

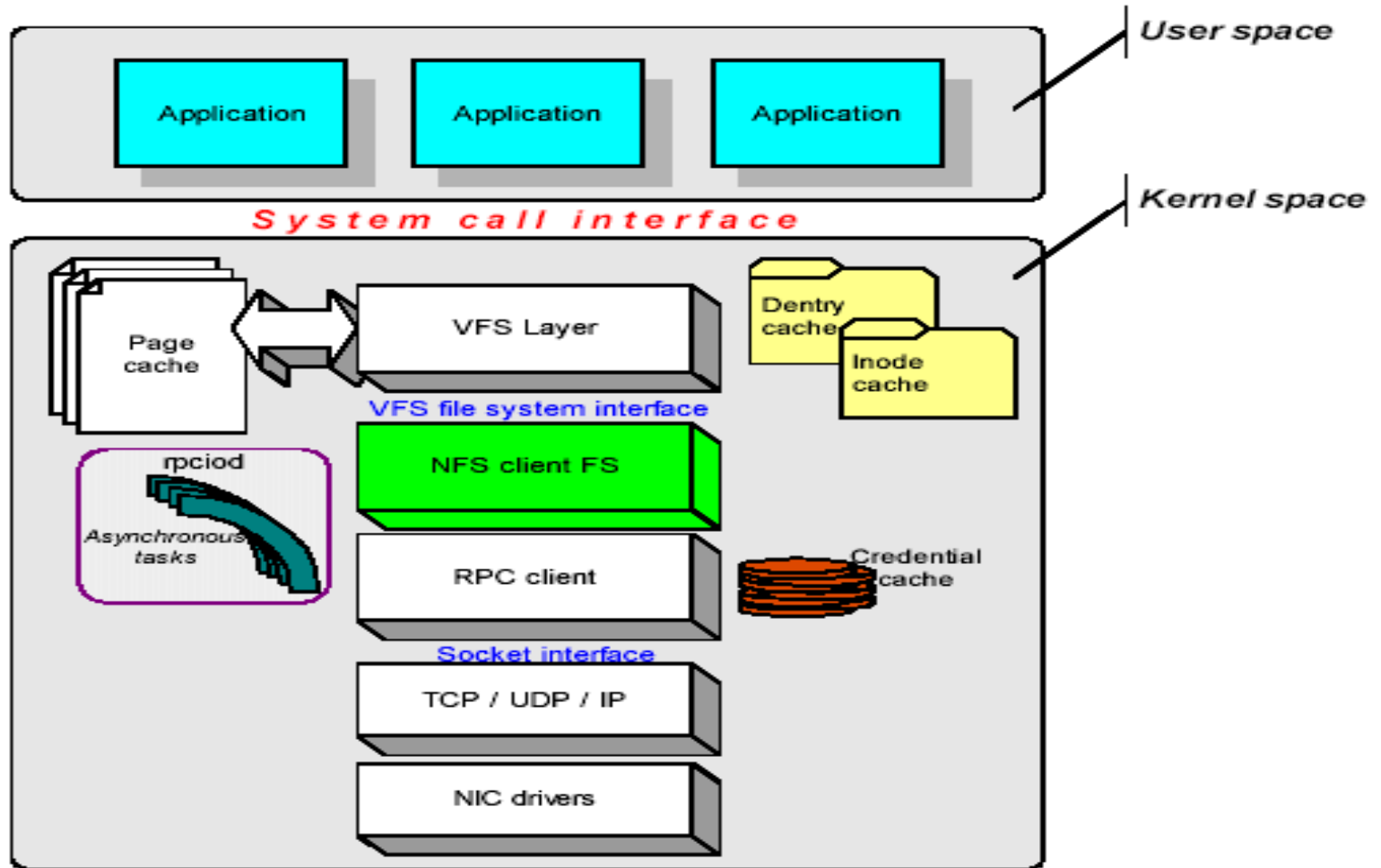


**Implementing Oracle11g Database over NFSv4 from a
Shared Backend Storage
Bikash Roy Choudhury
(NetApp)
Steve Dickson
(Red Hat)**

Overview

- Client Architecture
- Why NFS for a Database?
- Oracle Database 11g RAC Setup
- Mount Options Used
- Database Tuning
- Netapp and the Linux Community

Linux NFS Client Architecture



Linux NFS Client Architecture

■ **Layer 1 – Virtual File System**

- Adapts system calls to generic interface calls supported by all file systems

■ **Layer 2 - NFS Client File System**

- Adapts generic file system calls into NFS RPC requests to server

■ **Layer 3 - RPC Client**

- Converts NFS RPC calls into socket calls
- Byte ordering
- Waits for server replies
- Marshals and unmarshals of data structures

■ **Layer 4 - Linux® Network Layer**

- TCP / UDP / IP

■ **Layer 5 - Network Interface Layer**

- NIC drivers

Linux NFS Client Architecture

- **Linux® NFS Client Implementation**
 - Separates file system from RPC client
 - Integrated in other implementations
 - More efficient by using sockets directly
- Keep architecture in mind for debugging and performance tuning

Linux NFSv4 Client in the 2.6.18-87 Kernel

- Support NFS v4
 - NFSv4 ACLs support
 - use nfs4-acl-tools package or download from <http://www.citi.umich.edu/projects/nfsv4/linux/>
 - Converts the POSIX ACLs to NFSv4
 - Read and write delegations
 - Kerberos 5/5i
- Features not in 2.6.18 kernel
 - Replications
 - Migration support

Why NFS for Database?

- **Less Complex**
 - Ethernet connectivity model
 - Simple storage provisioning & backup
- **Reduce the Cost of Storage Provisioning**
 - Amortize storage costs across servers
 - FlexClone® helps cloning master DBs for Test & Dev. Areas
- **Improved Oracle Administration**
 - Single repository
 - Recovering from Snapshot™ quick and reliable
-

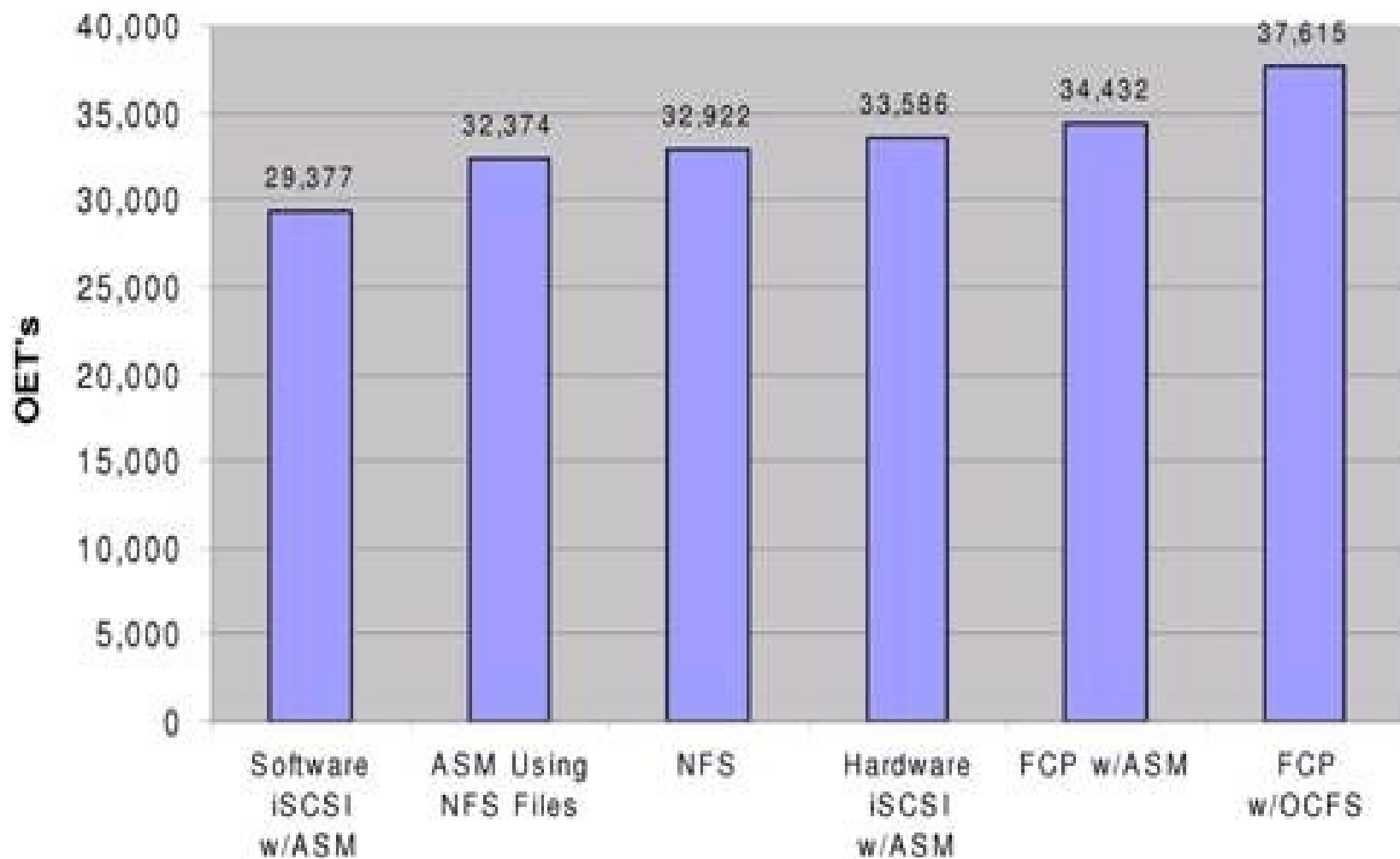
Why NFS for Database?

- **Better Performance**

- Data is cached just once, in user space, which saves memory – no second copy in kernel space.
- Metadata access for the clients are much quicker with less over-head
- Load balances across multiple network interfaces, if they are available.

Oracle Prefers NFS/NAS

Performance comparison with different Protocols



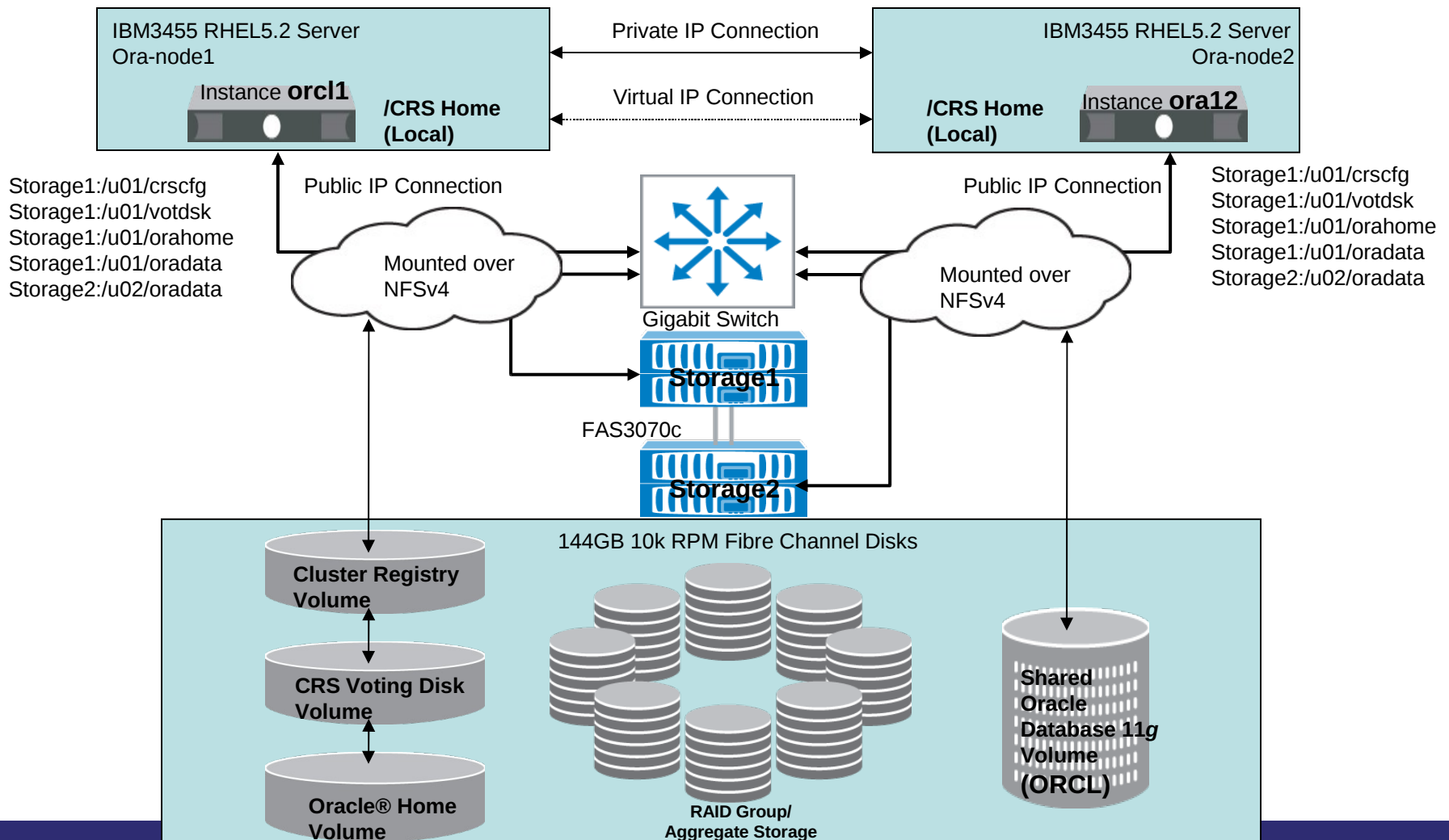
Why NFS Version 4 for a Database?

- NFSv4 will be the building block for scaling out implementations of Oracle11g over NFS.
 - Leased-based locking helps to clear or recover locks on event of a network or Oracle datafile outages.
 - Delegations would help performance for certain workloads
 - Referrals will allow a storage grid and a compute grid to mutually optimize I/O paths.
 - A storage system can tell a compute server which storage system can best service particular requests to facilitate grid-based scale-out.

Why Oracle11g over NFSv4

- **NFSv4 is the building block for all scale out implementations of Oracle11g over NFS.**
 - Leased-based locking
 - Helps to clear or recover locks on event of a network or Oracle datafile outages.
 - Referrals will allow a storage grid and a compute grid to mutually optimize I/O paths.
 - A storage system can tell a compute server which storage system can best service particular requests to facilitate grid-based scale-out.

Reference Architecture – 2 Node Oracle Database 11g RAC over NFSv4



Hardware Used for Oracle Database 11g RAC Setup

- Oracle® RAC nodes
 - x86_64 Dual Core 2.8Ghz AMD Opteron CPU
 - 10Gb RAM
 - 80Gb HDD SATA
 - 2Gb of Swap Space
- 1Gb (Gigabit) Switch
- NetApp® Storage
 - FAS3070 Cluster
 - 144Gb 10k RPM FC drives
 - 4Gb Fibre Channel back end shelf speed
 - DATA ONTAP 7.3

Software Used for Oracle Database 11g RAC Setup

- RHEL5 Update 2 – x86 64 bit architect
 - Update 2 was used due the the recent NFS performance enhancements
- Oracle® Database 11g database and clusterware
- Data ONTAP® 7.3 on NetApp® storage
- NFS Mounts are all over NFSv4

Service configuration for Oracle Database 11g RAC Setup

- Boot with non-XEN kernel
 - “libvirt” will be disabled
 - Creates interface call “virbr0” that has issues with Oracle® CRS install
- Disable “iptables” on the Linux® RAC nodes
- Synchronize Time with NTP on the RAC nodes and the NetApp® Storage

Network Transport used for Oracle Database 11g RAC Setup

- Use the TCP transport.
 - More reliable and low risk of data corruption and better congestion control compared to UDP
 - Retransmission happens in the transport layer instead of application layer
- Enlarge TCP window size for fast response
 - `net.ipv4.tcp_rmem = 4096 524288 16777216`
 - `net.ipv4.tcp_wmem = 4096 524288 16777216`
 - `net.ipv4.tcp_mem = 16384 16384 16384`
- Benefits:
 - This will increase the speed of the cluster interconnect and public network.

Mount Options Used for Oracle Database 11g RAC

- NFSv4 Protocol
 - Specify “-t nfs4” to ensure mounting over NFSv4
- Background mounts (bg)
 - Clients can finish booting without waiting for storage systems
- rsize=32768 wsize=32768
 - RHEL5.2 also supports 64k transfer size and up to 1Mb
- NetApp Storage
 - DATA ONTAP 7.3 uses up to 128kb block size

Mount Options Used for Oracle Database 11g RAC

- timeo
 - 600 is good for TCP
- Hard Mount
 - Default recommendation
 - Mandatory for data integrity
 - Minimizes the likelihood of data loss during network and server instability

Mount Options Used for Oracle Database 11g RAC

■ intr option

- Allows users and applications to interrupt the NFS client
- Be aware that this doesn't always work in Linux® and rebooting may be necessary to recover a mount point
- Use *soft* mount instead
- *Oracle has verified that using "intr" instead of "nointr" can cause corruption when a database instance is signaled (during a "shutdown abort")*

Mount Options for only Database mounts

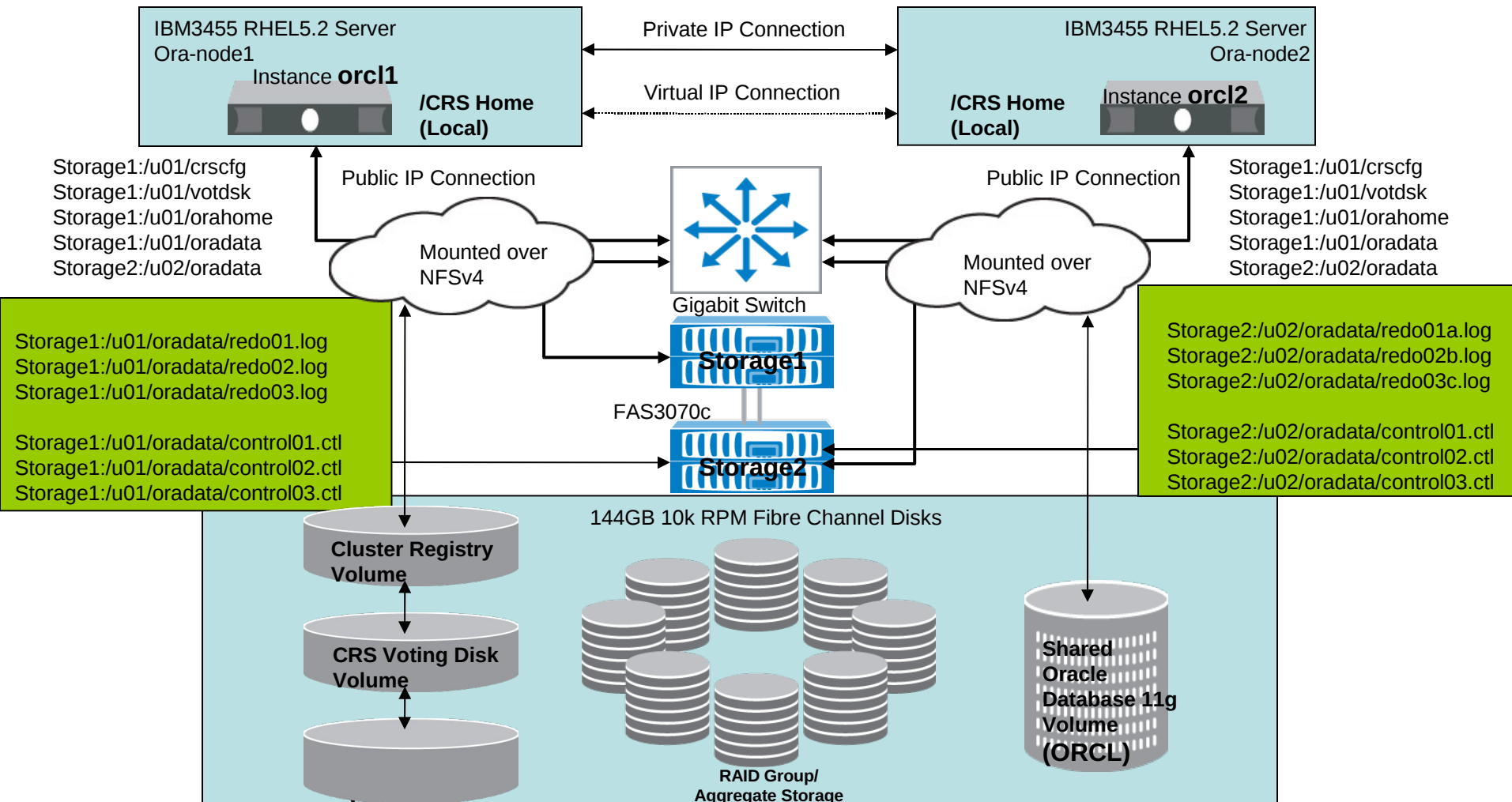
- “noac” option
 - Disables client side caching and keeps file attributes up to date with the NFS Server
 - Shorthand for “actimeo=0,sync”
 - **Bug** - https://bugzilla.redhat.com/show_bug.cgi?id=446083
 - **Patch** - <http://article.gmane.org/gmane.linux.nfs/20074>
- Set the “sunrpc.tcp_slot_table_entries” to 128
 - Benefits:
 - Removes a throttle between the Linux® nodes and the backend storage system
 - Allows a single Linux box to drive substantially more I/O to the backend storage system

ORACLE_HOME on Shared Storage

Benefits:

- Redundant copies are not needed for multiple hosts.
 - Extremely efficient in a test/dev environment where quick access to the Oracle® binaries from a similar host system is necessary.
- Disk space savings.
- It is easier to add nodes.
- Patch application for multiple systems can be completed more rapidly.
 - For example, if testing 10 systems that you want to all run the exact same Oracle DB versions, this is beneficial.

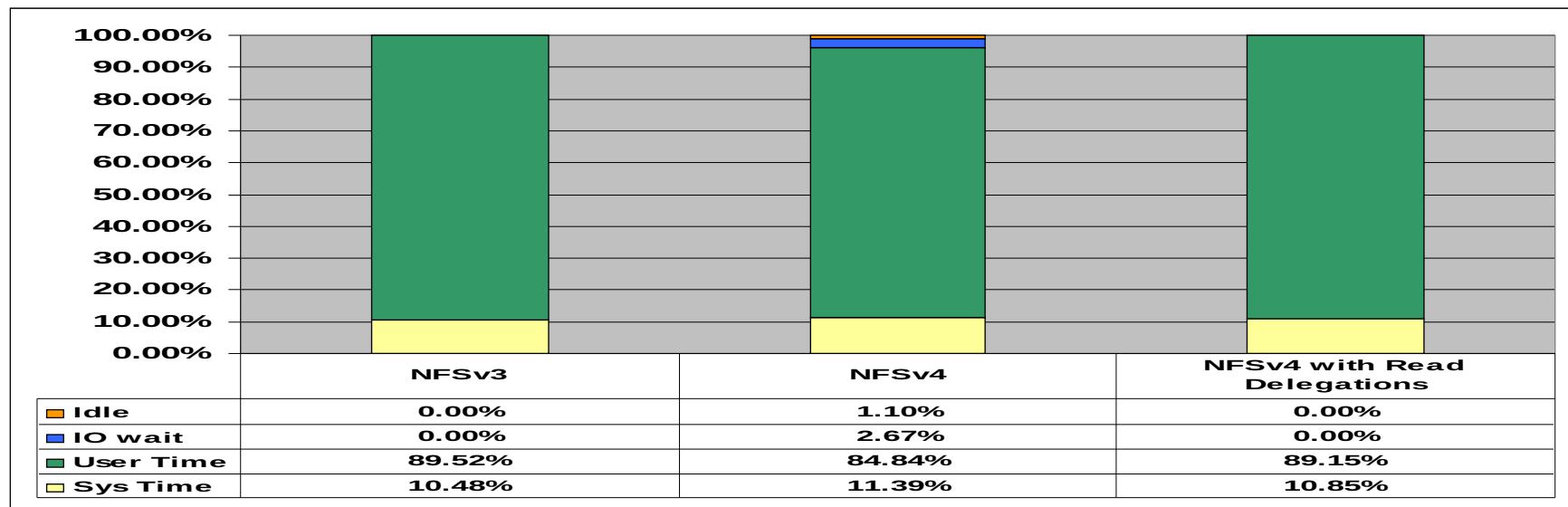
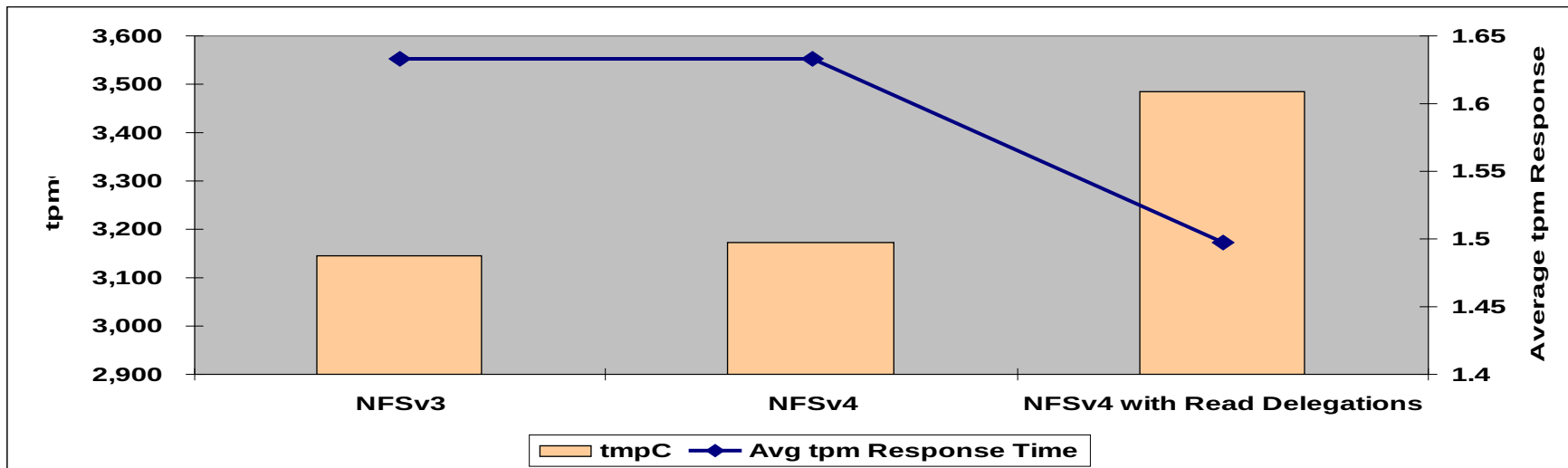
Reference Architecture – 2 Node Oracle Database 11g RAC over NFSv4

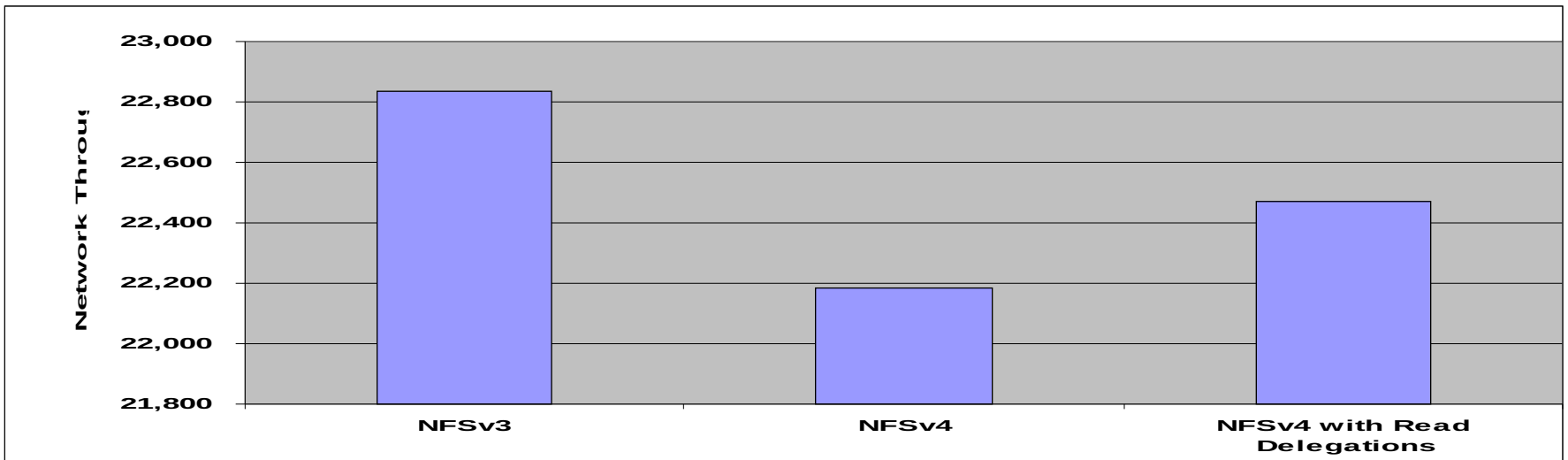
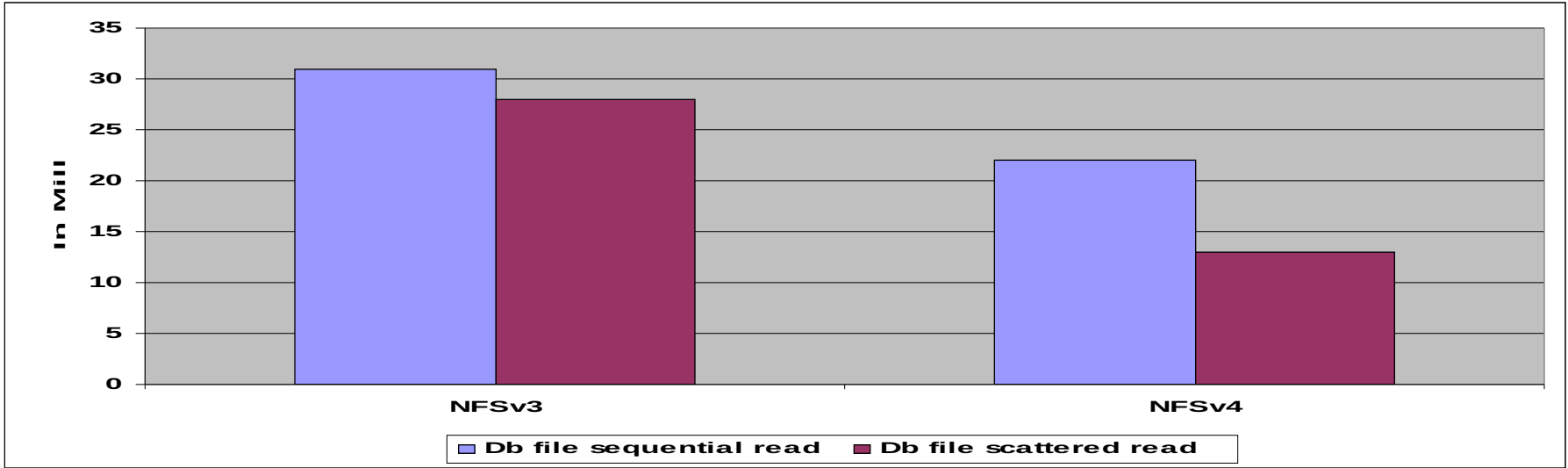


Oracle Database 11g CRS Timeout Settings - Best Practices

- OCR and CRS voting files have to be multiplexed
 - A copy of both the files has to reside on each storage
- Three CSS parameters have to be set
 - `misscount` – 120 seconds (30 secs default)
 - `disktimeout` – 200 seconds (default)
 - `reboottime` – 3 seconds (default)

NFsv3 & NFSv4 Comparison - Performance Analysis





NetApp's Linux Community

- NetApp's business model depends on superior client behavior and performance
- NetApp is driving Linux® Client Performance and scalability, sponsored by NetApp at CITI, Univ. of Michigan
- Build expertise with Linux clients and storage systems to help our customers get the most from our products
 - Explore and correct Linux NFS client and OS issues
 - Establish positive relationship with Linux community
 - Develop internal resources for customer-facing teams

NetApp's Linux Community

- Linux Certification Testing Results
 - Linux 10g/11g RAC testing over NFSv3/NFSv4
 - Linux FCP and iSCSI testing
 - Linux NFSv4 client support
 - Linux certification with NFS
 - Linux Best Practices document
 - <http://www.netapp.com/library/tr/3183.pdf>

Linux Leadership with NetApp

- Mature NetApp Solution for Oracle® on Linux®
 - Database Consolidation
 - High Availability
 - Backup and Recovery
 - Disaster Recovery
- Oracle Database 10g/11g certification with RedHat Linux and NetApp® Storage over NFSv3/NFSv4
- Partnership and Performance Testing Results
 - RedHat partnership agreement

ORACLE®



Q&A

Email: bikash@netapp.com
steved@redhat.com

BACKUP SLIDES

Mount Options for Oracle Database 11g RAC Components

- CRS and voting disk mount options
 - `rw,bg,hard,rsize=32768,wspace=32768,proto=tcp,noac,nointr,timeo=600`
- Oracle® Home and Oracle data mount options
 - `rw,bg,hard,rsize=32768,wspace=32768,proto=tcp,actimeo=0,noitr,timeo=600`

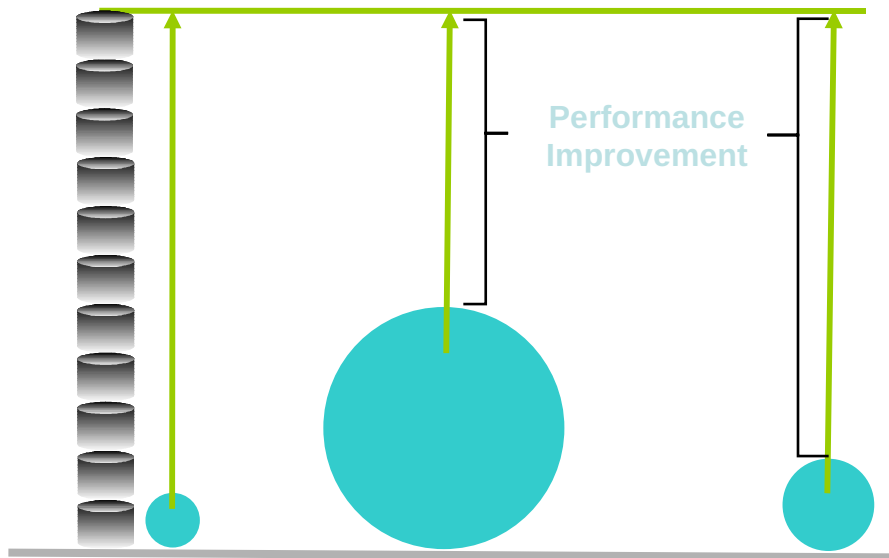
Required Linux RPMs to install Oracle Database 11g RAC

- binutils-2.15.92.0.2-21
- compat-db-4.1.25-9
- compat-libstdc++-33-3.2.3-47.3
- elfutils-libelf-0.97.1-3
- elfutils-libelf-devel-0.97.1-3
- glibc-2.3.4-2.25
- glibc-common-2.3.4.2-25
- glibc-devel-2.3.4.2-25
- gcc-3.4.6-3
- gcc-c++-3.4.6-3
- libaio-0.3.105-2
- libaio-devel-0.3.105-2
- libstdc++-3.4.6-3.1
- libstdc++-devel-3.4.6-3.1
- make-3.80-6
- pdksh-5.2.14-30.3
- sysstat-5.0.5-11
- unixODBC-devel-2.2.11-7.1
- unixODBC-2.2.11-7.1

Storage Resiliency – High Availability

- Clustered Failover in the event of hardware failure
- Less cluster failover/giveback times
- Transparent to NFS clients
- Nondisruptive Data ONTAP® upgrades without any user downtime
- Reduced TCO and maximized Storage ROI

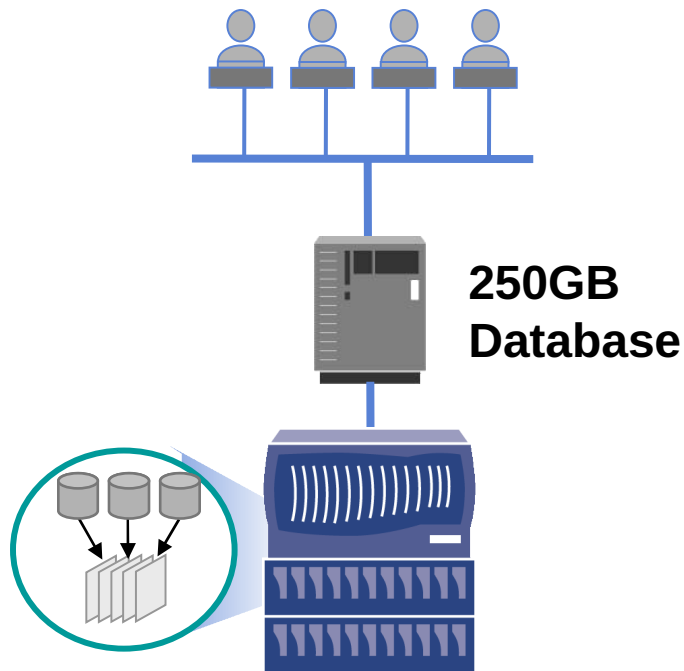
Database Performance Tuning with FlexVol



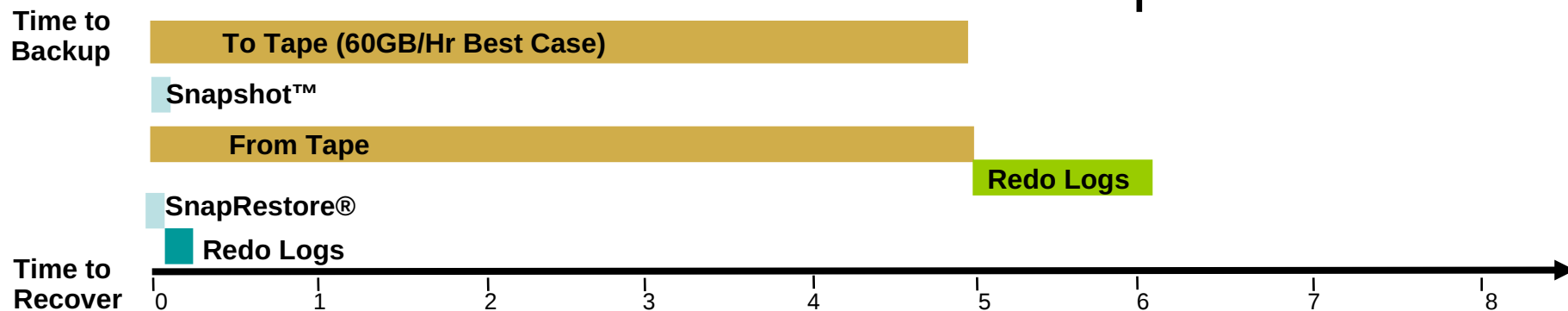
Benefits

- Improves database performance quickly and measurably
- Uses all available spindles for data and transaction logs
- Spindle sharing makes total aggregate performance available to all volumes
- Automatic load shifting

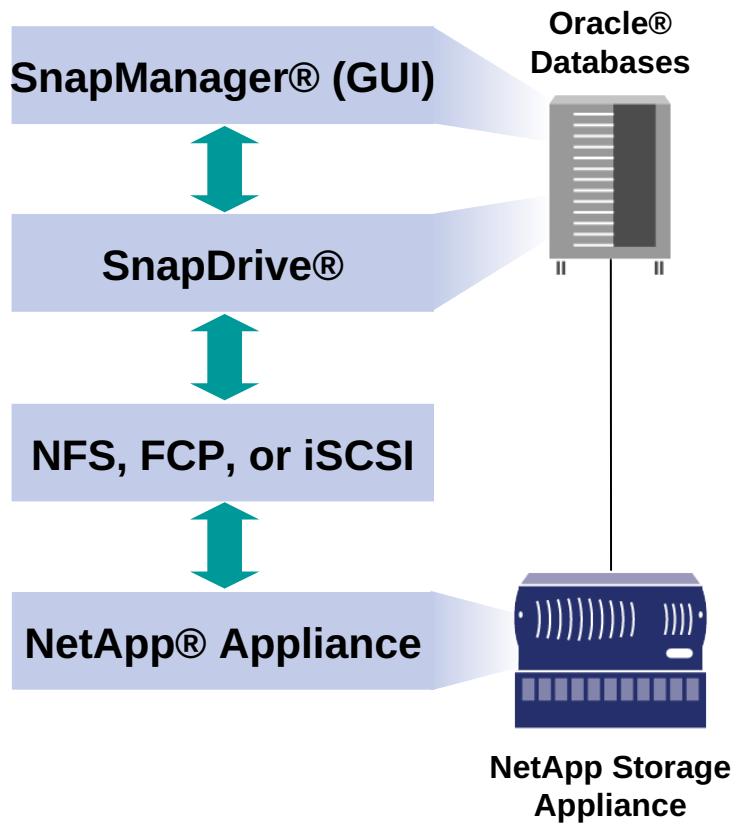
Backup and Recovery



- Significant time savings
- Stay online
- Reduce system and storage overhead
- Consolidated backups
- Back up more often



SnapManager for Oracle



- Automated, fast, and efficient
- Uptime AND performance
- Simplify backup, restore, and cloning
- Tight Oracle Database 10g integration
 - Automated Storage Manager (ASM)
 - RMAN