Under the Hoods of Cache Fusion, GES, GRD and GCS

Arup Nanda

Principal Database Architect

Starwood Hotels

About Me

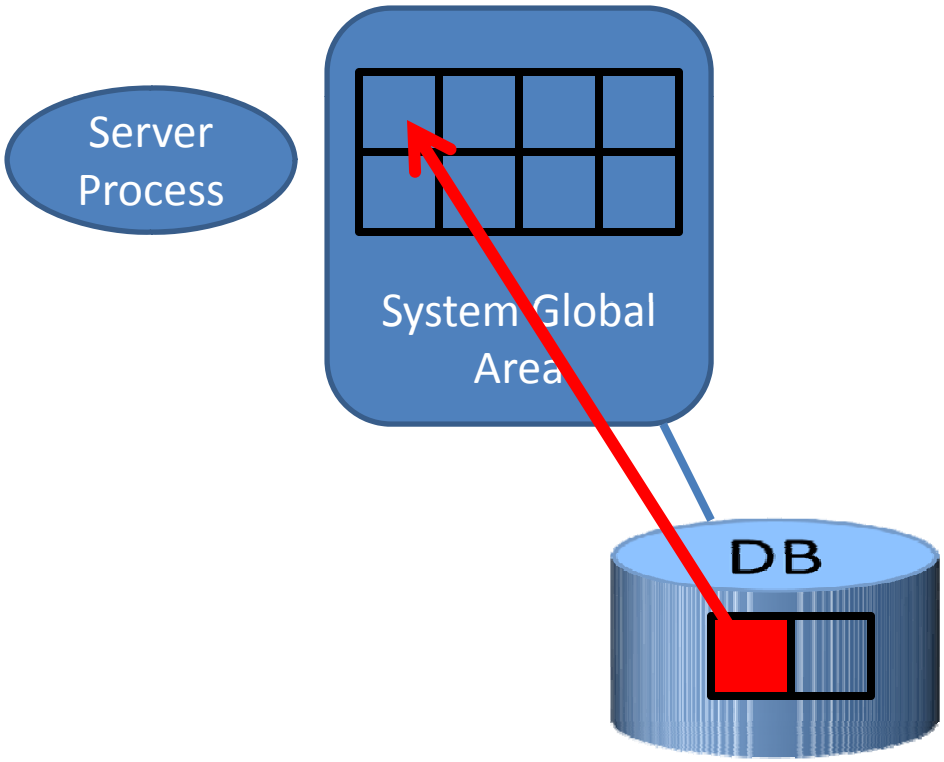
- Oracle Database Admin for 16 years
- RAC (and OPS) since 1999
- Troubleshoot, tune performance
- Developed and Teach a course: *RAC Performance Tuning*

Why this Session?

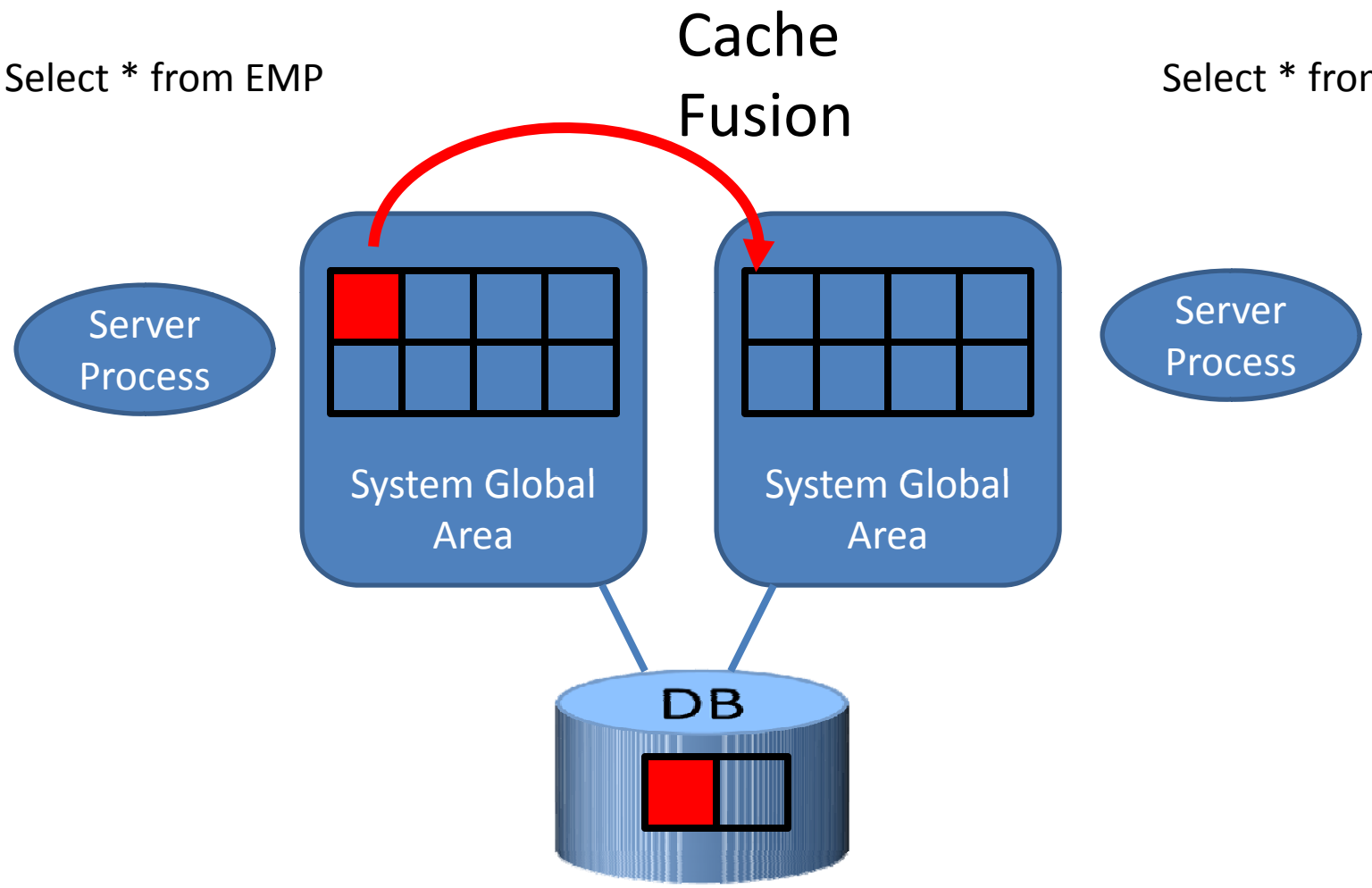
- If I have a 100MB database, I can have a 100 MB buffer cache and I never have to go to the disk, right?
- How does Cache Fusion know where to get the block from?
- How are block locks vary from row locks?
- I'm confused about Global Cache Service (GCS), Global Resource Directory (GRD) and Global Enqueue Service (GES)
- We will understand how all these actually work

Buffer Cache

Select * from EMP



RAC – More than 1 Buffer Cache



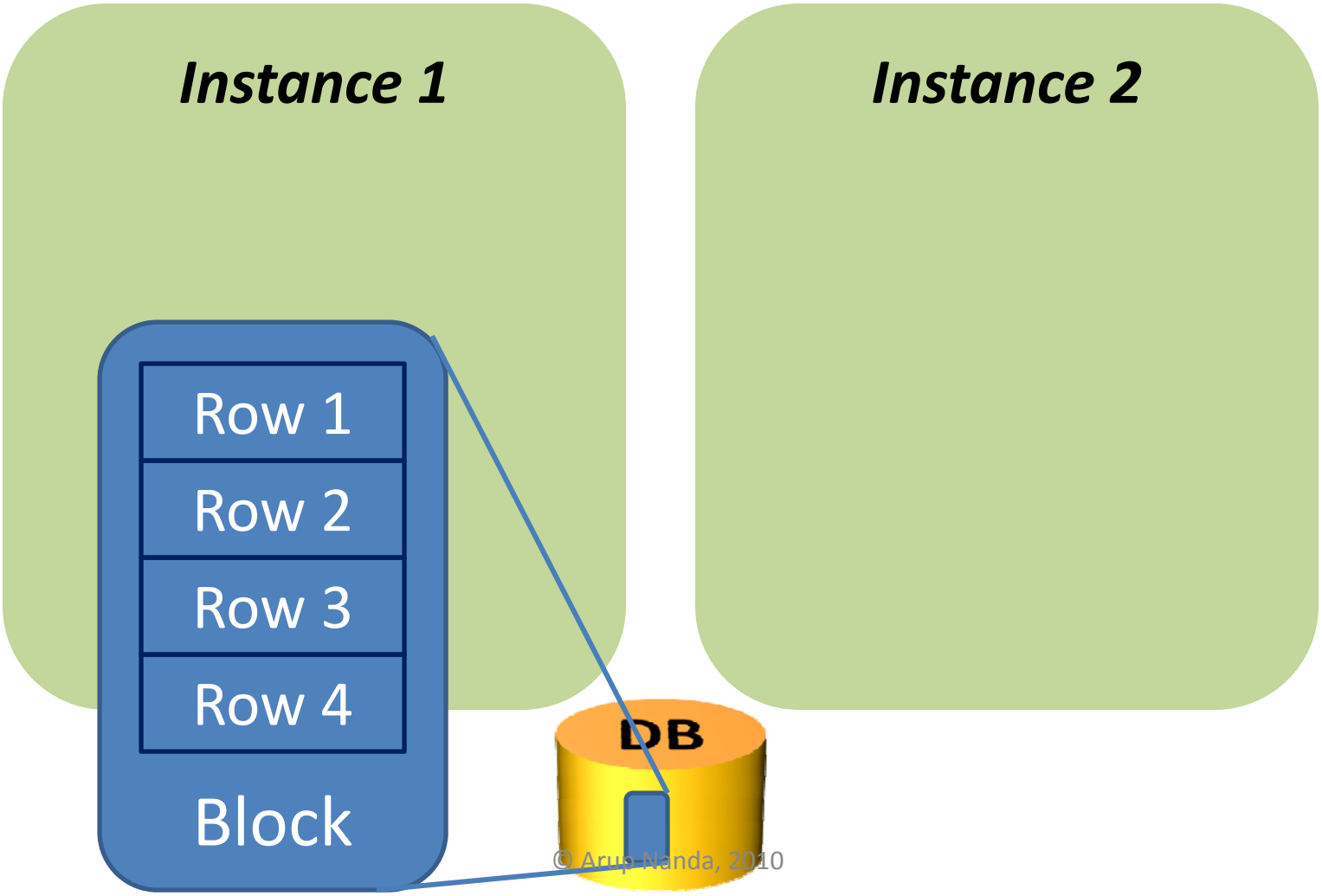
To Cache Fusion or Not?

- When a block is requested, the buffer cache is searched
- If not found, there are two options
  - Get from disk
  - Get from the other cache
- If found, there are three options:
  - Send the buffer to the user
  - Examine other caches for the presence of this buffer
  - Get from the disk
- How does it decide which option to take?

Buffer States

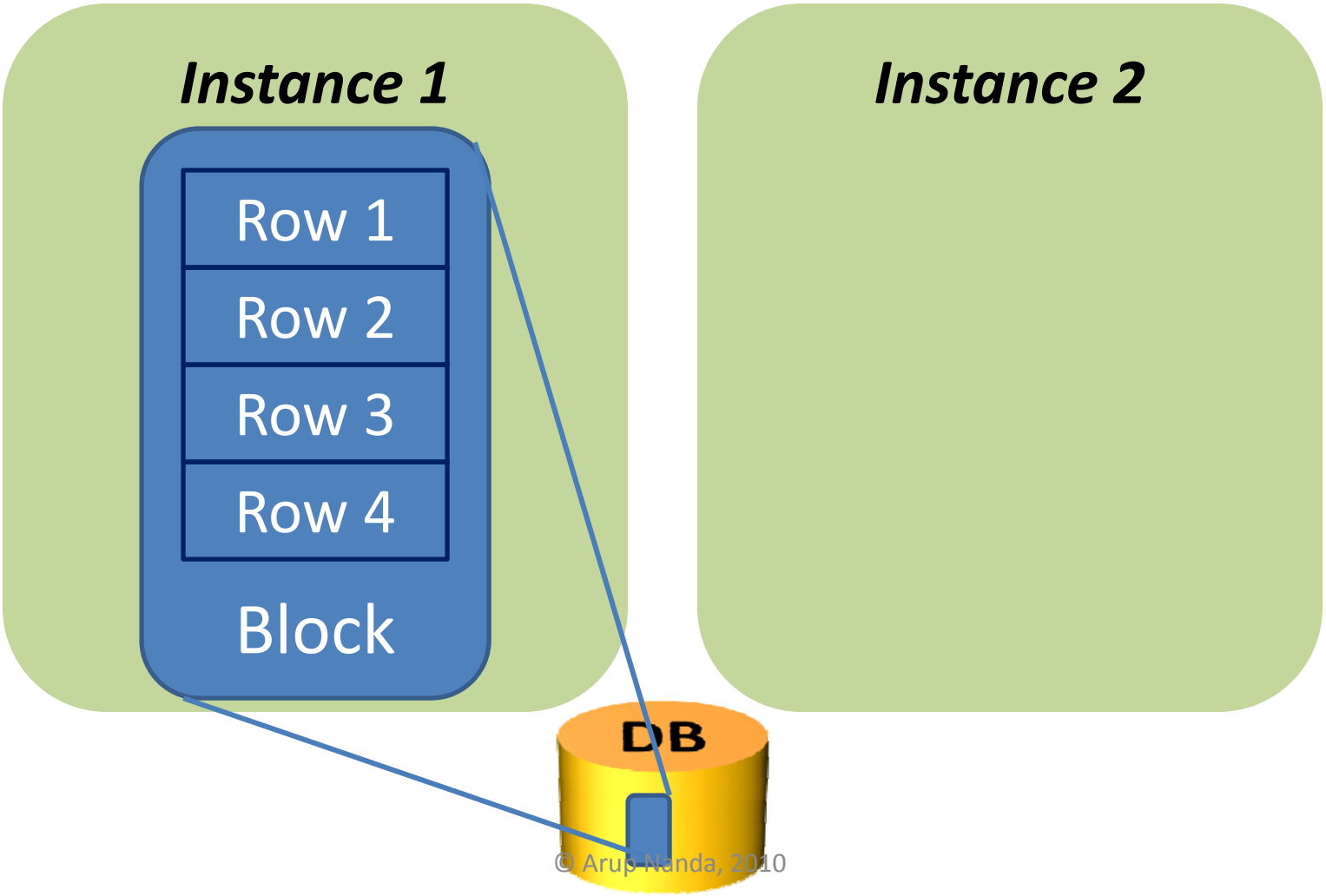
- The buffer can be retrieved in two modes
  - Consistent Read (CR)
  - Current
- There can be several CR copies of a buffer
- There can be only one current mode
  - For an instance
- Each current buffer is Shared Current
- Only one buffer in the entire cluster can be Exclusive Current

Block – Row Relationship



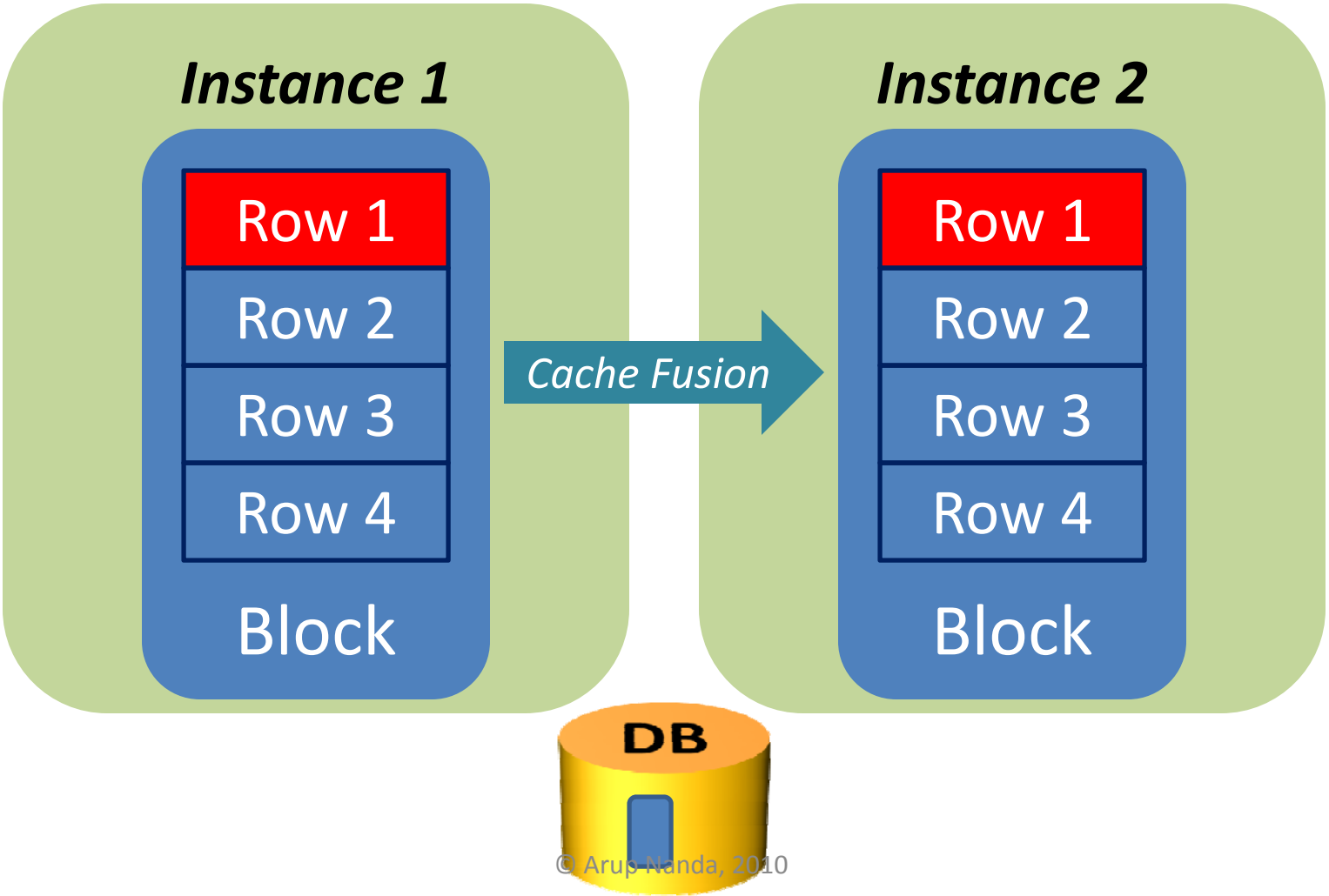
Update on One Instance

```
UPDATE ROW1 …
```

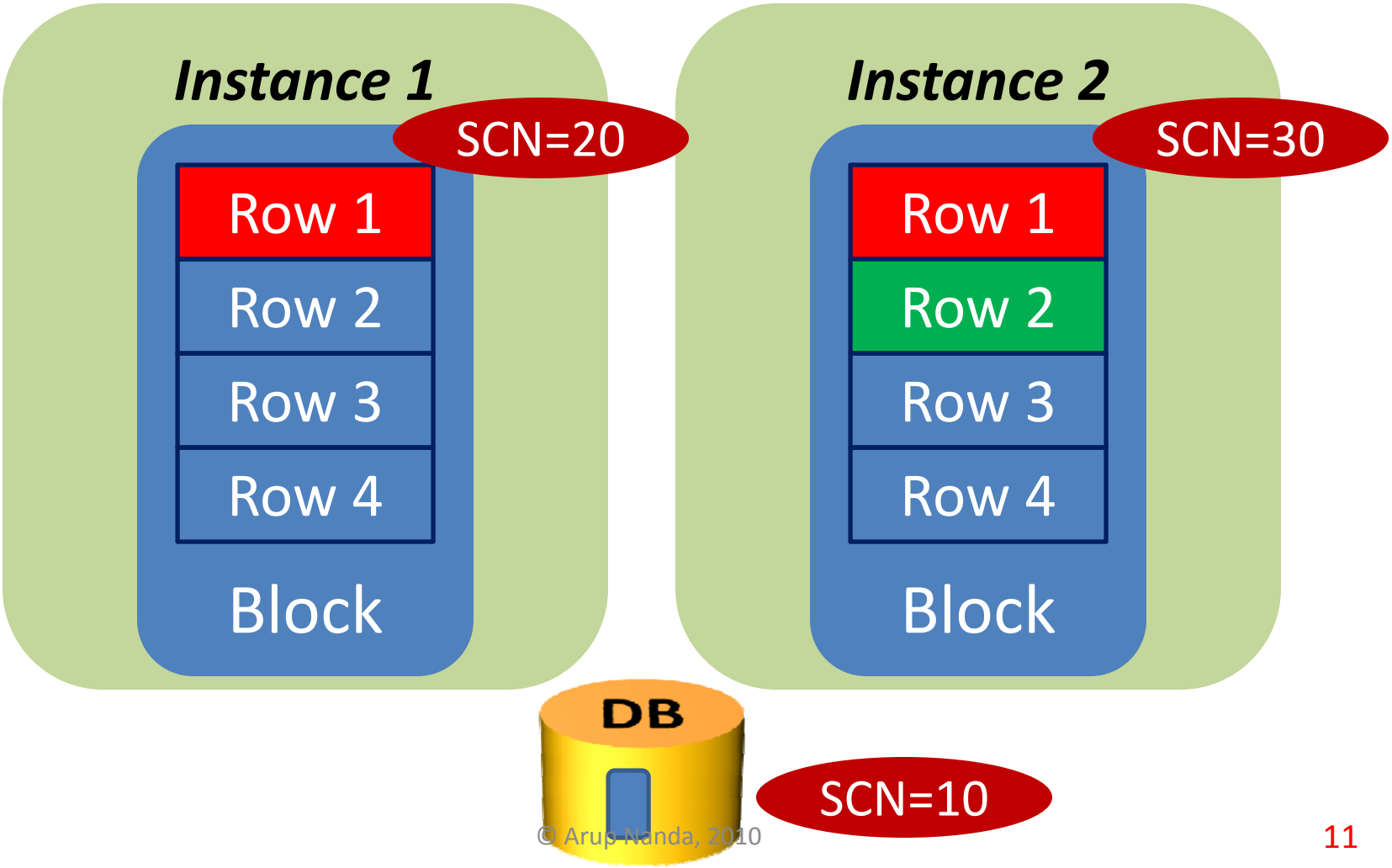


Update a Different Row on Node 2

```
UPDATE ROW2 …
```

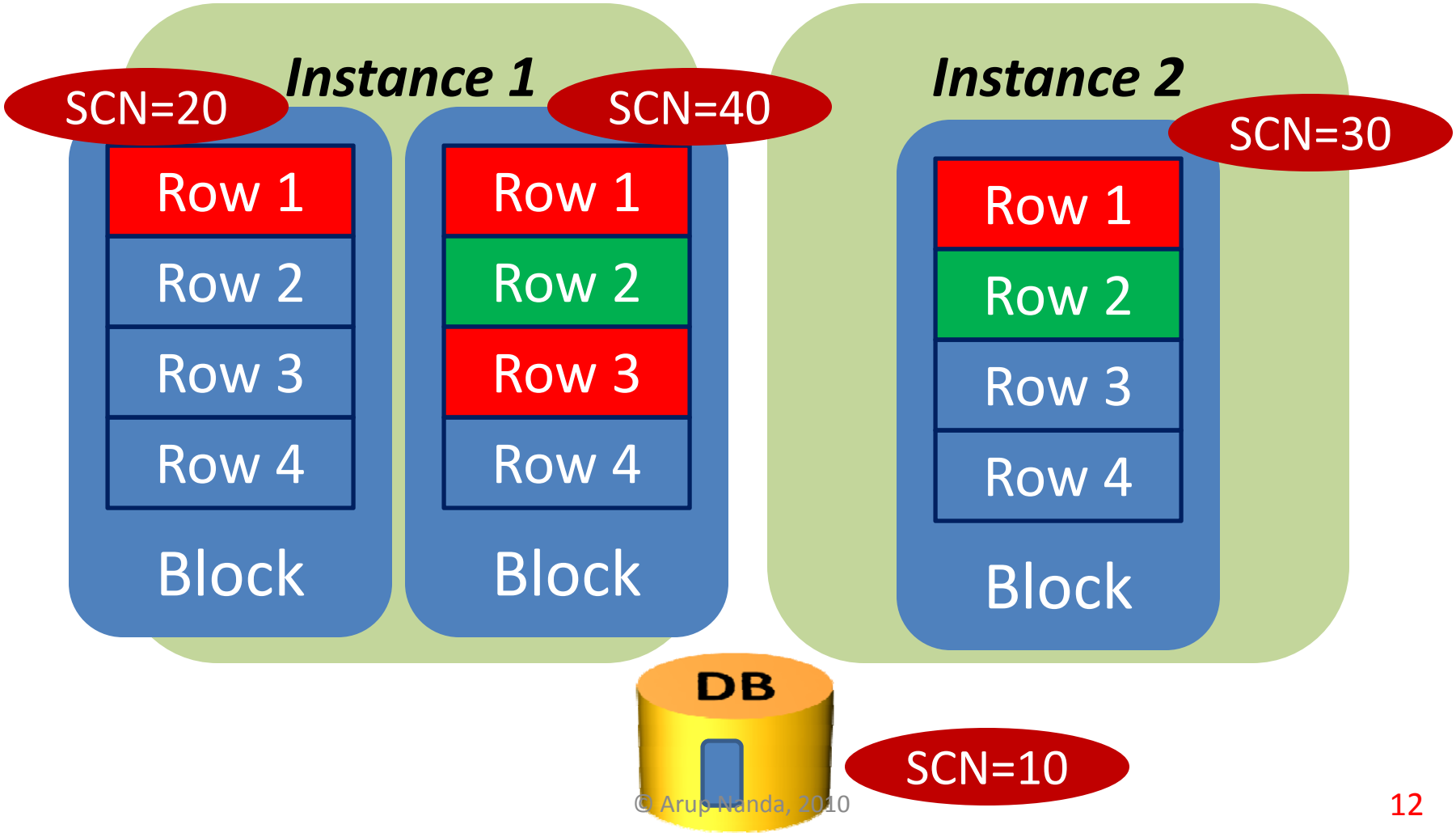


Buffer Versions



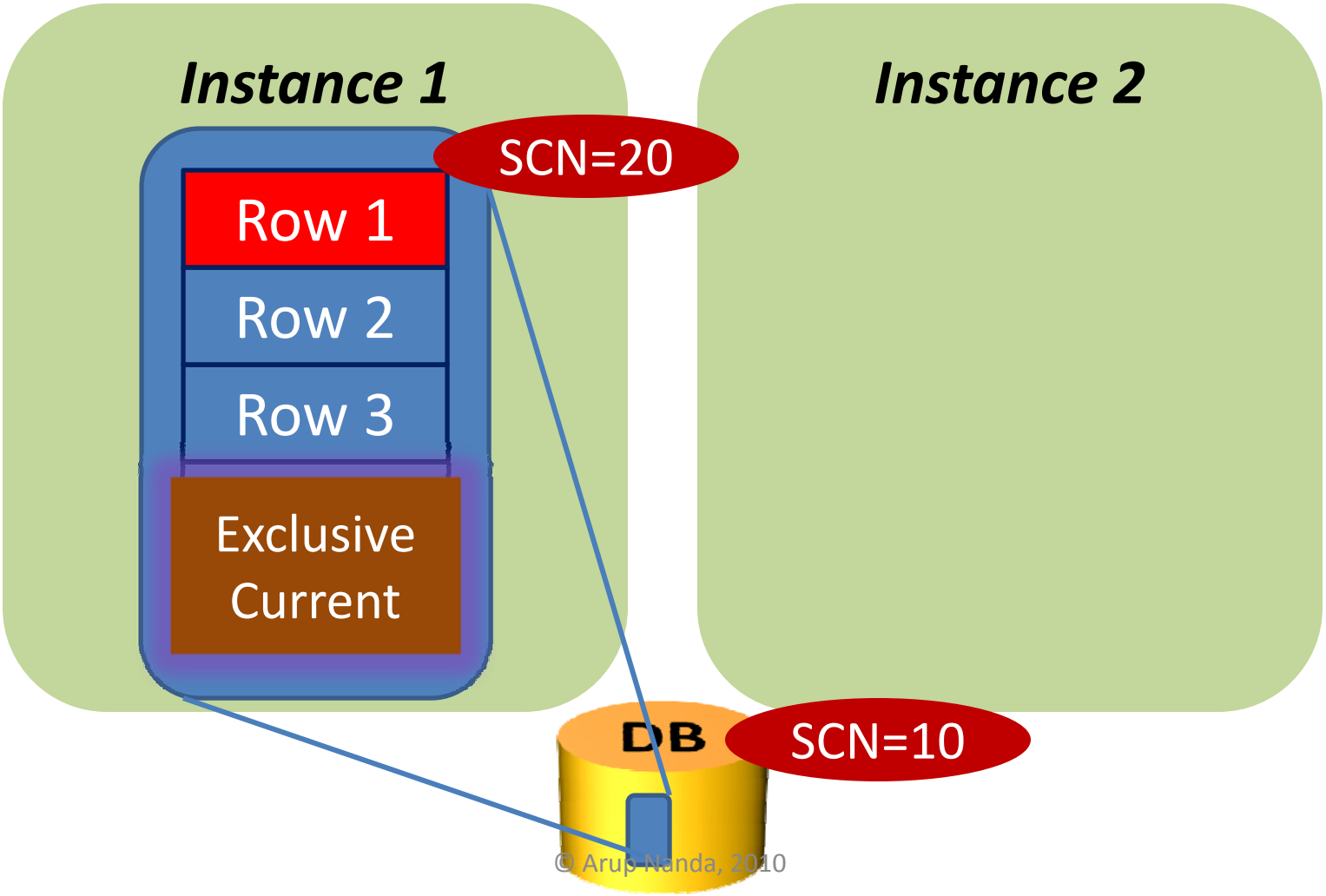
Buffer Versions

```
UPDATE ROW3 …
```



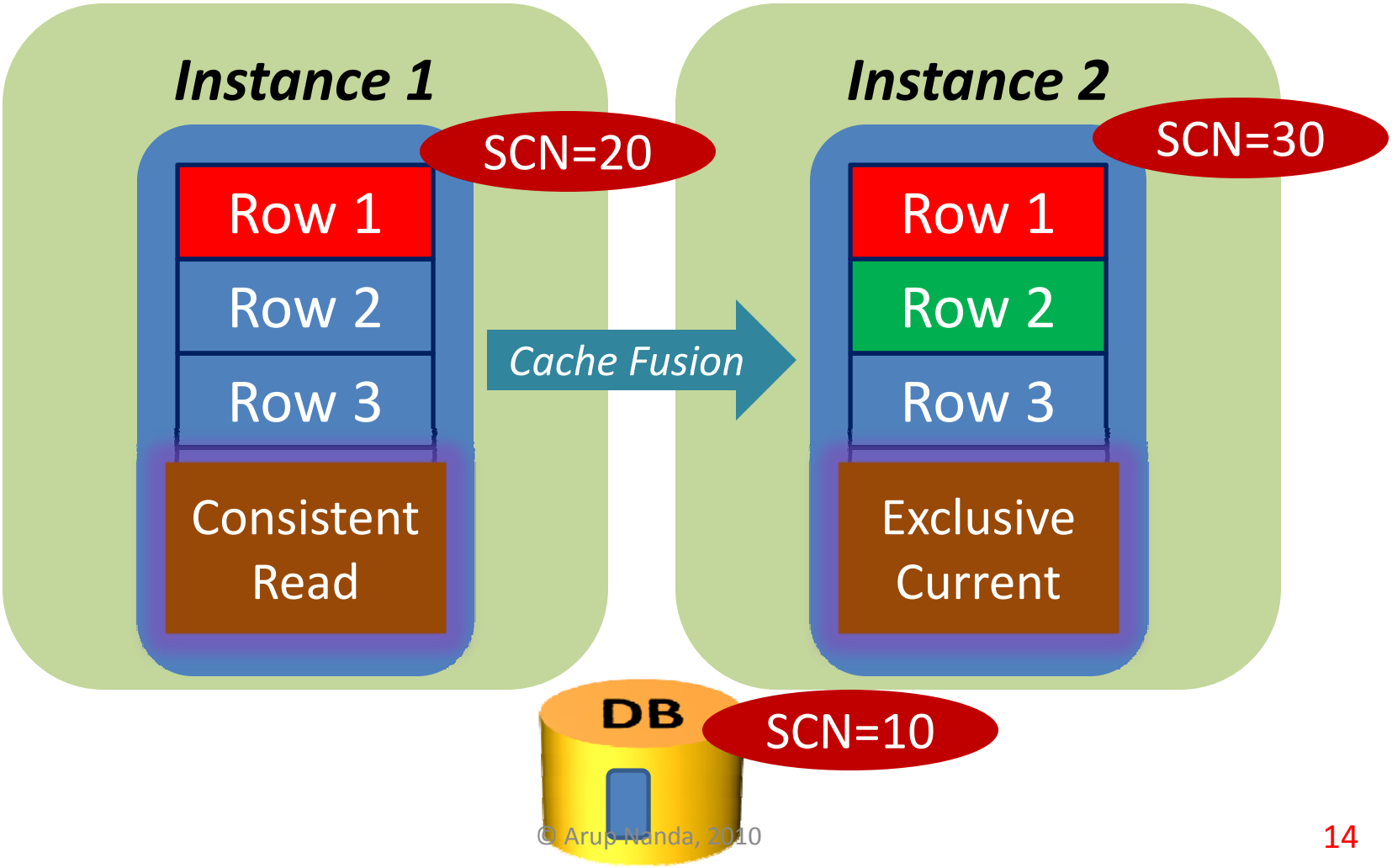
Buffer State 1

```
UPDATE ROW1 …
```



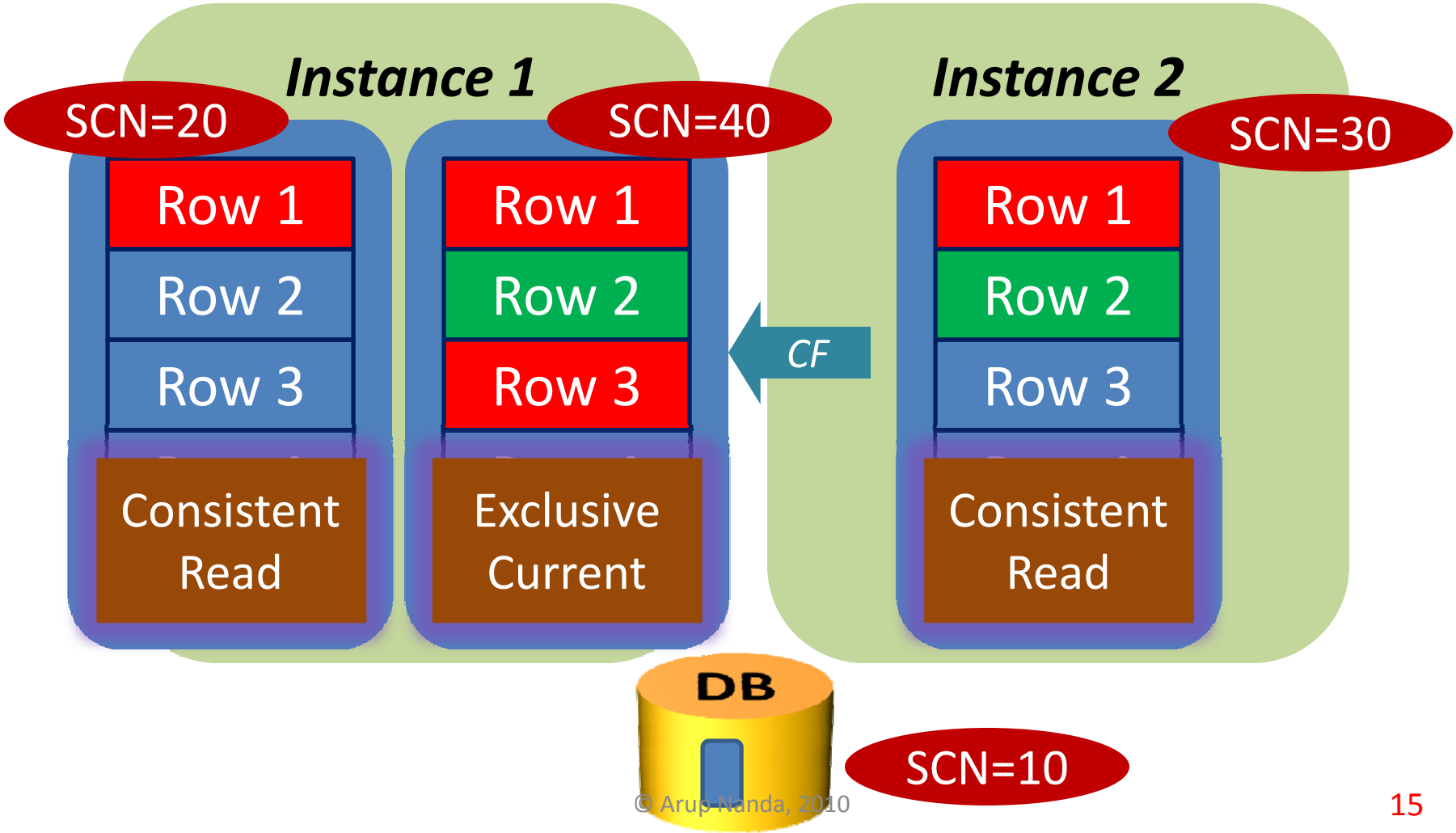
Update a Different Row on Node 2

```
UPDATE ROW2 …
```



Buffer Versions

```
UPDATE ROW3 …
```



Putting it all together

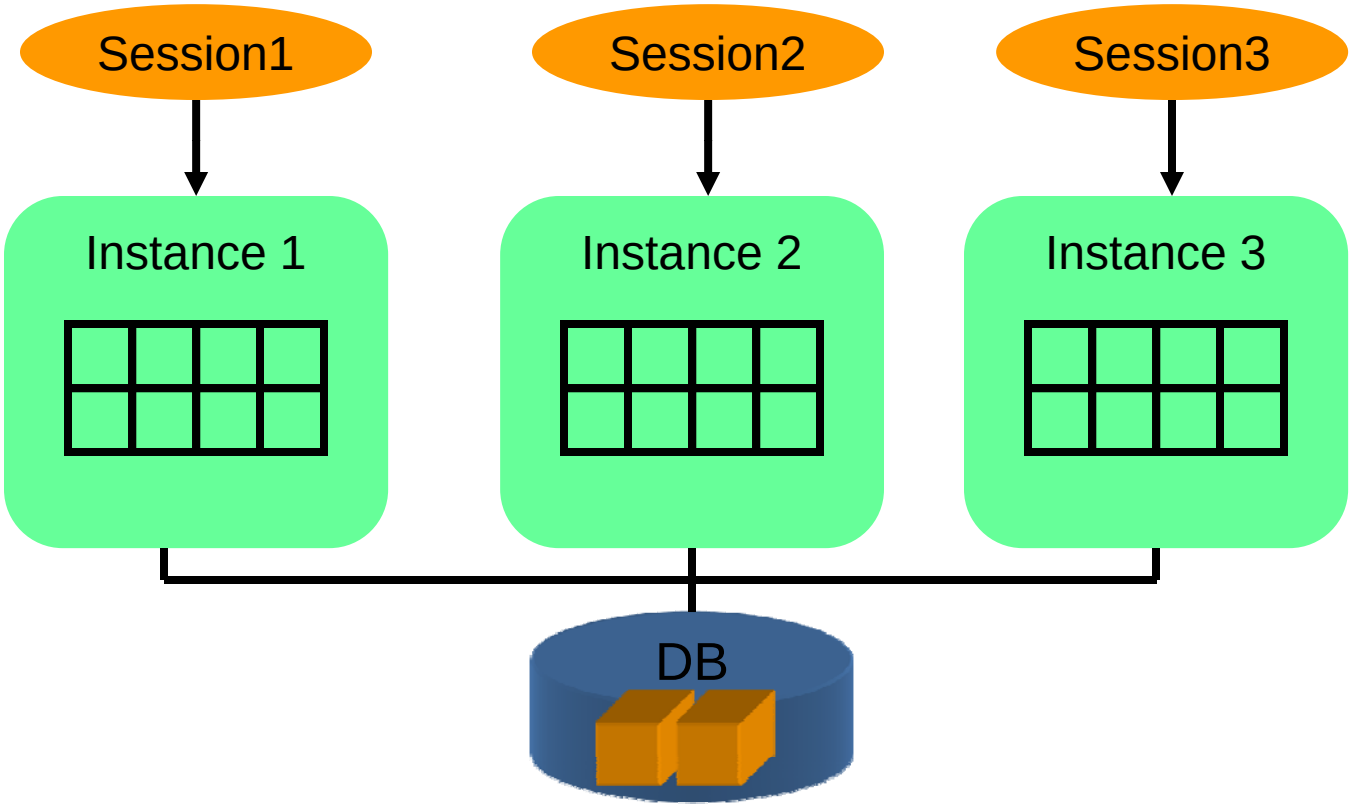
- When the block comes from the disk to the buffer cache
  - If the intent is to modify, it's gotten in CURRENT mode
  - If the intent is to read, it's gotten in CR mode
- There can be only one Shared Current per instance
  - Multiple SCURs in the cluster
- Many CR copies in the instance

Past Image

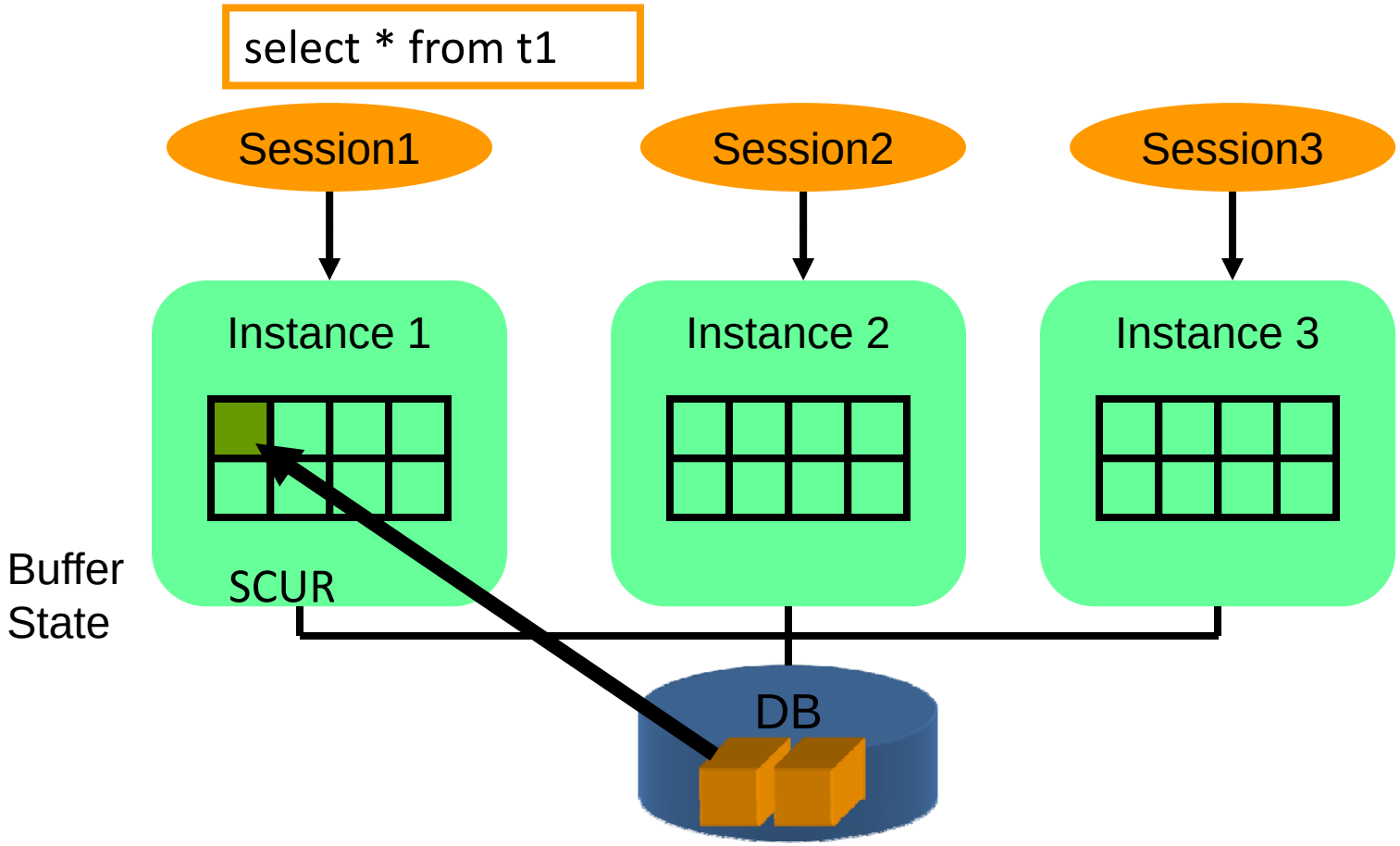
- Sequence of Events
  - Instance 1 has version 1 of the block
  - Instance 2 has version 2
  - Instance 2 updates the block -> current block changes
  - Instance 1 wants to update the block
  - Instance 2 prepares a copy of the block before sending it
- This “copy” is called a Past Image (PI) of the block
  - Note: the term Past Image is not documented in Oracle Manuals. It’s just widely understood and acceptable.

Cache Fusion in Operation

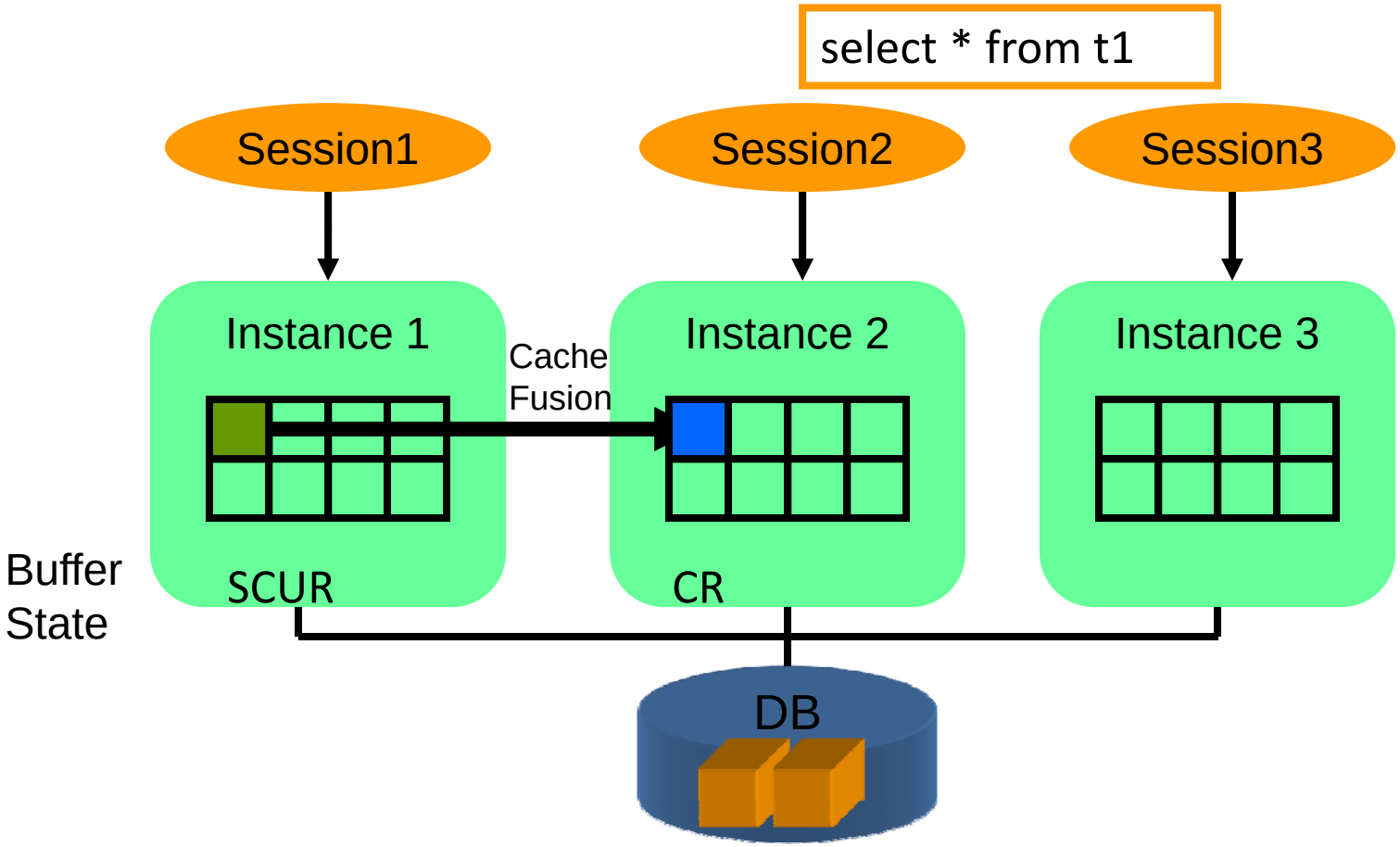
Time 0



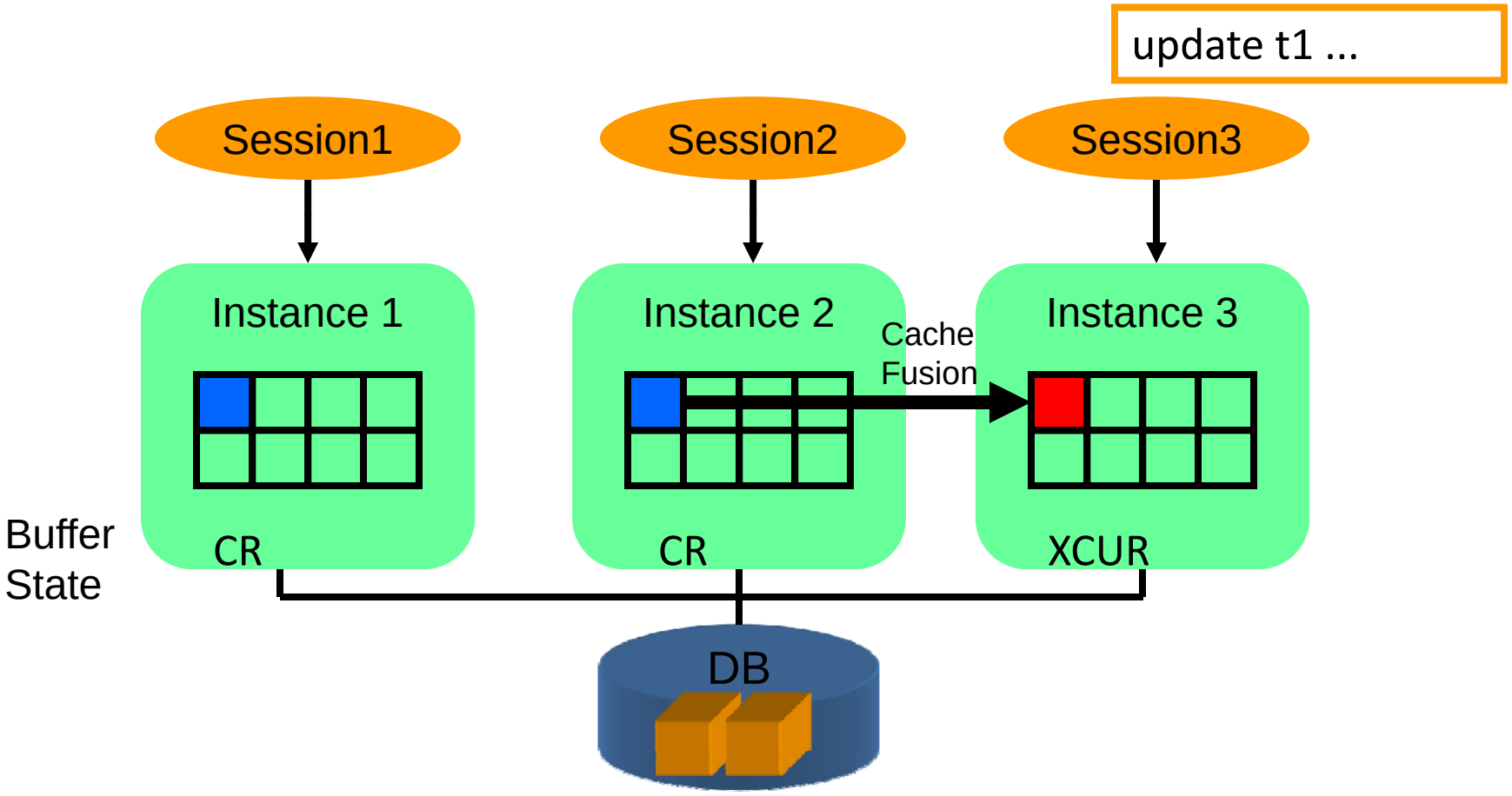
Time 1



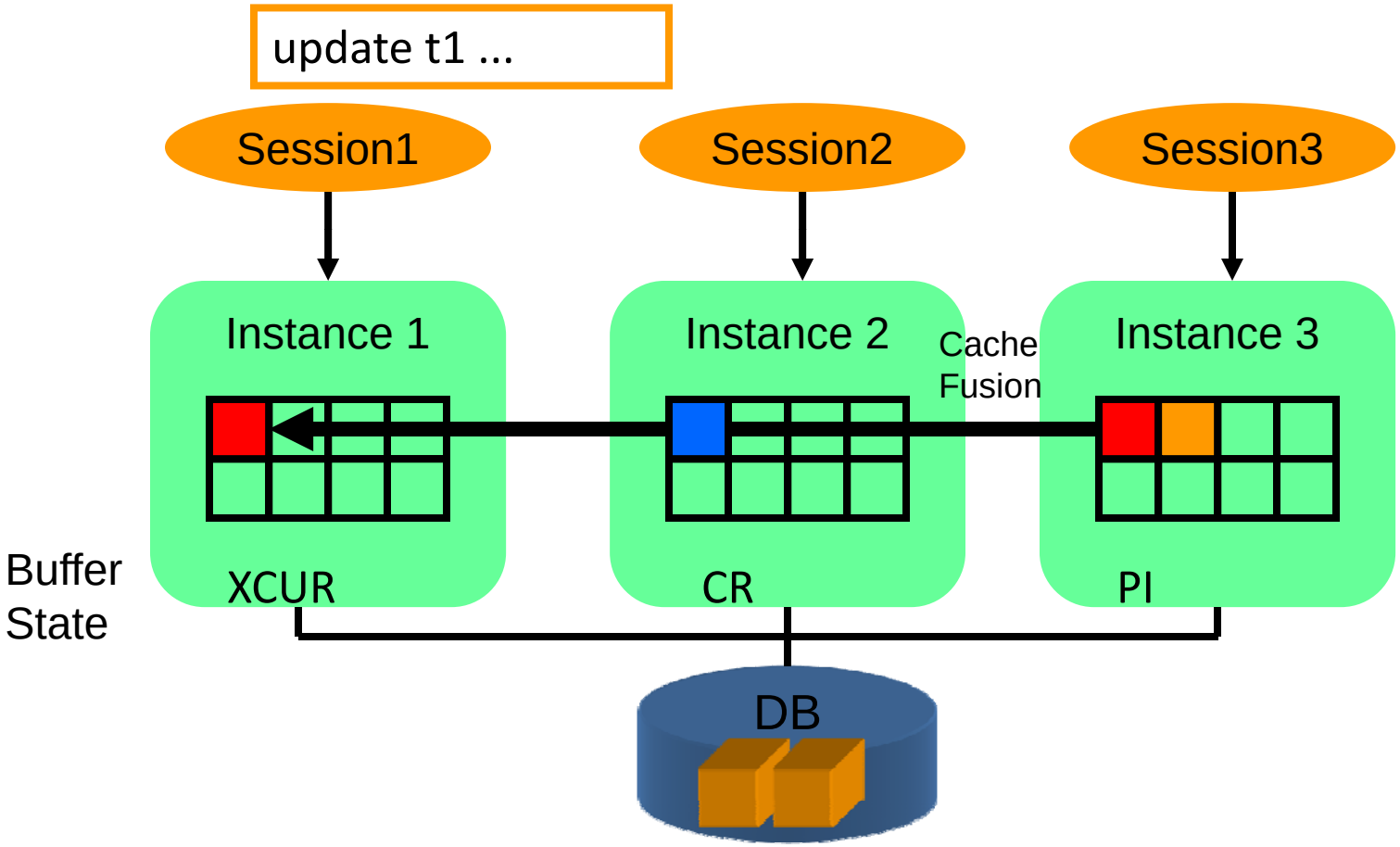
Time 2



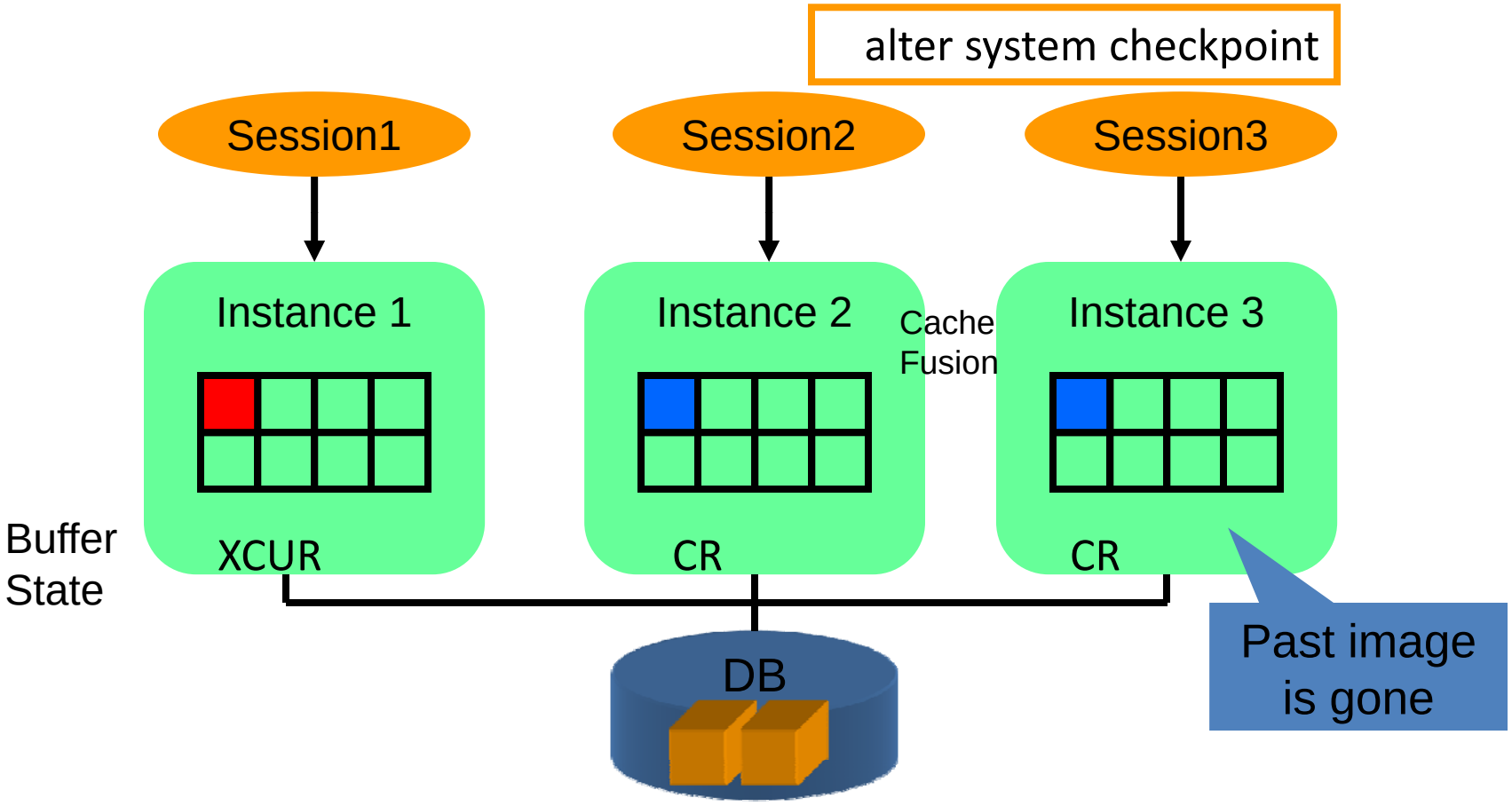
Time 3



Time 4



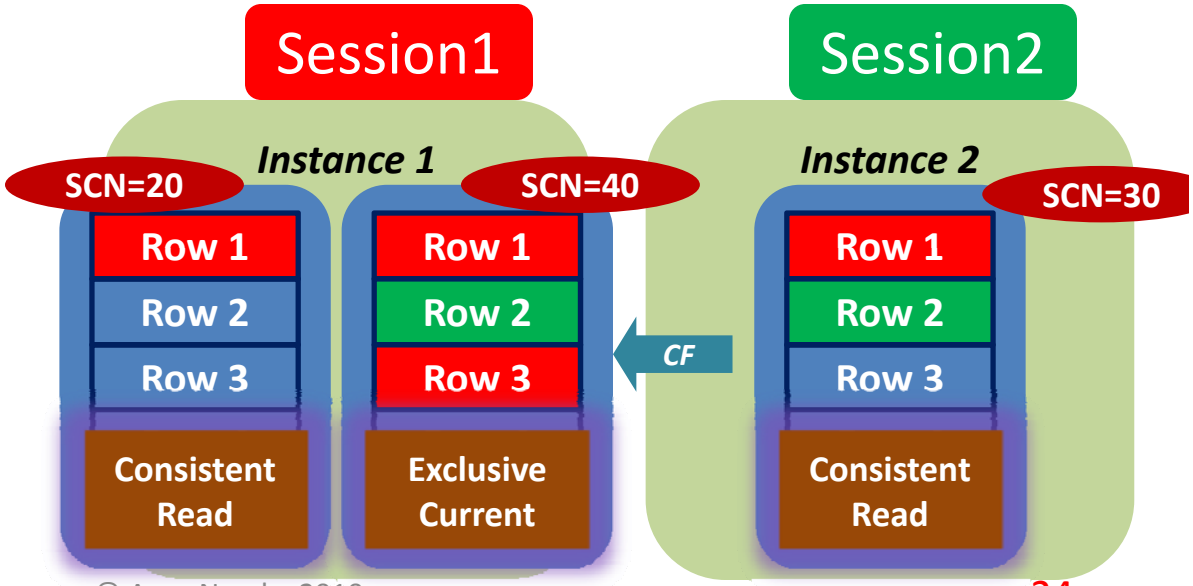
Time 5



Buffer Lock

- When an instance wants to change the state of the buffer from CR to Exclusive Current
  - It must get a lock on that buffer
  - This is called a Buffer Lock
  - Different from a row lock

Buffer Locks:
Instance 1 = *Exclusive*
Instance 2 = *None*
Row Locks:
Session 1 = *Row 1 and Row 3*
Session 2 = *Row 2*



Global Cache Service

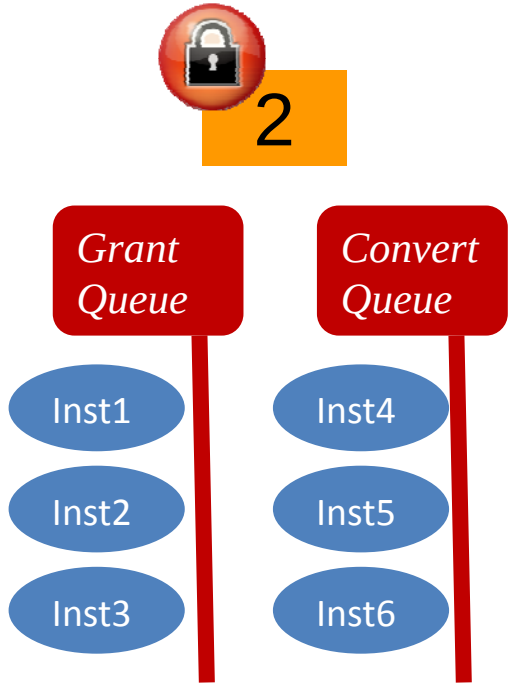
- Provides buffer from one instance to the other
  - But does not know who has what type of buffer lock

Global Enqueue Service

- Used to be called Dynamic lock Manager (DLM)
- Holds the information on the locks on the buffers
- Each lock has a name shown in V$LOCK_ELEMENT (or X$LE)
- This is different from row locking, which is on a specific row
- If a buffer is locked, the lock element name is shown in V$BH.LOCK_ELEMENT

Lock Queuing

- Each Buffer in a RAC instance has two queues
  - Grant Queue – the queue where the requesters are queued for the locks to be granted in a certain mode
  - Convert Queue – the queue where the granted requests are queued to be notified to the requesters
- The queues for a specific buffer are placed in a single instance



Master Instance

- The instance that has the Grant and Convert Queues of the Buffer is called the **Master Instance** of the Buffer
- A Buffer has only one Master
- The Master may change
  - Manually
  - By a process known as Dynamic Resource Mastering
- When an instance wants to get a lock, it has to check with the master

Global Resource Directory

- Someone has to keep a list of all buffers and where they are mastered
- This is called Global Resource Directory (GRD)
- GRD is present on all the instances of the cluster
- To find out the master:

```
select  b.dbablk, r.kjblmaster master_node
from x$le l, x$kjbl r, x$bh b
where b.obj = <DataObjectId>
and b.le_addr = l.le_addr
and l.le_kjbl = r.kjbllockp
```

Demo

In Summary

- Buffers are gotten in 2 modes
  - CURRENT - is need to be modified
  - CR - if selected only for reading
- Every time other node wants the buffer
  - it is copied to a new buffer and sent (CR processing)
- There can be only one current state of the buffer in an instance in Shared Mode
- Only one Exclusive Current in the Cluster
- The Exclusive/Shared Current Locks on the Buffer is handled by GES
- Each buffer has a master node that holds the lock Grant and Convert Queues
- GRD maintains information on the buffers-masters

Thank You!

My Blog: **arup.blogspot.com**
My Email: **arup@proligence.com**
Download the Scripts:
proligence.com/cfscripts.zip