# Flash Performance for Oracle RAC with PCIe Shared Storage

## A Revolutionary Oracle RAC Architecture

Authored by: Estuate & Virident | HGST

## Table of Contents

# A Revolutionary Oracle RAC Architecture

## Introduction

Oracle Real Application Clusters (RAC) allows Oracle Database to run any packaged or custom application unchanged across a server pool. This provides the highest levels of availability and the most flexible scalability. To keep costs low, even the highest-end systems can be built out of standardized, commodity parts.

Oracle Real Application Clusters provides a foundation for Oracle's Private Cloud Architecture. Oracle RAC technology enables a low-cost hardware platform to deliver the highest quality of service that rivals and exceeds the levels of availability and scalability achieved by more expensive mainframe SMP computers. By dramatically reducing administration costs and providing new levels of administration flexibility, Oracle RAC is the defacto architecture for enterprise database applications.

Until now 'PCIe shared storage' has been an oxymoron. But with Virident | HGST FlashMAX PCIe Solid State Disks (SSDs) coupled with FlashMAX ClusterPak software, Oracle RAC systems can gain the performance benefits of PCIe while enjoying the reliability and scalability of a SAN. This revolutionary architecture features microsecond latency Infiniband networks and Remote Direct Memory Access (RDMA) between nodes to deliver low latency and high IOPs, while preserving the integrity of the Oracle RAC platform.

The Virident | HGST FlashMAX ClusterPak storage management software suite features vHA for high availability replication between PCIe SSDs, vShare for storage pooling and ClusterCache for server-side caching in front of existing SAN or DAS installations.

This paper talks about this unique offering that makes it possible for Oracle RAC database to run on PCIe SSDs. The RAC cluster need not have any other shared storage, yet it still achieves the best performance, availability, scalability and cost savings for building out blocks of databases for the public or private cloud.

The Oracle RAC database uses the extended cluster configuration and ASM preferred reads to achieve High Availability (HA) and get the best performance out of the suggested configuration.

## RAC "Share Everything" Architecture

The Oracle RAC database has the "share everything" architecture. All data files, control files, SPFILEs, and redo log files in Oracle RAC environments must reside on cluster-aware shared disks so that all of the cluster database instances can access the shared storage. All database instances must see the same view of the cluster storage and the data files. The redo log files on the shared storage are used for instance recovery.

PCIe SSDS are directly attached to the server. They have some very desirable features that benefit databases that require high performance and cost effective solutions. Until now, it was not possible to use server attached PCIe SSDs as RAC storage. In fact, it was not common even for single server databases, due to some of the limitations of existing PCIe SSDs. Before we discuss the revolutionary solution from Virident, we will quickly list out advantages and disadvantages of server attached PCIe SSDs to better understand why the proposed solution is revolutionary.

PCI Express (PCIe or PCI-E) is a high-speed expansion card format that connects a computer with its attached peripherals. PCIe SSDs are the latest technology to be adopted by datacenters to store and transfer data in a very efficient way. Since these SSDs are attached to a server directly, they're known as "server attached PCIe SSDs."

The advantages of PCIe SSDs over array storage are:

- Performance: The biggest benefit of PCIe SSDs is increased performance. Not only does the PCIe interface have low latency for data transfer, it also bypasses any storage area networking to store or retrieve data. It is hence the fastest way of accessing data. It delivers microsecond latencies versus millisecond latencies for traditional SAN based storage.

- Energy savings: Server-attached PCIe SSDs eliminate the need for additional storage servers, hence saving on power and cooling. Traditional storage solutions for high throughput, low latency and high IOPSs need hundreds of hard disk drives (HDDs), Fibre Channel controllers and significant amounts of power and cooling.

- Space Savings: PCIe SSDs are compact and fit into the PCIe slot of a server. They eliminate the need for rack space, cooling, and power for storage servers.

The disadvantage of traditional PCIe SSDs are:

- Capacity: Traditional PCIe enterprise class PCIe SSDs are often on the order of 100 GBs rather than terabytes in size. System architects can quickly run out of available PCIe slots in a server when sizing for capacity alone.

- High Availability: If a server goes down, the storage is down as well. Hence there is an absolute need for mirroring and other high availability software to ensure service continuity.

- No sharing: Traditional PCIe SSDs do not have a "Share Everything" capability, so they simply cannot be used as storage for Oracle RAC databases.

The unique combination of FlashMAX hardware and FlashMAX Connect software from Virident eliminates the disadvantages of PCIe cards and provides a powerful solution for database architects.

There are two features that make Virident PCIe SSDs very compelling for use as storage for Oracle databases. They are:

- **Capacity:** Currently FlashMAX PCIe SSDs range from 550GB up to 4.8TB, and can be used as building blocks to reach any required capacity. Additional SSDs can be added to a single server, or across servers to increase both storage capacity and available CPU. Being compact with high capacity and high performance, there will be significant cost savings if you can eliminate expensive storage server altogether for your production systems.

- **High Availability:** FlashMAX ClusterPak vShare provides PCIe SSDs sharing capability through an Infiniband (IB) network. Remote servers can access every server's FlashMAX PCIe SSDs as shared storage using FlashMAX ClusterPak vShare. This meets the exact requirements of the "Share Everything" architecture for Oracle RAC database.

The Virident  FlashMAX ClusterPak vShare solution makes it possible to fully use all the advantages of PCIe SSDs with no disadvantages. The combination of vShare and Oracle RAC provides a solution that takes advantage of the best of both worlds.

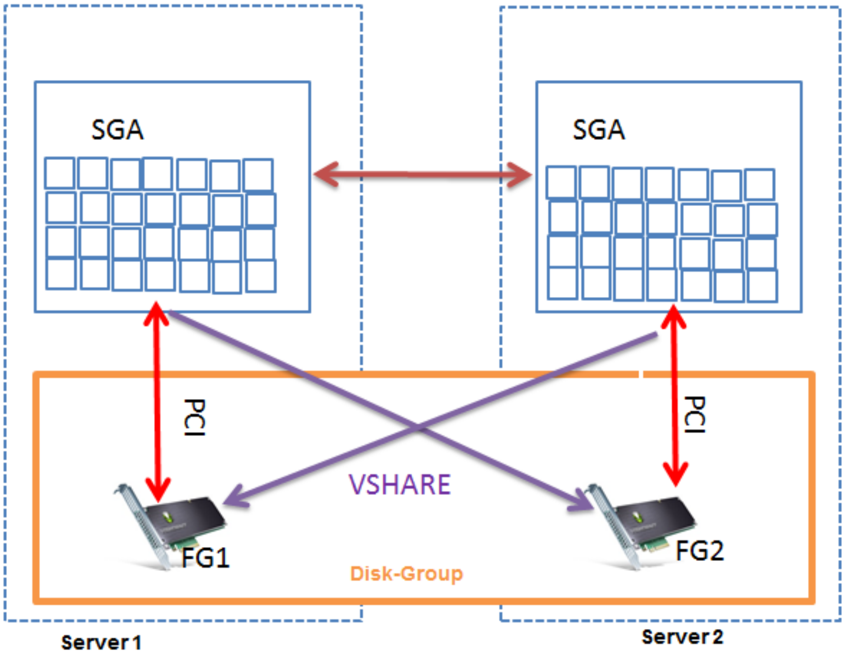
This unique RAC architecture has several advantages that make it a compelling solution for building out a cloud platform on which to deploy Oracle RAC.

## Oracle RAC on FlashMAX PCIe SSDs

The FlashMAX ClusterPak software suite redefines the enterprise storage architecture. It brings server-side flash into enterprise applications that previously relied on high-cost proprietary SAN (Storage Area Networking) hardware. FlashMAX ClusterPak software builds on the FlashMAX hardware platform and adds flash-aware network-level storage capabilities. Now you can use FlashMAX ClusterPak to create Virtual Flash Storage Networks on the server side. It allows you to deploy applications on flash storage with shared block-level access and high availability across servers, while keeping the low-latency advantages of the PCIe-attached flash.

The PCIe cards are directly attached to the database server. This configuration needs no network access to move data from SSD to server's memory and CPU to process the data. Virident | HGST's cards provide over a million IOPS with a disk-like capacity of up to 4.8 TB per SSD. What this means is, you have servers with effectively limitless I/O throughput. Applications which use Virident | HGST PCIe SSDs will often have the bottleneck moved from storage to CPU and memory. Note that most of the performance tuning historically involved reducing I/O contention and the bottleneck was mostly I/O.

Below is the image of 2 node RAC cluster with Virident FlashMAX Storage only



## RAC cluster with Virident | HGST FlashMAX

ASM built in mirroring is used to efficiently mirror database files across servers (failure groups). FlashMAX II cards on each server are used as separate failure groups.

- Servers 1 and 2 have FlashMAX PCIe cards which are configured with ASM normal redundancy.
- ASM handles the consistency of data across the two PCIe SSDs by means of ASM mirroring.
- A Diskgroup consists of logical shareable disks across the server. In our case, each FlashMAX SSD is exposed as a disk.
- Each Diskgroup has two failure groups, FG1 and FG2, with one on each node. So all data in server1:FG1 is also available in server2:FG2. Oracle ASM ensures that at least two copies of data are present, one on each failure group. This way, should a server go down (i.e. a failure group), there is no impact on data availability.
- Each failure group can contain one or more FlashMAX PCIe SSDs. The data will be evenly distributed across all the SSDs in a single failgroup.
- Both servers can access the failure groups for Reads and Writes.
- *ASM Preferred Read is set up for node affinity*. So all reads from server1 will only access data from FG1 and reads from server2 will access data from FG2 through the direct high speed PCIe bus which is attached to the server. Each failure group has mirror copy of data.

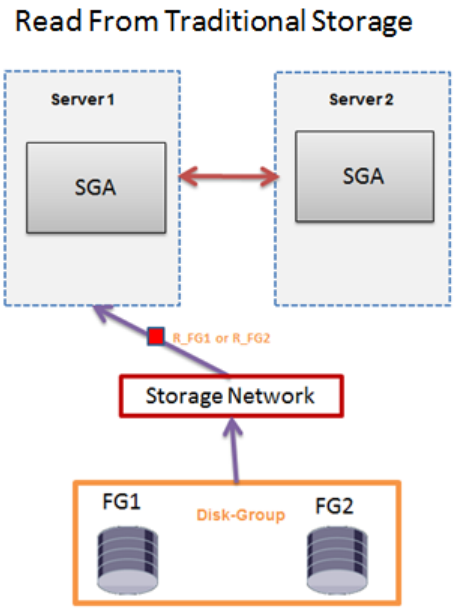
- Writes to server1 is written directly to FG1 through the PCIe bus and to FG2 through a high speed Infiniband interconnect.

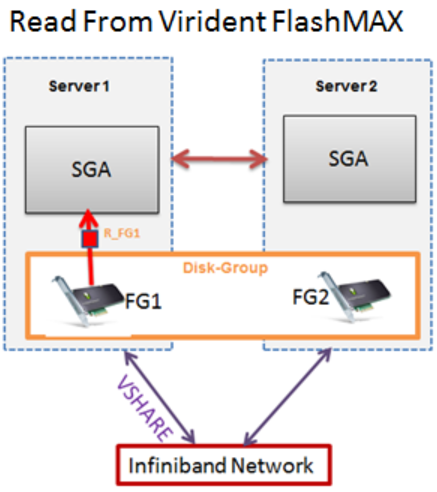
## Network

There are 2 networks between the servers. They are connected through Infiniband connectivity and run standard TCP over IB protocol and IB RDMA. The Infiniband network provides microsecond latencies and gigabytes of bandwidth.

- **RAC interconnect:** Oracle Clusterware requires that you connect the nodes in the cluster to a private network by way of a private interconnect. The private interconnect is a separate network that you configure between cluster nodes. This interconnect should be a private interconnect, meaning it is not accessible to nodes that are not members of the cluster. They are used as part of the RAC cluster to share and transfer data from memory to memory of the servers. In this architecture, for the highest performance, this interconnect is Infiniband.

- **vShare:** This software works similarly to RAC interconnect. It too will run on a separate private network. In traditional cluster storage, all the servers are connected to a storage network and one big storage rack connected to the hub. vShare on the other hand is a software that allows PCIe SSDs of all servers to be viewed as shared disks on every server. All servers are connected together through high speed Infiniband network. Since PCIe SSD storage is attached to the server, there is no requirement to connect to any other storage. In this architecture, the shared Infiniband fabric utilized by vShare has much lower latency and higher bandwidth than traditional storage network which usually run on Fibre Channel. Below figure shows the network connection to storage on traditional SAN storage and FlashMAX.





## ASM and Oracle files

Since Oracle release 11gR2, all Oracle related files can reside in the ASM. Prior to 11gR2 release, Oracle voting disks and OCR files which are part of the cluster earlier had to be kept outside ASM to help start and stop the RAC cluster. However, Oracle introduced a new process that makes it possible to have all files within ASM. You can still have the database configured with voting disks and OCR files outside ASM, but having all Oracle managed files within ASM is discussed here. Note that a two node RAC cluster will need a quorum voting disk. See Appendix for more information about this ASM functionality.

viriden
Forma

viriden
Delete

## The Architecture Advantage

Oracle RAC on vShare network has some fundamental differences in how the data is stored and retrieved. This design has advantages over traditional RAC storage. They are

- Unlimited IOPS with linear scalability
- Less I/O traffic through network.

### *Unlimited IOPS with linear scalability*

With this architecture, it is possible to have unlimited IOPS and linear scalability of IOPS due to low latency I/O of FlashMAX SSDs and vShare network.

In a simulated test, a single server was able to achieve 80,000 8K IOPS before it ran out of CPU resources. Adding another node to the test doubled it to 160,000 with both nodes pegged due to CPU. Now if additional RAC nodes were added to this setup, you are adding more CPU resources in addition to storage. These CPU resources could be used to access SSDs through vShare network or through the PCIe bus depending on where the data resides. This ability to read/write more blocks through additional nodes would mean more IOPS.

It is possible to simulate a scenario where all the IOPS across all the nodes access a single FlashMAX SSD on a server and max out the IOPS capacity on a single SSD. Practically this is unlikely to happen as data is distributed across all the RAC nodes and further distributed across diskgroups and SSDs within a single failure group. Note that when you double the FlashMAX SSDs within a single failure group, the maximum IOPS limit would double.

The point here is with vShare network, the total IOPS capacity of the RAC server is not limited by the IOPS limits of a single SSD. This is very different from closed SAN storage architectures where the total IOPS of a storage system is described in the specification of the H/W and systems with higher IOPS will be more expensive.

With Virident | HGST vShare network, you get low latency IO with IOPS increasing as more servers are added. This is a revolutionary change and it liberates the DBA and architects from working under the confines of storage performance and capacity limitations.

## The Write Advantage (50% less I/O traffic through the network)

In the above configuration, all the DB files are stored in an ASM diskgroup. The files are striped across the disks in each fail group and mirrored across failure groups. Every file on server 1 is present on server 2. When ASM writes to server1:FG1, it also writes to server2:FG2. This write data is sent to server2:FG2 through the vShare private network. The interconnect setup in our case is over Infiniband network which has very low latency and high bandwidth.

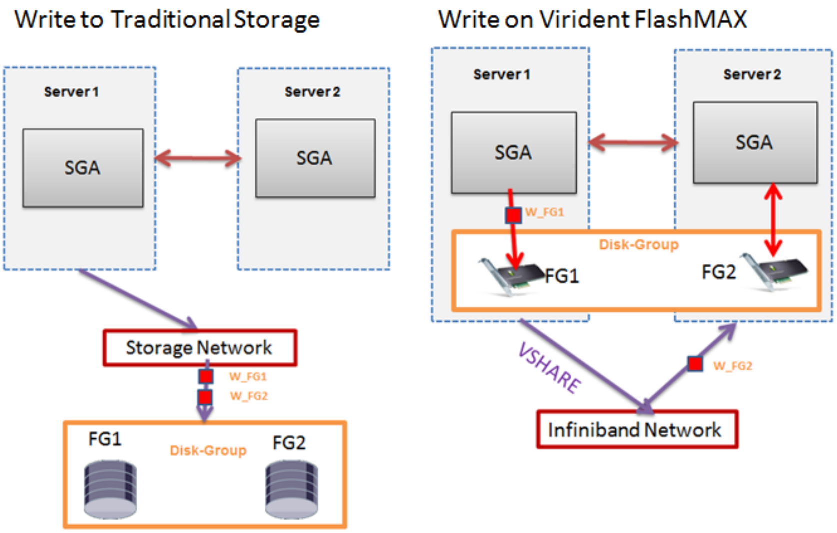
In a typical RAC configuration with shared storage, all servers are connected to the shared storage (SAN) through network switches. This would mean all the writes will have to pass through the network to be written to disks. In case of the above configuration, one of the writes is done directly on local PCIe storage and the other write passes over the network.

The advantage here is the reduction in write I/O traffic from server to the diskgroup by 50% than a typical SAN storage.

The writes are synchronous and an acknowledgment is returned to database instance after ensuring that all I/O are done. So the performance of a write will be equal to the time of the slowest disk. In our case, you will have similar I/O performance as any other shared storage connected through high speed network. Due to low write latencies of the FlashMAX and the low network latency of the Infiniband network compared to traditional SAN storage, the performance is likely to be much better. In addition to this, since the amount of data you pass through the network is half of typical shared storage, you can push a lot more writes through the network before you get any network bottleneck.

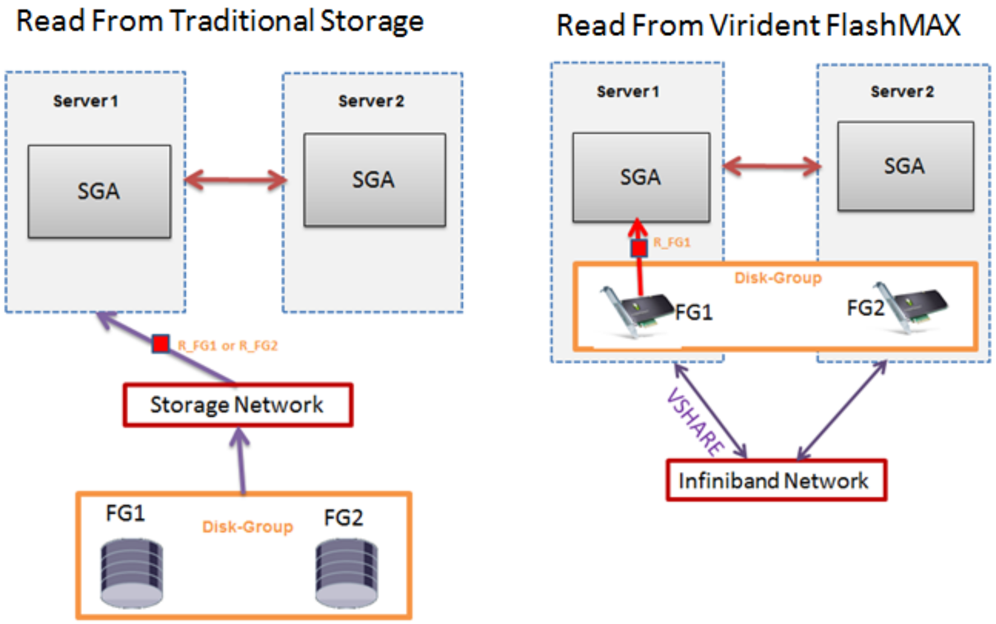
Infiniband networking has become an industry standard and is also used as a key component in high performance Oracle architectures to reduce latency and increase bandwidth for RAC interconnect. By utilizing it also for ClusterPak vShare, even higher performance is possible.

The figure below illustrates the path of a single Write I/O from server 1 in traditional RAC storage and FlashMAX.



## The Read Advantage (100% reads from local PCIe, zero traffic through vShare network)

When you configure ASM failure groups, it is more efficient for server1 to read from FG1 that is directly attached to the server than read data from the remote server through the network. Using preferred read failure groups configuration of ASM, all physical reads from the database are read from local PCIe cards. What this means is that you have 100% reduction of network traffic for data reads. This means that there is zero network latency to read data and access time for PCIe cards are in microseconds and not milliseconds which is typical of SAN storage. Below figure illustrates single block read from server1.



The advantages of this solution architecture are:

- High Availability
- Performance
- Scalability
- Cost Savings

## High Availability

One of the main reasons why traditional PCIe cards cannot be used in Oracle RAC configuration is that they do not support "Share Everything" architectures. Now with FlashMAX ClusterPak vShare, the biggest technical hurdle is eliminated. All data in server 1 is available in server 2. If one of the servers goes down, the application can still access the database and cluster through the surviving nodes. In our configuration, failure of a server will also mean failure of the attached storage and the ASM fail-group. This is why you should not have more than one failure group of a diskgroup per server.

The OCR is striped and mirrored similar to ASM Database Files. So we can now leverage the mirroring capabilities of ASM to mirror cluster files.

The Voting Disk is not striped but put as a whole on ASM Disks – if we use a redundancy of normal on the Diskgroup, 3 Voting Files are required, each on one ASM Disk.

In addition to this, there is a RAC cluster HA requirement that all servers need access to half or more of the voting disks in order to be part of the cluster. If the server loses access to 1/2 or more of all of your voting disks, then nodes get evicted from the cluster, or nodes kick themselves out of the cluster. For normal redundancy you can have only 3 voting disks in a single diskgroup. Therefore, you need to have at least 3 failure groups to store the voting disks: In a 2 node RAC cluster as above, we can only have 2 failure groups (voting disks cannot span multiple disk groups). In such cases the new quorum failure group concept is introduced. A quorum failure group is a special type of failure group and stores only the voting disk. The quorum failure group does not contain user data and are not considered when determining redundancy requirements. A quorum failure group can be created from standard NFS based or iSCSI disk that the RAC cluster can access. The space you need for voting disk is typically less than 300MB. The installation document shows the details of how this is configured.

## Performance

There are two major types of workload a database experiences:

- **On-Line Transaction Processing (OLTP):** The performance of OLTP systems is usually less physical reads and writes, but more focused on latency and IOPS.
- **Data Warehouse (DW):** The performance characteristic of DW application is measured in terms of throughput. This workload will have high physical reads to run queries on large amounts of data and batch writes to update the warehouse system.

### *OLTP Performance*

- **Physical Reads:** In an OLTP system, the amount of physical reads is dependent on the size of SGA. The more DRAM available, the less likely many physical reads/sec will be needed. Even the busiest OTLP (like TPC-C) benchmarks often do not need many physical reads. To give you an idea, the busiest OTLP systems with thousands of transactions a second might need less than 25,000 8k block size I/O per second. The FlashMAX SSDs tested are able to run in a simulated environment with more than 80,000 physical reads per second before CPU resources were exhausted. The physical reads were doubled to over 160,000 8k block reads when the test was run against both nodes. You can conclude that reads scale very well as you add nodes in the setup. By setting up ASM preferred reads, the data is directly read through the PCIe bus when the data required is in the local FlashMAX SSDs of the diskgroup. In such cases no reads pass through the vShare network. Even if one node requires a block from the other node's SGA, the data is transferred through RAC interconnect and is considered a logical read. If the data required is not available in the local FlashMAX SSDs, data can be read through the vShare network. The low latency of FlashMAX and the Infiniband network is in microseconds unlike traditional SAN storage, which has latency in milliseconds. As described earlier, when more FlashMAX SSDs or servers are added to the cluster, the low latency IOPS increases with no upper limit.

- **Physical Writes:** The physical writes are written to the local SSDs and the remote mirror through IB network. Writing over IB will take only a few microseconds longer than a local write. This is in contrast to the many milliseconds common in a storage network like Fibre Channel. The reason SSDs are popular is because of zero seek time, low latency. Chances are you will run out of CPU resources on the server before you can get any bottleneck on the system. Note that keeping OLTP database files and online redo log files on the same disk will not show any I/O spikes or difference in I/O latency since the bottleneck is not on the SSDs.

- **Network Bottleneck:** One might experience a network bottleneck on the IB cable connecting to the storage array when several RAC servers are performing I/O. However, as mentioned above, since the amount of data that needs to pass through the network is half for writes and zero for reads, you can put more than double the load before you reach any bottleneck compared to a typical SAN storage. More so, the IB interconnect used provides over 5GB/s of per-link bandwidth so only in the absolute largest of systems will this become a true bottleneck.

- **CPU Bottleneck:** The CPU is either busy doing work or waiting for I/O. Since the latency for I/O is in microseconds, the CPU is going to be a lot busier doing work with FlashMAX SSDs than when connected to SAN. The FlashMAX SSDs give effectively unlimited IOPS and microsecond latency, so the system is more likely going to have a CPU bottleneck before hitting any I/O bottleneck. In this case, you will be able to put a lot more workload on the system and get better TCO for the investment.

### *Data Warehouse Performance*

The data warehouse performance characteristics are high physical reads and high batch writes.

To speed up reads, it is typical for data warehouse queries to run using parallel queries. The query slave processes run in parallel and work on a subset of the data. The performance of the end query is highly dependent on the latency of I/O and degree of parallelism.

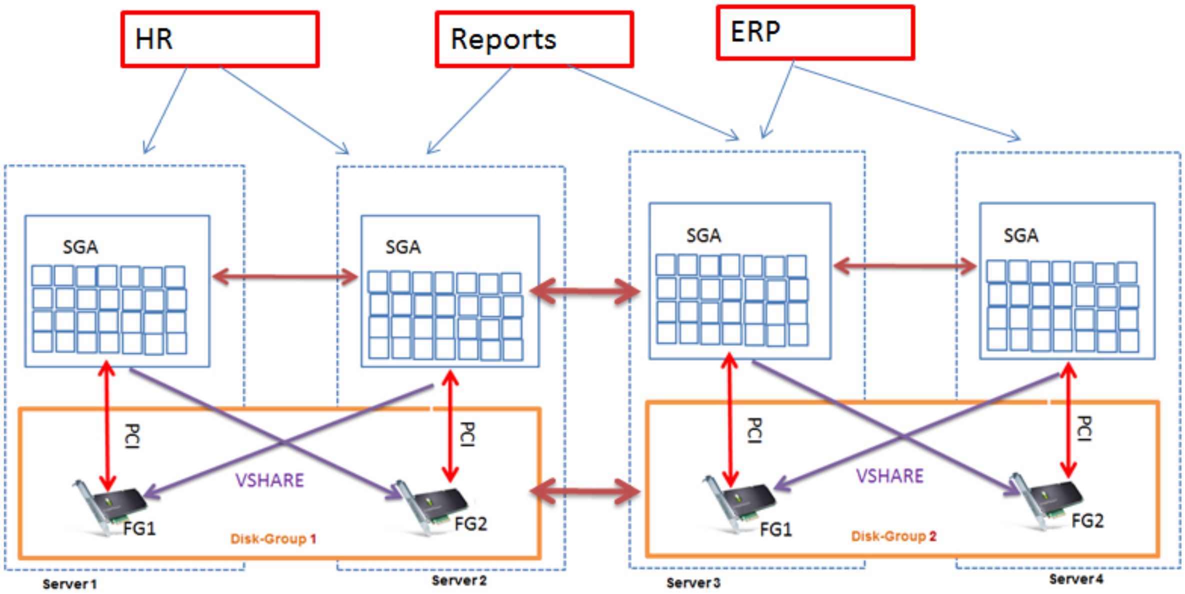
As all reads are 100% from the local SSDs, the latency of IO is very low. In addition to this, you will not observe any spike in latency when you run multiple concurrent parallel queries at the same time due to the ability of PCIe SSDs to provide low latency IOPS.

Oracle RAC parallel query can spawn slaves across servers and do the work on all servers and then return the processed data to the query coordinator. In our configuration, each parallel slave will work on local mirror copy and return data to the query-coordinator.

The batch write performance depends on the write latency, write IOPs and network latency. Having the database redo logs and data files on PCIe cards is going to have a significant impact on the latency due to the nature of SSDs. They have zero seek time compared to HDDs whose performance can vary depending on the load on the storage server. In a busy system the bottleneck is likely shift from IO to CPU and memory.

The performance of any database is always better when data is processed close to where it resides. In this architecture, data resides in the same physical enclosure where the processing is done. This greatly improves performance on regular RAC clusters.

## Scalability



Application scalability issues arise mainly due to storage capacity, CPU, or memory constraints. The requirement to scale can arise when you run out of any of these.

Note that with Virident FlashMAX SSDs, IO sizing is often not required.

IO sizing is usually one of the biggest issues DBA needs to confront on when scaling applications. When you add servers to accommodate more workload, you need to increase the I/O infrastructure as well. This would mean more storage servers, more Fibre Channel controllers, controller caches, etc.

### *Storage scalability*

As mentioned earlier, FlashMAX PCIe SSDs currently boast a capacity of up to 4.8TB each. This is sufficient for most of the OLTP applications. If the servers have more PCIe slots, then you can simply add additional PCIe cards as ASM disks to the same diskgroup. Each additional FlashMAX SSD will add an additional 4.8 TB to the diskgroup capacity. Note that you will need to add equal sized cards to the failure group as well.

Another option is to have a tiered storage. You can keep active data on flash disk groups and older less frequently used data in diskgroups on a SAN. Oracle 12c has new features like Automatic Data Optimization (ADO) that can automate the process.

### *CPU/Memory scalability*

If CPU or memory is the bottleneck, you will need to add more RAC nodes. RAC's core value proposition is its ability to scale out on low cost servers while maintaining excellent performance.

When you add additional nodes for consolidation of different applications, you can configure such that each application runs on its own disk group and configure the services such that the connection to the servers are only made to servers where you have a failure group you can access locally.

In case additional nodes added do not have data locally, the application still will scale but will have some additional latency to access data on remote node for physical reads. This really may not be significant for OLTP systems as the amount of physical reads is dependent on memory as well. With additional servers, you are also increasing the SGA size as well and very likely the physical reads/second requirement will be reduced.

## Cost Savings

The cost savings of this solution is a very important component that makes this architecture very compelling. The biggest cost savings arise from the fact that the CPU is busy doing work rather than waiting for I/O. Oracle RAC software is licensed per core, no matter how utilized that core maybe. By freeing the CPU from waiting for IO, more processing can be done with fewer numbers of cores. Alternatively, consolidations of multiple databases onto a smaller set of servers can significantly reduce cost.

To run an Oracle RAC database with comparable I/O performance on shared storage server, you would need multiple high end storage servers. In addition to the cost of these storage servers, you will certainly need additional system administrators to manage these storage systems. With FlashMAX SSDs attached to servers, this role of system administrator could easily be picked up by a DBA. There is only a one time setup to configure the SSDs and minimal ongoing maintenance.

The energy consumption of these systems is significantly less than RAC on traditional storage servers with similar performance characteristics.

The amount of rack space occupied by FlashMAX SSD storage systems is effectively zero as all the storage resides in the server case itself. This is a big cost savings in hosting and datacenters environments where there is additional cost for rack space consumed by the servers.

This solution can also be easily be implemented by smaller companies that need high end performance from a small energy efficient systems with no special datacenter in house.

## Conclusion

The ability to share PCIe SSDs across servers using FlashMAX ClusterPak makes it possible to combine the advantage of Oracle RAC database with the advantages of PCIe SSDs to create a very compelling architecture on which to run all databases.

## Appendix

Just like database files, Oracle Cluster Registry (OCR) files are stored in an ASM disk group and therefore utilize the ASM disk group configuration with respect to redundancy. For example, a normal redundancy ASM disk group will hold a two-way-mirrored OCR. A failure of one disk in the disk group will not prevent access to the OCR. The Oracle Cluster Registry is the central repository for all the resources registered with Oracle Clusterware. It contains the profile, state, and ownership details of the resources. This includes both Oracle resources and user-defined application resources. Oracle resources include the node apps (VIP, ONS, GSD, and Listener) and database resources, such as database instances, and database services. Oracle resources are added to the OCR by tools such as DBCA, NETCA, and srvctl.

Oracle only allows one OCR per disk group in order to protect against physical disk failures. When configuring Oracle Clusterware files on a production system, you could configure either normal or high redundancy ASM disk groups. You can, however, have up to five OCR disks in separate diskgroups.

Voting files are not stored as standard ASM files. They follow the ASM disk group configuration with respect to redundancy, but are not managed as normal ASM files in the disk group. Each voting disk is placed on a specific disk in the disk group. Voting files must reside in ASM in a single disk group. With normal redundancy, you need to have 3 voting disks and with high redundancy you will need 5 voting disks.

The files that are part of this ASM include Voting Disks, OCR files, Spfiles for ASM and the DB, Database files, Flashback files, redo log files, archive log files, rman backup files, export dump, password and wallet files.

With ASM normal redundancy, Oracle will need 3 voting disks in an ASM diskgroup. The OCR file is stored similar to any other oracle database file. It is striped and mirrored across all the disks in diskgroup. We need to point it to 3 ASM failure group of a diskgroup. In addition to this, there is a RAC cluster feature that if you lose 2/3 voting disks, then nodes get evicted from the cluster, or nodes kick themselves out of the cluster.

In our two node RAC configuration, the loss of a node will also mean you will lose the storage attached to it. So make sure that you have only one Failure group per node for each diskgroup.

In the case of a RAC cluster with 3 or more nodes, this is not an issue. You can point to 3 ASM disks (Verident SSDs) in each server of the same diskgroup to store the voting disks. Even if one of the server goes down, the surviving nodes can access 2/3 voting disks and hence will not get evicted.

In a 2 node RAC cluster, if you store 2 voting disks on one server and one on the other server, then during the failure of a server with 2 voting disks, the surviving node cannot access 2/3 voting disks and the surviving node will get evicted. For such cases a new quorum failure group concept is introduced. A quorum failure group is a special type of failure group. Disks in this failure group do not contain user data and are not considered when determining redundancy requirements. The quorum failure group can be created from standard NFS based or iscsi disk that the RAC cluster can access. That way, when one of the nodes goes down, the surviving node can access the quorum voting disk and voting disk on its storage. That will add to 2/3 voting disks and the node will not get evicted.

The issue while configuring a quorum diskgroup is, when we install a RAC using Oracle Universal Installer, the default installer does not have an option to create a quorum disk while installing using the GUI installer in both 11gr2 and 12c. The workaround for this situation is to first create extra asm disk (by partitioning the vShare disk) and install the RAC cluster and ASM diskgroups.

After that, you can manually move the voting disk to an nfs or iscsi destination. Quorum failure group can be created from standard NFS based or iscsi disk that the RAC cluster can access. The space you need for the voting disk typically is less than 300MB.

## References

- http://www.oracle.com/technetwork/products/clustering/overview/twp-rac11gr2-134105.pdf
- http://www.oracle.com/technetwork/database/enterprise-edition/extendedrac10gr2-131186.pdf

## Contact Information

Estuate
1183 Bordeaux Drive, Suite 4
Sunnyvale, CA  94089
Phone: 408-400-0680

Fax:
http://www.estuate.com

Virident | HGST, A Western Digital Company
500 Yosemite Drive, Suite 108
Milpitas, CA 95035-5444
Phone: (408) 503-0100

Fax: (408) 263-7760
http://www.virident.com

Estuate …