

NVIDIA DGX SuperPOD

Deployment Guide

Featuring NVIDIA DGX A100 and DGX H100 Systems

DG-11251-001 V15 2024-09-26 BCM 10.24.09

Contents

1.	Introduction	2
2.	Initial Point-to-Point Preparations	3
3.	Initial Cluster Setup	4
4.	Head Node Configuration	.25
5.	High Availability	.29
6.	Slurm Setup	.45
7.	Upgrade Infiniband Switches	.49
8.	SuperPOD Validation	.51

1. Introduction

This document details deploying NVIDIA Base Command[™] Manager (BCM) on NVIDIA DGX SuperPOD[™] configurations.

Physical installation and network switch configuration must be completed before deploying BCM. In addition, information about the intended deployment should be recorded in a site survey.

The DGX SuperPOD Deployment Guides contain essential information and should be consulted before deploying the configuration. The DGX H100 version of the deployment guide is available here: <u>https://docs.nvidia.com/dgx-superpod/index.html</u>

Note: NVIDIA DGX SuperPOD only supports Slurm and Kubernetes with RunAI.

2. Initial Point-to-Point Preparations

Follow the instructions provided to automatically assign MAC addresses from the given switch and switch ports. If you prefer to utilize the old method of mapping MAC addresses to IP addresses, you can skip this section.

When using the p2p-superpod-template, navigate to the first tab labeled "Notes," where you'll find four buttons.

- 1. "Create Table of Contents": This button generates a new tab called "Contents," listing all tabs with links (this option is for generating purposes only).
- "Generate P2P": This button creates a new tab called "p2p_ethernet" automatically populating content from existing tabs such as OOB, MGMT-InBand, and DGX-InBand into a single sheet.
- 3. "**Search & Replace**": This function utilizes the tab labeled "Alias" to search for text in the "p2p_ethernet" tab's Column C and replace it with the corresponding text from Column A.
- 4. **"Populate**": Automatically populate the FLOW and ALOCATE_IP columns based on data from the remaining columns.

Below button will ge	nerate		Below button will	search and replace	Below button v	vill generate
P2P_ETHERNET (new t	ab)					
t will combine data	from column [I	-,H,J,K]	data from the P2	P_ETHERNET	Additional data	fillup
ООВ			based on Alias ta	ab	data in the colu	umn FLOW and ALLOCATE_IP
GX-InBand			Seach column [0			
/IGMT-InBand			Replace value co	lumn [A]		
Generate	P2P		Search &	Replace		populate

Next, navigate to the **p2p_ethernet** tab. From there, access the *File* menu and choose the "*Downloads*" option, selecting the (.csv) format. Save the file and transfer it to a USB stick, along with Cumulus OS and IBSW OS.

3. Initial Cluster Setup

The deployment stage of a DGX SuperPOD consists of using BCM to provision and manage the Slurm cluster.

1. Configure the NFS server.

User home directories (home/) and shared data (cm_shared/) directories must be shared between head nodes (such as the DGX OS image) and must be stored on an NFS filesystem for HA availability. Because DGX SuperPOD does not mandate the nature of the NFS storage, the configuration is outside the scope of this document. This DGX SuperPOD deployment uses the NFS V3 export path provided in the <u>site</u> <u>survey</u>: /var/nfs/general.

The following parameters are recommended for the NFS server export file /etc/exports.

/var/nfs/general *(rw,sync,no_root_squash,no_subtree_check)

- 2. Configure the DGX systems to PXE boot by default.
 - a. Using either KVM or a crash cart, connect to the DGX system, enter the BIOS menu, and configure Boot Option #1 to be [NETWORK].

Aptio Setup Utili Main Advanced Chip	<mark>ty – Copyright (C) 2021 Amer</mark> set Security <mark>Boot </mark> Save &	r <mark>ican Megatrends, Inc.</mark> Exit Server Mgmt
Boot Configuration Setup Prompt Timeout Bootup NumLock State	1 [0n]	Sets the system boot order
FIXED BOOT ORDER Prio Boot Option #1 Boot Option #2 Boot Option #3 Poot Option #4	rities [Network] [Disabled] [Disabled]	
Boot Option #4 Boot Option #5 Boot Option #6	[Disabled] [Disabled] [Disabled]	↔: Select Screen ↑↓: Select Item

b. Ensure that other Boot Options are [Disabled] and go to the next screen.

c. Set Boot Option #1 and Boot Option #2 to use IPv4 for Storage 4-2 and Storage 5-2.

Aptio Setup Util.	ity – Copyright (C) 2021 Ar <mark>Boot</mark>	merican Megatrends, Inc.
Boot Option #1	[UEFI: PXE IPv4	Sets the system boot
	Adapter – OC:42:A1:OA:33:4B]	
BOOT UPTION #2	LUEFI: PXE IPV4 Mellanox Network	
	Adapter – OC:42:A1:74:F3:1F]	
Boot Option #3	[Disable]	
Boot Option #4	[Disable]	
Boot Option #5	[Disable]	++: Select Screen
Boot Option #6	[Disable]	↑↓: Select Item
Boot Option #7	[Disable]	Enter: Select
Boot Option #8	[Disable]	+/-: Change Opt.
Boot Option #9	[Disable]	F1: General Help
Boot Option #10	[Disable]	F2: Previous Values
Boot Option #11	[Disable]	F3: Optimized Defaults
Boot Option #12	[Disable]	F4: Save & Exit
and the second		ESC: Exit
Version 2.20.12	75. Copyright (C) 2021 Amer	rican Megatrends, Inc.
		AB

- d. Ensure that other Boot Options are [Disabled].
- e. Select Save & Exit.
- 3. On the failover head node and the CPU nodes, ensure that Network boot is configured as the primary option. Ensure that the Mellanox ports connected to the network on the head and CPU nodes are also set to Ethernet mode. This is an example of a system that will boot from the network with Slot 1 Port 2

and Slot 2 Port 2.



- 4. Download the BCM installer ISO.
- Burn the ISO to a DVD or to a bootable USB device.
 It can also be mounted as virtual media and installed using the BMC. The specific mechanism for the latter will vary by vendor.
- 6. Ensure that the BIOS of the target head node is configured in UEFI mode and that its boot order is configured to boot the media containing the BCM installer image.
- 7. Boot the installation media.



- 8. At the grub menu, choose Start Base Command Manager Graphical Installer.
- 9. Select Start installation on the splash screen.



10. Accept the terms of the NVIDIA EULA by checking I agree and then select Next.

Les through the	DISODICA /	
Ø 🔤 💿	10.180.217.154	
NIDIA.	Base Command Manager installer	v10.23.11 (UBUNTU2204)
NVDIA EULA	NVIDIA	
Kernel modules	NUTDTA AT PRODUCT ACREENENT	A
Hardware info	IMPORTANT NOTICE - PLEASE READ AND AGREE BEFORE USING NVIDIA AI PRODUCTS.	
Installation source	This AI Product Agreement is entered into between the entity you represent or	
Cluster settings	you individually if you do not designate an entity ("Customer") and NVIDIA Corporation ("NVIDIA"). This AI Product Agreement consists of the terms and	
Workload manager	conditions below and all documents attached to or referenced in this AI Product Agreement (together, the "Agreement"). The AI Product catalogs include products	
Network topology	that can be used without payment and paid products and services. Key terms are defined in Section 17.	
Head node	By using or registering to use AI Products, Customer is affirming that Customer	
Compute nodes	has read the Agreement and agrees to its terms. If Customer does not have the required authority to enter into the Agreement or if Customer does not accept	
BMC configuration	all the terms and conditions below, do not use (or register to use) AI Products.	
Networks	1. AI PRODUCTS OFFERINGS.	
Head node interfaces	1.1 Grant.	
Compute nodes interfaces	Subject to the terms of the Agreement, Customer's Order Agreement and Subscription or Perpetual license parameters, and payment of applicable fees,	
Disk layout	NVIDIA grants Customer a non-exclusive, non-transferable, non-sublicensable (except as expressly provided in the Agreement) license to do the following for	
Disk layout settings	the duration of the license:	
Additional software	1.1.1 install and use copies of AI Products,	
Summary	 1.1.2 create Derivative Samples and Derivative Models to develop and test services and applications, 	
Deployment	 1.1.3 configure the AI Product using the configuration files provided (as applicable), 	
	I agree	
	Continue remotely Load config	Show config Back Next
1024 x 769		

11. Accept the terms of the Ubuntu Server UELA by checking I agree and then select

Next.					
Ø 🔤 💿	10.180.217.154				
🧆 NVIDIA.	Base Command Manager installer v10.23.11 (UBUNTU22				
NVDIA EULA	Ubuntu Server 22.04				
Kernel modules					
Hardware info	Ubuntu is a collection of thousands of computer programs and documents created by a range of individuals, teams and companies.				
Installation source	Fach of these programs may come under a different licence. This licence policy describes the process that				
Cluster settings	we follow in determining which software will be included by default in the Ubuntu operating system.				
Workload manager	Copyright licensing and trademarks are two different areas of law, and we consider them separately in Ubuntu. The following policy applies only to copyright licences. We evaluate trademarks on a case-by-case				
Network topology	basis.				
Head node	Categories of software in Ubuntu The thousands of software packages available for Ubuntu are organised into four key groups or components:				
Compute nodes	main, restricted, universe and multiverse. Software is published in one of these components based on whether or not it meets our fee software philosophy, and the level of support we can provide for it. To addition.				
BMC configuration	software may be published for Ubuntu as a universal Linux snap package, in which case licenses are determined by the snap publisher and documented in the snap store.				
Networks	This policy only addresses the software that you will find in main and restricted, which contain software				
Head node interfaces	that is fully supported by the Ubuntu team and must comply with this policy.				
Compute nodes interfaces	Ubuntu 'main' component licence policy All application software included in the Ubuntu main component:				
Disk layout	Must include source code. The main component has a strict and non-negotiable requirement that application				
Disk layout settings	software included in it must come with full source code. Nust allow modification and distribution of modified copies under the same licence. Just having the source				
Additional software	code does not convey the same freedom as having the right to change it. Without the ability to modify software, the Ubuntu community cannot support software, fix bugs, translate it, or improve it.				
Summary	Ubuntu 'main' and 'restricted' component licence policy				
Deployment	All application software in both main and restricted must meet the following requirements: Must allow redistribution. Your right to sell or give away the software alone, or as part of an aggregate software distribution, is important because:				
	📄 I agree				
	Continue remotely Load config Show config Back Ne				
1024 x 768	[ති] A SSL 😪 🔿				

12. Unless instructed otherwise, select Next without modifying the kernel modules to be loaded at boot time.

Ø 🔤 O	10.180.217.154	usted Platform Modu		_⊌ л ×
	Base Command Manager installe		v10.23.11 (UBU	NTU2204)
NVDIA EULA	Kernel modules			Î
Kernel modules	In order to be able to use all the hardware, it is important	that the correct set of ker	nel modules are loaded	at boot-
Hardware info	time. The hardware in this machine has been probed and circumstances it is not necessary to modify the kernel mo	the kernel modules listed odules selection, but if you	below were loaded. Une wish to do so, you may	der most add or
Installation source	remove kernel modules here.			
Cluster settings				€ ⊕
Workload manager	Name	Parameters	Path	
Network topology	acpi_ipmi	-	-	\oslash
	acpi_power_meter	-	-	\oslash
nead node	acpi_tad	-	-	\oslash
Compute nodes	aesni_intel	-	-	
BMC configuration	autofs4		-	\oslash
Networks	bnxt_en	-		
Head node interfaces	cec			\oslash
Compute nodes interfaces	coretemp	-	-	
Disk layout	crc32_pclmul			\oslash
Disk layout settings	crct10dif_pclmul	-	-	
Additional software	cryptd	-	-	\oslash
Summary	crypto_simd	-	-	
Deployment	dca	-	-	\oslash
	drm	-	-	
	drm kms helper	-	-	0 -
		Continue remotely	Show config Bac	k Next

13. Verify that the Hardware info is correct and then select Next.

Ø 📰 0		10	.180.217.154	unanan hada 👘 🖬 🗖 🖉	\times
唑 NVIDIA.		Base Comma	and Manager installer	v10.23.11 (UBUNTU22))4)
NVDIA EULA	На	rdware info			Î
Kernel modules	The	following hardware has been	detected. If not all hardware i	has been recognized, you may go back to the kernel	
Hardware info	mod	ules configuration screen to	load extra kernel modules.		
Installation source	÷	Туре	Device	Model	
Cluster settings	Ð	Cd Rom			
Workload manager	÷	Keyboard	/dev/input/event2	HP Virtual Keyboard	
Network topology	Ð	Mouse	/dev/input/mice	HP Virtual Keyboard	
Head node			ens10f0np0	Network interface [bnxt_en]	
Compute nodes			ens10f1np1	Network interface [bnxt_en]	
RMC configuration	Ð	Network Interfaces	ens1np0	Network interface [mlx5_core]	
			ensznpo	Network interface	
Networks			ibs3	[mlx5_core[ib_ipoib]]	
Head node interfaces			/dev/sdb	30GB DataTraveler 3.0	
Compute nodes interfaces			/dev/sdb	30GB	
Disk layout		Storage	/dev/sda	959GB MR416i-a Gen10+	
Disk layout settings	0		/dev/nvme2n1	MZXLR6T4HALA-000H3	
Additional software			/dev/nvme1n1	MZXLR6T4HALA-000H3	
Summary			/dev/nvme0n1	MZXLR6T4HALA-000H3	
- · · ·					
Deployment	Ð	Storage Controllers		-	
			-	-	
400/ 7/0			60		
1024 x 708				10 🗄 SSL 🥑 🔾	

For example, that the target storage device and the cabled host network interfaces are present (in this case three NVMe drives are the target storage device, and ens1np0 and ens2np01 are the cabled host network interfaces).

14. On the Installation source screen, choose the appropriate source and then select Next. Running a media integrity check is optional.

	Base Command Manager installer		v10.23.11 (UBUNTU2204)	
NVDIA EULA	DVD/ISO/USB			-
Kernel modules	Salact from the list helpur			
Hardware info	 /dev/sdb (BCM Install Media) 			
Installation source	Run media integrity check			
Cluster settings				
Workload manager				
Network topology				
Head node				
Compute nodes				
BMC configuration	•			
Networks				
Head node interfaces				
Compute nodes interfaces				
Disk layout				
Disk layout settings				
Additional software				
Summary				
Deployment				
		Continue remotely	Show config Back Next	
				-

15. On the Cluster settings screen, enter the required information and then select Next.

🕸 📖 💿	10.180.217.154				
<mark>></mark> NVIDIA.	Base Command Manager installer		v10.23.11	(UBUNT	U22
IVDIA EULA	General cluster settings				
Cernel modules	Cluster name:				
lardware info	BCM 10.0 Cluster				
nstallation source	Organization name:				
luster settings	NVIDIA				
vorkload manager Ietwork topology	Administrator email:				
lead node					
ompute nodes	Send email to the administrator on first boot				
MC configuration	Time zone:				
letworks	(GMT-08:00) America/Los_Angeles				+
ead node interfaces	Time servers:				
ompute nodes interfaces	× 0.pool.ntp.org × 1.pool.ntp.org			>	× •
isk layout	Nameservers:				
isk layout settings	× 8888				
dditional software	Leave this field emoty if you intend to use DHCP for external network				
ummary	Search domains:				
eployment					
	Leave this field empty if you intend to use DHCP for external network				
	Conti	nue remotely	Show config	Back	N

16. On the Workload manager screen, choose None and then select Next.



17. On the Network topology screen, choose the network type for the data center environment and then select Next.



18. On the Head node screen, enter the Hostname, Administrator password, choose Other for Hardware manufacturer, and then select Next

Ø 0	10.180.217.154		
	Base Command Manager installer	v10.23.11 (UBUNTU2204)	
NVDIA EULA	Head node settings		
Kernel modules	Hostname:		
Hardware info	head-01		
Installation source	Administrator password:		
Cluster settings	•••••	Ø	
Workload manager	Confirm administrator password:		
Network topology		Ø	
Compute nodes	Hardware manufacturer	42	
BMC configuration	Other		
Networks			
Head node interfaces			
Compute nodes interfaces			
Disk layout			
Disk layout settings			
Additional software			
Summary			
Deployment		k	
	Continue remotel	ly Show config Back Next	
1024 x 768		🖄 🔒 SSL 🛛 🥪 🔿 🍅	

19. Accept defaults in the Compute nodes and then select Next.

\sim

	Base Command Manager installer v1	0.23.11 (UBUNTU2204)
NVDIA EULA	Compute nodes settings	
Kernel modules	Number of racks:	
Hardware info	1	
Installation source	Number of podes:	
Cluster settings		
Workload manager	1	
Network topology	Node start number:	
Head node	1	
Compute nodes	Node base name:	
BMC configuration	node	
Networks	Node digits:	
Head node interfaces	3	
Compute nodes interfaces		
Disk layout	Hardware manufacturer:	
Disk layout settings	Other	·
Additional software	15	
Summary		
Deployment		
	Continue remotely Show	config Back Next
1024 x 768	(ට 🔒 ss∟ 🕜 🔿 🆕 _

20. Ensure that the Node base name is node. Other values will be updated later in the installation.

21. On the BMC Configuration screen, choose No for both Head Node and Compute Nodes, and then select Next.

<mark>⊘</mark> nvidia.	Base Command Manager installer		v10.23.11 (UB	UNTU2204)
	MC configuration			
Kernel modules	ine configuration			
Hardware info	ead Node	Compute Node	S	
Installation source Will BM0	I head node have IPMI/iDRAC/iLO/CIMC compatible Cs?	Will compute nodes ha compatible BMCs?	ive IPMI/iDRAC/iLO/CI	MC
Cluster settings 🛛 🕞 Y	Yes No	○ Yes ● No		
orkload manager				
etwork topology				
ead node				
ompute nodes				
4C configuration				
tworks				
ad node interfaces				
npute nodes interfaces				
k layout				
k layout settings				
ditional software				
ımmary				
eployment				
		Continue remotely	Show config Ba	ack Next
1024 x 768			ି 🔒 SSL	0 🙂

These will be updated later in the post install stages.

22. On the Networks screen, enter the required information for internalnet, and then select Next.

	Base Command Manager installer	v10.23.11 (UBUNTU2204)
NVDIA EULA	Networks	
Kernel modules	The following IP networks have been pre-configured. Using the controls below,	, the network settings may be altered.
Hardware info	internaliset	A
Installation source	Name:	v
Cluster settings	internalnet	
Workload manager	Base IP address:	
Network topology	10.180.115.128	
Head node	Netmask	
Compute n <mark>o</mark> des	255 255 255 192//26)	× -
BMC configuration	255.255.255.152(26)	
Networks	Dynamic range start:	
Head node interfaces	10.180.115.168	
Compute nodes interfaces	Dynamic range end:	
Disk layout	10.180.115.183	
Disk layout settings	Domain name:	
Additional software	eth.cluster	
Summary	Gateway:	
Deployment	10.180.115.129	I
	By default the head node will be used as the default gateway.	-
	MTU: Continue remote	ly Show config Back Next
1024 x 768		 © _ A_SSL ⊘ () ()

Since a Type 2 network was specified, there are no other network tabs (for example, externalnet Or ipminet).

23. On the Head node interfaces screen, ensure that one interface is configured with the head node's target internalnet IP, and then select Next.

		Base Command Manage	r installer	v10.23	.11 (UBUNTU2204)
NVDIA EULA	Head nod	le network interf	aces		
Cernel modules					•
Hardware info	Interface	Network	IP addres	s	
nstallation source	ens1np0	× 👻 internalnet	× - \$10.180.	115.189	× - Ō
uster settings					
orkload manager					
letwork topology					
lead node					
ompute nodes					
MC configuration				k	
letworks				- 94	
ad node interfaces					
mpute nodes interfaces					
sk layout					
sk layout settings					
ditional software					
immary					
Deployment					
	_		Continue remot	ely Show confi	g Back Next
1024 x 768				റി	Assl ⊘ 🔿 🌰

Other interfaces will be configured by the post install script.

24. On the Compute node interfaces screen, leave the default entries, and then select Next.

	Ba	se Commar	nd Manager insta	aller	v10.	.23.11 (UBUNTU	J2204)
NVDIA EULA	Compute n	odes ne	twork inter	faces			
Kernel modules							۲
Hardware info	Interface		Network		IP offset		
Installation source	BOOTIF	× =	internalnet	\times $-$	0.0.0.0	× -	Ċ
Cluster settings	IP addresses for eac	of the interfa	aces on all nodes wil	l be assigned autom	atically from a cor	secutive range of	IP
Workload manager	addresses. The start	ing address in	the range is determ	nined by adding the	specified offset to	the base address o	of the
Network topology	to be assigned 10.14	1.0.1, and the	e second node 10.14	1.0.2.		o war cause ute fit:	at node
Head node							
Compute nodes							
BMC configuration							
Networks							
Head node interfaces							
)isk layout							
isk layout settings							
dditional software							
ummary							
eployment							
				Continue rem	notely Show c	onfig Back	Ne🏝
1024 x 768					<u>6</u>] 🔒 SSL 🥏	0 🈃

These will be updated post install.

25. On the Disk layout screen, select the target install location (in this case nvme@n1) and then select Next.



26. On the Disk layout settings screen, accept defaults and then select Next.



These settings will be updated later in the post installation steps.

27. In the Additional software screen, do not choose anything and then select Next.

	Base Command I	Manager Installer v10.23.11 (UBUNTU2	2204)
DIA EULA	Additional software		
nel modules	Please select from available so	ftware	
dware info			
tallation source	OFFD stack		
ster settings	OFED Stack		
rkload manager			
work topology			
ad node			
npute nodes			
C configuration			
works			
ad node interfaces	k		
npute nodes interfaces			
k layout			
k layout settings			
litional software			
nmary			
oloyment			
onfirm the in	formation on the s	© ≙ عد ⊗ ummary screen and then select Nex:	0 0
21x768 21x768 DIFITIAN DIA EULA mel modules refugare info	formation on the S Base Command I Summary Below is a brief summary of some of t	©	0 🍅 (t. 2204)
24 × 768 Difirm the in: DIA EULA rnel modules rdware info tallation source	formation on the S Base Command I Summary Below is a brief summary of some of the Primary internal interface IP: Primary internal interface IP:		0 🅲
24 × 768 Difirm the in NUDIA. DIA EULA role modules rdware info tallation source ster settings	formation on the S Base Command I Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone:	A sst Cummary Screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.8.8.8 Amorical as Angeles	(t . 2204)
24 × 768 Difirm the in NUDIA. DIA EULA ranel modules rdware info tallation source ster settings rkload manager	formation on the s Base Command I Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers:	A sst Cummary screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 3.8.8.8 America/Los_Angeles 0.ppol.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org	(† 1. 2204)
Ax768 Infirm the in NUDIA. NA EULA nel modules dware info callation source ster settings rkload manager work topology	formation on the s Base Command 1 Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Time servers: Time servers: Time servers: Head node hardware vendor	A sst Cummary screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.8.8.8 America/Los_Angeles 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org None other	(† 1. 2204)
A x 768 INVIDIA . INVIDIA . INA EULA Inal modules dware info allation source ter settings kload manager work topology d node	formation on the S Base Command I Summary Below is a brief summary of some of the Primary internal interface IP: Primary internal interface Netmask: Nameservers: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives	A sst Commany Screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.8.8.8 America/Los_Angeles 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org None Other Other /dev/nyme0n1	(t . 2204)
Ax768 Infirm the in NUDIA. NA EULA almodules dware info callation source ster settings rkload manager work topology ud node apute nodes	formation on the S Base Command I Summary Below is a brief summary of some of the Primary internal interface IP: Primary internal interface Netmask: Nameservers: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node BMC type Compute nodes BMC type	A sst Commany screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.8.8 America/Los_Angeles 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org None Other Other Other /dev/nvme0n1 No BMC No BMC	0 ()
Ax 768 Infirm the in Invibia. Inv	formation on the s Base Command I Summary Below is a brief summary of some of th Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node MC type Compute nodes BMC type		C 😃
Ax768 nfirm the in NUDIA. NA EULA nel modules dware info allation source ster settings kload manager work topology d node apute nodes c configuration works	formation on the s Base Command 1 Below is a brief summary of some of the Primary internal interface IP: Primary internal interface Netmask: Nameservers: Time servers: Time servers: Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node BMC type Compute nodes BMC type	A sst Cummary screen and then select Next Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 3.8.8.8 America/Los_Angeles 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org None Other Other Other /devinyme0n1 No BMC No BMC) 🔮
At x 768 TVIDIA. TVIDIA. DIA EULA nel modules rdware info tallation source ster settings rkload manager twork topology ad node mpute nodes C configuration tworks ad node interfaces	formation on the s Base Command 1 Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers: Workload manager Head node hardware vendor Install drives Head node BMC type Compute nodes BMC type	Cummary screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.5.5.5 America/Los_Angeles 0.pool.ntp.org, 1.pool.ntp.org None Other Other Other Other Mome Other No BMC No BMC) 🔮
24 × 768 Diffirm the information of the informatio	formation on the s Base Command I Summary Below is a brief summary of some of th Primary internal interface IP: Primary internal interface Netmask: Nameservers: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node BMC type Compute nodes BMC type	Cummary screen and then select Nex Cummary screen and then select Nex Manager installer V10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.8.8 America/Los_Angeles 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org None Other Other Other /dev/nvme0n1 No BMC No BMC No BMC) () () () () () () () () () () () () () () () (
AL & 768 Infirm the infirm Invibia. DIA EULA nel modules dware info callation source ster settings rkload manager work topology ad node npute nodes c configuration works ad node interfaces npute nodes interfaces k layout	formation on the s Base Command I Summary Below is a brief summary of some of th Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node MaC type Compute nodes BMC type	Cummary screen and then select Nex Cummary screen and then select Nex Commary screen) 🔮
24 × 768 Diffirm the in NUDIA. DIA EULA role modules rdware info tallation source aster settings rkload manager twork topology ad node mpute nodes C configuration tworks ad node interfaces mpute nodes interfaces k layout k layout settings	formation on the s Base Command I Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node BMC type Compute nodes BMC type	Cummary screen and then select Nex Commary screen) (b) (c) (c) (c) (c) (c) (c) (c) (c) (c) (c
24 × 768 Diffirm the in NVIDIA. DIA EULA The I modules rdware info tallation source ster settings rkload manager twork topology ad node mpute nodes tc configuration tworks ad node interfaces mpute nodes interfaces k layout k layout settings litional software	formation on the s Base Command 1 Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node BMC type Compute nodes BMC type	Cummary screen and then select Next Manager installer Anager installer (USUNTU2 Anager installer (USUNTU2 Anager installation settings that were selected: (USUNTU2 Anager installes (USUNTU2 (USUNTU2 Anager installes (USUNTU2 (USUNTU2 Anager installes (USUNTU2) () () () () () () () () () () () () () () (
24 × 768 Diffirm the in Diffirm the in Diff	formation on the s Base Command 1 Summary Below is a brief summary of some of t Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers Workload manager Head node hardware vendor Install drives Head node BMC type Compute nodes BMC type	Cummary screen and then select Nex Manager installer Culosed and a selected: Culosed and a s) () () () () () () () () () () () () () () (
24 × 768 Diffirm the in NUDIA. DIA EULA rinel modules rdware info tallation source ster settings rkload manager twork topology ad node mpute nodes (C configuration tworks ad node interfaces mpute nodes interfaces k layout k layout k layout settings ditional software mmary ployment	formation on the s Base Command I Summary Below is a brief summary of some of th Primary internal interface IP: Primary internal interface Netmask: Nameservers: Timezone: Time servers Workload manager Head node hardware vendor Compute nodes hardware vendor Install drives Head node hardware vendor Compute nodes BMC type Compute nodes BMC type	Cummary screen and then select Nex Manager installer v10.23.11 (UBUNTU2 the installation settings that were selected: 10.180.115.189 26 8.8.8 America.los_Angeles 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org None Other /dev/nvme0n1 No BMC No BMC	0 (

The Summary screen provides an opportunity to confirm the Head node and basic cluster configuration before deployment begins. This configuration will be updated/modified for DGX SuperPOD after deployment is complete. If any values do

not match expectations, use the Back button to navigate to the appropriate screen to correct any mistakes.

29. Once the deployment is complete, select Reboot.

NVDIA EULA Installation Kernel modules Overview of install Hardware info Parsing build Reboot now? Installation source Cluster settings Vorkload manager Network topology Vinstalling bear Network topology Head node Installing Board setting uburtu Server 22 04 base software images(s) Compute nodes Installing BCM packages to software images(s) Bik Configuration Vinstalling BCM packages to software images(s) Bik Configuration Vinstalling BCM packages to software images(s) Bik Loopt Vinstalling Installation Vinstalling Installation Vinstalling Installation Vinstalling Installation Compute nodes interfaces Summary Packages Additional software Show config Installation is complete		B	ase Command Mana	iger installer				
Kernel modules Overview of installs Overview of installs Overview of installs Hardware info Paring build Reboot now? Installation source Mouning CD Paritioning h Installing build No Person Cluster settings Installing head node BCM packages Compute nodes Installing bace distribution packages Compute nodes interfaces Cinstalling of installe in Stalling head node intelf RSF inges Retworks Creating node installer Stimg installation Initializing management daemon Compute nodes interfaces Initializing management daemon Compute nodes interfaces Compute node interfaces Initializing management daemon Disk layout Initializing management daemon Disk layout Initializing rebot after installation is complete Summary Initializing rebot after installation is complete 	NVDIA EULA	Installation	Installation	×				
Hardware info Parsing buil Additional source Parsing buil Parsing buil Parsing buil Parsing buil Parsing buil Parsing buil Parsing buil Parsing buil Parsing buil Pars	Kernel modules	Overview of installa	completed!					
Installation source Mounting CD Partitioning h Installing Ub Workload manager Installing head node BCM packages Configuring kemel and setting up bootloader Installing Boba node BCM packages to software image(s) Network topology Installing Boba node BCM packages to software image(s) Compute nodes Installing Offine selection of Python packages to software image(s) Installing Installer NFS image Installing installation Installing installation Installing management daemon Disk layout	Hardware info	✓ Parsing build	Report now?					
Cluster settings Partitioning h Installing hea No ves Installing lob No ves Installing hea Installing hea Configuring kernel and setting up bootloader Installing Ubuntu Server 22.04 base software image(s) Compute nodes Installing DBuntu Server 22.04 base software image(s) Compute nodes Installing CM packages to software image(s) BMC configuration Installing offline selection of Python packages Creating node installer NFS image Head node interfaces Finalizing installation Initializing management daemon Initi	Installation source	Mounting CD/	Reboot now:					
Workload manager installing obs installing hes Network topology installing hes installing dean dode BCM packages installing dean dode BCM packages installing dubuntu Server 22.04 base software image(s) Compute nodes installing Base distribution packages to software image(s) Read node interfaces installing BCM packages to software image(s) Installing BCM packages to software image(s) installing BCM packages to software image(s) installing BCM packages to s	Cluster settings	Partitioning h		_				
Network topology installing head node BCM packages Configuring kernel and setting up bootloader installing Ubuntu Server 22.04 base software image(s) Compute nodes installing BCM packages to software image(s) BMC configuration installing BCM packages to software images(s) installing BCM packages to software images(s) Networks Creating node installer NFS image Head node interfaces Finalizing installation Compute nodes interfaces Initializing management daemon Disk layout Initializing management daemon Disk layout attring office attring the poot after installation is complete Deployment 14/14 Automatically reboot after installation is complete Summary Back Reboot 	Workload manager	 Installing Ubu Installing heat 		No Yes				
Head node Configuring kernel and setting up bootloader Installing Ubuntu Server 22.04 base software image(s) Compute nodes Installing BCM packages to software image(s) Installing BCM packages to software image(s) Installing BCM packages to software image(s) Installing offine selection of Python packages Creating node installer NFS image Creating node installer NFS image Finalizing installation Initializing management daemon Disk layout Disk layout settings Additional software Summary Deployment 14/14 Initializing reboot after installation is complete Show config Installing Back Reboot	Network topology	🗸 Installing head	d node BCM packages					
Compute nodes	Head node	 Configuring ke Installing Ubu 	ernel and setting up boot ntu Server 22.04 base so	tloader oftware image(s)				
BMC configuration winstalling BCM packages to software images(s) winstalling offline selection of Python packages vinstalling installer NFS image Head node interfaces compute nodes interfaces Disk layout Disk layout Disk layout settings Additional software Summary Deployment 14/14 automatically reboot after installation is complete Show config Installing BCM packages to software images(s)	Compute nodes	 Installing base 	e distribution packages t	o software images(s)				
winstalling offline selection of Python packages Vetworks Creating node installer NFS image Head node interfaces Finalizing installation Compute nodes interfaces Initializing management daemon Disk layout Initializing management daemon Disk layout settings Additional software Summary Initializing reboot after installation is complete Show config Installing Back Reboot	BMC configuration	Installing BCM	I packages to software in	nages(s)				
Head node interfaces Finalizing installation Initializing management daemon Disk layout Initializing management daemon Initializing management daemon Disk layout Initializing management daemon Initializing management daemon Disk layout Initializing management daemon Initializing management daemon Summary Initializing management daemon Initializing management daemon	Networks	 Installing offling Creating node 	ne selection of Python pa e installer NFS image	ackages				
Compute nodes interfaces Disk layout Disk layout settings Additional software Summary Deployment 14/14 Automatically reboot after installation is complete Show config Install log Back Reboot	Head node interfaces	✓ Finalizing inst.	allation					
Disk layout settings Additional software Summary Deployment I4/14 Automatically reboot after installation is complete Show config Install log Back Reboot 1024 x 768	Compute nodes interfaces	🖌 Initializing ma	nagement daemon					
Disk layout settings Additional software Summary Deployment Automatically reboot after installation is complete Show config Install log Back Reboot 1024 x 768	Disk layout							
Additional software Summary Deployment Id / 14 Automatically reboot after installation is complete Show config Install log Back Reboot 1024 x 768 C	Disk layout settings				8			
Summary Deployment Automatically reboot after installation is complete Show config Install log Back Reboot 1024 x 768 C	Additional software							
Deployment 14/14 Automatically reboot after installation is complete Show config Install log Back Reboot 1024 x 768 Image: Complete to the complete tot to the complete to the complete tot to the complete	Summary							
Automatically reboot after installation is complete Show config Install log Back Reboot								
Show config Install log Back Reboot		Automatically re	boot after installation is	14/14 complete				
1024 x 768 (Ö A SSL 🖉 🔿 🖒					Show config	Install log Back	Reboot	
	1024 x 768					rði A ssl	00	

30. License the cluster by running the request-license and providing the product key. $_{\mbox{sudo}\mbox{-i}\mbox{ request-license}}$

Product Key (XXXXXX-XXXXXX-XXXXXX-XXXXXXX-XXXXXX):

31. Options:

4

- 1. If using the old method of MAC to IP allocation, skip line 32.
- 2. If employing the new method: Automatically detect MAC addresses based on switch and switchport, proceed to the next step.
- 3. Before advancing with the execution of the network automation application, certain prerequisites are necessary. Do as following:
 - Copy the "p2p_ethernet.csv" file from the USB stick to the following path: /cm/local/apps/bcm-superpod-network/config/p2p_ethernet.csv mv p2p_ethernet.csv /cm/local/apps/bcm-superpod-network/config/
- 32. Load the bcm-superpod-network module. module load bcm-superpod-network
- 33. Run the bcm-netautogen script.
- bcm-netautogen

Noticed New additional information has been provided.



Data was extracted from the p2p_ethernet.csv file to compute the quantities of Network Switches, DGX, IBSW, and PDUs. Accurate values must be provided during menu execution, which will be updated in future releases to utilize the count as physical cable connections.



The following generated files are important and contain data:

- Site network configuration /cm/local/apps/bcm-superpodnetwork/config/network-configuration.yml
- Site network allocations /cm/local/apps/bcm-superpodnetwork/config/network-allocations.yml

- Switch connection /cm/local/apps/bcm-superpod-network/config/switch-connections.yml
- IP Allocation Readme file /cm/local/apps/bcm-superpodnetwork/config/ip_allocations.md
- 34. Download and move cumulus-linux-5.10.0-mlx-amd64.bin and image-X86_64-3.11.2300.img to the following directory on the head node. Contact your TAM for access to the correct file and move the file to the following directory on the head node

```
mv cumulus-linux-5.10.0-mlx-amd64.bin
/cm/local/apps/cmd/etc/htdocs/switch/image/
mv image-X86_64-3.11.2300.img /cm/local/apps/cmd/etc/htdocs/switch/image/
```

- 35. Load the bcm-post-install module. module load bcm-post-install/
- 36. Run the bcm-pod-setup script.

These parameters were used:

- –C sets the base address of the computenet network.
- –S sets the base address of the storagenet network.
- -I sets the installation source.

bcm-pod-setup -C 100.126.0.0/16 -S 100.127.0.0/16 -I /dev/sdb

37. Check the nodes and their categories.

Extra options are used for device list to make the format more readable. %cmsh

```
%[bcm-head-01]%device list -f hostname:20,category:10
hostname(key) category
bcm-cpu-01 default
bcm-dgx-a100-01 dgx-a100
bcm-dgx-h100-01 dgx-h100
```

38. Confirm the config is correct for bcm-dgx-h100-01 / bcm-dgx-a100-01.

```
[bcm-head-01->device[bcm-dgx-h100-01]]% interfaces
```

[bcm-head-0	1->device[bcm-dgx-h100-01]->interfa	aces]% list		
Туре	Network device name	IP		Network	Start if
bmc	ipmi0		10.0.92.50	ipminet	
always					
bond	bond0 [prov]		10.0.93.12	dgxnet	
always					
physical	enp170s0f1np1 (bond0)	0.0.0.0			always
physical	enp41s0f1np1 (bond0)	0.0.0.0			always
physical	ibp154s0		100.126.5.14	ibnetcompute	always
physical	ibp170s0f0		100.127.2.2	ibnetstorage	always
physical	ibp192s0		100.126.6.14	ibnetcompute	always
physical	ibp206s0		100.126.7.14	ibnetcompute	always
physical	ibp220s0		100.126.8.14	ibnetcompute	always
physical	ibp24s0		100.126.1.14	ibnetcompute	always
physical	ibp41s0f0		100.127.1.2	ibnetstorage	always
physical	ibp64s0		100.126.2.14	ibnetcompute	always
physical	ibp79s0		100.126.3.14	ibnetcompute	always



39. Check the Ethernet Switch are in the devices.

cmsh >> device >>	list				
[bcm-head-01->device]% Type	list Hostname (key)	MAC	Category	Ip	Network
HeadNode PhysicalNode PhysicalNode PhysicalNode PhysicalNode PhysicalNode Switch Switch Switch Switch	bcm-head-01 slogin-01 slogin-02 storage-cpu-01 storage-dgx-01 storage-dgx-02 IPMI-01 SJC4-C25-TOR-04 TOR-01 TOR-02	B8:CE:F6:FF:98:A0 00:00:00:00:00:00:00 00:00:00:00:00:00	default default default dgx-a100 dgx-a100	10.0.105.189 10.0.105.140 10.0.105.141 10.0.105.139 10.0.0.11 10.0.0.12 10.0.255.194 10.0.128.61 10.0.128.61 10.0.128.60	internalnet internalnet internalnet dgxnet1 dgxnet1 loopback ipminet2 ipminet
Switch [bcm-head-01->device]%	ib01	E8:EB:D3:31:FE:D2		10.0.105.41	ipminet2

Validate under the "Type" Switches are added after executing *bcm-pod-setup*

40. Add Switch *credential*, under each IPMI, TOR and SPINE switch.

Example:

[bcm-head-01->device]% use IPMI-01				
[bcm-head-01->device[IPMI-01]]% accesssettings				
[bcm-head-01->device*[IPMI-01*]->accesssettings*]%	set	username	cumulus	
[bcm-head-01->device*[IPMI-01*]->accesssettings*]%	set	password	<pass></pass>	

commit

- quit
- 41. To allocate IP via switch port:
 - After running bcm-pod-setup, once all the networks, and devices object is added to the Bright
 - Make sure the IPMI switch is UP in the Bright before moving to the next step
 - Based on the switch and switch port configuration for each node, navigate to the device then nodes, and execute the below command:
 - **setmacviaswitchport** Set the MAC of a device via the MAC found on its switch ports

- It will access the switch and pull the MAC address based on the switch port allocation.



42. To gather UFM metrics

- Add UFM to the Bright with mgmt. IP address
- Make sure UFM Promethus exporter is enable in UFM
- ## You can check by curl command from bright:

curl http://<UFM-IP>:9001/metrics

Configure Bright with following: monitoring setup add prometheus UFM set urls https://<UFM-IP>:9001/metrics set -e NoPostAllowed yes nodeexecutionfilters active commit

Wait for (~2mins) for data to be collected
get measurables

To plot
monitoring labeledentity
list

Using the index value: instantquery <index value>

4. Head Node Configuration

This section addresses configuration steps to be performed on BCM head nodes.

Use the root (not cmsh) shell.

- 1. In /cm/local/apps/cmd/etc/cmd.conf, uncomment the AdvancedConfig parameter.
 AdvancedConfig = { "DeviceResolveAnyMAC=1" } # modified value
- Restart the CMDaemon to enable dependable PXE booting from bonded interfaces.
 # systemctl restart cmd

The cmsh session will be disconnected because of restarting the CMDaemon. Type connect to reconnect after the CMDaemon has restarted. Or enter exit and then restart cmsh.



Older method: The steps below are only necessary if you are employing the MAC to IP Allocation method. Newer methods omit assigning MAC addresses based on the ports **** MOVE to STEP 9**

3. The steps that follow are performed on the head node and should be run for all DGX systems.

Note: Double check the MAC address for each interface and the IP number for the bond0 interface. Mistakes here will be difficult to diagnose.

For DGX A100 systems, the commands should be like this code block.

```
# cmsh
% device
% use bcm-dgx-a100-01
% interfaces
% use enp225s0f1np1
% set mac B8:CE:F6:2F:08:69
% use enp97s0f1np1
% set mac B8:CE:F6:2D:0E:A7
% ...
% commit
For DGX H100 systems, the commands should be like this code block.
# cmsh
% device
% use bcm-dgx-h100-01
% interfaces
% use enp170s0f1np1
% set mac B8:CE:F6:2F:08:69
% use enp41s0f1np1
% set mac B8:CE:F6:2D:0E:A7
% ..
% commit
```

4. Verify the configuration.

```
This example is for a DGX A100 system. The output for a DGX H100 system is similar.
   [bcm10-headnode->device[dgx-01]]% get provisioninginterface
   bond0
   % interfaces
   % list
             Network device name IP
                                                             Network Start if
   Туре
                -----
                                                  -----
                                                  10.130.111.68 ipminet
              ipmi0
   bmc
   always
bond bond0 [prov]
physical enp225s0f1np1 (bond0) 0.0.0.0
physical enp97s0f1np1 (bond0) 0.0.0.0
                                             10.130.122.5 internalnet
                                                                              always
                                                                               always
                                                                                 always
5. Identify the nodes by setting the MAC address for the provisioning interface for
   each node to the MAC address listed in the site survey.
   % device
   % use bcm-dax-h100-01
   % set mac b8:ce:f6:2f:08:69
   % use bcm-dgx-h100-02
   % set mac 0c:42:a1:54:32:a7
   % use bcm-dgx-h100-03
   % set mac 0c:42:a1:0a:7a:51
   % use bcm-dgx-h100-04
   % set mac 1c:34:da:29:17:6e
   % foreach -c dgx-h100 (get mac)
   B8:CE:F6:2F:08:69
   0C:42:A1:54:32:A7
   0C:42:A1:0A:7A:51
   1C:34:DA:29:17:6E
6. If all the MAC addresses are set properly, commit the changes.
   % device commit
   % quit
7. Set the MAC addresses for the Ethernet interfaces.
   For control nodes connected to DGX A100 systems, use the following commands.
   % device
   % use bcm-cpu-01
   % interfaces
   % use ens2f0np0
   % set mac 88:e9:a4:92:26:ba
   % use ens2f1np1
   % set mac 88:e9:a4:92:26:bb
   % commit
```

```
For control nodes connected to DGX H100 systems, use the following commands.
```

```
% device
% use bcm-cpu-01
% interfaces
% use enp37s0np0
% set mac 88:e9:a4:92:26:ba
% use enp65s0np0
% set mac 88:e9:a4:92:26:bb
% commit
```

If the head node uses a bonded interface, use the following commands. You may need to reboot the head node and redo request-license steps

% device

```
% use headnode-01
% interfaces
% use ens1np0
% clear ip
% clear network
% add physical ens2np0
% set mac 88:e9:a4:20:18:d8
% add bond bond0
% append interfaces ens1np0 ens2np0
% set mode 1
% set network internalnet
% set ip 10.180.115.189
% ...
% ..
% set provisioninginterface bond0
% interfaces
% use ipmi0
% set ip 10.180.217.154
```

8. Set the IP address for the bond0 interface.

```
% device
% use bcm-cpu-01
% interfaces
% use bond0
% set ip 10.127.3.15
% commit
```

9. Power on and provision the cluster nodes.

For initial provisioning, the cluster nodes must be powered on either directly or by using a KVM. It will take several minutes for the nodes to go through their BIOS. After that, node status progress will be displayed as the nodes are being provisioned. Monitor the /var/log/messages and /var/log/node-installer log files to verify that everything is proceeding smoothly.

Newer method:

Prerequisite

- The switch must be in the **UP** position to retrieve the MAC address from the switch and switchport specified in the CSV file.
 - If any TORs are offline, Bright will be unable to retrieve MAC addresses from the switch.

Verify:

 Ensure that the weather node interfaces have been assigned a switch and switchport. Navigate to cmsh > device > use <node> and execute the "show" command.

 Switch ports
 IPW01091129 (Internation)

 • Interfaces
 IPW01091129 (Internation)

 • Next Step:
 <4 in submode>

- Reboot the Node
 - Detailed MAC addresses per node are not necessary as Bright will automatically detect them based on Switch and Switchport assignments.

5. High Availability

1. Verify that the head node has power control over the cluster nodes.

```
% device
% power -c dgx-h100 status
[-head1->device]% power -c dgx-h100 status
ipmi0 ...... [ ON ] bcm-dgx-h100-01
ipmi0 ...... [ ON ] bcm-dgx-h100-02
ipmi0 ...... [ ON ] bcm-dgx-h100-03
ipmi0 ...... [ ON ] bcm-dgx-h100-04
[bcm-head-01->device]%
```

2. Power off the cluster nodes.

The cluster nodes must be powered off before configuring HA.

% power -c dgx-h100 off			
ipmi0	[0FF] bcm-dgx-h100-01
ipmi0	[0FF] bcm-dgx-h100-02
ipmi0	[0FF] bcm-dgx-h100-03
ipmi0	[0FF] bcm-dgx-h100-04

- 3. Start the cmha-setup CLI wizard as the root user on the primary head node. # cmha-setup
- 4. Choose Setup and then select SELECT.

Please choose 'Setup' to enter the failover settings menu, 'Shared' Storage' to setup shared storage, 'Status' to view the failover status, if high availability has already been setup. Choose 'Help' to see a detailed description of the options available.
SeturConfigure failover seturShared StorageConfigure shared storageStatusView failover statusHelpCmha-setup help
<pre>SELECT> < QUIT ></pre>

5. Choose Configure and then select NEXT.



6. Verify that the cluster license information found cmha-setup is correct and then select CONTINUE.

	The following MAC addresses have been found in the license information:
	04:3F:72:E7:67:07 14:02:EC:DA:AF:18
	If they are correct, then please press 'Continue'.
	If not, one of the following has to be done:
	 If you have not activated your Product Key, please run request-license and follow instructions. If you have run out of licenses, please contact your reseller, or contact our support.
	Press 'BACK' to go back to the failover setup menu.
	88% - C <mark>ontlinue></mark> < back >
-	

7. Provide an internal Virtual IP address that is to be used by the active head node in the HA configuration.



8. Provide the name of the secondary head node and then select NEXT.

Name:	basepod-head2	
	< NEXT > < BA	CK >

9. Because DGX SuperPOD uses the internal network as the failover network, select $_{\mbox{\scriptsize SKIP}}$

This screen takes parameters of a dedicated failover network that will be created. Please keep in mind, to select a network name different from those that have been defined already. Name: Base address: 10.151.0.0 Netmask: 255.255.0.0 Domain name: failover.cluster	S
< NEXT > < SKTP > < BACK >	

10. Configure the IP addresses for the secondary head node that the wizard is about to create and then select NEXT.

ens1f1np1[10.130.122.0/24] ens10f1 [10.130.121.0/24] inpi0 [10.130.121.0/24]	10.130.122.254 10.130.121.254	10.130.122.253 10.130.121.253
ens10f0 [10.130.111.64/26]	10.130.111.126	10.130.111.125
< NEX	BACK	>

11. The wizard shows a summary of the information that it has collected. The VIP that will be assigned to the internal and external interfaces, respectively.



12. Select Yes to proceed with the failover configuration.

Proceed with failover setup? This will write the new setting to the failover config.
< ¥@s > < No >

13. Enter the root password and then select OK.

Please enter the mysql root password:	
 UX > <lancet></lancet> 	

14. The wizard implements the first steps in the HA configuration. If all the steps show OK, press ENTER to continue. The progress is shown here.

ing proce Enter to contained the progress is enter	••••		
Initializing failover setup on master	[0K]
Updating shared internal interface	[0K]
Updating shared external interface	[0K]
Updating extra shared internal interfaces	[0K]
Cloning head node	[0K]
Updating secondary master interfaces	[0K]
Updating Failover Object	[0K]
Restarting cmdaemon	[0K]
Press any key to continue			

15. When the failover setup installation on the primary master is complete, select OK to exit the wizard

The failover setup initialization on the primary master is done. Now boot the secondary master into the rescue environment and run the following command:
/cm/cm-clone-installfailover
and follow the instructions.
Once the installation has begun, select 'Install Progress' from the Failover setup menu, to see the installation progress of the clone machine. When the installation is complete, and the secondary master is up, select 'Finalize' from the Failover setup menu, to complete the failover setup process.
94% < 0K >

- 16. PXE boot the secondary head node and then select RESCUE from the grub menu. Because this is the initial boot of this node, it must be done outside of BCM (BMC or physical power button).
- 17. Select RESCUE from the grub menu.



19. After the secondary head node has booted into the rescue environment, run the /cm/cm-clone-install --failover command, then enter yes when prompted. The secondary head node will be cloned from the primary.



20. When cloning is completed, enter y to reboot the secondary head node. The secondary must be set to boot from its hard drive. PXE boot should not be enabled. 21. Wait for the secondary head node to reboot and then continue the HA setup procedure on the primary head node.

1 *Welcome to the Cluster Manager rescue environment*	
) I Creating failover/clone nodes:	
I × Install the secondary head node	
\$ /cm/cm-clone-installfailover	
1 * Create a clone of the primary head node	
\$ /cm/cm-clone-installclonehostname=new-hostname	
<pre>1 * Install the secondary (failover) head node and reboot automatically</pre>	
I \$ ∕cm∕cm-clone-installfailouerreboot	
I ≈ Help	
1 \$ ∕cm/cm-clone-installhelp	
<pre>biological and the second second</pre>	

22. Select finalize from the cmha-setup menu and then select NEXT. This will clone the MySQL database from the primary to the secondary head node.

Select 'Co	onfigure' to configure	failover setup, 'Finalize' to finalize	
'Clone Ins	stall' to see the insta if the failover conf	allation instructions for the secondary iguration has been completed. Choose	
'Install F	Progress', if the secon lover' to remove exist	ndary head node is being installed. Choose	
	Configure Clone Install	Configure failover setup View install instructions	
	Install Progress	View install progress	
	Undo Failover	Remove failover setup	
	Math	Main menu	
L			
	< NEXT >	< BACK >	

23. Select CONTINUE on the confirmation screen.

CONFIRM
CONFIRM
firmation is required to proceed.
following tasks will be performed:
under an and an hand and we address in andress der base
Update secondary nead node MAL address in cmdaemon database. Clana MySOL database(s) to the secondary head node
ctone hysic database(s) to the secondary head hode.
ss 'CONTINUE' to proceed with the finalize operation,
press 'BACK' to return to the failover Setup menu.

24. Enter the root password and then select OK.

Please enter the mysql root password: [**********	—
<mark>< OK ></mark> <cancel></cancel>	

25. The cmha-setup wizard continues. Press ENTER to continue when prompted.



Updating secondary master mac address [0K]
Initializing failover setup on bcm-head-02 [0K]
Stopping cmdaemon [0K]
Cloning cmdaemon database [0K]
Checking database consistency [0K]
Starting cmdaemon, chkconfig services [0K]
Cloning workload manager databases [0K]
Cloning additional databases[0K]
Update DB permissions [0K]
Checking for dedicated failover network [0K]
Press any key to continue		

26. The Finalize step is now completed. Select REBOOT and wait for the secondary head node to reboot.



27. The secondary head node is now UP.

% device list -f	hostname:20,cat	egory:12,ip:20),status:15					
hostname (key)	category	ip				statu	IS	
bcm-head-01		10.	.130.122.25	4	[UP]	
bcm-head-02		10.	.130.122.25	3	[UP]	
bcm-dgx-h100-01	dgx-h100	10.130.122.5	[DOWN]			
bcm-dgx-h100-02	dgx-h100	10.130.122.6	[DOWN]			
bcm-dgx-h100-03	dgx-h100	10.130.122.7	[DOWN]			
bcm-dgx-h100-04	dgx-h100	10.130.122.8	[DOWN]			

28. Select Shared Storage from the cmha-setup menu and then select SELECT. In this final HA configuration step, cmha-setup will copy the /cm/shared and /home directories to the shared storage and configure both head nodes and all cluster nodes to mount it.

	Welcome to the Bright Cluster Manager High Availability Setup Utility. Please choose 'Setup' to enter the failover settings menu, 'Shared Storage' to setup shared storage, 'Status' to view the failover status, if high availability has already been setup. Choose 'Help' to see a detailed description of the options available. Setup Status Help Configure failover setup Configure shared storage View failover status Cmha-setup help
_	<pre><guit></guit></pre>

29. Choose NAS and then select SELECT.

The following shared type of shared storag	storage solutions a je from the menu bel	re supported. Please select a ow.	
DAS DRBD	Network Attached S Direct Attached St Distributed Redund	<mark>torage</mark> orage ant Block Device	
	< <mark>SELECT></mark>	< BACK >	

- Please select resources that will be shared: [★] /cm/shared ★ /home NEXT > < BACK >
- 30. Choose both /cm/shared and /home and then select NEXT.

31. Provide the IP address of the NAS host, the paths that the /cm/shared and /home directories should be copied to on the shared storage, and then select NEXT. In this case, /var/nfs/general is exported, so the /cm/shared directory will be copied to 10.130.122.252:/var/nfs/general/cmshared, and it will be mounted over /cm/shared on the cluster nodes.

Please fill in NAS parameters	
NAS host: Path to /cm/shared: Path to /home:	10.130.122.252 /var/nfs/general/cmshared /var/nfs/general/home
< NEXT >	< BACK >

32. The wizard shows a summary of the information that it has collected. Select EXIT to continue.

NAS Host: 10.180.116.4 NAS Paramas: Path to /cm/shared: /bcmvikinghome Path to /home: /bcmvikingshared		NAS Host: 10.180.116.4 NAS Params: Path to ∕cm∕shared: ∕bcmvikinghome Path to ∕home: ∕bcmvikingshared	
	_		

33. When asked to proceed with the NAS setup, select Yes to continue. This will initiate a copy and update to fsexports.



35. When setup completes, press any key to finish HA setup. The progress is shown here:

1 5			
Copying NAS data	[0K]
Mount NAS storage	[0K]
Remove old fsmounts	[0K]
Add new fsmounts	[0K]
Remove old fsexports	[0K]
Write NAS mount/unmount scripts	[0K]
Copy mount/unmount scripts	[0K]
Press any key to continue			

36. cmha-setup is now complete. EXIT the wizard to return to the shell prompt.



37. Run the cmha status command to verify that the failover configuration is correct and working as expected.

The command tests the configuration from both directions: from the primary head node to the secondary, and from the secondary to the primary. The active head node is indicated by an asterisk.

```
# cmha status
Node Status: running in active mode
bcm-head-01* -> bcm-head-02
 failoverping [ OK ]
 mysql [ OK ]
            [ OK ]
 ping
 status
            [ OK ]
bcm-head-02 -> bcm-head-01*
 failoverping [ OK ]
             [ OK ]
 mysql
 ping
             [ OK ]
             [ OK ]
 status
```

38. Verify that the /cm/shared and /home directories are mounted from the NAS server.

```
# mount
   . . . some output omitted . . .
   10.130.122.252:/var/nfs/general/cmshared on /cm/shared type nfs4
   (rw, relatime, vers=4.2, rsize=32768, wsize=32768, namlen=255, hard, proto=tcp, timeo=600, retrans=2
   , sec=sys, clientaddr=10.130.122.253, local_lock=none, addr=10.130.122.252)
   10.130.122.252:/var/nfs/general/home on /home type nfs4
   (rw, relatime, vers=4.2, rsize=32768, wsize=32768, namlen=255, hard, proto=tcp, timeo=600, retrans=2
   , sec=sys, clientaddr=10.130.122.253, local_lock=none, addr=10.130.122.252)
39. Login to the head node to be made active and run cmha makeactive.
   # ssh bcm-head-02
   # cmha makeactive
   _____
   This is the passive head node. Please confirm that this node should become
   the active head node. After this operation is complete, the HA status of
   the head nodes will be as follows:
   bcm-head-02 will become active head node (current state: passive)
   bcm-head-01 will become passive head node (current state: active)
   _____
   Continue(c)/Exit(e)? c
   Initiating failover..... [ OK ]
   bcm-head-02 is now active head node. makeactive successful
40. Run the cmsh status command again to verify that the secondary head node has
   become the active head node.
   # cmha status
   Node Status: running in active mode
   bcm-head-02* -> bcm-head-01
     failoverping [ OK ]
    mysql [ OK ]
    ping
                [ OK ]
               [ OK ]
     status
   bcm-head-01 -> bcm-head-02*
     failoverping [ OK ]
     mysql [ OK ]
     ping
                [ OK ]
     status
               [ OK ]
```

41. Manually failover back to the primary head node by running cmha makeactive.

```
# ssh bcm-head-01
   # cmha makeactive
   _____
   This is the passive head node. Please confirm that this node should become
   the active head node. After this operation is complete, the HA status of
   the head nodes will be as follows:
   bcm-head-01 will become active head node (current state: passive)
   bcm-head-02 will become passive head node (current state: active)
   _____
   Continue(c)/Exit(e)? c
   Initiating failover..... [ OK ]
   bcm-head-01 is now active head node, makeactive successful
42. Run the cmsh status command again to verify that the primary head node has
   become the active head node.
                                               # cmha status
   Node Status: running in active mode
   bcm-head-01* -> bcm-head-02
    failoverping [ OK ]
    mysql [ OK ]
              [ OK ]
    ping
    ping [ UK ]
status [ OK ]
   bcm-head-02 -> bcm-head-01*
    failoverping [ OK ]
    mysql [ OK ]
ping [ OK ]
status [ OK ]
43. Power on the cluster nodes.
   # cmsh -c "device ; power -c dgx-h100 on"
   ipmi0 ..... [ ON ] bcm-dgx-h100-01
   ipmi0 ..... [ ON ] bcm-dgx-h100-02
   ipmi0 ..... [ ON ] bcm-dgx-h100-03
   ipmi0 ..... [ ON ] bcm-dgx-h100-04
44. This concludes the setup and verification of HA.
```

6. Slurm Setup

- 1. Update the interface names on the slogin nodes.
 - % device use slogin-01

If the slogin-01 does not have the expected interface names, update the interface names.

% use networkdevicename

% set networkdevicename new-name

- Assign the MAC addresses to the slogin nodes. device use slogin-01 set mac <MAC address>
- 3. Power on and install the slogin nodes.
- 4. Run the bcm-install-slurm script. Use the following parameters:
 - Installation source for the --bcm-media parameter. It can be either a USB or a path to a *.iso file.
 - Use the -A parameter to run the script in air-gapped mode.
 - If CMHA is set up but has failover ping errors, append --ignore-ha-errors.
 - If there is only one slogin node, append --ignore-missing-login-node.

bcm-install-slurm -A --bcm-media <path to installer image or usb device to mount>

5. Confirm that the slurmd file is present in DGX image before provisioning DGX nodes, if not create it.

The same file is needed for both DGX A100 and DGX H100 systems. This example is for DGX H100 systems. It is observed that NCCL tests with PMIX need this file.

```
cat /cm/images/dgx-os-6.3-h100-image/etc/sysconfig/slurmd
PMIX_MCA_ptl=^usock
PMIX_MCA_psec=none
PMIX_SYSTEM_TMPDIR=/var/empty
PMIX_MCA_gds=hash
```

6. Reboot the slogin and compute nodes.

```
cmsh
device
reboot -c slogin
reboot -c dgx-h100
```

7. To simplify the configuration, modify the slurmclient-gpu role to remove the slurm-client role and convert slurm-client-gpu to instead use that name.

```
cmsh
configurationoverlay
remove slurm-client
commit
use slurm-client-gpu
set name slurm-client
commit
```

8. For DGX A100 systems, clear the type value and set the correct core association with each GPU entry for maximum performance.

cmsh configurationoverlay use slurm-client roles use slurmclient genericresources use gpu0 clear type set cores 48-63,176-191 use gpu1 clear type set cores 48-63,176-191 use gpu2 clear type set cores 16-31,144-159 use gpu3 clear type set cores 16-31,144-159 use gpu4 clear type set cores 112-127,240-255 use gpu5 clear type set cores 112-127,240-255 use gpu6 clear type set cores 80-95,210-223 use gpu7 clear type set cores 80-95,210-223 commit

9. For DGX H100 systems, generic resources are set to autodetect.

```
Use this script.
```

```
cmsh
wlm
set gpuautodetect nvml
commit
configurationoverlay
use slurm-client
roles
use slurmclient
set gpuautodetect nvml
commit
genericresources
foreach * (remove)
commit
add autodetected-gpus
set name gpu
set count 8
set addtogresconfig yes
commit
```

Note: addtogresconfig is set by default to "yes" and does not need to be set explicitly.

Which should yield output like this.

The gres.conf file will be updated automatically by BCM—these settings align with the expectations of various scripts and tools in the NVIDIA ecosystem and will then maximize the compatibility of this environment with those scripts and tools.

10. If the /home directory is not mounted on the nodes, increase the number of retries. Due to a race condition between the bond@ interface being up and /home being mounted, sometimes /home will not be mounted. Increasing the number of retries should fix the issue.

```
cmsh
category
use dgx-h100
fsmounts
use /home
set mountoptions "x-systemd.mount-timeout=150,defaults,_netdev,retry=5,vers=3"
```

The pod setup might leave stale repos in an air gapped environment. In which case following files need to be removed manually on the login nodes

```
cd /etc/apt/sources.list.d/
```

Disable following repos:

mv local.list local.list.disabled mv cm.disabled cm.list mv cm-ml.disabled cm-ml.list mv /etc/apt/sources.disabled /etc/apt/sources.list

7. Upgrade Infiniband Switches

Check out the instructional <u>guide</u> on upgrading InfiniBand switches using the UFM appliance.

The ability to update the server software is applicable only for hosts (servers) with the UFM Agent.

To upgrade the software:

- Select a device.
- From the right-click menu, select Software Update.
- Enter the parameters listed in the following table.

Parameter	Description
Protocol	Update is performed via SCP protocol
IP	Enter the BCM head node bonded IP
Path	Enter the parent directory of the SCP directory structure for the Upgrade image. The path should not be an absolute path and should not contain the first slash (/) or trailer slash.
User	Name of the BCM head node root user
Password	Enter the host password

4

- Click Submit to save your changes.
- Right-click InfiniBand to "reboot" the switch.

8. SuperPOD Validation

After the installation is completed, using the bcm-validate-pod command, one can verify whether all the settings are applied correctly according to the specification.

```
root@vikinghead-01:~# module load bcm-post-install/
root@vikinghead-01:~# bcm-validate-pod --verbose
Checking slurm on head node:
Run: /cm/shared/apps/slurm/current/bin/sinfo --version
slurm 23.02.7
Passed
Checking ubuntu on head node:
Run: /usr/bin/lsb_release -a
Distributor ID: Ubuntu
                Ubuntu 22.04.2 LTS
Description:
Release:
                22.04
Codename:
                jammy
Passed
```

This module validates the completed pod setup with the expected configuration.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Materials (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA'S aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, DGX, DGX NVIDIA SuperPOD, and NVIDIA Base Command Manager are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Authors

AY, DS, JB, HS, LH, MK, ML, RA, RS, and VS.

Copyright

© 2023 NVIDIA Corporation. All rights reserved.