Pivotal™

TECHNICAL DOCUMENTATION

Pivotal Clustering Concepts Guide

Reference Architecture for implementing a Pivotal cluster on customer-supplied hardware

Rev: 01

GoPivotal, Inc. believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." GOPIVOTAL, INC. ("Pivotal") MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

November 2013

Hard Disk Configuration Guidelines 17
LUN Configuration 17
Master Node 17
Segment and ETL Nodes 17
SAN/JBOD Storage 18

Network Layout Guidelines 19
General Recommendations 19
Network Topology 19
Network Connections in a Greenplum Database Cluster 20
Admin Networks 20
Interconnect Networks 20
External Network Connections 21

Installation Guidelines 21
Racking Guidelines for a 42U Rack 21
Cabling 22

Switch Configuration Guidelines 22

IP Addressing Guidelines 22
IP Addressing Scheme for each Server and CIMC on the Admin Network 22
IP Addressing for Non-server Devices 23
IP Addressing for Interconnects using 10 Gb NICs 23
IP Addressing for Fault Tolerant Interconnects 23

Data Loading Connectivity Guidelines 24
VLAN Overlay 24
How the VLAN Overlay Method Works 24
Configuring the Overlay VLAN—An Overview 26
Step 1 – Virtual interface tags packets with the overlay VLAN 26
Step 2 – Configure the switch in the cluster with the overlay VLAN 26
Step 3 – Configure the switch ports connected to the customer network 26
Direct Connect to the Customer's Network 27
Routing 27

Validation Guidelines 28
Labels 28

Certification Guidelines 29
Network Performance Test 29
`gpcheckperf` 29
Retail Demo 29

Hardware Monitoring and Failure Analysis Guidelines 29
Fans/Temp 29
Memory 29
System Errors 29
Power 30

Appendix A: Pivotal Cluster Examples 31
Greenplum Database Cluster Example 31
Pivotal Hadoop Cluster Example 32
Pivotal Hadoop with External HDFS (Isilon) Cluster Example 33
Performance Difference on a Single RAID5 vs JBOD 33

Appendix B: Example Rack Layout 34

Appendix C: Using `gpcheckperf` to Validate Disk and Network Performance 35
Checking Disk Performance —`gpcheckperf` Output 35
Checking Network Performance —`gpcheckperf` Output 36

Introduction to Designing a Pivotal Cluster with Customer-supplied Hardware

The Pivotal Appliance provides a ready-made platform that strives to accommodate the majority of customer workloads. More and more, Pivotal Engineering is seeing cases where customers elect to build a cluster that satisfies a specific requirement or purpose. Platform and Systems Engineering publishes this framework to provide field personnel with a resource for assisting customers in this effort.

Objectives—Pivotal Clustering Concepts Guide

Field personnel can rely on this guide for:

- A clear understanding of what characterizes a certified Pivotal hardware platform
- A review of the three most common topologies with supporting Reference Architecture diagrams
- Pivotal-approved Reference Architecture that includes hardware recommendations and configuration, hard disk guidelines, network layout, installation, data loading, and verification
- Extra guidance with real-world Pivotal cluster examples (see Appendix A: Pivotal Cluster Examples on page 31)

Scope— Pivotal Clustering Concepts Guide

This document does:

- Provide recommendations for building a well-performing Pivotal cluster using the hardware guidelines presented
- Provide general concepts, but does not include specific tuning suggestions

This document does not:

- Promise Pivotal support for the use of third party hardware
- Imply that a good platform for Greenplum Database is a good platform for Pivotal Hadoop (and vice versa)
- Assume that the information herein applies to every site, but is subject to modification depending on a customer's specific local requirements
- Provide all-inclusive procedures for configuring Greenplum Database and Pivotal Hadoop. A subset of information is included as it pertains to deploying a Pivotal cluster.

Feedback and Quarterly Updates

Please send feedback and/or questions regarding content directly to jsargent@gopivotal.com.

Key Points for Review

What Is a Pivotal Engineering Approved Reference Architecture?

This approved Pivotal Reference Architecture comprises third party hardware tested with Pivotal software products. Pivotal maintains examples of current, approved Reference Architectures for support purposes, and as such is prepared to assist customers with cluster diagnostics and configuration assistance. Pivotal does not perform hardware replacement nor is Pivotal a substitute for the OEM vendor support for these configurations.

Note: Pivotal Engineering generally tests certified Reference Architectures with both Greenplum Database and Pivotal Hadoop. Contact platform-engineering@gopivotal.com with specific questions.

Why install on a non-Pivotal Appliance Platform?

The Pivotal Appliance strives to achieve the best balance between performance and cost while meeting a broad range of customer needs. There are some very valid reasons why customers may opt to design their own clusters.

Some possibilities include:

- Varying workload profiles that may require more memory or higher processor capacity
- Specific functional needs like virtualization, increased density, or Disaster Recover (DR)
- Support for radically different network topologies
- Deeper, more direct access for hardware and OS management

RECOMMENDATION FOR FIELD»
Pivotal engineering highly recommends implementing a Pivotal approved reference configuration if customers opt out of using the appliance. Customers achieve much greater flexibility with a reference configuration. For example, adding or removing devices is acceptable as long as the core functionality of the cluster remains intact.

Characteristics of a Supported Pivotal Hardware Platform

Commodity Hardware

This Pivotal-approved Reference Architecture allows customers to take advantage of the inexpensive, yet powerful commodity hardware that includes x86_64 platform commodity servers, storage, and Ethernet switches.

Pivotal recommends:

- **Chipsets or hardware used on other platforms**
    - NIC chipsets (like some of the Intel series)
    - RAID controllers (like StorageWorks)
- **Reference Motherboards/Designs**
    - There is a preference for machines that use reference motherboard implementations.
    - Although DIMM count is important, if a manufacturer integrates more DIMM slots than the CPU manufacturer qualifies; this places more risk on the platform.
- **Ethernet based interconnects (10 Gb) are:**
    - Highly preferred to proprietary interconnects.
    - Highly preferred to storage fabrics.

Manageability

Pivotal recommends:

- Remote, out-of-band management capability with support for ssh connectivity as well as web-based console access and virtual media
- Diagnostic LEDs that convey failure information. Amber lights are a minimum but an LED that displays the exact failure is more useful.
- Tool-free maintenance (i.e. can the cover be opened without tools, are parts hot swappable without tools, etc.)
- Labeling – Are components labeled (such as DIMMs) so it's easy to determine what part needs to be replaced
- Command-line, script-based interfaces for configuring the server BIOS, and options like RAID cards and NICs

Redundancy

Pivotal recommends:

- Redundant hot-swappable power supplies
- Redundant hot-swappable fans

Power and Density

Pivotal recommends:

- Less than 750W of power per 2Us of rack space
- 1 TB of usable space, per 1U of rack space

Note: In general, higher density and lower power work best.

Expansion

Pivotal recommends:

- Three or more PCI-e x8 expansion slots
- Onboard NIC count

RECOMMENDATION FOR FIELD»
For example, integrating more built in NICs is always useful since it can reduce the need for NICs via expansion cards.

Upgrade Options

Ask customers these questions:

- Are any of the internal components upgradeable?
- Does it make sense to move to a new CPU series while continuing to leverage the existing investment (disks, chassis, etc.)?

RECOMMENDATION FOR FIELD»
Great examples of this include Thumper or Thor upgrade modules.

Determining the Best Topology for a Customer's Needs

Traditional Topology

This configuration requires the least specific networking skills, and is the simplest possible configuration. In a traditional network topology, every server in the cluster is directly connected to every switch in the cluster. This is typically implemented over 10 Gb. This topology limits the cluster size to the number of ports on the selected interconnect switch.

For example, if the Cisco Catalyst 4948 is used, the cluster size will be limited to 46 segment nodes (45 is more typical to leave a port open for switch management). This is because the C4948 has 48 ports and the masters each need one port on each switch.

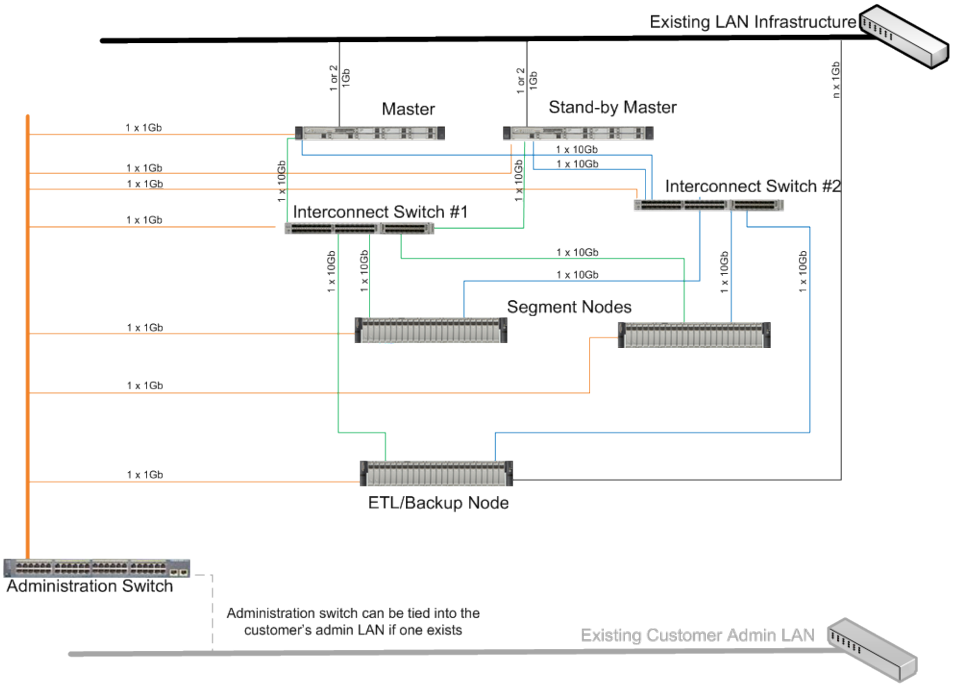

Figure 1. Reference Architecture Example 1 (Typical Topology)

Scalable Topology

Scalable networks implement a network core that allow the cluster to grow beyond the number of ports in the interconnect switches. Care must be taken to ensure that the number of links from the in-rack switches is adequate to service the core.

IMPORTANT » The switch port counts and the link count between the core and the in-rack switches dictates the maximum size of the cluster.

How to Determine the Maximum Number of Servers

For example, if thirty two port switches are used and four links per switch are required between the in-rack switches and the core, the maximum number of servers is determined by the following formula:

```
max-nodes = (nodes-per-rack * (core-switch-port-count / 4))
```

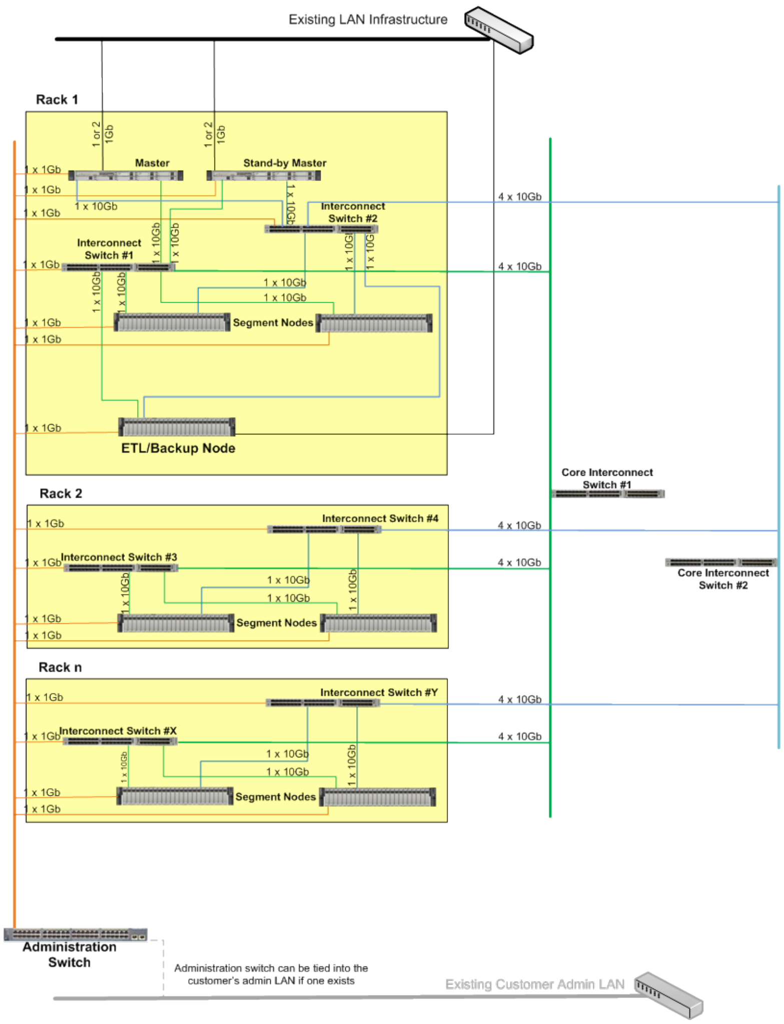

Figure 2. Reference Architecture Example 2 (Scalable Topology)

Fault-tolerant Topology

Fault tolerant networks are an expansion of scalable networks. In these topologies, the core switches and the in-rack switches are tied together such that the loss of any one link or any one switch has no effect on the network connectivity of any server in the cluster.

The maximum node count is the same here as for scalable configurations with one additional limitation. Typically peer links are required between the switches in this sort of configuration. These are connections between the core switches and between the pairs of in-rack switches that allow the switches to collectively managed linked ports across switches.

The number of ports needed to service these peer links must be accounted for in the plan. Usually, it is two or four links per pair of switches. This means that the cores port count is effectively reduced by two or four ports. Also, that the maximum number of nodes in a rack that can be connected is reduced by two or four. Using 32-port switches means that the core port count drops to 28 or 30 (depending on the number of peer links) and the max node count per rack to 24 or 26.

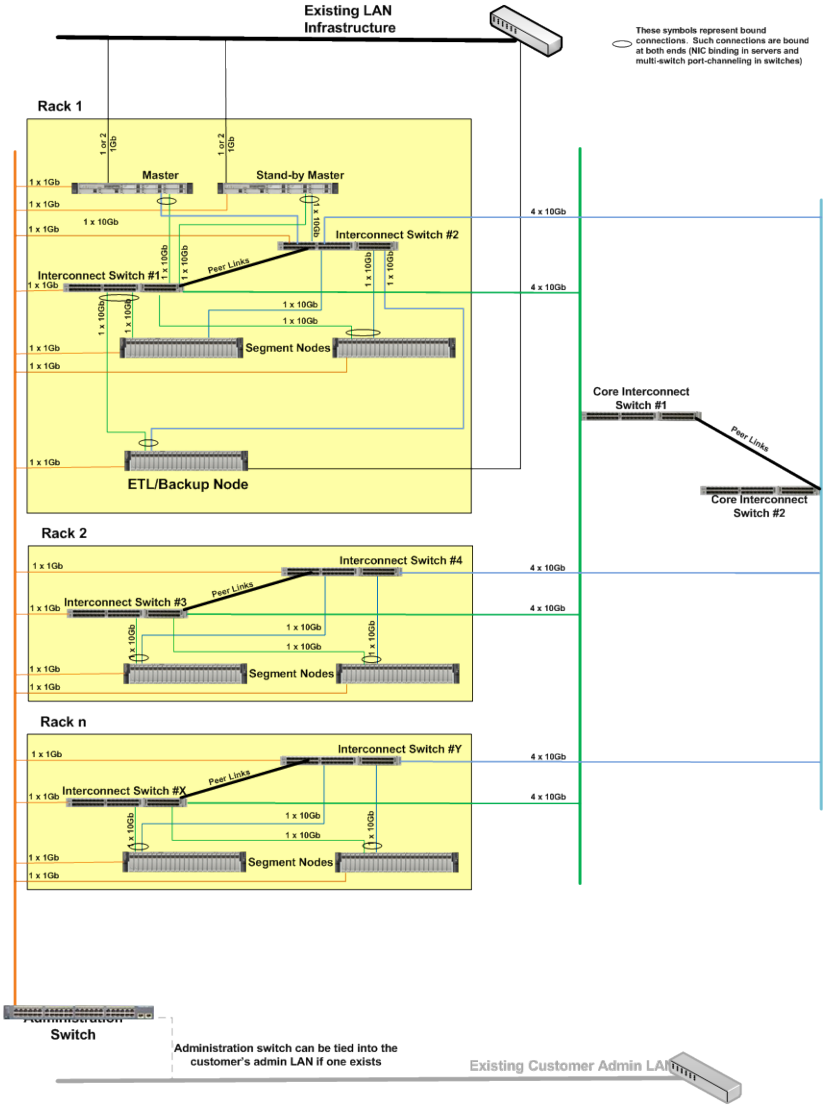

Figure 3. Reference Architecture Example 3 (Fault-tolerant Topology)

Pivotal approved Reference Architecture

Table 1 lists minimum requirements for a good cluster. Use gpcheckperf to generate these metrics.

Note: See Appendix C: Using gpcheckperf to Validate Disk and Network Performance on page 35 for example gpcheckperf output.

Minimum Server Guidelines

Table 1. Baseline numbers for a Pivotal cluster

Server Purpose	Processor	RAM	Storage Bandwidth	Network Bandwidth	Typical Case Size
GPDB					
Master Nodes (mdw & smdw) Users and applications connect to masters to submit queries and return results. Typically, monitoring and management of the cluster and the DB is performed through the master nodes.	Core count less important than clock speed	64 GB or more	> 600 MB/s read > 500 MB/s write	2x10 Gb NICs Multiple NICs	1U
Segment Nodes (sdw) Segment nodes store data and execute queries. They are generally node public facing.	Core count less important than clock speed	64 GB or more	> 900 MB/s read > 1000 MB/s write	2x10 Gb NICs Multiple NICs	2U
ETL/Backup Nodes (etl) Generally identical to segment nodes. These are used as staging areas for loading data or as destinations for backup data.	Core count less important than clock speed	64 GB or more	> 900 MB/s read > 1000 MB/s write	2x10 Gb NICs Multiple NICs	2U
GPHD and PHD					
Master Nodes Used for name nodes, job tracker, and other Pivotal Hadoop services **Note:** The above does not include all possible services. Refer to the appropriate Hadoop documentation for a complete list.	Core count less important than clock speed	128 GB or more	Capacity matters more than bandwidth	1x10 Gb NICs	1U
Worker Nodes Used to present hdfs and to perform compute jobs **Note:** The above does not include all possible services. Refer to the appropriate Hadoop documentation for a complete list.	Clock speed less important than core count	64 GB or more	> 900 MB/s read > 1000 MB/s write	1x10 Gb NICs	1U or 2U
Compute Nodes Compute nodes are used when hdfs is stored via an external hardware resource such as Isilon. These servers handle the compute work load only.	Clock speed less important than core count	64 GB or more	Minimal local storage required.	1x10 Gb NICs	1U or 2U
Other Other servers may be useful depending on implementation. These are generally used as access points for users submitting jobs.	No special processing capability is required.	64 GB or more	Storage bandwidth less important than capacity	1x10 Gb NICs	2U

Network Guidelines

Table 2. Administration and Interconnect Switches

Switch Purpose	Port Count	Port Type	Description
Administration Network Administration networks are used to tie together lights out management interfaces in the cluster and provide a management route into server and OS switches.	48	1 Gb	A layer-2/layer-3 managed switch per rack with no specific bandwidth or blocking requirements
Interconnect Network	48	10 Gb	Two layer-2/layer-3 managed switches per rack. These must have full bandwidth (i.e. all ports can operate at line rate) and be non-blocking (all ports equal with respect to any other port).

Physical Guidelines

Table 3. Racking, Power, and Density

Physical Element	Description
Racking	Generally, a 40U or larger rack that is 1200mm deep is required. Built in cable management is preferred. ESM protective doors are also preferred
Power	The typical input power for a GPDB or PHD rack is 4 x 208/220Volt, 30Amp, Single Phase circuits in the US. Internationally, 4 x 230Volt, 32Amp, Single phase circuits are generally used. This affords a power budget of ~9600VA of fully redundant power. Other power configurations are absolutely fine so long as there is enough energy delivered to the rack to accommodate the contents of the rack in a fully redundant manner.
Density	Typically, 16, 2U servers are accommodated per rack. Each of these servers must have a maximum draw under 550VA to allow for switches and master nodes where applicable in a 9600VA power budget. Up to 22, 1U servers is an alternate configuration (used primarily for Pivotal HD Compute Nodes). These nodes need to remain at near 400VA max draw per node to fit into the 9600VA budget and account for switches and other incidental devices in the rack.

Node Guidelines

Nodes must meet the following criteria.

OS Levels

At a minimum the following Operating Systems (OS) are supported.

- Redhat Linux 5
- Redhat Linux 6
- SUSE Enterprise Linux 10.2 or 10.3
- SUSE Enterprise Linux 11

IMPORTANT» Pivotal Hadoop requires RHEL6.

Setting Greenplum Database OS Parameters

Careful consideration must be given when setting Greenplum Database OS parameters. Refer to the latest released version of the *Greenplum Database Installation Guide* for these settings.

Greenplum Database Server Guidelines

Greenplum Database integrates three kinds of servers: Master servers, Segment servers, and ETL servers. Greenplum Database servers must meet the following criteria.

Master Servers

- Typically a 1U server
- Same processors, RAM, RAID card, and interconnect NICs as the segment servers
- Six to ten disks (eight is most common) organized into a single RAID5 group with one hot spare configured
- SAS 15k disks are preferred with 10k disks a close second
- All disks must be the same size and type
- Draw around 400W maximum
- Should be capable of read rates in gpcheckperf of 500 MB/s or higher (the faster the master scans, the faster it can generate query plans which improves overall performance)
- Should be capable of write rates in `gpcheckperf` of 500 MB/s or higher
- Should include enough additional network interfaces to be able to connect it to the customer network directly in the manner desired by the customer

Segment Servers

- Typically a 2U server
- The fastest available processors
- 64 GB RAM or more
- One or two RAID cards with maximum cache and cache protection (flash or capacitors preferred over battery)
- 2 x 10 Gb
- 12 to 24 disks organized into a RAID5 groups of six to eight disks with no hot spares configured (unless there are available disks after the RAID groups are constructed)

- SAS 15k disks are preferred with 10k disks a close second. SATA disks are preferred over Nearline SAS if SAS 15k or SAS 10k cannot be used. All disks must be the same size and type.
- Draw around 600W maximum
- A minimum read rate in `gpcheckperf` of 900 MBPS or higher (2000 MB/s is typical)
- A minimum write rate in `gpcheckperf` of 1,000 MB/s or higher (1200 MB/s is typical)

ETL Servers

- Typically a 2U server
- The same processors, RAM, and interconnect NICs as the Segment servers
- One or two RAID cards with maximum cache and cache protection (flash or capacitors preferred over battery)
- 12 to 24 disks organized into a RAID5 groups of six to eight disks with no hot spares configured (unless there are available disks after the RAID groups are constructed)
- SATA disks are a good choice for ETL as performance is less of a concern than storage for these systems typically
- Draw around 600W maximum
- Should be capable of read rates in `gpcheckperf` of 100 MB/s or higher (the faster the ETL servers scan, the faster queries data can be loaded)
- Should be capable of write rates in `gpcheckperf` of 500 MB/s or higher (the faster ETL servers write, the faster data can be staged for loading)

Additional Tips for Selecting ETL servers

ETL nodes can be any server that offers enough storage and performance to accomplish the tasks required. Typically, between 4 and 8 ETL servers are required per cluster. The maximum number is dependent on the desired load performance and the size of the Greenplum Database cluster.

For example, the larger the Greenplum Database cluster, the faster the loads can be. The more ETL servers, the faster data can be served. Having more ETL bandwidth than the cluster can receive is pointless. Having much less ETL bandwidth than the cluster can receive makes for slower loading than the maximum possible.

Pivotal HD Server Guidelines

Master Servers

- Typically a 1U server
- The same processors and interconnect NICs as the Worker Nodes
- As much RAM as possible (64 GB minimum)
- One RAID card with maximum cache and cache protection (flash or capacitors preferred over battery)
- Six to 10 disks (eight is most common) organized into a single RAID5 group with one hot spare configured
- Draw around 400W maximum per server
- Should include enough additional network interfaces to be able to connect it to the customer network directly in the manner desired by the customer

Worker Nodes

- Typically a 2U server
- The fastest available processors
- 64 GB RAM or more
- 2x10 Gb
- The storage controller should be able to configure RAID 1 mirrored pairs. It does not need to be write protected but, if it's not, it should not have cache.
- 12 to 24 disks organized into a one RAID1 pair and the remaining JBOD. The RAID1 pair will be for the operating system and swap.
- SATA disks are a good choice for Hadoop worker nodes
- Draw around 600W maximum

Pivotal Hadoop Manager Admin Node

The Pivotal Hadoop Manager Admin node is separate from cluster nodes especially if cluster size is greater than 15 to 20 nodes. The minimum hardware requirements are as follows

- 1 Quad code CPU
- 4 to 8 GB RAM
- 2x2 TB SATA disks
- 1GbE network connectivity

Additional Tips for Selecting Pivotal Hadoop Servers

Select the cluster node hardware based on the resource requirements of the analytics workload, and overall need for data storage. It is hard to anticipate the workload that may run on the cluster and so designing for a specific type of workload may lead to under-utilized hardware resources.

Pivotal recommends selecting the hardware for a balanced workload across different types of system resources. But also have the ability to provision more specific resources such as CPU, I/O bandwidth, and memory as the workload evolves over time.

Hardware and capacity requirements for cluster nodes may vary depending upon what service roles run on them. Typically failure of cluster slave nodes is tolerated by PHD services but disruptive if a master node fails. This can cause service availability issues. For these reasons it's important to provide more reliable hardware for master nodes (such as NameNode, YARN Resource manager, HAWQ master) to ensure higher cluster availability.

Deploying Hadoop Services

A test cluster usually comprises 3 to 5 nodes. Provision one of these nodes to run all of the services. The remaining nodes should be considered data nodes. When deploying a production cluster with more nodes, here are some guidelines for better performance, availability, and use:

Hadoop services Master roles: For example, HDFS NameNode, YARN ResourceManager and History Server, HBase Master, HAWQ Master, USS Namenode. These should reside on separate nodes. These services and roles require dedicated resources since they communicate directly with Hadoop client applications. Running Hadoop slave application tasks (map/reduce tasks) on the same node interferes with master resource requirements.

Hadoop services slave roles: For example, HDFS DataNode, YARN NodeManager, HBase RegionServer, HAWQ SegmentServer. These should reside on the cluster slave nodes. This helps provide optimal data access as well as better hardware use.

HBase requires Zookeeper: Zookeeper should have an odd number of zookeeper servers. This application does not need dedicated nodes and can reside on the master server with ~ 1 GB RAM and dedicated disk with ~ 1 TB of space.

Note: A minimum of three to five Zookeepers in production deployed on separate racks can support up to 1-2K nodes.

Hadoop Clients: For example, Hive, Pig etc. These install on separate gateway nodes depending on multi-user application requirements.

Hard Disk Configuration Guidelines

A generic server with 24 hot-swappable disks can have several potential disk configurations. Testing completed by Pivotal Platform and Systems Engineering shows that the best performing storage for supporting Pivotal software is:

- Four, RAID5 groups of six disks each (used as four filesystems), or
- Combined into one or two filesystems using logical volume manager

The following instructions describe how to build the recommended RAID groups and virtual disks for both master and segment nodes. How these ultimately translate into filesystems is covered in the relevant operating system Install Guide.

LUN Configuration

The RAID controller settings and disk configuration are based on synthetic load testing performed on several RAID configurations. Unfortunately, the settings that resulted in the best read rates did not have the highest write rates and the settings with the best write rates did not have the highest read rates.

The prescribed settings selected offer a compromise. Test results for these settings showed all the rates higher than the low side corresponding to the highest rates. In other words, these settings result in write rates lower than the best measured write rate but higher than the write rates associated with the settings for the highest read rate. The same is true for read rates. This is intended to ensure that both input and output are the best they can be while affecting the other the least amount possible.

Master Node

Master nodes have eight, hot swappable disks. Configure all eight of these into a single, RAID5 stripe set. Each of the virtual disks that carved from this disk group will use the following properties:

- 256k stripe width
- No Read-ahead
- Disk Cache Disabled
- Direct I/O

Virtual disks are configured in the RAID card's optional ROM. Each virtual disk defined in the RAID card will appear to be a disk in the operating system with a /dev/sd? device file name.

Segment and ETL Nodes

Segment nodes have 24, hot swappable disks. These can be configured in a number of ways but Pivotal recommends four, RAID 5 groups of six disks each (RAID5, 5+1). Each of the virtual disks that will be carved from these disk groups should use the following properties:

- 256k stripe width
- No Read-ahead
- Disk Cache Disabled
- Direct I/O

Virtual disks are configured in the RAID card's optional ROM. Each virtual disk defined in the RAID card will appear to be a disk in the operating system with a /dev/sd? device file name.

SAN/JBOD Storage

In some configurations it may be a requirement to connect up an external storage array due to the database size or server type being used by the customer. With this in mind it is important to set expectations upfront with the customer that SAN and JBOD storage will not perform as well as local internal server storage based on testing done internally within Pivotal Platform and Systems Engineering.

IMPORTANT » Some considerations to be taken into account if installing or sizing such a configuration are the following independent of the vendor of choice:

- Know database size and estimated growth over time
- Know the customer read/write % ratio
- Large block I/O is the predominant workload (512KB)
- Disk type and preferred RAID type based on the vendor of choice
- Expected disk throughput based on read and write
- Response time of the disks/JBOD controller
- Preferred option is to have BBU capability on either the RAID card or controller
- Redundancy in switch zoning, preferably with a fan in:out 2:1
- At least 8 GB Fibre Channel (FC) connectivity
- Ensure that the server used supports the use of FC, FCoE , or external RAID cards

In all instances where an external storage source is being utilized the vendor of the disk array/JBOD should be consulted to obtain specific recommendations based on a sequential workload. This may also require the customer to obtain additional licenses from the pertinent vendors.

Network Layout Guidelines

General Recommendations

All the systems in a Pivotal cluster need to be tied together in some form of high-speed data interconnect. The general rule of thumb for a Pivotal cluster is 20 percent of the maximum, theoretical I/O scan rate of each segment node will be transmitted over the interconnect. This clearly depends on workload but, for planning purposes, this number ensures that the cluster will be well served by the data interconnect.

Greenplum Database: A minimum of two 10 Gb NICs are used as a dedicated network interconnect between clustered servers. This is the path used for loading data and for accessing systems during queries. It should be as high speed and as low latency as possible and should not be used for any other purpose (i.e. not part of the general LAN).

Pivotal Hadoop: A minimum of one, 10 Gb NIC is recommended. In addition, two NICs bonded together, when used with switches that support multi-switch LAGs, is highly recommended. This is used as part of a dedicated network between the Hadoop cluster nodes. It may also be used as the interconnect for a Greenplum Database cluster. In this case, Hadoop is leveraged to load data for a more in-depth analysis than is available on Hadoop itself.

10 Gb networking is recommended. In general, Cisco, Brocade, and Arista switches are good choices as these brands include the ability to tie switches together in fabrics. Together with NIC bonding on the servers, this approach eliminates single points of failure in the interconnect networks Intel, QLogic, or Emulex based network interfaces tend to work best. Layer 3 capability is recommended. Layer 3 integrates many features that could be useful in a Greenplum Database or Pivotal Hadoop environment as well.

Note: The vendor hardware referenced above is strictly used as an example. Pivotal Platform and Systems Engineering is not specifying which product to use in the network.

FCoE switch support is also required if SAN storage is being utilized as well as support for Fibre snooping (FIPS).

Network Topology

A rule of thumb for network utilization is to plan for up to twenty percent of each server's maximum I/O read bandwidth as network traffic. This means a server with a measured 1500MB/s read bandwidth (as measured by `gpcheckperf`) might be expected to transmit 300MB/s. Previous Greenplum best practices required four, 1Gb interfaces for the interconnect to accommodate this load.

Current best practice suggests two, 10Gb interfaces for the cluster interconnect. This ensures that there is bandwidth to grow into, and reduces cabling in the racks. Using 10Gb also accommodates compression. Pivotal applications compress data on disk but uncompress it before transmitting to other systems in the cluster. This means that a 1500 MB/s read rate with a 4x compression ratio results in a 6000 MB/s effective read rate. Twenty percent of 6000 MB/s is 1200 MB/s which is more than a single 10Gb interface's bandwidth.

Network Connections in a Greenplum Database Cluster

A Greenplum Database cluster uses three kinds of network connections:

- Admin networks
- Interconnect networks
- External networks

Admin Networks

An Admin network ties together all the management interfaces for the devices in a configuration. This is generally used to provide monitoring and out-of-band console access for each connected device. The admin network is typically a 1 Gb network physically and logically distinct from other networks in the cluster.

Servers are typically configured such that the out-of-band or lights out management interfaces share the first network interface on each server. In this way, the same physical network provides access to lights out management and an operating system level connection useful for network OS installation, patch distribution, monitoring, and emergency access.

Switch Types

- Typically one 24 or 48-port, 1 Gb switch per rack and one additional 48-port switch cluster as a core
- Any 1 Gb switch can be used for the Admin network. Careful planning is required to ensure that a network topology is designed to provide enough connections and the features desired by the site to provide the kind of access required.

Cables

Use either cat5e or cat6 cabling for the Admin network. Cable the lights out or management interface from each cluster device to the Admin network. Place an Admin switch in each rack and cross-connect the switches rather than attempting to run cables from a central switch to all racks.

Note: Pivotal recommends using a different color cable for the Admin network (other network cables are typically orange).

Interconnect Networks

The Interconnect network ties the servers in the cluster together and forms a high-speed, low contention data connection between the servers. This should not be implemented on the general data center network as Greenplum Database interconnect traffic will tend to overwhelm networks from time to time. Low latency is needed to ensure proper functioning of the Greenplum Database cluster. Sharing the interconnect with a general network will tend to introduce instability into the cluster.

Typically two switches are required per rack, and two more to act as a core. Use two 10 Gb cables per server and eight per rack to connect the rack to the core.

Interconnect networks are often connected to general networks in limited ways to facilitate data loading. In these cases, it is important to shield both the interconnect network and the general network from the Greenplum Database traffic and visa-versa. Use a router or an appropriate VLAN configuration to accomplish this.

External Network Connections

The master nodes are connected to the general customer network to allow users and applications in to submit queries. Typically, this is done over a small number of 1 Gb connections attached to the master nodes. Any method that affords network connectivity from the users and applications needing access to the master nodes is acceptable.

Installation Guidelines

Each configuration requires a specific rack plan. There are single and multi-rack configurations determined by the number of servers present in the configuration. A single rack configuration is one where all the planned equipment fits into one rack. Multi-rack configurations require two or more racks to accommodate all the planned equipment.

Racking Guidelines for a 42U Rack

Consider the following if installing the cluster in a 42U rack.

- Prior to racking any hardware, perform a site survey to determine what power option is desired, if power cables will be top or bottom of the rack, and whether network switches and patch panels will be top or bottom of the rack.
- Install the KMM tray into tack unit 19.
- Install the interconnect switches into rack units 21 and 22 leaving a one unit gap above the KMM tray.
- Rack segment nodes up from first available rack unit at the bottom of the rack (see multi-rack rules for variations using low rack units).
- Install no more than sixteen 2U servers (excludes master but includes segment, and ETL nodes).
- Install the master node into rack unit 17. Install the stand-by master into rack unit 18.
- Admin switches can be racked anywhere in the rack though the top is typically the best and simplest location.
- All computers, switches, arrays, and racks should be labeled on both the front and back.
- All computers, switches, arrays, and racks should be labeled as described in the section on labels later in this document.
- All installed devices should be connected to two or more power distribution units (PDUs) in the rack where the device is installed.

When installing a multi-rack cluster:

- Install the interconnect core switches in the top two rack units if the cables come in from the top or in the bottom two rack units if the cables come in from the bottom.
- Do not install core switches in the master rack

Cabling

The number of cables required varies according to the options selected. In general, each server and switch installed will use one cable for the Admin network. Run cables according to established cabling standards. Eliminate tight bends or crimps. Clearly label all at each end. The label on each end of the cable must trace the path the cable follows between server and switch. This includes:

- Switch name and port
- Patch panel name and port if applicable
- Server name and port

Switch Configuration Guidelines

Typically, the factory default configuration is sufficient. Contact platform-engineering@gopivotal.com with specific questions regarding switches. The next quarterly update of this guide will include recommendations.

IP Addressing Guidelines

IP Addressing Scheme for each Server and CIMC on the Admin Network

Note: Pivotal's recommended IP address for servers on the Admin network uses a standard internal address space and is extensible to include over 1,000 nodes.

RECOMMENDATION FOR FIELD»
All Admin network switches present should be cross connected and all NICs attached to these switches participate in the 172.254.0.0/16 network.

Table 4. IP Addresses for Servers and CIMC

Host Type	Network Interface	IP Address
Primary Master Node	CIMC Etho	172.254.1.252/16 192.254.1.250/16
Secondary Master Node	CIMC Etho	172.254.1.253/16 172.254.1.251/16
Non-master Nodes in rack 1 (master rack)	CIMC Etho	172.254.1.101/16 through 172.254.1.116/16 172.254.1.1/16 through 172.254.1.16/16
Non-master Segment Nodes in rack 2	CIMC Etho	172.254.2.101/16 through 172.254.2.116/16 172.254.2.1/16 through 172.254.2.16/16
Non-master Segment Nodes in rack #	CIMC Etho	172.254.#.101/16 through 172.254.#.116/16 172.254.#.1/16 through 172.254.#.16/16

Note: Where **#** is the rack number.

- The fourth octet is counted from the bottom up. For example, the bottom server in the first rack is 172.254.1.1 and the top, excluding masters, is 172.254.1.16.
- The bottom server in the second rack is 172.254.2.1 and top 172.254.2.16. This continues for each rack in the cluster regardless of individual server purpose.

IP Addressing for Non-server Devices

Table 5 lists the correct IP addressing for each non-server device.

Table 5. Non-server IP Addresses

Device	IP Address
First Interconnect Switch in Rack	*172.254.#.201/16
Second Interconnect Switch in Rack	*172.254.#.202/16

* Where **#** is the rack number

IP Addressing for Interconnects using 10 Gb NICs

Table 6 lists the correct IP address for traditional or scalable interconnects based on 10 Gb NICs.

Table 6. Interconnect IP Addressing for 10 Gb NICS

Host Type	Physical RJ-45 Port	IP Address
Primary Master	1st port on PCIe card 2nd port on PCIe card	172.1.1.250/16 172.2.1.250/16
Secondary Master	1st port on PCIe card 2nd port on PCIe card	172.1.1.251/16 172.2.1.251/16
Non-Master Nodes	1st port on PCIe card 2nd port on PCIe card	172.1.#.1/16 through 172.1.#.16/16 172.2.#.1/16 through 172.2.#.16/16

Note: Where **#** is the rack number:

- The fourth octet is counted from the bottom up. For example, the bottom server in the first rack uses 172.1.1.1 and 172.2.1.1.
- The top server in the first rack, excluding masters, uses 172.1.1.16 and 172.2.1.16.

Each NIC on the interconnect uses a different subnet and each server has a NIC on each subnet.

IP Addressing for Fault Tolerant Interconnects

Table 7 lists correct IP addresses for fault tolerant interconnects regardless of bandwidth.

Table 7. IP Addresses for Fault Tolerant Interconnects

Host Type	Physical RJ-45 Port	IP Address
Primary Master	1st port on PCIe card	172.1.1.250/16
Secondary Master	1st port on PCIe card	172.1.1.251/16
Non-Master Nodes	1st port on PCIe card	172.1.#.1/16 through 172.1.#.16/16

Note: Where **#** is the rack number:

- The fourth octet is counted from the bottom up. For example, the bottom server in the first rack uses 172.1.1.1. The top server in the first rack, excluding masters, uses 172.1.1.16.
- The NICs connected to the interconnects are bonded into active-active bond configurations and connected to interconnect switches using multi-switch Link Aggregation Groups.

Data Loading Connectivity Guidelines

High-speed data loading requires direct access to the segment nodes, bypassing the masters. There are three ways to connect a Pivotal cluster to external data sources or backup targets:

- **VLAN Overlay**—The first and recommended best practice is to use virtual LANs (VLANS) to open up specific hosts in the customer network and the Greenplum Database cluster to each other.
- **Direct Connect to Customer Network**—*Only use if there is a specific customer requirement.*
- **Routing**—*Only use if there is a specific customer requirement.*

VLAN Overlay

VLAN overlay is the most commonly used method to provide access to external data without introducing network problems. The VLAN overlay imposes an additional VLAN on the connections of a subset of Pivotal cluster servers.

How the VLAN Overlay Method Works

Using the VLAN Overlay method, traffic passes between the cluster servers on the internal VLAN, but cannot pass out of the internal switch fabric because the external facing ports are assigned only to the overlay VLAN. Traffic on the overlay VLAN (traffic to or from IP addresses assigned to the relevant servers' virtual network interfaces) can pass in and out of the cluster.

This VLAN configuration allows multiple clusters to co-exist without requiring any change to their internal IP addresses. This gives customers more control over what elements of the clusters are exposed to the general customer network. The Overlay VLAN can be a dedicated VLAN and include only those servers that need to talk to each other; or the Overlay VLAN can be the customer's full network.

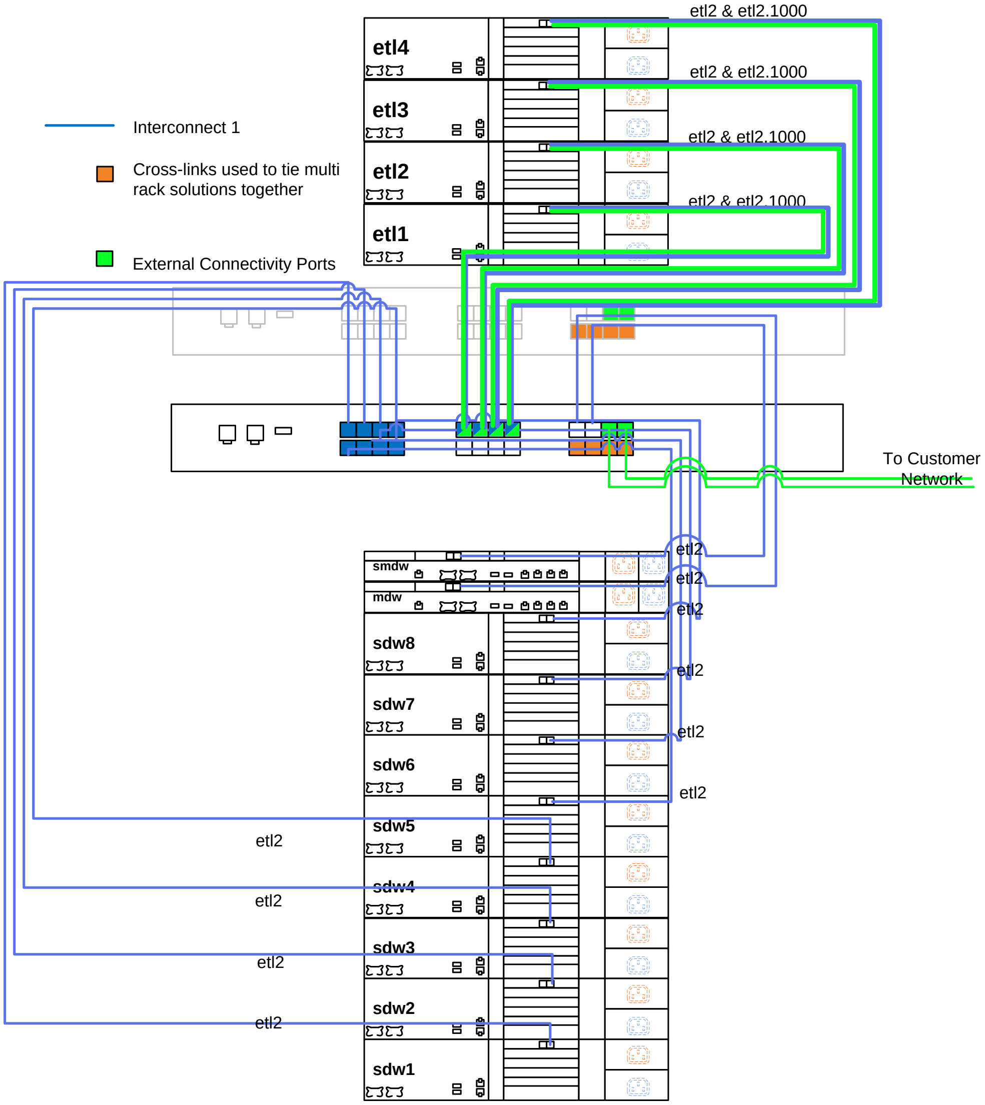

Figure 4. Basic VLAN Overlay Example: Configured with One Interconnect

Figure 4 shows a cluster with segment nodes (sdw1 – sdw8 + mdw and smdw) and four ETL nodes (etl1 – etl4). In this case, only the etl nodes are part of the overlay. It is not a requirement to have ETL nodes to use the overlay. Any of the servers in this rack or any rack of any other configuration may participate in the overlay.

The overlay in **Figure 4** is configured against the first interconnect. It is possible to configure two overlays, one on each interconnect, if more than two incoming ports are desired. These can be configured in a link aggregation group or not. It will depend on whether the customer wants port fault tolerance or not and whether the customer can configure the corresponding LAG on their side of the connection.

These can be configured as switchport mode trunk or access. It will depend on whether the customer wants the VLAN tag to be sent to their side or not. These ports can either filter or forward BPDU for spanning tree. In most cases, filter is most appropriate.

Configuring the Overlay VLAN—An Overview

Configuring the VLAN involves three steps:

1. Virtual interface tags packets with the overlay VLAN
2. Configure the switch in the cluster with the overlay VLAN
3. Configure the ports on the switch connecting to the customer network

Step 1 – Virtual interface tags packets with the overlay VLAN

Each server that is both in the base VLAN and the overlay VLAN has a virtual interface created that tags packets sent from the interface with the overlay VLAN. For example, suppose eth2 is the physical interface on an ETL server that is connected to the first interconnect network. To include this server in an overlay VLAN (VLAN 1000 as shown in **Figure 5**), the interface eth2.1000 is created using the same physical port but defining a second interface for the port. The physical port does not tag its packets but any packet sent using the virtual port is tagged with a VLAN (VLAN 1000 as shown in **Figure 5**).

Step 2 – Configure the switch in the cluster with the overlay VLAN

The switch in the cluster that connects to the servers and the customer network is configured with the overlay VLAN. All of the ports connected to servers that will participate in the overlay are changed to switchport mode converged and added to both the internal VLAN (199) and the overlay VLAN (1000) as shown in **Figure 5**.

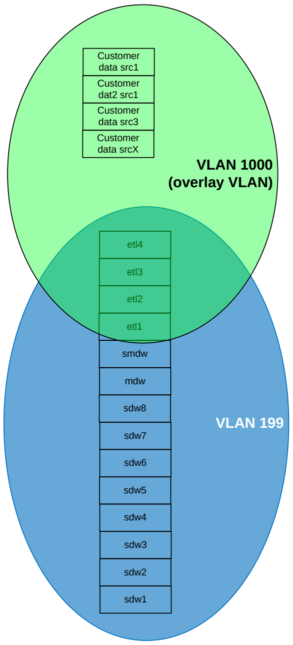

Figure 5. Configuring the VLAN Overlay

Step 3 – Configure the switch ports connected to the customer network

The ports on the switch connecting to the customer network are configured as either access or trunk mode switch ports (depending on customer preference) and added only to the overlay VLAN.

Direct Connect to the Customer's Network

Each node in the Greenplum Database cluster can simply be cabled directly to the network where the data sources exist or a network that can communicate with the source network. This is a brute force approach that works very well. Depending on what network features are desired (redundancy, high bandwidth, etc.) this method can be very expensive in terms of cabling and switch gear as well as space for running large numbers of cables. See **Figure 6.**

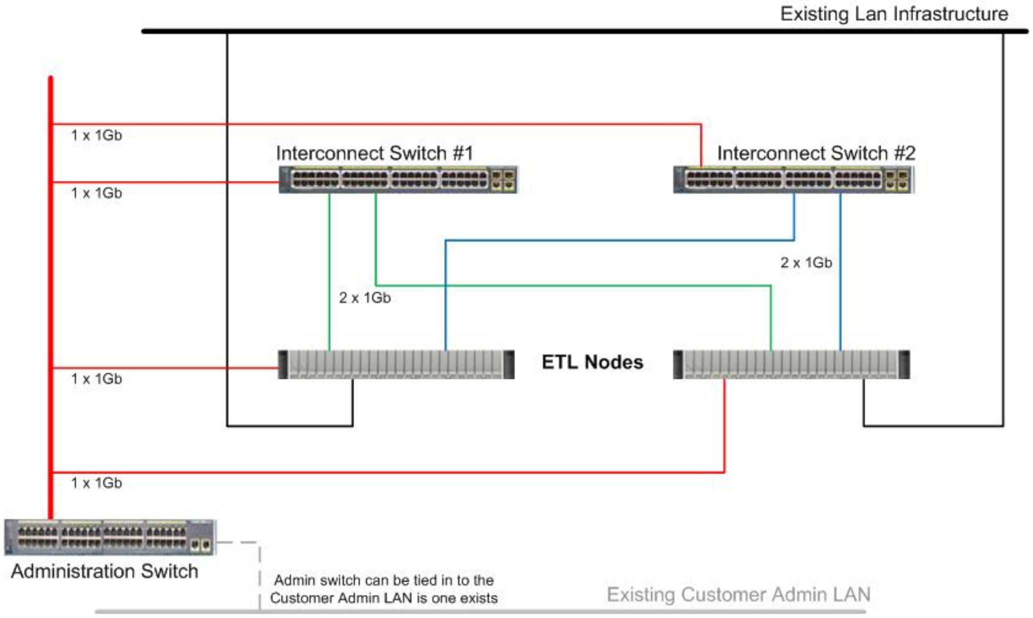
Existing Lan Infrastructure
1 x 1Gb
1 x 1Gb
Interconnect Switch #1
Interconnect Switch #2
2 x 1Gb
2 x 1Gb
ETL Nodes
1 x 1Gb
1 x 1Gb
Administration Switch
Admin switch can be tied in to the Customer Admin LAN is one exists
Existing Customer Admin LAN

Figure 6. Data Loading—Direct Connect to Customer Network

Routing

One way is to use any of the standard networking methods used to link two different networks together. These can be deployed to tie the interconnect network(s) to the data source network(s). Which of these methods is used will depend on the circumstances and the goals for the connection.

A router is installed that advertises the external networks to the servers in the Pivotal cluster. This method could potentially have performance and configuration implications on the customer's network.

Validation Guidelines

Most of the validation effort is performed after the OS is installed and a variety of OS-level tools are available. A checklist is included in the relevant OS installation guide that should be separately printed and signed for delivery and includes the issues raised in this section.

Examine and verify the following items:

- All cables labeled according to the standards in this document
- All racks labeled according to the standards in this document
- All devices power on
- All hot-swappable devices are properly seated
- No devices show any warning or fault lights
- All network management ports are accessible via the administration LAN
- All cables are neatly dressed into the racks and have no sharp bends or crimps
- All rack doors and covers are installed and close properly
- All servers extend and retract without pinching or stretching cables

Labels

Racks
Each rack in a reference architecture is labeled at the top of the rack and on both the front and back. Racks are named Master Rack or Segment Rack # where # is a sequential number starting at 1. A rack label would look like this:

Master Rack

Segment Rack 1

Servers
Each server is labeled on both the front and back of the server. The label should be the base name of the server's network interfaces.

In other words, if a segment node is known as sdw15-1, sdw15-2, sdw15-3, and sdw15-4, the label on that server would be sdw15.

sdw15

Switches
Switches are labeled according to their purpose. Interconnect switches are i-sw, administration switches are a-sw, and ETL switches are e-sw. Each switch is assigned a number starting at 1. Switches are labeled on the front of the switch only since the back is generally not visible when racked.

i-sw-1

Certification Guidelines

Network Performance Test

Verifies the line rate on both 10 Gb NICs

gpcheckperf

Run `gpcheckperf` on the disks and network connections within the cluster. As each certification will vary due to the number of disks, nodes and network bandwidth available the commands to run tests will differ.

See Appendix C: on page 35 for more information on the `gpcheckperf` command.

Retail Demo

The retail demo is an internally developed tool that applies a sample workload to a cluster. The workload includes data ingestion and queries with results from DCAV1 for a basis of comparison.

Hardware Monitoring and Failure Analysis Guidelines

In order to support monitoring of a running cluster the following items should be in place and capable of being monitored with information gathered available via interfaces such as SNMP or IPMI.

Fans/Temp

- Fan status/presence
- Fan speed
- Chassis temp
- CPU temp
- IOH temp

Memory

- DIMM temp
- DIMM status (populated, online)
- DIMM single bit errors
- DIMM double bit errors
- ECC warnings (corrections exceeding threshold)
- ECC correctable errors
- ECC uncorrectable errors
- Memory CRC errors

System Errors

- Post errors
- PCIe fatal errors
- PCIe non-fatal errors
- CPU machine check exception
- Intrusion detection
- Chipset errors

Power

- Power Supply presence
- Power supply failures
- Power supply input voltage
- Power supply amperage
- Motherboard voltage sensors
- System power consumption

Appendix A: Pivotal Cluster Examples

Table 8 through **Table 10** list good choices for cluster hardware based on Intel Sandy Bridge processor-based servers and Cisco switches.

Greenplum Database Cluster Example

Table 8. Hardware Components

Cluster Element	Description
Master Node Two of these nodes per cluster	1U server (similar to the Cisco C220M3): - 2 x E5-2670 processors (2.6GHz, 8 cores, 115W) - 128 GB RAM (8 x 8 GB, 1333 MHz DIMMs) - 1 x LSI 2208-based RAID card w/ 1 GB protected cache - 8 x SAS, 10 k, 6 G disks (typically 8x300 GB, 2.5”) Organized into a single, RAID5 disk group with a hot spare. Logical devices defined as per the OS needs (boot, root, swap, etc..) and the remaining in a single, large filesystem for data - 2 x 10 Gb, Intel or QLogic or Emulex based NICs - Lights out management (IPMI-based BMC) - 2 x 650W or higher, high-efficiency power supplies
Segment Node & ETL Node Up to 16 per rack. No maximum total count	2U server (similar to the Cisco C240M3): - 2 x E5-2670 processors (2.6GHz, 8 cores, 115W) - 128 GB RAM (8 x 8 GB, 1333 MHz DIMMs) - 1 x LSI 2208-based RAID card w/ 1 GB protected cache - 12 to 24 x SAS, 10k, 6G disks (typically 12x600 GB, 3.5” or 24x900 GB, 2.5”) Organized into two to four, RAID5 groups. Used either as two to four data filesystems (with logical devices skimmed off for boot, root, swap, etc.) or as one large device bound with Logical Volume Manager. - 2 x 10 Gb, Intel or QLogic or Emulex based NICs - Lights out management (IPMI-based BMC) - 2 x 650W or higher, high-efficiency power supplies
Admin Switch	Cisco Catalyst 2960S (2960S-48TS-L recommended) A simple, 48-port, 1 GB switch with features that allow it to be easily combined with other switches to expand the network. The least expensive, managed switch with good reliability possible is appropriate for this role. There will be at least one per rack.
Interconnect Switch	Cisco Nexus 55xx/22xx Switches The Nexus switch line allows for multi-switch link aggregation groups (called vPC), easy expansion, and a reliable body of hardware and operating system. The only caveat is that not less than 8 links between Nexus 2232 (10 Gb fabric extenders) and Nexus 55xx switches are required for Pivotal clusters. Less than this restricts network traffic to the degree that the cluster cannot function properly.
Rack	4-post, 42U, 1200mm deep rack with 19” internal rail dimensions (similar to the APC Netshelter SX model AR3305).
Power	4 x 208/220V, 30A, 1P power distribution unit (PDU) similar to APC model AP8841

Pivotal Hadoop Cluster Example

Note: Redundant NICs can be used for the servers in a Hadoop cluster. This has many advantages and creates a stronger, more robust environment. Hadoop is designed around the use of minimal hardware however so a reliable cluster can be created with non-redundant NICs.

Table 9. Hardware Components

Cluster Element	Description
Master Node Minimum four of these nodes per cluster. No maximum (depends on implementation)	1U server (similar to the Cisco C220M3): - 2 x E5-2670 processors (2.6 GHz, 8 cores, 115W) - 128 GB RAM (8 x 8 GB, 1333 MHz DIMMs) - 1 x LSI 2208-based RAID card w/ 1 GB protected cache) - 8 x SAS, 10k, 6G disks (typically 8x300 GB, 2.5”) Organized into a single, RAID5 disk group with a hot spare. Logical devices defined as per the OS needs (boot, root, swap, etc.) and the remaining in a single, large filesystem for data - 2 x 1 Gb, Intel or QLogic or Emulex based NICs or 1x 1 Gb and 1 x 10 Gb, Intel or QLogic or Emulex based NICs (preferred) - Lights out management (IPMI-based BMC) - 2 x 650W or higher, high-efficiency power supplies
Worker Node Up to 16 per rack (max two masters per rack with 16 workers). No maximum total count	2U server (similar to the Cisco C240M3): - 2 x E5-2670 processors (2.6GHz, 8 cores, 115W) - 128 GB RAM (8 x 8 GB, 1333 MHz DIMMs) - 1 x LSI 2008 or LSI 1068-based storage controller - 2 x SAS, 10k, 6g disks (typically 2x300 GB, 3.5”) and 10 x SATA, 5.4k, 3G disks (typically 10x3 TB, 3.5”) Organized into one mirrored pair for the boot, root, swap, and other OS structures (based on the two SAS disks) and individual disks with no RAID configuration for the remaining 10 disks. - 1 x 1 Gb, Intel or QLogic or Emulex based NICs or 1 x 10 Gb, Intel or QLogic or Emulex based NICs (10 Gb preferred) - Lights out management (IPMI-based BMC) - 2 x 650W or higher, high-efficiency power supplies
Admin Switch	Cisco Catalyst 2960S (2960S-48TS-L recommended) A simple, 48-port, 1 Gb switch with features that allow it to be easily combined with other switches to expand the network. The least expensive, managed switch with good reliability possible is appropriate for this role. There will be at least one per rack.
Interconnect Switch	Cisco Nexus 55xx/22xx Switches The Nexus switch line allows for multi-switch link aggregation groups (called vPC), easy expansion, and a reliable body of hardware and operating system. The only caveat is that not less than 8 links between Nexus 2232 (10 GB fabric extenders) and Nexus 55xx switches are required for Pivotal clusters. Less than this restricts network traffic to the degree that the cluster cannot function properly.
Rack	4-post, 42U, 1200 mm deep rack with 19” internal rail dimensions (similar to the APC Netshelter SX model AR3305).
Power	4 x 208/220V, 30A, 1P PDUs (similar to APC model AP8841)

Pivotal Hadoop with External HDFS (Isilon) Cluster Example

Note: Redundant NICs can be used for the servers in a Hadoop cluster. This has many advantages and creates a stronger, more robust environment. Hadoop is designed around the use of minimal hardware however so a reliable cluster can be created with non-redundant NICs.

Table 10. Hardware Components

Cluster Element	Description
Master Node Minimum four of these nodes per cluster. No maximum (depends on implementation)	1U server (similar to the Cisco C220M3): - 2 x E5-2670 processors (2.6GHz, 8 cores, 115W) - 128 GB RAM (8 x 8 GB, 1333 MHz DIMMs) - 1 x LSI 2208-based RAID card w/ 1 GB protected cache) - 8 x SAS, 10 k, 6G disks (typically 8x300 GB, 2.5”) Organized into a single, RAID5 disk group with a hot spare. Logical devices defined as per the OS needs (boot, root, swap, etc.) and the remaining in a single, large filesystem for data - 2 x 1 GB, Intel or QLogic or Emulex based NICs or 1x 1 GB and 1 x 10 GB, Intel or QLogic or Emulex based NICs (preferred) - Lights out management (IPMI-based BMC) - 2 x 650W or higher, high-efficiency power supplies
Worker Node Up to 22 per rack counting masters. No maximum total count	1U server (similar to the Cisco C220M3): - 2 x E5-2670 processors (2.6G Hz, 8 cores, 115W) - 64 GB RAM (8 x 8 GB, 1333 MHz DIMMs) - 1 x LSI 2208-based RAID card w/ 1 GB protected cache) - 4 x SAS, 10k, 6G disks (typically 4x300 GB, 2.5”) Organized into a single, RAID5 disk group with a hot spare. Logical devices defined as per the OS needs (boot, root, swap, etc..) and the remaining in a single, large filesystem for data - 1 x 1 GB, Intel or QLogic or Emulex based NICs or 1 x 10 GB, Intel or QLogic or Emulex based NICs (10 GB preferred) - Light’s out management (IPMI-based BMC) - 2 x 650W or higher, high-efficiency power supplies
Admin Switch	Cisco Catalyst 2960S (2960S-48TS-L recommended) A simple, 48-port, 1 GB switch with features that allow it to be easily combined with other switches to expand the network. The least expensive, managed switch with good reliability possible is appropriate for this role. There will be at least one per rack.
Interconnect Switch	Cisco Nexus 55xx/22xx Switches The Nexus switch line allows for multi-switch link aggregation groups (called vPC), easy expansion, and a reliable body of hardware and operating system. The only caveat is that not less than 8 links between Nexus 2232 (10GB fabric extenders) and Nexus 55xx switches are required for Pivotal clusters. Less than this restricts network traffic to the degree that the cluster cannot function properly.
Rack	4-post, 42U, 1200mm deep rack with 19” internal rail dimensions (similar to the APC Netshelter SX model AR3305).
Power	4 x 208/220V, 30A, 1P PDUs (similar to APC model AP8841)

Performance Difference on a Single RAID5 vs JBOD

Additionally on the Worker Node, there's a performance difference on a single RAID5 versus JBOD. The map-reduce shuffle output is written to local disk usually in a directory called "local." The more local disk a customer has, the faster it can write.

Appendix B: Example Rack Layout

Figure 7 shows an example rack layout with proper switch and server placement.

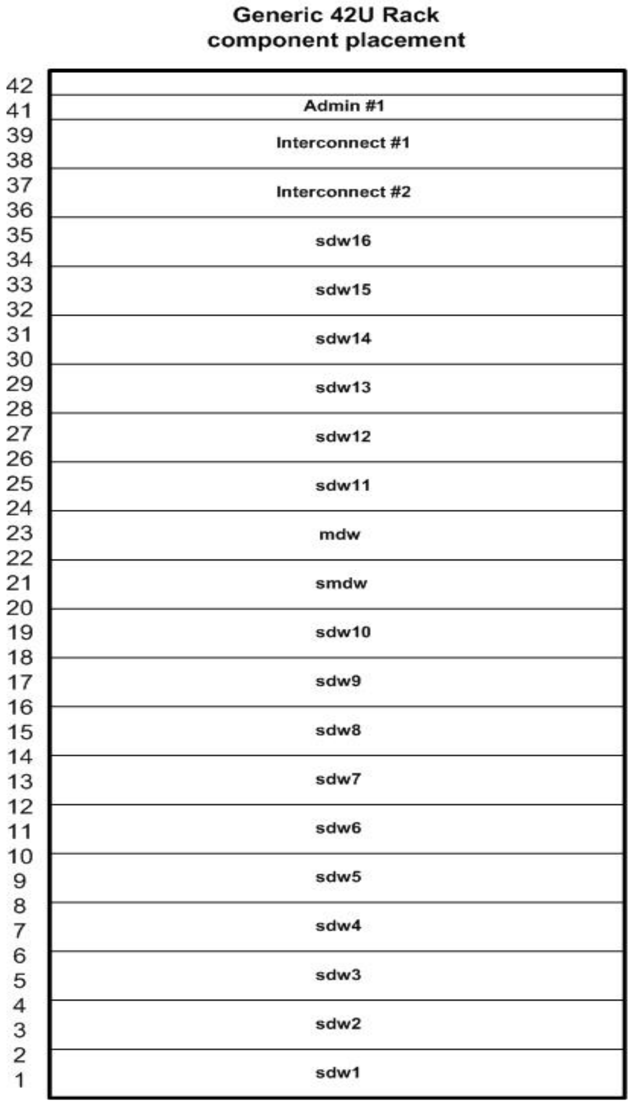

Figure 7. 42U Rack Diagram

Appendix C: Using gpcheckperf to Validate Disk and Network Performance

The following examples illustrate how gpcheckperf is used to validate disk and network performance in a cluster.

Checking Disk Performance — gpcheckperf Output

```
[gpadmin@mdw ~]$ gpcheckperf -f hosts -r d -D -d /data1/primary -d
/data2/primary -S 80G

/usr/local/greenplum-db/./bin/gpcheckperf -f hosts -r d -D -d
/data1/primary -d /data2/primary -S 80G

--------------------

-- DISK WRITE TEST

--------------------

--------------------

-- DISK READ TEST

--------------------

====================

== RESULT

====================

 disk write avg time (sec): 71.33

 disk write tot bytes: 343597383680

 disk write tot bandwidth (MB/s): 4608.23

 disk write min bandwidth (MB/s): 1047.17 [sdw2]

 disk write max bandwidth (MB/s): 1201.70 [sdw1]

 -- per host bandwidth --

  disk write bandwidth (MB/s): 1200.82 [sdw4]

  disk write bandwidth (MB/s): 1201.70 [sdw1]

  disk write bandwidth (MB/s): 1047.17 [sdw2]

  disk write bandwidth (MB/s): 1158.53 [sdw3]


 disk read avg time (sec): 103.17

 disk read tot bytes: 343597383680

 disk read tot bandwidth (MB/s): 5053.03

 disk read min bandwidth (MB/s): 318.88 [sdw2]

 disk read max bandwidth (MB/s): 1611.01 [sdw1]

 -- per host bandwidth --

  disk read bandwidth (MB/s): 1562.76 [sdw4]
  disk read bandwidth (MB/s): 1611.01 [sdw1]
  disk read bandwidth (MB/s): 318.88 [sdw2]
  disk read bandwidth (MB/s): 1560.38 [sdw3]
```

Checking Network Performance — **gpcheckperf** Output

```
[gpadmin@mdw ~]$ gpcheckperf -f network1 -r N -d /tmp
/usr/local/greenplum-db/./bin/gpcheckperf -f network1 -r N -d /tmp
-------------------
-- NETPERF TEST
-------------------
====================
== RESULT
====================
Netperf bisection bandwidth test
sdw1 -> sdw2 = 1074.010000
sdw3 -> sdw4 = 1076.250000
sdw2 -> sdw1 = 1094.880000
sdw4 -> sdw3 = 1104.080000

Summary:
sum = 4349.22 MB/sec
min = 1074.01 MB/sec
max = 1104.08 MB/sec
avg = 1087.31 MB/sec
median = 1094.88 MB/sec
```