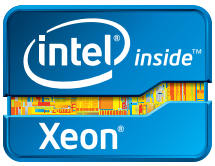
WHITE PAPER
Intel® Ethernet 10 Gigabit
Converged Network Adapters
Arista Networks Switches
Hadoop*

Hadoop* Clusters Built on 10 Gigabit Ethernet

A Common Man's Guide to Building a Balanced System on a Cost-Effective 10GBASE-T Foundation



Hadoop* is increasingly popular for processing big data. Dramatic improvements in mainstream compute and storage resources help make Hadoop clusters viable for most organizations. But to provide a balanced system, those building blocks must be complemented by 10 Gigabit Ethernet (10GbE), rather than legacy Gigabit Ethernet (GbE) networking. This study found success by building on a 10GBASE-T foundation that combines Arista switches, Intel® Ethernet 10 Gigabit Converged Network Adapters, and Intel® Xeon® processor-based servers.

IT organizations across all industries increasingly must handle, store, and analyze very large data stores, with operations on petabytes of data becoming relatively commonplace. Moreover, a recent study by IDC indicates that those data stores are growing at a compounded annual rate of 40 percent.[1]

Manipulating these large-scale bodies of data is increasingly complex because much of that data is unstructured, meaning it does not fit a predefined model (or relational database tables). Such data is also typically text-heavy, such as that generated by social media or search functionality, and conventional relational databases are poorly suited to its management and analysis.

The traditional tools for handling large bodies of unstructured data were proprietary, expensive, and required specialized expertise, putting them out of reach for most organizations. Today, commonly available solutions such as Hadoop* clusters provide excellent support for manipulating big data on commercial off-the shelf (COTS) servers. While building a Hadoop environment can be intimidating, it doesn't have to be.

If you or someone you love has been tasked with building a Hadoop cluster, take heart. This paper aims to provide the background you need to take the first steps painlessly.

Reducing the Pain and Cost of Handling Big Data

It's common for organizations to spend the money to collect large data sets then not to take the next step of extracting value from them. That disconnect comes from having to deal with the traditional cost and complexity of tools to manage and analyze unstructured data, as well as simply not knowing where to start. In reality, powerful x86 servers and Hadoop address the issues of cost and complexity, aided by the increasing cost effectiveness of 10GbE. This paper will help you get started by sketching plans of how to create a practical 10GBASE-T Hadoop cluster as a foundation you can build on.

The goal of this approach is to make the initial build as simple, affordable, and flexible as possible, while also providing the basis to justify and plot a course for future investment. Hadoop is an excellent choice of technology to transform the liability associated with a massive data store into an asset that can help drive your organization's success. It is available in both open source and commercial packages.

Introducing Hadoop

Hadoop is an Apache software framework that analyzes petabytes of unstructured data and transforms it into a more manageable form for applications to work on. Based on Google's MapReduce and distributed file system work, Hadoop is designed specifically to be deployed on commonly available, general-purpose network and server hardware.

Table of Contents

TAMING BIG DATA AT *THE NEW YORK TIMES*

The course of human innovation is powered by tinkering, as illustrated by this example of Derek Gottfrid using Hadoop at The New York Times several years ago.[2] "The Old Gray Lady," as the newspaper is respectfully (if quirkily) known, decided to make all 11 million of its public domain articles from 1851 to 1980 available as PDF files. Doing so required several TIFF images for each article to be scaled and stitched together, and Gottfrid set out to automate the process using code he built to run on Amazon Simple Storage Service (S3) and Elastic Compute Cloud (EC2).

Processing the large data set of TIFF images (4 terabytes) in a reasonable time frame required being able to run in parallel across multiple machines. To solve that problem, Gottfrid built his first Hadoop* cluster, which became the basis (to paraphrase his words) of transforming an adventure into a success. Using 100 Hadoop nodes on Amazon's EC2 cloud, the job completed in 24 hours at a cost of US$240 (not including network bandwidth costs).[3]

Hadoop is optimized for a specific class of tasks, such as indexing and sorting large data sets, data mining, log analytics, and image manipulation. It is not designed for real-time processing or process-intensive tasks that don't involve large amounts of data. Architecturally, Hadoop has two main parts:

- **Hadoop Distributed File System (HDFS)** uses a write-once, read-many model that breaks data into blocks that it spreads across many nodes for fault tolerance and high performance. In addition to the massive aggregate I/O across many nodes, performance is aided by large block sizes—about 128 MB, as compared to the more typical size in Linux* implementations of perhaps 4 KB.
- **MapReduce Engine** accepts jobs from applications through its *JobTracker* node, which divides the work into smaller tasks that it assigns to *TaskTracker* nodes. If it is connected to a network-topology-aware switching infrastructure, the JobTracker node boosts performance by intelligently keeping the work being done on a piece of data as close as possible to that data (within the same node if possible).

The Need for a Balanced System

Hadoop is designed and optimized for commonly available hardware. The pace of server innovation has continued unabated for many years, and mainstream systems now deliver massive processing power. To keep pace with that capability, it is vital to deploy Hadoop in the environment it was designed for, one that is balanced between compute, storage, and networking.

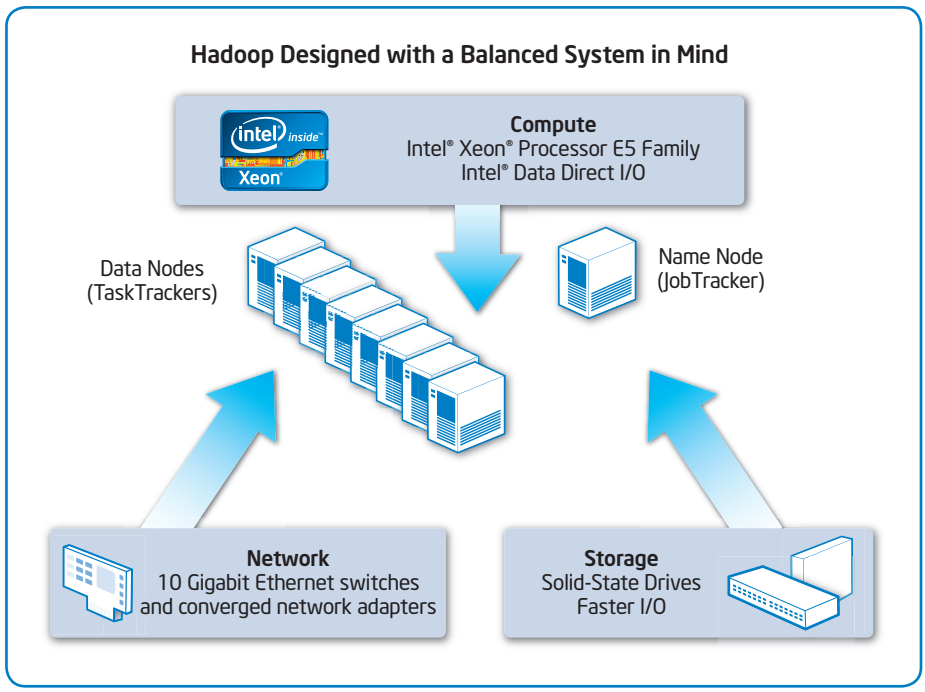


Figure 1. A question of balance.

Figure 1 illustrates key factors associated with creating and maintaining that balance.

- **Processor performance keeps climbing**. The Intel® Xeon® processor E5-1600/2600 product families deliver up to 80 percent higher performance than the previous generation.[4,5]
- **Storage speeds keep increasing, and latency and costs keep decreasing**. Performance improvements in storage I/O are ongoing, and the costs to gain the performance advantages of solid-state drives (SSDs) keep dropping. In 2011, for example, the price per GB of SSD storage fell to US$2.42 (versus US$40 in 2007), and some are projecting prices to fall below US$1.00 in 2012.[6]
- **Internode communications are driving up networking speeds**. The large-scale I/O requirements of a distributed server and storage architecture such as Hadoop demand high throughput. 10GbE switching is an excellent way to meet those needs cost effectively.

Server platform enhancements are not limited to processor performance. A particularly relevant aspect of the Intel Xeon processor E5 family architecture in the context of Hadoop, Intel® Data Direct I/O Technology (Intel® DDIO)[7] makes a substantial contribution to the platform's overall I/O performance. Intel DDIO intelligently directs I/O packets to the processor cache, skipping main system memory. This action can dramatically reduce latency and improve overall system bandwidth and power utilization by eliminating unneeded trips to memory.

In the area of networking for this balanced system, the performance of Gigabit Ethernet (GbE) implementations for Hadoop has been a major limiting factor to overall performance. Using the large block size means that, for example, when a packet is dropped and retransmitted, the system needs to handle a large piece of data, which strains network bandwidth in a GbE environment. 10 Gigabit Ethernet (10GbE) networking proves its value in Hadoop clusters through high observed levels of network utilization, demonstrating the benefit of the higher bandwidth.

Realizing the Full Benefits of 10 Gigabit Ethernet

A Hadoop infrastructure based on well-matched sets of building blocks includes 10GbE networking to support high-performance compute and storage resources. This requirement is occurring at the same time as other technical advances and market forces that are also helping to advance 10GbE. Perhaps the most significant factor has been the rise of large-scale virtualization; large numbers of virtual machines per physical host create demand for high network throughput, and 10GbE is a natural choice. Likewise, consolidating traffic—including multiple GbE interfaces, NFS, iSCSI, etc.—onto 10GbE switched networks drives the need for higher bandwidth.

As adoption in mainstream environments continues to accelerate, greater availability of 10GbE equipment helps drive down the prices of 10GbE network adapters and switches. That lower pricing helps drive further adoption, creating a virtuous cycle of widespread use and cost effectiveness. The cost per gigabit of bandwidth is now substantially lower with 10GbE than GbE, making 10GbE the logical choice for most server implementations.

INTEL® PLATFORM I/O ENHANCEMENTS

Building your Hadoop* servers on an all-Intel foundation gives you the benefit of synergies from cohesive engineering throughout the platform based on the Intel® Xeon® processor E5 family[5] and Intel® Ethernet Converged Network Adapters and controllers.

- **Intel® Integrated I/O** moves the I/O subsystem onto the processor, reducing server platform latency up to 30 percent[8] while increasing bandwidth by up to 2x[9] with support for the PCI Express* 3.0 specification.
- **Intel® Data Direct I/O Technology (Intel® DDIO)** allows I/O traffic to be directed straight to the processor cache (instead of through memory), reducing latency as well as reducing power consumption by allowing memory to stay in a lower power state longer.
- **Intel® Ethernet** products deliver innovation and trusted performance with trusted, reliable hardware and software drivers that draw from three decades of Ethernet leadership.

The Intel® Ethernet Controller X540 is Intel's latest 10 Gigabit Ethernet controller and the industry's first fully integrated 10GBASE-T controller, combining MAC and PHY into a single chip. Intel designed this controller for low-cost, low-power 10GBASE-T LAN on Motherboard (LOM) and converged network adapter (CNA) implementations. The Intel Ethernet Controller X540 includes advanced I/O virtualization (virtual machine device queues and support for PCI-SIG Single Root I/O Virtualization), storage over Ethernet (including NFS, iSCSI, and Fibre Channel over Ethernet), and support for I/O enhancements on the new Intel Xeon processor E5 family.

WHY 10GBASE-T?

The industry transition to 10 Gigabit Ethernet (10GbE) is spurred on by 10GBASE-T options such as the Arista 7050T switch and the Intel® Ethernet Controller X540, which deliver the following benefits:

- **Reduced deployment costs** to bring 10GbE to the broad market through integrated 10GbE, lower cable costs, and flexible cable lengths.
- **Simplified transition to 10GbE** from backward compatibility with the switching and cabling infrastructures of existing GbE networks.
- **Reduced complexity** through industry-standard twisted-pair copper cabling and support for advanced I/O virtualization and unified networking.

Ethernet switching innovation and product maturation of 10GBASE-T switching ports also play significant roles in reducing the deployment costs of Hadoop clusters, especially when compared to the cost of optical cabling and transceivers. Moreover, 10GBASE-T switching is compatible with many existing category 5 and 6 cabling racks and cabling trays, which helps avoid the need for a complete "rip and replace" during adoption.

10GbE switching also offers advancements over GbE switching beyond bandwidth capacity, in the areas of lower switching latency, better traffic balancing and failover, and greater scalability as clusters grow. The software capabilities of the switch also directly benefit Hadoop deployments, including rack awareness that enables the Hadoop JobTracker node to optimize placement of tasks, as described earlier in this paper.

Another instance of 10GbE innovation brought on by its widespread adoption, 10GBASE-T connectivity is now integrated onto mainstream servers through LAN on Motherboard (LOM) designs. By taking advantage of this LOM approach, end customers can benefit from 10GbE without additional expense beyond the cost of the core server platform. These solutions are powered by the Intel® Ethernet Controller X540, the industry's first fully integrated (MAC and PHY) 10GBASE-T controller designed for LOM and converged network adapters.

HIGH-BANDWIDTH, LOW-LATENCY, TOPOLOGY-AWARE SWITCHES FROM ARISTA NETWORKS

Arista recently announced and began shipping the Arista 7050T switch used in the proof of concept described in this paper—a 1RU wire-speed, low-power-consumption 10GBASE-T switching platform. It can be used both as a top-of-rack switch within Hadoop* clusters and as the spine switch (with layer-3 switching) when there are multiple racks.

As a spine switch, the 7050T offers the benefits of layer-3 switching (routing) to complement Hadoop's layer-3 file system, for better scalability when moving traffic between racks. Because it is topology-aware, Hadoop takes advantage of the topology intelligence provided by the 7050T to place compute tasks close to the relevant storage, enhancing efficiency. The 7050T switch also supports deep buffering, which can significantly increase performance as opposed to dropping and retransmitting packets when Hadoop over-subscribes links.

The 7050T switching platform provides deep packet buffering and open interfaces for rack-awareness integration, making it the ideal workhorse for scaling from small- to medium-size Hadoop infrastructures. The 7050T switching platform is fully compatible with the Intel® Ethernet Controller X540.

Stefanie, one of our systems engineers, spending some quality time with the cluster.

Assembling a Hadoop Cluster

In addition to demonstrating performance and scalability, this proof of concept (PoC) outlines a relatively simple approach to building a Hadoop cluster on 10GbE. While there is plenty of room for fine-tuning the cluster once it is up and running, this discussion demonstrates how to hit the ground running alongside a solid Hadoop environment that can grow with you.

Recipe for a Proof-of-Concept Hadoop Environment

Our Hadoop cluster design demonstrates the value of 10GbE networking in comparison to GbE. It is modest in scale (10 nodes), and it is based on commonly available components, as shown in Table 1. We used conventional hard drives, for example, instead of SSDs because traditional magnetic drives are in more general use. A single client machine was used to submit jobs to the cluster.

Table 1. Components in the Hadoop* test cluster

HARDWARE CONFIGURATION	
Network Switch	**Arista 7050T 48-port 1 Gbps/10 Gbps switch**
Compute Nodes: ▪ One head node (Name node, JobTracker) ▪ Nine worker nodes (Data nodes, TaskTrackers)	10 SuperMicro SuperServer* 1026T-URF servers (1U, two-socket): ▪ Intel® Xeon® processors 5690 ▪ 48 GB RAM ▪ Five 700 GB SATA hard drives @ 7200 RPM
Network Interface Cards	One Intel® Ethernet 10 Gigabit Converged Network Adapter (10GBASE-T) and one Intel® Ethernet Gigabit Server Adapter (1000BASE-T) per server
SOFTWARE CONFIGURATION	
Operating System	CentOS 6.2
Hadoop* Version	Cloudera distribution CDH3 (0.20.2)
Java* Development Kit (JDK)	Oracle JDK 1.7.0

For this configuration, we have coined the term "head node" to denote the machine that hosts both the Name node and the JobTracker. In a production deployment, it is desirable to have redundancy for the Name node, which is otherwise a potential single point of failure. Name node redundancy can be accomplished either by deploying on two separate machines or by dual-homing two server adapters to the switch. Note that, for our PoC, we did not provide redundancy using either of those approaches. In any case, the other nodes are designed to fail gracefully, so specific configuration for redundancy is not needed.

To help achieve our design goal of simplicity, we didn't make many changes to the default configuration, but some tuning was necessary. We implemented compression to avoid storage bottlenecks, which also had the effect of highlighting the performance of the network subsystem for testing purposes and comparison between GbE and 10GbE connectivity. The cluster used the default 128 MB block size and the default HDFS replication factor of 3, meaning that the Name node replicates incoming data to three Data nodes.

Alternate Cluster Config is a simple file that allows you to make changes to the Hadoop environment by changing parameters. This approach provides a way to experiment with different options and then to return easily to the defaults, which can provide a safe but powerful way of exploring Hadoop's capabilities.

Gauging the Results and Planning Next Steps

With the PoC cluster built, we next turned to testing that would gauge the performance differential between 10GbE and GbE. The first usage scenario we tested was to hypothetically release a spider on the Internet to gather a large amount of unstructured data, such as end-customer comments about a product. Once that data was gathered and resided on a client machine, the next step was a so-called PUT operation. That operation consisted of importing the data from the client machine into our Hadoop cluster's head node, followed by the head node breaking the data into blocks and replicating each one to three separate data nodes.

Test results of the PUT operation with various data set sizes using both GbE and 10GbE are shown in Figure 2. Note that these results focus on network performance and do not include the time spent processing the data with the MapReduce operation. Using 10GbE networking resulted in completion of the PUT operation in just one fifth the time of the same operation using GbE. Also note that for both 10GbE and GbE, the amount of time required to complete the operation scales linearly with the size of the data set, so the time required is roughly 80 percent less with 10GbE than with GbE. Thus, for data sets in the range of multiple terabytes, the differences could amount to several hours of waiting.

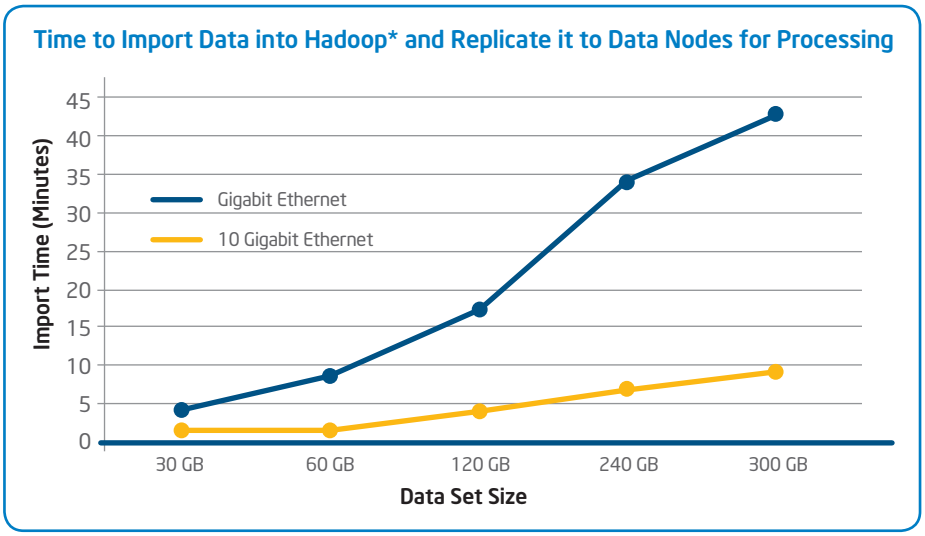


Figure 2. Hadoop* PUT operation completed in 80 percent less time using 10 Gigabit Ethernet, compared to Gigabit Ethernet.

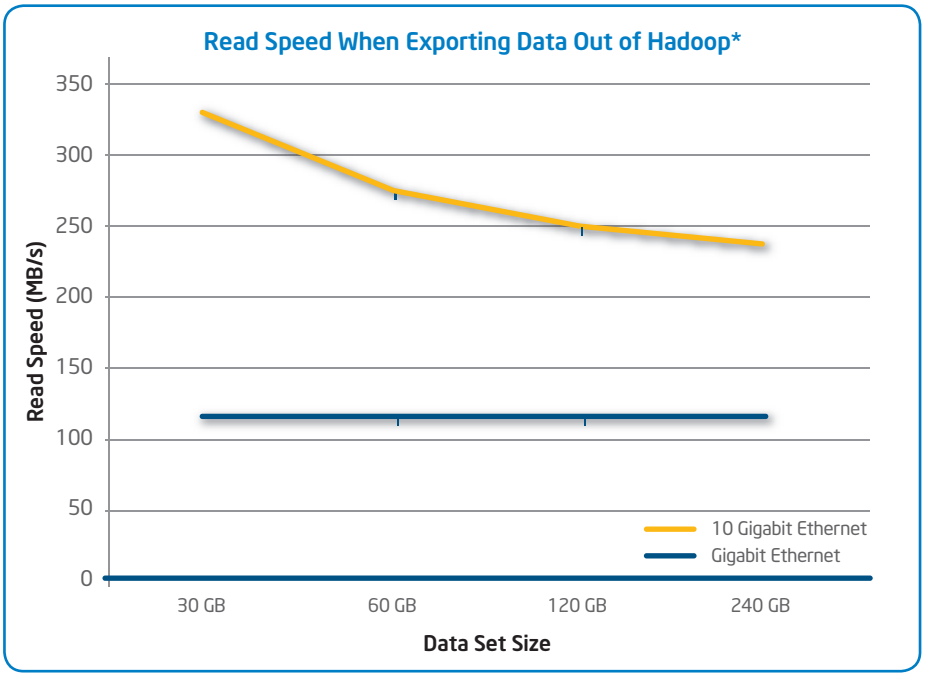


Figure 3. Hadoop* GET operation shows the clear advantage of 10 Gigabit Ethernet over Gigabit Ethernet, limited by a storage subsystem bottleneck.

The companion task to the PUT operation is a GET operation, which is to say, the process of pulling your data out of the Hadoop cluster. For example, to touch back on the example at *The New York Times* mentioned in the sidebar earlier, once the TIFF files have been stitched together to create PDFs, you would need to extract the data back out. Test results for a GET operation are shown in Figure 3.

While with the GET operation, we again see the clear advantage of 10GbE plumbing in the Hadoop cluster, the advantage starts to fall off a bit as the data size increases, due to limitations of the storage subsystem (the data does not include a 300 GB data set because of a "not enough space on local disk" error). We plan to explore in future work the positive effects on this bottleneck from advances in storage technology, such as various types of non-volatile memory as well as tiering strategies that involve putting non-volatile memory in front of large-scale disks. We expect the performance advantages of 10GbE to GET operations to be even clearer as these storage advances continue to develop.

Once you have built a simple test bed as described here, we expect that you will immediately see the advantages of 10GbE networking for Hadoop clusters, in terms of scalability and the ability to keep up with your workloads. You will also soon enough be looking for ways to tinker with your cluster, in search of the best results possible. We wouldn't have it any other way, and we recommend that you have a look at the Intel white paper, "Optimizing Hadoop Deployments"[10] as part of your journey. Enjoy.

Conclusion

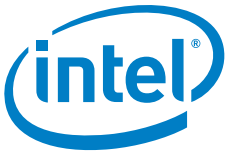
Building a simple Hadoop cluster is within the realm of everyday IT. To be sure, there are many details around fine-tuning the environment, but the foundation is straightforward, as described in this paper. Building an environment that is balanced between compute, storage, and network resources is fundamental to the success of the project. 10GbE has obvious advantages in this regard, as shown by the 80-percent time savings in importing data into your cluster using 10GbE compared to GbE.

Intel® Xeon processor-based servers, Intel® Ethernet 10GbE Converged Network Adapters, and Arista switches are the key hardware elements for building cost-effective Hadoop clusters. New 10GBASE-T products, including servers with LOM connections based on the Intel Ethernet Controller X540 and 10GBASE-T switches from Arista, continue to make 10GbE more cost effective. The latest Intel processors are designed for 10GbE speeds and beyond. Hadoop's linear scalability allows for your environment to grow in tandem with the size of your data sets, turning them decisively from liabilities into assets.

Learn more at
www.intel.com/go/ethernet

www.aristanetworks.com

SOLUTION PROVIDED BY:



[1] Figures reported by IDC, 2012.

[2] http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/.

[3] www.slideshare.net/dgottfrid/hadoop-world-oct-2009.

[4] Performance comparison using geometric mean of SPECint*_rate_base2006, SPECfp*_rate_base2006, STREAM*_MP Triad, and Linpack* benchmark results. Baseline geometric mean score of 166.75 on prior generation 2S Intel® Xeon® processor X5690 platform based on best published SPECrate* scores to www.spec.org and best Intel internal measurements on STREAM_MP Triad and Linpack as of 5 December 2011. New geometric mean score of 306.74 based on Intel internal measured estimates using an Intel® Rose City platform with two Intel® Xeon® processor E5-2690, Turbo and EIST Enabled, with Hyper-Threading, 128 GB RAM, Red Hat Enterprise Linux* Server 6.1 beta for x86_6, Intel® Compiler 12.1, THP disabled for SPECfp_rate_base2006 and enabled for SPECint_rate_base2006.

[5] Note that the testing reported on in this paper was conducted using the Intel® Xeon® processor 5600 series, the predecessor of the Intel® Xeon® processor E5 family.

[6] http://royal.pingdom.com/2011/12/19/would-you-pay-7260-for-a-3-tb-drive-charting-hdd-and-ssd-prices-over-time/.

[7] www.intel.com/content/www/us/en/io/direct-data-i-o.html.

[8] Intel measurements of average time for an I/O device read to local system memory under idle conditions. Improvement compares Intel® Xeon® processor E5-2600 product family (230 ns) versus Intel® Xeon® processor 5500 series (340 ns). Baseline Configuration: Green City system with two Intel® Xeon® processor E5520 (2.26 GHz, 4C), 12 GB memory @ 1333, C-States Disabled, Turbo Disabled, SMT Disabled, Rubicon* PCIe* 2.0 x8. New configuration: Meridian system with two Intel® Xeon® processor E5-2665 (C0 stepping, 2.4 GHz, 8C), 32 GB memory @1600 MHz, C-States Enabled, Turbo Enabled. The measurements were taken with a LeCroy* PCIe protocol analyzer using Intel internal Rubicon (PCIe 2.0) and Florin (PCIe 3.0) test cards running under Windows* 2008 R2 with SP1.

[9] 8 GT/s and 128b/130b encoding in PCIe* 3.0 specification enables double the interconnect bandwidth over the PCIe 2.0 specification. Source: www.pcisig.com/news_room/November_18_2010_Press_Release.

[10] software.intel.com/file/31124.

*Other names and brands may be claimed as the property of others. Printed in USA 0412/ME/MESH/PDF Please Recycle 327302-001US