SAN Setup for HP's Tru64™ UNIX

Ludwig Frohnsbeck

26. Decus Symposium 2003 Bonn

Tru64™ SAN Setup

- Purpose
  - Overview of setting up SANs in a Tru64 Unix environment
  - Tips to aid in management and troubleshooting
- Not a course in SAN fundamentals or a tutorial
- Why? Fibre channel is different from any other storage for Tru64 Unix

Objectives

- **Storage Challenges**
- **Products and Roadmaps**
- **Console Configuration**
- **Snapshots and Clones**
- **Connectivity**
  - **DRD and Multipathing**
  - **The Emx driver**
- **Management and configuration tools**

Storage Challenges

Storage challenges

Enabling business velocity

- Explosive data growth
- Critical data availability and integrity
- Managing complexity

- Open environment needs
- Intelligent connectivity
- Return on investment

Tru64 has exceptional and unique strengths

Imagine solving your toughest challenges through…

- Scalable I/O and files to handle **explosive growth**
- **Available and reliable data** in all circumstances
- Access to your data **regardless of where it stands**
- Complying with the **latest standards**
- Using **smart connectivity** to increase performance and availability
- Managing it all **automatically**
- **It's all possible today on *Tru64* UNIX!**

Tru64 UNIX Optimizes SANs

• Fast data recovery	• Snapshots • Tight integration with EVM
• Data availability and disaster tolerance	• DRM integrates with clusters
• Robust, qualified, huge SANs	• Big Asymmetric SAN System • 800 TB to petabytes
• Integrated management and automation	• SAN management from SysMan • Automatic detection and management of volume expansions

StorageWorks Fibre Channel Roadmap

Tru64 UNIX handles explosive growth

AdvFS can handle file sizes of up to 16TB

Fewer mount points for easier management of growth

Infinite connection possibilities with Fibre Channel support

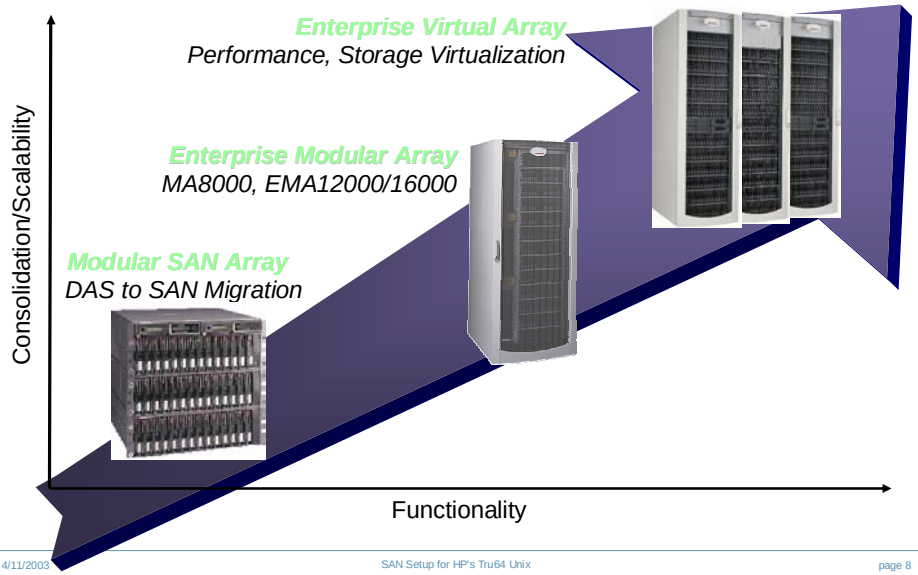
1TB physical volume
256 LUNs per target
256 targets per bus
255 buses

BASS (Big Asymmetric SAN System)

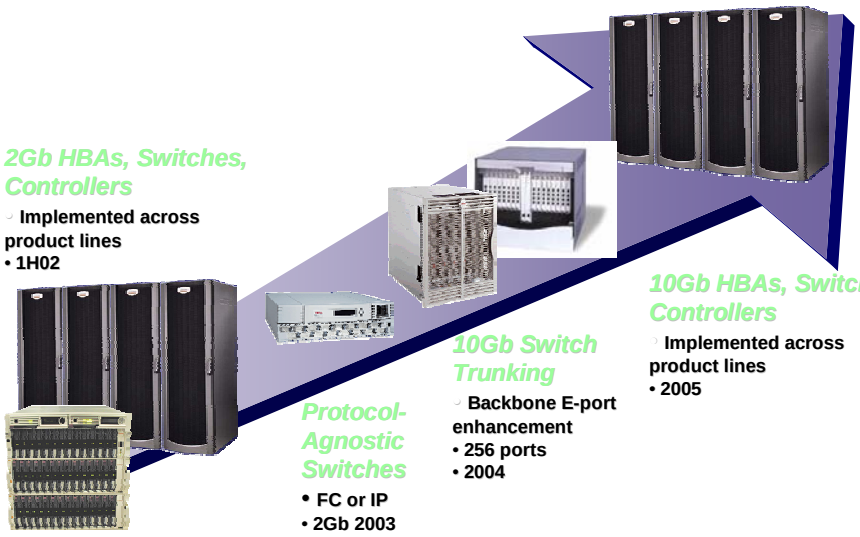
800 TB up to petabytes

LSM for increased performance in complex configurations

StorageWorks Array Family



StorageWorks Fibre Channel Roadmap



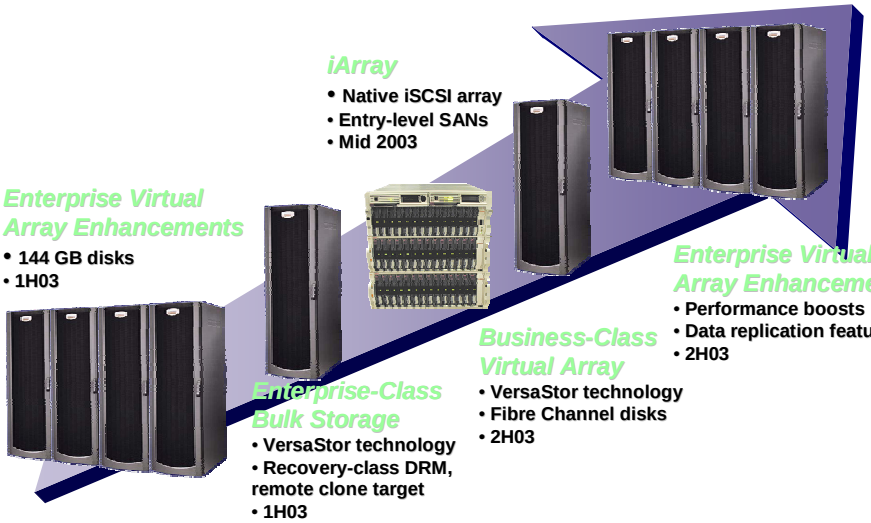
2Gb HBAs, Switches, Controllers
- Implemented across product lines
- 1H02

Protocol-Agnostic Switches
- FC or IP
- 2Gb 2003
- 10Gb 2004

10Gb Switch Trunking
- Backbone E-port enhancement
- 256 ports
- 2004

10Gb HBAs, Switches, Controllers
- Implemented across product lines
- 2005

StorageWorks Fibre Channel Roadmap



iArray
- Native iSCSI array
- Entry-level SANs
- Mid 2003

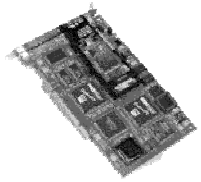
Enterprise Virtual Array Enhancements
- 144 GB disks
- 1H03

Enterprise-Class Bulk Storage
- VersaStor technology
- Recovery-class DRM, remote clone target
- 1H03

Business-Class Virtual Array
- VersaStor technology
- Fibre Channel disks
- 2H03

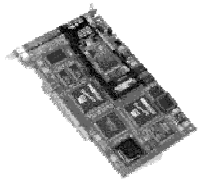
Enterprise Virtual Array Enhancements
- Performance boosts
- Data replication features
- 2H03

DS-KGPSA-CA (aka Emulex LP8000)



- 5V, 64-bit/33MHz PCI adapter
- Supports 1Gb fabrics
- Supports DS, ES and GS series of platforms
- DS10, DS20, DS25, ES40, ES45, GS80, GS160, GS320
- Full boot/dump support in Tru64 UNIX kernel

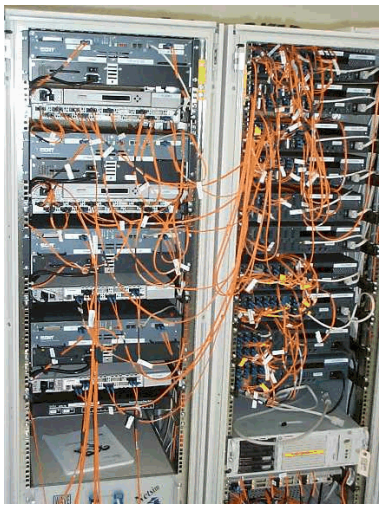
FCA2354 (DS-KGPSA-DA – aka Emulex LP9002L)



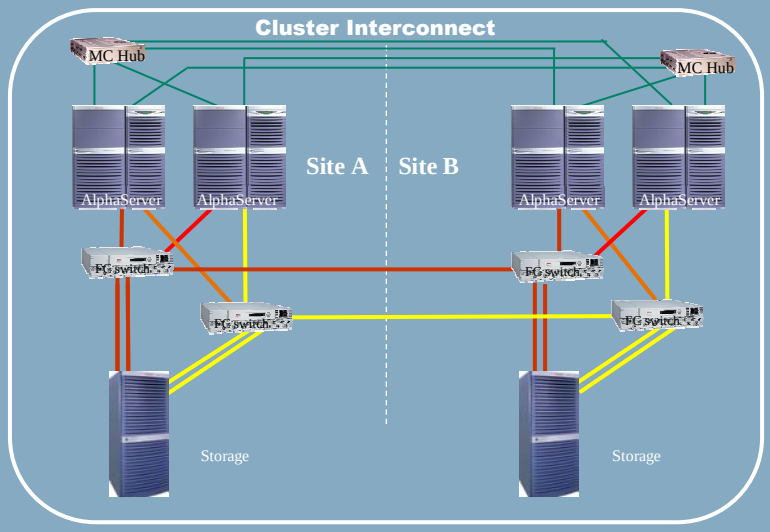
- 3.3V or 5V, 64-bit/66MHz PCI adapter
- Supports 2Gb fabrics
- Supports DS, ES and GS series of platforms
- DS10, DS20, DS25, ES40, ES45, GS80, GS160, GS320
- Full boot/dump support in Tru64 UNIX kernel

Document architecture and design

- SAN Architecture must be designed on paper *FIRST*
- Why document?
  - Might this picture be a good reason?
  - This is a good example of why you want to document
- This is one of the *most important* aspects of the Architecture process
  - This allows you to fully review and evaluate the design beforehand
- SAN is Not Documented?
- *SAN is Not Supported!!!*



Tru64 Unix disaster tolerant cluster



Fibrechannel configuration

Console Configuration

Tru64 Unix Console
Setting a WWID in NV memory

scan for fibre devices

map disk UDID 5

wwid (wwn)

paths (4)

```
P00>>>wwidmgr -show wwid
[0] UDID:5 WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-003a (ev:none)
[1] UDID:4 WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-0039 (ev:none)
[2] UDID:3 WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-0038 (ev:none)
[3] UDID:2 WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-0037 (ev:none)
[4] UDID:1 WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-0036 (ev:none)
[5] UDID:-1 WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-0046 (ev:none)
P00>>>wwidmgr -quickset -udid 5
Disk assignment and reachability after next initialization:

6000-1fe1-0000-0cb0-0009-9130-8234-003a

                    via adapter:        via fc nport:         connected:
dga5.1001.0.3.1     pga0.0.0.3.1        5000-1fe1-0000-0cb4   Yes
dga5.1002.0.3.1     pga0.0.0.3.1        5000-1fe1-0000-0cb2   No
dgd5.1001.0.2.0     pgd0.0.0.2.0        5000-1fe1-0000-0cb4   Yes
dgd5.1002.0.2.0     pgd0.0.0.2.0        5000-1fe1-0000-0cb2   No
```

Tru64 Unix Console accessing a fibre boot disk

mapped disk UDID 5 now appears

```
P00>>>show dev
dga5.1001.0.3.1 $1$DGA5 HSG80 V85F
dga5.1002.0.3.1 $1$DGA5 HSG80 V85F
dgd5.1001.0.2.0 $1$DGA5 HSG80 V85F
dgd5.1002.0.2.0 $1$DGA5 HSG80 V85F
dka0.0.0.1.1 DKA0 RZ2CA-LA N1H0
.
.
pga0.0.0.3.1 PGA0 WWN 2000-0000-c921-0d00
pgb0.0.0.5.1 PGB0 WWN 1000-0000-c920-cd9c
pgc0.0.0.1.0 PGC0 WWN 1000-0000-c920-a7ae
pgd0.0.0.2.0 PGD0 WWN 2000-0000-c921-07c4
pka0.7.0.1.1 PKA0 SCSI Bus ID 7
pkb0.7.0.2.1 PKB0 SCSI Bus ID 7 5.57
```

4 paths

Register disk as boot device

```
P00>>>set bootdef_dev dga5.1001.0.3.1, dga5.1002.0.3.1, dgd5.1001.0.2.1,
dgd5.1002.0.2.1
```


Tru64 OS Installation what to consider

- Only those FC devices whose SCSI-3 WWIDs are configured in the console will be presented in the installation menus.
  - Enforces console configuration before installation starts
  - As B/T/L meaningless, protects against selection of the wrong device

Tru64 OS Installation what to consider

- To help recognize HSG80 UNITs, installation menus will display HSG80 Unit IDENTIFIER

	Device Name	Size in GB	Controller Type	Disk Model	Location
1)	dsk0	4.0	SCSI	RZ2CA-LA	bus-0-targ-0-lun-0
2)	dsk1	4.0	SCSI	RZ2CA-LA	bus-0-targ-1-lun-0
3)	dsk2	1.0	SCSI	RZ26F	bus-1-targ-1-lun-0
4)	dsk3	2.0	SCSI	RZ28	bus-1-targ-4-lun-0
5)	dsk10	8.5	SCSI	HSG80	IDENTIFIER=133
6)	dsk12	8.5	SCSI	HSG80	bus-2-targ-2-lun-2

Tru64 UNIX mapping: WWIDs to device names

Identifying a storage device

•Tru64 makes use of WWID's within CAM

-Exports device special files to user-space

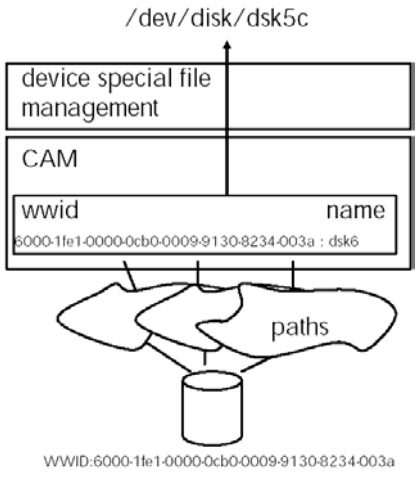
•WWID's are collected and stored by CAM

-WWID to DSF maps are stored in databases

•how CAM collects WWID's is specified in the DDR database.

-depends on each peripheral

-Tru64 support WWID's for devices that do not "have them"



Unique identification with „WWID“

example of an old device with a concocted WWID (rare)

```
# hwmgr -show scsi -did 0 -full
SCSI DEVICE DEVICE DRIVER NUM DEVICE FIRST
HWID:  DEVICEID HOSTNAME    TYPE   SUBTYPE OWNER   PATH FILE   VALID PATH
------------------------------------------------------------------------------
  17:  0        ernie       disk   none    2       2    dsk0   [0/3/0]

WWID:0410004c:"DEC RZ26 (C) DECPCB=412225056947(ZG25056947 );
HDA=0000030635357245"

     BUS  TARGET  LUN  PATH STATE
     ------------------------------
     0    3       0    valid
     2    3       0    valid
```


Unique identification with „WWID“

example of an new wwid capable device

```
host1 > hwmgr -get attr -id 133
     133:
     name = SCSI-WWID:01000010:6000-1fe1-0000-0cb0-0009-9130-8234-003a
     category = disk
     sub_category = generic
     architecture = SCSI
     phys_location = IDENTIFIER=5
     dev_base_name = dsk17
     capacity = 17768677
     block_size = 512
     model = HSG80
     boot_capable = 1
```

Tru64 Unix
mappings are stored in hardware databases

Hardware Component Databases

- /etc/dec_hwc_ldb (binary) *(CDSL)*
- /etc/dec_hwc_cdb (binary)

SCSI Device Database

- /etc/dec_scsi_db (binary) *(CDSL)*

Hardware Persistence Database

- /etc/dec_hw_db (binary) *(CDSL)*

Device Special File Data Files

- /etc/dfsl.dat (text) *(CDSL)*
- /etc/dfsc.dat (text)

Unique ID Database

- /etc/dec_unid_db (binary)

Preparing your data for backup

by using:

LSM
ADVFS
EVM

Using LSM in a SAN for backup

- **Logical Storage Manager**
  - Since V5.1A ability to mirror all Cluster Filesystems
  - Alternative to Hardware backup cloning, using volassist for backups
- ***volassist snapstart volumename***

  adds an extra plex to a volume and synchronizes it (mirror copy, may take a while); plex removed and used for backup later.
- ***volassist snapshot volumename tempname***

  creates a temporary volume (tempname) that is now detached.
- ***Continue application or remount volume***

  while backup temporary volume (tempname). When backup complete, remove the temp volume

  volume stop tempname; voledit –r rm tempname.

Hardware Cloning/Snapshotting and AdvFS

- In case of wrong usage you will be confronted with:
  - Unmountable file systems.
  - User data corruption.
  - AdvFS domain panics.
  - Tru64 UNIX kernel panics.
- Cannot clone individual filesets; must clone entire domain. Put filesets you want to back up or use together in the same domain.

What About Application Data Consistency?

- Only applies to single-LUN scenarios. (single-volume-domain)
- Before clone/snap is made, may want to quiesce or stop applications for application data consistency in clone/snap:
  - Oracle: online backup mode.
  - Flush all cached file data via *fsync(), O_SYNC*, *O_DSYNC*, *chfile –l on.*
  - Kill/halt application threads.
- Without application quiesce, clone/snap is “crash consistent” only.

Hardware based snapshots and Advfs

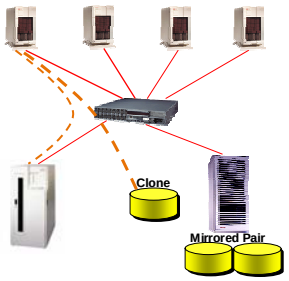
Whats the big picture

1. Quiesce (single-LUN) or unmount (multi-LUN).
2. Use HSG80 commands to create clone/snap and make it available. This creates new LUNs. (CLI or Cmdscript)
3. Use Tru64 UNIX commands to make new LUN(s) visible to the operating system as new devices.
4. Use Tru64 UNIX commands to create a populated /etc/fdmns/<new domain> directory. Do not use *mkfdmn.*
5. Mount filesystem(s).

SAN-Initiated Backup with EVM

Enterprise Volume Manager V2

- Full support for Tru64 UNIX and TruCluster Server V5 for controller-based snapshot/clone generation
  - AdvFS provides Freeze/Thaw to enable consistent clone or snapshot
  - LSM supports cloned volumes



Connectivity

- Multipath
- Multibus
- Load Balancing
- Overview of implementation in Tru64 and HP-UX

Multipath / Multibus / Load Balancing

- Multi-Path is the ability to connect more than one adapter to the same storage
- Multi-Bus is a more specific term , refers to the capability of devices to connect to multiple independent busses, or ports.
- Load Balancing is embedded in Tru64
  - Done on the KGPSA Adapter
  - Since V5.1 OS is starting a round robin to determine the path sending down an IO.

Multi – Pathing

HP-UX

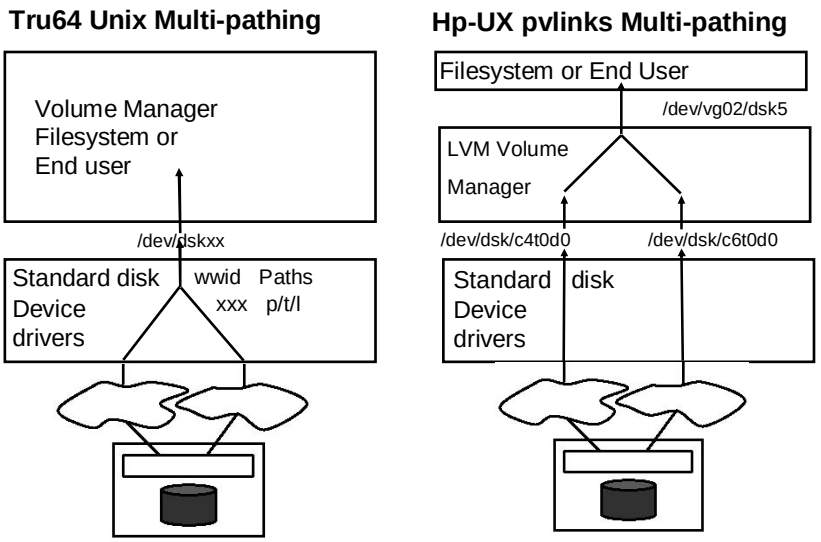
- not in base-os iostack drivers
  - devices have location dependent names "c*X*t*X*d*X*"
- multiple options at one of two levels

1. special layered driver associated with storage array on top of default drivers
   - HP-Classic SureStore AutoPath
   - Compaq-Classic StorageWorks SecurePath
   - EMC…
   - etc.
2. feature of volume manager
   - LVM pvlinks
   - VxVM dynamic multi-pathing (DMP)

Tru64 UNIX

- built-in to base-os iostack
  - transparently and automatically enabled for all scsi/fibre devices
  - devices have physical location independent names "dsk*XX*"
  - other methods are not required

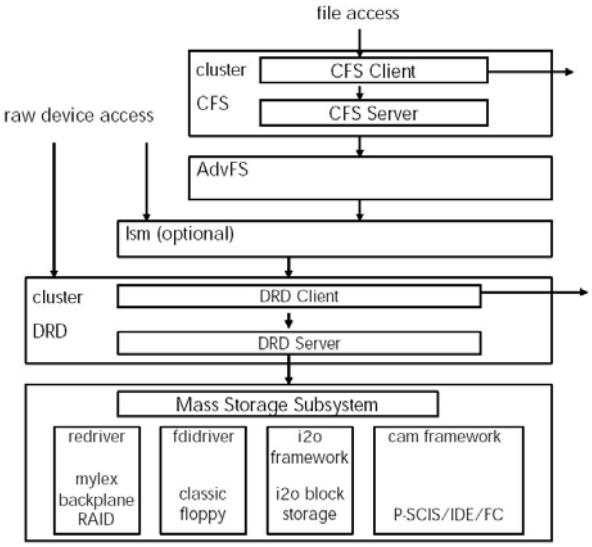
Multi – Pathing implementation



Multi – path capabilities

Description	HP-UX SureStore AutoPath	HP-UX LVM PVLinks	HP-UX VxVM DMP	Tru64 UNIX CAM
Fail Over: automatically fails over to an alternate path when the primary path is no longer available	√	√	√	√
Fail Back: automatically recognizes the newly available path when a failed path comes back up alive.	√		√	√
Active/Active - Static Load-balancing: Balances I/O load among all available paths with user selectable load-balancing policies.	√		√₂	√₂
Active/Active - Dynamic Load-balancing System automatically balances I/O load among all available paths based on run-time statistics such as device queue depths etc.				√

Tru64 IO - stack



what does the emx driver do?

- maps each adapter to a logical SCSI Bus
- logs into the fabric's F-Port if available (Switch)
- logs into all visible N-Ports in the fabric (Storage Endpoints)
- maps each FCP-Target N-Port to a SCSI Target
- receives SCSI commands from the peripheral driver
- sends SCSI commands wrapped in Fibre Channel frames to the devices

emx driver parameters in sysconfig

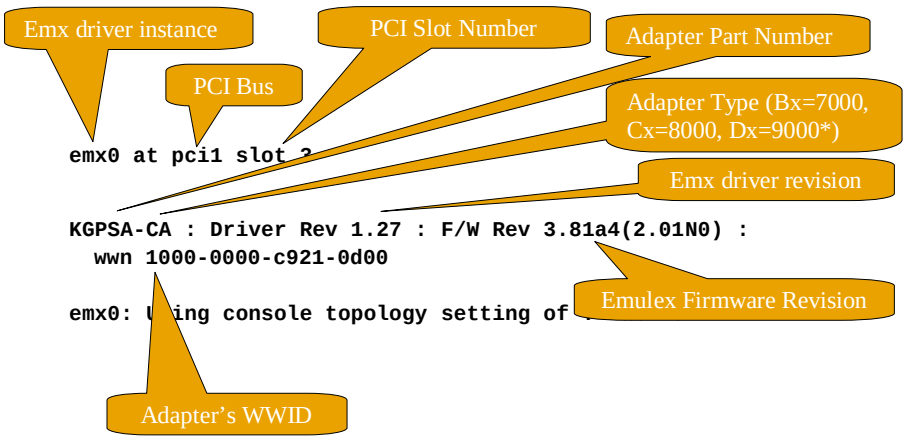
Driver Version

```
# sysconfig -q emx Driver_Version
emx:
Driver_Version = 1.32a
#
```

Number of HBAs being Managed

```
# sysconfig -q emx Num_Attached
emx:
Num_Attached = 2
#
```

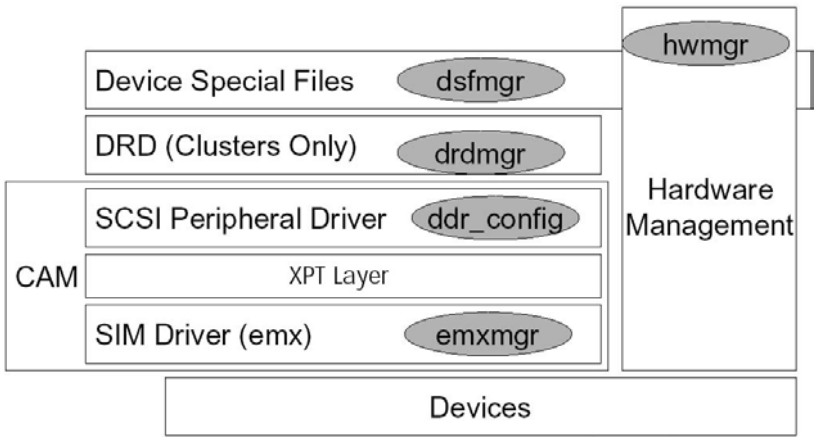

Emx messages during boot.



Hardware management tools

- hwmgr
- dsfmgr
- drdmgr
- emxmgr
- scu

Tru64 Tools



device and hardware management tools

device discovery and configuration

- automated device discovery
  - during boot process
  - or by automated polling of FC busses
- automated loading of drivers based on connected devices
- device special files created automatically

hwmgr

- used to view capabilities, connectivity and properties of devices
- used to modify properties of a disk device

dsfmgr

- used to view, verify and manage HW to DSF mappings

emxmgr example

```
# emxmgr -d

The available adapter instances are:

        emx0   emx2   emx3   emx4

foo> emxmgr -t emx0


emx0 state information:
  Link :connection is UP
        Point to Point
        Fabric attached
        FC DID 0x210513
  Link is SCSI bus 2 (e.g. scsi2)
        SCSI target id 255
        portname is 1000-0000-C921-0D00
        nodename is 2000-0000-C921-0D00
```

This adapter's link state

This adapter's Topology

An F – Port exists on this Fabric

Fabric assigned Fabric ID

CAM has assigned SCSI-BUS 2 to this adapter

The adapter's traget ID is 255

The adapter´s port and nodename

The scu command

- Moving a device to another path (controller failover)

```
# scu
scu> set nexus bus 3 target 1 lun 2
```

this should be a path on the controller
you have the lun preferred to

```
Device: HSG80, Bus: 3, Target: 1, Lun: 2, Type: Direct Access
scu> start
scu> quit
```


Adding a new device after boot

- "Connect" the device to the fabric or make it visible to the system.
- Run "hwmgr –scan scsi" to make the system poll for new devices.
- Run "dsfmgr –k" to add device special files for the new devices.
- Go for it!

Viewing events with EVM

- A graphical event viewer, fully integrated in the Sysman Application suite
- Set of command line utilities
  - ***evmwatch*** is used to monitor events as they occur
  - ***evmget*** to retrieve stored events from log files
  - ***evmsort*** to sort a set of retrieved events
  - ***evmshow*** to format the output of the retrieved events

Gathering information…

- kern.debug /var/adm/syslog/kern.log
- user.debug /var/adm/syslog/user.log
- daemon.debug /var/adm/syslog/daemon.log
- auth.debug /var/adm/syslog/auth.log
- syslog.debug /var/adm/syslog/syslog.log
- mail,lpr.debug /var/adm/syslog/misc.log
- binary.err /var/adm/binary.errlog
- msgbuf.err /var/adm/crash/msgbuf.savecore
- kern.debug /var/adm/messages
- kern.debug /dev/console

Questions ?