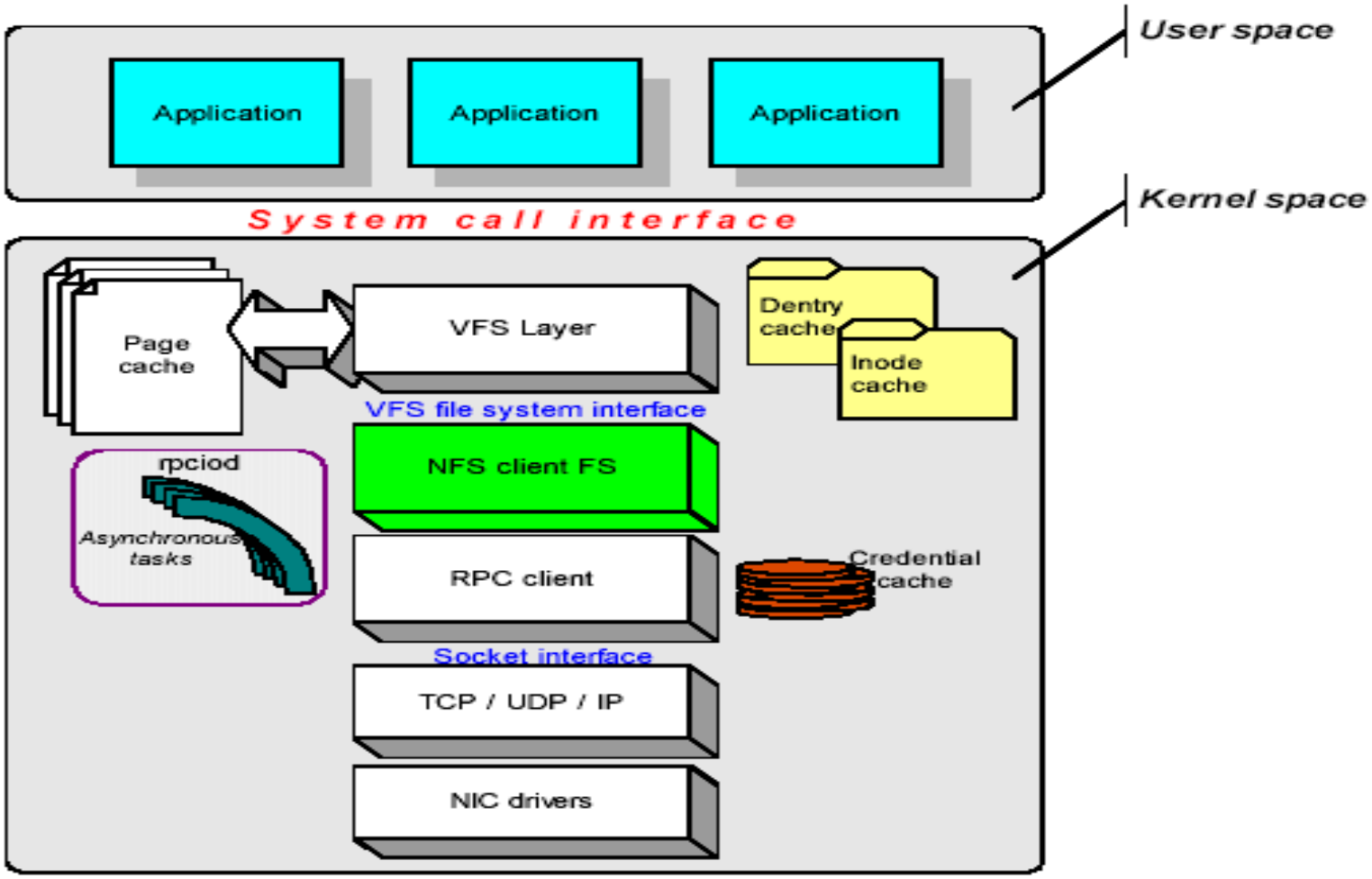
Implementing Oracle11g Database over NFSv4 from a Shared Backend Storage

Bikash Roy Choudhury

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 NFSv4 Client in the 2.6.18-88 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?

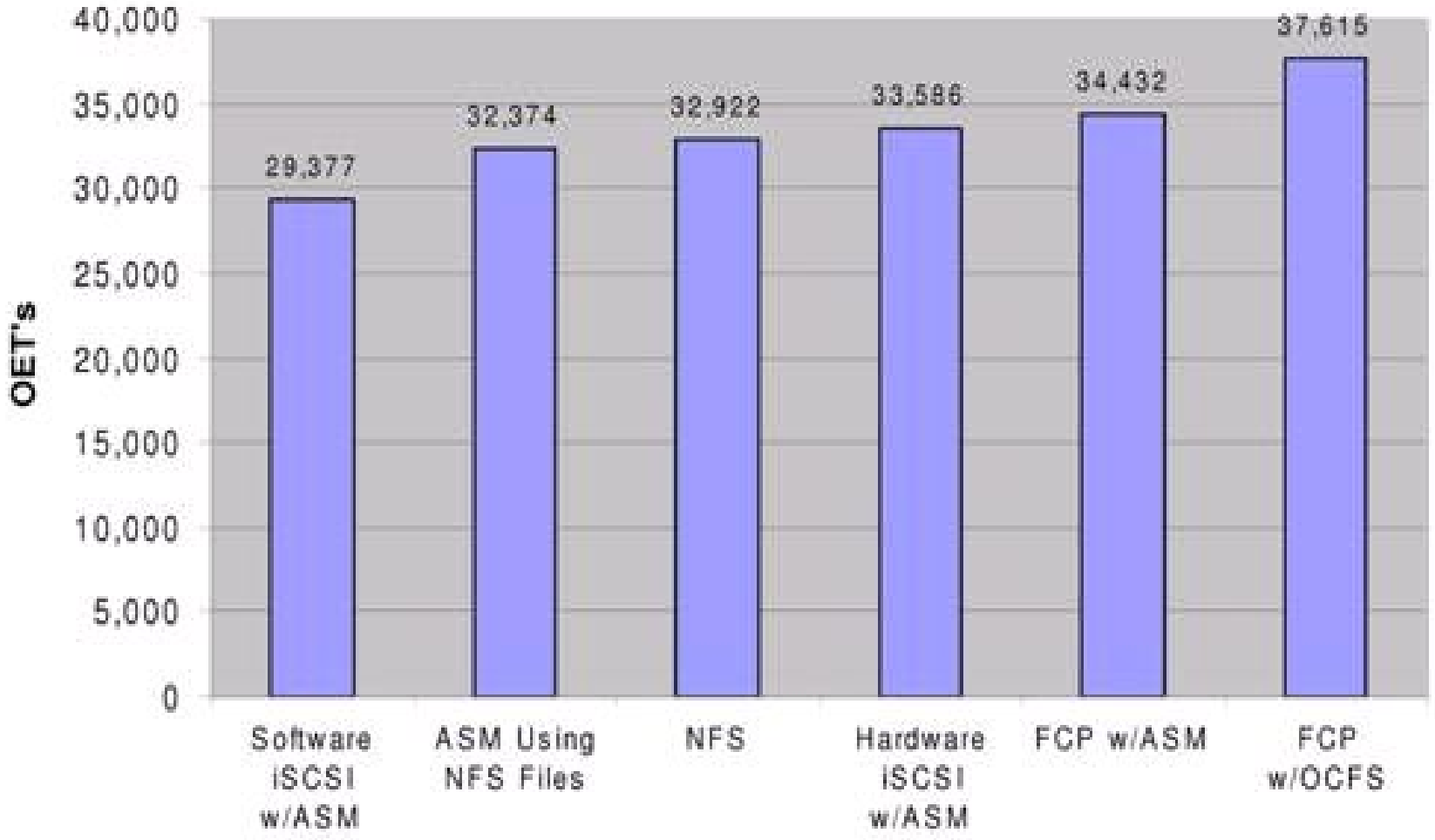
- **Less Complex**
  - Ethernet connectivity model
  - Simple storage provisioning & backup
- **Reduce the Cost of Storage Provisioning**
  - Amortize storage costs across servers
  - FlexClone® helps cloning master DBs for Test & Dev. Areas
- **Improved Oracle Administration**
  - Single repository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

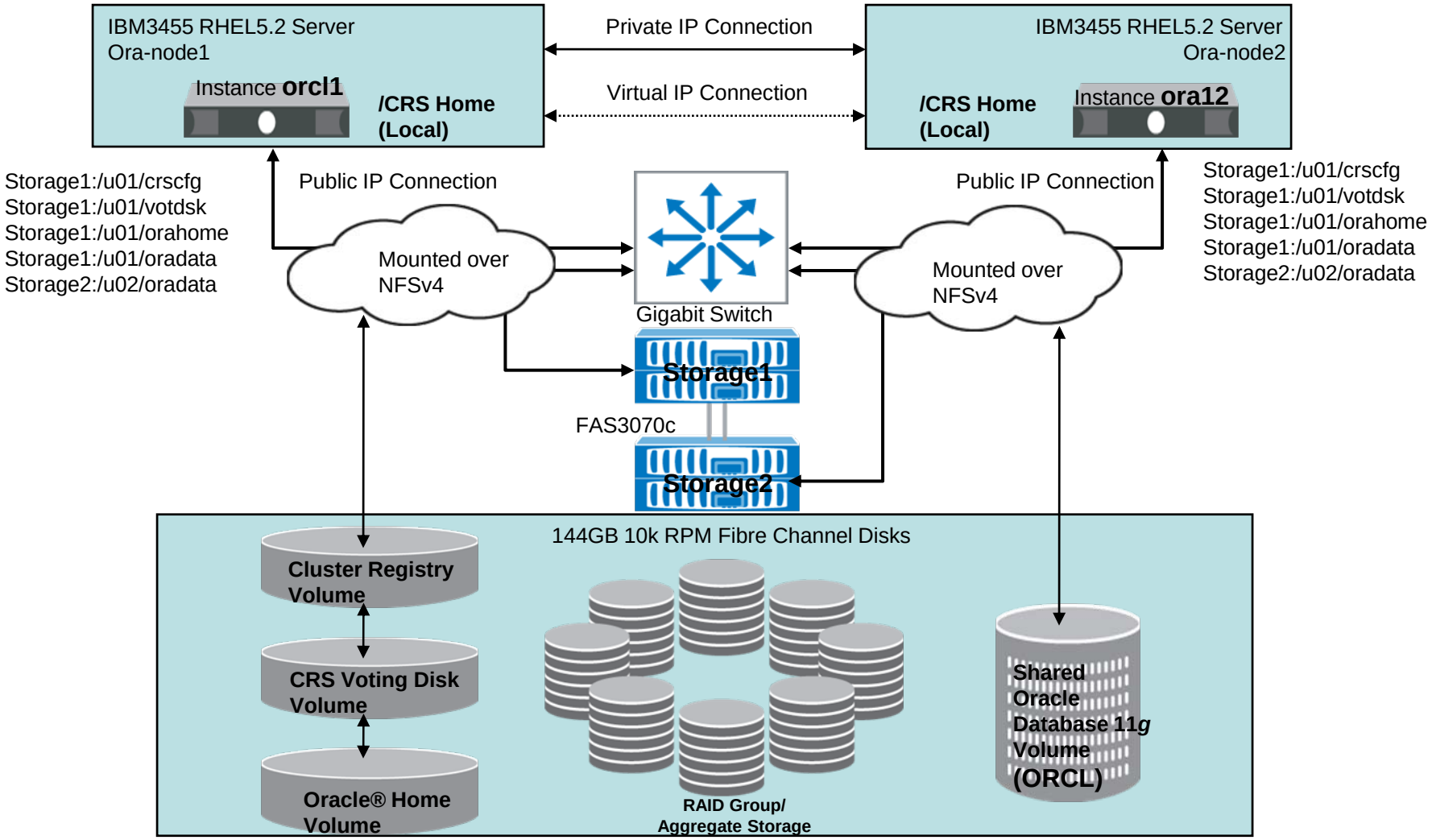
Performance comparison with different Protocols



Why 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.**
  - The redirection feature allows a storage grid and a compute grid to mutually optimize I/O paths.

2 Node Oracle11g RAC over NFSv4 -Reference Architecture



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

- **2.6.18-88.el5xen #1 SMP** – x86 64 bit
  - This kernel 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 Oracle11g 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 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 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**
  - 2.6.18-88 kernel 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 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 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")*
    - *"nointr"* is recommended

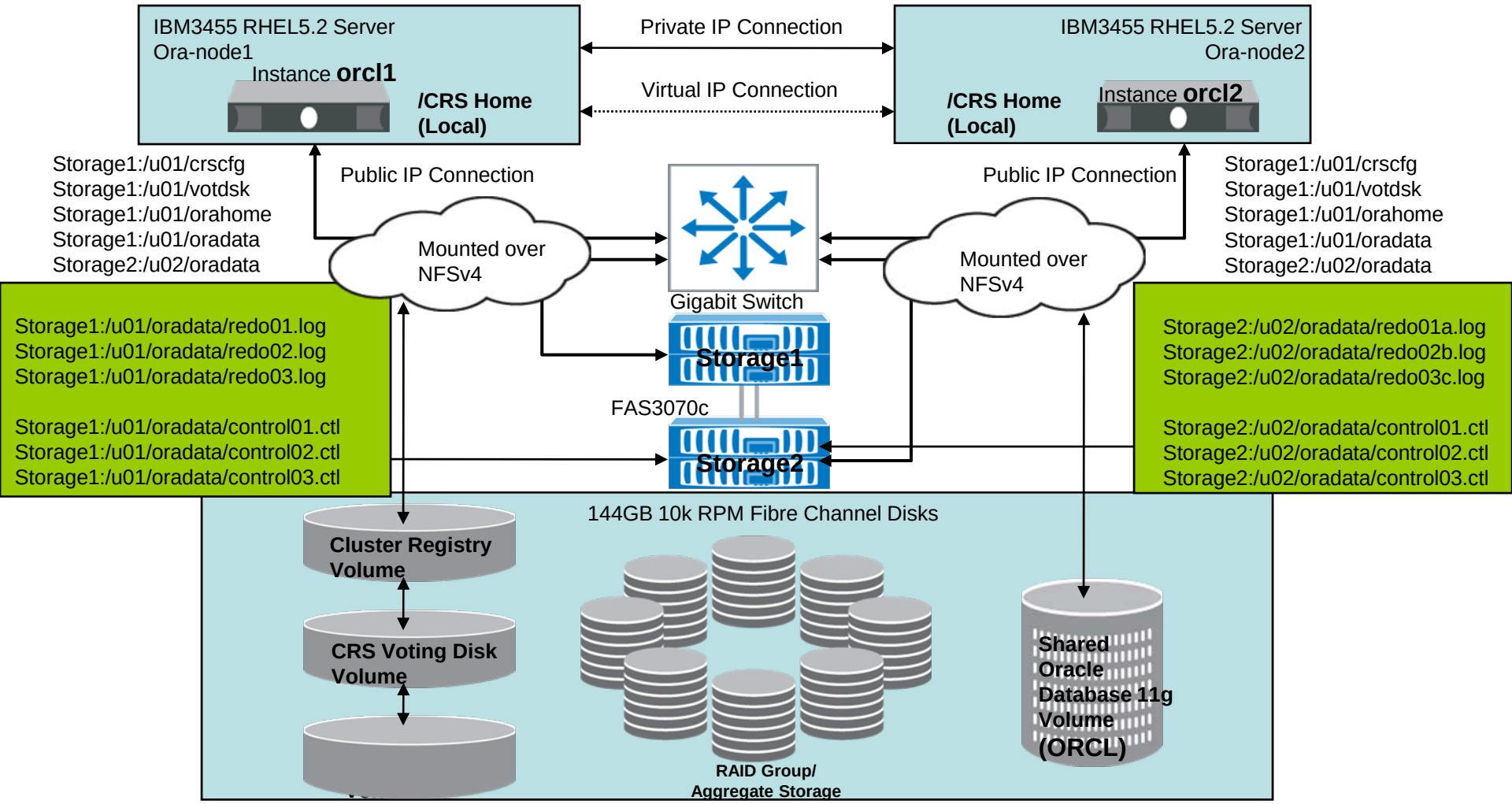
Mount Options for only Database mounts

- **"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.

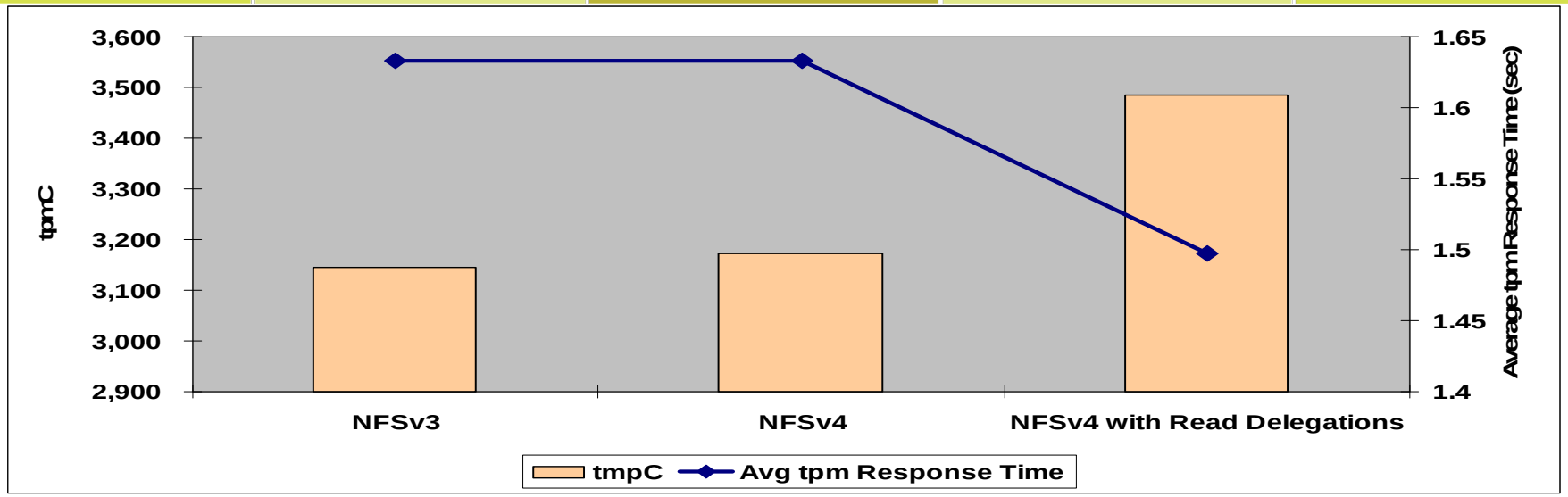
Reference Architecture – 2 Node Oracle Database 11g RAC over NFSv4

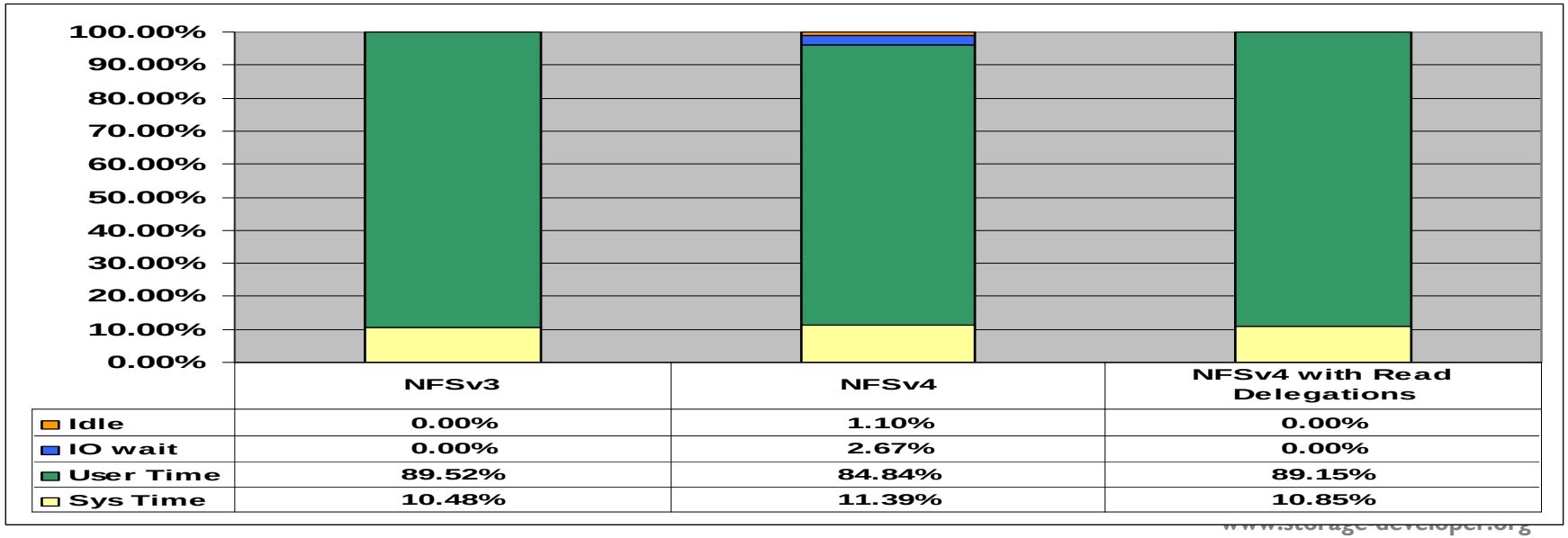


Oracle Database 11g CRS Timeout Settings – Best Practices

- 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)

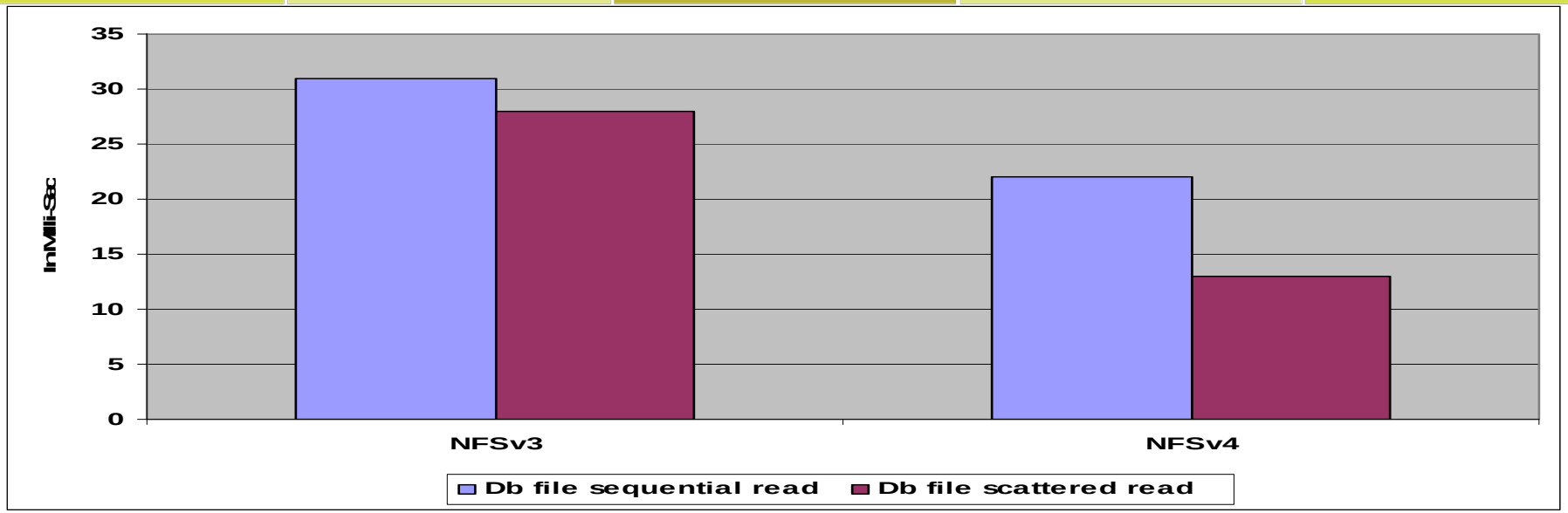
NFsv3 & NFSv4 Comparison – Performance Analysis

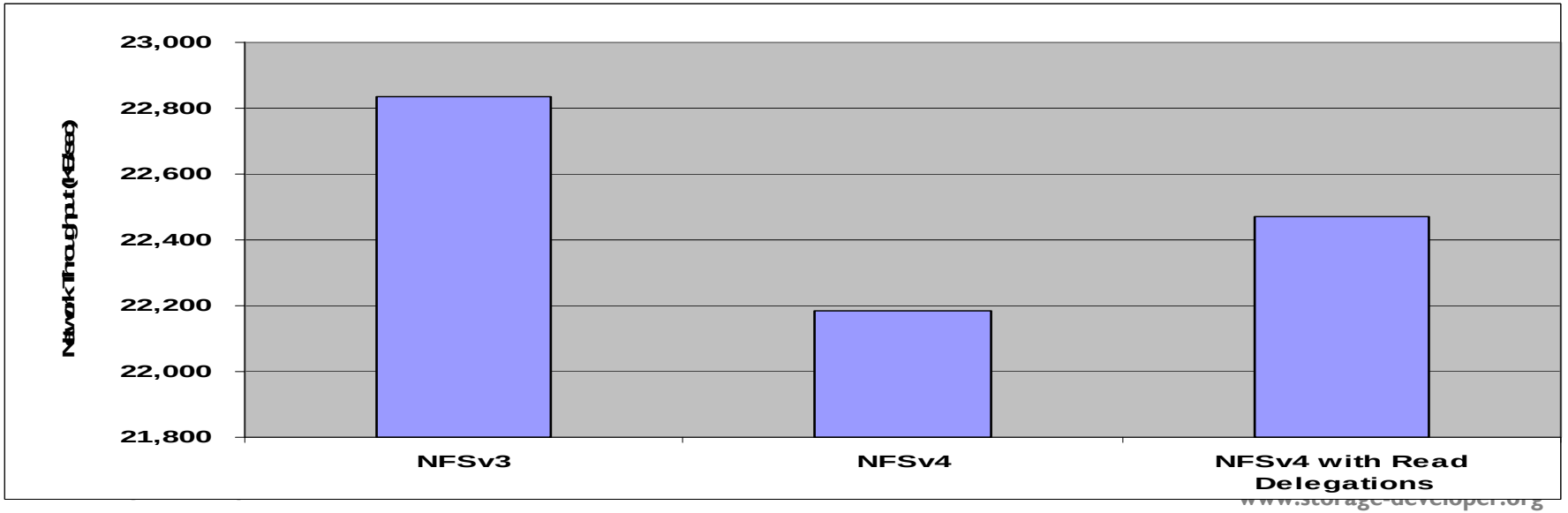




	NFSv3	NFSv4	NFSv4 with Read Delegations
Idle	0.00%	1.10%	0.00%
IO wait	0.00%	2.67%	0.00%
User Time	89.52%	84.84%	89.15%
Sys Time	10.48%	11.39%	10.85%

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 10*g*/11*g* 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**
- **Unbreakable and Enterprise ready**
  - –NetApp, Oracle, Oracle Enterprise Linux (OEL)
- **Partnership and Performance Testing Results**
  - RedHat partnership agreement

Thank You

Q&A

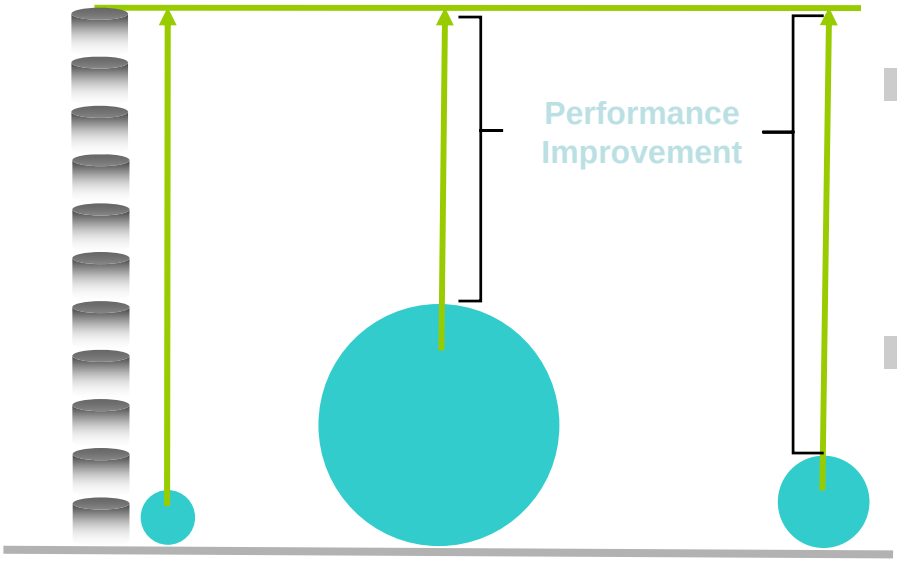
Email:bikash@netapp.com

BACKUP SLIDES

Storage Resiliency – High Availability

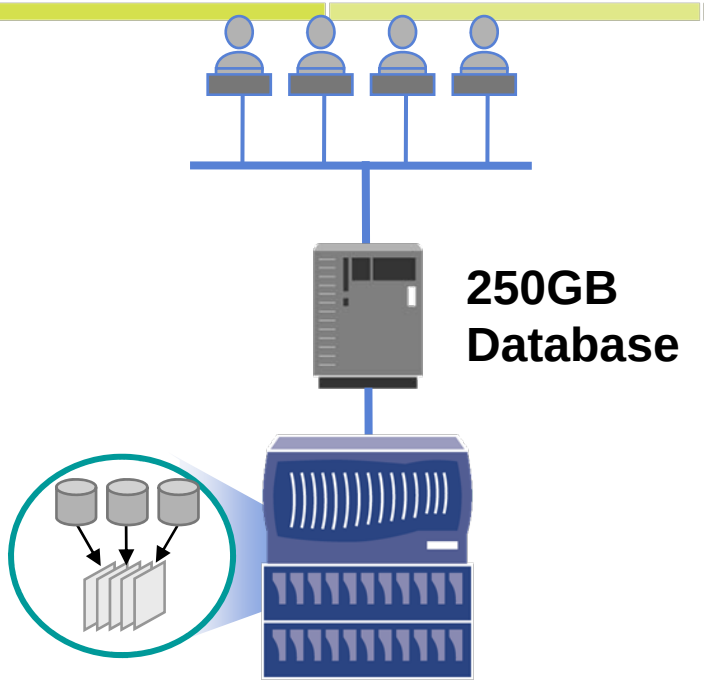
- Clustered Failover in the event of hardware failure
- Less cluster failover/giveback times
- Transparent to NFS clients
- Nondisruptive Data ONTAP® upgrades without any user downtime
- Reduced TCO and maximized Storage ROI

Database Performance Tuning with FlexVol

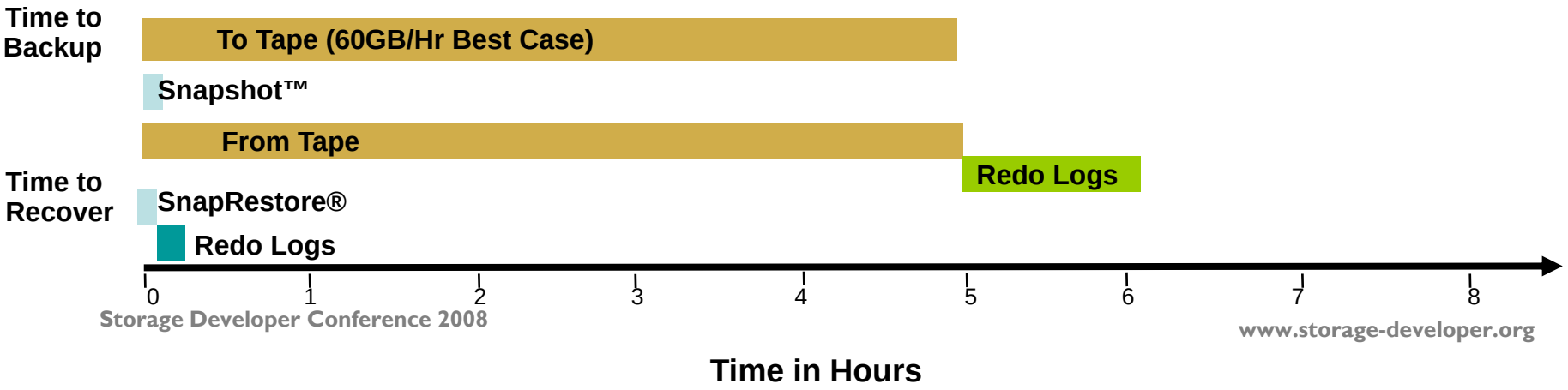


- Benefits
- Improves database performance quickly and measurably
- Uses all available spindles for data and transaction logs
- Spindle sharing makes total aggregate performance available to all volumes
- Automatic load shifting

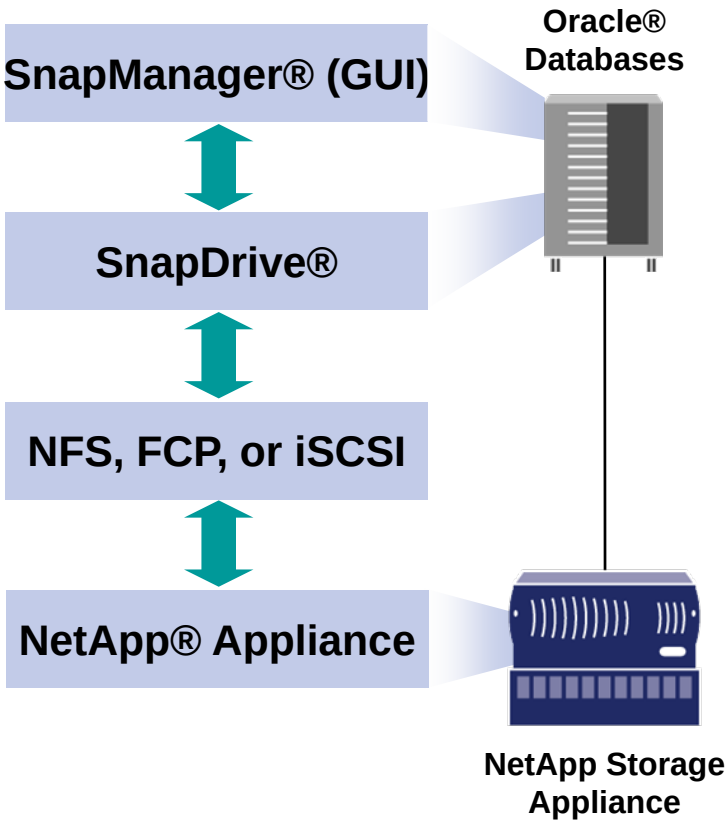
Backup and Recovery



- 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

Thank You