Scalable Performance: No-Frills Clusters for Power and Efficiency

Michael Feldman

Analyst paper
March 2015

EXECUTIVE SUMMARY

Scalable applications have become so prevalent in data centers that they are dictating new server designs. High performance computing (HPC), data analytics, web serving and internet services, while diverse in nature, all demand common features generally absent in legacy enterprise servers. To meet these demands, the most important criteria is ensuring seamless scale-out capabilities. Key server design points include performance, scalability, flexibility, and simplicity.

At the same time, power and other data center costs have become constraining factors for large-scale deployments. For HPC centers, facilities represents approximately 10 percent of the total budget, with power consumption, cooling, and building space together representing over 90 percent of that total[1]. Any infrastructure solutions, therefore, must also encapsulate energy-efficient operation and density.

The NeXtScale System, Lenovo's ultra-dense server line, has been designed specifically with these criteria in mind. The latest version supports the fastest Intel Xeon E5-2600 v3 processors and some of the highest performance memory modules on the market. Configurability is enabled by choice of processor, network type, and internal storage, as well as the addition of separate accelerator and storage nodes. There is also the option of water cooling, which provides one of the most efficient, low-cost solutions in the market. In aggregate, NeXtScale offers a very practical, high-performance solution for large-scale or small-scale deployments.

[1] HPC Budget Allocation Map: Industry Averages, Intersect360 Research, April 2014

TABLE OF CONTENTS

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

HYPER-SCALABILITY IS THE NEW NORMAL

As large-scale data centers are increasingly being used to encompass a growing array of applications, hyperscale computing is becoming the dominant form of information processing. For the purpose of this report, we'll define hyperscale as systems that can offer practical distributed computing solutions across a wide range of system sizes, from a few servers up to large-scale clusters containing thousands of nodes. Think hyper-scalability.

Internet services, web hosting, data analytics, and traditional high performance computing (HPC) represent some of the largest application sets that are being run on hyperscale systems, and together they constitute some of the most dynamic markets for servers. Customers range from small and medium-sized businesses up through the largest Fortune 500 companies. As a result of the diversity of applications and users, hyperscale computing is becoming established across nearly every public and private sector in the economy.

Intersect360 Research projects a 4 percent compound annual grow rate (CAGR) for servers between 2013 and 2018 for High Performance Technical Computing (HPTC) and High Performance Business Computing (HPBC)[2], which together comprise much of the hyperscale server market. Over the last five years, our end user surveys indicated that servers accounted for nearly half (48 percent) of all expenditures in HPTC and HPBC data centers[3], and those expenditures are expected to increase in absolute terms. It should therefore come as no surprise that server makers have been busy designing systems specifically optimized for such environments.

Cost containment is a primary goal of hyperscale computing. That applies to the small business with a four-server cluster doing data mining on its internal database as well as a 5000-node supercomputer performing petascale climate modeling. At both ends of the spectrum, cost considerations dominate, in the first case due to the inherent spending constraints at small businesses, in the second case due to the sheer size of the system.

The hallmark design criteria for all such systems is simplicity. It is employed to achieve cost effectiveness and operational efficiency, two characteristics that are paramount when either CAPEX or OPEX are driving factors. Simplicity is often the most difficult element in design, since it requires not just the elimination of non-essential features, but the careful assembly of what remains to create something cohesive.

[2] HPC Market Model and Forecast 2013 to 2018 Snapshot Analysis, Intersect360 Research, June 2014

[3] HPC Budget Allocation Map Industry Averages Snapshot Analysis, Intersect360 Research, August 2014

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

The basic approach is to build the servers from industry-standard parts, eliminate redundancies and other components that are not critical, and use integration or shared infrastructure to simplify design and assembly. For example, in most cases, redundant power supplies and fans are done away with, and those that remain tend to be shared across multiple server nodes.

Fault tolerance for these minimalist designs is achieved with the inherent redundancy of large numbers of servers in concert with software that is able to dynamically reallocate compute (or storage) to working hardware as components fail. This software-supported failover capability is at the heart of fault tolerance for all hyperscale environments.

All of these design elements lower costs (both up-front and operational), but they also maximize another important criteria: density. To underscore the prevalence of this characteristic, hyperscale systems are sometimes referred to simply as density-optimized servers. Typically this means two or more server nodes per 1U or rack space. Since this reduces the amount of data center square footage needed to house a given amount of hardware, it represents yet another cost-saving element of design.

However, the feature of density has created an additional challenge: heat dissipation. As processors and other components are more closely packed, heat generation becomes a problem – both for the longevity of those components and their ability to function optimally. In addition, conventional cooling using fans is expensive, since air is a relatively poor conductor of heat. As a result, there has been growing interest to incorporate water cooling into hyperscale servers. This is yet another challenge, given the inherent costs and complexities of integrating plumbing into electronic equipment.

Although simplicity is the driving force behind hyperscale servers, the diversity of scale-out applications implies a range of designs. Some hyperscale applications are compute-intensive, limited by processor performance or memory speeds and capacities, while others are data-intensive, where the bottleneck lies in the memory or I/O subsystems. This diversity has led to different design points targeted at specific classes of applications, but some vendors have built more general-purpose platforms that can be configured for a range of workloads. Such an approach allows customers to construct heterogeneous clusters of servers that can handle a variety of workloads, but does so with a platform that is interoperable.

NEXTSCALE: SIMPLICITY WITH FLEXIBILITY

One such platform is the NeXtScale System, Lenovo's server line for the HPC and hyperscale market. It incorporates configurability in two key dimensions, server type and a

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

choice of components. Specifically, there is the base NeXtScale chassis and compute nodes with the ability to easily and cost-effectively add storage and/or accelerators via NeXtScale's unique Native eXpansion (NeX) architecture. Depending upon the customer's relative needs for compute performance, compute throughput, and storage, various combinations of those servers may be used to populate a rack.

To further refine configurability, within each server type, there is a choice of processor, memory capacities, storage devices, and network type. Also, even though these systems are designed as stripped down servers, there exists the option to add in redundant power supplies and duplicate storage drives.

The NeXtScale nx360 M5 platform includes the latest Intel Xeon E5-2600 v3 ("Haswell EP") processors with double the memory capacity of the prior M4 generation. The system is based on industry standard components and is housed in standard-sized 19-inch racks.

Highly Efficient Water Cooling

Perhaps more significantly, the NeXtScale M5 offers a choice of air cooling or water cool technology (WCT), which makes it suitable for some of the most power-constrained data centers and for situations where electricity costs are paramount. Power costs have become especially relevant across the industry, given that power-demanding infrastructure is growing at the same time that electricity costs are increasing (at a rate that generally follows the overall inflation rate). In our most recent budget allocation site survey[4], it was revealed that power consumption comprises the highest percentage of facilities spending, with a 43 percent share.

The nx360 M5 WCT also allows customers to build systems with the highest frequency and core count E5-2600 v3 processors, which consume as much as 165 watts of power, as well as enabling performance-demanding customers to run the processors continuously with turbo boost enabled. Since that feature provides what amounts to a controlled overclocking, it allows users to speed-up a range of HPC applications that are compute limited.

Performance aside, Lenovo projects that the water-cooled option will deliver about 40 percent better energy efficiency in the datacenter than its air-cooled counterpart. That adds up quickly in cost savings, especially for larger scale deployments. While there are some additional upfront costs involved with setting up a water-cooled design in a facility, there can be significant savings in operational expenses that can provide lower TCO in most geographies.

[4] HPC Budget Allocation Map: Industry Averages Snapshot Analysis, Intersect360 Research, August 2013

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

A key advantage that Lenovo has exploited is the elimination of internal fans as a result of the more efficient water cooling which also drives down hardware costs, reduces noise, and increases reliability. In addition, the M5's direct water cool design is much more efficient than convection-based water cool designs available elsewhere. The water is routed through copper cooling tubes directly to each of the processors for maximum cooling efficiency, and it draws heat away from the memory and IO via copper heat pipes. Also since the system is designed to work with inlet temperatures of up to 45C (113F), no external water chillers are required. All of this can be accomplished with the same standard racks used for the air-cooled version.

More Performance, More Options

In addition to the water cooling capability, almost all the other upgrades to M5 compute server are designed to maximize compute density and performance. For example, with the 18-core E5-2600 v3 processors, the M5 will deliver up to 49 percent more compute performance than E5-2600 v2-based M4 servers[5]; and with 16 DIMM slots, each dual-socket node can house up to 512 GB of DDR4 memory at 2133 MHz to feed all the extra cores. External storage has been doubled as well, with support for four 2.5-inch drives. Configurability is also enhanced with a new RAID slot in the rear to free up the PCI in front for other devices.

If PCI is not desired, two front hot-swappable disk drives can take its place. For storage at the rear, there's a choice of a 3.5 inch hard drive, two 2.5-inch hard drives, or up to four 1.8-inch SSDs. Likewise for network configurability: an x16 mezzanine card is provided, which will support the latest network adapters for both FDR and EDR InfiniBand, as well as 10 Gigabit Ethernet.

Not all users will want to max out on capacity and performance. Nodes can be configured with four-core processors and as little memory and storage as necessary. So for applications that are primarily throughput oriented, such as web services, users can opt for low-cost E5-2600 v3 SKUs and 4 and 8 GB DIMMs.

For the compute server, density is achieved by housing 12 dual-socket nodes per 6U chassis. That translates to 84 nodes per rack. For systems outfitted with the highest core-count E5-2600 v3 processors (18-core), each rack can contain up to 3,024 cores.

[5] Based on SPECint_rate_base2006 benchmarks for NeXtScale nx360 M5, E5-2699 v3 compared to NeXtScale nx360 M4, E5-2697 v2, www.spec.org

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

Storage, Acceleration, and Software

The NeXtScale nx360 M5 compute node also supports connection to Native Expansion (NeX) trays, including Storage NeX and PCI NeX. Each storage node (Storage NeX) can house up to eight 3.5-inch hard disk drives, while the acceleration node (PCI NeX) can be outfitted with up to two high-power (up to 300 watt) NVIDIA GPUs or Intel Xeon Phi devices. Again, since compute, storage, and acceleration can be mixed and matched within a rack or across racks, systems can be built according to application storage and co-processing needs. No custom components are required to add storage and acceleration and there are no mid-plane dependencies.

An additional element to NeXtScale's flexibility is the variety of system software that is supported, much which is derived from open source and open standards. This includes software such as Linux (operating system), OpenStack (cloud computing), Puppet (configuration management), Ganglia (system monitoring), and xCAT (system management). Lenovo also offers robust commercial software such as IBM General Parallel File System (data management), and IBM Platform Computing (workload and resource management) software. In fact, IBM Platform LSF has unique energy-aware scheduling features that helps improve energy-efficiency and drives down the cost of computing.

NEXTSCALE: MARKET SUITABILITY

As a result of its simplicity, scalability, and ability to be configured for different application profiles, the NeXtScale M5 is well-suited across the principle hyperscale use cases: HPC, internet services, web hosting, and data analytics. Those span public and private sectors for technical computing (traditional HPC such as physics simulations, oil & gas exploration, bioscience modeling, defense/intelligence applications, and product design) and business computing (commercial analytics, financial trading, risk management, fraud detection, digital content creation, online gaming, and the whole spectrum of services and applications delivered via large-scale internet environments). They span traditional cluster set-ups, but especially on the business side, are increasingly represented by clouds, public and private.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

For **high performance technical computing**, NeXtScale supports a number of high-end components that map well to this performance-demanding domain, even up to the level of petascale systems. Typical configurations would incorporate:

- Two high performing Intel Xeon processors, up to 36 cores per node
- The fastest memory, up to 512 GB per node
- FDR/EDR InfiniBand
- Up to two GPUs or Xeon Phi devices for additional application speedup
- Air or water cooling technology

For **internet data centers** targeting loosely coupled, throughput-oriented applications, NeXtScale systems would be more likely to employ these less expensive components:

- Lower bin, more power-constrained Intel processors, in either one-socket or two-socket configurations
- Lower capacity memory – 4 GB and 8 GB DIMMs
- Onboard NIC for Gigabit Ethernet
- Two swappable drives
- A free PCI slot for other uses

For **large-scale enterprise applications**, such as business analytics, which are increasingly hosted in public or private clouds, needs can vary, but represent something of a mix of HPC and throughput computing requirements. A NeXtScale system built for this application set and environment would likely be configured with:

- Two high-performing or mid-performing Xeon processors
- At least 256 GB per node
- Ethernet or 10 Gigabit Ethernet
- A variety of storage configurations

NeXtScale can be delivered as individual server(s) or as an Intelligent Cluster, an integrated system offered by Lenovo where they build, integrate, and test the entire system at the rack level before delivery. Essentially they shrink-wrap the cluster so that it can be brought into production quickly at the customer site and supported as a single system with one part number (versus 1000's of individual parts). This is important across many domains, particularly in situations where additional capacity needs to be added quickly and seamlessly.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

NEXTSCALE: TECHNICAL COMPUTING USE CASE

Notre Dame University deployed a NeXtScale cluster at its Center for Research Computing (CRC) to support multiphysics simulations work and other scientific research at the university[6]. The cluster is made up of 84 NeXtScale nx360 servers, and is being used as an extension of an existing System x iDataPlex cluster housed at the center. According to the CRC director Dr. Jarek Nabrzysk, the system performed flawlessly during their evaluation testing, delivering "ideal" performance results.

The NeXtScale cluster also met Notre Dame's need in price/performance, maintenance ease, and extendibility. Although this particular system is x86 only, the CRE is exploring the idea of using GPUs to speed computations, which would be possible with NeXtScale's accelerator node. Open source compatibility was another key selling point, specifically, the support for xCAT, the center's system management tool.

CONCLUSION

Hyperscale computing infrastructure is different than that of traditional enterprise computing, but is predominate in the fast-growing segments of HPC and web-based computing. With the expansion of those two domains and the advent of scalable big data applications, hyperscale computing has become even more generalized, and that means that specialized designs of once-niche domains are becoming mainstream.

The challenges for server vendors are the conflicting demands of cost, compute efficiency, and application diversity. While the overriding criteria of cost drives design simplicity, density, and energy efficiency for these systems, varying application demands necessitates either multiple design points or a design that can be configured according to application profiles. In NeXtScale, we have the latter.

NeXtScale configurability supports a range of application criteria – compute-intensive, memory-intensive, and I/O-intensive – and does so while retaining cost and operational efficiencies. Systems can be built with processors of different speeds, core counts, and even architecture (multicore processors and manycore accelerators), along with memories of different capacities and networks with different bandwidth and latency characteristics; likewise, for storage capacities and configurations.

Besides applications flexibility, NeXtScale M5 adds a choice of air or water cooling technology that is suitable for computing at scale, while maintaining simplicity. As systems

[6] http://news.lenovo.com/images/20034/systemx_university_of_notre_dame_cs.pdf

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com

get larger and denser, and electricity costs rise, water cooling becomes an ever-more-viable option. With its low-cost and unobtrusive design, NeXtScale has one of the best water cooling solutions in the market.

Finally, Lenovo's deep supply chain and high operational efficiency creates a distinct advantage for the company as it enters the cost-sensitive hyperscale market. When combined with the skills and intellectual property acquired from IBM, along with its successful System x business, Lenovo offers a set of capabilities that places it in a unique position to serve the growing base of HPC and other hyperscale customers.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com|info@Intersect360.com