WHITE PAPER

Increased Hadoop™ Optimization Using Intel®-based SGI® Rackable™ Solutions

Overview

Apache™ Hadoop™ is a software framework that enables applications to be divided into many small tasks, each of which may be executed or re-executed on any node in a cluster. This allows applications to work in a 'Big Data' environment across clusters of many compute nodes. The term "Big" in Big Data refers to unprecedented quantities of information – terabytes, petabytes and more - of new and legacy data generated by today's fast-moving businesses and technologies. In many instances, data collected over the course of days or weeks exceeds the entire corpus of legacy data in a given domain. Examples abound in social media, federal, retail and financial services, and also in scientific disciplines such as genetics, astronomy and climate science.

Big Data is not just defined by the sheer volume of information, but also by the rate of growth of that data and how the IT industry and its customers are meeting the challenge of managing Big Data. While a number of technologies fall under the Big Data label, Hadoop is the Big Data mascot for unstructured data which can include raw text or binary, web pages, e-mail, documents, sensor data, click streams, and more. The impetus for Hadoop adoption is greatest when projects combine "Big Analytics" (comprehensive analysis of complex data) the massive, unstructured data sets typical in Big Data. Despite fast-growing deployment, Hadoop is still time-consuming to set up, deploy and use; building and running Hadoop jobs and queries is non-trivial for developers and analysts. They need deep understanding of Hadoop particulars, including but not limited to cluster sizing and structure, and MapReduce performance. Additionally, executing queries and analyzing results with Hadoop does not leverage existing Business Intelligence skills and tools, leading to the rise of a disruptive technology comprised of new tools to analyze the massive unstructured data and glean insight from it.

This paper describes sizing methodologies of an SGI® Rackable™ cluster for Hadoop; shares results from standard benchmarks executed on a cluster with SGI Rackable half-depth C2005-TY6 nodes using Cloudera distribution Apache Hadoop; explains a price/performance strategy for making optimum decisions; and, finally, describes the value proposition of a best-in-class SGI Rackable Hadoop solution.

TABLE OF CONTENTS

1.0 Big Data Challenges and Hadoop 3

2.0 Apache Hadoop & Associated Components 5
2.1 Apache Hadoop-Cloudera and Cloudera® products 5
2.2 Criteria for Deploying Hadoop 6

3.0 Sizing a Hadoop Cluster for optimal performance 6

4.0 Hadoop Standard Benchmark Performance 7
4.1 Benchmark Configurations 8
4.2 Benchmark Results 9
4.2.1 Terasort Benchmark 9
4.2.2 TestDFSIO Benchmark 10
4.2.3 WordCount Benchmark 11
4.2.4 Sort Benchmark 11

5.0 Case Study: SSD vs. HDD 12

6.0 Price/Performance 13

7.0 SGI Rackable servers for Hadoop 14

8.0 Summary 15

1.0 Big Data Challenges and Hadoop

A distinguishing feature of Big Data is a mixture of traditionally structured data together with massive amounts of unstructured information. The data can come from legacy databases and data warehouses, from web server logs of e-commerce companies and other high-traffic web sites, and from other machines and sensor nets.

Major challenges in dealing with Big Data are:

- **Volume:** Businesses today are dealing with gigabytes, terabytes to even petabytes of incoming information per day, (Figure 1).
- **Velocity:** The ability to perform analytics on thousands of transactions per second is becoming mission critical.
- **Variety:** Harnessing structured, semi-structured and unstructured information to gain insight has become a key business requirement.
- **Vitality:** Big Data analysis and predictive models need to be agile and adapt to changes as business demands evolve.

The *enablers* of Big Data processing are:

- Low cost hardware – especially CPUs and storage
- Disruptive new technology with NoSQL distant cousins such as Hadoop, CouchDB, and MongoDB, to name a few.
- High competitive pressure to derive meaningful information from massive content

The *solution* to Big Data challenges also requires:

- Empowering business users;
- Overcoming complexity;
- Accelerating time to value.

Apache Hadoop is a powerful open source disruptive technology that addresses the challenges in economics, flexibility and scalability of Big Data. However, with Hadoop, one needs to deal with other challenges like raw technology, complexity, resource requirements, and the lack of packaged applications to keep pace with the rapid adoption by large retail, federal, financial sectors, and social networking sites.

Hadoop forms the infrastructure foundation at leading social media companies like Facebook, LinkedIn and Twitter. It is the fastest growing Big Data technology, with 26% of organizations using it today in data centers and in the Cloud, and an additional 45% seriously considering its deployment.

For more information please review
http://karmasphere.com/Resource-Center/collateral.html?mainid=398;
http://karmasphere.com/Resource-Center/understanding-elements-big-data-wp.html

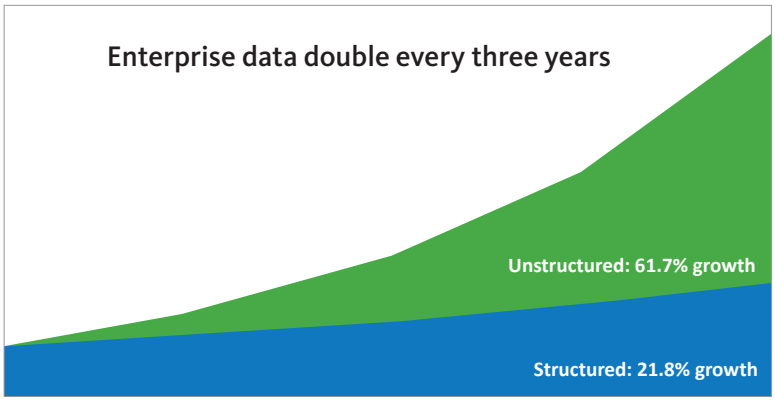

Figure 1: *Growth of structured and unstructured data*

(Sources: "Forrester Research", October 17, 2008 and

"IDC: Unstructured data will become the primary task for storage," October 29, 2008)

Not every massive data store or data-intensive segment is ready to embrace Big Data. However, numerous industries and segments stand out as leaders in their ability to deploy Big Data platforms and analytics (Figure 2).

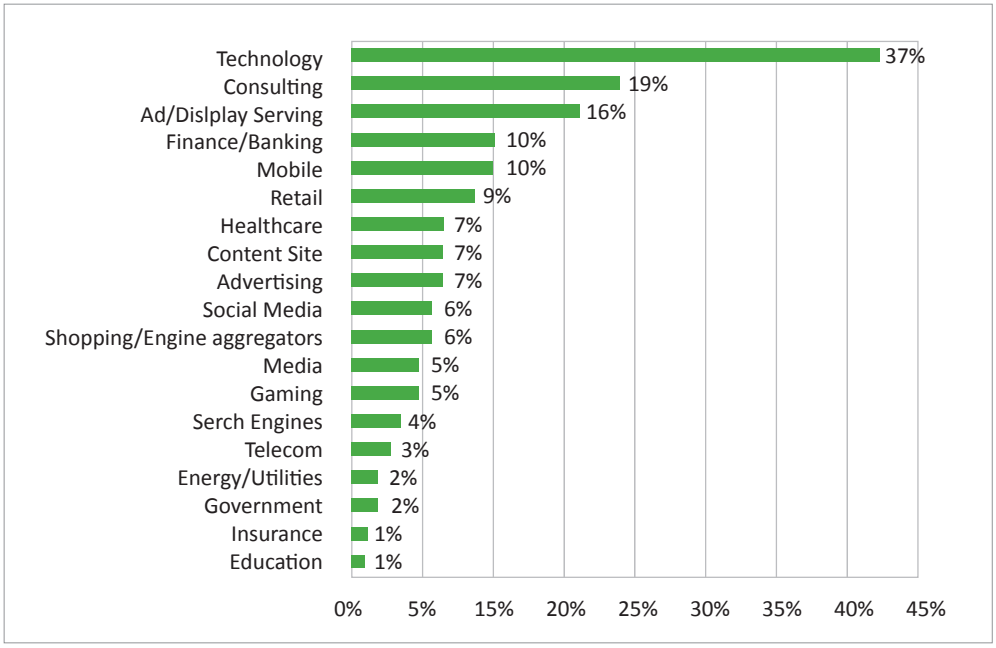

Figure 2: *Industries deploying Big Data*

(Source: Karmasphere, "Understanding the Elements of Big Data: More than a Hadoop Distribution," May 2011)

2.0 Apache Hadoop & Associated Components

- **Apache Hadoop**
  - Is an open source software framework written in Java®, which dramatically simplifies writing distributed, data intensive applications. It provides a distributed file system (HDFS™) for managing data, which is modeled after the Google File System, along with a map/reduce implementation (MapReduce™) that manages distributed computation.
- **HDFS**
  - Is based on Google® GFS (Google File System);
  - Provides redundant storage of massive amounts of data using inexpensive servers;
  - At load time, distributes data across all nodes (Figure 3);
  - Enables efficiency in MapReduce processing.
- **MapReduce**
  - Is a method for distributing a task across multiple nodes (Figure 3);
  - Processes data stored on each node, where possible;
  - Consists of two phases: Map and Reduce;
  - Allows for automatic parallelization and distributed fault-tolerance, usage of status and monitoring tools, and a clean abstraction interface for programmers.
- **HBase™, described as "The Hadoop database"**
  - Is a column-oriented data store;
  - Provides random, real-time read/write access to large amounts of data;
  - Allows one to manage tables consisting of billions of rows, with potentially millions of columns;
  - Runs on top of Hadoop, with data stored in the HDFS.
- **Hive™ is a data warehouse infrastructure built on top of Hadoop**
  - Allows users to query large datasets using a familiar, SQL-like language called HiveQL (Figure 3);
  - Turns HiveQL queries into MapReduce jobs using the Hive interpreter;
- **Pig™ provides a scripting language, known as PigLatin**
  - PigLatin can be used to query data stored in the HDFS;
  - PigLatin scripts are converted to MapReduce jobs by the Pig compiler;
- **Mahout™ provides a suite of machine learning libraries**
  - For data mining atop Hadoop platform.

For more information, please visit the Apache Hadoop website[1].

2.1 Apache Hadoop-Cloudera and Cloudera® products

- **Cloudera distribution Apache Hadoop (CDH)**
  - is an open, easy-to-install package including the Apache Hadoop core repository;
  - includes a stable version of Hadoop, plus critical bug fixes and solid new features from the development projects.
- **Cloudera HUE (Hadoop User Experience**)
  - Is a browser-based tool for cluster administration and job development;
  - Supports managing internal clusters as well as those running on public clouds;
  - Helps reduce development time.

For more information, please refer to the Cloudera web site http://www.cloudera.com/hadoop/.

[1]http://hadoop.apache.org/

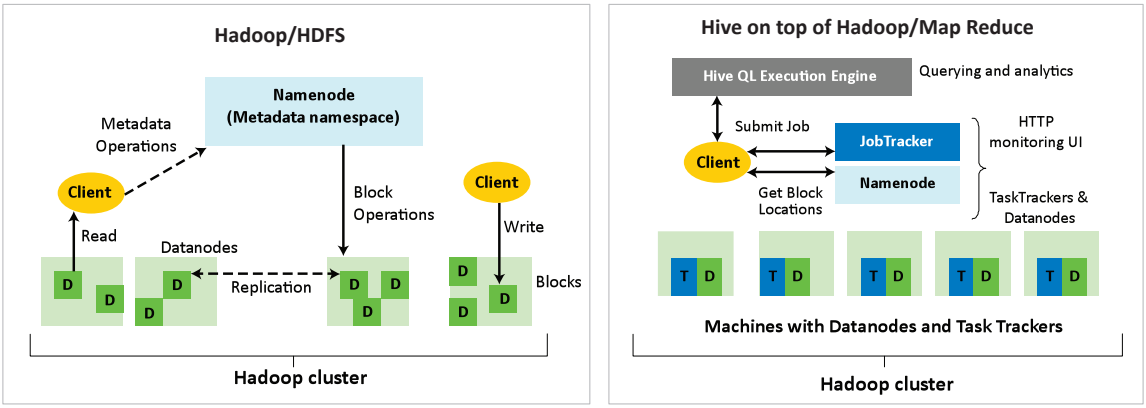

Figure 3: *Hadoop and Hive: Schematics*

2.2 Criteria for Deploying Hadoop

The deployment criteria for Hadoop depend on certain factors where:

- **Data Access is a pain point**
  - While processing data may be fast, read/write from hard disk drives (HDDs) may be slow. Hadoop provides the mechanism of parallel reads/writes of data distributed across many nodes, thus reducing access time.

- **Data Sharing is slow and tricky**
  - Hadoop uses a modern approach of distributing the data, instead of distributing the workload;
  - Data is spread throughout nodes in the cluster;
  - Hadoop allows for computation to be data-local;
  - Instead of bringing data to the processors, Hadoop brings processing to the data.

- **Application semantics are compatible with Hadoop**
  - They are written to scale in a distributed system architecture instead of a SMP architecture;
  - They are compatible with the Java MapReduce framework;
  - They have the ability to process unstructured and semi-structured data.

3.0 Sizing a Hadoop Cluster for optimal performance

If a workload meets the criteria for deploying on Hadoop, then the steps to size a Hadoop cluster can be articulated as follows:

1. **Define the problem**
   - Determine the volume of data that needs to be stored and processed;
   - Determine the rate at which the data is expected to grow;
   - Determine when the cluster needs to grow, and whether there will bea need for additional processing power, storage, tasks or memory;
   - Determine whether the data transfer rate needs to be considered with the growth of the cluster;
   - Determine power efficiency of servers with respect to their performance/watt metrics.

2. **Determine storage space requirements**
   - The number of disks should be based on the amount of useable Hadoop disk space required. Replication factor is 3 by default, so this comes to total raw space / 3;
   - Check on the rate of growth of data and try reducing the requirement of adding new machines every year. For example, depending on the net new data per year, it may be worthwhile using 10x 1Tb hard drives per server than using 4x 1TB drives per server, to accommodate larger amount of new data in the existing cluster.
   - For greater power efficiency and higher ROI over time, choose machines with more capacity. This helps to reduce the frequency of new machines being added.
3. **Design Slave Nodes**
   - A Slave Node runs as a DataNode plus TaskTracker daemons;
   - A minimum of 4 x 1TB hard drives (10-12 1TB drives is recommended), in a JBOD configuration is required. HDFS provides built-in redundancy by replicating blocks across multiple nodes;
   - Size the number of CPU cores, RAM and the drives in a node for an optimally balanced configuration.
4. **Design Master Nodes**
   - A Master Node runs either a NameNode daemon, a Secondary NameNode Daemon, or a JobTracker daemon.
   - It will be important that at least one copy of the NameNode's metadata is stored on a separate machine.
   - Losing the NameNode's metadata would mean all data in HDFS is lost. Future releases of Hadoop, will introduce a Backup NameNode, which should help to mitigate this.
   - The NameNode needs to be Carrier-class, not commodity hardware, and needs to have sufficient RAM (but not over-configured) to drive a large cluster. A good formula for required Namenode memory is to assume that one million data blocks requires 1GB of memory. However, having plenty of extra Namenode memory space is highly recommended, so that the cluster can grow without having to add more memory to the Namenode, requiring a restart. As an example, a NameNode with 32 GB RAM should be good enough to drive a 100 node cluster.
5. **Design network**
   - Use one Gigabit Ethernet switch per rack.
   - For high-bandwidth applications, consider splitting each physical rack into two logical racks.
   - Use network bonding between the data nodes as it helps performance when there is a thick network back plane with high throughput between the data nodes.
   - Use a low-latency 10GigE switch across multiple Racks

4.0 Hadoop Standard Benchmark Performance

This section describes the hardware and software configurations used to run various Hadoop standard benchmarks on an Intel® Xeon® 5600 processor series-based SGI® Rackable™ cluster comprised of half-depth C2005-TY6 servers, followed by benchmark results.
The following benchmarks have been executed:

1. **TestDFSIO**[3], which is a standard benchmark used to perform I/O stress tests for HDFS;
2. **Terasort**[4], which helps derive the sort time for 1TB or any other amount of data in the Hadoop cluster. It is a benchmark that combines testing the HDFS and MapReduce layers of a Hadoop cluster;
3. **WordCount**[5], which reads text files and counts how often words occur. The input and ouput are text files, each line of which contains a word and the count of how often it occurred, separated by a tab;
4. **Sort**[6], uses the MapReduce framework to sort the input directory into the output directory. The inputs and outputs must be Sequence files where the keys and values are BytesWritable.

[2] *http://www.sgi.com/pdfs/4193.pdf*
[3] *http://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/#testdfsio*
[4] *http://sortbenchmark.org/*
[5] *http://wiki.apache.org/hadoop/WordCount*
[6] *http://wiki.apache.org/hadoop/Sort*

4.1 Benchmark Configurations

The benchmark configuration is comprised of:

Hadoop Nodes:

- **Data Nodes:** 20x SGI Rackable C2005-TY6 servers each with
  - 2x Intel® Xeon® E5630 processor series (2.53 GHz, 4 cores), 6x 8GB DIMMS, HT enabled;
  - Disk Subsystem: 4x 1TB SATA HDD 7.2K RPM.

- **NameNode and JobTracker Nodes:** 3x SGI Rackable C2108-TY10 servers each with
  - 2x Intel® Xeon® X5675 processor series (3.06 GHz, 6 cores), 12x 8GB DIMMs, HT enabled;
  - Disk Subsystem: 4x 1TB SATA HDD 7.2K RPM;
  - 2x dual-port 10GigE HBAs.

- OS: RHEL ™ 6 .0 (2.6.32-71.el6.x86_64)
- Cloudera distribution Apache Hadoop 3 update 0 (Hadoop 0.20.2-cdh3u0)

Network:

- 2x SMC 8950EM 48-port GigE switches;
- 1x Brocade® TurboIron® x24 10GigE switch.

Test server:

- 1x SGI Rackable C2005-TY6 server with
  - 2x Intel® Xeon® L5640 processor series (2.26 GHz, 6 cores), 6x 8GB DIMMS, HT enabled;
  - Disk Subsystem: 4x 480 GB OCZ SSD;
  - 1x 1GigE NIC.

The benchmark solution stack with Cloudera distribution Apache Hadoop is shown in Figure 4.

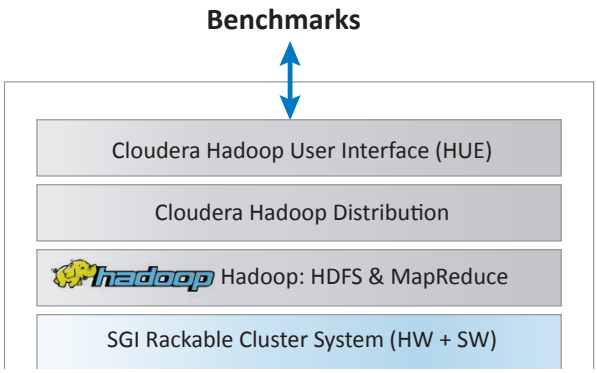

Figure 4: *Benchmarking the Cloudera distribution Apache Hadoop (CDH) solution stack*

4.2 Benchmark Results

4.2.1 Terasort Benchmark

Results as of the date that this paper was published (September, 2011) show that a 20-node cluster comprised of SGI® Rackable™ C2005-TY6 half-depth servers with Intel® Xeon® E5630 processor series (2.53 GHz, 4 cores/8 threads), 48GB memory, and 4x 1TB SATA HDDs running Cloudera distribution Apache Hadoop (CDH3u0) takes only 130 secs to complete a Terasort with a job size of 100GB (Figure 5).

Terasort scales super linearly from one to 20 nodes on an SGI Rackable C2005-TY6 cluster running Cloudera distribution Apache Hadoop (CDH3u0) and is 81% faster than an Oracle Sun X2270 cluster of similar size (Figures 6 and 7).

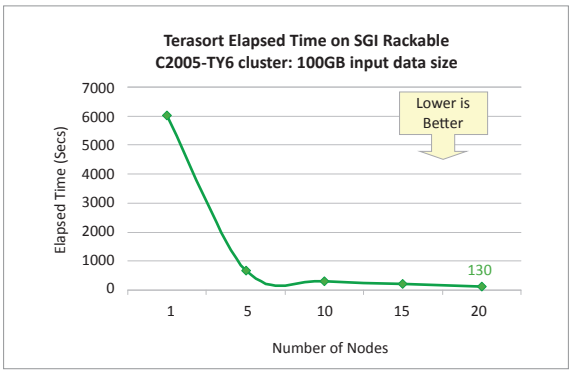

Figure 5: *Terasort Elapsed Time: SGI Rackable C2005-TY6 Hadoop cluster on Intel® Xeon® 5600 processors*

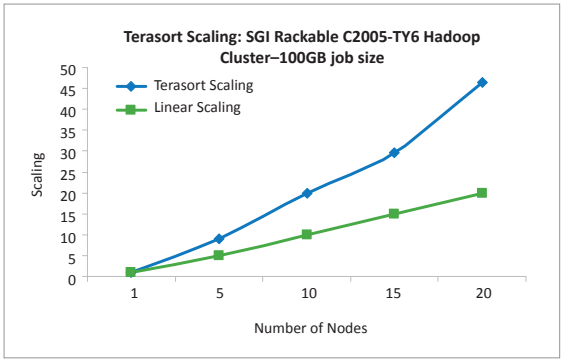

Figure 6: *Terasort Scaling: SGI Rackable C2005-TY6 Hadoop cluster on Intel® Xeon® 5600 processors*

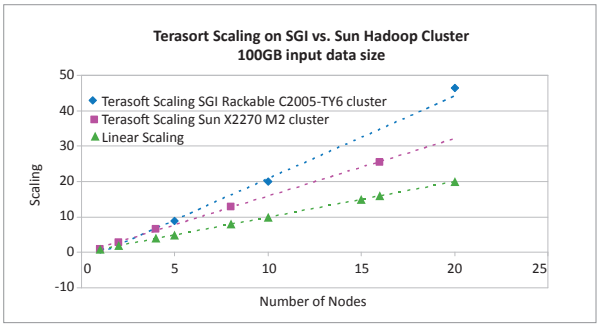

Figure 7: *Terasort Scaling on Hadoop: SGI Rackable C2005-TY6 cluster vs. Oracle Sun X2272 M2 cluster*[7]

[7] *http://sun.systemnews.com/articles/152/1/server/23549*

4.2.2 TestDFSIO Benchmark

Read and write performance with TestDFSIO scale in general, across the nodes and number of files on a SGI Rackable C2005-TY6 Hadoop cluster (Figures 8 and 9). Beyond 1000 files, response time becomes disk-limited. More than 4 spindles per node are required to process many files.

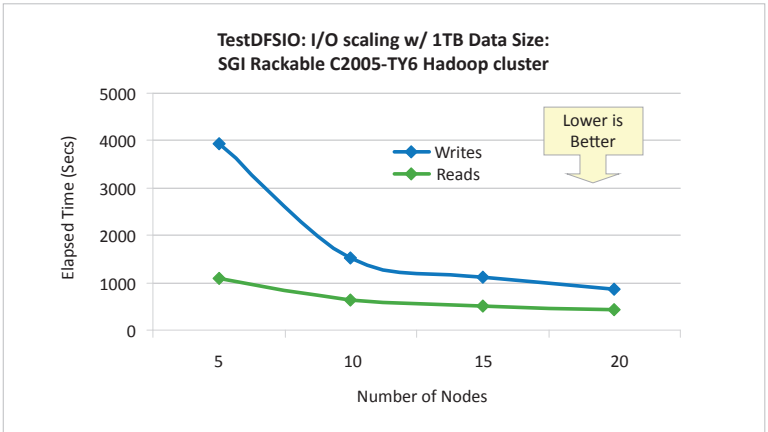

Figure 8: *TestDFSIO Scaling on 20-node SGI Rackable C2005-TY6 Hadoop cluster*

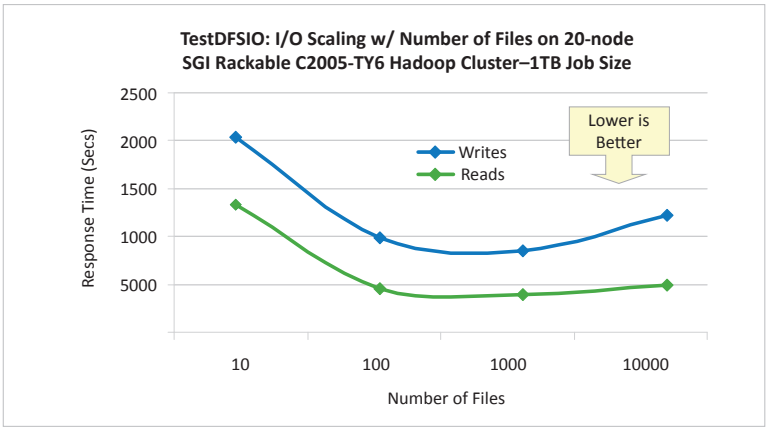

Figure 9: *TestDFSIO Scaling with Number of Files*

4.2.3 WordCount Benchmark

WordCount with a 1 TB job size scales linearly across a 20-node SGI Rackable C2005-TY6 Hadoop cluster, with 4.4 times capability to count occurrence of a word from text files, while using 4 times the number of nodes (Figure 10).

4.2.4 Sort Benchmark

In a 20-node SGI Rackable C2005-TY6 Hadoop cluster, sort time remains stable with sort sizes up to 150GB, while the best time was observed with 15 nodes (Figure 11). Further investigation and tuning is required on 20 nodes.

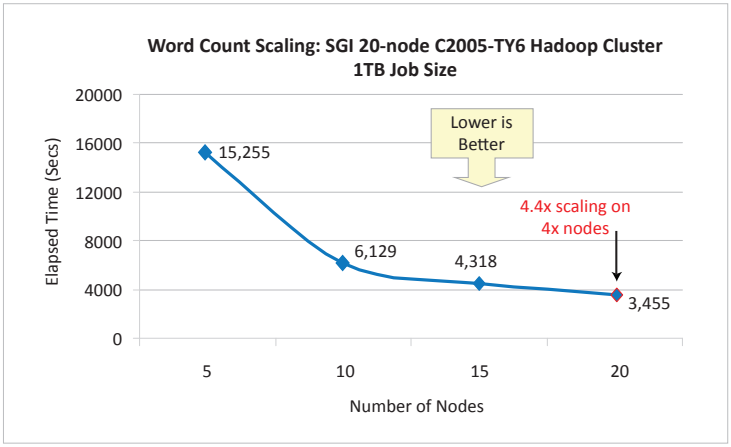

Figure 10: *WordCount Scaling on 20-node SGI Rackable C2005-TY6 Hadoop cluster*

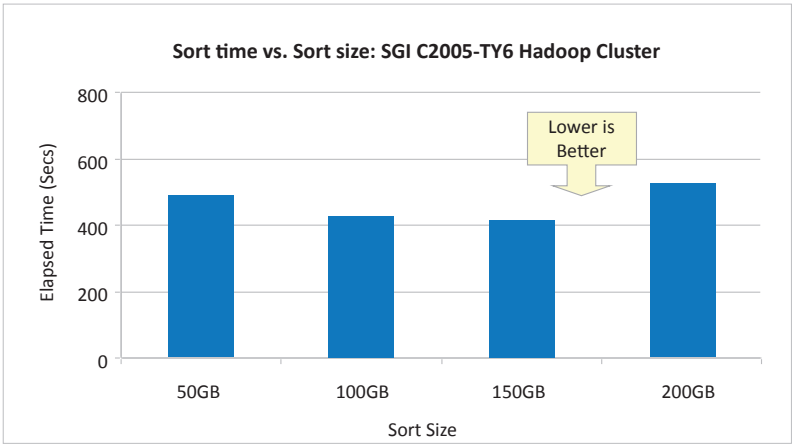

Figure 11: *Sort Time vs. sort sizes on SGI Rackable C2005-TY6 Hadoop cluster*

5.0 Case Study: SSD vs. HDD

This section describes a case study performed on a single SGI Rackable C2005-TY6 node to characterize performance of Hadoop components with HDDs vs SSDs, using Terasort and TestDFSIO. Results show that Terasort on an SGI Rackable C2005-TY6 node running Cloudera distribution Apache Hadoop is 79% faster on SSDs compared to HDDs (Figure 12). Similarly, TestDFSIO is 3x and 2x faster on SSDs vs. HDDs for reads and writes, respectively (Figures 14 and 15).

Terasort throughput projections on a 20-node cluster with nodes using SSDs and HDDs (Figure 13) have been calculated based on the measured data (Figure 12). This helps to estimate price/performance metrics with SSDs vs. HDDs (Section 6) for the cluster.

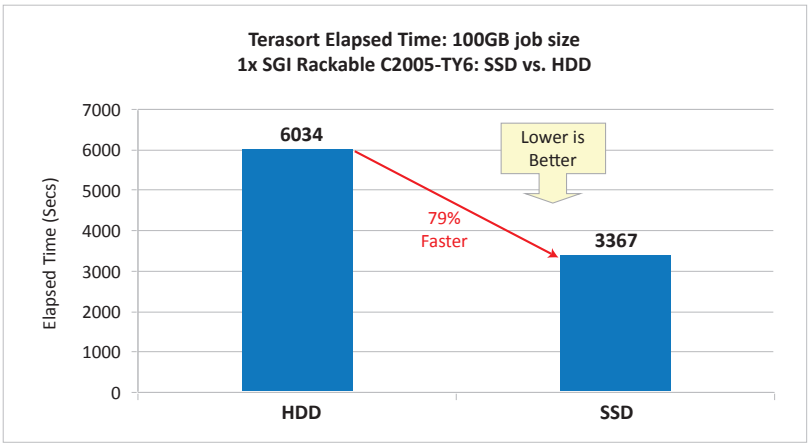

Figure 12: *Terasort: HDD vs. SSD on 1x SGI Rackable C2005-TY6 Hadoop cluster node*

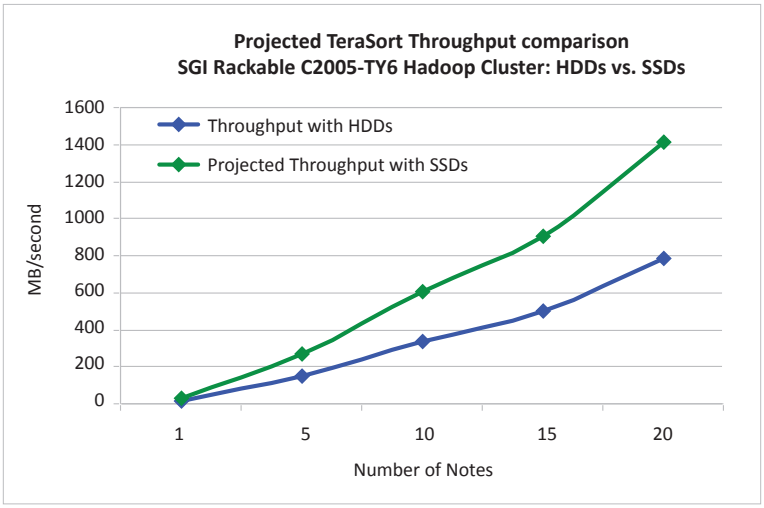

Figure 13: *Terasort Throughput Projections: 20-node SGI Rackable C2005-TY6 Hadoop cluster w/ HDDs vs. SSDs*

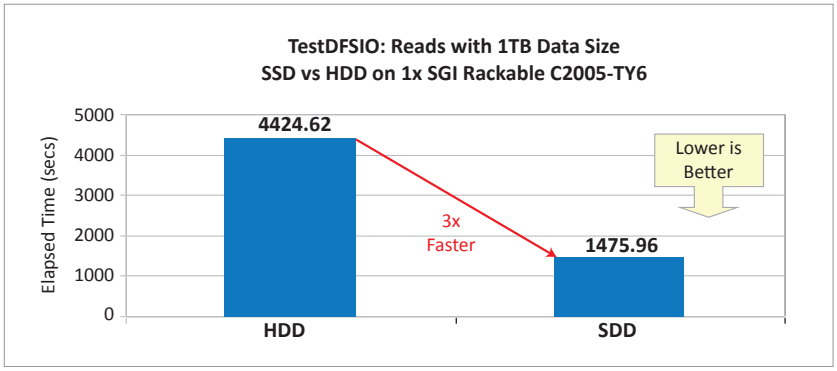

***Figure 14**: TestDFSIO: Read performance using HDD vs. SSD on 1x SGI Rackable C2005-TY6 Hadoop cluster node*

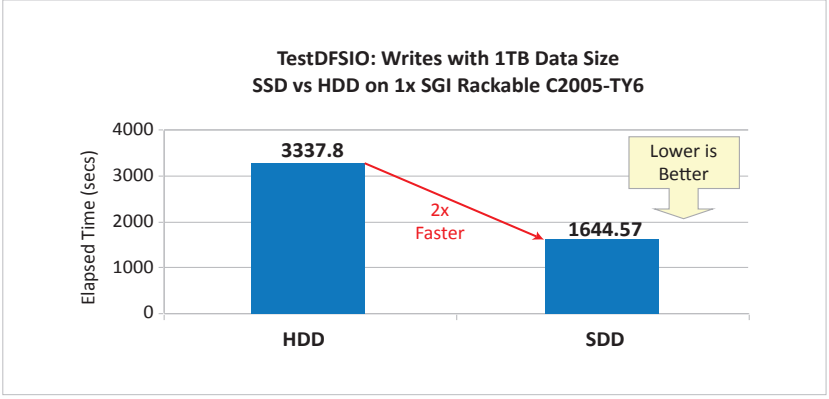

***Figure 15:** TestDFSIO: Write performance using HDD vs. SSD on 1x SGI Rackable C2005-TY6 Hadoop cluster node*

6.0 Price/Performance

This section describes the price/performance metrics of a 20-node Hadoop cluster using SGI Rackable C2005-TY6 nodes. Also, the variation of price/performance of HDDs vs. SDDs is calculated using the measured data and projections derived in Section 5 (Figure 13).

With HDDs on all of the 20 nodes, the measured price/performance is $3.44/TPM (as of September 2011), whereas with all SSDs in the same configuration, the projected price/performance is $3.74/TPM (as of September 2011), which is 8.5% higher than that of the HDD configuration (Pricing is available upon request). TPM is the Terasort-Throughput per minute.

Thus, a cluster with all HDDs may often be less expensive in Price/Performance than a cluster with all SSDs. As such, it is important to justify the need for SSDs depending on the workload being executed.

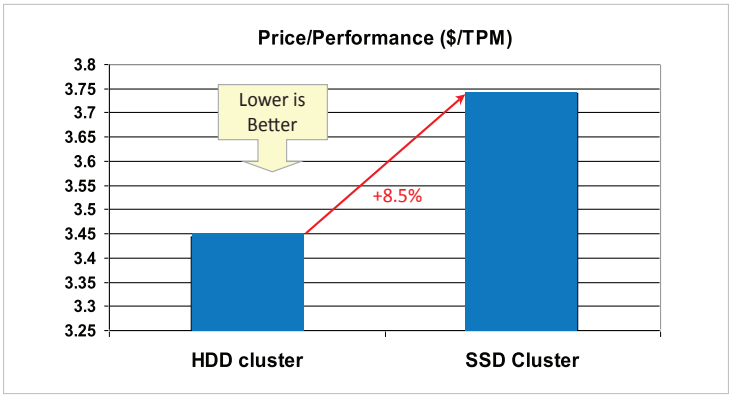

Figure 16: *Price/Performance ($/Terasort-Throughput per min): HDDs vs. SSDs on a 20-node SGI Rackable C2005-TY6 Hadoop cluster*

7.0 SGI Rackable servers for Hadoop

This section describes the value proposition of SGI Rackable Half-Depth Servers, in particular the C2005-TY6 nodes, for Hadoop.

The C2005-TY6 server with Intel Xeon 5600 processors has the following features and performance characteristics:

- Half-depth form factor for density;
- 10 x 2.5" hot-swap drives for 10 TB mixed content;
- HDD or Flash-enabled;
- AC, DC, or AC RPS power;
- Optimal SPECpower® ssj_ops/watt on single server[8];
- Super-linear scalability and record performance with Terasort on a Hadoop cluster;
- Affordable Price/Performance (Refer to Section 6).

Figure 17: *C2005-TY6 with Intel Xeon 5600 series Processors*

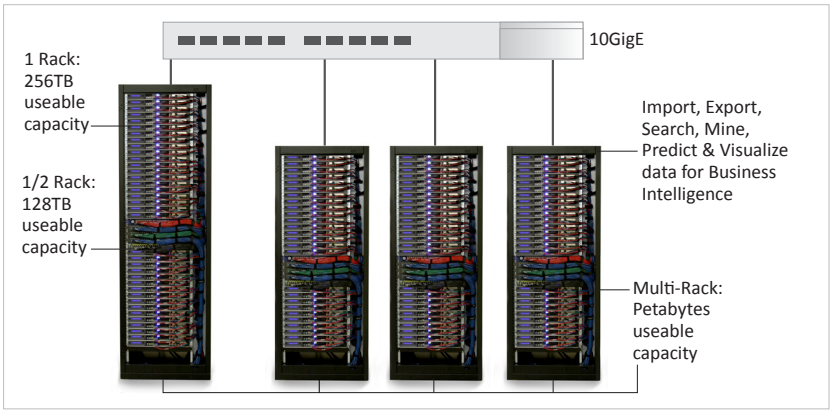

Figure 18: *SGI Rackable™ Half-Depth Servers in one or many Racks —Meeting the needs of daily "net new" data into Hadoop*

[8] *http://www.spec.org/power_ssj2008/results/res2011q1/power_ssj2008-20101210-00322.html*

8.0 Summary

As described in this paper, SGI® Rackable™ half-depth C2005-TY6 servers combine affordable price/performance, super-linear scalability, disk capacity, rack density, and power and cooling efficiency to serve especially well as Hadoop DataNodes. These characteristics also enable customers to easily add servers to an existing Hadoop cluster, scaling out to multiple racks while minimizing energy usage and maintaining optimal performance and high capacity in a data center level Hadoop deployment.

To summarize:

- SGI Rackable clusters can deal with challenges in volume, velocity, variety and vitality of unstructured data for deep analytics;
- Flexibility of choice of memory, compute, capacity, I/O, network latency, and low power of Rackable Rackmount servers helps to:
  - Optimize task assignment and expedite MapReduce tasks across distributed nodes with efficient parallelism;
  - Maintain filesystem metadata operations for HDFS;
  - Store large HDFS files and handle HDFS read/write requests;
  - Co-locate I/O with TaskTrackers for optimal data locality;
  - Achieve optimal performance-per-watt across all load levels.
- An SGI Rackable cluster is able to provide super-linear Terasort scalability with excellent price/performance as low as $3.44/TPM (as of September 2011), where TPM is the Terasort-Throughput per minute for a 20-node configuration with hard drives (HDDs);
- With an ecosystem of various analytical options on top of Hadoop, an SGI Rackable cluster can offer data integration capabilities and a best-in-class business intelligence solution over Hadoop.

SGI® Rackable servers for Hadoop are most suitable for social media/online gaming, financial trading, federal, telecommunications and other commercial markets, where large scalability, energy efficiency for server consolidation and extreme performance of Big Data analytics, are the primary goals.

Contact:

Sanhita Sarkar
Big Data Solutions & Performance
SGI, 46555 Landing Parkway Fremont, CA 94538

Corporate Headquarters
46600 Landing Parkway
Fremont, CA 94538
tel 510.933.8300
fax 408.321.0293
sgi.com

Global Sales and Support
North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.927.8000
Asia Pacific +61 2.9448.1463

© 2011 SGI. SGI and Rackable are registered trademarks or trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States or other countries. Intel and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. SPECpower® and SPECpower_ssj® are registered trademarks of the Standard Performance Evaluation Corporation (SPEC). All other trademarks mentioned herein are the property of their respective owners. 17112011 4333