Integrating a CRAY T3E into KFA's Supercomputer Complex

Jutta Docter, Zentralinstitut fuer Angewandte Mathematik Forschungszentrum Juelich (KFA), Juelich, Germany

ABSTRACT: *With the addition of a 512 node CRAY T3E to its vector processors CRAY T916/12-512 and CRAY M94 KFA will offer a unique supercomputer complex for scientific computing. The presentation will describe KFA's plans to implement a single system image for the CRAY supercomputers and the early experiences with the brand new operating system UNICOS/mk.*

Introduction

The Central Institute of Applied Mathematics (ZAM) at the Research Centre Juelich (KFA) is responsible for planning, installation and management of the central computer systems and networks for KFA and operates the supercomputers for the national Supercomputer Centre (HLRZ) which provides computer resources to researchers throughout Germany.

Cray Research parallel vector processors have been used at KFA since 1983; the installation of an Intel Paragon with 140 compute nodes in 1992 was the first step towards massively parallel systems.

These machines are available to a user community consisting of researchers in many different scientific fields within KFA, universities and German industry.

The Supercomputer Complex

The recent configuration comprises a CRAY M94/4-512 for fileserving and interactive work, which will be replaced by a CRAY J90 in the near future and a CRAY T916/12-512 for vector-intensive batch processing which has been installed in April 1996. A CRAY T3E LC512-128 (512 liquid cooled nodes with 128 MB of memory each) will soon complete KFA's supercomputer complex. A preliminary 128 node CRAY T3E system was installed in August 1996.

Single System Image

KFA wants to present the CRAYs as a single system image to its users. This includes access to the user's data from every system, only one password in the complex, job submission to all CRAYs from the user's workstation, identical look and feel, consistent job priority schemes, and a unique access to the documentation.

To simplify system administration additional goals are a unified central user administration, automated operating, and central accounting.

The filesystems for $HOME (incrementally backed up to tapes) and $ARCHIVE (a DMF managed filesystem with a capacity of 5 TeraBytes in an automated cartridge library) reside on the fileserver CRAY M94 and are NFS mounted on the CRAY T90 and CRAY T3E. Users have access to their data on all three CRAY supercomputers; user names, numerical uids and NIS managed passwords are identical. Local disks are available on each system under the symbolic names of $TEMP and $WORK. Files in $WORK are not backed up and they are removed if they have not been accessed for four weeks. This allows the user to keep data from one job execution to the next. All files created within a batch job in $TEMP are deleted after the job has finished. Users have to move all generated data that should be retained permanently to the fileserver.

Job submission and status inquiry is locally available on each system. To allow job submission and control over jobs from any workstation in the Internet a WEB interface has been developed.

A central profile, identical path structure, KFA specific job status displays, and remote inquiry commands for disk usage on the fileserver allow the users to view the UNICOS systems as one entity.

Batch jobs are scheduled in a priority scheme where users can decide to "pay" double for their jobs if they want them to be returned fast or to get a slower turnaround for only half the price on the PVP systems.

All Cray documentation is accessible on a documentation server which will be migrated to the new Cray DynaWeb documentation software. The online documentation does not at all eliminate the necessity of having a printable version of the manuals (also in non-US formats, like A4).

User administration for about 1000 users is done centrally using a KFA developed administration system. Accounting data is fed into a central database which manages the data of the central computers and provides the CRAYs with updated CPU quotas once a day. Jobs from users or user groups which have exceeded their quota are scheduled only if the CRAYs would run idle otherwise.

To provide maximum system availability all CRAY supercomputers at KFA are monitored by a workstation which automatically restarts a failing system or alerts system administrators or hardware technicians.

Preliminary Installation of the CRAY T3E

KFA received an early shipment of the CRAY T3E to be able to use this new massively parallel system as soon as possible. Therefore some of the components which will be part of the final configuration are not available yet.

The preliminary hardware consists of 136 processor elements (PEs), which serve as application (APP) PEs, operating (OS) PEs, and command (CMD) PEs. The initial distribution was 128 APP PEs, 6 CMD PEs and 2 OS PEs to allow running parallel jobs on 128 PEs. This CRAY T3E has only one Multi Purpose Node (MPN) and one GigaRing to serve five SCSI controllers with 32 SCSI disks (approx. 128 GB). A FDDI interface provides the communication with the fileserver and the outside world. The operating system is a pre-release of UNICOS/mk, the new microkernel based, UNICOS-like operating system for the CRAY T3E parallel systems. The first official release UNICOS/mk R1.3 is expected in October 1996.

Limitations of the new operating system, initial stability problems, and the limited I/O capabilities with only one MPN led to limit the number of users. Only applications that had been tested on a CRAY T3D in Berlin and other CRAY T3E development machines in Eagan, MN were allowed during the first weeks to increase the availability of the machine for production jobs.

First Experiences

The NFS software was included in the UNICOS/mk pre-release, so that the data from the fileserver and other workstations could be mounted and accessed right from the beginning.

Users access the CRAY T3E interactively to compile their applications and manage their files on the CMD PEs and execute parallel applications on the APP PEs. The number of APP PEs for one run is either fixed by compiling it into the executable or *malleable* and has to be specified on the *mpprun* command.

NQS on the CRAY T3E implements a subset which allows to schedule jobs on the APP PEs, but the interaction between NQS and the system is very limited. There is no control whether the number of PEs actually used by an application matches the number of PEs requested for the NQS job.

Monitoring tools like *xmppview* and *xsam* are restricted for operator usage at the moment. It is intended to disable them for general usage because they put too much load on the system. These tools need functioning interval specifications to reduce the load and the possibility to be run on workstations.

Initially the users' passwords were not NIS managed because NIS was not part of the pre-release. The installation of the later delivered NIS commands led to the first experience with upgrading a binary-only system. The NIS files and all locally modified files were added to copies of the original */root* and */usr* filesystems. Then the system was booted from this different set of filesystems to be able to switch back if problems were encountered running NIS.

Problems

The biggest problem during the first two weeks was an error in file handling causing intermittent data corruption on the local disks. Moving the *disk server* and the *packet server* to run on a third OS PE helped to work around a serious memory management error in the operating system—A hardware memory mover problem, which caused the system to crash, required to configure additional four CMD PEs as a workaround.

These actions leave only 123 APP PEs for running parallel applications. To be able to run 128 APP PE applications on a so called "128 PE CRAY T3E" the number of additional PEs necessary for the operating system and command execution for a stable and functional environment should be known and included within the system.

Other problems cause several "hanging" or "blocking" situations:

- The disk error recovery in UNICOS/mk automatically sets disks *down* if an unrecovered error occurs. If the disk contains a system filesystem this results in problems with executing commands or accessing system files and the system to hang.
- If one process of a parallel application receives an *abort* signal the other processes are not automatically removed and might block the PEs until the situation is detected by the user or a system administrator.
- Aborting user programs might leave *zombie* processes which can only be removed by rebooting the system.
- The timelimits from the user database were not checked correctly and had to be globally deactivated.
- NQS jobs are not aborted if the *mpp* timelimit was reached and may run unlimited.

Furthermore some of the problems appear under rare circumstances or are hard to reproduce. The usage of the *stream buffers* in special applications cause the system to panic. Operand range errors in system commands are hard to analyse if they occur only occasionally.

The number of SPRs opened for the CRAY T3E by KFA amounted to about 50 in the first month. Often problems are not reproducible on other CRAY T3Es because they are strictly related to a certain disk configuration or the number of installed PEs.

The number of interrupts, mainly caused by the software, is about one per day.

System Startup and Dumps

Integrating the CRAY T3E into KFA's automatic operating environment is not yet feasible because automatic rebooting into multi-user mode is not available at the moment.

Booting the CRAY T3E needs an experienced system administrator, because the interaction between the System WorkStation (SWS) for operating and maintenance of the CRAY T3E, the MPN, which has its own software and associated problems, and the GigaRing is rather complex.

To analyse the system failures the memory of the MPN, the OS PEs and CMD PEs, and eventually of APP PEs, which might have reported a problem, must be dumped. Writing the dumps and bringing up UNICOS/mk might take between 0.5 and up to 1.5 hours if problems occur.

Software Distribution

The *binary-only* software distribution, which is new to KFA, reduces the effectiveness of system administration and analysis. Access to source code is needed to better identify the problem area for reporting. The possibility of fixing a serious problem immediately with a local modification is completely out of reach.

In a new version of a UNICOS/mk archive, it is absolutely not obvious, which errors have been corrected or which modifications have been made. In addition Cray Research has to develop a method to distribute different versions of archives, containing special feature, e.g. including NIS or not, or archives at different "patch" levels.

MPP Usage Model at KFA

On PVP systems the memory is a resource which has to be managed and split among the applications; on MPP systems the APP PEs have to be split among the different user applications.

The allocation scheme on the CRAY T3E requires consecutive logical PEs for every application. The absence of process migration may lead to extensive waste of CPU time when the APP PEs area is fragmented and free PEs can't be allocated because of the fragmentation.

Interactive access to the CRAY T3E is necessary to compile and test new applications, but experience on other MPP systems has shown that best utilization of the PEs can be accomplished by running batch jobs. To combine both usage models the APP PEs are split (50:50) among interactive and batch usage during daytime. At night and on weekends all APP PEs will be accessible for batch use only.

Preliminary Scheduling Strategies

Interactive applications are killed in the evening to free the APP PEs for large NQS production jobs. PE utilization is optimized by running the jobs which require the most APP PEs prior to jobs requesting fewer PEs, because execution of batch jobs of different sizes may fragment the application area.

To reduce the loss of CPU time, when a job is aborted by a system failure, the maximum run time of a job will be limited to four hours at KFA until checkpointing is available. This forces the user to write "checkpointing" information to disk if the application needs more time. In addition the rather short timelimit allows to share the CPU time on the APP PEs among different users during the night.

At the moment the scheduling of the jobs is "first in first out", except that large jobs are scheduled at the beginning of the batch phase. The KFA job priority scheme and methods to prevent one user from monopolizing the queues by submitting many jobs at a time will be established later. With the increasing number of installed PEs and more users running applications of different sizes, more sophisticated scheduling strategies will have to be developed. Monitoring on which APP PEs applications have executed will be crucial to understand the behavior of the scheduling algorithm.

Future Plans and Requirements

The stepwise installation of the rest of the 512 APP PEs (+32 OS, CMD, and redundant PEs) of the CRAY T3E is planned for late 1996. The configuration will be completed when three MPNs are functional on the GigaRing, HiPPI and ATM connections are established and the I/O can be done to fibre channel disk arrays in 1997.

The main goal is to have an operating system, hopefully with UNICOS/mk R1.3, which provides a safe and stable environment. I/O performance should allow applications to read and write as much data as necessary as fast as possible.

The optimal utilization of the system will evidently rely on process migration and better scheduling algorithms. The new *political scheduler* developed by Cray Research is expected to help in this task.

Another desirable feature is checkpoint/restart although its usability will depend on the time it takes to write all APP PE's memory out to disks.

For testing new versions of the operating system the announced feature of partitioning the system in multiple regions of PEs, running different operating systems, will be very helpful.

Conclusion

The CRAY T3E hardware works without major problems. Few carefully selected users get very good results and see performance improvements over the CRAY T3D of a factor 3 - 4. The UNICOS/mk pre-release has been made usable by installing some workarounds, and a number of open problems are expected to be fixed in the official release.

It's too early give a final comment on the usability of the CRAY T3E because of the preliminary hardware and software situation, but hopefully Cray Research will provide the sites with a stable environment, even for large systems, as soon as possible.