Session ID S06

Oracle Real Application Clusters on Linux and IBM xSeries

Stephen Poon

IBM eserver xSeries
Technical Conference

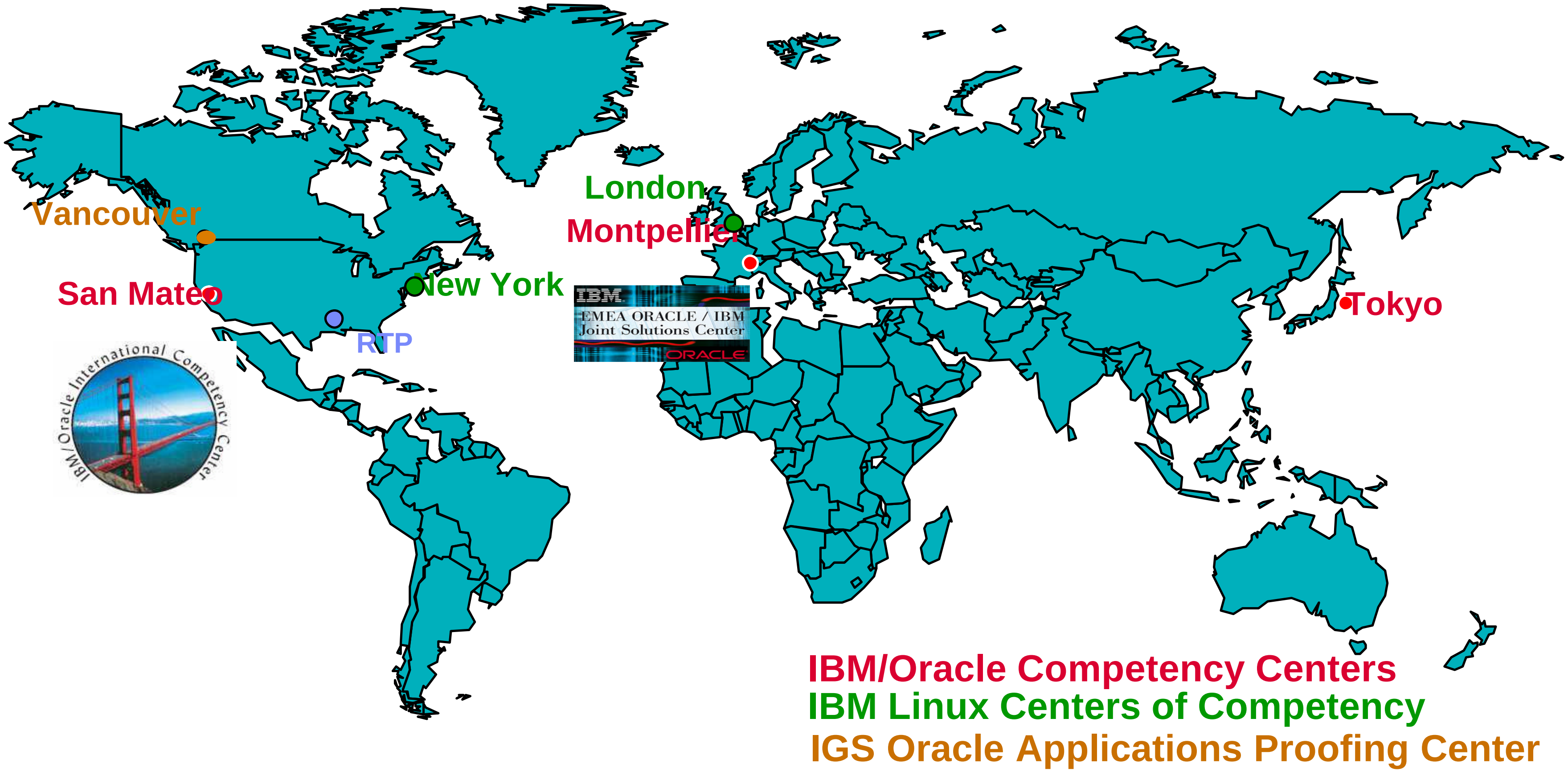
Aug. 9 - 13, 2004

Chicago, IL

Agenda

- Introduction
- Oracle Product Certification
- RAC Overview
- Oracle9*i* RAC Components
- Installing Oracle9*i* RAC
- TAF Configuration

Centers of Intellectual Capital



IBM/Oracle Competency Centers
IBM Linux Centers of Competency
IGS Oracle Applications Proofing Center

IBM / Oracle International Competency Center

- IBM / Oracle International Competency Center, San Mateo, CA
  - Created in 1997 and is IBM's focal point for Oracle technical communication
  - Responsible for sizing tool for Oracle Applications on IBM Servers
  - General sales support, education & customer briefings
  - Has computer laboratory with IBM Servers & Oracle Software
  - Responsible for content updates on IBM internal Oracle Web site (ISV Solution Link)
- ICC Technical projects
  - Validate and maintain Oracle E-Business Suite sizing methodology
  - Gain experience on installation and configuration of Oracle products
    - Basis for technical white papers
- Contact: ibmoracl@us.ibm.com

Oracle Product Certification

- Oracle certifies their products to the operating system and version
- Certifications published at:
  http://otn.oracle.com/support/metalink/index.html
  - Includes RAC Technologies Compatibility Matrix (RTCM) for Linux platforms
    Examples: Fibre Channel, Gigabit Ethernet
- Hardware vendors must certify their servers/storage to these operating systems
  http://www.pc.ibm.com/ww/eserver/xseries/clustering/parallel_server.html
- IBM TotalStorage
  "As a rule, Oracle does not certify or support third party storage products (disk, RAID, tape drive, HSM, etc.) Oracle assumes the underlying storage solution is reliable, and the storage vendors support their products directly."
  Source: http://otn.oracle.com/deploy/availability/htdocs/storage_overview.html
  - Ø IBM storage products certified for OSs by IBM can be deployed in Oracle environments

    FAStT Storage interoperability matrix - http://www.storage.ibm.com/disk/fastt/supserver.htm
    ESS Interoperability Matrix - http://www.storage.ibm.com/disk/ess/supserver.htm

Additional Information for Certification

A Certified Configuration may not be supported by IBM

An xSeries server can be regarded by Oracle as “certified” to run an Oracle product, and supported by the OS vendor (e.g. Red Hat has completed certification testing of a version of the OS)…..yet IBM does not support the OS on that server as IBM ServerProven testing has not yet completed.

Reference sites for hardware certification information

IBM ServerProven

http://www.pc.ibm.com/us/compat/nos/matrix.shtml
http://www.pc.ibm.com/us/compat/eserver.html

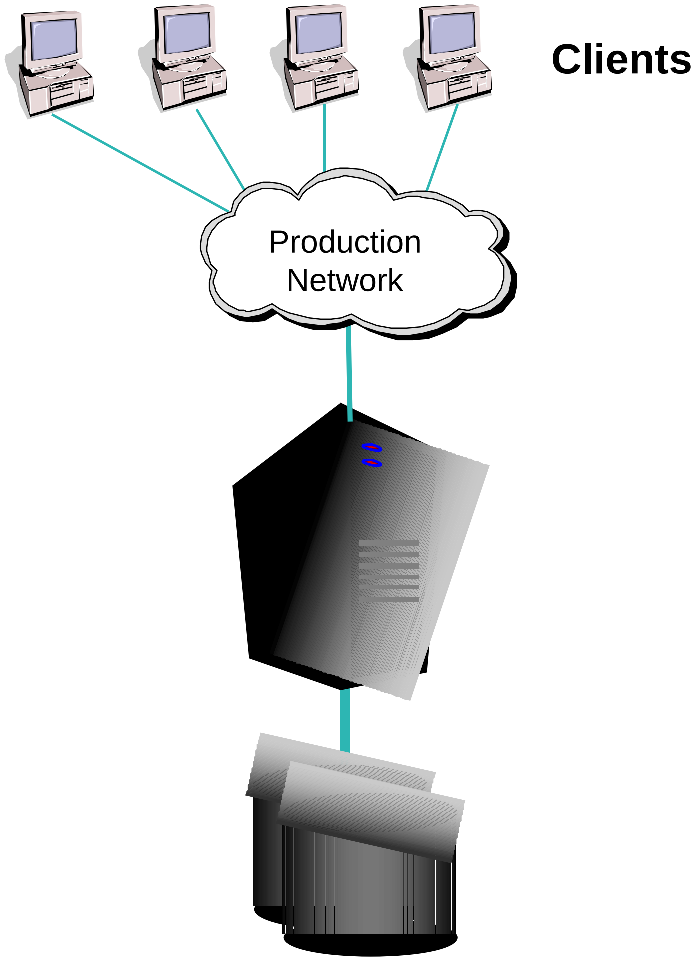
Red Hat

http://hardware.redhat.com/hcl/

SuSE

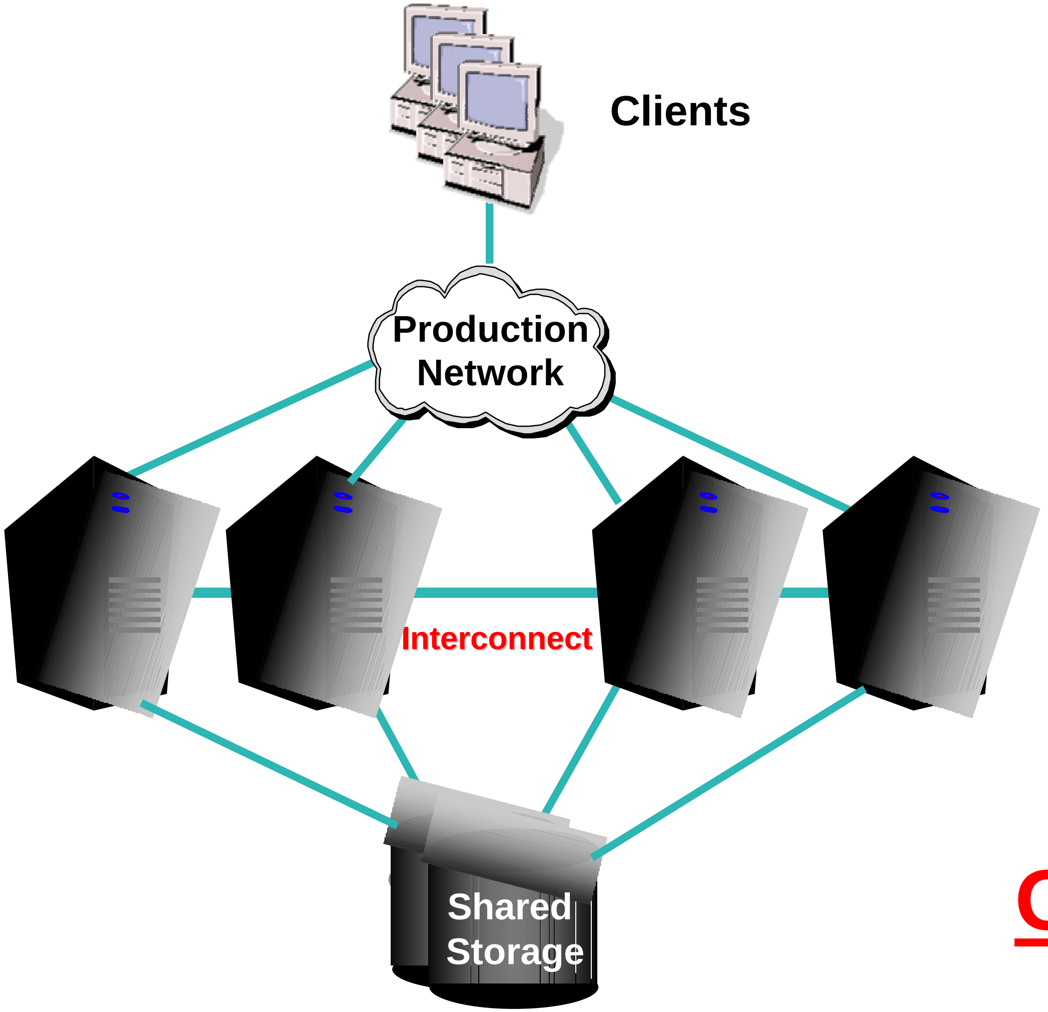
http://www.suse.com/de/business/certifications/certified_hardware/ibm/index.html

Oracle9*i* DB non-RAC



- One database node
- “scale up”
  - Add CPUs, memory
- Additional solutions required for high availability

Oracle9*i* DB RAC

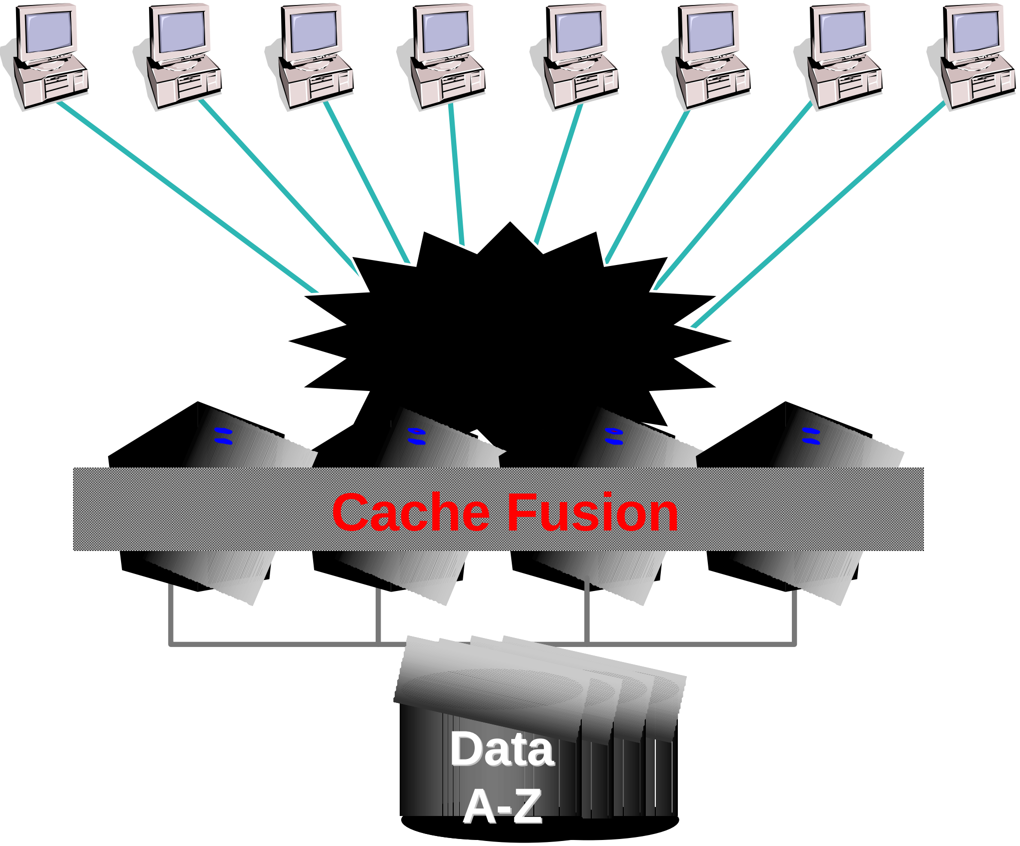


- 2 to n database nodes
- “scale out”
  - Add nodes
- “scale up”
  - Add CPUs, memory
- High availability
  - Transparent application failover
- Load balancing

One Database

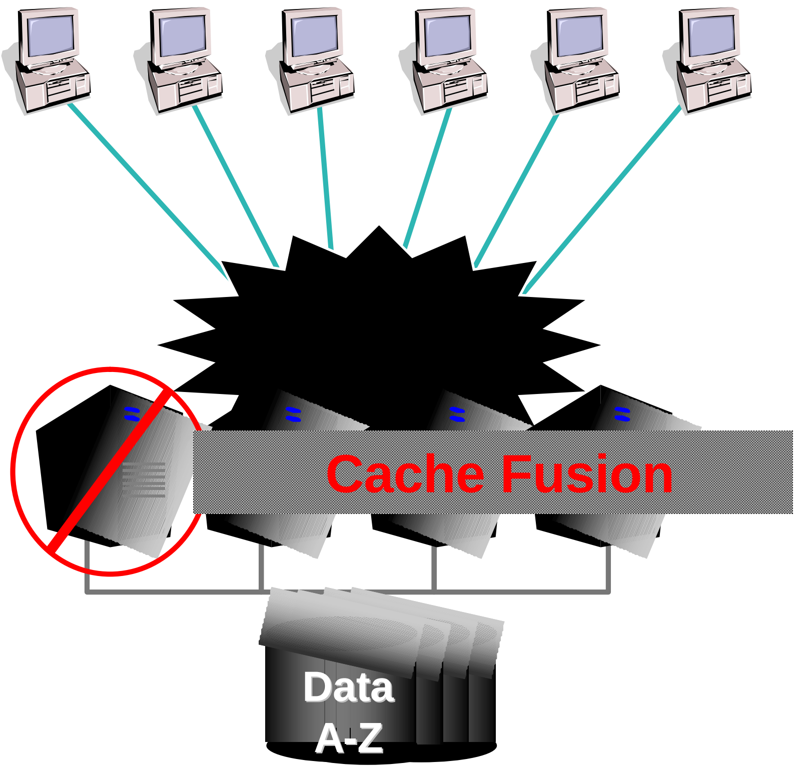
Oracle9*i* RAC Cache Fusion

- Cache Fusion provides high performance access to data from all caches within the cluster
- Scalability for all applications, across multiple computer systems

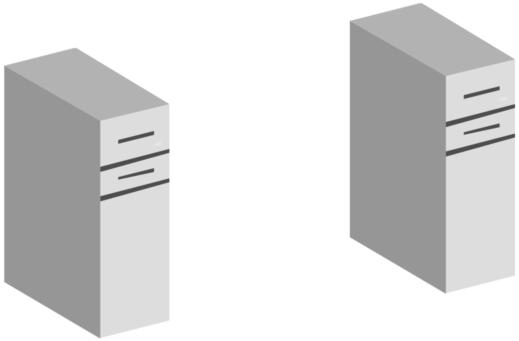


Oracle9*i* RAC Transparent Application Failover

- If a system fails, applications and users are automatically and transparently connected to another system
- Applications continue
- Login context is maintained
- Limitations
  - All applications must be TAF aware (i.e. use the OCI driver)
    - SQL*Plus
    - ODBC connection
    - JDBC Thick Driver (OCI driver)
    - Pro* - precompilers
  - Only select statement will continue
    - Must rollback for the update, insert and delete statements



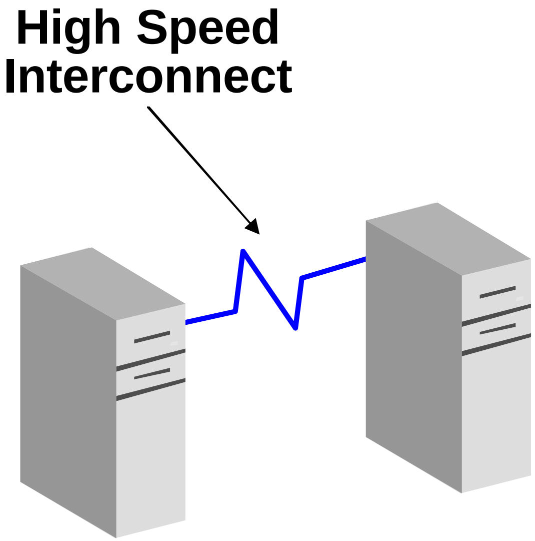
Oracle9*i* RAC Nodes



2 nodes minimum

- Intel or AMD based hardware
- 2 to n nodes
- Same operating system
  - Red Hat 2.1 Advanced Server
  - Red Hat 3.0 EL Advanced Server
  - UnitedLinux 1.0 (SuSE SLES8)
- Same patch levels
- CPU’s and Memory could be different

Oracle9*i* RAC Interconnect

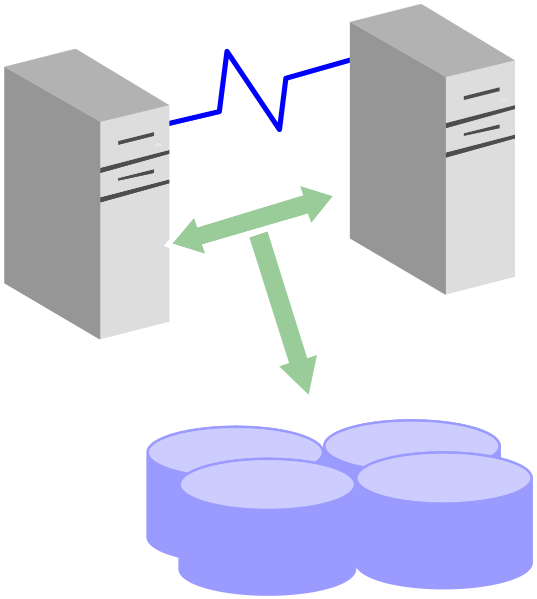


§ Private network used to transfer
  - § Cluster traffic
  - § Oracle Resource directory information
  - § Blocks to satisfy queries

§ Can use a standard network protocol such as TCP/IP

§ Best results are achieved using high speed interconnects
  - ü Gigabit
  - ü Infiniband not currently supported

Oracle9*i* RAC Shared Disk Subsystem



Shared disk subsystem

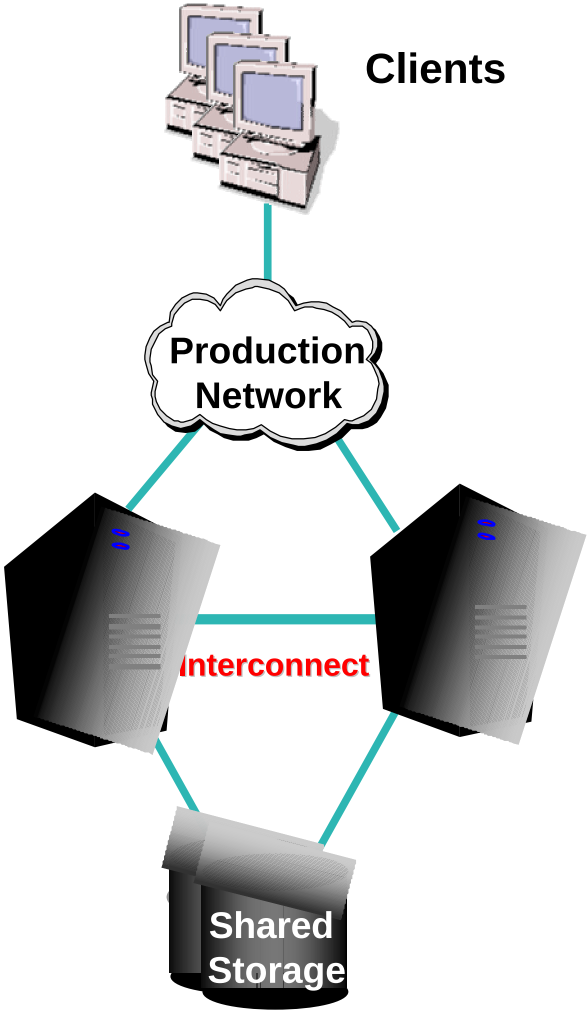
- § Simultaneous access by all nodes to disks used by the cluster software and Oracle database
- § Storage Area Networks (SAN)
  - ü Fiber Channel
    - – FAStT, ESS
  - ü Typically implemented with raw (unformatted) disk partitions
  - ü Clustered file systems
    - – Oracle Clustered File System (OCFS)
    - – Proprietary, vendor-certified cluster file systems
      - • Polyserve
      - • Veritas
      - • Red Hat GFS
- § Network-Attached Storage (NAS)
  - ü NFS
    - – NAS Gateway 500

Oracle9*i* RAC Storage

- § Oracle9*i* RAC requires shared storage for database files
  - ü Raw devices
  - ü Clustered file system
- § Raw devices
  - ü 255 maximum (with Linux 2.4 kernels)
  - ü 14 partitions per SCSI disk (logical drive)
    - § fdisk to create partitions
    - § Bind raw devices to logical partitions
    - § Symbolic links help simplify management
  - ü Extending partitions requires backup/restore
  - ü Requires more experience to administer
- § Clustered File System
  - ü Not part of operating system today
  - ü Simpler to administer (similar to regular filesystem)
  - ü Varying product functionality
  - ü Product examples: Oracle Clustered File System (OCFS), PolyServe, Veritas, Red Hat GFS, IBM GPFS

Oracle9*i* RAC Installation Overview

- § Linux setup (all nodes)
  - ü RPM requirements
  - ü Kernel requirements
  - ü Oracle environment setup
  - ü Network Setup
  - ü Storage driver installation
- § Storage setup
  - ü Create logical drives (LUNs)
  - ü Partition LUNs
  - ü If using OCFS
    - ø Format OCFS filesystems
  - ü If using raw devices
    - ø Map partitions to raw devices
- § Oracle product installation (one node)
  - ü Oracle Cluster Manager
  - ü Oracle9*i*
- § Create database



Oracle9*i* RAC Installation Documents / Papers

- § Oracle Metalink Docid: 184821.1**,** Step-By-Step Installation of 9.2.0.5 RAC on Linux
- § IBM “cookbooks”
  - ü Installing Oracle9i Database with Real Application Clusters on IBM xSeries Model 365 running Red Hat Enterprise Linux AS version 3, April 2004
  - ü Installing Linux and Oracle9i RAC on IBM ^ BladeCenter, March 2004
  - ü Available at

    http://www.pc.ibm.com/ww/eserver/xseries/clustering/parallel_server.html

    or from

    Ibmoracl@us.ibm.com

RHEL 3.0 prerequisites

- If installing Oracle9iR2 on RH EL 3 install the following RPMs:

```
su - root rpm –ivh
compat-db-4.0.14-5.i386.rpm
compat-gcc-7.3-2.96.122.i386.rpm
compat-gcc-c++-7.3-2.96.122.i386.rpm
compat-libstdc++-7.3-2.96.122.i386.rpm
compat-libstdc++-devel-7.3-2.96.122.i386.rpm
openmotif21-2.1.30-8.i386.rpm
setarch-1.3-1.i386.rpm
```

- Create symbolic links so that the older gcc will be used during the Oracle installation

```
su – root
mv /usr/bin/gcc /usr/bin/gcc323
ln -s /usr/bin/gcc296 /usr/bin/gcc
mv /usr/bin/g++ /usr/bin/g++323
ln -s /usr/bin/g++296 /usr/bin/g++
```

  - Can revert back to 323 after Oracle9*i* installation

- Reference: Oracle Metalink docid:252217.1 for more information

Red Hat Linux Setup Considerations

- RPM Requirements
  - server install + development packages (glibc-devel, gcc, kernel source…)
  - kernel source
  - telnet server, ftp server
  - If using OCFS, download and install ocfs rpm's

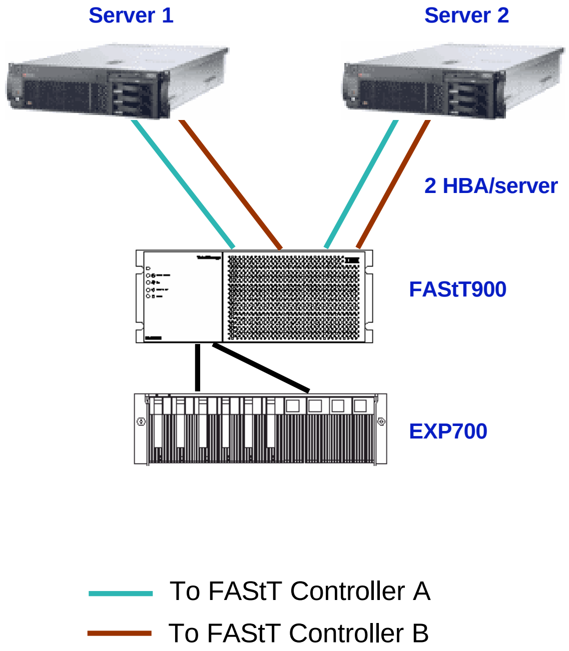
- Kernel requirements

  Red Hat Advanced Server 2.1 U2 or later
    - 2.4.9-e.25 or later
    - 2.4.9-e.xx - Uniprocessor kernel
    - 2.4.9-e.xxsmp - SMP kernel capable of handling up to 4GB of physical memory
    - 2.4.9-e.xxenterprise - SMP kernel capable of handling up to about 16GB of physical memory
    - 2.4.9-e.xxsummit – SMP kernel for xSeries 440/445

  Red Hat Enterprise Linux 3 AS U2 or later
    - 2.4.21-x.EL - Uniprocessor kernel
    - 2.4.21-x.Elsmp - SMP kernel capable of handling up to 16GB of physical memory
    - 2.4.21-x.Elhugemem - SMP kernel capable of handling beyond 16GB, up to 64GB

  Kernel parameters for Oracle
    - /etc/sysctl.conf
    - kernel.shmmax, kernel.sem
    - sysctl -p

IBM Fibre Channel Adapter Driver



- IBM Fibre Channel Adapter drivers
  - IBM FAStT FC-2 Host Bus Adapter non-failover device driver for Linux - V7.00.61
  - IBM FAStT FC-2 Host Bus Adapter Failover device driver for Linux – v7.00.61-fo
  - IBM TotalStorage FAStT Linux RDAC Software Package - v8.4

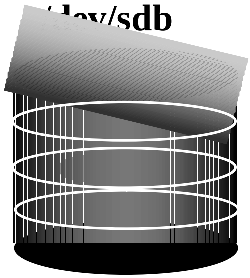
IBM FAStT Storage Manager v8.4 download files matrix

http://www-1.ibm.com/support/docview.wss?rs=0&uid=psg1MIGR-52951&loc=en_US

Shared Storage Configuration – Raw devices

- Using IBM FAStT Storage Manager
  - ü Create logical drives (LUNs)
  - ü Define host mappings
  - ü Logical drives appear to the server as scsi drives, e.g. /dev/sdb

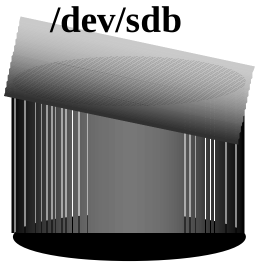
Using fdisk



/dev/sdb1
/dev/sdb2
/dev/sdb3
…
/dev/sdb15

/dev/raw/raw1
/dev/raw/raw2
/dev/raw/raw3
…

Using LVM



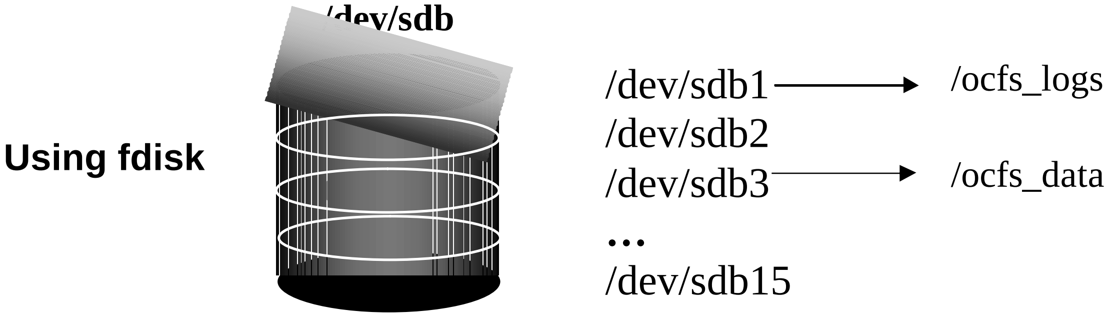
/dev/oravg/quorumlv
/dev/oravg/srvmlv
/dev/oravg/xxxlv
…

/dev/raw/raw1
/dev/raw/raw2
/dev/raw/raw3
…

Not supported by Oracle

Shared Storage Configuration - OCFS

- With OCFS
  - ü Use fdisk to partition logical drives
  - ü Use ocfstool to format one or more partitions
  - ü Mount OCFS filesystem
    - ocfstool
    - /etc/fstab
  - ü Create files on OCFS filesystem



Using fdisk

/dev/sdb

/dev/sdb1 → /ocfs_logs

/dev/sdb2

/dev/sdb3 → /ocfs_data

…

/dev/sdb15

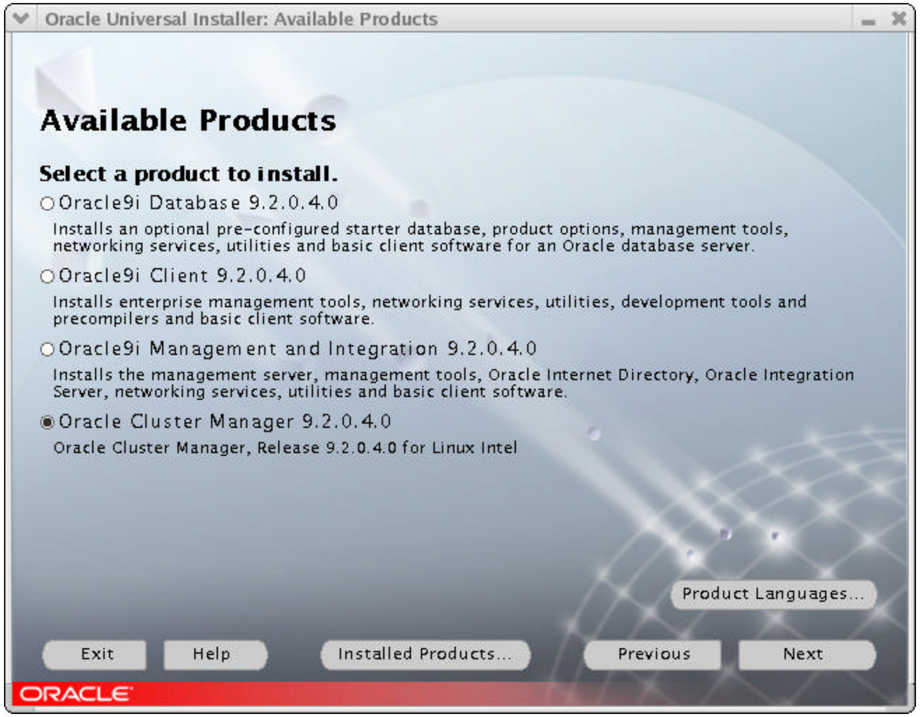
Oracle Cluster Manager Installation

- Preinstallation
  - Create quorum and srvm files or raw devices
  - Apply Oracle patch 3006854 (p3006854_9204_LINUX)
- ORACM 9.2.0.4 installation:
  - As oracle user, run Oracle Universal Installer (OUI) from product CD
  - Install Oracle Cluster Manager
    - The quorum file will be needed during this step
  - Quit OUI when done.
- Modify $ORACLE_HOME/oracm/admin/cmcfg.ora on all nodes
  - Add the following statement -

    KernelModuleName=hangcheck-timer
- Start Cluster Manager on all nodes
  - As root,

    ```
    export ORACLE_HOME=/oracle/92
    $ORACLE_HOME/oracm/bin/ocmstart.sh
    ```
- Verity that Cluster Manager starts

  ```
  ps –ef |grep oracm
  ```
  - Check log in $ORACLE_HOME/oracm/log/cm.log
- Add the startup and shutdown of oracm in the startup script (e.g. rc.local)



Oracle Universal Installer: Available Products

Available Products

Select a product to install.

○ Oracle9i Database 9.2.0.4.0

Installs an optional pre-configured starter database, product options, management tools, networking services, utilities and basic client software for an Oracle database server.

○ Oracle9i Client 9.2.0.4.0

Installs enterprise management tools, networking services, utilities, development tools and precompilers and basic client software.

○ Oracle9i Management and Integration 9.2.0.4.0

Installs the management server, management tools, Oracle Internet Directory, Oracle Integration Server, networking services, utilities and basic client software.

● Oracle Cluster Manager 9.2.0.4.0

Oracle Cluster Manager, Release 9.2.0.4.0 for Linux Intel

Oracle9*i* Installation

- Make sure the Oracle Cluster Manager is running on each node
- As oracle user, run the Oracle Universal Installer and install the 9.2.0.4 binaries
  - Select all cluster nodes
  - Use the custom option and select the products to install (be sure to select Real Application Clusters)
  - Chose not to create the database at this time
- When the popup window appears asking to run the root.sh script, perform the following:

  Create the following directory on all nodes:

  /var/opt/oracle

  Create the following directory on remaining nodes as oracle

  ```
  $ORACLE_HOME/network/log
  $ORACLE_HOME/network/agent/log
  $ORACLE_HOME/network/agent/reco
  $ORACLE_HOME/sysman/log
  $ORACLE_HOME/rdbms/audit
  $ORACLE_HOME/rdbms/log
  ```

  Run root.sh as root on each node, click Ok in the root.sh window

  When OUI launches the Network Configuration Assistant (netca), select Perform Typical Configuration
- Agent configuration will fail. This is a known problem and can be ignored at this time. Patch p3119415_9204_LINUX can be applied later to fix this problem.

Database creation

In order to use any RAC tool, GSD has to be started

gsdctl start

gsdctl stat

Configure listener

netca

Database creation

Scripts created with DBCA

With OCFS,

dbca –datafileDestination /oracle/oradata

With raw devices,

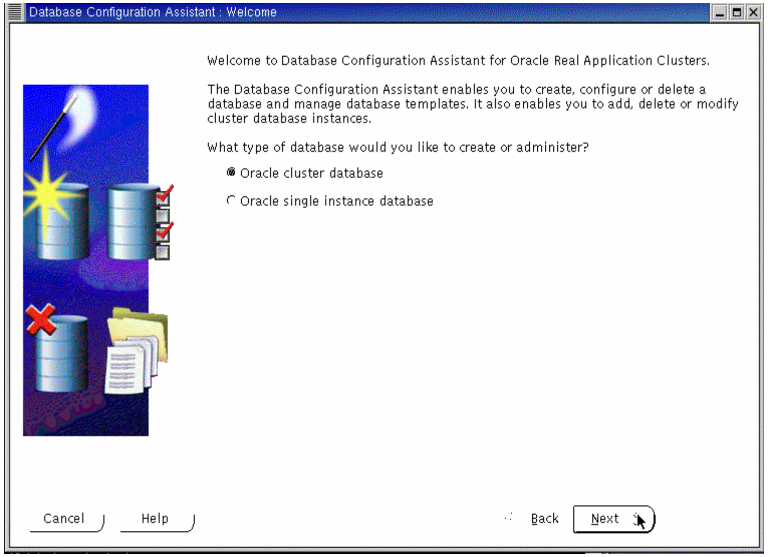
dbca

DBCA_RAW_CONFIG

Create required directories:

/oracle/oradata/{ORACLE_DATABASE_NAME}

/oracle/log/{ORACLE_DATABASE_NAME}



Database Configuration Assistant : Welcome

Welcome to Database Configuration Assistant for Oracle Real Application Clusters.

The Database Configuration Assistant enables you to create, configure or delete a database and manage database templates. It also enables you to add, delete or modify cluster database instances.

What type of database would you like to create or administer?

● Oracle cluster database

○ Oracle single instance database

TAF Configuration

Failover Modes

- Method

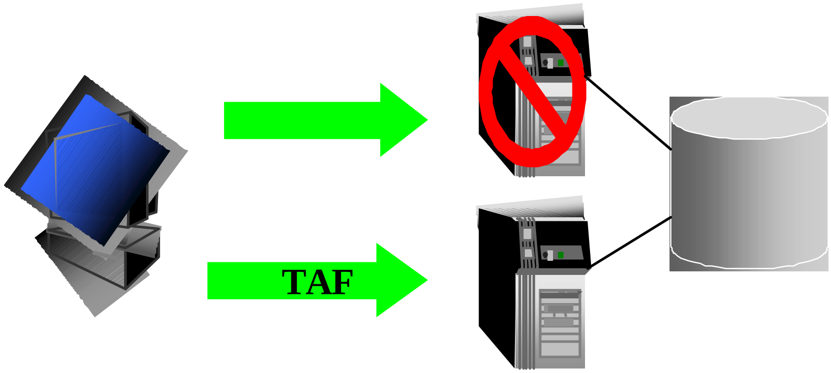
  BASIC: connections are established at failover time.

  PRECONNECT: Parallel connections are made to the primary and backup instance providing faster failover. However, the backup instance must be able to handle the same connection load as the primary.

- Type

  SESSION: If a connection is lost, a new session is automatically created for the user on the backup. SELECT statements are not recovered.

  SELECT: SELECT statements are preserved and fetches will continue on the backup after the failure.



Where to configure TAF

- LISTENER.ORA on the server side

  $ORACLE_HOME/network/admin

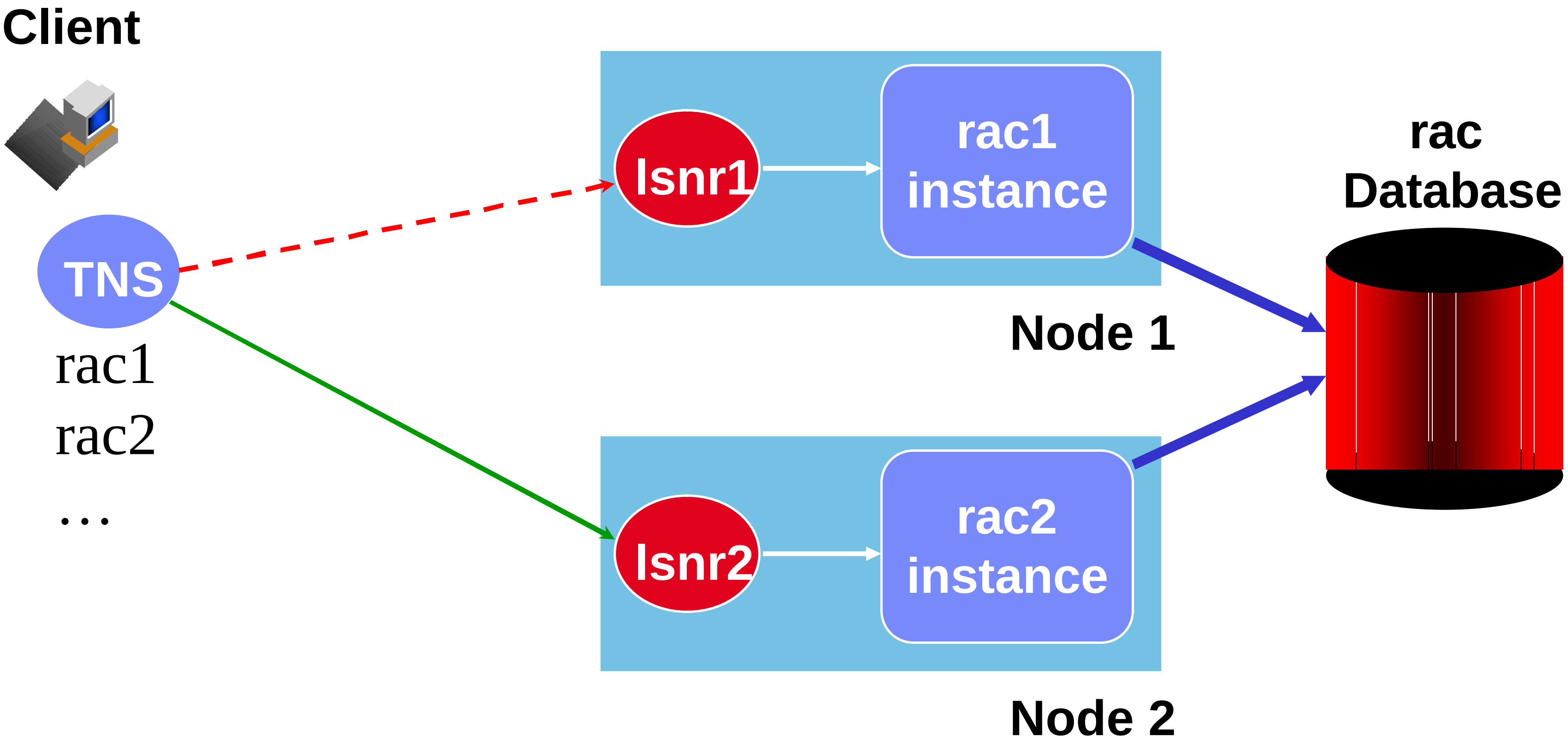
- Mainly **TNSNAMES.ORA** on the client side

  $ORACLE_HOME/network/admin

Connect Time Failover

- Automatically retries the connection

  Uses the next entry in the address list in tnsnames.ora

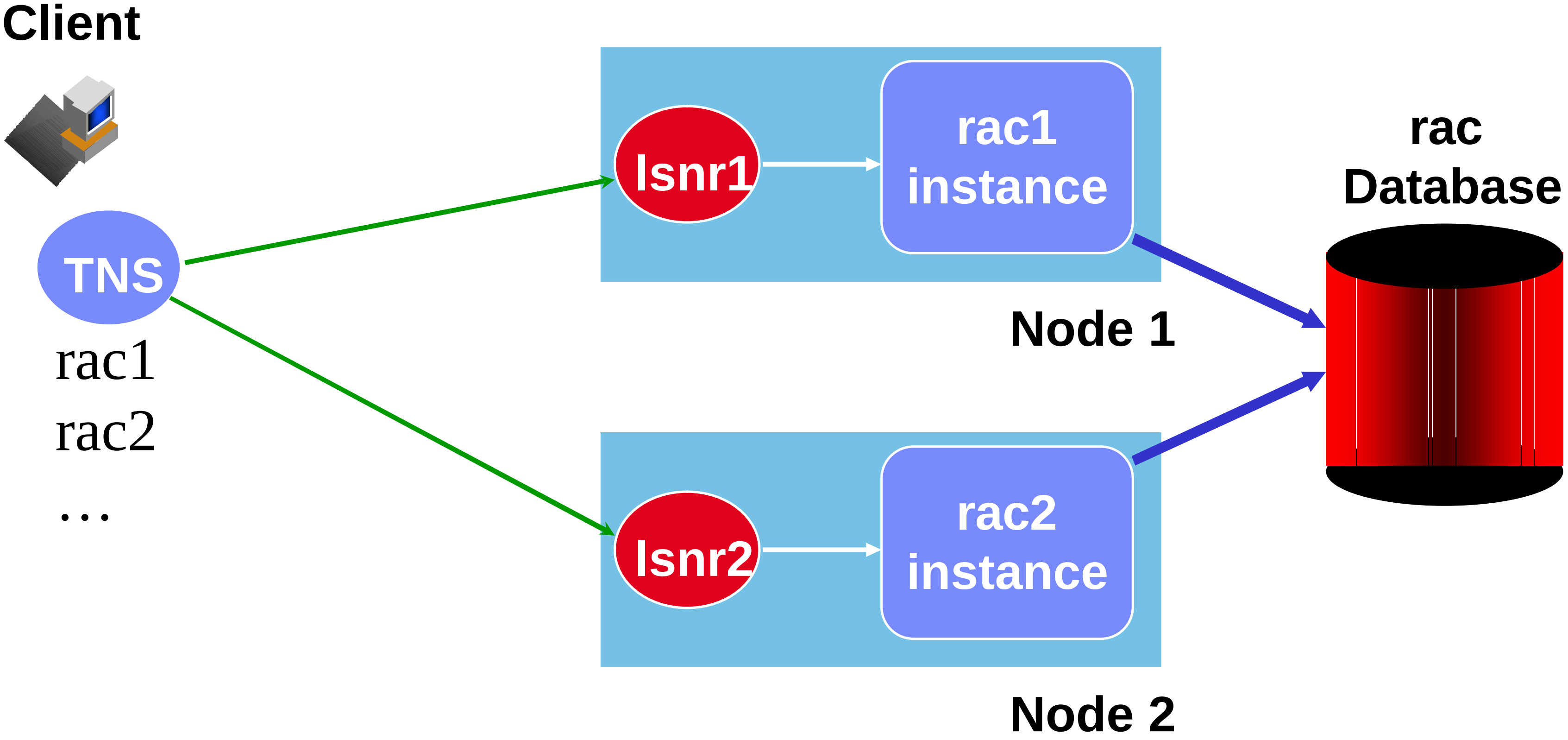


Connect Time Failover Example

```
RAC =
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = rac1)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = rac2)(PORT = 1521))
      (LOAD_BALANCE = off)

      (FAILOVER = ON)

      (CONNECT_DATA =

            (SERVICE_NAME = rac)

            (failover_mode = (type=select)(method=basic))

    )
 )
```

Pre-Connected Clients

- Automatically connected to both nodes, with only one connection (node A or B) used for the transactions.

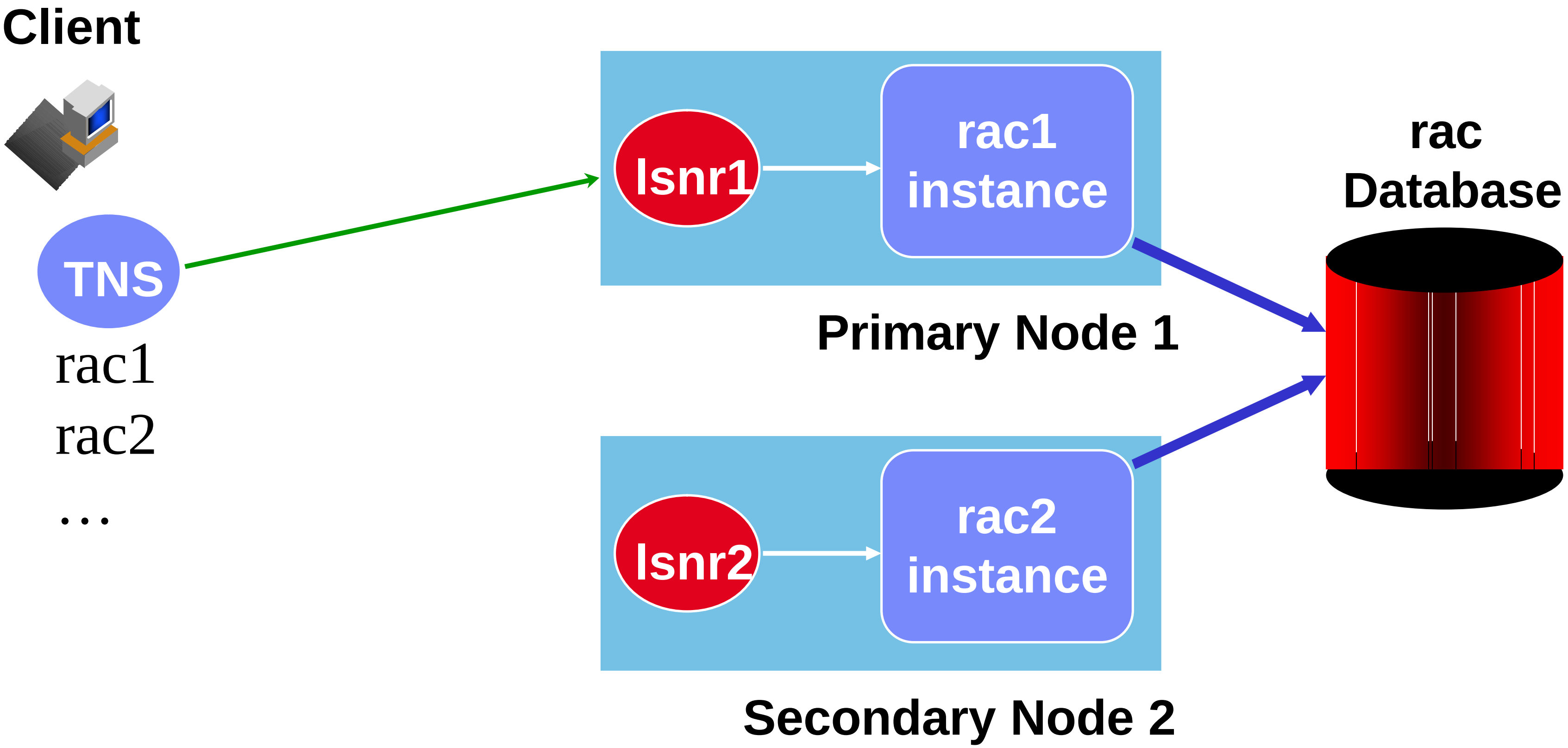


Pre-Connected Clients Example

```
RAC =
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = rac1)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = rac2)(PORT = 1521))
      (LOAD_BALANCE = off)

      (FAILOVER = ON)

      (CONNECT_DATA =

            (SERVICE_NAME = rac)
            (failover_mode = (type=select)(method=preconnect))
    )
  )
```

Primary And Secondary Instance

- Always connect first to the primary node specified in the tnsnames.ora if available.



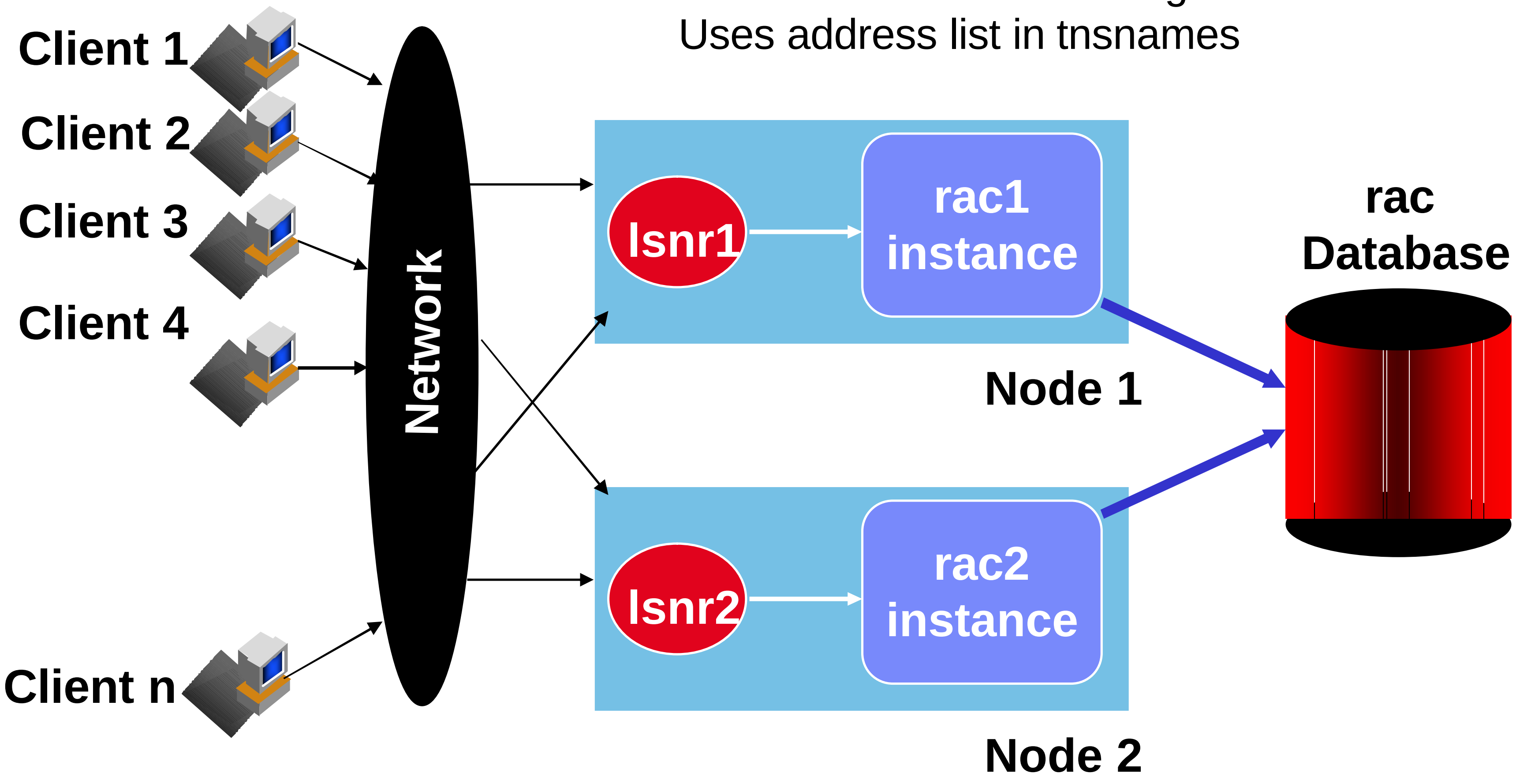
Primary And Secondary Instance Example

```
RAC_pre_ded =

(DESCRIPTION=
(LOAD_BALANCE=off)
(FAILOVER=on)
(ADDRESS=(PROTOCOL=TCP)(Host=rac1)(Port=152
    1))
(ADDRESS=(PROTOCOL=TCP)(Host=rac2)(Port=152
    1))
(CONNECT_DATA=
(SERVICE_NAME=RAC)

(INSTANCE_ROLE=primary)

(SERVER=dedicated)
(FAILOVER_MODE=
(BACKUP=RAC_pre_ded_secondary)
(TYPE=SESSION)
(METHOD=preconnect)
(RETRIES=180)
(DELAY =5)))
)
```

```
RAC_pre_ded_secondary =

(DESCRIPTION=
(LOAD_BALANCE=off)
(FAILOVER=on)
(ADDRESS=(PROTOCOL=TCP)(Host=rac1)(Port=152
    1))
(ADDRESS=(PROTOCOL=TCP)(Host=rac2)(Port=152
    1))
(CONNECT_DATA=
(SERVICE_NAME=RAC)

(INSTANCE_ROLE=secondary)

(server=dedicated)
(FAILOVER_MODE=
(BACKUP=RAC_pre_ded)
(TYPE=SESSION)
(METHOD=preconnect)
(RETRIES=48)
(DELAY =900)))
)
```

Client Side Load Balancing

- Clients connect to instance using random method
  Uses address list in tnsnames

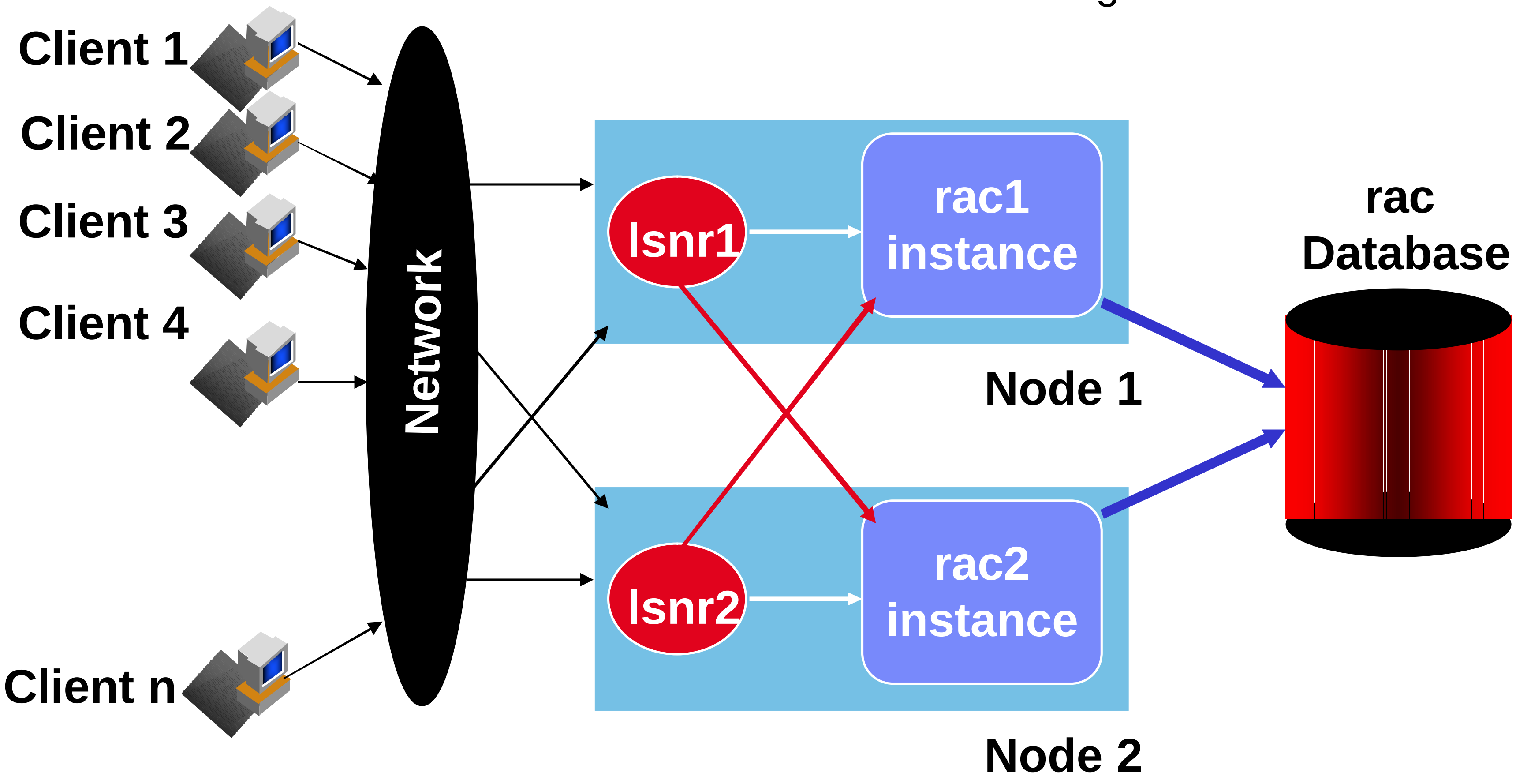


Client Side Load Balance Example

```
RAC =
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = rac1)(PORT = 1521))
      (ADDRESS = (PROTOCOL = TCP)(HOST = rac2)(PORT = 1521))

      (LOAD_BALANCE = ON)

      (FAILOVER = ON)

      (CONNECT_DATA =

            (SERVICE_NAME = rac)
            (failover_mode = (type=select)(method=basic))

    )
 )
```

Listener Load Balancing

- Listeners balance load using CPU / user load



Listener Load Balancing

- Load balancing facilities to allow client connections to be distributed among multiple listeners, dispatchers, instances and nodes
- Balancing the number of active connections
- No single component becomes overloaded.
- PMON sends updated service_register loading information to the listener, i.e. load, maxload, instance_load, instance_maxload.

Listener Load Balancing …

- init.ora Parameters
  - Service_names
    - Oracle Service is a logical way to represent an application:
    - Example: Sales Oracle Service and HR Oracle Service
  - Instance_name
  - Dispatchers
  - Local_listener
  - Remote_listener

TAF Hints & Tips

- Optimizing Switchover Time with TAF Parameters
  - Failover Type = Select
    - allow selects to be ‘preserved’
  - Failover Method = Preconnect
    - pre-established connection to the backup instance
      - Works with primary and secondary instance connection. Don’t use with Load Balancing, this can result in being connected to the same instance twice
  - Retries & Delay
    - Avoids false fail-over but increases failover time

Resources and Tools To Use

xSeries Oracle Website:

http://www.pc.ibm.com/ww/eserver/xseries/clustering/parallel_server.html

xSeries servers

http://www.pc.ibm.com/us/eserver/xseries

xSeries Linux Sales site on System Sales Info And Resources

Competitive, Education and Certification, Customer Presentations, Press, Sales and Technical Support Resources, etc

http://www-1.ibm.com/partnerworld/sales/systems

(Docname: xlinuxskbp)

xSeries Performance Benchmarks Links

http://www.pc.ibm.com/ww/eserver/xseries/benchmarks/

Certification Links

Summary of Linux Distributor Certifications
Public-updated weekly on Wednesday
http://www.pc.ibm.com/us/compat/nos/cert.shtml

Summary of IBM ServerProven (NOS) Testing
Public-updated weekly on Wednesday
http://www.pc.ibm.com/us/compat/nos/matrix.shtml

IBM Software for Linux (Public)
http://www.ibm.com/software/linux

IBM xSeries Update for Linux monthly e-newsletter

http://www.pc.ibm.com/us/eserver/xseries/linux_update/

The Linux Linux website for Business Partners
http://www.ibm.com/partnerworld/linux

Apply to become a Leader for Linux

Receive up to $5K in matched co-marketing Funds, and more
http://www.ibm.com/partnerworld/linux and click "Leaders for Linux"

Linux at IBM

http://www.ibm.com/linux

Customer Reference Database

http://www.ibm.com/eserver/success

IBM TotalStorage FAStT interoperability matrix
http://www.storage.ibm.com/disk/fastt/supserver.htm