White Paper

NVDIA® DGX™ Data Center Reference Design

Easy Deployment of DGX Servers for Deep Learning

2018-07-19

© 2018 NVIDIA Corporation.

Contents

Abstract

The NVIDIA® DGX™ family of supercomputers for artificial intelligence (AI) and the NVIDIA GPU Cloud (NGC) Deep Learning (DL) containers provide the world's best solution for AI software development. However, many organizations are unsure how to size, install, and manage a DGX infrastructure for large-scale AI development efforts. The DGX POD reference architecture provides a blueprint for easy sizing and deployment of DGX servers for large-scale DL. A DGX POD is a single data center rack containing up to nine NVIDIA DGX-1™ servers, storage, networking, and NVIDIA AI software to support workgroups of AI developers. Multiple DGX PODs can be combined into a cluster to support larger workgroups. The DGX POD reference architecture is based on the NVIDIA DGX SATURNV AI supercomputer which has 1000 DGX-1 servers and powers autonomous vehicle software and internal AI R&D across NVIDIA research, graphics, HPC, and robotics.

This white paper is divided into the following sections:

- AI Workflow and Sizing
- NVIDIA AI Software
- DGX POD Design
- DGX POD Installation and Management

© 2018 NVIDIA Corporation.

1. AI Workflow and Sizing

A typical AI software development workflow follows the steps shown in the following figure:

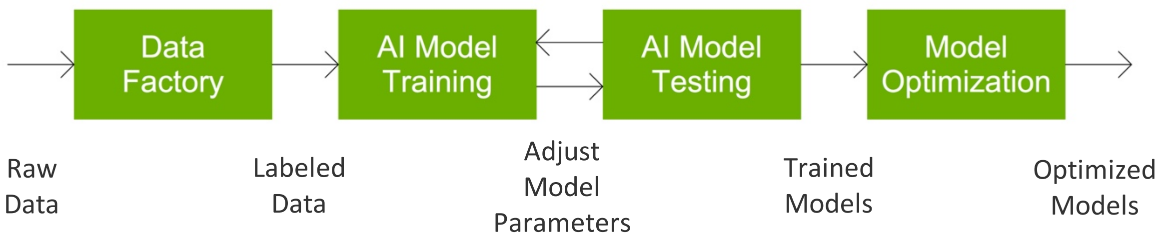


Figure 1. AI workflow

The workflow is detailed as follows:

1. A data factory collects raw data and includes tools used to pre-process, index, label, and manage data.
2. AI models are trained with labeled data using a DL framework from the NVIDIA GPU Cloud (NGC) container repository running on DGX servers with Volta Tensor Core GPUs.
3. AI model testing and validation adjusts model parameters as needed and repeats training until the desired accuracy is reached.
4. AI model optimization for production deployment (inference) is completed using the NVIDIA TensorRT optimizing inference accelerator.

Sizing DL training is highly dependent on data size and model complexity. A single DGX-1 can complete a training experiment on a wide variety of AI models in one day. For example, the autonomous vehicle software team at NVIDIA developing NVIDIA DriveNet uses a custom Resnet-18 backbone detection network with 960x480x3 image size and trains at 480 images per second on a DGX-1 server allowing training of 120 epochs with 300k images in 21 hours. Internal experience at NVIDIA has shown that five developers collaborating on the development of one AI model provides the optimal development time. Each developer typically works on two models in parallel thus the infrastructure needs to support ten model training experiments within the desired TAT (turn-around-time). A DGX POD with nine DGX-1 servers can provide one day TAT for model training for the five-developer workgroup. During schedule-critical times, multi-node scaling can reduce turnaround time from one day to four hours using eight DGX-1 servers. Once in production, additional DGX-1 servers will be necessary to support on-going model refinement and regression testing.

© 2018 NVIDIA Corporation.

2. NVIDIA AI Software

NVIDIA AI software running on the DGX POD provides a high-performance DL training environment for large scale multi-user AI software development teams. NVIDIA AI software includes the DGX operating system (DGX OS), cluster management and orchestration tools, NVIDIA libraries and frameworks, workload schedulers, and optimized containers from the NGC container registry. To provide additional functionality, the DGX POD management software includes
third-party open-source tools recommended by NVIDIA which have been tested to work on DGX PODs with the NVIDIA AI software stack. Support for these tools can be obtained directly through third-party support structures.

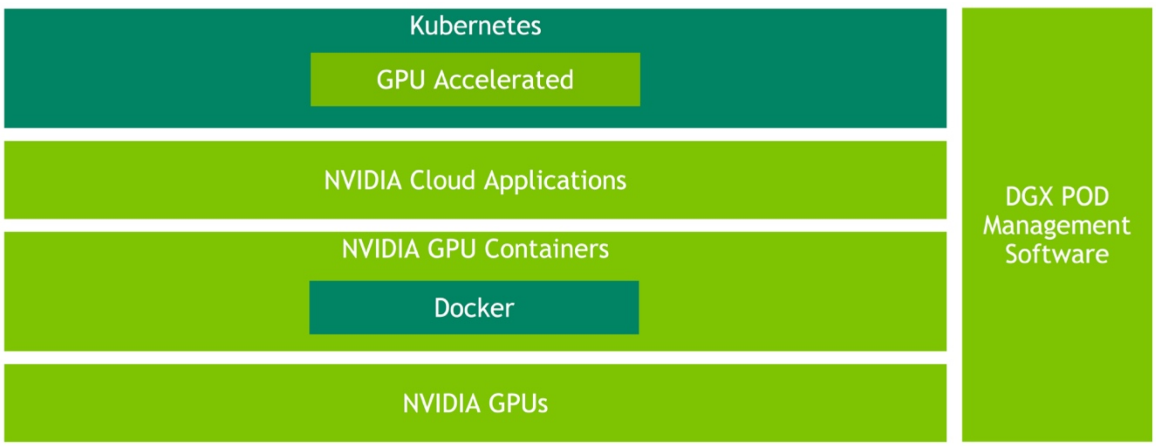


Figure 2. NVIDIA AI software

The foundation of the NVIDIA AI software stack is the DGX OS, built on an optimized version of the Ubuntu Linux operating system and tuned specifically for the DGX hardware.

The DGX OS software includes certified GPU drivers, a network software stack, pre-configured NFS caching, NVIDIA data center GPU management (DCGM) diagnostic tools, GPU-enabled container runtime, NVIDIA CUDA® SDK, cuDNN, NCCL and other NVIDIA libraries, and support for NVIDIA GPUDirect™. The DGX OS software can be automatically re-installed on demand by the DGX POD management software.

The management software layer of the DGX POD (Figure 3) is composed of various services running on the Kubernetes container orchestration framework for fault tolerance and high availability. Services are provided for network configuration (DHCP) and fully-automated DGX OS software provisioning over the network (PXE).

© 2018 NVIDIA Corporation.

Monitoring of the DGX POD utilizes Prometheus for server data collection and storage in a time-series database. Cluster-wide alerts are configured with Alertmanager, and DGX POD metrics are displayed using the Grafana web interface. For sites required to operate in an air-gapped environment or needing additional on-premises services, a local container registry mirroring NGC containers, as well as Ubuntu and Python package mirrors, can be run on the Kubernetes management layer to provide services to the cluster.

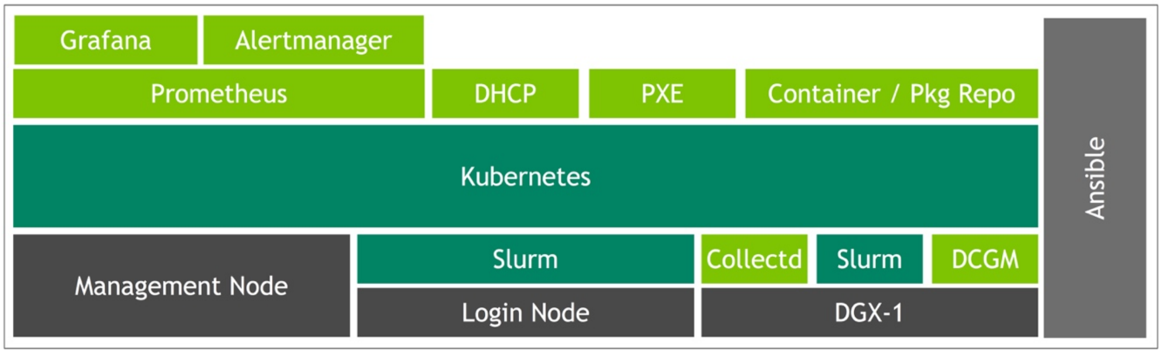


Figure 3. DGX POD management software

The DGX POD software allows for dynamic partitioning between the nodes assigned to Kubernetes and Slurm such that resources can be shifted between the partitions to meet the current workload demand. A simple user interface allows administrators to move DGX-1 servers between Kubernetes- and Slurm-managed domains.

Kubernetes serves the dual role of running management services on management nodes as well as accepting user-defined workloads, and is installed on every server in the DGX POD. Slurm runs only user workloads and is installed on the login node as well as the DGX compute nodes. The DGX POD software allows individual DGX-1 servers to run jobs in either Kubernetes or Slurm. Kubernetes provides a high level of flexibility in load balancing, node failover, and bursting to external Kubernetes clusters, including NGC public cloud instances. Slurm supports a more static cluster environment but provides advanced HPC-style batch scheduling features including multi-node scheduling that some workgroups may require. With the DGX POD software, idle systems can be moved back and forth as needed between the Kubernetes and Slurm environments. Future enhancements to Kubernetes are expected to support all DGX POD use cases in a pure Kubernetes environment.

User workloads on the DGX POD primarily utilize containers from NGC (Figure 4), which provides researchers and data scientists with easy access to a comprehensive catalog of GPU-optimized software for DL, HPC applications, and HPC visualization that take full advantage of the GPUs. The NGC container registry includes NVIDIA tuned, tested, certified, and maintained containers for the top DL frameworks such as TensorFlow, PyTorch, and MXNet. NGC also has third-party managed HPC application containers, and NVIDIA HPC visualization containers.

© 2018 NVIDIA Corporation.

Figure 4. NGC overview

Management of the NVIDIA AI software on the DGX POD is accomplished with the Ansible configuration management tool. Ansible roles are used to install Kubernetes on the management nodes, install additional software on the login and DGX-1 servers, configure user accounts, configure external storage connections, install Kubernetes and Slurm schedulers, as well as performing day-to-day maintenance tasks such as new software installation, software updates, and GPU driver upgrades.

The software management stack and documentation are available as an open source project on GitHub at:

https://github.com/NVIDIA/deepops

© 2018 NVIDIA Corporation.

3. DGX POD

The DGX POD (Figure 5) is an optimized data center rack containing up to nine DGX-1 servers, twelve storage servers, and three networking switches to support single and multi-node AI model training and inference using NVIDIA AI software.

There are several factors to consider when planning a DGX POD deployment in order to determine if more than one rack is needed per DGX POD. This reference architecture is based on a single 35 kW high-density rack to provide the most efficient use of costly data center floorspace and to simplify network cabling. As GPU usage grows, the average power per server and power per rack continues to increase. However, older data centers may not yet be able to support the power and cooling densities required; hence the three-zone design allowing the DGX POD components to be installed in up to three lower-power racks.

The DGX POD is designed to fit within a standard-height 42 RU data center rack. A taller rack can be used to include redundant networking switches, a management switch, and login servers. This reference architecture uses an additional utility rack for login and management servers, and has been sized and tested with up to six DGX PODs. Larger configurations of DGX PODs can be defined by an NVIDIA solution architect.

A primary 10 GbE (minimum) network switch is used to connect all servers in the DGX POD and to provide access to a data center network. The DGX POD has been tested with an Arista switch with 48 x 10 GbE ports and 4 x 40 GbE uplinks. VLAN capabilities of the networking hardware are used to allow the out-of-band management network to run independently from the data network, while sharing the same physical hardware. Alternatively, a separate 1 GbE management switch may be used. While not included in the reference architecture, a second 10 GbE network switch can be used for redundancy and high availability. In addition to Arista, NVIDIA is working with other networking vendors who plan to release switch reference designs compatible with the DGX POD.

A 36-port Mellanox 100 Gbps switch is configured to provide four 100 Gbps InfiniBand connections to the nine DGX-1 servers in the rack. This provides the best possible scalability for multi-node jobs. In the event of switch failure, multi-node jobs can fall back to use the 10 GbE switch for communications. The Mellanox switch can also be configured in 100 GbE mode for organizations that prefer to use Ethernet networking. Alternately, by configuring two 100 Gbps ports per DGX-1 server, the Mellanox switch can also be used by the storage servers.

With the DGX family of servers, AI and HPC workloads are fusing into a unified architecture. For organizations that want to utilize multiple DGX PODs to run cluster-wide jobs, a core InfiniBand switch is configured in the utility rack in conjunction with a second 36-port Mellanox switch in DGX POD.

© 2018 NVIDIA Corporation.

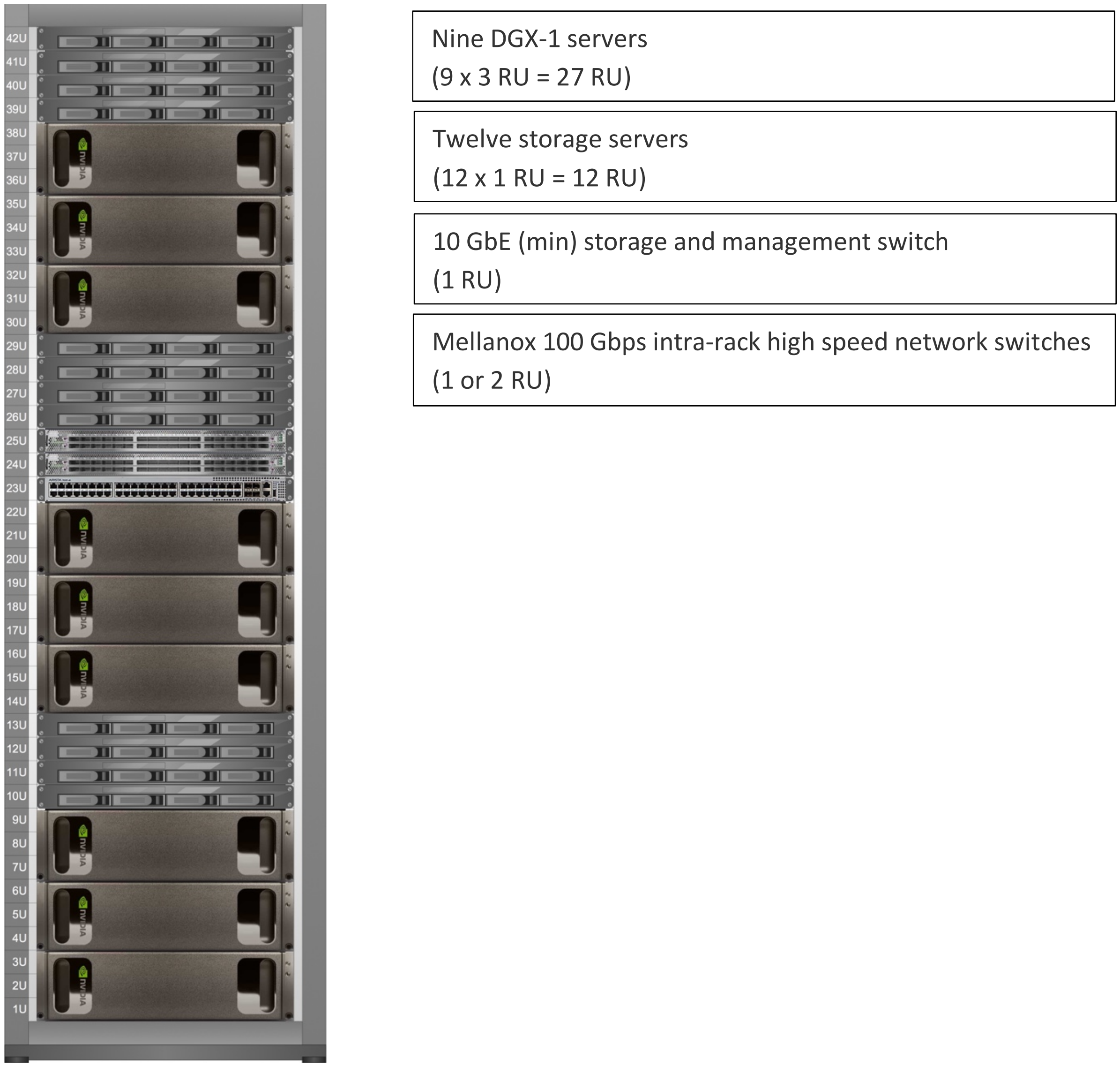


Figure 5. Elevation of a DGX POD

Storage architecture is important for optimized DL training performance. The DGX POD uses a hierarchical design with multiple levels of cache storage using the DGX-1 SSD and additional cache storage servers in the DGX POD. Long-term storage of raw data can be located on a wide variety of storage devices outside of the DGX POD, either on-premises or in public clouds.

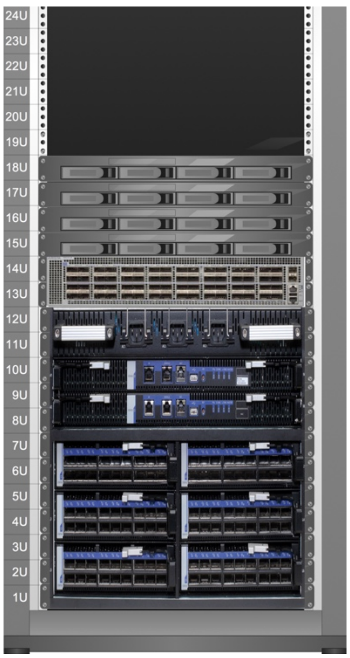
The DGX POD baseline storage architecture consists of standard NFS on the storage servers in conjunction with the local DGX SSD cache. Additional storage performance may be obtained by using the Ceph object-based file system or other caching file system on the storage servers.

© 2018 NVIDIA Corporation.

The DGX POD is also designed to be compatible with a number of third-party storage solutions, see the reference architectures from DDN, NetApp, and Pure Storage for additional information. NVIDIA is also working with other storage vendors who plan to release DGX POD compatible reference architectures.

While based on the DGX-1 server, the DGX POD has been designed in a modular fashion to support the NVIDIA DGX-2™ server which starts shipping in Q3 of 2018. Each of the three compute zones in a DGX POD is designed such that the three DGX-1 servers in the compute zone can be replaced with a single DGX-2 server.

A partial elevation of a DGX POD utility rack is shown in Figure 6.



Login server which allows users to login to the cluster and launch Slurm batch jobs[1].
(1 RU)

Three management servers running Kubernetes server components and other DGX POD management software[2].
(3 x 1 RU = 3 RU)

Optional multi-POD 10 GbE storage and management network switches
(2 RU)

Optional multi-POD clustering using a Mellanox 216 port EDR InfiniBand switch.
(12 RU)

Figure 6. Partial elevation of a DGX POD utility rack

[1] To support many users, the login server should have at two high-end CPUs, at least 1 TB of memory, two links to the 100 Gbps network, and redundant fans and power supplies.
[2] These servers can be lower performance than the login server and can be configured with mid-range CPUs and less memory (128 to 256 GB).

© 2018 NVIDIA Corporation.

4. DGX POD Installation and Management

Deploying a DGX POD is similar to deploying traditional servers and networking in a rack. However, with high-power consumption and corresponding cooling needs, server weight, and multiple networking cables per server, additional care and preparation is needed for a successful deployment. As with all IT equipment installation, it is important to work with the data center facilities team to ensure the DGX POD environmental requirements can be met.

Additional DGX site requirements are detailed in the *NVIDIA DGX Site Preparation Guide* but important items to consider include:

Area	Design Guidelines
Rack	• Supports 3000 lbs of static load • Dimensions of 1200 mm depth x 700 mm width • Structured cabling pathways per TIA 942 standard
Cooling[3]	• Removal of 119,420 BTU/hr • ASHRAE TC 9.9 2015 Thermal Guidelines “Allowable Range”
Power	• North America: A/B power feeds, each three-phase 400V/60A/33.2kW (or three-phase 208V/60A/17.3 kW with additional considerations for redundancy as required) • International: A/B power feeds, each 380/400/415V, 32A, three-phase – 21-23kW each.

[3] Via rack cooling door or data center hot/cold aisle air containment

Table 1. Rack, cooling, and power considerations for a 35 kW DGX POD

Figure 7 shows the server components and networking of the DGX POD. Management servers, login servers, DGX compute servers, and storage communicate over a 1 or 10 Gbps Ethernet network, while login servers, DGX compute servers and optionally storage can also communicate over high-speed 100 Gbps Infiniband or Ethernet. The DGX compute servers shown here are running both Kubernetes and Slurm to handle varying user workloads.

© 2018 NVIDIA Corporation.

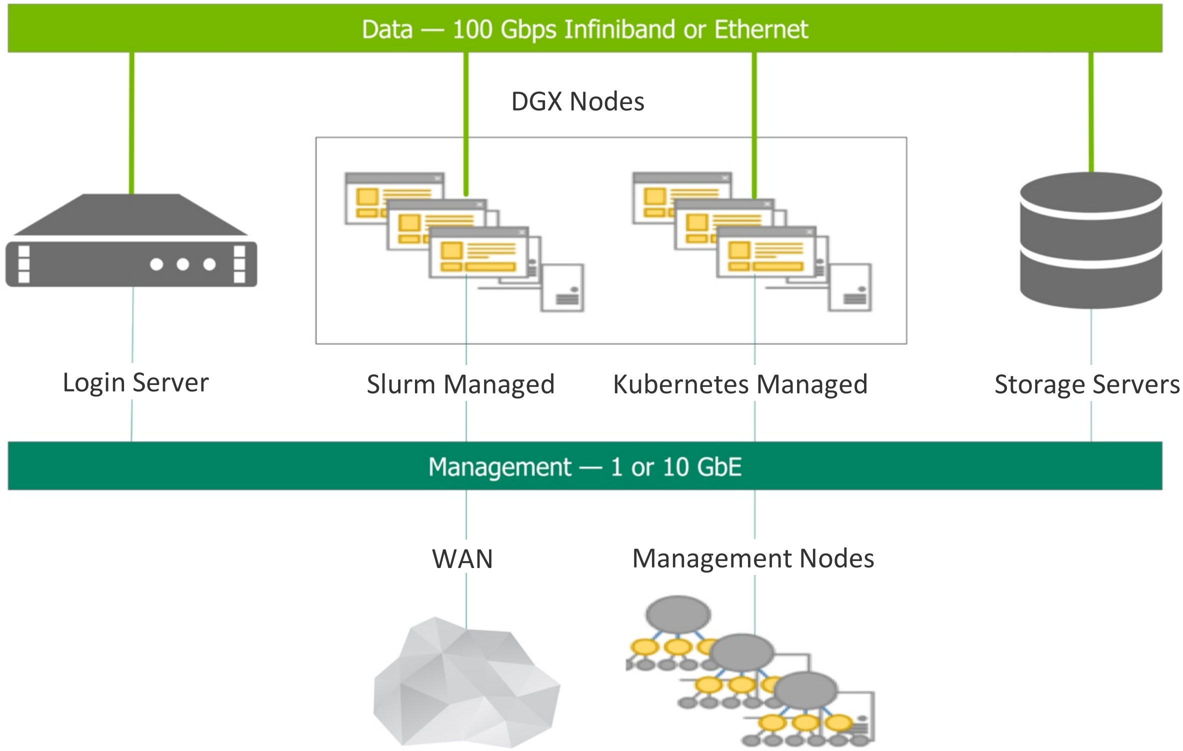


Figure 7. DGX POD networking

Whether deploying single or multiple DGX PODs, it is best to use tools that take care of the management of all of the nodes. Use the NVIDIA AI software stack to provide an end-to-end solution from server OS installation to user job management.

Installing the NVIDIA AI software stack on DGX PODs requires meeting a basic set of hardware and software prerequisites, setting local configuration parameters, and deploying the cluster following step-by-step instructions. Optional installation steps include configuring multiple job schedulers and integrating with enterprise authentication and storage systems.

Once the PODs are powered-on and configured, verify all the components are working correctly. This can be an involved process. During this operational checkout, the following should be verified:

- Compute hardware, networking, and storage are operating as expected
- Power distribution works properly under maximum possible load
- Additional site tests as may be required for the data center

The first step to validate DGX-1 server installation is performed using the diagnostic feature of DCGM. This tests many different GPU functions including memory, PCIe bus, SM units, memory bandwidth, and NVLink. DCGM will draw close to maximum server power during operation and thus is a good stress test of server power and cooling. By running this test simultaneously across all nodes using Kubernetes, it will stress the rack-level power, cooling, and airflow.

The complete DCGM test takes approximately ten minutes to run on each DGX-1. This test should be run back to back for several hours to stress-test the DGX POD and ensure correct completion of each iteration.

To test DL training, the NGC containers have built-in example scripts to check several different families of networks including image classification and language modeling via LSTMs. For example, the NGC TensorFlow Container contains scripts to test the major ImageNet Networks (Resnet-50, Inception, VGG-16, etc.) and allows for scalable server testing with real data or in synthetic mode.

In addition, test any local applications that are planned to be run on the DGX POD. All performance results measured during server installation should be saved and used during routine retesting of the DGX POD to verify performance consistency as servers are modified and updated.

Day-to-day operation and ongoing maintenance of a DGX POD is greatly simplified by the NVIDIA AI software stack. Typical operations such as installing new software, performing server upgrades, and managing scheduler allocations and reservations can all be handled automatically with simple commands.

© 2018 NVIDIA Corporation.

5. Summary

The DGX POD reference architecture provides organizations a blueprint to simplify deployment of GPU computing infrastructure to support large-scale AI software development efforts. A single DGX POD supporting small workgroups of AI developers can be grown into an infrastructure supporting thousands of users. The DGX POD reference architecture is based on the NVIDIA DGX SATURNV AI supercomputer which has 1000 DGX-1 servers and powers autonomous vehicle software and internal AI R&D across NVIDIA research, graphics, HPC, and robotics.

The DGX POD has been design and tested by using specific storage and networking partners. In addition to those mentioned in this paper, NVIDIA is working with additional storage and networking vendors who plan to publish DGX POD-compatible reference architectures using their specific products.

Because most installations of a DGX POD will require small differences such as cable length to integrate into your data center, NVIDIA does not sell DGX POD as a single unit. Work with an authorized NVIDIA Partner Network (NPN) reseller to configure and purchase a DGX POD.

Finally, this white paper is meant to be a high-level overview and is not intended to be a step-by-step installation guide. Customers should work with an NPN provider to customize an installation plan for their organization.

© 2018 NVIDIA Corporation.

Legal Notices and Trademarks

Legal Notices

ALL INFORMATION PROVIDED IN THIS WHITE PAPER, INCLUDING COMMENTARY, OPINION, 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 MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification 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. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification. NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk. NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs. No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. 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 specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, Tesla, NVLink, DGX, DGX-1, and GPUDirect are trademarks or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright © 2018 NVIDIA Corporation. All rights reserved.