ORACLE DATABASE 12c

Oracle Clusterware 12c Release 2 Technical Overview

ORACLE WHITE PAPER | MARCH 2017

ORACLE

Table of Contents

Introduction

Oracle Clusterware enables the clustering of otherwise independent servers so that they co-operate as a single system. As a cluster, these servers then provide the integrated foundation upon which Oracle Real Application Cluster (RAC) databases and user applications can take advantage for high availability and scalability.

The Cluster of servers is coordinated via the Oracle Clusterware, with cluster resources made available as required in support of the high availability requirements of the Oracle RAC databases and applications running on the cluster, on one or more of the clustered servers, or nodes. Introduced in Oracle 10g Release 1, Oracle Clusterware has evolved and broadened its capabilities to meet the demand for a more versatile and more capable infrastructure.

The introduction of Oracle Clusterware 12c Release 2 further enriches the already broad array of available features and functions. These enrichments include new deployment architectural options, advances in the management of Clusterware resources, the control of failure scenarios, diagnosis of those failures and the reduction of operational requirements.

Oracle Clusterware 12c Release 2 introduces the following new features:

- *Cluster Domains* – a new cluster architecture aimed at multi-cluster deployments, consolidating the management and storage functions common to those clusters
- *Node Weighting for Split Brain Resolution* – an intelligent approach to resolving split brain conditions, emphasizing survival based upon workload viability and critical resources
- *Cluster Resource Groups* – easing the management of related resources in support of application high availability and management in clustered environments
- *Cluster Activity Log* – a log of cluster wide events that can be queried to review activities across the cluster, benefiting diagnostic efforts and operations
- *Highly Available GNS* – enhancing GNS server deployments for secure, highly available GNS lookups
- *VM Manager Agent for Grid Infrastructure* – provides the ability for Oracle Clusterware to request the restart or reboot of non-clustered VM's that are not releasing cluster resources, thus impacting the operations of the cluster
- *Enhancements for Policy-Managed Deployments* – capabilities have been added to ensure the even dispersion of cluster resources at startup and during failover, and to enhance the what-if predictive functions (introduced in Oracle Clusterware 12c Rel 1) to provide explanations of the predictions for Policy-Managed deployments.

Cluster Domains

Newly introduced in Oracle Clusterware 12c Rel 2 is the Cluster Domain architecture. This is an optional deployment model in addition to the long-established cluster architecture now known as the Standalone Cluster. The Cluster Domain[1] architecture enables simpler, easier deployments, reduced storage management effort and performance gains for I/O operations, especially useful when managing a larger estate of Oracle Clusters.

Standalone Cluster

The Standalone Cluster consists of one or more cluster nodes configured with locally available shared storage, a private interconnect, local instances of Automatic Storage Management (ASM) for managing that shared storage, and the Management Database housed in the Grid Infrastructure Management Repository (GIMR) for cluster health and diagnostic information.

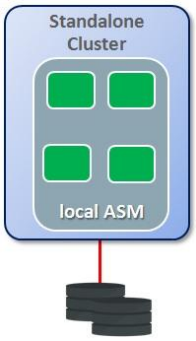

Figure 1: Standalone Cluster

Cluster Domains

A Cluster Domain is actually a grouping of clusters. A Cluster Domain consists of a single Domain Services Cluster and a number of Member Clusters (hosting applications or databases) that utilize services offered on the Domain Services Cluster. Centralized and consolidated services are hosted by the Domain Services Cluster, and consumed by the Member Clusters that are registered with that Domain Services Cluster.

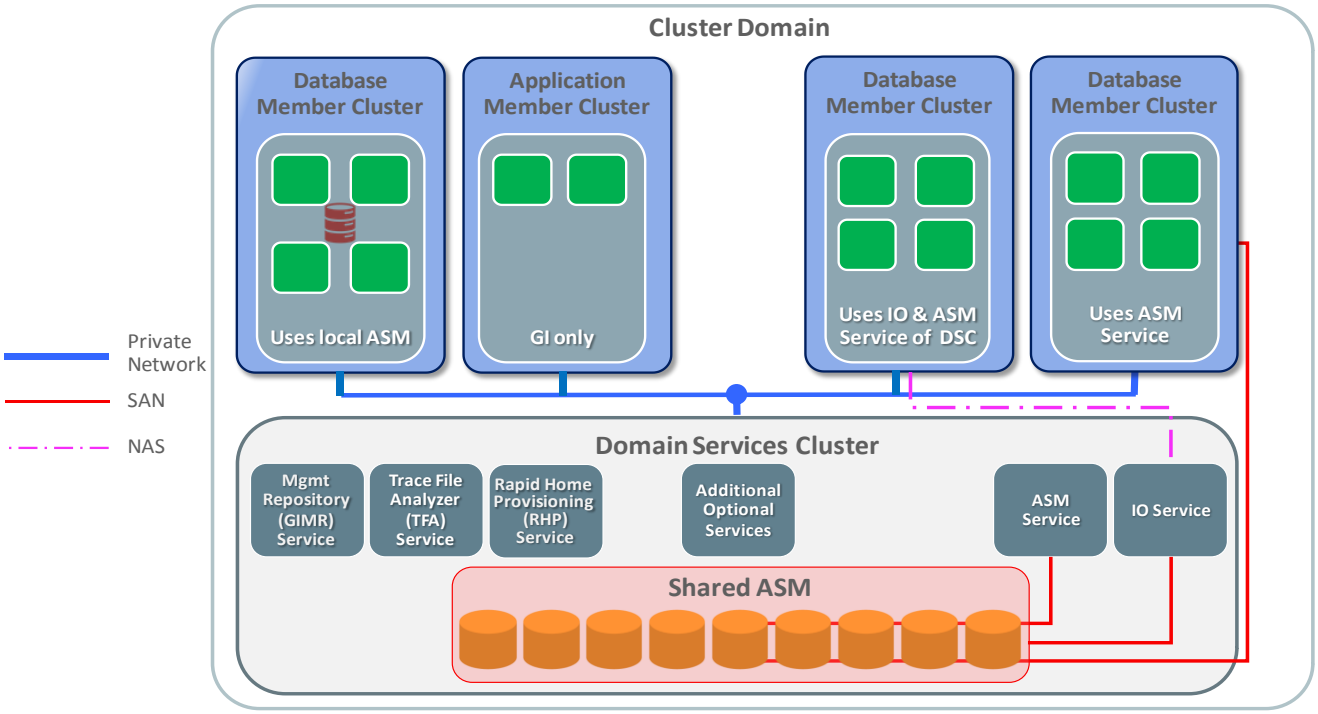

Figure 2: Cluster Domain Architecture

[1] For information about Cluster Domains, refer to "*Oracle Grid Infrastructure 12c Release 2 – Cluster Domains*"

Node Weighting for Split Brain Resolution

Without better understanding of what is critical or of higher priority to the customer's workload, Oracle Clusterware has always resolved split brain conditions in favor of the cluster cohort containing the node with the lowest node number (i.e. which node first joined the cluster). This somewhat arbitrary approach is simple and effective, but does not take into account which of the surviving cluster cohorts might be more viable or required for critical processing.

The Node Weighting feature has been introduced to enhance the resolution of split brain conditions in order to take into account the viability of the cohort to support workloads, and the criticality of that cohort for conducting work. The key to this approach is to qualify when this split brain resolution should take effect. In other words, how does Oracle Clusterware assess the viability and criticality of the surviving cluster cohorts as the result of a split brain condition?

Consider the surviving cluster cohorts in terms of supporting the ongoing workload assigned to the cluster.

If one cluster cohort is bigger than the others, then that cohort will more likely be capable of processing more work, or is currently processing more work and thus should not be interrupted. Instead, that cluster cohort should fence the other cluster cohorts while it continues to process its current workload.

What is a Split Brain condition?

"*a failure condition based on servers not communicating and synchronizing their data to each other*"

– from Wikipedia (https://en.wikipedia.org/wiki/Split-brain_(computing))

In practical terms, for Oracle Clusterware, a split brain condition arises when Oracle Clusterware believes that there is a communication failure between nodes in the cluster.

This can mean that one or more network cards are faulty, there is a break in the network on which the private interconnect is configured, that there is a software fault preventing the communications from being processed, or that there could be resource constraints impeding those communications. For one or more of these problems, Oracle Clusterware will perceive the communications failure as a split-brain condition.

Resolution of the split brain condition will most often result in one of the surviving viable groups of nodes (or cluster cohorts) forcing the fencing of the remainder of the cluster in order to restore cluster communications.

If the cohorts are of equal size, then they must be evaluated on very basic levels in terms of supporting their workloads. These basic criteria might consist of whether they have access to one or more ASM instances and whether they have viable public networks. Without these basic conditions met, it is assumed that the cluster cohort cannot support a workload and should be fenced.

However, the question arises, if there exists two or more viable cluster cohorts, then how does Oracle Clusterware decide which should survive and which should be fenced?

Oracle Clusterware will take into account indicators of criticality. These include singleton resources (if those resources only exist on one node, then it is assumed that that node is critical to a particular workload). Singleton resources are used to differentiate between equally viable cluster nodes and cohorts as they are assumed to indicate support for a designated workload. (If these same resources were distributed across the cluster nodes, then they would not indicate a differentiating workload).

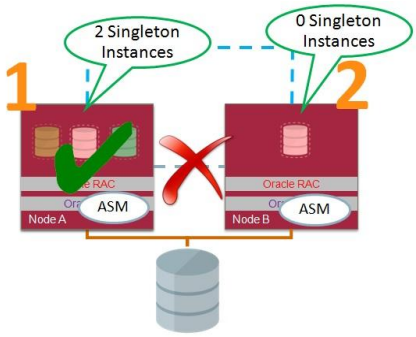

Figure 3: Counting Singleton Resources

In addition, some cluster resources can be designated as workload generating (i.e. they are critical to some workload – set the USER_WORKLOAD attribute to 'yes'). Any resource can be so designated, though it is usual for these to actually be indicative of where the workload is being processed. Thus, if a cluster node or cohort has resources that are attributed to be workload generating, then they are given a higher priority in terms of the split brain resolution.

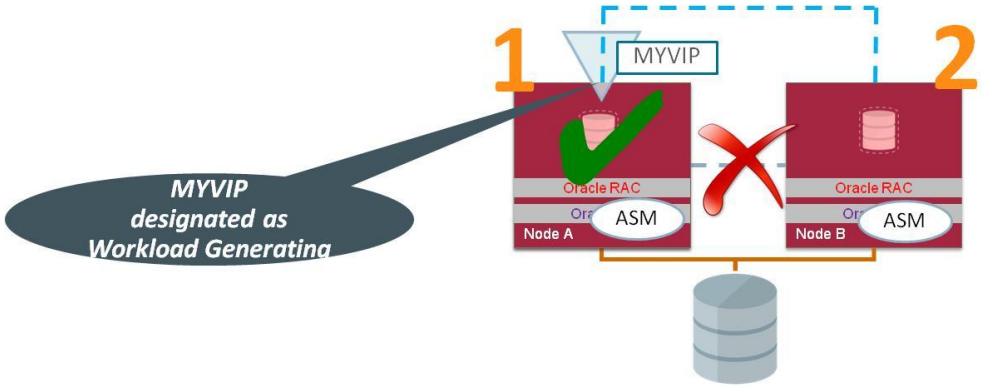

Figure 4: Counting Workload Generating Resources

Most significantly, the cluster administrator is now able to directly designate the criticality of cluster nodes and cluster resources. This allows the administrator to effectively set a priority for which cluster cohorts should survive a split brain condition in terms that make sense to the business.

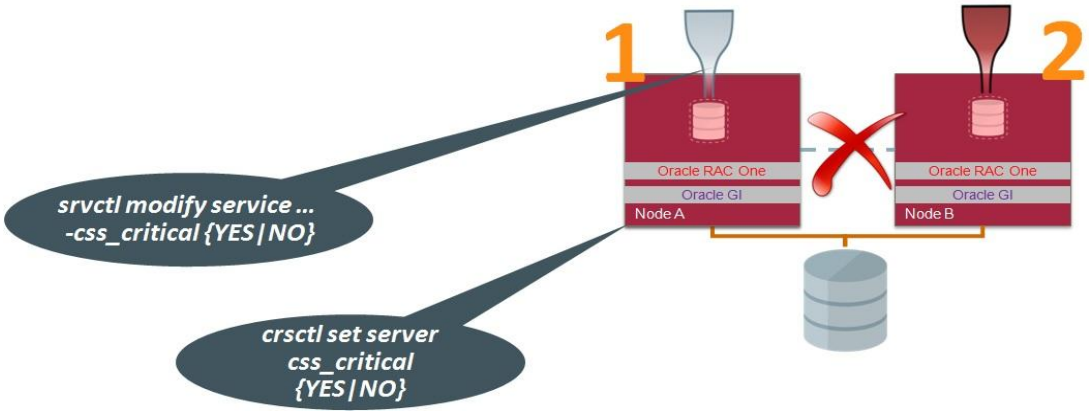

Figure 5: Designating CSS_CRITICAL for Cluster Nodes and Resources

With the Node Weighting feature, cluster administrators and DBA's can now be assured that Oracle Clusterware will take into account both the viability and the criticality of the cluster cohorts in resolving split brain conditions.

Cluster Resource Groups

Cluster Resource Groups provides the cluster administrator with the capability to intuitively model, manage and monitor a group of resources as a single composite entity. The focus is on enhancing the support for applications and non-database processes or programs by grouping the cluster resources in support of that application, and then being able to define dependencies for that entire group in terms of other resources, and within that group for dependencies between the application resources.

As an example, a resource group might be created for a deployment of Oracle GoldenGate. In order to manage the availability of Oracle GoldenGate and the resources it requires to function, the resources would be added to the resource group: for the application itself, for the NFS-mounted trail files, and for an APPVIP created to access the Oracle GoldenGate application processes. The dependencies for this set of resources would then be defined at the group level, possibly including the database instances that must be accessed (for read or write).

The benefit for the cluster administrator is that now a set of related resources can be grouped together, enabling easier management of those resources by only once defining their common behavior and common dependencies, thus managing them as a single unit.

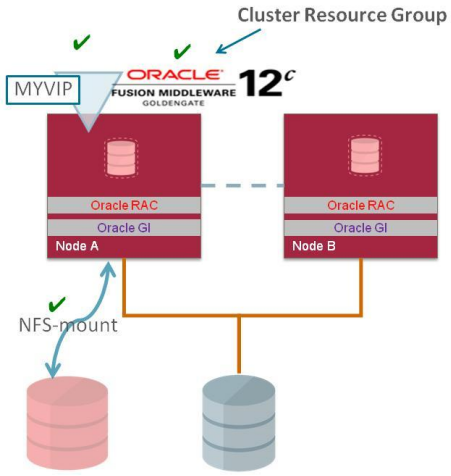

Figure 6: Cluster Resource Group

Cluster Activity Log

The Cluster Activity Log (CALOG) has been introduced as a vehicle for reviewing the history of cluster events provided in a sequence of related activities. The log is maintained and coordinated cluster-wide, enabling the cluster administrator to view activities across the cluster as they occurred, keyed to a causal event or events.

The log can be reviewed as a general sequence or according to a specific event. It may also be accessed for a given period of time, or on a continuous basis. Many query options have been added to the crsctl command to enable administrators a wide range of options.

In the example shown at right, the crsctl query against CALOG is requesting activity records from before a significant date. This will display ALL the records preceding this date. Consequently, the administrator could then query all events associated with a 'parent' activity (i.e. showing only those events that occurred as a consequence of that specific activity).

How to query the CALOG entries?

$ crsctl query calog -beforetime "2017-02-09 15:09:46.522-07:00"
2017-02-09 15:06:51.106-07:00 : Server 'tclust01' has been assigned to pool 'Free'. : 143198681110210633/0/1 :
2017-02-09 15:09:39.228-07:00 : Resource 'ora.net1.network' has been registered. : 143198681110610633/117/1 :
2017-02-09 15:09:40.933-07:00 : Resource 'ora.tclust01.vip' has been registered. : 143198697922810633/136/1 :
2017-02-09 15:09:41.827-07:00 : Resource 'ora.ons' has been registered. : 143198698093310633/143/1 :
2017-02-09 15:09:46.522-07:00 : Resource 'ora.ons' has been modified. : 143198698182710633/194/1 :

Format of output records is:
DATE & TIME (YYYY-MM-DD HH24:MI:SS[.FF][[+-]HH:MM]): Event text: ACTID
ACTID is a generated sequential identifier for related Clusterware events, use it to track related events

Such as,
crsctl query calog -filter "actid == 143198681110210633/0/1"
will only return events associated with that event.

The log data is automatically collected at each of the cluster nodes, buffered, and then written to the cluster's Management Database.

Highly Available Grid Naming Service

As the number of Oracle Clusters has proliferated in support of Oracle RAC Databases and mission-critical applications, and the introduction of new features with Oracle Grid Infrastructure 12c, there has been an increasing requirement for Grid Naming Service (GNS) lookups. If the number and frequency of lookups reaches the capacity of the GNS server there is the potential for delays in GNS request processing to impact cluster activity.

To meet this increasing demand and avoid negatively impacting cluster activity, the GNS server architecture has been enhanced to provide high availability of lookup and other services to the clients by running multiple instances of GNS with different roles. There will be one primary and multiple secondary instances. All the updates from the clients will be serviced by primary instance. The lookup queries will be processed by both primary and secondary instances. Secondary instances will act as backup for primary instance and can be promoted to the primary role whenever an existing primary fails or is removed.

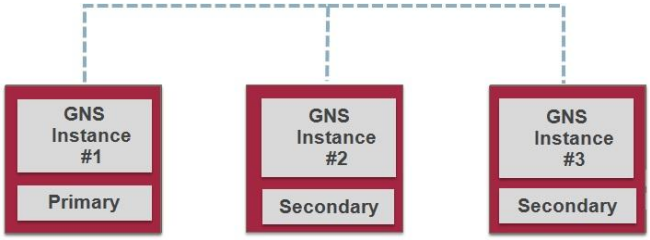

Figure 7: Highly Available Grid Naming Service (HA GNS)

VM Manager Agent for Grid Infrastructure

For Oracle Virtual Machine (OVM) deployments there can be requirements to interact with VM's outside the direct control of Oracle Clusterware. One such case might be for one or more otherwise independent VM's to be started or stopped depending upon the availability of the Oracle Cluster (i.e. such as a firewall, or a mid-tier transaction monitor, for instance). Another case is that of a VM hosting an application that accesses the database on the Oracle Cluster. If that application or the VM itself were to hang, it might be holding resources on the Oracle Cluster indefinitely.

To address these circumstances, the VM Manager Agent for Grid Infrastructure has been introduced, though only for Oracle Virtual Machine environments. Oracle Clusterware can be configured to act at startup, shutdown or under failure conditions. In such cases, the VM Manager Agent will request that the Oracle VM Manager directly control the VM in question. The client VM is referred to as a 'black box' VM, since it is unknown and immaterial to Oracle Clusterware what might be running on that VM.

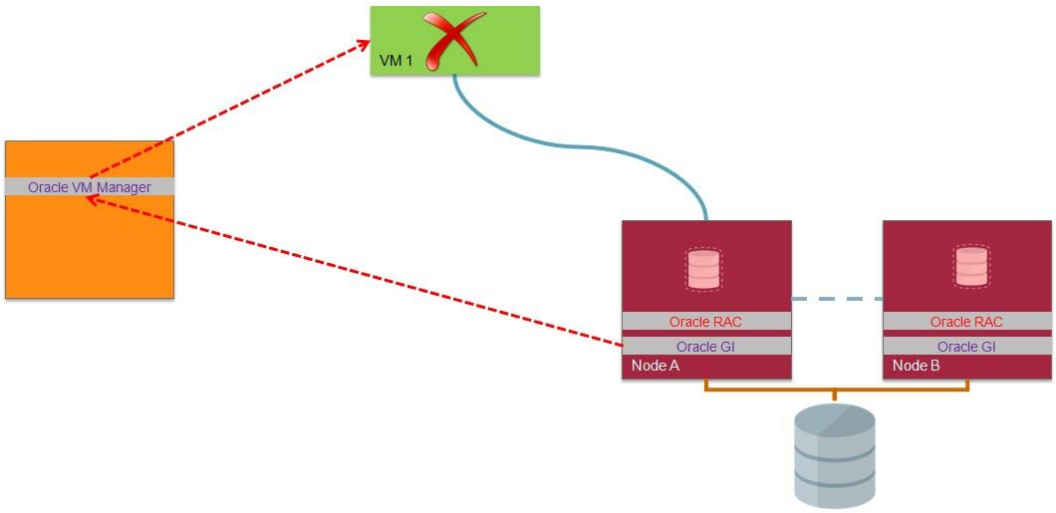

Figure 8: VM Manager Agent for Grid Infrastructure

Policy-Managed Enhancements

Load-Aware Resource Placement

With Oracle Clusterware 12c Rel 1, consolidation in policy-managed environments was enhanced to distribute application resources across server pools in a random fashion, usually as a round-robin dispersion during startup and failover processing. With the newest release, Clusterware enables a more intelligent dispersal of workloads across the server pool, essentially making load-aware resource allocations.

Oracle Clusterware 12c Rel 2 introduces a way to indicate the anticipated CPU and memory consumption of an application resource. By using these indicators, Oracle Clusterware is now able to distribute the application resources more intelligently, even taking into account the possible overprovisioning that is common in consolidated environments.

The following resource attributes have been added to allow cluster administrators to indicate the anticipated CPU and memory requirements for those resources:

Attribute	Units	Description
WORKLOAD_CPU	Number of CPU's	The CPU count. Default:0, which means "uninitialized"
WORKLOAD_MEMORY_TARGET	MB's	The target amount of memory to be allocated for the resource. Default:0, which means "uninitialized".
WORKLOAD_MEMORY_MAX	MB's	The hard maximum amount of memory to be allocated for the resource. Default:0, which means "uninitialized".
WORKLOAD_CPU_CAP	Number of CPU's	The integer percentage that is the maximum utilization of the workload CPUs the resource requires. The valid range is 0..100. Default:0.

Why-If Analysis

The What-If predictive analysis was introduced in Oracle Clusterware 12c Rel 1, providing cluster administrators and DBA's with the ability to predict what would be the result of significant changes in the infrastructure (either in the number of cluster nodes or in the policy governing the server pool). This functionality allowed the cluster administrators and DBA's to discover the consequences of those changes before they were implemented, thus reducing the likelihood of unexpected problems.

The Why-If Analysis for Policy-Managed environments, new in Oracle Clusterware 12c Rel 2, adds to this functionality by providing insight into why the predicted consequences would follow the actions being analyzed. In simple terms, for instance, '*if we add a server to the server pool, **what** would be the impact, and **why** would this be the case?*'

Summary

Oracle Clusterware 12c Release 2 enhances the already feature-rich Oracle Clusterware offering, by providing vast improvements in manageability and deployment, better mitigation and control for failure scenarios.

The introduction of the Cluster Domains to provide simplified, easier and faster management and deployment of multi-cluster infrastructures is key for larger cluster installations. By using this new cluster architecture, customers will be able to consolidate storage management, offload the storage and operation of the Management Databases, and provide an infrastructure designed for rapid cluster deployments and maintenance. This service-oriented infrastructure is both versatile and flexible, allowing for deployment models along lines of business, operational requirements, geographic location, or departmental boundaries.

In addition, as clustered systems have become so widespread, there is more and more focus applied to the possibility and outcome of failure scenarios, whether they occur due to hardware, software, or other causes. With the new Node Weighting functionality, the split brain resolution algorithm is not only more intelligent in supporting the greatest proportion of the workload, it also takes into account what the customer designates as critical to that workload. This is particularly important in consolidated environments, where not all workloads might be consider equal, and hard choices must sometimes be made.

Other features and functionality have been added to Oracle Clusterware to further improvement manageability, such as Cluster Resource Groups, operational support, with the introduction of the Cluster Activity Log, Highly Available GNS and Virtual Manager Agent for Grid Infrastructure. Last, but not least, there are the enriched functionality for Policy-Managed environments, with Load-Aware Resource Placement and Why-If analyzes.

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

blogs.oracle.com/oracle
facebook.com/oracle
twitter.com/oracle
oracle.com

Integrated Cloud Applications & Platform Services

Oracle Clusterware 12c Release 2
New Features Overview
March 2017
Author: Ian Cookson