UDOC

Full Hardware Specification

Deployment System Architecture and Data Operations Core for the Sovereign AI Oversight System

Prepared for G.O.D.S Holdings (Pty) Ltd
Sashin J. Singh — Founder & Systems Architect
Draft engineering package · Version 1.0 · 24 March 2026

Purpose	Vendor-neutral hardware and facility specification for an AI oversight platform that embeds governance controls into AI infrastructure, applications, and decision pipelines.
Reference intent	National backbone, sovereign cloud, private enterprise, and air-gapped deployments.
Design anchors	Sovereignty first; audit by default; zero trust; quantum-ready; interoperable.
Engineering stance	Detailed enough to procure against, but still adaptable to local supplier availability and client jurisdiction.

Contents

1. Executive design brief

This specification translates the UDOC whitepaper and the G.O.D.S systems package into a concrete deployment architecture for the AI oversight system. The design assumes that every consequential AI workload must be observable, classifiable, explainable, reviewable by humans where required, and cryptographically auditable without relying on foreign cloud operators for sovereign workloads.

The result is not a single application server. It is a layered sovereign platform made of an embedded governance fabric, an ingestion and policy plane, an immutable data operations core, a post-quantum-capable security plane, and an operator environment for compliance, oversight, and incident response.

This hardware specification is written to support four deployment forms: a national backbone, a sovereign in-country cloud deployment, a private enterprise deployment, and a compact edge deployment for ministries, agencies, or regulated organisations that need local enforcement near the AI system itself.

The key objective is to make AI governance impossible to bypass by policy alone. It must be technically embedded into the infrastructure stack and the software delivery path.

2. Non-negotiable technical constraints

- Sovereign-critical data must be processable in a fully air-gapped mode, with no dependency on public cloud infrastructure.
- Every significant AI event must generate an immutable, cryptographically verifiable audit record and a lineage record.
- High-risk and critical AI systems must be capable of mandatory human oversight enforcement, suspension, and override logging.
- The governance path must add minimal latency to production decisioning; synchronous enforcement paths should target a sub-50 ms governance overhead where required.
- Primary key material must remain in client-controlled hardware security modules; support access must be temporary, authorised, and fully logged.
- Architecture must support at least 99.9% uptime for enterprise deployments and 99.95% for government backbone deployments through multi-zone or multi-site redundancy.
- The software estate must be updateable through signed, offline-capable release pipelines for air-gapped environments.

3. Reference deployment models

Deployment models

Model	Primary use	Sovereignty level	Minimum footprint	Notes
National backbone	Government-wide AI oversight, identity, welfare, border, justice, regulated state AI	Maximum	Two sites + secure build/signing enclave	Supports full air-gap, dual HSM, dual-site data replication, central compliance authority.
Sovereign cloud	Regulated industries, parastatals, provincial	High	One primary site + one DR site	Client-approved in-country operator. No foreign hyperscaler

	programmes			dependency for sensitive classes.
Private enterprise	Banks, insurers, hospitals, utilities, large enterprises	High	One primary cluster + backup target	Client-owned or contracted infrastructure. Same control model, reduced site count.
Edge governance node	Agency, ministry, OT/edge, branch site, disconnected field system	Targeted local sovereignty	1–3 appliances per site	Embeds governance capture, local caching, signed export, and store-and-forward to core.

4. Hardware architecture overview

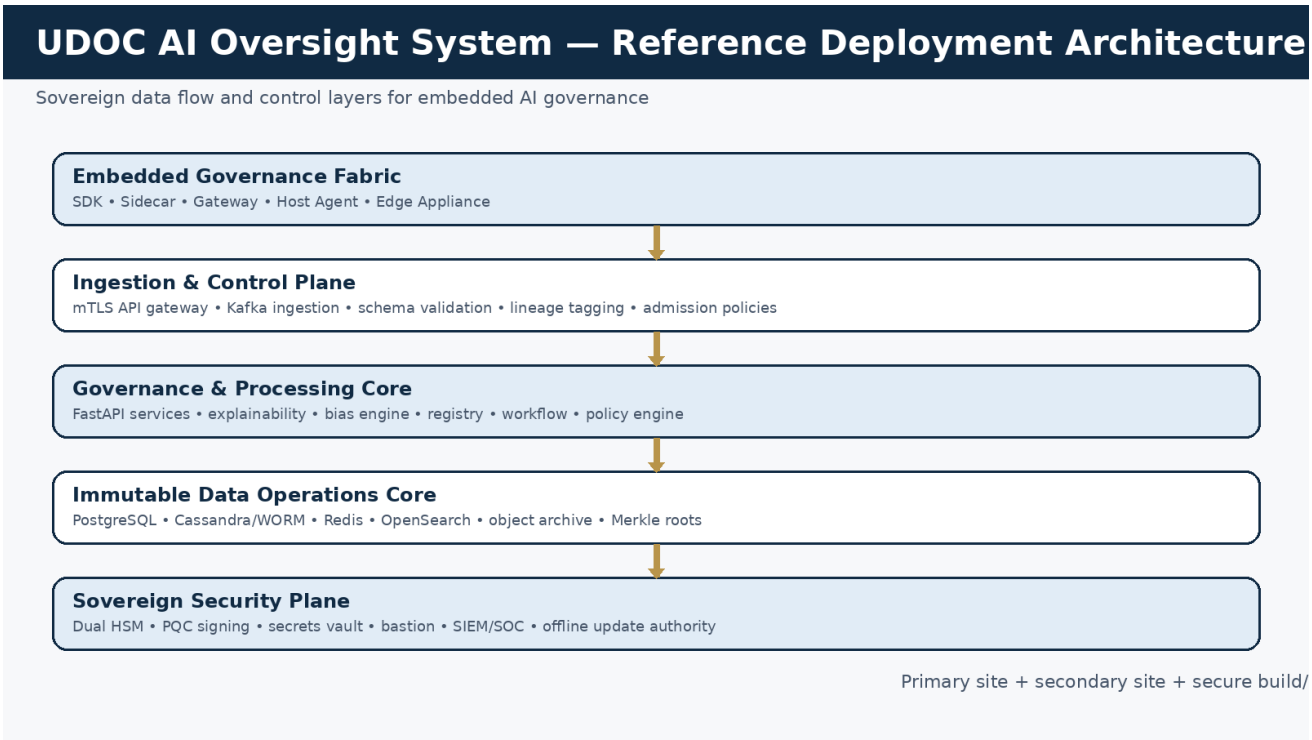

Figure 1. Reference layered architecture for the UDOC AI oversight deployment.

The hardware estate is divided into five planes. Each plane has different performance, storage, and failure-domain requirements.

Hardware planes

Plane	Primary functions	Hardware emphasis	Failure priority
Embedded governance fabric	SDK gateways, sidecars, host agents, branch appliances, model registration hooks	Low-latency compute, strong local buffering, trusted identity	High
Ingestion and control	API ingress, mTLS	Fast NVMe, 25/100GbE,	Critical

plane	termination, Kafka brokers, schema registry, admission control	predictable write latency	
Governance and processing core	FastAPI services, policy engine, explainability, bias engine, workflow, oversight queue	Balanced CPU/RAM, optional GPU for advanced analytics	Critical
Immutable data operations core	PostgreSQL, Cassandra/WORM, Redis, OpenSearch, object archive	Storage endurance, replication, consistent IO, retention management	Critical
Security and operations plane	HSM, vault, bastion, SIEM, Prometheus/Grafana, secure update vault	Trust isolation, management network, tamper-evident ops	Critical

5. Detailed node classes and minimum bill of materials

All server specifications below are vendor-neutral class definitions. Equivalent Dell, HPE, Lenovo, Supermicro, Fujitsu, or local sovereign OEM systems may be used provided the class requirements are met. For government backbone deployments, use enterprise hardware with 5-year next-business-day or 24x7x4 support, redundant power, remote management, TPM 2.0, and signed firmware.

Node classes

Node class	Qty	Chassis	CPU	RAM	Storage	Network	Primary role
Kubernetes control plane	3 per site	1U or 2U	16–24 physical cores	128–192 GB	2 × 960 GB SSD (RAID1)	Dual 10/25GbE + 1GbE OOB	etcd, API server, schedulers, cluster control
Ingress/API nodes	2–4 per site	1U	24–32 cores	128 GB	2 × 1.92 TB NVMe	Dual 25GbE	HAProxy/ Envoy, API gateway, auth edge, admission policies
Governance worker nodes	4–8 per site	2U	32–64 cores	256–512 GB	4 × 3.84 TB NVMe	Dual 25/100GbE	FastAPI services, policy engine, explainability, workflow, scanning

Analytics/ GPU nodes	0–2 per site	2U	32–64 cores	512 GB	2 × 3.84 TB NVMe	Dual 100GbE	1–2 data-centre GPUs with 48–80 GB VRAM each for heavy explainability, embeddings, or model forensics
PostgreSQL HA nodes	3 per site	2U	32–64 cores	512 GB	8 × 3.84 TB NVMe (RAID10) + mirrored boot	Dual 25GbE	Registry, workflow, config, metadata, transactional state
Kafka broker nodes	5 per site	1U/2U	24–32 cores	128–256 GB	6 × 3.84 TB NVMe	Dual 25GbE	Event stream, replay, compliance trail, fan-out
Cassandra audit nodes	6 per site	2U	24–32 cores	256 GB	8 × 7.68 TB NVMe + boot mirror	Dual 25GbE	Append-only audit log, Merkle-linked storage, 10-year retention
Redis nodes	3 per site	1U	16–24 cores	128 GB	2 × 1.92 TB SSD	Dual 10/25GbE	Session, rate limits, queues, short-lived state
OpenSearch data nodes	3–5 per site	2U	24–32 cores	256 GB	8 × 3.84 TB NVMe	Dual 25GbE	Full-text audit search, dashboards, alerting
Object archive nodes	4 per site	4U	16–24 cores	128 GB	12 × 16–20 TB NL-SAS + SSD cache	Dual 25GbE	Immutable object storage, signed report exports, snapshots, long-term archive
Prom/ SIEM/ops nodes	2–3 per site	1U/2U	16–24 cores	128–256 GB	4 × 1.92 TB SSD	Dual 10/25GbE	Observability, alerting, log aggregatio

							n, SOC tooling
Bastion/ admin nodes	2 per site	1U	8–16 cores	64 GB	2 × 960 GB SSD	Dual 10GbE + isolated OOB	Jump host, break-glass access, admin tools, signed session capture

5.1 Kubernetes control and worker classes

Control-plane servers must not host heavy application workloads in production. They should run Kubernetes control services, GitOps agents, and cluster policy enforcement only.

Governance worker nodes host the FastAPI microservices, rule engines, workflow services, scanning services, and registry backends. Reserve 20–25% compute headroom for failover and rolling upgrades.

Analytics/GPU nodes are optional for phase 1. Add them when live counterfactual testing, embedding analysis, model forensics, or large explainability workloads exceed CPU-only capacity.

5.2 Database and immutable audit classes

- PostgreSQL nodes should use enterprise NVMe with power-loss protection, mirrored boot, battery-backed cache where applicable, and synchronous replication between primary and standby within site.
- Cassandra audit nodes should avoid RAID across the data volume; rely on Cassandra replication and node-level redundancy. Use JBOD or RAID0 per device only where supported by the operational model.
- Kafka nodes require low-latency NVMe and enough broker count to tolerate at least two simultaneous failures while maintaining quorum and replication health.
- Object archive nodes should expose S3-compatible immutability and legal hold capability for signed exports, snapshots, and archive bundles.

6. Data operations core

The data operations core is the heart of the platform. It is where the platform earns the claim that audit, lineage, registry state, bias signals, and compliance evidence are technically durable rather than merely application logs.

Data operations core by data class

Subsystem	Data type	Write pattern	Retention target	Hardware rule
PostgreSQL	Registry, workflows, users, config, oversight queue, metadata	Transactional read/write	7 years for governance metadata unless longer required	Use HA trio with NVMe RAID10 and synchronous replication.
Kafka	Decision events, lifecycle events,	Append stream	Short-to-medium	5 brokers minimum,

	integration events		retention + replay	replication factor 3, 25GbE, isolated broker storage.
Cassandra/WORM	Immutable audit events and verification records	Append-only, wide-column	10 years minimum	6 nodes minimum, separate audit keyspace, compaction tuned for write-heavy immutable data.
Redis	Session, rate limit, transient queue, cache	Ephemeral in-memory	Hours to days	3 nodes minimum, persistence for selected queues only.
OpenSearch	Searchable audit and compliance indexes	Indexed ingest	6–24 months hot/warm then archive	Separate ingest and data roles where volume is high.
Object archive	Signed reports, snapshots, evidence packs, exports	Immutable object	7–10+ years	4 nodes minimum, erasure coding or replication, legal hold and versioning enabled.

Recommended logical storage separation:

- Hot zone: NVMe-only transactional and stream storage for PostgreSQL, Kafka, Cassandra commitlog/data, Redis persistence, and hot OpenSearch indexes.
- Warm zone: lower-cost SSD or high-capacity NVMe for OpenSearch warm indexes, replay topics, and nearline evidence packages.
- Archive zone: immutable object storage for signed reports, snapshot chains, monthly Merkle root publications, and exported regulatory packs.
- Offline archive: optional tape or removable-media vault for yearly cold archive sealing in air-gapped state deployments.

6.1 Cassandra immutable audit pattern

Use six audit nodes per site in a 3× replication topology. Partition audit records by jurisdiction, month, and system family to keep partitions bounded and predictable. Each record stores event hash, previous hash, Dilithium signature reference, classification, actor class, and retention metadata. Merkle roots should be computed at fixed intervals and written to PostgreSQL metadata plus object archive for independent verification.

6.2 PostgreSQL governance state pattern

Deploy a three-node PostgreSQL cluster with one primary, one synchronous standby, and one asynchronous or witness/standby node depending on the site topology. Patroni or an equivalent orchestrator should control failover. Keep audit payloads out of PostgreSQL except for hashes and references; the database should hold the searchable state and coordination metadata rather than becoming a log sink.

6.3 Kafka event pattern

Use five brokers per site as the base national-backbone starting point. Separate topics by jurisdiction, client, risk class, and functional stream. Enable idempotent producers, ACLs, mTLS, and immutable topic policy for governance event topics. Mirror selected topics to the secondary site or retain a signed relay export if operating in a hard air-gap.

7. Embedded governance fabric for all AI infrastructure

To embed AI governance into all AI infrastructure and software, the platform must be delivered as a fabric rather than a single dashboard. The fabric attaches to applications at five different points depending on the technical environment.

Governance attachment methods

Attachment	Where used	Minimum local spec	What it enforces	Failure behavior
SDK / in-process library	Custom applications, model-serving code, batch jobs	Negligible beyond host app	Registration hooks, event capture, decision hash, policy checks	Fail-closed for critical use; configurable fail-open for low-risk telemetry.
Sidecar container	Kubernetes workloads, model-serving pods, APIs	2–4 vCPU, 4–8 GB RAM per pod pair	mTLS, local queueing, policy calls, lineage stamping, export to core	Buffers locally and retries; can block high-risk responses.
Host agent	VMs, bare metal model hosts, legacy services	4–8 vCPU, 8–16 GB RAM per host	Process hooks, local logs, drift scan, model registration, route controls	Continues capture during core outage and forwards later.
Gateway appliance	Shared services, SaaS edge, branch connectivity	8–16 vCPU, 32–64 GB RAM, mirrored SSD	Protocol translation, local policy enforcement, topic relay, certificate pinning	Store-and-forward with signed queue; can sever non-compliant connections.
Edge governance node	Disconnected sites, OT, field offices, ministries	16–32 vCPU, 64–128 GB RAM, 2–4 TB NVMe	Local registry mirror, audit cache, signed exports, reviewer queue	Runs autonomously for defined disconnected period; syncs when link returns.

A practical implementation should standardise on these four deployment packages:

- UDOC Agent: a lightweight host process for Windows/Linux servers and VMs.
- UDOC Sidecar: a container image deployed alongside AI inference or business-decision services.
- UDOC Gateway: a hardened appliance or VM used to terminate mTLS, attach lineage tags, and bridge older systems.

- UDOC Edge Node: a compact three-node or single-node appliance cluster used where disconnected or sovereign-local operation is mandatory.

7.1 Edge governance appliance specification

Edge appliance classes

Class	Use case	CPU	RAM	Storage	Networking	Notes
Compact edge	Single agency office or branch	8–16 cores	32–64 GB	2 × 1.92 TB SSD	Dual 10GbE + LTE/5G optional OOB	Best for local relay and signed queueing only.
Standard edge	Provincial department, hospital group, field cluster	16–24 cores	64–128 GB	2 × 3.84 TB NVMe + mirrored boot	Dual 10/25GbE	Supports local registry mirror and short-term audit retention.
Resilient edge trio	Disconnected high-value site	3 nodes of 16–24 cores each	64 GB each	2 × 3.84 TB NVMe each	Dual 10GbE each	Forms local micro-cluster with quorum, local review queue, and sync-on-return.

8. Network, security, and sovereign key architecture

The network is segmented into discrete trust zones. East-west traffic is authenticated and policy controlled. North-south traffic is terminated at ingress gateways or air-gap transfer stations. The out-of-band management network is physically and logically separate from the production network.

Reference network zones

Zone	Systems	Connectivity rule	Minimum hardware
Zone 0 — OOB management	iDRAC/iLO/BMC, switch management, PDU, console	Accessible only from bastion over dedicated management network	1GbE management switching, isolated jump hosts, session recording
Zone 1 — Edge / ingress	Firewalls, load balancers, ingress/API nodes, mTLS termination	Only approved inbound channels; no direct DB access	HA firewall pair, HA load balancer pair, DDoS-aware perimeter
Zone 2 — App / governance	FastAPI services, policy engine, review services, registry services	Can talk to Kafka/Redis/PostgreSQL via allow-listed ports only	25/100GbE leaf switching, micro-segmentation
Zone 3 — Data core	PostgreSQL, Cassandra, Kafka, OpenSearch, object store	No direct user access; service-to-service only	Dedicated low-latency switching, encryption in transit
Zone 4 — Security plane	HSM, vault, PKI, signing	Restricted to signing and	Dedicated secure

	service	secret workflows	enclave or appliance segment
Zone 5 — Secure transfer	Offline package import/export, USB scanning, release vault	Manual and dual-control only	Kiosk, malware scanning, signed media process, optional data diode

8.1 HSM and PQC key architecture

- Deploy two network-attached HSMs per primary site and two per secondary site for government backbone deployments.
- Government backbone deployments should target FIPS 140-3 Level 3 HSM. Enterprise minimum can be FIPS 140-2 Level 2 where legally acceptable.
- Use a strict hierarchy: master key in HSM, KEK tier, DEK tier, per-session keys. Split custody of master-key activation across at least 2 of 3 officers.
- Separate signing service from application services. Applications submit hashes or signing requests; private key material never leaves HSM boundaries.
- Use hybrid TLS 1.3 for transition years where classical and PQC key exchange operate together. Make mTLS mandatory for inter-service traffic in sovereign deployments.
- Publish Merkle roots and signing metadata into archive storage and a public or regulator-visible key registry where policy permits.

8.2 Secure build and offline update authority

Air-gapped and sovereign deployments need a separate secure build/signing enclave. This enclave creates release bundles, signs them, scans transfer media, records hashes, and produces an import manifest for the production site.

Secure build enclave minimum hardware

Component	Qty	Minimum spec	Purpose
Build runners	2	24–32 cores, 128 GB RAM, mirrored SSD, TPM 2.0	Reproducible builds, SBOM generation, image/package compilation
Artifact registry	2	16–24 cores, 128 GB RAM, 8 TB SSD usable	Signed package and container storage
Signing station	2	16 cores, 64 GB RAM, attached HSM or smartcard	Release signature creation and approval workflow
Malware scan kiosk	1	16 cores, 64 GB RAM, multiple removable-media interfaces	Offline media validation before import
Release vault	1	Immutable storage appliance or object store	Retention of signed release packages, manifests, and import logs

9. Storage, backup, archive, and retention architecture

Backups are not enough. The platform needs distinct mechanisms for snapshots, point-in-time recovery, immutable evidence retention, and independent verification.

Protection layers

Layer	Technology pattern	Frequency	Retention	Recovery use
PostgreSQL WAL + snapshots	Continuous WAL archiving + storage snapshots	Continuous + every 4 hours	35–90 days hot; monthly sealed copy	Operational restore, point-in-time recovery
Kafka topic durability	Broker replication + mirrored export	Continuous	Topic-specific	Broker recovery, replay, downstream reprocessing
Cassandra repair + snapshots	Scheduled incremental repair + snapshots	Daily to weekly	30–90 days local + monthly archive	Node rebuild, forensic restore
Object archive immutability	Versioned bucket / WORM object store	Continuous	7–10+ years	Regulatory pack retention, signed proof preservation
Offline seal set	Quarterly archive to tape or secure removable media	Quarterly	Long-term vault	Sovereign legal hold, catastrophic recovery

- For a government backbone, target 1-hour RPO and 4-hour platform RTO at system level, with database failover significantly faster inside a site.
- Run monthly restore tests across PostgreSQL, Cassandra evidence bundles, Kafka replay, and object-archive export validation.
- Never store the only copy of Merkle roots, signing manifests, or regulator exports inside the same fault domain as the originating service.

10. Facility, power, cooling, and physical security

Facility baseline requirements

Area	Requirement	Reason
Power	Dual utility feed where available, dual PDUs per rack, online UPS for 4 hours minimum, N+1 generators	Matches sovereign-backbone continuity target and whitepaper power baseline.
Cooling	N+1 cooling, hot/cold aisle containment, continuous temp/humidity monitoring	Protects dense NVMe/database racks and optional GPU nodes.
Racks	42U lockable racks, front/rear access control, cable management, labelled power/network paths	Supports auditability and maintainability.
Physical access	CCTV, access logs, two-person integrity for HSM and transfer	Required for critical operations

	room operations	and sovereign evidence chain.
Fire protection	Clean-agent suppression, zoned detection, emergency shutdown procedures	Protects electronics without water damage.
Room separation	Separate rooms or cages for production, OOB, security enclave, and transfer room	Reduces insider risk and accidental cross-connect.

Indicative rack power planning:

- Control/network rack: 4–6 kW.
- Governance compute rack: 6–12 kW depending on GPU density.
- Data core rack: 8–14 kW due to NVMe-rich databases and brokers.
- Archive/ops rack: 5–9 kW.
- Design room and cooling for peak load plus 20% headroom.

11. Availability, disaster recovery, and operational targets

Operational targets

Metric	Target	Hardware / architecture implication
Platform uptime	99.9% enterprise / 99.95% backbone	At least two fault domains, rolling upgrades, HA network edge, redundant HSM and storage paths
Governance latency overhead	<50 ms for synchronous path	Keep policy engines on fast worker nodes close to ingress; reserve CPU headroom
Audit verification	<500 ms typical	Hot Cassandra partitions, Merkle index in fast store, low-latency search metadata
Standard compliance report	<60 s	Parallel report workers, indexed data, hot OpenSearch and PostgreSQL metadata
Full audit extract (12 months)	<10 min	Archive indexing, batch export workers, object-store staging area
Incident report	<5 min	Predefined templates, hot metadata, queued signing service

- Primary and secondary sites should each hold enough independent capacity to run the essential governance functions during the loss of the other site.
- Use reserved capacity rather than running every cluster at 80–90% utilisation. For failover-sensitive clusters, stay below 60% steady-state utilisation.
- Run quarterly failover exercises for PostgreSQL, Kafka, control-plane recovery, HSM switchover, and offline package import.

12. Reference site build: primary, secondary, and secure build enclave

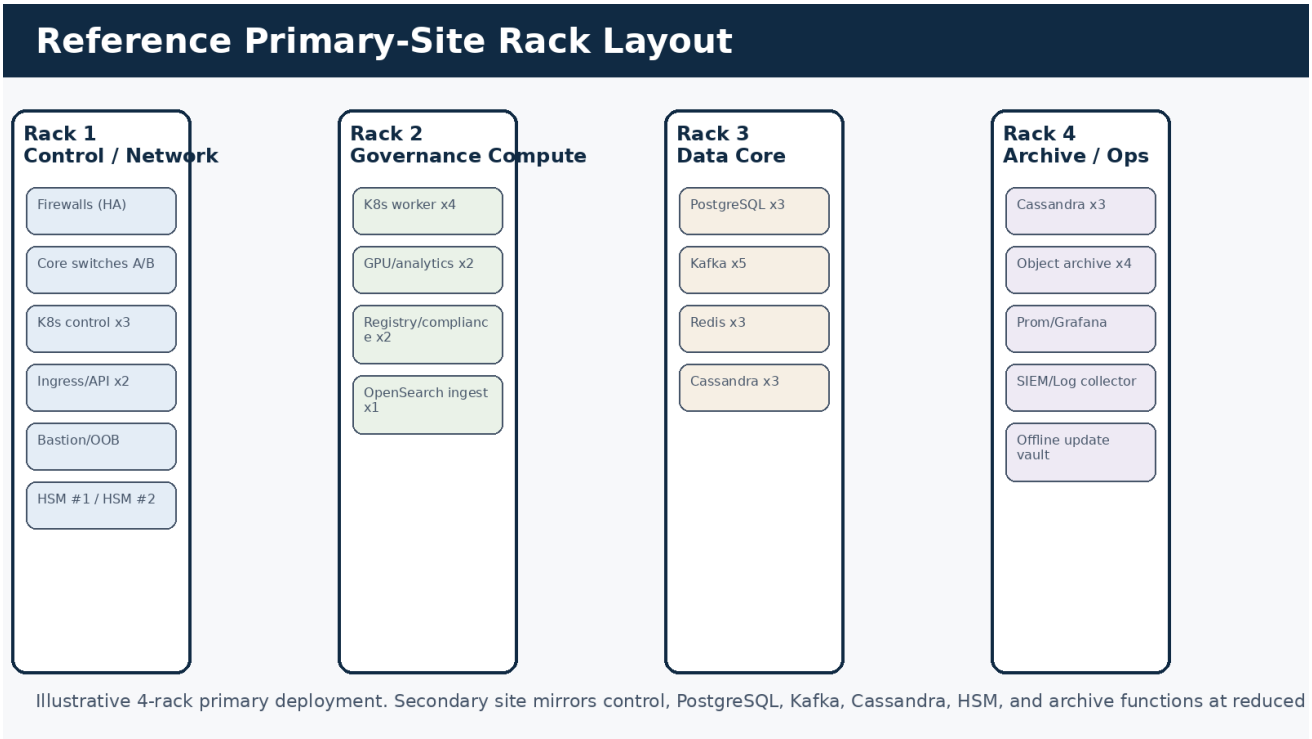

Figure 2. Illustrative four-rack primary-site layout.

A serious national deployment should be built from three coordinated environments:

18. Primary production site. Runs the live control plane, governance plane, data operations core, and security plane.
19. Secondary production or DR site. Maintains replicated or mirrored capacity for failover, restore, and continuity. It should not be a passive afterthought; it must be testable and promotable.
20. Secure build/signing enclave. Stays separated from production, especially for air-gapped deployments, and produces all signed release material.

Reference site summary

Environment	Indicative racks	Key systems	Notes
Primary site	4 racks	Full stack including HSM, data core, archive, ops	Scale to 5–6 racks if multiple ministries or heavy model forensics are expected.
Secondary site	2–3 racks	Reduced mirror of control, PG, Kafka, Cassandra, HSM, archive	May omit GPU and some OpenSearch tiers initially.
Secure build enclave	1 rack	Build, registry, signing, malware scan, release vault	May live in separate secure room or separate facility.

13. Capacity tiers and phased rollout

Capacity tiers

Tier	Target scope	Primary site minimum	Secondary site minimum	Recommended start
Tier 0	Proof-of-capability / pilot	3 control+worker nodes, 3 PG/Kafka combined, no Cassandra cluster yet if pilot only	1 backup target	Use only for internal validation, not full sovereign claim.
Tier 1	Enterprise regulated deployment	3 control, 4 workers, 3 PG, 3 Kafka, 3 Cassandra, 3 Redis/OpenSearch shared	1–2 racks DR	Suitable for bank, insurer, hospital network, provincial programme.
Tier 2	National backbone starter	3 control, 6 workers, 3 PG, 5 Kafka, 6 Cassandra, 3 Redis, 3–5 OpenSearch, 4 object archive, dual HSM	2–3 racks mirror	Recommended baseline for serious government rollout.
Tier 3	National multi-domain backbone	Tier 2 plus extra worker, data, archive, and edge relay capacity	Full mirror with independent quorum	Use when multiple ministries or many AI systems are onboarded concurrently.

Phased rollout:

21. Phase 1: Build the secure build/signing enclave, control plane, PostgreSQL cluster, Kafka starter ring, and governance worker nodes.
22. Phase 2: Add Cassandra immutable audit cluster, object archive, OpenSearch, and HSM-backed signing service.
23. Phase 3: Add secondary site, edge nodes, and formal DR testing.
24. Phase 4: Add GPU analytics nodes, larger archive tiers, and broader ministry or enterprise integrations.
25. Phase 5: Expand edge governance appliances and disconnected-site support.

14. Acceptance criteria and operational readiness checklist

- All critical services survive the loss of any single server without loss of platform availability.
- PostgreSQL failover completes inside the agreed database RTO and application reconnection is automatic.
- Kafka and Cassandra continue operating after the loss of at least one broker/node pair per fault domain.
- HSM switchover and signer failover are demonstrably testable and logged.
- Offline release import process produces signed import manifests and retains a full chain of custody.

- Governance sidecar or agent deployment templates exist for Kubernetes, VM, and legacy gateway scenarios.
- Immutable archive export can be produced, signed, and validated independently from the live UI.
- Monthly restore tests, quarterly DR drills, and annual physical penetration tests are scheduled in the operating model.
- All physical racks, power feeds, PDUs, patch panels, and switch ports are labelled and mapped to an asset register.
- Capacity dashboards report CPU, RAM, disk latency, queue lag, broker health, repair status, and signing-service health for every environment.

Appendix A. Sample rack layout

See Figure 2. The primary-site layout should keep network edge, compute, data core, and archive/operations physically grouped but logically segmented. Where the facility allows, place the security plane and transfer room in separate locked cages or rooms.

Appendix B. Data classification to hardware treatment map

Classification treatment

Classification	Deployment rule	Crypto rule	Storage rule	Connectivity rule
SOVEREIGN-CRITICAL	Air-gapped only or fully disconnected sovereign enclave	AES-256 + PQC key exchange/signing, client-held HSM keys	No public cloud, no shared object archive with lower classes	No public internet exposure; controlled transfer only
SOVEREIGN-HIGH	Sovereign in-country infrastructure only	AES-256, mTLS, HSM-backed keys	Dedicated logical separation, immutable audit path	No foreign provider, firewall allow-list only
GOVERNANCE	Sovereign or private cloud	AES-256 at rest, signed records	Immutable store + searchable metadata + archive	Private network or mTLS internet if approved
OPERATIONAL	Approved cloud or private infra	AES-128 minimum, stronger preferred	Shorter retention, anonymised where possible	Standard secure enterprise connectivity
PUBLIC	Any approved infra	TLS in transit	No special storage restrictions	May be internet accessible if intended

Appendix C. Procurement notes and local adaptation guidance

- Do not lock the specification to a single OEM unless the client procurement environment demands it. Define equivalence by CPU class, RAM, storage endurance, NIC speed, PSU redundancy, support SLA, TPM, and firmware controls.
- For South Africa or other in-country sovereign deployments, prefer data-centre operators and support contracts that keep hardware replacement, logistics, and spare holding within jurisdiction.

- Buy spare disks, NICs, PSUs, and at least one spare server per critical class for sovereign or air-gapped environments where vendor lead time is a strategic risk.
- Where budget is constrained, reduce OpenSearch and GPU capacity first. Do not underbuild HSM redundancy, PostgreSQL HA, Kafka quorum, or immutable audit storage.
- In a pilot or enterprise deployment, the architecture may begin with fewer nodes, but the topology should still reflect the final trust zones and upgrade path.