TECHNICAL OVERVIEW

CONSIDERATIONS FOR SCALING GPU-READY DATA CENTERS

New Rules and Best Practices for Running Deep Learning Workloads in the Modern AI Data Center

NVIDIA.

TABLE OF CONTENTS

Abstract

Enterprise and hyperscale data centers are increasingly being built around workloads using Artificial Intelligence (AI) and computationally intensive Deep Neural Networks (DNNs) with massive amounts of data. The level of computation required is significant and benefits greatly from the power of GPUs. They're massively parallel, optimized for high memory bandwidth, and designed for the AI-class matrix multiplication and analytics needed for fast data insight. Data centers that support GPU servers with dense, high-power racks featuring advanced cooling techniques like water cooling and hot aisle containment use significantly less floor space. They also provide much higher efficiency and performance, as well as lower overall power usage for these advanced workloads. This paper describes best practices for making a data center 'GPU-ready' with a focus on power, cooling, and architecture, including rack layout, system and network architecture, and storage. Using examples of computationally intensive workloads on NVIDIA® DGX-1™ Systems for deep learning and NVIDIA Tesla® V100 GPU Accelerators, this paper provides a guide to minimizing spend. It also provides tips to ensuring that a data center is optimized for NVIDIA GPUs to run today's advanced workloads at scale.

New Rules for AI Data Centers

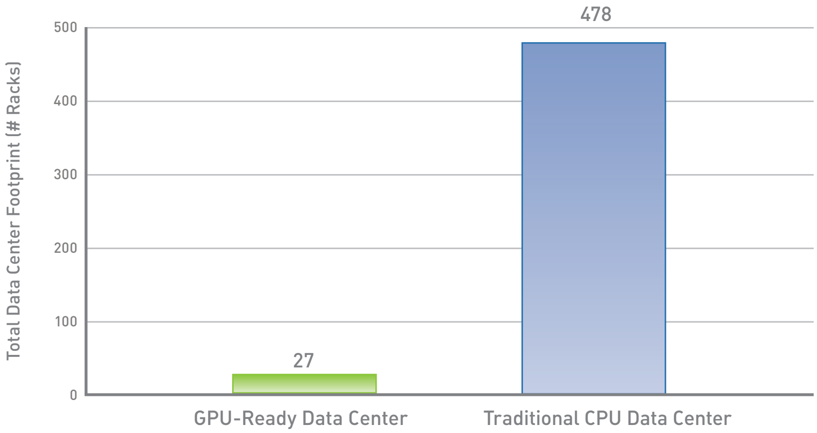
Today's data centers rely mainly on servers with one or two CPU sockets running general-purpose workloads. As the drive for faster data insight grows, new computing paradigms using AI workloads are becoming commonplace. Using computationally intensive DNNs for these workloads is only feasible with the massive performance gains from new types of servers based on GPU technology. This paper focuses on how to design, deploy, manage, and monitor data centers for optimum efficiency and performance using GPU-based technologies.

AI/deep learning workloads run in two modes of operation: DNN training and inference. GPU-based servers provide many benefits for DNN workloads. These include significantly higher performance per server and substantially better performance per watt, delivering lower overall data center power usage with a fraction of the number of racks.

A single high-density GPU server can match the performance of dozens of CPU-based servers. The charts below show comparable clusters of GPU vs. CPU server racks running typical workloads. Chart 1 illustrates AI Research workloads, where 27 NVIDIA DGX-1 racks (666 KW) offer the same performance as 478 racks (12,054 KW) of CPU-only systems. Chart 2 illustrates AI Batch Production, where 34 NVIDIA DGX-1 racks (656 KW) compare to 1602 CPU-only racks (34,944 KW). Chart 3 assumes Mixed Workloads, where 30 NVIDIA DGX-1 racks (648 KW) compare to 1119 CPU-only servers (24,752 KW). Assuming a similar volume of AI/deep learning and HPC workloads, a GPU-ready data center needs only 1/40 the footprint and 1/20 the power of a traditional CPU-only data center.

Chart 1: Chart 1 shows a GPU-ready AI research data center heavily focused on AI training and algorithm development with dense computational resources. This data center provides some resources dedicated to data preparation and AI inference.

AI Research Data Center: Maximum Performance Training with Extended Workloads



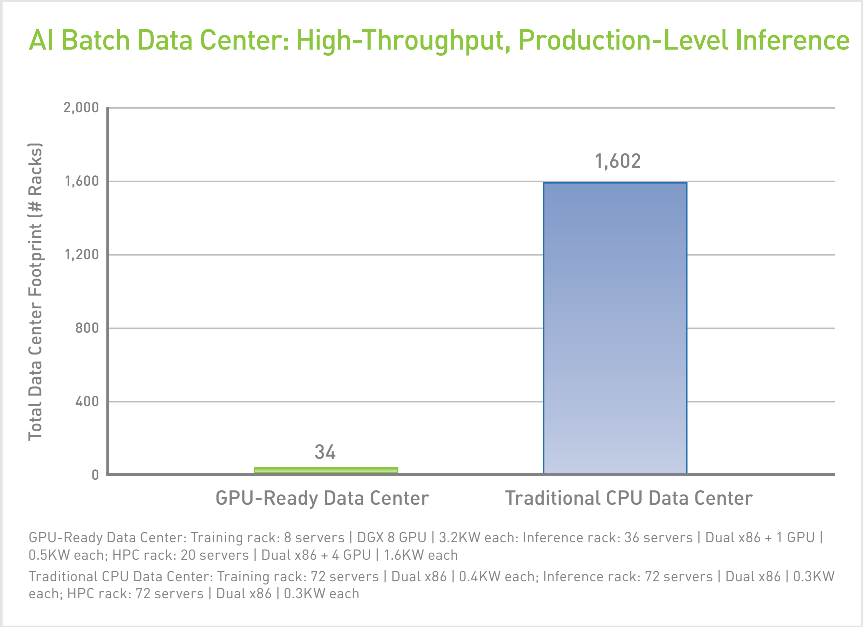
GPU-Ready Data Center: Training rack: 8 servers | DGX 8 GPU | 3.2KW each; Inference rack: 36 servers | Dual x86 + 1 GPU | 0.5KW each; HPC rack: 20 servers | Dual x86 + 4 GPU | 1.6KW each

Traditional CPU Data Center: Training rack: 72 servers | Dual x86 | 0.4KW each; Inference rack: 72 servers | Dual x86 | 0.3KW each; HPC rack: 72 servers | Dual x86 | 0.3KW each

Chart 1

[1] NVIDIA Performance Lab

Chart 2: Chart 2 shows a production AI inference GPU-ready data center that is focused mostly on AI Inference in a large-scale production environment. This data center provides some resources dedicated to data preparation and AI training.

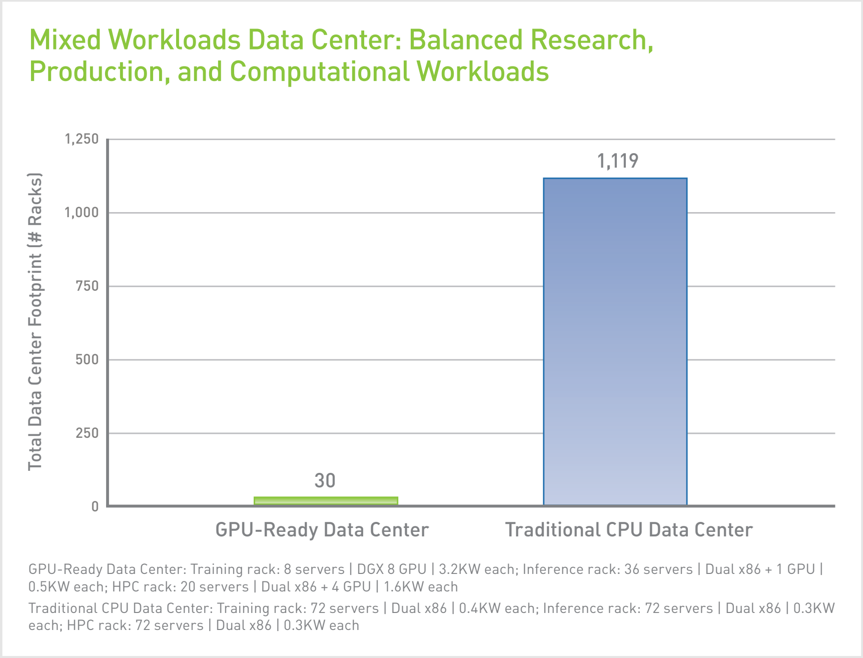

AI Batch Data Center: High-Throughput, Production-Level Inference

GPU-Ready Data Center: Training rack: 8 servers | DGX 8 GPU | 3.2KW each: Inference rack: 36 servers | Dual x86 + 1 GPU | 0.5KW each; HPC rack: 20 servers | Dual x86 + 4 GPU | 1.6KW each

Traditional CPU Data Center: Training rack: 72 servers | Dual x86 | 0.4KW each; Inference rack: 72 servers | Dual x86 | 0.3KW each; HPC rack: 72 servers | Dual x86 | 0.3KW each

Chart 2

Chart 3: Chart 3 shows a GPU-ready data center designed to run Mixed Workloads with a combination of AI batch or interactive research and production operation using a mix of AI training, inference, and computational resource.


Mixed Workloads Data Center: Balanced Research, Production, and Computational Workloads

GPU-Ready Data Center: Training rack: 8 servers | DGX 8 GPU | 3.2KW each; Inference rack: 36 servers | Dual x86 + 1 GPU | 0.5KW each; HPC rack: 20 servers | Dual x86 + 4 GPU | 1.6KW each

Traditional CPU Data Center: Training rack: 72 servers | Dual x86 | 0.4KW each; Inference rack: 72 servers | Dual x86 | 0.3KW each; HPC rack: 72 servers | Dual x86 | 0.3KW each

Chart 3

In addition to efficiency, GPU-based servers provide significant Total Cost of Ownership (TCO) savings for a large multi-node system. Table 1 below shows a deployment of NVIDIA DGX-1 systems in contrast with 250 CPU-based servers. The scenario reflects the three-year total cost of ownership inclusive of the servers, networking (10 Gigabit Ethernet and InfiniBand), power, co-location[2], and systems administration. The dramatic reduction in physical infrastructure enabled by the dense computational footprint of the NVIDIA DGX-1 creates a TCO advantage over traditional CPU-based systems.

[2] https://en.wikipedia.org/wiki/Colocation_centre

Table 1: TCO Comparison of NVIDIA DGX-1 to CPU Servers

Assuming 2RU dual-socket CPU server @ $10,000/each, 48-port 10 GBe switch, 36-port IB switch; 13 racks total footprint of CPU solution; approx. 318 kW total power with 1.5 data center PUE; $0.085/kWh, 25% sys admin per year for NVIDIA DGX-1 @ $250k/yr loaded labor rate; one (1) sys admin per year for CPU environment.

	NVIDIA® DGX-1™ SYSTEMS (1 SERVER)	CPU SERVER ENVIRONMENT (250 SERVERS)
Up-Front Capital Expenses		
Server (OTP)	$149,000	$2,500,000
Network & Cables (OTP)	$16,280	$187,600
Recurring Operating Expenses		
Power (3 yrs)	$7,153	$710,835
Colo (3 yrs)	$43,200	$1,774,800
Sys Admin OpEx (3 yrs)	$187,500	$750,000
Support and Maintenance (3yrs)	$63,698	$1,125,000
TOTAL 3 YR COST	$466,831	$7,048,235

Table 1

The much denser compute capability of NVIDIA GPU-based servers provides three-year cost savings described in the table above. They also require 15-32 kW of power and cooling per rack, which is typically higher than today's average data center design point; many of today's cloud data centers have power distribution and cooling infrastructures designed to handle only 5-10 kW/racks. Open Compute Project[3] (OCP) designs, for example, define workload-specific servers and custom rack designs.

The OCP V2 rack has two 6.6 kW power shelves[4] that limit power to 13 kW/rack. This only allows four dense GPU servers per rack, thereby losing the advantage of density gains.[5] The need for compute has driven hyperscale data centers like Facebook Prineville to grow from 10,000 (2008) to 30,000 (2009) to 60,000 servers today and add 487,000 square feet to a 307,000 square-foot facility (13 football fields total). Currently, the growth is in linear space versus an increase in density. The growth also drove a linear increase in their overall network investment of $3.63 billion in 2015, up from $3.02 billion in 2014.[6] Increasing compute capability with dense servers can greatly reduce floor space and network requirements.

The same problems apply to all sizes and types of data centers. According to Rick Villars, Vice President, Data Center & Cloud, IDC, "Typical enterprise data centers have configured their power systems to deliver less than 8 KW per rack, while leading cloud service providers with denser designs deliver closer to 12 KW per rack. For their next-generation data centers, IDC believes these companies are targeting around 30 KW per rack as they plan for a dramatic increase in real-time analytic and cognitive workloads that require the inclusion of dense GPU capacity in their compute pools."

Thinking Differently About Scaling with GPUs

Deep Neural Networks (DNNs) are the core of today's AI applications and can have thousands of layers, hundreds of thousands of neurons,

[3] http://www.opencompute.org/

[4] OCP V2 Power Shelf Spec

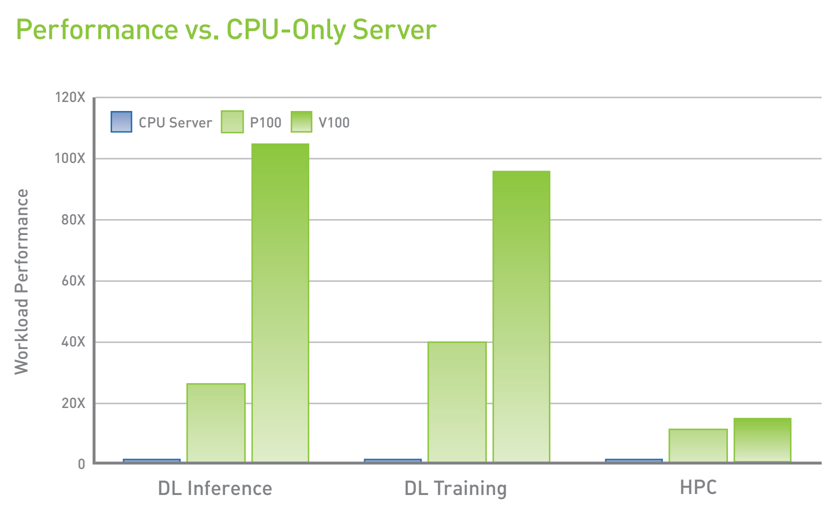
[5] http://www.DataCenterdynamics.com/content-tracks/open-data-center/ocp-summit-facebook-refreshes-its-servers/97937.article

[6] http://www.DataCenterknowledge.com/the-facebook-data-center-faq/

and millions of connections. The impressive performance of today's AI models is achieved by training these large DNNs with Gigabytes or Terabytes of data across hundreds of computational iterations to find the most accurate set of weights. GPUs drive AI with massively parallel compute and optimized high-memory bandwidth, enhanced for AI-class matrix multiplication and convolution. While GPU systems provide much higher performance per system than typical CPU-only systems, they also drive greater density and power requirements.

The drive for faster and more accurate insight with larger AI models also requires performance beyond a single-GPU system. Scaling AI and other heavy workloads to multiple servers involves executing an application across many servers with minimum bottlenecks to ensure high performance. In contrast to scaling with traditional lightweight CPU-only servers, the greatest GPU system benefits are seen when starting with compute-dense servers designed with many GPUs per server before scaling to multiple GPU servers. Dense GPU servers are the ideal data center building blocks for multi-server deep learning training workloads.

Chart 4: GPU-based systems provide significant performance gains for AI and HPC workloads over CPU-only systems, reducing footprint in data centers and increasing performance density of each compute rack. In addition, increased density means less systems and much better scaling efficiency of large workloads.



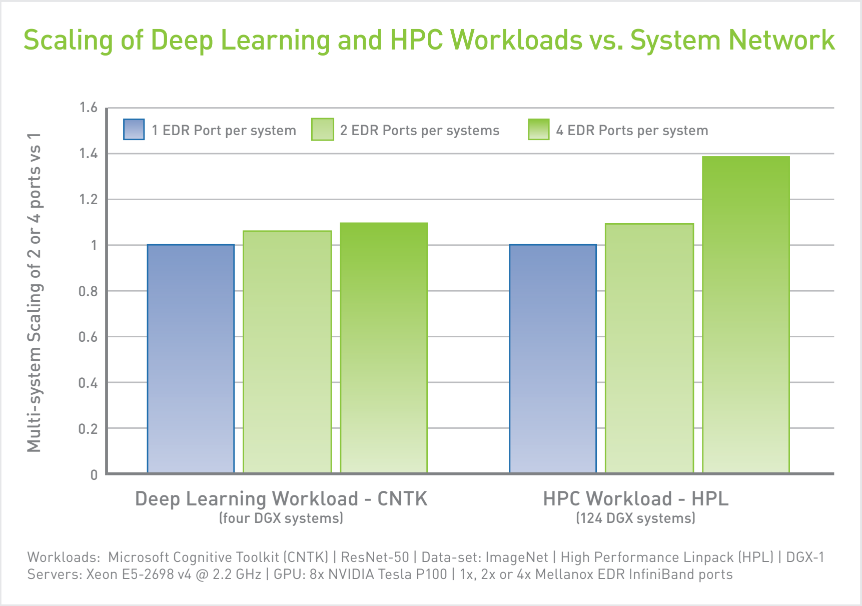
Deep Learning Inference Workload: ResNet 50 Inference | ImageNet Dataset | CPU Server: Xeon E5-2690 v4 | GPU Servers: Add 1x NVIDIA Tesla P100 or V100

Deep Learning Training Workload: ResNet 50 | ImageNet Dataset | CPU Server: Dual Xeon E5-2699 v4 | GPU Servers: Add 8x NVIDIA® Tesla® P100 or V100

HPC Workload: VASP | b.hr105 dataset | CPU Server: Dual Xeon E5-2690 v4 | GPU Servers: Add 4x NVIDIA Tesla P100 or Tesla V100

Chart 4

Massive computation of AI compute elements also requires strong networks between systems to ensure scalable performance. Chart 5 below shows a comparison of performance versus number of network ports per system and large tradeoffs when using different numbers of network connections to each system.



Workloads: Microsoft Cognitive Toolkit (CNTK) | ResNet-50 | Data-set: ImageNet | High Performance Linpack (HPL) | DGX-1 Servers: Xeon E5-2698 v4 @ 2.2 GHz | GPU: 8x NVIDIA Tesla P100 | 1x, 2x or 4x Mellanox EDR InfiniBand ports

Chart 5

Chart 5: Network interconnect is critical when designing with very high-performance GPU systems and can have a large impact on multi-node performance. Clusters based on DGX-1 systems that use four InfiniBand links per node can provide 20% performance gains for DL workloads and 40% performance gains for HPC workloads over the same systems when using only one InfiniBand link per system.

Chart 5 features both CNTK for deep learning training (multi-node CNTK, ResNet50)[8] and HPL for computational science HPC workloads. These rely on heavy computation and high-performance communication for best performance and both lose multi-system performance when using less network ports per system.[9] The graph shows that using four ports per system provides almost 40% more performance for HPC and 10% more performance for CNTK deep learning workloads. 10% performance gain may seem small at first. But the cost for this can be much less than 10% of the total system costs to implement the correct interconnect, reducing bottlenecks and providing more consistent performance across the board for both compute and storage access.

New GPU-Ready Data Center Best Practices

POWER AND COOLING

Solving large-scale infrastructure problems means considering compute, power, and cooling density together. Several of today's cooling solutions provide improved performance per-watt and performance per-dollar and leverage higher densities in the data center. These techniques include:

- Hot or cold aisle containment
- Rear-door water heat exchangers
- Component-level water cooling

These advanced cooling techniques provide a significant benefit with GPU servers to minimize power and floor space needs and increase performance efficiency. Table 2 below shows trade-offs between cooling solutions.

[8] NVIDIA Performance Labs

[9] NVIDIA Performance Labs

Table 2: Sample GPU-Ready Server Configurations

COOLING SOLUTION	COOLING TYPE	RACK POWER	SOLUTION SIZE
Traditional Air Cooling	Air	8kW	52 racks
Hot/Cold aisle containment	Air	15 kW	28 racks
Rear Door Heat Exchangers	Air+Water	35 kW	12 racks
Direct water cooling	Water	60 kW	7 racks

Table 2

Today's data centers with AI and data-focused workloads can drive different needs for GPU servers, and further optimization can be made based on workload type. Table 3 below shows sample GPU server configurations focused on DNN, analytics, and HPC workloads with corresponding power, rack, and cooling. GPU servers (eight GPUs per server in NVIDIA® DGX-1™) for DNN training benefit greatly from extremely dense GPU servers and racks. This greatly reduces floor space and cabling requirements if an adequate cooling system can be configured.

Table 3: Sample GPU-Ready Server Configurations

COMPUTE RACKS		DNN TRAINING/BATCH INFERENCE	DNN REAL-TIME VIDEO INFERENCE	DATA ANALYTICS	HPC
Sample Server Model		3u 8 GPU system - NVIDIA DGX-1	1/2u 1 GPU system	4u 8 GPU system	1u 4 GPU system
Compute	CPU:	2 high-end	2 low-to-mid	2 high-end	2 high-end
	GPU:	8x NVIDIA Tesla V100	1x NVIDIA Tesla V100/low-power	8x NVIDIA Tesla V100	4x NVIDIA Tesla V100
System Memory		512-1024 GB	128-256 GB	512-1024 GB	256-512 GB
Network	Internal:	NVIDIA NVLink™	PCIe	NVLink	PCIe
	Multi-node:	100 GB InfiniBand	10 GB Ethernet	25 GB Ethernet	100 GB Ethernet
Servers/ Rack		4 to 8	36 to 72	4 to 10	10 to 20
Power/ Server (W)		3,200	500	2,400	1,500
Power/ Rack (KW)		32	18	32	15-30
Less Dense Racks		4 servers 12.8 KW 1,340 CFM	36 servers 18.0 KW 1,320 CFM	4 servers 9.6 KW 1,000 CFM	12 servers 18 KW 1,890 CFM
-Cooling Solutions		Air - Partition Water RDHX	Air - open aisle	Air - Partitions Water - RDHX Direct water	Air - open aisle
Dense Rack		8 servers 25.6 KW 2,630 CFM	72 servers 36 KW 2,650 CFM	8 servers 19.2 KW 2,050 CFM	24 servers 36 KW 3,800 CFM
-Cooling Solutions		Air - Partition Water RDHX Direct water	Air - Partitions Water - RDHX	Air - Partitions Water - RDHX Direct water	Air - Partition Water RDHX

Table 3

From Table 3, some items are important to note about GPU servers versus traditional CPU data centers.

> GPU-based servers require much higher air flow per server to maintain the highest performance. It's critical to ensure that air flow in and through the racks properly accounts for the higher volume of air and temperature difference. Gaps between equipment must be blocked, and airflow within the data center must be carefully designed to ensure that hot air returns to the chillers and does not stay in the data center, raising intake temperatures.

- High-power density racks need special care to ensure that power and cooling are properly balanced across the server, as well as the rack. Characterizing peak power loads in your racks is important to ensure overload scenarios during peak power consumption don't cause issues. Power should be properly load-balanced across nodes and servers so unexpected power surges don't cause nodes to fail. Higher-density power racks—from 32 kW up to 50-60 kW—using multiple 208V/3-phase/60 A or 415 V/240 V/3-phase/30 A power circuits per rack are ideal. In addition, higher voltages are more stable and efficient, providing lower power-operating expense.
- Also consider rear-door cooling, component-level liquid cooling, and immersion. Liquid-cooled systems can be used to conduct up to 3,500 times more heat[10] than air-cooled systems. Component-level liquid cooling can also capture between 60-80% of server heat and reduce costs by 50%, which allows for a 2-5X increase in density. Even when using water-based, rack-level heat-exchanging cooling systems, it's still important to guarantee that the hot air is removed from the rack and doesn't continue to circulate into the front of the servers.
- A "rule-of-thumb" metric of 100 cfm/kW of server load with a 5% overhead for air leakage and short cycling was used to calculate server air-flow requirements. The total Cubic Feet Per Minute (cfm) used was 105 cfm/kW of server load for heat rejection.

HGX SERVER REFERENCE ARCHITECTURE FOR GPU SERVERS

With the rapid pace of innovation in GPU technology, server architecture becomes increasingly important. The NVIDIA HGX-1 hyperscale GPU accelerator architecture[11] has been widely deployed in the world's largest cloud service providers, and elements of that design are found in many enterprise-class GPU servers, including the NVIDIA DGX-1. These platforms are optimized to deliver industry-leading performance for AI and data analytics workloads. Items considered in the HGX Reference Architecture include:

- PCIe and NVLink topologies for GPU, CPU, network, and storage interconnects
- CPU-to-GPU ratios
- System memory capacity
- Local storage, including SSD and NVME

Because NVIDIA DGX-1 is NVIDIA's first platform to deliver deep learning software performance optimization, customers are assured that such platforms will always provide the highest levels of performance. NVIDIA software libraries like NCCL (NVIDIA Common

[10] http://www.pge.com/includes/docs/pdfs/mybusiness/energysavingsrebates/incentivesbyindustry/DataCenters_BestPractices.pdf

[11] https://www.nvidia.com/en-us/data-center/hgx-1/

Collectives Library) are optimized for the PCI and NVIDIA NV Link topologies of the HGX Reference Architecture.

Maximizing GPU density within a server provides the highest level of performance for GPU-accelerated applications, including deep learning training, data analytics, databases, and high-performance computing. Most GPU-accelerated applications scale well to eight GPUs per server with properly configured CPU, memory, networking, and local storage. Deep learning frameworks, including MXNet, TensorFlow, Caffe2, and Microsoft's Cognitive Toolkit, all scale well to eight GPUs. Some HPC applications have not been optimized to scale beyond two or four GPUs, so fewer than eight GPUs may be optimal if your workload is dominated by HPC applications.

Balanced performance of the NVIDIA HGX Reference Design is ensured with:

> Sufficiently powerful CPUs, typically two high-end x86 CPUs to match eight-GPU performance.

> System memory configured to be at least 2x GPU memory with 4x being optimal for deep learning training. GPU-accelerated data analytics and databases generally benefit from as much system memory as can be configured in the server.

> For distributed or multi-node deep learning training, use a minimum of one 100 GB network interface card (NICs) supporting RDMA configured for every two GPUs. These NICs should be located on the same PCIe switch as the GPUs.

> Network topology that supports GPUDirect Peer 2 Peer transfers from GPU to GPU inside a system across NVIDIA NVLink and GPUDirect RDMA between GPUs in multiple systems across InfiniBand.

> SSD and NVME local storage configured on the same PCIe switch, or as close as possible, to the GPUs.

COMPUTE NETWORK RECOMMENDATIONS

Scaling beyond individual servers requires communication networks that provide high bandwidth, low latency, and high efficiency. When building your data center, consider using 100 GB Ethernet, EDR (100 GB) or HDR (200 GB) InfiniBand[12] for these compute networks.

Ethernet networks can approach InfiniBand performance and efficiency in many cases. Think about the following:

> To minimize the load of the Ethernet adapter on your CPU, consider using adapters that support TCP offload.

[12] http://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf

- The Ethernet switch architecture should support cut-through communications.
- Use network adapters that support Remote Direct Access Memory (RDMA) for the highest performance and most efficient transfers.
- Create layer-two networks using a spine-leaf topology, large uplinks, and fewer switches to minimize bottlenecks due to link congestion. Networks designed using a spine-leaf[13] topology provide a cost-effective way to build networks with high-bisection bandwidth—a key characteristic for efficient scaling of distributed applications.
- Use the fewest number of layer-three networks to minimize bottlenecks due to routing.
- Consider designs that localize traffic for systems intended for running scalable applications.

For the highest multi-server GPU performance, InfiniBand is specifically architected to support high-compute, multi-server applications. This is an industry standard that provides high-bandwidth and low-latency communications for scaling applications across nodes. It's ubiquitous in the HPC community as the technology used to connect both small (less than 20 nodes) to extremely large (thousands of nodes) clusters. Consider the following options when designing your InfiniBand network:

- Use full fat-tree networks to maximize the total cluster bandwidth of the network.
- Use multiple InfiniBand connections per node for dense GPU nodes to maximize performance.

To achieve multi-server scaling performance, it's critical to balance the bandwidth of traffic between GPUs inside a node with traffic between multiple servers. Table 4 below compares two multiple-server systems.

Table 4: Relative Multi-Node Computational Code Performance with Different High-Speed Interconnects

EXAMPLE 8 SERVER SYSTEM	SERVER	NETWORK TECHNOLOGY	BANDWIDTH IN/OUT OF EACH SERVER	TOTAL MULTI-SERVER BANDWIDTH[14] (8 SERVERS)	RELATIVE APPLICATION PERFORMANCE BETWEEN SOLUTIONS[15]
NVIDIA DGX-1 8 GPU servers, 160 GB/s internal GPU-to-GPU bandwidth		10 GB Ethernet (1 port per system)	2 GB/s per system	16 GB/s total	1X
		100 GB EDR InfiniBand (4 ports per system)	47 GB/s per system	376 GB/s total	2X

Table 4

In Table 4, because the internal bandwidth between GPUs in each system is 160 GB/s, it's critical to maintain balance between communications within the node and off node. The EDR solution provides 47 GB/s of off-node bandwidth that's 20X the performance of the 10 GB Ethernet-based solution. Plus, it's a much better balance for high-computational workloads, resulting in 2X real multi-server application performance.

[13] http://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/white-paper-c11-737022.html

[14] Bisection bandwidth is the total bandwidth available between two halves of a networked cluster system. It is determined by splitting the system network down the center and adding the bandwidth of all the links that were split.

[15] Comparison based on average performance gains between several computation codes when run using each type of network.

STORAGE ARCHITECTURE

As an organization scales out their GPU-enabled data center, there are many shared storage technologies that pair well with GPU applications. Because the performance of a GPU-enabled server is so much greater than a tradition CPU server, special care needs to be taken to ensure that the performance of a storage system isn't a bottleneck to advanced workloads.

Workload properties need to be considered because they can drive different access patterns and data types. Running parallel HPC applications may require the storage technology to support multiple processes accessing the same files simultaneously. Accelerated analytics require storage technologies with support for many threads and quick access to small pieces of data. Vision-based deep learning that accesses images and video used in classification, object detection, or segmentation is dominated by reads and requires high streaming bandwidth, fast random access, or fast memory mapped (mmap) performance. Other deep learning techniques like recurrent networks working with text or speech can require any combination of fast bandwidth with random and small files.

For deep learning, the ability to cache previously read data is paramount for maximizing training performance. Deep learning training maximizes accuracy by iterating over the data multiple times. It's not uncommon for a training exercise to consist of at least 100 iterations. If data is cached locally, then shared storage doesn't need to be accessed for each iteration. The local memory and local disk can be used to cache data depending on the file system technology. It's best to match the capacity and performance needs of the local cache with the needs of your deep learning applications.

Table 5 below shows general guidelines of the storage architecture for different GPU-enabled workloads. As always, it's best to understand your own applications' requirements to design the optimal storage system.

Table 5: Storage Architectures

USE CASE	ADEQUATE READ CACHE?	NETWORK TYPE RECOMMENDED	NETWORK FILE SYSTEM OPTIONS
Data Analytics	N/A	10 GBe	Object-storage, NFS, or other system with good multi-threaded read and small file performance
HPC	N/A	10/40/100 GBe, InfiniBand	NFS or HPC targeted file system with support for large numbers of clients and fast single-node performance, support multi-threaded writes
Deep learning, 256x256 images	Yes	10 GBe	NFS or storage with good small file support
Deep learning, 1080p images	Yes	10/40 GBe, InfiniBand	High-end NFS, HPC file system or storage with fast streaming performance
Deep learning, 4K images	Yes	40 GBe, InfiniBand	HPC Filesystem, high-end NFS or storage with fast streaming performance capable of 3+ GB/s per node
Deep learning, uncompressed Images	Yes	InfiniBand, 40/100 GBe	HPC Filesystem, high-end NFS or storage with fast streaming performance capable of 3+ GB/s per node
Deep learning, datasets that are not cached	No	InfiniBand, 10/40/100 GBe	Same as above, aggregate storage performance must scale to meet the all applications simultaneously

Table 5

Lastly, this discussion has only discussed performance needs. Reliability, resiliency, and manageability are as important as the performance characteristics. When choosing between different solutions that meet your performance needs, make sure that you've considered all aspects of running a storage system and the needs of your organization to select the solution that will provide the maximum overall value.

SYSTEM RUNTIME MONITORING AND MANAGEMENT

It's important that your system monitoring and management tools are GPU-aware. Systems must be able to monitor the temperature, clock rate, GPU memory usage, and other key GPU parameters. If your existing management tools lack GPU monitoring capabilities, or for additional GPU specific monitoring, you should use the NVIDIA Data Center GPU Manager (DCGM)[16].

DCGM is a complete suite of enterprise-grade tools for managing the accelerated data center. IT managers can implement system policies, monitor GPU health, diagnose system events, and maximize data center throughput. There's a number of tools that have already integrated DCGM, including **Bright Cluster Manager**, **Altair's PBSWorks**, **IBM Spectrum LSF**, **Adaptive Computing**, **SchedMD,** and **Univa**.

DCGM provides monitoring of GPU operation to minimize impact on overall performance, performance variability, and node health. Monitoring GPU temperatures prevents power throttling due to thermal extremes. Integrating DCGM into your scheduling software will provide accurate measurements of GPU utilization and throughput on a per-job instance. Running periodic GPU health and diagnostic checks using DCGM will also help to proactively identify components requiring service—allowing you to maximize uptime.

Other important system metrics to monitor in dense GPU nodes include fan speed, chassis and component temperature, system error logs (in particular logs associated with the PCIe bus), power supply state, and power consumption for each power supply. The Intelligent Platform Management Interface (IPMI) has long been a standard way of providing management and monitoring capabilities of these server components. It provides a wealth of information about the health of your servers. The IPMI sensors will give you an insight to the health of your server and often tell you when your servers are starting to fail.

[16] http://www.nvidia.com/object/data-center-gpu-manager.html

Summary

Enterprise and hyperscale data centers are increasingly being built around data focused workloads using Artificial Intelligence (AI) with computationally-intensive Deep Neural Networks (DNNs) and massive amounts of data. The level of computation required is significant and benefits greatly from the power of GPUs, which are massively parallel, optimized for high-memory bandwidth, and designed for the AI-class matrix multiplication, convolution, and analytics needed for fast data insight.

GPU systems provide much higher performance per system than typical CPU-only systems. When deployed in large scale data centers, they offer higher performance, better performance per watt, and faster time-to-solution, with a fraction of the compute racks. To realize these savings in a GPU-ready data center requires a more advanced approach to design and operation in these key areas:

- **Design Data Centers** to support much higher power densities. For highest efficiency, consider racks hosting 30 KW to 50 KW per rack and controlled temperature airflow into the systems. Also, consider liquid cooling at either the rack level or the component level to improve cooling efficiency ongoing costs. Review compute, power, and cooling density together. For example, component-level cooling allows higher densities in the data center and provides improved performance per watt and performance per dollar.
- **Build System Architectures** in your data center for data and AI-focused workloads that support large computation and high I/O throughputs. With the performance increases realized by GPUs, it's necessary to re-evaluate all system subcomponents to minimize bottlenecks including networking and storage.
- **Use Data Center Network and Storage Architectures** that provide high bandwidth, low-latency, and high efficiency to avoid bottlenecks for high-performance AI deep learning, accelerated analytics, and HPC workloads. These multisystem GPU workloads drive large data transfers and require robust, low-contention networks to achieve good scaling.
- **Boost System Monitoring and Management** of critical components to meet the demands of dense GPU systems and applications that efficiently scale across multiple nodes. Multi-system workloads are gated by the slowest system in the job, so consistent performance of all systems is key or fast systems will be waiting for slower ones to complete.

The principles of GPU-ready data center design laid out in this white paper are key to removing bottlenecks, reaching maximum performance and efficiency, and achieving the true capabilities of NVIDIA GPU systems.

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.