Building Oracle on Flash Arrays

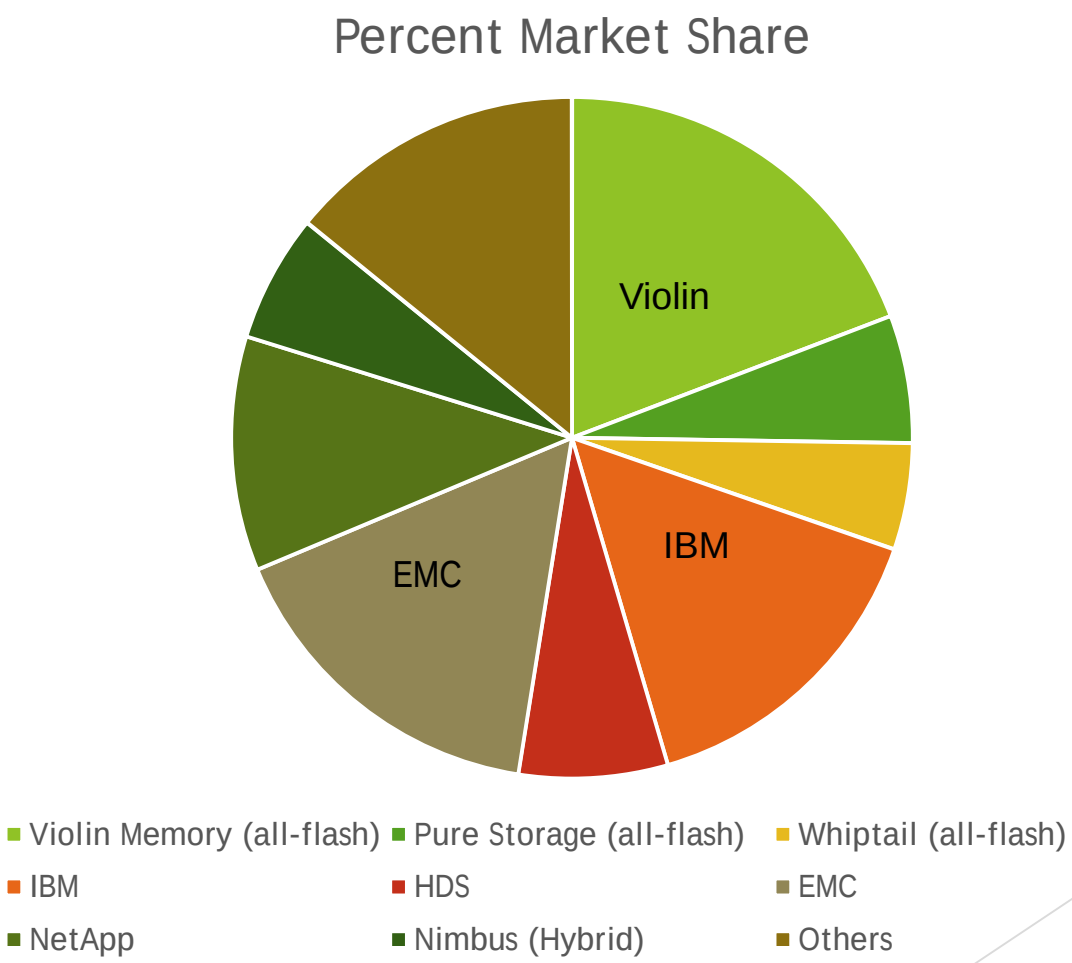
Gilbert Standen
Oracle Engineering Team
Violin Memory Corporation

Presented at NYOUG
March 12, 2014

About Presenter

- 20 Years Oracle DBA by 2015
- Presenter NYOUG
- Presenter Oracle Communities with 20:20 Foresight, Australia
- Build/Upgrade Architect:
  - Pharmaceutials (Pfizer, Schering-Plough, Purdue, McNeil)
  - Financials (Currenex-State Street, Federated Investors, Interactive Data)
  - Government (USDA, USGS, DOI)
- Can be reached at **gstanden@vmem.com**
- Work in a Great Company…Violin Memory
- Work with the awesome Violin Memory Oracle Team!

Violin Memory
The All-Flash Array Market Leader



The Violin Oracle Team "Past and Present"

This presentation is made possible by

Randy Cook

Nathan Fuzi

Chris Buckel

Ashminder Ubhi

Matt Morris

Jon Bennett, Violin Memory, Founder & CTO

Violin Memory Incorporated

NYOUG, its Members, and Corporate Sponsors

Caryl Lee, Michael Olin

THANK YOU ALL FOR ATTENDING THIS PRESENTATION!

I BELIEVE IT WILL BE WELL WORTH YOUR TIME.

Examples of Other Verified All-Flash Database Solutions

- Running Oracle Virtual Machine (OVM) on flash storage
- Running Oracle Applications on flash storage
- **Running Red Hat Enterprise Linux 6 on flash storage**
- Running AIX and PowerVM on flash storage
- Running Vmware ESXi on flash storage
- Running Microsoft SQL Server 2012 on flash storage
- Running Cassandra on flash storage
- Running Hadoop and Hbase on flash storage
- Running IBM GPFS on flash storage
- And more…but this paper is about Oracle DB on flash
- Contact me for help with any of these or any other idea you have !!

Why Oracle on Flash?

- Dramatic Improvement of Application Response
- Diminishing Results from Code Tuning
- Huge Growth in Data Quantity
- Lower Energy Consumption / Green
- CPU Utilization, Count, Licensing Fee Improvements
- A 32-core server can often be downgraded to 16-core after all-flash storage implementation resulting in an annual savings of $47,500 x 16 cores = $760,000 yearly savings on Oracle per-core licensing fees, while delivering faster performance also.
- All-flash storage ROI can be lightening fast too.

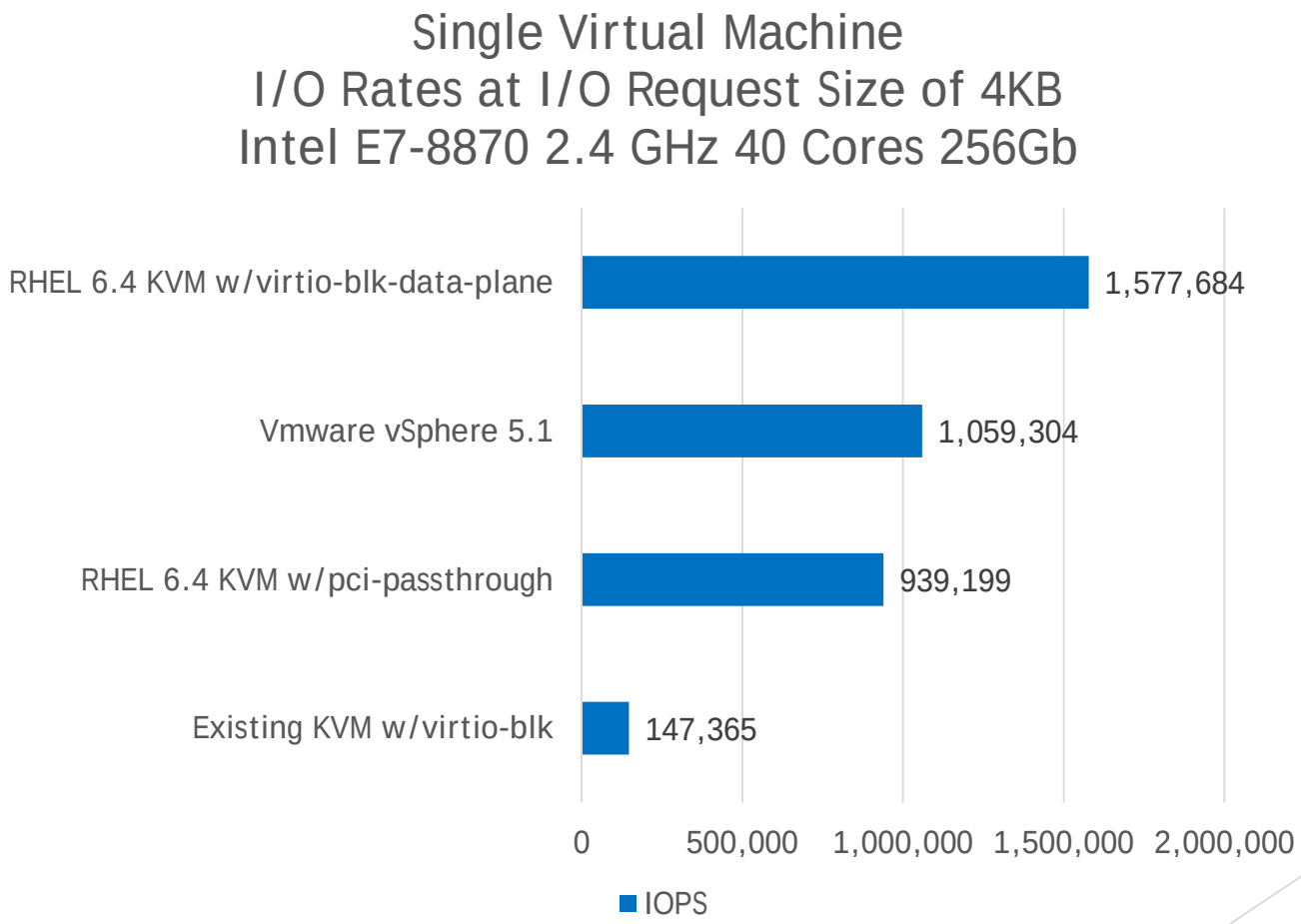
Options for Oracle Low-Latency on Flash Bare Metal & Virtualized

- Bare metal server on flash storage array.
- VM on flash storage array (KVM, Vsphere)
- Flash caching “helper” physical devices to speed up reads for legacy spinning platter disk.

Can VM's run low-latency applications on all-flash ?

- Yes they can, and Yes they do !
- There are really only two choices for high-IOPS, low-latency VMs/hypervisor running on all-flash storage
  - RedHat OpenStack
  - VMWare Vsphere
- This presentation posits that OpenStack is the best choice.
- RedHat OpenStack KVM Hypervisor provides sole-source clear vendor support for all stack layers: OS, hypervisor, VM.
- KVM holds 5 of the 6 SPECvirt_sc2013 VM IOPS world records.
- VMWare Vsphere holds 1 of the 6 SPECvirt_sc2013 IOPS world records, and is a distant second to the KVM current world record.
- VMWare Vsphere is also an option but results in multiple vendors for hypervisor, OS, and VM, and so support blame game is possible.
- OpenvSwitch software defined network (SDN) can be leveraged in both solutions to build an all-software solution.

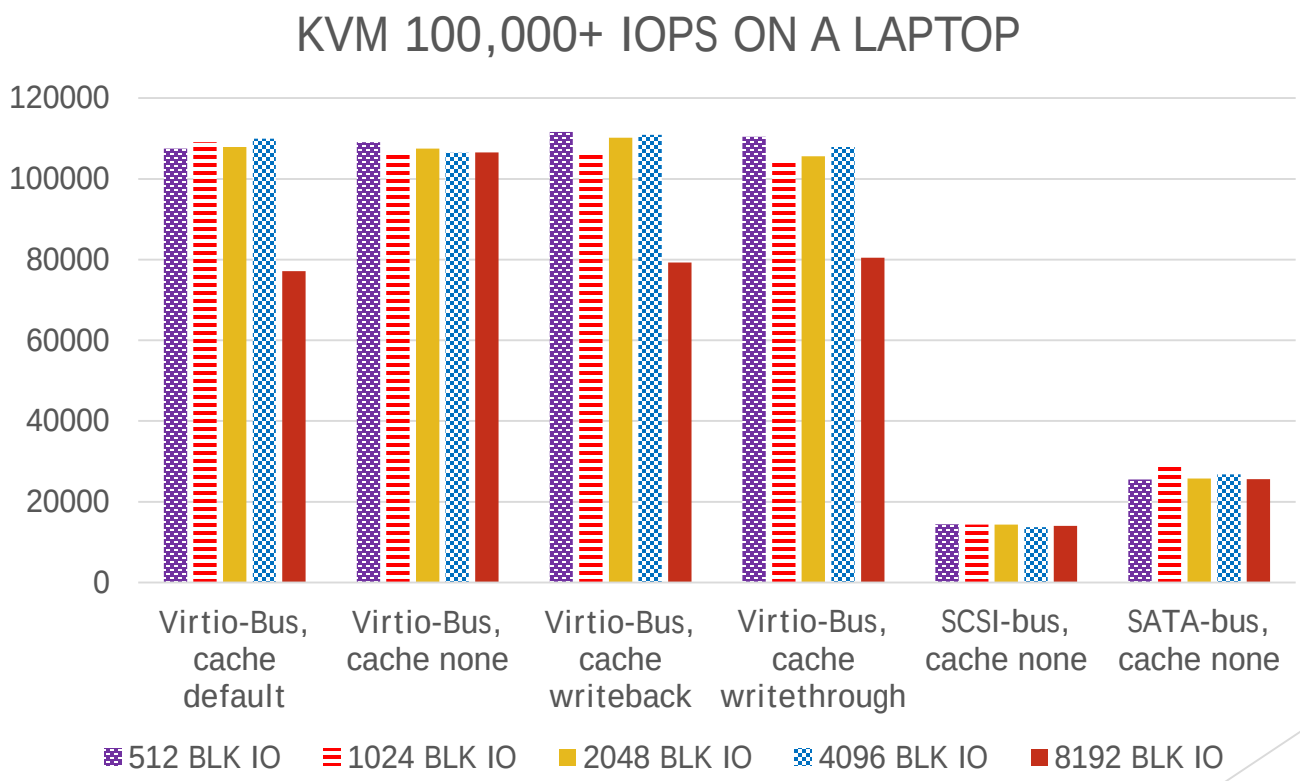
KVM Breaks 1.5M+ IOPS in a VM



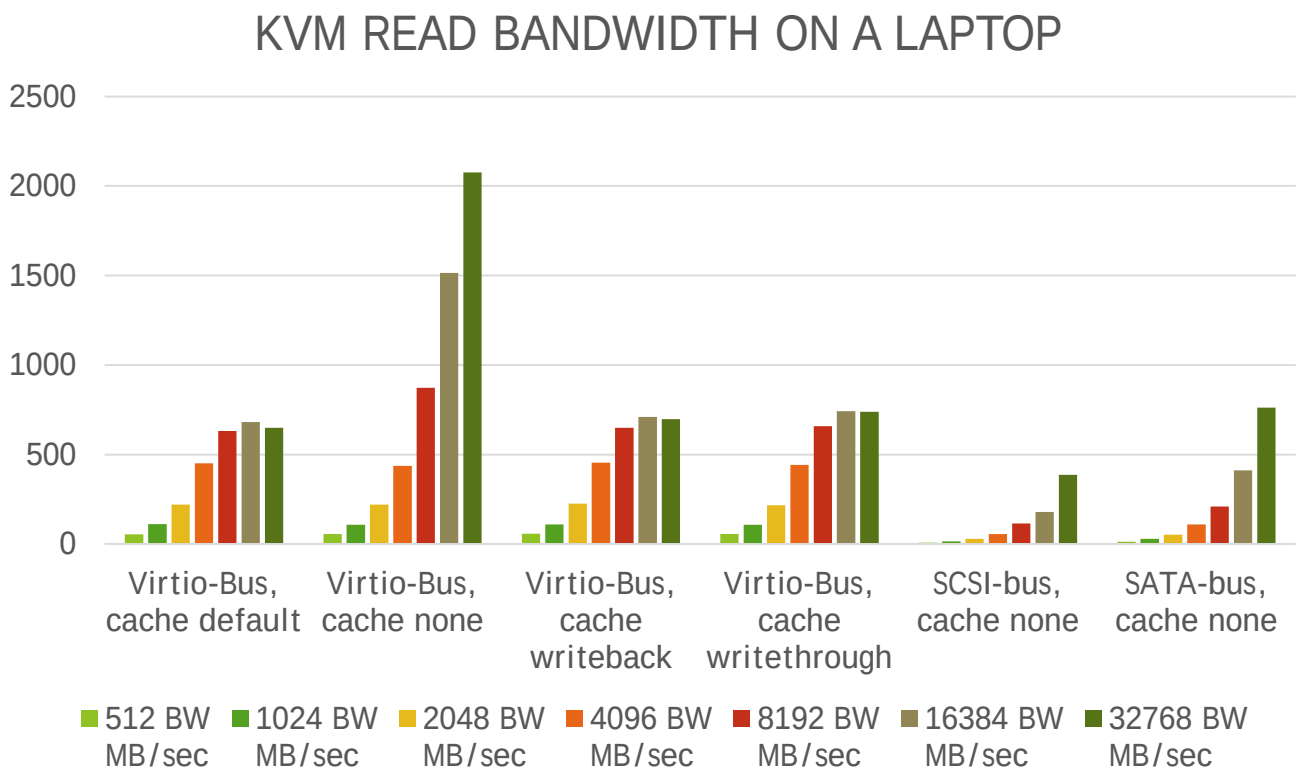
Virtualized Low Latency

- To give you an idea of what is possible with all-flash storage + KVM + virtio-blk-data-plane, let's look at a VM setup that I built on my laptop and its peformance.
- Specifications and Equipment:
  - Toshiba Q-Pro THNSNJ512GCST toggle-MLC 19nm NAND Flash SSD
  - Lenovo W520 mobile workstation 32Gb RAM, Sandy Bridge CPU
  - Intel Quad Core i7-2720QM 2.20 GHz (VT-d, VT-x, EPT)
  - Ubuntu Linux 3.11.10.3-iommu-pci #1 SMP x86_64 (custom-built)
  - Supported in KVM version 1.4 and higher
  - IOMMU and PCI_STUB custom-built into host linux kernel
  - Oracle Enterprise Linux 6.5 guest
  - KVM *.img raw disk image file for VM backing

KVM
Fig. 1: virtio-blk-data-plane IOPS



KVM
Fig. 2: virtio-blk-data-plane BW



KVM virtio-blk-data-plane

- Culmination of collaborative effort between:
  - IBM Linux Technology Center Performance Team
  - Red Hat KVM DevelopmentTeam
- Dedicated thread for each image device together with Linux AIO host kernel support for IO processing bypassing QEMU block layer
- Avoids time-consuming global mutex lock acquire inside QEMU
- The virtio-blk-data-plane also exploits the *ioeventfd / irqfd* mechanism, which decouples the IO processing from the guest execution.
- Only raw image files are supported (typical extension is *.img)
- Storage migration, hot unplug, I/O throttling, image streaming, and drive mirroring are currently not supported

- "***The most exciting development in QEMU in the last five years!!***"
–Anthony Liguori, QEMU Maintainer

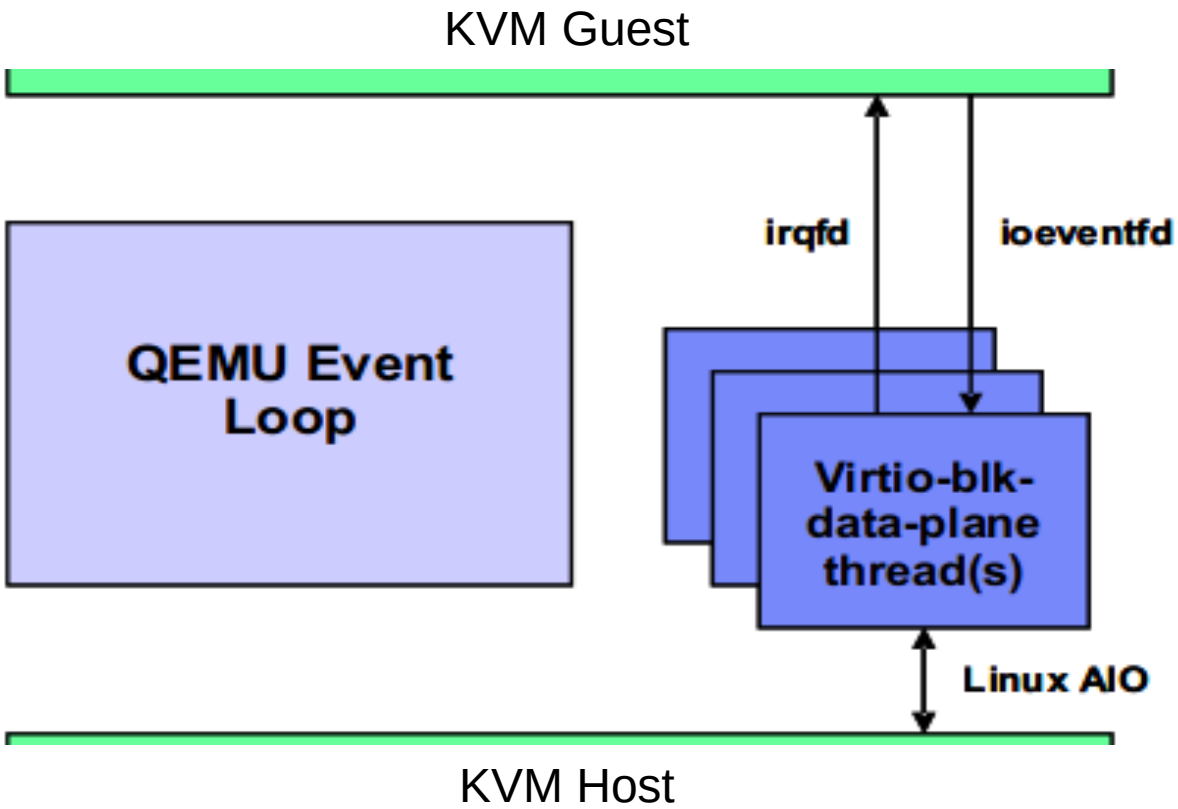


Figure: schematic virtio-blk-data-plane

RedHat OpenStack KVM with virtio-blk-data-plane provides the best single-vendor-supported world-record IOPS performance on par with bare metal, making OpenStack/KVM the low-latency virtualization platform of choice.

Technology Alternatives

- Engineered vRAID NAND Flash SAN
  - Integrated vRAID switched-memory Raid 3
  - Ensures all I/O serviced by all raid group members (VIMMs)
  - vRAID with Raid 3 ensures there is no parity-write penalty
  - vRAID ensures that VIMM erase will never slow flash RW.
  - Erase/update in parallel with sustained steady IOPS
  - Allows control of block size independent of user IO size
  - 4k data+1k parity aligns to 4+1 / 5-vimm raid groups
- Traditional HDD form-factor “SSD” SAN
  - Legacy SATA slower controllers outpaced by NAND
  - Legacy external raid solutions using legacy aggregation
  - Serial erase/update with write-cliff effects

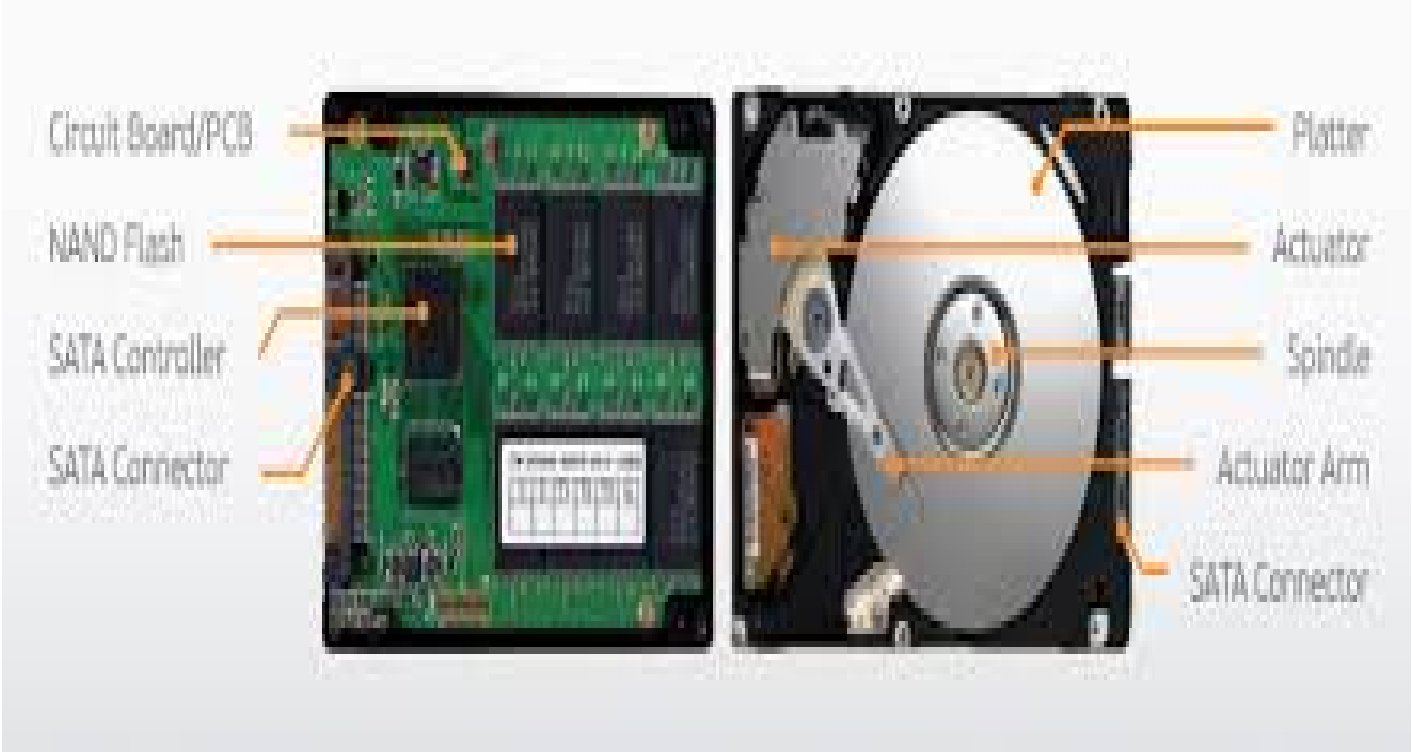
Example vRAID NAND Flash Array

Storage engineered to maximize NAND flash performance

Storage built for NAND to address and use its unique features

Legacy HDD NAND Form Factor

Legacy HDD Enclosure modified to support NAND Flash



- Typically uses a software layer to handle GC
- Suffers from write-cliff effects
- **CANNOT** guarantee that erase/update wnever affects reads
- Uses some form of log file system to provide product features
- Has relatively more inefficient GC leading to write-cliff
- Drives are aggregated using legacy equipement.

vRAID

Conceptualizing Hardware vRAID in Flash Storage Arrays

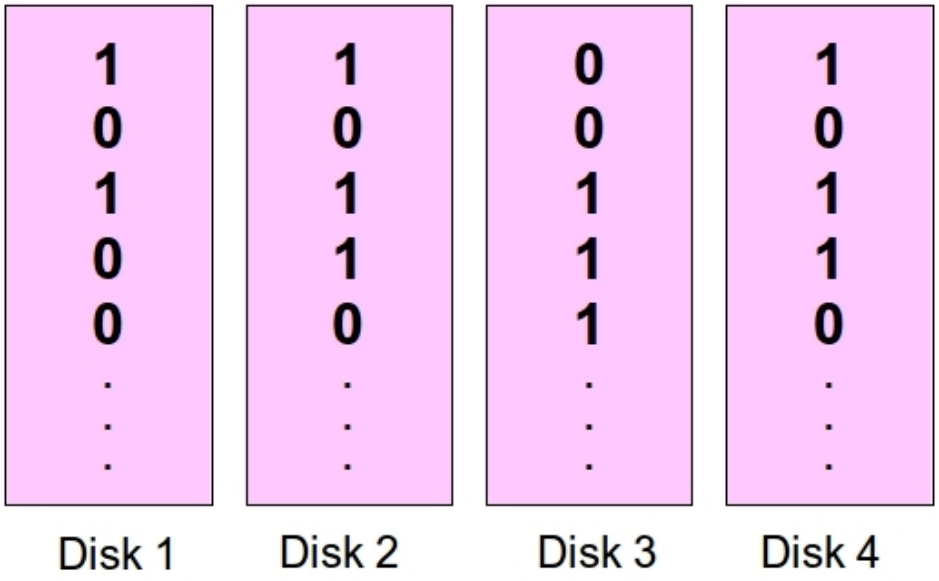
Performance Optimization Considerations

Flash Storage vs. Spinning Platter Storage

- Flash storage performs best when
  - (a) block erase is avoided, for example, an empty array where block erase is not yet needed
  - (b) block erase is bypassed, for example, when using a single parity disk during erase
  - (c) reads are massively parallel, such as from a raid array.

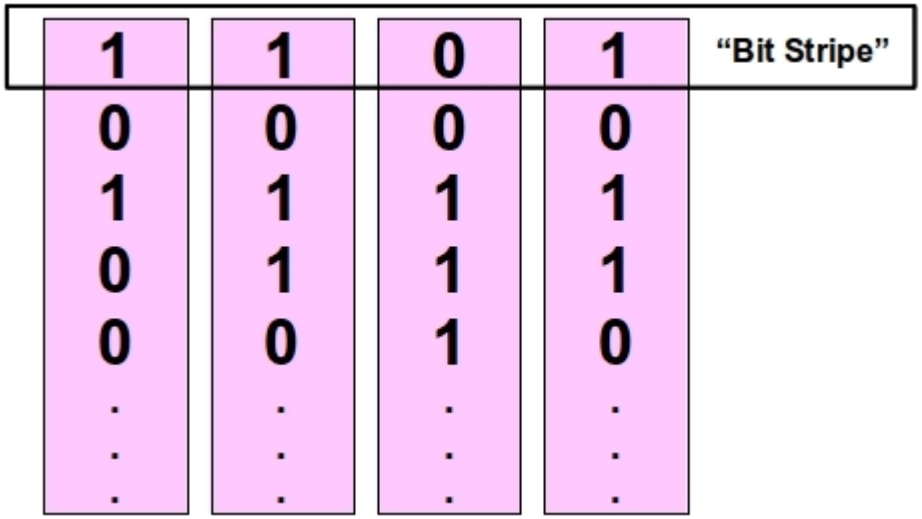
- Spinning storage performs best when
  - (a) seeks are not needed, for example R/W of large-stripe highly-sequential data
  - (b) seeks are massively parallel, such as from a raid array.

Figure 1. The Data VIMMs of a 5-disk RAID 3 Array



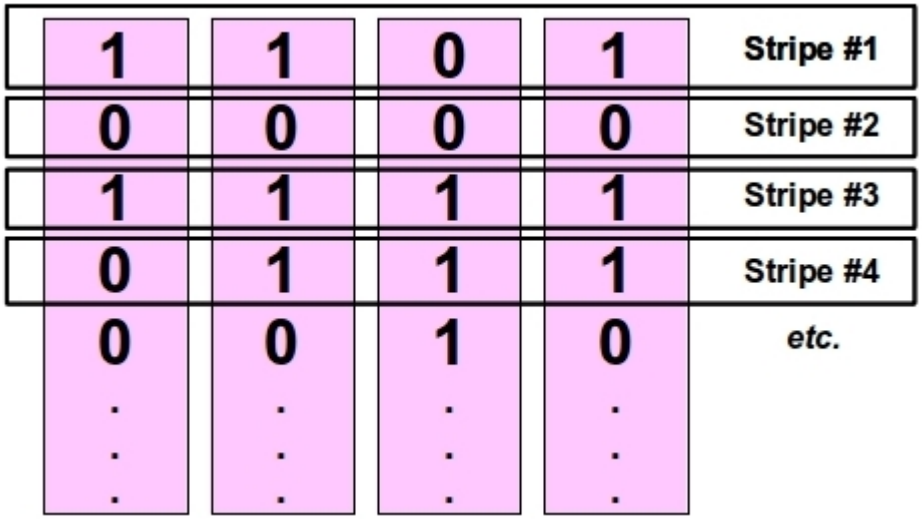
The data on a disk VIMM is sequences of bits, 1's and 0's, i.e. binary data. Conceptually, we can represent the four VIMMs of the array as four silos of bits.

Figure 2. A Bit Stripe Across the RAID3 Data VIMMs



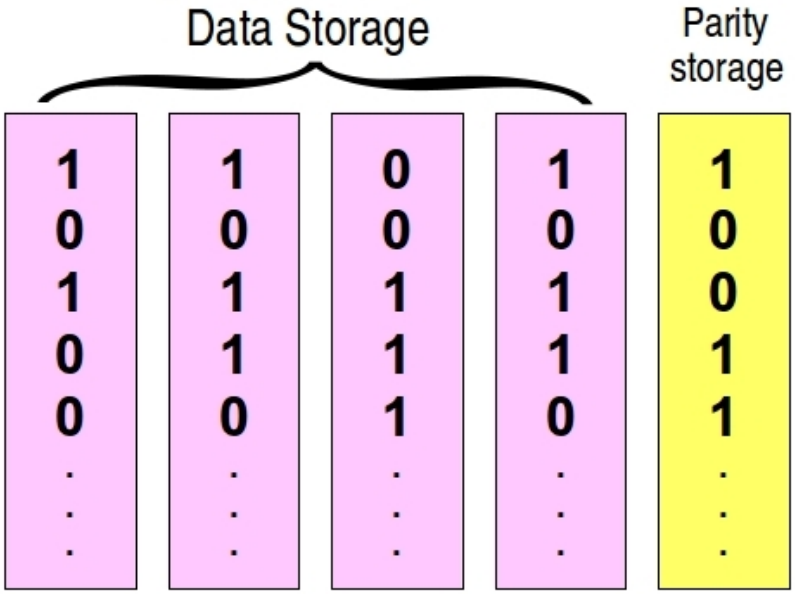
The array ingests a 4K block and splits it into 4 x 1K blocks sharing the same logical page address, and distributes those blocks across the 4 array storage devices, Which in flash storage are called VIMMs and are bus-mounted cards which have flash on them. Conceptually, we can create a "bit stripe" across the four VIMMs holding the 1K blocks of that 4K block as shown above.

Figure 2. A Bit Stripe Across the RAID3 Data VIMMs



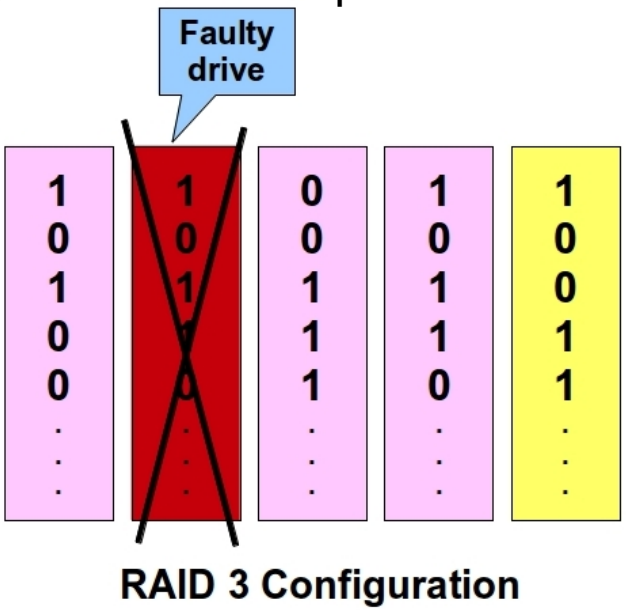
Stripes exist for all the 4K blocks that have been in 1K blocks distributed across the VIMMs, as shown conceptually above.

Figure 2. A Bit Stripe Across the RAID3 Data VIMMs



Stripe #1 from the top is odd (1+1+0+1) so the parity is 1.
Stripe #2 from the top is even (0+0+0+0) so the parity is 0.
The parity disk stores these computed parity values for each stripe.

Figure 2. A Bit Stripe Across RAID3 Data VIMMs

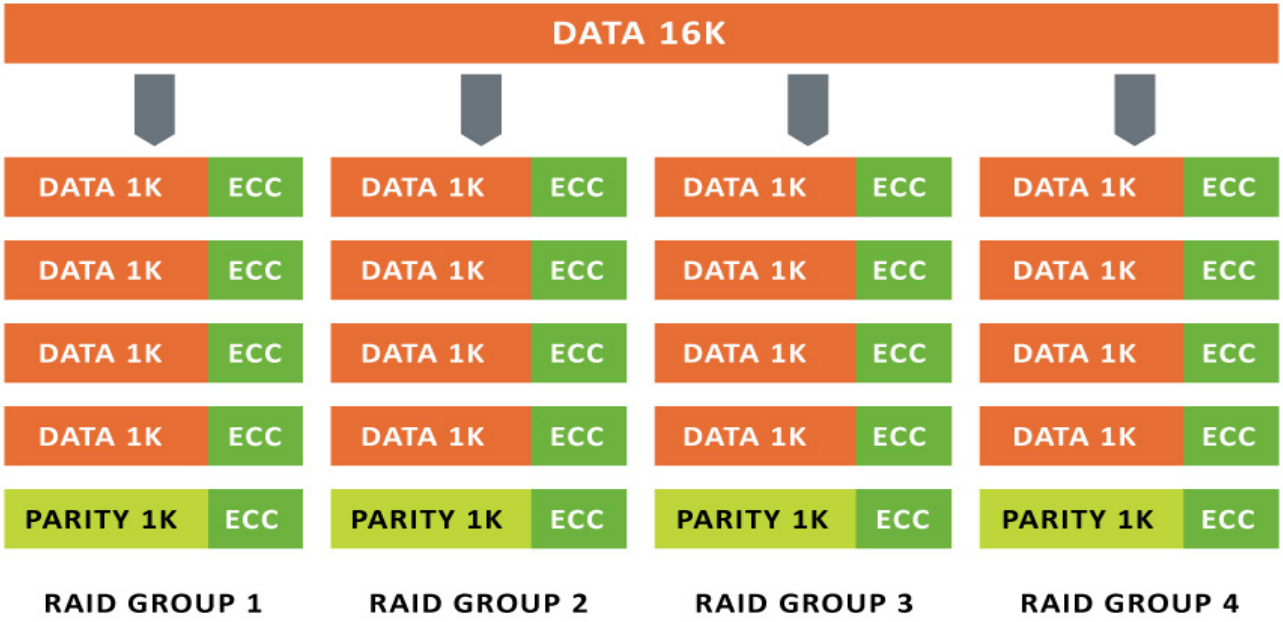


Without disk #2, we now have 1, 0 and 1 on the remaining **data** VIMMs for Stripe #1 totaling decimal 2, an even number. The parity for that stripe was 1, which means the missing bit on the faulty VIMM had to have been a 1 to produce an odd sum, so once the faulty storage device is replaced, parity can perform this VIMM reconstruction for all the data stripes on the replaced VIMM component.

Loss of the Parity Disk & The Hot Spare & Other Considerations

- vRAID maintains hot spare VIMMs (4 of them) to replace any failing disk in the RAID3 arrays using auto-detection and automatic replacement.
- The hot spare VIMM is used to replace a failing VIMM, as well as a failing parity VIMM automatically when needed, which is handled by auto-detection in the array.
- RAID3 is bad for ***random access*** on legacy spinning disks because all the disks in a stripe are accessed for every operation.
- RAID3 is excellent for ***random writes*** to flash because each 4K write writes (4Kdata +1K parity) 5K to flash not the (4Kdata+4Kparity=) 8K of RAID5 or (4Kdata+4Kparity+4Kq) 12K of RAID6
- RAID3 is excellent for vRAID erase hiding because the read from parity to reconstruct the erasing VIMM has 0 overhead, i.e. 4K is read from flash to return 4K user data, in a 4+1 RAID5 system 16K would have to be read from flash to return 4K user data, in a 23+2 RAID6 system (popular with some competitors to vRAID) it would take 96K of reads from flash to return 4K user data.

vRAID Distributes 1K Blocks for a16K Incoming Server Block Across the



Linux Flavors – Which to Choose?

- Strongly recommended to build Oracle DB on flash on Oracle Enterprise Linux or the RHEL OEL compatible kernel. Why?
- Entire software stack from OS to database layer 100% Oracle fully-vendor-controlled and fully-vendor-supported, sole-source one-stop-shop for all support needs, which is important when implementing a database with 4K sector size storage such as NAND flash or Advanced Format drives.
- Oracle Enterprise Linux 6 is the only Linux 6 which has a fully-vendor-supported ASMLib which is the recommended I/O pathway for Oracle on 4K sector storage **at this time**.
- The ASMLib solution for RHEL6 is not formally supported by Oracle, nor is it supported formally by Red Hat. Using RHEL6 for an Oracle all-flash implementation is often done with excellent results but introduces a stack which is partially supported by RedHat, partially by Oracle, and some gray areas not clearly supported at all (e.g. kmod-oracleasm)

Oracle Database Block Sizes

- Flash is optimized for 4K (4096 byte) blocks
- Any db_block_size that is a multiple of 4K is suitable for the deployment of Oracle on 4K sector size flash storage
- Why did we use larger blocks in Oracle DB (for example in Data Warehouse/Reporting solutions). Partly due to POOR I/O latency we needed to move more data per read event.
- With Flash solutions, this problem is removed, and I/O no longer is a bottleneck, so even Data Warehouse reporting databases can be potentially be operated with smaller Oracle block sizes provided that chaining and row-migration will not be a factor in downsizing the block size.

How About Jeopardy ? …

Dinos and Divas	Hardware Software	Oracle Flash Stack	Oracle Storage	Odds and Ends
10	10	10	10	10
20	20	20	20	20
30	30	30	30	30
40	40	40	40	40
50	50	50	50	50

Dinos and Divas – 10 Points

QUESTION:

- This was the first method for storing Oracle database data prior to the release of Oracle ASM, and is a less-advantageous, not-recommended choice of data storage method for deploying Oracle databases on flash storage, and is a Dinosaur.

ANSWER:

- What are Oracle database files stored on a non-ASM journaled file system.

Dinos and Divas – 20 Points

QUESTION:

- These versions of Linux 6 vended by Oracle should be used when deploying Oracle on all-flash because they are the only versions of Linux 6 which provide a fully-vendor-supported and updated distribution of ASMLib, which is strongly recommended when deploying Oracle on 4K format storage devices, such as NAND flash or Advanced Format devices.

ANSWER:

- What are Oracle Enteprise Linux 6 and the Red Hat Compatible Kernel for Oracle Linux 6

Dinos and Divas – 30 Points

QUESTION:

- This legacy enterprise database storage technology, when compared to flash storage technology, provides very poor random-read latency, provides relatively good sequential, large-stripe R/W latency, uses far more electrical energy, typically needs a flash cache helper product to meet modern low-latency performance requirements.

ANSWER:

- What is spinning platter disk. Real Dinosaurs.

Dinos and Divas – 40 Points

QUESTION:

- This Oracle-provided backup utility bundled with all Oracle database offerings, should be implemented when deploying Oracle, and its' block checking feature should be implemented, and this utility should pull backups of archived redo logs to secondary storage at least every 15 minutes, with the database running in ARCHIVELOG.

ANSWER:

- What is RMAN. Divo.

Dinos and Divas – 50 Points

QUESTION:

•This type of enterprise flash storage uses legacy external raid aggregators, legacy raid strategies, legacy disk protocols, and legacy HDD form factor enclosures which have been modified to hold flash instead of spinning disk platters and actuator arms, and which typically crashes into the “write-cliff” upon reaching capacity utilization, thus degrading IOPS performance relative to the high IOPS which might have been seen in a preliminary POC benchmark test.

ANSWER:

•What are legacy form factor SSD drive arrays.

Hardware Software – 10 Points

QUESTION:

- This patented, integrated flash raid3 solution handles “garbage collection” and write amplification, reading and writing in parallel, for industry-leading benchmark sustained IOPS with negligible write-cliff effects, and is industry-leading in heavy write environments where reading and writing to flash must occur in parallel with a constant IOPS delivery curve.

ANSWER:

- What is vRAID

Hardware Software – 20 Points

QUESTION:

- This property of NAND flash storage causes latency and results in more writes to the NAND storage than are actually required, and affects to varying degrees all enterprise SSD storage systems which deliver NAND flash storage via legacy external raid controllers, aggregators, and protocols, but is minimized and largely eliminated by vRAID technology.

ANSWER:

- What is Write Amplification.

Hardware Software – 30 Points

QUESTION:

•This linux utility when used with various switches such as [rereadpt | getpbsz | getbsz | getss”] will show you how the Linux OS **currently** reports your flash block device properties, such as, respectively, [reread partition table | report physical block size | report logical block size | report sector size] so that these properties can be verified as the **true properties** before attempting to present storage to the DB.

ANSWER:

•What is “blockdev”.

Hardware Software – 40 Points

QUESTION:

- This Linux memory management paradigm is an alternative to Oracle AMM and should be used with Linux Oracle SGA sizes greater than 10Gb to reduce unnecessary CPU cycles and minimize swapping of page table entries

ANSWER:

- What are Linux Hugepages

Hardware Software – 50 Points

QUESTION:

- When deploying Oracle RAC on flash, the interconnect ideally should at a minimum have this speed rating (although multiple-bonded 1Gb/s Ethernet may be feasible).

ANSWER:

- What is 10Gb/s Ethernet.

Oracle Flash Stack – 10 Points

QUESTION:

•This linux OS flavor includes a fully-supported IO layer and device persistence kernel module, providing end-to-end single-vendor full-stack-support for 4K flash sector size storage, and provides full KVM virtualization with virtio-blk and is RECOMMENDED for deploying Oracle on bare metal or virtualized KVMs.

ANSWER:

•What is Oracle Enterprise Linux 6 and the OEL Red Hat Compatible Linux (both with vendor-supported ASMLib)

Oracle Flash Stack – 20 Points

QUESTION:

•This version of Oracle Enterprise Edition database software was the first version of Oracle to offer major support features for storage devices and flash arrays and is the first version which supports a “true” 4K sector size together with several 4K bug fixes Oracle database implementation on 4K flash and is STRONGLY RECOMMENDED as the minimum version for Oracle deployment on flash.

ANSWER:

•What is Oracle Enteprise Edition 11gR2

Oracle Flash Stack – 30 Points

QUESTION:

- This proprietary Oracle logical volume manager (LVM) is recommended for implementations of Oracle on flash storage devices for performance, manageability, and supportability and really should be the **first and only choice** when compared to say, 3rd-party LVMs, file-system-based Oracle, etc.

ANSWER:

- What is the Oracle Automatic Storage Manager (ASM) 11gR2 and higher.

Oracle Flash Stack – 40 Points

QUESTION:

•This kernel module is provided by Oracle for Oracle Enterprise Linux 6 (OEL6) only and was introduced to address both device persistence and IO_DIRECT deficiencies in the Linux kernel, as well as providing its' own proprietary IO layer which delivers fully-vendor-supported 4K IO efficiently and safely, which 4K Linux native IO cannot claim at this time make due to an unpublished Oracle bug.

ANSWER:

•What is ASMLib for Oracle Enterprise Linux 6.

Oracle Flash Stack – 50 Points

QUESTION:

- This value for the db_block_size(n) parameters in the 11gR2 database is the minimum size that should be used when building an Oracle database on flash storage devices. All db_block_sizes in the 11gR2 database on flash storage should be multiples of this basic block size.

ANSWER:

- What is 4K (4096 bytes) db_block_size.

Oracle Storage – 10 Points

QUESTION:

- This ASMLib parameter should be set to “true” (case matters – must be lower case!) when presenting 512-byte emulation mode LUNs from an 4K physical sector size flash storage device. This ASMLib parameter is best set using the command “oracleasm configure -b” (yes, “b” is the switch for “logical” i.e. 512-byte size) to avoid the case issue.

ANSWER:

- What is ORACLEASM_USE_LOGICAL_BLOCK_SIZE

Oracle Storage – 20 Points

QUESTION:

- The redo log BLOCKSIZE should always be set to 4K when running oracle on 4K flash storage for performance reasons. In certain configurations of OS and flash storage, the redo BLOCKSIZE parameter cannot be set to 4K without first setting this oracle "_" instance initialization parameter.

ANSWER:

- What is “_DISK_SECTOR_SIZE_OVERRIDE”

Oracle Storage – 30 Points

QUESTION:

- This value of the ASM initialization parameter "asm_diskstring" MUST be used when using 4K flash in order to ensure that the IO facilities of ASMLib are being used. Note, using "/dev/oracleasm/disks" does NOT provide the same IO facilities as this parameter.

ANSWER:

- What is "ORCL:*".

Oracle Storage – 40 Points

QUESTION:

- This logical LUN blocksize also known as an emulation layer size for Oracle on flash storage devices should be used for ALL Linux OS versions [except for OEL 5(2.6.32) and OEL 6 which CAN use the 4k (4096 byte) size].

ANSWER:

- What are 512-byte logical “emulated” LUN blocks.

Oracle Storage – 50 Points

QUESTION:

- This blocksize specification for an ASM diskgroup, as described in Oracle bug # 14626924, must be used for any ASM diskgroup that will be used to store the Oracle spfile.

ANSWER:

- What is 512-byte block size.

Odds and Ends – 10 Points

QUESTION:

- When creating online redo logs on flash storage for an Oracle 11gR2 database this blocksize must be used to avoid degrading maximum possible performance.

ANSWER:

- What is 4K (4096-byte) blocksize.

Odds and Ends – 20 Points

QUESTION:

- This category of Oracle database professionals employed by your prospective all-flash storage vendor should be involved from the vendor from the very beginning in all preliminary and final planning stages for your Oracle on flash benchmarking and POC.

ANSWER:

- What is a vendor Oracle DBA team

Odds and Ends – 30 Points

QUESTION:

- Any of these storage protocols can be used in a POC with a storage array running vRAID, unlike many legacy SSD solutions, which often do not support all these storage protocols.

ANSWER:

- What are Fiber Channel, Infiniband, iSCSI

Odds and Ends – 40 Points

QUESTION:

- When comparing flash storage solutions from different vendors in a “Proof of Concept” implementation, where arrays are installed in your datacenter, this test factor MUST be implemented in order to measure ACTUAL performance of the array including the action of the garbage-collection.

ANSWER:

- What is “pre-loading the array with random addresses (and not sequential copy!)” so that garbage collection is activated and factored into performance.

Odds and Ends – 50 Points

QUESTION:

- These important command line parameters for CREATE DISKGROUP and CREATE LOGFILE GROUP, respectively, set compatibilities with 4K sector and Advanced Format devices.

ANSWER:

- What are SECTOR_SIZE (asm) and BLOCKSIZE (redo logs).

Wrap-Up

- Thank you for attending this presentation!
- A word about AWR's and evaluating them…
- Questions?
- You can contact me through any of the following methods:

  - 914-261-4594
  - gstanden@vmem.com
  - Blackberry Messenger (BBM) Pin 334C46F7 (now on IOS and Android!)
  - LinkedIn
  - gil_standen (yahoo IM)
  - Standen98 (AIM)