OpenVMS Cluster Overview

Andreas Polze

Cluster Configurations

- An OpenVMS Cluster systems is a group of OpenVMS systems, storage subsystems, interconnects, and software that work together as one virtual system.
  - An OpenVMS Cluster system can be homogeneous, that is, all systems are the same architecture (Alpha, VAX, or the HP Integrity servers)
  - An OpenVMS Cluster can be heterogeneous, that is, a combination of two architectures with all systems running OpenVMS.
  - Alpha and VAX or Alpha and HP Integrity
- In an OpenVMS Cluster system, each system:
- Shares processing resources, queues, and data storage
- Can boot or fail independently
- Runs the OpenVMS operating system
- In addition, an OpenVMS Cluster system is managed as a single entity.

Cluster Benefits

- Resource sharing
  - Multiple systems can access the same storage devices, so that users can share files clusterwide.
  - You can also distribute applications, batch, and print-job processing across multiple systems.
  - Jobs that access shared resources can execute on any system.
- Availability
  - Data and applications remain available during scheduled or unscheduled downtime of individual systems.
  - A variety of configurations provide many levels of availability up to and including disaster-tolerant operation.
- Flexibility
  - OpenVMS Cluster computing environments offer compatible hardware and software across a wide price and performance range.

Cluster Benefits (contd.)

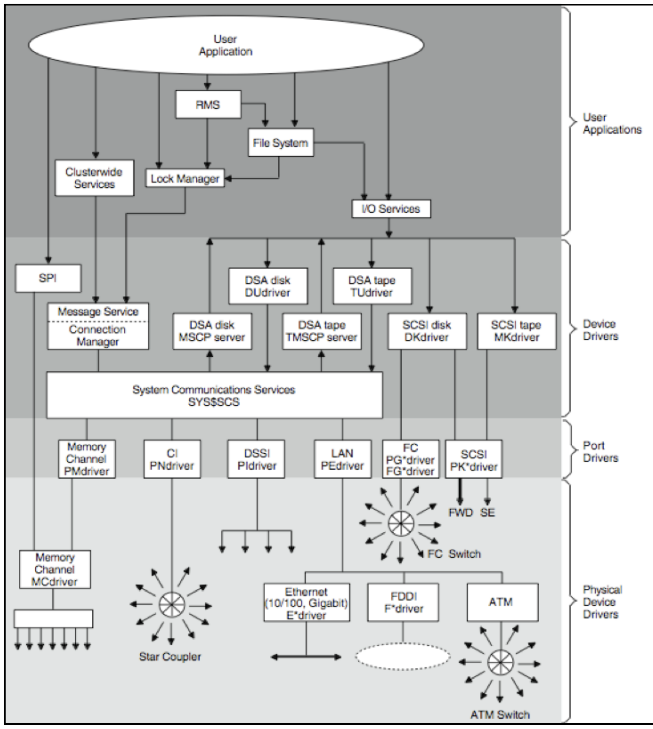
- Scalability
  - You can add processing and storage resources without disturbing the rest of the system.
  - The full range of systems, from high-end multiprocessor systems to smaller workstations, can be interconnected and easily reconfigured to meet growing needs.
- Ease of management
  - OpenVMS Cluster management is efficient and secure. Because you manage an OpenVMS Cluster as a single system, many tasks need to be performed only once.
  - OpenVMS Clusters automatically balance user, batch, and print work loads.
- Open systems
  - Adherence to IEEE, POSIX, OSF/1, Motif, OSF DCE, ANSI SQL, and TCP/IP standards provides OpenVMS Cluster portability and interoperability.

Operating System Components

- Record Management Services (RMS) and OpenVMS file system
  - Provide shared read and write access to files on disks and tapes in an OpenVMS Cluster environment.
- Clusterwide process services
  - Enables clusterwide operation of OpenVMS commands, such as SHOW SYSTEM and SHOW USERS, as well as the ability to create and delete processes clusterwide.
- Distributed Lock Manager
  - Synchronizes access by many users to shared resources.
- Distributed Job Controller
  - Enables clusterwide sharing of batch and print queues, which optimizes the use of these resources.
- Connection Manager
  - Controls the membership and quorum of the OpenVMS Cluster members.

Operating System Components (contd.)

- SCS (System Communications Services)
  - Implements OpenVMS Cluster communications between nodes using the OpenVMS System Communications Architecture (SCA).
- MSCP server
  - Makes locally connected disks to which it has direct access available to other systems in the OpenVMS Cluster.
- TMSCP server
  - Makes locally connected tapes to which it has direct access available to other systems in the OpenVMS Cluster.



Hardware and Operating System Components

OpenVMS Cluster Networking Components

- DECnet-Plus
  - A network transport is necessary for internode communication.
- Distributed File Service (DFS)
  - Software to let you communicate and share resources among systems over extended distances.
- LAT software
  - Used with terminal server hardware to support Ethernet-based character cell terminals. During a system failure, LAT software automatically makes a connection to one of the remaining systems.
- Advanced Server for OpenVMS and PATHWORKS for OpenVMS
  - Client and server networking software that links PC systems into OpenVMS Cluster systems.
- TCP/IP Services for OpenVMS software
  - Provides Network File System (NFS) server capabilities for OpenVMS and supports Internet networking protocols.

Storage Enhancement Software

- Optional storage enhancement software improves the performance or availability of storage subsystems.
- Examples include:
  - Volume Shadowing for OpenVMS (redundant arrays of independent disks [RAID] level 1)
  - DECram for OpenVMS (random access memory [RAM] disk)
  - RAID Software for OpenVMS (supports RAID level 0 arrays (disk striping) and RAID level 5 arrays (disk striping with parity)
  - Hierarchical Storage Manager (HSM)
- http://www.hp.com/go/openvms

Cluster Configuration

- An OpenVMS Cluster system consisting of Alpha and VAX systems cannot contain more than 96 (combined total) systems.
  - An OpenVMS Cluster system consisting of Alpha and I64 systems cannot contain more than 16 (combined total) systems with a maximum of 8 I64 systems for OpenVMS Version 8.2.
  - Larger clusters will be supported in subsequent releases.
- An Alpha and a VAX system can not boot from the same system disk.
  - System disks are architecture specific and can be shared only by systems of the same architecture.
  - Cross-architecture satellite booting is supported. Alpha satellites (clients) can boot from a VAX boot server, and VAX satellites (clients) can boot from an Alpha boot server.
- Every OpenVMS node must be able to communicate directly with every other OpenVMS Cluster node.

Configuring OpenVMS Clusters for Availability

Availability Requirements:

- Conventional
  - For business functions that can wait with little or no effect while a system or application is unavailable.
- 24 x 365
  - For business functions that require uninterrupted computing services, either during essential time periods or during most hours of the day throughout the year. Minimal down time is acceptable.
- Disaster tolerant
  - For business functions with stringent availability requirements. These businesses need to be immune to disasters like earthquakes, floods, and power failures.

How OpenVMS Clusters Provide Availability

- Multiple systems to share access to resources
  - Direct access
    - Connect disk and tape storage subsystems to CI and DSSI interconnects rather than to a node. The shutdown or failure of a system has no effect on the ability of other systems to access storage.
  - Served access
    - Storage devices attached to a node can be served to other nodes in the OpenVMS Cluster (MSCP, TPMSCP). However, the shutdown or failure of the serving node affects the ability of other nodes to access storage.
- Redundancy of major hardware components
  - Systems, Interconnects, Adapters
  - Storage devices and data
- Software support for failover between hardware components
- Software products to support high availability

Mechanism	What Happens if a Failure Occurs	Type of Recovery
DECnet–Plus cluster alias	If a node fails, OpenVMS Cluster software automatically distributes new incoming connections among other participating nodes. Manual.	Users who were logged in to the failed node can reconnect to a remaining node. Automatic for appropriately coded applications. Such applications can reinstate a connection to the cluster alias node name, and the connection is directed to one of the remaining nodes.
I/O paths	With redundant paths to storage devices, if one path fails, OpenVMS Cluster software fails over to a working path, if one exists.	Transparent, provided another working path is available.
Interconnect	With redundant or mixed interconnects, OpenVMS Cluster software uses the fastest working path to connect to other OpenVMS Cluster members. If an interconnect path fails, OpenVMS Cluster software fails over to a working path, if one exists.	Transparent.
Boot and disk servers	If you configure at least two nodes as boot and disk servers, satellites can continue to boot and use disks if one of the servers shuts down or fails. Failure of a boot server does not affect nodes that have already booted, providing they have an alternate path to access MSCP served disks.	Automatic

Mechanism	What Happens if a Failure Occurs	Type of Recovery
Terminal servers and LAT software	Attach terminals and printers to terminal servers. If a node fails, the LAT software automatically connects to one of the remaining nodes. In addition, if a user process is disconnected from a LAT terminal session, when the user attempts to reconnect to a LAT session, LAT software can automatically reconnect the user to the disconnected session.	Manual. Terminal users who were logged in to the failed node must log in to a remaining node and restart the application.
Generic batch and print queues	You can set up generic queues to feed jobs to execution queues (where processing occurs) on more than one node. If one node fails, the generic queue can continue to submit jobs to execution queues on remaining nodes. In addition, batch jobs submitted using the /RESTART qualifier are automatically restarted on one of the remaining nodes.	Transparent for jobs waiting to be dispatched. Automatic or manual for jobs executing on the failed node.
Autostart batch and print queues	For maximum availability, you can set up execution queues as autostart queues with a failover list. When a node fails, an autostart execution queue and its jobs automatically fail over to the next logical node in the failover list and continue processing on another node. Autostart queues are especially useful for print queues directed to printers that are attached to terminal servers..	Transparent

Availability Strategies

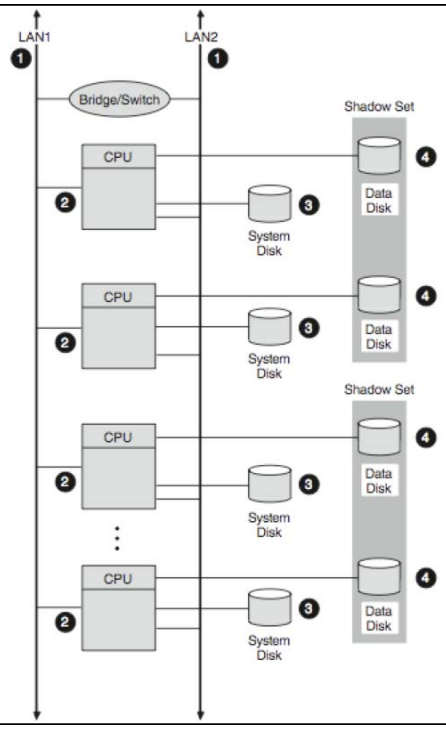
- Eliminate single points of failure
  - Make components redundant so that if one component fails, the other is available to take over.
- Shadow system disks
  - The system disk is vital for node operation. Use Volume Shadowing for OpenVMS to make system disks redundant.
- Shadow essential data disks
  - Use Volume Shadowing for OpenVMS to improve data availability by making data disks redundant.
- Provide shared, direct access to storage
  - Where possible, give all nodes shared direct access to storage. This reduces dependency on MSCP server nodes for access to storage.

Availability Strategies (contd.)

- Minimize environmental risks
  - Provide a generator or uninterruptible power system (UPS) to replace utility power for use during temporary outages.
  - Configure extra air-conditioning equipment so that failure of a single unit does not prevent use of the system equipment.
- Configure at least three nodes OpenVMS
  - Cluster nodes require a quorum to continue operating. An optimal configuration uses a minimum of three nodes so that if one node becomes unavailable, the two remaining nodes maintain quorum and continue processing.
- Configure extra capacity
  - For each component, configure at least one unit more than is necessary to handle capacity.
  - For crucial components, keep resource use sufficiently less than 80% capacity so that if one component fails, the work load can be spread across remaining components without overloading them.

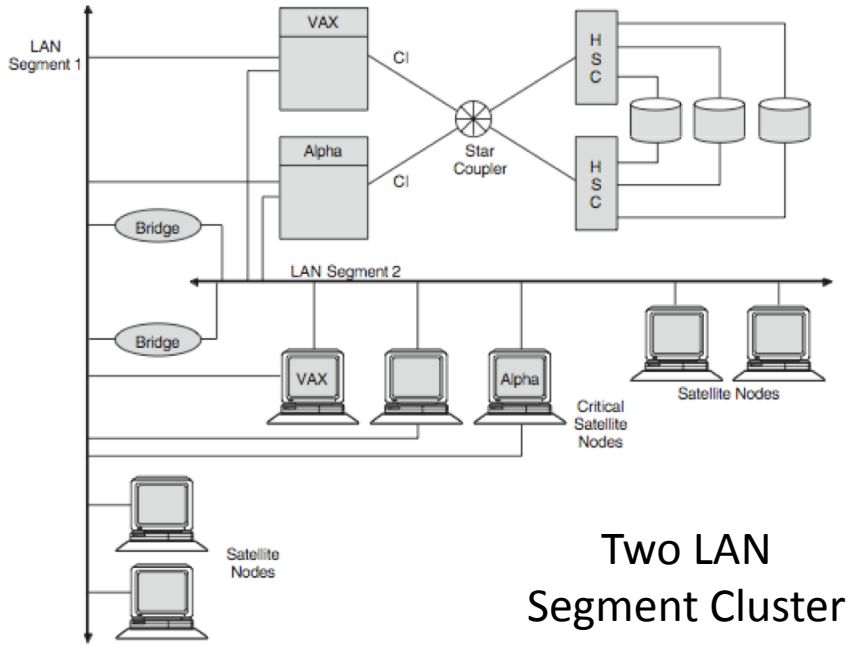
Availability Strategies (contd.)

- Keep a spare component on standby
  - For each component, keep one or two spares available and ready to use if a component fails.
  - Be sure to test spare components regularly to make sure they work.
  - More than one or two spare components increases complexity as well as the chance that the spare will not operate correctly when needed.
- Use homogeneous nodes
  - Configure nodes of similar size and performance to avoid capacity overloads in case of failover.
  - If a large node fails, a smaller node may not be able to handle the transferred work load.
  - The resulting bottleneck may decrease OpenVMS Cluster performance.
- Use reliable hardware
  - Consider the probability of a hardware device failing. Check product descriptions for MTBF (mean time between failures). In general, newer technologies are more reliable.



LAN OpenVMS Cluster System

- Advantages
  - Lowest cost of all the sample configurations shown.
  - Some potential for growth in size and performance.
  - The LAN interconnect supports the widest choice of nodes.
- Disadvantages
  - No shared direct access to storage. The nodes are dependent on an MSCP server for access to shared storage.
  - Shadowing disks across the LAN nodes causes shadow copies when the nodes boot.
  - Shadowing the system disks is not practical because of boot-order dependencies.

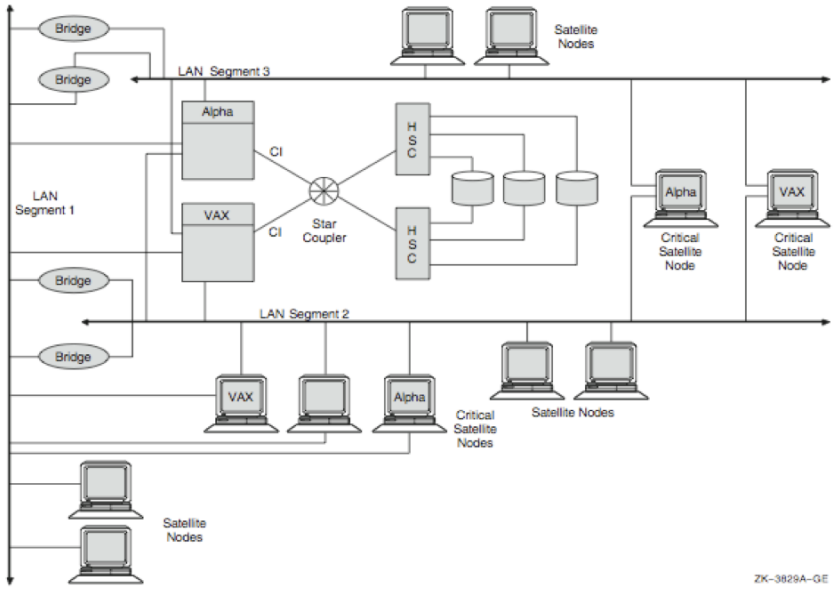


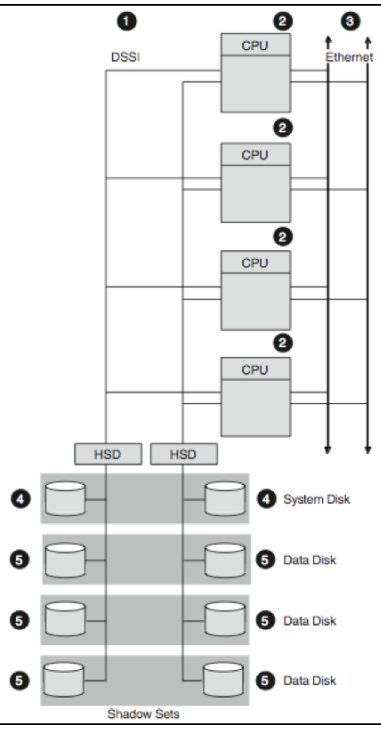
Two LAN Segment Cluster

Discussion

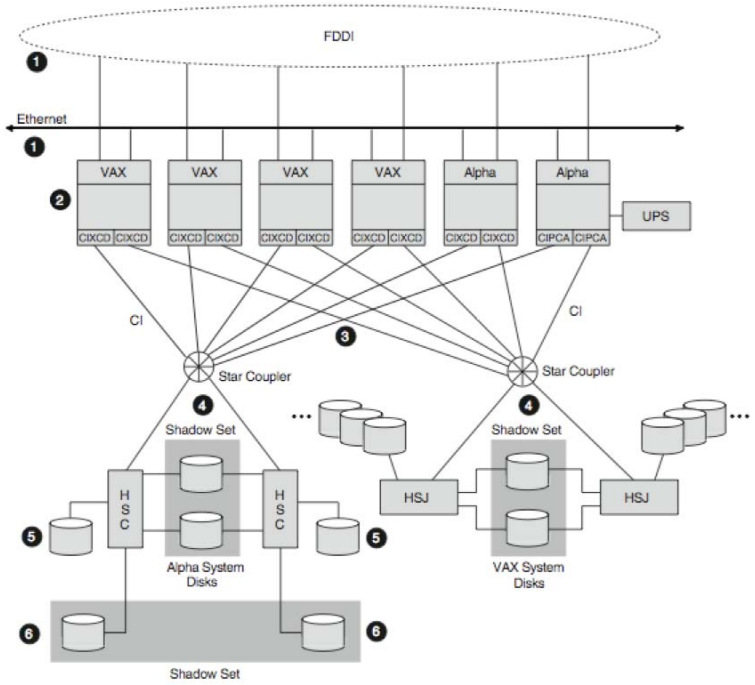
- Connecting critical nodes to multiple LAN segments provides increased availability in the event of segment or adapter failure. Disk and tape servers can use some of the network bandwidth provided by the additional network connection. Critical satellites can be booted using the other LAN adapter if one LAN adapter fails.
- Connecting noncritical satellites to only one LAN segment helps to balance the network load by distributing systems equally among the LAN segments. These systems communicate with satellites on the other LAN segment through one of the bridges.
- Only one LAN adapter per node can be used for DECnet and MOP service to prevent duplication of LAN addresses.

Configuring three LAN segments



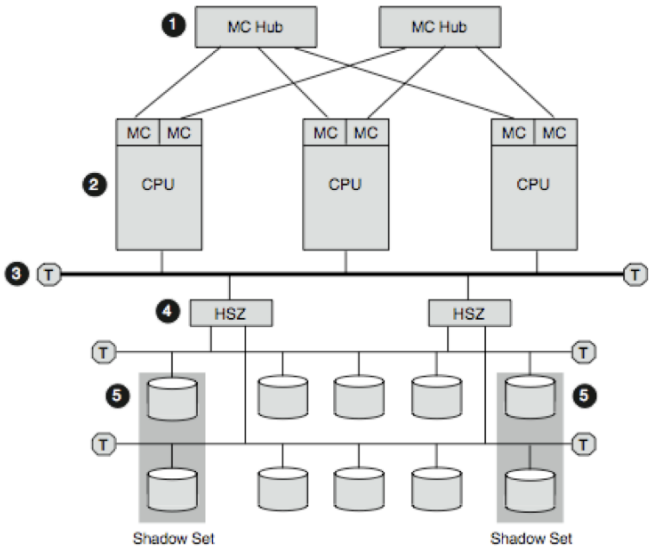


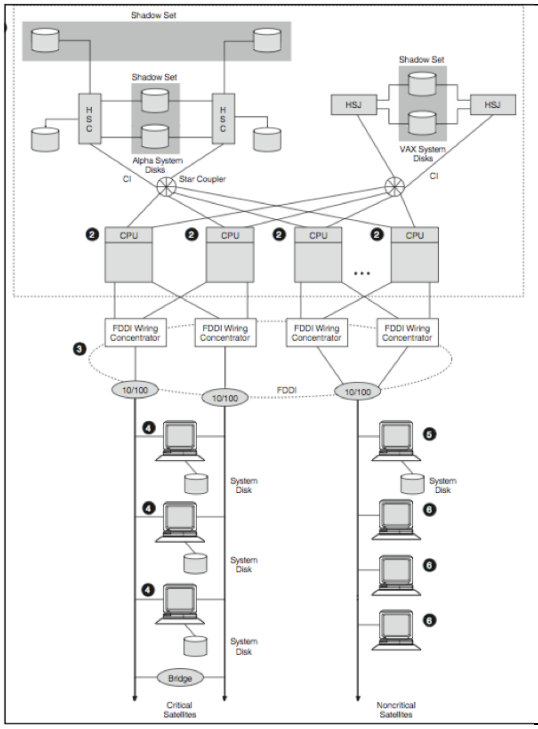
DSSI OpenVMS Cluster System



CI Open VMS Cluster

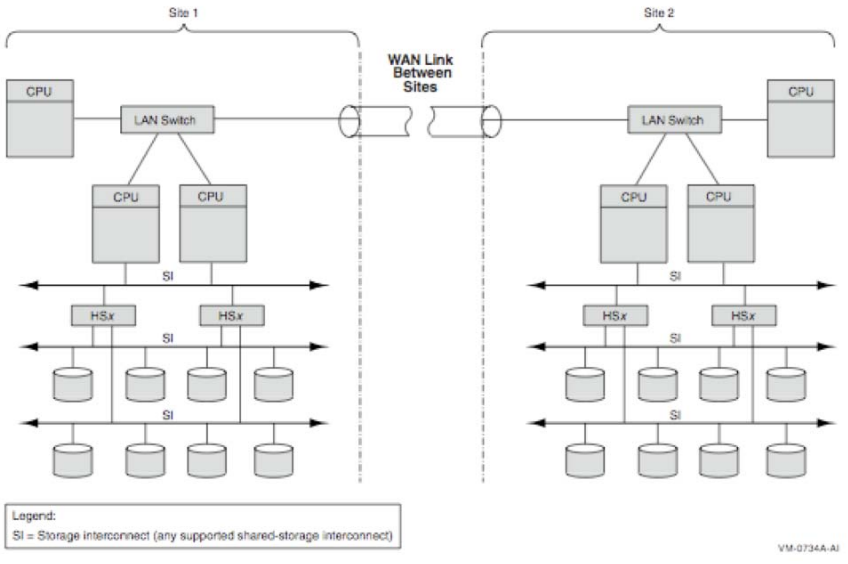
Memory Channel Cluster





Availability in Cluster with Satellites

Multiple Site Cluster with WAN link



Reference

- Guidelines for OpenVMS Cluster Configurations Order Number: AA–Q28LH–TK
  January 2005 OpenVMS Cluster availability, scalability, and system management benefits are highly dependent on configurations, applications, and operating environments. This guide provides suggestions and guidelines to help you maximize these benefits.

- http://h71000.www7.hp.com/doc/os83_index.html
- http://h71000.www7.hp.com/doc/82FINAL/6318/aa-q28lh-tk.PDF