SOLUTION BRIEF

Memblaze Partners with Mellanox to Advance PBlaze SSD-based High Performance Oracle RAC

As stated in Oracle's Engineered Systems evolution plan, NVMe SSD storage is being employed in the application of enterprise solutions. The outstanding features of NVMe SSD such as high bandwidth, high IOPS and low latency are rapidly enhancing the performance of databases. Mellanox has partnered with the leading Chinese flash memory producer, Memblaze, to advance the high performance open architecture Oracle RAC solution.

THE SOLUTION

PBlaze4 Series is Memblaze's Generation IV solid state drives for enterprise applications. This series features not only outstanding quality, but provides guaranteed consistent service and high reliability which are great help to meet the challenges in modern data centers where the equipment always needs to operate 24x7. PBlaze4 increases the speed of data centers in terms of database access, virtualization, CDN, cloud computing and many other applications.

Leveraging InfiniBand's proven scalability and efficiency, Mellanox helps our clients easily build large clusters that can be extended to thousands of nodes. Mellanox is well-known for providing world leading technologies featured with low latency, high bandwidth, high transmission rate, extremely low CPU usage, remote direct memory access (RDMA) and optimized communication offload, etc., has become a leading solution provider for large-scale, end-to-end, high-speed interconnect technologies..

- InfiniBand switch provides a high speed of up to 200Gb/s on each port, enabling the computing cluster to extended to tens of thousands of nodes
- InfiniBand switch can be used to achieve high server efficiency as well outstanding application performance through high bandwidth and a low latency of less than 90ns
- Most cost-effective solution with up to 40Gb/s-200Gb/s operation without any error

In this comprehensive solution, distributed data servers and the Oracle database server architecture are connected via an InfiniBand network. In each data server, a PBlaze4 SSD disk is installed, which uses iSCSI protocol to export the data to the block device. The database server also uses all of the PBlaze4 SSD block devices through iSCSI protocol and uses Oracle ASM to make data redundant to secure it. In actual data testing, HammerDB runs on the application server, the number of virtual users is adjusted to verify the functionality of OLTP TPMc.

InfiniBand is a high-speed, multi-purpose, network architecture designed for I/O transmission networks, such as LAN for storage or cluster networks. In the high-performance computing field, InfiniBand has become an advanced industry standard. InfiniBand implements an application messaging service called channel I/O. And, in order to achieve high performance under some specific environments, it eliminates the complex network stack operations

KEY BENEFITS

- **Fully Supports NVMe:** NVMe 1.1, PCIe 3.0
- **Sustainable High Performance:** Outstanding QoS, continued high IOPS and bandwidth, as well consistent low latency.
- **Verified Data Reliability:** ECC, RAIN, pSLC, enhanced power-fail protection and temperature protection
- **Hot Plugging:** SFF-8639 interface and front maintenance
- **High Return on Investment (ROI):** Occupies a small amount of memory and CPU, but increases storage density, resulting in a reduced total cost of ownership (TCO)

INFINIBAND SWITCH BENEFITS

- InfiniBand switch provides high speeds of up to 200Gb/s on each port, enabling the computing cluster to extended to tens of thousands of nodes.
- InfiniBand switch can be used to achieve high server efficiency as well outstanding application performance through high bandwidth and a low latency of less than 90ns.
- This is the most cost-effective solution with up to 40Gb/s-200Gb/s operation without any error.

"In recent years, the rapid development of flash technology has been adopted by the majority of industry users, with the flash performance and capacity to be further enhanced, it will become increasingly popular and generate unprecedented pressure on the data center network. The Mellanox end-to-end, high-speed, low-latency network will undoubtedly be the key to solving this issue."

Zhibo Tang, CEO of Memblaze

within the operating system. For a pair of applications that have turned on InfiniBand features, the InfiniBand protocol will create a channel to connect them, sending messages directly to each other by sending and receiving queues. These sending and receiving queues will map memory to the user's state space of the application, known as remote DMA (RDMA).

PERFORMANCE TESTS

For the test architecture we created 5000 data warehouses on the application server, run HammerDB to generate the TPCC transactions, and then sent these transactions to the database server via Ethernet. The database will then process these requests, and write the logs and data re-generated by Oracle to a data server through the InfiniBand network, and eventually write to the PBlaze4 SSD device. With flexible and strong InfiniBand networking technology, plus decoupling calculation and storage function modules, the computing nodes and storage nodes can be expanded easily. In addition, Memblaze PBlaze4 supports hot plugging technology, enabling the addition of storage resources dynamically while reducing maintenance downtime.

- Application Note x 1
  - Dell™ PowerEdge R720 rack server
  - 1 x Intel® Xeon® Processor E5-2630 (6 cores) v3 CPU
  - 2 x 8GB DRAM
  - CentOS 6.5
- Database Node x 3
  - Dell™ PowerEdge™ R720 rack server
  - 2 x Intel® Xeon® Processor E5-2680 (12 cores) v3 CPU
  - 8 x 16GB DRAM
  - CentOS 6.5
- Data Server x 3
  - Dell™ PowerEdge™ R730xd rack server
  - 2 x Intel® Xeon® ProcessorE5-2630(6 cores) v3 CPU
  - 8 x 8GB DRAM
  - 1 x Memblaze 3.2T PBlaze4
  - CentOS 6.5
- Network
  - Mellanox 100Gb/s InfiniBand EDR Switch
  - Mellanox ConnectX-4 100Gb/s network adapter
  - Mellanox LinkX high-speed cable
- Test Tools
  - HammerDB 2.19, TPC-C test tool
- Software
  - Oracle® RAC 12C, database software
  - STGT, block device export

TPM Test Results

Figure 3 clearly shows the results of TPM tests where different numbers of virtual users were specified. These results include the transactions that the users submitted per minute and the transactions that the user rolled back per minute. NOPM (shown in orange) indicates the number of events generated in new order per minute. From Figure 3, we can see that when the number of virtual users exceeded 230, the number of TPMs began to drop, at which point the database server CPU utilization reached 100% and formed a bottleneck.

As shown in Figure 4, when the number of virtual users reaches 230, the number of transactions handled by the database in one second is up to 16,555.9 with IOPS reaching 9,513.5 and a data transmission rate of 100.4MB. PBlaze4 SSD can easily handle this load since most I/O requests are buffered by the memory. At this point, in the database server, CPU utilization has reached 100%.

HammerDB runs a TPMc diagram that runs sequence tests automatically according to the number of virtual users, a TPM of 1,575,840 corresponds to a status where 230 virtual users are running simultaneously. With the increase in the number of virtual users, TPMc draws a smooth curve, which means good scalability.

Figure 1. Illustration of How RDMA Works

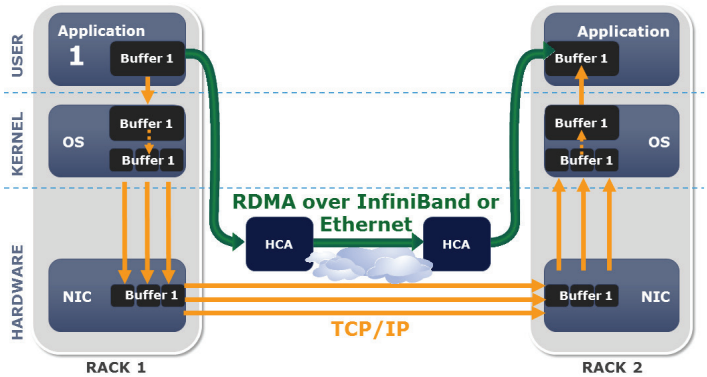

Figure 2. Test Topology Architecture

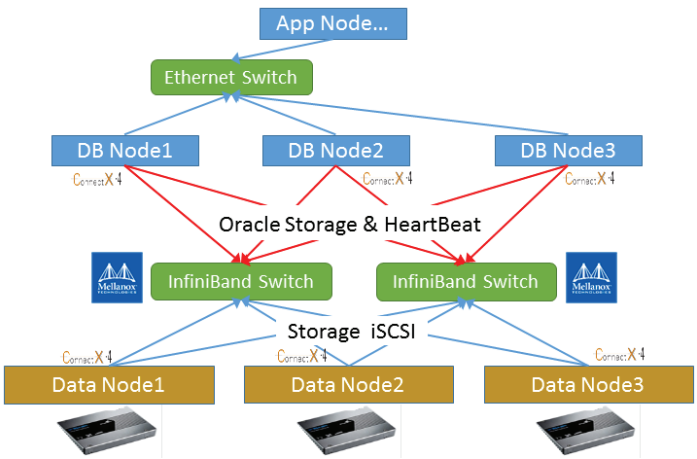

Figure 3. Benchmark Test Results from Different Virtual Users

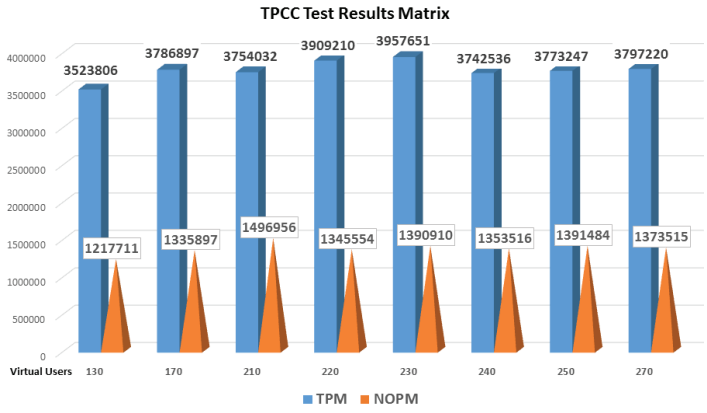

Figure 4. AWR Report of Database Instance (with 230 virtual users)

Load Profile

	Per Second	Per Transaction	Per Exec	Per Call
DB Time(s):	70.7	0.0	0.00	0.01
DB CPU(s):	29.4	0.0	0.00	0.00
Redo size (bytes):	93,830,550.3	5,667.5		
Logical read (blocks):	1,955,351.1	118.1		
Block changes:	527,323.3	31.9		
Physical read (blocks):	4,807.9	0.3		
Physical write (blocks):	8,039.1	0.5		
Read IO requests:	4,799.6	0.3		
Write IO requests:	4,713.9	0.3		
Read IO (MB):	37.6	0.0		
Write IO (MB):	62.8	0.0		
Global Cache blocks received:	2,585.6	0.2		
Global Cache blocks served:	2,657.3	0.2		
User calls:	12,687.0	0.8		
Parses (SQL):	7,062.7	0.4		
Hard parses (SQL):	0.3	0.0		
SQL Work Area (MB):	3.0	0.0		
Logons:	0.4	0.0		
Executes (SQL):	341,660.9	20.6		
Rollbacks:	27.8	0.0		
Transactions:	16,555.9			

Test Conclusion

According to the HammerDB test results, the CPU performance of the database node can be the bottleneck of system performance. The Memblaze PBlaze SSD storage and Mellanox high-speed network show excellent performance in the Oracle database environment, and the overall TPMc performance and scalability of the comprehensive solution are outstanding.

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end Ethernet and InfiniBand intelligent interconnect solutions and services for servers, storage, and hyper-converged infrastructure. Mellanox offers a choice of high performance solutions: network and multicore processors, network adapters, switches, cables, software and silicon, that accelerate application runtime and maximize business efficiency for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage, network security, telecom and financial services. More information is available at www.mellanox.com.

About Memblaze

Founded in 2011 and based in Beijing, Memblaze is a technological company with innovative genes. It focuses on providing outstanding enterprise-level solid-state memory products and solutions in fields such as IT, internet, communication and cloud computing.
http://www.memblaze.com/en/

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

15-52427SB
Rev1.0