

Oracle^(R) 10gR2 Real Application Cluster Installation Using Sun Cluster Software and Sun StorageTek^(R) 5000 Series NAS Appliance

ASPE
Disk Products Group
Sun Microsystems, Inc.

© 2005 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, CA 95054 USA

All rights reserved.

This product or document is protected by copyright and distributed under licenses restricting its use, copying, distribution, and decompilation. No part of this product or document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any. Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD appliances, licensed from the University of California.

Sun, Sun Microsystems, Sun StorEdge, the Sun logo, are trademarks, registered trademarks, or service marks of Sun Microsystems, Inc. in the U.S. and other countries.

UNIX is a registered trademark in the United States and other countries, exclusively licensed through X/Open Company, Ltd.

Windows is a registered trademark of Microsoft Corporation in the United States and other countries.

All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

The OPEN LOOK and Sun's Graphical User Interface was developed by Sun Microsystems, Inc. for its users and licensees. Sun acknowledges the pioneering efforts of Xerox in researching and developing the concept of visual or graphical user interfaces for the computer industry. Sun holds a non-exclusive license from Xerox to the Xerox Graphical User Interface, which license also covers Sun's licensees who implement OPEN LOOK GUIs and otherwise comply with Sun's written license agreements.

RESTRICTED RIGHTS: Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g)(2)(6/87) and FAR 52.227-1987), or DFAR 252.227-7015(b)(6/95) and DFAR 227.7202-3(a). DOCUMENTATION IS PROVIDED AS IS AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS HELD TO BE LEGALLY

Table of Contents

Introduction.....	1
Sun StorageTek 5000 Series NAS Appliance.....	1
Sun Cluster for Oracle RAC.....	1
Supported Configuration.....	2
Sun Cluster Quorum Device(QD)	3
SCSI Reservation	3
Internet SCSI (iSCSI).....	4
Oracle RAC with Sun Cluster Implementation overview.....	4
Stage 1 : Host Configurations	6
Stage 2a : File Volume Configuration on StorageTek NAS Appliance.....	7
Stage 2b : iSCSI LUN Configuration on StorageTek NAS and Cluster Nodes.....	8
Stage 3 : Sun Cluster 3.1 U 4 setup.....	9
Stage 4 : Installing Oracle RAC Framework for Sun Cluster.....	10
Stage 5 : Oracle Environment Setup.....	11
Stage 6 : Oracle 10gR2 Clusterware Installation.....	13
Stage 7 : Oracle RAC Database binaries Installation.....	14
Database Creation.....	15
Reference.....	16
Appendix.....	17

Introduction

This technical report describes the procedure to implement a two-node Oracle10gR2^(R) Real Application Cluster (RAC) with Sun Cluster Software in a SolarisTM10 U2 Operating environment using Sun StorageTek^(R) 5000 NAS appliance for storage.

The intended audience are system and/or database administrators. The reader of this document is assumed to have a fair understanding of Solaris, Sun Cluster software, Oracle Clusterware, iSCSI and NFS protocols.

Note : The term *Sun StorageTek NAS Appliance* is synonymously used with *Sun StorageTek NAS 5320 Appliance*.

Sun StorageTek 5000 Series NAS Appliance for Oracle

Some of the key advantages by implementing Oracle RAC using Sun StorageTek 5000 Series NAS Appliance are :

1. All the different types of Oracle files such as binaries, data files, cluster files, backup sets etc., can be stored on the NAS Appliance and can be shared across multiple nodes using NFS protocol.
2. There is no requirement for additional volume management and file management. It is File system based storage which are NFS shared and very easy to manage.
3. Controller based RAID-5 data protection.
4. Business Continuity protection features such as Checkpoint Software (Snapshot) and File Replicator (Remote mirroring) provide ways to protect the data from failures. Both these features are certified to work with Oracle10g/9i databases.
5. Oracle RAC implementations on Sun StorageTek NAS appliance is certified by Oracle.
6. Dynamic addition of space to the file volume without any down time for Oracle RAC.

Sun Cluster for Oracle RAC

Sun Cluster is part of the Solaris Enterprise System. Sun Cluster software offers the premier availability platform for improving the predictability and resilience of business critical applications – such as Oracle databases in a clustered environment. Sun Cluster framework for Oracle RAC provides the enterprise with increased manageability for RAC deployments on both SPARC and X64 platforms.

The following are some of the advantages while implementing Oracle RAC within Sun Cluster framework :

Network Interconnects & node synchronization

- Better fault management and availability due to tighter integration with Solaris kernel. Heart beats are more reliable and tunable.
- With redundant private interconnect (up to 6), offers better communication and fault tolerance by doing transparent fail-over of traffic from failed link(s) to the working link(s). This is configured as part of Sun Cluster installation, aided by the auto-discovery capability in the installation tools.
- Network interface names could be different from node to node, allowing flexibility of using different NICs model or placement of NICs in server expansion slots, etc. At runtime, Sun Cluster automatically plumb up the interconnect interfaces and assign network address and subnet attributes.
- The interconnects can be dynamically modifiable and new interconnects also can be added in real-time with no down time to Oracle RAC.
- Only one interface name *clprivnet0* which is configured by Sun Cluster will be used by the Oracle CRS for internode communication. This pseudo interface is uniform across the nodes.
- Cache fusion traffic is striped over these multiple interconnects resulting in higher network throughput.

- Sun Cluster's RDT (Reliable Datagram Transport) along with Solaris RSM (Remote Shared Memory) on a specialized interconnect hardware SCI (Scalable Coherent Interface) provides an optimized IPC transport for Oracle RAC with better throughput and latency in comparison with Oracle RAC's default UDP/IP transport layer.
- Sun Cluster takes care of nodes time synchronization during installation and keeps the time across the node synchronized using Network Time protocol which is critical for Oracle Clusterware operations.

I/O Fencing

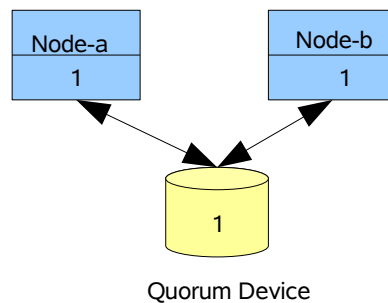
- The Quorum device management by Sun Cluster which is a form of fencing is activated on a two node configuration to detect split-brain and also the health of participating nodes of a cluster. If a node is deemed unhealthy, the node not only evicted from cluster membership, but no writes are allowed from the node to the shared storage to ensure data integrity.
- Sun Cluster enables fail-fast behavior in the kernel I/O module so that on the failed node, upon receiving error from the storage, the driver panics the node.

Supported Configuration

Configuration	Value
Operating System	Solaris 10 Update 2 Solaris 10 / Solaris 10 U 1 with Patch119090-20 (Sparc) / 119091-20 (x86) or above
Cluster Configuration	Two node cluster (SPARC / X64)
Server configuration	Sparc / X64 64 bit At least 2GB of memory
Network interface (Node interconnects)	GigE (1000 Mbit)
Sun Cluster	3.1 U 4 / 3.2
Oracle Clusterware / Database	10.1.0.4 or above 10.2.0.2 or above
Storage	Sun StorageTek 5000 series NAS Appliance StorageTek NAS OS version 4.20 or above RAID-5 Configuration
Protocol	NFSv3
Sun Cluster Quorum device	iSCSI LUN of Sun StorageTek 5000 NAS (or) A shared FC storage (direct attached / SAN)

Sun Cluster Quorum Device(QD)

In Sun Cluster, the mechanism that determines node participation is known as a *quorum*. Each cluster node is counted towards a quorum vote. For the cluster to be operational, the vote requirement is $N/2 + 1$ where N is the node count. In a two-node cluster, in order for the cluster to survive, it requires at least 2 votes. Without a third quorum, in the event of one of the cluster node is down, the entire cluster would be down. In order to avoid that condition, a shared storage device is identified and configured as the quorum device. The device must be a shared disk that can be accessed by both the nodes of the cluster. Sun Cluster configurations use quorum devices to maintain data and resource integrity by preventing amnesia and split-brain problems when the cluster node attempts to join the cluster.



The requirements :

- Two-node cluster must have a quorum device. For other topologies, it is optional.
- Configure odd number of quorum devices. To ensure the quorum devices to have completely independent failure pathways.
- There is no real size requirement for the quorum device. It is also possible to assign the device which could contain user data as the quorum device.

To implement a two-node Sun Cluster with StorageTek NAS Appliance, the following methods can be used.

1. Use iSCSI Lun from StorageTek NAS appliance as shared block device for quorum device.
2. Use shared disk from direct attached FC storage (or) SAN Lun for quorum device.
3. Using Sun Cluster 3.2 version, a Quorum Server can be configured.

In this document, the procedure for method (1) which uses iSCSI as the quorum device is explained. If the user choose to use a non-iSCSI lun as QD, the same procedure can be followed excluding Stage-2b which describes iSCSI configuration.

SCSI Reservation

Sun Cluster uses the SCSI reservation mechanism in conjunction with quorum principle for data integrity protection. When the cluster is formed, one node takes responsibility of the quorum device and is tagged as the owner. The other node is tagged as capable of becoming owner. The keys are written to the quorum device by both the nodes and so, the nodes are allowed to be part of the cluster. If a node leaves the cluster, the other node clears the keys that belong to the node that left the cluster. The node is fenced and no I/O can happen from the node. When that node rejoins the cluster, the keys are re-registered and enabled access. Sun Cluster uses emulated PGR (persistent group reservation) for SCSI-2 devices. For cluster nodes with more than two nodes, the shared storage device must be SCSI-3 capable.

Internet SCSI (iSCSI)

iSCSI is a protocol that enables transport of block data over IP network. This protocol doesn't require any special network infrastructure that is typically required for block device access from devices such as Fibre channel. Sun StorageTek NAS appliance has the capability to provide block level access for the file volume.

The iSCSI luns are created and granted access to the nodes in the StorageTek NAS appliance. After that, the iSCSI configuration has to be done at the nodes. The iSCSI luns are discovered by the nodes either using *static* or *dynamic (target)* method. In static method, device is discovered by providing the iSCSI name which is given by the NAS appliance for that lun. In target discovery method, all the iSCSI luns that are created in the NAS appliance can be accessed by the node.

Oracle RAC with Sun Cluster Implementation overview

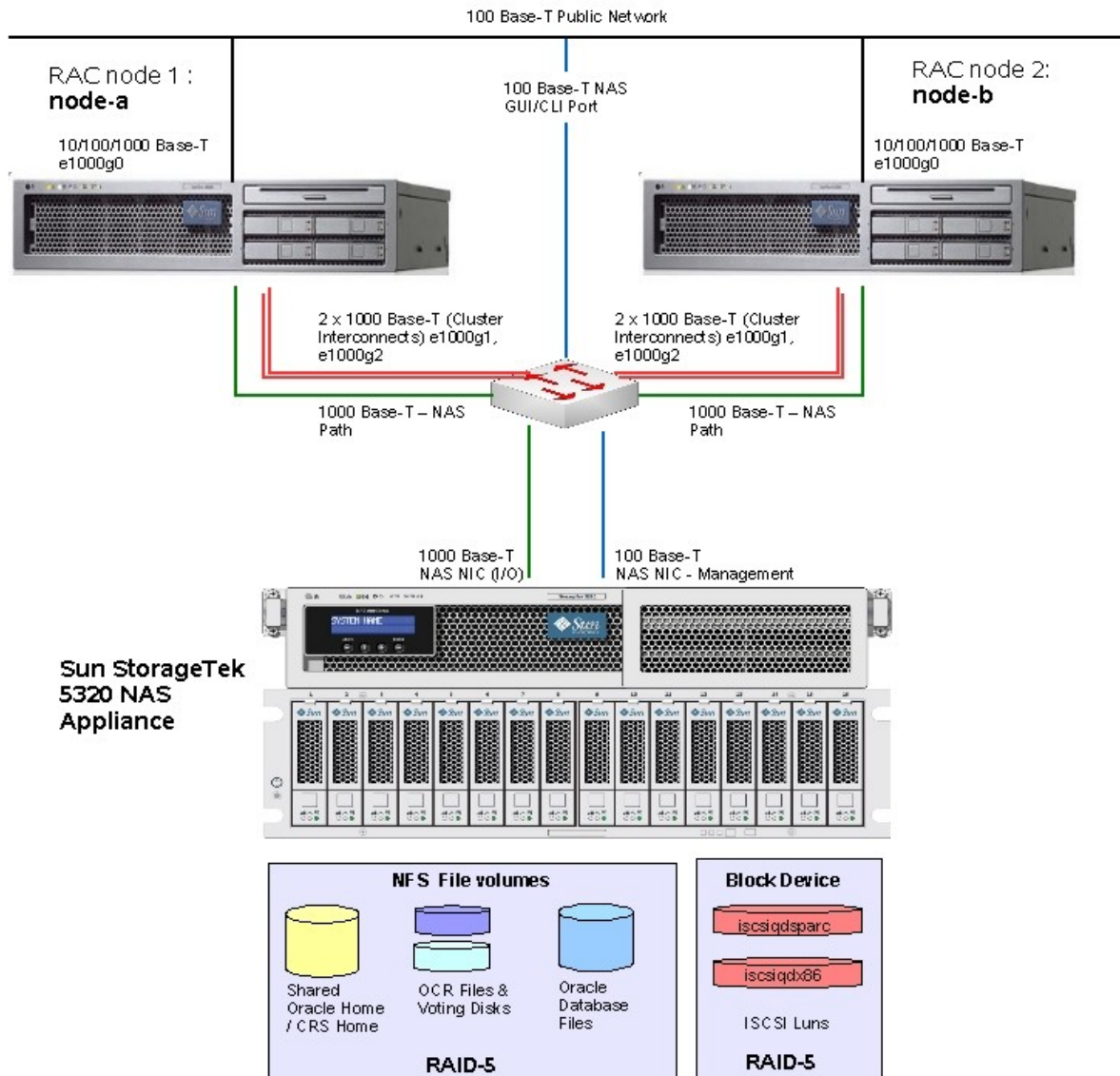
For this proof of concept, StorageTek NAS 5320 Appliance is used to store all Oracle files including Oracle database files, OCR, Voting and Oracle binaries. For quorum device, iSCSI lun from StorageTek NAS is used. Sun Cluster binaries are stored local to the nodes.

In our test environment, a fresh install of Solaris 10 U 2 was done on both the cluster nodes. A slightly different procedure have to be followed for the X86 and Sparc architectures. Those are explained wherever applicable.

Note : The steps provided here are to be considered only as a guideline and not to be substituted for the install procedures provided by the individual components. For details, please refer to the installation documents for Sun Cluster 3.1, Oracle RAC agent for Sun Cluster 3.1, Oracle Clusterware, Oracle Database, StorageTek 5320 NAS Appliance and Solaris operating system. Also, verify the compatibility matrix between Solaris versions, Oracle versions and Sun Cluster versions for dependencies.

The steps and procedures explained in this document are based on Sun Cluster 3.1 Update 4 version. For Sun Cluster 3.2 version, please refer to the documentation for that steps need to be followed.

Oracle 10gR2 Real Application Cluster on Solaris 10 Update 2 Using Sun Cluster 3.1 and Sun StorageTek 5320 NAS Appliance



Oracle 10gR2 RAC with Sun Cluster on StorageTek NAS Appliance

Stage 1 : Host Configurations

The following steps are to be performed on both the nodes.

Physical configurations

1. Two cluster nodes with similar architecture (X86 or Sparc).
2. One 500MB partition for globaldevice configuration on each node.
3. At least two private interconnects (GigE) between them via gig-E switch.
4. Sun StorageTek NAS appliance with two Gig-E network interface for data path. The third NIC is used for management.

Environment configurations

1. Enable *telnet* sessions and verify :

```
# svcadm -v enable svc:/network/telnet:default
# svcs -a | grep telnet
```

2. Enable *rlogin*/*rsh* sessions & verify :

```
# inetadm -e svc:/network/login:rlogin
# inetadm -e svc:/network/shell:default
# inetadm -e svc:/network/rexec:default
# inetadm | egrep rlog
```

3. Set the following parameters in */etc/system* file. With Solaris 10, a lot of shared memory and semaphore settings are obsoleted. The *shmmax* value is to specify the maximum shared memory that can be allocated for the Oracle instance. It is recommended to have this value to at least 50% of the physical memory.

```
set noexec_user_stack=1
set shmsys:shminfo_shmmax=4294967295
```

4. With Solaris 10, the multipath software (Sun Traffic Manager) is built-in the kernel itself. Verify whether the multipath option in */kernel/drv/tp.conf* file. If the *mpxio-disable="yes"*, then change the value to "no" .

5. At least two physical Gigabit Ethernet connections are required between the cluster nodes for private interconnects to be used by both Oracle and Sun Cluster. There is no need to configure private hostnames. As long as the physical connection is available, Sun Cluster plumbs the interface up and assigns IP addresses and names *clprivnet<#>*.

Edit */etc/hosts* and add the cluster nodes and VIP nodes.

```
# Internet host table
#
127.0.0.1      localhost
10.8.54.54    node-a   loghost   # Primary Cluster node
10.8.54.57    node-b               # Secondary Cluster Node

# VIP Nodes
10.8.54.80    node-a-vip
10.8.54.81    node-b-vip
```

Stage 2a : File Volume Configuration on StorageTek NAS Appliance

- It is recommended to install the latest StorageTek NAS O/S version (at this time of writing, 4.20) and any patches – if available from <http://sunsolve.com>.
- Configure network interface for StorageTek NAS with at least two Gigabit NIC interfaces. It is recommended to connect each interface to a different switch for high availability. The role for those interfaces must be *Primary*.
- It is also recommended to configure one Gigabit interface for *File Replicator* - which is a premier feature for remote replication which replicates volumes to another StorageTek NAS Appliance. The role for that interface must be “Mirror”.
- Create one RAID group (LUN) using the drives from a tray. Each RAID group stripes the data across the drives that were chosen.
 - Create a LUN using 14 x 146GB drives.
- Create three partitions (which also creates file volume), for Oracle binaries, database files and Clusterware files. It is also possible to use same file volume for database and clusterware files – since they have similar NFS mount options.
 - a) For Database with 200GB
 1. Create 1GB file volume (partition) *oradata*
 2. Create one 200GB segment *seg1*.
 3. Attach that segment to the file volume *oradata*.
 - b) For Oracle binaries , create 20GB filevolume *oraclebin*.
 - c) For OCR and Voting Clusterware files, *external redundancy* is used with one copy of OCR and Voting disk . Create 2GB filevolume *clusterware*.
- From the GUI menu-> Unix Configuration -> Configure NFS -> add the hosts and create host group by adding the two nodes into a group.


```
Host group : scracnas
Hosts : node-a, node-b
```

The alternate way is to use the “General” host group which grants access of these volumes to any node.
- Configure NFS export for the file volumes from GUI menu -> click on file volume -> right click and choose properties.
 1. Access : *Read/Write*
 2. Map root user : *root*
 3. Host : Choose *Host Group* and from drop-down, choose the *scracnas* .

Stage 2b : iSCSI LUN Configuration on StorageTek NAS and Cluster Nodes

The following is the procedure to create and use the iSCSI lun as quorum device. For this proof-of-concept, the iSCSI device is going to be used purely for quorum purposes and not for storing any data. No iSNS server is configured.

On Sun StorageTek NAS Appliance:

1. Using web based UI, point the URL to NAS appliance and, create file volume (*iscsiqd*) with 10MB size.
2. Go to iSCSI Configuration screen and add the nodes in the Access list to allow access to the cluster nodes. (add *node-a* and *node-b*)
3. Go to Configure iSCSI lun screen and add a new iSCSI lun
 - From the drop-down volume list. Choose the filevolume *iscsiqd*.
 - Choose size smaller than the filevolume size (5 MB)
 - After this step, the NAS box adds the iSCSI lun and assigns name.
iqn.1986-03.com.sun:01:000e0c9f0afe.44D4C0E4.iscsiqd

On both the cluster nodes :

1. **Only for Solaris 10 and Solaris 10 Update 1** : It is strongly recommended to update the iSCSI drivers to the latest patch for proper functioning. Obtain the latest *SUNWiscsir* and *SUNWiscsiu* patches (which can be obtained from <http://sunsolve.sun.com>). As at time of this writing, the following patches need to be applied :

```
119090-20 (Sparc)
119091-20(X64)
```

```
# patchadd -d 119090-20
```

```
PKGINST:  SUNWiscsir
NAME:      Sun iSCSI Device Driver (root)
CATEGORY:  system
ARCH:      sparc
VERSION:   11.10.0,REV=2005.01.04.14.31
BASEDIR:   /
VENDOR:    Sun Microsystems, Inc.
DESC:      Sun iSCSI Device Driver
```

NOTE : No iSCSI driver patches are required for Solaris 10 Update 2.

2. Enable the static discovery mode. It is recommended to use static discovery method.

```
#iscsiadm modify discovery --static enable
```

3. Add the iSCSI static configuration. For example,

```
# iscsiadm add static-config \
iqn.1986-03.com.sun:01:000e0c9f0afe.44D4C0E4.iscsiqd,10.8.11.250
```

4. Discover the iSCSI device and run *format* command to verify whether the device is seen by the host.

```
# devfsadm -i iscsi
# format
Searching for disks...done
```

AVAILABLE DISK SELECTIONS:

- 0. c0t0d0 <SUN72G cyl 14087 alt 2 hd 24 sec 424>
/pci@1c,600000/scsi@2/sd@0,0
- 1. c0t1d0 <SUN72G cyl 14087 alt 2 hd 24 sec 424>
/pci@1c,600000/scsi@2/sd@1,0
- 2. c2t6080020FFF9E6D400000000300000000d0 <SUN-StorageTekNAS-4.11 cyl 2248 alt
2 hd 32 sec 256> [/scsi_vhci/ssd@g6080020fff9e6d400000000300000000](#)

Note : Refer to Solaris Documentation (or) man page of *iscsiadm* command for various options that can be used to discover, view and configure iSCSI luns.

Stage 3 : Sun Cluster 3.1 U 4 setup

NOTE : Please refer to Sun Cluster Installation Manual for detailed step by step procedure to install Sun cluster Software. The following are the broad overview of steps that are done :

1. On both the nodes, create 500MB partition. Then create UFS filesystem on that partition and mount using the mountpoint */globaldevices*.

```
# newfs /dev/rdisk/c0t1d0s1
# mkdir /globaldevices
# mount /globaldevices
```

Make the following entry into */etc/vfstab*.

```
/dev/dsk/c0t1d0s1      /dev/rdisk/c0t1d0s1      /globaldevices  ufs      2  yes
```

2. **For X86 bases systems only** : Do the following on both the nodes.

- a. Make a copy of the file as backup.

```
# cp /boot/solaris/filelist.ramdisk filelist.ramdisk.orig
```

- b. Add the following entry in the file.

```
# echo "etc/cluster/nodeid" >> /boot/solaris/filelist.ramdisk
```

Sun Cluster Installsteps on node-a:

1. Install the sun cluster software. From the installation CD or location, run the command *scinstall* command to initiate the installation.

```
# cd <CD mount>
/sparc/stack/Solaris_sparc/Product/sun_cluster/Solaris_10/Tools
# ./scinstall
```

2. Choose to install , *Only this node at this time* option
3. Provide the cluster name. For example, *scracnas*
4. After the completion of the steps, the system reboots.

5. When the system comes back up, run `/usr/cluster/bin/scsetup` and perform the following :
 - a) Add the second node information. This is to authorize the other node to be added to this cluster.
 - b) Add the transport cables and enable them (if the private interconnects are directly connected between the hosts instead of connection via transport junction).

Sun Cluster installsteps on node-b:

1. Install the Sun Cluster software by issuing `scinstall` command.


```
# cd <CD mount>
/sparc/stack/Solaris_sparc/Product/sun_cluster/Solaris_10/Tools
# ./scinstall
```
2. Choose “Add node to the existing cluster” option.
3. Provide the cluster name `scracnas`, the primary cluster node name `node-a`, the interconnect information. Sun Cluster authenticates the information, installs the binaries and reboots the node.

Quorum device configuration from any of the cluster node. For example, from node-a :

1. Once both the systems are up, run `scdidadm` command to find out the shared storage to be used as Quorum device.

```
[root@node-a]# scdidadm -L
1      node-a:/dev/rdisk/c0t0d0 /dev/did/rdsk/d1
2      node-a:/dev/rdisk/c0t1d0 /dev/did/rdsk/d2
3      node-a:/dev/rdisk/c2t6080020FFF9E6D400000000300000000d0 /dev/did/rdsk/d3
3      node-b:/dev/rdisk/c3t6080020FFF9E6D400000000300000000d0 /dev/did/rdsk/d3
4      node-b:/dev/rdisk/c0t0d0 /dev/did/rdsk/d4
5      node-b:/dev/rdisk/c1t0d0 /dev/did/rdsk/d5
```

Sun Cluster assigns `/dev/did/rdsk/d#` numbers for all the devices the cluster node sees. Note that `/dev/did/rdsk/d3` is assigned to the same iSCSI WWN seen by both the nodes – though each node assigned unique controller number.

2. Using `/usr/cluster/bin/scsetup` command, Choose to *add quorum device* and add `d3`.
3. With that step, the basic infrastructure for Sun Cluster 3.1 Update 4 is ready.

NOTE : Make sure that the devices that are being used are not shared with other cluster environments.

Stage 4 : Installing Oracle RAC Framework for Sun Cluster

On both the nodes:

1. **For SPARC Based systems only :**
 - a) Sun Cluster dynamic lock manager packages (UDLM) are required to be added before we install the Oracle RAC framework.


```
# cd <CD mount>/ agents/components/SunCluster_Oracle_RAC_FRAMEWORK_3.1/
Solaris_10/Packages
# pkgadd -d . SUNWscor SUNWscum SUNWudlm SUNWudlmr
```

Note : SUNWjscor and SUNWcscor are for Japanese and chinese character support. Add those if necessary.

b) Oracle RAC patches must be added at this time. The patch is available with the Oracle Clusterware CD. This package is provided by Oracle to interface with the vendor clusterware product.

```
# cd <Oracle 10gR2 Clusterware CD mount >/racpatch
# cp ORCLudlm.tar /tmp
# cd /tmp
# tar -xvf ./ORCLudlm
# pkgadd -d . ORCLudlm
```

2. For X86 Based systems only :

a) Add Oracle RAC Framework agent. Note that there is no need to add UDLM packages.

```
# cd <CD mount>/agents/components/SunCluster_Oracle_RAC_FRAMEWORK_3.1/
Solaris_10/Packages
# pkgadd -d . SUNWscor SUNWscum
```

3. Add Hardware Raid agent if the RAID is managed externally – such as Sun StorageTek NAS Appliance.

```
# cd <sparc/agents/components/SunCluster_Oracle_RAC_HWRAID_3.1/Solaris_10/Packages
# pkgadd -d . SUNWschwr
```

4. From one of the cluster node, run the *scsetup* command to add *Sun Cluster support for Oracle RAC* data service which then proceeds to add the *rac-framework-rg* resource group, place them in managed state and then bring them online.

```
# scsetup
```

5. Verify the Sun Cluster Setup and verify the status of *resources rac_framework, rac_udlm*(for Sparc), *rac_hwraid* are Online. (refer to Appendix)

```
# scstat -g
```

Stage 5 : Oracle Environment Setup

The following are done to setup the environment for the Oracle user on both the cluster nodes. Make sure that same *groupid* and *userid* is provided on both nodes. Please refer to the Oracle documentation for details.

Type of Files	Mount point	Location	File Volume in NAS
ORACLE_HOME & ORA_CRS_HOME	/oracle/products/db/10.2.0 /oracle/oracrs/10.2.0	NAS (NFS) / Local	oraclebin
OCR/Voting Location	/clusterware/ocr/ocr.dbf /clusterware/voting/voting.dbf	NAS (NFS)	clusterware
RAC Database Files	/oradata	NAS (NFS)	oradata

On each node, make the directories and assign ownership and access rights.

```
# mkdir /orahome
```

```
# mkdir /oracle
# mkdir /clusterware

# groupadd -g 2000001 oinstall
# groupadd -g 2000002 dba
# useradd -u 5000001 -d /orahome -g "oinstall" -G "dba" -m -s /bin/ksh
oracle
# chown -R oracle:oinstall /orahome /oracle /clusterware
# chmod -R 775 /orahome /oracle /clusterware
# passwd oracle < Provide same password on both nodes>
```

Make entries in */etc/vfstab* so that the volumes can be mounted automatically upon system reboots. Make sure that the *rsize* and *wsize* are atleast 32K. Direct I/O must be used on OCR, Voting and database files. Direct I/O should not be used for Oracle binaries mount point.

As oracle user, make sure to add the host names in *~/.rhosts* . Otherwise, rsh may fail and causes “*User Equivalence Failed*” during Oracle's *runcluvfy* command.

Mount Options :

Type of File	NFS Mount Options
Oracle Clusterware files (Voting and OCR) & Oracle Database Files	rw,bg,hard,nointr,rsize=32768,wsiz=32768,proto=tcp,vers=3,noac,forcedirectio
Oracle Clusterware binaries, Oracle Database binaries	rw,bg,hard,nointr,rsize=32768,wsiz=32768,proto=tcp,vers=3,noac,suid

Descriptions for the NFS mount options :

Mount Options	Description
<i>hard / soft</i>	If “hard” mounted, the Oracle server does not crash if the underlying file system is unavailable (either due to planned or unplanned down time for NAS appliance). If “soft” mounted, then the Oracle server may crash if the database volumes – especially for control and redo log volumes are unavailable. It is strongly recommended to use “hard” mount for Oracle databases.
<i>rw / ro</i>	The mode of file system access. read-write or read-only.
<i>suid</i>	This option instructs the Oracle Database server to honor the set-uid bit on files mounted at the mount point. If you are placing any of the Oracle binaries onto the StorageTek NAS Appliance then this setting must be used.
<i>intr</i>	This option makes the connection interruptible. When combined with the “hard” option, this parameter enables the administrator to interrupt the application that is holding the NFS connection.
<i>forcedirectio</i>	This option enables Direct I/O for all files under the mount point and

Mount Options	Description
	allows the Log Writer to efficiently transmit the data to storage without the latency associated with splitting I/O into page size transfers. Recommended for data files and redo logs. <i>Note: This option is not to be used on Oracle binary location or Cluster Registry Files.</i>
bg	This option indicate that mount is to retry in the background if the server's mount daemon does not respond.
vers	The “vers=” option enables the administrator to specify the NFS version to use. There are two versions of NFS: version 2 and version 3. Version 3 supports additional file operations that version 2 did not have.
proto	It is the type of IP connection to be used. Sun and Oracle Suggest using TCP. TCP enforces important requirements, such as packet ordering and packet acknowledgment for reacting to various network anomalies.
rsize & wsize	These are to set the read and write packet sizes. It should be a multiple of 512 bytes. For oracle databases, set it to 32768 (32K).
llock	This NFS mount option was introduced to address performance issues by utilizing local file locking rather than NFS' Network Lock Manager (NLM). This has the potential to dramatically improve performance.
timeo	Timeout in 1/10 of the second. Specifying 600, it is 60 seconds.

Sun StorageTek 5000 NAS appliance have the business continuity premium features such as snapshot and remote replication which can be implemented for RAC database. Please refer to the documents under <http://www.sun.com/storagetek/oracle>

Stage 6 : Oracle 10gR2 Clusterware Installation

All the steps need to be performed from one of the node – unless and otherwise specified.

1. Load the Oracle 10gR2 Clusterware CD and run the cluster preinstall check.

```
$ ./runclufy.sh stage -pre crsinst -n node-a,node-b -verbose
```

Since we are leveraging Sun Cluster, all the necessary cluster packages are checked by the utility. The checking would fail on VIP nodes but that can be ignored at this time.

2. For X86 based Solaris, it is mandatory to run the *rootpre.sh* from the Clusterware install CD (or) image.

```
# cd <location>/clusterware/rootpre
# ./rootpre.sh
```

3. Set the DISPLAY variable and launch the universal installer GUI.

```
$ ./runIntaller
```

4. In the cluster configuration page, ithe GUI displays “*clusternode1-priv*” and “*clusternode2-priv*” as private node names which are derived from Sun Cluster.

5. In the subsequent page, change the following

- a) Host IP/name of the public network as “*public*”
- b) *clprivenet0* as “*private*”
- c) The rest of the private interfaces as “*do not use*”

6. Choose *External Redundancy* and provide one location for OCR file.

```
/clusterware/ocr/ocr.dbf
```

7. Choose *External Redundancy* and provide one location for Voting disks.

```
/clusterware/voting/voting.dbf
```

Note : *External Redundancy* is used because, the NAS appliance uses RAID-5 to stripe data across drives. If additional protection is required, then it is recommended to create RAID groups out of different set of drives and create filevolume to store additional copies of clusterware files.

8. Run *oralnstRoot.sh* and *root.sh* when prompted for on both the nodes one at a time and not to be run in parallel. The script formats the voting and OCR files. On the second node, while the *root.sh* configures the clusterware file, it attempts to configure VIP by doing a silent launch of *vipca*. Depending on the IP address (if the address is of RFC1918 compliant), the registration of resources such as VIP, GDS and ONS may not occur.

9. If the previous step was run successful without errors, skip this step. Otherwise, from one of the node, as root user, set the DISPLAY parameter and launch *vipca* utility.

```
# $ORA_CRS_HOME/bin/vipca
```

Provide the correct VIP addresses for both the nodes. The GUI then registers the VIP,GSD and ONS resources to the CRS. Then it starts those resources.

The Oracle Clusterware installation can be validated using the *crs_stat* command (after running the *root.sh* from the last node). With that step, the Oracle Clusterware installation is complete.

Stage 7 : Oracle RAC Database binaries Installation

Please refer to Oracle documentation for step-by-step details on Oracle RAC DB installation. It is possible to install the Oracle database binaries on either shared home on NAS or local to the nodes.

- If a shared Oracle home is used, then ORACLE_HOME must be the same across nodes. A soft link can be created which points to the NAS directory (or) use the same mount point.


```
/oracle/products/db/10.2.0
```
- As oracle user, initiate the universal installer for Oracle RAC database binary installation


```
$ export DISPLAY=<node>:0.0
$ cd <Database binary location>/.runInstaller
```
- Provide the following information when asked for :
 1. \$ORACLE_HOME location to install Oracle RDBMS binaries (*/oracle/products/db/10.2.0*). Make sure that the directory is empty and no other RAC or Oracle binaries are installed in that directory.
 2. Provide system group – either “dba” or “oinstall”
 3. Choose to install “Enterprise Edition”
 4. Choose to install “RAC database binaries” and make sure that both the nodes are listed and checked.
 5. You may choose to install “Database Software Only” at this time.

The database binaries would be installed in the directory location `/oracle/products/db/10.2.0` which is shared between nodes. Hence no remote copy operation occurs. When prompted, as root user, run the `/oracle/products/db/10.2.0/root.sh` on both the nodes one after the other. With that step, the installation of Oracle RAC binaries is complete.

Database Creation

The last step is to create the Oracle RAC database – which can either be created using “*dbca*” GUI tool or using manual scripts. The control files and database files are shared amount the RAC instances. Each instance will have its own Online redo logs. At the event of a node failure, the other surviving node performs the instance recovery which makes Oracle RAC infrastructure highly available.

If *dbca* is used, it will also make appropriate entries in *listener.ora*, *tnsnames.ora* and *sqlnet.ora* files. They are, by default, would be in `$ORACLE_HOME/network/admin` directory. Refer to Appendix for sample files.

All the database files such as control files, online redo logs and datafiles are stored in a single filevolume. That way, it becomes easier to add storage to one volume. It is recommended to use another filevolume to store archived redo logs and FRA – preferably using a different LUN.

The following are the requirements for the Oracle RAC Database.

1. Each instance has its own redo log files (threads) and shares the control files and database files. For easier and efficient management, it is sufficient to have one large file volume to store all the database files and online redo logs. If additional cluster node is added, then an additional set of redo logs with a new thread number are to be created and enabled.
2. Each instance name and instance number must be unique. The `$ORACLE_HOME/rdbms/log` directory (by default) is used as repository for files such as alert logs and trace files. The file names typically have the instance name.
3. The startup and shutdown of listener and database instances are handled by Oracle Clusterware and not by Sun Cluster. To bring them up automatically after system bootups, make the appropriate entry into `/var/opt/oracle/oratab` file :
`$ORACLE_SID:$ORACLE_HOME:<N|Y>`

Conclusion

Sun StorageTek NAS offers a perfect and easy platform for Oracle RAC deployment for easier implementation and maintenance. StorageTek NAS also offers iSCSI luns that can be used for data as well as quorum disk for Cluster. Using Sun Cluster 3.1 with Oracle Clusterware provides superior availability, scalability, stability and performance to the Oracle 10g RAC infrastructure.

Reference

Sun Cluster Data Service for Oracle Real Application Clusters guide for Solaris OS (Part # 819-0583-10)
Sun StorageTek 5320 NAS Documentation (<http://www.sun.com/storagetek/nas/5320/>)
Sun StorageTek 5320 Oracle Usage Documentation (<http://www.sun.com/storagetek/oracle>)
Oracle 10.2.0.1 Clusterware Documentation
Oracle 10.2.0.2 Database Documentation
Solaris 10 Documentation

Appendix

/etc/hosts

```
# cat /etc/hosts
#
# Internet host table
#
127.0.0.1      localhost
172.20.98.85   node-a        loghost
172.20.98.86   node-b
## VIP Nodes
172.20.200.200 node-a-vip
172.20.200.201 node-b-vip
```

For X86, /boot/solaris/filelist.ramdisk

```
# cat filelist.ramdisk
etc/rtc_config
etc/system
etc/name_to_major
etc/driver_aliases
etc/name_to_sysnum
etc/dacf.conf
etc/driver_classes
etc/path_to_inst
etc/mach
etc/devices/devid_cache
etc/devices/mdi_scsi_vhci_cache
etc/devices/mdi_ib_cache
kernel
platform/i86pc/biosint
platform/i86pc/kernel
boot/solaris.xpm
boot/solaris/bootenv.rc
boot/solaris/devicedb/master
boot/acpi/tables
etc/cluster/nodeid
```

/etc/vfstab

```
# Oracle Binaries
b20-5310a:ora10gbin - /oracle nfs - yes
rw,bg,hard,nointr,rsize=32768,wspace=32768,vers=3,noac,proto=tcp,suid
# OCR and Voting Files
b20-5310a:clusterware - /clusterware nfs - yes
rw,bg,hard,nointr,rsize=32768,wspace=32768,vers=3,noac,proto=tcp,forcedirectio
# Database Files
b20-5310a:oradata - /oradata nfs - yes
rw,bg,hard,nointr,rsize=32768,wspace=32768,vers=3,noac,proto=tcp,forcedirectio
```

ifconfig command

```
# ifconfig -a
lo0: flags=2001000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4,VIRTUAL> mtu 8232 index 1
    inet 127.0.0.1 netmask ff000000
bge0: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 2
    inet 172.20.98.85 netmask ffffffff broadcast 172.20.98.255
    groupname sc_ipmp0
    ether 0:14:4f:2a:93:3a
bge0:1: flags=1040843<UP,BROADCAST,RUNNING,MULTICAST,DEPRECATED,IPv4> mtu 1500 index
2
    inet 172.20.200.200 netmask ffffffff broadcast 172.20.200.255
bge1: flags=1008843<UP,BROADCAST,RUNNING,MULTICAST,PRIVATE,IPv4> mtu 1500 index 4
    inet 172.16.0.129 netmask ffffffff broadcast 172.16.0.255
    ether 0:14:4f:2a:93:3b
bge2: flags=1008843<UP,BROADCAST,RUNNING,MULTICAST,PRIVATE,IPv4> mtu 1500 index 3
    inet 172.16.1.1 netmask ffffffff broadcast 172.16.1.127
    ether 0:14:4f:2a:93:3c
clprivnet0: flags=1009843<UP,BROADCAST,RUNNING,MULTICAST,MULTI_BCAST,PRIVATE,IPv4>
mtu 1500 index 5
    inet 172.16.193.1 netmask ffffffff broadcast 172.16.193.255
    ether 0:0:0:0:0:1
lo0: flags=2002000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv6,VIRTUAL> mtu 8252 index 1
    inet6 ::1/128
bge1: flags=2008841<UP,RUNNING,MULTICAST,PRIVATE,IPv6> mtu 1500 index 4
    inet6 fe80::214:4fff:fe2a:933b/10
    ether 0:14:4f:2a:93:3b
bge2: flags=2008841<UP,RUNNING,MULTICAST,PRIVATE,IPv6> mtu 1500 index 3
    inet6 fe80::214:4fff:fe2a:933c/10
    ether 0:14:4f:2a:93:3c
```

For X86, scstat -g (notice no rac_udlm resources)

```
# scstat -g

-- Resource Groups and Resources --

          Group Name          Resources
          -----
Resources: rac-framework-rg   rac_framework rac_hwraid

-- Resource Groups --

          Group Name          Node Name          State
          -----
Group: rac-framework-rg      node-b             Online
Group: rac-framework-rg      node-a             Online

-- Resources --

          Resource Name       Node Name          State      Status Message
          -----
Resource: rac_framework      node-b             Online     Online
Resource: rac_framework      node-a             Online     Online

Resource: rac_hwraid         node-b             Online     Online
Resource: rac_hwraid         node-a             Online     Online
```

For Sparc system,

```
# scstat -g

-- Resource Groups and Resources --

      Group Name          Resources
      -----
Resources: rac-framework-rg  rac_framework rac_udlm rac_hwraid

-- Resource Groups --

      Group Name          Node Name          State
      -----
Group: rac-framework-rg    node-a             Online
Group: rac-framework-rg    node-b             Online

-- Resources --

      Resource Name       Node Name          State      Status Message
      -----
Resource: rac_framework  node-a             Online     Online
Resource: rac_framework  node-b             Online     Online

Resource: rac_udlm       node-a             Online     Online
Resource: rac_udlm       node-b             Online     Online

Resource: rac_hwraid     node-a             Online     Online
Resource: rac_hwraid     node-b             Online     Online
```

Sparc, scstat command

```
# scstat
-----

-- Cluster Nodes --

      Node name          Status
      -----
Cluster node:  node-a             Online
Cluster node:  node-b             Online

-----

-- Cluster Transport Paths --

      Endpoint          Endpoint          Status
      -----
Transport path:  node-a:bge2       node-b:bge2       Path online
Transport path:  node-a:bge1       node-b:bge1       Path online

-----

-- Quorum Summary --
```

```

Quorum votes possible:      3
Quorum votes needed:       2
Quorum votes present:      3

-- Quorum Votes by Node --

          Node Name          Present Possible Status
          -----          -
Node votes:   node-a          1         1      Online
Node votes:   node-b          1         1      Online

-- Quorum Votes by Device --

          Device Name          Present Possible Status
          -----          -
Device votes: /dev/did/rdisk/d7s2 1         1      Online
-----

-- Device Group Servers --

          Device Group          Primary          Secondary
          -----          -
-----

-- Device Group Status --

          Device Group          Status
          -----          -
-----

-- Multi-owner Device Groups --

          Device Group          Online Status
          -----          -
-----

-----

-- Resource Groups and Resources --

          Group Name          Resources
          -----          -
Resources: rac-framework-rg   rac_framework rac_udlm rac_hwraid

-- Resource Groups --

          Group Name          Node Name          State
          -----          -
Group: rac-framework-rg     node-a            Online
Group: rac-framework-rg     node-b            Online

-- Resources --

```

Resource Name	Node Name	State	Status Message
Resource: rac_framework	node-a	Online	Online
Resource: rac_framework	node-b	Online	Online
Resource: rac_udlm	node-a	Online	Online
Resource: rac_udlm	node-b	Online	Online
Resource: rac_hwraid	node-a	Online	Online
Resource: rac_hwraid	node-b	Online	Online

-- IPMP Groups --

Node Name	Group	Status	Adapter	Status
IPMP Group: node-a	sc_ipmp0	Online	bge0	Online
IPMP Group: node-b	sc_ipmp0	Online	bge0	Online

iSCSI command examples (in the nodes) :

```
#iscsiadm list discovery
Discovery:
  Static: disabled
  Send Targets: disabled
  iSNS: disabled
```

To Enable static configuration:

```
# iscsiadm modify discovery --sendtargets disable
# iscsiadm modify discovery --static-config enable
#iscsiadm add static-config iqn.1986-
03.com.sun:01:000e0c6754fa.456E271F.iscsiqdsparc,172.20.98.198
iqn.1986-03.com.sun:01:000e0c6754fa.456f1153.iscsiqdx86,172.20.98.198
# iscsiadm list static-config
```

To disable static configuration :

```
# iscsiadm modify discovery --static-config disable
```

To enable dynamic (send targets) discovery

```
# iscsiadm modify discovery --sendtargets enable
# iscsiadm add discovery-address 172.20.98.198:3260
# iscsiadm list discovery-address -v 172.20.98.198:3260
```

To disable dynamic (send targets) discovery

```
# iscsiadm modify discovery --sendtargets disable
```

Oracle Clusterware Verification :

```
-bash-3.00$ crs_stat -t
Name                Type                Target              State              Host
-----
ora....e-a.gsd      application         ONLINE              ONLINE              node-a
ora....e-a.ons      application         ONLINE              ONLINE              node-a
ora....e-a.vip      application         ONLINE              ONLINE              node-a
ora....e-b.gsd      application         ONLINE              ONLINE              node-b
ora....e-b.ons      application         ONLINE              ONLINE              node-b
ora....e-b.vip      application         ONLINE              ONLINE              node-b
```

Setting the misscount (as root user)

```
$ORA_CRS_HOME/bin/crsctl set css misscount 600
```

iSCSI commands

```
# iscsiadm modify discovery --sendtargets {enable| disable}
# iscsiadm add discovery-address <NAS ip address>:3260
# iscsiadm list discovery
# iscsiadm list discovery-address
# iscsiadm list discovery-address -v <NAS ip address>:3260

# iscsiadm modify discovery --static {enable | disable}
# iscsiadm {add | remove} static-config <iscsi name>,<NAS ip address>
# iscsiadm list static-config
```

\$ORACLE_HOME/tnsnames.ora (Clients & RAC Nodes)

```
racdb1 =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = node-a)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SID = racdb1)
    )
  )

racdb2 =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = node-b)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SID = racdb2)
    )
  )
```

\$ORACLE_HOME/listener.ora (RAC Nodes)

```

LISTENER_node-a =
  (DESCRIPTION_LIST =
    (DESCRIPTION =
      (ADDRESS_LIST =
        (ADDRESS = (PROTOCOL = IPC)(KEY = EXTPROC))
        (ADDRESS = (PROTOCOL = TCP)(HOST = node-a)(PORT = 1521))
      )
    )
  )

SID_LIST_LISTENER_node-a =
  (SID_LIST =
    (SID_DESC =
      (SID_NAME = PLSExtProc)
      (ORACLE_HOME = /oracle/products/db/10.2.0)
      (PROGRAM = extproc)
    )
    (SID_DESC =
      (GLOBAL_DBNAME = racdb)
      (ORACLE_HOME = /oracle/products/db/10.2.0)
      (SID_NAME = racdb1)
    )
  )

LISTENER_node-b =
  (DESCRIPTION_LIST =
    (DESCRIPTION =
      (ADDRESS_LIST =
        (ADDRESS = (PROTOCOL = IPC)(KEY = EXTPROC))
        (ADDRESS = (PROTOCOL = TCP)(HOST = node-b)(PORT = 1521))
      )
    )
  )

SID_LIST_LISTENER_node-b =
  (SID_LIST =
    (SID_DESC =
      (SID_NAME = PLSExtProc)
      (ORACLE_HOME = /oracle/products/db/10.2.0)
      (PROGRAM = extproc)
    )
    (SID_DESC =
      (GLOBAL_DBNAME = racdb)
      (ORACLE_HOME = /oracle/products/db/10.2.0)
      (SID_NAME = racdb2)
    )
  )

STARTUP_WAIT_TIME_LISTENER=0
CONNECT_TIMEOUT_LISTENER=10
TRACE_LEVEL_LISTENER=OFF
TRACE_DIRECTORY_LISTENER=/oracle/products/db/10.2.0/network/trace
TRACE File_LISTENER=listener.trc
LOG_DIRECTORY_LISTENER=/oracle/products/db/10.2.0/network/log
LOG_FILE_LISTENER=listener.log

```