WHITE PAPER

April 2014

Driving IBM BigInsights Performance Over GPFS Using InfiniBand+RDMA

Executive Summary

The purpose of this study was to review the capabilities of IBM General Parallel File System (GPFS) as a file system for IBM BigInsights Hadoop deployments and to test the performance advantages of Mellanox's Remote Direct Memory Access (RDMA) for BigInsights applications using GPFS. To provide a basis of comparison, tests were run comparing the use of GPFS with Apache Hadoop Distributed File System (HDFS). Benchmark results show GPFS improves application performance over HDFS by 35% on the analytics benchmark (Terasort benchmark), 35% on write tests and 50% on read tests using the Enhanced TestDFSIO benchmark.

This paper provides details on the test architecture, methodology, results and conclusions reached during the study.

Background

"Big Data" has become the hot "buzz word" in contemporary data store and analysis discussions. "Big Data" can be described as data that does not easily fit into a traditional Relational Database Management System (RDBMS) because of the sheer volume of data or because it does not have a well-defined structure of rows and columns or clearly defined data types. Large Web 2.0 companies were the first to encounter these enormous data sets and their seminal efforts have evolved into the Apache Hadoop framework. The rest of the industry is now in the midst of a tremendous transformation trying to cope with ever growing and more complex data. Companies are searching for ways to extract actionable intelligence from this vast data set in the shortest time possible. This means that system architects need to figure out how to address the challenges of capturing, curating, managing and processing this data. IBM is solving the challenge of Big Data with the innovative and powerful IBM BigInsights product built on an underlying Hadoop architecture.

The Hadoop framework is composed of two core blocks. The first is the Hadoop Distributed File System (HDFS) that provides a scalable storage mechanism for vast amounts of data. The second is a framework that provides an analytics component called MapReduce that organizes the mining of data stored on the file system. The Hadoop architecture is going through a rapid adoption phase presenting a new set of challenges that customers are now facing. The architecture presented in this document can provide a solution for continuous data ingestion or write heavy workloads problems.

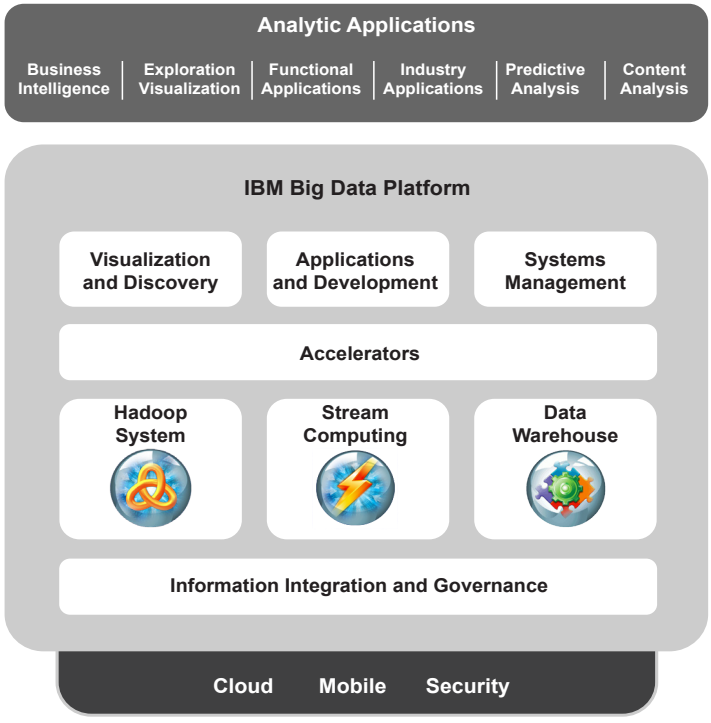

Figure 1: IBM Big Data Platform

File Systems Architecture

While HDFS is a component of the Apache Hadoop package it has several short-comings which can be overcome by replacing HDFS with another file system. One such approach offered by IBM with BigInsights is the IBM General Parallel File System (GPFS).

General Parallel File System

GPFS is an IBM product which was first released in 1998. GPFS is a high performance enterprise class distributed file system. Over the years it has evolved to support a variety of workloads and can scale to thousands of nodes. GPFS is deployed and used in many enterprise customer production environments to support machine critical applications.

GPFS Applications:

- Digital media
- Engineering design
- Business intelligence
- Financial analytics
- Seismic data processing
- Geographic information systems
- Scalable file serving

GPFS Features:

- Seamless capacity expansion to handle the extreme growth of digital information and improve efficiency through enterprise wide, interdepartmental information sharing
- High reliability and availability to eliminate production outages and provide disruption-free maintenance and capacity upgrades
- Performance to satisfy the most demanding applications
- Policy-driven automation tools to ease information lifecycle management (ILM)
- Extensible management and monitoring infrastructure to simplify file system administration
- Cost-effective disaster recovery and business continuity
- POSIX (Portable Operating System Interface) compliant

Hadoop Distributed File System

HDFS is the intrinsic distributed file system provided as part of the Apache Hadoop package. HDFS data is distributed across the local disks of multiple computers which are networked together. The computers containing the data are referred to as data nodes. Data nodes are typically homogeneous with each having multiple hard disks. Each file written to HDFS is split into smaller blocks and distributed across the data nodes. By default, HDFS is configured for 3-way replication to ensure data reliability.

HDFS Features

- Simple installation
- Block size & replication can be changed during the file creation
- Inherent reliability
- Designed for Hadoop workloads write once and read many times

Network Architecture

InfiniBand & RDMA

InfiniBand provides a messaging service that applications can access directly without requiring the operating system. Compared to a TCP/IP byte-stream oriented transport, InfiniBand eliminates the need for a complex exchange between an application and the network. Direct access means that an application does not need to rely on the operating system to transfer messages. This "application-centric" approach to computing is the key differentiator between InfiniBand and TCP/IP networks.

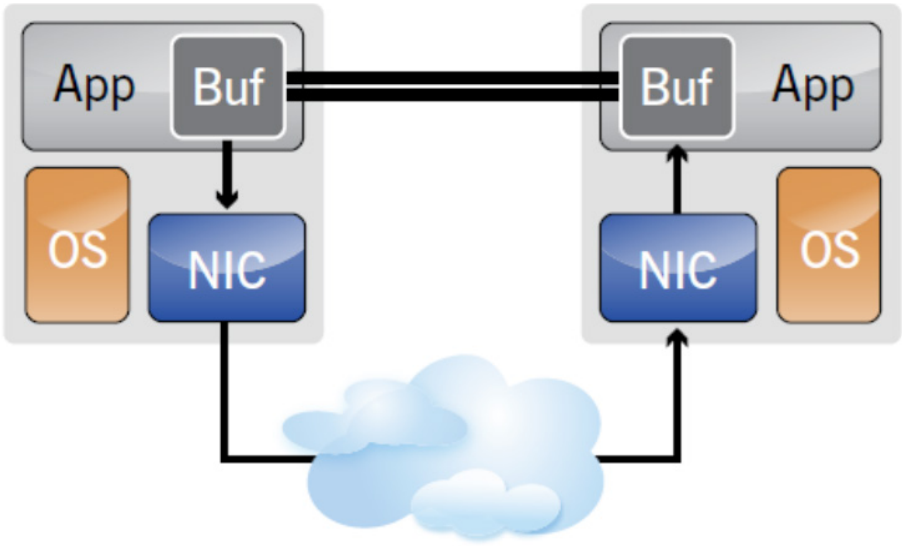

Figure 2: Messaging architecture in InfiniBand fabric

The current speed of InfiniBand FDR is 56Gb/s with communication latency of less than 1us from application to application.

A key capability of InfiniBand is Remote Direct memory Access (RDMA). RDMA provides direct application level access from the memory of one computer into the memory of another computer without requiring any services from the operating system on either computer. This enables high-throughput, low-latency, and low CPU transport overhead node to node communication which is critical in massively parallel computing clusters. As shown in the InfiniBand architecture in Figure 3 the software transport interface sits just above the transport layer. The software transport interface defines the methods and mechanisms that an application needs to take full advantage of the RDMA transport service.

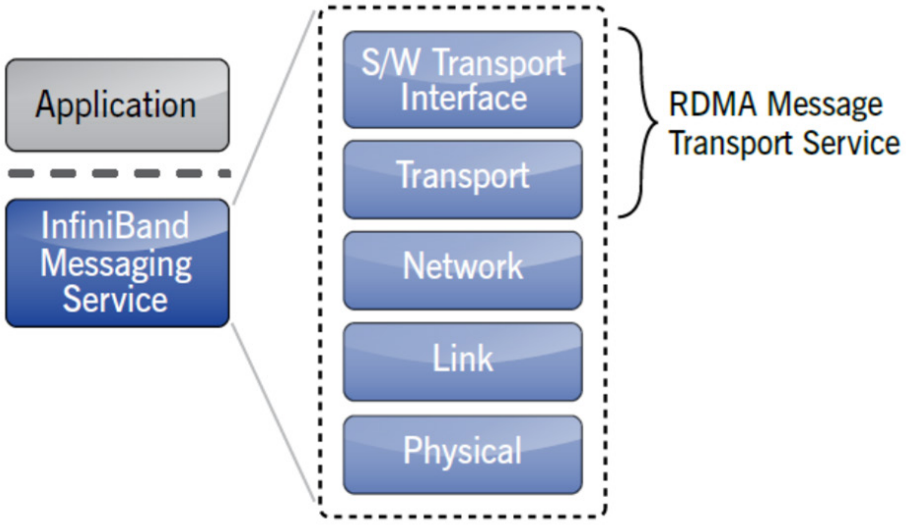

Figure 3: InfiniBand Architecture

SX6036 Switch System Overview

The SX6036 switch systems provide the highest-performing fabric solutions in a 1RU form factor by delivering 4.032Tb/s of non-blocking bandwidth to High-Performance Computing and Enterprise Data Centers, with 200ns port-to-port latency. Built with Mellanox's latest SwitchX®-2 InfiniBand switch device, the SX6036 provide up to 56Gb/s full bidirectional bandwidth per port. This stand-alone switch is an ideal choice for top-of-rack leaf connectivity or for building small to medium sized clusters. It is designed to carry converged LAN and SAN traffic with the combination of assured bandwidth and granular Quality of Service (QoS).

The SX6036 with Virtual Protocol Interconnect (VPI) supporting InfiniBand and Ethernet connectivity provide the highest performing and most flexible interconnect solution for PCI Express Gen3 servers. VPI simplifies system development by serving multiple fabrics with one hardware design. VPI simplifies today's network by enabling one platform to run both InfiniBand and Ethernet subnets on the same chassis.

ConnectX®-3 Pro Dual-Port Adapter with Virtual Protocol Interconnect® Overview

ConnectX-3 Pro adapter cards with Virtual Protocol Interconnect (VPI) supporting InfiniBand and Ethernet connectivity provide the highest performing and most flexible interconnect solution for PCI Express Gen3 servers used in Enterprise Data Centers, High-Performance Computing, and Embedded environments. Clustered data bases, parallel processing, transactional services and high-performance embedded I/O applications will achieve significant performance improvements resulting in reduced completion time and lower cost per operation.

IBM BigInsights

IBM BigInsights V2.1 release introduced support for GPFS File Placement Optimizer (FPO). GPFS FPO is a set of features that allow GPFS to support map reduce applications on clusters with no shared disk.

GPFS-FPO Benefits

- Locality awareness so compute jobs can be scheduled on nodes containing the data
- Chunks that allow large and small block sizes to coexist in the same file system to make the most of data locality
- Write affinity allows applications to dictate the layout of files on different nodes to maximize both write and read bandwidth
- Distributed recovery to minimize the effect of failures on ongoing computation

The table below shows the benefits of GPFS over HDFS to put the performance expectation into perspective.

Feature	HDFS	GPFS
Different file size handling	Supports only very large block sizes.	Wide range of supported file sizes with stable performance across the file size spectrum.
POSIX Compliance	No	Yes
Management and High Availability	No support for high availability metadata until version 2.0, ad-hoc management tool	Complete high availability support, standard interface and automation tools for management

Improving Hadoop Resiliency and Performance

When using HDFS, as shown in Figure 4 below, the metadata is managed by one master node which is called the NameNode. A NameNode failure may mean complete data loss. However, architectures such as IBM BigInsights provide a High Availability (HA) design to eliminate this Single Point of Failure (SPOF). Specifically, IBM BigInsights includes support for NameNode HA implemented since BigInsights V2.1 release. Some implementations use Network File System (NFS) with a Secondary NameNode which does not provide a complete HA but is a reasonably quick manual recovery option in case of a NameNode failure.

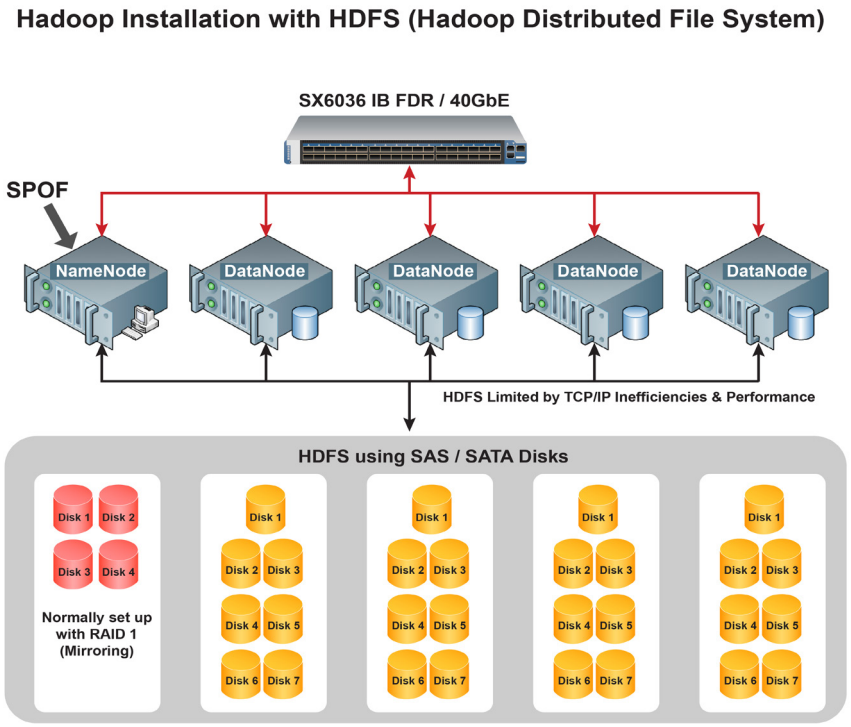

Figure 4: Hadoop HDFS Architecture

As of the time of writing this paper, HDFS 1.2 does not support tiered storage (mixing of different sizes and types of storage devices). That means that the file system treats all storage identically and is unable to optimize overall application performance based on the type of storage being used. For example, the slowest component in a Hadoop node is the spinning disk which provides max throughput of about 100-200MB/s per device. During a continuous data ingest, these disks become the bottleneck and using fast SSD's may not be cost effective when storing petabytes of data.

GPFS offers a solution to these challenges. Figure 5 shows that the very concept of a NameNode goes away when GPFS is used for Hadoop map reduce. Every node has equal access to metadata and metadata can be replicated in the same manner as data for reliability. Furthermore GPFS provides the flexibility of mixing storage devices. There can now be SSD's in the 1st tier, fast 10K RPM SAS storage in the 2nd tier, and 7.2K RPM SATA drives in the 3rd tier.

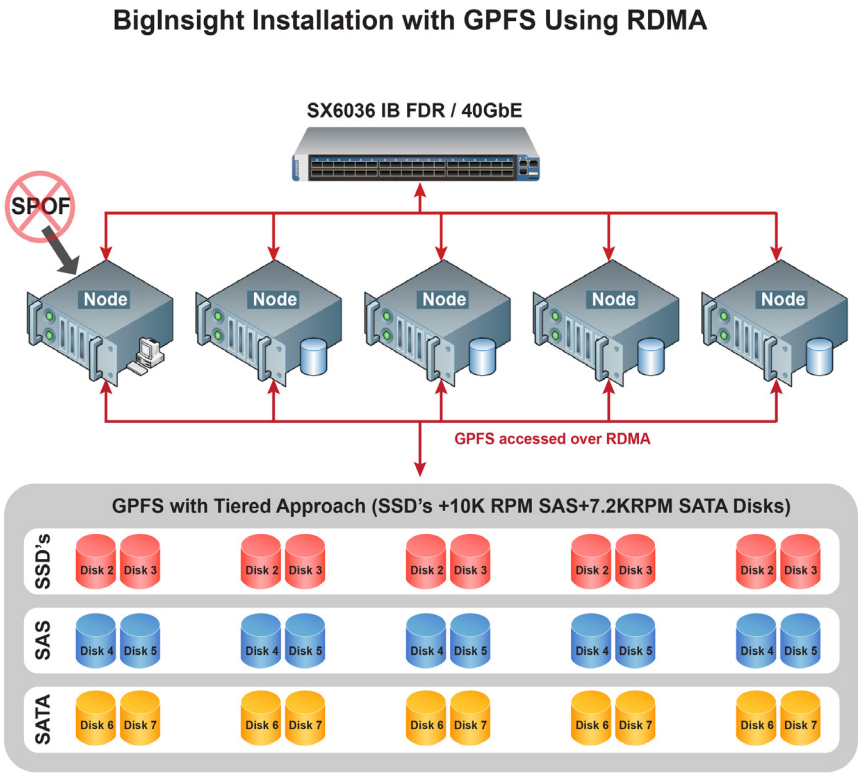

Figure 5: Hadoop GPFS Architecture

Using GPFS for MapReduce workloads has many benfits over HDFS:

- Eliminates the single point of failure of the NameNode without requiring a costly high availability design
- Provides tiered storage so you can place data on the right type of storage (SSD, 15k SAS , 7200 RPM NL/SAS or any other block device)
- Use native InfiniBand RDMA for better throughput, lower latency and more CPU cycles for your application

To get the best price performance for MapReduce workloads use IBM BigInsights with GPFS and leverage SSD, SAS and SATA drives capabilities.

Benchmark Setup

Two equally sized five node clusters with IBM BigInsights were used for comparison, one had HDFS and the other GPFS. Both clusters had file system metadata stored in SSD's. The goal was to compare Hadoop Map Reduce performance on HDFS vs. GPFS.

Both clusters used the same underlying Infiniband fabric. At the time of this test HDFS does not support RDMA so TCP/IP over Infiniband (IPoIB) was used for node to node communication. The GPFS cluster setup used the RDMA capabilities of the file system.

The complete technical details and nuances of the deployment will be published in a separate technical document.

Results

TeraSort Benchmark Results

TeraSort is a well-known Hadoop benchmark. TeraSort uses a map reduce job to sort 1TB of data as quickly as possible. TeraSort stresses both the file system and MapReduce layers of Hadoop. Figure 6 shows the 1TB TeraSort results. The test ran 35% faster on GPFS than it did on HDFS.

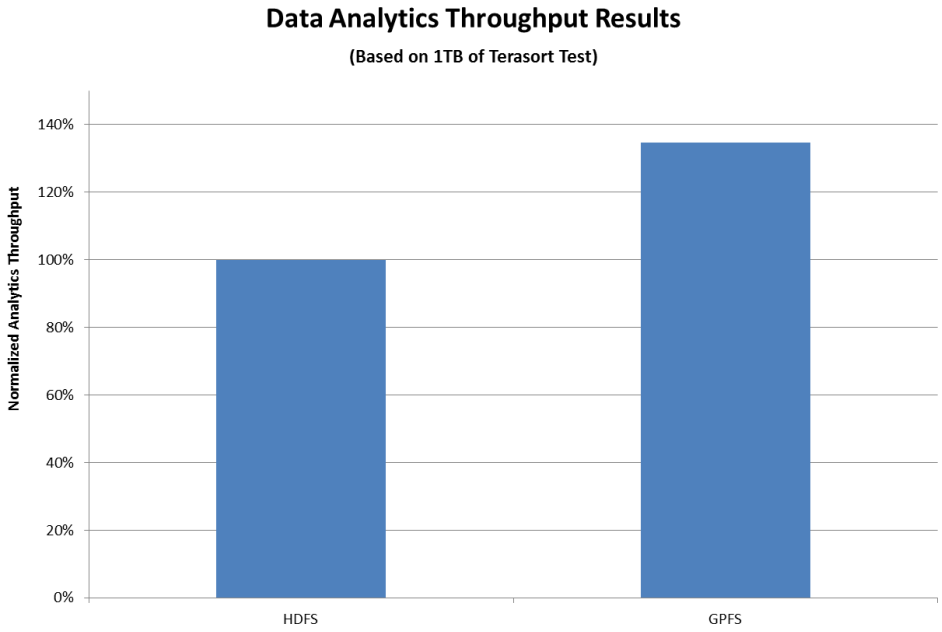

Fig 6: Data analtyics performance results comparison,higher numbers show better results

HiBench Enhanced DFSIO Results

HiBench is a benchmark suite that contains a variety of Hadoop tests. Enhanced DFSIO, from the HiBench suite, tests file system performance for sequential read and write workloads measuring the maximum throughout of the file system. Fig 7 shows the write throughput performance using the E-DFSIO test, as in the TeraSort test the cluster using GPFS provided 35% faster throughput than the HDFS cluster. Figure 8 shows the read throughput performance using E-DFSIO test. In this test GPFS was 50% faster than HDFS.

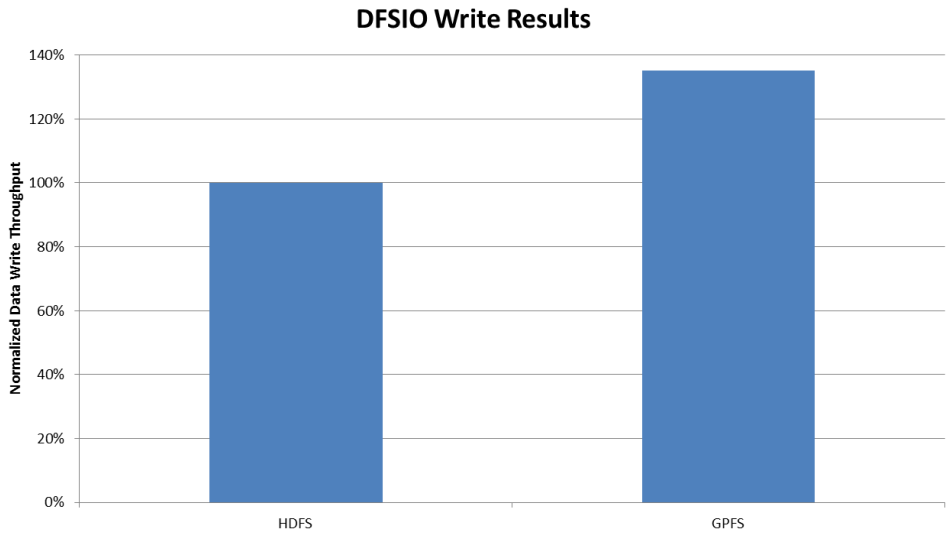

Figure 7: Write performance, higher numbers show better results

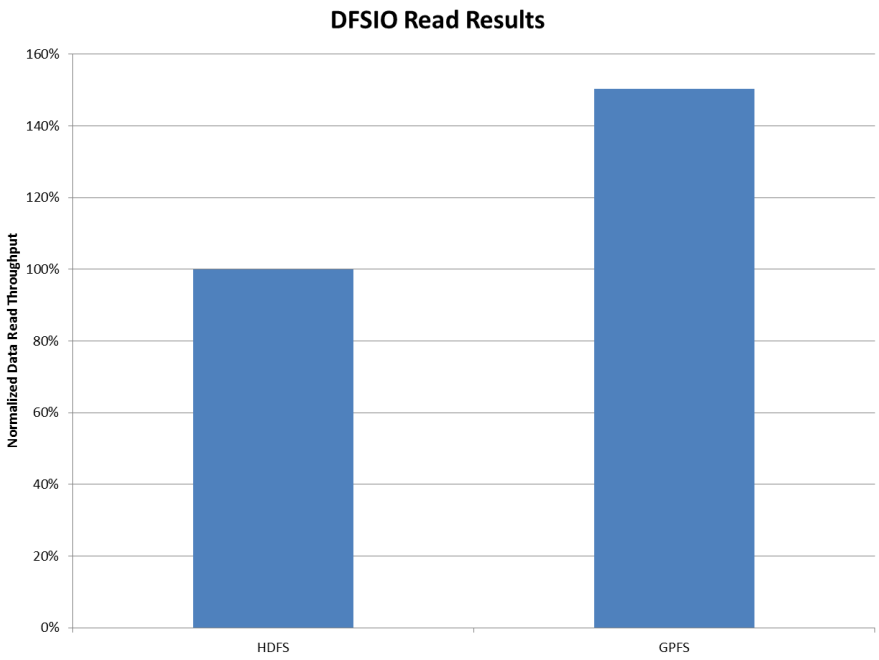

Figure 8: Read performance, higher numbers show better results

Conclusion

The E-DFSIO tests demonstrate, on the benchmark systems used, that there is a clear IO performance benefit to using GPFS instead of HDFS with a 35% performance improvement on write and a 50% performance improvement on read. The goal of this project was to measure the benefits of GPFS utilizing InfiniBand RDMA for a continuous data ingest (write) workload. In addition, GPFS performs better on the analytics portion (based on the Terasort benchmark) with 35% gain demonstrating the combined workload benefits of data retrieval, storage and data analytics.

On the benchmark system, GPFS provided better performance than HDFS. The measured performance improvements along with the fact that GPFS is POSIX compliant, contains a rich set of administrative tools and has a proven track record of reliability make it a compelling file system choice for map reduce workloads. The high performance requirements for Big Data applications are fulfilled with an infrastructure provided by Mellanox interconnects, the enterprise capabilities of IBM Biginsights and IBM GPFS.

Special Notice

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any performance data contained in this document were determined in various controlled laboratory environments and are for reference purposes only. Customers should not adapt these performance numbers to their own environments as system performance standards. The results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

For more information

http://www.mellanox.com/related-docs/prod_ib_switch_systems/SX6036.pdf

http://www.mellanox.com/page/products_dyn?product_family=161&mtag=connectx_3_pro_vpi_card

http://www.mellanox.com/pdf/products/SwitchSystem_Brochure.pdf

For more information about IBM InfoSphere BigInsights, visit:

ibm.com/software/data/infosphere/biginsights

For more information about IBM GPFS FPO, visit:

ibm.com/systems/software/gpfs

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com