SYMBEEO

PUBLIC

COMPARISON

AI Inference: One Rack or Many Servers?

Published comparison

CAPABILITY	DOCUMENT	VERSION	DATE
AI & Data	CMP-AI-001	1.2	Nov 6, 2026

WRITTEN BY

Diksha Shah, Principal AI Infrastructure Architect

© 2026 Symbeeo. All rights reserved.

What this document is. Six ways to serve a bank's AI models inside its own hall, held to one written requirement: 3,000 internal users, four production models, every request kept in the hall, the largest model whole in one memory domain. Three GB300 NVL72 racks meet it, each with a condition. Two eight-GPU servers do not. One, NVIDIA's next rack, draws more power than the hall allows.

Document control

Document	CMP-AI-001	Version	1.2
Issued	Nov 6, 2026	Status	Issued by Symbeeo
Classification	Public	Supersedes	1.1, Nov 4, 2026
Capability	AI & Data	Author	Diksha Shah
Role	Principal AI Infrastructure Architect		
Credentials	NVIDIA Certified Professional: AI Operations, NVIDIA Certified Associate: AI Infrastructure and Operations, Google Cloud Professional Cloud Architect, AWS Certified Data Engineer		
Approval	Version 1.0 issued by Symbeeo, Oct 24, 2026. Version 1.1 issued by Symbeeo, Nov 4, 2026. Version 1.2 issued by Symbeeo, Nov 6, 2026. Every mark carries its source; prices are not published.		

Contents

1 · The requirement 3
2 · The options 4
3 · Conformance, clause by clause 5
4 · The hardware criteria 6
5 · Price and due diligence 6
6 · The verdict 6
7 · Sources 7
8 · The assessor 7

© 2026 Symbeeo. All rights reserved. · Public

CONVENTIONS

Marks. MET the vendor's own current document shows the clause is met · **MET WITH A CONDITION** met only if the named condition is confirmed at the quote · **OPEN** the documents read do not settle it; the reason is given · **NOT MET** the vendor's own figure falls short.

Source tags. [S] read with curl from the vendor's own current page or document on October 24, 2026; N6 on November 6, 2026 · [S*] from a summary of the vendor's page where the page itself could not be opened; provisional, never settles a mark · [A] an inference, stated as one · [T] a figure set by the requirement. No price, discount or price figure appears in this document.

Table 1. Revision history

Date	Version	Change
Oct 24, 2026	1.0	First issue.
Nov 4, 2026	1.1	Set-aside table corrected: Supermicro and Lenovo listed separately.
Nov 6, 2026	1.2	Option F corrected on R-2 from OPEN to NOT MET. Version 1.1 said NVIDIA had not published the rack power; NVIDIA's datasheet, created October 15, 2026, rates the rack at 156 to 227 kW, above the 142 kW clause. No other mark changed.

1 · The requirement

The requirement comes from Symbeeo's design study Inference cluster on GB300 NVL72 for a bank's document models. It names no product. Each clause names the option expected to split on it.

Before any clause: the hall. No inference request or response leaves the bank's hall. Any option that serves from outside the hall is set aside at stage 1. [T]

Table 2. The requirement

Clause	The buyer's words	Expected to split
R-1	The largest model runs whole inside one GPU memory domain of at least 20 TB, with no inference traffic between servers, and each of the two racks can carry all four models alone. [T]	Eight-GPU servers, whose memory domain ends at the server
R-2	Each rack stays within 142 kW and is cooled by the facility water already in the hall. [T]	Air-cooled servers; any rack above 142 kW or whose power is not published
R-3	Each GPU has its own network link of at least 800 Gb/s to the bank's fabric. [T]	Servers whose network cards are chosen separately
R-4	One supplier supports the whole system, hardware and software, as one case. [T]	Systems whose software support runs shorter than the hardware's

The criteria used. Conformance to the requirement comes first. Beside it, the hardware set: rack fit, power draw, lead time and availability (section 4). Five-year cost is a supporting column, never a criterion.

© 2026 Symbeeo. All rights reserved. · Public

2 · The options

Assessed, all from makers on Symbeeo's line card:

Table 3. The options

Option	System as assessed	Why it is in
A	Dell PowerEdge XE9712: NVIDIA GB300 NVL72 in Dell Integrated Rack Scalable Systems	The system the design study is drawn on
B	NVIDIA GB300 NVL72 by HPE	The same NVIDIA rack from a second maker
C	NVIDIA DGX GB300	NVIDIA's own rack system
D	Cisco UCS C880A M8: eight NVIDIA HGX B300 GPUs per server	The many-servers answer, air-cooled
E	Dell PowerEdge XE9785L: eight AMD Instinct MI355X GPUs per server	The many-servers answer on AMD, liquid-cooled
F	NVIDIA Vera Rubin NVL72	NVIDIA's next rack, now in production

Set aside or not assessed, with the reason:

Table 4. The options

Platform	Reason
AWS, Google Cloud	Stage 1: inference would leave the hall
Equinix	Stage 1: colocation is not the bank's hall
AMD's rack-scale platform (Helios)	Not assessed: amd.com returned errors on four tries, October 22 to 24, 2026; read when AMD's own page answers
Supermicro, Lenovo	Not assessed in this edition
Makers not on the line card (for example Gigabyte)	Not assessed; Symbeeo assesses what it can supply and support

© 2026 Symbeeo. All rights reserved. · Public

3 · Conformance, clause by clause

Each cell gives the mark, the fact and its source key. Keys resolve in section 7.

Table 5. Conformance, clause by clause

Clause	A · Dell XE9712 GB300 NVL72	B · HPE GB300 NVL72	C · NVIDIA DGX GB300	D · Cisco C880A M8	E · Dell XE9785L MI355X	F · NVIDIA Vera Rubin NVL72
R-1 one memory domain, 20 TB	**MET.** 72 Blackwell Ultra GPUs and 36 Grace CPUs per rack [S, D2]; 20 TB of GPU memory in one NVLink domain [S, N1]	**MET.** 72 Blackwell Ultra GPUs and 36 Grace CPUs on fifth-generation NVLink [S, H1]; 20 TB of GPU memory in the GB300 NVL72 [S, N1]	**MET.** 72 GPUs, 20 TB GPU memory, 37 TB fast memory [S, N3]	**NOT MET.** Eight HGX B300 GPUs per server [S, C1]; HGX B300 total memory 2.1 TB [S, N4]	**NOT MET.** Eight MI355X GPUs of 288 GB each [S, D3]: 2.3 TB per server [A, derived]	**MET.** 72 Rubin GPUs, 20.7 TB of HBM4 [S, N5]
R-2 within 142 kW, facility water	**MET.** GB300 NVL72 rack: up to 142 kW, eight 33 kW power shelves [S, N2]; XE9712 direct liquid cooling [S, D1]. At the budget, with no margin	**MET.** Eight 33 kW power shelves; cold plates on CPUs and GPUs, HPE direct liquid cooling [S, H1]; up to 142 kW for the GB300 NVL72 rack [S, N2]	**MET.** Liquid-cooled rack [S, N3]; up to 142 kW for the GB300 NVL72 rack [S, N2]	**OPEN.** Air-cooled [S, C1]; twelve 3.2 kW supplies per server [S, C1]. The hall's air capacity is not stated	**OPEN.** Direct-to-chip liquid cooling, 33 kW power shelves [S, D3]; rack total not published on the page read	**NOT MET.** Liquid-cooled, 45 °C inlet [S, N5]; NVIDIA's datasheet rates the rack at MaxP 227 kW, MaxQ 187 kW, MaxLPS 156 kW [S, N6], each above 142 kW
R-3 800 Gb/s per GPU	**MET WITH A CONDITION.** ConnectX-8 gives "800 gigabits per second" to each GPU in the GB300 NVL72 [S, N1]; Dell's page lists no network card, so confirm the module at the quote [A]	**MET.** Two ConnectX-8 IO modules per four-GPU tray, "2x NVIDIA ConnectX-8 (800 Gbps)" [S, H1]	**MET.** 72 ConnectX-8 ports at 800 Gb/s [S, N3]	**MET WITH A CONDITION.** Eight ConnectX-8 on the GPU board [S, C1]; 800 Gb/s per port inferred [A]	**OPEN.** Up to eight PCIe slots for east-west cards [S, D3]; speed depends on the cards chosen	**MET.** 0.45 TB/s per GPU, bi-directional [S, N5]: 1.8 Tb/s each way [A, derived]
R-4 one supplier, hardware and software	**MET WITH A CONDITION.** Dell ProSupport covers "hardware and software" [S, D1]; confirm hours at the quote	**MET WITH A CONDITION.** Three years of hardware repair, 90 days of software support; longer support through HPE Services [S, H1]. Put it on the order	**MET WITH A CONDITION.** "Three-year business-standard hardware and software support" [S, N3]; confirm hours at the quote	**MET WITH A CONDITION.** Three-year hardware warranty, 90-day software warranty; Cisco Success Tracks augment it [S, C1]. Put it on the order	**MET WITH A CONDITION.** Dell ProSupport, hardware and software [S, D3]	**OPEN.** No carried maker's offer was read

© 2026 Symbeeo. All rights reserved. · Public

Clause	A · Dell XE9712 GB300 NVL72	B · HPE GB300 NVL72	C · NVIDIA DGX GB300	D · Cisco C880A M8	E · Dell XE9785L MI355X	F · NVIDIA Vera Rubin NVL72
Stage 2 standing	**Conforms, with a condition:** network module and support hours	**Conforms, with a condition:** software support on the order	**Conforms, with a condition:** support hours	**Set aside at stage 2:** R-1 not met	**Set aside at stage 2:** R-1 not met	**Not passed to stage 3:** over 142 kW

4 · The hardware criteria

Table 6. The hardware criteria

Criterion	A · Dell XE9712	B · HPE	C · NVIDIA DGX	D · Cisco C880A M8	E · Dell XE9785L	F · Vera Rubin
Rack fit	48RU IR9048 rack; each sled 1RU [S, D1]	48U rack-scale system [S, H1]	NVIDIA MGX rack [S, N3]	10RU per server [S, C1]	IR7000 rack, 44 OU; 3 OU per node [S, D3]	OPEN
Power draw	Up to 142 kW per rack [S, N2]	Eight 33 kW shelves [S, H1]; up to 142 kW [S, N2]	Up to 142 kW per rack [S, N2]	Twelve 3.2 kW supplies, N+N [S, C1]	1,400 W per GPU [S, D3]	156 to 227 kW per rack [S, N6]
Lead time	OPEN: from distributor data at stage 4	OPEN	OPEN	OPEN	OPEN	OPEN
Availability	Dell shipped its first GB300 NVL72 racks in July 2025 [S, D2]; allocation checked at stage 4	OPEN	OPEN: checked at stage 4	HGX B300 "shipping now" [S, N4]; allocation at stage 4	OPEN	Shipping to AI labs, clouds and hyperscalers [S, N4]; enterprise allocation OPEN

5 · Price and due diligence

Stage 3, price. Price is weighed only among the three options that passed stage 2: Dell's XE9712 racks, HPE's GB300 NVL72 and NVIDIA's DGX GB300. No price figure is published here. They are not ranked: the buyer's own quote for two racks, their networking and their support decides between them. The other three are not evaluated on price; cost never rescues an option that has not met the requirement.

Stage 4, due diligence. Lead time and allocation for two racks are checked with authorized distribution before the order. Nothing in this document states a delivery date.

6 · The verdict

One rack, not many servers. Three GB300 NVL72 racks meet the requirement: Dell's XE9712 (confirm the network module and support hours at the quote), HPE's (longer software support on the order) and NVIDIA's DGX GB300 (confirm support hours). The buyer's quote decides between them.

What splits the field is the memory domain. An eight-GPU server holds about 2 TB in one domain, whether NVIDIA (Cisco's C880A M8) or AMD (Dell's XE9785L). The bank's largest model needs 20 TB in one domain, so both are set

© 2026 Symbeeo. All rights reserved. · Public

aside, whatever their price.

One draws too much power. NVIDIA's Vera Rubin NVL72 meets the memory and network clauses, but NVIDIA's datasheet rates the rack at 156 to 227 kW, above the hall's 142 kW, so it does not go on to price.

7 · Sources

Read with curl under a user agent that names Symbeeo and a contact address, one request at a time, no sign-in, on October 24, 2026, with retries after pauses; N6 on November 6, 2026. The date in brackets is the document's own.

- **N1** NVIDIA GB300 NVL72 product page and specifications. https://www.nvidia.com/en-us/data-center/gb300-nvl72/
- **N2** NVIDIA NVL72 AI Factory enterprise reference architecture, system hardware and components (September 15, 2026). https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html
- **N3** NVIDIA DGX GB300 product page and specifications. https://www.nvidia.com/en-us/data-center/dgx-gb300/
- **N4** NVIDIA HGX platform page and specifications. https://www.nvidia.com/en-us/data-center/hgx/
- **N5** NVIDIA Vera Rubin NVL72 product page and specifications. https://www.nvidia.com/en-us/data-center/vera-rubin-nvl72/
- **N6** NVIDIA Vera Rubin datasheet, linked from that page (marked October 2026; file created October 15, 2026). https://dam-cdn.nvd.orangelogic.com/AssetLink/56p68o47y6f1yucpubl0ump1c0aif422.pdf
- **D1** Dell PowerEdge XE9712 product page and technical specifications. https://www.dell.com/en-us/shop/ipovw/poweredge-xe9712
- **D2** Dell blog, first NVIDIA GB300 NVL72 delivery (July 3, 2025). https://www.dell.com/en-us/blog/dell-delivers-market-s-first-nvidia-gb300-nvl72-to-coreweave/
- **D3** Dell PowerEdge XE9785L product page and technical specifications. https://www.dell.com/en-us/shop/ipovw/poweredge-xe9785l
- **H1** HPE QuickSpecs, NVIDIA GB300 NVL72 by HPE (version 3, September 8, 2026), read through HPE's public support document service. https://support.hpe.com/hpesc/public/api/document/a50009244enw
- **C1** Cisco UCS C880A M8 Rack Server data sheet (updated August 13, 2026). https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c880a-m8-rack-server-ds.html

Tried and not read. hpe.com timed out on four tries from October 22 to 24, so HPE's QuickSpecs were read through HPE's support document service. amd.com returned errors on four tries, including the MI355X and MI400 pages; AMD's rack-scale platform stays unassessed until AMD's own page can be read.

8 · The assessor

WRITTEN BY

Diksha Shah

Principal AI Infrastructure Architect

NVIDIA Certified Professional: AI Operations, NVIDIA Certified Associate: AI Infrastructure and Operations, Google Cloud Professional Cloud Architect, AWS Certified Data Engineer

Issued by Symbeeo.

© 2026 Symbeeo. All rights reserved. · Public

SYMBEEO

Each vendor's tool shows you their option. We show you all of them against one requirement.

Compared. Then Delivered.

hello@symbeeo.com · symbeeo.com

© 2026 Symbeeo. All rights reserved.

Product and company names and logos are their owners' trademarks, shown only to identify the lines Symbeeo sources. Naming a vendor or showing its logo does not mean the vendor sponsors or approves Symbeeo's comparisons or any recommendation Symbeeo makes.

Figures are as published by their sources on the dates given in this document; figures marked [A] are assumptions.

CMP-AI-001 · Version 1.2 · Nov 6, 2026