Oracle Exadata Database Machine v2

Hüsnü Şensoy
Global Maksimum Data & Information Technologies
Founder, VLDB Expert
husnu.sensoy@globalmaksimum.com

ORACLE ACE Director

Agenda

- Introduction
  - Why do we need Exadata ?
- Understanding Exadata v2
  - Exadata v2 Hardware
  - Exadata v2 Software
- Conclusion

husnu.sensoy@globalmaksimum.com

Global Maksimum & Exadata v2

- Only company in Turkey having IB interconnected RAC 11g implementation experience on Linux x86-64bit.
- Only company in Turkey having sufficient consultancy experience (more than 120 TB conventional system data) on Exadata v2
  - Physical & Architecture Design
  - Migration
  - Performance Optimization
  - Backup & Recovery Architectures Design
- Trains customers, Oracle partners, and Oracle employees all over the Europe
- Strong joint relation with Oracle Platinum Partners, Oracle Development Team Head Office, and IB technology leaders.
- **X-Migrator** service provider for high capacity customers.

husnu.sensoy@globalmaksimum.com

Oracle Exadata Database Machine v2

Introduction

husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

Exadata for All

- Engineers
  - To learn that «The mechanic with a hammer thinks that all problems are nail»
- Customers
  - Shorter setup time
- Non-Exadata Customers
  - More stable Oracle releases
- Oracle
  - Easy to manage/standardize its code repository

husnu.sensoy@globalmaksimum.com

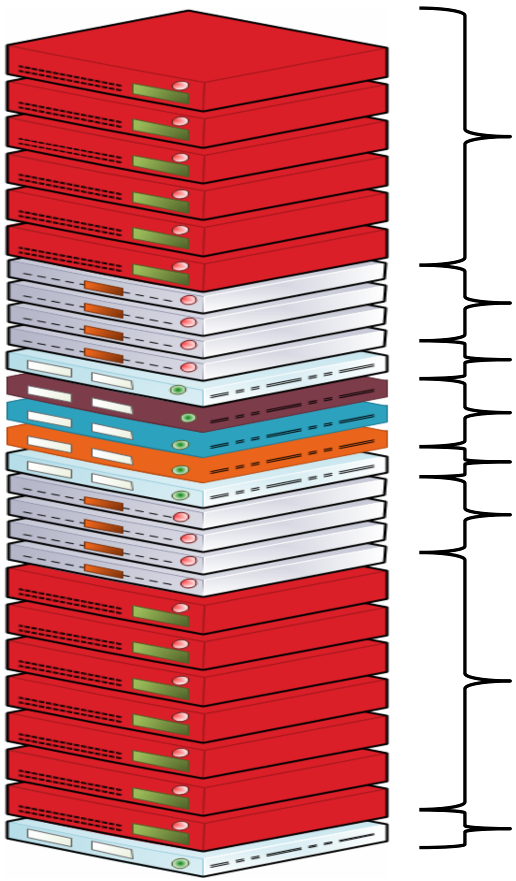
Oracle Exadata Database Machine v2

Exadata v2 Hardware

husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

Exadata v2 from 10.000 ft



48-port Gigabit Ethernet Switch

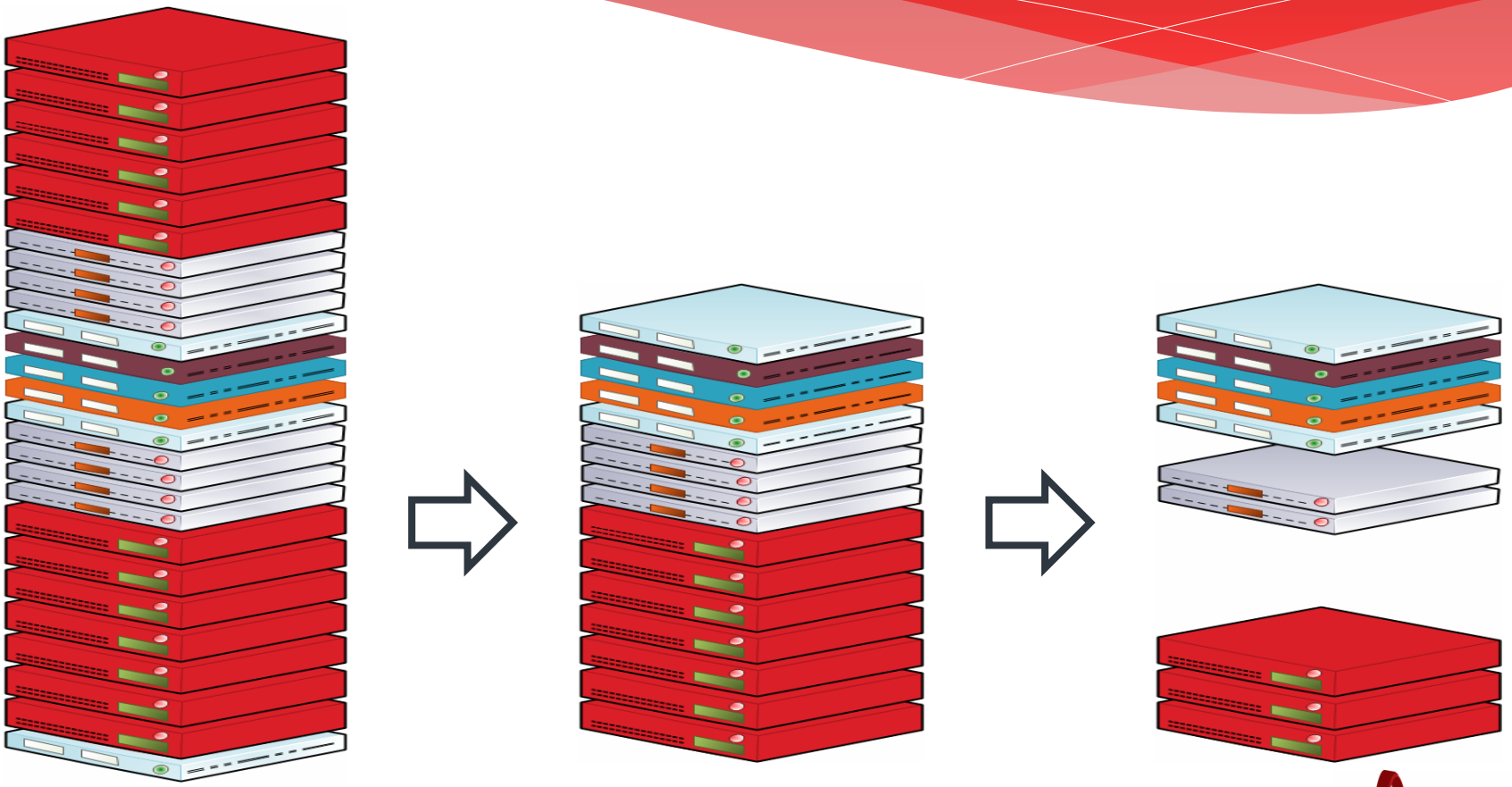
Rackmount KMM Keyboard with TFT monitor

Sun Datacenter 36-port Managed QDR IB Switched

KVM IP Console Switch

husnu.sensoy@globalmaksimum.com

Full → Half → Quarter



husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

Capacity & Performance

- 28 TB of User Data
- 5.3 TB of RAW Flash
- 21 GB/s
- 1.000.000 IOPS

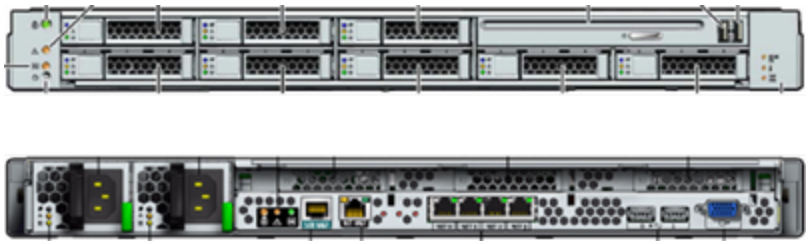
- 14 TB of User Data
- 2.6 TB of RAW Flash
- 10.5 GB/s
- 500.000 IOPS

- 6 TB of User Data
- 1.1 TB of RAW Flash
- 4.5 GB/s
- 225.000 IOPS

husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

Sun Fire™ X4170 as RAC Node



- 2 socket Quad Core
  - 2.53 GHz
  - 2 Hyper-Threads
  - So, `CPU_COUNT=16`
- 18 DDR3 DIMM Slots
  - 72 GB@800 MHz (2x3x3x4 GB)
- 4 10/100/1000Base-T Ethernet ports
  - **NET0** : Management
  - **NET1** : Public Network
  - **NET2** : Public Network
  - **NET3** : -

husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

Sun Fire™ X4275 as Storage Node

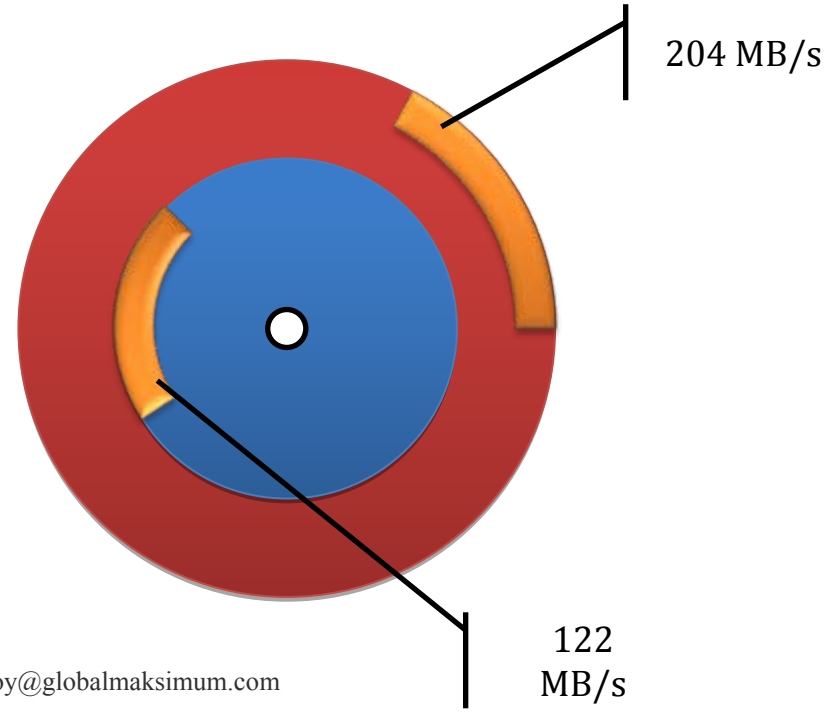
- 2 socket Quad Core
  - 2.53 GHz
  - 2 Hyper-Threads
- 6 DDR3 DIMM Slots
  - 24 GB@1066 MHz (2x3x1x4 GB)
- HDD Storage
  - 12x600 GB 15K RPM SAS disks
  - 12x2 TB 7.2K RPM SATA disks
- 4 Sun Flash Accelerator F20 PCIe Cards

husnu.sensoy@globalmaksimum.com

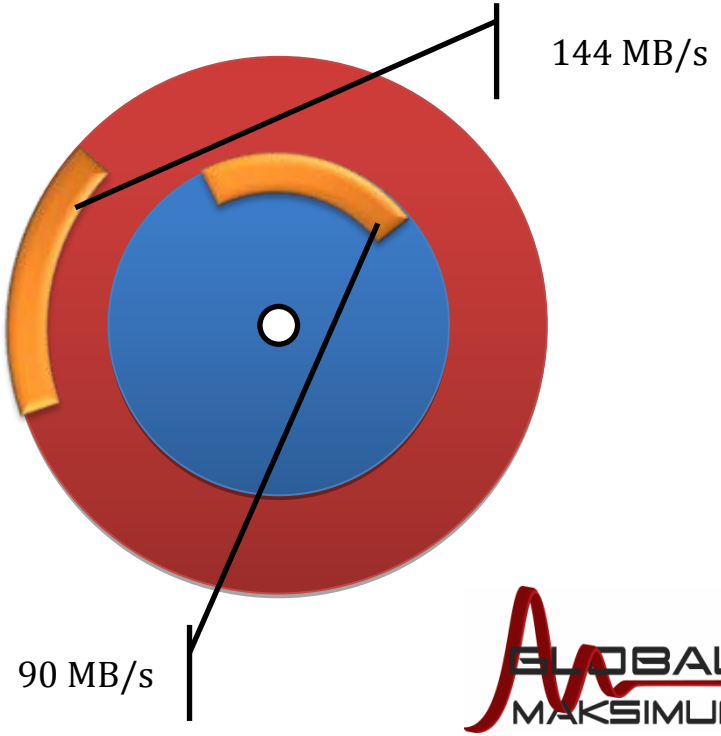
GLOBAL MAKSIMUM

HDD Sequential Read Performance

15K RPM SAS



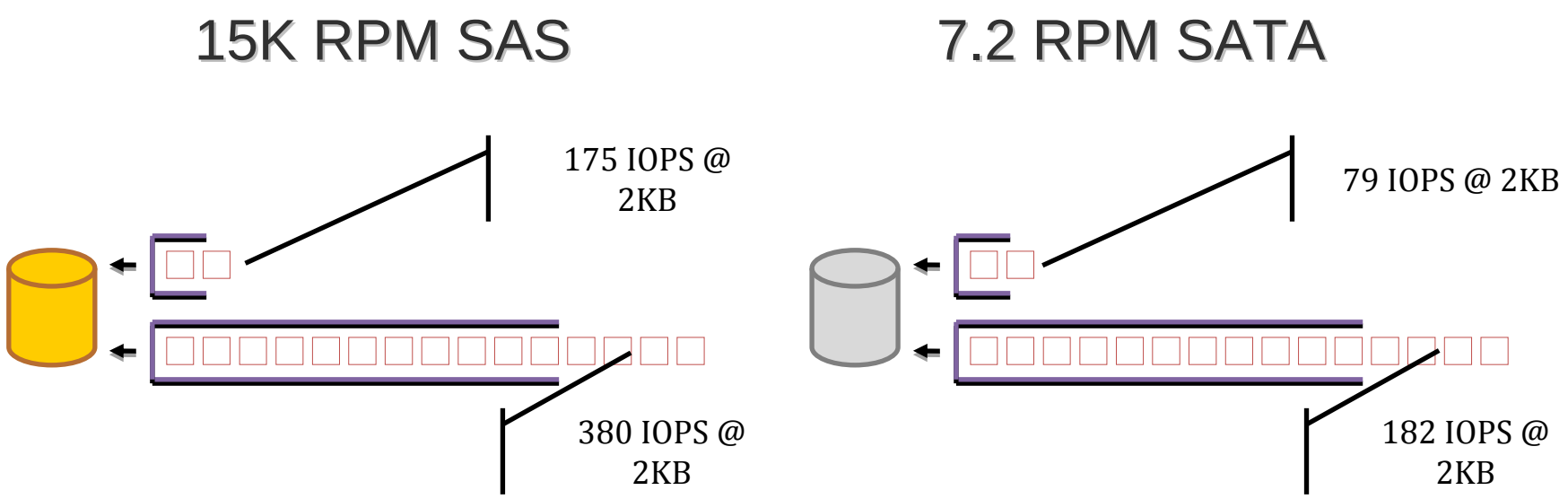
7.2 RPM SATA



husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

HDD Random Read Performance

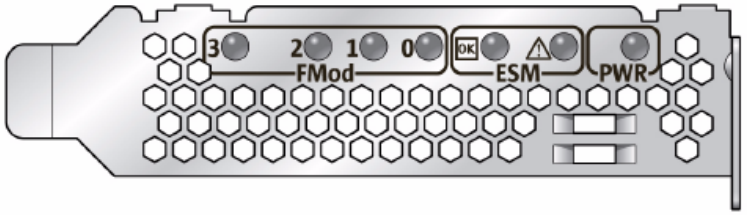


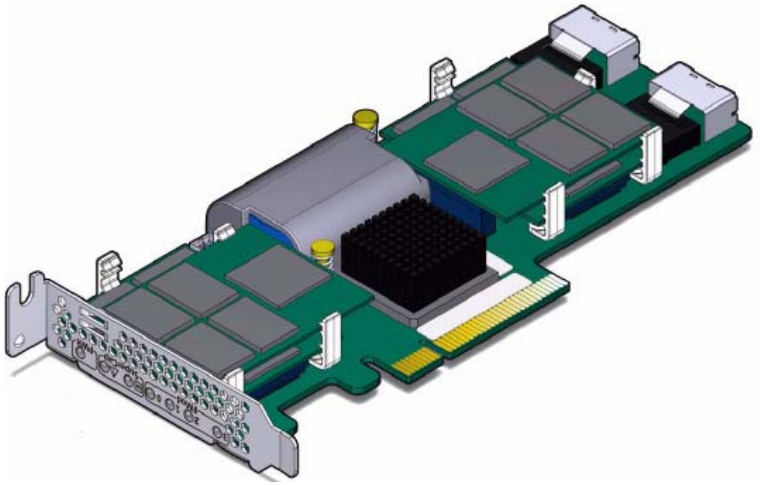
husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM

F20 PCIe Card

- Neither SATA nor SAS interface SSD driver. But a PCIe card having a embedded SAS controller managing 4 Solid State Flash Disk Modules (*FMod*) each of 24 GB size.
  - Embedded controller will expose 16 (`4 cards x 4 FMod`) Linux devices.
    - `/dev/sdn`
- SuperCap Power Reserve (**E**nergy**S**torage**M**odule) provides write-back operation mode.
  - ESM should be enabled for optimal write performance
  - Should be replaced in every two years.
  - Can be monitored using various tools like ILOM
- 4K sector boundary for *Fmods*
- Each *FMod* consists of several NAND modules best performance can be reached with multithreading (32+ thread/*FMod* etc)





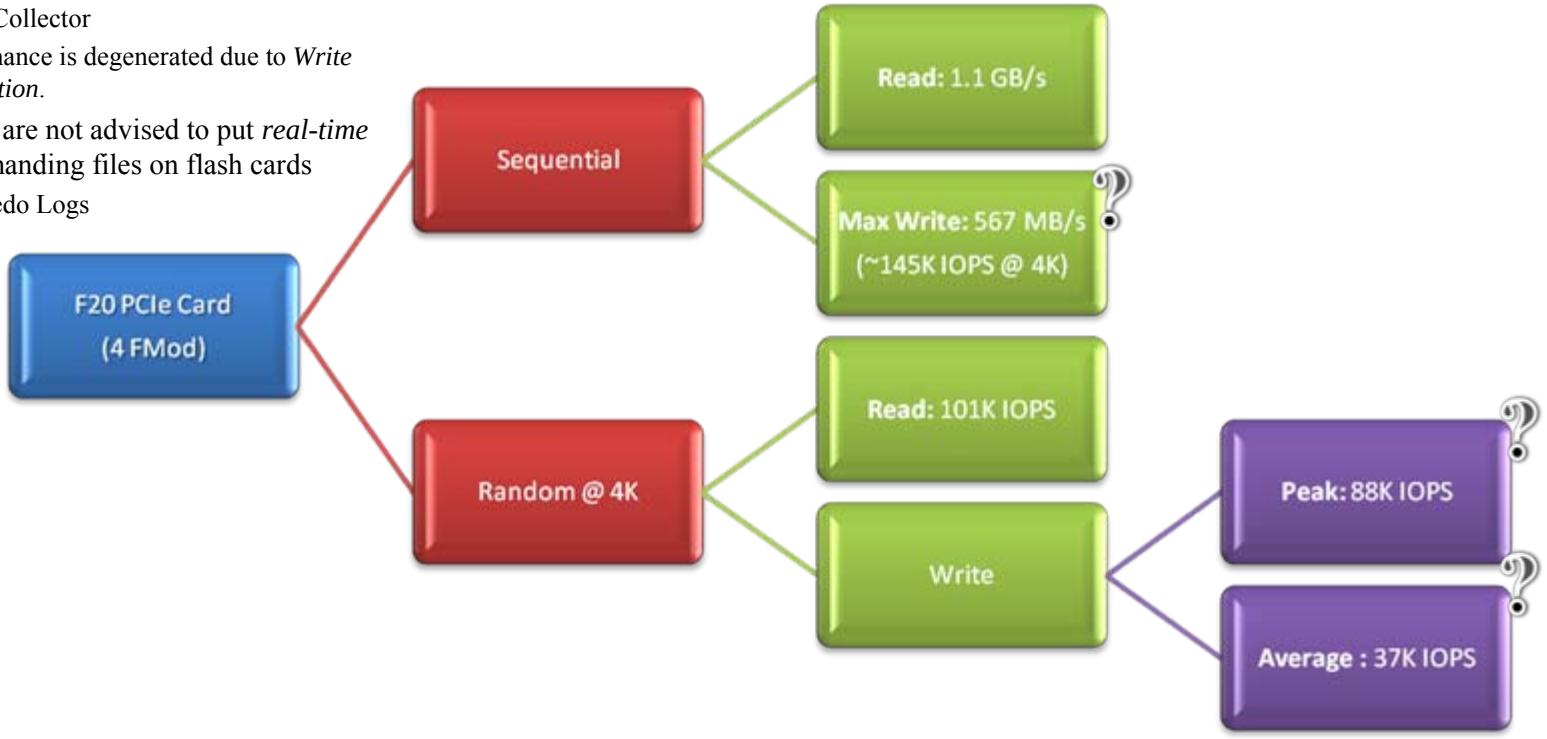
husnu.sensoy@globalmaksimum.com

F20 Performance

Random Write Performance Degeneration

- As the flash cache get full (sustained write)
  - Wear Leveling
  - SLC Update Mechanism : Delete + Write
  - Garbage Collector

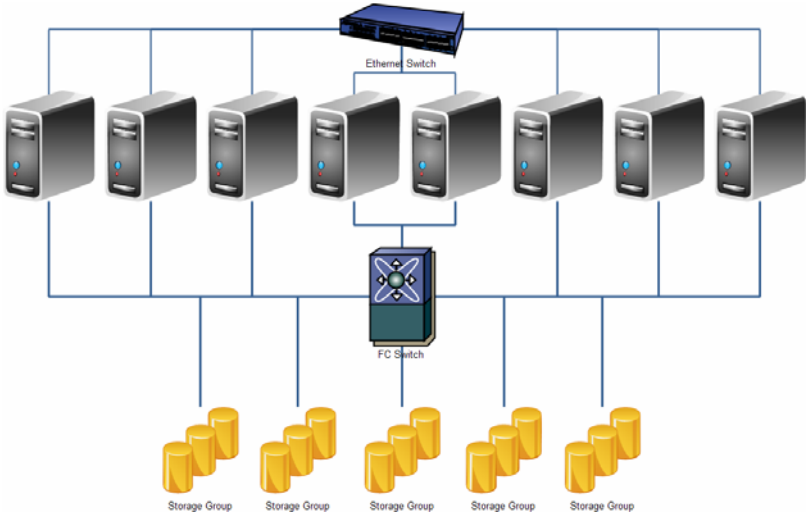
  write performance is degenerated due to *Write Amplification*.
- That's why you are not advised to put *real-time performance* demanding files on flash cards
  - Online Redo Logs



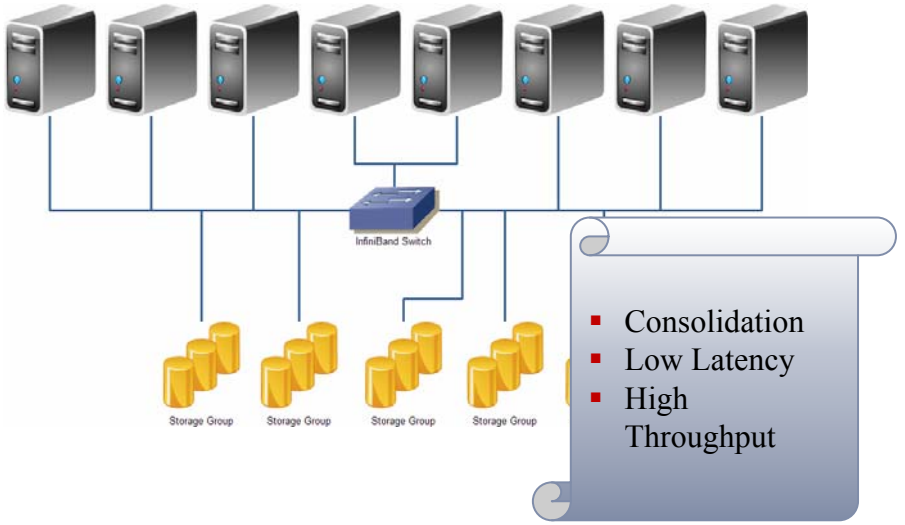
husnu.sensoy@globalmaksimum.com

InfiniBand

Classical Data Center

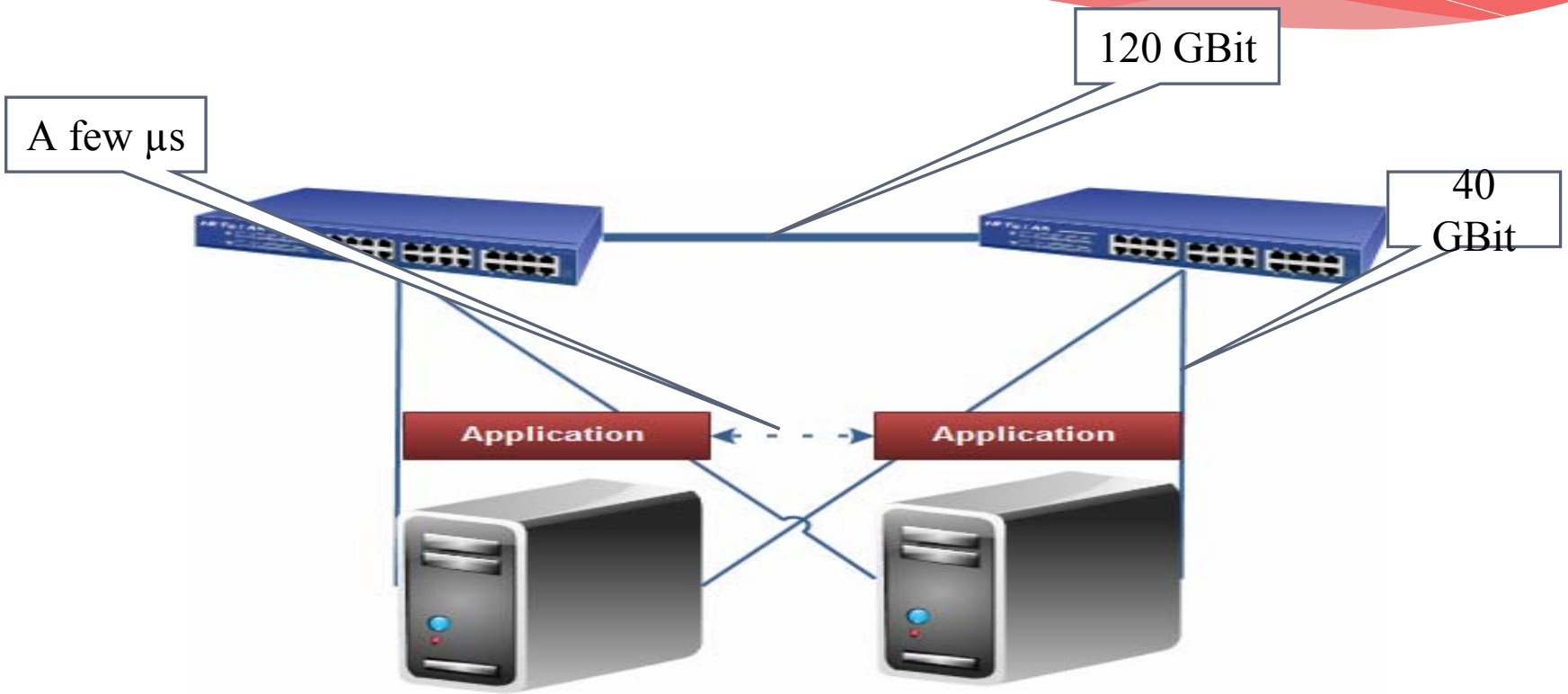


IB Data Center



husnu.sensoy@globalmaksimum.com

InfiniBand Performance



husnu.sensoy@globalmaksimum.com

A Pruned Stack of InfiniBand

husnu.sensoy@globalmaksimum.com

Reliable Datagram Socket

- Oracle has worked on applicability of some alternatives
  - *IPoIB* – high CPU overhead, same unreliable delivery (UDP)
  - *SDP* – connection oriented
- RDS
  - 50% less CPU than IPOIB, UDP
  - ½ Latency of UDP (no user-mode acks)
  - Decoupled from user-mode CPU loading
  - Passes all Oracle regression tests in < 2 wks !!!!
  - Supports fail-over across and within HCAs
- With in Exadata cluster RDS
  - is used for cache fusion (*bcopy*).
  - is used to request an I/O from storage cell (*bcopy*).
  - is used for data shipment from storage node to RAC nodes (*RDMA*).
  - is not used for RAC heartbeat over cluster (*TCP*).

Oracle Exadata Database Machine v2

Exadata v2 Software

husnu.sensoy@globalmaksimum.com

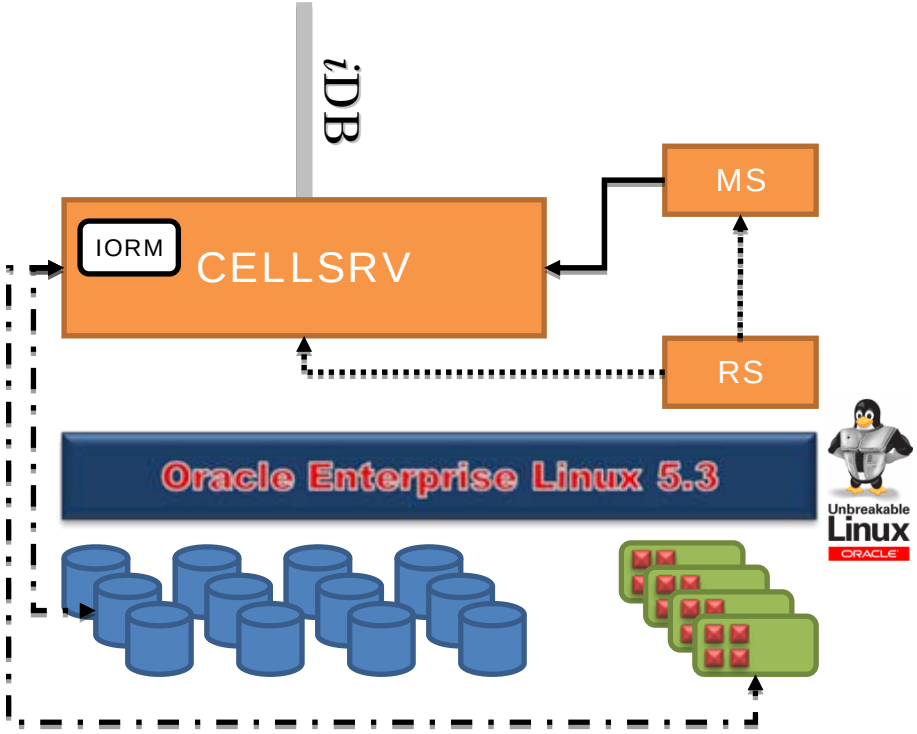
GLOBAL MAKSIMUM

Exadata v2 Specific Software

- Smart Scan
- Storage Indexes
- Smart Flash Cache
- I/O Resource Manager
- Exadata Hybrid Columnar Compression (EHCC)

husnu.sensoy@globalmaksimum.com

Storage Cell Soft Components



- CELLSRV
  - Multithreaded block server
    - Buffer cache reads
    - Smart scans
  - Performs I/O Resource Management
  - Gather operational statistics
  - Communicates over *iDB* with the clients.
- MS
  - OC4J application
  - Provides functionalities for
    - Cell management
    - Cell administration
    - Aler generation
- RS
  - First process becoming live in storage cell.
  - Work as a hang analyzer for CELLSRV and MS

husnu.sensoy@globalmaksimum.com

What is Smart Scan ?

- Smart Scan is initially formed to be column and row filtering based on projection and predicates.
- But this was just the seed idea. Today Smart Scan can also do
  - Projection (column) filtering
  - Predicate (row) filtering
    - `SELECT * FROM v$sqlfn_metadata WHERE offloadable = 'YES';`
  - Preparation of bloom filters for join
  - Smart Incremental backup
  - Scan on encrypted data
  - Smart File Creation
    - RMAN Restore
    - Tablespace Creation
    - File Grow
  - Scoring for Data Mining
    - All data mining scoring functions are offloaded

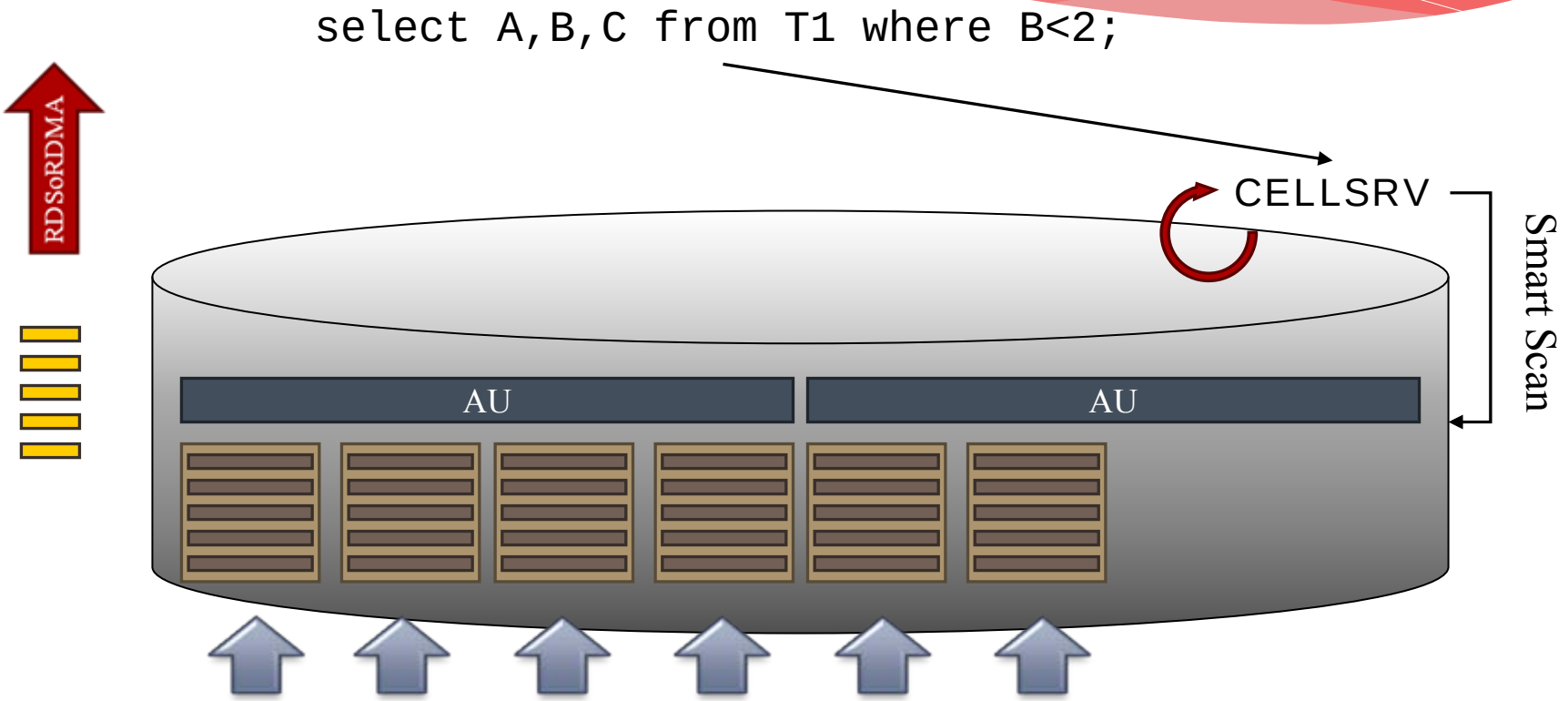
husnu.sensoy@globalmaksimum.com

Storage Index

- Smart Scan is about saving RAC node CPUs during I/O processing, but storage index is about saving the processors of Exadata storage cells.
- Storage Index is not something first used in Exadata. It is borrowed from *Netezza ZoneMap*.
  - Oracle's SI is in memory
- It is about filtering out for a super set of actual result set.

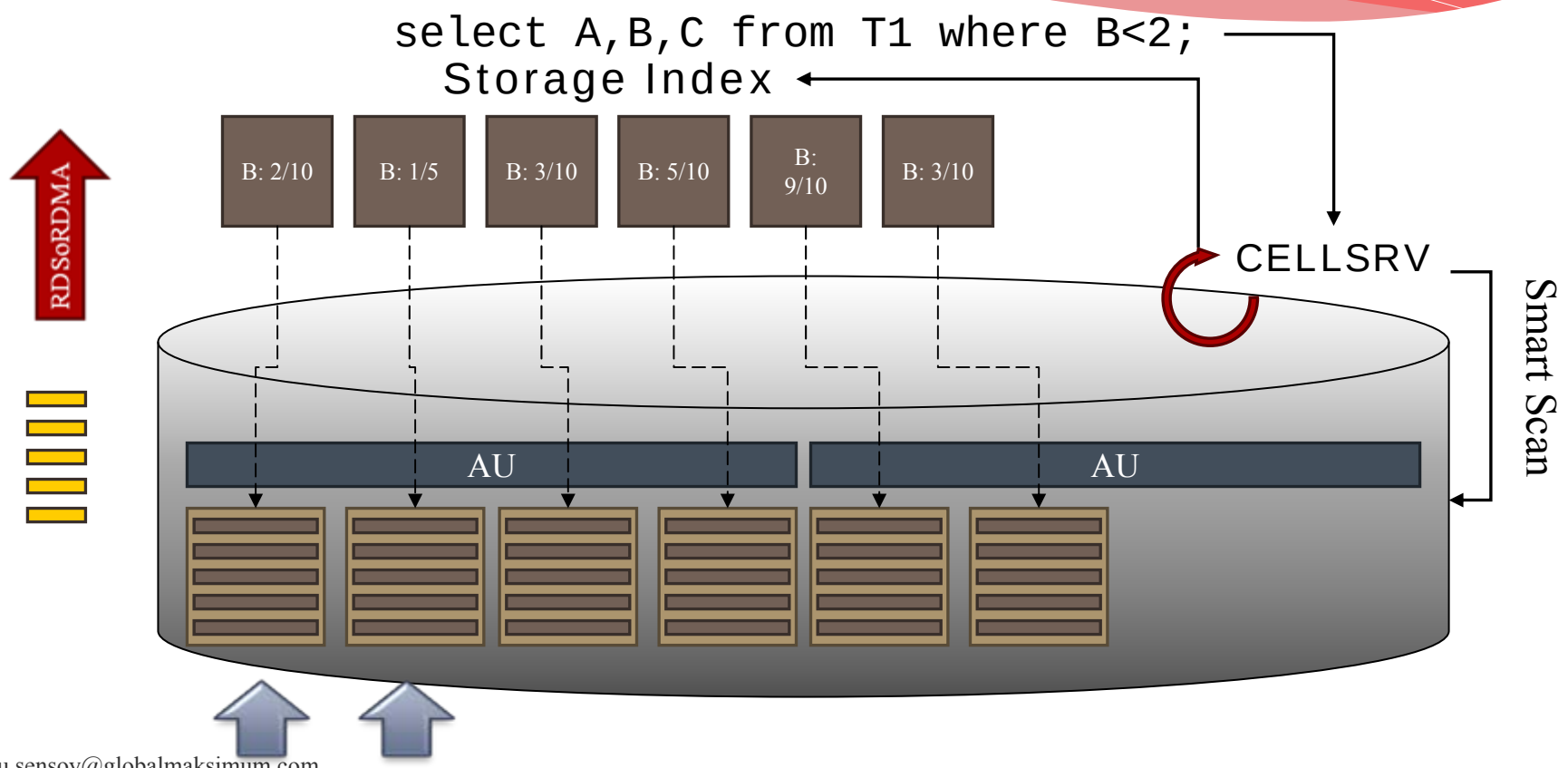
husnu.sensoy@globalmaksimum.com

First Execution



husnu.sensoy@globalmaksimum.com

Go to Coffee Break & Next Executions



husnu.sensoy@globalmaksimum.com

Smart Flash Cache

- Smart Flash Cache is the idea to cache some storage cell data into F20 flash cache drives so that subsequent access will not require disk access.
- Oracle uses flash cache in write-through mode:
  - Don’t confuse this with internal operation of the F20 cards.
  - Although F20 cards are persistent mediums, CELLSRV will not ack the client until write ack returns from disk I/O.
- Smart Flash Cache is logically a non-persistent medium meaning that it’s content will be lost/useless in case of a cell reboot.
  - That’s mainly because the content of a flash cache is kept in a hash table by CELLSRV.

husnu.sensoy@globalmaksimum.com

Good/Bad Things for Caching

Good Things

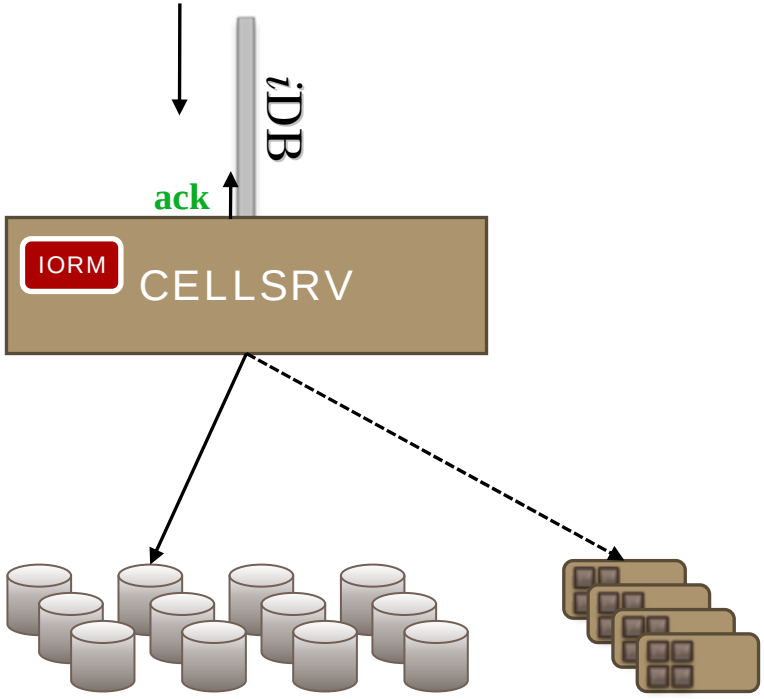
- Frequently accessed data and index blocks.
- Control file reads and writes.
- File header reads and writes.

Bad Things

- I/Os to mirror copies.
- Backup-related I/O
- Data Pump I/O
- Data file formatting.
- Redo Write Operations.

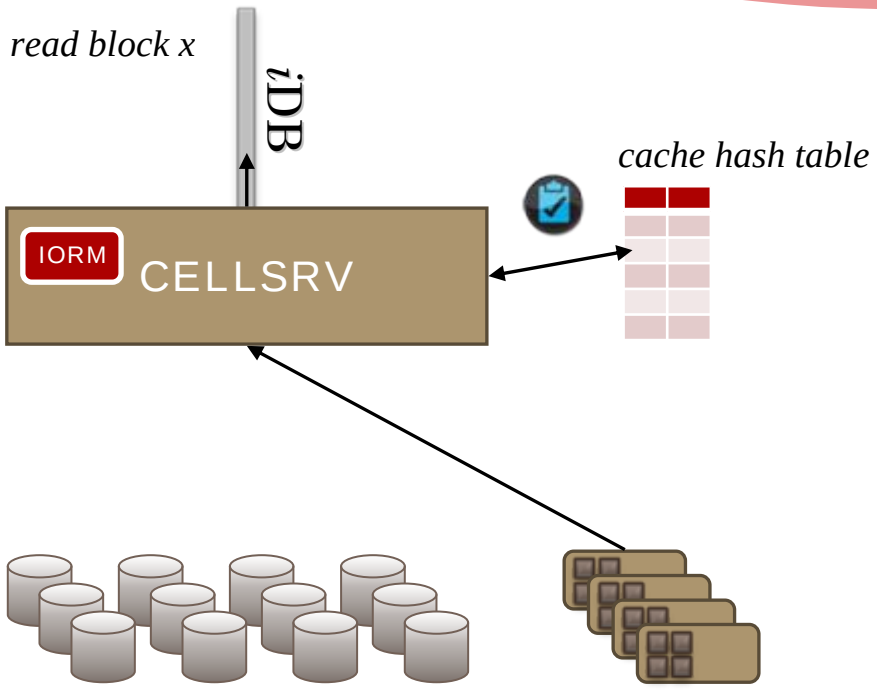
husnu.sensoy@globalmaksimum.com

Smart Flash Cache Write



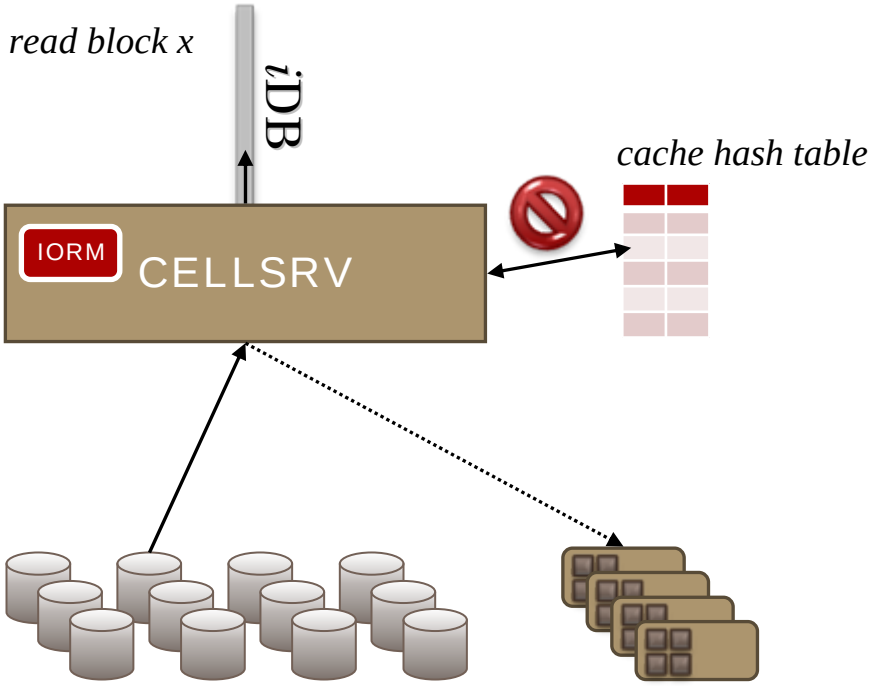
husnu.sensoy@globalmaksimum.com

Smart Flash Cache Read Hit



husnu.sensoy@globalmaksimum.com

Smart Flash Cache Read Miss



husnu.sensoy@globalmaksimum.com

Q & A

husnu.sensoy@globalmaksimum.com

GLOBAL MAKSIMUM