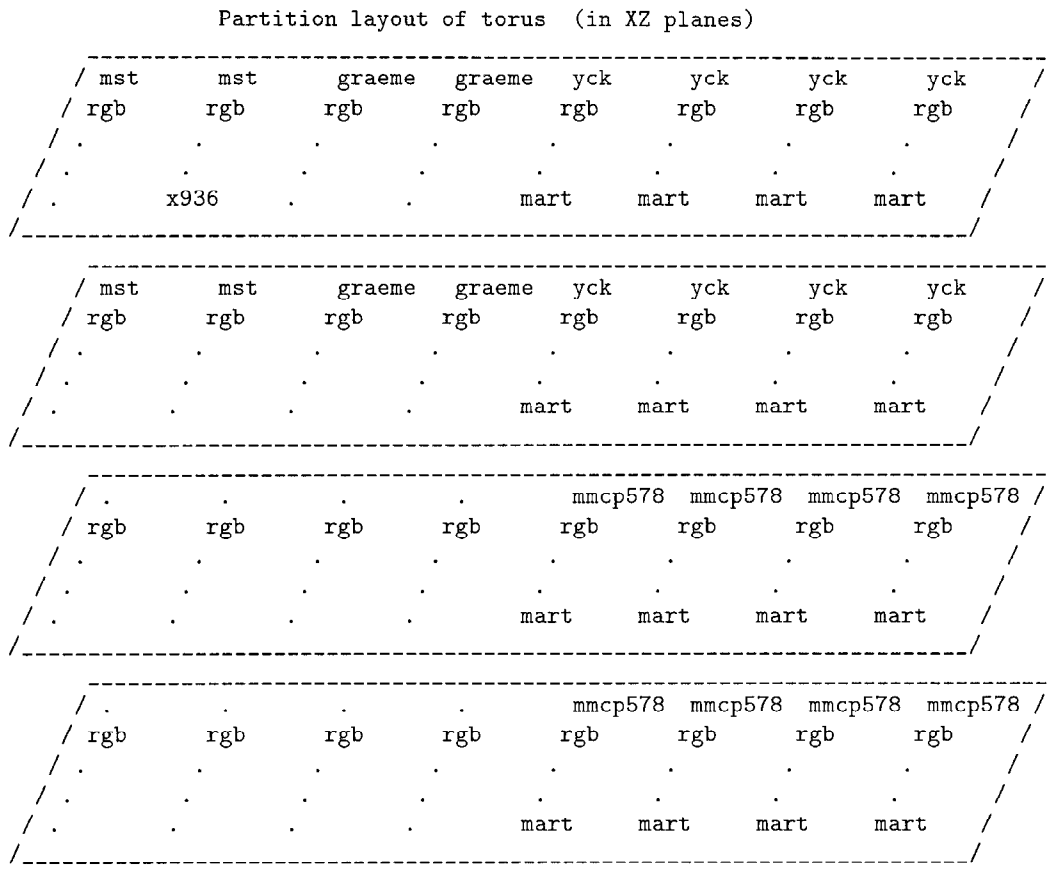
POOL_64 base node: 0x400 (indicated by *64)

POOL_256 base node: 0x000 (indicated by *256)

```
                    Interval: 10.0      03/07/95 14:46:09    Page: 0  Line: 0
Main
help   torus   pools   warn   clear   quit   refresh   shell



                Partition layout of torus  (in XZ planes)
     ------------------------------------------------------------------------
    / gnae03   gnae03   gnae03   gnae03   pzl      pzl      pzl      pzl     /
   / trf      trf      trf      trf      trf      trf      trf      trf     /
  / mmcp565  mmcp565  mmcp565  mmcp565  matthias matthias matthias matthias/
 / reync    reync    reync    reync    reync    reync    reync    reync   /
/ reync    reync    reync    reync    reync    reync    reync    reync   /
/_______________________________________________________________________/
     ------------------------------------------------------------------------
    / gnae03   gnae03   gnae03   gnae03   pzl      pzl      pzl      pzl     /
   / trf      trf      trf      trf      trf      trf      trf      trf     /
  / mmcp565  mmcp565  mmcp565  mmcp565  matthias matthias matthias matthias/
 / reync    reync    reync    reync    reync    reync    reync    reync   /
/ reync    reync    reync    reync    reync    reync    reync    reync   /
/_______________________________________________________________________/
     ------------------------------------------------------------------------
    / gnae03   gnae03   gnae03   gnae03   mmcp578  mmcp578  mmcp578  mmcp578 /
   / trf      trf      trf      trf      trf      trf      trf      trf     /
  / graeme   graeme   njgp     njgp     matthias matthias matthias matthias/
 / reync    reync    reync    reync    reync    reync    reync    reync   /
/ reync    reync    reync    reync    reync    reync    reync    reync   /
/_______________________________________________________________________/
     ------------------------------------------------------------------------
    / gnae03   gnae03   gnae03   gnae03   mmcp578  mmcp578  mmcp578  mmcp578 /
   / trf      trf      trf      trf      trf      trf      trf      trf     /
  / graeme   graeme   njgp     njgp     matthias matthias matthias matthias/
 / reync    reync    reync    reync    reync    reync    reync    reync   /
/ reync    reync    reync    reync    reync    reync    reync    reync   /
/_______________________________________________________________________/

User       PID  Program  State  Pool     Flags  Shape- XYZ   (base)  Elapsed
-------- ------ -------- ------ -------- ------ ------------------- ---------
reync      1266 Amber128 Active POOL_256 BR      128=16x 4x 2(x000) 00:25:52
pzl        4563 mag      Active POOL_64  I        16= 8x 2x 1(x428) 00:09:38
mmcp578    5598 gbmeson  Active POOL_64  I        16= 8x 2x 1(x408) 00:05:18
gnae03     5775 ns3d     Active POOL_64  I        32= 8x 4x 1(x400) 00:04:52
graeme     6690 clons.e  Active POOL_256 I         8= 4x 2x 1(x200) 00:01:02
mmcp565    6732 gbmeson  Active POOL 256 B        16= 8x 2x 1(x220) 00:00:57
matthias   6834 pboff    Active POOL_256 B        32= 8x 4x 1(x208) 00:00:55
njgp       6987 nbarot   Active POOL_256 I         8= 4x 2x 1(x204) 00:00:10
trf        6885 bluemoon Active POOL_256 B        64=16x 4x 1(x300) 00:00:01
```

'mppmon' output displaying status of the machine during the 'DAY' configuration, with a typical mix of 4 batch and 5 interactive programs running. Because the batch jobs are not requiring all 256-PEs in the 256-PE pool, two interactive jobs have been able to start within that pool.

```
                    Interval: 10.0      03/09/95 11:58:17    Page: 0  Line: 0
Main
help   torus   pools   warn   clear   quit   refresh   shell
```




```
User       PID  Program  State  Pool     Flags  Shape- XYZ   (base)  Elapsed
-------- ------ -------- ------ -------- ------ ------------------- ---------
rgb       91477 PROGRAM  Active POOL_256 B        64=16x 4x 1(x300) 00:56:41
mart      96409 gb32     Active POOL_256 I        32= 8x 4x 1(x008) 00:22:24
mst       96822 mp_lestu Active POOL_64  I         8= 4x 2x 1(x420) 00:19:45
yck       97088 comfort9 Active POOL_64  I         8= 4x 2x 1(x42c) 00:17:03
graeme    97606 clons.e  Active POOL_64  I         8= 4x 2x 1(x424) 00:13:35
yck       97988 comfort9 Active POOL_64  I         8= 4x 2x 1(x428) 00:09:54
mmcp578   98293 gbmeson  Active POOL_64  I        16= 8x 2x 1(x408) 00:07:55
peer      96467 smear.12 Wait   POOL_256 Ba~p    128=16x 4x 2       00:22:09
apr       98738 nwchem   Wait   POOL_256 Baq      64=16x 4x 1       00:01:18
```

Because a 32-PE job has been running within the 256-PE pool and the 64-PE pool had been full, a 32-PE interactive job was able to start within the 256-PE pool. A 128-PE program was able to start (there were less than the NQS global PE limit) but the program could not be placed, and so it entered WAIT state blocking any subsequent NQS jobs, but not blocking any interactive work in the 64-PE pool.

```
                    Interval: 10.0      03/09/95 11:59:00    Page: 0  Line: 0
Main
help   torus   pools   warn   clear   quit   refresh   shell
```

```
                Partition layout of torus  (in XZ planes)
     ------------------------------------------------------------------------
    / mst      mst      graeme   graeme   yck      yck      yck      yck     /
   / rgb      rgb      rgb      rgb      rgb      rgb      rgb      rgb     /
  / apr      apr      apr      apr      apr      apr      apr      apr     /
 / peer     peer     peer     peer     peer     peer     peer     peer    /
/ peer     peer     peer     peer     peer     peer     peer     peer    /
/_______________________________________________________________________/
     ------------------------------------------------------------------------
    / mst      mst      graeme   graeme   yck      yck      yck      yck     /
   / rgb      rgb      rgb      rgb      rgb      rgb      rgb      rgb     /
  / apr      apr      apr      apr      apr      apr      apr      apr     /
 / peer     peer     peer     peer     peer     peer     peer     peer    /
/ peer     peer     peer     peer     peer     peer     peer     peer    /
/_______________________________________________________________________/
     ------------------------------------------------------------------------
    / mmcp578  mmcp578  mmcp578  mmcp578  mmcp578  mmcp578  mmcp578  mmcp578 /
   / rgb      rgb      rgb      rgb      rgb      rgb      rgb      rgb     /
  / apr      apr      apr      apr      apr      apr      apr      apr     /
 / peer     peer     peer     peer     peer     peer     peer     peer    /
/ peer     peer     peer     peer     peer     peer     peer     peer    /
/_______________________________________________________________________/
     ------------------------------------------------------------------------
    / mmcp578  mmcp578  mmcp578  mmcp578  mmcp578  mmcp578  mmcp578  mmcp578 /
   / rgb      rgb      rgb      rgb      rgb      rgb      rgb      rgb     /
  / apr      apr      apr      apr      apr      apr      apr      apr     /
 / peer     peer     peer     peer     peer     peer     peer     peer    /
/ peer     peer     peer     peer     peer     peer     peer     peer    /
/_______________________________________________________________________/
```

```
User       PID  Program  State  Pool     Flags  Shape- XYZ   (base)  Elapsed
-------- ------ -------- ------ -------- ------ ------------------- ---------
rgb       91477 PROGRAM  Active POOL_256 B        64=16x 4x 1(x300) 00:57:23
mst       96822 mp_lestu Active POOL_64  I         8= 4x 2x 1(x420) 00:20:27
yck       97088 comfort9 Active POOL_64  I         8= 4x 2x 1(x42c) 00:17:45
graeme    97606 clons.e  Active POOL_64  I         8= 4x 2x 1(x424) 00:14:18
yck       97988 comfort9 Active POOL_64  I         8= 4x 2x 1(x428) 00:10:36
mmcp578   98293 gbmeson  Active POOL_64  I        16= 8x 2x 1(x408) 00:08:37
mmcp578   98976 gbmeson  Active POOL_64  I        16= 8x 2x 1(x400) 00:00:38
apr       98738 nwchem   Active POOL_256 B        64=16x 4x 1(x200) 00:00:28
peer      96467 smear.12 Active POOL_256 BR      128=16x 4x 2(x000) 00:00:28
```

Shortly afterwards, the interactive job within the 256-PE pool terminated, and both waiting batch jobs were able to start.

```
                     Interval: 10.0      03/09/95 20:18:31    Page: 0  Line: 0
Main
help   torus   pools   warn   clear   quit   refresh   shell



                Partition layout of torus  (in XZ planes)
      ------------------------------------------------------------------------
     / ws       ws       ws       ws       ws       ws       ws       ws      /
    / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
   / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
  / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
 / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
/______________________________________________________________________/
      ------------------------------------------------------------------------
     / ws       ws       ws       ws       ws       ws       ws       ws      /
    / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
   / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
  / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
 / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
/______________________________________________________________________/
      ------------------------------------------------------------------------
     / ws       ws       ws       ws       ws       ws       ws       ws      /
    / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
   / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
  / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
 / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
/______________________________________________________________________/
      ------------------------------------------------------------------------
     / ws       ws       ws       ws       ws       ws       ws       ws      /
    / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
   / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
  / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
 / blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03   blaw03  /
/______________________________________________________________________/

User       PID  Program  State  Pool     Flags  Shape- XYZ   (base)  Elapsed
-------- ------ -------- ------ -------- ------ ------------------- ---------
blaw03     1040 advpg.ou Active POOL_256 BR      256=16x 4x 4(x000) 00:20:05
ws         1237 PROGRAM  Active POOL_64  B        64=16x 4x 1(x400) 00:17:21
```

'mppmon' output displaying status of the machine during the 'NIGHT' configuration, with the 256-PE pool being used by a single 256-PE production program and the 64-PE pool being used by a single 64-PE production program. There would have been interactive access to the 64-PE pool until 2000 hours

```
                    Interval: 10.0      03/06/95 18:11:02    Page: 0  Line: 0
Main
help   torus   pools   warn   clear   quit   refresh   shell


                Partition layout of torus  (in XZ planes)
      ------------------------------------------------------------------------
     / mmcp578  mmcp578  mmcp578  mmcp578  expc01   expc01   expc01   expc01  /
    / matthias matthias matthias matthias matthias matthias matthias matthias/
   / ws       ws       ws       ws       ws       ws       ws       ws      /
  / matthias matthias matthias matthias matthias matthias matthias matthias/
 / jhol     jhol     jhol     jhol     jhol     jhol     jhol     jhol    /
/_____________________________________________________________________/
      ------------------------------------------------------------------------
     / mmcp578  mmcp578  mmcp578  mmcp578  expc01   expc01   expc01   expc01  /
    / matthias matthias matthias matthias matthias matthias matthias matthias/
   / ws       ws       ws       ws       ws       ws       ws       ws      /
  / matthias matthias matthias matthias matthias matthias matthias matthias/
 / jhol     jhol     jhol     jhol     jhol     jhol     jhol     jhol    /
/_____________________________________________________________________/
      ------------------------------------------------------------------------
     / mmcp565  mmcp565  mmcp565  mmcp565  expc01   expc01   expc01   expc01  /
    / matthias matthias matthias matthias matthias matthias matthias matthias/
   / ws       ws       ws       ws       ws       ws       ws       ws      /
  / matthias matthias matthias matthias matthias matthias matthias matthias/
 / jhol     jhol     jhol     jhol     jhol     jhol     jhol     jhol    /
/_____________________________________________________________________/
      ------------------------------------------------------------------------
     / mmcp565  mmcp565  mmcp565  mmcp565  expc01   expc01   expc01   expc01  /
    / matthias matthias matthias matthias matthias matthias matthias matthias/
   / ws       ws       ws       ws       ws       ws       ws       ws      /
  / matthias matthias matthias matthias matthias matthias matthias matthias/
 / jhol     jhol     jhol     jhol     jhol     jhol     jhol     jhol    /
/_____________________________________________________________________/

User       PID  Program  State  Pool     Flags  Shape- XYZ   (base)  Elapsed
-------- ------ -------- ------ -------- ------ ------------------- ---------
jhol       3022 cgmd_m_g Active POOL_256 BR       64=16x 4x 1(x000) 00:09:33
matthias   3130 pboff    Active POOL_256 B        32= 8x 4x 1(x100) 00:09:31
matthias   3118 pboff    Active POOL_256 B        32= 8x 4x 1(x108) 00:09:29
matthias   3624 pboff    Active POOL_256 B        64=16x 4x 1(x300) 00:07:30
mmcp578    3697 gbmeson  Active POOL_64  I        16= 8x 2x 1(x420) 00:07:03
ws         4341 PROGRAM  Active POOL_256 B        64=16x 4x 1(x200) 00:04:01
mmcp565    5237 gbmeson  Active POOL_64  I        16= 8x 2x 1(x400) 00:00:37
expc01     5760 sheep30  Active POOL_64  I        32= 8x 4x 1(x408) 00:00:07
```

'mppmon' output displaying typical status of the machine during the early part of a Monday evening. The 64-PE pool is being used by interactive users and the 256-PE pool is being used by smaller jobs as the job size priority has been reversed to drain out the backlog of 64-PE jobs.