

How to Deploy IBM OpenPower Cluster for Accelerated Databases

Version 1.0

INTRODUCTION

This document along with referenced links describes a comprehensive set of instructions, rules, and automation tools for building an OpenPOWER-based cluster tuned for Accelerated Databases. Examples of such accelerated databased include Kinetica's GPUdb.

Step 1 lists the hardware needed for this solution, including BOM. Step 2 designates the lab infrastructure configuration parameters needed to build solution. Step 3 is a high level description of the deployer node and initial preparation. Step 4 specifies a generic hardware build instructions including BOM contents for each building block and build rules (racking and network options). Step 5 describes the bare metal machine genesis processes used in Accelerated Database solutions. Step 6 is any desired post Genesis configuration(currently all within the automation). Step 7 is Accelerated Database operational manager or software orchestrator (not supported in first release). Finally, step 8 is for Kinetica's GPUdb specific instructions and documentation.

HIGH LEVEL DEPLOYMENT STEPS

EACH STEP BELOW IS DESCRIBED IN MORE DETAIL BELOW

1	Acquire the hardware
2	Choose your configuration parameter for the solution
3	Prepare the deployer node
4	Rack and cable the hardware
5	Configure the cluster (Genesis)
6	Complete any post Genesis configurations
7	Install open source management software (not yet supported for this deployment kit)
8	Install the applicable Accelerated DB software (not automated in this deployment kit)

STEP 1: ACQUIRE THE HARDWARE

Go [here](#) for the Bill of Materials list of required parts:

Go [here](#) to contact an IBM representative for order/purchase assistance.

STEP 2: CHOOSE YOUR BASIC CONFIG PARAMETERS

To facilitate faster automated configuration of the overall solution, collect the following parameters before starting. This data will be edited into a config.yml file which will in turn be used to automatically configure and deploy the entire solution.

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

Go [here](#) to see more options in the config.yml file.

STEP 3: PREPARE THE DEPLOYER NODE

The deployer node is used to obtain the latest software and deployment tools from Github and populate the cluster. You can establish the deployer node as a temporary or permanent server. It can be any Power8 LC or x86 server with the following minimum characteristics:

- 2 cores and 32G RAM
- 3 Network Interface connections: IPMI, 10G (highspeed), 1G(Mgmt).
- Ubuntu 16.04 LTS before beginning with deployment.

If you do not already have Ubuntu, you can obtain it from the following:

- Power8-LC servers: <https://www.ubuntu.com/download/server/power8>
- For x86 servers: <http://releases.ubuntu.com/16.04.1/>

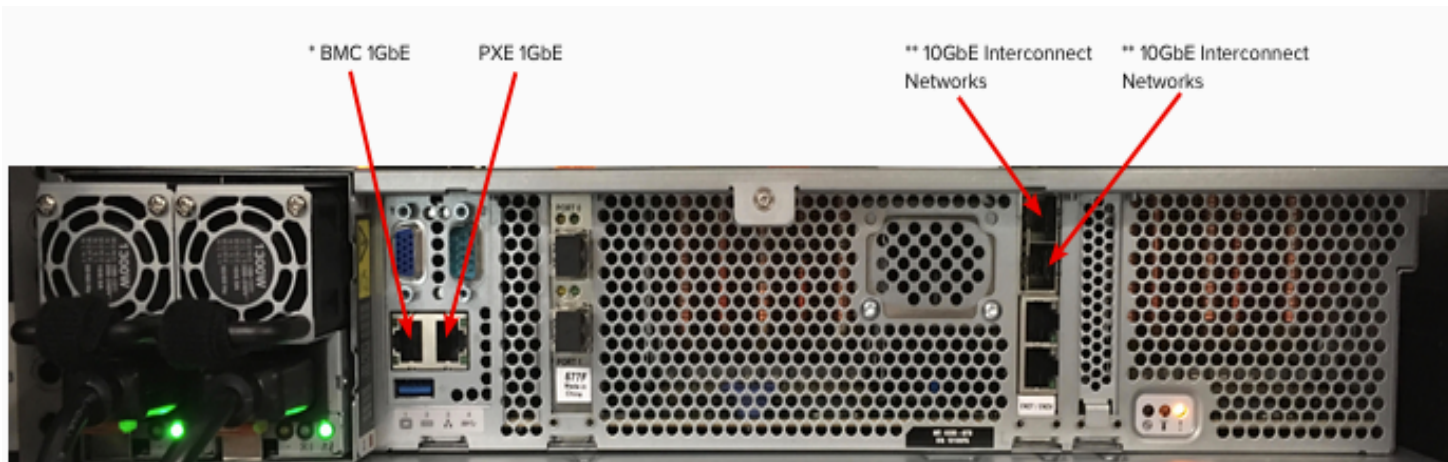
STEP 4: RACK AND CABLE THE CLUSTER

GPU-accelerated databases are best optimized with Power's unique high-speed NVLink bus between its CPU and NVIDIA GPUs. As a result, this design prescribes the IBM S822LC for HPC server (product ID: 8335-GTB) to enable the best possible database performance.

Go [here](#) to see its specific configuration.

The Power server's network adapters should be configured as shown in the illustrations below for a system named MIN.

Back view:MIN



Note:

* While these servers are capable of sharing ports(multi-function port), automation requires the port to be set up as BMC data only.

** If bonded networks are desired, two dual port adapters are required.

RACKING THE COMPONENTS

Instead of providing a long list of step-by-step instructions to build up the rack or servers and switches, 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:

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

Reserve U42 for accessibility.

Reserve U37-U41 for rack-to-rack network switches to enable multi-rack expansion. This will reduce the maximum number of data servers containable by the first rack.

Ensure that any component inserted above 32U are approved by your compliance team to prevent rack tipping.

Reserve U24-U26 for in-rack network switches. Multiple switches enable ...

Reserved for TF4 Display if selected

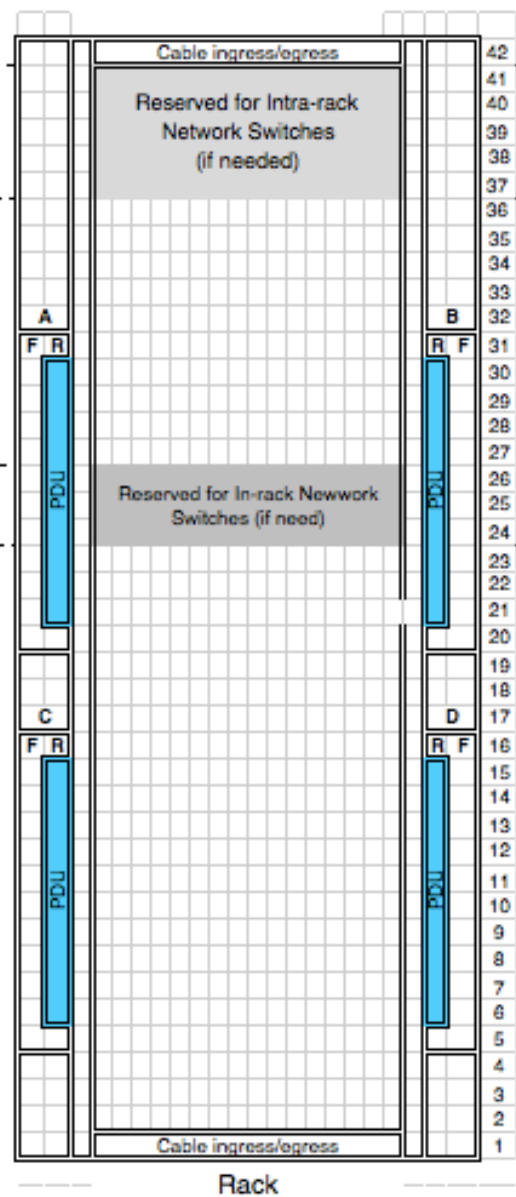
General rules:

- Place the 1st and heaviest components at U2 and work up.
- Keep common units (same MTMs, building blocks) together
- Follow native MTM unique racking requirement.

Solution-specific rules:

- Place seeding node / license node at U2
- Place and cable between 4-18 2U servers per rack

Reserve U1 for accessibility



ADDITIONAL RACKING RULES

This version of the accelerated DB design supports in-server storage only. A future version will add support for external storage.

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

- 10G/40G/IB data plane switch in **26U** (parts 8828-E36, 8831-NF2, or 7120-64C)
- Leave **U25** open. Reserve for later use of short-depth devices (see switch placement restriction above).
- 1G management plane switch in **24U**.

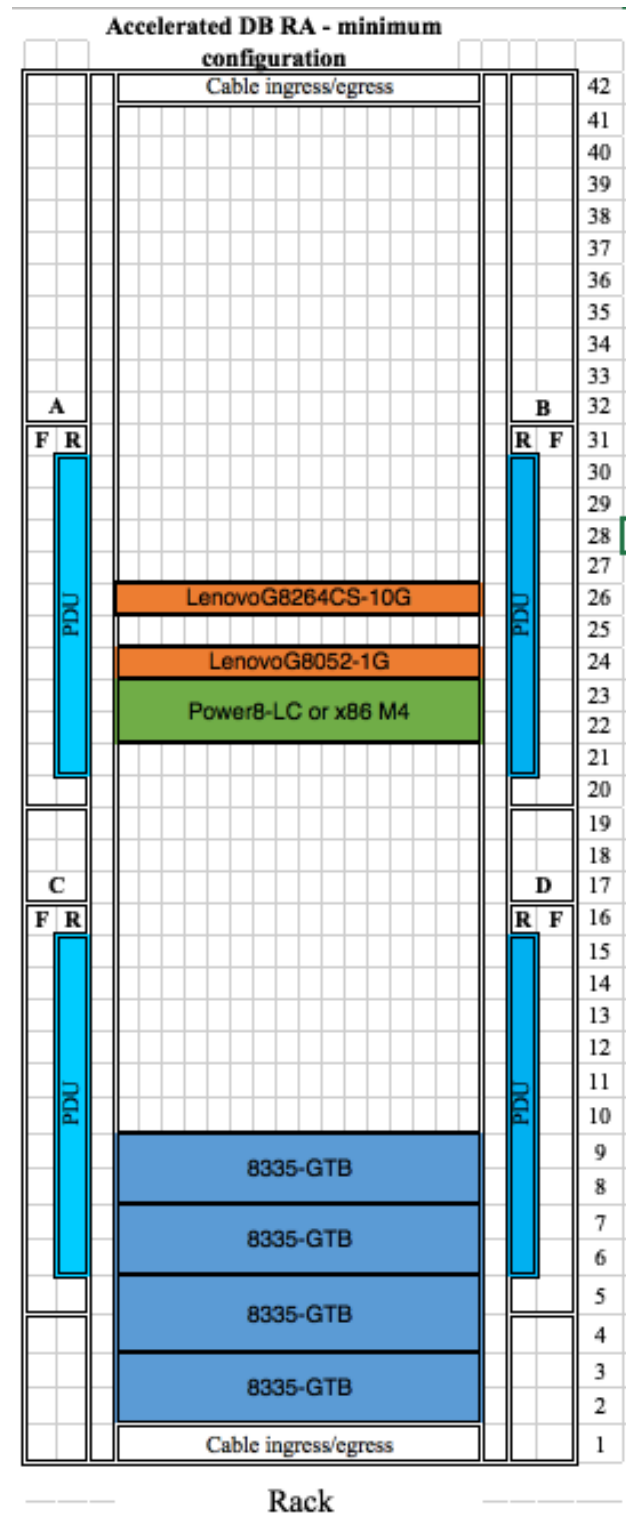
Place the inter-rack (spine) switches in slots **U37-U41** as follows:

- You can design the modular rack block as (every 2 racks if required). See the **48x High Speed TOR Switch Option**.

No more racks can be ordered than what is needed to hold the servers.

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.

RESULTING EXAMPLE: MINIMUM FOUR 8335-GTB SYSTEM CLUSTER

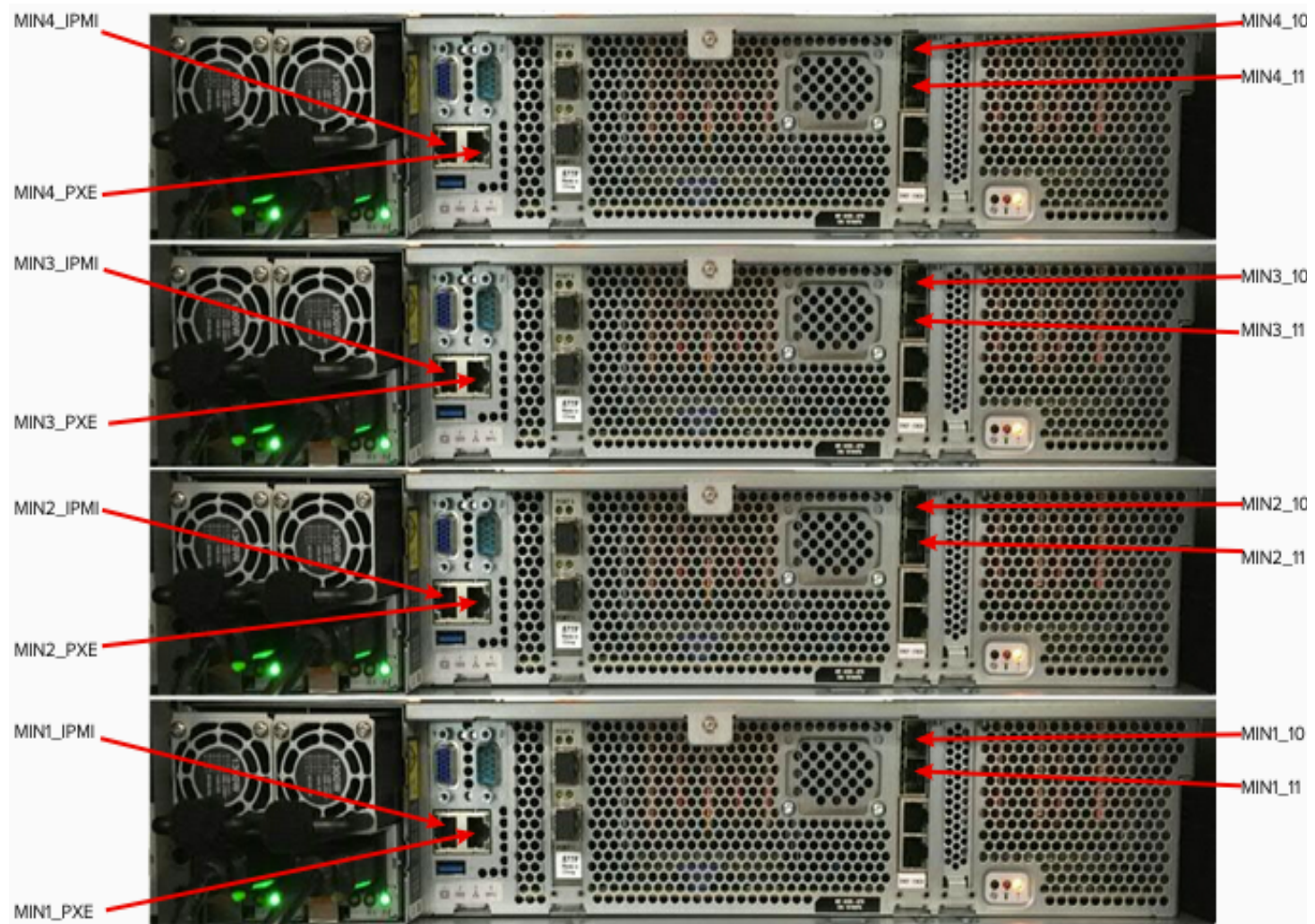


CABLE TOGETHER ALL COMPONENTS

To follow these cabling rules, simply cable the like labels on the servers to the applicable network switches. Shown here is the approach for a four server cluster configuration.

EXAMPLE: MINIMIM FOUR 8335-GTB SYSTEM CLUSTER NETWORKING

Rear server view with labels for four servers named MIN1, MIN2, MIN3, and MIN4...



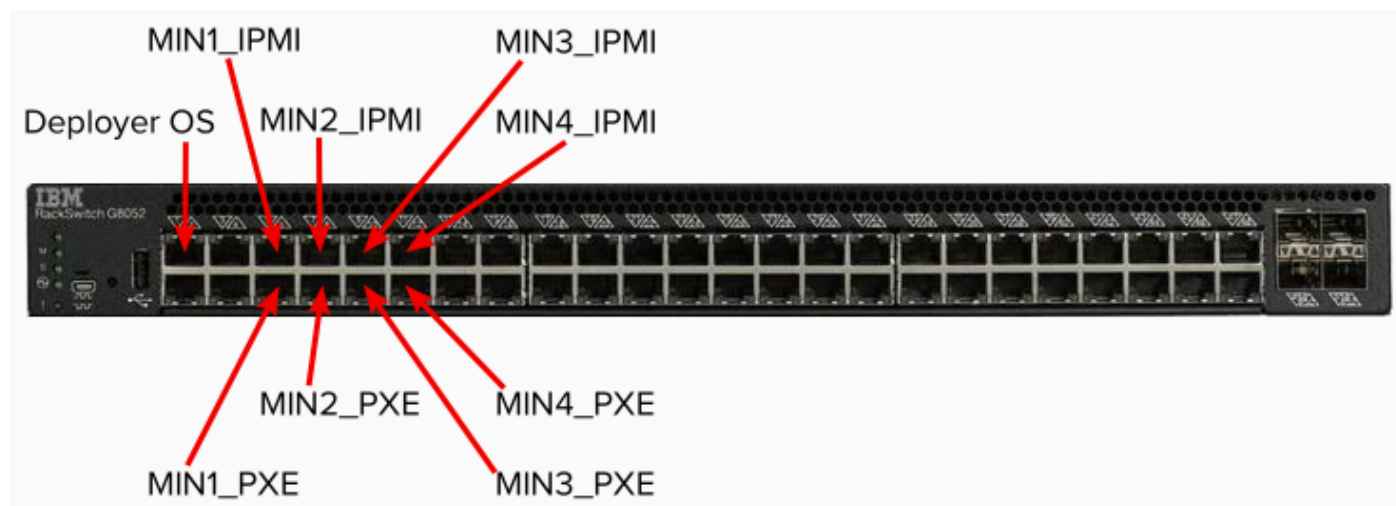
Network switch port view. The Labels on the ports shown on the network switch below correspond to the server labels above and also to what you would see in the config.yml and inventory.yml.

In this example, 40GbE-to-4x10GbE fanout cables were used

Lenovo G8264CS - 10GbE data network switch cabling scheme below:



Lenovo G8052 - 1GbE management network switch cabling scheme below:



STEP 5: CONFIGURE THE CLUSTER (GENESIS)

This section covers the power on, initialization, configuration and installation of a cluster solution. This deployment kit provides an automated method to quickly and more predictably go from assembly to a tuned operational state of the cluster's infrastructure. We call it *hardware genesis*.

While genesis occurs once at the beginning of the cluster solution lifecycle, the open-sourced automation scripts are available and can be reused for maintenance and cluster expansion.

GENESIS AUTOMATICALLY INITIALIZES & CONFIGURES THE HARDWARE BY...

A	Reading the config.YML 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 deployer node
D	Deploying needed firmware and operating system images to the server nodes
E	Configuring the network switches (VLANs, MLAG) – <i>some manual steps required</i>
F	Configuring all server management and data nodes (network interfaces, GPU drivers,...)

At genesis completion, control of the cluster is transferred to the operational manager (or the software orchestrator) of choice, if supported. Throughout the cluster life cycle, persistent ansible playbooks provide runtime services to the operational manager.

Go [here](#) to see the genesis deployment README to learn more.

Go [here](#) to see the Accelerated Database README for procedure overview and deployment automation procedures. Deployment procedures include clone repository, update config.yml and two automation scripts (*install.sh* and *deploy.sh*.)

OBTAIN THE DEFAULT CONFIGURATION FILE

The genesis automation uses a config file to specify the target cluster configuration. The deployment tooling 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 deployment process.

Go [here](#) to see the **generic** master copy of the latest version of the config file.

Go [here](#) for a copy of the OpenPOWER Accelerated Database Small Cluster config file.

TAILOR THE CONFIGURATION FILE FOR YOUR ENVIRONMENT

The *config.yml* file contains a lot of configuration information. To enable a cluster tailored to your environment, edit the YAML file with the configuration parameters you collected in Step 2, replacing the **RED** text with your data. The excerpt below focuses on the lines to edit.

Editable Portions of the Config.yml file

```
~ ~ ~ ~ ~ bunch of licencing comment and YAML ~ ~ ~ ~ ~
ipaddr-mgmt-network: 192.168.3.0/24    ← Should this be a red field?
ipaddr-mgmt-switch:
  rack1: 192.168.3.5                  ← Type your management switch IP address here.
ipaddr-data-switch:
  rack1: 9.3.3.178                    ← Type your data switch IP address here.
~ ~ ~ ~ ~ bunch of YAML and comments ~ ~ ~ ~ ~
networks:
  external:
    description: Organization site or external network
    addr: 9.3.3.0/24                  ← Type your XXX here.
    broadcast: 9.3.3.255              ← Type your XXX here.
    gateway: 9.3.3.1                  ← Type your XXX here.
    dns-nameservers: 9.3.1.200        ← Type your XXX here.
    dns-search: aus.stglabs.ibm.com
    method: static
    eth-port: eth10
  interconnect:
    description: Private 10G Data Network to Interconnect Cluster
    addr: 10.0.0.0/24
    broadcast: 10.0.0.255
    method: static
```

```

    eth-port: eth11
    ~ ~ ~ ~ ~ bunch of YAML and comments ~ ~ ~ ~ ~
node-templates:
  controller1:
    hostname: min                ← Type your XXX here.
    userid-ipmi: ADMIN          ← Type your XXX here.
    password-ipmi: admin        ← Type your XXX here.
    cobbler-profile: ubuntu-16.04.1-server-ppc64el
    ~ ~ ~ ~ ~ bunch of YAML and comments ~ ~ ~ ~ ~

```

THE INVENTORY FILE

The Inventory file is a YAML text file that contains the entire inventory of the cluster taken during the Genesis process.

Go [here](#) for the **generic** master copy of the latest inventory file.

The file consists of the network switches and server nodes. The Switches data structure indicates the types of switches (Management, Spine, or Leaf), their IP addresses, and associated login credentials. The following is a sample inventory data structure specifying the attributes of the management and leaf switches.

```

switches:
  mgmt:
    - hostname: mgmtswitch1
      ipv4-addr: 192.168.3.5
      rack-id: rack1
      userid: admin                ← Type your XXX here????
      password: mspassword        ← Type your XXX here????
  leaf:
    - hostname: leafswitch1
      ipv4-addr: 192.168.3.6
      rack-id: rack1
      userid: joeleaf
      password: joeleafpassword

```

The server Nodes data structure specifies the type of node controller, its network properties, and its system architectural information (ppc64 vs x86). A snippet of such an inventory data structure follows below.

Controller1:

```
- hostname: min-1
  userid-ipmi: ADMIN
  password-ipmi: admin
  port-ipmi: 15
  port-pxe: 16
  port-eth10: 21
  port-eth11: 22
  mac-ipmi: 70:e2:84:14:0a:10
  ipv4-ipmi: 192.168.3.107
  rack-id: rack1
  template: controller2
  architecture: ppc64
  chassis-part-number: 8335-GTB
  chassis-serial-number: 1004C9A
  mac-pxe: 70:e2:84:14:0a:12
  ipv4-pxe: 192.168.3.108
  external-addr: 9.3.3.5
  interconnect-addr: 10.0.0.4
reference-architecture:
  gpudb_nvidia_playbook:
    description: playbook for installing nvidia for gpudb
    cuda_deb: /tmp/cuda-repo-ubuntu1604-8-0-local_8.0.35-1_ppc64el.deb
    driver_level: nvidia-361
    dkms_deb: /tmp/dkms_2.2.0.3-2ubuntu14_all.deb
```

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

STEP 6: COMPLETE ANY POST GENESIS CONFIGURATIONS

For this solution, there are a couple of procedures that need to be completed post-Genesis. These have been automated within deploy.sh referenced in Accelerated Database README and described in STEP 5.

COMPLETE NETWORKING CONFIG

Performed during the deploy.sh.

Each node will have a 10G network interface for cluster interconnect and external connection.

INSTALL NVIDIA DRIVER

Performed during the `deploy.sh`.

For GPUdb only the Nvidia driver is installed, but full CUDA is available locally on the deployment node in the `~/accelerated-db/playbooks/packages` directory.

STEP 7: INSTALL OPEN SOURCE MANAGEMENT SOFTWARE

This deployment kit does not yet auto-deploy open source management tooling for this cluster. *A future version of this kit will add this feature.* Installing your own management tooling will need to be performed independent of this deployment kit.

STEP 8: INSTALL THE APPLICATION SOFTWARE

Since today's GPU-accelerated Database tools are proprietary today, this deployment kit does not automatically deploy them. *We will add more pointers to those solutions as they support this configuration.*

GPUDB

Kinetica, Inc. is the commercial provider of GPUDB. Go to their site or contact them to get help deploying their database onto this cluster.

Go [here](#) to Kinetica's website.

Go [here](#) to see Kinetica's overview and architecture.

Go [here](#) to see Kinetica's installation instructions.

Go [here](#) to see how to run Kinetica once installed.

APPENDIX

REFERENCE LINKS

Go [here](#) for 8335-GTB system Redbook(technical overview)

Go [here](#) for further 8335-GTB reference material via IBM Knowledge Center.

Go [here](#) for Accelerated-db GitHub repository(currently IBM internal only)

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>.