Hewlett Packard Enterprise

HPE OpenVMS Clusters

Contents

Technical white paper

Introduction

Clustering technology is widely adopted in the IT industry for high performance and also to provide high availability. Multi-site clustering involves forming a single cluster from two different data centers that are separated geographically. This technology is adopted by enterprise customers in verticals like financial, healthcare, and manufacturing for high availability and business continuity. It also gives the advantage of scale-out architecture. In a converged infrastructure, clustering provides the required attributes of robustness, scalability, and simplified management.

This article will brief about the diverse technical aspects (server, networking, storage) involved in setting up multi-site OpenVMS Clusters with built-in redundancy to avoid single point of failure in any of the components. Mission-critical computing requires stringent recovery-point objective (RPO) and recovery-time objective (RTO). It requires expertise to build a mission-critical environment that can avoid any single point of failure in the system. The article will provide a high-level overview of various software components of OpenVMS clustering technology along with management strategies and the security principles for the environment. The article will describe different cluster network transports and different models like active-standby and active-active clusters used in mission-critical computing. The various concepts and techniques will be elaborated with a real-world example of multi-site clustering used in mission-critical environments.

Introduction to clustering

Clustering gives essential attributes of robustness, scalability, and availability to the system. In a converged infrastructure, clusters are deployed as a part of a multi-tier solution. Clusters provide the reliable environment for the middleware providers to implement highly reliable and available servers. A cluster system is a set of closely coupled nodes cooperating using the cluster software and the operating system so as to share the resources among all the users in the system. A cluster system gives the advantage of a "single system" view for the end users. Clusters software includes a subset of components that interact and interface so as to achieve the clustering functionality. A cluster is built with a well-defined set of functionality like the ability to dynamically add new nodes to the cluster, share the resources among all the processes or users on various nodes in the clusters, detect failure of a node, and perform the necessary cluster transition to bring nodes back to the steady state. In a converged infrastructure, clustering plays a polymorphic role by providing the ability to do infrastructure automation, resource pools, and scale-out architecture along with high availability and high performance.

System or software stack—Role of cluster

Cluster as a software solution exhibits a polymorphic attribute and reusability by different layers in the software stack. Clusters provide the needed attributes of scalability, availability, and resiliency across the software stack.

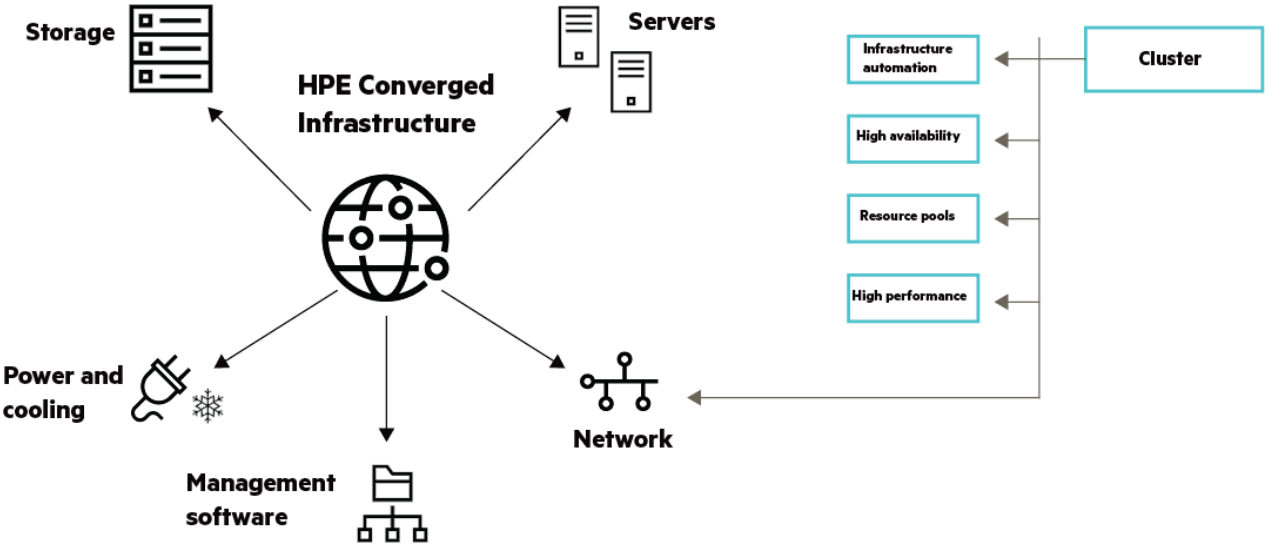

Figure 1. Cluster in converged infrastructure

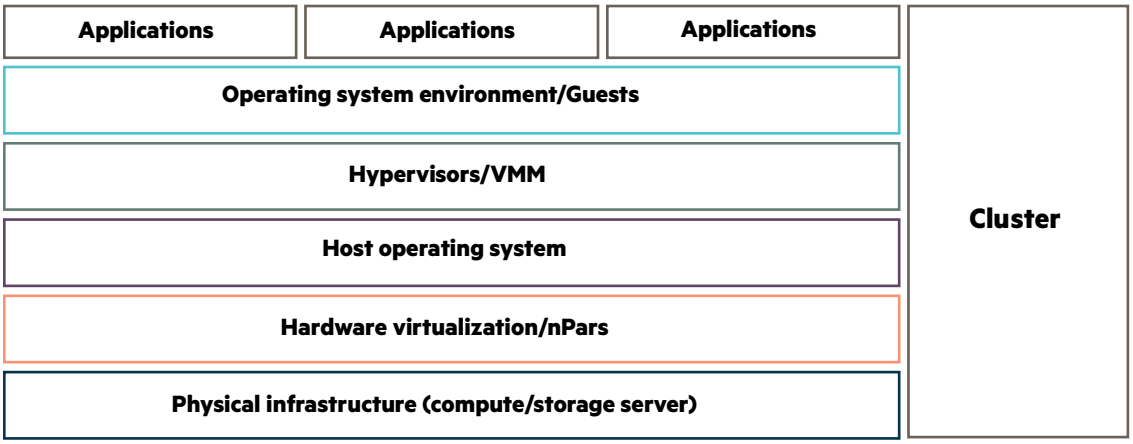

Figure 2. System or software stack—Role of cluster

High performance clusters and high availability clusters

High performance clusters (HPC) are used in a wide variety of scientific research, space, and manufacturing and technology industries. HPC involves building a large cluster farm of multiple nodes giving high performance processing capabilities. An example of such an infrastructure would be search engines processing millions of search queries per second. Load balancers are used for distributing the load across multiple servers in the cluster.

High availability clusters are used for ensuring business continuity by running critical applications and having the necessary capability to fail-over to different nodes in the cluster should a failure occur. This technology is used to ensure business continuity in industries like healthcare, financial institution, and transportation.

HPE OpenVMS Clusters

OpenVMS clustering technology, a proven clustering technology considered as "gold standard" for disaster tolerance and hailed for its security and reliability, is adopted by major customers of Hewlett Packard Enterprise for whom downtime is never an option. OpenVMS Clusters can safely operate over a distance of 500 miles. OpenVMS Cluster over IP introduced with OpenVMS V8.4 enhances the capability of multi-site clustering capability by using IP protocol for cluster communications. Host-based volume shadowing for OpenVMS provides customers the unique advantage with storage replication by having up to six members in a single shadow set. The HPE Availability Manager software complements by giving the necessary management support for OpenVMS Cluster operating environment.

Components of OpenVMS Cluster software

The various components in OpenVMS clustering software involve connection manager, lock manager, cluster communication, and a cluster-wide file system.

Connection manager is responsible for coordinating the cluster. Connection managers are distributed on all cluster nodes and collectively decide upon cluster membership. A quorum scheme is defined and enforced by the connection manager so as to avoid split-brain syndrome and cluster partitioning. A quorum service is provided by a node in the cluster and can also be provided by a storage disk with direct connectivity to all nodes in the cluster. An OpenVMS Cluster can have up to 96 nodes in a single cluster, and each node can have up to 32 cores.

Distributed lock manager (DLM) is the basis for resource sharing across different servers in a cluster environment. Each resource in the cluster environment is known by a unique name that is shared by the cooperating processes in the cluster to synchronize the access. The required synchronization is achieved using the services provided by the lock manager. The DLM forms a building block for a cluster-wide distributed file system. The DLM provides the ability to share resources in a cluster environment so that applications can execute coherently in a synchronous manner when needed.

Cluster communication involves the ability to use the transport reliably and efficiently for exchanging messages between the nodes of the cluster. Quality of service (QoS) parameters like latency and bandwidth define the choice of interconnects like InfiniBand or Ethernet for cluster communication. OpenVMS supports Ethernet for cluster communication, heartbeats or hellos are exchanged between multiple nodes, and loss of heartbeats triggers the appropriate events forcing a cluster transition. OpenVMS Cluster network interconnect or system communication architecture (SCA) uses both layer 2 with protocol ID 60-07 directly over Ethernet and layer 3 (IP) for long distance cluster communication. Intra-Cluster Communication **(ICC)** provides services to applications on various systems in a cluster to exchange messages efficiently. This is ideal for financial institutions looking for low latency communication. It follows a client server paradigm and provides a programming interface for applications. ICC has the ability to do connection management between the nodes and relies on SCA for the reliable delivery of messages.

Cluster-wide file system provides processes, running in multiple nodes in a cluster, simultaneous access to a common block of storage while ensuring data integrity. The mutual exclusion that is required for data integrity during a simultaneous update is enforced by strict locking rules in a distributed model or by having dedicated server architecture.

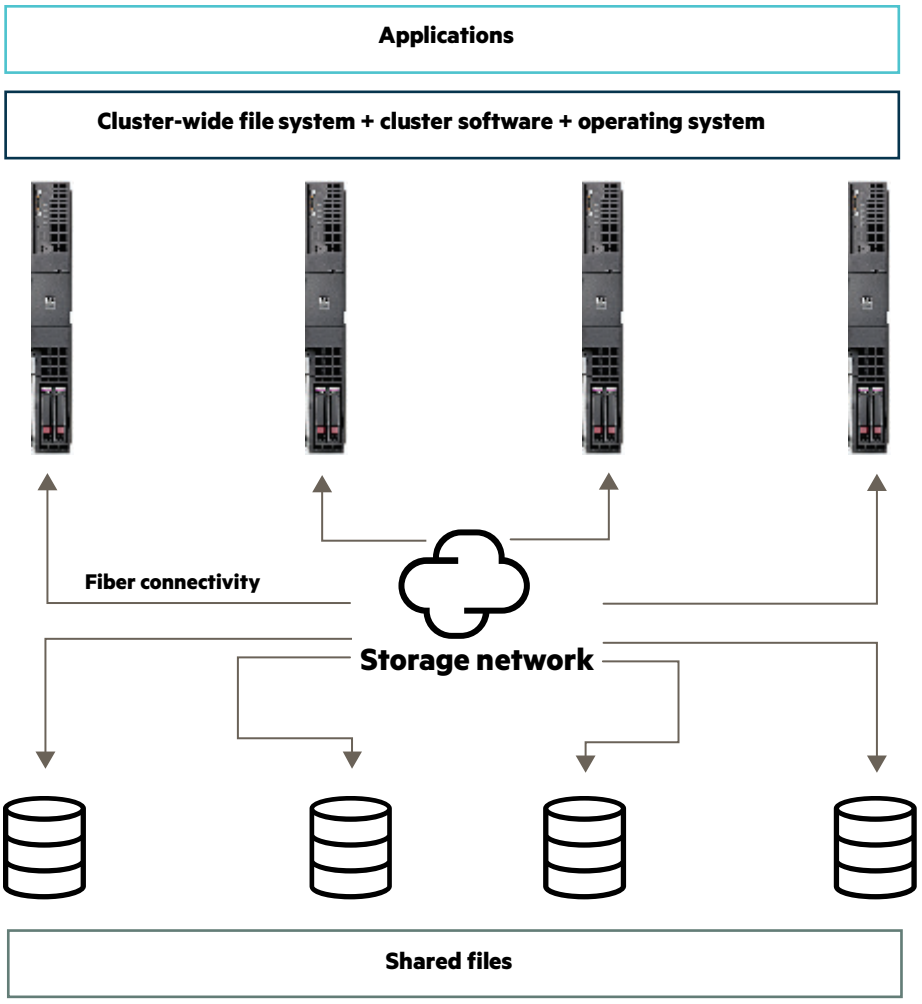

Figure 3. Cluster-wide file system

Multi-site OpenVMS Clusters

Multi-site clustering involves a group of nodes distributed in different sites working together as a single cluster environment. This technology is ideal for enterprise customers in verticals like financial, healthcare, and manufacturing to provide high availability and business continuity. Multi-site clustering also gives the advantage of a scale-out architecture. Inter-site distance can vary from a few miles to hundreds of miles based on business requirement and the QoS definitions of the applications. Latency due to speed of light is an important factor for long distance multi-site clusters, with greater distances causing greater latency. It is also important from a performance and QoS standpoint to localize the application with necessary resources at the local site and build the failover capability with multiple sites. For example, it is possible for applications and users on a site to use the site local storage for read and write, and provide storage-based multi-site replication and failover to the storage at a remote site should failure be detected in the local storage. Multi-site clusters are built with the ability to scale out applications by distributing the data sets across the set of nodes in the cluster. For example, in the case of stock exchanges, the data sets could be a group of stocks, and for healthcare, it can be patient records grouped by different criteria. Multi-site cluster formation includes the ability to detect and to discover nodes beyond the virtual LAN (VLAN) segment in an IP only network. The ability to add and remove nodes dynamically without any disruption to the cluster is an important aspect to be able to get more compute capacity and enable a scale-out architecture for applications. It also provides the ability to perform rolling updates to various nodes in the cluster without having to interrupt cluster services to the user community.

Multi-site clusters involve active-standby model, where applications active at one site will be standby at remote site and vice versa. A failover is forced when a node failure is detected and cluster nodes undergo a transition. Multi-site clusters can also be active-active with applications executing simultaneously at different sites. Clustering software provides the necessary capability to load balance the multiple available networks with bandwidth aggregation and transparent failover capabilities. During a node or site failure, a transparent migration of the applications and redistribution of the data sets among the existing set of nodes in the cluster provide the necessary failover capabilities. Cluster transition is invoked by cluster software and can be completed in seconds and arbitrated by a Quorum site. It gives the ability to achieve the defined service-level agreements (SLAs) with the ability to withstand failure in any single component, server, or site in the system.

Storage replication

Storage replication using synchronous and asynchronous techniques is deployed for protecting against storage failures. The replication can be achieved using either host-based methodology or by having a controller-based replication. Clusters also provide the ability to serve storage using protocol similar to NFS between the nodes using the cluster transport. OpenVMS provides Mass Storage Control Protocol **(MSCP)** and is used for serving locally attached storage to different nodes in the cluster. This enriches the multipathing capabilities for accessing the storage devices across the site. For example, if the direct path to the storage on a remote site using the storage network has a problem, then access to remote storage can happen over the cluster network and this transition is transparent to the application. The HPE Volume Shadowing for OpenVMS software works with OpenVMS Cluster software to provide host-based storage replication; Volume Shadowing binds multiple physical disks into a single virtual unit (shadow set). Applications issue I/O to the virtual unit and shadowing software replicates data across the multiple physical disks transparently.

Management

OpenVMS Clusters provide a single system view of all nodes and hence simplified management for system managers. OpenVMS Satellite systems enable cluster nodes to boot from a common boot server over the network, which makes it easier to scale without additional management overhead. Cluster management involves being able to effectively manage the multi-site data centers remotely and to provide fault isolation and recovery capabilities using the management software. A single pane of glass for real-time monitoring of entire cluster nodes along with the various elements like network and storage provide a distinct advantage to system managers. Cluster-wide alerts and logging capabilities are necessary to understand the sequence of events leading to a failure and isolate the root cause, and corrective action can be taken by improving either the software or infrastructure, or both. Management software and tools (HPE Availability Manager, T4, MONITOR) provide the ability to monitor the health of the system, their peak and average loading, and headroom available on different nodes in the cluster. Management tools provide the knobs to control the cluster transition during site or node failures. Cluster management software makes it convenient for system administrators to manage the environment from any location. The HPE Availability Manager software provides the necessary management capabilities to the OpenVMS Cluster-operating environment.

Security

The model for providing security in an IP and WAN environment is:

- Isolating IP subnets conveying cluster communications subnets from the WAN IP environment
- Ensuring that cluster communications over insecure WAN links will be encrypted and authenticated

Standard firewall technique would be applicable. Customers whose intranet spans multiple sites must (and normally do) use secure private links, or site-site encryption such as virtual private network (VPN) type tunneling, between the firewalls. Firewall and dedicated network for cluster communication along with encryption can be adopted for internal security. IPSec can be enabled between the sites using site-site IPSec capabilities of the HPE Networking routers; thus, all IP-based cluster communications would be secure without any other external units. Security within the LAN is also provided by isolating cluster traffic using VLAN tagging in the networks.

Cluster authorization for all users in a multi-site environment is managed with ease by having a common cluster authorization file.

Virtualization

Virtualization improves efficiency in server utilization thus giving the opportunity for consolidation and optimization. The ability to achieve high availability along with efficient utilization of server resources can be achieved with a combination of clustering and virtualization. A cluster can be formed with guest nodes and brings about multiple advantages. For example, at Quorum site, a guest node can be performing the role of a Quorum node. Guest nodes can be clustered for applications running on highly available guest systems. Further host nodes can be clustered to provide high availability to applications running on hosts. OpenVMS V8.4 supports OpenVMS as guest virtual machines (VMs) on the HPE Integrity VM software running on HP-UX host giving the advantage to combine the leading clustering and virtualization technologies.

Cloud computing

Cloud computing has emerged as a paradigm shift, redefining the information technology industry. Cloud technology defines the various service offerings for infrastructure, platform, and software as service to end users. The scale of cloud deployments can be large, such as managing a network of thousands of VMs within single or multiple data centers. The ability to provide a reliable, guaranteed QoS is a strong requirement and differentiates one service provider from another. It is a challenge to scale and meet the growing demands by using the resources in an optimal way and also to provide the guaranteed QoS. Multi-site clustering provides the ability to scale, and create a reliable and highly available platform for cloud computing. It also makes it possible to move mission-critical applications to a cloud environment.

Multi-site OpenVMS Cluster reference architecture

A typical disaster tolerant multi-site clustering involves two main data centers and a third site with a single node acting as a Quorum site. Multiple sites are connected with network services typically provided by Telco operators. Enterprise servers typically include multiple 10GbE interfaces and dedicate a single interface explicitly for cluster communication by forming mesh networks among the cluster nodes, which helps in lowering latency and provides higher bandwidth. Techniques like auto port aggregation (logical LAN) provide the necessary failover capabilities at the network layer. QoS is guaranteed with dedicated network switches for cluster communication and separating cluster traffic from other traffic in the data center. Multi-site networking involves extended LAN (layer 2) or IP Network (layer 3) services for cluster communication. OpenVMS Clusters also have the ability to use TCP or User Datagram Protocol (UDP) along with IP multicasting for long distance cluster communication.

Multi-site OpenVMS clustering is built with QoS parameters like latency and congestion avoidance so as to guarantee acceptable applications performance. Latency due to speed of light is a key attribute that defines the maximum inter-site distance for multi-site clusters for applications to provide the necessary end-user experience. RPO and RTO are defined by the business and are used as strict guidelines for defining and setting up various components in the environment. Mission-critical environments set very strict objectives, which enforces redundancy at every component. The various components involve server, network, and storage, in addition to systems in the data centers. The ability to deal with multiple failures without affecting the users provides the necessary mission-critical capabilities for the enterprise customers. In addition to automatic failover, system managers are alerted by sending automatic alarms bringing in the necessary action to repair or replace the faulty components. In this reference architecture, we have two sites and each of these sites consists of nodes of a single cluster with the third site acting as a Quorum site. Jobs are distributed based on each site so as to achieve load balancing and the flexibility to scale out. OpenVMS distributed queue manager provides the ability to distribute batch jobs to various nodes in the cluster. Storage replication is used to replicate transactions in either direction. HPE Volume Shadowing for OpenVMS is used to replicate storage across the site by creating a shadow set consisting of the physical disks as a mix of local disks and disks at remote site. The reads can be satisfied with the physical disks at the local site by setting the "READ_COST" value appropriately for the members of the shadow set. This gives the advantage of faster reads in a multi-site environment. Volume Shadowing has the ability to track the modified blocks using in-memory bit maps. This is used to speed up the merge operation for bringing the members to a consistent state following a failure.

In case of a site failure, jobs and transactions are transparently migrated to the other site with no intervention. Once the site is restored, storage at both sites is synchronized to ensure data is consistent. An OpenVMS Cluster provides a set of system tunable (sysgen parameters) that can be used to customize the cluster transition and recovery process. The "RECNXINTERVAL" sysgen parameter defines the number of seconds elapsed before a node is removed from the cluster in the event of a failure. The "TIMVCFAIL" sysgen parameter specifies the time required to detect a communication failure between the nodes. The "PE4" sysgen parameter specifies the hello interval and listen timeouts for the cluster communication. The mentioned sysgen parameters can be customized for the environment to achieve a cluster failover of less than 10 seconds. The SCA Control Program (SCACP) management interface provides the ability to automatically calculate the maximum window size between the nodes based on the distance between the nodes for better performance. The various sysgen parameters give maximum flexibility to system managers and architects to manage the environment based on their requirement. Refer to the cluster manuals for more information on the parameters. In addition, HPE Availability Manager can be used to adjust the Quorum during cluster transition.

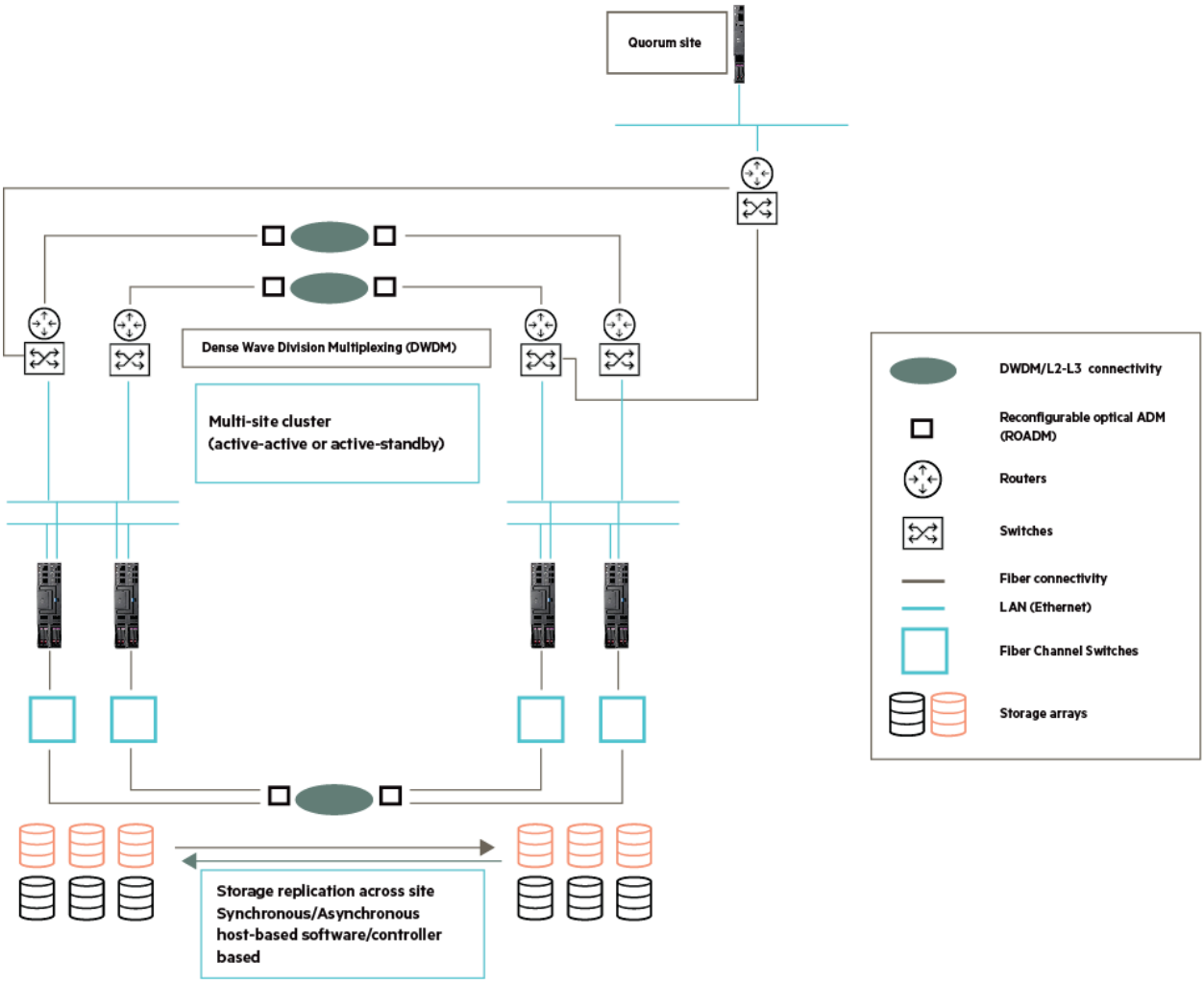

Figure 4. Multi-site OpenVMS Cluster reference architecture

Latency

A business that requires a high level of availability and protection against any site level failure will require distributing their technology data centers in multiple locations. Multi-site clustering between the data centers is characterized by longer distances between the sites for better protection and at the same time shorter distances to achieve expected performance levels. Application performance is characterized by inter-site latency when the transaction involves completion of task (say I/O) both locally as well as at the remote site. Latency between nodes separated by a distance is actually a measure of the circuit distance that connects the two sites.

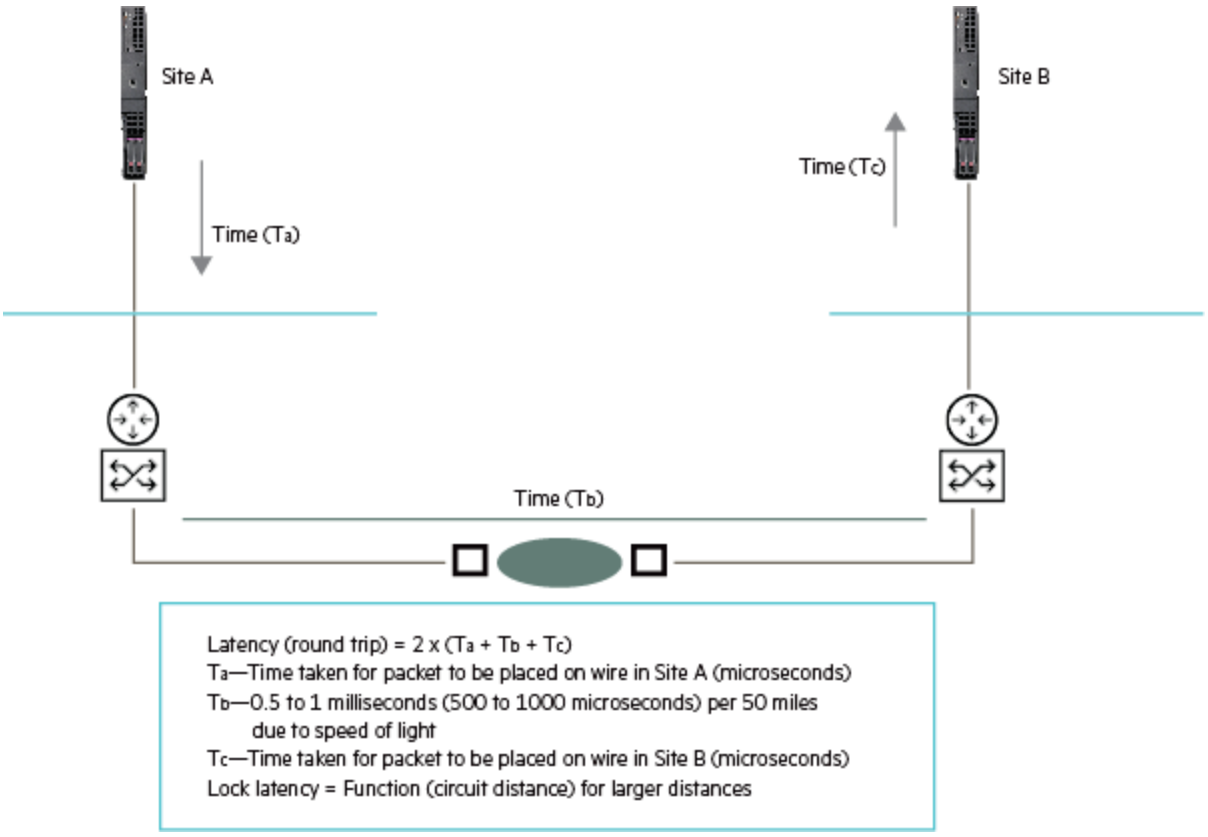

Figure 5. Inter-site latency

Clusters—Mesh topology

Hewlett Packard Enterprise enables mission-critical computing infrastructure with the blade servers combined with the HPE Virtual Connect technology. HPE Blade server technology forms an ideal platform for cluster computing because of the ability to scale out with more number of processors and nodes in addition to the simplified management software. HPE Integrity Server Blades are able to meet the requirements of simplicity, scalability, and resiliency for the mission-critical workloads. HPE Integrity Blades with 10 Gbps Ethernet is able to provide latency in the order of microseconds between the nodes in the enclosure. In addition, cluster failover and transition can be achieved transparently within a few seconds giving uninterrupted service for the end user using the applications. HPE Integrity Blade servers provide multiple LANs on Motherboard (LOM) and each LOM can be used to create an explicit connection with one node in the cluster to form a mesh topology providing high bandwidth and low latency for intra-node communication in the cluster. This is a viable approach for nodes that are in the same enclosure using Ethernet for cluster communication. The OpenVMS Fast Path technique allows every device to have affinity with a particular CPU so as to improve performance. This technique can be used to have the LAN device used for cluster communication and the Port Emulator (PE) device (cluster communication) affinitize to the same CPU to achieve better performance (low CPU overhead and low latency).

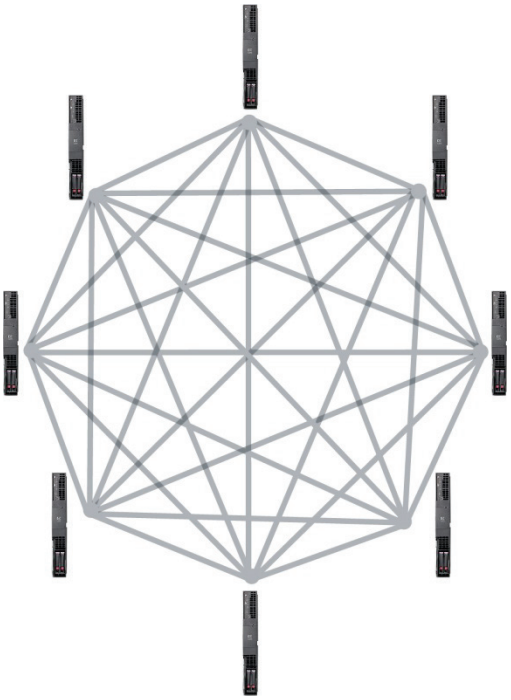

Figure 6. Clusters—Mesh topology

Summary

OpenVMS Clusters continue to provide the mission-critical capabilities for enterprise customers. OpenVMS Clusters have continued to evolve with newer improvements and support for the HPE Integrity i2 server environment. Customers get the distinct advantage of higher resiliency and lower TCO by migrating their infrastructure to the HPE Integrity environment. Customers can take advantage of the migration support available for multi-architecture cluster (Alpha-Integrity i2) servers to move from Alpha to Integrity in a phased manner. OpenVMS Clusters give the flexibility to build highly available and disaster tolerant IT infrastructure.

Learn more at

hpe.com/docs/openvms
hpe.com/servers
hpe.com/networking
hpe.com/storage

4AA4-7138ENW, October 2016, Rev. 1

Hewlett Packard Enterprise