# 

# **IBM Cluster Recipe: Accelerated Database**

Version

12/06/16

Document owner:

Jorge Yanez

[jorgeyan@us.ibm.com](mailto:jorgeyan@us.ibm.com)

Content Contributors:

Anh Dang

Irving Baysah

Jorge Yanez

Contributors and Reviewers:

Jonathan Dement

Raymond Harrington

Ted Maeurer

[1 Introduction 3](#_Toc468826527)

[2 Pre-Requisites 4](#_Toc468826528)

[2.1 Cluster Genesis High Level Architecture 4](#_Toc468826529)

[2.2 Deployer Node: Server/Laptop 4](#_Toc468826530)

[2.3 Pre-build Customizations 5](#_Toc468826531)

[3 Physical Build Instructions 6](#_Toc468826532)

[3.1 Building Blocks – BOM 6](#_Toc468826533)

[3.1.1 IBM OpenPower 8335-GTB (Minsky) 6](#_Toc468826534)

[3.1.1.1 Node Build Instruction – Slot Placement 6](#_Toc468826535)

[3.2 Racking Rules 7](#_Toc468826536)

[3.2.1 Racking Recommendations and Restrictions 8](#_Toc468826537)

[3.2.1.1 Racking Example for a Four 8335-GTB System Cluster 10](#_Toc468826538)

[3.3 Physical Networking Plug Rule and Options 11](#_Toc468826539)

[3.3.1 Four 8335-GTB System Cluster Illustration – block diagram of the rear of the system network ports 11](#_Toc468826540)

[3.3.2 Physical Network Plugging Example 12](#_Toc468826541)

[4 Hardware Cluster Genesis 13](#_Toc468826542)

[4.1.1 Deployment README 13](#_Toc468826543)

[4.1.2 The Configuration File 13](#_Toc468826544)

[4.1.3 The Inventory File 15](#_Toc468826545)

[5 Operational Manager and Software Orchestrator 16](#_Toc468826546)

[6 Solution Specific Applications and Frameworks 17](#_Toc468826547)

[6.1 Post Genesis – Networking 17](#_Toc468826548)

[6.2 Post Genesis – Install Nvidia driver 17](#_Toc468826549)

[6.3 GPUdb 17](#_Toc468826550)

[6.3.1 Install Software Application 17](#_Toc468826551)

[6.3.2 Running Solution Application 17](#_Toc468826552)

[7 Notices 18](#_Toc468826553)

# Introduction

This document along with the Accelerated DB BOM.xlsxdescribes the fully working set of instructions, rules, and automations for building Accerated Database solutions on OpenPower systems. This document is designed to facilitate Kinetica’s GPUdb

Chapter 2 describes the cluster build overall flow. Chapter 3 specifies a generic hardware build instruction including: BOM contents for each building block, build rule (racking and network options). Chapter 4 described the bare metal machine genesis processes used in Accelerated Database solutions. Chapter 6 is Accerated Database operational manager or software orchestrator(not supported in first release). Finally, Chapter 6 is for Kinetica’s GPUdb specific instructions and documentation.

# Pre-Requisites

## Cluster Genesis High Level Architecture

1. Aquire all the hardware.
2. Rack and Cable your hardware (new switches require some manual config). – 3
3. Setup the **deployer node.** – 4.1.3
4. Fill out (edit) the **config.yml** file. – 4.1.2
5. Apply power to nodes to bring up BMCs which will fetch an IP address from the deployer via DHCP.
6. Deployer node reads MTM, serial number via IPMI and node MAC addresses from switches. Info is written to inventory file (also YAML format).
7. Via IPMI, program the nodes to PXE load the appropriate OS image.
8. Configure switches (vlans created, MLAG setup)
9. Post configure nodes
   * Configure network interfaces
   * Install Nvidia driver for P100 GPU module
   * Install GPUdb
10. Push **inventory.yml** to /var/oprc on all cluster nodes?

## Deployer Node: Server/Laptop

Any Power8-LC or x86 Server with min 2 core 32G of memory

Requires 3 Network Interface connections: IPMI, 10G (highspeed), 1G(Mgmt). This node can be installed as temporary or permanent. Desired deployer node, whether Power8-LC or x86, should be installed with Ubunutu 16.04 LTS before beginning with deployment.

For Power8-LC:

<https://www.ubuntu.com/download/server/power8>

For x86:

<http://releases.ubuntu.com/16.04.1/>

## Pre-build Customizations

IBM manufacturing preloads the software on the server(s) prior to shipping. To facilitate the preloaded software's integration into the existing environment, additional information is required about the cluster environment.

The following information needs to be collected in order to provision the cluster before the software stack can be loaded. This information will be used as input to the toolkit **“config.yml”** file. Each solution will have its own config file as needed.

Please see the **Readme** –file included in the toolkits for more details and options….

The following information are expected to have available when update the **“config.yml”** file

|  |  |  |
| --- | --- | --- |
| **Item** | **Description** | **Remarks** |
| 1 | **Domain Name** | For example: ibm.com |
| 2 | **Upstream DNS server(s)**  - A DNS server is configured inside the cluster. For those names that could *not* be resolved by themselves, a set of upstream DNS servers are needed in the customer's network. | For example:  \* 4.4.4.4, 8.8.4.4 as the default public upstream DNS servers. |
| 3 | **Host name for the Deployment Node** | For example, depnode |
| 4 | **Management network IP address for the deployment node** | For example, 192.168.3.3/24 |
| 5 | **Data network IP address for the deployment node** | For example, 10.0.0.1/24 |
| 6 | **Mgmt IPaddr of 1G Network switch** | For example, 192.168.3.5 |
| 7 | **Mgmt IPaddr of High Speed Data Network switch** | For example, 9.3.3.178 |
| 8 | **Default Logins/Passwords** | BMC network, OS Mgmt network |
| 9 | **Host name and IP addresses for servers other than the Deployment node** | For example:   |  |  |  | | --- | --- | --- | | **Name** | **IP in management network** | **IP in data network** | | 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 | |

# Physical Build Instructions

## Building Blocks – BOM

It is suggested to build the recipe based on the existing supported build block as specified in the ***<Accelerated DB BOM 102716.xlsx>*** located at:

[*https://github.ibm.com/open-power-solution-genesis/accelerated-db/tree/initial\_dev/docs*](https://github.ibm.com/open-power-solution-genesis/accelerated-db/tree/initial_dev/docs)

### IBM OpenPower 8335-GTB (Minsky)

*<* ***Accelerated DB BOM 102716.xlsx > >>>>Minksy 8335-GTB Tab***

#### Node Build Instruction – Slot Placement





\* Port sharing capable – not currently supported in Genesis

\*\* Need two dual port adapters if bonded networks desired.

## Racking Rules

This toolkit suggesting the following Racking rule which focused on Modulation per Rack. Multiple-racks configuration will be a replication at per rack level. This rule provides:

* Consistency
* Expandability
* Common senses to optimized racking for : servicing, re-use , shipping, cooling

*<* ***Accelerated DB BOM 102716.xlsx > >>>>Racking Rule Tab***

### Racking Recommendations and Restrictions

### 

|  |
| --- |
| **Common Rules** |
| 1) Bottom 1U of rack must always be open (FC ER1B).  2) Top 1U of rack must always be open (FC ER1T).  3) All configurations require the following placement rules:   1. (10G/40G/IB) 8828-E36/E36 / 8831-NF2 / 7120-64C placement in 26U 2. U25 un-used- (can be used if need for short depth devices .see switch placement restriction above)   c) (1G Mgtm) 7120-48E placement in 24U  4) Minimum qty = 1 of Servers per rack; if a client wants an empty rack, it can be ordered separately and not be part of the HPC Cluster configuration  5) No more racks may be ordered than what is needed to support the servers that are ordered  6) If more than 4 PDUs are needed, the 2 additional Horizontal PDUs placement in 40U and 41U. Spine Switches take priority over additional PDUs, and if 40U and 41U are occupied by Spine Switches, additional Horizontal PDUs placement in next available slots  **Recomendations**   * Modular Rack design at the **Per Rack** Level   + Each Rack Contained Inter-rack network switches (Leaf switches) (place in between U24-U26)   + Intra-rack connectivity is done via Spine Switch configuration (place spine switches in U37-U41)   Notes: You can design the Modular Rack block as (every 2 racks if required) –recommending see ***48x High Speed TOR Switch Option***. |
|  |

#### Racking Example for a Four 8335-GTB System Cluster



\*No external storage support for first release, only local disk storage available.

## Physical Networking Plug Rule and Options

### Four 8335-GTB System Cluster Illustration – block diagram of the rear of the system network ports



### Physical Network Plugging Example

Names on the network switch ports correspond to the four 8335-GTB cluster illustration in 3.3.1.

Example of the physical plugging order of a four 8335-GTB cluster in 3.3.1



Magnified view of the physical ports on the IBM Rackswitch G8264CS 10GbE data network switch for a four 8335-GTB system cluster(3.3.1).





Magnified view of the physical ports on the IBM Rackswitch G8052 1GbE management network switch for a four 8335-GTB system cluster(3.3.1).



# Hardware Cluster Genesis

The hardware genesis process involves the power on, initialization, configuration and installation of a cluster solution. It provides an automated approach to building a cluster solution from initial assembly to an operational state. The genesis process occurs once at the beginning of the cluster solution lifecycle. However, automation open source scripts are accessible and can be reused for the maintenance and expansion of the cluster solution.

* Automation of hardware initialization and configuration.
  + Hardware detection
  + Power control
  + Firmware install
  + OS install
  + Network configuration of node and switches
  + Create/offer inventory

After the Genesis process completes, control of the cluster is transferred to the operational manager (or software orchestrator) of choice. Throughout the cluster life cycle, persistent ansible playbooks provide runtime services to the operational manager.

### Deployment README

[*https://github.ibm.com/open-power-solution-genesis/accelerated-db/blob/initial\_dev/README.md*](https://github.ibm.com/open-power-solution-genesis/accelerated-db/blob/initial_dev/README.md)

### The Configuration File

The Config file is a YAML text file used 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. The **generic** master copy of the latest version of the config file is located at:

[*https://github.com/open-power-ref-design/cluster-genesis/blob/master/config.yml*](https://github.com/open-power-ref-design/cluster-genesis/blob/master/config.yml)

Working config.yml for the four 8335-GTB cluster illustration in 3.3.1 is located at:

[*https://github.ibm.com/open-power-solution-genesis/accelerated-db/blob/initial\_dev/config.yml.p8dep.4min*](https://github.ibm.com/open-power-solution-genesis/accelerated-db/blob/initial_dev/config.yml.p8dep.4min)

\*\*\* User would have to change IP addresses in **RED** to their desired network subnets.

*---*

# Copyright 2016 IBM Corp.

#

# All Rights Reserved.

#

# Licensed under the Apache License, Version 2.0 (the "License");

# you may not use this file except in compliance with the License.

# You may obtain a copy of the License at

#

# <http://www.apache.org/licenses/LICENSE-2.0>

#

# Unless required by applicable law or agreed to in writing, software

# distributed under the License is distributed on an "AS IS" BASIS,

# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.

# See the License for the specific language governing permissions and

# limitations under the License.

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"

ipaddr-mgmt-network: 192.168.3.0/24

ipaddr-mgmt-switch:

rack1: 192.168.3.5

ipaddr-data-switch:

rack1: 9.3.3.178

redundant-network: false

userid-default: ubuntu

password-default: passw0rd

userid-mgmt-switch: admin

password-mgmt-switch: admin

userid-data-switch: admin

password-data-switch: admin

networks:

external:

description: Organization site or external network

addr: 9.3.3.0/24

broadcast: 9.3.3.255

gateway: 9.3.3.1

dns-nameservers: 9.3.1.200

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

node-templates:

controller1:

hostname: min

userid-ipmi: ADMIN

password-ipmi: admin

cobbler-profile: ubuntu-16.04.1-server-ppc64el

os-disk:

- /dev/sdj

- /dev/sdi

name-interfaces:

mac-pxe: eth15

mac-eth10: eth10

mac-eth11: eth11

ports:

ipmi:

rack1:

- 5

- 7

- 9

- 11

pxe:

rack1:

- 6

- 8

- 10

- 12

eth10:

rack1:

- 1

- 3

- 5

- 7

eth11:

rack1:

- 2

- 4

- 6

- 8

networks:

- external

- interconnect

### The Inventory File

The Inventory file is a YAML text file that contains the entire inventory of the cluster taken during the Genesis process. The generic master copy of the latest inventory file is located at

[*https://github.com/open-power-ref-design/cluster-genesis/blob/master/master\_inventory.yml*](https://github.com/open-power-ref-design/cluster-genesis/blob/master/master_inventory.yml)

The file consists of the Switches and the Nodes. The Switches data structure indicates the types of switches (Management, Spine, or Leaf) and their associated login credentials and IP addresses. 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*

*password: yggdrasil*

*leaf:*

*- hostname: leafswitch1*

*ipv4-addr: 192.168.3.6*

*rack-id: rack1*

*userid: joeleaf*

*password: joeleafpassword*

The Nodes data structure indicates the type of node controller and its associated network communication properties. The file also provide the system architectural information (ppc64 vs x86). The following is a snippet of inventory data structure specifying the attributes of one of the controllers.

*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*

# Operational Manager and Software Orchestrator

Not supported on this release

# Solution Specific Applications and Frameworks

## Post Genesis – Networking

Performed during the deploy.sh

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

## Post Genesis – Install Nvidia driver

Performed during the deploy.sh

For GPUdb only the driver is installed, but full CUDA is available local within the cluster.

## GPUdb

[*http://www.kinetica.com/*](http://www.kinetica.com/)

[*http://www.kinetica.com/docs/overview/arch.html*](http://www.kinetica.com/docs/overview/arch.html)

### Install Software Application

No Ansible playbook yet, for Install instructions please see:

[*http://www.kinetica.com/docs/install/index.html#*](http://www.kinetica.com/docs/install/index.html)

### Running Solution Application

o Ansible playbook yet, for execution instructions please see:

[*http://www.kinetica.com/docs/gpudbAdmin/startgpudb.html*](http://www.kinetica.com/docs/gpudbAdmin/startgpudb.html)

# 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](http://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>.