Data Center Best Practices with DGX B200

Steven Hambruch DCA, Distinguished Engineer, Data Center Infrastructure Architect

Dennis O’Brien DCEP, CISSP, RCDD, Senior Data Center Deployment Engineer

Agenda

Introduction to DGX SuperPOD Architecture

- Overview of the DGX SuperPOD Hardware, Connectivity, and Resource Requirements

The Economy of Data Center Resources

- Discussion of the balance and contention caused by deploying high density workloads in the data center

Best Practices for Optimized Deployments

- Planning for Power, Cooling, and Space requirements

- Summary / Q&A

Introduction to DGX B200 Data Center Architecture

Overview of DGX Hardware, Connectivity, and Resource Requirements

DGX B200

One universal building block for the AI data center

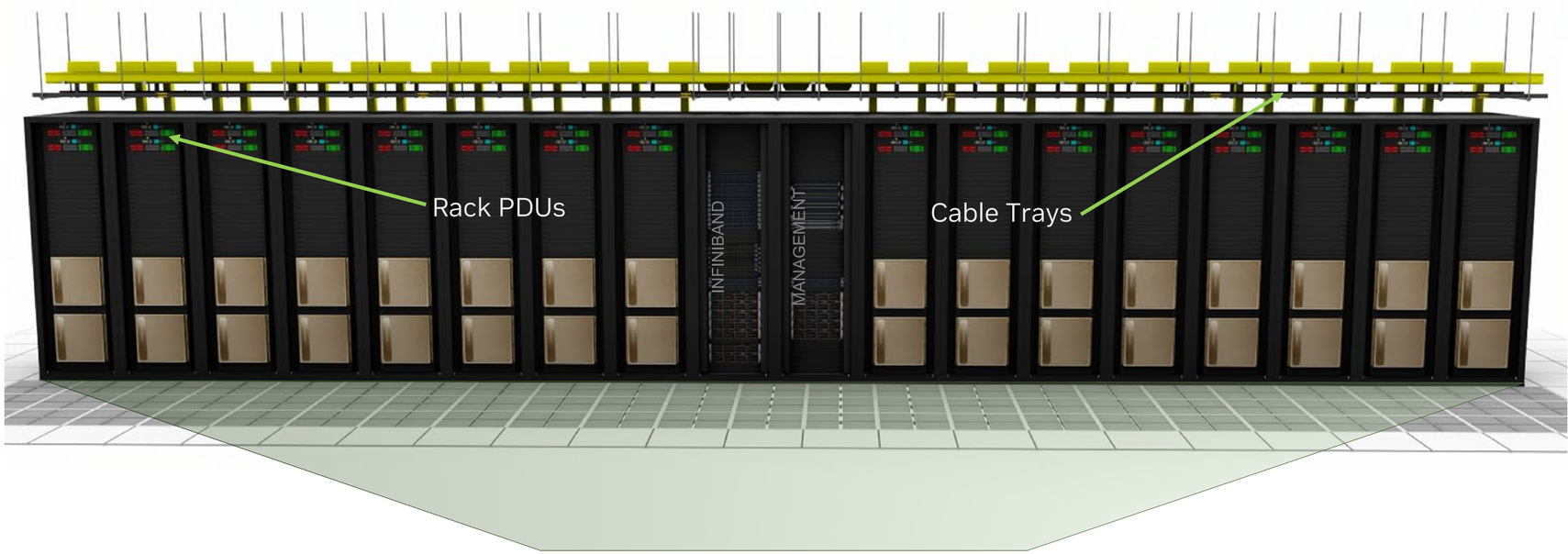
DGX B200 System Specifications	
Estimated System Power	**14.3 kW Max**
Estimated System Weight	>287.6 lbs. (130 kgs)*
System Dimensions	10 Rack Units (RU) Height: 17.5 in (444 mm) Width: 19.0 in (482.2 mm) Length: 35.3 in (897.1 mm)
Operating Temperature	5ºC to 30ºC (41ºF to 86ºF)
Cooling	Air
Operating Altitude	3048 meters (10,000 feet) maximum 5° to 35°C: altitude 0 – 1000 ft 5° to 30°C: altitude 1000 – 5000 ft 5° to 25°C: altitude 5000 – 10000 ft
Estimated Acoustic Noise Operating	Acoustic Power (LWA,m)* 25 ˚C /77 ˚F = 96 dB 30 ˚C / 86 ˚F = 101 dB

*Estimated. Actual specification to be determined

DGX SuperPOD Data Center Infrastructure

Power, Cooling, Cabling and Layout

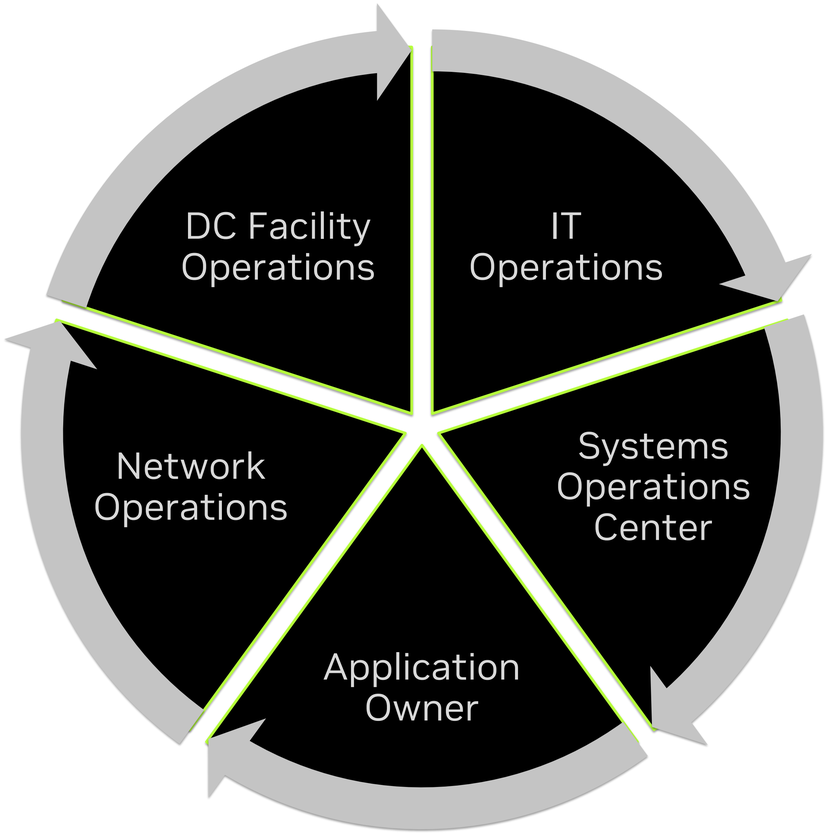
- A DGX Scalable Unit (SU) consists of up to 32 DGX B200 systems for 458kW per SU
- A DGX SuperPOD is comprised of multiple interconnected Scalable Units (Using a common InfiniBand spine and storage)
- A NVIDIA Unified Fabric Manager (UFM) appliance displaces one DGX B200 Server in the SuperPOD deployment pattern, resulting in a maximum of 127 DGX B200 Servers per full SuperPOD
- Performance and cost optimized
- Two air-cooled DGX B200s per 48U/52U rack
  - Four air-cooled DGX B200s per 52U rack in high density deployment pattern
- Proximity of the cabinets is governed by the IB cable distance limits
- InfiniBand Fabric drives rack to rack cable distance requirements



DGX B200 Scalable Unit (SU)

Planning a data center deployment

Equal consideration must be given to multiple domains in a DGX B200 deployment



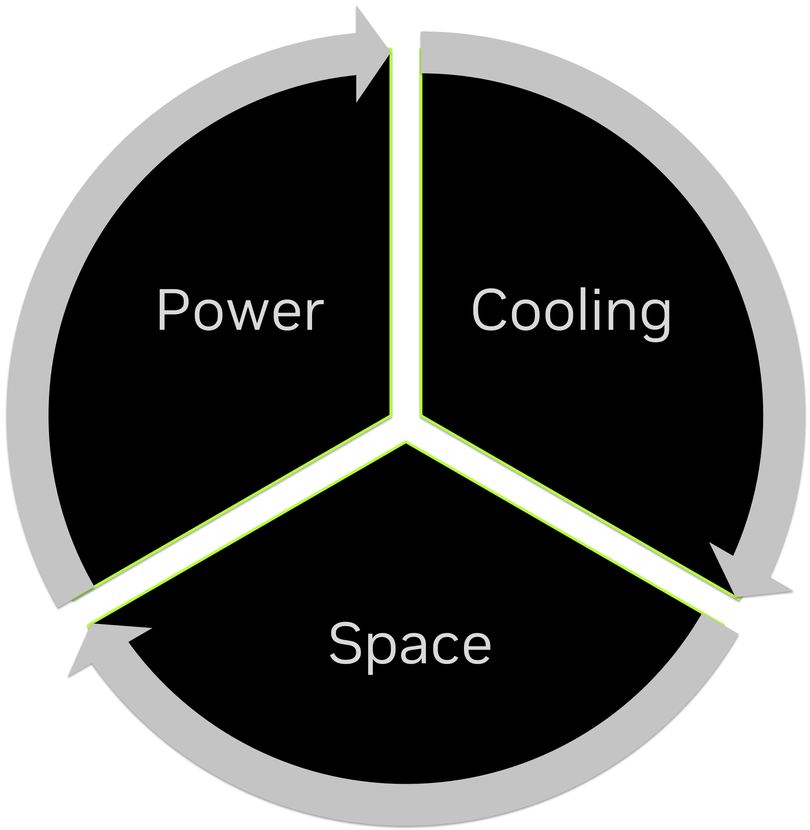
- A well-planned deployment will involve aligning multiple constituencies, including
  - Data Center Facility Operations
  - IT Operations
  - Systems Operations Center
  - Application Owner
  - Network Operations
- Solutions to resource constraints in one domain may impact planning, operations, personnel, and budgets in another domain

The Economy of Data Center Resources

Deploying high density workloads in the Data Center

The Economy of Data Center Resources

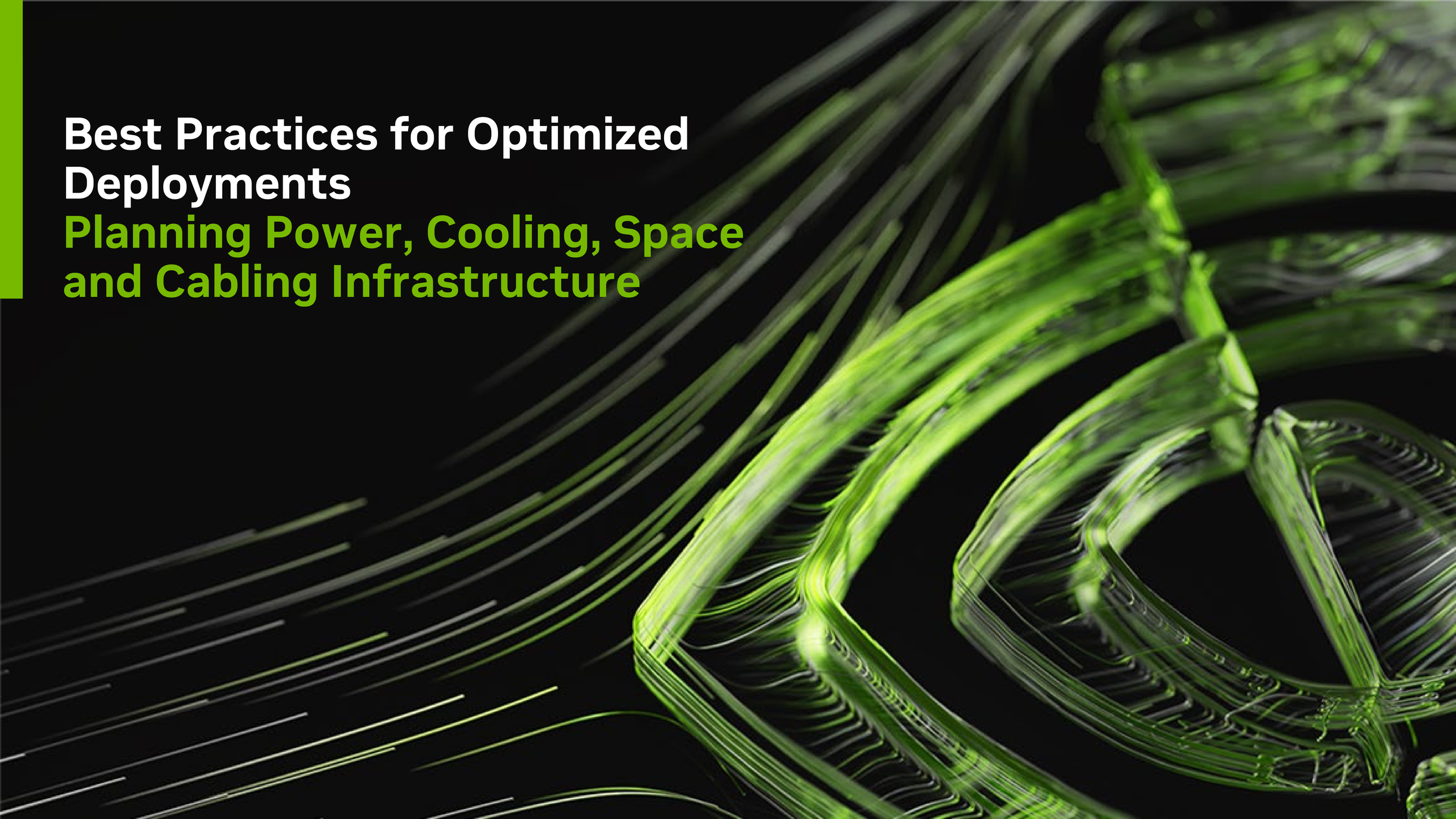
Deploying High Density Workloads in the Data Center



- Data Centers have finite resources
  - These limited resources create scarcity in the face of demand
  - This drives the need to optimize resource utilization for efficiency
- The three main resource constraints are
  - Power
  - Cooling
  - Space
- These resources are interrelated
  - Excess demand on one resource could negatively impact other resources
  - These interrelationships can also impact InfiniBand cable lengths

Best Practices for Optimized Deployments

Planning Power, Cooling, Space and Cabling Infrastructure



Planning Power and Cooling per Rack

DGX B200 Rack Power Profiles

- Power Capacity Constraints
  - N Deployment Pattern
    - N = 2 power circuits
    - Each circuit must be rated to handle 50% of the peak demand of the Rack, factoring in any required breaker deratings
  - The typical deployment pattern will be based on 2 systems per rack
  - Deployment patterns based on 4 systems per rack are an option at specialized data center sites that are designed for extreme density air cooled deployments
    - Inquire about NVIDIA's DGX-Ready Data Center Program to connect with a partner to host your DGX infrastructure
  - Supplemental cooling apparatus such as rear door heat exchangers and in-row coolers are not typically suitable for use with DGX B200 systems

Minimum Specifications	
4 B200 Systems Per Rack	2 B200 Systems Per Rack
57.2 kW	28.6kW
52U Rack	42U Rack**
2x 415V 3Φ 60A	2x 380V 3Φ 32A
8580 CFM*	4290 CFM*

*At Sea Level, based on estimated demand of 150CFM per kilowatt. Actual requirements will vary by site.
**52U Rack is recommended and shown.

Power Provisioning

Common distribution schemes compatible with DGX B200 racks

Phase	Distribution Voltage	Line Voltage	Amps	Breaker Derating	Rack Power Capacity kW* (Two Circuits Provisioned)	Maximum Supported DGX B200 Systems per Rack**	Peak Server Demand per Circuit kW**	Stranded Capacity at Peak Demand kW**
3Φ	380	219	32	80.0%	32	2	28.6	3.4
3Φ	400	230	32	80.0%	34	2	28.6	5.1
3Φ	415	240	32	80.0%	35	2	28.6	6.4
3Φ	415	240	60	80.0%	66	4	57.2	8.4
3Φ	415	240	63	80.0%	69	4	57.2	11.6
3Φ	380	219	32	100.0%	40	2	28.6	11.4
3Φ	400	230	32	100.0%	42	2	28.6	13.5
3Φ	415	240	32	100.0%	44	2	28.6	15.1
3Φ	415	240	60	100.0%	82	4	57.2	24.7
3Φ	415	240	63	100.0%	86	4	57.2	28.8

* 0.95 power factor.

** Based on a two circuit N power provisioning scheme.

Planning Power Deployments

With Traditional Dual Source/rPDU Power Provisioning

- A DGX B200 system includes 6 Power Supply Units (PSUs)
  - 5 of the 6 PSUs must be energized for the system to operate
  - The system can operate if a single internal power supply unit is de-energized, but will not operate if more than one power supply unit is de-energized, regardless of upstream power redundancies
    - This is a critical Data Center design consideration
- Due to the internal N+1 PSU configuration of the systems, traditional power provisioning redundancy models are not effective
  - *In this model, power provisioning to the rack should achieve N, and should be understood as delivering ONLY N, even if the provisioning is comprised of dual feeds from discrete sources*

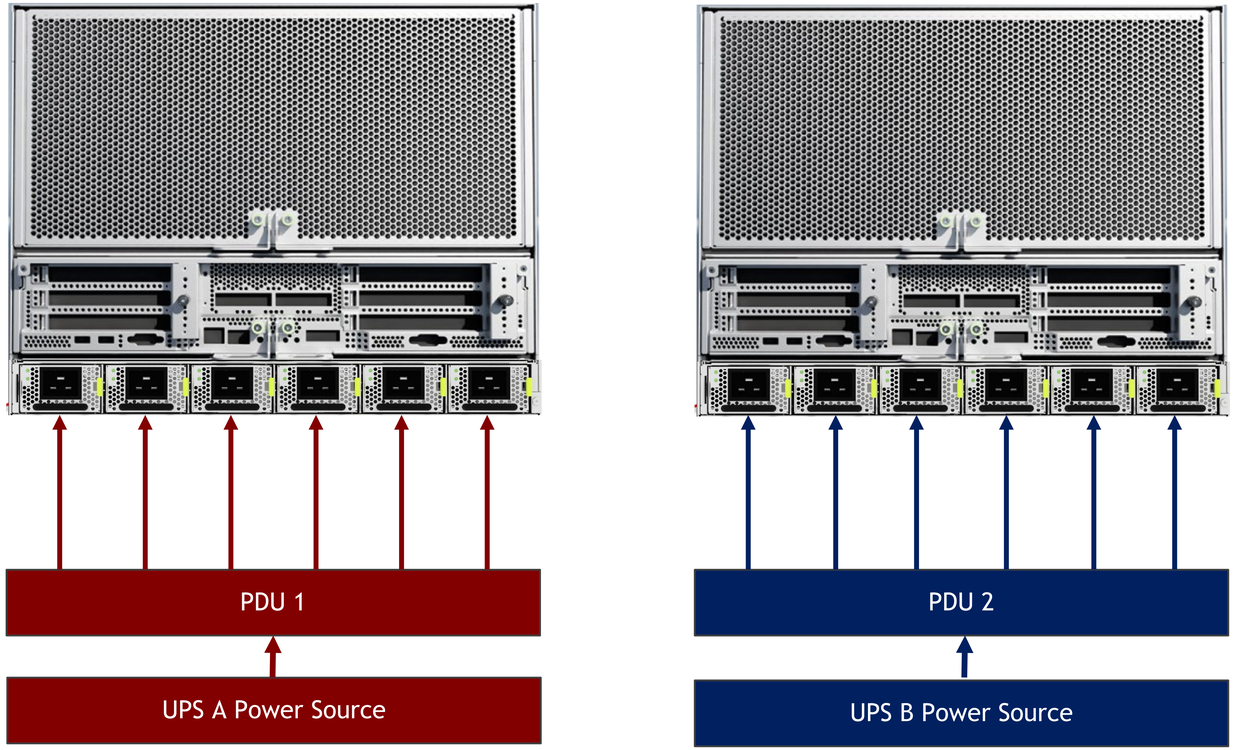
N Configuration with Two Discrete Power Sources

Creates the fewest interdependencies

A failure of either PDU will impact only half of the systems in the rack

A failure of either UPS will impact only half of the systems

This is the most optimal configuration when only two UPS power sources are available at the rack



Planning Power Deployments

With Single Source/rPDU Power Provisioning

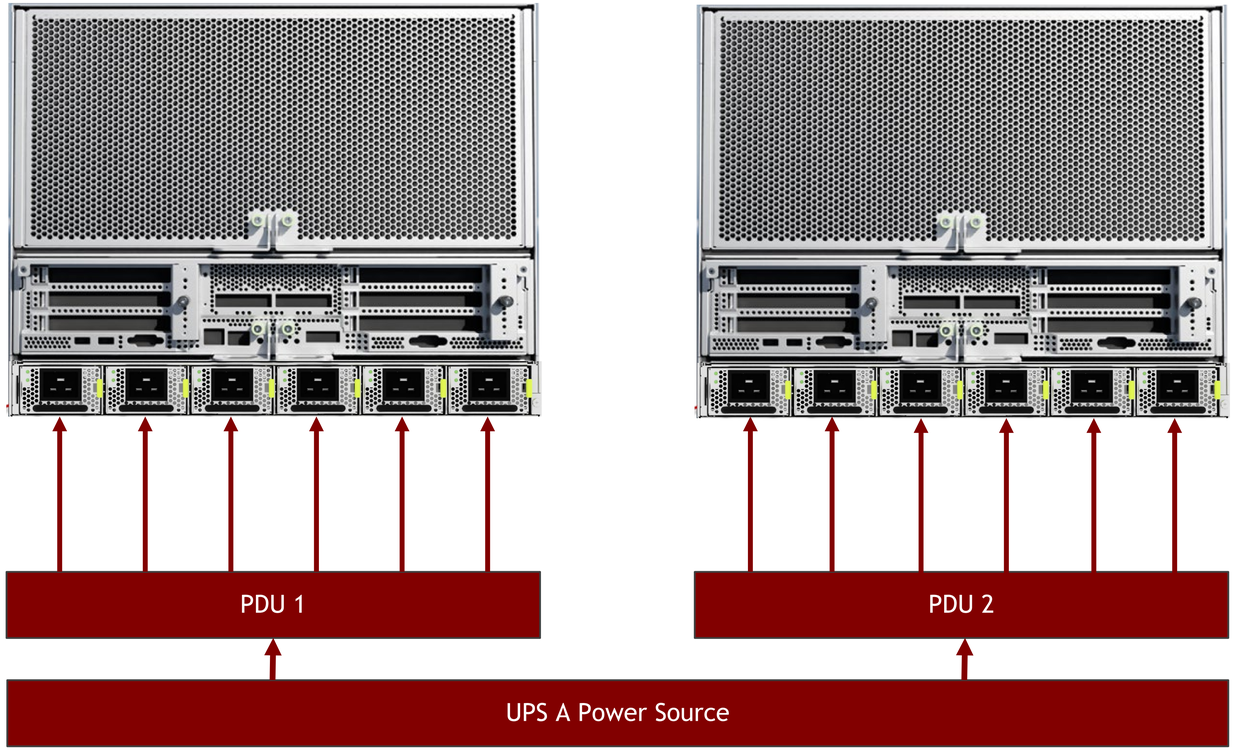
- A DGX B200 system includes 6 Power Supply Units (PSUs)
  - 5 of the 6 PSUs must be energized for the system to operate
  - The system can operate if a single internal power supply unit is de-energized, but will not operate if more than one power supply unit is de-energized, regardless of upstream power redundancies
    - This is a critical Data Center design consideration
- Due to the internal N+1 PSU configuration of the systems, traditional power provisioning redundancy models are not effective
  - *In this model, power provisioning to the rack should achieve N, and should be understood as delivering ONLY N, even if the provisioning is comprised of dual rack PDUs*

N Configuration with Single Power Source

A PDU level failure impacts only half of the systems in the rack

A UPS failure will affect all of the systems

Provisioning from a single UPS source increases the probability of disruption due to maintenance or failure events.



Planning Power Deployments

With Dual Interleaved Source/rPDU Power Provisioning

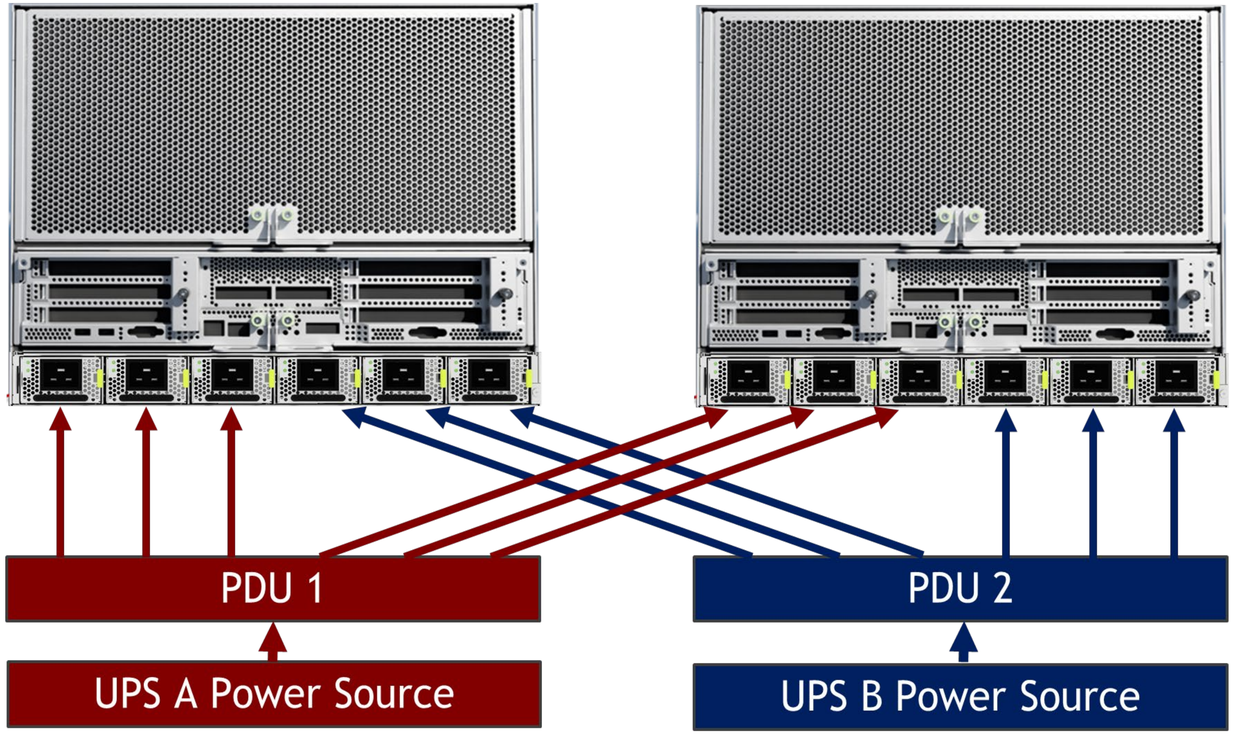
- A DGX B200 system includes 6 Power Supply Units (PSUs)
  - 5 of the 6 PSUs must be energized for the system to operate
  - The system can operate if a single internal power supply unit is de-energized, but will not operate if more than one power supply unit is de-energized, regardless of upstream power redundancies
    - This is a critical Data Center design consideration
- Due to the internal N+1 PSU configuration of the systems, traditional power provisioning redundancy models are not effective
  - *In this model, power provisioning to the rack should achieve N, and should be understood as delivering ONLY N, even if the provisioning is comprised of dual feeds from discrete sources*

N Configuration with Multiple Interleaved Power Sources

Creates the most interdependencies

A failure of either PDU will impact all of the systems in the rack

A failure of either UPS will impact all of the systems



Due to the N+1 PSU design, interleaving power from two sources to a single system increases the probability of disruption rather than reducing it. This model is not recommended.

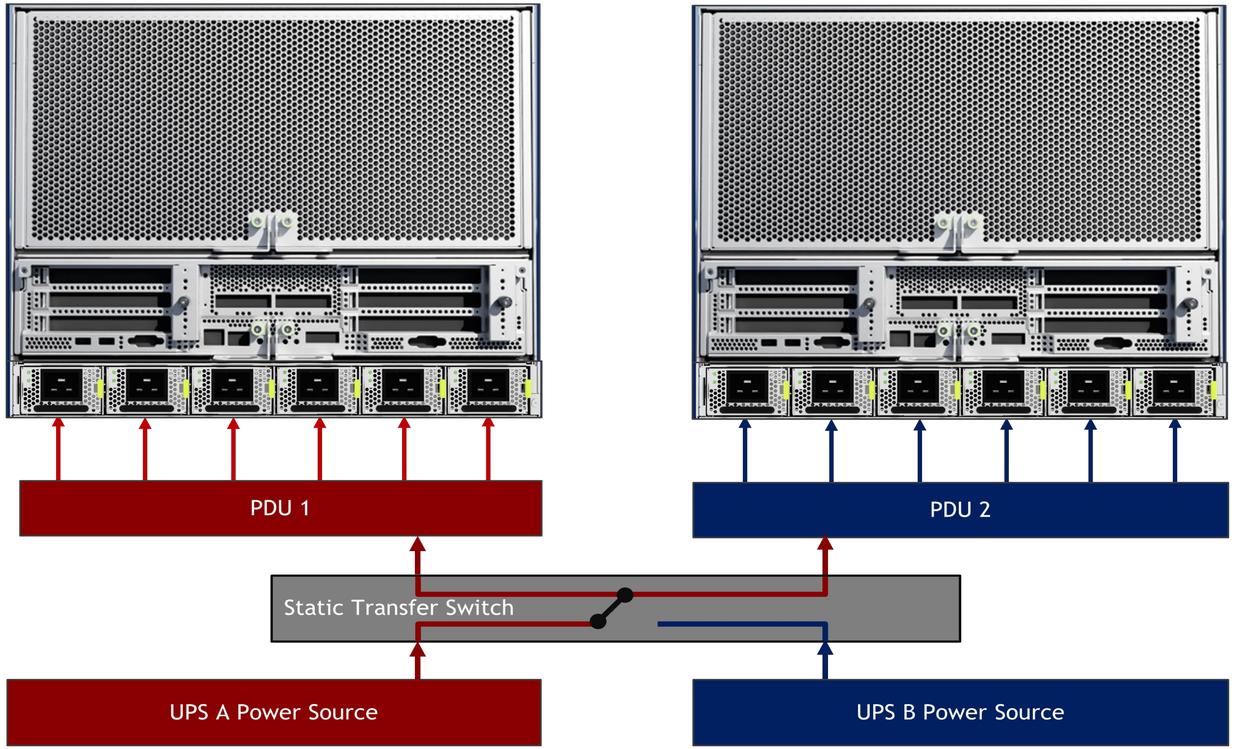
Planning Power Deployments

N+1 Configuration with Dual UPS power sources and Static Transfer Switch

- A DGX B200 system includes 6 Power Supply Units (PSUs)
  - 5 of the 6 PSUs must be energized for the system to operate
  - The system can operate if a single internal power supply unit is de-energized, but will not operate if more than one power supply unit is de-energized, regardless of upstream power redundancies
    - This is a critical Data Center design consideration
- Due to the internal N+1 PSU configuration of the systems, more complex power provisioning models are necessary if high availability is a system design priority
- Static Transfer Switches (STS) will power the rack from a single primary source, then switch to an alternate source if the primary source is disrupted.
- STS systems are generally considered a stopgap solution
- In-Rack STS systems are not currently supported

N+1 Configuration with Static Transfer Switch

Intended to improve availability
Static Transfer Switch becomes single point of failure
A failure of any single PDU will impact only one system
A failure of any single UPS source will not impact the systems
May increase stranded power
Adds complexity and cost to power architecture
May create load imbalance on upstream UPS power paths



NVIDIA has not tested or validated any static transfer switch product for compatibility with DGX B200 systems and does not currently recommend this type of solution. It is possible that switchover events may cause system damage, adverse system behavior, or workload failures.

Planning Power Deployments

N+1 Configuration with “6 to make 5” power sources

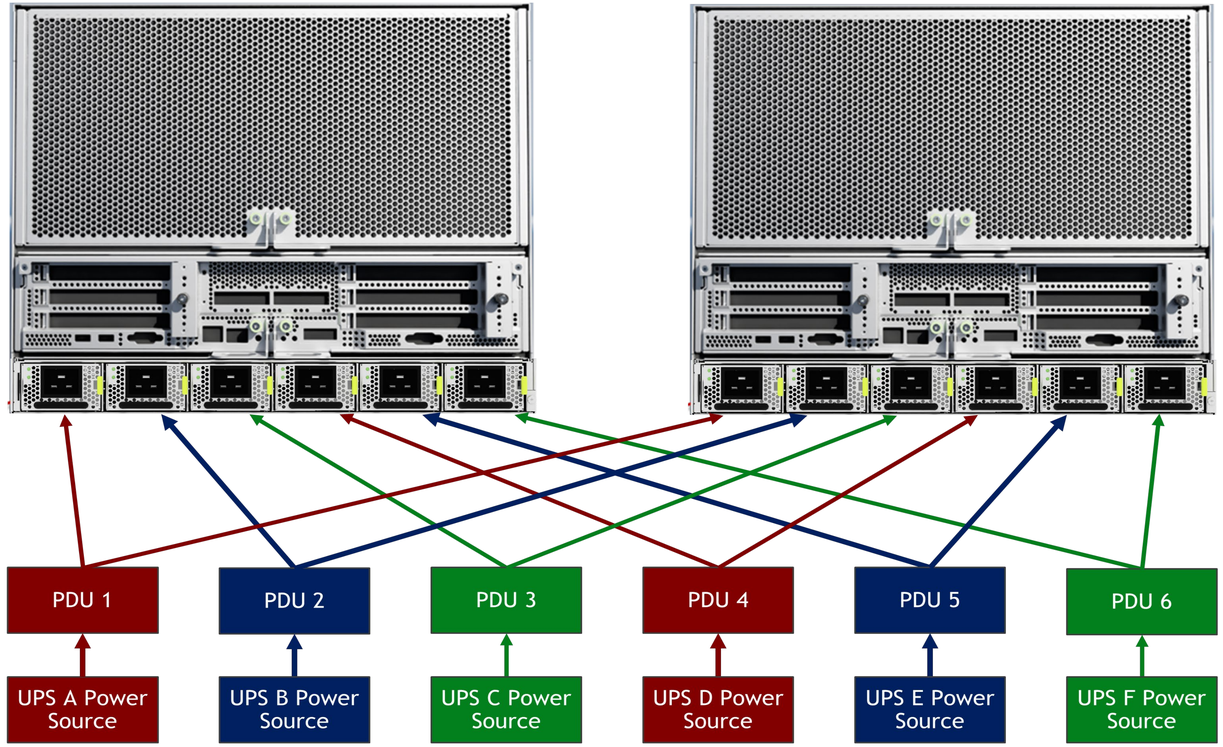
- A DGX B200 system includes 6 Power Supply Units (PSUs)
  - 5 of the 6 PSUs must be energized for the system to operate
  - The system can operate if a single internal power supply unit is de-energized, but will not operate if more than one power supply unit is de-energized, regardless of upstream power redundancies
    - This is a critical Data Center design consideration
- Due to the internal N+1 PSU configuration of the systems, more complex power provisioning models are necessary if high availability is a system design priority

N+1 Configuration with “6 to make 5” Power Sources

Provides higher availability than dual sourced models
Creates no interdependencies
A failure of any single PDU will not impact the systems
A failure of any single UPS source will not impact the systems
May increase stranded power
Adds complexity and cost to power architecture
Many sites are not equipped to provision 6 discrete UPS sources to the rack

This is the optimal configuration for high availability

Due to rack space constraints, this option may be limited to rack densities of 2 systems per rack



Strategies for Cooling Optimization and Scalability

Environmental Parameters

- Each DGX B200 system requires up to approximately 2145 CFM of supply air at sea level
  - Supply air should generally comply with ASHRAE Class A1 specifications, but with a maximum supply air temperature of 30°C (86°F).
- For optimal cooling efficiency and system performance, air contamination should be mitigated through filtration
  - Minimum of MERV 13 (or EN779-2012 M6/F7, or ISO 16890 ePM1-50%) rated filter
  - Rack area should meet the cleanliness level of the ISO 14644-1 Class-8 standard
    - Maximum particulate counts not to exceed 3,520,000 @ 0.5 µm/m3 for longer than 15 minutes.

ASHRAE specifications

Range	Class	Dry-Bulb Temperature	Humidity Range, Non-Condensing	Maximum Dew Point
Recommended	All A	64.4–80.6 °F 18–27 °C	41.9 °F to 60% RH and 59 °F DP 5.5 °C to 60% RH and 15 °C DP	59 °F 15 °C
Allowable up to 30 °C for DGX B200 Systems	A1	59–89.6 °F 15–32 °C	20–80% RH	62.6 °F 17 °C
Allowable per ASHRAE for various other classes of data center and telecom environments	A2	50–95 °F 10–35 °C	20–80% RH	69.8 °F 21 °C
	A3	41–104 °F 5–40 °C	10.4 °F DP and 8–85% RH -12 °C DP and 8–85% RH	75.2 °F 24 °C
	A4	41–113 °F 5–45 °C	10.4 °F DP and 8–90% RH -12 °C DP and 8–90% RH	75.2 °F 24 °C
	B	41–95 °F 5–35 °C	8–80% RH	82.4 °F 28 °C
	C	41–104 °F 5–40 °C	8–80% RH	82.4 °F 28 °C

ISO 14644-1 standard for air cleanliness classifications

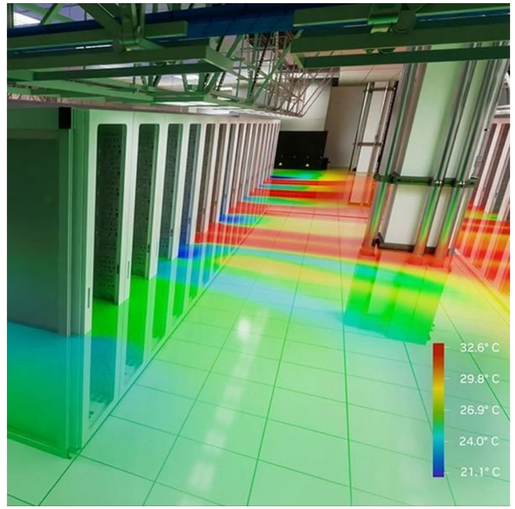
Class	Particle Size[1]					
	> 0.1 µm	> 0.2 µm	> 0.3 µm	> 0.5 µm	> 1 µm	> 5 µm
1	10	2				
2	100	24	10	4		
3	1,000	237	102	35	8	
4	10,000	2,370	1,020	352	83	
5	100,000	23,700	10,200	3,520	832	29
6	1,000,000	237,000	102,000	35,200	8,320	293
7				352,000	83,200	2,930
8				3,520,000	832,000	29,300
9				35,200,000	8,320,000	293,000

Uncertainties related to the measurement process require that data with no more than three significant figures be used in determining the classification level.

Strategies for Cooling Optimization and Scalability

Data Center Best Practices

Aisle Containment Systems



Computational Fluid Dynamics

- Start with a solid foundation
  - Computational Fluid Dynamics (CFD) Modeling and Analysis
  - Aisle Containment
  - Blanking Panels
  - Brush Grommets
  - Humidification / Dehumidification
  - Proper air filtration
  - Perforated Floor Tile flow rate audits with anemometer

Blanking Panels

Brush Grommet

Perforated Floor Tile Flow Rate Analysis

Cabling

DGX Infrastructure Support

SU Count	Cluster Size # of nodes	Cluster Size # of GPUs	Leaf Switch Count	Spine Switch Count	Compute + UFM Node Cable Count	Spine-Leaf Cable Count
1	31	248	8	4	252	256
2	63	504	16	8	508	512
3	95	760	24	16	764	768
4	127	1016	32	16	1020	1024

- There is a large concentration of cables in the central Management Racks and in the supporting cable pathways overhead. Pre-planning for the cabling infrastructure is a critical part of the design process.
- Areas of analysis include:
  - Top of rack cable penetration access
  - Cable tray types
  - Cable tray width and depth
  - Cable tray fill ratios
  - Structural load/weight planning
  - Cable management devices

Cabling

DGX Cabling Types

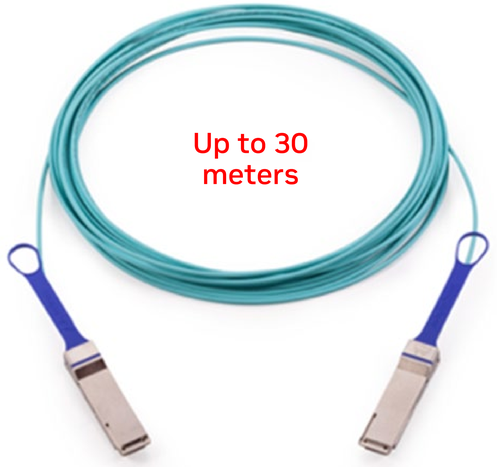
Passive DAC (Direct Attached Copper) Cabling

A high-speed twinaxial (two shielded conductors) electrical cable with a SFP or QSFP connector on each end. The term ‘passive’ means that there is no active signal processing circuitry or amplification of the electrical signal built into the cable itself.

Passive DAC Cables are intended for shorter distances of up to 5 meters when being used for InfiniBand networks.

AOC (Active Optical Cables)

Active Optical Cables are comprised of two components, a multimode or single mode fiber optic cable, and an SFP or QSFP active optical transceiver on each end. Unlike traditional optical transceivers and cables, AOC cable components are permanently bonded together, creating a complete cable assembly. cable.

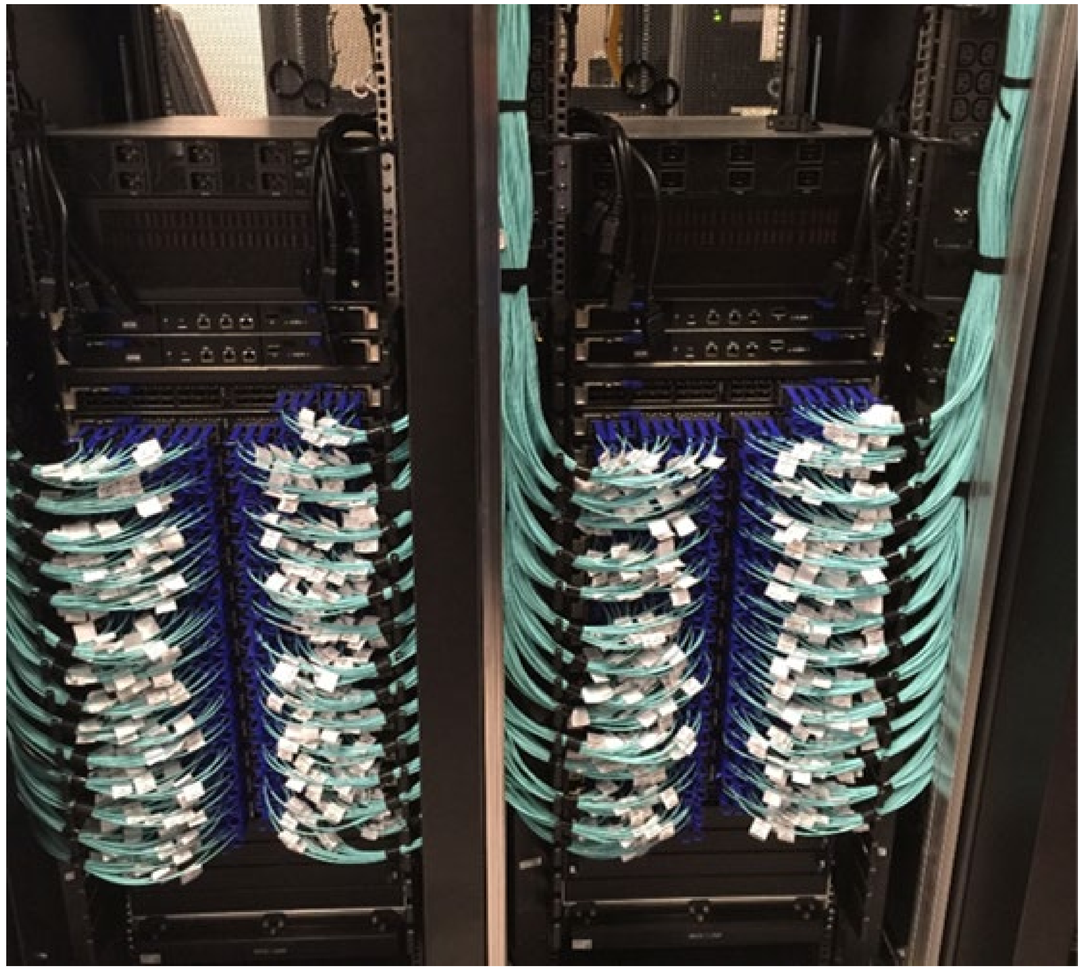


Active Optical Cables are used for longer distances of up to 30 meters when being used for InfiniBand networks.

Cabling

DGX Infrastructure Support

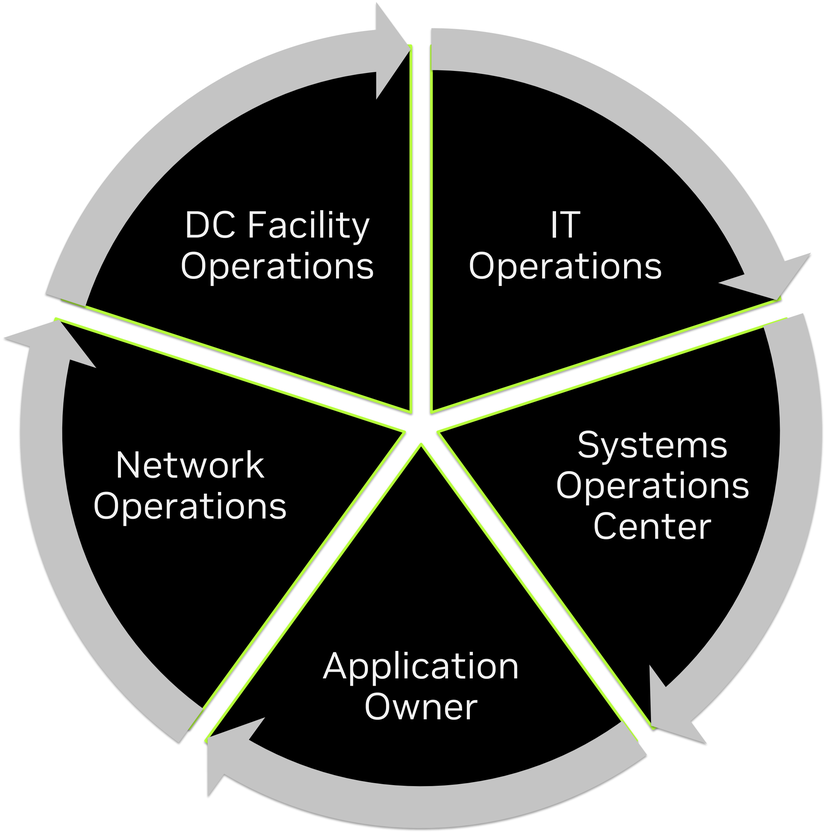
- Bundle cables together in groups of relevance (for example, ISL (Inter-Switch Link) cables and uplinks to core devices) to ease management and troubleshooting.
- Use cables of correct length.
- Allow for bend radiuses and intra-rack cable routing in cable length calculations
- Keep cable runs under 90% of the max distance supported for each media type, as specified in the applicable standard.
- Keep copper and fiber bundles and runs separated.
- Install spare cables in advance for future replacement of damaged cables.
- Use color coding of the cable ties. The colors should indicate the endpoints. Place labels at both ends, as well as along the run.

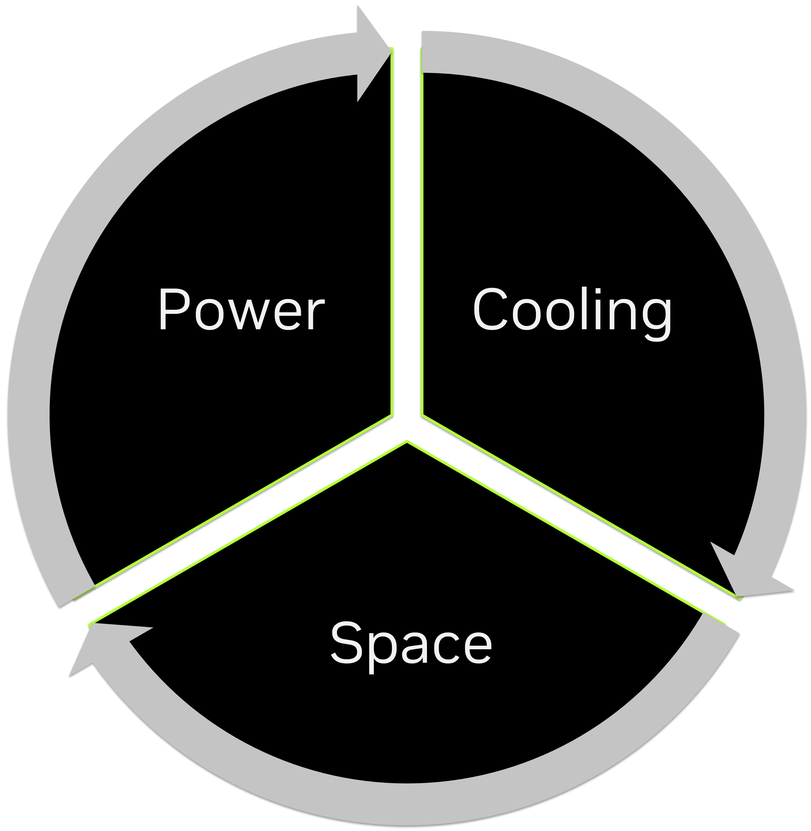


Example of proper cable management

Summary

Align domain owners and plan across a budget of data center resources





Frequently Asked Questions

1. Does Nvidia specify certain cabinet dimensions or parameters?
    A. Yes. While Nvidia does not specify a particular brand or model of rack to be used, it does specify that Racks must conform to EIA-310 standards for enclosed racks with 19" EIA mounting. Cabinets must be at least 24" x 44" (600 mm × 1100 mm) in size, and at least 48U tall. NVIDIA recommends 30” x 48” x 52 U (700 mm × 1200 mm) racks.
2. Can InfiniBand cable length limitations be extended using patch panels or top of rack switches, or using patch panels for structured cabling?
    A. No. InfiniBand is an extremely high-performance architecture, and its cable length limitations are based on signal attenuation and latency of the entire signal path, not just a segment of cable. Patch panels exacerbate signal attenuation, and intermediary top of rack switches add latency.
3. Does Nvidia specify certain rack PDU types?
    A. Yes. Since rack PDUs must conform to the power provisioning available in each data center, and the brands available in each market region, NVIDIA does not specify a particular brand or model. However, we recommend the NVIDIA SuperPOD design include rPDUs which have an Integrated smart module, Network Interface, RestAPI interface, Ports for temperature and sensor probes, Locking receptacles, PDU level Metering, Remote outlet switching per receptacle, and optionally Red and Blue exterior colors for visual circuit differentiation.
4. How many Scalable Units can I deploy?
    A. A typical SuperPOD can contain up to 4 Scalable Units, the reference architecture is designed for up to 64 scalable units.
5. Can I utilize empty space in DGX racks for other IT Equipment?
    A. No. A DGX Scalable Unit is an engineered solution that should not cohabitate with unrelated equipment in a shared rack. Additionally, If racks are spaced apart to aggregate cooling capacity, no unrelated equipment should be deployed in the spaces between DGX B200 Racks


NVIDIA