White Paper

Considerations When Building a Multi-Node Environment Based on NVIDIA DGX-1

November 2017 - Initial Release
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of Contents</td>
<td>2</td>
</tr>
<tr>
<td>Preface</td>
<td>3</td>
</tr>
<tr>
<td>Benefits for IT Administrators</td>
<td>3</td>
</tr>
<tr>
<td>Considerations for Multi-Node Scale</td>
<td>4</td>
</tr>
<tr>
<td><strong>Useful Insights into Multi-node Scaling</strong></td>
<td>6</td>
</tr>
<tr>
<td><strong>Scaling a DGX-1 Multi-Node Environment</strong></td>
<td>7</td>
</tr>
<tr>
<td>Scalability Bottlenecks Impacting Traditional Architectures</td>
<td>7</td>
</tr>
<tr>
<td><strong>Multi-Node Design Guidance</strong></td>
<td>8</td>
</tr>
<tr>
<td>SATURNV: Lessons in Building a DGX-1 Multi-Node Architecture</td>
<td>8</td>
</tr>
<tr>
<td>InfiniBand network topology</td>
<td>8</td>
</tr>
<tr>
<td>Storage</td>
<td>10</td>
</tr>
<tr>
<td>Reference Configurations</td>
<td>12</td>
</tr>
<tr>
<td>Planning for Scale - How to Get There</td>
<td>12</td>
</tr>
<tr>
<td>Small-sized Multi-Node Configuration: Up to 12 Nodes</td>
<td>13</td>
</tr>
<tr>
<td>Medium-sized Multi-Node Configuration: Up to 36 Nodes (1 compute “POD”)</td>
<td>14</td>
</tr>
<tr>
<td>Large-sized Multi-Node Configuration: Up to 144 Nodes (4 compute PODs)</td>
<td>15</td>
</tr>
<tr>
<td>Data Center Power and Cooling Considerations</td>
<td>16</td>
</tr>
<tr>
<td><strong>DGX Management Considerations</strong></td>
<td>16</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>18</td>
</tr>
<tr>
<td><strong>Appendix</strong></td>
<td>19</td>
</tr>
<tr>
<td>DGX-1 Overview</td>
<td>19</td>
</tr>
<tr>
<td>System Architecture</td>
<td>19</td>
</tr>
<tr>
<td><strong>LEGAL NOTICE</strong></td>
<td>20</td>
</tr>
</tbody>
</table>
Preface

As deep learning training and other computationally-intensive workloads become increasingly prevalent in your data center, the need to scale-out GPU-accelerated computing infrastructure becomes an important challenge that the IT administrator must address. Deep learning, for example is a fundamentally different workload than traditional enterprise applications running on x86-based CPU servers, necessitating consideration of specific networking, storage, and infrastructure management approaches that have been proven to enable improved scalability, performance, and cost-effective manageability. In this document we’ll explore these considerations using deep neural networks (DNN) as our example of a GPU-accelerated workload for which DGX-1 can be deployed in a multi-node cluster, and how to approach it.

Benefits for IT Administrators

NVIDIA DGX-1 offers your data scientists unprecedented GPU-powered capacity to train DNNs in a fraction of the time taken by traditional CPU servers, using a fraction of the infrastructure (nodes, racks, power, cables, switching, cooling). This shift to dense GPU-compute nodes represents a dramatic savings for you, the IT administrator or architect seeking to integrate deep learning workloads within your data center, in terms of both CapEx and and IT administrative OpEx.
Considerations for Multi-Node Scale

Many organizations are adopting deep learning to improve competitiveness. As in most cases small projects that are successful will grow. Applications grow to handle larger volumes of data, solve more complicated questions, and require faster time to solution. When this growth is required, applications need access to more compute resources. Where a single GPU or single server may have been adequate for smaller problems, larger problems need to access more resources than a single system can provide.

To scale an application across multiple nodes, the application needs to have multiple processes running across multiple nodes coordinated to act as one. Scaling applications across multiple systems pose unique challenges uncommon to traditional data centers. Bottlenecks in any component of the system will impact an application’s ability to scale effectively. To ensure maximum performance, systems need to designed to minimize the bottlenecks.

Figure 1 below is a large configuration based on 128 NVIDIA DGX-1 systems. The NVIDIA DGX-1 is a 3U 8 GPU system highly optimized for deep learning with 2 high-end Intel Xeon CPUs, 8 NVIDIA V100 GPUs and 7TB of local SSD RAID storage. The 128 node configuration uses high speed InfiniBand and 10Gb/s Ethernet networking to interconnect the DGX-1 systems for compute and storage. The topology of these networks is optimized for optimum scaling of deep learning type workloads.

Figure 1. NVIDIA DGX-1 Deep Learning Data Center Reference Architecture (144 servers)
Implementation of large configurations like the one above require focus in many areas including:

- What bottle-necks might impact multi-node scalability?
- How does power density at the rack and data center change with GPU computing?
- What does a DGX-1 configured in a multi-node environment look like?
- What are the best practices for server networking?
- What are the best practices for storage?
- What are some multi-node reference architectures that can guide cluster design?
- What tools can one use for GPU server management and scheduling?
- How do traditional approaches to server scaling impact deep learning workload performance?

In this document, we’ll explore these questions, and provide guidance to IT architects, administrators and managers who need to support their data science teams with a practical, cost-effective approach for scaling deep learning using NVIDIA DGX-1 in their data center. Many of the concepts and approaches discussed here represent best practices following when scaling similar workloads such as high-performance computing, with some important distinctions that make deep learning unique.

This document assumes the reader is familiar with the core concepts and technologies employed in NVIDIA DGX-1, and is familiar with the basic principles of deep learning. For technical information on NVIDIA DGX-1 please refer to the [DGX-1 Architecture Whitepaper](http://www.nvidia.com/object/dgx-1-system-architecture-whitepaper.html). To learn more about the fundamentals of deep learning, please consult the [NVIDIA Deep Learning Institute](https://www.nvidia.com/en-us/deep-learning-ai/education/).

It is also recommended that the reader familiarize themselves with the broader considerations at-play when designing a data center that’s optimized for GPU-accelerated workloads, like deep learning, high-performance computing and more. The whitepaper “*Considerations for Scaling GPU-Ready Data Centers*” is a valuable pre-cursor to this whitepaper, and can help data center architects redesign their facilities with a focus on power, cooling, networking, and storage. Additionally useful is the “*Best Practices for DGX Frameworks and Containers User Guide*”. It provides recommendations to help administrators and users work with Docker, extend frameworks, and administer and manage DGX products.

Finally, this document focuses primarily on the hardware architecture needed to support multi-node scaling, however software plays an equally important role. NVIDIA recommends using containerized software applications to simplify workload manageability and to keep a consistent environment across all nodes. NVIDIA provides DGX system customers with many performance-optimized containers for popular DL frameworks that incorporate the software-engineering expertise of NVIDIA’s own teams and the framework developers with

---

whom they collaborate. Please refer to the DGX-1 Architecture Whitepaper and the “DGX Deep Learning Software Brief” to learn more.

**Useful Insights into Multi-node Scaling**

NVIDIA engineers and solution architects have engaged a wide spectrum of clientele, focused on developing best of breed architectures for multi-node scalability of DGX-1 infrastructure. Over the course of these engagements, we’ve assembled a number of important insights gained from the field as well as proven solution design approaches that now form our recommended best practices.

As an implementer of a multi-node DGX-1 environment, the following insights deserve careful consideration, and may be useful in helping you recognize the important differences between traditional server workloads, and GPU-accelerated deep learning.

**Overall cluster**
Traditional server workloads do not provide a useful comparison for deep learning. As we’ll explore in this document, significant differences in networking and storage, for example, necessitate clean-slate thinking when it comes to architecture and capacity planning. High performance computing (HPC) is the closest “cousin” to deep learning from a resource demand perspective, and having access to HPC expertise can aid in your design process. Be aware that even with HPC, the similarities are limited.

**Rack design**
Deep learning workload tends to drive compute nodes to close to the operational limits, unlike other workload where more operating headroom is often assumed. As such proper airflow is crucial to maximizing per-node performance, as well as consistent performance across a multi-node cluster.

**Networking**
As with HPC, InfiniBand is considered the networking medium of choice offering the bandwidth and low latency needed for deep learning. Designers should carefully evaluate the switching infrastructure that connects nodes, and based on expected performance and scale goals, allow for using the maximum number of per-node InfiniBand connections possible, in order to avoid future bottlenecks.

**Storage**
Deep learning workloads drive the need for huge read bandwidth capacity, as well as repeated hits on the same data, many times (tens to hundreds, or even more) during a training run. DGX-1 read cache becomes especially critical to performance as a result. Additionally, the data in question can be very large, ranging from tens of thousands of objects to millions, occupying

---

terabyte levels of storage footprint. Finally, the dataset size can have a large variance, depending on the type of input data and the amount of data being leveraged.

**Facilities**

Multi-node clusters of DGX-1 architected for deep learning, drive increased power density in the data center, and will necessitate higher per-rack wattage your design assumption. The upside of this platform is that its power efficiency (FLOPS/watt) is dramatically higher than traditional CPU-based infrastructure, enabling planners to consolidate data center footprint while increasing computational capacity. Today’s GPU-accelerated data center will comfortably operate at near-maximum power specification levels, as is typical with similar HPC workload.

**Software**

Ensuring scalability goes well beyond capable hardware, and necessitates an intelligent software stack that includes components that are cluster-aware. The NVIDIA Communications Collective Library (NCCL2) provides fast routines for multi-GPU, multi-node acceleration, with automatic topology detection, and optimizations for today’s popular deep learning frameworks. Additionally, it should be noted that not all frameworks scale as efficiently as others. Selecting the right frameworks and versions is important for realizing the true scalability of a multi-node DGX-1 system. Lastly, the right job scheduling software must be selected, and must be topology-aware in order to help balance the benefits of multi-node scaling versus confining job execution to specific infrastructure pods, or smaller clusters.

**Scaling a DGX-1 Multi-Node Environment**

**Scalability Bottlenecks Impacting Traditional Architectures**

As developers seek to increase application performance through parallelism, they quickly find that there are many challenges to achieving good scalability. First, the cost of communication between threads of an application result in a significant drop-off in training performance as more GPUs are added to a node. In standard servers, communication between GPUs is limited by the PCIe bus, which can only move data at approximately 1% of the rate at which data can be transferred within GPU memory. Second, communicating between GPUs on different servers is impacted by typical data center networking, ie: 10Gb/s Ethernet. In addition, making uninformed assumptions about communication patterns can result in unnecessary traffic between GPUs on the same and different systems.

The DGX-1 server was designed to overcome these bottlenecks and maximize application scalability and performance. First, the DGX-1 server was the first platform to adopt NVLink, an ultra-high bandwidth path between GPUs within the node. NVLink provides 300GB/s of bandwidth within the node, allowing for better application scaling than using the PCIe bus.
Second, DGX-1 is equipped with four Infiniband ports on each system, providing 100 GB/s of bandwidth between systems.

In addition, the NVIDIA software technologies GPUDirect RDMA\(^5\) and NCCL\(^6\) provide optimized methods of communication between GPUs. GPUDirect RDMA provides a method for communicating between GPUs on different systems without the need to involve CPU or system memory at either end of the transfer. NCCL2 provides optimized collective data movement operations, with automatic topology detection to scale deep learning training across multiple GPUs and nodes. This enables programmers to develop code that can automatically take advantage of multi-GPU and multi-node training in the most efficient way possible. Within a node, NCCL2 will move data between GPUs using NVLink to maximize bandwidth. Between nodes, it will leverage GPUDirect RDMA and all of the InfiniBand links to provide for the fastest possible inter-node data transfers.

**Multi-Node Design Guidance**

**SATURNV: Lessons in Building a DGX-1 Multi-Node Architecture**

SATURNV is NVIDIA’s first AI supercomputer specifically designed to bring the world closer to the era of exascale computing. It’s architecture is based on DGX-1 as a fundamental building block, and integrates a cluster of 125 DGX-1 nodes, hosting 1,000 NVIDIA Tesla GPUs, delivering the equivalent computing power of over 30,000 x86 servers. The experience gained by NVIDIA’s deep learning and IT teams in constructing SATURNV, have created a valuable foundation of architectural guidance that can be employed in designing scale-out DGX-1 architectures, even those much smaller by comparison.

**InfiniBand network topology**

A fundamental requirement of SATURNV is the ability to ingest data as fast as possible and enable passing of data between nodes stretched across the cluster. Building a scale-out cluster for deep learning can be considered similar (from a design perspective) to high-performance computing (HPC) workloads. InfiniBand, which is considered the defacto standard in HPC networking due to it’s high bandwidth, incredibly low latency and low CPU overhead, is therefore the most technically suitable option for networking a DNN deployed over a multi-node cluster. In order to enable strong scaling across the SATURNV cluster, a high performance two tier InfiniBand switching architecture was developed. This included an InfiniBand leaf (level 1) switch accommodating up to 4 DGX nodes per switch, each with up to 4 InfiniBand EDR connections to the switch. These 16 links are fully connected via 16 uplinks to the InfiniBand top level (level 2) switch (sometimes referred to as the root or spine switch). See Figure 2

\(^5\)https://devblogs.nvidia.com/parallelforall/benchmarking-gpudirect-rdma-on-modern-server-platforms/
\(^6\)https://developer.nvidia.com/nccl
Figure 2: InfiniBand network topology

The premise of this fat-tree topology is to enable any-to-any interconnectivity without interference or added latency between nodes. While this topology offers full bi-section bandwidth with minimum contention between nodes, and unconstrained bandwidth from a node up to the level 2 switch, in many bases a 2:1 oversubscription with half the links between level 1 and level 2 switches provides a good balance of cost versus performance. In any case, it is recommended to have a minimum of 2 connected ports per DGX node (with at least one InfiniBand connection per CPU), but 4 connections are preferred to enable seamless scale with reduced loss in relative DNN performance as workload demand increases (see figure below).

Reducing the number of InfiniBand ports used in each system will result in lower relative performance (ex: anywhere from 10 to 40% loss when reducing from four ports to one as seen in testing on various workloads including High-Performance Linpack and Microsoft Cognitive Toolkit). The effect of connected port count on scaling is shown in Figure 3
Figure 3: The impact of connected IB port count on system scaling

Another consideration that impacts performance is the use of GPUDirect RDMA\(^7\). This technology provides the ability to transfer data directly between GPUs in two different systems across the InfiniBand network without involving the CPU or system memory. This maximizes bandwidth between cluster nodes, while minimizing CPU overhead, for reduced latency and significantly increased performance of multi-node systems. Enabling RDMA requires use of at least 2 IB connections to all nodes in the configuration.

Storage

DNNs involve very large datasets and storage transactions that are read-dominated at the beginning of each epoch (an epoch is defined as one complete pass-through of the dataset, inclusive of multiple iterations of model parameter updates). This differs from typical HPC applications which are write-intensive. Deep learning training is usually static, involving large groups of random reads, accessed repeatedly, since the same data is used for training over and over. To support this storage access profile, the DGX-1 includes four SSDs in a RAID 0 configuration for a total of 7TB local SSD RAID set-up as a local NFS read cache, that enables accelerated data streaming of repeated read operations.

\(^7\) https://developer.nvidia.com/gpudirect
**NFS vs Parallel File Systems**

It is recommended that an NFS appliance is implemented for long term data storage. NFS offers a cost-effective, simple distributed file system approach for network-attached storage. The anticipated read transaction volume and the number of expected nodes in the environment will influence the selection of storage architecture. In the section that follows, we’ll explore three scenarios that offer various tiers of scalability, and how this impacts storage. Parallel file systems such as IBM Spectrum Scale introduce an increased level of complexity, and while the capability has been proven for over a decade in HPC computing, it is still considered a nascent technology for a deep learning workload. These solutions provide the benefit of performance traditionally only found in an enterprise-grade SAN infrastructure or those employed in high-performance computing environments. Parallel file systems offer higher-speed access to large volumes of data that are distributed (striped) across multiple servers and disks, and read in parallel, supporting multiple concurrent users and jobs, with greater speed than traditional NFS. As such it is recommended to consider parallel file systems where the cluster node count is expected to be large, for example 32, 64 or more nodes.

In multi-node environments, it’s recommended to copy all data or cache all data local to each DGX-1 node at the beginning of a job since DL workloads access the same data many times during a long training job. The DGX-1’s SSD NFS cache greatly reduces the stress on central shared storage. Since these SSDs are intended for caching, you must set up your own NFS drives for long term data storage. Consult the DGX-1 [user guide](http://images.nvidia.com/content/technologies/deep-learning/pdf/DGX-1-UserGuide.pdf) for instructions describing how to mount the NFS onto the DGX-1, and how to cache NFS reads using the DGX-1 SSDs.

<table>
<thead>
<tr>
<th>DEEP LEARNING USE CASE</th>
<th>DGX-1 CACHING CAPABILITY</th>
<th>DGX-1 NETWORK TYPE RECOMMENDED</th>
<th>NETWORK FILE SYSTEM OPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL, 256x256 images</td>
<td>63 million images</td>
<td>10 GbE</td>
<td>NFS or storage with good small file support</td>
</tr>
<tr>
<td>DL, 1080p images</td>
<td>13 million images</td>
<td>10 GbE, InfiniBand</td>
<td>High-end NFS, HPC filesystem or storage with fast streaming performance</td>
</tr>
<tr>
<td>DL, 4K images</td>
<td>5 million images</td>
<td>10 GbE, InfiniBand</td>
<td>HPC Filesystem, high-end NFS or storage with fast streaming performance capable of &gt; 2 GB/s per node</td>
</tr>
<tr>
<td>DL, uncompressed images, 1080p</td>
<td>1 million images</td>
<td>10 GbE, InfiniBand</td>
<td>HPC Filesystem, high-end NFS or storage with fast streaming performance capable of &gt; 2 GB/s per node</td>
</tr>
<tr>
<td>DL, Datasets too large to cache</td>
<td>N/A</td>
<td>10 GbE, InfiniBand</td>
<td>Same as above, aggregate storage performance must scale to meet</td>
</tr>
</tbody>
</table>

Table 1 shows general guidelines for choice of storage configuration based on workload / use case. In general it’s recommended to review these suggestions with your storage architect to determine the configuration best suited to your environment.
Reference Configurations

Planning for Scale - How to Get There

IT teams should work in consultation with the data science team, to define performance and capability requirements of the environment being proposed, including workload and scaling objectives. It is important to note that DGX-1 offers an incredibly dense compute footprint, delivering almost one petaflop of deep learning performance (FP16) in just 3 rack units of space. Ensuring the maximum effective utilization of each DGX-1 node, and by extension, the maximum return on your investment, relies on careful consideration of the data center characteristics at play in a DGX-1 deployment, including air flow, cooling and power. This document offers considerations and approaches that should be evaluated for varying tiers of deployment size.

For example, organizations that envision an end-game that encompasses fewer than 13 nodes, should consider the "small" system configuration detailed below, using it as a building block. Organizations that can envision the need for over 12 nodes, should begin with the DGX performance-optimized deployment construct, or "POD" (detailed below), which allows for large scale using a repeatable model based on laying down the right switching infrastructure from day one.

From a storage perspective, an NFS appliance with 10GbE network interfaces may be sufficient for the "small" multi-node configuration. For larger configurations with multiple PODs, it may be necessary to implement a parallel filesystem, as described earlier. Additionally, large configurations should assume the use of InfiniBand storage interfaces to ensure sufficient throughput with low latency.

Let's now explore three possible configuration approaches that can be employed to enable DGX scale. Each has been formulated with the goal of ensuring unconstrained DNN training performance for environments expected to scale to a prescribed number of nodes (up to 12, 36, and 144). As part of these configurations, we’ll include definition of a “POD” that provides a basic modular unit of infrastructure that can be scaled to support higher node counts. These suggestions should be measured against the specific goals and cost-objectives an organization must consider, and tailored as necessary.
Small-sized Multi-Node Configuration: Up to 12 Nodes

The “small” configuration (Figure 4) has been designed to enable seamless scalability across a cluster of nodes expected to grow to a maximum of 12 nodes. It allows for full bi-section bandwidth for each group of 6 nodes via 4 InfiniBand EDR links per node and a 2:1 bandwidth ratio between the two sets of 6 nodes. The small configuration fits entirely within two data center racks. However, if the available data power cannot readily support the prescribed 19.2 kW per rack, this configuration can be split into 3 or 4 racks if necessary. Additionally, a “mini-small” configuration can be built based on a configuration leveraging 6 nodes with 1 InfiniBand switch. The maximum node count assumed for a single InfiniBand switch is assumed to be 9, based on a 36 port device. Beyond 9 nodes, a second InfiniBand switch is required. Note that this configuration does not support IB network redundancy and assumes that network failures will be handled at the application level.

- 12 node cluster, full interconnect in groups of 6
  - 4 InfiniBand EDR links per node
  - 6 DGX-1 nodes per rack (~19.2 kW per rack)

**Bill of Materials**

- Two racks
- 12 DGX-1 nodes
- Compute: 2 Mellanox SB7700 36-port EDR IB Compute leaf switches (4 ports per DGX-1)
- Storage: 1 Ethernet 24 port 10Gb Ethernet Storage switch (2 ports per DGX-1) + additional uplinks for storage connection
- Management: 1 Ethernet 24 port 1Gb Ethernet switch (1 port per DGX-1)

Figure 4: Small-sized Multi-Node Configuration
Medium-sized Multi-Node Configuration: Up to 36 Nodes (1 compute “POD”)

The “medium” configuration (Figure 5) defines a DGX “POD” of up to 36 nodes. PODs can be replicated to achieve greater scale, as shown in the “large” configuration that follows. The DGX POD is a scalable cluster configured for full bisection network bandwidth performance, servicing 36 nodes installed across 6 data center racks. It assumes 19.2kW per rack, with 6 nodes in each rack. Similarly, it utilizes 4 InfiniBand EDR links per node, but employs a larger director switch, with allowance for addition of more PODs via unused switch ports.

![Diagram of DGX-1 compute POD with InfiniBand 216 port Director Switch and Ethernet connections](image)

**Figure 5:** Medium-sized multi-node configuration

36 node cluster fully interconnected
- 4 InfiniBand EDR links per node
- 6 racks with 6 DGX-1 nodes per rack (~19.2kW racks)

**Bill of Materials**
- 6 racks
- 36 DGX-1 nodes
- Mellanox EDR InfiniBand EDR network
  1. One 216-port CS7520 Mellanox InfiniBand EDR modular core switch
  2. 144 InfiniBand EDR optical cables from systems to core switch
  3. 72 ports open to expand to other PODs
- 2 Ethernet 48 port 10Gb Ethernet switch (2 ports per DGX-1)
- 1 Ethernet 48 port 1Gb Ethernet switch (out-of-band network, 1 port per DGX-1)
Large-sized Multi-Node Configuration: Up to 144 Nodes (4 compute PODs)

The “large” configuration (Figure 6) implements 4 of the PODs (medium configuration) described earlier, comprised of 144 DGX nodes in total, distributed across 24 racks. As with the small and medium configurations, the large configuration offers full bisection network bandwidth within each POD. Inter-POD connectivity is via an InfiniBand top level switch, resulting in 2:1 oversubscription of bandwidth between PODs. It is recommended that when scheduling a training run, the job should be resourced to run within a POD, and avoid straddling more than one POD. This will maximize performance, and reduce inter-POD traffic across the top level switch.

Figure 6: Large-sized multi-node configuration

- 4 compute Pods
  1. Each 36 system compute Pod cluster fully interconnected
  2. 4 InfiniBand EDR links per system
  3. 6 DGX-1 nodes per rack (~19.2kW racks)

Bill of Materials

- 24 racks (4 compute PODs with 6 racks each)
- 144 DGX-1 systems (4 PODs with 36 DGX-1 each)
- Mellanox InfiniBand EDR network
  1. Leaf switches, one per POD
4 Mellanox CS7520 EDR Infiniband 216-port modular leaf switches (4 compute PODs)

2. Top level switch - Connects multiple PODs / leaf switches
   - 144 InfiniBand EDR optical cables from systems to top level switch
   - Two options for the top level switch
     a. Option A - Mellanox CS7510 324-port InfiniBand EDR switch
     b. Option B - (8) Mellanox SB7700 36-port InfiniBand EDR switches
       - 16 Ethernet 48 port 10Gb Ethernet switch (2 per compute POD, 2 ports per DGX-1)
       - 8 Ethernet 48 port 1Gb Ethernet switch (out-of-band network, 1 per compute POD, 1 port per DGX-1)

Data Center Power and Cooling Considerations

Deploying a DGX-1 scale-out environment requires careful consideration of compute, power and cooling density together. For example, component-level cooling allows higher densities in the Data Center, providing improved performance per watt and performance per dollar.

We recommend designing for higher density power per rack, from 32kW, up to 50-60kW, using multiple 208V/3-phase/60A or 400V/230V/3-phase/30A power circuits per rack. Higher voltages are more stable and efficient, thus lowering power OpEx.

DGX Management Considerations

Some organizations will require implementation of a management / scheduling solution for their DGX-1 cluster. In such scenarios, it is recommended for implementers to consult with available 3rd party ISV solutions familiar to the HPC domain, that can be used to assist with managing a DGX-1 multi-node cluster. More information on these options can be found at https://developer.nvidia.com/cluster-management

Data Center GPU Manager (DCGM) Integration with ISV Management Solutions

NVIDIA has worked closely with the ISV partners (listed below) to integrate DCGM\(^9\) into industry-leading cluster and workload management solutions. This integration provides IT admins richer GPU management capabilities, improved UI experience, and higher throughput and resiliency with optimized job scheduling.

Bright Cluster Manager\(^\text{TM}\) version 8.0, includes DCGM support, expanding its GPU management to the next level. With the integration of DCGM, Bright Cluster Manager can plot graphs of GPU metrics at much higher accuracy, and seamlessly correlate job statistics with system-level metrics. In addition, Bright Cluster Manager can provide non-invasive health checks while jobs are running as well as deep

GPU diagnostics available to administrators through Bright’s burn-in environment, used for cluster validation during infrastructure refresh.

**Altair PBS Professional® HPC workload management software** provides DCGM integration, offering the following key new functionality:

- Pre-job node risk identification and GPU resource allocation;
- Automated monitoring of node health;
- Reduced job terminations due to GPU failures;
- Increased system resilience via intelligent routing decisions;
- Increased job throughput via topology optimization;
- Optimized job scheduling through GPU load and health monitoring.

**IBM® Spectrum LSF™** provides extensive support for NVIDIA GPUs and is adding support for DCGM. The next release of LSF version 10.1 with DCGM integration will enable the following capabilities:

- Intelligent topology aware job placement of applications onto healthy GPU’s
- Monitoring of GPU health
- CPU-GPU Affinity
- GPU Power Management and support for synchronous boost.
- Integrated GPU accounting

In addition to these options, implementers can consider **Slurm Workload Manager**. While it isn’t currently integrated with DCGM, customers with experience using this solution in the HPC realm may find it useful in managing/scheduling their DGX environment..
Summary

The transition from a traditional data center to one that is optimized for modern, GPU-accelerated AI and deep learning, requires careful consideration of the architecture needed for scale. As deep learning workloads become increasingly computationally intensive, best practices for storage and networking must be considered, and valuable insights can be derived from industry-leading examples of multi-node scale such as the DGX SATURNV.

NVIDIA DGX-1 incorporates a number of system technologies that aid in multi-node scale, including NVLink, GPUDirect RDMA and NCCL2 - all of which can work in concert to help the IT administrator deliver predictable performance with scale. Additionally, the plan for building a multi-node system should consider these factors:

- **Recognize that DNN workloads are inherently more stressful than HPC and other workloads**, in how they consume data center resources. The stated power draw of a DGX-1 node is 3200W, and the assumption should be made that the node will draw power very close to that rate under normal operating conditions. As a result, designers should plan for higher density power per rack, and associated cooling demand
- **Implement a non-blocking networking fabric** that ensures high-bandwidth and low latency between nodes in a workload cluster, using InfiniBand networking deployed in a fat-tree topology, with a minimum of 2 connected InfiniBand EDR ports per node
- **Storage must support very large datasets, accounting for the read-intensive nature of DNN’s**. Cluster design should leverage DGX-1 local SSD cache for accelerated data streaming on read operations, along with NFS appliances using 10GbE interfaces for small cluster configurations. Consideration should be given to parallel file systems for high performance storage across across clusters that are expected to be large, ie: 32+ nodes
- **Proven configuration models can help ensure seamless scalability of DNN workloads**: “small” multi-node configurations for up to 12 nodes, “medium” configurations for up to 36 nodes, and “large” configurations for up to 144 nodes. The medium and large configurations can be easily built using the 36-node “POD” construct as a building block

The guidance provided here can be a valuable foundation in ensuring that IT administrators and the data science teams they support, can maximize the value and performance gained from their DGX-1 investment, yielding accelerated results and insights for their organizations.
Appendix

DGX-1 Overview

System Architecture

The NVIDIA DGX-1 is designed for maximizing the speed with which DNNs can be trained. It is based on an integrated system design that includes eight NVIDIA Tesla V100 GPU’s connected in a hybrid cube-mesh topology. DGX-1 maximizes multi-GPU and multi-node training performance using several enabling technologies, including:

- **NVLink interconnect of the system-level GPU complex**, providing a 10X speed-up over traditional PCIe Gen3 based interconnect
- **480 GB of SSD for O/S boot, and 4x1.92 TB SAS SSD** configured as a RAID 0 striped volume for high-performance, high-bandwidth streaming of large training datasets
- **Four InfiniBand 100 Gbps Extended Data Rate (EDR) ports** which provide multi-node interconnect with very high bandwidth and low latency, while also enabling GPUDirect Remote Direct Memory Access (RDMA) for GPU to GPU data transfer between systems
- **Dual 10 GbE network interfaces** for connecting the system to external storage, as well as interactive sessions
- **Out-of-Band Management** providing remote management for lights-out administration
- **DGX software** includes a complete containerized stack that includes popular deep learning frameworks, each optimized for maximum training performance, including dependent libraries and drivers
LEGAL NOTICE

Notice

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.

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

NVIDIA, the NVIDIA logo, CUDA, Pascal, Tesla, NVLink, and DGX 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 © 2017 NVIDIA Corporation. All rights reserved.