Lenovo Big Data Reference Architecture for Hortonworks Data Platform Using System x Servers

Last update: 12 December 2017
Version 1.1
Configuration Reference Number: BDAHWKSXX62

Describes the RA for Hortonworks Data Platform, powered by Apache Hadoop and Apache Spark

Solution based on the powerful, versatile Lenovo System x3650 M5 server

Deployment considerations for high-performance, cost-effective and scalable solutions

Contains detailed bill of material for different servers and associated networking

Dan Kangas
Weixu Yang
Ajay Dholakia
Brian Finley

LENOVO PRESS

Table of Contents

1 Introduction

This document describes the reference architecture for Hortonworks Data Platform (HDP), a distribution of Apache Hadoop with enterprise-ready capabilities. It provides a predefined and optimized Lenovo hardware infrastructure for the Hortonworks Data Platform. The intended audience is IT professionals, technical architects, sales engineers, and consultants to assist in planning, designing, and implementing the Hortonworks big data solution using Lenovo hardware. It is assumed that you are familiar with Hadoop components and capabilities. For more information about Hadoop, see “Resources” on page 40.

Lenovo and Hortonworks worked together on this document and the reference architecture that is described herein was validated together by Lenovo and Hortonworks. This predefined configuration provides a baseline for a big data solution, which can be modified, based on the specific customer requirements, such as lower cost, improved performance, and increased reliability.

The Hortonworks Data Platform, powered by Apache Hadoop, is a highly scalable and fully open source platform for storing, processing and analyzing large volumes of structured and unstructured data. It is designed to deal with data from many sources and formats in a very quick, easy and cost-effective manner. Hortonworks expands and enhances this technology to withstand the demands of your enterprise, adding management, security, governance, and analytics features. The result is that you obtain a more enterprise ready solution for complex, large-scale analytics.

2 Business problem and business value

This section describes the business problem that is associated with big data environments and the value that is offered by the Hortonworks Data Platform solution and Lenovo hardware.

2.1 Business problem

By 2012, the world generated 2.5 million terabytes (TB) of data, daily - a level that is expected to increase to 44 zettabytes (44 trillion gigabytes by 2020). In all, 90% of the data in the world today was created in the last two years alone. This data comes from everywhere including posts to social media sites, digital pictures and videos, purchase transaction records, cell phone GPS signals, and from sensors used to gather climate information. This data is big data!!

Big data spans the following dimensions:

- **Volume**: Big data is enormous – in size, quantity and/or scale. Enterprises are awash with data, easily amassing terabytes and even petabytes of information.
- **Velocity**: Often time-sensitive, big data must be used as it is streaming into the enterprise to maximize its value to the business.
- **Variety**: Big data extends beyond structured data, including unstructured data of all varieties, such as text, audio, video, click streams, and log files.

Big data is more than a challenge; it is an opportunity to find insight into new and emerging types of data to make your business more agile. Big data also is an opportunity to answer questions that, in the past, were beyond reach. Until now, there was no effective way to harvest this opportunity. Today, Hortonworks uses the latest big data, fully open sourced technologies such as the Apache SPARK in-memory processing capabilities in addition to the standard MapReduce scale-out capabilities, and all based on a centralized architecture (YARN) to open the door to a world of possibilities.

2.2 Business value

Hadoop is used to reliably manage and analyze large volumes of structured and unstructured data. Hortonworks enhances this technology by adding management, security, governance, and analytics features.

How can businesses process tremendous amounts of raw data in an efficient and timely manner to gain actionable insights? Hortonworks allows organizations to run large-scale, distributed analytics jobs on clusters of cost-effective server hardware. This infrastructure can be used to tackle large data sets by breaking up the data into “chunks” and coordinating data processing across a massively parallel environment. After the raw data is stored across the nodes of a distributed cluster, queries and analysis of the data can be handled efficiently, with dynamic interpretation of the data formatted at read time. The bottom line: businesses can finally grasp massive amounts of untapped data and mine that data for valuable insights in a more efficient, optimized, and scalable way.

Hortonworks HDP that is deployed on Lenovo System x servers with Lenovo networking components provides superior performance, reliability, and scalability. The reference architecture supports entry through high-end configurations and the ability to easily scale as the use of big data grows. A choice of infrastructure components provides flexibility in meeting varying big data analytics requirements.

3 Big Data Requirements

The functional and non-functional requirements for this reference architecture are desribed in this section.

3.1 Functional requirements

A big data solution must support the following key functional requirements:

- Ability to handle various workloads, including batch and real-time analytics
- Industry-standard interfaces so that applications can work together seamlessly
- Ability to handle large volumes of unstructured, structured and semi-structured data
- Support a variety of client interfaces

3.2 Non-functional requirements

Customers require their big data solution to be easy, dependable, and fast. The following non-functional requirements are key:

- Easy:
  - Ease of development
  - Easy management at scale
  - Advanced job management
  - Multi-tenancy
  - Easy to access data by various user types
- Dependable:
  - Data protection with snapshot and mirroring
  - Automated self-healing
  - Insight into software/hardware health issues
  - High Availability (HA) and business continuity
- Fast:
  - Superior performance
  - Scalability
- Secure and governed:
  - Strong authentication and authorization
  - Kerberos support
  - Data confidentiality and integrity

4 Architectural overview

Figure 1 shows the main features of the Hortonworks reference architecture that uses Lenovo hardware. Users can log into the Hortonworks client-side from outside the firewall by using Secure Shell (SSH) on port 22 to access the Hortonworks Utility Machines from the corporate network. Hortonworks provides several interfaces that allow administrators and users to perform administration and data functions, depending on their roles and access level. Hadoop application programming interfaces (APIs) can be used to access data. Hortonworks APIs can be used for cluster management and monitoring.

Hortonworks data services, management services, and other services run on the nodes in cluster. Storage is a component of each data node in the cluster. Data can be incorporated into Hortonworks storage through the Hadoop APIs or network file system (NFS), depending on the needs of the customer.

A database is required to store the data for Ambari, hive metastore, and other services. Hortonworks provides an embedded database for test or proof of concept (POC) environments and an external database is required for a supportable production environment.

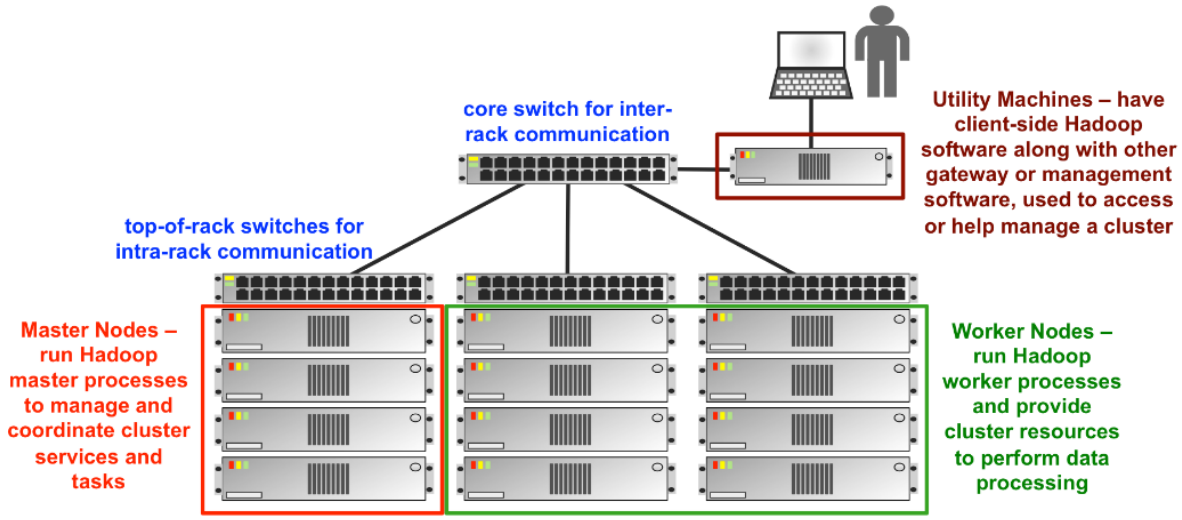

Figure 1. Hortonworks Architecture Overview

5 Component model

Hortonworks Data Platform provides features and capabilities that meet the functional and nonfunctional requirements of customers. It supports mission-critical and real-time big data analytics across different industries, such as financial services, retail, media, healthcare, manufacturing, telecommunications, government organizations, and leading Fortune 100 and Web 2.0 companies.

Hortonworks Data Platform is the industry's only truly secure, enterprise-ready, open source Apache Hadoop distribution based on a centralized architecture (YARN). It addresses the complete needs of “data-at-rest”, it powers real-time customer applications and it delivers robust analytics that accelerate decision making and innovation.

The Hortonworks Data Platform for big data can be used for various use cases from batch applications that use MapReduce or Spark with data sources, such as click streams, to real-time applications that use sensor data.

Figure 2 shows the Hortonworks Hadoop collection of software frameworks which make up the Hortonworks distribution of Apache Hadoop. Many of these Hadoop components are optional and provide specific functions to meet the requirements of customers.

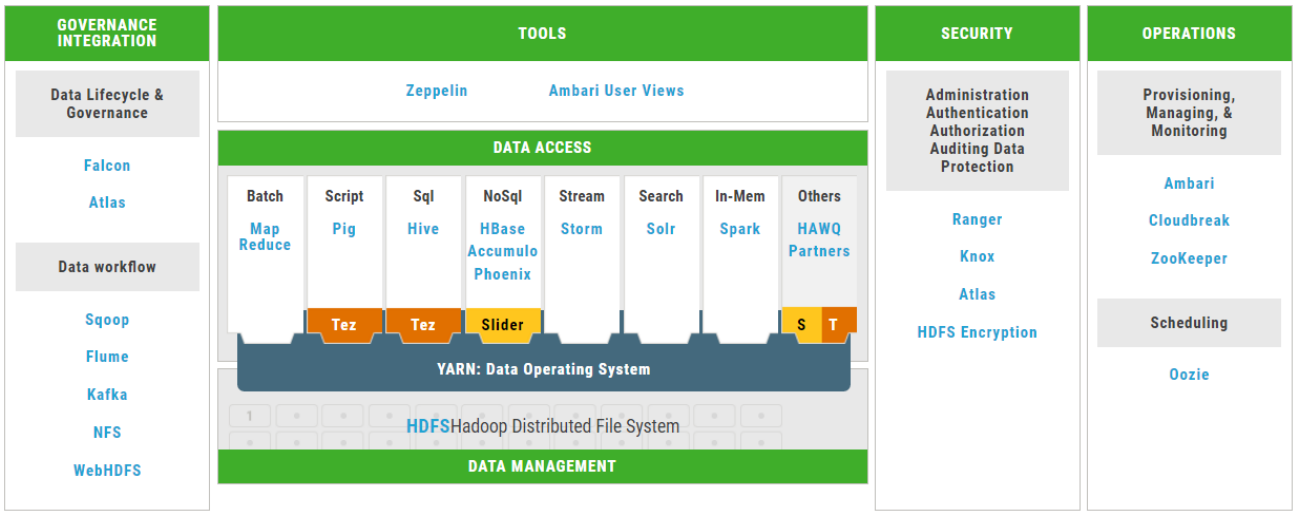

Figure 2. Hortonworks Hadoop Collection of Software Frameworks

Hortonworks Data Platform contains the following components:

Data Management Components

YARN and Hadoop Distributed File System (HDFS) are the cornerstone components of Hortonworks Data Platform. While HDFS provides the scalable, fault-tolerant, cost-efficient storage for your big data lake, YARN provides the centralized architecture that enables you to process multiple workloads simultaneously. YARN provides the resource management and pluggable architecture for enabling a wide variety of data access methods.

Data Access Components

Hortonworks Data Platform includes a versatile range of processing engines that empower you to interact with the same data in multiple ways, at the same time. This means applications can interact with the data in the best way: from batch to interactive SQL or low latency access with NoSQL. Emerging use cases for data science, search and streaming are also supported with Apache Spark, Storm and Kafka. Other components include: Hive, Tez, Pig, Hbase and Accumulo.

Data Governance & Integration Components

HDP extends data access and management with powerful tools for data governance and integration. They provide a reliable, repeatable, and simple framework for managing the flow of data in and out of Hadoop. This control structure, along with a set of tooling to ease and automate the application of schema or metadata on sources is critical for successful integration of Hadoop into your modern data architecture. The components include: Atlas, Falcon, Oozie, Scoop, Flume and Kafka.

Security Components

Security is woven and integrated into HDP in multiple layers. Critical features for authentication, authorization, accountability and data protection are in place to help secure HDP across these key requirements. Consistent with this approach throughout all of the enterprise Hadoop capabilities, HDP also ensures you can integrate and extend your current security solutions to provide a single, consistent, secure umbrella over your modern data architecture. These components include Knox, Ranger and Ranger KMS.

Operations Components

Operations teams deploy, monitor and manage a Hadoop cluster within their broader enterprise data ecosystem. Apache Ambari simplifies this experience. Ambari is an open source management platform for provisioning, managing, monitoring, and securing the Hortonworks Data Platform. It enables Hadoop to fit seamlessly into your enterprise environment. These components include Ambari and Zookeeper.

Cloud Component

Cloudbreak, as part of Hortonworks Data Platform and powered by Apache Ambari, allows you to simplify the provisioning of clusters in any cloud environment including Amazon Web Services and Microsoft Azure. It optimizes your use of cloud resources as workloads change.

Spark

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. With Spark running on Apache Hadoop YARN, developers everywhere can now create applications to exploit Spark's power, derive insights, and enrich their data science workloads within a single, shared dataset in Hadoop.

The Hadoop YARN-based architecture provides the foundation that enables Spark and other applications to share a common cluster and dataset while ensuring consistent levels of service and response. Spark is now one of many data access engines that work with YARN in HDP.

Spark is designed for data science and its abstraction makes data science easier. Data scientists commonly use machine learning – a set of techniques and algorithms that can learn from data. These algorithms are often iterative, and Spark's ability to cache the dataset in memory greatly speeds up such iterative data processing, making Spark an ideal processing engine for implementing such algorithms.

For more information on all of the Hortonworks HDP Projects, see the following website:

http://hortonworks.com/apache/

The Hortonworks solution is operating system independent. Hortonworks HDP 2.5 supports many 64-bit Linux operating systems:

- Red Hat Enterprise Linux (RHEL),
- CentOS
- Debian
- Oracle Linux
- SUSE Linux Enterprise Server (SLES)
- Ubuntu.

For more information about the versions of supported operating systems, see this website:

https://docs.hortonworks.com/HDPDocuments/Ambari-2.2.1.1/bk_Installing_HDP_AMB/content/_operating_systems_requirements.html

6 Operational model

This section describes the operational model for the Hortonworks reference architecture. To show the operational model for different sized customer environments, four different models are provided for supporting different amounts of data. Throughout the document, these models are referred to as half rack, full rack, and multi-rack configuration sizes. The multi-rack is three times larger than the full rack.

A Hortonworks Data Platform deployment consists of cluster nodes, networking equipment, power distribution units, and racks. The predefined configurations can be implemented as is or modified based on specific customer requirements, such as lower cost, improved performance, and increased reliability. Key workload requirements, such as the data growth rate, sizes of datasets, and data ingest patterns help in determining the proper configuration for a specific deployment. A best practice when a Hortonworks cluster infrastructure is designed is to conduct the proof of concept testing by using representative data and workloads to ensure that the proposed design works.

6.1 Hardware description

This reference architecture uses Lenovo System x3650 M5 and x3550 M5 servers and Lenovo RackSwitch G8052 and G8272 top of rack switches.

6.1.1 Lenovo System x3650 M5 Server - Data Node

The Lenovo System x3650 M5 server (as shown in Figure 3 Figure 3) is an enterprise class 2U two-socket versatile server that incorporates outstanding reliability, availability, and serviceability (RAS), security, and high-efficiency for business-critical applications and cloud deployments. It offers a flexible, scalable design and simple upgrade path to 26 2.5-inch hard disk drives (HDDs) or solid-state drives (SSDs), or 14 3.5-inch HDDs with doubled data transfer rate through 12 Gbps serial-attached SCSI (SAS) internal storage connectivity and up to 1.5 TB of TruDDR4 Memory. On-board it provides four standard embedded Gigabit Ethernet ports and two optional embedded 10 Gigabit Ethernet ports without occupying PCIe slots.

Figure 3. Lenovo System x3650 M5

Combined with the Intel Xeon processor E5-2600 v4 product family, the Lenovo x3650 M5 server offers an even higher density of workloads and performance that lowers the total cost of ownership (TCO). Its pay-as-you-grow flexible design and great expansion capabilities solidify dependability for any kind of workload with minimal downtime.

The x3650 M5 server provides internal storage density of up to 128 TB in a 2U form factor with its impressive array of workload-optimized storage configurations. It also offers easy management and saves floor space and power consumption for most demanding use cases by consolidating storage and server into one system.

The reference architecture recommends the storage-rich System x3650 M5 model for the following reasons:

- Storage capacity: The nodes are storage-rich. Each of the 14 configured 3.5-inch drives has raw capacity up to 8 TB and each, providing for 112 TB of raw storage per node and over 1000 TB per rack.
- Performance: This hardware supports the latest Intel Xeon processors and TruDDR4 Memory.
- Flexibility: Server hardware uses embedded storage, which results in simple scalability (by adding nodes).
- More PCIe slots: Up to 8 PCIe slots are available if rear disks are not used, and up to 2 PCIe slots if both Rear 3.5-inch HDD Kit and Rear 2.5-inch HDD Kit are used. They can be used for network adapter redundancy and increased network throughput.
- Better power efficiency: Innovative power and thermal management provides energy savings.
- Reliability: Lenovo is first in the industry in reliability and has exceptional uptime with reduced costs.

For more information, see the Lenovo System x3650 M5 Product Guide:

https://lenovopress.com/lp0068-lenovo-system-x3650-m5-e5-2600-v4

6.1.2 Lenovo RackSwitch G8272 - Data Switch

Designed with top performance in mind, Lenovo RackSwitch G8272 is ideal for today's big data, cloud, and optimized workloads. The G8272 switch offers up to 72 10 Gb SFP+ ports in a 1U form factor and is expandable with six 40 Gb QSFP+ ports. It is an enterprise-class and full-featured data center switch that delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center grade buffers keep traffic moving. Redundant power and fans and numerous HA features equip the switches for business-sensitive traffic.

The G8272 switch (as shown in Figure 4) is ideal for latency-sensitive applications. It supports Lenovo Virtual Fabric to help clients reduce the number of I/O adapters to a single dual-port 10 Gb adapter, which helps reduce cost and complexity. The G8272 switch supports the newest protocols, including Data Center Bridging/Converged Enhanced Ethernet (DCB/CEE) for support of FCoE and iSCSI and NAS.

Figure 4. Lenovo RackSwitch G8272

The enterprise-level Lenovo RackSwitch G8272 has the following characteristics:

- 48 x SFP+ 10GbE ports plus 6 x QSFP+ 40GbE ports
- Support up to 72 x 10Gb connections using break-out cables
- 1.44 Tbps non-blocking throughput with low latency (~ 600 ns)
- Up to 72 1Gb/10Gb SFP+ ports
- OpenFlow enabled allows for easily created user-controlled virtual networks

For more information, see the Lenovo RackSwitch G8272 Product Guide:

https://lenovopress.com/tips1267-lenovo-rackswitch-g8272

6.1.3 Lenovo System x3550 M5 Server - Management Node

The Lenovo System x3550 M5 server (as shown in Figure 5) is a cost- and density-balanced 1U two-socket rack server. The x3550M5 features a new, innovative, energy-smart design with up to two Intel Xeon processors of the high-performance E5-2600 v4 product family processors a large capacity of faster, energy-efficient TruDDR4 Memory, up to twelve 12Gb/s SAS drives, and up to three PCI Express (PCIe) 3.0 I/O expansion slots in an impressive selection of sizes and types. The server's improved feature set and exceptional performance is ideal for scalable cloud environments.

Figure 5. Lenovo System x3550 M5

For more information, see the Lenovo System x3550 M5 Product Guide:
https://lenovopress.com/lp0067-lenovo-system-x3550-m5-e5-2600-v4

6.1.4 Lenovo RackSwitch G8052 - Management Switch

The Lenovo System Networking RackSwitch G8052 (as shown in Figure 6) is an Ethernet switch that is designed for the data center and provides a simple network solution. The Lenovo RackSwitch G8052 offers up to 48 1 GbE ports and up to 4 10 GbE ports in a 1U footprint. The G8052 switch is always available for business-critical traffic by using redundant power supplies, fans, and numerous high-availability features.

Figure 6. Lenovo RackSwitch G8052

Lenovo RackSwitch G8052 has the following characteristics:

- A total of 48 1 GbE RJ45 ports
- Four standard 10 GbE SFP+ ports
- Low 130W power rating and variable speed fans to reduce power consumption

For more information, see the Lenovo RackSwitch G8052 Product Guide:
https://lenovopress.com/tips1270-lenovo-rackswitch-g8052

6.1.5 Lenovo RackSwitch G8272 - Data Switch

Designed with top performance in mind, Lenovo RackSwitch G8272 is ideal for today's big data, cloud, and optimized workloads. The G8272 switch offers up to 72 10 Gb SFP+ ports in a 1U form factor and is expandable with six 40 Gb QSFP+ ports. It is an enterprise-class and full-featured data center switch that delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center grade buffers keep traffic moving. Redundant power and fans and numerous HA features equip the switches for business-sensitive traffic.

The G8272 switch (as shown in Figure 7) is ideal for latency-sensitive applications. It supports Lenovo Virtual Fabric to help clients reduce the number of I/O adapters to a single dual-port 10 Gb adapter, which helps reduce cost and complexity. The G8272 switch supports the newest protocols, including Data Center Bridging/Converged Enhanced Ethernet (DCB/CEE) for support of FCoE and iSCSI and NAS.

Figure 7. Lenovo RackSwitch G8272

The enterprise-level Lenovo RackSwitch G8272 has the following characteristics:

- 48 x SFP+ 10GbE ports plus 6 x QSFP+ 40GbE ports
- Support up to 72 x 10Gb connections using break-out cables
- 1.44 Tbps non-blocking throughput with low latency (~ 600 ns)
- Up to 72 1Gb/10Gb SFP+ ports
- OpenFlow enabled allows for easily created user-controlled virtual networks

For more information, see the Lenovo RackSwitch G8272 Product Guide:

https://lenovopress.com/tips1267-lenovo-rackswitch-g8272

6.1.6 Lenovo RackSwitch G8316 - Cross-Rack Switch

The RackSwitch G8316 (as shown in Figure 8) is a 40 Gigabit Ethernet (GbE) switch that is designed for the data center, providing speed, intelligence, and interoperability on a proven platform. It is an ideal aggregation class switch for connecting multiple RackSwitch G8264 class switches with their 40Gb uplink ports.. RackSwitch G8316 provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center grade buffers keep traffic moving. Hot-swappable, redundant power and fans, along with numerous high-availability features, enable the RackSwitch G8316 to be available for business-sensitive traffic. 16 ports of 40 Gb Ethernet with QSFP+ transceivers are provided while each port can optionally opearte as four 10Gb ports using the 4x 10 Gb SFP+ break-out cable.

Figure 8 - Lenovo RackSwitch G8316

For more information, see the Lenovo RackSwitch G8316 Product Guide:

https://lenovopress.com/tips0842

6.1.7 x3550/x3650 Network Adapter Options

A wide selection of 10Gb and 1Gb network adapters are supported in the data nodes and master nodes beyond the adapters used in this reference architecture. Table 1 below is an example listing of the adapters currently offered from the Lenovo x3550/x3650 Configurator tool as of the writing of this reference architecture.

Reference the web link below for the latest listing. Click "Add Options" for model of interest, then click the "Expansion Options" tab, then scroll to "Communication Adapters".

Lenovo x3550/x3650 Network Adapter Options

Table 1. Supported Network Adapters

Adapters supported on x3550/x3650
10Gb Data Network
Broadcom NetXtreme II ML2 Dual Port 10GbaseT
Broadcom NetXtreme II ML2 Dual Port 10GbE SFP+
Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter
Broadcom NetXtreme 2x10GbE BaseT Adapter
Emulex VFA5 ML2 Dual Port 10GbE SFP+ Adapter
Emulex VFA5.2 ML2 Dual Port 10GbE SFP+ Adapter
Emulex VFA5 2x10 GbE SFP+ PCIe Adapter
Emulex VFA5.2 2x10 GbE SFP+ PCIe Adapter
Intel X540 ML2 Dual Port 10GbaseT Adapter
Intel x520 Dual Port 10GbE SFP+ Adapter
Intel X540-T2 Dual Port 10GBaseT Adapter
Intel X550-T2 Dual Port 10GBase-T Adapter
Intel X710-DA2 ML2 2x10GbE SFP+ Adapter
Mellanox ConnectX-3 10GbE Adapter
Qlogic 8200 Dual Port 10GbE SFP+ VFA
1Gb Systems Management Network
Broadcom NetXtreme I Dual Port GbE Adapter
Broadcom NetXtreme I Quad Port GbE Adapter
Broadcom NetXtreme 2xGbE BaseT Adapter

Intel I350-T2 2xGbE BaseT Adapter
Intel I350-T4 4xGbE BaseT Adapter
Intel I350-T4 ML2 Quad Port GbE Adapter
Intel I350-F1 1xGbE Fiber Adapter

6.2 Cluster Architecture

The Hortonworks reference architecture is implemented on a set of nodes that make up a cluster. A Hortonworks cluster consists of two types of nodes, data nodes and master nodes. Data nodes use System x3650 M5 servers with locally attached storage and master nodes use System x3550 M5 servers.

Data nodes run data (worker) services for storing and processing data.

Master nodes run the following types of services:

Management (control) services for coordinating and managing the cluster
Miscellaneous services (optional) for file and web serving

6.2.1 Data nodes

Table 2 lists the recommended system components for data nodes.

Table 2. Data node configuration

Component	Data node configuration
System	System x3650 M5
Processor	2 x Intel Xeon processor E5-2680 v4 2.4GHz 14-core
Memory - base	256GB: 8x 32GB 2400MHz RDIMM
Disk (OS)	2x 2.5” HDD or SSD
Disk (data)	4 TB drives: 14x 4TB NL SATA 3.5 inch (56 TB Total) 6TB drives; 14x 6TB NL SATA 3.5 inch (84 TB total) 8 TB drives: 12x 8TB NL SATA 3.5 inch (96 TB Total)
HDD controller	OS: ServeRAID M1215 SAS/SATA Controller HDFS: N2215 SAS/SATA HBA
Hardware storage protection	OS: RAID1 HDFS:None (JBOD). By default, Hortonworks maintains a total of three copies of data stored within the cluster. The copies are distributed across data servers and racks for fault recovery.

Hardware management network adapter	Integrated 1GBaseT (IMM interface)
Data network adapter	Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter

The Intel Xeon processor E5-2680 v4 is recommended to provide sufficient performance. A minimum of 256 GB of memory is recommended for most MapReduce workloads. Two sets of disks are used: one set of disks is for the operating system and the other set of disks is for data. For the operating system disks, RAID 1 mirroring is recommended.

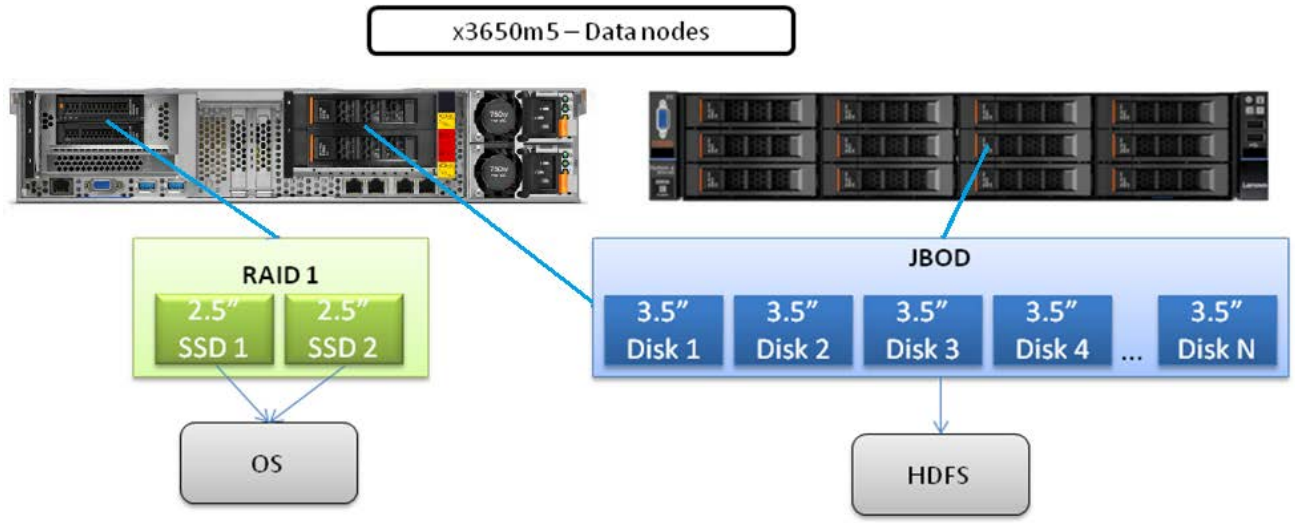

Figure 9. Data node disk assignment

Each data node in the reference architecture has internal directly attached storage. External storage is not used in this reference architecture. Available data space assumes the use of Hadoop replication with three copies of the data, and 25% capacity reserved for efficient file system operation and to allow time to increase capacity, if needed.

In situations calling for additional storage, the design approach allows for either the use of higher capacity drives for a storage rich environment, or an increase in the number of nodes for a linearly scalable compute and storage solution.

When increasing data disk capacity, there is a balance between performance and increased capacity on a node. For some workloads, increasing the amount of user data that is stored per node can decrease disk parallelism, creating a bottleneck at that node which negatively affects performance. Increasing drive size also affects rebuilding and repopulating the replicas if there is a disk or node failure. Higher density disks or nodes results in longer rebuild times. Drives that are larger than 4 TB are not recommended based on this balance of

drive capacity versus performance. To increase capacity and maintain good I/O disk performance, the number of nodes in the cluster should be increased.

For high IO throughput, the data node can be configured with 24 2.5-inch SAS drives, which have less storage capacity but much higher IO throughout. In such cases, it is recommended to use 3 host bus adapters in the configuration.

For the HDD controller, Just a Bunch of Disks (JBOD) is the best choice for data storage on a Hortonworks cluster. It provides excellent performance and, when combined with the Hadoop default of 3x data replication, also provides significant protection against data loss. The use of RAID with data disks is discouraged because it reduces performance and the amount data that can be stored. Data nodes can be customized according to client needs.

Although a minimum of three data nodes are required by Hadoop with three replica copies of data,. three data nodes should only be used for test or Proof of Concept (POC) environments. A practical minimum of nine data nodes are required for production environment to reduce risk of losing multiple nodes at a time, and other performance and availablity reasons.This reference architecture and component bill of materials representes a balanced system, a useful blend of disk, CPU and memory capacity, and I/O performance. Flexibility exists for customers to create speciality node types as required to meet workload requirements. This includes modifying the disk, CPU and memory sizes, and cluster node count to create nodes which are skewed toward higher disk I/O performance, more storage capacity, or higher compute performance.

6.2.2 Master nodes

The Master node is the nucleus of the Hadoop Distributed File System (HDFS) and supports several other key functions that are needed on a Hortonworks cluster. The Master node hosts the following functions:

- YARN ResourceManager: Manages and arbitrates resources among all the applications in the system.
- Hadoop NameNode: Controls the HDFS file system. The NameNode maintains the HDFS metadata, manages the directory tree of all files in the file system, and tracks the location of the file data within the cluster. The NameNode does not store the data of these files.
- ZooKeeper: Provides a distributed configuration service, a synchronization service, and a name registry for distributed systems.
- JournalNode: Collects, maintains, and synchronize updates from NameNode.
- HA ResourceManager: Standby ResourceManager that can be used to provide automated failover.
- HA NameNode: Standby NameNode that can be used to provide automated failover.
- Other hadoop component management services: HBase master, HiveServer2, and Spark History Server.

Table 3 lists the recommended components for a Master node. Master nodes can be customized according to client needs.

Table 3. Master node configuration

Component	Master node configuration
System	System x3550 M5

Processor	2 x Intel Xeon processor E5-2650 v4 2.2 GHz 12-core
Memory - base	128 GB – 8 x 16 GB 2133 MHz RDIMM (minimum)
Disk (OS / local storage)	OS: 2x 2.5” HDD or SSD Data: 8 x 2TB 2.5” HDD
HDD controller	ServeRAID M5210 SAS/SATA Controller
Hardware management network adapter	Integrated 1GBaseT IMM Interface
Data network adapter	Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter

The Intel Xeon processor E5-2650 v4 is recommended to provide sufficient performance. A minimum of 128 GB of memory is recommended. A choice of 240GB and 480GB SSD drives is suggested for the operating system and local storage.

The Master node uses 10 HDDs for the following storage pools:

- Two drives are configured with RAID 1 for operating system
- Two drives are configured with RAID 1 for NameNode metastore
- Four drives are configured with RAID 10 for database
- One drive is configured with RAID 0 for ZooKeeper
- One drive is configured with RAID 0 for Quorum Journal Node store

This design separates the data stores for different services and provides best performance. SSD and PCIe flash storage can be used to provide improved I/O performance for the database.

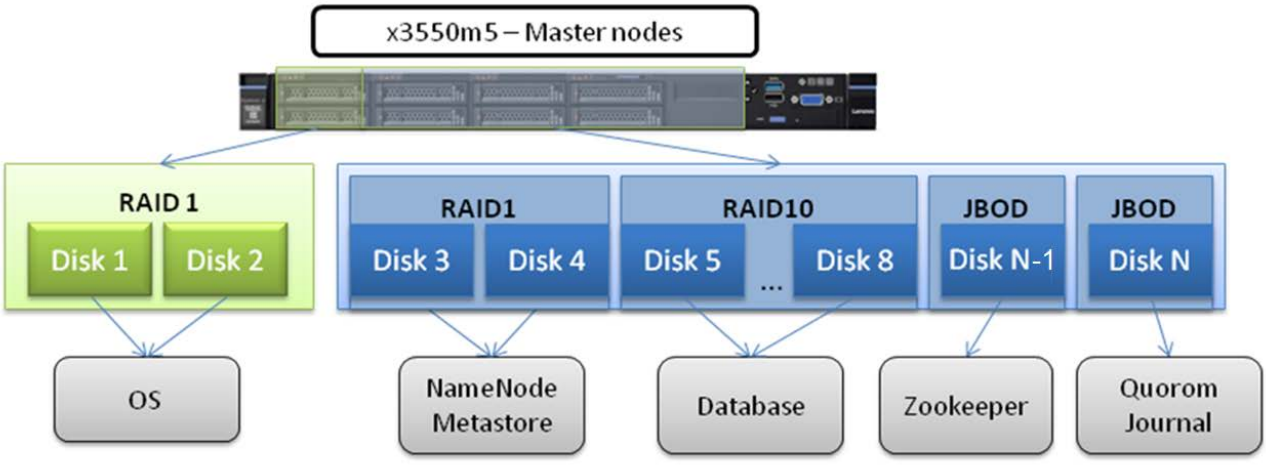

Figure 10. Hortonworks Master node disk assignment

Because the Master node is responsible for many memory-intensive tasks, multiple Master nodes will be needed to split functions. For most implementations, the size of the Hortonworks cluster is a good indicator of how many Master nodes are needed. Table 4 provides a high-level guideline for a cluster that provides HA NameNode and ResourceManager failover when configured with multiple Master nodes. For a medium size cluster with 20 - 100 data nodes, consider the use of a better configuration for the Master nodes, such as more memory and CPU core.

Table 4. Number of Master nodes

Number of Data Nodes	Number of Master nodes	Breakout of function
< 100	3	1 - Journal Node, ZooKeeper 2 - ResourceManager, HA Hadoop NameNode, JournalNode, ZooKeeper 3 - HA ResourceManager, Hadoop NameNode, JournalNode, ZooKeeper
> 100	5	1 - Journal Node, ZooKeeper 2 -ResourceManager, HA Hadoop NameNode, JournalNode, ZooKeeper 3 -HA ResourceManager, Hadoop NameNode, JournalNode, ZooKeeper 4 -JournalNode, ZooKeeper, other roles 5 -JournalNode, ZooKeeper, other roles

Note: If there are multiple racks, Master nodes should be separated across racks.

If you plan to scale up the cluster significantly, it might be best to separate out each of these functions from the beginning, even if the starting configuration is smaller and requires fewer Master nodes. The number of Master nodes can be customized based on specific needs.

Table 5. Service Layout Matrix *

Node		Master Node	Master Node	Master Node	Data Nodes
Service/ Roles	**ZooKeeper**	ZooKeeper	ZooKeeper	ZooKeeper	
	HDFS	NN,QJN	NN,QJN	QJN	Data Node
	YARN	RM	RM	History Server	Node Manager
	Hive			MetaStore, WebHCat, HiveServer2	
	Management			Ambari-server, Oozie, Metrics Monitor	Ambari-agent
	Security			Ranger KMS	
	Search				Solr
	Spark				Runs on YARN
	HBASE	HMaster	HMaster	HMaster	Region Servers

* This service layout matrix is an appropriate starting point for a balanced system. Variations on this layout may be suitable for node specializations such as incorporating data lake nodes versus analytic nodes within the cluster.

Installing and managing the Hortonworks Stack

The Hadoop ecosystem is complex and constantly changing. Hortonworks makes it simple so enterprises can focus on results. Using Apache Ambari with its web user interface is the easiest way to administer Hadoop in any environment. Ambari is a management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters. Ambari, as part of the Hortonworks Data Platform, allows enterprises to plan, install and securely configure HDP making it easier to provide ongoing cluster maintenance and management, no matter the size of the cluster,

Reference Hortonworks latest Installation documentation for detailed instructions on Installation

6.3 Systems management

Systems management includes cluster system management and hardware management.

Cluster systems management uses Apache Ambari, which places the management services on separate servers than the data servers. Because the Master node hosts important and high-memory functions, it is important that it is a powerful and fast server. The recommended Master nodes can be customized according to client needs.

Hardware management is addressed with the Lenovo XClarity Administrator, which runs on the dedicated systems management node. Lenovo XClarity is an agentless centralized resource management solution that that is aimed at reducing complexity, speeding response, and enhancing the availability of Lenovo server systems. The solution seamlessly integrates into Lenovo M5 rack servers. Through an uncluttered, dashboard-driven GUI, XClarity provides automated discovery, monitoring, firmware updates, pattern-based configuration management, hypervisor operating system deployments. Lenovo XClarity also features extensive REST APIs that provide deep visibility and control via higher-level cloud orchestration and service management software tools.

Figure 11 shows the Lenovo XClarity Administrator interface where cluster hardware can be managed (Lenovo M5 rack servers, Flex System components, Storage components, etc) and are accessible on the common dashboard. Lenovo XClarity Administrator is a virtual appliance that is quickly imported into a virtualized environment server configuration.

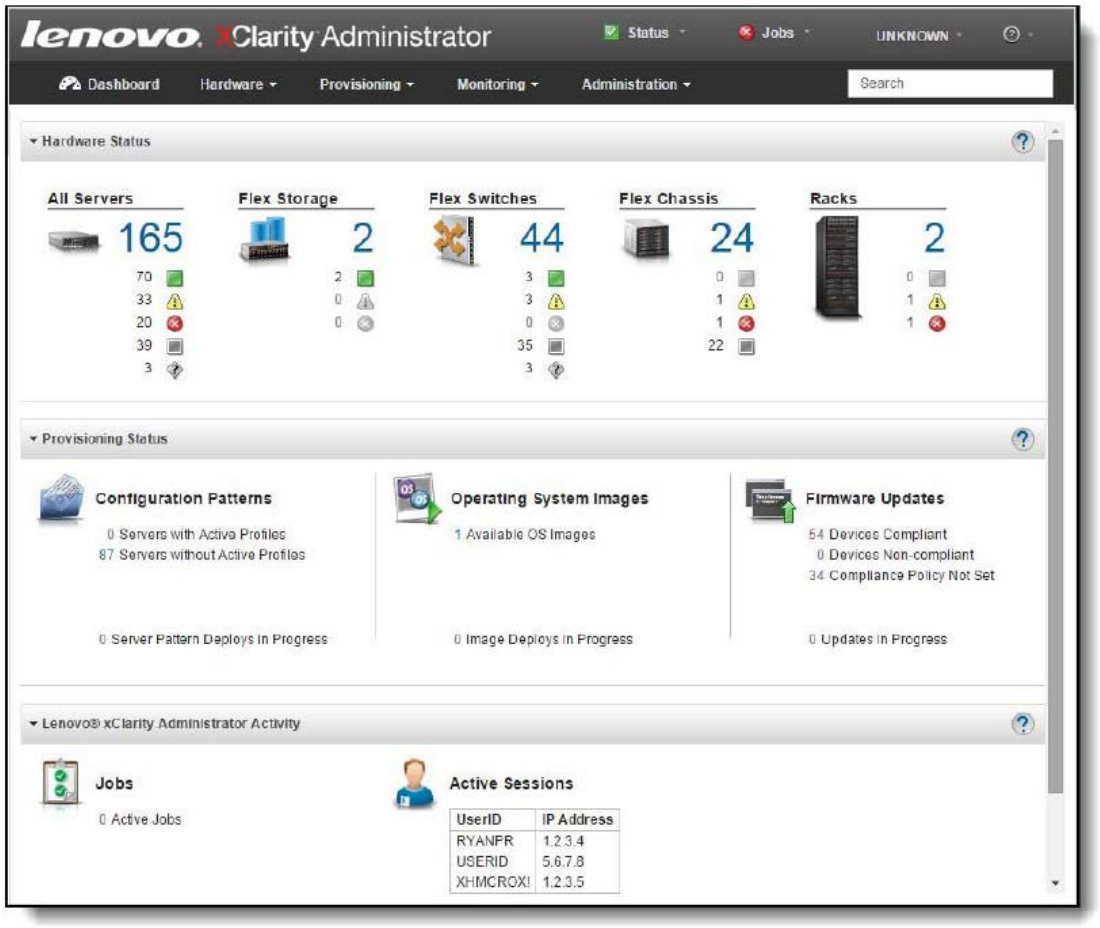

Figure 11. XClarity Administrator interface

In addition, xCAT provides a scalable distributed computing management and provisioning tool that provides a unified interface for hardware control, discovery, and operating system deployment. It can be used to facilitate or automate the management of cluster nodes. For more information about xCAT, see “Resources” on page 40.

6.4 Networking

Regarding networking, the reference architecture specifies two networks: a data network and an administrative or management network. Figure 12 shows the networking configuration for Hortonworks Data Platform.

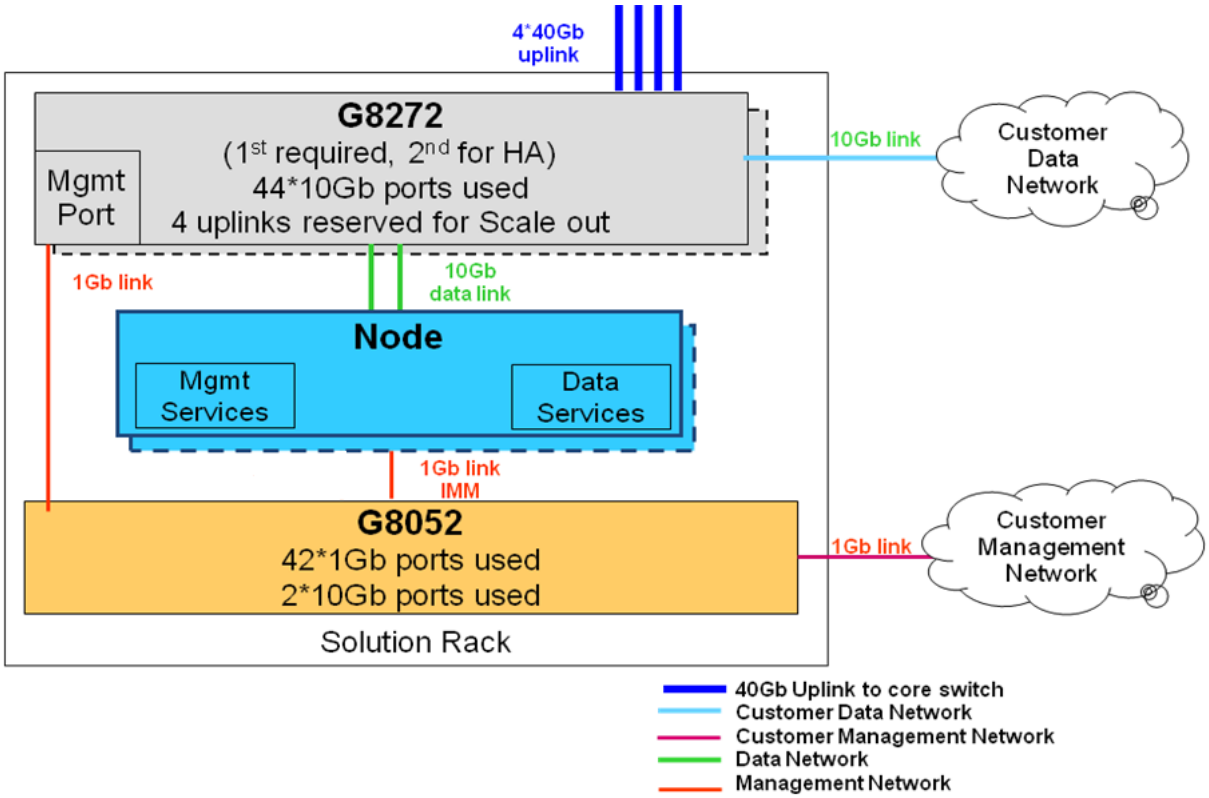

Figure 12. Hortonworks Network Configuration

6.4.1 Data network

The data network is a private cluster data interconnect among nodes that is used for data access, moving data across nodes within a cluster, and importing data into the Hortonworks cluster. The Hortonworks cluster typically connects to the customer's corporate data network.

Two top of rack switches are required; one for out-of-band management and one for the data network that is used by Hortonworks. At least one 1GbE switch is required for out-of-band management of the nodes. The data switch should be 10GbE, depending on workload requirements.

The recommended 1 GbE switch is the Lenovo RackSwitch G8052. The 10 Gb Ethernet switch can provide extra I/O bandwidth for better performance. The recommended 10 GbE switch is the Lenovo System Networking RackSwitch G8272.

The two Broadcom 10 GbE ports of each node are link aggregated to the recommended G8272 rack switch for better performance and improved HA. The data network is configured to use a virtual local area network (VLAN).

6.4.2 Hardware management network

The hardware management network is a 1 Gb Ethernet network that is used for out-of-band hardware management. Through the integrated management module II service processor (IMM2) located within the System x3650 M5 server, remote out-of-band management is possible for hardware-level maintenance of cluster nodes, such as: node deployment, UEFI configuration, firmware updates, power status and state changes (Hadoop has no dependency on the IMM2).

The hardware management network is typically connected directly to the customer's administrative network Based on customer requirements, the operating system administration links and management links can be segregated onto separate VLANs or subnets..

The reference architecture requires one 1 Gb Ethernet top-of-rack switch for this hardware management network. Administrators can access all of the node's Operating System in the cluster through this customer network/administration network, as shown in Figure 12. The hardware management link connects either to the dedicated IMM2 port or to the first port on the x3550/x3650 integrated 1GBaseT adapter to operate in a shared mode.

With the x3550/x3650 hardware, the first 1Gb network port can operate in shared mode where a single ethernet cable and a single 1Gb switch port per node can be used to access both the IMM2 and the operating system for administration. This eliminates consuming a second ethernet cable and 1Gb switch port per node for accessing both the IMM2 and Linux operating system.

6.4.3 Multi-rack network

The data network in the predefined reference architecture configuration consists of a single network topology. For cross rack communication between multiple racks, a Lenovo RackSwitch G8316 core switch per cluster is required. In this case, the second Broadcom10 GbE port can be connected to the second Lenovo RackSwitch G8272. The over-subscription ratio for G8272 is 1:2.

Figure 13 shows how the network is configured when the Hortonworks cluster is installed across more than one rack. The data network is connected across racks by two aggregated 40 GbE uplinks from each rack's G8272 switch to a core G8316 switch.

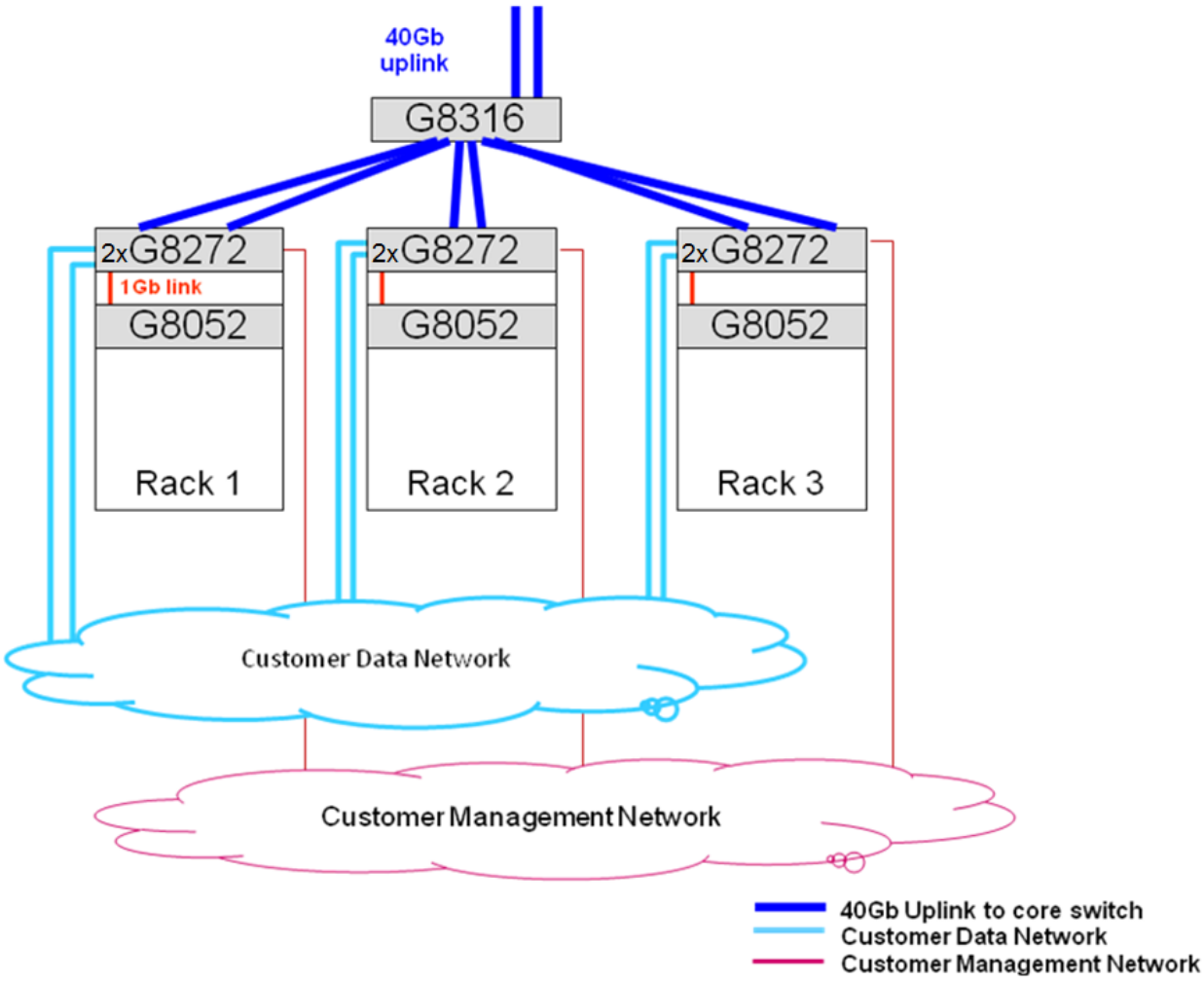

Figure 13. Hortonworks Cross Rack Network Configuration[1]

A 40GbE switch is recommended for interconnecting the data network across multiple racks. Lenovo System Networking RackSwitch G8316 is the recommended switch. A best practice is to have redundant core switches for each rack to avoid a single point of failure. Within each rack, the G8052 switch can optionally be configured to have two uplinks to the G8272 switch to allow propagation of the administrative or management VLAN across cluster racks through the G8316 core switch. For large clusters, the Lenovo System Networking RackSwitch G8332 is recommended because it provides a better cost value per 40 Gb port than the G8316. Many other cross rack network configurations are possible and might be required to meet the needs of specific deployments or to address clusters larger than three racks.

[1] One G8272 can be used but for High Availability two G8272 are recommended

If the solution is initially implemented as a multi-rack solution, or if the system grows by adding racks, the nodes that provide management services are distributed across racks to maximize fault tolerance.

6.5 Predefined cluster configurations

The intent of the predefined configurations is to ease initial sizing for customers and to show example starting points for four different-sized workloads. A Proof of Concept (POC) or test rack configuration consists of three nodes (the absolute minimum required) and a pair of rack switches. The half rack configuration consists of nine nodes and a pair of rack switches. The full rack configuration (a rack fully populated) consists of 18 nodes and a pair of rack switches. The example multi-rack configuration contains a total of 57 nodes; in each rack there are 19 nodes, a Master node and a pair of data switches.

Table 6 lists rack configurations for the Hortonworks reference architecture. The table also lists the amount of space for data and the number of nodes that each predefined configuration provides. Storage space is described in two ways: the total amount of raw storage space when 4 TB or 8 TB drives (raw storage) are used and the amount of space for the data that the customer has (available data space). Available data space assumes the use of Hadoop replication with three copies of the data and 25% capacity that is reserved for intermediate data (scratch storage). The estimates that are listed in Table 6 do not include extra space that is freed up by using compression because compression rates can vary widely based on file contents.

Table 6. Rack configurations

	Pre-defined configurations		**Example Multi-rack**
	Half rack	**Full rack**	
Raw storage (4TB)	504 TB	1008 TB	3192 TB
Available data space (4TB)	126 TB	252 TB	798 TB
Raw storage (8 TB)	1008 TB	2016 TB	6384 TB
Available data space (8 TB)	252 TB	504 TB	1592 TB
Number of Data Nodes	9	18	57
Number of Master nodes	3	3	3
Number of Racks	1	1	3
Number of 10 GbE cables	24	42	120
Number of 1 GbE cables	14	23	66

Figure 14 shows an overview of the architecture in two different one-rack sized clusters without network redundancy: a half rack and a full rack. Figure 15 shows a multi-rack-sized cluster without network redundancy.

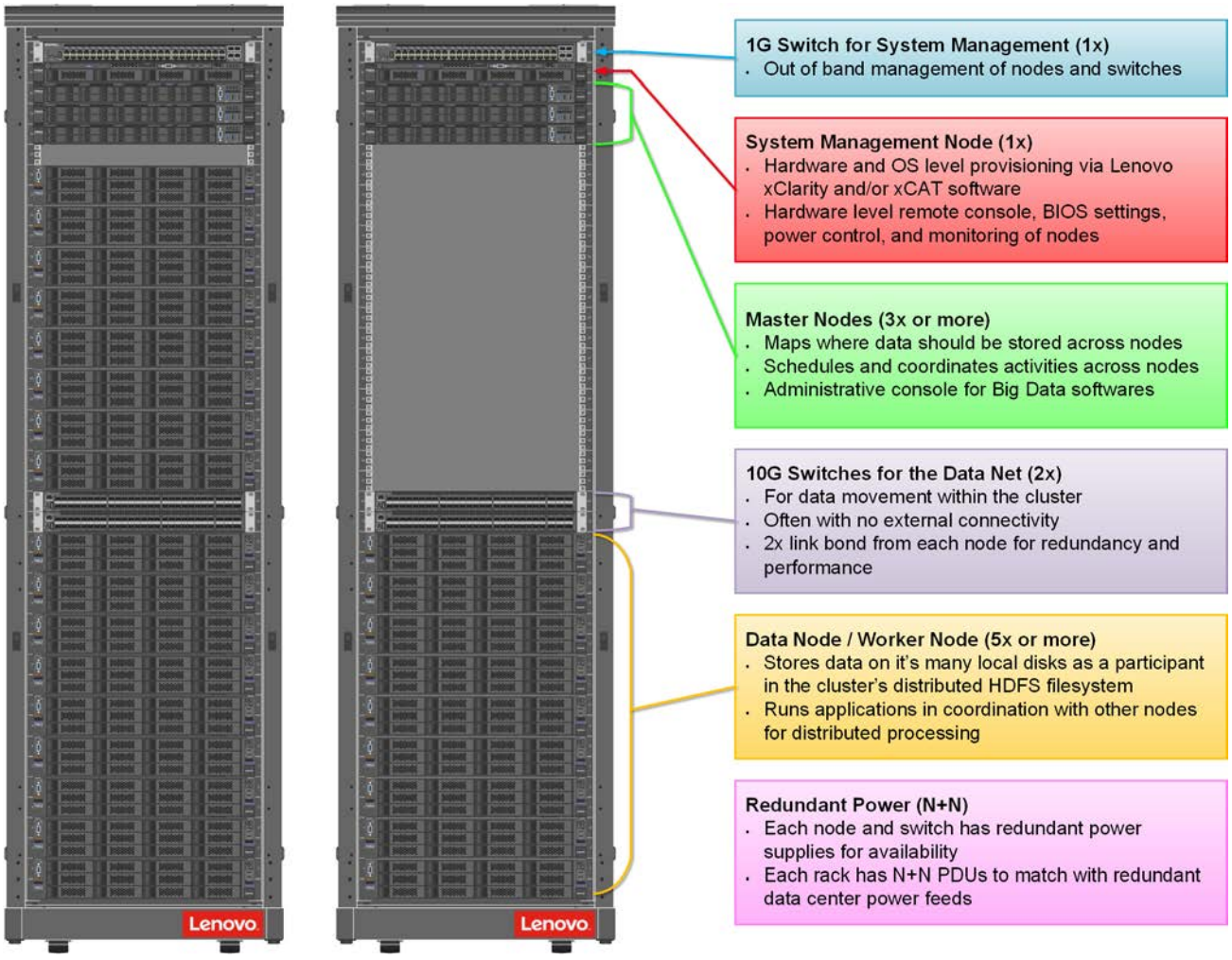

Figure 14. Half rack and full rack Hortonworks predefined configurations

Multi-Rack
(53 Data Nodes)

Figure 15. Hortonworks Multi-Rack configuration

7 Deployment considerations

This section describes other considerations for deploying the Hortonworks solution.

7.1 Increasing cluster performance

There are two approaches that can be used to increase cluster performance: increasing node memory and the use of a high-performance job scheduler and MapReduce framework. Often, improving performance comes at increased cost and you must consider the cost-to-benefit trade-offs when increasing performance.

In the Hortonworks predefined configuration, node memory can be increased to 1025 GB by using 16 x 64GB RDIMMs. An even larger memory configuration can provide greater performance, depending on the workload.

7.2 Designing for lower cost

There are two key modifications that can be made to lower the cost of a Hortonworks reference architecture solution. When lower-cost options are considered, it is important to ensure that customers understand the potential lower performance implications. A lower-cost version of the Hortonworks reference architecture can be achieved by using lower-cost node processors and lower-cost cluster data network infrastructure.

The node processors can be substituted with the Intel Xeon E5-2630 v4 2.2 GHz 10-core processor. This processor supports 1600 MHz and 1866 MHz RDIMMs, which can also lower the per-node cost of the solution.

The use of a lower-cost network infrastructure can significantly lower the cost of the solution, but can also have a substantial negative effect on intra-cluster data throughput and cluster ingest rates. To use a lower cost network infrastructure, use the following substitutions to the predefined configuration:
Within each cluster, substitute the Lenovo RackSwitch G8316 core switch with the Lenovo RackSwitch G8272.

7.3 Designing for high ingest rates

Designing for high ingest rates is difficult. It is important to have a full characterization of the ingest patterns and volumes. The following questions provide guidance to key factors that affect the rates:

- On what days and at what times are the source systems available or not available for ingest?
- When a source system is available for ingest, what is the duration for which the system remains available?
- Do other factors affect the day, time, and duration ingest constraints?
- When ingests occur, what is the average and maximum size of ingest that must be completed?
- What factors affect ingest size?
- What is the format of the source data (structured, semi-structured, or unstructured)? Are there any data transformation or cleansing requirements that must be achieved during ingest?

To increase the data ingest rates, consider the following points:

- Ingest data with MapReduce, which helps to distribute the I/O load to different nodes across the cluster.
- Ingest when cluster load is not high, when possible.
- Compressing data is a good option in many cases, which reduces the I/O load to disk and network.

- Filter and reduce data in earlier stage saves more costs.

7.4 Designing for in-memory processing with Apache Spark

Methods from the Lenovo Big Data Reference Architecture for Hortonworks apply for general Spark considerations as well; however, there are additional considerations. Conceptually, Spark is similar in nature to high performance computing.

It is important that memory capacity be carefully considered, as both the execution and storage of Spark should be able to reside fully in memory, to achieve maximum performance. Disk access, for storage or caching, is very costly to Spark processing. The memory capacity considerations are highly dependent on the application. To obtain an estimate, load an RDD of a desired dataset, into cache, and evaluate the consumption. Generally, for workloads with high execution and storage requirements, higher memory capacity becomes more important.

Additional considerations for memory configuration include the bandwidth and latency requirements. Applications with high transactional memory usage should focus on DIMM configurations that result in four DIMMs per channel using dual rank DIMMs. The following table provides ideal data node memory configurations for bandwidth/latency sensitive workloads.

Table 7. Memory configurations for x3650 data nodes

Capacity	DIMM Description	Feature	Quantity
128GB	16GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	ATCA	8
256GB	32GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	ATCB	8
384GB	32GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	ATCB	12
512GB	32GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	ATCB	16
768GB	64GB TruDDR4 Memory (4Rx4, 1.2V) PC4-19200 PC4 2400MHz LP LRDIMM	ATGG	12
1024GB	64GB TruDDR4 Memory (4Rx4, 1.2V) PC4-19200 PC4 2400MHz LP LRDIMM	ATGG	24

Similarly, processor selection may vary based on the level of desired parallelism for the workloads. For example, Apache recommends 2-3 tasks per CPU core. Large working sets of data can drive memory constraints, which can be alleviated through further increasing parallelism, resulting in smaller input sets per task. In this case, higher core counts can be beneficial. Naturally, the nature of the operations is considered, as they may be simple evaluations or complex algorithms.

7.5 Estimating disk space

When you are estimating disk space within a Hortonworks cluster, consider the following points:

For improved fault tolerance and performance, Hortonworks Data Platform replicates data blocks across multiple cluster data nodes. By default, the file system maintains three replicas.
Compression ratio is an important consideration in estimating disk space and can vary greatly based on file contents. If the customer's data compression ratio is unavailable, assume a compression ratio of 2.5:1.
To ensure efficient file system operation and to allow time to add more storage capacity to the cluster if necessary, reserve 25% of the total capacity of the cluster.

Assuming the default three replicas maintained by Hortonworks Data Platform, the raw data disk space, and the required number of nodes can be estimated by using the following equations:

Total raw data disk space required = (User data, uncompressed) * (4 / compression ratio)

Total data nodes required = (Total raw data disk space) / (Raw data disk per node)

You should also consider future growth requirements when estimating disk space.

Based on these sizing principals, Table 8 shows an example for a cluster that must store 500 TB of uncompressed user data. The example shows that the Hortonworks cluster needs 800 TB of raw disk to support 500 TB of uncompressed data. The 800 TB is for data storage and does not include operating system disk space. A total of 15 nodes are required to support a deployment of this size.

Table 8. Example of storage sizing with 4TB drives

Description	**Value**
Size of uncompressed user data	500 TB
Compression ratio	2.5:1
Size of compressed data	200 TB
Storage multiplication factor	4
Raw data disk space needed for Hortonworks cluster	800 TB
Storage needed for Hortonworks Hadoop 3x replication	600 TB
Reserved storage for headroom	200 TB
Raw data disk per node (with 4TB drives)	56 TB
Minimum number of nodes required (800/56)	15

7.6 Scaling considerations

The Hadoop architecture is linearly scalable but it is important to note that some workloads might not scale completely linearly.

When the capacity of the infrastructure is reached, the cluster can be scaled out by adding nodes. Typically, identically configured nodes are best to maintain the same ratio of storage and compute capabilities. A Hortonworks cluster is scalable by adding System x3650 M5 data nodes, Master nodes, and network switches. As the capacity of racks is reached, new racks can be added to the cluster.

When a Hortonworks reference architecture implementation is designed, future scale out should be a key consideration in the initial design. There are two key aspects to consider: networking and management. These aspects are critical to cluster operation and become more complex as the cluster infrastructure grows.

The cross rack networking configuration that is shown in Figure 13 provides robust network interconnection of racks within the cluster. As racks are added, the predefined networking topology remains balanced and symmetrical. If there are plans to scale the cluster beyond one rack, a best practice is to initially design the cluster with multiple racks (even if the initial number of nodes fit within one rack). Starting with multiple racks can enforce proper network topology and prevent future re-configuration and hardware changes. As racks are added over time, multiple G8316 switches might be required for greater scalability and balanced performance.

Also, as the number of nodes within the cluster increases, so does the number of tasks in managing the cluster, such as updating node firmware or operating systems. Building a cluster management framework as part of the initial design and planning ahead for the challenges of managing a large cluster pays off significantly in the long run.

Proactive planning for future scale out and the development of cluster management framework as a part of initial cluster design provides a foundation for future growth that can minimize hardware reconfigurations and cluster management issues as the cluster grows.

7.7 High Availability (HA) considerations

When a Hortonworks cluster on Lenovo servers, is implemented, consider availability requirements as part of the final hardware and software configuration. Typically, Hadoop is considered a *highly reliable* solution. Hadoop and Hortonworks best practices provide significant protection against data loss. Generally, failures can be managed without causing an outage. There is redundancy that can be added to make a cluster even more reliable - to make it highly available. Consideration must be given to hardware and software redundancy.

7.7.1 Networking considerations

Optionally, a second redundant switch can be added to ensure HA of the hardware management network. The hardware management network does not affect the availability of the Hortonworks Hadoop file system functionality, but it might affect the management of the cluster; therefore, availability requirements must be considered.

To support HA in the network, link aggregation is used between the 10 Gb ports of a server network adapter and the top-of-rack switch. Virtual Aggregation Groups (vLAG) can be used between redundant switches.

7.7.2 Hardware availability considerations

The redundancy of each individual data node is not necessary with Hadoop. HDFS default 3x replication provides built-in redundancy and makes loss of data unlikely if a single node fails. If Hadoop best practices are used, an outage from a data node failure is extremely unlikely as the workload can be dynamically re-allocated. The loss of a data node will not cause a job to fail. Multiple Master nodes are recommended because if there is a failure, function can be moved to an operational Master node and normal cluster operation continues. Having multiple Master nodes does not automatically resolve the issue of the NameNode being a single point of failure. For more information, see “Software availability considerations”.

Within racks, switches and nodes must have redundant power feeds with each power feed connected from a separate PDU.

7.7.3 Storage availability

HDFS 3x replication provides more than sufficient protection. Higher levels of replication can be considered if needed.

Hortonworks also provides manual or scheduled snapshots of volumes to protect against human error and programming defects. Snapshots are useful for rollback to a known data set.

7.7.4 Software availability considerations

Operating system availability is provided by using RAID1 mirrored drives for the operating system.

NameNode HA is recommended and can be achieved by using three master nodes. Active and standby nodes communicate with a group of separate daemons called JournalNodes to keep their state synchronized. When any namespace modification is performed by the active NameNode, it durably logs a record of the modification to most of these JournalNodes. The standby NameNode can read the edits from the JournalNodes and is constantly watching them for changes to the edit log. As the standby Node sees the edits, it applies them to its own namespace.

HA configuration of external database is recommended to avoid single point of failure. An external database is required for Ambari, Hive metastore and for other functions. Embedded databases should only be used for test or POC environment.

7.8 Migration considerations

If migrating data or applications to Hortonworks Data Platform is required, you must consider the type and amount of data to be migrated. Most data types can be migrated, but you must understand migration requirements to confirm viability. Hortonworks provides tools to move data between external SQL databases and Hadoop.

Other considerations should be given to whether applications must be modified to use Hadoop functionality. Significant effort might be required in some cases.

8 Appendix: Bill of Materials

This appendix includes the Bill of Materials (BOM) for different configurations of hardware for the Big Data Solution from Hortonworks deployments. There are sections for master nodes, data nodes, and networking.

The BOM includes the part numbers, component descriptions, and quantities. Table 6 on page 23 lists how many core components are required for each of the predefined configuration sizes.

The BOM lists in this appendix are not meant to be exhaustive and must always be verified with the configuration tools. Any discussion of pricing, support, and maintenance options is outside the scope of this document.

This BOM information is for the United States; part numbers and descriptions can vary in other countries. Other sample configurations are available from your Lenovo sales team. Components are subject to change without notice.

8.1 Master node

Table 9 lists the BOM for the Master node.

Table 9. Master node BOM

8869AC1	Lenovo System x3550 M5	1
A5AK	System x3550 M5 Slide Kit G4	1
A5AG	System x3550 M5 PCIe Riser 1 (1x LP x16 CPU0)	1
A4Z6	Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter	1
ATKR	x3550 M5 Base Chassis 10x2.5	1
A3YZ	ServeRAID M5210 SAS/SATA Controller	1
A3Z2	ServeRAID M5200 Series 2GB Flash/RAID 5 Upgrade	1
ATCB	32GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	8
ATLY	Intel Xeon Processor E5-2650 v4 12C 2.2GHz 30MB Cache 2400MHz 105W	1
ATMN	Addl Intel Xeon Processor E5-2650 v4 12C 2.2GHz 30MB 2400MHz 105W	1
A5AY	System x 750W High Efficiency Platinum AC Power Supply	1
A5AY	System x 750W High Efficiency Platinum AC Power Supply	1
A5A0	System x3550 M5 10x 2.5" HS HDD Kit	1
A1ML	Integrated Management Module Advanced Upgrade	1
ATL3	System Documentation and Software-US English	1
ATKT	x3550 M5 MLK Planar	1
5978	Select Storage devices - configured RAID	1

ASBK	1.2TB 10K 12Gbps SAS 2.5" G3HS 512e HDD	6
A2KB	Primary Array - RAID 10	1
6316	Rack power cable - 2.0m, 125-250V, C13 to IEC 320-C14 (WW)	2
2305	Rack Installation of 1U Component	1
ATKX	x3550 M5 Label GBM	1
ATKW	x3550 M5 Fan Filler	1
ATRW	System x M5 Rear USB Port Cover	1
2302	RAID Configuration	1
A47F	Super Cap Cable 925mm for ServRAID M5200 Series Flash	1
A52A	2U Bracket for Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter	1

8.2 Data node

Table 10 lists the BOM for the Data node.

Table 10. Data node BOM

8871AC1	Lenovo System x3650 M5	1
ATG2	System Documentation and Software-US English	1
A5GE	x3650 M5 12x 3.5" HS HDD Assembly Kit	1
A5GL	System x3650 M5 Rear 2x 3.5" HDD Kit (Cascaded)	1
A3YY	N2215 SAS/SATA HBA	1
A5GH	System x3650 M5 Rear 2x 2.5" HDD Kit (Independent RAID)	1
ATE4	System x3650 M5 Planar BDW	1
ATEN	Intel Xeon Processor E5-2650 v4 12C 2.2GHz 30MB Cache 2400MHz 105W	1
ATFD	Addl Intel Xeon Processor E5-2650 v4 12C 2.2GHz 30MB 2400MHz 105W	1
A5FV	System x Enterprise Slides Kit	1
ATE0	System x3650 M5 12x 3.5" Base without Power Supply BDW	1
A5EU	System x 750W High Efficiency Platinum AC Power Supply	1
A45W	ServeRAID M1215 SAS/SATA Controller	1
A5EU	System x 750W High Efficiency Platinum AC Power Supply	1
ATEA	System x3650 M5 EIA L - Blank	1

A4Z6	Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter	1
A483	Populate and Boot From Rear Drives	1
5978	Select Storage devices - configured RAID	1
A5VQ	4TB 7.2K 12Gbps NL SAS 3.5" G2HS 512e HDD	14
A2K7	Primary Array - RAID 1	1
A577	120GB SATA 2.5" MLC G3HS Enterprise Value SSD	2
ATCB	32GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	8
6316	Rack power cable - 2.0m, 125-250V, C13 to IEC 320-C14 (WW)	2
2306	Rack Installation >1U Component	1
ATE2	System x3650 M5 System Agency Label	1
ATE3	System x3650 M5 System Level Code	1
ATRG	system x M5 rear USB port cover	1
2302	RAID Configuration	1
A52A	2U Bracket for Broadcom NetXtreme Dual Port 10GbE SFP+ Adapter	1
A5FT	System x3650 M5 Power Paddle Card	1
A5G1	System x3650 M5 EIA Plate	1
A5V5	System x3650 M5 Right EIA for Storage Dense Model	1
ASQA	System x3650 M5 Rear 2x 2.5" HDD Label (Independent RAID-Riser1)	1

8.3 Systems Management Node

Table 11 lists the Management node BOM.

Table 11. Systems Management node BOM

8869AC1	Lenovo System x3550 M5	1
A5AK	System x3550 M5 Slide Kit G4	1
ATKS	x3550 M5 Base Chassis 4x3.5	1
A3YZ	ServeRAID M5210 SAS/SATA Controller	1
ATCA	16GB TruDDR4 Memory (2Rx4, 1.2V) PC4-19200 CL17 2400MHz LP RDIMM	1
ATLU	Intel Xeon Processor E5-2609 v4 8C 1.7GHz 20MB Cache 1866MHz 85W	1

A5AX	System x 550W High Efficiency Platinum AC Power Supply	1
A5AX	System x 550W High Efficiency Platinum AC Power Supply	1
A5A4	System x3550 M5 4x 3.5" HS HDD Kit	1
A1ML	Integrated Management Module Advanced Upgrade	1
ATL3	System Documentation and Software-US English	1
ATKT	x3550 M5 MLK Planar	1
5978	Select Storage devices - configured RAID	1
A2K7	Primary Array - RAID 1	1
A22P	1TB 7.2K 6Gbps NL SATA 3.5" G2HS HDD	2
6316	Rack power cable - 2.0m, 125-250V, C13 to IEC 320-C14 (WW)	2
2305	Rack Installation of 1U Component	1
ATKX	x3550 M5 Label GBM	1
ATKW	x3550 M5 Fan Filler	2
ATRW	System x M5 Rear USB Port Cover	1
2302	RAID Configuration	1
A595	ODD Filler	1

8.4 Management/Administration network switch

Table 12 lists the BOM for the Management/Administration network switch.

Table 12. Management/Administration network switch BOM

Code	Description	Quantity
7159HC1	Lenovo RackSwitch G8052 (Rear to Front)	1
ASY2	Lenovo RackSwitch G8052 (Rear to Front)	1
A3KR	Air Inlet Duct for 442 mm RackSwitch	1
A3KP	Adjustable 19" 4 Post Rail Kit	1
6201	1.5m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
2305	Rack Installation of 1U Component	1

8.5 Data network switch

Table 13 lists the BOM for the data network switch.

Table 13. Data network switch BOM

Code	Description	Quantity
7159HCW	Lenovo RackSwitch G8272 (Rear to Front)	1
ASRD	Lenovo RackSwitch G8272 (Rear to Front)	1
ASTN	-SB- Air Inlet Duct for 487 mm RackSwitch	1
6201	1.5m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
A3KP	Adjustable 19" 4 Post Rail Kit	1
2305	Rack Installation of 1U Component	1

8.6 Rack

Table 14 lists the BOM for the rack.

Table 14. Rack BOM

Code	Description	Quantity
1410HPB	Intelligent Cluster 42U 1100mm Enterprise V2 Dynamic Rack	1
A2M8	Intelligent Cluster 42U 1100mm Enterprise V2 Dynamic Rack	1
5895	1U 12 C13 Switched and Monitored 60A 3 Phase PDU	4
2202	Cluster 1350 Ship Group	1
2304	Rack Assembly - 42U Rack	1
2310	Cluster Hardware & Fabric Verification - 1st Rack	1
AU8K	LeROM Validation	1
4271	1U black plastic filler panel	2
4275	5U black plastic filler panel	3

Different cluster sizing leaves different unused rack space; therefore, consider the use of blank plastic filter panels for the rack to better direct cool air flow.

The number of PDUs in the rack depends on the server numbers in the rack. Four PDUs should be used for the half rack configuration and six PDUs for a full rack.

8.7 Cables

Table 15 lists the BOM for the cables available for each node.

Table 15. Cables BOM

Code	Description
AT2S	-SB- Lenovo 3m Active DAC SFP+ Cables
A3RG	0.5m Passive DAC SFP+ Cable
3791	0.6m Yellow Cat5e Cable
A51N	1.5m Passive DAC SFP+ Cable
3794	10m Yellow Cat5e Cable
A51P	2m Passive DAC SFP+ Cable
3793	3m Yellow Cat5e Cable
A4RA	CAT5E IntraRack Cable Service
40K8951	1.5m Yellow Cat5e Cable

9 Acknowledgements

This reference architecture document has benefited very much from the detailed and careful review comments provided by colleagues at Lenovo and Hortonworks.

Lenovo technical review

- Brian Finley – Principal Architect, Big Data

Hortonworks technical review

- Ali Bajwa

10 Resources

For more information, see the following resources:

Lenovo System x3650 M5 (Hortonworks Data Node):

- Product page: shop.lenovo.com/us/en/systems/servers/racks/systemx/x3650-m5/
- Lenovo Press product guide: lenovopress.com/tips1193

Lenovo System x3550 M5 (Hortonworks Master node):

- Product page: shop.lenovo.com/us/en/systems/servers/racks/systemx/x3550-m5/
- Lenovo Press product guide: lenovopress.com/tips1194

Lenovo RackSwitch G8052 (1GbE Switch):

- Product page: shop.lenovo.com/us/en/systems/browsebuy/%20rackswitch-g8052.html
- Lenovo Press product guide: lenovopress.com/tips0813

Lenovo RackSwitch G8272 (10GbE Switch):

- Product page: shop.lenovo.com/us/en/systems/browsebuy/lenovo-rackswitch-g8272.html
- Lenovo Press product guide: lenovopress.com/tips1267

Lenovo XClarity Administrator:

- Product page: shop.lenovo.com/us/en/servers/thinkserver/system-management/xclarity
- Lenovo Press product guide: lenovopress.com/tips1200

Hortonworks:

- Hortonworks Data Platform (HDP): http://hortonworks.com/products/data-center/hdp/
- Hortonworks products: http://hortonworks.com/products/
- Hortonworks services: http://hortonworks.com/services/
- Hortonworks solutions: http://hortonworks.com/solutions/
- Hortonworks training: http://hortonworks.com/training/ Open source software:
- Hadoop: hadoop.apache.org
- Spark: spark.apache.org
- Flume: flume.apache.org
- HBase: hbase.apache.org
- Hive: hive.apache.org
- Oozie: oozie.apache.org
- Mahout: mahout.apache.org
- Pig: pig.apache.org
- Sqoop: sqoop.apache.org
- ZooKeeper: zookeeper.apache.org

xCat: xcat.sourceforge.net

11 Document history

Version 1.0	1/6/2017	Initial release version
Version 1.1	2/15/2107	Updated memory size in Table 2 and Table 10

12 Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and Rescue and Recovery are trademarks of Lenovo.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual

user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.