An Oracle White Paper
October 2013

Oracle RAC and Oracle RAC One Node on Extended Distance (Stretched) Clusters

ORACLE

Introduction

Oracle Real Applications 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 as well as demands increase, the question arises whether Oracle RAC or Oracle RAC One Node over a distance is a viable solution.

Oracle RAC on Extended Distance (Stretched) Clusters is an architecture that provides extremely fast recovery from a site failure and allows for all nodes, in all sites, to actively process transactions as part of a single database cluster. While this architecture has been successfully implemented many times over the last decade, it is critical to understand where this architecture fits best, especially in regards to distance, latency, and the 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. An active / active Oracle RAC architecture fits best where the two datacenters are located relatively close (<100km) and where the costs of setting up a low latency and dedicated direct connectivity between the sites for Oracle RAC has already taken place, which is why it cannot be used as a replacement for a disaster recovery solution such as Oracle Data Guard or Oracle GoldenGate.

Oracle RAC One Node completes the Extended Distance Cluster offering. While using the same infrastructure like Oracle RAC (i.e. Oracle Clusterware and Oracle ASM), Oracle RAC One Node is best used in environments or for applications that do not require the degree of HA or scalability that an active / active Oracle RAC database provides, but which benefit from an integrated and fully supported failover solution for the Oracle database.

As Oracle RAC One Node maintains only one instance during normal operation, it can also be used in environments, in which an insufficient latency prevents Oracle RAC to function optimally for example.

Oracle RAC on Extended Distance Clusters provides a greater high availability than a local Oracle RAC implementation. However, it does not provide full Disaster Recovery. Feasible separation is a 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. Oracle RAC on Extended Distance Clusters does not protect from human errors or corruptions in the shared storage, either, as an Oracle RAC system, even on Extended Distance Clusters, is still a tightly coupled and enclosed system.

For comprehensive data protection, including protection against corruptions and regional disasters, Oracle recommends using Oracle Data Guard together with Oracle RAC as building blocks of Oracle's Maximum Availability Architecture (MAA)[1]. For active-active databases deployed across geographically separated data centers, the MAA recommendation is to use Oracle GoldenGate. Note that Data Guard and GoldenGate also provide additional benefits such as minimizing downtime for maintenance activities such as upgrades and migrations.

This paper discusses the potential of the extended cluster architecture, covers the required components and design options that should be considered during implementation, reviews empirical performance data over various distances, explores supported and non-supported configurations, and lists the advantages that Oracle Data Guard provides to this solution.

[1] For more information regarding MAA, see: http://www.oracle.com/goto/maa

What is an Extended Distance Oracle RAC Cluster?

As the name implies, an Extended Distance Cluster is a cluster, in which most or all the nodes are not local and typically set up with a certain distance between them. Clusters of this kind have been referred to by many names, including "campus clusters", "metro clusters", "geo clusters", "stretched clusters" and "extended clusters". Some of these names imply a vague notion of distance range.

For this paper, this type of configuration will be referred to as either Oracle RAC on "Extended Distance Clusters" or "Stretched Clusters", while the names will be used synonymously. If Oracle RAC One Node is concerned in a certain context, respective references will be made.

Unlike classic Oracle RAC implementations, which are primarily designed as scalability and high availability solution that resides in a single data center, it is possible – under certain circumstances – to build and deploy an Oracle RAC system in which the nodes are separated by greater distances.

For example, if a customer has a corporate campus, they might want to place the individual Oracle RAC nodes in separate buildings. This configuration provides a higher degree of disaster tolerance, in addition to the normal Oracle RAC high availability, since a fire in one building would not, if properly set up, stop the database from processing. Similar, many customers have two data centers in reasonable proximity (<100km) which are already connected by a direct, ideally non-routed, high speed link(s) and are often on different power grids, flood plains, and the like.

Practically, there is, however, another characteristic of Extended Distance Clusters or Stretched Clusters, at least when implemented for an Oracle RAC Database. This aspect of an Extended Distance Cluster derives from the storage configuration and setup and is therefore independent of the physical distance between the nodes in the cluster.

Assuming that an Extended Distance Cluster maintains servers that are physically dispersed to protect the system as a whole from local server failures, similar considerations can lead to a storage setup that foresees (at least two) independent storage arrays (SAN / NAS systems) in different locations. As a matter of fact, probably all Extended Distance Clusters use this kind of storage setup. However, often one will find that the necessity to protect the storage arrays leads to an architecture that is also used in Extended Distance Clusters, while the nodes in the cluster are actually in rather close proximity.

Over the last decade of implementing Oracle RAC on Extended Distance Clusters, the latter aspect might have been a driver for implementing such architecture more often than the idea of protecting the system as a whole from server failures. While for the configuration of such systems the main driver is of minor importance it needs to be pointed out that such a configuration cannot replace a disaster recovery solution, protecting the shared storage from more than just a failure.

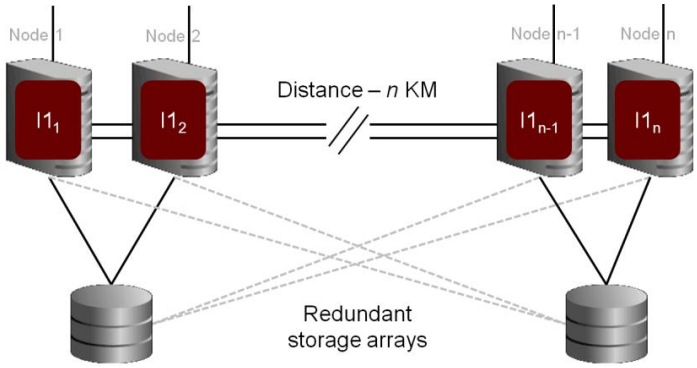

Figure 1: Oracle RAC on Extended Distance Clusters Characteristics

Benefits of Oracle RAC on Extended Distance Clusters

Implementing an Oracle RAC database on a cluster, in which some of the nodes are located in a different site, provides the two main advantages listed below, for which it is typically implemented.

Full utilization of resources

Being able to distribute any and all work (which assumes an active/ active Oracle RAC deployment, not Oracle RAC One Node), including writes, across all nodes, including running a single workload across the whole cluster, allows for the greatest flexibility in terms of resources. Since an Oracle RAC Database is one physical database used across the distance, there is neither any lag in data freshness, nor any requirement for implementing conflict schemes. However, as an Oracle RAC database is one database, there is likeliness that despite the extended architecture a failure might affect both sides at the same time. This is the reason, why Oracle does not recommend an Oracle RAC on Extended Distance Cluster architectures as a replacement for a disaster recovery solution such as Oracle Data Guard.

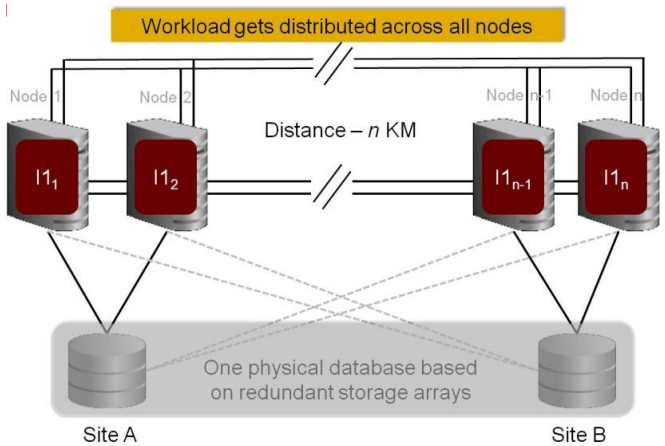

Figure 2: Distributed Workload in an Extended Distance Oracle RAC Cluster

Rapid Recovery

Should one site fail, for example because of a fire at that site, all work can be routed to the remaining site that can take over the databases processing with nearly no interruption for most users connected.

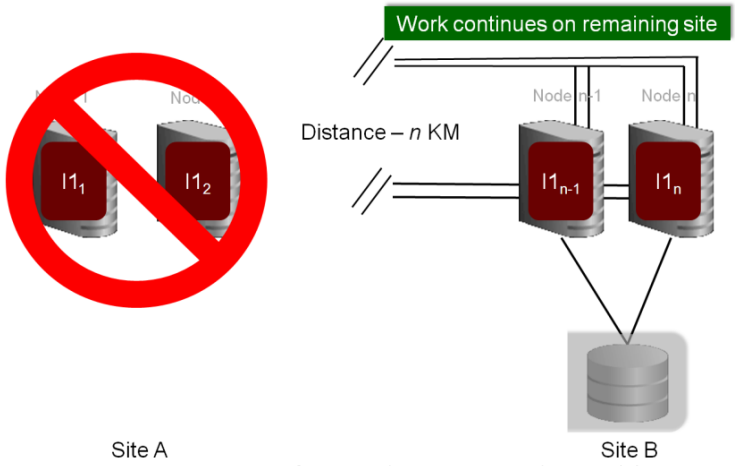

Figure 3: Nearly Uninterrupted Database Operation using Oracle RAC

Protection from a broad range of, but not all disasters

While not a full disaster recovery (DR) solution for the reasons mentioned, an Extended Distance Oracle RAC or Oracle RAC One Node deployment will provide protection from a broad range of disasters. For a full DR protection Oracle recommends deploying an Oracle RAC together with a local and a remote Oracle Data Guard setup as described in the Maximum Availability Architecture (MAA).

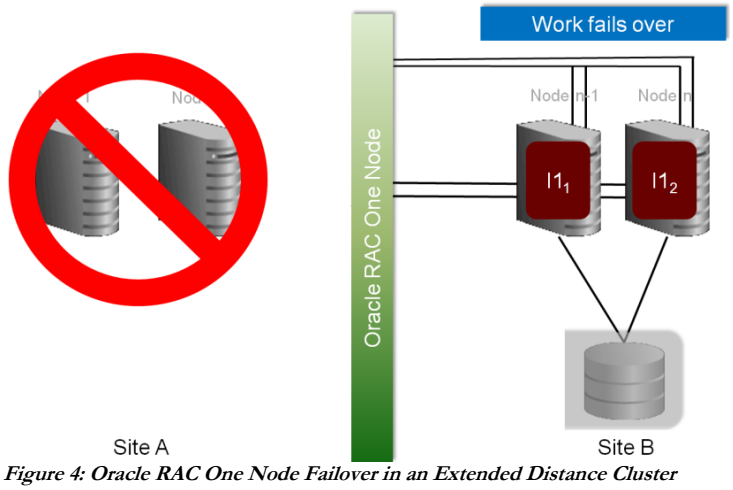

Figure 4: Oracle RAC One Node Failover in an Extended Distance Cluster

Components and Design Considerations

Before discussing the details of an Oracle RAC on Extended Distance Cluster configuration, it is important to understand that an Oracle RAC implementation on such a configuration is not considered a different installation scenario.

This means that the Oracle Universal Installer (OUI), which is used to install (and configure) the main parts of the Oracle RAC stack (Oracle Clusterware, Oracle ASM, Oracle Grid Infrastructure – 11g Release 2 version –, and the Oracle Database Home), cannot be instructed to install those components into an Extended Distance Cluster by simply selecting an installation option.

On the contrary, when planning to install an Oracle RAC on an Extended Distance Cluster, it is crucial to prepare the system so that the distance between the nodes is actually transparent to the Oracle RAC Database. Any network latency optimization discussed in the following is mainly performed to achieve this goal. Regarding the storage setup, similar applies. Ideally, it is transparent to the cluster that the data is mirrored across distances and between storage arrays.

Consequently, Oracle does not certify Oracle RAC on Stretched Cluster configurations. Oracle RAC certification is generally performed on the OS level considering technology requirements, but not on the hardware level. Particular attention needs to be paid to the configuration used for an Extended Distance Oracle RAC cluster as far as network latency and storage response time is concerned.

In general, an Oracle RAC on an Extended Distance Cluster is very similar to an Oracle RAC setup within a single site. To build an Oracle RAC database in an Extended Distance Cluster environment the following needs to be considered:

- One set of nodes is typically placed in site A
- The other set of nodes is typically placed in site B
- A tie breaking voting disk needs to be placed in a third site
    - Special configurations can be used for this site, discussed later
- A host based mirroring solution should be used to host the data on site A and site B and to make sure that the data between the two sites is kept in sync.
    - Disk array based mirroring solutions generally lead to an active / passive configuration between the sites and are therefore not recommended.
- A fast and dedicated connectivity between the nodes and the sites is required for the Oracle RAC inter-instance communication (e.g. a dedicated wavelength on Wavelength Division Multiplexing over Dark Fiber)
    - Using Oracle RAC One Node only would allow for a slightly less powerful connection between the sites, if needed, but is generally not recommended.

Oracle recommends using an Oracle Only Stack for Oracle RAC Extended Distance clusters. Third party cluster solutions can be used in addition, but will complicate the stack and might even add further requirements that could otherwise be avoided. The same applies to third party storage mirroring or cluster solutions. These solutions are generally discouraged and must not be used as a replacement for a full DR solution. Any tests referred to in the following have been performed using a bare metal deployment (physical servers) on an Oracle Only software stack.

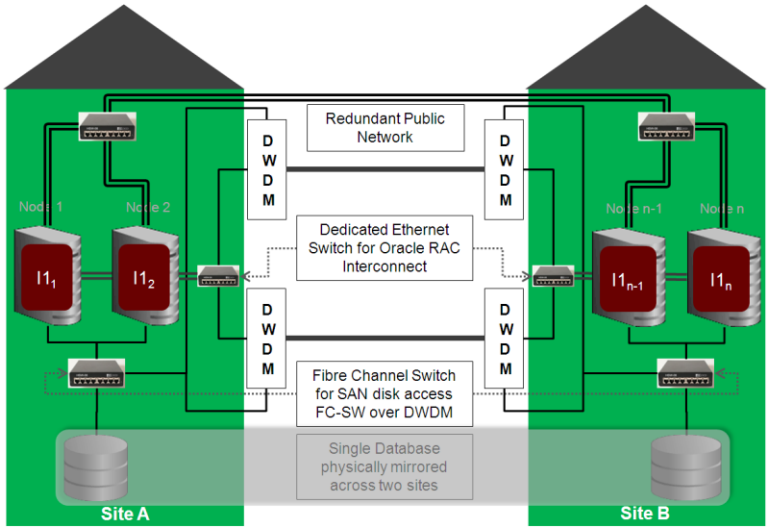

Figure 5: Oracle RAC on Extended Distance Cluster Architecture Overview

Latency and Empirical Performance Results

The Oracle cluster interconnect is a crucial component of the architecture. The cluster interconnect is not only used to ping nodes to ensure their activeness and responsiveness, it is also used to ship data (blocks) between Oracle RAC instances belonging to the same database (Oracle RAC One Node databases do not use data shipping during normal operation). As such, the interconnect can be viewed as a memory channel between the Oracle RAC database instances, requiring a respective throughput and low latency. The network interface used for this purpose should therefore be private, dedicated, and support a low latency protocol.

> **Note**: tests discussed in this paper were performed based on Oracle9i Real Application Clusters (RAC) using traditional, disk based storage as well as dedicated network channels as outlined in figure 5. More recent storage systems utilize flash and provide much lower local IO response times, which leads to greater percentage degradation based on cluster distance. In addition, many improvements to the cache fusion algorithm and Oracle ASM have been made as well as in the networking area, which could improve performance. Individual tests prior to production implementation are therefore highly recommended.

Interconnect latency directly affects the time it takes to access blocks in the cache of remote nodes, and thus it directly affects application scalability and performance. Local interconnect traffic (i.e. between nodes on one site) is generally in the 1-2 ms range and improvements (or degradations) can have a big effect on the scalability the application. I/O latencies tend to be in the 8-15ms range, and are also affected by the additional latencies introduced with distance. This is why figure 5 assumes a separate channel for the storage I/O.

A typical architecture as shown in figure 5 also assumes that various network types are used. Within one site of the cluster, Ethernet or InfiniBand networks are commonly used. To bridge the distance between the sites, DWDM connections have shown to be a common choice. Improvements in network technology in the last years were mainly targeted on the local connectivity. Instead of using multiple 1Gb Ethernet network connections, 10Gb Ethernet connections or InfiniBand connections have become more popular, allowing for less dedication in favor of converging communication types (e.g. public, private, and storage communication, if applicable) on less physical hardware in general.

For Oracle RAC environments, converging communication is less favorable and details will be discussed in subsequent chapters of this paper. It needs to be noted therefore that the performance tests discussed in the following have been performed using a dedicated network connection per communication channel, which especially means that the private interconnect communication was isolated on a private, dedicated network as generally required for Oracle RAC environments.

Based on these tests, the following classification and categorization can be concluded:

- For **distances up to 50km** (cable length) Oracle RAC on Extended Distance Clusters can be considered as a solution regardless of the application used, assuming that an architecture as outlined in figure 5 has been put in place.
    - No explicit tests have been performed to determine the threshold up to which Ethernet connections can be used to bridge distances between sites. However, experience shows that depending on the application, using Ethernet connections up to 5km should be possible. Under certain circumstances, 10km can be spanned.
    - InfiniBand connections are limited a few hundred meters of distance and therefore cannot be used for longer distances. Using repeaters or other components to strengthen the signal and thereby to allow for longer distances using InfiniBand require explicit tests, as those components can have a negative effect on the latency.
    - Oracle RAC One Node databases are subject to the same classification.
- For **distances up to 100km** (cable length) Oracle RAC on Extended Distance Clusters is a possible solution, but performance tests using the application(s) to be deployed on the system should be performed. Again, this still assumes that an architecture as outlined in figure 5 has been put in place.
    - Oracle RAC One Node databases require less performance tests over this distance. For Oracle RAC One Node databases, it is fair to assume that Extended Distance Clusters over this distance are generally a considerable solution regardless of the application used, depending on IO performance, and assuming that an architecture as outlined in figure 5 is used.

Note: There is no magic barrier for the distance; latency just keeps increasing. Write intensive applications are generally more affected by distance then read intensive applications.

Various partners have tested Oracle RAC on Extended Distance Clusters. These tests include tests performed by HP and Oracle over distances of 0, 25, 50, and 100 kms. Tests were also performed by the EMEA Oracle/IBM Joint Solution Center over distances of 0, 5, and 20 kms. In addition, Symantec Veritas performed tests over distances of 0, 20, 40 and 80kms. All tests were based on an OLTP application test. Some tests included unit tests of individual components. The unit test results from the HP/Oracle testing will be used to illustrate what happens on the component level.

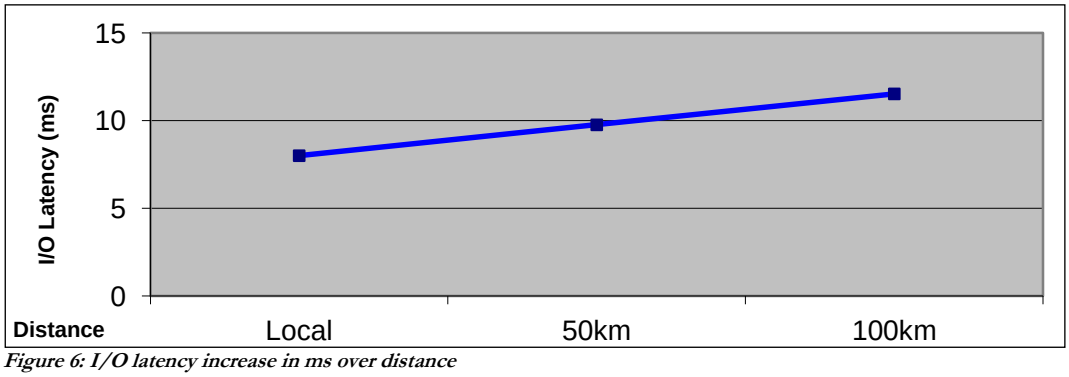

Figure 6: I/O latency increase in ms over distance

Figure 6 shows the effects of distance on I/O latency with buffer-to-buffer credits (BBC). BBC allows for a greater number of unacknowledged packets on the wire, thus allows for greater parallelism in the mirroring process. As distances increase, especially with high traffic volumes, these BBC can make a huge difference. For example, when the tests above where run without the additional BBC, I/O latency at 100km was 120-270% greater than local I/O latency, instead of 43% as shown in the chart above. These numbers are consistent with the results from the Oracle/IBM testing which had 20-24% throughput degradation on I/O Unit tests at 20 km when BBC were not used.

Interconnect Traffic Unit Test Results

Tests at both, high and low load levels, and with one or two interconnect, have shown that there is a latency 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.

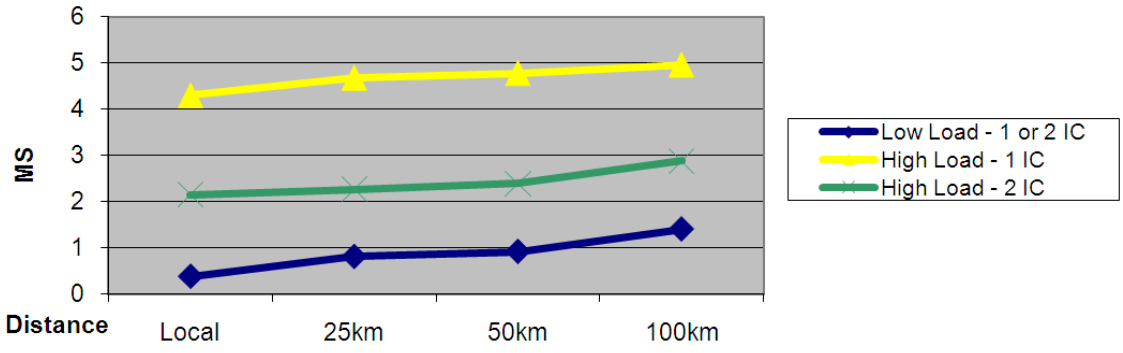

Figure 7: Interconnect Unit Traffic Test Results

Overall Application Impact

Unit tests are useful, but in reality, it is more important how a "real-world" application reacts to the increased latencies induced by distance. Having three independent sets of tests provide a more complete picture than each individual unit test. A summary of each test is provided and full details can be found in the appendix of this paper, ordered by the respective vendor who performed the tests.

> **Note**: tests discussed in this paper were performed based on Oracle9i Real Application Clusters (RAC) using traditional, disk based storage as well as dedicated network channels as outlined in figure 5. More recent storage systems utilize flash and provide much lower local IO response times, which leads to greater percentage degradation based on cluster distance. In addition, many improvements to the cache fusion algorithm and Oracle ASM have been made as well as in the networking area, which could improve performance. Individual tests prior to production implementation are therefore highly recommended.

The following shows the overall performance impact on applications due to distance measured as a percentage of local performance.

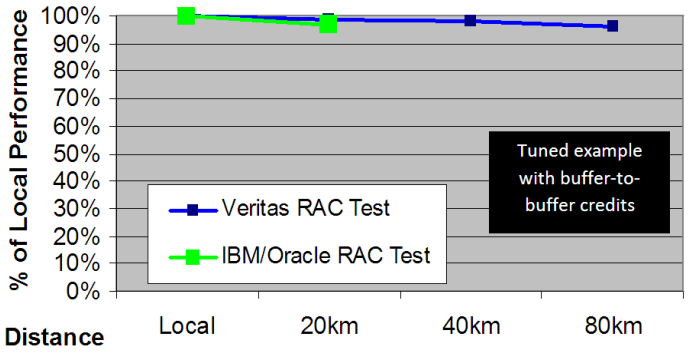

Figure 8: Tuned example – "overall performance impact on applications due to distance"

The IBM/Oracle tests were performed using a representative workload, which was set up by running the SwingBench workload with proper use of BBC. These tests showed a 1% degradation for read transactions and 2-8% degradation for most write transactions at 20 km distance. The average single transaction encountered 2% degradation.

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

Other tests were performed without using buffer-to-buffer credits nor directing reads to the local disks first. Combined with a very contentious application, this resulted in some impact at 25 m (10%), but significant degradation at 50km-100km. Further testing would be needed to determine why the 50 & 100km numbers are similar, but the 0, 25 and 100km numbers form a very clear linear slope. With an appropriate BBC configuration these numbers would be expected to significantly improve and be closer to the Veritas and Oracle/IBM numbers.

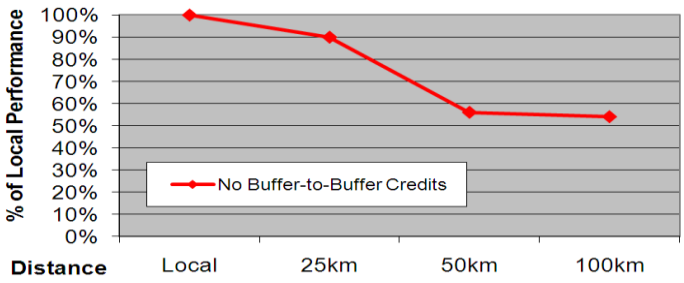

Figure 9: Untuned example – "overall performance impact on applications due to distance"

Real-life applications are expected to follow the IBM/Oracle & Veritas examples demonstrated earlier – at the best. In reality they will probably have more interconnect traffic and thus suffer slightly more from the distance than measured during the tests performed. Each of these results is for a particular application with a particular setup. Other applications will be affected differently depending on their access pattern. the basic idea is that as distance increases, IO and Cache Fusion message latency increases. The limitations come from a combination of the inefficiencies and latency added by each switch, router or hub a message must be handled by.[2] As previously stated, Dark Fiber can be used to achieve connections greater than 10km without repeaters.

Conclusion and a note on Oracle RAC One Node

While there is no magic barrier to how far Oracle 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 Oracle RAC at a greater distance, performance tests using the specific application are recommended.

Oracle RAC One Node does not suffer from an increase in distance between the nodes per se, as under normal operations, it would only run one instance at a time and hence Cache Fusion communication between instances will not occur. Exceptions apply during the time of an Online Database Relocation and in case Oracle RAC One Node is converted to full Oracle RAC. However, storage connectivity and distance should still be considered, especially if Host Based Mirroring is used to mirror the data between sites. **This is the reason why this paper considers Oracle RAC One Node subject to the same distance limitations**.

Given these numbers, it can be concluded that Oracle RAC using Extended Distance Clusters at distances under 50km performs acceptably in general, while performance tests are suggested for distances over 50km, up to 100km. Implementations of more than 100km are not recommended.

[2] To avoid large latencies the configuration should only have a switch at each site, WDM devices, and a direct uninterrupted physical connection in between. No additional routed network.

Implementation Details and Behavioral Background

Oracle RAC on Extended Distance Clusters does not constitute a different type of cluster, neither is there a special installation option that one can choose from. This means that as far as the configuration of the system is concerned, the main goal is to hide the fact that Oracle RAC is now operating over distance. On the other hand, this means that the basic configuration remains the same, including its requirements. Attention must be paid when configuring the network and storage connectivity for Extended Distance Oracle RAC environments.

Network Connectivity

For Oracle RAC environments it has always been stated that the network used as the Interconnect, the SAN connectivity and the public network are ideally kept on separate *dedicated* channels, each providing simple redundancy for full high availability. Redundant connections must not share the same Dark Fiber (if used), switch, path, or even building entrances. These channels should not be shared by any other communication channel or link. It needs to be kept in mind that cables can be cut, so physically separated paths should be part of the overall network design, which is outlined in figure 10 below.

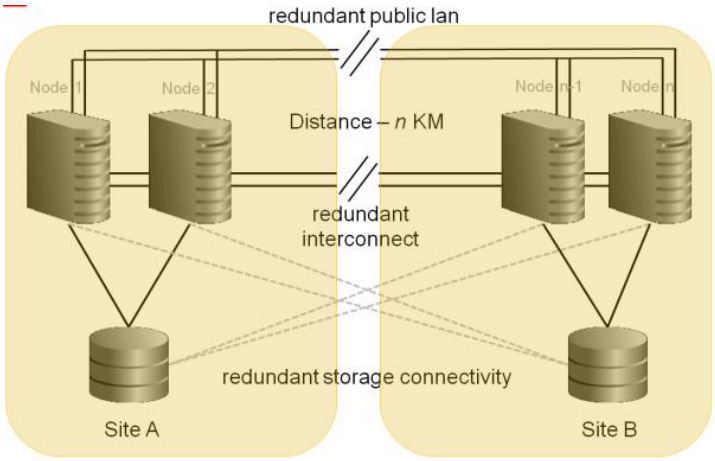

Figure 10: Overall Network Design

Enhancements in network technology over the last years along with an increasing cost pressure on IT departments have made it more and more interesting as well as common to use shared network components and converge different types of network communication on the same physical network, mostly preserving redundancy in the same way.

Typical examples include, but are not limited to, using 2* 10GB Ethernet connections for public network communication as well as storage access, isolating the two using VLANs[3]. Other combinations and even network virtualization technologies could be considered in this context.

For Oracle RAC on Extended Distance Clusters the same restrictions as for the non-Stretched version apply. In addition, it needs to be noted that Oracle has not tested and will not test such configurations in extended environments. Using converged networks as well as VLANs in Extended Distance Oracle RAC clusters will have an impact on the scalability as well as the stability of the system, as failures of those networks will affect more than one communication channel at the same time.

While failure scenarios can be considered, predicted as well as documented, saturation, interferences, or general scalability issues on converged networks are harder to foresee. The distance between the nodes in an Extended Distance Clusters and the thereby potentially increased latency on the network can further increase the likeliness for issues. If converged network configurations cannot be avoided on Extended Distance Oracle RAC clusters, performance and reliability tests should be performed.

Consequently, direct, non-shared, point-to-point cables (see latency discussion in previous chapter) are ideal. Normal SAN and Ethernet Connections are limited to 10km for point to point communication. WDM over Dark Fibre networks allow these to be much further apart while still maintaining the low latency of a direct connection. The disadvantage of Dark Fibre networks on the other hand is costs. This is why they are typically only an option if they already exist between the two sites.

Multiple Subnet Support in Stretched Clusters

The difference between a multiple subnet support and a separate subnet support lies in the way the respective subnets are distributed across the (nodes in the) cluster. "Multiple Subnet Support" assumes that the subnets are laid out horizontally across the nodes in the cluster so that all the nodes can be accessed by multiple subnets at any time. This setup is fully supported with Oracle RAC 12c.

A "Separate Subnet Support", especially in Extended Distance Cluster environment, assumes that the two sites will reside in different subnets so that subsets of nodes (one set in site A, one in site B typically) could be accessed using different subnets. Such separated subnet support is not supported for either the public lan or the Oracle private interconnect, not even with Oracle RAC 12c.

[3] For general guidelines on using VLANs with Oracle RAC, see http://www.oracle.com/technetwork/products/clusterware/overview/interconnect-vlan-06072012-1657506.pdf

Depending on the version of Oracle RAC, multiple subnets are supported for the Oracle private interconnect. While versions prior to Oracle RAC and Oracle Grid Infrastructure 11g Rel. 2, Patch Set One (11.2.0.2) would only allow for one subnet on the interconnect, this restriction has been lifted with this version, assuming that network redundancy is enabled using the Redundant Interconnect Usage Feature available in 11.2.0.2 and later versions.

For database versions prior to Oracle RAC and Oracle Grid Infrastructure 11g Release 2, Patch Set One (11.2.0.2) multiple subnets for the interconnect are not supported, since these versions cannot use the Redundant Interconnect Usage Feature.[4]

For the public network "Multiple Subnet Support" has been introduced with Oracle RAC and Oracle Grid Infrastructure 11g Release 2 (11.2.0.1). The idea behind this feature was to allow applications other than databases, managed in a consolidated cluster environment, to be accessed using Virtual IPs (VIPs) of another subnet (different from the one used for the database). The implementation of this feature can be extended so that even databases can be accessed via various subnets at the same time.

With Oracle RAC 12c, multiple subnet support, including multiple SCAN support (with one SCAN per subnet) for the public lan and thereby client connectivity to the database via multiple subnets is fully supported. [5]

Storage Connectivity and Setup

Oracle RAC on Extended Distance Clusters by definition has multiple active instances on nodes in different locations (which the exception of Oracle RAC One Node). For availability reasons the data needs to be located at least in two different sites and therefore one needs to look at alternatives for mirroring the storage in addition to a full disaster recovery solution.

When looking at the various options available, one will find that there are really only two principle types of storage mirroring solutions one can choose from. All other solutions can be considered a subset or a certain incarnation of one of those two principle types, which are:

1. Host Based Mirroring
2. Array Based Mirroring

[4] More Information on the Redundant Interconnect Usage Feature and its requirements as well as implications can be found in the Oracle RAC Documentation and My Oracle Support

[5] For more information, see http://www.oracle.com/technetwork/products/clustering/overview/scan-129069.pdf

For the purpose of this paper, a solution is considered a host based mirroring solution as soon as certain software is deployed on the servers of the cluster and used to mirror data across storage arrays located in different sites.

Oracle recommends using host based mirroring over any other storage mirroring or clustering solution. Oracle's Automatic Storage Management (ASM) is the host based mirroring solution recommended by Oracle, as it provides all means necessary to manage storage in an Extended Oracle RAC environment.

For the purpose of this paper, a solution is considered an array based mirroring solution, if some form of intelligence (typically software) deployed on the storage array level is used to mirror (often also referred to as to "replicate") data between them.

With the introduction of storage fabrics, storage clustering or "cloud storage" solutions, and other, "intelligent" storage solutions a clear distinction between a host based and an array based mirroring solution is often hard to draw. In general, those are all storage solutions and would therefore need to provide certain high availability (HA) features, which is the main aspect to be considered when it comes to using such a solution in Stretched Oracle RAC clusters.

Once the HA capabilities of a particular solution have been evaluated, one will often find that the respective solution is a subset or a certain incarnation of one of the two base types and can therefore be considered as such. None of these, solutions, however, can replace a full disaster recovery solution such as Oracle Data Guard, as the base assumption is that an enclosed system is used, in which the data is simply mirrored, which could lead to corruptions being replicated within the system (depending on the nature of the corruption)

Host Based Mirroring (Active / Active Storage)

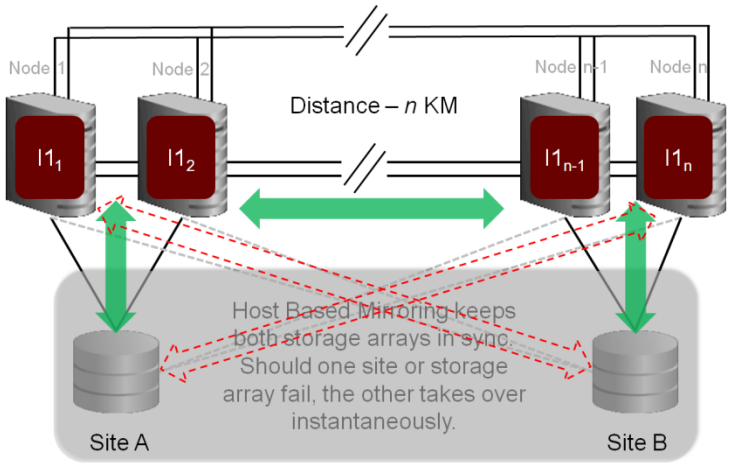

Figure 11: Host Based Mirroring Typical Setup

When using Host Based Mirroring, a typical setup can be outlined as follows:

- Use at least two SAN/FC storage subsystems, one in each site.
  - Cabling as illustrated in figure 11 – note the cross-cabling.
  - If only two sites are used, use one storage array on each site, which is the assumed standard for this paper.
- Standard, cluster-aware, host based mirroring software is deployed all nodes in the cluster, configured to work across both storage arrays.
- The setup is as such that writes are propagated at the OS level to both disk arrays, across all disks used by the database.
- The intention is to make these disks appear as one single set of disks independent of their location.
- These kinds of Logical Volume Managers (LVM) are typically tightly integrated with a cluster software solution, which needs to be certified with Oracle Clusterware to be used for an Oracle RAC on extended distance cluster setup.

Host Based Mirroring is recommended for Oracle RAC on Extended Distance Clusters as well as for local implementations (non-stretched) as it is the only solution that has proven to meet all high availability requirements with respect to mirroring data across more than one storage array without causing an active / passive setup or any other form of manual intervention. Oracle ASM is the recommended Host Based Mirroring solution for Oracle Databases[6]. Nevertheless, a Host Based Mirroring solution cannot replace a full DR solution.

Array Based Mirroring (Active / Passive Storage)

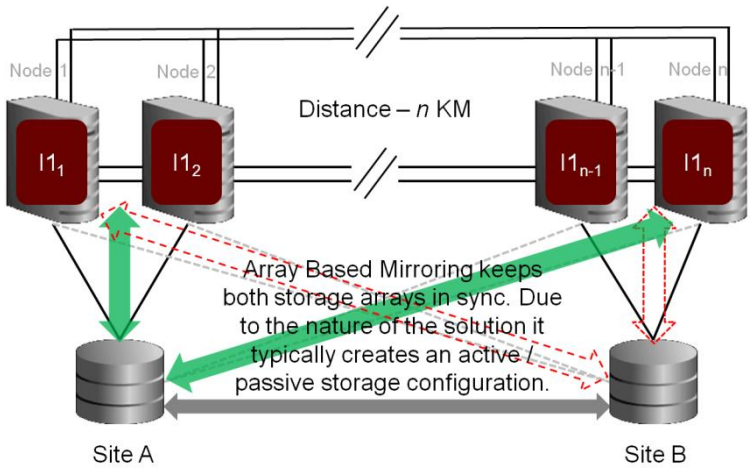

Figure 12: Array Based Mirroring Typical Setup

[6] For more information on using Oracle ASM for host based mirroring, see page 24f in this paper.

When using Array Based Mirroring solutions, it needs to be noted that most array based mirroring solutions create an active / passive configuration, as illustrated in figure 12 by the fact that the servers in "Site B" do not have an active link (green arrow) to the storage in "Site B".

The reason being is that most array based mirroring solutions typically can only activate the disks used to receive the mirrored data in "Site B" upon failure of "Site A". Some array based mirroring solutions provide processes to observe the sites and to initiate an automatic failover once an issue is detected while staling I/O for the time of the activation. However, these operations typically exceed the time permitted to interrupt I/O for the Oracle RAC cluster implementation and are in general subject to further limitations considering site failures[7].

Such configurations will therefore result in a full cluster outage due to these operations. This is the reason, why Oracle does not recommend array based mirroring solutions for an extended setup.

In addition, performance can be impacted in such setups due to additional work in the storage array because of mirroring, but more importantly due to the fact that IO from the secondary site has to cross the 'distance' 4 times[8] before control can be returned.

Despite some improvements in this area, there has not been a single array based mirroring solution that provides the same level of storage high availability as a host based mirroring solution, while neither solution provides a full DR protection for the database. Most array based mirroring solutions will also fall short, if tested against all sort of failure scenarios. Therefore, Oracle discourages using third party storage based mirroring solutions for either Extended Oracle RAC environments or disaster recovery.

Storage Clustering Solutions and other Storage Fabrics

For the purpose of this paper, storage fabrics, storage clustering or "cloud storage" solutions as well as other, "intelligent" storage solutions are used synonymously and are mainly characterized by the fact that some sort of intelligence is used to represent the storage under a single entry point to the servers.

[7] See chapter "Node Management in Extended Distance Clusters" for more information

[8] The I/O will have to travel from the Secondary host to primary storage, then from the primary storage to secondary storage, then from the secondary storage to primary storage, and finally from the primary storage to secondary host. All need to be synched to ensure no data loss.

Technically, this is typically achieved by using some kind of intelligent "storage head" as illustrated in figure 13.Behind this "storage head", storage arrays can be allocated as required and as the storage solution supports. The intelligence provided by this head can be of various levels and depends on the individual solution. Typically, these heads will be responsible for mirroring, which can be implemented in various ways in such a setup. For the purpose of this paper, such storage clustering solutions are therefore considered a more sophisticated implement of an array based mirroring solution.

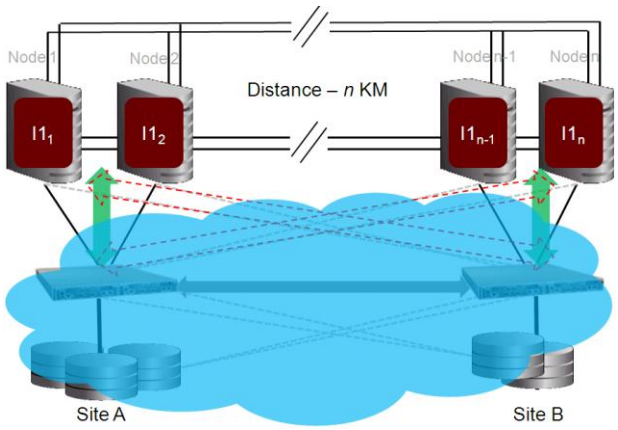

Figure 13: Cloud Storage / Storage Fabrics - An Illustration

Obviously, an Extended Distance Cluster demands that the "intelligent storage head" is available in both (all) sites, as a failure of this device would lead to an inaccessibility of the storage if not kept redundant. Most of these solutions allow for redundant setups.

In addition, the storage solution needs to be able to understand that data needs to be mirrored across sites. This is especially important if more than one storage array is used in one location, as just mirroring across arrays may not be sufficient in this case. The system needs to be configured in such a way that mirroring is performed across arrays in different sites or locations, as otherwise site failures would be unrecoverable.

The specific implementation of the cloud storage is irrelevant for its use with Oracle RAC on Extended Distance Clusters as long as it fulfills the HA requirements mentioned and is able to cover certain failure scenarios, especially site failures.

At the moment of writing, no cloud or storage fabric system on the market provides a sustaining solution that will guarantee a non-interruptive operation under an Extended Oracle RAC setup[9].

[9] See chapter "Node Management in Extended Distance Clusters" for more information

Therefore and considering the fact that these solutions cannot provide full Disaster Recovery support for the Oracle Database in general, Oracle discourages the use of any third party storage mirroring, replication or storage clustering solution in combination with an Extended Oracle RAC setup.

Why One Single Storage Solution in One Location is Insufficient

While it is possible to implement Oracle RAC on Extended Distance Clusters with a storage array in only one location / on only one site, it is not advisable to use such a configuration. The main reason is that should the site with the single storage array fail, storage is no longer available to any surviving node and hence the whole cluster becomes unavailable. This defeats the purpose of having Oracle RAC nodes in different locations.

Summary – Storage Configuration and Support

Oracle does not certify Oracle RAC on Extended Distance Clusters. The current Oracle RAC Certification and Technology Matrixes apply to Oracle RAC on Extended Distance Clusters in the same way. Of course, the information provided needs to be adapted to the special use case (e.g.: using directly attached storage in an Extended Distance Cluster does not make sense, regardless of whether or not this cluster is used for Oracle RAC).

It also needs to be noted that Oracle RAC certification is hardware independent. That means, storage that meets the technology requirements as listed in the Certification and Technology Matrix can generally be used with Oracle RAC on Extended Distance Clusters (with some exceptions as stated).

Plenty of storage solutions can therefore be used with Oracle RAC on Extended Distance Clusters, but Oracle discourages the use of any third party storage mirroring, replication or storage clustering solution in combination with an Extended Oracle RAC setup.

Oracle ASM provides all the means necessary to manage storage for an Extended Oracle RAC setup. In addition, Oracle Data Guard or Oracle GoldenGate should be used to provide full disaster recovery protection for the Oracle Database.

Node Management in Extended Distance Clusters

There is no special configuration option to enable an Oracle RAC on Extended Distance Clusters. Hence, standard Oracle RAC behavior with respect to node management (fencing) to prevent split-brain situations as well as data corruption remains the same whether or not the Oracle RAC cluster is installed locally or on an Extended Distance Cluster[10].

The physical location of the cluster quorum (in case of Oracle Clusterware known as the Voting Disks or Files), however, has a bigger impact on the design of an extended distance cluster than it would on a local cluster. For local clusters, details about the cluster quorum mechanism do not need to be considered in general, as the cluster software is designed to make the process of deploying and using the cluster quorum entities somewhat "fool proof" in order to prevent split-brain scenarios[11] as well as to ensure that the biggest available sub-cluster remains operational, should a system failure occur. Once the cluster nodes are physically dispersed or multiple storage arrays hosting a file based cluster quorum (a.k.a. Voting File) are used, the initial configuration is no longer that simple.

That is why cluster solutions supporting extended distance cluster will typically also support using a tie-breaker mechanism located in a third location. For a typical Oracle RAC on Extended Distance cluster with 2 sites (site A and site B), this allows both sites to be equal while the third location acts as an arbiter should either site fail or connectivity be lost between them. Figure 14 illustrates such a setup.

In an extended Oracle RAC setup, Oracle therefore recommends to host three Oracle Clusterware managed Voting Files; one in each data center hosting the Oracle RAC servers and one in a third location, independent of those data centers.

As most customers may not have the ability to host a third SAN in the independent third location, Oracle allows accessing the third Voting File using standard NFS as described in this paper: "*Using standard NFS to support a third voting file for extended cluster configurations*"[12]

---

[10] For a general overview of Oracle Clusterware based node management see: http://www.oracle.com/technetwork/database/clusterware/overview/oracle-clusterware-11grel2-owp-1-129843.pdf

[11] A split brain occurs when two portions of the cluster stop coordinating and start acting independently. This can easily lead to data corruption. Therefore, clustering software is carefully written to prevent split-brain situations from occurring. Should nodes start to misbehave or to stop communicating with each other, some nodes will be evicted from the cluster until normal operation is resumed and while ensuring that the biggest possible sub-cluster survives.

[12] For more information, see: http://www.oracle.com/technetwork/products/clusterware/overview/grid-infra-thirdvoteonnfs-131158.pdf

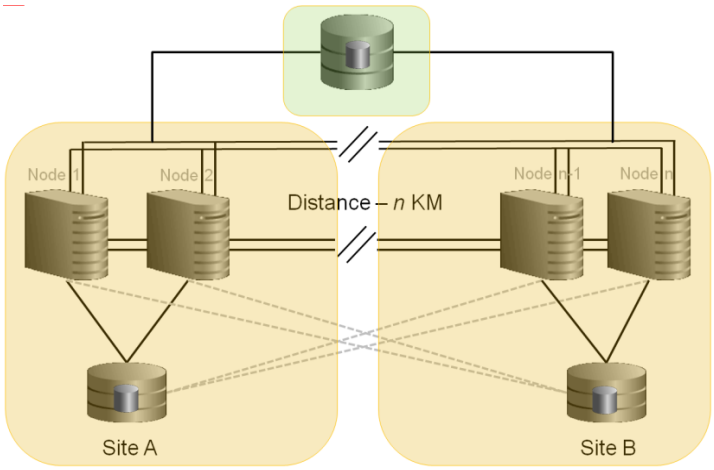

Figure 14: Extended Distance Cluster with Quorum File / Voting File in Third Location

This setup is supported as mentioned in the paper and for extended Oracle RAC cluster configurations only. It is assumed that the third location is not used for data storage, just for hosting a third Voting File in above mentioned configuration. From a cluster perspective all Voting Files are alike. The third location based Voting File cannot be distinguished and will therefore not be treated differently.

Given these assumptions, the third location does not need to observe the same network requirements as the inter-data center connectivity, which needs to meet Oracle RAC standards as outlined in the Oracle RAC documentation or in this paper: "*Oracle RAC and Oracle Clusterware Interconnect VLAN Deployment Considerations*"[13]

Oracle has not tested the impact of distance on the third location based Voting File, as it is generally assumed that the third location is ideally in similar distance to the data centers as the two data centers are located to each other. However, customers can choose any distance for the third location, as long as the accessibility to the third Voting File is ensured and considering that increasing distance to the third location from the data centers increases the risk for failures on the path that not only cause delays, but also other failures that might either be recognized as IO errors or simply as a (NFS) freeze.

Minimum network requirements for the connectivity to the third location can therefore be derived from the access patterns to Voting Files in an Oracle cluster. In general, Oracle Clusterware accesses a Voting File every second for read and write with less than a kilobyte of data read or written. An acknowledgement of the write I/O needs to be received in 200 seconds under normal operations (long disk timeout) and 27 seconds during a reconfiguration in the cluster (short disk timeout).

[13] For general guidelines on using VLANs with Oracle RAC, see http://www.oracle.com/technetwork/products/clusterware/overview/interconnect-vlan-06072012-1657506.pdf

In order to meet those response time requirements (latencies) for accessing the third Voting File hosted on standard NFS or using iSCSI based connectivity, Oracle recommends setting up the system to meet half of the lower latency requirement in average. In other words, the connectivity to the third location should ensure that the Voting File write I/O can be acknowledged in 27/2 seconds (approx. 14 seconds), providing a minimum average throughput of at least 128 Kbps.

Storage Clustering and the Dueling Cluster Framework Issue

While a third location for a third Voting File is crucial for Oracle Clusterware, it needs to be noted that any clustering solution requires a tie-breaker mechanism to resolve split brain-like situations resulting from a communication failure between the sites of the extended distance setup. This is a simple necessity when operating in such configurations and various solutions will use different approaches.

The problem is that two independent clustering solutions (i.e. solutions that are not tightly integrated) can make autonomous decisions. This fact applies when using a storage clustering solution as well as for Host Based Mirroring solutions that rely on cluster solutions that are not integrated with Oracle Clusterware. The latter setup is not supported therefore. In case of Oracle Clusterware, this issue is also referred to as the dueling cluster framework problem.

Using a storage clustering solution under an Oracle RAC in an Extended Distance Cluster with one or more Voting File(s) (from Oracle Clusterware perspective) hosted on the storage fabric is subject to the dueling cluster framework problem, which in short can be described as follows.

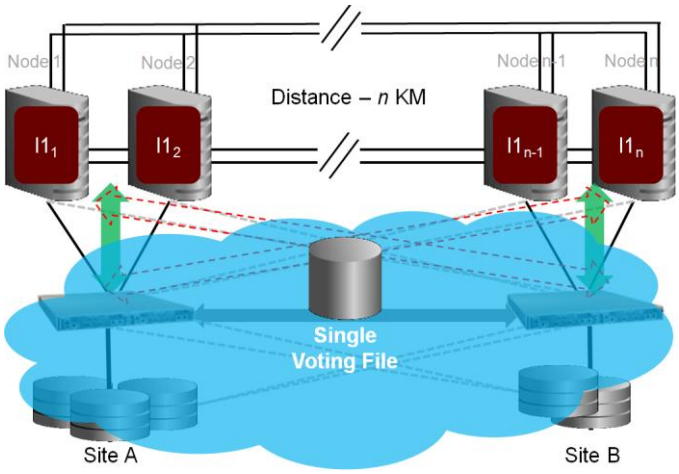

Figure 15: Single Voting File used on Storage Cluster

Oracle Clusterware relies on the accessibility of the Voting File(s). If a node in the cluster cannot access the majority of the Voting File(s), the node(s) to which this applies is (are) immediately removed from the cluster (evicted / fenced). If only one Voting File is used, accessibility to this one file needs to be ensured in order to prevent a reboot of individual nodes or even the whole cluster.

Certain storage clustering solutions or storage fabric solutions claim to ensure accessibly to the one Voting File used for Oracle Clusterware at any point in time, while they themselves are subject to situations, in which they need to use a tie-breaker to determine which a data center site needs to be taken down in order to ensure data consistency.

In this case, the storage clustering solution will make a decision and prevent access to one site in the dual-site setup as shown in figure 15 as soon as a communication failure between the sites occurs (if a site goes down completely, no decision needs to be made). The result is that the nodes in the site that has been determined to be taken down will immediately lose access to the Voting File and hence will be evicted from the cluster.

Oracle Clusterware continues to independently monitor the communication between the sites and will perform corrective actions as required in order to ensure data consistency as it would in a local Oracle RAC setup. Due to the fact that the clustering solutions used in the stack are not integrated and thus operate on the stack independently, **autonomous decisions are made and the clustering solutions can decide to evict opposite sites in a dual-site setup, resulting in a complete cluster outage**.

Some storage cluster or storage fabric vendors claim that they mitigate or even solve the problem by allowing manual adjustment to the fencing algorithm so that the storage cluster solution would make the same decision as Oracle Clusterware assuming that both clusters have the same view on the system.

As this approach requires manual adjustment and potentially ongoing re-adjustment over the lifetime of a cluster and the algorithm that Oracle Clusterware uses to determine which node(s) are to be evicted from the cluster is subject to change without notice, Oracle discourages the use of storage clustering solutions under an Oracle RAC on an Extended Distance Cluster.

Oracle-Only-Stack Advantages

The Oracle RAC stack is designed to optimally operate in Extended Distance Cluster setups. Various components in the Oracle RAC stack have been enhanced to facilitate its use in an Extended Distance Cluster. In addition, Oracle Clusterware (part of Oracle Grid Infrastructure) has been tightly integrated with Oracle Data Guard and GoldenGate in order to provide a complete disaster recovery solution, which Oracle RAC on Extended Distance Clusters alone cannot provide.

Oracle Clusterware

Oracle Clusterware has been enhanced in many ways to provide easier setup and management for Oracle RAC on Extended Distance Cluster setups. Some of the enhancements are listed below.

- The Oracle Cluster Registry (OCR) can be stored in multiple locations.
    - Starting with Oracle Clusterware 11g Release 2, the OCR can be stored in multiple disk groups in addition to being mirrored within the same disk group for maximum protection as well as simplified recovery after a cluster outage.[14]
- Oracle Clusterware supports multiple Voting Files in multiple locations
    - Starting with Oracle Clusterware 11g Release 2, Voting Files can be stored in Oracle Automatic Storage Management (ASM), thereby immediately benefitting from the mirroring capabilities of Oracle ASM, which are essential to operate an Oracle RAC on Extended Distance Cluster setups.
    - Oracle ASM has been enhanced in this context to understand that a third location for a third Voting File based on a standard NFS destination is used as a quorum device.[15]
- Oracle Clusterware provides an integrated mechanism to utilize redundant networks for the interconnect, known as the Redundant Interconnect Usage Feature, often referred to as the HAIP-feature. This feature is particularly interesting for Oracle RAC on Extended Distance Cluster setups, as it foresees redundancy across subnets.
- While "separated subnets per data center / site" remains unsupported, Oracle RAC 12c fully supports the use of multiple subnets in the cluster.

Oracle Automatic Storage Management (ASM)

Traditionally, Oracle ASM's built in mirroring capabilities have been used to efficiently mirror all database files across both sites in a dual-site Oracle RAC on Extended Distance Cluster setup. Storage at each site must be set up as separate failure groups and use ASM mirroring to ensure that at least one copy of the data is stored in each site.

---

[14] For more information, see My Oracle Support note 106293.1:
"How to restore ASM based OCR after complete loss of the CRS diskgroup on Linux/Unix systems"

[15] For more information, see:
http://www.oracle.com/technetwork/products/clusterware/overview/grid-infra-thirdvoteonnfs-131158.pdf

Using Oracle ASM based mirroring in an extended Oracle RAC configuration nearly eliminates the risk of certain corruptions spreading across both array / sites and thereby makes it the preferred and recommended solution over other mirroring technologies; especially storage mirroring solutions that work on the bits-and-bytes level with no understanding of database block formats and consistency.

A unique benefit of Oracle ASM based mirroring is that the database instance is aware of the mirroring. For more detailed information, see the chapter on "*Oracle ASM Recovery from Read and Write I/O Errors*"[16] in the Oracle documentation, which states: "*For many types of logical corruptions such as a bad checksum or incorrect System Change Number (SCN), the database instance proceeds through the mirror side looking for valid content and proceeds without errors. If the process in the database that encountered the read can obtain the appropriate locks to ensure data consistency, it writes the correct data to all mirror sides.*"

Using Oracle ASM based mirroring in addition to database inherent block corruption detection and prevention mechanisms, such as DB_BLOCK_CHECKSUM[17] and DB_BLOCK_CHECKING[18], enabled on an appropriate level (see documentation for details), protects from a wide range of logical as well as physical corruptions, especially in an extended Oracle RAC environment, in which data is mirrored across storage arrays.

Combining those technologies with Oracle Flashback Technologies to protect from Database-wide logical corruptions caused by human or application errors means that an (extended) Oracle RAC system provides a fairly comprehensive protection against most corruptions and failures.

However, as an Oracle RAC system is still a tightly coupled system using a shared data approach, Oracle's recommendation to achieve the most comprehensive data corruption prevention and detection is to use Oracle Data Guard with physical standby databases in addition[19].

---

[16] Oracle ASM Recovery from Read and Write I/O Errors:
http://docs.oracle.com/cd/E11882_01/server.112/e18951/asmdiskgrps.htm#OSTMG94132

[17] DB_BLOCK_CHECKSUM:
http://docs.oracle.com/cd/E11882_01/server.112/e40402/initparams049.htm

[18] DB_BLOCK_CHECKING:
http://docs.oracle.com/cd/E18283_01/server.112/e17110/initparams046.htm

[19] For more information on "Preventing, Detecting, and Repairing Block Corruption: Oracle Database 11g", see: http://www.oracle.com/technetwork/database/focus-areas/availability/maa-datacorruption-bestpractices-396464.pdf

From Oracle Database 11g onwards several enhancements are available with ASM to specifically provide better support for Oracle RAC on Extended Distance Clusters:

- Partial Re-silvering: With the fast re-sync option, full re-silvering is no longer required for ASM mirrors should a temporary loss of connectivity between the sites occur. The amount of time for which re-silvering information is maintained is configurable
- Local Reads: IO read requests can be configured via the `ASM_PREFERRED_READ_FAILURE_GROUPS` parameter to go to the local mirror instead of going to any available mirror. Reading from both mirrors is better for shorter distances as all IO cycles are fully utilized. Local mirrors are better for further distances as all reads are satisfied locally.
- Starting with Oracle ASM 12c, Oracle Flex ASM can be used in Extended Oracle RAC environments, further enhancing the availability of database instances across the cluster.
    - As a general configuration recommendation when using Oracle Flex ASM, the number of Oracle Flex instances should be increased to span all nodes in the cluster (using *srvctl modify asm –count ALL*).

Conclusion

Oracle RAC on Extended Distance Clusters is an attractive alternative architecture that allows for scalability, rapid availability, and even for some very limited disaster recovery protection with all nodes being fully active in both (all) sites.

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

While this architecture provides a slightly better level of High Availability than a local Oracle RAC configuration, 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 an Oracle RAC and Oracle Data Guard combination provides.

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

Appendix A: Table of Figures

Oracle RAC and Oracle RAC One Node on Extended Distance (Stretched) Clusters
October 2013
Author: Markus Michalewicz
Original version: Erik Peterson et al.

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

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together