SYMBEEO

PUBLIC

DESIGN STUDY

Sovereign AI stack for a public agency, no data leaving the premises

Complete design on current platforms

CAPABILITY	INDUSTRY	SETTING
AI & Data	Public Sector	Government agency
DOCUMENT	**VERSION**	**DATE**
AI-004	1.3	Oct 10, 2026

WRITTEN BY

Diksha Shah, Principal AI Infrastructure Architect

© 2026 Symbeeo. All rights reserved.

Document control

Document	AI-004	Version	1.3
Issued	Oct 10, 2026	Status	Issued by Symbeeo
Classification	Public	Supersedes	1.2, Sep 27, 2026
Capability	AI & Data	Industry	Public Sector
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 6
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 chief information officer of a government agency under a data-residency mandate, who must run language and document models for staff without any request, weight or log crossing the agency boundary, and must prove it to an auditor.

Typical buyer [A]. A government agency under a data-residency mandate: 4,000 staff, one primary data center and one secondary 30 km away, 3 model families (document summarization, search over internal records, an internal assistant), an audit regime that requires every inference to be attributable to a user and retained for 7 years.

What is in. An air-gapped enclave in the primary data center; Red Hat OpenShift with OpenShift AI as the platform; the model registry inside the enclave; GPU nodes on Dell PowerEdge; inference logging; the offline update path. **What is out.** Model training; the agency's end-user identity provider (integrated, not designed); the secondary site (DC design studies); network security design (SEC design studies).

Assumptions. Red Hat is the platform vendor, Dell the server vendor, NVIDIA the accelerator vendor [A]. The enclave has no route to the internet; updates arrive on removable media through a controlled transfer station [A]. Models are open-weight and enter the enclave through the same station [A].

Related documents, read Sep 18, 2026. Red Hat OpenShift AI product page (https://www.redhat.com/en/products/ai/openshift-ai) · Red Hat OpenShift AI Self-Managed 3.5 documentation index, read Sep 27, 2026 (https://docs.redhat.com/en/documentation/red_hat_openshift_ai_self-managed/3.5) · Red Hat OpenShift AI Self-Managed 3.5, deploying models on the single-model serving platform, read Sep 27, 2026 (https://docs.redhat.com/en/documentation/red_hat_openshift_ai_self-managed/3.5/html/deploying_models/deploying_models) · Red Hat OpenShift AI Self-Managed life cycle, read Sep 27, 2026: 3.5 general availability Aug 27, 2026, full support to Mar 29, 2027, Extended Update Support to Feb 28, 2028; 3.4 and 3.3 in full support without EUS; 3.2 and 3.0 end of life (https://access.redhat.com/support/policy/updates/rhoai-sm/lifecycle) · Dell PowerEdge XE7745 product page (https://www.dell.com/en-us/shop/ipovw/poweredge-xe7745) · Dell press release, Nov 17, 2025, enterprise AI with NVIDIA (https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2025~11~dell-technologies-and-nvidia-advance-enterprise-ai-innovation.htm) · NVIDIA RTX PRO 6000 Blackwell Server Edition (https://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/) · Dell PowerEdge R7725 product page (https://www.dell.com/en-us/shop/ipovw/poweredge-r7725) · Dell AI Factory for Government with NVIDIA, RTX PRO 6000 Blackwell Server Edition configuration on Dell

© 2026 Symbeeo. All rights reserved. · Public

PowerEdge servers, datasheet, read Sep 20, 2026 (https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/dell-ai-factory-with-nvidia-rtx-pro-6000-on-dell-poweredge-servers-datasheet.pdf).

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. Dell AI Factory for Government with NVIDIA: endorsed RTX PRO 6000 Blackwell Server Edition configuration on PowerEdge XE7745 and XE7740, 8 GPUs per node, 4-node and 4–32-node configurations, air-gapped deployment named	Dell Technologies · Mar 24, 2026 (the PDF's document date; no date is printed on the page), read Sep 27, 2026	The XE7745 with 8 × RTX PRO 6000 as a government configuration (R5–R7); about 8 kW per node for hall power; air-gapped deployment as a supported case (R1)	https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/dell-ai-factory-with-nvidia-rtx-pro-6000-on-dell-poweredge-servers-datasheet.pdf

2 · Requirements register

Table 3. Requirements register

#	Requirement	Value	Source
R1	Enclave	air-gapped; no external API in the data path	[T]
R2	Platform	Red Hat OpenShift AI Self-Managed 3.5, disconnected mode; full support to Mar 29, 2027, Extended Update Support to Feb 28, 2028	[S] Red Hat product page: "on-premise and disconnected mode support"; [S] Red Hat life cycle, read Sep 27, 2026
R3	Model registry	inside the enclave	[T]; [S] model registry is a platform component
R4	Serving runtime	vLLM NVIDIA GPU ServingRuntime for KServe, RawDeployment	[S] Red Hat 3.5 deploying models, read Sep 27, 2026
R5	GPU node	Dell PowerEdge XE7745, 4U, 8 × PCIe Gen5 double-width up to 600 W	[S] Dell XE7745
R6	GPU	NVIDIA RTX PRO 6000 Blackwell Server Edition, 96 GB GDDR7, up to 600 W	[S] NVIDIA
R7	GPU count	24 GPUs across 3 nodes	[A]
R8	Node CPU and memory	2 × AMD EPYC 9005, up to 192 cores each; up to 3.072 TB DDR5	[S] Dell XE7745
R9	Node local storage	up to 8 × E3.S Gen5 NVMe, 122.88 TB max	[S] Dell XE7745

© 2026 Symbeeo. All rights reserved. · Public

#	Requirement	Value	Source
R10	Inference logging	every request: user, model, version, timestamp, prompt and response hash; retained 7 years	[T] every inference logged; [A] fields and retention
R11	Guardrails	input and output filtering on the assistant model	[S] Red Hat "Guardrails" feature; [A] applied here
R12	Availability	99.5 % monthly for the assistant endpoint	[A]
R13	Update cadence	platform and model updates by media transfer, monthly	[A]

Thirteen rows; eight carry a vendor figure, one is a design target, four are design assumptions.

3 · Architecture overview

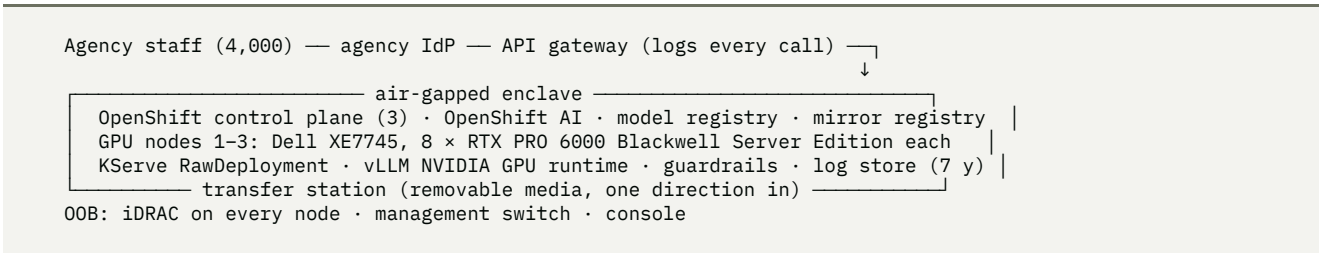


Figure 1. Air-gapped AI enclave

One enclave, one OpenShift cluster, three 8-GPU nodes, a mirror registry for images and a model registry for weights, and a gateway that writes a log line before it forwards a request.

Table 4. Architecture overview

Role	Product (exact model)	Qty	Capacity used	Why this unit
Platform	Red Hat OpenShift AI Self-Managed 3.5 (disconnected)	1 cluster	3 model families	Disconnected mode, model registry, KServe and vLLM are platform features [S]
Control plane	Dell PowerEdge R7725, 2U, 2 × AMD EPYC 9005	3	control plane and registries	Current Dell 2U AMD EPYC 9005 node [S] R7725 page
GPU node	Dell PowerEdge XE7745, 4U	3	8 of 8 double-width slots [S]	Dell names XE7745 with RTX PRO 6000 Blackwell Server Edition [S] press release
GPU	NVIDIA RTX PRO 6000 Blackwell Server Edition, 96 GB	24	24 [A]	4 PFLOPS FP4, MIG up to 4 instances, 600 W fits the XE7745 slot [S]
Serving	KServe RawDeployment with vLLM NVIDIA GPU ServingRuntime	per model	—	Listed runtime [S]
Log store	object store inside the enclave, 7-year retention	1	—	Vendor to be chosen at first build [A]

© 2026 Symbeeo. All rights reserved. · Public

4 · Design by subsystem

4.1 Enclave and platform.

OpenShift AI Self-Managed 3.5 in disconnected mode [S], images mirrored into an in-enclave registry, operators installed from the mirror [A]. No route to any external endpoint exists from any node; egress is denied at the enclave boundary (R1) [T].

4.2 GPU nodes.

Dell PowerEdge XE7745, 4U, two AMD EPYC 9005 processors, up to 3.072 TB DDR5, eight PCIe Gen5 double-width slots at up to 600 W, 3,200 W Titanium PSUs [S]; eight NVIDIA RTX PRO 6000 Blackwell Server Edition per node, 96 GB GDDR7, 1,597 GB/s, 4 PFLOPS FP4 [S]. Rejected: an NVL72 rack — no facility liquid cooling in the agency hall and a 3-model estate does not need a 72-GPU NVLink domain [A].

4.3 Model registry.

The OpenShift AI model registry [S] holds every version that has entered the enclave; a model not in the registry cannot be served [A]. Weights enter through the transfer station with a signed manifest checked at import [A].

4.4 Serving.

KServe RawDeployment with the vLLM NVIDIA GPU ServingRuntime [S]; MIG at up to four instances per GPU [S] for the smaller document model [A]; the assistant model on whole GPUs [A]. Guardrails filter inputs and outputs on the assistant endpoint [S] feature, [A] applied.

4.5 Inference logging.

The Red Hat serving documentation read for this document does not describe request logging; logging is therefore placed at the API gateway, which writes user, model, version, timestamp and prompt and response hashes before forwarding, and retains 7 years [A]. Rejected: logging in the model server — not a documented platform feature [S] absent.

4.6 Updates.

Monthly: mirror the operator catalog and images on a connected staging host, transfer by media, verify signatures, apply [A]. Model updates follow the same path (4.3).

4.7 Management and out-of-band.

iDRAC on every node, a management switch and console on their own VLAN, inside the enclave [A].

5 · Failure behavior

Table 5. Failure behavior

Failure	Behavior	Source
One GPU fails	KServe reschedules the replica on another GPU in the cluster	[A]
One GPU node fails	8 of 24 GPUs lost; endpoints continue on 16; replicas rebalance	[A]
One control-plane node fails	cluster continues on 2 of 3	[A]

© 2026 Symbeeo. All rights reserved. · Public

Failure	Behavior	Source
Model registry unavailable	serving continues; no new deployments until restored	[A]
Gateway log store full	gateway refuses requests rather than serve unlogged (R10)	[T]
Transfer station unavailable	no updates; serving unaffected	[A]

6 · Limits

- Inference logging is at the gateway, not in the model server; a request that bypasses the gateway is not logged. The enclave network must make the gateway the only path.
- The GPU count of 24 is a design assumption; it is sized at first build from the three models' measured throughput.
- Three GPU nodes is below Dell's smallest listed AI Factory block, a 4-node configuration (§1c); the node count is confirmed against that block at first build.
- Disconnected updates are monthly by media; a critical fix waits for the next transfer unless the agency runs an out-of-cycle one.
- The platform is pinned at OpenShift AI Self-Managed 3.5, the current release on Sep 27, 2026 and the only 3.x release with Extended Update Support (to Feb 28, 2028) [S]; a later release enters only through the monthly update path (4.6).
- No secondary-site failover in this design; loss of the primary hall stops serving.

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-004 · Version 1.3 · Oct 10, 2026