QORINAI

SOFTWARE BRIEFING · NOVEMBER 2026

PRIVATE AI ON QORINAI GPUS

The software that turns our GPUs into private AI you control.

Our own enterprise-AI platform, IntraLLM, and the partners who run and orchestrate the clusters underneath it — one stack, from the GPU to the user, inside your own network.

QORINAI × Amaze Intelligence
Operating team · private GPU clusters · deploy, fine-tune, serve

amaze

QORINAI × GreenThread AI
AI factory OS · multi-model inference engine

GreenThread



QORINAI

AI democratises knowledge. Private AI keeps it yours.

Enterprises are moving from public model APIs to open-weight models they run themselves — for control, for data protection and for cost. Asia-Pacific is where that compute is being built fastest.

THE SHIFT

57 %

of large enterprises now use open-weight models

THE REGION

+78 %

APAC data-centre capacity, 32 → 57 GW by 2030

THE ASSET AT RISK

~40 %

of employee prompts to public AI tools contain sensitive data

THE DECISION

80 %

of APAC's top 2,000 companies will prioritise AI sovereignty by 2028

Your organisation's IP and know-how is capital. The question is whether it stays inside the company — or leaks out through someone else's API.

Sources: Mozilla State of Open Source AI 2026 · JLL 2026 · Cyberhaven 2026 · IDC APAC (as cited in the Amaze × GreenThread roundtable pack)

QORINAI

From the GPU to the user — one accountable chain.

QORINAI builds and finances the GPU capacity. Our partners operate and orchestrate it. IntraLLM puts it in front of your people. You deal with one counterparty and one service level.

Your teams	Assistants, agents, SQL Agent, chat and APIs — behind your identity provider
QORINAI IntraLLM	Business Assistant · Process Automation · Business Assurance · BYOM · Model Repository · Cluster Manager
GreenThread AI	AI factory OS + multi-model serving: up to 5× models per GPU, sub-500 ms switching
Amaze Intelligence	Designs, builds and runs the private clusters — 24×7 operations, break-fix, ISO 27001
QORINAI GPUs & halls	NVIDIA HGX B300-class nodes from Hyperscalers, Tier III facilities; QINARI owns the fleet, procurement and capital

WHAT THIS MEANS FOR YOU

1. **Your data never leaves**

   Inference, fine-tuning and retrieval run on GPUs inside your network or a private QORINAI tenancy.

2. **Flat, forecastable cost**

   Dedicated GPUs instead of per-token bills that grow with every user and agent.

3. **Any open-weight model**

   Bring your own or use the curated repository; swap models without re-platforming.

Illustrative stack; partner roles are governed by the respective written agreements

PART 1

QORINAI IntraLLM

Your own AI, inside your own network.

IntraLLM is QORINAI's turn-key enterprise-AI platform: run and train your own large language and multimodal models on-premises or in a private cloud — hardware, GPU orchestration, model repository, assistants and automation in one console.

QORINAI-owned

Developed with our strategic shareholder Hyperscalers; licensed per deployment

One console

Web console for IT, assistants and agents for users, APIs for developers

Runs anywhere

Your data centre, a QORINAI private tenancy, or the sovereign cloud

QORINAI

Every enterprise will train its own models. The knowledge should stay inside.

01 · IP RETENTION

Knowledge that stays when people leave

Models trained inside the network turn staff experience into a permanent, queryable company asset.

02 · VALUE CREATION

Every automated workflow adds value

Owning the models lets IT streamline, automate and quality-assure process after process — and the capability compounds.

03 · SOVEREIGNTY

Compliance by architecture, not by contract

No enterprise data or prompts go to external proprietary LLMs — privacy and sovereignty obligations are met by design.

04 · SECURITY

An AI that defends the network it lives in

Cyber-assurance workloads inspect logs, configurations and tickets without any of it going offsite.

05 · ECONOMICS

Predictable cost at enterprise scale

Per-token API bills grow with every user and agent; dedicated GPUs give a flat, forecastable monthly cost.

06 · LATENCY & BANDWIDTH

Keep the data next to the GPUs

Inference and training beside the data avoid slow, costly transfers and keep GPU utilisation high.

QORINAI

Eight enterprise-AI functions in one console.

Organised the way an IT department buys and operates: business-facing modules on top, model management in the middle, cluster operations underneath.

Business Assistant

Internal assistant grounded in company documents, systems and policies — plus a customer-facing assistant on the same governed stack.

INTERNAL · CUSTOMER

Process Automation

Connects the model to ERP, CRM, ticketing and any system; agents read screens, fill forms and complete workflows end-to-end.

ERP → ANY SYSTEM

Business Assurance

Quality-assures construction documents, cyber posture and marketing output against your own standards.

CONSTRUCTION · CYBER · MARKETING

Bring Your Own Model

Load, fine-tune and serve proprietary or fine-tuned models alongside the curated repository.

BYOM · FINE-TUNE

Model Repository

Pre-trained LLM and multimodal models ready to serve — GLM-4.6V, Qwen 2.5 and DeepSeek-class reasoning models.

LLM · LMM · OPEN-WEIGHT

Cluster Manager

Clustering, GPU job orchestration and live monitoring — utilisation, memory, power and temperature per GPU.

ORCHESTRATION · MONITORING

Support

Australian engineering support for the platform, the models and the hardware — one number for the whole stack.

AU ENGINEERS

SQL Agent

Ask the database in plain English; the agent writes, runs and explains the SQL across five dialects.

NATURAL LANGUAGE → SQL

QORINAI

Bottom-up, from the GPU to the user.

Most "enterprise AI" products are a chat window that calls someone else's API. IntraLLM is the whole stack — hardware, scheduler, models and interface engineered and supported together.



INTRALLM CONSOLE · CLUSTER MANAGER

L5 — Interfaces for admins and user groups

Web console for IT; assistants, agents and SQL Agent for business users; APIs for developers — behind your identity provider.

L4 — Model repository

Curated pre-trained LLM and LMM models plus your own — versioned, access-controlled, served inside the network.

L3 — Transformers & custom extensions

Serving and fine-tuning runtime, INT8 / FP8 quantisation, retrieval pipelines, enterprise connectors.

L2 — Clustering & GPU job orchestration

Multi-node scheduling of inference and training, per-GPU telemetry, alerting and capacity planning.

L1 — Hardware

NVIDIA HGX B300 / H200 / H100 NVL, AMD Instinct, RTX PRO — built and burned in by Hyperscalers, deployed by QORINAI.

Proprietary platform · QORINAI Pty Ltd · developed with Hyperscalers

QORINAI

What the platform does — and a local model that beat the cloud.

GLM-4.6V

Multimodal on-premises

Image, document and video understanding with a 128K context — about 150 pages or an hour of video in one pass. Reads SAP, CRM and ticketing screens, interprets drawings and network diagrams.

SQL AGENT

Natural language to SQL

"Which products have never been purchased?" becomes an optimised query, run locally and explained in plain language — for finance, healthcare and government.

BROWSER & DESKTOP AUTOMATION

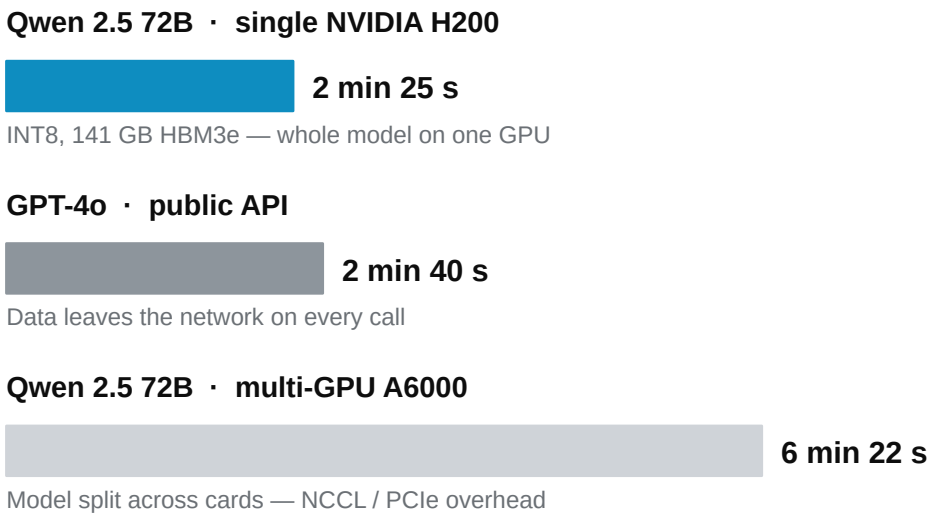
Agents that drive real systems

Natural-language instructions configure products on vendor websites, complete forms and extract data from legacy applications — on local models.

BENCHMARK · BROWSER-AUTOMATION TASK

One H200 running a local model finished first.

Configure a vehicle on a manufacturer's website and complete the enquiry form from natural-language instructions.



Hyperscalers internal benchmark, single run, indicative only; results vary by task, quantisation and model version

QORINAI

Own the model, the GPUs and the data path.

PUBLIC AI API · RENT SOMEONE ELSE'S MODEL

- Prompts and documents leave the country
- Per-token cost grows with every user and agent
- Provider can change, throttle or retire the model
- Your interactions may train their next model
- No control over where inference physically runs

INTRALLM ON QORINAI INFRASTRUCTURE

- Runs in your hall, a QORINAI private tenancy or the sovereign cloud
- Flat monthly cost on dedicated GPUs, billed in AUD
- Open-weight and proprietary models under your version control
- No training on your data — ever
- Australian hardware, warranty, engineers and SLA

Appliance in your data centre

- Pre-configured GPU nodes built by Hyperscalers, IntraLLM installed and burned in
- Single node to multi-rack; air-gapped if required
- Full Australian warranty

Private tenancy in a QORINAI hall

- Dedicated GPUs in a private cage or suite, operated 24×7
- 30–250 kW liquid-cooled racks
- 99.9 % node SLA, hot spares, your identity provider

Sovereign cloud

- Managed IntraLLM on QORINAI's Australian GPU cloud
- Reserved capacity, AUD billing, no training on your data
- Start with one GPU; move on-premises later without re-platforming

Pricing: platform licence per deployment plus the hardware or hosting underneath it

PART 2

QORINAI × amaze

Fast, secure and private AI infrastructure.

Amaze Intelligence is QORINAI's operating team and a strategic shareholder in QINARI Pty Ltd, the vehicle that owns the GPU fleet. It designs, builds and runs the private clusters — and provides the software layer that lets a customer put a cluster to work rather than merely rent one.

20 years

Australian cloud infrastructure, sovereign from the outset

ISO 27001

Certified; Sydney and Melbourne regions

99.9 %

Financially backed uptime SLA, measured monthly

Deploy your models, fine-tune, inference and productise with ease.

Bring any open-weight or self-hosted model onto your own GPU cluster — Llama, Qwen, DeepSeek, Mistral, Gemma and more.

01
Deploy
Bring any open-weight or self-hosted model.

02
Fine-tune
Adapt it to your data on dedicated GPUs.

03
Inference
Serve it privately, at production scale.

04
Productise
Ship it into your products and teams.

WHO IT IS BUILT FOR · TEAMS PUTTING AI INTO PRODUCTION

Model trainers
Adapt open-weight models to your domain, language or regulated data.

- Fine-tuning and adaptation
- Distillation
- Continual training

High-volume inference
Ship AI features where latency and data residency matter.

- Production inference
- Dedicated endpoints
- Burst capacity

Enterprise agentic workloads
Run agentic teams on predictable, secure GPUs.

- Agentic workforces
- Inference platforms
- Business harnesses

From proof of concept to a 1,024-GPU supercluster.

Turnkey: Amaze deploys into AI data centres in partnership with Equinix, offers GPU clusters on a flexible OpEx model, and runs the DC, hardware, break-fix and cluster operations 24×7.

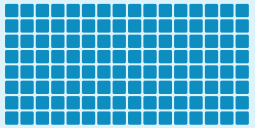
Single node



8

GPUs · Single-node proof of concept

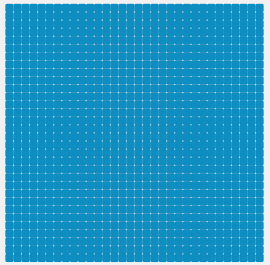
Small cluster



128

GPUs · Production capacity

Large cluster



1,024

GPUs · Large-model training and inference

NEXT STEP · DISCOVERY WORKSHOP

1. Define your problem statement.
2. Map workloads and data.
3. Leave with a cluster plan.

ISO 27001 certified

Ecosystem: NVIDIA, Dell, Gigabyte compute · DDN, VAST AI data · Equinix facilities

PART 3

QORINAI × GreenThread

The AI factory OS and multi-model inference engine.

GreenThread runs AI on your own GPUs: it provisions and operates every node, packs many models onto each GPU, and serves them to your teams as private chat and Anthropic- / OpenAI-compatible endpoints.

< 500 ms

Model switching time

up to 5×

More models per GPU

up to 70 %

Lower inference infrastructure cost

Everything you need to run AI on your own GPUs.



OPERATE

Factory OS

Provision, monitor and operate every GPU node.

SERVE

Serving

Up to 5× more models per GPU, switched in under 500 ms.

CHAT

Assist & Chat

A private chat workspace for your teams, on your own models.

CONSUME

Endpoints

Anthropic- and OpenAI-compatible. Multi-tenant. White-label ready.

Runs on any site — on-premises, in colocation, or with a provider such as Amaze Intelligence on QORINAI capacity.

Dashboard shown is illustrative (GreenThread)

From metal to models — plug-and-play, on hardware you control.

Like a managed model service such as Bedrock, but on your own infrastructure: models and tokens as a service, with no idle GPUs.

WHAT YOU RUN TODAY

- GPU servers
- Networking and fabric
- Drivers and firmware
- Kubernetes and schedulers
- Model runtimes

GreenThread

Factory OS

+ Serving

Takes the complexity out.

WHAT YOUR TEAMS GET

Chat and reasoning	READY
Code assistant	READY
Document search	READY
Vision	READY
Agents and tools	READY
Guardrails	READY

Tokens, by team · no idle GPUs

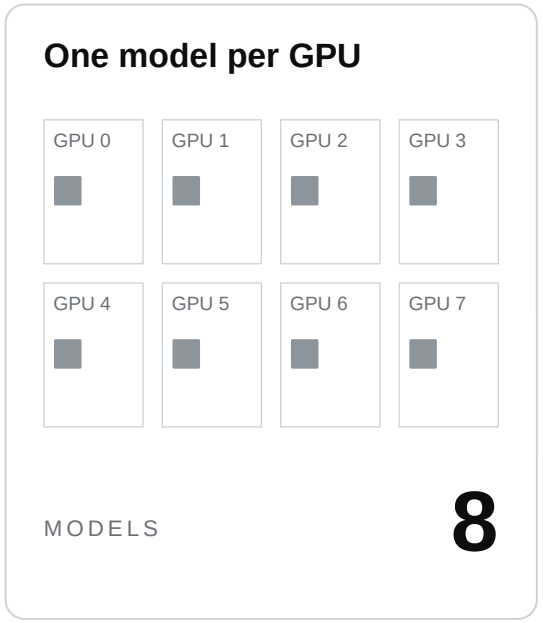
GreenThread × QORINAI

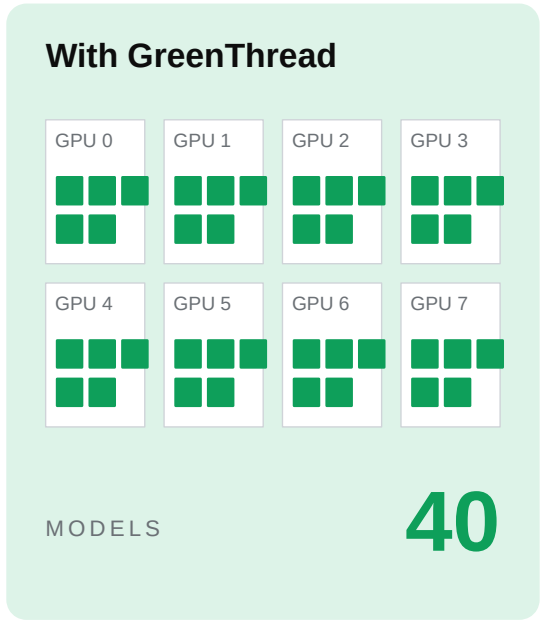
Same hardware. Up to five times the models.

Most GPUs run one model, then sit idle. GreenThread's multiplexing holds many models warm on every GPU and routes the right model to every request — more speed, better unit economics on every GPU.

ONE SERVING LAYER, EVERY WORKLOAD

- chat
- code
- search
- vision
- agents
- policy





ONE GPU NODE · ILLUSTRATIVE

< 500 ms
Model switching time

up to 5×
More models per GPU

up to 70 %
Lower inference infrastructure cost

Any size
Scales to fit your infrastructure

GreenThread figures; savings are "up to" and depend on your model mix

One factory. Your choice of compute.

GreenThread Factory

CPU 01 | CPU 02 | CPU 03

NICo | GitOps | FluxCD | Renovate

FLEET HEALTH · LIVE

Monitor. Detect. Resolve.

Prometheus | Grafana | AutoRMA

PROVISION + OPERATE →

Managed GPU Metal

ssh

GPU server 01 — GPU

GPU server 02 — GPU

Ready to use. No Kubernetes required.

Managed GPU Kubernetes

kubectl

workload-1: CPU · CPU · CPU; GPU · GPU

workload-2: CPU · CPU · CPU; GPU · GPU

Slurm | GreenThread Serving | Your apps

Metal or managed Kubernetes.

Fleet health, built in.

Every GPU stays with customers.

Built for how you deploy AI — scales to fit your infrastructure.

NEOCLOUDS AND MSPS

Offer AI inference to your customers.

Turn raw GPU compute into a managed inference platform. White-label, multi-tenant, Anthropic- and OpenAI-compatible.

PRIVATE ENTERPRISE

Move off managed AI services.

Stop paying per token on Bedrock and Azure AI Studio. The same managed experience on your own GPUs, at a fraction of the cost.

UNIVERSITIES AND RESEARCH

Stretch every GPU dollar.

Share GPUs across departments and run dozens of models on the hardware you already fund.

RUN IT ANYWHERE

Deploy on-premises

Your data centre. Your GPUs.

Run in colocation

Right next to your data.

With Amaze Intelligence on QORINAI capacity

Dedicated GPUs, no build.

Five tests for production AI — and how the stack answers them.

01	**Who controls and secures your GPUs?**	Dedicated GPUs in your hall, a QORINAI private tenancy or sovereign cloud; operated by an ISO 27001 team.
02	**Does every GPU earn its ROI?**	Multiplexing holds up to 5× models per GPU; flat monthly cost replaces per-token bills.
03	**Do you own your inference?**	Open-weight and proprietary models under your version control — no training on your data.
04	**Is it modular and flexible?**	Metal or managed Kubernetes; appliance, tenancy or cloud; one node to 1,024 GPUs.
05	**Can you adapt and deploy new models?**	Bring any open-weight model; fine-tune on dedicated GPUs; switch models in under 500 ms.

Next step: a discovery session — define the problem statement, map your workloads, models and data, and leave with a cluster and efficiency plan.

QORINAI

SEE IT ON REAL GPUS

Bring your documents and databases. We will run your AI inside your network.

Test-drive IntraLLM in HyperLabs or on a QORINAI tenancy — with Amaze Intelligence running the cluster and GreenThread AI serving the models.

CONTACT

QORINAI Pty Ltd

enquiry@qorinai.ai · qorinai.ai

A/200 Bourke Road, Alexandria NSW 2015, Australia · ACN 702 633 789

amaze

GreenThread

Partner figures are as stated by the respective partners and are indicative. This presentation is for discussion only and is not an offer.