1 System Boot

I. Connectivity

A. Physical: complete SCI ring cabling

In order for the host to successfully probe all SCI nodes, the ring must be physically intact, i.e. completely cabled.

B. Electrical: host, E-Boxes must be powered up separately

Formerly, there were synchronization problems associated with the mechanical necessity to power up the host and ECHaPs at different times. We believe this is no longer a problem - you should be able to power them up in any order and each will ‘wait’ for the others it needs to complete booting.

I. Connectivity

A. Physical: complete SCI ring cabling

B. Electrical: host, E-Boxes must be powered up separately

- o former SCI node synchronization problems
- o now OK

II. Configuration

A. Host (still) serves all M16 downloads, RAID and vpartab definitions

B. ax_bootcfg must reflect H/W configuration for *next* bootup

```
EBOX-0
        Memory=256M
        Disks=10
        Diags=Yes
```

1. Connectivity and Configuration

II. Configuration

A. Host (still) serves all M16 downloads, RAID and vpartab definitions

Look for download M16 images in /usr/AXbase/images, RAID, vpartab files, and iftab in /usr/AXbase/etc.

B. ax_bootcfg must reflect H/W configuration for *next* bootup

Here's a sample ax_bootcfg file for a single-node system:

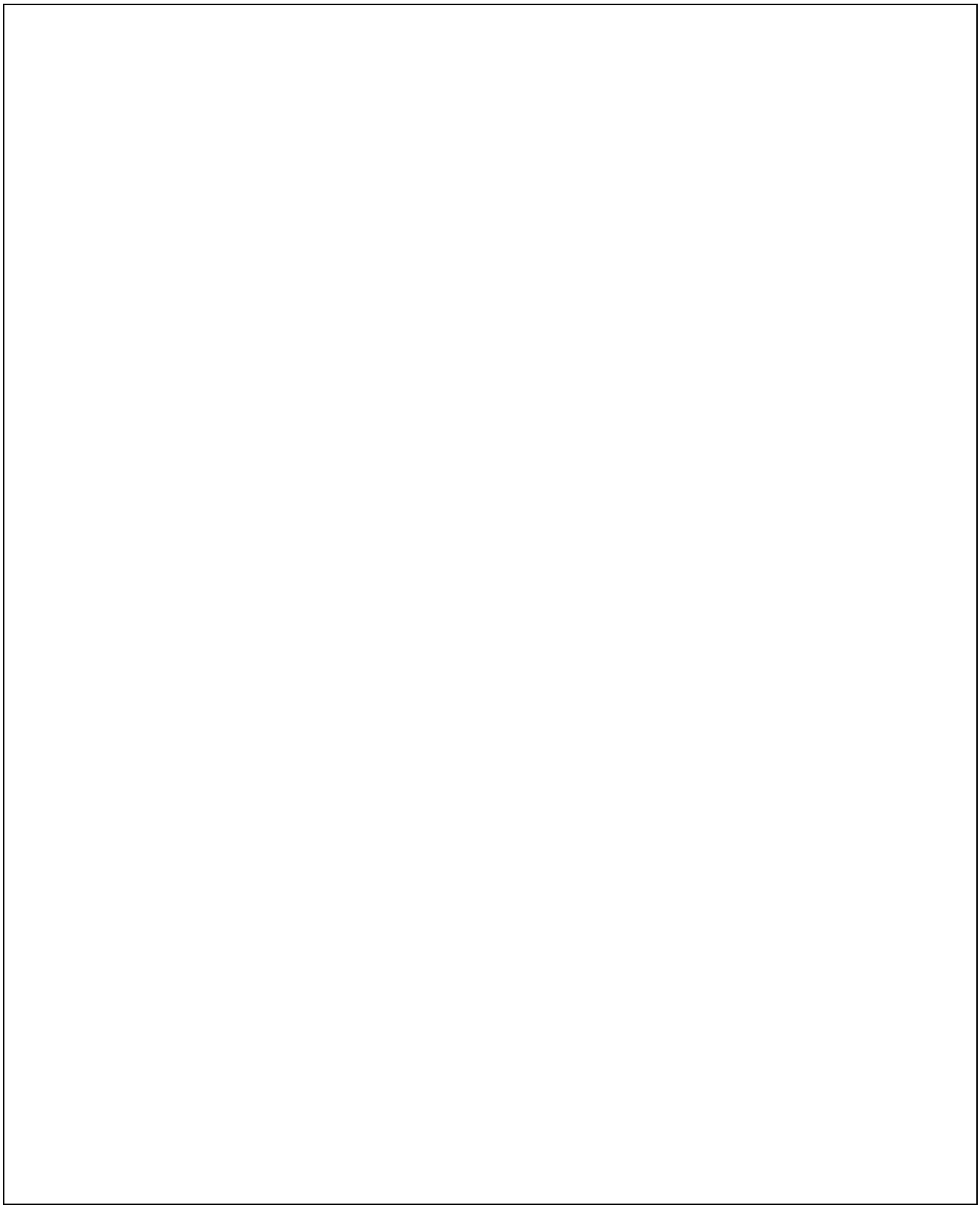
```
EBOX-0
        Memory=256M
        Disks=10
        Diags=Yes
```

Specifying memory and disks allows the host to estimate the length of time it 'should' take for this EBox to become visible, i.e. to boot itself to the EMon prompt, spin up its disks and to be able to be probed by the host. Diags specifies whether or not diagnostics are being run on boot; if 'Yes', the time estimate for this EBox to boot is increased accordingly.

If this file is not present, the system chooses a reasonable default (but see below).

If the system is to be taken down for a hardware upgrade, this file should be modified to reflect the new configuration *beforehand*.

▲ Parachute: If you take down the system and make hardware modifications before remembering that you forgot to edit this file, use 'boot -AUSPEX'. Auspex services will not be started, and you will be able to edit the file and reboot the system to allow the new configuration to be recognized.



Slidetitle

III. Boot Flow

A. Boot options from the kadb and OBP prompts

1. Booting without providing Auspex services

You may simply boot vanilla Solaris all the way to run level 3 with

ok boot -AUSPEX

No LFS, RAID or virtual partitions, or Auspex NFS service will be available. You may edit files by hand, use the host Ethernet interface for diagnostic purposes, or perform other ‘offline’ activities.

To support this option, /etc/init.d/auspex checks for it early; if in effect, it creates the /.AXIPC_DIAGMODE file and exits. This bypasses the remainder of Auspex boot logic and allows Solaris to boot normally.

▲ Parachute: boot -AUSPEX is absolutely essential for recovering from whoops-this-configuration-just-won’t-boot-and-I’m-truly-stuck-type (WTCJWBAITST) problems.

2. Booting without taking core dumps

Under certain conditions, you may wish to bypass taking a core dump after a system crash. To do this, boot with

ok boot -NODUMP

To support this option, /etc/init.d/auspex checks for it right after -AUSPEX; if in effect, it creates the /.AXIPC_NODUMP file before continuing.

3. After a crash (when you *do* want core dumps!)

a. For the alpha release, we will be configuring systems to boot kadb and without auto-boot? set. After a crash, you should be at the kadb prompt:

```
kadb[0]:
```

Typing ‘:c’ here tells kadb to ‘continue’ and will allow the system to take dumps from all nodes and the host.

III. Boot Flow

A. Boot options from the kadb and OBP prompts

o boot -AUSPEX (‘Not starting Auspex Services...’)

o boot -NODUMP

o :c (kadb) and sync (ok)

B. Vanilla Solaris (pre)boot

o boot file, size

o banner (including jumbo patch bug ID, if any)

o early Solaris devices: ata, host network interface

2. Boot Flow

Typing ‘$q’ here tells kadb to quit and brings you back to the OBP prompt:

```
kadb[0]: $q
ok
```

If you quit the kernel debugger, you have two choices. If you still want cores, you may type ‘sync’. This will cause all host file systems to sync as best they can, and core dumps to be taken from all nodes and the host. If you wish to avoid taking core dumps, simply boot with ‘boot -NODUMP’ or, if you wish, ‘boot -AUSPEX’ here.

NOTE: Once you use any method other than :c or sync to boot, you have lost the opportunity to take meaningful system core dumps. The SCI nodes will be reset and state information from the crash will have disappeared.

4. In the final release, kadb will not be booted by default. You then have only the sync or boot -* options available after a system crash.

B. Vanilla Solaris (pre)boot

From the host processor’s point of view, booting starts with the early stages of booting Solaris. The following messages are typical:

```
Rebooting with command: boot
Boot device: disk  File and args: kadb
kadb: kernel/unix
Size: 276441+61156+70372 Bytes
/platform/sun4u/kernel/unix loaded - 0x96000 bytes used
SunOS Release 5.6 Version Generic_105181-05 [UNIX(R) System V
Release 4.0]
Copyright (c) 1983-1997, Sun Microsystems, Inc.
ata-model: QUANTUM FIREBALL ST3.2A
ata-model: TOSHIBA CD-ROM XM-6002B
ata_clear_interrupt problem  4
WARNING: interrupt level 4 not serviced
configuring network interfaces: hme0.
Hostname: lima2 # .....................end of Solaris (pre)boot
Starting Auspex IPC at Init Level S
NOTICE: Link Controller LC2 found
NOTICE: Rev C PSB found
```

III. Boot Flow (cont'd)

C. Auspex run level S

- o S50drvconfig configures SCI (ax_ipcd will verify)
- o all other run levels launch from /etc/init.d/auspex (links)
- o ax_ipcd forks KPROC, SMON, PROBE, UTEST

D. Auspex run level 2

- o ax_ipcd again if needed
- o ax_storage scan
- o ax_loadvpar ('Loading virtual partitions ...')
- o ax_mkdev ('Populating /dev tree with active FSP...')
- o ax_fsck_scratch -s
- o ax_fsck_mount

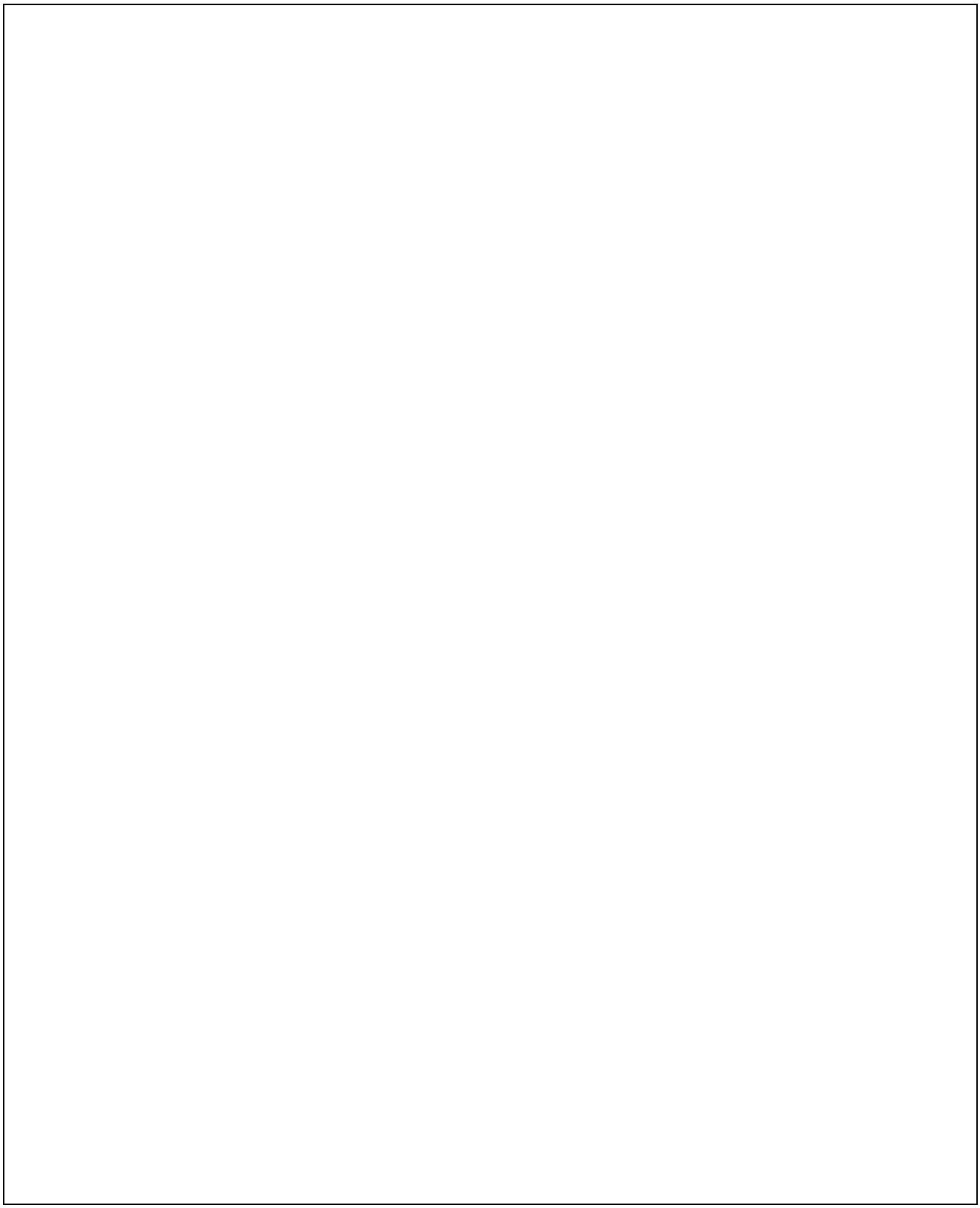
3. Boot Flow (cont'd)

```
WARNING: sci address 605d8000 605e0000 615e4000

NOTICE: Host is the SCRUBBER node

WARNING: sci: high level handler required.
System needs 84 secs delay
NOTICE: Number of nodes is 1
NOTICE: Performing SCI configuration
ASSIGNED NEW NODE ID - OLD = ff0a, NEW = ffea
ASSIGNED NEW NODE ID - OLD = ffea, NEW = ff0a
Peer Node Fifo Address is e2000000
sci_config: NO SCI SOFT CONFIG CHANGES
registered dbg vec trap
 have instance 1 of DE 21140.
 have instance 2 of DE 21140.
 have instance 3 of DE 21140.
 have instance 4 of DE 21140.
 have instance 1 of IPFDDI .
 have instance 1 of ALT GB_ENET .
 set AXIPC_ONLINE
Auspex IPC online
configuring Auspex network interfaces:
ifconfig: SIOCSIFBRDADDR: afe0: Bad address
 afe0.
The system is coming up.  Please wait. # end of Auspex run level S
Starting Auspex IPC at Init Level 2
Waiting for RAID arrays to be scanned...
ax_storage: Scan complete on fsp0
Loading virtual partitions...
Populating /dev tree with active FSP device nodes...
Checking Auspex FileSystems
checking ufs filesystems
/dev/rdsk/c0t0d0s5: is clean.
/dev/rdsk/c0t0d0s7: is clean.
checking for crash dump...
add net default: gateway 10.80.8.1
NIS domainname is lab.auspex.com
starting rpc services: rpcbind keyserv ypbind done.
Setting default interface for multicast: add net 224.0.0.0: gateway
lima2
syslog service starting.
Print services started.
volume management starting. # ......... end of Auspex run level 2
Starting Auspex IPC at Init Level 3
The system is ready.

lima2 console login:
```



Slidetitle

C. Auspex run level S

1. /etc/rcS.d/S50drvconfig configures SCI. Normally, it's during this process that the host pauses to wait for all ECHaPs to come up to the EMon prompt (described in a later section).

2. ax_ipcd

a. 'ax_ipcd -p <images>' verifies that SCI has been configured (or otherwise initiates configuration) and M16 nodes to be downloaded. It's during this process that the host pauses to wait for all ECHaPs to come up to the EMon prompt (described in a later section).

b. 'ax_ipcd -c start 0' causes daemons to start only.

3. ax_ipcd forks daemons AX_KPROC, AX_SMON, AX_PROBE, AX_UTEST.

a. AX_KPROC: M16 'kernel process' on the host. Listens for M16 messages from other M16 nodes, including name daemon messages.

b. AX_SMON: System monitor process. Monitors for initial SCI soft configuration failure, and for runtime IPC state changes (panic, reset, dump, etc.).

c. AX_PROBE: M16 node probe process. Continually probes and maintains information regarding other M16 nodes.

d. AX_UTEST: Unit test process. Expected to be removed before final release, though it may well be present in alpha and beta releases. Any messages regarding 'ut_worker' or 'unit test worker' originate here.

4. Configure Auspex devices

a. drvconfig: Builds all nodes in /devices for attached Auspex devices and applies permissions (only to any newly-created nodes) as indicated in /etc/minor_perm. See drvconfig(1M). Straight Solaris.

b. devlinks: Makes links in /dev to device nodes placed in /devices by drvconfig. Consults /etc/devlink.tab for specifications: device type, device name, minor fields. See devlinks(1M). Straight Solaris.

III. Boot Flow (cont’d)

E. Auspex run level 3

- o ax_nfsd (Yay!)
- o ax_mrestore
- o Now serving NFS and fully booted

4. Boot Flow (cont’d)

At the end of run level S, normally the SCI ring is configured, M16 nodes have been downloaded, and the host is monitoring IPC activity.

D. run level 2

1. try ax_ipcd again (if it failed previously)

2. ax_timed

3. ax_storage scan

a. Advises GAM to scan for RAID arrays and initialize state.

4. ax_loadvpar (`Loading virtual partitions...')

5. ax_mkdev (`Populating /dev tree with active FSP...')

a. Complements devlinks by creating devices in /dev/[r]axmrd, /dev/[r]axvp, /dev/raxmt, /dev/raxac

6. ax_fsck_scratch -s - configure FSP scratch device for fsck

7. ax_fsck_mount - mount, if necessary first fscking, file systems in /usr/AXbase/etc/lfstab

7. ax_statd

At the end of run level 2, normally all FSP file systems (except for mirrors) and virtual partitions are initialized, fscked and ready for use.

E. run level 3

1. try ax_ipcd yet again (if it failed previously)

2. ax_nfsd (yay! We can actually serve files now)

3. ax_mrestore (‘Restoring Auspex Mirror File Systems...')

4. ax_errd

IV. SCI Configuration

A. Characterize hardware

B. Read /etc/aus_node.config

- **o binary file of known size**
- **o contains M16 node/PCI config info**
- **o checksummed and backed up**

C. Compute maximum expected EBox delay

- **o ASCII file**
- **o contains info on memory size, disks, diags for each EBox**
- **o 60-sec. default is not enough for all configurations!**

D. Wait for all ECHaPs to come up

- **o to be visible to entire SCI ring**

5. SCI Configuration

At the end of run level 3, the NPs are ready for NFS service, mirrored file systems are initialized and the system has finished booting.

V. Boot Performance

1. BIOS - full memory scrub (2-3 mins. on 1Gb)
2. EPOST - memory test which may be shortened
3. ECHaP perf - possibly no improvement for now
4. Host - unknown as of now

6. Boot Performance

IV. SCI configuration

1. Characterize hardware (Link Controller, PSB chip revision)

2. Read /etc/aus_node.config to count E-Boxes

a. binary file of known size

b. contains information on all M16 nodes and PCI devices in this configuration

c. checksummed

▲ Parachute: This file is important! The system maintains (and if necessary uses) a backup in /etc/aus_node_backup.config.

3. Compute expected maximum delay for E-Boxes to come up

a. From /etc/ax_bootcfg as shown above.

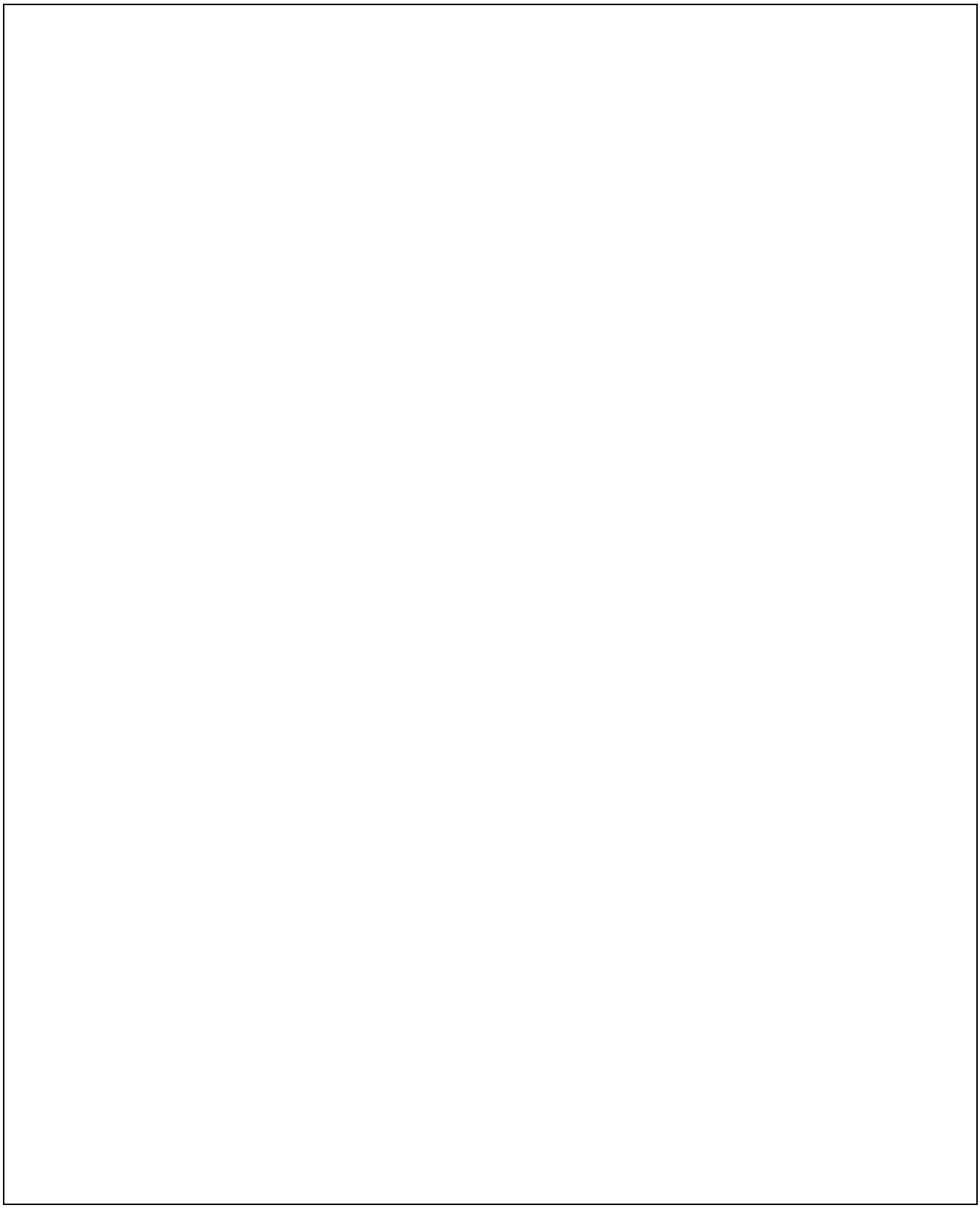
b. If ax_bootcfg is not found, default is currently 60 secs. This is not enough for a 1Gb node!

4. Wait for *all* ECHaPs to come up to EMon prompt
(otherwise they are not visible to the SCI ring)

Notes:

1. SCI cable location independence.
Currently, the only supported way to add nodes is at the 'end' of the SCI ring. Insertions inside a ring may cause the new node to be recognized as the node formerly occupying its place. In progress ...

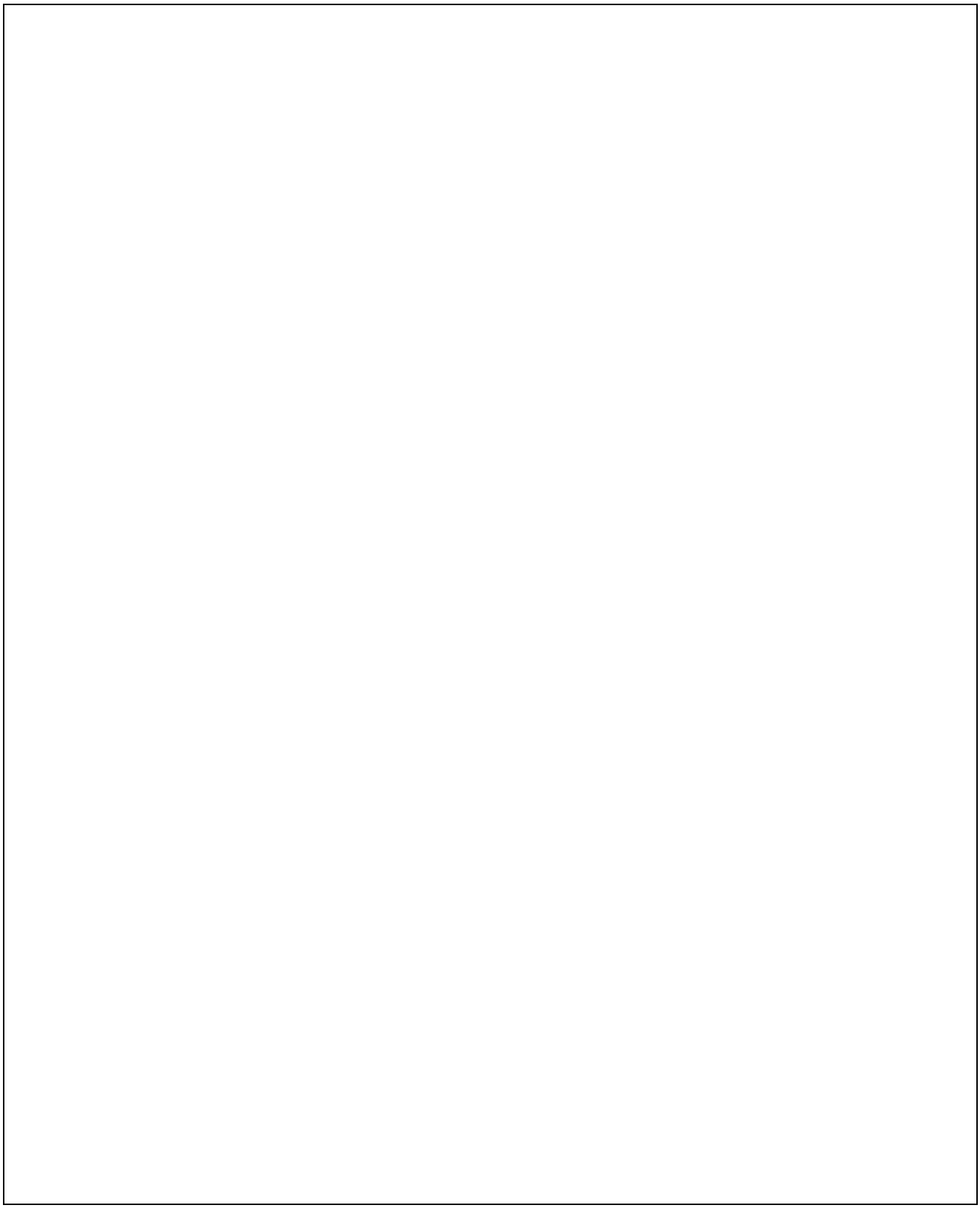
2. SCI switchouts.
Switching one bad node for a new node requires manual intervention (ax_bootcfg editing).



Slidetitle

V. Boot Performance

1. BIOS - full memory scrub (2-3 mins. on 1Gb)
2. EPOST - memory test which may be shortened
3. ECHaP perf - may be no improvement for now
4. Host - unknown as of now



Slidetitle