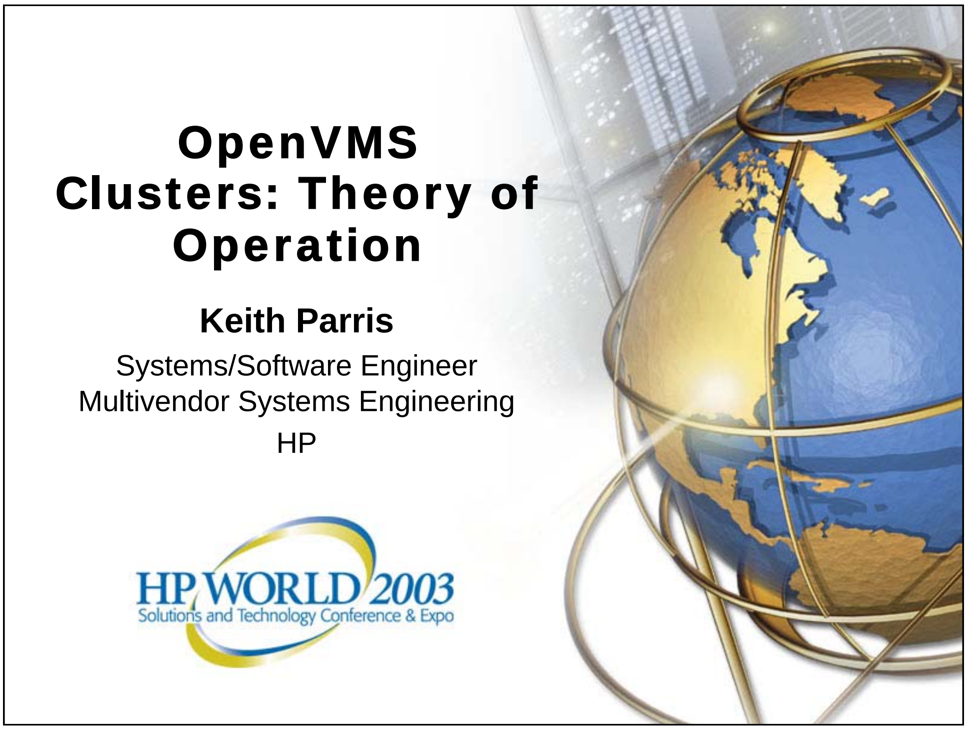
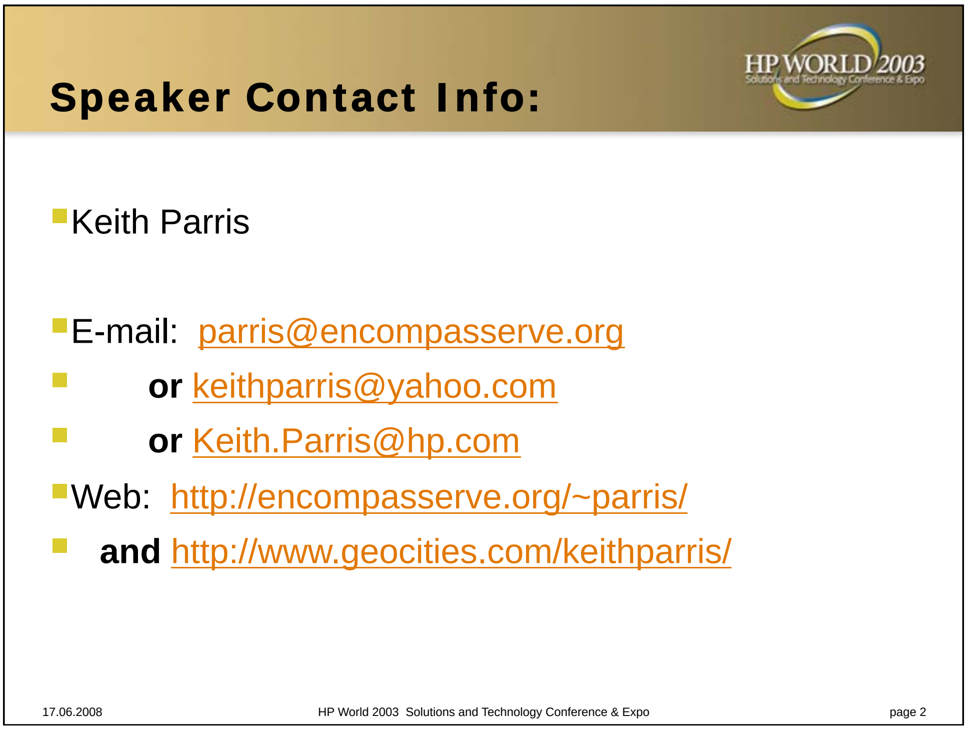


OpenVMS Clusters: Theory of Operation

Keith Parris

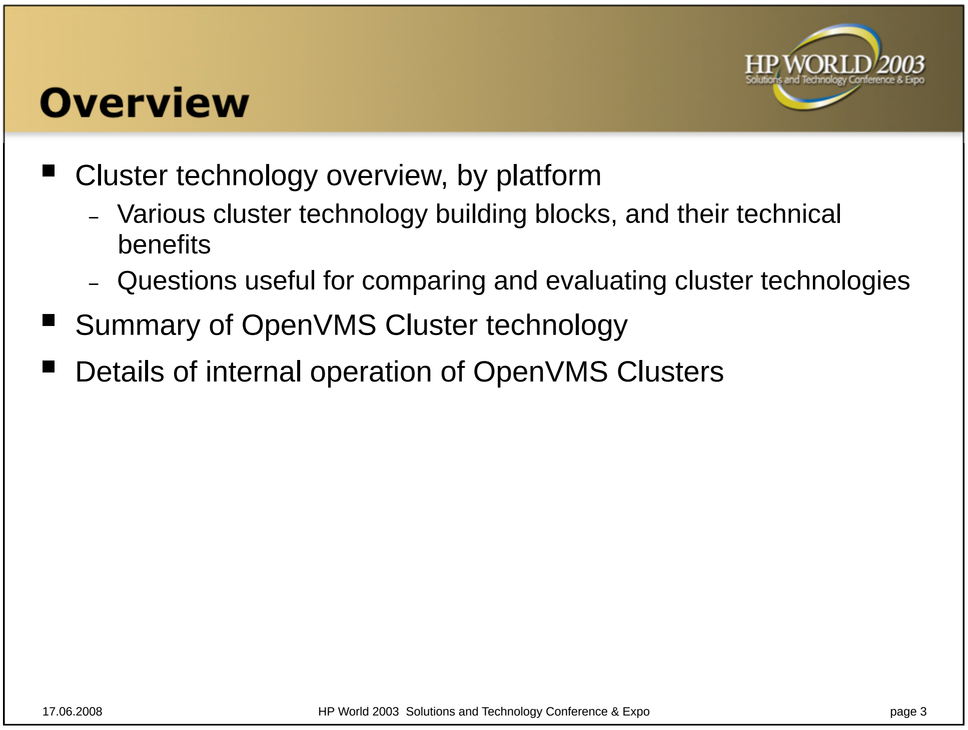
Systems/Software Engineer
Multivendor Systems Engineering
HP

HP WORLD 2003
Solutions and Technology Conference & Expo



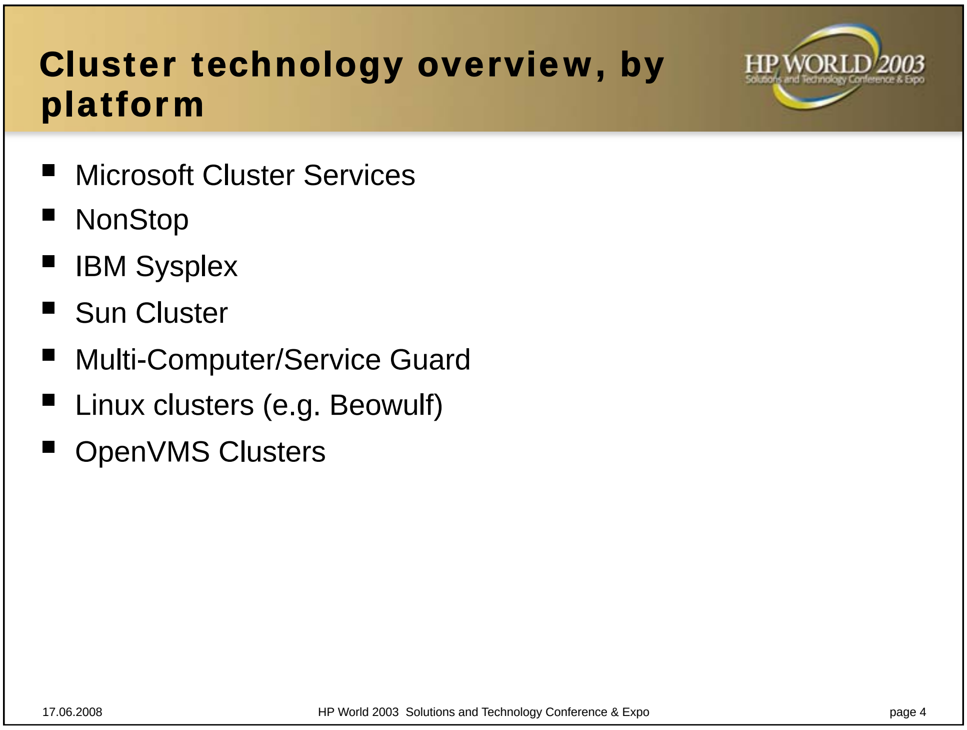
Speaker Contact Info:

- Keith Parris
- E-mail: parris@encompasserve.org
- **or** keithparris@yahoo.com
- **or** Keith.Parris@hp.com
- Web: http://encompasserve.org/~parris/
- **and** http://www.geocities.com/keithparris/



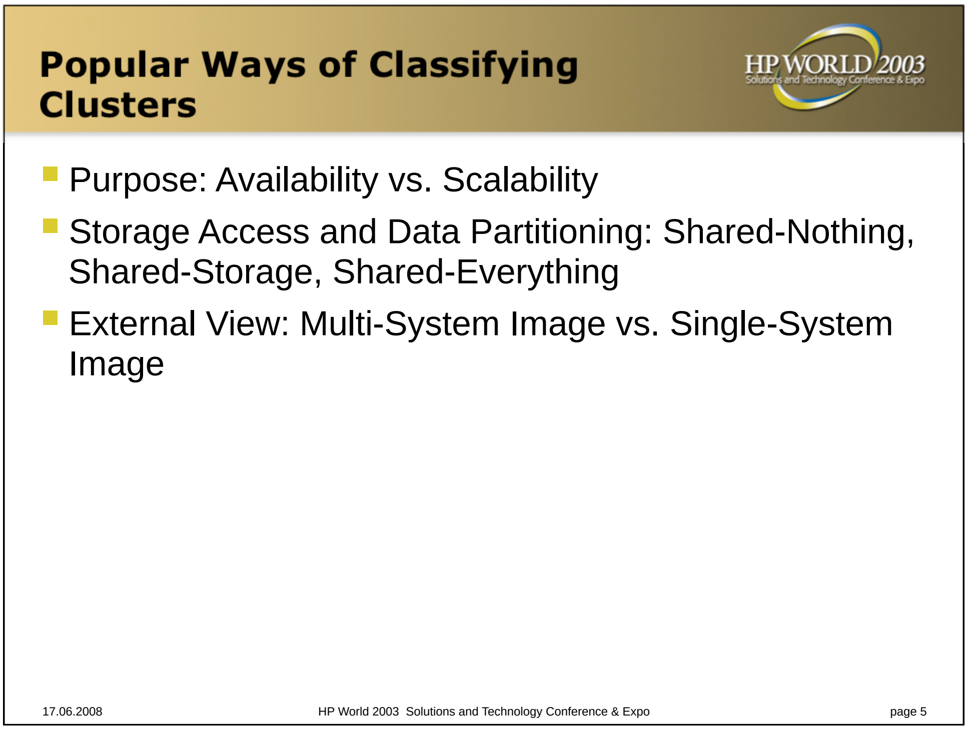
Overview

- Cluster technology overview, by platform
  - Various cluster technology building blocks, and their technical benefits
  - Questions useful for comparing and evaluating cluster technologies
- Summary of OpenVMS Cluster technology
- Details of internal operation of OpenVMS Clusters



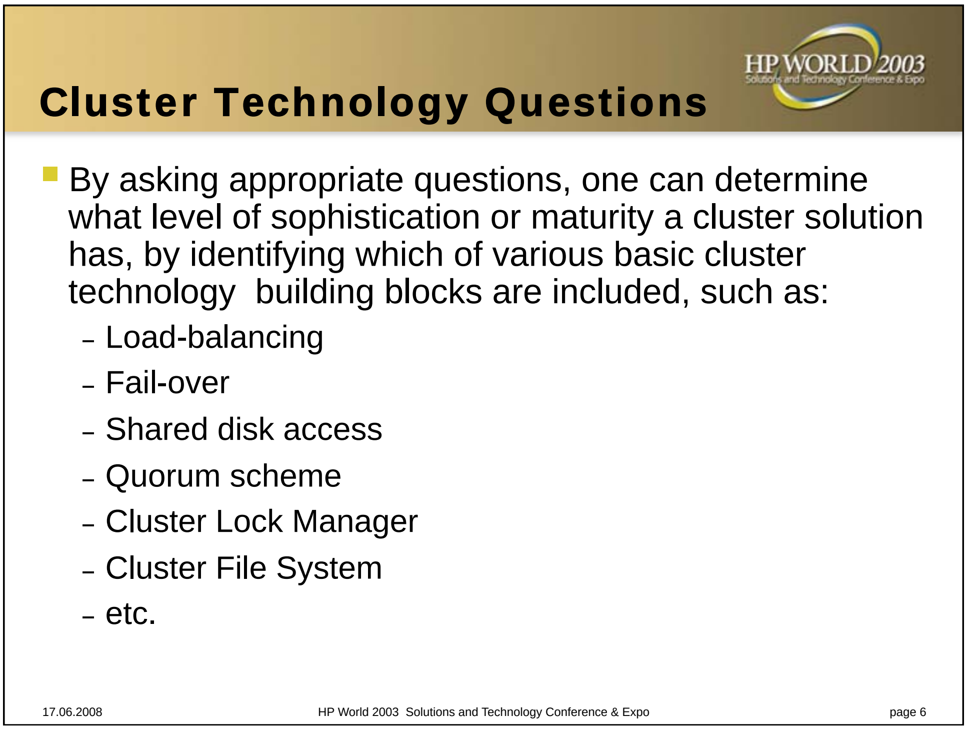
Cluster technology overview, by platform

- Microsoft Cluster Services
- NonStop
- IBM Sysplex
- Sun Cluster
- Multi-Computer/Service Guard
- Linux clusters (e.g. Beowulf)
- OpenVMS Clusters



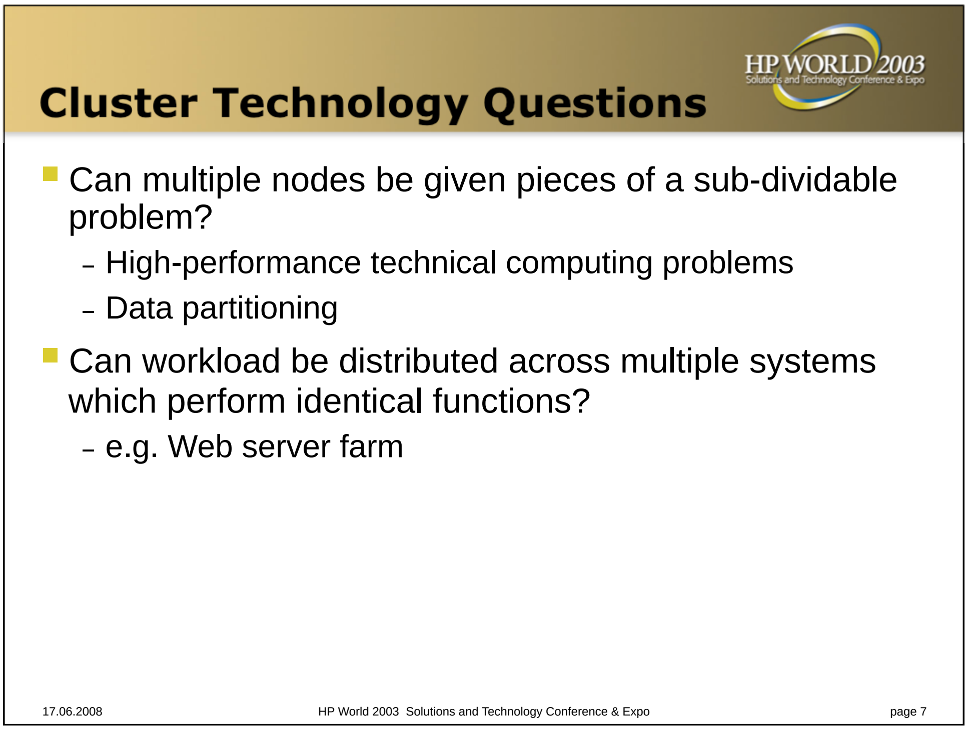
Popular Ways of Classifying Clusters

- Purpose: Availability vs. Scalability
- Storage Access and Data Partitioning: Shared-Nothing, Shared-Storage, Shared-Everything
- External View: Multi-System Image vs. Single-System Image



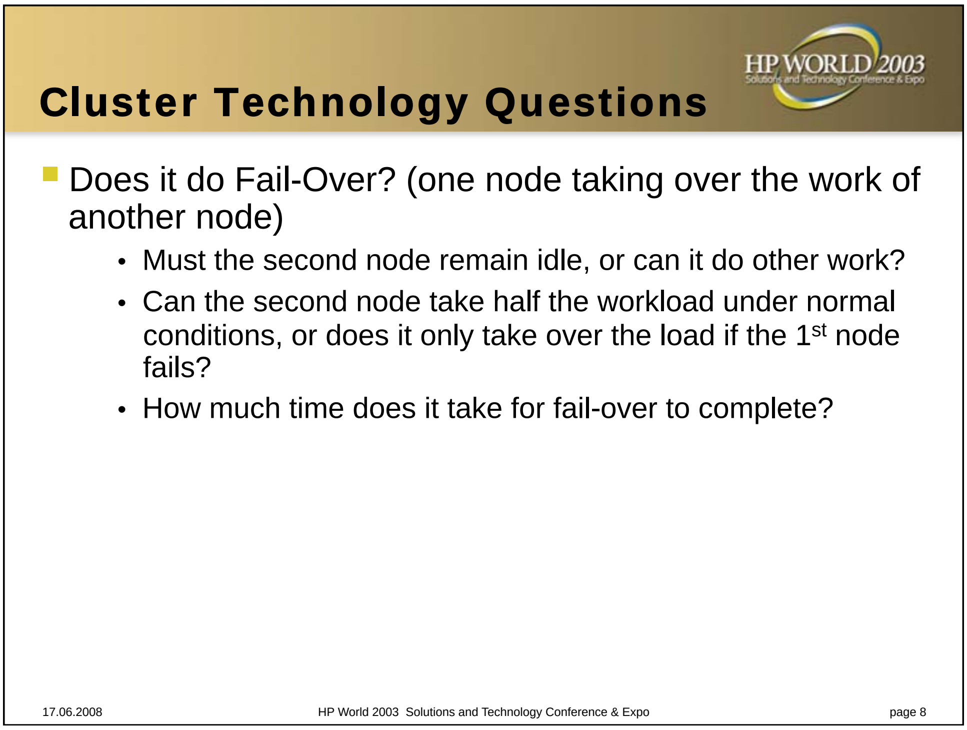
Cluster Technology Questions

- By asking appropriate questions, one can determine what level of sophistication or maturity a cluster solution has, by identifying which of various basic cluster technology building blocks are included, such as:
  - Load-balancing
  - Fail-over
  - Shared disk access
  - Quorum scheme
  - Cluster Lock Manager
  - Cluster File System
  - etc.



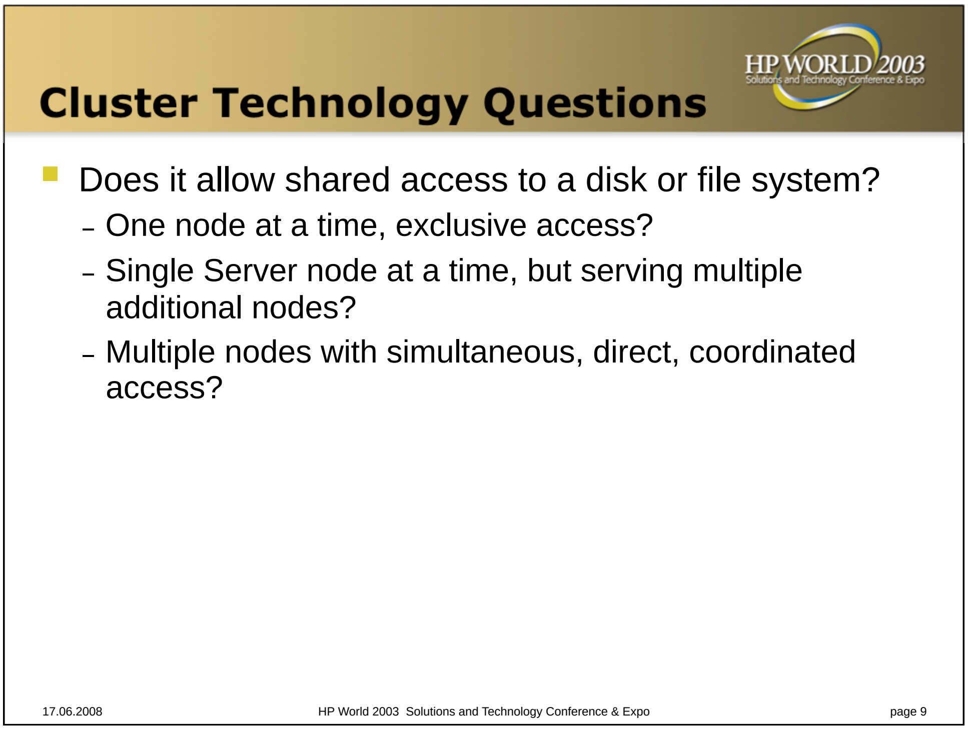
Cluster Technology Questions

- Can multiple nodes be given pieces of a sub-dividable problem?
  - High-performance technical computing problems
  - Data partitioning
- Can workload be distributed across multiple systems which perform identical functions?
  - e.g. Web server farm



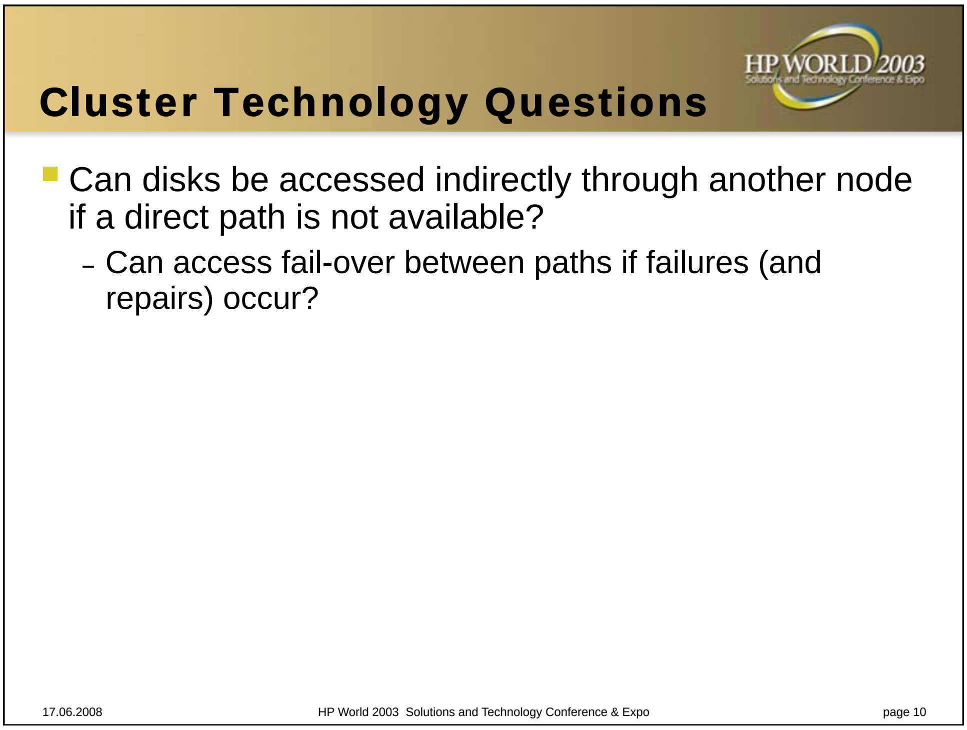
Cluster Technology Questions

- Does it do Fail-Over? (one node taking over the work of another node)
  - Must the second node remain idle, or can it do other work?
  - Can the second node take half the workload under normal conditions, or does it only take over the load if the 1st node fails?
  - How much time does it take for fail-over to complete?



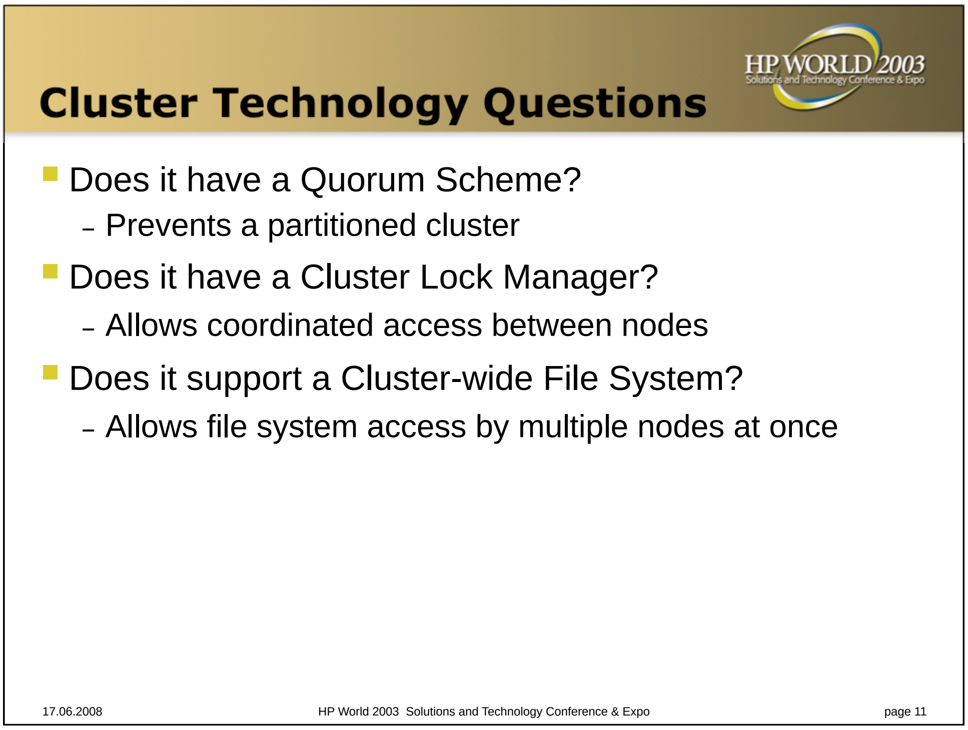
Cluster Technology Questions

- Does it allow shared access to a disk or file system?
  - One node at a time, exclusive access?
  - Single Server node at a time, but serving multiple additional nodes?
  - Multiple nodes with simultaneous, direct, coordinated access?



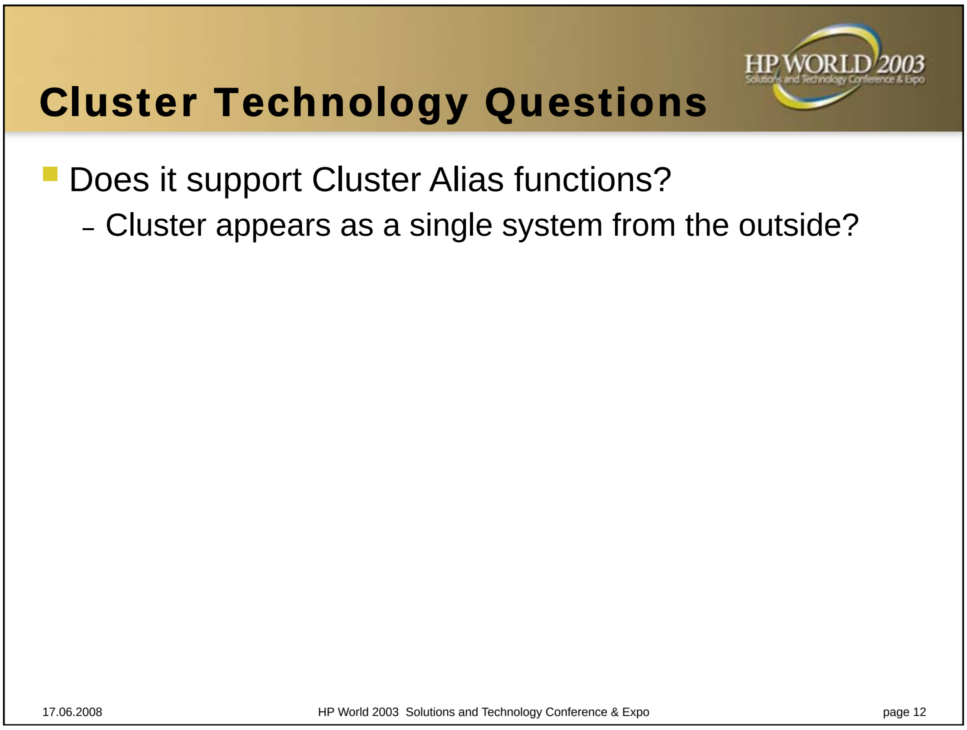
Cluster Technology Questions

- Can disks be accessed indirectly through another node if a direct path is not available?
  - Can access fail-over between paths if failures (and repairs) occur?



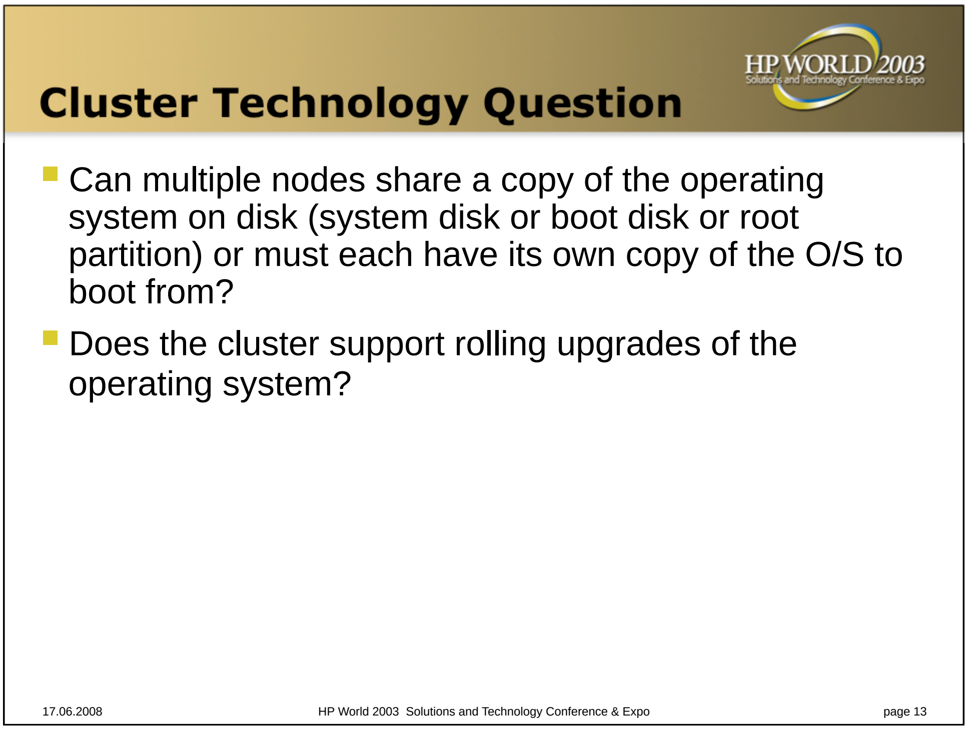
Cluster Technology Questions

- Does it have a Quorum Scheme?
  - Prevents a partitioned cluster
- Does it have a Cluster Lock Manager?
  - Allows coordinated access between nodes
- Does it support a Cluster-wide File System?
  - Allows file system access by multiple nodes at once



Cluster Technology Questions

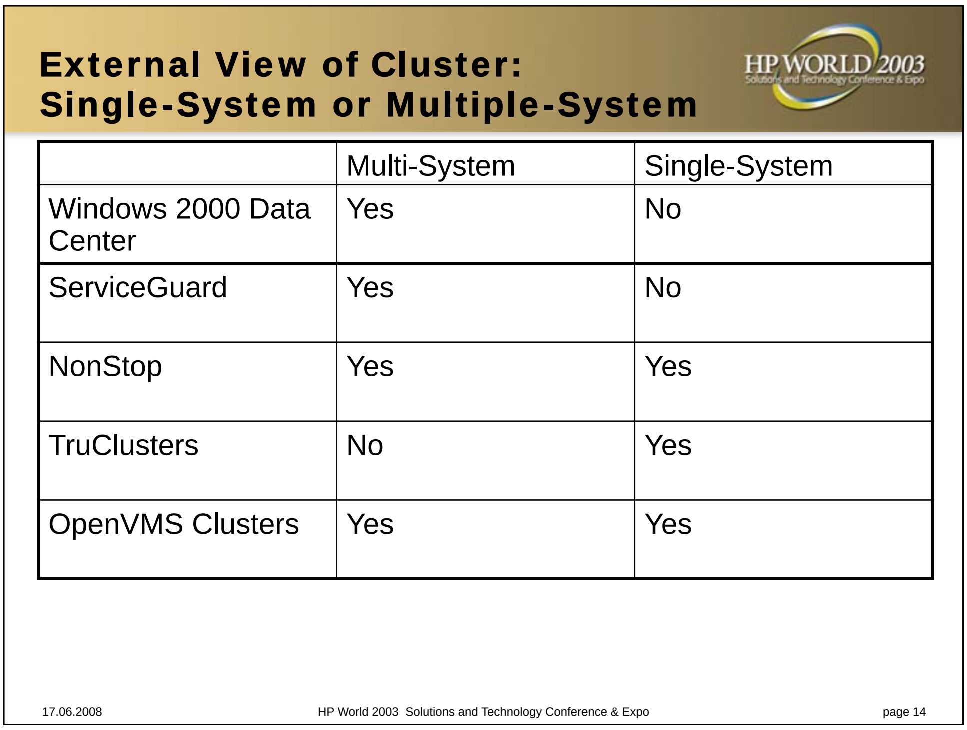
- Does it support Cluster Alias functions?
  - Cluster appears as a single system from the outside?



Cluster Technology Question

- Can multiple nodes share a copy of the operating system on disk (system disk or boot disk or root partition) or must each have its own copy of the O/S to boot from?
- Does the cluster support rolling upgrades of the operating system?

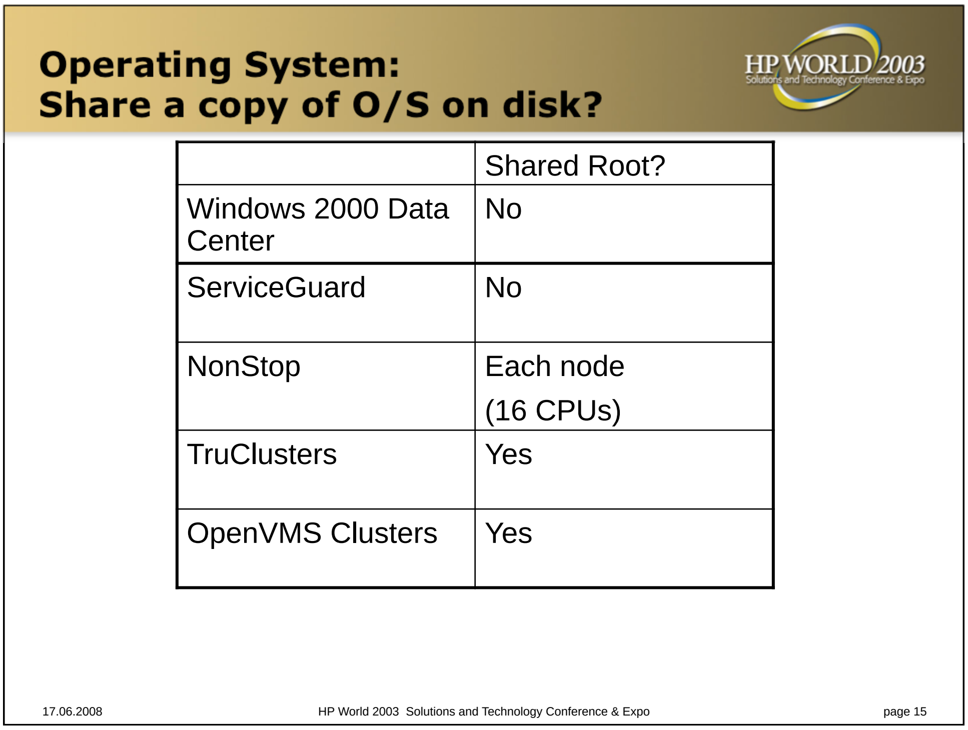
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 13



External View of Cluster: Single-System or Multiple-System

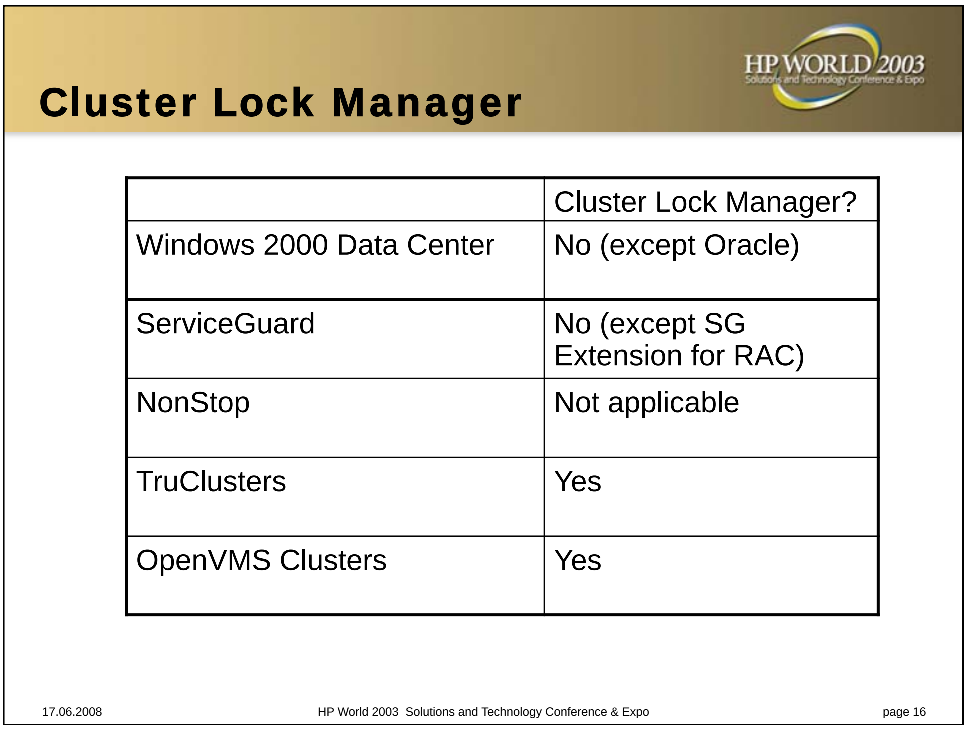
	Multi-System	Single-System
Windows 2000 Data Center	Yes	No
ServiceGuard	Yes	No
NonStop	Yes	Yes
TruClusters	No	Yes
OpenVMS Clusters	Yes	Yes

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 14



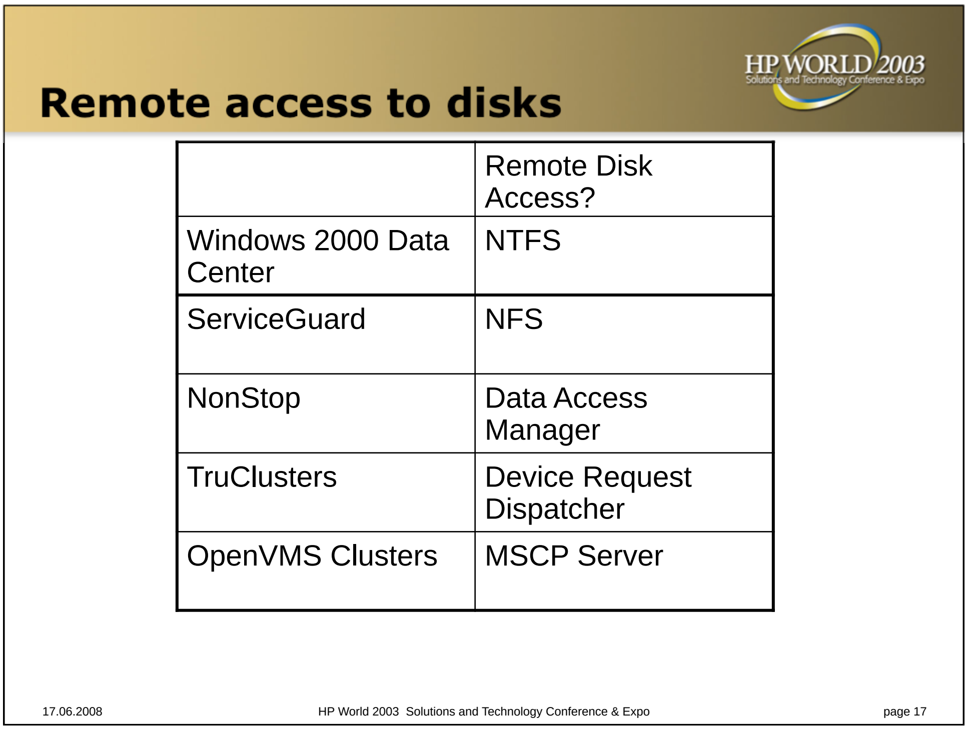
Operating System: Share a copy of O/S on disk?

	Shared Root?
Windows 2000 Data Center	No
ServiceGuard	No
NonStop	Each node (16 CPUs)
TruClusters	Yes
OpenVMS Clusters	Yes



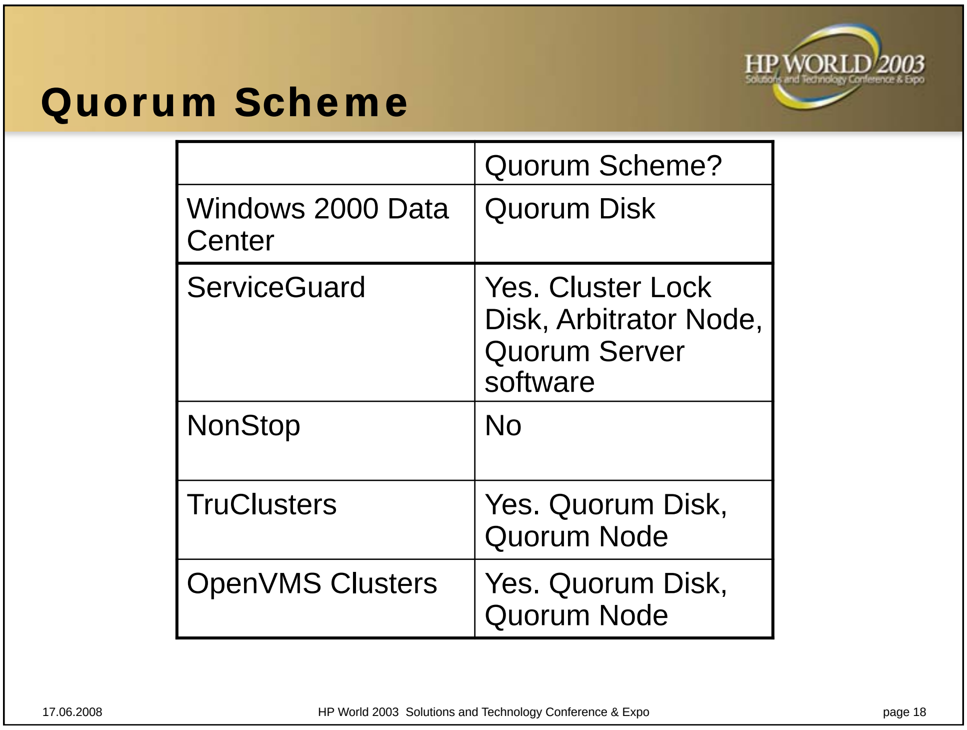
Cluster Lock Manager

	Cluster Lock Manager?
Windows 2000 Data Center	No (except Oracle)
ServiceGuard	No (except SG Extension for RAC)
NonStop	Not applicable
TruClusters	Yes
OpenVMS Clusters	Yes



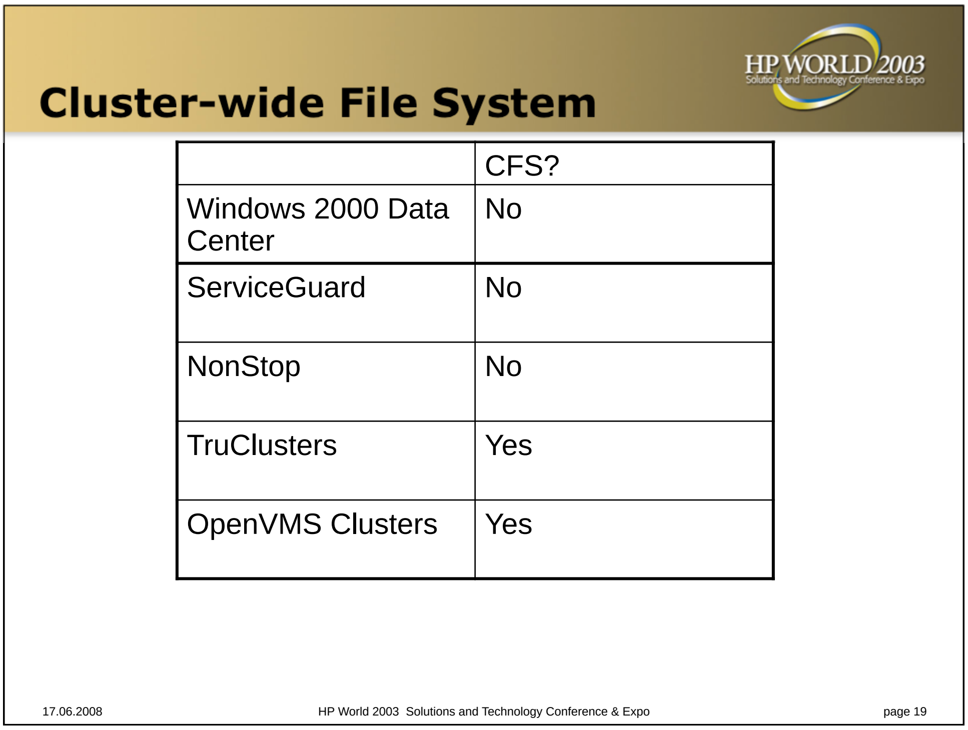
Remote access to disks

	Remote Disk Access?
Windows 2000 Data Center	NTFS
ServiceGuard	NFS
NonStop	Data Access Manager
TruClusters	Device Request Dispatcher
OpenVMS Clusters	MSCP Server



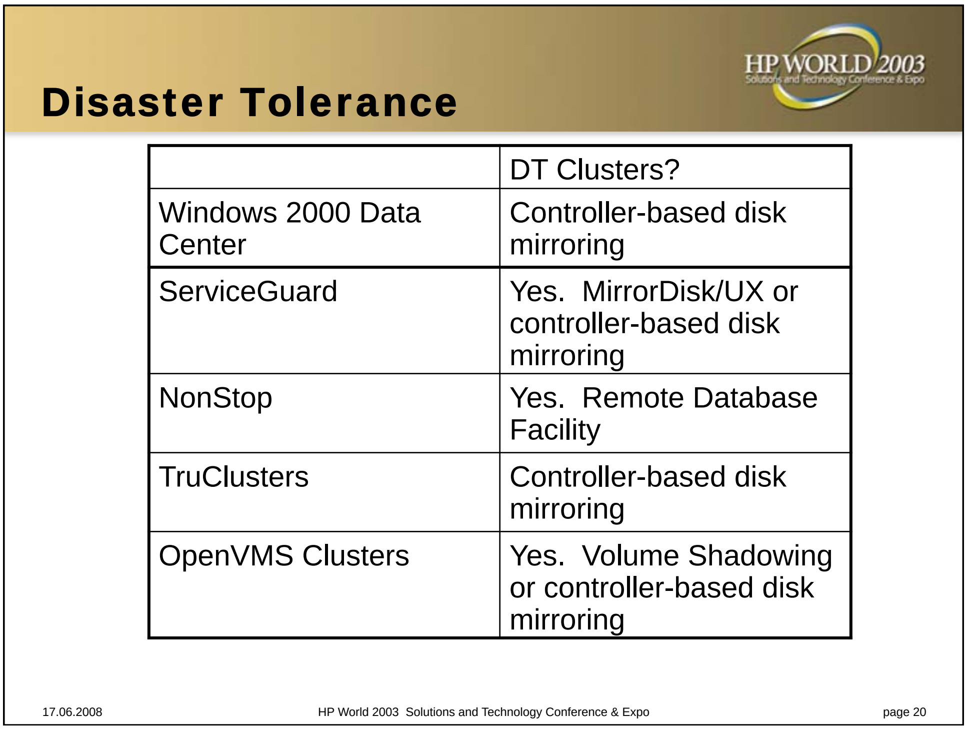
Quorum Scheme

	Quorum Scheme?
Windows 2000 Data Center	Quorum Disk
ServiceGuard	Yes. Cluster Lock Disk, Arbitrator Node, Quorum Server software
NonStop	No
TruClusters	Yes. Quorum Disk, Quorum Node
OpenVMS Clusters	Yes. Quorum Disk, Quorum Node



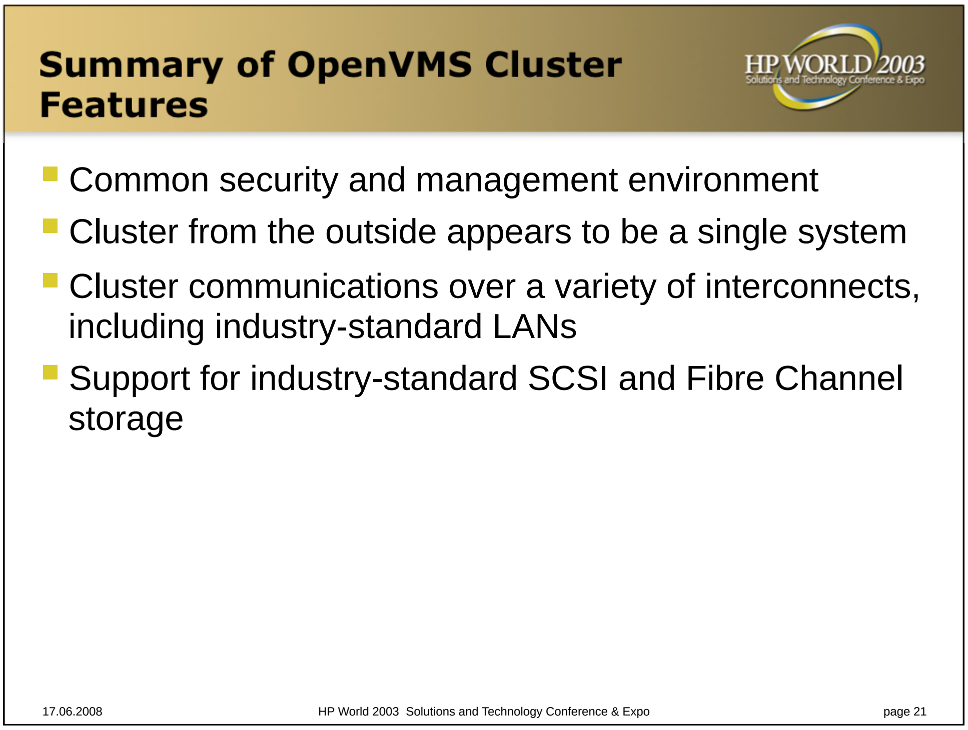
Cluster-wide File System

	CFS?
Windows 2000 Data Center	No
ServiceGuard	No
NonStop	No
TruClusters	Yes
OpenVMS Clusters	Yes



Disaster Tolerance

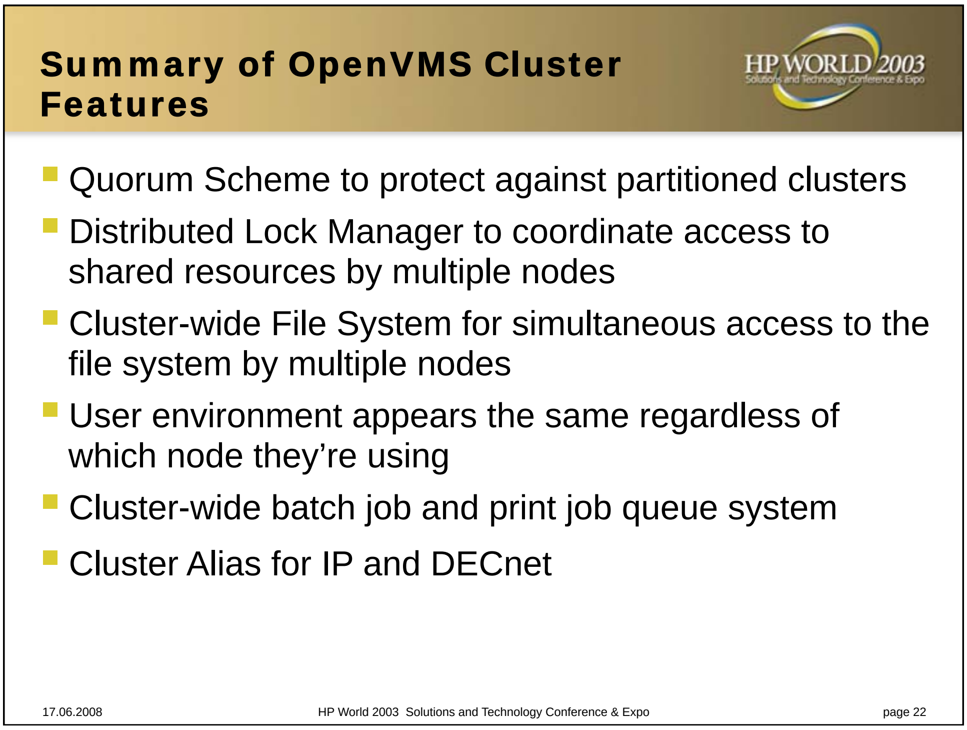
	DT Clusters?
Windows 2000 Data Center	Controller-based disk mirroring
ServiceGuard	Yes. MirrorDisk/UX or controller-based disk mirroring
NonStop	Yes. Remote Database Facility
TruClusters	Controller-based disk mirroring
OpenVMS Clusters	Yes. Volume Shadowing or controller-based disk mirroring



Summary of OpenVMS Cluster Features

HP WORLD 2003
Solutions and Technology Conference & Expo

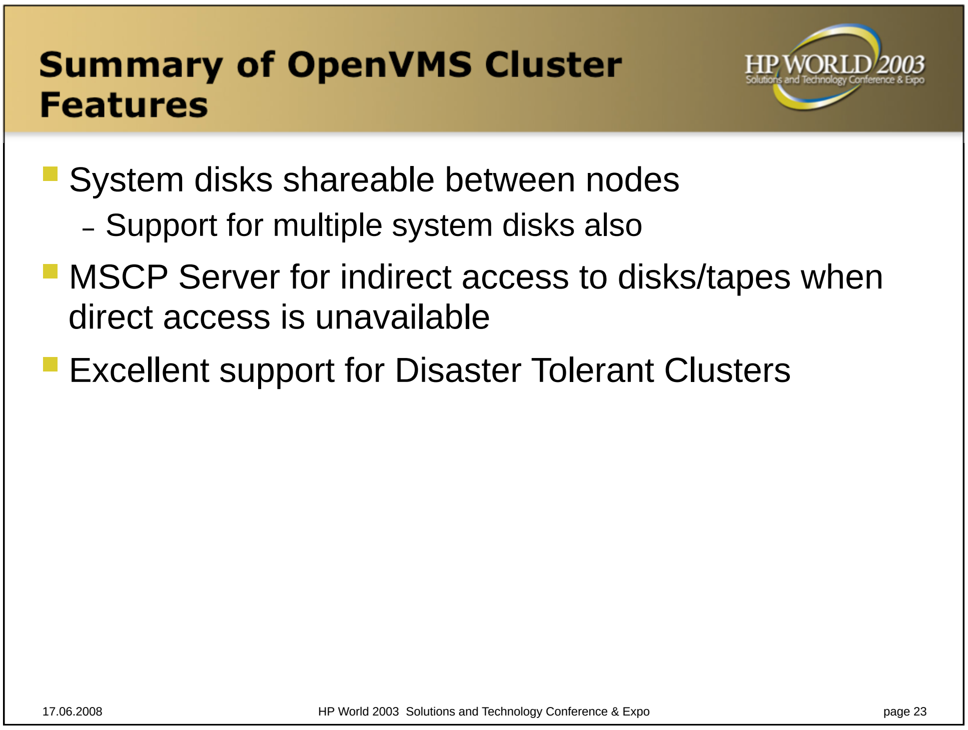
- Common security and management environment
- Cluster from the outside appears to be a single system
- Cluster communications over a variety of interconnects, including industry-standard LANs
- Support for industry-standard SCSI and Fibre Channel storage



Summary of OpenVMS Cluster Features

HP WORLD 2003
Solutions and Technology Conference & Expo

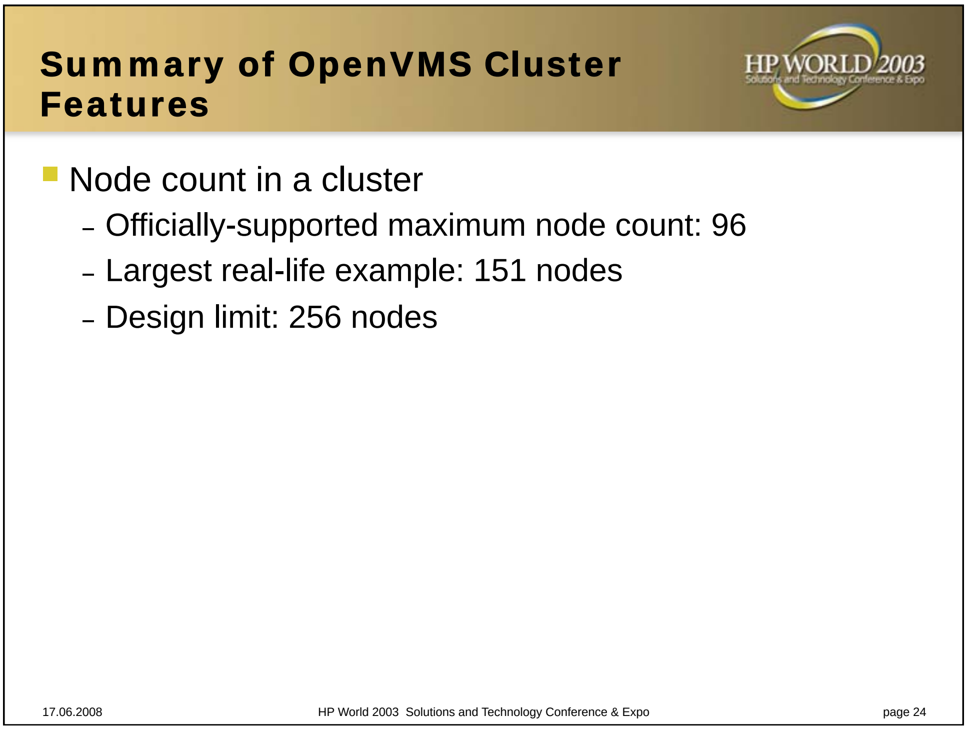
- Quorum Scheme to protect against partitioned clusters
- Distributed Lock Manager to coordinate access to shared resources by multiple nodes
- Cluster-wide File System for simultaneous access to the file system by multiple nodes
- User environment appears the same regardless of which node they’re using
- Cluster-wide batch job and print job queue system
- Cluster Alias for IP and DECnet



Summary of OpenVMS Cluster Features

HP WORLD 2003 Solutions and Technology Conference & Expo

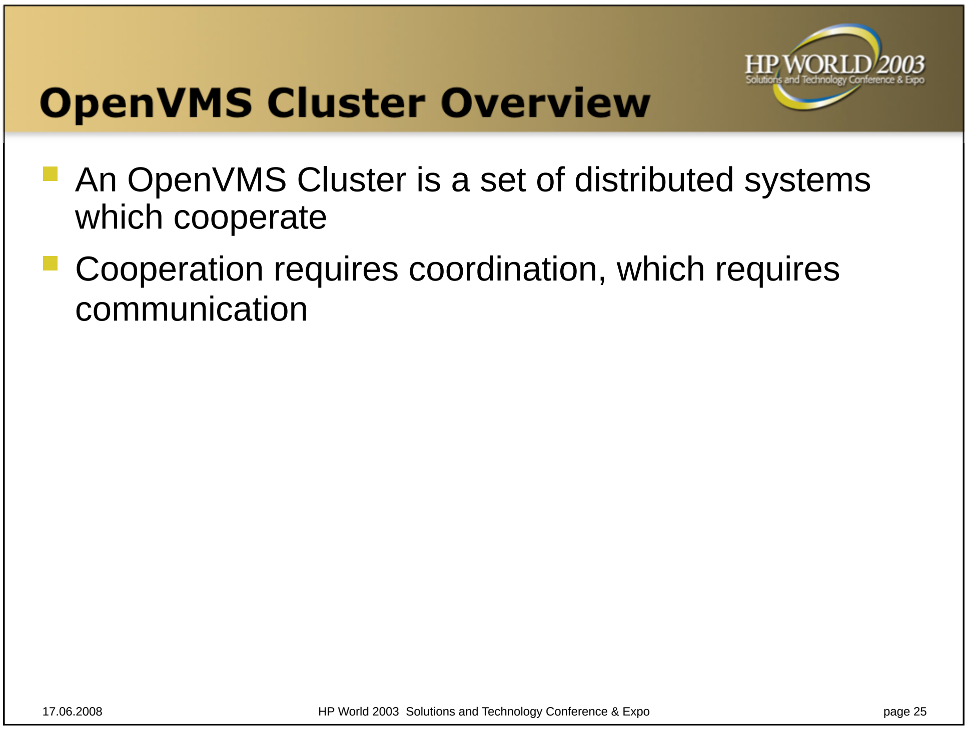
- System disks shareable between nodes
  - Support for multiple system disks also
- MSCP Server for indirect access to disks/tapes when direct access is unavailable
- Excellent support for Disaster Tolerant Clusters



Summary of OpenVMS Cluster Features

HP WORLD 2003 Solutions and Technology Conference & Expo

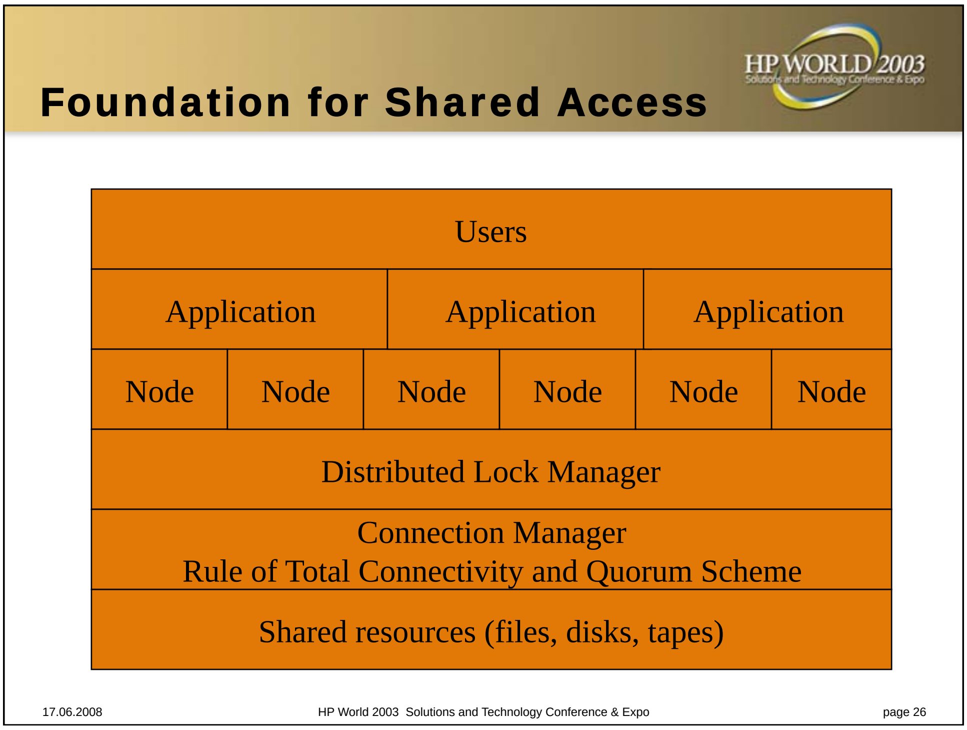
- Node count in a cluster
  - Officially-supported maximum node count: 96
  - Largest real-life example: 151 nodes
  - Design limit: 256 nodes



OpenVMS Cluster Overview

- An OpenVMS Cluster is a set of distributed systems which cooperate
- Cooperation requires coordination, which requires communication

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 25



Foundation for Shared Access

Users

Application | Application | Application

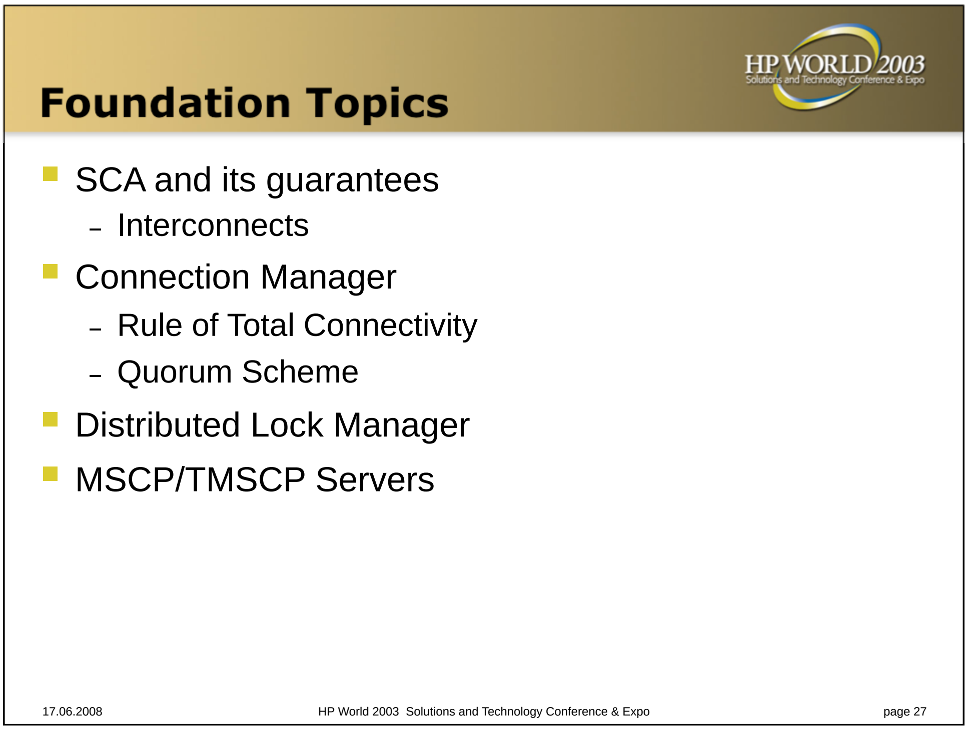
Node | Node | Node | Node | Node | Node

Distributed Lock Manager

Connection Manager
Rule of Total Connectivity and Quorum Scheme

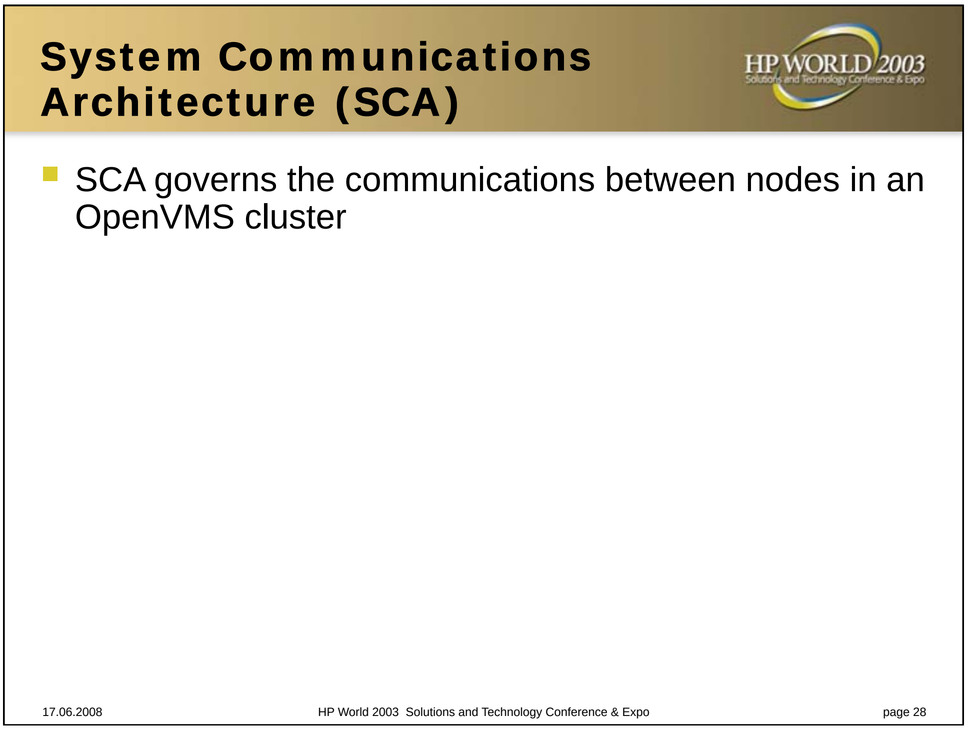
Shared resources (files, disks, tapes)

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 26



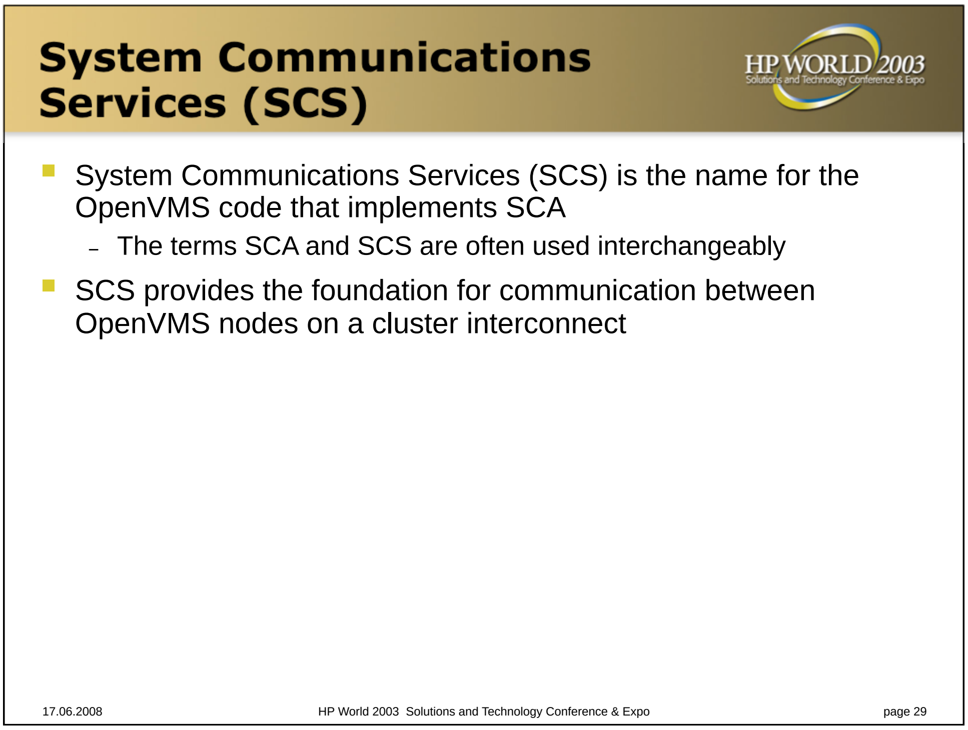
Foundation Topics

- SCA and its guarantees
  - Interconnects
- Connection Manager
  - Rule of Total Connectivity
  - Quorum Scheme
- Distributed Lock Manager
- MSCP/TMSCP Servers



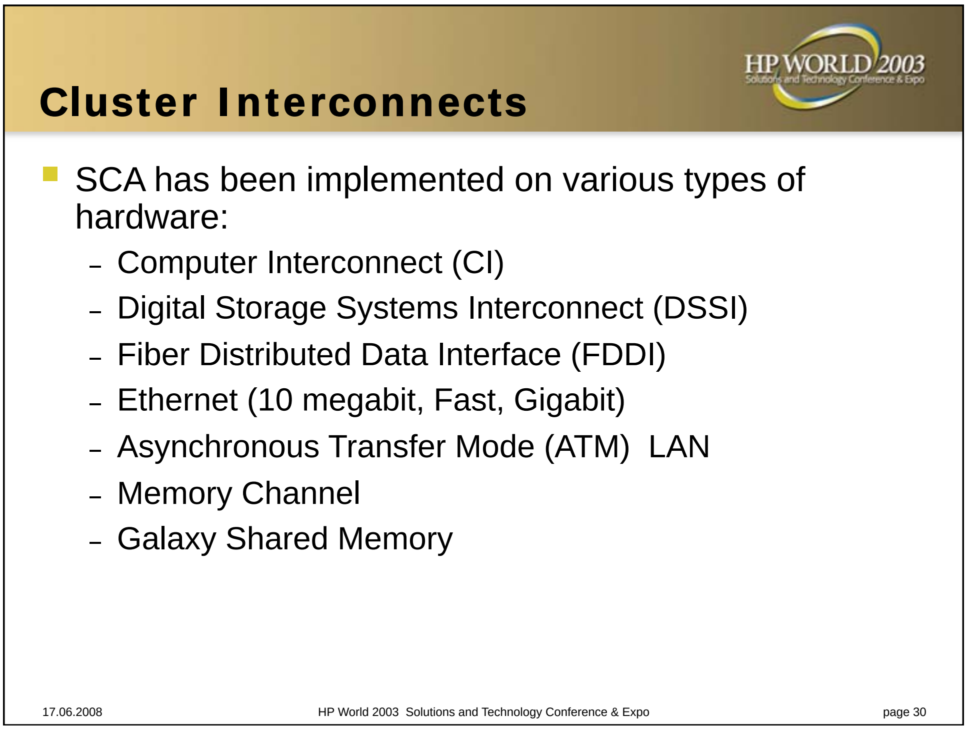
System Communications Architecture (SCA)

- SCA governs the communications between nodes in an OpenVMS cluster



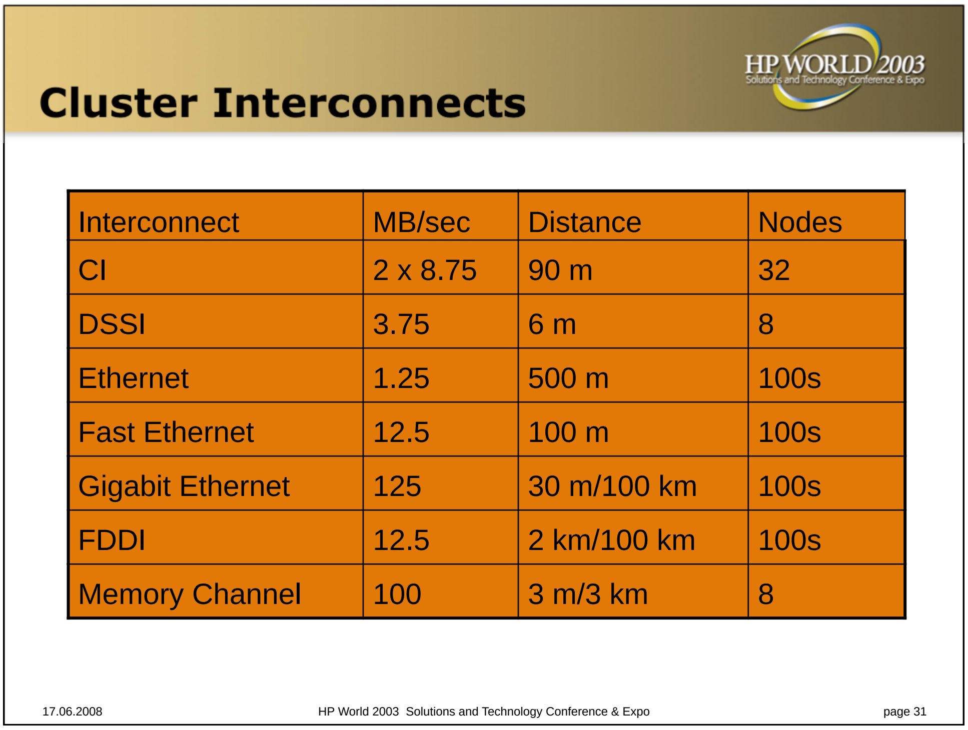
System Communications Services (SCS)

- System Communications Services (SCS) is the name for the OpenVMS code that implements SCA
  - The terms SCA and SCS are often used interchangeably
- SCS provides the foundation for communication between OpenVMS nodes on a cluster interconnect



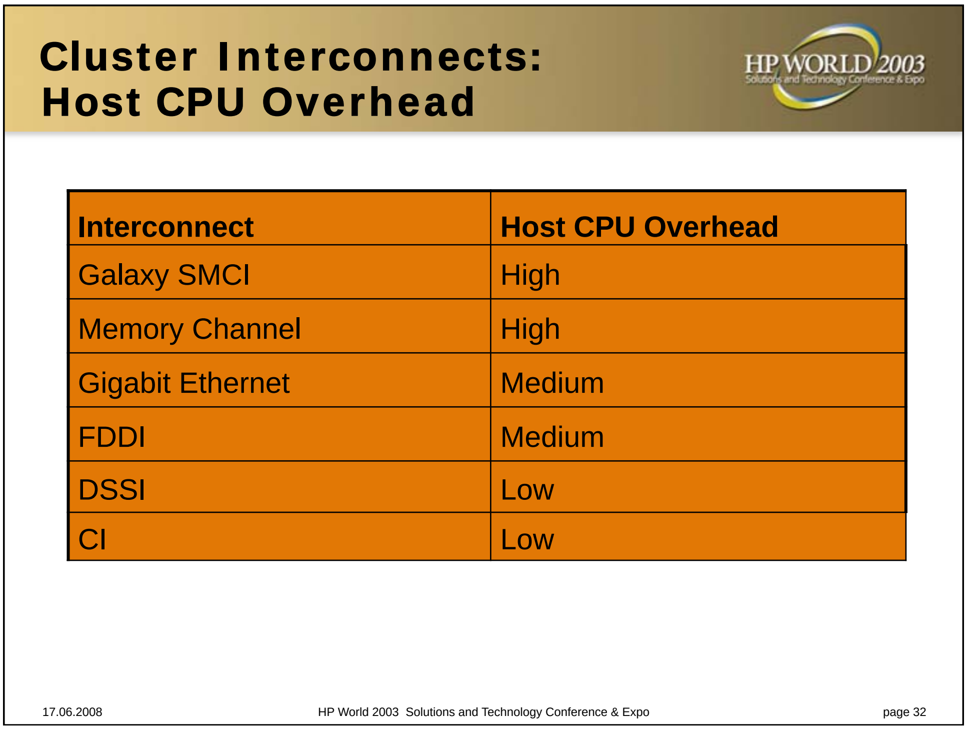
Cluster Interconnects

- SCA has been implemented on various types of hardware:
  - Computer Interconnect (CI)
  - Digital Storage Systems Interconnect (DSSI)
  - Fiber Distributed Data Interface (FDDI)
  - Ethernet (10 megabit, Fast, Gigabit)
  - Asynchronous Transfer Mode (ATM) LAN
  - Memory Channel
  - Galaxy Shared Memory



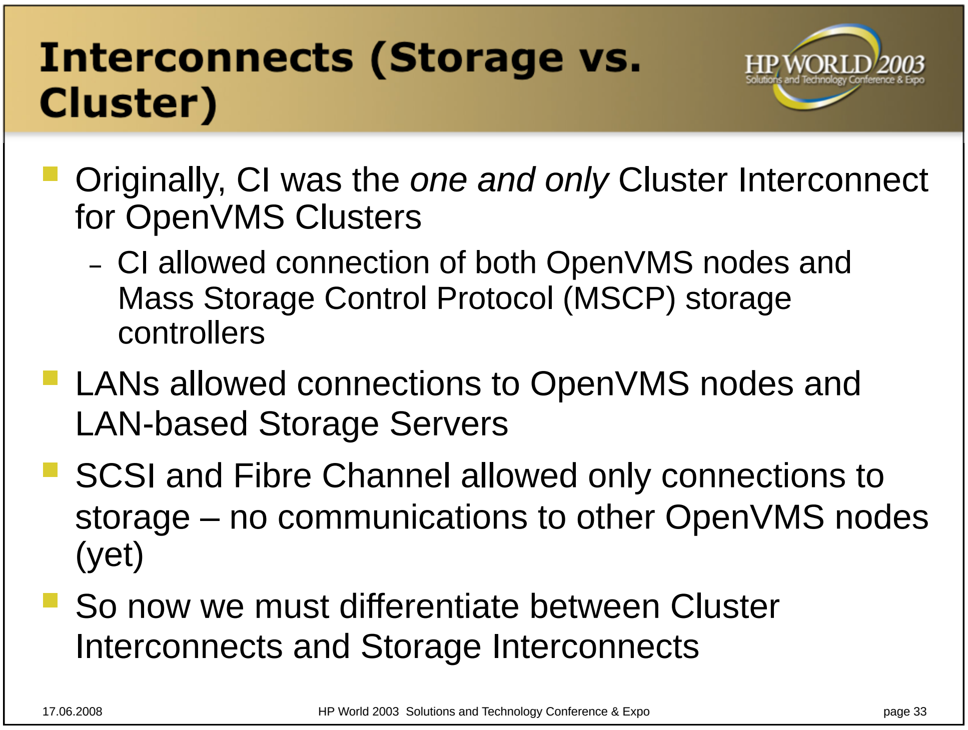
Cluster Interconnects

Interconnect	MB/sec	Distance	Nodes
CI	2 x 8.75	90 m	32
DSSI	3.75	6 m	8
Ethernet	1.25	500 m	100s
Fast Ethernet	12.5	100 m	100s
Gigabit Ethernet	125	30 m/100 km	100s
FDDI	12.5	2 km/100 km	100s
Memory Channel	100	3 m/3 km	8



Cluster Interconnects: Host CPU Overhead

Interconnect	**Host CPU Overhead**
Galaxy SMCI	High
Memory Channel	High
Gigabit Ethernet	Medium
FDDI	Medium
DSSI	Low
CI	Low

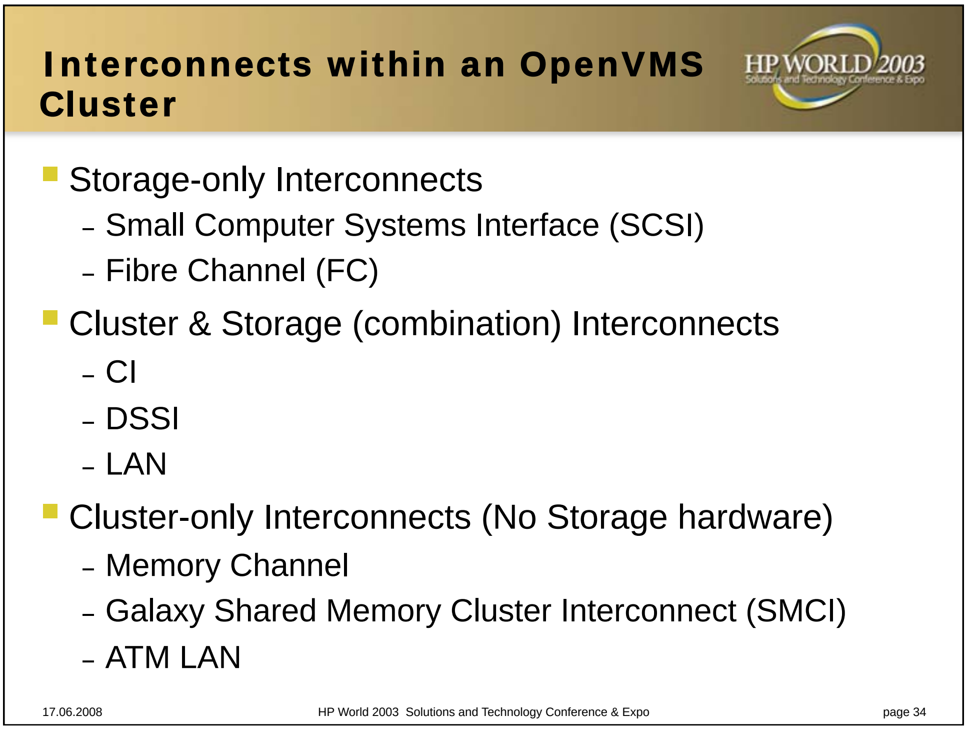


Interconnects (Storage vs. Cluster)

HP WORLD 2003 Solutions and Technology Conference & Expo

- Originally, CI was the *one and only* Cluster Interconnect for OpenVMS Clusters
  - CI allowed connection of both OpenVMS nodes and Mass Storage Control Protocol (MSCP) storage controllers
- LANs allowed connections to OpenVMS nodes and LAN-based Storage Servers
- SCSI and Fibre Channel allowed only connections to storage – no communications to other OpenVMS nodes (yet)
- So now we must differentiate between Cluster Interconnects and Storage Interconnects

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 33

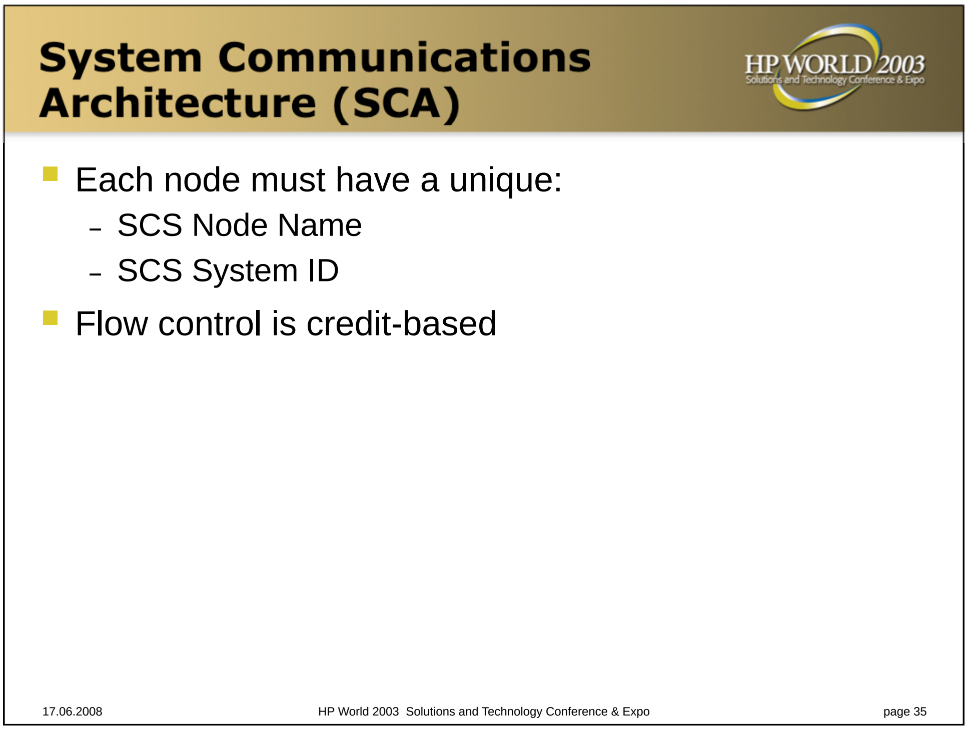


Interconnects within an OpenVMS Cluster

HP WORLD 2003 Solutions and Technology Conference & Expo

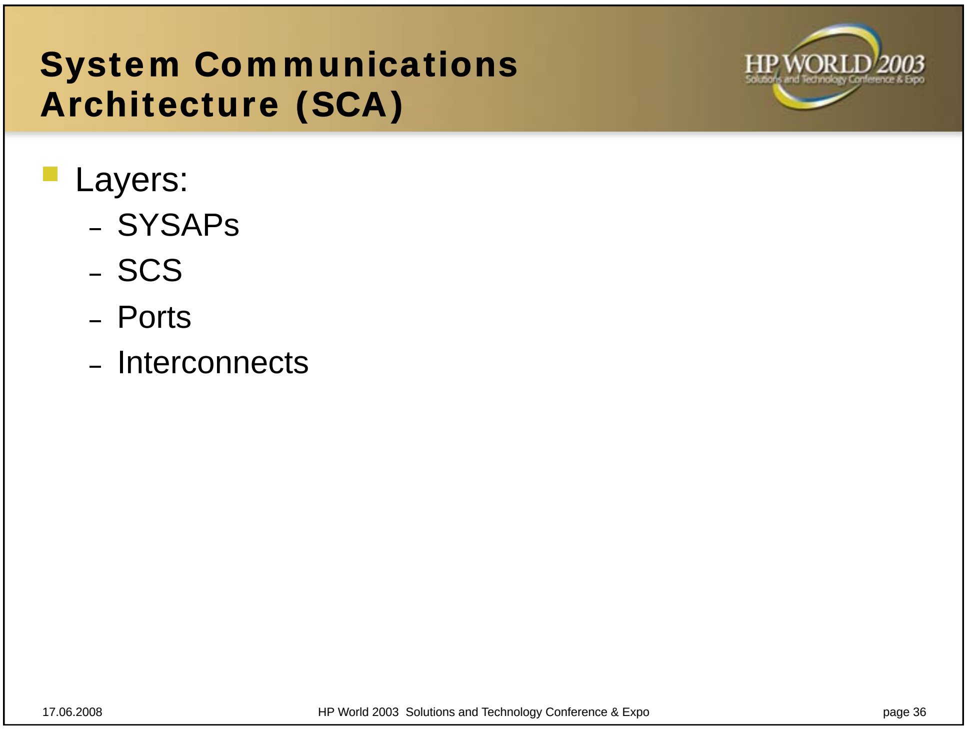
- Storage-only Interconnects
  - Small Computer Systems Interface (SCSI)
  - Fibre Channel (FC)
- Cluster & Storage (combination) Interconnects
  - CI
  - DSSI
  - LAN
- Cluster-only Interconnects (No Storage hardware)
  - Memory Channel
  - Galaxy Shared Memory Cluster Interconnect (SMCI)
  - ATM LAN

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 34



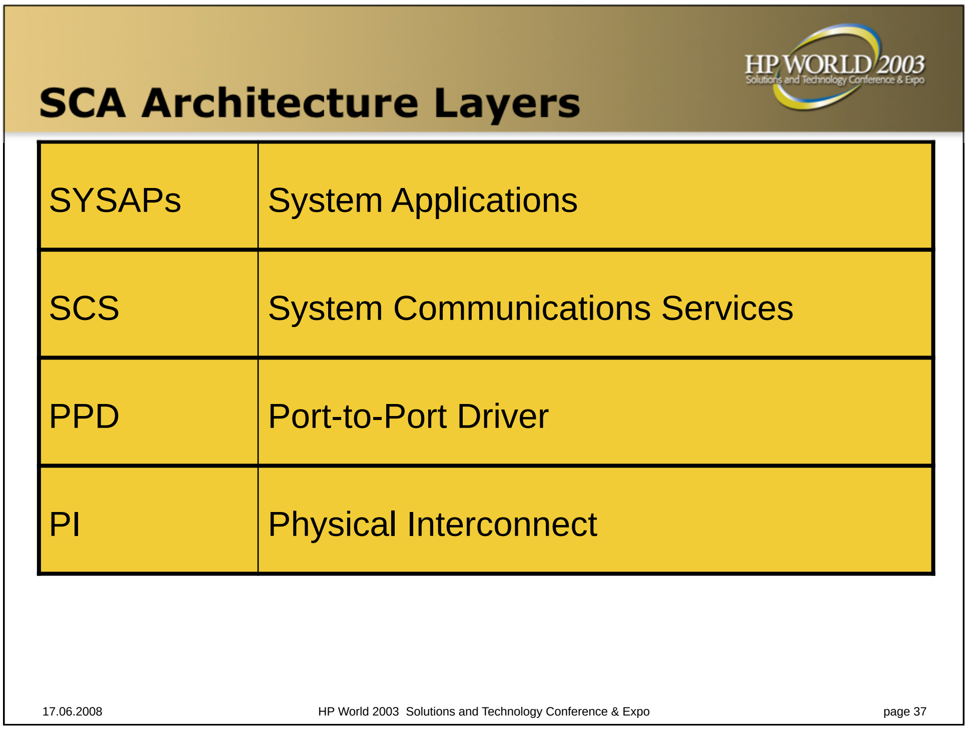
System Communications Architecture (SCA)

- Each node must have a unique:
  - SCS Node Name
  - SCS System ID
- Flow control is credit-based



System Communications Architecture (SCA)

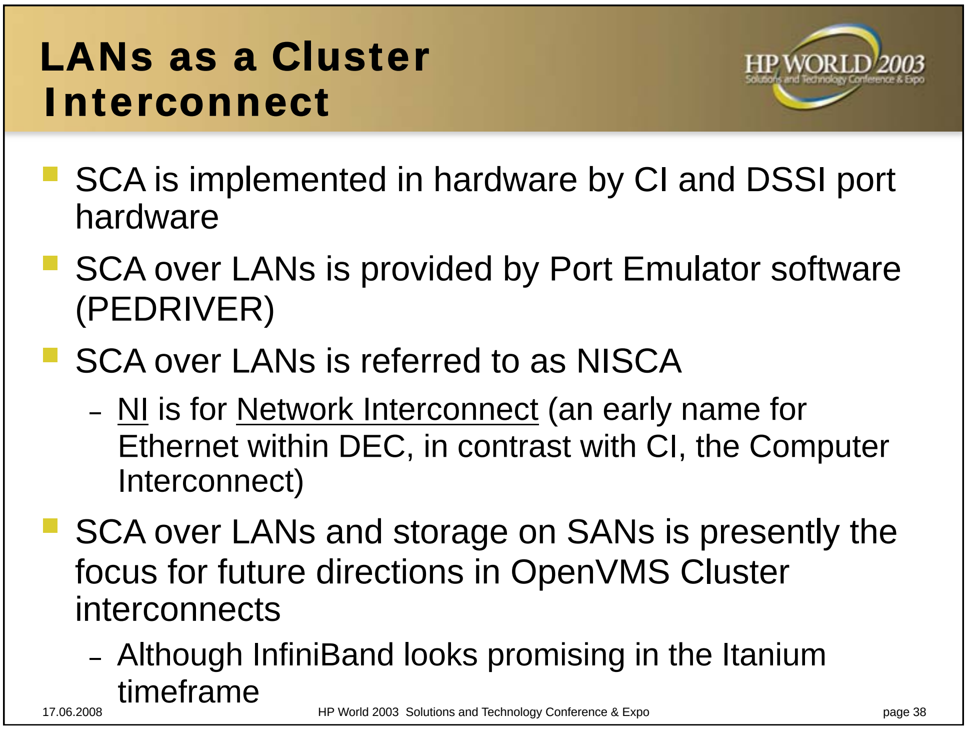
- Layers:
  - SYSAPs
  - SCS
  - Ports
  - Interconnects



SCA Architecture Layers

SYSAPs	System Applications
SCS	System Communications Services
PPD	Port-to-Port Driver
PI	Physical Interconnect

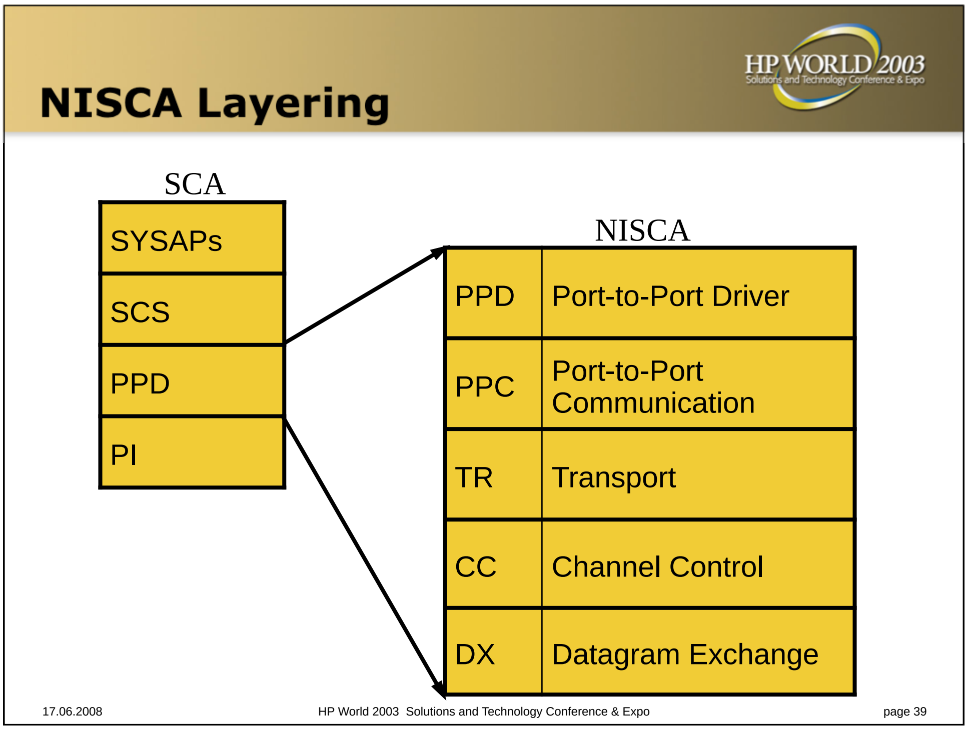
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 37



LANs as a Cluster Interconnect

- SCA is implemented in hardware by CI and DSSI port hardware
- SCA over LANs is provided by Port Emulator software (PEDRIVER)
- SCA over LANs is referred to as NISCA
  - NI is for Network Interconnect (an early name for Ethernet within DEC, in contrast with CI, the Computer Interconnect)
- SCA over LANs and storage on SANs is presently the focus for future directions in OpenVMS Cluster interconnects
  - Although InfiniBand looks promising in the Itanium timeframe

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 38



NISCA Layering

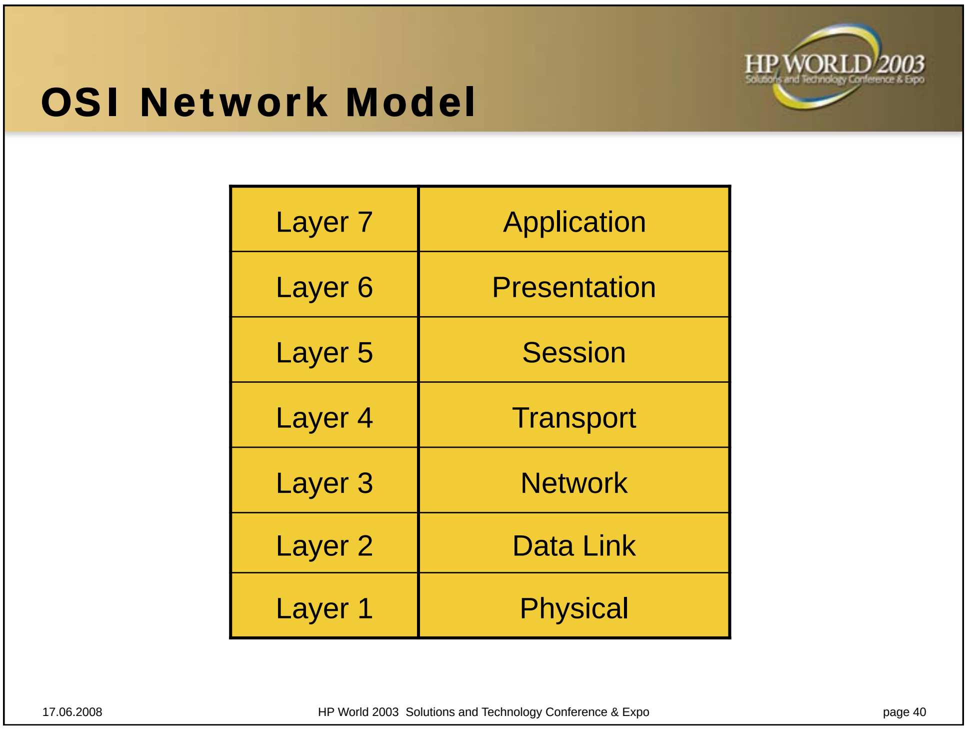
SCA

- SYSAPs
- SCS
- PPD
- PI

NISCA

PPD	Port-to-Port Driver
PPC	Port-to-Port Communication
TR	Transport
CC	Channel Control
DX	Datagram Exchange

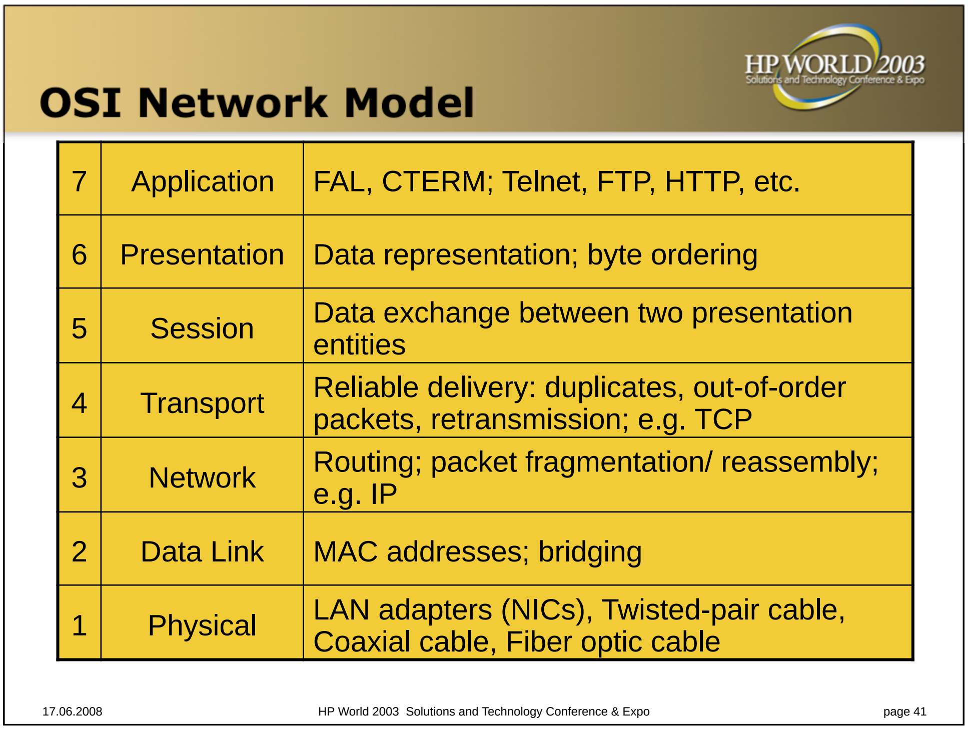
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 39



OSI Network Model

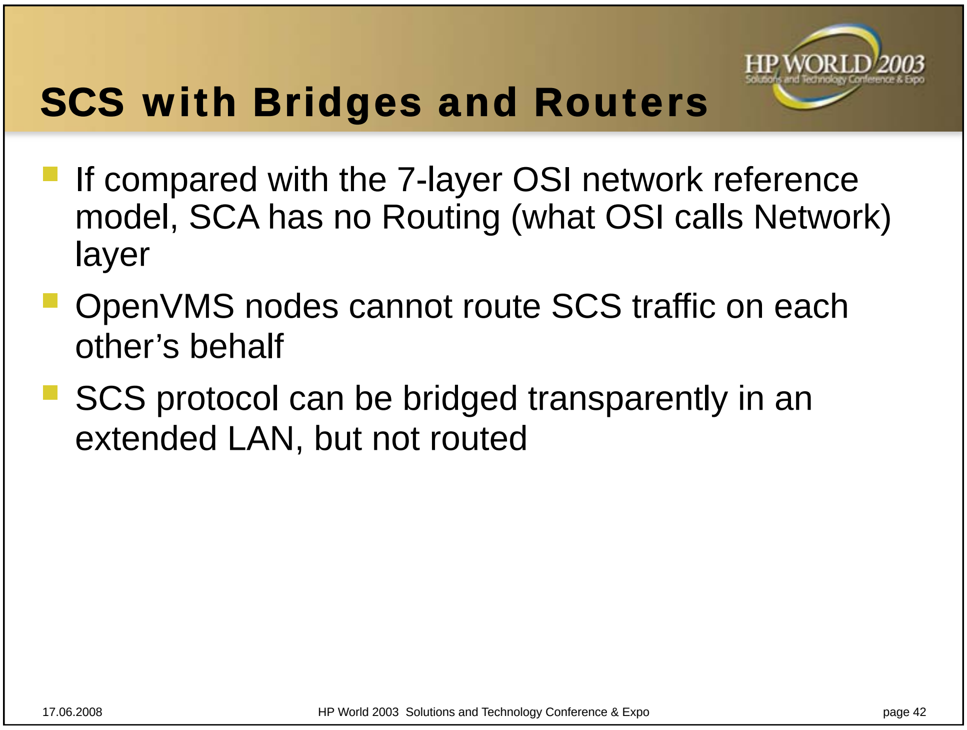
Layer 7	Application
Layer 6	Presentation
Layer 5	Session
Layer 4	Transport
Layer 3	Network
Layer 2	Data Link
Layer 1	Physical

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 40



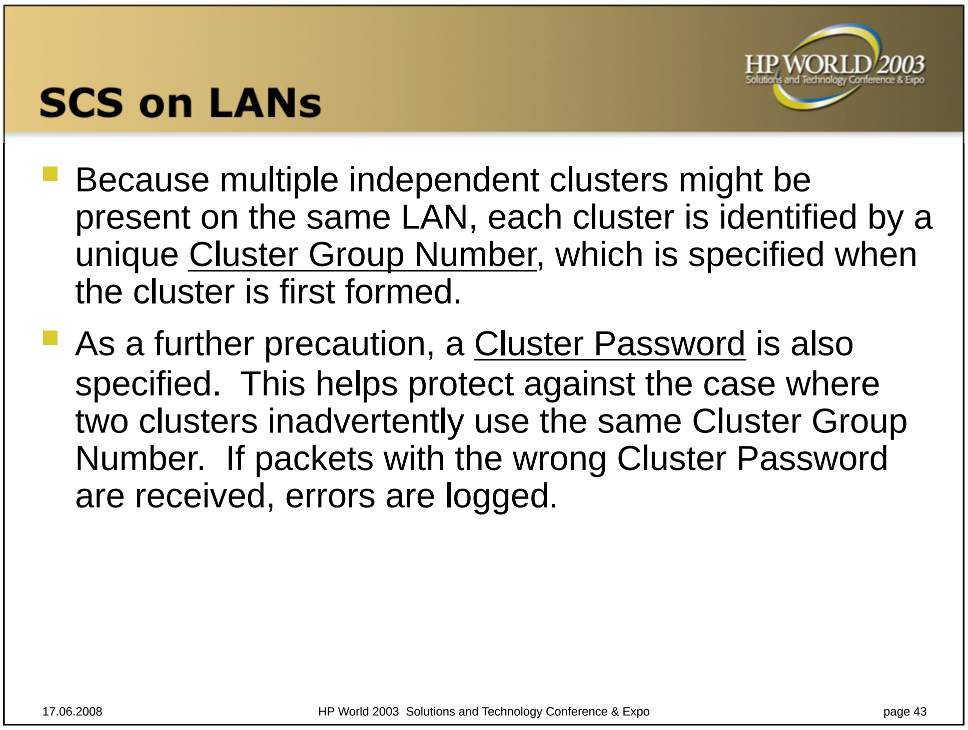
OSI Network Model

7	Application	FAL, CTERM; Telnet, FTP, HTTP, etc.
6	Presentation	Data representation; byte ordering
5	Session	Data exchange between two presentation entities
4	Transport	Reliable delivery: duplicates, out-of-order packets, retransmission; e.g. TCP
3	Network	Routing; packet fragmentation/ reassembly; e.g. IP
2	Data Link	MAC addresses; bridging
1	Physical	LAN adapters (NICs), Twisted-pair cable, Coaxial cable, Fiber optic cable



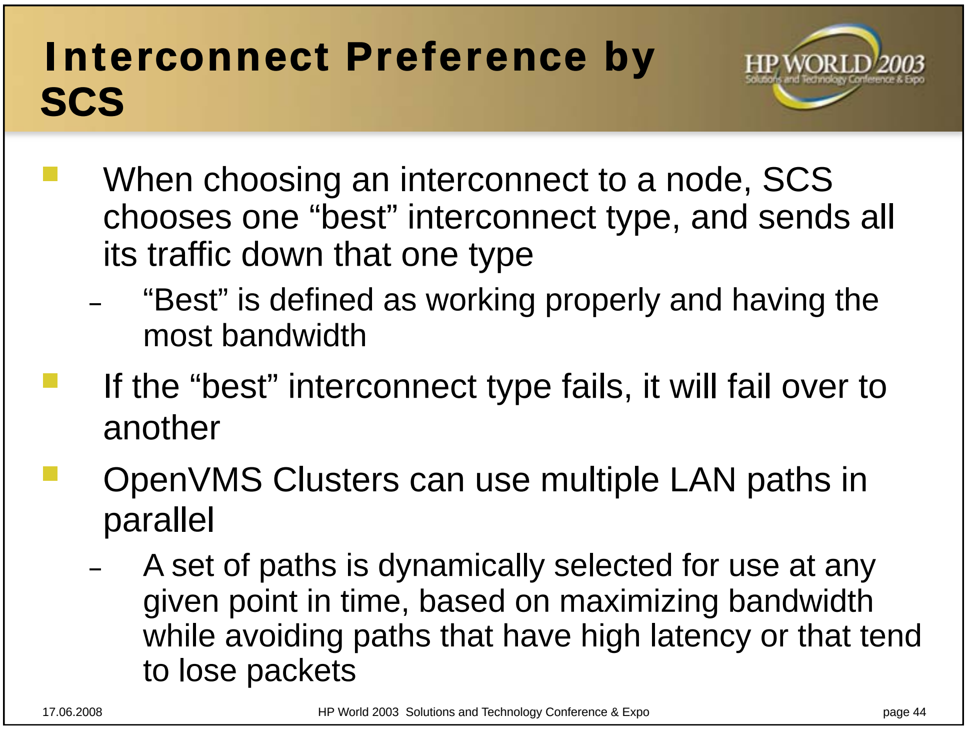
SCS with Bridges and Routers

- If compared with the 7-layer OSI network reference model, SCA has no Routing (what OSI calls Network) layer
- OpenVMS nodes cannot route SCS traffic on each other’s behalf
- SCS protocol can be bridged transparently in an extended LAN, but not routed



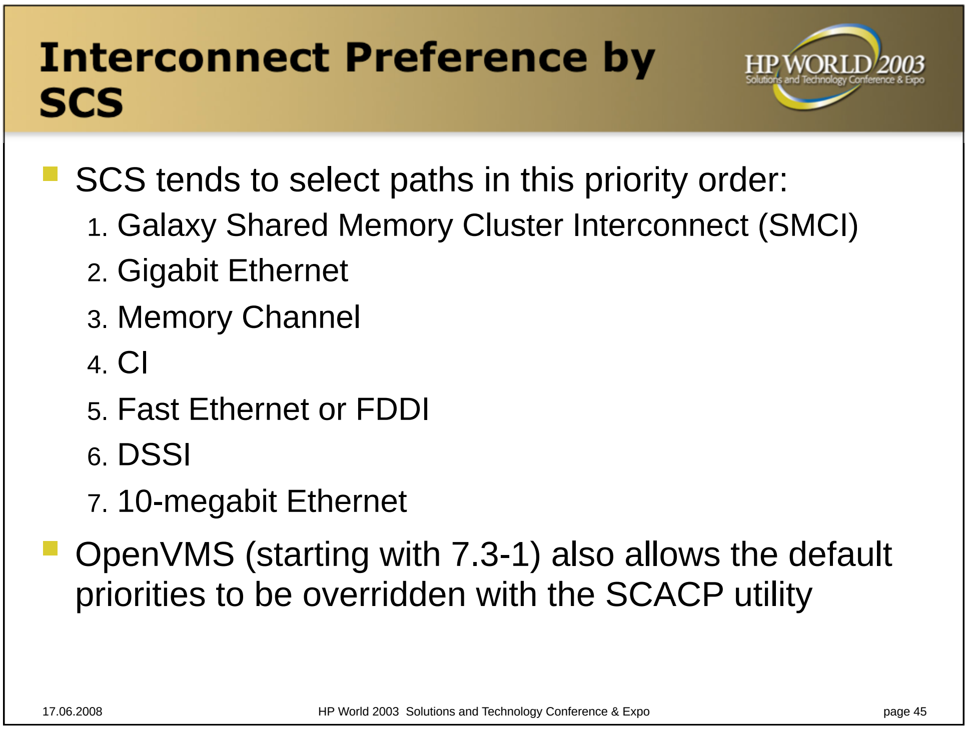
SCS on LANs

- Because multiple independent clusters might be present on the same LAN, each cluster is identified by a unique Cluster Group Number, which is specified when the cluster is first formed.
- As a further precaution, a Cluster Password is also specified. This helps protect against the case where two clusters inadvertently use the same Cluster Group Number. If packets with the wrong Cluster Password are received, errors are logged.



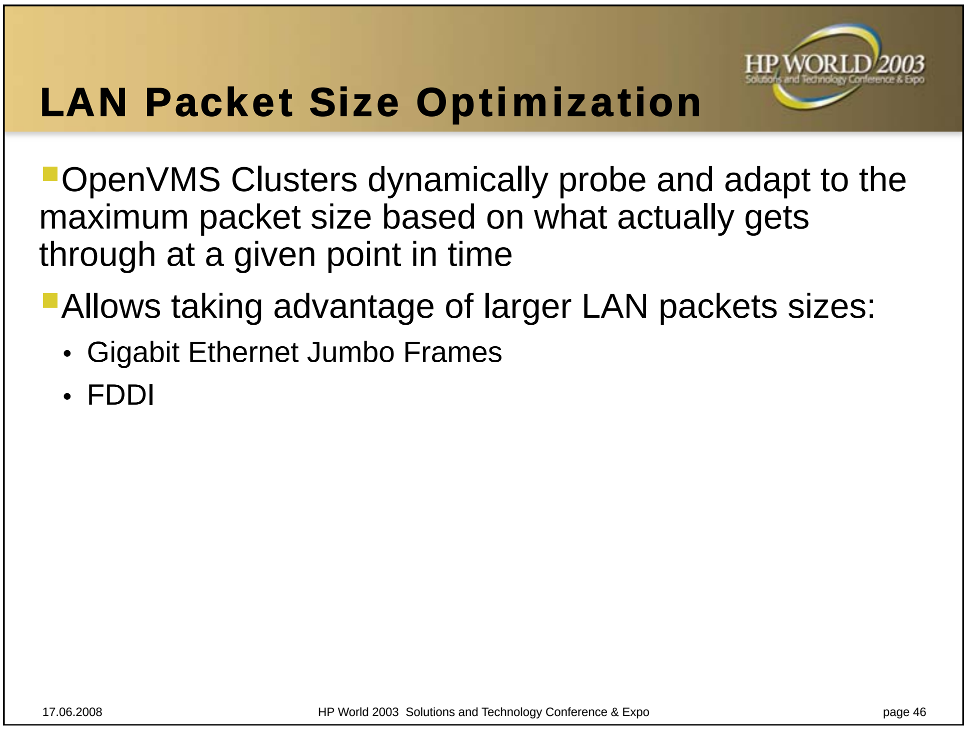
Interconnect Preference by SCS

- When choosing an interconnect to a node, SCS chooses one “best” interconnect type, and sends all its traffic down that one type
  - “Best” is defined as working properly and having the most bandwidth
- If the “best” interconnect type fails, it will fail over to another
- OpenVMS Clusters can use multiple LAN paths in parallel
  - A set of paths is dynamically selected for use at any given point in time, based on maximizing bandwidth while avoiding paths that have high latency or that tend to lose packets



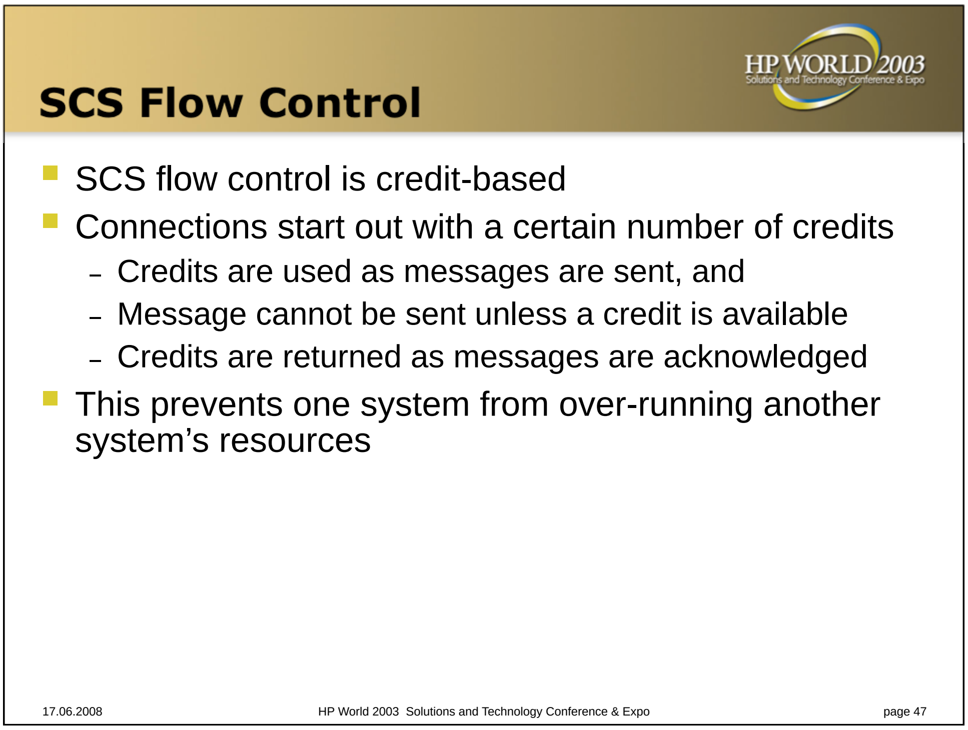
Interconnect Preference by SCS

- SCS tends to select paths in this priority order:
  1. Galaxy Shared Memory Cluster Interconnect (SMCI)
  2. Gigabit Ethernet
  3. Memory Channel
  4. CI
  5. Fast Ethernet or FDDI
  6. DSSI
  7. 10-megabit Ethernet
- OpenVMS (starting with 7.3-1) also allows the default priorities to be overridden with the SCACP utility



LAN Packet Size Optimization

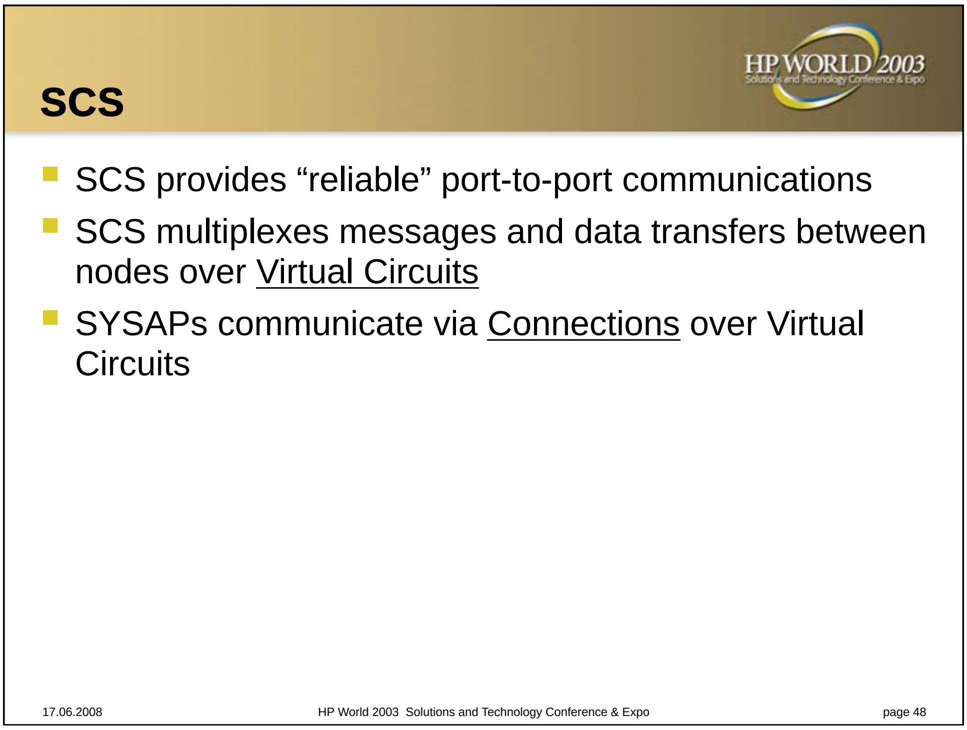
- OpenVMS Clusters dynamically probe and adapt to the maximum packet size based on what actually gets through at a given point in time
- Allows taking advantage of larger LAN packets sizes:
  - Gigabit Ethernet Jumbo Frames
  - FDDI



SCS Flow Control

- SCS flow control is credit-based
- Connections start out with a certain number of credits
  - Credits are used as messages are sent, and
  - Message cannot be sent unless a credit is available
  - Credits are returned as messages are acknowledged
- This prevents one system from over-running another system’s resources

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 47



SCS

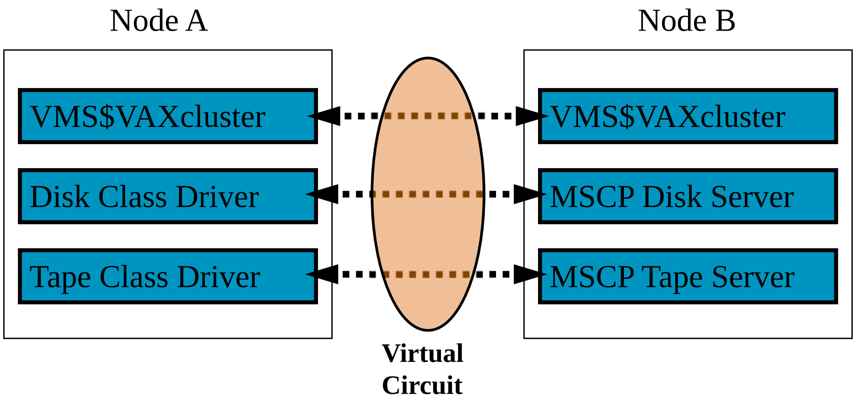
- SCS provides “reliable” port-to-port communications
- SCS multiplexes messages and data transfers between nodes over Virtual Circuits
- SYSAPs communicate via Connections over Virtual Circuits

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 48

Virtual Circuits

- Formed between ports on a Cluster Interconnect of some flavor
- Can pass data in 3 ways:
  - Datagrams
  - Sequenced Messages
  - Block Data Transfers

Connections over a Virtual Circuit



Node A

VMS$VAXcluster

Disk Class Driver

Tape Class Driver

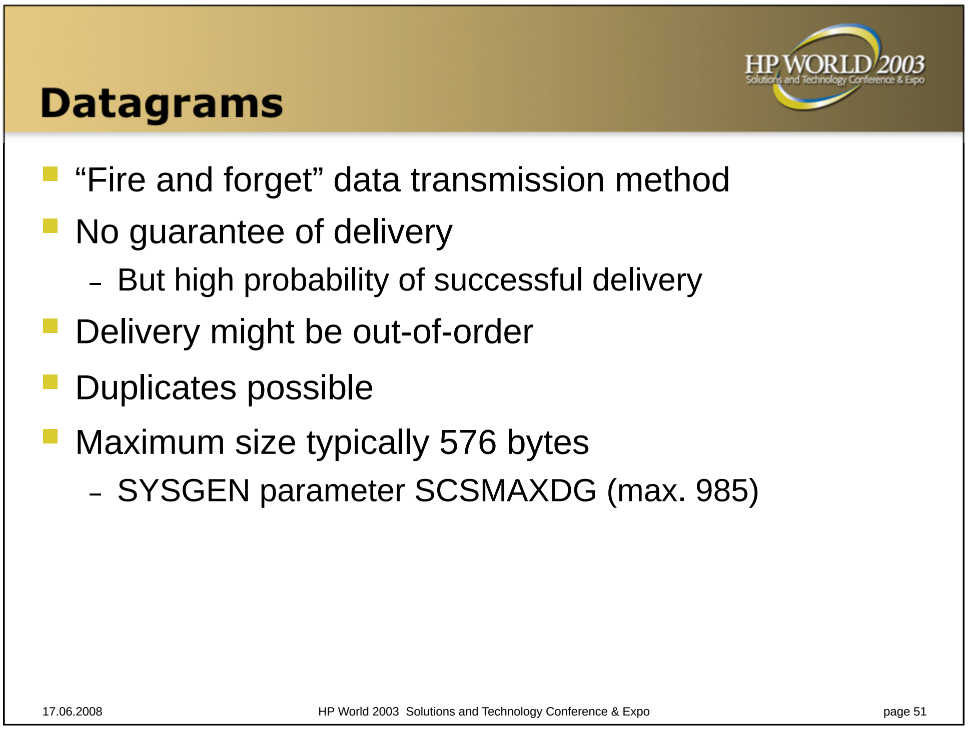
Node B

VMS$VAXcluster

MSCP Disk Server

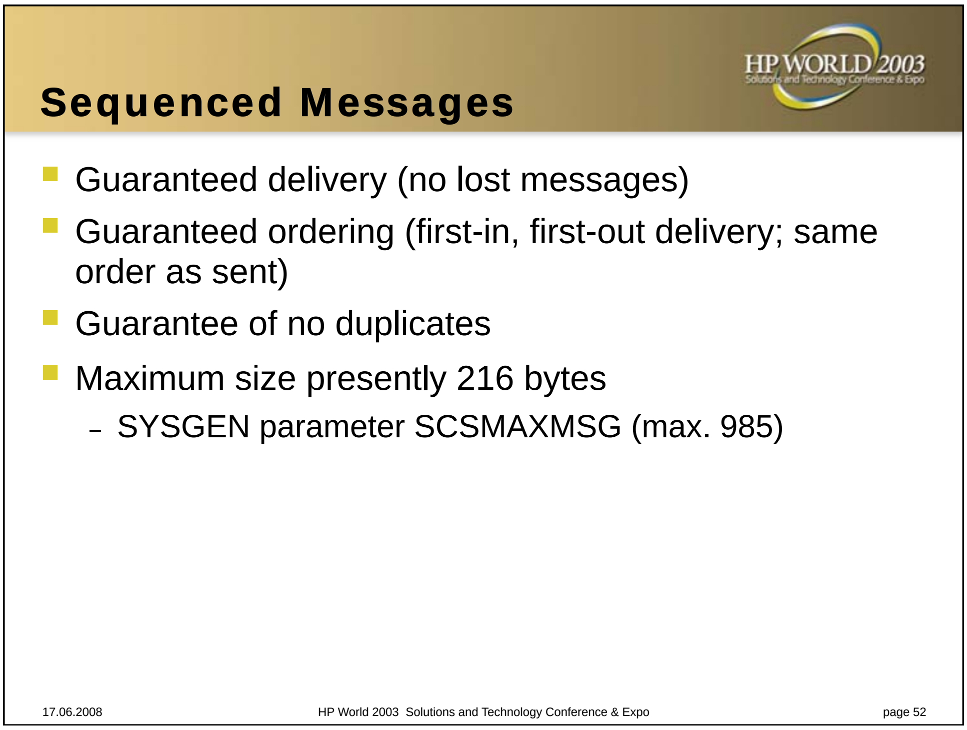
MSCP Tape Server

Virtual Circuit



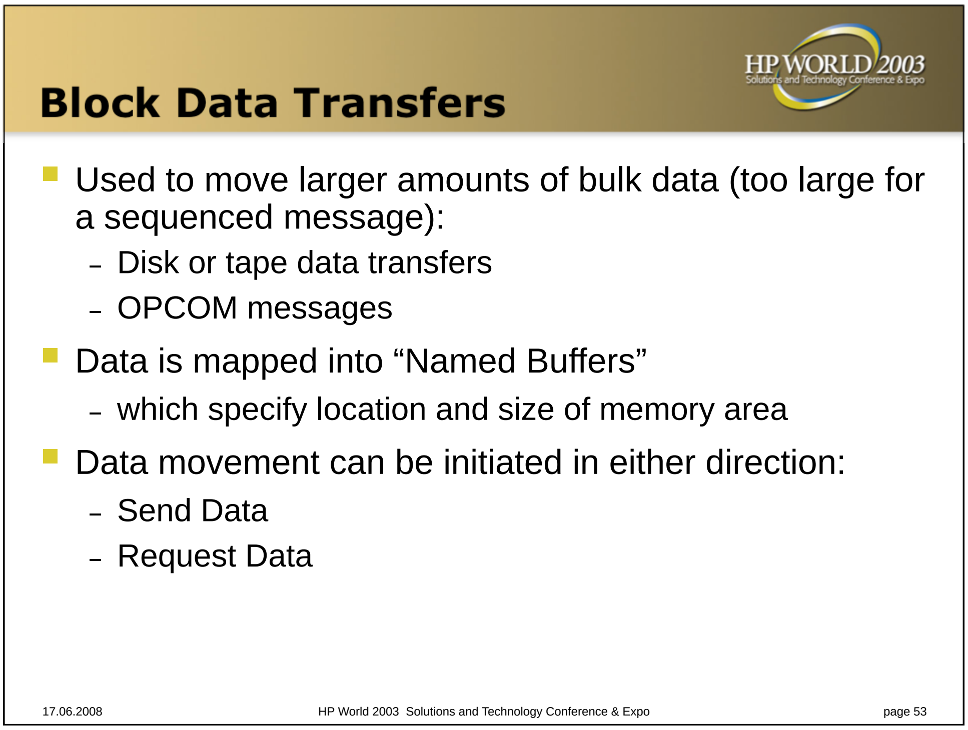
Datagrams

- “Fire and forget” data transmission method
- No guarantee of delivery
  - But high probability of successful delivery
- Delivery might be out-of-order
- Duplicates possible
- Maximum size typically 576 bytes
  - SYSGEN parameter SCSMAXDG (max. 985)



Sequenced Messages

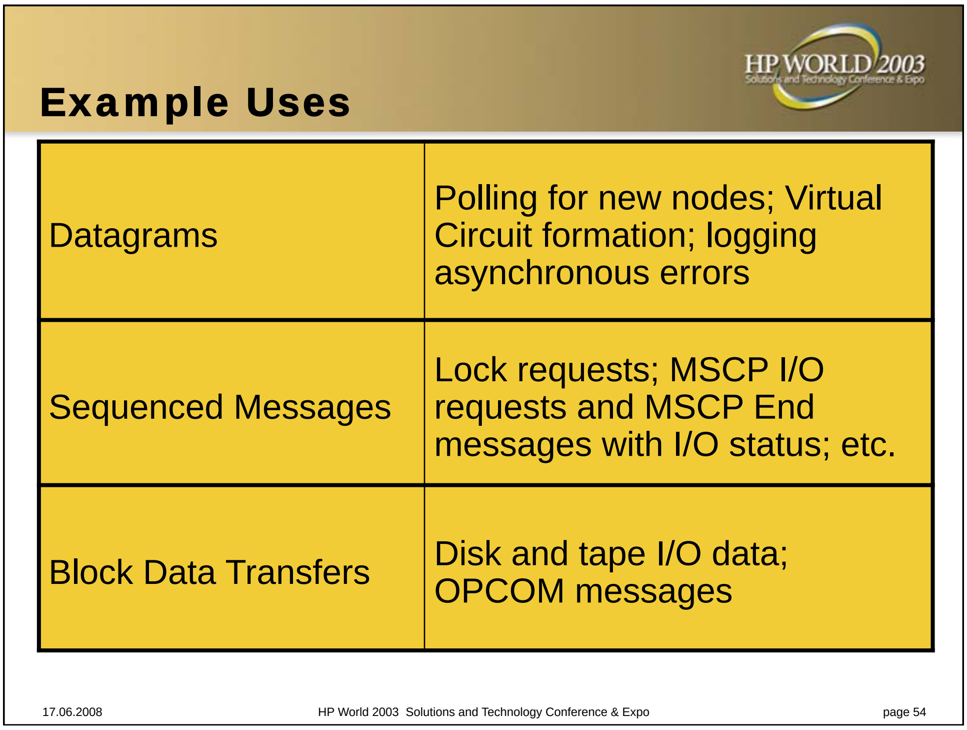
- Guaranteed delivery (no lost messages)
- Guaranteed ordering (first-in, first-out delivery; same order as sent)
- Guarantee of no duplicates
- Maximum size presently 216 bytes
  - SYSGEN parameter SCSMAXMSG (max. 985)



Block Data Transfers

- Used to move larger amounts of bulk data (too large for a sequenced message):
  - Disk or tape data transfers
  - OPCOM messages
- Data is mapped into “Named Buffers”
  - which specify location and size of memory area
- Data movement can be initiated in either direction:
  - Send Data
  - Request Data

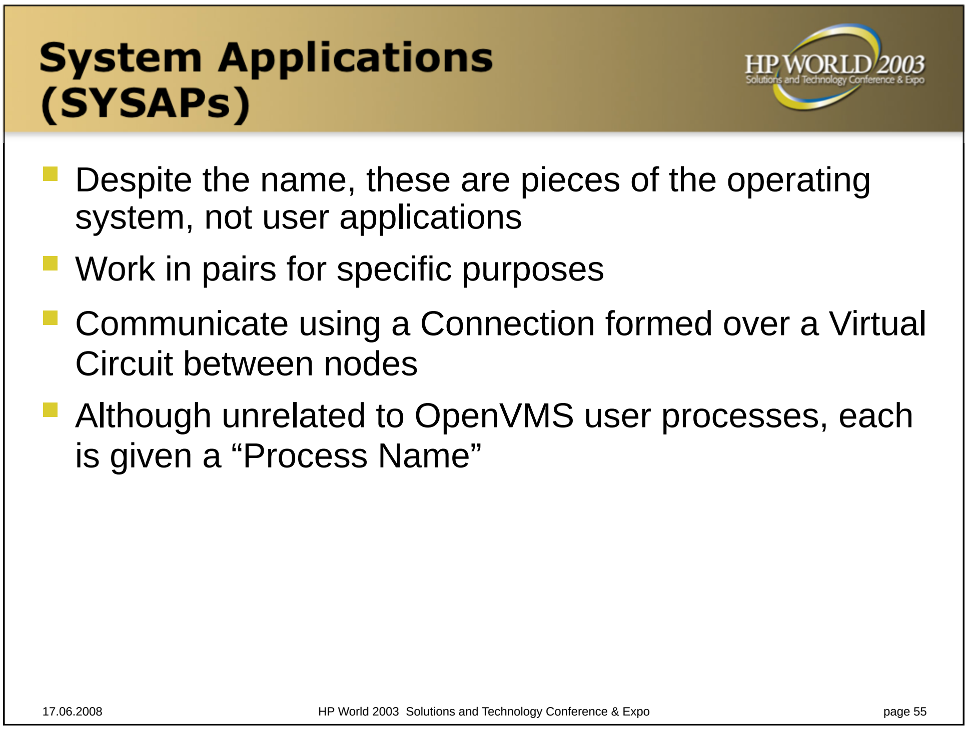
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 53



Example Uses

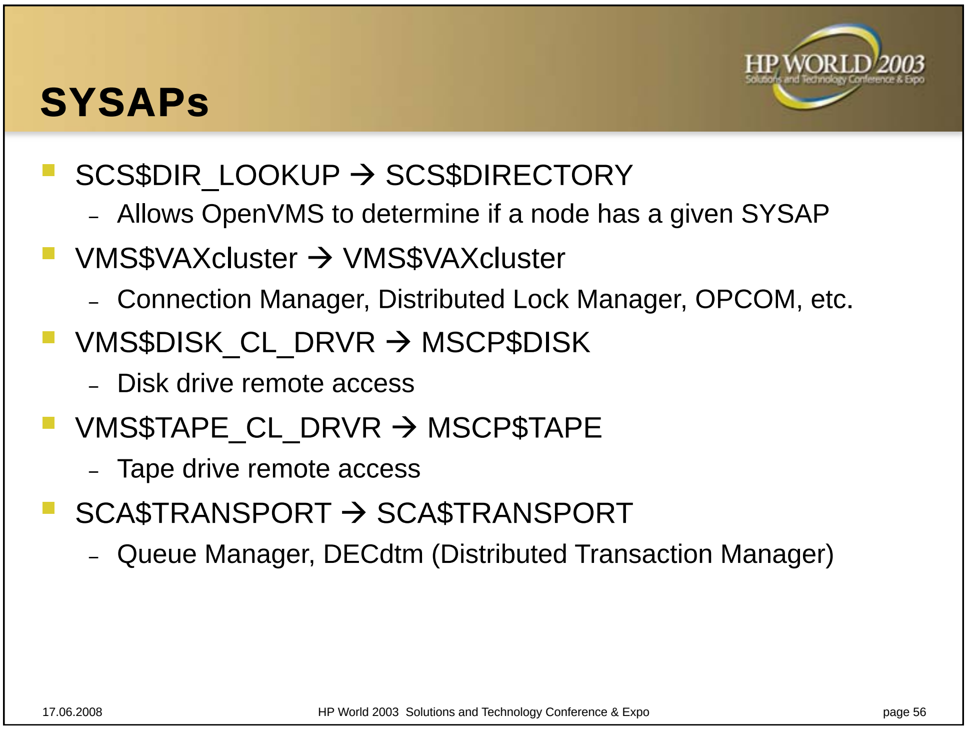
Datagrams	Polling for new nodes; Virtual Circuit formation; logging asynchronous errors
Sequenced Messages	Lock requests; MSCP I/O requests and MSCP End messages with I/O status; etc.
Block Data Transfers	Disk and tape I/O data; OPCOM messages

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 54



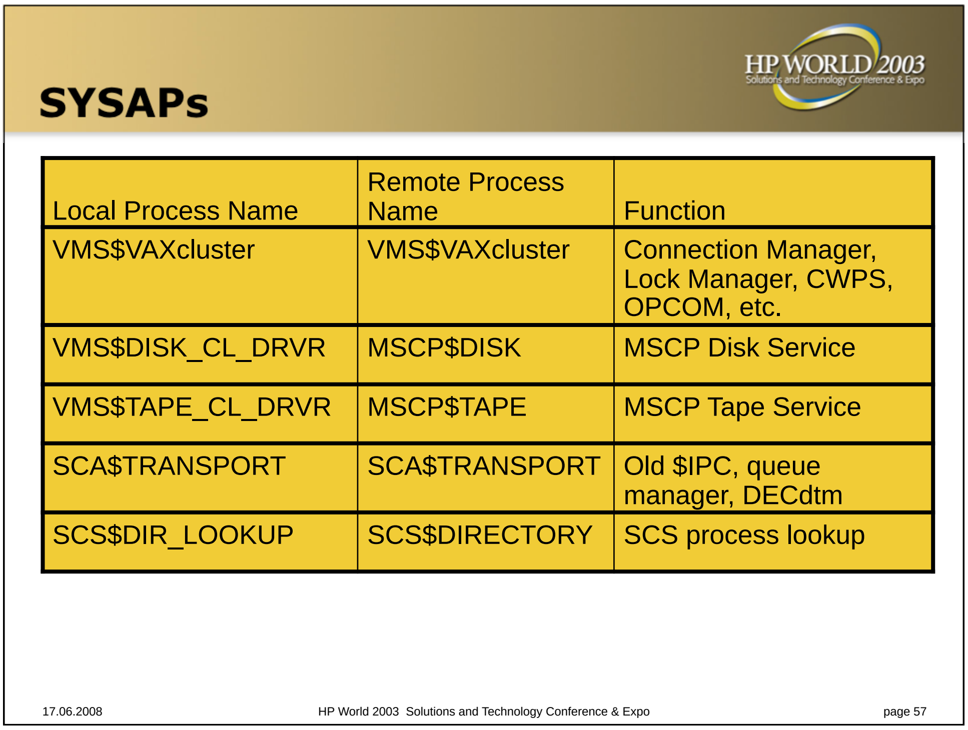
System Applications (SYSAPs)

- Despite the name, these are pieces of the operating system, not user applications
- Work in pairs for specific purposes
- Communicate using a Connection formed over a Virtual Circuit between nodes
- Although unrelated to OpenVMS user processes, each is given a “Process Name”



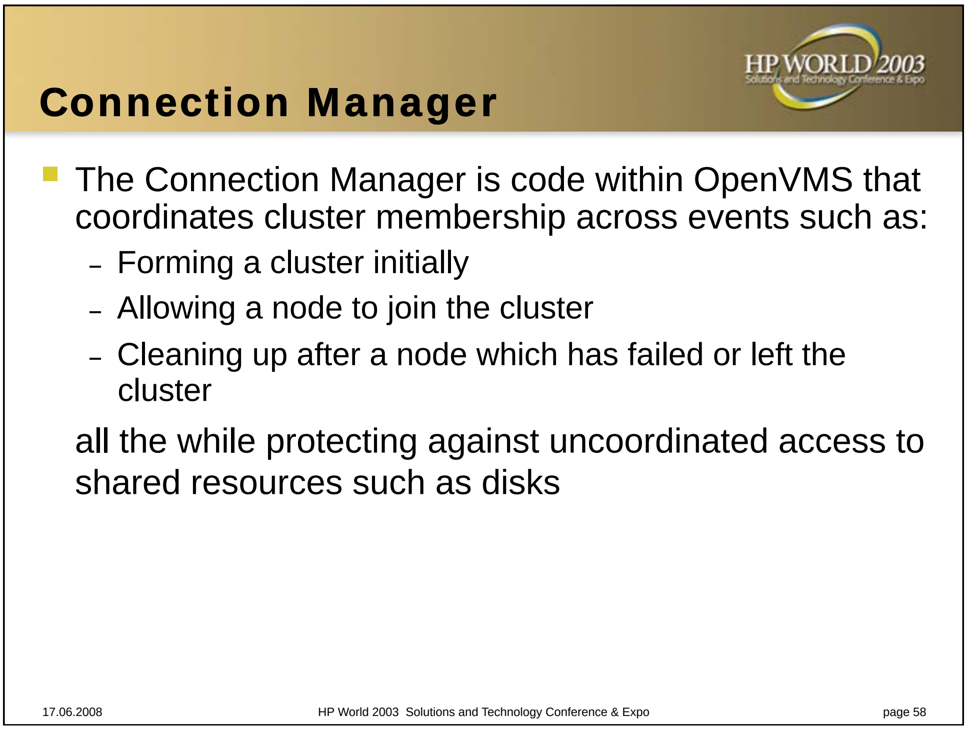
SYSAPs

- SCS$DIR_LOOKUP → SCS$DIRECTORY
  - Allows OpenVMS to determine if a node has a given SYSAP
- VMS$VAXcluster → VMS$VAXcluster
  - Connection Manager, Distributed Lock Manager, OPCOM, etc.
- VMS$DISK_CL_DRVR → MSCP$DISK
  - Disk drive remote access
- VMS$TAPE_CL_DRVR → MSCP$TAPE
  - Tape drive remote access
- SCA$TRANSPORT → SCA$TRANSPORT
  - Queue Manager, DECdtm (Distributed Transaction Manager)



SYSAPs

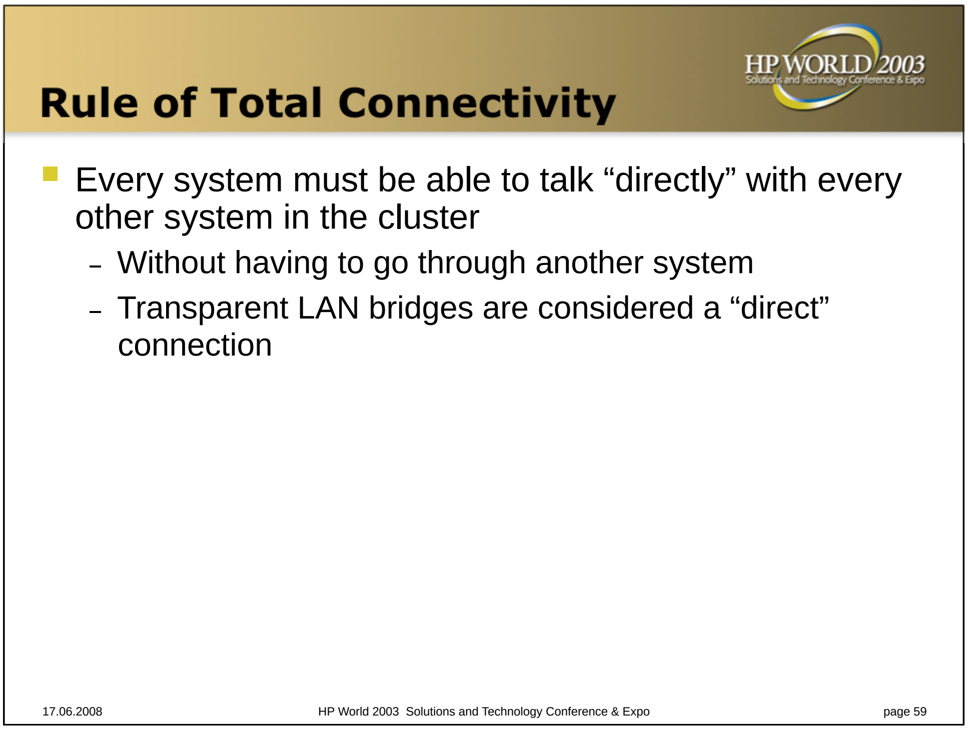
Local Process Name	Remote Process Name	Function
VMS$VAXcluster	VMS$VAXcluster	Connection Manager, Lock Manager, CWPS, OPCOM, etc.
VMS$DISK_CL_DRVR	MSCP$DISK	MSCP Disk Service
VMS$TAPE_CL_DRVR	MSCP$TAPE	MSCP Tape Service
SCA$TRANSPORT	SCA$TRANSPORT	Old $IPC, queue manager, DECdtm
SCS$DIR_LOOKUP	SCS$DIRECTORY	SCS process lookup



Connection Manager

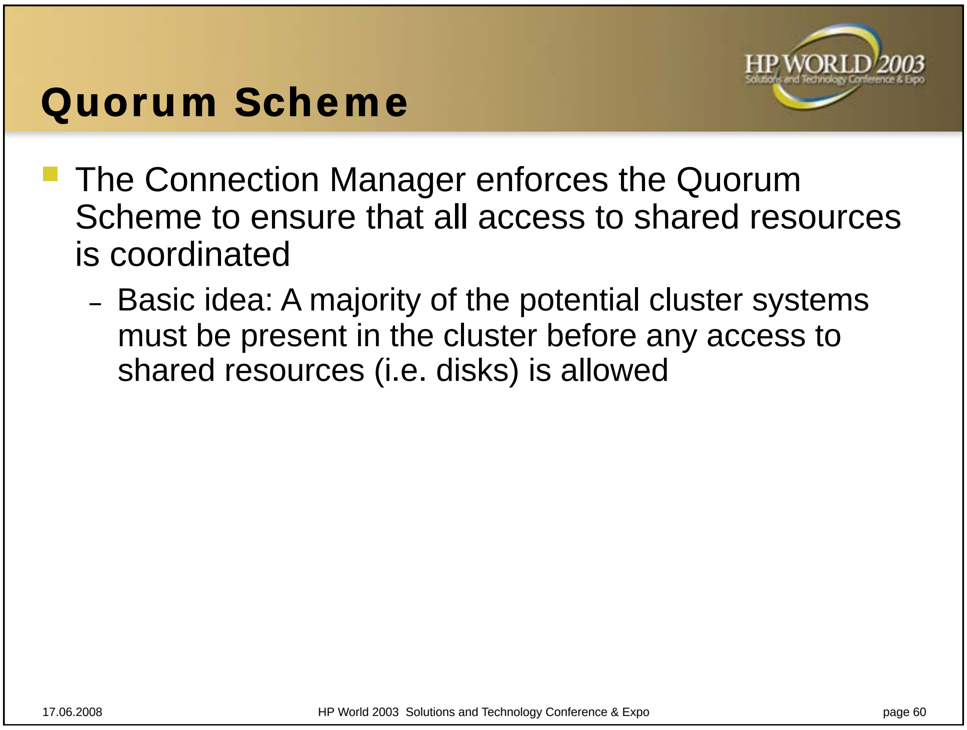
- The Connection Manager is code within OpenVMS that coordinates cluster membership across events such as:
  - Forming a cluster initially
  - Allowing a node to join the cluster
  - Cleaning up after a node which has failed or left the cluster

  all the while protecting against uncoordinated access to shared resources such as disks



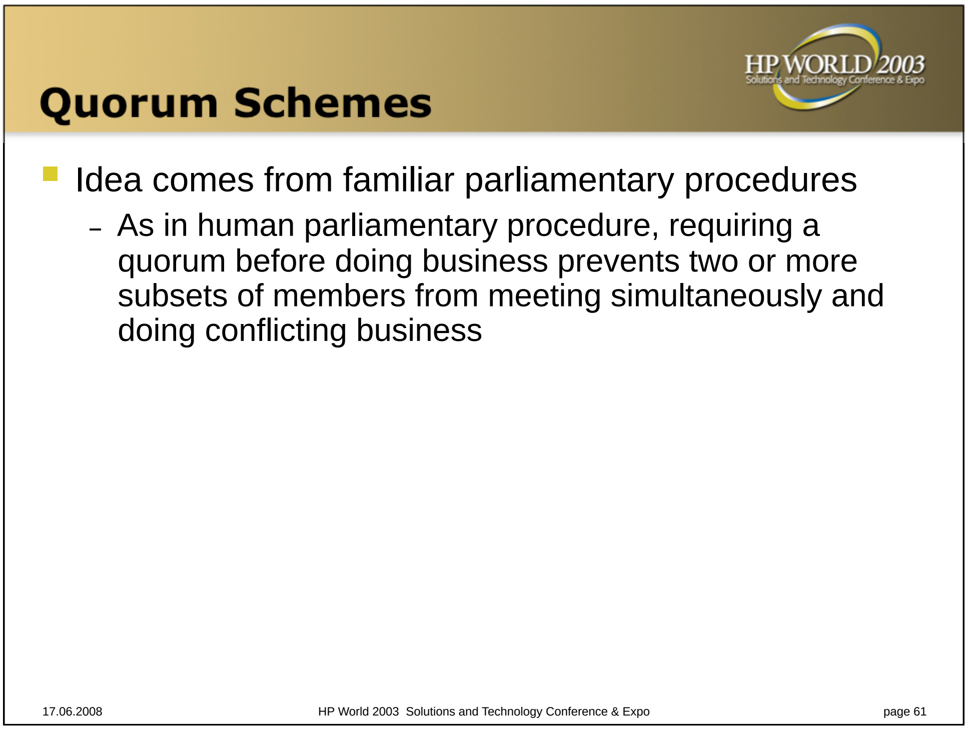
Rule of Total Connectivity

- Every system must be able to talk “directly” with every other system in the cluster
  - Without having to go through another system
  - Transparent LAN bridges are considered a “direct” connection



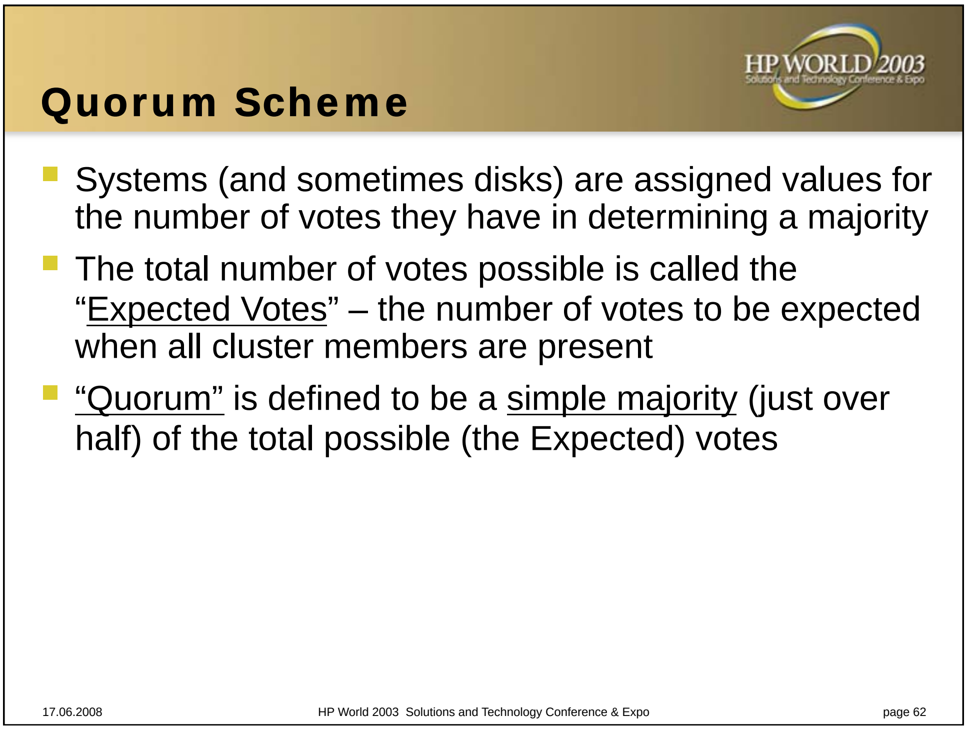
Quorum Scheme

- The Connection Manager enforces the Quorum Scheme to ensure that all access to shared resources is coordinated
  - Basic idea: A majority of the potential cluster systems must be present in the cluster before any access to shared resources (i.e. disks) is allowed



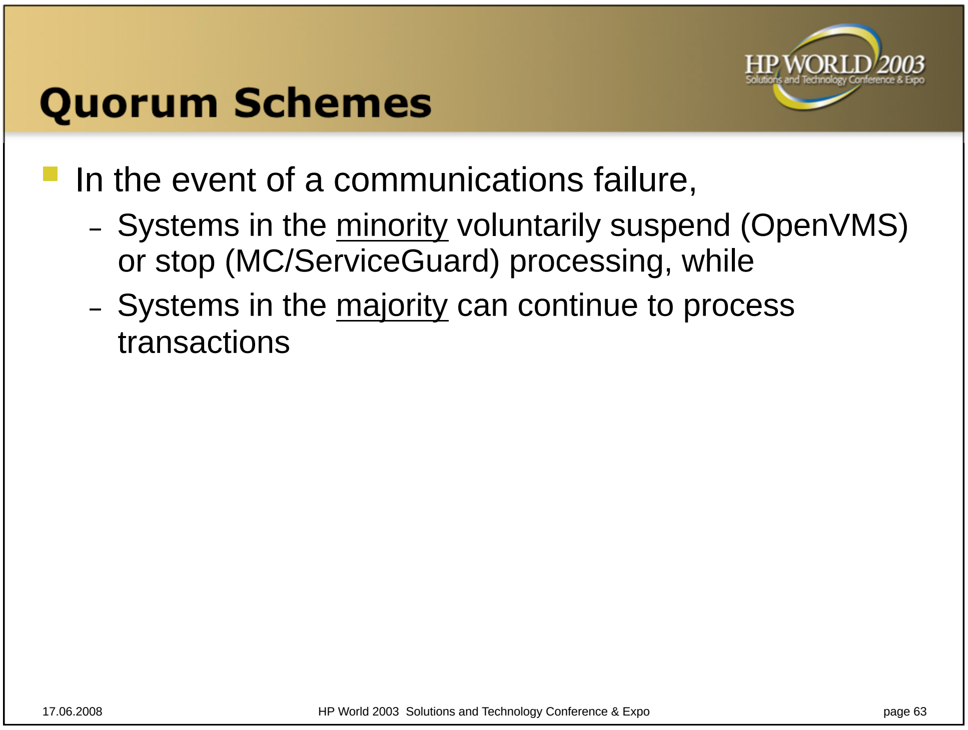
Quorum Schemes

- Idea comes from familiar parliamentary procedures
  - As in human parliamentary procedure, requiring a quorum before doing business prevents two or more subsets of members from meeting simultaneously and doing conflicting business



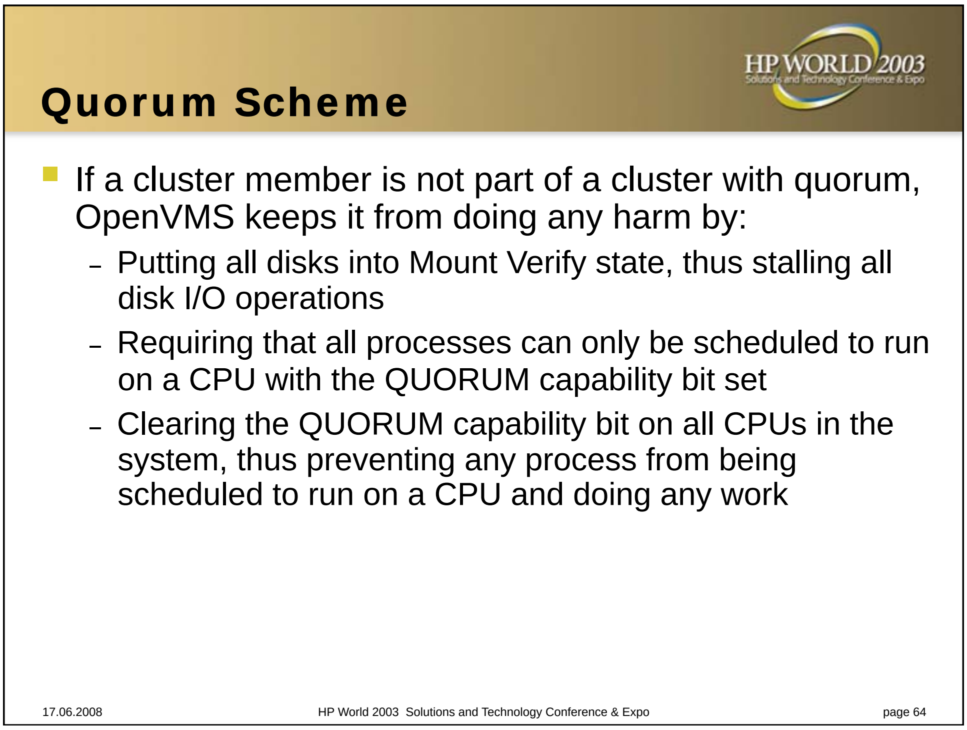
Quorum Scheme

- Systems (and sometimes disks) are assigned values for the number of votes they have in determining a majority
- The total number of votes possible is called the "Expected Votes" – the number of votes to be expected when all cluster members are present
- "Quorum" is defined to be a simple majority (just over half) of the total possible (the Expected) votes



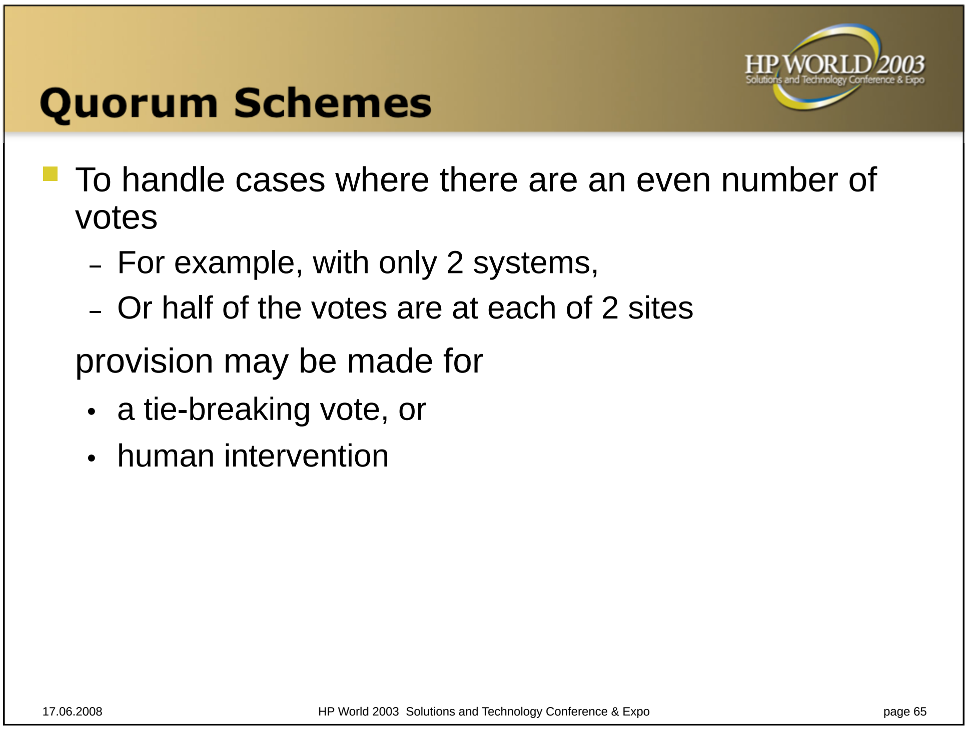
Quorum Schemes

- In the event of a communications failure,
  - Systems in the minority voluntarily suspend (OpenVMS) or stop (MC/ServiceGuard) processing, while
  - Systems in the majority can continue to process transactions



Quorum Scheme

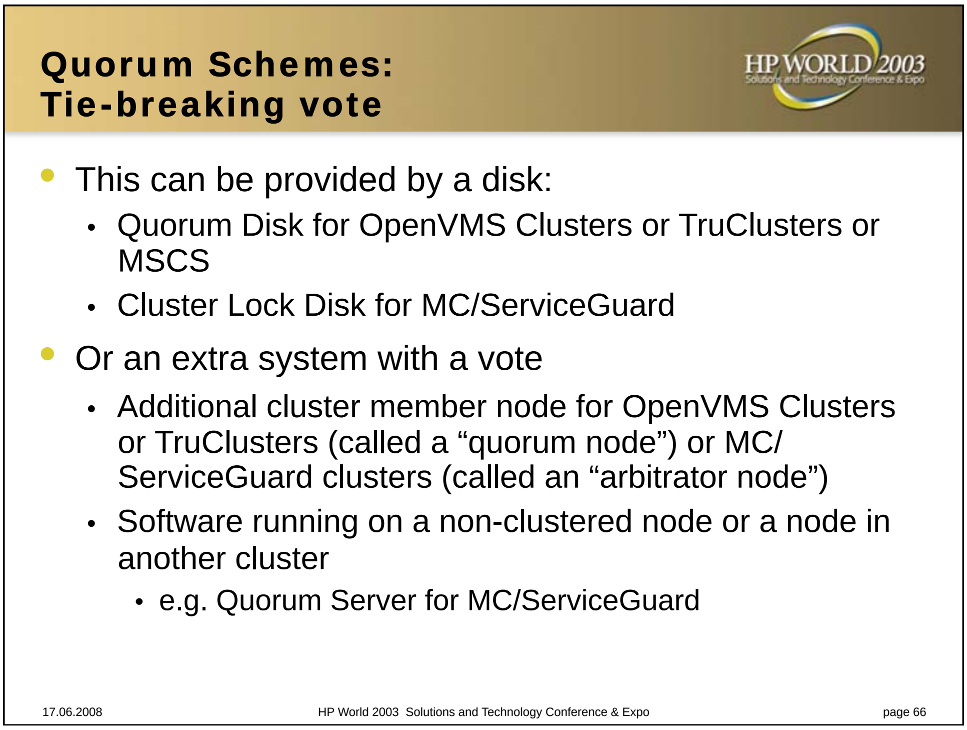
- If a cluster member is not part of a cluster with quorum, OpenVMS keeps it from doing any harm by:
  - Putting all disks into Mount Verify state, thus stalling all disk I/O operations
  - Requiring that all processes can only be scheduled to run on a CPU with the QUORUM capability bit set
  - Clearing the QUORUM capability bit on all CPUs in the system, thus preventing any process from being scheduled to run on a CPU and doing any work



Quorum Schemes

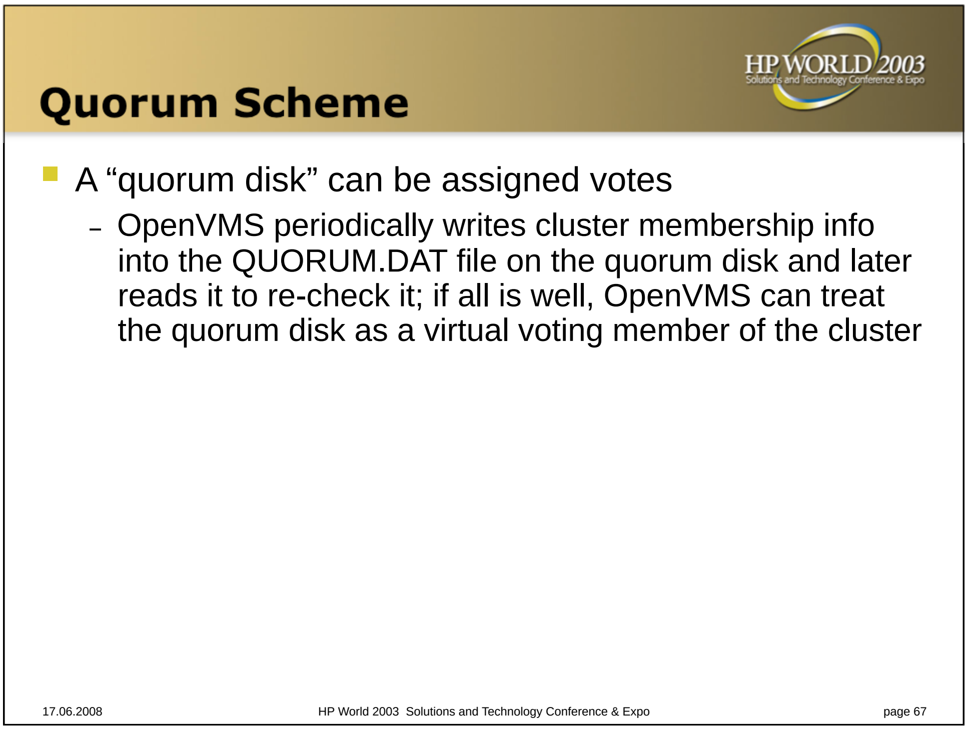
- To handle cases where there are an even number of votes
  - For example, with only 2 systems,
  - Or half of the votes are at each of 2 sites

  provision may be made for
  - a tie-breaking vote, or
  - human intervention



Quorum Schemes: Tie-breaking vote

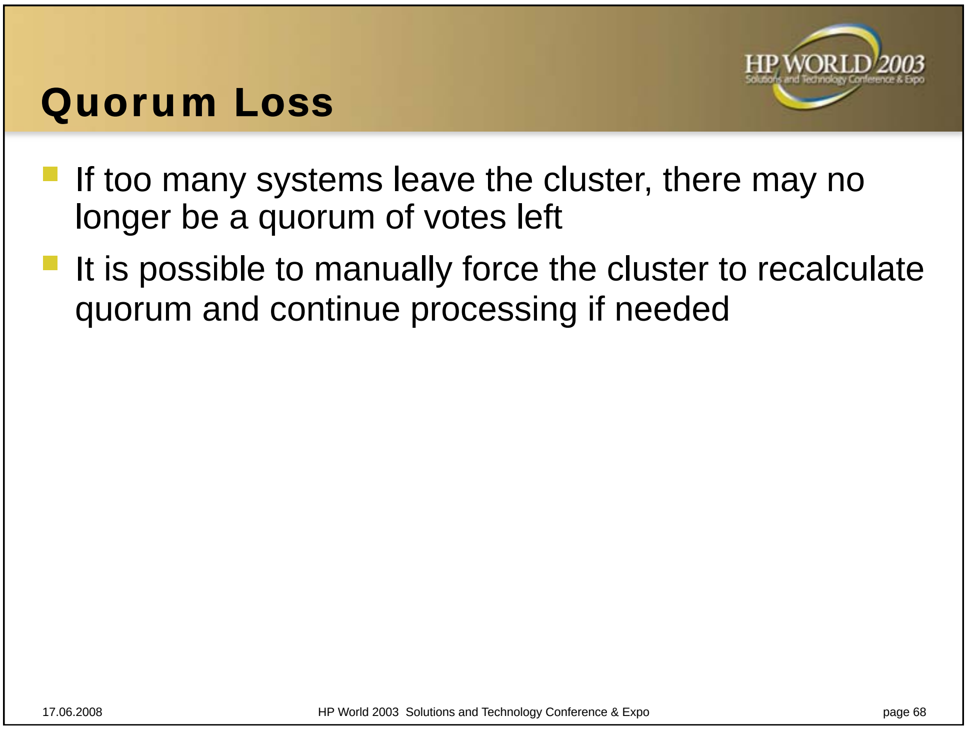
- This can be provided by a disk:
  - Quorum Disk for OpenVMS Clusters or TruClusters or MSCS
  - Cluster Lock Disk for MC/ServiceGuard
- Or an extra system with a vote
  - Additional cluster member node for OpenVMS Clusters or TruClusters (called a “quorum node”) or MC/ ServiceGuard clusters (called an “arbitrator node”)
  - Software running on a non-clustered node or a node in another cluster
    - e.g. Quorum Server for MC/ServiceGuard



Quorum Scheme

- A “quorum disk” can be assigned votes
  - OpenVMS periodically writes cluster membership info into the QUORUM.DAT file on the quorum disk and later reads it to re-check it; if all is well, OpenVMS can treat the quorum disk as a virtual voting member of the cluster

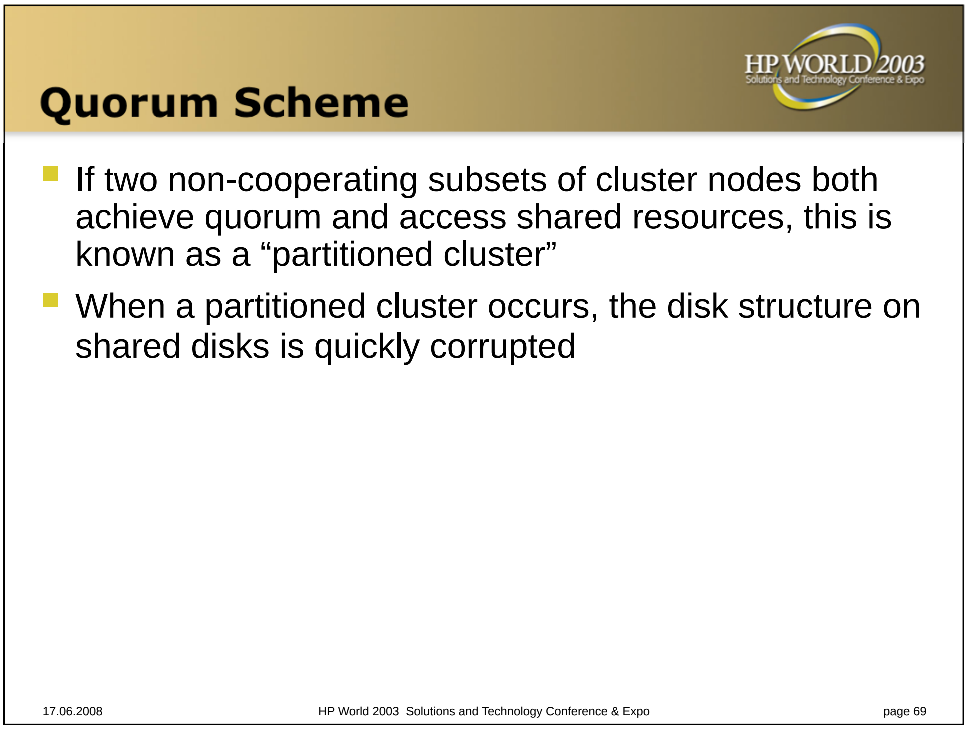
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 67



Quorum Loss

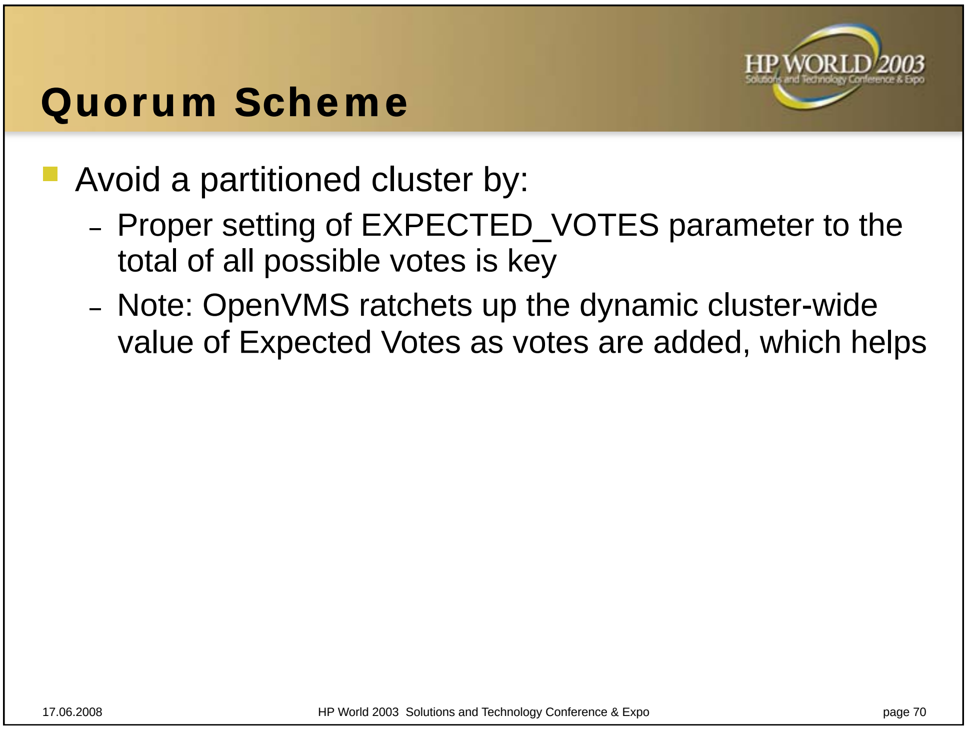
- If too many systems leave the cluster, there may no longer be a quorum of votes left
- It is possible to manually force the cluster to recalculate quorum and continue processing if needed

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 68



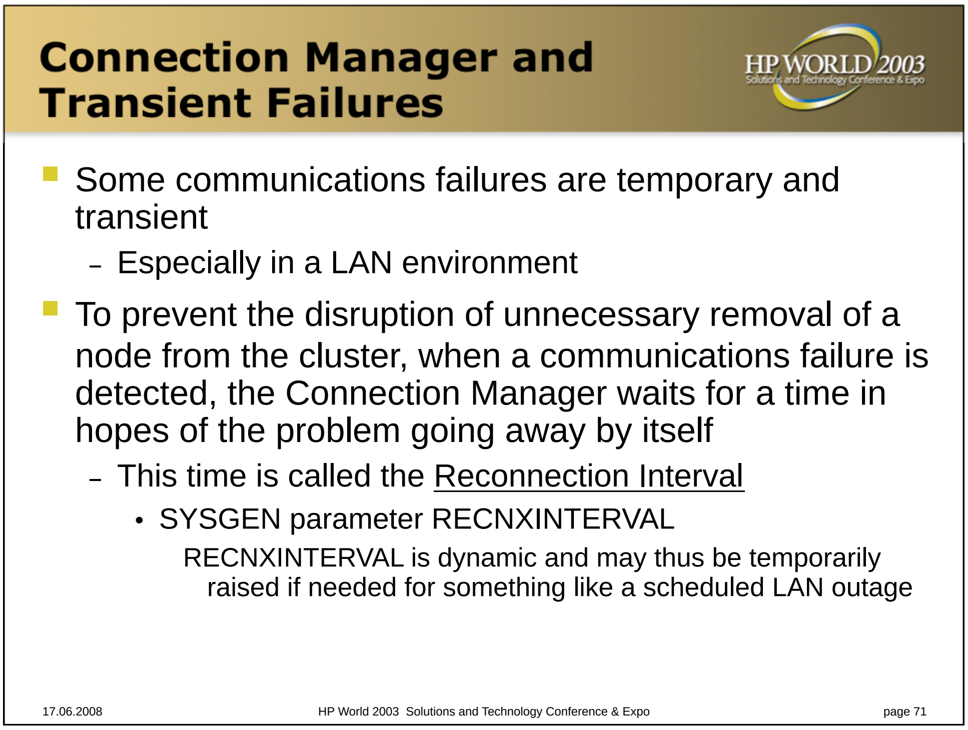
Quorum Scheme

- If two non-cooperating subsets of cluster nodes both achieve quorum and access shared resources, this is known as a “partitioned cluster”
- When a partitioned cluster occurs, the disk structure on shared disks is quickly corrupted



Quorum Scheme

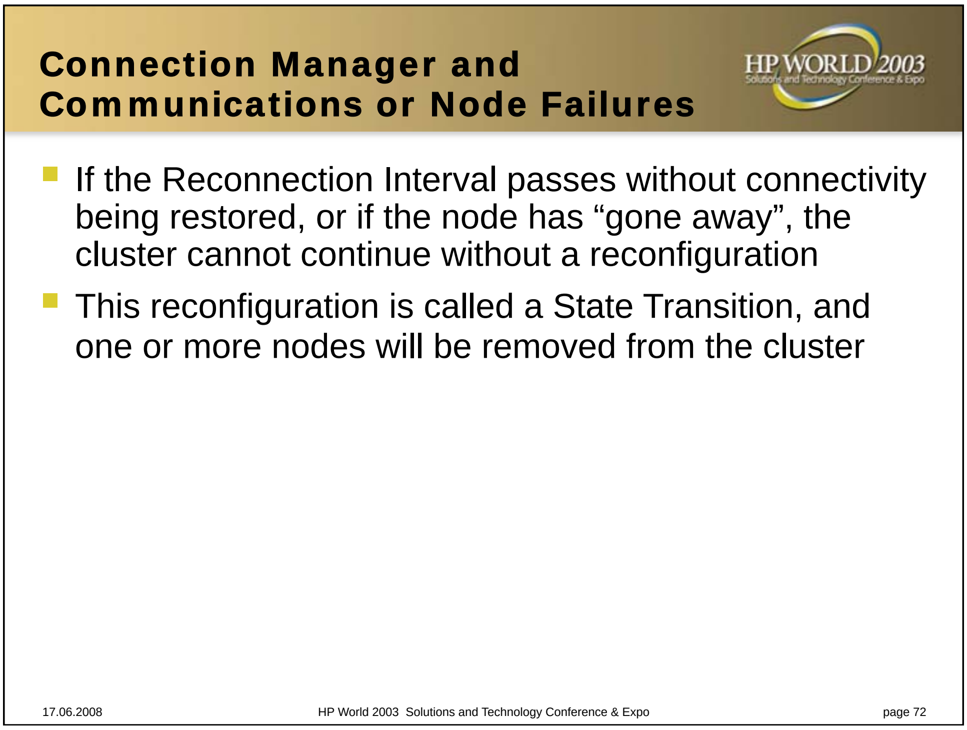
- Avoid a partitioned cluster by:
  - Proper setting of EXPECTED_VOTES parameter to the total of all possible votes is key
  - Note: OpenVMS ratchets up the dynamic cluster-wide value of Expected Votes as votes are added, which helps



Connection Manager and Transient Failures

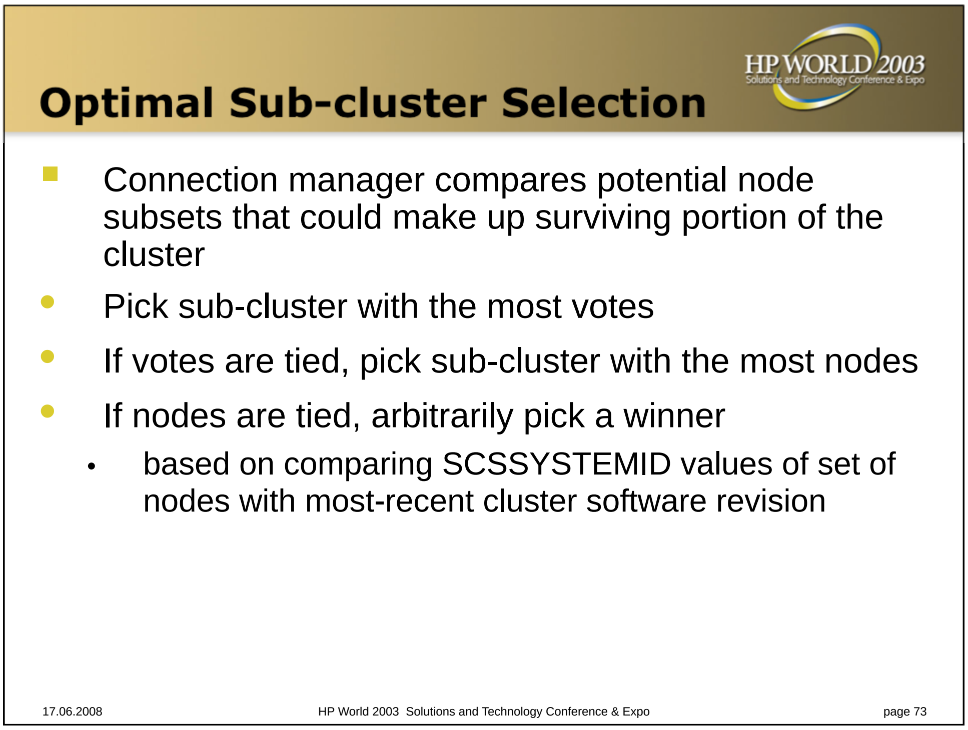
- Some communications failures are temporary and transient
  - Especially in a LAN environment
- To prevent the disruption of unnecessary removal of a node from the cluster, when a communications failure is detected, the Connection Manager waits for a time in hopes of the problem going away by itself
  - This time is called the Reconnection Interval
    - SYSGEN parameter RECNXINTERVAL

      RECNXINTERVAL is dynamic and may thus be temporarily raised if needed for something like a scheduled LAN outage



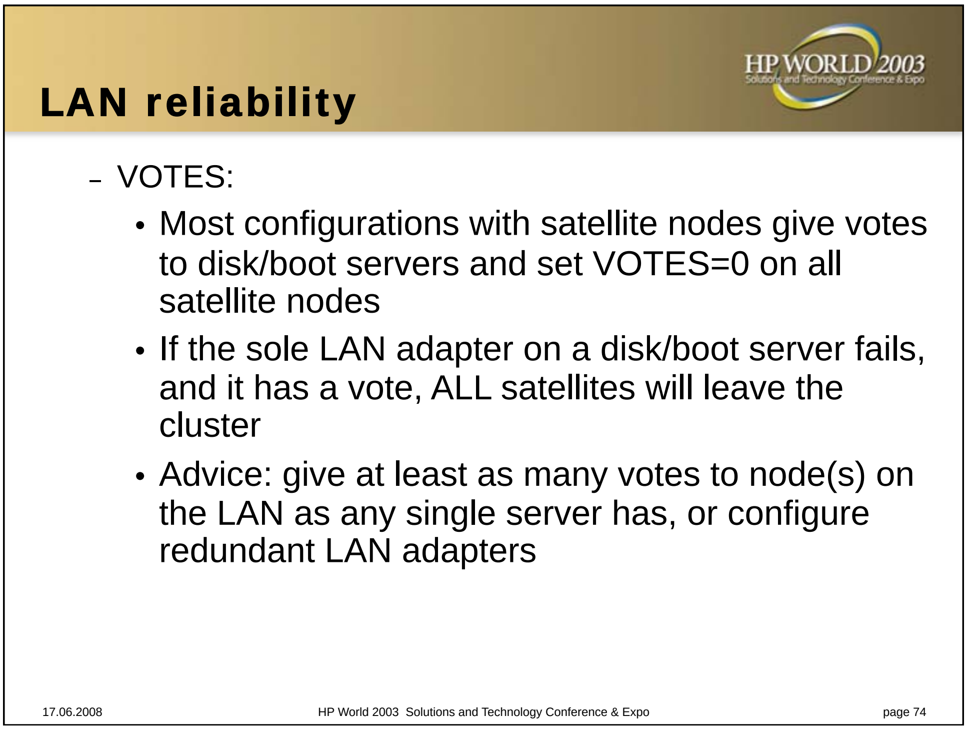
Connection Manager and Communications or Node Failures

- If the Reconnection Interval passes without connectivity being restored, or if the node has “gone away”, the cluster cannot continue without a reconfiguration
- This reconfiguration is called a State Transition, and one or more nodes will be removed from the cluster



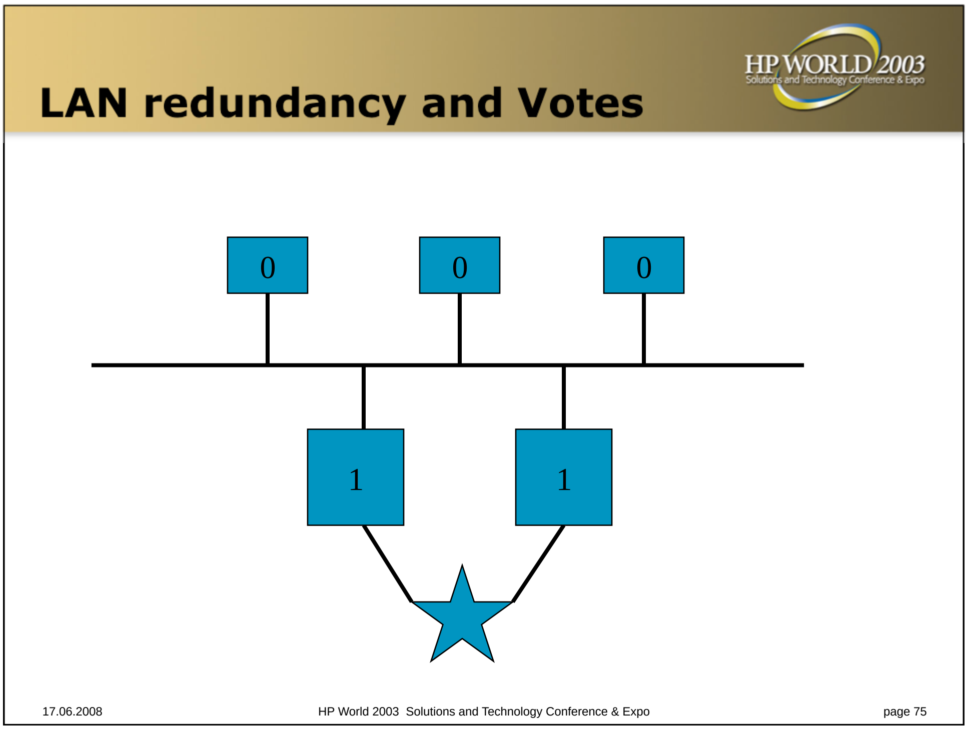
Optimal Sub-cluster Selection

- Connection manager compares potential node subsets that could make up surviving portion of the cluster
- Pick sub-cluster with the most votes
- If votes are tied, pick sub-cluster with the most nodes
- If nodes are tied, arbitrarily pick a winner
  - based on comparing SCSSYSTEMID values of set of nodes with most-recent cluster software revision



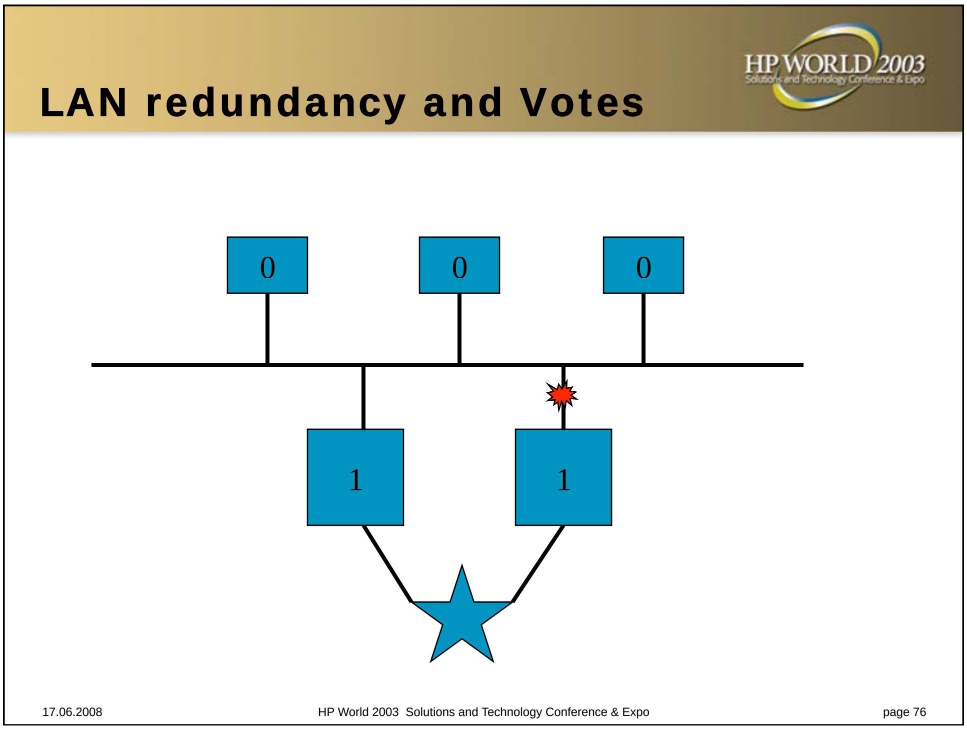
LAN reliability

- VOTES:
  - Most configurations with satellite nodes give votes to disk/boot servers and set VOTES=0 on all satellite nodes
  - If the sole LAN adapter on a disk/boot server fails, and it has a vote, ALL satellites will leave the cluster
  - Advice: give at least as many votes to node(s) on the LAN as any single server has, or configure redundant LAN adapters



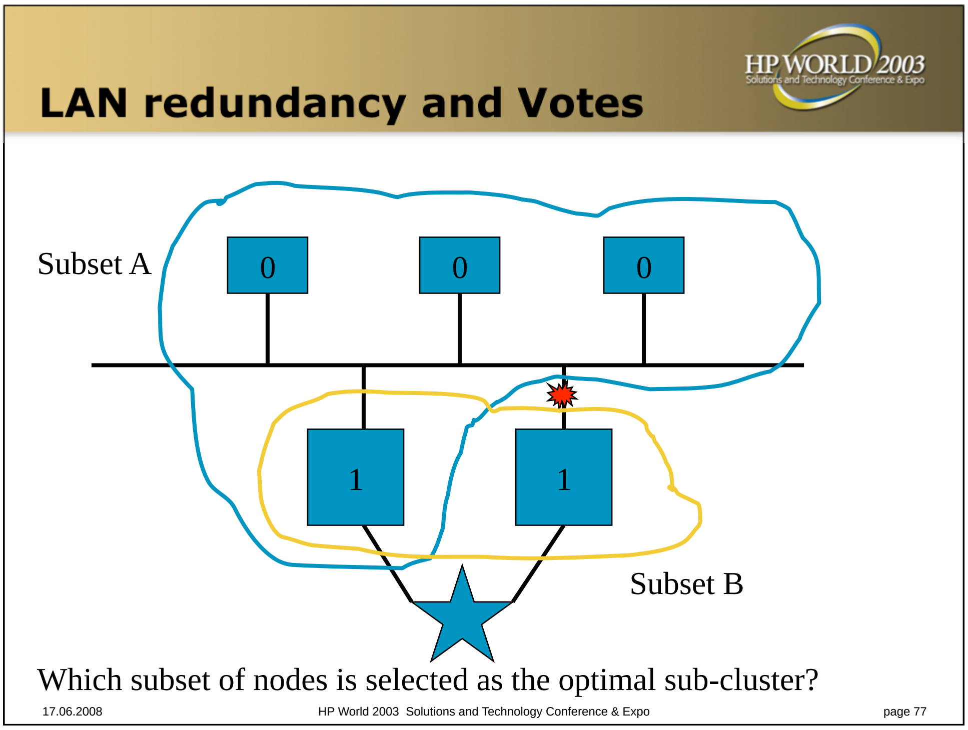
LAN redundancy and Votes

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 75



LAN redundancy and Votes

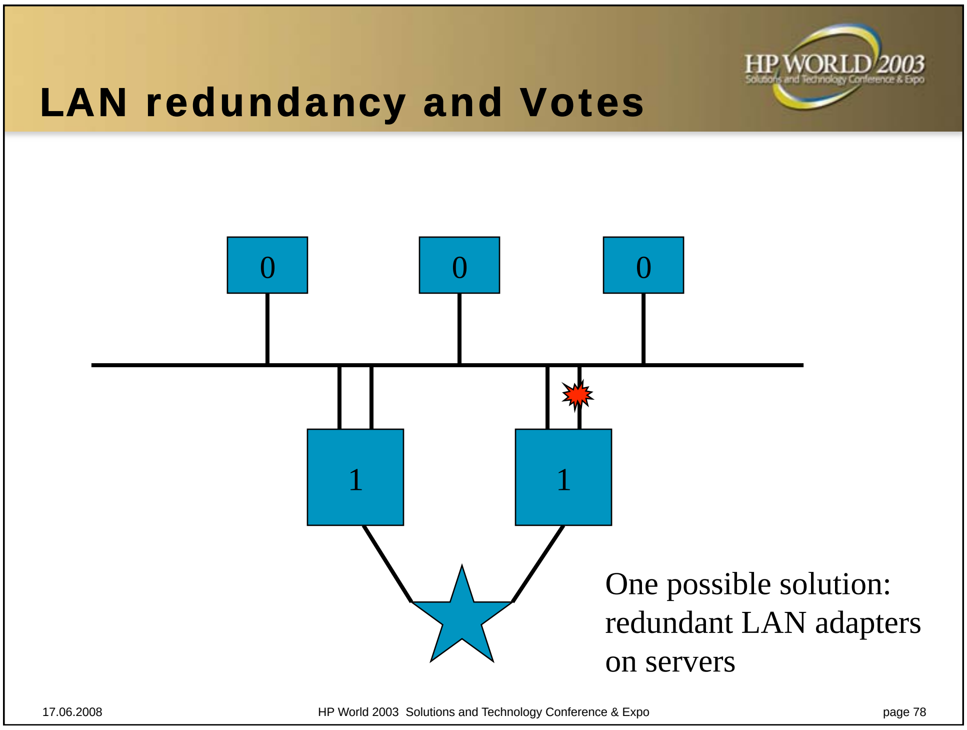
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 76



LAN redundancy and Votes

Which subset of nodes is selected as the optimal sub-cluster?

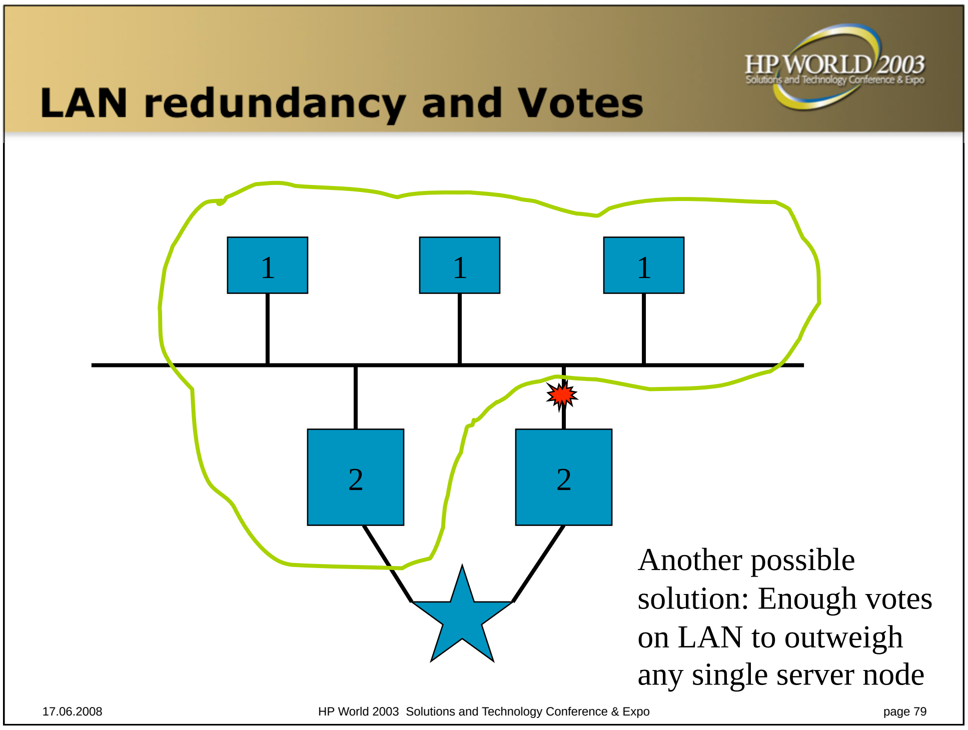
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 77



LAN redundancy and Votes

One possible solution: redundant LAN adapters on servers

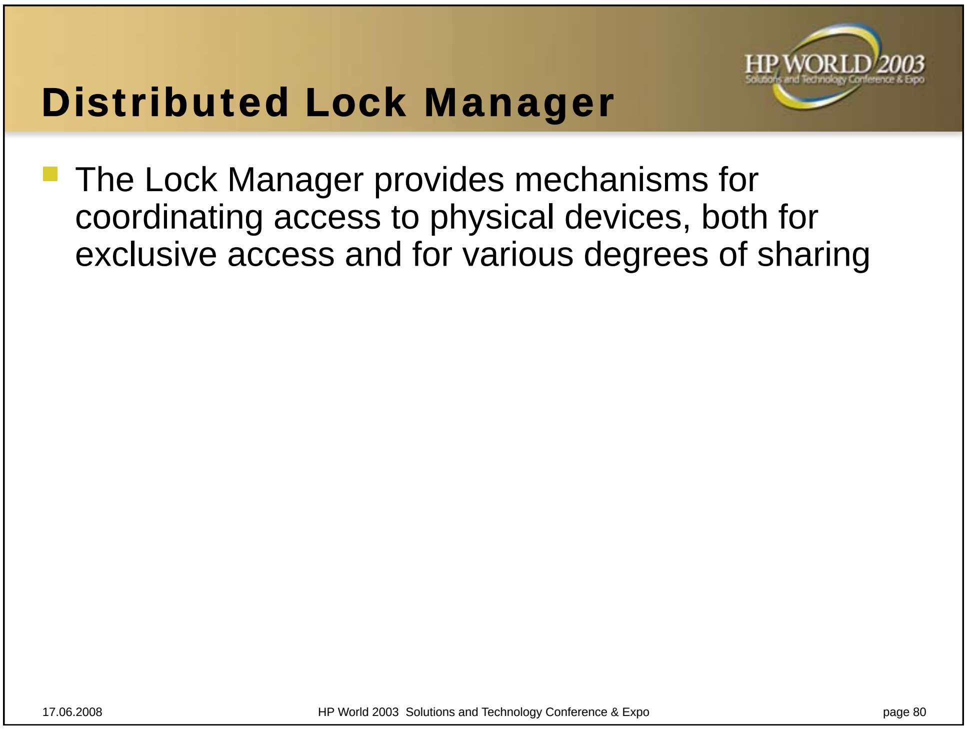
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 78



LAN redundancy and Votes

Another possible solution: Enough votes on LAN to outweigh any single server node

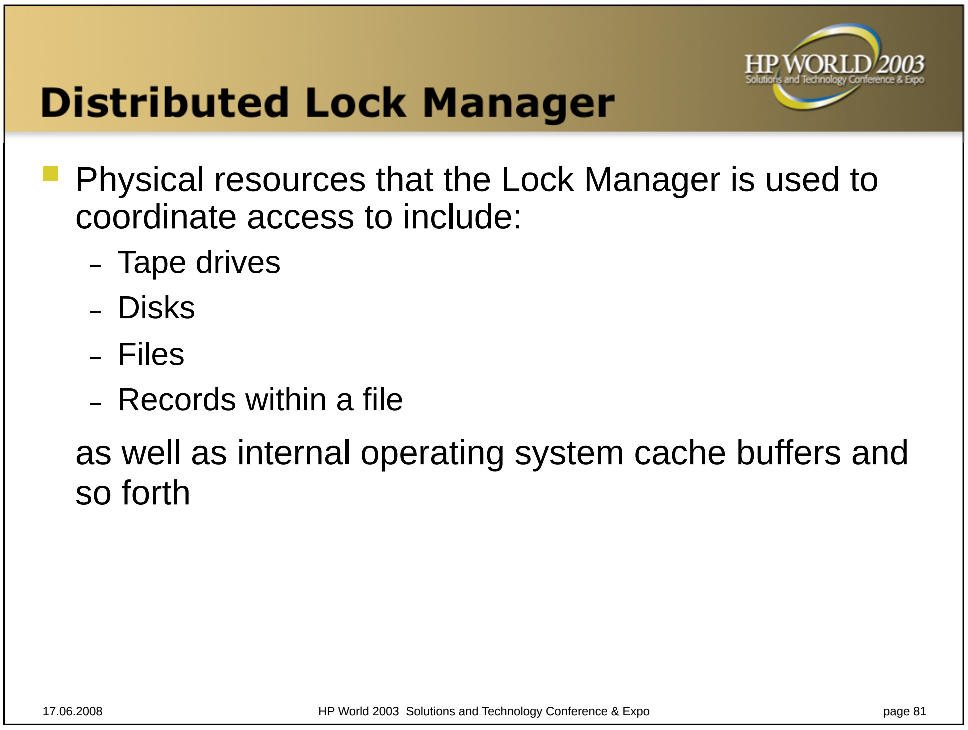
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 79



Distributed Lock Manager

- The Lock Manager provides mechanisms for coordinating access to physical devices, both for exclusive access and for various degrees of sharing

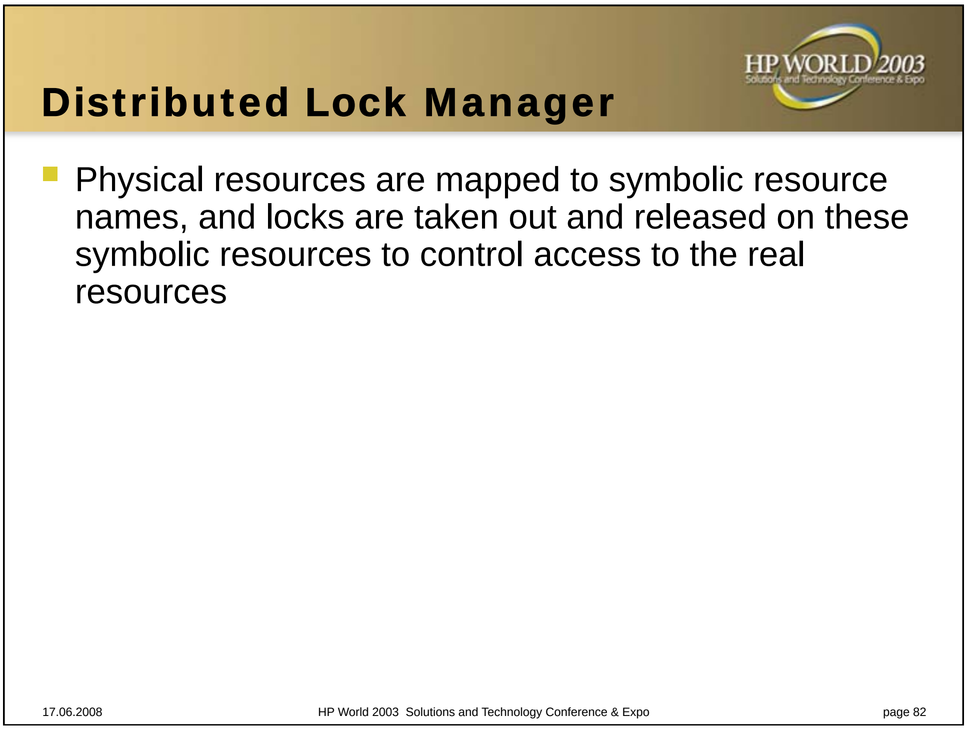
17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 80



Distributed Lock Manager

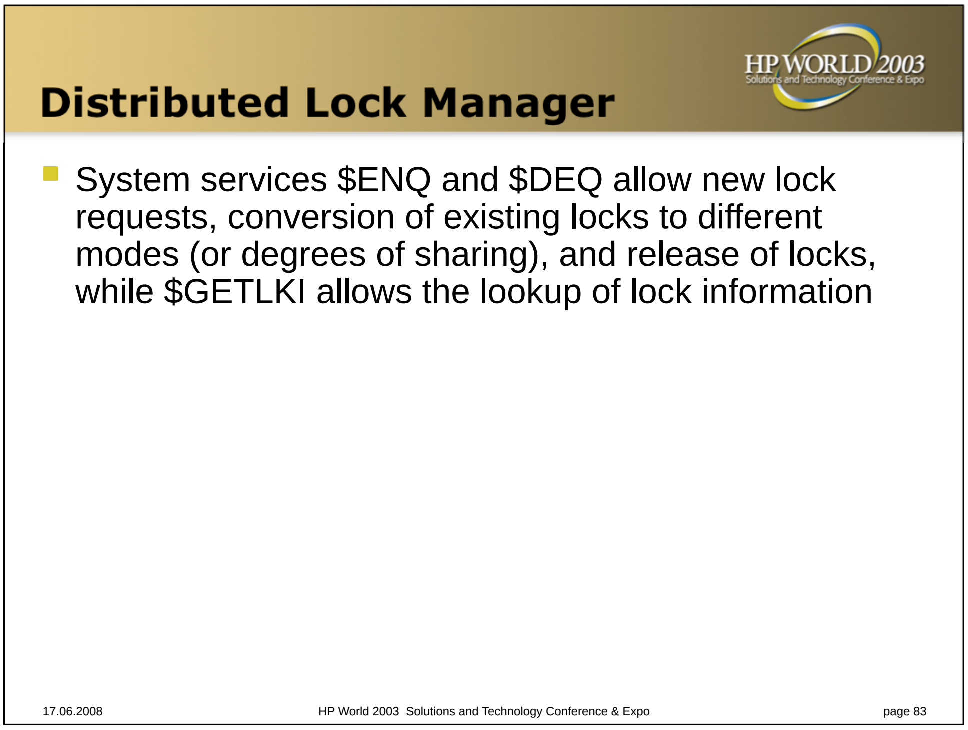
- Physical resources that the Lock Manager is used to coordinate access to include:
  - Tape drives
  - Disks
  - Files
  - Records within a file

  as well as internal operating system cache buffers and so forth



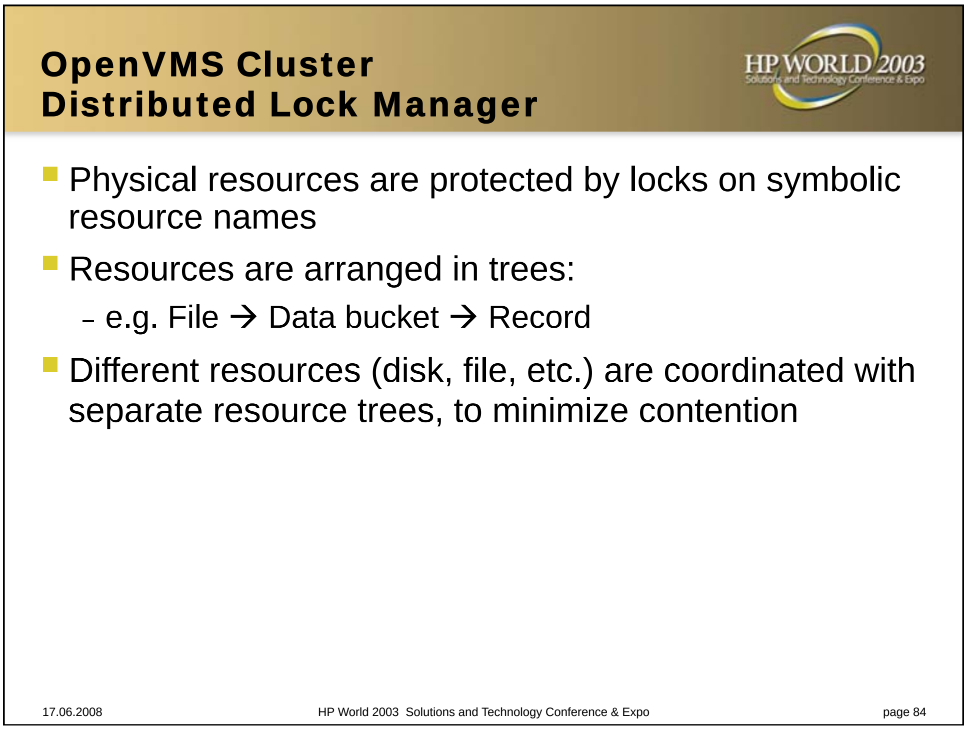
Distributed Lock Manager

- Physical resources are mapped to symbolic resource names, and locks are taken out and released on these symbolic resources to control access to the real resources



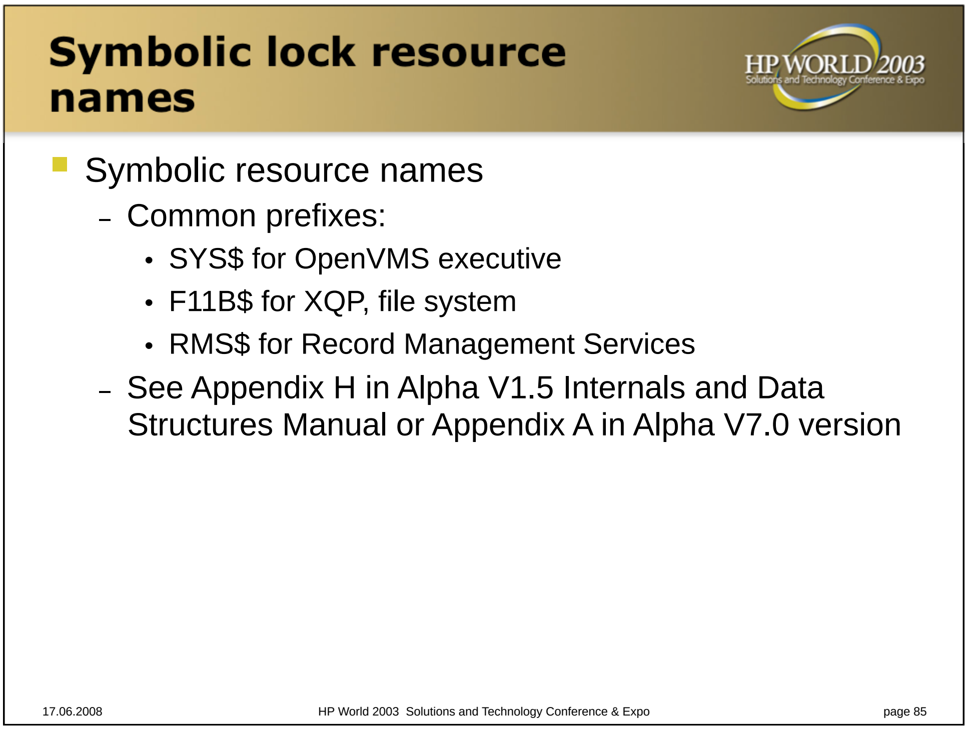
Distributed Lock Manager

- System services $ENQ and $DEQ allow new lock requests, conversion of existing locks to different modes (or degrees of sharing), and release of locks, while $GETLKI allows the lookup of lock information



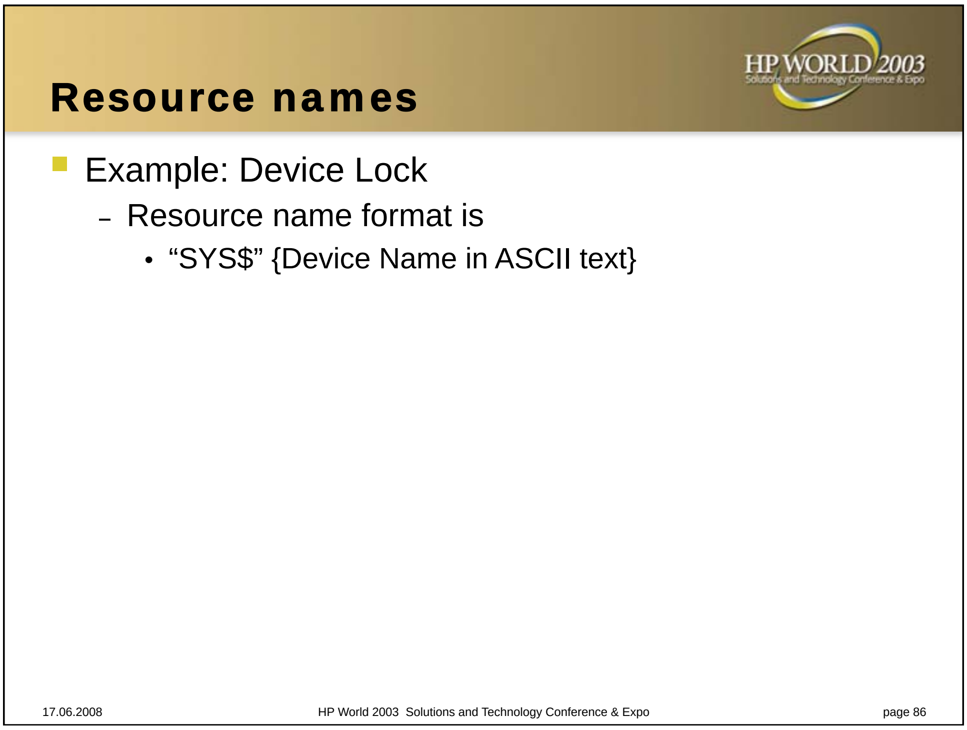
OpenVMS Cluster Distributed Lock Manager

- Physical resources are protected by locks on symbolic resource names
- Resources are arranged in trees:
  - e.g. File → Data bucket → Record
- Different resources (disk, file, etc.) are coordinated with separate resource trees, to minimize contention



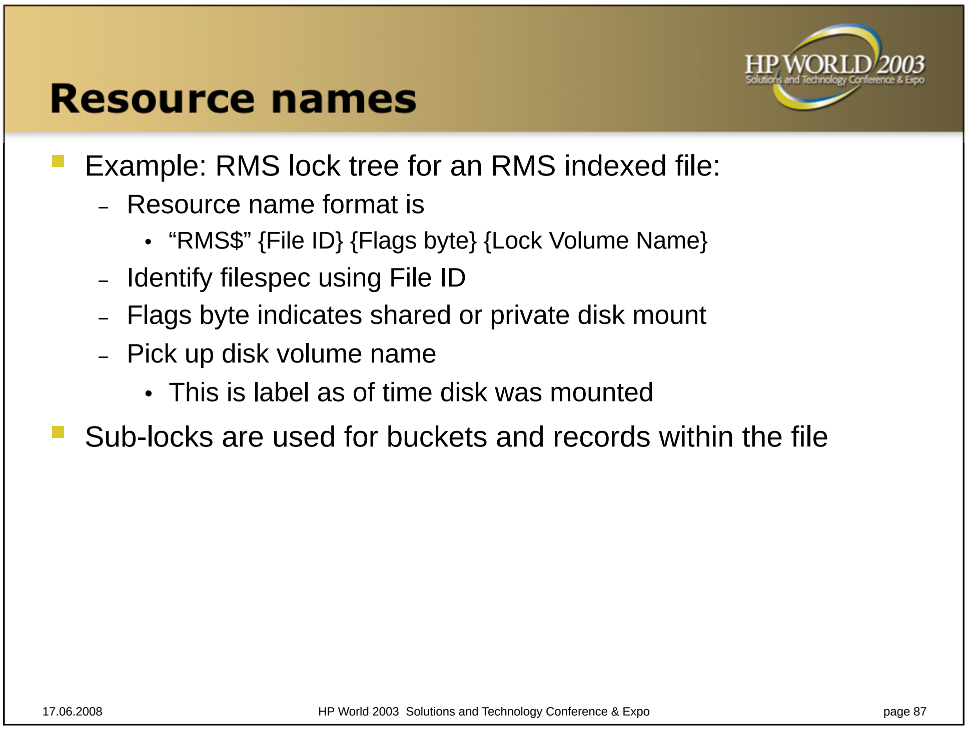
Symbolic lock resource names

- Symbolic resource names
  - Common prefixes:
    - SYS$ for OpenVMS executive
    - F11B$ for XQP, file system
    - RMS$ for Record Management Services
  - See Appendix H in Alpha V1.5 Internals and Data Structures Manual or Appendix A in Alpha V7.0 version



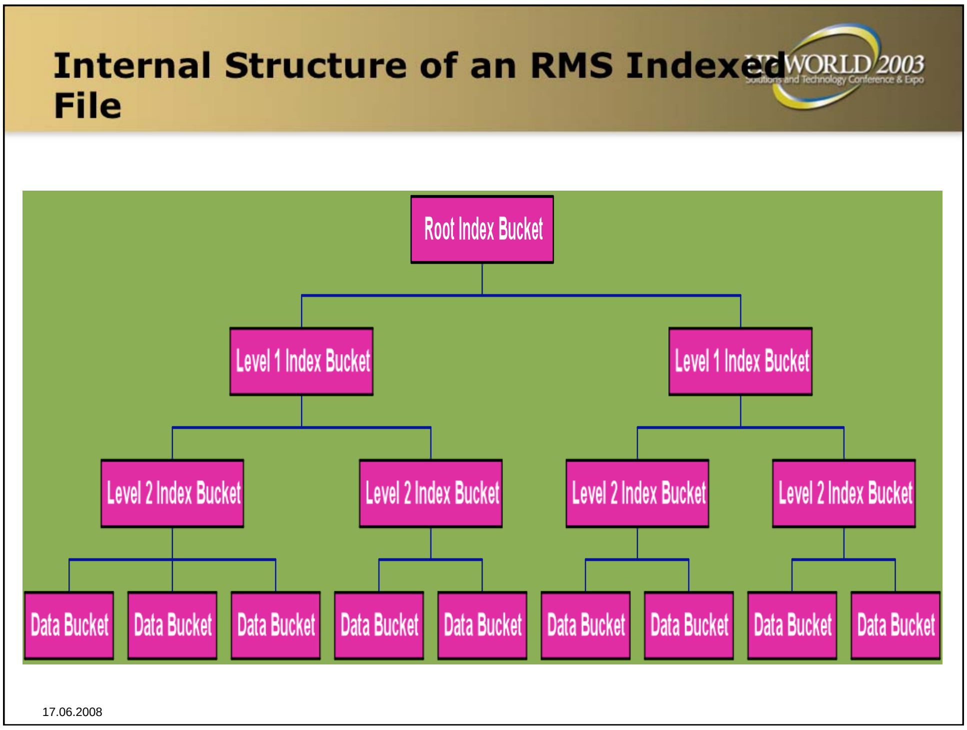
Resource names

- Example: Device Lock
  - Resource name format is
    - “SYS$” {Device Name in ASCII text}



Resource names

- Example: RMS lock tree for an RMS indexed file:
  - Resource name format is
    - “RMS$” {File ID} {Flags byte} {Lock Volume Name}
  - Identify filespec using File ID
  - Flags byte indicates shared or private disk mount
  - Pick up disk volume name
    - This is label as of time disk was mounted
- Sub-locks are used for buckets and records within the file



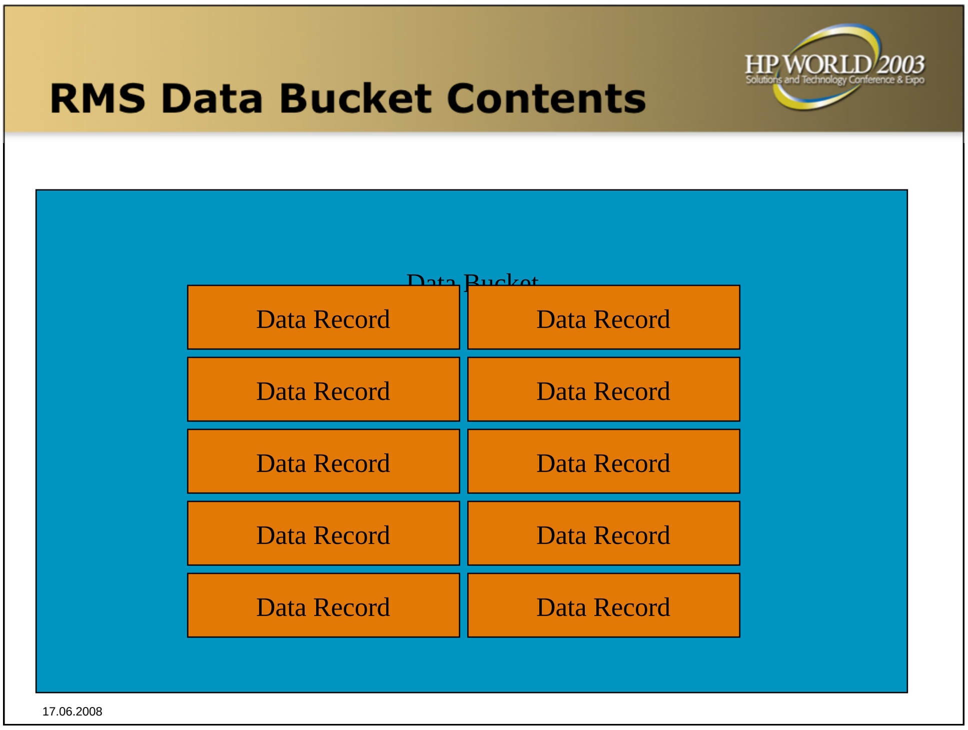
Internal Structure of an RMS Indexed File

Root Index Bucket

Level 1 Index Bucket
Level 1 Index Bucket

Level 2 Index Bucket
Level 2 Index Bucket
Level 2 Index Bucket
Level 2 Index Bucket

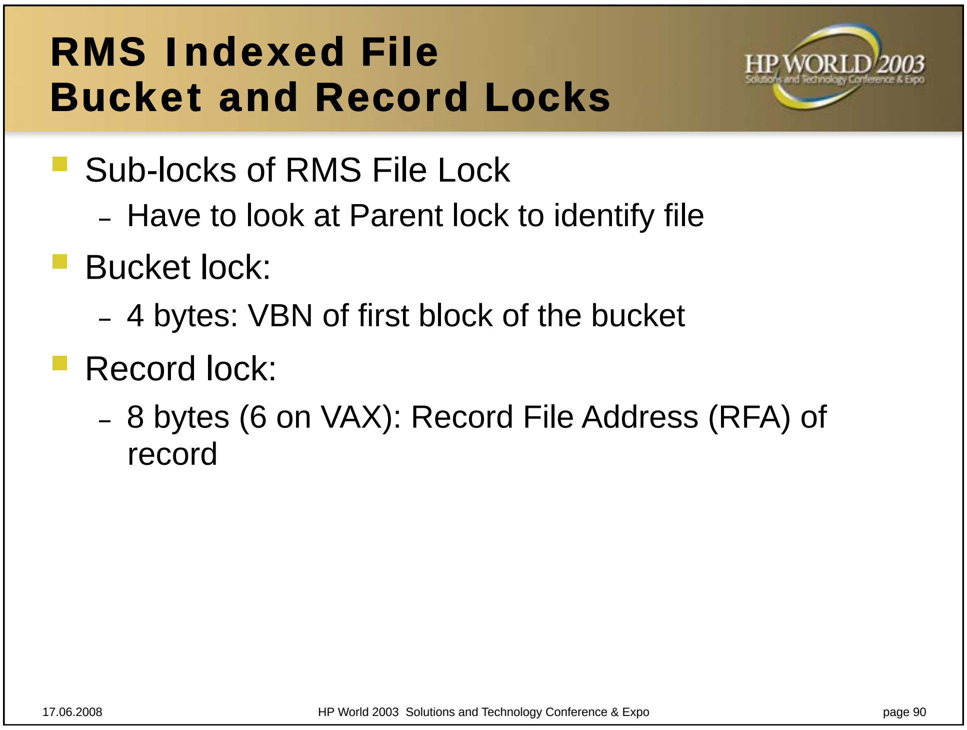
Data Bucket
Data Bucket
Data Bucket
Data Bucket
Data Bucket
Data Bucket
Data Bucket
Data Bucket
Data Bucket



RMS Data Bucket Contents

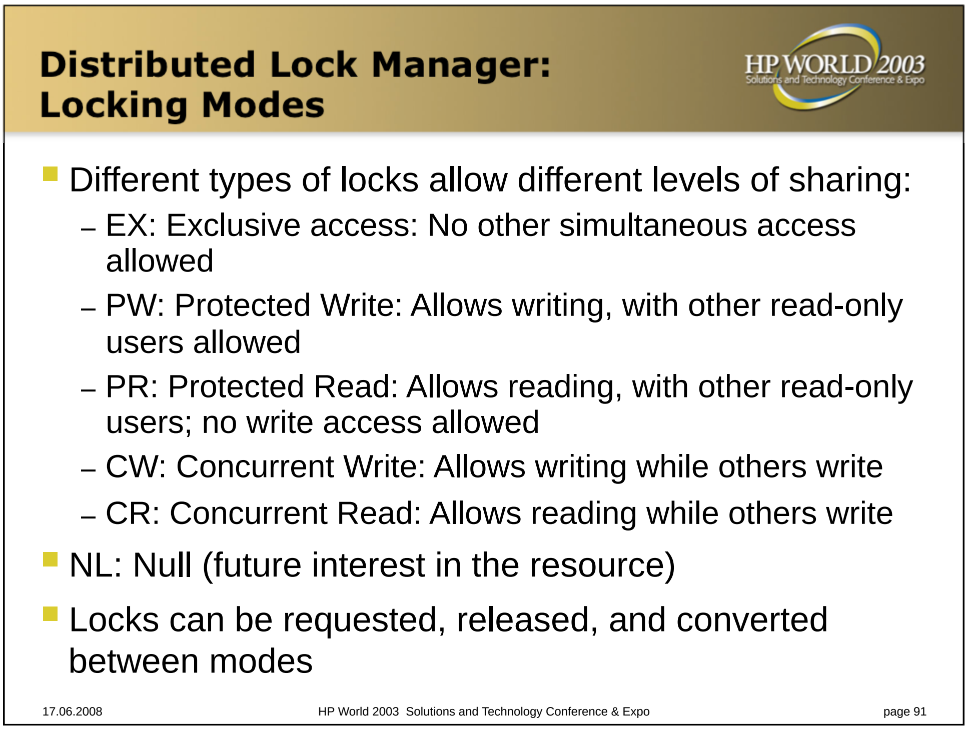
Data Bucket

Data Record	Data Record
Data Record	Data Record
Data Record	Data Record
Data Record	Data Record
Data Record	Data Record



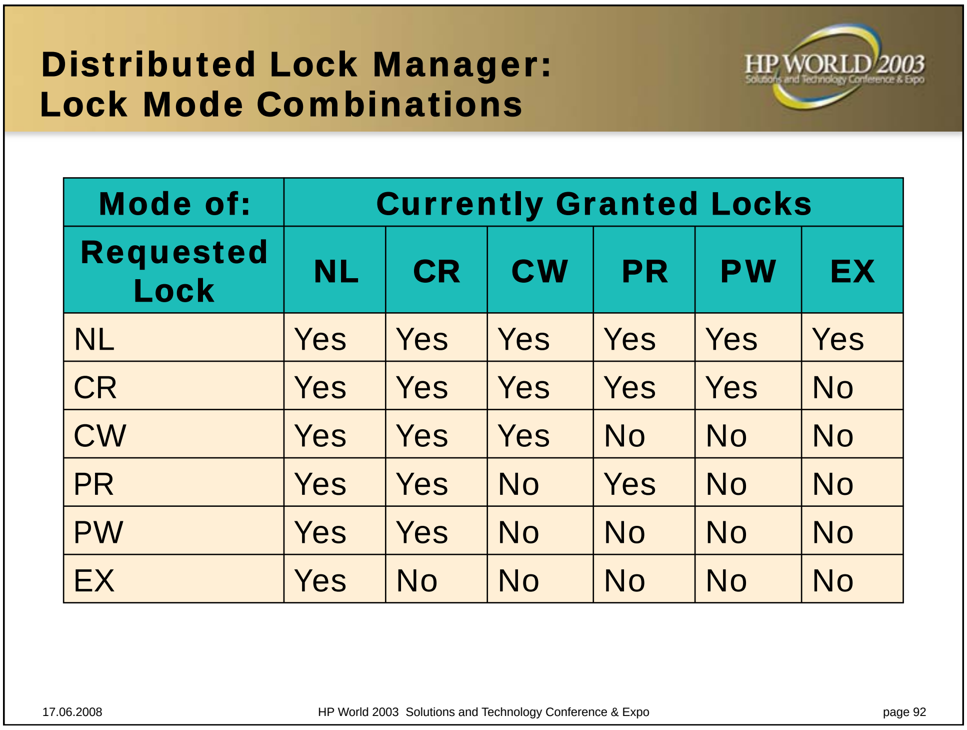
RMS Indexed File Bucket and Record Locks

- Sub-locks of RMS File Lock
  - Have to look at Parent lock to identify file
- Bucket lock:
  - 4 bytes: VBN of first block of the bucket
- Record lock:
  - 8 bytes (6 on VAX): Record File Address (RFA) of record



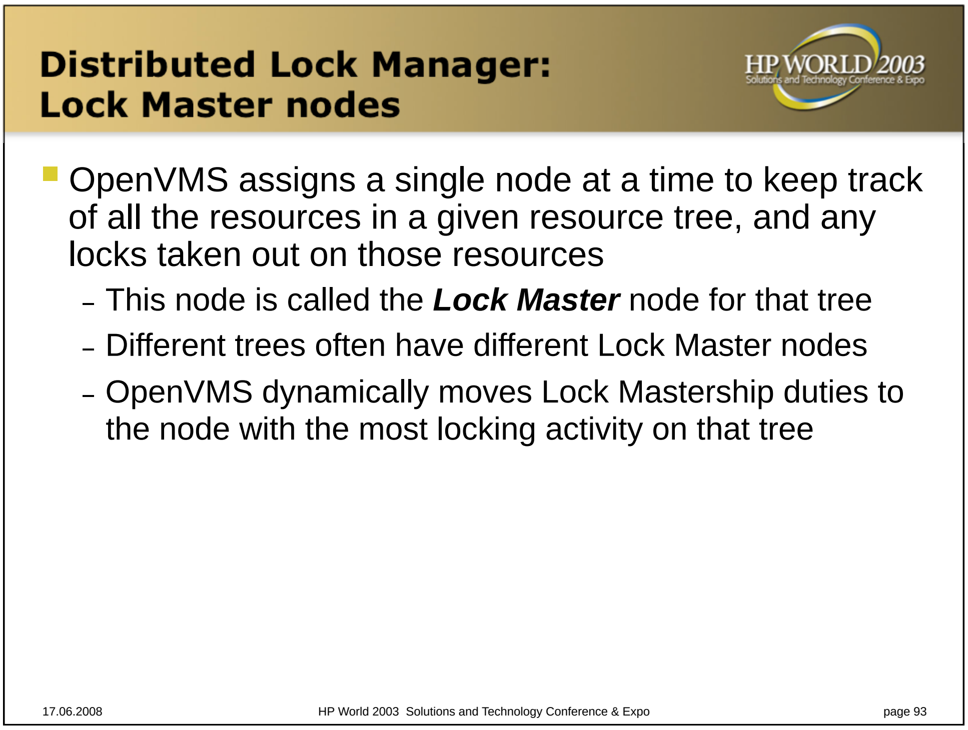
Distributed Lock Manager: Locking Modes

- Different types of locks allow different levels of sharing:
  - EX: Exclusive access: No other simultaneous access allowed
  - PW: Protected Write: Allows writing, with other read-only users allowed
  - PR: Protected Read: Allows reading, with other read-only users; no write access allowed
  - CW: Concurrent Write: Allows writing while others write
  - CR: Concurrent Read: Allows reading while others write
- NL: Null (future interest in the resource)
- Locks can be requested, released, and converted between modes



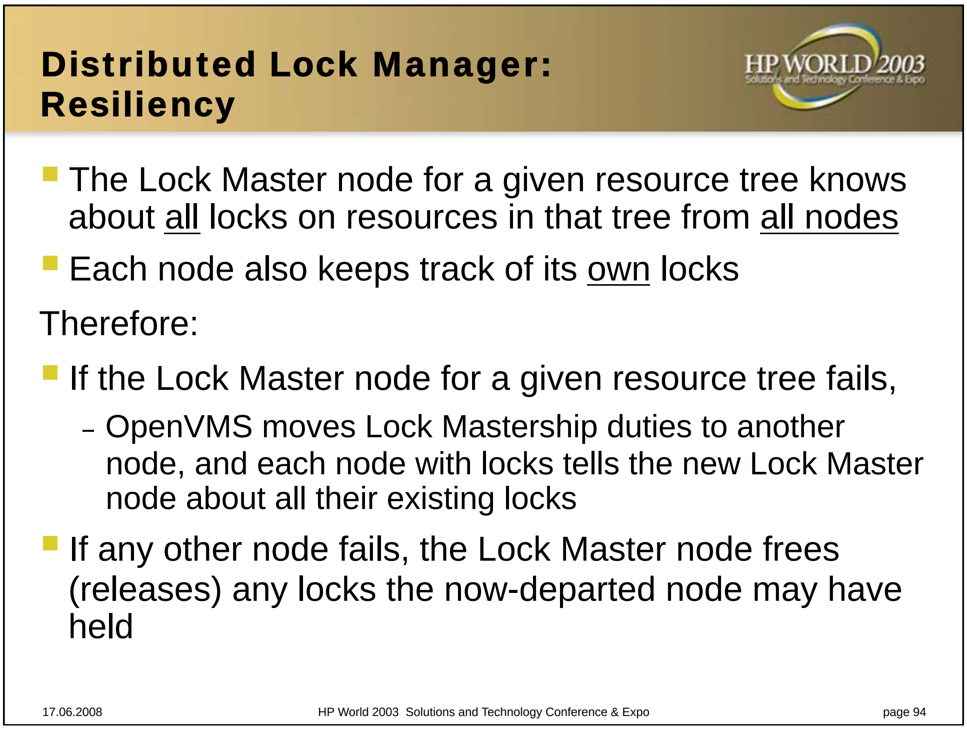
Distributed Lock Manager: Lock Mode Combinations

Mode of:	Currently Granted Locks					
Requested Lock	NL	CR	CW	PR	PW	EX
NL	Yes	Yes	Yes	Yes	Yes	Yes
CR	Yes	Yes	Yes	Yes	Yes	No
CW	Yes	Yes	Yes	No	No	No
PR	Yes	Yes	No	Yes	No	No
PW	Yes	Yes	No	No	No	No
EX	Yes	No	No	No	No	No



Distributed Lock Manager: Lock Master nodes

- OpenVMS assigns a single node at a time to keep track of all the resources in a given resource tree, and any locks taken out on those resources
  - This node is called the ***Lock Master*** node for that tree
  - Different trees often have different Lock Master nodes
  - OpenVMS dynamically moves Lock Mastership duties to the node with the most locking activity on that tree

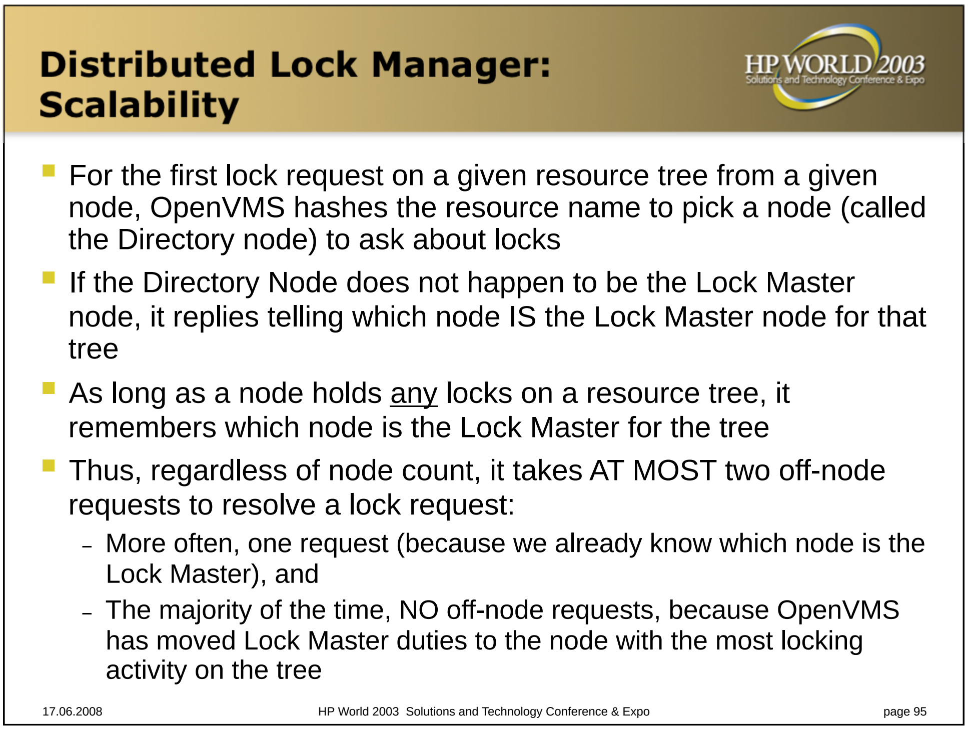


Distributed Lock Manager: Resiliency

- The Lock Master node for a given resource tree knows about all locks on resources in that tree from all nodes
- Each node also keeps track of its own locks

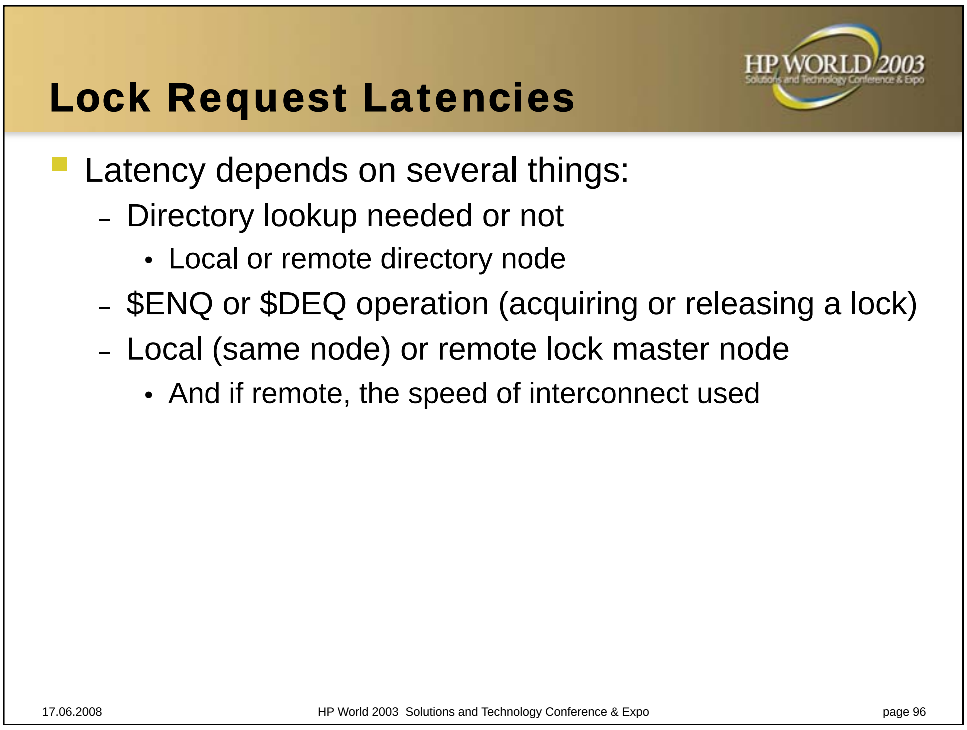
Therefore:

- If the Lock Master node for a given resource tree fails,
  - OpenVMS moves Lock Mastership duties to another node, and each node with locks tells the new Lock Master node about all their existing locks
- If any other node fails, the Lock Master node frees (releases) any locks the now-departed node may have held



Distributed Lock Manager: Scalability

- For the first lock request on a given resource tree from a given node, OpenVMS hashes the resource name to pick a node (called the Directory node) to ask about locks
- If the Directory Node does not happen to be the Lock Master node, it replies telling which node IS the Lock Master node for that tree
- As long as a node holds any locks on a resource tree, it remembers which node is the Lock Master for the tree
- Thus, regardless of node count, it takes AT MOST two off-node requests to resolve a lock request:
  - More often, one request (because we already know which node is the Lock Master), and
  - The majority of the time, NO off-node requests, because OpenVMS has moved Lock Master duties to the node with the most locking activity on the tree



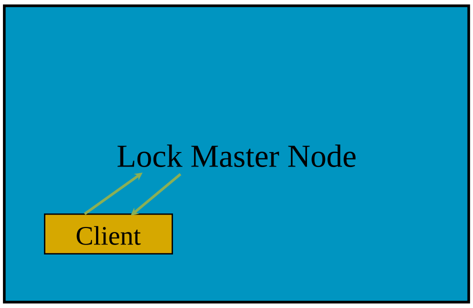
Lock Request Latencies

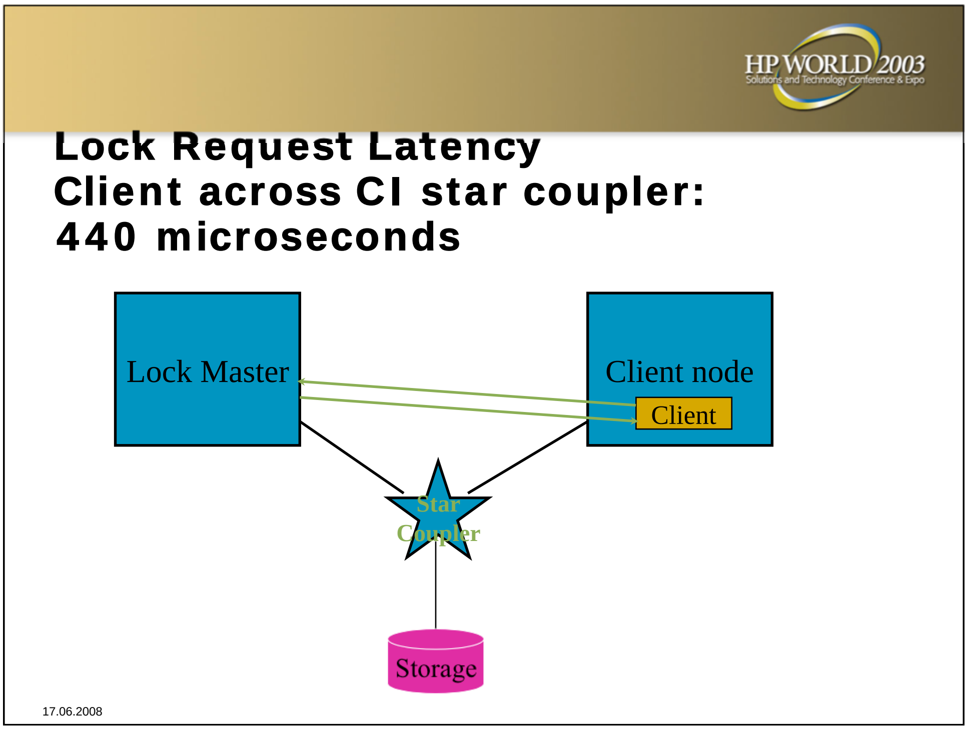
- Latency depends on several things:
  - Directory lookup needed or not
    - Local or remote directory node
  - $ENQ or $DEQ operation (acquiring or releasing a lock)
  - Local (same node) or remote lock master node
    - And if remote, the speed of interconnect used

Lock Request Latencies

- Local requests are fastest
- Remote requests are significantly slower:
  - Code path ~20 times longer
  - Interconnect also contributes latency
  - Total latency up to 2 orders of magnitude higher than local requests

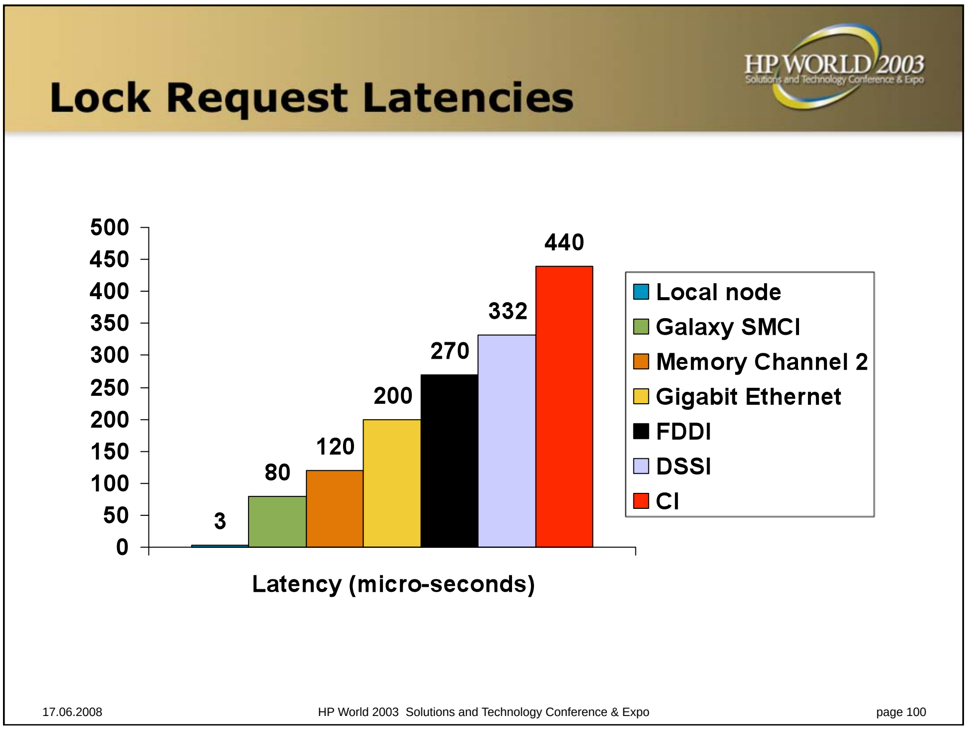
Lock Request Latency
Client process on same node:
2-6 microseconds





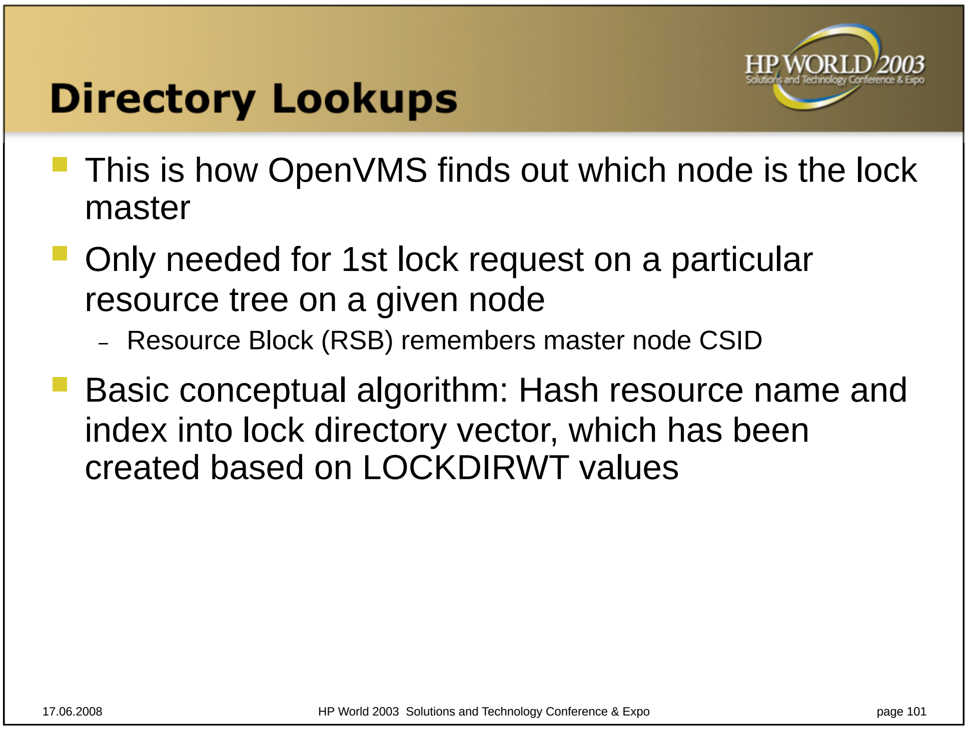
Lock Request Latency Client across CI star coupler: 440 microseconds

17.06.2008



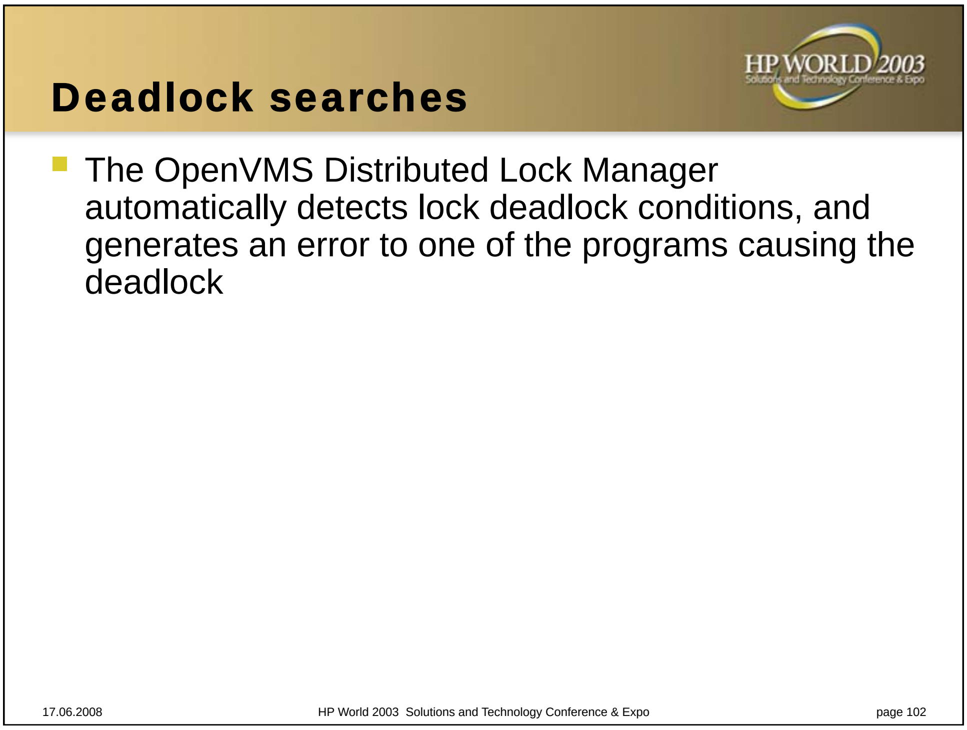
Lock Request Latencies

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 100



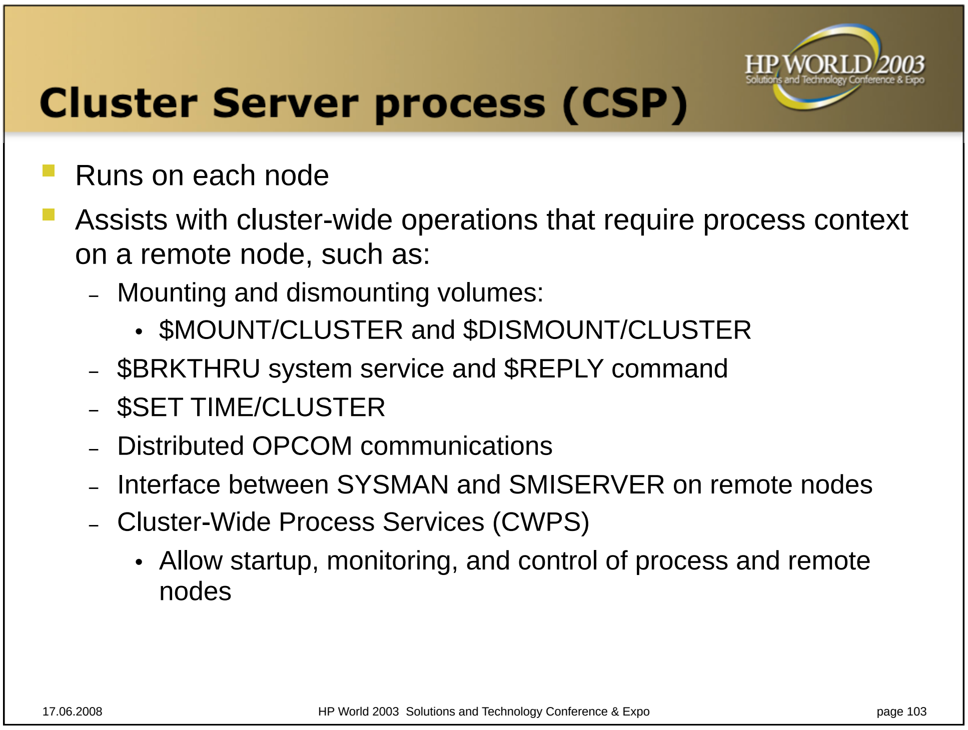
Directory Lookups

- This is how OpenVMS finds out which node is the lock master
- Only needed for 1st lock request on a particular resource tree on a given node
  - Resource Block (RSB) remembers master node CSID
- Basic conceptual algorithm: Hash resource name and index into lock directory vector, which has been created based on LOCKDIRWT values



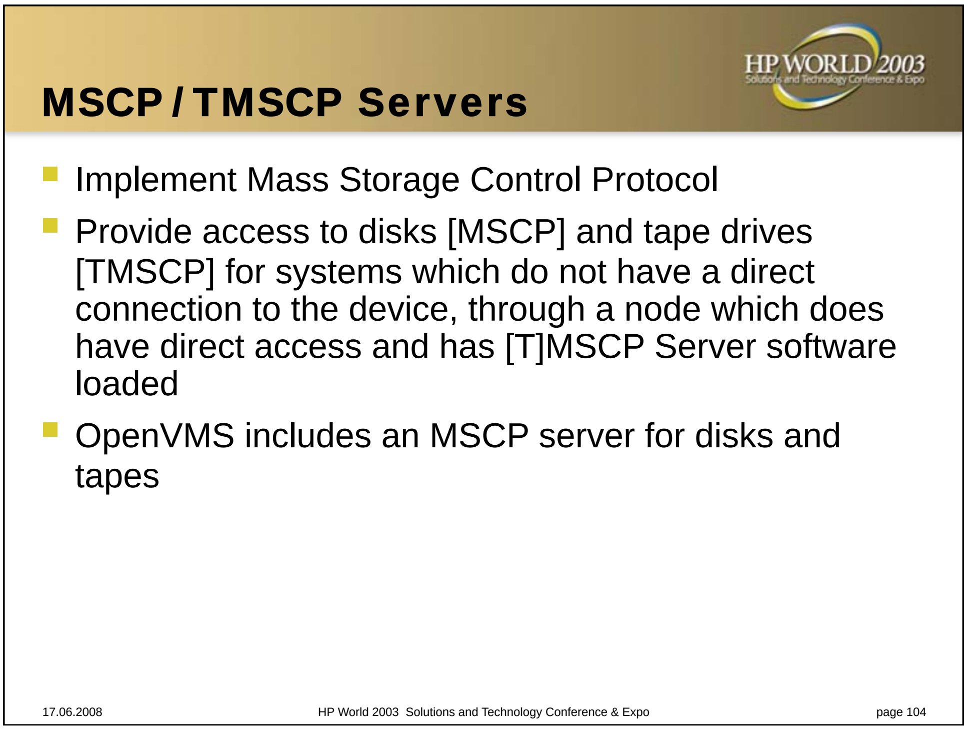
Deadlock searches

- The OpenVMS Distributed Lock Manager automatically detects lock deadlock conditions, and generates an error to one of the programs causing the deadlock



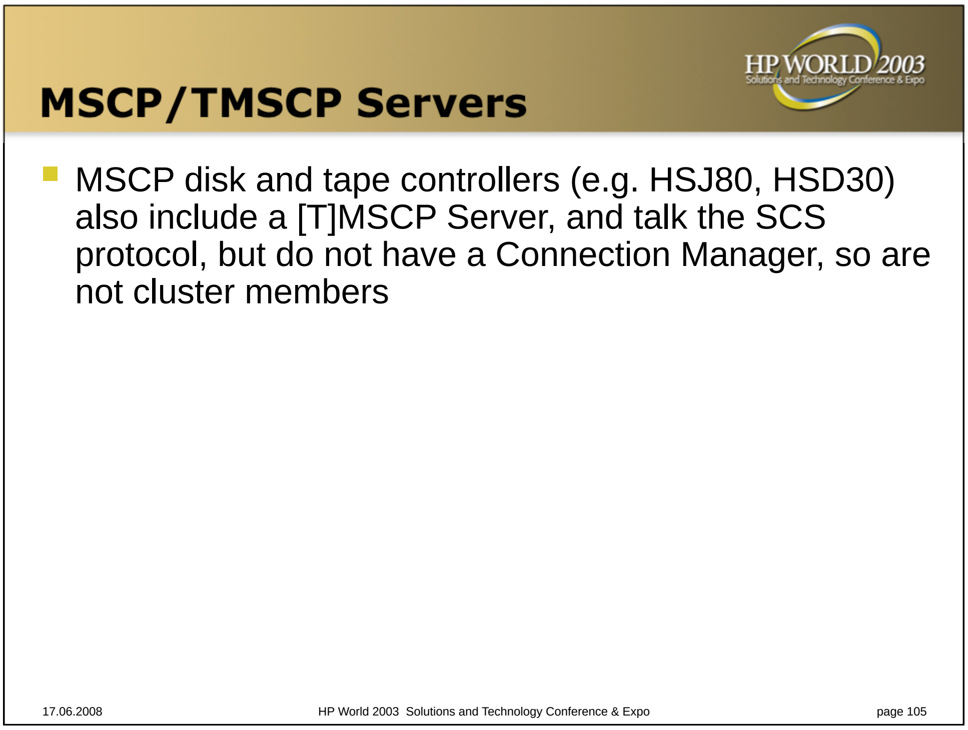
Cluster Server process (CSP)

- Runs on each node
- Assists with cluster-wide operations that require process context on a remote node, such as:
  - Mounting and dismounting volumes:
    - $MOUNT/CLUSTER and $DISMOUNT/CLUSTER
  - $BRKTHRU system service and $REPLY command
  - $SET TIME/CLUSTER
  - Distributed OPCOM communications
  - Interface between SYSMAN and SMISERVER on remote nodes
  - Cluster-Wide Process Services (CWPS)
    - Allow startup, monitoring, and control of process and remote nodes



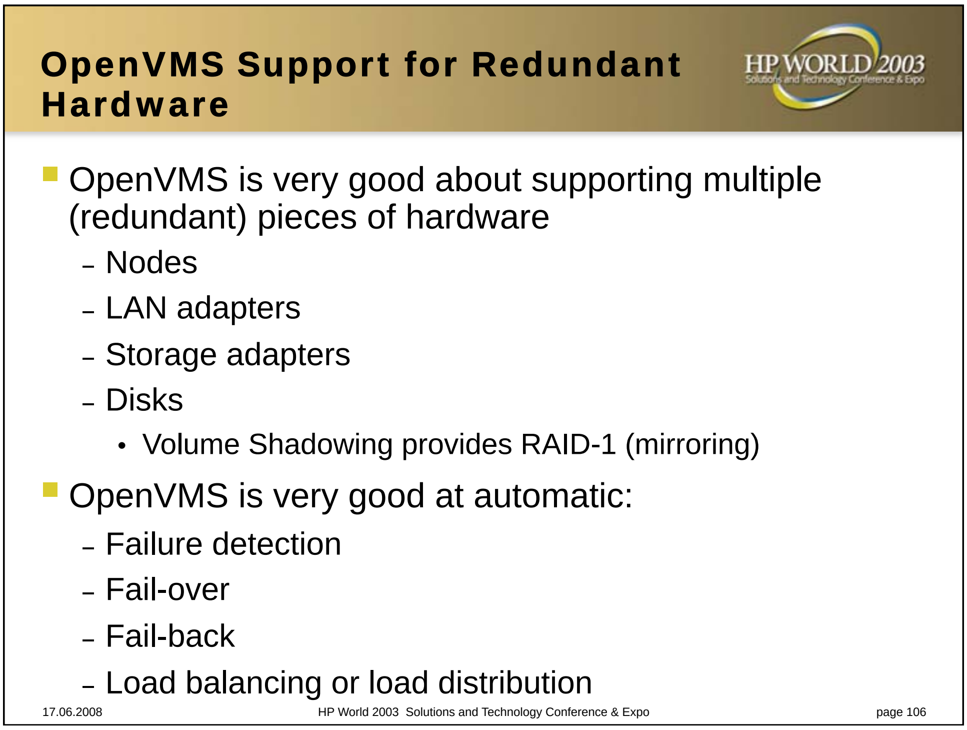
MSCP / TMSCP Servers

- Implement Mass Storage Control Protocol
- Provide access to disks [MSCP] and tape drives [TMSCP] for systems which do not have a direct connection to the device, through a node which does have direct access and has [T]MSCP Server software loaded
- OpenVMS includes an MSCP server for disks and tapes



MSCP/TMSCP Servers

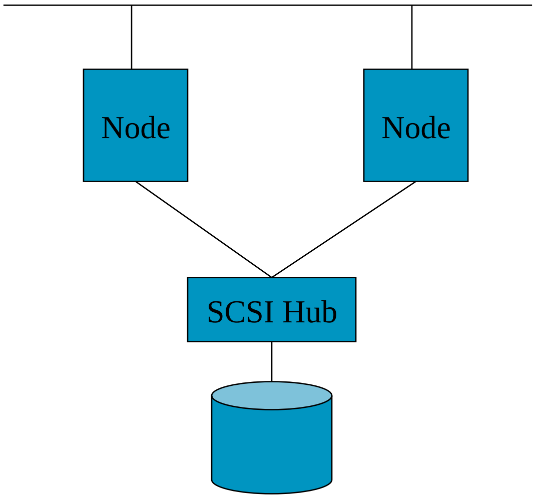
- MSCP disk and tape controllers (e.g. HSJ80, HSD30) also include a [T]MSCP Server, and talk the SCS protocol, but do not have a Connection Manager, so are not cluster members



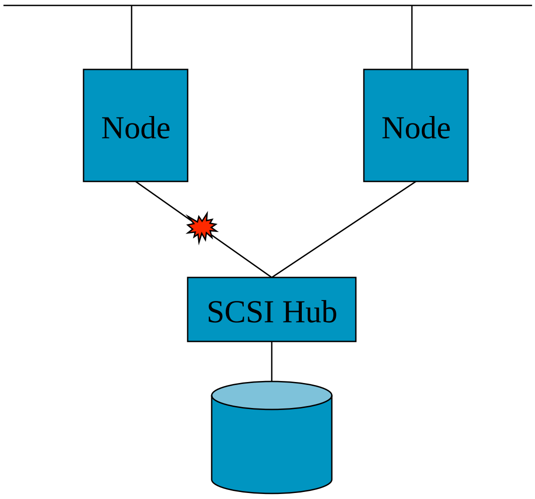
OpenVMS Support for Redundant Hardware

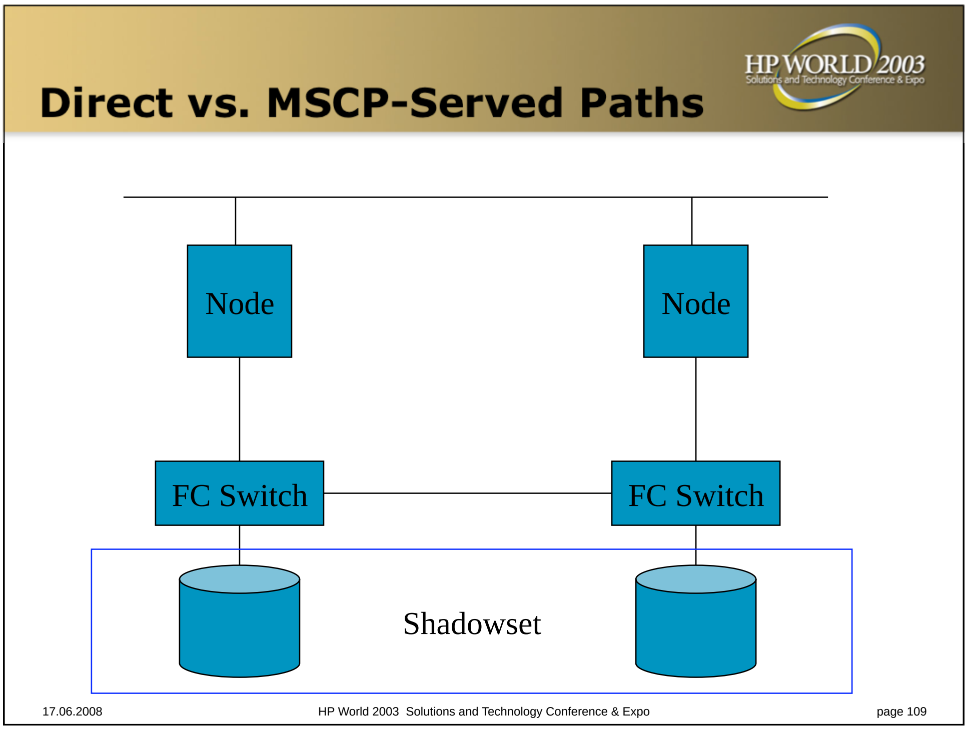
- OpenVMS is very good about supporting multiple (redundant) pieces of hardware
  - Nodes
  - LAN adapters
  - Storage adapters
  - Disks
    - Volume Shadowing provides RAID-1 (mirroring)
- OpenVMS is very good at automatic:
  - Failure detection
  - Fail-over
  - Fail-back
  - Load balancing or load distribution

Direct vs. MSCP-Served Paths



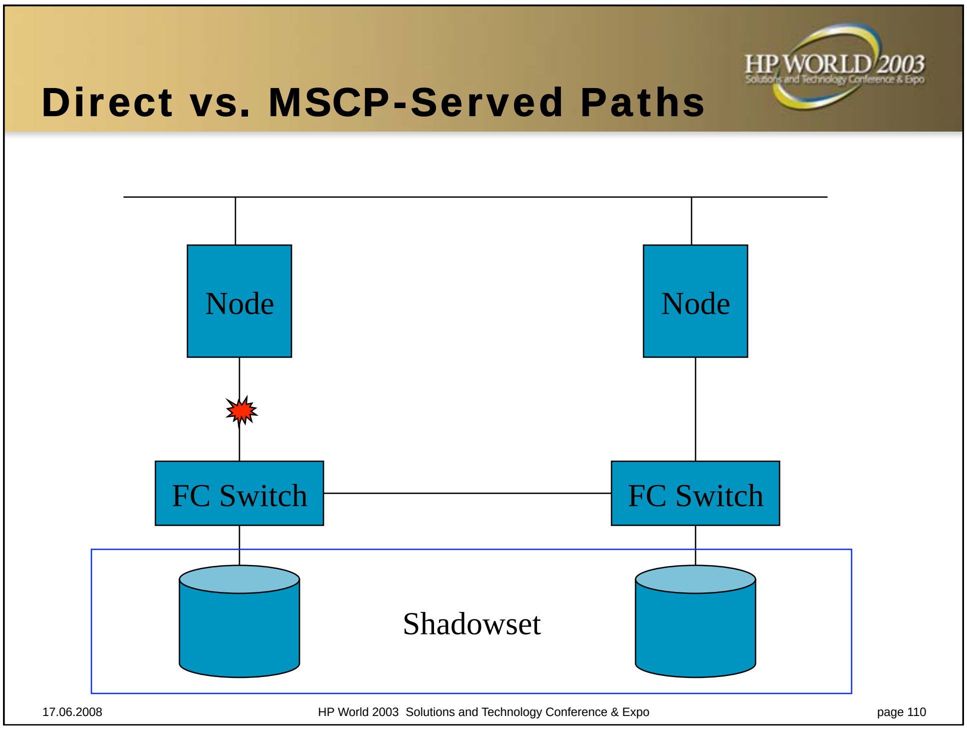
Direct vs. MSCP-Served Paths





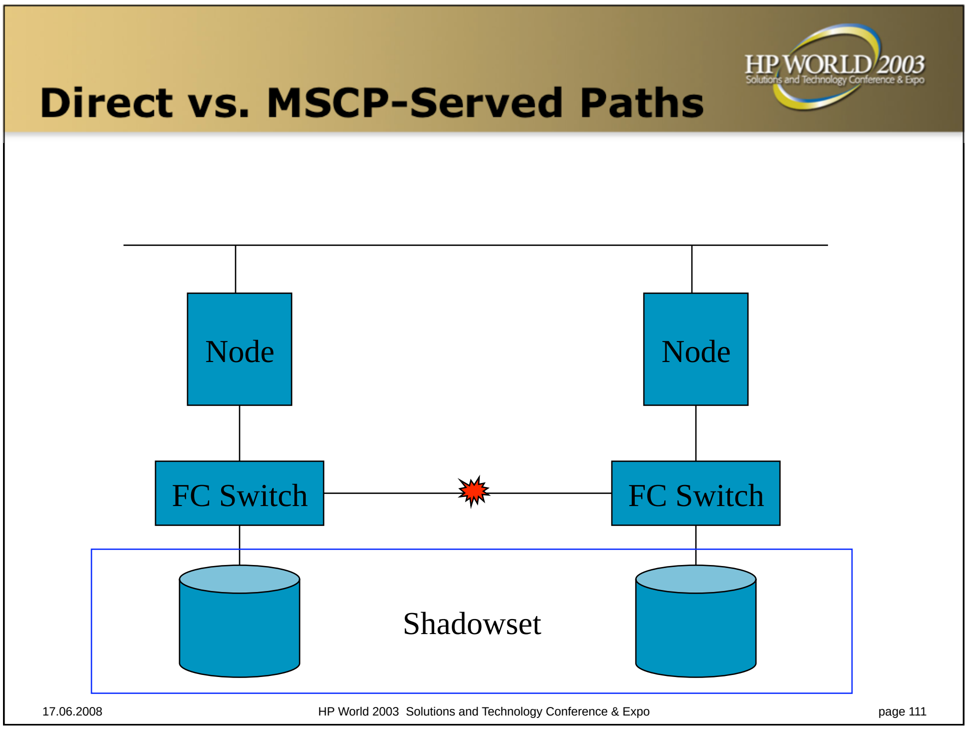
Direct vs. MSCP-Served Paths

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 109

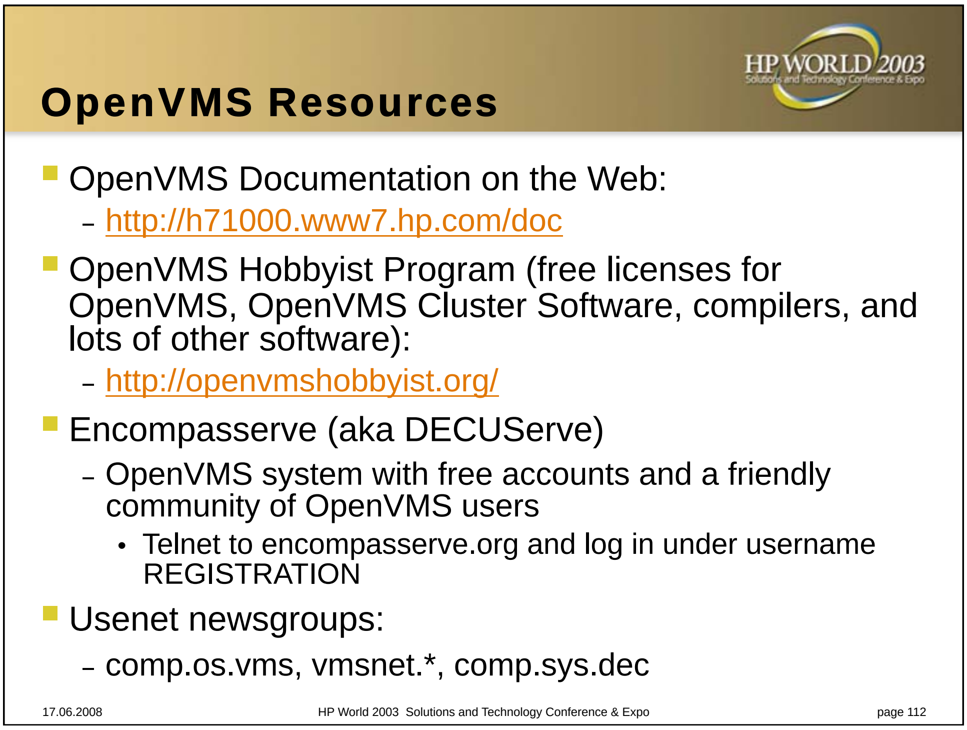


Direct vs. MSCP-Served Paths

17.06.2008 HP World 2003 Solutions and Technology Conference & Expo page 110

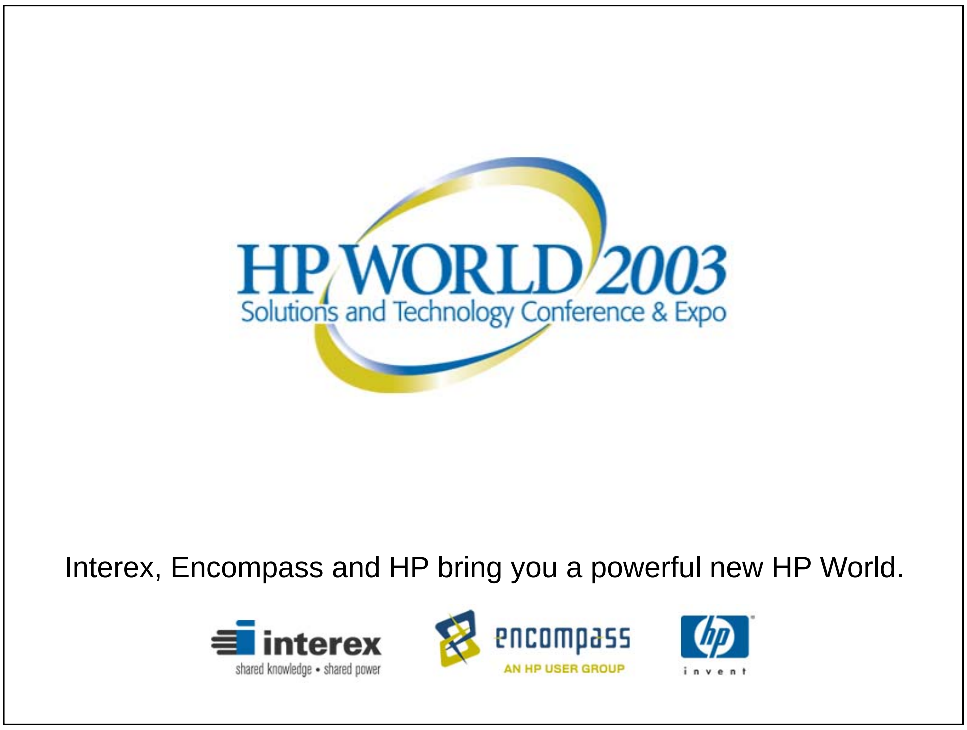


Direct vs. MSCP-Served Paths



OpenVMS Resources

- OpenVMS Documentation on the Web:
  - http://h71000.www7.hp.com/doc
- OpenVMS Hobbyist Program (free licenses for OpenVMS, OpenVMS Cluster Software, compilers, and lots of other software):
  - http://openvmshobbyist.org/
- Encompasserve (aka DECUServe)
  - OpenVMS system with free accounts and a friendly community of OpenVMS users
    - Telnet to encompasserve.org and log in under username REGISTRATION
- Usenet newsgroups:
  - comp.os.vms, vmsnet.*, comp.sys.dec


HP WORLD 2003
Solutions and Technology Conference & Expo
interex
shared knowledge • shared power
encompass
AN HP USER GROUP
hp
i n v e n t

Interex, Encompass and HP bring you a powerful new HP World.