IBM
Hortonworks Data Platform on IBM Power

Reference Architecture and Design
Version 2.0

IBM Power Development

August 30, 2018

POL03270USEN-00

Table of Contents

1 Introduction

1.1 Purpose of Document

This document is intended to be used as a technical reference by IT professionals who are defining and deploying solutions for Hortonworks Data Platform (HDP) on IBM® Power® clusters (hereafter referred to as "HDP on Power"). This document describes the architecture for the HDP on Power along with a related reference design that complies with the architecture. The architecture is intended to serve as a guide for designs. The reference design is intended to provide a useful example configuration that can be used to more easily construct suitable designs for specific deployments.

1.2 Document Content and Organization

The core content of this document consists of an architecture and a reference design for an HDP on Power solution. This document provides context and background by beginning with a review of the **objectives**, **scope**, and **requirements** that apply to the solution. The **architecture** follows with an outline of **key concepts**, followed by a presentation of the architecture – covering the primary elements and how they are composed to form the solution. Finally, a **reference design** is presented that conforms to the architecture.

1.2.1 Architecture versus Design

Within this document, a relevant distinction is made between architecture and design.

1.2.1.1 Architecture

Architecture in this document and context refers to key concepts, components, roles and responsibilities, models, structures, boundaries, and rules, which are intended to guide and govern the designs for the solution and the components that comprise the solution.

Consistent with a good architecture, the elements included in the architecture are intended to remain largely intact and relatively stable over time (as compared to the underlying designs). For components that are not compliant, the architecture provides the direction toward which these components should evolve. For components that are compliant, the architecture provides the boundaries within which designs can further evolve, improve, and otherwise achieve various goals.

It is a goal of the architecture to supply the right balance of prescriptiveness (to help ensure success and goals are achieved) and latitude (to allow designers and developers as many degrees of freedom as possible).

Throughout the architecture, references to preferred or recommended design selections are sometimes included for clarity and convenience, but these should not be considered as restrictive.

1.2.1.2 Design

Design represents a fairly specific description of a solution that is sufficient to allow the solution to be realized. For this solution, the reference design in this document (see section 7 “Reference Design 1.0A - Minimum Production Configuration” on page 29) describes the specific components and elements that comprise the solution, specific variations of the solution, and how these are to be interconnected, integrated, and configured.

1.2.2 Key Influences

This version of this architecture was influenced by preceding related work (specifically, the Hadoop and Hortonworks architecture and design principles and best practices, IBM Data Engine for Hadoop and Spark, and IBM Data Engine for Analytics). To the extent possible, consistency in terminology was maintained from those works, and this architecture document bridges to those works as appropriate.

1.3 References

[1] IBM Hortonworks Data Platform on IBM Power - Reference Architecture and Design - Version 1.0
https://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=POL03270USEN&

[2] IBM Hortonworks Data Platform on IBM Power with IBM Elastic Storage Server –Reference Architecture and Design - Version 1.0
https://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=POL03281USEN&

[3] Hortonworks. (2018). HDP Documentation Version 3.0
http://docs.hortonworks.com/HDPDocuments/HDP3/HDP-3.0/index.html

[4] Apache Software Foundation. (2017). Welcome to Apache Hadoop
http://hadoop.apache.org/

[5] Hortonworks. (2017). Hortonworks Data Platform (HDP)
https://hortonworks.com/products/data-center/hdp/

[6] Mellanox IBM Quick Reference Guide Solutions
http://www.mellanox.com/oem/ibm/rel_docs/IBM_qrsg.pdf

2 Objective

The objective of this architecture and reference design is to provide a guide for those designing systems to host HDP installations using IBM Power Systems™. Primary users of such a system are data scientists and data analysts who do not necessarily possess deep IT skills. Primary administrators are those who manage and maintain the infrastructure, those who manage and maintain the Hadoop platform, and those who manage and maintain the applications used by the primary users.

2.1 Scope

For the purpose of defining the scope of this architecture (and providing some basic orientation), it is necessary to understand the high level architectural elements (see section 5 “Architecture - Overview” on page 10). While all of these elements are part of the solution addressed by this architecture, the scope of this architecture and reference design does not cover all these elements equally. Specifically:

2.1.1 Data

This architecture covers how this data is *hosted* and *accessed*. The form and nature of the data is not within the scope of this architecture except to note that it may be of any form that is consistent with the Hadoop platform data models—typically data stored within the Hadoop Distributed File System (HDFS).

2.1.2 Applications

This architecture covers how these applications are *hosted* and *executed*. The form and nature of these applications is not within the scope of this architecture except to note that they may be any form consistent with the Hadoop platform application models—typically distributed applications that are run across a cluster, often using a MapReduce programming model.

2.1.3 Platform

This architecture prescribes the HDP suite as the platform. However, it is important to note that this architecture does not prescribe any particular architecture or design *within* the HDP installation itself, and a suitable design of a HDP software installation (for example, specific choices of HDP components to include, distribution of these components across nodes, configuration of these components, and so on) is necessary for a complete and operational environment. This architecture assumes common design patterns and best practices for the HDP installation as recommended by Hortonworks, but the intent and scope of this architecture is to provide a hosting environment that can accommodate any reasonable HDP design.

2.1.4 Infrastructure

The Infrastructure hosts the Platform (HDP) directly, and it provides the next level hosting and access mechanisms for the Data and the next level hosting and execution for the Applications. The Infrastructure is the primary scope of this architecture and reference design.

3 Requirements

This section contains the requirements that apply to this solution. The following list is intended to be as complete as possible, including requirements that are not, or may not be, satisfied in this version of the architecture and reference design. This helps to ensure a fuller and more complete representation of the requirements that apply in general. Requirements that are not addressed directly by this revision are marked as [tbd].

The solution must:

- Provide a solution that supports any reasonable HDP design pattern.
- Provide efficient use of system resources.
- Provide industry-leading performance.
- Provide industry-leading price-performance.
- Provide a highly scalability infrastructure. Deployments ranging from a few nodes to hundreds of nodes must be supported. Deployments ranging from a few TB to several PB must be supported.
- Be well integrated. The elements of the solution must be configured and connected such that they form a coherent and well-functioning system.
- Be easy to extend and grow. It is expected that many deployments will begin with a base set-up of capacity that will need to grow as usage and demand increases. Further, it is a requirement that most extensions to the system be sufficiently easy to accomplish so that a client's administrators can accomplish them without requiring specialized services or consultants.
- Be highly reliable and resilient. This requirement has multiple dimensions and applies at multiple levels, but common requirements with respect to reliability and resilience generally apply.
- Include options for, and accommodate both, Ethernet and InfiniBand® network fabrics. [tbd, Ethernet only in this version]
- Include options for GPU accelerators.
- Accommodate a wide range of system capacities. Specifically:
  - A client must be allowed to select the compute capacity of the solution within a wide range of limits.
  - A client must be allowed to select the storage capacity of the solution within a wide range of limits.
  - A client must be allowed to select the network capacity of the solution within a wide range of limits
  - A client need not be generally prevented from specifying a configuration that may be (or appear to be) unbalanced or otherwise suboptimal. While this architecture provides guidance for specifying a "good" configuration, it is the responsibility of the parties that configure a system to ensure that it meets the client's requirements for capacity, performance, resilience, and so on.
- Accommodate a reasonable set of configuration variations. It is recognized that clients have varying requirements. It is also recognized that clients have varying preferences or policies with respect to vendors, Linux distributors or similar suppliers, and that they may also have legacy infrastructure. Specific variations that are recognized as requirements include:

- A reasonable set of Linux distributions must be supported. The specific set that is chosen will likely need to accommodate nimble, early adopters (for example, distributions that rapidly release support that is relevant to Power) and those who require reliable, long-term support (for example, distributions that have good enterprise characteristics). [tbd, Red Hat only in this version]
- Environments that range from minimally resilient to as highly resilient as the solution components allow must be supported.

The reference architecture and/or designs must:

- Define a minimum configuration for a production environment.
- Define variations for *balanced*, *performance*, and *storage dense* options.
- Define recommendations for default, minimum, and maximum parameters.

3.1 Non-requirements

The following items are not requirements. In other words, they need not be part of the solution. These are listed explicitly for clarity (for example, to help reduce the need for assumptions) and to assist the reader in understanding the capabilities of the solution. Note that some non-requirements, and limitations on requirements, are also included in the requirements section mentioned earlier.

The solution need *not*:

- Provide a unified (that is, *single pane of glass*) management capability. For example, the management of the infrastructure may be distinct and separate from the management of the HDP installation.
- Support any particular standard beyond those that are commonly supported within an IT infrastructure.

4 Concepts

4.1 Terminology - Common Usage

Several terms are introduced and used through this document. This section contains some of the more important or relevant of these, along with an initial description of their common or ordinary meaning. Note that as the architecture is described in the later sections, some of these terms can be used and defined in a more precise manner to refer to elements of the architecture.

4.1.1 Solution

A *solution* refers to the set of elements assembled and configured to accomplish some purpose. Within this document, the solution is the complete, operational environment that stores and processes the client data.

4.1.2 System

A *system* refers to the totality of elements that forms the solution (for example, hardware, software, servers, storage, network).

4.1.3 Cluster

A *cluster* refers to a collection of servers or nodes. Within this document, a cluster refers to the entire collection of nodes that are included in the system.

4.2 Roles

This architecture recognizes the following roles for people that interact with the system.

4.2.1 User

A User submits jobs (runs applications) to obtain results. A Data Analyst and a Data Scientist are common examples of Users for this solution.

4.2.2 Application Developer

The Application Developer creates the analytics applications to be run by the Users.

4.2.3 Platform Admin

The Platform Admin configures, manages, and maintains the HDP installation. Ambari is a key component for this role, providing a primary tool for the administration of the HDP components and installation.

4.2.4 Infrastructure Admin

The Infrastructure Admin administers and manages the server, storage, network, operating systems, and system software. In practice, this role can be divided by specialty (for example, storage administration, network administration), but the nature of the role remains similar across specialties.

5 Architecture - Overview

It is useful to start with a very basic, high-level architectural overview to provide some context and orientation for the following content.

5.1 Elements

5.1.1 Data

Data in this context is the client data (typically "big data") specific to a client which the client wishes to analyze. This architecture covers how this data is *hosted* and *accessed*. The form and nature of the data is not within the scope of this architecture except to note that it may be of any form consistent with the Hadoop platform data models –typically data stored within the Hadoop Distributed File System (HDFS).

5.1.2 Applications

Applications in this context are the big data analytics applications specific to a client's data and the analysis the client wishes to perform. This architecture covers how these applications are *hosted* and *executed*. The form and nature of these applications is not within the scope of this architecture except to note that they may be of any form consistent with the Hadoop platform application models – typically distributed applications that are run across a cluster, often using a MapReduce programming model.

5.1.3 Platform

The Platform in this context is the application-level software that creates the environment which provides the first-level hosting of the client data and analytics applications and the means to run these applications and access the data. This is essentially a Hadoop-style *platform*. This architecture prescribes the HDP suite as this platform. Though HDP support for Power began with HDP version 2.6, this architecture version requires HDP version 3.0 or later.

However, it is important to note that this architecture does not prescribe any particular architecture or design *within* the HDP installation proper, and a suitable design of a HDP software installation (for example, specific choices of HDP components includes distribution of these components across nodes, configuration of these components) is necessary for a complete and operational environment. This architecture assumes common design patterns and best practices for the HDP installation as recommended by Hortonworks, but the intent and scope of this architecture is to provide a hosting environment that can accommodate any reasonable HDP design.

5.1.4 Infrastructure

Infrastructure is the set of hardware -- servers, storage, network --, and all of the system software, firmware, and infrastructure management software that are necessary to host and support the these elements. The Infrastructure hosts the Platform (HDP) directly, and it provides the next level hosting and access mechanisms for the Data and the next level hosting and execution for the Applications. The Infrastructure is the primary scope of this architecture.

5.2 Composition

The elements such as data, application, platform and infrastructure can be viewed in three layers.

The first and top-most layer is the application layer that includes the Applications and the Data. These elements are the content typically provided by a client, and these are the elements most relevant to the primary purpose of the system – analysis of large amounts of data. The Applications directly access (read and write) the Data.

Second is the platform layer, which is mostly composed of the HDP components. It may also include other third-party software components that serve various functions within the Platform. The Platform directly handles the task of hosting (execution) of the Applications by orchestrating their execution as a part of a Job. The Platform also directly hosts the Data by providing the Hadoop Distributed File System (HDFS) into which the Data is placed. The Platform also serves as the primary interface and touchpoint for Users of the system.

The third and bottom-most layer is the infrastructure layer that consists of the hardware and software elements mentioned earlier. The Infrastructure hosts the platform layer elements (especially HDP) directly, and it provides the next level hosting (the storage) and access mechanisms for the Data and the next level hosting and execution (the servers and OS) for the Applications. The Infrastructure layer also contains all of the network elements, software, and hardware, to provide connectivity for the Nodes in the system.

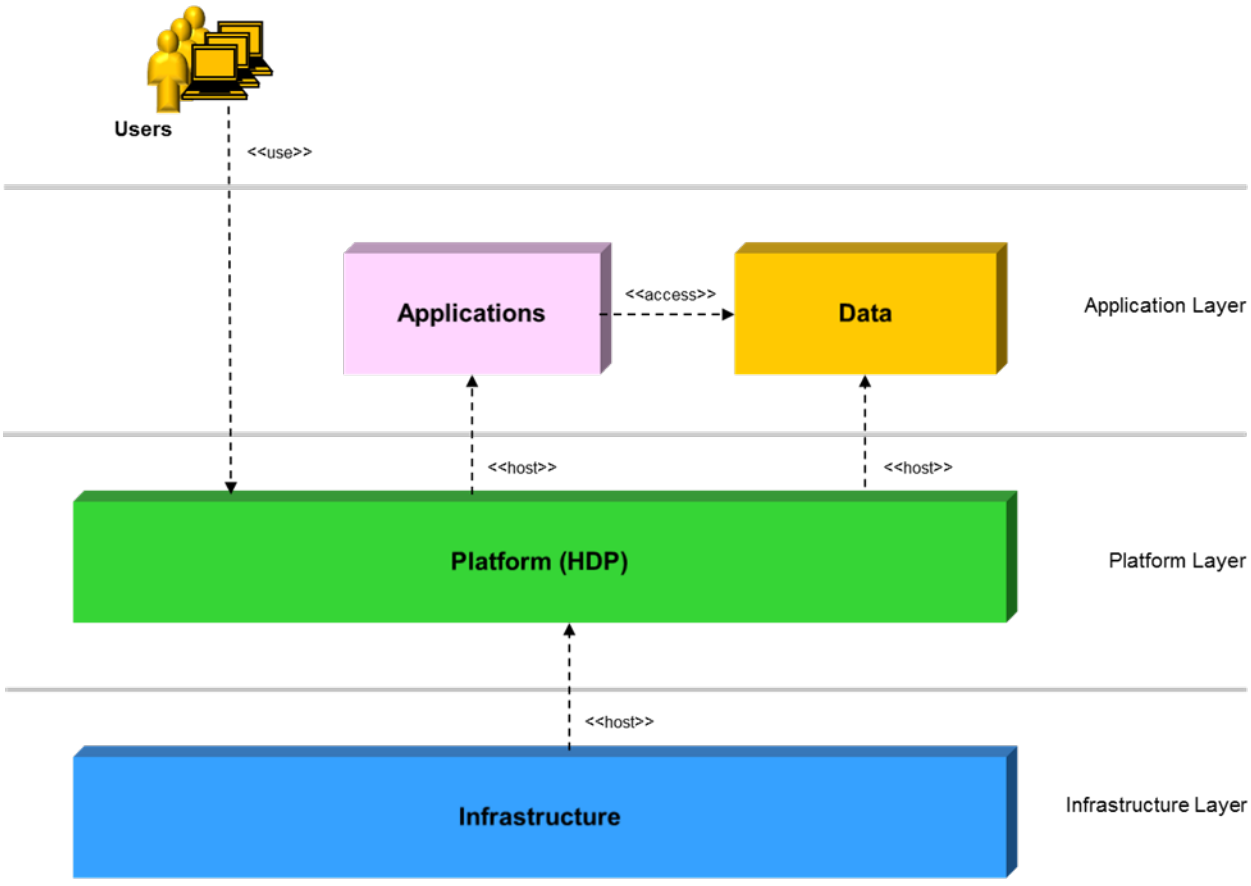

Figure 1. Architecture Overview - Top Level Elements

6 Architecture

This section describes the system architecture. The Platform and Infrastructure layers are treated together in this section as distinct but related layers of the system.

This section starts with the elements of the architecture and then covers their assembly into a system. Some readers may find it more instructive to start with a review of the resulting system before reviewing the elements. These readers are referred to section 6.3 “Summary Views” on page 27.

6.1 Elements

This section describes the elements relevant to this architecture.

6.1.1 Functions

6.1.1.1 Management Function

Most of the HDP components are Management Functions that run as services on Nodes in the Cluster. Management Functions accomplish a variety of purposes in the Cluster such as submission and control of job execution across the Cluster nodes and monitoring of other services to facilitate failover. Examples include YARN (ResourceManager, Node Manager), Oozie, and Zookeeper. Most of the Management Functions run on the Master Nodes.

6.1.1.2 Storage Function

Some of the HDP components are Storage Functions that run as services on Nodes in the Cluster. Examples include the Name Node and Data Node services for HDFS. Storage Functions are distributed across most (often all) of the nodes in the Cluster. The primary role of the Storage Functions is to realize HDFS (both server and client roles).

6.1.1.3 Edge Function

Some of the HDP components are Edge Functions that run as services on nodes in the Cluster. Edge Functions can be roughly divided into two categories: access control functions (for example, Knox) and functions which handle data movement into and out of the Cluster (for example, Flume or Sqoop). Most of the Edge Functions run on the Edge Nodes.

6.1.2 Hadoop Distributed File System (HDFS)

HDFS is the filesystem used to hold the Data. A variety of Storage Functions are used to realize HDFS in the Platform.

6.1.3 External Data Source

Data moving in and out of the Cluster will do so from and to some External Data Source. The specific nature and requirements for this External Data Source are out of scope for this architecture – limited only by the capability of the data import and export functions that connect to it. This element is specifically noted as the volume and the rate of data import and export are often important design criteria.

6.1.4 Cluster

The Cluster is the set of Nodes in the System.

6.1.5 Nodes

A Node is a server, and its associated system software, that is used by the Platform to host Functions and accomplish its role. The Platform design (Hadoop and HDP specifically) recognizes that it is running on a Cluster infrastructure, and concepts such as Nodes are explicitly visible and handled at this layer. Nodes are categorized into the following primary types:

6.1.5.1 Worker Node

A Worker Node serves two primary roles: First, each Worker Node contains some physical storage which is used for HDFS, and it hosts some Storage Functions that allow it to manage this storage as part of HDFS. These Storage Functions communicate and cooperate to form the distributed filesystem across the collection of Worker Nodes in the Cluster. Second, a Worker Node is also used by the Management Functions to execute Applications which are parts of Jobs. Job execution is typically distributed across multiple Worker Nodes to provide parallel execution of the Job.

There are typically three or more (often many more) Worker Nodes in a Cluster. Three Worker Nodes provides the ability to directly support the common HDFS replication factor of three.

Worker Nodes are usually the most common Node type in a Cluster, accounting for perhaps 80-90% (or more) of the Nodes in the Cluster.

6.1.5.2 Master Node

A Master Node is used to host Management Functions and some Storage Functions. There are typically one or more Master Nodes in a Cluster, and three Master Nodes are common to provide basic high availability (HA) capability.

6.1.5.3 Edge Node

An Edge Node serves as the host for functions which require both an "external" and an "internal" connection. Edge Nodes commonly provide the pathway into and out of a Cluster from any "external" person or element. Topologies vary, but in many deployments the Master and Worker Nodes have only internal (that is, private) connections to each other, and access to the cluster is controlled and routed through functions running on the Edge Nodes. One common case is User access through a component such as Knox that requires an external connection for the User which in turn allows a User to access internal functions (as selected and authorized). Another common case is data import and export using components such as Flume and Sqoop which require an external connection to some external data source and an internal connection to HDFS. Such components handle the import of Data into or the export of Data out of HDFS.

There are typically one or more Edge Nodes in a Cluster, and two Edge Nodes are common to provide basic high availability (HA) capability.

6.1.5.4 Utility Node

A Utility Node is more general Edge Node – specifically an Edge Node that also runs some Management Functions, for example Ambari. However, within this reference architecture, this term is not used further, and Edge Nodes are defined such that they may also host Management Functions.

6.1.5.5 Machine Learning/Deep Learning Node

A Machine Learning / Deep Learning node is a specialty worker node that can be added as an optional node to support Machine Learning, Deep Learning, or other workloads that use GPU capabilities.

6.1.5.6 Other Specialty Nodes

In addition to the Node types mentioned earlier, other specialty Nodes can be introduced to the Cluster to provide dedicated hosts and serve special functions. These can normally be simply considered as special cases of these node types, so specialty Node types are not explicitly covered further in this reference architecture.

6.1.5.7 System Management Node

The System Management Node is a server that hosts the software that accomplishes the provisioning and management of the Infrastructure. The System Management Node is not visible or used by the Platform. It is used exclusively for Infrastructure and Cluster-level purposes.

6.1.6 Platform Manager

The complexity of a Platform deployment typically requires a dedicated manager to accomplish the initial provisioning and ongoing maintenance and monitoring of the various components across the Cluster. This is commonly provided by Ambari for Hadoop Platforms, and this architecture assumes Ambari as the Platform Manager. Ambari is commonly hosted on an Edge Node.

6.1.7 Cluster Manager

The Cluster Manager accomplishes the initial provisioning and ongoing monitoring and management of the Infrastructure. This architecture specifies Genesis (for provisioning) and OpsMgr (for monitoring and management) as the Cluster Manager. The Cluster Manager is hosted on the System Management Node.

6.1.8 Operating System

6.1.8.1 Linux

Linux is an operating system instance which is installed on all Nodes within this architecture. Only Red Hat Enterprise Linux (little endian) instances are presently supported within this architecture.

6.1.9 Server

A Server is the virtual or physical element that forms the foundational layer of a Node. Only physical servers are presently recognized by this architecture.

6.1.9.1 Physical Server

A Physical Server is a physical element that provides foundational compute capability for nodes in the system. IBM POWER9™ processor-based servers are the only Physical Servers presently recognized by this architecture.

6.1.10 Management Processor

A Management Processor is the physical element that is embedded in a Physical Server that provides service access, power control, and related functions for the Physical Server. Baseboard management controllers (BMCs) are the only Management Processors presently recognized by this architecture.

6.1.11 Network Subsystem

The Network Subsystem is the collection of logical networks and the physical elements (for example, switches, cables) that host and realize them. This architecture provides much latitude with respect to the network architecture and design for the solution. There are few mandatory requirements, and many aspects of the networking apply across architecture and design and across Platform and Infrastructure. This architecture notes the mandatory requirements where applicable, and it includes a set of architectural suggestions for the networking. These suggestions should not be considered as prescribed by the architecture because any Network Subsystem design that meets the requirements can be used. However, providing some network guidance is typically helpful for the design process, and it is expected that many deployments will follow one of a few patterns.

At the Platform level, the Network Subsystem must meet the following connectivity requirements:

- All Nodes in the Cluster must have high-speed connectivity to each other. This high-speed connectivity is typically provided by the Data Network (refer to the following section).
- All Edge Nodes in the Cluster must have a connection to allow User access, Platform Admin access and connections to the External Data Sources. This connectivity is typically provided by the Campus Network (refer to the following section).

The architectural suggestions for these requirements take the form of two logical networks at the Platform level:

6.1.11.1 Data Network

The Data Network is a (typically) private network that is used to provide higher speed, higher bandwidth, and/or lower latency communication between the Worker Nodes, the Master Nodes, and the Edge Nodes. It is the primary network over which the Nodes communicate to accomplish Platform-level operations.

While the Data Network is typically private, to facilitate the transfer of large amounts of data into and out of the system, the Data Network may also be bridged directly to other external (to the system) client networks to provide a more optimal data transmission path. There are many considerations that apply if such a bridge is created, and this option is not discussed further in this architecture as routing external traffic through Edge Nodes is suggested as a preferable approach.

6.1.11.2 Campus Network

The Campus Network is the primary path for Users to access the system. Users typically connect initially to or through an Edge Node connected to the Campus Network. The Campus Network is also the default path for data to be imported and exported from the system. This data import and export also typically go through an Edge Node connected to the Campus Network.

- At the Infrastructure level, the Network Subsystem must meet the following additional requirements:
- The System Management Node must have a connection to allow Infrastructure Admin access. This requirement is typically met by the Management Network (refer to the following section).

- The System Management Node must have connectivity to a network which can be used by the Cluster Manager, hosted on the System Management Node, to accomplish provisioning of the other Nodes. The System Management Node must be able (capable and authorized) to provide Dynamic Host Configuration Protocol (DHCP) service on this network. This requirement is typically fulfilled by the Provisioning Network (refer to the following section).
- All Nodes, except for the System Management Node, must have connectivity to a network that can be used by the Cluster Manager to provision them. These Nodes must be able to obtain a DHCP address on this network. This requirement is typically fulfilled by the Provisioning Network (refer to the following section).
- The System Management Node must have an OS-level connection that allows access to the Management Processors (BMCs) of the other Nodes in the Cluster. This connectivity is required for the Cluster Manager to accomplish power control and similar operations on the other Nodes. This requirement is typically fulfilled by the Service Network (refer to the following section).
- The Switches must have a connection that allow Infrastructure Admin access to their respective management interfaces. This is typically accomplished by connecting to the Management Network (refer to the following section).

The architectural suggestions for these requirements take the form of three additional logical networks at the Infrastructure level:

6.1.11.3 Management Network

The Management Network is the commonly defined network used by administrators or other privileged persons to access the infrastructure or other elements that are not intended to be accessible to Users. The Management Network may be merged with the Campus Network in some environments (that is, the Campus Network may also be used for Management).

6.1.11.4 Provisioning Network

The Provisioning Network is a private network that is used by the Cluster Manager to accomplish the provisioning of Nodes within the system and subsequent basic monitoring of these nodes.

6.1.11.5 Service Network

The Service Network is a private network that is used to access the management processors of the servers within the system. For this architecture, only BMC-based servers are recognized, so this is the network that is used to connect to the BMCs. This is the network over which persons or other elements (for example, the Cluster Manager) access the BMCs to accomplish operations such as power control.

Physically, the system includes switches as part of the Network Subsystem.

6.1.11.6 Switches

Switches are the physical elements that realize the Networks in the Network Subsystem.

Physically, the collection of switches (and their configurations) that realize these networks is largely left as a design consideration, with the following constraints:

- The Data Network may be Ethernet or InfiniBand®.
- The other networks (Campus Network, Management Network, Provisioning Network, and Service Network) must be Ethernet.

The Data Network is typically a higher speed network that is used to support interconnection of the Worker Nodes. Because of this, the switches that support the Data Network are typically distinct and separate from those that support the other networks. The Data Network switches typically support 10 Gb or higher data rates, while the balance of the network switches typically require only 1 Gb.

6.2 Composition

6.2.1 Functions and Nodes

As a first order view, the collection of Functions at the Platform layer cooperate and interact to realize the Platform. The Platform hosts the Applications and Data, and it provides interfaces, services, and touchpoints for Users and Admins. The relationships and interactions of the various Functions is a significant and complex topic in its own right which is covered the by the HDP and Hadoop architectures. The particular selection of Functions and how they are distributed across the collection of Nodes is an important architectural and design choice for this layer of the system. However, most details in this regard are not relevant to the scope of this architecture, and the interested reader is referred to Hortonworks education materials for more information on the HDP architecture within this layer.

As a second order view, the Platform can be considered to be composed of a set of Functions that are hosted across the set of Nodes forming the Cluster. The particular selection of Functions and how they are distributed across the collection of Nodes is an important architectural and design choice for this layer of the system, and many variations in this regard are possible. However, this architecture does not prescribe any particular layout for these Platform Functions and their hosting choices. See Figure 2 for an example of this view of the Platform layer.

What is perhaps most relevant to this architecture is the nature of the Nodes (and support elements) that are provided to the Platform as part of the Infrastructure. The primary role of the Infrastructure can be viewed as providing the set of Nodes that the Platform requires. These Nodes must have suitable characteristics as required by the Platform (for example, suitable connectivity and bandwidth between the Nodes, appropriate storage capacity, sufficient resilience).

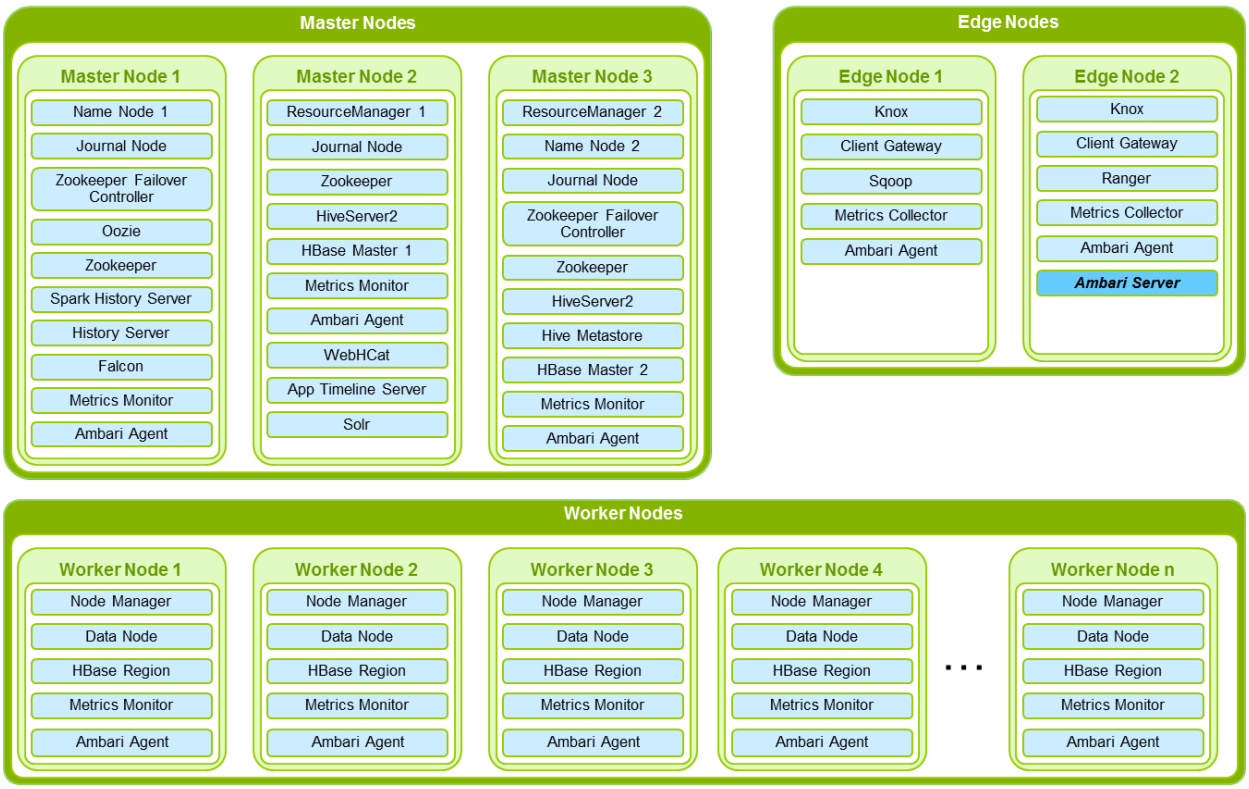

Figure 2. Example of Platform Layer View - Functions and Nodes

6.2.2 Node Composition

The Infrastructure elements above are configured and interconnected to form a Cluster. The Nodes are provisioned with an Operating System and are configured to have an appropriate network connectivity. The Switches are similarly configured to support the required network connectivity and realize the Networks described earlier.

6.2.2.1 Worker Nodes

Worker Nodes host Storage Functions that collectively cooperate and realize a critical portion of HDFS – including the physical storage for HDFS. Worker Nodes also host Applications as orchestrated by Management Functions.

Each Worker Node consists of a POWER9 processor-based server with a Linux OS. The Linux OS must be certified to operate on POWER9 processor-based servers, and it must be certified by Hortonworks as a supported OS for the particular version of HDP to be used. Currently, RHEL 7.5 for POWER9 is the OS that meets these requirements. The hardware within the Server includes processor, memory, storage, and networking components appropriate for the Worker Node role. Architecturally, these hardware components may be of any size and type which is compatible with the rest of the elements in the Infrastructure and that meet the system level requirements (for example, total HDFS storage capacity). Specific selection of the Server model and its hardware components is left as a design choice. However, the following architectural guidance is offered:

- Overall system performance and behavior is strongly influenced by the design of the Worker Nodes. Thus, the design and configuration of the Worker Nodes should be considered early in the design process, with significant attention to the particular requirements of the particular deployment.

- Worker Nodes are frequently optimized for performance of the Storage Functions and for performance when running Applications. This commonly leads to the following recommendations:
    - Higher CPU core counts and clock rates – often the maximum offered by the particular Server model chosen.
    - Larger memory sizes. 128 GB or more per Node is common.
    - As many storage drives as possible provided to HDFS. Ten or more drives per Node is common. To obtain the required storage capacity per Node, it is preferred to use more drives with smaller capacity versus fewer drives with larger capacity.
    - High performance storage controllers are preferred, but significant RAID capability is not required as the HDFS storage is typically configured as just a bunch of disks (JBOD).
    - Significant network bandwidth to the Data Network. 20 GbE per Node or better is common.
- Worker Nodes generally need not be configured for high availability characteristics. The HDP Functions and architecture and the HDFS architecture tolerate significant failures within the collection of Worker Nodes. Thus, Worker Node components can typically be chosen which optimize performance and capacity versus resilience.
- Every Worker Node is typically configured with the same hardware.

6.2.2.2 Master Nodes

Master Nodes host most of the Management Functions and some of the Storage Functions.

Each Master Node consists of a POWER9 processor-based server with a Linux OS. The Linux OS must be certified to operate on POWER9 processor-based servers, and it must be certified by Hortonworks as a supported OS. Currently, RHEL 7.5 for POWER9 is the OS which meets these requirements. The hardware within the Server includes processor, memory, storage, and networking components appropriate for the Master Node role.

Architecturally, these hardware components may be of any size and type which is compatible with the rest of the elements in the Infrastructure and which meet system level requirements. Specific selection of the Server model and its hardware components is left as a design choice. However, the following guidance is offered:

- Master Nodes should generally be configured to have good availability characteristics. The HDP Functions tolerate some of the failures within the Master Nodes, but this resilience is not complete, and the failure of a Master Node can be disruptive. So, it is recommended that the hardware choices provide good resilience where possible.
- Master Nodes typically have somewhat lower hardware demands than Worker Nodes. Master Nodes can be configured with the same hardware as the Worker Nodes if it is required to have one Node configuration in the Cluster and allow the Servers for each Node type to be interchangeable. However, processor, memory, and network configurations can be the same or somewhat less than what is configured for the Worker Nodes. Storage demands are also generally different:
    - Storage on a Master Node is typically configured in a more traditional manner. The set of drives is typically configured using RAID 5 or RAID 10, and all drives are presented to the OS. Less storage capacity is typically required than on a Worker Node, and fewer drives with larger capacity may be used.

- High performance storage controllers are preferred, and significant RAID capability is preferred.

- Every Master Node is typically configured with the same hardware. Exceptions are common, however, as a Master Node which is chosen to host some Function with larger memory or larger storage requirements may be configured differently.

6.2.2.3 Edge Nodes

Edge Nodes host Edge Functions and some Management Functions.

Each Edge Node is typically composed like a Master Node. A common exception is that an Edge Node that is intended to be used for Data import and export may be configured to have additional network adapter capacity (for example, 2x 10 GbE for internal connections to the Data Network, plus an additional 2x 10 GbE connectivity to the external network used to access the External Data Sources).

6.2.2.4 System Management Node

The System Management Node is a more modestly sized Node with lower hardware demands. To host the Cluster, the following is typically sufficient:

- 16 CPU cores and any clock rate above 2GHz
- 32 GB of memory
- 4 TB of usable storage (for example, 2x 4 TB RAID 1)
- A basic RAID controller

6.2.2.5 Machine Learning/Deep Learning Nodes

The Machine Learning or Deep Learning node can be used when performance is critical for Machine Learning, Deep Learning, or other workloads that are enabled to run on GPUs. The IBM Power System AC922 server, for example, which supports GPUs could be deployed as a Machine Learning or Deep Learning node.

6.2.3 Node Counts

In the limit, all of the Functions for a system can be hosted on a single Node serving the role of all of the Node types mentioned earlier, but such an environment is not generally useful or appropriate for any practical deployment. In this architecture, it is considered an absolute minimum requirement to have at least one Node of each primary type described earlier – that is, one each of Worker Node, Master Node, Edge Node, and System Management Node. Further, common usage modes for this environment are such that the environment will normally be a true cluster-oriented environment with multiple Worker Nodes, Master Nodes, and Edge Nodes (only one System Management Node is required, even for large Clusters).

An earlier section (section 6.1.5 “Nodes” on page 13) provides some guidance on Node counts. The following section offers some additional practical guidance for Node counts for environments that are intended for production use.

6.2.3.1 Worker Nodes

Eight (8) Worker Nodes minimum. HDFS is hosted on the Worker Nodes, and it is typically configured to have three replicas of the Data to provide resilience and performance. Thus, a minimum of three Worker Nodes are required to have even basic resilience with a common HDFS configuration. However, with only three Nodes, each Node has a complete set of the Data, and while this is helpful for performance, it does not result in performance which is representative of what would be seen as the Cluster grows. Thus, a minimum of eight Worker Nodes are recommended for any environment intended for production use.

6.2.3.2 Master Nodes

Three (3) Master Nodes minimum. Three Master Nodes are required to provide basic HA capability. As the number of Worker Nodes increase, the number of Master Nodes typically increases to provide the additional capacity to manage the larger number of Worker Nodes. The following table (Figure 3) provides some guidance from Hortonworks on appropriate Master Node counts for various Cluster sizes.

Cluster Size Type	Number of Nodes in the Cluster	Number of Master Nodes
Tiny	< 8	
Mini	8 - 16	3
Small	17 - 40	4 - 6
Medium	41 - 120	7 - 9
Large	121 – 512	10 - 12
Jumbo	> 512	consulting required

Figure 3. Suggested Master Node Counts for Various Cluster Sizes

6.2.3.3 Edge Nodes

One (1) Edge Node minimum. An Edge Node allows provides a control point for User access, and it provides dedicated capacity to handle the Data import and export. It also provides a convenient host for Ambari. It is technically possible to operate without an Edge Node, but this is not recommended for any production environment. The number of Edge Nodes typically increase with increasing Cluster size as the demands on the Edge Nodes increase similar to the demands on the Master Nodes.

6.2.3.4 Machine Learning/Deep Learning Nodes

One or more Machine Learning/Deep Learning nodes may be added to the cluster to support running Machine Learning/Deep Learning workloads or workloads that use GPU capabilities for acceleration.

6.2.4 Cluster Types

For the purposes of characterizing some of the primary variations and usage modes for the system, the following Cluster Types are defined.

6.2.4.1 Balanced

A "Balanced" cluster is cluster for which the design choices reflect a general balance between the primary characteristics of the system – especially performance, capacity, and price.

6.2.4.2 Performance

A "Performance" cluster is a cluster for which the design choices reflect more preference for increased Application performance (especially, versus capacity or price).

6.2.4.3 Storage Dense

A "Storage Dense" cluster is cluster for which the design choices reflect more preference for increased storage capacity (esp. vs. performance).

6.2.5 Network Patterns

At the Platform layer, it is useful to introduce some primary, suggested design patterns for the Network Subsystem. All of these are acceptable network patterns within this architecture, and the appropriate choice is a function of the particular requirements and desire characteristics for a particular design. A full discussion of the advantages and disadvantages of each model and the relevant trade-offs is beyond the scope of this reference architecture.

6.2.5.1 Partial-Homed Network (Thin DMZ)

This is a common pattern that provides a good balance between separating and segregating the various network traffic and providing convenient access and communication paths.

In the Partial-Homed Network model, the Data Network is private, and all Nodes are connected to it. The Campus Network is the "public" network over which Users access the Edge Nodes. Only the Edge Nodes are connected to the Campus Network, and all User and Admin access and transfers of Data to and from External Data Sources are directed through the Edge Nodes.

Advantages of this pattern:

- Cluster traffic is well isolated from other network traffic
- Master and Worker Nodes are more securely positioned

Disadvantages of this pattern:

- No direct access is provided to the Master or Worker Nodes
- Access to UIs not on the Edge Nodes must be configured through the Edge Nodes (for example, using Knox)

One variation on this pattern allows the Master Nodes to also be connected to the Campus Network, allowing direct access by Users and Admins. The implications of this variation are fairly obvious and not discussed further here.

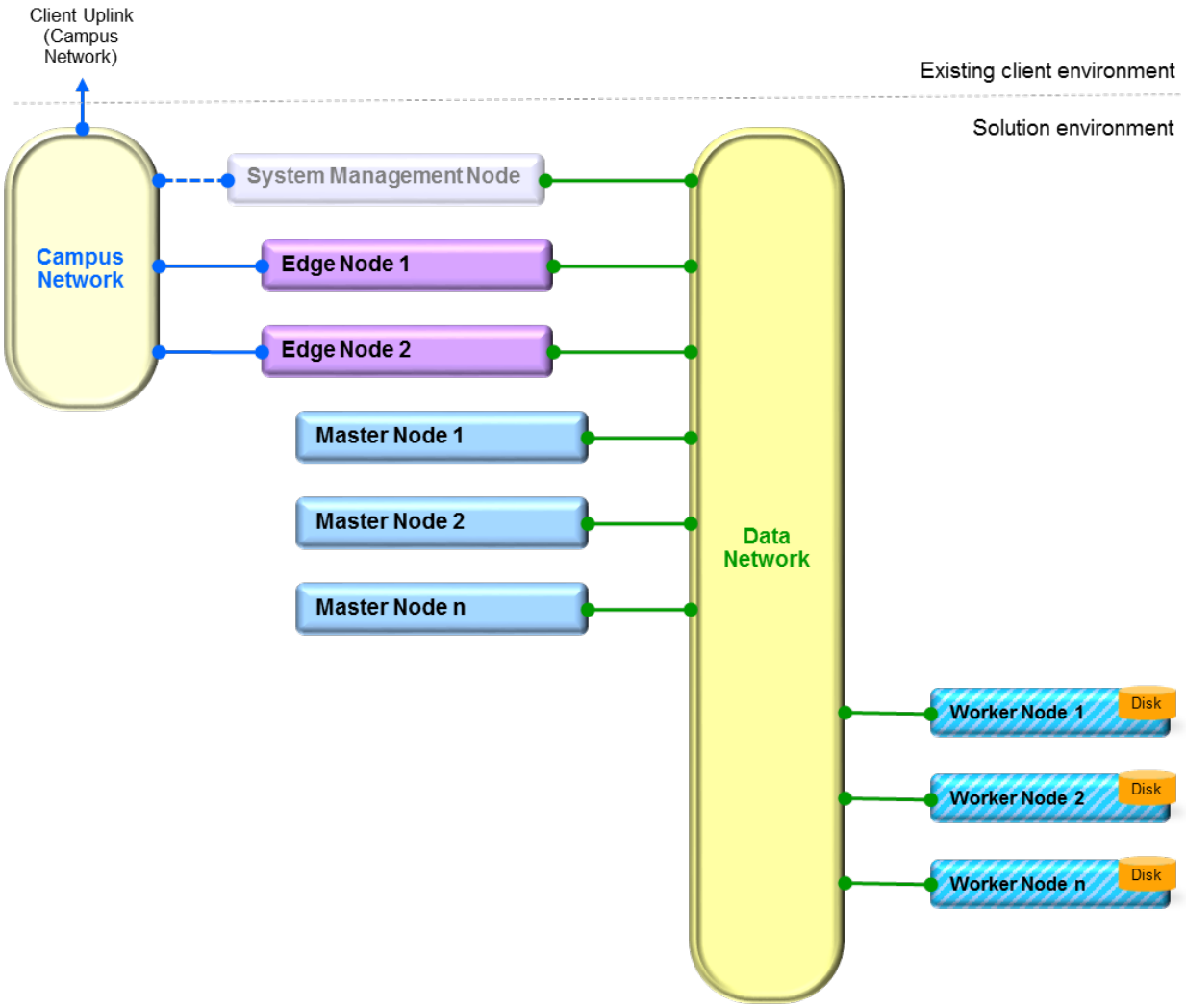

Figure 4. Partial-Homed Network

6.2.5.2 Dual-Homed Network

This pattern can be considered as an extension of the Partial-Homed Network pattern.

In the Dual-Homed Network model, the Data Network is private and all Nodes are connected to it. The Campus Network is the "public" network and all Nodes are also connected to it. This provides more connection options as User may be directed through Edge Nodes, but they may also be directed directly to interfaces on the Master Nodes. Admins can access all Nodes directly from the campus Network. Data transfers still typically go through Functions hosted on the Edge Nodes, but this is not as controlled as other paths into the Cluster.

Advantages of this pattern:

- Cluster traffic is well isolated from other network traffic
- External access to all Nodes is possible

Disadvantages of this pattern:

- Configuration is more complex

- Name and resolution must be well-considered to ensure that traffic is routed as desired
- Master and Worker Nodes are less securely positioned

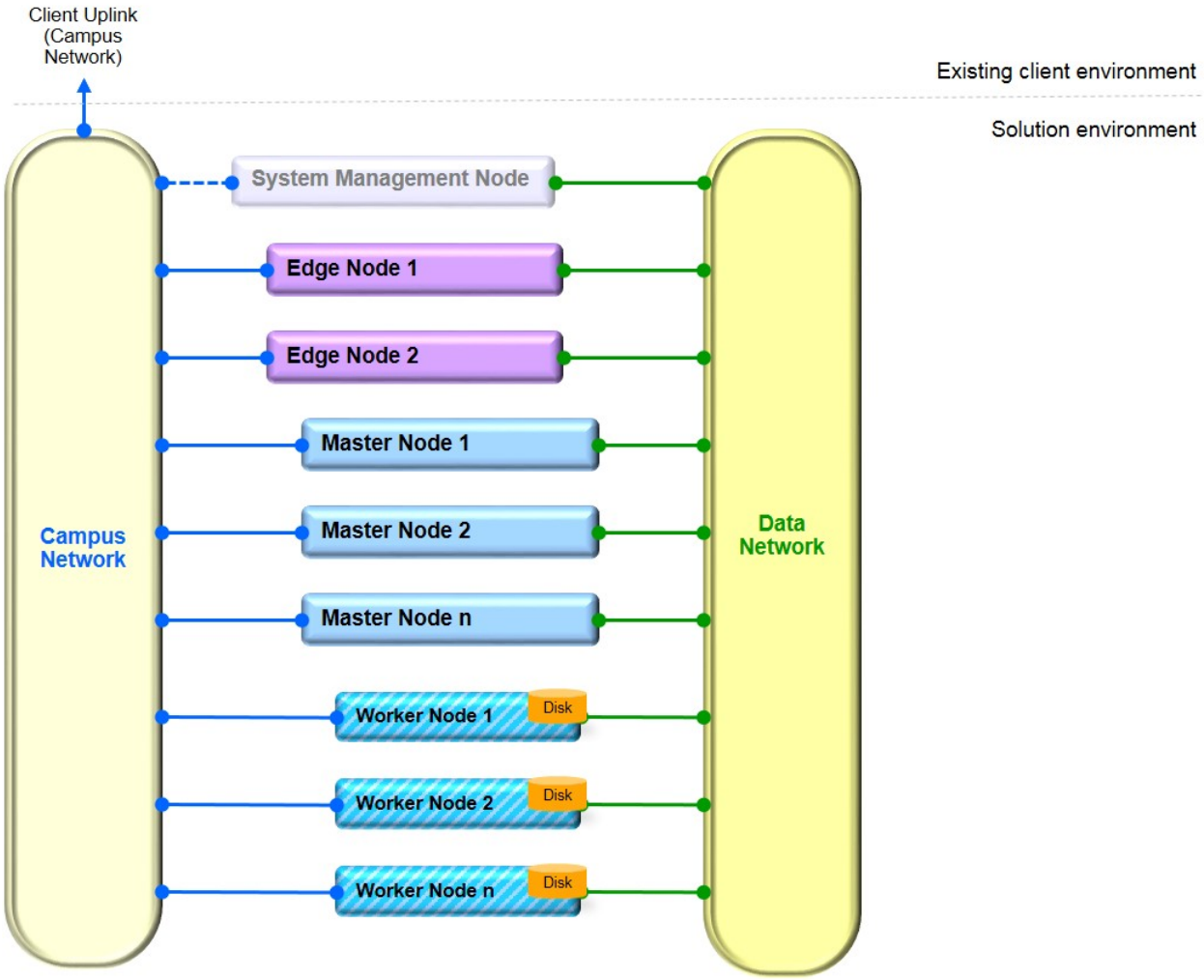

Figure 5. Dual-Homed Network

6.2.5.3 Flat Network

In the Flat Network model, the Data Network and the Campus Network are combined. All Nodes are connected to this combined network. This model provides a much simpler topology than the other models as all traffic is directed over a single network. This provides the same connection options as with the Dual-Homed Network model. The primary difference is that there is no dedicated Data Network over which the intra-Cluster traffic flows.

Advantages of this pattern:

- Configuration is simpler
- External access to all Nodes is possible

Disadvantages of this pattern:

- Cluster traffic is not isolated from other network traffic
- Master and Worker Nodes are less securely positioned

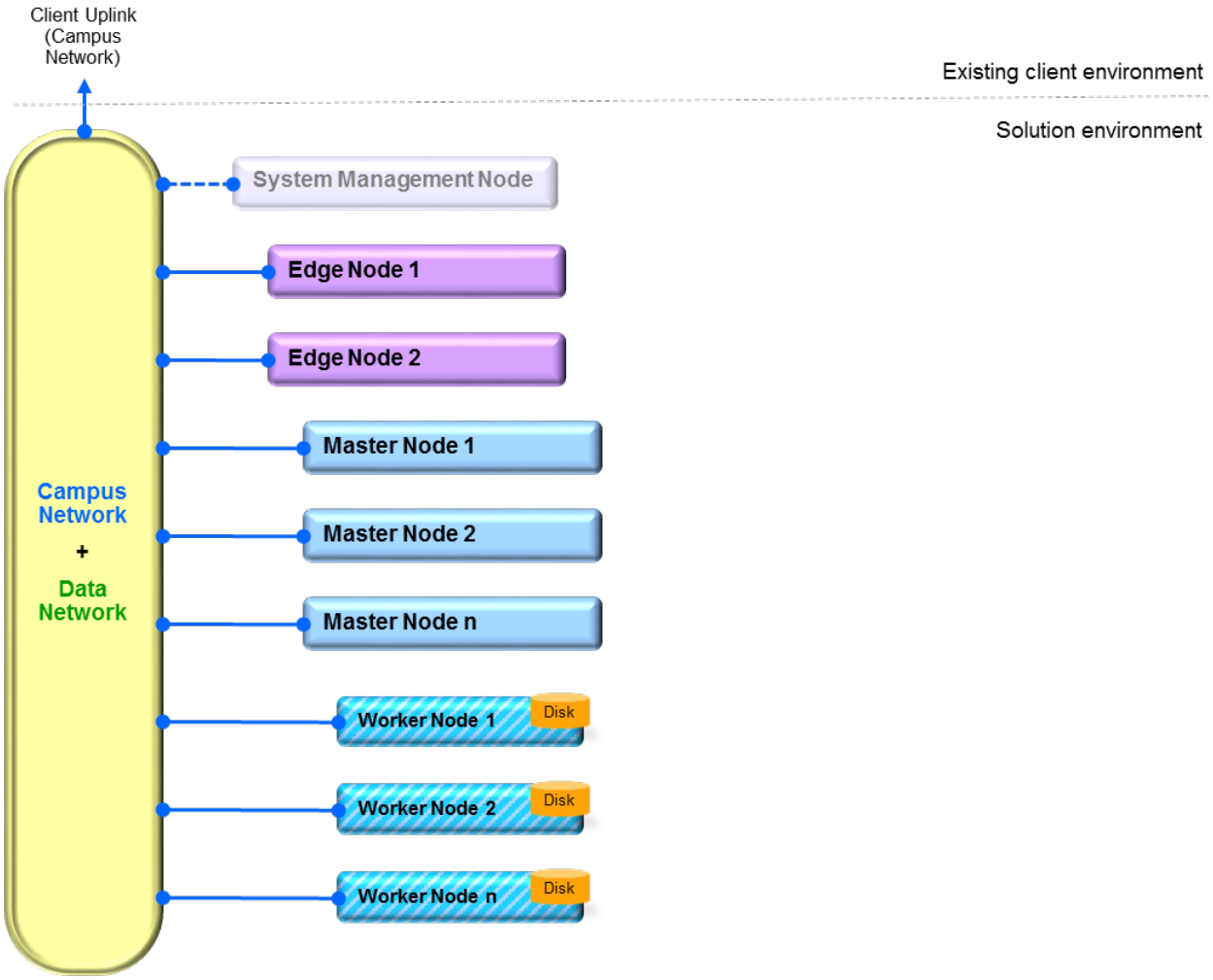

Figure 6. Flat Network

6.2.5.4 DMZ Network

In the DMZ Network model, the network is constructed similar to a Flat Network, but the Nodes are grouped and firewalls are inserted into the network to control access to and traffic between the groups. Similar to the Flat Network model, the Data Network and the Campus Network are combined, and all Nodes are connected to this combined network. The firewalls are configured to allow only the required access and traffic at each control point.

Advantages of this pattern:

- Base network configuration is simpler (similar to the Flat Network)
- Services access can be selectively controlled and at concentrated points
- Master and Worker Nodes are more securely positioned

Disadvantages of this pattern:

- Firewalls and firewall configuration are required (for example, punching holes for ports)
- Cluster traffic is not isolated from other network traffic
- Firewalls can be a bottleneck and must be sized and managed properly

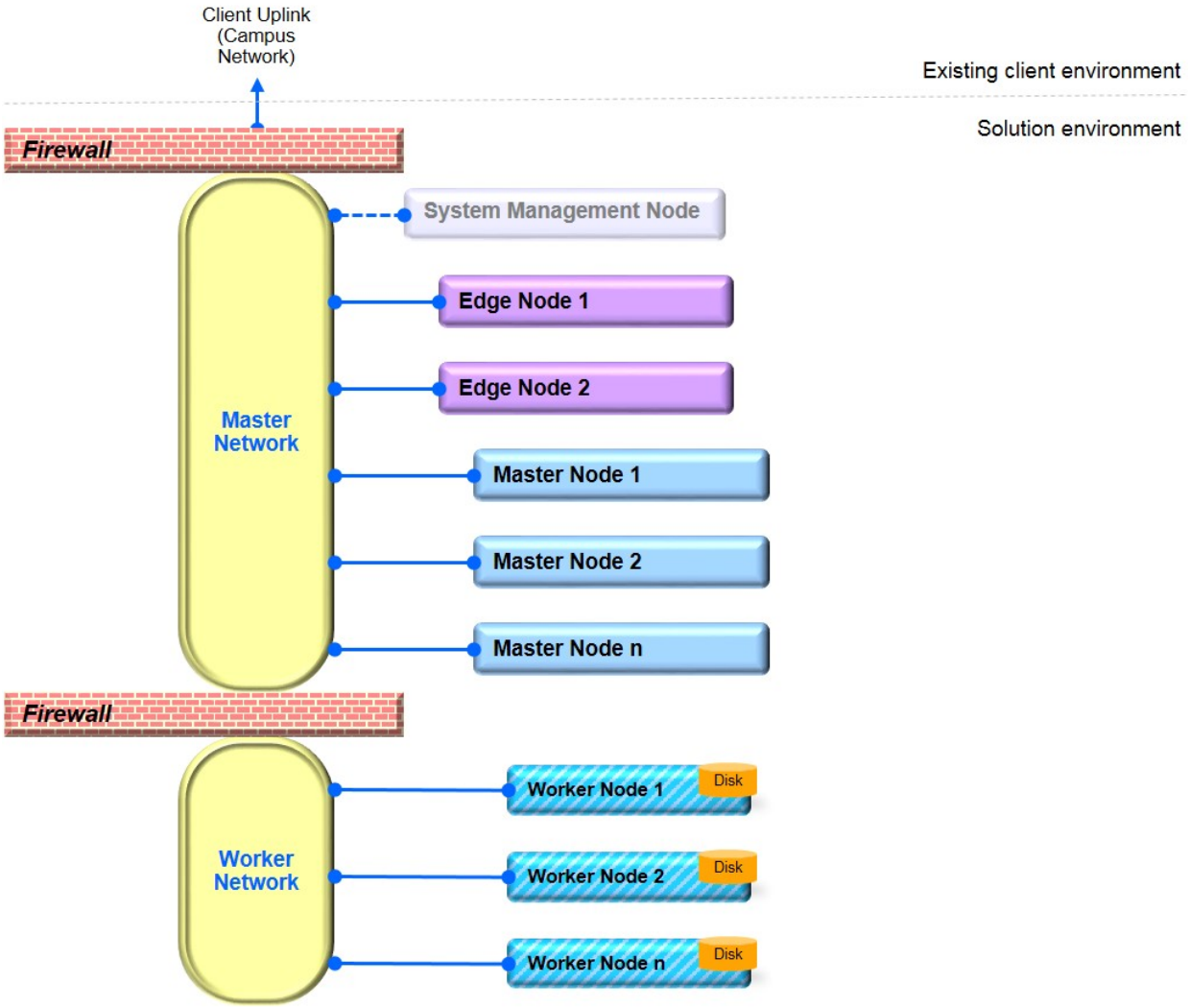

Figure 7. DMZ Network

6.2.6 Network - Other

The remaining points related to the Network Subsystem and the connections of each Node type to the various logical networks are essentially design choices and a function of the particular Platform level network pattern which is chosen (see section 6.2.5 “Network Patterns” beginning on page 22).

6.3 Summary Views

Given the above information, some summary views of the system can be presented.

A basic User/Application-level view of the system is depicted in Figure 8.

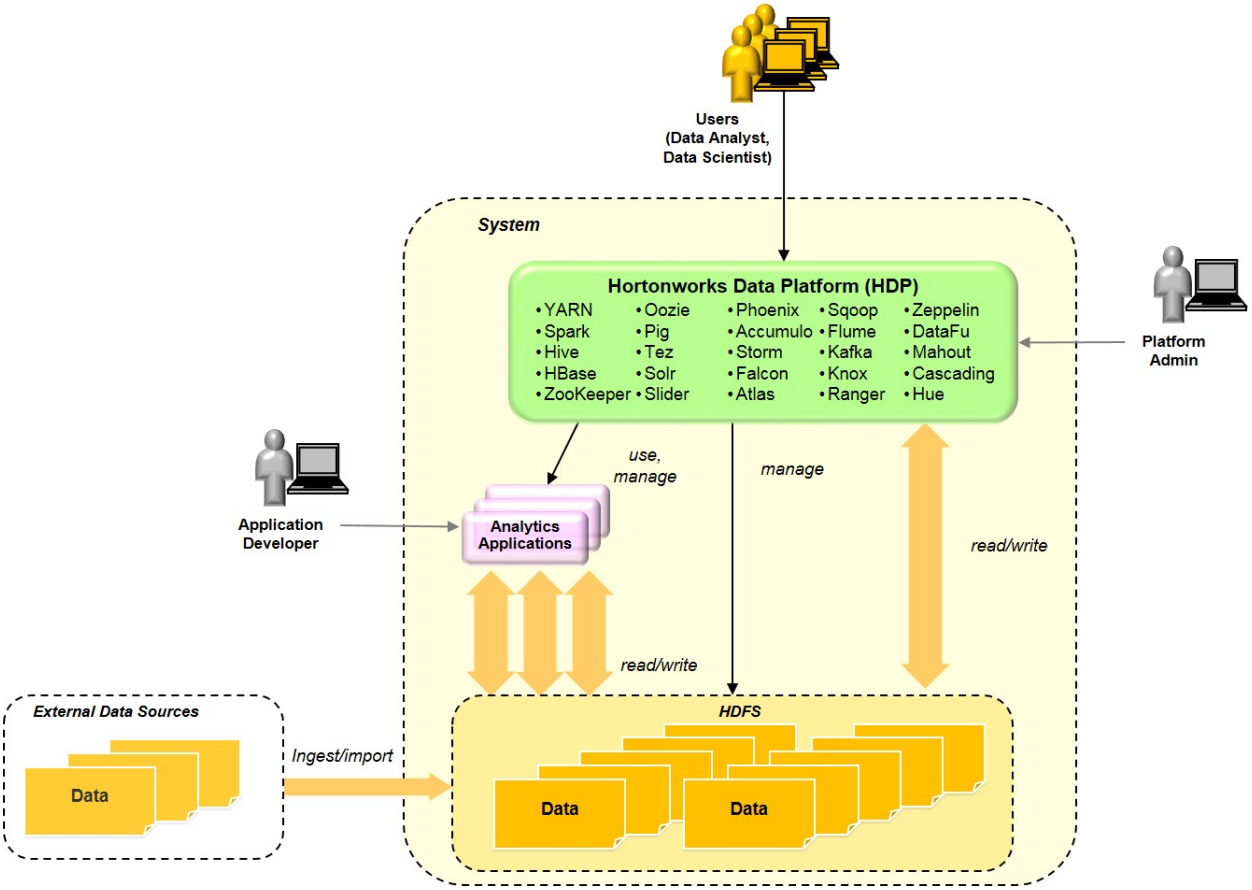

Figure 8. User/Application Level View

Figure 9 and Figure 10 provide two Infrastructure/Hosting views of the system.

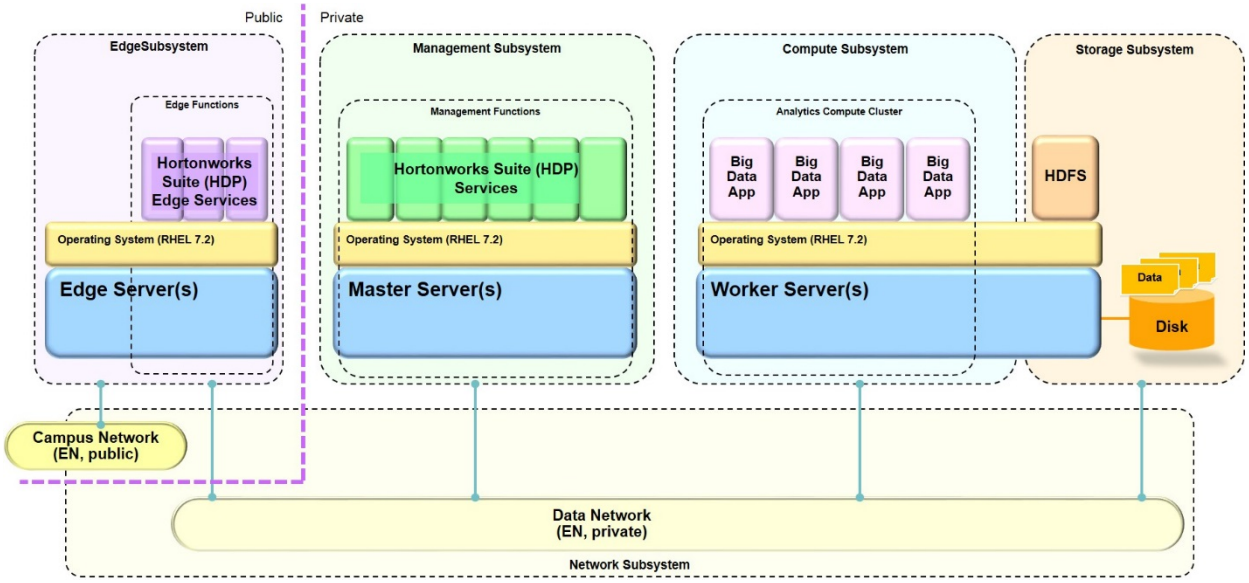

Figure 9. Infrastructure/Hosting View - Simplified

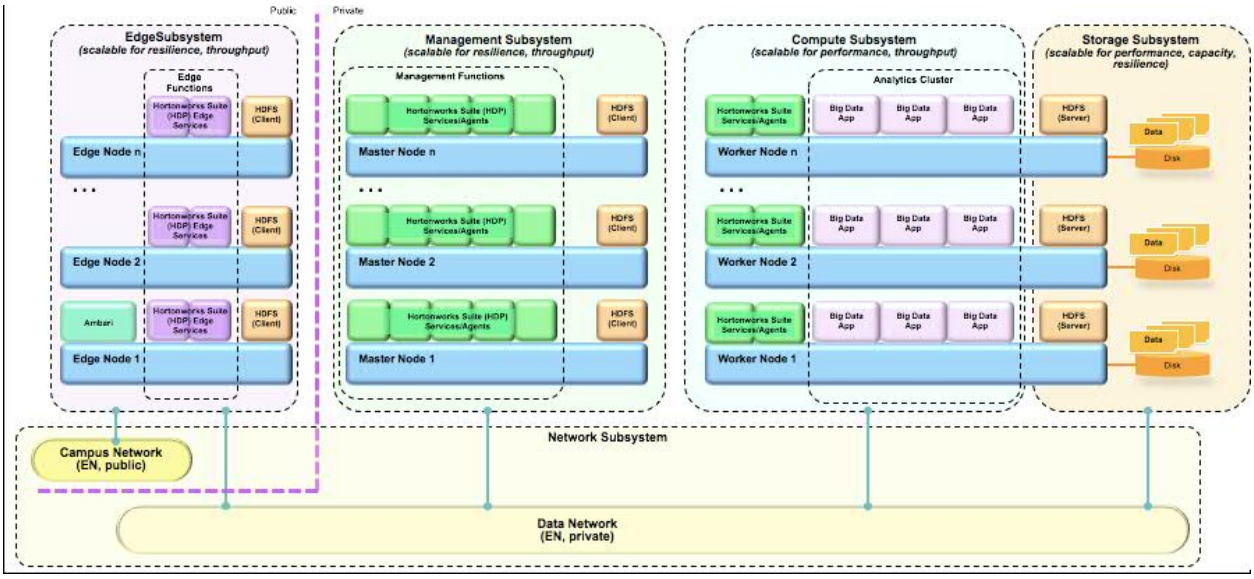

Figure 10. Infrastructure/Hosting View - More Detail

7 Reference Design 1.1A - Minimum Production Configuration

This section describes a reference design for this solution. It is an example of a system design that complies with the architecture explained in the earlier section. This reference design is considered to be a "minimum production" configuration as it is designed and sized with a minimum set of elements that would be generally appropriate for consideration as a minimum starting point for a production deployment.

This reference design is intended as a reference only. Any specific design, with appropriately sized components that are suitable for a specific deployment, requires additional review and sizing that is appropriate for the intended use.

7.1 Platform Software

Hortonworks Data Platform (HDP) version 3.0 is specified as the Platform software for this design.

7.2 Node Configurations

7.2.1 Hardware and OS Configurations

This design selects a "Balanced" Cluster Type. The specific Node configurations for each Node type for this design are listed in Figure 11. It lists the configuration parameters for other Cluster Types ("Performance" and "Storage Dense") which are referenced later.

	System Mgmt Node	Master Node	Edge Node	Worker Node		
Cluster Type	*All*	*All*	*All*	*Balanced*	*Performance*	*Storage Dense*
Server Model	1U LC921	1U LC921	1U LC921	2U LC922	2U LC922	2U LC922
# Servers (Min/Default/Max)	1 / 1 / 1	3 / 3 / Any	1 / 1 / Any	4 / 8 / Any	4 / 8 / Any	4 / 8 / Any
Sockets	2	2	2	2	2	2
Cores (total)	32	40	40	44	44	44
Memory	32GB	256GB	256GB	256GB	512GB	128GB
Storage - HDD (front)	2x 4TB HDD	4x 4TB HDD	4x 4TB HDD	12x 4TB HDD	8x 4TB HDD	12x 10TB HDD
Storage - SSD (front)					+ 4x 3.8TB SSD	
Storage - HDD (rear for OS)				2x 1.2TB HDD	2x 1.2TB HDD	2x 1.2TB HDD
Storage Controller	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)
Network* - 1 GbE	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)	Internal (4 ports OS)
Cables* - 1 GbE	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)
Network** - 10 GbE	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)	2x 2-port Intel (4 ports)	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)
Cables** - 10 GbE	2 cables (DACs)	2 cables (DACs)	4 cables (DACs)	2 cables (DACs)	2 cables (DACs)	2 cables (DACs)
Operating System	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks. See Section 7.4.1 for details.
** The 10Gbe network infrastructure hosts the data network.

Figure 11. Node Hardware and OS Configurations

7.2.2 Node Counts

7.2.2.1 Worker Nodes

Eight (8) Worker Nodes are specified for this design. This count yields a reasonable Cluster-level performance behavior and a sufficient number of Nodes for HDFS to reasonably distribute and replicate the Data. The minimum required is four Worker Nodes.

7.2.2.2 Master Nodes

Three (3) Master Nodes are specified for this design. Three Master Nodes allows the HDP Functions to be distributed such that a basic HA configuration for the Management Functions exists for the system.

7.2.2.3 Edge Nodes

One (1) Edge Node is specified for this design. One Edge Node is the minimum requirement for this architecture, and this selection represents no HA configuration for Edge Functions.

7.2.2.4 System Management Nodes

One (1) System Management Node is specified for this design.

7.3 Functions-to-Nodes Mapping

Figure 12 specifies a services layout selected for the mapping of Functions to Nodes for this design. This layout is not strictly part of the design as previous noted in section 5.1.3 on page 10, and it is included here as an example. Any appropriate layout of the HDP services across the Nodes is permitted, and such a Platform level design is a relevant and necessary part of realizing a complete and operational environment.

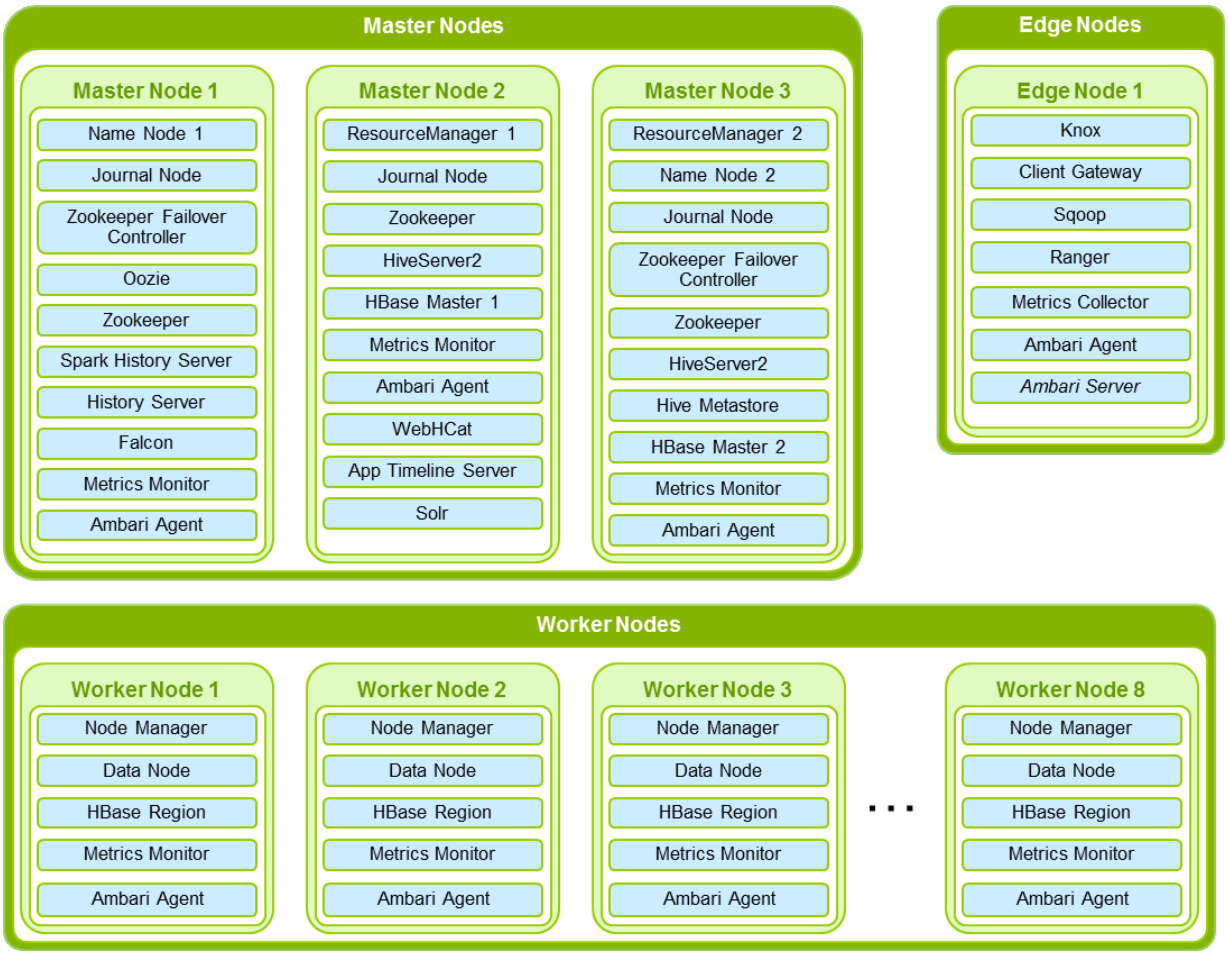

Figure 12. Functions-to-Nodes Mapping

7.4 Network Subsystem

This system design includes the following network design. At the Platform level, the "Partial-Homed" network pattern is specified for this design, and three Infrastructure level logical networks are included in the network topology as distinct and separate networks

The networks are realized by two pairs of Ethernet switches. One pair is 10 Gb, and it realizes the Data Network. The other pair is 1 Gb, and it realizes the balance of the networks. The switch pairs exist to provide resilience to the network and additional bandwidth for the Data Network.

The following choices apply to the network design. The specific virtual LAN (VLAN) numbers are arbitrary except for the VLAN 1 selection -- representing a common case where the data center management network is a simple 'flat' network carried on the default VLAN (1) of existing client switches.

Note: In the network diagrams in the following sections, EN means Ethernet.

7.4.1 Logical Mappings

7.4.1.1 Data Network

The Data Network is private (within this system) and is assigned to VLAN 77. The servers in the system present untagged traffic to the switches for this network. This network is realized by the 10 Gb switch pair.

7.4.1.2 Campus Network

The Campus Network is shared (outside of this system) and is assigned to VLAN 22. This network is realized by the 1 Gb switch pair and uplinked into the existing client network infrastructure. The servers in the system present *tagged* traffic to the switches for this network.

7.4.1.3 Management Network

The Management Network is shared (outside of this system) and is assigned to VLAN 1. This network is realized by the 1 Gb switch pair and uplinked into the existing client network infrastructure. The servers in the system present *tagged* traffic to the switches for this network. This network is also used to carry the management traffic for the other (10 Gb) switches.

7.4.1.4 Provisioning Network

The Provisioning Network is private (within this system) and is assigned to VLAN 88. This network is realized by the 1 Gb switch pair. The servers in the system present *untagged* traffic to the switches for this network. This is done to more conveniently support NetBoot, which is used to provision the nodes in the Cluster.

7.4.1.5 Service Network

The Service Network is private (within this system) and is assigned to VLAN 110. This network is realized by the 1 Gb switch pair. The BMCs in the system present untagged traffic to the switches for this network. The BMC-to-switch connections are dedicated to this function. The System Management Node also has an OS level connection to this network to accomplish power control of the nodes during provisioning.

Figure 13 and Figure 14 depict the logical network topology for this reference design. Figure 14 excludes the Provisioning and Service Network from the diagram -- allowing a simpler and cleaner rendering that better illustrates the primary connectivity for the Nodes.

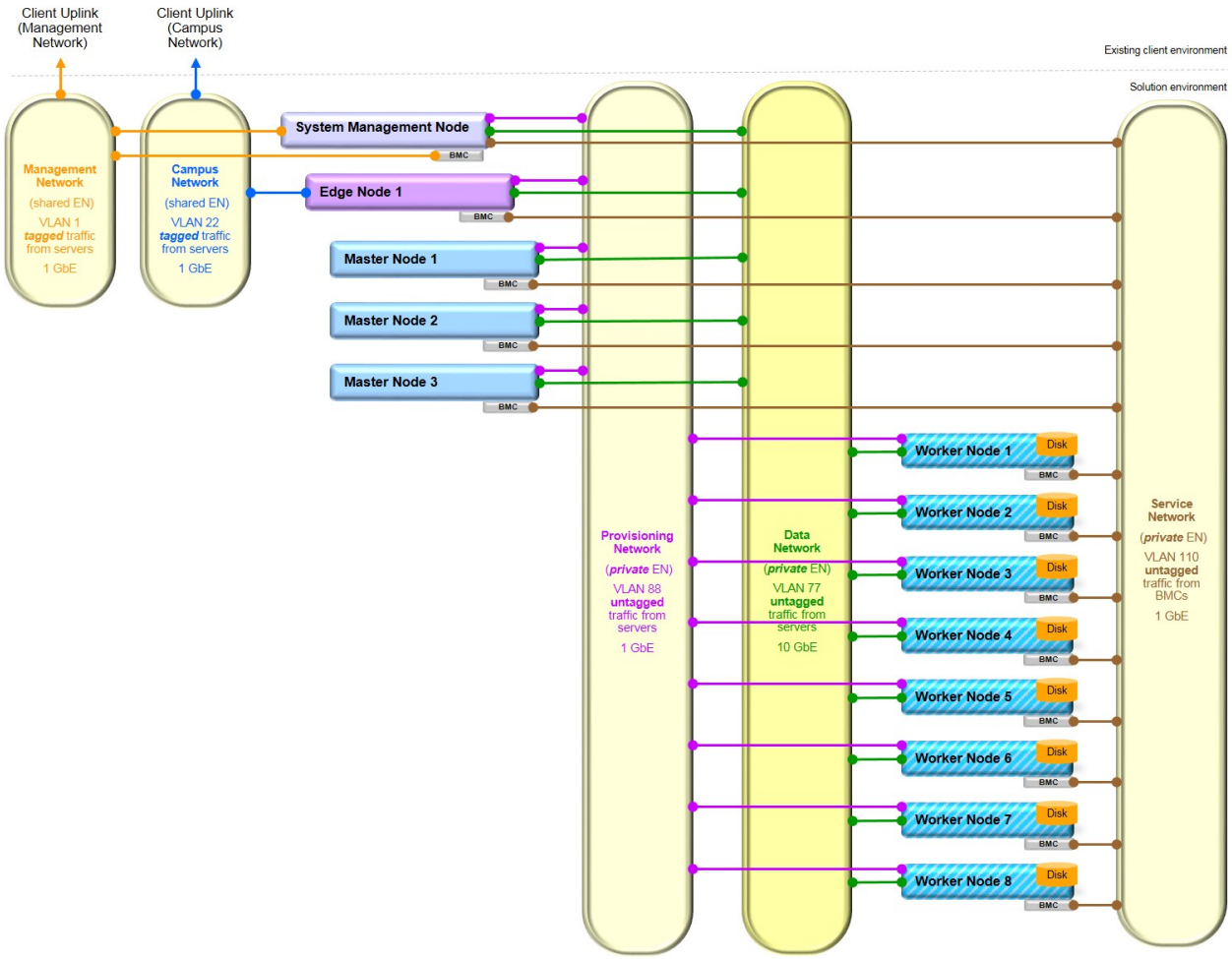

Figure 13. Network Design - Logical View - All Networks

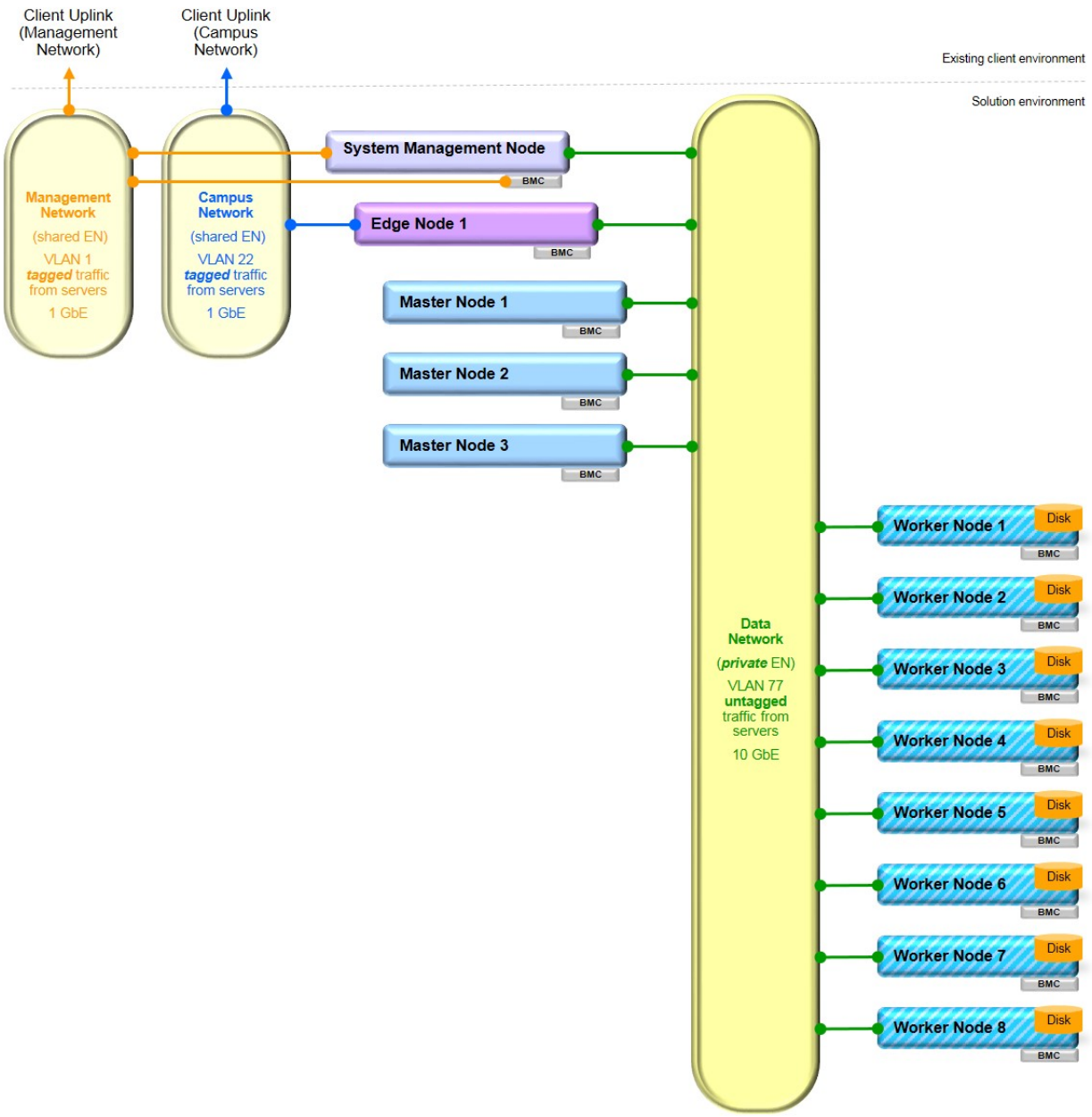

Figure 14. Network Design - Data, Campus, and Management Networks

7.4.2 Cabling

The physical cabling for each server in the system is identical. Likewise, the switch configurations for the ports for each node are identical. This provides consistency and reduces the opportunity for error. It also provides flexibility for special situations that might arise. Using this consistent physical cabling, each server is configured (within its OS) to connect to the appropriate network in a manner that is consistent with the logical view in the previous section.

7.4.2.1 1 Gb Networks

The connection between each server and the switches for the Campus Network, Management Network, Provisioning Network, and the OS connection for the Service Network (System Management Node only) is carried over two physical links (cables) to the 1 Gb switches. This provides a redundant path that is used to provide resilience for these networks. The logical networks that are listed earlier are trunked over this pair of links -- minimizing the need for dedicated links for these networks. This pair of links is configured for link aggregation using Link Aggregation Control Protocol (LACP) on the server and on the switch. Further, the 1 Gb switch pair is configured (and cabled with MLAG (multi-chassis link aggregation), which allows the links to be aggregated across the pair of switches. On the server side, the two network interfaces for these two cabled ports are bonded together (LACP mode). IP address configuration is applied to the bond interface for the native VLAN (88), and the VLAN-based interfaces with IP addresses are added for 1 Gb traffic that requires tagging (VLANs 1, 22, and 110).

The 1 Gb switches also host the Service Network. The Service Network is different than the other 1 Gb networks in that each server has a single dedicated link between its BMC interface and one of the switches. The BMC interfaces are connected to just one of the switches (Management Switch A). The System Management Node also requires an OS level connection to the Service to accomplish power operations to the other servers in the system that it can provision.

See Figure 15 for a diagram of the cabling design for the 1 Gb Networks.

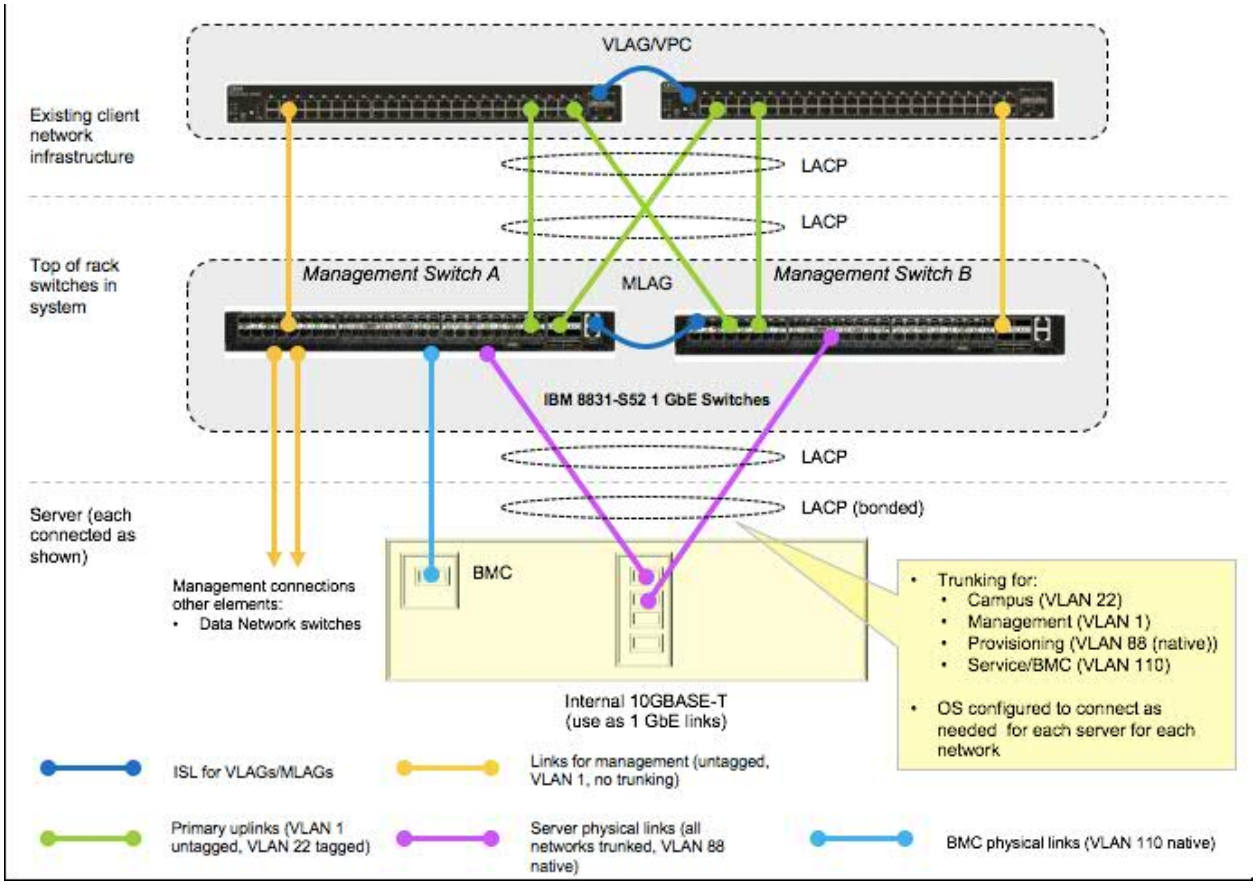

Figure 15. 1 Gb Network Cabling - Physical Schematic View

7.4.2.2 10 Gb Network

The connection between each server and the switches for the Data Network is carried over two physical links (cables) to the 10 Gb switches. This provides a redundant path that is used to provide resilience for these networks, as well as increased bandwidth (up to 20 Gb) between the Node (especially, Worker Nodes). With only a single logical network, no trunking or tagging is required, and the switch ports are simply configured to place the traffic from the servers on VLAN 77 as the native VLAN. Similar to the 1 Gb links, this pair of links is configured for link aggregation using LACP on the server and on the switch. The 10 Gb switch pair is similarly configured (and cabled with an IPL (Inter Peek Link)) for MLAG, which allows the links to be aggregated across the pair of switches. See Figure 16 for a diagram of the cabling design to the 10 Gb switches.

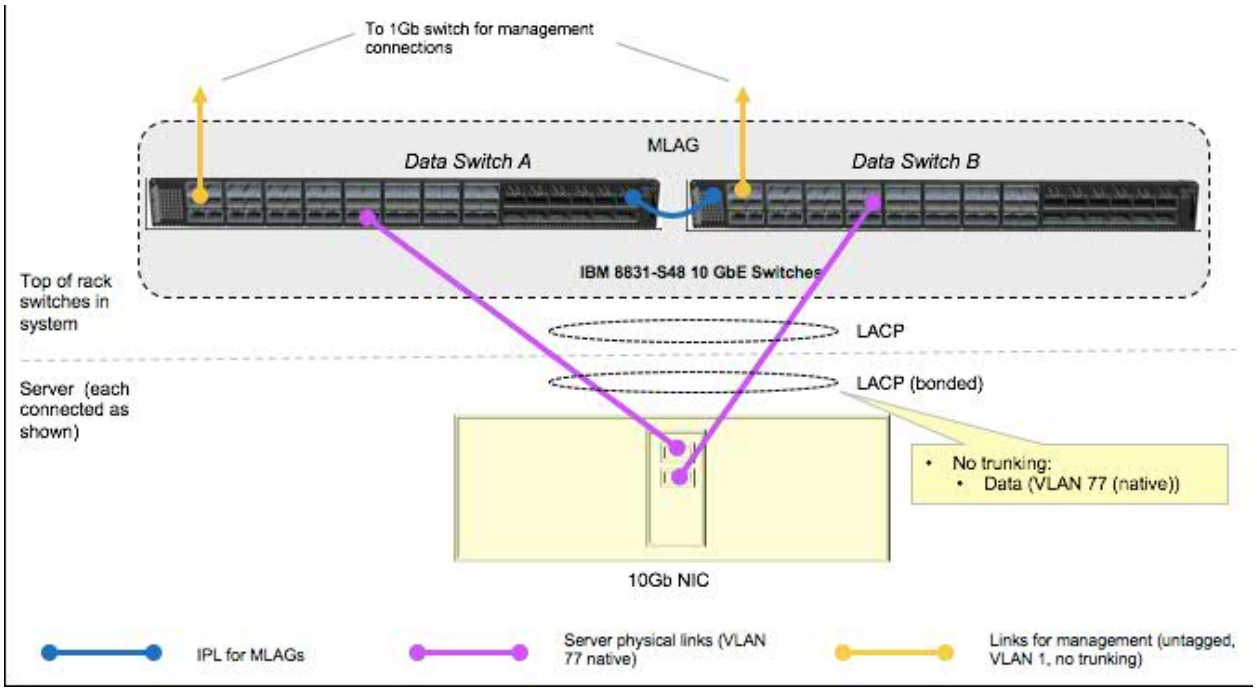

Figure 16. 10 Gb Cabling - Physical Schematic View

7.4.3 Other Considerations

7.4.3.1 NetBoot

The Provisioning Network is used to accomplish NetBoot for some provisioning operations. This creates some additional considerations that must be handled. Specifically, the driver that is used during the NetBoot process on the target node typically does not support LACP. As a result, the switches which realize the Provisioning Network must be configured to accommodate this fact. Recent switch firmware (for example, IBM Networking OS 7.9 and later) allows the ports in an LACP group to be configured to tolerate the case in which a server does not support LACP, as often occurs during NetBoot (reference the "lacp suspend-individual" option in the applicable IBM Networking OS command reference).

7.4.3.2 Dynamic Host Configuration Protocol (DHCP)

This design provides DHCP for two of the networks in the system. The Cluster Manager is configured to provide DHCP for the Service Network and the Provisioning Network.

7.5 Physical Configuration - Rack Layout

Figure 17 shows the physical layout of the system within a rack. All of the components for this reference design fit within a single 42U rack with space for growth and expansion.

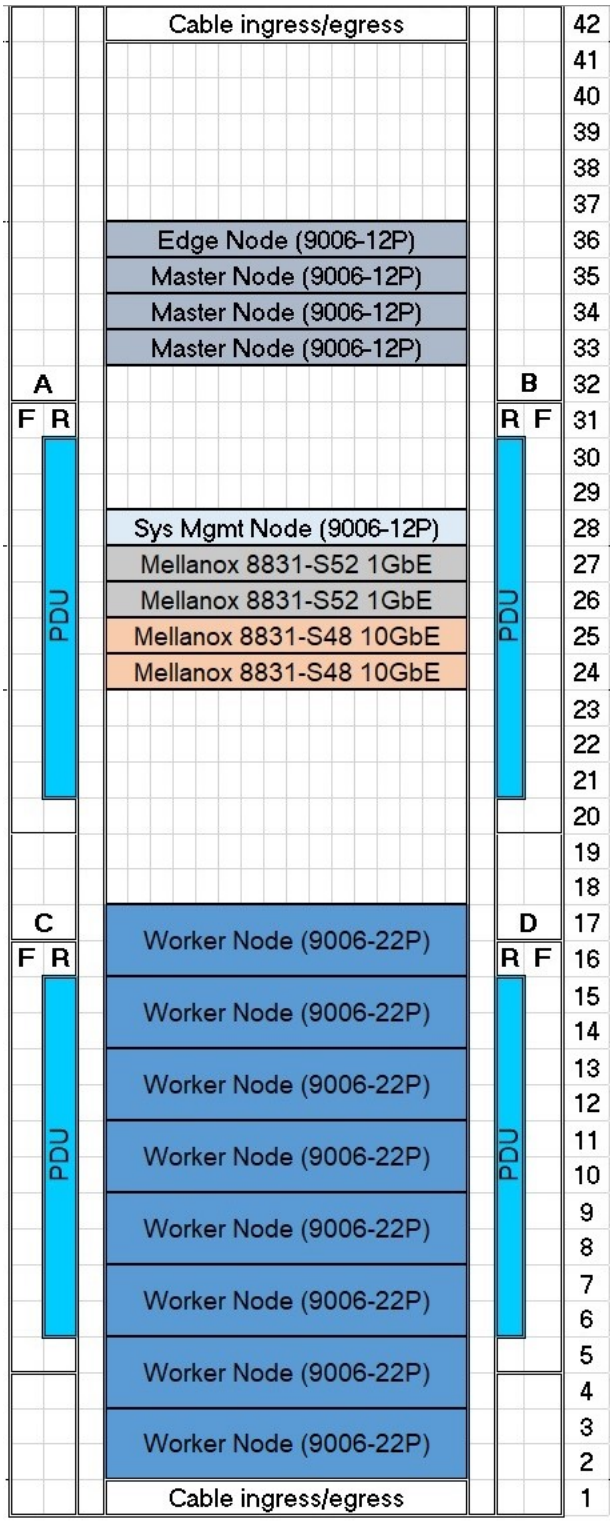

Figure 17. Physical Rack Layout

7.6 Hardware Features for e-config

Following are the e-config hardware features for this design. A full e-configuration will include the system software (RHEL 7.5 for POWER9) and services. The HDP software is supplied by Hortonworks. When configuring the Nodes, adjust the quantity of Nodes as needed. Also, adjust the size and types of drives or other features to meet your specific requirements. In the following figures, ** means that the Feature Code will be determined through based on the processor selection.

9006	22P		HDP Worker Node	8
		EHDT	HDP on Power Solution	1
		EHDX	HDP Worker Node (HDP on Power Solution)	1
		2147	Primary OS - Linux	1
		4650	Rack Indicator- Not Factory Integrated	1
		9442	New Red Hat License Core Counter	**
		9446	3rd Party Linux License Core Counter	**
		EC16	Baremetal	1
	Adapter	EKA2	PCIe3 2-port 10GbE SFP+ Adapter, based on Intel XL710	1
	Backplane	EKBG	2-Socket 2U 12 LFF/SFF 4 NVMe Direct Attach Fab Assembly	1
		EKC1	3m 10Gb SFP+, Copper Passive	2
	Drives	EKD2	4 TB 3.5-inch SAS 12Gb/s HDD NONSED WrtCache	12
		EKDQ	1.2 TB 2.5-inch SAS 12Gb/s HDD NONSED WrtCache	2
		EKLM	1.8m (6-ft) Power Cord, 200-240V/10A, C13	2
	Memory	EKMG	32GB DDR4 Memory	8
	Processor	EKPE	22-core 2.6 GHz POWER9 Processor	2
		ERBZ	Single Packaging Request Indicator	1
		ESC5	Shipping and Handling	1

Figure 18. Worker Nodes- Balanced (Quantity 8) - Hardware Features

9006	12P		HDP Master Node	3
		EHDT	HDP on Power Solution	1
		EHDW	HDP Master Node (HDP on Power Solution)	1
		2147	Primary OS - Linux	1
		4650	Rack Indicator- Not Factory Integrated	1
		9442	New Red Hat License Core Counter	**
		9446	3rd Party Linux License Core Counter	**
		EC16	Baremetal	1
	Adapters	EKA2	PCIe3 2-port 10GbE SFP+ Adapter, based on Intel XL710	1
	Backplane	EKBC	2-Socket 1U LFF NVMe Fab Assembly Base	1
		EKC1	3m 10Gb SFP+, Copper Passive	2
	Drive	EKD2	4 TB 3.5-inch SAS 12Gb/s HDD NONSED WrtCache	4
		EKLM	1.8m (6-ft) Power Cord, 200-240V/10A, C13	2
	Memory	EKMG	32GB DDR4 Memory	8
	Processor	EKP7	20-core 2.13 GHz POWER9 Processor	2
		ERBZ	Single Packaging Request Indicator	1
		ESC5	Shipping and Handling	1

Figure 19. Master Nodes (Quantity 3) - Hardware Features

9006	12P		HDP Edge Node	1
		EHDT	HDP on Power Solution	1
		EHDV	HDP Edge Node (HDP on Power Solution)	1
		2147	Primary OS - Linux	1
		4650	Rack Indicator- Not Factory Integrated	1
		9442	New Red Hat License Core Counter	**
		9446	3rd Party Linux License Core Counter	**
		EC16	Baremetal	1
	Adapters	EKA2	PCIe3 2-port 10GbE SFP+ Adapter, based on Intel XL710	2
	Backplane	EKBC	2-Socket 1U LFF NVMe Fab Assembly Base	1
		EKC1	3m 10Gb SFP+, Copper Passive	4
	Drive	EKD2	4 TB 3.5-inch SAS 12Gb/s HDD NONSED WrtCache	4
		EKLM	1.8m (6-ft) Power Cord, 200-240V/10A, C13	2
	Memory	EKMG	32GB DDR4 Memory	8
	Processor	EKP7	20-core 2.13 GHz POWER9 Processor	2
		ERBZ	Single Packaging Request Indicator	1
		ESC5	Shipping and Handling	1

Figure 20. Edge Node (Quantity 1) - Hardware Features

9006	12P		System Management Node	1
		EHDT	HDP on Power Solution	1
		EHDU	System Management Node (HDP on Power Solution)	1
		2147	Primary OS - Linux	1
		4650	Rack Indicator- Not Factory Integrated	1
		9442	New Red Hat License Core Counter	**
		9446	3rd Party Linux License Core Counter	**
		EC16	Baremetal	1
	Adapter	EKA2	PCIe3 2-port 10GbE SFP+ Adapter, based on Intel XL710	1
	Backplane	EKBC	2-Socket 1U LFF NVMe Fab Assembly Base	1
		EKC1	3m 10Gb SFP+, Copper Passive	2
	Drive	EKD2	4 TB 3.5-inch SAS 12Gb/s HDD NONSED WrtCache	2
	Memory	EKMA	8GB DDR4 Memory	4
		EKLM	1.8m (6-ft) Power Cord, 200-240V/10A, C13	2
	Processor	EKP6	16-core 2.2 GHz POWER9 Processor	2
		ERBZ	Single Packaging Request Indicator	1
		ESC5	Shipping and Handling	1

Figure 21. System Management Node (Quantity 1) - Hardware Features

8831	S52		IBM Ethernet Switch (48x1Gb+4x10Gb) - Switch A	0
		EHDT	HDP on Power Solution	1
		1111	3m, Blue Cat5e Cable	*
		1118	3m, Yellow Cat5e Cable	*
		4650	Rack Indicator- Not Factory Integrated	1
		6458	Power Cord 4.3m (14-ft), Drawer to IBM PDU (250V/10A)	2
		ECBG	0.5m (1.6-ft), IBM Passive DAC SFP+ Cable	0
		EU36	1U AIR DUCT and Rack Mount Kit for S52	1
8831	S52		IBM Ethernet Switch (48x1Gb+4x10Gb) - Switch B	0
		EHDT	HDP on Power Solution	1
		1111	3m, Blue Cat5e Cable	*
		1118	3m, Yellow Cat5e Cable	0
		4650	Rack Indicator- Not Factory Integrated	1
		6458	Power Cord 4.3m (14-ft), Drawer to IBM PDU (250V/10A)	2
		ECBG	0.5m (1.6-ft), IBM Passive DAC SFP+ Cable	2
		EU36	1U AIR DUCT and Rack Mount Kit for S52	1
8831	S48		Networking TOR Ethernet Switch MSX1410-B2F - Switch A	1
		EHDT	HDP on Power Solution	1
		1111	3m, Blue Cat5e Cable	2
		4650	Rack Indicator- Not Factory Integrated	1
		6458	Power Cord 4.3m (14-ft), Drawer to IBM PDU (250V/10A)	2
		EB40	0.5m FDR IB/40GbE Copper QSFP	0
		EDT6	1U AIR DUCT FOR S48	1
8831	S48		Networking TOR Ethernet Switch MSX1410-B2F - Switch B	1
		EHDT	HDP on Power Solution	1
		1111	3m, Blue Cat5e Cable	2
		4650	Rack Indicator- Not Factory Integrated	1
		6458	Power Cord 4.3m (14-ft), Drawer to IBM PDU (250V/10A)	2
		EB40	0.5m FDR IB/40GbE Copper QSFP	2
		EDT6	1U AIR DUCT FOR S48	1
7965	94Y		Rack	1
		EHDT	HDP on Power Solution	1
		6654	4.3m (14-Ft) 1PH/24-30A Pwr Cord	4
		7188	Power Dist Unit-Side Mount, Universal UTG0247 Connector	4
		9002	Ship Empty Feature	1
		EC01	Rack Front Door (Black)	1
		EC02	Rack Rear Door	1
		EC03	Rack Side Cover	1
		ELC0	0.38M, WW, UTG to UTG INTERNAL JUMPER CORD (PIGTAIL)	4
		ER1B	Reserve 1U at Bottom of Rack	1
		ER1T	Reserve 1U at top of rack	1
		ERLR	Left/Right PDU Redundancy	1
		ESC0	Shipping and Handling - No Charge	1

Figure 22. Switches and Rack - Hardware Features

7.7 Design Variations

The following variations are included as part of this reference design. Each of these variations brings with it some trade-offs that may be non-obvious or difficult to quantify. If any of these variations are applied, care should be taken to ensure that the resulting behavior and characteristics of the system meet the requirements of the deployment.

7.7.1 Node Configurations

- A Cluster Type of *Performance* may be selected, and the node configuration outlined in Figure 11 on page 29 may be substituted for the Worker Nodes. This variation alters the characteristics of the system to favor better performance.
- A Cluster Type of *Storage Dense* may be selected, and the node configuration outlined in Figure 11 on page 29 may be substituted for the Worker Nodes. This variation alters the characteristics of the system to favor higher storage density.
- SATA disk drives may be used for any of the HDDs for a Node Type. This may be done for any or all of the Node types in the Cluster. However, if done, this substitution is typically most appropriate to apply to the Worker Nodes first and the Master Nodes last. This variation trades some performance and reliability, availability, and serviceability (RAS) characteristics for lower price.
- CPU for a Node type is assumed to be two sockets. It may be reduced to as low as 16 core processors by using one socket. This variation trades performance for lower price. If a single socket processor option is chosen, note that other features of the server may not be available or other capacities (for example, maximum memory) may be reduced.
- Memory for a Node type may be increased up to 512 GB. 512 GB is the maximum memory available for the Server models in this reference design. This variation may improve performance, and it typically increases price.
- Memory for a Node type may be reduced down to 128 GB. 128 GB is recommended as the minimum memory for Worker, Master, and Edge Nodes This variation typically lowers price, and it may reduce performance.
- HDD sizes may be increased up to 8 TB per drive. This variation increases the total storage capacity with a reduction in performance likely when compared to the same capacity spread over a larger number of smaller drives.
- HDD sizes may be decreased down to 2 TB per drive. This variation reduces the total storage capacity with an increase in performance likely when compared to the same capacity spread over a fewer number of smaller drives.

7.7.2 Node Counts - Increasing

Additional Worker Nodes, Master Nodes, and/or Edge Nodes may be specified.

7.7.2.1 Additional Worker Nodes

Additional Worker Nodes may be specified to increase system capacity and/or system performance. Worker Node counts up to several hundred Nodes may be added before some limits may need to be considered and additional design consulting is required. To specify additional Worker Nodes, it is largely a matter of the following factors:

- Deciding how many Worker Nodes are required (for example, see “Appendix A - Sizing” on page 45)
- Adding an appropriate number of Master Nodes and Edge Nodes to handle the increased number of Worker Nodes
- Specifying the additional physical infrastructure for the additional Nodes (for example, racks, PDUs)
- Scaling the network design appropriately for the total number of Nodes

7.7.2.2 Additional Master Nodes

Additional Master Nodes may be specified to provide additional hosting capacity or performance for Management Functions or to allow Management Functions to be distributed differently or more sparsely across the Master Nodes. For large Clusters, dedicated Master Nodes for some of the more critical Management Functions is often appropriate.

7.7.2.3 Additional Edge Nodes

Additional Edge Nodes may be specified to support more Users or to provide additional Data import or export capacity.

7.7.3 Node Counts - Decreasing

When using this reference design, it is not recommended to reduce the Node counts below the numbers listed. A system can function with fewer Nodes, but reducing the Node counts begins to introduce some distortions in the way the system operates. For example, reducing the number of Master Nodes below three does not allow the HA related services to operate normally.

7.7.4 Network Configurations

- The 1 GbE network may be hosted on a single 1 GbE switch. This variation trades some resilience for lower price.
- The 10 GbE network may be hosted on a single 10 GbE switch. This variation trades performance and resilience for lower price.
- The 10 GbE switches may be realized with Mellanox SX1410 switches (IBM 8831-S48). This substitution is strongly recommended for deployments that are greater than two racks or are expected to grow to greater than two racks. Deployments that are greater than two racks will typically require additional sets of 10GbE top-of-rack switches and a substantial switch interconnect design. These Mellanox switches have multiple (12x) 40 GbE ports that can used for ISLs and/or uplinks to the next level switches (for example, aggregation layer switches).
- If additional network bandwidth is needed to the servers for the Data Network, 25 GbE connections can be used instead of the 10 GbE connections. You can use the Mellanox 8831-25M network switches (Feature Code EKAU) to host the Data Network along with matching 25GbE-capable transceivers (Feature Code EB47) and cables (Feature Codes EB4J, EB4K, EB4L, and EB4M).

8 Reference Design 1.1B - Minimum Proof-of-Concept Configuration

This section describes a reference design for this solution for proof-of-concept (POC) testing where a smaller cluster is sufficient and more cost effective, and a redundant network is unnecessary. This reference design may be considered a *minimum POC* configuration as it is designed and sized with a minimum set of elements that would be generally appropriate as a minimum starting point for a test deployment.

This reference design is intended as a reference only and has not been specifically tested by IBM. Any specific design, with appropriately sized components that are suitable for a specific deployment, requires additional review and sizing that is appropriate for the intended use.

8.1 POC Configuration

Figure 23 shows a POC configuration for a minimally functioning environment with cost-sensitive variations described after the figure.

	System Mgmt Node	Master/Edge Node	Worker Node
Cluster Type	*All*	*All*	*POC*
Server Model	1U LC921	1U LC921	2U LC922
# Servers (Min)	1	1	3
Sockets	2	2	2
Cores (total)	32	40	44
Memory	32GB	256GB	256GB
Storage - HDD (front) **Storage - SSD (front)**	2x 4TB HDD	4x 4TB HDD	4x 4TB HDD
Storage - HDD (rear for OS)			
Storage Controller	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)	MicroSemi PM8069 (internal)
Network* - 1 GbE **Cables* - 1 GbE**	Internal (4 ports OS) 3 (2 OS + 1 BMC)	Internal (4 ports OS) 3 (2 OS + 1 BMC)	Internal (4 ports OS) 3 (2 OS + 1 BMC)
Network - 10 GbE** **Cables** - 10 GbE**	1x 2-port Intel (2 ports) 2 cables (DACs)	2x 2-port Intel (4 ports) 4 cables (DACs)	1x 2-port Intel (2 ports) 2 cables (DACs)
Operating System	RHEL 7.5 for P9	RHEL 7.5 for P9	RHEL 7.5 for P9

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks. See Section 7.4.1 for details.

** The 10Gbe network infrastructure hosts the data network.

Figure 23. POC Hardware and OS Configuration

- One server could be used to host the services for the edge and master nodes. More nodes can be added as needed.
- The systems management node is outside the HDP cluster so that it can provision the cluster.
- Three Worker Nodes are sufficient for a POC configuration. More nodes can be added as needed.
- For Worker Nodes, start with 4 HDD disks as a starting point. Select the disk size based on initial dataset size and increase the quantity up to 12 disks total based on expected growth. See the Appendix A - Sizing section for information on how to estimate the necessary storage space needed.
- A redundant network is commonly unnecessary for a POC configuration. This simplifies the network design and reduces the number of network adapters required.
- Less expensive network adapters could be used (for example, PCIe2 2-port 10 GbE BaseT RJ45 Adapter or leverage the adapter that is already built into the server). Using this adapter requires compatible network rail switches.

Appendix A - Sizing

Sizing a system for HDP is a significant and complex topic. Sizing is relevant in a number of dimensions. For example, in terms of requirements, factors like throughput, response time, ingest rate, and so on may be relevant. In terms of the system design, parameters like numbers of each Node type, processor cores total and per Node, memory total and per Node, HDFS storage capacity total and per Node, network adapter ports and bandwidth, and so on must be chosen.

A complete treatment of the design and sizing process for HDP on Power is beyond the scope of this reference architecture. *Note that some skilled consulting is typically required to properly size a Cluster for a particular client deployment.* However, some guidance for one common process – sizing for storage capacity – is provided in the following section for reference.

A.1 Data Capacity Driven Sizing

This section outlines the process for a "data capacity driven" sizing, and provides an example applying this process.

A.1.1 Process

Refer to the following process for a data capacity driven sizing.

1. Gather client requirements.
   - Usable storage capacity (UsableStorageCapacity) needed (HDFS content)
   - Usage modes/cases expected
   - Storage calculation input parameters (DataReplicationFactor, IntermediateDataUplift, CompressionRatio, Freespace; see Formula 1)
   - Addition demand on Master Nodes expected – especially database operations
   - Number of Users expected
   - Data ingest rate expected
   - Networking preferences or policies
   - Availability requirements
2. Choose the Cluster Type (for example, Balanced, Performance, Storage Dense) as a function of the usage modes/cases expected.
3. Choose the drive size and number of drives per server for Worker Nodes (typically fully populate all drive bays on Worker Nodes).
4. Calculate the amount of raw storage per Worker Node (StoragePerWorkerNode).
5. Calculate the total system raw storage capacity needed (RawStorageCapacity; see Formula 1).
6. Calculate the total number of Worker Nodes (see Formula 2).
7. Choose the number of Master Nodes as a function of the number of Worker Nodes (see Figure 3 on page 21). If necessary, increase the Master Node count to handle any additional demand expected on Master Nodes
8. Choose the number of Edge Nodes as a function of the number of Users and expected Data ingest rate.

9. Choose the logical network topology, typically based on client networking preferences or policies (see section 6.2.5 “Network Patterns” on page 22).
10. Choose the network switches and network redundancy preferred for each network class.
11. Confirm or adjust the configuration for each Node type – beginning with the reference configuration for each node type for the chosen Cluster Type (see Figure 11 on page 29)
12. Confirm or adjust the network link capacities and switch capacities as appropriate.
13. Confirm or adjust Node counts to meet availability requirements.
14. Confirm or adjust any selections based on growth expectations or initial headroom required.

```
RawStorageCapacity = ( ( UsableStorageCapacity * DataReplicationFactor + IntermediateDataUplift )
                                / CompressionRatio ) + Freespace
```

Formula 1. Raw Storage Capacity Calculation

```
NumberOfWorkerNodes = RawStorageCapacity / StoragePerWorkerNode
```

Formula 2. Number of Worker Nodes Calculation

A.1.2 Example

1. Gather client requirements.
   - UsableStorageCapacity = 150 TB
   - Usage modes/cases expected – general data analytics

   Storage calculation input parameters:
   - DataReplicationFactor = 3
   - IntermediateDataUplift = 35%
   - CompressionRatio = 2
   - Freespace = 20%
   - Addition demand on Master Nodes expected - none
   - Number of Users expected – n/a
   - Data ingest rate expected – n/a
   - Networking preferences or policies – none
   - Availability requirements – basic HA
2. Choose Cluster Type = Balanced
3. Choose 12x 4TB drives per Worker Nodes
4. Calculate StoragePerWorkerNode = 12 * 4 TB = 48 TB
5. Calculate RawStorageCapacity = (150 TB * 3 * (1 + 0.35)) / 2 * (1 + 0.20) = 365 TB
6. Calculate NumberOfWorkerNodes = 365 TB / 48 TB = 7.59 => round up to 8 worker nodes
7. Choose three Master Nodes (basic HA for Management Functions)

8. Choose two Edge Nodes (enable basic HA at edge)
9. Choose Partial-Homed (Thin DMZ) network topology
10. Choose two 10 GbE switches for data network (two Mellanox 8831-S48) and one 1 GbE switch for other networks (one Mellanox 8831-S52)

	1	2	3	4				
	System Management Node	Master Node	Utility Node	Worker Node	Server Cour (HDP)	Server Cour (Total)	Data Storag Total (Raw	Notes
				Form 1 - HDD only				
	smn	mn	un	wn/an/dn				
Server count	1	3	2	8	13	14	384 TB	
	1U LC921	1U LC921	1U LC921	2U LC922				
Cores	16	40	40	44				
Memory	64GB	256GB	256GB	256GB				
Storage - cluster - HDD				12x 4TB (48TB)				JBOD for HDFS
Storage - local - HDD	2x 4TB (4TB usable)	4x 4TB (8TB usable)	4x 4TB (8TB usable)	small partition on above				RAID 1 or 10
Storage controller	MicroSemi PM8069	MicroSemi PM8069	MicroSemi PM8069	MicroSemi PM8069				
1GbE	4 internal	4 internal	4 internal	4 internal				
10GbE	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)	1x 2-port Intel (2 ports)				
10GbE cables	2 cables	2 cables	2 cables	2 cables				

2x Mellanox 8831-S48 for Data Network
1x Mellanox 8831-S52 for Other (1GbE) Networks

Figure 24. Sizing Example - Resulting Configuration

Appendix B - Multi-Rack Considerations

Configurations that require more than one rack introduce some additional factors which must be considered. Most of these considerations are the same as those which apply to other multi-rack cluster deployments. These considerations include:

- Providing additional physical infrastructure for the additional Nodes and switches (for example, racks, PDUs, and so on)
- Scaling and designing the network appropriately for the total number of Nodes
- Distributing Master Nodes and Edge Nodes across racks to improve availability

The first item is largely a matter of choosing the number of nodes per rack, choosing where to place the switches, and configuring sufficient power for the components in the rack. The second item is beyond the scope of this reference architecture, and network design consulting is recommended for any configuration that exceeds more than a very few racks. The third is somewhat specific to an HDP deployment, and some general points are offered in this regard in the following section.

B.1 An Extensible Three-Rack Configuration

The following section outlines a three-rack configuration that can form a set of building blocks that can be extended to construct a configuration of many racks. It is based on the reference design explained earlier -- extended to add additional Nodes and capacity.

B.1.1 Rack-Level Building Blocks

The first rack (Rack A) in the configuration hosts the first set of Master and Edge Nodes for the system, some of the Worker Nodes, the 1 Gb switches which serve the first rack, and the 10 Gb switches which serve the first two racks.

The second rack (Rack B) in the configuration hosts more Worker Nodes and the 1 Gb switches which serve the second rack. It does not need any additional 10 Gb switches as the set of 10 Gb switches in Rack A are able to serve this rack.

The third rack (Rack C) in the configuration hosts more Master, Edge, and Worker Nodes, the 1 Gb switches which serve the third rack, and another set of 10 Gb switches which serve this third rack and which can serve a fourth rack if one were to be added to the configuration. It is essentially the same as the first rack (Rack A) minus the System Management Node.

B.1.2 Extending the Cluster

In this configuration Rack B and Rack C can be considered as building blocks that can be replicated and added to the configuration in an alternating manner to extend and grow the Cluster. For example, a fourth rack would be like Rack B, a fifth rack like Rack C, and so on.

As the Cluster grows, additional Master and Edge Nodes are required. The proportion of Master and Edge Nodes required in any given Cluster will vary, and the distribution of Master and Edge Nodes across the racks can also be adjusted from the pattern shown to meet the specific requirements of a deployment.

B.1.3 Network Design

As noted above, the network design for a multi-rack configuration is beyond the scope of this reference architecture. However, it is noted that it is typically appropriate and preferred to have significant interconnect bandwidth between the 10 Gb switches. The Mellanox 8831-S48 switches are indicated in this configuration as they have a significant number (12x) of 40 Gb ports which are very useful for such interconnects and uplinks. In addition, some rack space is reserved in this configuration for the inter-rack switches. The amount of rack space actually required will depend upon the specific network design.

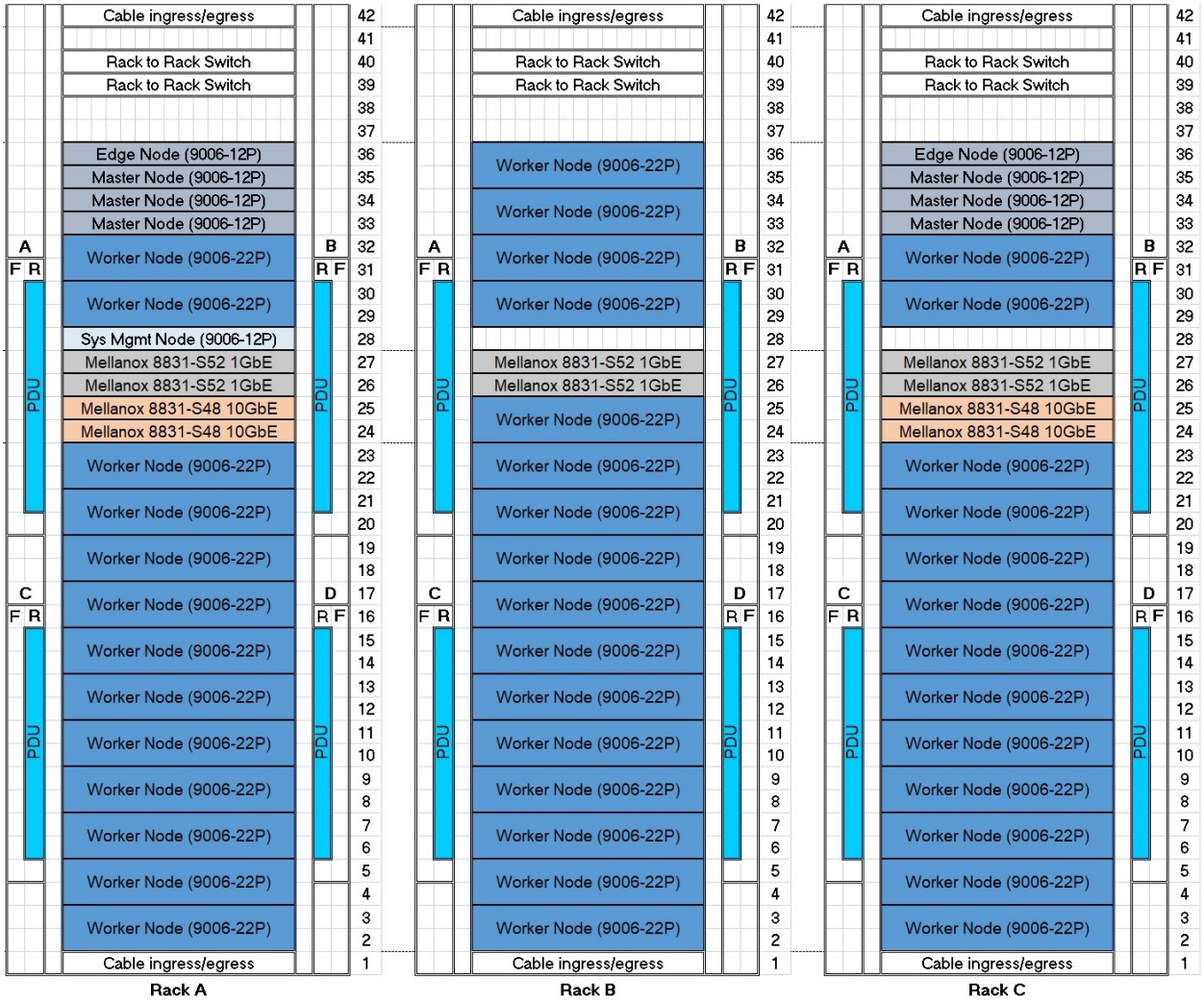

Figure 25. Three Rack Cluster and Rack-Level Building Blocks

Appendix C - Self-Encrypting Drives Considerations

Environments that need self-encrypting drives (SEDs) require a number of configuration adjustments and special considerations.

SEDs can be used in the Worker Nodes where the data that requires encryption is stored.

- A maximum of 12 SEDs can be configured on one Power LC922 server, and they must be in the front drive slots.
- SEDs are not supported in the rear slots.
- To maximize internal storage, use 12 HDDs that support SED in the front drive slots along with the configuration mentioned next.
- Using SEDs requires the LFF NVMe Fab Assembly backplane (Feature Code EKBJ) in order to have full access to all 12 drive bays from a single LSI host bus adapter (HBA).
- In addition to the backplane, using the SED feature of an SED drive requires one LSI MegaRAID HBA feature to be ordered (recommend Feature code EKEH for LC922 2U or EKAH for LC921 1U).
- In addition to the HBA adapter, you must order one of the following two Feature Codes to unlock the LSI SafeStore feature:
  - Feature Code EKWB (the software license) or
  - Feature Code EKWC (the LSI hardware key)
- The SED feature cannot be utilized with LSI HBA Feature Codes EKEB or EKAB and cannot be used without an adapter.

SEDs can also be used on the other Nodes to protect the operating system, HDFS, and so on. For Power LC921, four SEDs are supported per backplane, so more than one backplane may need to be configured.

Appendix D - Notices

This information was developed for products and services that are offered in the USA.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
United States of America

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM 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 IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

Portions of this code are derived from IBM Corp. Sample Programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at www.ibm.com/legal/copytrade.shtml.

Terms and conditions for product documentation

Permissions for the use of these publications are granted subject to the following terms and conditions.

Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

Personal use

You may reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You may not distribute, display or make derivative work of these publications, or any portion thereof, without the express consent of IBM.

Commercial use

You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not make derivative works of these publications, or reproduce, distribute or display these publications or any portion thereof outside your enterprise, without the express consent of IBM.

Rights

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any information, data, software or other intellectual property contained therein.

IBM reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by IBM, the above instructions are not being properly followed.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

IBM MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Appendix E - Trademarks

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Hortonworks and HDP are registered trademarks or trademarks of Hortonworks, Inc. and its subsidiaries in the United States and other jurisdictions.

Apache Hadoop, Hadoop, and Apache are either registered trademarks or trademarks of the Apache Software Foundation in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Red Hat is a registered trademark of Red Hat, Inc.

InfiniBand is a trademark and service mark of the InfiniBand Trade Association

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