ORACLE

Database Tables to Storage Bits: Data Protection Best Practices for Oracle Database

Gurmeet Goindi,
Principal Product Manager, Oracle

ORACLE

Agenda

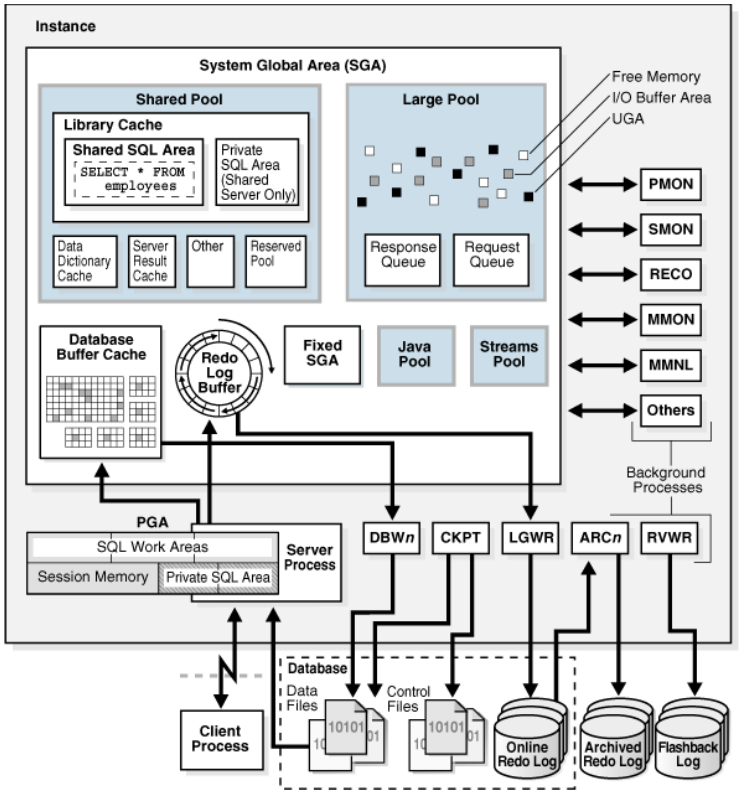
- Database & Storage Architecture
- Data Protection with Storage Technologies
- Storage-based Data Protection – Database Implications
- Database-Integrated Data Protection
- Summary

Quick Show of Hands

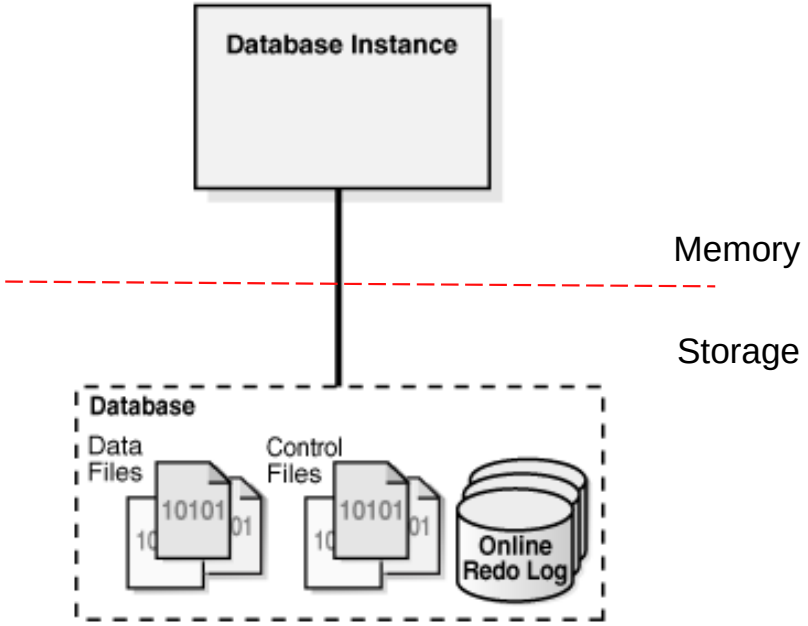
- How many Database administrators?
- How many Storage administrators?
- ☺ Neither? Example, Managers?

Oracle Database Architecture

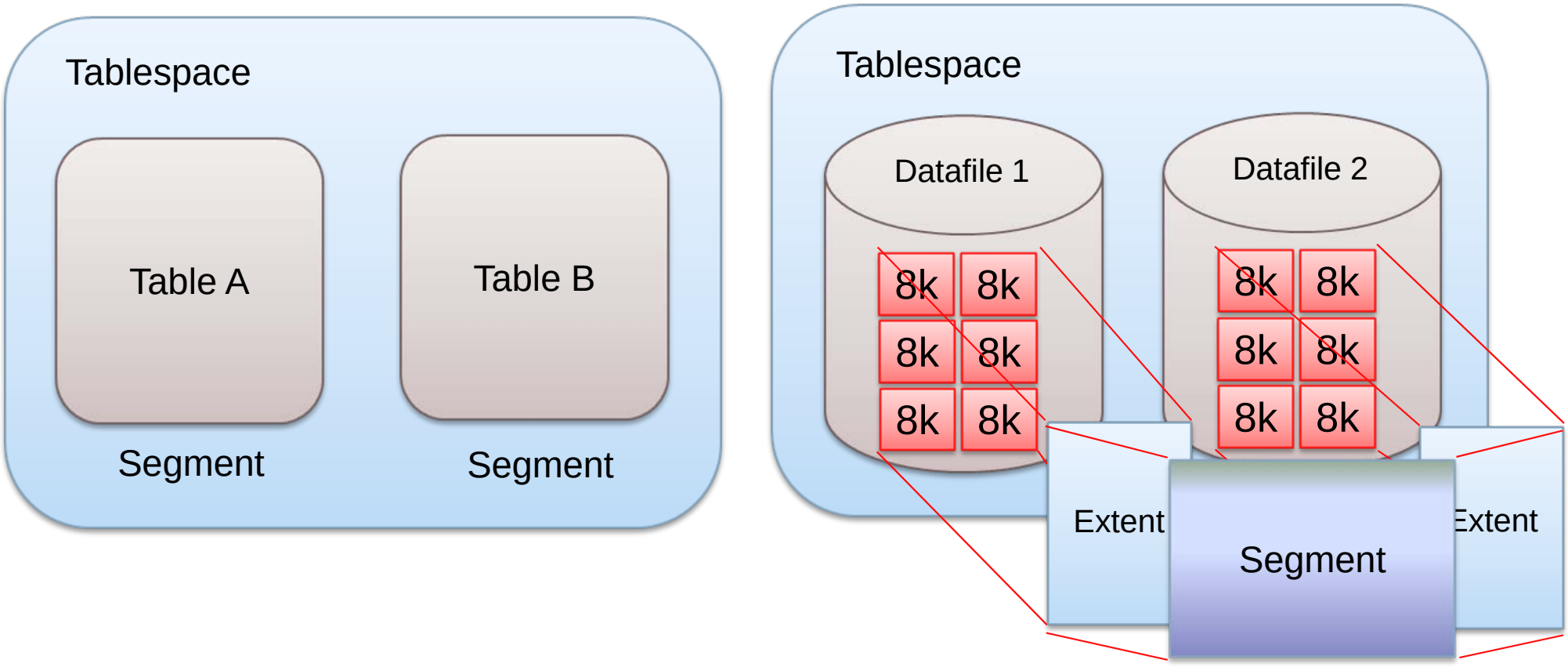
Logical View



Physical View

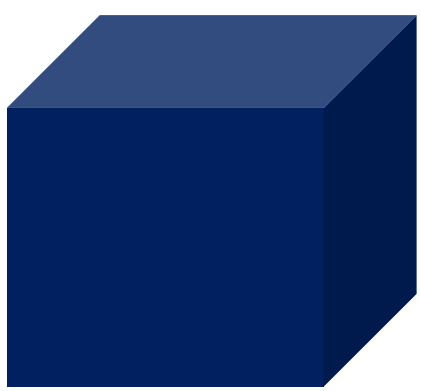


Logical and Physical Database Structures

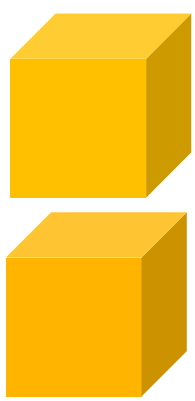


Logical and Physical Database Structures

Summary



Segments



Extents

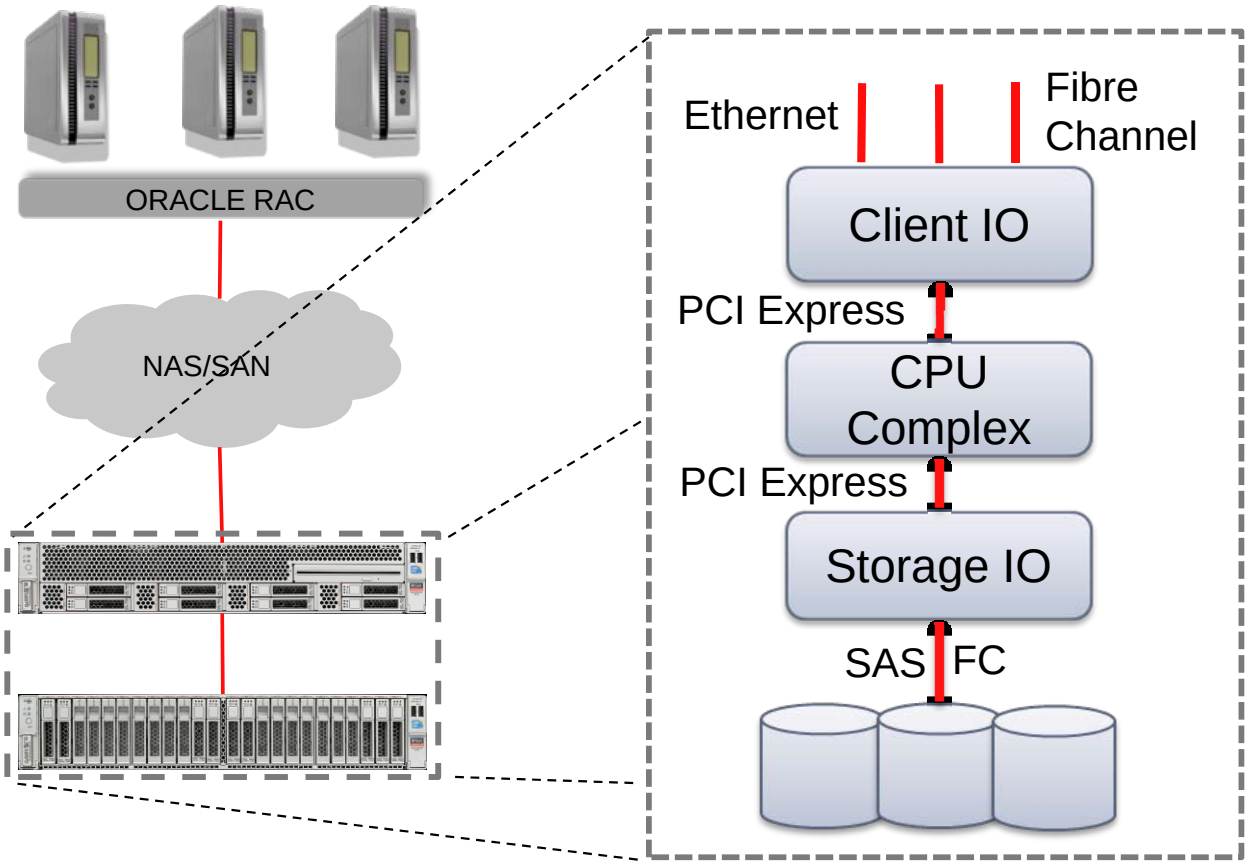
Oracle Data Blocks

Disk Blocks

Compute

Storage

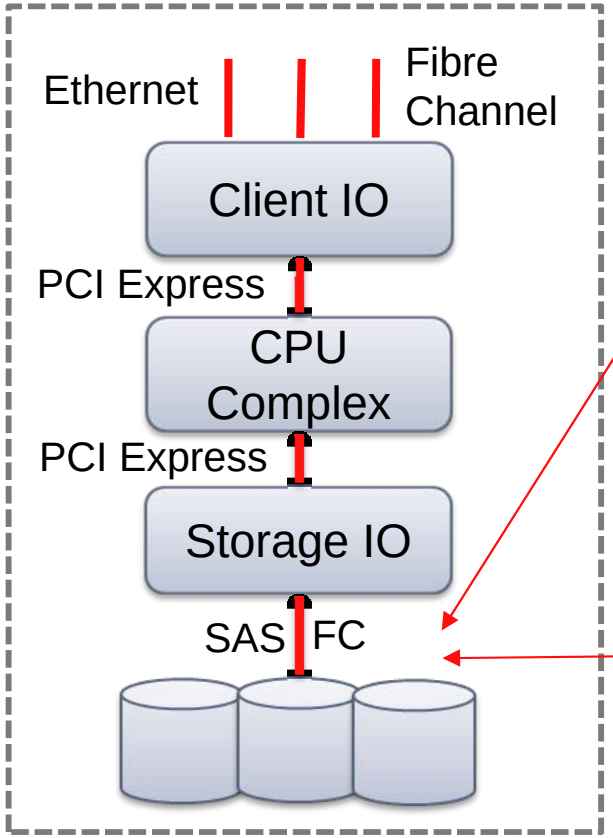
The Anatomy of Storage Infrastructure



- Storage controllers/heads are specialized servers
- Running hundreds of thousands of lines of code
- Most setups require redundant configuration of controllers, disk shelves, switches, etc.

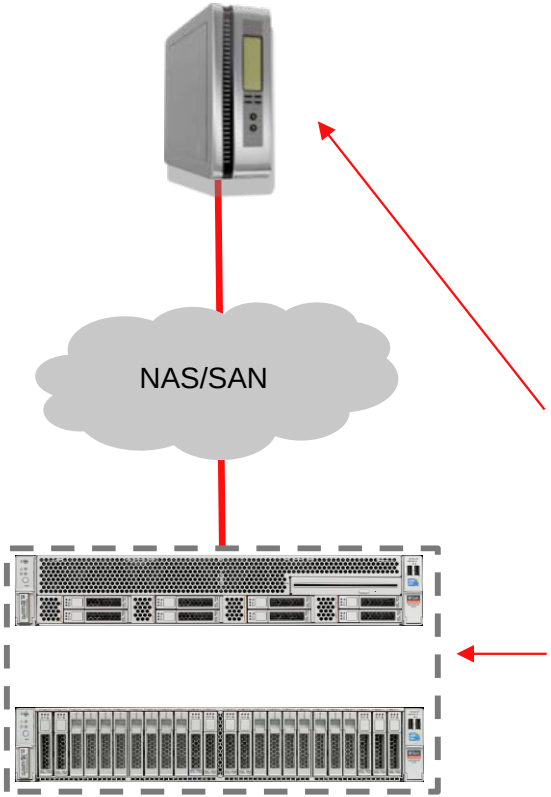
Exposed to similar failures as your server hardware and software

The Anatomy of a Storage Controller



- Client traffic comes in at 10GbE or 4-8Gb FC
- Storage backend is usually slower
- Further, one client request usually results in multiple IO requests to drives
- A write is only acknowledged once written to disk or some form of Non Volatile Memory
- Data flows through many interfaces, each interface can have bugs, add corruptions, etc.
- For performance:
  - Add more drives and/or add faster drives
  - Introduce Flash in the architecture

So … What About Flash in Storage Architecture



- What's special about Flash
  - Performance
  - No moving parts
  - Expensive, but getting cheaper
- Server Attached Flash
  - PCI Express Attached IO cards
  - Extremely fast, improves performance for currently active data
- Flash in Storage Controller
  - Flash is faster than Hard Disk Drives
  - Flash in server is faster than Flash in the storage controller

Server Attached Flash

Standalone Servers with Local Flash Cards

- **Typical deployment:**
  - Standalone x86 server(s)
  - Single instance (non-RAC) databases
  - 1-10 TB local Flash
  - Database stored on Flash
  - Ex: HP DL580 G7 + Fusion-io
- **ANY component failure = loss of database access**
  - CPU, memory, O/S, Flash, etc.
  - Local second server with identical Flash recommended ($$$ x 2)

Good for performance, Not so good for HA

Agenda

- Database & Storage Architecture
- Data Protection with Storage Technologies
- Storage-based Data Protection – Database Implications
- Database-Integrated Data Protection
- Summary

Storage Solutions for Data Protection

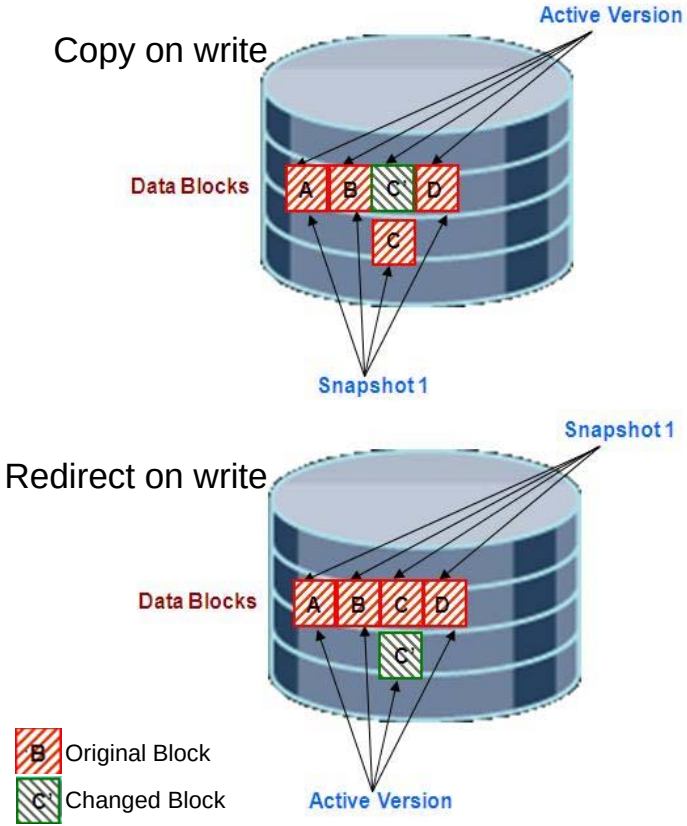
Traditional Approach to Datacenter Disruptions

Backups & Snapshots
Human Errors
Storage Failures
Data Corruptions

Storage Replication
Systemwide Failures
Site or Network Failures
Natural Disasters

Storage Solutions for Data Protection

Snapshots Overview



- Point-in-time, read-only logical copy of your data
- Space efficient because only changed blocks are stored
- Multiple implementations:

Copy on write	Redirect on write
Array copies the original block before changing it	Writes are directed to new blocks
High write overhead	Low write overhead
EMC BCV	NetApp Snapshots

Storage Solutions for Data Protection

Snapshots Limitations

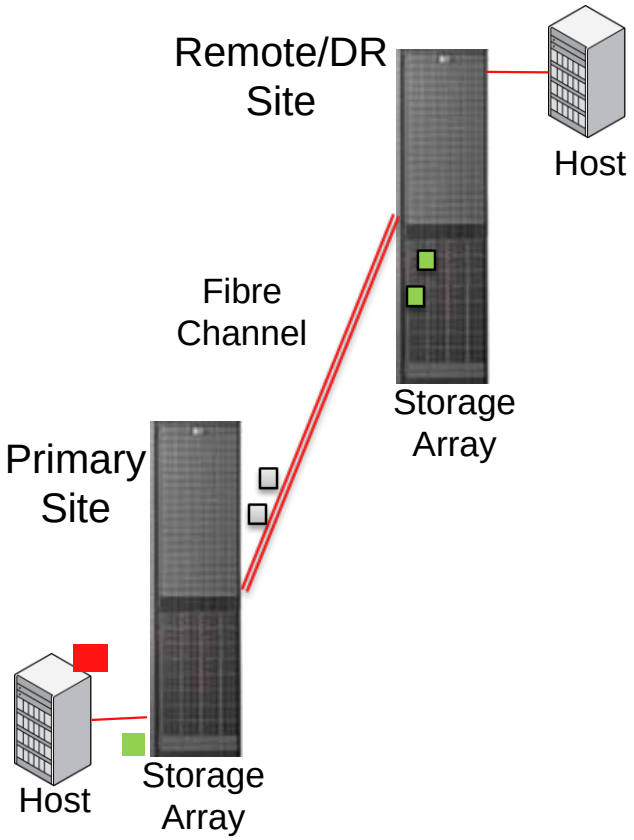
- Snapshots are application data-format unaware and cannot validate application data
- Snapshots reside on the same array as the source data, so they are vulnerable to failures that affect the storage array
- Snapshots are rendered useless in a data loss or corruption scenario
- A corrupted block can potentially affect a series of Snapshots
- Reconstruction has the same performance penalty as that of a full copy
- Restoring a Snapshot can invalidate all Snapshots taken after it

Here is the rub: Snapshots are NOT Backups !!

ORACLE

Storage Solutions for Data Protection

Synchronous Mirroring



- Update made to a data volume on primary site is synchronously replicated to a data volume on secondary site
- The data volume on the secondary site is not mountable for most implementations
- The distance between the sites is typically less than 100km, though the technology limits it to 200km
- Deploying FC links is expensive and complicated. Bandwidth is usually shared by multiplexing Multiple FC channels using a DWDM switch($$)

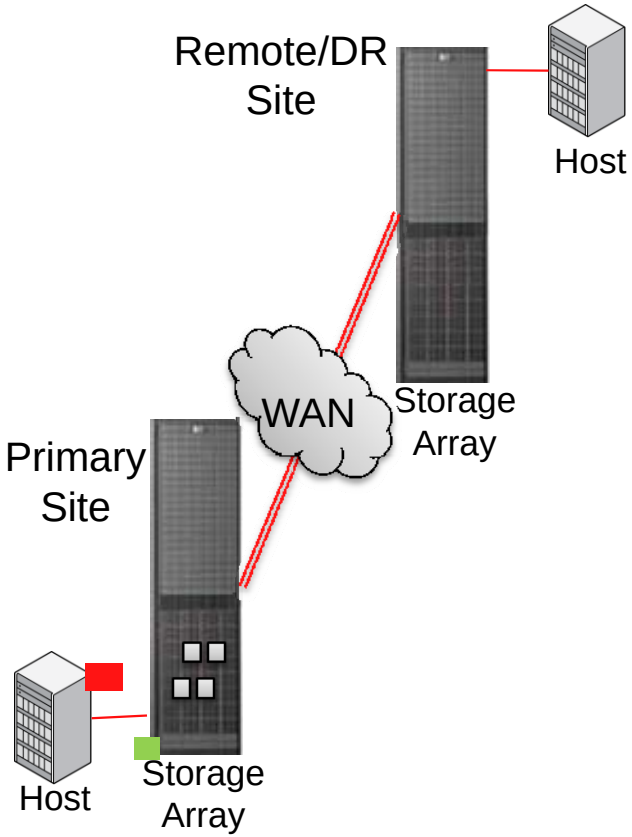
Storage Solutions for Data Protection

Synchronous Mirroring: Limitations and Challenges

- Physical Limitations
  - Cable degrades the signal, the signal can only be transmitted in the range of ~200km without regenerating
  - Latency for light travelling in a fiber cable is 1 ms for every 100 km
  - FC is an acknowledgement-based protocol, hence latency if transmitting each frame will have at least 2 ms latency
  - An IO request from an application will involve transmitting multiple such frames
- Protocol Limitations
  - FC uses credit based algorithm for flow control (buffer credits)
  - A transmit port can send only as many frames as the number of available buffer credits
  - Ineffective management of buffer credits will affect availability and performance

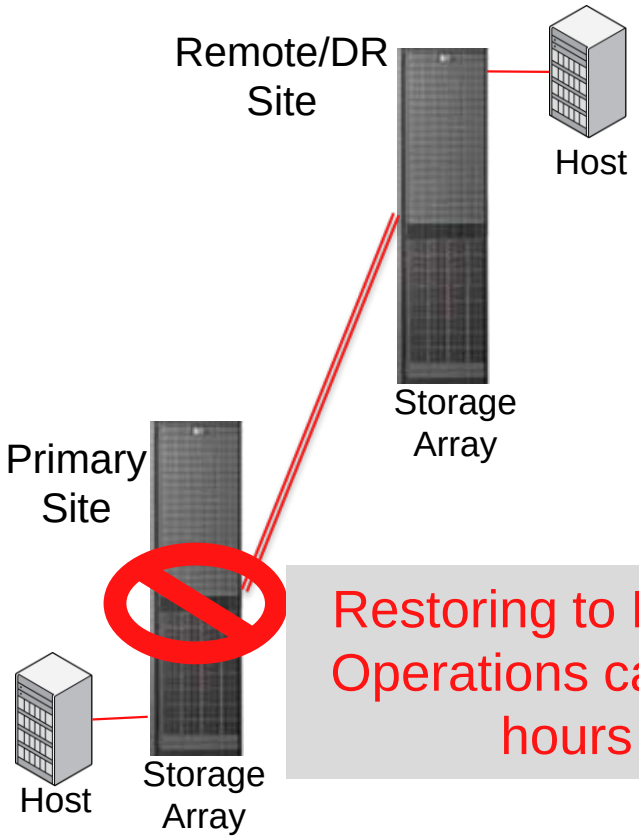
Storage Solutions for Data Protection

Asynchronous Storage Replication



- An Update made to a data volume on the primary site is transmitted to a data volume on the secondary site at a later point in time
- The data volume on the secondary site might be mountable
- Can easily support distances greater than 500 km
- Not the same level of protection as synchronous mode
- The link between the two sides is typically an IP based WAN link (FCIP)
- In case of FCIP the networking gear must handle FC over IP, and also have advanced QoS features

In the Event of a Disaster



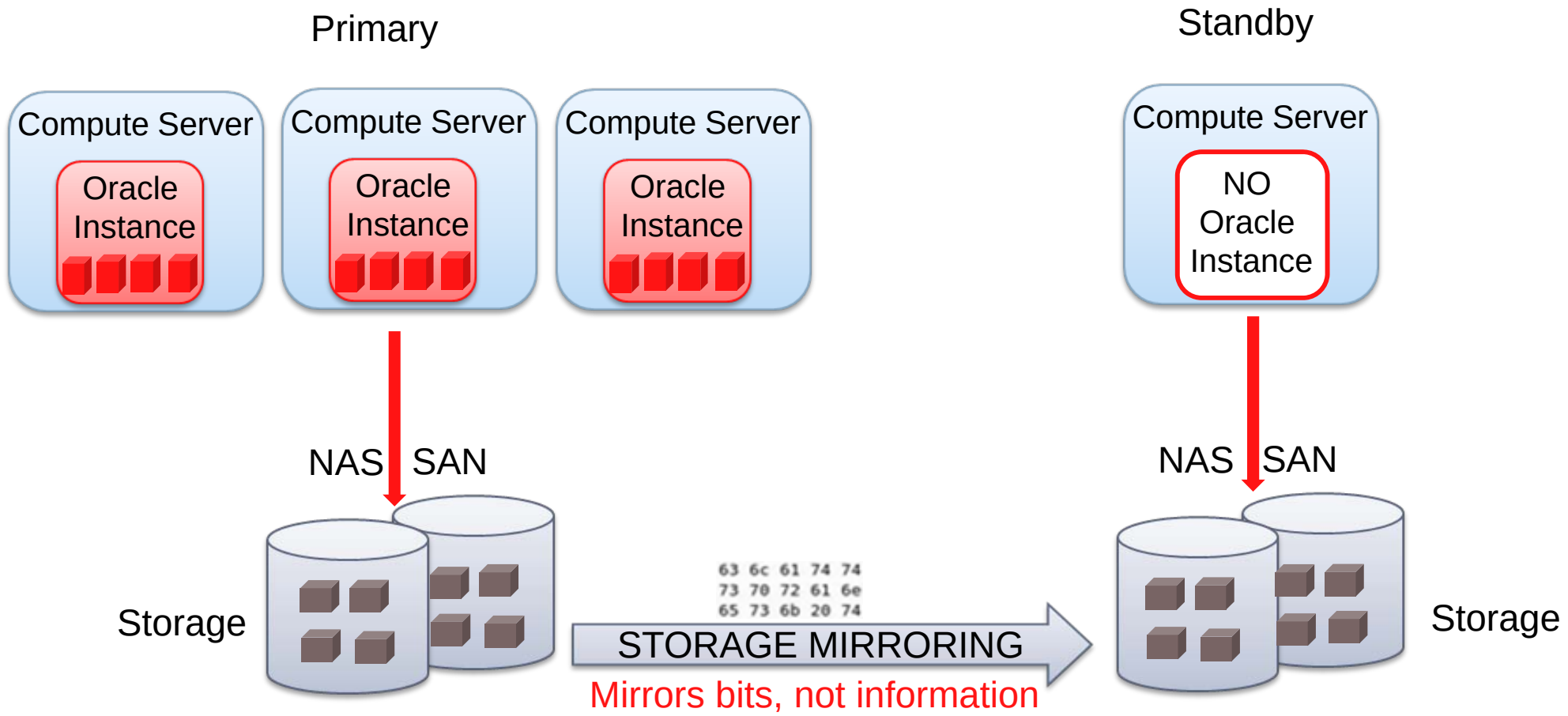
Restoring to Normal Operations can take hours

- Remote system will detect an outage
- The storage resources at the remote site will have to be enabled for writes
- The Remote host will perform the required procedure to access the storage. For example a UNIX host will need to do the following:
  - Import the Volume group
  - Activate the logical Volumes
  - Sanity check the Files System (fsck)
  - Mount the Volumes
  - Restart the application

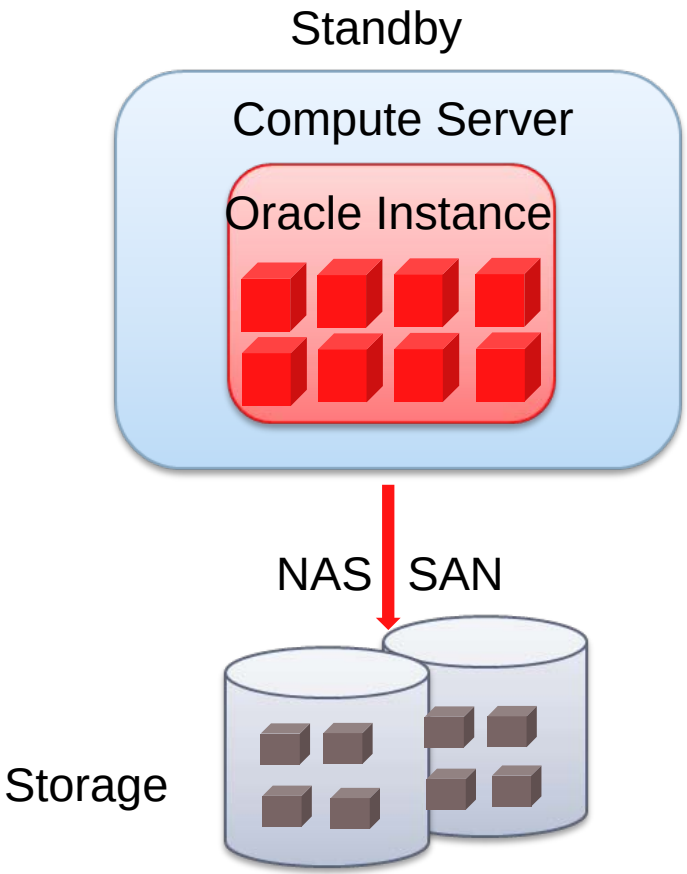
Agenda

- Database & Storage Architecture
- Data Protection with Storage Technologies
- Storage-based Data Protection – Database Implications
- Database-Integrated Data Protection
- Summary

Storage Level Data Protection for Database



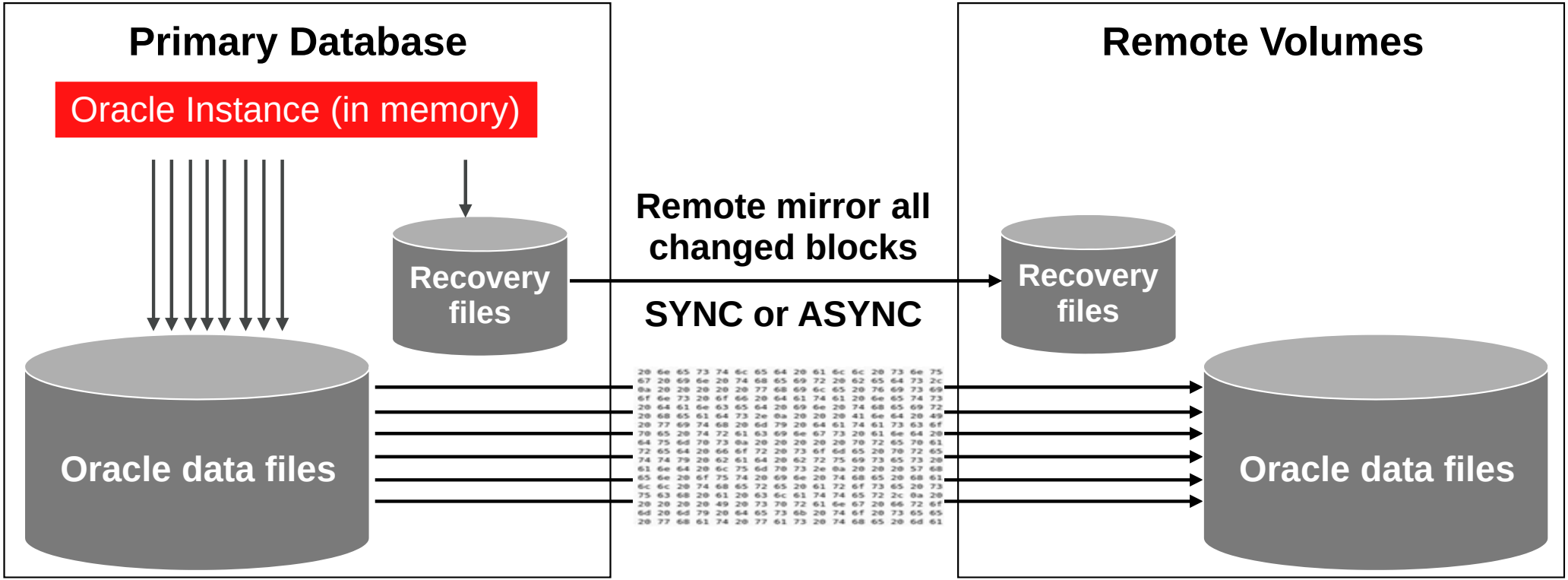
Database Recovery in Event of a Disaster



- Activating Storage Replica
  - Failure Detection
  - Follow Vendor specific process to bring DR storage online (can take hours)
- Bring database to a consistent state
  - Roll Forward (redo)
  - Transaction Recovery
- Open database for applications

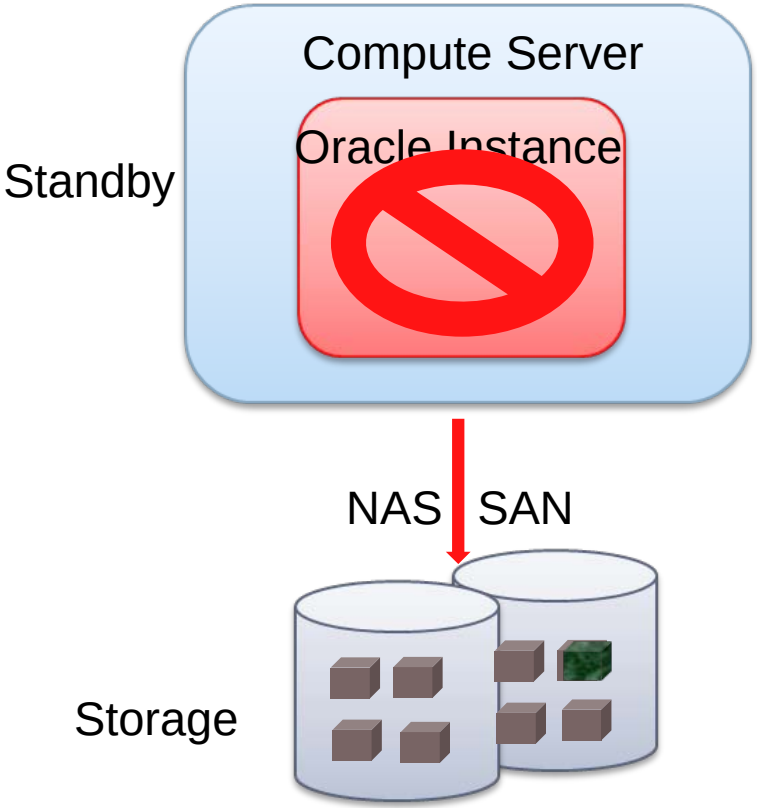
Storage Remote-Mirroring: Fundamental Flaw

Zero Oracle Awareness, Mirrors Every Write for Real-Time Protection



Protection From Localized Failures

Silent corruptions and disk failures



- Bit Error Rates for memory, media and network elements have remained almost constant over last few years
- The density of bits packed in those media has increased. The rate of processing these bits has also increased
- Hence the probability of corrupting these bits keeps increasing.
- HBA and disk drives all have firmware and firmware have bugs
- Disk Drive firmware sometimes misrepresent the state of a write – Lost Writes

Corruption Propagation

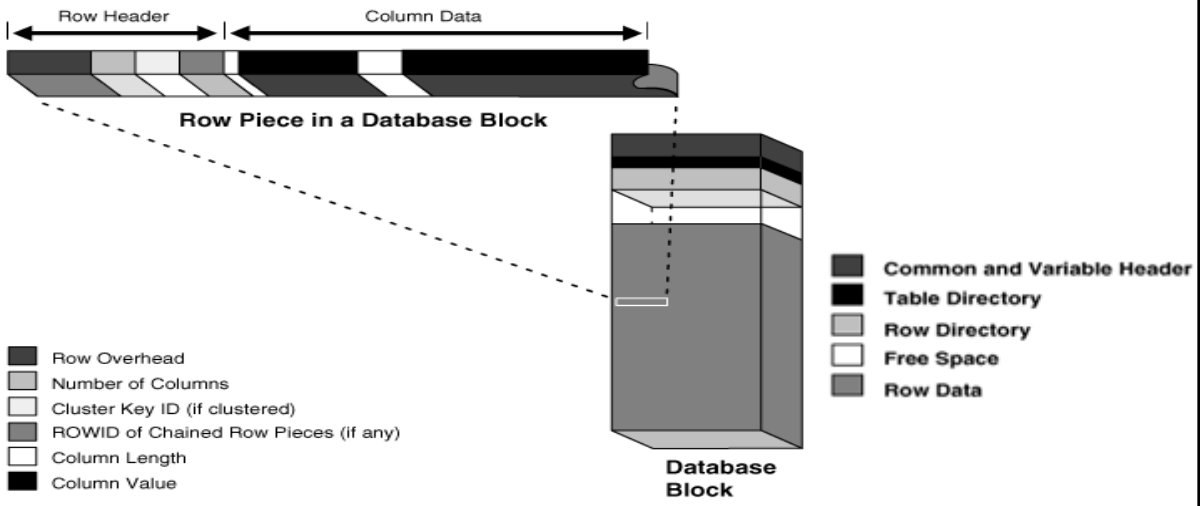
What is my exposure ??

Storage Remote Mirroring…
blocks are just bits on a disk

```
Block 3941 (0x0f65)
      0  1  2  3  4  5  6  7  8  9  a  b  c  d  e  f
000  20 20 20 54 77 61 73 20 74 68 65 20 6e 69 67 68
010  74 20 62 65 66 6f 72 65 20 73 74 61 72 74 2d 75
020  70 20 61 6e 64 20 61 6c 6c 20 74 68 72 6f 75 67
030  68 20 74 68 65 20 6e 65 74 2c 0a 20 20 20 20 20
040  6e 6f 74 20 61 20 70 61 63 6b 65 74 20 77 61 73
050  20 6d 6f 76 69 6e 67 3b 20 6e 6f 20 62 69 74 20
060  6e 6f 72 20 6f 63 74 65 74 2e 0a 20 20 20 54 68
070  65 20 65 6e 67 69 6e 65 65 72 73 20 72 61 74 74
080  6c 65 64 20 74 68 65 69 72 20 63 61 72 64 73 20
090  69 6e 20 64 65 73 70 61 69 72 2c 0a 20 20 20 20
0a0  20 68 6f 70 69 6e 67 20 61 20 62 61 64 20 63 68
0b0  69 70 20 77 6f 75 6c 64 20 62 6c 6f 77 20 77 69
0c0  74 68 20 61 20 66 6c 61 72 65 2e 0a 20 20 20 54
0d0  68 65 20 73 61 6c 65 73 6d 65 6e 20 77 65 72 65
0e0  20 6e 65 73 74 6c 65 64 20 61 6c 6c 20 73 6e 75
0f0  67 20 69 6e 20 74 68 65 69 72 20 62 65 64 73 2c
100  0a 20 20 20 20 20 77 68 69 6c 65 20 76 69 73 69
110  6f 6e 73 20 6f 66 20 64 61 74 61 20 6e 65 74 73
120  20 64 61 6e 63 65 64 20 69 6e 20 74 68 65 69 72
130  20 68 65 61 64 73 2e 0a 20 20 20 41 6e 64 20 49
140  20 77 69 74 68 20 6d 79 20 64 61 74 61 73 63 6f
150  70 65 20 74 72 61 63 69 6e 67 73 20 61 6e 64 20
160  64 75 6d 70 73 0a 20 20 20 20 20 20 70 72 65 70
170  72 65 64 20 66 6f 72 20 73 6f 6d 65 20 70 72 65
180  74 74 79 20 62 61 64 20 62 72 75 69 73 65 73 20
190  61 6e 64 20 6c 75 6d 70 73 2e 0a 20 20 20 57 68
1a0  65 6e 20 6f 75 74 20 69 6e 20 74 68 65 20 68 61
1b0  6c 6c 20 74 68 65 72 65 20 61 72 6f 73 65 20 73
1c0  75 63 68 20 61 20 63 6c 61 74 74 65 72 2c 0a 20
1d0  20 20 20 20 49 20 73 70 72 61 6e 67 20 66 72 6f
1e0  6d 20 6d 79 20 64 65 73 6b 20 74 6f 20 73 65 65
1f0  20 77 68 61 74 20 77 61 73 20 74 68 65 20 6d 61
```

Checksum is the only validation method

Only Application aware data consistency check can guarantee end to end data integrity.
Eg: Oracle Data Guard



Far superior than storage level checksum

Storage Based Data Protection for Databases

Summary – Risks and Limitations

Attributes	Storage Based Solution
Transactional Integrity	Not Guaranteed
Time to recover to normal operations	In hours
Complexity	High
Use of Network Resources	Inefficient
Idle DR Resources	Mostly
Cost of the Solution	Expensive
Management Overhead	High
Risk of a Resume Generating Event ☺	Very High

Agenda

- Database & Storage Architecture
- Data Protection with Storage Technologies
- Storage-based Data Protection – Database Implications
- Database-Integrated Data Protection
- Summary

Oracle Integrated Data Protection

- Oracle recommendation: Maximum Availability Architecture (MAA)
  - Vastly more intelligent data protection than offered by storage layer
  - Architectural principle: Fault Isolation
- Examples of MAA's unique value-proposition
  - Database-optimized Physical Replication
  - Surgical Protection from Corruptions – e.g. Lost Writes
  - Zero Downtime Auto-repair of Corrupted Blocks
  - Efficient Correction of Human Errors

Maximum Availability Architecture (MAA)

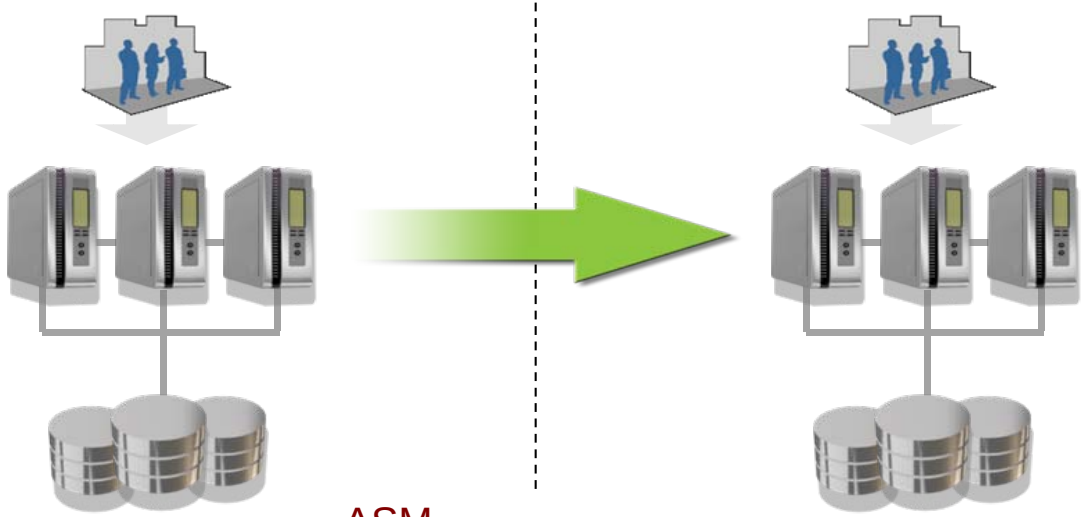
Low-Cost, Integrated, Fully Active, High ROI

Production Site

RAC
– *Scalability*
– *Server HA*

Flashback
– *Human error correction*

Online Redefinition, Edition-based Redefinition, Data Guard, GoldenGate
– *Minimal downtime maintenance, upgrades, and migrations*



ASM
– *Volume Management*

RMAN & Fast Recovery Area
– *On-disk backups*

Active Replica

Active Data Guard
– *Data Protection, DR*
– *Query Offload*

GoldenGate
– *Active-active*
– *Heterogeneous*

Oracle Secure Backup
– *Backup to tape / cloud*

Maximum Availability Architecture (MAA)

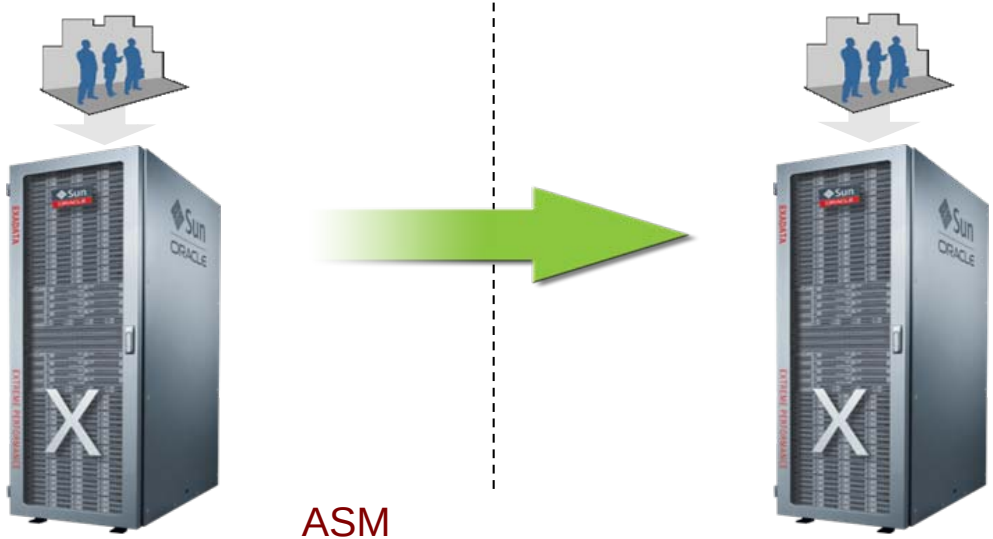
Low-Cost, Integrated, Fully Active, High ROI

Production Site

RAC
– *Scalability*
– *Server HA*

Flashback
– *Human error correction*

Online Redefinition, Edition-based Redefinition, Data Guard, GoldenGate
– *Minimal downtime maintenance, upgrades, and migrations*



ASM
– *Volume Management*

RMAN & Fast Recovery Area
– *On-disk backups*

Active Replica

Active Data Guard
– *Data Protection, DR*
– *Query Offload*

GoldenGate
– *Active-active*
– *Heterogeneous*

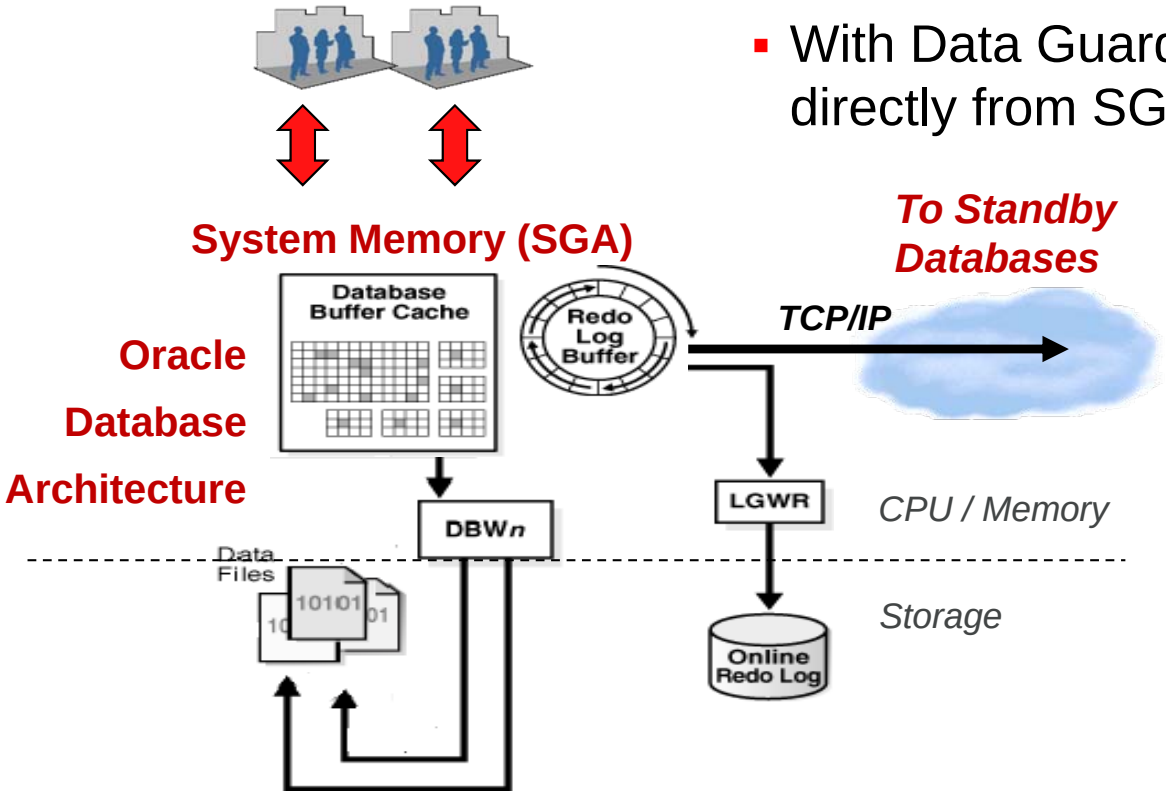
Oracle Secure Backup
– *Backup to tape / cloud*

Oracle Integrated Data Protection

- Oracle recommendation: Maximum Availability Architecture (MAA)
  - Vastly more intelligent data protection than offered by storage layer
  - Architectural principle: Fault Isolation
- Examples of MAA's unique value-proposition
  - Database-optimized Physical Replication
  - Surgical Protection from Corruptions – e.g. Lost Writes
  - Zero Downtime Auto-repair of Corrupted Blocks
  - Efficient Correction of Human Errors

Database Integrated Physical Replication

Data Guard: Why A Big Deal

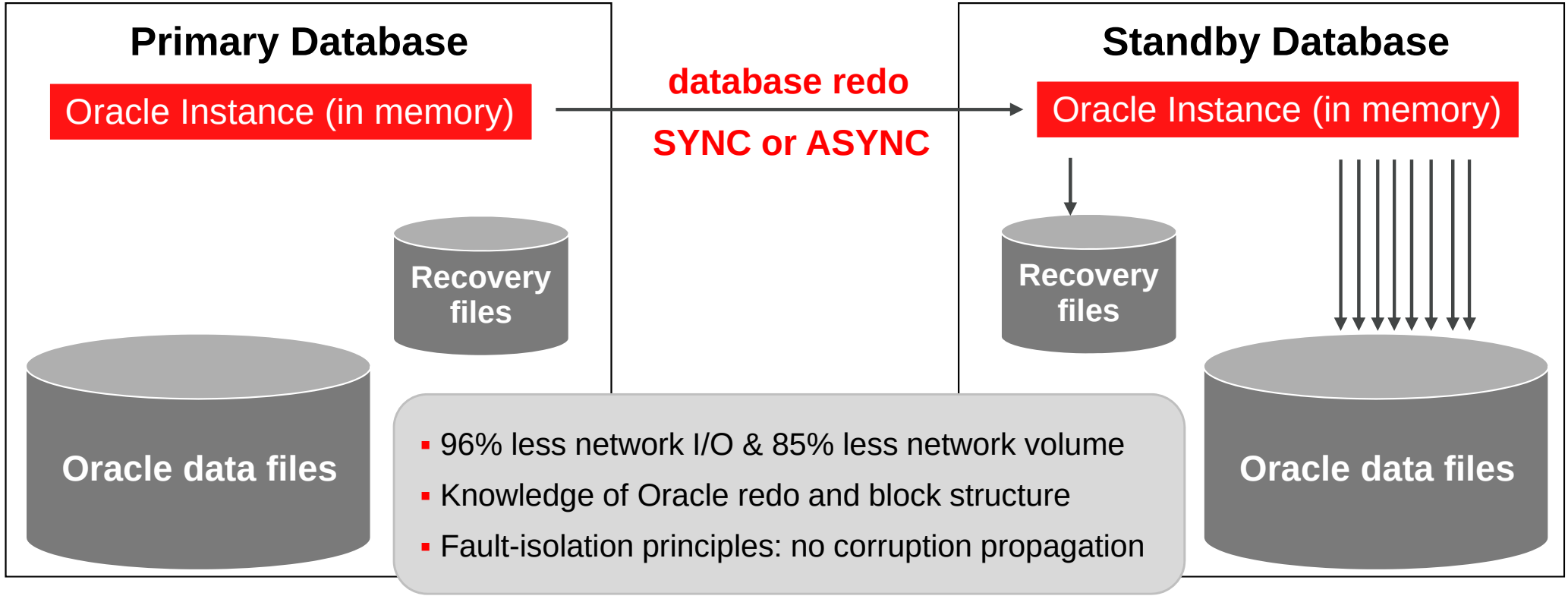


- With Data Guard, redo blocks are transmitted directly from SGA: like a memcpy over the network

- Better performance since no disk I/O
- Better isolation from lower layer faults
- Better network utilization: only redo blocks sent over the network
- Transactional consistency always maintained
- Upon database failover, apps simply reconnect to new primary database

Data Guard, Compared to Storage Mirroring

Oracle Aware – Simple, Efficient, Physical Replication



Oracle Integrated Data Protection

- Oracle recommendation: Maximum Availability Architecture (MAA)
  - Vastly more intelligent data protection than offered by storage layer
  - Architectural principle: Fault Isolation
- Examples of MAA's unique value-proposition
  - Database-optimized Physical Replication
  - Surgical Protection from Corruptions – e.g. Lost Writes
  - Zero Downtime Auto-repair of Corrupted Blocks
  - Efficient Correction of Human Errors

Outage at Financial Services Company

- Production database alert.log:
  - **Errors in file /opt/app/oracle/admin/dg/bdump/dg1.trc:**
  - **ORA-01186 : file 93 failed verification tests**
  - **ORA-01251 : Unknown File Header Version read for file number 93**

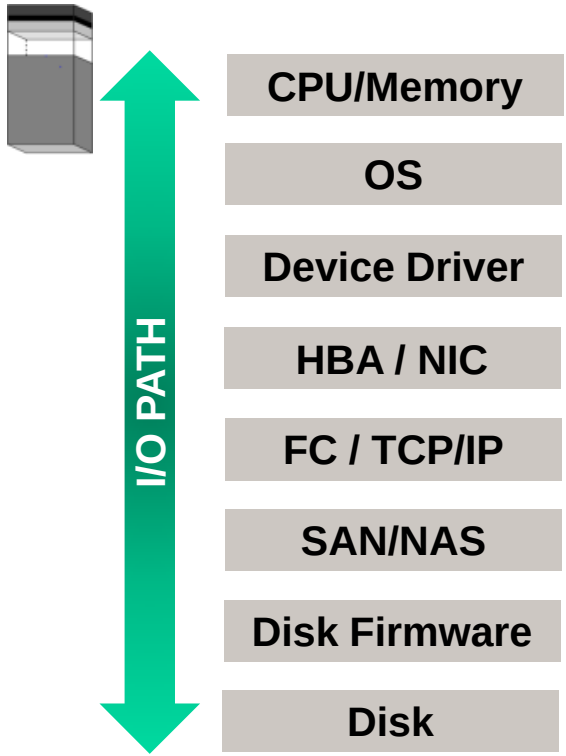
- ***ORA-01251*** **- Corrupted file header. This could be caused due to missed read or write or hardware problem or process external to oracle overwriting the information in file header.**

- Database crashed – customer-facing applications down
  - Trade confirmation, new accounts, customer account information

Data Corruption Protection

Integrated with the Database

- Only the database system has intrinsic knowledge of its block structure
- Only the database can perform true end-to-end validation as block traverses the system stack
- Validation checks can be scaled across related processes such as transaction management, backup & recovery, mirroring, etc.



Oracle-aware Corruption Protection

Built-in Data Validation

- Database can check data, detect and repair corruptions
  - Checksum validation to detect data and redo block corruption
  - Checks semantic integrity of data blocks (Oracle knows Oracle)
  - Detects writes acknowledged but really lost by the I/O subsystem
  - Administrator can configure the level of checking
  - Can be configured for data blocks / data + index blocks
- Specific technologies provide additional validation
  - RMAN validates while doing backup and recovery
  - ASM validates using mirroring copies
  - Data Guard validates with standby database

The Problem of Lost Writes

One of the Nasty Ones!

- Lost Writes – because of storage layer bug / malfunction, a write operation has really failed, but acknowledged as successful
- Database now has stale blocks
  - Subsequent transactions may access the stale blocks
  - May update the same block / other blocks based on stale contents, with serious business implications
    - *Forwarding confidential information to a terminated employee?*
    - *Investment on a stock with multiple sell orders?*
    - *Issuing an incorrect press release?*
- Database may continue running for days, till various ORA-600s

Lost Write Real Life Example

From Support Database: Lengthy Outage for Multi-TB Database

- Problems first appear at the standby, but standby data is safe

> *ORA-00600: internal error code, arguments: [3020] , [648], [1182463], [2719091455], []*
> *ORA-10567 : Redo is inconsistent with data block (file# 648, block# 1182463)*
> ***Recovery interrupted!***

- 4 days later: Production Database still down

> *Noticed odd query results on production*
> *Noticed ORA-600 errors on production this morning for which SGA Heapdump was uploaded.*
> *New info : I was rebuilding an index. After a few minutes, the database took an unexpected crash.*
> ******please help. it's very urgent, production is down.***

The Problem of Lost Writes

Solution!

- Remedy – A basis of comparison with valid blocks
- Oracle solution – Data Guard Physical Standby
- Controlled by **DB_LOST_WRITE_PROTECT**
  - **TYPICAL:** buffer cache read operations logged in redo, for read-write tablespaces
  - **FULL:** buffer cache read operations logged in redo also for read-only tablespaces
- When lost write protection enabled, SCNs of incoming redo blocks from primary database compared to SCNs of blocks on physical standby

The Problem of Lost Writes

Detection by Data Guard

- Primary(SCN) lower than Standby(SCN) implies lost-write error on primary database:
  - `ORA-00752: recovery detected a lost write of a data block`
  - `ORA-10567: Redo is inconsistent with data block (file# 7, block# 26)`
  - `ORA-10564: tablespace TBS_2`
  - `ORA-01110: data file 7: '/oracle/dbs/btbs_21.f'`
  - `ORA-10561: block type 'TRANSACTION MANAGED DATA BLOCK', data object# 57503`
- In such a case, best option: failover to the standby database
  - `SQL> ALTER DATABASE ACTIVATE STANDBY DATABASE;`
- Additional information:
  - MOS Note 1265884.1 - *Resolving ORA-752 or ORA-600 [3020] During Standby Recovery*

Demonstration

Lost Write Protection

Lost Write Protection By Data Guard

http://www.oracle.com/technetwork/database/features/availability/demonstrations-092317.html

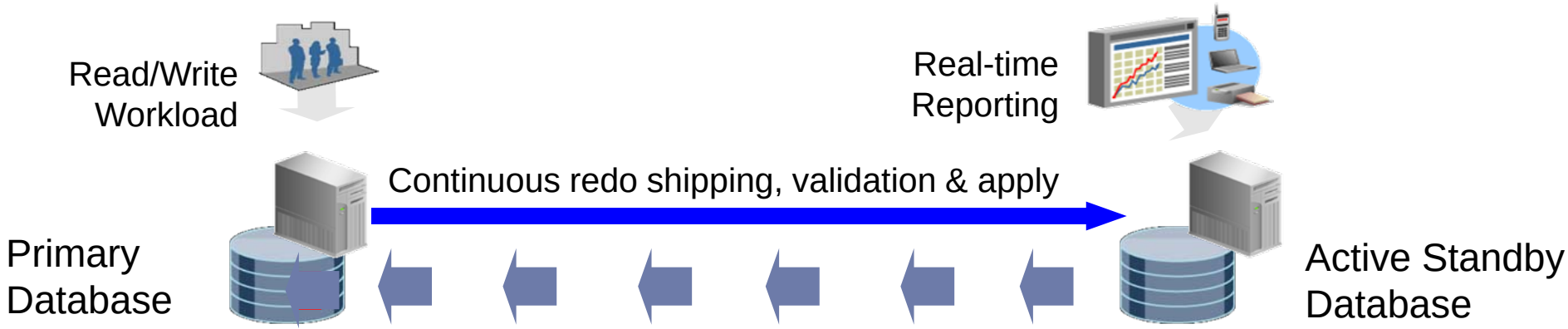
Oracle Integrated Data Protection

- Oracle recommendation: Maximum Availability Architecture (MAA)
  - Vastly more intelligent data protection than offered by storage layer
  - Architectural principle: Fault Isolation
- Examples of MAA's unique value-proposition
  - Database-optimized Physical Replication
  - Surgical Protection from Corruptions – e.g. Lost Writes
  - Zero Downtime Auto-repair of Corrupted Blocks
  - Efficient Correction of Human Errors

Active Data Guard: Auto-Block Repair

High Availability by Repairing Corruptions Online

- Automatic Block Repair
  - When Oracle detects corrupt blocks at primary database, it repairs online by copying good version from an active standby database (& vice versa)
  - Transparent to the user and application



Demonstration

Auto-Corruption Fix

Active Data Guard
Auto Block Repair

http://www.oracle.com/technetwork/database/features/availability/demonstrations-092317.html

Auto Block Repair … In Real Life

Email From Customer For A Tier-1 Deployment

- *As you know, we have put one of our physical standby database in "open read only" mode to make use of "ABMR" (Automatic Block Media Recovery) feature. One incident happened last night … I thought I will share this information with you. Physical standby Database which we opened in read only mode reported below messages in alert.log.*

```
Tue Nov 15 22:00:04 2011
Automatic block media recovery requested for (file# 99, block# 369368)
Automatic block media recovery successful for (file# 99, block# 369368)
Errors in file /home/oracle/admin/physdby_ora_10751_ORAOOP.trc  (incident=560569):
ORA-01578: ORACLE data block corrupted (file # 99, block # 369368)
ORA-01110: data file 99: '+DATA/data2/data2.dbf'
```

- *From operational perspective we see that, Physical standby database which was in open read only mode has seen block corruption, and ABMR helped us to recover this block from primary. Thrilled to see that ABMR works!!*

Auto Block Repair … Insights

Don’t Leave Home Without It!

- For your production databases, deploying an Active Data Guard standby is a good thing!
- You may not need to run real-time reports, but …
  - Active Data Guard standby, in real-time, protects your production database from block corruption, in an app transparent manner
  - As we saw earlier, this also works vice-versa
- Does not matter whether redo transport mode is SYNC or ASYNC – as long as corresponding block can be applied within timeout threshold (default=60 secs), Auto Block Repair will be successful

Oracle Integrated Data Protection

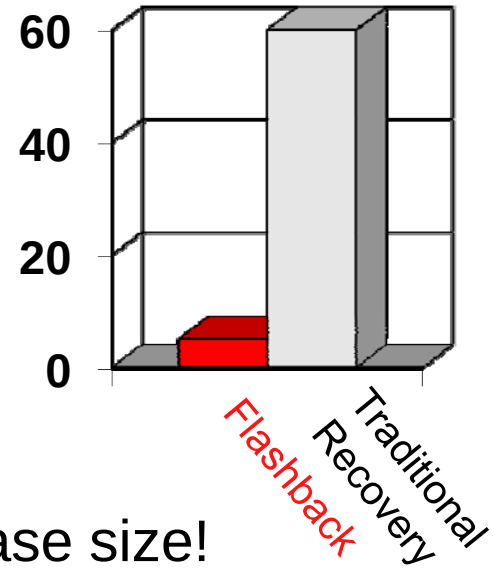
- Oracle recommendation: Maximum Availability Architecture (MAA)
  - Vastly more intelligent data protection than offered by storage layer
  - Architectural principle: Fault Isolation
- Examples of MAA's unique value-proposition
  - Database-optimized Physical Replication
  - Surgical Protection from Corruptions – e.g. Lost Writes
  - Zero Downtime Auto-repair of Corrupted Blocks
  - Efficient Correction of Human Errors

IOUG Survey: Causes for Unplanned Downtime

To Err Is Human

Flashback Technologies

Fast, Granular Error-Correction



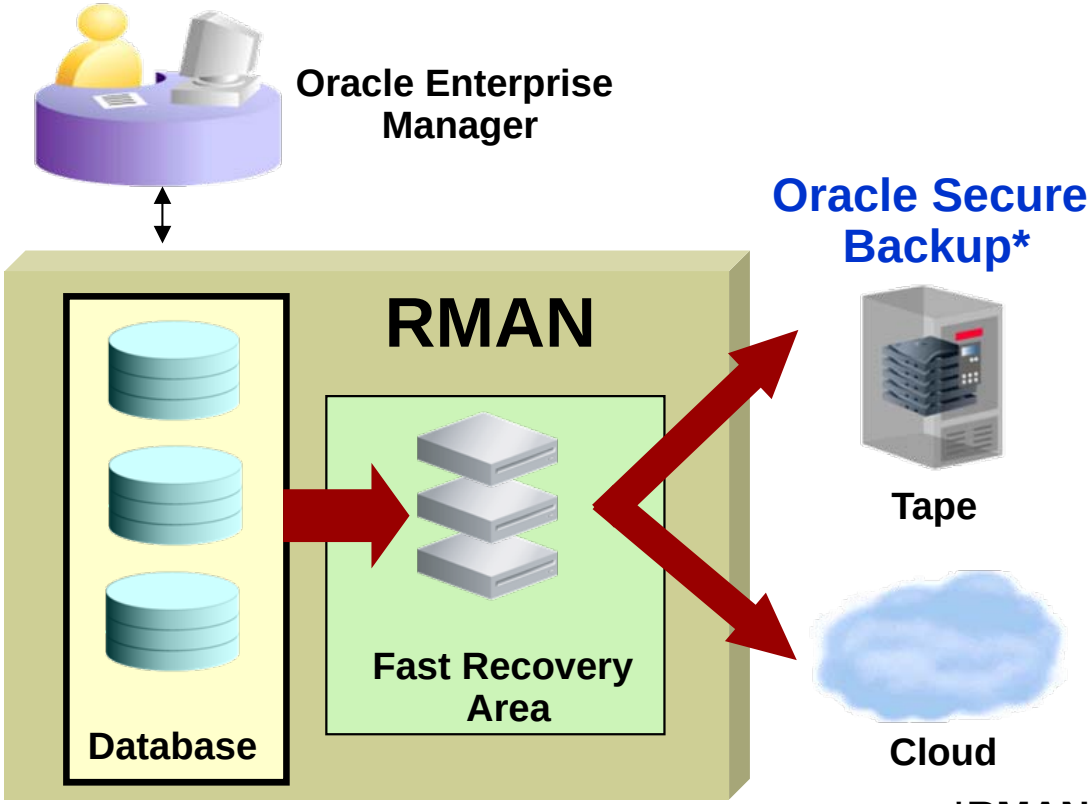
- Flashback revolutionizes error correction:
  - View ‘good’ data as of a point in time before error
- Time/work to rewind data depends on the work done since the error happened, instead of the database size!

Correction Time = Error Time + ~~f(DB_SIZE)~~

- Simple: **`SQL> flashback database to <timestamp>;`**
- Flexible: Flashback Query, Table, Transaction, Database, Drop

Oracle Recovery Manager (RMAN)

Oracle-Integrated Backup & Recovery Engine



- Intrinsic knowledge of database file formats and recovery procedures
  - Block validation
  - Online block-level recovery
  - Tablespace/data file recovery
  - Online, multi-streamed backup
  - Unused block compression
  - Native encryption
- Integrated disk, tape & cloud backup leveraging the Fast Recovery Area (FRA) and Oracle Secure Backup

***RMAN also supports leading 3rd party media managers**

Database Integrated Data Protection

Contrasting with a Storage based Data Protection Solution

Attributes	Storage Based Solution	Database Integrated Solution
Transactional Integrity	Not Guaranteed	Guaranteed
Time to recover to normal operations	In hours	Seconds
Complexity	High	Minimal
Use of Network Resources	Inefficient	Efficient
Idle DR Resources	Mostly	Never
Cost of the Solution	Expensive	Low
Management Overhead	High	Low
Risk of a Resume Generating Event ☺	Very High	Very Low

Agenda

- Database & Storage Architecture
- Data Protection with Storage Technologies
- Storage-based Data Protection – Database Implications
- Database-Integrated Data Protection
- Summary

HA and Data Protection are in Oracle's DNA

Oracle's Design Approach

- High Availability is ingrained in Oracle's software development process
- Bottom's Up approach – the solution is highly available since each component is highly available
- Each new feature adheres to Maximum Availability guidelines

Resources

- **OTN HA Portal:**
  http://www.oracle.com/goto/availability
- **Maximum Availability Architecture (MAA):**
  http://www.oracle.com/goto/maa
- **MAA Blogs:**
  http://blogs.oracle.com/maa
- **Exadata on OTN:**
  http://www.oracle.com/technetwork/database/exadata/index.html
- **Oracle HA Customer Success Stories on OTN:**
  http://www.oracle.com/technetwork/database/features/ha-casestudies-098033.html


ORACLE®