OpenVMS Technical Journal V8

Disaster Tolerance Proof of Concept: HP Volume Shadowing for OpenVMS and Oracle9*i*® RAC over Extended Distances

Melanie Hubbard - OpenVMS Sustaining Engineer
Carlton Davis- OpenVMS Partner Relationship Manager
Craig Showers - OpenVMS Solutions Center Manager
John Andruszkiewicz - OpenVMS Engineer
Keith Parris - Systems Engineer

Overview

In today's high-risk world, the goal of information technology's contribution to business continuity is to ensure that an enterprise's data, transactions, and IT infrastructure continue to be available, regardless of adverse conditions. This is generally known as *disaster tolerance*. These conditions can go as far as the loss of an entire data center. When considering various business continuity implementations, an enterprise must take into account the potential cost of losing its IT capabilities. A generally recommended implementation for mitigating this risk is to deploy at least two geographically separate data centers, often at distances measured in hundreds of miles.[1]

The measure of a disaster-tolerant environment is its ability to keep working as the database or operating system recovers from any failure. The overall goal for this project is to demonstrate that HP Volume Shadowing for OpenVMS along with Oracle9*i*® RAC, as well as the connectivity technology from LightSand and Digital Networks, can ensure high availability and data integrity over a variety of distances, including those formally supported by OpenVMS Clusters, as well as distances longer than what is currently supported.[2]

---

[1] *Interagency Paper on Sound Practices to Strengthen the Resilience of the U.S. Financial System* Federal Reserve System [Docket No. R-1128], Department of the Treasury Office of the Comptroller of the Currency [Docket No. 03-05], Securities and Exchange Commission [Release No. 34-47638; File No. S7-32-02]. Refer to http://www.sec.gov/news/studies/34-47638.htm for more information.

[2] The current maximum supported distance for OpenVMS is 250 km (150 mi); the OpenVMS Disaster Tolerant Cluster Solution (DTCS) services package generally supports internode distances up to 500 mi (800 km) but can span greater distances depending on requirements and

These distances can be thought of as representing what is sometimes referred to as a Local or Campus configuration (0 mi/0 km – 18 mi/30 km between nodes), a Regional configuration (upwards of 372 mi/600 km between nodes), and a Geographically Extended configuration (upwards of 621 mi/1000 km between nodes). This proof of concept validates that long distance disaster-tolerant Oracle RAC on OpenVMS systems can be successfully deployed.

For more information on LightSand Communications, Inc., Digital Networks, and Oracle RAC, see Appendix B.

Business Needs

When data security and availability are critical to their success, enterprises require a computing solution that protects their information systems from disasters such as power outages, earthquakes, fires, floods, or acts of vandalism. The effects of a disaster range from temporary loss of availability to outright physical destruction of an entire facility and its assets. In the event of such a disaster, the system design must allow organizations to shift their information-processing activities to another site as quickly as possible, with minimal loss of function and data. Therefore, procedures for disaster recovery must be predictable, well-defined, and immune to human error.

Disaster tolerance is characterized by a short recovery time (low Recovery Time Objective -- RTO) and avoidance of data loss (low Recovery Point Objective -- RPO). In a disaster-tolerant system based on this approach, redundant, active servers and client interconnects are located at geographically separated sites. Should the environment at one site suffer a disaster, applications that were running at the now-disabled site can continue to run at the surviving site.

While the cost of lost IT capabilities is the key consideration, affordability is also a factor. Deploying the network infrastructure over a long distance, depending on its configuration, can be prohibitively expensive, especially when it has to handle cluster traffic, Oracle RAC operations, and data replication. This is an analysis that should be taken by any organization considering a disaster-tolerant IT environment. Cost-benefit, however, is not the subject of this proof of concept. Rather, it is intended to demonstrate the operational capability of one specific and popular architecture over distances that are considered appropriate for disaster tolerance.

Business Solution

HP OpenVMS, Digital Networks, and LightSand Communications, Inc. have joined to test Oracle9*i* RAC in disaster-tolerant configurations over a variety of distances using Volume Shadowing[3] technology under the control of an OpenVMS host system. The goal for this proof of concept was to observe and record the behavior of an Oracle9*i* RAC server on a clustered OpenVMS system using Volume Shadowing across extended distances. The distances tested are 0 mi (0 km), 372 mi (600 km), and 621 mi (1000 km).

OpenVMS cluster uptimes are often measured in years. Active-active cluster technology at both the operating system and database level (Oracle RAC) and cluster-aware applications provide the capability to shut down individual systems for proactive reasons with zero application availability impact. *Multi-site* clusters can include the capability to proactively shut down an entire site with zero application availability impact. No applications or end-user connections would need to fail over as they would already be running on other systems.

In the event of unexpected or unplanned outages, only the connections to that single system would be impacted from an availability perspective. Other servers would continue to process their existing connections. The failed server connections would then automatically reconnect to other servers that are already running in the cluster. Because the applications and storage devices are already running

---

circumstances. Basic cluster protection and data protection can be between distances as great as 60,000 mi (97,000 km), however, the latency at greater distances may not be acceptable for specific customer implementations.

[3] HBVS uses RAID-1 technology. Refer to http://h71000.www7.hp.com/openvms/products/volume-shadowing/index.html for more information.

and available on the shared file system on other servers, the failed server connections fail over extremely quickly.[4]

The connectivity technology used in this proof of concept provides an example of a cost-effective *multi-site* network configuration. There are multiple ways to simulate a long-distance cluster without the actual expense of real long-distance inter-site links; such as delaying packets and thus simulating distance via latency. The Spirent/AdTech product, the Shunra STORM network emulator, or a PC running the free NIST Net software from the National Institutes of Technology can be used to delay traffic. Most of these tools can delay IP traffic only, not LAN traffic in general.[5] Since OpenVMS Clusters use the SCS (sometimes called SCA) protocol on LANs and SCS is not an IP family protocol, you need a method to convert SCS traffic into IP format.

One method of converting SCS traffic is to use routers (such as Cisco) running a Layer 2 Tunneling Protocol version 3 (L2TPv3) tunnel to encapsulate LAN traffic (including SCS) and send it over IP. The ability of the LightSand boxes to bridge both Fibre Channel and LAN (including SCS) traffic over IP provided a solution for encapsulating SCS without the need for routers and an L2TPv3 tunnel. Thee latter, less complicated configuration was chosen for our series of tests.

Business Summary

Results demonstrate that, under testing, Oracle RAC in conjunction with Volume Shadowing works across extended distances. All components continued to function without interruption over distances of up to 621 mi (1000 km). Longer distances could be used but every environment will be different depending on such factors as workload, transaction size, required transaction rate, user response requirements, database size, and site hardware. Each site would need to evaluate the effect of latency on their application and make decisions based on their current functional needs.

The chart for Test 4E7, Host-Based Minimerge (HBMM) Forced Merge, shows a noticeably longer time to return to steady state than the other HBMM tests. This was due to the fact that, although both nodes had bitmaps enabled, only one bitmap was active and it was running on the node that crashed. Effective disaster-tolerant configurations must ensure that multiple active bitmaps are defined. This is done using the DCL command shown under Test 4E1, and fully documented in the Volume Shadowing/Host Based Minimerge documentation referenced in that test description.

Test timing data for all distances tested shows that the time it takes for a Shadow Set member to return to steady state using HBMM (Host Based Minimerge) is significantly less than that for a full merge; therefore, HBMM is the best option to return members to steady state.

Test Environment

The hardware and software were configured in a two-node disaster-tolerant configuration that would be typical of an enterprise meeting the stringent requirements for a live remote datacenter DT environment. This included commercially configured servers, storage, and infrastructure components that are frequently seen in production environments. In addition, a network delay component was added to realistically emulate network latency behavior over the test distances.

See Appendix A for a list of the selected hardware and software configurations for this proof of concept. Figure 1 shows the hardware configuration.

[4] Refer to www.hp.com/go/openvms/availability for more information.

[5] We believe that with a new release of firmware, the Shunra box can delay ordinary LAN traffic, not just IP. If LAN traffic in general can be delayed, then a simpler test environment is possible -- you merely need to connect the delay box between a couple of LANs. Storage traffic can be put on the LAN and delayed through the same box, either via MSCP-serving or using a technique called "SAN Extension." MSCP-serving can be used for remote storage access in an OpenVMS Cluster. If it is desired to bridge Fibre Channel SANs between sites instead of using MSCP Serving as the primary remote disk access method, then any of the SAN Extension boxes on the market which can bridge Fibre Channel over a LAN (or an IP network running on a LAN) can be used.

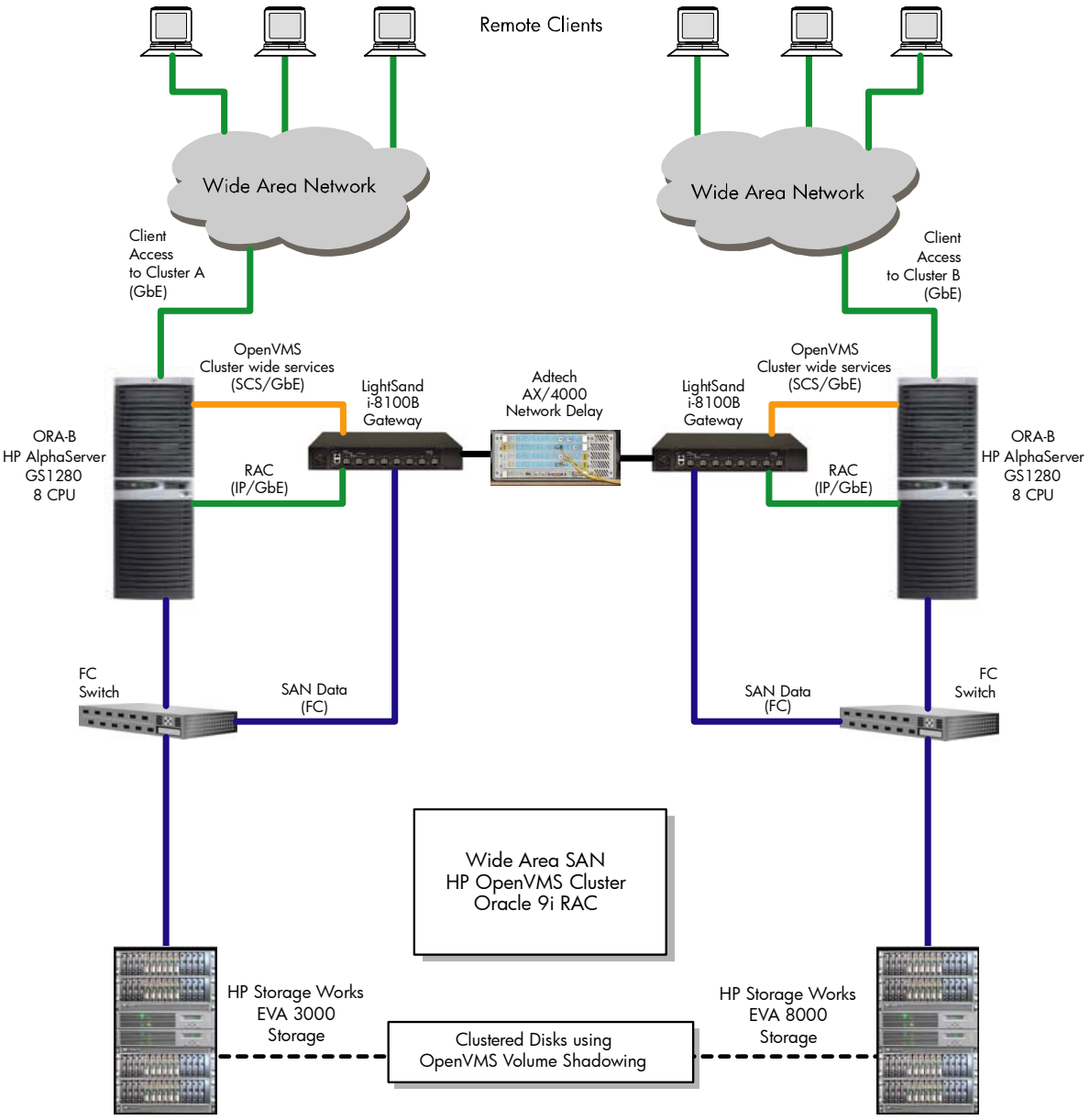

Figure 1 - Hardware Configuration Diagram

Test Description

The testing goal was to provide a specific, repeatable work load (via Swingbench) running on various Volume Shadowing configurations with the OpenVMS Cluster nodes separated by various distances and record the results of the work load. The work load generated by Swingbench simulated typical order-entry functions of 300 remote clients and generating 600,000 transactions (sometimes referred to as 600K transactions).

The testing goal was accomplished by having a baseline copy of the system and database created before the tests were started, which was restored at the beginning of each test run to ensure a common starting point for each test configuration. Also, the network delay hardware was set for the distance that was being tested before each test run.

The data collected for analysis from each run included:

- The output of the Swingbench benchmark software, which includes the total number of transactions, total time to completion, as well as the minimum, maximum, and average response times for each type of transaction.
- The output of T4, an OpenVMS timeline tracking tool, which captures and consolidates important OpenVMS system performance statistics.
- The output of system- and device-specific DCL commands, which shows the current state of those devices.

Note that no performance tuning was done at any time during these tests nor was it ever intended to be done. Tuning recommendations are typically very specific to an application and the system it is running on, and that was not part of this project.

Initial tests were done and a decision made that the site identifier for the disks should be defined as local to their node, and that each node would have a unique identifier. This ensures that the read cost to the local Volume Shadow members is optimized. This is the default behavior of Volume Shadowing. Refer to the HP Volume Shadowing[6] for OpenVMS documentation for complete details.

There are a countless number of test variations which can be run, using varying numbers of local and remote members of each Shadow Volume Set. We chose the following tests to represent an appropriate cross-section of those variations.

The following tests were performed for various Shadow Set Member (SSM) configurations and their results were recorded.

- Steady State: Two Shadow Set Members (Test 7A1)
- Full Copy State: Two Shadow Set Members (Test 2B1)
- Minicopy Logging/Recovery: Two Shadow Set Members (Test 2C1)
- Full Forced Merge State: Two Shadow Set Members (Test 4D2)
- Host-Based Minimerge (HBMM) Demand Merge: Two Shadow Set Members (Test 4E1)
- Host-Based Minimerge (HBMM) Forced Merge: Two Shadow Set Members (Test 4E7)
- Host-Based Minimerge (HBMM) Demand Merge: Three Shadow Set Members (Test 4F1)
- Host-Based Minimerge (HBMM) Forced Merge: Three Shadow Set Members (Test 6F5)

Test Results

For each test, two tables are shown. The first table shows the total time (in `hh:mm:ss` format) to complete 600,000 transactions for each distance specified. The distances are emulated by adding known network latency or delays (in milliseconds) in the data transmission between the nodes.[7] The second table shows the time it takes to return the shadow set member to steady state (fully copied or merged).

The distance equivalents are as follows:

```
00 ms = 0 mi/0 km   03 ms = 372 mi/600 km   05ms = 621 mi/1000 km
```

Test 7A1. Steady State: Two SSMs

This test consists of a two-member Shadow Set Volume, with one disk local to each node and one disk remote to each node. The test is started only after each disk is 100% copied (also known as in a Steady State) and is a full member of the Shadow Volume set.

[6] Refer to http://h71000.www7.hp.com/openvms/products/volume-shadowing/index.html for more information.

[7] Although the speed of light is used as the general speed of data through a network, there is some minor additional delay due to the laws of physics. Thus, the speed of data across the network is not the theoretical speed of light 0.66 ms/100 km per round trip, but instead a slightly slower (and generally accepted) speed of 0.5 ms/100 km per round trip. Thus, a 5 ms delay emulates a distance of 1000 km between nodes.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**7A1 – Time for Completion (HH:MM:SS)**
00 /00	00	1:32:04
372 /600	03	5:00:00
621 /1000	05	8:01:11

The following table shows the amount of time to complete a full (100%) disk copy and return to steady state before the transactions are started.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:10:00	00:50:00	01:20:00
372 /600	03	00:10:00	00:50:00	01:25:00
621 /1000	05	00:15:00	01:35:00	02:15:00

Test 2B1. Full Copy State: Two SSMs

This test consists of a two-member Shadow Set Volume, with one disk local to each node and one disk remote to each node. One disk is mounted and is a full member of the set. The test is started and then the second member is mounted.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**2B1 – Time for Completion (HH:MM:SS)**
00 /00	00	1:40:16
372 /600	03	5:57:39
621 /1000	05	8:24:58

The following table shows the amount of time it took for the disks to become full members (this is for a copy operation starting after the test is running) and return to steady state.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:02:00	00:12:30	00:07:54
372 /600	03	00:14:12	01:02:23	00:25:42
621 /1000	05	01:08:47	02:23:47	01:13:47

Test 2C1. Minicopy Logging/ Recovery: Two SSMs

This test consists of a two-member Shadow Set Volume, with one disk local to each node and one disk remote to each node. Both members are 100% copied (in a Steady State). At that time, the remote member is removed from the Shadow Set volume (using the OpenVMS DCL command:

`DISMOUNT <Device-name> /POLICY=MINICOPY`)

The test is then started and run until completion. The remote member is then mounted and a minicopy takes place.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**2C1 – Time for Completion (HH:MM:SS)**
00 /00	00	1:24:18
372 /600	03	5:27:48
621 /1000	05	8:50:45

The following table shows the amount of time to return the second shadow volume member to steady state at the end of the task completion using the `MINICOPY` policy.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:01:48	00:00:04	00:01:32
372 /600	03	00:04:01	00:00:24	00:04:32
621 /1000	05	00:07:33	00:01:01	00:07:05

Test 4D2. Full Forced Merge State: Two SSMs

This test consists of a two-member Shadow Set Volume, with one disk local to each node and one disk remote to each node. Both members are 100% copied (in a steady state). The merge state can be entered when a system crashes (using the VU mounted crash or by using a DCL command requesting a merge). The system crash will force Oracle9*i* RAC to fail over to the other system in the cluster; therefore, Oracle TAF must be enabled and functioning. After the start of the test, one system is crashed, which forces a demand merge of the remaining Shadow Set members.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**4D2 – Time for Completion (HH:MM:SS)**
00 /00	00	1:22:57
372 /600	03	5:03:08
621 /1000	05	8:49:55

The following table shows the amount of time to resolve the demand merge on the remaining node and return to steady state.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:08:51	00:16:36	00:23:54
372 /600	03	00:09:19	00:17:20	00:27:45
621 /1000	05	00:17:01	00:35:57	00:47:39

Test 4E1. Host-Based Minimerge (HBMM) Demand Merge: Two SSMs

The E and F tests are based on OpenVMS Volume Shadowing Host-Based Minimerge functionality. The HBMM policy creates a selected number of bitmaps on selected systems in the cluster. For this test, `SYS1` is the Active Instance and `SYS2` is the Passive Instance. A policy is assigned to each of the Shadow Set Volumes. The Active Instance manages all Minimerge recovery operations by virtue of the following HBMM policy declarations shown below.

Before the test started, a run was done with an extremely high threshold value, to help establish an appropriate reset threshold. We determined for our testing purposes that a reset value of 1,000,000 blocks was appropriate for our system. This value reflected approximately 50% of the writes done during a test run. Refer to http://h71000.www7.hp.com/news/hbmm.html for more information. A policy with that value was created and assigned to the Shadow Volumes using the following DCL commands:

```
$! Create a policy named RECOVERY_on_ACTIVE_NODES_DSAn
SET SHADOW /POLICY=HBMM=(master_list=(SYS1,SYS2),
reset_threshold=1000000) /NAME=RECOVERY_on_ACTIVE_NODES_DSAn

$! Associate a policy named HBMM_DSAn with DSAn:
SET SHADOW DSAn:/POLICY=HBMM=RECOVERY_on_ACTIVE_NODES_DSAn
```

Test 4E1 consists of a two-member Shadow Set Volume, with one disk local to each node and one disk remote to each node. Both members are 100% copied (in a Steady State). The test run was started and before completion, a demand merge was initiated with the following OpenVMS DCL command:

```
SET SHADOW/DEMAND_MERGE DSAn:
```

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**4E1 – Time for Completion (HH:MM:SS)**
00 /00	00	1:25:47
372 /600	03	5:07:05
621 /1000	05	8:34:04

The following table shows the amount of time for the demand merge to be completed and return to steady state.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:00:28	00:01:39	00:00:49
372 /600	03	00:00:07	00:00:05	00:00:05
621 /1000	05	00:00:29	00:01:36	00:00:47

Test 4E7. Host-Based Minimerge (HBMM) Forced Merge: Two SSMs

This test is similar to test 4E1 in all ways, except that instead of starting the merge by a DCL command, the merge is started by a forced crash of one of the nodes at some point after the start of the test run.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**4E7 – Time for Completion (HH:MM:SS)**
00 /00	00	1:23:54
372 /600	03	5:41:45
621 /1000	05	8:25:45

The following table shows the amount of time for the forced merge to be completed and return to steady state.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:09:02	00:18:21	00:17:36
372 /600	03	00:08:19	00:15:03	00:16:27
621 /1000	05	00:26:21	00:47:47	00:17:13

Test 4F1. Host-Based Minimerge (HBMM) Demand Merge: Three SSMs

This test is similar to test 4E1 in all ways, except that there are now two local full Shadow Set Members and one remote full Shadow Set Member, making a total of three disks instead of two. A demand merge is initiated by DCL command.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**4F1 – Time for Completion (HH:MM:SS)**
00 /00	00	1:28:35
372 /600	03	5:36:09
621 /1000	05	8:51:06

The following table shows the amount of time for the demand merge to be completed and return to steady state.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:00:48	00:01:30	00:00:29
372 /600	03	00:00:06	00:00:05	00:00:05
621 /1000	05	00:00:40	00:01:57	00:00:06

Test 6F5. Host-Based Minimerge (HBMM) Forced Merge: Three SSMs

This test is similar to test 4F1 in all ways, except that instead of starting the merge by a DCL command, the merge is initiated by a forced crash of one of the nodes at some point after the start of the test run. Also, there is one local full Shadow Set member and two remote full Shadow Set members.

The following table shows the amount of time to complete 600,000 transactions for the specific delays and distances.

Simulated distance (mi/km)	**Network Latency (ms)**	**6F5 – Time for Completion (HH:MM:SS)**
00 /00	00	1:23:31
372 /600	03	5:42:00
621 /1000	05	8:03:21

The following table shows the amount of time for the forced merge to be completed and return to steady state.

Simulated distance (mi/km)	**Network Latency (ms)**	**DSA10 – 5 GB – 28% full (HH:MM:SS)**	**DSA14 – 10 GB – 54% full (HH:MM:SS)**	**DSA20 – 5 GB – 14% full (HH:MM:SS)**
00 /00	00	00:00:13	00:00:20	00:00:21
372 /600	03	00:00:11	00:00:14	00:00:07
621 /1000	05	00:00:09	00:00:07	00:00:10

Appendix A

This appendix lists the selected hardware and software configurations for this proof of concept.

Hardware Configuration

This section lists the hardware configuration used for this proof of concept.

Servers

- 2 AlphaServer GS1280 (EV7, 1.5GHz) systems
- 8 CPU
- 8 GB memory

Storage

- EVA3000 and EVA8000 configured as RAID 1 volumes
- KGPSA disk fibre cards, PCI 133 MHz bus, 2 GB fibre speed
- DSGGB SANswitch connecting the disk fibre, SANswitch 2/16

Network

- LightSand i-8100B gateway
- Adtech AX/4000 network delay simulator[8]
- DEGXA Network Interface Card (NIC), PCI 133 MHz bus, 1 GB network speed
- Procurve 9308m network switch with EP J4895a 100/1000T modules and J4885a EP Mini-GBIC

Software Configuration

This section lists the software configuration used for this proof of concept.

Operating System - OpenVMS 7.3-2:

- HP Volume Shadowing for OpenVMS
- Latest system patches; HBMM patches are required
- TCPIP 5.4, ECO5

Database

Oracle9*i* RAC Version 9.2.0.5, running as active-active instances.

Load Generator

Swingbench 2.1f (for more information on Swingbench, see Appendix B).

Appendix B

This appendix describes the various vendors involved in this project as well as the software used for testing.

Digital Networks

Digital Networks has a rich history in providing network infrastructure for OpenVMS clusters in Disaster Tolerant environments. LightSand Communications Inc. has been an industry leader in SAN Over Distance products and technology. The recent technology partnership between Digital Networks and LightSand has generated the capability to support SAN extensions that will concurrently support OpenVMS Cluster infrastructure with HP Volume Shadowing for OpenVMS over the same physical wide area interconnect.

LightSand

LightSand Communications, Inc. is a company pioneering SAN connectivity and routing solutions. The LightSand S-8100B gateway used for testing provides cluster connectivity, IP connectivity, and Fibre Channel (FC) connectivity. In addition to FC SAN connectivity, the LightSand 8100 family switches support network connectivity for both Layer 2 Ethernet, including non-IP Cluster SCS traffic, and Layer 3 IP traffic over the same physical wide area interconnect.

Oracle RAC

Oracle9*i* Real Application Clusters (RAC) is an option to the Oracle9*i* Database Enterprise Edition, Release 2. It is a cluster database with a shared cache architecture that overcomes the limitations of traditional shared nothing and shared disk approaches to provide highly scalable and available database solutions. Oracle9*i* RAC allows large transactions to be separated into smaller ones for fast parallel execution, providing high throughput for large workloads. Through the introduction of a quorum disk, network failure and node failure are detected and resolved faster, resulting in faster completion of cluster reconfiguration. Lock remastering due to instance failure and instance recovery are concurrent. Failover capability is consolidated and enhanced to provide more robust and generic solutions. Oracle9*i* RAC can now function as a failover cluster with active instances on all nodes. It does require supporting clusterware software from the operating system that manages the cluster.

[8] This component was used for the test environment and would not be needed in a production environment.

Swingbench

Swingbench 2.1f was used as a load generator with typical order-entry functions[9] of 300 remote clients and 600,000 transactions.

Swingbench is an extensible database benchmarking harness designed to stress test Oracle databases via Java Database Connectivity (JDBC). It consists of a load generator, a coordinator, and a cluster overview. The software enables a load to be generated and the transaction response times are recorded. It was written internally by Oracle developers, primarily to demonstrate Real Application Clusters, but can also be used to demonstrate functionality such as online table rebuilds, standby databases, and online backup and recovery. It is not an official Oracle product, but is available for general use at no cost to the end user.

The code that ships with Swingbench includes two benchmarks: OrderEntry and CallingCircle. The testload used for this project is the OrderEntry. It is based on the oe schema that ships with Oracle9*i* Database and Oracle Database 10*g*. It has been modified so that the Spatial, Intermedia, and Oracle9*i* schemas do not need to be installed. It can be run continuously, that is, until you run out of space. It introduces heavy contention on a small number of tables and is designed to stress interconnects and memory. Both benchmarks are heavily CPU-intensive. The entire framework is developed in Java and as a result can be run on a wide variety of platforms. It also provides a simple API to allow developers to build their own benchmarks.

Glossary

600K	600,000 'typical' Order Entry transactions generated by Swingbench, the Oracle informal load-generating software (http://www.dominicgiles.com/swingbench.php). (This is the software that will be used to generate the I/O traffic that will be monitored in this project.)
Active	Clients are connected to that node.
Active-Active	A description of a type of Oracle RAC operational configuration composed of two nearly identical infrastructures logically sitting side by side. In this type, one node acts as a primary to a database instance and another one acts as a secondary node for failover purposes. At the same time, the secondary node acts as the primary for another instance and the primary node acts as the backup and secondary node. Clients typically connect in a distributed or round-robin fashion to both nodes.
Active-Passive	A description of a type of Oracle RAC operational configuration composed of two nearly identical infrastructures logically sitting side by side. One node hosts the dataset service or application, to which all clients connect, while the other rests idly waiting in case the primary system goes down. Upon failure, the primary server gracefully turns over control of the database and application to the other server or node who in turn becomes the primary server.
Copy State	Duplicate data on a source disk to a target disk. At the end of a copy operation, both disks contain identical information. Copy can be a full copy or a minicopy. Only Full Members and Merge Members can service user read requests.

[9] Typical order-entry functions are: new customers, product order, process order, browse of order, and browse for products.

Demand Merge	Merge process initiated by the DCL command `SET SHADOW/DEMAND_MERGE`.
Full SSM	SSM that is a full shadow set member, responding to user read and write I/O.
Forced Merge	Merge process initiated by system failure or crash.
HBMM	Host-Based Minimerge
HBVS	Host-Based Volume Shadowing
Local FCPY SSM	SSM that is a copy shadow set member, responding to user write I/O. SSM requires a full copy operation to make it a full member.
Local Full SSM	SSM that is a full shadow set member, responding to user read and write I/O. SSM is at the same site that the Active Oracle Instance is running at.
Local Full Merge SSM	SSM that is one of several merge shadow set members, responding to user read I/O using merge semantics and requiring full merge operation to transition VU to a steady state.
Local MCPY SSM	SSM that is a copy shadow set member, responding to user write I/O. SSM requires a minicopy operation to make it a full member, per master bitmap. Logging of all write I/O to shadow sets with bitmaps use messages that generate SCS traffic.
Local Minimerge SSM	SSM that is one of several merge shadow set members responding to user read I/O using merge semantics, which requires a Minimerge operation to transition VU to a steady state.
Merge State	Compare data on shadow set members and to ensure that inconsistencies are resolved. Merge can be a Full Merge or a Minimerge.
NIST	OpenSource software available from the National Institute of Standards and Technology that introduces a delay in transmitting data across a network in order to simulate long distances between hardware.
OpenVMS Server	Cluster member that serves directly connected devices to other cluster members.

Oracle RAC failover	The ability to resume work on an alternate instance upon instance failure.
Oracle TAF	Run-time failover that enables client applications to automatically reconnect to the database if the connection fails
Passive	No clients are connected to that node.
Remote FCPY SSM	SSM that is a copy shadow set member responding to user write I/O. SSM requires a full copy operation to make it a full member. Remote SSM is at a site that is not physically local to the Local node. In our testing, the remote site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Remote Full Merge SSM	SSM that is one of several merge shadow set members responding to user read I/O. Using merge semantics requires full merge operation to transition VU to a steady state. Remote SSM is at a site that is not physically local to the Local node. In our testing, the remote site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Remote Full SSM	SSM that is a full shadow set member responding to user read and write I/O. Remote SSM is at a site that is not physically local to the Local node. In our testing, the remote site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Remote MCPY SSM	SSM that is a copy shadow set members responding to user write I/O. SSM requires a minicopy operation to make it a full member. Remote SSM is at a site that is not physically local to the Local node. In our testing, the remote site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Remote Minimerge SSM	SSM that is one of several merge shadow sets responding to user read I/O. Using merge semantics requires a Minimerge operation to transition VU to a steady state. Remote SSM is at a site that is not physically local to the Local node. In our testing, the remote site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
RPO	Recovery Point Objective. The maximum acceptable data loss in the event of a system outage.
RTO	Recovery Time Objective. The maximum acceptable time between a system outage and the continuation of operations.

Served FCPY SSM	SSM that is a copy shadow set member responding to user write I/O. SSM requires a full copy operation to make it a full member. Served SSM is at a site that is not physically local to the Local node. In our testing, the served site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site. SSM is served to this system by another OpenVMS system in this cluster via a WAN.
Served Full Merge SSM	SSM that is one of several merge shadow set members responding to user read I/O. Using merge semantics requires a full merge operation to transition VU to a steady state. Served SSM is at a site that is not physically local to the Local node. In our testing, the served site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Served Full SSM	SSM that is a full shadow set member responding to user read and write I/O. Served SSM is at a site that is not physically local to the Local node. In our testing, the served site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Served MCPY SSM	SSM that is a copy shadow set member responding to user write I/O. SSM requires a minicopy operation to make it a full member. Served SSM is at a site that is not physically local to the Local node. In our testing, the served site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Served Minimerge SSM	SSM that is one of several merge shadow set members responding to user read I/O. Using merge semantics requires a Minimerge operation to transition VU to a steady state. Served SSM is at a site that is not physically local to the Local node. In our testing, the served site was simulated (via network delay) at distances of 600 km and 1000 km away from the local site.
Site ID	A `sysgen` value that volume shadowing uses to determine the best device to perform read I/O operations, thereby improving applications performance. This nonzero value indicates to the shadowing driver the site location of the specified shadow set or virtual unit (`DSAnnnn`). A value of zero is not considered valid.
SSM	Shadow Set Member. A SSM can either be a Full Member (no copy or merge in progress) or a Copy Member (copy in progress) or a Merge Member (merge in progress).
Steady State	A VU which has no SSMs in either a Merge or a Copy state.
T4	Timeline data gathering tool that makes use of OpenVMS MONITOR and creates `.CSV` files, which contain the MONITOR data gathered via batch jobs. This data can then be displayed using TLVIZ or made use of by any program that can read a `.CSV` file. For more information, refer

	to: http://h71000.www7.hp.com/openvms/products/t4/index.html
TLVIZ	Timeline Visualization tool that displays T4 data in graphical format.
VU	Virtual Unit. Represents the physical devices that make up the mounted Shadow Set.

For more information

- For questions, contact:
  openvms-info@hp.com and put DT RAC POC on the subject line.
- For Digital Networks, refer to:
  www.digitalnetworks.net
- For LightSand, refer to:
  www.lightsand.com
- For Oracle RAC, refer to:
  www.oracle.com
- For Swingbench, refer to:
  http://www.dominicgiles.com/swingbench.php