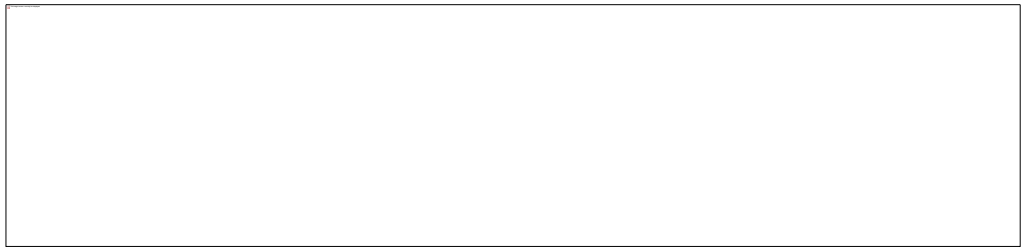
RAC for Beginners

Arup Nanda

Longtime Oracle DBA

(and a beginner, always)



April 7-11, 2013 – Colorado Convention Center

RMOUG Training Days Attendees Receive the Lowest Possible Price to attend COLLABORATE! - $1,295 - $600 Savings

- Choose “Member Group Rate”
- Enter “RMOUG” into Discount Code Box (Non IOUG Members)
- Enter “RMOUGM” into Discount Code Box (IOUG Members)
- Enter Zip Code into Hotel Registration Code Box
- Must Register Before March 6th!

Visit IOUG at Booth #18 for more information!

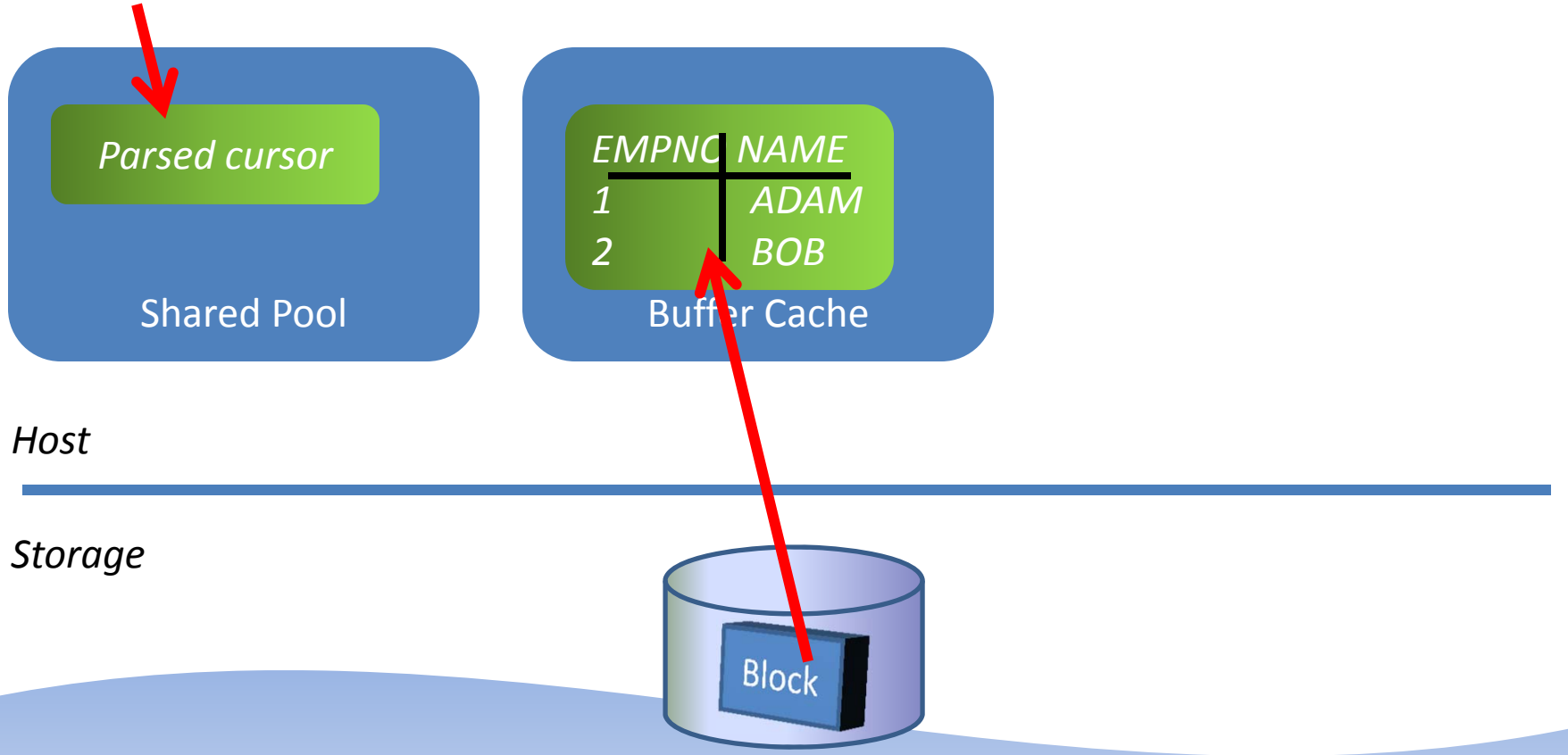
www.collaborate13.ioug.org

What Is This?

- RAC of often a confusing topic for most DBAs
- What exactly is RAC and how it is different
- Terminology galore
  - Cache Fusion
  - Global Locking
  - Global Enqueue Service (GES)
  - Global Cache Service (GCS)
  - Interconnect
  - Virtual IP (VIP)
  - SCAN IP, SCAN Listener
- You will learn about all these and more

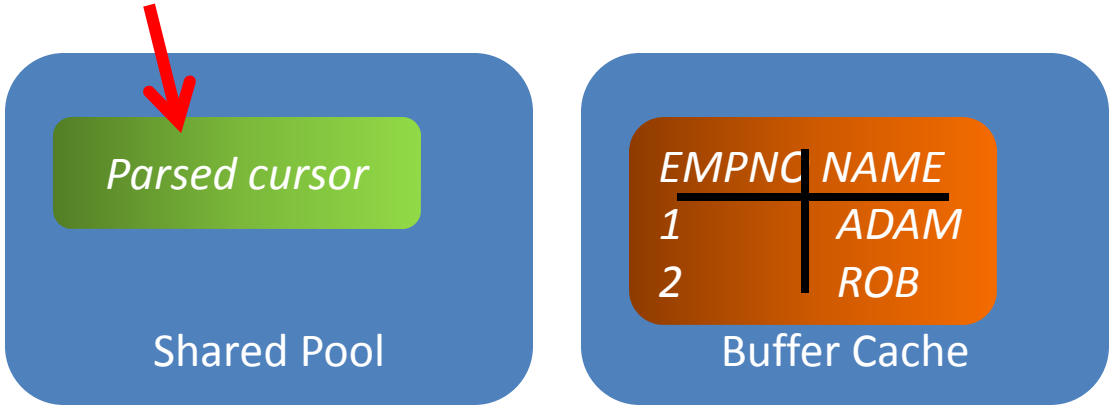
Database and Instance

```
update EMP
set NAME = 'ROB'
where EMPNO = 1
```



Dirty Buffer

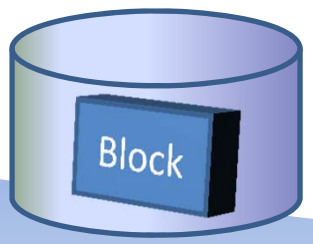
```
update EMP
set NAME = 'ROB'
where EMPNO = 1
```



This is known as a *dirty buffer* because the copy in the buffer cache is different from the block on the disk.

Host

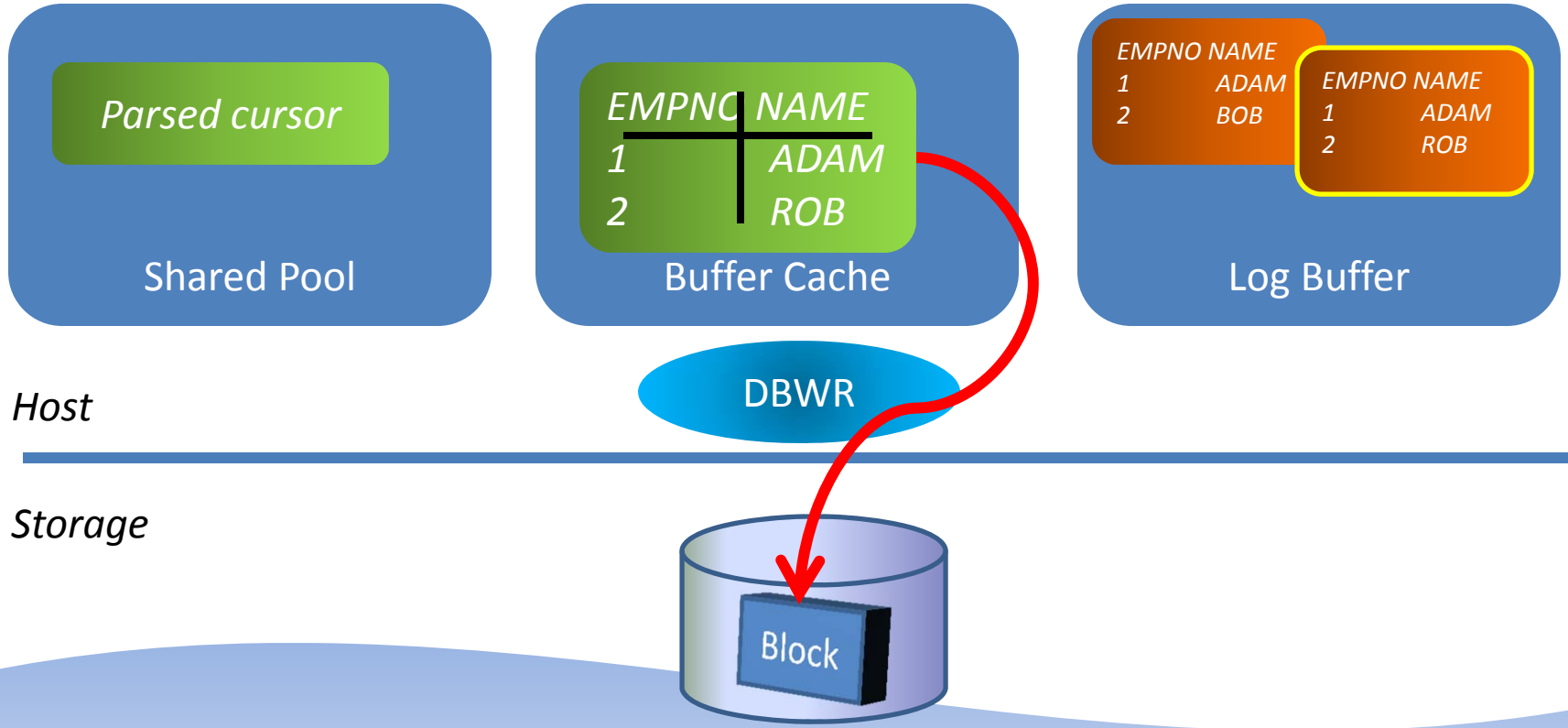
Storage



Buffer Writer

```
ALTER SYSTEM CHECKPOINT;
```

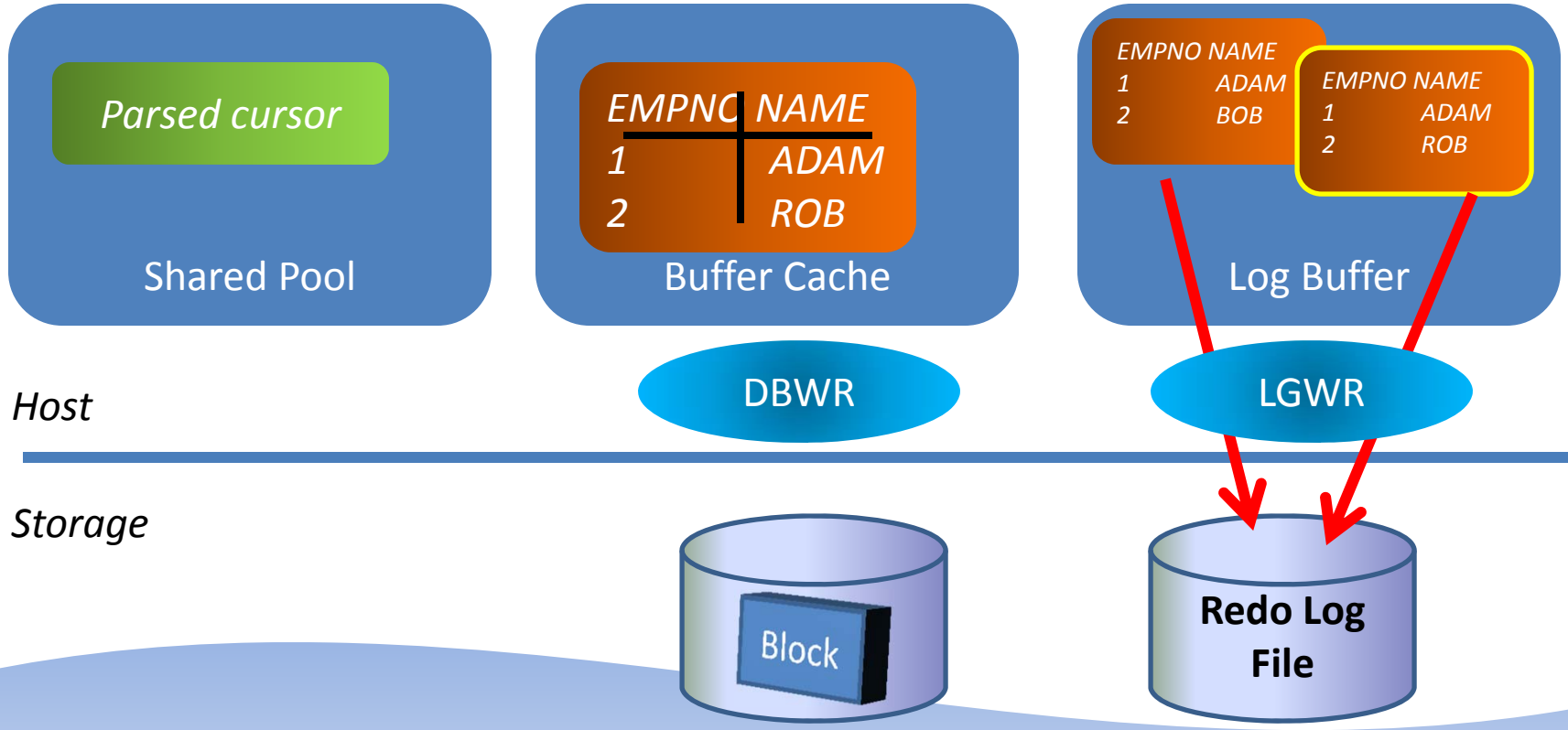
Process Database Buffer Writer (DBWR) writes dirty buffers to the disk.



Log Writer

```
COMMIT;
```

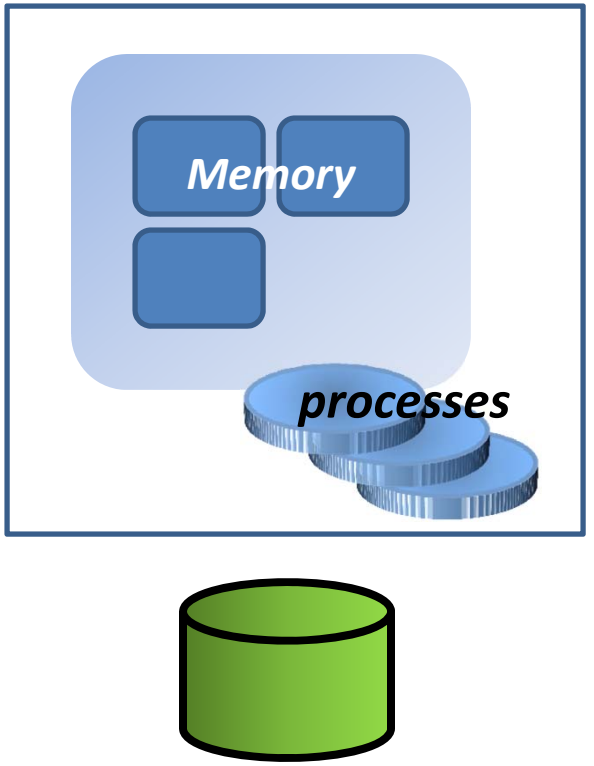
Process Log Writer (LGWR) writes the log buffers to the Online Redo Log File.



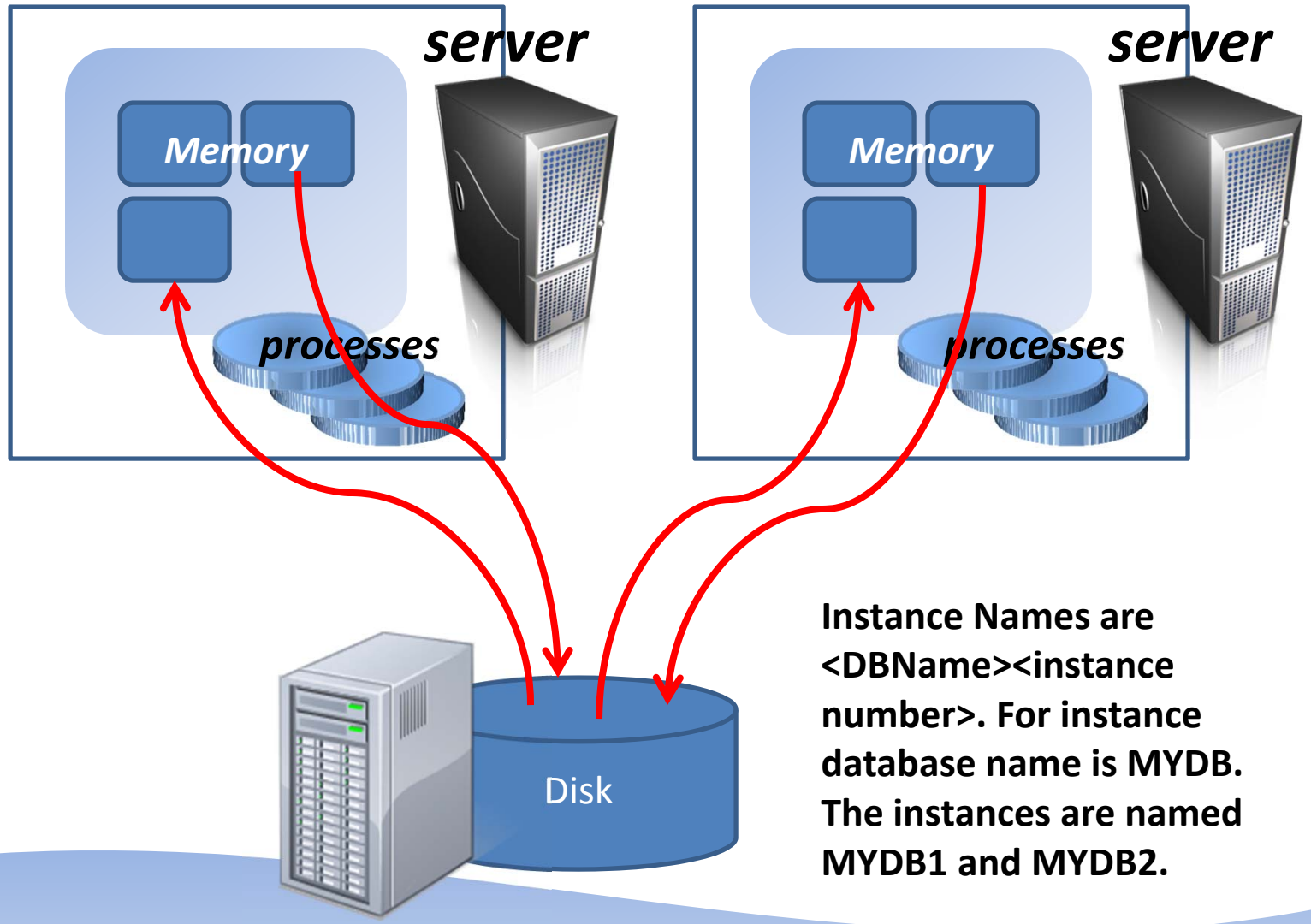
Database Instance

- Processes
  - Database Buffer Writer (DBWn)
  - Log Writer (LGWR)
  - Process Monitor (PMON)
  - System Monitor (SMON)
  - … and many more
- Memory Areas
  - Buffer Cache
  - Shared Pool
  - Large Pool

The processes and memory areas are known as an Oracle Instance

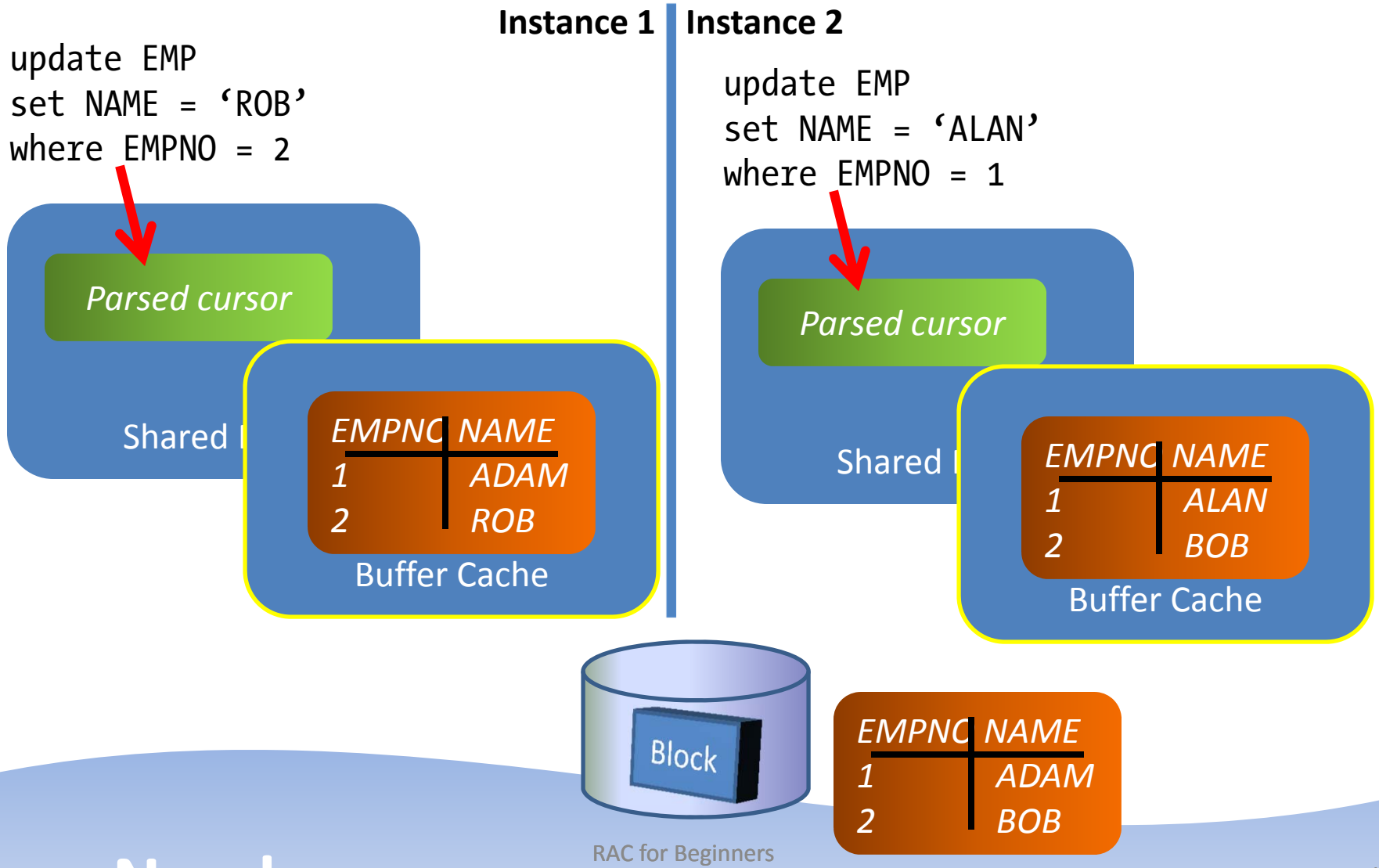


Instances



Instance Names are <DBName><instance number>. For instance database name is MYDB. The instances are named MYDB1 and MYDB2.

Multiple Buffer Caches



Buffer Caches

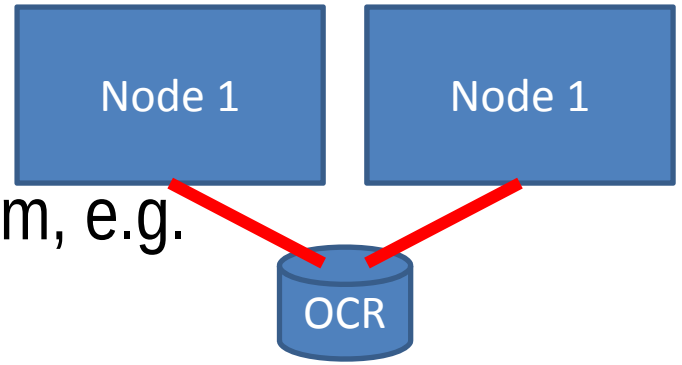
- Each Buffer Cache in RAC is unique
  - They are not replicated
  - They are not even synchronized
- This is not an “extended cache”, or “unified cache”.
- When one instance requests a block from the disk, it may come from
  - The disk
  - Or, another instance’s buffer cache
  - **Cache Fusion** is concept behind moving one buffer to the other

Interconnect

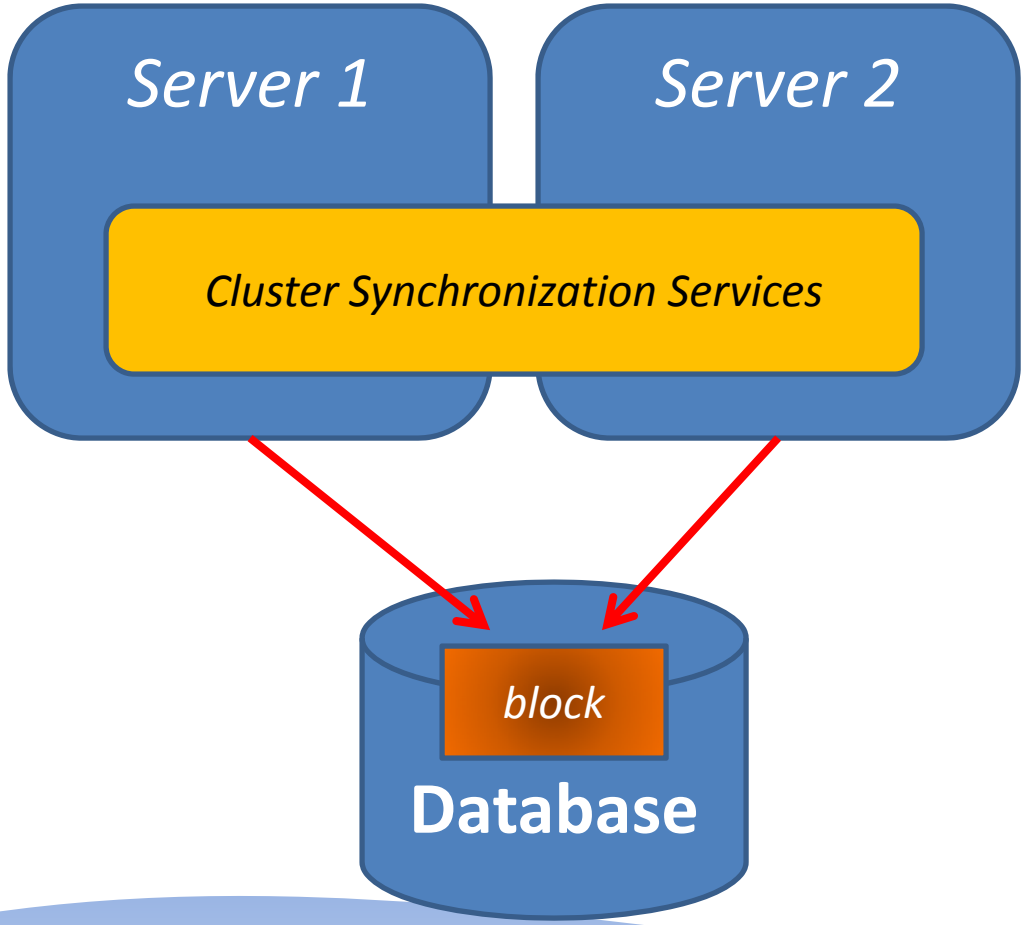
- Instance 1 gets a buffer from another
- The buffer is physically transported
- The medium of transport is known as **Interconnect**
- But that's not the only function of the Interconnect
- Nodes communicate to one another
  - To make sure they are alive
- But that brings up a vital question – how does one node know what other nodes are there?

Cluster Repository

- A special file holds the information on the members of the cluster
- This is called **Oracle Cluster Repository** (OCR)
- Must be on a shared storage
- Can be on a file in a cluster filesystem, e.g. Veritas
- In 11.2, should be on an ASM diskgroup.
- Pre-11.2: a shared raw device

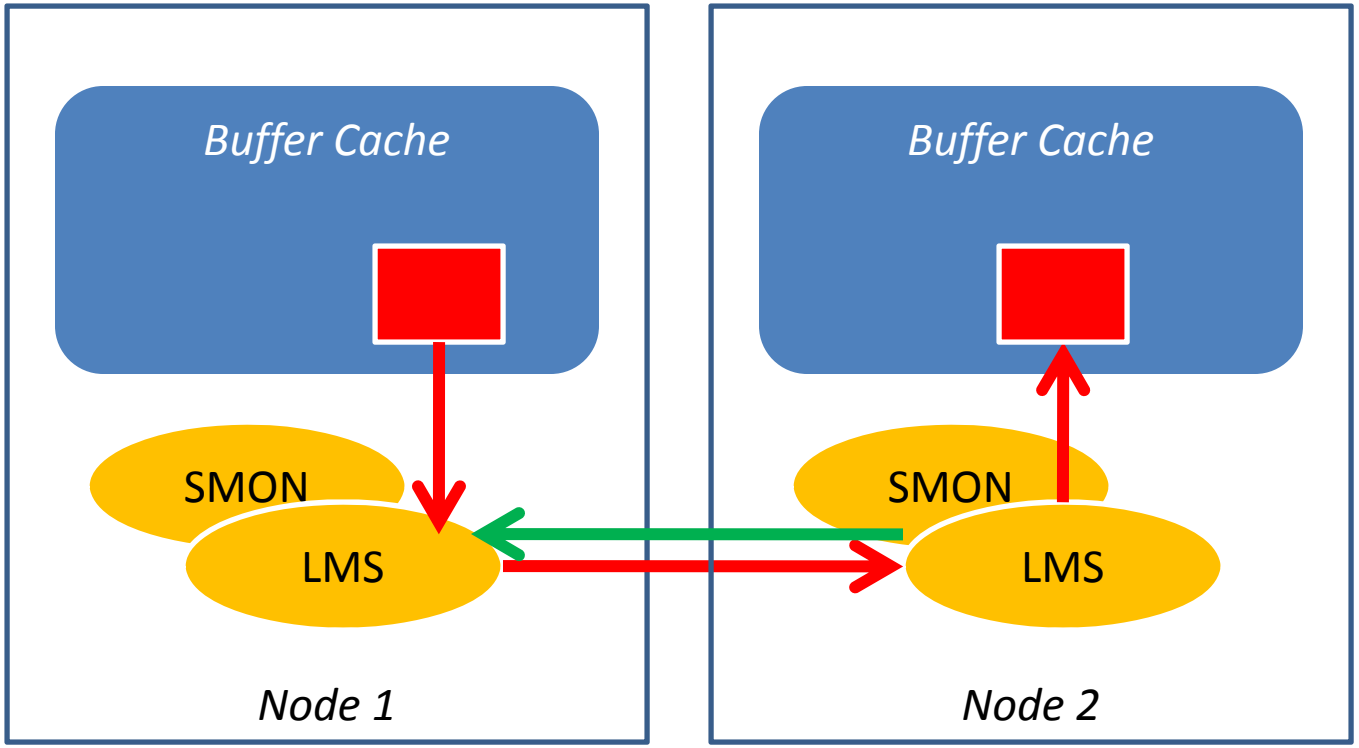


Cluster Operation



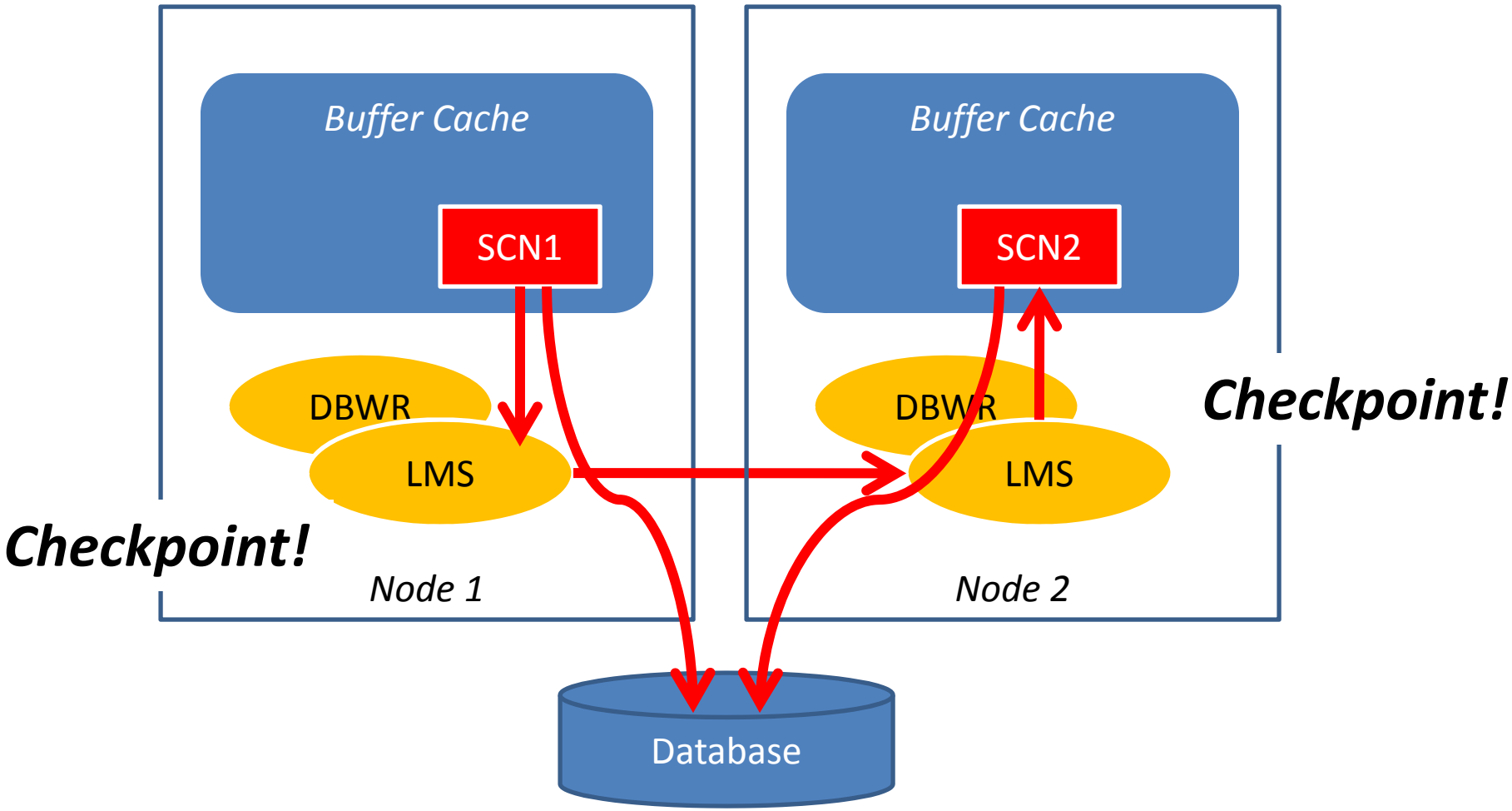
This is WRONG! There is no such thing (process, memory, etc.) that exists across all the nodes of the cluster

Cache Fusion



When node 2 wants a buffer, it sends a message to the other instance. The message is sent to the LMS (Lock Management Server) of the other instance. LMS then sends the buffer to the other instance. LMS is also called Global Cache Server (GCS).

Cluster Coordination



DBWR must get a lock on the database block before writing to the disk. This is called a Block Lock.

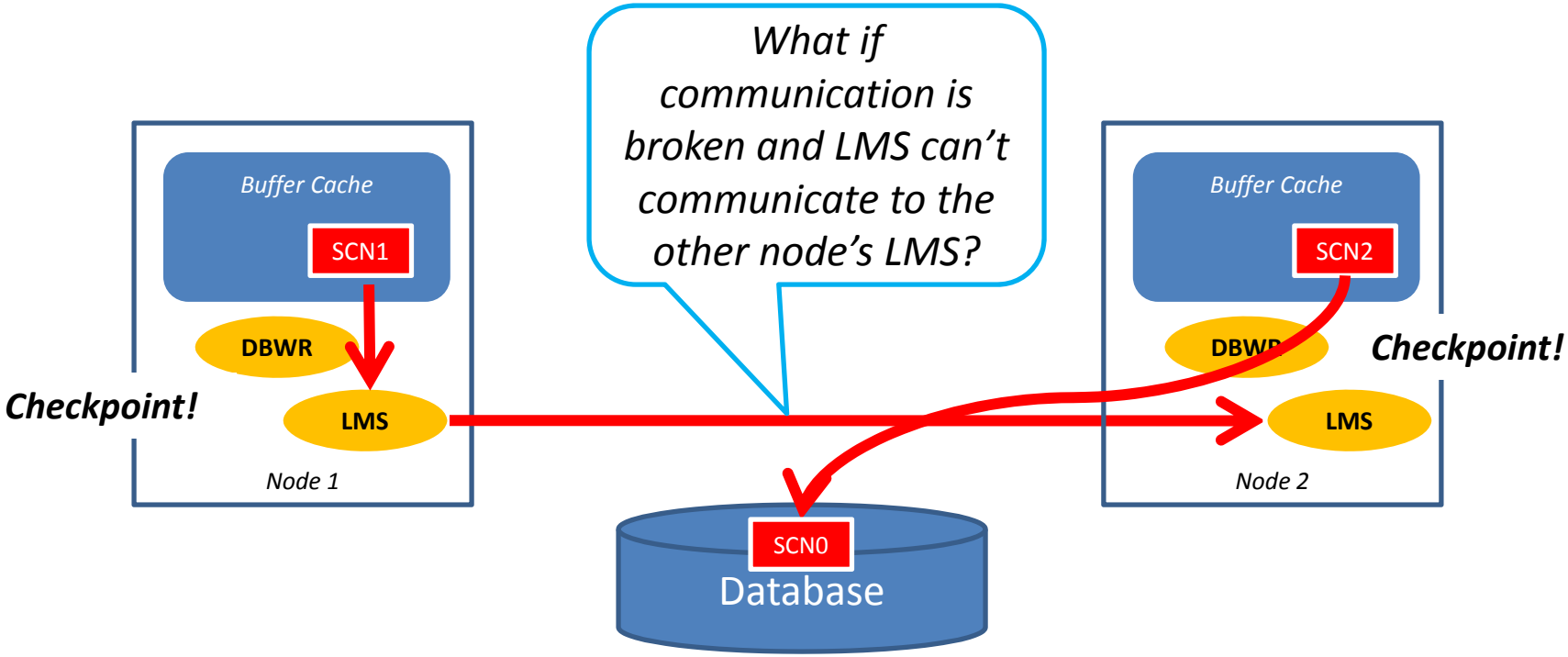
Block Lock

- Before modifying any buffer in the buffer cache, or writing the buffer to the disk,
- The instance must acquire a Exclusive Current **Block Lock** on the buffer
  - This is regardless of the various locks on the rows inside the block
- The request and response for the Block Lock are sent through the LMS processes
  - Over the Interconnect
- The request and response queues for a specific block is stored in a single node
  - Called the **Master Instance of the Block**

Global Structures

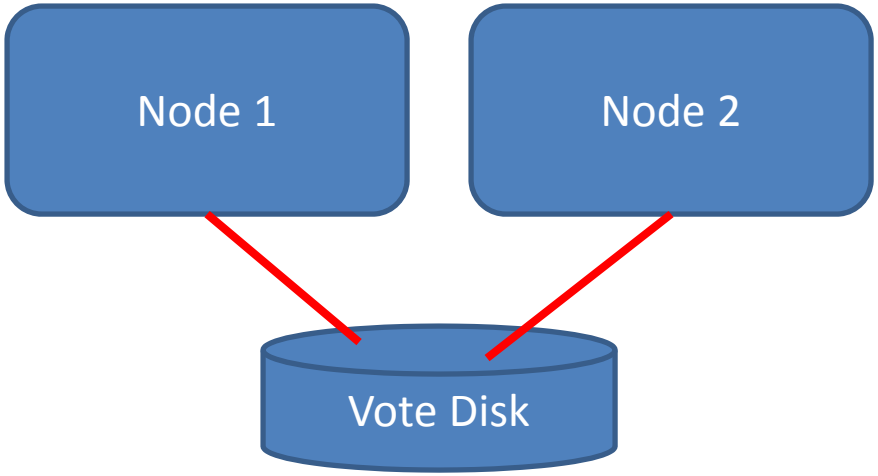
- Global Cache Service (GCS)
  - The overall mechanism to transfer buffers between nodes
- Global Enqueue Service (GES)
  - To coordinate locking between nodes
  - a.k.a. Distributed Lock Manager (DLM)
- Global Resource Directory (GRD)
  - The memory structure that records the location of the block lock for a specific block

Split Brain



In this case, Node 2 will think the other node is down and it's the only node in the cluster. Therefore it will not need to coordinate with any other node. Node 1 will also assume the same. They will write the buffer to the disk independently – leading to corruption. This is known and Split Brain Syndrome.

Voting Concept



In case of network failure between nodes, each node tries to grab the vote disk. Whoever grabs it becomes master and asks other nodes to join the cluster. Whoever does not join the cluster or does not get the message commits suicide so that it can't make updates.

Vote Disks

- Vote disks
  - can be files in cluster filesystem
  - can be raw devices
  - must be files in ASM diskgroup in 11.2
  - Don't contain anything of value
  - Preferably be more than 1 … odd number

Threads

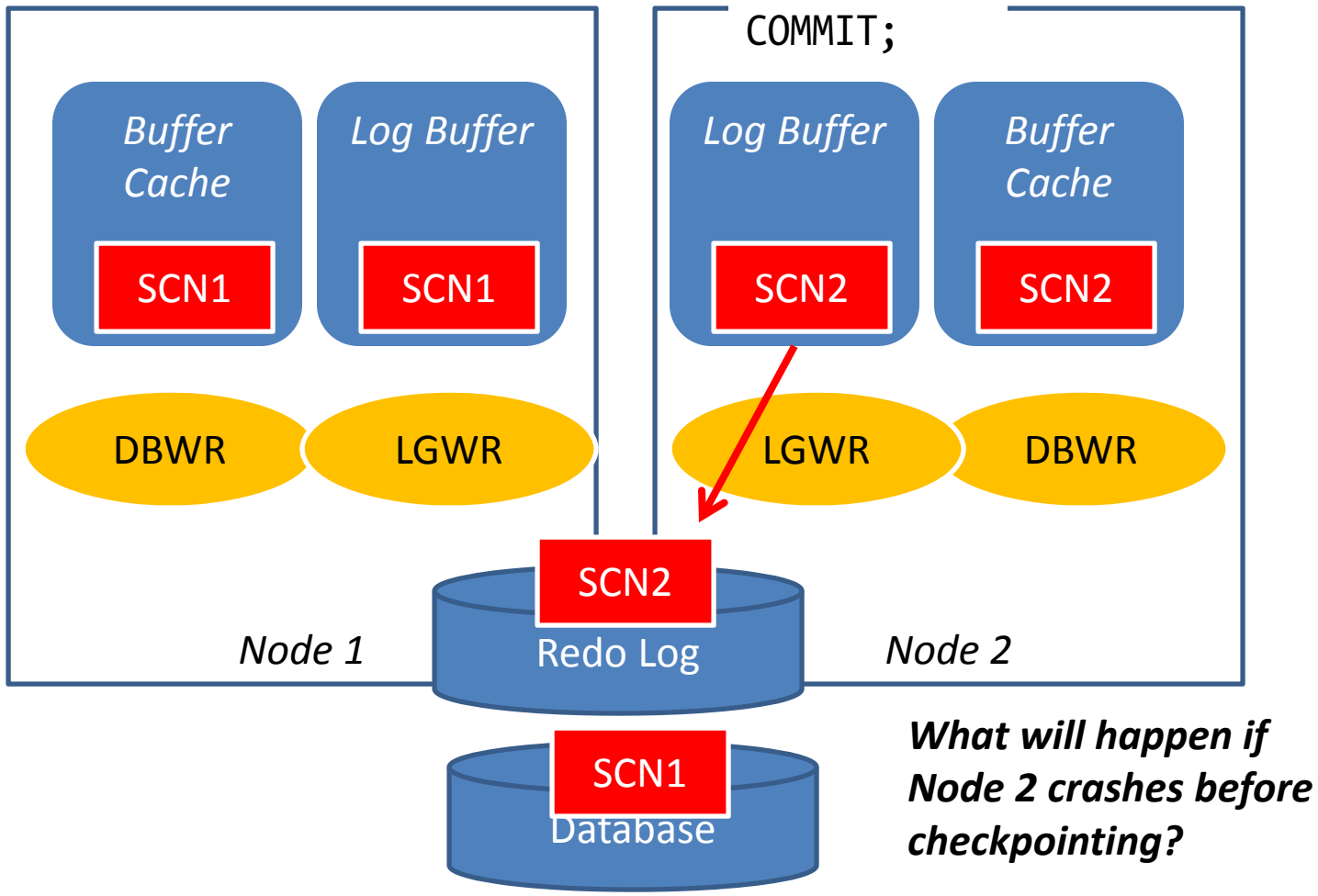
- Each instance has a log buffer
- Each instance has a LGWR
- Commits
  - Flush the Log Buffer to the Redo Log
  - Must be fast
  - Coordination of log buffer flushing (and locking) will only slow it down
- Therefore each instance has its own redo log.
- Each instance has a Thread, which corresponds to a Redo Log group

Init.ora File

```
Inst1.thread=1
Inst2.thread=2
```

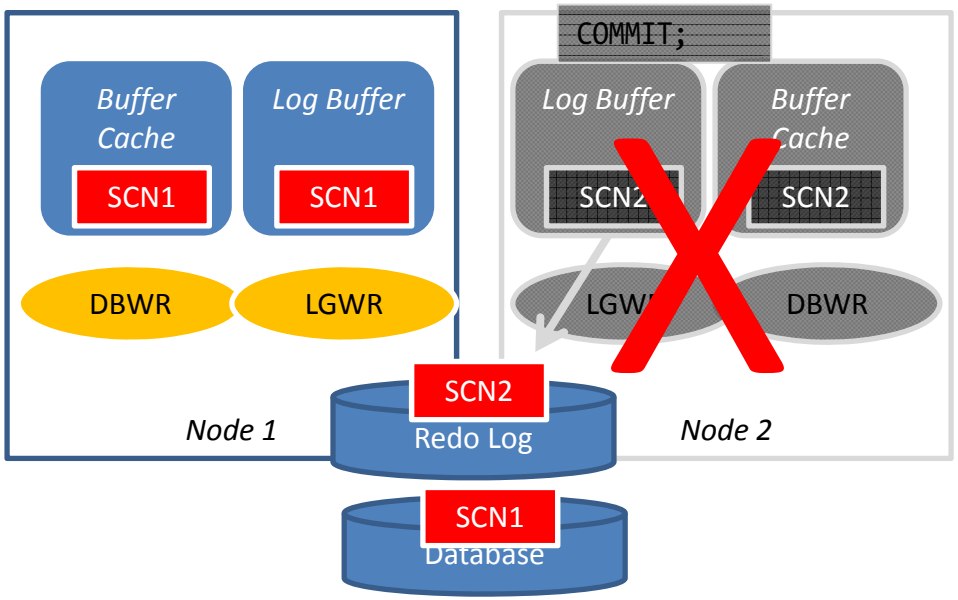
```
select thread#, group#
from v$log;

   THREAD#     GROUP#
---------- ----------
         1          1
         1          2
         2          3
         2          4
```

Instance Recovery



What will happen if Node 2 crashes before checkpointing?

RAC Instance Recovery

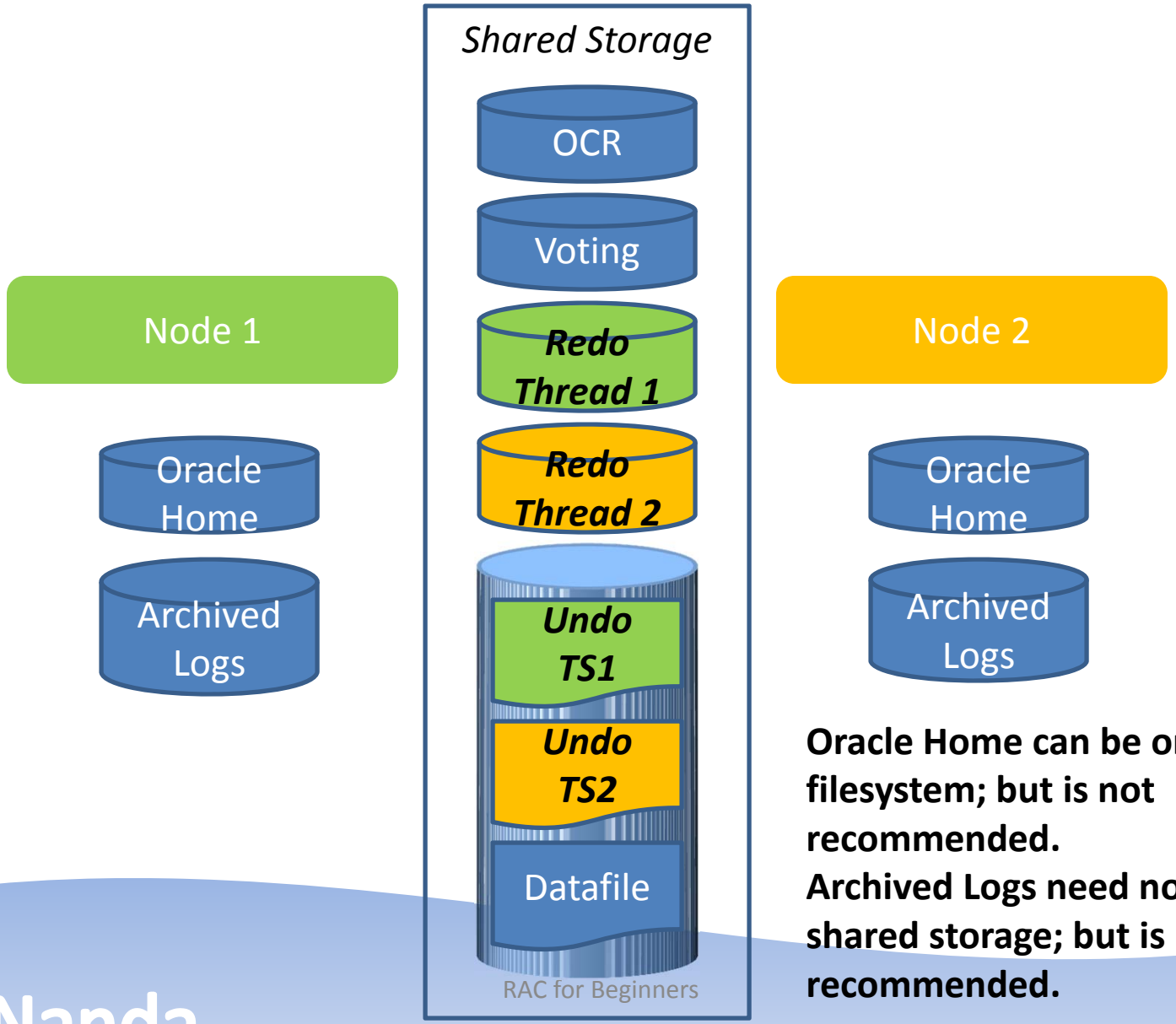


1. In this case, Instance 1 must recover instance 2 (which is dead now), by reading its Redo Log Files.
2. Therefore Redo Log Files must be common to both the nodes, just like data files

Undo Tablespace

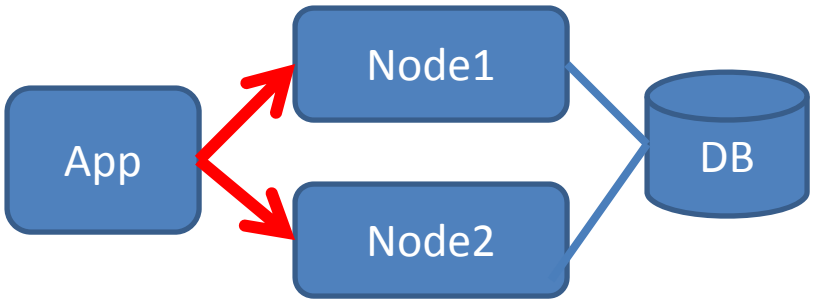
- When data modification occurs, the past image is stored in the undo segments
  - (aka Rollback Segments)
- Each instance has its own undo tablespace
- But the undo tablespace must be on a shared storage inside the database for all instances to access it.

RAC Storage Layout



Oracle Home can be on shared filesystem; but is not recommended.
Archived Logs need not be on shared storage; but is recommended.

RAC TNS Entry



1. If node1 is down, the client will automatically try node2, since it is the next in line.
2. By default the client will try node1 and node2 randomly
3. As long as one of the nodes is up, the client will be able to connect

TNSNAMES.ORA File Entry

```
CONN1 =
  (DESCRIPTION =
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = node1) (PORT = 1521)
    )
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = node2) (PORT = 1521)
    )
    (CONNECT_DATA =
      (SERVICE_NAME = SRV1)
    )
  )
```

Service Name

- It's another "gateway" into the database
- If service name SRV1 is defined on node1 and not on node2, the client will not connect to node2.
- You can control which application connects to which node by defining the services
- SRVCTL is the command to create service

TNSNAMES.ORA File Entry

```
CONN1 =
  (DESCRIPTION =
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = node1) (PORT = 1521)
    )
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = node2) (PORT = 1521)
    )
    (CONNECT_DATA =
      (SERVICE_NAME = SRV1)
    )
  )
```

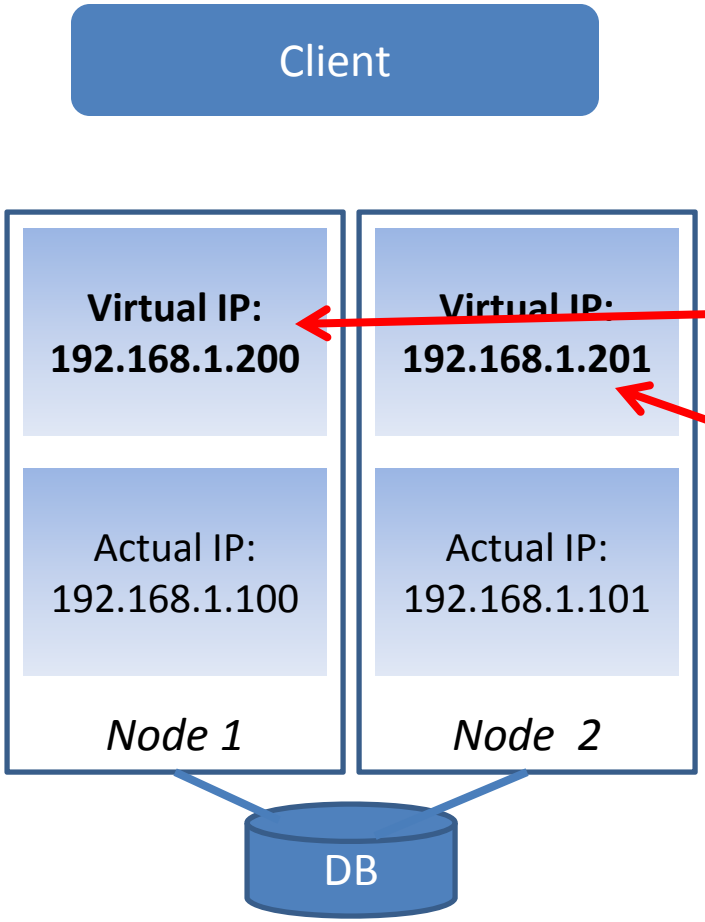
The Problem with TCP

- If node1 is down, the client will wait until the TCP/IP timeout, which could be as much as 30 seconds.
- The failover will be delayed if the node1 is actually down

TNSNAMES.ORA File Entry

```
CONN1 =
  (DESCRIPTION =
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = node1) (PORT = 1521)
    )
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = node2) (PORT = 1521)
    )
    (CONNECT_DATA =
      (SERVICE_NAME = SRV1)
    )
  )
```

Virtual IP



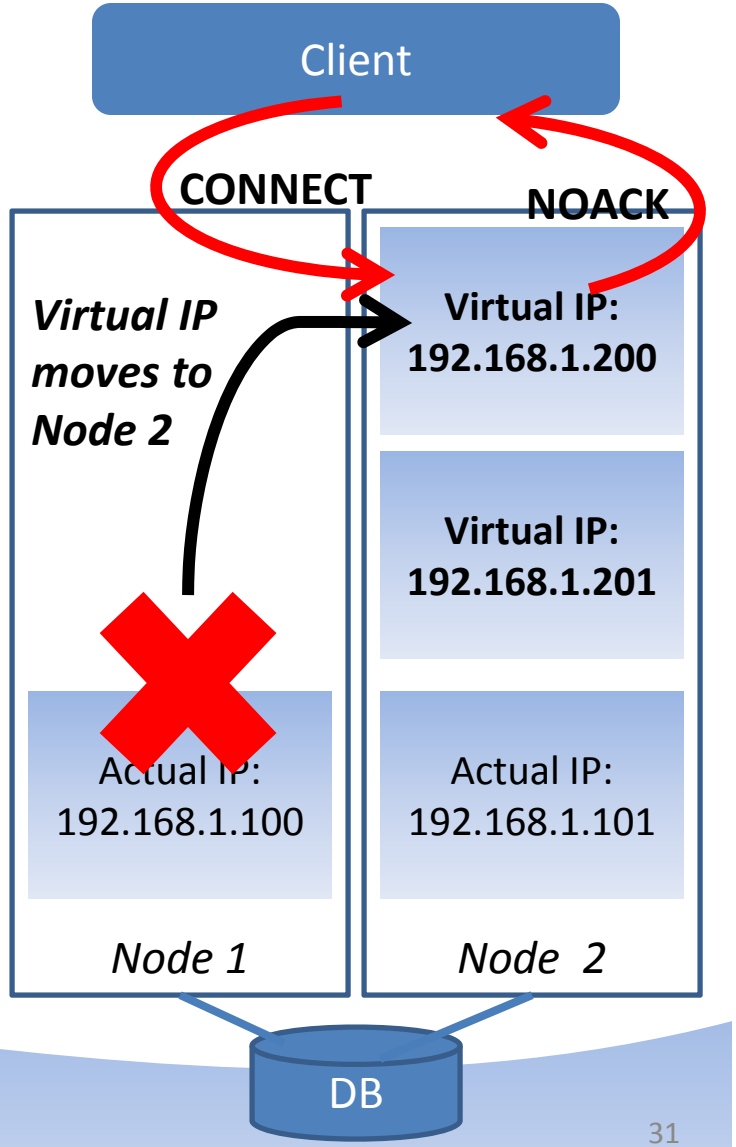
Normally, Virtual IP of the node is up on that node

TNSNAMES.ORA File Entry

```
CONN1 =
  (DESCRIPTION =
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = vip1) (PORT = 1521)
    )
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = vip2) (PORT = 1521)
    )
    (CONNECT_DATA =
      (SERVICE_NAME = SRV1)
    )
  )
```

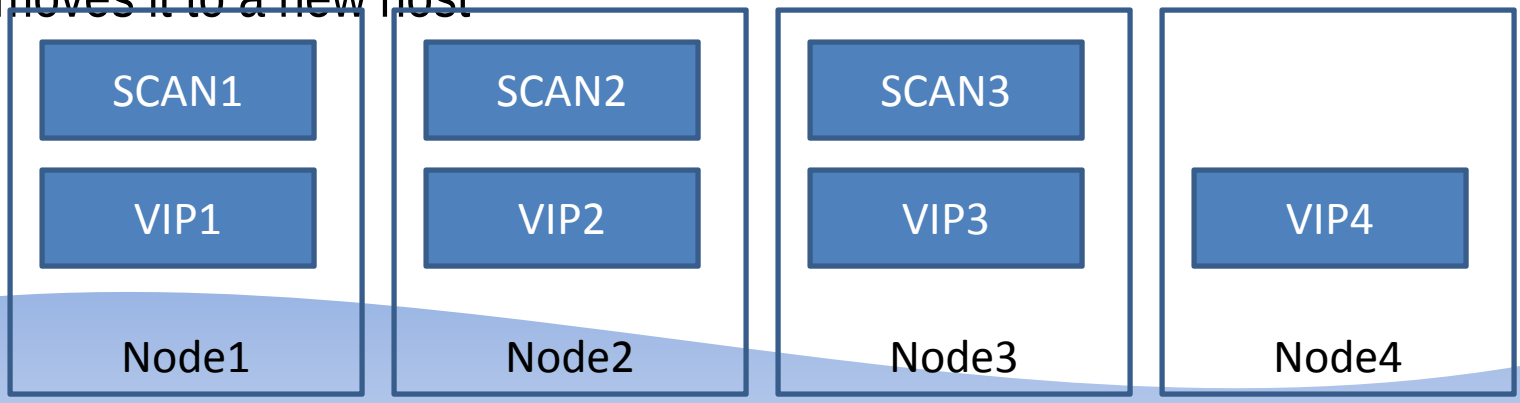
Virtual IP under Failure

1. When Node 1 fails, VIP of node 1 comes up on node 2
2. When client attempts to connect to the VIP1, it goes to node2
3. But it does not connect to node2
4. Instead the CRS sends a NOACK signal (i.e. not alive signal immediately
5. The client then tries the next host in the list
6. This way, the failover does not have to wait for TCP timeout

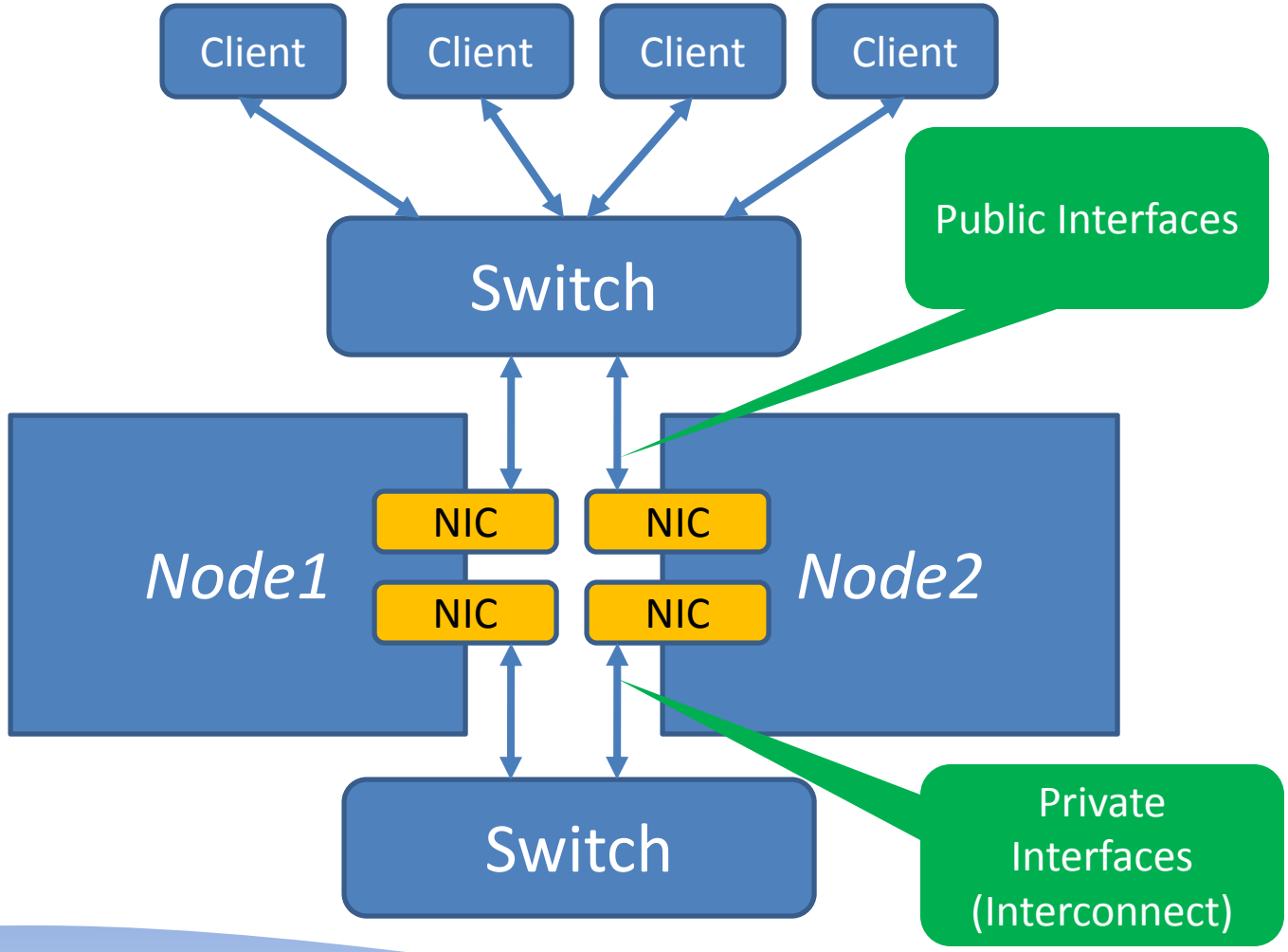


SCAN IP

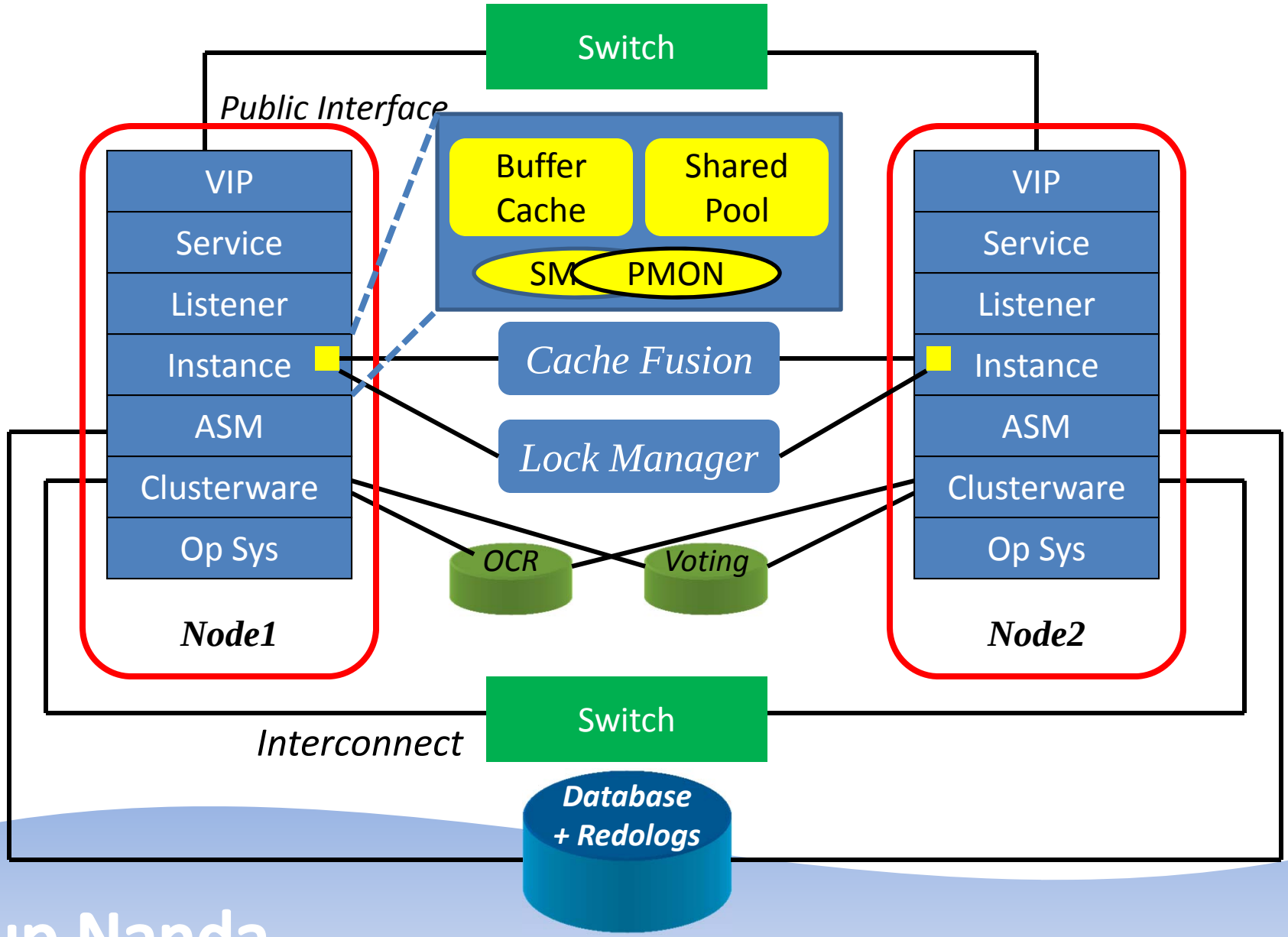
- When you add a node to the cluster
  - You have to update the TNSNAMES.ORA file
- In 11.2 Single Client Access Name (SCAN) solves the problem
  - You can define a hostname (e.g. myscan) which resolves to three different IPs (called SCAN IPs)
  - The clients connect to the SCAN hostname
  - The listener directs to one of the instances
  - When the node hosting a SCAN IP crashes, Clusterware automatically moves it to a new host



RAC Network Layout



Putting it all together



Common Questions

- *Q: Can we have nodes with different OS'es in the same cluster, e.g. Linux and Windows?*
  - A: No. They must be same, e.g. Linux on both
- *Q: Can we have different number of CPUs the nodes of a cluster?*
  - A: Yes, but that is not advisable. All nodes should be similar for equal load balancing

How Do I Learn

- **Virtual Box** (100% free)
  - Allows you to create two "virtual" hosts on a single machine, could be your laptop
  - Install Linux on each host
    - http://www.oracle.com/technetwork/server-storage/virtualbox/downloads/index.html
  - Eliminates the shared storage problem
  - Create 11gR2 RAC on Virtual Box
    - http://www.oracle-base.com/articles/11g/OracleDB11gR2RACInstallationOnOEL5UsingVirtualBox.php
  - Virtual Box Help
    - http://forums.virtualbox.org/

Thank You!

My Blog: **arup.blogspot.com**
My Tweeter: **arupnanda**