Oracle Real Application Clusters on Extended Distance Clusters

Updated for Oracle RAC 10g Release 2

An Oracle White Paper
October 2006

ORACLE

Oracle Real Application Clusters on Extended Distance Clusters

Oracle Real Application Clusters on Extended Distance Clusters

EXECUTIVE OVERVIEW

Oracle Real Application Clusters (RAC) is a proven mechanism for local high availability (HA) for database applications. It was designed to support clusters that reside in a single physical datacenter. As technology advances, customers are looking at the viability of using RAC over a distance.

Can RAC be used over a distance, and what does this imply? RAC on Extended Distance Clusters is an architecture that provides extremely fast recovery from a site failure and allows for all nodes, at all sites, to actively process transactions as part of single database cluster. While this architecture creates great interest and has been successfully implemented, it is critical to understand where this architecture best fits especially in regards to distance, latency, and degree of protection it provides.

The high impact of latency, and therefore distance, creates some practical limitations as to where this architecture can be deployed. This architecture fits best where the 2 datacenters are located relatively close (<~100km) and where the extremely expensive costs of setting up direct cables with dedicated channels between the sites has already been taken.

RAC on Extended Distance Clusters provides greater high availability than local RAC but it may not fit the full Disaster Recovery requirements of your organization. Feasible separation is great protection for some disasters (local power outage, airplane crash, server room flooding) but not all. Disasters such as earthquakes, hurricanes, and regional floods may affect a greater area. Customers should do an analysis to determine if both sites are likely to be affected by the same disaster.

For comprehensive protection against disasters including protection against corruptions, and regional disasters Oracle recommends the use of Data Guard with RAC as described in the Maximum Availability Architecture (MAA).[1] Data Guard also provides additional benefits such as support for full rolling upgrades across Oracle versions.

Configuring an extended distance cluster is more complex than a local cluster. Specific focus needs to go into node layout, quorum disks, data disk placement, and other factors discussed in this paper.

Implemented properly, this architecture can provide greater HA than a local RAC database. This paper will address the necessary components, the benefits and limitations of this architecture, and will highlight some actual customer examples.

INTRODUCTION

Oracle's Real Application Clusters (RAC) is designed primarily as a scalability and availability solution that resides in a single data center. It is possible, under certain circumstances, to build and deploy a RAC system where the nodes in the cluster are separated by greater distances. For example if a customer has a corporate campus they might want to place the individual RAC nodes in separate buildings. This configuration provides a degree of disaster tolerance, in addition to the normal RAC high availability, since a fire in one building would not, if properly set up, stop database processing.

This paper discusses the potential benefits that attract customers to this type of architecture, covers the components required and design considerations that should be considered when implementing, reviews empirical performance data over various distances, and covers the additional advantages that are provided by an Oracle Data Guard solution. Finally it looks at several case studies from actual production customer implementations.

Clusters, where all the nodes are not local, have been referred to by many names including campus clusters, metro clusters, geo clusters, stretch clusters and extended clusters. Some of these names imply a vague notion of distance range.

Throughout this paper this type of configuration will be referred as RAC on Extended Distance Clusters.

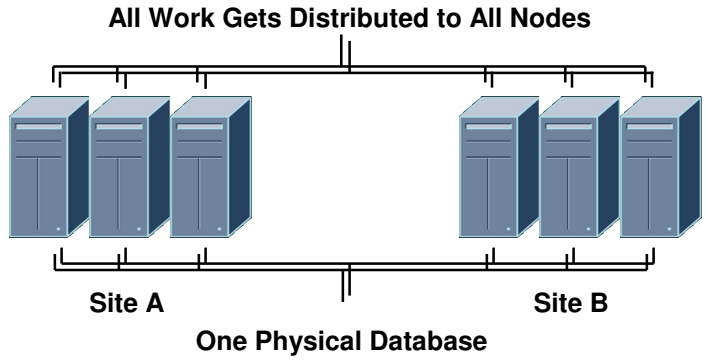
This paper is intended to provide a deeper understanding of the topic and to allow one to determine if this type configuration is applicable and appropriate.

BENEFITS OF RAC ON EXTENDED DISTANCE CLUSTERS

Implementing a RAC database on a cluster where some of the nodes are located a different site, is attractive to customers for the two main advantages it provides

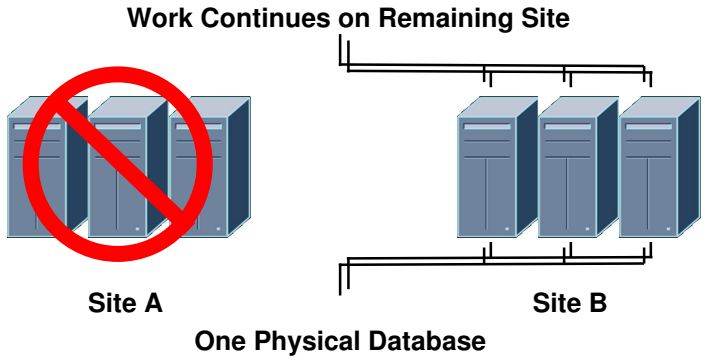
Full utilization of resources

Being able to distribute any and all work across all nodes, including running as a single workload across the whole cluster, allows for the greatest flexibility in usage of resources.



Extreme Rapid Recovery

Should one site fail, for example because of a fire at a site, all work can be routed to the remaining site that can very rapidly (< 1-2 minutes) take over the processing.

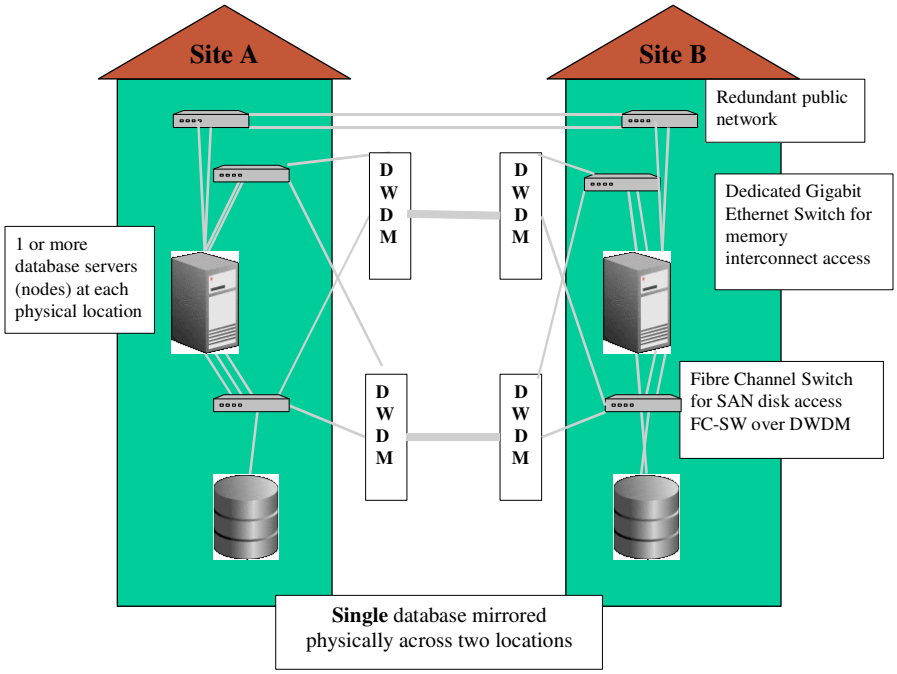


COMPONENTS & DESIGN CONSIDERATIONS

RAC on an Extended Distance Cluster is very similar to a RAC implementation at a single site.

To build a RAC database on an Extended Distance Cluster environment you will need to.

- Place one set of nodes at Site A
- Place the other set of nodes at Site B



- Use fast dedicated connectivity between the nodes/buildings for RAC cross instance communication (Dense Wavelength Division Multiplexing (DWDM or Dark Fiber) is optional)
- Use host or array based mirroring to allow you to host all the data on both sites and keep it synchronously mirrored.

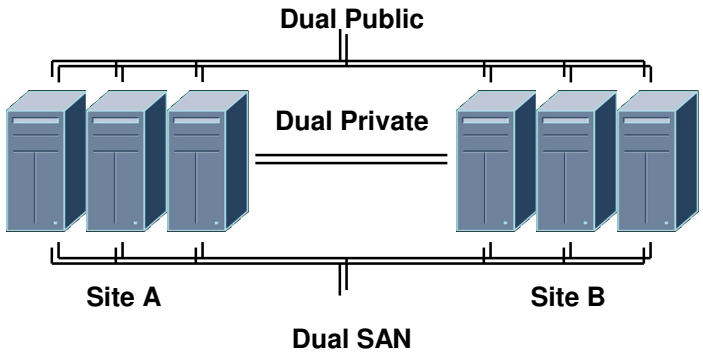
Details of the components, and design considerations, follow.

Connectivity

Networking requirements for a distance cluster are much greater than that of a normal Wide Area Network (WAN) used for Disaster Recovery. This plays in two aspects: necessary connections and latency.

Necessary Connections

Interconnect, SAN, and IP Networking need to be kept on separate *dedicated* channels, each with required redundancy. Redundant connections must not share the same Dark Fiber (if used), switch, path, or even building entrances. Keep in mind that cables can be cut.



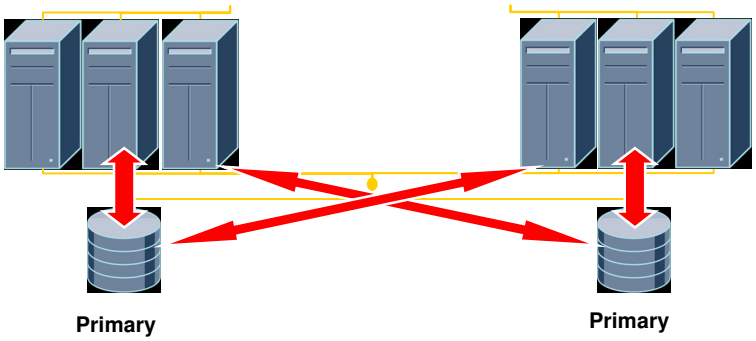
The SAN and Interconnect connections need to be on direct point-to-point cables (see effects of latency in the next section). Traditional networks are limited to about 10 km if you are to avoid using repeaters. Dark Fiber networks allow the communication to occur without these repeaters. Since latency is limited, Dark Fiber networks allow for a greater distance in separation between the nodes. The disadvantage of Dark Fiber networks are they can cost hundreds of thousands of dollars, so generally they are only an option if they already exist between the two sites.

Latency effects and performance implications of distances are discussed in the Latency & Empirical Performance Results Chapter.

Storage

RAC on Extended Distance Clusters by definition has multiple active instances on nodes at different locations. For availability reasons the data needs to be located at both sites, and therefore one needs to look at alternatives for mirroring the storage.

Host Based Mirroring (Active/Active Storage)

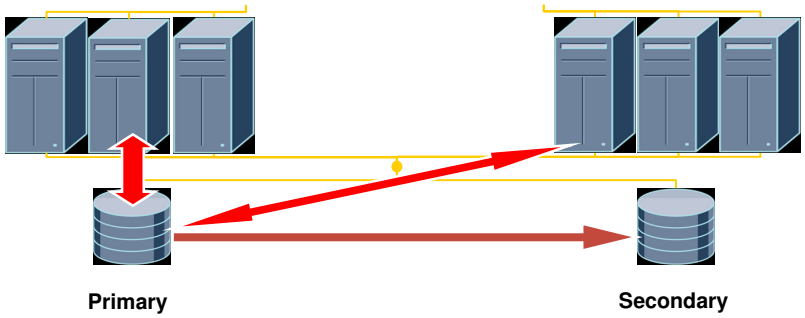


- Use two SAN/FC storage subsystems, one co-located with each node.
- Standard, cluster aware, host based (OS level) mirroring software is implemented across both disk systems. With this, system writes are propagated at the OS level to both sets of disks, making them appear as single set of disks independent of location. These Logical Volume Managers (LVM) need to be tied closely with the clusterware. Examples of these include Veritas CVM, HP-UX Mirror Disk/UX, & Oracle's Automatic Storage Management (ASM).
- While there may be a performance impact[1] from doing host based versus array based mirroring, this is the preferred configuration from an availability perspective. When we refer to RAC on Extended Distance Clusters in this paper, it generally refers to this active/active storage configuration.

[1] Host based mirroring requires CPU cycles from the host machines. Array based mirroring offloads this work to the storage layer. Advantage or disadvantage of this depends on which layer you either have spare cycles or it is more cost effective to add cycles.

⚠ **CAUTION: Array Based Mirroring generally implies a primary/secondary storage site solution. Should the primary storage location fail, all instances will crash and need to be restarted once the secondary storage is made active. Array based mirroring requires a switch be made from receiving changes at the remote side to functioning as local disk. From an HA viewpoint it is recommended to instead do Host Based mirroring as it does not require a manual restart.**

Array Based Mirroring (Active/Failover Storage)



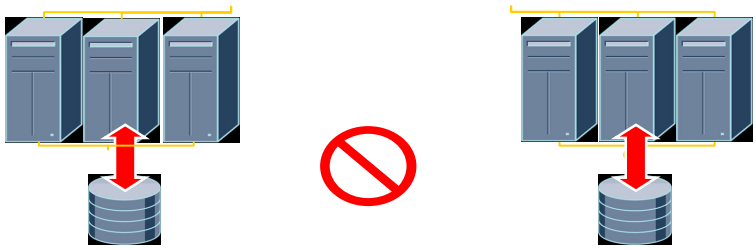
- Use two SAN/FC storage subsystems, one co-located with each node and each is cross cabled to both nodes
- One storage subsystem has all the live database files on it, all writes are sent to this system
- The second storage subsystem has an array based mirror mechanism (i.e. EMC's SRDF, HP's CA, etc.) of the first storage subsystems files
- Performance impacts in this case come from both doing additional work in the storage array for the mirroring, but more importantly by I/Os from the secondary site having to cross the 'distance' 4 times[2] before they return control.
- In this case additional cycles are consumed in the storage arrays to do the mirroring, and additional latency introduced for the 'secondary' site I/O as it needs to first come to the primary storage and

Why not have just a single storage location?

While it is possible to implement RAC on Extended Distance Clusters with storage on only one site, should the site with the storage fail, storage is no longer available to any surviving nodes, and the whole cluster becomes unavailable. This defeats the purpose of having had the RAC nodes at different locations.

[2] Secondary host to primary storage, primary storage to secondary storage, secondary storage to primary storage, primary storage to secondary storage. All need to be synch to ensure no data loss.

Cluster Quorums, or Ensuring Survival of One Part of the Cluster:



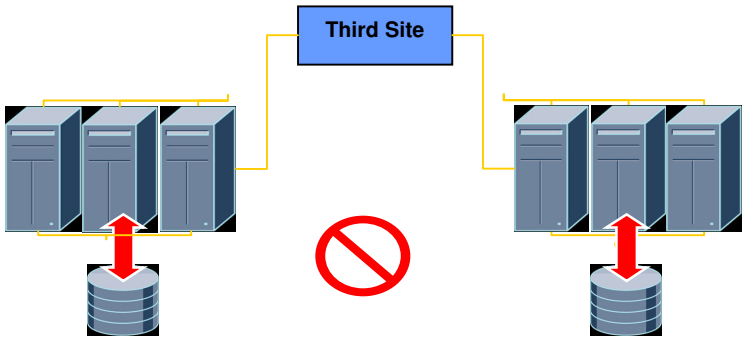
Cluster quorum mechanisms have a bigger impact on the design of an extended distance cluster than they would on a local cluster.

⚠ CAUTION: Extended RAC implementations without a third site for tie breaking quorum, require making one site a 'primary' site and the other a secondary. Then should the primary site fail, the secondary site will *require a manual restart*.

When a local cluster is being built, one need not worry much about how quorum mechanisms work. Cluster software is designed to make the process fool proof both for avoiding split brains[3] and for giving the best odds for a portion of the cluster to survive when communication failure between the nodes occurs.

Once one takes the nodes of the cluster and separates them, things are no longer so simple. They have a tie breaking mechanism that must be located someplace.

Alternatively all cluster software support putting a tie-breaker at a third site. This allows both sites to be equal and the third site can act as an arbitrator should either fail or connectivity is lost between the sites. Because of the HA implications, the 3 site implementation is highly recommended.



Setting up voting disks across sites should only be done directly via the clusterware software. They should not be mirrored remotely otherwise as this could potentially result in a dual active database scenario.

Depending on the clusterware provider, the third site may not have the same connectivity requirements and may be connectable via a WAN. Some quorum mechanisms may also require a balanced number of nodes across the sites. More detailed discussion and examples of quorum mechanisms, and the alternatives for implementing the third site are discussed in Appendix A.

---

[3] A split brain is when two portions of the cluster stop coordinating and start doing actions on their own. This would usually lead to a database corruption, so clustering software and Oracle are both carefully written to avoid a split brain situation from occurring.

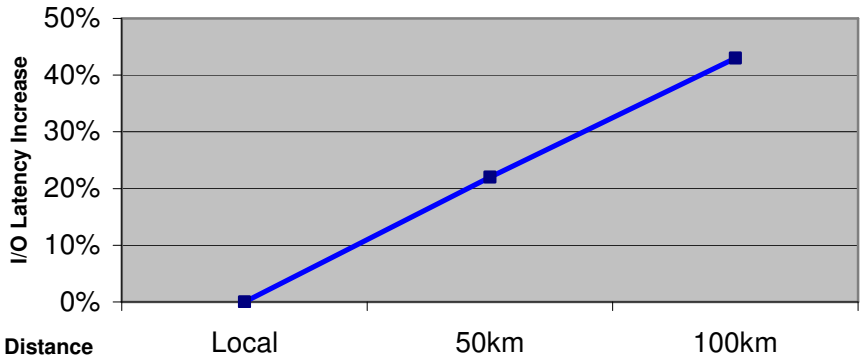
Latency & Empirical Performance Results

Oracle Real Application Clusters requires that the cluster interconnect (and thus Cache Fusion) have a dedicated low latency network. A dedicated network is required to ensure consistent response times and avoid the loss of the cluster heartbeat, which can cause nodes to be kicked out of the cluster. Interconnect latency directly affects the time it takes to access blocks in the cache of remote nodes, and thus directly affects application scalability and performance. Local interconnect traffic is generally in the 1-2 ms range and improvements (or degradations) can have a big effect on the scalability levels of the application. I/O latencies tend to be in the 8-15ms range, and are also affected by the additional latencies introduced with distance.

Various partners have tested RAC on Extended Distance Clusters. These tests include ones done by Mai Cutler (HP) and Stefan Pommerenk (Oracle) at 0,25,50, and 100 km; tests done by Paul Bramy (Oracle), Christine O'Sullivan (IBM), Thierry Plumeau (IBM) at the EMEA Joint Solutions Center Oracle/IBM at 0,5, and 20 km; and tests done by Veritas at 0, 20, 40 and 80km. All included a full OLTP application test and some included unit tests of the individual components.

The unit tests results from the HP/Oracle testing will be used to illustrate what happens at each component level.

Figure 2: I/O Latency Increase Over Distance

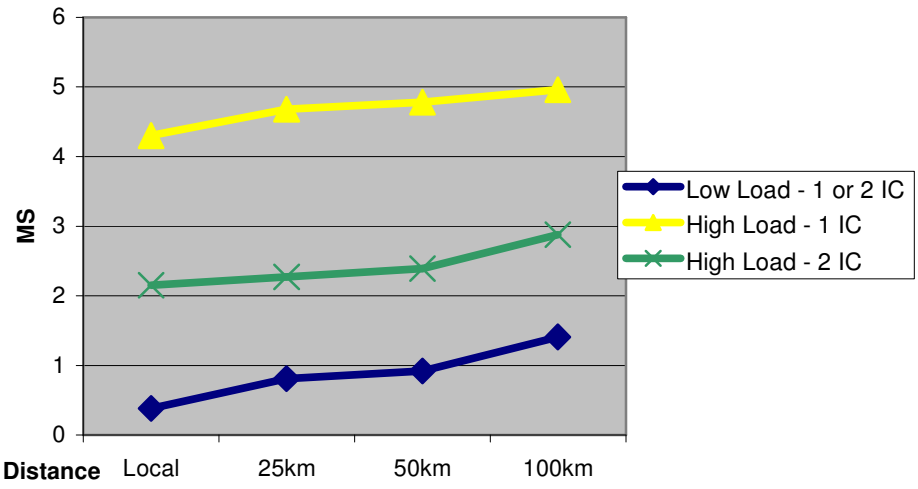


This figure shows the effects of distance on I/O latency with SAN Buffer Credits. SAN Buffer Credits allow a greater number of unacknowledged packets on the wire, thus allow greater parallelism in the mirroring process. As distances increase, especially with high traffic volumes, these SAN Buffer Credits can make a huge difference. For example when the tests above where run without the additional SAN Buffer Credits, I/O Latency at 100km was 120-270% greater than local, instead of 43% in the chart above. The folks at the IBM/Oracle EMEA Joint

Solutions Center recommend 1 SAN Buffer Credit for each 2 kilometers.[4] These numbers are consistent with the results from the Oracle/IBM testing which had 20-24% throughput degradation on I/O Unit tests at 20km when SAN Buffer Credits where not set.

Interconnect Traffic Unit Test Results

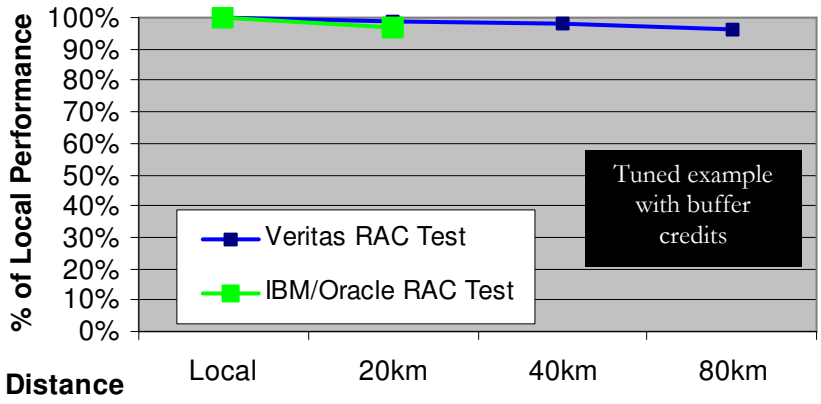
Tests at both high and low load levels, and with one or two interconnects, show that there is an increase of about 1 ms at 100km. While Cache Fusion traffic is not as sensitive to distance as I/O latency, the effect of this latency increase can be as significant



[4] Bramy, O'Sullivan, Plumeau & the EMEA Joint Solutions Center, Oracle 9i RAC Metropolitan Area Network implementation in an IBM pSeries environment.

Overall Application Impact

Unit tests are useful, but the final real impact comes down to how a full application reacts to the increased latencies induced by distance. Having three independent sets of tests provides a more complete picture than each individual test. A summarization of each test is provided, and full details can be seen in the paper by each respective vendor listed in references.

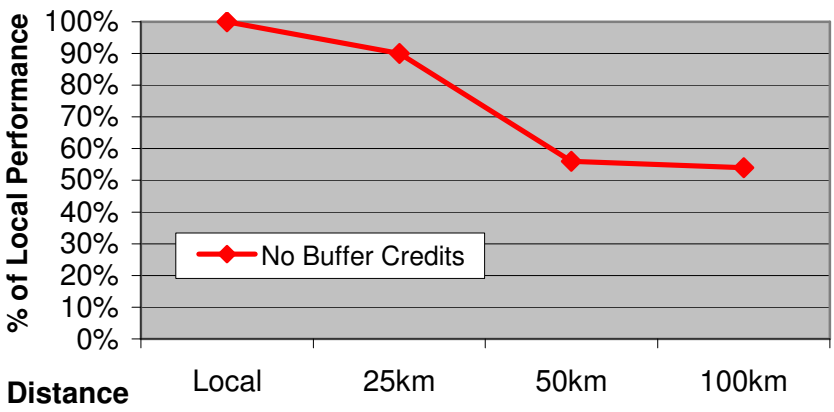


The IBM/Oracle tests performed a representative workload, which was accomplished by running the SwingBench workload with proper use of SAN Buffer Credits. These tests at 20km showed 1% degradation for read transactions, 2-8% degradation for most write transactions. The average single transaction resulted in 2% degradation.

Veritas used another well-known OLTP workload, and set it up in a manner in which it was highly scalable. These tests done at 0, 20, 40, and 80km showed that the application suffered minimal performance loss (4% in their worst case at 80km).

Extended RAC implementations without a third site for tie breaking quorum, require making one site a 'primary' site and the other a secondary. Then should the primary site fail, the secondary site will *require a manual restart.*. Other tests where done without having SAN Buffer Credits set. Combined with a very contentious application, this resulted in minimal impact at 25km (10%), but significant degradation at 50km-100km. Further testing is needed to determine why the 50 & 100km numbers are similar, but the 0, 25 and 100km numbers form a

⚠ CAUTION: Not using SAN Buffer Credits can cause serious application performance degradation for greater distances



very nice linear slope. With appropriate SAN Buffer Credits these numbers would be expected to significantly improve and be closer to the Veritas and Oracle/IBM numbers.

Real life applications are expected at best to follow the IBM/Oracle & Veritas examples demonstrated earlier. In reality they will probably have more interconnect traffic and thus suffer slightly more from the distance. For example a real life example done at Comic Relief in the United Kingdom by Mike Hallas and Rob Smyth from Oracle. Those results showed that a cluster with an 8km distance between nodes has roughly 10% degradation in service versus running the application on a local cluster.[5]

Each of these results is for a particular application with a particular setup. Other applications will be affected differently, but the basic idea is that as distance increases, IO and Cache Fusion message traffic latency increases. The limitations come from a combination of the ideal network speed minus inefficiencies and additional latency added by each time the network goes through a switch, router or hub.[6] As was previously stated, Dark Fiber can be used to achieve connections greater than 10km without repeaters.

While there is no magic barrier to how far RAC on an Extended Distance Clusters can function, it will have the least impact on performance at campus or metro distances. Write intensive applications are generally more affected than read intensive applications. If a desire exists to deploy RAC at a greater distance, performance tests using the specific application are recommended.

From these numbers I am generally comfortable with a RAC on Extended Distance Clusters at distances under 25km, concerned about performance at 50km, and skeptical at 100km or more. There is no magic barrier for the distance; latency just keeps getting worse. Write intensive applications are generally more affected then read intensive applications.

---

[5] Hallas & Smyth *Comic Relief Red Nose Day 2003 (RND03)*, I*nstalling a Three-Node RAC Cluster in a Dual-Site Configuration using an 8 Km DWDM Link*, Issue 1, April 2003

[6] If a direct connection does not exist, over 1 km switches should be used instead of hubs, as hubs experience an exponential degradation over distance (80% already at 1km) (Algieri & Dahan, page 44)

HARDWARE VENDOR SPECIFICS

The hardware vendor should approve the cluster configuration that is implemented, including the disk mirroring mechanisms used, quorum device placement, and interconnect redundancy.

Below we provide examples of support by HP, Sun and IBM for a general extended distance cluster environment. Other configurations may also be supported, please contact them for specifics.

Sun

The campus cluster configurations started being supported with in Sun Cluster 3.0 In Sun Cluster 3.1 support is now expanded on a wide variety of the newer Sun storage devices, Fast Ethernet or Gigabit Ethernet, media converters for extending the interconnect up to 10 kilometers and the supported node count has been increased to 8. Details of this are covered in Sun Blueprints document in the references section.

This paper is not RAC specific, but there are many customers in production on this platform listed in the customer's section.

HP

Tru-64: Support for extended distance clusters has existed for a good many years, and this is the environment being used by some of the production customers referenced below.

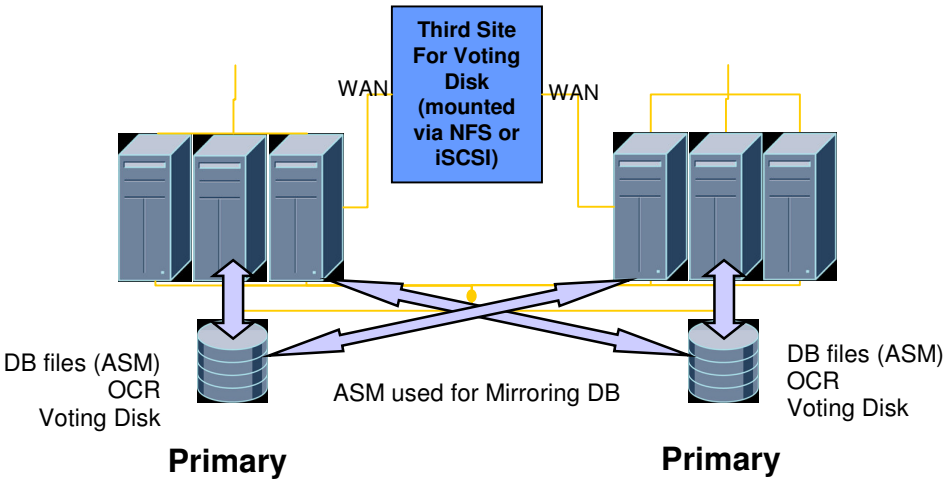
HP-UX: HP's offering on HP-UX is called Extended Serviceguard clusters, and is offered on either 2 or 3 site configurations. The number of nodes supported varies from 2-16, and is dependent on storage technology and distance. HP has tested and can therefore support a single cluster whose nodes (and disks) are separated by a distance of up to 100 kilometers using CFS, CVM, and SLVM. Joint Oracle and HP test results at 25, 50, and 100 kilometer distance are included in this paper and where presented at Oracle World San Francisco in 2003 (see references).

IBM

The Oracle/IBM Joint Solution Center has successfully tested RAC with pSeries, iSeries and xSeries servers. Tests have been done on both AIX & Linux, and have used either AIX Mirroring (on AIX) or ASM (on Linux or AIX) to keep the disks in sync. They also have built practical experience with real world production customers. Their tests included both detailed high availability and performance components. Their research has great detail on DWDM and fiber networking, and detailed configuration on was what used in the testing (see references).

FULL ORACLE STACK

Starting in Oracle Database 10g Release 2, one is now able to create an extended cluster on any OS using standard Oracle components. The Oracle Clusterware can be used for integrity and Automatic Storage Management (ASM) for mirroring.



Oracle Clusterware

Starting with the version of Oracle Clusterware released with Oracle Database 10g Release 2, Oracle provides direct support for mirroring of the Oracle Cluster Repository (OCR), as well as supporting multiple voting disks.

To setup an extended RAC with Oracle Clusterware:

1. OCR must be mirrored across both sites using Oracle provided mechanisms.
2. Voting disks we have a preferably 2 voting disks at each site, and tie-breaking voting disk at a third site. This third site only need be a supported NFS device over a WAN. On most platforms this is still currently the same as full RAC support (i.e. using something like a NetApp filer) but support for Generic NFS is in progress and is currently available on Linux.[7]

ASM

ASM built in mirroring can be used to efficiently mirror the rest of the database files across both sites. Storage at each site much be setup as seperate failure groups and use ASM mirroring, to ensure at least one copy of the data at each site.

---

[7] **Roland Knapp, Daniel Dibbets, Amit Das,** Using standard NFS to support a third voting disk on a stretch cluster configuration on Linux, September 2006

Two minor limitations do exist with ASM mirroring which are not necessarily present when using other cluster software:

1. ASM does not currently provide partial resilvering. Should a loss of connectivity between the sites occur, one of the failure groups will be marked invalid. When the site rejoins the cluster, the failure groups will need to be manually added. This will not impact normal operations.
2. ASM does not currently provide a mechanism for local reads. I/O read requests to an ASM group will be satisfied from any available mirror. Some other cluster software (Veritas, Sun Cluster, etc) do provide reads from the local mirror. Except for extended clusters that are very far apart this should not have much of an impact.

Solutions for both of these are planned for a future release of ASM.

COMPARISON WITH A LOCAL RAC AND DATA GUARD REMOTE SITE

Here is a comparison of a RAC over an Extended Distance Cluster versus a local RAC cluster for HA and Data Guard for DR.

Comparison Summary

	RAC on Extended Distance Clusters	**RAC + DG**
Needed Nodes	2	3
Active Nodes	All	One Side Only DG site can be used for reporting purposes
Recovery from Site Failure	Seconds, No Intervention Required	Seconds, No Intervention Required[8]
Performance Hit See charts	Minor to Crippling	Insignificant to Minor in same Cases
Network Requirements	High cost direct dedicated network w/ lowest latency. Much greater network bandwidth	DG Sync - High cost direct dedicated network w/ lowest latency. DG Async Shared commercially available network. Does not have low latency requirements.
Effective Distance	Campus & Metro	Country and Continental-Wide distances
Disaster Protection	Host, building, and localized site failures	Host, building, localized site failures, Database Corruptions Local and wider area Site Disasters
Costs	Very High Network Costs	Additional Nodes

Strengths of RAC on Extended Distance Clusters

All Nodes Active

One of the main attractions for an Extended Distance Cluster environment is that all nodes can be active, and dedicated nodes are not required for disaster recovery. Thus instead of a minimum of 2 RAC clusters required in full RAC+DG architecture, 1 RAC cluster can be used. One note of comment: in a RAC+DG architecture, the DR site can be used for other purposes including reporting and decision support activities.

In environments with larger number of nodes, some advantage is still gained from having all nodes able be an active part of the same cluster

[8] Assuming you are using Fast-Start Failover in Oracle Database 10g Release 2 onwards

Fast Recovery

Prior to Oracle 10g Release 2 the biggest advantage of RAC on Extended Distance Clusters is that when a site fails, it is possible to recover quickly with no manual intervention needed. With Data Guard, when the primary site fails, failover is generally manually instantiated. In Oracle Database 10g Release 2, Fast-Start Failover was introduced as a Oracle Data Guard feature that automatically, quickly, and reliably fails over to a designated, synchronized standby database in the event of loss of the primary database, without requiring manual intervention to execute the failover. This also requires a third arbitrating site.

Now in the event of server failures, both RAC and Data Guard with Fast-Start Failover can accomplish the failover in a few seconds, requiring no manual intervention.

Costs

The biggest attraction of Extended Distance Clusters is in its potential to reduce costs. By being able to have all nodes active, it is possible to get scalability, very high availability and DR with just 2 nodes.

While one could get away with just one mirror copy of the data at each site, this would be a risky proposition when one site becomes unavailable. Two mirrors should be kept at each site, totaling 4 copies of the data (same as w/ RAC + Data Guard).

Cost increments can be incurred by the higher bandwidth and specialized communication needs of an extended distance cluster environment. Dark Fiber for example can easily cost *hundreds of thousands of dollars*. Additional costs can come from reduced performance, and the potential need to implement a third site[9] for the quorum disk.

Strength of local RAC + Data Guard at a remote site

No Performance Hit

A Data Guard environment can be setup to be asynchronous, which allows data to be transferred across a great distance and still have from none to a minimal impact on the performance of the primary environment. Of course in an asynchronous configuration you no longer have the guarantee of zero data loss.

With RAC, the sites are much more tightly coupled, thus any latencies involved have a greater affect because of the separation of the two sites. Details of this were discussed in the latency section (Page 6). Furthermore, the latency affects the data transfer between caches. Data Guard only sends redo data, and thus is less sensitive to network latency.

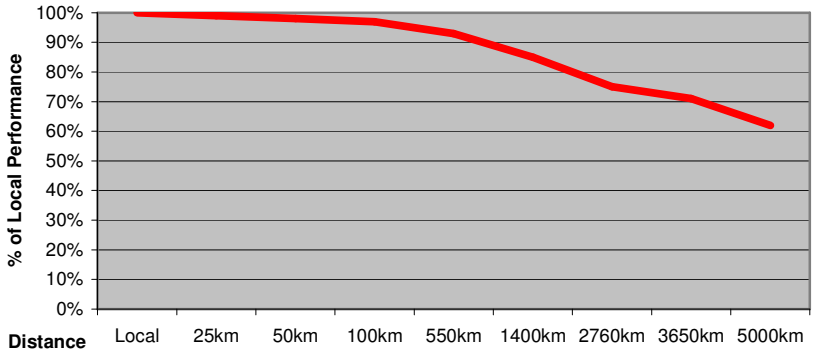
---

[9] This can be negligible for large corporations with multiple locations. For example with the HP quorum server this can be any site with IP access, running a very small server.

To show the difference performance and bandwidth impact of Data Guard versus full mirroring, it is useful to look at internal analysis of Oracle's corporate e-mail systems. Here it was demonstrated that 7 times more data was transmitted over the network and 27 times more I/O operations were performed using a remote mirroring solution, compared to using Data Guard. [10]

Keeping in mind the performance impact caused by distance with RAC on an Extended Distance Cluster for a well-known OLTP workload (Figure 4), it is interesting to compare this to some other performance impact tests for synchronous Data Guard[11] with another well known OLTP workload. These tests show that the impact of distance on Data Guard is much less, even allowing distance to be taken to thousands of km, something that would be impossible with RAC.

Performance Degradation for Data Guard Sync



Now why is there such a difference? The impacts of latency are actually quite different and occur in different layers: with Data Guard, the impact is on the synchronous I/O from lgwr and network I/O for redo transmission, whereas with RAC on Extended Cluster, the impact is on the synchronous I/O from dbwr, lgwr and network.

Greater Disaster Protection

An Extended Distance Cluster scenario does not provide full disaster recovery (DR) as distance between the sites is limited. In reality it is more of an extended HA solution as it does get you some degree of separation between the sites.

In DR design it is important to avoid common utility failure (water, electricity), being on the same flood plain, or being part of a larger location that can all be damaged by the same jumbo jet. In an earthquake zone the general

---

[10] from Oracle Data Guard and Remote Mirroring Solutions, Oracle Technology Network (OTN)

[11] from Oracle9*i* Data Guard Log Transport Services and Performance Characterization by Rabah Mediouni and Rick Anderson

recommendation is 300 km at right angles to the main fault line. Hurricanes and wars can take out very large areas. Terrorism brings more unpredictable effects.

So for example if the two sites are in a non-flooding non-earthquake zone, not under a flight path and each has independent automatic standby generators and self-contained cooling then 1Km may be ample except perhaps in times of war, terrorism, hurricanes, etc.

Data Guard is able to function more efficiently and at a much greater distance, and is in general a more complete DR solution. Many of its advantages come from the fact that Data Guard does not depend upon remote-mirroring to synchronize the replica at the remote site. Specifically:

- Data Guard Redo Transport Services provide the database performance and network utilization advantages described above.
- Unlike remote-mirroring, Data Guard Apply Services validate redo data before it is applied to data files at the remote site. This validation isolates the remote database from hardware-induced data file corruptions that can occur at the primary location or during the transmission of data to the remote site.
- Data Guard can provide a delayed copy to protect against user errors (important in Oracle9i but not in Oracle Database 10g when Flashback Database is used).
- Oracle Databae 10g Rolling Upgrades with Data Guard also provide the ability to reduce downtime during planned outages.

Costs

A RAC approach with only Data Guard at the remote site requires less network bandwidth and these networks do not need to be as redundant or with such extreme low latencies they would need for a RAC environment on Extended Distance Clusters.

Other Limitations of RAC on Extended Distance Clusters

- All vendors do not offer these configurations and the amount of testing they have done varies.
- Quorum implementations in some platforms (HP-UX) require that there is an equal number of nodes at each site.

CONCLUSION

RAC on Extended Distance Clusters is an attractive alternative architecture that allows scalability, rapid availability, and even some very limited disaster recovery protection with all nodes fully active.

This architecture can provide great value when used properly, but it is critical that that the limitations are well understood. Distance can have a huge effect on performance, so keeping the distance short and using costly dedicated *direct* networks are critical.

While this is a greater HA solution compared to local RAC, it is not a full Disaster Recovery solution. Distance cannot be great enough to protect against major disasters, nor does one get the extra protection against corruptions and flexibility for planned outages that a RAC and Data Guard combination provides.

While this configuration has been deployed by a small number of customers, thorough planning and testing is recommended before attempting to implement.

APPENDIX A: DETAILED QUORUM EXAMPLES

Clusters are designed so that in the case of a failure of communication between any 2 subsets of nodes of the cluster, at most one sub-cluster will survive and thus avoid corrupting the database.

The "At Most" in the last phrase is key. If the clusterware cannot guarantee after a failure that only one sub-cluster will survive, then all sub-clusters go down. You cannot assume that sub-clusters will be able to talk to each other (a communication failure could be the cause of needing to reform the cluster).

How the clusterware handles quorum affects how one should layout an extended cluster. Some cases require a balanced number of nodes at each of the 2 main sites, while all cases require a third site to locate the tie-breaking device for higher availability.

The following examples will help you to understand the details of these restrictions, as well as get a better understanding of how quorum works.

HP Serviceguard / Sun Cluster example

Quorum is achieved here by giving each node a vote, and a quorum device (normally a disk or server) acts a tiebreaker to make sure only one side gets the majority.

Veritas Storage Foundation for RAC (fomerally DBE/AC) example:

With Veritas SFRAC, nodes don't get votes but instead all nodes race for access to 3 coordinator disks. Because of the algorithm used, larger subsets of nodes will get to the coordinator disks quicker thus are more likely to survive.

In a 2-site environment, one would not want both sides to survive, as this would quickly cause corruptions. Therefore one side must be able to form a quorum, and the tie breaking vote must exist on one side or the other. This ends up creating a primary and a secondary site. Should the primary site fail, the secondary site will not have a quorum and will shut down. In this case a manual reconfiguration is required and this should be practiced and well rehearsed.

In a 3-site implementation, quorum can be redistributed so that any 2 sites left can have a majority of votes or coordinator disks to ensure that the cluster survives.

Oracle Clusterware example:

The following example applies to when only Oracle Clusterware is used (i.e. on Linux and Windows in Oracle9i and on all platforms in Oracle10g when a third party clusterware is not used in conjunction with Oracle Clusterware).

By design, shared disk cluster nodes have 2 ways to communicate with each other, thru the interconnect network and shared disk sub system. Many vendor 's clusterware monitor cluster availability only based upon the network heartbeat., but depend upon SCSI timeouts for detecting disk failures to one or all nodes,

these timeouts can take up to 15 minutes..
Oracle Clusterware uses the concept of a voting disk and a heartbeat to monitor the cluster through both the disk subsystem and the interconnect. This helps Oracle Clusterware to resolve asymmetric failures very effectively without resorting to SCSI timeout mechanisms.

This method of using the voting disk actively helps protect against heterogeneous failures (where one node sees the cluster as being fine but others do not) but it also means that the 'voting disk' must be accessible at all times, from all nodes or the cluster will fail, and the location of 'voting disk' will make that site primary.

The 'voting disk' file should be mirrored locally for high availability purposes. Multiple voting disks setup via Oracle Clusterware are not mirrors, but members of a group for which you need to achieve a quorum to continue. Thus a local mirror is good. They should not be mirrored remotely as part of an extended cluster as this could allow two sub clusters to continue working after a failure and potentially lead to a split-brain or diverging database situation.

APPENDIX B: CUSTOMERS USING RAC ON EXTENDED DISTANCE CLUSTERS

The Rover Group completed the first known implementation with a similar architecture in the mid 1990's using Oracle7 Parallel Server. Since then other clients have implemented it with Real Application Clusters including the following examples:

The list below includes many of the known production customers running RAC on an extended cluster. Because Oracle9i has been around for a longer period it has a larger set of the production customers. Today the majority of new customers implementing are doing so using Oracle 10g, Oracle Clusterware and using ASM to mirror the data between the sites.

Names in italics have been modified to only show country or region and industry.

Name	Release	Nodes	Platform	OS	Clusterware	Stretch Distance (KM)
Italian Financial Services firm	10g	20	IBM	AIX	HACMP	0.2
Groupe Diffusion Plus	10g	2	IBM	AIX	Oracle	0.5
Austrian IT Services Provider	10g	2	IBM	AIX	HACMP	1
European Electronics firm	9i	2	IBM	AIX	HACMP	8
US Police Department	9i	2	IBM	AIX	HACMP	3
European Government	9i	2	IBM	AIX	HACMP	8
US Broadcaster	9i	2	IBM	AIX	HACMP	0.2
Austrian Hospital	9i	2	IBM	AIX	HACMP	0.6
Brazilian Credit Union Network	9i	3	IBM	AIX	HACMP	10
UzPromStroyBank	9i	2	IBM	AIX	HACMP	1.7
Daiso Sangyo	10g	2	HP	HP-UX	Oracle	10
US Fortune 100 firm	9i	2	HP	HP-UX	HP Service Guard	2
Brazilian Hospital	9i	2	HP	HP-UX	HP Service Guard	0.5
Italian Manufacturer	10g	4	HP	Linux	Oracle	0.8
Swedish Automotive Parts	10g	2	IBM	Linux	Oracle	2
Austrian Health Provider	10g	2	IBM	Linux	Oracle	0.3
Thomson Legal	10g	8	Sun	Linux	Oracle	1
North American Lottery	9i	4	HP	OpenVMS		10
German Telecom	10g	4	Sun	Solaris	Sun Cluster	5
European Bank	10g	2	Sun	Solaris	Oracle	5
European Mobile Operator	9i	3	Sun	Solaris	Veritas Cluster	48
Comic Relief	9i	3	Sun	Solaris	Sun Cluster	8
German Bank	9i	2	Sun	Solaris		12
European Mail	9i	2	Sun	Solaris	Veritas Cluster	12
European Government	9i	2	Sun	Solaris	Sun Cluster	0.4
UK University	9i	2	Sun	Solaris	Sun Cluster	0.8
Italian Telco	9i	2	Sun	Solaris	Sun Cluster	2
Austrian Railways	9i	2	HP	Tru64	TruCluster	1.5
Nordac/ Draeger	9i	4	HP	Tru64	TruCluster	0.3
University of Melbourne	9i	3	HP	Tru64	TruCluster	0.8
European Electronics Components firm	10g	2	IBM	Windows	Oracle	0.5
Spanish Health firm	9i	6	Dell	Windows	Oracle	25

REFERENCES

Roland Knapp, Daniel Dibbets, Amit Das, Using standard NFS to support a third voting disk on a stretch cluster configuration on Linux, September 2006

EMEA Joint Solutions Center Oracle/IBM, 10gRAC Release2 High Availability Test Over 2 distant sites on xSeries, July 2005

Paul Bramy (Oracle), Christine O'Sullivan (IBM), Thierry Plumeau (IBM) at the EMEA Joint Solutions Center Oracle/IBM, *Oracle9i RAC Metropolitan Area Network implementation in an IBM pSeries environment*, July 2003

Veritas, *VERITAS Volume Manager for Solaris: Performance Brief – Remote Mirroring Using VxVM*, December 2003

HP Oracle CTC, Extended Serviceguard cluster configurations. Detailed configuration information for extended RAC on HP-UX clusters, Novemer 2003

Mai Cutler (HP), Sandy Gruver (HP), Stefan Pommerenk (Oracle) *Eliminate the Current Physical Restrictions of a Single Oracle Cluster,* OracleWorld San Francisco 2003

Joseph Algieri & Xavier Dahan (HP)**,** *Extended MC/ServiceGuard cluster configurations (Metro clusters),* Version 1.4, January 2002

Michael Hallas and Robert Smyth, *Comic Relief Red Nose Day 2003 (RND03), Installing a Three-Node RAC Cluster in a Dual-Site Configuration using an 8 Km DWDM Link,* Issue 1, April 2003

Lawrence To, *Oracle Database 10g Release 2: Roadmap to Maximum Availability Architecture (MAA),* April 2006

Michael T. Smith, *Oracle Database 10g Release 2 Best Practices: Data Guard Redo Transport & Network Configuration,* August 2006

Oracle Technology Network, *Oracle Data Guard and Remote Mirroring Solutions*

Joseph Meeks, Michael T. Smith, Ashish Ray, Sadhana Kyathappala, *Fast-Start Failover Best Practices: Oracle Data Guard 10g Release 2,* November, 2005

Tim Read, Architecting Availability & Disaster Recovery Solutions, Sun BluePrints™ OnLine, April 2006

Oracle Ral Application Clusters on Extended Distance Cluster
October 2006
Author: Erik Peterson
Reviewers: Daniel Dibbets, Bill Bridge, Joseph Meeks

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com