Abstract—As IT organizations are pursuing database High Availability (HA) solutions to ensure and protest critical commercial data, the challenge is to leverage the three-fold key dimensions: cost, performance and availability. A successful solution needs to integrate Database Management System (DBMS) seamlessly with back-end storage and offer good performance and customer data protection. Through extensively investigating plenty of productions in the market, Oracle database is a leading solution for business critical application. As a cost effective solution, it could reduce business risk and achieve data availability, performance, and Return on Investment (ROI). On the other aspect, high-end storage solution, Storage Attach Network (SAN) based on Fibre Channel protocol can be easy to deploy with Oracle database and match end-user’s requirements. In this paper, we will present a design of high available database on new reference architectures. The implementation of design is based on Oracle database with Dell Acceleration Appliances for Databases (DAAD), which introduces flash technology by Fusion-IO. To demonstrate practical performances, we built a proof-of-concept platform and compared the platform with a traditional 96-SAS-drive platform. The results show that our approach can deliver more than 1-million random Input/output Operations Per Second (IOPS) that is 27 times faster than the traditional platform and it can also achieve a 96% reduction in the latencies, thus showing the scalability of the approach with massive database nodes. Therefore, our approach provides not only an ultra-fast storage solution to boost database performances but it also offers a flexible and high available design to achieve zero-downtime database.

Keywords— clouding computing; high availability; database; storage; flash technology

I. INTRODUCTION

Cloud Computing is becoming a prevailing provision of computing solutions, where the on-demand high-performance computing resources could be efficiently provisioned on a shared environment. In such emerging Cloud Computing platforms, the database is essentially built to offer the high quality data services such as DBaaS (Database as a Service) [1, 2]. As usual, the Cloud users need to access the large volumes of diversified data. To address the needs for data services, the Cloud providers must deploy a high performance and high available database system which can support the substantial data access and guarantee the availability of critical data. For examples, banks, financial institutes, stock exchanges, even companies where the downtime of core database needs to be minimized, must design and implement their database with high availability guaranteed.

To build DBaaS in a shared infrastructure, multiple databases usually have to be consolidated into a multitenant architecture. In this scenario, to implement high availability features for database Cloud is more essential and challengeable because (1) the infrastructure downtime means the downtime for all databases, which can cause a great business impact; (2) it is hard to find a downtime for maintenance that meets every database’s requirements. Our approach is to provide a fully redundant hardware design to support high availability database. The benefits of high availability database include (1) hardware upgrade and maintenance does not cause database shut down as the data is mirrored on other parts of the storage; (2) each database can be updated one-by-one without any negative effects to other databases in the system.

Through sharing resources in the Cloud environment, the approach of consolidating multiple databases can not only reduce the hardware costs, but also maximize the utilization of systems and improve the aggregate performances. From a practical perspective, the performance of individual databases is limited by the bottleneck of the infrastructural components—processor, memory, network as well as storage. Especially in
A fundamental approach to build a HA infrastructure is to deploy redundant and often idle hardware and software resources. In addition to the main goal of downtime prevention, high availability includes a wide range of design factors. Key dimensions of a comprehensive HA architecture include: (1) Data availability and protection: ensuring data access and preventing data loss to keep business continuity; (2) Performance: delivering sufficient response time for business operations; (3) Cost: reducing deployment, management, and support costs to preserve company resources. Successful HA design begins with understanding the service levels required by the business along each of these dimensions [6, 7].

In the following part, basic principles and techniques utilized in providing HA in computer systems are explained. The content of this part is based on the works in refining high availability concepts for modern computing systems [8-13]. A fault can be defined as an unexpected behavior of a system and can be visible and non-visible. The term failure is a fault that is externally visible to the end user. There are many different causes for failures in database systems: (1) system overloading can cause failures. An example is 99% transactions finished in a two seconds window; (2) protective shut down can cause failures, since the system forces a failure to prevent permanent damage; for example, when high temperature readings is observed unusually; (3) further causes are catastrophic events such as flood, tornado. The term outage is used to describe any kinds of deviations from specified system behavior, whether it is expected or unexpected. All faults and failures can be classified as unplanned outages, while specified system functionality, such as software upgrades and system maintenance, are planned outages.

With the analysis of the various root causes of the potential system outages, we start to go through replication mechanism, a redundancy strategy. When a component fails, the redundant component replaces it. The variation of configurations can be categorized as follows: (1) Cold Standby requires an automatic procedure to put the redundant component into service, when service is interrupted. A cold standby solution typically loses the component state and it provides hardware redundancy; (2) Hot Standby requires full component state replication and an automatic fail-over procedure from the failed to the redundant component without any component states lost. A hot standby solution provides both hardware and software redundancy. Cold and hot standby are active/standby configurations commonly used in HA computing. The number of standby components may be increased to tolerate more than one failure at a time. The more complicated active/active configurations require more than one redundant system components to be active, i.e., to accept and execute state change requests. In general, active/active configurations include two or more active components that offer the same capabilities and maintain a common global component state using virtual synchrony or distributed control [8]. There is no service interruption and state loss, since active services run in a virtual synchrony without the need to failover [11].

To measure the availability of system, we need to introduce the term: Service Level Agreement (SLA). SLA determines the kinds of applications to develop, and HA systems provide the hardware and software framework in which these applications
can work effectively to provide the needed level of service. Here, HA implies a service level in which both planned and unplanned system outages do not exceed a small stated value. It is very common to see that SLA is defined by the percentage of the availability time, for example, 99.95%, which means by annual only 4 hours and 22 minutes downtime is allowed for the system. SLA also specifies where the downtime is allowed. For example, the downtime window is the first Saturday from 8pm to 10pm every quarter.

In the rest of the section, we review two popular disk technologies used in the storage. Traditional storage uses Hard Disk Drives (HDD) as the main nonvolatile devices; HDD is essentially metal platters with a magnetic coating. Due to mechanical limitation, the speed of random read/write is hardly to improve and reliability is also a major concern. Today, the capacities of 2.5-inch and 3.5-inch drives have grown from multiple megabytes to multiple terabytes (TB), an increase of millions fold. Current 3.5-inch HDD tops out at 10TB, with 2.5-inch drive at 3TB max. Compared to HDD, Solid State Disk (SSD) stores the data on interconnected flash memory chips and retains the data when no power present. The 2.5-inch SSD capacity currently tops out at 4TB. As a low cost solution, HDD is still widely used in variation of storage solutions, but SSD based flash technology provides more durable and much faster data services to the customers [14-20].

III. EXISTING SOLUTIONS

With the analysis of the various root causes of the potential system failures and the technology of achieving HA systems, we will review existing software and hardware approaches to provide HA database and HA storage.

A. HA Database

In this session we discuss various architecture options for HA database architecture design.

As an active/active model, Oracle Real Application Clusters (RAC) provides the following features: (1) protect database availability against up to N-1 server failure; (2) reduce planned downtime for hardware, OS, software upgrade; (3) add nodes or remove nodes based on demands of capacity; (4) balance application loads (Fig. 1). When one node fails, the workload will be re-assigned to the rest of working nodes to achieving load balancing and data availability. In the meanwhile, based on the real scenario, adding more hardware resource to avoid system overloading or removing the node to replace or upgrade are easy and will not impact the whole system. From the hardware perspective, each of the server nodes, which run RAC instances, can be setup to fully redundant connects to the back-end database storage to avoid any failures on each components: Host Bus Adaptor (HBA) cards, cables and network switches of database storage.

The other database model, active/standby, is the second option, for example, Oracle RAC one node database. It is a single node database and no load balancing between nodes. It requires the same system architecture as RAC: network and storage. The system utilization of RAC one node is not good as RAC due to the standby server is idle. The reason we use RAC one node as alternative to RAC is because of its low cost.

B. HA Storage

The oldest storage redundancy technology should be Redundancy Array of Inexpensive Disks (RAID), which is a data storage virtualization method that multiple disk drive components put into a single logical unit for the purposes of data redundancy and performance improvement. The term “RAID” is presented by David Patterson et al. at the University of California at Berkeley in 1987. A number of standard schemes have evolved and they are called RAID levels. Originally, there were five RAID levels, but many variations have evolved. RAID levels and their associated data formats are standardized by the Storage Networking Industry Association (SNIA) in the Common RAID Disk Drive Format (DDF) standard [21]. RAID is widely used to provide data redundancy in hardware and software layer. When a disk failed, the data can be recovered from the rest disks to achieve better data availability [22, 23].

As more expensive storage solutions, NAS and SAN, provide more stable data access and data recovery. Traditional NAS solutions reduce the possibility that a single point of failure can cause the data served by the cluster to become inaccessible to client systems, but are employed for testing and developing database due to their performance limitations. Performance limitations arise due to additional overhead of managing the network file systems and their lack of scalability options. The NAS system has to be scalable and equipped with features such as a Scalable File System bundled with redundant cache optimized controllers in order to fit into database storage system.

Another network storage solution is SAN, which only provides block level of data access through Fibre Channel, iSCSI (Internet Small Computer System Interface) and ATA over Ethernet and HyperSCSI. SAN opens to clients as disk volume and it can be formatted and mounted by volume management utilities. For example, HA-SAN storage solution by Hewlett-Packard includes 8Gb (Gigabit/s) Fibre Channel (FC) SAN switches and a pair of disk-array storage servers operate in an active/standby configuration, however both servers are actively taking on their full IO processing capabilities at all times. All connections on the SAN network are fully redundant [24].

To balance the cost, performance, and data availability, we present our reference architecture (DAAD) based on Fusion-IO flash technology and Fibre Channel-based SAN to achieve our
HA enabled storage solution for database. Key design principles: (1) utilizing 4 ioMemory cards, which integrate flash memory with the disk controller, on each server and achieving best IO performance; (2) providing up to 12Tbytes storage with data redundancy; (3) considering performance and data availability, SAN is a better choice for the solution. Next section will describe the infrastructure of the solution.

IV. INFRASTRUCTURE

This section discusses the combination of an RAC database with DAAD and the configuration method used. Our approach is to establish a solution for RAC databases based on the Fibre Channel version of DAAD with the HA option enabled. However, this approach also applies to a single node non-RAC database system [4, 5].

A. Solution Architecture

In the solution, RAC database uses a clustered pair of Fibre Channel DAAD nodes as the shared storage to store database files as well as the Oracle Clusterware Registry (OCR) and voting disk files of Oracle Grid Infrastructure. ION accelerator appliance, by Fusion-IO, enables HA storage architecture that prevents the Grid Infrastructure and the RAC database from a single point of failure of hardware and software components.

Fig. 2 shows the architecture of RAC database on the DAAD. The database nodes connect to the storage appliance through a Fibre Channel network. The database nodes are connected to the private network for the cluster heartbeat and the data synchronization between the database nodes. All the database nodes are connected to the public network which allows users applications and Database Administrators (DBAs) to connect to the database nodes. The public network is usually connected to the main corporate network through Internet.

B. Physical Architecture and Network Software

To ensure the HA of the infrastructure, the network and storage IO paths are redundant. The following diagram (Fig. 2) shows the physical implementation of our approach.

From the top down, the two public switches are connected to the RAC database nodes and the network backbone of the data center or corporate network to which all the applications are also connected. Through this network, the applications send the SQL statements to the database server nodes and the database server nodes send the database query results back to the applications. Two switches are dedicated for private network. The private network is a non-routable network and should not be connected to other networks, particularly the public network.

As shown, the Fibre Channel network that connects the RAC database server nodes and DAAD consists of two 16Gb Fibre Channel switches. Each database server node is connected to both of these Fibre Channel switches, as is each DAAD node. The database server nodes are connected to DAAD through the Fibre Channel network. This configuration ensures that there is no single point of failure in the storage network.

C. DAAD Appliance

As shown in Fig. 3, in this configuration, two appliance nodes are connected to two Fibre Channel switches. To increase storage network bandwidth, each appliance node is configured with two dual-port 16Gb Fibre Channel host bus adaptors (HBAs) that together provide four total Fibre Channel links split between the two Fibre Channel switches. It also ensures that the entire storage network is free of a single point of failure.

Each appliance server has 4 ioMemory cards (Fig. 3), which are 3TB each card. DAAD also offers a standalone option that consists of a single appliance node. This option, we achieve HA on ioMemory card level with a RAID 10 configuration by mirroring two ioMemory cards with another two ioMemory cards. This configuration prevents any data loss due to the unlikely failure of an ioMemory card. As a result of the mirroring, the usable capacity of each node will be reduced to 50% of the original capacity, namely 12TB to 6TB per appliance node. However, a single appliance node presents a single point of failure, so the highly available configuration is more desirable.

For mission critical database applications, the acceleration appliance provides the HA architecture to avoid any database downtime or data loss in the event of a failure or maintenance of the appliance node. This HA configuration consists of clustered acceleration appliance nodes that store redundant copies of the data. The storage volumes from this pair of appliance nodes are mirrored so that the data stored in one node of the pair is replicated to the corresponding mirrored node of the same pair. Mirroring and data replication can be implemented by ION HA Clustering, a powerful and effective array-based solution for high availability. ION HA clustering is enabled by ION HA software. It uses a private 40Gb point-to-point network that links the clustered appliance nodes. ION HA replicates all the data block changes between two clustered appliance nodes over this interconnect network using this point-to-point connection. The advantage of our approach is that HA is provided on storage level. ION HA clustering is used to implement HA for an RAC database configuration since it can provide the external redundancy for the Oracle Grid Infrastructure voting disk needed to ensure that the voting disk is accessible to the RAC Database nodes even when one of the appliance nodes is offline.
D. Fusion-IO Storage

Each Appliance node has four 3.0TByte ioMemory cards. In the reference configuration with ION HA clustering enabled, the four ioMemory cards (ioDrive1-4) of one Appliance node are mirrored by four ioMemory cards (ioDrive1-4) of another node in the HA cluster.

In the storage configuration, we create a storage pool for each ioMemory card and create two volumes for each storage pool. Each volume has a primary appliance node and a secondary node. As in Fig. 4, volumes V1, V3, V5 and V7 use Appliance node 1 as the primary node and volumes V2, V4, V6 and V8 use Appliance node 2 as the primary node to present the volumes to the database nodes. When the database servers update data on the volumes, updates load onto the primary nodes and are then replicated to their mirrored volumes on the secondary node. For example, the update on volume V1 will first come to the V1 of ioDrive1 on Appliance node 1, then the updates will be replicated through the 40Gb Ethernet HA links to the ioDrive1 on Appliance node 2. This design allows us to balance the workloads evenly over two appliance nodes.

In the test platform, the volume V1 is created with size 1.5TB on storage pool ioDrive1 with node1 as the primary node and node2 as the secondary node. Then we create eight 1.5TB volumes V1, V3, V5, and V7 with node1 as the primary node and volumes V2, V4, V6, and V8 with node2 as the primary node as shown in Fig. 4. Each of these eight volumes is created with the same size as they will be presented to the database nodes to form Oracle ASM (automatic storage management) disks of an ASM disk group.

V. EXPERIMENTAL EVALUATION

We use three different tools to measure the IO performance of the storage appliance and the overall performance of the entire RAC database environment, they are: (1) Iometer; (2) CALIBRATE_IO; (3) HammerDB. The database configuration for performance studies is based on the reference configuration described above. The configuration consists of four database nodes and two-nodes DAAD with HA enabled [25-32].

Iometer is used to compare the storage IO performance of the appliance against a baseline configuration that used 96x 15k SAS disks in a standard Fibre Channel SAN. We use four access patterns to ascertain the performance: (1) 4k Random Read; (2) 8k Random 70/30 Read/Write; (3) 1MB Sequential Read; (4) 1MB Sequential Write.

CALIBRATE_IO is a procedure in Oracle Database. We use this procedure to measure the performance of a database and studied the performance scalability by the number of database nodes for the following metrics: (1) Latency; (2) Max IOPS; (3) Max MB/s.

HammerDB is used to measure the TPC-C-like performance of the database and the scalability of the performance by the number of database nodes. The performance metrics we used to analyze the performance of the system are: (1) Transaction Per Minute (TPM); (2) New Orders per Minute (NOPM); (3) Average Response Time (AVT).

A. Iometer

Immediately we see that the DAAD pushes some very high random IOPS and sequential throughput numbers. Table 1 lists the detailed data from both the DAAD configuration and the 96-Disk Baseline.

The data in Table 1 and Fig. 5 demonstrates the significant performance improvements of the DAAD over the baseline. For example, in the 4k Random Read test the DAAD produces 27x the IOPS of the 96-Disk Baseline configuration while reducing Latency by 96%. The detailed latency data is showed in Table 2. For sequential read and write, DAAD only offer about 3X and 4X speed-up due to the performance of block-read and write on HDD is much better than that of random read and write. Considering the workload of our target customers, DAAD can achieve huge speed-up, but slightly increase storage investment [33-36].

B. CALIBRATE_IO

CALIBRATE_IO provides information similar to that offered by Iometer except that the data access patterns used are provided by Oracle and are intended to be representative of common database workloads. Max IOPS is designed to simulate the workload of OLTP through a transactional database, while Max MB/s is simulating the workload of online analytical processing (OLAP) through an analytics database. In each case, the DAAD is much faster than the baseline of standard rotating disks. There is a limitation in this test, however, in that it was designed before the implementation of high-performance, low-latency storage solutions like the DAAD. The latency is reported as an integer value in microseconds. Therefore, when the latency is less than one
microsecond the reported number is zero. In testing the DAAD, we never saw a non-zero latency result. It still shows the incredible speed of the Fusion-IO ION appliance, even with high throughput.

CALIBRATE_IO measures a combination of disk performance and database performance, it is beneficial to test how the performance scales across multiple nodes showed in Fig. 6 and Fig. 7. We ran the tool four times, adding a node after each set of tests for a total node count of four. We note from this test that performance is greatly improved by scaling from one node to four nodes, while latency is still not measurable with this test. At four nodes we are not yet able to saturate the DAAD and would likely see some increase in performance above four nodes.

C. HammerDB

The DAAD with Oracle Database shows strong scaling in our primary application benchmark while keeping Average Response Time (ART) very low. The use of PCIe flash storage in the DAAD alongside the low-latency Fibre Channel storage fabric allows for very large increases in performance over traditional storage while driving very low response times.

HammerDB allows us to simulate (though not exactly duplicate) a traditional TPC-C database workload and gain an understanding of how the appliance performs in a real world environment. Typically we tune our generated load so that 10 milliseconds (ms) is an acceptable average response time.

NOPM is typically the most consistent measure of performance, as it is database-independent and does not fluctuate as much as GTPM. As shown in Fig. 8, the peak NOPM reached almost 960,000 with four database nodes. The peak TPM reached over 2.5 million with four database nodes. Throughout the testing, the ART remained well below 10ms. Detailed data is included in Table 3.

VI. CONCLUSION

In this paper, we present the design and implementation based on Oracle RAC Database with Dell Acceleration Appliances for Database (DAAD). The results of performance studies are given in terms of the storage IO performance and they are compared with the baseline system that is based on 96x 15k SAS drives. The reference architecture achieved: (1) 27 times IOPS and 96% reduction of response time for 4k random read, and (2) 2.6 times MB/S throughout for 1 MB sequential read. When we run the TPC-C-like performance benchmark using the HammerDB testing tool with four RAC database nodes, the peak New Orders per Minute (NOPM) reached almost 960,000 with an ART less than 8ms. The test data also shows the near-linear scalability of database performance with increase of database server nodes.

IT organizations as well as Cloud database providers can take the following benefits of our approach: (1) high performances: delivering over 1 million random IOPS, sub-millisecond latency for OLTP transactional type workloads, and tens of gigabytes per second bandwidth for OLAP sequential reads and writes in Decision Support Systems and Data Warehouses. (2) More than 9x application acceleration for demanding application and user workloads. This approach can enable substantial concurrent uses with quicker response times and faster batch processing. (3) A scalable storage solution that allows customers to add additional pairs of the two-node HA configuration, which seamlessly integrates with any database systems, to scale the performance and capacity. The handy instructions could help the practitioners to construct a HA database.
VII. ACKNOWLEDGMENT

We would like to acknowledge that Yuxiang Gao was a system engineer in the Global Solutions Engineering Group at Dell Inc. from April 2011 to February 2014. We also would like to acknowledge that Meikang Qiu is supported by NSF 1457506.

VIII. REFERENCE


Kai Yu is a senior principal engineer and technologist in Dell Oracle Solutions Engineering lab. He focuses on design and architecting of various IT solutions. In April 2010, he became one of 51 Oracle ACE Directors (as in April, 2010). He has published 100+ technical presentations and 25 technical papers. He was awarded the 2012 Oracle Excellence Award: Technologist of the Year: Cloud Architect by Oracle Magazine. He is a co-author of Apress book “Expert Oracle RAC12c”. He holds M.S. degree in Computer Science in University of Wyoming, and M.S. degree and B.S. Computer Science in Huazhong University of Science and Technology, China.

Yuxiang Gao was a senior HPC consultant in the Global Solutions Engineering Group at Dell Inc. until 2014. He received his PhD degree in Applied Mathematics from Stony Brook University in 2010 and B.S. degree in Mathematics in Peking University in 1996. He focuses on accelerator-based cluster design and performance analysis.

Peng Zhang is a senior research associate in Biomedical Engineering Department at Stony Brook University. He received his PhD degree in Applied Mathematics from Stony Brook University in 2012, and M.S. and B.S. degrees in Parallel Computing and Mathematics from Nankai University in 2006 and 2003. He is working on the multidisciplinary programs for simulating complex biological systems on modern high performance computing (HPC) systems. He focuses on optimizing large-scale parallel applications on modern HPC systems, for offering computationally affordable solutions.

Meikang Qiu received the BE and ME degrees from Shanghai Jiao Tong University, China, and the MS and PhD degrees of computer science from University of Texas at Dallas, in 2003 and 2007, respectively. Currently, he is an Associate Professor of Computer Science at Pace University. He has worked at Chinese Helicopter R&D Institute, IBM, etc. He is an ACM Senior member. His research interests include cyber security, embedded systems, cloud computing, smart grid, microprocessor, data analytics, etc. A lot of novel results have been produced and most of them have already been reported to research community through high quality journal (such as IEEE Transactions on Computer, ACM Transactions on Design Automation, IEEE Transactions on VLSI, and JPDC) and conference papers (ACM/IEEE DATE, ISSS+CODES and DAC). He has published 3 books, 170þ peer-reviewed journal and conference papers (including 70 journal articles, 100 conference papers), and 3 patents. His research is supported by NSF. He has won Naval Summer Faculty Award in 2012 and Air Force Summer Faculty Award in 2009. His paper about cloud computing has been ranked 1st in Top 25 Hottest Articles published in JPDC (Journal of Parallel and Distributed Computing, Elsevier) the whole year of 2012. He is the receipt of ACM Transactions on Design Automation of Electrical Systems (TODAES) 2011 Best Paper Award selected from a 3-year window. In recent four years, he has won another 4 Conference Best Paper Awards (IEEE/ACM ICESS’12, IEEE GreenCom’10, IEEE EUC’10, IEEE CSE’ 09). He is a Senior Member of the IEEE.