White paper

A Reference Model for High Performance Data Analytics(HPDA) using an HPC infrastructure

Discover how to reshape an existing HPC infrastructure to run High Performance Data Analytics (HPDA) workloads with the performance of HPC and the usability and flexibility of a Big Data stack.

Content

Executive Summary

High Performance Data Analytics (HPDA) is an emerging technology which has been gaining increased traction and adoption in various industry segments in this cloud and IoT era. Today and in the near future, accelerated computing on data intensive workloads is not only applicable to the traditional HPC simulation and algorithm space, but also e-commerce, financial and economic segment where unstructured data is increasing and on-time fraud detection or on-time information is critical. This paper gives an insight in to the benefits of the amalgamation of the two technologies, Data Analytics and High Performance Computing (HPC), specific use cases where it is applicable and performance benchmark results substantiating the economic value of high performance data analytics.

With tremendous growth expected in this accelerated computing on data analytics, adoption of HPDA will transform organizations and the way they work by giving them the ability to obtain insights faster, increased agility with higher productivity, enabling them to stay competitive. Fujitsu's experience and expertise in providing solutions in both HPC and Big Data customer environments combined with the FUJITSU Integrated System PRIMEFLEX for HPC and FUJITSU Integrated System PRIMEFLEX for Hadoop helps us co-create flexible solutions for users HPDA requirements. This whitepaper elaborates on the overall HPC/HPDA coexistence concept, the components involved and explains the general solution architecture. The paper also expounds on a cost effective reference model which enables companies to leverage and utilize their existing HPC infrastructure to run HPDA workloads effectively.

The Emergence of High Performance Data Analytics (HPDA)

HPC offers immense potential for data-intensive business computing. However as data explodes in velocity, variety, and volume, it is increasingly difficult to scale compute performance using enterprise class servers and storage linearly. More than 80% of the data is expected to be unstructured. Hyperion forecasts that by 2025 the global data sphere will grow to 163 zettabytes (that is a trillion gigabytes). That's ten times the 16.1ZB of data generated in 2016[1] HPC is well positioned to enable Big Data use cases through all three phases of typical workflows: data capture and filtering, analytics, results visualization. In addition to the three phases, the speed of computation matters just as much as the scale. In order to unlock the full potential of Big Data, we have to pair it with "big compute", or HPC. More industry segments can benefit as much from converged Big Data and HPC used in advanced simulation workloads, where the data sets are enormous, the queries and comparisons intensive, and visualizations are complex. We already have companies in e-commerce, financial and commercial space using HPDA to obtain real-time or near real time information on security fraud and anomaly detection. There has been a constant effort in identifying the segments that would further benefit from HPDA adoption by Hyperion and other firms. And we see new segments getting added. Hyperion predicts that the HPDA market will grow robustly to a $5.43 Billion by 2020[2].

Use cases

The power of high performance computing has been used in Big Data not only in scientific simulation, but also in various other industry segments. Any use case where workloads require accelerated computing coupled with data-intensive workloads qualifies for using HPDA architecture. Some of the factors which have extended the usage are:

- More input data with higher velocity from sensors, IoT devices, application servers, databases, etc
- More output data for integration and analysis
- The need for executing smarter mathematical models and algorithms
- Real-time or near real-time requirements

Key industries where adoption of HPDA would lead to better productivity and help them stay ahead of competition are:

1) E-commerce and commercial
2) Weather and climate modeling
3) Traditional HPC environment – scientific modeling, educational and research segment

All the above mentioned industry segments would have existing HPC infrastructure resources to run their simulation workload. Hence, one of the most effective use cases has been where a customer leverages their existing high performance computing infrastructure to run non-simulation Big Data workloads to achieve the performance required for real time information analysis at the best price performance. The solution can also be easily scaled out by adding more compute resources, if required.

[1] Hyperion White Paper, sponsored by Seagate, Data Age 2025, April 2017

[2] Hyperion Future Scape: Worldwide High Performance Data Analysis 2017 Predictions

Understanding High performance Data Analytics workload

Parallel environment workloads can be of different types depending on the velocity[3], data flow, data set, information input and information to be retrieved. The combination of these factors defines the workload and the process required to obtain the best results. High performance data analytics is applicable where the workload is massively parallel in terms of data intensive and compute intensive. Data intensive workloads would consist of both structured and unstructured data from all devices, including IoT devices, sensors, etc which is high velocity data and requires global analysis.

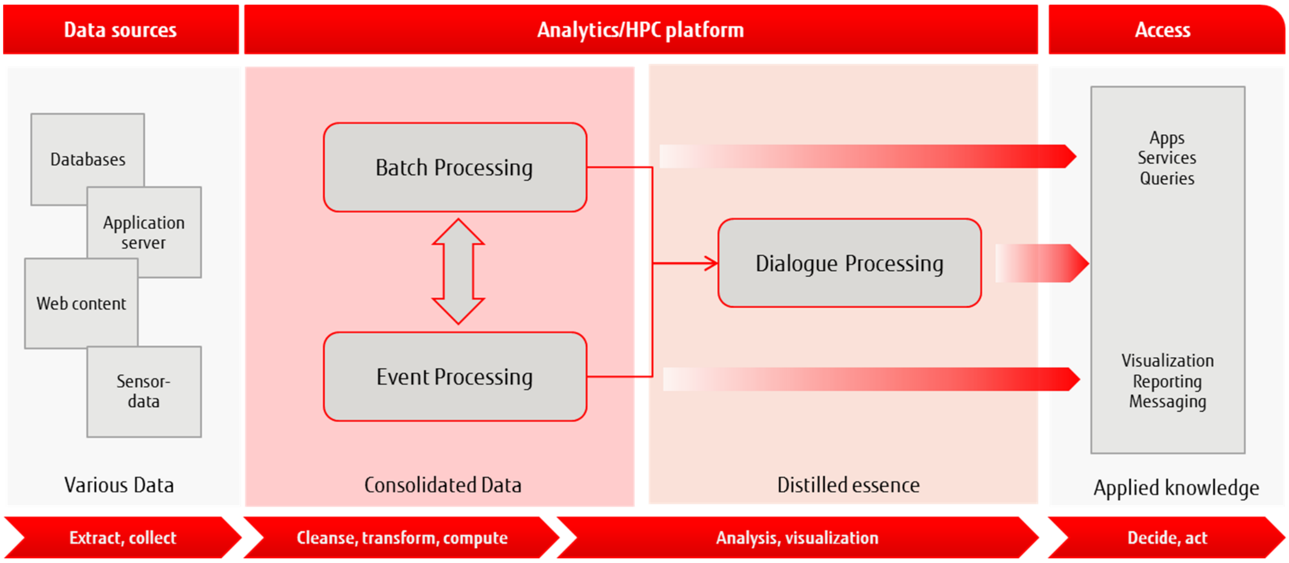
When the workload characteristics include requirements for processing small sets of high velocity input data sets and retrieving low velocity, massive data set workloads, utilizing HPC resources on a Big Data application would be more effective, providing higher performance real time results. These workloads are highly network and storage dependent as well; hence the system requirements for the workload include the following:

Driven by storage dependency:	**Driven by network dependency:**
• On node storage capability • Parallel file systems • Multi-tier storage capacity	• Remote Direct Memory Access (RDMA) capability • High speed interconnect • Capacity tuning and workload balance

High performance data analytics process flow

Utilizing HPC resources on Big Data applications like Hadoop creates a high performance data analytic configuration. In a standard Big Data application workflow, the data is collected and analyzed to get actionable insights near-real time. Time required to retrieve information is dependent on the rate at which data is collected and the complexity of analysis required. Similarly, in a standard HPC simulation workflow, the data is consolidated and collected, to which complex numerical models are run and post processing is conducted to derive the desired output.

When there is amalgamation of both HPC and Big Data technology, the platform would have the power of HPC resources to run complex workloads on a Big data application, processing and storing large data sets. The process flow is as shown in the chart below:



Why and When to implement an HPDA architecture

When it comes to the data analytics solutions available in the market, Hadoop and Spark solutions have become synonymous with these. In this section we are essentially trying to explain when it makes sense to stick with a standard data analytic solution and when using a HPDA architecture makes sense.

[3] Velocity of data is essentially the rate at which data is generated

As mentioned, one of the main tools is Hadoop, an open source Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the project sponsored by the Apache Software Foundation. Hadoop also leverages the Map reduce framework (developed by Google), which provides the scheduling, distribution, and parallelization services. The MapReduce framework gets input data from the Hadoop Distributed File System (HDFS). This provides rapid data transfer rates between nodes. However, in comparison to HPC parallel file systems like FUJITSU Software FEFS, a Parallel File System or Lustre or FEFS it is rather slower. They also have the ability to efficiently support the random reading of small files more efficiently than HDFS. HDFS works well for large datasets. The standard Big Data infrastructure has limitations in scaling concurrent jobs whilst keeping the performance level intact and additionally it does not support scalable network topologies like multi-dimensional meshes which can help boost performance. Meanwhile, HPC production platforms can process thousands of jobs every hour without compromising performance.

The other solution is Spark, which is a second generation tool from the Apache project. It is a framework for performing general data analytics on distributed computing clusters like Hadoop. It provides in memory computations for increased speed and data processing over MapReduce, hence it's a higher performing solution. However, the solution is heavy on memory resources for workloads which require real time results on high velocity data sets. The expected productivity required, cannot be obtained by simply re-writing Hadoop's communication system in message passing interface (MPI), supporting Lustre in Hadoop, or implementing InfiniBand communication layer in Hadoop. Hence by combining HPC hardware, software resources and Hadoop or Spark brings Big Data application to high performance computing configurations, providing high performance, agile and scalable solution. Combing best of the two worlds, High performance data analytics has the greatest potential to resolve the computation speed and scale required for these workloads.

Fujitsu has extensive experience in providing solutions for data analytics with the FUJITSU Integrated System PRIMEFLEX for Hadoop solution. Hence we recognize the value of Hadoop in providing a cost effective data analytics solutions. However, this whitepaper focuses on users with an existing HPC infrastructure, showing how they can leverage that infrastructure to construct a scalable, agile, highly performant HPDA environment.

Fujitsu's Approach

The HPC world is moving towards the concept of dynamic deployments, which will enable data analytic frameworks to be quickly configured and run on supercomputers, and then torn down to make way for something else. This usage model will not be confined specifically to supercomputers. However, this will involve multiple systems using multiple frameworks to get the answer.

Fujitsu's approach is to co-create a solution to let the users integrate their analytic pipeline into the HPC workflow, utilizing the existing HPC infrastructure, with smart data movement between the different software components. It is not just a dynamic deployment of the framework but how to execute the HPC production workload and get all components coordinated, especially with the data movement and data staging.

The approach includes:

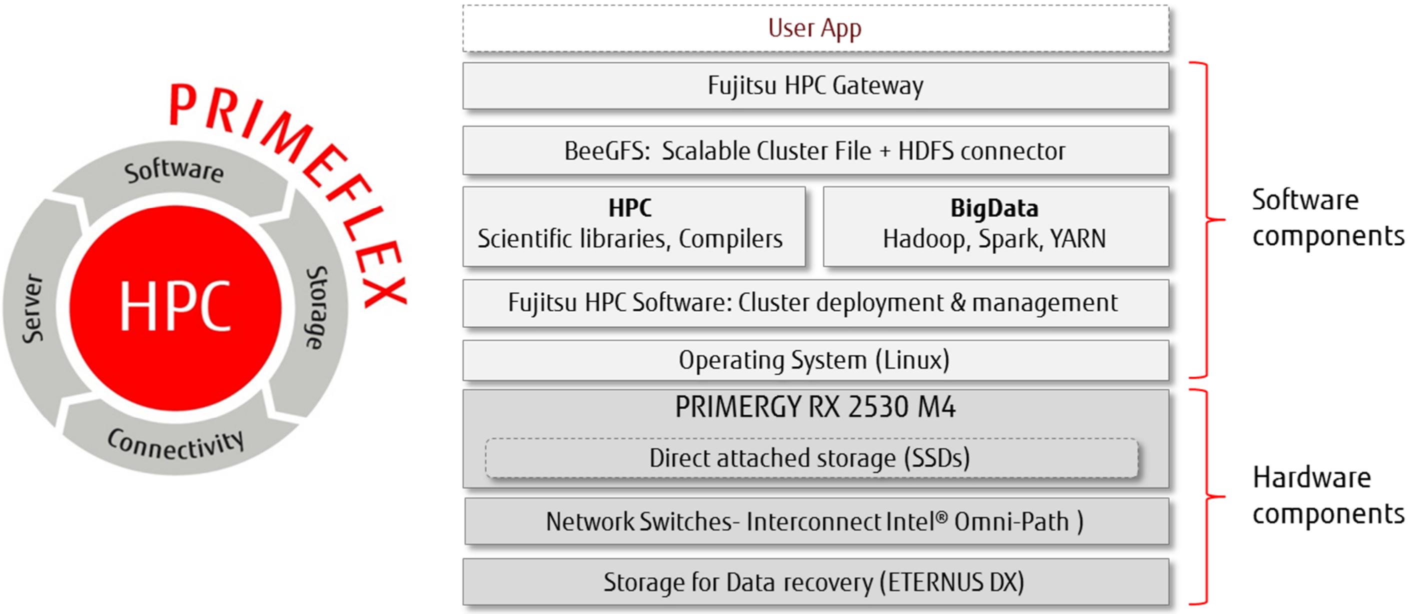
- Identify technologies to:
  - Build a platform architecture by utilizing existing infrastructure resources ,focusing on simplicity and the best price / performance ratio
  - Support for data distribution across intra and inter-node memory hierarchies
  - Support the combination of HPC workloads, including serial to massively parallel applications, simultaneously with the more structured process allocation in HPDA applications
  - Identify the processes needed to make existing applications run effectively on systems with large scale many-core processors
- Maintain portability of the solution
  - It is mandatory to ensure that the benefits of these efforts are not limited to the current platforms but can be re-used and expanded for future generations
  - Capitalize on architecture design and system optimization techniques and internal "know-how" of Fujitsu engineers with high end computing techniques
  - Build a software stack enforcing language/runtime interoperability as well as the most efficient production environment
- Capitalize on Fujitsu's years of business experience with both Integrated systems on Big Data Analytics and HPC
  - Strong link and cooperation with major HPC players globally.
  - Market recognition of Fujitsu as a major player in Big Data solutions.
  - With the Fujitsu Dynamic Infrastructures approach, Fujitsu offers a full portfolio of IT products, solutions and services, ranging from clients to data center solutions, Managed Infrastructure and Infrastructure as a Service.

Reference Model

HPDA Reference Model represents the smart integration and performance advantages of an HPC infrastructure along with Big Data and data analytic technology. If there is an existing HPC infrastructure or FUJITSU Integrated System PRIMEFLEX for HPC running, the HPDA reference model could be built on the existing infrastructure resources and scaled further based on requirement. Or a new PRIMEFLEX for HPC –HPDA infrastructure could be built from scratch. In both these solutions, a traditional HPC cluster is augmented with Hadoop tools needed for Big Data and Data Analytic processing, enabling traditional HPC workloads to co-exist with HPDA processes. In addition, an HPC PFS (parallel file system) is configured with the HDFS connector to enable HPDA applications to seamlessly access data alongside HPC applications.

Such approach enables construction of agile systems which deliver effective utilization of both HPC and Data Analytics workloads, leading to optimized TCO and improved ROI.

High Performance Data Analytics integrated solution stack


PRIMEFLEX
Software
Server
HPC
Storage
Connectivity
User App
Fujitsu HPC Gateway
BeeGFS: Scalable Cluster File + HDFS connector
HPC
Scientific libraries, Compilers
BigData
Hadoop, Spark, YARN
Fujitsu HPC Software: Cluster deployment & management
Operating System (Linux)
PRIMERGY RX 2530 M4
Direct attached storage (SSDs)
Network Switches- Interconnect Intel® Omni-Path)
Storage for Data recovery (ETERNUS DX)
Software components
Hardware components

Compute nodes

The need is to run both complex and simple computation on a set of flexible HPC cluster nodes. The Fujitsu PRIMERGY RX2530 1U dual processor server is an optimal workhorse server which brings with it great flexibility and expandability. With 24 DIMMs and broad support for the latest Intel® XEON processors, the RX2530 is ideally placed to support a variety of workloads and ensures it can scale to meet the most demanding processing ad memory requirements.

Direct attached storage for HPDA processing

SSDs are preferred to ensure the storage performance is met. Intel SSDs are used for local storage on compute nodes. The SSD shows a very positive impact on the results, both with standard file system and parallel file system. Both structured and unstructured data processing is done in HPDA architecture. Recommended SSD size to memory ratio is 3:1 and the ratio of SSD size per memory per core is 33.33 for optimal performance.

File System

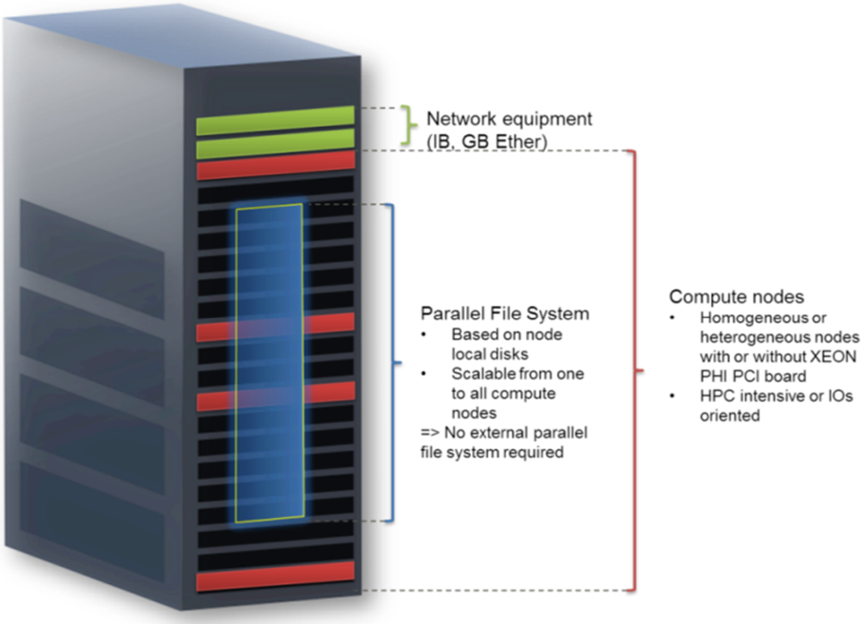
With High Performance Data Analytics solution, a single parallel file system based on BeeGFS is configured with an HDFS connector function to offer the best possible performance to both HPC and HPDA workloads. Data redundancy is a not standard feature of the HPC parallel file system therefore we have included a Permanent project storage layer within the reference architecture.
Data needing to be retained is copied from the local PFS to the permanent project store layer.

Network

The high speed interconnect of the HPC architecture based on InfiniBand/Omni-Path is used to ensure inter node communication is maximized and data movement to the permanent project storage is achieves the highest throughput.
By using such an efficient inter-node link one can continue to scale the Big Data problem knowing it will execute as efficiently as possible.

Cluster resource management

Fujitsu HPC Software stack is the recommended software solution for cluster management and monitoring. The solution helps simplify the installation, management, and ongoing maintenance of the overall stack. The choice of either SLURM or PBS Pro with Hadoop integration as the batch job manager can be made.


Network equipment (IB, GB Ether)
Parallel File System
• Based on node local disks
• Scalable from one to all compute nodes
=> No external parallel file system required
Compute nodes
• Homogeneous or heterogeneous nodes with or without XEON PHI PCI board
• HPC intensive or IOs oriented

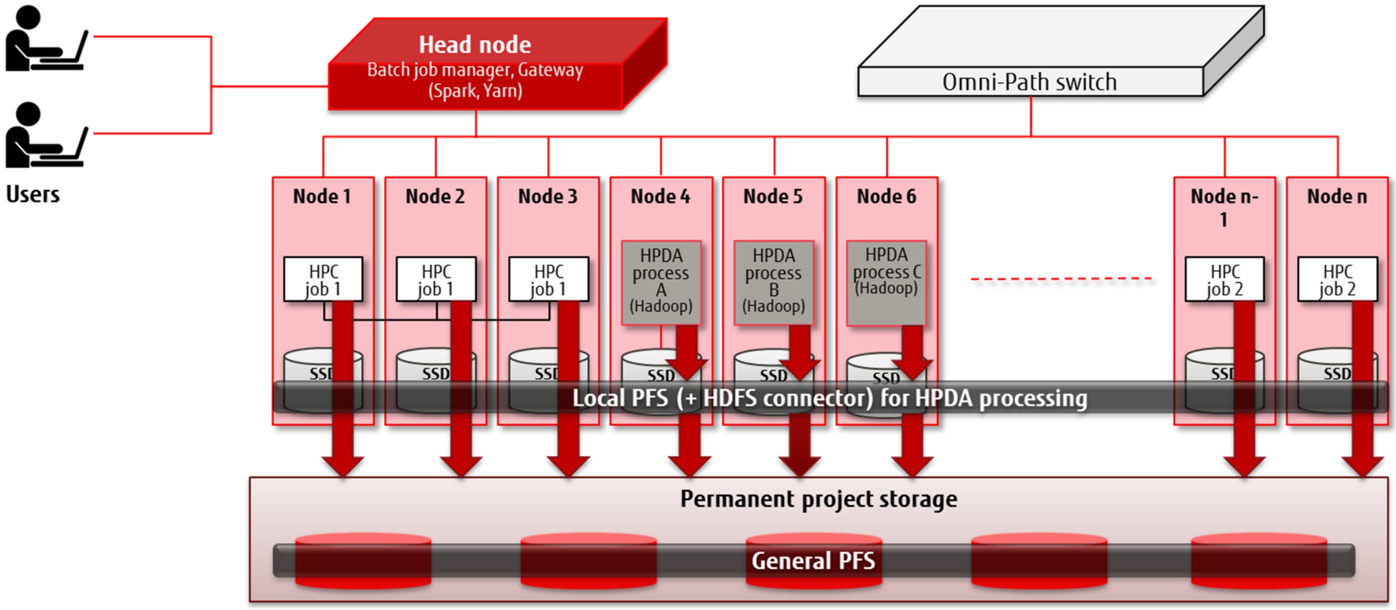
Gateway

FUJITSU's HPC Gateway brings the simplicity of web access, the expertise from agile application workflows and clarity of active reporting to tune business processes and better management of project outcomes. The intuitive web user interface includes direct simulation monitoring, data access and collaboration. New and occasional users as well as practiced HPC users will find the interface for preparing, launching and monitoring their work highly effective. Combined with the pre-built packages from the Application Catalogue, they will have at hand the most productive and coherent HPC workplace in the market today.

General System architecture – A shared HPC and HPDA environment

The fundamental architecture used for the HPC and HPDA combined solution is shown. The main attributes of the HPDA components are the integration of the job submission on the head node with the batch system enabling HPDA users the ability to submit work directly to the HPC infrastructure.

A local PFS is configured as an HDFS compatible file system, providing high speed access to a data for HPDA processes. A permanent storage layer is also built with a PFS to enable both HPC and HPDA jobs a location for retaining long term data.



Proposed configurations for various data sizes

A selection of recommended configurations based on medium to extra-large data sizes is given below. These system configurations can be considered as guidelines which implement optimal configurations with best price performance.

Reference Hardware components

Item	Configuration 1 Medium	Configuration 2 Large	Configuration 3 Extra Large
Data Size	50TB	100TB	200TB
Cluster size	8 nodes	12 nodes	16 nodes
Node type	PRIMERGY RX2530		
Cores per node	24 cores	28 cores	32 cores
Cluster cores	192	336	512
Memory	1.5 TB	3.4 TB	6 TB
Direct Attached Storage (SSD)	50TB	112.5TB	200TB
SSD to RAM ratio	33:1		
Interconnect switches	Intel Omni-Path 100Gb/s single switch		
Permanent storage	2x DX200	4x DX200	8x DX200
Fill 50% of SSD 1:10 hours	200TB	600TB	1600TB

Reference Software components

Item	Software component
Cluster software	Fujitsu HPC Software Stack
Batch Job Manager	SLURM or PBS Professional or PBS Professional OSS or PBS Professional
cal storage layer	BeeGFS or Lustre
PFS support for HDFS	For BeeGFS: BeeGFS HDFS connector For Lustre: HDFS connector
Data Analytics engines	Hadoop, Spark, YARN
Batch integration (for running YARN processes as batch jobs)	Scheduling connector for Hadoop

Advantage of HPDA Reference Model

1) Non-intrusive deployment on existing HPC platform
   - The additional packages are compatible with standard open SSF or Fujitsu HCS cluster management tools
   - Normal HPC operation is kept un-interrupted
   - Agile infrastructure which could be easily scaled based on the requirement
2) Allow to run both intensive compute and/or data analytic through the same cluster management tool (SLURM)
   - Standard batch jobs under the control of SLURM scheduler with fair share of the resource
3) The recommended architecture provides high end performance of the HPC platform to speed up Hadoop, with both high speed interconnect and parallel file system which brings unprecedented performance to boost efficiency

Performance evidence

To prove the performance benefits of this solution we ran a standard TeraSort benchmark on identical hardware but varying from a standard SPARK / HDFS environment to one using SLURM and BeeGFS. This follows the guidelines of the reference model.The environment used a data analytics scheduling connector to the Slurm job batch manager to manage the data analytics processes like a HPC application as well as BeeGFS with the HDFS connector for the local PFS storage which was responsible for the data storage component, replacing HDFS.

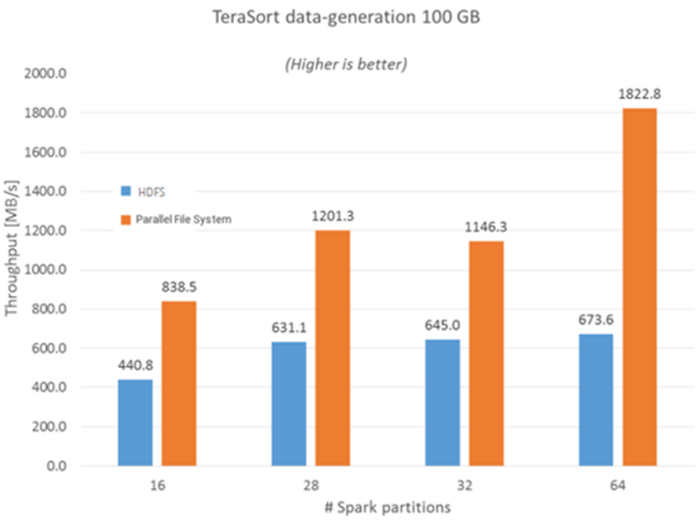
The goal of the TeraSort benchmark is to sort massive amounts of data as fast as possible. It is a benchmark that combines testing for the storage and HDFS as well as MapReduce layers of the platform.

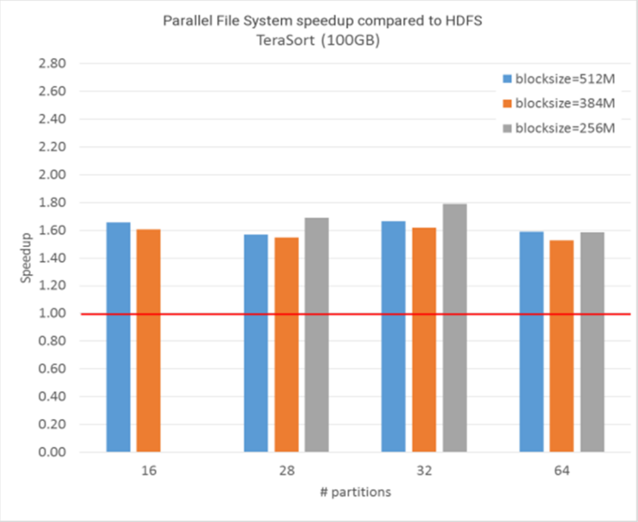
The reference model solution shows a far better throughput than the standard HDFS configuration, dramatically reducing the needed generation and analysis time.

This benchmark ran on a Fujitsu PRIMERGY reference platform consisting of 8x compute nodes with dual socket mother boards equipped with Intel Broadwell processors and 128GB main memory. Storage devices used were standard 400 GB Intel SSDs, with the parallel file system built using the 8 compute nodes as data servers.

Phase 1 - Data generation
Data generation is the first phase of the TeraSort benchmark. It generates random data that can be conveniently used as input data for a subsequent TeraSort run. Using the local parallel file system clearly provides a boost here, even if this step only generates random data to be used at the second stage of the sort program.

As shown in the diagram (right), when using an identical hardware platform the parallel file system solution exhibits far better data throughput than the standard HDFS solution, leading to dramatically reduced generation time.





Phase 2 - Data Sort
The second step of the TeraSort Benchmark consists in sorting all randomly generated data. With a very coherent behavior the clear advantage of using the local parallel file system was reflected during the sort performance evaluation.
Similarly to the results of Phase 1 the parallel file system solution contributed to a significant performance boost for the Phase 2 results. The diagram (left) shows a consistent speedup of about 1.6 times compared to the standard HDFS storage solution.

These results confirm the validity and business benefits for the moderate investment needed to implement a HPDA environment on top of an existing HPC infrastructure.

Conclusion

The use of accelerated computing or high performance computing on advanced commercialized data analytics has been increasing and expected to grow exponentially. In simple terms these are Big Data workloads that are complex and time critical and hence benefit from using HPC resources to obtain fast communication among processing elements and RAM as well as data storage access times. Utilize your existing HPC infrastructure and resources, without disrupting simulation workloads to run Hadoop or Spark Big Data applications to obtain the best of both worlds. This solution makes use of the hardware and software designed to run complex models and provide high performance combined with Big Data application capable of running large data sets and storage to deliver the most insightful, actionable results, real time.
A number of research firms working with government and commercial organizations have already reported they have saved millions of dollars per year by using HPC resources for advanced analytics. Take advantage of FUJITSU's years of experience working with Enterprises on their Big Data and High Performance Computing workloads. FUJITSU Integrated System PRIMEFLEX for HPC provides an ideal scalable solution to combine HPC and Data Analytics workloads on a single HPC infrastructure.