

ServiceGuard and Oracle 9i RAC

Asghar Riahi

Senior Technical Consultant
HP Technical Services



Objectives

- What is HA
- Overview of MC/ServiceGuard
- Latest News
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O

Objectives (Cont.)

- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)

Prerequisites

1. System administration expertise with HP-UX using command-line interfaces
 - 10.X and 11.X
 - File system layout
 - Software distributor usage
 - Journaled File System
2. Expertise with the command-line interface for LVM
3. Expertise with system hardware configuration (ioscan, lsssf, lanscan, ifconfig, netstat, etc.)
4. Basic network administration expertise
5. MC/ServiceGuard expertise
6. Oracle basic knowledge (not a must, but helpful)

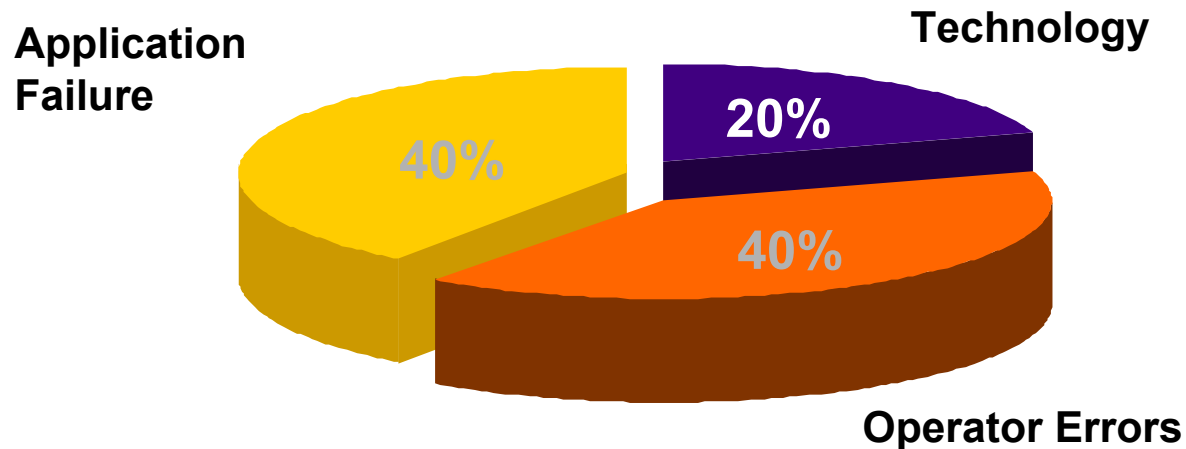
- **What is HA**

- Overview of MC/ServiceGuard
- Latest News
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Causes of Failure






industry reports on causes of unplanned downtime



Source: GartnerGroup, October 1999






Causes and Solutions for SPOF's

SPOF

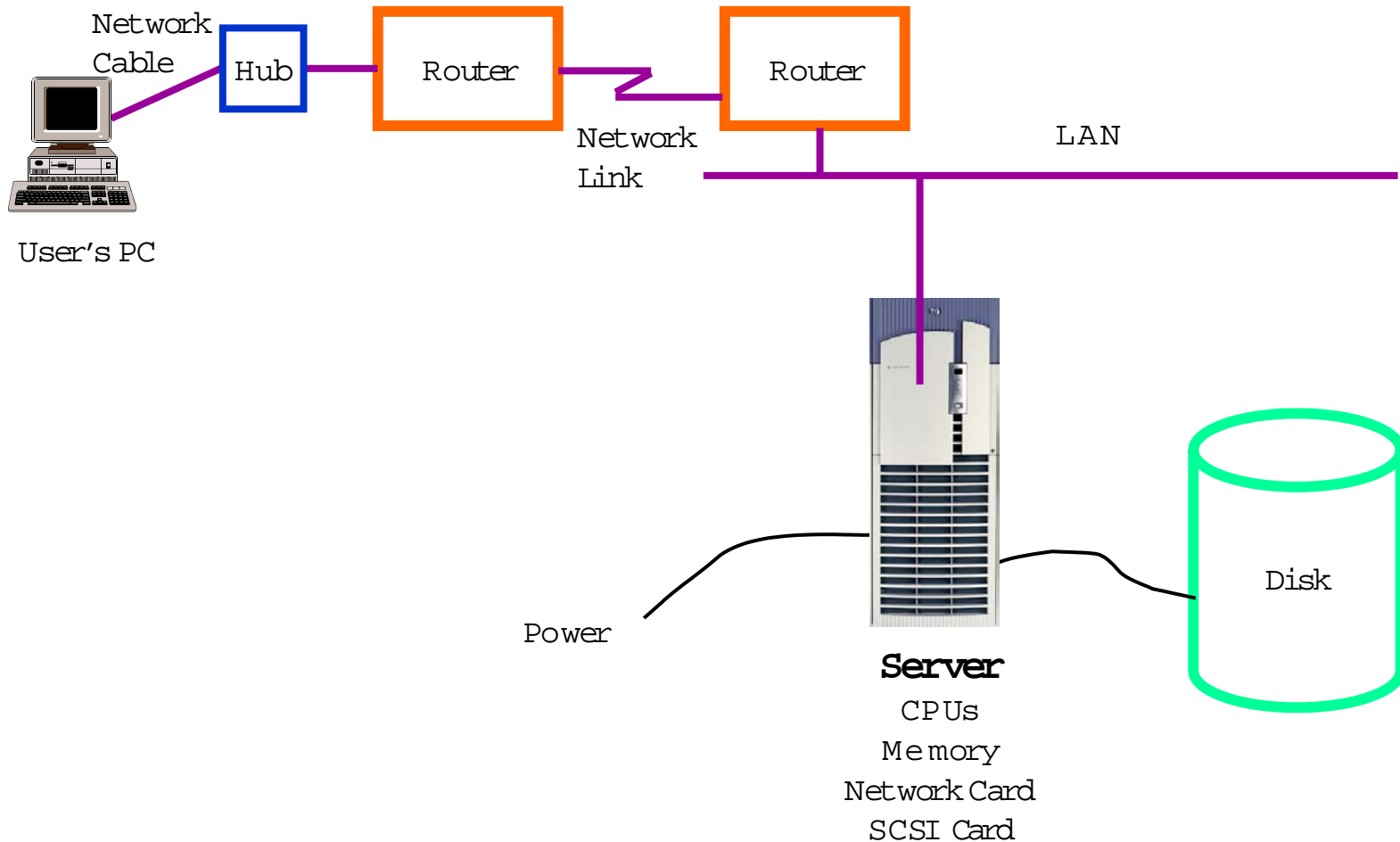
-  CPU, Memory
-  Disks
-  Interface Cards
-  LAN, NICs
-  Power



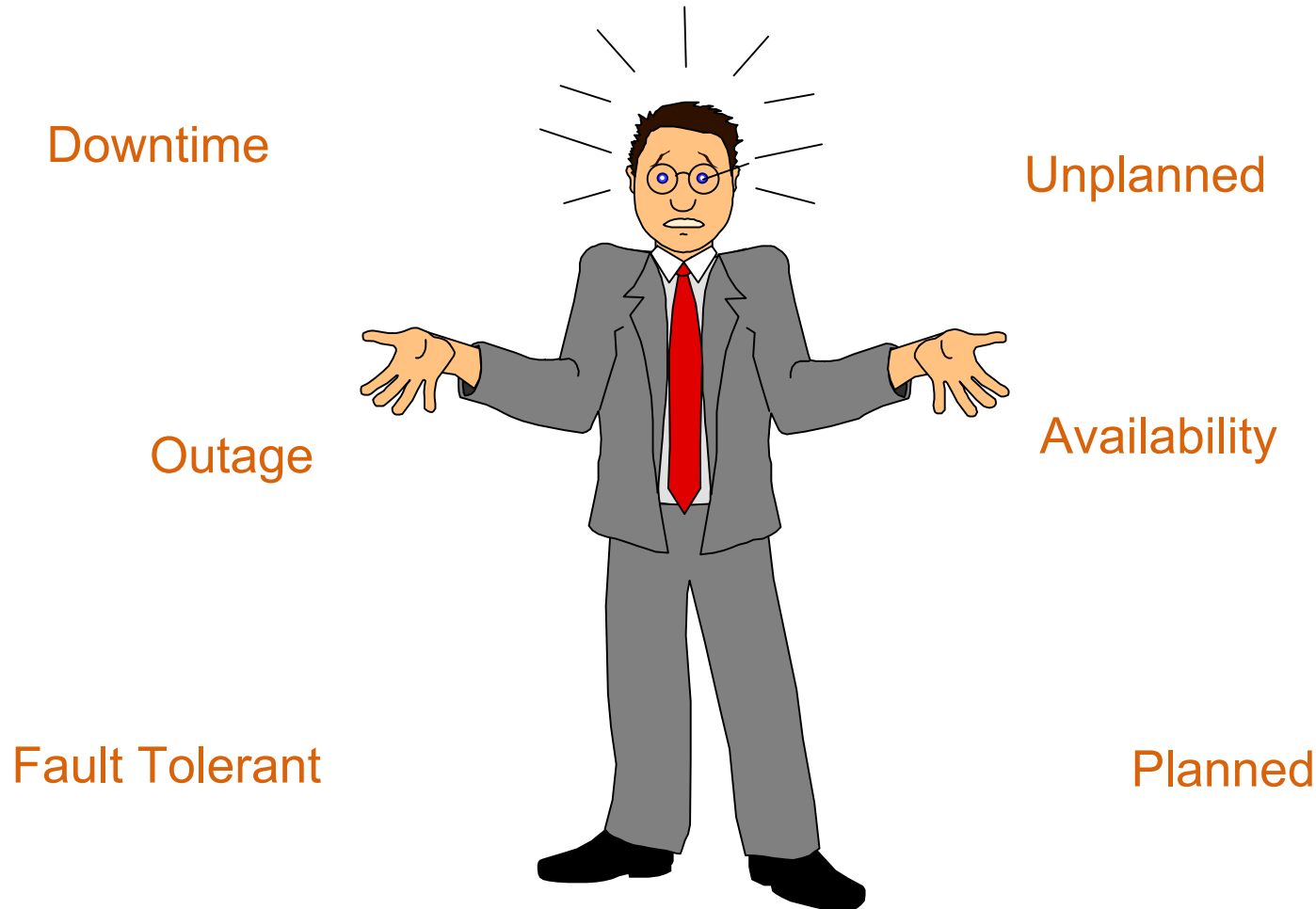
Potential Solution

-  Clusters
-  Mirroring and RAID
-  Mirroring and PV Links
-  Redundant LANs and LANIC
-  UPS, Diesel generator

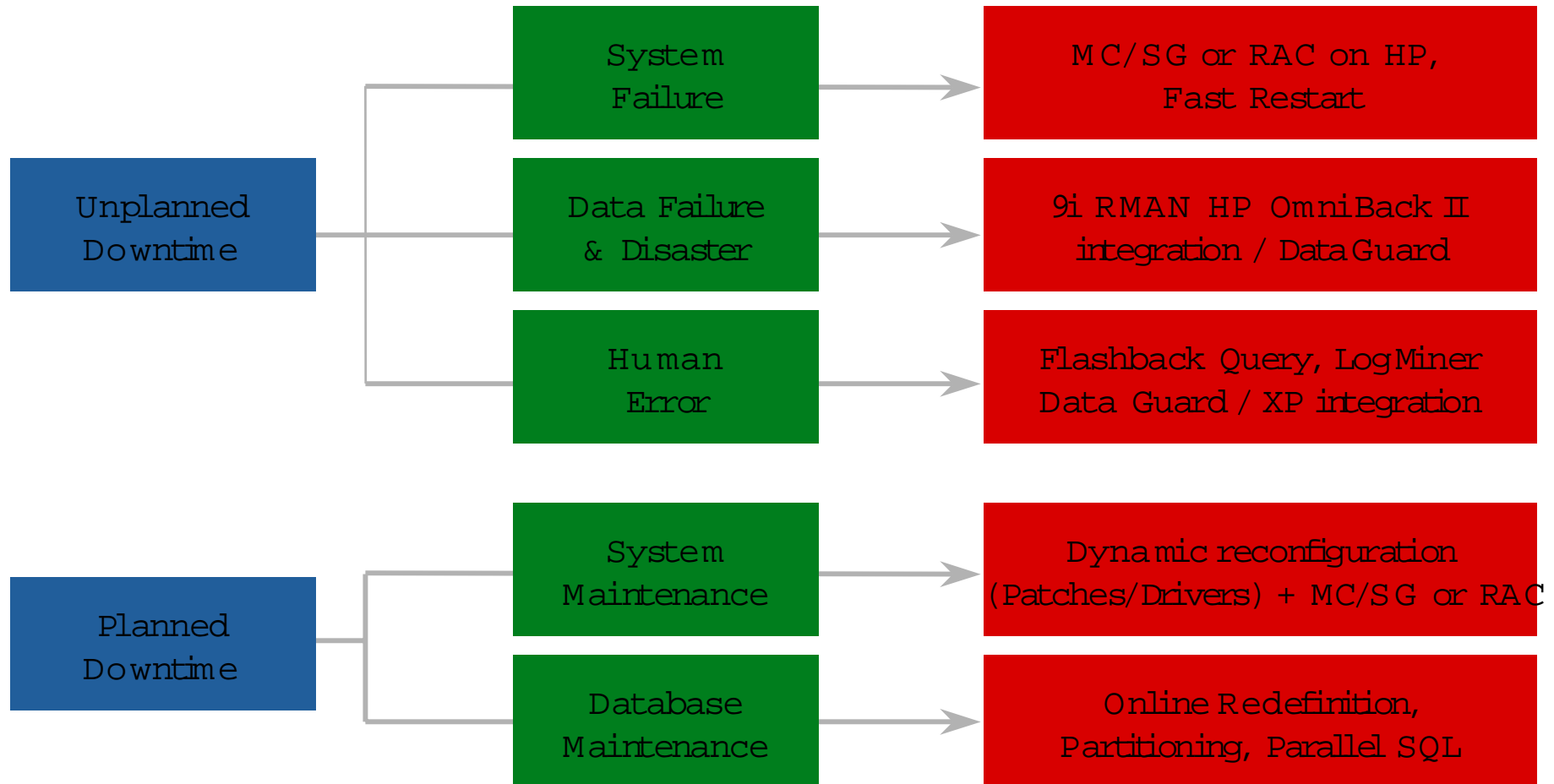
Single Points of Failure for an Application



High Availability Terms



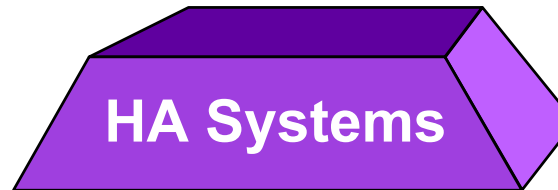
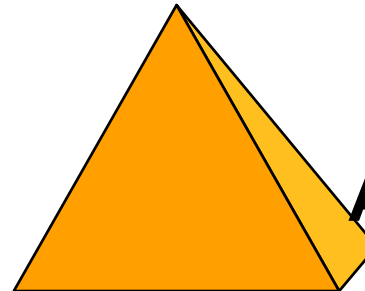
HP and Oracle handle the cause of downtime



High Availability Solutions

Future HP Products
Transaction Process Monitoring
5 Nines: 5 Minutes

**Continuously
Available Systems**



HA Systems

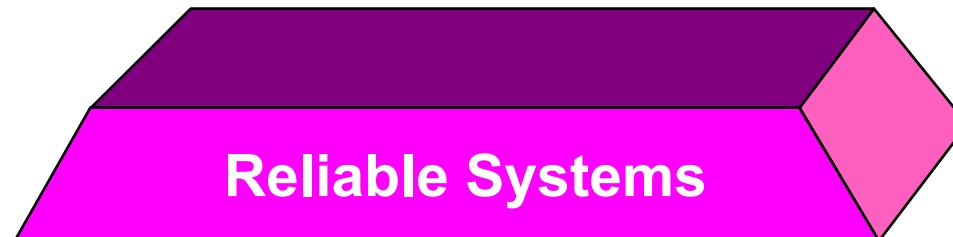
MC/SG
SG/OPS
PRM

Cluster Manager (GUI)
ClusterView

MirrorDisk/UX
XP512/ EMC/ Arrays
Advanced JFS
Hot Swap Disk



Protected Data



Reliable Systems

HP 9000
HP Peripherals
HP-UX

- What is HA
- **Overview of MC/ServiceGuard**
- Latest News
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Features and Benefits of MC/ServiceGuard

- **Completely transparent to applications**
- **Intelligent cluster reconfiguration after node failure**
 - **Data Integrity: No 'split-brain' syndrome**
 - **Dynamic formation of new, viable cluster**
- **Flexible load balancing**
- **Facilitates online hardware and software updates**
- **Highly available Enterprise Cluster**
 - **Fast switching of applications to alternate node (<60 seconds for basic system resources with JFS)**
 - **LAN failure protection (very fast local switch to standby LAN adapter inside same node)**

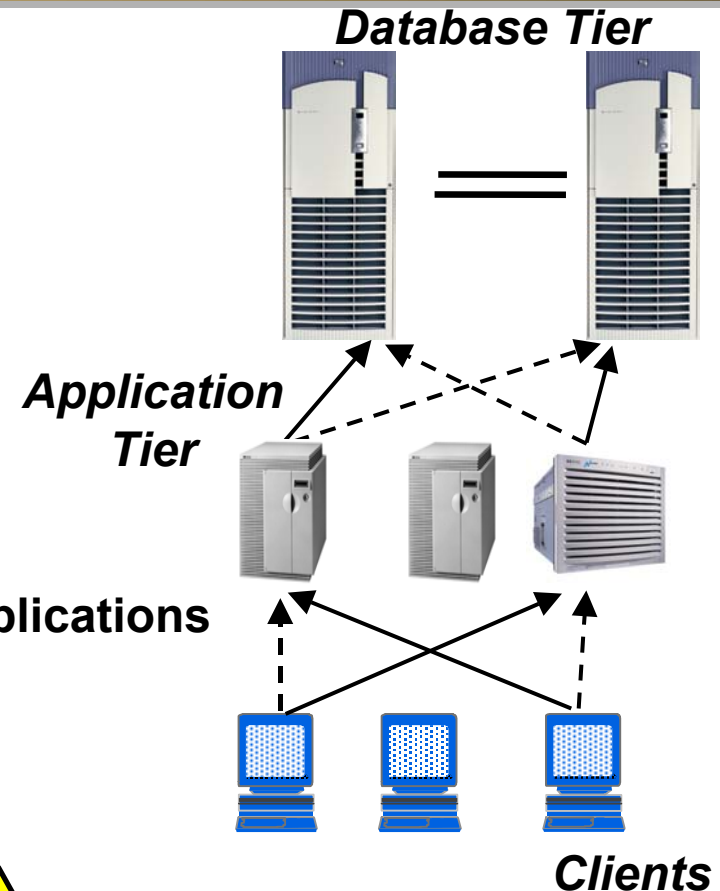
Features and Benefits of MC/ServiceGuard (Cont.)

- **Application Packages**

- Easy application management
- Flexible recovery options

- **No idle resources**

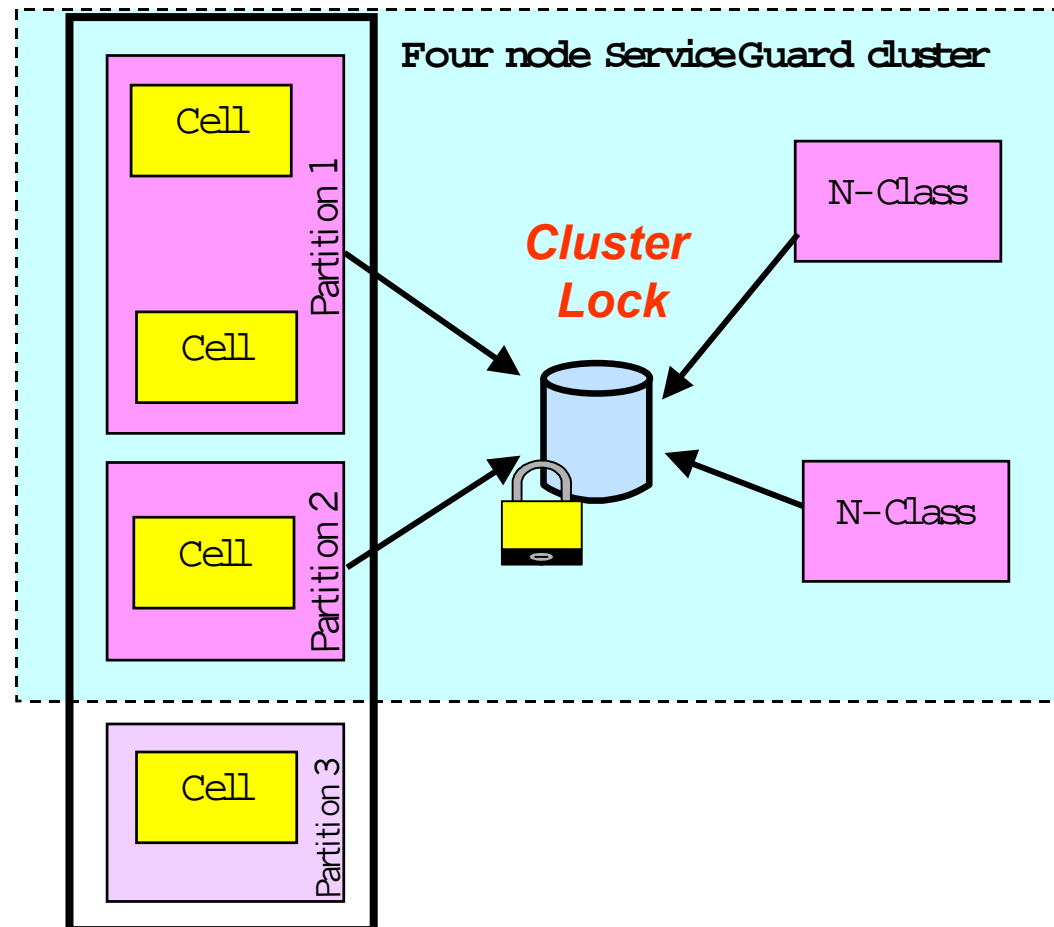
- All systems run mission-critical applications



Mixed Clusters with 11.0 and 11i Supported

- Mixed clusters where some nodes are running HP-UX 11.0 and some nodes running HP-UX 11i is supported – requires all nodes running the same SG revision (e.g. 11.09 or 11.13)
- **EXAMPLE** where two nodes in a SuperDome are running HP-UX 11i and two N-class nodes are running 11.0

SuperDome System



MC/ServiceGuard History



- Initial release of MC/SG
 - 4 nodes
 - Packages & Relocatable IP
- New Features of 10.03 SG (HP-UX 10.01)
 - service restart option
 - cmmigrate command (SwitchOver to MC/SG)
- New Features of 10.04 SG (HP-UX 10.10)
 - SAM GUI for MC/SG
 - 8 node support
 - RS-232C link for 2-node heartbeat
 - Security Enhancements (/etc/cmcluster/cmclnodelist)
- New Features of 10.05 SG (HP-UX 10.10)
 - support for ClusterView
- New Features of 10.10 SG (HP-UX 10.20)
 - online node and package configuration
 - Support for Event Monitoring Services (EMS)
 - New MC/SG commands: cmscancl, cmgetconf, cmdeleteconf
- New Features of 11.00 SG (HP-UX 11.00)
 - 16 nodes
 - New Failover and Failback Policies
 - Support for V-Class
 - Support for HP-UX 11.00
- New Features of 11.13
 - New Hardware support (Superdome, etc.)
 - Mixed clusters supported
 - VERITAS Volume Manager
 - Max packages increased to 60
 - APA auto-mode support
 - ServiceGuard Manager (GUI)
 - MC/ServiceGuard Linux (2h01)
- New Features of 11.14
 - Quorum Server
 - Parallel fsck
 - 150 Package support
 - Campus Cluster Extended distance
 - VLAN support
 - IPv6Support Phase I
 - HA Announcements
 - Including ECM Toolkit template for SG/OPS 9i integration

- What is HA
- Overview of MC/ServiceGuard
- **Latest News**
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Latest News

Serviceguard and SGeRAC A.11.15 available in Summer '03

- **Release on HP-UX 11i v1 (11.11) on AR0603**
- **Will release with HP-UX 11i v2 (11.23)**
 - HP-UX 11i v2 will provide support for new Itanium2 based servers
- **IPV6 support**
- **SGeRAC**
 - support for Oracle 10i (timing dependent on Oracle release of 10i)
 - Version rolling upgrade allows the database to be up and active while new versions of SG/SGeRAC are deployed within the cluster

Serviceguard A.11.15

- **New features in Serviceguard (A.11.15) for HP-UX 11.I (PA):**

- IPv6 support

- Automatic probe of networks with run commands

- Options to limit storage checks for faster performance

- Flight Recorder

- **Quorum Server (A.02.00)**

- Support HP-UX and Linux clusters

- Runs with Serviceguard (A.11.15) for HP-UX

- Runs with Serviceguard (A.11.14.02) for Linux

- **Serviceguard Manager (A.03.00.01)**

- Run on HP-UX, Linux, or Windows/NT

- Supports Serviceguard (A.11.15) on HP-UX

- Supports Serviceguard (A.11.4.02) on Linux

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- **SGeRAC**
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



HP ServiceGuard Extension for Oracle9i RAC



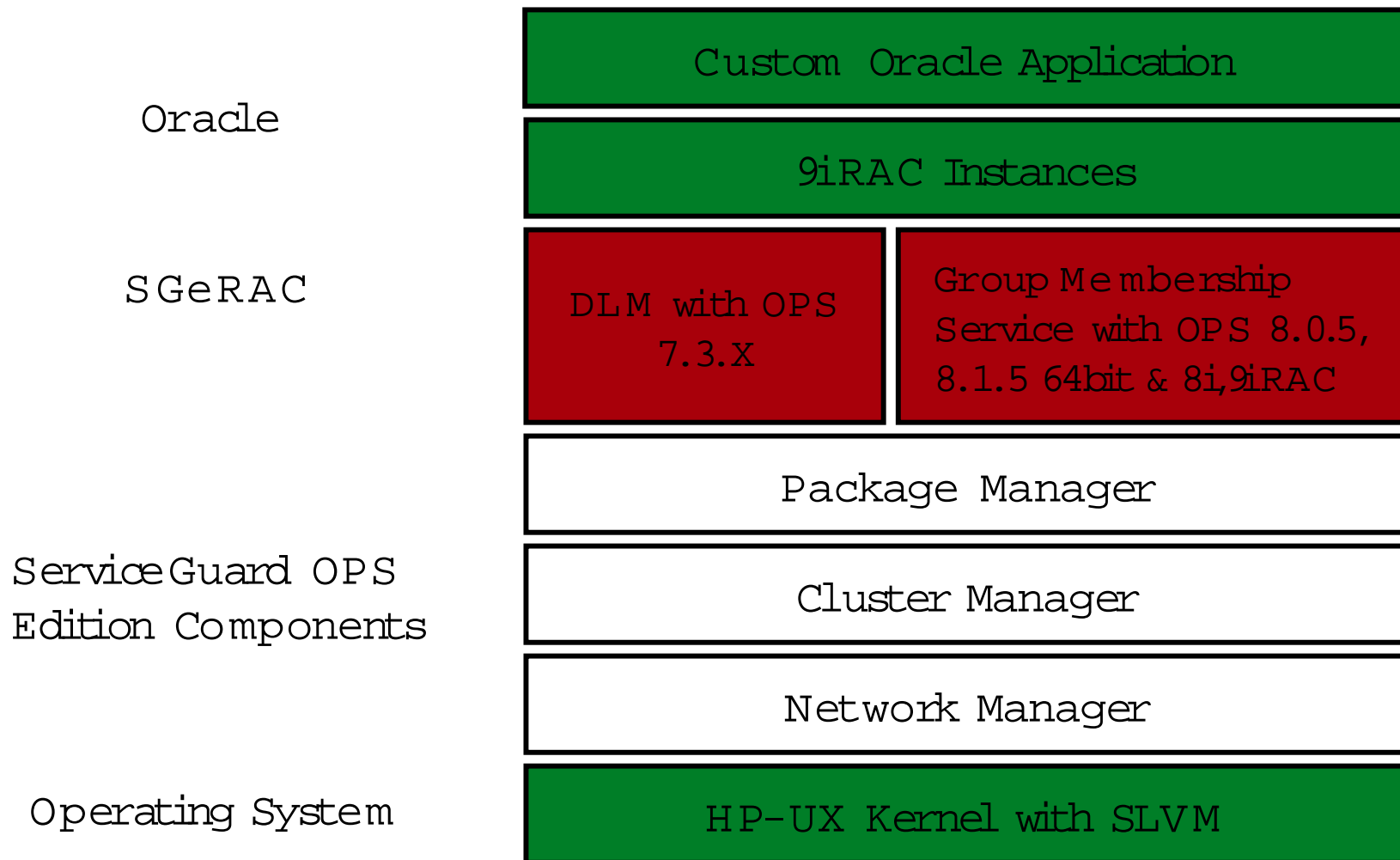
- What is ServiceGuard Extension for RAC?
- Dynamic Cluster Reformation
- ServiceGuard OPS Edition A.11.14 Features
- Configuring a SG OPS Edition
- Volume Managers for Data Management
- ServiceGuard Manager

What is ServiceGuard Extension for RAC



- formerly ServiceGuard OPS Edition, and before MC/Lockmanager
- a product especially designed for Oracle9i RAC
- new product structure with version 11.14.01 (Itanium) & 11.15 (PA-RISC) : SGeRAC as an add-on product to MC/ServiceGuard
- cluster management is provided with MC/SG
- shared disk activation comes with SGeRAC

MC/SG & SGeRAC Software Stack



ServiceGuard OPS Edition Daemons



Nine daemon processes are associated with
ServiceGuard

- /usr/sbin/cmclcd - ServiceGuard Cluster Daemon
- /usr/sbin/cmclconfd - ServiceGuard Configuration Daemon
- /usr/sbin/cmlogd - ServiceGuard Syslog Log Daemon
- /usr/sbin/cmlymd - Cluster Logical Volume Manager Daemon
- /usr/sbin/cmomd - Cluster Object Manager Daemon
- /usr/sbin/cmsnmpd - Cluster SNMP subagent (optionally running)
- /usr/sbin/cmsrvassisd - ServiceGuard Service Assistant Daemon
- /usr/sbin/cmtaped - ServiceGuard Shared Tape Daemon
- /usr/sbin/qs - ServiceGuard Quorum Server Daemon
- /usr/sbin/cmngmsd - Group Membership Daemon

SG & Oracle Release Matrix

	SG-OPS	HP-UX
9.2 RAC 64bit IPF	11.14.01	11.22
9.2 RAC 64bit	11.14 11.13	11.0 11.11
9.0.1 RAC 64bit	11.14 11.13 11.09	11.0 11.11
8.1.7 OPS 32/64bit	11.14 11.13 11.12 11.09	11.0 11.11

SGeRAC Product Numbers

- T1859BA SGeRAC for Itanium2
- B5161FA ServiceGuard OPS Edition Software & License
- B5158FA ServiceGuard OPS Edition Documentation
- B3935DAN A customer will be able to return the MC/SG License for full credit towards the cost of Serviceguard OPS Edition.

Dynamic Cluster Reformation

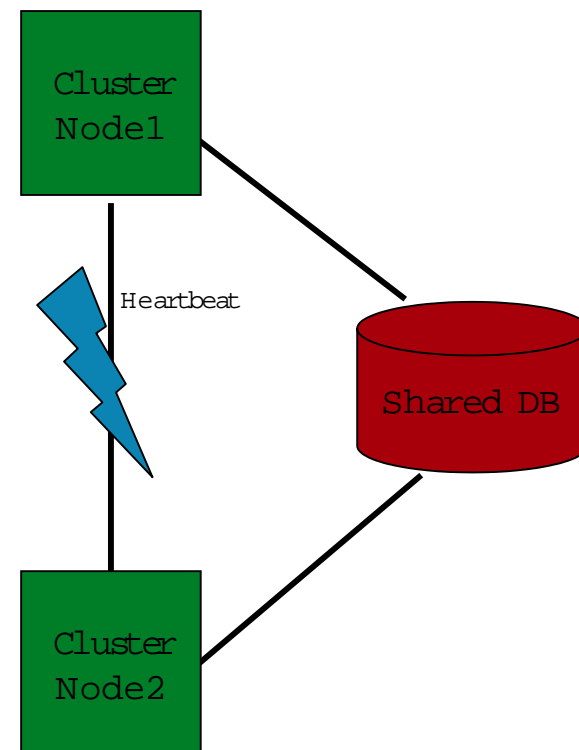
= a dynamic reformation is a temporary change in cluster membership that takes place as nodes join or leave a running cluster

Examples which requires Re-formation of the cluster:

- An SPU or network failure was detected on an active node.
- An inactive node wants to join the cluster.
- The cluster manager daemon has been started on that node.
- The system administrator halted a node.
- A node halts because of a package or service failure.
- Heavy network traffic prohibited the heartbeat signal from being received by the cluster

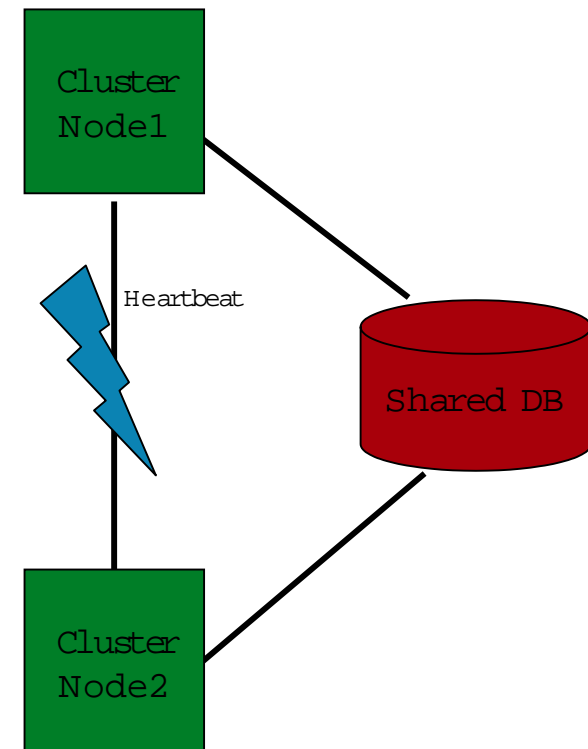
Split-Brain Syndrome

- cluster re-formation requires a cluster quorum of a strict majority (>50%)
- Split-brain: both halves of a previously running cluster have 50%
- in split-brain situations the use of a tie-breaker (cluster lock disk) is necessary.
- the cluster lock is implemented either by means of a lock disk or a quorum server.



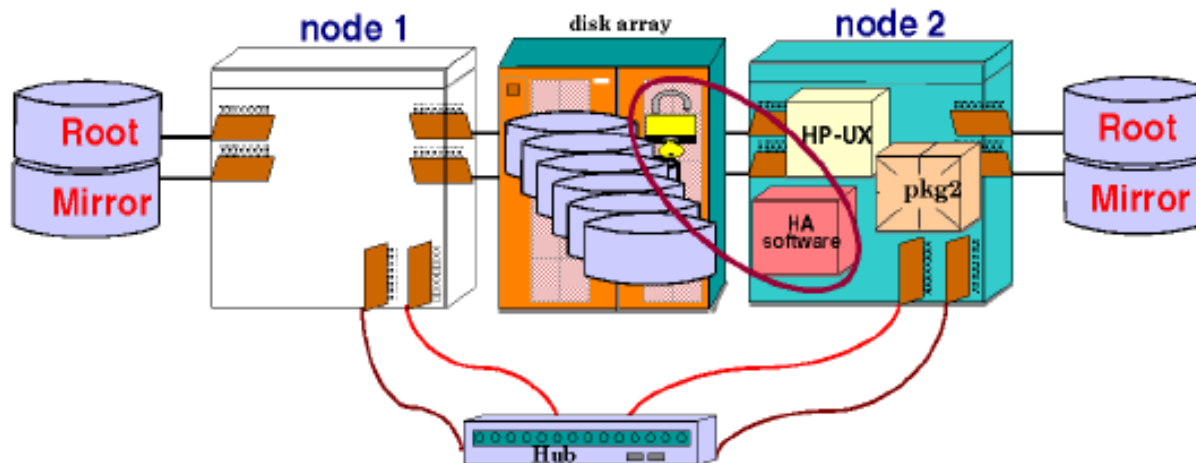
Split-Brain Syndrome (Cont.)

- without a cluster lock, a failure will cause the cluster to halt.
- two-node clusters require a cluster lock.
- in clusters larger than 3 nodes a cluster lock is strongly recommended.



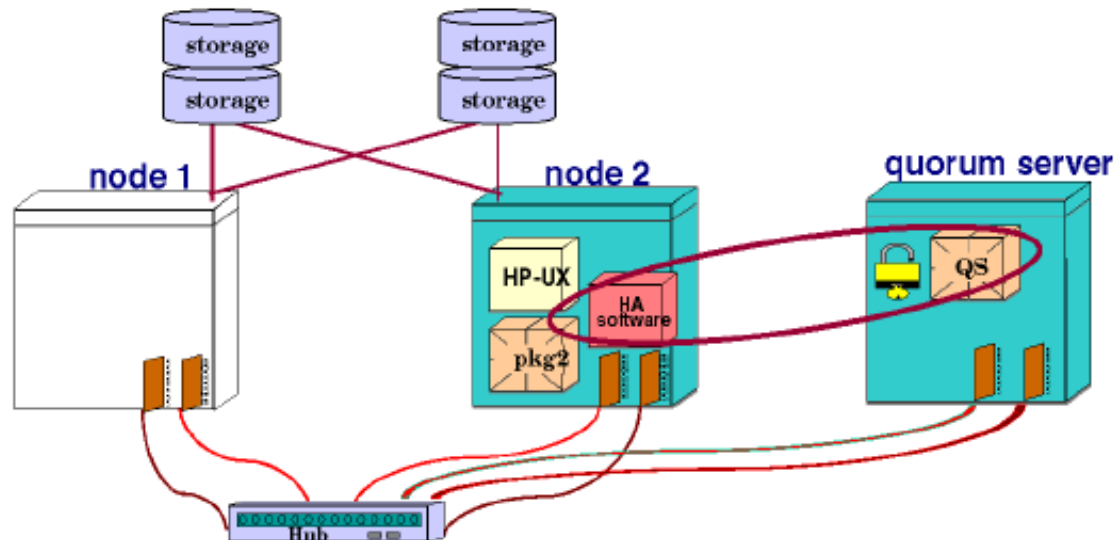
Cluster Lock: Lock Disk

- Can be used for clusters up to 4 nodes
- The cluster lock disk is a special area on a LVM disk located in a volume group that is sharable by all nodes in the cluster.



Cluster Lock: Quorum Server

- Can be used for clusters of any size.
- The quorum server is a process that runs on an HP-UX host that is external to the cluster.



- HP-UX 11.0 and 11i
- maximum of sixteen nodes in a cluster
- up to 200 relocatable package IP addresses per cluster are supported
- a total of 127 RAC instances per cluster is supported.
- multiple 9iRAC instances on the same node are supported.
- Quorum Server is supported as a type of cluster lock different from the lock disk.
- Software mirroring under LVM with MirrorDisk/UX is now supported once again for RAC data in a two-node cluster.

- support for HP Automatic Port Aggregation (APA).
support of VERITAS Volume Manager 3.2+ for HP-UX (VxVM), including the Cluster Volume Manager (CVM).
starting with A.11.13 online node addition/deletion are supported with Veritas CVM. This feature remains unsupported for clusters that use LVM.
- The same applies to adding, deleting and modifying packages.
- Even rolling patchset upgrades are supported.

- support for the HyperFabric product (B6257AA) is provided for the following configurations:

two-node point-to-point configuration.

configurations larger than two nodes with HyperFabric LAN interfaces linked through HyperFabric switches.

heartbeat and local LAN failover are not supported.

for Oracle 9i, UDP and HMP protocol both are supported for Oracle DLM traffic.

for OPS 8i, only UDP protocol is supported for Oracle DLM traffic.

SGeRAC A.11.15 (PA):

- Delivered on the June'03 Application Release
- SGeRAC (A.11.15 on HP-UX 11.11), unbundled from Serviceguard (Serviceguard sold and installed separately)
- Fast detection of Oracle instance failures
- Database and Group Membership provided for Serviceguard Manager
- Pre-enabled Rolling Upgrade for future releases
- New SGeRAC manual

Serviceguard Extension for RAC (A.11.15 on HP-UX 11.11)



SG unbundled from SGeRAC (SG + SGeRAC = SG OPS Edition)

- New name - Serviceguard Extension for RAC (SGeRAC), formerly known as SG/OPS Edition
 - With version A.11.15 for HP-UX 11.11
- New product number: T1859BA
- The old product number B5161FA will stay on CPL for (11.0)
For SGeRAC on 11i - Order new product number.
- New product structure requires SG license (B3935DA)
- New documentation product number: T1862BA
- Supports Mission Critical Operating Environment bundle

Note: All Serviceguard features and limitations will apply to SGeRAC, unless otherwise specified

Serviceguard OPS Edition / SGeRAC Summary



- Supports the following versions of Oracle OPS/RAC (Real Application Cluster).
 - SGOPS 11.13, 11.14 and SGeRAC (PA) 11.15 supports: OPS 8.1.7, RAC 9.0.1, RAC 9.2.0
 - SGeRAC 11.14.01 supports IPF RAC 9.2
 - OPFS 8.1.7, RAC GUARD 9i
- Support for HMP/HyperFabric for 9i RAC DLM (Distributed Lock Manager) communication – high speed, low latency network protocol
- Veritas CVM (Cluster Volume Manager)
 - 4-node CVM cluster on HPUX 11i
 - 10km extended cluster
 - on-line node and storage addition (see whitepaper, URL listed at end of slideset)

Serviceguard OPS Edition / SGeRAC Summary (Cont.)



- Up to 16 nodes 9i RAC cluster when SLVM (plus patch) is available on HP-UX 11.11 only
 - Patch rolling upgrade
 - Support QS, parallel fsck, VLAN, compatible with IPv6 OS, 150 package support
 - Partitioning (nPartitions, vPars):
 - support on rp5470, rp7400, rp7410, rp8400 and SuperDome servers
- * Refer to latest configuration guide and release notes for latest information*

quick reference of versions & product numbering

HP-UX HA Products	HP-UX 11.0	HP-UX 11.11	HP-UX 11i v1.6	HP-UX 11i v2
Serviceguard	B3935DA	B3935DA	B3935DA	T1905BA
SG manual	B3936EA	B3936EA	B3936EA	T1906BA
SGeRAC	B5161FA	T1859BA	T1859BA	T1907BA
SGeRAC manual	B5158FA	T1862BA	T1862BA	T1908BA
SGeSAP	B7885BA	B7885BA	N/A	T2357BA
ECMT	B5139DA	B5139DA	B5139DA	T1909BA
NFS Toolkit	B5140BA	B5140BA	N/A	B5140BA
Mirrordisk/UX	B2491BA	B2491BA	B2491BA	B2491BA

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- **Different LVM Versions**
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Different LVM Versions

- Cluster LVM (CLVM)
 - superset of 10.x/11.x LVM
 - need MC/ServiceGuard Edition to activate new features
 - exclusive activation (-a e option)
 - cluster ID
- In 10.x and 11.x the LVM commands are shipped with HP-UX; however, either SG or SG OPS is required to enable some features.

Different LVM Versions

- Shared LVM (SLVM)
part of 10.X, 11.0 LVM, needs MC/SG OPS Edition to activate new features
includes SG cluster features
friendlier vgexport and vgimport (means -s option)
shared VG activation (read and write)
cluster ID
- Standard LVM or MirrorDisk/UX features supported for 2 node cluster
bad block relocation is supported
lvsplit and lvmerge is supported
striping done with lvcreate -i or -l option (is supported starting with 11.0 and patch)
Mirror Write Cache is supported

New Export & Import Options

`vgexport -p -s -m <mapfilename> <vg_name>`
create mapfile without removing the VG,
and save the VGID for the `vgimport` command

`vgimport -s -m <mapfilename> <vg_name>`
scan for disks that have the same VG ID as in
the mapfile

Special Volume Group Options



`vgchange -c <mode> <vgname>`

`y`

write cluster ID in VGDA

`n`

remove cluster ID from VGDA

`vgchange -c y -S y <vgname>`

mark VG as shareable (MC/LM only)

`vgchange -c n -S n <vgname>`

mark VG as non-shareable (MC/LM only)

Volume Group Activation

`vgchange -a <mode> <vgname>`

Modes:

n	deactivate
y	activate normal (non-exclusive, non-shared, read/write)
r	activate read-only (10.X only)
e	activate exclusive read/write (MC/SG & MC/LM only)
s	activate shared read/write (MC/LM only)

New Display Information

The SLVM version of "vgdisplay" additionally shows:

- cluster ID
- shareable status
- server / client
- nodes which have activated the VG

LVM Configuration

- edit the /etc/lvmrc file:
 AUTO_VG_ACTIVATE=0
- custom_vg_activation() should contain *only* non-shared & non-exclusive access VGs
- vgcfgbackup must be done for the cluster lock volume group(s) after cluster configuration
- nodes must be members of the same cluster to use exclusive and shared activation
- VG names & minor numbers *must* be unique and consistent throughout the cluster

Additional Information:

Latest Support Information on SG OPS can always be found at:

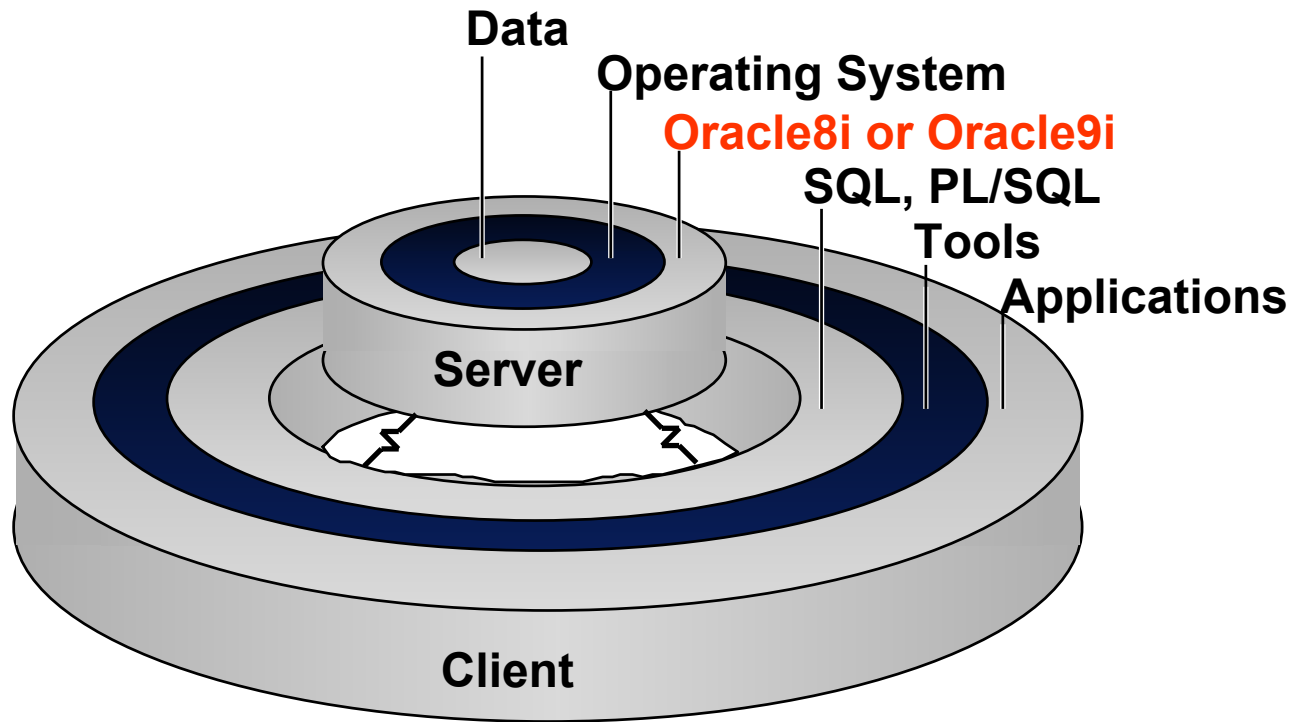
<http://www.hp.com/products1/unix/highavailability/ar/serviceops/index.html>

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- **Oracle architecture and functionality**
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Oracle instance architecture

Overview



Oracle instance architecture (cont.)

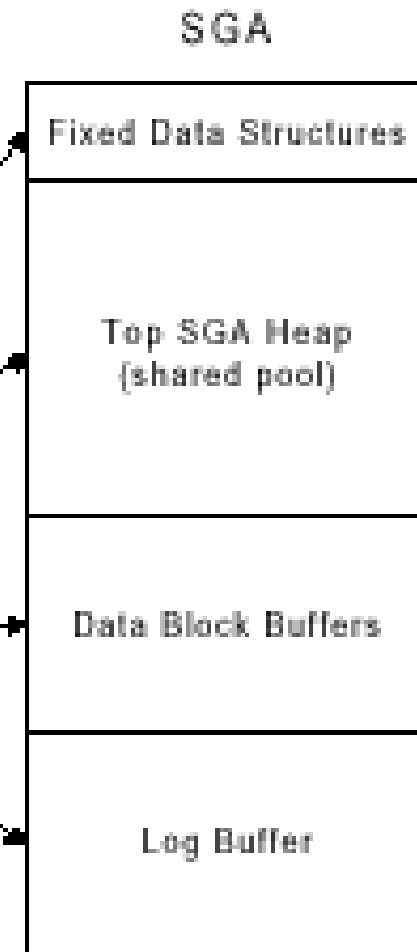
What Oracle consists of

- Oracle Software
- Oracle Instance
 - *SGA (System Global Area)*
 - *Oracle background processes*
 - *Shadow processes (user processes)*
 - *Shared Memory*
- Oracle Database (physical structure -data files)
- Oracle Database (logical structure - tablespaces)
- Client/Server architecture

Oracle instance architecture (cont.)

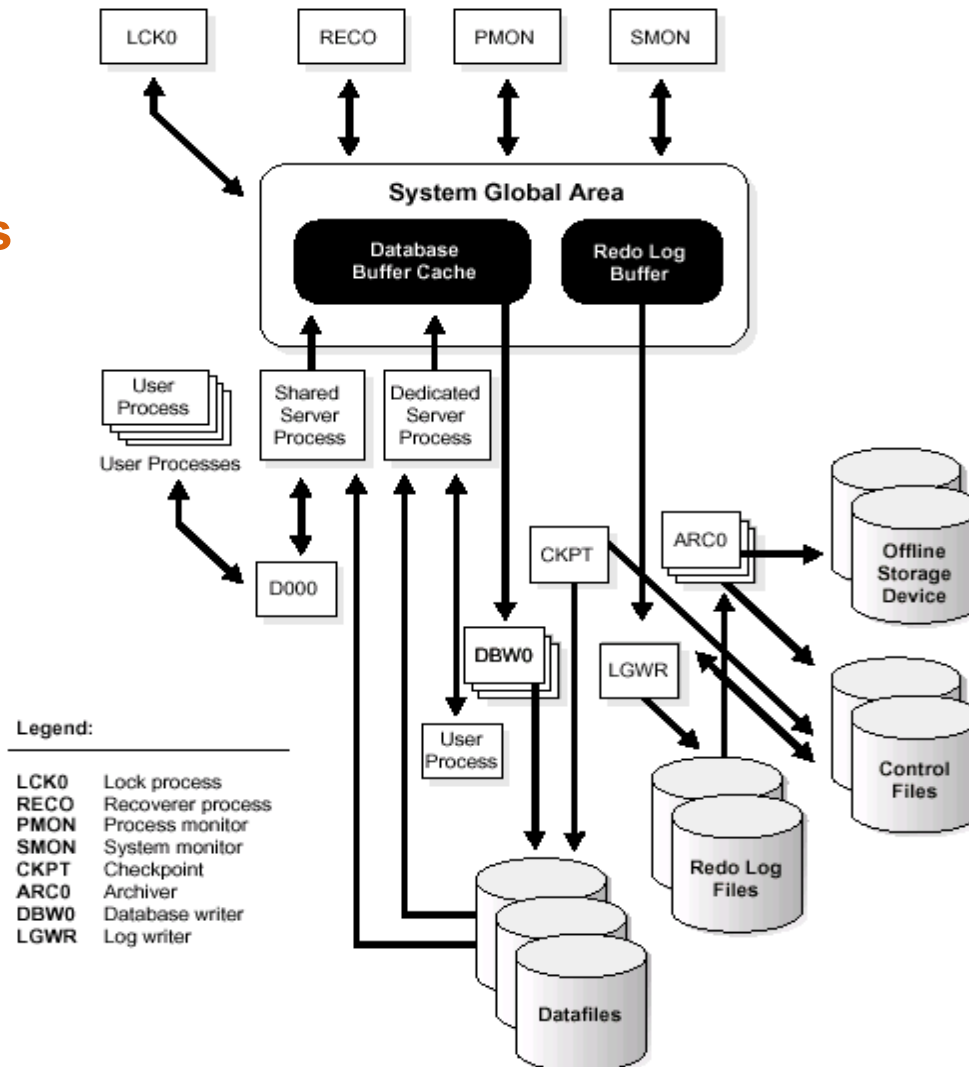
SGA (in Memory)

Total System Global Area	4435156 bytes
Fixed Size	47152 bytes
Variable Size	3970212 bytes
Database Buffers	409600 bytes
Redo Buffers	8192 bytes



Oracle instance architecture (cont.)

Oracle Processes

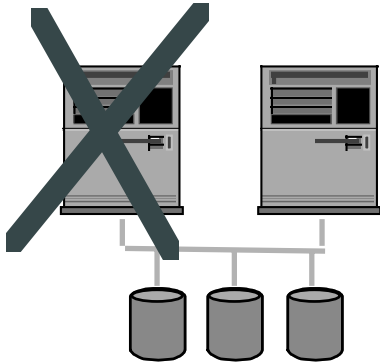


Oracle Parallel Server Overview

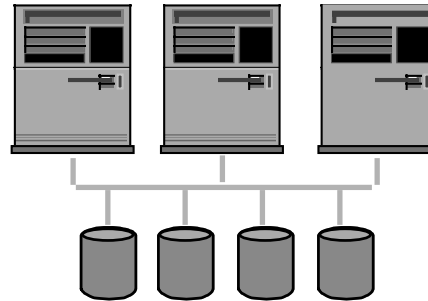


- Introduction to Oracle Parallel Server
- Differences between single Oracle and OPS
- Oracle Parallel Server Architecture
- 9i RAC on HP

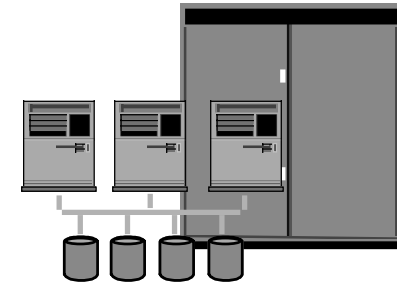
• Why Oracle Parallel Server?



**High
Availability**



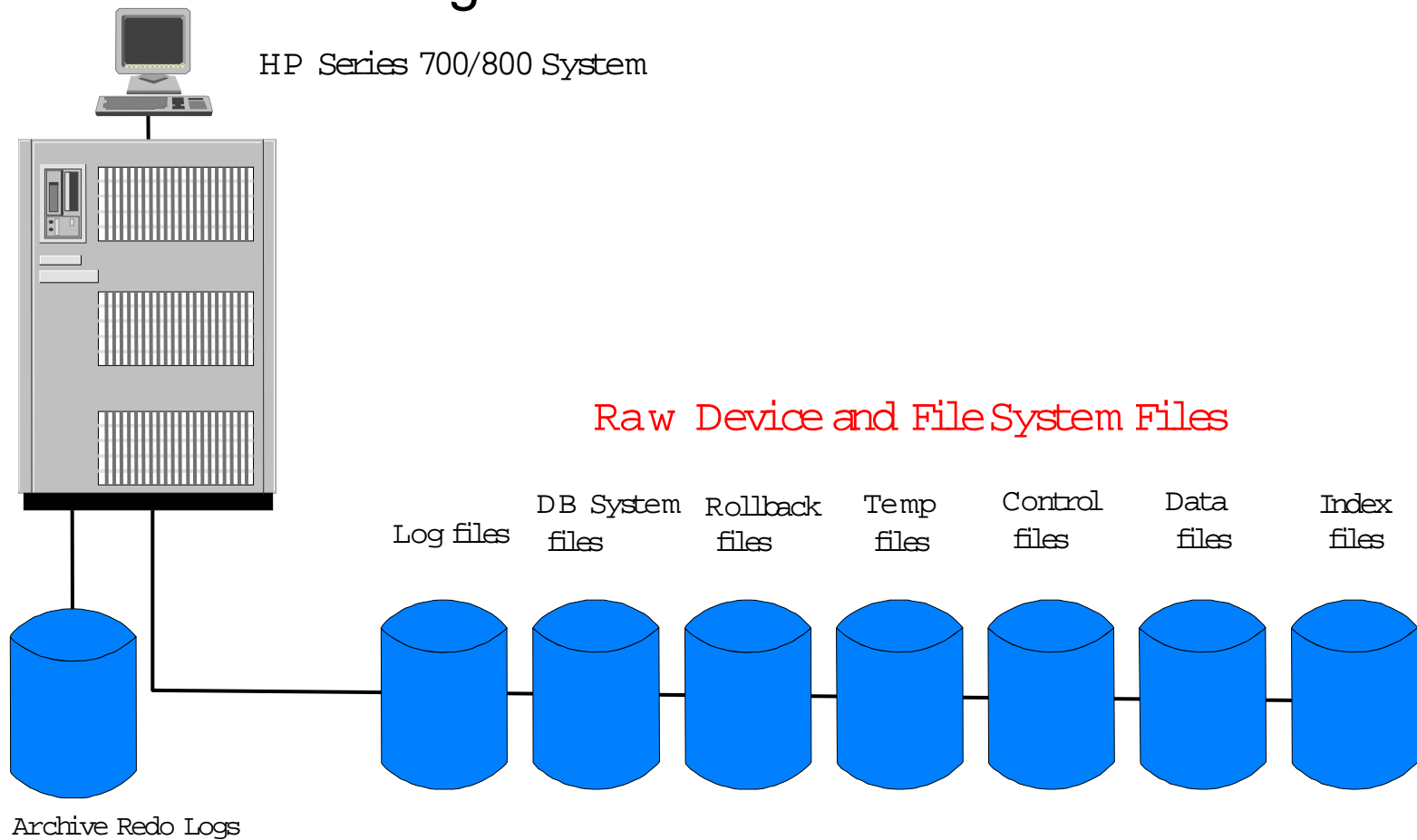
**Size of Databases
Performance
Throughput**



**Costs
(Consolidation of
database)**

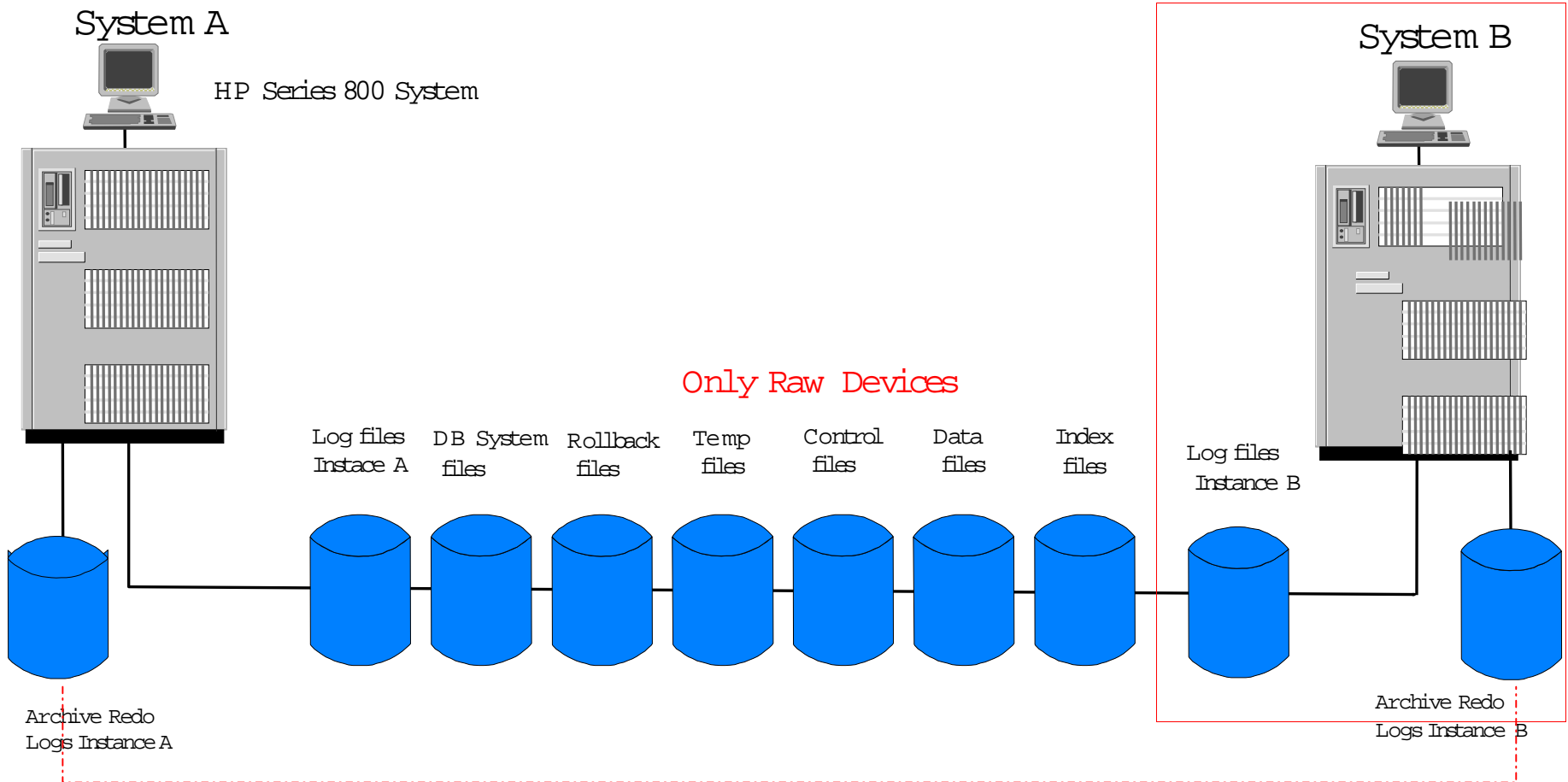
Differences between single Oracle and OPS

Single Oracle Database



Differences between single Oracle and OPS (cont.)

Oracle Parallel Server Database

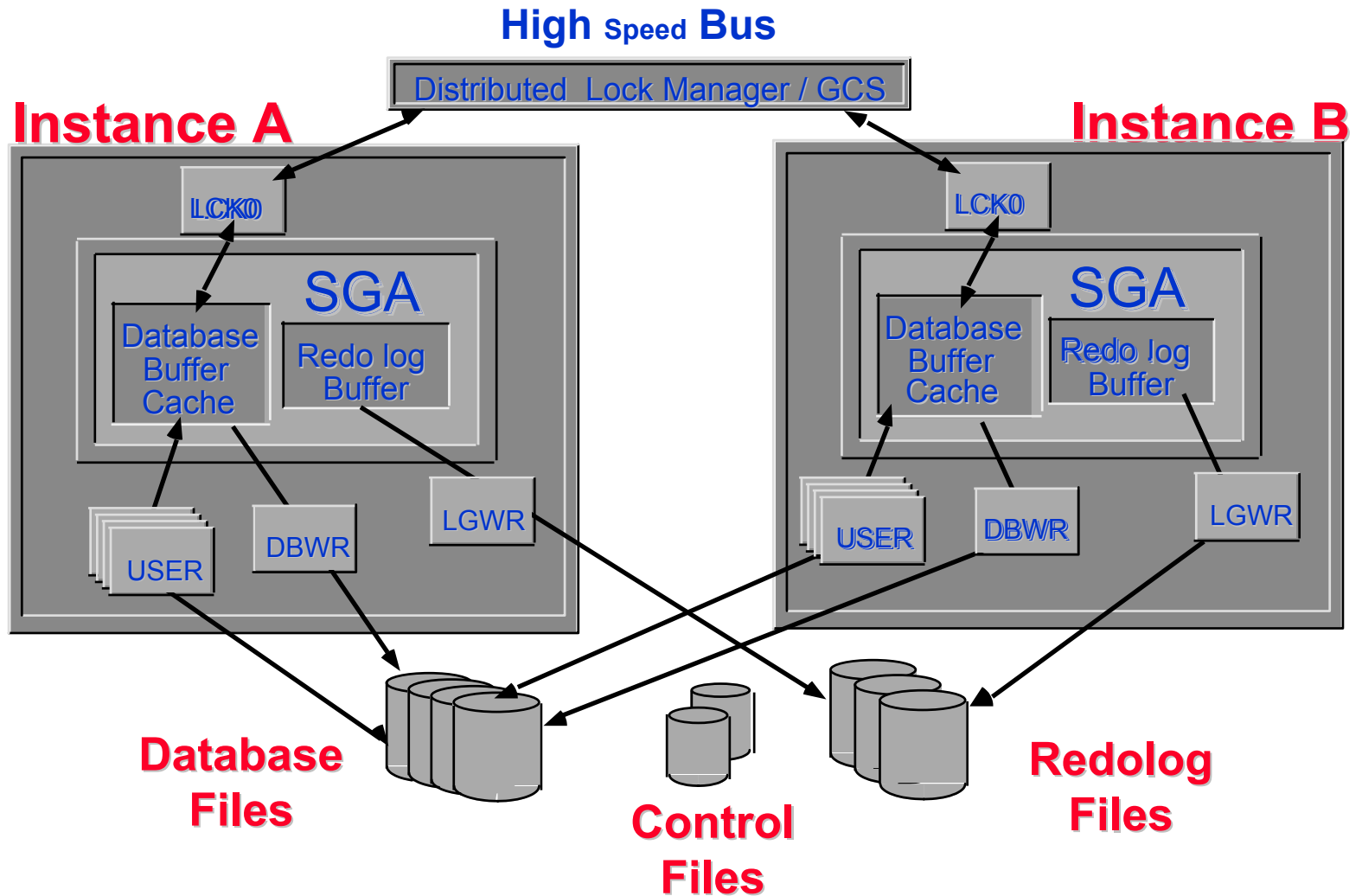


Differences between single Oracle and OPS (cont.)

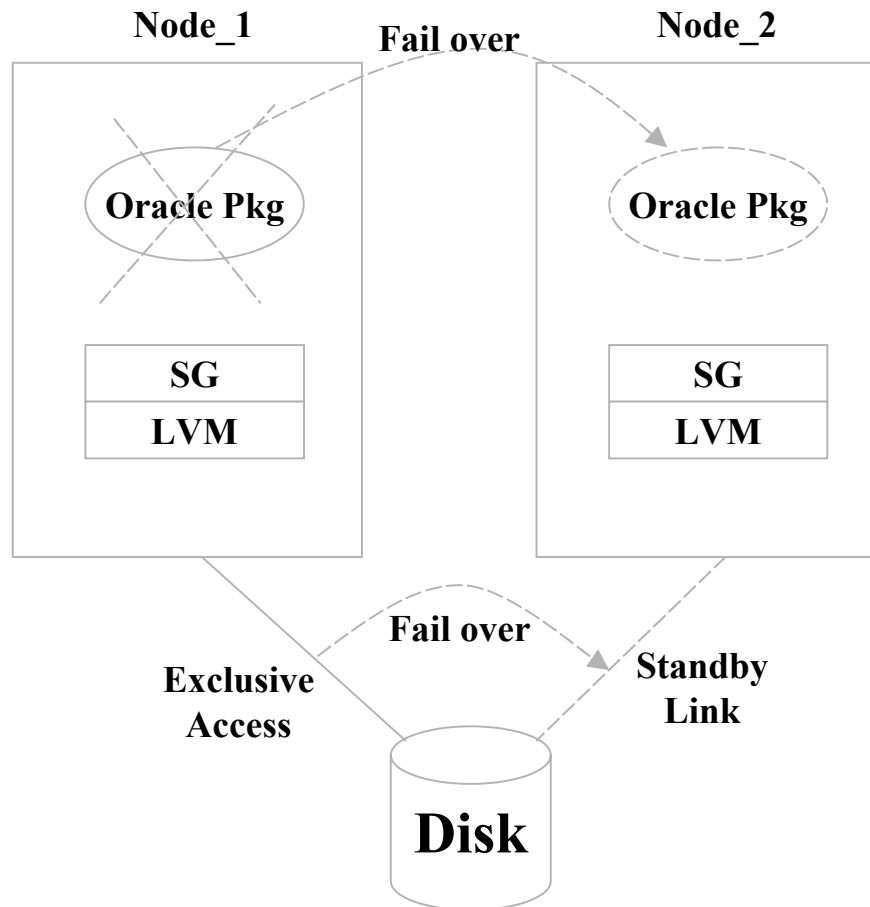
Summary:

- Each Instance has it's own Redo Log Files
- Each Instance has it's own Archive Redo Logs
- A Distributed Lock Manager ensures Data integrity
- An additional Process LCK0 requests Locks from DLM
- **IMPORTANT!** There is just ONE Control File for all OPS Instances!

OPS Architecture

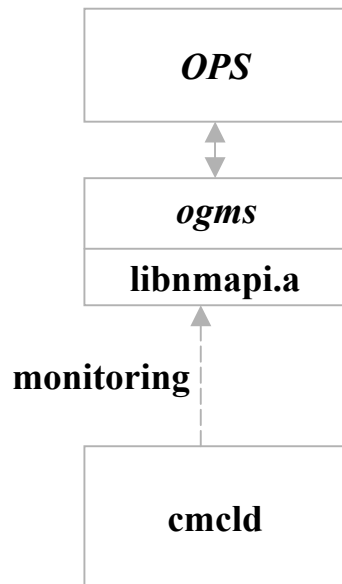


Oracle in SG cluster



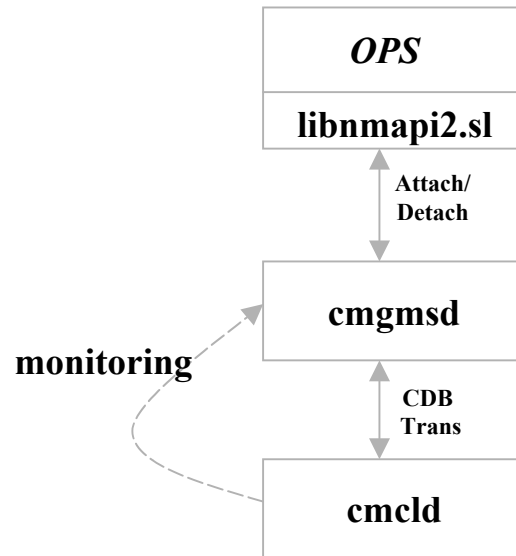
OPS 8.0.x Architecture on HP-UX 11.0 and beyond

(**obsolete 10/01**)



- *libnmapi.a* is linked to *ogms* (Oracle Group Membership Service).
- OPS gets cluster information via *ogms*.
- *ogms* is from Oracle but is started by *cmcld* when the node joins the cluster.
- *cmcld* starts up and shut down *ogms* as well as monitors the health of *ogms*.
- HP DLM is no longer used.

OPS 8.1.x / 9i Architecture



- libnmapi2 - dynamically linked with OPS process.
- cmgmsd - manager group membership.
- cmcld - startup/shutdown cmgmsd and monitor the health of cmgmsd

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- **ServiceGuard OPS Edition Concepts & Configuration**
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)

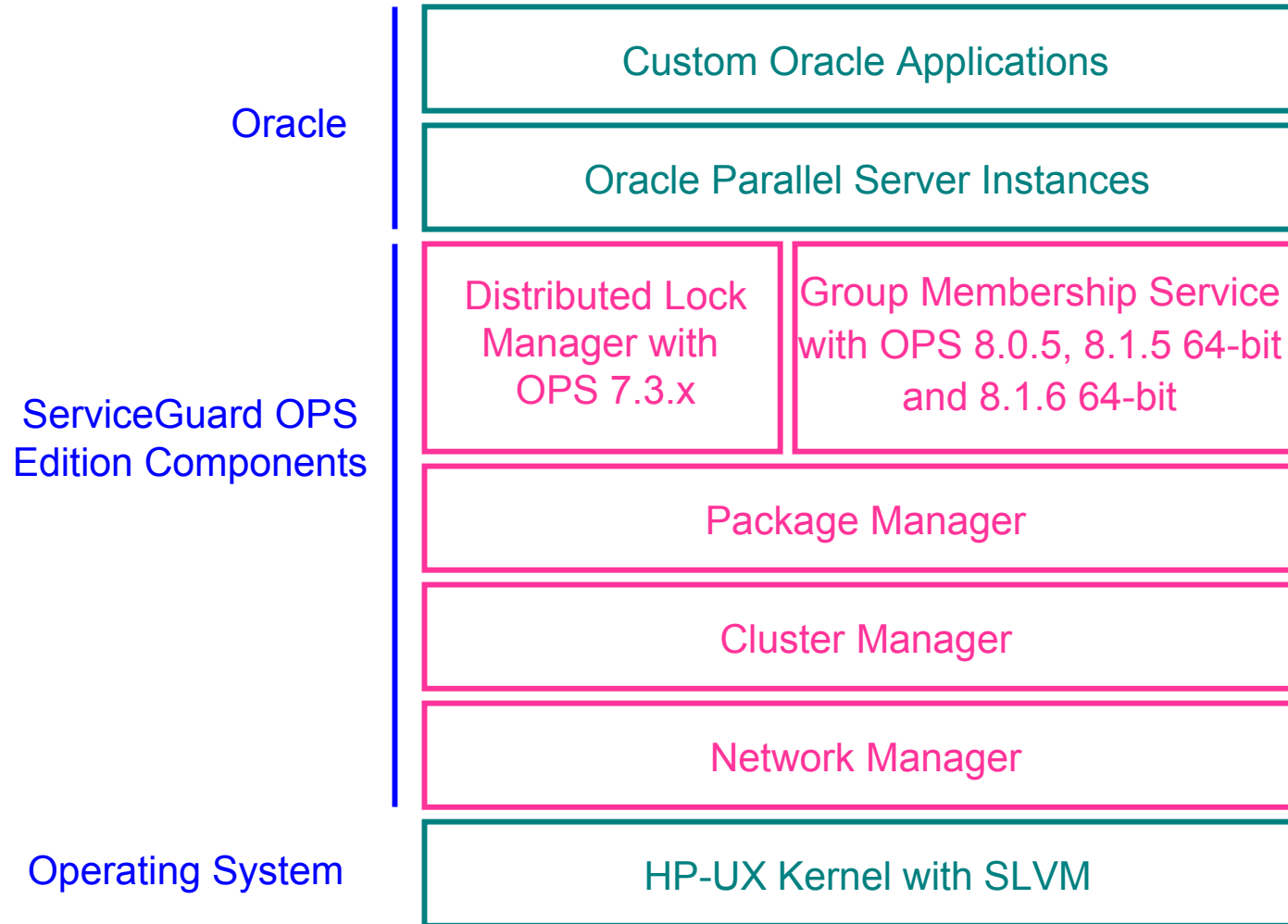


ServiceGuard OPS Edition Concepts & Configuration



- HP-UX administration know how
 - UNIX administration
 - Network basics
 - Strong LVM knowledge
- Cluster Administration knowledge
 - HP Cluster concepts
 - Installation and configuration
- Strong Oracle RDBMS administration know how (single instances)
- Oracle Parallel Server administration

SG OPS Software Components and Libraries

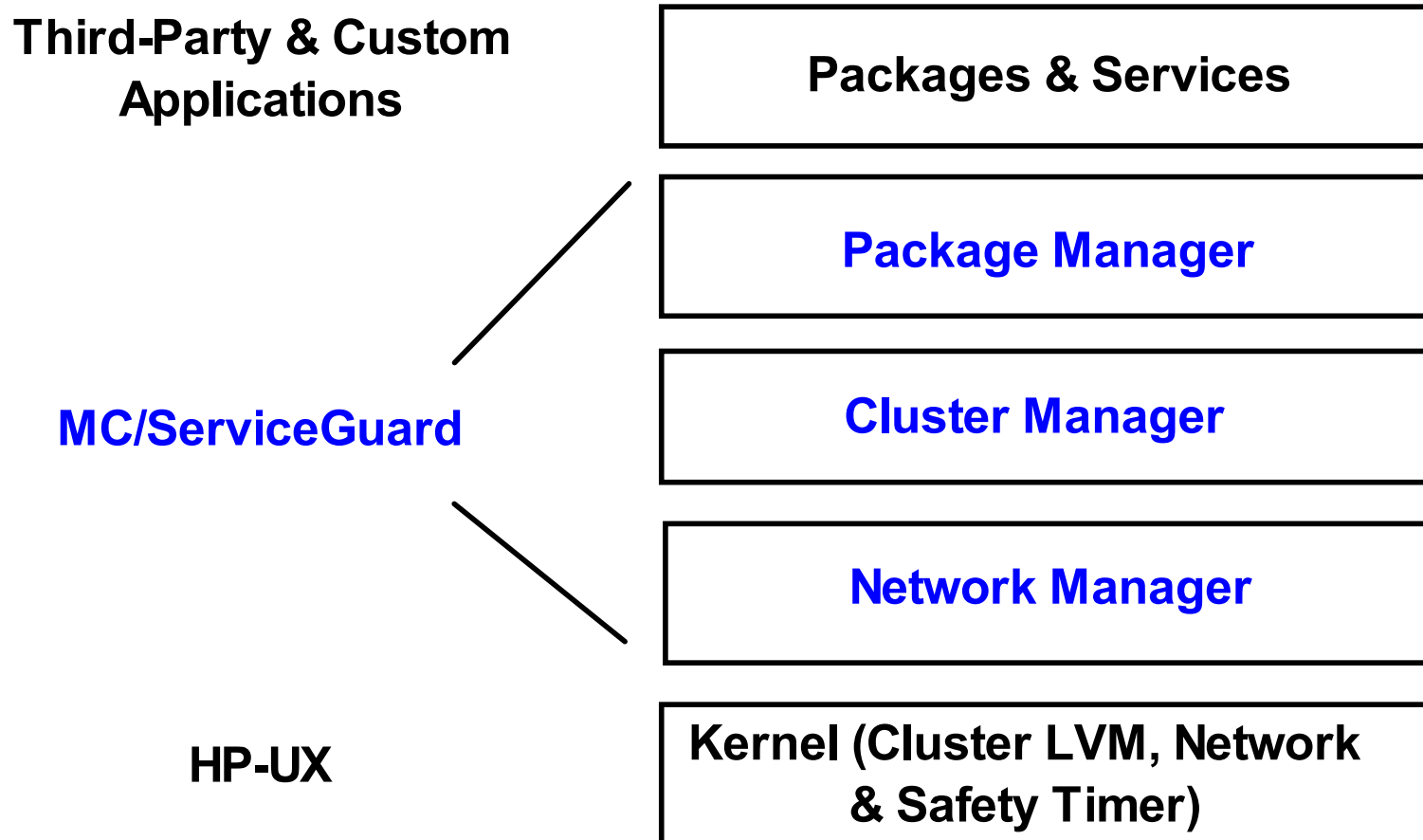


What is ServiceGuard OPS Edition?

- a.k.a. MC/LockManager.
- A product especially designed for Oracle Parallel Server (OPS).
- SG OPS Edition = SG + DLM + NMAPI Lib.
- cmcld shipped in SG OPS Edition shares the same code as the one in regular SG but some features are **disabled**.
- Shared Logical Volume Manager features are enabled with SG OPS

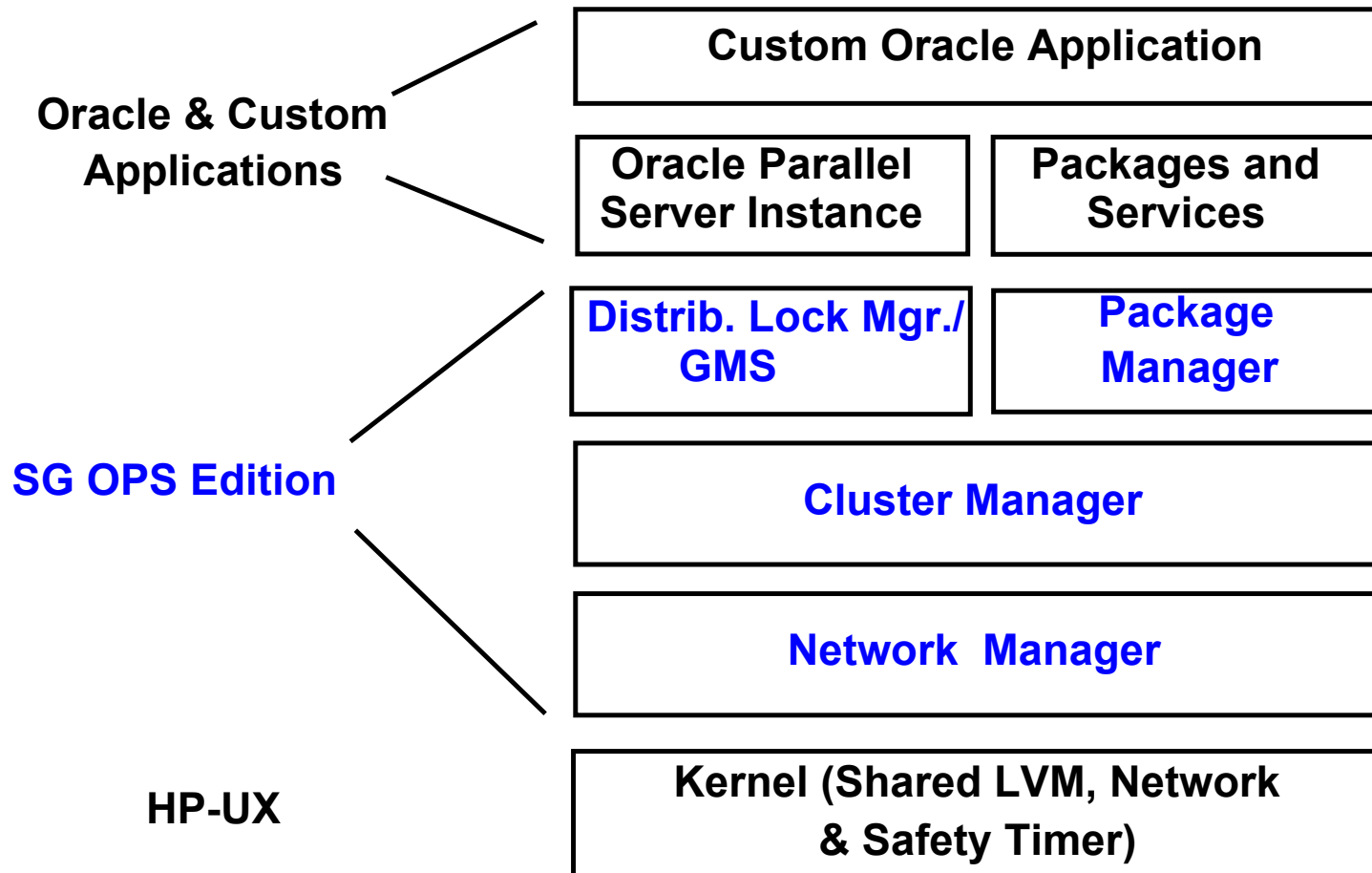
Introduction to SG OPS Edition (cont.)

• MC/ServiceGuard Components

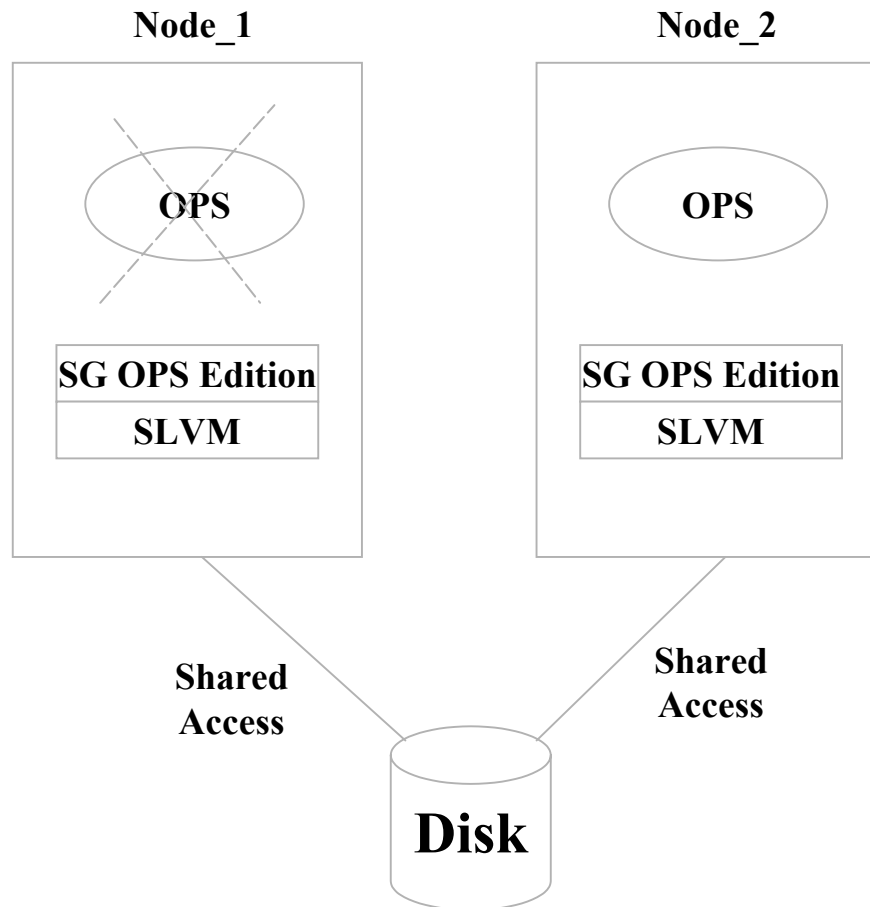


Introduction to SG OPS Edition (cont.)

SG OPS Edition Components

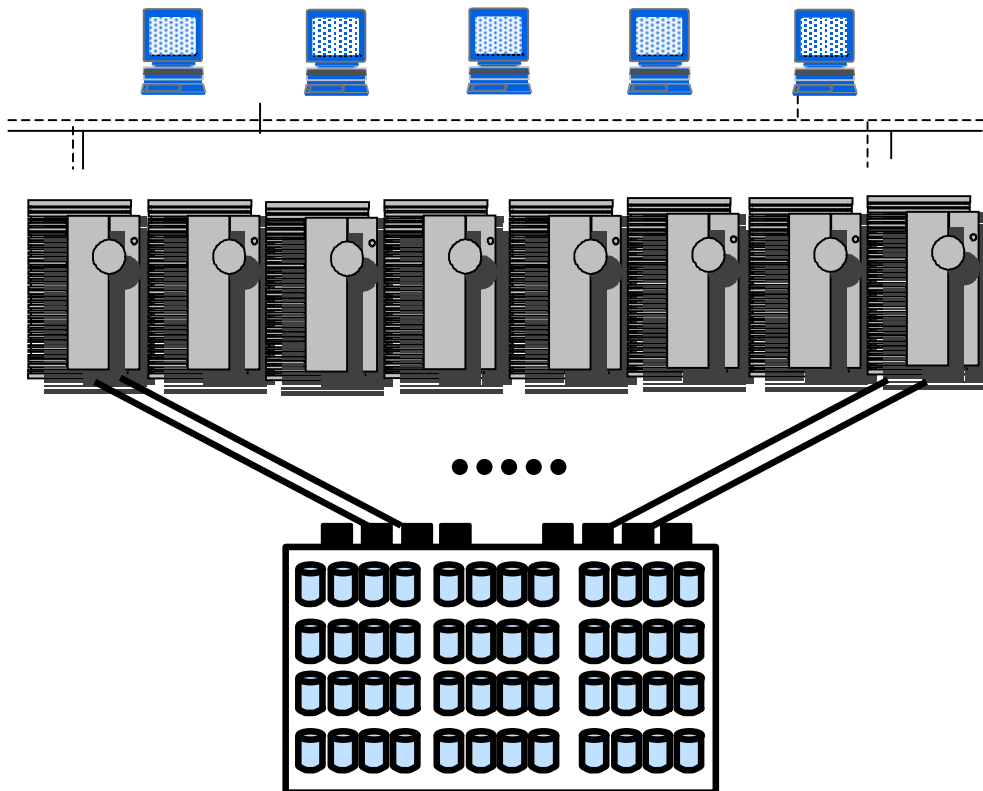


OPS in SG Cluster



ServiceGuard OPS Edition

HP-UX 11.0 and OPS 8.0.5+



- 11.0 Support
- DLM contained in Oracle kernel
- Oracle 8 support
- 64 bit support (OPS 8.0.5)
- Multiple Instance support (OPS 8.0.5)

Steps to Configure the Cluster (LM 11.0x)

1. Install ServiceGuard OPS Edition.
2. Configure each node's disk resources & cluster lock disk using SLVM.
3. Using SAM or the command line:
Edit the `/.rhosts` or `/etc/cmcluster/cmclnodelist` file to permit root access by all nodes in the cluster (SG OPS 10.07 and beyond).
4. Generate and edit a Cluster ASCII configuration file.
5. Verify the configuration file.
6. Generate and distribute a binary cluster configuration file to all nodes.
7. Enable Automatic startup and shutdown for the cluster and OPS by configuring the packages for the OPS instances.

Configuring SG OPS 11.x (cont.)



Start up Sequence with Lockmanager 11. x and Oracle Group Membership Service (ogms):

- 1) cmclds determine members of active cluster
- 2) cmcld starts gms deamon

Configuring SG OPS 11.x (cont.)



cluster.ascii file 11.0x

```
# *****
# ***** HIGH AVAILABILITY CLUSTER CONFIGURATION FILE *****
# ***** For complete details about cluster parameters and how to *****
# ***** set them, consult the cmquerycl(1m) manpage or your manual. *****
# *****
# Enter a name for this cluster. This name will be used to identify the
# cluster when viewing or manipulating it.
```

```
CLUSTER_NAME          adp_cluster
```

```
# Cluster Lock Device Parameters. This is the volume group that
# holds the cluster lock which is used to break a cluster formation
# tie. This volume group should not be used by any other cluster
# as cluster lock device.
```

```
FIRST_CLUSTER_LOCK_VG          /dev/vg_adp
```

```
# Definition of nodes in the cluster.
# Repeat node definitions as necessary for additional nodes.
```

```
NODE_NAME          adhp1
  NETWORK_INTERFACE lan0
    HEARTBEAT_IP          130.130.100.231
  NETWORK_INTERFACE lan3
  NETWORK_INTERFACE lan4
  NETWORK_INTERFACE lan5
  NETWORK_INTERFACE lan6
  FIRST_CLUSTER_LOCK_PV          /dev/dsk/c5t0d0
```


Configuring SG OPS 11.x (cont.)



cluster.ascii file 11.0x

```
# List of serial device file names
# For example:
# SERIAL_DEVICE_FILE           /dev/tty0p0

# Possible standby Network Interfaces for lan0: lan3,lan5,lan4,lan6.

NODE_NAME                adh2
NETWORK_INTERFACE lan0
HEARTBEAT_IP              130.130.100.232
NETWORK_INTERFACE lan3
NETWORK_INTERFACE lan4
NETWORK_INTERFACE lan5
NETWORK_INTERFACE lan6
FIRST_CLUSTER_LOCK_PV     /dev/dsk/c3t0d0
# List of serial device file names
# For example:
# SERIAL_DEVICE_FILE           /dev/tty0p0

# Possible standby Network Interfaces for lan0: lan4,lan3,lan6,lan5.

# Cluster Timing Parmeters (microseconds).

HEARTBEAT_INTERVAL        1000000
NODE_TIMEOUT               2000000
```

Configuring SG OPS 11.x (cont.)



cluster.ascii file 11.0x

```
# Configuration/Reconfiguration Timing Parameters (microseconds).

AUTO_START_TIMEOUT  600000000
NETWORK_POLLING_INTERVAL      2000000

# Package Configuration Parameters.
# Enter the maximum number of packages which will be configured in the cluster.
# You can not add packages beyond this limit.
# This parameter is required.
MAX_CONFIGURED_PACKAGES      6

# List of cluster aware Volume Groups. These volume groups will
# be used by package applications via the vgchange -a e command.
# For example:
# VOLUME_GROUP              /dev/vgdatabase.
# VOLUME_GROUP              /dev/vg02.

VOLUME_GROUP                /dev/vg_moi

# List of OPS Volume Groups.
# Formerly known as DLM Volume Groups, these volume groups
# will be used by OPS cluster applications via
# the vgchange -a s command. (Note: the name DLM_VOLUME_GROUP
# is also still supported for compatibility with earlier versions.)
# For example:
# OPS_VOLUME_GROUP          /dev/vgdatabase.
# OPS_VOLUME_GROUP          /dev/vg02.

OPS_VOLUME_GROUP            /dev/vg_adp
```

Configuring SG OPS 11.x (cont.)



cluster.ascii file 11.0x

```
# DLM parameters.  
# When using Oracle Parallel Server versions prior to 8.0, set  
# DLM_ENABLED to YES, enter values for the other parameters as  
# specified in your Oracle documentation, and set GMS_ENABLED to NO below.
```

```
DLM_ENABLED                YES  
DLM_CONNECT_TIMEOUT        30000000  
DLM_PING_INTERVAL          20000000  
DLM_PING_TIMEOUT           60000000  
DLM_RECONFIG_TIMEOUT       300000000  
DLM_COMMFAIL_TIMEOUT       270000000  
DLM_HALT_TIMEOUT           240000000
```

```
# GMS parameters.  
# When using Oracle Parallel Server version 8.0.5 (64-bit), set  
# DLM_ENABLED to NO above, set GMS_ENABLED to YES below, and  
# enter values for the other parameters as specified in your  
# Oracle documentation.  
# When using 8.1.5 or later, set DLM_ENABLED and GMS_enabled to NO!
```

```
GMS_ENABLED                NO  
GMS_CONNECT_TIMEOUT        30000000  
GMS_LOCATION
```

Configuring SG OPS 11.x (cont.)



DLM_Enabled YES or NO

** Start DLM up when cluster comes up*

DLM_Reconfig_Timeout (microseconds)

** How long to wait for cmdlmd and cmdlmmond to reconfigure*

DLM_Ping_Interval (microseconds)

** How often does cmcld ping cmdlmmond*

DLM_Ping_Timeout (microseconds)

** How long without a ping before we give up*

DLM_Halt_Timeout (microseconds)

** How long to wait for the DLM halt script to complete*

DLM_Connect_Timeout (microseconds)

** How long to wait for the DLM run script to complete*

DLM_Commfail_Timeout (microseconds)

** How long to wait for a reconfiguration to complete*

Note: These parameters are obsolete with SG OPS 11.13!

Configuring SG OPS 11.x (cont.)



New Parameters with MC/LM 11.0x and OPS 8.0.5+

GMS_Enabled YES or NO

** Start GMS up when cluster comes up*

GMS_Connect_Timeout 30 sec

** How long to wait for GMS to come up*

GMS_LOCATION

** Directory where the ogms binary is located (Typically \$ORACLE_HOME/bin)*

Configuring SG OPS 11.x (cont.)



Security entries (LM 11. x)

- Starting with 8.0.5 you need to add an entry to the **cmclnodelist** file that gives each node in the cluster oracle user name security access for each configured Oracle user.

The format is:

first node name	Oracle user name
second node name	Oracle user name

Configuring SG OPS 11.x (cont.)



Parameters Setting for different OPS

	DLM_ENABLED	GMS_ENABLED
OPS 7.x	Yes	No
OPS 8.0.x	No	Yes
OPS 8. and 9i RAC	No	No
OPS 8 Ref Imp (32-bit 8.1.5 OPS only)	Yes	No

Configuring SG OPS 11.x (cont.)



- **Automatic Startup with OPS 8.x + Native**

- OPS Instances is started with a SG OPS Package
- Package should be non-failover
- Just the node on which the instance is running should be configured to run the OPS package
- Set the PACKAGE_SWITCHING_ENABLED parameter to YES
- NODE_FAILFAST_ENABLED should also be set to NO.

Configuring SG OPS 11.x (cont.)



OPS Instance package example:

- The OPS instance package configuration consists of usually 3 files
 - * OPS_instance1.conf Package configuration file
 - * OPS_instance1.cntl Package control file
 - * oracle.sh Enterprise Cluster Master Toolkit file for Oracle
- The package is allowed to switch but will not, since it has just one node configured

Configuring SG OPS 11.x (cont.)



oracle.sh from Cluster Master Toolkit (Declaration part)

```
ORA_7_3_X=no
ORA_8_0_X=yes
SID_NAME=OPS1
ORACLE_HOME=/oracle/product/8.1.6
SQLNET=yes
NET8=yes
LISTENER_NAME=LISTENER
LISTENER_PASS=
MONITOR_INTERVAL=30
PACKAGE_NAME=OPS1pgk
TIME_OUT=10
set -A MONITOR_PROCESSES ora_smon_${SID_NAME} ora_pmon_${SID_NAME}\
ora_lgwr_${SID_NAME} ora_dbw0_${SID_NAME} ora_lmd0_${SID_NAME} ora_lmon_${SID_NAME}
HOST=`hostname`
DATE=`date`
PATH=${ORACLE_HOME}/bin:/sbin:/usr/bin:/usr/sbin:/etc:/bin:export
ORACLE_SID=${SID_NAME}
export ORACLE_HOME
```

Upgrading an SG OPS Cluster



- Upgrading a Lockmanager Environment has to be planned very carefully, because the OPS has to be brought down! (this is true for LM 10.x and 11.x clusters)

Upgrading an SG OPS Cluster (cont.)



1. Switch Packages from node A to node B and Stop the OPS Instance.
2. Deactivate OPS VG's from shared mode.
3. Stop cluster on node A.
4. Upgrade the OS.
5. Upgrade ServiceGuard OPS Edition.
6. Relink Oracle.
7. Reconfigure the cluster to a one-node cluster, but don't start it yet including the packages !!! (Use a new cluster name!)
8. Now you have to stop the packages and OPS on node B.
9. Unshare the VG's, remove the cluster id and shutdown the entire cluster.

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- **Group Membership Service**
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Cluster Membership is...

- A cluster can be defined as several loosely-coupled systems grouped together by monitoring each other's health via some heart beat mechanism.
- Cluster membership represents the health of each member in the cluster.
- High-availability is achieved by failing over applications (defined as package) running on the down node to a healthy node within the cluster in the event of node crash.

Group Membership Service (cont.)

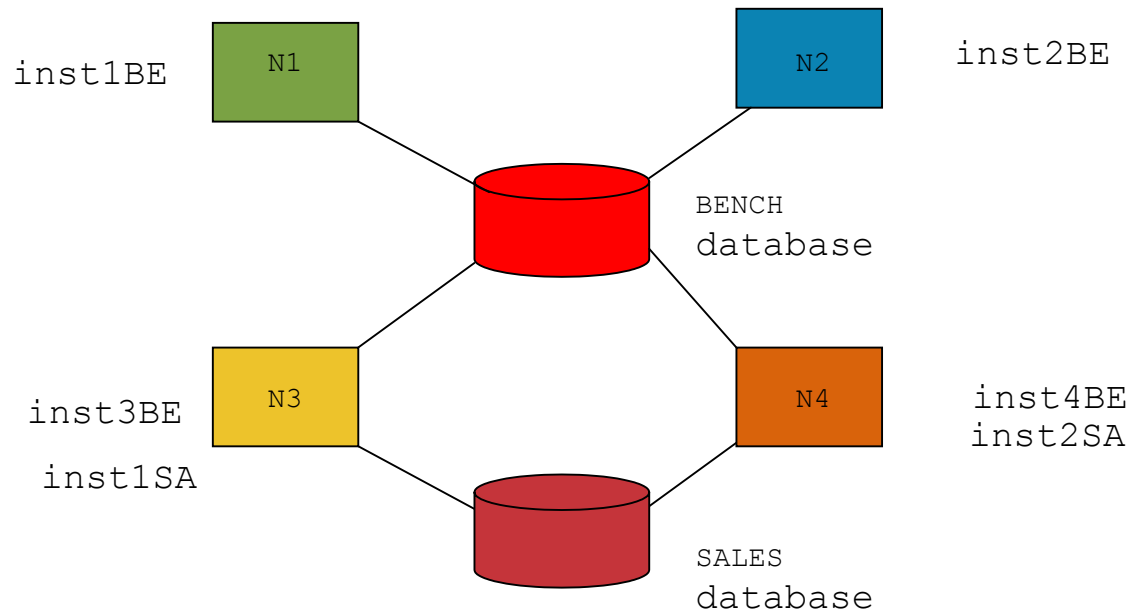


Group Membership Service is...

- A Process Monitoring facility built on top of ServiceGuard.
- Process group is a cluster-wide entity identified by a unique name. Every group member is assigned a unique id.
- Any application that is interested in knowing the status of other instances running in the cluster can participate in Group Membership Service.
- Oracle Parallel Server 8i is currently using this service.

Group Membership Service (cont.)

Example of an SG-OPS cluster with GMS



Group Membership Service (cont.)



Example of an SG-OPS cluster with GMS

- Use `cmviewcl -v -l group`
to see the
groups

Group Membership Service (cont.)



How Group Membership Service Works

- When a process requests to join a group, an unique id is assigned to this new member and other members in the group are notified by receiving a new group membership status. (Incarnation is associated with each bitmap.)
- Once a process becomes a member of the group, it will be informed any change in the group membership.
- When a process no longer wants to participate in the group, it can request to detach from the group and as a result, other members in the group are notified by this event.

Group Membership Service (cont.)

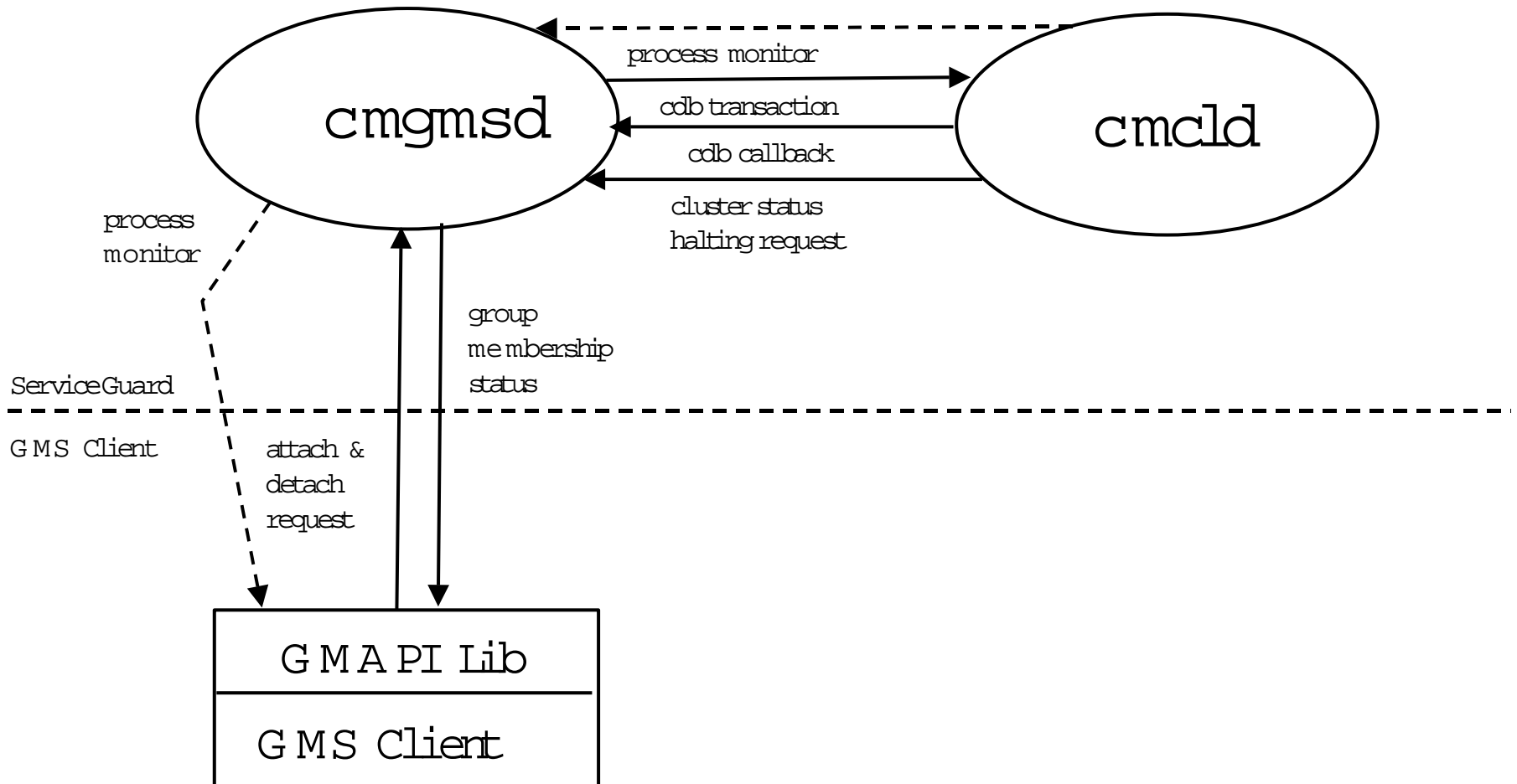


Events Triggering Group Membership Change

- A process joins the group.
- A process leaves the group.
- A group member died. Group Membership Service monitors group members and automatically de-registers group member if a member died.
- Cluster membership change due to node crash. Group Membership Service assumes that all group members running on the down node are gone.

Group Membership Service (cont.)

Architectural Overview



Group Membership Service (cont.)



Role of cmgmsd

- Add/remove group members.
- Deliver group membership status to group members.
- Monitor local group members. Automatically remove group member when the process goes away.
- Listen to cluster membership status from cmcld. Perform recovery to de-register group members running on the failed node.

Group Membership Service (cont.)



Role of cmcld

- Start-up/shutdown cmgmsd.
- Monitor cmgmsd. If cmgmsd goes away, cmcld toc's the machine.
- Provide cluster membership status to cmgmsd.
- Provide CDB as a cluster-wide data storage for distributing group membership. Adding/removing a group member is accomplished by performing a CDB transaction.

Group Membership Service (cont.)

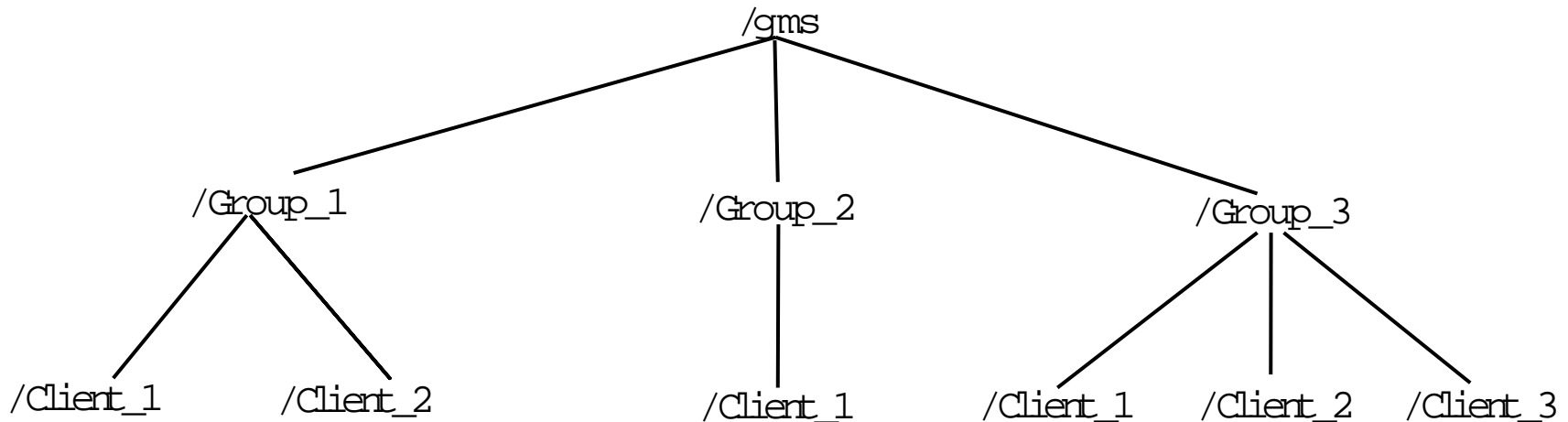


Role of GMAPI

- Provide a thin programming interface for client program to receive Group Membership Service.
- Handle the communication between client program and cmgmsd.
- The library is linked into the client process.

Group Membership Service (cont.)

CDB Tree



“\gms” is the root of the tree.

The picture represents three groups.

“group_1” has two members (client_1 and client_2).

“group_2” has one member (client_1).

“group_3” has three members (client_1, client_2 and client_3).

Group Membership Service (cont.)



Member Registration/De-registration

- Client program sends a request to cmgmsd and waits for a reply.
- cmgmsd issues a CDB transaction to add/remove the CDB object that represents the member. For adding a new member, if this is the first member joining the group, a group object will be added. For removing a member, if this is the last member in the group, the group object is also removed.
- cmgmsd sends a reply back to the client program based on the status of the transaction.

Group Membership Service (cont.)



Delivering Group Membership Status

- Several events (i.e. cluster membership change, member crash and member joining/leaving) lead to the change in group membership. cmgmsd submits a CDB transaction to change the group membership information.
- As the result of the transaction, a CDB callback is delivered to cmgmsd running on each node. Based on the information provided by the CDB callback, each cmgmsd sends out the new group membership status to its local group members. The group incarnation is increased by 1 every time there is a change in the group membership status.
- Group members periodically check for incoming new group membership status.

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- **Volume Groups and async I/O**
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Volume Groups and async I/O

1. Shared VGs are needed for

- * All database data files
 - * All log files
 - * Control files
- } Owned by oracle user

2. Work with Oracle consultant or DBA to determine size and layout of disk and LVs

3. Use striping to distribute workload across many disks.

Volume Groups and async I/O (cont.)



- For OPS it is highly recommended to use async I/O
- Configuring async I/O (Oracle 7.x.x, 8.x.x):
 1. include asyncdsk driver to unix kernel
 2. mknod /dev/async c 101 0x0
 3. chown oracle:dba /dev/async
 4. chmod 660 /dev/async
 5. Set use_async_io in the init.ora file to TRUE (this step is only for Oracle 7.x.x), with Oracle 8 async I/O is true by default.

Note: If you use more than one db_writer, async I/O will not be used! (only for 7.3.x)

Volume Groups and async I/O (cont.)

- After introduction of patch PHKL_22744 (already superseded by patch **PHKL_25906**)
- New async security features are enabled
- Oracle dies on startup with PMON failure
- Create file /etc/privgroup and add the following line:
dba MLOCK

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- **Unix Environment**
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



OFA - Oracle Flexible Architecture

- Directory structure
 - multiple Oracle Homes (e.g. different Releases)
- \$ORACLE_BASE
 - /u01/app/oracle is the default ORACLE_BASE directory
- \$ORACLE_HOME
 - /product/<release>
is the name of the Oracle home by default,
where <release> is e.g. 8.1.6 or 7.3.4

UNIX environment (cont.)

- Unix User: oracle
- Unix group: dba
- SW user: oinstall

Asynchronous I/O

- The asynchronous I/O pseudo-driver on HP-UX allows the Oracle Server to perform I/O to raw disk partitions using an asynchronous method, resulting in less I/O overhead and higher throughput.
- To enable the asynchronous I/O driver for Oracle set the `initsid.ora` parameter **disk_async_io** to **TRUE**.

UNIX environment (cont.)

UNIX Kernel Parameters

dbc_max_pct	2
dbc_min_pct	1
maxdsiz	0X07B000000
maxdsiz_64bit	0X00000000400000000
maxfiles	2048
maxfiles_lim	2048
maxssiz	0X017F00000
maxssiz_64bit	0X040000000
maxswapchunks	8192
maxtsiz	0X040000000
maxtsiz_64bit	0X00000000400000000
maxuprc	8192
maxusers	2100
msgmap	1463
msgmax	32768

UNIX environment (cont.)

UNIX Kernel Parameters

msgmnb	65535
msgmni	1461
msgssz	128
msgtql	1461
npty	6000
nstrpty	60
nstrtel	400
num_tachyon_adapters	5
semmap	1024
semmni	2044
semmns	2048
semmnu	2048
semume	1024
semvmx	65534
shmmax	0X200000000
shmmni	100
shmseg	10

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- **Oracle 9i RAC architecture**
- 7 Steps to Configure the SG OPS Cluster
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)





cluster solutions
(MC/SG & Oracle9i RAC)

Cold Failover with MC/ServiceGuard

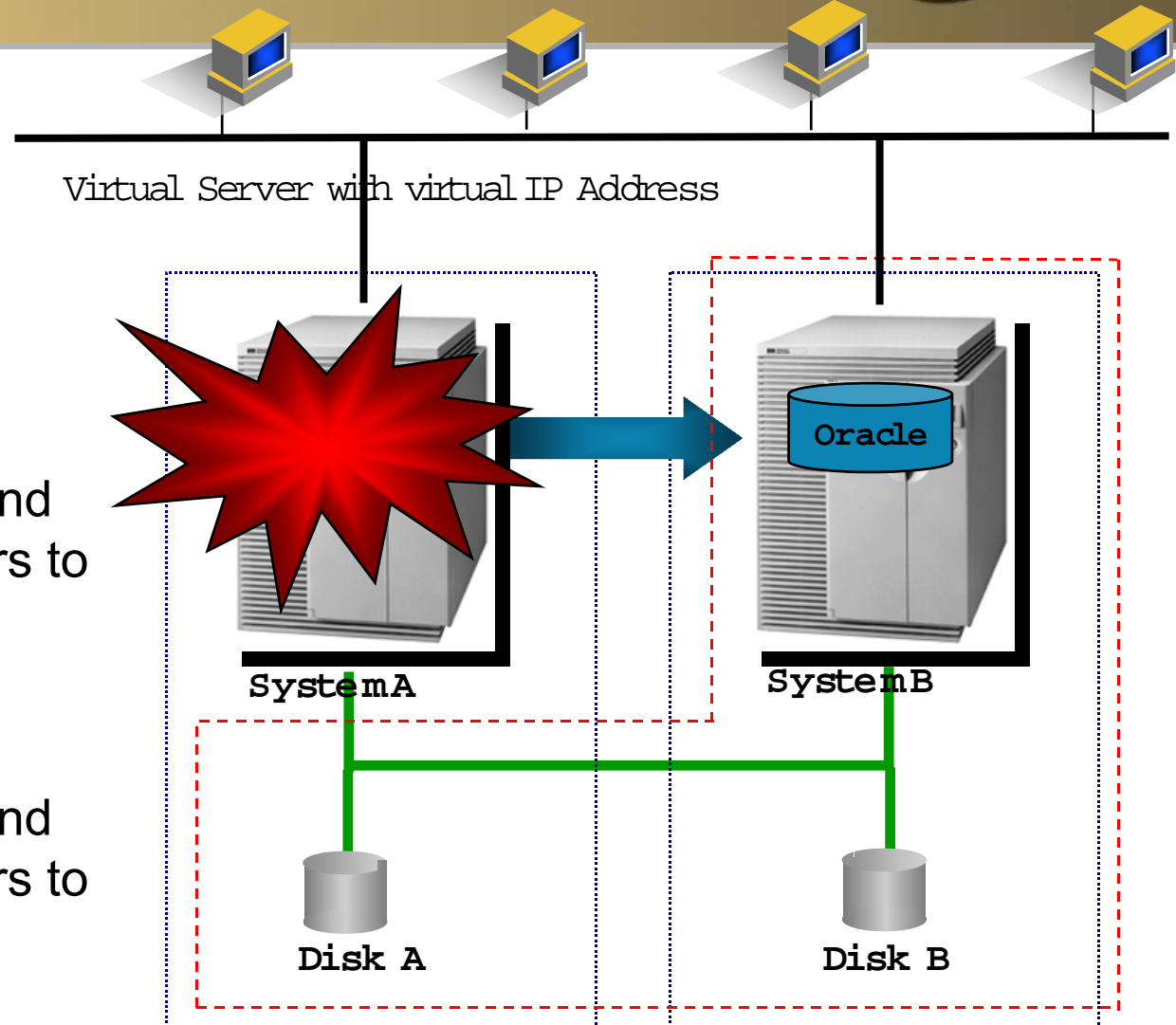
- Non-Shared Database
- Provides 16 node failover solution

Before Failover:

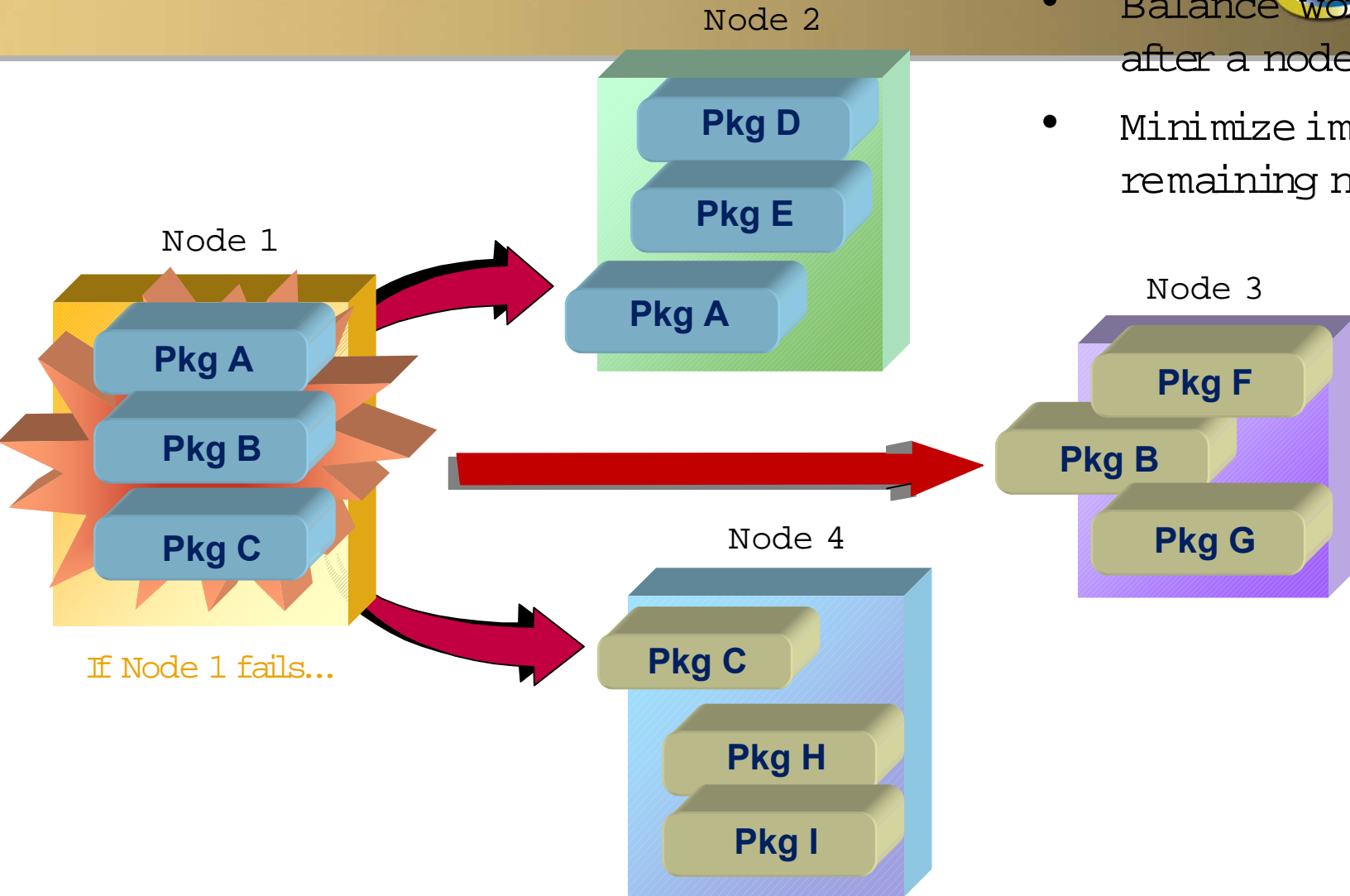
- Virtual IP Address and Networkname refers to System A

After Failover:

- Virtual IP Address and Networkname refers to System B

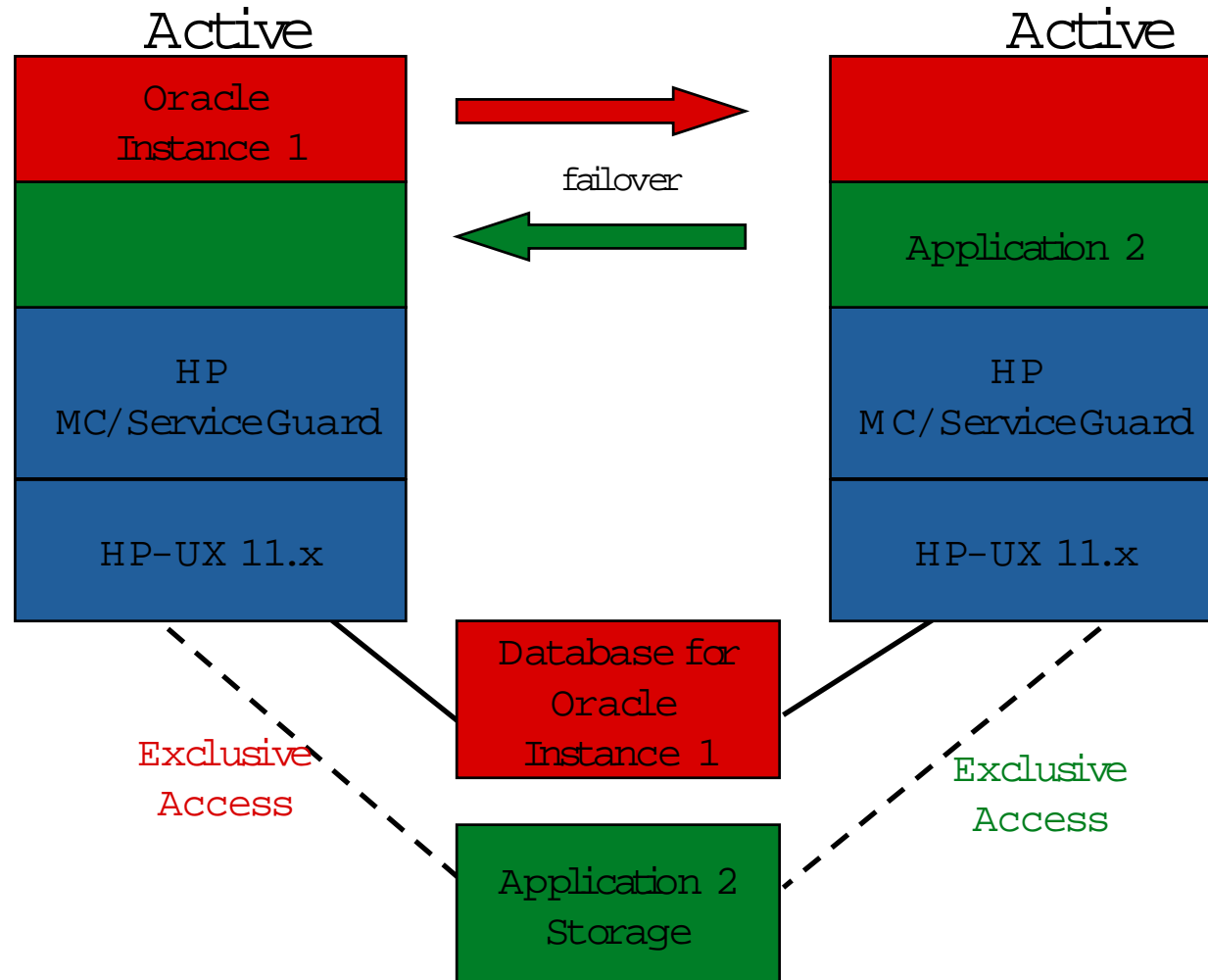


Cold Failover with MC/ServiceGuard



- Balance workload after a node failure
- Minimize impact on remaining nodes

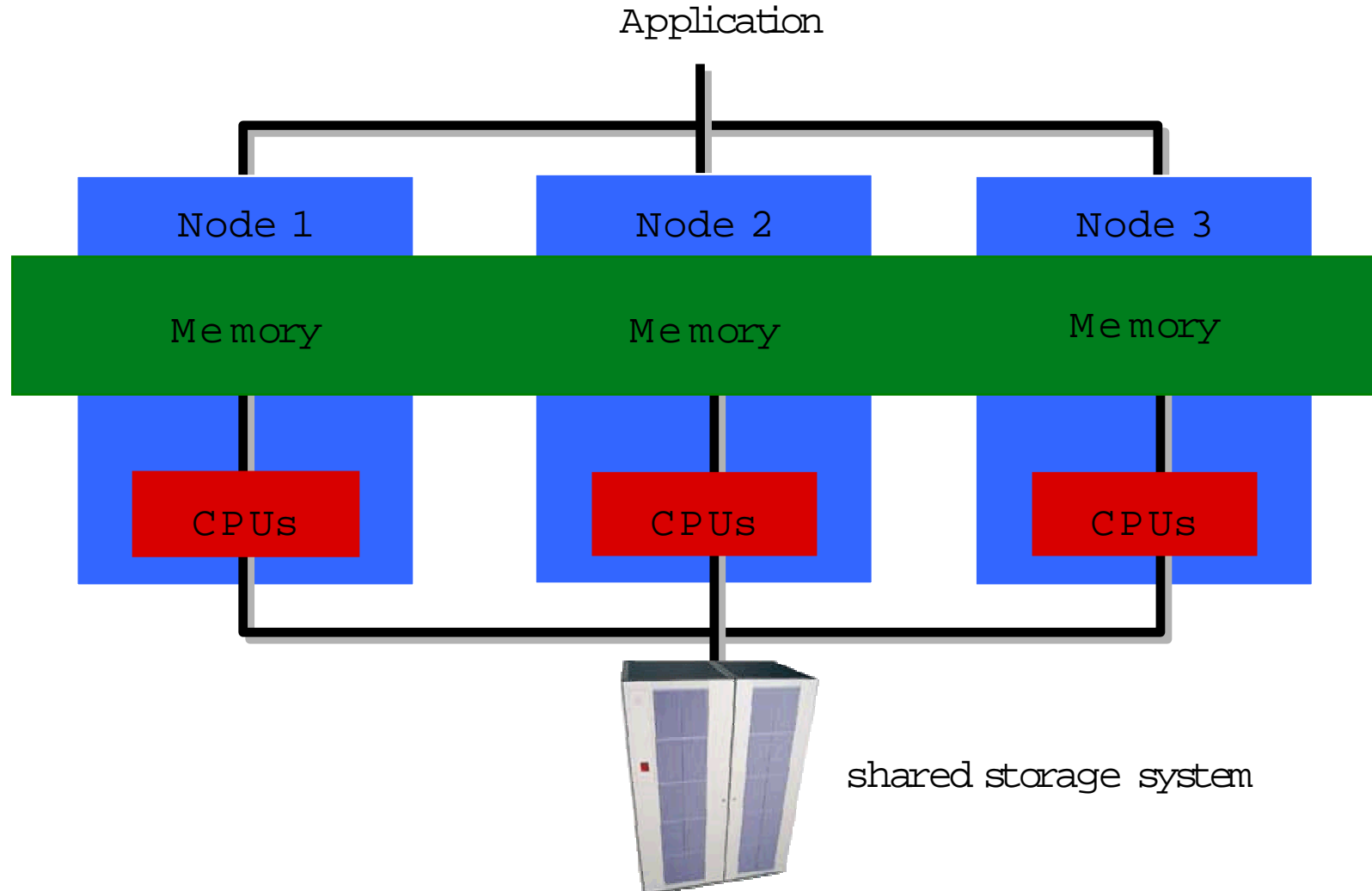
MC/ServiceGuard Software Stack



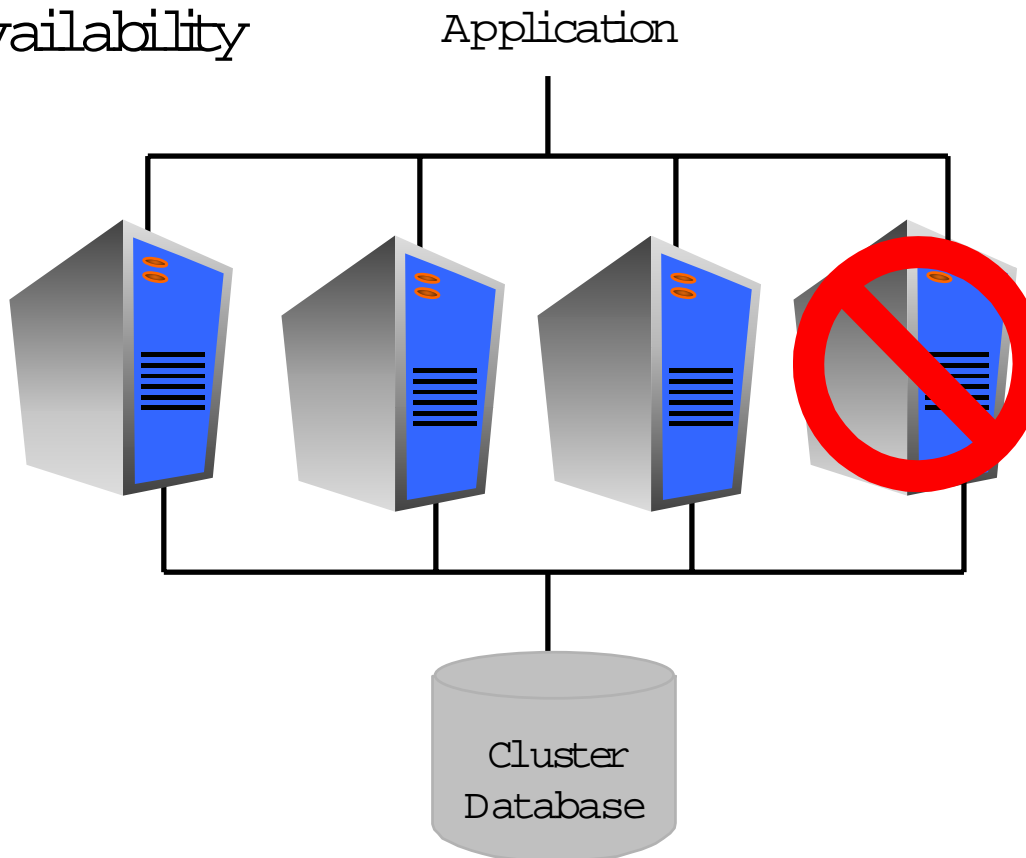
Limits of Cold Failover

- Scalability of cluster is limited to scalability of one server
- Cold failover is slow, as many time consuming tasks must be performed as part of failover
 - moving and mounting logical volumes
 - starting the oracle instance
 - opening the data files
- After failover, all sessions must re-establish connections to the database
- After failover, the instance caches are cold introducing a performance brownout

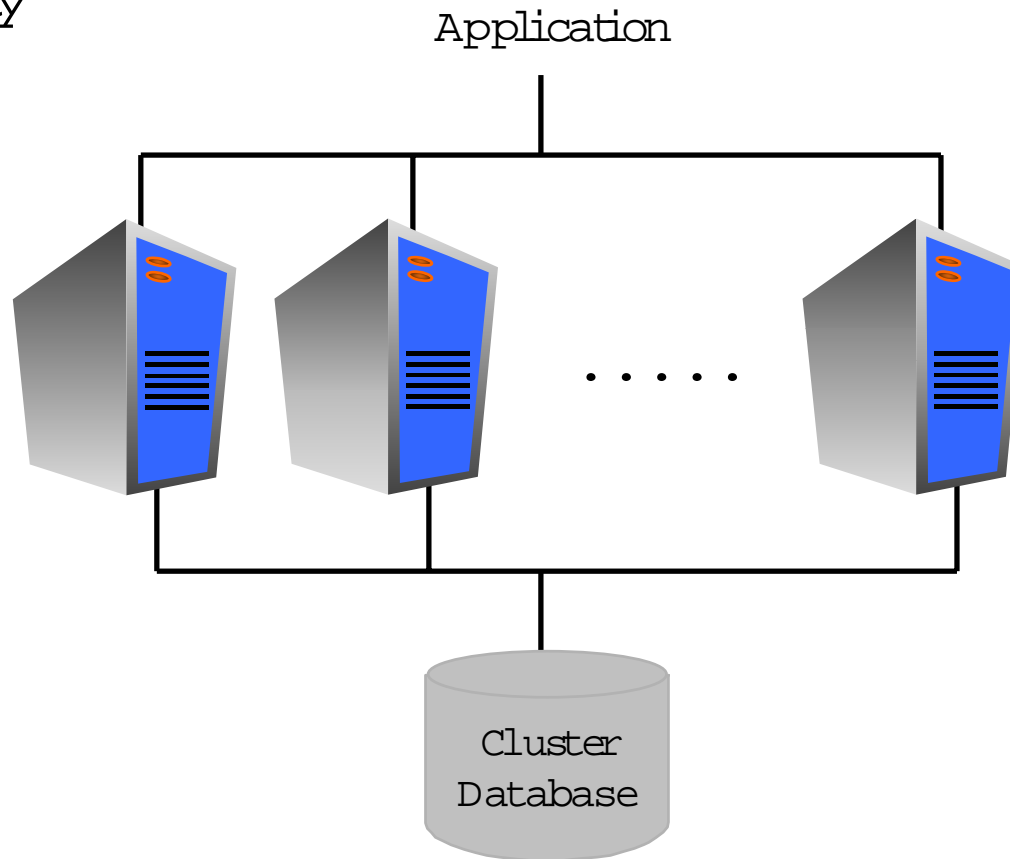
Oracle9i Real Application Clusters



a) High Availability



b) Scalability

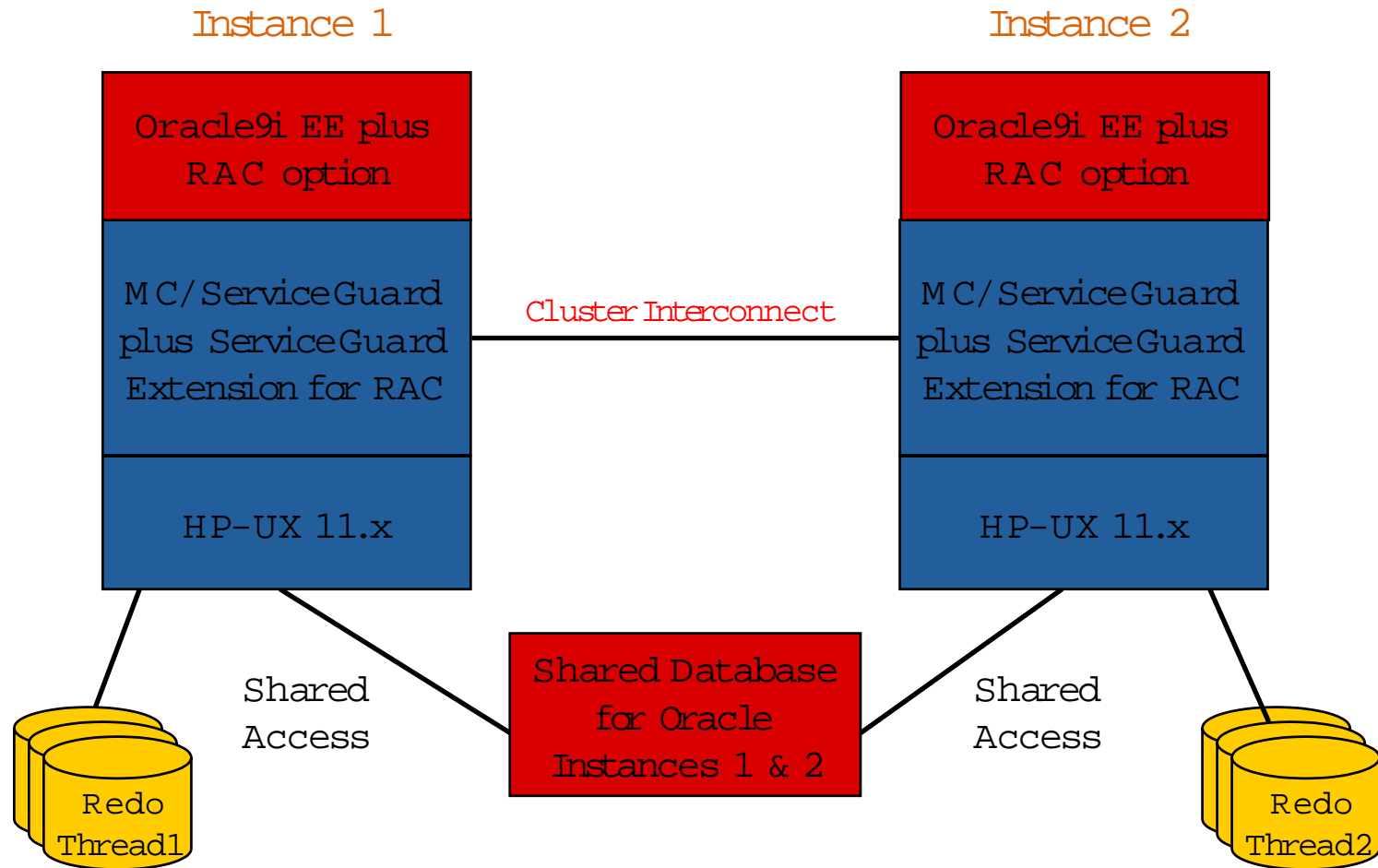


Oracle9i RAC improves upon Cold Failover

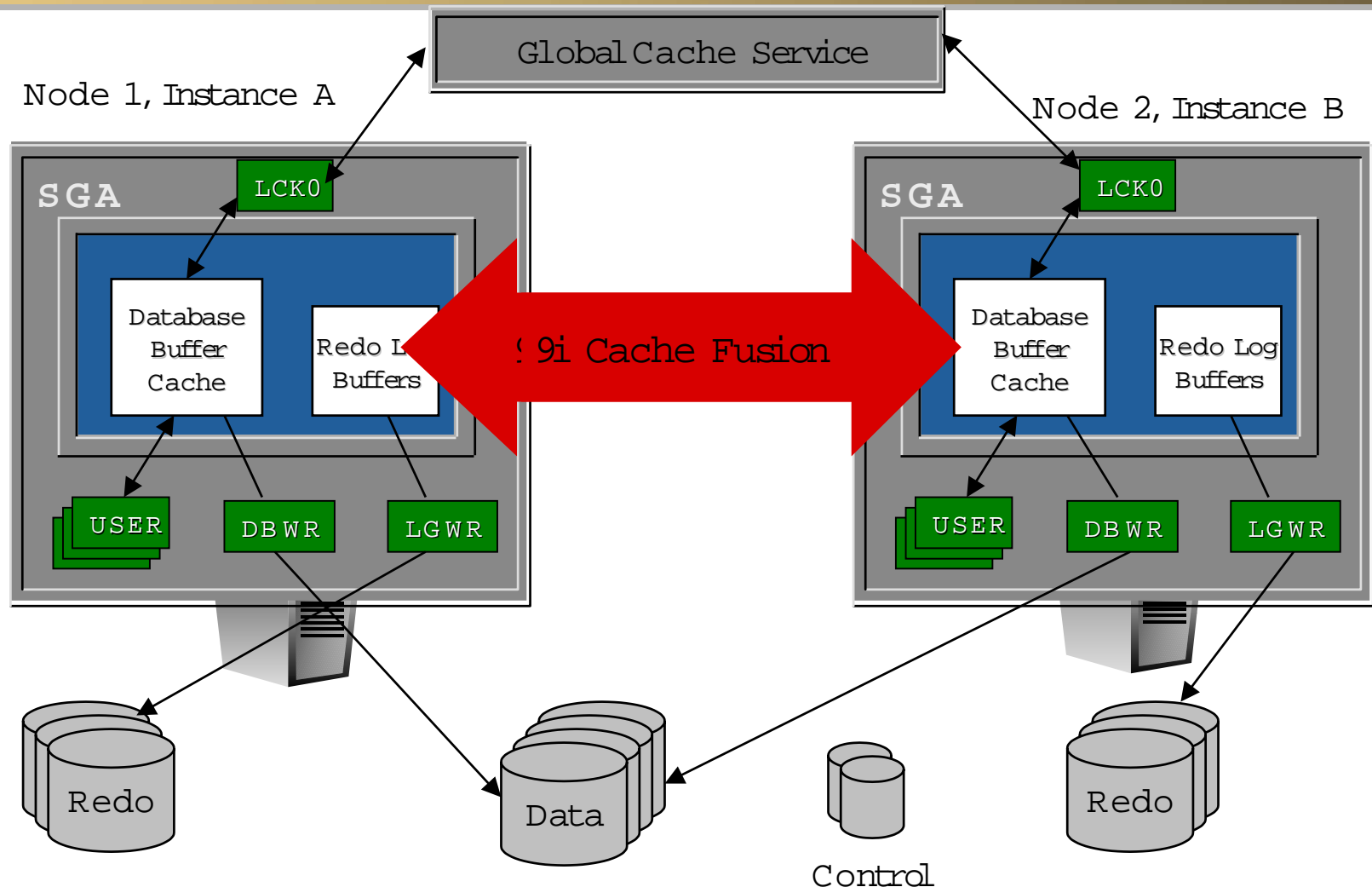


- All systems are active and access the same database
- Increase Scalability beyond one system
- Much faster failover and switchover – no need to move volumes or restart the Oracle instance
- Can preconnect users to the secondary instance for fast restart without logon
- Load can be distributed
- Caches are warm on the secondary instance – better performance after failure
- Requires raw devices or CFS

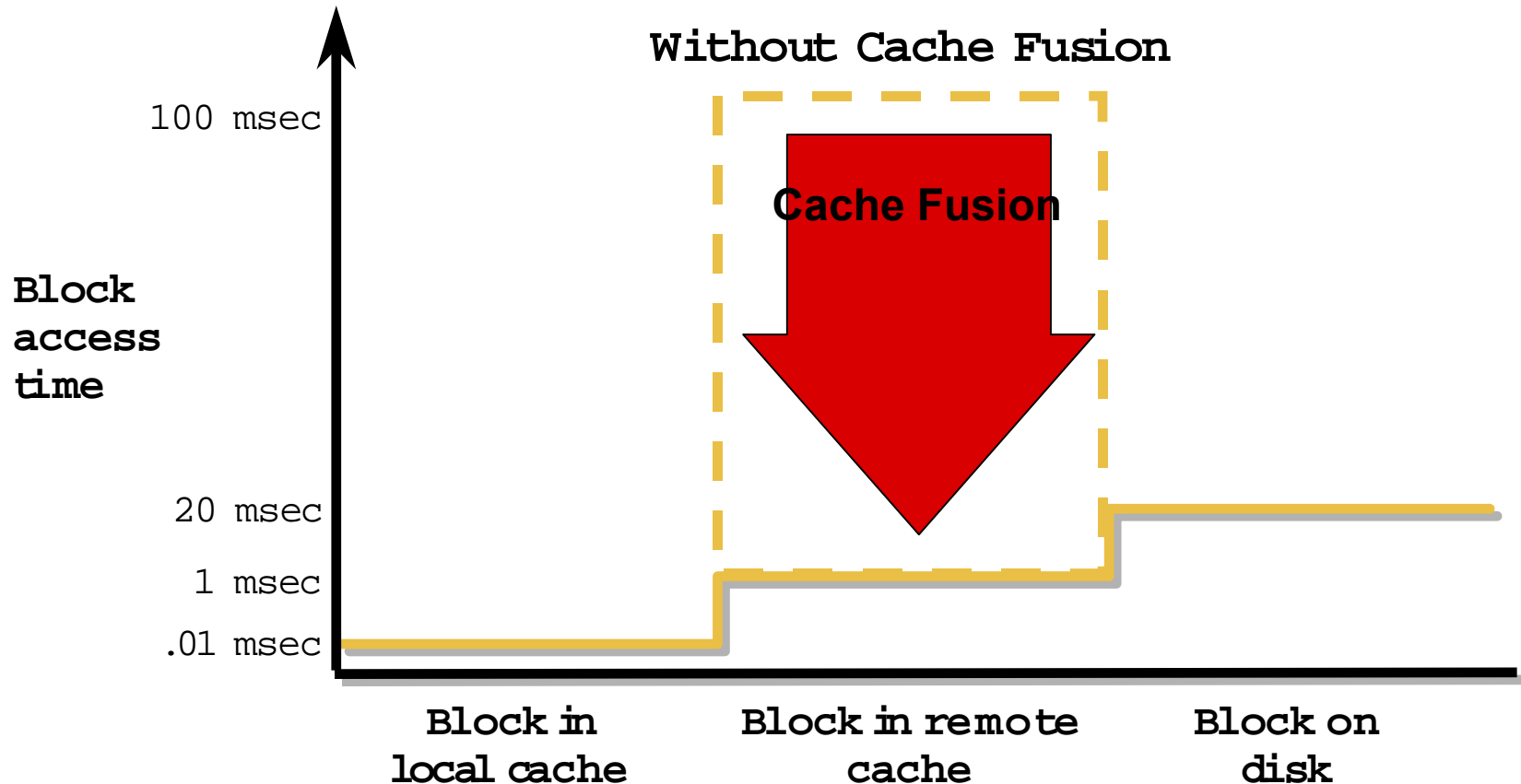
Oracle9i RAC Software Stack



Oracle9i RAC Architecture



Oracle9i RAC Cache Fusion Performance

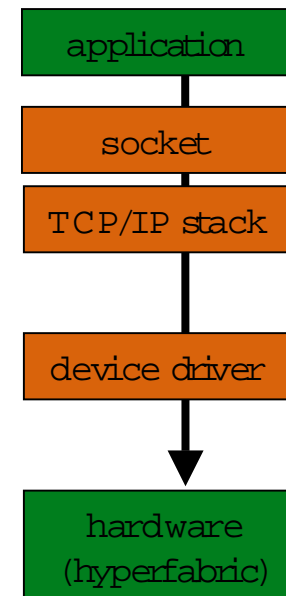


HP's best interconnect for 9iRAC: HMP over Hyperfabric2

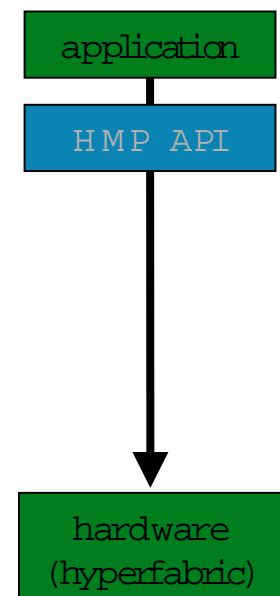


- HyperFabric2 bandwidth increased to 4 Gb/s
- Fast and flexible data center and multi-system application interconnect
- Can be implemented either point-to-point or in switched configuration
- Hyper Messaging Protocol (HMP) first implementation of low latency, high-bandwidth protocol with performance improvement over UDP
- HMP was co-developed by HP and Oracle for exemplary performance
- Lower latency to 22-25 μ sec with HMP through bypassing the kernel services (direct memory access)

standard TCP/IP

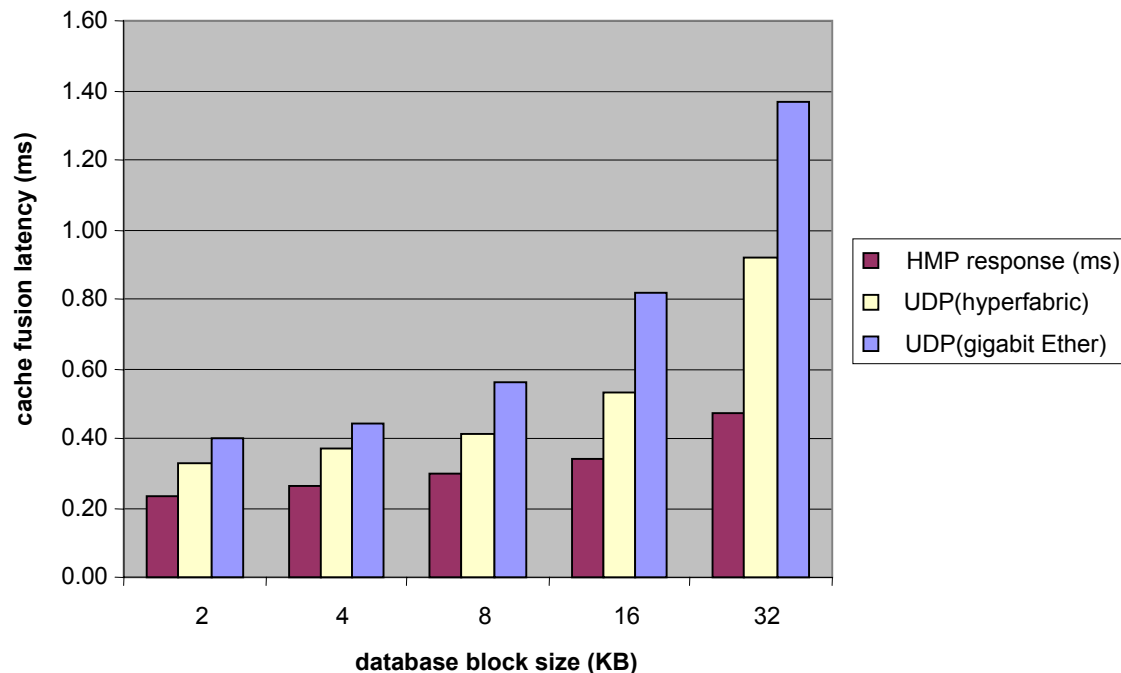


H M P



Cache Fusion Latency

Cache Fusion Latency
Hyperfabric(HMP) vs. GigabitEther(UDP)
HP N4000 8CPU(440MHz), 2node 9iR2 RAC



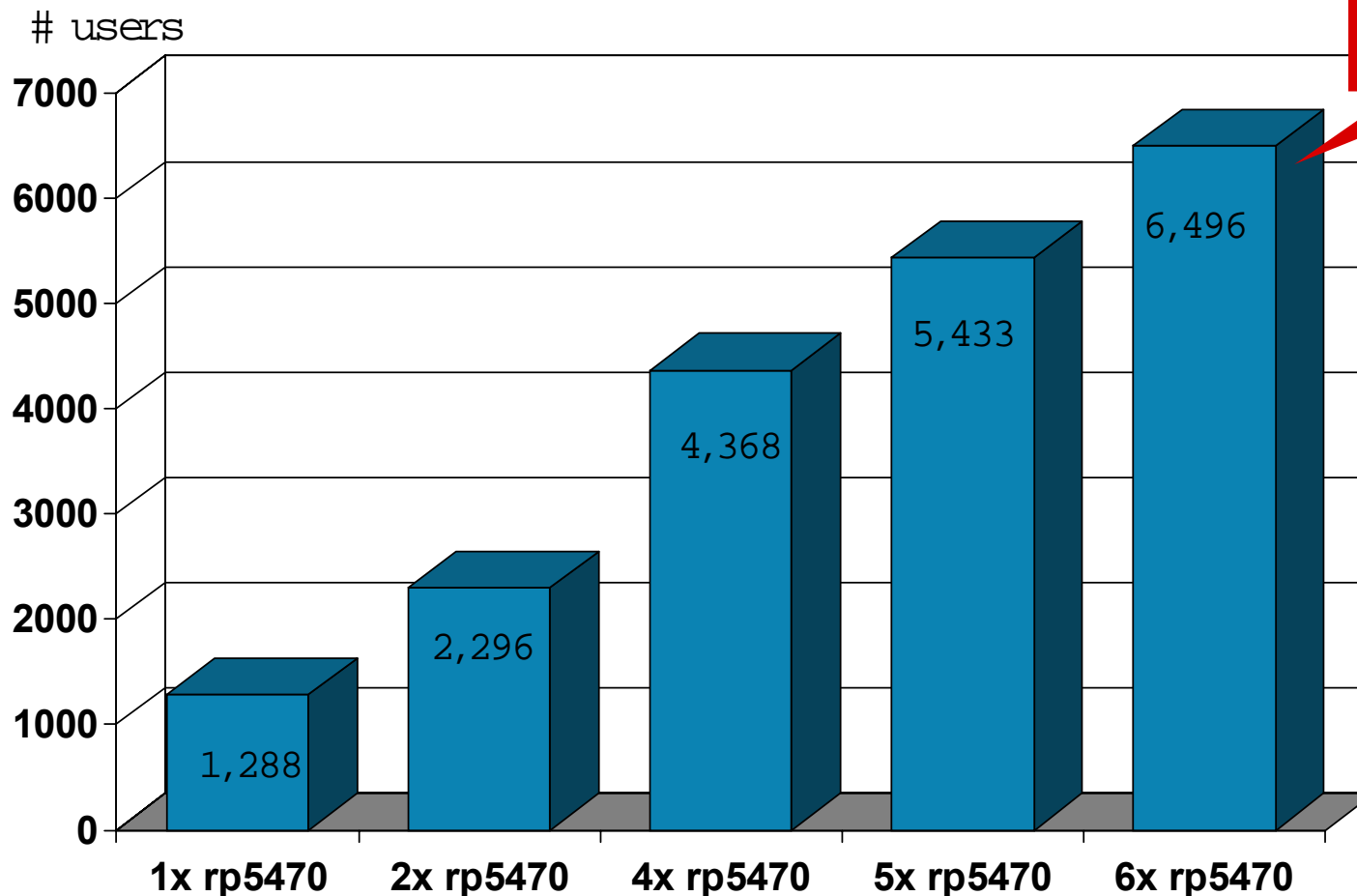
- HMP/HyperFabric vs. UDP/HyperFabric:

30 % (2k) – 49% (32k) faster

- HMP/HyperFabric vs. UDP/GbEthernet:

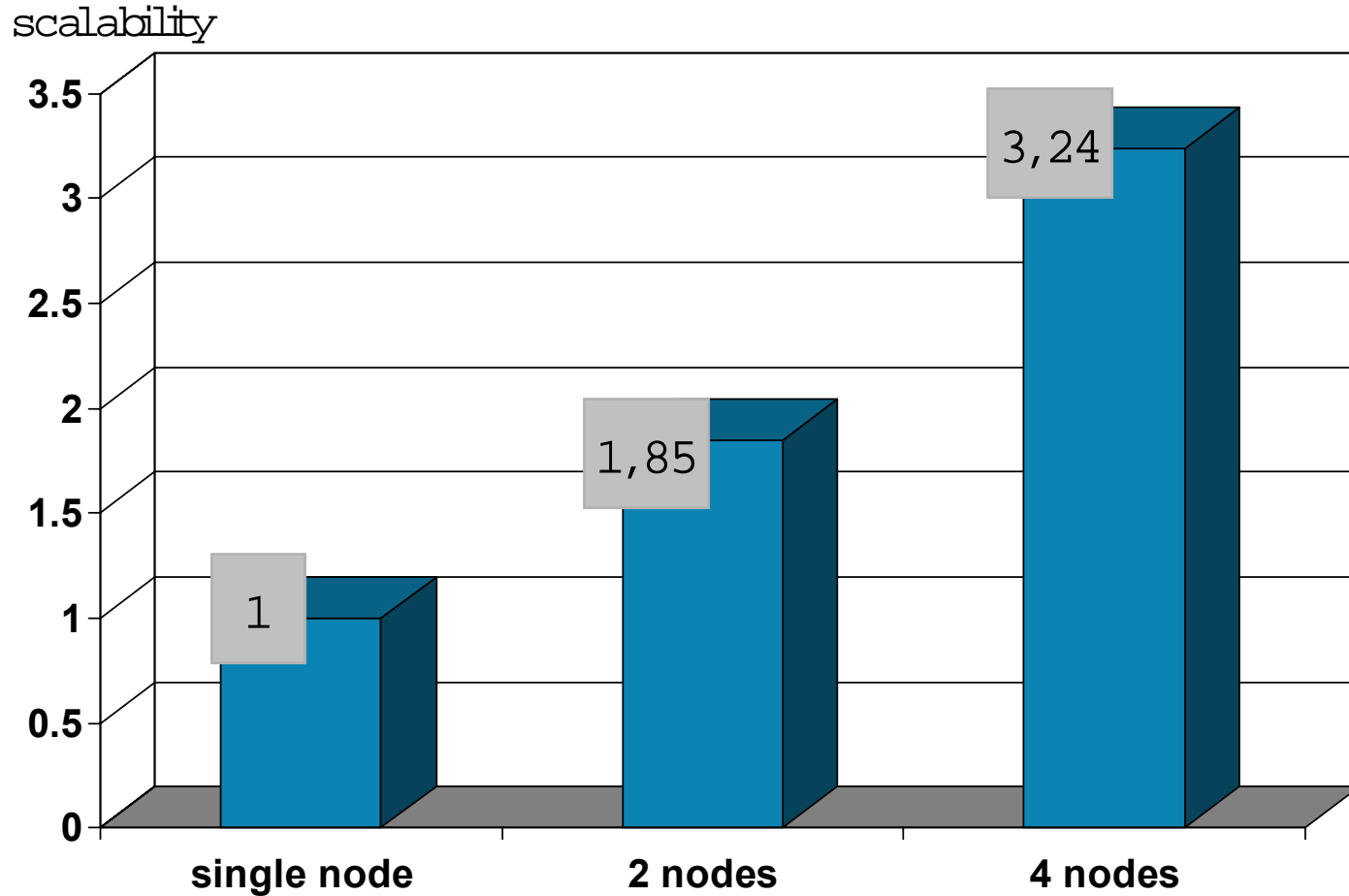
40 % (2k) – 66% (32k) faster

Oracle9i Cluster Scalability with Oracle Applications 11i



http://www.oracle.com/apps_benchmark/html/index.html?results.html

SAP scalability with 9i RAC



- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- **7 Steps to Configure the SG OPS Cluster**
- SLVM versus CVM
- ServiceGuard Manager
- Transparent Application Failover (TAF)



7 Steps to Configure the SG OPS Cluster



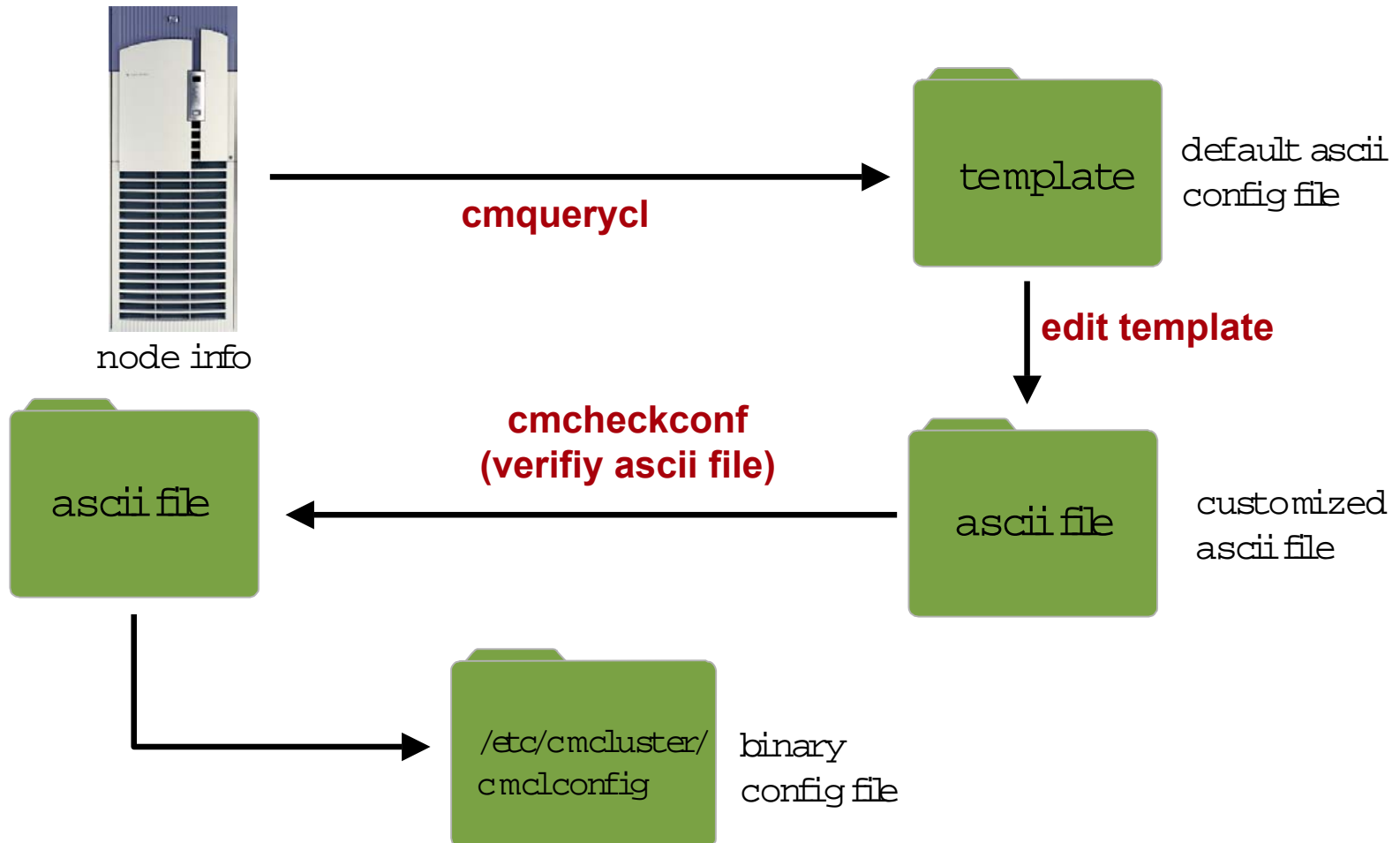
1. Install ServiceGuard OPS Edition
2. Configure each node's disk resources & cluster lock disk
3. Using SAM or the command line: edit the `/.rhosts` or `/etc/cmcluster/cmclnodelist` file to permit root access by all nodes in the cluster.
4. Generate and edit a cluster ASCII configuration file

7 Steps to Configure the SG OPS Cluster (Cont.)



5. Verify the configuration file
6. Generate and distribute a binary cluster configuration file to all nodes
7. Enable Automatic startup and shutdown for the cluster and RAC by configuring the packages for the RAC instances

Cluster Lock: Lock Disk



Step4: Editing Ascii Cluster Config File

```
# *****  
# ***** HIGH AVAILABILITY CLUSTER CONFIGURATION FILE *****  
# ***** For complete details about cluster parameters and how to *****  
# ***** set them, consult the ServiceGuard manual. *****  
# *****
```

```
# Enter a name for this cluster. This name will be used to identify the  
# cluster when viewing or manipulating it.
```

```
CLUSTER_NAME                oracle
```

Cluster Lock Parameters

```
#  
# The cluster lock is used as a tie-breaker for situations in which a running  
# cluster fails, and then two equal-sized sub-clusters are both trying to form a  
# new cluster. The cluster lock may be configured using either a lock disk or a  
# quorum server.  
#  
# You can use either the quorum server or the lock disk as a cluster lock but not  
# both in the same cluster.  
#  
....
```

```
# Lock Disk Parameters. Use the FIRST_CLUSTER_LOCK_VG and FIRST_CLUSTER_LOCK_PV  
# parameters to define a lock disk. The FIRST_CLUSTER_LOCK_VG is the LVM volume  
# group that holds the cluster lock. This volume group should not be used by any  
# other cluster as a cluster lock device.
```

```
FIRST_CLUSTER_LOCK_VG      /dev/vg_slvm
```

Step4: Editing Ascii Cluster Config File (cont.)

```
# Definition of nodes in the cluster.
# Repeat node definitions as necessary for additional nodes.

NODE_NAME          oracle7
NETWORK_INTERFACE  lan0
    HEARTBEAT_IP      15.136.25.207
NETWORK_INTERFACE  lan1
    HEARTBEAT_IP      10.0.0.7
FIRST_CLUSTER_LOCK_PV  /dev/dsk/c5t0d0

# Warning: There are no standby network interfaces for lan0.
# Warning: There are no standby network interfaces for lan1.

NODE_NAME          oracle8
NETWORK_INTERFACE  lan0
    HEARTBEAT_IP      15.136.25.208
NETWORK_INTERFACE  lan1
    HEARTBEAT_IP      10.0.0.8
FIRST_CLUSTER_LOCK_PV  /dev/dsk/c5t0d0

# Warning: There are no standby network interfaces for lan0.
# Warning: There are no standby network interfaces for lan1.
```

Step4: Editing Ascii Cluster Config File (cont.)

```
# Cluster Timing Parameters (microseconds).

# The NODE_TIMEOUT parameter defaults to 2000000 (2 seconds).
# This default setting yields the fastest cluster reformations.
# However, the use of the default value increases the potential
# for spurious reformations due to momentary system hangs or
# network load spikes.
# For a significant portion of installations, a setting of
# 5000000 to 8000000 (5 to 8 seconds) is more appropriate.
# The maximum value recommended for NODE_TIMEOUT is 30000000
# (30 seconds).

HEARTBEAT_INTERVAL          1000000
NODE_TIMEOUT                2000000

# Configuration/Reconfiguration Timing Parameters (microseconds).

AUTO_START_TIMEOUT 600000000
NETWORK_POLLING_INTERVAL 2000000

# Package Configuration Parameters.
# Enter the maximum number of packages which will be configured in the cluster.
# You can not add packages beyond this limit.
# This parameter is required.
MAX_CONFIGURED_PACKAGES    5
```

Step4: Editing Ascii Cluster Config File (cont.)

```
# List of cluster aware LVM Volume Groups. These volume groups will
# be used by package applications via the vgchange -a e command.
# Neither CVM or VxVM Disk Groups should be used here.
# For example:
# VOLUME_GROUP                /dev/vgdatabase
# VOLUME_GROUP                /dev/vg02

# List of OPS Volume Groups.
# Formerly known as DLM Volume Groups, these volume groups
# will be used by OPS cluster applications via
# the vgchange -a s command. (Note: the name DLM_VOLUME_GROUP
# is also still supported for compatibility with earlier versions.)
# For example:
# OPS_VOLUME_GROUP            /dev/vgdatabase
# OPS_VOLUME_GROUP            /dev/vg02

OPS_VOLUME_GROUP              /dev/vg_slvm
```

Step7: Automatic Startup with 9i RAC



- SG OPS packages are coordinating the startup and shutdown of RAC instances.

Just the node on which the instance is running should be configured to run the RAC package.

You must create the RAC instance package with a `PACKAGE_TYPE` of `FAILOVER`, but the fact that you are entering only one node ensures that the instance will only run on that node.

Set the `AUTO_RUN` parameter to `YES` (formerly `PKG_SWITCHING_ENABLED`). This setting determines that the RAC instance will start up as the node joins the cluster.

Set `NODE_FAILFAST_ENABLED` to `NO`.

Enterprise Cluster Master Toolkit for Oracle



- The Enterprise Cluster Master Toolkit for Oracle is a set of templates and scripts that allow you to configure SG packages for Oracle databases.

To simplify the creation of RAC instance packages, you can use the Oracle template provided with the separately purchasable ECM Toolkits product (B5139BA).

Use the special oracle.sh script that is provided, and follow the instructions that appear in the README file.

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluster
- **SLVM versus CVM**
- ServiceGuard Manager
- Transparent Application Failover (TAF)



Types of Volume Managers

- ServiceGuard OPS Edition allows a choice of the following volume managers for data storage:
 1. HP-UX Shared Logical Volume Manager (SLVM) and (optionally) MirrorDisk/UX
 2. VERITAS Volume Manager (VxVM) & VERITAS Cluster Volume Manager (CVM)

- Logical Volume Manager (LVM) is the legacy storage management product on HP-UX.

Included with the operating system, LVM is available on all cluster nodes.

It supports the use of MirrorDisk/UX, which is an add-on product that allows disk mirroring with up to two mirrors.

The ServiceGuard cluster lock disk also is configured using a disk configured in an LVM volume group.

ServiceGuard OPS Edition is shipped with the HP-UX Shared Logical Volume Manager (SLVM), which allows you to activate LVM volume groups for shared write access on multiple nodes configured as a RAC cluster.

VERITAS Volume Manager (VxVM)



- VERITAS VxVM 3.1 (B.03.10.05) was integrated and released with HP-UX 11i in Dec. 2000.
- Latest version VxVM 3.5 and CVM were introduced for HP-UX 11i in Sep 2002.
- For ServiceGuard OPS Edition customers, a CVM two-node shared write activation license is included along with the Base VxVM Volume Manager. However, mirroring and four-node capability require additional licenses.

Prod No	Product Name	Licensing
BASE-VxVM	Base HP Veritas Volume Manager	Free of charge
B9116AA	HP Veritas Volume Manager	Add-on, pay for license
B9117AA	HP VERITAS Cluster Volume Mgr	

SLVM versus CVM

Product	Pros	Cons
SLVM	<ul style="list-style-type: none"> Legacy System is robust and familiar to HP-UX users. Provided free with SGeRAC Supports up to 16 nodes per cluster in shared mode Supports multiple heartbeat subnets 	<ul style="list-style-type: none"> Limited Mirroring Support All shared volume groups must be deactivated before online node configuration. Support for multiple disk data path via PV links with active/passive
CVM	<ul style="list-style-type: none"> Shared disks can remain active during online node configurations Enhanced set of volume management features including software mirroring and RAID0/1. Dynamic Multi-Pathing for active/active devices 	<ul style="list-style-type: none"> Can only be used with up to 4 cluster groups No support RAID5 software data protection. Supports only a single heartbeat subnet. Requires purchase of additional license.

- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluste
- SLVM versus CVM
- **ServiceGuard Manager**
- Transparent Application Failover (TAF)



ServiceGuard Manager



HP ServiceGuard Manager is an intuitive and easy-to-use Java™-based GUI.

Available on HP-UX, Windows and Linux

Will be delivered free with ServiceGuard.

Provides a graphical map and tree of ServiceGuard clusters, nodes and packages

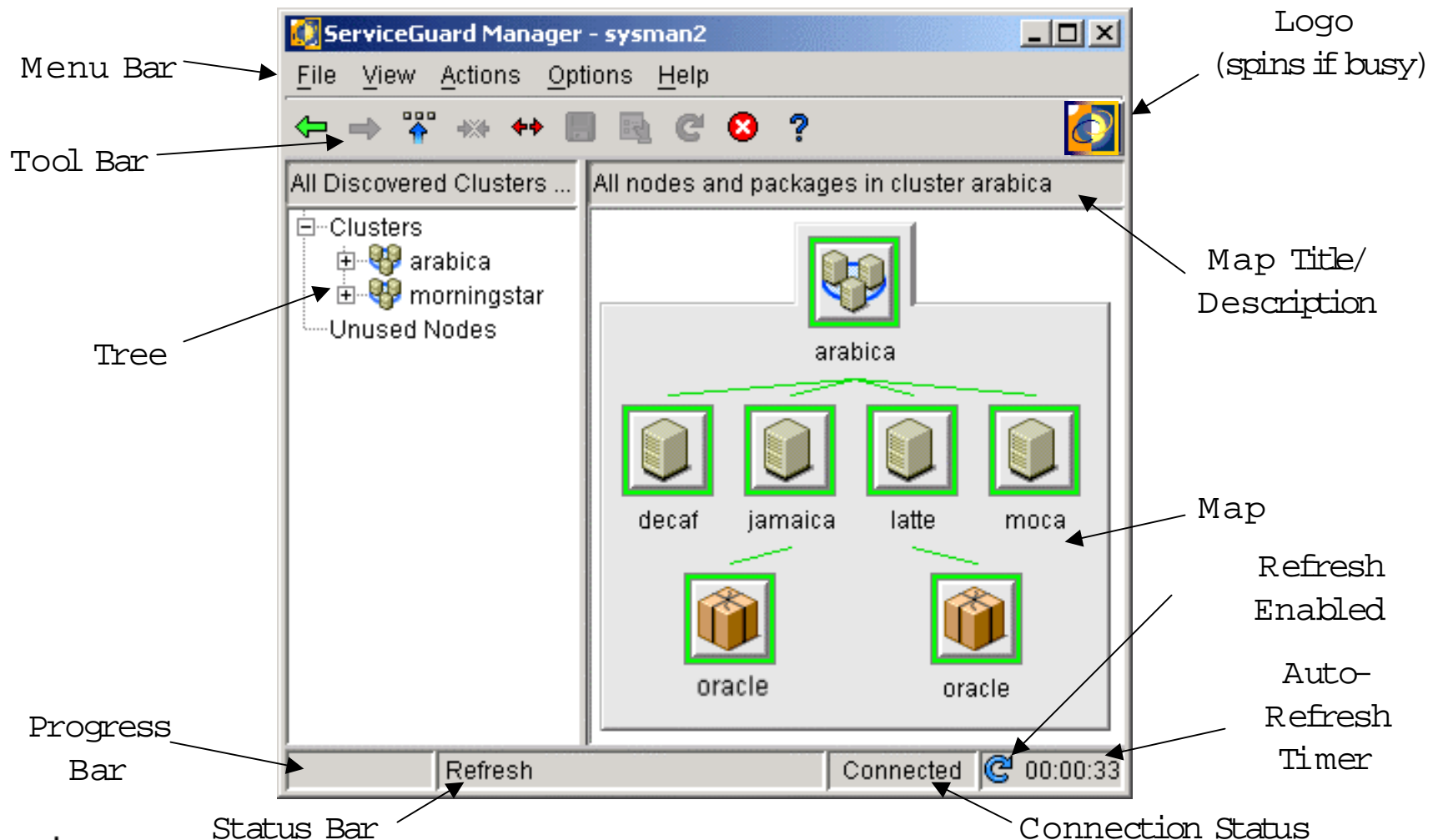
Provides administrative services for changing ServiceGuard cluster node, and packages (run, and halt)

Uses color-coded icons to show status information about a cluster, node or package

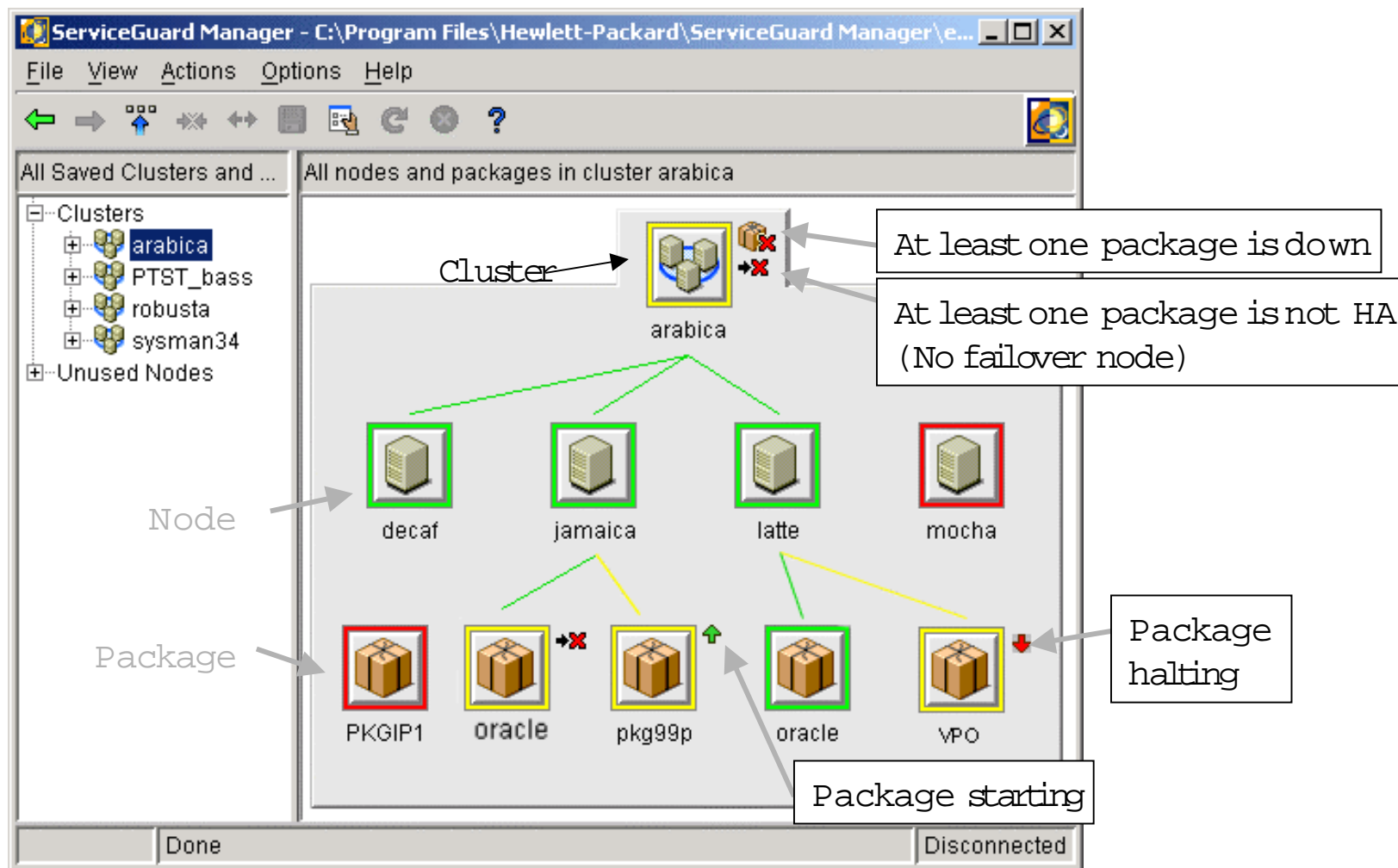
Visually display more clusters from a single point

Understand and resolve cluster issues quickly & accurately

ServiceGuard Manager User Interface



Map Objects and Badges

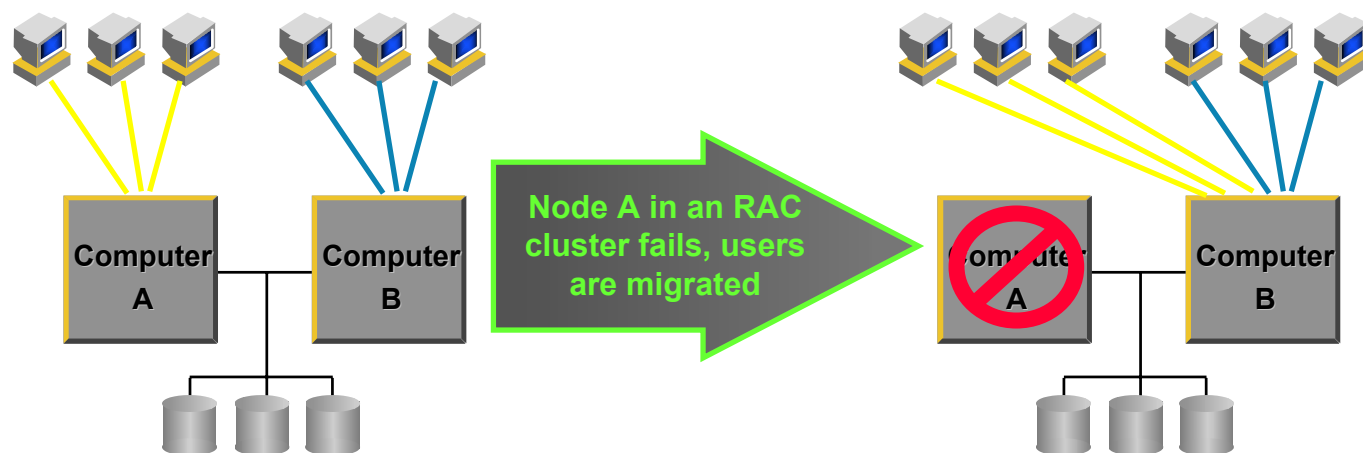


- What is HA
- Overview of MC/ServiceGuard
- ServiceGuard OPS / RAC Edition architecture
- SGeRAC
- Different LVM Versions
- Oracle architecture and functionality
- ServiceGuard OPS Edition Concepts & Configuration
- Group Membership Service
- Volume Groups and async I/O
- Unix Environment
- Oracle 9i RAC architecture
- 7 Steps to Configure the SG OPS Cluste
- SLVM versus CVM
- ServiceGuard Manager
- **Transparent Application Failover (TAF)**



Oracle9i provides a high availability architecture that provides transparent client failover capability:

- Little or no user downtime
- Applications and users are automatically and transparently reconnected to another system
- Applications and queries continue uninterrupted
- Login context maintained



Transparent Application Failover & Load balancing



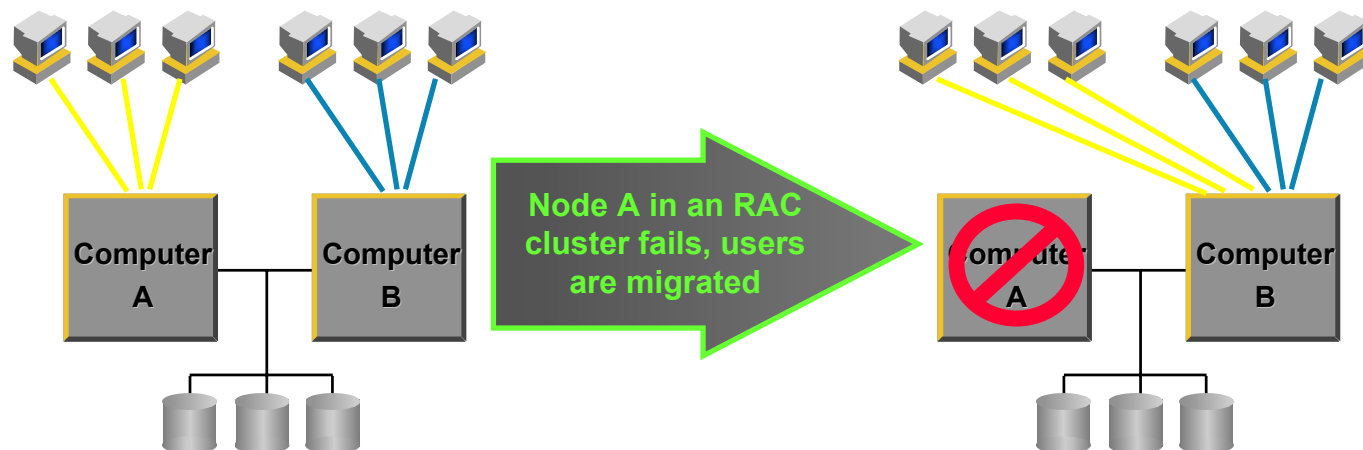
- **Transparent Application Failover (TAF)**
- **Client side load balancing**
- **Server side load balancing**
- **Combined Client + Server side load balancing**

What do you want to avoid?

- Database connections are stateful, simple network failover is insufficient
- Even after a fast database recovery clients were forced to exit their application and reconnect to the database
 - Very intrusive
 - Work lost

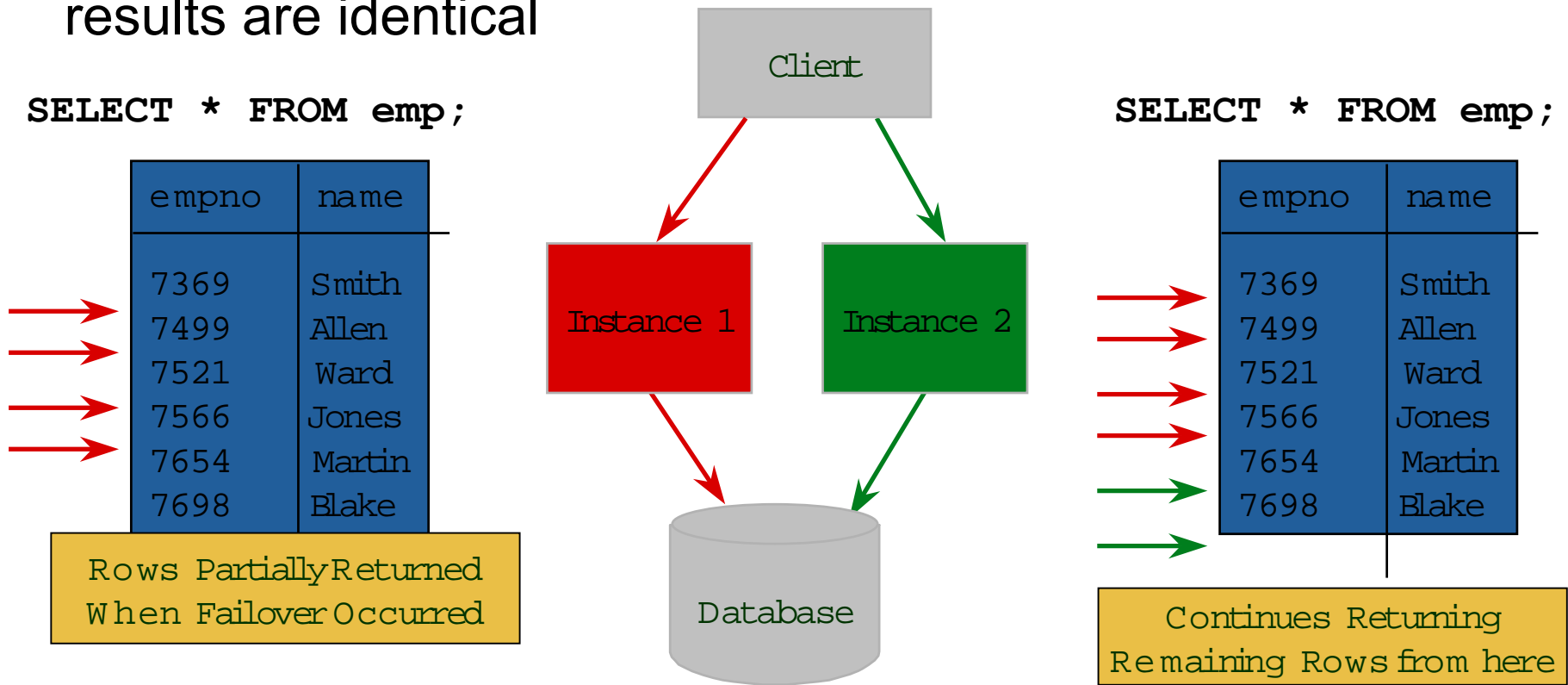
TAF, Overview

- Oracle9i provides a high availability architecture that provides transparent client failover capability
 - Little or no user downtime
 - Applications and users are automatically and transparently reconnected to another system
 - Applications and queries continue uninterrupted
 - Login context maintained



TAF Select Failover

- Failover allows the application to continue execution or fetching
- Leverages Oracle's multi-versioning read consistency to ensure results are identical



- TAF was designed for Real Application Clusters but can also be used for:
 - Replicated database
 - Single instance failover (transparent crash recovery!)
 - Data Guard

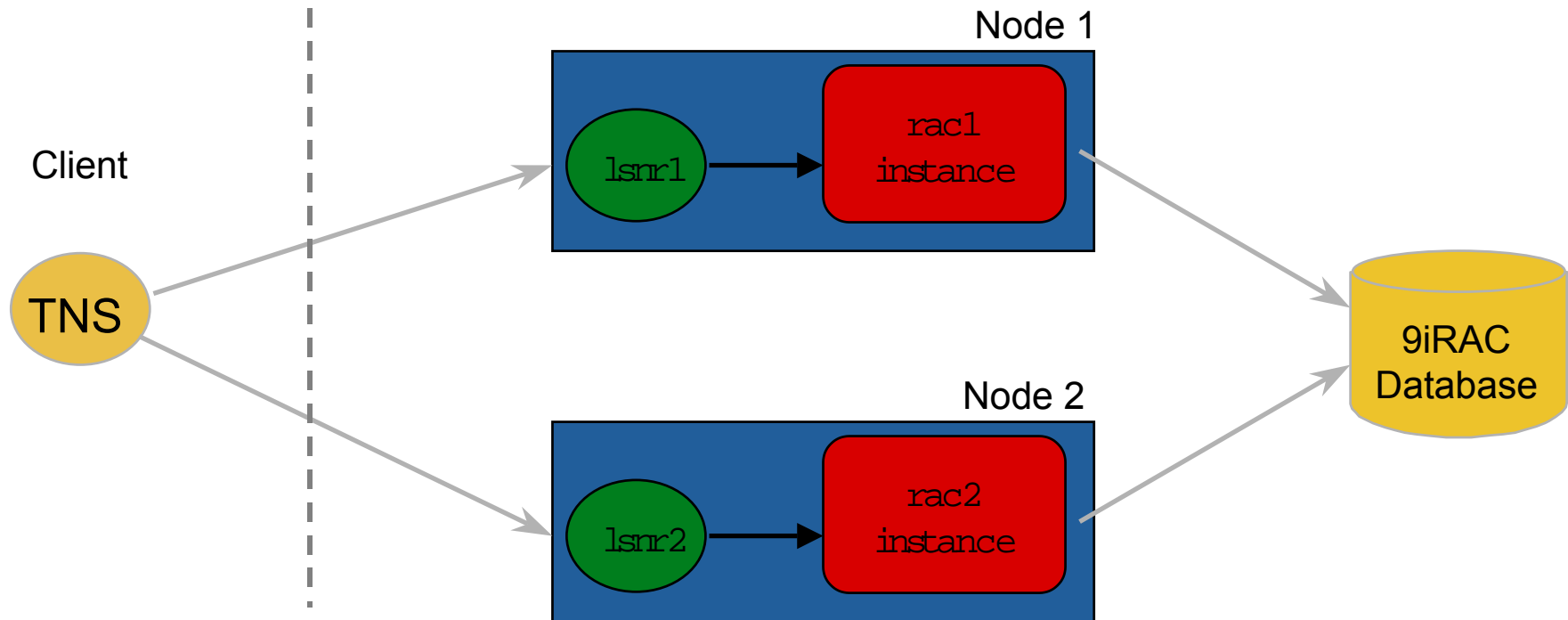
- Limitations:
 - All the applications must be TAF aware (i.e. use the OCI driver) like SQL*Plus, ODBC connection, JDBC Thick Driver (OCI driver), Pro* - precompilers
 - Applications that are not compatible w/ TAF restrictions: Forms, Reports, JDBC thin driver

Characteristics of TAF

- TAF protects or failover:
 - Client/server connection
 - User session state
 - Active cursors (select statements) that have begun to return results
- Not failed over:
 - Active update transactions
 - PL/SQL server-side package variables
- All client applications that use Oracle Call Interface version 8 (OCI8) or higher, implicitly or explicitly, are failover aware:
 - SQL*Plus,
 - ODBC, JDBC/OCI
 - Oracle precompilers,
 - applications using OCI8 and onwards interface and calls

Connect Time Failover

- Automatically retries the connection (uses the next entry in the address list in **tnsnames.ora**)



- Optionally, a client application call-out can be provided that will be executed after instance failure
- This can be used to:
 - Notify users that failover is in progress
“Please Stand By” message
 - Restore user session environment
for example:
`ALTER SESSION SET NLS_LANGUAGE = French;`
- The callback is an application function that the OCI library will automatically call during failover processing
- Totally optional

Where Can You Use TAF?

■ Failover within or between:

- Single instance Oracle (!)
- Real Application Clusters
- RAC Guard (ex OPFS)
- Fail Safe clusters (NT/W2K)
- Replicated databases
- Data Guard (aka Standby) databases

■ Single instance availability

- Transparent database crash and recovery

■ Planned downtime

- Graceful shut down of an instance after all users transactions are completed
- Prescheduled maintenance and orderly user migration

■ DBA driven load balancing

- Load balancing with a predesigned failover strategy

Three “Levels” of TAF Functionality

- TYPE=SESSION, METHOD=BASIC (Login Failover)
 - Client is automatically logged into surviving node of cluster
- TYPE=SELECT (Statement Failover)
 - Node failure occurs during query
 - Client fails over to a surviving node and is logged in
 - Query replayed on surviving node but only rows not returned during the original query execution are returned
- METHOD=PRECONNECT (“Fast” Session Failover)
 - Client connected to two instances at session establishment
 - Avoid impact of “login storm” during failover to surviving node in Real Applications Cluster

TAF, What to Expect After Failover

■ FAILOVER_TYPE = SESSION

- For SELECT, the user does not have to re-connect
- For any uncommitted DML (i.e. insert, update, delete)
 - ORA-25402: transaction must roll back
 - rollback
 - re-issue the statement

■ FAILOVER_TYPE=SELECT

- For a large query, the user will not see any interruption

TAF Retrying a Connection

```
sales.us.acme.com=  
(DESCRIPTION=  
  (ADDRESS=  
    (PROTOCOL=tcp)  
    (HOST=sales1-server)  
    (PORT=1521))  
  (CONNECT_DATA=  
    (SERVICE_NAME=sales.us.acme.com)  
    (FAILOVER_MODE=  
      (TYPE=select)  
      (METHOD=basic)  
      (RETRIES=20)  
      (DELAY=15))))
```


Oracle Net waits 15 seconds before trying to reconnect again.
Oracle Net attempts to reconnect up to 20 times.

Use it also for single-instance crash recovery!!

TAF Pre-Establishing a Connection

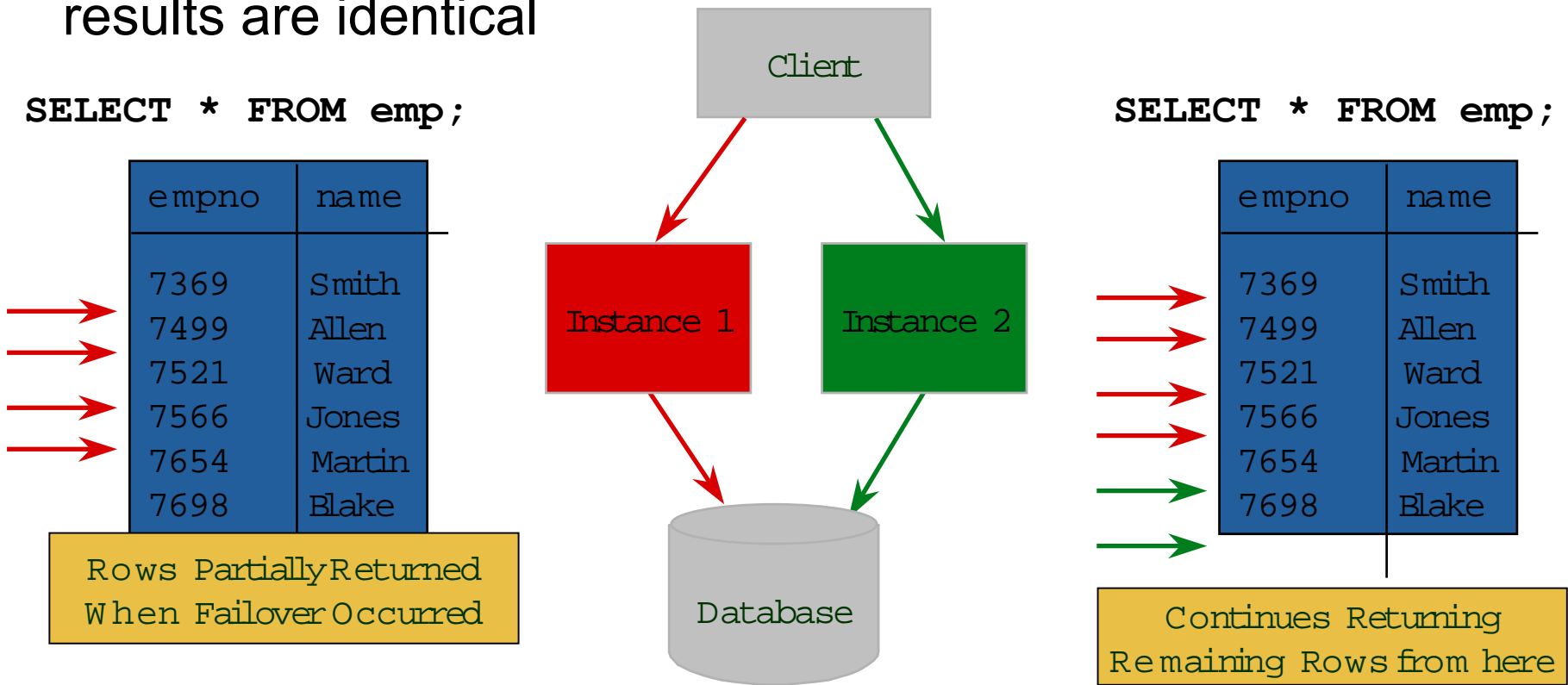
```
sales1.us.acme.com=  
(DESCRIPTION=  
(ADDRESS=(PROTOCOL=tcp)(HOST=sales1-server)(PORT=1521))  
(CONNECT_DATA=  
(SERVICE_NAME=sales.us.acme.com)  
(INSTANCE_NAME=sales1)  
(FAILOVER_MODE=  
(BACKUP=sales2.us.acme.com)(TYPE=select)(METHOD=preconnect)))))
```

```
sales2.us.acme.com=  
(DESCRIPTION=  
(ADDRESS=(PROTOCOL=tcp)(HOST=sales2-server)(PORT=1521))  
(CONNECT_DATA=  
(SERVICE_NAME=sales.us.acme.com)  
(INSTANCE_NAME=sales2)  
(FAILOVER_MODE=  
(BACKUP=sales1.us.acme.com)(TYPE=select)(METHOD=preconnect)))))
```



TAF Select Failover

- Failover allows the application to continue execution or fetching
- Leverages Oracle's multi-versioning read consistency to ensure results are identical



```
SELECT MACHINE, FAILOVER_TYPE, FAILOVER_METHOD, FAILED_OVER,  
COUNT(*) FROM V$SESSION  
GROUP BY MACHINE, FAILOVER_TYPE, FAILOVER_METHOD, FAILED_OVER;
```

The output before failover resembles the following:

MACHINE	FAILOVER_TYPE	FAILOVER_M	FAI	COUNT(*)
-----	-----	-----	---	-----
sales1	NONE	NONE	NO	11
sales2	SELECT	PRECONNECT	NO	1

The output after failover is:

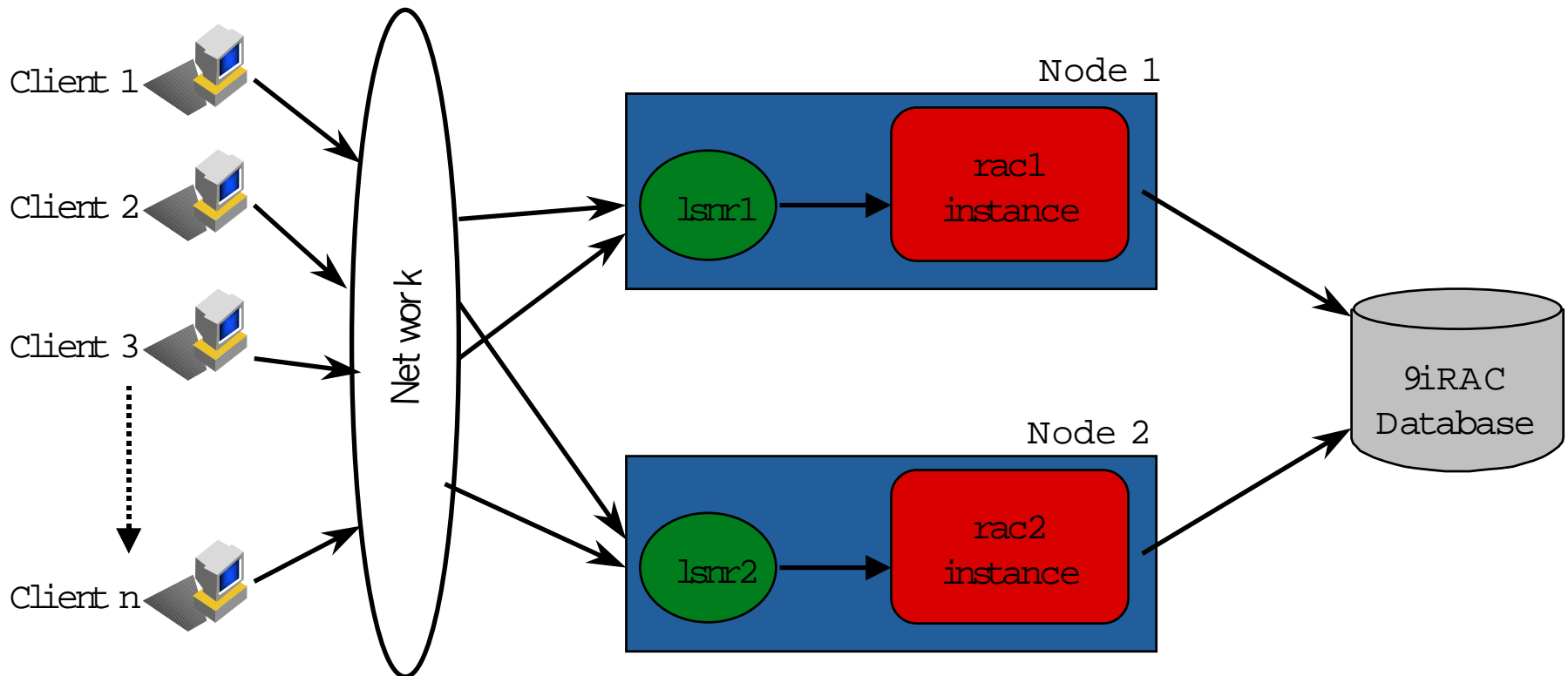
MACHINE	FAILOVER_TYPE	FAILOVER_M	FAI	COUNT(*)
-----	-----	-----	---	-----
sales2	NONE	NONE	NO	10
sales2	SELECT	PRECONNECT	YES	1

Client Side Load Balancing

- Enables clients to randomize connection requests among the listeners
- Oracle9i Net Services progresses through the list of protocol addresses in a random sequence balancing the load on the various listeners
- Without client load balancing, Net Services progresses through the list of protocol addresses sequentially until one succeeds

Client Side Load Balancing

- Clients connect to instance using random method (uses address list in tnsnames.ora)



TAF Failover and Client Load Balancing

```
sales.us.acme.com=  
(DESCRIPTION=  
  (FAILOVER=on)  
  (ADDRESS_LIST=  
    (ADDRESS=  
      (PROTOCOL=tcp)(HOST=sales1-server)(PORT=1521))  
    (ADDRESS=  
      (PROTOCOL=tcp)(HOST=sales2-server)(PORT=1521))  
    (LOAD_BALANCE=on)  
  )  
(CONNECT_DATA=  
  (SERVICE_NAME=sales.us.acme.com)  
  (FAILOVER_MODE=  
    (TYPE=select)  
    (METHOD=basic))))
```

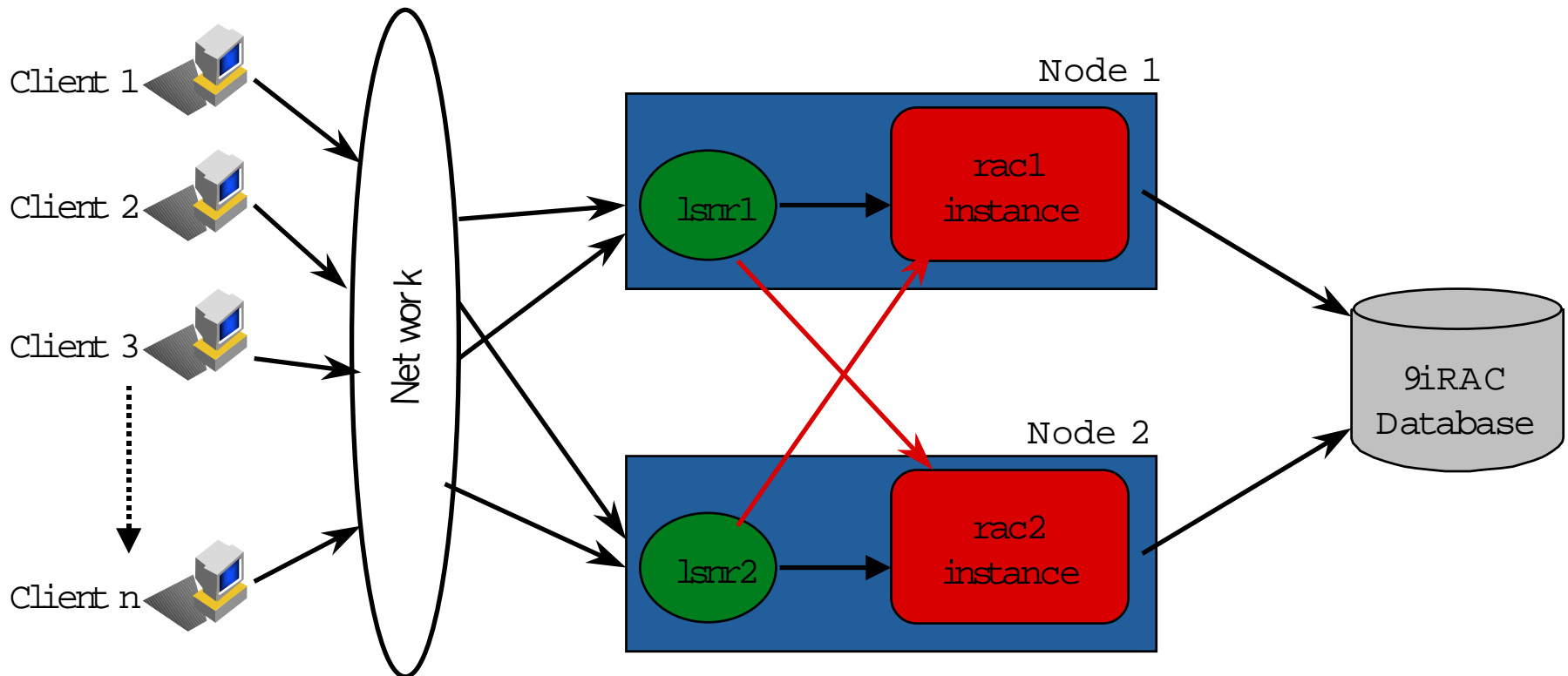
Oracle Net connects **randomly** to one of the listeners on sales1-server or sales2-server that redirects based on load information.

Server side load balancing

- Improves connection performance by balancing the number of active connections among multiple instances (and dispatchers / Shared Server)
- Due to dynamic service registration, a listener is always aware of all instances (and dispatchers) regardless of their locations
- Depending on the load information, a listener decides which instance to send the incoming client request to

Listener Load Balancing

- Listeners balance load using CPU/user load



Load Balancing: Listener registration

- PMON dynamically registers the services to the specified listeners with:
 - service names for each running instance of the database
 - instance names of the database
 - service handlers (dispatchers and dedicated servers) available for each instance

Connection vs. Node load balancing

- Oracle8 - Connection load balancing:

1. Least loaded node (#connections)
2. Least loaded instance
3. Least-loaded dispatcher for that instance

- Oracle9i/8i:

Instead of using the number of active connections as the arbitrator, the actual **operating system load of a node** is used. The node loading value is operating system dependant, indeed not all ports support this functionality.

Prerequisites

1. Static parameters "INSTANCE_NAME" and "SERVICE_NAMES" should be set in "init<SID>.ora". An instance can have only one "INSTANCE_NAME," but more than one "SERVICE_NAMES".
2. It is not necessary to have a "LISTENER.ORA" unless one wants to listen on another PORT (other than 1521).
3. "TNSNAMES.ORA" is required unless names server is used. Use "SERVICE_NAME" instead of "SID_NAME" .

Remote Instance Listener Registration

- If using a non default environment (other ports), to have remote instances registered with the listener, you need to have `local_listener` in your `init.ora` file. In addition, you need an entry `remote_listener` & local listeners

Example `init.ora`:

```
remote_listener=LISTENERS_RAC  
rac1.local_listener='LISTENER_rac1'  
rac2.local_listener='LISTENER_rac2'  
rac3.local_listener='LISTENER_rac3'
```

Server side `tnsnames.ora`

```
LISTENERS_RAC =  
  (DESCRIPTION =  
    (ADDRESS_LIST =  
      (ADDRESS = (PROTOCOL = TCP)(HOST = hp1)(PORT = 1522))  
      (ADDRESS = (PROTOCOL = TCP)(HOST = hp2)(PORT = 1522))  
      (ADDRESS = (PROTOCOL = TCP)(HOST = hp3)(PORT = 1522)))
```

```
LISTENER_rac1 = ...
```

```
LISTENER_rac2 = ...
```

```
LISTENER_rac3 = ...
```

Example: tnsnames.ora

```
RAC.HP.COM =  
(DESCRIPTION =  
  (load_balance=on)  
  (failover=on)  
  (ADDRESS_LIST =  
    (ADDRESS = (PROTOCOL = TCP)(HOST = host1)(PORT = 1521))  
    (ADDRESS = (PROTOCOL = TCP)(HOST = host2)(PORT = 1521))  
  )  
  (CONNECT_DATA =  
    (SERVICE_NAME = rac)  
    (failover_mode =  
      (type = select)  
      (method=basic)  
    )  
  )  
)
```

Things To Be Aware Of ...

■ Service Registration Delay:

- If startup is 1) Listener 2) Instance then service registration is immediate.
- If startup is 1) Instance 2) Listener then PMON fails the initial registration and retries

Workaround: “ALTER SYSTEM REGISTER” upon listener startup.

Instance Listener Registration F&Q



Q? At what frequency does the instance try to register with the Listener? Can this be changed/configured?

A: Almost every minute. No.

Q? Does the instance update the Listener? If yes, at what frequency? Can the frequency be changed/configured?

A: Yes, the instance updates the Listener every 10 minutes apart from registering/unregistering during startup/shutdown. No, the frequency is not tunable.

Q? What happens when the instance is still not registered and a client tries to connect? Will the database then register with the Listener?

A: The client will not connect [ORA-12514]. No, the instance will wake up at its own interval and registers.

Combined Client + Server side load balancing



RAC_no_client_side_load_balance

RAC_no_client_side_load_balance =

(DESCRIPTION =

(LOAD_BALANCE = yes)

(ADDRESS = (PROTOCOL = TCP)(HOST = hp1)(PORT = 1521))

(ADDRESS = (PROTOCOL = TCP)(HOST = hp2)(PORT = 1521))

(ADDRESS = (PROTOCOL = TCP)(HOST = hp3)(PORT = 1521))

(ADDRESS = (PROTOCOL = TCP)(HOST = hp4)(PORT = 1521))

(CONNECT_DATA =

(SERVICE_NAME = rac)

)

)

RAC_client_side_load_balance

RAC_client_side_load_balance =

(DESCRIPTION =

(LOAD_BALANCE = yes)

(ADDRESS = (PROTOCOL = TCP)(HOST = hp1)(PORT = 1521))

(ADDRESS = (PROTOCOL = TCP)(HOST = hp2)(PORT = 1521))

(ADDRESS = (PROTOCOL = TCP)(HOST = hp3)(PORT = 1521))

(ADDRESS = (PROTOCOL = TCP)(HOST = hp4)(PORT = 1521))

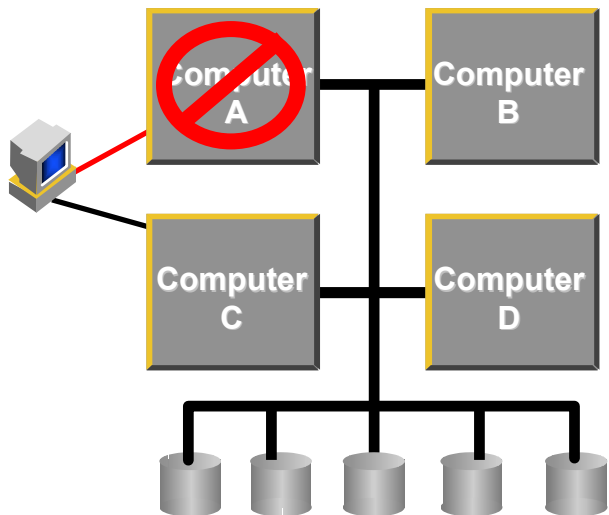
(CONNECT_DATA =

(SERVICE_NAME = rac)

)

)

TAF, Summary



- Transparent Application Failover is a powerful tool for improving application availability
 - Provides near continuous application availability
- Many occasions that would cause substantial interruption to users can be prevented
- Uninterrupted execution of selects across failover
- Application developers can improve failover functionality using callbacks
- Failover can be used for a variety of different purposes, including planned maintenance and upgrades



Interex, Encompass and HP bring you a powerful new HP World.

