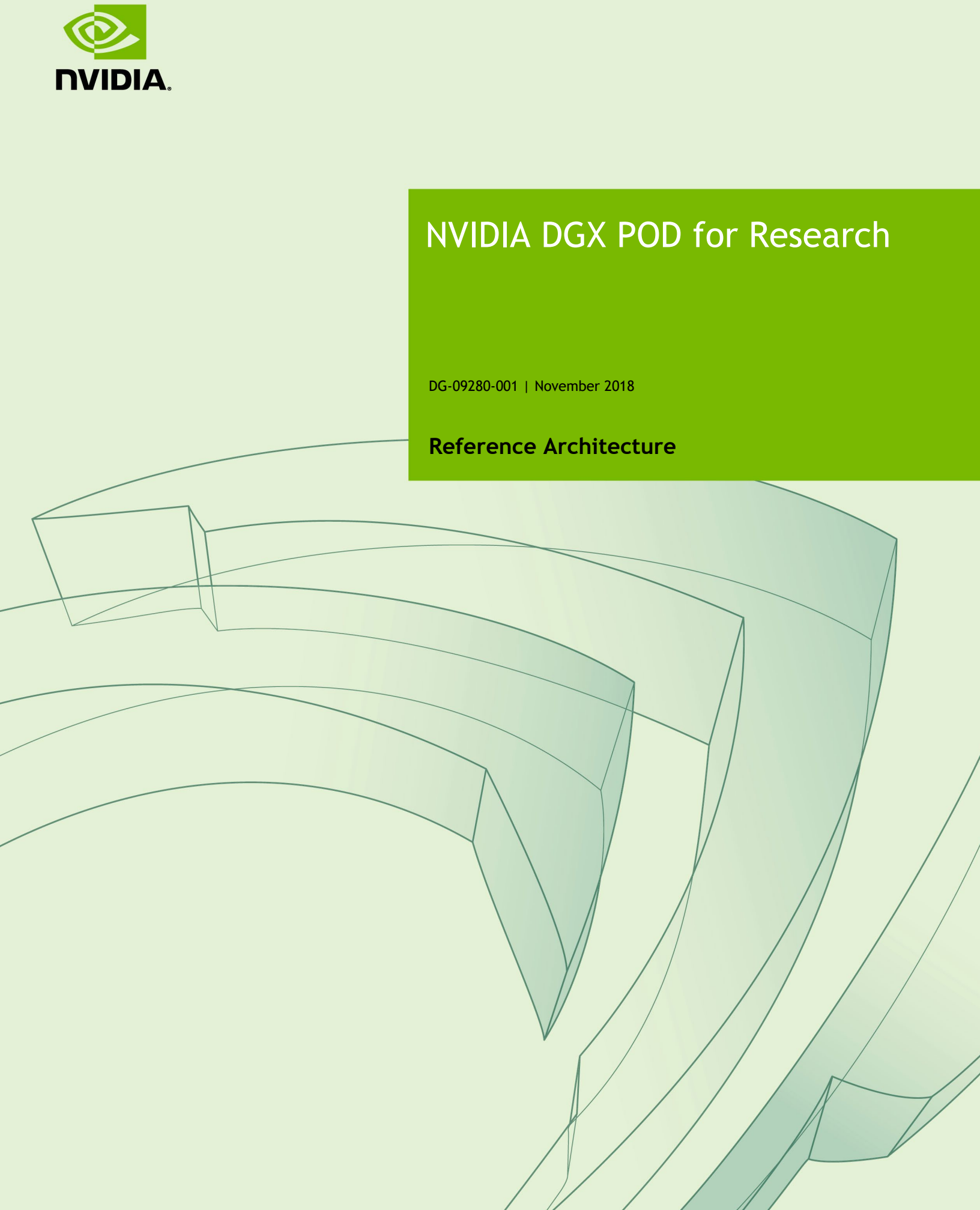


NVIDIA DGX POD for Research

DG-09280-001 | November 2018

Reference Architecture

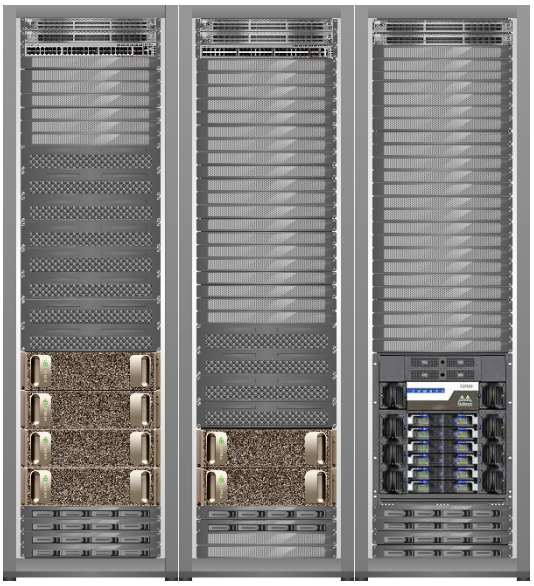
Document Change History

DG-09280-001

Version	Date	Authors	Description of Change
01	2018-11-09	Bradley Palmer, Griffin Lacey, Michael Knox, Jonathan Bentz, and Robert Sohigian	Initial release

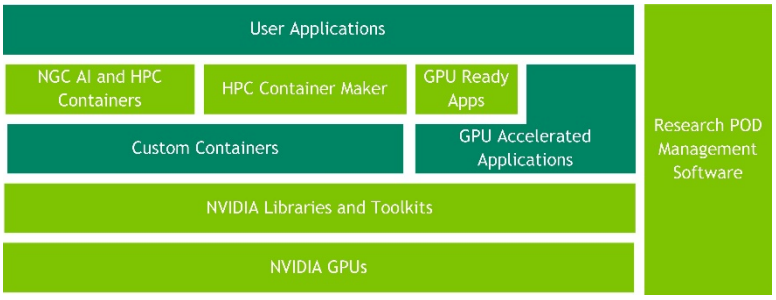
Abstract

The NVIDIA® DGX POD™ for Research (Research POD) reference architecture provides a blueprint for university High Performance Computing (HPC) centers to design a computing resource that is cost effective, designed for the future, and sized to support a wide variety of researchers and applications. The NVIDIA DGX family of supercomputers with NVIDIA Tesla® V100 GPUs allows university HPC centers to lower the cost of computing infrastructure by utilizing a single fused architecture supporting accelerated HPC and artificial intelligence (AI) applications, as well as other emerging fields of research.



Research POD rack elevations

A Research POD includes: one or more racks of DGX and SCX servers, storage, networking, NVIDIA GPU Cloud (NGC) deep learning (DL) and HPC containers, and Research POD management servers.



Research POD software stack

The Research POD reference architecture is based on the NVIDIA DGX SATURNV AI supercomputer which has over 1000 DGX servers and powers internal NVIDIA AI R&D including autonomous vehicle, robotics, graphics, HPC, and other software domains.

Contents

Mixed Workload Sizing

Universities have a long and distinguished history of developing and enabling on-campus HPC for researchers. Traditionally, these researchers have come from the fields of fundamental and applied science such as chemistry, biology, physics, engineering, computer science, CAE, and CAD. The developers of scientific codes were early adopters of accelerated computing using the NVIDIA family of GPUs. A majority of traditional HPC codes now support acceleration on GPUs, to varying degrees of maturity. The adoption of GPUs into HPC codes is transformative as researchers envision computing larger problem sizes, studying wider parameter sets, and leveraging larger datasets.

In the past few years, the success of AI research using GPUs has revolutionized how new scientific research is being performed and has grown the community supported by university HPC centers. This support requires hardware and software resources that can efficiently scale to provide computing cycles for current applications as well as the large suite of ever-evolving software. The current community includes mixed workloads, such as those who are training AI models, running community or commercial HPC applications, developing their own codes, or visualizing large datasets. Planning the infrastructure to accommodate such diversity is a challenging task, and success is not always easy to define. To start thinking about how to approach this problem, consider how researchers spend their time. Despite large diversities in workload, most will follow a similar general workflow, as seen in Figure 1.

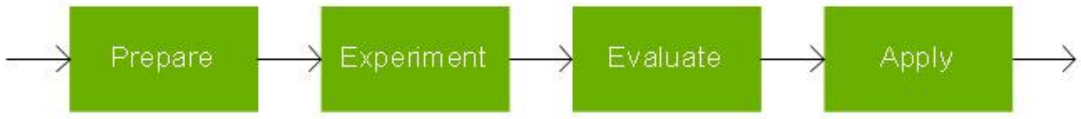


Figure 1. Typical workflow for HPC and AI researchers

Typical AI and HPC workflows might follow these steps:

1. **Prepare**: Pre-work necessary to run experiments.

   AI: Collecting raw data and using tools to pre-process, index, label, and manage that data.

   HPC: Determining the input parameters for an application. This may involve running smaller exploratory simulations.

2. **Experiment**: The scientific procedure.

AI: Using a DL framework to train AI models with a hyperparameter configuration on large cleaned datasets.

HPC: Running HPC applications for analyzing input data or performing a simulation.

3. **Evaluate**: Assessment of the experimental results.

   AI: Testing the trained DL model parameters against a withheld test dataset using an evaluation metric (e.g. classification accuracy, character error rate).

   HPC: Post-processing and can involve a great deal of visualization or other data processing methods

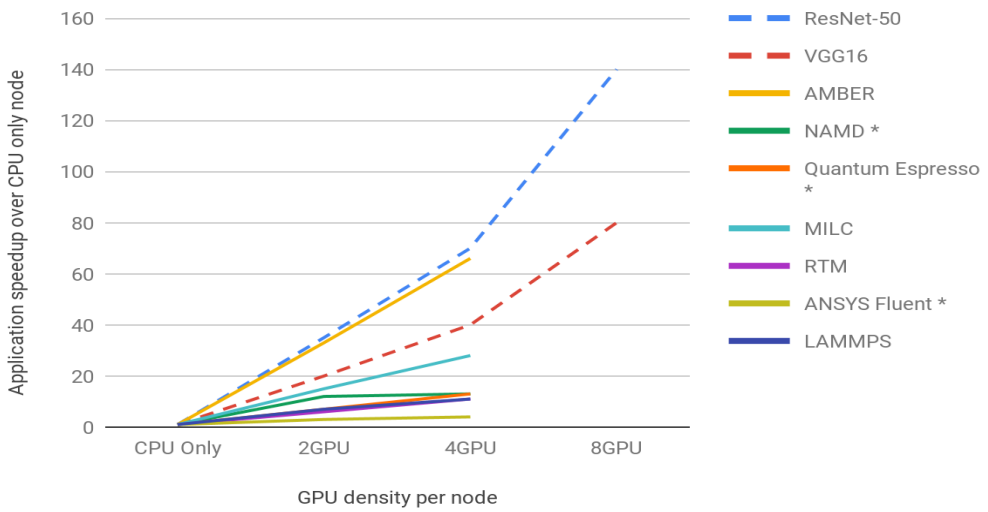
4. **Apply**: How the evaluation can be used to improve future experiments.

   AI: Improving the DL experiments either by iterating through the experiment phase with better hyperparameters, or for production deployment, using an optimization tool (e.g. TensorRT) to improve inference performance.

   HPC: Applying the knowledge from the previously completed jobs to guide their research and determine future simulations

Though the experiment phase is typically where one interacts most with large computing infrastructures, the increasing sizes of datasets and experiments mean most phases of the workflow will benefit from acceleration. This allows researchers to iterate more quickly through their workflows, improving both their productivity and the quality of their solutions. The size of experiments in AI and HPC depend on many factors (e.g. data size, model complexity), making sizing an infrastructure for mixed workloads a complex task. To start sizing an infrastructure to maximize the success of its researchers, the best node configuration is determined from the distribution of workloads.

The distribution of HPC and AI workloads will vary by university. Some researchers are currently or will be highly involved in AI research, while others have a primary focus on traditional HPC. The infrastructure needs to be scalable to the anticipated number of simultaneous users to minimize wait times. Examples of how different AI and HPC applications scale across multiple GPUs are shown in Figure 2. From the slope of the lines in the figure, one observes that AI applications scale better than HPC applications (on average) up to four-GPU nodes. Furthermore, AI applications continue to scale well up to eight-GPU nodes.



Performance reflects NVIDIA NVLink™ SXM2 V100, * denotes PCIe V100 performance.[1]

Figure 2. AI (dashed line) and HPC (solid line) application speedups over CPU nodes.

To better visualize this difference, consider Figure 3 which averages these results over the various AI and HPC applications.

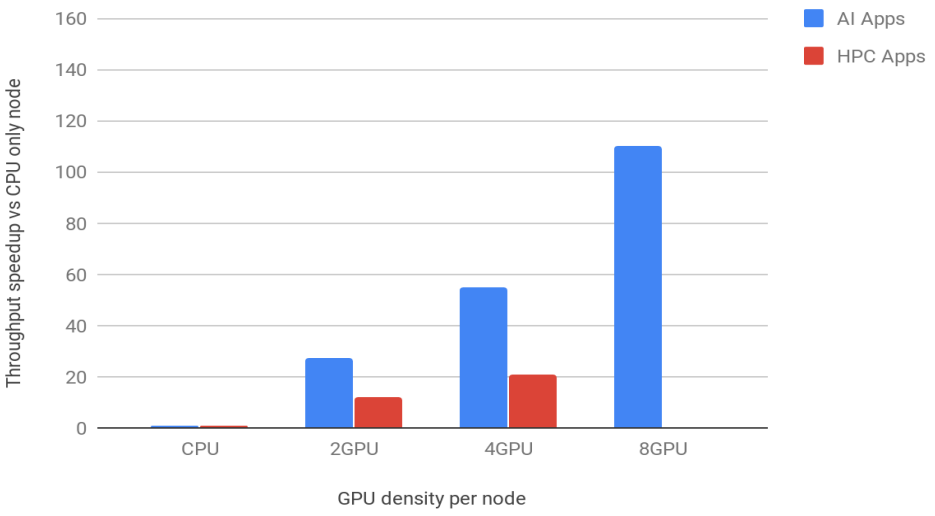


Figure 3. AI and HPC application throughput increases over CPU-only nodes[2]

[1] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/v100-application-performance-guide.pdf

[2] https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/v100-application-performance-guide.pdf

When designing infrastructure for mixed workloads, it's important to determine the right balance of node density. Recommendations for both HPC and AI centric workloads are described in Table 1.

Recommended Sizing Guidelines	Research Computing Focus	
	HPC (SCX-E4[1])	AI (DGX-1)
Server GPU Capacity	4-GPU	8-GPU
User Ratio (Users:Servers)	2:1	1:1

1. http://images.nvidia.com/content/pdf/gpu-accelerated-server-platforms.pdf

Table 1. Infrastructure sizing for teams with HPC or AI workloads

For AI workloads, users are likely to run long, multi-GPU jobs due to their complex models, large datasets, and mature software tools that readily support multi-GPU experiments. Those AI users who use smaller models and datasets will still benefit from dense GPU nodes for running many small concurrent experiments, such as those for hyperparameter optimization. As a result, AI workloads are best suited for high ratios of GPUs to CPUs. Based on experience deploying both internal and external GPU clusters at NVIDIA, the recommendation is that a node with eight GPUs and dual CPU sockets is the most effective configuration for AI workloads. For example, the autonomous vehicle software team at NVIDIA developing NVIDIA DriveNet[3] uses a custom Resnet-18 backbone detection network with a 960x480x3 image size and trains at 480 images per second on an eight-GPU node. This configuration allows for training of 120 epochs with 300k images in 21 hours. The NVIDIA DGX-1™ server, a node with dual CPU sockets and eight Tesla V100 GPUs which utilizes NVLink technology is the recommended server platform for AI.

An eight-GPU architecture is equally well-suited for a few of the most widely used scientific computing applications. For example, Amber 18.6 achieves 7.00 ns/day for the RuBisCo protein simulation on a node with such a configuration. This represents a 700X speed-up over a CPU-only result. While not all HPC codes can take advantage of eight-GPUs and this is an exceptional result, developers are continually porting and optimizing new and existing applications. Currently NVIDIA tracks hundreds of GPU accelerated applications[4], many of which are run in academic and high-performance research computing environments.

[3] https://www.nvidia.com/en-us/self-driving-cars/

[4] https://www.nvidia.com/en-us/data-center/gpu-accelerated-applications/catalog/

For typical HPC workloads the applications tend to scale well up to four GPUs in a single node. The efficiency of this scaling is improving as the more computationally intensive components of the applications are ported to the GPU. There are some HPC applications that either do not yet take advantage of GPUs or are inherently not suited for GPU computing. These non-GPU applications should either be run on CPU-only nodes, or, utilizing job scheduler features, run on the CPUs of the GPU nodes while leaving the GPUs open for execution of GPU accelerated HPC applications. When considering a wide range of HPC applications, nodes with four GPUs and two CPU sockets optimizes for the current scalability in HPC applications as well as future application development. The NVIDIA SCX-E4 server, a node with dual CPU sockets and four Tesla V100 GPUs which also uses NVLink technology, is the recommended server platform for HPC.

Scaling is an important factor to consider when sizing an infrastructure. Trends in both AI and HPC show that scaling up model size and complexity across multi-GPUs improves acceleration and tends to also improve results[5]. While many workloads are moving in this direction researchers tend to only scale their workloads as large as the infrastructure allows. Thus, infrastructure choices may unintentionally limit the size and types of workloads that can be executed on a system. To avoid this limitation, one must be aware of the researchers who push the limits of scaling as well as those who also tend to be the largest consumers of infrastructure resources.

[5] Deep Learning Scaling is Predictable, Empirically

Research POD Software

NVIDIA provides a rich set of software tools, SDKs, and libraries to accelerate HPC and AI applications. The foundation of the software stack consists of the CUDA libraries such as cuSOLVER (sparse and direct solvers) and cuDNN (deep learning and neural networks).

NVIDIA software tools (Figure 4) running on Research PODs provide a high-performance environment for large scale multi-user mixed workload environments. The overall GPU cluster software stack includes cluster management and orchestration tools, workload schedulers, optimized containers from the NGC registry, and additional tools that embody best practices for running containers, particularly with GPUs, libraries and frameworks. The NGC catalog contains many GPU-accelerated containers, including NVIDIA-optimized deep learning software, third-party managed HPC applications, NVIDIA HPC visualization tools, and partner applications.

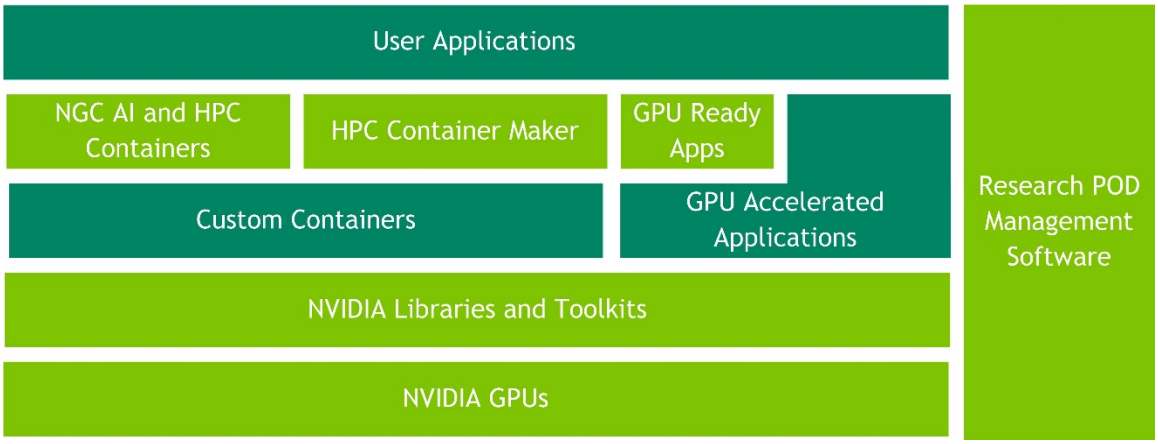


Figure 4. Research POD software stack

The Research POD management software is a set of software tools based on best practices from internal use and customer experience. The focus of these tools is on cluster management and orchestration of the cluster workload. This software, named "DeepOps", is publicly available at:

https://github.com/NVIDIA/deepops

The Research POD software stack provides a fully configurable infrastructure that runs atop standard RHEL/CentOS and Ubuntu Linux operating systems and includes the NVIDIA GPU driver, CUDA Toolkit, GPU management tools, and the NGC Software Containers. These tools can be used to integrate a Research POD with an existing environment or deploy an entirely new cluster from scratch.

The Research 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. Management of the NVIDIA software on the Research 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 GPU compute nodes, configure user accounts, configure external storage connections, install Kubernetes and Slurm schedulers, as well as perform day-to-day maintenance tasks such as new software installation, software updates, and GPU driver upgrades.

When possible, researcher GPU workloads on the Research POD will benefit from utilizing NGC containers. These containers provide 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, including NAMD, GROMACS, RELION, and ParaView.

HPC Container Maker (HPCCM) can be used to containerize applications that are not already available on NGC. HPCCM encapsulates into modular building blocks the best practices of deploying core HPC components and containers to reduce container development effort, minimize image size, and take advantage of image layering. HPCCM recipes provide more portable, higher-level building blocks that separate the concerns of choosing what to include in a container from the low-level details of the container specification file. Using the HPCCM building blocks, it's easy to generate optimized container specification files for Docker and Singularity.

NVIDIA GPU-Ready App quick start guides help you get up and running quickly on GPUs with a simple set of instructions for a wide range of accelerated applications such as Amber, GROMACS, and MILC. Each guide provides recommendations on how to build, install, run, test, and benchmark the application. One can use the GPU-Ready App quick start guides to get the best performance for the applications on NVIDIA GPUs.

Research POD Design

The optimal architecture for a research computing cluster is highly dependent on the characteristics of the workload. The workload distribution between HPC and AI is going to vary by university. Some are or will be highly involved in AI research, while others may have a primary focus on traditional HPC. The architecture needs to be scalable to the anticipated number of simultaneous user applications to minimize wait times.

The Research POD is an optimized data center rack containing compute servers, storage servers, and networking switches to support single and multi-node AI and HPC workloads using NVIDIA software. For this reference architecture, the SCX-E4 server is used for primarily HPC workloads and the DGX-1 server for primarily AI workloads.

The density of a server will depend on the diversity of the workloads. Higher capacity servers (i.e. DGX-1 servers) are ideal for AI workloads. Lower capacity servers (i.e. SCX-E4 servers) are ideal for HPC and balanced workloads. As the number of GPUs per node increases, it is important to consider the host memory and CPU. The CPUs need to execute any non-GPU accelerated portions of the workload and are important for maintaining adequate data flow to and from the GPUs. Nodes should have at least eight CPU cores for each GPU and the host memory on a node should be at least two times the total GPU memory on the system. For example, a compute node with four 32 GB Tesla V100 GPUs should have, at a minimum, 32 CPU cores and 256 GB of memory.

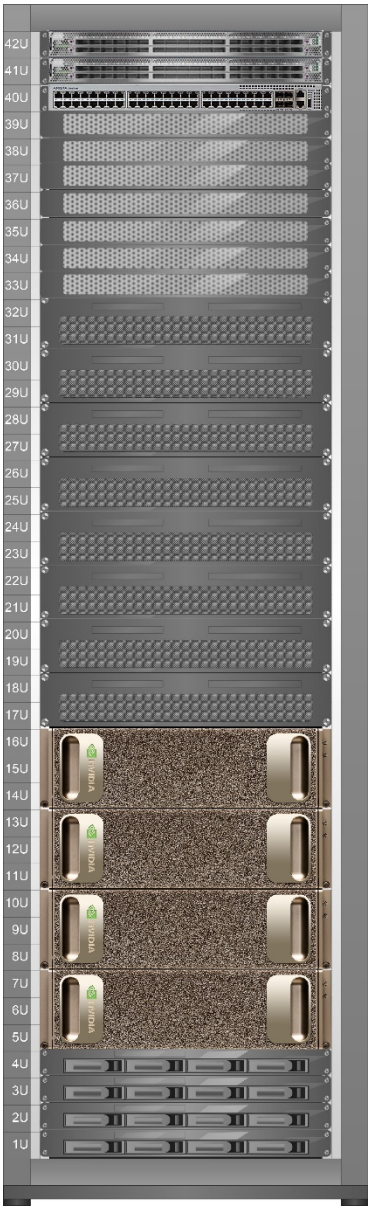
The Research POD is designed to fit within a standard-height 42 RU data center rack. This reference architecture includes login and management servers in the initial rack, but additional racks would not include these servers. Multi-rack configurations of PODs can be defined by an NVIDIA solution architect. The POD architecture can be scaled up or down to fit the expected workload mix with the smallest POD containing one of each server type and serving one AI user and two HPC users.

A primary 10 GbE (minimum) network switch is used to connect all servers in the Research POD. 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, or to provide additional compute server connections. NVIDIA is working with networking vendors who plan to release switch reference designs compatible with the Research POD.

A 36-port Mellanox 100 Gbps switch is configured to provide 100 Gbps EDR InfiniBand connections between the compute servers in the rack. The DGX-1 servers have two InfiniBand connections each which provides excellent scalability for multi-node AI jobs. This configuration provides 18 switch ports for uplink. Note that the Mellanox switch can also be configured in 100 GbE mode for organizations that prefer to use Ethernet networking.

Research POD (35 kW)

In Figure 5, the SCX and HGX compute configurations are used to implement a single Research POD rack for 16 HPC researchers and four AI researchers.



Mellanox 100 Gbps intra-rack high-speed network switches
1 or 2 RU

10 GbE (min) storage and management switch
1 RU

Eight SCX-E4 servers (primarily for HPC workloads)
16 RU

Four DGX-1 servers (primarily for AI workloads)
12 RU

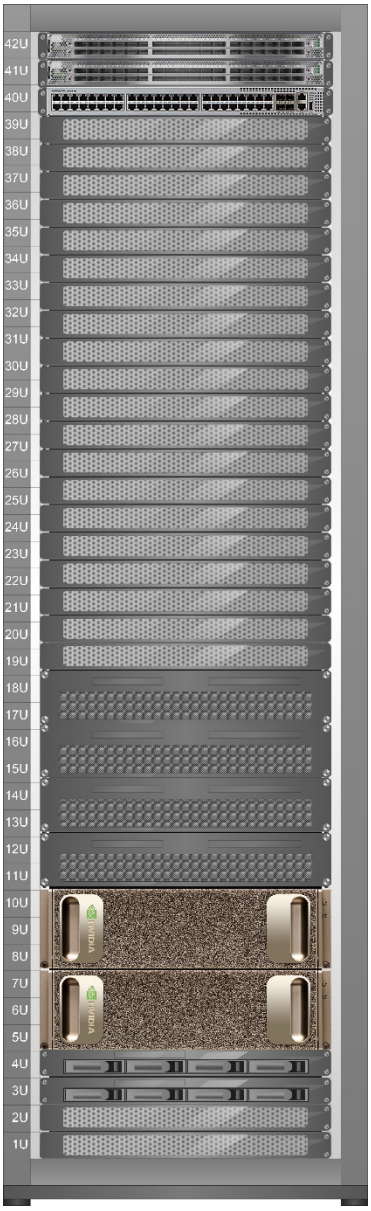
Four storage servers (for DGX-1 server caching)
4 RU

Figure 5. Elevation of a Research POD for 16 HPC researchers and four AI researchers

There are several factors to consider when planning a Research POD deployment to determine if more than one rack is needed per POD. This reference architecture is based on a single 35 kW high-density rack to provide the most efficient use of costly data center floor space and to simplify network cabling. As GPU usage grows, the average power per server and power per rack continues to increase, the servers may need to be distributed to multiple lower-power racks.

Research POD (18 kW)

In Figure 6, an 18kW rack supporting eight HPC researchers and two AI researchers is illustrated.



Mellanox 100 Gbps intra-rack high-speed network switches
1 or 2 RU

10 GbE (min) storage and management switch
1 RU

Four SCX-E4 servers (primarily for HPC workloads)
8 RU

Two DGX-1 servers (primarily for AI workloads)
6 RU

Four storage servers (for DGX-1 server caching)
2 RU

Figure 6. Elevation of a Research POD for eight HPC researchers and two AI researchers

Research POD Utility Rack

For larger configurations, login and management servers as well as management and clustering switches can be housed in a utility rack.

A partial elevation of a utility rack is shown in Figure 7.

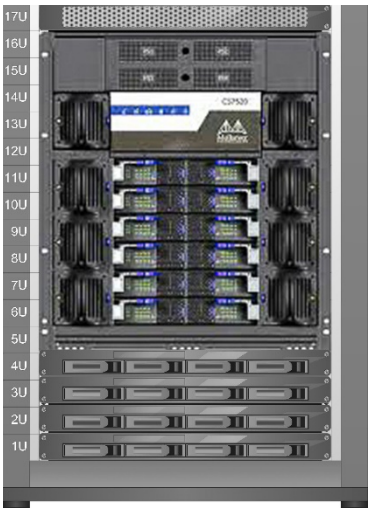


Figure 7. Partial elevation of a Research POD utility rack

[1] To support many users, the login server should have 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).

Research POD Installation and Management

Deploying a Research POD is like 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 POD environmental requirements (Table 2) can be met.

Area	Design Guidelines	
	Research POD (35 kW) rack	**Research POD (18 kW) rack**
Rack	Supports 3000 lbs. of static load	Supports 1200 lbs. of static load
	•Dimensions of 1200 mm depth x 700 mm width •Structured cabling pathways per TIA 942 standard	
Cooling[1]	Removal of 119,420 BTU/hr	Removal of 59,030 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.	• North America: A/B power feeds, each three-phase 208V/60A/17.3 kW • International: A/B power feeds, each 380/400/415V, 32A, three-phase –21-23kW each.
1. Via rack cooling door or data center hot/cold aisle air containment		

Table 2. Rack, cooling, and power considerations for Research POD racks

Figure 10 illustrates the network topology for the Research POD. This network architecture allows for a fully non-blocking InfiniBand connection to a network of additional POD racks. The cache storage for the DGX-1 servers is InfiniBand-connected. The 10 GbE network provides connections to additional storage servers and POD racks for management and storage I/O. The external connections to the management and login servers is accomplished using a VLAN.

Management servers, login servers, compute servers, and storage communicate over a 1 or 10 Gbps Ethernet network, while login servers, compute servers and optionally storage, can also communicate over high-speed 100 Gbps InfiniBand or Ethernet. The compute servers shown here are running both Kubernetes and Slurm to handle varying user workloads.

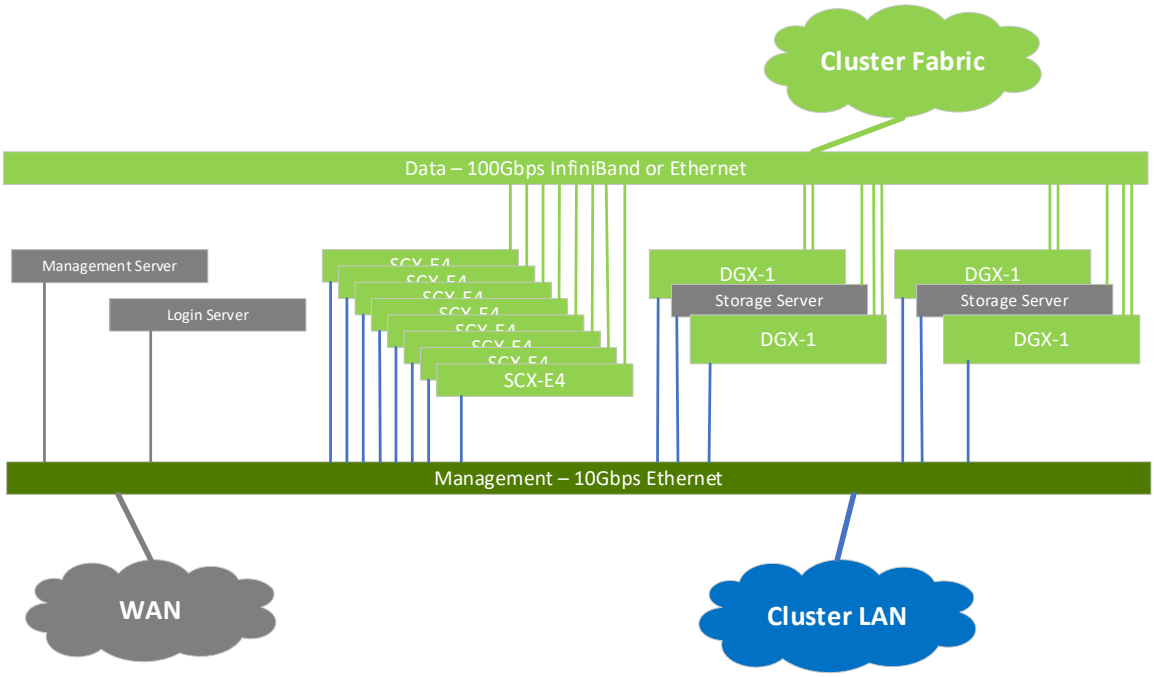


Figure 8. Network topology for a Research POD

For organizations that want to utilize multiple Research PODs to run cluster-wide jobs, additional switches or a core InfiniBand switch can be configured in a utility rack to complete a fat tree topology either with or without a blocking factor. Multi-rack configurations of PODs can be defined by an NVIDIA solution architect

Storage architecture is important for optimized performance, especially for AI training. Long-term storage of raw data can be located on a wide variety of storage devices outside of the Research POD, either on-premises or in public or private clouds. Many workloads, in particular AI, benefit from having local caching available on the compute nodes. This can be accomplished using local SSDs on the compute nodes which are made accessible for user managed caching. Alternatively, when the shared persistent storage is NFS based, client-side caching can be accomplished with per node SSDs and

the NFS cachefilesd service. I/O patterns for AI are very heavily read oriented, so client read caching proves to be very effective.

HPC workloads tend to use a high performance clustered filesystem which is accessed via the InfiniBand fabric. While not as common, some HPC application can utilize local node disks for checkpoint restarts and intermediate data storage.

The Research POD baseline storage architecture consists of in-rack NFS storage servers used in conjunction with the local DGX SSD cache. Additional storage performance may be obtained by using one of the distributed filesystems listed at this link:

https://docs.nvidia.com/deeplearning/dgx/bp-dgx/index.html#storage_parallel

The Research POD is also designed to be compatible with several third-party storage solutions, many of which are documented at this link:

https://www.nvidia.com/en-us/data-center/dgx-reference-architecture/

Summary

The described Research POD reference architecture is intended as a starting point for designing an effective GPU accelerated computing infrastructure which can serve a wide variety of researchers in university HPC and AI. This architecture is intended to scale as the demand for GPU computing resources increases.

When designing new computing infrastructure, it is critically important to understand current needs as well as future trends in application development and emerging fields of science. As described, the overall trend in both AI and HPC workloads is a demand for increasing amounts of input data, larger AI network models, larger HPC problems sizes, and generally more computation per job. NVIDIA's recommendation of SCX-E4 nodes for HPC and DGX-1 nodes for AI balances today's workloads with a view towards the more demanding workloads of the future.

The Research POD reference architecture is designed for integration into typical university data centers. However, the actual integration will require customization and as such, NVIDIA does not sell the Research POD as a single unit. Instead, it is recommended to work with an authorized NVIDIA Partner Network (NPN) reseller to configure and purchase a Research 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.

Legal Notices and Trademarks

Notices

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation (“NVIDIA”) does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. 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 publication supersedes and replaces all other specifications for the product that may have been previously supplied.

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 regarding 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 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.

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. 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 NVIDIA terms and conditions of sale for the product.

Trademarks

DGX, Tesla, NVLink, NVIDIA and the NVIDIA logo 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.

Copyright