Technical report:

Oracle RAC 10*g* Release 2 and Oracle Cluster Synchronization Services on IBM System Storage N series

Best practices for high availability

Document NS3555-0

April 11, 2008

Table of contents

Abstract

*Enterprises rely on Oracle databases for their high availability. IBM System Storage N series solutions enable effective Oracle database clustering. This paper covers the best practices for setting various Oracle Cluster Synchronization Services parameter values specifically for Oracle RAC 10*g *Release 2 on an IBM N series storage cluster in an NFS environment.*

Problem statement

Highly available configurations of Oracle® Real Application Cluster (RAC) 10*g* Release 2 (Oracle RAC 10*g* R2) utilizing an IBM® System Storage™ N series storage cluster may experience downtime in a scenario where the initial cluster formation or cluster reconfiguration happens during the time of storage failover. This paper addresses this issue and provides a recommended and validated solution to the above problem in a network file system (NFS) environment.

Assumptions

The paper assumes readers are familiar with Oracle RAC 10*g* R2 and the operation of IBM N series storage cluster systems. The paper also assumes that readers are familiar with installation of Oracle patch-sets and one-off patches and any relevant operating system (OS) patches. It is also required that the reader has a general understanding of IBM N series storage clustering technology, networking terminology, and implementations.

The server environment

In the environment used for the purposes of this paper, the servers are running a Red Hat Enterprise Linux® OS (RHEL AS 4 U3). The database is Oracle RAC 10*g* R2 with Oracle Cluster Ready Services (CRS). This is a certified configuration and, as such, the components presented in this paper have to be used in the same combination to gain support from all parties involved. The only exception to this is the application of certain patches (as defined and required by all the vendors in this configuration). This document will also cover the patches and recommendations of Cluster Synchronization Services (CSS) parameter values for running Oracle RAC 10*g* R2 on an IBM N series storage cluster in NFS environment.

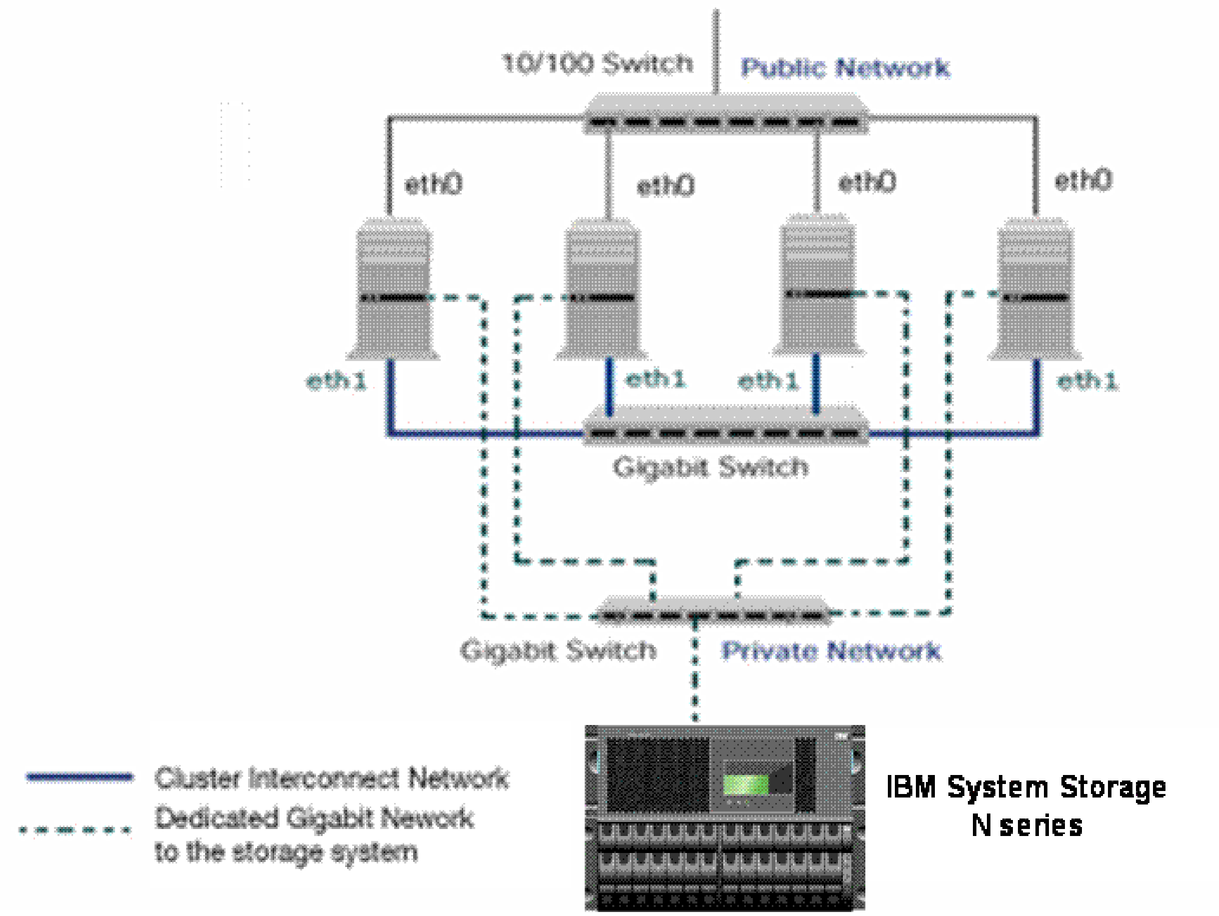

*Figure 1) Oracle RAC 10*g *R2 using four nodes with IBM N series storage.*

Figure 1 illustrates a typical four-node configuration of Oracle RAC10*g* R2 utilizing an IBM N series storage cluster. This is a scalable configuration and allows users to scale horizontally and internally in terms of processor, memory, and storage.

As shown in the network diagram, it is recommended that you dedicate a private network connection between the Oracle RAC 10*g* R2 servers and the N series storage. This is accomplished using a dedicated gigabit network (with a gigabit switch) to the N series storage. A dedicated network connection is beneficial for the following reasons:

- In an Oracle RAC 10*g* R2 environment, it is important to eliminate any contentions and latencies.
- Providing a separate network ensures security.

Requirements

Hardware used for tests

Cluster nodes:

- Four servers running Linux OS (2.6-based kernel)
- One 10/100 Base-TX Ethernet PCI adapter (Public IP)
- One 10/100/1000 Base-T Ethernet PCI adapter (for private interconnect)
- One 10/100/1000 Base-T Ethernet PCI adapter (connected to N series storage).

Storage infrastructure:

- Two IBM System Storage N series N7800 systems with IBM System Storage N series with Data ONTAP® 7.2
- One gigabit switch with at least four ports
- One gigabit NIC in the system
- One or more disk shelves, based on the disk space requirements.

Software used for tests

For all four nodes in the participating cluster unless specified otherwise:

- Linux OS (2.6-based kernel)
- Oracle 10*g* R2 (Enterprise Edition).

Setup for N series storage cluster

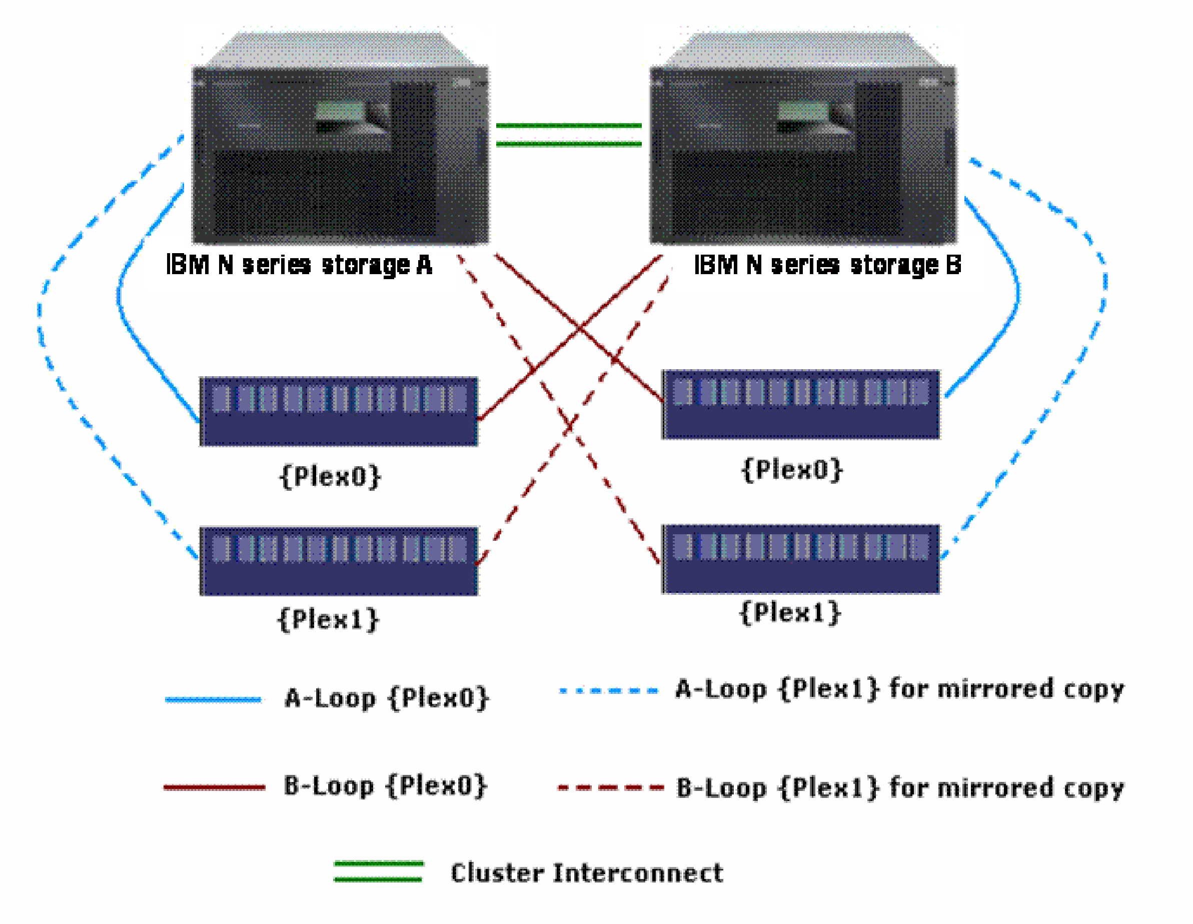

Figure 2) Hardware setup on mirrored active/active controllers.

The storage configuration is a mirrored active/active controller configuration of N series N7800 systems. The words failover and takeover, failback and giveback are also used interchangeably throughout the document. The words node and partner described in a cluster pair refer to a storage controller.

When one partner fails or becomes impaired, a takeover occurs and the partner node continues to serve the failed node's data.

When the failed node is functioning again, the administrator initiates a giveback command that transfers resources (failedover resources) back to the original partner node and the original node resumes normal operation, serving its own data.

Please do not configure both of the IBM N series storage systems for automatic giveback; giveback needs to be initiated manually by the administrator during planned downtime as the giveback process takes a longer time than the takeover process.

1. Configure an IBM N series storage system running Data ONTAP 7.2 or above and with cluster, NFS, IBM System Storage N series with SnapMirror®, Snapmirror_sync, syncmirror_local and IBM System Storage N series with SnapRestore® license keys.
2. The cluster failover parameters on both the N series storage systems should have the following values:

CF.GIVEBACK.AUTO.ENABLE	OFF
CF.GIVEBACK.CHECK.PARTNER	ON
CF.TAKEOVER.DETECTION.SECONDS	15
CF.TAKEOVER.ON_FAILURE	ON
CF.TAKEOVER.ON_NETWORK_INTERFACE_FAILURE	ON
CF.TAKEOVER.ON_PANIC	ON
CF.TAKEOVER.ON_SHORT_UPTIME	ON

3. Create and export volumes for storing Oracle database files on the storage:

Create three volumes on the storage (Data1) as listed below.

oradata ---Oracle datafiles and control files
ora10g ---CRS files
oralogs ---database logs, a copy of control file and archive logs.

To create volumes, use the following command at the IBM N series storage console:

```
Data1> vol create oradata 14
```

Note: Volume oradata was created with 14 disks and volumes oralogs and orahome with eight disks each. You can create your volumes based on your workload needs.

Edit the /etc/exports file on the N series storage (Data1) by adding the following entries to that file:

```
/vol/orahome -anon=0
/vol/oradata -anon=0
/vol/oralogs -anon=0
/vol/ora10g  -anon=0
```

Execute the following command at the storage system console:

```
Data1> exportfs -a
```

Note: It is recommended that you use flexible volumes in your database environment. IBM System Storage N series with FlexVol™ technology pools storage resources automatically and enables you to create multiple flexible volumes on a large pool of disks. This flexibility means you can simplify operations, gain maximum spindle utilization and efficiency, and make changes quickly and seamlessly.

The database volume layout discussed in this document was defined for certification purposes and your setup may vary depending upon requirements.

Setup for Oracle RAC 10*g* R2

The installation of Oracle RAC 10*g* R2 on four nodes is not in the scope of this document.

Oracle RAC 10*g* R2 CSS parameters

With different patch-sets of Oracle 10*g* R2, there exist different timeout parameters that are used by CSS while accessing storage data. This documentcovers the following Oracle 10*g* R2 patch-set versions:

- Oracle Database 10.2.0.1
- Oracle Database 10.2.0.1 + Patch for Bug 4896338
- Oracle Database 10.2.0.2
- Oracle Database 10.2.0.3.

Oracle Database 10.2.0.1

There is only one CSS parameter available in this version of Oracle and it is called miscount; it represents the maximum time in seconds that a heartbeat can be missed before entering into cluster reconfiguration to evict the node, and the maximum time allowed for a voting file I/O to complete. The default value for misscount is 60 seconds.

Oracle Database 10.2.0.1 + Patch 4896338 and Oracle Database 10.2.0.2

There is a bug and patch (4896338, for very low brownout) with Oracle Database 10.2.0.1. Oracle Database 10.2.0.2 has a fix for this bug.

Three CSS parameters are available in 10.2.0.2 and 10.2.0.1 + patch for bug 4896338; they are as follows:

- misscount – again, which represents the maximum time in seconds that a heartbeat can be missed before entering into a cluster reconfiguration to evict the node.
- disktimeout – the maximum amount of time allowed for a voting file I/O to complete; if this time is exceeded the voting disk will be marked as offline.
- reboottime – the amount of time allowed for a node to complete a reboot after the CSS daemon has been evicted.

Default values for these parameters are as follows:

- misscount = 60 seconds
- disktimeout = 200 seconds
- reboottime = 3 seconds

Using "crsctl get css disktimeout / reboottime" will not show the parameter value unless you modify it explicitly. You can check the parameter values using ocssd.log under the $CRS_HOME directory.

CRS internally calculates two parameters, namely diskshorttimeout and disklongtimeout (which can be checked in ocssd.log), where:

- diskshorttimeout = misscount - reboottime–this value is used during reconfiguration and initial cluster formation as a timeout for voting file I/O to complete.
- disklongtimeout = disktimeout—this value is used during normal operation of RAC as a timeout for voting file I/O to complete.

Oracle Database 10.2.0.3

This version also has same parameters as that of Oracle Database 10.2.0.2; also the default values are the same as Oracle Database 10.2.0.2. There is slight difference in the internal calculation of the parameter values. If disktimeout is less than the misscount value, then during cluster formation and throughout cluster operation, misscount - reboottime is considered as disktimeout and the modified parameter disktimeout is ignored.

That is, in Oracle Database 10.2.0.3, diskshorttimeout = disklongtimeout if the CSS disktimeout parameter is less than the CSS misscount.

Recommendations for Oracle 10*g* R2 CSS parameter values with N series storage

As diskshorttimeout = misscount - reboottime; and if misscount and reboottime are kept as default values (i.e., 60 seconds and 3 seconds, respectively); then the time for accessing a voting file will be considered as 57 seconds by CSS, so If the reconfiguration happens during the N series storage takeover or giveback process there are chances of a CRS reboot taking place. Hence, following are the recommended values for CSS timeout parameters for Oracle RAC 10*g* R2 to work smoothly during the N series takeover and giveback process.

1. Oracle Database 10.2.0.1

   misscount = 120 seconds (default is 60 seconds)

2. Oracle Database 10.2.0.1 + Patch for Bug 4896338

   misscount = 120 seconds (default is 60 seconds)
   disktimeout = 200 seconds (default)
   reboottime = 3 seconds (default)

3. Oracle Database 10.2.0.2

   misscount = 120 seconds (default is 60 seconds)
   disktimeout = 200 seconds (default)
   reboottime = 3 seconds (default)

4. Oracle Database 10.2.0.3

   misscount = 120 seconds (default is 60 seconds)
   disktimeout = 200 seconds (default)
   reboottime = 3 seconds (default)

All the above recommendations are for a Linux OS.

Note: The stock version of Oracle 10*g* R2 in versions lower than 10.2.0.2 does not provide all the configurable CSS parameters. Hence, it is advisable to upgrade to version 10.2.0.2 or higher.

Appendix

Commands to check / modify CSS parameters:

crsctl get css misscount —	to check misscount value
crsctl get css disktimeout —	to check disktimeout value
crsctl get css reboottime —	to check reboottime value
crsctl set css misscount 120 —	to set misscount to 120 seconds
crsctl set css disktimeout 200 —	to set disktimeout to 200 seconds
crsctl set css reboottime 3 —	to set reboottime to 3 seconds

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

FlexVol, Data ONTAP, Network Appliance, the Network Appliance logo, SnapMirror, SnapRestore and SnapVol are trademarks or registered trademarks of Network Appliance, Inc., in the U.S. and other countries.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.