Sovereign AI: Production AI Inside Your Perimeter

Open-weight models served from GPU hardware you operate, on a Kubernetes platform your team already knows, behind a governance boundary that makes external AI usage a written policy. The architecture below is the stack THNKBIG runs in production — metal to model. Page 2: the full platform diagram.

THE FOUR LAYERS

GPU compute + cooling — METAL

`Dell PowerEdge XE · Supermicro · Cisco AI PODs`

Validated GPU nodes sized to a measured cooling envelope. Facility work is the long-lead item — engineer power and cooling before the first rack powers on.

Kubernetes — PLATFORM

`RKE2 · Harvester · ArgoCD GitOps`

Declarative, GitOps-operated orchestration. The same pipeline that ships your applications ships your models — versioned, reviewed, rolled back the same way.

GPU scheduling + inference — SERVING

`NVIDIA GPU Operator · vLLM · Ollama · Open WebUI`

Automated driver lifecycle; open-weight models served with continuous batching and autoscaling; a self-hosted interface the whole organization can use.

AI Ontology boundary — GOVERNANCE

`Policy egress · audit trail · typed schemas`

The layer most builds skip: which data classes may reach which models is written policy. Frontier APIs become controlled, audited egress — not a default data path.

CROSS-CUTTING

Security & compliance
`NeuVector · STIG · air-gap ready`
Inference inherits the controls your platform already passes audit with.

Observability & cost
`Prometheus · Grafana · cost per inference`
Sovereign economics only work if utilization is a number you manage.

COOLING ENVELOPE — DECIDE THIS FIRST

APPROACH	RANGE	REALITY
Air / hot-aisle	≤ ~30kW/rack	Works in existing facilities; caps out as density grows
Direct-to-chip liquid	30–100kW+	Required for dense racks; plumbing is the long-lead item
Validated designs	Engineered	Cisco AI PODs / Dell / Supermicro — cooling math pre-done

WORKLOAD SPLIT (HYBRID IS THE END STATE)

DIMENSION	SOVEREIGN	HOSTED API
Data	Never leaves your boundary	Leaves under provider terms
Cost	Fixed; ~$0 marginal/token	Per-token, forever
Best fit	Steady, sensitive volume	Spiky, frontier-hard tasks

READINESS CHECKLIST

- Inventory the data classes your teams currently send to hosted AI tools
- Map compliance obligations: HIPAA, ITAR, residency, client confidentiality
- Profile workloads: tokens/day, latency targets, steady vs. spiky demand
- Measure your cooling envelope before ordering GPUs
- Confirm your platform team's Kubernetes + GitOps readiness
- Draft the governance boundary: what crosses, to whom, logged how

ABOUT THNKBIG

THNKBIG is a US-based Kubernetes and AI infrastructure consultancy headquartered in Austin, TX. Senior engineers — no offshore handoffs — designing, building, and operating open-source platforms for regulated and IP-sensitive enterprises nationwide.

Kubernetes Consulting & Platform Engineering · AI Infrastructure & Self-Hosted LLMs · Cloud Migration & Modernization · Security · Zero Trust · STIG/FedRAMP · DevOps · GitOps · Observability · Team Training & Enablement

THE THNKBIG TECHNOLOGIES GROUP, LLC
14205 N Mopac Expy, 5th Floor
Austin, TX 78728

Book a discovery call
www.thnkbig.com/contact
(512) 706-9553

Full guide
www.thnkbig.com/resources/sovereign-ai
US-based · Serving enterprises nationwide

DUNS
#108901579