# TruCluster on HP-UX: How does it look?

Greg Yates (gry@hp.com)
Consultant
Hewlett-Packard

# Disclaimer

The information in this presentation is for a product that is still in development. Consequently the plans presented here are subject to change.

# Introduction

- This presentation will focus on the IPF (Itanium) platform
- TruCluster CFS (Cluster File System)
- TruCluster SSI (Single System Image)
- AdvFS File System

# TruCluster on HP-UX

- Since we are integrating TruCluster into HP-UX, some of the ‘feel’ of TruCluster/Tru64 will be different.  The idea was to integrate the product into HP-UX as seamlessly as possible.

# TCR CFS and TCR SSI

- There are two distinct offerings of TruCluster on HP-UX
  - TCR CFS -- CFS but no shared root. The CFS-served cluster file system is /etc/cfs plus a user-available mount point called /cfs. This will ship earlier than the SSI product, in the 11.31 time-frame.
  - TCR SSI -- Traditional TruCluster SSI. Complete shared cluster file systems (shared root) and user file systems.

# TruCluster CFS

- Provided in the 11.31 release time-frame
- CFS for user data file systems (no shared root)
  - User data in /cfs
  - Some cluster-wide system and TCR data in /etc/cfs
- Greatly enhances Serviceguard package manager, since data file systems can be accessible from all members
- Private file systems and volume managers
- LAN Cluster Interconnect
- Clusterwide UIDs and GIDs (option)
- Clusterwide PIDs (Legacy and Large PIDs)

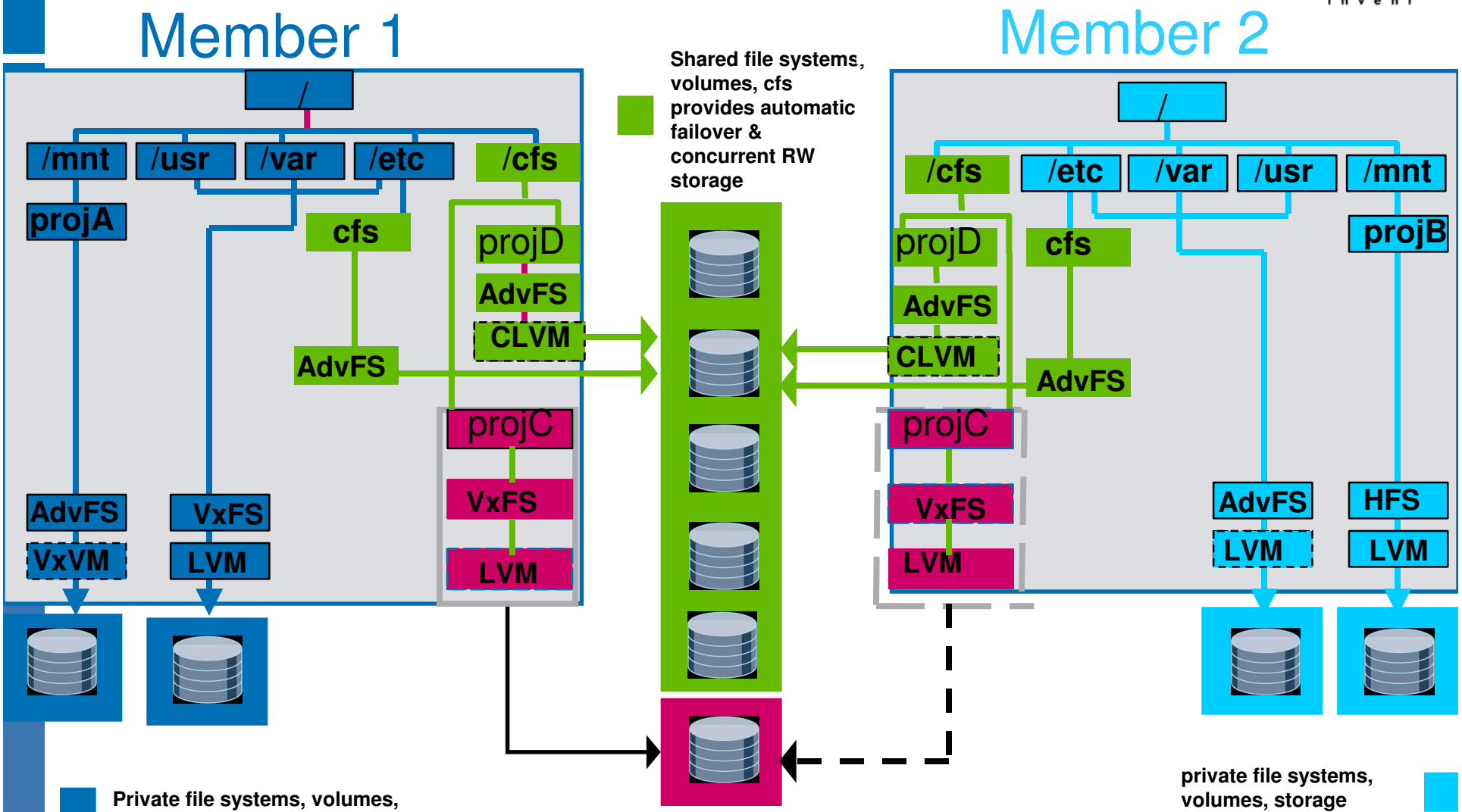
# TruCluster CFS

- Serviceguard Package Manager for Application Availability
  - Supports applications using private file systems, and volume managers as well as CFS
- HA shared storage
- Each member installed and maintained separately (including Applications)

# TruCluster CFS

- User data file systems (AdvFS) can be mounted cluster-wide under /cfs
- DRD can be configured to be cluster-wide for storage devices
- Cluster-wide device namespace (for CFS devices)
- Initially 4 node max (16 CPUs or less); 2 nodes if more than 16 CPUs and initially only supporting IPF

# TruCluster CFS



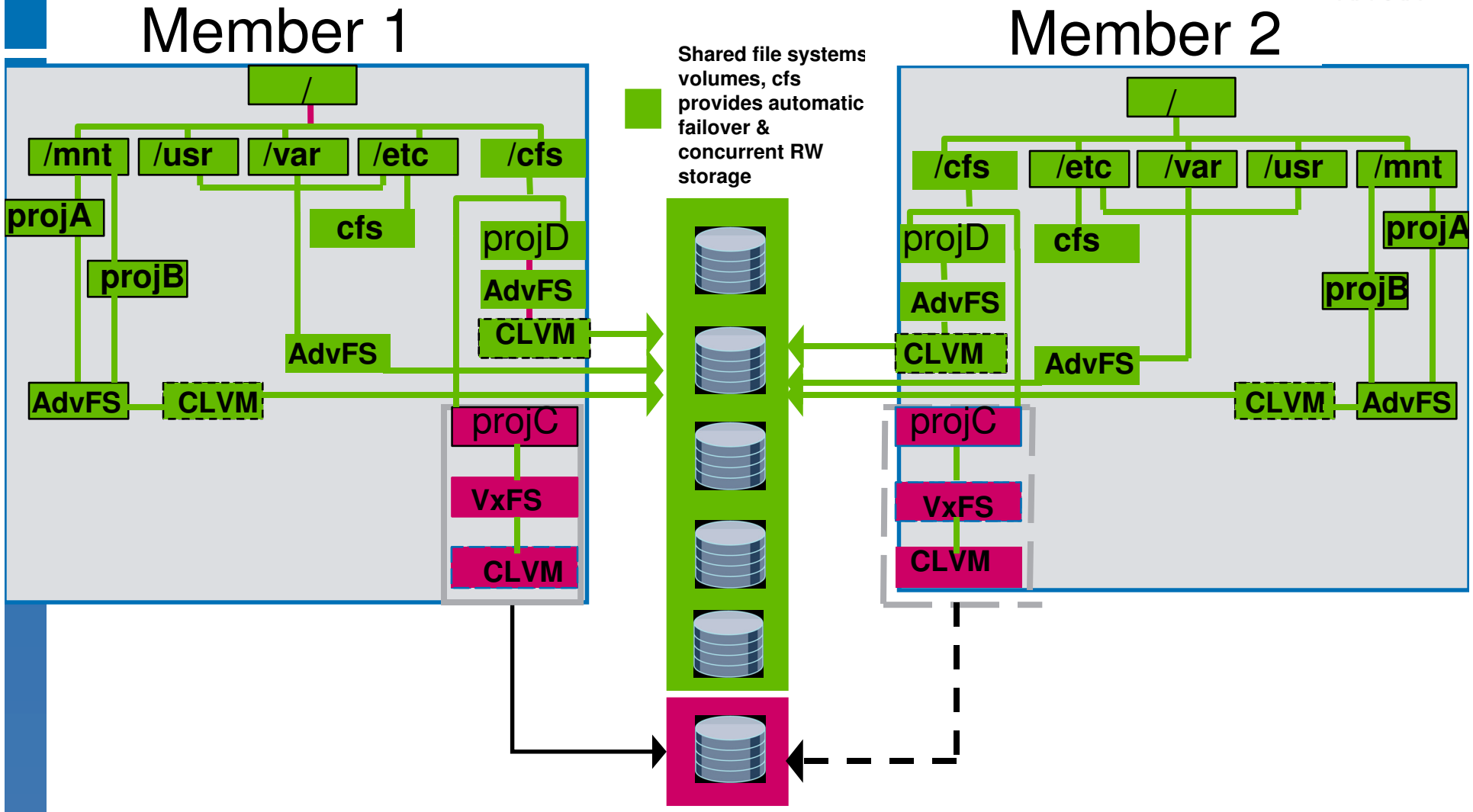
# TruCluster SSI

- Follow-on after initial 11.31 release although there will be some SSI enablers in 11.31
- Shared root option (selectable:  choose CFS or SSI during clu_create)
- SSI Management
  - Management of down nodes
  - Single application installation and management
- Install Once (since shared root)

# TruCluster SSI

- Clusterized LVM
- Clusterized NFS
- Cluster Alias
- Parallel Remote Execution
- VPAR (virtual partition) as a cluster member

# TruCluster SSI




Application and associated server only file system projC configured with VxFS and LVM stack, failover from member 1 to member 2 using SGPM scripting

# TruCluster Roadmap

- 2005 -- 11.31 and TCR CFS
- 2006 -- Add TCR SSI as a choice point (also Cluster Alias, Clusterized NFS, etc.)

# TruCluster – The Big Picture

- TCR CFS and TCR SSI (distinct configuration options)
- Multi-node (4 member) TruCluster Configuration
- Shared Root CFS with AdvFS
- Distributed Lock Manager
- Kernel Group Services
- Membership Manager (AKA Connection Manager)
- LAN as Cluster Interconnect
- Device Request Dispatcher
- Cluster Alias

# TruCluster – The Big Picture

- Cluster-wide events (EVM)
- ICSNET network driver
- Cluster APIs
- Symmetrical storage device configuration based upon 11.31 IO stack
- clu_create, clu_add_member, clu_delete_member
- Cluster File System (CFS) failover
- Application failover
- Cluster Network Adapter Failover

# Cluster File System

- Same view from each member
- Client/Server (in most cases)
  - Each file system or AdvFS domain is served by one cluster member
  - A member can be a client for some domains and a server for others
  - A member can transition between roles transparently
  - Server coordinates caches, meta-data updates
    - Client reads directly from storage, writes through server
- Full X/OPEN, POSIX file system semantics with binary compatibility

# Cluster File System

- Cluster-wide cache coherency
- Transparent file system failover and recovery
- Integrated with Cluster Alias for NFS server

# Member-specific files

- So, how do I keep member-specific configuration info separate?
- CDSLs – Context Dependent Symbolic Link
- For example the /stand directory:

```
lrwxr-xr-x 1 root  sys 26 Jun 30 19:22
  /stand -> ./.cdsl_files/{memb}/stand
```

- {memb} represents the member number and is translated real-time
- Create your own CDSLs with mkcdsl(1M)

# Cluster Management

- clu_create/clu_add_member/clu_delete_member – These work largely as before (Tru64) -- see later slide
- cluamgr – Set up and configure the cluster alias
- clu_get_info – Display cluster information (such as member number, status, etc.
- clu_type – Display cluster type (CFS or SSI)

# Cluster Management

- cmrunpkg, cmmodpkg -- Serviceguard package manager commands control package placement. (Very much like CAA.)
- kctune – similar to sysconfig, display/modify cluster subsystem attributes
- EVM – The event management suite of commands is much the same as on Tru64

# clu_* utilities differences

- The type specified to clu_create determines whether the cluster being created is CFS or SSI
- There will be a complete command-line interface for all three utilities. Both the command-line and .cfg file interfaces are completely noninteractive
- clu_delete_member supports a .cfg file (it did not on Tru64)
- A command-line option allows you to check the correctness of the config parameters without actually performing the clu_* action

# Device Naming

- For now, device naming follows the traditional HP-UX style of /dev/[r]dsk/cXtYdZ.
- The delivered product will have generic device naming such as /dev/[r]disk/diskN.

# Cluster Alias

- View Cluster as a single system
- Can set up multiple aliases
- Transparent node/adapter failover
- Changed commands
  - cluamgr -c option -- cluster; -m -- member(s); -S -- traffic statistics (new);
- Virtual subnets for alias no longer supported (very rarely used)
  - -r changed

# Managing applications -- Serviceguard

- If an application works on a single HP-UX system, it will work (on at least one node at a time) in a Cluster as a single-instance application
- Familiar Serviceguard
- Base services such as Sendmail and Printing

# System Management Interface

- Web-based client for multi-cluster
- “SAM” for single node/single cluster
- Goal is (no surprise) that system management in a Single System Image (SSI) cluster continues to look like a single system. No need to add users on all members, mount file systems on each member, install software on each member, etc.

# Product Features by Configuration

| Feature | TCR CFS | TCR SSI |
|---|---|---|
| Cluster-wide file systems | Customer selects which user-data file systems use CFS. AdvFS is required for file systems under CFS use. Small set of system files also use CFS. | All file systems, including the root, are CFS. AdvFS file systems are highly available with CFS. |
| Cluster-wide device names | All storage devices have cluster-wide location independent device names | Same |
| Cluster-wide device access | Administrator selects subset of storage for cluster-wide access | All storage devices have cluster-wide access |
| Cluster alias | One alias by default. Customers can configure multiple aliases. | Same |
| Application failover with SGPM | Serviceguard packages can failover between members and can be configured to use any HP-UX supported file system/volume manager pair. | Serviceguard packages can failover between members and can be configured to use any file system and CLVM. |

# Management Capabilities by Configuration

| Feature | TCR CFS | TCR SSI |
|---|---|---|
| System management | OS management must be performed per member. | OS management is generally applied to all cluster members. |
| Management of down members | Management operations must be applied to down member when they next boot. | Since most management operations apply to the cluster, down members inherit the changes automatically when they boot. |
| Installation of software | Software must be installed on each member. | Software is installed once for the entire cluster. |

# Hardware/Software configurations

| Feature | TCR CFS | TCR SSI |
|---|---|---|
| Platforms | IA for 11.31 time-frame. | IA after 11.31 |
| Volume managers | Support for VxVM and LVM for non-CFS file systems in 11.31. LVM support for CFS as a follow-on. | Clusterized LVM supported. |
| File Systems | All HP-UX supported file systems are available for private access. AdvFS can be mounted for cluster-wide access. | All HP-UX supported file systems are available as server_only (access by only one member at a time). AdvFS file systems are cluster wide. |
| Storage Options | FC for physically shared storage. Any HP-UX supported storage privately attached. | Same |
| Interconnects | LAN (Gb) with IB to follow.<br>IB will be supported for RAC messaging. | Same |
| Node count | 4 nodes in 11.31. | Same |

# AdvFS

# AdvFS – The Big Picture

- Logging and Transactions
- POSIX functionality
- AdvFS tools integrated into HP-UX toolset
- Support for HP-UX root and boot file sytems (/, /stand, /etc, …)
- Direct I/O
- B+ tree directory indices
- Integration with HP-UX patented Read Ahead Solution

# AdvFS – The Big Picture

- Multiple Volume File System Management – including dynamic volume add and remove
- Online file system expansion (extendfs, fsadm) as a result of underlying volume/LUN expansion
- File Defragmentation and Migration
- VFAST Automated Policy Engine
- Freeze/Thaw in support of array snapshots and snapclones

# AdvFS – The Big Picture

- Performance
  - Online defragmentation
  - Rebalancing
  - Self-tuning (hot file, balancing, defragmenting, etc.)
  - B+ tree index
  - DirectIO
- Sizes (16TB single file; 512TB file system)
- Serviceability (monitoring, fine grain fault isolation, on-line verification)

# AdvFS – Some differences

- No concept of a fileset on HP-UX (correspondingly, the fileset-related commands/switches are gone)
  - This means no multi-fileset domains
- No frag file
- No file striping

# AdvFS – Some differences

- Unified Buffer Cache (UBC) is replaced with Unified File Cache (UFC)
- No clonefset, use hardware cloning/snapshots instead
- /etc/fdmns is replaced with /dev/advfs
- /dev/advfs
  - default (block device)
  - Link to storage is actually in the .stg subdirectory

# AdvFS command (changes from Tru64)

| | |
|---|---|
| • addvol | fsadm addvol |
| • rmvol | fsadm rmvol |
| • advscan | fsadm scan |
| • chfsets | fsadm chfs |
| • chvol | fsadm chio |
| • defragment | fsadm defrag |
| • migrate | fsadm migrate |
| • mount -o extend | fsadm extend |

# AdvFS command (changes from Tru64)

| | |
|---|---|
| • savemeta | fsadm savemeta |
| • showfdmn | fsadm (-F advfs) |
| • showfile | fsadm getattr |
| • switchlog | fsadm mvlog |
| • vfast | fsadm autotune |

# AdvFS command (changes from Tru64)


| | |
|---|---|
| • advfsstat | advstat |
| • diskusg | advdiskusg |
| • salvage | advsalvage |
| • mktrashcan | advtrashcan -m |
| • rmtrashcan | advtrashcan -r |
| • shtrashcan | advtrashcan |
| • nvbmtpg,nvlogpg nvtagpg,vfilepg vsbmpg | advvods (various switches) |

# AdvFS command (changes from Tru64)

| | |
|---|---|
| • rvdump | advrdump |
| • rverstore | advrrestore |
| • vdump | advdump |
| • vrestore | advrestore |

# AdvFS command (changes from Tru64)

| | |
|---|---|
| • fixfdmn | fsck |
| • verify | fsck |
| • mkfdmn | mkfs/newfs |
| • rmfdmn | fsadm rmfs |
| • mountlist | mount |
| • tag2name | ncheck |

# AdvFS/Cluster Demo

- AdvFS multi-volume functionality
- Cluster-wide mount
- Add a user
- Install software (Apache – multi-instance)
- Cluster Alias (distribution of connections)
- Oracle (single-instance)

# Resources

- Session 3849 -- 08/16/200405:15PM “Nuts and Bolts of Enhanced Security Management for Tru64 UNIX”
- Session 3847 -- 08/18/2004 at 11:00AM “Best Practices for Patching and Upgrading Tru64 UNIX and TruClusters
- Session 3761 -- 08/19/2004 at 01:30PM “A Side-by-Side Comparison of Tru64 UNIX and HP-UX 11i v2 Hardware Management”
- Session 3878 -- 08/20/2004 11:00AM “Tru64 UNIX v5.1x TruCluster Recovery Techniques”
- http://hp.com/go/hp-ux

# **Hear more** about our service offerings by visiting us in the **Solutions Showcase**!

- HP Web Support Tools
- HP Active Savings Tool
- HP Education
- HP Business Continuity & Availability Solutions
- HP Adaptive Enterprise Agility Assessment
- HP Financial Services
- HP IT Consolidation Solutions
- HP Radio Frequency Identification (RFID)

# HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

RECOMMENDED TRAINING VENUE FOR THE

**HP Certified Professional**

# Extra slides

# General HP-UX Information-gathering utilities

- machinfo (IA-only) – display information about CPUs, firmware, memory, etc.
- tddiag – display tons of configuration data… tons (this utility ships with the Tachyon FC card)
- fcmsutil (/dev/td0) – get detailed information about your fiber channel
- kctune | grep -E "advfs|cfs|clua|cnx|ics|kgs|drd|clubase" – display settings of various cluster (and file system) settings

# Devices -- Troubleshooting

- cfgv,cfgc,cfgs -- display/modify the devices on a system (to see the disk devices cfgv -r component -c disk). Similar to hwmgr on Tru64.
- dfsv,dfsc,dfss -- display/modify device names. Similar to dsfmgr in Tru64.
- lssf -- get bus, target, LUN, path info for a particular device (lssf /dev/dsk/c4t0d3). Similar to the file command in Tru64.
- diskinfo -- provide some useful information about the disk, like the size (diskinfo /dev/rdsk/c10t0d6)
- fstype -- display the file system type (fstyp /dev/rdsk/c10t0d6s2)

# Devices -- Troubleshooting

- devnm -- find out what storage is associated with a particular file system (devnm . Or devnm /var). If the underlying file system is AdvFS, you'll get the domain name (not the actual devices).