IBM and Mellanox

Enabling Highly Available, Elastic Storage for Complex Modeling, Research and Analytics

Executive Summary

In today's new era of computing, the need for rapid analytics is no longer being meet by the resistance of having to process complex computations. Providing the powerful infrastructure needed to run large and complex simulations quickly and effectively has been answered by multicore processors, clustered scale out computing and the lower price of RAM.

The compute power required for solving these type of calculations has been available for some time now. However, managing the growing amount of unstructured data that is being produced by these calculations is now the new challenge. With a wide range of research disciplines that require supercomputer capabilities growing; including climate change modeling, physics and life sciences studies, and research clusters for high-frequency trading, the delivery of large amounts of unstructured data is growing by leaps and bounds. Often times, this data is known as "scratch data", which is used during the simulation run, but doesn't always need to be stored for long term access. The most important criteria of a storage infrastructure is that it must match the capabilities of enterprise reliability together with supercomputing high bandwidth and capacity. It's important to single out bandwidth for this paper as we'll be discussing the use of a 56 Gb Ethernet infrastructure, compared to 40 Gb Ethernet. And as for storage in these environments goes, performance is typically network bound, therefore the achievable performance depends heavily on the network technology used and its scaling capabilities.

Mellanox 56Gb Ethernet delivers greater results, at the same cost as 40Gb

- 40% Greater Throughput
- 2.5X Greater Processing Power
- Decreased Recovery Times
- Sub- µsec Latencies
- Highest Message Rates
- Linear Scalability

IBM Elastic Storage Server

IBM Elastic Storage Servers (ESS) is a high-performance, General Parallel File System (GPFS) network shared-disk solution that is perfectly suited to provide fast and reliable access to data from advanced clustered servers. Applications running on clusters servers can readily access files using standard file system interfaces, while the same file can be accessed concurrently from multiple servers or protocols. IBM ESS is designed to use one or more building blocks, where a building block is a pair of servers with a shared disk enclosure attached. The GPFS software acts as the basis file system for the storage server, which uses a "declustered" RAID technology to deliver outstanding throughput, extreme data integrity, with enhanced data protection and faster rebuild times. This combination shorten the time-to-delivery for compute result. As in the supercomputing world, which differs from traditional enterprise storage where data concerns are about the availability of data, protecting it and safe archiving, it's about delivering results. The ability to use stored data is more about answering questions, in a meaningful and often times, very short time frame.

To understand how the IBM ESS can deliver on the storage requirements for enterprise and supercomputing, compare to traditional storage arrays, it resembles more of a clustered supercomputing environment itself with a front end of powerful servers as it's interface. The main building blocks are high-performance Power Systems servers with two 10 cores POWER8 processors which are further accelerated and loaded with RAM and SSDs, and with expansion drawers with many storage slots that can accommodate a broad range of storage devices, including both small and large format HDD and SSDs. Due to its capabilities of delivering reduced latency, high-performance Ethernet or InfiniBand connections and extreme data integrity, the ESS has the fastest rebuild times. In order to provide unparalleled I/O performance for unstructured data, data striping is performed across multiple disks and to multiple nodes, enabling high-performance metadata scans that assist in achieving the fastest time to answers.

Throughput Requirements

Accessing data is done over a TCP/IP or InfiniBand connection with networking choices that include 10 Gb Ethernet, 40 Gb Ethernet, FDR and EDR InfiniBand. But little is known about the 56 Gb Ethernet interface offered by Mellanox. This provides the IBM ESS 40 percent more throughput for the same cost of 40 Gb Ethernet. In computational & analytic clusters, requirements for reliable, rapid access to a vast data repository is demanded. The more bandwidth within the storage infrastructure, the better the ability a system has to answer storage demands for larger amounts of unstructured data. After all, if time-to-delivery of an answer or result is the dominant requirement that you seek, then a high-performance storage solution is the answer. At 56 Gb Ethernet, this easily surpasses even the greatest speeds achievable by Fibre Channel storage which is advertised at 32 Gb but delivers a line rate of merely 28.05 Gb per second transfer rate when you take into account encoding, interframe spacing, and frame headers. Running the IBM ESS at 56Gbps, makes it the fastest production storage device available.

Latency and Message Rates

Trading platforms, especially those supporting algorithmic trading, require latencies of less than 5 microseconds with an extremely low level of packet loss. Mellanox, the world leading provider of high-speed Ethernet and low latency networking solutions, is the first to introduce 56GbE speeds for adapters. Making the speed jump from 10 or 40, to 56GbE will have an exponential increase in delivery of message rates. This ensures that the network and connections to servers are not creating a bottleneck that impede optimum time to solution well into the future. Ultra-low latencies, as low as a few microseconds for the adapter and less than 300 nanoseconds for the switch, equates to extremely high Packet Per Second (PPS) rates of near 3 million PPS. The faster speeds of 56GbE and lower latency provides for better execution rates and allows for more in-line processing.

Complex Research Studies

Since its beginnings, the GPFS file system has been deployed on a wide range of cluster types and sizes, with the larger clusters serving the scientific computing needs from universities to national research laboratories, and small-to-medium and large sized clusters serving High Performance Compute (HPC) applications. The distributed locking architecture is a good match for these scalable, general file-serving applications, especially when workloads consisting of a large collection of different systems accessing different sets of files. File access through GPFS is just as efficient as a local file system, but with a unique ability to scale to meet growing bandwidth and capacity demands. This allows for the IBM ESS solutions outstanding performance that can be complimented with 56 Gb Ethernet for the highest performing solution available for researchers as they conduct complex studies or modeling.

Decrease Recovery Time

In the larger storage environments that are necessary to keep pace with the research fields that are utilizing supercomputers for analyzing and providing the fastest response times, a small but significant proportion of hard drives have the potential of failing every week. The more terabytes of data stored, the more disks, the more likely a failure is to occur. In these environments, a disk failure is the norm, rather than a rare and exceptional case. In a typical RAID scenario, when a disk failure occurs, all of the remaining disks become critical to the rebuild process. However, in a declustered array scenario that is utilized by GPFS, only a small percentage of the disks become critical to the rebuild. Rebuilding just those critical stripes to get the system to a non-critical state takes only a few minutes. The rebuild of the remaining non-critical aspects of the drives take place only when there is no user I/ O activity on the system. This avoids any performance impacts on user applications and is an essential feature in high performance, large capacity file systems such as the IBM ESS.

In a conventional RAID protected storage array, it may take up to 12 hours to rebuild a single TB drive, and the I/O and processing power required during the rebuild can negatively affect the subsystems processing power, performance and overall chews into available bandwidth. The IBM ESS solution enables rebuilds of a TB disk, to a non-critical state, in under an hour. This is possible due to the GPFS Native RAID (GNR) an advanced, spare space disk layout scheme that uniformly spreads or “declusters” data through a software RAID algorithm that is integrated into the ESS I/O servers. This eliminate the disadvantages that are seen when using standard asymmetrical RAID algorithms used in most typical storage arrays. By using a refined data algorithm and spreading the user data, redundancy information, and spare space across a much greater amount of operating disks, and utilizing 56 GbE speeds, allows for the largest possible bandwidth and significantly shortens the time required to recover from disk failures.

Conclusion

The IBM ESS with GPFS and GNR represent, first and foremost, a high-performance storage solution for organization that want to reduce the time to results that must be delivered from a very large set of data pools. The successful collaboration between IBM and Mellanox brings several interconnect solutions including 10 GbE, 40 GbE, FDR and EDR InfiniBand, that are sure to work for any environment. However, for those clusters supporting business analytics, high-performance computing applications from climate modeling to tornado simulation, with databases such as IBM DB2®, in big data MapReduce analytics, gene sequencing, digital media and scalable file serving, 56 GbE is available to provide a 40% increase in bandwidth at no additional cost. This is by far the fastest time to answer solution you’ll find on the market today.

Mellanox TECHNOLOGIES

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