

High Density Reference Design for AI

AI Reference Design #003

360kW Cluster at up to 41kW per rack air cooled

A full data center environment designed for a 360kW solution with 2N redundancy, using NVIDIA DGX H100 System.

Design Principles

- 1. Design both Power and Cooling together** to optimize AI infrastructure.
- 2. Power is at a premium.** Eliminate stranded power by aligning AI clusters to data center 'capacity blocks.'
- 3. Consider total cost of ownership, redundancy, and blast radius** in AI power and cooling designs, and the tradeoffs among them.
- 4. Handle AI workload surges** though system-level controls and power and cooling buffers.
- 5. Design for a mix of liquid and air cooling,** which have a combined and interdependent impact on the ability to remove heat.
- 6. Design for the future.** AI power density will rapidly grow toward 500kW per rack.



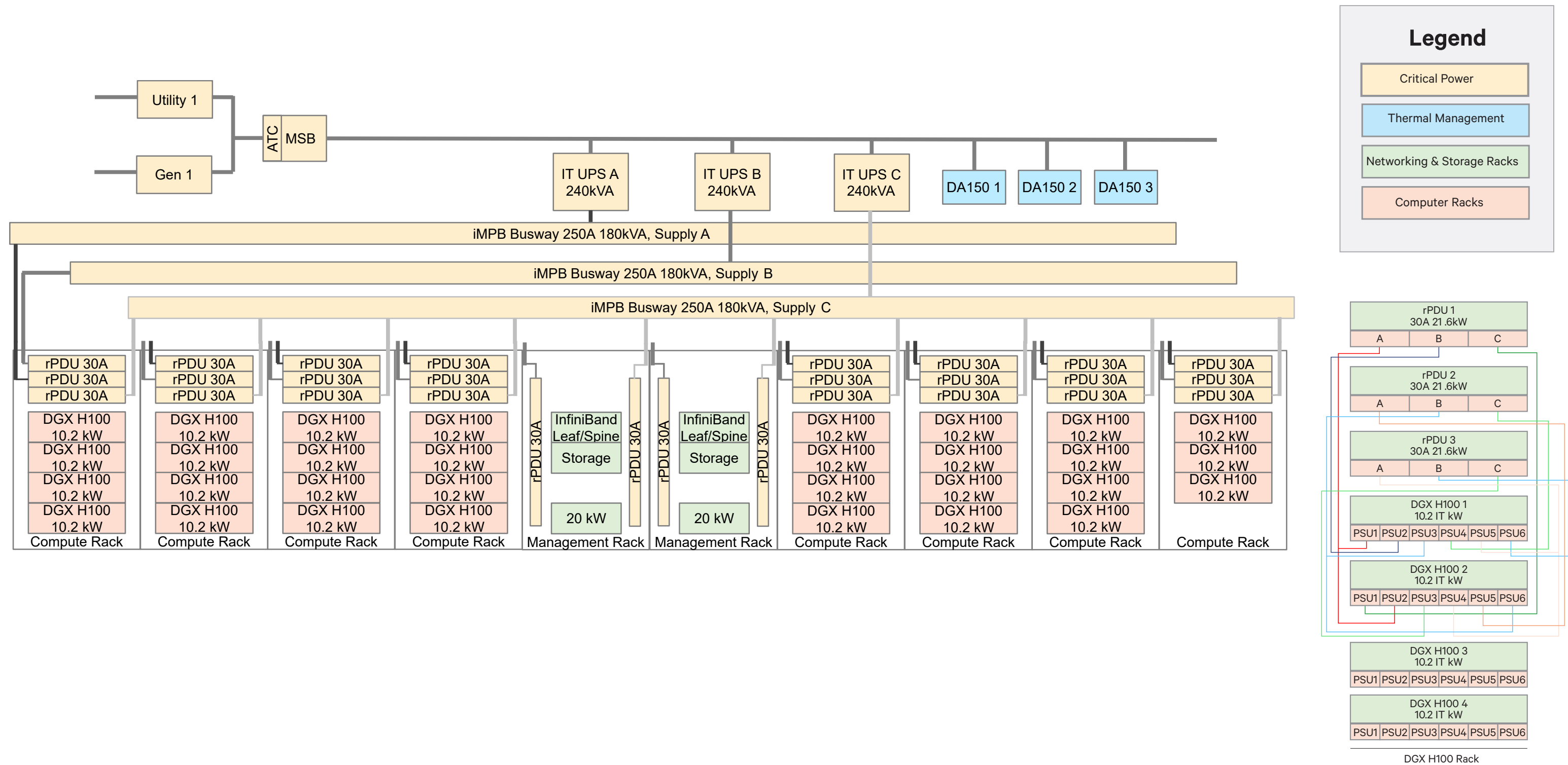
image is for representation only

Design Summary

Solution Capacity	360kW
Redundancy	Enhanced N+1 for power, N+1 for cooling
Total IT Load	358kW
Rack Density (IT load)	7 racks at 41kW, 1 rack at 30kW, 2 racks at 20kW
Rack(s)	10
Cooling Topology	100% Air (perimeter units)
Perimeter air cooling	DA150
Rack PDU(s)	30A horizontal rack PDU, 3 per compute rack 30A vertical rack PDU, 2 per management rack
Busway	250A busway, 3 per POD 30A tap-off boxes, 3 per rack
UPS for IT	240kVA UPS, Enhanced N+1, 3 per POD



Reference Design Schematic



References: <https://www.nvidia.com/en-us/data-center/dgx-superpod-gb200/>

Vertiv Components of the Design

Component	Details
UPS for IT	3x 240kVA Enhanced N+1 UPS for POD
Busway and Tap-off Boxes (TOB)	3x 250A busway for POD, 3x 20A TOB per compute rack, 2x 30A TOB per management rack
Rack PDUs for compute racks	3x 30A horizontal rack PDUs per rack
Rack PDUs for management racks	2x 30A vertical rack PDUs per rack
Air Cooling Unit	3x Air-cooling unit for POD
Rack Enclosures – Compute	8x 48U 800×1200
Rack Enclosures – Management	2x 48U 800×1200