The Joint Research Councils' Supercomputing Unit

B W Davies, R G Evans, T Daniels, D J Rigby

SERC Rutherford Appleton Laboratory

Abstract

A brief description is given of the Joint Research Councils Supercomputer Unit, a Cray X-MP/48 installation for the use of academic researchers in the UK. Operational experience and scientific use of the machine is discussed.

Introduction

The history of academic supercomputing in the UK dates probably from the provision of the original Ferranti 'Atlas' at the Atlas Computer Laboratory in 1964. 'Atlas' was a 1 MIP (one million instruction per second) machine with a main memory storage of 256 Kbytes. In 1979 the Science and Engineering Research Council (SERC) purchased facilities on the first of the current generation of 'supercomputers' the Cray-1, operated at Daresbury Laboratory. The Daresbury Cray-1 moved to the University of London computer Centre in 1983 (later to be joined by a second Cray-1) and Manchester acquired a Cyber 205 at about the same time. London, Manchester and the Atlas Centre are currently the three national supercomputer centres.

At about the time of the rapid increase in the US NSF supercomputer activities, the UK academic funding bodies (ie the Advisory Board for the Research Councils, the University Grants Committee and the Computer Board) set up a working party under Prof John Forty of the University of Warwick (now Vice Chancellor of Stirling University) to investigate the UK requirement. The working party set out to look at the need for supercomputing facilities over the whole of academic research, in the social sciences and humanities, as well as the traditional areas of science and engineering.

The Forty working party looked at a variety of disciplines and concluded that there was a strong scientific case for the central provision of a supercomputer of the most powerful type available. A technical sub- group looked at the technical options including a variety of manufacturers and architectures and concluded that a tried and proven system such as the Cray X-MP/48 was the best option. The potential of highly parallel systems to provide high power at low cost was recognised and the Forty report recommended support to develop more advanced hardware and software on machines of this type, particularly those of UK manufacture.

Since a supercomputer would necessarily be 'remote' for most of its users it was regarded as essential to upgrade the UK academic network 'JANET' and the panel endorsed a £5M programme proposed by the Computer Board.

As a result of the Forty report the ABRC made available the necessary capital for a Cray X-MP/48 to be installed at the Rutherford Appleton Laboratory where it is operated by SERC on behalf of all the Research Councils and the British Academy.

On a futuristic note, the Forty report recommended that in three years the Cray X-MP should be augmented by the new state of the art machine, eg a Cray-3 or Cray Y-MP, and thereafter a new machine should be acquired every three years.

The Cray X-MP/48 was delivered to the Atlas Centre on December 3 1986. Installation and commissioning by Cray engineers were completed on January 12 1987. Acceptance tests were completed on February 1 and a trial service to users started the next day. Peer Review access and control was introduced on April 1, and the computer was formally opened by the Secretary of State for Education and Science, the Right Honourable Kenneth Baker MP, on April 15. From installation stage to the end of this first period the commissioning and operation of the facility has been very successful. As this paper demonstrates, the large number of users attached and the progress to date fully justify the decision to procure a Cray X-MP/48.

Hardware Description

The major components of the hardware installation at the Atlas Centre are shown in Figure 1. As its name implies the Cray X-MP/48 has four central processors and 8 Mwords of real memory. Each processor has a 8.5 nsec cycle time and with both vector floating point units active has a peak performance of 235 MFlops. If it is possible to assign all four cpu's to an individual job then the X-MP/48 is almost a 1 GFlop machine. In addition to the 8 Mwords of memory the machine is provided with 32 Mwords of 'solid state disk' (SSD) with a bandwidth of 1000 Mbytes/sec. The system software enables the user to treat the solid state disk exactly as a conventional disk drive.

Twelve DD49 disk drives, each with a capacity of 1.2 Gbytes and a data transfer rate of 40 Mbytes/sec, provide a total of 14 Gbytes of 'conventional' disk storage for the X-MP/48. Past experience has shown that disk storage capacity is one of the major bottlenecks limiting performance of supercomputers, particularly if data is staged out to the front-end in significant quantities. The amount available on the X-MP was judged adequate to support currently executing jobs, but was clearly inadequate to hold more than a small fraction of the permanent filestore. An alternative solution to staging data to the front end was found by exploiting an existing Masstor M860 storage system.

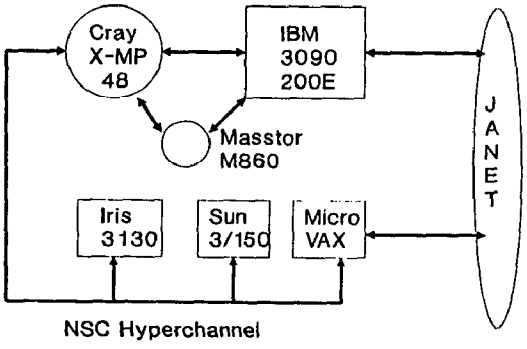

Figure 1

Our model of M860 is capable of storing 110 Gbytes on reusable cartridges containing 66 feet of 2.7 inch tape recorded using a helical scan rotary head. An automatic cartridge retrieval mechanism means the device is able to retrieve data from any part of its store in less than 20 seconds without operator intervention. Soon after installing the X-MP it was established that the M860 could be successfully attached to its IBM channel interface, where it appears as a bank of 3420 tape drives. The problem of driving the control part of the M860 (to instruct it which cartridge to mount) was quickly solved by trapping the mount messages issued by COS to the station software in the IBM front-end, and using the existing driver running under VM to control the M860. Thus control is exercised from the IBM front end, but data flows directly from the M860 to the Cray. (We understand Cray Research are considering a similar technique to support such devices in a general way under UNICOS). With these arrangements the Cray archive product was found to work to the M860 without modification. It provides a means of migrating data from disk to M860 automatically and from M860 to disk on demand. The resulting two level 'disk store' is completely invisible to users and provides the capacity to hold virtually the entire file store on line.

The system was introduced early in 1987 and has provided a reliable and valuable service since then. As a result data staging to the front-ends is completely negligible.

The success of this scheme is limited only by the capacity of our M860, which is shared with the IBM service. At present around 30 Gbytes are used by the Cray, although this is likely to grow. We are currently looking with great interest at automated cartridge stores which use standard IBM 3480 cartridges as an alternative means of providing higher capacity for minimum cost, although models of the M860 with Terabyte capacities are now available. An extremely large filestore is seen as an essential component of a supercomputing service for the 1990's.

System Software

The Cray X-MP/48 runs the standard Cray batch job operating system COS version 1.16. With the relatively small SSD size on this installation the SSD roll-out feature first introduced in COS1.16 has been essential in maintaining good high priority turnaround to jobs requesting large SSD allocations. The intention is at some future date to migrate to UNICOS since it provides the interactive facilities which are increasingly regarded as an essential feature of all advanced computing systems, however until 1989 UNICOS will lack some features such as archive support. It is possible to run 'guest' operating systems on the Cray X-MP but at present with rather inflexible allocations of cpu and memory. For this reason we do not anticipate running concurrent COS and UNICOS user services but we will run UNICOS in this mode as soon as possible for familiarisation.

An unusual feature of our installation is that the Cray job scheduling and accounting is largely controlled by the IBM VM front-end. This has the advantage that changes to the scheduling algorithm are easy to install since they do not involve any changes to COS. This was seen as being particularly valuable during the start-up period when changes to job scheduling were expected to be required frequently.

The scheme adopted relies on manipulating the COS job priority by commands from a locally written scheduler running under VM. The scheduler is notified of the entry and exit of jobs by messages sent via the station from locally written versions of the CHARGES and ACCOUNT programs which are invoked at the start and end of every job. Jobs in the input queue are held at priority zero which ensures they do not begin execution until required. The scheduler then simply sets the priority to suitable non-zero value whenever it wants to place a job into execution. The algorithm which determines the order of execution of work is then totally separate from COS. The dispatching of jobs once they enter execution is controlled by COS in a completely standard way.

The information about a job required by the scheduler is held under VM in an SQL database, and can easily be interrogated by users. Accounting data is held in a similar way, also under VM. These systems are very similar to the scheduling and accounting systems used to control our VM batch and MVS services, permitting much of the code to be common.

Front End Services

For most users the Cray is front-ended by an IBM 3090/200E running VM/CMS which also provides computing services to several hundred SERC and university users. In June 1988 IBM announced that the Atlas Centre was to be its UK centre for supercomputing, and we anticipate installing a 3090/600E with six vector facility processors later this year.

A microVAX running VMS is provided for general Cray use but is limited to about 50 users. Any university user in the UK can submit Cray jobs from his local machine provided that it supports the Job Transfer and Manipulation Protocol (JTMP) or File Transfer Protocol (FTP) defined on the JANET network. Currently about 90% of Cray jobs come from the IBM front end, about 5% via JTMP/FTP and 5% from the VAX.

Sun 3/160 and Silicon Graphics Iris 3130 workstations provide a UNIX front end service but only on the RAL site. The VAX and UNIX machines are connected to the Cray by a NSC hyperchannel.

The Sun 3/160 provides users with access to the FORGE software for analysing and optimising Cray FORTRAN source code. Currently we have a single user licence for FORGE on the Sun, but we are collaborating with Pacific Sierra on a multi user VM/CMS version which will make FORGE more asily useable by our remote users. The Iris 3130 is a high performance colour graphics workstation intended for interactive manipulation and display of data generated on the Cray. Currently the interaction is only at the level of file transfer using the Cray station software, but tighter coupling should be possible when UNICOS is introduced.

The Cray X-MP/48 is operated by staff of the Atlas Centre of the Rutherford Appleton Laboratory on behalf of all the Research Councils. There is a management committee, the Atlas Centre Supercomputer Committee (ACSC), chaired by Professor A J Forty. The interface between the Atlas Centre and its users is via a User Meeting convened by Professor Pert of York who is an ex-officio member of the ACSC. There is also a bi-monthly users' newsletter, ARCLIGHT.

Peer Review and the Computing Workload

Access to the computer, other than for use by industry or other paying customers, is via Peer Reviewed grant application to one of the Research Councils or to the British Academy. There are also arrangements, delegated to the Head of the Atlas Centre, for giving pump-priming access of up to five cpu hours of Cray time to potential grant applicants who wish to run small jobs in order to assess the amount of computing to request in a subsequent application.

Figure 2 shows the demand approved by Peer Review in cpu hours per month over the period April 1987 to March 1991, broken down according to scientific discipline. The capacity of the Cray X-MP/48 is about 2200 cpu hours per month, and so commitment from the beginning of 1988 is approaching the machine capacity. Since there are further announcements of awards at regular intervals it is likely that in a few months' time the machine will be over-committed and decisions will be needed on the levels of over allocation that should be made. The apparent decline in demand from 1989 onwards is simply due to a number of early grant awards being for two years only.

Performance

The Cray X-MP/48 hardware has proved very reliable in service, granted that the amount of preventive maintenance required (6 hours per week) is well above the norm for conventional computers. The Cray operating system software initially failed on average six times per week, but is now running with only one or two failures per week which is about the average for other Cray X-MP installations. Our statistics on machine availability are shown in Figure 3.

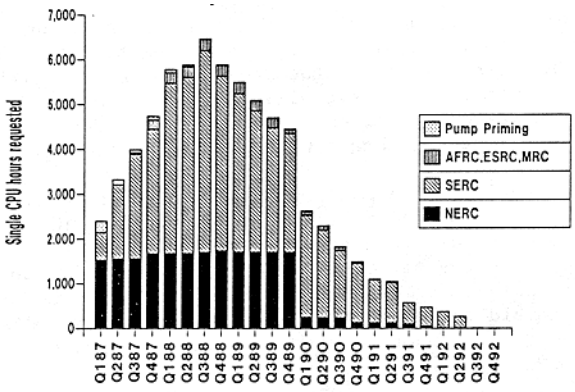

Figure 2

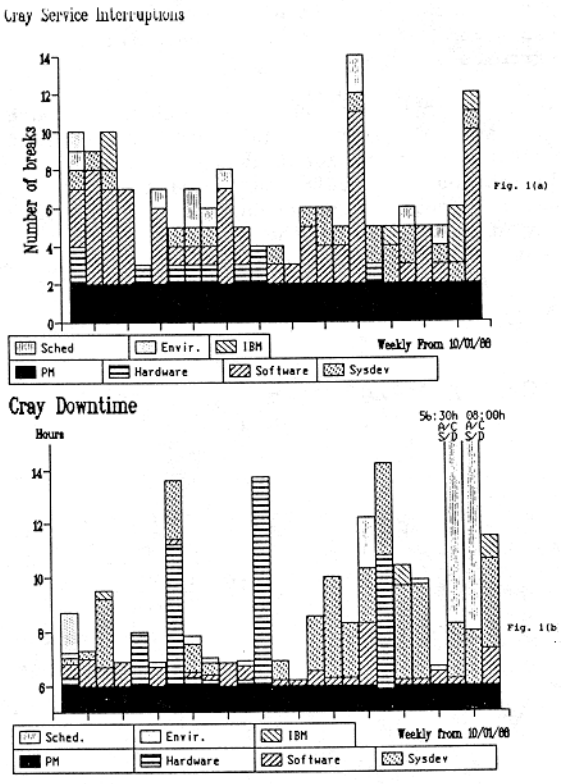

Figure 3

Since August 1987 it has been possible to monitor the performance that users are extracting from the Cray on individual jobs. Figure 4 shows a histogram of the average MFlops (millions of floating point operations) per processor second attained. The average weighted by job time is 39.5 Megaflops, a significant number of jobs are achieving over 100 Megaflops, and a few jobs reach 200 Mflops. Almost all these jobs run on a single processor of the Cray X-MP/48, for which the theoretical maximum performance is about 235 Mflops. The statistics on Mflops are broken down according to scientific discipline, but the only conclusion we would wish to make from this is that users (typically the environmental sciences and some chemists) with large production codes are well optimised.

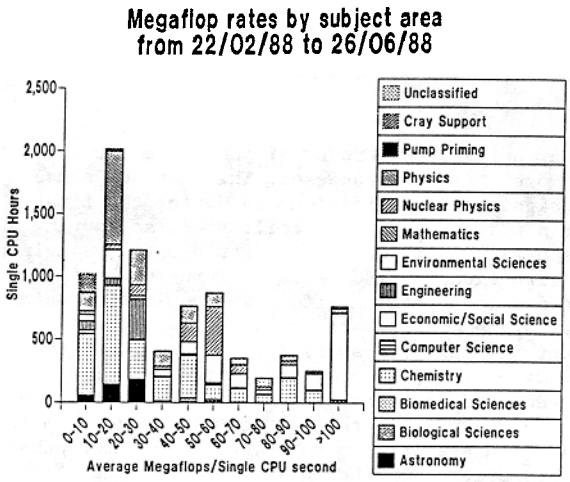

Figure 4

One of the features that distinguishes the Cray X-MP/48 from the supercomputers at ULCC and UMRCC is its larger memory. As is to be expected users are beginning to exploit this, and Figure 5 shows distribution of job memory sizes in the period August-November 1987 which shows substantial activity up to 4 Mwords. The lack of jobs above 4 Mwords causes us some concern that users with the largest jobs may be receiving poor turnaround.

To sustain efficient overall performance of the Cray X-MP/48 as job memory sizes grow, users will have to be encouraged to make their large memory jobs use more than one processor by using 'multi-tasking' and 'micro-tasking' techniques. Otherwise for example it would be possible for the 8 Mword memory to be occupied by a sequence of pairs of 3-4 Mword jobs each of which used only one processor. In such circumstances two of the machine's four processors would be idle and half the overall processing power would be unused.

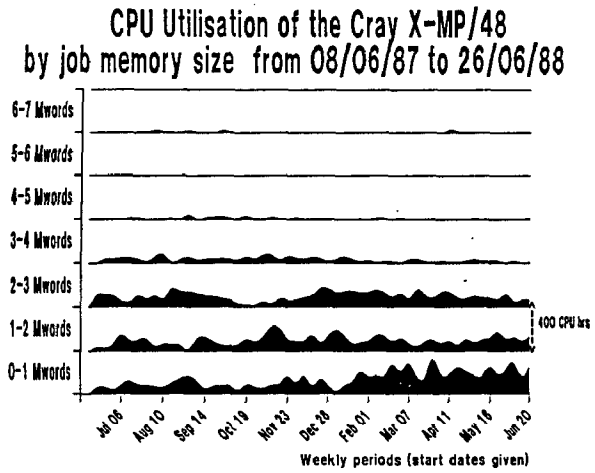

Figure 5

So far only a few percent of jobs have attempted to use more than one processor. This has not caused any significant inefficiencies during the first year's operation of the facility. Also, during this period individual users would not have gained any significant performance or turn-round time benefits from putting effort into multiprocessing. Staff at the centre are now working to increase the amount of multiprocessing activity on the machine. Assistance is being provided in the techniques to be used, and pressure will be brought to bear on users by changes in the job charging algorithm that will encourage large memory jobs to use more than one processor. On the positive side, some of the largest projects have already spontaneously adapted their programs for multiprocessing on the basis that this is a prerequisite for making real progress in the longer term. It is likely that this view will gradually become accepted by other large projects as they complete the first round of advances that the Cray has enabled and they look to their longer term plans.

Prior to release COS 1.16 of the Cray operating jobs using the Solid State Device for file storage were not 'rolled-out' by jobs of higher priority. This has meant that some users have had some difficulties with high priority jobs being blocked by low priority work. Following the introduction of COS 1.16 the scheduling of the SSD as a generic resource has greatly improved and users have responded very favourably.

Scientific Support

A small Advanced Research Computing Unit (ACRU) has been set up at the Atlas Centre to provide skilled computational science support for users of the Cray X-MP/48 and to foster the use of supercomputing in new areas including collaboration with industry. It is hoped to build the unit up to about 12 staff.

Contacts with Industry

Approaches have been made to potential commercial users in the industrial and finance sectors and there are indications that, as the number of packages available on the machine becomes greater, so commercial interest will grow. Generally speaking commercial use requires value added services rather than raw computing power.

Two meetings have been organised in conjunction with the Society of British Aerospace Companies to examine areas of cooperation with the academic sector in the aircraft and aero engine fields and this has led to continuing collaboration with Cranfield and a research position funded by Rolls Royce in conjunction with Pembroke College, Oxford.

The Scientific Programme

Since the JRCSU provides a service to all the UK universities and Research Councils its scientific programme is naturally very diverse and cannot be described comprehensively. The distribution of machine use by scientific discipline, shown in Figure 6 shows an equitable spread over many subject areas. Some of the larger, or more interesting programmes are:

* A large body of theoretical chemists use both ab initio and molecular mechanics methods for instance in rotational/vibrational band structure calculations for identification of transient species, potential energy surfaces for explosives, study of catalysts and the interactions of small molecules and proteins for pharmacological purposes.

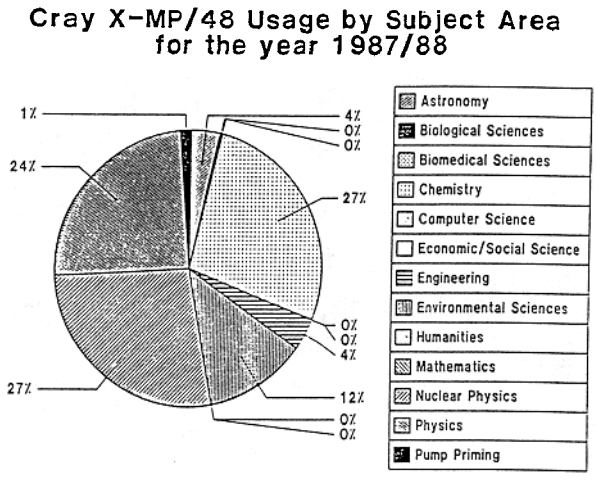

Figure 6

* Protein crystallographers are integrating the X-ray observations with molecular dynamics modelling to give a mutual constraints analysis and structure refinement time down from six months to a few hours of X-MP time.

* Atomic physics calculations now find a variety of applications in other areas of science, for example in astrophysical (stellar) opacities, in thermonuclear fusion plasmas (JET) and in the emerging science of X-ray lasers.

* Band structure calculations of the new high temperature super- conductors are being carried out for a variety of stoichiometries while fluids such as liquid neon and liquid helium where quantum effects dominate are also being studied.

* Astrophysical applications include atomic physics calculations, precision modelling of earth satellite orbits, gravitational many body simulations and pulsar searches involving 32 million point Fourier transforms.

* Environmental modelling is a large user of machine time with three large projects studying the North Sea, the Antarctic Ocean and Global Atmospheric Modelling

* Crack propagation in metals is simulated by high resolution finite element methods with elastic/plastic flow models. This is being applied to such critical structures as PWR pressure vessels.

* From the social sciences area, the Cray X-MP is being used to study the regional statistics on cancer cases, and establishing a bias-free significance for apparent geographical 'clusters'.

Conclusions

1. There is a high level of demand from within the UK universities for a machine of this performance.

2. Users are limited in the speed with which they can take up awards of machine time by outside factors such as recruitment of post-doctoral research workers.

3. The use of the M860 as a filestore is easy, works well, is well liked by users and reduces the load on the front end by a large factor. It also allows the Cray disks to be used mainly for temporary files.

4. Users with large production codes will put a substantial effort into code performance, while others must trade off machine time versus human time.