OpenVMS Cluster Interconnect and Feature Update

Andy Schneider
OpenVMS Business Management
Hewlett-Packard

Agenda

- OpenVMS Cluster Interconnects
  - Status of existing Cluster interconnects
  - Future Cluster interconnects
- OpenVMS Clusters on Integrity Servers
  - Mixed-Architecture Configurations
  - Evolution of Cluster support
- New Features
  - Host-Based MiniMerge
  - Other new features

Cluster Interconnect Update

- What is a cluster interconnect?
- Existing interconnects
  - Dual-Purpose (SCS plus Storage)
  - SCS-Only
  - Storage-Only
- What are the interconnect futures

What is a Cluster Interconnect?

- **Connection between multiple systems in an OpenVMS cluster, performing one of the following functions**
  - System Communication Services (SCS) traffic
    - MSCP/TMSCP Data Serving, Lock Manager, Intra-Cluster Communications
  - Shared Direct Storage Access
    - Disks, Tapes, SAN's
- **Connections can be either**
  - Local distance (measured in meters)
  - Long distance (measured in kilometers)
    - "Disaster Tolerant"
- Depending on the Specific Interconnect, the connection can do SCS only, Storage Only, or Both!

Dual-Purpose Cluster Interconnects Today

- **DSSI**
  - End Of Life over 5 years ago
  - No support for new AlphaServers
- **CI**
  - Still embraced by some customers for SCS traffic
  - Star Coupler, CIPCA/CIXCD's still being sold, but…
    - **CIPCA EOL Buy made – less than 2 years coverage**
    - **CIXCD's, Star Coupler - near exhaustion of inventory**
  - HSJ80 "kicker" for CI storage in 2000
  - Local-distance limitation for interconnect
    - **Won't be moving CI to Integrity - Alpha platform support only!! Must use MSCP serving to see CI storage on Integrity servers**

SCS-Only Interconnects Today

- **Network Interconnects**
  - 10/100 Ethernet
    - Continued support on all new servers
  - Gigabit Ethernet
    - Preferred Cluster Interconnect Today
    - High Bandwidth Potential - large packets, distance
    - New DEGXA is vast improvement over previous DEGPA
    - Primary SCS Interconnect for use on Integrity Servers
  - FDDI/ATM
    - Limited use today, typically in DT configurations, not on Integrity
- **Memory Channel**
  - MC2 supported today; Distance to 3km w/Fibre
  - No support on Integrity Servers
  - **End of Life extended due to large buy to extend Tru64 MC support - several years supply.**

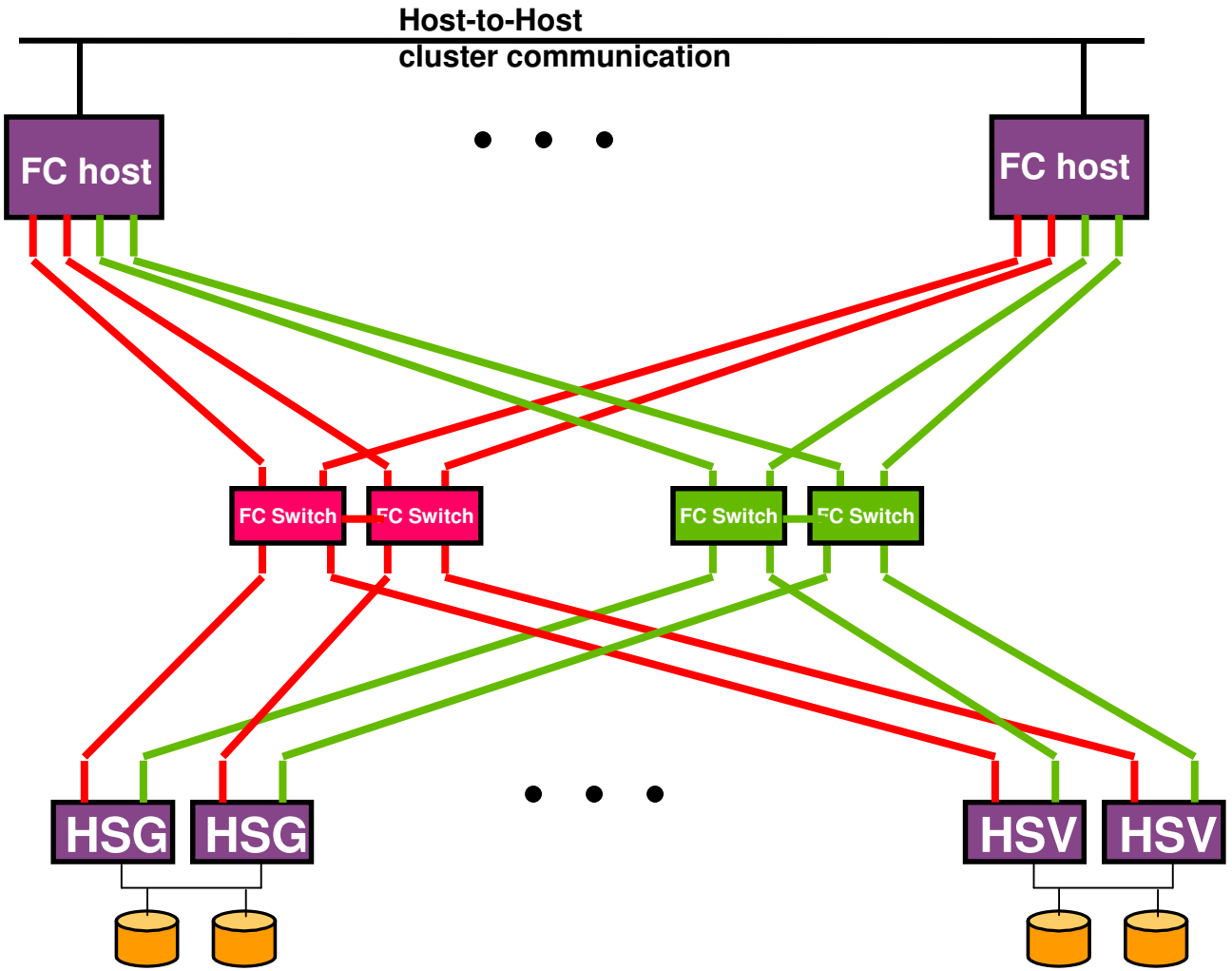
Storage-Only Interconnects

- **Parallel SCSI**
  - UltraSCSI multi-host support on Alpha today
    - 4 nodes w/5-port hub
  - HSZ22/HSZ70/HSZ80 Solutions Supported (RA7000,etc)
  - Future SCSI support by OpenVMS will NOT be multi-host capable
    - **Backplane RAID** (SmartArray 53xx/64xx generation)
      - Supported on both Alpha and Integrity Servers
    - Low-end external SCSI-3,4 Solutions
      - Available on both Alpha and Integrity Servers
        - Utilize core I/O integration where possible
      - Direct-Attached ONLY!

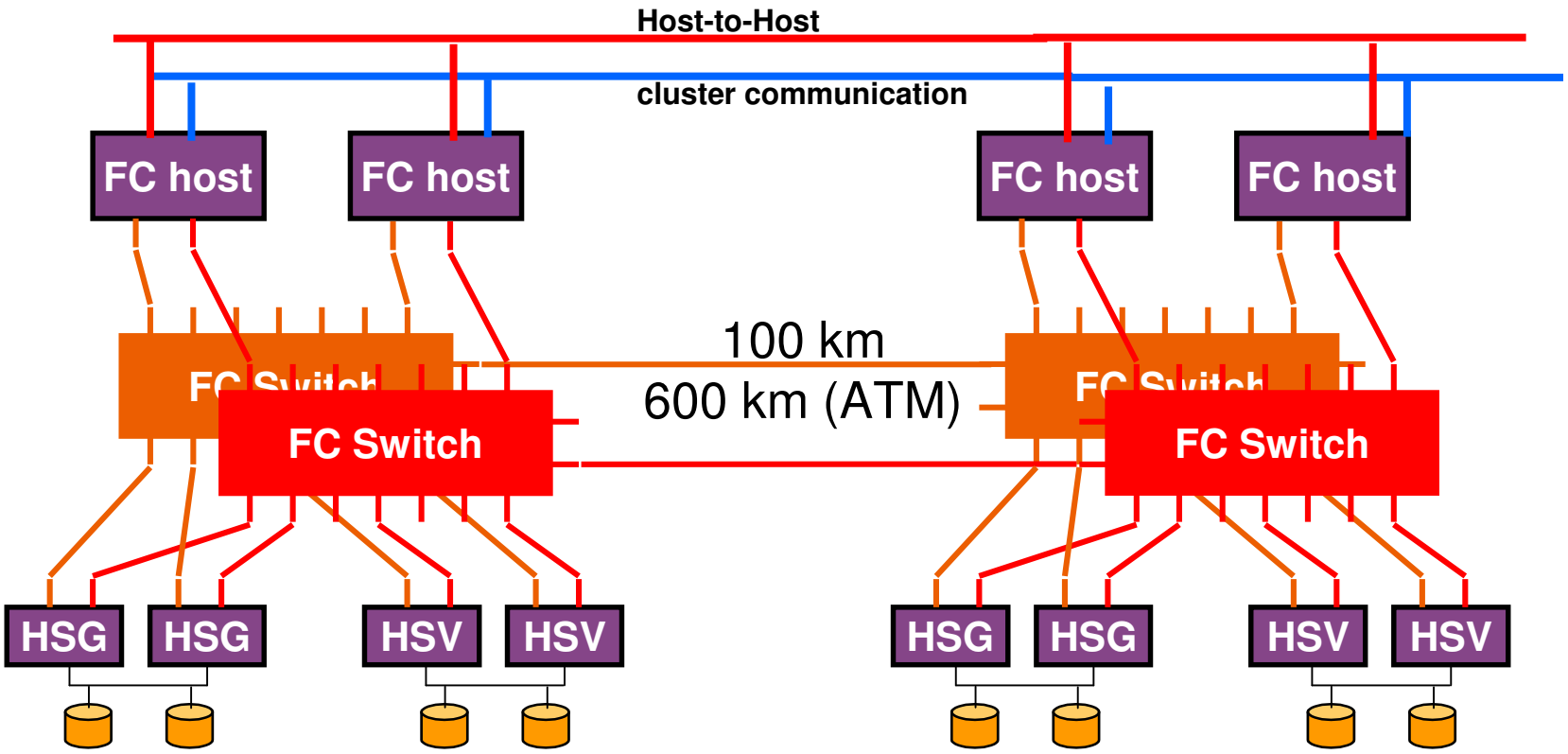
Storage-Only Interconnects (cont.)

- **Fibre Channel Switched Fabric**
  - Fully supported today w/V7.2-1 and above
  - Supports up to physical limits of configurations
  - Is the basis for all HP SAN Future offerings
  - Support for Long-Distance direct storage access
    - DRM/CA - Progressing technology with OpenVMS helping to drive requirements
    - Shadowing - full support in V7.3-1 w/MultiPath failover to served path, and Host-Based MiniMerge (Q3 CY2004)
  - Infrastructure can be complex and expensive, but MAY be able to handle more than just storage

Typical Fibre Channel-based Cluster



Disaster Tolerance - A Multi-site FC Cluster



Cluster Interconnect Futures

“FC as a LAN” project - planned for delivery in V8.3 (2005) on Alpha and Integrity

FEATURES:

- Utilization of existing FC adapters as LAN devices to do SCS (host-to-host) communications as well as Storage
- Will use existing FC infrastructure and coexist with storage traffic - just as CI does today!
- **Target is support for DT environments** where single pipe is important for both SCS and Storage.
- Local distance configurations should still use dedicated SCS interconnects (Gigabit Ethernet) for greater performance.

- Last major AlphaServer cluster interconnect planned.
  - Future work here on Integrity Servers only

V7.3-2 Cluster Interconnect Performance Considerations

- Gigabit Ethernet (DEGXA) w/Jumbo frames is currently highest performance LAN interconnect.
- FibreChannel LAN emulation best used where cost savings are more important than latency & CPU utilization.
- FC LAN Application Examples:
  - Disaster Tolerence:
    - 2 inter-site links instead of 4 is an enormous cost savings.
    - Inter-site link delays are dominated by media latencies.
  - Hot-standby links for data center redundancy - Only use existing FC infrastructure when LAN infrastructure fails.

Integrity Server Cluster Interconnect Possible Futures:

- New Low-Latency Cluster Interconnect (SCS) Investigation:
- Investigation based on reviewing HP-UX cluster interconnect strategy & plans as well as general industry trends.
- OpenVMS Requirements:
  - In general, surprisingly close to HP-UX requirements (eg: industry standard, low-latency, high-bandwidth, “intelligent”...
  - Add’l OpenVMS unique requirements/goals:
    - IP routeable (Strong customer need is recognized.)
    - Minimize risk of another Memory Channel (if possible).
    - Can do only 1 major interconnect project at a time (resource reality).

– HP-UX roadmap has ServiceGuard using Infiniband late 2004, and integrated TruClusters using Infiniband and iWARP technologies starting in late 2005

Infiniband and iWARP Technical Comparison

Feature	IB	iWARP
IP routeable:	Future(via IP switch port)	Day 1 (it's IP!!!)
Latency(now):	~6.5 uS	~10 uS (faster w/ 10Gb when SCTP replaces TCP
Physical Media:	Unique Parallel Copper ->$$	It's GbE: CAT-5 & fiber
Fabric:	Unique -> $$	Commodity (Ethernet)
Distance:	Tbd (via IP when sw w/ IP ports avail)	The world
Speed		
Industry Std API:	IT-API (User & Kernel)	IT-API (User & Kernel) w/ iWARP extensions.
Code Size	55,000 lines of C	TBD

Infiniband and iWARP Viability Comparison

Item	IB	iWARP
Market Position:	High End: Max. Perf @ a price	Commodity: High Perf LAN & SAN
Expected Volumes:	Low (HPTC, Mid-High End clusters)	High-Vy High (Clusters, iSCSI, TCP offload,…)
Price:	Very High due to Fabric,	High, but potential for rapid decrease w/ volume
# Component Suppliers:	Few (2-3), all smaller houses.	Many: All major NIC vendors
Systems Users:	hp, SUN, IBM, Windows	hp, SUN, IBM, Windows, Linux
Applicability beyond clustering:		-iSER (iSCSI Enabled for RDMA) - TCP offload - NFS over RDMA - Sockets Direct - growing

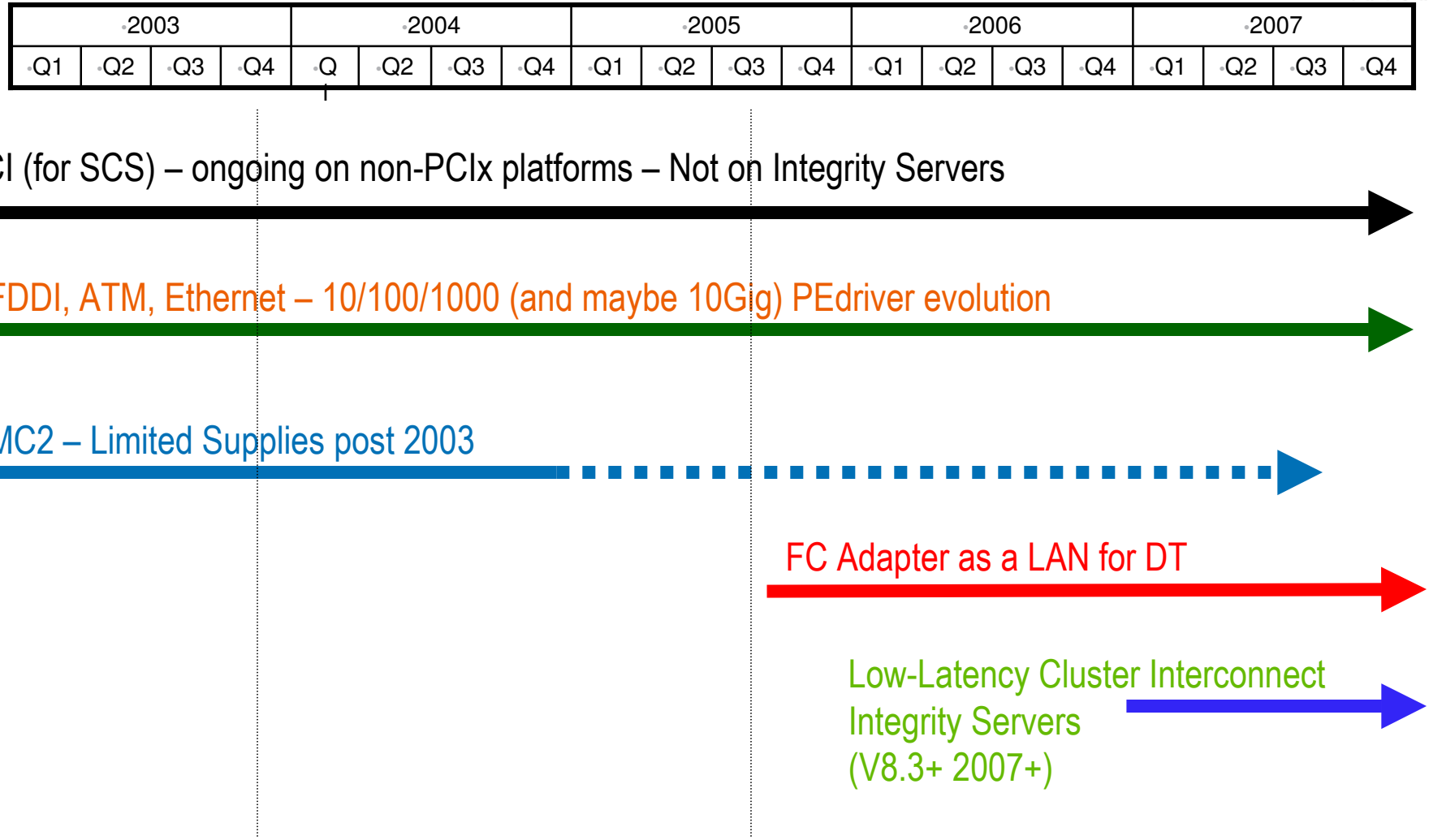
Infiniband v. iWARP – What makes sense for OpenVMS?

- Infiniband has performance going for it, but is limited to a cluster interconnect model (similar to Memory Channel)
- iWARP has inherent IP routability, long-term IP application capabilities, good performance, long distance, & the strongest viability metrics.
- Doing either is going to be a big project, most likely OpenVMS support is '07.
  - Picking one would not preclude doing the other later.
- HP-UX is going with iWARP as longer-term solution after Infiniband.
- OpenVMS Target - Go with iWARP

Preliminary Investigation – Not Committed as an OpenVMS Project at this point in time!!!

Q: What is the need for a Cluster Interconnect of this caliber given that Gigabit Ethernet exists?

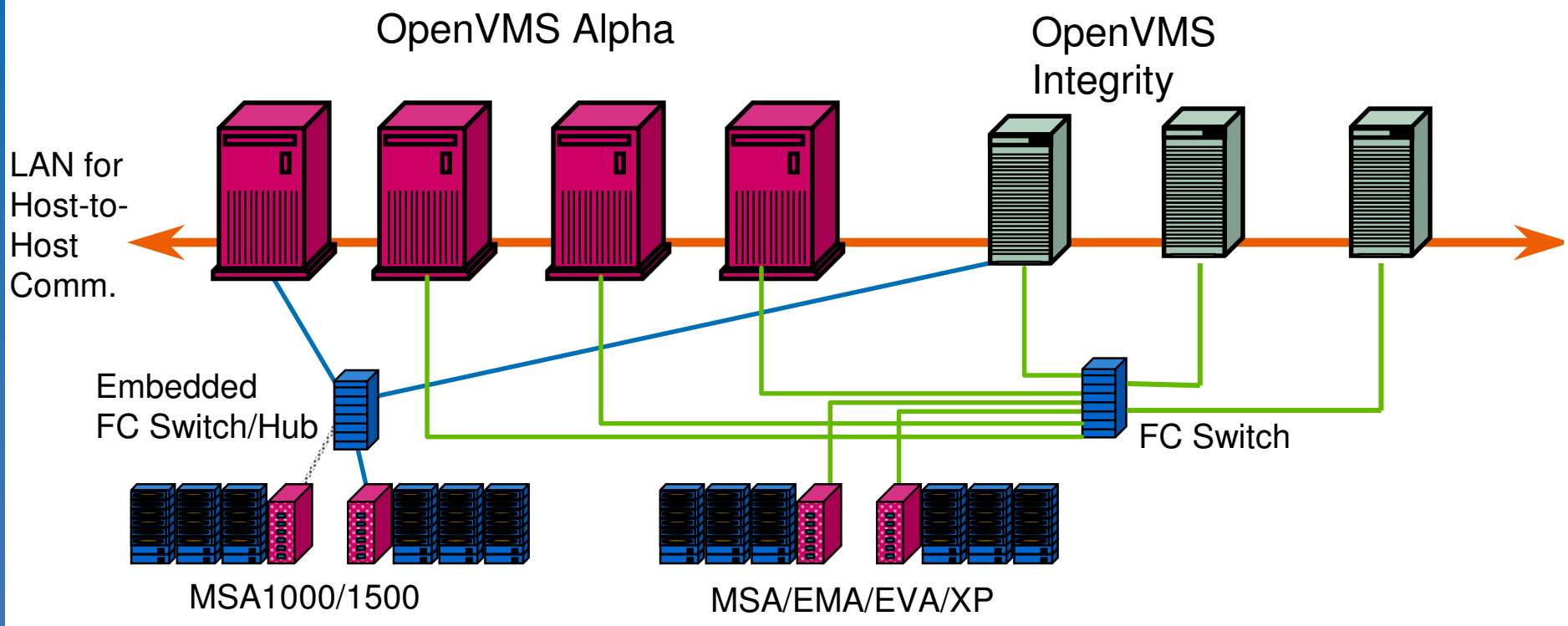
OpenVMS Interconnect Roadmap



OpenVMS Clusters on Integrity Servers

Storage in a Mixed-Architecture OpenVMS Cluster Environment

•Using Current-Generation FC Storage



Departmental Fibre Channel Storage

8 Node Limit (replaces multi-host SCSI)
2 Node for FCAL (release TBD)

Enterprise Fibre Channel Storage

8 Nodes MSA, 8-96 nodes others

OpenVMS Integrity Option Rollout Plan

2004

Version 8.2

- **Storage**
  - U160 & U320 SCSI (non-shared)
  - Fibre Channel
  - SmartArray Backplane RAID
- **Storage Solutions**
  - HSG, EVA, MSA1000, XP
- **LAN**
  - 5701(10/100/1000)
  - 8255x (10/100/1000)
- **Graphics**
  - ATI Radeon 7000 equiv.
  - ATI Radeon 7500/3D
- **Audio (**tbd)

2005

Version 8.3

- **Storage**
  - U160 & U320 SCSI (non-shared)
  - Fibre Channel
  - SmartArray Backplane RAID
- **Storage Solutions**
  - HSG, EVA, MSA1000, XP
- **LAN**
  - 5701(10/100/1000)
  - 8255x (10/100/1000)
- **Graphics**
  - Yes (tbd)
- **Audio** (tbd)

OpenVMS Integrity Clusters Rollout Plan

Mid 2004

Version 8.2 Field Test

- 2-4 Node Clusters
- Cluster Interconnect
  - Gigabit Ethernet
  - 10/100 Ethernet

Q4 2004

Version 8.2

- Up to 16 Nodes
- Expanded Functionality
- Cluster Interconnect
  - Gigabit Ethernet
  - 10/100 Ethernet

V8.2 (Warranted)
V8.2 & V7.3-2 Alpha (Warranted)

2005

Version 8.3

- Full Functionality
- Cluster Satellite Boot
- Cluster Interconnect
  - Gigabit Ethernet
  - 10/100 Ethernet
  - Fibre Channel LAN
  - 10 Gigabit Ethernet*
  - Next-Generation Interconnect – post V8.3

V8.x (Warranted)
V8.x & V8.2 & V7.3-2 Alpha (Migration)

* - If HP Program is Ready

Host Based Volume Shadowing Features

- *V7.3-2 Features*
- *Host Based MiniMerge*

V7.3-2 HBVS Projects

Dynamic Volume Expansion (DVE)

- Storage arrays today can dynamically expand volumes
- OpenVMS could not utilize the expanded volume without re-initializing
- OpenVMS dynamic volume expansion allows allocation of extra bitmap space at init time and then later enables volume size expansion while the device is mounted
  - $ init /limit 1dga100: name *! Allocates 1TB bitmap*
  - $ mount/cluster 1dga100: name *! Mount and use volume*
  - *(expand physical volume size using management tools)*
  - $ set volume 1dga100: /size=xxxxxx *! Set new size*
- For volumes already initialized
  - $ dismount/cluster 1dga100: *! Dismount volume*
  - $ mount 1dga100: name *! Privately mount volume*
  - $ set volume 1dga100: /limit *! Allocate new bitmap*
  - $ mount/cluster 1dga100: name *! Cluster mount volume*
  - *(expand physical volume size using management tools)*
  - $ set volume 1dga100: /size=xxxxxx *! Set new size*

V7.3-2 HBVS Projects

Dissimilar Device Shadowing (DDS)

- Prior to V7.3-2, shadowed volumes had to be identical in size
  - This made it impractical to shadow between different types of storage arrays
- Different sized volumes can now be shadowed
  - The size of the smallest volume determines the size of the shadow set
  - The shadow set can only be mounted on V7.3-2 nodes
  - The size of the shadow set can be grown using DVE

V7.3-2 & V8.2 – Host Based MiniMerge

What is Mini Merge?

- Full merge requires comparing entire shadow set
  - But only blocks with I/O in progress need to be merged
- MSCP Based mini merge
  - Supported on HSCxx / HSDxx / HSJxx controllers
  - Controller tracks in-progress writes
  - Host can get list of writes from the controller
- Host Based Mini Merge (HBMM)
  - Utilizes Write BitMap capability in OpenVMS
  - Selected cluster hosts track recent writes using write bitmap
  - Bitmap is periodically reset to flush out old writes
  - Contents of bitmap drive mini merge operation

V7.3-2 & V8.2 – Host Based MiniMerge

Write Bitmap for Mini Merge

- Write bitmap originally released on V7.2-2 for mini copy
  - Each system has an in-memory bitmap to track write I/O
  - Each bitmap has one system that is the master
  - 2KB memory per Gbyte of storage per bitmap per system
- There are 6 bitmaps are available (per VU) for mini merge use
  - There are also 6 bitmap slots reserved for mini copy use
- HBMM recovery must be by a bitmap master system
- Use multiple bitmap masters for availability of the bitmap after a system crash

V7.3-2 & V8.2 – Host Based MiniMerge

Host-Based Mini Merge Project

- Project began in mid-2003 to produce host-based MiniMerge capability for Fibre Channel
  - Rather than rely upon history-logging from FC controllers
- Currently in final Beta test of **host-based** MiniMerge solution, based on Write Bit Map technology (from MiniCopy)
- Will support **ALL** FC types (HSG, HSV, MSA, XP)
- Does NOT require any storage firmware assists

V7.3-2 & V8.2 – Host Based MiniMerge

Host Based Mini Merge Release Plans

- Version 7.3-2 as a remedial kit
  - V73-2R-HBMM
  - Estimated Availability in Aug/Sept 2004
- Version 8.2 (both Alpha and Integrity Servers)
  - Integral part of the Volume Shadowing capability in this release
- No plans for VAX support
- No plans for backport to releases earlier than V7.3-2

V7.3-2 PE Driver FastPath Tuning Recommendations

- Keep PE & LAN drivers off of the primary CPU.
- Ensure that high CPU load devices are assigned to different CPUs:
  - TCP/IP
  - PEdriver
  - FibreChannel adapters
  - Gb LAN adapters
- Try to keep devices sharing the same driver image on the same CPU
  (Keeps driver code & common data structures in single cache.)

V7.3-2 PE Driver Lock/Latency Tuning Recommendations

- Try to minimize # of CPU-CPU work transfers by such things as:
  - IF dedicated lock manager's CPU isn't near saturation, assign PEdriver to the dedicated lock manager CPU. Assign LANdrivers to a different CPU
  - IF dedicated lock manager's CPU is near saturation, Assign PE & LAN drivers to the same CPU, or LAN on 1 cpu, PE on another.

OpenVMS Cluster Roadmap

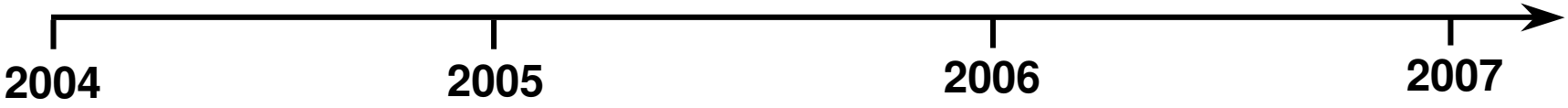
Version 7.3-2
- Dynamic Volume Expansion
- Dissimilar Device Shadowing (DDS) support
- Cluster Performance Enhancements
- PE Driver Fastpath

V8.2 - Mixed-Architecture Alpha/Integrity Clusters
- Add-in new systems to existing clusters!
- Tap into existing FC storage fabrics

Mini Merge Support for HBVS
- New Host-Based MiniMerge
  - V8.2 Release, V7.3-2 Remedial Kit
  - Will support HSG, HSV, MSA, XP

New Cluster Interconnects (Post V8.2)
- LAN over FC Adapter (DT Capabilities)
- Next-generation Low-Latency Interconnect (Integrity Servers Only)

OpenVMS Storage HW Roadmap



MSA Storage Family
- MSA1000 Switched Support during H2 2003
- MSA1000 2-node Arbitrated Loop late 2004
- Support of new MSA solutions during 2004+

EVA Storage Family
- Support new Continuous Access version mid-2004
- Support new EVA product array late 2004

XP Storage Family
- Ongoing support of larger cluster configs
- Support of broader XP Software tools 2004
- Support of updates to XP Array offerings 2004

Storage Adapter Support
- U320 Backplane RAID (SmartArray) H2 2004
- LP10000 Dual Channel FC – Q2 CY 2004

OpenVMS Integrity Storage Support (CY 2004 and beyond)
- FC support to connect into existing SANs at V8.2
  - EVA, MSA, EMA and XP Solutions
- SmartArray support identical to Alpha
- Direct-Attached SCSI support ONLY for internal and non-shared external storage
- Utilize adapters that will be supported by multiple OS environments

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