How to Deploy an IBM OpenPOWER Cluster for Hortonworks Data Platform

Version 1.0

Introduction

This document and provides a set of instructions, guidelines, and links to automation tools for deploying (provisioning) an IBM® Power-based cluster that is designed to host an installation of Hortonworks Data Platform (HDP). A complete solution consists of the cluster infrastructure and an installation of HDP software on that cluster infrastructure. The scope of this document covers the deployment of the cluster infrastructure – hardware (server, storage, network), systems software (operating systems), and management software. Installation of the HDP software is not covered by this document. It is recommended that a Hortonworks consultant be retained to accomplish the HDP software installation.

A variety of Power-based configurations are suitable for the cluster infrastructure for an HDP solution. The reference design section of the Hortonworks Data Platform on IBM Power – Reference Architecture and Design – Version 1.0 provides additional details and information regarding HDP on Power configurations generally. This deployment guide selects a variant of the "Minimum Production Configuration" as the example configuration to deploy.

This document is part of the "Hortonworks Data Platform on IBM Power Solution Deployment Kit" (hereafter referenced as the "HDP on Power Solution Deployment Kit"), and this document contains references and links to other materials in the deployment kit.

High-level deployment steps

The following steps accomplish the deployment of this solution.

1	Acquire the hardware
2	Rack and cable the hardware
3	Choose the basic configuration parameters
4	Prepare the switches
5	Acquire the system software
6	Prepare the deployer node
7	Configure the cluster using the Cluster Genesis tool
8	Complete the post-Cluster Genesis configuration
9	Install Hortonworks Data Platform (HDP)

Step 1: Acquire the hardware

Refer to the Design Proposal for an overview of the configuration and the Bill of Materials for a list of the parts required for the configuration.

For assistance with ordering and purchasing, access the following link to contact an IBM representative.

https://www-01.ibm.com/marketing/iwm/dre/signup?source=MAIL-power&disableCookie=Yes

Step 2: Rack and cable the hardware

Step 2a: Rack the servers and switches

Physically install the servers and switches in the rack. See Figure 1 on page 3 for recommended rack locations for the servers and switches in this configuration.

Details to accomplish this step for the servers may be found in the IBM Knowledge Center in the following sections:

- "8001-12C (IBM Power System S821LC) > Installing and configuring the system > Installing the IBM® Power® System S821LC (8001-12C) system"
- "8001-22C (IBM Power System S822LC for Big Data) > Installing and configuring the system > Installing the system"

Details to accomplish this step for the switches may be found at the following links:

- http://systemx.lenovofiles.com/help/topic/com.lenovo.rackswitch.g8052.doc/G8052_IG.pdf
- http://www.mellanox.com/related-docs/user_manuals/1U_HW_UM_SX1710_SX1410.pdf

Single Rack
Minimum Production Config

HDP
1smn+ 3mn + 1en + 8wn

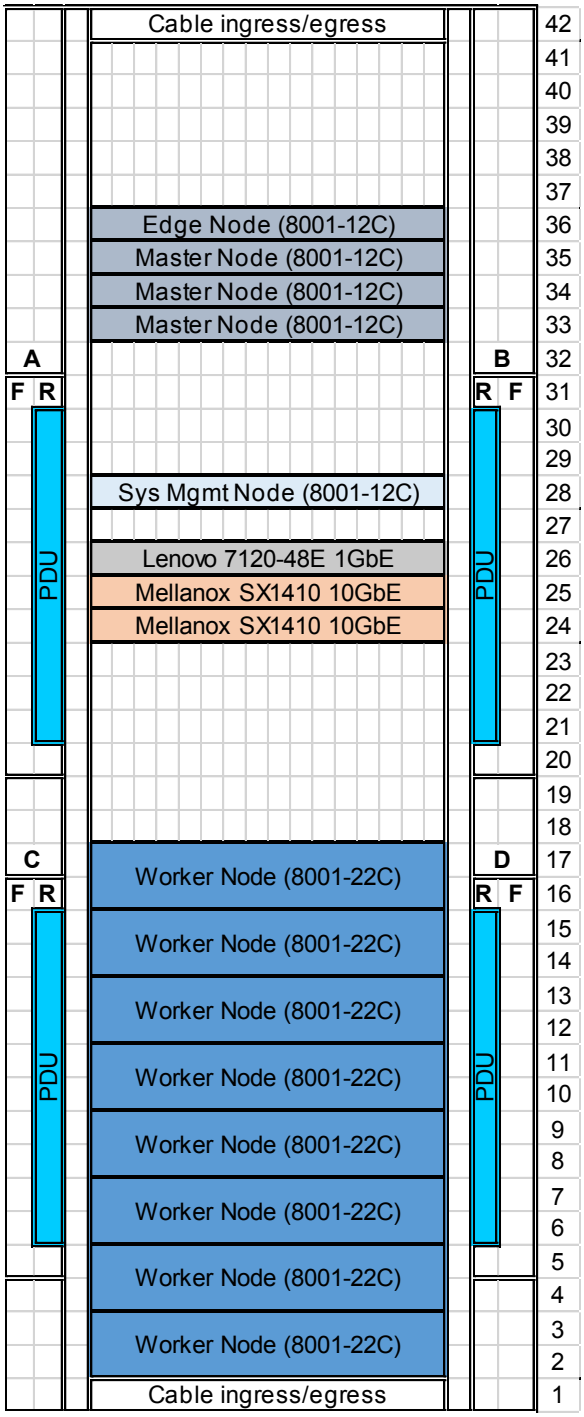

Figure 1. Recommended Rack Locations

Step 2b: Install electrical power cables

Install the appropriate electrical power cables between the servers and switches and the PDUs *within the rack*. The links in the previous step include some guidance regarding the installation of these power cables.

Install the appropriate electrical power cables from the PDUs to the external electrical power source. **NOTE: Installation of these external electrical power cables and applying electrical power to the rack is beyond the scope of this document, and clients should retain a qualified electrician to accomplish this step.**

Step 2c: Install network cables

Install the network cables for the solution. Refer to Figure 2 on page 5, Figure 3 on page 6, Figure 4 on page 7, and Figure 5 on page 8. The servers and switches for this solution are depicted with the relevant network ports on each identified and labeled. For the connections between the servers and switches, each label is listed twice in the figures, indicating two ports which are to be connected with a network cable. Install a cable for each connection indicated.

> Example: The first 10Gb port on Worker Node 1 is on port 1 of the the adapter installed in PCI slot 5. This port is labeled “wn-1 10Gb A” in Figure 2. Data Switch A has the matching label on port 18 indicated in Figure 5. Connect these two ports with a 3m 10Gb SFP+ copper passive cable (included in the BOM, feature code EKC1).

The network cable connections between the switches (management interface connections and IPLs) are depicted as cable connections between the relevant ports. Install a cable for each such cable connection indicated.

Cabling of the uplinks from the Management Switch is covered later in “Step 4: Prepare the switches”.

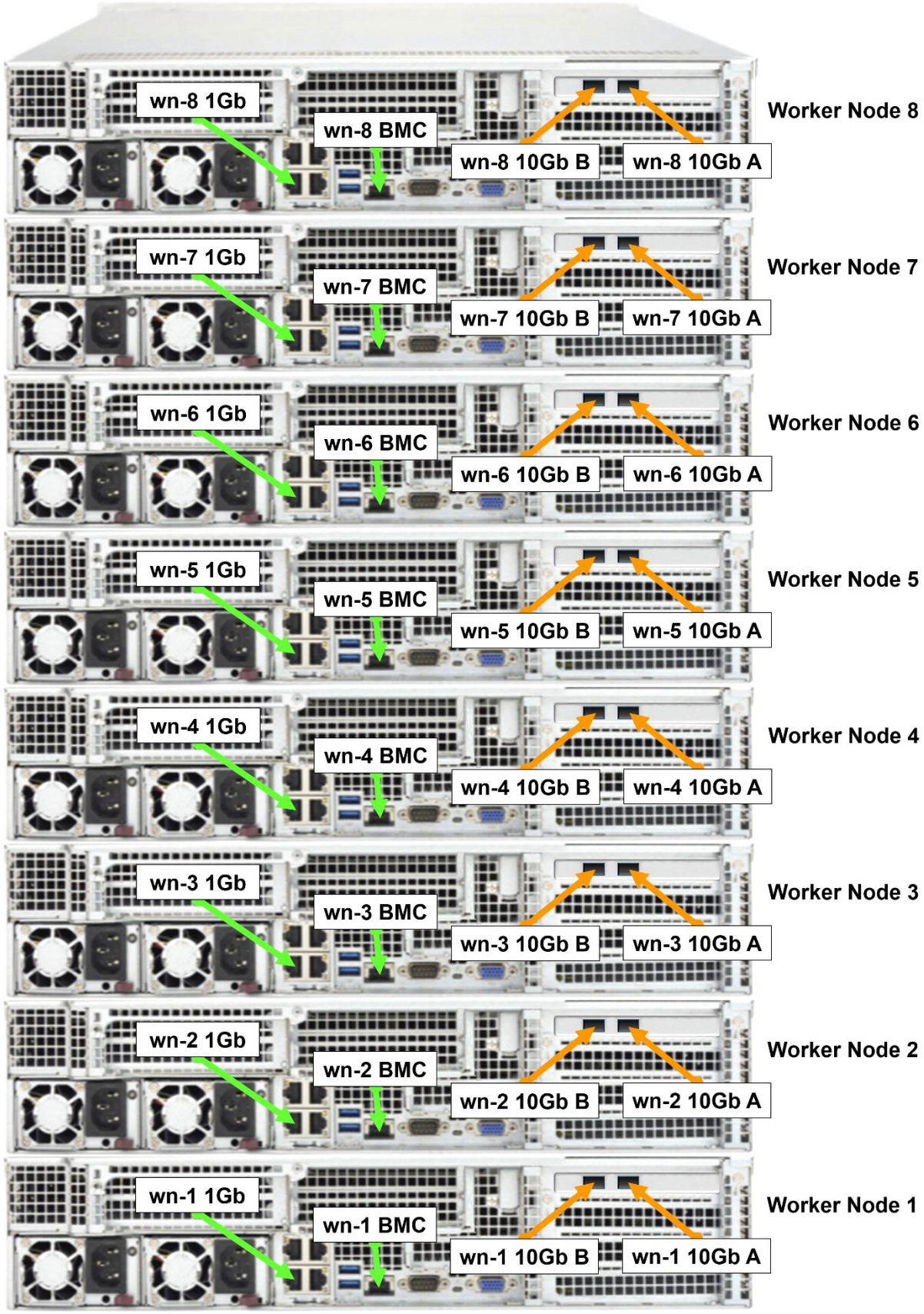

Figure 2. Worker Nodes – Network ports

Figure 3. Master Nodes, Edge Node, System Management Node - Network ports

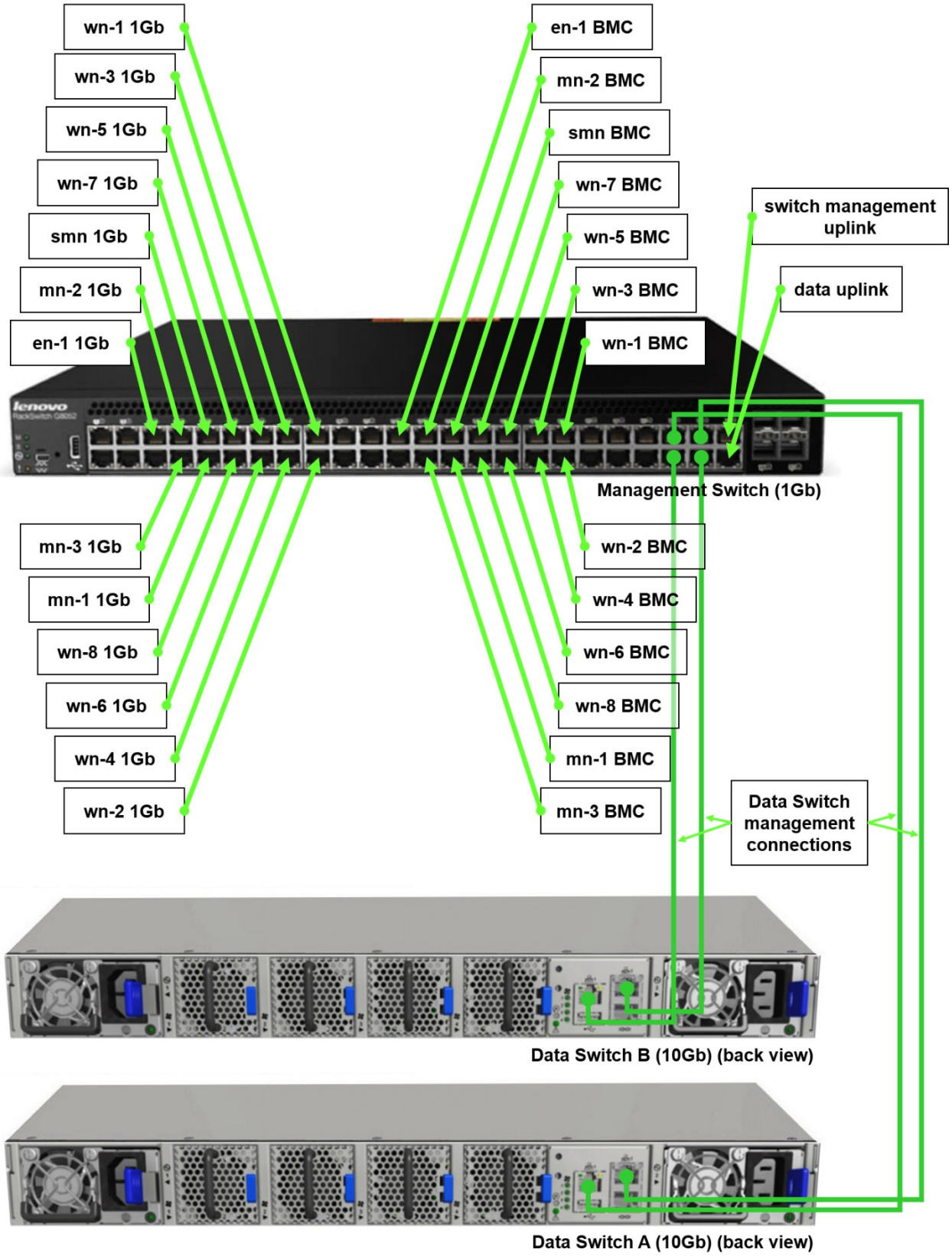

Figure 4. Management Switch - Network ports

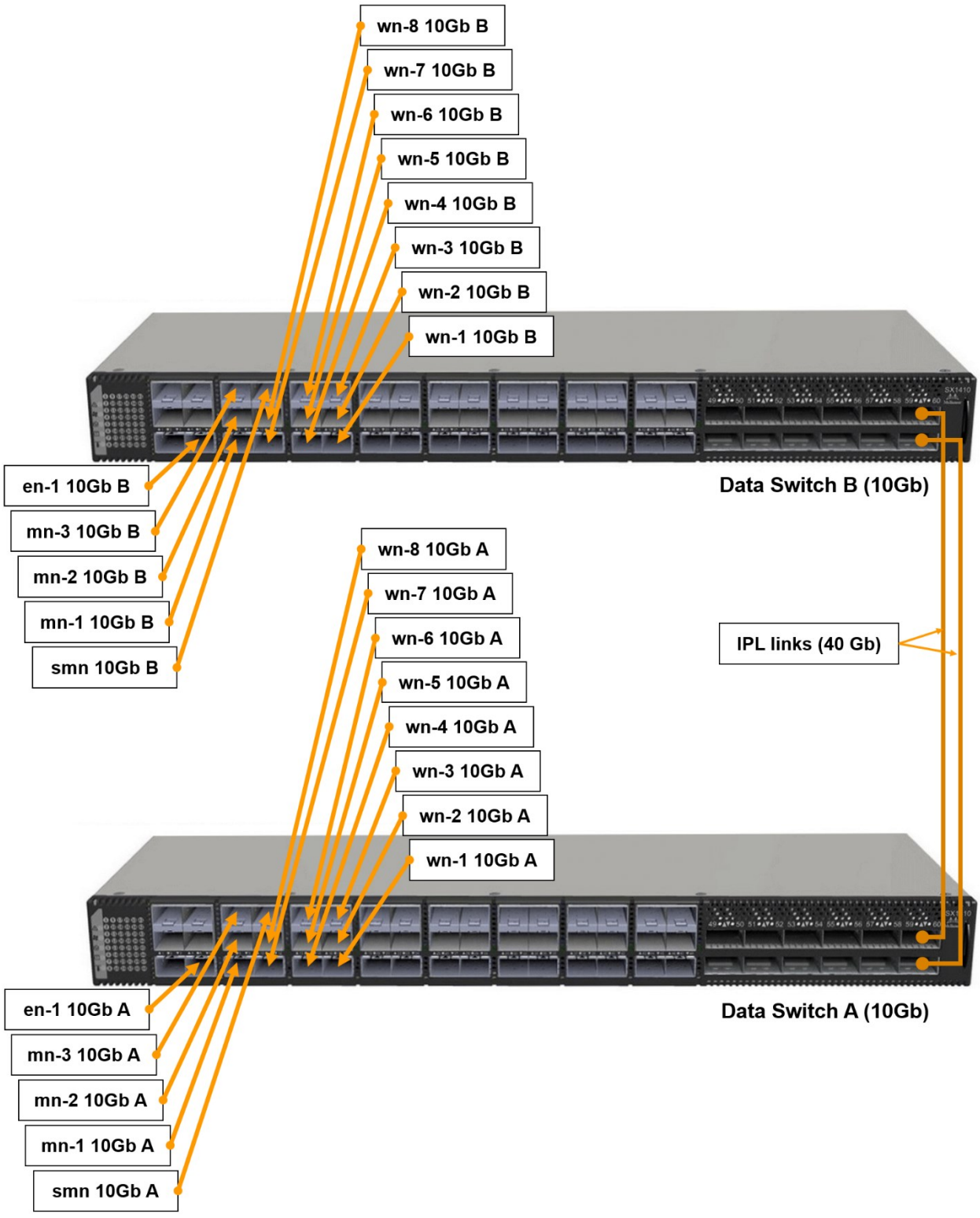

Figure 5. Data Switches - Network ports

Step 3: Choose the basic configuration parameters

Collect or choose the parameters listed in Table 1 on page 10. In a later step, most of these parameters are entered into the configuration file, which is then used during the automated configuration and deployment processes; Some of these parameters are used for manual configuration.

Parameter	Description	Example
Domain name	The domain name for cluster nodes	ibm.com
Upstream DNS servers	Upstream DNS servers	4.4.4.4, 8.8.8.8
Deployment node hostname	The hostname of the deployer node	smn
Deployer node initial IP address	The initial IP address of the deployer node on the Management Network to be used by automation processes; Labeled *cidr-mgmt-switch-external-dev* in the config.yml; cidr format	10.19.60.25/24
Management switch initial IP address	The initial management IP address of the management switch on the Management Network to be used by automation processes; Labeled *ipaddr-mgmt-switch-external* in the config.yml	10.19.60.11
Internal Management Network subnet	The subnet definition of the Internal Management Network; Labeled *ipaddr-mgmt-network* in the config.yml	172.19.190.0/24
Internal Management Network VLAN ID	The VLAN ID for the Internal Management Network; Labeled *vlan-mgmt-network* in the config.yml	190
Provisioning/Service Network subnet	The subnet definition of the Provisioning and BMC Network; Cluster Genesis chooses and assigns addresses from this subnet to nodes which it provisions; Labeled *ipaddr-mgmt-client-network* in the config.yml	172.19.188.0/24
Provisioning/Service Network VLAN ID	The VLAN ID for the Provisioning and BMC Network; Labeled *vlan-mgmt-client-network* in the config.yml	188
Management switch IP address	The IP address to be applied for management access to the management switch; Must be on the Internal Management Network subnet; Labeled *ipaddr-mgmt-switch* in the config.yml file	172.19.190.11
Data switch IP addresses	The IP addresses to be applied to the management interfaces of the data switches; Must be on the Internal Management Network subnet; Labeled *ipaddr-data-switch* in the config.yml file in example below.	172.19.190.13 172.19.190.14

MLAG VIP address	The MLAG virtual IP for the data switches; Must be on the Internal Management Network subnet; Labelled *ipaddr-mlag-vip* in the config.yml file	172.19.190.214
Data Network subnet	The subnet definition for the Data Network; Labeled *networks/data/addr* in the config.yml file	172.19.66.0/24
Data Network broadcast IP address	The broadcast IP for the Data Network subnet; Labeled *networks/data/broadcast* in the config.yml file	172.19.66.255
Data Network static IP range	The range of IP addresses, within the Data Network subnet, from which automation processes will choose IP addresses to configure on provisioned nodes; Labeled *networks/data/available-ips* in the config.yml file	172.19.66.220 172.19.66.239
Management switch login	User ID and password to be used by automation processes to access the management switch (1Gb)	admin/admin
Data switch login	User ID and password to be used by automation processes to access the data switches (10Gb)	admin/admin
Server IPMI login	User ID and password to be used by automation processes to access the server BMCs	ADMIN/ADMIN
Node OS login	User ID and password to be configured in the OS on provisioned nodes	user/passw0rd
Node hostname stem	The hostname stem for hostnames to be configured on provisioned nodes; Each node type (master, edge, worker) must have a unique stem; Automation processes append “-1”, “-2”, “-3”, etc. to the hostname stem to form a full, unique hostname	mn, en, wn

Table 1. Basic configuration parameters

Step 4: Prepare the switches

Prepare the network switches as follows to facilitate the deployment:

1. Reset all switches to factory default configurations
2. Configure IP addresses on each switch to allow admin access over the Management Network. (Refer to the basic configuration parameters.)
   a. For the 1GbE switch, this is a general IP address within VLAN 1.
   b. For the 10GbE switches (Mellanox SX1410), create these IP addresses on the MGMT1 ports. Cluster Genesis automation configures and uses the MGMT0 ports.
3. Uplink the 1GbE switch to the existing network environment
   a. Configure the uplink ports on the 1GbE switch
   b. Cable the uplinks to the existing network environment
4. On the 10Gb switches, manually disable ("shutdown") the IPL ports (ports 59 and 60).

Step 5: Acquire the system software

Red Hat Enterprise Linux (RHEL) 7.2 is the system software (operating system) required for all nodes in this solution. Acquire a copy of the RHEL 7.2 iso.

RHEL 7.2 also requires a specific patch to properly support network boots on the particular servers used in this solution. See RHEL 7.2 Network install on IBM Power 8001-12C and 8001-22C and follow the instructions in sections 1.1 and 1.3 to rebuild the initramfs. Name the new file "initrd-i40e.img". This file is required in later steps.

Step 6: Prepare the deployer node

The System Management Node serves as the deployer node in this configuration.

Step 6a: Install RHEL 7.2

Install RHEL 7.2 on the deployer node (the System Management Node). This may be accomplished in any manner convenient for the environment (e.g. manually using a USB storage device and serial connection).

Step 6b: Configure base networking

Configure base networking on the deployer node as follows:

1. Configure the physical port that is cabled to the 1GbE switch with an IP address for the Management Network (e.g. 10.19.60.25). Do not configure any VLANs for this port. Genesis automation uses this same physical port to configure additional network connections in later steps.

2. Configure internet access for the deployer node over the Management Network. The requirements to accomplish this depend upon the nature of the existing network to which this cluster is uplinked. Typically (and minimally), this includes:
   a. Configue a default gateway or route to the internet.
   b. Configure public DNS access.
3. Disable Network Manager.

Step 7: Configure the cluster using the Cluster Genesis tool

Step 7a: Download the HDP on Power Solution Deployment Kit

On the deployer node, log in as a non-root user. Choose or create a workspace directory to contain the various toolkits, deployment kits, and materials that are downloaded and created in following steps. Change to this directory as the shell working directory.

```
[admin@smn ~]$ mkdir toolkit-workspace
[admin@smn ~]$ cd toolkit-workspace
[admin@smn toolkit-workspace]$
```

Using git, download the HDP on Power Solution Deployment Kit into the workspace directory.

```
[admin@smn toolkit-workspace]$ git clone https://github.com/open-power-ref-design/hdp-solution
```

This deployment kit provides scripts, updates, and other materials required in following steps.

Step 7b: Stage inputs for the install and deployment processes

Run the “stage_inputs.sh” script (from within the HDP on Power Solution Deployment Kit) to stage the inputs required for the install, configuration, and update of the toolkits. The RHEL 7.2 iso and associated initrd.img created previously (refer to section “Step 5: Acquire the system software”) are specified explicitly on the script invocation. Other input files are identified and created automatically.

```
[admin@smn toolkit-workspace]$ cd hdp-solution
[admin@smn hdp-solution]$ ./stage_inputs.sh --iso <path to RHEL-7.2 iso> --initrd <path to modified initramfs file>
```

Example:

```
[admin@smn hdp-solution]$ ./stage_inputs.sh --iso /home/admin/master/iso-
RHEL-7.2-GA/RHEL-7.2-20151030.0-Server-ppc64le-dvd1.iso --initrd
/home/admin/master/images/RHEL-7.2-20151030.0-Server-ppc64le-dvd1/initrd-
i40e.img
```

The “hdp-solution-inputs” directory is created in the workspace directory, and all of the input files are staged within that directory.

```
[admin@smn hdp-solution]$ ll ..
total 4
drwxrwxr-x  4 genesis genesis 4096 Jun 24 12:04 hdp-solution
drwxrwxr-x  6 genesis genesis  117 Jun 24 12:07 hdp-solution-inputs
[admin@smn hdp-solution]$
```

Step 7c: Customize the configuration file for the environment

The *config.yml* file holds configuration information that is used by the automation processes which follow to accomplish this specific deployment. The *config.yml* file for this deployment is staged in the “hdp-solution-inputs/” directory, and its contents are in YAML format.

Edit the *config.yml* file to incorporate the base configuration parameters collected or chosen earlier in “Step 3: Choose the basic configuration parameters” on page 9. Refer to Figure 6 on page 14, Figure 7 on page 15, and Figure 7 on page 15 for the specific updates required. For each value indicated in red, find the value in the config.yml file and replace the value with the associated base configuration parameter.

```
...
ipaddr-mgmt-network: 172.19.190.0/24 ← Internal Management Network subnet
vlan-mgmt-network: 190               ← Internal Management Network VLAN ID

ipaddr-mgmt-client-network: 172.19.188.0/24 ← Provisioning/Service Network subnet
vlan-mgmt-client-network: 188               ← Provisioning/Service Network VLAN ID

...
ipaddr-mgmt-switch:
    hdp-rack1: 172.19.190.11         ← Management switch IP address
...
cidr-mgmt-switch-external-dev: 10.19.60.25/24 ← Deployer node initial IP address
ipaddr-mgmt-switch-external:
    hdp-rack1: 10.19.60.11           ← Management switch initial IP address
ipaddr-data-switch:
    hdp-rack1:
        - 172.19.190.13              ← Data switch IP address (switch A)
        - 172.19.190.14              ← Data switch IP address (switch B)
ipaddr-mlag-vip:
    hdp-rack1: 172.19.190.214        ← MLAG VIP address

...
userid-mgmt-switch: admin            ← Management switch login (userid)
password-mgmt-switch: admin          ← Management switch login (password)
userid-data-switch: admin            ← Data switch login (userid)
password-data-switch: admin          ← Data switch login (password)
...
```

Figure 6. Configuration file: Infrastructure level networking and switch parameters

```
...
networks:
...
    data:
        description: Data Network
        bond: bond1
        addr: 172.19.66.0/24              ← Data Network subnet
        available-ips:
            - 172.19.66.220 172.19.66.239 ← Data Network static IP range
        broadcast: 172.19.66.255          ← Data Network broadcast IP address
...
```

Figure 7. Configuration file: Data Network parameters

```
...
userid-default: user                  ← Node OS login (userid)
password-default: passw0rd            ← Node OS login (password)

...
node-templates:
    master:
        hostname: mn                  ← Node hostname stem for Master Nodes
        userid-ipmi: ADMIN            ← Node IPMI login (userid)
        password-ipmi: ADMIN          ← Node IPMI login (password)
...
    edge:
        hostname: en                  ← Node hostname stem for Edge Nodes
        userid-ipmi: ADMIN            ← Node IPMI login (userid)
        password-ipmi: ADMIN          ← Node IPMI login (password)
...
    worker:
        hostname: wn                  ← Node hostname stem for Worker Nodes
        userid-ipmi: ADMIN            ← Node IPMI login (userid)
        password-ipmi: ADMIN          ← Node IPMI login (password)
...
```

Figure 8. Configuration file: Node parameters

Step 7d: Modify and customize Inputs

At this point in the process, all of the inputs required to guide and accomplish the remaining automation steps of the deployment are staged within the “hdp_solution_inputs” directory. However, if additional modifications or customizations to the deployment are desired, such changes must be done before proceeding with the next step.

Step 7e: Finalize inputs

Run the “finalize_input.sh” script to do final adjustment and reconciliation of the inputs.

```
[admin@smn hdp-solution]$ ./finalize_inputs.sh
```

Step 7f: Install Cluster Genesit Toolkit

Install the Cluster Genesis Toolkit by running the “install_genesis.sh” script.

```
[admin@smn hdp-solution]$ ./install_genesis.sh
```

When the above script asks “OK for Cluster Genesis to add setup statements to your .bashrc file?”, answer “y” to update your bash profile.

Run the following command to refresh your shell environment and apply the setup statements noted above.

```
[admin@smn hdp-solution]$ . ~/.bashrc
```

Step 7g: Deploy the cluster

Deploy the cluster using the Cluster Genesis “gen deploy” command.

```
[admin@smn hdp-solution]$ gen deploy
```

The above command accomplishes the base provisioning (incl. OS installation) for the nodes in the cluster, and this step typically runs for an extended period of time (more than 30 minutes). During execution, the following may be required:

1. Entering a sudo password at various points in the process
2. Manually powering on or power cycling each server to be provisioned.
3. If the automated provisioning processes failed to obtain BMC (IPMI) connections to all of the servers, additional manual configuration of the BMCs may be required. Specifically, the server BMCs may need to be reconfigured to obtained an IP address via DHCP. Refer to https://www.ibm.com/support/knowledgecenter/en/linuxonibm/liabw/liabwresetDHCP.htm for more information.

Step 8: Complete the post-Cluster Genesis configuration

Step 8a: Do post-deploy configurations

Do post-deploy configurations using the Cluster Genesis “gen post-deploy” command.

```
[admin@smn hdp-solution]$ gen post-deploy
```

The primary effect of this command is to configure the Data Network and the server interfaces to that network.

After the above command completes, manually enable (‘no shutdown’) the IPL ports (ports 59 and 60) on the 10Gb switches.

Step 8b: Configure the user external network interface to the cluster

Public access for HDP users to the cluster is directed through the Edge Node. On the Edge Node, configure an IP address for the Campus Network. This interface must be VLAN tagged by the node for the Campus Network, and this VLAN is typically part of the client’s existing network environment.

Step 9: Install Hortonworks Data Platform (HDP)

To complete the deployment of the HDP on Power solution, Hortonworks Data Platform software must be installed on the cluster. The installation of the HDP software is beyond the scope of this document. It is recommended that a Hortworks consultant be retained to accomplish this step.

References

The following links and documents provide more information relevant to this document:

- IBM Cluster Genesis Documentation
- Hortonworks HDP Home Page

Appendix A: How to add a node to the cluster before deployment

The configuration includes the following: 8 worker nodes, 3 master nodes, and 1 edge node. An additional node for any of these node types may be added to the configuration prior to doing the automated deployment. To do so:

1. Select an available rack location for the additional node.
2. Select the network switch ports to which the node is to be cabled. It is recommended that the port selections follow the same pattern used for the other nodes in the configuration (see “Step 2c: Install network cables” on page 4).
3. Rack and cable the additional node consistent with the instructions in “Step 2: Rack and cable the hardware” beginning on page 2.
4. Update the config.yml file to indicate the additional node. Specifically, add to the config.yml the switch-side port number for each port (four total) that is connected to the additional node. Within the “node-templates” section, under the node type subsection which applies (i.e. “worker”, “master”, or “edge”), add the switch-side port numbers to the list of ports for each network port group. The new ports may be added to any position in the lists, but ***each of the ports for the additional node must be added in the same position in each port group list*** (e.g. the additional port for each port group must be added to the beginning of the list for each port group). The port groups are:

    - “pxe”, which represents the 1Gb, OS-level connections to each nodes
    - “ipmi”, which represents the 1Gb connections to the BMCs of each nodes
    - “eth10”, which represents the first of the 10Gb connections to each node
    - “eth11”, which represents the second of the 10Gb connections to each node

The updates to the config.yml must be completed as part of “Step 7d: Modify and customize Inputs” on page 16. The racking and cabling should be completed as part of “Step 2: Rack and cable the hardware” on page 2, but it must be completed prior to beginning “Step 7g: Deploy the cluster” on page 17.

Example: An additional master node is to be added to the configuration. The node is physically installed in rack location U32. The next available switch port for each port group (next lower port number) is selected for the connections to the additional node. Specifically:

- The **1Gb port** on the node (“pxe”) is cabled to the **1Gb switch port 23**
- The **BMC port** on the node (“ipmi”) is cabled to the **1Gb switch port 5**
- The **10Gb port A** on the node (“eth10”) is cabled to the **10Gb Data Switch A port 5**
- The **10Gb port B** on the node (“eth11”) is cabled to the **10Gb Data Switch B port 5**

The config.yml is updated as shown in Figure 9 as part of “Step 7d: Modify and customize Inputs” on page on page 16, and the additional node is subsequently provisioned as part of the cluster.

```
...
node-templates:
    master:       ← Within the “node-template” section for “master” nodes...
...
        ports:
            pxe:
                hdp-rack1:
                    - 23 ← Add entry for 1Gb port connection on additional node
                    - 25
                    - 26
                    - 27
            ipmi:
                hdp-rack1:
                    - 5  ← Add entry for BMC port connection on additional node
                    - 7
                    - 8
                    - 9
            eth10:
                hdp-rack1:
                    - 5  ← Add entry for 10Gb port A connection on additional node
                    - 7
                    - 8
                    - 9
            eth11:
                hdp-rack1:
                    - 5  ← Add entry for 10Gb port B connection on additional node
                    - 7
                    - 8
                    - 9
...
```

Figure 9. config.yml changes for ‘add a node’ example

Printed in the United States of America May 2017

IBM Systems
294 Route 100, Building SOM4
Somers, NY 10589-3216

The IBM home page can be found at **ibm.com®**.

Version 1.0
August 2, 2017