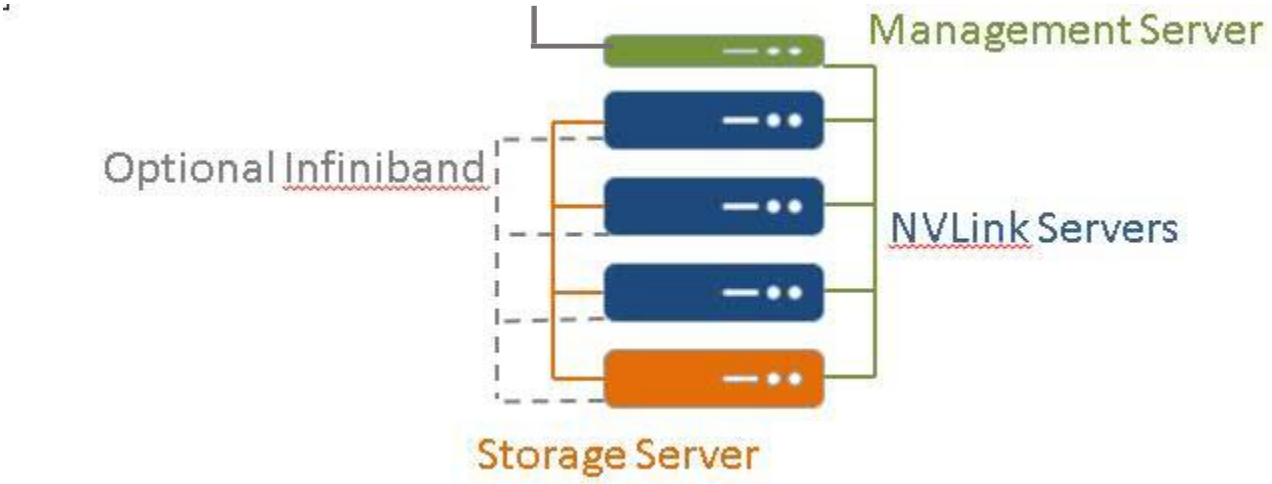
How to deploy IBM OpenPower Cluster for Deep Learning

A Multiple Node Deployment for TensorFlow

Solution Summary

This solution consists of this document, a prescribed Bill of Materials (BoM), and a deployment configuration file. The Bill of Materials document provides a description and representation of a PowerAI TensorFlow installation across multiple OpenPOWER servers. It provides information such as model numbers and feature codes to simplify the ordering process and it provides the racking and cabling rules for the preferred layout of the servers, switches and cables.

A high-level diagram representation of a minimal deployment of the solution is below:



The hardware consists of NLink capable Power8 servers with an additional server for shared storage and an additional server for deployment and management tasks. In the design, there are dedicated networks for storage (**orange**) and management (**green**). The solution includes an option for an additional high speed, low latency InfiniBand network (**dashed grey**) for tensor communication.

The Management server is used to manage the configuration and, optionally, automatically deploy the entire solution. This deployment automation is accomplished

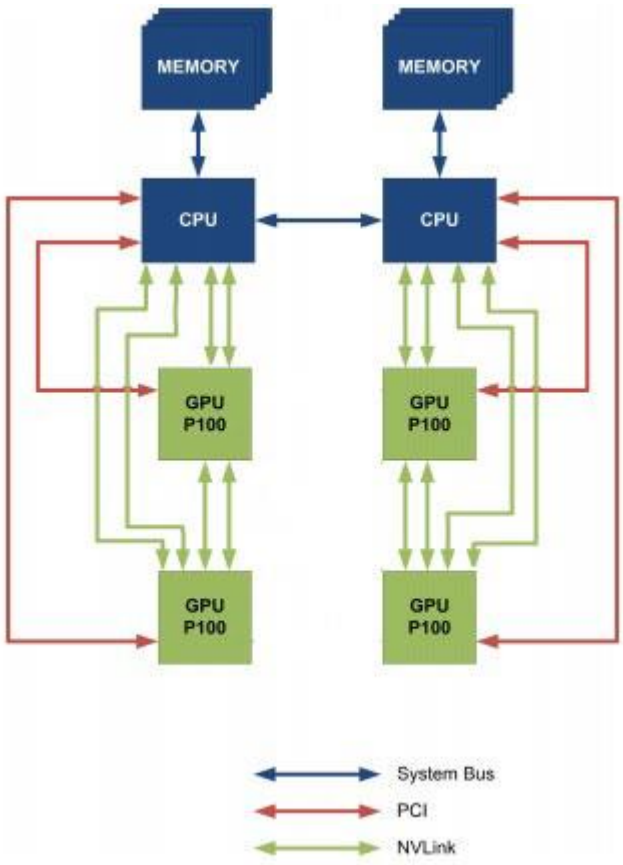
using a tool called **Cluster Genesis** which uses the configuration file as input to install the nodes and subsequently install and configure the needed software for the solution.

The included deployment configuration file provides a mapping of servers and switches to software for deployment. Software is mapped by assigning each server a software *role*. A TensorFlow training environment requires an entity to perform a parameter serving role for updating and sharing parameter data. In addition to this **Parameter Server** role, there are also the **Storage Server**, and **Training Compute** roles.

The Parameter Server role and the Storage Server role are assigned to the same physical server: The Storage Server. The NVLink servers are given the training compute role.

Compute Node Architecture

The Power8 Training Compute servers have a unique architecture that includes the high bandwidth NVLink bus between each of two Power8 CPUs and the pair of GPUs they are attached to. NVLink is NVIDIA's high speed interconnect technology for GPU-accelerated computing. The NVLink provides an increased bandwidth to GPUs as well as a reduction in latency and code paths lengths. The high-level architecture design of the compute node servers is described in the diagram below.



The initialization of the GPUs is done through the PCIe interfaces shown above. The PCIe interfaces also contain side band communication for status, power management, and so on. Once the GPU is up and running, however, all data communications is done using the NVLink bus.

High Level Deployment Steps

	EACH STEP BELOW IS DESCRIBED IN MORE DETAIL BELOW
1	Acquire the hardware
2	Rack and cable the hardware
3	Prepare the deployer node
4	Choose your configuration parameter for the solution
5	Configure the cluster (Genesis)

Step 1: Acquire the hardware

Go to the link below for the Bill of Materials list of required parts.

http://github.com/open-power-ref-design/deep-learning/documentation

If you do not already have the needed parts, contact an IBM representative to help.

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

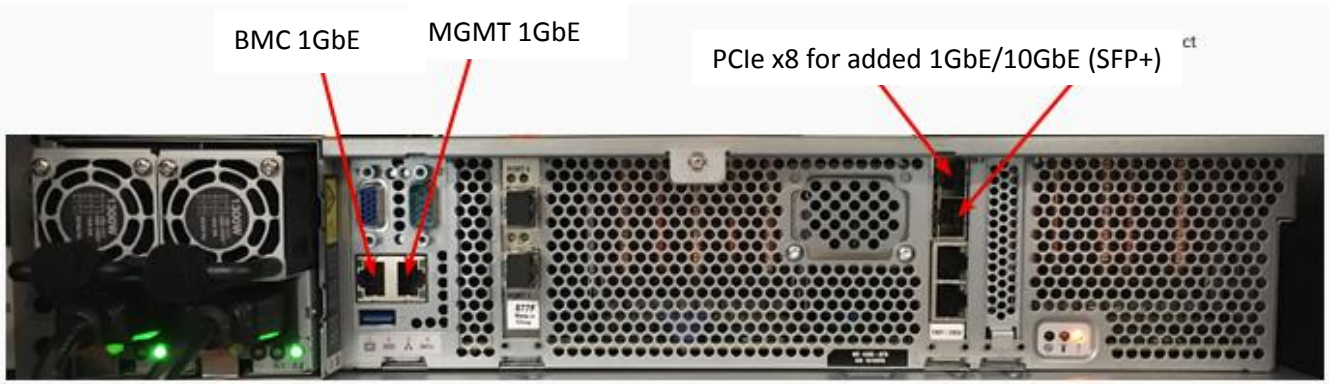
Step 2: Rack and cable the hardware

This section describes the hardware details, port listings and cable connection information needed to properly cable the solution.

Hardware orientation

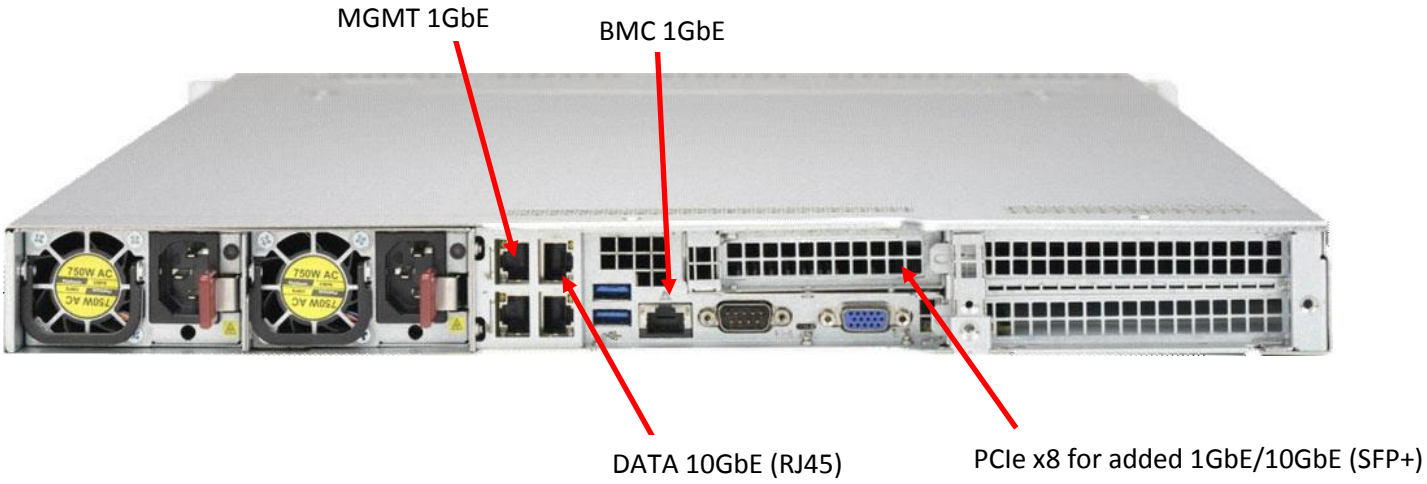
Compute Node

The illustration below shows the back of the compute nodes (“S822LC for HPC”) and its base networking. Note that while these servers can share the BMC service port (a multi-function port), the automation requires the port to be set up for BMC data only.



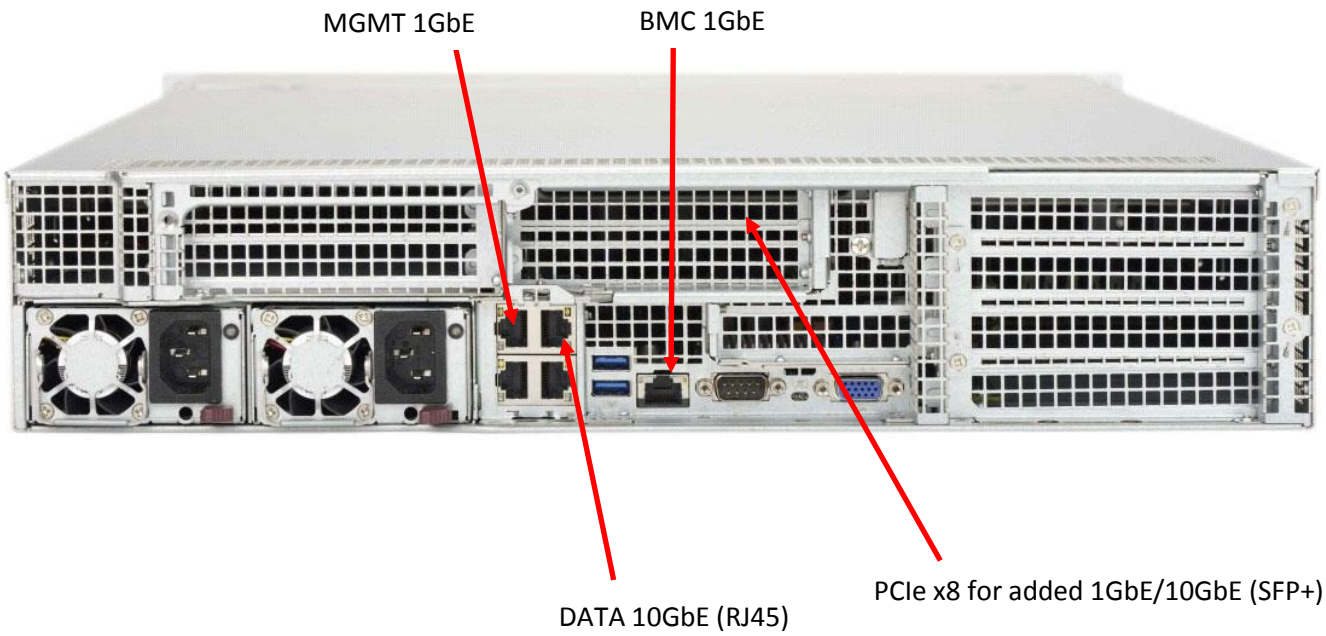
Management Node

The management node (S812LC) rear view is shown below. The four native Ethernet (RJ45) ports support 1GbE, 10GbE, or 100MbE and can be configured for separate speeds. This solution also includes a 10GbE card (SFP+) in the PCI3 Slot3. Use the appropriate port based on the switch being used. The 10GbE connection from the management node is optional and can be used to transfer data into the cluster faster, if desired.



Storage Node

The storage node (S822LC) rear view is shown below. The four native Ethernet ports (RJ45) support 1GbE, 10GbE, or 100MbE and can be configured for separate speeds. The solution also includes a 10GbE card (SFP+) in the PCI3 Slot3. Use the appropriate port based on the switch being used.



Racking the components

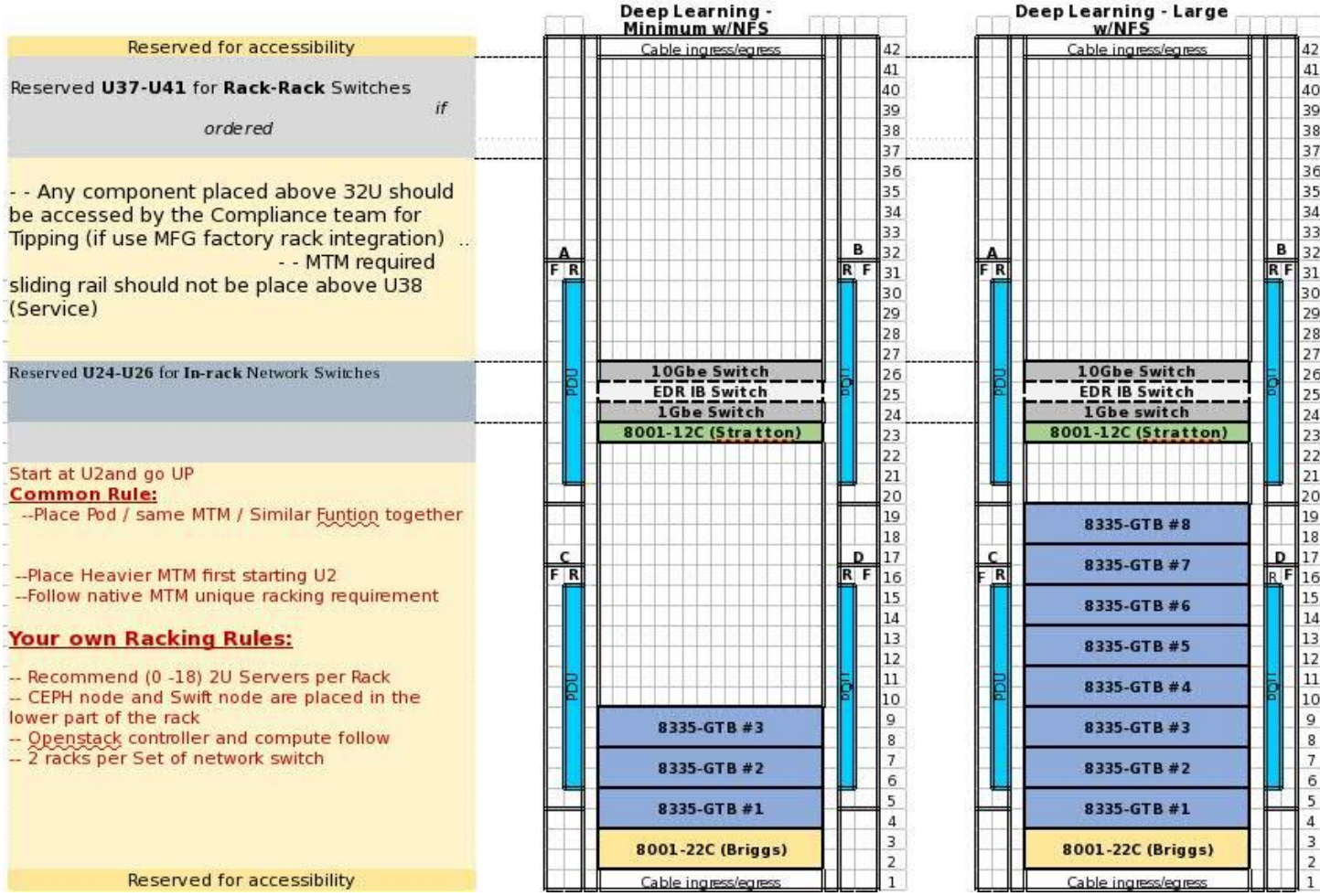
This section specifies racking rules that specify both where to place the servers and switches

And where to connect the cables.

These suggested racking rules focus on enabling:

- Rack modularity
- Consistency
- Expandability
- Ease of servicing, repurposing, shipping, and cooling

Suggested Racking Rule



Place the intra-rack (leaf) network switches in U24-U26 as follows:

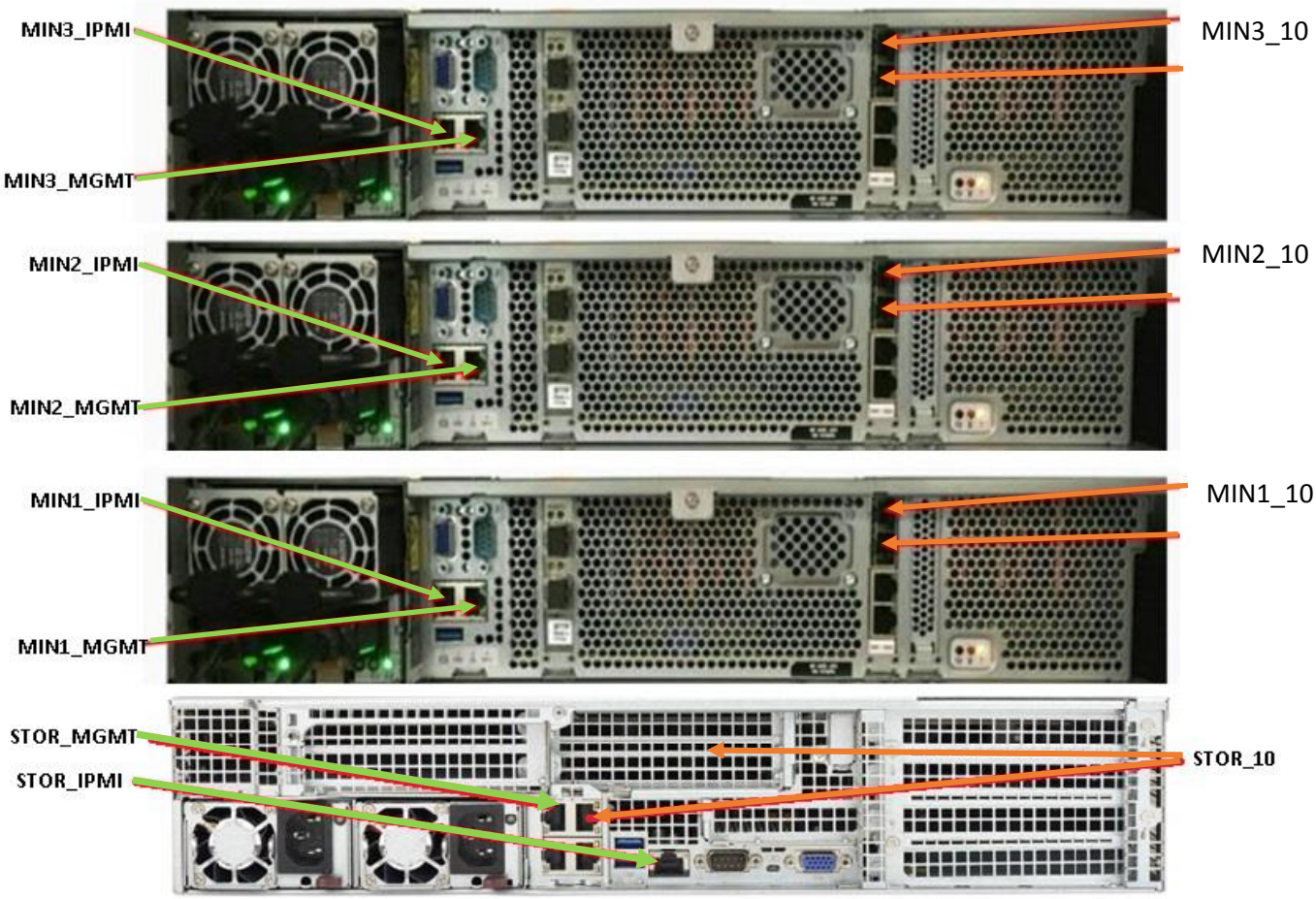
- 10G storage network switch in U26
- EDR IB tensor network switch in U25
- 1G management network switch in U24

Place inter-rack (spine) switches in slots U37-U41

If more than 4 PDUs are needed, place 3 horizontal PDUs in 40U and 41U. Spine switches take priority over additional PDUs. If 40U and 41U are occupied by spine switches, place horizontal PDUs in next available slots.

Cable together the components

Rear server view with labels for three compute nodes names MIN1, MIN2, and MIN3 along with the storage node. The IPMI enabled service port and the onboard 1GbE network port connect to the management network (**green**), while the 10GbE port connects to the storage network (**orange**).



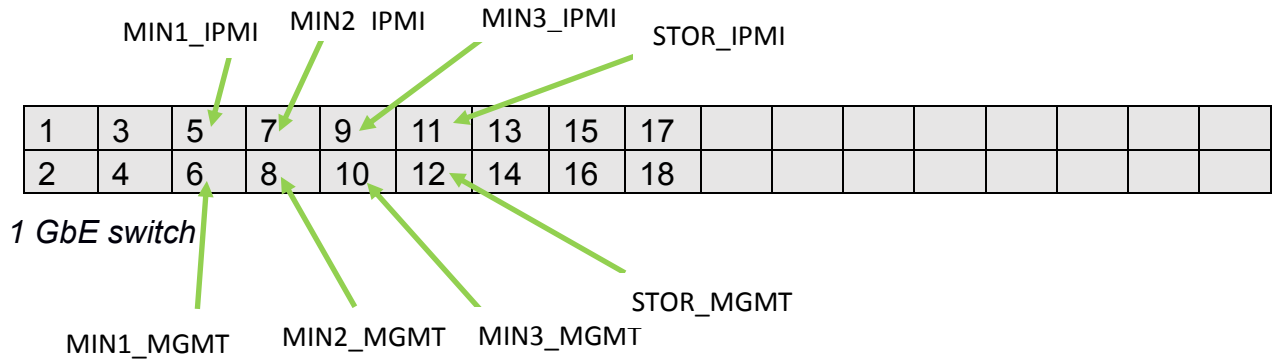
The Optional InfiniBand adapters connect to any available ports on the InfiniBand switch.

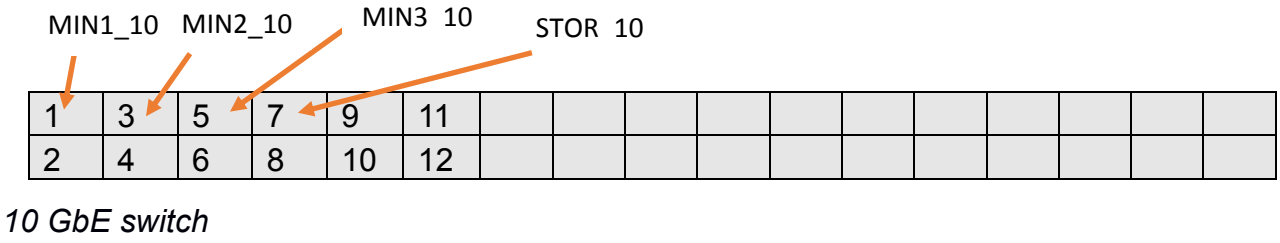
Network Switches

The Cluster Genesis automation tooling manages the included network switches. It uses the switch access to identify MAC addresses of the connected servers. The configuration file provides a mapping of systems to ports. The physical cabling needs to reflect this port mapping that is in the configuration file. Currently Genesis supports Lenovo switches in the automation. (see the BOM)

The network switch view is below. The 1GbE switch connects to both the IPMI and management ports on each server. The connections below reflect what is defined in the default configuration file. This can be customized as needed.

The 10GbE switch creates the storage networks and connects to a 10GbE port on each server.





If the InfiniBand option is being used, there will be a third (InfiniBand) switch. Genesis does not currently manage InfiniBand switches, so the port connection on the switch do not matter from server to server.

Mellanox InfiniBand switches have been tested and are recommended.

Step 3: Choose your basic config parameters

To facilitate the automation process of the solution, the parameters in the following table need to be collected. This data will be edited into the configuration file which will be used to automatically configure and deploy the entire solution.

<table>
<tr><th>Parameter</th><th>Description</th><th>Example</th></tr>
<tr><td>Domain Name</td><td></td><td>Ibm.com</td></tr>
<tr><td>Upstream DNS Servers</td><td>While a DNS server is configured within the cluster, upstream DNS servers need to be defined for names the cannot otherwise be resolved.</td><td>*4.4.4.4, 8.8.8.8 as default public upstream DNS servers</td></tr>
<tr><td>Deployment Node Host Name</td><td>What do you want to call your deployment node?</td><td>mgmtnode</td></tr>
<tr><td>Management network IP address</td><td>Management for the cluster happens on its own internal network.</td><td>192.168.3.3.24</td></tr>
<tr><td>Data network IP address</td><td>Labeled interconnect in the config file</td><td>10.0.0.1/24</td></tr>
<tr><td>Management switch IP address</td><td>Labeled ipaddr-mgmt-switch in the config file</td><td>192.168.3.5</td></tr>
<tr><td>Data switch IP addresses</td><td>Labeled ipaddr-data-switch in the config file</td><td>9.3.3.178</td></tr>
<tr><td>Default login data</td><td>Both IDs and passwords</td><td>BMC network, OS Mgmt network</td></tr>
<tr><td>Data node hostnames and IPs addresses</td><td>Each node in the cluster needs a hostname and IP address for each of the management and data networks.</td><td>
<table>
<tr><th>Name</th><th>Management IP</th><th>Data IP</th></tr>
<tr><td>Min-1</td><td>192.168.3.102</td><td>10.0.0.2</td></tr>
<tr><td>Min-2</td><td>192.168.3.104</td><td>10.0.0.4</td></tr>
<tr><td>Min-3</td><td>192.168.3.106</td><td>10.0.0.6</td></tr>
<tr><td>Min-4</td><td>192.168.3.108</td><td>10.0.0.8</td></tr>
</table>
</td></tr>
</table>

Step 4: Prepare the management node

The management server is the multi-homed head node of the cluster. It has access to any external networks as well as the internal cluster networks. Additionally, when using Cluster Genesis for automated deployment, the management node will automatically check out the latest software and deployment tools to install and populate the cluster. We've defined the management node as a 1U OpenPOWER server, but it can also be an x86 server with at least 32GB RAM.

Ubuntu 16.04 is needed on the management server. You can obtain it from the following locations:

- Power8 servers: https://www.ubuntu.com/download/server/power8
- X86 servers: https://releases.ubuntu.com/16.04.1/

Check the “Manual Cluster Deployment” section for instructions for manually installing Ubuntu to OpenPOWER servers using the Ubuntu netboot resources., if needed.

Step 5: Deploy the cluster (Genesis)

A correct and complete configuration file allows the Cluster Genesis tooling to power on, initialize, configure and install the cluster. This deployment kit provides an automated method to quickly and more predictably go from physical rack and stack to an operational state.

	GENESIS AUTOMATICALL INITIALIZES AND CONFIGURES THE HARDWARE BY…
A	Reading the configuration files with edited environment-specific changes
B	Driving the BMCs to populate the IP addresses to the nodes
C	Detecting and populating relevant config data to the management node
D	Deploying needed operating system images to the server nodes
E	Configuring the network switches (some manual steps are required)
F	Installing required software once the nodes in the cluster are up

Check out the solution and obtain the default configuration file

Use the git clone command to check out the solution to the management server.

```
$ git clone https://www.github.com/open-power-ref-design/deep-learning
```

The genesis automation uses a configuration file written in YAML to specify the target cluster configuration. The deployment uses this YAML text file to specify the IP address

locations of the managed switches and the system nodes attached to the switches as well as other useful details for the deployment process.

The generic master copy of the configuration file, which assumes a 4-node cluster (3 compute, 1 storage/parameter) is located here:

https://www.github.com/open-power-ref-design/deep-learning/config.yml.tfdist.4min

Tailor the configuration file for the environment

The configuration file contains a lot of information. To enable a cluster tailored to a specific environment, edit the configuration file with the parameters that were collected in step 2. Refer to the illustration below and replace the RED text with your data.

```
reference-architecture:
  nvidia_playbook:
    description: playbook for installing cuda repository
    cuda_deb: “packages/cuda8.deb”
    cudnn5_deb: “packages/cudnn5.deb”
    cudnn_dev_deb: “packages/cudnn_dev.deb”
    dkms_deb: “packages/dkms.deb
  powerai_playbook:
    description: Playbook for installing Power machine learning
frameworks
      mldl_deb: “packages/mldl.deb”
      powerai_frameworks:
        - tensorflow
      sameples_src: “static/tensorflow/distributed/samples”
      samples_dest: “/opt/DL/tensorflow/samples”
      ofed_driver: “packages/ofed.tgz”
      models_tree: https://guthub.com/tensorflow/models.git
```

The `powerai_frameworks` section (above) tells Genesis to install the frameworks listed. Currently the frameworks supported in PowerAI include: ibm-caffe, nv-caffe, bvlc-caffe, chainer, tensorflow, theano, and torch. The `ofed_driver` parameter is required when using the optional InfiniBand network.

```
ipaddr-mgmt-network: 192.168.3.0/24
ipaddr-mgmt-switch:
     rack1: 192.168.3.5
ipaddr-data-switch:
     rack1: 10.0.3.178
redundant-network: false
userid-default: ubuntu
password-default: password
userid-mgmt-switch: admin
password-mgmt-switch: admin
userid-data-switch: admin
password-data-switch: admin

nfs:
     nfs_network: interconnect
     nfs_server: parameter
     nfs_clients: worker
     nfs_shares:
      -  /data
```

The `ipaddr-mgmt-network` parameter defines the network space for the management network. The `ipaddr-*-switch` parameters have the actual management port addresses for the management and data switches themselves. The user and passwords for the two switches are also needed.

The next section defines the networks in the solution. Each network gets a network space, method (static vs dhcp) and a specified network interface name.

```
networks:
  external:
     addr: 1.2.3.0/24
     broadcast: 1.2.3.255
     gateway: 1.2.3.1
     dns-nameservers: 1.2.3.200
     dns-search: example.com
     method: static
     eth-port: eth10
  interconnect:
    description: Private 10G Storage Network
    addr: 10.0.0.0/24
    broadcast: 10.0.0.255
    method: static
    eth-port: eth11
  mgmt-pxe:
    description: Management Network
    method: dhcp
    eth-port: eth15
```

InfiniBand Option

If the InfiniBand option is being used, a few sections need to be in the configuration file, in addition to the `ofed_driver` parameter described above. A stanza for the InfiniBand network needs to be added to the networks section, and that network name needs to be referenced in each of the node template sections. See below for specific additions needed.

```
networks:
    infiniband:
      description: Private IPoIB Network
      addr: 10.0.1.0/24
      broadcast: 10.0.1.255
      method: static
      eth-port: ib0
```

```
node-templates
  worker:
    networks:
      - external
      - infiniband
      - interconnect
      - mgmt-pxe
  parameter:
    networks:
      - external
      - infiniband
      - interconnect
      - mgmt-pxe
```

The configuration file has now been customized for the environment. For this document, we'll refer to this new file as **myconfig.yml.**

Additional Non-automatable steps

There are a few steps that are not automatable today. These include software dependencies that are currently behind a required login and/or manual click-through license acceptance. Complete the steps below before running the cluster automation to manually download the required software dependencies.

Create the management network bridge

During deployment, the Cluster Genesis tool creates a container that runs the boot time services. A bridge must be created for this container to have access to the management network. Creation of this bridge is a manual step.

On the management node, edit the `/etc/network/interfaces` file and add the lines below. The RED text should be customized for the environment. Note that the interface in the `bridge_ports` stanza (enP1p3s0f0) is the interface connected to the management network.

```
auto br0
iface br0 inet static
      address 192.168.3.3
      netmask 255.255.255.0
      bridge_ports enP1p3s0f0
```

NVIDIA CuDNN

NVIDIA's CuDNN library is a prerequisite, however it requires a user registration for

download. To download NVIDIA's latest CuDNN packages, visit the NVIDIA website.

```
https://developer.nvidia.com/cudnn
```

Login or register for NVIDIA's Accelerated Computing Developer Program. Download the following deb files:

- cuDNN v5.1 Runtime Library for Ubuntu16.04 Power8 (Deb)
- cuDNN v5.1 Developer Library for Ubuntu16.04 Power8 (Deb)

Copy the .deb file to the management server and export the location as the CUDNN5 and CUDNN5_DEV environment variables so Genesis will know the location.

For example:

```
export CUDNN5=/home/Ubuntu/Downloads/libcudnn5_5.1.5-
1+cuda8.0_ppc64el.deb
export CUDNN5_DEV=/home/ubuntu/Downloads/libcudnn5-dev_5.1.5-
1+cuda8.0+ppc64el.deb
```

Mellanox OFED

If using the optional InfiniBand network, download Mellanox OFED from the Mellanox website. Visit the link below to download the correct package. There is a click-through table at the bottom of the page in the *Downloads* tab. Download the .tgz version of the package

```
http://www.mellanox.com/page/products_dyn?product_family=26&mtag
=linux_sw_drivers
```

Once downloaded, copy the .tgz file to the management server and export the location as the MLX_OFED environment variable so that Genesis can find it. For example:

```
export MLNX_OFED=/home/Ubuntu/Downloads/MLNX_OFED_LINUX-4.0-
2.0.0.1-ubuntu16.04-ppc64le.tgz
```

Note: if using the InfiniBand option, a session manager is required. It is recommended to enable the InfiniBand session manager on the IB switch. InfiniBand switch management is not currently included in Genesis.

Perform the deployment (Genesis)

Run the install script.

```
$ install.sh
```

The install script checks out Cluster Genesis from its own github repository, applies patches, and downloads the various dependent packages needed for the install. These dependencies include cuda, PowerAI, and a few specific Ubuntu packages needed during the automated deployment.

Once the install.sh is run cleanly, start the automated Genesis deployment:

```
$ deploy.sh myconfig.yml
```

The Inventory File

During the Genesis deploy, an entire inventory of the cluster is taken and written into a separate YAML file. This file is the de-facto database for the cluster. It consists of the network switches and server nodes. The data structure for the switches indicates the type of switch (management, data), their IP addresses and associated login credentials. Similarly, the server nodes entries also contain access information, descriptive information, and general details. This file should not be edited manually.

After Genesis is complete, this inventory file is located on the deployment node in `/var/oprc/inventory.yml`.

Step 6: Complete post-Genesis configuration

TensorFlow run scripts

The Genesis automation will fully install the nodes with an operating system and needed software, including TensorFlow. It also will place a run script on each node at:

```
/opt/DL/tensorflow/samples/dist_driver.sh
```

The script is personalized for each node and can be used to run distributed models on the cluster. Note that not all TensorFlow models implement distributed training, they must be explicitly written and tuned to take advantage of the distributed TensorFlow functions.

Shared Storage

A shared NFS shared storage disk is created by Genesis and the TensorFlow models library is checked out there. The default mounted NFS share location for these models is:

```
/data/models
```

There is a distributed version of Inception V3 included in the model library. ImageNet is a common academic data set to use with the Inception V3 model and can be downloaded from:

```
http://www.image-net.org
```

Appendix A: Resources

Resources on distributed TensorFlow models

TODO

Appendix B: Notices

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

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

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

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
United States of America

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

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

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.
This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.
Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.
IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

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

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

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

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on
generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.
Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and

cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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

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

Portions of this code are derived from IBM Corp. Sample Programs.
© Copyright IBM Corp. 2016. All rights reserved.

Trademarks
IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" (www.ibm.com/legal/copytrade.shtml).
Java™ and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

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

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

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

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

Rights

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any information, data, software or other intellectual property contained therein.
IBM reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by IBM, the above instructions are not being properly followed.

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

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

IBM Online Privacy Statement

IBM Software products, including software as a service solutions, (“Software Offerings”) may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user, or for other purposes. In many cases, no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering’s use of cookies is set forth below.

This Software Offering does not use cookies or other technologies to collect personally identifiable information.

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, see IBM’s Privacy Policy at http://www.ibm.com/privacy and IBM’s Online Privacy Statement at http://www.ibm.com/privacy/details in the section entitled “Cookies, Web Beacons and Other Technologies”, and the “IBM Software Products and Software-as-a-Service Privacy Statement” at http://www.ibm.com/software/info/product-privacy.