Automated Error Reporting on the Cray T3D

Michael W. Brown, HPC Operations and Service Manager, Edinburgh Parallel Computing Centre, The University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom

ABSTRACT: *The University of Edinburgh operates a national service on a 512-processor Cray T3D, a Y-MP4E and a J916 on behalf of the UK academic and research community. Very high availability and servicability are demanded from these systems, but reductions in operations cover has meant that automated error detection and reporting is required so that anomalies can be detected and systems staff alerted during shifts not covered by machine room operations staff. This paper describes the work done by the University and local Cray Research staff in which hardware and software monitoring is undertaken continuously, and problems detected and signalled automatically to systems staff.*

Site Overview

On behalf of the UK Engineering and Physical Sciences Research Council (EPSRC), the managing agent for academic supercomputing in the United Kingdom, the Edinburgh Parallel Computing Centre (EPCC) at the University of Edinburgh has run a peer-reviewed national service for Grand Challenge science in a Cray T3D system to the British academic and research community since July 1994.

The original 256 processors and the Y-MP4E front-end system were installed in April 1994, and the configuration received its first substantial upgrade in December 1994 when an additional 64 processors were installed in the T3D.

The operation of the service has already been described [1] and [2] but there have since been a number of substantial changes to the service, with the upgrade of the T3D to 384 processors in October 1995 and finally to 512 processors in February 1996.

In addition, the front-end system (a Y-MP4E) was augmented by a 10-processor J90 system in October 1995 and at the same time the total on-line disc capacity was doubled, with in excess of 512 Gbytes being shared across both systems.

Attached to the J90 system is an IBM 3494 Automatic Tape Library containing 2000 cartridge slots. This system contains both 3490E (800 Mbyte) and 3590 (10 Gbyte) drives, and with the migration to the higher density cartridges will have a capacity of 20 Tbytes.

The Y-MP and J90 systems exist only to service the requirements of the T3D, and because of the extensive inter-dependency between the system due to the distribution between and cross-mounting of the user filesystems, the maintenance of the Y-MP and J90 services are of equal importance, and the T3D service cannot operate without either. This is extremely significant when the need for effective automatic monitoring is discussed later.

Within the configuration, the respective roles of the Y-MP and J90 are as follows:

Y-MP

- Front-end host for T3D
- Batch job server for all T3D jobs
- Fileserver for large workspace filesystems using fast (but expensive) IPI discs

J90

- Fileserver for all user /home filesystems using slow (but cheap) SCSI discs
- Batch job server for all T3D development, pre and post- processing work
- Interface to IBM 3494 mass storage system under the control of DMF
- Server for interactive editing and compilation

A schematic diagram of the various HPC systems at EPCC is appended (diagram 1).

In addition, a Cray T3E system will be installed at the end of 1996. This system is being operated by EPCC on behalf of the Particle Physics and Astronomy Research Council (PPARC) for dedicated use by the UKQCD community. This system shall be closely-coupled with the existing machines as there will be a

high degree of data-exchange between the T3D, T3E and J90 systems, with data-storage under DMF on the IBM 3494.

Reasons for unattended running

The reasons for having operational cover at University sites like EPCC have changed greatly with the move away from large central facilities and systems requiring a high degree of manual intervention in ensuring effective throughput of batch work. Developments in tape technology, with a substantial move towards automation robotic systems, have meant that the necessity to have full-time operations cover has greatly diminished.

The Cray systems at EPCC run substantially without the need for any intervention by operations staff, with regular NQS configuration changes and reloads of the T3D to change the pool structure being performed automatically under the control of `cron`. The bulk of tape requirements are satisfied by small autoloaders and the large IBM 3494 Automatic Tape Library.

The need to make substantial savings in staff costs, and the redeployment of some operations staff into non-shift based support roles has forced a reduction in operational cover from the 24 hours/3 shift cover previously necessary. Although operational cover remains for two shifts per day, Monday until Friday, for more that 50%of the week the machine room is unmanned. As it turns out, the busiest time for the Cray T3D (when it is exclusively running production batch work), corresponds to the times that there is no operational cover in the machine room.

When the service on the T3D was started, it was clear that absolute reliance could not be made on having operational cover to ensure the integrity of the service, thus the need for effective error detection and recovery without the need for direct intervention by operations staff was necessary.

Deficiencies in Cray provision

Before the T3D was even delivered to EPCC, it was noticed that in the event of a problem on the T3D that the front-end host would not be automatically notified. It was hoped that OPI running on the OWS would be able to interpret a failure indication initiated from the T3D, but no such interface was ever implemented. Unless somebody was watching a mppview type display permanently, a crash on the T3D could remain undetected for a considerable length of time. For a system that was designed to run 24 hours out of 24, clearly this was not acceptable and a locally-designed, implemented and supported mechanism was necessary to enable fault conditions on the T3D to be detected as they arose, and signalled to operations and systems staff.

Detection of failures on the Cray T3D

Most problems detected on the T3D are written to the `/usr/spool/mpp/mppsyslog` file. This file contains a large amount of information about the starting and stopping of each user application as well as error information. Information on memory errors (both singlebit and multibit) is written in addition to the binary version of the `mppsyslog` file which may be interrogated by olhpa, but this does not give a very effective way of signalling errors, and in any case most T3D error conditions may not be detected in this way.

The number and type of error conditions that may be signalled to the `mppsyslog` file is quite large, and although some conditions are non-fatal the bulk of errors signalled are fatal to the operation of the system. In particular, a multibit memory error will always fail the user application that is running on the processor and the partition will shutdown although the system will continue in operation. If the error occurred not in the user part of the memory, then in addition to the partition shutting down, the entire T3D will fail as the barrier network shuts down. This lack of resiliance has caused the loss of much user time on the system, although the recently announced `softpanic' feature in UNICOS-MAX 1.3.0.3 adds reliance to the system in certain cases.

Typical error messages are appended at the end of the paper.

On the T3D signalling an unrecovered error, the relevent information is written to the mppsyslog file and the system is stopped. No other action is taken by UNICOS-MAX, and UNICOS remains ignorant of the problem.

Monitoring and reporting requirements

With the inability of UNICOS or AIR to detect and report T3D errors, it was necessary to start a local exercise to:

1. Devise a method to enable the Y-MP to detect all T3D-related errors;
2. Devise a method of interpreting the signalled errors and informing operations and systems staff.

In addition, to cover for the possibility of problems on the Y-MP itself taking the system down when there was no operator cover, it was desirable to:

1. Devise a method to interpret and signal any Y-MP failures to operations and systems staff;
2. Devise a method to interpret and signal infrastructure equipment failures to operations and systems staff.

In order to undertake this, it was first necessary to identify the problems that could cause the failure of the service, especially during the hours of unattended running. Because of the close-coupling of the T3D and the Y-MP (and later the J90), any failure on the Y-MP side that could cause the loss of the T3D service had to be considered. The following possible reasons were thus identified:

T3D

1. T3D hardware or software failure (as reported in the `mppsyslog` file)
2. Loss of communication between Y-MP and T3D through the IO gateways
3. Loss of input power
4. Automatic shutdown due to failures in the cooling circuits
5. Automatic shutdown due to smoke alarms

Y-MP

1. UNICOS failure
2. Loss of input power
3. Automatic shutdown due to failures in the cooling circuits
4. Automatic shutdown due to smoke alarms
5. Filling up of critical filesystems (`/, /tmp, /usr/spool etc`)
6. Unrecovered disc errors

J90

1. UNICOS failure
2. Loss of input power
3. Filling up of critical filesystems (`/, /tmp, /usr/spool etc`)
4. Unrecovered disc errors

Implementation

It was decided that the Y-MP should be used to monitor the T3D and its own filesystems etc, but to use the MWS supplied with the Y-MP to monitor the existence of the Y-MP as well as to interrogate the WACS (Warning and Control System) on both the Y-MP and T3D systems. With the subsequent installation of the J90 system, and the need to monitor it as closely as the Y-MP and T3D, it was decided to use the J90 to monitor itself as much as possible, and to use the Y-MP to monitor the existence of the J90 itself.

The most convenient method of alerting systems staff to problems was judged to be the issue of automatic radio-paging calls backed up with electronic mail and messages sent via the msgd system.

The monitoring tasks are thus distributed over three systems, the Y-MP, the J90 and the MWS, but are based on a consistent set of scripts.

A schematic diagram of the monitoring system is appended (diagram 2).

Monitoring under the control of the `patrol` system

Most of the monitoring is undertaken by a generic script called `patrol`, which is called every ten minutes under the control of cron on the Y-MP, J90 and MWS.

The `patrol` script runs in turn a series of checking scripts, each of which checks the status of a particular part of the system (such as a scan of the `mppsyslog` file for failure patterns). The result of each check is written to a series of status files. Once the checking scripts have been run, the status files are examined in turn to see if an alert trigger has been set.

In the event of a trigger being set, the failure code for the contingency is derived (see table below) and a check made to see if this contingency is already active (i.e. to see if it has already been detected by a previous run of the `patrol` system).

If the contingency is not active, then a the process of initiating a paging call is made. This is handled by the MWS which has a dedicated modem attached to it, and the initiating system (Y-MP, J90 or MWS itself) will send the request to initiate the paging calls by mail to a special user id on the MWS.

If the contingency was active (i.e. already detected, but not closed down), a further paging request is not made unless a period (currently two hours) has elapsed since the last paging call was initiated. This is to prevent the issue of repeated calls every ten minutes. If two hours has elapsed, then duplicate calls are initiated.

Monitoring under the control of `fsmon`

Additional monitoring of critical filesystems is undertaken under the control of the regular UNICOS file system monitor.

Critical filesystems are those judged to be essential for the operation of the service, and include `/, /tmp,` and `/usr/spool.` If any of these filesystems become full, it is likely that NQS will fail, or (worse) that user jobs may start, and immediately fail thus losing the entire contents of the job queues. Such an eventuality is catastrophic for the user service.

fsmon runs every few seconds, and on reaching the `warning' level (typically 90%), mail messages are sent to systems, CRI and operations staff but no additional action is taken. If the `critical' level is reached (typically 95%), all NQS queues are stopped and paging calls initiated. The method of initiating the paging calls is the same as under the patrol system, with each type of contingency allocated a unique numbered identifier.

Contingency identifiers

Each type of contingency is allocated a unique five-digit numerical identifier, and in addition each identifier is prefixed by `1' or '2'. Because of some paging calls being lost (typically by a number of calls being initiated simultaneously, and the modem becoming engaged), each call is sent twice with a period of two minutes between the calls. The prefix is to identify which calls are successfully made and received.

Paging calls and mail messages are sent to systems staff and on-site Cray Research staff, and mail messages only to shift operators. In addition, messages are sent to the msgd system, for reading by the oper command.

The numerical identifiers are as follows:

Initiated from Y-MP:

20000 test message from Y-MP
20102 IOG not responding
20110 potential error indicated in `mppsyslog`
20200 Y-MP `/tmp` almost full
20201 Y-MP `/usr/adm` almost full
20202 Y-MP / almost full
20204 Y-MP `/usr/spool` almost full
20210 Y-MP unrecovered disk errors
30101 J90 not responding to Y-MP

Initiated from J90:

30000 test message from J90
30200 J90 `/tmp` almost full
30201 J90 /`usr/adm` almost full
30202 J90 / almost full
30203 J90 /`arch` almost full
30204 J90 /`usr/spool` almost full
30210 J90 unrecovered disk errors

Initiated from MWS:

40000 test message from from MWS
40101 T3D DC power off
40102 T3D nert level dropped
40103 T3D MG set over-temperature
40104 T3D smoke alarm
40201 Y-MP cooling unit off
40203 Y-MP input power off
40104 Y-MP smoke alarm
40303 IOS-E input power off
40304 IOS-E smoke alarm
50101 No response from Y-MP

The `patrol` system automatically clears down contingencies once they have been corrected, but not before.

Monitoring done by the Y-MP

The Y-MP has four distinct monitoring roles:

(1) T3D software and hardware

It was clear that the only method for T3D monitoring and reporting that could be adopted without a major software exercise from Cray Research, was to provide scripts that searched through the current `mppsyslog` file for various failure patterns, and then to signal these in a suitable manner.

Under the control of `patrol`, a script is run which searches for the following patterns:

ALERT
PANIC PPE
Uncorrectable
deadman
wiremat
MultiBit
YPE_PT_REQUEST

These correspond to all known problems that are recorded within the `mppsyslog` file that are fatal to the operation of the T3D.

Previously, individual patterns were seached for with different scripts, but the result could be a multiplicity of paging requests issued within a very short time, as certain T3D failures provoke more than one of the above failure patterns.

The `mppsyslog` file is rolled-over at each T3D reload (typically once per weekday to change the pool configuration), so the file is never so long that the searches take an inordinate amount of time.

(2) Communication with the T3D

Under the control of `patrol`, the mppping command is run to check on the existence of the T3D down the two IOG (IO Gateways).

The loss of either gateway will cause applications running on the T3D to fail.

(3) Y-MP discs and filesystems

The filesystem monitor fsmon is run every few seconds to check on the status of certain critical filesystems.

Under the control of `patrol`, the Y-MP discs are checked for unrecovered errors, by a scan of the output of `pddstat`. In addition, a call is made of df to check whether certain critical filesystems are over-full, this is done as a backup to the checking done by `fsmon`.

(4) UNICOS running on the J90

Under the control of `patrol`, the health of UNICOS on the J90 system, and the existence of the network that joins the J90 and the Y-MP is checked by running a remote shell from the Y-MP on the J90. This health check was originally done by running ping on the Y-MP, but it was found that the J90 IOS could still respond positively even if UNICOS iteslf was down.

If the J90 is unavailable, then it is likely that user work on the T3D would be adversely affected because of the cross-mounting of filesystems between the two systems.

Monitoring done by the MWS

The MWS has three distinct monitoring roles:

(1) Environmental monitoring of the T3D

Under the control of `patrol`, the status of the environmental conditions as measured by the T3D WACS is checked. A call is made to /`cri/bin/nwacsdump`, and various error conditions are checked for.

(2) Environmental monitoring of the Y-MP and IOS-E

Under the control of `patrol`, the status of the environmental conditions as measured by the Y-MP and IOS-E WACS is checked. A call is made to /`cri/bin/nwacsdump`, and various error conditions are checked for.

(3) UNICOS running on the Y-MP

Under the control of `patrol`,the health of UNICOS on the Y-MP system is checked by a call of `ping` being initiated.

Monitoring done by the J90

The J90 has one distinct monitoring role:

(1) J90 discs and filesystems

The filesystem monitor fsmon is run every few seconds to check on the status of certain critical filesystems.

Under the control of `patrol`, the J90 discs are checked for unrecovered errors, by a scan of the output of `pddstat`. In addition, a call is made of df to check whether certain critical filesystems are over-full, this is done as a backup to the checking done by `fsmon`.

Contingencies not monitored

The principal eventuality that is not detected is the loss of power to the entire machine room. If power is lost to the T3D or Y-MP, the MWS will receive an indication from the WACS system, and initiate paging calls.

If power is lost to the J90, the Y-MP will no longer be able to run a remote shell on the J90, and a contingency would be raised, but if power was lost to the machine room, then without the installation of a dedicated UPS to supply the MWS and the modem, no such calls may be made.

Consideration is being given to entending the patrol system to enable a controlled shutdown to take place if the T3D WACS starts to generate warnings in advance of a total shutdown. Typically, a problem in the chilled-water system will cause the module input manifold temperatures to rise first to a warning level, and then to a critical one at which point the T3D powers itself off. If the T3D loses power, and the connections to the HISPs on the Y-MP have not been properly terminated, UNICOS will panic.

It should be possible to search for such warning conditions and instruct the MWS to set the switches and terminate the HISPs before the temperature reaches the critical level and the power-down initiated.

Adaptability and availability

The monitoring system is easily adaptable for use at other sites, and is not specific to the monitoring of a MPP system. The search for each type of contingency is done by a dedicated script, the monitoring system thus could be tailor-made to monitor the critical components of any Cray site (on a site that was totally dependent on the operation of its tape drives, a script could be written for operation under the `patrol` system that checks for swdn messages in `tpstat` output, for instance).

The entire monitoring system may be made available to any Cray site on request, with the only caveat being the usual disclaimers and no support committment.

Conclusions

Automatic monitoring on the EPCC Cray systems was introduced for two principal reasons:

1. To enable T3D errors to be detected since UNICOS was incapable of doing so and reporting these errors in a useful manner.
2. To enable the monitoring of as many critical parts of the entire service as possible, due to the `lights out' nature of the operation of the service for the bulk of the time.

In both these respects, the automatic monitoring implemented and developed has been extremely successful.

While the Y-MP and J90 systems have been generally very reliable, there have been a significant number of errors (hardware and software) that have occurred on the T3D system. None of these errors were intercepted or reported by UNICOS, and only the existence of an automatic monitoring system has enabled these errors to be detected by systems staff and acted upon. The consequence of a T3D problem occurring during `out of hours' can be very far-reaching, as any failure of the T3D hardware (with the sole exception of a multi-bit error that occurs in the application part of the memory on a PE) or software is immediately catastrophic as far as the continued operation of the machine is concerned.

Until very recently, there has been no graceful degradation of the system as a result of processors being lost, or barrier wires downed, and the failure of one PE running in one small partition in an interactive pool of processors that is logically totally distinct from a large batch pool elsewhere in the system, will bring the entire system down.

Although grossly inconvenient to staff, the ability to be paged out-of-hours has meant that downtimes which may have lasted overnight or all weekend, have been substantially minimised.

Many thousands of T3D processor hours have been saved as a result.

Acknowledgements

This work was undertaken under the terms of the contract between the Engineering and Physical Sciences Research Council and the University of Edinburgh with respect to the operation of the Cray systems at the University of Edinburgh.

Particular thanks are due to Stuart Wilson, the on-site analyst from Cray Research for his support, advice and expert knowledge, and who has been principally involved in the design and implementation of the automatic monitoring system.

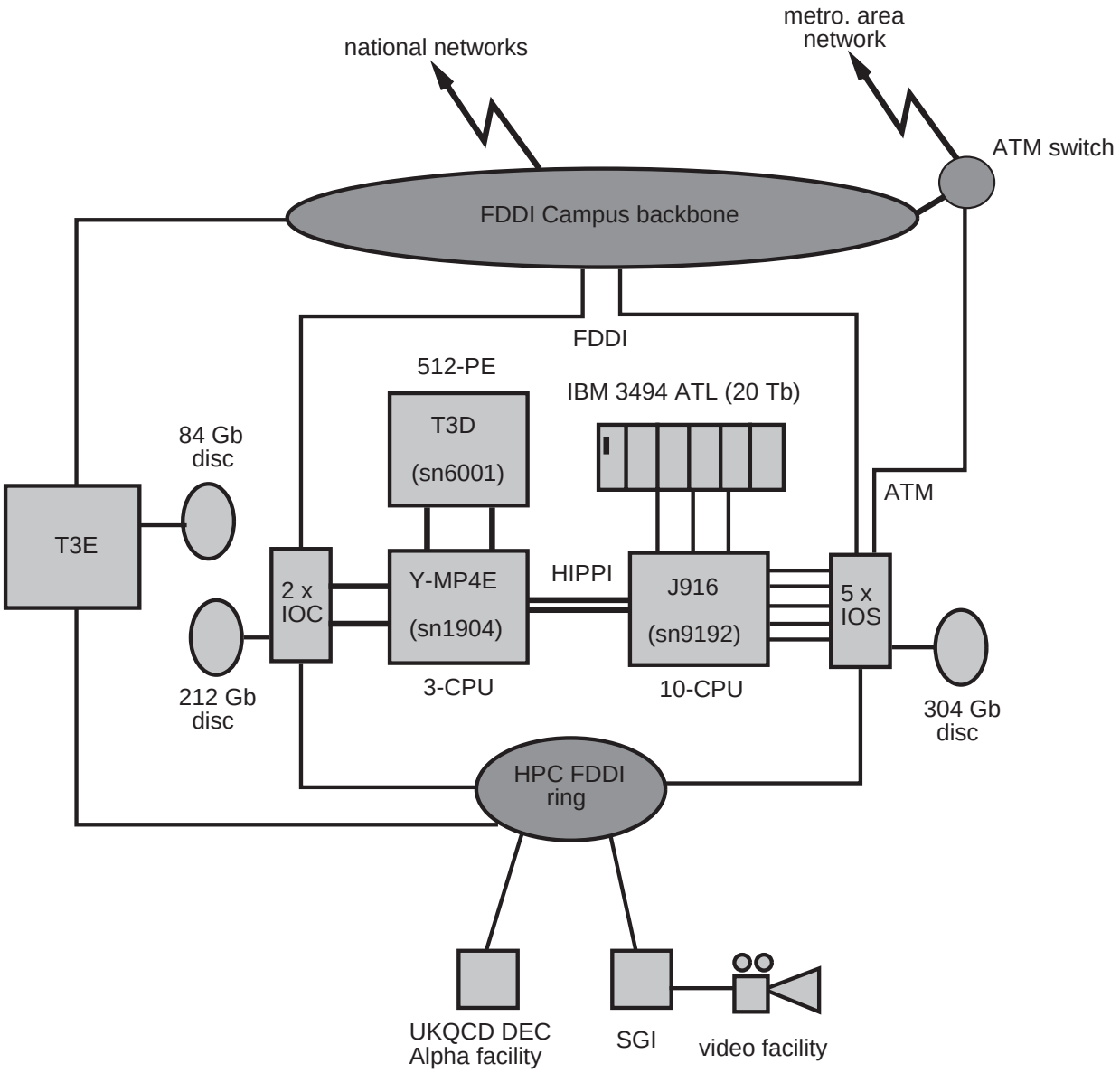
The author is on full-time secondment from the Edinburgh University Computing Services to EPCC and gratefully acknowledges the support of the Director of Computing and Information Technology Services and the Vice-Principal for Academic and Information Services.

References

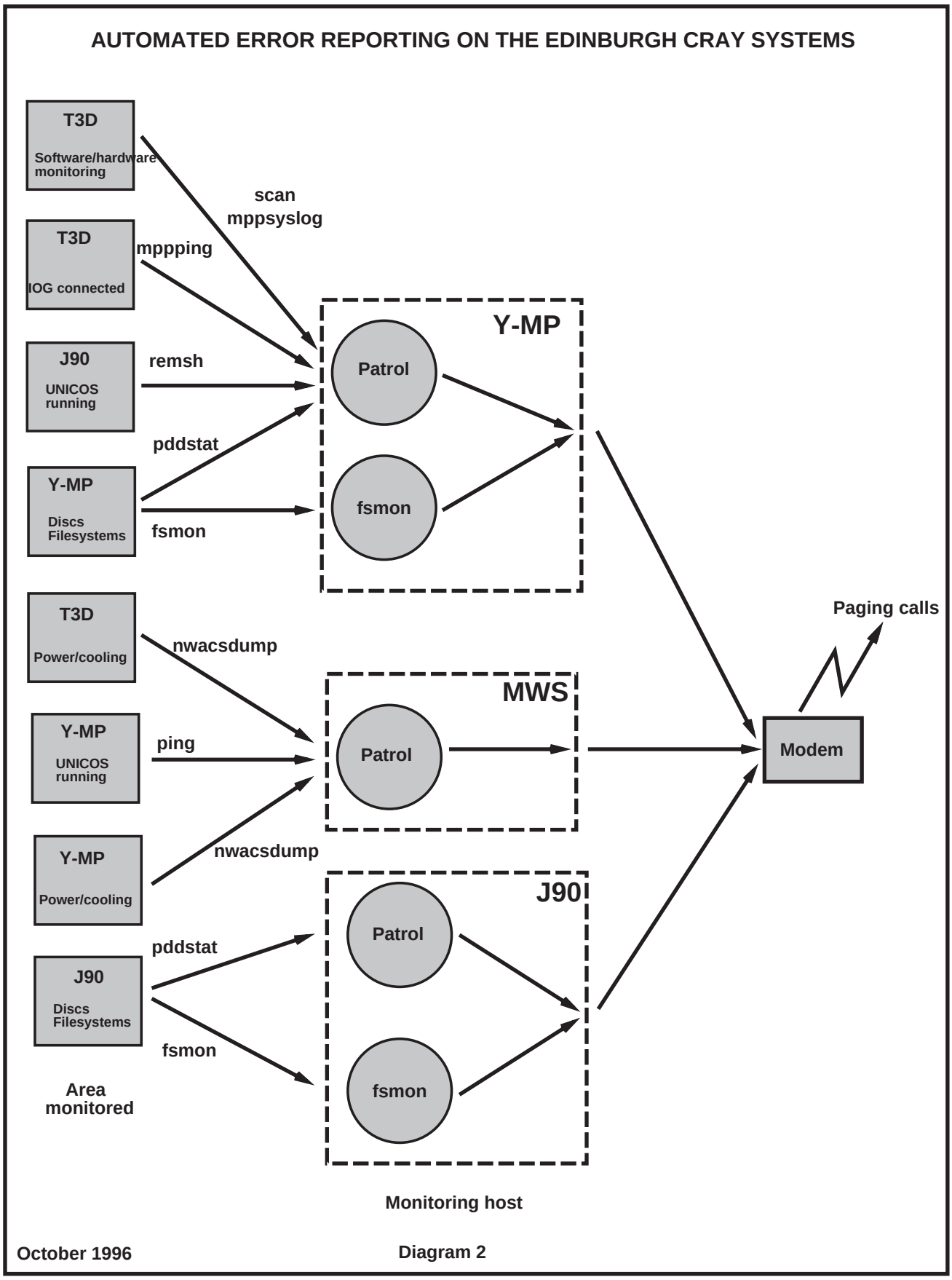
[1] Operation of the Cray T3D as a National Facility, M.W.Brown, Proc. 35th Cray User Group, March 1995

[2] Maximising Job Throughput Under NQS on the Cray T3D, M.W.Brown, Proc. 36th Cray User Group, September 1995

HPC FACILITIES AT THE UNIVERSITY OF EDINBURGH



October 1996 **Diagram 1**


AUTOMATED ERROR REPORTING ON THE EDINBURGH CRAY SYSTEMS
T3D
Software/hardware monitoring
T3D
IOG connected
J90
UNICOS running
Y-MP
Discs Filesystems
T3D
Power/cooling
Y-MP
UNICOS running
Y-MP
Power/cooling
J90
Discs Filesystems
Area monitored
scan mppsyslog
mppping
remsh
pddstat
fsmon
nwacsdump
ping
nwacsdump
pddstat
fsmon
Y-MP
Patrol
fsmon
MWS
Patrol
J90
Patrol
fsmon
Monitoring host
Modem
Paging calls
October 1996
Diagram 2

Examples of error messages placed in /usr/spool/mpp/mppsyslog

These are all that the system provides in the way of direct information in the event of various fatal conditions.

```
09/21/96 04:29:47   -  Barrier circuit 2 is unavailable because nodes have been disabled.
09/21/96 04:29:47   -  Barrier circuit 3 is unavailable because nodes have been disabled.
09/21/96 04:29:47   -  Partition O Agent PANIC: Unexpected SIGMTKILL received
09/21/96 04:29:47   -  Type = MPP_USER Priority = ALERT Subtype = 0003
09/21/96 04:29:47   -  SANITY: PPE 118 (LPE 018) : Deadman timeout occurred
09/21/96 04:29:47   -  SANITY: Disabled nodes (1)
09/21/96 04:29:48   -  Partition O Force exit message sent to PE OxO
09/21/96 04:30:45   -  Type = MPP_USER Priority = ALERT Subtype = 0003
09/21/96 04:30:45   -  SANITY: PPE 217 (LPE 117) : Deadman timeout occurred
09/21/96 04:30:45   -  SANITY: Disabled nodes (1)
09/21/96 04:32:23   -  Barrier circuit O is unavailable because nodes have been disabled.
09/21/96 04:32:23   -  Barrier circuit 1 is unavailable because nodes have been disabled.
09/21/96 04:32:26   -  Type = MPP_USER Priority = ALERT Subtype = 0003
09/21/96 04:32:26   -  SANITY: PPE 000 (LPE 000) : Deadman timeout occurred
09/21/96 04:32:26   -  SANITY: PPE 001 (LPE 001) : Deadman timeout occurred
                       (message repeated very every PE in the partition)
09/21/96 04:32:26   -  SANITY: PPE 73e (LPE 33e) : Deadman timeout occurred
09/21/96 04:32:26   -  SANITY: PPE 73f (LPE 33f) : Deadman timeout occurred
09/21/96 04:32:26   -  SANITY: Disabled nodes (128)
09/21/96 04:33:49   -  Partition O Exit response YPE_PT_REQUEST failed: Connection timed out
09/21/96 04:33:49   -  mppd: Disabling nodes in partition 0:
09/21/96 04:33:52   -  Partition O Agent core file copied to
                       /core/MPPAGENT.0921043349/core.47919
09/21/96 04:34:13   -  Partition O is released.


08/07/96 16:10:43   -  PPE 710 (LPE 710): MultiBit Uncorrectable memory error
08/07/96 16:10:43   -     (LPE 710): signature 80000062e25bO988
08/07/96 16:10:43   -     first event 30c649fd2dd last event 30c649fd2dd event countOOOl
08/07/96 16:10:43   -     ECC FILL_ADDR Ox2e25bO988: syndromes: hi = OxO (No error), lo =
                       Ox18 (Unknown)
08/07/96 16:10:44   -  PPE 710 (LPE 710) : User application experienced uncorrectable memory
                       error (ps Ox8 pc Ox200013d8dO syn OxfffffcO2e25bO988).

08/07/96 16:13:03   -  Channel/Gateway 1 has been disabled because of a timeout.
08/07/96 16:13:03   -     Logical Channel 0, Sequence number 0.
08/07/96 16:14:51   -  Partition 15 Exit response YPE_PT_REQUEST failed: Connection timed out
08/07/96 16:14:51   -  mppd: Disabling nodes in partition 15:
08/07/96 16:14:51   -  Barrier circuit O is unavailable because nodes have been disabled.
08/07/96 16:14:51   -  Barrier circuit 1 is unavailable because nodes have been disabled.
08/07/96 16:14:51   -  Barrier circuit 2 is unavailable because nodes have been disabled.
08/07/96 16:14:51   -  Barrier circuit 3 is unavailable because nodes have been disabled.
08/07/96 16:14:51   -  Partition 15 is released.
```