SYMBEEO

PUBLIC

DESIGN STUDY

Inference cluster on GB300 NVL72 for a bank's document models

Complete design on current platforms

CAPABILITY	INDUSTRY	SETTING
AI & Data	Financial Services	Bank
DOCUMENT	**VERSION**	**DATE**
AI-002	1.3	Oct 10, 2026

WRITTEN BY

Diksha Shah, Principal AI Infrastructure Architect

© 2026 Symbeeo. All rights reserved.

Document control

Document	AI-002	Version	1.3
Issued	Oct 10, 2026	Status	Issued by Symbeeo
Classification	Public	Supersedes	1.2, Sep 27, 2026
Capability	AI & Data	Industry	Financial Services
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	Approved by Edwin Ozer, Founder, Oct 10, 2026. Issued by Symbeeo, Oct 10, 2026. Design study: every figure is vendor-published, a design target or a marked assumption.		
Sign-off	"An engagement under this design is signed by the practice lead before any build proceeds."		

Contents

1 · Scope 3
1c · Public implementations this design follows 4
2 · Requirements register 4
3 · Architecture overview 5
4 · Design by subsystem 5
5 · Failure behavior 6
6 · Limits 7
7 · Lessons learned 7
8 · The author 7

© 2026 Symbeeo. All rights reserved. · Public

CONVENTIONS

Source tags. [S] figure read from the vendor's own current page on Sep 18, 2026 · [S*] figure from the vendor's page summary where the page itself could not be opened · [A] design assumption, to be confirmed by the signing engineer · [T] design target taken from the study record or derived from it. No figure comes from any prior engagement, file or asset. A row with more than one tag is counted once: as a published figure if any tag is [S] or [S*], otherwise as an assumption if any tag is [A], otherwise as a design target.

Table 1. Revision history

Date	Version	Change
Sep 18, 2026	1.0	First issue.
Sep 20, 2026	1.1	Public implementations added; figures reconciled with published sources.
Sep 27, 2026	1.2	Figures and platform details rechecked against published sources.
Oct 10, 2026	1.3	Approval line now reads: approved by the Founder, issued by Symbeeo. Diksha Shah remains the author.

1 · Scope

Who it is for. The head of infrastructure of a bank that runs document-understanding and fraud models on its own floor, who must show a regulator that inference traffic never leaves the building and show finance that the rack budget holds.

Typical buyer [A]. A bank running document and fraud models in its own hall: one data-center hall with liquid-cooling capacity already installed, 4 production models (2 document extraction, 1 fraud scoring, 1 general LLM for internal assistants), 3,000 internal users, 24 × 7 serving with a 4-hour nightly window for model rollouts.

What is in. Two NVIDIA GB300 NVL72 racks delivered as Dell PowerEdge XE9712 in Dell Integrated Rack Scalable Systems; the in-rack NVLink domain; the north-south Ethernet to the bank's network; the serving layer and its tokens-per-second targets; power and cooling at the rack; the failure behavior. **What is out.** Model training and fine-tuning; the data-preparation pipeline; the bank's identity and network security design (SEC design studies); storage for training corpora.

Assumptions. NVIDIA is the accelerator vendor and Dell the system integrator [A]. The hall has facility water for direct liquid cooling [A]. Models are served as FP4-quantized checkpoints where the vendor runtime supports it [A]. Data residency forbids any inference API outside the hall [T].

Related documents, read Sep 18, 2026. NVIDIA GB300 NVL72 product page (https://www.nvidia.com/en-us/data-center/gb300-nvl72/) · NVIDIA NVL72 AI Factory reference architecture, system hardware and components (https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html) · Dell PowerEdge XE9712 product page (https://www.dell.com/en-us/shop/ipovw/poweredge-xe9712) · Dell blog, first GB300 NVL72 delivery (https://www.dell.com/en-us/blog/dell-delivers-market-s-first-nvidia-gb300-nvl72-to-coreweave/) · Dell Integrated Rack Scalable Systems (https://www.dell.com/en-us/shop/storage-servers-and-networking-for-business/sf/integrated-rack-scalable-systems) · NVIDIA NVL72 AI Factory reference architecture, index, read Sep 20, 2026 (https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/index.html).

© 2026 Symbeeo. All rights reserved. · Public

1c · Public implementations this design follows

Deployment rows are real public deployments, each dated. Where no public deployment matches, the row is Vendor documentation: the vendor's document that this design, Symbeeo's own, is built on.

Table 2. Public implementations this design follows

Implementation	Who published it · date	What it settles for this design	URL
Vendor documentation. NVIDIA NVL72 AI Factory enterprise reference architecture, smallest size: 2 racks, 36 trays, 144 Blackwell Ultra GPUs	NVIDIA · last updated Sep 15, 2026	The two-rack, 144-GPU scale of R1; SN2201 OOB and a separate north-south fabric; the rack power figure in R8 and R9	`https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/index.html`

2 · Requirements register

Table 3. Requirements register

#	Requirement	Value	Source
R1	Racks	2 × GB300 NVL72, on-premises, liquid-cooled	[T]
R2	GPUs per rack	72 Blackwell Ultra GPUs, 36 Grace CPUs	[S] NVIDIA GB300 NVL72
R3	GPU memory per rack	20 TB HBM3e, up to 576 TB/s	[S] NVIDIA GB300 NVL72
R4	Fast memory per rack	37 TB (GPU + CPU)	[S] NVIDIA GB300 NVL72
R5	NVLink domain bandwidth	130 TB/s across 72 GPUs	[S] NVIDIA GB300 NVL72
R6	FP4 inference per rack	1,440 PFLOPS with sparsity, 1,080 dense	[S] NVIDIA GB300 NVL72
R7	Network per GPU	800 Gb/s via ConnectX-8 SuperNIC	[S] NVIDIA GB300 NVL72
R8	Rack power, vendor figure	up to 142 kW per rack; 8 power shelves of 33 kW	[S] NVIDIA NVL72 RA components
R9	Rack power, design budget	up to 142 kW per rack, equal to R8	[S] NVIDIA NVL72 RA components (`https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html`)
R10	Cooling	direct liquid cooling, facility water at the rack	[S] Dell XE9712 DLC; [A] water available
R11	Serving target	tokens-per-second target per model, held under R9	[T]
R12	Data residency	no inference request or response leaves the hall	[T]
R13	Availability of serving	99.9 % monthly, measured at the API gateway	[A]
R14	Rollout window	4 hours nightly for model version changes	[A]

© 2026 Symbeeo. All rights reserved. · Public

Fourteen rows; nine carry a vendor figure, three are design targets, two are design assumptions.

3 · Architecture overview

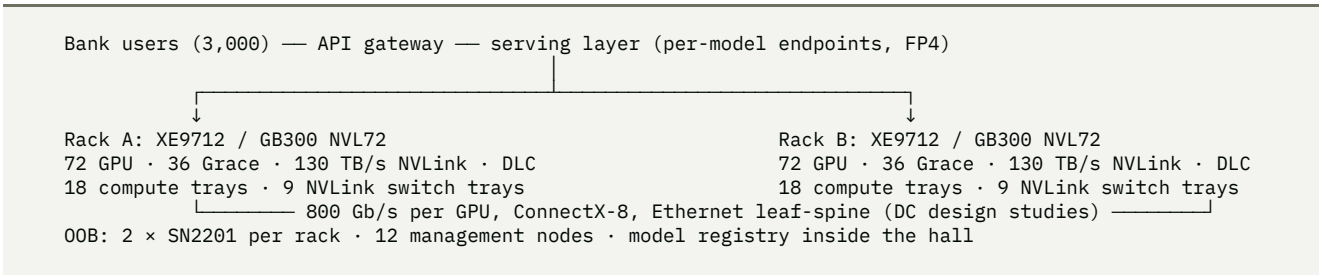


Figure 1. Inference racks and serving layer

Two identical NVL72 racks, each a single 72-GPU NVLink domain; models pinned to a rack so that a rack failure loses capacity, not a model; north-south Ethernet only, no inter-rack model sharding in this design.

Table 4. Architecture overview

Role	Product (exact model)	Qty	Capacity used	Why this unit
GPU rack	Dell PowerEdge XE9712 with NVIDIA GB300 NVL72	2	72 GPUs, 37 TB fast memory per rack [S] NVIDIA GB300 NVL72	Dell's rack-scale Blackwell Ultra system with a 72-GPU NVLink domain [S]; Dell is the design's integrator [A]
Compute tray	XE9712 sled, 4 Blackwell Ultra GPU + 2 Grace	36	18 per rack [S] NVIDIA RA	Defined by the rack; not a choice
NVLink switch tray	fifth-generation NVLink, 2 NVSwitch ASICs per tray	18	9 per rack, 130 TB/s [S]	Defined by the rack
Network adapter	NVIDIA ConnectX-8 SuperNIC, 800 Gb/s	144	4 per compute tray [S]	Defined by the rack
DPU	NVIDIA BlueField-3 B3240, dual-port, 480 Gb/s aggregate	36	1 per compute tray [S]	Defined by the rack; carries storage and management traffic
OOB switch	NVIDIA SN2201	4	2 per rack [S]	Defined by the reference architecture
Rack enclosure	Dell Integrated Rack Scalable Systems, IR7000 (ORv3 21-inch)	2	up to 504 kW per rack [S] Dell IRSS	Dell's delivery format for XE9712 with GB300 NVL72 [S]
Management nodes	x86, per NVIDIA RA	12	control plane [S]	Defined by the reference architecture

4 · Design by subsystem

4.1 GPU racks.

Two Dell PowerEdge XE9712 racks, each NVIDIA GB300 NVL72: 72 Blackwell Ultra GPUs, 36 Grace CPUs, 2,592 Arm Neoverse V2 cores, 20 TB GPU memory, 37 TB fast memory, 130 TB/s NVLink [S]. Each rack is one NVLink domain.

© 2026 Symbeeo. All rights reserved. · Public

Rejected: HGX B300 8-GPU servers — a 72-GPU NVLink domain fits the largest model in one memory space without inter-node traffic [A].

4.2 Model placement.

Four models, each replicated on both racks; the two document models and the fraud model share one rack's GPUs by partitioning without MIG at the serving layer; the general LLM takes the remainder [A]. Tokens-per-second per model is set at first build from a benchmark on the delivered hardware and recorded here [T]; no figure is claimed now.

4.3 Serving layer.

Per-model endpoints behind one API gateway inside the hall; FP4 checkpoints where the runtime supports them, otherwise FP8 (720 PFLOPS FP8 per rack [S]) [A]. The model registry, the gateway and every log stay in the hall (R12) [T].

4.4 Network.

800 Gb/s per GPU through ConnectX-8 [S] into an Ethernet leaf-spine designed under the DC design studies; BlueField-3 B3240 at 480 Gb/s per tray [S] carries storage and management. No inter-rack NVLink; the two racks are independent failure domains [A].

4.5 Power and cooling.

Vendor figure: up to 142 kW per rack, 8 power shelves of 33 kW, six 5.5 kW PSUs per shelf [S]. Direct liquid cooling on the XE9712 sleds [S]. The design budget per rack is the vendor figure, up to 142 kW (R9) [S]; the hall must provision it.

4.6 Management and out-of-band.

2 × SN2201 OOB switches per rack and 12 management nodes, per the NVIDIA reference architecture [S]; reachable before the serving network exists [A].

4.7 Rollout.

New model versions land in the in-hall registry, are promoted to one rack in the 4-hour window, then to the second the following night [A].

5 · Failure behavior

Table 5. Failure behavior

Failure	Behavior	Source
One GPU fails in a rack	serving layer drains that GPU's replicas; rack continues at 71/72	[A]
One compute tray fails	4 GPUs and 2 Grace CPUs lost; rack continues; tray replaced	[S] tray composition; [A] behavior
One NVLink switch tray fails	NVLink domain degraded; rack drained to the other rack until replaced	[A]
One power shelf fails	33 kW of 8 × 33 kW lost; rack throttles or drains by policy	[S] shelf figure; [A] policy
Facility water loss	both racks shut down on thermal limit; serving stops	[A]
One whole rack lost	all four models remain on the other rack at half capacity	[A]

© 2026 Symbeeo. All rights reserved. · Public

6 · Limits

- Tokens-per-second per model is a target set at first benchmark; no figure is stated here.
- No inter-rack model sharding; a model larger than one rack's 37 TB fast memory is out of scope.
- Facility water and the liquid-cooling loop are assumed present; if absent, the design changes platform.
- The NVIDIA product page does not list cooling in its specification table; the liquid-cooling statement is from the page overview and the Dell XE9712 page.

7 · Lessons learned

None yet — this is a design study; populated after the first build against it.

8 · The author

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

Signed off by Edwin Ozer, Founder.

© 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.

AI-002 · Version 1.3 · Oct 10, 2026