Pivotal

PRODUCT DOCUMENTATION

# Pivotal™ Greenplum Database®

Version 4.3

## Clustering Concepts Guide

Rev: A02

# Notice

## Copyright



Revised August 2016 (4.3.9.0)

# Contents

# Chapter 1

# Introduction

The EMC Data Computing Appliance provides a ready-made platform that strives to accommodate the majority of customer workloads. One of Pivotal Greenplum's strongest value propositions is its ability to run on practically any modern-day hardware platform. More and more, Pivotal Engineering is seeing cases where customers elect to build a cluster that satisfies a specific requirement or purpose.

Pivotal Platform Engineering publishes this framework as a resource for assisting customers in this effort.

## Objectives

This guide can be used for:

- A clear understanding of what characterizes a recommended platform for running Pivotal Greenplum Database
- A review of the two most common topologies with supporting recommended architecture diagrams
- Pivotal recommended reference architecture that includes hardware recommendations, configuration, hard disk guidelines, network layout, installation, data loading, and verification
- Extra guidance with real-world Greenplum cluster examples (see *Pivotal Cluster Examples*)

## This document does:

- provide recommendations for building a well-performing Pivotal cluster using the hardware guidelines presented
- provide general concepts without specific tuning suggestions

This document does not:

- promise Pivotal support for the use of third party hardware
- 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 Pivotal Greenplum. A subset of information is included as it pertains to deploying a Pivotal cluster.

## Greenplum Terms to Know

| | |
|---|---|
| **master** | A server that provides entry to the Greenplum Database system, accepts client connections and SQL queries, and distributes work to the segment instances. |
| **segment instances** | Independent PostgreSQL databases that each store a portion of the data and perform the majority of query processing. |
| **segment host** | A server that typically executes multiple Greenplum segment instances. |
| **interconnect** | Networking layer of the Greenplum Database architecture that facilitates inter-process communication between segments. |

## Feedback and Updates

Please send feedback and/or questions regarding content to cluster-build@pivotal.io.

# Chapter 2

# Key Points for Review

## What is Pivotal Engineering Recommended Architecture?

This Pivotal Recommended Architecture comprises generic recommendations for third party hardware for use with Pivotal software products. Pivotal maintains examples of various implementations internally to aid in assisting customers in cluster diagnostics and configuration assistance. Pivotal does not perform hardware replacement, nor is Pivotal a substitute for the OEM vendor support for these configurations.

## Why Install on an OEM Vendor Platform?

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

Some possibilities are:

- Varying workload profiles that may require more memory or higher processor capacity
- Specific functional needs like public/private clouds, increased density, or disaster recovery (DR)
- Support for radically different network topologies
- Deeper, more direct access for hardware and OS management
- Existing relationships with OEM hardware partners

Pivotal Engineering highly recommends following Pivotal architecture guidelines if customers opt out of using the appliance and discussing the implementation with a Pivotal Engineer. Customers achieve much greater reliability when following these recommendations.

# Chapter 3

# Characteristics of a Supported Pivotal Hardware Platform

## Commodity Hardware

Pivotal believes that customers should 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 across many platforms
  - NIC chipsets (like some of the Intel series)
  - RAID controllers (like LSI or StorageWorks)
- Reference motherboards/designs
  - Machines that use reference motherboard implementations are preferred.
  - Although DIMM count is important, if a manufacturer integrates more DIMM slots than the CPU manufacturer specifies, more risk is placed 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 (the cover can be opened without tools, parts are hot-swappable without tools, etc.).
- Labeling – components such as DIMMs are labeled so it's easy to determine which 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
- Redundant network connectivity
- Hot-swappable drives
- Hot-spare drives when immediate replacement of failed hardware is unavailable

# Determining the Best Topology

## 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 Ethernet. This topology limits the cluster size to the number of ports on the selected interconnect switches. 10Gb ports on the servers are bonded into an active/active pair and route directly to a set of switchesconfigured using MLAG (or comparable technology) to provide a redundant high speed network fabric.

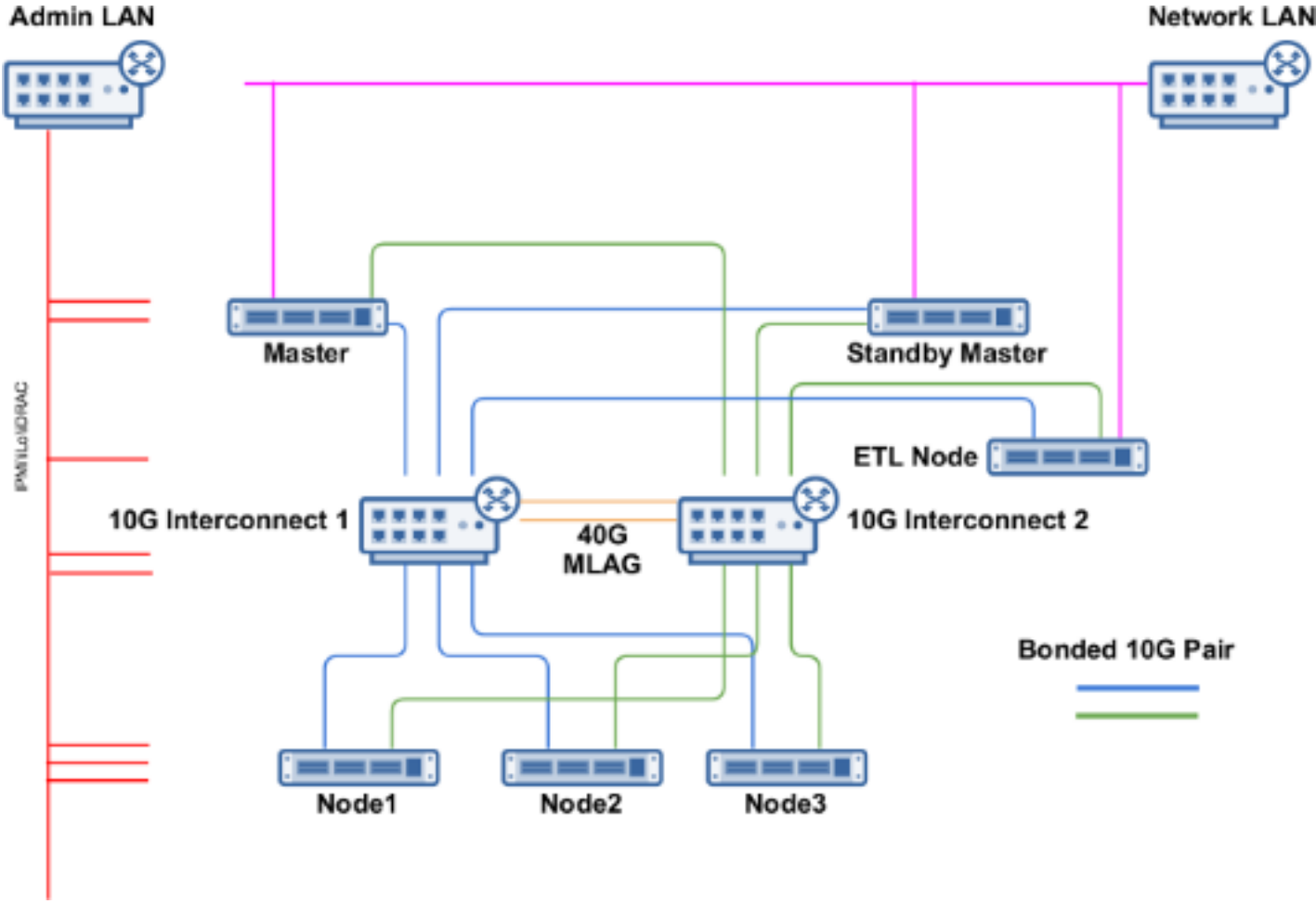



**Figure 1: Recommended Architecture Example 1 (Typical Topology)**

## Scaleable Topology

Scalable networks implement a network core that allows 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.

### How to Determine the Maximum Number of Servers

For example, each rack can hold 16 servers and you determine that the core switches each have 48 ports. Of these ports 4 are used to create the MLAG between the two core switches. Of the remaining 44 ports,

networking from a single set of interconnect switches in a rack uses 4 links per core switch, 2 from each interconnect switch to each of the core switches. The maximum number of servers is determined by the following formula:

```
max-nodes = (nodes-per-rack * ((core-switch-port-count - MLAG port utilization) /
          rack-to-rack-link-port-count))
176 = ( 16 * ( (48 - 4) / 4 ) )
```

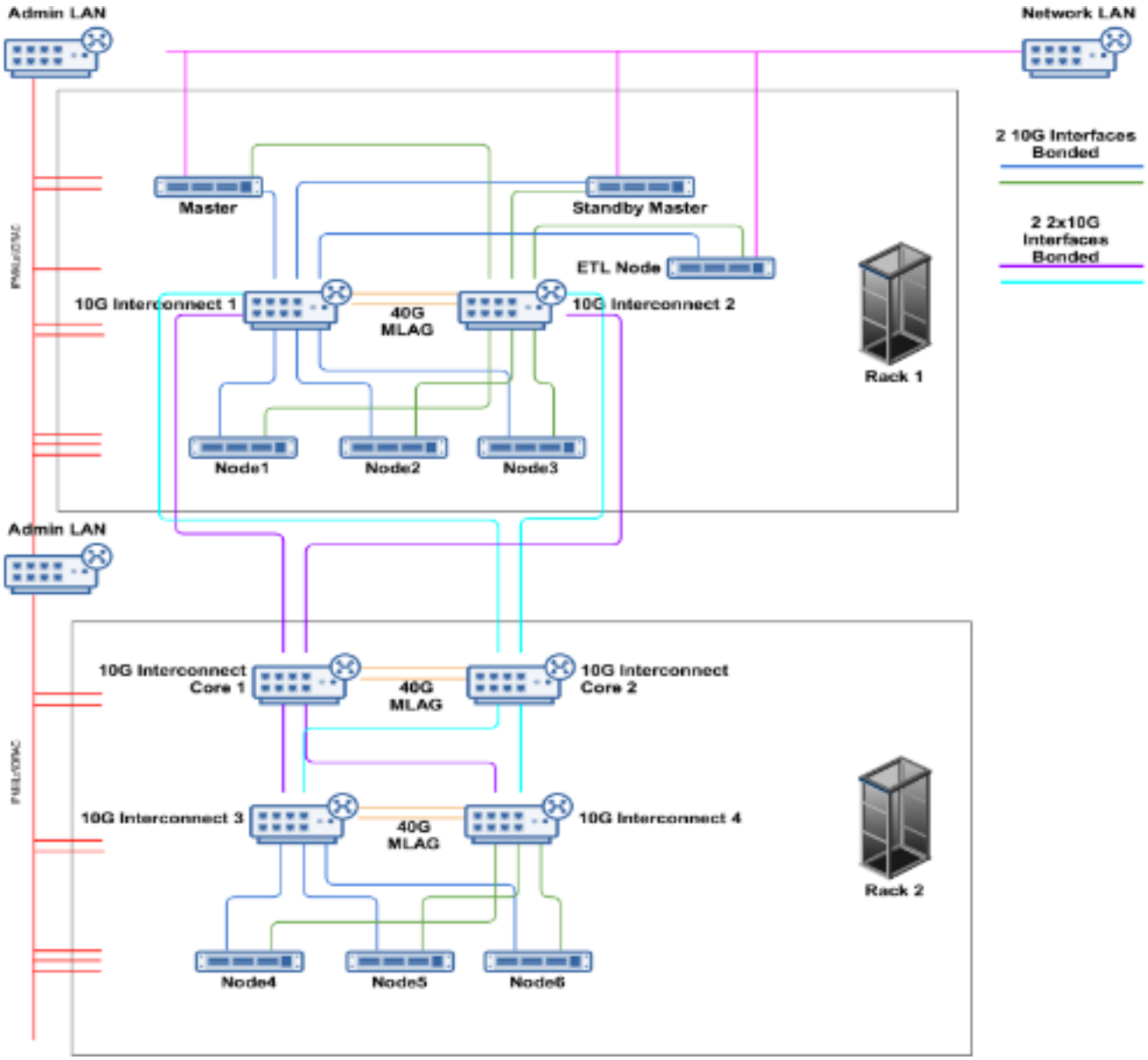



**Figure 2: Recommended Architecture Example 2 (Scalable Topology)**

# Chapter 4

# Pivotal Approved Recommended Architecture

# Minimum Server Guidelines

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

See Appendix C: Using `gpcheckperf` to Validate Disk and Network Performance for example `gpcheckperf` output.

**Table 1: Baseline Numbers for a Pivotal Cluster**

| Server Purpose | Processor | RAM | Storage Bandwidth | Network Bandwith | Typical Case Size |
|---|---|---|---|---|---|
| **Master Nodes (mdw & smdw)**<br>Users and applications connect to masters to submit queries and return results. Typically, monitoring and managing the cluster and the database is performed through the master nodes. | 8+ physical cores at greater than 2GHz clock speed | > 256GB | > 600MB/s Read<br>> 500MB/s Write | 2x10Gb NICs<br>Multiple NICs | 1U |
| **Segment Nodes (sdw)**<br>Segment nodes store data and execute queries. They are generally not public facing.<br>Multiple segment instances run on one segment node. | 8+ physical cores at greater than 2GHz clock speed | > 256GB | > 2000MB/s Read<br>> 2000MB/s Write | 2x10Gb NICs<br>Multiple NICs | 2U |
| **ETL/Backup Nodes (etl)**<br>Generally identical to segment nodes. These are used as staging areas for loading data or as destinations for backup data. | 8+ physical cores at greater than 2GHz clock speed | > 64GB or more | > 2000 MB/s Read<br>> 2000 MB/s Write | 2x10Gb NICs<br>Multiple NICs | 2U |

# Network Guidelines

**Table 2: Administration and Interconnect Switches**

| Switch Purpose | Port Count | Port Type | Description |
|---|---|---|---|
| **Administration Network**<br>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 | 1Gb | A layer-2/layer-3 managed switch per rack with no specific bandwidth or blocking requirements. |
| **Interconnect Network** | 48 | 10GB | Two layer-2/layer-3 managed switches per rack. All ports must have full bandwidth, be able to operate at line rate, and be non-blocking. |

**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 Pivotal Greenplum rack is 4 x 208/220 V, 30 amp, single phase circuits in the US. Internationally, 4 x 230 V, 32 amp, single phase circuits are generally used. This affords a power budget of ~9600 VA of fully redundant power.<br>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. |

# Node Guidelines

## OS Levels

At a minimum the following operating systems (OS) are supported:

- Red Hat/CentOS Linux 5*
- Red Hat/CentOS Linux 6
- Red Hat/CentOS Linux 7**
- SUSE Enterprise Linux 10.2 or 10.3
- SUSE Enterprise Linux 11

*RHEL/CentOS 5 will be unsupported in the next major release

**support for RHEL/CentOS 7 is near completion, pending kernel bug fixes

For the latest information on supported OS versions, refer to the *Greenplum Database Installation Guide*.

## Setting OS Parameters for Greenplum Database

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

## Greenplum Database Server Guidelines

Greenplum Database integrates three kinds of servers: master servers, segment hosts, and ETL servers. Greenplum Database servers must meet the following criteria.

## Master Servers

- 1U or 2U server. With less of a need for drives, rack space can be saved by going with a 1U form factor. However, a 2U form factor consistent with segment hosts may increase supportability.
- Same processors, RAM, RAID card, and interconnect NICs as the segment hosts.
- Six to ten disks (eight is most common) organized into a single RAID5 group with one hot spare configured.
- SAS 15k or SSD disks are preferred with 10k disks a close second.
- SATA drives are acceptable in solutions oriented towards archival space over query performance.
- All disks must be the same size and type.
- 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 have sufficient additional network interfaces to connect to the customer network directly in the manner desired by the customer.

## Segment Hosts

- Typically a 2U server.
- The fastest available processors.
- 256 GB RAM or more.
- One or two RAID cards with maximum cache and cache protection (flash or capacitors preferred over battery). RAID cards should be able to support full read/write capacity of the drives.
- 2 x 10 Gb NICs.

- 12 to 24 disks organized into two or four RAID5 groups. Hot spares should be configured, unless there are disks on hand for quick replacement.
- 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.
- A minimum read rate in `gpcheckperf` of 300 MB/s per segment or higher. (2000 MB/s per server is typical.)
- A minimum write rate in `gpcheckperf` of 300 MB/s or higher (2000 MB/s per server is typical.)

## Additional Tips for Segment Host Configuration

The number of segment instances that are run per segment host is configurable, and each segment instance is itself a database running on the server. A baseline recommendation on current hardware, such as the hardware described in Appendix A, is **8 primary segment instances per physical server**.

A set of memory parameters will be determined when installing the database software that depend upon the amount of RAM selected for each segment instance. While these are not platform parameters, it is the platform that determines how much memory is available and how the memory parameters should be set in the software. Refer to the online calculator (*http://greenplum.org/calc/*) to determine these settings.

Refer to Appendix D for further reading on segment instance configuration.

## 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 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 typically less of a concern than storage for these systems.
- Should be capable of read rates in `gpcheckperf` of 100 MB/s or higher. (The faster the ETL servers scan, the faster query 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.

# Hard Disk Configuration Guidelines

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

- four RAID5 groups of six disks each (used as four file systems), or
- combined into one or two file systems 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 file systems is covered in the relevant operating system's installation 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 offer a compromise. 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.

LUNs for the system should be partitioned and mounted as /data1 for the first LUN and additional LUNs should follow the same naming convention while incrementally increasing the number (/data1, /data2, /data3 … /dataN). All file systems should be formatted as xfs and follow the recommendations set forth in the *Pivotal Greenplum Database Installation Guide*.

## Master Server

Master servers (primary and secondary) have eight, hot-swappable disks. Configure all eight disks into a single, RAID5 stripe set. Each of the virtual disks that are carved from this disk group should have 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 Servers

Segment servers 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 have 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.

## SSD Storage

Flash storage has been gaining in popularity. Pivotal has not had the opportunity to do enough testing with SSD drives to make a recommendation. It is important when considering SSD drives to validate the sustained sequential read and write rates for the drive. Many drives have impressive burst rates, but are unable to sustain those rates for long periods of time. Additionally, the choice of RAID card needs to be evaluated to ensure it can handle the bandwidth of the SSD drives.

## SAN/JBOD Storage

In some configurations it may be a requirement to use 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 understand that, based on testing by Pivotal Platform and Systems Engineering, SAN and JBOD storage will not perform as well as local, internal server storage.

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

- Know the database size and the estimated growth over time
- Know the customer’s 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 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

All the systems in the Greenplum cluster need to be tied together in some form of dedicated, high-speed data interconnect. This network is used for loading data and for passing data between systems during query processing. It should be as high-speed and low-latency as possible, and it should not be used for any other purpose (i.e., it should not be part of the general LAN).

A rule of thumb for network utilization in a Greenplum cluster 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 2000MB/s read bandwidth (as measured by `gpcheckperf`) might be expected to transmit 400MB/s. Greenplum also compresses some data on disk but uncompresses it before transmitting to other systems in the cluster, so a 2000 MB/s read rate with a 4x compression ratio results in an 8000 MB/s effective read rate. Twenty percent of 8000 MB/s is 1600 MB/s which is more than a single gigabit interface's bandwidth.

To accommodate this traffic, 10 Gb networking is recommended for the interconnect. Current best practice suggests two 10 Gb interfaces for the cluster interconnect. This ensures that there is bandwidth to grow into, and reduces cabling in the racks.

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 network. Intel, QLogic, or Emulex network interfaces tend to work best. Layer 3 capability is recommended since it integrates many features that are useful in a Greenplum Database environment.

> **Note:**
>
> The vendor hardware referenced above is strictly mentioned as an example. Pivotal Platform and Systems Engineering does not specify which products to use in the network.

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

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. It 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 kinds 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.

## 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 tends 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 tends 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 to submit queries. Typically, this is done with 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 rack 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.

# IP Addressing Guidelines

## IP Addressing Scheme for the Admin Network

An admin network should be created so that system maintenance and access work can be done on a network that is not the same as cluster traffic between the nodes.

> **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.

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 |
|---|---|---|
| Secondary Master Node | CIMC | 172.254.1.252/16 |
| | Eth0 | 172.254.1.250/16 |
| Secondary Master Node | CIMC | 172.254.1.253/16 |
| | Eth0 | 172.254.1.251/16 |
| Non-master Segment Nodes in rack 1 (master rack) | CIMC | 172.254.1.101/16 through 172.254.1.116/16 |
| | Eth0 | 172.254.1.1/16 through 172.254.1.16/16 |
| Non-master Segment Nodes in rack 2 | CIMC | 172.254.2.101/16 through 172.254.2.116/16 |
| | Eth0 | 172.254.2.1/16 through 172.254.2.16/16 |
| Non-master Segment Nodes in rack # | CIMC | 172.254.#.101/16 through 172.254.#.116/16 |
| | Eth0 | 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

The following table 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

The Interconnect is where data is routed at high speed between the nodes.

**Table 6: Interconnect IP Addressing for 10 Gb NICS**

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

The following table 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.

# 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 the 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.

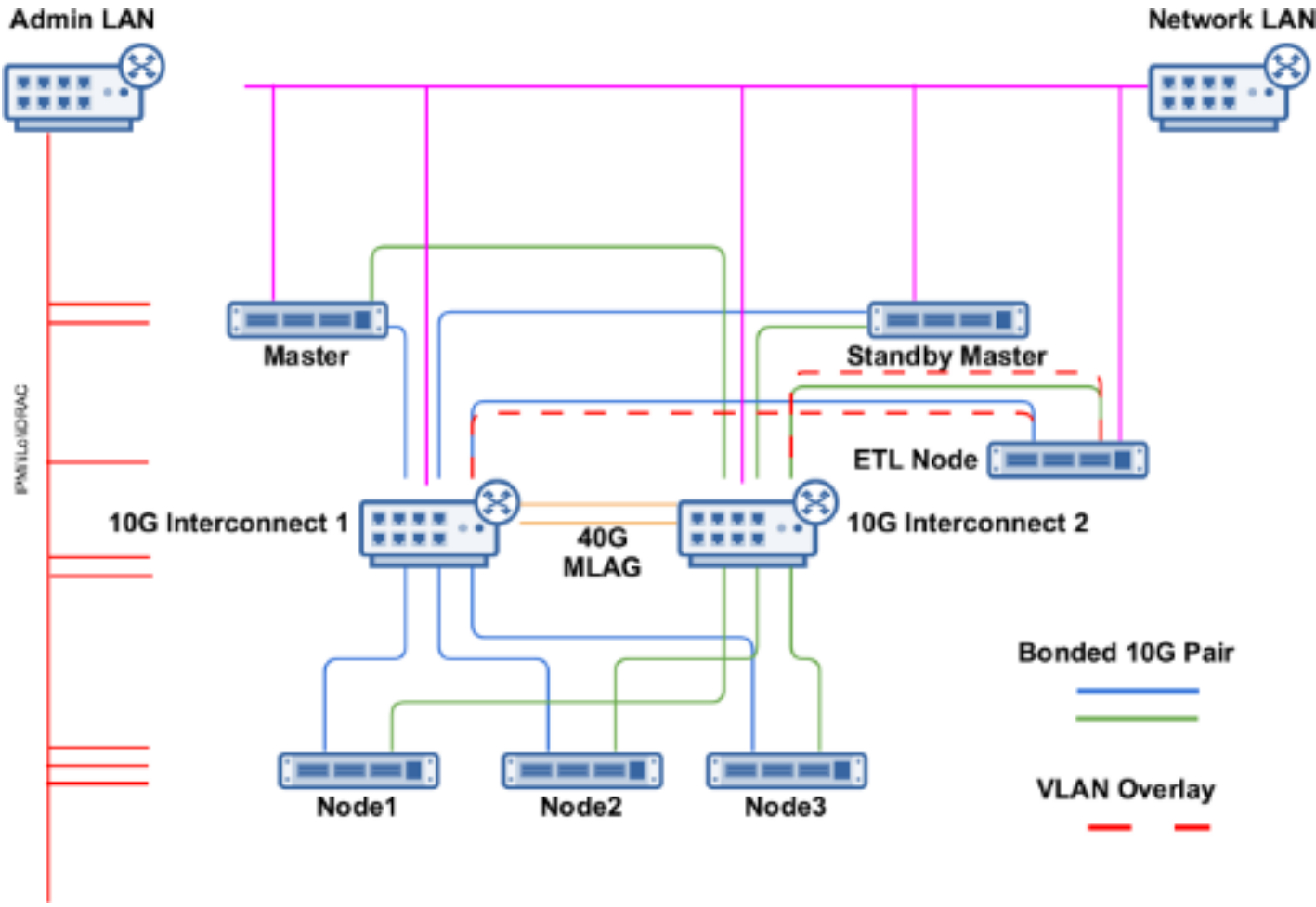



**Figure 3: Basic VLAN Overlay Example**

This figure shows a cluster with 3 segment hosts, master, standby master and ETL host. In this case, only the ETL host is part of the overlay. It is not a requirement to have ETL node use the overlay, though this is common in many configurations to allow data to be staged within a cluster. Any of the servers in this rack or any rack of any other configuration may participate in the overlay if desired. The type of configuration will depend upon security requirements and if functions within the cluster need to reach any outside data sources.

## *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 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.

## 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).

## 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.

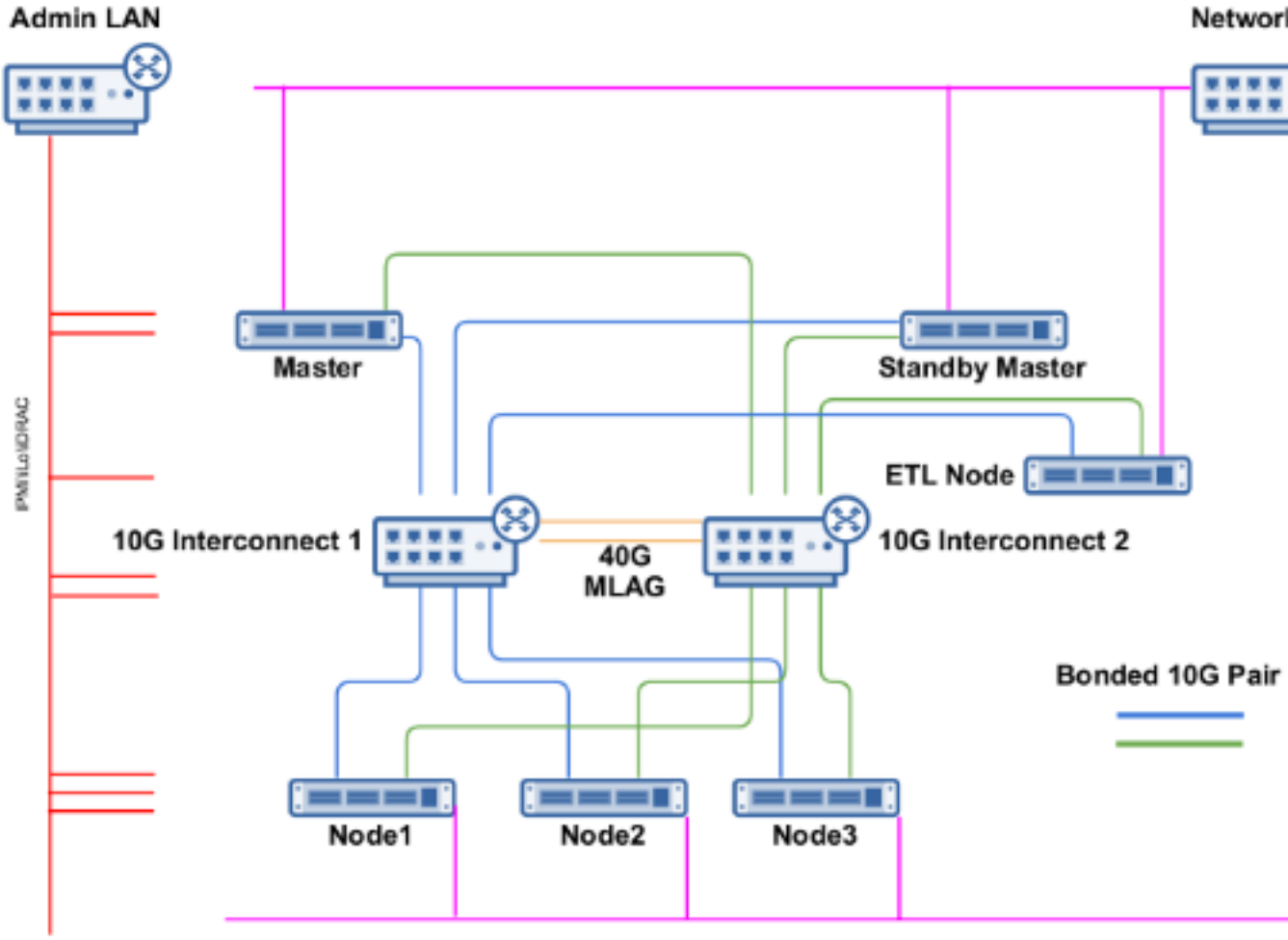



**Figure 4: Data Loading—Direct Connect to Customer Network**

### 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 Greenplum 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

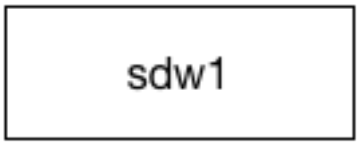
### Racks

Each rack in a Recommended 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:


### Servers

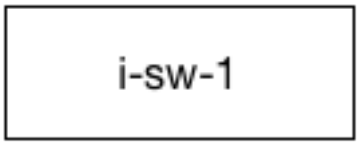
Each server is labeled on both the front and back of the server. The label should be the hostname of the server.

In other words, if a segment node is known as sdw15, the label on that server would be 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.




## *Certification Guidelines*

### Network Performance Test gpcheckperf

Verifies the line rate on both 10 Gb NICs.

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 *Using gpcheckperf to Validate Disk and Network Performance* for more information on the `gpcheckperf` command.

# 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

The following table lists good choices for cluster hardware based on Intel Sandy Bridge processor-based servers and Cisco switches.

**Table 8: Hardware Components**

| Cluster Element | Description |
|---|---|
| Master Node<br><br>Two of these nodes per cluster | 1U server (similar to the Dell R630):<br>• 2 x E5-2680v3 processors (2.5GHz, 12 cores, 120W)<br>• 256 GB RAM (8 x 16 GB)<br>• 1 x RAID card w/ 1 GB protected cache<br>• 8 x SAS, 10 k, 6 G disks (typically 8x600 GB, 2.5”)<br>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 file system for data<br>• 2 x 10 Gb Intel, QLogic, or Emulex based NICs<br>• Lights out management (IPMI-based BMC)<br>• 2 x 650W or higher, high-efficiency power supplies |
| Segment Node & ETL Node<br><br>Up to 16 per rack. No maximum total count | 2U server (similar to the Dell R730xd):<br>• 2 x E5-2680v3 processors (2.5GHz, 12 cores, 120W)<br>• 256 GB RAM (8 x 16 GB)<br>• 1 x RAID card w/ 1 GB protected cache<br>• 12 to 24 x SAS, 10k, 6G disks (typically 12x600 GB, 3.5” or 24x1.8TB, 2.5”)<br>Organized into two to four RAID5 groups. Used either as two to four data file systems (with logical devices skimmed off for boot, root, swap, etc.) or as one large device bound with Logical Volume Manager.<br>• 2 x 10 Gb Intel, QLogic, or Emulex based NICs<br>• Lights out management (IPMI-based BMC)<br>• 2 x 650W or higher high-efficiency power supplies |
| Admin Switch | Cisco Catalyst 2960 Series<br><br>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 is appropriate for this role. There will be at least one per rack. |

| Cluster Element | Description |
|---|---|
| Interconnect | Arista 7050-52<br>The Arista switch line allows for multi-switch link aggregation groups (called MLAG), easy expansion, and a reliable body of hardware and operating system. |

# Appendix B

## Example Rack Layout

The following figure is an example rack layout with proper switch and server placement.

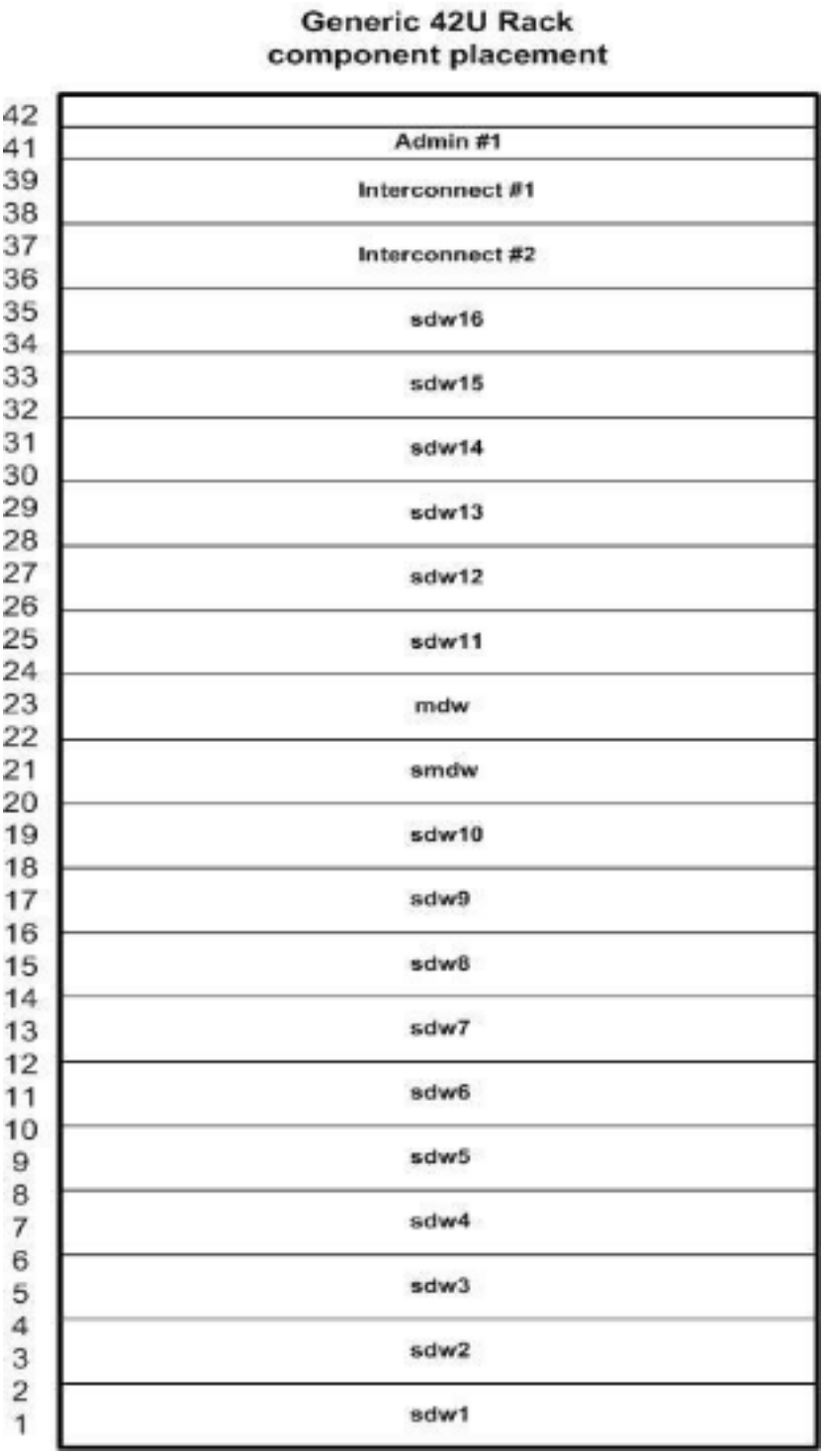



**Figure 5: 42U Rack Diagram**

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

```
-- per host bandwidth --
```

## 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
```

# Appendix D

# Pivotal Greenplum Segment Instances per Server

## Understanding Greenplum Segments

Greenplum segment instances are essentially individual databases. In a Greenplum cluster there will be a Greenplum master server which dispatches work to be done to multiple segment instances. Each of these instances will reside on segment hosts. Data for a table is distributed across all of the segment instances and when a query is executed that requests data it is dispatched to all of them to execute in parallel. Those instances that actively process the query are referred to as the primary instances. A Greenplum cluster in addition will be running mirror instances, one paired to each primary. The mirrors do not participate in answering queries; they are just performing data replication, so that if a primary should fail its mirror can take over processing in its place.

When planning a cluster, it is important to understand that all of these instances are going to accept a query in parallel and act upon it. Therefore there must be enough resources on a server to facilitate all of these processes running and communicating with each other at once.

## Segments Resources Rule of Thumb

A general rule of thumb is that for every segment instance (primary or mirror) you will want to provide at least:

1 core
200 MB/s IO read
200 MB/s IO write
8 GB RAM
1GB network throughput

A segment host with 8 primary and 8 mirror instances would have:

16 cores
3200 MB/s IO read
3200 MB/s IO write
128 GB RAM
20GB network throughput

These numbers have proven to provide a reliable platform for a variety of use cases and give a good baseline for the number of instances to run on a single server. Pivotal recommends a maximum of 8 primary and 8 mirror instances on a server even if the resources provided are sufficient for more.

Pivotal has found that allocating a ratio of 1 to 2 physical CPUs per primary segment works well for most use cases; it is not recommend to drop below 1 CPU per primary segment. Ideal architectures will additionally align NUMA architecture with the number of segments.

## Reasons to reduce the number of segment instances per server

- A database schema that uses partitioned columnar tables has the potential to generate a large number of files. For example, a table that is partitioned daily for a year will have over 300 files, one for each day. If that table additionally has columnar orientation with 300 columns it will have well over 90,000 files representing the data in that table on one segment instance. A server that is running 8 primary instances with this table would have to open 720,000 files if a full table scan query were issued to that table. Systems that make use

of partition columnar tables may benefit from a lesser number of segment instances per server if data is being used in a way that requires many open files.

- Systems that span large numbers of nodes create more work for the master to plan queries and do coordination of all of the segments. In systems spanning two or more racks consider reducing the number of segment instances per server.
- When queries require large amounts of memory reducing the number of segments per server increases the amount of memory available to any one segment.
- If the amount of concurrent query processing causes resources to run low on the system, reducing the amount of parallelism on the platform itself will allow for more parallelism in query execution.

## Reasons to increase the number of segment instances per server

- In low concurrency systems increasing the segment instance count will allow each query to utilize more resources in parallel if system utilization is low.
- Systems with large amounts of free RAM that can be used by the OS for file buffers may benefit from increasing the number of segment instances per server.

# Appendix E

# Pivotal Greenplum on Virtualized Systems

## General understanding of Pivotal Greenplum and virtualization

Greenplum Database is a parallel processing software. This means that the Pivotal Greenplum software often does the same process at the same time across a cluster of nodes. Virtualization is frequently used to centralize systems so that they will be able to share resources, taking advantage of the fact that software often utilizes resources sporadically, allowing those resources to be over-subscribed. Greenplum Database will not function well in an oversubscribed environment because all segments become active at once during query processing. In that type of environment, the system is prone to bottlenecks and unpredictable behavior that could result from being unable to access resources the system believes it has been allocated.

With this in mind, as long as the system meets the requirements set forth in the installation guide, Greenplum is supported on virtual infrastructure.

## Choosing the number of segment instances to run per VM

The recommended hardware specifications are quite large and may be hard to achieve in a virtual environment. In these cases each VM should have no more than 1 primary and 1 mirror segment for every 2 CPUs, 32 GB of RAM, and 300MB/s of sequential read bandwidth and write bandwidth. Thus a VM with 4 CPU, 64GB RAM, and 1GB/s sequential read and write would be able to host 2 primary segment instances and 2 mirror segment instances.

While it is possible to create segment host VMs that only host a single primary segment instance, it is preferred to have at least two or more primary segment instances per VM. Certain queries that perform tasks such as looking for uniqueness can cause some segment instances to perform more work, and require more resources, than other instances. Grouping multiple segment instances together on one server can mitigate some of these increased resource needs by allowing a segment instance to utilize the resources allocated to the other segment instances.

## VM Environment Settings

VMs hosting Greenplum Database should not have any auto-migration features turned on. The segment instances are expecting to run in parallel and if one of them is paused to coalesce memory or state for migration the system can see it as a failure or outage. It would be better to take the system down, remove it from the active cluster and then introduce it back into cluster once it has been moved.

Special care should be given to understand the topology of primary and mirror segment instances. No set of VMs that contain a primary and its mirror should run on the same host system. If a host containing both the primary and mirror for a segment fails, the Greenplum cluster will be offline until at least one of them is restored to complete the database content.

# Appendix
# F

# Additional Helpful Tools

## Yum Repository

Configuring a YUM repository on the master servers can make management of the software across the cluster more efficient, particularly in cases where the segment nodes do not have external internet access. More than one repository can make management easier, for example one repository for OS files and another for all other packages. Configure the repositories on both the master and standby master servers.

## Kickstart Images

Kickstart images for the master servers and segment hosts can speed up implementation of new servers and recovery of failed nodes. In most cases where there is a node failure but the disks are good, reimaging is not necessary because the disks in the failed server can be transferred to the new replacement node.