
RAC War Stories

Caleb Small, BSc, ISP

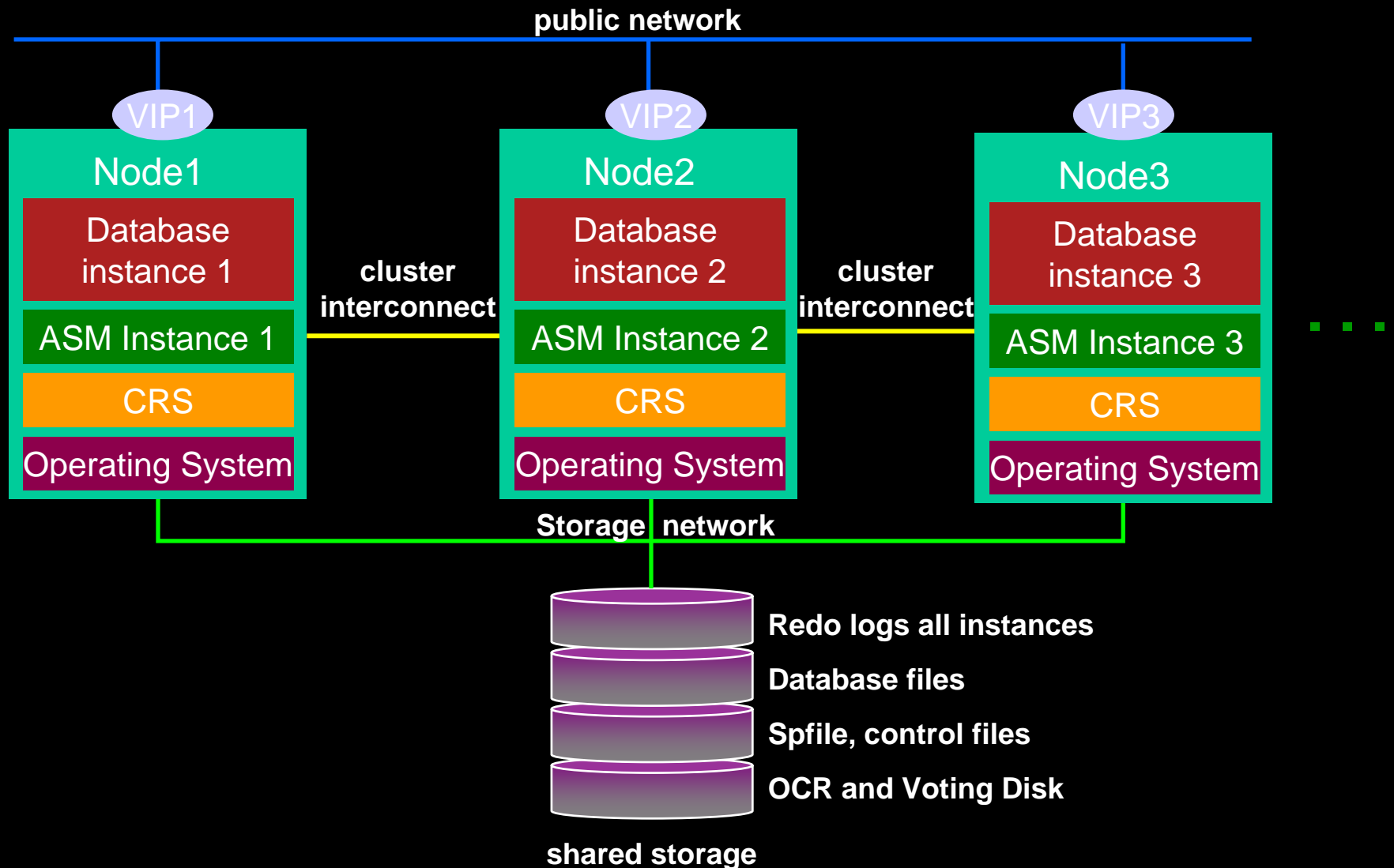
Caleb@Caleb.com

www.Caleb.com/dba

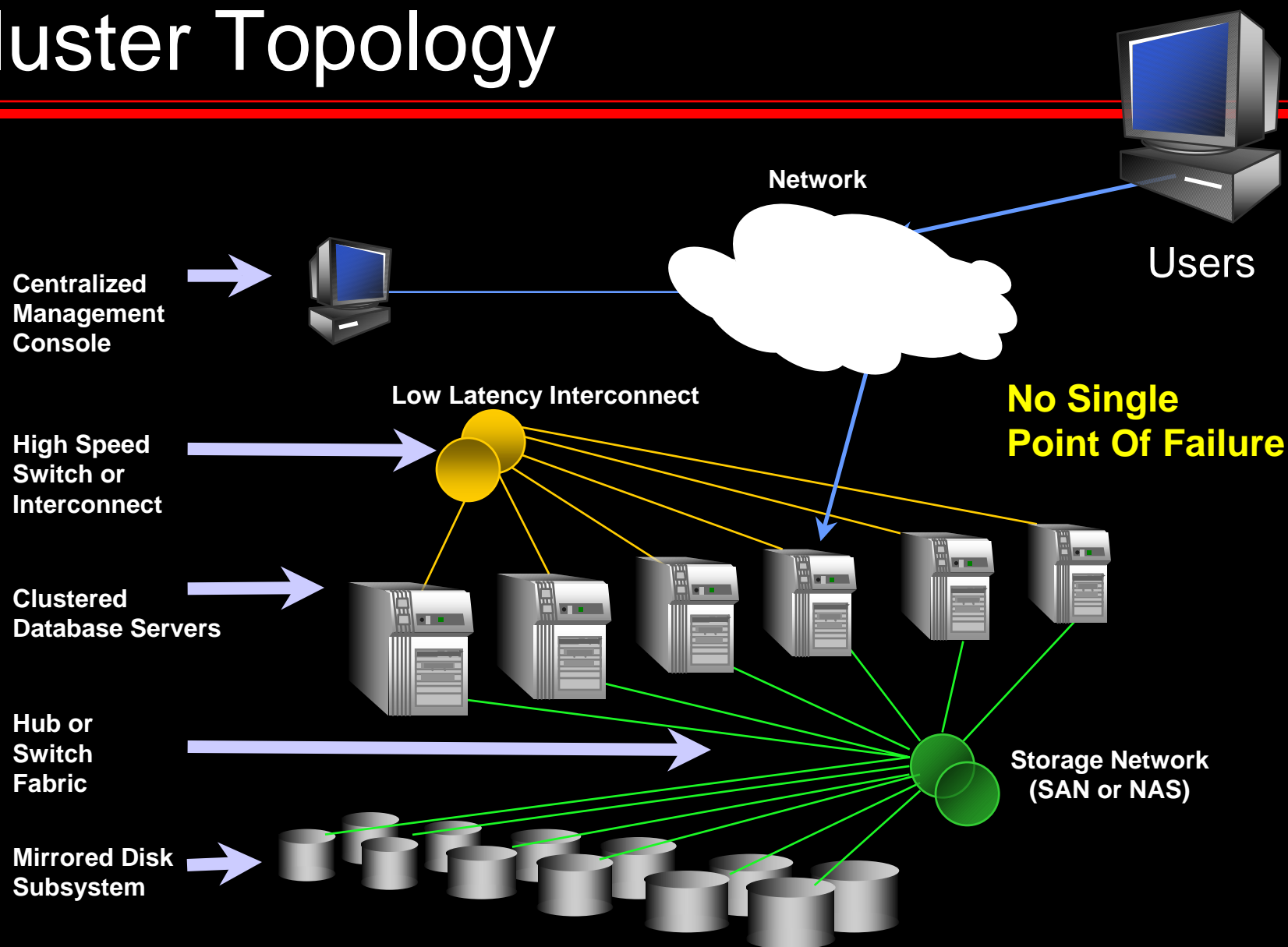
Most Common Features

- Red Hat/Oracle Linux 4 – 64bit
- 2 to 4 node RAC clusters
- Automatic RMAN backups
- Data Guard disaster recovery
- Grid Control management
- Data Pump
- Partitioning
- Transportable tablespace

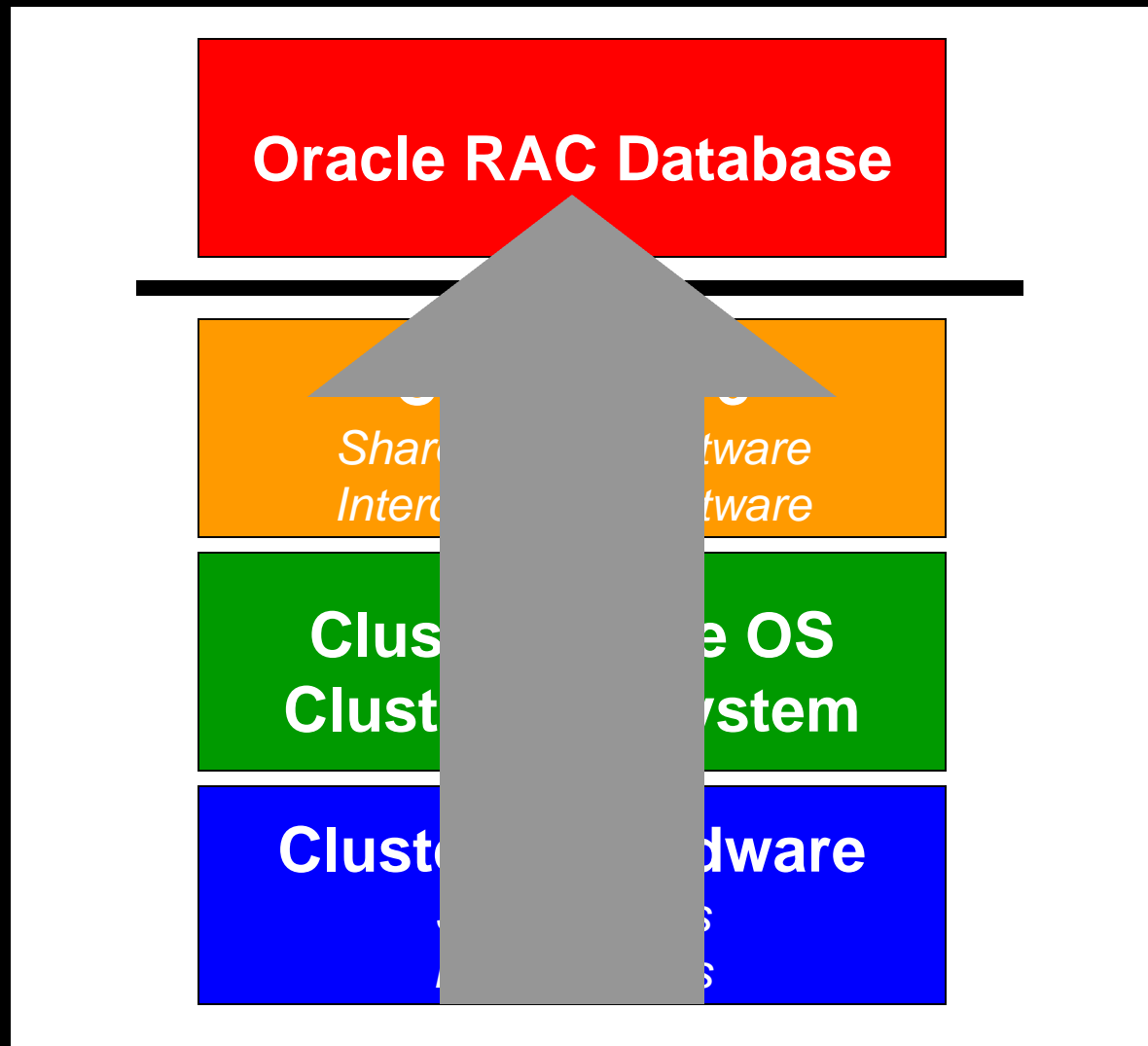
Cluster Topology



Cluster Topology



The "Black Line"



Shared Storage Decisions

| Storage Option | OCR and Voting Disks | Oracle Software Installation | Database | Recovery |
|------------------------------------|----------------------|------------------------------|----------|----------|
| Automatic Storage Management (ASM) | No | No | Yes | Yes |
| OCFS | Yes | No | Yes | Yes |
| OCFS2 | Yes | Yes | Yes | Yes |
| GPFS* for Linux on IBM POWER PC) | Yes | Yes | Yes | Yes |
| Local Storage | No | Yes | No | No |
| NFS File System | Yes | Yes | Yes | Yes |
| Shared Raw Partitions | Yes | No | Yes | No |

* IBM General Parallel File System. Other CFS supported include DBE/AC (Veritas) and Tru64 CFS

ASM Limitations

- ASM does not provide:
 - Clusterware files *
 - ASM Spfile *
 - Alert & log files
 - Script file
 - Staging area
 - Other misc files

*** Requires cluster file system**

Secondary Cluster File Solution

- Oracle recommends shared raw
- Individual mini-LUNs on SAN
- Block vs Character devices
 - 2.4 vs 2.6 kernel
 - rawdevices vs O_DIRECT
- UDEV and device persistence
 - UDEV rule
 - LUN probing script
- Not suitable for a general purpose file area (eg. log files, scripts)

Shared Storage Challenges

- Installing HBAs and drivers
- Multipathing
- Presentation to Linux
- Device Persistence
- Primary partitions (1 MB offset)
- ASM Lib
 - Mark disks on primary, discover on secondary nodes
- OCR/Vote raw devices

Network Challenges

- Installing hardware
- Vendor specific drivers
- Bonding
- Non-routable “public” IPs
- Messing with parameters
 - MTU=9000 aka Jumbo Frames
- Kernel Parameters for failover

```
net.ipv4.tcp_keepalive_time  
net.ipv4.tcp_keepalive_intvl  
net.ipv4.tcp_retries2  
net.ipv4.tcp_syn_retries
```

Calculate Kernel Parameters

- Start with:

```
max_sga_size  
processes
```

- Calculate shared memory parameters

```
SHMMAX, SHMMNI, SHMALL  
*      soft      memlock 3401728  
*      hard      memlock 3401728
```

- Calculate semaphore values

```
SEMMSL = PROCESSES + 10  
SEMMNS = SEMMSL * SEMMNI  
SEMOPM = SEMMSL  
SEMMNI  = 128 or greater
```

Huge Pages Memory Optimization

- Configure Huge Page Pool

```
$ sysctl -w vm.nr_hugepages=1661
vm.nr_hugepages = 1661
```

```
$ grep Huge /proc/meminfo
HugePages_Total: 1661
HugePages_Free: 1661
Hugepagesize: 2048 kB
```

- Verify SGA withdrawal

```
$ ipcs -m
```

```
----- Shared Memory Segments -----
key          shmid      owner      perms      bytes
0x5451cfe8  65538     oracle     600        132120576
0x59e98e98  98307     oracle     600        1612709888
```

Usual Oracle pre-install

- Create oracle user and groups (same id)
- Install cvuqdisk for the cluster verify utility
- Modify kernel parameters for Oracle
- Configure hangcheck timer module and nscd
- Create oracle install directory and set permissions
- Set permissions on shared storage dirs
- Install environment files for user oracle
- Modify the /etc/pam.d/login file for security
- Configure SSH

Order of Installation - Oracle

- Install Clusterware into \$CRS_HOME
- Install ASM software only into \$ASM_HOME
- Install Database software only into \$DB_HOME
- Patch Clusterware
- Patch ASM software
- Patch Database software
- Repeat for patch#5679560, and quarterly update
- Create a Listener in the \$DB_HOME directory
- Create an ASM instance - register with Listener
- Create a Database - register with Listener

Oracle Clusterware Install

- OUI is cluster aware
- Installed on primary node only
- Binaries pushed out to other nodes
- Shared storage for OCR and Vote
- Will fail if “public” IPs are non-routable
 - Manual intervention – run vipca
- Will fail if any reqd packages missing
 - Check carefully before starting
- May fail with OraInventory permissions
 - `chmod 644 /etc/oraInst.loc`

Oracle DB Binaries Install

- Install for ASM and DB
- Installed on primary node only
- Binaries pushed out to other nodes
- Separate (2nd and 3rd) ORACLE_HOME
- Do not create ASM instance or sample Database
- Be *very* patient during last few screens!

Apply 10.2.0.3 Patch

- Stop all services
- Run OUI on primary node only
- Patch CRS first
- Root script fails, fix permissions

```
chmod 755 $CRS_HOME/lib/libclntsh.so.10.1
chmod 755 $CRS_HOME/lib32/libclntsh.so.10.1
```
- Re-run root script
- Patch ASM
- Patch DB
- Upgrade any existing databases

Apply Patch #5679560

- Stop all services
- Execute scripts on each node
- Rolling upgrade is possible
- Patch CRS and ASM homes
- Repeat for DB home
- Post patch steps

Apply Critical Patch Updates

- April CPU pre-requisites:
 - Upgrade Opatch to 10.2.0.3.1
 - Bug patch 5240469 – x86_64
 - Doesn't fix the bug
 - Hand edit \$OH/bin/genoccish
`LD_FLAGS_ARCH=-m32`
- Run scripts on each node
- Patch each DB home, not CRS
- Almost a rolling upgrade...
- Run DB upgrade scripts on one instance only

Configure Oracle Net

- Run NETCA to create listener prior to creating ASM or RAC database
- NETCA is cluster aware and will create listener on all nodes
- Balance of Oracle Net configuration (eg. DB connect string and services) will be done by DBCA
- Special configurations for load balance and failover (tnsnames)

Create ASM Instance

- OH and PATH must point to ASM
- Run dbca
- Create ASM instance
- Create disk groups
- Change disk discovery string:
/dev/oracleasm/disks/*

Create RAC Database

- DBCA is cluster aware
- Database is created on shared storage
- Includes spfile
- Each instance has it's own:
 - Undo tablespace
 - Thread of redo log groups
 - Addn'l space in SYSAUX

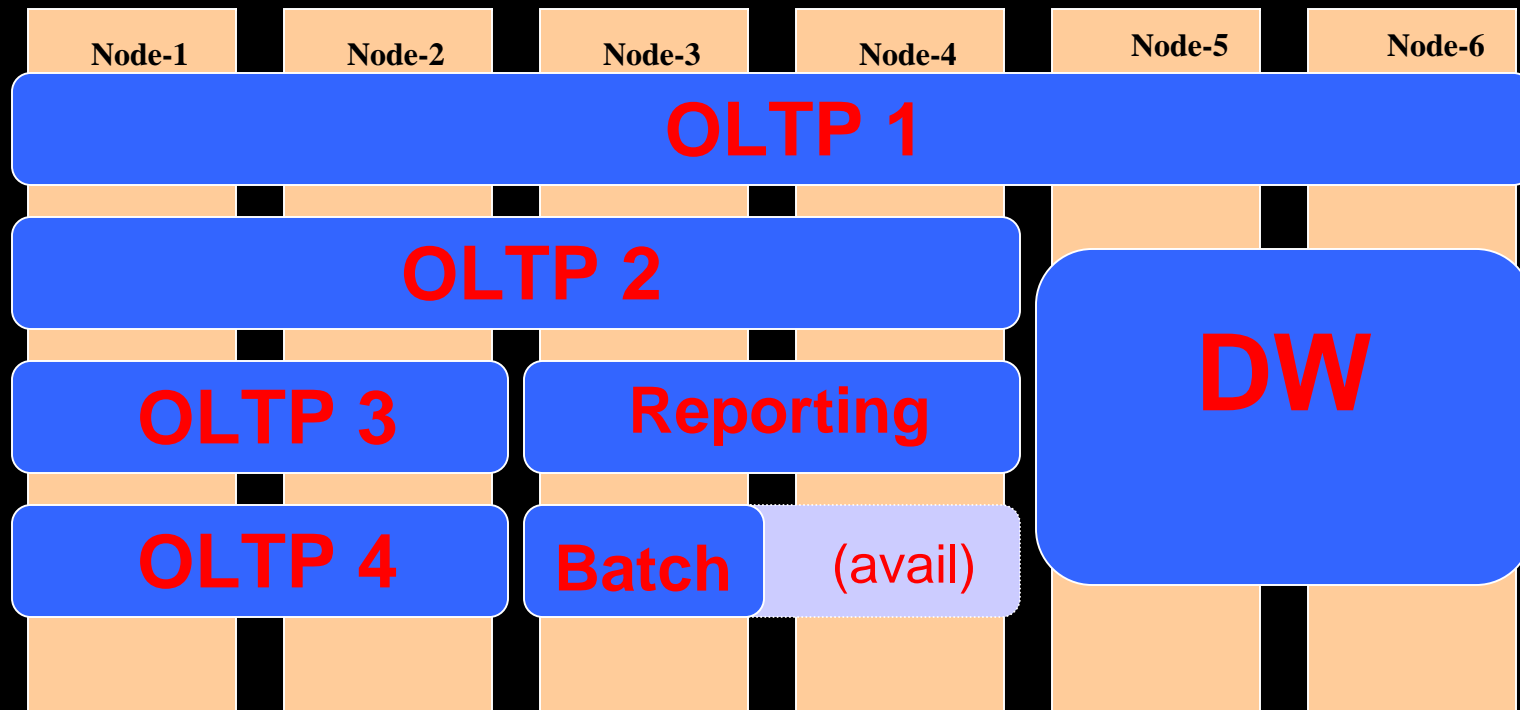
Post Installation

- Backup OCR and Voting Disk
- Configure OCR auto backup
- Verify recovery settings
- Set up RMAN automatic backup
- Verify async I/O
- Enable block checking
- Consider flashback database

Services

- Applications with similar needs connect to a Service rather than a database or instance
- One Service can be spread across multiple 'preferred' and 'available' instances
- Multiple Services (eg. oltp, batch, reports) can be defined and mapped to different instances
- When an instance fails, the Service is not necessarily interrupted
- Services are required to take advantage of High Availability features

Services

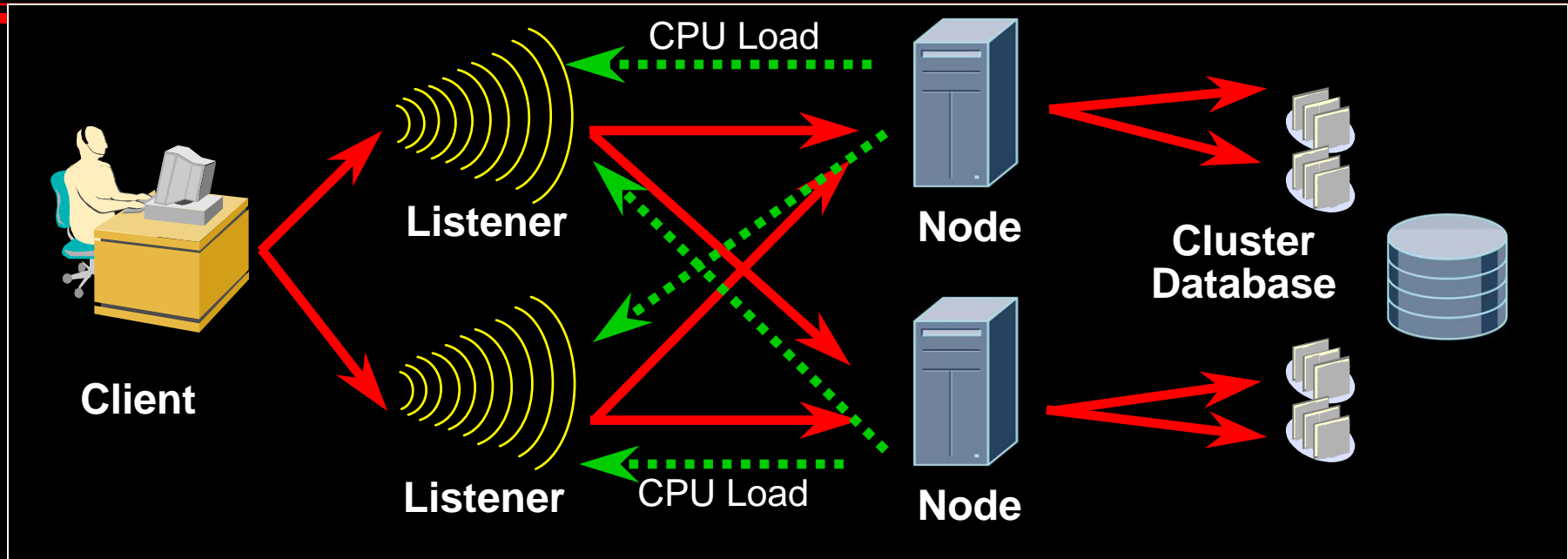


Services can be mapped to Resource Plans

Load Balancing & Failover

- Connection Load Balancing
- Transparent Application Failover (TAF)
- Fast Application Notification (FAN)
- Fast Connection Failover (FCF)
- Load Balancing Advisory
- Runtime Connection Load Balancing
- Database Services
- Adaptive Parallel Query

Connection Load Balancing



- Databases register with listeners when started (PMON)
- Nodes report CPU usage back to registered listeners (SMON)
- Client randomly distributes connections across listeners
- Listener chooses least used node and redirects connection
- Supports both Shared Server and Dedicated Server Configurations
- Uses built-in Load Balancing Advisory

Transparent Application Failover

- Defined as part of a service
- If an instance fails, connection is re-established to a surviving instance
- Session continues...
- Queries resume from last row returned
- Transactions are rolled back
- Session state is reset to default (eg. alter session, variables, etc)

Fast Application Notification - FAN

- Method by which applications are notified immediately of cluster changes (eg. node UP/DOWN events, workload)
- Applications can react immediately
- Part of High Availability Framework
- Uses Advanced Queuing (AQ) and Oracle Notification Server (ONS)

Fast Application Notification - FAN

FAN can be utilized in three ways:

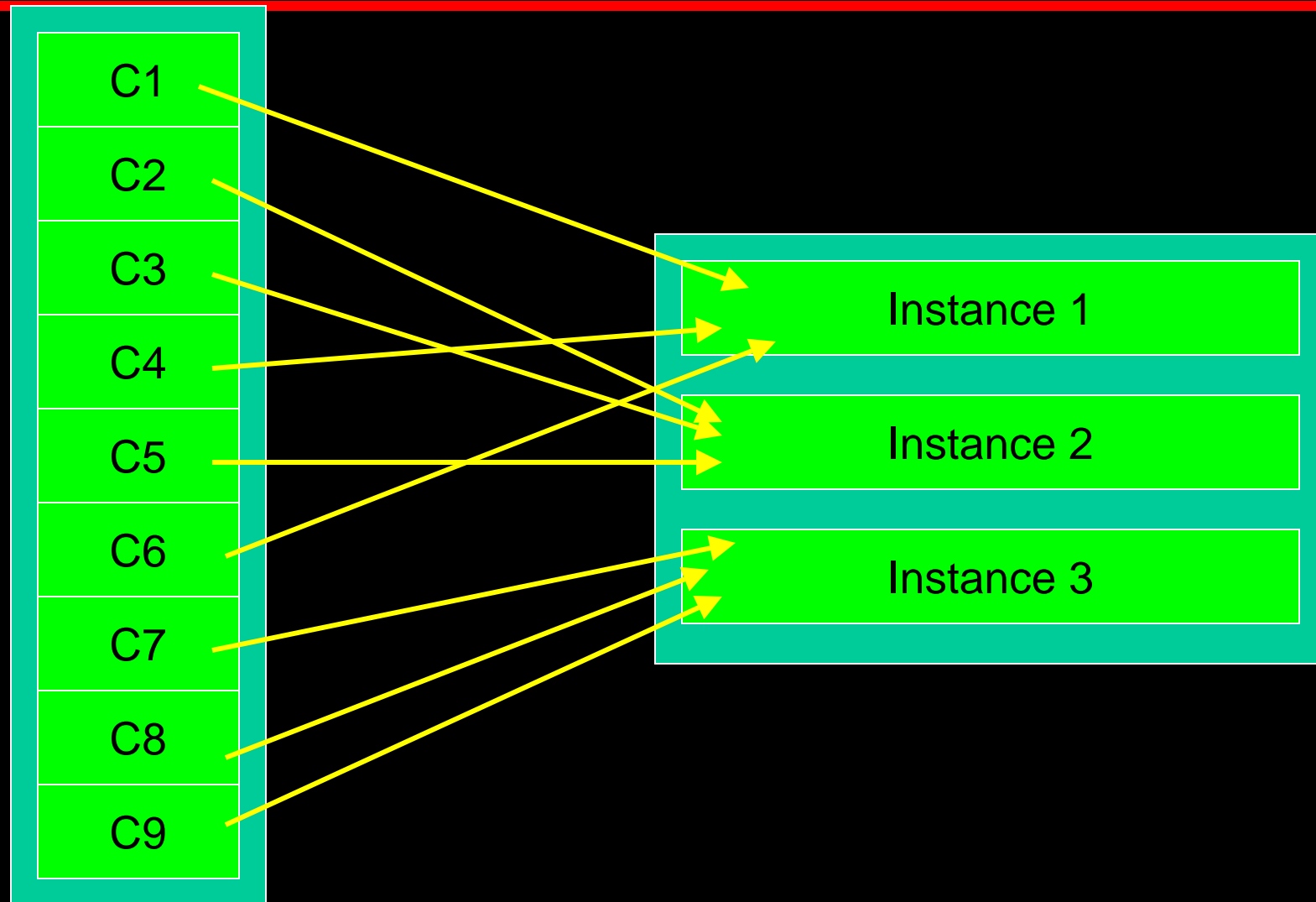
1. Use FCF client - JDBC, ODP.NET, OCI
2. Server side callouts
3. Application can subscribe to FAN events using the ONS API.

Also integrated with Listener and Connection Manager at the Net Services Layer

Fast Connection Failover - FCF

- Subscriber of FAN events
- Built-in to:
 - JDBC (thick & thin)
 - ODP.NET
 - OCI
- Cleans up connections when failures occur
- Distributes work requests across available instances

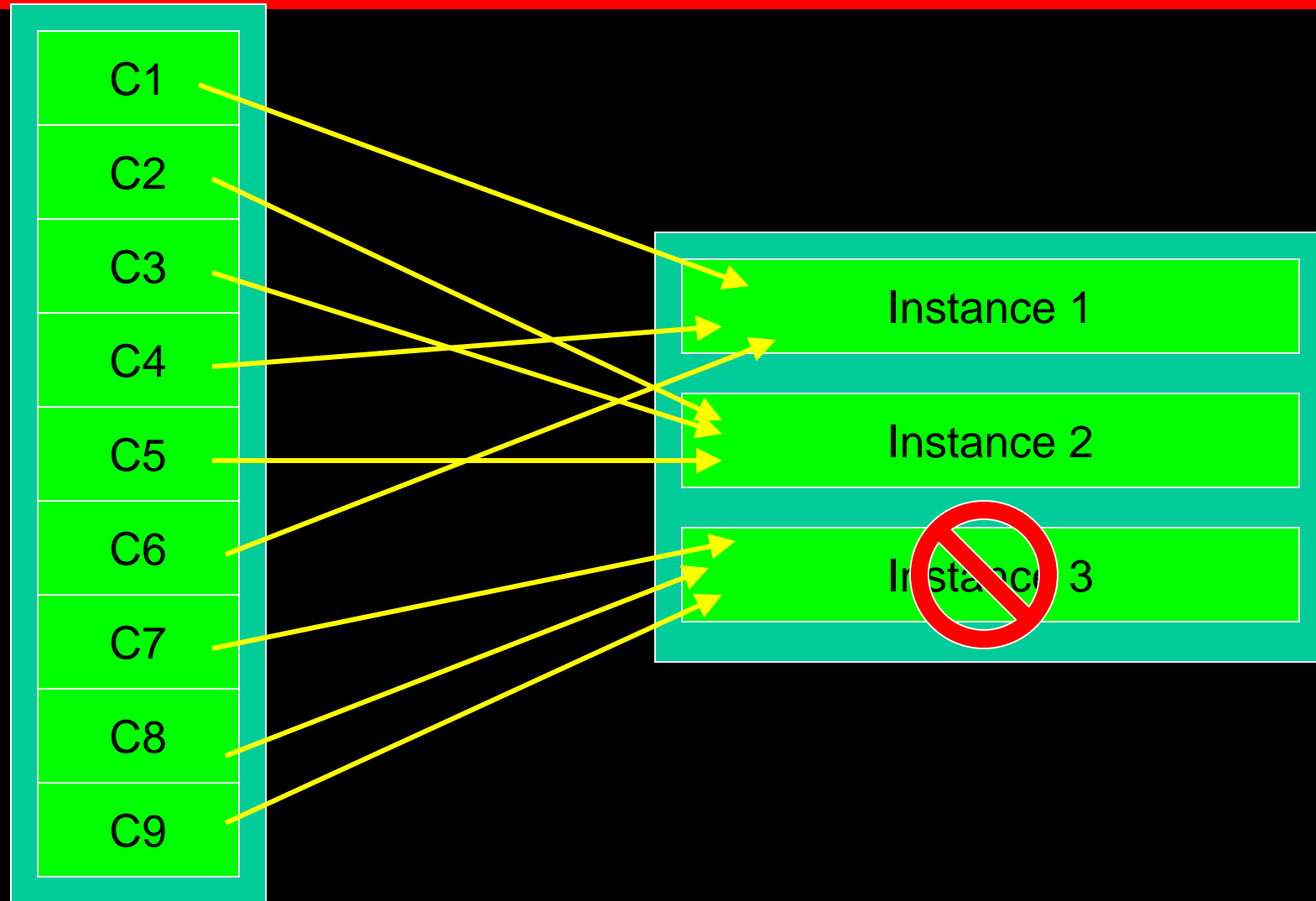
Node Joins . . . Clients Join



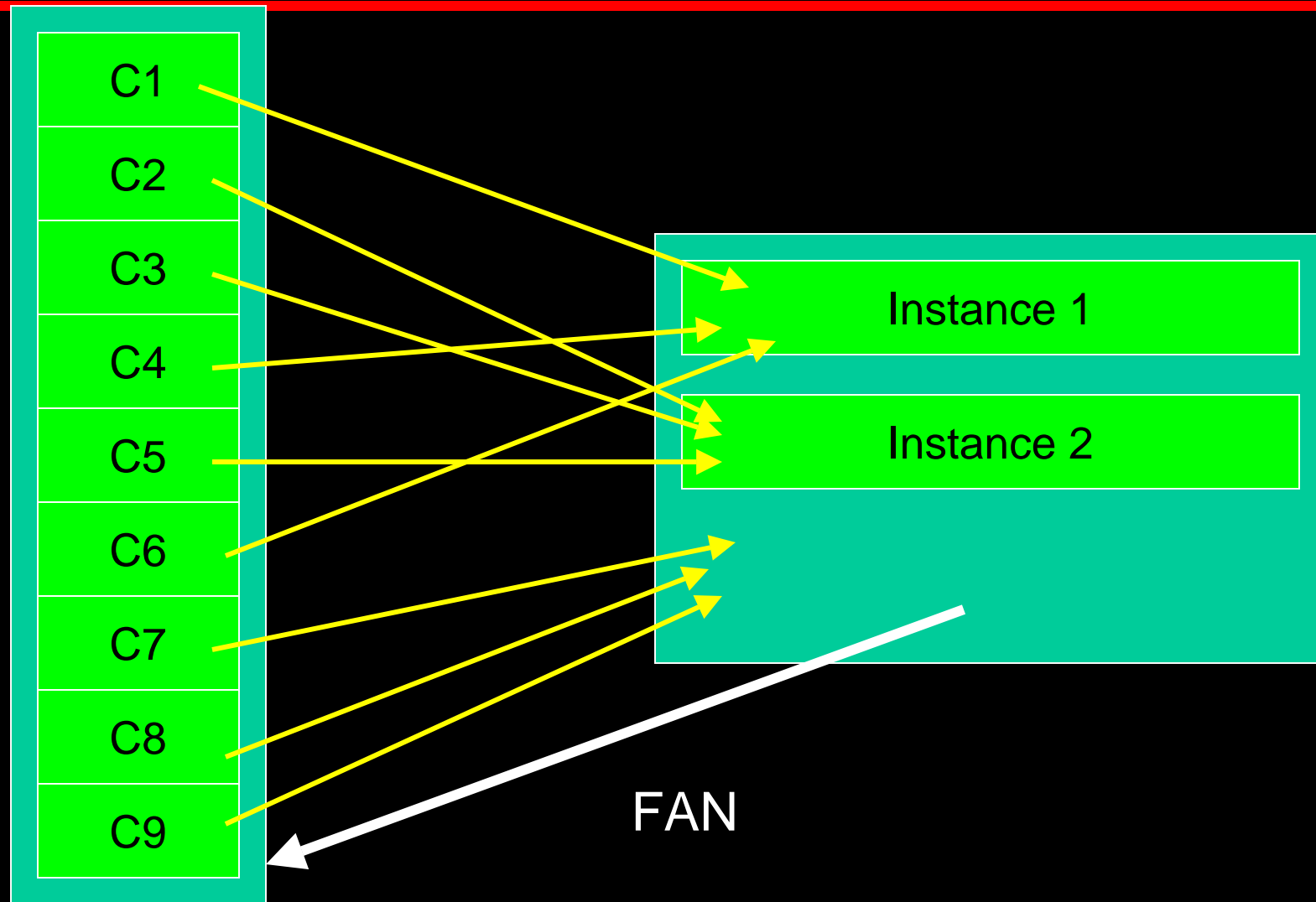
Rebalancing Upon Instance-Client Join

- Desire: Create new connections to new instances and potentially disable some old ones.
- Method:
 - 10g Fast Application Notification (FAN)
 - Using the HA **up** and **down** events for RAC services
 - 9i DBMS session disconnect (Manual steps)

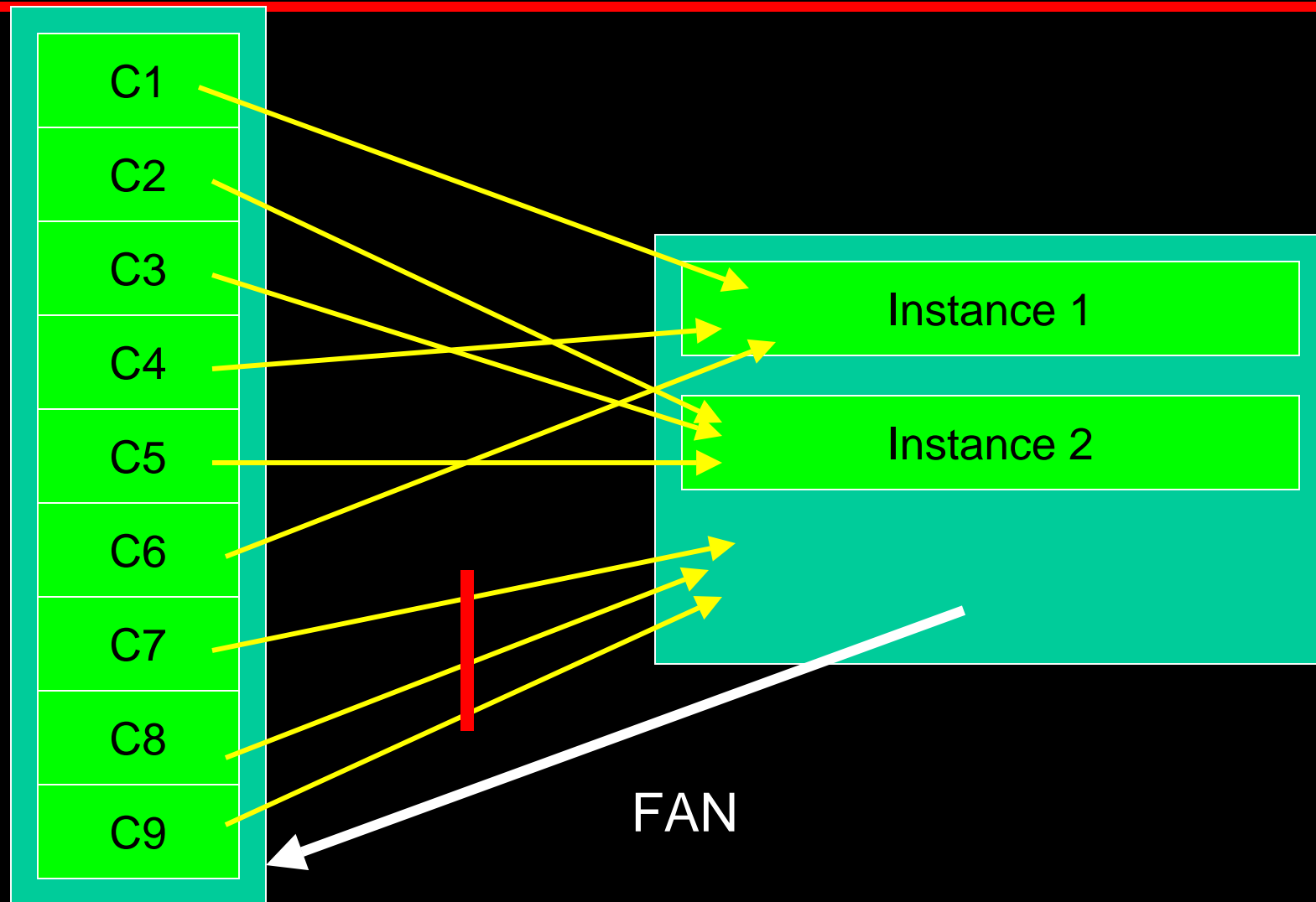
Node Dies



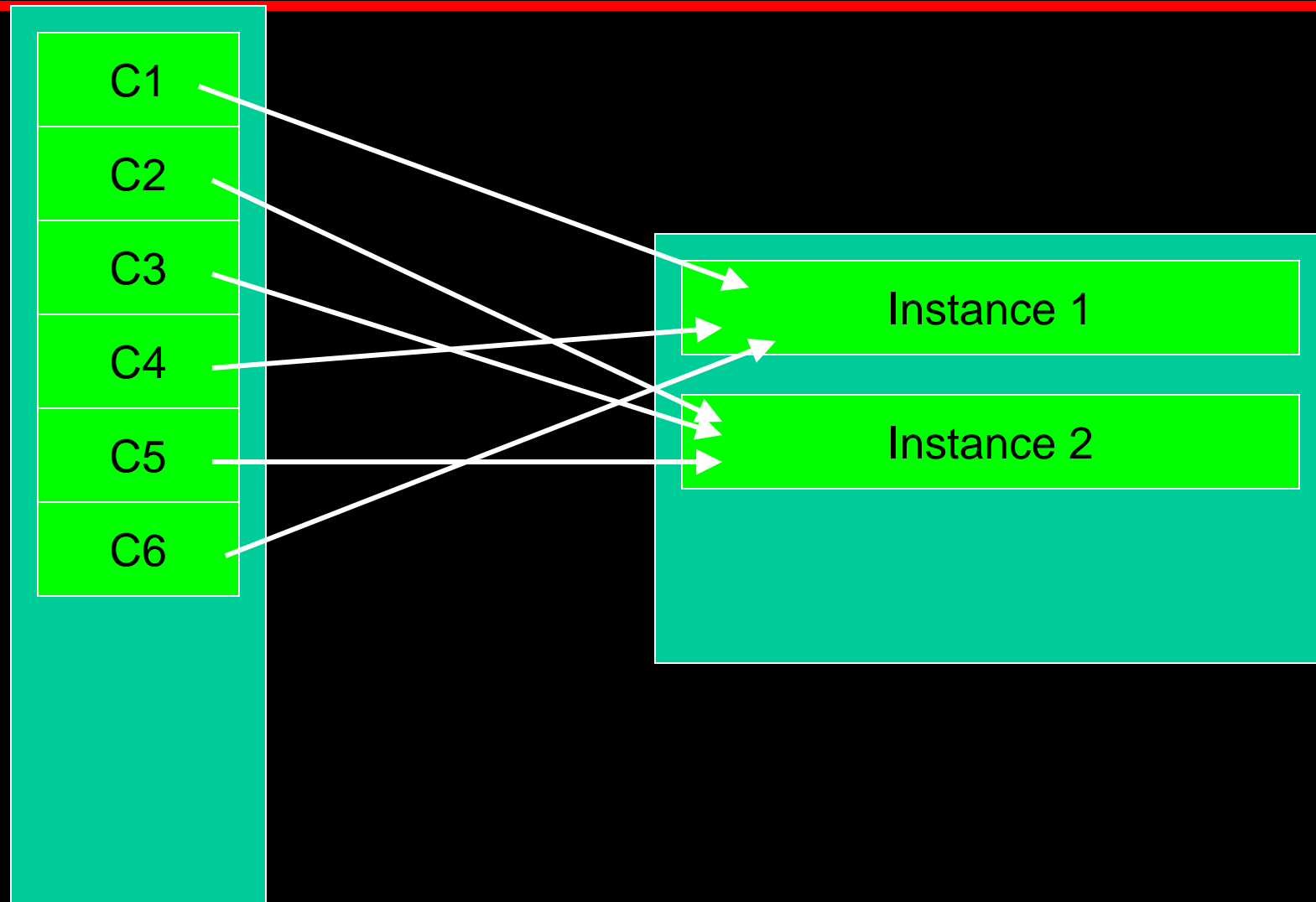
FAN Notifies Connection Pool



Connections Broken



Connections Removed From Pool



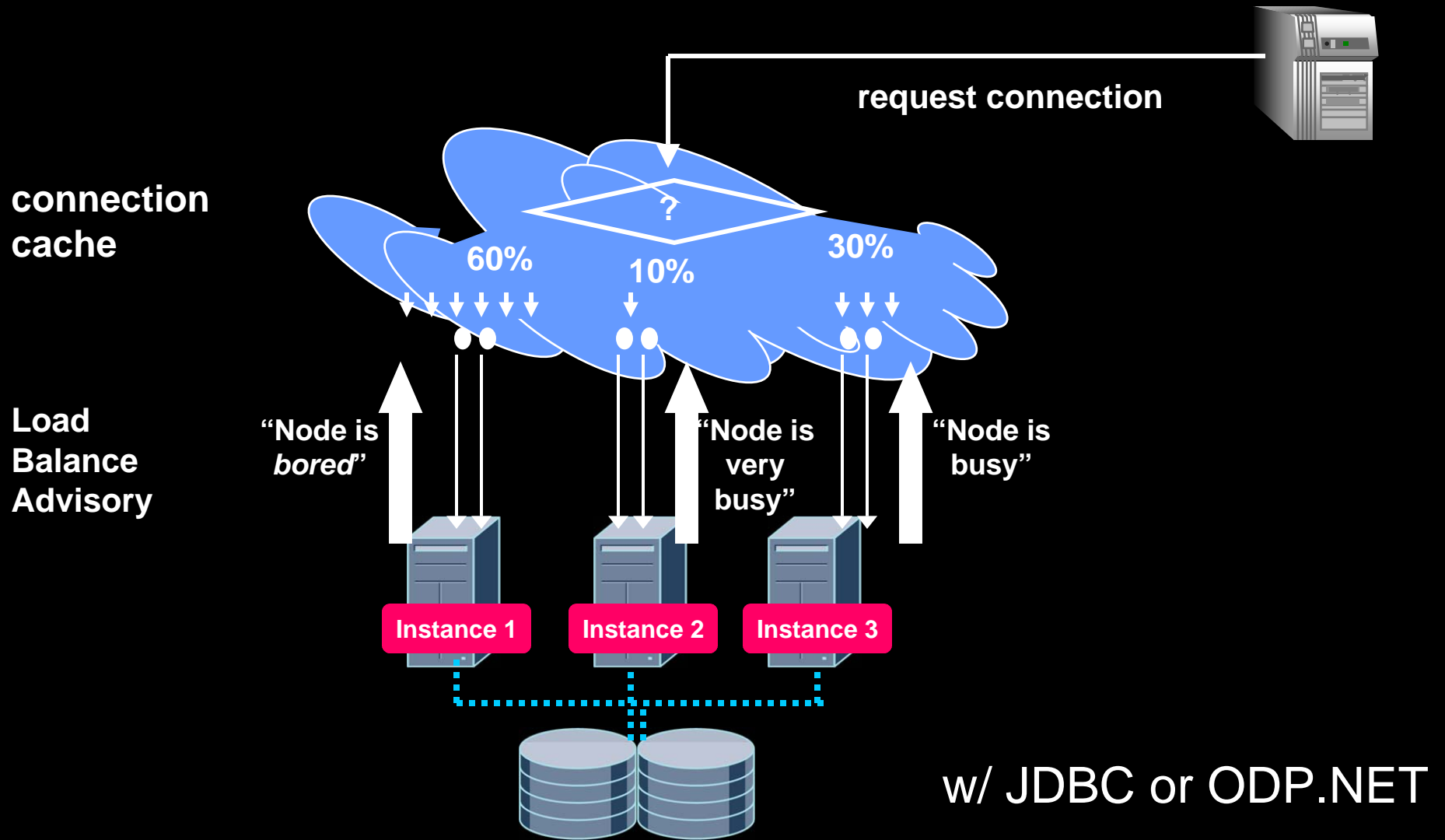
Load Balancing Advisory

- Monitors workload activity for a service across all instances in the cluster
- Analyzes the service level for each instance based on defined goal (service time or throughput)
- Publishes FAN events describing the current service level, amount of work to be sent to each instance and data quality flag
- Integrated with AWR
- If no goal defined, no load balancing events sent.

Runtime Connection Load Balancing

- Subscriber of FAN events
- Works with the connection pool (eg. App Server layer)
- Distributes “units of work” across instances in a cluster according to the Load Balancing Advisory
- Most suitable instance is chosen for each work request

Runtime Connection Load Balancing



Monitoring and Tuning

- All traditional tuning practices apply
- RAC specific pages in DB Console & AWR
- GV\$... dynamic performance views
- Latency & load on the interconnect
- Very little RAC specific tuning, but:
 - Partition application across nodes
 - Services & resource plans
 - Partition data (instance #)
 - Re-design indexes (rev key, inst #)
 - Re-visit block sizes

Recommended Resources

- Oracle Technology Network
 - <http://otn.oracle.com>
- Tahiti
 - <http://tahiti.oracle.com>
- Metalink
 - <http://metalink.oracle.com>
- Oracle RAC Special Interest Group
 - <http://www.oracleracsig.org>
- Morgan's Library
 - www.psoug.org (DBMS_SERVICE, DBMS_UTILITY, RAC)
- Network Appliances (NetApp)
 - www.netapp.com/partners/oracle
 - www.netapp.com/library/tr/3349.pdf
- Oracle RAC Best Practices on Linux White Paper
November 2003, Kirk McGowan, Roland Knapp

RAC War Stories

Caleb Small, BSc, ISP

Caleb@Caleb.com

www.Caleb.com/dba