Exadata: from Beginner to Advanced in 3 Hours

Arup Nanda

Longtime Oracle DBA

(and now DMA)

Why this Session?

- If you are
  - an Oracle DBA
    - Familiar with RAC, 11gR2 and ASM
  - about to be a Database Machine Administrator (DMA)
- How much do you have to learn?
- How much of you own prior knowledge I can apply?
- What's different in Exadata?
- What makes it special, fast, efficient?
- Do you have to go through a lot of training?

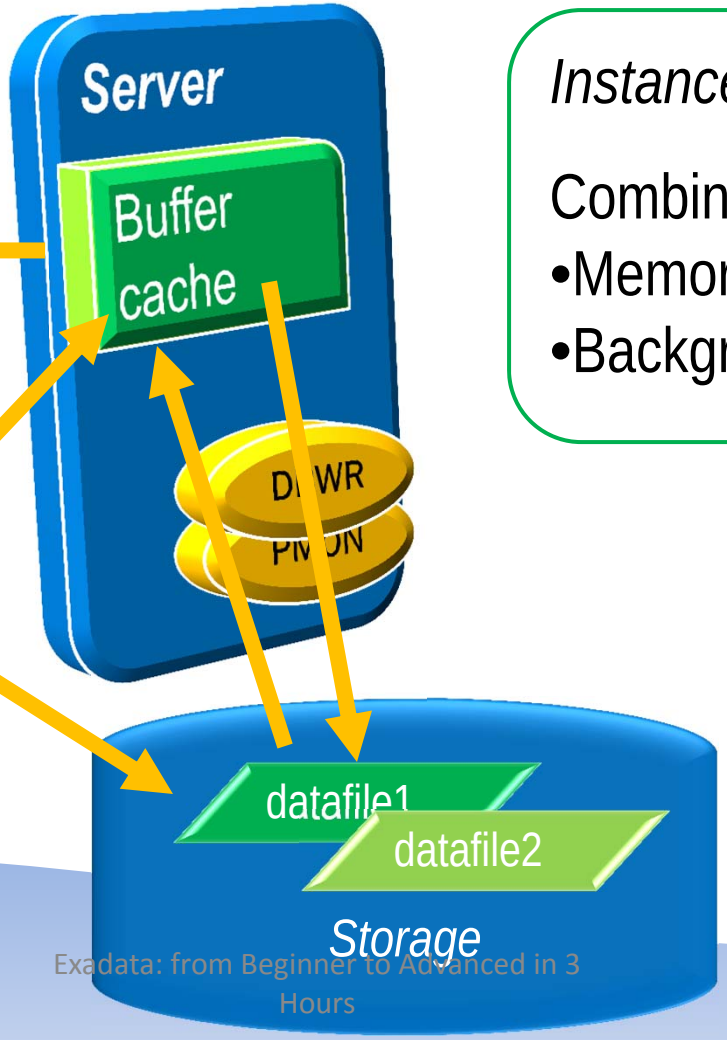
What is Exadata

- It is like an *appliance* containing
  - Storage, Flash Disks, Database Servers, Infiniband Switches, Ethernet Switches, KVM (some models)
- But it is *not* an appliance. Why?
  - additional software to make it a better database machine
  - Components can be managed independently
- That's why Oracle calls it a **Database Machine** (DBM)
- And **DMA** – Database Machine Administrator

Anatomy of an Oracle Database

```
SELECT NAME
FROM CUSTOMERS
WHERE STATUS ='ANGRY'
```

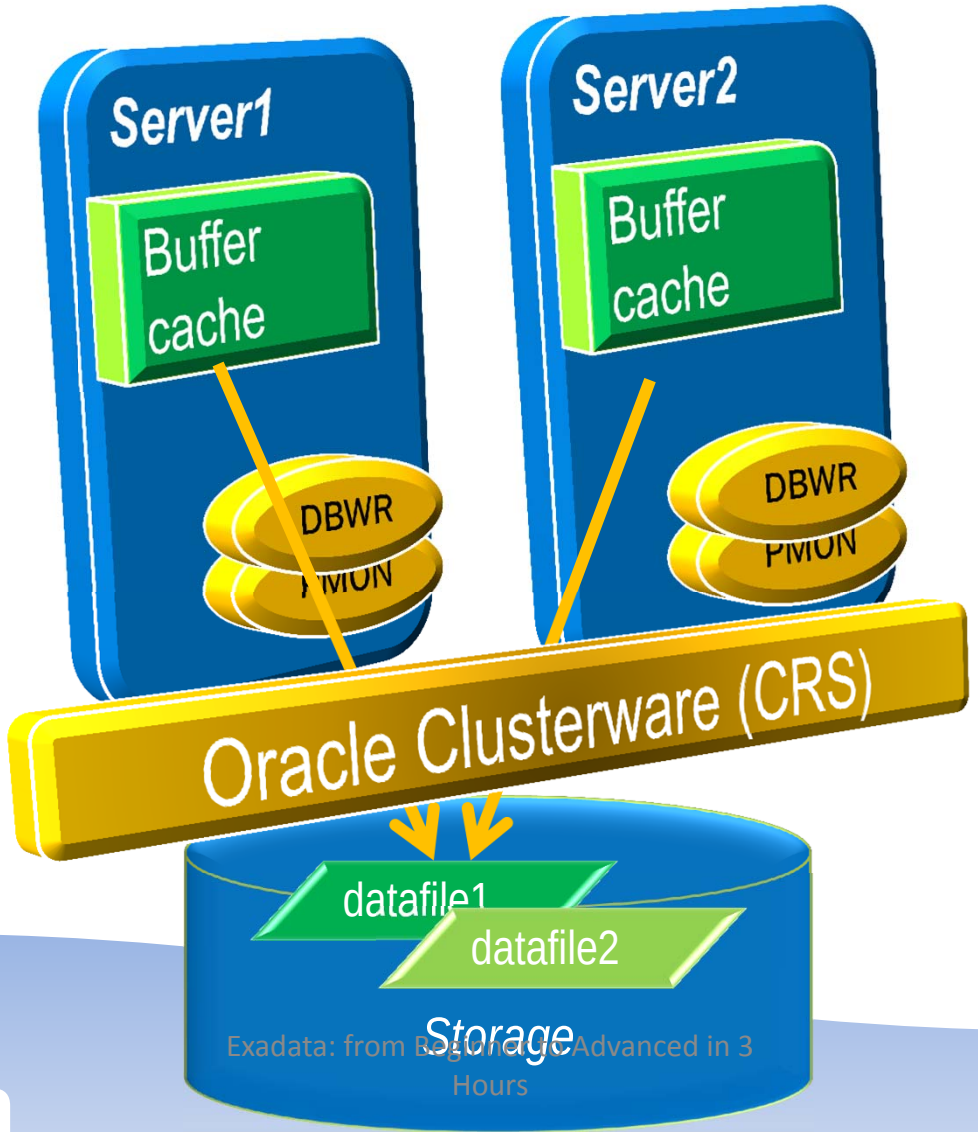
```
UPDATE
CUSTOMERS
SET BONUS = 1M
WHERE STATUS ='ANGRY'
```



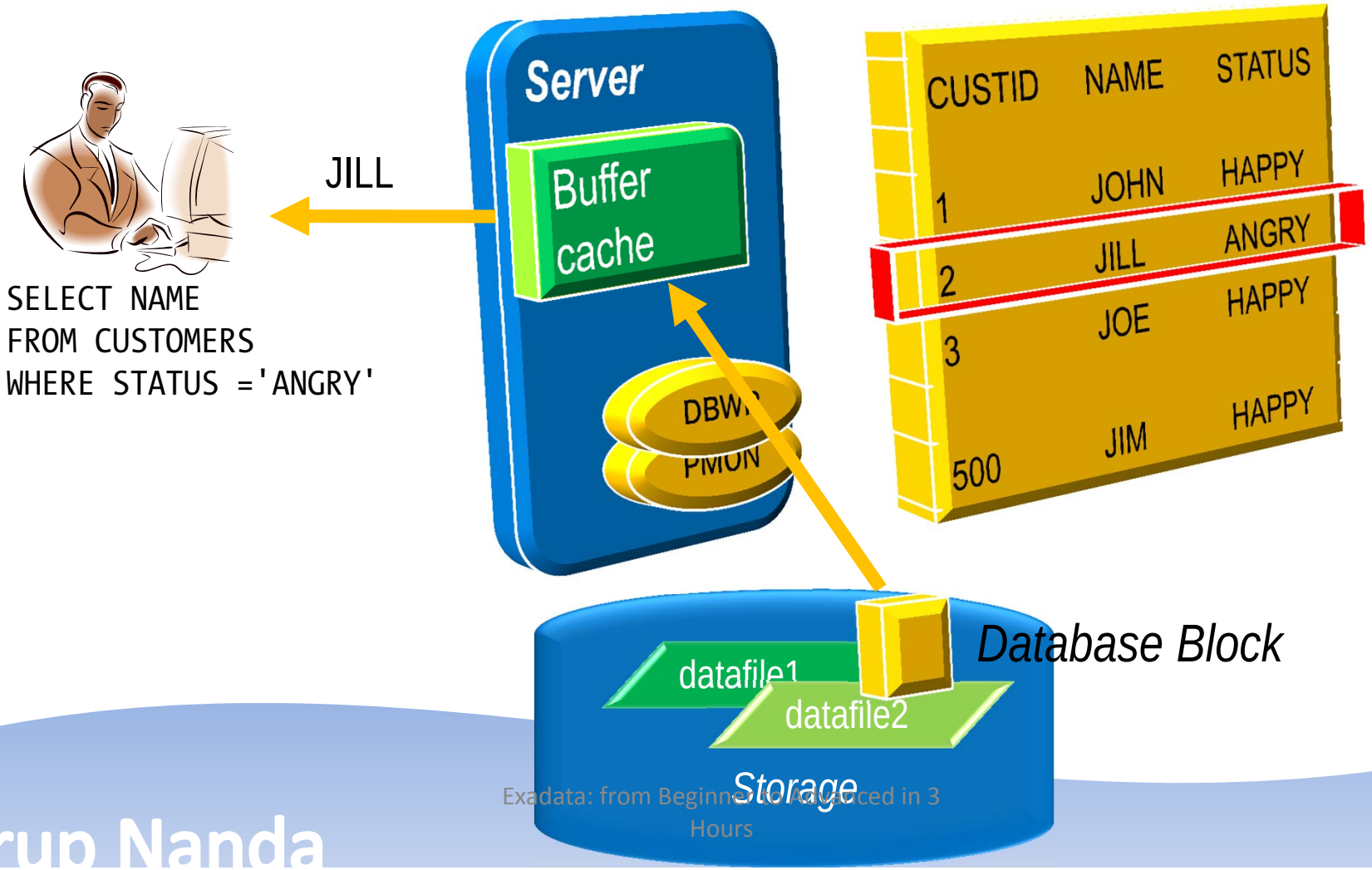
Instance

Combination of
- Memory Areas
- Background Processes

RAC Database



Query Processing



Components for Performance

- CPU
- Memory
- Network
- I/O Controller
- Disk

Less I/O = better performance

What about SAN Caches?

- Success of SAN caches is built upon predictive analytics
- They work well, if a small percentage of *disk* is accessed most often
  - The emphasis is on *disk*; not *data*
- Most database systems
  - are way bigger than caches
  - need to get the data to the memory to process

    --> I/O at the disk level is still high
- Caches are excellent for filesystems

  ☞ or very small databases

What about In-Memory DBs

- Memory is still more expensive
- How much memory is enough?
- You have a 100 MB database and 100 MB buffer cache
- The whole database will fit in the memory, right?
- NO!
- Oracle database fills up to 7x DB size buffer cache

http://arup.blogspot.com/2011/04/can-i-fit-80mb-database-completely-in.html

The Solution

- A typical query may:
  - Select 10% of the entire storage
  - Use only 1% of the data it gets
- To gain performance, the DB needs to shed weight
- It has to get less from the storage
  - ☞ Filtering at the storage level
  - ☞ The storage must be cognizant of the data

```
SELECT NAME
FROM CUSTOMERS
WHERE STATUS ='ANGRY'
```

Filtering should be Applied Here

CPU

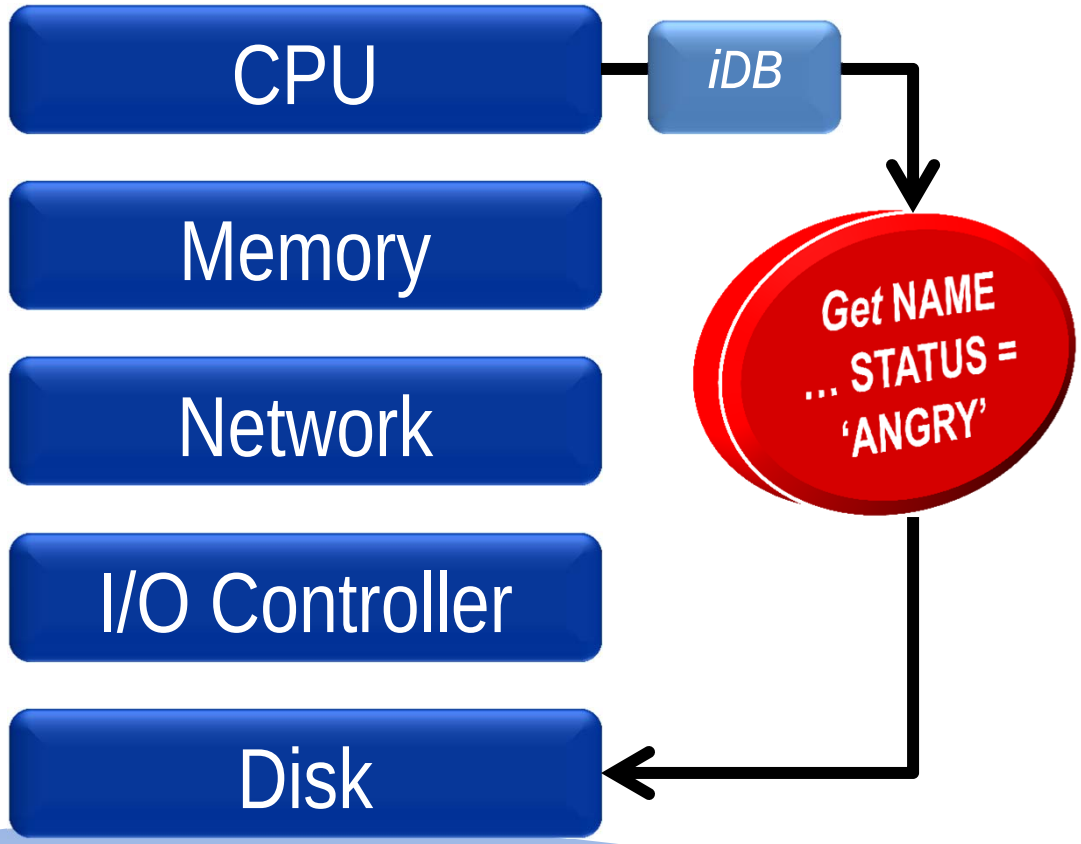
Memory

Network

I/O Controller

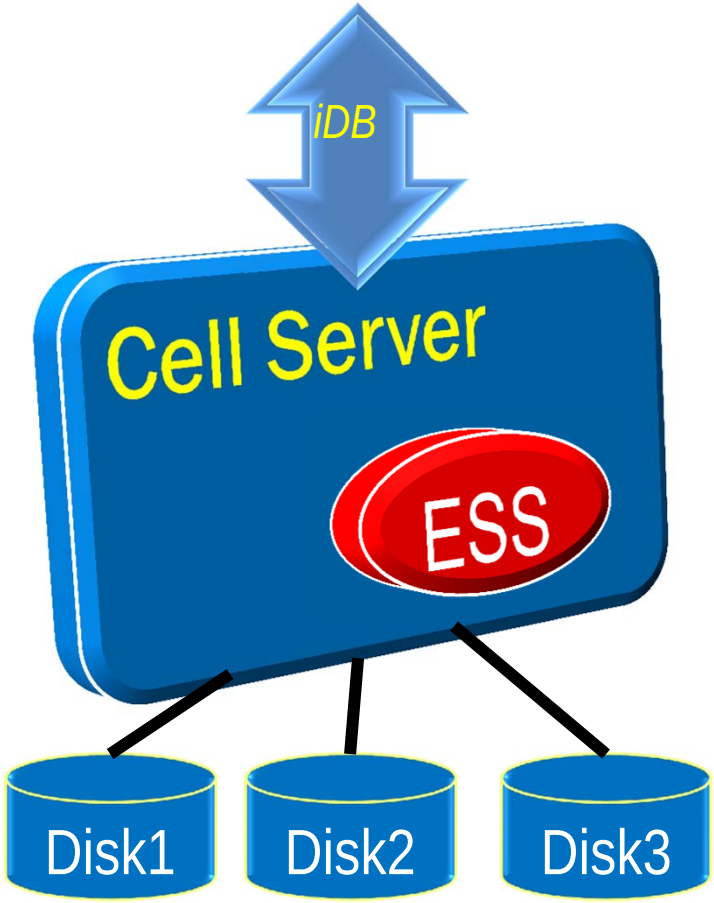
Disk

The Magic #1



The communication between CPU and Disk carries the information on the query – columns and predicates. This occurs as a result of a special protocol called iDB.

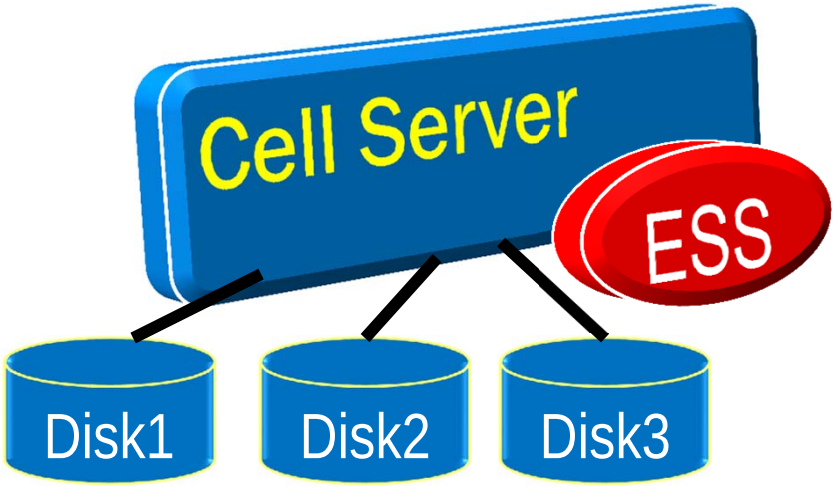
Magic #2 Storage Cell Server



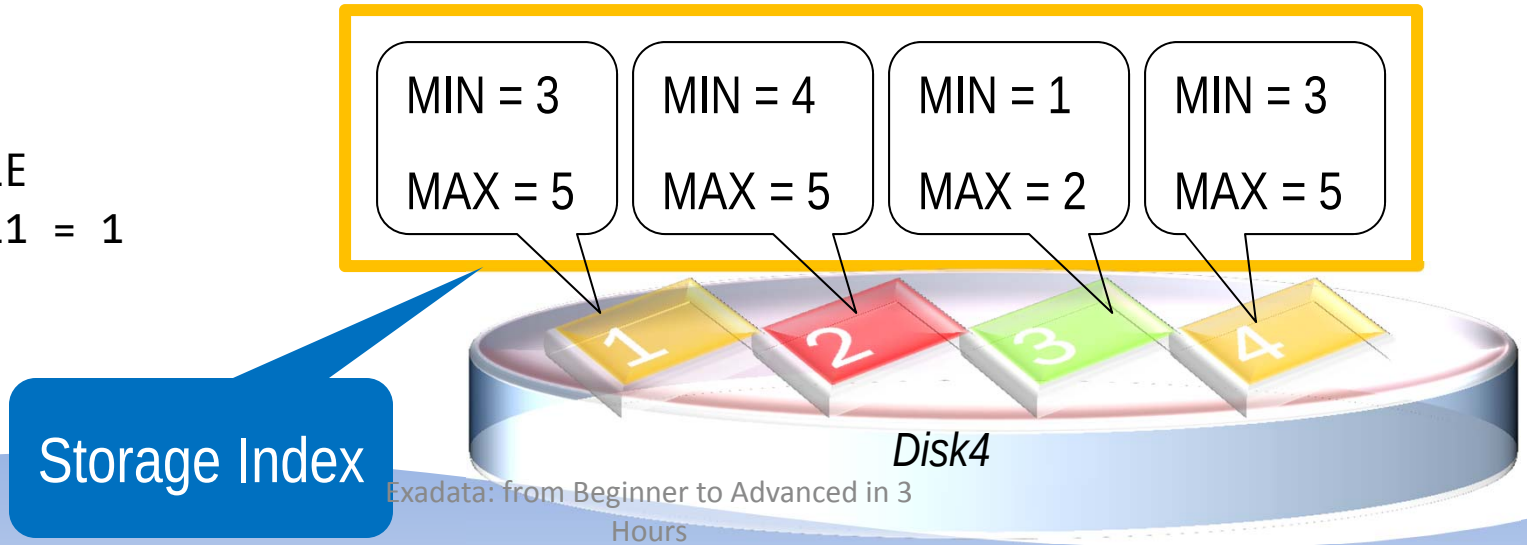
- Cells are Sun Blades
- Run Oracle Enterprise Linux
- Software called Exadata Storage Server (ESS) which understands iDB

Magic #3 Storage Indexes

Storage Indexes store in memory of the Cell Server the areas on the disk and the MIN/MAX value of the column and whether NULL exists. They eliminate disk I/O.




```
SELECT …
FROM TABLE
WHERE COL1 = 1
```



Checking Storage Index Use

```
select name, value/1024/1024 as stat_value
from v$mystat s, v$statname n
where s.statistic# = n.statistic#
and n.name in (
  'cell physical IO bytes saved by storage index',
  'cell physical IO interconnect bytes returned by smart
  scan’)
```

Output

```
STAT_NAME  STAT_VALUE
---------- ----------
SI Savings    5120.45
Smart Scan    1034.00
```

Checking Offloading of an SQL

```
select
        sql_id,
        child_number child#,
        plan_hash_value plan_hash,
        executions execs,
        (elapsed_time/1000000)/decode(nvl(executions,0),0,1,executions)/

decode(px_servers_executions,0,1,px_servers_executions/decode(nvl(executions,0),0,1,
executions)) avg_elapsed_time_in_secs,
        px_servers_executions/decode(nvl(executions,0),0,1,executions) avg_par_deg,
        decode(io_cell_offload_eligible_bytes,0,'No','Yes') Offloaded,

decode(io_cell_offload_eligible_bytes,0,0,100*(io_cell_offload_eligible_bytes-
io_interconnect_bytes)

/decode(io_cell_offload_eligible_bytes,0,1,io_cell_offload_eligible_bytes)) "%age IO
Saved",
        buffer_gets/decode(nvl(executions,0),0,1,executions) avg_lio
from v$sql
where sql_text like <SQL Statement Comes Here>%'
```

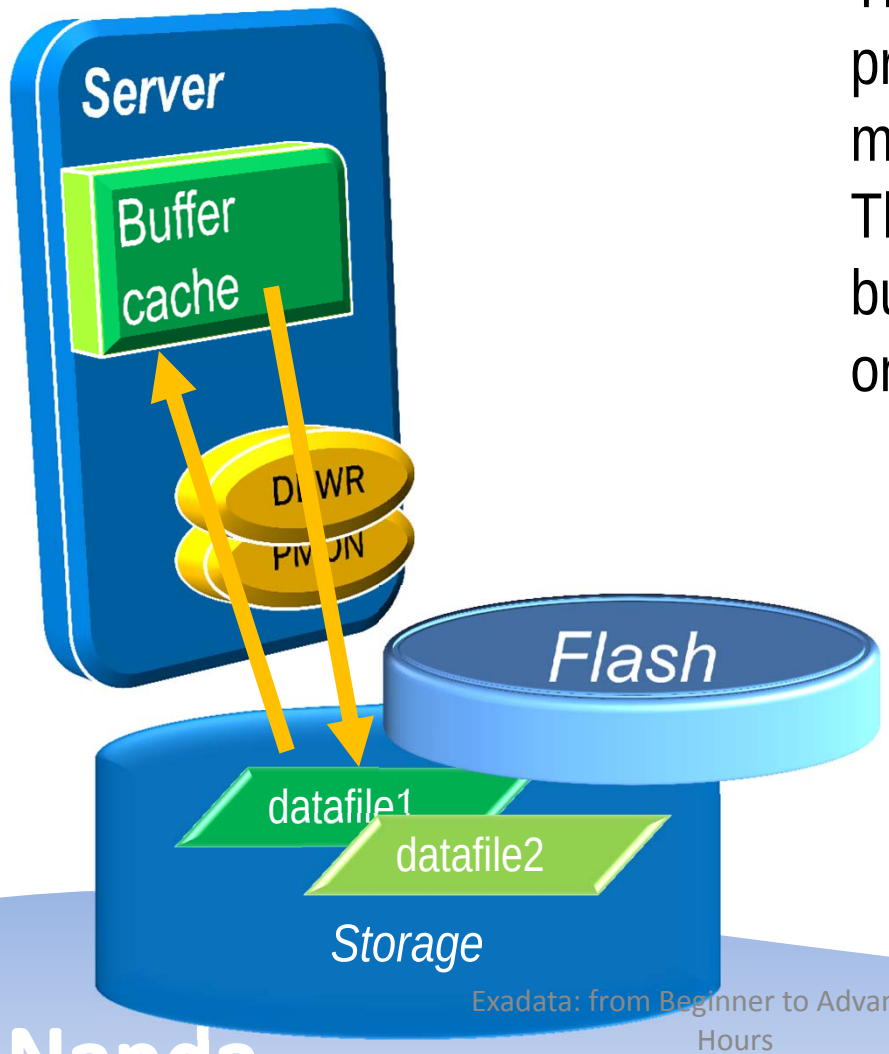
Why Not?

- Pre-requisite for Smart Scan
  - Direct Path
  - Full Table or Full Index Scan
  - > 0 Predicates
  - Simple Comparison Operators
- Other Reasons
  - Cell is not offload capable
    - The diskgroup attribute cell.smart_scan_capable set to FALSE;
  - Not on clustered tables, IOTs, etc.

Disabling Smart Scans

```
cell_offload_processing =
false;
_kcfis_storageidx_disabled =
true;
```

Magic #4 Flash Cache



These are flash cards presented as disks; not memory to the Storage Cells. They are similar to SAN cache; but Oracle controls what goes on there and how long it stays.

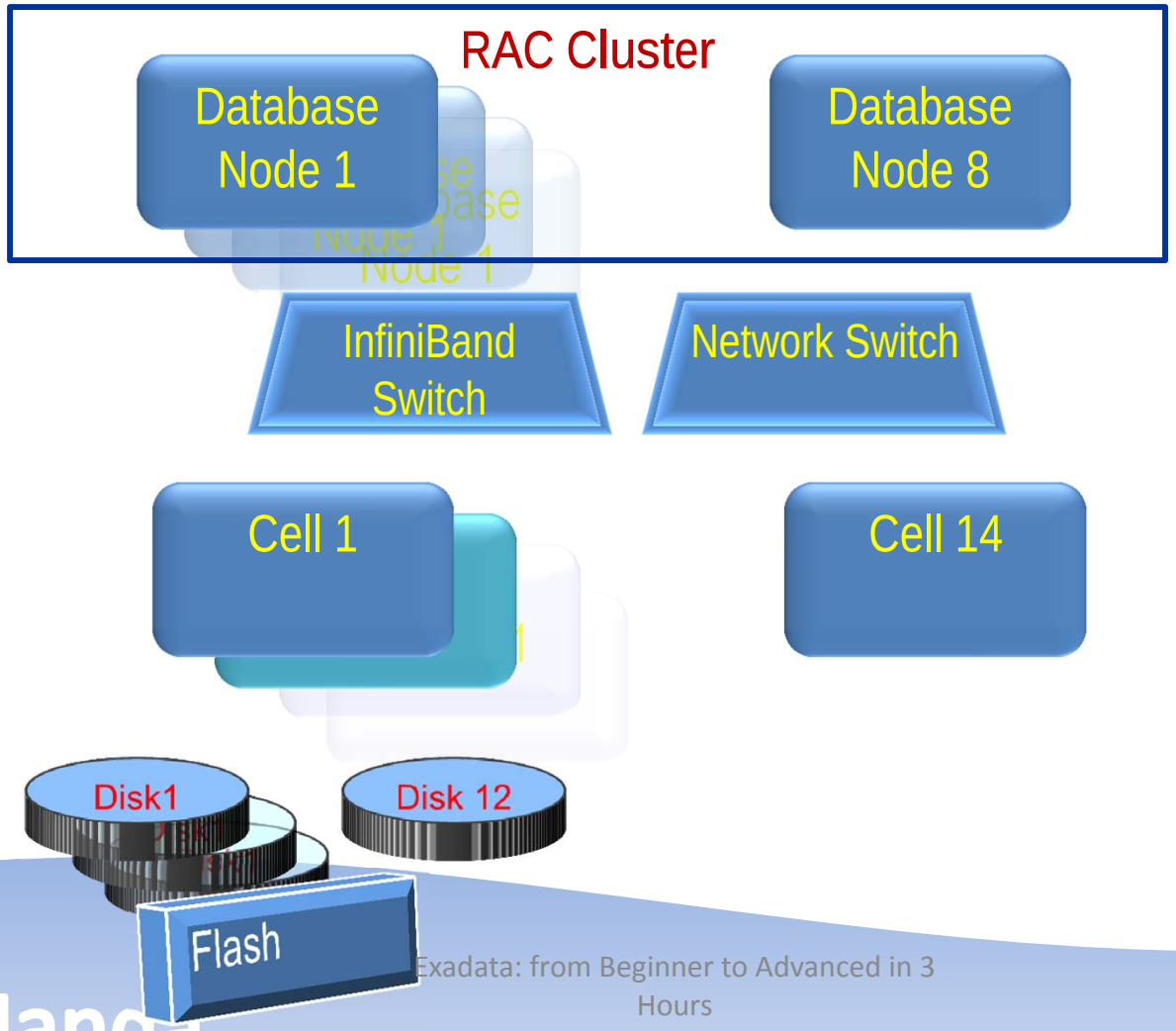
Magic #5 Process Offloading

- Bloom Filters
- Functions Offloading
  - Get the functions that can be offloaded
    - `V$SQLFN_METADATA`
- Decompression
  - (Compression handled by Compute Nodes)
- Virtual Columns

Components

CPU Memory	*Database Node* (Sun Blade. OEL) Oracle 11gR2 RAC
Network	*InfiniBand Switch*
I/O Controller Disk	*Storage Cell* Exadata Storage Server Disks, Flash

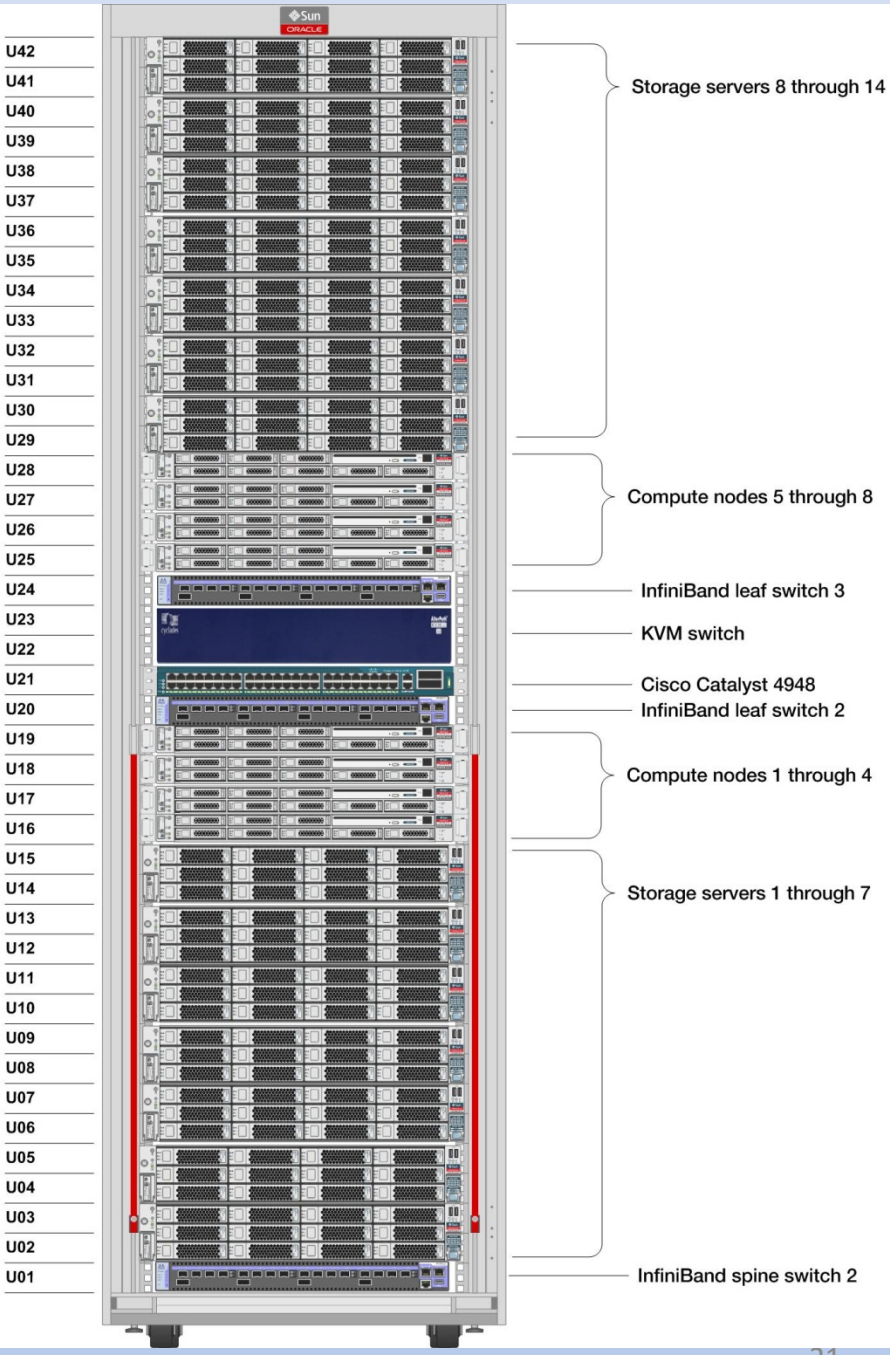
Put Together: One Full Rack



Clients connect to the database nodes.

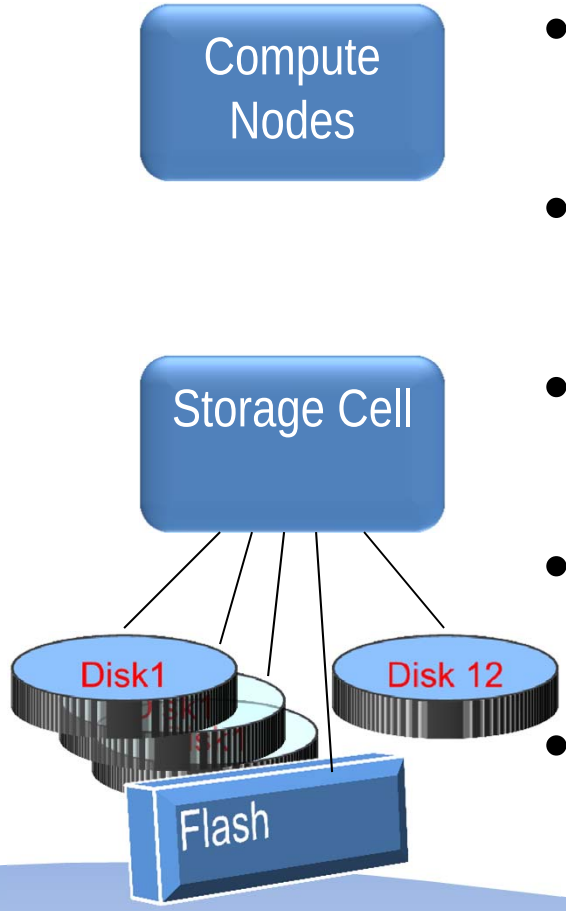
How it Looks

	X2-2 Qtr	X2-2 Half	X2-2 Full	X2-8 Full
Number of Compute Nodes	2	4	8	2
Total Compute Node Processor Cores	24	48	96	160
Total Compute Node Memory	196 GB	384 GB	768 GB	4 TB
Number of Storage Servers	3	7	14	14
Number of SAS Disks in Storage	36	84	168	168
Storage Capacity - HP	21.6 TB	50.4 TB	100.8 TB	100.8 TB
Storage Capacity - HC	108 TB	252 TB	504 TB	504 TB
Number of InfiniBand Switches	2	3	3	3



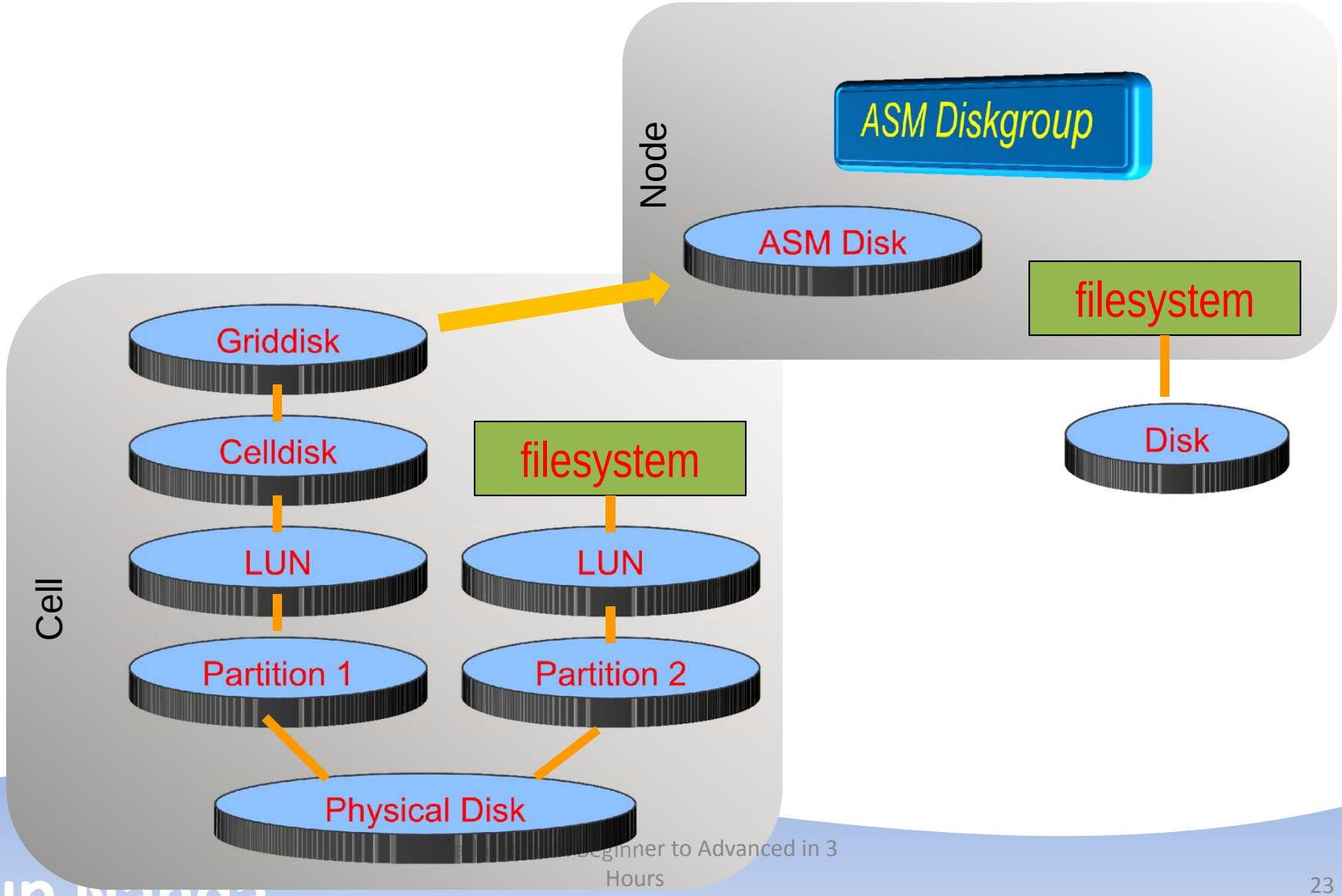
Source: upcoming book Exadata Recipes by Clarke from Apress

Disk Layout

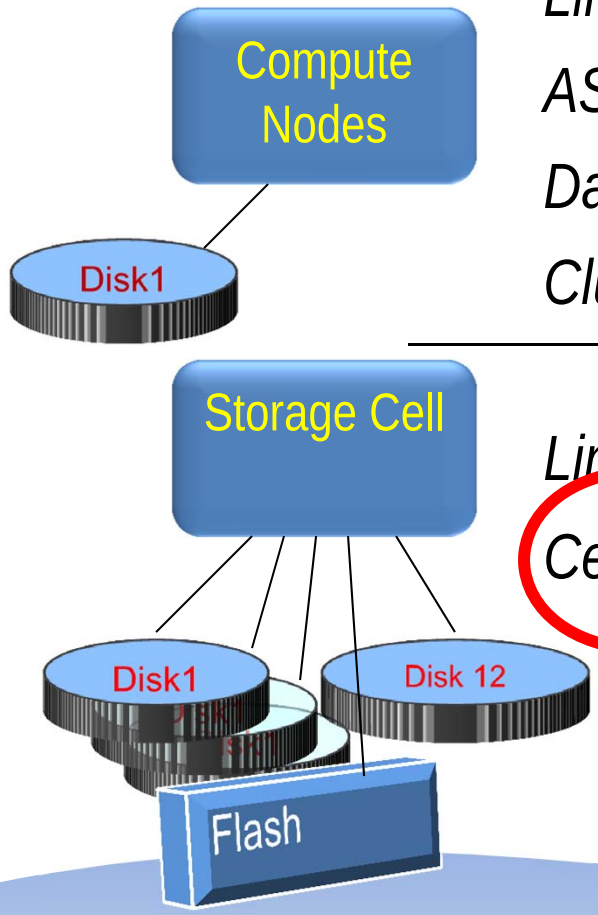


- Disks (hard and flash) are connected to the cells.
- The disks are partitioned at the cell
- Some partitions are presented as filesystems
- The rest are used for ASM diskgroups
- All these disks/partitions are presented to the compute nodes

Disk Presentation



Command Components



Linux Commands – vmstat, mpstat, fdisk, etc.

*ASM Commands – SQL*Plus, ASMCMD, ASMCA*

Database Commands – startup, alter database, etc.

Clusterware Commands – CRSCTL, SRVCTL, etc.

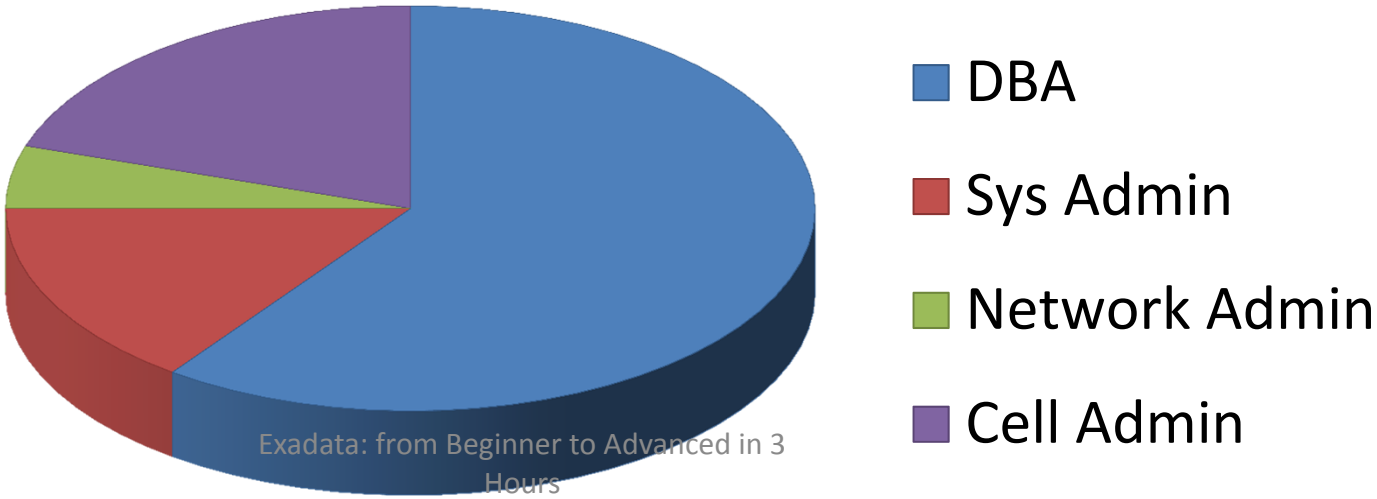
Linux Commands – vmstat, mpstat, fdisk, etc.

CellCLI – command line tool to manage the Cell

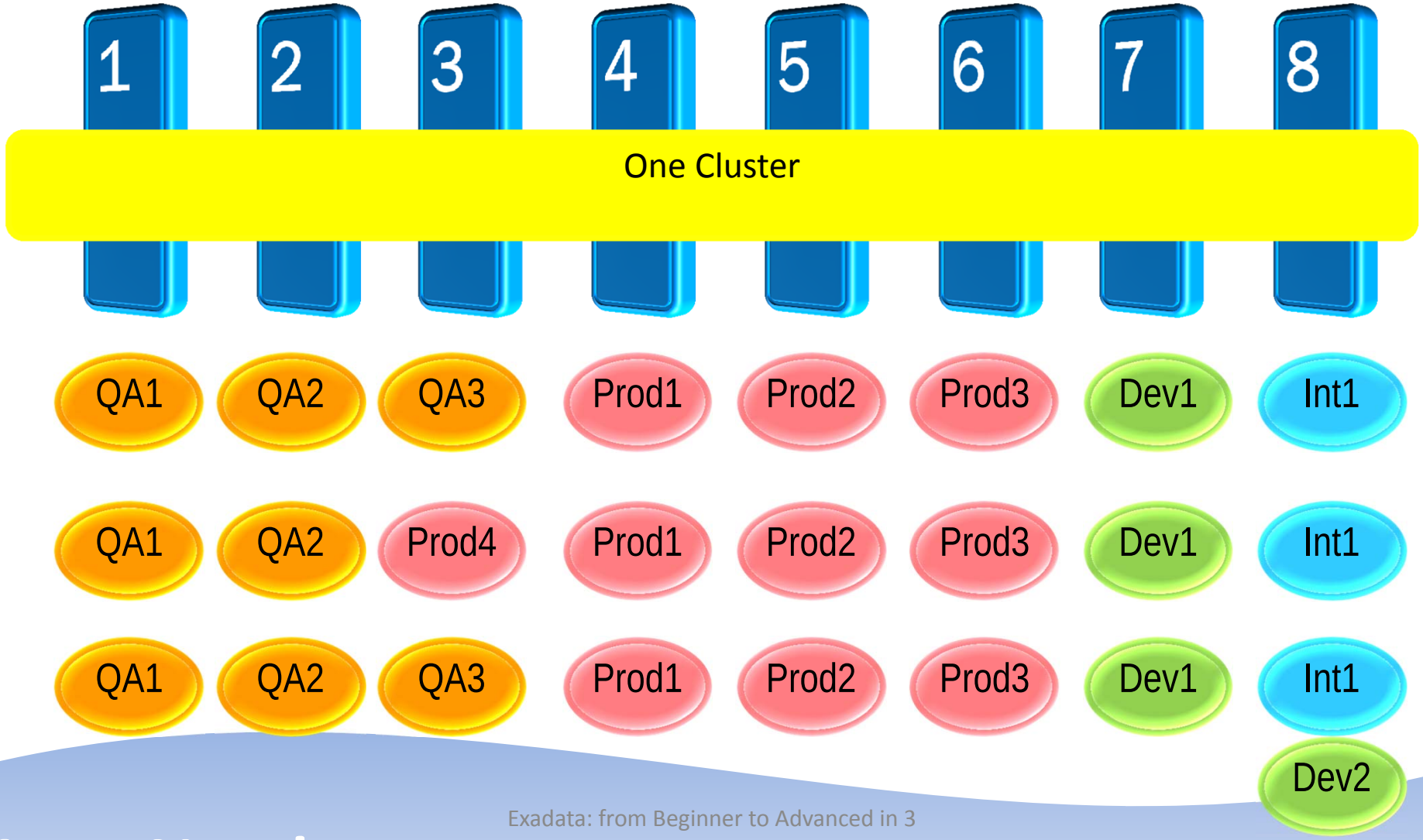
5-part Linux Commands article series
http://bit.ly/k4mKQS
4-part Exadata Command Reference article series
http://bit.ly/lljFl0

Administration Skills

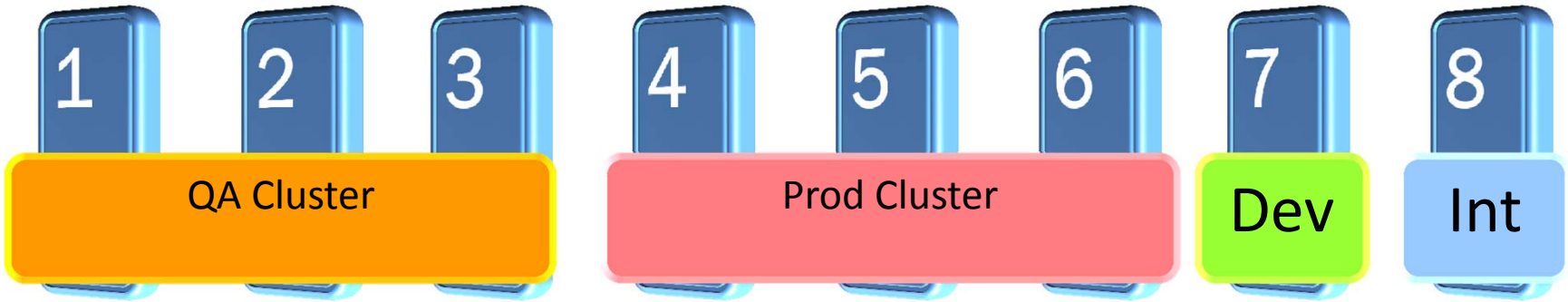
Skill	Needed
System Administrator	15%
Storage Administrator	0%
Network Administrator	5%
Database Administrator	60%
Cell Administration	20%

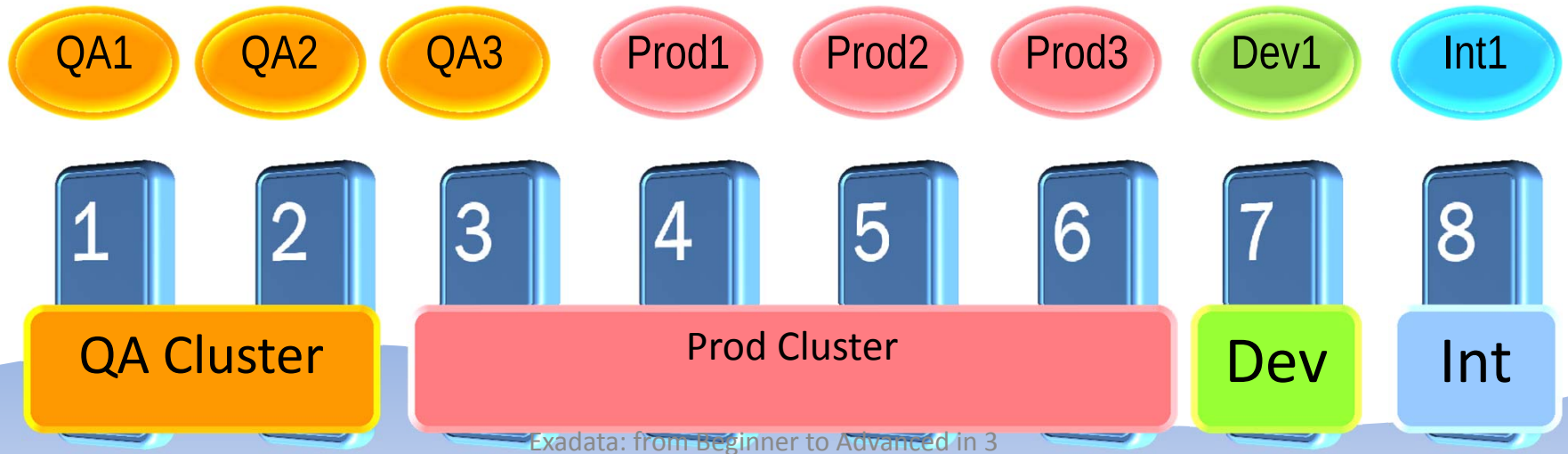


One Cluster?

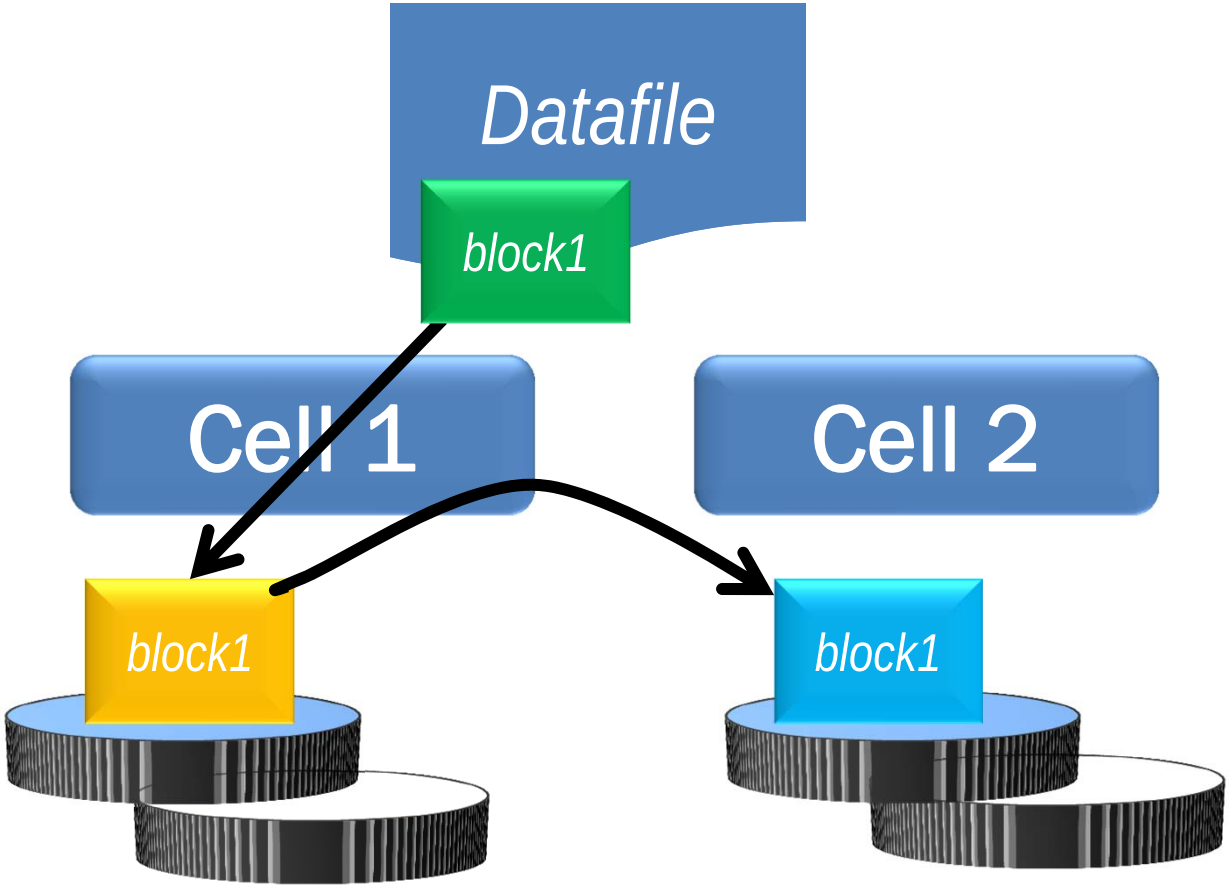

1
2
3
4
5
6
7
8
One Cluster
QA1
QA2
QA3
Prod1
Prod2
Prod3
Dev1
Int1
QA1
QA2
Prod4
Prod1
Prod2
Prod3
Dev1
Int1
QA1
QA2
QA3
Prod1
Prod2
Prod3
Dev1
Int1
Dev2

Many Clusters?

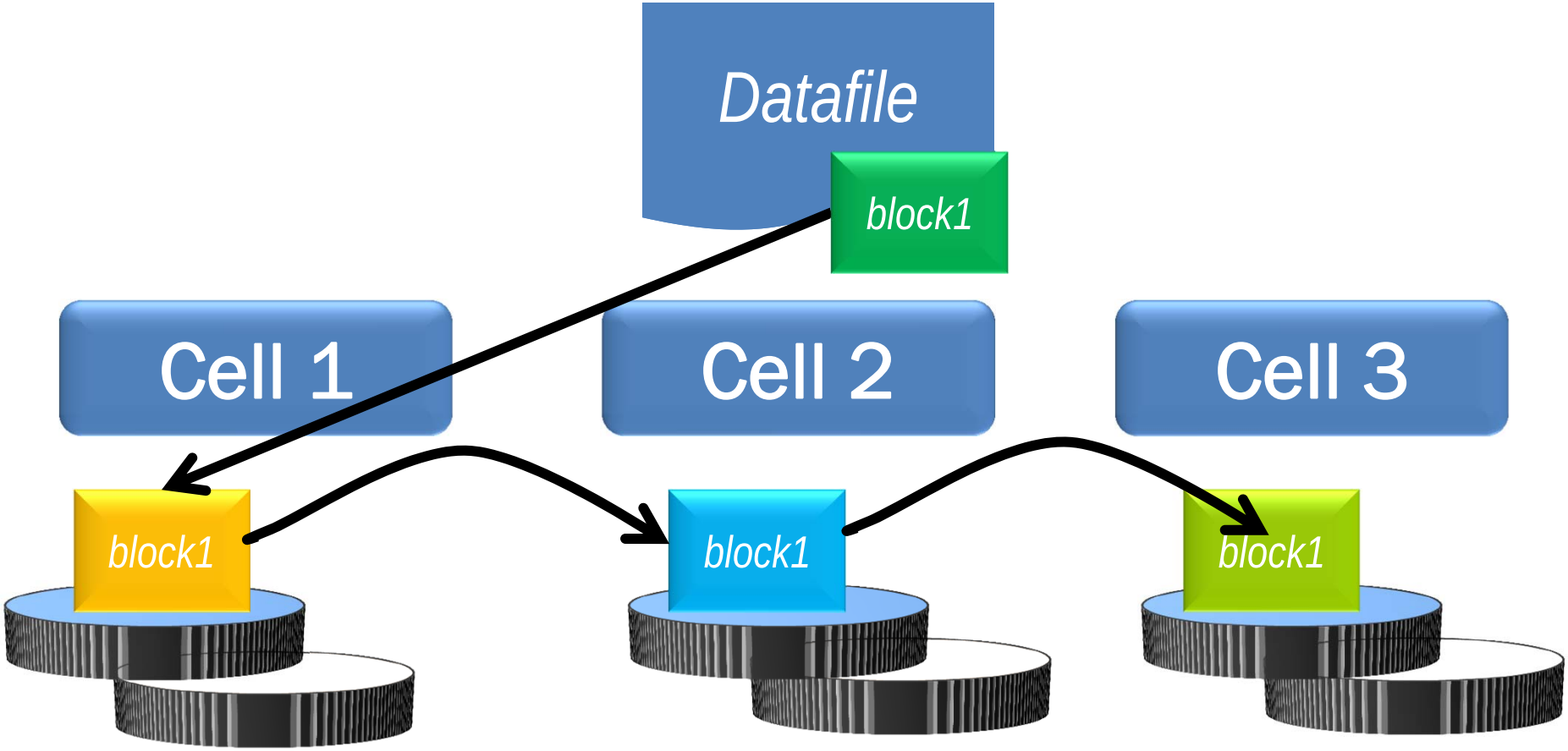




Disk Failures



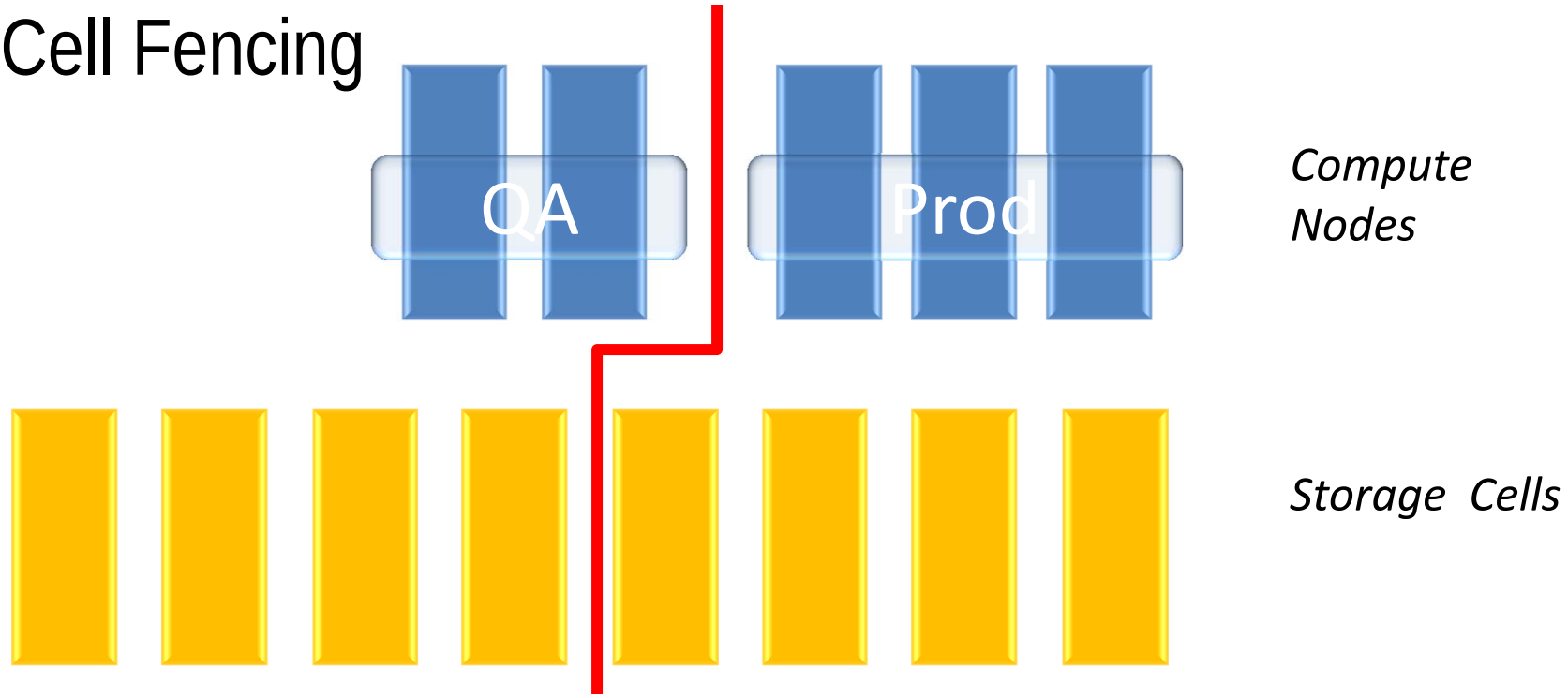
Disk Failures



High Redundancy

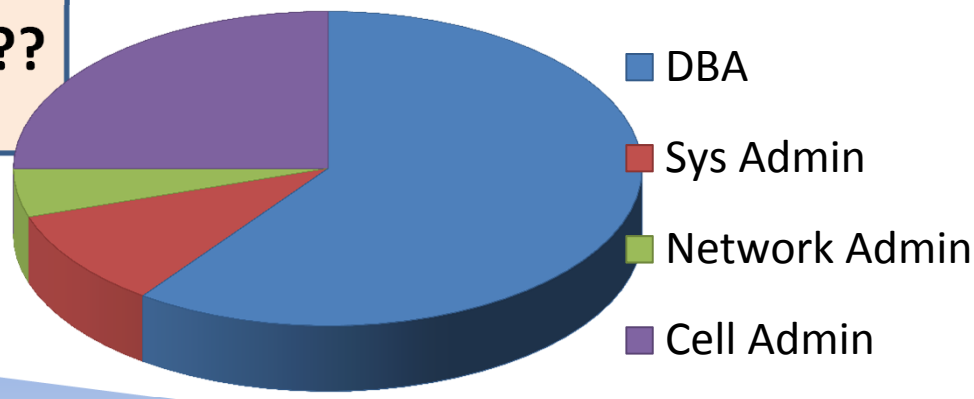
Playing Nice

- Database Resource Manager
- I/O Resource Manager
- Cell Fencing

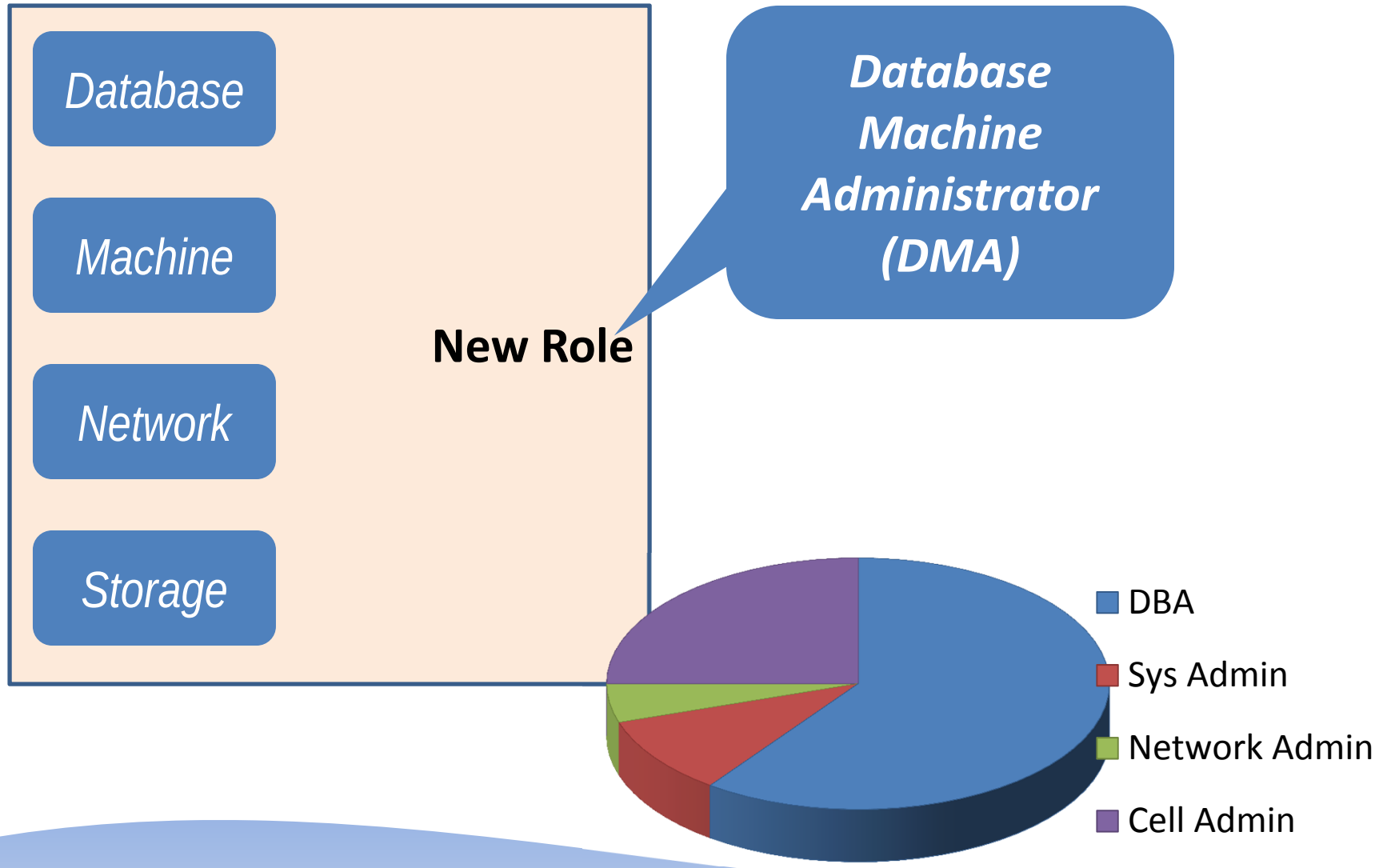


Divide and Conquer

Database	**DBA**
Machine	**System Admin**
Network	**Network Admin**
Storage	**??**



Combined Skills



Other Questions

Q: Do clients have to connect using Infiniband?

A: No; Ethernet is also available

Q: How do you back it up?

A: Normal RMAN Backup, just like an Oracle Database

Q: How do you create DR?

A: Data Guard is the only solution

Q: Can I install any other software?

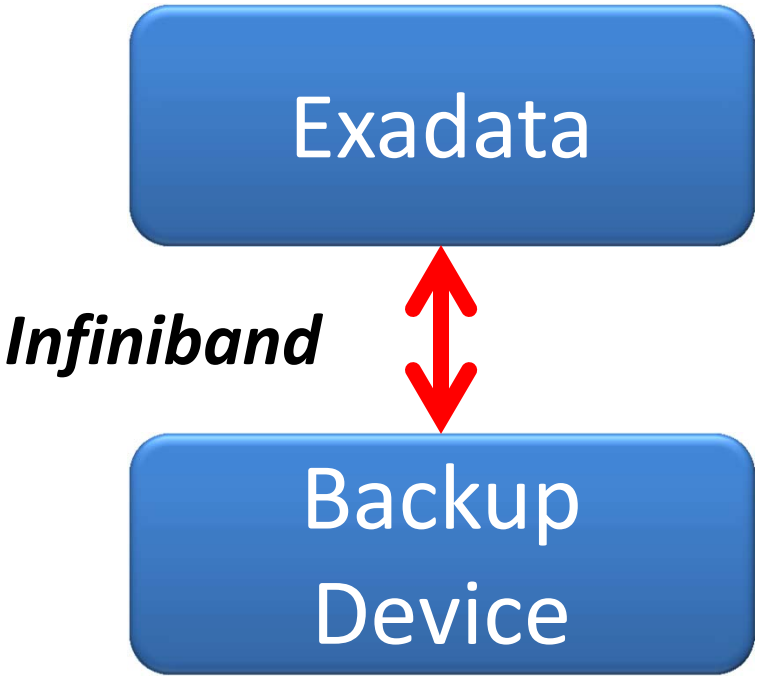
A: Nothing on Cells. On nodes – yes

Q: How do I monitor it?

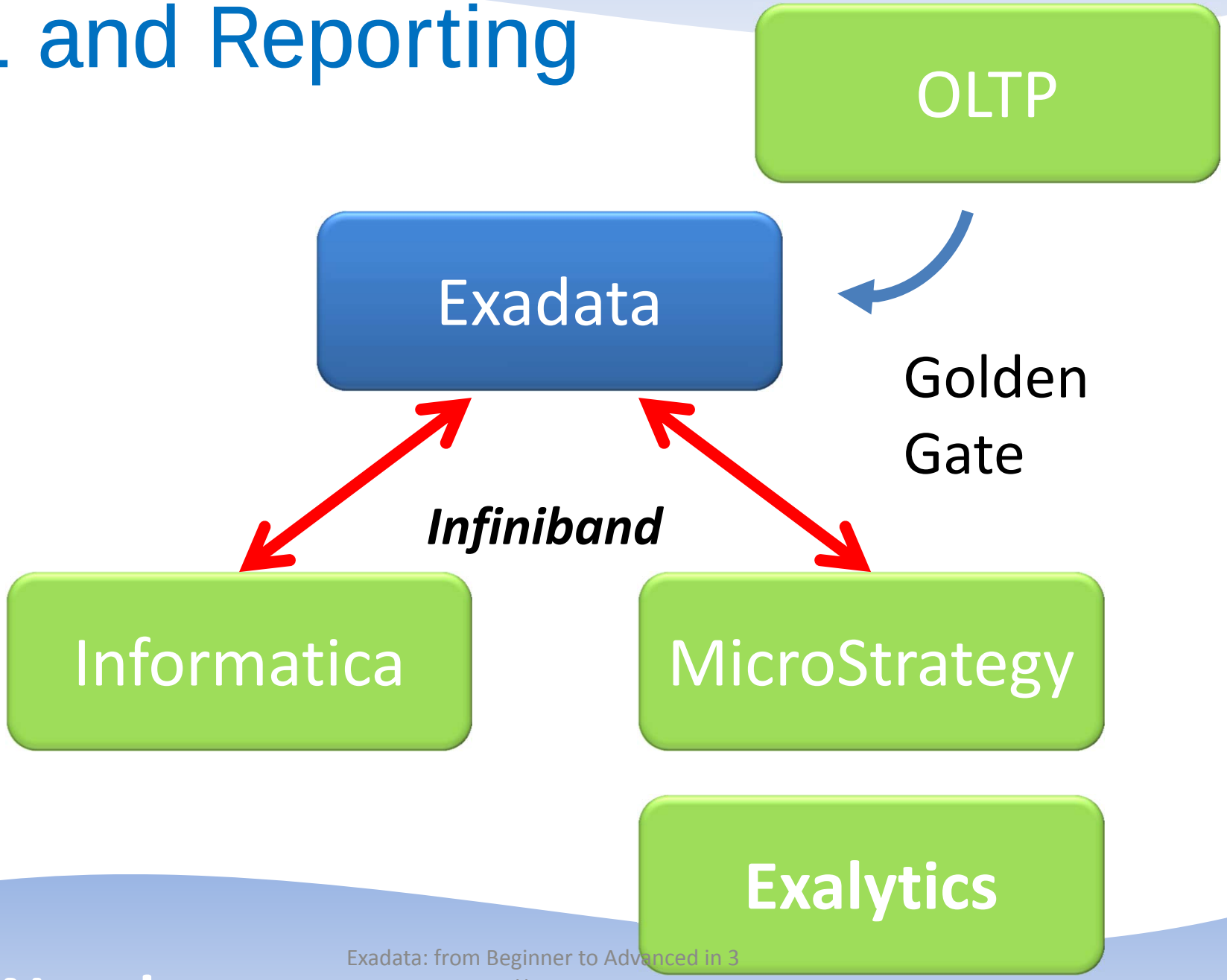
A: Enterprise Manager, CellCLI, SQL Commands

Backup and DR

- No SAN connectivity
- Only NAS
  - Infiniband
  - Tape , Disk Pool
- DR
  - No Storage Level Replication
  - Only Data Guard
  - Supplemental Logging
  - Force Logging
  - http://www.oracle.com/technetwork/database/features/availability/maa-wp-dr-dbm-130065.pdf
- Golden Gate



ETL and Reporting



Overall Activities

- Physical Aspects
  - Delivery, power, network components, etc.
- Layout Planning
- Installation and Configuration
- Data Migration
- Administration
  - Who manages it
  - Backup and Disaster Recovery
- Application Development

Power Delivery Units

- Over or Under the unit
- Power Requirement
  - Single-Phase Low Voltage Americas / Japan / Taiwan)
  - Single-Phase High Voltage (EMEA & APAC (excluding Japan / Taiwan)
  - Three-Phase Low Voltage (Americas / Japan / Taiwan)
  - Three-Phase High Voltage (EMEA & APAC (excluding Japan / Taiwan)

Network Ports

- NET0
  - Admin Interface
- NET1, NET2
  - Network Access to Nodes
- NET3
  - Backup Network
- IB
  - Infiniband Network
  - IP Addr: Qtr Rack: 5; Half Rack: 11; Full rack: 22

Installation Activities

1. Configuration Worksheet
2. Pre-delivery Survey
3. Generate config files
4. Run checkip.sh

ACS

5. Power on and validate components
6. Configure KVM
7. Configure IB
8. Configure Cisco Switch

Oracle HW

Installation, contd.

9. Configure IP to PDUs
10. Validate Storage Cells
11. Validate Compute Nodes

Oracle HW

12. Config files from USB
13. Firstboot and applyconfig.sh
14. Stage Oracle Software on Node 1
15. Run OneCommand

ACS

Summary

- Exadata is an Oracle Database running 11.2
- The storage cells have added intelligence about data placement
- The compute nodes run Oracle DB and Grid Infra
- Nodes communicate with Cells using iDB which can send more information on the query
- Smart Scan, when possible, reduces I/O at cells even for full table scans
- Cell is controlled by CellCLI commands
- DMA skills = 60% RAC DBA + 15% Linux + 20% CellCLI + 5% miscellaneous

Resources

- My Articles
  - 5-part Linux Commands article series http://bit.ly/k4mKQS
  - 4-part Exadata Reference article series http://bit.ly/lljFl0
- OTN Page on Exadata
  - http://www.oracle.com/technetwork/database/exadata/index.html
- Tutorials
  - http://www.oracle.com/technetwork/tutorials/index.html
- OTN Exadata Forum
  - https://forums.oracle.com/forums/forum.jspa?forumID=829
- Exadata SIG
  - http://www.linkedin.com/groups?home=&gid=918317

Thank You!

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