CLABBY ANALYTICS

Research Report

Ultra-Dense Hyper-Scale x86 Computing: The Lenovo NeXtScale System

Executive Summary

For years high performance computing (HPC) was largely associated with large-scale scientific workloads (for example: pharmaceutical research) and with applied technical computing where it was used primarily for engineering and simulations in electronic design automation, automotive/aerospace engineering and petroleum discovery/analysis. In late 2010 and throughout 2011, however, *Clabby Analytics* noticed a distinct shift in the HPC market as workloads such as digital media, new life sciences applications, various financial services applications, on-demand cloud computing services and analytics made their way onto HPC servers. In February, 2012, we decided to publish a white paper that described this trend.

Since 2012, we've seen the arrival of new types of systems designed to accelerate HPC processing: "ultra-dense" hyper-scale servers. These new designs replaced clusters that consisted of racks of full wide, 1U servers. Some of these clusters used specialized non-standard-sized racks and were preconfigured with the vendor's preferred switch. And these clusters were generally serviced from the back (in the "hot" aisle). Buyers took on the responsibility for the configuration, reliability testing, and optimization of these designs. Further, these designs lacked density. Hyper-scale designs have changed all of this.

The availability of ultra-dense, hyper-scale designs such as Lenovo's NeXtScale System have changed system design dynamics in the HPC marketplace. NeXtScale systems are based on industry standard components including standard racks, networking, and memory and can easily be preconfigured and optimized for performance and energy-efficiencies. These servers feature dense compute nodes (1/2 wide, 1U 2 socket servers) designed for scale-out computing. When sold as an Intelligent Cluster these systems are factory integrated and pre-tested for reliability, faster deployment and time to value.

In this *Research Report*, *Clabby Analytics* takes a closer look at Lenovo's NeXtScale system design. What we find is a system that:

- Has been designed for raw throughput and performance;
- Has been positioned to handle HPC, cloud, grid, and managed hosted workloads– as well as a wide range of new workloads including computational analysis, upstream/downstream processing, next-generation genomics, satellite ground stations, video capture and surveillance, 3-D computer modeling, social media analysis, data mining/unstructured information analysis, financial "tick" data analysis, and large-scale real-time customer relationship management;

- Provides customers with a great deal of flexibility in configuration and component choices (including a wide range of processors, integrated storage, accelerators, choice of standard racks and network switches);
- Can be managed by a variety of workload, resource, and data management tools;
- Offers greater density and higher throughput than traditional 1U scale-out designs.

The NeXtScale nx360 Compute Node and n1200 Enclosure

The NeXtScale nx360 M5 compute node features Xeon E5-2600 v3 processors with up to 18 processor cores per chip in a two socket configuration that is ½ wide (so two of these half-wide nodes can fit into a traditional 1/U socket space). A closer look at this node shows plenty of on chip L2 and L3 Cache on the processor; access to up to 6 TB of local storage; access to additional storage using the storage native expansion (NEX) tray; and various input/output options (see Figure 1).

Figure 1 – Technical Specifications: Lenovo NeXtScale nx360 M5

Form factor/height	Half-wide 1U
Processor	Two Intel Xeon E5-2600 v3 series
Memory	16 DDR4 LP, 512 GB maximum with 32 GB LP RDIMM
Local Storage	Choice of one 3.5-inch hard disk drive (HDD), two 2.5-inch HDDs/solid-state drives (SSDs) (simple swap), or four 1.8-inch SSDs. Optional two front hot-swap 2.5-inch HDDs.
PCIe Native Expansion (NEX) Tray	Two PCIe-based accelerator cards (GPU, Xeon Phi)
Storage Native Expansion (NEX) Tray	Seven 3.5-inch SAS/SATA HDDs, up to 42 TB maximum
Internal RAID	Onboard SATA controller with RAID options
USB ports	One internal USB key
Ethernet	Two built-in 1 Gigabit Ethernet (GbE) ports standard
Input/output	Two ML2 ports for InfiniBand FDR or 10 GbE, two 10 GbE one PCIe (x16 PCI Express 3.0)
Power management	Rack-level power capping and management via Extreme Cloud Administration Toolkit (xCAT)
Systems management	1x shared port with 1GbE per 1/2 wide server

Source: Lenovo – September 2014

The NeXtScale enclosure is a 6U standard rack design that contains 12 bays, six hot-swappable power supplies, 10 hot-swappable power supplies, as well as fan and power controllers.

NeXtScale System Evaluation

When we evaluate HPC or high-performance cloud computing cluster system designs, we look closely at how the system has been optimized for performance. We also look at system density, efficiency, flexibility, reliability, serviceability, and manageability.

Performance Optimization

In addition to fast, efficient Xeon E5-2600 v3 processors (which can operate in turbo mode for even faster processing), the nx360 M5 compute node features high performance 2133Mhz memory (up to 512 GB per node), fast hard drives, optional high performance

solid state drives, and fast network connections (internal dual 1GbE standard connections to external data as well as options for dual port 10 GbE or FDR IB [Infiniband]).

"Density" Is Important

Denseness refers to the practice of packing more and more computing power into ever smaller system envelopes (such as ½ width computing nodes). Denseness is important from for two reasons:

1. It reduces the amount of real estate that a system occupies (footprint); and,
2. It reduces the amount of supporting infrastructure (networking components, management facilities, cooling, etc.) needed to support the system design.

> ***The goal with denseness is to pack as much processing power and storage as possible into as small a system envelope as possible in order to reduce the real estate footprint of a given system – as well as to reduce the number of additional components needed to support the core computing/storage components. From a density perspective, Lenovo's NeXtScale packs twice as much computing power into a 1U form factor as compared to traditional HPC cluster designs.***

Efficiency

Network throughput and power efficiency have been major focal points in the design of NeXtScale systems. From a networking perspective, traffic is driven through top of rack – there is no in-chassis input/output and associated overhead to deal with. Additionally, compute, storage and acceleration nodes can all be housed in a common footprint (again helping to eliminate hops, thus further improving efficiency). From a power perspective, NeXtScale uses a shared resource approach to distribute power, cooling, and mechanical resources. It has six power supply units to ensure balanced 3phase power delivery.

It is also worth mentioning that Lenovo's NextScale design supports water cooling. As far back as 2008, Clabby Analytics has been encouraging enterprises to evaluate water cooling (see here and here – it is our belief that water cooling is about 4,000 times more efficient than air cooling). With the latest M5 version, NeXtScale now offers a direct water cool design as an alternative and more efficient method of cooling the system and gaining the highest levels of energy efficiency and power savings.

Flexibility

The NeXtScale chassis measures 6U, a size that can accommodate all sorts of new and forthcoming storage designs as well as next generation microservers. Further, the NeXtScale chassis can be placed into industry standard racks; can use 3rd party switches (this is important to IT buyers who have commitments to certain network vendors); and NeXtScale can also use industry standard memory components and accelerators.

Reliability

NeXtScale systems can be self-integrated or use the "Intelligent Cluster" manufacturing process to factory integrate the cluster and pre-test system components for reliability. Manufacturing separates systems into small functional groups and runs Linpack tests across them. Results are analyzed and components either pass or fail. This work takes place at Lenovo – not at the customer site – so failed parts are replaced during the manufacturing process, thus helping improve the reliability of a system before deployment.

Serviceability
NeXtScale provides front "cold aisle" access to most components – making it possible for most maintenance work to be done in a cooler climate than can be found in the "hot aisle" in the rear of most systems. NeXtScale uses a tool-less access approach. Servers can be removed without unplugging power. Network cables and switches can also be accessed from the front – and power and other LEDs all face front for easy viewing. Also, errors that can occur when cables need to be replaced are reduced.

Manageability
One of the biggest differentiators of NeXtScale compared to other systems architectures is that NeXtScale is focused on throughput with minimal management overhead. NeXtScale has been designed to focus on executing workloads – which means NeXtScale focuses on computing and requires little management intervention while doing so. IT customers who want to deeply micromanage a NeXtScale can use tools and utilities like IBM Platform Computing and IBM General Parallel File System (GPFS) offered by Lenovo or simply use tools already found in their data center to manage x86 servers and clusters – but they may find that micro-management is not necessary for this type of environment.

Caris Life Sciences: A Customer Success Story
Clabby Analytics recently had the opportunity to interview Caris Life Sciences regarding their use of NeXtScale architecture. This company provides molecular profiling for patients – analyzing DNA, RNA, proteins, nucleotides and other elements of the human genome in an effort to determine the best course of action to achieve a better outcome for a given patient's cancer or other complex disease. Caris' tests and analyzes extremely large volumes of structured and unstructured data (Big Data) – and require high speed, high performance computing (HPC), a high performance clustered file, data tiering, and advanced storage and workload management tools.

The data sets that Caris Life Sciences analyzes are huge. The human genome has 22000 genes (approximately 1 billion data points). Caris runs a 600 gene test on tumors, and these tests generate 1500 records that are then run through a replication process for further screening. Further, Caris researchers are constantly expanding the number of genes that can be analyzed so the array of tests – and the amount of analysis – is constantly expanding. Today, each patient generates between 40 and 100 GB of information (depending on the tests used) – and Caris has, to date, served over sixty-five thousand patients so the Caris database is immense. Tomorrow, with a broader scope of analysis such as new Caris blood assays that may generate one or two terabyte sized files, and with even more patients, the Caris database will become even larger, and patient record sizes will continually increase to perhaps as much as 5GB of data per patient.

To perform its analysis, Caris needs to deal with both structured and unstructured data. The company performs next generation gene sequencing; it performs variant calling -offsetting information from a norm to identify a variant; and it uses images to examine bunches of nucleotides. The company's new blood assay examines DNA and RNA sequences and conducts protein analysis in an effort to match biomarker information (currently using 70 different molecular markers) with the proper drug treatment. Caris Molecular Intelligence Service is used to determine how much protein is present in a normal patient profile versus that of a cancer patient.

To analyze large data sets and manage large volumes of data, the company has opted to use NeXtScale clusters because of its raw throughput and performance characteristics along with IBM Platform HPC and IBM GPFS for efficient workload and data management capabilities. NeXtScale systems offer high performance with low systems management overhead, high utilization, while minimizing latency. Further, with NeXtScale systems, Caris has been able to deploy twice as many servers per floor tile as compared with x86-based blade solutions. The company uses less energy because fewer components are needed to drive NeXtScale clusters as compared to other architectures; because of the energy efficiency of the components used; and because power usage can easily be notched back during slow times). Finally, the company has been able to improve its time-to-result (the speed at which solutions are derived) due to processing efficiency and low communications overhead.

At present, the Caris NeXtScale system uses twenty x86 cores with 256 GB of solid state disk. The use of this large amount of solid state disk (SSD) is important because solid state drives offer significantly more input/outputs per second (IOPS) as compared with traditional mechanical hard disk drives (HDD). This additional speed enables Caris to read large amounts of data faster – and it enables results to be achieved significantly more quickly than if traditional hard disks had been used in the top tier of the storage hierarchy.

Summary Observations

NeXtScale systems have been designed for high performance while minimizing overhead and latency – all in a very dense design envelope. When compared to other system designs, IT buyers who chose a NeXtScale solution can expect to be able to:

- Deploy twice as many servers per floor tile;
- Use less energy (because fewer components are needed to drive NeXtScale, because of the energy efficiency of the components used in NeXtScale, and because power usage can be notched back during slow times);
- Reduce cooling costs (because air and water cooling can be used – and water is thousands of times more efficient as a conductor than air);
- Improve *time-to-result* (the speed at which solutions are derived) due to processing efficiency and low communications overhead;
- Improve system reliability (due to extensive system pre-testing); and,
- Reduce overall systems cost by purchasing industry standard components (such as industry standard racks and memory) – and by having to purchase far fewer components that traditional "mission-critical" server environments (such as blade architectures).

Businesses looking for a dense, high performance systems environment that has been pre-tested for reliability, preconfigured for fast deployment, and designed to scale easily – and that is easy to deploy and service – should be looking closely at this ultra-dense hyper-scale server architecture.

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 846-6662

© 2015 Clabby Analytics
All rights reserved
May, 2015

Clabby Analytics is an independent technology research and analysis organization. Unlike many other research firms, we advocate certain positions — and encourage our readers to find counter opinions — then balance both points-of-view in order to decide on a course of action. Other research and analysis conducted by Clabby Analytics can be found at: www.ClabbyAnalytics.com.