



Extended Distance RAC

Eliminating the current physical restriction of Oracle Real Application Cluster

Mai Cutler
HP Oracle Development Alliance Lab

Sandy Gruver
HP Oracle Advanced Technology Center

Stefan Pommerenk
Oracle Cluster and Parallel storage Technology



- Why Extended Serviceguard RAC cluster
- Review of RAC architecture
- What is DWDM and Dark Fiber?
- Architecture used for the tests
- Description of tests and results

Why Extended Serviceguard RAC cluster?

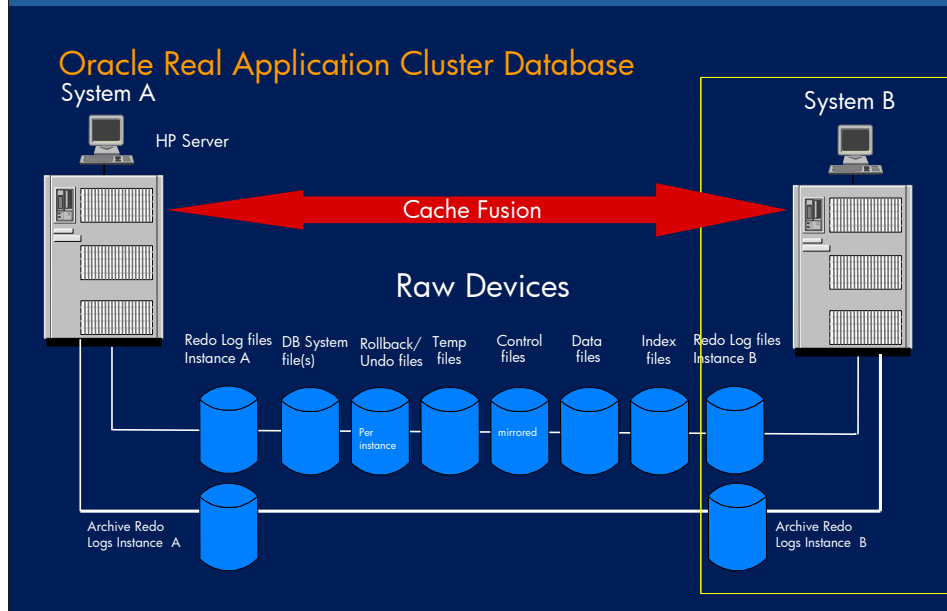
Why have customers asked for this solution



A need for a single RAC solution which can span an extended distance (up to 100km)

This solution:

- Is implemented as a single RAC database spanning two geographically distant sites
- Is simpler than maintaining a standby
 - Disaster tolerance
- Ensures no data loss in case of a failure at one site
- Combines HA with DT and fully utilizes all resources



Just a review of Oracle single instance and Real application cluster multiple instances.

With a single instance Oracle (non-RAC), Oracle manages a single set of data for multiple users.

Integral to the performance of Oracle is the management of a large shared memory area. Since accesses to memory take nanoseconds while accesses to disk take milliseconds, the performance of any operation is directly proportional to the location of the data - whether in shared memory or on disk.

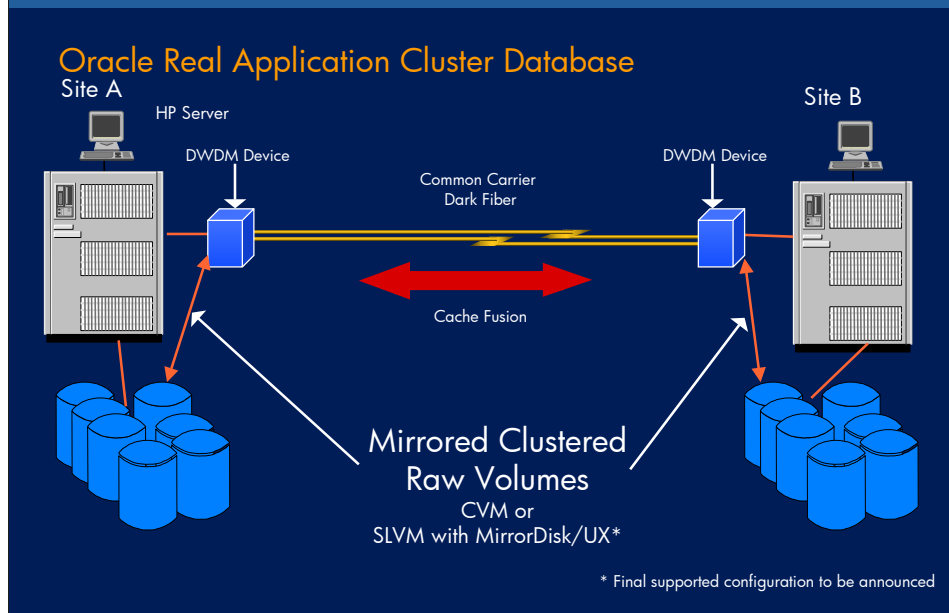
With RAC ... As before there is one copy of data ...

(mouse click) As before there is one copy of data ... that is shared by two or more Oracle instances on two or more servers.

(mouse click) again data can be managed as raw or as clustered file system data.

Since data locality is so important to Oracle performance, Oracle introduced full cache fusion with Oracle 9i RAC implementing high speed memory to memory data passage - (mouse click)

Review of Real Application Clusters on HP-UX Extended Serviceguard clustered database



Typically, Real Application Cluster databases share a single set of storage and are located on servers in the same data center.

What we're here to review are the Extended ServiceGuard and Oracle RAC tests, done jointly by HP and Oracle, using disk mirroring and DWDM technology to extend the reach of the cluster by up to 100km.

This configuration allows two data centers, separated by up to 100km, that are connected using DWDM (dense wave division multiplexing) equipment and dark fiber (mouse click).

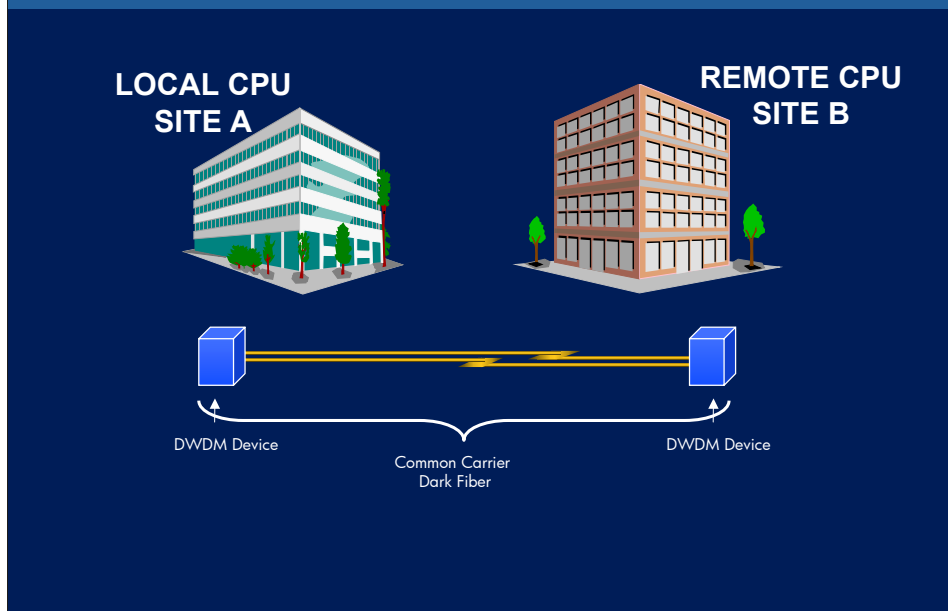
Each data center has its own set of storage which is synchronously mirrored using either CVM mirroring or SLVM with MirrorDiskUX (mouse click). All traffic between the two sites is sent thru the DWDM and carried on dark fiber. This includes mirrored writes, network and heartbeat traffic, and

(mouse click) memory to memory data passage.

HP and Oracle wanted to test this solution to discover the performance differences between a RAC installation at a single data center and one in an extended configuration.

What is DWDM and Dark Fiber

Dark Fiber with DWDM



DWDM is optoelectronic technology

Can simultaneously transmit multiple separate optical signals through a single optical fiber thinner than a human hair

Similar to radio or TV broadcasting where each station broadcasts on a different frequency

The maximum distance allowed between a DWDM device pair depends on the particular DWDM vendor product used, but can reach distances as high as the 100-120 km range.

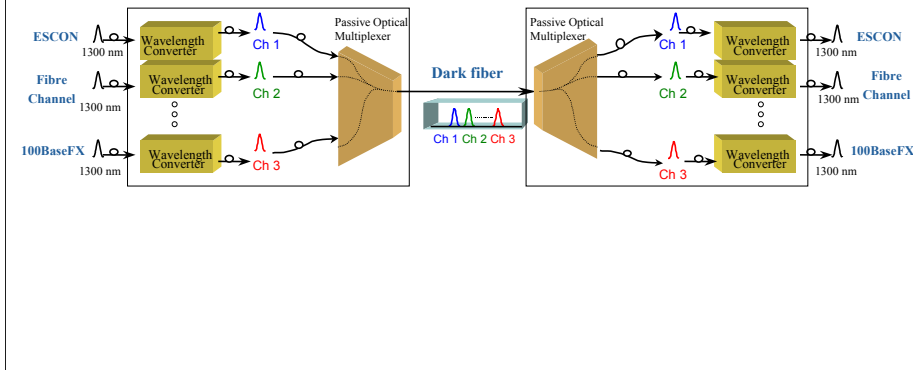
Dark Fiber is a single fiber optic cable or strand, thinner than a human hair

- sold by telecom provider
- has no light or signal (is “dark”) until customer puts a signal on it.

DWDM (Dense Wavelength Division Multiplexing)

Wavelength Division Multiplexing is a technology that uses multiple lasers and transmits several wavelengths of light (lamdas) simultaneously over a single optical fiber. Each signal travels within its unique color band, which is modulated by the data (text, voice, video, etc.). WDM enables the existing fiber infrastructure of the telephone companies and other carriers to be dramatically increased. DWDM enables a single optical fiber to simultaneously carry multiple traffic-bearing signals, thereby increasing the capacity of a fiber many times over. DWDM systems can support more than 150 wavelengths, each carrying up to 10 Gbps. Such systems provide more than a terabit per second of data transmission on one optical strand, thinner than a human hair.

Diagram of DWDM and Dark Fiber



- The DWDM device multiplexes several of these converted optical inputs over the same fiber optic cable
- The destination DWDM device reverses the process. It de-multiplexes the signals and converts them back to their original wavelength.
- Dense Wave Division Multiplexing (DWDM) technology provides the capability to extend the Fibre Channel link distance.
- Network and disk communication share DWDM links.
- We're presenting our findings to date for ServiceGuard Extended Distance Clusters with RAC at distances up to 100 km using DWDM technology. Since this work is currently under investigation, we ask that you contact HP and Oracle for the supported configurations.

What has been tested

Extended Serviceguard Cluster with DWDM



- Previously tested and certified

- * Extended Distance Serviceguard with
 - * LVM with MirrorDisk/UX or
 - * VxVM with mirroring
 - * (supported up to 100 km)
- * Quorum Server, arbitrator nodes or dual cluster locks for arbitration
- * Extended Distance Serviceguard Extension for RAC (SGeRAC) cluster configurations of 2 or 4 nodes using Veritas CVM, for distances of up to 10 km

- Current tests

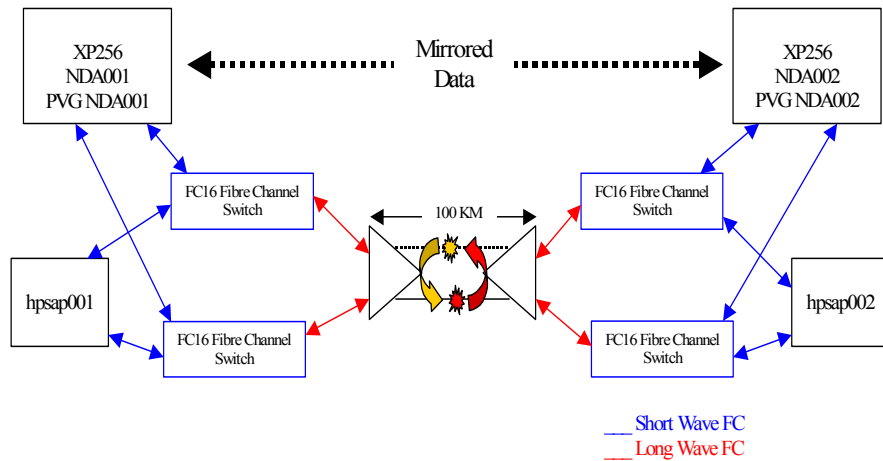
- * Extended Distance SGeRAC cluster configurations of 2 nodes using Veritas CVM, for distances of up to 100 km
- * Follow-on tests with SLVM and Mirrordisk/UX

Failure tests

- Host failure
- Storage device failure
- Inter-site link failure
- Data center failure

What has been tested

DWDM - Fibre Channel Testing

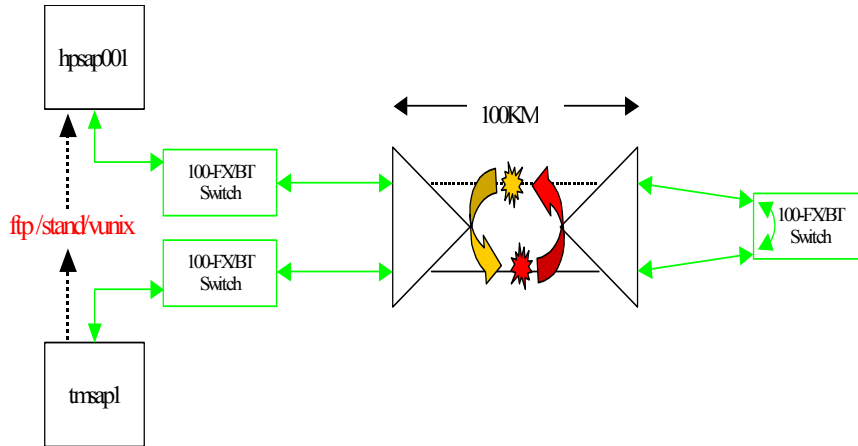


- Fibre Channel links operated properly through link failure scenarios
- Tests were run during data transfer and data integrity was maintained

- The test methodology for checking Fibre Channel over DWDM consisted of creating a Volume Group from the LDEVs on both nda001 and nda002 and importing this VG to both test nodes. Once this was completed, a series of manual tests were conducted which moved data to the VG and tested the integrity of the data after a series of different LVM operations and failure scenarios were executed.

What has been tested

DWDM - 100BaseFX Testing



- Network links operated properly through various link failure scenarios
- Tests were run during data transfer and data integrity was maintained

These test scenarios consisted of testing 100BaseT/100BaseFX connectivity over DWDM by setting up a network between two systems in the test environment and running this network (100BaseFX) over the DWDM link.

What has been tested

Disaster Tolerant Cluster Solutions for Oracle on HPUX



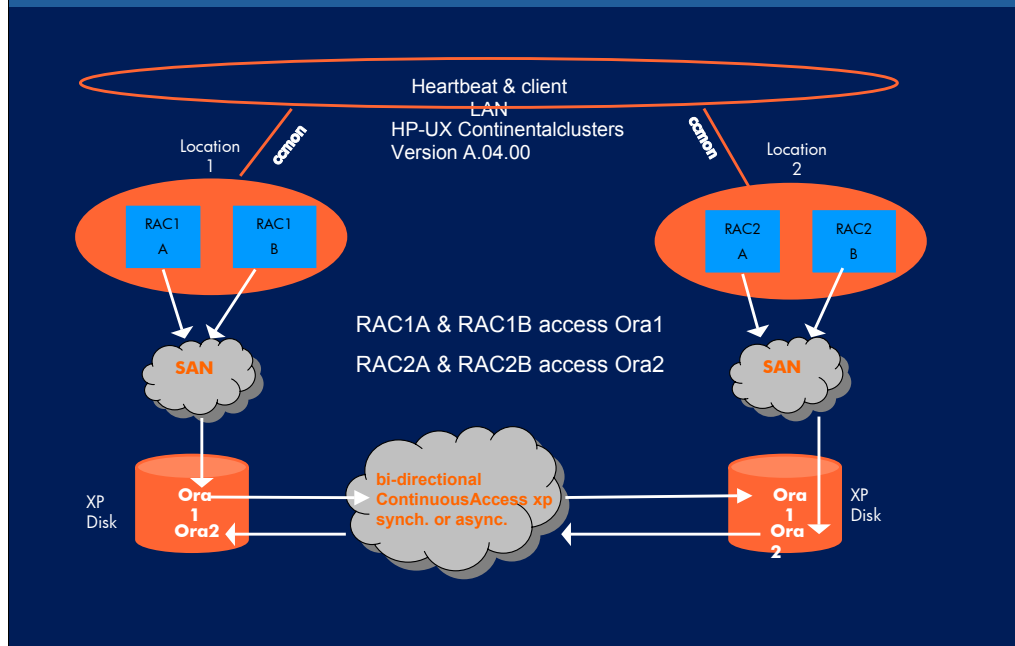
HP-UX 11.11 (PA-RISC) Oracle 9i RAC and Disaster Tolerance						
	Disaster Tolerant RAC Solutions			Extended RAC		
Topology	SGeRAC		Veritas DBeAC	SGeRAC with Continentalclusters	SGeRAC Stretched 2 arrays	Extended SGeRAC 2 arrays*
Distance	Local DC		Local DC	Unlimited	10 kms	100kms
Volume Manager	SLVM	Veritas CVM	Veritas CVM	SLVM	CVM	SLVM CVM
CFS	-	-	VCFS	-	-	-
SW Mirroring	MirrorDisk UX	CVM with VxVM Mirroring	TBD	N/A	CVM with VxVM Mirroring	MirrorDisk UX CVM Mirroring
HW Mirroring	N/A	N/A	N/A	CA XP	N/A	N/A
Storage	VA, XP, EMC, EVA	VA, XP, EMC	XP	XP	VA, XP, EMC	TBD
# Nodes	16	4	TBD	16 per cluster / total of 32 nodes	2 or 4	2
Bi-directional Failover	N/A	N/A	N/A	Yes	Yes	Yes

* Final supported configuration to be announced

This chart details the various solutions that have been tested and certified with ServiceGuard and Oracle.

Disaster tolerant solutions

2 RAC sites protected with Continentalclusters and CA



What is HP ContinentalClusters

A ContinentalClusters provides a disaster tolerant solution in which distinct ServiceGuard clusters are separated by large distances, with wide area networking used between them.

It basically works with any data replication mechanism - software based or disk array based. HP ContinentalClusters provides a pre-integrated data replication solution using high performance data replication mechanisms HP Continuous Access XP or EMC SRDF, and also offers an integration tool for using Oracle Standby Database.

ContinentalClusters A.04.00 that runs on HP-UX 11.11 currently supports Oracle10g RAC instances in the environment only using SLVM for volume management and XP CA for data replication - VxVM/CVM volume management and EMC SRDF data replication are not currently supported in this configuration.

For Oracle10g RAC, this means that we have one Oracle9i RAC cluster running in data center 1, called the primary cluster. If this primary cluster fails, then all Oracle instances will be restarted by ContinentalClusters in data center 2. This cluster at the data center 2 is also referred to as recovery cluster. This failover between the two clusters requires a manual issue of a single command line command (cmrecoverl).

Extended Serviceguard RAC cluster Test Configuration



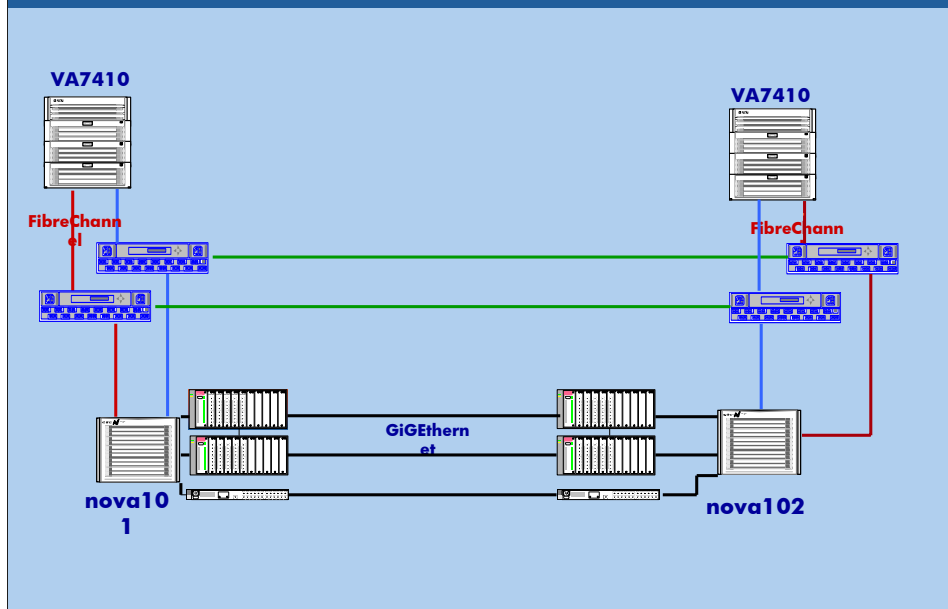
- 2 rp7410s (8 CPUs, 32GB RAM)
- 2 Tachyon XL2 FibreChannel HBAs.
- 3 Fibre-based GB Ethernet interfaces
- 4 ULTRA SCSI internal disks(73GB)

- HP-UX B.11.11 and September 2002 bundle
- ServiceGuard OPS Edition A.11.14
- Veritas VxVM 3.5
- Veritas CVM 3.5
- Oracle 9.2.0.3 with RAC option, UDP/IPC

The configuration shows a solution where the cluster components reside in separate data centers. The dual cluster lock disks are required – one at each center – to ensure recovery from an entire data center failure.

Architecture used for the tests

Local Serviceguard Cluster with RAC



The project starts with a local cluster of 2 nodes. The shared 9i database is provided by 2 identical VA7410s. The database is comprised of data files residing on one VA7410. Redundant (mirror) copies of the datafiles are created with host-based mirroring functionality of Veritas VxVM. The mirror datafiles reside on similar LUN/diskgroup configuration on the 2nd VA7410. Other mirroring methods are available for different storage systems – such as Continuous Access for XP or MirrorDisk/UX if SLVM is used.

VA7410 (HP StorageWorks Virtual Array) information – A6218A, 2GB cache/controller.

_ 4 DS2405 with 15*36GB at 15K rpm.

SAN is a key component of an Extended ServiceGuard cluster. In this configuration, 4 SAN switches are used to provide the interswitch links between the 2 VA7410s as well as for the servers. 2 switches are used at each side to eliminate SPOF. The servers are connected to the ISL switches in the way which allow both servers see and have access to the 2 storage systems at the same time.

ISL Information: fc 1/2GB 16B, 2.125 Gb/sec line, full duplex.

64Gb/sec end-to-end and 2112 byte/payload.

Three GigEthernet networks are used: 1 for Heartbeat and 2 for cluster interconnects.

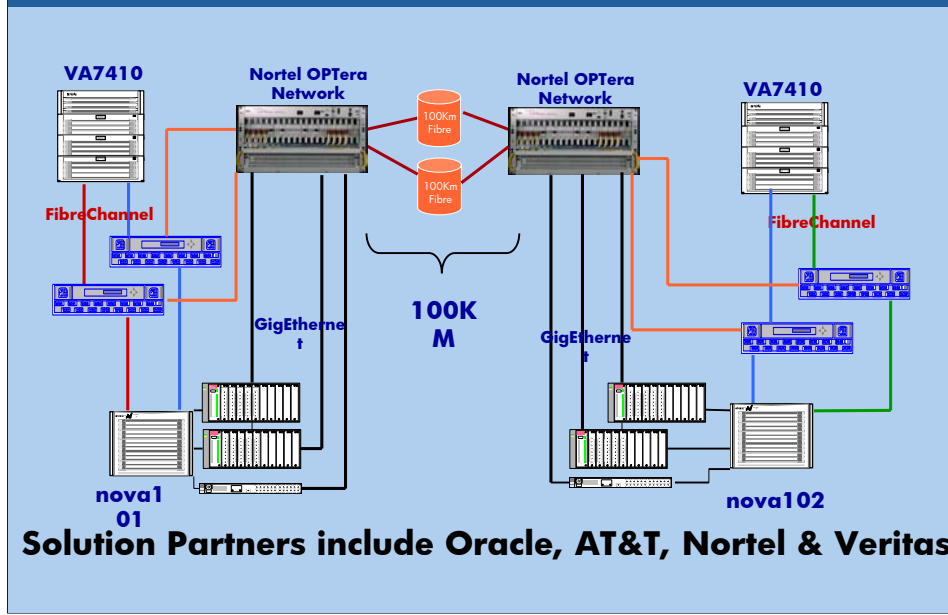
Network switch information: Procurve 8000M

Both network and storage connections to the servers are optical fibre-based to allow direct connections to the DWDM switch without the need for a copper-to-fiber signal converter. DWDM can handle different wavelengths: shortwave (780/850nm and 50/125 or 62.5/125 microns multimode fibers) longwave (1330nm, 9/125 micron fibers). We used shortwave (SC) for GigE connection and shortwave (LC) connection for the Fibre-connection our setup.

The configuration shows a solution where the cluster components reside in separate data centers. The dual cluster lock disks are required – one at each center – to ensure recovery from an entire data center failure.

Architecture used for the tests

Extended Serviceguard Cluster with RAC over DWDM



Once the baseline data was collected from the local setup, the Nortel OPTera Network switches and fibers Provided by ATT were put in place which connect the 2 sites. The fibre distance of 100km was tested first Then reduced to 50km and finally to 25km.

HP does not require any particular vendor's DWDM equipment to be used. The customer is responsible for the selection and maintenance of any DWDM equipments. HP can provide a list of tested but not certified DWDM vendor's equipment. In our setup, we NORTEL OPTera DWDM switches along with ATT-supplied fibers.

DWDM (Dense Wavelength Division Multiplexing)

Wavelength Division Multiplexing is a technology that uses multiple lasers and transmits several wavelengths of light (lamdas) simultaneously over a single optical fiber. Each signal travels within its unique color band, which is modulated by the data (text, voice, video, etc.). WDM enables the existing fiber infrastructure of the telephone companies and other carriers to be dramatically increased. DWDM enables a single optical fiber to simultaneously carry multiple traffic-bearing signals, thereby increasing the capacity of a fiber many times over. DWDM systems can support more than 150 wavelengths, each carrying up to 10 Gbps. Such systems provide more than a terabit per second of data transmission on one optical strand, thinner than a human hair.

Description of tests

Test Scenarios



- IPC tests
- I/O tests
- OLTP-like workload tests
- Failover Tests

1- CRTEST definition

CRTEST is a micro-level performance benchmark for measuring raw IPC throughput. The test first updates a set of blocks in a hash cluster table on instance 1, and then an increasing number of clients running SELECT are started on instance 2. These queries cause a message to be sent from instance 2 to instance 1 and instance 1 will return a CR block. CR fairness down converts were disabled for this test, so this whole test just nothing but 'send a message' and 'receive a CR block' from the client's perspective.

The following extract from a Statspack report nicely illustrates the workload, at a 1-minute snapshot with a CR block receive rate of 12763 blocks/sec.:

STATSPACK report for

Instance Activity Stats for DB: TPCC Instance: tpcc2 Snaps: 388 -389

Statistic	Total	per Second	per Trans
execute count	791,400	12,764.5	395,700.0
gcs messages sent	791,347	12,763.7	395,673.5
global cache cr blocks received	791,350	12,763.7	395,675.0

The CRTEST (IPC) tests were done using one GigEthernet for cluster interconnect (1IC), then with two interconnect (2IC).

The modified TPCC workload was used to simulate OLTP application on this configuration. Only one interconnect was used.

2- I/O Test description

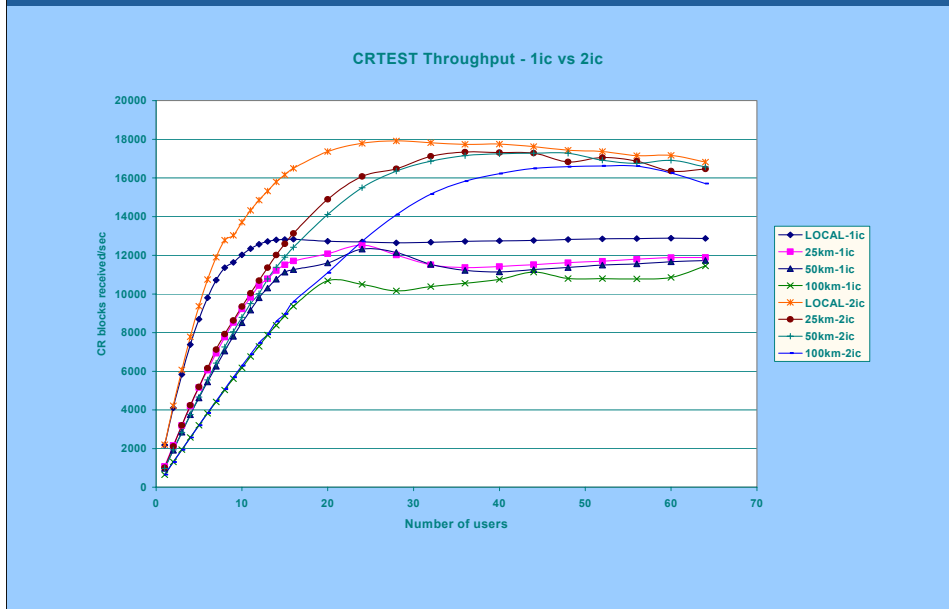
The I/O tests were done using Diskbench(DB). It is a disk subsystem performance measurement tool. DB can measure properties like IOs/sec and Mb/sec for different disk subsystems. The tests were using a mixed of 60% read and 40% write. In this particular test, only a single device was used for the Read/Write test. To be more representative of the performance in the actual - other variations could be used such as: all devices in the array or the file systems. In our limited time, we want to see the effect of the distance with respect to i/o in the simplest setting.

3- Failover tests

During failover tests, storage connections either via the ISL switch or directly to the VA were removed one at a time. We observed the availability of the Oracle instances as well as the resynchronization of the data volumes as the reconnections are restored.

4- TPC-C test: a modified TPCC workload was used to simulate OLTP transactions over the cluster interconnect as well as over DWDM network.

Test results – CRTEST throughput



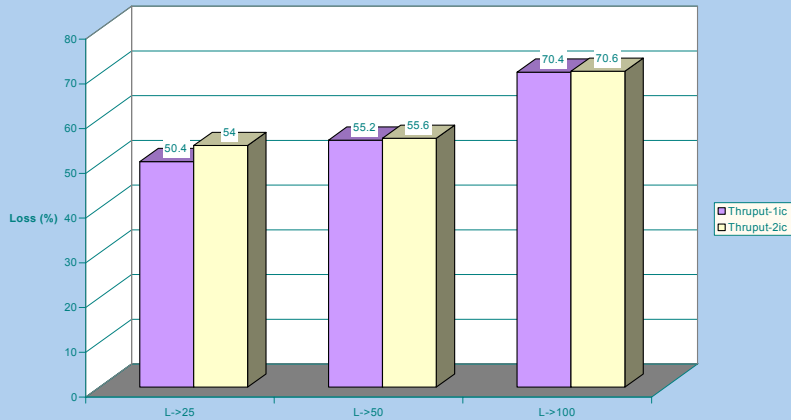
The tests shown that the usage of 2 interconnects is beneficial particularly at higher load for both LOCAL and REMOTE cases. The impact of the distance causes the drop in the throughput at all distances. The drop, however, is more than compensated when additional interconnect was added. The tables below show the percentage changes at lowest and highest loads for throughput.

Although the addition of the 2nd interconnect does not provide 2X throughput, it does increase it by 40-45% gain. It is apparent that the highest load the throughput comparison looks a lot better than at low load. This could be due to the resources are not fully utilized at low load, to compensate for the additional overhead introduced by the distance at this load level.

Test results – CRTEST throughput

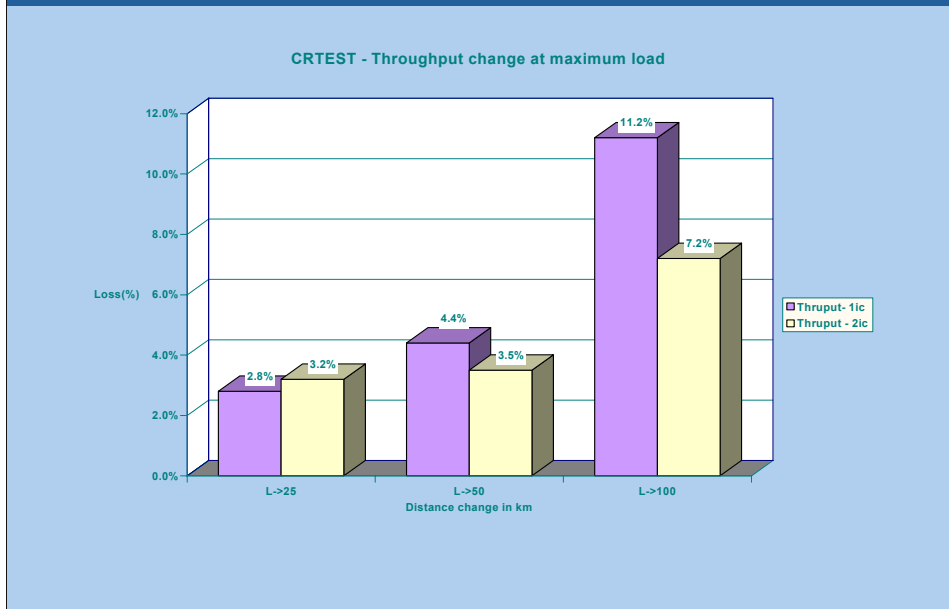


CRTEST - Throughput change at lowest load



At low load of one user, the throughput change is quite substantial at 50%-70% reduction in the number of blocks received. The percentage change does not differ much with the addition of the second interconnect.

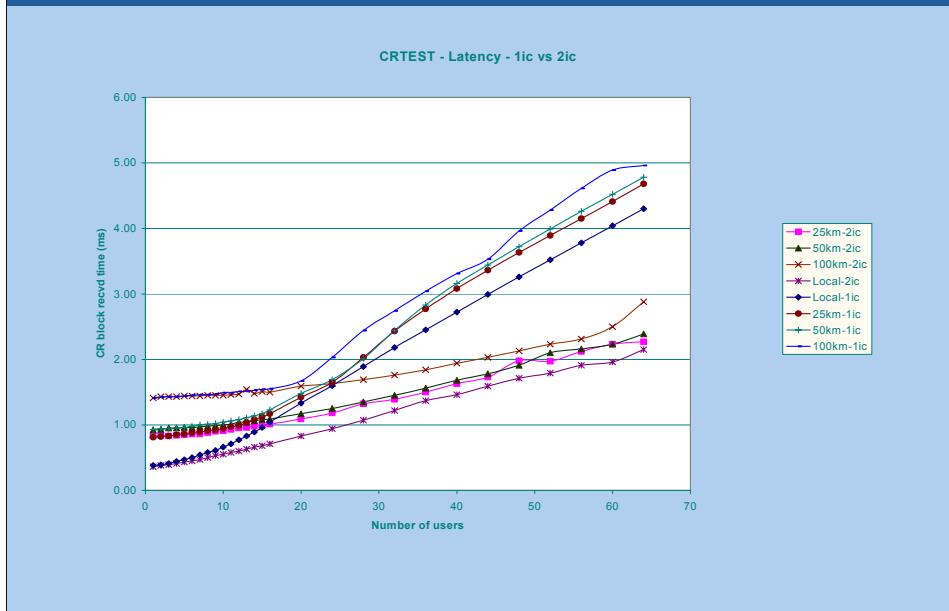
At the higher load – maximum at 64 users, the loss is more reasonable from 3 thru 11%. The change can be said insignificant with the usage of 2 interconnects.



As expected, the latencies degrade as the distance increases. In general, the addition of the second interconnect improves the latency at distances. At maximum load, the latency with 2 interconnects is 42-51% lower than with one interconnect.

Although, the usage of the second interconnect consistently outperforms a single interconnect configuration, the latency worsen over longer distance at 50km or higher when compare with the latency in a local with same number of interconnects.

Test results – CRTEST - IPC latency



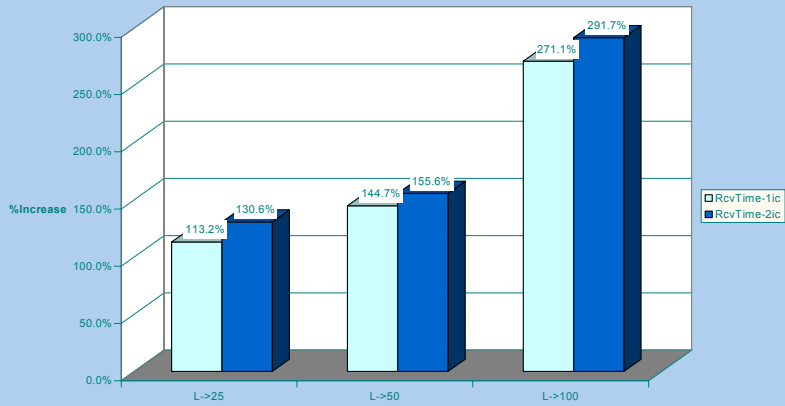
As expected, the latencies degrade as the distance increases. In general, the addition of the second interconnect improves the latency at distances. At maximum load, the latency with 2 interconnects is 42-51% lower than with one interconnect.

Although, the usage of the second interconnect consistently outperforms a single interconnect configuration, the latency worsen over longer distance at 50km or higher when compare with the latency in a local with same number of interconnects.

Test results – CRTEST - IPC latency

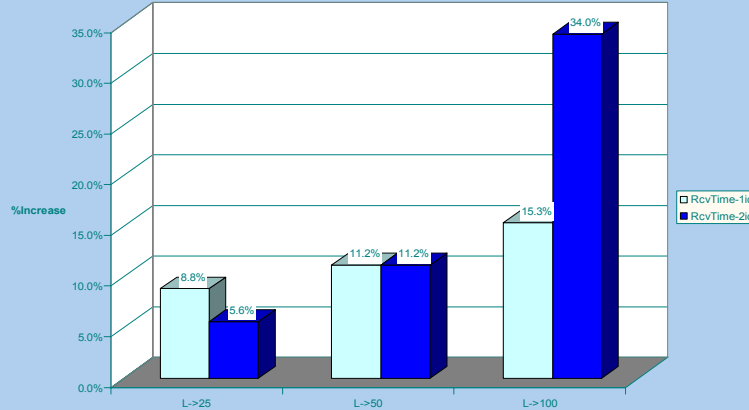


CRTEST-Latency change at lowest load



At lowest load with one user, the latency increases by almost 300% going from Local to 100km. The latency does not differ much with the addition of the 2nd interconnect.

CRTEST-Latency change at highest load



As expected, the latencies degrade as the distance increases. In general, the addition of the second interconnect improves the latency at distances. At maximum load, the latency with 2 interconnects is 42-51% lower than with one interconnect.

Although, the usage of the second interconnect consistently outperforms a single interconnect configuration, the latency worsen over longer distance at 50km or higher when compare with the latency in a local with same number of interconnects.

Test results – IPC latency

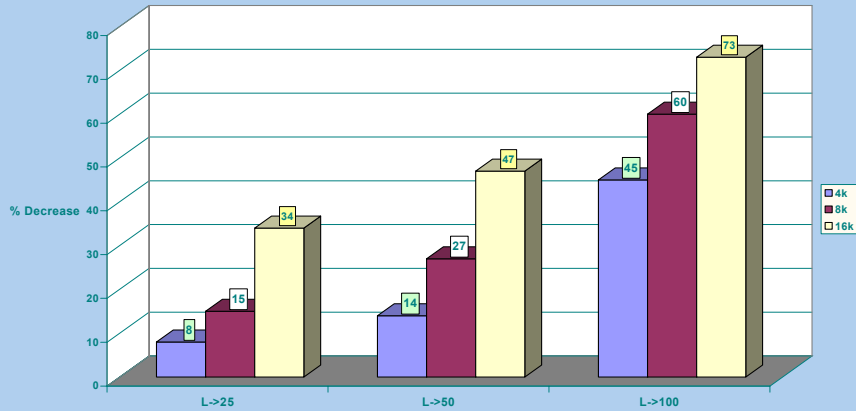


- IPC latency at lowest and highest load with 1 and 2 cluster interconnects.

	Local 1ic	Local 2ic	25km 1ic	25km 2ic	50km 1ic	50km 2ic	100k m 1ic	100km 2ic
CR time (ms)	0.38 4.30	0.36 2.15	0.81 4.68	0.83 2.27	0.92 4.78	0.93 2.39	1.41 4.96	1.41 2.88

Standard propagation delay over DWDM is 4.5 microsecond/km. Over 100km, we expect to add 0.45 ms to the latency. The DWDM protocol requires the conversion of the optical signal at each end of the fiber, therefore the additional overhead.

I/O bandwidth change vs Distance vs Blocksize

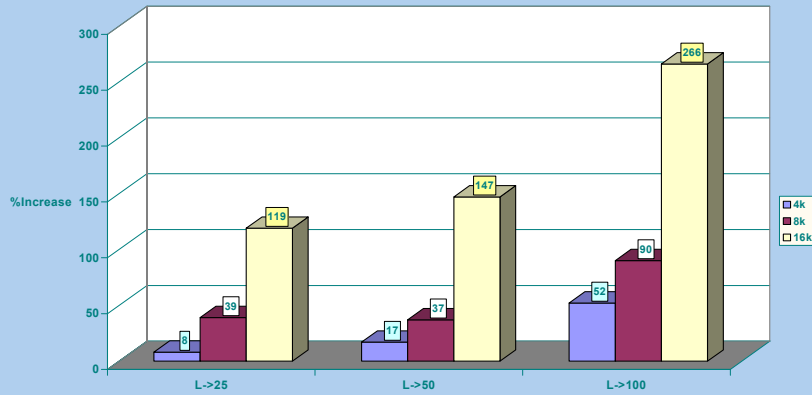


In the I/O tests, the bandwidth and the latency worsen over distances as expected. The interesting observation is the degradation is more severe with larger block sizes, as shown in the graph for 4k, 8k and 16k.

Test results – I/O tests

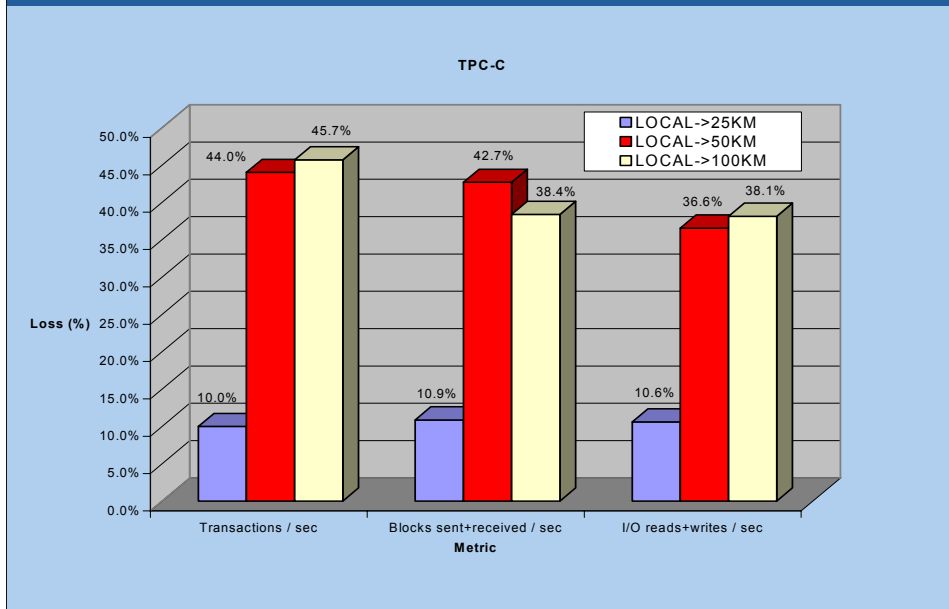


I/O Latency vs Distance vs Blocksize



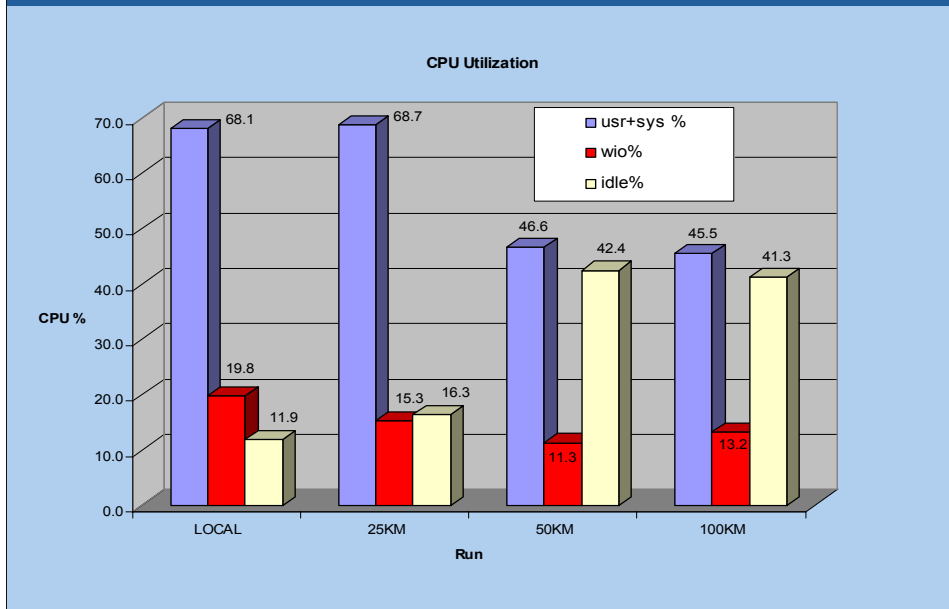
Similar degradation in latency is observed – bigger block size, worse degradation.

Test results – application test



The chart shows the loss in the key performance metrics of the modified TPCC workload: Transaction/sec (TPS), Number of Blocks Sent/Received/sec as well as the number of I/O Reads/Writes per sec. At 25km, the loss is about 10% for each of the metric. At 50 and 100km, the performance metrics degrade to about 50-60%.

Test results – application test



The chart shows the CPU utilization during the application tests. At shorter distances (local and 25km) systems are about 70% busy. At longer distances, due to contentions for other resources – IPC and I/O, the systems stay idle much more – going from 11% to 40% at 100km.

- Host failure
- Storage Device failure
- Inter-site link failure
- DWDM link failure

We repeated the same failure tests which were done earlier.

- Host failure – a server was shutdown. Oracle database became inaccessible for a couple minutes ,stayed up and resumed be accessible
- Storage device failure – Oracle instances became inaccessible for a couple minutes then resumed to be accessible. Devices on disconnected shown as “NO DEVICE”.
- Inter-Site Link failure –
- DWDM link failure – The cluster continued to stay up, ServiceGuard or RAC noted no difference in the configuration.

- Solutions display HA characteristics of traditional RAC /Serviceguard cluster – Oracle and systems continue to be available if a component fails.
- Results in throughput and latency over distances are as expected – better performance at shorter distances.
- Usage of multiple cluster interconnects is beneficial to improve the overall performance.
- Assessment of your application with respect to high availability and performance is important.
 - Performance results differ with applications characteristics.*

* performance results will vary depending on your particular configuration and application

The Extended RAC solution over DWDM tests show the HA characteristics of SG cluster – allowing the server and RAC continue to run/accessible when failure occurs on one of the components.

The overall performance degrades as expected over distances. The results of the basic tests (IO and IPC) shown that at 25km the performance degrades about 8% and 3% respectively. The application tests in the same environment shown a 10% overall degradation.

Similarly, at 50km, the basic I/O and IPC tests shown 11% and 17% degradation respectively, whereas the application tests shown 42% degradation.

One can expect that if both IPC and I/O tests yield reasonable performance change, the application would perform in an acceptable manner at the same distance.

Additionally, the test results shown that assessment of the application characteristics/workload in a local configuration is critical prior to implementation over extended distances. The application performance differ with its design and workload. Tuning and consideration of hardware components would improve the application performance at longer distances. Lastly, the trade-off between performance and high availability is key element in the design decision.

What has been tested

Disaster Tolerant Cluster Solutions for Oracle on HPUX



HP-UX 11.11 (PA-RISC) Oracle 9i RAC and Disaster Tolerance						
	Disaster Tolerant RAC Solutions			Extended RAC		
Topology	SGeRAC		Veritas DBBeAC	SGeRAC with Continentalclusters	SGeRAC Stretched 2 arrays	Extended SGeRAC 2 arrays *
Distance	Local DC		Local DC	Unlimited	10 kms	100kms
Volume Manager	SLVM	Veritas CVM	Veritas CVM	SLVM	CVM	SLVM CVM
CFS	-	-	VCFS	-	-	-
SW Mirroring	MirrorDisk UX	CVM with VxVM Mirroring	TBD	N/A	CVM with VxVM Mirroring	MirrorDisk UX CVM Mirroring
HW Mirroring	N/A	N/A	N/A	CA XP	N/A	N/A
Storage	VA, XP, EMC, EVA	VA, XP, EMC	XP	XP	VA, XP, EMC	TBD
# Nodes	16	4	TBD	16 per cluster / total of 32 nodes	2 or 4	2
Bi-directional Failover	N/A	N/A	N/A	Yes	Yes	Yes

* Final supported configuration to be announced

This chart details the various solutions that have been tested and certified with ServiceGuard and Oracle.

- Product literature, whitepapers and case studies including extended distance solutions:
www.hp.com/go/serviceguard
- Whitepapers and product documentation
<http://docs.hp.com/hpux/ha/>

