T3D Experiences at ECMWF

Graham Holt, European Centre For Medium-Range Weather Forecasts, Reading, Berkshire, RG2 9AX, England

ABSTRACT: *This presentation describes the limited experience Computer Operations staff have of providing a service on a T3D, and describes the problems that are expected when the time comes to monitor and control a T3D in an operational environment. At present there is no operational work running on the T3D but this does not alter the problem of how to determine if the service provided is a good one and if the code that the users are generating is efficient. In truth it is not clear how to determine what the most important problems are and how adapt traditional skills to the solution of problems on a parallel processing system.*

By highlighting the way the C90 is monitored and managed, and comparing this with the lack of information available about the work running on the T3D, it is hoped the reader will understand why the lack of raw data and hence the lack of tools on the T3D is such an issue for a site like ECMWF.

1 Glossary

In this article the pre-UNICOS 8 names of the C90 monitoring tools provided by Cray analysts are used namely **systop**, **sysmon**, **syscall**, and **crayperf**. This is done because not all of the features provided in these tools has been made available via xsam. In addition local scripts namely scansds and newscan are used very frequently and if further information on the tools is required information can be provided.

2 The Presentation

Slide 1. T3D experiences at ECMWF.

GRAHAM HOLT

.Operations Supervisor.

.C90 System administrator.

As Computer Operations Supervisor at ECMWF with special responsibility for the services supported by the C90, I have some knowledge of the way the C90 is configured and used. During the last 2 years both the system configuration and the users jobs have been modified to make each more suited to the other.

Slide 2. The task.

To monitor and manage the computer service provided on the T3D and to assess the likely problems when operational tasks are to be run with jobs from in-house and remote site users.

Thus I was given the task of ensuring the shiftleaders and operators could monitor the T3D and could perform the usual tasks - system dumps and restarts. The skills of shift staff have been improved through experience so that they can now monitor and manage the C90 very efficiently. In theory this should be possible on the T3D as well.

Slide 3. My work as Operations Supervisor.

.Monitor C90 performance.and operational forecasts.
.Detect, document and correct problems.
.Schedule C90 batch work to maximise performance and to meet operational targets.

The shift staff at ECMWF do a lot more than monitor the C90. They monitor many other systems and all of the tasks that are part of the Operational Suite. The forecast is a very important task but there are hundreds of other tasks too. Changes are

frequently made both to programs, libraries, compilers and operating systems. Inevitably problems occur and it is the shift staff who initially detect, document and correct these problems. They also schedule the C90 batch work alongside the operational work to maximise the cpu utilisation without compromising the operational schedule. Needless to say the fair-shares are stacked in favour of the operational work.

Slide 4. C90 System Administrator.

.Alter queue user/runlimits.
.Modify nice values.
.Create complexes.
.Monitor memory and SDS use.
.Maximise cpu utilisation.
.Minimise disk I/O.
.Advise analysts of users needs.

As C90 system administrator I am responsible for altering C90 queue limits, queue and batch job nice values. I create complexes if needed and regularly monitor the way UNICOS manages main memory, SDS, the cpu's and I/O. As user's needs change I keep the systems analysts aware of the way the system responds to the change in the type of work submitted. If users have problems with job turnround the problem is passed to me for initial investigation.

Slide 5. Cray systems at ECMWF.

.C90 16/256 512 MW SSD.
Batch work user and operational.

.YMPEL 6/128.
Interactive and batch user.

.YMP2E / T3D 128 PEs (8mw)
Interactive and batch for research development only.

There are 3 Cray systems at ECMWF.

The C90 which has 16 cpus, 256 mw of memory and a 512 mw SSD. It runs batch work, user and pre-operational during the day and user, pre-operational and operational overnight.

The YMPEL which has 6 cpus and 128 mw memory. This system is configured to support a mixture of interactive and batch work. Additionally user can debug T3D programs interactively on this system.

The YMP2E / T3D which is being used to develop and test weather forecast models equivalent to those run on the C90. Some of this is done interactively and some in batch mode.

The T3D system has been reasonably reliable, 7 events resulting in loss of service during the last 7 months, longest downtime 4 hours, total downtime about 10 hours.

Slide 6. Model performance.

.What are the problems.
.Is the T3D running well.
.How can I tell?
.Ask an analyst?
.No, ask the user!

The developers of weather models on the T3D have been working for some months now. The model is running 75% faster than it was to start with but there are still concerns about model performance. But what are the problems. Is the T3D running well? How can I tell ? On the C90 the answer would be see an analyst. On the T3D the answer is ask the user!

Slide 7. C90 vs T3D performance (display not shown here).

This slide was produced by David Dent of ECMWF based on benchmark tests conducted on Cray systems at ECMWF and elsewhere. As you can see from this slide an equivalent performance can be obtained from a 128 PE T3D and just over 4 cpus on a C90. Alternatively it takes about 400 T3D pe's to run a forecast at the same speed as it would using 15 cpus on the C90. But the T3D model is not doing any postprocessing where-as the C90 based model is postprocessing so the comparison is not totally accurate.

Slide 8. What else do we know.

.Operational models take 1.5 Hours on the C90 and 6 Hours on the T3D.

This equates to about 15 M/flops per PE on the T3D. The alpha chip is rated at 150M/flops. Thus the model averages 10% of peak speed.

THIS IS A PROBLEM!

Looking at the problem slightly differently. David Dent tells me that on the T3D the T213 L31 operational model would take about 6 hours to run. On a dedicated C90 it would take about 90 minutes. 4 times faster on the C90. Now the C90 is rated at 6 G.flops which gives the T3D a rating of 1.5 G.flops. Very roughly speaking 15 M.flops per pe.

But the alpha chip is rated at 150 M.flops so the T3D version of the operational model, admittedly still under development runs at an average of 10% of peak performance.

This for me is a BIG problem. 10% of peak is way too low.

If I could find a way to run the job faster then I could provide more users with the same service or existing users could be given a better service.

Slide 9. The problem revealed?

.What limits T3D performance?

.How can improvements be made.

ANSWER : I do not know.

So what is it that limits T3D performance. How can I help identify where improvements can be made using the tools provided ? Are there any tools ?

If for example PVM has a problem and the message passing between the YMP2E host process and the T3D nodes fail then the T3D tasks will do nothing until time limit is reached.

As far as I can tell there are no commands available on the YMP2E that tell me what the T3D is doing. There are no tools that show what is happening inside the PE's. It is possible to find out what the YMP2E is doing on behalf of the T3D but not what the T3D is doing.

Slide 10. Performance monitoring.

.Computer Operations staff need monitoring tools. Why is this?

.Operational forecasts run between 17:00 and 05:30 every night.

.Shift staff are on duty to detect, document and overcome problems if they can.

Over the years that Cray systems have been in use at ECMWF shift staff have been encouraged to use the tools that both Cray and ECMWF analysts have created. Shift staff monitor what is happening in memory, ldcache, SDS and on swap devices. They also monitor the time taken for jobs to run. Variable run-times have historically been a feature of Cray programs and gaining an understanding of why this is has enabled Computer Operations staff to understand the weaknesses and to make a positive contribution when system reconfigurations have been necessary. Additionally because the shift staff are on duty 168 hours a week shift staff have been encouraged to describe the problems they see very accurately and they can do this because of the tools at their disposal.

Operational work is run almost exclusively between 15:15 and 05:30 with the critical period starting at about 17:00. Thus during the critical period the shift staff are usually on their own. They are encouraged to detect, document and overcome as many problems as they can but there are analysts on-call at all times if the problem is beyond the operators abilities. The operators skills in detecting and solving problems (even if it is just a server reboot) are not limited to Cray systems. Any system that has a problem, any network device be it LAN or WAN that fails, is a problem for an operator or the shiftleader.

Slide 11. Is the T3D that different ?

.The T3D is no different to other systems.

.Monitoring will identify problems.. This is the first step towards improving the service provided.

Is the T3D to be treated differently. I do not think so. Monitoring any system will identify problems. Once problems have been solved the service provided will be improved. The T3D is a 'young' system. But when parallel systems become more complex (e.g. when an IOS is attached) it will be essential that the activities of the system can be monitored, particularly when operational work is running. I just cannot accept the idea of a system that cannot be monitored.

Slide 12. Improved service means...

.Improved turnround
.Shorted development period
.Better forecasts
.Faster code
.Faster recovery after failure.
.Reduced delays to schedules.

It is very important to continue to improve the services provided. As a service improves so benefits emerge. Users find that there is an improved turnround. They are able to develop their programs more quickly or more development can be done. Thus better forecasts are made possible and that leaves time to create faster code. Thus it is possible to provide a more reliable service to paying customers, as recovery time after failure is reduced as are delays to operational schedules.

Slide 13. C90 monitoring has provided great dividends.

***** this slide shows the C90 cpu utilisation chart for February 1995.

idle time,	system time,	user time
1.11%	4.50%	94.39%

C90 monitoring and management by operators is now finely tuned. The skills that the shift staff have are of great benefit to all users. The system is well used. Idle time is about 1%. Each

job can be monitored and at critical times jobs are monitored constantly.

Slide 14. shift staff monitor and manage.

Shift staff monitor and manage C90
resources as follows:

systop - percentage system/user time per process.

sysmon - memory, swap and SDS maps.

syscall - total number of system calls and
percentage cpu time used..

crayperf - 3 swap maps, memory map,,
cpu graph and swap graphs.

jstat -j 'reqID' - displays for each process
the user time and system time,
the command and system call.

ldcache -l /dev/dsk/'filesystem' shows the
ldcache hit rate.

systop - a display of all active processes. This is run constantly with a refresh time of 10 seconds. Thus it is possible to see how much cpu time on average has been used by each process, how much is user time and how much is system time. This tool displays user time as ******* and system time as ---------.

A process showing ----------**** is bad

A process showing -********** is good.

jstat -j "reqID" . The command jstat -j 'reqID' shows for all processes for a specified job - user time, system time, command and system call. Using systop and jstat bad processes can be identified and correlated to the batch job. The job can be watched, analysts can be called, users can be called, action can be taken.

syscall - displays the sum of the cpu utilisation for each system call in descending order.

sysmon - has a number of displays (I/O to cache and disk, maps of memory, swap and SDS, logical and physical disk transfers and kernel information).

crayperf (xsam) - shows the same as sysmon but graphically.

ldcache -l /dev/dsk/'filesystem' - shows that there is either enough or too few ldcache blocks assigned.

The use of these tools by myself and shift staff were influential in the process of tuning the C90 to the work to be run.

Slide 15. Good monitoring tools.

These enable shift staff to notice the between
a poor service and a good service.

They encourage pro-active operating.

If you see the problem as it happens you can
fix it much earlier.

So experience has shown that good monitoring tools allow shift staff to gain an in-depth knowledge of the work that runs and the way UNICOS manages resources. By a process of constant involvement they just 'know' that is happening and almost instinctively are able to correlate variations in system activity with periods of bad service and of good service. They become pro-active and look for problems as soon as unusual events take place.

Slide 16. Pro-active operating means ..

.Finding problems
.Investigating problems
.Documenting problems
.Minimising operational delays.
.Identifying inefficient code.
.Chasing solutions to problems.
.Working closely with analysts, users
and User Support staff.

So at ECMWF shift staff have become pro-active. They keep a watchful eye on all displays and they detect problems almost intuitively. I have documented all that I know about the work and the system weaknesses and as a result I spend much less time monitoring the C90. The shift staff now detect problems, investigate problems, document problems and solve many of them without assistance. This minimises operational delays. They often detect jobs that are not performing well, often jobs with a corrupted control block that will loop until cpu time-limit reached. They contact the users, call analysts, inform user support and work with these people to ensure the problem is fully documented so that further analysis is much easier.

Slide 17. Everyone needs monitoring tools.

All staff involved in solving problems need tools. Frequently they are created by analysts and are passed on to shift staff.

Tools need raw data to work with.

There is no raw data..

This is what is missing from the T3D.

Thus shift staff are members of a wider team and like analysts require tools to help in the detection of, the understanding of and the solution of problems. But as there is no raw data available on the workings of the T3D, analysts cannot create monitoring tools and there is nothing shift staff can do.

Slide 18. Operators need to be able to .

.Take a snapshot of each PE to see what it is doing/waiting for.
.Look at the history of the events in each PE (sar). Has the PE done useful work lately?
.Send signals to a job asking that the PEs are released to enable operational work to run.

Shift staff need to be able to interrogate the T3D.

It should be possible for an operator or analyst to find out what each pe for a given job is doing or waiting for. An equivalent command to jstat -j "reqID".

It should also be possible for an operator or analyst to look at the history of activity in a specified pe to see what the performance of a process or pe is like. It would be of great benefit to know how much useful work a PE has done.

When running the T3D operationally the release of PE's is a major problem if the work done so far by the job in execution is to be saved. Checkpointing is not a good solution. What is needed is signal processing. An operator should be able to tell a job to tidy up, create a restart file and finish so that the processing that the job has done is not lost. At ECMWF users are told to create restart files regularly so that the time lost when a job is killed is minimised. This is a weak solution. Signal processing is a better answer and this takes the responsibility away from the user and ensures processing time is not wasted.

Slide 19. Summary

History has shown that the service provided on Cray systems can be improved if problems can be detected and investigated by Computer Operations staff.

This can only be done if raw data is provided for analysis..

The T3D provides no data and this is a severe weakness.

The basic principle is this. Computer systems never work exactly as the manufacturer says they will. Computer Operation's job is to detect the problems and capture the data needed for others to understand and correct the problems. Good data capture enables good tools to be created. This results in early problem detection, in better solutions and ultimately in a better service to users.

Each year the service provided to the users should improve. When parallel systems are in more general use I will need to know that jobs are running well and to find out what the problems are by monitoring the service provided.

This concludes my presentation but please note for those interested in the fine details of IFS model performance on T3D systems, I have copies of the paper written on this subject that was presented by staff and consultants at the workshop on Parallel Computing at ECMWF in November 1994.