



WHITEPAPER

# Boosting Hadoop Performance with Emulex OneConnect™ 10GbE Network Adapters





### ■ Abstract

Hadoop is designed to run on commodity servers and is optimized for 1 Gigabit Ethernet (GbE) networks. But today's commodity servers are delivering a substantial leap in processing and I/O performance. With recent and ongoing advancements in processor and storage technology, the legacy 1GbE network has become the bottleneck in Hadoop clusters composed of commodity servers. Emulex Corporation, the leading 10GbE ports supplier, has tasked the Emulex Advanced Development Organization to analyze the performance benefits of 10GbE versus 1GbE for Hadoop clusters.

The focus of this paper is to study the impact of migrating to a 10GbE network when importing/migrating data into or rebuilding the cluster. Network throughput is measured for various workloads and it will be shown that a 10GbE network can transfer data into a Hadoop cluster up to four times faster than a 1GbE network. A brief discussion on optimizing Hadoop parameters for 10GbE networks is included. This paper is intended for those in the planning stages of implementing a new Hadoop cluster or those who are seeking ways to improve performance for an existing one.



**Table of Contents**

**Abstract** ..... 2

**Introduction** ..... 4

**The Moving Bottleneck** ..... 5

**Hadoop on 10GbE** ..... 6

**Fine Tuning Hadoop** ..... 7

**The Setup** ..... 7

    Hardware/software configuration ..... 7

    Cluster configuration ..... 8

**The Tests** ..... 8

**The Results** ..... 9

    Importing data with 1GbE ..... 9

*Single client, single operation* ..... 9

*Single client, multiple operations* ..... 9

*Multiple clients, multiple operations* ..... 10

    Importing data with 10GbE Emulex OneConnect™ OCe11102 ..... 11

*Single client, single operation* ..... 11

*Single client, multiple operations* ..... 11

*Multiple clients, multiple operations* ..... 12

    Comparison ..... 13

**Conclusion** ..... 14

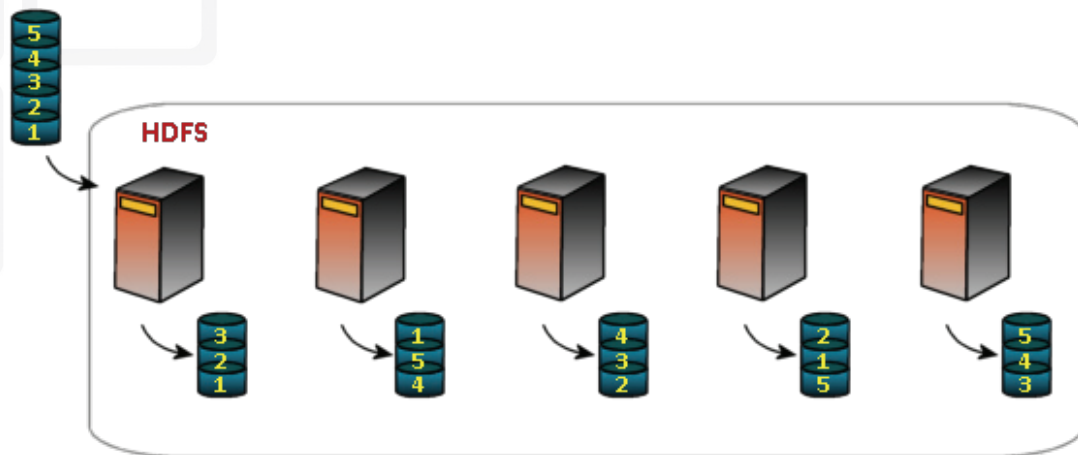


### Introduction

In the last decade, we have witnessed the creation of data sets, which have sizes well beyond the ability of common data processing tools to effectively store and analyze. In 2010, Facebook announced that with 21 petabytes<sup>1</sup> of storage, they had the largest Hadoop cluster in the world. In 2011, they announced that the data had grown to 30 petabytes. According to Cisco<sup>2</sup> by 2014, global Internet traffic will increase to 767 exabytes. This is equivalent to 191 billion DVDs. The industry has dubbed this trend “Big Data”. IDC and Gartner predict that by 2012, approximately 2.43 zettabytes of unstructured data will be constructed. Gartner predicts that, “Through 2015, more than 85 percent of the Fortune 500 organizations will fail to effectively exploit Big Data for competitive advantage.”

Big Data requires massively parallel processing (MPP)<sup>3</sup>, distributed file systems, and scalable storage systems. Hadoop, a software framework designed to meet Big Data’s requirements, is modeled after Google’s MapReduce and Google File System (GFS), and is being developed by the open source community. It has already been adopted by more than 150 businesses that offer services based on it or around it. Simply put, Hadoop is a distributed, scalable, and fault tolerant data storage and processing system running on clusters of commodity servers. Hadoop’s data storage services are provided through HDFS (Hadoop Distributed File System), as shown in Figure 1, and its processing system, a high performance parallel data processing service, is provided through map-reduce.

Figure 1  
HDFS with  
a replication  
factor of 3



<sup>1</sup> Zettabyte (ZB) = 10<sup>21</sup>, Exabyte (EB) = 10<sup>18</sup>, Petabyte (PB) = 10<sup>15</sup>, Terabyte (TB) = 10<sup>12</sup>, Gigabyte (GB)=10<sup>9</sup>

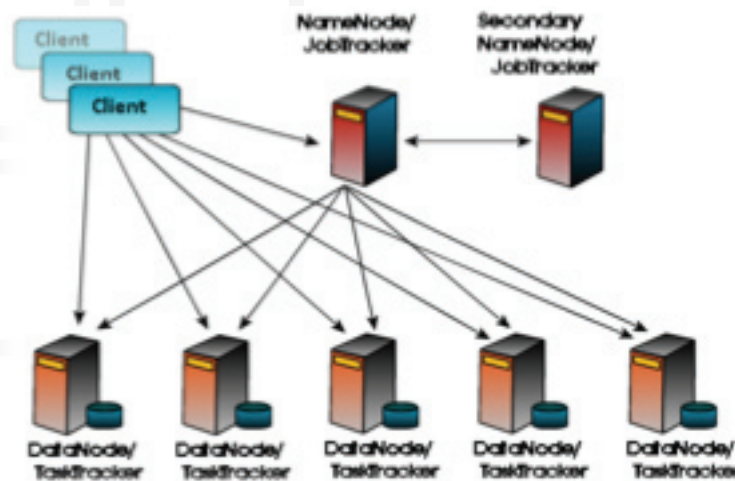
<sup>2</sup> [http://newsroom.cisco.com/dlls/2010/prod\\_060210.html](http://newsroom.cisco.com/dlls/2010/prod_060210.html)

<sup>3</sup> A massively parallel computer is a distributed memory computer system which consists of many individual nodes, each of which is essentially an independent computer in itself, and in turn consists of at least one processor, its own memory, and a link to the network that connects all the nodes together. Such computer systems have many independent arithmetic units or entire microprocessors that run in parallel. The term massive connotes hundreds if not thousands of such units. Nodes communicate by passing messages, using standards such as MPI. (see [http://en.wikipedia.org/wiki/Massively\\_parallel](http://en.wikipedia.org/wiki/Massively_parallel)).

<sup>4</sup> <http://wiki.apache.org/hadoop/PoweredBy>



With a master/slave architecture, an HDFS cluster is comprised of a single master server called the NameNode, and a number of slave servers, or DataNodes as shown in Figure 2. The NameNode manages the file system namespace and controls access to files residing on DataNodes. Each DataNode manages storage attached to the nodes they run on. Files are divided into one or more blocks. The blocks are then stored in a set of DataNodes. File system namespace operations such as opening, closing, and renaming files and directories in addition to block mapping, are performed by the NameNode. The DataNodes service read/write requests from the file system's clients. The DataNodes also service block creation, deletion, and replication requests by the NameNode.



**Figure 2**  
Hadoop topology.

### The Moving Bottleneck

Hadoop is designed to run on commodity servers and is optimized for a 1GbE network. But today's commodity servers are delivering much more processing and disk I/O performance. The Intel Nehalem micro architecture, with its memory and I/O subsystem, provides a leap in performance, thereby justifying replacement of older servers. Intel 2<sup>nd</sup> generation Core i7 processors offer four cores capable of running eight threads and delivering a maximum memory bandwidth of 25.6GB/s on a DMI bus.<sup>5</sup> PCIe 3.0's 8GT/s bit rate effectively delivers double PCIe 2.0's bandwidth. DDR4 memory modules with bus frequencies of 2,133 MT/s regular speed, and 3,200 MT/s enthusiast speed will be introduced in 2012. In December 2009, Micron Technology announced the world's first Solid State Drive (SSD) using a six gigabits per second (Gbit/s) or 600 (MB/s) SATA interface.

Capable of ingesting much more data at a faster rate than ever before, multi-core multi-threaded processors, teamed with fast DDR memory, expanded system memory space, and an increased number of faster and larger internal system drives, are quickly shifting the bottleneck to the legacy 1GbE network, which is arguably the slowest component of today's systems.

<sup>5</sup> A bus is a subsystem that transfers data between computer components or between computers. Types include front-side bus (FSB), which carries data between the CPU and memory controller hub; direct media interface (DMI), which is a point-to-point interconnection between an Intel integrated memory controller and an Intel I/O controller hub on the computer's motherboard; and Quick Path Interconnect (QPI), which is a point-to-point interconnect between the CPU and the integrated memory controller. [en.wikipedia.org/wiki/Bus\\_\(computing\)](http://en.wikipedia.org/wiki/Bus_(computing))



Even though 10GbE was introduced in the early 2000s, the majority of server connections in data centers are still at 1Gbps, mostly due to cost considerations. Multiple gigabit connections have been used to achieve higher bandwidths. However, port shipments of 10GbE adapters are on the rise. By 2013, shipments of 10GbE are expected to exceed 1GbE shipments. 10GbE ports are growing based on:

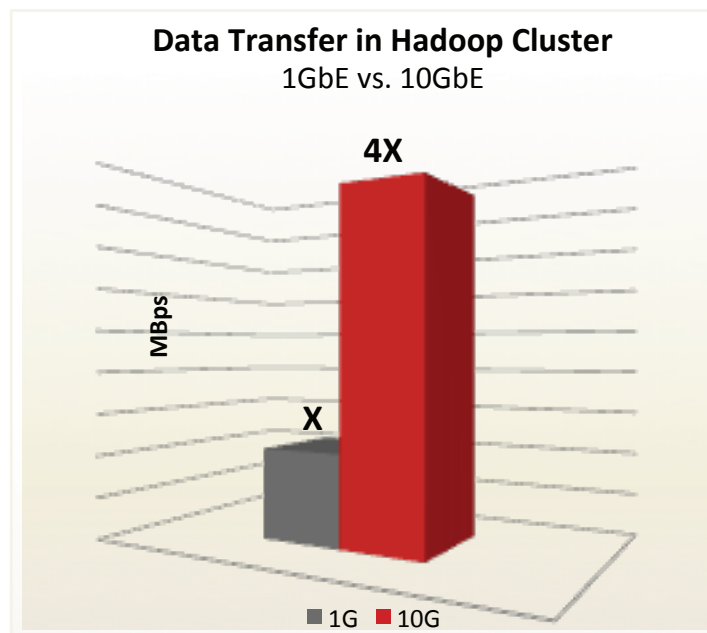
- New servers, such as HP G7 servers, feature 10GbE LAN on motherboards (LOM)
- Support for twisted-pair cable using 10GBaseT adapters is driving down costs
- Technologies, such as Emulex Universal Multi-Channel, HP's Flex10, and IBM's VFA enable parsing up a 10GbE pipe into multiple variable Ethernet ports
- New applications such as storage convergence with FCoE or iSCSI require 10GbE capacity
- Blade servers with high network attached requirements are lowering connectivity costs and are quickly adopting 10GbE as a standard
- Virtual Server technology and its requirement for a minimum of six 1GbE ports are making 10GbE attractive

As the volume goes up, the cost of 10GbE is coming down, making it more attractive for Hadoop clusters. However, what are the performance benefits of a wider network pipe for Hadoop clusters? Emulex Corporation, the leading 10GbE ports supplier, has tasked the Emulex Advanced Development Organization (ADO) to analyze the performance benefits of 10GbE versus 1GbE for Hadoop clusters.

### Hadoop on 10GbE

As Big Data increases in size and scope, the demand to increase processing power, data capacity and storage performance increases. To meet this demand, the network I/O performance must scale to support larger and faster systems. As evidenced in tests performed by the ADO, migrating from a 1GbE to a 10GbE network using an Emulex OneConnect OCe11102 quadrupled Hadoop's data transfer speed (see Figure 3).

**Figure 3**  
Emulex OneConnect  
OCe11102 delivers up  
to 4x the throughput.





Today's commodity servers allow for increased DataNode capacity, greater processor performance and storage access speed. However, very little of the generational performance gains are utilized while 1GbE network pipes remain congested. Offering four times the performance of traditional 1GbE networks, 10GbE is the necessary and obvious solution to meet the growing demands of Big Data.

### Fine Tuning Hadoop

Hadoop administrators have discovered a need to fine tune Hadoop data clusters through more than 200<sup>6</sup> cluster-wide and job-specific parameters, such as replication factor, number of parallel map/reduce tasks to run, number of parallel connections for transferring data, file system block size, map output compression, maximum tasks for a TaskTracker, etc. Since one size does not fit all, administrators must understand the workload and where to expect bottlenecks.

Most often, a workload's resource demands are not equally distributed, and can be categorized as either CPU or I/O intensive. For example, indexing, searching, grouping, decoding/decompressing, and data importing/exporting are I/O intensive. Machine learning, complex data/text mining, natural language processing, and feature extraction are CPU intensive workloads. Furthermore, Hadoop and HDFS are, by default, optimized for a 1GbE network. Parameter adjustments are necessary to optimize performance when implementing a 10GbE network. Also, for best results, the operating system, and particularly the local file system, must be carefully selected and adjusted.

### The Setup

To examine the deficiencies imposed by legacy 1GbE networks currently being shipped, entry-level servers with processor and memory specifications similar to those of Open Compute Project<sup>7</sup> v1.0 were used.

#### Hardware/software configuration

The ADO tests were performed with the following cluster configuration (see Figure 4):

- Server specifications:
  - HP ML350 G6:
    - *Dual Processors (Quad core Intel Xeon 2GHz)*
    - *16 GB DDR3*
    - *Broadcom 1G NetXtreme BCM5715*
    - *Emulex OneConnect™ OCe11102 Ethernet Adapter*
- Storage:
  - SATA II 500GB 7200rpm Disk Drives, 6 per node
  - HP Smart Array G6 RAID Controller (JBOD - No RAID configured)
- OS and Software:
  - Ubuntu 64 bit
  - Hadoop HDFS

<sup>6</sup> 53 available in hdfs-site.xml, 57 available in core-site.xml, and 124 available in mapred-site.xml

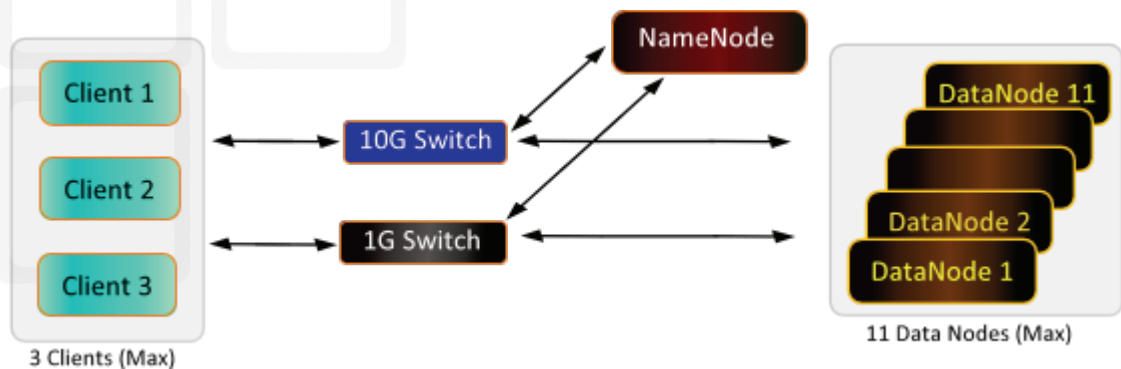
<sup>7</sup> <http://opencompute.org/projects/intel-motherboard/>



### Cluster configuration

- 15 servers with discrete roles:
  - 1 NameNode
  - 11 DataNodes
  - 3 Clients
- 1GbE and 10GbE Switches

**Figure 4**  
Cluster configuration.



### The Tests

Big Data is being generated at staggering rates and it has to be ingested at extremely high speeds before it can be analyzed. The goal of this study was to measure the raw HDFS throughput when ingesting data into a Hadoop cluster. Using HDFS 'put' operations, the data transfer rate for a 1GbE network was measured for an eleven (11) DataNode cluster. The number of clients was increased from one to three while the number of HDFS 'put' operations was incrementally increased as shown below. Each 'put' operation transferred a 5GB file to the cluster. The replication factor was set to three.

- 1 Client, 11 DataNodes, 1, 2, 4, 6, and 8 'put' operations
- 2 Client, 11 DataNodes, 1, 2, 4, 6, and 8 'put' operations per client (2, 4, 8, 12, and 16 operations total)
- 3 Client, 11 DataNodes, 1, 2, 4, 6, and 8 'put' operations per client (3, 6, 12, 18, and 24 operations total)

The tests were then repeated for a 10GbE network.

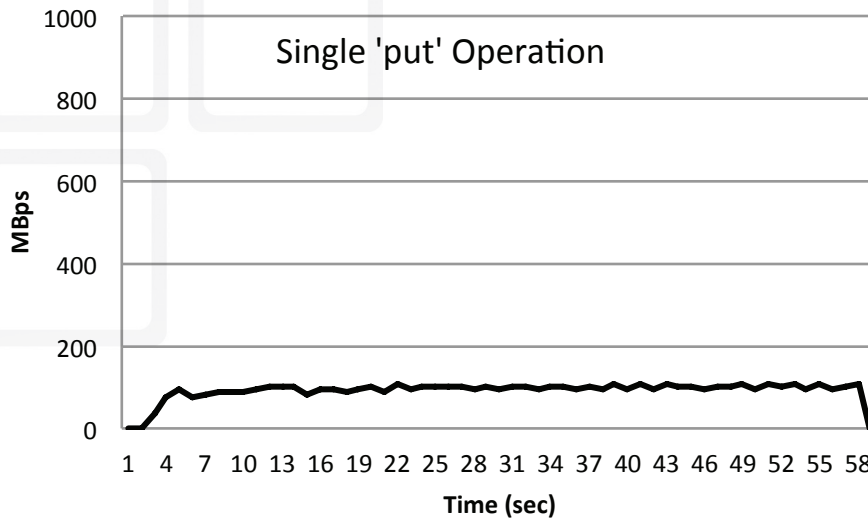


### The Results

#### Importing data with 1GbE

##### Single client, single operation

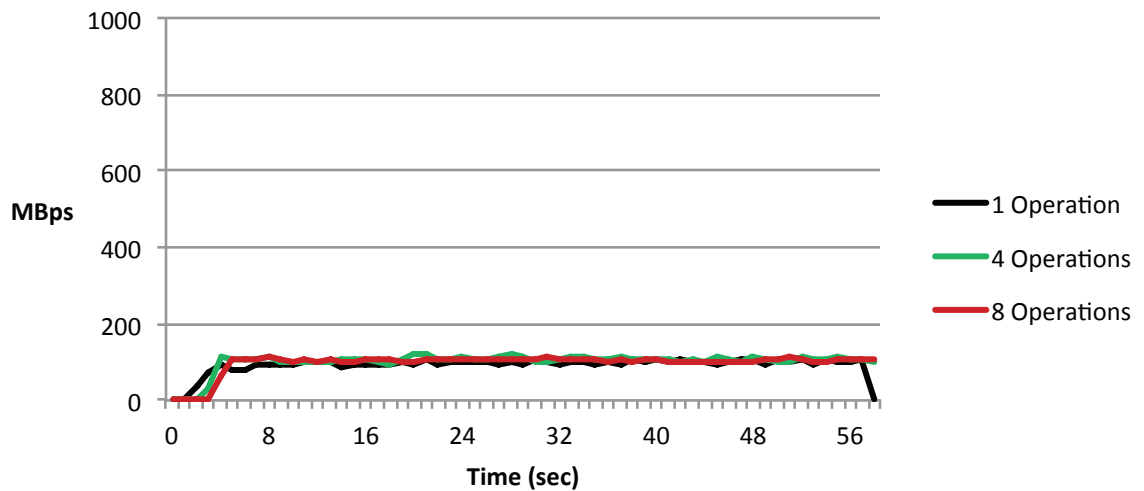
Data transfers into a 1GbE Hadoop cluster maximally use the network when there is one 'put' operation running on a single client server. Tests showed that HDFS efficiently transfers data to multiple DataNodes within the cluster, maintaining an average throughput of 108MBps out of the client server, as shown in Figure 5.



**Figure 5**  
Transfer rate:  
single client,  
single operation,  
1GbE network  
is fully utilized.

##### Single client, multiple operations

However, if more than one 'put' operation is running on a client server, the 1GbE network becomes a bottleneck in the system. As seen in the graph in Figure 6 below, increasing the number of 'put' operations did not increase the throughput out of the client server since it was restricted by the 1GbE network connection.



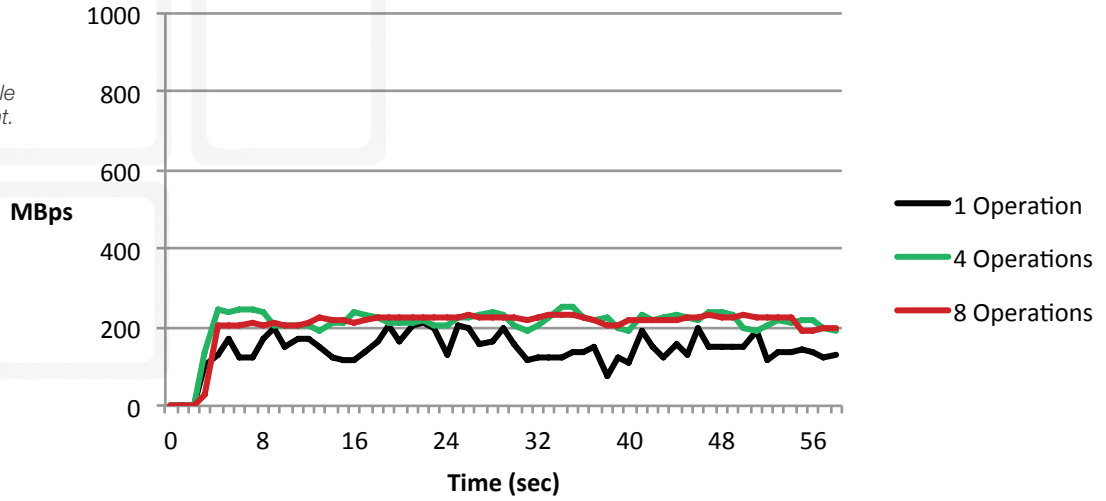
**Figure 6**  
1GbE transfer  
rate: single  
client, multiple  
operations.



**Multiple clients, multiple operations**

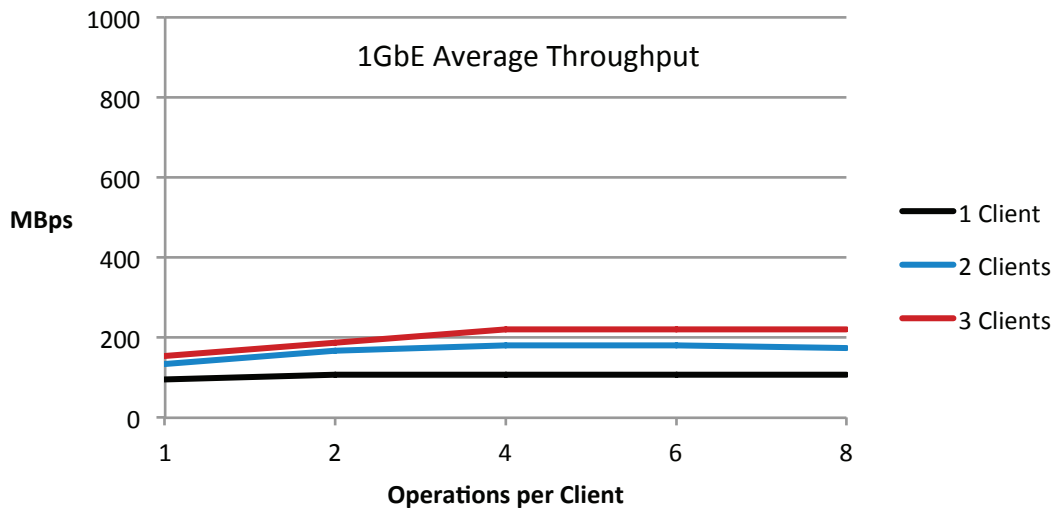
Since most implementations of Hadoop clusters utilize more than one client, the number of client servers was incrementally increased. It was expected that adding more clients would result in proportionally higher total throughput. However, with three clients each running eight 'put' operations for a total of 24 operations, 1GbE network achieved an average throughput of 225MBps. That was only slightly higher than twice the 1GbE maximum throughput (see Figure 7). As network load is increased, 1GbE reached saturation, creating a bottleneck in the system.

**Figure 7**  
1GbE transfer rate:  
three clients, multiple  
operations per client.  
Throughput does  
not improve.



As the number of clients and 'put' operations running on each client was increased, the 1GbE network flatlined (see Figure 8).

**Figure 8**  
1GbE flatlines  
as the number  
of operations is  
increased.





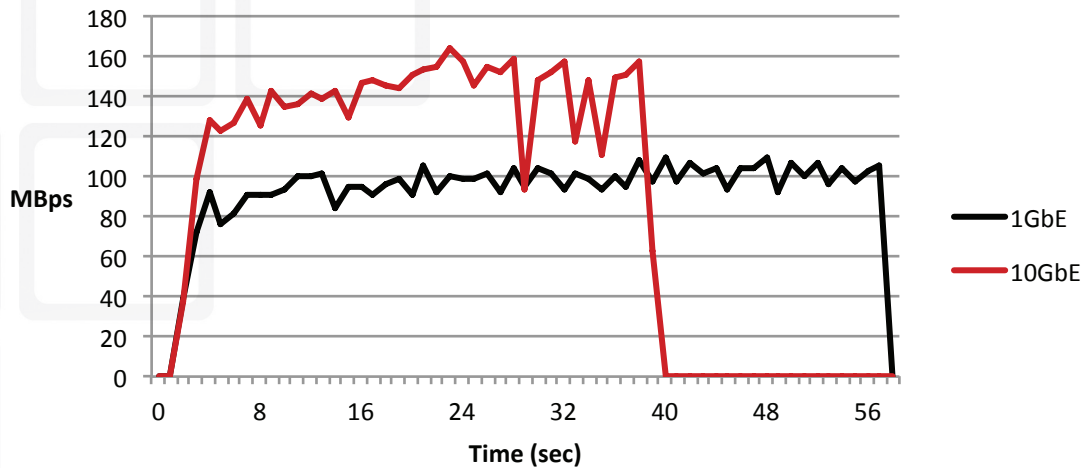
### Importing data with 10GbE Emulex OneConnect OCe11102

To analyze the performance benefits of 10GbE versus 1GbE for Hadoop clusters, the previous tests were repeated using a 10GbE Emulex OneConnect OCe11102 Ethernet Adapter.

#### Single client, single operation

As shown in Figure 9, with one HDFS 'put' operation running on a single client server, a 10GbE network showed an immediate performance improvement of 50 percent over the 1GbE network. 10GbE completed the data transfer in less than three quarters of the time.

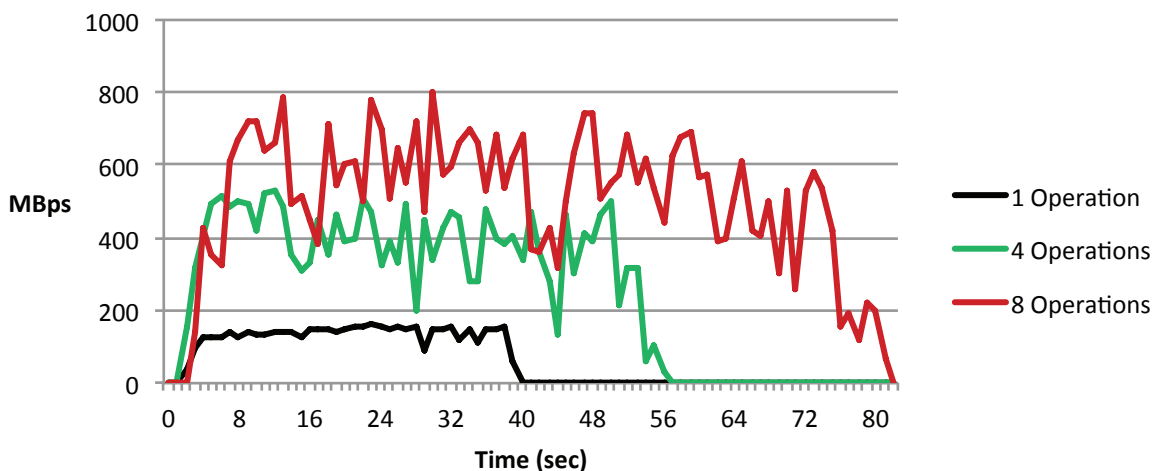
**Figure 9**  
Saving time with Emulex OneConnect OCe11102.



#### Single client, multiple operations

Unlike the 1GbE network, as the number of operations was increased, the 10GbE network did not limit the transfer rate. As seen in Figure 10, increased network load was met with increased throughput. With one client server running eight 'put' operations, a transfer rate of approximately 800MBps was achieved.

**Figure 10**  
Transfer rate: single client, multiple operations. 10GbE network is not the bottleneck.

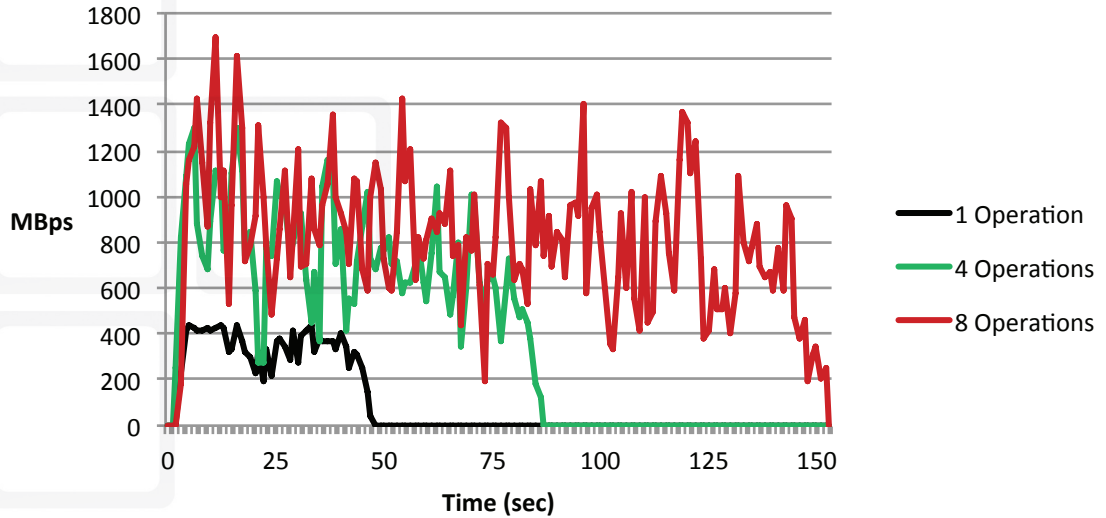




**Multiple clients, multiple operations**

As shown in Figure 11, increasing the number of clients and operations did not result in network saturation.

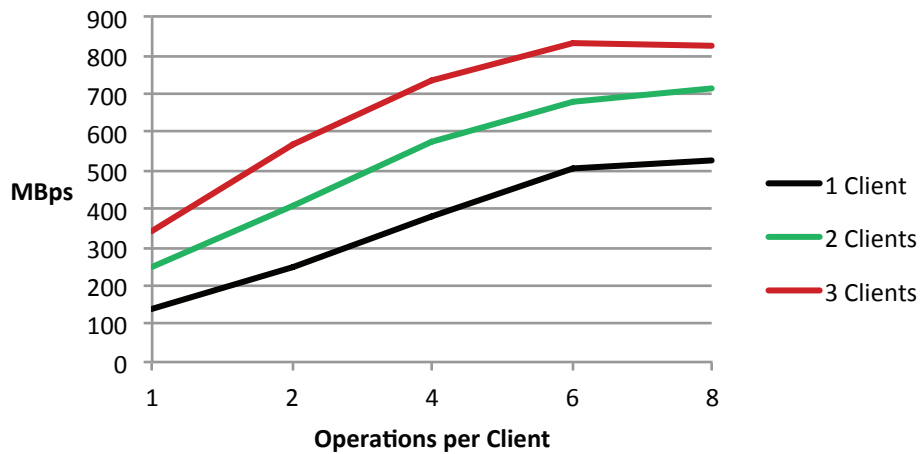
**Figure 11**  
10GbE network,  
3 clients  
and multiple  
operations  
per client.  
Throughput is  
not limited.



Unlike the 1GbE network (see Figure 8), as the number of clients and operations was increased, throughput was also increased (see Figure 12).

**10GbE Average Throughput**

**Figure 12**  
Data transfer rate  
vs. number of  
'put' operations.

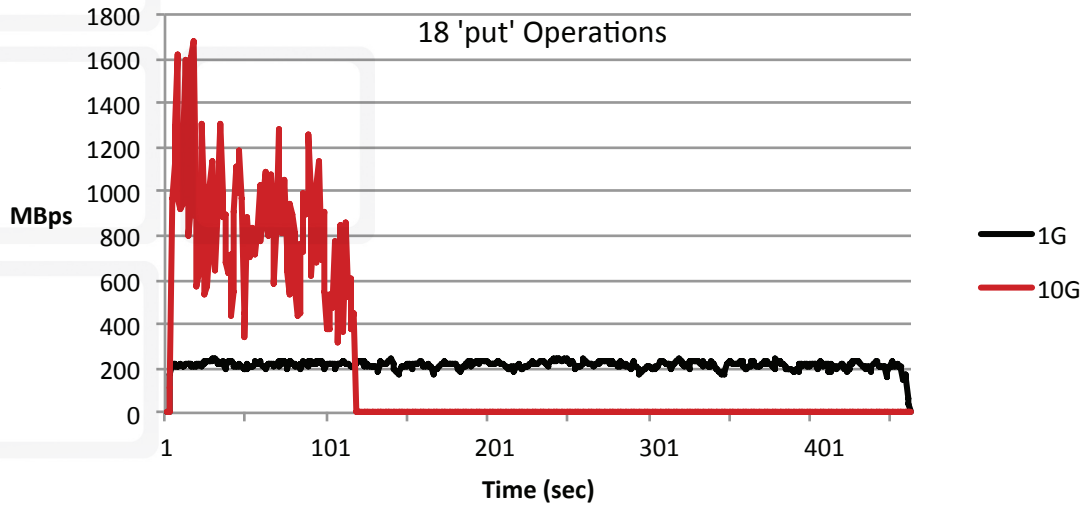




**Comparison**

Cluster ingest speed was measured by running six 'put' operations on each of three client servers to write a total of 270GB of data (90GB with a replication factor of three) to an 11 node cluster. The results are shown in Figure 13 and Table 1.

**Figure 13**  
Emulex OneConnect OCe11102 gets the job done four times faster than 1GbE.

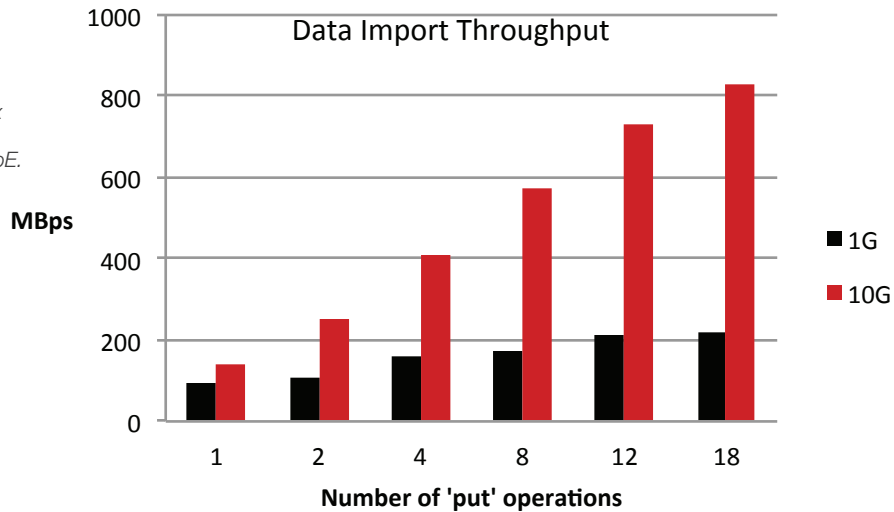


**Table 1**  
10GbE delivers throughput and saves time.

Network Type	Average Throughput (MBps)	Maximum Throughput (MBps)	Time to complete (seconds)
1GbE	216	251	453
10GbE	831 (3.85x better)	1,674 (6.7x better)	115 (3.94x better)

Increased 'put' operations are handled by the 10GbE network but the 1GbE network throttles (see Figure 14).

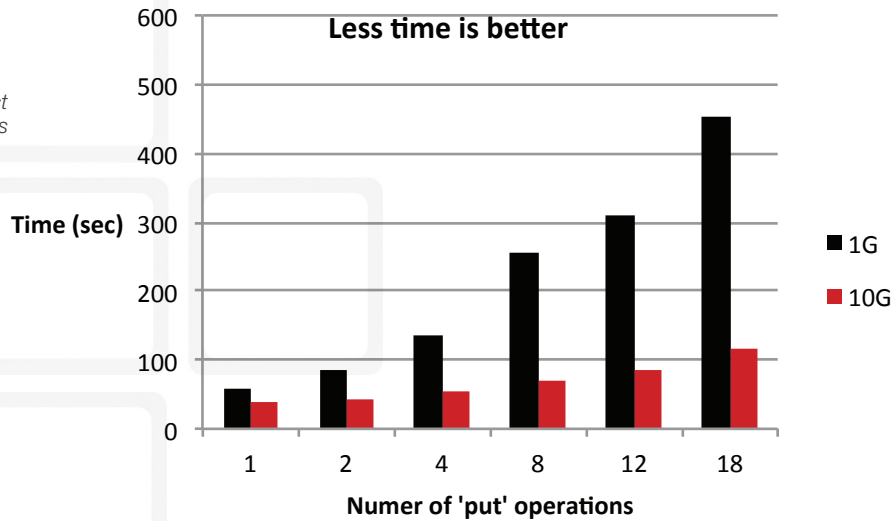
**Figure 14**  
1GbE vs. Emulex OneConnect OCe11102 10GbE.



The 1GbE network slowed exponentially when the load was only incrementally increased, as shown in Figure 15.



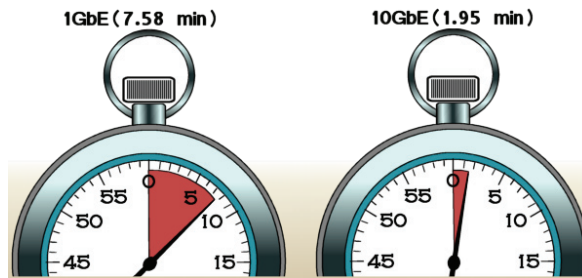
**Figure 15**  
Emulex OneConnect  
OCe11102 performs  
four times better  
under load.



**Conclusion**

Performance of Hadoop can be increased to a great degree by tuning some parameters based on the cluster specifications, input data size, and processing complexities. Nevertheless, with huge leaps in processor, RAM, and storage performance in today's commodity servers' building blocks, yesterday's 1GbE network has become the bottleneck in Hadoop clusters. Boosting throughput with a 10GbE network infrastructure improved overall Hadoop performance but required fine tuning of certain Hadoop parameters. In the ADO study, data imports (HDFS 'put' operations) were four times faster when migrated from a 1GbE to 10GbE network, as shown in Figure 16.

**Figure 16**  
HDFS is four times  
faster with Emulex  
OneConnect  
OCe11102 10GbE  
adapters.



With increased network throughput, data ingestion and data replication take less time. As a result, more CPU cycles become available for data analysis, improving system response time and reducing the required number of cluster nodes. Cluster size reduction and increased transaction response time result in significant cost reductions in implementing, operating, and maintaining a Hadoop cluster.



Emulex is the market share leader in 10GbE Ethernet ports. This market share dominance is being driven by companies like HP and IBM who have chosen Emulex as the standard LOM on their new servers. Exceptional across-the-board performance for bandwidth, IOPS, latency and CPU-offload drove the selection of Emulex. In a survey of 130 Emulex Ethernet Adapter customers, 94 percent were satisfied or very satisfied with the OneConnect Ethernet Adapter performance. Emulex enterprise-proven reliability is also a key driver with 97 percent of the customers indicating they were satisfied or very satisfied with the OneConnect Ethernet Adapter reliability.

Emulex, the leader in converged networking solutions, provides enterprise-class connectivity for servers, networks and storage devices within the data center. The Company's product portfolio of Fibre Channel host bus adapters, network interface cards, converged network adapters, controllers, embedded bridges and switches, and connectivity management software are proven, tested and trusted by the world's largest and most demanding IT environments. Emulex solutions are used and offered by the industry's leading server and storage OEMs including, Cisco, Dell, EMC, Fujitsu, Hitachi, Hitachi Data Systems, HP, Huawei, IBM, NEC, NetApp and Oracle. Emulex is headquartered in Costa Mesa, Calif., and has offices and research facilities in North America, Asia and Europe. Emulex is listed on the New York Stock Exchange (NYSE:ELX). News releases and other information about Emulex is available at [www.Emulex.com](http://www.Emulex.com)

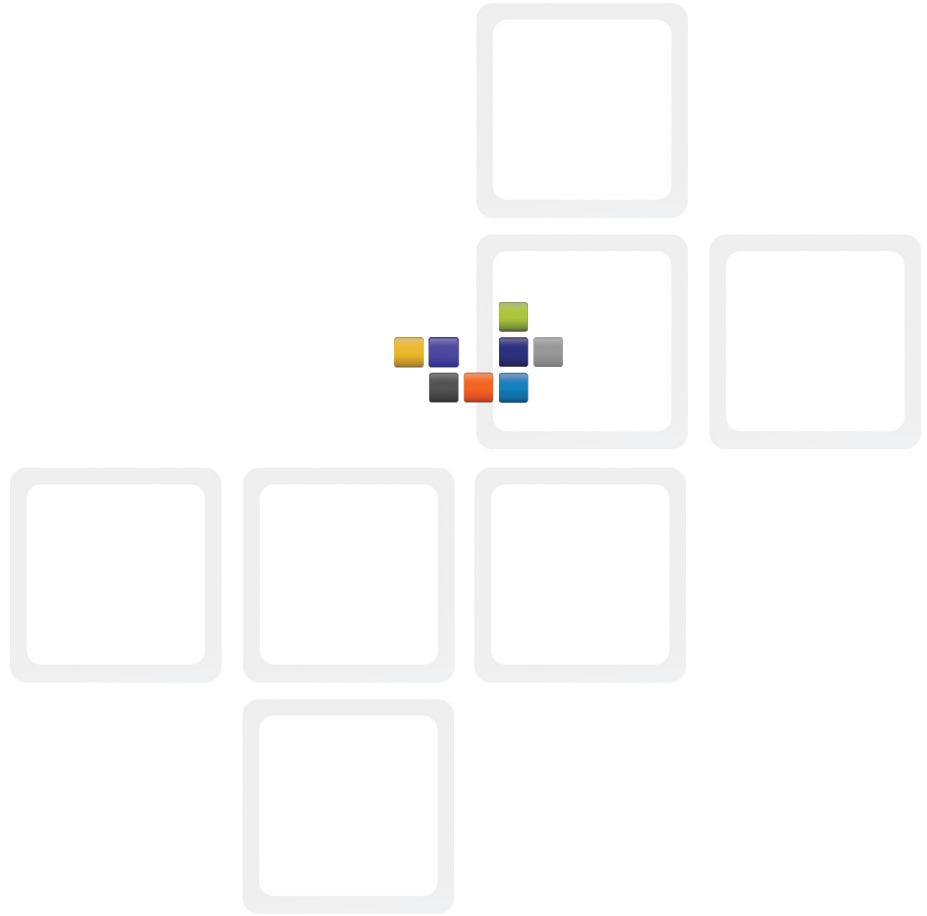
### Authors

Niels Reimers, senior architect in the Advanced Development Organization, keeps one foot in the near-term future and one in longer term. In his near-term role as customer advocate, he works with storage array vendors to identify their requirements and champion these requirements within Emulex. His longer-term role is to foster innovation within Emulex by managing the Innovation Portal and facilitating Ideation Events. The Innovation Portal provides a mechanism for employees around the world to post ideas with a framework for tracking and routing these ideas. Ideation Events are focused and organized brainstorming sessions, bringing together the great minds within Emulex to analyze a specific problem area and come up with creative solutions. Reimers has more than 20 years of experience in disk drive, storage array and IO controller development at companies including Memorex, HP/Agilent, Quantum and Emulex.

Artem Gavrilov joined Emulex in 2010 as a senior architect in the Advanced Development Organization and provides software expertise for company innovations. Currently, he is focusing on Cloud technologies and Hadoop. Artem held multiple architect and lead positions at companies including i365 (Seagate), and Microsoft. Gavrilov has a Master's degree in computer science and 15 years of industry experience in storage, virtualization, cloud computing and search.

Murthy Kompella joined Emulex in October 2006 and serves as senior director of engineering and senior architect in the Advanced Development Organization. Prior to Emulex's acquisition of Sierra Logic, Kompella served as Sierra Logic's lead architect and designer. During his career, Kompella has been responsible for more than ten enterprise storage controller chips involving Fibre Channel, SAS, SATA and PCI Express industry standards and was awarded seven patents for enterprise storage solutions he designed. Prior to Sierra Logic, Kompella held technical positions at Agilent and HP. He has 21 years of industry experience and holds a Master's degree in electrical engineering with a focus on digital systems from Osmania University in India.

©2012 Emulex, Inc. All rights reserved. This document refers to various companies and products by their trade names. In most, if not all cases, their respective companies claim these designations as trademarks or registered trademarks. This information is provided for reference only. Although this information is believed to be accurate and reliable at the time of publication, Emulex assumes no responsibility for errors or omissions. Emulex reserves the right to make changes or corrections without notice. This report is the property of Emulex and may not be duplicated without permission from the Company.



[www.emulex.com](http://www.emulex.com)

**World Headquarters** 3333 Susan Street, Costa Mesa, California 92626 +1 714 662 5600  
**Bangalore, India** +91 80 40156789 | **Beijing, China** +86 10 68499547  
**Dublin, Ireland**+35 3 (0)1 652 1700 | **Munich, Germany** +49 (0) 89 97007 177  
**Paris, France** +33 (0) 158 580 022 | **Tokyo, Japan** +81 3 5325 3261  
**Wokingham, United Kingdom** +44 (0) 118 977 2929