Chapter 10 VAXclusters

Previous chapter by Stuart Rance . . .

This chapter assembled by Richard Penn, September 1993

10.1 Cluster Registrations

10.1.1 Benefits of Cluster Registration

(How to justify to your customer the need to take some nodes down and spend half a day on site poking about the cluster).

1. Verifies that all parts of the cluster are compatible.
2. Makes future add-ons more likely to work.
3. Makes future S/W upgrades more likely to work.
4. Enables FCO planning so no low rev options get missed.
5. Improves quality of RDC input to problems.
6. Improves quality of TSC/CSC input to problems.

All clusters should be registered when they are installed. A copy of the registration should be left on site so that it can be easily updated whenever there is a significant change.

10.1.2 How to Register your Cluster

1. Come to support with a list of ALL system serial numbers and the customer name and address, we will give you some blank forms to fill in and allocate a CLUK number for the cluster.
2. Take the forms to site, allow plenty of time and be aware that you may need to shut some systems down to check on revisions.
3. Please complete ALL parts of the forms.
   - SYSGEN> SHOW /CLUSTER for SCSID, SCSNOD, Quorum, Votes, ALLoclass
   - $ SHOW LOGICAL SYS$SPECIFIC for boot device and root
   - Use the latest VAXCLUSTER revision document to determine the VAXcluster revision level. If you don't know how to do this then see the next section.

4. When you have completed all of the forms bring them to support and ask someone to go through them with you to check them. Do not just leave them on my desk 'cos they might get lost!

10.1.3 Documentation

The term **CLUSTER** has over the years grown from being a small collection of computers in quite close proximity, to large numbers of Vaxes at quite large distances away from one another. To help cope with the ever increasing diverse problems that can occur the following sources of infromation will be of help.

Table 10–1: Documentation

Part No.	Title	Decscription
EK-VAXCP-TM-01	VAXcluster Principles	Good for definative answers.
EK-VAXCP-UP-001	V5.4 VAXcluster Priciples Update.	Good for definative answers.
AA-LA27C-TE	VAXcluster Manual	Management of Clusters.
EK-VSTIT-RM-002	VAXcluster Troubleshooting Part 1.	Introduction to Troubleshooting
EK-VSTCP-RM-002	VAXcluster Troubleshooting Part 2.	Common Procedures.
EK-VSTFP-RM-002	VAXcluster Troubleshooting Part 3.	Specific Flows and Procedures.
EK-410AB-MG-002	DSSI VAXcluster	Installation and Troubleshooting.
EK-VAXCS-CG.005	VAXcluster Systems	VAXCluster Configuration Guidelines.
AA-LA46A-TE	SHOW Cluster Manual	Online view of Cluster.
EA-P2388-02	VAXcluster System Quorum	Last hardcopy distributed.
AA-PBTVA_TE	Volume Shadowing Manual	Host Based Volume Shadowing

The Cluster notesfile is very active and useful for both failure and configuration questions and is CSSE32::CLUSTER. A much valued source of information.

10.1.3.1 Quorum

The publication **QUORUM** features new technology, products, procedures for the VAXCluster environment. It also documents the latest VAXCluster Revision Matrix and is usually quicker then fiche. The distribution in hardcopy of Quorum has finnished, however it is still available to Customers over AES. Quorum can now be accessed from a STARS database on TIMA, providing your TIMA node has the DISPATCHES database available either locally or remote. It is also available under VTX Readers Choice.

10.1.4 Cluster Revision Levels.

Option revisions are calculated using a sort of tree structure.

- To check whether your VAXcluster is at revision K1 you must determine the revision of all the components, e.g all 11780's must be at rev 8, all HSC70's must be at rev B5, all TA81's must be at rev C.
- To determine if your 11780 is at rev 8 you must check the revision of all components of the 11780, e.g. WCS126.PAT, KA780 at rev 7.
- To determine if your KA780 is at rev 7 you must check the revision of all components of the KA780, e.g. M8235 rev N1.

There are extensive revision tables which enable you to calculate the revision of any option IF you know the revisions of all of its parts. The following example tree shows how the CLUSTER revision is affected by the revision of a single module.

```
                     CLUSTER
                        |
 +--------+-------+-----+---+----------+--------+-------+------+----etc.
 |        |       |         |          |        |       |      |
 |        |       |         |          |        |       |      |
VAX      VAX   VAX11780    HSC        VMS      DISK    DISK   TAPE
                  |
                  |
     +------+-----+---------+---------+-----etc.
     |      |     |         |         |
     |      |     |         |         |
   MS780  KA780 DW780   WCS12?.PAT  ADAPTOR
            |
            |
    +-------+------+------+------+------etc.
    |       |      |      |      |
    |       |      |      |      |
  M8237   M8236  M8235  M8234  M8233
```

The VAXcluster revision level R1, included VMS Version 5.4-3, VAX 9000-110, VAX9000-300, VAXft 410, VAXft 610, DEMFA, DEFCN, RA71, RA72, RF30, RF31, RF31F, RF71, RF72, TF857, HSC60, HSC90, KFQSA.
The latest VAXcluster revision level (as of May 1993) is S1, this revision includes VMS Version 5.5, VAX 6000-600,MicroVAX Model's 30, 40, 80, Vax4000-500, RF35, RF73, KFMSA, DECps.

You can either come to Support and borrow a hardcopy of this document to photocopy or you can find it on Pink Microfiche - headed VAXCLUSTER REVISION CONTROL DOCUMENT EP-VXCLS-RM-R1. (latest version on microfiche is S1). There are also pink fiche with detailed revision information on other options which you may find useful.

The notesfile VCSESU::Cluster_rm has revision matrix for individual options.

10.2 CI.

10.2.1 CI node numbering.

There has been much confusion about what numbers CI nodes should have. Any numbering scheme is equally acceptable and if the system manager has any preference then this should always be followed. Where the installation engineer has a choice I would suggest the following.

Number VAX CI nodes upwards starting from 1.
Number HSC CI nodes downwards starting from F.
Only use node 0 when you are running out of node numbers.

Please note this is a suggestion only. Install the cluster in any supportable way that the customer wants it configured.

The advantages of this numbering scheme are that if you have a cluster common system disk then each VAX can boot from a root with the same number as its CI node number, you can add VAX or HSC nodes up to the maximum number without having to renumber existing nodes, if you accidentally boot using a brand new piece of console media you will not boot into some other systems root (root 0 unused).

10.2.2 CI cabling.

All nodes in the VAXcluster should have 4 cables joining them to the Star Coupler. These cables are a transmit and receive pair for the A path and a transmit and receive pair for the B path.

The maximum length of any CI cable is 45 metres. This is therefor the maximum distance any node (HSC or VAX) can be from the star coupler.

The minimum bend radius for CI cables is 4 inches.

Table 10–2: Part numbers for ordering CI cables

70-18541-xx †	1 cable
BNCI-xx †	1 pair cables
BNCIA-xx †	2 pairs cables

† xx = 10, 20, or 45 metres.

Short cables and attenuators may be used for testing a single node without using a star coupler. These enable you to attach the transmit to the receive so that the port can send itself loopback datagrams. You cannot connect a vax directly to an HSC (even if you use the attenuators to prevent transceiver damage) since this would not allow loopback datagrams and the HSC power up diagnostics would fail.

4 x 70-18430-00	(short CI cables, could use ordinary CI cables)
2 x 12-19907-01	(attenuators)
2 x 70-18771-00	(8 node Star Coupler bocx)

10.2.3 Link modules and settings.

The Link module is the part of the CI interface that talks to the CI.
Module switches or backplane jumpers define the CI NODE ID as well as the TICK TIME (also known as DELTA TIME or QUIET SLOT).
Each node has to have a unique ID, but the TICK TIME must be the same on all nodes.
All clusters should be set to 10 TICK, traditionally the FCO used a 6 node limit as justification to replace all L0100 below rev E. Now all new CI adapters are set to 10 tick **CIXCD, HSJ40, etc.** Proactive chaging will help new installations, 10 TICK is seen as being more reliable ,and whats more the FCO is still current.
Clusters with a CIXCD must be set to 10 TICK.
Clusters with a HSJ40 must be set to 10 TICK.

10.2.3.1 L0100

L0100 leds	GREEN Carrier detect RED Internal loop enabled
L0100 switches	Both switch packs must be set the same. All switches left = node 0

Pre rev E L0100s are preset to 7 TICK, and should no longer be used.

L0100 rev E1 has jumper at E177; Pin 11-12 for 7 TICK; Pin 9-10 for 10 TICK.
L0100 rev E2 has a switch at the top centre of the module; OFF for 7 TICK; ON for 10 TICK.

10.2.3.2 L0118

The L0118 is a replacement for the L0100.
The L0118 can be set for 10 TICK or 7 TICK.
The L0118 is a MUST if the cluster is more than 16 nodes, or if there is an HSC60/90.

Switch settings are . . .

L0118 S1	Node address
L0118 S2	Node address
L0118 S3-1 OFF, S3-2 OFF, S3-3 OFF, S3-4 OFF	7 TICK, 16 NODE
L0118 S3-1 OFF, S3-2 OFF, S3-3 OFF, S3-4 ON	10 TICK, 16 NODE
L0118 S3-1 ON, S3-2 OFF, S3-3 OFF, S3-4 OFF	7 TICK, 32 NODE
L0118 S3-1 ON, S3-2 OFF, S3-3 OFF, S3-4 ON	10 TICK, 32 NODE

Note

THE TABLE FOR L0118 SWITCH S3 SETTINGS IN THE HSC POCKET REFERENCE CARD (EK-HSCPK-RC-001) PAGE 14 IS COMPLETELY INCORRECT

10.2.3.3 Caution when adding new CI nodes

Because of the L0100 FCO for 6 node and greater clusters it is important to realise that when adding a new node to an already existing cluster of more than 6 nodes you must set the link module on the new node to the correct number of ticks.

Also if adding a node to an existing cluster which brings the number of nodes up to six you may have to FCO some of the other nodes or at least change all their tick settings.

10.2.4 CI780

The CI780 is used in 11780, 11782, 11785 and 8600 systems.

The CI780 operates on the SBI at TR14 BR4 and has a dedicated backplane. To examine the CI780 registers use addresses 2001C000 to 2001C014 (assuming standard addressing).

Table 10–3: CI780 Module Utilisation

slot 1	L0104	(ISI)	SBI interface
slot 2	L0102	(IDP)	data path
slot 3	L0101	(IPB)	packet buffer
slot 4	empty		
slot 5	L0100/L0118	(ILI)	link interface
slot 6	empty		
CI780 backplane	70-17654		
CI bulkhead assembly	70-18526-8A		

10.2.5 CI750

The CI750 is used in 11750 systems only.

To examine CI750 registers use addresses F3E000 to F3E014. (Assuming standard addressing). If F3E000 responds but F3E900 does not then the CIPA is faulty.

The CI750 consists of a L0009 module in the CPU backplane, a CIPA box in a separate cab and a CIPA bus which joins the two together.

10.2.5.1 L0009.

The CI750 interface to the CMI is by way of the L0009.

- CPU slots 7, 8 and 9 can hold a CI750, RH750, DW750 or DR750.
- The CI750 CMI address jumpers on the L0009 backplane pins are normally out, this selects address F3E000, I/O slot 15.
- All BG jumpers should be installed on the CPU backplane
- See Section 3.5 for information on CMI arbitration levels.

10.2.5.2 CIPA bus.

The CIPA bus consists of a pair of cables. The CI750 CIPA bus is different to the CIBCI CIPA bus, as the connectors at the CPU end are different.

CI750 CIPA bus	70-19602-00	(Complete cable assembly)
CIBCI CIPA bus	17-01027-01	(Complete cable assembly)

On the CI750, P1 connects to the middle connector on the L0009, P2 to the lower connector, both should have the red stripe up.

10.2.5.3 CIPA box.

This box is used on both the CI750 and the CIBCI. Jumpers on the CIPA box are not inserted although they can be seen on one side of their backplane pins on slot A5.

Table 10–4: CIPA (CI750 and CIBCI) Module Utilisation

CIPA slot 3	L0100/L0118	(ILI)	link interface
CIPA slot 4	L0101	(IPB)	packet buffer
CIPA slot 5	L0400	(CDP)	data path
CI backplane	54-15555		
PSU	H7202D		
CI bulkhead cable assembly	70-18526-8A		

10.2.6 CIBCI

The CIBCI may be used in any BI based system. It consists of a two module BI interface (the BICA), a CIPA box identical to that used in the CI750, and a cable assembly joining the two together (CIPA bus).

For information about the CIPA bus see Section 10.2.5.2

For information about the CIPA box see Section 10.2.5.3

The CIBCI on an 8200 is normally at BI node 5. Registers are at addresses 2000A000 upwards, if 2000A000 responds but 2000A100 fails to respond then the CIPA is at fault.

Table 10–5: BICA (CIBCI) Module Utilisation

T1017	"BAC" Adapter control
T1018	"BAD" Adapter data
17-01029-01	3 inch "under the bottom" cables
17-01029-02	3 and 3/4 inch "under the bottom" cables

10.2.7 CIBCA-A, CIBCA-B

The CIBCA is the replacement for the CIBCI. It will allow direct connection from the BI bus to the CI bus, without a CIPA box in between. The CIBCA adapter consists of two separate hardware modules and two intermodule cables.

Table 10–6: CIBCA Module Utilisation

CIBCA-A	CIBCA-B	Description
T1015	T1045	Port controller
T1025	T1046	Link/packet buffer
17-01504-01	17-01504-01	2 inch cable
17-01504-02	17-01504-02	0.7 inch cable
17-01473-01	17-01473-01	Internal CI bulkhead cable assembly

The CIBCA does not use a L0100/L0118 link module, the CI NODE ADDRESS is obtained from jumpers on the T1015 backplane slot, T1045 backplane slot for the CIBCA-B. The node address and its complement must be configured exactly the same. Various other backplane jumpers select other functions of the CIBCA. Refer to CIBCA-A user guide EK-CIBCA-UG-001 for backplane jumper pinning or CIBCA-B user guide EK-CIBCB-UG-001. In fact the jumper locations and meanings are the same for CIBCA-A and CIBCA-B.

The CIBCA BI NODE id is selected by an encapsulated plug on the BI backplane.

The CIBCA-A uses the microcode file CIBCA.BIN must be resident on the system console boot device. The EEPROM Programming and Update utility - EVGDA.EXE is used to update the CIBCA EEPROM.

To dump the contents of the CIBCA registers in event of a failure
(Assuming CIBCA at node 5 on NBIA#0, BI#0)

```
>>>E/P/L 2000A000/next:10
```

10.2.8 CIXCD

Table 10–7: CIXCD Module Utilistation

CIXCD	Part No.	Description
9XXX		
CIXCD-AA	T2080 54-20225-01 17-02894-01	XMI CI adapter Header Card assembly. VAX9000 bulkhead cable assembly
6XXX		
CIXCD-AB	T2080 54-20225-01 17-02894-02	XMI CI adapter Header Card assembly VAX6000 bulkhead cable assembly
76XX		
CIXCD-AC	T2080-YA 54-20225-01 17-*****-**	XMI CI adapter Header Card asssembly VAX7000 bulkhead cable assembly

The CIXCD is single module CI adapter for the XMI backplane, older versions use EEproms, below rev E06. Rev E06 and above use FLASHproms

The CIXCD firmware is CIXCD.BIN and is loaded by a diagnostic, EVGEA.

Care should be taken when upgrading CIXCD's as a rev issue problem exsists around modules above rev E06 (FLASHPROM) and diagnostic EVGEA rev2.1, the microcode loader.
Do not use EVENT FLAG 3 with EVGEA rev 3.1 or rev 4.0, for the same reason as rev 2.1.
The current ver of EVGEA is rev 4.2 shipped with VAXPAX48.

Please reference the 7000 chapter re the updating and testing of CIXCD on the 7000 platform be they **VAX or AXP** based.

10.3 CI microcode

Table 10–8: CI uCode Revisions

Adapter	uCode rev	VAXPAX
CI780.BIN	8	
CIBCA.BIN	8.5	45
CIBCB.BIN	4007.4002	45
CIXCD	70 (decimal)	49

All CI adapters, except the CIBCA-A/B and CIXCD, use the same CI780.BIN microcode file. Part of the CI microcode is in ROM on one of the CI modules. Part of the microcode is in a file called CI780.BIN and is loaded into the CI when it is initialised. All VAXs should now be running revision 8 of CI780.BIN microcode.

The CIBCA-A uses a different microcode file called CIBCA.BIN. All VAXs should now be running revision 80005 of CIBCA.BIN microcode.

The CIBCA-B uses a different microcode file called CIBCB.BIN. But this file is not loaded at every boot like other CI interfaces because it is held in EEPROM on the CIBCA-B. Use EVGDA.EXE to upgrade the EEPROM with a new version.

Like the CIBCA-B the CIXCD firmware is held in EEprom or Flashprom and does not need to be loaded each time a VAX boots. EVGEA is the firmware loader for CIXCD.BIN.

10.3.1 Copying CI microcode files to/from console.

The CI microcode file lives on the console media. It should be copied to and from the console using EXCHANGE. It is essential to copy it using the switch /TRANSFER=BLOCK this says to exchange that the file has a non-standard organisation and it should copy each block exactly as it is without any attempt to format the data. If you copy the file without this switch then the CI will not work.

To copy from console to VMS . . .

```
$ EXCHANGE COPY CSA1:CI780.BIN CI780.BIN /TRANSFER=BLOCK
```

To copy from VMS to console . . .

```
$ EXCHANGE COPY CI780.BIN CSA1: /TRANSFER=BLOCK
```

To check that /TRANSFER=BLOCK was used each time . . .

```
$ EXCHANGE DIR CSA1:CI780.BIN
```

and check for a size of 36 blocks, or

```
$ Run EVGAA
```

which will check the S/W revision level and report 'garbage' values if the microcode has been corrupted.

10.3.2 Checking CI microcode revisions.

To check the revision of CI microcode on any node in a live cluster.

```
$ SHOW CLUSTER /CONTINUOUS
ADD RP_REVISION
```

For revision 8 CI780.BIN microcode this should give 80007. For revision 3 CIBCA.BIN it should show 50003.

To check the revision of a CI780.BIN console microcode file.

```
$ EXCHANGE COPY _CSA1:CI780.BIN CI780.BIN /TRANSFER=BLOCK
$ DUMP/BLOCK=(START:36,END:36) CI780.BIN
```

This will print out a table of numbers. The bottom line will look like one of the following.

```
Rev 5 and earlier
1F6605A1 D0000000 00000000 00000000 00000000 00000000 00000000 00000000
      |
  rev level

Rev 6 and later
1A6600C1 D0001F00 07A1C21F 5F6600A1 F0000000 00000000 00000000 00000000
                     |
                 rev level
```

To check the revision of a CIBCA.BIN or CIBCB.BIN console microcode file.

```
$ EXCHANGE COPY _CSA1:CIBCA.BIN CIBCA.BIN /TRANSFER=BLOCK
$ DUMP/BLOCK=(START:1,END:1) CIBCA.BIN
```

```
The printout will contain 8 columns of 8 digits followed by some text.
                                              EEPROM version
                                                |
                                                |
xxxxxxxx .....6 More..... xxxxxxxx  CIBCA.BIN 0003.'.p....- 000000
xxxxxxxx .....columns.... xxxxxxxx  ........................ 000020
xxxxxxxx .......of....... xxxxxxxx  ........................ 000040
xxxxxxxx .....digits..... xxxxxxxx  ........................ 000060
xxxxxxxx ................ xxxxxxxx  ........................ 000080
xxxxxxxx ................ xxxxxxxx  ........................ 0000A0
xxxxxxxx ................ xxxxxxxx  ........................ 0000C0
xxxxxxxx ................ xxxxxxxx  ........................ 0000E0
xxxxxxxx ................ xxxxxxxx  0005.j.[.....0......... 000100
xxxxxxxx ................ xxxxxxxx  ..|.................... 000120
                                      |
                                      |
                                    Functional version
```

For the above to work on a Nautilus you have to copy the microcode file from the hard disk (CSA3:) to a floppy disk (in CSA1:) using the PRO software. Alternatively, you can dump it to the PRO console as follows:

```
PRODCL> DUMP/LONG/BLOCK:1:1/ASCII/HEX CIBCA.BIN
```

Note that block numbers are now in octal. (IE:use 44 instead of 36 for CI780.BIN)

10.4 CI Diagnostics

- For information on how to boot diagnostics see Section 10.5.
- For information on HSC diagnostics see Section 10.11.
- For information on where to place diagnostics see Section 10.9.

10.4.1 Attaching CI adapters.

Generally the Autosizer (EVSBA) will attach the CI adaptor for you if you are trying to run stand-alone (level 3) diagnostics. It requires the microcode to be loaded and some versions of EVSBA attach the CI780 at the wrong TR level.

11750

```
DS> ATTACH CI750 HUB PAA0 15 4 n
                          15 = i/o slot
                           4 = BR
                           n = node
```

1180/11782/11785

```
DS> ATTACH CI780 HUB PAA0 14 4 n
                          14 = TR
                           4 = BR
                           n = node number
```

8600/8650

```
DS> ATTACH SBIA  HUB SIO    assuming CI is on first SBI
DS> ATTACH CI780 SIO PAA0 14 4 n
                          14 = TR
                           4 = BR
                           n = node number
```

6XXX with a CIBCA-A/B

```
DS> ATTACH DWMBA  HUB DWMBAn m b
                          n = 0 to F
                          m = XMI node number  (XMI slot)
                          b = BI node number   (BI plug)
DS> ATTACH CIBCA DWMBAn PAA0 b x c
                          n = 0 to F
                          b = BI node number   (BI plug)
                          x = BR level
                          c = CI node number
```

6XXX with a CIXCD

```
DS> ATTACH CIXCD  HUB PAA0 m b
                          m = XMI node number  (XMI slot)
                          b = CI node number
```

7XXX
Please ref the 7XXX chapter.

8200/8250/8300/8350

```
DS> ATTACH CIBCx HUB PAA0 5 4 n
                          x = I or A
                          5 = BI node number  (BI plug)
                          4 = BR level
                          n = CI node number  (L0100/L0118 switches)
```

8500/8530/8550/8700/8800

```
DS> ATTACH NBIA HUB NBIAn n
                          n = Adapter number (Physical slot)
                              0 or 1 for 8800/8700 (usually 0)
                              1 for 8500/8530/8550
```

```
DS> ATTACH NBIB NBIAn NBIBx x y
                          n = Adapter number (from above)
                          x = BI number (Physical slot)
                              0 to 3 for 8800/8700
                              2 to 3 for 8500/8530/8550
                          y = BI node number (backplane plug)
```

```
DS> ATTACH CIBCx NBIBy PAA0 5 4 n
                          x = I or A
                          y = BI number (from above)
                          5 = BI node number  (backplane plug)
                          4 = BR level
                          n = CI node number  (L0100/L0118 switches)
```

9XXX with a CIXCD

```
DS> ATTACH XJA HUB XJA0 0 8

DS> ATTACH CIXCD  XJA0 PAA0 m b
                          m = XMI node number  (XMI slot)
                          b = CI node number
```

10.4.2 Level 3 Diagnostics

Be very careful if using pre rev 19 diagnostic tapes. Some of these diagnostics were renamed when rev 19 diagnostic tape came out. This was because they used to be 11780 only and so were called ESxxx but they now run on many CPU types, so they were renamed EVxxx.

See Chapter 12, VAX DIAGNOSTICS for revision information.

ECCGA	CI750 repair level	1		TU58 53
ECCGB	" " "	2		"
ECCGC	" " "	3		"
ECCGD	" " "	4		TU58 54
ECCGE	" " "	5		"
EVCKA	CIBCI repair level	1		
EVCKB	" " "	2		
EVCKC	" " "	3		
EVCKD	" " "	4		
EVCKE	" " "	5		
EVCKF	" " "	6		
EVCGA	CI780 repair level	1		RX63
EVCGB	" " "	2		"
EVCGC	" " "	3		"
EVCGD	" " "	4		"
EVGCA	CIBCA-A Repair Level	1		
EVGCB	" " "	2		
EVGCC	" " "	3		
EVGCD	" " "	4		
EVGCE	" " "	1		
EVGCK	" Repair Level Microcode	1		
EVGCL	" " " "	2		
EVGCM	" " " "	3		
EVGCN	" " " "	4		
EVGDA	CIBCA EEPROM Programming and Update Utility			RX50 FJ11
EVGEE	CIBCA-B Repair Level	1		
EVGEF	" " "	2		
EVGEG	" " "	3		
EVGEA	CIXCD Functional Test			
EVGEB	CIXCD Update and Verify			
EVGAA	CI functional diag	1	RX72	TU58 54
EVGAB	" " "	2	"	" "
EVGAC	" " "			

10.4.3 EVGAA

Diagnostics EVGAA and EVGAB run on ALL CI adapters. They require the CI to be connected to a star coupler or loopback cables to be installed. They also require the CI microcode to be loaded.

To cause EVGAA or EVGAB to load the microcode you must set event flag 1 prior to starting the diagnostic - **MAKE SURE THAT THE MICROCODE IS ON THE DEFAULT LOAD DEVICE BEFORE STARTING THE DIAGNOSTIC.** Note that this flag may get cleared by a run command. On an 8200/8300 you must leave the Diagnostic floppy in, the console floppy is in RT11 format and cannot be read by the CI diagnostics.

Event flag 1 is always cleared by RUNning a diagnostic . The Diagnostic supervisor will automatically ABORT the previous diagnostic for you when you run a new one, and the ABORT command clears **EVENT FLAGS** . So the following will not work,

```
DS> SET EVENT FLAG 1
DS> RUN EVGAA
```

but the following will work.

```
DS> LOAD EVGAA
DS> SET EVENT FLAG 1
DS> START
```

If there is a cluster using the same star coupler then both EVGAA and EVGAB may intermittently fail any/all of the tests. Failures that I have seen include Data Structure Errors, "Unable to remove entries from Receive queue" and others. If you get intermittent failures from these diagnostics then you will have to remove your CI cables from the star coupler and use loopback cables. You will need two attenuators **(12-19907-01)** and either four short CI cables **(70-18430-00)** or the normal four CI cables that were connected to the star coupler.

Always remove the transmit cables first, and the recieve cables last. The reverse is true when reconnecting CI cables, always make sure the recieve cables are connected first and transmit cables last.
The reason is that todays CI adapters are only to willing to broadcast their exsistance, especially if they are deaf to all other nodes on the CI.

You should never connect the transmit and receive cables directly together or connect the transmit cable of one node directly to the receive cable of another node. This can damage the link module. Attenuators or a star coupler reduce the signal strength and prevent damage.

ECCGE, EVCGD, EVCKF and EVGCE can be run with or without loopback cables. None of the other repair level diagnostics need loopbacks. By default the loopback cable is not needed. If you want to test the loopback then . . .

```
DS> START /SECTION = EXTMLOOP       or
DS> START /SECTION = EXTERNALLOOP
```

depending on which diagnostic you are running.

10.4.4 Level 2 (Online) diagnostic.

There is only one CI diagnostic that can run whilst VMS is up. EVXCI comes on a separate tape to the rest of VAXPAX and must be installed with VMSINSTAL. I have never managed to successfully install and run it. If you manage to do so then please contact me with details of how it is done!

10.5 Booting systems via the CI.

All systems that boot via the CI need to initialise the CI BEFORE booting. This initialisation includes loading the CI microcode from the console media, this means that the console floppy/TU58 must be left in until the system has booted. See Section 10.3.1 for how to update the console media.

10.5.1 Register Setups.

When booting any VAX the details of how to boot are passed to VMB.EXE in registers R0 to R5. How you deposit numbers into the registers varies for each CPU type.

11750

```
 >>> B/800 DDA0
BOOT58> D/G 0 20
```

SCORPIO (82xx, 83xx)

```
>>> B/800 CSA1
BOOT58> D/G 0 20
```

1178x, 86x0, NAUTILUS (85xx, 87xx, 88xx)

```
>>> DEPOSIT R0 20
```

Register	Contents †	Meaning
R0	00000020	(CI port device)
R1		(CPU type dependent)
	0000000E	(11780,11782,11785,8600, TR number of CI)
	00F3E000	(11750, CMI address of CI)
	000000xy	(BI systems, x=BI number, y=BI node number)
R2	0000000x	Boot via HSC at CI node number x
	00000x0y	x = CI node number of primary HSC to try
		y = CI node number of secondary HSC to try
		ONLY SPECIFY ONE HSC IF BOOTING DS>
R3	000000nn	boot from drive nn
	80ss00nn	boot from shadow set number ss.
		use drive nn if shadow set not already mounted.
		DO NOT BOOT PRE-REV-27 DS> FROM A SHADOW SET
R4	00000000	not used
R5	r0004000	boot VMS from root r
	r0004010	boot diagnostics from root r
	r0004001	conversational boot VMS from root r

† Don't forget these are all HEX numbers. eg for DUA37, DEPOSIT R3 25.

After setting up the registers you should LOAD VMB.EXE at a start address of 200 and then start it at 200.

10.6 Configurations and Performance

Configuration Rules for CI VAXcluster Systems.

- The maximum number of CPU's that can be connected to a CI is 16.
- A CPU can have multiple CI adapters.
- Different types of CI adapters can not be mixed in the same CPU.
- Up to five Star Couplers can be used in a Vaxcluster system.
- A CPU cannot be attached to two Star Couplers that are not in the same VAXcluster system.
- A Star Coupler cannot be attached to two CPU's that are in different VAXcluster systems.
- Each CI adapter connected to the same Star Coupler must have a unique node number.
- Adapters in the same CPU may have the same node number if they are connected to different Star Couplers.
- Storage devices cannot be dual ported between HSC's that are located on different Star Couplers.

Table 10–9: Maximum no. of CI Adapters per VAXCluster CPU.

CPU	CI750	CI780	CIBCI	CIBCA-A	CIBCA-B	CIXCD
VAX 11-750	1					
VAX 11-780		1				
VAX 11-785		1				
VAX 6000-xxx				1	4	4
VAX 7000-xxx						3
VAX 82xx,83xx			1	1	1	
VAX 85xx,87xx,88xx			1	1	2	
VAX 86xx		2				
VAX 9000-xxx						6

CI Adapter Bandwidths with Differing Packet Sizes.

Table 10–10: CI Adapter Bandwidths

BLOCKS	CI780	CIBCI	CIBCA-A	CIBCA-B	CIXCD
1	.57	.45	.45	.54	2.1
4	1.56	.87	1.04	1.48	5.7
16	2.3	1.0	1.1	2.0	7.5
128	2.8	1.0	1.2	2.6	8.0

Bandwidths are in megabytes per second.

10.6.1 DSSI

Configuration Rules for DSSI VAXcluster Systems.

- As many as three Q-bus systems with at least one VAX 4000 Model 100 or higher system.
- Up to three VAX 6000 systems.
- Any combination (totaling three) of VAX 6000 and VAX 4000 Model 100 or higher systems.
- As many as four VAX 6000/7000/10000 series systems.
- All CPU's connected to the same DSSI segment must be memebers of the same VAXcluster system.
- DSSI VAXcluster CPU's must also be connected by another interconnect for DECnet communications, for example Ethernet. DECnet communications are not supported over the DSSI bus, but are required for VAXcluster operation.
- Each DSSI adapter in a single CPU must be connected to a different DSSI segment.
- The KFQSA Q-bus-to-DSSI adapter cannot be used for VAXcluster communications between CPU's on a DSSI segment. An alternate interconnect, for example Ethernet, must be provided.
- For a more complete list off DSSI configuration rules the stars article **Guide to DSSI VAXcluster Configurations** by Ken Robb and David Weeks in the **V5_VMS** Stars database on TIMA.

Table 10–11: Maximum no. of DSSI Adapters per CPU.

CPU	EDA640	EDA660	EDA670	SWIFT	KFQSA	KFMSA
MicroVAX 3300/3400	1				2	
MicroVAX 3500/3600/3800/3900					2	
MicroVAX 2					2	
VAX 4000-200		1			2	
VAX 4000-300			2		2	
VAX 6000						6
VAX 7000						12
VAXft				4		

Table 10–12: DSSI Adapters performance characteristics.

	EDA640	EDA660	EDA670	SWIFT	KFQSA	KFMSA
I/O's per second	390	800	800	390	190	1600/2
Mbytes per second	1.9	3.1	3.1	1.9	1.5	5.5/2

10.7 Dual Porting.

10.7.1 HSC to HSC.

Any DSA disk drive may be dual-ported between any two HSCs. Both port select buttons may be pushed at the same time. The drive can either be available to both HSCs (no port select lights on, both HSCs can see the drive) or it can be online to ONE HSC in which case the other HSC cannot see it. (As indicated by HSC> SHOW DISK)

If the HSC which has the drive online fails, or the drive develops a fault on one path, or the drive is dismounted, or the lit port button is released then the drive will become unavailable on the original port and available on the other port. Mount verification code within VMS takes advantage of this and - transparently to the user - remounts the drive on the other HSC.

Any two HSCs which share a dual ported disk drive must have the same disk allocation class. This is how VMS knows that two DUAn's are the same drive.

It is quite possible to have 3 or more HSCs which all have the same allocation class with multiple disk drives dual ported between each pair of HSCs.

If two HSCs have the same allocation class then they should not have different drives with the same unit number! The easiest way to implement this is to give all DSA drives in the cluster unique unit numbers.

Until VMS V4.6 dual ported DSA tapes should only be made available on one path. If you attempt to fail a tape over between HSCs then you may get very confusing results - up to and including HSC crashes and tape drives that cannot be accessed on either port!

The HSC parameter TAPE ALLOC is used to set allocation class for tapes in a similar way to DISK ALLOC for disks.

10.7.2 UDA50 and clones.

A disk drive may be dual ported between a UDA50, KDA, KDB or other similar interface and an HSC or another UDA etc. VMS V4 does not support this configuration, only one port button may be pressed. If it is necessary to switch one of these drives to the other port then it must be manually dismounted from all nodes which can see it, switched to the other port and then manually remounted again. This practice is strongly discouraged and may cause large numbers of drive command timeouts to be logged.

From VMS V5 it is supported to dual port a disk drive between two non-HSC controllers. You may have both port buttons pressed and it is possible for the drive to fail over from one system to another (both systems must be serving the disk to the cluster and both systems must have the same non-0 allocation class). It is still not supported to have a drive dual ported between an HSC and a non-HSC controller.

10.7.3 MASSBUS to MASSBUS

Any massbus disk drive - except for a VMS system disk - may be simultaneously on line on both ports. Both VAX's connected to any drive must have the same (non zero) allocation class (sysgen ALLOCLASS parameter). Automatic failover is provided for all nodes in the cluster EXCEPT the two nodes physically connected to the drive. These nodes can only access the drive through the direct massbus link EVEN if that massbus link fails. The dual ported drive MUST HAVE THE SAME NAME ON BOTH SYSTEMS. (e.g. it cannot be DRA1 on one system and DRB1 on the other).

Each system connected to the drive must execute the sysgen MSCP command. This command loads the mscp software which allows the vax to act as an mscp server. They must then $SET DEVICE/SERVED ddan: (where ddan is the device name). Finally the drive can be either mounted separately by each node that wishes, or any node can $MOUNT/CLUSTER ddan: which causes all nodes to mount the drive.

From VMS V5, the recommended way to load the MSCP server and SET DEVICE /SERVED is to use the sysgen parameters MSCP_LOAD and MSCP_SERVE_ALL.

10.8 VMS

Note

Many of the things discussed in this section involve doing things which should only be done by the system manager. Do not take the responsibility of changing any VMS parameter, shutting down a system, mounting a disk etc. on a cluster. It could lead to disaster. This information is here to help you check the state of parameters and see what the customer might be doing wrong that could affect you.

10.8.1 Shutdown Options.

When running shutdown on a node in a vaxcluster there are four options that you can specify, these are REMOVE_NODE, CLUSTER_SHUTDOWN, REBOOT_CHECK and SAVEFEEDBACK (VMS V5 only)

10.8.1.1 Remove_node

The intention of this option is that after this node has shut down the rest of the cluster should re-compute quorum and allow it to fall. (Normally quorum is not allowed to fall, only to go up). This is so that if the node is going to be down for some time the rest of the cluster can stand the failure of another node. As of VMS 4.7 it does not always have the desired effect, due to logic in the software designed to prevent the possibility of a partitioned cluster. It is better to shut down the node to be removed and then type

```
$SET CLUSTER/QUORUM
```

on one of the remaining nodes. If you have lost quorum due to a node shutting down, **AND YOU KNOW THAT THERE IS NO POSSIBILITY OF THE CLUSTER BEING PARTITIONED,** then you may be able to use the technique described in Section 10.8.2 to attempt to recover.

10.8.1.2 Cluster_shutdown

The intention of this option is to allow all nodes to shut down at the same time, without one of them being left without quorum so that it cannot shut down. **IT DOES NOT CAUSE ALL NODES TO SHUTDOWN.** It causes the node that is being shutdown to hang just before it leaves the cluster and to wait for all other nodes to reach the same state, at which point all nodes will shut down together.

10.8.1.3 Reboot_check

The intention of this option is to allow you to check that essential files that are needed to boot successfully do really exist, BEFORE the node shuts down. It is a good idea to ALWAYS specify this option, in addition to any other options you want.

10.8.1.4 Savefeedback

This option is used to save information about the running system for use by Autogen in resetting sysgen parameters.

10.8.2 Hung Clusters.

10.8.2.2 Loss of Quorum

If the cluster has hung due to a loss of quorum it is sometimes possible to cause the cluster to re-compute quorum and continue running.

IF the problem is lack of quorum then the console terminals will show

```
%CNXMAN, quorum lost - blocking activity.
```

If you know why quorum has been lost, **AND YOU ARE QUITE SURE THAT THERE IS NO POSSIBILITY OF THE CLUSTER BEING PARTITIONED** then you may be able to cause the node to continue by the following process. `Set the console system to local enable`

```
^P
>>>  H                  (Halt CPU)
>>>  D/I 14 C           (Request a software interrupt at IPL C
>>>  C                  (Continue CPU)
IPC> Q                  (Recompute quorum allowing it to fall)
IPC> EXIT               (And back to VMS...   )
```

If you ever feel that this may be necessary you should explain to the system manager about the dangers of a partitioned cluster and allow them to take the responsibility of deciding whether to continue.

10.8.2.2 Gradual loss of activity.

If over a small period of time an element, a cpu, or the whole cluster system seems to lockup (sounds familiar), it is quite possible that the problem is due to lack of resources. Usually software resources, but can be inspired by hardware faults, or by software features, or parameters wrongly set.

To get a handle on a loss of resource problem, entails getting aview of the problem from every individual node in the cluster. If the degradation of system is slight, the CSC may be able to diagnose on the live system. However once degradation occurs it usually starts a spiral down to total non-userability, in this case you have to generate a crashdump file on all nodes in the cluster.

By now most VAXes in the Cluster are not responding, even to the console. How to gather the information needed.

1. **HALT** all vaxes with ^P and or HALT.
2. Set all vaxes to Power On/Reboot **HALT**
3. Individually on each VAX
   - Force a crash dump to be taken. ie **@CRASH** etc.
   - Allow dump to finnish writing to file.
   - Boot vax in standalone mode.
   - Write SYSDUMP.DMP file to tape.
   - Shutdown vax.
   - **NEXT VAX**
4. Once all the data is gathered, the Cluster can then be rebooted as per normal. Depending on the nature of the problem depends on how long before the system starts to degrade again.
5. Get dumpfiles analysed.

All of the above takes a great deal of time and customers will be reluctant to undertake the data gathering, however this is usually the only handle on the problem, and if persistant will have to be done in the end.

A new system management tool **DECamds**, Availability Manager for Distributed Systems has been especially developed to be of use in these situations. DECamds working via private protocols (from a host, VCS?) at elevated IPL, allows the RMDRIVER to be used in :-

Crisis Intervention

- Unhanging of nodes and clusters.
- Suggesting further investigations and fixes.
- Logs all known problems.

System Resource Watchdog

- Real time system resource monitoring.

- User defined period of monitoring.
- Logs system status changes.
- Can be customized to focus on set resorces.

DECamds would be very useful in the above scenario, by possibly highlighting the problem, and the means to effect an action when all normal user interfaces are not responding.

Customers have to purchase DECamds, part no. QL-GW3A*-AA.

10.8.3 Show Cluster Utility.

Information about a running cluster can be obtained by typing $ SHOW CLUSTER/CONTINUOUS on any system in the cluster. The help file is extensive and further information about this utility is in VMS V4.4 reference manual 5B (System Management) or VMS V5.0 System Management Volume 4 (Performance).

Amongst other things this utility is the easiest way to find the hardware and software revisions of your CI microcode (ADD RP_REVISION, see Section 10.3.2, the states of CI cables (ADD CABLE_STATUS, should be A-B for all ports), the votes held by each node and the current quorum. Use ADD HW_VERS to display the SID for VAXes and hardware configuration for HSC's . . .

```
$ SHOW CLUSTER /CONTINUOUS
Command > add rp_rev
Command > add hw_vers

View of Cluster from system ID 42086  node: WELMTS
+----------------------------------------------+----------+-----------+
|                   SYSTEMS                    | MEMBERS  | CIRCUITS  |
+--------+--------------------------+----------+----------+-----------+
|  NODE  |         HW_VERS          | SOFTWARE |  STATUS  | RP_REVIS  |
+--------+--------------------------+----------+----------+-----------+
| WELMTS | 000000000400FFE206FF532F | VMS V4.7 | MEMBER   |     70007 |
| HSC015 | 0000001B3282282282282282 | HSC V350 |          |       22B |
| WELCHS | 000000000000000005902F14 | VMS V4.7 | MEMBER   |     70007 |
| WELCAS | 000000000000000005202B14 | VMS V4.7 | MEMBER   |     30002 |
| HSC014 | 0000001B3282282282282282 | HSC V350 |          |       22B |
+--------+--------------------------+----------+----------+-----------+
                                  \ /
                    Requestor 2,  282 = K.SDI, MC rev 40
```

For a VAX the HW_VERS is the 32 or 64 bit system ID (depending on CPU type). For an HSC the HW_VERS should be broken down into 3 nibble groups. Each of these groups shows the type and revision of one requestor. The right hand three nibbles show REQUESTOR 2, the left hand 3 nibbles are for REQUESTOR 9. xx2 is for a K.SDI, xx3 is for a K.STI, so in the example above, both HSCs have the following configuration . . .

Req Number	Type	Microcode Version
2	K.SDI	28(hex), 40(dec)
3	K.SDI	28(hex), 40(dec)
4	K.SDI	28(hex), 40(dec)
5	K.SDI	28(hex), 40(dec)
6	K.SDI	28(hex), 40(dec)
7	K.STI	1B(hex), 27(dec)

Those of you who are alert will by now have noticed that these HSC's have the K.STI modules at the HIGHEST requestor numbers, they should of course be at the LOWEST requestor numbers, I'll tell the AR when I see him!

It is worth playing with this tool sometime when you don't have problems to see what you can display.

10.8.4 Shadow Sets

10.8.4.1 Controller Based Volume Shadowing

CBVS or Phase 1 volume shadowing is the traditional HSC Based volume shadowing. SCSI disks on a K.SCSI will not be supported in CBVS.

10.8.4.2 Host Based Volume Shadowing

HBVS or Phase 2 volume shadowing, extends shadowing to DSA disks of the same physical geometry, on a single system or located anywhere in a VAXcluster system.

Distributed , not centralized, shadowing. HBVS maintains virtual units in a distributed fashion on each node in the cluster, (or on a single system). HBVS uses the following system interconnects.

- CI - Computer Interconnect
- NI - Ethernet
- DSSI - Digital Small System Interconnect
- MI - A combination of one of the above.

HBVS provides for shadowing caperbilites across different controllers.

Devices that cannot be shadowed include:

- SCSI devices on the VAX3100 and the VAX3150.
- MicroVAX 2000 RD disks.
- Older disk devices such as:
    - Massbus, RM03/05, RP06/07
    - RK07
    - RL02

10.8.4.3 Volume Shadowing SYSGEN Parameters

SHADOWING - 0 to 3

- 0 - No shadowing is enabled; DUDRIVER is used.
- 1 - Phase 1 shadowing only; DSDRIVER is used.
- 2 - Phase 2 shadowing only; SHDRIVER and DUDRIVER are used.
- 3 - Both phase 1 and 2 shadowing are enabled; SHDRIVER is used with DSDRIVER during migrations from phase 1 to phase 2 volume shadowing.

SHADOW_SYS_DISK - 0 or 1

Used by Phase 2 shadowing only. When 0, the system disk is not a member of a shadow set. When set to a 1, the system disk is a member of a phase 2 shadow set.

SHADOW_SYS_UNIT - 0 or 9999

Used by Phase 2 shadowing only. Is an integer value that is the virtual unit number of the system disk.

SHADOW_MAX_COPY - 1 to 42

Used by Phase 2 shadowing only. The value controls how many parallel copy threads are allowed or a given node. The default is 4 but even this can be too many, for say a 6300 and 2 KDB50's.

10.8.4.4 HBVS Performance part 1

HBVS performance based on a two member shadow set residing on an HSC, as opposed to CBVS.

- Under steady state conditions.
  - Read is equal to or better than phase 1.
  - Write is the same as Phase 1.
- During a MERGE copy... When a Vax crashes.
  - Read is significantly better then phase 1.
  - Write is equal to or better then phase 1.
  - Merge copies can take a very long time. Days have been known, please check with CSC for latest patch kit.
- During FULL copy...
  - Read is equal to or better than Phase 1.
  - Write is equal to or better than Phase 1.
  - Full copy (MOUNT) takes about as twice as long as PHASE 1.

10.8.4.5 HBVS Performance part 2

To help speed up Full and Merge copies, the **CONTROLLER ASSIST** functionality has been introduced on some disk controllers.

- Full Copy Assist.
  - All HSC's except HSC50.
  - Needs Cronic v6.5
  - All shadow set members must be on the same HSC at copy time.
  - Copy speed simlar to Phase 1.
- Merge Assist
  - HSC needs Cronic V6.5
  - RF73,RF35, and later (not RF72 and earlier.)
  - Merges takes less then 60 seconds.
  - Not used on system disks.

The increase in performance in Merge copies is by the controller employing a technique called **Write Logging**, where the controller keeps a log of all writes to its disks that are members of a HBVS shadow set .

10.8.4.6 Volume Shadowing General Considerations

- The physical configuration rules vary between Phase 1 (CONTROLLER based) and Phase 2 (HOST based).
- A shadow set is a VIRTUAL disk.
- Each shadow set has one or more identical PHYSICAL disk drives which are its "members".
- There is no "master" member of a shadow set, all members are equal.
- Every write goes to all members of the shadow set and does not finish till they have all been successfully written
- Each read comes from only one member, the HSC optimises shadow set reads to get the data as fast as possible.
- If there is an error whilst using a shadow set the error log entry will be logged against the PHYSICAL drive, shadow sets never log errors against the VIRTUAL drives.
- The VIRTUAL disk is mounted, the PHYSICAL disks are not mounted, they are not available for mounting because they are members of the shadow set.
- If you DISMOUNT one of the physical drives this does not close any open files, it just removes that drive from the shadow set.
- To use shadow sets the shadowing licence must be installed and SYSGEN> parameter SHADOWING must have a non-zero value. (On VMS V4.x this parameter was called VMS7). Also see SHADOW_SYS_DISK and SHADOW_SYS_UNIT.
- A Quorum disk cannot be shadowed, as system processes which read and write the QUORUM.DAT file, do so without using VMS Distributed Lock Manager synchronization.
- All changes in the composition of a shadow set are done by the mount verification code and are logged on the Console terminal.
- If you have problems with shadow sets then examine the following.
  1. Console printouts from all systems
  2. Console printouts from all HSCs
  3. Errorlogs from all systems
  4. Operator logs from all systems where console printout not available. (Look for mount verification messages).
- If you repair a failed shadow set member and return it to the customer, they must mount it back into the shadow set again before it will be used.
- Forced Errors are fixed by fetching the data from another member.

10.8.5 UETP

UETP can be very useful for testing devices for which you do not have diagnostics. If you want to use UETP to test a single device then look in the DSA chapter of DATADOC for instructions.

VMS has two different UETP accounts. One of these is the standard SYSTEST account which you log in to for running UETP. The other is called SYSTEST_CLIG and is used by UETP on a remote node during the cluster phase of UETP to test served MSCP disks, lock management and other clusterwide software. Before running UETP you should make sure that this account is usable (most system managers set the DISUSER flag on this account. You need to . . .

```
UAF> MODIFY SYSTEST_CLIG /FLAG=NODISUSER
```

before running UETP and then

```
UAF> MODIFY SYSTEST_CLIG /FLAG=DISUSER
```

afterwards.

If you get an error whilst starting UETP that tells you the SYSTEST account has the wrong privileges or quotas then you should look in the VMS V4.4 Guide to VAX/VMS Software Installation, page 7-21. Chapter 7 of this guide has lots of other useful information about UETP. From VMS V5.0 this information is in the CPU type specific Installations and Operations manual in a chapter titled "Running UETP".

10.8.6 Sysgen Parameters.

There are a lot of sysgen parameters relating to the CI or CLUSTER. The following parameters are the most important. If any of these is not correct then the system should be booted with a conversational boot and they should be modified before continuing. Failure to do this could result in having to reboot the whole cluster to put the problem right again.

Sysgen parameters should normally only be altered by editing MODPARAMS.DAT to define the new value and then using autogen. **ASK the system manager to do this if you think it is necessary. Do not do so yourself.**

To check whether AUTOGEN has been run on a system type out the file called SYS$SYSTEM:PARAMS.DAT look at the values for SID (should equal your SID register), Version (should be the version of VMS still running), and MEMSIZE (should be the same as the number of pages from a $ SHOW MEMORY).

VAXCLUSTER †

Controls whether the system should join or form a vaxcluster. This parameter accepts the following three values:

1. Specifies that the system will not participate in a vaxcluster.
2. Specifies that the system should participate in a vaxcluster only if hardware supporting SCS is present (CI, UDA, HSC50).
3. Specifies that the system should participate in a vaxcluster

Should always be 2 on systems intended to run in a vaxcluster.

SCSNODE

Specifies the SCS system name. This parameter is not dynamic. You should use a name that is the same as the decnet node name (limited to six characters, must be LETTERS and NUMBERS only). The name must be unique among all systems in the cluster.

Once a VAX (or HSC) node has been recognised by another node in the cluster, you cannot change the SCSSYSTEMID or SCSNODE parameter without changing both.

SCSSYSTEMID

Specifies the lower-order 32 bits of the 48-bit system identification number. This parameter is not dynamic and must be the same as the decnet node number. It must be different on every node. DECNET node number is normally written as AAA.NNN This corresponds to an SCSSYSTEMID of (1024 x AAA) + NNN.

SCSSYSTEMIDH

The high-order 16 bits of the 48 bit system identification number.

†These parameters should normally have the same value on ALL nodes in the cluster.

VOTES

Specifies the number of votes towards a quorum to be contributed by the node. The default value is 1.

QUORUM †

This parameter is used by VMS V4 only. It specifies an initial setting for the dynamic quorum value. This setting is a numeric value that is an estimate of the correct quorum value to be used and should be greater than half of the total expected votes. The default value is 1. To prevent cluster partitions should be set to (TOTAL NUMBER OF VOTES IN CLUSTER + 1) / 2

EXPECTEDVOTES †

Specifies a setting taht is used to derive the initial quorum value. This setting is the sum of all the VOTES held by potential cluster members.
By defualt the value is 1. The connection manager sets a quorum value to a number that will prevent cluster partitioning. To calculate quorum, the system uses the following formula:
Estimated Quorum = (EXPECTED VOTES + 2) / 2

DISK_QUORUM †

The name, in ASCII, of the quorum disk. ASCII spaces indicate that no quorum disk is being used. Must not be a shadowed disk, must be a PHYSICAL DEVICE NAME, not a logical name.

QDSKVOTES †

Specifies the number of votes contributed to the cluster votes total by a quorum disk. The maximum is 127, the minimum is 0, and the default is 1.

NISCS_CONV_BOOT

This is a new parameter for VMS Version 5. Specifies whether a satellite node should be allowed to perform a conversational boot. This is intended to increase security in NI Clusters and Mixed interconnect clusters where a workstation may have it's console terminal in a physically insecure place.

NISCS_LOAD_PEA0 †

This is a new parameter for VMS Version 5. Causes the PEDRIVER to be loaded. Must be set to 1 if on all nodes if ANY node is using the ethernet instead of the CI for VAXcluster communications.

NISCS_PORT_SERV

This is a new parameter for VMS Version 5. Any node using a DEQNA should set this parameter to a value of 2. This causes all VAXcluster packets sent on the ethernet to this node to be checksummed.

MSCP_LOAD

This is a new parameter for VMS Version 5. This causes the MSCP server to be loaded very early in the boot sequence. It is a replacement for the V4 `SYSGEN> MSCP` command.

MSCP_SERVE_ALL

Specifies MSCP disk-serving functions when the MSCP server is loaded. The default value 0 specifies that no disk is served. A value of 1 specifies that all available disks are served. A value of 2 specifies that only locally connected (non-HSC) disks are served. It is a replacement for the V4 `$ SET DEVICE /SERVED Ddan:`

†These parameters should normally have the same value on ALL nodes in the cluster.

ALLOCLASS

Specifies a numeric value to be assigned as the allocation class for the node. If the system has no dual ported drives it should be 0. Otherwise it must be the same.on all systems sharing dual ported drives, and different on all systems not sharing dual ported drives. If you get this one wrong it could cause data corruption.

PANOPOLL

Disables polling if set to 1 (the default is 0). Disabling polling enables you to boot a system from a private system disk and isolate it from CI activity. You may want to do this following repairs to verify that the system runs properly before introducing it into the hardware cluster. Never set PANOPOLL to 1 while a system is participating in a cluster, if a system is being booted from an HSC, or if it is being booted in order to join a cluster.

PASANITY

Controls whether the port sanity timer is enabled to permit remote systems to detect a system that has been halted or hung at IPL 7 or above for 99 seconds. This parameter is normally set to 1 and should only be set to 0 when debugging with XDELTA. (Used for debugging Device Drivers and other high Priority code) or whilst troubleshooting. PASANITY is a dynamic parameter (altered the next time the port is initialised) and has a default value of 1. If it is set to 0 and you halt the VAX then you must INITIALISE and UNJAM to stop the adaptor responding to incoming CI traffic.

RECNXINTERVAL †

Specifies in seconds, the interval during which the connection manager attempts to reconnect a broken connection to another VMS system. If a new connection cannot be established during this period, the connection is declared irrevocably broken, and either this system or the other must leave the cluster. This parameter trades faster response to certain types of system failures against the ability to survive transient faults of increasing duration. This may need to be set to higher vales to allow certain CPU types to complete a warm restart. (e.g. 300 seconds if you have an 8800 with battery backup in the cluster).

PASTIMOUT †

Specifies the basic interval at which the CI port driver wakes up to perform time-based bookkeeping operations. It is also the period after which a start handshake datagram is assumed to have timed out. On VMS V4.4 and later it is also used to set the timeout period for the CI sanity timer and virtual circuit timeout timer. See Section 10.10.1.2, CI port Timeout, Virtual Circuit Timeout.

PAPOLLINTERVAL †

Specifies in seconds, the polling interval the Computer Interconnect (CI) port driver uses to poll for a newly booted system, a broken port-to-port virtual circuit, or a failed remote node. PAPOLLINTERVAL is a dynamic parameter with a minimum value of 1, a maximum value of 32767, and a default value of 15. On VMS V4.5 and later it is also used to set the timeout period for the CI port timeout. See Section 10.10.1.2, CI port Timeout, Virtual Circuit Timeout.

TIMVCFAIL

Specifies the time required for an adapter or virtual circuit failure to be detected. Digital recommends that the default value be used. Digital also recommends that this value be lowered in VAXclusters of five CPU's or less, and that a dedicated NI segment be used for cluster I/O.

†These parameters should normally have the same value on ALL nodes in the cluster.

You should normally use AUTOGEN to set sysgen parameters. Any parameters that you want to set to specific values should be placed in a file called MODPARAMS.DAT in the SYS$SYSTEM directory, before running AUTOGEN. These should normally include **VAXCLUSTER, SCSNODE, SCSSYSTEMID, VOTES, EXPECTEDVOTES, NISCS_LOAD_PEA0, NISCS_CONV_BOOT, MSCP_LOAD, MSCP_SERVE_ALL, ALLOCLASS, DISK_QUORUM, QDSKVOTES, SHADOWING (VMS7)**

In addition the following may be defined in MODPARAMS.DAT (these are not sysgen parameters but they are used by autogen for calculating some sysgen parameters).

NUMHOST	The number of VAX/VMS systems that are members of the cluster.
NUMSERVER	The number of disk or tape servers seen by this system that use the MSCP protocol (HSC50s, VAXes running MSCP software, UDA50's).

10.8.7 VMS Ver 5.5 SYSGEN Parameter changes

10.8.7.1 Changes to CI Port Sanity Timeout

A CI adapter considers the node its attached to, to be **dead** if that node fails to interact with that node for an **excessive** period. An adapter whose sanity timer has expired enters the disabled state. The adapter does not respond to any type of packet except IDreq. The CI port driver periodically resets this timer. Pre VMS ver 5.5

PASANITY = max(PASTIMOUT,2xPAPOLLINTERVAL)

- PASTIMOUT=PA Sanity Time Out.
- PAdriver polls every PAPOLLINTERVAL for the exsistance of remote ports.

VMS ver 5.5 and later

PASANITY = max(1sec,2/3 TIMVCFAIL)

- TIMVCFAIL (defaults = 1600, in units of 10mS) equates to 16 seconds.
- PASTIMOUT is still present but no longer serves any purpose.

10.8.7.2 Changes to Virtual Circuit Check

PAdriver used VC_Check, to check out virtual circuits in periods of no activity, in pre ver 5.5 VMS, this interval was calculated the same way as PASanity. In Post ver 5.5 :

- VC_Check = 1/3 TIMVCFAIL

This functionality allows for only one port failure to be logged versus many Virtual Circuit Failures reported all around a cluster.

10.9 Cluster common system disk.

There are lots of possibilities for confusion here. If you don't know what you are doing then ASK someone who does. It's very easy to get it wrong.

From V5.0 of VMS all system disks have the cluster common disk format.

On a normal VMS V4 system disk there is only one directory pointed to by the logical names SYS$MANAGER, SYS$SYSTEM etc. On a cluster common system disk there are TWO directories for each of these logical names.

One of these directories is a COMMON directory, this is where files that are shared by the whole cluster are put. The other directory is a SPECIFIC directory, this is for files that are only used by one node.

Whenever you create a new file it is put in the SPECIFIC directory, unless you force it to go to the COMMON directory. When you try to read, run or edit an existing file the system first looks in the SPECIFIC directory and if the file cannot be found there it looks in the COMMON directory.

For example, consider a cluster with 3 systems, using SYS1, SYS2, and SYS3 for their SPECIFIC roots. The following SYS$MANAGER directories exist.

```
[SYS1.SYSMGR]
[SYS2.SYSMGR]
[SYS3.SYSMGR]
[V4COMMON.SYSMGR] also known as [SYS1.SYSCOMMON.SYSMGR] etc.
```

If you log in to SYS2 and create a new file called SYS$MANAGER:NOTICE.TXT, it will go into [SYS2.SYSMGR]. You could now log into SYS1 and type

```
$ DIRECTORY SYS$MANAGER:NOTICE.TXT
```

and you would not see the file. If you want to see the file from all systems then you must

```
$ RENAME [SYS2.SYSMGR]NOTICE,TXT SYS$COMMON:[SYSMGR]NOTICE.TXT
```

Now the same directory command from SYS1 would find the file.

The following logical names are created at boot time to help you navigate round the cluster common system disk.

SYS$COMMON   points to [V4COMMON], i.e. [SYSn.SYSCOMMON]

SYS$SPECIFIC   points to the SPECIFIC root for your node, i.e. [SYSn]

SYS$SYSROOT   points to sys$specific first, then sys$common

SYS$MANAGER, SYS$SYSTEM, SYS$MAINTENANCE point to SYS$SYSROOT:[SYSMGR] etc.

```
           += user directories
           +=   "        "
           +=   "        "
           |
           |
           +-badblk.sys          .
           |                     .  plus various other
           +-indexf.sys          .  required directories
           |                     .
           +-bitmap.sys          .
           |
           |
           |
           +-sysexe.dir           These are here for compatibility
           |                      with earlier versions of VMB.EXE
           +-sysmaint.dir  <..........................................
           |                                                         .
           |                                                         .
           |                               +---sysmaint.dir      <....
           |                               |                         .
           |              +-syscommon.dir+-+---sysmgr.dir      <...  .
           |              |              . |                      .  .
000000.dir-+              +-sysmaint.dir . +---sysexe.dir    <..  .  .
           |              |              .                     .  .  .
           +-sys0.dir---+-sysmgr.dir     .                     .  .  .
           |              |              .                     .  .  .
           +-sysn.dir     +-sysexe.dir   .                     .  .  .
           |                             .                     .  .  .
           |                             .                     .  .  .
           +-v4common.dir--+ <............                     .  .  .
                           |                                   .  .  .
                           +--sysexe.dir       <................  .  .
                           |                                      .  .
                           +---sysmgr.dir      <...................  .
                           |                                         .
                           +---sysmaint.dir    <......................
```

Directories joined by dotted lines in the diagram are the SAME directory. So if (for example) you delete [SYS0.SYSCOMMON.SYSEXE]SYS.EXE;* you will find that you no longer have [SYS1.SYSCOMMON.SYSEXE]SYS.EXE or [V4COMMON.SYSEXE]SYS.EXE

The only system directories shown in the diagram are SYSEXE, SYSMGR, and SYSMAINT. There are others but they are treated the same as SYSMGR.

10.9.1 Where to put the Diagnostics?

On a cluster common disk you should put all common files in the SYS$COMMON directory. So ALL diagnostics should be in SYS$COMMON:[SYSMAINT] and there should be NO diagnostics in SYS$SPECIFIC:[SYSMAINT].

All node specific files should go into SYS$SPECIFIC, so you should put any node specific LOGIN.COM procedures and a node specific CONFIG.COM into SYS$SPECIFIC:[SYSMAINT]

The diagnostic supervisor does not understand search lists. This means that if you put all the diagnostics into SYS$COMMON:[SYSMAINT] and type the following.

```
$ SET DEFAULT SYS$MAINTENANCE
$ RUN ESSAA
DS> DIR
```

You will get a file not found error message. The workaround to this is to put a CONFIG.COM into SYS$SPECIFIC:[SYSMAINT] this config.com should have the following lines at the end.

```
DS> SET LOAD SYS$SYSDEVICE:[SYSMAINT]
DS> EXIT
```

The one drawback of this method is that if you boot standalone diagnostics from this disk you will not be able to find the CONFIG.COM files very easily. SO create a CONFIG.COM in SYS$COMMON:[SYSMAINT] which is a text file saying where all of the config.com files are, e.g.

```
DS> ! SYS$COMMON:[SYSMAINT]CONFIG.COM
DS> !
DS> SET FLAG VERIFY
DS> !
DS> !
DS> ! This disk is a Cluster common system disk for cluster CLUKxxxx
DS> !
DS> ! CONFIG.COM files can be found as follows.
DS> !
DS> ! for 11750 850012345, CI node 1, [SYS1.SYSEXE]CONFIG.COM
DS> ! for 8800  870099012, CI node 2, [SYS2.SYSEXE]CONFIG.COM
DS> ! for 11780 870041256, CI node 3, [SYS3.SYSEXE]CONFIG.COM
DS> !
DS> CLEAR FLAG VERIFY
DS> EXIT
```

10.10 Error Messages.

This section covers error messages that you may find on the CONSOLE, some of them may also be found in the Operator Log file (SYS$MANAGER:OPERATOR.LOG), the errorlog file (SYS$ERRORLOG:ERRLOG.*) or the crash dump file (SYS$SYSTEM:SYSDUMP.DMP).

10.10.1 %PAA0 errors

These error messages are generated by the CI port driver. When possible the PADRIVER will also generate an entry in the errorlog. There are a lot of different possible errors of this type, many of these are self explanatory and I will not include them here, some of them are very rare and those I have also not included.

All %PAA0 errors are listed in the VMS VAXCLUSTER MANUAL (in volume 1B of the VMS V5 System Managers manual set).

10.10.1.1 Port Error Bit(s) Set - CNF/PMC/PSR

%PAA0, Port Error Bit(s) Set - CNF/PMC/PSR xxxxxxxx/xxxxxxxx/xxxxxxxx

The device driver has found an error status in the CI adaptor registers. Only three registers are printed out, the CONFIGURATION register, the PORT MAINTENANCE and CONTROL register and the PORT STATUS register. I have included here the meaning of the bits in those registers. If the PSR register finishes with a value of 40 you have a Maintenance Timer Expiration (also known as Sanity Timer Expiration), this is the most common cause of this error that I have seen. See Section 10.10.1.1.1.

Some of these bits are only used on some CPU types, for example a Multiple Transmitter Fault will only occur on a CI780. Bits marked "May be 0" should be ignored. There are two tables for the CNF register, the first is for the CI750 and CI780, the second is for the CIBCI.

Table 10–13: CI750/CI780 CNF register

Bit 31	SBI parity Fault
Bit 30	SBI Write sequence fault
Bit 29	SBI Unexpected Read Data Fault
Bit 28	May be 0
Bit 27	SBI Multiple Transmitter Fault
Bit 26	SBI Transmitter during fault
Bit 25	May be 0
Bit 24	May be 0
Bit 23	CI750/CI780 Port is Powering Down (set by ACLO or microcode, cleared by power up or microcode)
Bit 22	CI750/CI780 Port is powered up (set by ACLO going, cleared by microcode or bit 23 getting set)
Bit 21	May be 0
Bit 20	CMI nonexistant memory SBI C/A timeout
Bit 19	CMI Read lock timeout, fixed by PCS rev 99 microcode SBI Read data timeout
Bit 18	SBI Error confirmation received
Bit 17	CI750/CI780 Read data substitute error (Hard memory error)
Bit 16	CI750/CI780 Corrected Read Data (Soft memory error)
Bit 15	CI750 CIPA parity error
Bit 14	CI750/CI780 Diagnose (allows access to diagnostic only registers)
Bit 13	CI750 CIPA timeout
Bit 12	CI750 No CIPA (Bit 12 clear = CIPA present, powered up and initialised)
Bit 11	May be 0
Bit 10	CI750/CI780 Transmit ACLO (this Adapter will cause a CPU ACLO)
Bit 9	CI750/CI780 Transmit DCLO (this Adapter will cause a CPU DCLO)
Bit 8	Powerfail Disable (prevents bits 9 and 10 from functioning)
Bits 7 - 0	CI750/CI780 Adapter Code = 38 hex

Table 10–14: CIBCI CNF register

Bit 31	CIBCI CIPA parity error
Bit 30	CIBCI BIC adapter parity error
Bit 29	CIBCI VAXBI Parity Error
Bit 28	CIBCI VAXBI Busy Error
Bit 27	CIBCI power down (see CI750/CI780 CNF bit 23)
Bit 26	CIBCI power up (see CI750/CI780 CNF bit 22)
Bit 25	CIBCI Diagnose. (Allow access to internal diagnostic registers)
Bit 24	CIBCI CIPA DCLO (CIPA disconnected or powered off)
Bit 23	CIBCI CIPA not connected
Bit 22	CIBCI Maintenance Go (Used by diagnostics)
Bit 21	May be 0
Bit 20	May be 0
Bits 19 - 16	CIBCI received BI command
Bits 15 - 0	CIBCI Read only contents of BCMR

Table 10–15: PMC register

Bit 31	CIBCI Parity Error (one of bits 14 to 8 is set)
Bits 30 - 15	May be 0
Bit 15	CIBCA/CI750/CI780 Parity Error (one of bits 14 to 8 is set)
Bit 14	Control Store Parity Error
Bit 13	Local Store Parity Error (if less than 1 per week then ignore!!)
Bit 12	Receive Buffer Parity Error
Bit 11	Transmit multiple parity error
Bit 10	CIPA Bus parity error
Bit 9	CI780 SBI Output parity Error (datapath to SBI module) CI750/CIBCI Output parity error (Error is in CIPA box) CIBCA II Parity Error (Parity error on II bus READ)
Bit 8	CI750/CI780/CIBCI Transmit buffer parity error CIBCA BI bus timeout
Bit 7	CI750/CI780/CIBCI Uninitialised state, the microcode needs to be loaded! CIBCA Halt Sequencer, causes the microcode to stop/start
Bit 6	CI750/CI780/CIBCI Programmable starting address, starts the microcode
Bit 5	CIBCA Maintenance Interrupt Enable (for Diagnostic use)
Bit 4	Wrong Parity (for Diagnostic use)
Bit 3	CI750/CI780/CIBCI Maintenance Interrupt Flag (for Diagnostic use)
Bit 2	CI750/CI780/CIBCI Maint inter enable (see bit 5)
Bit 1	Maintenance Timer Disable
Bit 0	Maintenance Initialise /Start (clears all errors, leaves port in initialised state)

Table 10–16: PSR register

Bits 31 - 16	May be 0
Bit 15	CIBCA No response error (summary of bits 1-8 and 10)
Bit 14	to Bit 12 May be 0
Bit 11	May be 0
Bit 10	CIBCA Uninitialised (see PMC register bit 7)
Bit 9	CIBCA Maintenance Interrupt Flag (see PMC register bit 3)
Bit 8	CIBCA Maintenance Error (any parity error in PSR)
Bit 7	Miscellaneous Error
Bit 6	Sanity Timer/Maintenance Timer expired. This is virtually never the result of a fault in the CI adapter, it means that the CI believes the VAX is hung. Could be VAX H/W or S/W. See Section 10.10.1.1.1.
Bit 5	Memory System Error (e.g. RDS or NXM). More info in CNF or Bus error register and Port Failing Address Register.
Bit 4	Data Structure Error (could be H/W or S/W problem) More info in Port Error Status Register.
Bit 3	Port Initialisation Complete
Bit 2	Port Disable Complete
Bit 1	Message Free Queue Empty (could be H/W or S/W problem)
Bit 0	Response Queue Available (not an error, info for PADRIVER)

10.10.1.1.1 Maintenance/Sanity Timer Expiration

%PAA0, Port Error Bit(s) Set - CNF/PMC/PSR xxxxxxxx/xxxxxxxx/xxxxxx40

This timer is sometimes called the Maintenance Timer and sometimes the Sanity Timer. They are the same thing.

When the CI is working it receives incoming messages from the CI cables, acknowledges them and stores them in memory for the PADRIVER to deal with. If the VAX is hung or halted then the PADRIVER will not process these messages but the CI will continue to acknowledge them so other nodes will believe this node is still OK.

The CI Maintenance timer checks that the PADRIVER is still alive. Before VMS V4.5 it was set to a fixed 99 second timeout. From VMS V4.5 it is set by SYSGEN parameter PASTIMEOUT and defaults to 5 seconds. If the CI does not detect any PADRIVER activity for this period of time then it sets bit 6 of the Port Status Register, initialises itself, stops responding to incoming messages and generates an interrupt.

This error is **ALMOST NEVER** caused by faulty CI hardware, it is caused by the CI detecting that the VAX is hung/halted. It may be caused by a hardware or software problem on the VAX. If you cannot find out what the cause of the problem was from the errorlog and there is no clue on the console of the VAX then you could try increasing PASTIMEOUT to cause the VAX to hang for longer before the CI is initialised or setting PASANITY to 0 which will prevent this timeout from EVER occuring. **BEWARE, this can turn a single VAX failure into an entire cluster hang.** If you are desparate enough to try something like this then you should have already spoken to support and read the section on Sysgen Parameters below.

10.10.1.2 CI port Timeout, Virtual Circuit Timeout.

%PAA0, CI Port Timeout
%PAA0, Virtual Circuit Timeout - Remote Port xxx

These timeouts were first implemented in VMS V4.3 and VMS V4.4.

CI Port Timeout means that the PADRIVER has not seen any incoming messages on the CI for PAPOLLINERVAL seconds, it causes the CI port to be re-initialised. It is not the same as a maintenance timer expiration, this is the Software detecting that the hardware has hung, the maintenance timer expiration is the Hardware detecting that the software has hung!

Virtual Circuit Timeout means that no CI packets have been received **FROM A PARTICULAR NODE** for PASTIMEOUT seconds.

These timeouts can be caused by faulty CI hardware (very unlikely), or S/W problems, or SYSGEN parameters not set to the correct values. Systems suffering from these symptoms should first have sysgen parameter changes to try to correct the problem. See VAX Stuff Issue No. 10, July 1987 page 21 for more information. Increasing PASTIMEOUT will reduce the incidence of Virtual Circuit Timeouts, increasing PAPOLLINTERVAL will reduce the incidence of CI port timeouts. If you need to do this then CONTACT SUPPORT FIRST. **NEVER CHANGE SYSGEN PARAMETERS ON A CUSTOMERS SYSTEM** without first speaking to support /your manager /the customer's system manager.

10.10.1.3 Path has gone from GOOD to BAD

%PAA0, Path #n has gone from GOOD to BAD - REMOTE PORT XXX

There are three possible reasons for getting this message.

1. Most commonly some remote system has been shut down or powered off, there is nothing wrong on this system and may well be nothing wrong on any system.
2. If this PORT has just been initialised and it is faulty you may see this message for a path THAT WAS NEVER GOOD! The assumption by the port driver is that the path was working and when it discovers that it isn't it reports this error.
3. The path may well have suddenly failed, this is not very common.

If the error occurs on one path at a time at infrequent intervals (less than one or two per week) with no other symptoms then **DISREGARD** it.

If these errors occur in a pair - i.e. both paths to another node go bad, followed by the PORT or SOFTWARE shutting the virtual circuit, then **DISREGARD** them. If there are other error messages at the same time then troubleshoot those other errors.

If this error occurs frequently for **ONE** path then you probably have a faulty (or missing) CI cable.

If this error occurs for **BOTH** paths every time you boot the system then you probably have a faulty CI module or corrupt CI microcode.

10.10.1.4 Cables have gone from Uncrossed to Crossed.

%PAA0, Cables have gone from UNCROSSED to CROSSED - REMOTE PORT XXX

If you get this message every time you boot then your CI cables are crossed over (swap the two A path cables for the two B path cables).

If you get two per week or less with no other symptoms then **DISREGARD** these errors. They will be fixed by an FCO some time in the future.

10.10.1.5 Remote system conflicts with known system

%PAA0, Remote System Conflicts with Known System - Remote Port xxx

The VAX or HSC with CI node number xxx (WHICH HAS JUST BOOTED) has the same SCSNODE or SCSSYSTEMID as some other VAX or HSC which has been in the cluster. This node will refuse to communicate with the new node until it is shut down and rebooted with a new SCSNAME and SCSSYSTEMID.

The usual reasons for getting this error are booting a system using some other-systems root directory, or changing the SCSNODE or SCSSYSTEMID on a node without changing **BOTH** to previously unused values.

This applies equally to HSC and VAX nodes.

10.10.1.6 Port/Software has closed virtual circuit.

%PAA0, Port has Closed Virtual Circuit - REMOTE PORT XXX
%PAA0, Software is closing Virtual Circuit - REMOTE PORT XXX

This may be the result of a system shutdown, fatal bugcheck, powerfail or some other problem. Look at the consoles on all systems and HSCs for an explanation and look at the errorlog entry for this error to try to gain more information. Also look at the errorlog entries on all systems for the time period just before this error.

10.10.2 %CNXMAN errors.

These error messages are written to the console by the Connection Manager which is the part of VMS responsible for arbitrating which VAXs are cluster members, what value quorum has at any instant etc.

They reflect changes in which nodes are members of the cluster, whether the quorum disk can be found and whether the cluster has quorum. The connection manager also sends these messages to OPCOM for broadcast to all operator terminals and writing to the operator log.

There is a description of all of these messages and what they mean in appendix B of the VAX/VMS Guide to Vaxclusters (this is one of the A5 size VMS manuals which were distributed with VMS V4).

Most of these messages are self-explanatory, the following are included here because they are not so obvious.

%CNXMAN, Deleting CSB for system NODENAME

The node has just rebooted and the old data structures are being deleted so that new ones can be created. This is a perfectly normal message and does not indicate an error condition.

%CNXMAN, Timed out lost connection to system NODENAME

The node has been out of communication for more than RECNXINTERVAL seconds, all locks that that node used to hold will now be re-allocated and if it attempts to rejoin the cluster without rebooting it will be forced to do a
FATAL BUGCHECK, CLUEXIT, Node voluntarily exiting cluster.

10.10.3 FATAL BUGCHECK, CLUEXIT

This bugcheck is reported on the console as

FATAL BUGCHECK, VERSION = V4.X CLUEXIT, Node voluntarily exiting cluster
or
FATAL BUGCHECK, CLUEXIT, Node exiting cluster

What it means is that the rest of the cluster has rejected this node for some reason. It is almost always caused by a node leaving the cluster for more than RECNXINTERVAL seconds and then attempting to continue. See Section 10.10.2, %CNXMAN errors. .

To troubleshoot these bugchecks you do not need a crash dump, you need to find out why the node left the cluster and returned. Look at the errorlogs and console printouts for all systems to see what was happening before the bugcheck.

If you cannot see what the cause was then the evidence you need to keep is the following.

- Console printouts from ALL VAXs and HSCs.
- Copy of sys$manager:operator.log for any VAX whose console printout was unavailable.
- Copy of sys$errorlog:errlog.* On ALL VAXs for the period just before and just after the bugcheck

Another possible cause of this bugcheck is trying to boot a VAX running a version of VMS more than one release different to the rest of the cluster, (e.g. booting a VMS V4.5 node into a cluster with a system running VMS V4.3).

10.11 HSC

After a long period of stability the number and variety of HSC's has dramatically increased over the last couple of years. This section attempts to enable an engineer to have enough information to work confidently in this changing workspace.

The notes file for the HSC is SSDEVO::HSCPHASE0, it is very active and well worth the reading.

10.11.1 HSC Documentation

Description	Part No.
Pocket Service Guide	EK-HSCPK-RC-004
Installation Manual	EK-HSCMN-IN-002
Fault Isolation Manual	EK-HSCFI-MN
Service Manual	EK-HSCMA-SV-003
Suplementry Service Information for HSC 65/95	EK-HS695-SI
Cronic V8.2 Release Notes	AA-GMFAP-TK
User Guide v8.0	AA-PFSQA-TK
Orange Maintenence Guide	QP-906-GZ

10.11.2 Part Numbers

Table 10–17: Box Part No's

HSC50/70 Operator Control Panel	54-15286
HSC50 Power Controller	70-20613
HSC70 Power Controller (881-B)	30-24374-02 (Fuse = +L-10575)
HSC70 Floppy Disk Drive	30-24962-01
Main PSU	70-20033-04 rev J01
Auxiliary PSU	70-20184-02 rev F01

Table 10–18: Module Part No's and Configuration of HSC Controller Models

Module Name	Part No.	HSC95	HSC65	HSC90	HSC60	HSC70	HSC40	HSC50
	K.CI							
Port processor	L0107-YA	No	No	No	No	Yes	Yes	Yes
Port processor	L0124-AA	Yes	Yes	Yes	Yes	No	No	No
Port link	L0100 Rev E2	No	No	No	No	Yes	Yes	Yes
	L0118-00	No	No	No	No	Yes	Yes	Yes
Port link	L0118-YA	Yes	Yes	Yes	Yes	No	No	No
Port buffer	L0109	No	No	No	No	Yes	Yes	Yes
Port buffer	L0125-YA	Yes	Yes	Yes	Yes	No	No	No
	P.ioj/c							
I/O control processor	L0105-00	No	No	No	No	No	No	Yes
I/O control processor	L0111-00	No	No	No	No	Yes	No	No
I/O control processor	L0111-YA	No	No	No	No	No	Yes	No
I/O control processor	L0111-YC	No	No	Yes	No	No	No	No
I/O control processor	L0111-YD	No	No	No	Yes	No	No	No
I/O control processor	L0142-YC	Yes	No	No	No	No	No	No
I/O control processor	L0142-YD	No	Yes	No	No	No	No	No
	M.std*							
Memory	L0106-AA	No	No	No	No	No	No	Yes
Memory	L0117-AA	No	No	No	No	Yes	Yes	No
Memory	L0123-AA	No	No	Yes	Yes	Opt	Opt	No
Memory	L0123-BA	Yes	Yes	No	No	No	No	No
	M.cache							
32 Mbyte cache	L0121-AA	No	No	Opt	Opt	No	No	No
64 Mbyte cache	L0121-BA	Yes	Yes	No	No	No	No	No
	K.S*I				**Requestors**			
K.SDI Disk	L0108-YA	Opt	Opt	Opt	Opt	Opt	Opt	Opt
K.STI Tape	L0108-YB	Opt	Opt	Opt	Opt	Opt	Opt	Opt
K.SI Disk/Tape	L0119-YA	Opt	Opt	Opt	Opt	Opt	Opt	Opt
K.SI Disk(8 port)	L0119-YB	Opt	Opt	Opt	Opt	No	No	No
K.SCSI Disk	L0131-YA	Opt	Opt	Opt	Opt	Opt	Opt	No

10.11.3 Cronic

10.11.3.1 Distribution Procedures, MDDS

The upgrade from one version of Cronic to another, was at one time an FCO, this is no longer the case. Cronic is now a customer installable software product. To recieve the latest release of the Cronic Operating System and Documentation, the customer must have a contract line item number :-

HSC50	QA-QX930-AE
HSC40/60-95	QA-QX926-AE

the same as for VMS. The customer has to pay for this privilege, so Cronic software can no longer be given away freely. If a customer is not receiving his Cronic Software, or needs the latest version (due to maintenence issues), the Geographic Unit Manager has to be informed and an entry placed on ECSO (to log the transaction), before an adhoc version wil be supplied, with only limited documentation.

10.11.3.2 HSC V8.40

Released September 94, includes support for the following new devices:

- RZ26L, RZ26B, EZ54R/EZ58R
- RRD43/RRD43/RRD44 CD-ROM, RWZ52 optical disk drive
- TZ87/TZ875/TZ877, TZ87N/TZ875N/TZ877N, STK4220
- TL820, RW524/RW530/RW534,RW536

10.11.3.3 HSC V8.20

Released May 93, for HSC40/60-95 Cronic now includes support for the RZ74 on the K.SCSI.

10.11.3.4 HSC V8.10

Cronic now includes support for the HSC9X-SX, K.SCSI requestor, and the RZ26 disk drive in a BA350 Storageworks Shelf.

10.11.3.5 HSC V7.00

Released to support the HSC 65/95, optimised for these HSC's only, as are the v700 Offline diagnostics.

10.11.3.6 HSC V6.50

New functionality provides for Controller assists in full and merge copies of HBVS shadow sets. New utility SCTSAV simplifies process for installing upgrades to Cronic software. Support for ESE50 and RA73 devvices.

10.11.3.7 HSC V4.1A

For HSC50's only. It is a field patched version of v4.10 the last distributed version of Cronic for the HSC50. No new functionality is envisiged for Cronic on the HSC50.

10.11.4 General

1. Since Cronic v6.0 the module configuration of the HSC is checked at boot time for legal module configurations. The HSC type (40/60-95) is also checked to see if the diskette has been configured for a HSC of another type (40/60-95) , a warning and instructions will be given if this is the case.
2. If you change the system diskette or tape on an HSC then you MUST boot the HSC with the online button OUT and change the system parameters BEFORE you allow any of the VAX's to see it. **Otherwise you may have to reboot thewhole cluster.**
3. If the online button is lit then there is at least one VAX with a connection to the HSC. If the online button is out then no NEW connections can be formed, but existing connections will not be broken (and the light will not go out) unless the VAX or the HSC is rebooted.
4. If ILDISK and/or ILEXER fail on a disk drive but internal micro-diagnostics and on-line operation are both OK, then try running FORMAT and reformatting JUST THE DBN's (not the customer data area).
5. VTDPY will give information on cable states, virtual circuits and SCS connections, disk and tape states (online, available, non-existent) in a constantly updated format. **You should not leave VTDPY running on an HSC all the time, it puts a heavy overhead on the HSC CPU.**

10.11.5 Using the HSC console terminal.

Each HSC has a console terminal for communicating with the operator or engineer. This terminal may be an LA12 or a VT220 with a slave LA75, the default terminal baud rate is 9600.

If the HSC is running its normal software then you must type ^Y on the terminal to get its attention. There is a switch inside the HSC marked ENABLE/SECURE. If this switch is in the secure position then you can only give certain commands from the terminal. **BEFORE TURNING THE HSC50 SWITCH TO THE ENABLE POSITION YOU MUST TYPE A FEW CHARACTERS ON THE TERMINAL.** If you don't then you may accidentally put the HSC50 into ODT mode. This *break* functionality has been removed in later revisions of P.IOJ/C modules .

If you manage to put the HSC into ODT mode by accident you will get one of two prompts depending on whether you have the HSC ODT running under RT11 or the micro-ODT which is driven by microcode on the P.IOC. If the HSC remains in either ODT for too long then the internal queues will get too large and the HSC will be requested to crash and reboot by one of the VAXs. (See Section 10.11.11.

The Prompt for micro-ODT is	@	To return to normal type	P
The prompt for HSC ODT is	*	To return to normal type	;P

To use the console terminal it must be set up correctly. The correct setups are given in the HSC Pocket Guide pages 37-40.

If you have to use some other hard copy terminal with the HSC50 then you can alter the baud rate of the console port to 300 Baud by removing jumper W3 on the L0105 (see HSC Pocket Guide pages 37-40). .

If you have to use some other baud rate for the HSC70 console terminal then use a 9600 baud terminal temporarily, type SET TERM /PERM 300 (for example) then remove the 9600 baud terminal and you can use a 300 baud terminal.

10.11.5.1 Set Host /HSC

You can use any vax terminal as though it were the HSC console

```
$ SET HOST/HSC hscname
```

will connect you to the HSC. Use control \ to return to VMS. This is possible for interactive use only. Not for command files or batch jobs. It is also not supported for running VTDPY, ILEXER, ILDISK and KSUTIL. Also backup or restore may only be run from ONE terminal at a time on any HSC.

To enable this to work the system manager must use sysgen to connect the FYDRIVER.

```
SYSGEN> CONNECT FYA0 /NOADAPTER
```

10.11.6 System Parameters

Before allowing an HSC to join or rejoin a VAXCluster you must check the following parameters.

THESE PARAMETERS SHOULD ONLY NORMALLY BE CHANGED BY THE SYSTEM MANAGER. IF YOU GET THEM WRONG THE CONSEQUENCES CAN BE DISASTROUS.

If any one of these is changed whilst the cluster is running then you **MUST** change all of them to previously unused values.

ID	Must be the same as last time this HSC booted
NAME	Must be the same as last time this HSC booted
CI NODE NUMBER	Must be the same as last time this HSC booted
DISK ALLOCATION CLASS	Must be the same as any other HSC which shares dual ported disk drives with this HSC. Must be different to any HSC which does not share disks. Must be 0 if no dual ported disks. IF YOU SET THIS WRONG YOU MAY GET A FATAL BUGCHECK (invalid disk configuration), OR DATA CORRUPTION, OR CONTINUOUS MOUNT VERIFICATION.
TAPE ALLOCATION CLASS	Must be the same as any other HSC which shares dual ported tape drives with this HSC. Before VMS V4.6 dual ported tapes were not supported.

10.11.7 Installing HSC Cronic operating system.

When installing an HSC or updating the HSC operating System cassette please be aware that the Software Control Table written on the diskette MUST be tailored to suit each HSC on the cluster. The following parameters need to be checked and changed as required **BEFORE THE HSC IS ALLOWED ONLINE** . . .

- System ID
- Name
- Host enables
- Dxxx Host Access
- Disk Allocation Class
- Tape Allocation Class

To install new code on an existing HSC follow these steps.

1. Obtain the System Managers permission to power off the HSC.
2. Type "SHO ALL" on the HSC console to obtain a printout of existing set up.
3. Use the A and B port button to failover any drives to alternate HSCs (do not do this step for SHADOW set members).

4. Press and release the ONLINE button to prevent VAXes from discovering this HSC when it reboots.
5. Power Fail the HSC (This will cause dual ported shadow sets to fail over).
6. Check that the new system tape/floppy is write enabled, then insert it in the HSC.
7. Press and hold the FAULT button, keep this button pressed till you see the message INIPIO-I-Booting . . . on the HSC console terminal.
8. Power the HSC on. Make sure that the ONLINE button is still in the out (offline) position.
9. When the HSC finishes booting run SETSHO and set parameters to the same value as those printed out in step 1. Type HELP for assistance with individual SETSHO commands. Each customer should have an HSC USER GUIDE with additional information on system parameters.
10. When you are completely satisfied that all parameters have the correct values, press and release the ONLINE button to put the HSC back online.

Each HSC on the cluster should have a unique ID (numbers only), and a unique name (numbers and letters only). Crash dumps should be enabled to the console only.

If changes are made to the ID or Name on an hsc in a running cluster each host will need to be rebooted otherwise although a connection will be made to the HSC when it is put Online the hosts will not see any disks/tapes attached to it (VMS ver. 3.x only) or will keep making and breaking connection (VMS ver. 4.x,5.x).

Other parameters may be changed as required by the customer or for fault- finding purposes but the defaults work OK.

If you are changing multiple parameters then type RUN SETSHO to a HSC> prompt, then make ALL of the changes, then EXIT SETSHO. This way you only reboot the HSC once for all of the changes.

To aid upgrading cronic software there is a new utility **SCTSAV**. You need to use the version for the cronic you're *coming from,* either **SCTV65** or **SCTV70.** Please see Release notes.

10.11.8 Performance Considerations.

Disks and Tapes must be configured on the Requestors as per HSC Software Release Notes. The highest priority requestor is on the R/H side (nearest CPU).

Table 10–19: Relative Device Speeds

Device	Relative Speed	Priority
TA90 or TA91 tape drive	Fastest	Highest
TA85 or TA857 tape drive	.	.
RA90 or RA92 disk drive	.	.
RA73 disk drive	.	.
ESE or EP-ESE Solid State Disk	.	.
RA82 disk drive	.	.
RA81 disk drive	.	.
RA71or RA72 disk drive	.	.
RA60 disk drive	.	.
RA70 disk drive	.	.
RA80 disk drive	.	.
RA73 disk drive	.	.
TA78/79/81 tape drives	Slowest	Lowest
SCSI disk drives (K.SCSI)	Delay-Tolerant	Lowest

- There is only one data path thru an HSC requestor at any instant in time.

- Devices of like speed should be connected to a requestor, this enables the requestor to make optimum use of the HSC backplane bandwidth.
- If the disks are dual-ported, configure the other HSC to be a mirror of the other, same disks to same requestor etc. With this configuration you can use the Preferred Path functions under VMS to load balance the HSC's. Not only will this spread the I/O thruput across two HSC's it will also make the failover functionality less traumatic as only half the disks have to failover. Also you know that what the disks are going to failover to, is also alive and well.
- No two members of the same shadow set should be connected to the same requestor. Not only do you loose performance, reads and writes are backed up behind one another, the failover functionality to other HSC's has been seen to fail.
- If all of the above has been achieved try to grade the usage of the disks on each requestor, so that only one heavily used disk is on requestor, along with one less heavily used disk. If the HSC partner is like wise configured, if failover takes place no one requestor has four heavily used disks.
- The 8 port requestor HSC9X-FA, is for connectivity only, connecting eight heavily used disks to it would cause a bottle-neck thru the one data path.
- With the advent of HSC Cache, programs have been developed to enable a customer to see if cache will be of a benfit to his disk usage, this is before the necessary upgrade to aquire cache. Also, when cache is installed in a HSC it can be monitored.

10.11.9 HSC Hardware Maintenence Features.

- You can have up to three requesters in an HSC50 without an auxiliary power supply. ALL HSC70s come with an auxiliary power supply as standard.
- HSC power shooting procedures have changed over the years, as older AUX PSU's were set to deliver +5.1volts, this is not the case with curent rev PSU's.
- HSC50 power cables, are not compatible with the new power controller, 881B, use the following part numbers are used in the HSC70...
  Cable part no. 17-01276-01 for the PSU's cable to 881B.
  Cable part no. 17-01276-02 for the Blower's cable to 881B.
- The LINK module L0118 must be in the backplane for the +12v to come up. No load on the -5.2v means no +12v. Also with just the AUX supply running the power light will not come on as it requires all 3 voltages to be up.
- A red light on an HSC50 module does not necessarily mean that that module is faulty. These lights are controlled by the P.IOC and may be turned on during the bootstrap process, or whilst running offline diagnostics. Also the red light on the PILABUFF module (L0109) is controlled by the K.PLI (L0107-YA), and the red light on the LINK module (L0100/L0118) simply shows that the local loopback is enabled.
  Generally if only one red led is on it probably shows a faulty module, if more than one red led is on then it should be safe to disregard them.
- Logistics have been known to ship the P.IOJ module with incorrect jumper settings with regard to HSC40/70 functionality, this could occur in the other simlar configurations, 60/90 and 65/95. Besides the jumper a rom is also unique to each CPU module of type.

10.11.10 HSC Diagnostics

10.11.10.1 Bootstrap diagnostics

The HSC will automatically boot on power up, it will also automatically reboot after a failure which causes a system crash. It can also be rebooted by a host sending a CI reset packet which generates an init signal to the P.IOC. To perform a manual reboot the enable/secure switch must be in the enable position, than press and release the init button.

Whilst the init button is pressed it provides a loopback connection for the console terminal so if the terminal is totally dead it is possible to press the init button, test that the terminal will echo and then release the init button to cause a reboot. There is a very good flowchart of the HSC bootstrap on pages 32-34 of the Orange Maintenance Guide, refer to this if the HSC will not boot.

Each module in the HSC has a red led which is turned on by the P.IOC asserting an init signal to that module and is turned off when the module completes its initialisation self-test.

10.11.10.2 Offline diagnostics

The HSC has its own suite of offline diagnostics. These are all on one offline diagnostic tape/floppy which should be on site with every HSC. Copies are available from support if any site needs them, they are not distributed with the HSC software. PLEASE check that these diagnostics are available on all of your sites so that they can be used when needed.

To boot the offline diagnostics put the tape into a TU58/RX33 drive and press the init button on the OCP.

The offline diagnostic monitor announces itself with an ODL> prompt you can then run any of the diagnostics or type any of the following commands.

ODL> SIZE	This is very useful. It produces a map of all the K's found and their slot numbers. It also gives you a memory map. ALWAYS run it before running any other tests, so you know what you have got.
ODL> HELP	Prints the help file. Uses lots of time and paper
ODL> TEST	Loads and starts a diagnostic
ODL> LOAD	
ODL> EXAMINE	Useful if you know of any patches (I don't)
ODL> DEPOSIT	
ODL> START	

To run the diagnostics type . . .

ODL> CONFIGURE	will print configuration map of HSC
ODL> TEST MEM	uses P.IOC to test memory (very slow)
ODL> TEST MEM BY K	uses one of the requesters to test memory (much quicker but wont test m.prog)
ODL> TEST REFRESH	tests memory refresh
ODL> TEST K	tests a single requester
ODL> TEST BUS	interactive test of all k's, ocp, tu58 and p.ioc, this is a good overall confidence test.
ODL> TEST OCP	tests the lights and switches

A common test sequence would be

```
ODL> SIZE
ODL> TEST BUS   (specify all Ks when asked)
```

All of the memory diagnostics ask for a starting and ending address to test. If you accept the offered defaults you will get non-existent memory errors. To get the correct addresses for the system you are on, just type SIZE to the offline diagnostic ODL> prompt.

For details of offline diagnostics see Orange Maintenance Guide pages 101 to 111.

10.11.10.3 Inline diagnostics

The HSC inline diagnostics run under the control of the normal HSC system software. Some of them can be run by issuing a RUN command from the console terminal, others are automatically scheduled to run by the HSC software. For full details of any of these inline diagnostics see the HSC inline diagnostic user documentation on microfiche. The inline diagnostics are:-

PRMEMY which tests the cbus, dbus and p.ioc parity logic. It does not test the memory. If it fails it causes the HSC to reboot. It is a periodically scheduled diagnostic and cannot be run by operator command.

PRKSDI and PRKSTI which test the K.SDI's and K.STI's. · These are also periodically scheduled diagnostics. Whenever they run they look for an idle requester and cause it to execute one of its micro-diagnostics. If they fail they cause the HSC to reboot.

CIRESP which responds to incoming test messages from the VAX CI exerciser diagnostic. CIRESP is never initiated by the HSC. To be runnable it must be loaded at boot time. See the HSC user guide for details on how do this.

ILMEMY which is a memory test. Whenever a requester detects a memory parity error in the data memory the offending memory buffer is placed on the suspect memory list. ILMEMY tests the memory on the suspect memory list. If ILMEMY fails, or detects that the memory buffer has been placed on the suspect list twice since the last reboot then the memory is placed on the bad memory list. It will not then be used, even after a reboot, until the bad memory list is cleared by operator intervention. (See the HSC user guide for details on how to do this). Because the bad memory list is on the system tape/floppy it is necessary that this device should always be left in the HSC in a write enabled state during normal operation. The load device is also used to store the reason for an exception if the system crashes.

ILTU58 tests a single TU58 tape drive. To run it type `HSC> RUN ILTU58`. It will prompt "Device name of TU58 to test (A) [] ?". Allowable responses are DD0: DD1: DD2: DD3: or LB:, but the normal HSC50 only has DD0 and DD1. LB refers to the TU58 from which the HSC50 was booted. The diagnostic does not require a scratch TU58, it uses a dedicated file on any of the standard HSC50 TU58's. It should not be run for extended periods of time as it continuously reads and writes the same block and would damage the media. The diagnostic will halt after any error message, except for "retries required" or "Data Compare Error". This is also a PERIODIC diagnostic and runs all by itself every now and again.

10.11.10.4 Testing disk and tape drives

Please see <REFERENCE>(ra_disks\full) for information on testing disk drives.

10.11.10.4.1 Running Iltape

ILTAPE is a tape transport and formatter diagnostic. It will test any available tape transport. It will also test any formatter that is available or offline. When you run it Iltape will prompt for a unit number. If you want to run a complete test of the K.STI and formatter then type Xm where m is any number. Iltape will then prompt you for a requester and port number. Alternatively answer Tnnn to test tape nnn. Iltape will give further prompts asking whether you want to run formatter diagnostics, whether you want to test the transport what speed and density you want to use etc. To default remaining questions type ^Z. Eventually it will stop asking questions and go away to test something. It is possible to enter a user defined sequence for testing tape transports. For details of how to do this see the HSC inline diagnostics user documentation on microfiche.

10.11.11 HSC "System Crashes"

There is lots of useful information about HSC Software Crashes in the Orange Maintenance Guide, pages 55 to 66.

If one of the Controllers on an HSC detects an irrecoverable problem it will interrupt at level 7. This causes the HSC to "Crash" and output a crash dump.

On rebooting the console will report (this duplicates much of the crash dump).

```
Last soft init caused by level 7K interrupt (trap thru 134).

From Process xxxxxx              \ these values are
PC xxxxxx                        / not interesting.
Status of requestors (1-9):

xxx   xxx   xxx   xxx   xxx   xxx   xxx   xxx   xxx
```

Possible values for xxx are

001	K.CI, not faulty
002	K.SDI, not faulty
203	K.STI, not faulty
377	Empty slot

Any other number indicates a fault detected by a K, see Orange Maintenance Guide page 59 for further info.

10.11.11.1 Host Clear.

One particular code to watch out for is 177, this is the "error" reported by a K.CI when it receives a "HOST CLEAR" command from a VAX. Only the K.CI can report this particular status. If you have previously done a SET NODUMP HOST then the HSC will simply reboot and report

```
Last soft init caused by host clear from CI node n.
```

This MAY be the result of . . .

- Exhausted retries on a Disk or Tape
- A VAX believes the HSC is insane.
- A CI problem.
- The HSC has been left in ODT too long. (See Section 10.11.5).
- Other transient problems anywhere in the software or hardware. (Sometimes the VAX errorlog will give a clue to this).

Sometimes the responsible node is INCORRECTLY reported as node 0, if node 0 is not a VAX then it cannot have caused a HOST clear.

10.11.11.1.1 Troubleshooting HOST CLEARs

1. Check the HSC errorlog (console printout) for any errors. If you find any then repair those!
2. Check the errorlogs of ALL vaxes in the cluster for any errors. If you find any then fix those!
3. If the HSC successfully reboots then it is almost certainly an INNOCENT party to the HOST CLEAR. There has probably been either a CI problem or a DISK/TAPE problem.
4. If the Host Clear is only intermittent and there are no other errors then ignore it!

Identifying which node requested a HOST CLEAR crash

This is an example HSC crash dump printout.
(..... represents where I have left something out to save space)

```
....................
....................caused by ( 134   )  Kint
....................
Status of requestors (1-9):
000177    any    any    any    any    any    any    any    any
....................
....................
Control area for slot  000001
nnnnnn
....................
....................
nnnnnn
nnnnnx  --17th Line of slot 1 shows the responsible NODE number.
nnnnnn
....................
....................
```

HOST_CLEAR status

The setting of HOST_CLEAR as DUMP or NODUMP is NOT SHOWable. Please ensure that on software upgrades you include "SET NODUMP HOST" as HSC crashdumps worry customers and engineers.

10.11.12 K.SCSI

The K.SCSI is the latest requestor to be manufactured for the HSC family of storage controllers. It is designed to enable customers having exsisting CI/HSC based disk subsystems to have a migration path to the new StrorageWorks SCSI based disk ARRAY controllers. As the device bridges two types of technology, and its quite a devide CI/HSC/DSA - SCSI, care has to be taken that the customer does not see any degradation in service of the product due to lack knowledge or training. This is the subject of great debate, the two storage representatives of PTG, Stuart Rance and Chris Loane are driving a country wide task force, to overcome training shortfalls where possible, seminars - roadshows - OJT and training.

The K.SCSI is a HSC requestor that uses's a SCSI bus interface to connect to its disks. This SCSI bus can be either single ended, or differential, but not both , detection is by cable connection. The BA350-SA shelf is used to house the RZ26-VA or RZ74-VA disk drives that are the only supported devices at present. No tape device support is available at present.

10.11.12.1 K.SCSI Documentation

Description	Part No.
HSC65/95 Supplementry Installation Info.	EK-HS695-IN
HSC65/95 Supplementry Service Info.	EK-HS695-SI
BA350 User's Guide	EK-BA350-UG
BA350 Configuration Guide	EK-BA350-CG
BA350 SBB User's Guide	EK-SBB35-UG
BA350-SA Shelf Users Guide	EK-350SA-UG
SCSI An Overview	EK-SCSIS-OV
SCSI A Developers Guide	EK-SCSIS-SP
DWZZA-VA Product Reference Card	EK-DWZVA-RC
StorageWorks Cookbook	Technical Library
SCSI Cookbook	Technical Library

Richard Hough of the Diagnosis group at Viables maintains a repository for all manner of StorageWorks documents on MAJERE::STORAGEWORKS. Richard is also the author of the StorageWorks cookbook which has a significant amount of information on the K.SCSI in it.

10.11.12.2 Part Nos.

Table 10–20: K.SCSI and BA350 part nos.

Part No.	Description
L0131-AA	K.SCSI Data channel (requestor)
54-22877-01	K.SCSI Transistion Cable Board (TCB)
BN21H-xx	SCSI hi-density A cable, male straight/straight
BA350-SA	StorageWorks Shelf
BA350-HA	131watt universal ac input power supply.
BA350-HB	131watt universal dc input power supply.
BA35X-MB	SCSI Terminator for BA350.
BA35X-MC	SCSI Jumper for BA350
BA35X-MA	Fan carrier assembly.
RZ26-VA	3.5" 1.05 GB disk.
RZ74-VA	5.25" 3.5 GB disk.
DWZZA-VA	DWZZA bus adapter SBB carrier assembly.
47-00157-01	3.5 SBB Carrier Extractor tool.
47-00158-01	5.25 SBB Carrier Extractor tool.

10.11.12.3 Single Ended or Differential Bus

In terms of the K.SCSI, SE is used whenever the SCSI bus is going to remain inside the HSC cabinet, this is both a maximum length of cable consideration (4 meters terminator to terminator), and of secure earthing (due to parity bus). The Differential bus is selected when the storage shelf is outside the HSC cabinet, in another HSC, using the DWZZA-VA SCSI bus adapter . This allows for the disk drives to be **DUAL PATHED**.

10.11.12.4 Storage Shelves.

The BA350-SA is the StorageWorks shelf that is used to house and power the RZ26-VA / RZ74-VA disk drives. It consists of a box with a backplane that is 1meter of SCSI bus. By use of the terminator/jumper on the rear of the backplane the bus can be configured as one or two busses. It can have redundent AC power supplys, DC power supply, or Battery backup. Depending on configuration 5-7 drives maximum can be installed in the BA350.

10.11.12.5 3.5 and 5.25 SBB's

The RZ26-VA is a 3.5 SCSI disk drive packaged in a plastic carrier, having a unique connector that plugs into a BA350-SA backplane slot and picks up its SCSI ID from the slot.

The RZ74-VA is a 5.25 SCSI disk drive packaged in a plastic carrier, having a unique connector that plugs into a BA350-SA backplane slot, the SCSI ID is determined by switches on the RZ74-VA. As a 5.25 SBB it takes up 3 slots in a BA350 shelf, as a SCSI bus can have up to 7 target drives, you would expect to split the BA350 SCSI bus when using these devices.

The above two devices are deemed **HOT SWAP** by HSC Engineering, this has only been recently defined. Up until this time it was considered necessary to **Quiese** the SCSI bus.

10.11.12.6 DWZZA

The DWZZA-VA is a DWZZA in a 3.5 SBB carrier, placed in slot 0 that is used to convert the Differential Bus from a (remote) K.SCSI, to Single Endded SCSI of the BA350 storage shelf. Differential SCSI bus length can be 20 meters long.

When used in the above scenario, the DWZZA must have its own termination disabled as it will be the third terminator and in the middle of the effective SE SCSI bus.

10.11.12.7 Dualpathed vs Dualport

The HSC has in the past always used dual ported disk drives, this allows seperate paths from one of two controllers to control, and pass data to and from the disk drive. SCSI drives have only one path, via the SCSI bus interface to obtain commands and data in the packet format of SCSI protocol. The use of the Differential bus from a K.SCSI in one HSC to another HSC's Storage shelf, allows for the (local) HSC to have maintenence work carried out on it without loosing connectivity to the disks. Ala dualpathed. The two K. scsi's that are effectively connected to the **ONE** SCSI bus, maintain a dialogue between each other to effect failover.

10.11.12.8 Shadowing

The K.SCSI does not support Controller Based Volume Shadowing, however after defining the SCSI disks into MSCP logical units, seen by VMS, then Host Based Volume Shadowing can be employed to shadow disks of like geometry.

10.11.12.9 KSUTIL

KSUTIL (chapter 14 of the HSC User's Guide) is the utility that is the user interface to the K.SCSI. This creates a virtual terminal on the HSC, because of this you cannot use the SET HOST/HSC function to run KSUTIL, but VCS can be used in the normal way. The KSUTIL code runs in the K.SCSI and provides the normal functions that you would expect to have to do to SCSI disk.

- **Configuration Utility**- Assigns SCSI devices to MSCP unit numbers.
- **Operator Control Panel**- Provides front panel display and functionality.
- **Format,Verify, and Initialize Utility**-Creates scripts that execute these functions.
- **Reassign Block Utility**- Forces a block replacement. .
- **Exercise Utility**- Exercises one or more SCSI devices.

Normal HSC utilitys like ILEXER and DKUTIL run on the MSCP assigned drives of the K.SCSI as any normal HSC disk. The assigning of MSCP drive numbers to K.SCSI disk drives should bear some resemblence to their physical configuration to aid identification.

10.11.12.10 DSA like.

Normally SCSI disk drives do not support the functionality of the Digital Storage Architecture, BBR/RCT/FCT etc... To ensure that this level of reliability is not lost, the K.SCSI writes **meta data** on to the innermost tracks of the disk, enabling tracking of forced errors and block replacements etc.. This makes the disk **non-transportable** to any other SCSI controller, execpt those that employ the same protocols. The HSJ40 is one such controller.

10.11.12.11 StorageWorks Errorlog Analysis Tool

The decoding of errorlog information of drives on the K.SCSI is greatly enhanced by using the tool SWEAT, written by Chris Loane. It will also decode errorlog information for HSJ40 disks. A copy of SWEAT can be obtained from Comics::Disk$Tech:[Storageworks]SWEAT_v21b.BCK

10.12 HSJ40, SW800/SW500

The HSJ40 is a CI based disk array controller, using SCSI disk, and eventually tape devices. All hardware is based on the StorageWorks BA350 enclosures, mounted in 800mm Datacentre cabinets. The HSJ40 has very simlar functionality to the HSC familey of disk controllers. It differs quite significantly in that it can configure more then one disk into a logical MSCP unit. At present this functionality allows only STRIPESET configurations, in the future, by purchasing other versions of the firmware customers will be able to configure RAID storage sets.

In Dual-Redundant controller configurations (2 x HSJ40) accessing one set of devices, Load balancing and Failover, HSClike functionality, is possible. In Dual-Redundant configurations the maximum no. of disk drives configurable falls from 42 to 36, as the 2nd HSJ40 takes up a SCSI id on all six of the controllers SCSI busses. Failover is accomplished by an internal comms path, used by the controllers to kill the partner and assume its I/O operations, including using its read cache, if it determines the **other_controller** is in some way faulty.

10.12.1 HSJ40 Software Dependancies

Operating System	" with limlits "	Latent	"Fully Supported"
OpenVMS VAX	V5.5-1/V5.5-2	V6.0	Coral
OpenVMS AXP		V1.5	Epsilon
OSF/1 AXP		V1.2A	Morgan

10.12.2 HSJ40 Documentation

Description	Part No.
HSJ40 User's Guide	EK-HSFAM-UG-A01
HSJ40 Service Manual	EK-HSFAM-SG-A00
HSJ40 Firmware Release Notes rev T047	EK-HSFAA-RN-A01
HSJ40 Firmware Release Notes rev E05B	EK-HSFAA-RN-B01
HSJ40 Firmware SPD ver 1.0A (E05B)	AE-PYTGA-TE-10A
StorageWorks SW800 Cabinet Installation Guide	EK-SW800-IG-A01
StorageWorks SW500 Cabinet Installation Guide	EK-SW500-IG-A01?
BA350 User's Guide	EK-BA350-UG
BA350-MA Controllers User Guide	EK-350MA-UG-A02
BA350 Configuration Guide	EK-BA350-CG
BA350 SBB User's Guide	EK-SBB35-UG
BA350-SA Shelf Users Guide	EK-350SA-UG
SCSI An Overview	EK-SCSIS-OV
SCSI A Developers Guide	EK-SCSIS-SP
StorageWorks Cookbook	Technical Library
SCSI Cookbook	Technical Library

Richard Hough of the Diagnosis group at Viables maintains a repository for all manner of StorageWorks documents on MAJERE::STORAGEWORKS. Richard is also the author of the StorageWorks cookbook which has a significant amount of information on the HSJ40 in it.

The notesfile for the HSJ40 is SSDEVO::HSJ40_PRODUCT, it is very active and well worth the following.

10.12.3 HSJ40 and BA350 Part Nos.

Table 10–21: Part Nos.

Part No.	Description
70-30097-01	HSJ40 Controller module
54-22229-02	HSJ40 16Mbyte read cache module.
54-22229-01	HSJ40 32Mbyte read cache module.
17-03427-02	HSJ40 internal CI cables.
BN21H-xx	SCSI hi-density A cable, male straight/straight
BA350-SA	StorageWorks Shelf
BA350-MA	StorageWorks Controller Shelf
BA350-SA	StorageWorks Shelf
BA350-HA	131watt universal ac input power supply.
BA350-HB	131watt universal dc input power supply.
BA35X-MB	SCSI Terminator for BA350.
BA35X-MC	SCSI Jumper for BA350
BA35X-MA	Fan carrier assembly.
RZ26-VA	3.5" 1.05 GB disk.
RZ74-VA	5.25" 3.5 GB disk.

10.12.4 HSJ40 Operational Features

It is very important to document the physical, logical, and Cluster configurations of the HSJ40/SW800 at Installation time and at any time that changes are made after installation. This could reduce any problems arising from a failure of an HSJ40 and the need for its replacement. If it is only a single controller configuration this is the only means of reconfiguring the disks to the controller.

10.12.4.1 Cluster HW Issues

Any one HSJ40 controller will take up one CI node ID no., this is set by firmware commands, as is the enabling of the CI paths on or off. A Dual_Redundant HSJ40 configuration will therefore take up 2 different CI node, ID no's.

Below is the subject of a BLITZ TD: 1446, dated the 20-Aug-1993.

The SW810-AA/AB (HSJ40) will not auto-boot with VAX or DEC 7000/10000.

VAX/DEC 7000/10000 console system currently does not support auto-boot functionality on the SW810-AA/AB (HSJ40).Customer expectation set that full HSJ40 functionality would be available in August'93. Problem is not well defined between VAX/DEC 7000 console and HSJ40. These restrictions apply to console code version 3.0, VAXVMS version 5.5-2, and HSJ40 code version V1.0 & V1.0B.

The HSJ40 running HSOF v10A only supports the following CI adapters ,

- CIXCD
- CIBCA-B

Future support may include,

- CI780
- CIBCI

The CIBCA-A will not be supported.

Any one SW800 cabinet can support 3 Dual_Redundant HSJ40 contoller configurations along with their associated 6 BA350 StorageWorks Shelves, making a total of 21 BA350-XX's, 18 SCSI cables, 6 CI cable sets, and possibly 2 Cabinet Distribution Units (power controllers). The word BUSY springs to mind.

10.12.4.2 OCP

The OCP is the home of 7 LED lit push switches, one (GREEN), is the RESET button and also acts as the HSJ40 STATE light. The six (AMBER), are the PORT QUIESCE buttons, labled 1 - 6 used to quiese its associated SCSI bus so that WARM SWAP replacement of SBB's can take place without disturbence to the rest of the HSJ40/SW800.

The OCP leds are also used to indicate particular failures of the HSJ40 and are decoded in the Service Manual.

The Port quiesce buttons are also used to indicate that the devices on its SCSI bus are not configured the way the NVRAM says that they are.

10.12.4.3 Console

The HSJ40 has an MMJ (EIA 423) socket for VT2XX like terminal to be attached, 9600 baud is the default but can be set by firmware commands, to your choice. The HSJ40 comes with neither cable or terminal.

If a working HSJ40 in Dual_Redundant configuration fails, it will not interface to the console terminal until it has been restarted by the Other_Controller, the one that killed it.

The HSJ40 HSOF can also be accessed by remotely from a VAX using the following DCL command.

```
$ SET HOST/LOG=CLIsession.info/DUP/SERVER=MSCP$DUP/TASK=CLI HSJ40nodename
```

The /log qualifier will keep a record of your session so it can be edited if necessary to provide configuration documentation. To return control to the VAX, type **EXIT** to the CLI prompt.

10.12.4.4 HSOF

The HSOF is distributed by PCMCIA program card media only, the HSJ40 needs the PCMCI to be inserted at all times otherwise it stops functioning, period.

The HSOF supports the following utilitys.

- VTDPY - HSC like.
- DIRECT - Displays a Directory of the program card.
- TILX - Tape Inline Exerciser Utility.
- DILX - Disk Inline Exerciser Utility
- CLI - Operating System Software.

As the documentaion for the CLI has been for along time in DRAFT state it is worth getting hold of the most current version possible. The CLI consists of six basic command sets.

- Failover commands.
- Controller commands.
- Device commands.
- Storageset commands.
- Logical Unit commands.
- Diagnostic Utility commands.

The CLI has a HELP command, using the (?) character will prompt CLI to give you a choice of commands/switches at that point.

Table 10–22: Part Nos. for HSOF

Part No.	Description
QL-0W9A*-**	HSOF License
QA-0W9A*-**	SWKS HSJ40 MSC v1.0 PCRM Card
QA-0W9A*-GZ	HSOF Documentation Kit
QT-0W9A*-**	HSOF Product Services

10.12.4.5 SHADOWsets - RAID 1

At present the HSJ40 does not support RAID1 storagesets configured within the HSJ40 as MSCP logical units. However MSCP logical units configured within the HSJ40, single disks or Stripesets can be shadowed with an equivelent logical unit with Host Based Volume Shadowing.

10.12.4.6 STRIPEsets - RAID 0

The maximum size of a Stripeset is 8.5GB with VMS below v6.0. How the customer wishes to configure his stripeset is very flexible, the HSJ40 can configure a disk, to a maximum of 5, from any one of the 6 SCSI busess that the HSJ40 supports, into a Stripeset. The trade offs are raw throughput - as against majority resilence as RAID 0 has no, R redundancy.

- Member disks of a Stripeset, on diferent SCSI busses will be fast, a power failure in any one BA350 would take out all Stripesets that have a disk from that BA350 configured in them.
- Memeber disks of a Stripeset, all on one SCSI bus may suffer from some small throughput delay, however a power failure in a BA350 will only take out the one Stripeset.
- Power failure's to one BA350 shelf or one SW800 cabinet can be provided for by installing extra power circuits. The SW800 comes wired to have an extra Cabinet Distribution Unit installed, to power an extra BA350-HA, the ac input PSU SBB in the BA350 shelves.

10.12.4.7 RAIDsets

RAID technology allows independant disks to be constructed into a logical array seeen as one disk by the operating system.

Table 10–23: Levels of RAID

RAID Level	Description
RAID 0	Striping - Has no Redundancy features.
RAID 1	Shadowing/Mirroring - Redundancy and Fastest
RAID 3	Storage array, with one dedicated parity disk.
RAID 5	Storage array, with distributed parity across disks.

10.12.4.8 RAID Evaluator

To assist with the configuring of RAID sets, engineering have introduced a RAID Evaluator package to run on a PC. This is primarily a tool to help the introduction of the DECRaid 5 software, soon to be introduced to OpenVMS, but can be used to understand performance considerations of different RAID levels. Given that you know your I/O characteristics and the hardware available you can simulate a given RAIDset and determin how it will perform, without actually having to invest in the time to build it.

The evaluator resides in KBOMFG::APPL4:[Storage$public.eval] for the time being.

10.12.4.9 3.5 and 5.25 SBB's

The RZ26-VA is a 3.5 SCSI disk drive packaged in a plastic carrier, having a unique connector that plugs into a BA350-SA backplane slot and picks up its SCSI ID from the slot.

The RZ74-VA is a 5.25 SCSI disk drive packaged in a plastic carrier, having a unique connector that plugs into a BA350-SA backplane slot, the SCSI ID is determined by switches on the RZ74-VA. As a 5.25 SBB it takes up 3 slots in a BA350 shelf, as SCSI busses can have up to 7 target drives, you would expect to split the BA350 SCSI bus when using these devices.

10.12.4.10 DSA like.

Normally SCSI disk drives do not support the functionality of the Digital Storage Architecture, BBR/RCT/FCT etc... To ensure that this level of reliability is not lost, the HSJ40 writes **meta data** on to the innermost tracks of the disk, enabling tracking of forced errors and block replacements etc.. This makes the disk **non-transportable** to any other SCSI controller, execpt those that employ the same protocols. The K.SCSI is one such controller.

On the HSJ40 you can initilize the SCSI disk in a *transportable* fashion, this loses the customer the benifits of DSA, but the disk should be able to be seen by non-intelligent SCSI controllers, such as VS3100's.

10.12.4.11 StorageWorks Errorlog Analysis Tool

The decoding of errorlog information of drives on the K.SCSI is greatly enhanced by using the tool SWEAT, written by Chris Loane. It will also decode errorlog information for HSJ40 disks. A copy of SWEAT can be obtained from Comics::Disk$Tech:[Storageworks]SWEAT_v21b.BCK

10.12.4.12 WARM SWAP

The HSOF v1.0A does not support Warm swapping of the HSJ40 controller module.

The HSJ40 does support WARM swapping of an individual RZ** diskdrive, however the bus, in which the disk drive is installed in, has to be quiesced.

WARM swapping is where the disk/shelf has power to it but no I/O activity is taking place on the SCSI bus.