Cloudera HDP on IBM Power with IBM Elastic Storage Server

Reference Architecture and Design

Version 3.0

IBM Power Development

May 26, 2020

POL03281USEN-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 Cloudera HDP on IBM® Power® clusters with IBM Elastic Storage Server (hereafter referred to as "HDP on Power with ESS"). This document describes the architecture for HDP on Power with ESS and two reference designs that comply with the architecture. The architecture is intended to serve as a guide for designs. The reference designs are intended to provide useful example configurations 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 reference designs for HDP on Power with ESS solutions. 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, two **reference designs** are presented that conform to the architecture.

1.2.1 How to Use this Document

For readers who wish to **produce a design** for an HDP on Power with ESS solution, the architecture portion of this document provides guidance and much of the fundamental material (incl. references to related materials and other sources) that should be understood to produce a satisfactory design. The reference designs in this document provide examples that can both assist with understanding the architecture and provide design elements and various design patterns that may be adopted when producing the design of interest.

For readers who wish to **adopt and adapt a design** for an HDP on Power with ESS solution, the architecture portion of this document may be largely skipped, and one of the reference designs may be used as a starting point for the design of interest. In this case, the "design variations" section of each reference design provides guidance regarding how the design may be modified to potentially meet the requirements for the design of interest. For design modifications that are more significant, selected sections of the architecture may be referenced for guidance on design options for the particular topic of interest.

For readers who wish to use this document as **general education** material for HDP on Power with ESS solutions, much of the material will be useful in that regard. Note, however, that most of the material is organized as a technical reference versus an instructional document, and more than one pass in reading may be required as many concepts and interrelationships exist that are clearer once broader understandings of the solution space are established.

It should be also noted that this document is not a 'how to' document for the installation and configuration procedures and processes relevant to deploying a solution. Such procedures and processes are numerous and important (especially when doing a deployment), and some applicable references are provided, but these are generally out of scope for this document.

1.2.2 Architecture versus Design

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

1.2.2.1 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.

1.2.2.2 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 designs in this document (section 5 “Reference Design 2.0A – 18 Node Base Configuration” on page 44 and section 6 “Reference Design 2.0B – 30 Node Server Dense Configuration” on page 64) describe the specific components and elements that form the solutions, specific variations of the solution, and how these are to be interconnected, integrated, and configured.

1.2.3 Architecture Diagrams

Architecture diagrams in this document are either UML or more traditional forms. Both of these diagram forms are used, and some concepts are presented in both forms to assist readers in understanding the architecture.

1.2.3.1 UML

UML class diagrams are useful for describing an architecture. The class diagrams that are used to define this architecture are a more general and a somewhat more loosely defined form of those commonly used in object-oriented related activities (for example, OOA and OOD). They are, however, very useful in defining and describing the kinds of elements that exist in the architecture and many of their relationships. A class diagram provides a representation that is usually more compact and precise, and one that can more clearly represent the range of possibilities (for example, multiplicity) that applies to the architecture. It will be clear that some liberties were taken with the UML notations, but the intent is to communicate the architecture to the reader as clearly and efficiently as possible.

The elements that are represented in these UML diagrams are all conceptual, representing architectural elements. The classes do not represent actual software classes, nor do the operations represent actual methods or functions.

1.2.3.2 Traditional

Traditional block diagrams (for example, component and stack diagrams) and also useful for describing an architecture. These are routinely less precise and usually less expressive and less complete than class diagrams. However, block diagrams are often more familiar and intuitive, and they can often represent useful examples and concepts.

1.2.4 Key Influences

This version of this architecture was influenced by preceding related work. Specifically, the Hadoop and Cloudera 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 Relationship to Previously Published Architecture

This is the third architecture published for HDP on Power. The second architecture [2] uses a traditional Hadoop "distributed storage" model where the storage that hosts the Data is distributed across the Worker Nodes and connected locally to each Worker Node. This architecture uses a "shared storage" model where all of the storage for the Data is factored into a separate, shared element (specifically an ESS) that is shared by all of the other Nodes in the Cluster. Use of a shared storage model – realized with ESS – is a primary feature introduced with this architecture. Cloudera acquired Hortonworks in January 2019, hence all the references to Hortonworks Data Platform has been changed to Cloudera HDP.

This architecture and associated reference designs *add* to the architectural and design options for deploying HDP on Power, and the previously published architecture and reference design [6] remain valid and appropriate for HDP on Power deployments.

1.4 References

[1] Cloudera HDP on IBM Power - Reference Architecture and Design - Version 2.0.
https://www.ibm.com/downloads/cas/PVYEOA0E

[2] Cloudera HDP on IBM Power with IBM Elastic Storage Server –Reference Architecture and Design - Version 2.0.
https://www.ibm.com/downloads/cas/3NYZWDNZ

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

[4] Cloudera (2020) HDP
https://www.cloudera.com/products/hdp.html

[5] IBM. (2017). IBM Spectrum Scale (*IBM Knowledge Center*)
https://www.ibm.com/support/knowledgecenter/en/STXKQY/ibmspectrumscale_welcome.html

[6] IBM. (2017). Cloudera Data Platform with IBM Spectrum Scale (a*n IBM Reference Guide*)
https://www.redbooks.ibm.com/redpapers/pdfs/redp5448.pdf

[7] IBM. (2020). IBM Elastic Storage Server (ESS) (*IBM Knowledge Center*)
https://www.ibm.com/support/knowledgecenter/en/SSYSP8/sts_welcome.html

[8] IBM File Object Solution Design Engine

https://fileobjectsolutiondesignstudio.ibm.com/

[9] Mellanox IBM Quick Reference Guide Solutions
https://www.mellanox.com/oem/ibm

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™ and ESS. 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 3 "Architecture - Overview" on page 6). 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 for the Platform. This architecture also prescribes IBM Spectrum Scale™ and ESS extensions and modifications to the HDP suite – replacing the native HDFS installation within HDP. A suitable design for the Platform 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. For the HDP components that are not affected by the Spectrum Scale and ESS extensions and modifications, this architecture does not prescribe any particular architecture or design within that portion of the HDP installation. Further, this architecture assumes common design patterns and best practices for the HDP installation as recommended by Cloudera, and a primary intent and scope of this architecture is to provide a hosting environment that can accommodate any reasonable HDP design, including the Spectrum Scale and ESS extensions.

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 Applications. The Infrastructure is a primary scope of this architecture and reference designs.

3 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. This overview factors the System into three layers, and the architecture sections which follow provide more detailed descriptions of each layer. These descriptions provide increasing levels of detail and progressively expand the architectural views – ultimately providing a complete architecture representation that captures the relevant abstractions and relationships.

3.1 Elements

This section describes the elements relevant to this architecture.

3.1.1 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).

3.1.2 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.

3.1.3 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 Cloudera 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. Refer to the Cloudera website for HDP documentation [5].

This architecture also prescribes Spectrum Scale and ESS extensions and modifications to the HDP suite – replacing the native HDFS installation within HDP. A suitable design for the Platform software installation (for example, specific choices of HDP components to include, distribution of these components across nodes, configuration of these components) is necessary for a complete and operational environment. For the HDP components that are not affected by the Spectrum Scale and ESS extensions and modifications, this architecture does not prescribe any particular architecture or design within that portion of the HDP installation. For the Spectrum Scale and ESS extensions, this architecture covers these in later sections.

3.1.4 Infrastructure

Infrastructure is the set of hardware (including servers, storage, network) and all of the system software, firmware, and infrastructure management software that are necessary to host and support 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 a primary scope of this architecture.

3.2 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, extended with Spectrum Scale. 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), implemented via Spectrum Scale, 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.

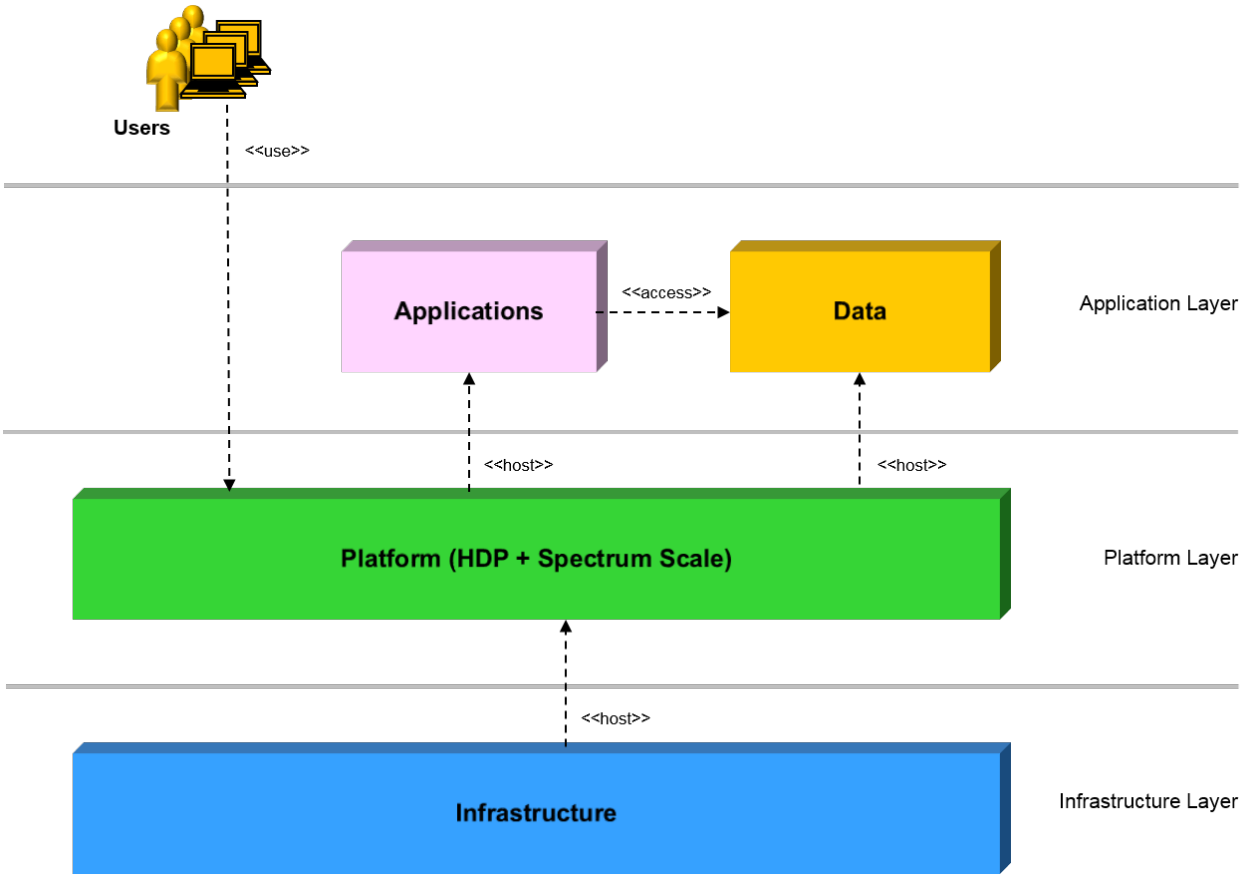

Figure 1. Architecture Overview – Top-Level Elements

4 Architecture

This section describes the architectural elements, followed by a description of how these are composed into various architectural views. It concludes with a discussion of some operations that are uniquely affected by this architecture and some notes relevant to sizing a System.

4.1 Elements

This section describes the elements relevant to this architecture at the Application, Platform, and Infrastructure layers.

4.1.1 Roles

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

4.1.1.1 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.

4.1.1.2 Application Developer

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

4.1.1.3 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.

4.1.1.4 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, and so on), but the nature of the role remains similar across specialties.

4.1.2 Data

Data are key elements in the Application Layer. At the Application Layer, Data may simply be viewed as the data that is consumed as input and produced as output by Applications. Initially, Data is typically loaded into the Data Store by an ingest process that copies the Data from an External Data Source.

4.1.3 Application

Applications are another key element in the Application Layer. One or more Applications are executed as part of a Job.

4.1.4 Job

A Job is a logical element that represents an analytic operation that a User wishes to execute. A User submits a Job and monitor its execution using HDP Functions.

4.1.5 Data Store

The Data Store is the element within this architecture that hosts (stores) the Data. At the Application Layer, the Data Store provides an HDFS interface for use by Applications. At the Platform Layer, the Data Store adds a Spectrum Scale File System implementation which underlies the HDFS interface. This structure of the Data Store allows it to be fully compatible with HDFS clients while gaining the advantages of Spectrum Scale as the native file system implementation.

4.1.5.1 Elastic Storage Server

The IBM Elastic Storage Server (ESS) is the specific Data Store recognized by this architecture. The ESS provides a native Spectrum Scale File System abstraction. The ESS is configured as one Spectrum Scale Cluster and the HDP Cluster is configured as a separate Spectrum Scale Cluster. The HDP Cluster (as a Spectrum Scale Cluster) remote mounts the relevant Spectrum Scale File System(s) hosted on the ESS to obtain access to the Data in the ESS.

The ESS also includes a user interface for admins to manage the ESS (for example, create file systems, monitor hardware, and so on).

The ESS is a modular component within the architecture, and any model of the ESS may be used within the System. Further, more than one ESS may be used to realize the Data Store for a System. Internal details of the structure of the ESS are suppressed from this architecture to the maximum degree possible, with such details becoming relevant only at a design or implementation level.

4.1.6 File Systems

The file systems for the Data within this architecture are factored into two primary and interrelated parts. First is the HDFS *abstraction* that is presented to Applications and Platform layer services and components that operate as HDFS clients. Second is the Spectrum Scale *implementation* that underlies the HDFS abstraction and serves as the native file system that hosts the Data.

Other file system interfaces may be provided as part of the ESS implementation. These other interfaces may be considered as alternate abstractions, paths, and mechanisms by which the Data in the Data Store may be accessed (typically by external-to-HDP functions). These other interfaces are fully compatible with this architecture, but they are not discussed further in this document. Refer to Elastic Storage Server documentation for more information on this point.

4.1.6.1 Hadoop Distributed File System (HDFS)

HDFS is the file system *abstraction* used to hold the Data. A variety of HDP and Spectrum Scale Storage Functions are used to realize the HDFS abstraction within the Platform. These components are widely hosted across most (typically all) of the HDP Nodes.

4.1.6.2 IBM Spectrum Scale File System

IBM Spectrum Scale File System is the native file system *implementation* used to host the Data. Spectrum Scale components are installed in the Worker, Master, and Edge Nodes to efficiently support these Nodes operating as clients of the ESS. The existence and operation of the Spectrum Scale file system is transparent to the clients, Applications, and Functions which are using the HDFS abstraction.

Spectrum Scale File System components are also installed on the ESS in this architecture as part of the file system implementation. These are an integral part of the ESS.

4.1.7 ESS Administrator User Interface

The ESS Administrator User Interface provides an interface into the ESS that is used by Administrators to manage and monitor the ESS (for example, create File Systems, monitor storage utilization, monitor performance). Within this architecture, the ESS is a modular element that has management requirements separate and distinct from the rest of the Platform and the HDP Cluster. The ESS Administrator User Interface provides this capability. The ESS Administrator User Interface is an integral part of the ESS, and it is hosted on the Elastic Management Server within the ESS.

It is relevant to note that the ESS Administrator User Interface includes both Platform layer functions (for example, creating a File System), and Infrastructure Layer functions (for example, managing and monitoring the physical Nodes within the ESS).

4.1.8 External Data Source

Data moving into and out of the Cluster does 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, and it is limited only by the capability of the data import/export functions which connect to it.

This element is specifically noted as the volume and rate of data import and export and the path and mechanisms used may be important design criteria. Within this architecture, Data may be moved into Data Store through the Edge Nodes as in the traditional Hadoop model, or it may be written directly to the Data Store (that is, data may be written directly into the ESS without handling by any Hadoop level functions or routing through any HDP Node).

4.1.9 HDP Functions

Multiple Functions exist within the Platform layer. Many of these are visible to the Application layer for use by Applications and Users. Most of these Functions are provided as components of HDP, with some additional components provided by Spectrum Scale.

4.1.9.1 Management Function

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

4.1.9.2 Storage Function

Some of the HDP components are Storage Functions that provide the HDFS storage abstraction and supporting services (for example, the NameNode service). Spectrum Scale provides some additional Storage Functions to support the native Spectrum Scale File System implementation. Various Storage Functions are hosted on most (typically all) of the HDP Nodes in the HDP Cluster.

4.1.9.3 Edge Function

Some of the HDP components are Edge Functions which run as services on some of the HDP Nodes. 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 Edge Functions run on the Edge Nodes.

4.1.10 Spectrum Scale Functions

Three Spectrum Scale Functions are added to the Platform to add Spectrum Scale capability and replace HDFS as the native file system implementation in this architecture.

4.1.10.1 Spectrum Scale Client

The Spectrum Scale Client provides Spectrum Scale file system access to the HDP Nodes. The Spectrum Scale Client is installed on all HDP Nodes.

4.1.10.2 Spectrum Scale Server

The Spectrum Scale Server runs Spectrum Scale server code as a member of Spectrum Scale cluster. This allows the HDP Cluster to be configured as a Spectrum Scale Cluster with three quorum nodes. Two Edge Nodes run Spectrum Scale server code which allows them to also provide direct Spectrum Scale access outside of the HDP Cluster. HDP Master Nodes also run Spectrum Scale server code.

4.1.10.3 HDFS Transparency Connector

The HDFS Transparency Connector is the Spectrum Scale Function that bridges between the HDFS components in the Platform (especially the HDFS Client code) and rest of the Spectrum Scale components (especially the Spectrum Scale Client). The HDFS Transparency Connector is essential to supporting the HDFS abstraction for clients and a Spectrum Scale implementation for the native file system. The HDP Transparency Connector is installed on all HDP Nodes.

4.1.10.4 Spectrum Scale Ambari Management Pack

The Spectrum Scale Ambari Management Pack is a Spectrum Scale component that handles the installation and monitoring of the other Spectrum Scale components and Functions that are added to the Platform in this architecture. The Spectrum Scale Ambari Management Pack is separately installed and configured on the same server that hosts Ambari (typically an Edge Node). This module was formerly called the *GPFS Ambari Integration Module*.

4.1.11 Cluster

A Cluster is some collection of Nodes in the System.

4.1.11.1 HDP Cluster

The HDP Cluster is the subset of the Nodes (HDP Nodes) that host HDP services and components.

4.1.11.2 Spectrum Scale Cluster

A Spectrum Scale Cluster is a logical collection of Nodes, each of which hosts some Spectrum Scale components, that are grouped to form an element relevant to the operation of Spectrum Scale. This architecture includes two Spectrum Scale Clusters: One Spectrum Scale Cluster runs on HDP nodes, and another Spectrum Scale Cluster runs within the ESS. The Spectrum Scale cluster running on HDP nodes does not use any local storage for the Spectrum Scale file system. It just remote mounts the relevant Spectrum Scale File System(s) hosted on the ESS to obtain access to the Data in the ESS.

4.1.12 Nodes – HDP Cluster

A Node in the HDP Cluster 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 the Platform layer. Nodes in the HDP Cluster are categorized into the following primary types:

4.1.12.1 Worker Node

A Worker Node is 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. Within this architecture, Worker Nodes obtain Data from the Data Store (for example, no Data is persistently stored on the Worker Nodes' storage).

An HDP Cluster must have at least one Worker Node, but an HDP Cluster typically has more (often many more) Worker Nodes. 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.

4.1.12.2 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 a minimum of three Master Nodes are commonly used to provide basic high availability (HA) capability.

4.1.12.3 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 the HDP 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 HDP 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.

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

4.1.12.4 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.

4.1.12.5 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.

4.1.12.6 System Management Node

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

4.1.13 Nodes – Elastic Storage Server

An Elastic Storage Server includes a set of Nodes (Servers). As noted previously, within this architecture, the *internal* structure of the ESS is intentional hidden wherever possible. However, in some cases, the Nodes with the Storage Server must be considered.

4.1.13.1 Storage Nodes

A Storage Node within the ESS is a Server that directly provides Data access to clients of the ESS. There are typically two Storage Nodes within an ESS (containing one ESS building block).

4.1.13.2 Elastic Management Server

The Elastic Management Server within the Storage Server is a Server which provides basic internal management functions for the ESS and its components and which hosts the UI functions used by the Storage Administrator.

4.1.14 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.

4.1.14.1 Ambari

The Platform Manager function is commonly provided by Ambari for Hadoop Platforms, and this architecture requires Ambari as the Platform Manager.

4.1.15 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.

4.1.16 Operating System

4.1.16.1 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 by this architecture.

4.1.17 Server

A Server is the physical element that forms the foundation of a Node. Only physical servers and IBM POWER9 based servers are presently recognized by this architecture.

4.1.18 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 for the HDP Cluster. The ESS may include any Management Processor that is consistent with its definition.

4.1.19 Network Subsystem

The Network Subsystem proper is an Infrastructure layer concept. However, the Application and Platform layers have some implicit requirements for connectivity between the HDP Nodes and between the HDP Nodes and the Data Store that are useful to note here. Network topologies may vary, but specifically (and most commonly), the HDP Cluster requires a networking infrastructure that supports "external" and "internal" connections. The internal connections are for connections between all of the Nodes in the HDP Cluster and to the Data Store, and these internal connections are typically restricted to the HDP Cluster and Data Store. The external connections are for providing access for Users using the System and Admins managing the Platform. Most details are deferred to the Infrastructure layer, but it is useful to note that for the Application and Platform layers, the internal connections are provided by a Data Network (typically a high-speed network), and the external connections are provided by a Campus Network.

Architecturally, note that the scope and definition of the Network Subsystem excludes any ESS *internal private* networks which are typically part of its internal structure. Relevant ESS internal networks are included as part of a specific design and examples are included in the reference designs in this document.

4.1.19.1 Data Network

The Data Network is a (typically) private network that is used to provide high speed, high bandwidth, and/or low latency communication between the HDP Nodes and between the HDP Nodes and the Data Store (specifically the Storage Nodes within the ESS). The Application Network may be Ethernet or InfiniBand®.

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.

4.1.19.2 Campus Network

The Campus Network is the primary path for Users to access the system. Users access HDP Functions over the Campus Network. The Campus Network is also the default path for movement of Data from an External Data Source to the Data Store.

4.1.19.3 Management Network

The Management Network is the commonly defined network that is 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 not be distinct from the Campus Network in some client environments (that is, the Campus Network may also be used for Management).

4.1.19.4 Provisioning Networks

The Provisioning Networks are private networks that are used by the System Management Node and the Elastic Management Server to accomplish the provisioning of nodes within the system and subsequent basic monitoring of these nodes. The HDP Cluster and the ESS have separate and independent Provisioning Networks.

4.1.19.5 Service Network

The Service Network is a private network that is used to access the management processors (BMCs or FSPs) of the servers within the system. Specifically, for servers that are managed by an HMC, this is the private network that is used by the HMC to connect to the FSPs of those servers. For servers that are not managed by an HMC, this is the network over which persons or other elements access the Management Processors to accomplish operations such as power control of Servers. The HDP Cluster and the ESS have separate and independent Service Networks.

4.1.20 Switches

Switches host the Networks defined by this architecture. Specifics regarding the number and nature of the Switches in the System are largely a design choice. However, this architecture makes some assumptions in this regard to allow factoring of the elements into useful groups which are defined as part of the racking and physical layout of the System (see section 4.4.5 “Racking and Physical Layout” on page 33).

4.1.20.1 Data Switch

One or more Data Switches host the Data Network. Specifics regarding the number and nature of the Data Switches are largely a design choice, but some architectural guidance is included.

4.1.20.2 Utility Switch

One or more Utility Switches host the Campus Network and Management Network for the HDP Cluster and the ESS. These Utility Switches also host the Provisioning Network and Service Network for the HDP Cluster. Specifics regarding the number and nature of the Utility Switches are largely a design choice, but some architectural guidance is included.

4.1.20.3 ESS Internal Management Switch

The ESS may include one or more Internal Management Switches that host that ESS Provisioning Network and ESS Service Networks or which provide similar internal connectivity for the ESS.

4.2 Composition – Overview

It is useful to start with some higher-level views of the architecture to provide some context and orientation. Refer to Figure 2 for a class diagram providing an initial, high-level summary view of the architecture. This view represents some of the most important elements of this architecture and their primary relationships. It is described first to allow a high-level and simplified view that provides context, orientation, and a useful base for the more detailed views and representations in the following sections.

The HDP Cluster and the ESS are important composite elements in this architecture at the Platform and Infrastructure layers. Starting with these:

- The HDP Cluster hosts Jobs which the User creates and submits. These Jobs use the Applications which in turn use (access) Data.
- The HDP Cluster consists of multiple HDP Nodes.
- Each HDP Node includes some HDP Functions and some Spectrum Scale Functions running on an instance of Linux hosted on a Server.
- The HDP Cluster uses the Network Subsystem to access the ESS and to allow Users and Admins access to the HDP Cluster.
- The ESS is the Data Store within this architecture – that is, the ESS is the specific Data Store chosen for this architecture.
- The Data Store hosts the Data for the System, and it provide an HDFS abstraction to the Application layer.
- The ESS provides a Spectrum Scale File System abstraction to the HDP Cluster which the Spectrum Scale Functions use to access and manipulate the Data.
- The ESS uses the Network Subsystem to serve data to its clients (HDP Nodes in the HDP Cluster) and to allow Admins access to the ESS for management and monitoring.

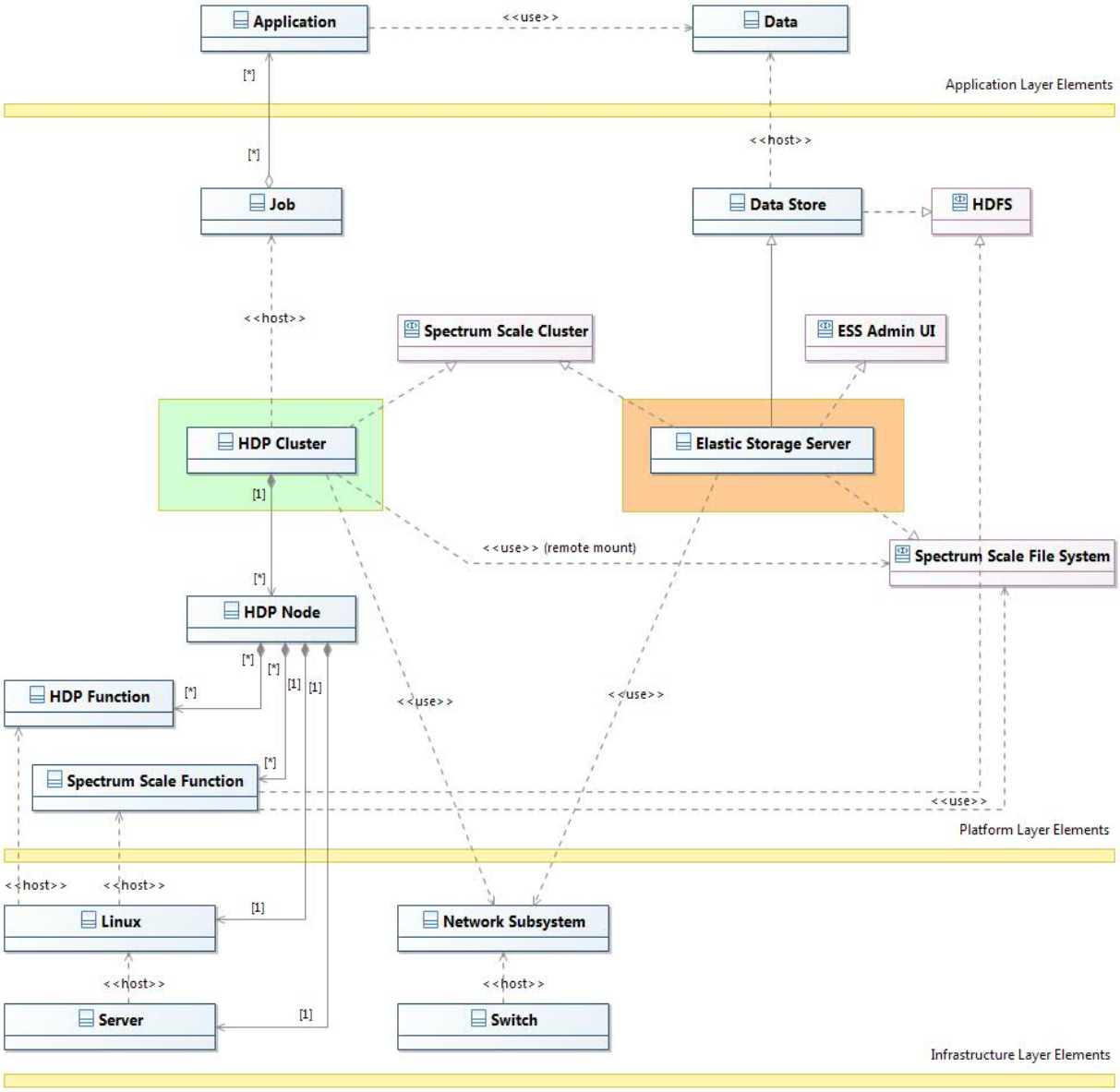

Figure 2. Class Diagram – All Layers – Summary

Refer to Figure 3 for a class diagram providing a more detailed and comprehensive view of this architecture. Details are more legible in additional diagrams which follow, and descriptions are in the sections which follow. However, Figure 3 illustrates an overview of how all of the elements at all of the layers relate. Readers may find it most useful to only lightly review this diagram on first reading and refer back to this detailed view of the architecture after the information in the focus sections is understood.

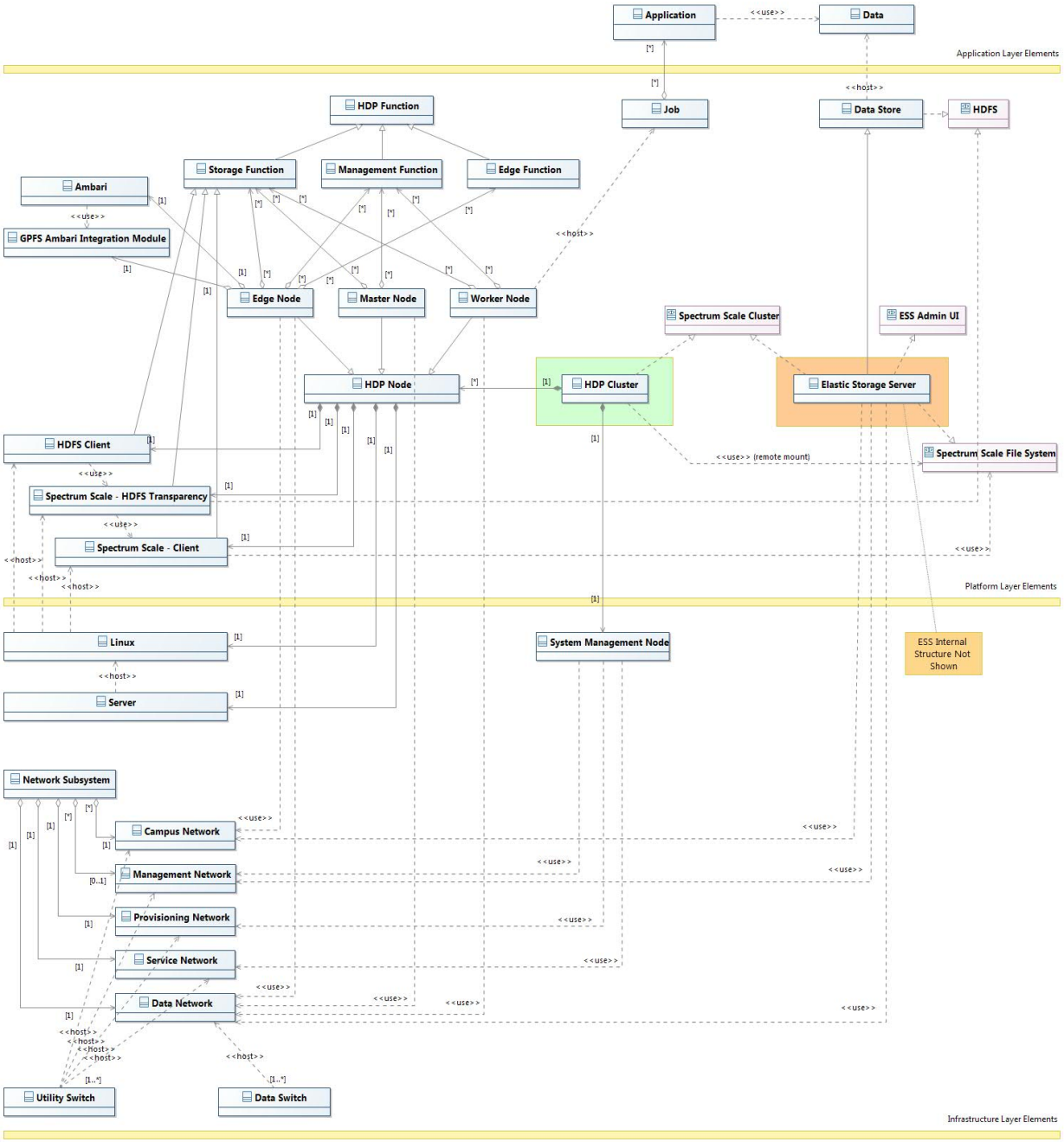

Figure 3. Class Diagram - All Layers – Detail

4.2.1 System Summary View

A summary representation of the System that is useful as a higher level of abstraction is depicted in Figure 4. This view cuts across layers and other boundaries and represents only selected parts of the System, but it factors the System into four parts that are often useful to consider independently or to enable some simplifying focus.

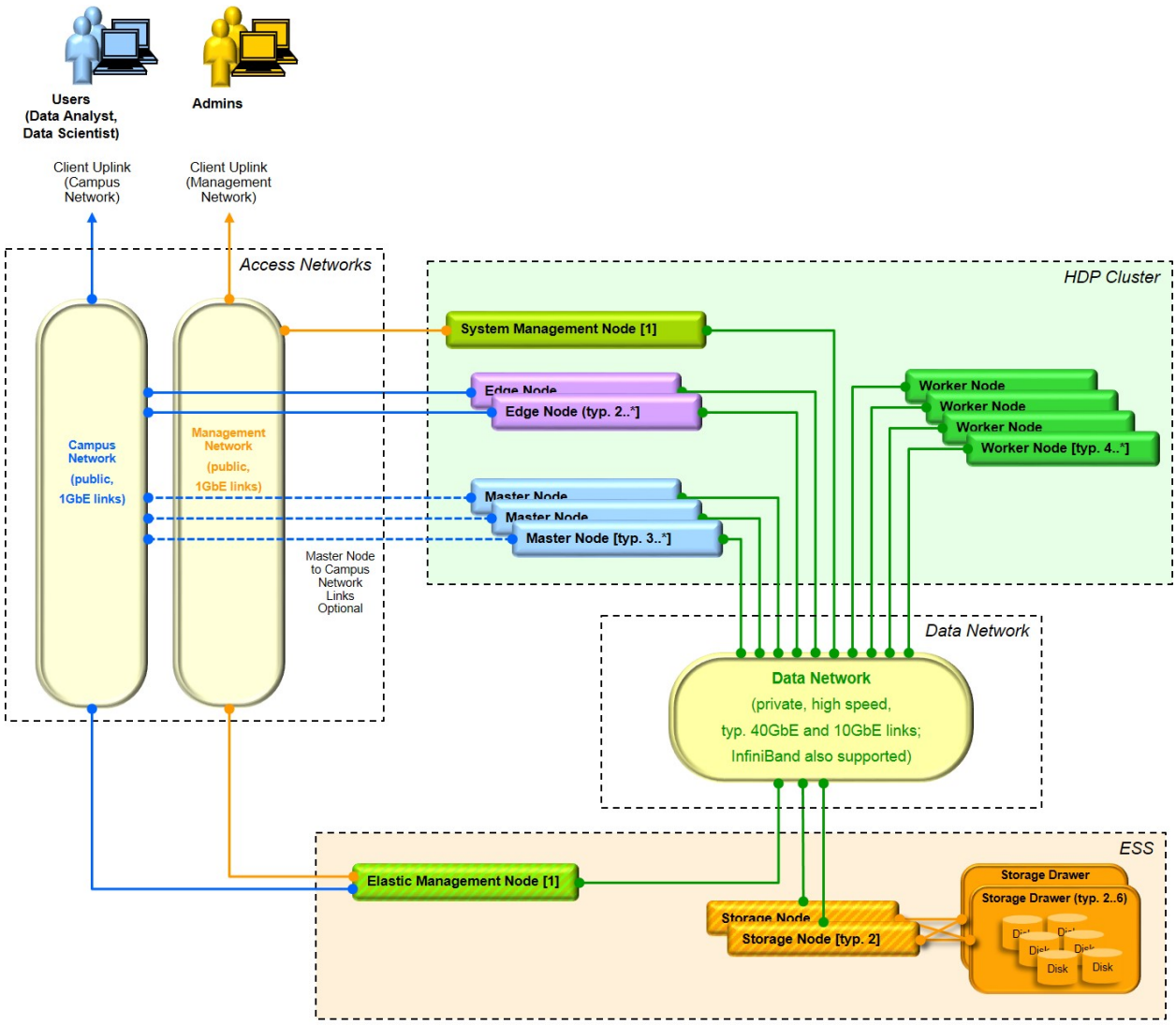

Figure 4. System Summary View

4.3 Composition – Application and Platform Layer View

Refer to Figure 5 for a class diagram representing an Application and Platform layer view within this architecture. Refer also to Figure 6 and Figure 7 which zoom in on the Platform to provide focus on the HDP Cluster and ESS portions of the Platform respectively. Refer to Figure 8 and Figure 9 for more traditional diagrams representing Application and Platform layer views for this architecture. Within this view, there is emphasis on describing the unique features of this architecture – specifically the ESS and related Spectrum Scale Functions.

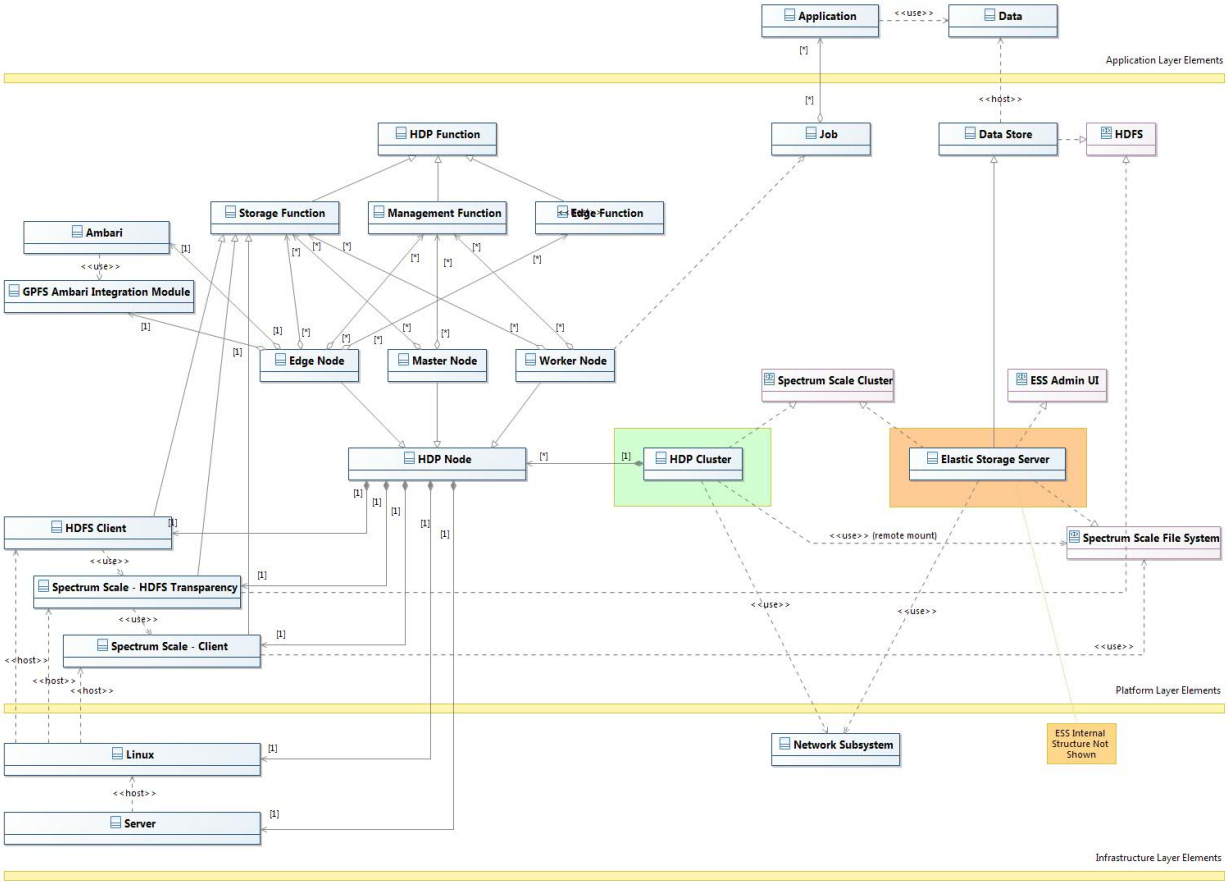

Figure 5. Class Diagram - Application and Platform View

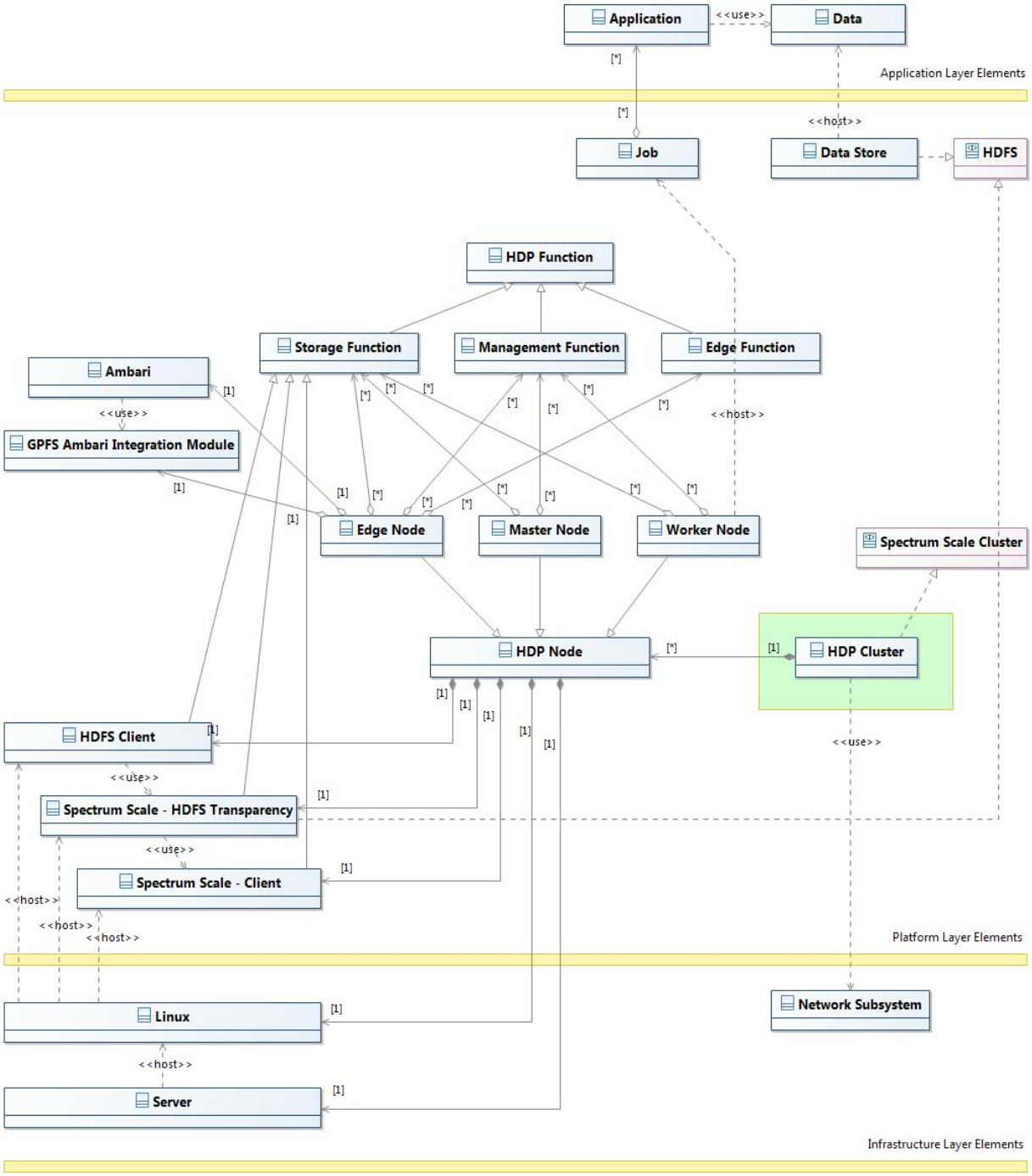

Figure 6. Class Diagram - Application and Platform View - HDP Cluster Focus

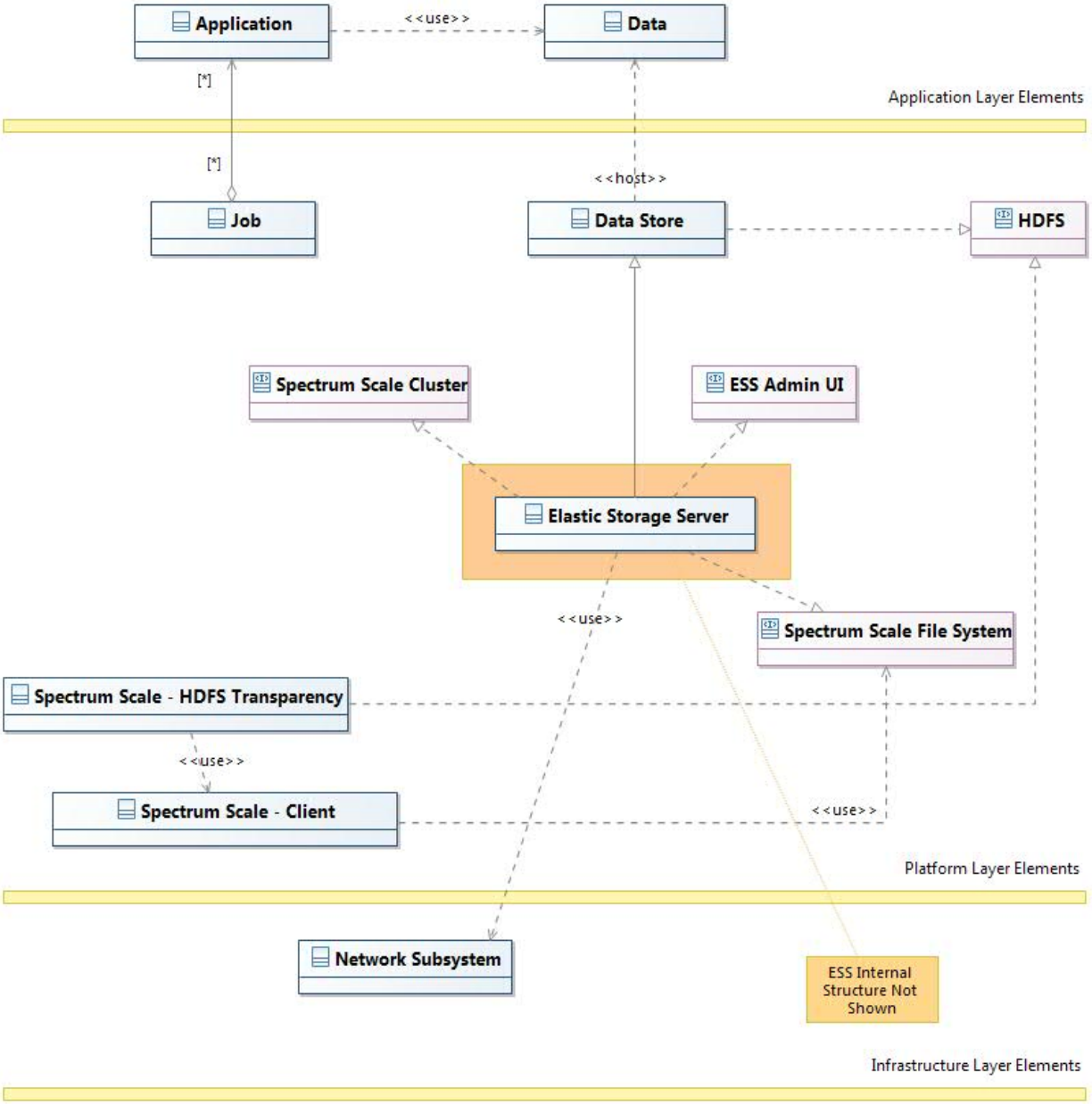

Figure 7. Class Diagram - Application and Platform View - ESS Focus

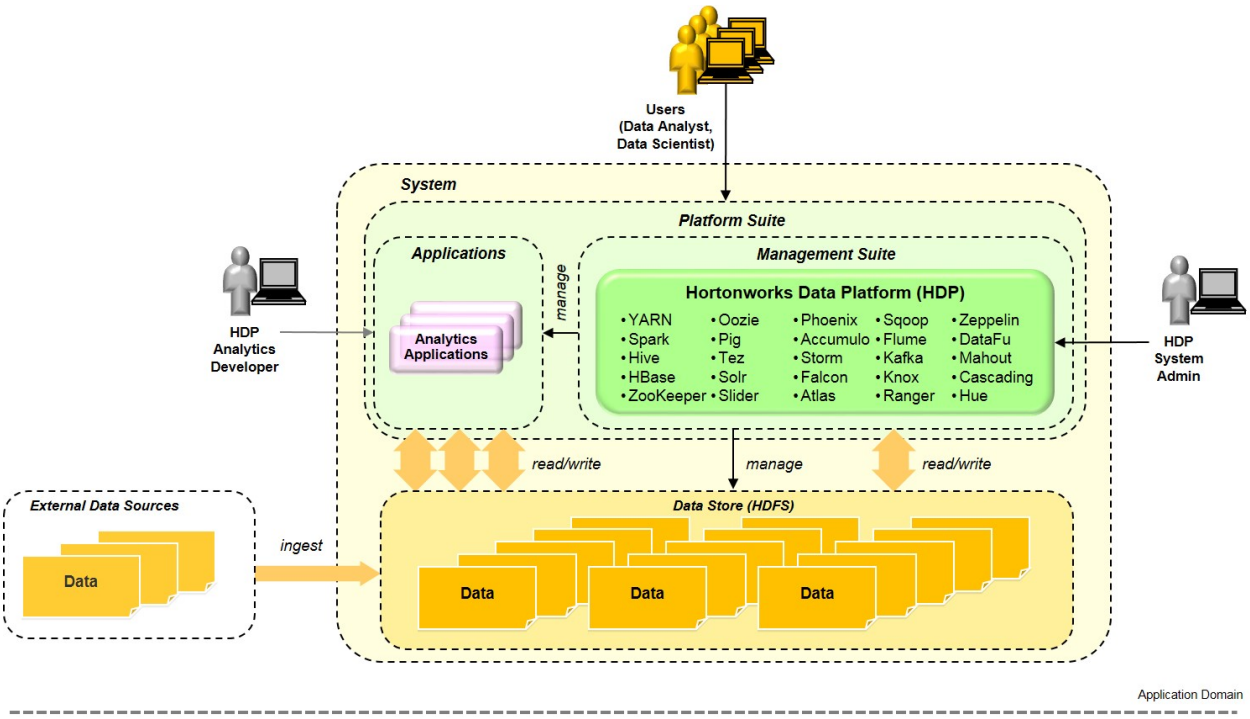

Figure 8. Architecture – Application View – Overview

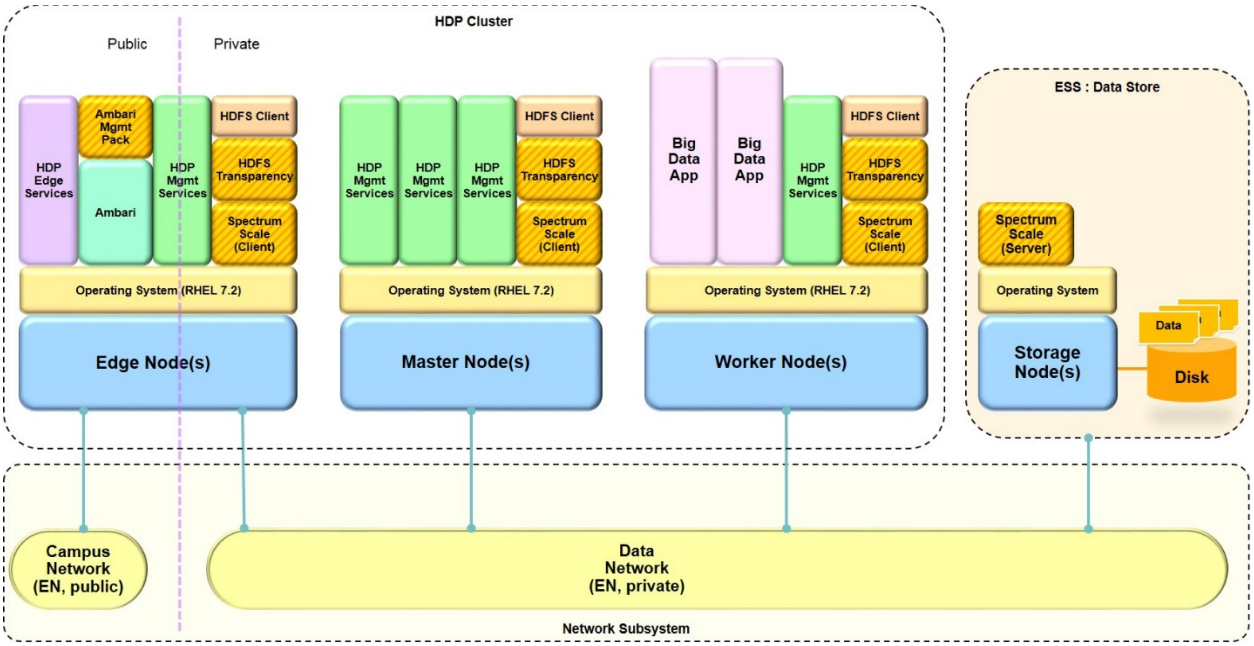

Figure 9. Architecture – Platform View – Node Types and Hosting Focus

At the Application layer, from a User point of view, the primary items with which the User works are the Data and the Applications, and the Platform logically hosts the Applications and Data. The Platform also provides primary access for the User in the form of the HDP Management and Edge Functions which allow a User to accomplish tasks such as submission and monitoring of Jobs. The "platform" represented by HDP allows the User to have a convenient and simple abstraction of the System.

At the Platform layer, it is useful to consider the Platform in two primary parts: the HDP Cluster and the ESS. At this level of abstraction, the ESS provides the Data Store for the System, and the HDP Cluster provides the balance of the Platform capabilities. Within the Platform layer, the HDP Cluster uses the ESS as the Data Store of the System, and the ESS is distinct and separate from the HDP Cluster. (This is in contrast to the traditional HDP "distributed storage" model where the HDP Cluster also provides the Data Store for the System.)

4.3.1 HDP Cluster

The HDP Cluster is a collection of HDP Nodes that collectively hosts all of the HDP Functions (Edge Functions, Management Functions, and Storage Functions) and the Spectrum Scale Functions that are of particular interest to this architecture. As a first order view, the collection of Functions at the Platform layer cooperate and interact to realize the Platform. The Platform provides interfaces, services, and touchpoints for Users and Admins. The relationships and interactions of the various HDP Functions is a significant and complex topic in its own right which is covered by the HDP and Hadoop architectures. The particular selection of HDP Functions for a System is an important architectural and design choice for this layer of the System. Further, the Spectrum Scale Functions which are added to the HDP Cluster are particularly relevant to this architecture, and these are discussed further below. However, most details regarding the other HDP Functions are not relevant to the scope of this architecture, and the interested reader is referred to Cloudera education materials for more information on the HDP architecture within this layer (refer to [4] and [5]).

As a second order view, the HDP portion of the Platform can be considered to be composed of a set of Functions which are hosted across the set of Nodes that form the HDP Cluster. How these Functions are distributed across the collection of HDP Nodes is another 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 10 for an example of this view of the Platform layer.

One of the most relevant parts of this architecture is the nature of the Nodes (and support elements) which 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. These Infrastructure layer considerations are a primary focus of this architecture.

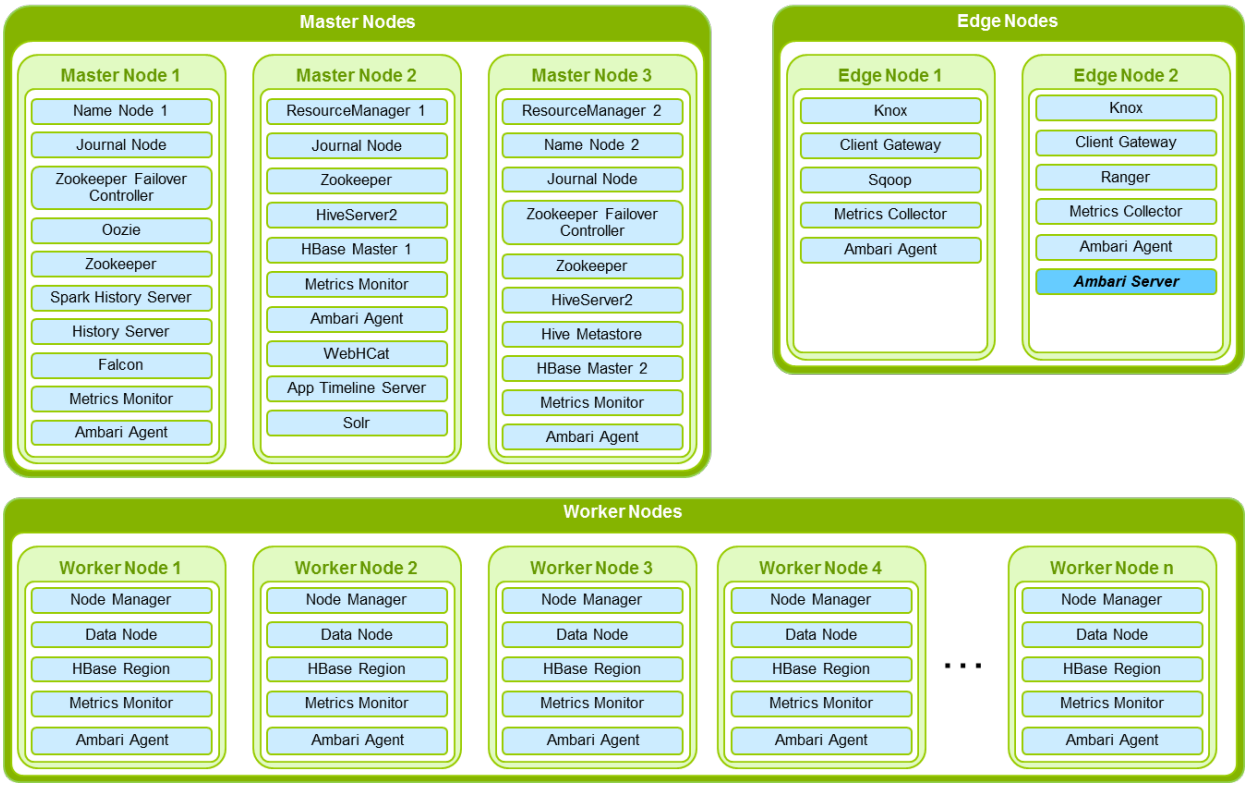

Figure 10. Platform Layer View – HDP Functions and Distribution Across Nodes - Example

4.3.1.1 Spectrum Scale Functions

While most of the Functions hosted by the HDP Cluster need not be considered further by this architecture, the Spectrum Scale Functions that are added to the HDP portion of the Platform warrant additional description. Specifically, each HDP Node adds the Spectrum Scale HDFS Transparency and Spectrum Scale Client components to the Data access stack. The resulting stack has three primary layers: Starting with the topmost layer, these are the HDFS Client code (part of HDP), the HDFS Transparency Connector, and the Spectrum Scale Client code (this last component may be installed with either a "client" *license* or a "server" *license*). The specific details of how these are configured is beyond the scope of this document, but additional information and direction are available with the Spectrum Scale documentation. Refer to [7], [8], and [9].

To install and configure the above Spectrum Scale Functions and configure the modified Storage Functions within the HDP Cluster, a third Spectrum Scale Function – the "Ambari Management Pack" -- is required to handle the installation and management of the other Spectrum Scale components. The Spectrum Scale Ambari Management Pack is installed with Ambari, and it extends the capabilities of Ambari to accomplish the above.

4.3.2 ESS

The ESS is a modular element that realizes the Data Store for the System. Architecturally, one or more ESSes may serve as the Data Store, and any ESS model or combination of ESS models may be used. It is relevant to note that the HDFS interface (abstraction) of the Data Store is actually realized by the HDFS Transparency Functions hosted on the HDP Cluster. Thus, it is perhaps more precise to describe the Data Store as realized by the combination of ESS with the Spectrum Scale HDFS Transparency Function supported by related Spectrum Functions.

Within this architecture, the ESS is configured as one Spectrum Scale Cluster and the HDP Cluster is configured as a separate Spectrum Scale Cluster. The HDP Cluster (as a Spectrum Scale Cluster) "remotely mounts" the relevant Spectrum Scale File System(s) hosted on the ESS (another Spectrum Scale Cluster) to obtain access to the Data in the ESS. The HDP Nodes operate as File System clients of the ESS (using the HDFS abstraction or Spectrum Scale natively, depending upon the component).

As a modular element, the ESS is managed directly (that is, not through the HDP Functions). The appropriate Admins use the ESS Administrator User Interface to create Spectrum File Systems and monitor and manage the operation of the ESS. General best practices for operating and managing the ESS apply for this architecture. Refer to [8] and [10] for more information regarding configuration and management of the ESS for HDP.

4.3.3 Network Subsystem

At the Platform layer, various network patterns may be used successfully for an HDP deployment. The most common options are enumerated and described in section "Appendix A - Network Patterns" on page 87. However, to allow this architecture to be more specific and complete, one pattern – the "Partial-Homed" pattern – is chosen and prescribed. The Partial-Homed pattern meets many common requirements and provides several desirable characteristics. In summary, this pattern consists of two primary networks:

- A private, high-speed network – the Data Network – to which all of the HDP Nodes and the ESS are connected and which they used to accomplish most operations.
- A public network – the Campus Network – which is used for User and Admin access to the System. Only selected Nodes (typically Edge Nodes and perhaps Master Nodes) are attached to the Campus Network.

See Figure 11.

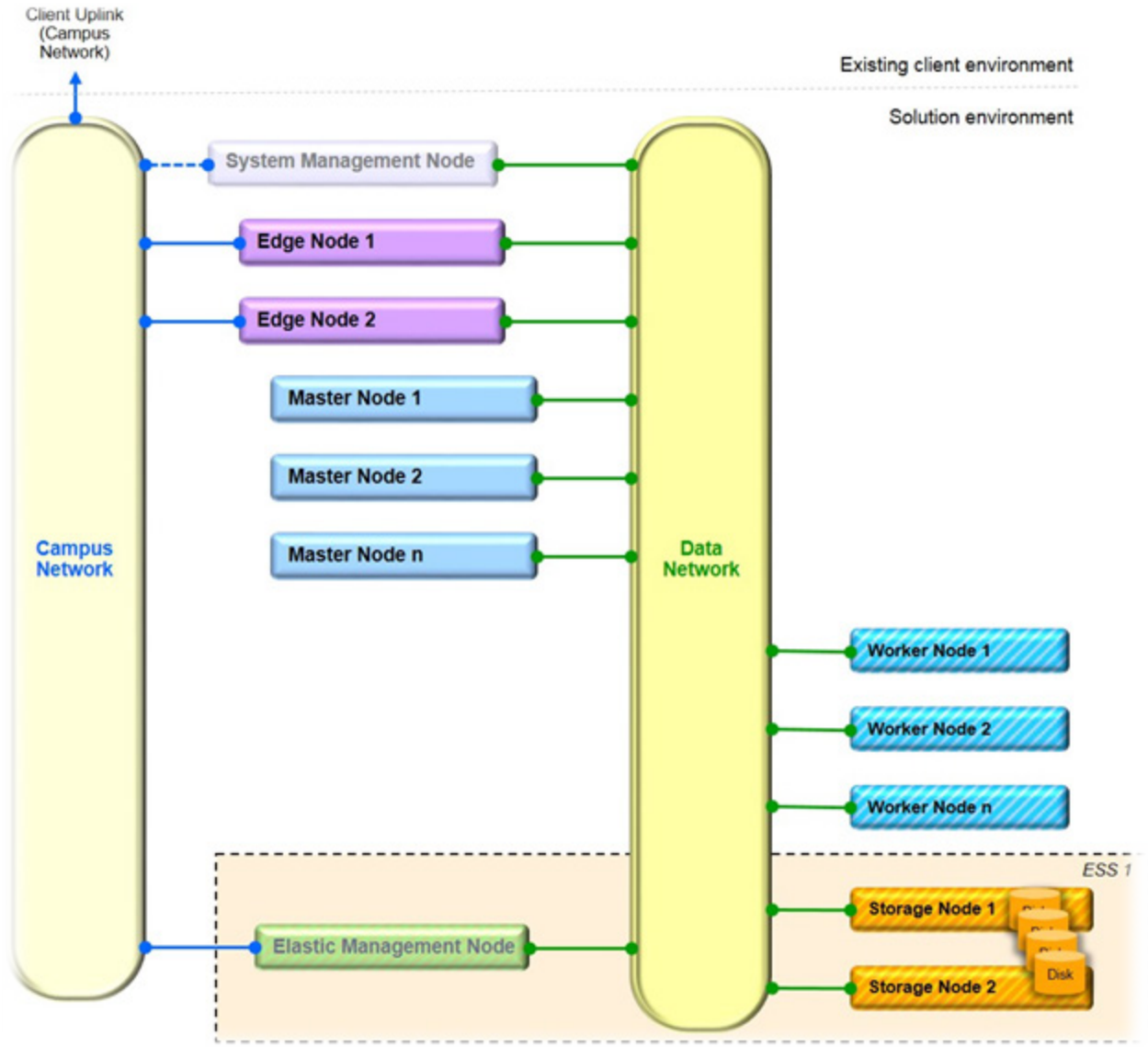

Figure 11. Partial-Homed Network

Even though this architecture prescribes the Partial-Homed network pattern, designers are encouraged to select another pattern if it better meets the client requirements. These other patterns referenced above can typically be applied with good results. Selecting a different network pattern does imply changes which propagate into various other areas of this architecture and into the related reference designs, so some additional investigation and work is required, but most of the changes are reasonably apparent and straightforward.

4.4 Composition – Infrastructure View

4.4.1 Node Composition

The HDP related elements above are configured as Nodes and interconnected to form the HDP Cluster. The Nodes are provisioned with an Operating System and are configured to have an appropriate network connectivity.

4.4.1.1 Worker Nodes

Worker Nodes host Jobs, which in turn run Applications, as directed by HDP Management Functions. The Jobs running on the Worker Nodes use Storage Functions (the Data access stack described previously) to access Data which is hosted on the ESS.

Each Worker Node consists of an IBM 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 Cloudera as a supported OS for the particular version of HDP to be used. Currently, RHEL 7.6 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 that is compatible with the rest of the elements in the Infrastructure and that meet the system level requirements. 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 deployment.
- Worker Nodes are frequently optimized for performance and Data access speeds 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. At least 128 GB and often more memory per Node is common.
  - Local storage for intermediate processing (for example, scratch space and/or shuffle space). The optimal amount and type (for example, HDD or SSD) of Storage is highly dependent upon the particular workload characteristics, but four 2.4 TB HDD drives is a useful initial guideline. A particular design may choose to RAID these local drives, and a straightforward capacity versus resilience design choice may be made in this regard. All of the local storage is presented to the OS.
  - High performance storage controllers are preferred.
  - Significant network bandwidth to the Data Network. 25 GbE (or two 25 GbE links) per Node or better is common.
- Worker Nodes generally need not be configured for high availability characteristics. The HDP Functions and Hadoop architecture tolerate significant failures within the collection of Worker Nodes. Thus, Worker Node components can typically be chosen to optimize performance versus resilience.
- Every Worker Node is typically configured with the same hardware.

4.4.1.2 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 Cloudera as a supported OS. 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 that 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 failures within the Master Nodes, but this resilience is not complete, and the failure of a Master Node can be disruptive. Thus, 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 a single 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 like traditional servers which host general application software. The set of drives is typically configured using RAID 5 or RAID 10, and all drives are presented to the OS.
  - 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 than other Master Nodes.

4.4.1.3 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, two 25GbE for internal connections to the Data Network, plus an additional two 25GbE connectivity to the external network used to access the External Data Source).

4.4.1.4 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:

- 12 CPU cores and any clock rate above 2GHz
- 32 GB of memory
- 4 TB of usable storage (for example, two 2.4 TB)

4.4.1.5 Machine Learning/Deep Learning Node

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.

4.4.2 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 4.1.12 "Nodes – HDP Cluster" 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.

4.4.2.1 Worker Nodes

Four (4) Worker Nodes minimum. In this architecture the Worker Nodes are dedicated as the hosts for Jobs running the Applications. Thus, the Worker Node count must provide sufficient compute capacity generally for Jobs and be sufficient in number and capacity to meet operational requirements for throughput and response time when one or more Worker Nodes is unavailable due to an outage. Note that in contrast to clusters with a traditional "distributed storage" model, the Worker Node count may be chosen independently from the Data Store capacity of the System.

4.4.2.2 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 12) provides some guidance from Cloudera 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 12. Suggested Master Node Counts for Various Cluster Sizes

4.4.2.3 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.

4.4.2.4 Machine Learning/Deep Learning Nodes

One or more Machine Learning or 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.

4.4.3 Cluster Types

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

4.4.3.1 Balanced

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

4.4.3.2 Performance

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

4.4.3.3 Server Dense

A "Server Dense" cluster is a cluster for which the design choices reflect more preference for maximizing the density of the components (especially Servers) in the Server Racks.

4.4.4 Network Subsystem

At the Infrastructure layer, the Network Subsystem is the collection of logical networks and physical elements (for example, switches, cables) that host and realize them. The logical networks are specified by this architecture and consist of the networks listed previously (see section 4.1.19 "Network Subsystem" on page 15).

The Campus Network and the Data Network are introduced at the Platform layer to satisfy its connectivity requirements – following the Partial-Homed network topology described previously. At the Infrastructure layer, the Management Network, the Provisioning Network, and the Service Network are added. The Management Network is a public network (that is, external to the HDP Cluster) which provides access for Admins to various Infrastructure elements (for example, System Management Server, Elastic Management Server, and Switches). The Provisioning Network and the Service Network are private networks (that is, internal to the HDP Cluster) and used only for internal operations within the HDP Cluster.

The Campus Network and the Management Network can be considered to be "access" networks for the Platform and Infrastructure layers respectively – providing access to Users and Admins to relevant elements in the System. The other networks can be considered "private" networks with the Data Network providing connectivity between the HDP Nodes and the ESS and the Provisioning Network and the Service Network strictly contained within and used by the HDP Cluster to accomplish internal operations.

The ESS typically includes some *internal private* networks and switches as part of its internal structure, but these are strictly contained within the ESS and treated only as required as part of a design.

The specific collection of Switches (and their configurations) that realize these networks is largely left as a design consideration, with the following requirements:

- The Campus Network and Management Network are hosted by the collection of Utility Switches in the System.
- The Provisioning Network and Service Network for the HDP Cluster are hosted by the collection of Utility Switches in the System.
- The Data Network is hosted by the collection of Data Switches in the System.

The Utility Switches are typically 1GbE Switches, and the Data Switches are typically 10Gb or better Switches. The Data Switches may be Ethernet or InfiniBand, but Ethernet is assumed for the remainder of this document.

Refer to Figure 13 for a more detailed depiction of the Network Subsystem including the Provisioning Network and Service Network internal to the HDP Cluster as well as an example of the internal private network details for an ESS.

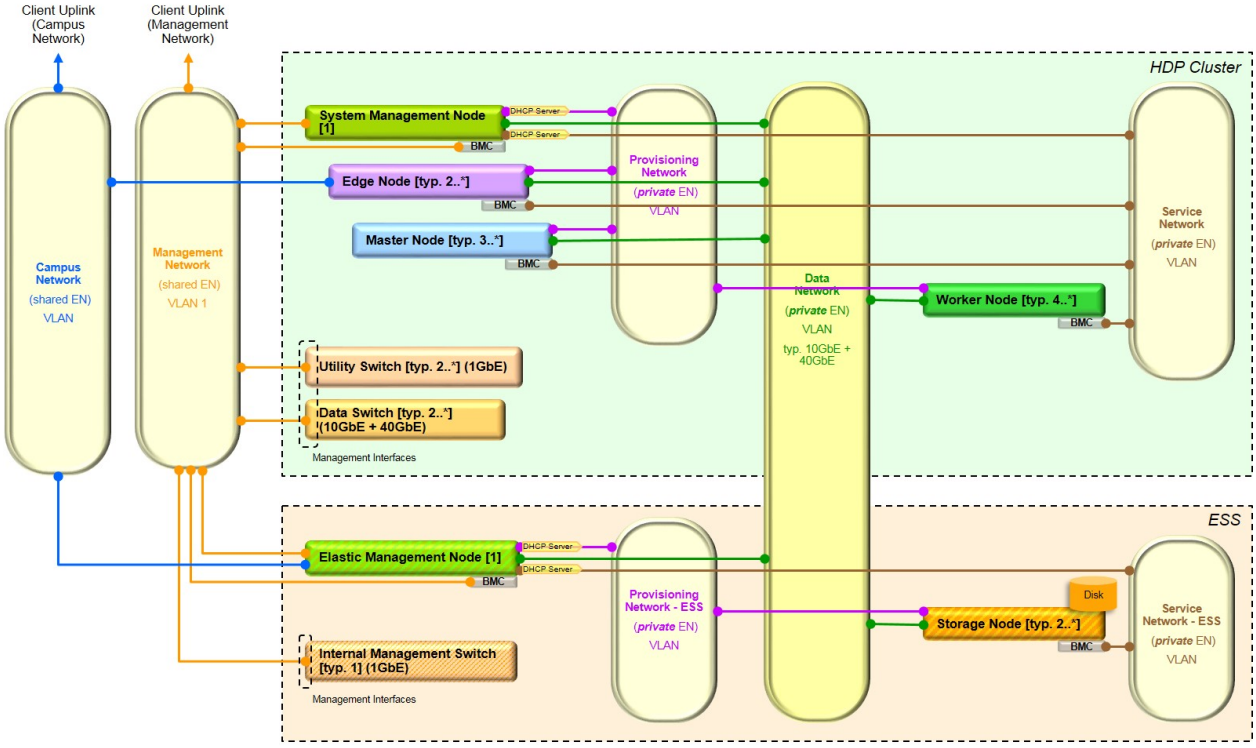

Figure 13. Network Subsystem – Logical Networks – Detail

4.4.5 Racking and Physical Layout

This architecture includes racking guidelines which assist with a more organized deployment. These racking guidelines should not be considered as strict architectural requirements, but as recommendations for designs which can facilitate the physical layout of a System and which allow more orderly management and extension of a System.

These racking recommendations do not specifically cover the following, which should also be incorporated into any design to ensure the relevant requirements are met.

- PDU (Power Distribution Unit) and power capacity calculations
- Distribution of specific Nodes or Node types to satisfy any particular capacity or resilience requirement (for example, tolerance for an outage of an entire rack)
- Inter-rack cabling details – especially cable lengths and technology choices to accommodate distance requirements

The following logical groups of physical elements (servers, I/O drawers, and switches) which are to be rack installed are defined. These groups are defined independently from the racking to more conveniently describe the racking recommendations. The base recommendations are based upon using 2U servers for the Worker Nodes. A variation of these recommendations is included later based upon using 1U servers for the Worker Nodes.

4.4.5.1 Primary Server Group

A Primary Server Group consists of the following:

- A set of HDP Nodes – typically 1-18 Nodes of any type which form the HDP Cluster.
- A set of Utility Switches that serve the Group – typically one or two Utility Switches.
- A set of Data Switches – typically two Data Switches – that serve the Primary Server Group plus potentially one Secondary Server Group.

4.4.5.2 Secondary Server Group

A Secondary Server Group consists of the following:

- A set of HDP Nodes – typically 1-18 Nodes of any type which form the HDP Cluster.
- A set of Utility Switches that serve the group – typically one or two Utility Switches.

The content of a Secondary Server Group is the same as the content of a Primary Server Group minus the Data Switches. A Secondary Server Group is always associated with a Primary Server Group. The association is based upon the fact that the Servers in the Secondary Server Group use (are cabled to) the Data Switches contained in the associated Primary Server Group.

4.4.5.3 ESS Group

An ESS Group consists of all of the elements which form an ESS. This includes:

- All Storage Nodes
- All Storage Drawers
- The Elastic Management Server
- All ESS Internal Switches

An ESS Group excludes:

- Any Utility Switches
- Any Data Switches
- Any Nodes that are part of the HDP Cluster

Note:

- An ESS Group may include multiple ESS *building blocks* forming a single ESS.
- A System may include more than one ESS Group.
- An ESS Group may have more elements than can be contained within a single rack.

The smallest System consists of one ESS Group and one Primary Server Group. As a larger System is defined (or as a System grows over time), additional ESS Groups may be added and additional Server Groups may be added. ESS Groups and Server Groups may be added independently as needed to scale the storage and compute capacity independently. Server Groups are included in an alternating fashion as follows: first a Primary Server Group then a Secondary Server Group (which is associated with the previously added Primary Server Group).

Using the Groups defined above, following are the racking guidelines:

4.4.5.4 Independent Rack Approach

This first and recommended approach racks the ESS Groups and the Server Groups independently and does not mixed ESS and Server Groups within any rack. This is natural and most convenient as the native rack type for the ESS is different than the native rack model used for the typical POWER9 servers used for these designs.

4.4.5.4.1 Rack Assignments

The following rack assignments are recommended:

- ESS Groups:
- Place the first ESS Group into ESS Rack 1.
- For each additional ESS Group, if the group fits within the available space within an ESS Rack, include it in that ESS Rack. Otherwise, add another ESS Rack to hold the ESS Group.
- Server Groups:
- Place the first Primary Server Group into Server Rack 1.
- Place the associated Secondary Server Group into Server Rack 2.
- Place the next Primary Server Group into Server Rack 3.
- Place the next associated Secondary Server Group into Server Rack 4.
- And so on

4.4.5.4.2 Switch Associations and Cabling

Each Server Rack includes the Utility Switches that serve that rack. All of the HDP Nodes within the rack are cabled to the Utility Switches within the same rack.

Each Server Rack that contains a Primary Server Group includes the Data Switches that serve that Primary Server Group and potentially one associated Secondary Server Group. All of the HDP Nodes within the Primary Server Group are cabled to the Data Switches that are within the same rack. All of the HDP Nodes within the Secondary Server Group are cabled to the Data Switches in the rack containing the Primary Server Group.

Each ESS Group is internally connected and cabled to its internal switches consistent with the design of the particular ESS model. The Elastic Management Server in each ESS Group is uplinked to at least one of the Utility Switches contained in the first Server Rack. This uplink carries Campus and Management traffic to and from the ESS. The Storage Nodes within each ESS Group are uplinked to one of the Data Switch pairs racked in one of the Server Racks.

When there is more than one pair of Utility Switches (that is, more than one Server Rack), these pairs of Utility Switches must be interconnected. For this architecture, the Utility Switch interconnections are handled within the existing Switch layers – for example, an aggregation or distribution switch layer is not added. For larger Systems, a more complex network design is often required, and such a design likely includes aggregation layer or spine switches for the Utility Switches. Custom design assistance is recommended for such cases.

When there is more than one pair of Data Switches, these pairs of Data Switches must also be interconnected. The interconnection between pairs of Data Switches requires an extended network design that typically includes aggregation layer or spine switches. These are not covered in this architecture. Custom design assistance is recommended for such cases

See Figure 14 for an example of Groups and racking consistent with the above. Other details related to the racking and cabling specifics are considered as design-level choices within this document.

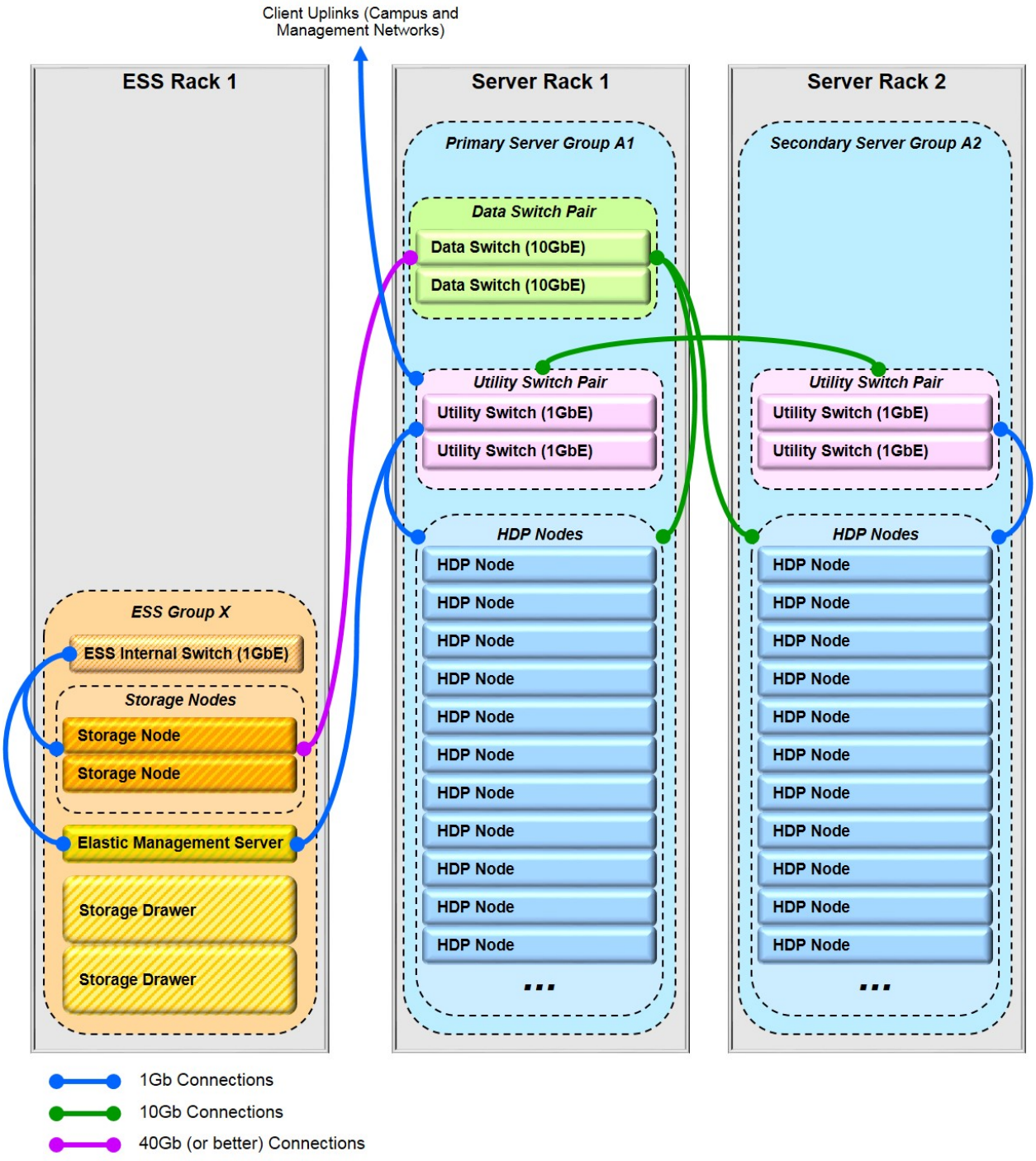

Figure 14. Group and Rack Assignments - Three Rack Example

4.4.5.5 Independent Rack Approach - 1U Worker Node Variation

When using 1U servers for the Worker Nodes, greater rack density for the servers can be achieved. Most of the above concepts apply, modified as follows:

- Only ESS Groups and Primary Server Groups are used. Secondary Server Groups are not used.
- A Primary Server Group may typically contain up to thirty (30) servers.
- The Data Switches for each Primary Server Group serve only that Server Group.
- Rack assignments and switch associations and cabling are as noted above except that every Server Group added is a Primary Server Group (that is, alternating between Primary and Secondary Server Groups is not required). Thus, every HDP Node is cabled to the Utility Switches and Data Switches within its rack.

4.4.5.6 Shared Rack Approach

The Groups defined above may be assigned to racks differently to meet other requirements. Perhaps the most relevant special case is the sharing of a single rack by an ESS Group and a Primary Server Group. Specifically, if the size of an ESS Group is sufficiently small, and the size of a Primary Server Group is also sufficiently small, these two Groups may be installed within a single rack. Note that it may be necessary to obtain alternative rack mounting hardware if the particular rack chosen does not match the native rack mounting expectations for a component. Otherwise such a racking is acceptable, and the System may still be extended in both server and storage capacity as described in the previous section. See Figure 15 for an example.

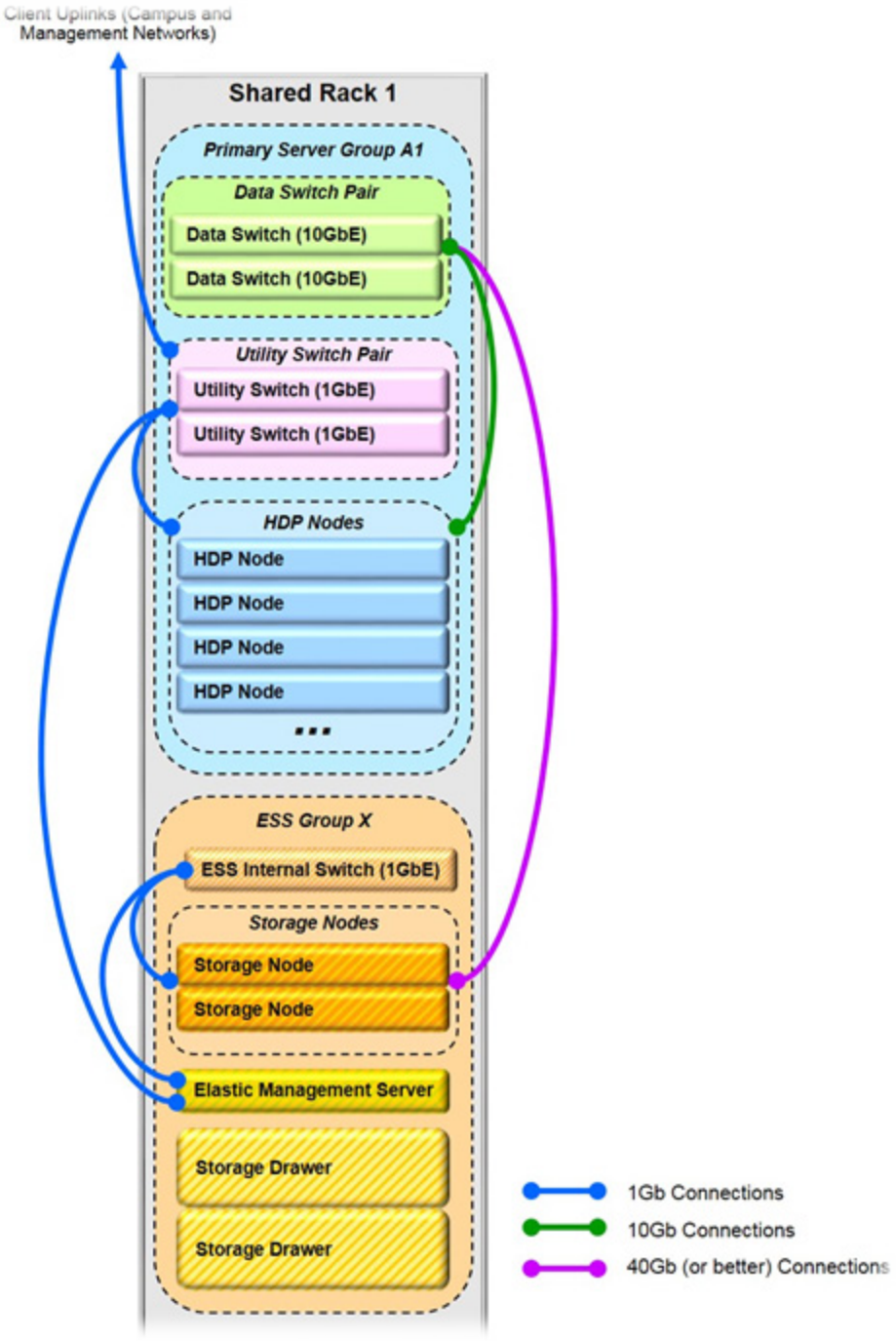

Figure 15. Group and Rack Assignments – Shared Rack Approach – Single Rack Example

4.5 Operations

4.5.1 Data Ingest and Data Sharing

Within this architecture, Data ingest and Data sharing (that is, sharing outside of the Cluster) may be handled as is commonly done with a traditional HDP/Hadoop style cluster. Specifically, Functions hosted on the Edge Nodes, using HDFS, serve as the portal through which Data is ingested into the Data Store and the path through which external elements access Data contained within the Data Store. However, Spectrum Scale and ESS enable additional Data ingest and Data sharing paths and mechanisms. With Spectrum Scale installed on each HDP Node, Functions running on these Nodes can bypass the HDFS abstraction and access the Data on the ESS using any other mechanism provided by Spectrum Scale – for example, POSIX-compliant APIs or the command line, native Spectrum Scale operations. With ESS, access to the Data can be achieved through means and paths completely outside of the HDP Cluster – for example, protocol nodes or other Spectrum Scale clients outside of the HDP Cluster.

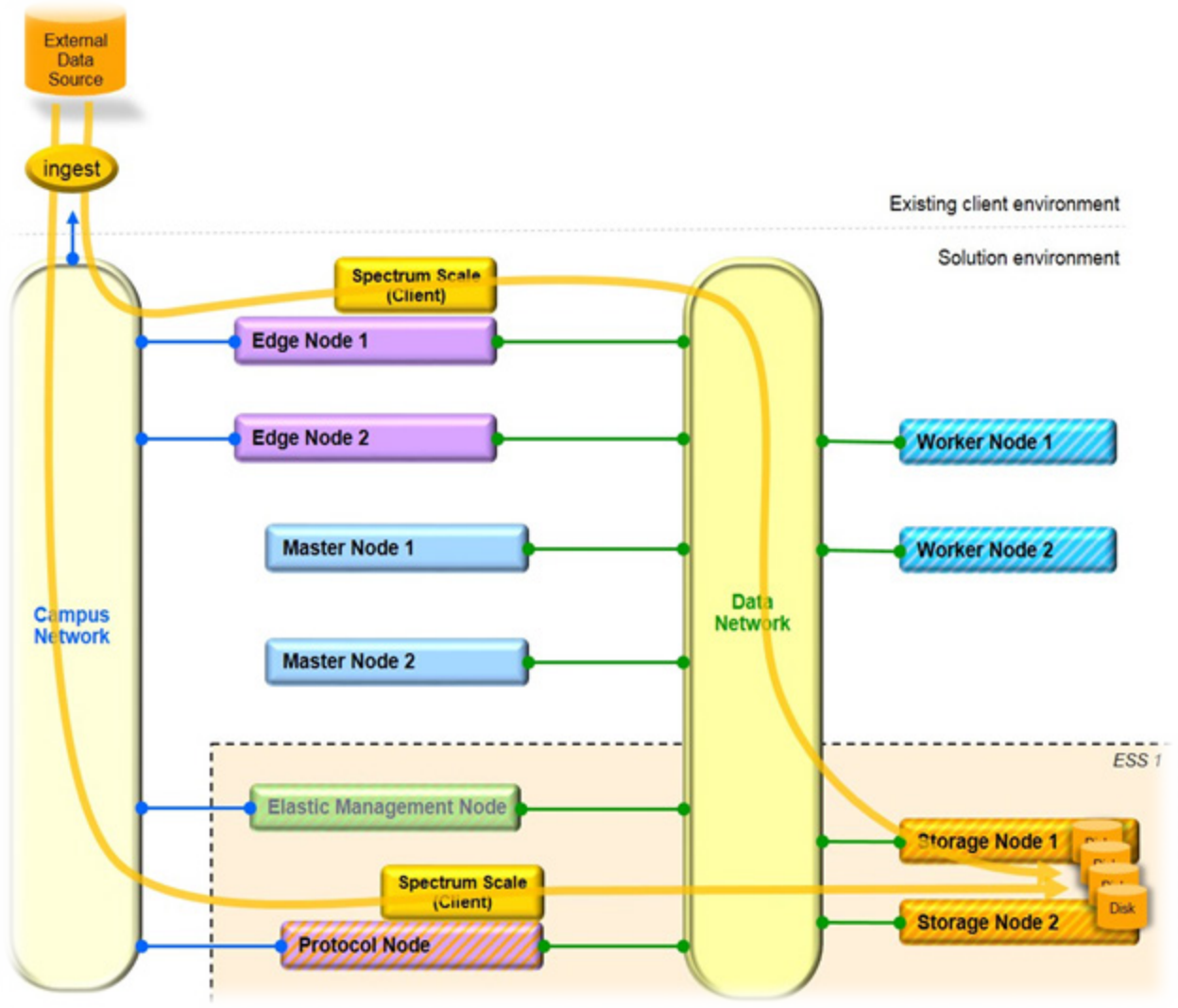

Figure 16. Alternate Data Ingest Paths with Spectrum Scale and ESS

4.5.2 Back-up and Replication

The gathering of the Data under the ESS provides the opportunity to back-up and replicate the Data using any mechanisms available with the ESS – for example, Spectrum Scale replication, Spectrum Protect, or Spectrum Archive EE. These are often richer and more capable mechanisms than those available through the HDP environment alone.

To back-up or replicate an entire HDP environment, note that not all data of interest may be contained in the Data Store (that is, on the ESS). Depending upon how they are configured, some data or meta-data (for example, for databases) may be stored locally on the Master Nodes. Properly including this data in a complete backup may require additional steps beyond replicating the ESS storage contents. This issue can be mitigated in some cases by choosing to place (host) such data or metadata on the ESS versus locally on an HDP Node.

4.6 Sizing

A relevant and significant part of the design process includes sizing various portions and components of the System. Full treatment of the sizing topic and all of the various factors that may influence the sizing is beyond the scope of this reference architecture but following are some points of guidance.

4.6.1 Storage and Compute Capacity

A primary feature of this architecture, compared with the traditional Hadoop distributed storage model, is the decoupling of the storage and processing capacity such that each can be sized independently from the other.

4.6.1.1 Storage Capacity

Storage capacity is provided by the ESS serving as the Data Store in this architecture. Obtaining the desired storage capacity is mostly a matter of choosing an appropriate ESS model and appropriate features within that model (especially disk drive sizes). The ESS is modular component within this architecture, and all models and configurations of the ESS, including flash based models, are available and suitable to include in a System. Many resources are available to guide the designer in this area (for example, [10]), and the reader is encouraged to use those as primary resources. The following guidelines are offered as general summary guidance appropriate to this solution.

4.6.1.1.1 Multiple ESSs versus a Single ESS

In most cases where this solution is deployed, it is most convenient to choose an ESS model and configuration that provides all of the storage capacity under a single ESS. This is typically most cost effective and most convenient to manage.

Multiple ESSs may be used for the Data Store, and it may be necessary in the case where capacity is later added to a System. However, this creates multiple elements which must be managed.

Multiple ESSs are typically most applicable for the following:

- A separate or remote site is used for back-up or recovery of the ESS within this solution.
- Heterogeneous storage is needed. For example, an ESS all flash model for high performance and an ESS model with HDDs for high capacity.

4.6.1.1.2 Growth

The ESS capacity may be expanded through any means supported by the ESS generally. This may include adding one or more building blocks to an ESS in the System or by adding another ESS to the System.

4.6.1.2 Compute Capacity

Compute capacity is primarily provided by the Worker Nodes in the HDP Cluster. Compute capacity is generally chosen to achieve certain performance criteria – for example, throughput and response time. Obtaining the desired compute capacity is generally a matter of choosing a suitable Worker Node configuration and choosing a number of Worker Nodes that provides satisfactory performance. Choosing an appropriate compute capacity is often more difficult than choosing the storage capacity of the System as the nature of the specific workload to be used can significantly affect the compute performance. Benchmarking, using the expected production workload and Data or ones very similar, is often the most effective method for estimating an appropriate compute capacity.

4.6.2 Bandwidth

4.6.2.1 Data Pipeline

An important bandwidth consideration relates to the data access path between the HDP Nodes (especially the Worker Nodes) and the ESS over the Data Network. Several elements are relevant to this path and provide parameters and components that must be chosen appropriately. This data access path can be considered as a "pipeline" with various stages for which the various parameters may be chosen to best satisfy the particular client requirements. The following stages represent a useful factoring of the pipeline.

4.6.2.1.1 Client Demand

"Client demand" represents the aggregate demand of the HDP Nodes. For most Systems, this value is dominated by the Worker Node demand, but the demand from the other HDP Node types should also be included. This value is calculated from input assumptions based upon the nature of the workload.

4.6.2.1.2 Client Network Interface

"Client network interface" represents the aggregate network interface capacity of the HDP Nodes. This value is calculated based upon the network adapter configurations chosen for the HDP Nodes.

4.6.2.1.3 Data Network Infrastructure

"Data Network infrastructure" represents the total bandwidth available between the HDP Nodes and the ESS. For a single set of Data Switches, this represents the total bandwidth of the Switches (as published by the switch vendor). For more complex Switch topologies, the available bandwidth may need to be adjusted to represent uplink or crosslink capacities which may reduce the total end-to-end bandwidth.

4.6.2.1.4 ESS Network Interface

"ESS network interface" represents the aggregate network interface capacity of the Storage Nodes on the ESS. This value is calculated from a) the number and type of network adapters selected for the ESS configuration, and b) the number of links that are connected from each adapter to the Data Switches.

4.6.2.1.5 ESS Supply

"ESS supply" represents the aggregate, maximum data serving capacity/performance for the particular ESS selected – independent from the network interfaces. This value may be calculated or obtained from various ESS resources such as [11] and [10].

Figure 17 depicts the various stages combined into a generic Data pipeline view.

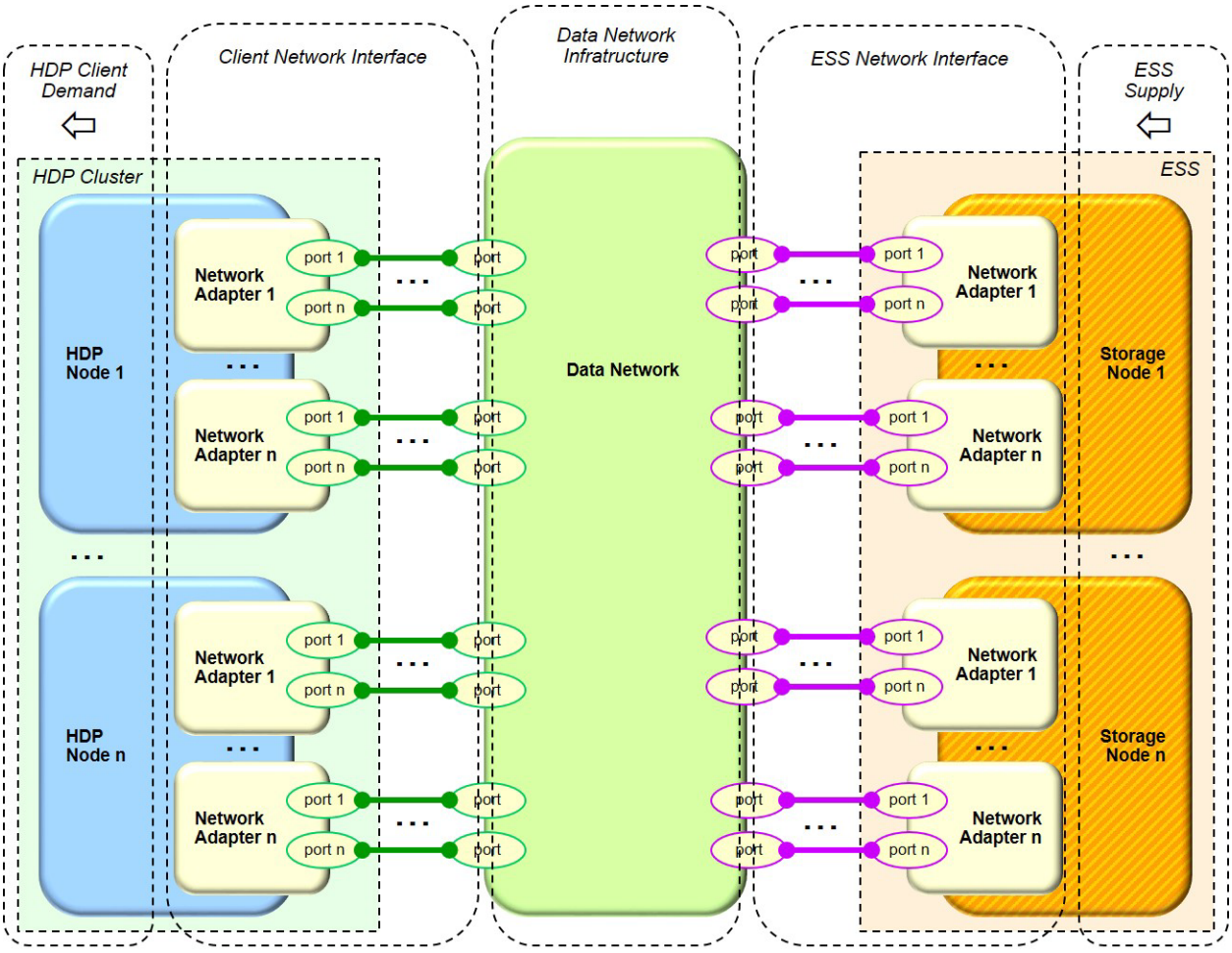

Figure 17. Data Pipeline - Generic view

The end-to-end data rate and pipeline utilization is effectively limited by the smallest data rate for any stage in the pipeline (noting that some inter-stage thresholding may need to be applied before summing the across the paths in a given stage). For any given System design, it is generally recommended that the ESS network interface be designed such that it is not the limiting stage for the ESS end of the pipeline (that is, the capacity of the ESS network interface is recommended to meet or exceed the maximum data serving rate of the ESS). Similarly, it is generally recommended that the client network interface be designed such that it is not the limiting stage for the client end of the pipeline (that is, the capacity of the client network interface is recommended to meet or exceed the maximum demand expected by the client Nodes). That said, which stage is acceptable as the limiting stage is largely a design choice.

Examples of Data pipeline calculations are included with each of the reference designs in this reference architecture.

5 Reference Design 2.1A – 18 Node Base 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 may be considered to be a base configuration with a single ESS Group and a single Primary Server Group. It may be considered as a useful starting point for a base production configuration.

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.

5.1 HDP Node Configurations

5.1.1 Hardware Configurations

This design selects a "Balanced" Cluster Type. The specific Node configurations for each Node type for this design are listed in Figure 18. It includes configuration parameters for another Cluster Type ("Performance") which is referenced later. The "Server Dense" Cluster Type is listed for comparison only and not directly applicable to this reference design.

	System Mgmt Node	Master Node	Edge Node	Worker Node		
Cluster Type	*All*	*All*	*All*	*Balanced*	*Performance*	*Server Dense*
Server Model	1U LC921	1U LC921	1U LC921	2U LC922	2U LC922	1U LC921
# 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	40
Memory	32GB	256GB	256GB	256GB	512GB	256GB
Storage - HDD (front)	2x 4TB HDD	4x 4TB HDD	4x 4TB HDD	4x 4TB HDD		4x 4TB HDD
Storage - SSD (front)					4x 3.8TB SSD	
Storage - HDD (rear for OS)				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 18.1 LC922/LC921 Hardware Configuration for HDP

	System Mgmt Node	Master Node	Edge Node	Worker Node
Cluster Type	*All*	*All*	*All*	***Storage Dense – ESS***
Server Model	2U IC922	2U IC922	2U IC922	2U IC922
# Servers (Min/Default/Max)	1 / 1 / 1	3 / 3 / Any	1 / 1 / Any	4 / 8 / Any
Sockets	1	2	2	2
Cores (total)	12	40	40	40
Memory	32GB	256GB	256GB	256GB
Storage Backplane (Front)	1	1	1	3
Storage - HDD (front) **Storage - SSD (front)**	2x 2.4TB HDD	4x 2.4TB HDD	4x 2.4TB HDD	6x 2.4TB HDD
OS Storage - HDD (front)				2x 2.4TB HDD
Storage Controller	1x Broadcom 9300-8i	1x Broadcom MegaRAID 9361-8i 1x Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1x Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1x Broadcom 9305-16i
Network* - 1 GbE	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 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)
Network - 10 GbE**	1x 2-port (2 ports)	1x 2-port (2 ports)	2x 2-port (4 ports)	1x 2-port (2 ports)
Cables - 10 GbE**	2 cables (DACs)	2 cables (DACs)	4 cables (DACs)	2 cables (DACs)
Operating System	RHEL 7.6 for P9	RHEL 7.6 for P9	RHEL 7.6 for P9	RHEL 7.6 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 18.2 IC922Hardware Configuration for HDP

As an alternative to the IBM Power System LC921 and Power LC922 server models, the Power L922 (9008-22L) or IC922 (9183-22X) server can be used for the nodes instead. See Appendix B for Power L922 and IC922 specific considerations.

5.1.2 Node Counts

5.1.2.1 Worker Nodes

Twelve (12) Worker Nodes are specified for this design. Twelve Worker Nodes provide significant processing capacity while allowing the HDP Cluster to be contained within a single rack.

5.1.2.2 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.

5.1.2.3 Edge Nodes

Two (2) Edge Node is specified for this design. This selection represents a basic HA configuration for Edge Functions.

5.1.2.4 System Management Nodes

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

5.2 ESS Configuration

This design specifies one (1) IBM GL2S Elastic Storage Server to serve as the Data Store. The following configuration selections are made for the ESS:

- 4TB HDDs (664TB raw disk capacity; ~443TB usable disk capacity)
- Two (2) 10, 56 or 100GbE network adapters per ESS Storage Node, one port per adapter used
- One (1) internal management switch, 1 GbE (MTM=8831-S52, Mellanox AS4610)

This ESS model is BMC based and does not include an HMC.

5.3 Software

For this reference design, the following software is specified. (Note that some other software versions are compatible with the architecture. See also [5] and [6]).

5.3.1 Operating System Software

RHEL 7.6 for POWER9 is specified as the operating system software for all HDP Nodes.

5.3.2 Platform Software

Cloudera HDP version 3.1.5 is specified as the Platform software for this design.

5.3.3 Spectrum Scale Components

The following Spectrum Scale components for the HDP Cluster are specified for this design:

- IBM Spectrum Scale 5.0.1.1 – Standard Edition
- IBM Spectrum Scale Transparency Connector 3.0.0-0
- IBM Spectrum Scale Ambari Management Pack 2.7.0.0

These are the minimum versions required. These are used to install required Spectrum Scale modules on HDP nodes. IBM Spectrum Scale software in IBM Elastic Storage Server (ESS) comes pre-installed. No additional Spectrum Scale licenses are required beyond those included with ESS for this reference architecture.

The Spectrum Scale software for the ESS is separate from the above and included with the ESS configuration.

5.4 Network Subsystem

This design specifies a logical network design that follows the architecture guidelines. Specifically, 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 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.

5.4.1 Logical Networks – HDP Cluster

5.4.1.1 Data Network

The Data Network is private (within this system) and assigned to VLAN 77. The servers in the system present untagged traffic to the switches for this network. This network is hosted by the Data Switches.

5.4.1.2 Campus Network

The Campus Network is shared (outside of the System) and assigned to VLAN 22. This network is hosted by the Utility Switches and uplinked into the existing client network infrastructure. The servers in the system present *tagged* traffic to the switches for this network.

5.4.1.3 Management Network

The Management Network is shared (outside of this system) and assigned to VLAN 1. This network is hosted by the Utility Switches and uplinked into the existing client network infrastructure. The servers in the system present *tagged* traffic to the switches for this network. This network also carries management traffic for the management interfaces for all of the Switches in the System.

5.4.1.4 Provisioning Network

The Provisioning Network is private (within this system) and assigned to VLAN 88. This network is hosted by the Utility Switches. The servers in the system present *untagged* traffic to the switches for this network. Configuring for untagged traffic more conveniently supports NetBoot, which is used to provision the Nodes in the HDP Cluster.

5.4.1.5 Service Network

The Service Network is private (within this system) and assigned to VLAN 188. This network is hosted by the Utility Switches. 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 HDP Nodes during provisioning.

5.4.2 Logical Networks – ESS

The Data Network, Campus Network, and Management common to the HDP Cluster and the ESS. Three is one instance of each of these networks, and they are described as part of the HDP Cluster above. The ESS has external (to the ESS) connections to these three networks.

The ESS also contains private internal networks to accomplish its installation and configuration and infrastructure level management. These are not visible outside of the ESS, and they are described here only to provide a complete network design for the System – consistent with the specific ESS model selected.

5.4.2.1 Provisioning Network – ESS

The Provisioning Network – ESS is private (within the ESS) and assigned to VLAN 91. This network is hosted by the ESS Internal Management Switch. The servers in the system present *untagged* traffic to the switches for this network. Configuring for untagged traffic more conveniently supports NetBoot, which is used to provision the Nodes in the HDP Cluster.

5.4.2.2 Service Network - ESS

The Service Network – ESS is private (within the ESS) and assigned to VLAN 191. This network is hosted by the ESS Internal Management Switch. The BMCs in the ESS present untagged traffic to this Switch for this network. The BMC-to-switch connections are dedicated to this function. The Elastic Management Server also has an OS level connection to this network to accomplish power control of the ESS Nodes.

Figure 19 and Figure 20 depict the logical network topology for this reference design. Figure 20 excludes the Provisioning Networks and the Service Networks from the diagram -- allowing a simpler and cleaner rendering that better illustrates the connectivity to the shared networks.

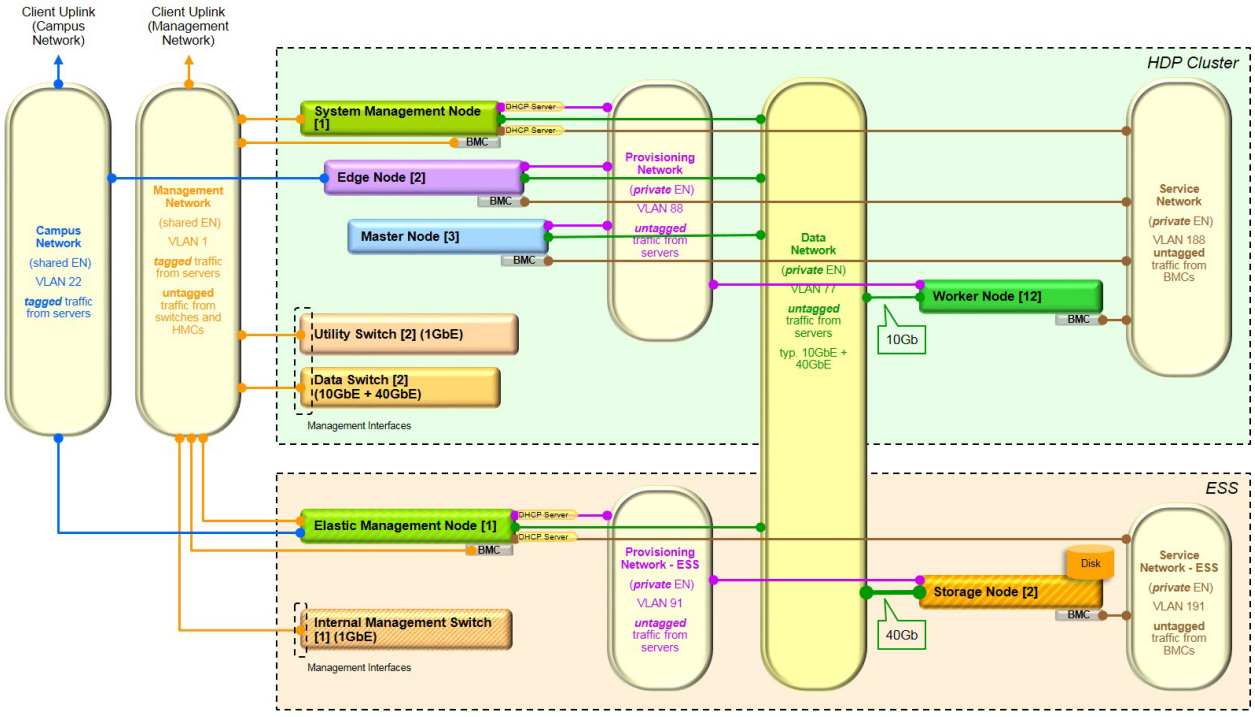

Figure 19. Network Design - Logical View - All Networks

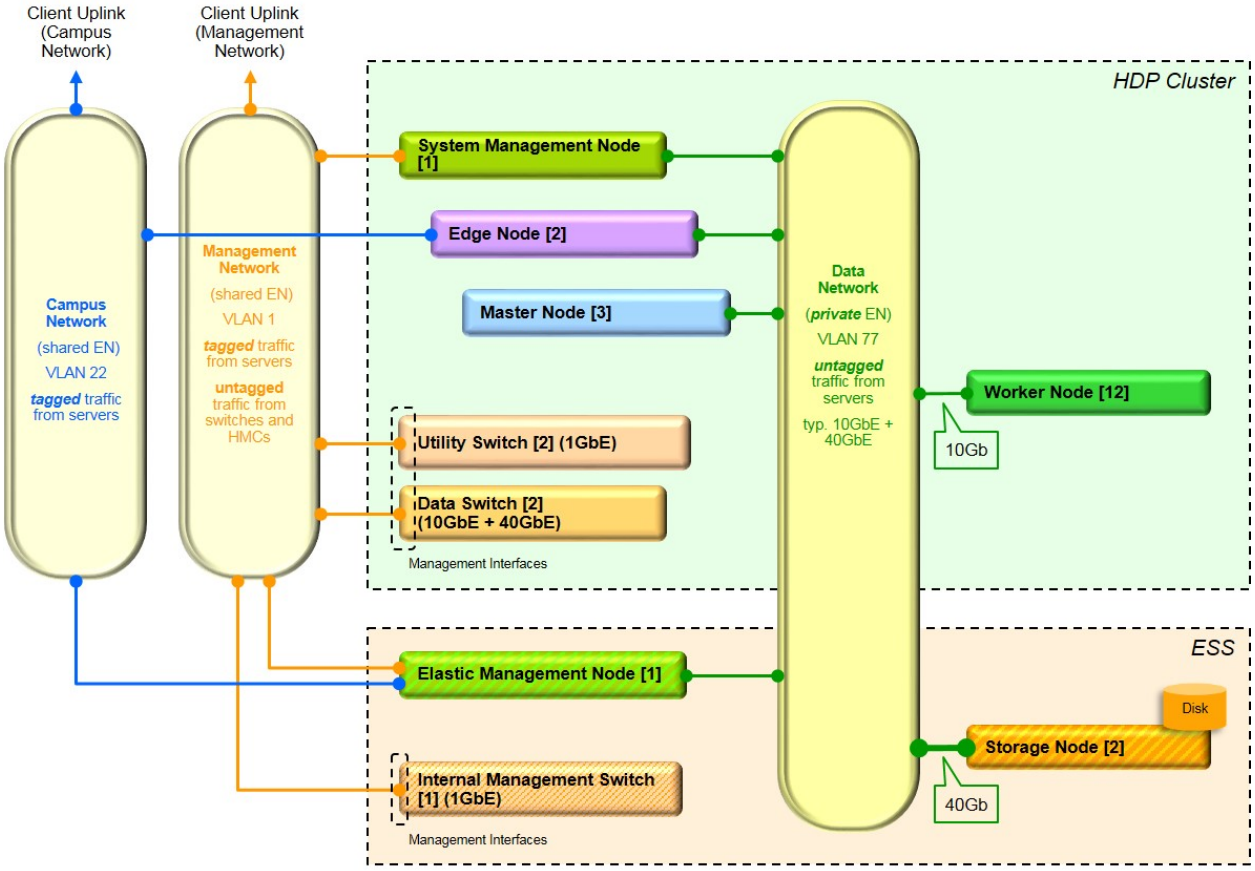

Figure 20. Network Design – Logical View – Data, Campus, and Management Networks

5.4.3 Switches

5.4.3.1 Utility Switches

Two (2) IBM 8831-S52 (Mellanox AS4610) switches are specified for the Utility Switches in this design. These are configured as a redundant pair with MLAG (multi-chassis aggregation link).

5.4.3.2 Data Switches

Two (2) IBM 8831-25M (Mellanox SN2410) switches are specified for the Data Switches in this design. These are configured as a redundant pair with MLAG.

5.4.3.3 ESS Internal Management Switch

One (1) IBM 8831-S52 (Mellanox AS4610) switch is specified as part of the ESS internal configuration.

5.4.4 Cabling

The physical cabling for each Server in the HDP Cluster follows a consistent pattern, and the switch-side port configurations for each HDP Node is typically the same. This provides consistency and reduces the opportunity for error. It also provides flexibility for special situations that might arise. Using 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.

5.4.4.1 Utility Switches (1Gb)

The connection for each HDP Node to 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 Utility 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. 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 1Gb traffic that requires tagging (VLANs 1, 22, and 177).

The Utility Switches also host the Service Network. The Service Network is different than the other 1Gb 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 (Utility 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.

The Utility Switch pair is configured and cabled with IPLs (Inter Peek Links) for MLAG, which allows the links to be aggregated across the pair of Switches.

See Figure 21 for a diagram of the cabling design for the 1Gb Networks.

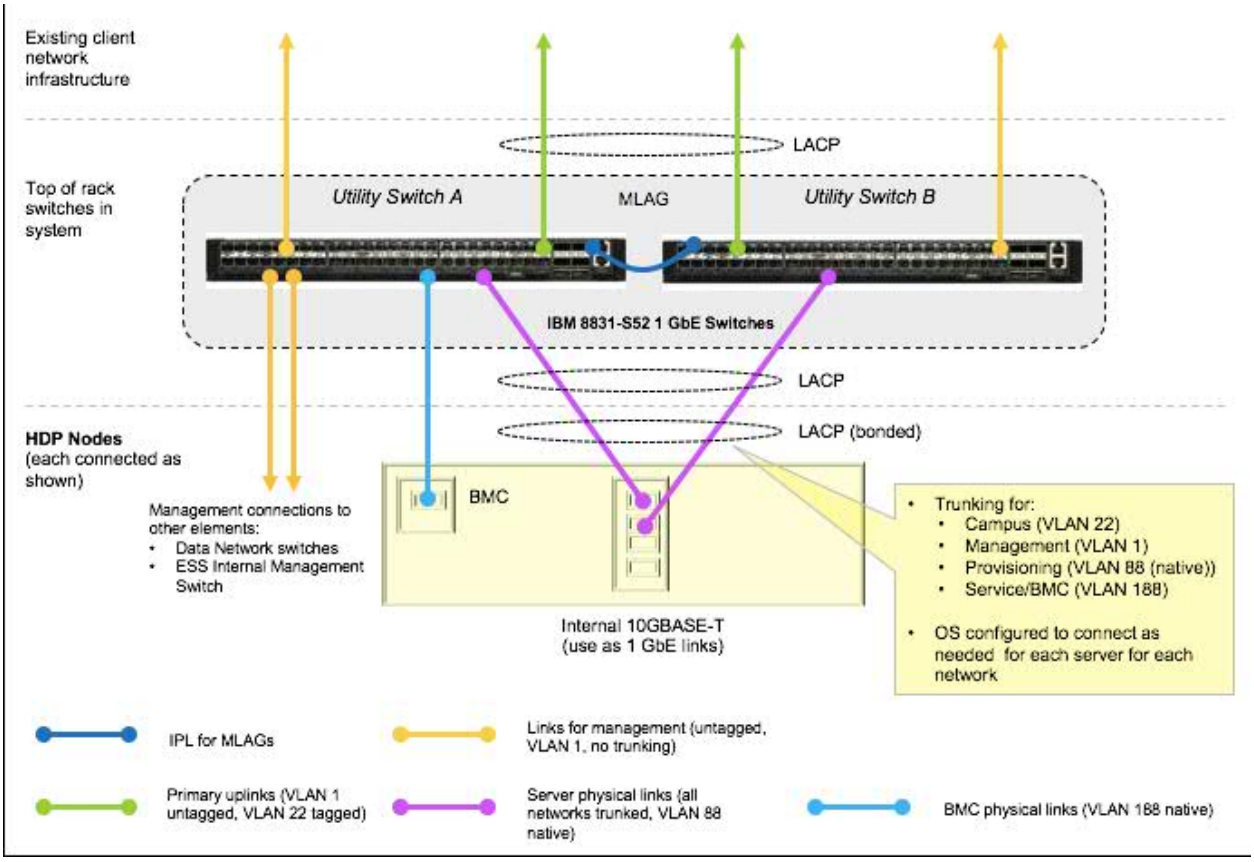

Figure 21. Utility Switch Cabling - Physical Schematic View

5.4.4.2 Data Switches (25 Gb and 100 Gb Links)

The connection between each server and the switches for the Data Network is carried over two physical links (cables) to the Data Switches. This provides a redundant path that is used to provide resilience for these networks, as well as increased bandwidth (up to 50 Gb) for each HDP 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 1Gb links, this pair of links is configured for link aggregation using LACP on the server and on the switch. The Data Switch pair is similarly configured (and cabled with an IPL) for MLAG, which allows the links to be aggregated across the pair of switches. See Figure 22 for a diagram of the cabling design to the Data Switches

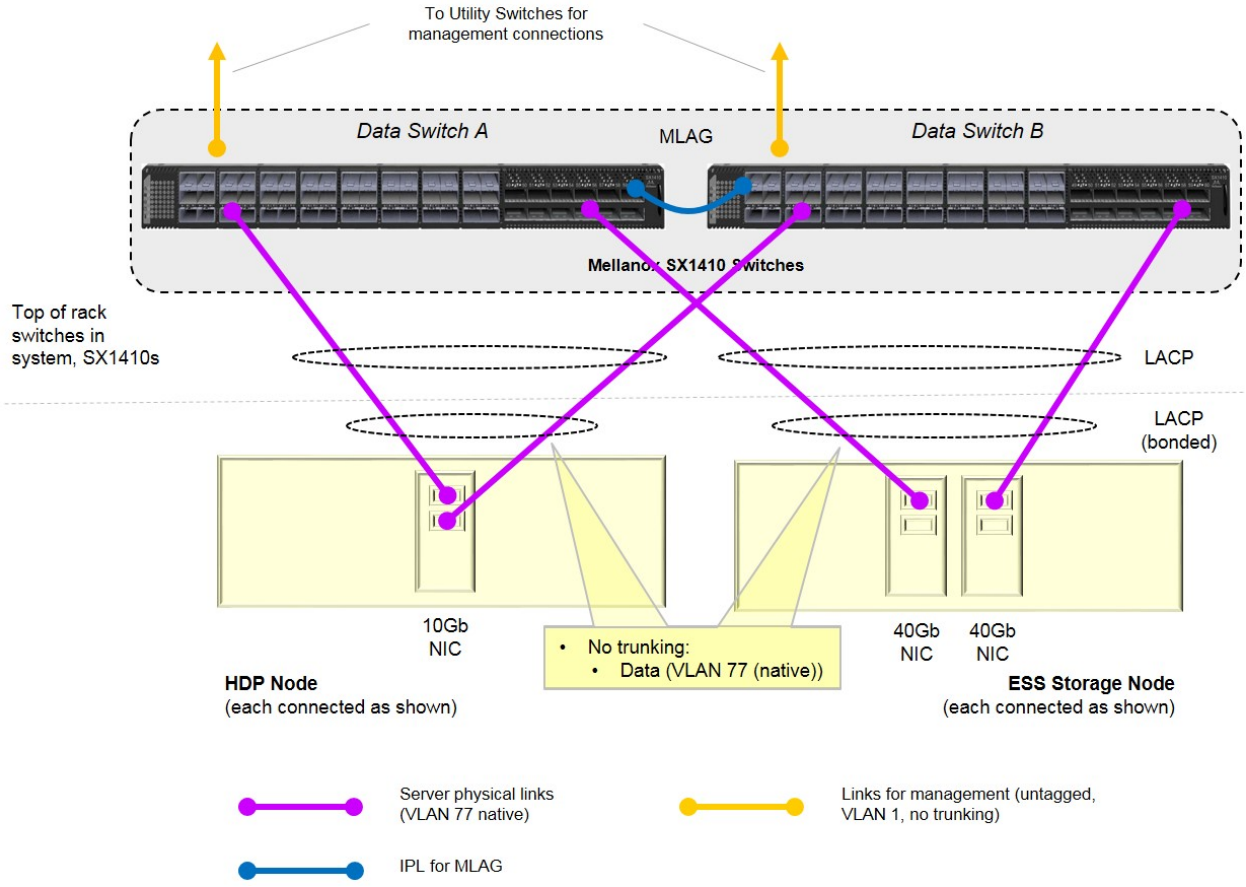

Figure 22. Data Switch Cabling - Physical Schematic View

5.4.5 Other Considerations

5.4.5.1 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).

5.4.5.2 Dynamic Host Configuration Protocol (DHCP)

This design provides DHCP for two of the networks in the System. The System Management Node is configured to provide DHCP for the Service Network and the Provisioning Network.

5.5 Data Pipeline Calculations

An overview of the data pipeline and the primary stages is depicted in Figure 23.

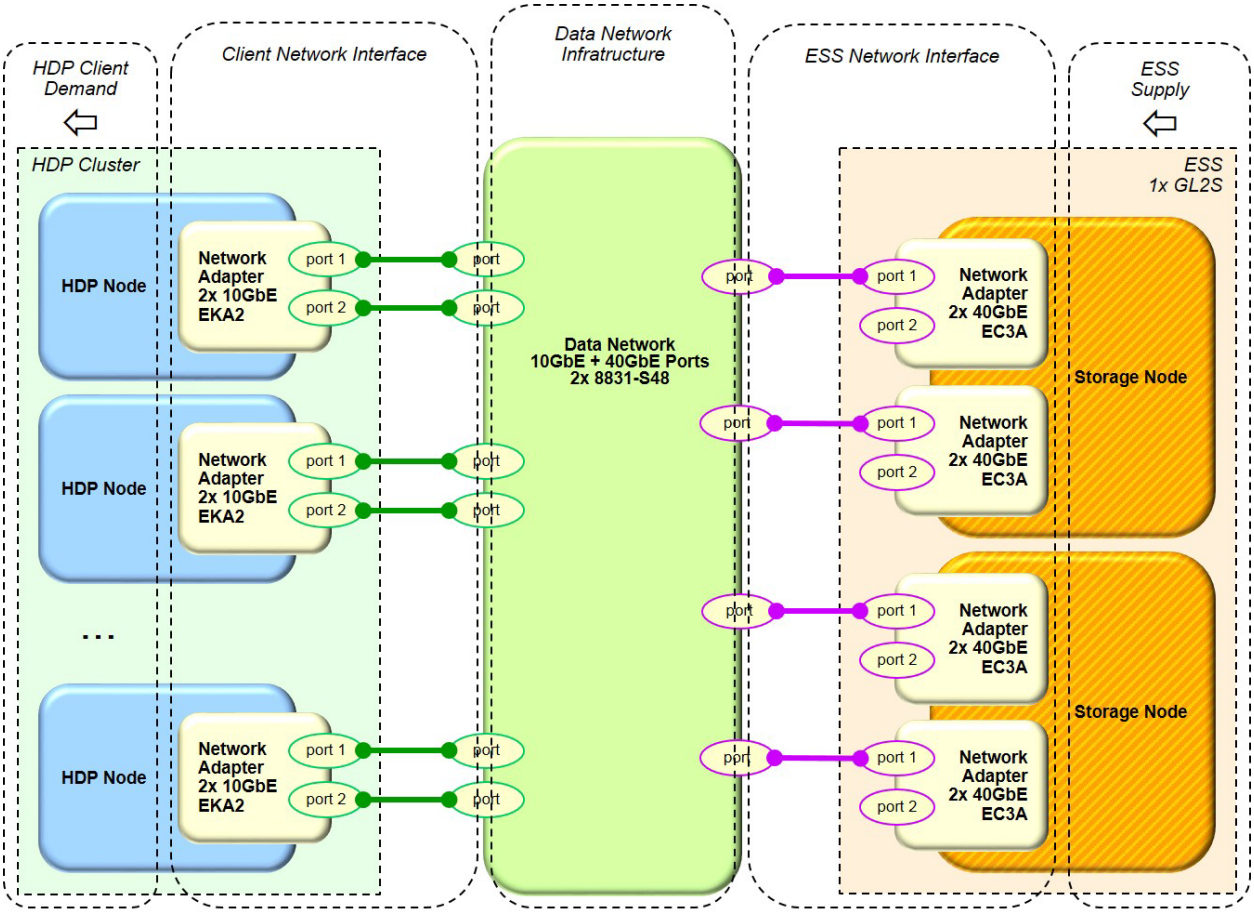

Figure 23. Data Pipeline Overview

The following calculations apply to the relevant stages of the data pipeline in this design.

5.5.1 Client Demand

The total client demand is estimate to be 6.4 GBps – all from HDP Nodes as follows.

5.5.1.1 Worker Nodes

For each Worker Node, the peak demand is estimated to be 440 MBps [1]. With 12 Worker Nodes, the total demand is 12 * 440 MBps = 5.3 GBps.

5.5.1.2 Master Nodes

For each Master Node, the average demand is estimated to be one half of the Worker Node demand or 220 MBps. With three Master Nodes, the total demand is 3 * 220 MBps = 0.7 GBps.

5.5.1.3 Edge Nodes

For each Edge Node, the average demand is estimated to be one half of the Worker Node demand or 220 MBps. With two Edge Nodes, the total demand is 2 * 220 MBps = 0.4 GBps.

5.5.2 Client Network Interface

Each client has the same network interface structure. Specifically, one network adapter with two 25GbE connections. Assuming 7% ethernet overhead, the network interface capacity for one HDP Node is then two connections * 10 Gbps * 0.93 = 5.75 GBps. Thus, the total client network interface capacity is 98.75 GBps.

(The single HDP Node network interface capacity is higher than the demand from any one of the HDP Nodes, so no adjustment is required in this regard.)

5.5.3 Data Network Infrastructure

The Data Network infrastructure for this design consists of two 8831-25M (Mellanox SN2410) switches. The switches are configured as a redundant pair with MLAG, so using Mellanox specifications, the aggregate bandwidth for the switching infrastructure is calculated to be 2 * 1.92 Tbps = 480 GBps.

5.5.4 ESS Network Interface

The ESS in this design includes two Storage Nodes, each configured identically. Specifically, each Storage Node has two network adapters, and each adapter has two 10/56/100 GbE ports. However, the specific adapter selected is a PCIe3 x8 which has a maximum interface bandwidth of 985 MBps/lane or 63 Gbps total. Thus, each adapter can only supply one connection with a full 56Gbps data rate at a time. Given the above, it is a design choice to use one port per adapter and include two adapters in each Storage Node to provide two 56 GbE connections at full, constant data rate. Assuming 7% Ethernet overhead, the ESS network interface capacity is therefore two connections * 56 Gbps * 0.93 * 2 Storage Nodes = 26.04 GBps.

[1] This is an input assumption based upon some common benchmark measurements. For any specific client design, demand estimates that are appropriate for the particular client and workload should be used.

5.5.5 ESS Supply

The rate at which the ESS GL2S can supply data is estimated to be 40 GBps per encloser.

Figure 24 provides a graphical comparison of the above calculations. For the data pipeline, the end-to-end data rate and pipeline utilization is effectively limited by the smallest data rate for any stage in the pipeline. For this design, this is the client demand at 6.4 GBps. It should be noted that the above calculations are generally conservative in that the supply and demand rates are estimated as peak/maximum rates, but with multiple parallel data paths, the peak demand from each client is not likely to occur simultaneously.

For this reference design, it was a goal to provide a pipeline rate that satisfies the maximum estimated client demand rate. Elements such as the ESS model and ESS network interface components and network connections were chosen such that the client demand was not limited by any other element in the Infrastructure. These calculations confirm this design goal is met.

The above calculations also provide information regarding performance headroom and growth headroom. For example, if the workload on the HDP Cluster is changed such that the data demand increases, the Infrastructure has sufficient headroom to accommodate that increase, up to ~ 12 GBps of demand from the HDP Cluster (~88% increase). Alternatively, more Worker Nodes may be added to the HDP Cluster, and the rest of the System has sufficient capacity to serve data to them. For example, an additional twelve (12) Worker Nodes could be added (bringing the total client demand to 11.7 GBps), before other limits in the pipeline (specifically, the ESS supply rate at 12.0 GBps) may start to become a limiting factor.

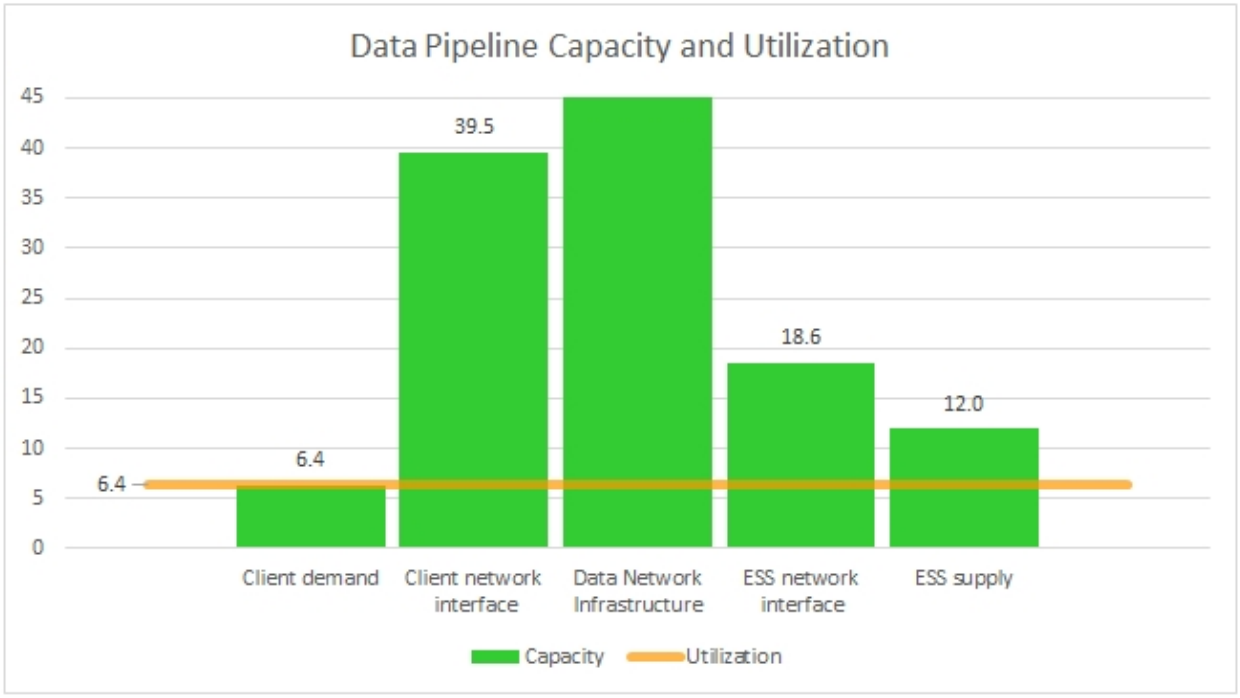

Figure 24. Data Pipeline Capacity and Utilization

5.6 Physical Configuration - Rack Layout

Figure 25 shows the physical layout of the System within its racks. All of the components for this reference design fit within two 42U racks.

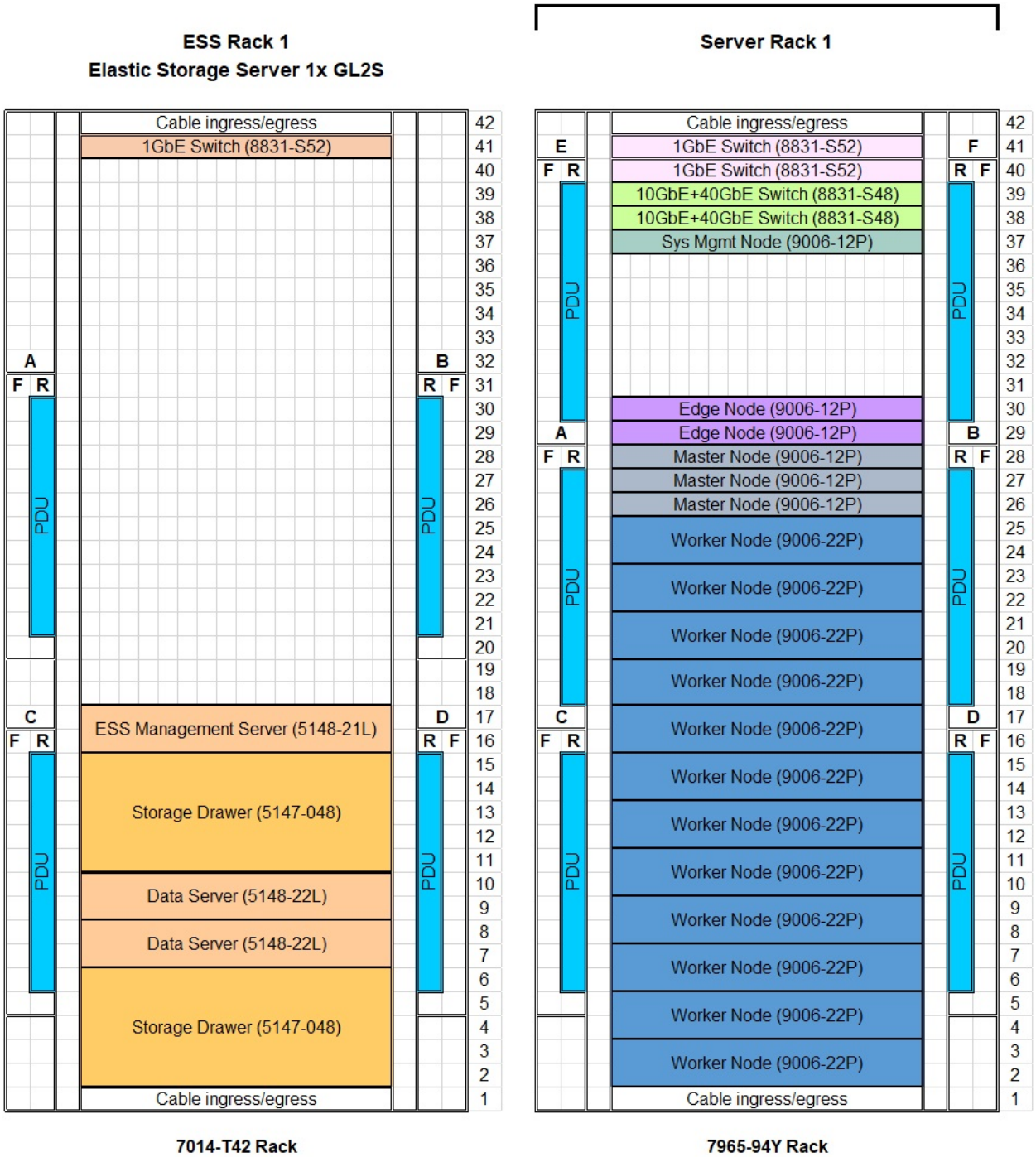

Figure 25. Physical Rack Layout

5.7 Hardware Features for e-config – HDP Cluster

Following are the e-config hardware features for the HDP Cluster in this design – including the Utility Switches and the Data Switches. A full e-configuration includes the system software (for example, RHEL) and services. The HDP software is supplied by Cloudera. The Spectrum Scale software for the HDP Nodes is obtained separately from IBM. 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 based on the processor selection. When configuring Worker Nodes shown in Figure 26, adjust the configuration including the disks to match the requirements.

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	4
		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 26. Worker Nodes (Quantity 12) - 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 27. 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 28. Edge Node (Quantity 2) - 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 29. 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 30. Switches and Server Rack - Hardware Features

5.8 Hardware Features for e-config – ESS

Following are the e-config hardware features for the ESS in this design. A full e-configuration includes the system software (for example, RHEL) and Spectrum Scale software.

Much of the ESS configuration is specified by e-config as part of the ESS solution definition. Items which should be specifically customized or selected are highlighted in the lists below.

In Figure 31, the ESS Management Server must use a network adapter that matches the data network adapter on the ESS Data Server. Thus, in the eConfig for the ESS Management Server, specify Feature Code EC3A instead of EL3X.

Product	Description	Qty
5148-21L	ESS Management Server:5148 Model 21L	1
10	One CSC Billing Unit	10
266	Linux Partition Specify	1
1111	CAT5E Ethernet Cable, 3M BLUE	2
1115	CAT5E Ethernet Cable, 3M GREEN	1
1118	CAT5E Ethernet Cable, 3M YELLOW	1
2147	Primary OS - Linux	1
4651	Rack Indicator, Rack #1	1
5000	Software Preload Required	1
5771	SATA Slimline DVD-RAM Drive	1
6665	Power Cord 2.8m (9.2-ft), Drawer to IBM PDU, (250V/10A)	2
9300	Language Group Specify - US English	1
9442	New Red Hat License Core Counter	10
EC16	Open Power non-virtualized configuration	1
EJTT	Front Bezel for 12-Bay BackPlane	1
EL1A	AC Power Supply - 900W	2
EL3T	Storage Backplane 12 SFF-3 Bays/DVD Bay	1
EL3X	PCIe3 LP 2-port 10GbE NIC&RoCE SFP+ Copper Adapter	1
EL4M	PCIe2 LP 4-port 1GbE Adapter	1
ELAD	One Zero Priced Processor Core Activation for Feature #ELPD	10
ELD5	600GB 10K RPM SAS SFF-3 Disk Drive (Linux)	2
ELPD	10-core 3.42 GHz POWER8 Processor Card	1
EM96	16 GB DDR4 Memory	2
EN03	5m (16.4-ft), 10Gb E'Net Cable SFP+ Act Twinax Copper	2
ESC0	S&H - No Charge	1
ESS0	ESS 5U84 Storage Solution Specify	1

Figure 31. ESS Management Server (Quantity 1) - Hardware Features

Product	Description	Qty
5148-22L	ESS GLxS Data Server:5148 Model 22L	1
10	One CSC Billing Unit	10
266	Linux Partition Specify	1
1111	CAT5E Ethernet Cable, 3M BLUE	1
1115	CAT5E Ethernet Cable, 3M GREEN	1
1118	CAT5E Ethernet Cable, 3M YELLOW	1
2147	Primary OS - Linux	1
4651	Rack Indicator, Rack #1	1
5000	Software Preload Required	1
5771	SATA Slimline DVD-RAM Drive	1
6665	Power Cord 2.8m (9.2-ft), Drawer to IBM PDU, (250V/10A)	2
9300	Language Group Specify - US English	1
9442	New Red Hat License Core Counter	20
EC16	Open Power non-virtualized configuration	1
EC3A	PCIe3 LP 2-Port 40GbE NIC RoCE QSFP+ Adapter	2
ECBP	7m (23.1-ft), IBM Passive QSFP+ to QSFP+ Cable (DAC)	2
ECCT	3M 12GB/s SAS Cable W/Universal Key	4
EGS1	Solution Sub System #1 Indicator	1
EJTQ	Front Bezel for 8-Bay BackPlane	1
EL1B	AC Power Supply - 1400W (200-240 VAC)	2
EL3W	2U SAS RAID 0,5,6,10 Controller + Back plane	1
EL4M	PCIe2 LP 4-port 1GbE Adapter	1
ELAD	One Zero Priced Processor Core Activation for Feature #ELPD	20
ELD5	600GB 10K RPM SAS SFF-3 Disk Drive (Linux)	2
ELPD	10-core 3.42 GHz POWER8 Processor Card	2
EM96	16 GB DDR4 Memory	16
ESA5	LSI SAS Controller 9305-16E 12GB/S host bus adapter	4
ESC0	S&H - No Charge	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS1	ESS GL2S Solution Specify (4TB HDD) - 664 TB Raw Disk Capacity	1
ESSS	Spectrum Scale for ESS Standard Edition Indicator	1

Figure 32. ESS Data Server (Quantity 2) - Hardware Features

Product	Description	Qty
5147-084	ESS 5U84 Secondary Storage for Elastic Storage Server 1:ESS 5U84 Storage for Elastic Storage Server	1
4651	Rack Indicator, Rack #1	1
AG00	Shipping and Handling - No Charge	1
AJG0	4TB Enterprise HDD	84
EFD0	MFG Routing Indicator for Rochester/Shenzhen	1
EGS1	Solution Sub System #1 Indicator	1
EN42	Storage Subsystem ID 02	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS1	ESS GL2S Solution Specify (4TB HDD) - 664 TB Raw Disk Capacity	1

Figure 33. ESS Secondary Storage (Quantity 1) - Hardware Features

Product	Description	Qty
5147-084	ESS 5U84 Primary Storage for Elastic Storage Server 1:ESS 5U84 Storage for Elastic Storage Server	1
4651	Rack Indicator, Rack #1	1
AG00	Shipping and Handling - No Charge	1
AJG0	4TB Enterprise HDD	82
AJG3	800GB SED SSD	2
EFD0	MFG Routing Indicator for Rochester/Shenzhen	1
EGS0	Primary Unit Indicator	1
EGS1	Solution Sub System #1 Indicator	1
EN41	Storage Subsystem ID 01	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS1	ESS GL2S Solution Specify (4TB HDD) - 664 TB Raw Disk Capacity	1
SVC0	IBM Systems Lab Services implementation services	12

Figure 34. ESS Primary Storage (Quantity 1) - Hardware Features

Product	Description	Qty
8831-S52	Switch 1:8831 Model S52	1
4651	Rack Indicator, Rack #1	1
6665	Power Cord 2.8m (9.2-ft), Drawer to IBM PDU, (250V/10A)	2
ESS0	ESS 5U84 Storage Solution Specify	1
EU36	1U Air Duct and 4 Post Rack Mount Rail Kit	1
7014-T42	Rack 1:Rack Model T42	1
4651	Rack Indicator, Rack #1	6
6069	Front door (Black) for High Perforation (2m racks)	1
6098	Side Panel (Black)	2
6654	PDU to Wall Powercord 14', 200-240V/24A, UTG0247, PT#12	4
9300	Language Group Specify - US English	1
EPTG	SUBSTITUTE BASE 9188 AC PDU WITH BASE VERSION EPTJ AC PDU	1
EPTJ	ADDTNL PDU, WW, 1-PH 24/48A, 1-PH 32/63A, 3-PH 16/32A, 9XC19 OUTPUTS, SWITCHED, UTG624-7 INLET	3
ER18	Rack Content Specify: 8247-21L - 2EIA	1
ER1B	Reserve 1U at Bottom of Rack	1
ER1T	Reserve 1U at Top of Rack	1
ER1V	Rack Content Specify: 8831-S52 - 1EIA	1
ERLR	Left/Right PDU Redundancy	1
ESC0	Shipping and Handling - No Charge	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS1	ESS GL2S Solution Specify (4TB HDD) - 664 TB Raw Disk Capacity	1

Figure 35. ESS Internal Switch and Rack - Hardware Features

5.9 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.

5.9.1 Node Configurations

1. A Cluster Type of *Performance* may be selected, and the node configuration outlined in Figure 18 on page 45 may be substituted for the Worker Nodes. This variation alters the characteristics of the system to favor better performance.
2. 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.
3. 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.
4. 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.
5. 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.
6. HDP Node HDD sizes may be increased up to 8 TB per drive. This variation increases the local storage capacity for the Node which may increase performance.
7. HDP Node HDD sizes may be decreased down to 2 TB per drive. This variation reduces the local storage capacity for the Node which may decrease performance.
8. Additional HDDs may be added to the Worker Nodes. This variation increases the local storage capacity for the Worker Nodes and is likely to provide better performance than adding the same additional capacity by increasing drive size.

5.9.2 HDP Node Counts - Increasing

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

5.9.2.1 Additional Worker Nodes

Additional Worker Nodes may be specified to increase compute capacity and processing 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:

1. Deciding how many Worker Nodes are required.

2. Adding an appropriate number of Master Nodes and Edge Nodes to handle the increased number of Worker Nodes
3. Specifying the additional physical infrastructure for the additional Nodes (for example, racks, PDUs)
4. Scaling the network design appropriately for the total number of Nodes

5.9.2.2 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.

5.9.2.3 Additional Edge Nodes

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

5.9.3 HDP Node Counts – Decreasing

Nodes counts may be reduced to the minimums listed in Figure 18 on page 8 with a corresponding reduction in System performance and capacity.

When using this reference design, it is not recommended to reduce the Node counts below the minimum numbers listed in Figure 18 on page 45. A System can function with fewer Nodes, but reducing the Node counts further 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 as commonly expected.

5.9.4 ESS

5.9.4.1 Storage Capacity

Storage capacity can largely be chosen independently from the other elements within this architecture. For a given ESS model, changing the storage capacity is largely a matter of selecting a different drive size from the choices provided. For example, the GL2S which was selected for this design offers 4TB, 8TB, and 10TB drive options providing 664TB, 1328TB, and 1660TB raw storage capacity respectively. Any of these drive options may be selected for the ESS configuration without typically requiring the balance of this design to be altered.

5.9.4.2 Network Interfaces

It is a recommended best practice to configure the ESS network interfaces such that the interfaces have a bandwidth that equals or exceeds the maximum data rate that the ESS can supply. This helps ensure that the ESS network interfaces are not the limiting factor in the data pipeline (initially or as the Cluster grows). Configuring substantially more network bandwidth than the ESS maximum data serving rate is not typically useful unless additional physical links are desired for resilience. Configuring less bandwidth than the ESS maximum data serving rate may result in reduced performance or limits to future growth.

5.9.4.3 Other Models

Any ESS model may be selected to be included in the design of a System. However, each particular ESS model brings specifications that must be factored into the larger System design. One of the most important is the maximum data serving rate that the particular model is capable of supplying. This influences the ESS network interface configuration which in turn influences the design of the Network Subsystem and the particular switch selections and port allocations. Further, these affect the maximum number of HDP Nodes clients that can be served or the performance that will be realized. Thus, a different ESS model may be selected for this design, but the associated implications of that particular model choice must be factored into the overall System design.

5.9.4.4 Additional ESSs

The Data Store may be realized by more than one ESS. Additional ESSs may be added initially to the System design, or they may be added later to increase storage capacity. From the HDP Cluster point of view the additional ESSs offer additional storage capacity. From a management point of view, each ESS is a separately managed element.

In addition to the above, the effect of adding an ESS is similar to the effect of choosing a different ESS model. Thus, the associated implications of having more than one ESS must be factored into the overall System design

5.9.5 Network Configurations

1. The 1GbE networks may be hosted on a single Utility Switch. This variation trades some resilience for lower price.
2. The Data Network may be hosted on a single Data Switch. This variation trades performance and resilience for lower price.
3. 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).

6 Reference Design 2.1B – 30 Node Server Dense Configuration

This section describes another reference design for this solution. It is an example of another system design that complies with the architecture explained in the earlier section. This design varies from the above reference design in the following primary points:

- This design specifies 1U servers (9006-12P) for the Worker Nodes. This substitution achieves greater rack density for the servers that are part of the HDP Cluster. Specifically, in each 2U of server rack space, this design includes two 9006-12P servers which have a combined 40 cores and up to a combined 1TB of memory compared with one 9006-12P server which has 22 cores and up to 512GB of memory. Thus, this design offers more server cores and more server memory for each U of rack space.

  Note, however, this substitution trades-off some performance on each server as the 2U server (9006-22P) offers more cores per server (22 versus 20), and higher CPU clock rates (2.89 GHz versus 2.095 GHz). Thus, for a particular deployment, which server is preferable is largely dependent upon which parameters and metrics (for example, space/performance, price/performance, power/performance) are most important for that deployment.

- The ESS model specified is a GL4S, providing increased storage capacity and increased data serving rate (estimated 24 GBps for the GL4S versus 12 GBps for the GL2S).

Much of the design information for this reference design is similar to the previous reference design, and significant portions are identical. However, note that some important differences apply and the information in the following sections is intended to be a complete and independent description for this reference design.

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.

6.1 HDP Node Configurations

6.1.1 Hardware Configurations

This design selects a “Server Dense” Cluster Type. The specific Node configurations for each Node type for this design are listed in Figure 36.1 and 36.2. It lists the configuration parameters for other Cluster Types (“Balanced” and “Performance”) which are listed for comparison only and not directly applicable to this reference design.

	System Mgmt Node	Master Node	Edge Node	Worker Node		
Cluster Type	*All*	*All*	*All*	*Balanced*	*Performance*	*Server Dense*
Server Model	1U LC921	1U LC921	1U LC921	2U LC922	2U LC922	1U LC921
# 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	40
Memory	32GB	256GB	256GB	256GB	512GB	256GB
Storage - HDD (front)	2x 4TB HDD	4x 4TB HDD	4x 4TB HDD	4x 4TB HDD		4x 4TB HDD
Storage - SSD (front)					4x 3.8TB SSD	
Storage - HDD (rear for OS)				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 36.2. HDP configuration with LC922/LC921

	System Mgmt Node	Master Node	Edge Node	Worker Node
Cluster Type	*All*	*All*	*All*	***Storage Dense – ESS***
Server Model	2U IC922	2U IC922	2U IC922	2U IC922
# Servers (Min/Default/Max)	1 / 1 / 1	3 / 3 / Any	1 / 1 / Any	4 / 8 / Any
Sockets	1	2	2	2
Cores (total)	12	40	40	40
Memory	32GB	256GB	256GB	256GB
Storage Backplane (Front)	1	1	1	3
Storage - HDD (front)	2x 2.4TB HDD	4x 2.4TB HDD	4x 2.4TB HDD	4x 2.4TB HDD
Storage - SSD (front)				
OS Storage - HDD (front)				2x 2.4TB HDD
Storage Controller	1x Broadcom 9300-8i	1x Broadcom MegaRAID 9361-8i 1x Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1x Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1x Broadcom 9305-16i
Network* - 1 GbE	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 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)
Network** - 10 GbE	1x 2-port (2 ports)	1x 2-port (2 ports)	2x 2-port (4 ports)	1x 2-port (2 ports)
Cables** - 10 GbE	2 cables (DACs)	2 cables (DACs)	4 cables (DACs)	2 cables (DACs)
Operating System	RHEL 7.6 for P9	RHEL 7.6 for P9	RHEL 7.6 for P9	RHEL 7.6 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 362. HDP configuration with IC922

Node Counts

6.1.1.1 Worker Nodes

Twenty-two (22) Worker Nodes are specified for this design. Twenty-two Worker Nodes provide significant processing capacity while allowing the HDP Cluster to be contained within one rack.

6.1.1.2 Master Nodes

Five (5) Master Nodes are specified for this design. Five Master Nodes allows the HDP Functions to be distributed to provide significant HA protection and sufficient Master Node capacity for a 30 Node Cluster.

6.1.1.3 Edge Nodes

Two (2) Edge Node is specified for this design. This selection represents a basic HA configuration for Edge Functions.

6.1.1.4 System Management Nodes

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

6.2 ESS Configuration

This design specifies one (1) IBM GL4S Elastic Storage Server to serve as the Data Store. The following configuration selections are made for the ESS:

- 4TB HDDs (1332 TB raw disk capacity; ~888 TB usable disk capacity)
- Three (3) 10, 56 or 100 GbE network adapters per Storage Node, one port per adapter used
- One (1) internal management switch, 1 GbE (MTM=8831-S52, Mellanox AS4610)

This ESS model is BMC based and does not include an HMC.

6.3 Software

For this reference design, the following software is specified. (Note that some other software versions are compatible with the architecture. See also [5] and [6].)

6.3.1 Operating System Software

RHEL 7.6 for POWER9 is specified as the operating system software for all HDP Nodes.

6.3.2 Platform Software

Cloudera HDP version 3.1.5 is specified as the Platform software for this design.

6.3.3 Spectrum Scale Components

The following Spectrum Scale components for the HDP Cluster are specified for this design:

- IBM Spectrum Scale 5.0.1.1+ – Standard Edition
- IBM Spectrum Scale Transparency Connector 3.0.0-0
- IBM Spectrum Scale Ambari Management Pack 2.7.0.0

These are used to install required Spectrum Scale modules on HDP nodes. IBM Spectrum Scale software in IBM Elastic Storage Server (ESS) comes pre-installed. No additional Spectrum Scale licenses are required beyond those included with ESS for this reference architecture.

The Spectrum Scale software for the ESS is separate from the above and included with the ESS configuration.

6.4 Network Subsystem

This design specifies a logical network design that follows the architecture guidelines. Specifically, 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 following choices apply to the network design. The specific 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.

6.4.1 Logical Networks – HDP Cluster

6.4.1.1 Data Network

The Data Network is private (within this system) and assigned to VLAN 77. The servers in the system present untagged traffic to the switches for this network. This network is hosted by the Data Switches.

6.4.1.2 Campus Network

The Campus Network is shared (outside of the System) and assigned to VLAN 22. This network is hosted by the Utility Switches and uplinked into the existing client network infrastructure. The servers in the system present *tagged* traffic to the switches for this network.

6.4.1.3 Management Network

The Management Network is shared (outside of this system) and assigned to VLAN 1. This network is hosted by the Utility Switches and uplinked into the existing client network infrastructure. The servers in the system present *tagged* traffic to the switches for this network. This network also carries management traffic for the management interfaces for all of the Switches in the System.

6.4.1.4 Provisioning Network

The Provisioning Network is private (within this system) and assigned to VLAN 88. This network is hosted by the Utility Switches. The servers in the system present *untagged* traffic to the switches for this network. Configuring for untagged traffic more conveniently supports NetBoot, which is used to provision the Nodes in the HDP Cluster.

6.4.1.5 Service Network

The Service Network is private (within this system) and assigned to VLAN 188. This network is hosted by the Utility Switches. 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 HDP Nodes during provisioning.

6.4.2 Logical Networks – ESS

The Data Network, Campus Network, and Management common to the HDP Cluster and the ESS. Three is one instance of each of these networks, and they are described as part of the HDP Cluster above. The ESS has external (to the ESS) connections to these three networks.

The ESS also contains private internal networks to accomplish its installation and configuration and infrastructure level management. These are not visible outside of the ESS, and they are described here only to provide a complete network design for the System – consistent with the specific ESS model selected.

6.4.2.1 Provisioning Network – ESS

The Provisioning Network – ESS is private (within the ESS) and assigned to VLAN 91. This network is hosted by the ESS Internal Management Switch. The servers in the system present *untagged* traffic to the switches for this network. Configuring for untagged traffic more conveniently supports NetBoot, which is used to provision the Nodes in the HDP Cluster.

6.4.2.2 Service Network - ESS

The Service Network – ESS is private (within the ESS) and assigned to VLAN 191. This network is hosted by the ESS Internal Management Switch. The BMCs in the ESS present untagged traffic to this Switch for this network. The BMC-to-switch connections are dedicated to this function. The Elastic Management Server also has an OS level connection to this network to accomplish power control of the ESS Nodes.

Figure 37 and Figure 38 depict the logical network topology for this reference design. Figure 38 excludes the Provisioning Networks and the Service Networks from the diagram -- allowing a simpler and cleaner rendering that better illustrates the connectivity to the shared networks.

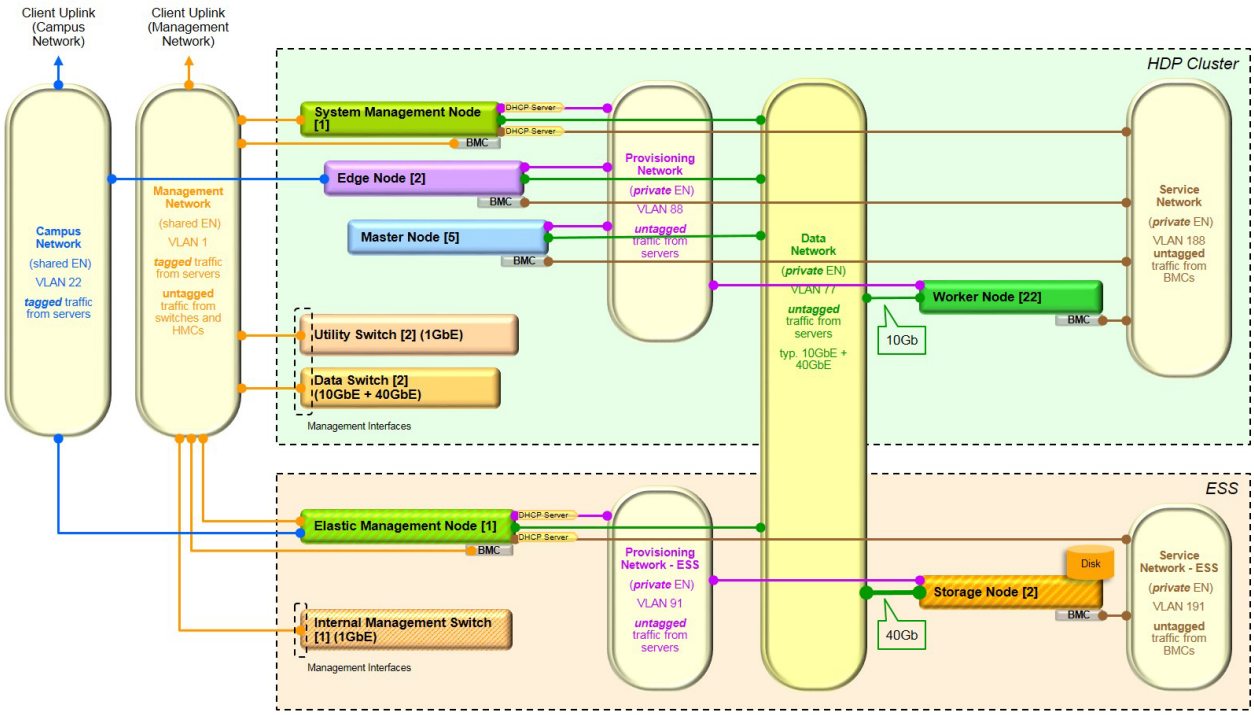

Figure 37. Network Design - Logical View - All Networks

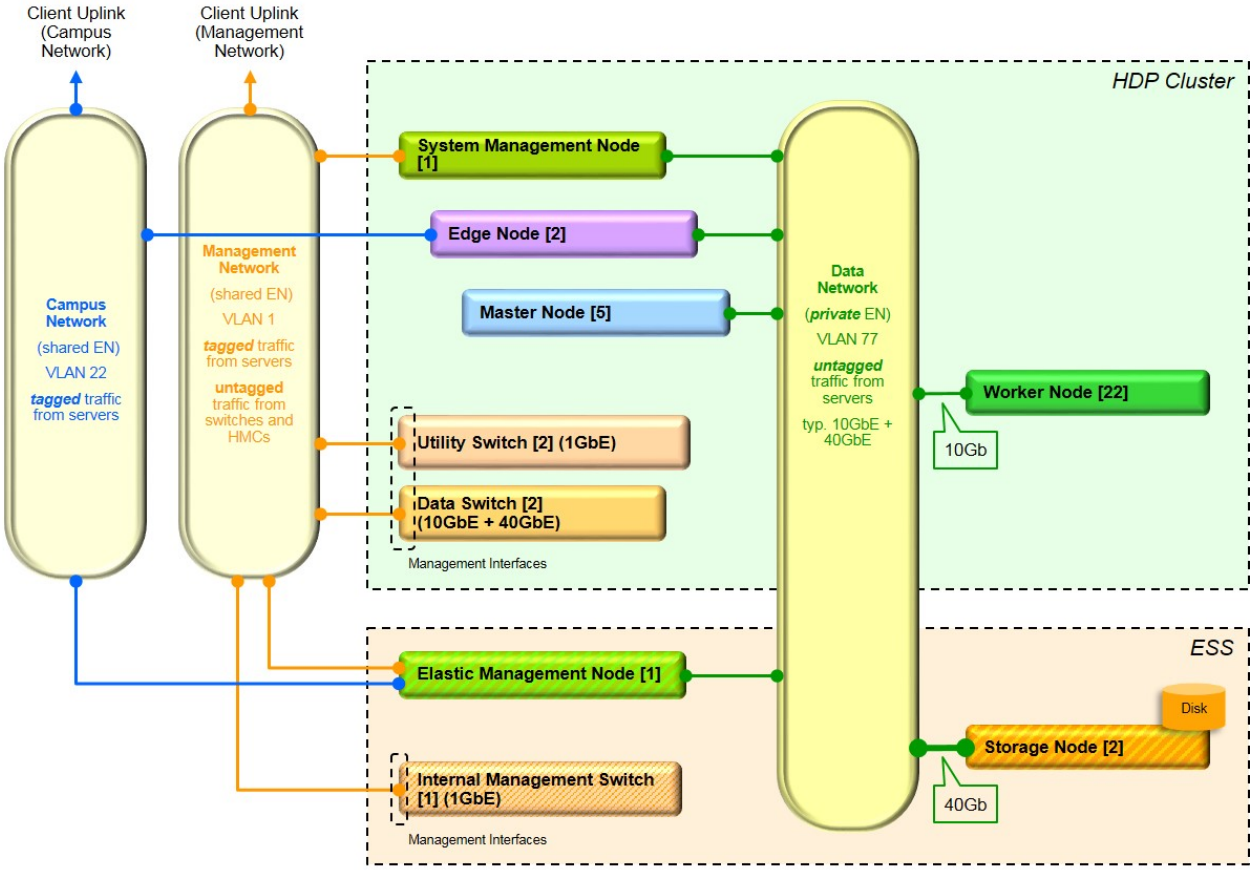

Figure 38. Network Design – Logical View – Data, Campus, and Management Networks

6.4.3 Switches

6.4.3.1 Utility Switches

Two (2) IBM 8831-S52 (Mellanox AS4610) switches are specified for the Utility Switches in this design. These are configured as a redundant pair with MLAG.

6.4.3.2 Data Switches

Two (2) IBM 8831-25M (Mellanox SN2410) switches are specified for the Data Switches in this design. These are configured as a redundant pair with MLAG.

6.4.3.3 ESS Internal Management Switch

One (1) IBM 8831-S52 (Mellanox AS4610) switch is specified as part of the ESS internal configuration.

6.4.4 Cabling

The physical cabling for each Server in the HDP Cluster follows a consistent pattern, and the switch-side port configurations for each HDP Node is typically the same. This provides consistency and reduces the opportunity for error. It also provides flexibility for special situations that might arise. Using 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 described in the previous section.

6.4.4.1 Utility Switches (1 Gb)

The connection for each HDP Node to 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 Utility 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 LACP on the Server and on the Switch. IP address configuration is applied to the bond interface for the native VLAN (88), and VLAN based interfaces with IP addresses are added for 1Gb traffic that requires tagging (VLANs 1, 22, and 177).

The Utility Switches also host the Service Network. The Service Network is different than the other 1Gb networks in that each server has a single dedicated link between its BMC interface and one of the Switches. The BMC interfaces are divided equally across the Switches (Utility Switch A provides a connection for half of the BMCs and Utility Switch B provides a connection for the other half). This is necessary to maximize the Utility Switch port utilization as the number of Utility Switch ports is a limiting factor for the number of Servers that can be installed in one rack. 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.

The Utility Switch pair is configured and cabled with IPLs for MLAG, which allows the links to be aggregated across the pair of Switches.

See Figure 39 for a diagram of the cabling design for the Utility Switches.

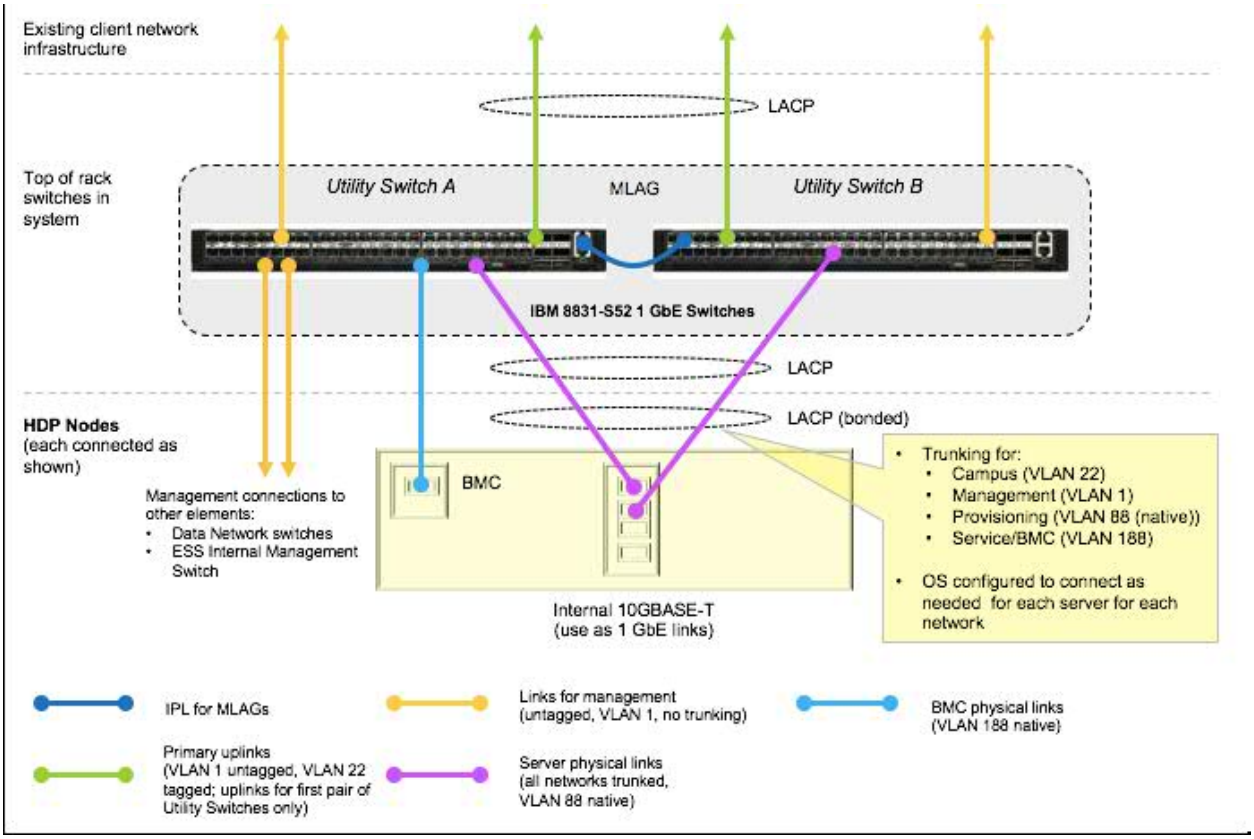

Figure 39. Utility Switch Cabling - Physical Schematic View

6.4.4.2 Data Switches (25Gb and 100Gb Links)

The connection for each HDP Node to the Data Network is carried over two physical links (cables) to the Data Switches. This provides a redundant path that is used to provide resilience for these networks, as well as increased bandwidth (up to 50 Gb) for each HDP 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 Utility Switch links, this pair of links is configured for link aggregation using LACP on the Server and on the Switch.

The connection for each Storage Node to the Data Network is carried over three physical links (cables) to the Data Switches. This provides a redundant path that is used to provide resilience for these networks, as well as increased bandwidth (up to 75Gb) for each Storage Node. 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 Utility Switch links, this pair of links is configured for link aggregation using LACP on the Server and on the Switch.

The Data Switch pair is configured and cabled with IPLs for MLAG, which allows links to be aggregated across the pair of Switches.

The two pairs of Data Switches are also crosslinked to each other using seven 25 Gb links. This allows the Cluster to be have full connectivity within the Cluster, independent from any uplinks.

See Figure 40 for a diagram of the cabling design for the Data Switches.

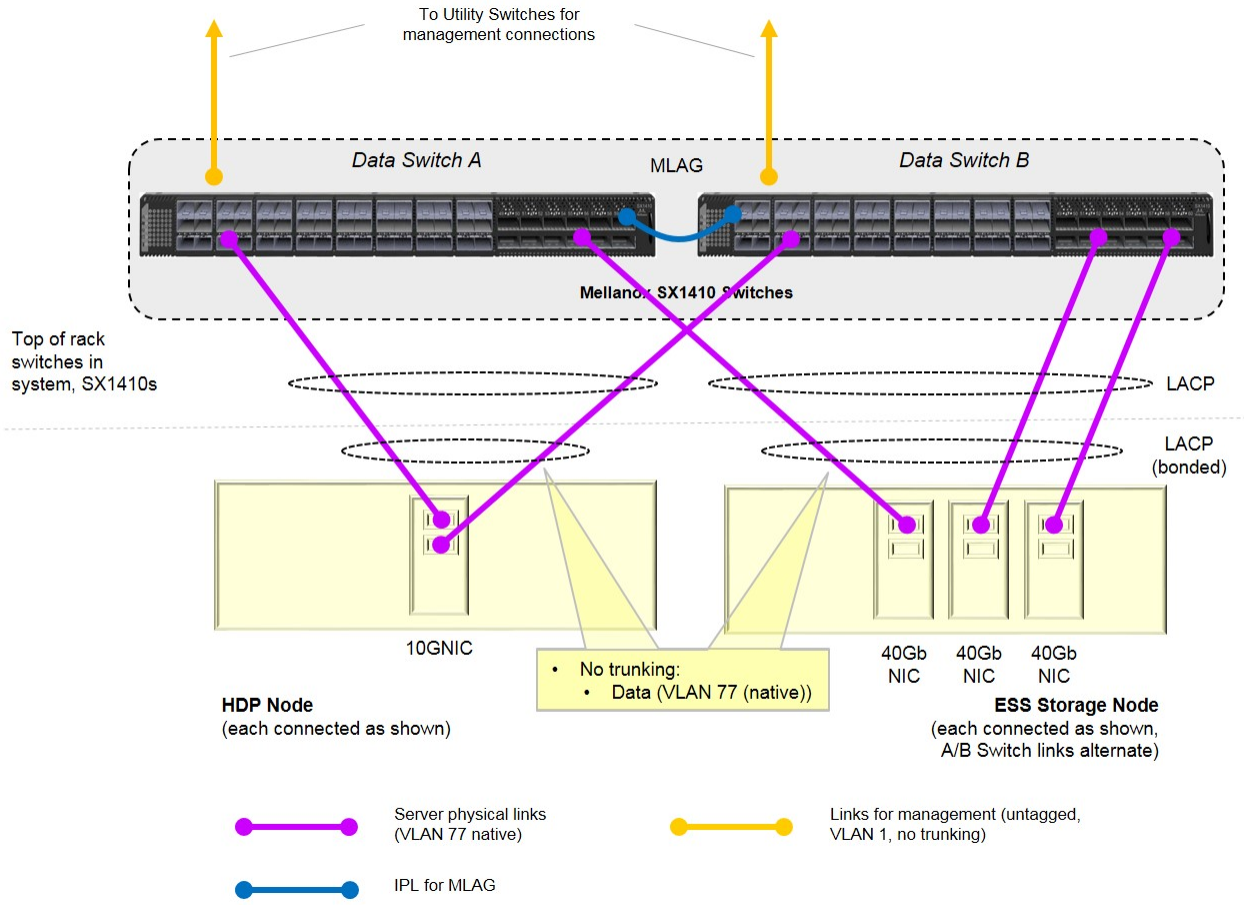

Figure 40. Data Switch Cabling - Physical Schematic View

6.4.5 Other Considerations

6.4.5.1 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).

6.4.5.2 Dynamic Host Configuration Protocol (DHCP)

This design provides DHCP for two of the networks in the System. The System Management Node is configured to provide DHCP for the Service Network and the Provisioning Network.

6.5 Data Pipeline Calculations

An overview of the data pipeline and the primary stages is depicted in Figure 41.

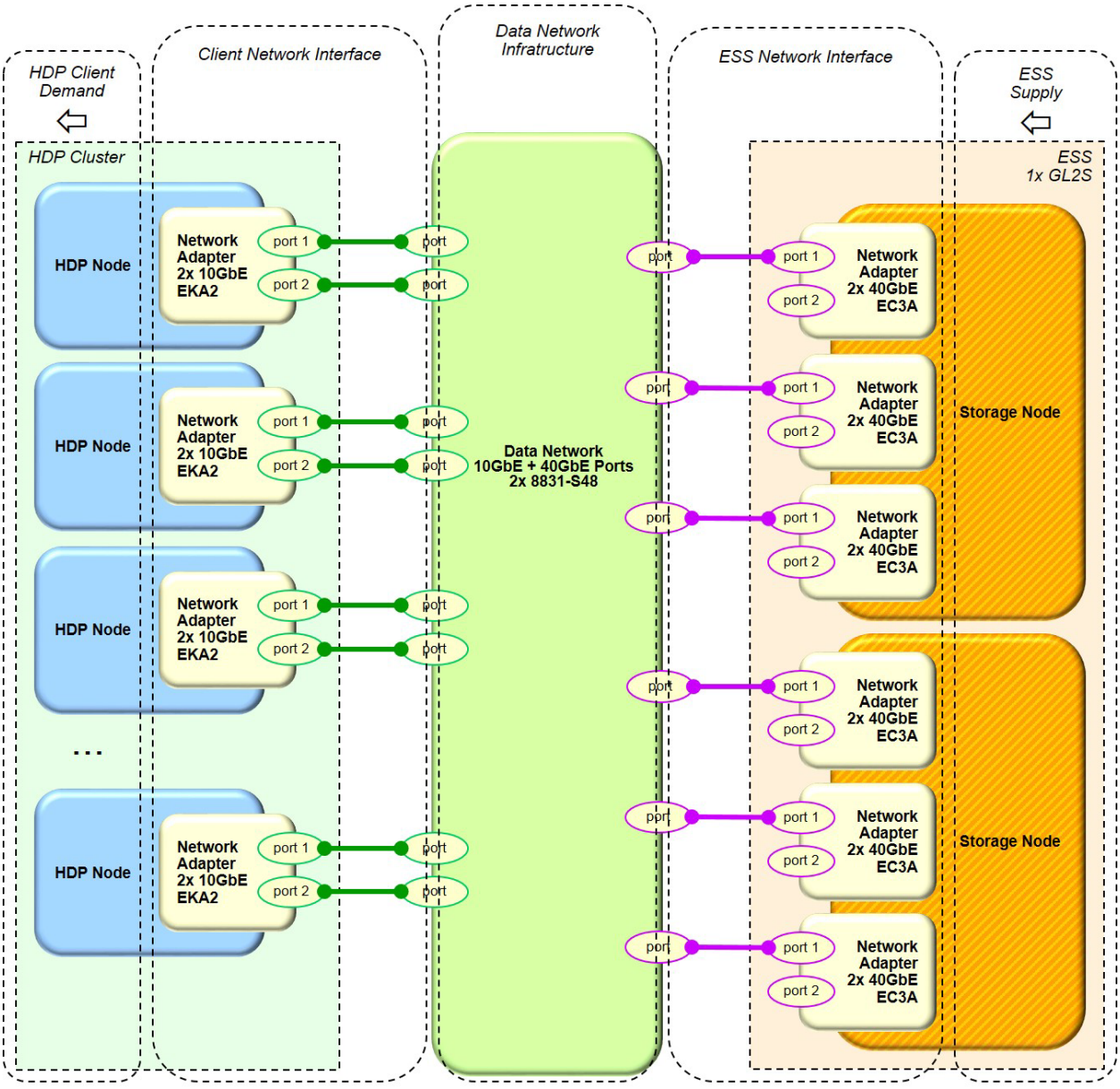

Figure 41. Data Pipeline Overview

The following calculations apply to the relevant stages of the data pipeline in this design.

6.5.1 Client Demand

The total client demand is estimated to be 11.2 GBps – all from HDP Nodes as follows.

6.5.1.1 Worker Nodes

For each Worker Node, the peak demand is estimated to be 440 MBps [2]. With 22 Worker Nodes, the total demand is 22 * 440 MBps = 9.7 GBps.

6.5.1.2 Master Nodes

For each Master Node, the average demand is estimated to be one half of the Worker Node demand or 220 MBps. With five Master Nodes, the total demand is 5 * 220 MBps = 1.1 GBps.

6.5.1.3 Edge Nodes

For each Edge Node, the average demand is estimated to be one half of the Worker Node demand or 220 MBps. With two Edge Nodes, the total demand is 2 * 220 MBps = 0.4 GBps.

6.5.2 Client Network Interface

Each client has the same network interface structure. Specifically, one network adapter with twp 10 GbE connections. Assuming 7% ethernet overhead, the network interface capacity for one HDP Node is then 2 connections * 10 Gbps * 0.93 = 2.3 GBps. Thus, the total client network interface capacity is 67.4 GBps.

(The single HDP Node network interface capacity is higher than the demand from any one of the HDP Nodes, so no adjustment is required in this regard.)

6.5.3 Data Network Infrastructure

The Data Network infrastructure for this design consists of two 8831-S48 (Mellanox SX1410) switches. The switches are configured as a redundant pair with MLAG, so using Mellanox specifications, the aggregate bandwidth for the switching infrastructure is calculated to be 2 * 1.92 Tbps = 480 GBps.

6.5.4 ESS Network Interface

The ESS in this design includes two Storage Nodes, each configured identically. Specifically, each Storage Node has three network adapters, and each adapter has two 10, 56 or 100 GbE ports. Given the above, it is a design choice to use one port per adapter and include three adapters in each Storage Node to provide three GbE connections at full, constant data rate. Assuming 7% ethernet overhead, the ESS network interface capacity is therefore 3 connections * 25 Gbps * 0.93 * 2 Storage Nodes = 27.9 GBps.

6.5.5 ESS Supply

The rate at which the ESS GL4S can supply data is estimated to be 24 GBps.

---

[2] This is an input assumption based upon some common benchmark measurements. For any specific client design, demand estimates that are appropriate for the particular client and workload should be used.

Figure 42 provides a graphical comparison of the above calculations. For the data pipeline, the end-to-end data rate and pipeline utilization is effectively limited by the smallest data rate for any stage in the pipeline. For this design, this is the client demand at 11.2 GBps. It should be noted that the above calculations are generally conservative in that the supply and demand rates are estimated as peak/maximum rates, but with multiple parallel data paths, the peak demand from each client is not likely to occur simultaneously.

For this reference design, it was a goal to provide a pipeline rate that satisfies the maximum estimated client demand rate. Elements such as the ESS model and ESS network interface components and network connections were chosen such that the client demand was not limited by any other element in the Infrastructure. These calculations confirm this design goal is met.

The above calculations also provide information regarding performance headroom and growth headroom. For example, if the workload on the HDP Cluster is changed such that the data demand increases, the Infrastructure has sufficient headroom to accommodate that increase, up to ~ 24 GBps of demand from the HDP Cluster (~114% increase). Alternatively, more Worker Nodes may be added to the HDP Cluster, and the rest of the System has sufficient capacity to serve data to them. For example, an additional twenty-nine (29) Worker Nodes could be added (bringing the total client demand to ~24 GBps), before other limits in the pipeline (specifically, the ESS supply rate at 24 GBps) may start to become a limiting factor.

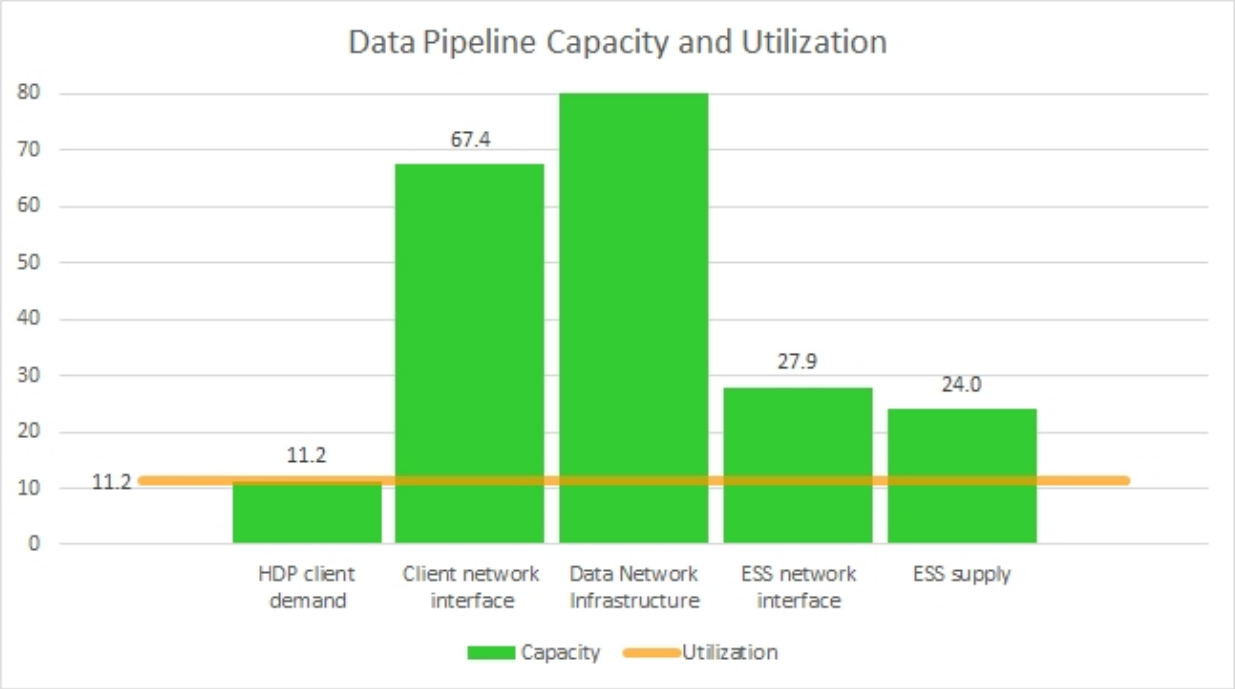

Figure 42. Data Pipeline Capacity and Utilization

6.6 Physical Configuration - Rack Layout

Figure 43 shows the physical layout of the System within its racks. All of the components for this reference design fit within three 42U racks.

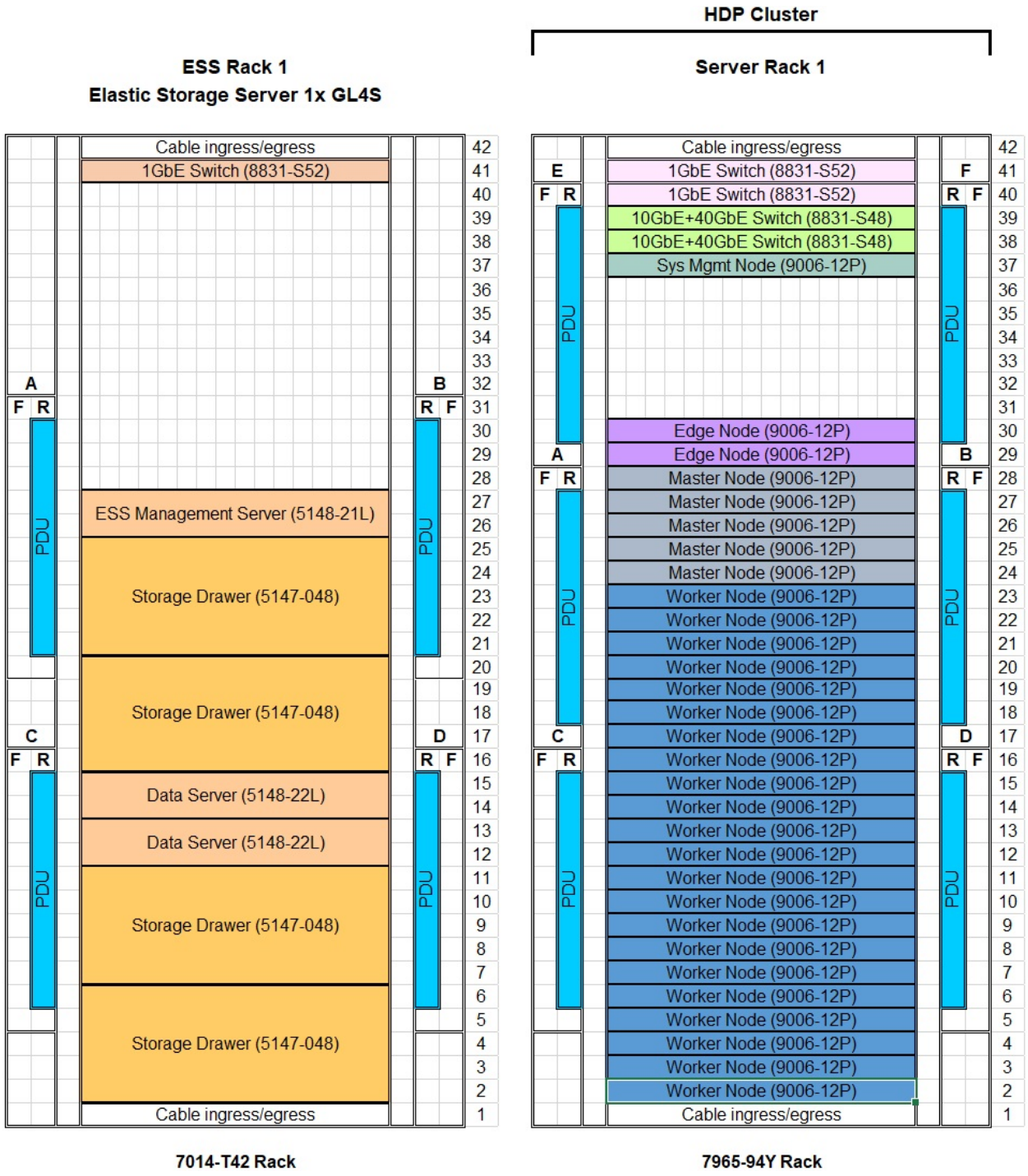

Figure 43. Physical Rack Layout

6.7 Hardware Features for e-config – HDP Cluster

Following are the e-config hardware features for the HDP Cluster in this design – including the Utility Switches and the Data Switches. A full e-configuration includes the system software (for example, RHEL) and services. The HDP software is supplied by Cloudera. The Spectrum Scale software for the HDP Nodes is obtained separately from IBM. 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 based on the processor selection.

9006	12P		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
	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 3.5in SAS HDD	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 44. Worker Nodes (Quantity 22) - 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 45. Master Nodes (Quantity 5) - 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 46. Edge Node (Quantity 2) - 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 47. 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 48. Switches and Server Rack (Quantity 1) - Hardware Features

6.8 Hardware Features for e-config – ESS

Following are the e-config hardware features for the ESS in this design. A full e-configuration includes the system software (for example, RHEL) and Spectrum Scale software.

Much of the ESS configuration is specified by e-config as part of the ESS solution definition. Items which should be specifically customized or selected are highlighted in the lists below.

In Figure 49, the ESS Management Server must use a network adapter that matches the data network adapter on the ESS Data Server. Thus, in the eConfig for the ESS Management Server, specify Feature Code EC3A instead of EL3X.

Product	Description	Qty
5148-21L	ESS Management Server:5148 Model 21L	1
10	One CSC Billing Unit	10
266	Linux Partition Specify	1
1111	CAT5E Ethernet Cable, 3M BLUE	2
1115	CAT5E Ethernet Cable, 3M GREEN	1
1118	CAT5E Ethernet Cable, 3M YELLOW	1
2147	Primary OS - Linux	1
4651	Rack Indicator, Rack #1	1
5000	Software Preload Required	1
5771	SATA Slimline DVD-RAM Drive	1
6665	Power Cord 2.8m (9.2-ft), Drawer to IBM PDU, (250V/10A)	2
9300	Language Group Specify - US English	1
9442	New Red Hat License Core Counter	10
EC16	Open Power non-virtualized configuration	1
EJTT	Front Bezel for 12-Bay BackPlane	1
EL1A	AC Power Supply - 900W	2
EL3T	Storage Backplane 12 SFF-3 Bays/DVD Bay	1
EL3X	PCIe3 LP 2-port 10GbE NIC&RoCE SFP+ Copper Adapter	1
EL4M	PCIe2 LP 4-port 1GbE Adapter	1
ELAD	One Zero Priced Processor Core Activation for Feature #ELPD	10
ELD5	600GB 10K RPM SAS SFF-3 Disk Drive (Linux)	2
ELPD	10-core 3.42 GHz POWER8 Processor Card	1
EM96	16 GB DDR4 Memory	2
EN03	5m (16.4-ft), 10Gb E'Net Cable SFP+ Act Twinax Copper	2
ESC0	S&H - No Charge	1
ESS0	ESS 5U84 Storage Solution Specify	1

Figure 49. ESS Management Server (Quantity 1) - Hardware Features

Product	Description	Qty
5148-22L	ESS GLxS Data Server 1:5148 Model 22L	1
10	One CSC Billing Unit	10
266	Linux Partition Specify	1
1111	CAT5E Ethernet Cable, 3M BLUE	1
1115	CAT5E Ethernet Cable, 3M GREEN	1
1118	CAT5E Ethernet Cable, 3M YELLOW	1
2147	Primary OS - Linux	1
4651	Rack Indicator, Rack #1	1
5000	Software Preload Required	1
5771	SATA Slimline DVD-RAM Drive	1
6665	Power Cord 2.8m (9.2-ft), Drawer to IBM PDU, (250V/10A)	2
9300	Language Group Specify - US English	1
9442	New Red Hat License Core Counter	20
EC16	Open Power non-virtualized configuration	1
EC3A	PCIe3 LP 2-Port 40GbE NIC RoCE QSFP+ Adapter	3
ECBP	7m (23.1-ft), IBM Passive QSFP+ to QSFP+ Cable (DAC)	3
ECCT	3M 12GB/s SAS Cable W/Universal Key	8
EGS1	Solution Sub System #1 Indicator	1
EJTQ	Front Bezel for 8-Bay BackPlane	1
EL1B	AC Power Supply - 1400W (200-240 VAC)	2
EL3W	2U SAS RAID 0,5,6,10 Controller + Back plane	1
EL4M	PCIe2 LP 4-port 1GbE Adapter	1
ELAD	One Zero Priced Processor Core Activation for Feature #ELPD	20
ELD5	600GB 10K RPM SAS SFF-3 Disk Drive (Linux)	2
ELPD	10-core 3.42 GHz POWER8 Processor Card	2
EM96	16 GB DDR4 Memory	16
ESA5	LSI SAS Controller 9305-16E 12GB/S host bus adapter	4
ESC0	S&H - No Charge	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS4	ESS GL4S Solution Specify (4TB HDD) - 1336 TB Raw Disk Capacity	1
ESSS	Spectrum Scale for ESS Standard Edition Indicator	1

Figure 50. ESS Data Server (Quantity 2) - Hardware Features

Product	Description	Qty
5147-084	ESS 5U84 Secondary Storage for Elastic Storage Server 1:ESS 5U84 Storage for Elastic Storage Server	1
4651	Rack Indicator, Rack #1	1
AG00	Shipping and Handling - No Charge	1
AJG0	4TB Enterprise HDD	84
EFD0	MFG Routing Indicator for Rochester/Shenzhen	1
EGS1	Solution Sub System #1 Indicator	1
EN42	Storage Subsystem ID 02	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS4	ESS GL4S Solution Specify (4TB HDD) - 1336 TB Raw Disk Capacity	1

Figure 51. ESS Secondary Storage (Quantity 3) - Hardware Features

Product	Description	Qty
5147-084	ESS 5U84 Primary Storage for Elastic Storage Server 1:ESS 5U84 Storage for Elastic Storage Server	1
4651	Rack Indicator, Rack #1	1
AG00	Shipping and Handling - No Charge	1
AJG0	4TB Enterprise HDD	82
AJG3	800GB SED SSD	2
EFD0	MFG Routing Indicator for Rochester/Shenzhen	1
EGS0	Primary Unit Indicator	1
EGS1	Solution Sub System #1 Indicator	1
EN41	Storage Subsystem ID 01	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS4	ESS GL4S Solution Specify (4TB HDD) - 1336 TB Raw Disk Capacity	1
SVC0	IBM Systems Lab Services implementation services	12

Figure 52. ESS Primary Storage (Quantity 1) - Hardware Features

Product	Description	Qty
8831-S52	Switch 1:8831 Model S52	1
4651	Rack Indicator, Rack #1	1
6665	Power Cord 2.8m (9.2-ft), Drawer to IBM PDU, (250V/10A)	2
ESS0	ESS 5U84 Storage Solution Specify	1
EU36	1U Air Duct and 4 Post Rack Mount Rail Kit	1
7014-T42	Rack 1:Rack Model T42	1
4651	Rack Indicator, Rack #1	8
6069	Front door (Black) for High Perforation (2m racks)	1
6098	Side Panel (Black)	2
6654	PDU to Wall Powercord 14', 200-240V/24A, UTG0247, PT#12	4
9300	Language Group Specify - US English	1
EPTG	SUBSTITUTE BASE 9188 AC PDU WITH BASE VERSION EPTJ AC PDU	1
EPTJ	ADDTNL PDU, WW, 1-PH 24/48A, 1-PH 32/63A, 3-PH 16/32A, 9XC19 OUTPUTS, SWITCHED, UTG624-7 INLET	3
ER18	Rack Content Specify: 8247-21L - 2EIA	1
ER1B	Reserve 1U at Bottom of Rack	1
ER1T	Reserve 1U at Top of Rack	1
ER1V	Rack Content Specify: 8831-S52 - 1EIA	1
ERLR	Left/Right PDU Redundancy	1
ESC0	Shipping and Handling - No Charge	1
ESS0	ESS 5U84 Storage Solution Specify	1
ESS4	ESS GL4S Solution Specify (4TB HDD) - 1336 TB Raw Disk Capacity	1

Figure 53. ESS Internal Switch and Rack (Quantity 1) - Hardware Features

6.9 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.

6.9.1 Node Configurations

1. 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 RAS characteristics for lower price.
2. CPU for a Node type may be reduced to as low as 16 core processors. 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.
3. 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.
4. 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.
5. HDP Node HDD sizes may be increased up to 8 TB per drive. This variation increases the local storage capacity for the Node which may increase performance.
6. HDP Node HDD sizes may be decreased down to 2 TB per drive. This variation reduces the local storage capacity for the Node which may decrease performance.

6.9.2 HDP Node Counts - Increasing

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

6.9.2.1 Additional Worker Nodes

Additional Worker Nodes may be specified to increase compute capacity and processing 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:

1. Deciding how many Worker Nodes are required
2. Adding an appropriate number of Master Nodes and Edge Nodes to handle the increased number of Worker Nodes
3. Specifying the additional physical infrastructure for the additional Nodes (for example, racks, PDUs)
4. Scaling the network design appropriately for the total number of Nodes

6.9.2.2 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.

6.9.2.3 Additional Edge Nodes

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

6.9.3 HDP Node Counts – Decreasing

Nodes counts may be reduced to the minimums listed in Figure 18 on page 8 with a corresponding reduction in System performance and capacity.

When using this reference design, it is not recommended to reduce the Node counts below the minimum numbers listed in Figure 36 on page **Error! Bookmark not defined.**. A System can function with fewer Nodes, but reducing the Node counts further 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 as commonly expected.

6.9.4 ESS

6.9.4.1 Storage Capacity

Storage capacity can largely be chosen independently from the other elements within this architecture. For a given ESS model, changing the storage capacity is largely a matter of selecting a different drive size from the choices provided. For example, the GL4S which was selected for this design offers 4TB, 8TB, and 10TB drive options providing 1336TB, 2672TB, and 3340TB raw storage capacity respectively. Any of these drive options may be selected for the ESS configuration without typically requiring the balance of this design to be altered.

6.9.4.2 Network Interfaces

It is a recommended best practice to configure the ESS network interfaces such that the interfaces have a bandwidth that equals or exceeds the maximum data rate that the ESS can supply. This helps ensure that the ESS network interfaces are not the limiting factor in the data pipeline (initially or as the Cluster grows). Configuring substantially more network bandwidth than the ESS maximum data serving rate is not typically useful unless additional physical links are desired for resilience. Configuring less bandwidth than the ESS maximum data serving rate may result in reduced performance or limits to future growth.

6.9.4.3 Other Models

Any ESS model may be selected to be included in the design of a System. However, each particular ESS model brings specifications that must be factored into the larger System design. One of the most important is the maximum data serving rate that the particular model is capable of supplying. This influences the ESS network interface configuration which in turn influences the design of the Network Subsystem and the particular switch selections and port allocations. Further, these affect the maximum number of HDP Nodes clients that can be served or the performance that will be realized. Thus, a different ESS model may be selected for this design, but the associated implications of that particular model choice must be factored into the overall System design.

6.9.4.4 Additional ESSs

The Data Store may be realized by more than one ESS. Additional ESSs may be added initially to the System design, or they may be added later to increase storage capacity. From the HDP Cluster point of view the additional ESSs offer additional storage capacity. From a management point of view, each ESS is a separately managed element.

In addition to the above, the effect of adding an ESS is similar to the effect of choosing a different ESS model. Thus, the associated implications of having more than one ESS must be factored into the overall System design

6.9.5 Network Configurations

1. The 1GbE networks may be hosted on a single Utility Switch. This variation trades some resilience for lower price, and *each Server Rack is then limited to approximately 22 Nodes*.
2. The Data Network may be hosted on a single Data Switch. This variation trades performance and resilience for lower price.
3. If additional network bandwidth is needed to the servers for the Data Network, 100 GbE connections can be used instead of the 25 GbE connections. You can use the matching 100GbE-capable transceivers (Feature Code EB59) and cables (Feature Codes EB5J, EB5K, EB5L, EB5M).

Appendix A - 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.

A.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.

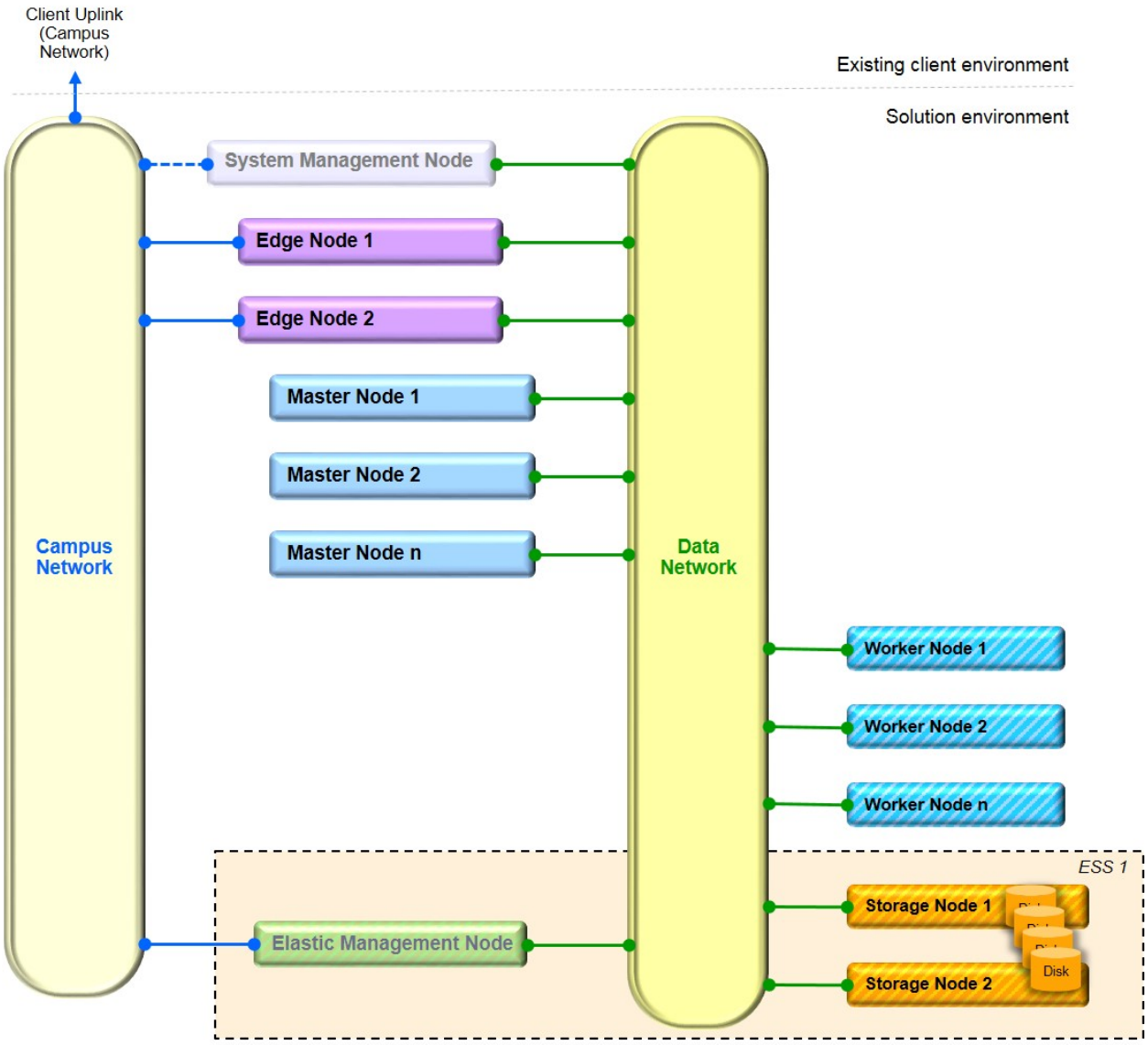

Figure 54. Partial-Homed Network

A.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 are available.

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

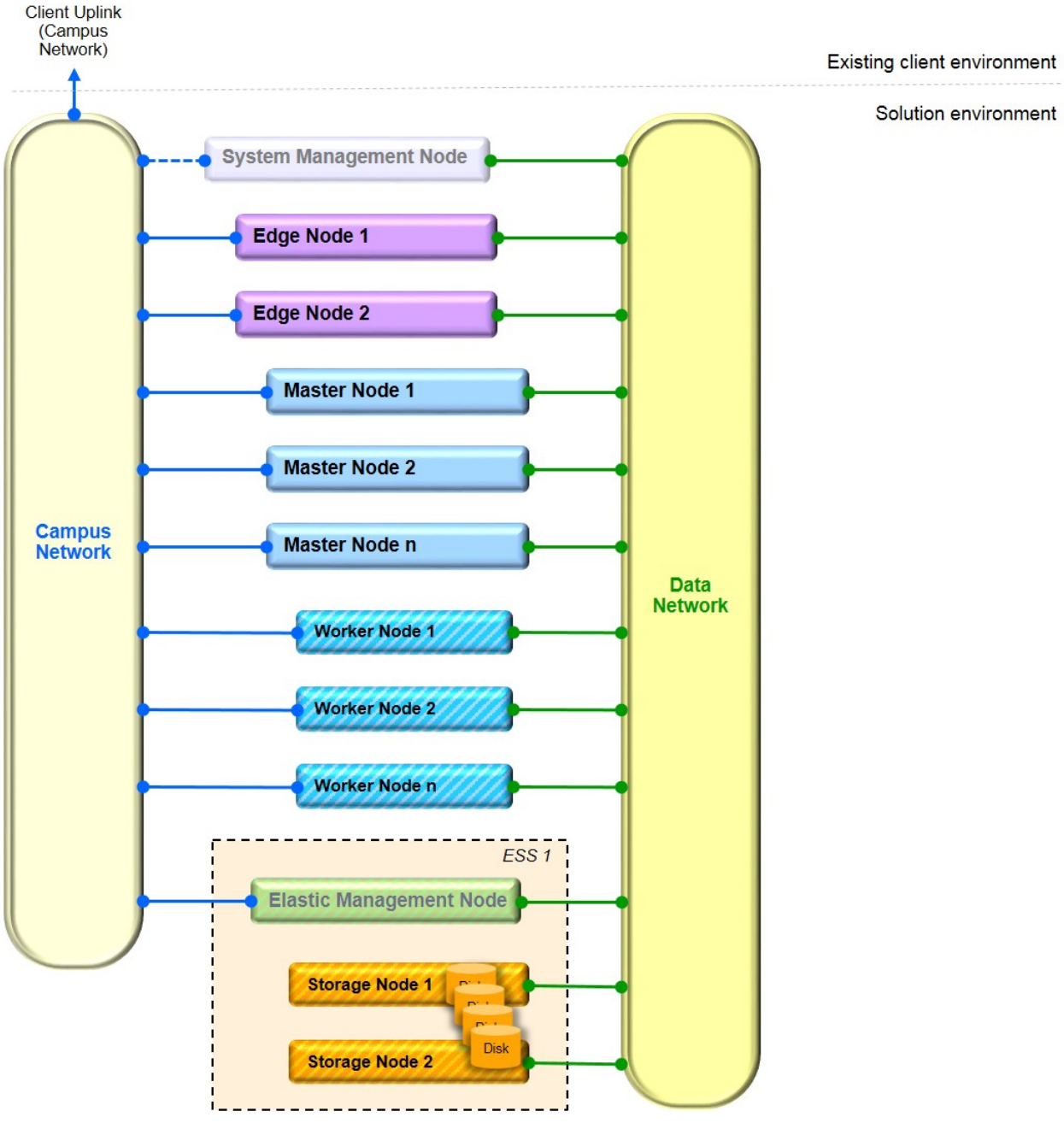

Figure 55. Dual-Homed Network

A.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

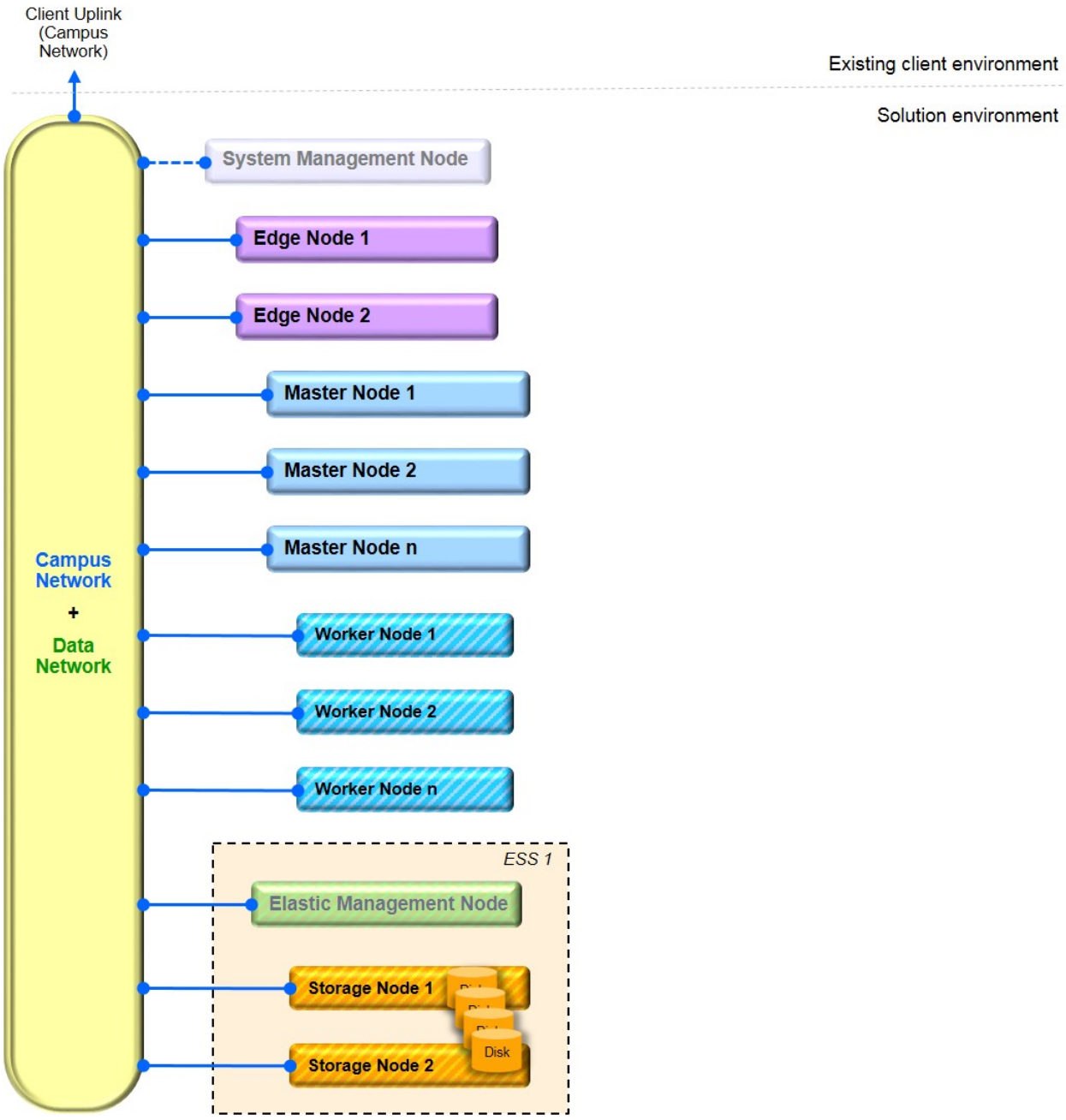

Figure 56. Flat Network

A.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

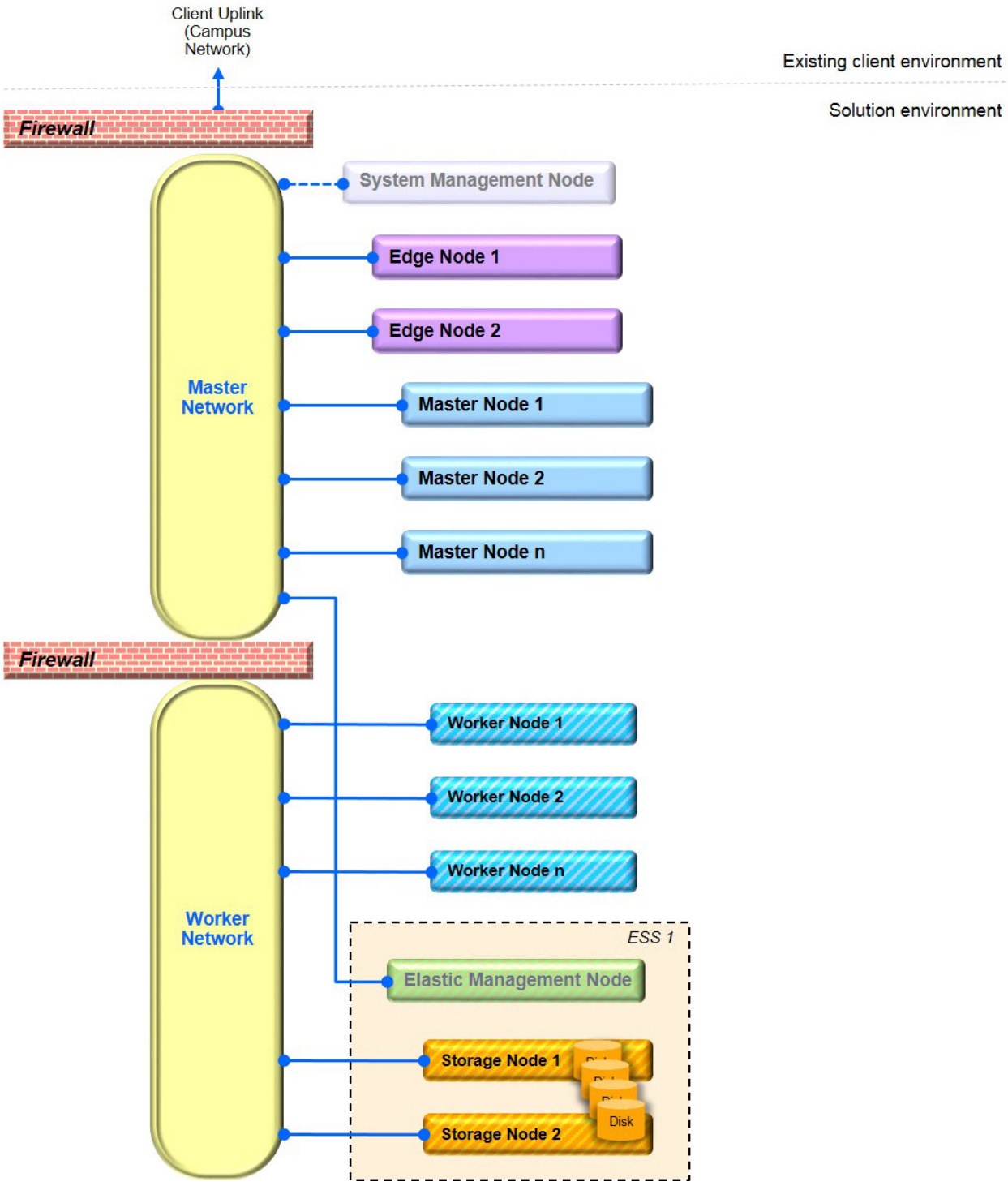

Figure 57. DMZ Network

Appendix B – Other POWER9 Server Considerations

IBM Power9 L922

As an alternative to the IBM Power System LC921 and Power LC922 server models, the Power L922 (9008-22L) server can be used for the nodes instead. Using the Power L922 server model requires IBM PowerVM® and the following configuration details will differ from Figure 18. The Power L922 server model has not been tested with HDP by IBM internally, and therefore, it is recommended that you test this environment with your workload before deploying in production.

- Work nodes (production environments) use:
  - One LPAR per server
  - One HDP worker node per server
  - Two 12-core ("big core") socket
  - Memory options: 256 GB (16x 16 GB DIMM), 512 GB (16x 32GB DIMM), or 1TB (16x 64GB DIMM)
  - Eight internal SFF HDD or SSD drives for the OS and HDP temp and shuffle data
  - One storage adapter (internal), one Data Network adapter, one Admin 1 GbE adapter
- Worker nodes (development/test environments) use:
  - Two LPARs per server using split backplane (Feature Code EJ1H)
  - Two HDP worker nodes per server, or 1 HDP worker node + 1 HDP Master/Edge Node
  - Two 8-, 10-, or 12-core ("big core") socket
  - Memory options: 256 GB (16x 16 GB DIMM) or 512 GB (16x 32 GB DIMM)
  - Each LPAR: Four internal SFF HDD or SSD drives for the OS and HDP temp and shuffle data
  - Each LPAR: One storage adapter (internal), one Data Network adapter, one Admin 1 GbE adapter
- Master nodes
  - Three or more servers. Each server has two LPARs using split backplane (Feature Code EJ1H)
  - Two 12-core ("big core") socket
  - 512 GB (16x 32 GB DIMM)
  - Each LPAR: Four internal SFF HDD or SSD drives for the OS and HDP Master Node services
  - Each LPAR: One storage adapter (internal), one Data Network adapter, one Admin 1GbE adapter
  - Assign three HDP Master Nodes over three servers using one LPAR per server. Other LPARs can be used for Edge Nodes, Utility Nodes, or HDF Master Nodes if HDF services are part of an existing or new HDP cluster.

IBM Power9 IC922

As an alternative to the IBM Power System LC921 and Power LC922 server models, the Power IC922 (9183-22X) server can be used for the nodes instead. The following configuration details will differ from Figure 18.

- Work nodes (production environments) use:
  - One HDP worker node per server
  - Two 20-core ("small core") socket
  - Memory options: 256 GB (16x 16 GB DIMM), 512 GB (16x 32GB DIMM), or 1TB (16x 64GB DIMM)
  - Twenty-four internal SFF HDD or SSD drives for the OS and HDP temp and shuffle data
  - Two storage adapters, one Data Network adapter, one Admin 1 GbE adapter (internal)
- Worker nodes (development/test environments) use:
  - One HDP worker nodes per server
  - Two 12-, 16-, or 20-core ("small core") socket
  - Memory options: 256 GB (16x 16 GB DIMM) or 512 GB (16x 32 GB DIMM)
  - Four to Eight internal SFF HDD or SSD drives for the OS and HDP temp and shuffle data
  - One storage adapter, one Data Network adapter, one Admin 1 GbE adapter (internal)
- Master nodes
  - One HDP master node per server
  - Two 20-core ("small core") socket
  - 512 GB (16x 32 GB DIMM)
  - Four internal SFF HDD or SSD drives for the OS and HDP Master Node services
  - One storage adapter, one Data Network adapter, one Admin 1GbE adapter (internal)

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 8 on Power IC922. 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.

Cloudera and HDP are registered trademarks or trademarks of Cloudera, 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.

Mellanox is a registered trademark of Mellanox Technologies, Ltd.

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.