Lenovo Big Data Reference Architecture for the MapR Converged Data Platform

Last update: **30 March 2017**

Configuration reference number: BDAMAPRXX53

Describes the reference architecture for MapR Converged Data Platform

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

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

Contains detailed predefined configurations for different servers and associated networking

Dan Kangas (Lenovo)

Weixu Yang (Lenovo)

Ajay Dholakia (Lenovo)

James Sun (MapR)

Table of Contents

1 Introduction

This document describes the reference architecture for the Big Data Solution based on the MapR Converged Data Platform. It provides a predefined and optimized hardware infrastructure for the MapR Converged Enterprise Edition, the commercial edition of the MapR Converged Data Platform that supports enterprise-grade features for business-critical production deployments. This reference architecture provides the planning, design considerations, and best practices for implementing the MapR Converged Data Platform with Lenovo products.

The Lenovo and MapR teams worked together on this document, and the reference architecture that is described herein was validated by both Lenovo and MapR.

MapR brings the power of data analytics to the enterprise. MapR platform services are the core data handling capabilities of the MapR Converged Data Platform and consist of MapR-FS, MapR-DB, and MapR Streams. The enterprise-friendly design provides a familiar set of file and data management services, including a global namespace, high availability (HA), data protection, self-healing clusters, access control, real-time performance, secure multi-tenancy, and management and monitoring.

On top of the platform services, MapR packages a broad set of Apache™ Hadoop® open source ecosystem projects that enable big data applications. The goal is to provide an open platform that lets you choose the right tool for the job. MapR tests and integrates open source ecosystem projects such as Apache Drill, Apache Hive™, Apache Pig, Apache HBase™ and Apache Mahout™, among others.

The predefined configuration provides a baseline configuration 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 intended audience of this document is IT professionals, technical architects, sales engineers, and consultants to assist in planning, designing and implementing the big data solution with Lenovo System x hardware. It is assumed that you are familiar with Apache Hadoop components and capabilities.

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 MapR solution that uses Lenovo hardware.

2.1 Business problem

Data sets grow rapidly - in part because they are increasingly gathered by low cost and numerous information-sensing mobile devices, aerial devices (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s. In 2015 data consumed in the world was estimated to be an astronomical 8 zettabytes (1 zettabyte - 1 trillion terabytes) but by 2020 it's expected to have increased significantly again to be 40 zettabytes (ZB).

Big data spans the following dimensions:

- **Volume:** Big data comes in one size: enormous! Enterprises are awash with data, easily amassing terabytes to petabytes of information. For example, human generated data through social media is growing at a 10x rate and machine generated is growing at a 100x rate.
- **Velocity:** Often time-sensitive, big data must be used as it is streaming into the enterprise to maximize its value to the business. Data ingestion comprises both batch and real-time loading methods.
- **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.

2.2 Business value

MapR provides the industry’s only converged data platform that uniquely allows applying analytical insights to operational processes in real time to create competitive advantage for our customers. MapR is a data platform that converges historically separate product segments/categories in order to enable extraordinary new value never before possible. The MapR Converged Data Platform (“MapR Platform” or “MapR”) is powered by the industry’s fastest, most reliable, secure, and open data infrastructure that dramatically lowers TCO and enables global real-time data applications.

The key benefits of MapR are to ensure customers’ production success for Hadoop, Apache Spark™ and much more. MapR was engineered for the data center with IT operations in mind. MapR serves business-critical needs that cannot afford to lose data, must run on a 24x7 basis, require immediate recovery from node and site failures – all with a smaller data center footprint. MapR supports these capabilities for the broadest set of data applications from batch analytics to interactive querying and real-time streaming.

MapR deployed on Lenovo System x servers with Lenovo networking components provides superior performance, reliability, and scalability. This 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 Requirements

The functional and non-functional requirements for the MapR reference architecture are described in this section.

3.1 Functional requirements

A big data solution supports the following key functional requirements:

- Various application types, including batch and real-time analytics
- Industry-standard interfaces, so that existing applications can work
- Real-time streaming and processing of data
- Various data types and databases
- Various client interfaces
- Large volumes of data

3.2 Non-functional requirements

Customers require their big data solution to deliver business value without significant overhead. The following non-functional requirements are key:

- Simplified:
    - Ease of development
    - Easy management at scale
    - Advanced job management
    - Multi-tenancy for users, data, and applications
- Reliable:
    - Data protection with snapshot and mirroring
    - Automated self-healing
    - Insight into software/hardware health and issues
    - Mission-critical high availability (HA) and business continuity (99.999% uptime)
- Fast and scalable:
    - Real-time streaming capabilities
    - Superior NoSQL database agility
    - Web-scale file system performance
    - Linear scale-out on a distributed architecture
- Secure:
    - Strong authentication and authorization
    - Kerberos and broad user registry support
    - Data confidentiality and integrity
    - Comprehensive data security auditing

4 Architectural overview

The MapR Converged Data Platform integrates Hadoop, Spark, and Apache Drill with real-time database capabilities, global event streaming, and scalable enterprise storage to power a new generation of big data applications. The MapR Platform delivers enterprise grade security, reliability, and real-time performance while dramatically lowering both hardware and operational costs of your most important applications and data. This MapR Converged Data Platform solution is shown below.

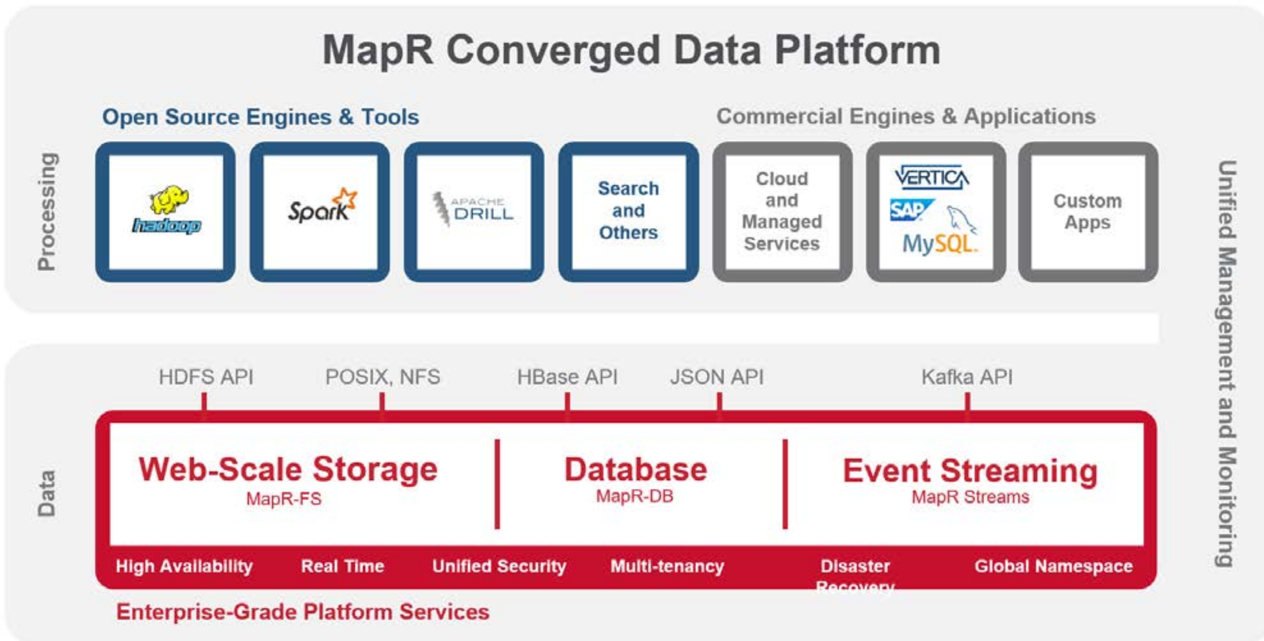

Figure 1: The MapR Converged Data Platform

5 Component model

MapR supports dozens of open source projects and is committed to using industry-standard APIs to provide a frictionless method of developing and deploying exciting new applications that can meet the most stringent production runtime requirements.

5.1 Enterprise-Grade Platform Services

MapR platform services are the core data handling capabilities of the MapR Platform and consist of MapR-FS, MapR-DB, and MapR Streams. Its enterprise friendly design provides a familiar set of file and data management services, including a global namespace, high availability (HA), data protection, self-healing clusters, access control, real-time performance, secure multi-tenancy, and management and monitoring.

A few of the key attributes of the MapR converged platform are shown below:

- Linux operating systems
  Red Hat Linux and SUSE Linux are supported with the Lenovo System x reference architecture.

- MapReduce
  MapR provides high performance for MapReduce operations on Hadoop and publishes performance benchmarking results. The MapR architecture is built in C/C++ and harnesses distributed metadata with an optimized shuffle process, enabling MapR to deliver consistent high performance. Both the classic MapReduce and YARN frameworks are supported by MapR.

- File-based applications
  MapR is a Portable Operating System Interface (POSIX) system that fully supports random read-write operations. By supporting industry-standard Network File System (NFS), users can mount a MapR cluster and run any file-based application, written in any language, directly on the data residing in the cluster. All standard tools in the enterprise including browsers, UNIX tools, spreadsheets, and scripts can access the cluster directly without any modifications.

- NoSQL Database
  MapR has removed the trade-offs that organizations face when looking to deploy a NoSQL solution. Specifically, MapR-DB delivers ease of use, reliability, and performance advantages for NoSQL applications. It provides native multi-model support for JSON document and wide column data models. MapR-DB also provides scalability, strong consistency, and continuous low latency with an architecture that eliminates compaction delays or background consistency corrections (“anti-entropy”).

- Stream processing
  MapR provides a simplified architecture for real-time stream computational engines such as Spark Streaming, Apache Flink®, and Apache Storm™. Streaming data feeds can be written directly to the MapR Platform for long-term storage and MapReduce processing.

5.2 Open Source Engines and Tools

MapR packages a broad set of Apache open source ecosystem projects that enable big data applications. The goal is to provide an open platform that lets you choose the right tool for the job. MapR tests and integrates open source ecosystem projects such as Apache Drill, Apache Hive™, Apache Pig, Apache HBase™ and Apache

Mahout™, among others.

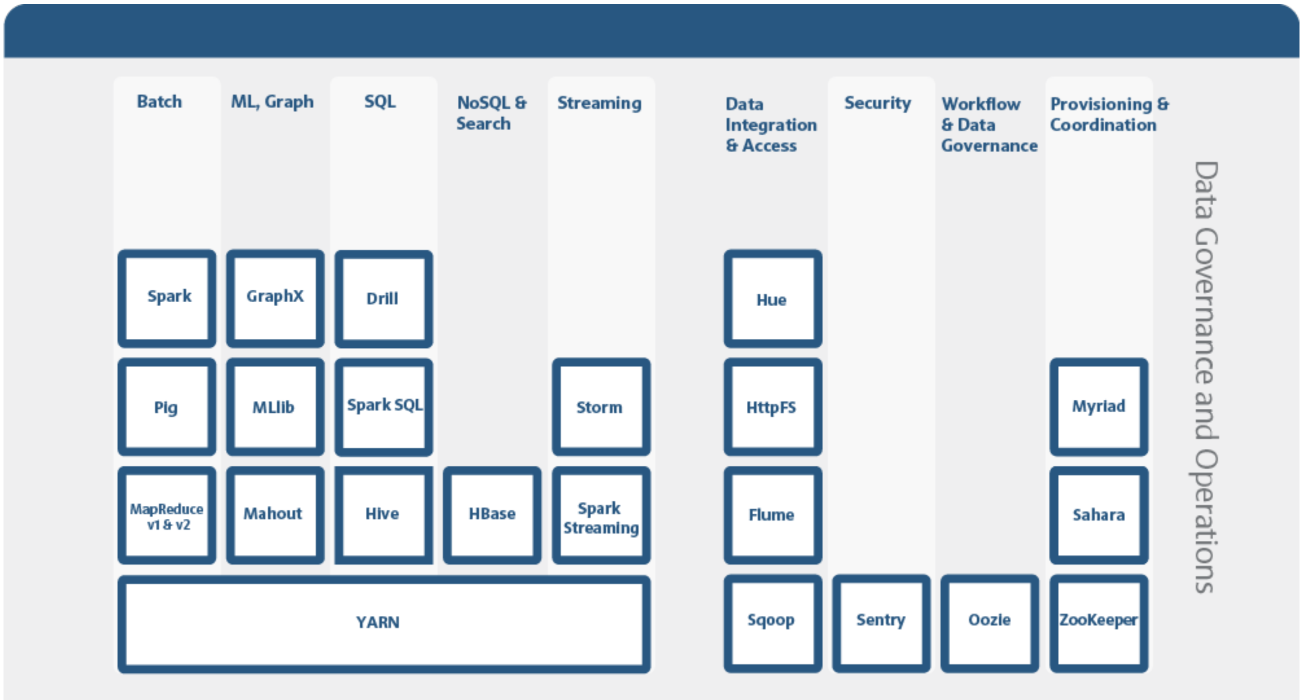

Figure 2: MapR key components for open source engines and tools

- Batch Processing
  [Apache Spark] provides fast and general engine for large-scale data processing. [Apache Pig] is a language and runtime for analyzing large data sets, consisting of a high-level. MapReduce/Yarn is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Machine Learning
  Popular machine learning libraries are available in the MapR Platform which includes Apache Mahout and Spark MLlib.

- SPL-on_Hadoop
  There are a number of applications that support SQL access against data contained in MapR. MapR is also leading the development of Apache Drill that brings American National Standards Institute (ANSI) SQL capabilities to Hadoop. Drill enables data analysts to perform self-exploratory tasks on any type of data stored in Hadoop, including JSON and Parquet files. [Apache Hive] is a data warehouse system for Hadoop that facilitates easy data summarization, ad hoc queries, and the analysis of large data sets stored in Hadoop compatible file systems. Spark SQL offers real-time in-memory processing to gain insights to your data.
- NoSQL
  Apache HBase is supported in addition to the core MapR-DB (HBase API compatible) to handle NoSQL workloads.
- Streaming
  Storm and Spark Streaming support is available in addition to the core MapR Streams (Kafka API compatible) to handle event streaming for Internet of Things (IoT) data streaming on the edge devices.

- Data Integration and Access
  Hue, HttpFS, Sqoop, and Flume are all supported for efficiently ingesting and accessing data in a cluster.

- Apache Flume – A distributed, reliable, and highly available service for efficiently ingesting large amounts of data into a cluster.

- Sentry - Sentry is supported for enforcing fine grained role-based authorization to data and metadata stored on a Hadoop cluster.

- Workflow and Data Governance – Oozie is the workflow coordination manager which is supported with MapR.

- Provisioning and Coordination
  Apache ZooKeeper is distributed service for maintaining configuration information, providing distributed synchronization, and providing group service. Sahara is a simple means to provision a data-intensive application cluster (Hadoop or Spark) on top of OpenStack. Apache Myriad enables the co-existence of Apache Hadoop and Apache Mesos on the physical infrastructure. By running Hadoop YARN as a Mesos framework, YARN applications and Mesos frameworks can run side-by-side, dynamically sharing cluster resources

5.3 Commercial Engines and Applications

One of the key developer benefits of the MapR Platform is its basis on well-known, open APIs and interfaces. This enables commercial software vendors such as SAP, SAS, HP, and Cisco to easily deploy large scale applications onto the MapR Platform. It also means that even small teams of developers can create enterprise-grade software products by exploiting the built-in protections of the MapR Platform as well as mature commercial processing engines.

Lastly, details about the various MapR and Linux version operability can be found at: http://maprdocs.mapr.com/home/InteropMatrix/r_os_matrix.html?hl=os,matrix

6 Operational model

This section describes the operational model for the MapR reference architecture.

To show the operational model for different sized customer environments, this reference architecture describes four different models for supporting different amounts of data. Throughout the document, these models are referred to as starter rack, half rack, full rack, and multi-rack configuration sizes. The full bill of material for building these models is available as a predefined configuration which can be used for ordering purposes or as a starting point for customizing the configuration.

A MapR deployment consists of cluster nodes, networking, power, and racks. The predefined configurations can be implemented as-is or modified based on specific customer requirements, such as lower cost, higher CPU performance, increased storage, or 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 MapR 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 servers and Lenovo RackSwitch G8052 and G8272 top of rack switches.

6.1.1 Lenovo System x3650 M5 Server

The Lenovo System x3650 M5 server (as shown in 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 Gb serial-attached SCSI (SAS) internal storage connectivity and up to 1.5TB 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 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) per virtual machine. Its pay-as-you-grow flexible design and great expansion capabilities solidify dependability for any kind of virtualized workload, with minimal downtime.

The Lenovo x3650 M5 server provides internal storage density of over 100TB 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 storage virtualization 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 3.5-inch drives has raw capacity up to 8 TB and each of two 2.5-inch drives has raw capacity of 1.8 TB for over 100TB per node and over a petabyte 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:

http://lenovopress.com/lp0068

6.1.2 Lenovo RackSwitch G8052

The Lenovo System Networking RackSwitch G8052 (as shown in Figure 4) is an Ethernet switch that is designed for the data center and provides a virtualized, cooler, and simpler 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-sensitive traffic by using redundant power supplies, fans, and numerous high-availability features.

Figure 4: Lenovo RackSwitch G8052

Lenovo RackSwitch G8052 has the following characteristics:

- Forty-eight 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 RackSwitch G8052 Product Guide: lenovopress.com/tips0813.

6.1.3 Lenovo RackSwitch G8272

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 four 40 Gb QSFP+ ports. It is an enterprise-class and full-featured data center switch that deliver 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 5) is ideal for latency-sensitive applications, such as client virtualization. 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 5: 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 very 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 RackSwitch G8272 Product Guide: lenovopress.com/tips1267.

6.1.4 Lenovo RackSwitch G8332 - Cross-Rack Switch

The RackSwitch™ G8332 (as shown in Figure 6) 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 G8332 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 G8332 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 6: Lenovo RackSwitch G8332

For further information on the G8332 switch visit this link:

[Lenovo RackSwitch G8332]

6.2 Cluster nodes

The MapR reference architecture is implemented on a set of nodes that make up a cluster. Nodes use System x3650 M5 servers with locally attached storage. MapR runs well on a homogenous server environment with no need for different hardware configurations for data services. Server nodes can run three different types of services which are MapR data services, MapR management services and other optional MapR services.

Unlike other Hadoop distributions that require different server configurations for management nodes and data nodes, the MapR reference architecture requires only a single MapR node hardware configuration. Each node is then configured to run one or more of the mentioned services.

6.2.1 Predefined Configuration Summary

Table 1 lists the recommended system components for cluster nodes.

Table 1: Node predefined configuration

Component	Predefined configuration
System	System x3650 M5
Processor	2 x Intel Xeon Processor E5-2680 v4 2.4Ghz 14-core
Memory – base	256 GB – 8 x 32GB 2400MHz RDIMM (minimum)
Disk (OS)	2x, 2.5 inch 120GB SSD
Disk (data)	• 4 TB drives: 14x 4TB NL SATA 3.5 inch (56 TB Total) • 6 TB drives: 14x 6TB NL SATA 3.5 inch (84 TB Total) • 8 TB drives: 14x 8TB NL SATA 3.5 inch (112 TB Total)

HDD controller	OS: ServeRAID M1215 SAS/SATA Controller MapR-FS: N2215 SAS/SATA HBA
Hardware storage protection	OS: RAID1 MapR-FS: software only (JBOD HDDs). By default, MapR 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 Adapter
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 with 512 GB or more recommended for Drill, Spark, and memory-intensive MapReduce workloads.

6.2.2 Storage Configuration

Two sets of disks are used, one set of disks is for operating system and the other set of disks is for data. For the operating system disks, RAID 1 mirroring should be used.

Each node in the reference architecture has internal storage. External storage is not used in this reference architecture. Available data space assumes the use of MapR 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 where higher storage capacity is required, the main design approach could be to increase the amount of data disk space per node. Using 6 TB drives instead of 4 TB drives increases the total per node data disk capacity from 56 TB to 84 TB, a 50% increase. However, when increasing data disk capacity, there is a trade-off with the node's storage performance (IO throughput). For some workloads, increasing the amount of user data that is stored per node can *decrease* disk parallelism and negatively affect 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 higher rebuild times. Drives that are larger than 4 TB are not recommended for best disk performance due to this balance of capacity versus performance. In this case, *higher capacity should be achieved by increasing the number of nodes in the cluster.*

For the case where higher IO throughout from each node is required, the data nodes can be configured with 24 2.5-inch SAS drives, (which may have less storage capacity than 3.5" drives) but much higher IO throughput due to parallelism from higher drive count and faster SAS technology.

For the HDD controller selection, just-a-bunch-of-disks (JBOD) is the best choice for a MapR cluster. It provides excellent performance and when combined with the 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.

Lenovo storage adapters provide a true JBOD configuration for best performance. In cases where only a RAID controller is available with no JBOD configuration, RAID0 may be configured but with a single HDD per RAID array. This will most closely emulate the JBOD configuration. Multiple HDDs in a single RAID0 array are not recommended nor required since a failure of a single HDD will cause all HDDs in that array to go off-line.

6.2.3 Minimum Node Count

A minimum of three data nodes are required as MapR has three copies of data by default. Three data nodes should be used for test or proof of concept (POC) environments only. A minimum of five data nodes are required for the production environment to allow continued operation if there are data node failures.

6.2.4 Node Service Layout

The location of various MapR services on certain nodes depends on the total number of nodes in the cluster. MapR is very flexible in its ability to use any node for any MapR service. Here we recommend three MapR service layout templates for small, medium and large deployments to address high availability (HA). For additional information on planning and installing a MapR converged data platform, reference this link: http://maprdocs.mapr.com/home/install.html.

Single Rack Deployment

In this scenario, we recommend using two CLDB, web and resource manager services in nodes 1 and 2. Zookeeper services are spread across nodes 3, 4 and 5. All nodes in the cluster have the base services such as file server, node manager, spark, and NFS gateway services installed.

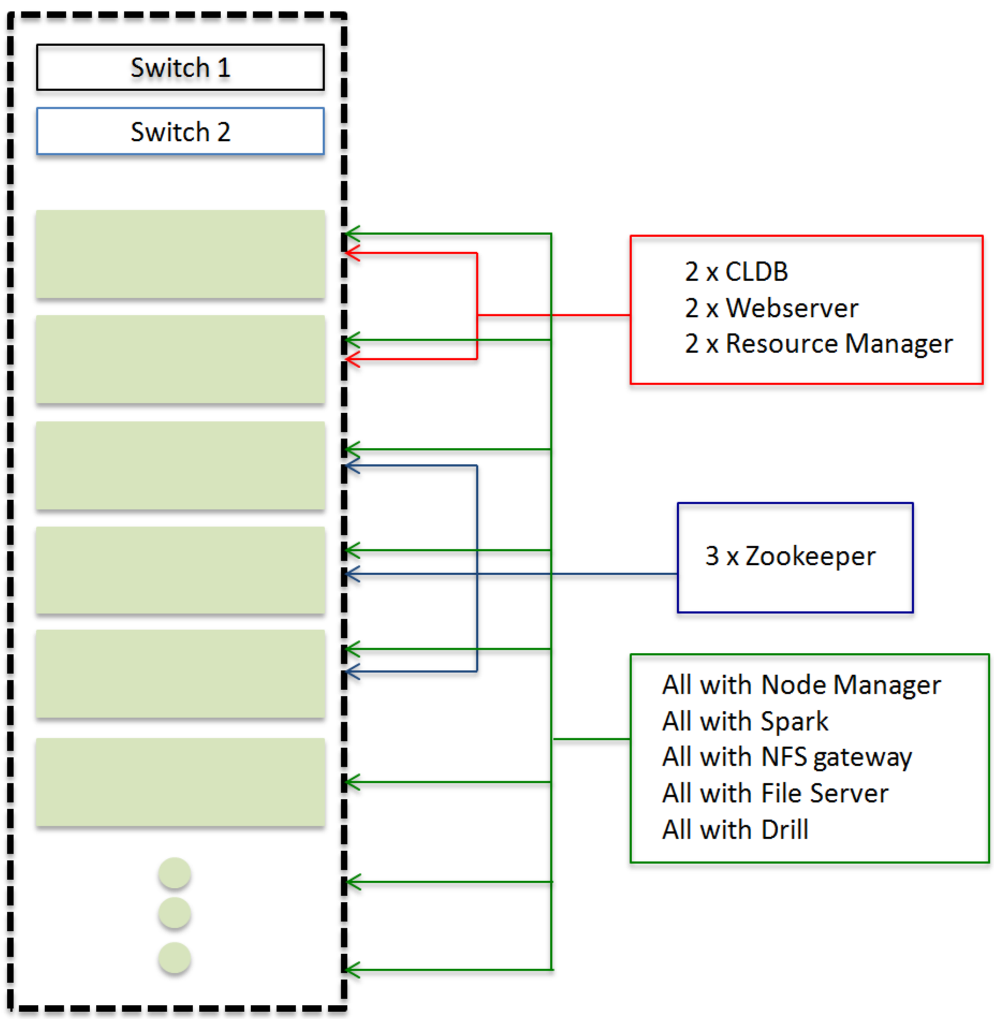

Figure 7: Single Rack Service Layout

Medium Size Deployment (Two Racks)

In this scenario, similar to single rack deployment, we recommend using two CLDB, web and resource manager services in node 1 and 2 in rack 1. Zookeeper services are spread across node 3, 4 and 5 in rack 1. All nodes in the cluster have the base services such as file server, node manager, spark, and NFS gateway services installed.

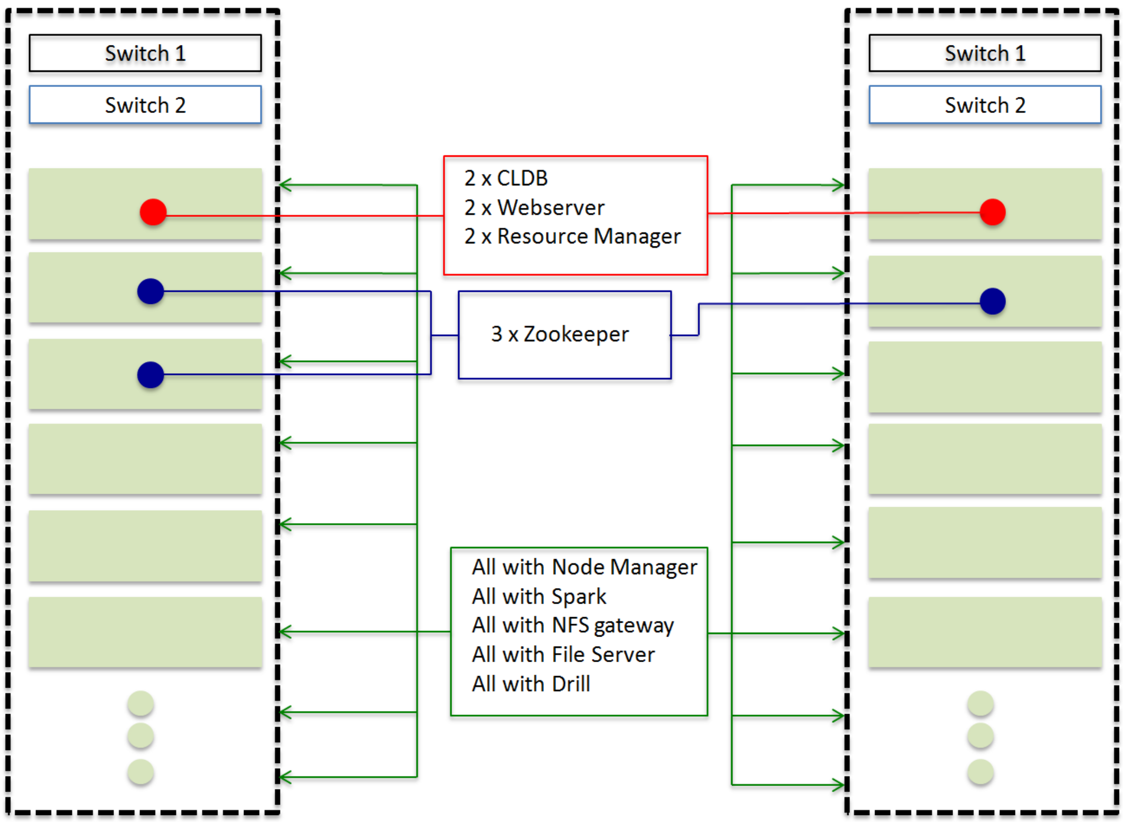

Figure 8: MapR Two Rack Service Layout

Large Multi Rack Deployment

In this scenario, taking into consideration the rack failure. We recommend to spread the main services across racks using three CLDB, web and resource manager services in node 1 across the racks 1, 2 and 3. Zookeeper services are deployed in node 2 across racks 1, 2 and 3. All nodes in all racks in the cluster have the base services such as file server, node manager, spark, and NFS gateway services installed.

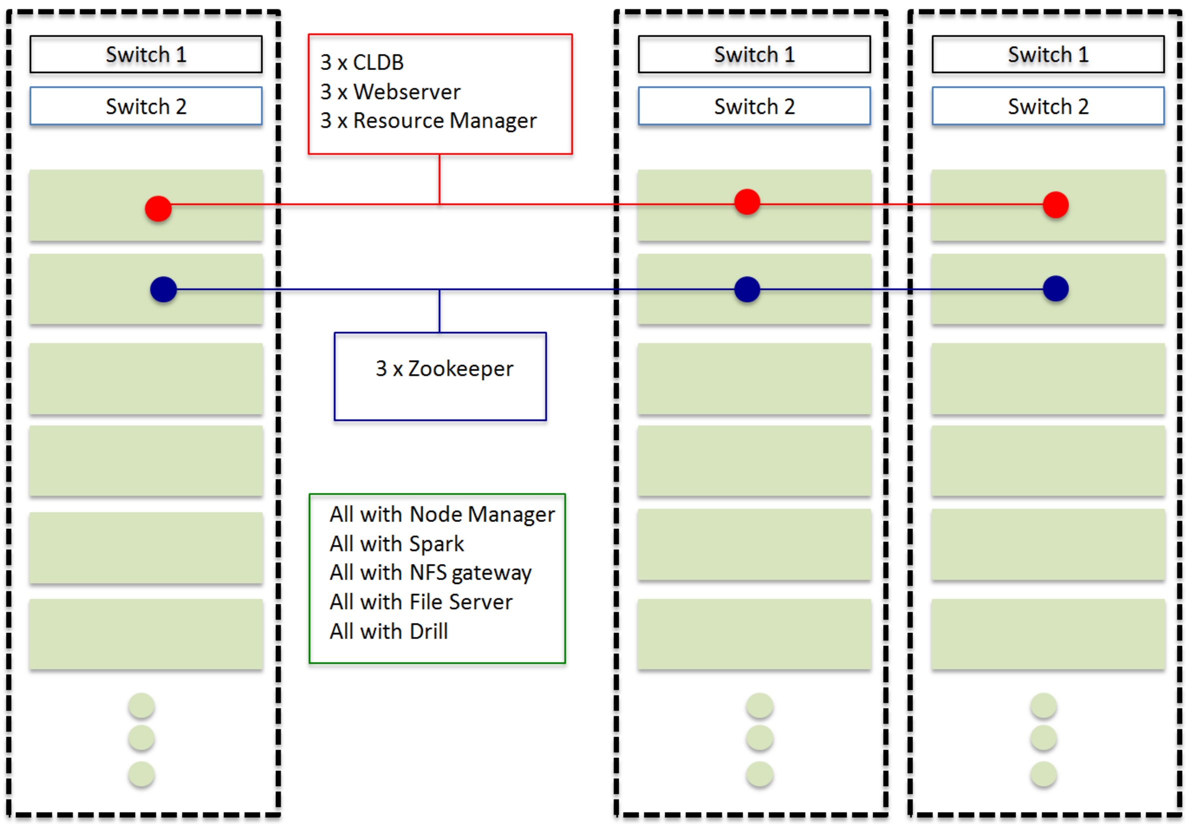

Figure 9: MapR Large Multi Rack Service Layout

6.3 Systems management

Systems management includes cluster systems management and hardware management.

Cluster systems management uses MapR Control System (MCS), which places the management services on servers across the cluster.

Hardware management uses the Lenovo XClarity™ Administrator which is a centralized resource management solution that reduces complexity, speeds up response, and enhances the availability of Lenovo® server systems and solutions.

The Lenovo XClarity Administrator provides agent-free hardware management for Lenovo's System x® rack servers and Flex System™ compute nodes and components, including the Chassis Management Module (CMM) and Flex System I/O modules. Figure 10 shows the Lenovo XClarity administrator interface, in which Flex System components and rack servers are managed and are seen on the dashboard. Lenovo XClarity Administrator is a virtual appliance that is quickly imported into a virtualized environment server configuration.

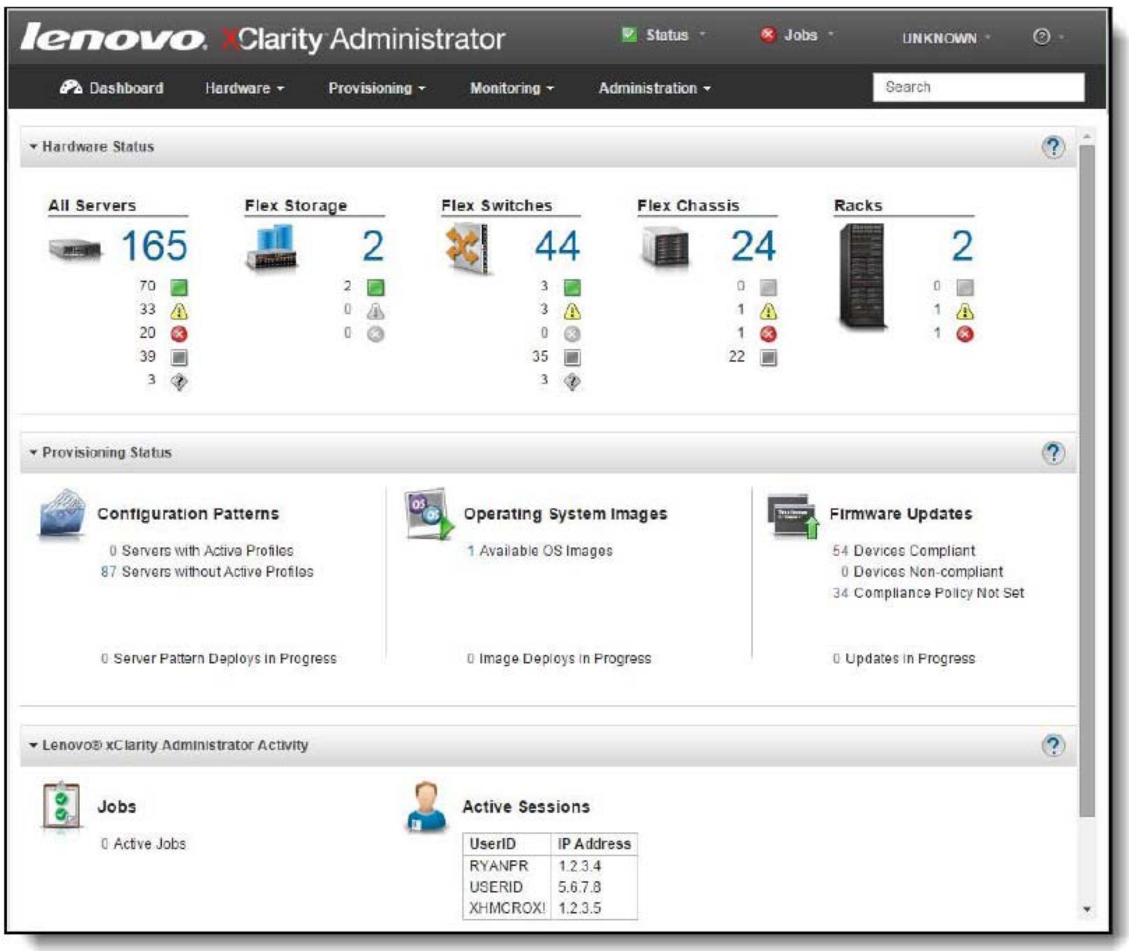

Figure 10: XClarity Administrator interface

Also, 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 and automate the management of cluster nodes. For more information about xCAT, see the Resources section at the end of this document.

6.4 Networking

Regarding networking, the reference architecture specifies two networks: a data network and an administrative or management network. Figure 11 shows the networking configuration for.

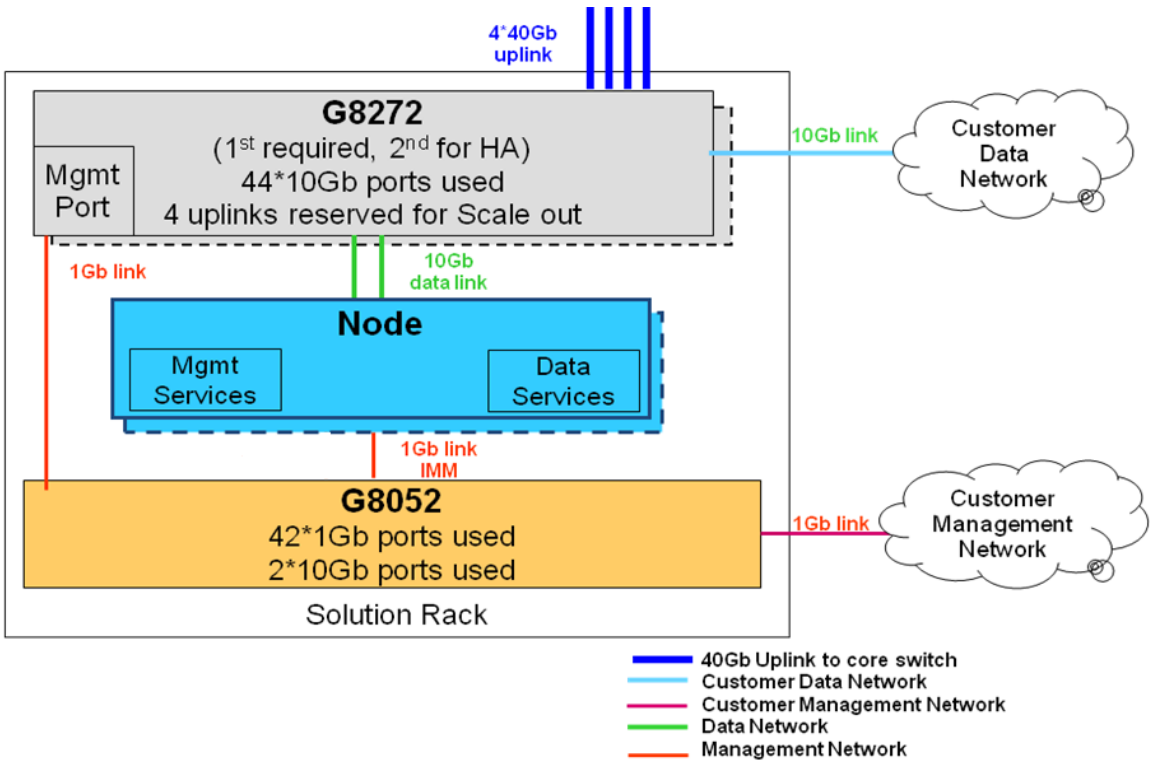

Figure 11: MapR network configuration

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 MapR file system (MapR-FS). The MapR cluster typically connects to the customer's corporate data network.

Two top of rack switch is required for the data network that is used by MapR. Two 10GbE switches are used. to provide the required network bandwidth between nodes. 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 1GbE network that is used for in-band operating system administration and out-of-band hardware management. In-band administrative services, such as SSH or Virtual Network Computing (VNC) that is running on the host operating system enables access to cluster nodes by system administrators for tasks such as software upgrades and problem determination. Through the integrated management modules II (IMM2) within the System x3650 M5 server, out-of-band management enables the hardware-level management of cluster nodes, such as hardware failure alert notifications, remote control of the node for problem determination, and firmware updates.

MapR has no dependency on the IMM2. Based on customer requirements, the administration links and management links can be segregated onto separate VLANs or subnets. The management network is typically connected directly to the customer's administrative network. When the in-band administrative services on the host operating system are used, MapR is configured to use the data network only. By default, MapR uses all the available network interfaces. 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 M5 servers, 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 consists of a single network topology. For cross rack communication between multiple racks, Lenovo RackSwitch G8332 core switches per cluster is recommended. In this case, the second Broadcom 10 GbE port can be connected to the second Lenovo RackSwitch G8272.

Figure 12 on page 20 shows how the network is configured when the MapR cluster is installed across more than one rack. The data network is connected across racks by two aggregated 40GbE uplinks from each rack's G8272 switch to a core G8332 switch.

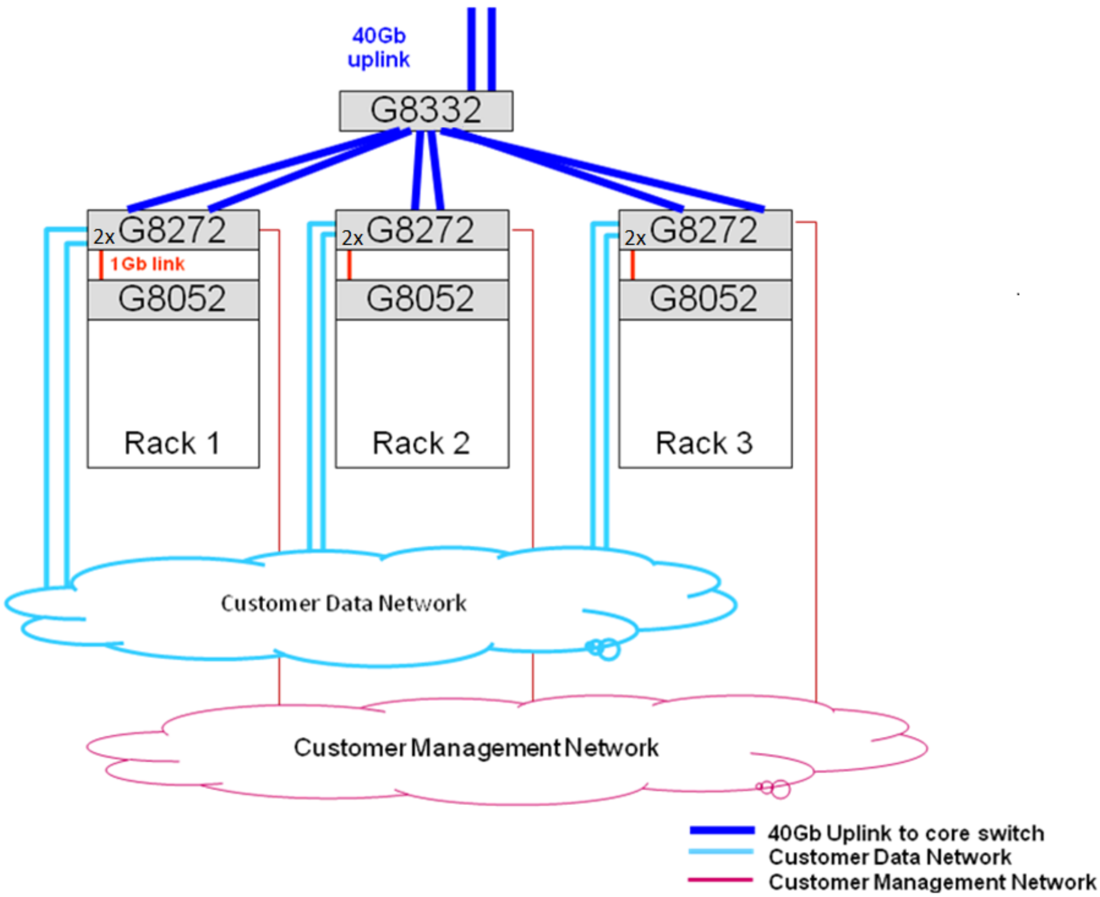

Figure 12: MapR cross rack network configuration[1]

A 40GbE core switch is recommended for interconnecting the data network across multiple racks. Lenovo System Networking RackSwitch G8332 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/management VLAN across cluster racks through the G8332 core switch. Many other cross rack network configurations are possible and may be required to meet the needs of specific deployments or to address clusters larger than three racks.

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

---

[1] To simplify the diagram, only one G8272 is drawn in the diagram, and in recommended configuration, two G8272 are configured for HA.

6.5 Predefined cluster configurations

The intent of the four predefined configurations in this reference architecture is to ease initial sizing for the customer and to show example starting points for different sized workloads. The reference architecture is not limited to these four cluster sizes:

- The starter rack configuration consists of three nodes and a pair of rack switches and is only recommended for non-production or proof of concept purposes due to lack of data redundancy if a node fails.
- The half rack configuration consists of ten nodes and a pair of rack switches.
- The full rack configuration (a rack fully populated) consists of 19 nodes and a pair of rack switches.
- The multi-rack contains a total of 57 nodes; 19 nodes and a pair of switches in each rack. A G8332 aggregation switch is included.

Table 2 lists the four predefined configurations for the MapR 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 6 TB drives (raw storage) are used and the useful amount of data space available. Available data space assumes the use of MapR replication with three copies of the data, and 25% capacity that is reserved for efficient file system operation and to allow time to increase capacity if needed. Available data space might increase significantly with MapR automatic compression. The estimates that are listed in Table 2 do not include extra space that is freed up by using compression because compression rates can vary widely based on file contents.

Table 2: Predefined configurations

	Starter Rack (non-prod)	Half Rack	Full Rack	Multi-Rack
Raw storage (4TB)	168 TB	560 TB	1120 TB	3360 TB
Available data space (4TB)	42 TB	140 TB	280 TB	840 TB
Raw storage (6TB)	252 TB	840 TB	1680 TB	5040 TB
Available data space (6TB)	63 TB	210 TB	420 TB	1260 TB
Number of nodes	3	10	19	57
Number of Racks	1	1	1	3
Number of 10 GbE cables	6	20	38	114
Number of 1 GbE cables	6	20	38	114

The number of nodes that are required in the cluster to support these four predefined configurations are shown in Table 2. These are the estimates for highly available clusters. Three nodes will support 63 TB of total data. Ten nodes are needed to support a deployment for 210 TB of data, and so on.

Figure 13 shows an overview of the architecture in two different one-rack sized clusters without network redundancy: a half rack and a full rack. Figure 14 shows a multi-rack-sized cluster without network redundancy.

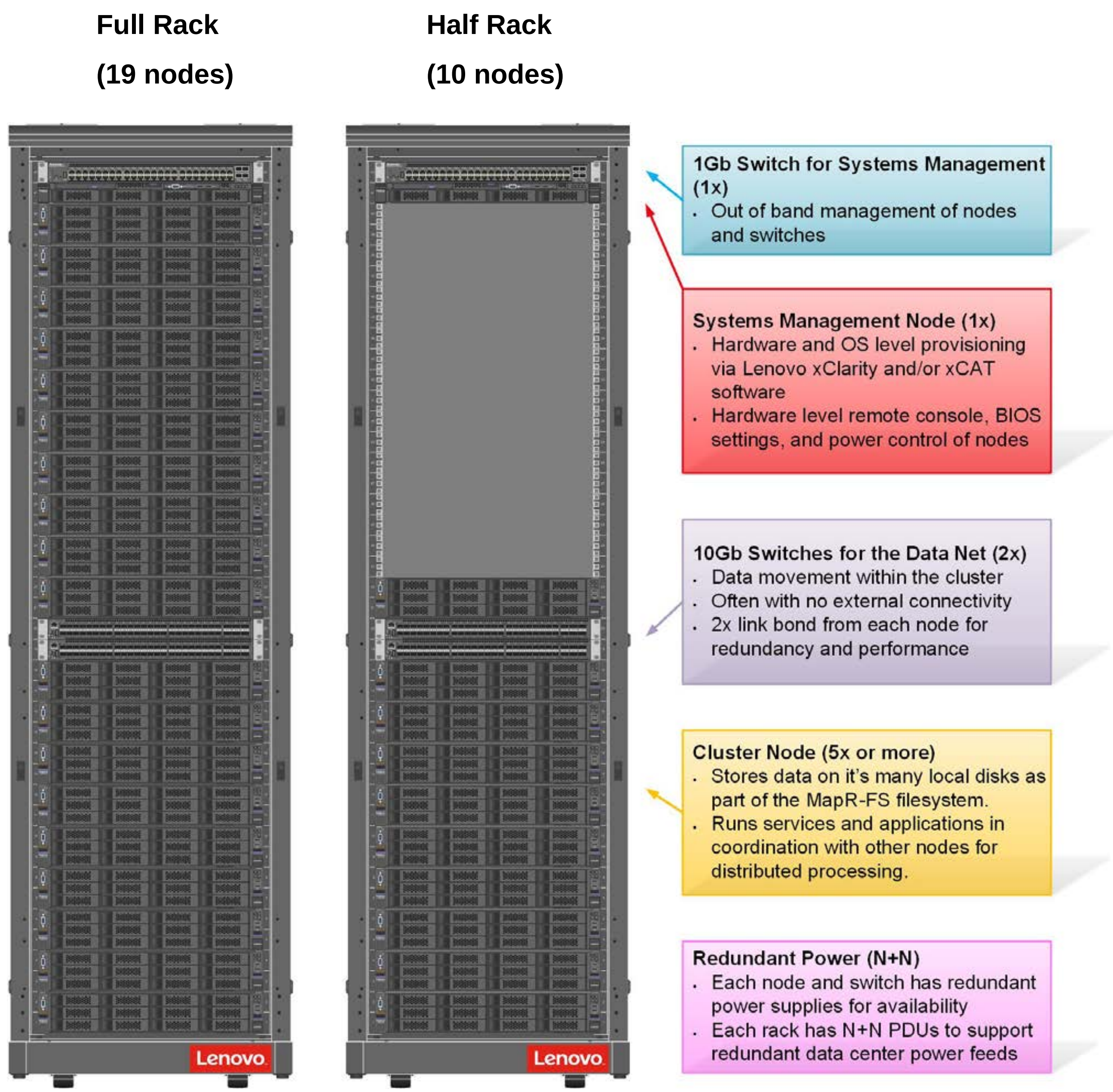

Figure 13: Half rack and full rack MapR predefined configurations

Multi-rack

(57 nodes)

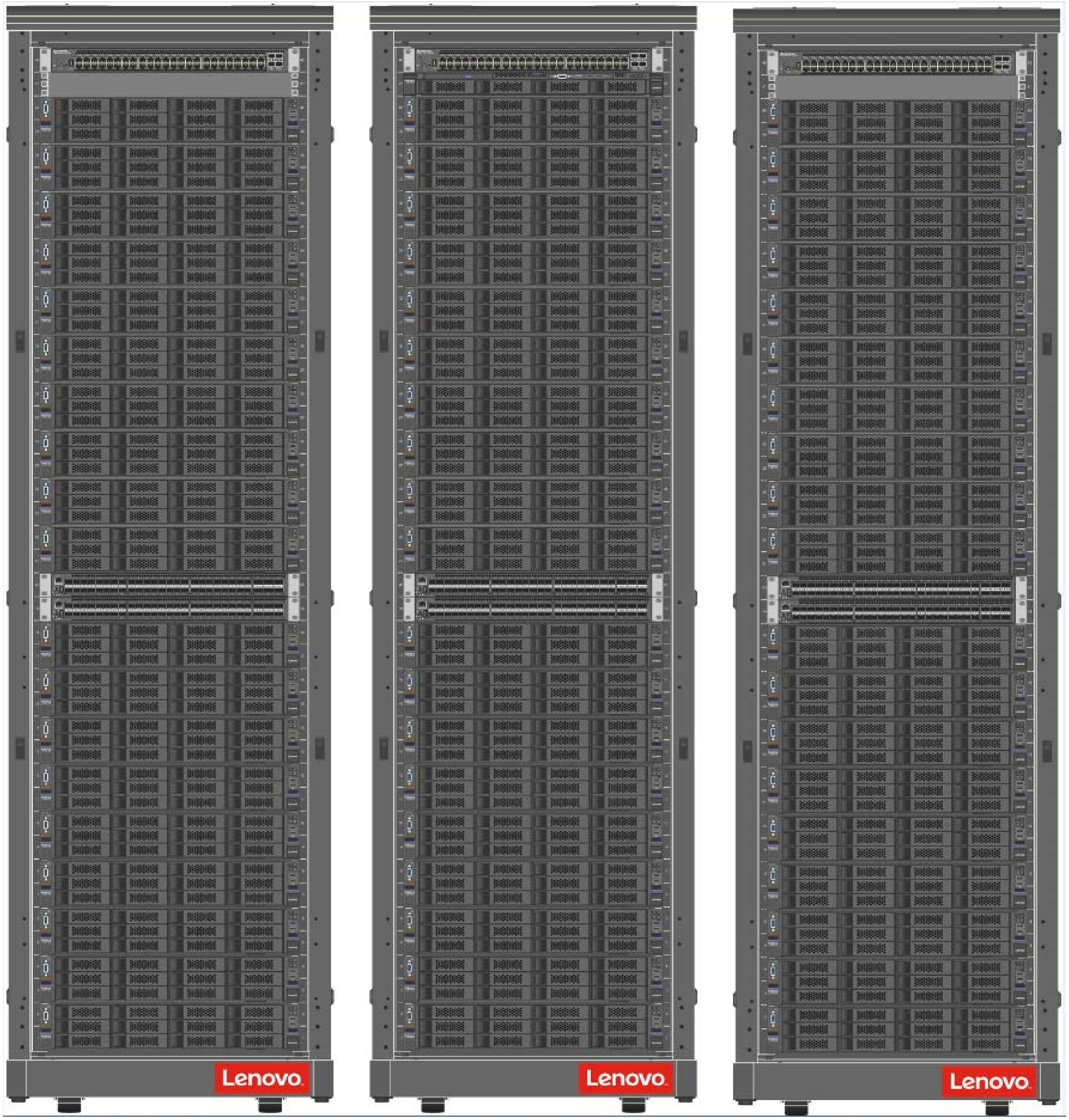

Figure 14: Multi-rack MapR predefined configuration

7 Deployment considerations

This section describes other considerations for deploying the MapR 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 of designing for higher performance.

In the MapR predefined configuration, node memory can be increased to 512 GB by using 16 32GB 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 MapR reference architecture solution. When lower-cost options are considered, it is important to understand the potential lower performance implications. A lower cost version of the MapR reference architecture can be achieved by using lower cost node processors (CPUs) and lower cost cluster data network infrastructure.

The node processors can be substituted with the Intel 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.

Lowering the network infrastructure cost is possible, but can also have a substantial negative effect on intra-cluster data throughput and cluster ingest rates. The following substitution can be made from this reference architecture for a multi-rack configuration:

- Within each cluster, substitute the Lenovo RackSwitch G8332 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, 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 jobs which helps to distribute the I/O load to different nodes cross the cluster.
- Ingest when cluster loading is not high, where 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 Estimating disk space

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

- For improved fault tolerance and performance, the MapR file system replicates data blocks across multiple cluster 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. MapR provides automatic compression. Available data space might increase significantly with MapR automatic compression. 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 the MapR file system, the raw data disk space, and the required number of nodes can be estimated by using the following equations:

Total raw data disk space = (User data, uncompressed) * (4 / compression ratio)
Total required nodes = (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 principles, Table 3 shows an example for a cluster that must store 500 TB of un-compressed user data. The example shows that the MapR cluster will need to have 800TB of raw disk storage 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 nine nodes, or nearly a half rack are required to support deployment of this size.

Table 3: Example of storage sizing with 6 TB 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 MapR cluster	800 TB
- Storage needed for MapR-FS 3x replication	600 TB
- Reserved Storage for headroom	200 TB
Raw data disk per node (with 6 TB drives)	84 TB
Minimum number of nodes required	10

7.5 Scaling considerations

The MapR 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 MapR cluster is scalable by adding System x3650 M5 nodes and network switches and adding management services and optional services on those nodes as required. The MapR architecture allows linear scalability to trillions of files and thousands of petabytes.

As the capacity of racks is reached, new racks can be added to the cluster. When a MapR reference architecture implementation is designed, future scale out should be a key consideration in the initial design - both: 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 in Figure 12 on page 20 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 just one of the racks). Starting with multiple racks can enforce proper network topology and prevent future re-configuration and hardware changes. When new racks are added over time, multiple G8332 switches might be required for greater scalability and balanced performance.

Also, as the number of nodes within the cluster increases, so do many of the tasks of managing the cluster, such as updating node firmware or operating systems. Building a cluster management framework as part of the initial design and proactively considering 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.

7.6 High Availability (HA) considerations

When a MapR cluster on System x is implemented, consider availability requirements as part of the final hardware and software configuration. Typically, a standard Hadoop deployment has some high availability features, but MapR enhancements make it more highly available for mission-critical environments. MapR best practices provide significant protection against data loss. MapR ensures that failures are managed without causing an outage. There is redundancy that can be added to make a cluster even more reliable. Some consideration must be given to hardware and software redundancy.

7.6.1 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 MapR-FS or Hadoop functionality, but it might affect the management of the cluster; therefore, availability requirements must be considered.

MapR provides application-level Network Interface Card (NIC) bonding for higher throughput and high availability. Customers can either choose MapR application-level bonding or OS-level bonding and switch-based aggregation of some form matching the OS bonding configuration when using multiple NICs. Virtual Link Aggregation Groups (vLAG) can be used between redundant switches. If 1Gbps data network links are used, it is recommended that more than one is used per node to increase throughput.

7.6.2 Hardware availability considerations

With no single point of failure, redundancy in server hardware components is not required for MapR. MapR automatically and transparently handles hardware failure resulting in the loss of any node in the cluster running any data or management service. The default three-way MapR replication of data ensures that no data is lost because two additional replicas of data are maintained on other nodes in the cluster. MapReduce tasks from failed nodes are automatically started on other nodes in the cluster. Failure of a node running any management service is automatically and transparently recovered as described in the following services.

- All ZooKeeper services are available for read operations, with one acting as the leader for all writes. If the node running the leader fails, the remaining nodes will elect a new leader. Most commonly, three ZooKeeper instances are used to allow HA operations. In some large clusters, five ZooKeeper instances are used to allow fully HA operations even during maintenance windows that affect ZooKeeper instances. The number of instances of ZooKeeper services that must be run in a cluster depends on the cluster's high availability requirement, but it should always be an odd number. ZooKeeper requires a quorum of (N/2)+1 to elect a leader where N is the total number of ZooKeeper nodes. Running more than five ZooKeeper instances is not necessary.
- All CLDB services are available for read operations, with one acting as the write master. If the node running the master CLDB service goes down, another running CLDB will automatically become the master. A minimum of two instances is needed for high availability.
- One ResourceManager service is active. Other ResourceManager instances are configured but not running. If the active ResourceManager goes down, one of the configured instances automatically takes over without requiring any job to restart. A minimum of two instances is needed for high availability.
- All NFS servers are active simultaneously and can present an HA NFS server to nodes external to the cluster. To do this, specify the virtual IP addresses for two or more NFS servers for NFS high availability. Additionally, use round-robin Domain Name System (DNS) across multiple virtual IP addresses for load balancing in addition to high availability. For NFS access from within the cluster, NFS servers should be run on all nodes in the cluster and each node should mount its local NFS server.
- The MapR web server can run on any node in the cluster to run the MapR Control System. The web server also provides a REST interface to all MapR management and monitoring functions. For HA, multiple active web servers can be run with users connecting to any web server for cluster management and monitoring. Note that even with no web server running, all monitoring and management capabilities are available using the MapR command line interface.
- Within racks, switches and nodes have redundant power feeds with each power feed connected from a separate PDU.

7.6.3 Storage availability

RAID data disk configuration is not necessary and should be avoided in MapR clusters. The use of RAID causes a negative impact on performance. MapR provides automated setup and management of storage pools. The three-way replication provided by MapR-FS provides higher durability than RAID configurations because multiple node failures might not compromise data integrity.

If the default 3x replication is not sufficient for availability requirements, the replication factor can be increased on a file, volume, or cluster basis. Replication levels higher than 5 are not normally used. Mirroring of MapR volumes within a single cluster can be used to achieve very high replication levels for higher durability or for higher read bandwidth. Mirrors can be used between clusters as well. MapR efficiently mirrors by only copying changes to the mirror. Mirrors are useful for load balancing or disaster recovery.

MapR 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.

Lenovo storage adapters provide a true JBOD configuration for best performance. In cases where only a RAID controller is available with no JBOD configuration, RAID0 may be configured but with a single HDD per RAID array. This will most closely emulate the JBOD configuration. Multiple HDDs in a single RAID0 array are not recommended nor required since a failure of a single HDD will cause all HDDs in that array to go off-line.

7.6.4 Software availability considerations

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

The MapR Platform is unique because it was designed with a “no NameNode” architecture for high availability. MapR is designed with no single point of failure and no single bottleneck for data access. With MapR, the file metadata is replicated, distributed, and persistent, so that there is no data loss or downtime even in the face of multiple disk or node failures.

The MapR ResourceManager HA improves recovery time objectives and provides for a self-healing cluster. Upon failure, the MapR ResourceManager automatically restarts on another node in the cluster. NodeManagers can automatically pause and then reconnect to the new ResourceManager. Any currently running jobs or tasks continue without losing any progress or failing.

You can easily set up a pool of NFS nodes with HA and failover using virtual IP addresses. If one node fails, the virtual IP addresses will be automatically reassigned to the next NFS node in the pool.

It is also common to place an NFS server on every node where NFS access to the cluster is needed.

7.7 Migration considerations

If migration of data or applications to MapR is required, you must consider the type and amount of data to be migrated and the source of the data being migrated. Most data type can be migrated, but you must understand the migration requirements to verify viability. Standard Hadoop tools such as *distcp* (distributed copy) can be used to migrate data from other Hadoop distributions. For data in a POSIX file system, you need to NFS or FUSE client mount the MapR cluster and use standard Linux commands to copy the files into the MapR cluster. Either Sqoop or database import/export tools with MapR NFS can be used to move data between databases and MapR.

You also need to consider whether applications must be modified to use Hadoop functionality. With the MapR read/write file system that can be mounted by a standard NFS client, your applications continue to work with no code changes, so you can completely avoid the significant effort of rewriting applications to conform to Hadoop APIs.

8 Predefined Configurations (Bill of Material)

The predefined Lenovo hardware configurations discussed in this document for the MapR 5.2 Converged Data Platform are available in the System x and Cluster Solutions configurator (x-config) tool. The pre-configurations are the full bill of material with part number, descriptions, and quantities used in this reference architecture and can be used for ordering purposes. Table 2 on page 21 summarizes the core components specified for each of the predefined configuration sizes.

The pre-configurations can be viewed, modified, and saved locally with the x-config tool. The installation link for x-config is below:

https://lesc.lenovo.com/products/hardware/configurator/worldwide/bhui/asit/install.html

A summary for installing x-config on your workstation is below:

1. Access x-config webpage at the above link.

2. On the Workstation requirements tab, use provided link to install Java

3. On Install/Update tab, install x-config application onto your workstation

4. Launch x-config: Windows Start - All Programs - x-config - System x and Cluster Solutions configurator

X-config can be used to perform these functions:

- View and modify and validate Lenovo/MapR Converged Data Platform configurations
- Open predefined *.cse or *.xml files stored locally on your workstation
- Translate user designs into configurations that can be priced, ordered, and built
- View graphical system diagrams to review expansion options, explore configuration alternatives and validate solution level characteristics
- View list prices as the configuration is being constructed
- Output hardware configuration data into a human readable XLS file

9 Acknowledgements

This reference architecture has benefited very much from the detailed and careful review provided by:

Lenovo business review

- Prasad Venkatachar – Sr. Solutions Product Manager

MapR technical review

- Dale Kim, Sr. Director, Product Marketing

MapR business review

- Tracy Lothringer, Sr Director, Worldwide Technology Alliances
- Christine Sheets, Sr. Director, Global OEM and Embedded Technology

10 Resources

For more information, see the following resources:

- Lenovo System x3650 M5:
  - Product page: shop.lenovo.com/us/en/systems/servers/racks/systemx/x3650-m5/
  - Lenovo Press product guide: http://lenovopress.com/lp0068
- 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
- MapR:
  - MapR main website: www.mapr.com
  - MapR products: www.mapr.com/products
  - MapR editions overview: www.mapr.com/products/mapr-distribution-editions
  - MapR architecture overview: www.mapr.com/why-hadoop/why-mapr/architecture-matters
  - MapR blogs: www.mapr.com/blog
  - MapR Resources: www.mapr.com/resources
  - MapR products and differentiation: mapr.com/why-mapr
  - MapR documentation: maprdocs.mapr.com
  - MapR technical training: www.mapr.com/training
  - MapR getting started: mapr.com/products/hadoop-download/
- Open source software:
  - Hadoop: hadoop.apache.org
  - Pig: pig.apache.org
  - Cascading:www.cascading.org
  - Spark:spark.apache.org
  - Apache Tez:tez.apache.org
  - Mahout: mahout.apache.org
  - Hive: hive.apache.org
  - Impala:https://impala.incubator.apache.org/
  - Drill:drill.apache.org
  - Solr :lucene.apache.org/solr
  - Sqoop: sqoop.apache.org
  - Flume: flume.apache.org
  - Hue:gethue.com
  - Sentry: sentry.incubator.apache.org
  - Oozie: oozie.apache.org
  - ZooKeeper: zookeeper.apache.org
  - Sahara:wiki.openstack.org/wiki/Sahara
- Other resources
  - xCat: xcat.sourceforge.net

11 Document History

Version	Date	Notes
1.0	9/17/2015	Initial release version
2.0	3/30/2017	Update for Mapr5.2 and Lenovo x3650 Broadwell servers

Trademarks and special notices