Table of Contents Author Index Home

Get Acrobat Reader Get Postscript Version Get PDF Version

CRAY T3E at the Research Centre Juelich - Delivering GigaFlops Around the Clock

Jutta Docter
Zentralinstitut fuer Angewandte Mathematik
Forschungszentrum Juelich GmbH, Juelich, Germany

ABSTRACT:

Scientists from different research areas made intensive use of the first CRAY T3E system in Germany already during the initial learning phase. The 512 CRAY T3E nodes delivered in December 1996 and a growing number of users require automated and optimized scheduling of batch and interactive work. A status update of the CRAY T3E at the Research Centre Juelich is given and the implemented scheduling mechanisms are described.

Introduction

The Central Institute of Applied Mathematics (ZAM) at the Research Centre Juelich was one of the first computer centres in the world to install a CRAY T3E in August 1996. A large community of researchers throughout Germany have been provided with computer resources for parallel applications since then. The CRAY T3E is imbedded into a supercomputer complex with a CRAY M94/4-512 used mainly as file server and for interactive work and a CRAY T916/12-512 for vector-intensive batch processing. The CRAY M94 will be replaced by a CRAY J90se in the near future. This description of the actual situation and status should help other installations to avoid running into the same problems and will demonstrate SGI/Cray why a production environment like the one established at the Research Centre Juelich has different needs than development machines.

Hardware Installations

A CRAY T3E with 128(+8) PEs was installed in August 1996 with additional 8 PEs following in October to allow execution of 128 PE applications. In November the CRAY T3E was upgraded to 256(+16) PEs. Lack of performance in accessing data on the fileserver disks could be improved by activating the HiPPI connection to the file server. Performance problems with the local SCSI disks were reduced by a second Multi Purpose Node (MPN).

In December 1996 256(+16) additional PEs were installed as a separate machine including a duplication of the peripheral hardware. The second CRAY T3E was dedicated to one user and only one application running a so-called safe environment to test if this mode of operation was more stable than the multi-user mode on the first CRAY T3E under UNICOS/mk. Unfortunately the hardware of the second CRAY T3E was less stable than the first, so that a real comparison was not possible.

The year 1997 began with a hardly better situation. Some of the MPN problems were addressed by exchanging the system support bridges in the MPNs in January. In February 1997 the two CRAY T3E frames were coupled into one system and a full hardware checkout was performed including an upgrade of those interconnect chips to I4-chips, which are connected to the GigaRing.

The actual configuration consists of 512(+32) PEs, 512 PEs for parallel applications, 6 PEs for the operating system, 20 PEs for command execution, and 6 redundant PEs. 64 SCSI disks connected to 4 MPNs provide 256 GBytes of temporary storage. HiPPI and FDDI connect the CRAY T3E to the network and to the file server, where the permanent user data resides.

Software History

UNICOS/mk has undergone multiple upgrades in the last months including the first official release 1.3 and release 1.4 which officially provided the capability to run 512 PE applications and was also intended to improve stability. Due to frequent software interrupts a new archive was installed almost every week. At present the CRAY T3E is running UNICOS/mk 1.4.260 and NQE/NQS 3.1.

The System WorkStation (SWS) now runs SWS-ION 2.9. It has also been upgraded several times which often meant changes in the behaviour of the configuration, dump, and restart procedures. The available method now of dumping all PEs to MPN disks provides much more data for error analysis, but one has to be aware that one 512 PE dump needs up to 4 GB of disk storage.

The frequent software changes have still prevented the training of non-T3E system administration personnel to manage the machine after a crash.

Interrupts

The average interrupt rate per month has decreased from 40 in 1996 to about 24 per month in 1997, while the number of active users per month has increased from 20 to about 60. In March an eight days acceptance period was completed with less than 10% downtime. The interrupt rate is still not acceptable for a production mode and SGI/Cray will have to strive to reach the goal of "only one interrupt per month" by the end of the year.

The CRAY T3E performs well while it is up and running. Unfortunately UNICOS/mk is currently neither robust nor resilient. Problems with a single PE, with the network or in an application may bring the whole system down, or worse, into a state where only a part of the system is functioning properly. The latter situation is difficult to detect and diagnose. The only way to recover is to reboot the CRAY T3E. In addition the cause of the problems often seems to be not well understood.

This situation leads to loss of productivity for users trying to develop or optimize their parallel applications, to frustration, and to mistrust in the technology. In addition production time is lost due to several other reasons. It takes at least up to one hour to take a dump and restart the system. The failing jobs have to rerun. The autobooter which was developed locally to restart the CRAY T3E without manual interaction can not always clear the situation if the same job crashes the system immediately after restart again.

To reduce downtime the Research Centre Juelich had established an additional on-call coverage seven days a week from 8 a.m. to 12 p.m. from January to March 1997 which was in addition supported by the on-site SGI/Cray personnel. This saved over 600 hours of production time. Expectations of more stability and a permanently improving autoboot mechanism should make this arrangement obsolete.

Problem Detection and Error Reporting

Offline and online diagnostic programs don't always detect hardware problems. Sometimes user applications are more likely to fail in such a situation. In many cases it is not obvious if a problem is caused by hardware or software, like multiple bit errors which were first corrected by replacing memory daughterboards on the PEs but finally turned out to be a software problem. Frequent hardware changes complicate the dump analysis because it is important to know exactly the actual status of the machine: which PEs are mapped out, how many links are disabled, etc.

About 250 hardware and software problems have been reported and closed since September 1996. Up to 100 incidents from critical problems to minor and design requests are still open or not yet reported due to the serious backlog of system dumps to be analysed and issues to be addressed. Due to the amount of open SPRs, the category critical was not even sufficient. A Top Ten List was created to identify the most critical problems, which then have a chance to get fixed more quickly.

The documentation which problem is fixed in which release/archive is not at all adequate. SGI/Cray should more efficiently take into concern that most of the offered features like limits, accounting, logging, etc. are immanently necessary for a site running a CRAY T3E in production mode with many users. These features should also be activated and used on the development machines to verify their functionality in an early state.

Intensive end-to-end testing is necessary, for example, if limits are changed in the UNICOS/mk kernel, it has to be verified that NQS works consistently with the new interfaces and that the accounting information is still valid.

Specific Problems

- *User applications might crash the system* .

  - As we experienced during the first months, some applications using *stream buffers* could crash the system. Therefore usage of the stream buffers is disabled in the User Data Base (UDB) for all users sacrificing performance. Selected important applications are thoroughly tested during dedicated times to verify that they are *stream safe* . Only after this test the usage of stream buffers is enabled for that particular user. This is a time consuming activity and guarantees by no means that a slight modification of the code will not crash the system.

  - Other applications have caused system break downs due to the allocation of too much memory. This should only result in messages like *no more memory available* or *operand range error* and terminating the application. Ironically the fix for that problem was just one byte long.

  - Some 512 PE applications combined with heavy I/O cause timing problems in the torus, leading to *probable auto-prod* interrupts, which are not always reproducible.

- *Reconfiguration of PEs is still not reliable* .

  - One complete weekend was lost because disabling only the PE but not the links of that node, which had been a proven method so far, failed after the upgrade to 1.4.2 and it was difficult to detect that this was the problem causing all the effects. The process of mapping out PEs and links is very complex and needs improvement and automation.

  - Some 512 PE applications do not run any more with an increasing number of PEs mapped out (e.g. 3) although enough PEs are available. The program does not get loaded completely and blocks the PEs forever. It is no real comfort that this application runs successfully in case only one PE is mapped out.

- *Failing PEs no longer result in crashes* .

  - In earlier versions of UNICOS/mk a faulty PE crashed the system consistently, now the PEs just fail while the application seems to continue to run. This is a minimal advantage during the day, when only one user is affected and a maintenance period can be scheduled to repair or map out the faulty PE and reboot the system. At night or during the weekend production time is lost when a situation like this hangs a 512 PE application and this is not detected from any available monitor, because the CRAY T3E seems to be pretty much alive.

  - Users who have a good knowledge about the range of results they expect from their computations reported inexplainable values, which were not reproducible on a functioning machine afterwards. This leads to the assumption that PEs might perform wrong calculations before they actually fail. There is no chance to verify that for users who can't estimate if their results are right or wrong without repeating each run.

- *Hardware stability*

  - The Multi Purpose Node connections to the SCSI disks have caused trouble with disk I/O. A first improvement was the exchange of the system support bridges, but *DMA timeout* and *DMA hangs* , followed by automatic MPN resets are still seen until another upgrade of the MPNs will be available.

  - The current power supplies show many power glitches which might cause PEs to fail. A complete exchange of the power supplies is announced to solve this hardware problem.

- The expectation in the MTBF of a processor is a lot higher in a 544 PE system than for a single workstation because the MTBF of the whole system is impacted by a factor of 544. Unfortunately the MTBF of the DEC Alpha chip is not as good as requested. Frequent exchanges of processor elements or other frequent hardware changes might have a negative influence on the overall stability.

- *Limits*

  - The functionality of limits in UNICOS/mk has evolved with the different releases. Release 1.4.2 is the first to essentially meet the needs of a production environment with many different users. The definition of e.g. the time limit changed from being the product of PE numbers and requested time to the requested connect time. This should be clearly documented together with the consitency of units to define, display or check memory limits. The changes to external specifications unfortunately do impact the users. Definition and testing should always involve all areas where limits or units are relevant from the kernel and the UDB to NQS and accounting.

Effects in/on the CRAY Complex

As the CRAY T3E itself gets more robustness against failures new more subtile problems appear with other serious effects. Two examples:

- NQS stopped all queues because *the process table was full* and no new job was started. This seems logical to avoid more problems in running jobs, but as long as this happens at night a lot of production time might be lost. Redefining the size of the process table had no effect. The issue was finally corrected by a kernel modification.
- The CRAYs at the Research Centre Juelich are coupled tightly so that problems on the CRAY T3E also effect the other supercomputers. The CRAY T3E e.g. sets file locks on the fileserver which are not cleared. The file lock table on the fileserver overflows and it must be rebooted impacting all users/jobs in the CRAY complex (M94, T912, T3E).

Scheduling

The usage of NQS batch jobs is favoured to get a maximum of PE utilization. Long phases of batch production on the whole machine alternate with phases which allow interactive testing on about a quarter of the PEs.

To allow execution of 512 PE applications all running applications must be terminated at the beginning of the batch period (night/weekend) . There are NQS queues defined for 512, 256, 128, 64, 32, 8 PE and CMD applications. Big jobs have priority over jobs requesting less PEs and jobs requesting less than 10 minutes are preferred to jobs of the same size with a longer time limit (max. four hours). The priority of the NQS queues are defined accordingly and scheduling starts with big jobs to smaller jobs to avoid the splitting of the application partition.

Every class of jobs runs for a certain time to distribute the cpu time over the submitted workload. During the day the number of PEs reserved for batch usage decreases and smaller jobs will run in advantage of an increasing interactive partition until the next batch cycle starts. Interactive applications are limited to 128 PEs and a maximum of 15 minutes connect time respectively an equivalent hereof.

PE scheduling without gang scheduling is more optimal if the workload is known at the beginning of a batch phase. If big jobs are submitted and scheduled while the machine is occupied with smaller jobs terminating at different times a starting NQS job might wait for a consecutive partition leaving a lot of PEs idle. There is no coordination between NQS and the scheduling of the PEs. More sophisticated scheduling algorithms like policy driven scheduling will have to be developed. These functions which are already partly implemented need robust and reliable entry points to easily implement site specific scheduling policies.

Plans

More and more users develop bigger and bigger applications leading to a serious amount of 512 PE applications requesting the whole machine. It is required to make the CRAY T3E robust against stress factors like those seen in large applications using barriers and doing heavy I/O. The installation of UNICOS/mk 1.5 and SWS-ION 2.10 as soon as possible will hopefully be the next step in this direction. Hardware improvements in the area of processor elements, links, power supplies and multi purpose nodes are also expected to improve the stability and robustness of the CRAY T3E. The political scheduler and other new features introduced with the coming releases will have to be evaluated and adapted to the production needs. The data from system accounting will have to be delivered into the central data base for accounting at the Research Centre Juelich.

Conclusion

The CRAY T3E hardware is less stable than expected and more reliable diagnostics and automated reconfiguration mechanisms are required. Users can solve new large problems on the CRAY T3E as long as interrupts don't prevent them from executing jobs. The CRAY T3E is a permanent challenge because of the permanently evolving hardware and software. SGI/Cray will have to provide the sites with a stable environment, especially for large systems, as soon as possible.

Table of Contents | Author Index | CUG Home Page | Home