



THE QUANTUM ALLIANCE

Sovereign Advanced Computing Infrastructure

Internal Technical Architecture & Know-How Manual

The Quantum Alliance Corporation — Latin America
Version 1.0 · June 2026

STRICTLY CONFIDENTIAL

Table of Contents

TQA Sovereign Advanced Computing Infrastructure

Internal Technical Architecture & Know-How Manual

The Quantum Alliance Corporation — Latin America

Version 1.0 · June 2026 · STRICTLY CONFIDENTIAL

Document control

Field	Value
Classification	Strictly Confidential — Internal / Advisor / Technical Partner use only
Owner	Office of the President / Chief Architecture
Audience	TQA technical leadership, prospective CTO/CIO/CISO advisors, infrastructure & delivery partners, vendor evaluation, technical investors
Status	v1.0 — Reference baseline. Supersedes the standalone "Cloud Orchestrator Architecture Proposal" as the controlling architecture
Companion deliverables	(2) *TQA Sovereign AI, Data & Quantum Infrastructure — Client Offering* (deck); (3) *TQA Control Center* (interactive demo + Figma spec); (4) Reference architecture in Figma
Source basis	TQA Investor Deck/Memo (2026), Solution End-to-End (June 2026), Cloud Orchestrator Architecture Proposal, Technology Strategy (R&D), plus external cited research (June 2026)

How to read this document

Throughout, three labels are used deliberately and kept separate, per TQA's quality rules:

- **FACT** — verifiable from an internal source document or an external citation (URL given).
- **ASSESSMENT** — TQA architecture analysis and engineering judgment built on those facts.
- **ASSUMPTION / OPEN** — something we believe but have not yet validated, or a decision still owed.

Nothing in the quantum or PQC sections should be read as a claim of present-day "quantum advantage." Where the honest answer is "not yet," the document says so — that discipline is itself a competitive asset with sophisticated buyers.

MASTER EXECUTIVE SUMMARY

1. What the solution is

TQA is building the first **sovereign advanced-computing layer for Latin America**: a single control plane that orchestrates **sovereign micro-data centers, AI GPU clusters, and quantum-computing access**, with a **native post-quantum security plane** underneath and an **AI delivery layer (QuantOS)** on top that stands up **client-owned Small Language Models** inside the client's own environment. The strategic identity is deliberate and repeated in the source material: *"We do not build the steel. We make it sovereign, intelligent and secure."* TQA is an **asset-light orchestrator and delivery company**, not a capital-heavy data-center owner.

This manual converts that one-line thesis into a defensible end-to-end reference architecture, a modular commercial offering, a build/buy/partner plan, a comparables analysis, seven rounds of red-team challenge, and an implementation roadmap.

2. Why it matters now

Three forces converge on the same eighteen-month window (FACT, internal deck + external research):

- **The post-quantum mandate is legally concrete.** NIST finalized FIPS 203/204/205 in August 2024; NIST IR 8547 deprecates RSA-2048/ECC-P-256 by **2030** and disallows classical public-key crypto by **2035**; NSA CNSA 2.0 sets a 2030–2035 transition. "Harvest now, decrypt later" makes long-lived government, banking and energy data **already at risk today**. There is no regional PQC provider to migrate to.
- **AI hit the compute and power wall.** GPU supply is rationed, data-center electricity roughly doubles to ~945 TWh by 2030 (IEA), and ~30% of generative-AI workloads shift to local Small Language Models by 2028. Enterprises increasingly need **owned models on owned compute**.
- **Data is coming home.** The US CLOUD Act and China's National Intelligence Law mean that choosing a hyperscaler or a Chinese cloud is choosing **which foreign jurisdiction** can compel your data — not "sovereign vs exposed." A Microsoft France executive admitted under oath (June 2025) he could not guarantee French data would never be handed to US authorities. Latin America hosts only ~4.8% of global data-center capacity.

Latin America has **40+** of the world's advanced-computing hubs' equivalent demand and **zero** sovereign operators. This is an unserved, not underserved, market, and the first operator to build the backbone defines the category.

3. The architecture we recommend

A **seven-plane reference architecture**, deployable in six models (TQA hub, in-client, third-party colo, hybrid, sovereign-local, federated multi-hub), and consumable as **ten modules** the client can buy whole or in parts:

1. **Control Plane** — the TQA orchestrator: identity, tenancy, quota, self-service, workflow, unified billing/metering across every site and substrate.
2. **Compute Plane** — a **deliberate two-substrate design**: CloudStack + KVM + Ceph for general-purpose multi-tenant VM/IaaS, and **bare-metal Kubernetes + the NVIDIA GPU Operator for AI/GPU workloads**. This is the single most important architectural correction to the original Cloud Orchestrator proposal (see below).
3. **Data Plane** — an open, sovereign, air-gappable lakehouse: Ceph object storage + Apache Iceberg + a self-hostable catalog (Polaris/Nessie) + Trino/Spark, with OpenMetadata/Soda governance and lineage.
4. **AI Plane (QuantOS)** — SLMs (7B–14B, QLoRA-tuned, Apache/MIT weights), enterprise RAG with **GraphRAG traceability**, vector search, model registry, guardrails, human-in-the-loop, agentic workflows; vLLM + KServe serving; MLflow + Langfuse ops.
5. **Quantum Plane** — IonQ access (Forte/Tempo class), a quantum job-management API, hybrid quantum-classical workflows, and — sold honestly — **quantum-inspired optimization on classical hardware as the near-term revenue, not "quantum advantage now."**
6. **Security & PQC Plane** — Zero Trust (NIST SP 800-207), IAM/ABAC, secrets, **non-US customer-held HSM root of trust (HYOK)**, hybrid PQC (X25519MLKEM768) on OpenSSL 3.5 / liboqs, crypto-agility, CBOM-based PQC-readiness service.
7. **Observability & Operations Plane** — telemetry, logging/SIEM, SRE/NOC/SOC, FinOps/metering, incident response, DR — cross-cutting all planes.

4. The modules TQA can sell

Module 1 Sovereign Data Foundation · Module 2 Secure Private/Hybrid Compute · Module 3 Sovereign AI Layer (QuantOS) · Module 4 Orchestration & Control Plane · Module 5 Security & Post-Quantum Readiness · Module 6 Quantum Access · Module 7 Enterprise Integration · Module 8 Industry Solutions · Module 9 Operations/Managed Service · Module 10 Commercial/Marketplace. The client can contract the whole stack or any subset; each module has a setup component, a recurring/managed component, and a usage-metered component.

5. What TQA can sell immediately (real today)

- **PQC-readiness service** (cryptographic discovery → CBOM → migration roadmap → hybrid pilots). Requires no novel cryptography, creates lock-in as system-of-record for the client's crypto estate, and is uniquely urgent given the 2030 deadline. **The single most sellable, lowest-technical-risk product in the portfolio.**
- **QuantOS sovereign-AI delivery** — client-owned SLMs + GraphRAG on the client's own infrastructure, wired into SAP/ServiceNow/Salesforce/SCADA/OSIsoft PI. The oil & gas reference workflow compresses a 4–7-day cycle to ~5 minutes (internal claim — to be substantiated with a reproducible benchmark).

- **Quantum-inspired optimization** for energy dispatch, logistics and portfolio/risk — deployable on classical hardware today, labeled honestly.
- **Sovereign infrastructure anchor contracts** — the Buenos Aires \$12M / 3-year model, where revenue is contracted ahead of capex.

6. What TQA must build

The **crypto-agility abstraction**, the **unified control/billing plane** that federates the two compute substrates and meters across sites, the **QuantOS delivery layer** (RAG/GraphRAG + citation/traceability pipeline + SLM lifecycle), and the **PQC-readiness/CBOM tooling workflow**. These are the defensible IP. TQA should **integrate and operate** the open stack — not author inference engines or table formats.

7. What TQA must buy / partner

GPU silicon and the NVIDIA stack (unavoidable for competitive AI; sovereignty does not extend to the transistor); a **non-US HSM** (Utimaco/Thales-class) as the cryptographic root of trust; colocation and energy partners (LatAm's time-to-power advantage is the real moat); IonQ for quantum and the IonQ networking/QKD pivot (ID Quantique, Capella) for quantum-safe communications; open-weights model partners (Mistral/IBM Granite/Qwen) rather than building a frontier model.

8. Principal risks

(1) **Two-plane operational complexity** — running CloudStack and bare-metal Kubernetes doubles the platform surface; mitigated by shared KVM/Ceph skills, GitOps, unified observability. (2) **GPU multi-tenant isolation in client DCs / colo** — mitigated with MIG + BlueField DPUs + Confidential Computing, dedicated GPUs for high-assurance tenants. (3) **Talent depth in Latin America** for niche platforms — mitigated by leading with the deepest-talent Kubernetes plane and structured training. (4) **Quantum overclaiming** — mitigated by the bright-line honesty rules in Part 3.17. (5) **Single-counterparty concentration** (the neocloud cautionary tale) — diversify anchor demand. (6) **Vendor support concentration** (ShapeBlue for CloudStack) — in-house expertise plus open-source insurance.

9. The deliverables this engagement produces

(1) **This internal technical manual** (Word + Markdown) and the **reference architecture in Figma**; (2) a **client-facing offering deck** (PPTX); (3) the **TQA Control Center** interactive demo (HTML prototype + Figma spec) with 5- and 15-minute demo scripts; (4) a **comparables and research dossier** with citations; (5) **seven red-team rounds** with the improved architecture; (6) an **implementation roadmap** (30/60/90 days; 6/12/24 months).

10. Decisions owed by leadership (detailed in Part 9)

8. **Confirm the two-plane compute design** (CloudStack IaaS + bare-metal K8s) as the controlling architecture, retiring the single-CloudStack proposal. *(Recommended: yes.)*

9. **Confirm "vanilla Kubernetes by default, OpenShift only on paid client demand."**
10. **Select the HSM vendor** for the sovereign root of trust (non-US requirement).
(Recommended shortlist: Utimaco, Thales.)
11. **Approve the honest quantum/PQC messaging line** as binding on all external material.
12. **Fix the public unit-economics claims** — the "~11B tokens/month SLM breakeven" is baseline-dependent and over-precise; restate it before it reaches a technical buyer.
13. **Choose the MVP beachhead** — recommend Neuquén Net-Zero + a PQC-readiness anchor, with Buenos Aires as the infrastructure reference unit.
14. **Set the certification target** — pursue/define a LatAm sovereign-cloud accreditation early (the OVHcloud SecNumCloud playbook).

PART 0 — SOURCE ANALYSIS

This part records exactly what the internal source documents say, what we infer, what is missing, and what contradictions or gaps must be resolved. It is the audit trail for every downstream architectural and commercial decision.

0.1 Documents analyzed

#	Document	Type	What it governs
S1	TQA Investor Deck (2026)	Strategy / fundraising	Vision, problem, market, IonQ, business model, financials
S2	TQA Investor Memo (May 2026)	Strategy / financial	Four-layer business architecture, P&L, layer economics
S3	Solution End-to-End High Level (June 2026)	Product narrative	QuantOS, orchestrator over three substrates, PQC plane, financing logic
S4	Cloud Orchestrator Architecture Proposal	Technical (vendor)	CloudStack/KVM/Ceph four-plane private-cloud design
S5	TQA Technology Strategy (R&D)	Research	Quantum R&D pillars, hub IT & energy infrastructure
S6	Brand assets (logo, palette)	Design	Black + mint-green (#A0F0D0) identity, orbital-globe mark

0.2 What the documents say (extracted)

Strategic thesis (S1–S3). Latin America rents its intelligence and borrows its security: institutions train on foreign clouds, ship sensitive data abroad, and own none of the result; encrypted data is harvested now to be decrypted later; critical workloads are increasingly barred from leaving national borders but the only compute able to run them sits outside them. TQA's answer is a single orchestrator over three substrates (micro data centers, AI GPU clusters, quantum nodes), quantum-safe by design, asset-light by structure.

Product (S3). QuantOS is an **AI delivery layer, not a chatbot**: it stands up agentic workflows inside the client environment, wired into SAP, ServiceNow, Salesforce, SCADA and OSIsoft PI. The output is a **client-owned model — weights and data owned by the client** — with regulatory traceability. Stated economics: past ~11B tokens/month, owning the model beats frontier-API rates; an oil & gas analysis cycle of 4–7 days collapses to ~5 minutes.

Quantum anchor (S1, S3). TQA is the **exclusive regional execution partner for IonQ**. Executed term sheet (March 6, 2026): \$55M total package, 256-qubit 6th-gen system (H2 2027), Edge Appliance live Q3 2026, IonQ 20% capacity buyback, hardware financed by IonQ rather than invoiced.

Business model (S1, S2). Four compounding layers: Infrastructure (\$138M Y5) → Solutions (\$478M) → Platforms (\$252M) → Ecosystem/IP (\$27M); P&L total \$1,097M at Year 5, 61.5% operating margin; Year-1 contracted \$8.5M (~84% of Y1). Pipeline \$112.5M, 42 leads, 6 countries.

Cloud Orchestrator proposal (S4). A four-plane private-cloud design: Management (CloudStack HA + DB cluster + API gateway), Compute (KVM general + GPU pods, N+1, live migration, NUMA-aware), Storage (Ceph 3-node, replica=3, block/object/file/NFS, S3 via RadosGW), Network (8-VLAN segmentation, firewall HA active-active, OOB isolation, Zero Trust default-deny). Security section covers MFA, SSO SAML/OIDC, RBAC, JWT, WAF, secrets vault, IDS/IPS.

Technology strategy (S5). Quantum R&D across four pillars (architectures, communication/networking, sensing/metrology, software/algorithms), plus hub IT and energy-infrastructure requirements; positions a Patagonia (Neuquén) quantum/R&D hub.

0.3 What we assume (and must validate)

- **ASSUMPTION:** "Sovereign micro data centers" are partner/colo facilities TQA orchestrates, not TQA-owned concrete (consistent with the asset-light claim). *Validate the colo/partner pipeline per hub.*
- **ASSUMPTION:** The IonQ "Edge Appliance" is a co-located rack giving low-latency access/queueing to IonQ cloud systems, not an on-prem quantum computer. *Validate exact deliverable and SLAs with IonQ.*
- **ASSUMPTION:** QuantOS today is an integration/delivery layer over open-source components (vLLM, RAG, vector DB) plus TQA glue, not a from-scratch model or engine. *Confirm current codebase maturity.*
- **ASSUMPTION:** "Net-zero / carbon accounting" Neuquén work is a data + AI + reporting solution on the sovereign data foundation, with optional quantum-inspired optimization. *Confirm scope with COPADE/Provincia.*

0.4 What is missing (gaps to close)

15. **No reference architecture above the single-DC orchestrator.** S4 describes one private cloud; there is no multi-hub federation, no control/billing plane spanning sites, no data/AI/quantum/PQC plane engineering. *This manual supplies it (Part 2–3).*
16. **GPU/AI substrate under-specified.** S4 treats GPUs as KVM pods; it does not address Kubernetes, the NVIDIA GPU Operator, MIG partitioning, or multi-tenant GPU isolation — the core of a competitive AI service.
17. **PQC plane named but not engineered.** S3 asserts "post-quantum security plane" but no algorithms, key custody model, HSM choice, crypto-agility design, or migration methodology.
18. **Data architecture absent.** "Sovereign data" is a slogan; there is no lakehouse/lineage/governance design, residency model, or ingestion pattern.
19. **Enterprise integration unspecified.** SAP/SCADA/PI/core-banking are named as connection points without patterns, and SAP's June-2026 ODP-RFC restriction is a live risk not addressed.
20. **Unit-economics fragility.** The "~11B tokens/month" SLM breakeven is over-precise and baseline-dependent (see Part 3.10 and Part 7 Round 6).
21. **Operations model (SRE/NOC/SOC), DR/BC, SLAs, compliance/audit** — not defined.
22. **Multi-tenancy & metering** across the federation — not defined.

0.5 Contradictions / tensions to manage

- **"We don't build the steel" vs a CapEx-heavy private-cloud proposal (S4).** The Cloud Orchestrator proposal reads like an owned single-DC build; the corporate thesis is asset-light orchestration across partner infra. **Resolution:** treat S4 as the *blueprint for a hub cell* that can be deployed on partner/colo hardware or inside a client DC — not as a mandate to own data centers. The control/billing plane (Part 3.22–3.24) is what makes it asset-light and multi-site.
- **Financing narrative differs across documents.** The older deck (S1) implies a larger equity raise to buy hardware; the newer Solution/Memo (S3) reframes to "\$6M + \$6M equity for the company, hardware financed by IonQ, debt unlocked by contracts." **Resolution:** the S3 "golden rule" (equity never touches depreciating iron) is the controlling commercial logic and is strongly supported by the external GPU-depreciation evidence (Part 6).
- **Quantum framing risk.** Marketing language ("quantum-safe by default," quantum nodes) must not drift into implying present-day quantum advantage. **Resolution:** the honesty rules in Part 3.17 are binding.
- **Pipeline figures differ between deck versions** (e.g., Neuquén ~\$22M vs commercial track ~\$10M net-zero). **Resolution:** treat the commercial track's working numbers as authoritative for proposals; reconcile in the Master Working Context.

0.6 What we must validate before external publication

(a) IonQ deliverable specifics and SLAs; (b) colo/energy partner LOIs per planned hub; (c) the 4–7-day → 5-minute O&G benchmark, reproducibly; (d) the SLM breakeven restatement; (e) Bolivia/Colombia data-localization legal specifics (flagged by research as secondary-sourced); (f) current QuantOS codebase maturity vs the capabilities described here.

PART 1 — AGENT WAR ROOM

Per the engagement brief, the work was organized as a war room of specialist squads (the "army of agents"). Below is the operating structure, what each squad analyzed, and its principal conclusion. Squads A–N map to the brief; four were executed as live, separately-tasked deep-research agents with full external citations (M-class research), feeding the architecture squads.

1.1 Squad roster and verdicts

A. Strategy & Sovereign Infrastructure. Analyzed the sovereignty thesis against the CLOUD Act, the LatAm policy landscape, and the neocloud market structure. **Verdict:** the defensible claim is **jurisdiction + operational sovereignty + local key custody**, not data residency alone; the "US-or-China is still a foreign overlord" symmetry argument is TQA's sharpest wedge; PQC + quantum bundled with sovereignty is genuine white space no incumbent occupies regionally.

B. Enterprise Architecture. Designed the seven-plane reference architecture and the six deployment models; resolved the asset-light-vs-owned tension via the hub-cell + federation pattern. **Verdict:** decouple compute into two substrates; federate via a thin control plane; keep hub cells reproducible so they deploy identically in TQA hubs, client DCs, and colo.

C. Data Architecture. Specified the open sovereign lakehouse (Ceph + Iceberg + Polaris/Nessie + Trino/Spark + OpenMetadata/Soda). **Verdict:** only an open, self-hostable stack can be simultaneously air-gappable and client-owned; avoid MinIO (archived), proprietary table formats, and phone-home catalogs.

D. AI / ML / LLM. Specified QuantOS: 7B–14B SLMs, QLoRA on distilled data, vLLM + KServe serving, hybrid retrieval + GraphRAG traceability, evaluation gates, HITL. **Verdict:** the **graph-backed citation/traceability pipeline is the real differentiator**; fix the breakeven messaging; lead with Apache/MIT-licensed weights for genuine client ownership.

E. GPU / Compute Infrastructure. Established the bare-metal-Kubernetes-for-AI rule, MIG partitioning, and DPU tenant isolation. **Verdict:** GPUs belong on bare-metal Kubernetes under the NVIDIA GPU Operator; CloudStack 4.21 GPU support is preview-grade and must not be the AI plane.

F. Cloud Orchestration. Ran the eight-platform substrate comparison. **Verdict:** keep CloudStack + KVM + Ceph as the VM/IaaS plane (Argentina's ARSAT national cloud is the in-country precedent); add bare-metal Kubernetes for AI; do **not** build on Broadcom-VMware (cost/lock-in/sovereignty); MicroCloud/Proxmox for edge.

G. Storage / Network / Observability. Validated Ceph as the unified storage substrate; specified the segmented network, observability stack, and SIEM. *Verdict:* keep S4's Ceph and VLAN segmentation; add full observability (metrics/logs/traces) and SIEM as a first-class plane, not an afterthought.

H. Security / Compliance / PQC. Engineered the Zero Trust + sovereignty + PQC plane. *Verdict:* HYOK with a **non-US, customer-held HSM** makes compelled decryption technically impossible — TQA's single strongest argument; hybrid PQC (X25519MLKEM768) on OpenSSL 3.5/liboqs; build the crypto-agility abstraction; the CBOM/PQC-readiness service is the most sellable near-term product.

I. Quantum Computing. Produced the honest quantum posture. *Verdict:* IonQ is a genuine trapped-ion fidelity leader but still NISQ/pre-fault-tolerant and behind Quantinuum on logical qubits; sell **access + quantum-inspired optimization + quantum-safe comms (the IonQ QKD pivot)** now, never "advantage now."

J. Enterprise Integrations. Specified ingestion patterns per system. *Verdict:* SAP is the make-or-break, hostile integration (June-2026 ODP-RFC restriction) — standardize on SLT + CDS/OData or a certified connector; OT/PI is ingest-only at the L3.5 DMZ behind a diode; Salesforce/ServiceNow are routine; banking gives best-structured data but tightest residency.

K. Product / UX / Demo. Defined the TQA Control Center information architecture, 25 screens, and demo scripts (delivered in Phase 3). *Verdict:* one operator pane across all planes; dark-premium aesthetic on the mint-green brand; the demo must make sovereignty **visible** (key custody, data residency, PQC score).

L. Commercial / Offering / Pricing. Defined the ten modules and the setup/recurring/usage packaging. *Verdict:* lead with PQC-readiness + QuantOS as land-and-expand; price on outcomes and capacity, not commodity GPU-hours where prices are collapsing.

M. Research / Benchmark / Comparables. Executed as four live cited-research agents (orchestration; GPU/sovereign cloud; data/AI/SLM/RAG; PQC/quantum). *Verdict:* feeds Part 6; the market has bifurcated into capital-heavy GPU owners and asset-light orchestrators, and TQA's lane is the asset-light one NVIDIA itself retreated to.

N. Red Team / Challenge. Ran the seven challenge rounds (Part 7). *Verdict:* the architecture survives challenge with specific corrections recorded per round; the biggest residual risks are operational complexity, talent, and concentration — all mitigable.

1.2 How the squads interlock

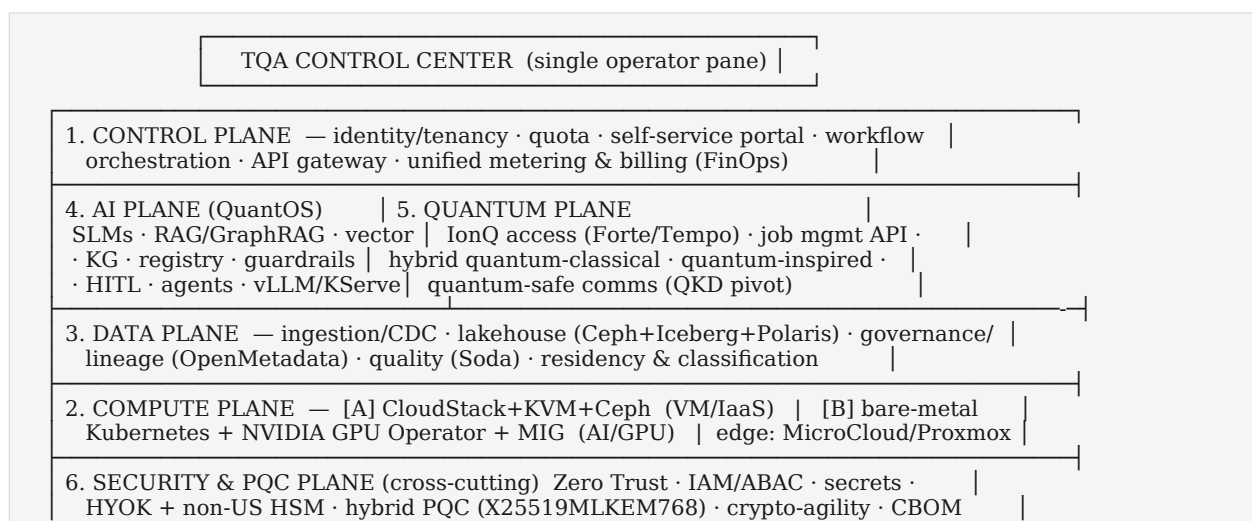
Strategy (A) sets the sovereignty claim → Security/PQC (H) makes it technically true (HYOK + jurisdiction) → Data (C) and AI (D) make it valuable (client-owned models on client-owned data) → Compute (E/F) and Storage/Network (G) make it run → Orchestration/Control (F + B) makes it sellable and multi-site → Integrations (J) make it real inside enterprises → Quantum (I) and Commercial (L) make it differentiated and monetizable → Research (M) and Red Team (N) keep it honest. The reference architecture in Part 2 is the synthesis.

PART 2 — REFERENCE ARCHITECTURE

2.1 Architecture principles (the rules every decision obeys)

23. **Sovereign by construction, not by label.** For every workload we can name *where the data lives, where the model runs, who holds the keys, and who can audit it.* If we cannot, it is not sovereign.
24. **Asset-light by structure.** TQA owns software, contracts, orchestration IP and key custody; partners/colos own the depreciating iron. Equity never touches hardware.
25. **Two substrates, one control plane.** VM/IaaS and AI/GPU are different physics; unify them at the control/billing/identity layer, not the hypervisor.
26. **Open core, no dead-ends.** Prefer Apache/MIT/foundation-governed components; every client must have a credible exit path (open formats, standard APIs). Sovereignty must not become a new lock-in.
27. **Crypto-agile and quantum-safe by default.** Algorithms are configuration, not architecture. Hybrid PQC on by default.
28. **Modular and reproducible.** A "hub cell" is a repeatable unit that deploys identically in a TQA hub, a client DC, or a colo. The client can buy the whole stack or any module.
29. **Honest about maturity.** Quantum value is sold as access + quantum-inspired + quantum-safe comms today; advantage is roadmap. AI value requires data, training and integration — stated plainly.
30. **Operate for 24/7.** HA, DR, SLAs, SRE/NOC/SOC and observability are designed in from day one, not bolted on.

2.2 The seven planes (end-to-end)



7. OBSERVABILITY & OPERATIONS PLANE (cross-cutting) metrics/logs/traces · SIEM · SRE/NOC/SOC · incident mgmt · DR/BC · SLA & audit

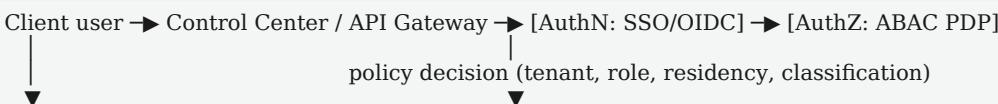
▲ all of the above instantiate as a repeatable "HUB CELL", deployable in:
| TQA hub · client DC · third-party colo · hybrid · sovereign-local · federated

Planes 6 and 7 are **cross-cutting**: every other plane is enforced by Zero Trust/PQC and observed by the operations plane. Planes 4 and 5 sit side-by-side because AI and quantum workflows interleave (a classical AI pipeline can call a quantum or quantum-inspired optimization step).

2.3 Plane-by-plane summary (engineered in depth in Part 3)

- **Control Plane** — TQA orchestrator. Federates identity (SSO/OIDC), tenancy (account/domain/project), quota, a self-service portal, workflow orchestration, an API gateway (REST/JSON, signed requests, rate-limit/WAF), and a **unified metering/billing service** that reconciles CloudStack Usage Server (VM plane) + Kubecost/OpenCost (AI plane) + quantum job-seconds + PQC-service seats into one chargeback view.
- **Compute Plane** — two substrates. *[A]* CloudStack + KVM + Ceph for general VMs, tenant networks, live migration. *[B]* Bare-metal Kubernetes + NVIDIA GPU Operator for training/inference with MIG hardware partitioning. Edge footprints use MicroCloud or Proxmox.
- **Data Plane** — ingestion (Debezium/Kafka/Airbyte) → lakehouse (Ceph object + Iceberg tables + Polaris/Nessie catalog + Trino/Spark/DuckDB compute) → governance (OpenMetadata lineage, Soda/Great Expectations quality) with residency tags and classification driving placement.
- **AI Plane (QuantOS)** — SLM registry and lifecycle (QLoRA fine-tuning, distillation, eval gates), vector + knowledge-graph retrieval (Qdrant/pgvector + Neo4j), hybrid search + reranking, citation/traceability pipeline, guardrails, HITL, agentic workflows, served via vLLM + KServe, observed via MLflow + Langfuse.
- **Quantum Plane** — a quantum job-management API abstracting IonQ (and, where it helps a client, Braket/Azure backends), hybrid quantum-classical orchestration, a quantum-inspired optimization service on classical hardware (the near-term revenue), and a quantum-safe-comms track aligned to the IonQ QKD pivot.
- **Security & PQC Plane** — Policy Decision/Enforcement points, ABAC policy, microsegmentation, secrets/vault, the **HYOK + non-US HSM root of trust**, hybrid-PQC TLS and signing, crypto-agility abstraction, and the CBOM-driven PQC-readiness workflow.
- **Observability & Operations Plane** — OpenTelemetry metrics/logs/traces, a SIEM, SRE runbooks, a NOC for infrastructure and a SOC for security, incident management, DR/BC orchestration, and the SLA/audit evidence store.

2.4 Control flow (how a request travels)

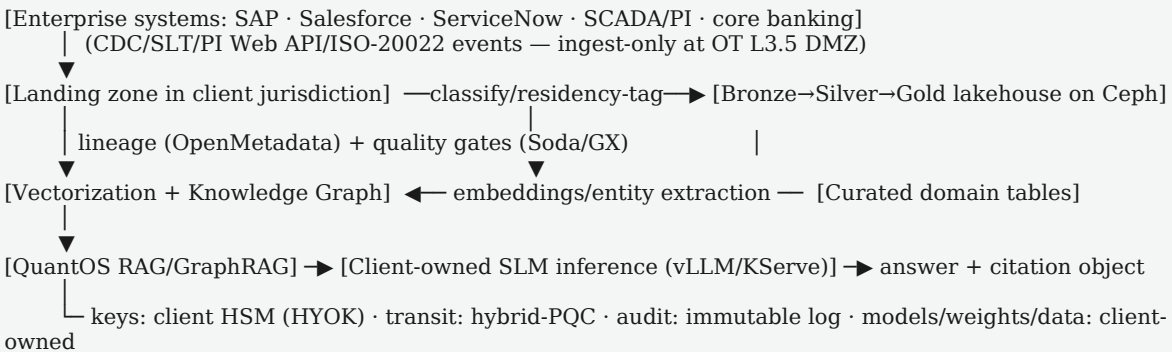


Self-service request (e.g. "fine-tune SLM on dataset X in-country")

- ➔ Data Plane: resolve dataset residency/classification → stage in client-DC lakehouse only
- ➔ Compute Plane[B]: schedule QLoRA job on MIG GPU partition (tenant-isolated)
- ➔ AI Plane: register model in client-owned registry; run eval gates; attach citation pipeline
- ➔ Security/PQC: keys from client HSM (HYOK); all transit hybrid-PQC; every action audit-logged
- ➔ Observability/Billing: meter GPU-seconds + storage + seats → unified chargeback

Every step is policy-gated and audit-logged; nothing crosses a residency boundary without an explicit policy decision.

2.5 Data flow (sovereign by construction)



The crucial property: raw data and the model never leave the client's jurisdiction or control; TQA orchestrates and operates, the client owns.

2.6 The six deployment models

Model	Where compute/data live	When to use	Sovereignty posture
TQA Hub	TQA-orchestrated partner/colo hub in-country	Clients without their own DC; shared sovereign capacity	TQA operates; client holds keys (HYOK); in-country jurisdiction
In-Client	Inside the client's own data center	Banks, government, energy with existing DCs and strict residency	Client-owned hardware; TQA delivers/operates as managed service
Third-Party Colo	Independent colo (neutral provider)	Fast standup, no client DC, energy-advantaged sites	TQA operates; jurisdiction = colo's country; keys client-held
Hybrid	Critical local + burst to public cloud	Spiky workloads where non-sensitive burst is economical	Sensitive data/models stay local; only non-sovereign workloads burst
Sovereign-Local (air-	Fully disconnected	Defense/intel, highest-	No external connectivity;

Model	Where compute/data live	When to use	Sovereignty posture
gapped)	enclave	assurance government	entirely client-controlled
Federated multi-hub	Several hubs across LatAm under one control plane	TQA network scale; cross-border enterprise; capacity pooling	Per-hub residency honored; control plane federates without moving data

ASSESSMENT: the **In-Client** and **Colo** models are what make TQA asset-light and let it deploy faster than capital-heavy hyperscalers; the **hub-cell** reproducibility (Part 2.2) is what makes all six the **same product** with different placement.

2.7 Client maturity paths (how a client adopts)

- **Crawl** — "Make me quantum-safe and show me my data." PQC-readiness (CBOM + roadmap) + Sovereign Data Foundation. Low risk, urgent, fast ROI. No GPU commitment.
- **Walk** — "Give me a client-owned copilot." QuantOS SLM + GraphRAG on one or two enterprise systems, in-client or hub. First AI value; begins the stickiness.
- **Run** — "Run my sovereign AI estate." Multi-system integration, agentic workflows, managed compute (hub or in-client), full observability and SOC, quantum-inspired optimization for a vertical use case.
- **Fly** — "We're a sovereign platform." Federated capacity, vertical SaaS platform, quantum pilots, marketplace, IP licensing.

2.8 Final architecture recommendation (the controlling decision)

Adopt the seven-plane architecture with the two-substrate compute plane as the controlling reference, superseding the single-CloudStack Cloud Orchestrator proposal. Keep CloudStack + KVM + Ceph for VM/IaaS (proven, open, in-country precedent via ARSAT); add bare-metal Kubernetes + NVIDIA GPU Operator + MIG for AI/GPU; federate both via a thin TQA control/billing plane; make the PQC plane real with HYOK + a non-US HSM + hybrid PQC; and instantiate everything as a reproducible hub cell deployable across the six models. The detailed engineering of each plane, the vendor comparisons, the build/buy/partner calls, and the risks follow in Part 3.

PART 3 — INTERNAL TECHNICAL DOCUMENT

This part is the engineering core. Sections 3.1–3.40 follow the brief's required structure. FACT/ASSESSMENT/ASSUMPTION labels are used throughout; external citations are collected in Part 6.

3.1 Executive technical summary

TQA's technical product is a **federated, multi-tenant, sovereign advanced-computing platform** expressed as seven planes and ten modules, instantiated as a reproducible hub cell across six deployment models. The defensible engineering IP is concentrated in four places: (1) the **control/billing plane** that federates two compute substrates and meters across sites; (2) the **QuantOS delivery layer** (RAG/GraphRAG + citation/traceability + SLM lifecycle); (3) the **crypto-agility abstraction + PQC-readiness workflow**; and (4) the **sovereignty governance model** (HYOK + non-US HSM + jurisdiction). Everything else is integration of best-of-breed open components. The platform sells from the security/data layers (urgent, low-risk) upward into AI and quantum (differentiated, higher-margin).

3.2 Strategic architecture principles

(See Part 2.1 — eight principles.) The three that most shape engineering trade-offs: **sovereign by construction** (drives HYOK + open formats + residency-aware placement), **two substrates one control plane** (drives the compute split and the federation layer), and **open core no dead-ends** (drives the rejection of proprietary table formats, Broadcom-VMware, and any phone-home dependency).

3.3 Full end-to-end reference architecture

(Diagram and flows in Part 2.2–2.6.) The architecture is a layered stack with two cross-cutting planes (security/PQC and observability/operations). Requests enter through the control plane, are policy-gated by ABAC, and fan out to data, compute, AI and quantum planes; every action is keyed by the client HSM, transported under hybrid PQC, metered, and audit-logged.

3.4 Architecture diagram descriptions (for the Figma build)

The Figma reference set (companion deliverable 4) renders: (D1) the **seven-plane stack**; (D2) the **two-substrate compute plane**; (D3) the **sovereign data flow**; (D4) the **control/security flow**; (D5) the **six deployment models** as placement variants of one hub cell; (D6) the **module map**; (D7) the **client maturity path**; (D8) the **PQC migration flow**; (D9) the **hybrid quantum-classical workflow**. Each diagram uses the black/mint-green brand, with planes as horizontal bands and the control center as the top pane.

3.5 Layer-by-layer system breakdown

Plane	Core components (open)	Commercial options	TQA-built glue
Control	Keycloak (IAM), Kong/APISIX (gateway), Temporal (workflow),	—	Federation + unified billing/metering

Plane	Core components (open)	Commercial options	TQA-built glue
	custom portal		
Compute A	CloudStack, KVM, Ceph	ShapeBlue support	Hub-cell IaC, tenant templates
Compute B	Kubernetes, NVIDIA GPU Operator, MIG, Kueue	OpenShift (on demand)	GPU scheduling policy, tenant isolation
Data	Ceph RGW, Iceberg, Polaris/Nessie, Trino/Spark/DuckDB, OpenMetadata, Soda	—	Residency-aware placement, classification
AI (QuantOS)	vLLM, KServe, MLflow, Langfuse, Qdrant/pgvector, Neo4j, BGE reranker	NIM (client-elected)	RAG/GraphRAG + citation pipeline, SLM lifecycle
Quantum	Qiskit/Braket SDKs, IonQ API	IonQ access, Azure/AWS quantum	Job-mgmt API, quantum-inspired service
Security/PQC	Keycloak, Vault, OpenSSL 3.5, liboqs, EJBCA	Non-US HSM (Utimaco/Thales)	Crypto-agility layer, CBOM workflow
Observability/Ops	OpenTelemetry, Prometheus/Mimir, Loki, Grafana, Wazuh/OpenSearch SIEM	—	Runbooks, SLA/audit store

3.6 Deployment models

(Detailed in Part 2.6.) Engineering note: the same IaC (Terraform/OpenTofu + Ansible/Helm + GitOps via Argo CD) provisions a hub cell in any of the six models; the difference is a placement profile (network egress, key location, connectivity, federation membership), not a different codebase. This is what keeps the asset-light promise credible operationally.

3.7 Client maturity paths

(Detailed in Part 2.7 — Crawl/Walk/Run/Fly.) Each stage maps to a contract structure (Part 3.40 commercial) and a technical footprint, so sales and delivery share one model.

3.8 Data architecture

Design (ASSESSMENT, grounded in cited research). A 100% client-owned, air-gappable lakehouse is now technically credible end-to-end:

- **Ingestion.** Log-based CDC via **Debezium on Kafka Connect**; **Airbyte (self-hosted)** or Meltano for the connector long tail; **Flink** for stateful streaming. Batch via Spark. Never **Fivetran** (cloud-only — conflicts with residency).
- **Storage.** **Ceph RGW** (S3-compatible) is the sovereign object substrate — foundation-governed, no single-vendor capture. *Do not* default to MinIO (OSS repo archived/maintenance mode early 2026).
- **Table format.** **Apache Iceberg** (the de-facto open table format; REST catalog API gives clients a vendor-neutral exit path). Delta only where an installed base demands it; Hudi for upsert-heavy CDC.
- **Catalog.** **Apache Polaris** (REST, multi-engine, self-hostable) or **Nessie** (git-style branching). Unity Catalog OSS where multi-format is needed.
- **Compute/query.** Trino (interactive), Spark (ETL), DuckDB (embedded).
- **Governance & lineage.** **OpenMetadata** (column-level lineage, data contracts) over DataHub for easier self-host. **Quality:** Soda (declarative monitoring) + Great Expectations (ingestion gates).

How the ten data questions are answered. Data enters via CDC/streaming/batch; is **classified** at landing (public/internal/confidential/regulated) and **residency-tagged**; cleaned/validated by quality gates; governed by OpenMetadata policies; stored bronze → silver → gold on Ceph; vectorized into Qdrant/pgvector and entity-extracted into Neo4j; connected to models via QuantOS RAG; sovereignty preserved by residency-aware placement + HYOK; lineage and audit are continuous and immutable.

What must be OSS vs SaaS. The sovereign core (object storage, table format, catalog, compute, governance/audit) is self-hosted in the client jurisdiction; only non-sensitive conveniences may be SaaS, and even those self-hosted for a true air-gap.

MVP vs Phase 2. MVP: Ceph+Iceberg+Polaris+Trino/Spark/DuckDB + OpenMetadata + Soda. Phase 2: Nessie branching, Great Expectations gates, audit hardening, data-mesh domain decomposition for multi-domain clients.

3.9 AI architecture (QuantOS)

Serving & LLMops. **vLLM + KServe** (LLMInferenceService) on the bare-metal Kubernetes plane + NVIDIA GPU Operator; **MLflow** (tracking/registry) + **Langfuse** (LLM observability). Avoid as *core* dependencies the recurring taxes TQA sells against: **NVIDIA NIM/AI Enterprise** (~\$4,500/GPU/yr) and **Seldon** (relicensed BSL — not OSS); offer NIM only as a client-elected paid add-on. HF **TGI is in maintenance mode** — do not start new deployments on it. Phase 2 adds **llm-d** (multi-node, heterogeneous GPU), **Ray/KubeRay** (distributed/agent), **Kueue** (multi-tenant GPU quotas), **Kubeflow Pipelines** (retraining), optional **TensorRT-LLM** backend.

When to use what (the decision the brief demands).

- **RAG** when the client needs current, citable, access-controlled knowledge from their own corpus — the default for enterprise copilots and regulated answers.
- **Fine-tuning (QLoRA/LoRA)** when the client needs domain tone, format, or task behavior the base model lacks — not for injecting facts (that's RAG's job).

- **SLM (7B–14B)** as the default product unit: fits a single A100/H100, cheap to serve, genuinely client-ownable on Apache/MIT weights.
- **Frontier external model** only for non-sensitive, low-volume, high-complexity tasks where no local model suffices, via the hybrid model — never with regulated data.
- **Hybrid** as the steady state: local SLM for the bulk, RAG for knowledge, optional frontier call for edge cases, with a router enforcing data-classification policy.
- **Distillation** to "pay frontier rates once, own the capability forever" — a frontier teacher generates filtered synthetic data to train the client's SLM.

Quality, hallucinations, cost, privacy. Evaluation gates (RAGAS faithfulness ≥ 0.75 as a grounding gate; DeepEval CI regression; TruLens live monitoring) + domain gold sets + HITL. Hard product rules: **no answer ships without a citation object; abstention is a first-class behavior.** Cost metered per workload (GPU-seconds + tokens). Privacy by construction: data and weights never leave the client jurisdiction; keys are client-held.

3.10 SLM / RAG / meta-learning architecture

SLM landscape & licensing (FACT). Lead with permissive weights for genuine ownership: **IBM Granite 3.x (Apache 2.0), Qwen2.5 7B/14B (Apache 2.0 for most sizes), Phi-4 (MIT), Mistral/Ministral (Apache 2.0)**; frame Llama's Community license honestly. Hardware: 7B ~4–5 GB at Q4 (any 24 GB card); 14B fits a 24 GB GPU at 4-bit; one 80 GB A100/H100 serves any 1–14B at FP16 with batching. Tuning: **QLoRA** (~14 GB for a 7B, recovers ~80–90% of full-FT quality) is the default; distillation is the proven recipe for domain capability.

The breakeven correction (this must be fixed before any external use). The internal "~11B tokens/month" SLM breakeven is **directionally sound but baseline-dependent and over-precise.** It holds only versus the *cheapest hosted-open APIs*; versus **premium frontier APIs (GPT-4o/Claude), owning a GPU wins at ~0.5–1.5B tokens/month.** **Approved restatement:** "A tuned SLM beats premium frontier APIs at ~0.5–1.5B tokens/month per GPU, and beats even the cheapest API tier in the high-single-digit-billions/month range — and the residency/control value pulls the effective breakeven lower for regulated clients." * Quoting "11B" unqualified invites a teardown.

Meta-learning / continuous learning. Feedback loops: capture HITL corrections and high-signal interactions → curated fine-tuning datasets → periodic QLoRA refresh → eval-gated promotion in the model registry. This is **continual fine-tuning with governance**, not online learning on raw traffic (which would leak and drift). Each refresh is versioned, evaluated, and reversible.

3.11 GPU cluster architecture

The rule (FACT-backed). GPUs for AI belong on **bare-metal Kubernetes** under the **NVIDIA GPU Operator** (driver, container toolkit, device plugin, DCGM, MIG Manager as pods). The advanced scheduling TQA needs — **MIG hardware partitioning (up to 7 isolated instances per GPU on A100/H100/H200/B200), DRA (GA in K8s 1.34 / OpenShift 4.21), NIM/KServe autoscaling** — is delivered in Kubernetes, not in any IaaS layer. **Do not** nest the GPU Kubernetes layer inside CloudStack GPU-passthrough VMs (NVIDIA's own tooling assumes bare-metal GPU nodes).

Workload placement.

- **GPU bare metal:** training, fine-tuning, large-batch and latency-sensitive inference.
- **GPU virtualized (vGPU/MIG-backed):** VDI, light inference, dev — acceptable via CloudStack vGPU for non-critical cases.
- **CPU:** data engineering, ETL, classical analytics, orchestration, most enterprise app logic.
- **Kubernetes-required:** all model serving, agentic workflows, autoscaled inference.
- **Bare metal required:** multi-node training (InfiniBand/RoCE fabric), confidential-computing tenants.

Multi-tenant GPU isolation (the sovereign crux). Default to **dedicated GPUs or MIG partitions per tenant; ban naive cross-tenant time-slicing** (no memory/fault isolation; documented side-channel risk). For high-assurance: **MIG + H100/Blackwell Confidential Computing (TEE, VRAM/PCIe encryption, attestation) + BlueField DPUs** for zero-trust bare-metal isolation (the CoreWeave/NCP pattern) — essential when operating inside client DCs and colo.

Scaling, reservation, monetizing idle, metering. Scale per client via Kueue quotas and node pools; reserve capacity against signed contracts (never build ahead of demand); monetize idle via the value-cascade (frontier training → production inference → batch/analytics) and by routing partner idle capacity; meter per-GPU-second and per-MIG-instance via Kubecost/OpenCost into the unified billing plane.

3.12 Orchestration architecture

The decision (the brief's central orchestration question). Two substrates, federated:

- **VM/laaS: CloudStack + KVM + Ceph** — keep it. Open (Apache 2.0), low lock-in, mature multi-tenancy (Accounts/Domains/Projects), native GPU-aware Usage Server, **direct Argentine sovereign precedent (ARSAT national cloud)**, ShapeBlue São Paulo support, light enough to deploy in client DCs and colo.
- **AI/Containers: vanilla bare-metal Kubernetes + NVIDIA GPU Operator** by default; **OpenShift only as a paid, client-facing SLA tier** on demand.
- **Edge: MicroCloud or Proxmox** for small single-tenant footprints.

Why not the alternatives (summary; full matrix Part 3.26). OpenStack — operationally heaviest, wrong trade for distributed client-DC/colo deployment (use only if a client mandates SCS/OpenStack sovereignty alignment). VMware Cloud Foundation — **disqualified** post-Broadcom on cost (documented 800–1,500%+ increases), lock-in, and a CISPE "not sovereign" determination; the VMware exodus is TQA's market. Nutanix — credible HCI but proprietary/per-core lock-in, wrong values fit. Proxmox — weak multi-tenancy and ~32-node scale ceiling; edge only. SUSE Harvester — architecturally closest single-pane (VMs+containers+MIG, GA Jan 2026) but too young + Rancher Prime price hike; run a 90-day PoC, don't bet the core yet.

Multi-tenant, self-service, APIs, integration. Multi-tenancy at the control plane (Keycloak identity + per-tenant CloudStack projects + per-tenant K8s namespaces/vClusters); self-service via the portal + signed REST API gateway; integration with billing (unified metering), observability (per-tenant dashboards), and security (ABAC policy at every boundary).

3.13 Private cloud architecture

The hub cell's private-cloud substrate keeps the strong bones of the Cloud Orchestrator proposal (S4) and corrects its gaps:

Take from S4: Ceph 3-node distributed storage (block/object/file, replica=3, S3 via RadosGW), 8-VLAN segmentation with default-deny, firewall HA active-active, OOB (IPMI/BMC) isolation, CloudStack HA management + DB cluster + API gateway, KVM live migration, NUMA-aware placement.

Add/change: the bare-metal Kubernetes GPU plane (S4 treats GPUs as KVM pods only); explicit MIG/DPU tenant isolation; the data/AI/PQC/observability planes entirely (absent in S4); the federation/control plane spanning hubs; storage tiering (NVMe hot → HDD warm → object cold) and erasure-coded pools for capacity; backup/DR targets (VLAN 70) wired to a tested DR runbook (Part 3.20).

3.14 Enterprise integration architecture

Per-system patterns (FACT-grounded).

- **SAP — make-or-break, actively hostile.** SAP Note 3255746 (enforced June 2026) blocks ODP-RFC for third-party extraction. **Standardize on SLT (trigger-based) + CDS-view/ODP-OData landed locally**, or a certified connector (Theobald/BryteFlow/Qlik). Budget the OData performance penalty (~10× slower than RFC). Treat SAP's cloud-leaning "BDC" path as a **sovereignty red flag**.
- **Salesforce / ServiceNow — routine.** Salesforce Bulk API 2.0 + Change Data Capture; ServiceNow Table API + Export/Import Sets, OAuth, 15–60 min incremental pulls.
- **OT / SCADA / OSIsoft–AVEVA PI — safety boundary.** Egress only via **PI Web API / PI Integrator (native Kafka target) / MQTT Sparkplug B**, ingest-only at the **Purdue Level 3.5 DMZ behind a data diode**; never write back to plant control.
- **Core banking — best data, tightest residency.** **ISO 20022** structured messages (mandatory for cross-border SWIFT since Nov 2025) are ideal AI input; consume via CDC/events and keep strictly in-country.

Cross-cutting plumbing. Self-hosted **Debezium + Kafka + Flink + Airbyte/Meltano** as the event/connector backbone; identity federation via OIDC/SAML; an event bus (Kafka) for near-real-time sync. **Priority connectors (product base):** SAP, Salesforce, ServiceNow, OSIsoft/AVEVA PI, core-banking ISO 20022, plus a generic REST/OData and Kafka connector. The rest are Phase 2 / professional-services.

3.15 (continued in 03b) — Security, PQC, Quantum, Observability, Operations, DR, Compliance, Multi-tenancy, API, Billing, Vendor comparison, Build/Buy/Partner, Alternatives, Risks, Open decisions, Roadmap, Team, Costs, Appendix, Glossary.

3.15 Security architecture

Zero Trust (NIST SP 800-207). "Never trust, always verify"; protect resources not perimeters; per-session access; dynamic policy. Logical core: **Policy Decision Point** (Policy Engine + Administrator) and **Policy Enforcement Points** at every resource boundary, with **microsegmentation** down to the workload. Aligned to CISA's Zero Trust Maturity Model v2.0 pillars (Identity, Devices, Networks, Applications/Workloads, Data).

IAM. ABAC (NIST SP 800-162) as the primary model — subject/resource/action/environment attributes drive fine-grained, context-aware decisions (the natural fit for dynamic ZTA policy), with RBAC for coarse role bundling. SSO via SAML/OIDC (Keycloak); MFA mandatory; least privilege (SP 800-53); **dynamic, short-lived credentials** and a secrets vault (HashiCorp Vault) with rotation and per-secret access.

Network & application. Microsegmentation per VLAN/namespace; encrypted east-west; default-deny firewalls; IDS/IPS; WAF + signed API requests + JWT validation at the gateway; no hardcoded secrets.

Data protection. In transit: hybrid-PQC TLS (Part 3.16). At rest: AES-256 with keys from the client HSM; envelope encryption; field-level encryption for regulated data. Everything audit-logged to an immutable store.

3.16 Post-Quantum Cryptography architecture

Standards (FACT). NIST finalized **FIPS 203 (ML-KEM)**, **204 (ML-DSA)**, **205 (SLH-DSA)** in Aug 2024; **FIPS 206 (FN-DSA/Falcon)** expected late-2026/early-2027; **HQC** (code-based backup KEM) selected Mar 2025, standard ~2027. **Deadlines:** NIST IR 8547 deprecates RSA-2048/ECC-P-256 by **2030**, disallows classical public-key by **2035**; NSA CNSA 2.0 runs 2030–2035. **Harvest-now-decrypt-later** makes long-lived data at risk today (Mosca's inequality).

TQA PQC plane design.

31. **Hybrid by default** — classical + PQC (IETF TLS **X25519MLKEM768**) so an attacker must break both; aligned to NIST/BSI/ANSSI transition guidance. Built on **OpenSSL 3.5** (native ML-KEM/ML-DSA/SLH-DSA, hybrid by default, LTS to 2030) and **Open Quantum Safe (liboqs/oqs-provider)** for stacks not yet on 3.5.
32. **Root of trust = non-US, customer-held HSM** (Utimaco "Quantum Protect" — German, field-activatable, ahead on CAVP; or Thales Luna). This preserves the CLOUD-Act argument (Part 3 sovereignty).
33. **Crypto-agility abstraction (TQA-built IP)** — algorithm choice is a *configuration*: pluggable providers, negotiation, config-driven swaps, so adding FN-DSA/HQC in 2027 is a config change, not a re-architecture. NIST/NCCoE frame this as the part you must own.

34. **PQC-readiness service (the sellable product).** Five stages: **discovery/inventory** → **CBOM (CycloneDX)** → **risk assessment (prioritized by data confidentiality lifespan)** → **hybrid pilots** → **remediation** backed by crypto-agility. Requires no novel crypto; creates lock-in as system-of-record for the client's cryptographic estate.

Build/Buy/Partner. *Buy:* validated primitives (OpenSSL 3.5, liboqs, AWS-LC) + non-US HSM. *Buy/self-host:* open CLM/discovery (EJBCA) on TQA-controlled infra (avoid hosting a US vendor's SaaS at the trust root). *Build:* the crypto-agility abstraction + CBOM workflow. *Partner:* quantum-entropy sources and IonQ QKD adjacency.

3.17 Quantum computing architecture (and the honesty rules)

IonQ reality (FACT). Trapped-ion (Yb-171), all-to-all connectivity, best-published two-qubit fidelity (99.99% via Oxford Ionics), long coherence — but **~1,000–10,000× slower gates** than superconducting (physics, not engineering), and **still NISQ / pre-fault-tolerant**. Current flagship **Tempo (~100 physical qubits, #AQ 64, Sept 2025)**; available via direct API + **AWS Braket, Azure Quantum, Google Cloud**. 2025–26: acquisitions pivoting toward **quantum networking + quantum-safe comms** (Oxford Ionics, ID Quantique, Capella satellite QKD, Lightsynq, SkyWater foundry). Quantinuum currently leads on demonstrated **logical** qubits (Helios, 48 logical from 98 physical).

TQA quantum plane. A **quantum job-management API** abstracting IonQ (and Braket/Azure where it helps a client), **hybrid quantum-classical orchestration** (classical pipeline calls a quantum/quantum-inspired step), a **quantum-inspired optimization service on classical hardware** (the near-term revenue), an **education/upskilling layer**, and a **quantum-safe-comms track** aligned to the IonQ QKD pivot (QKD-distributed keys feeding HYOK/HSM custody = a coherent "sovereign quantum-safe communications" story unique to the IonQ partnership).

Realistic use cases. *Today (mostly quantum-inspired / classical hardware):* logistics routing, energy dispatch/unit-commitment, portfolio/risk optimization — deployable now, labeled honestly. *Hardware pilots (research/upskilling):* small optimization (QAOA), chemistry/ML sampling. *Future (fault-tolerant, ~2029–2030+):* cryptanalysis (drives PQC urgency), large-scale chemistry/materials, finance Monte-Carlo.

Binding honesty rules (apply to ALL external material). (1) Never call quantum-inspired "quantum computing." (2) Always distinguish **logical vs physical** qubits. (3) Always cite a **strong classical baseline** (Gurobi/CPLEX, not "greedy"). (4) Sell IonQ hardware as **access + research + sovereign regional availability**, not production advantage. (5) Never stake credibility on IonQ's 2028–2030 logical-qubit roadmap in government proposals. **Honesty here is a competitive moat with sophisticated buyers.**

3.18 Observability and operations architecture

Observability. OpenTelemetry instrumentation across all planes; metrics (Prometheus/Mimir), logs (Loki), traces (Tempo), dashboards (Grafana); per-tenant views surfaced in the Control Center. GPU telemetry via DCGM. AI observability via Langfuse. Cost via Kubecost/OpenCost + CloudStack Usage Server reconciled in billing.

SIEM & security ops. Wazuh/OpenSearch (or equivalent) SIEM; correlation rules mapped to MITRE ATT&CK; immutable audit store feeding compliance evidence.

3.19 SRE / NOC / SOC model

A **follow-the-sun** model scaled to footprint: an **SRE** function owning reliability (SLOs, error budgets, runbooks, on-call), a **NOC** for infrastructure health (capacity, hardware, network, DR drills), and a **SOC** for security monitoring and incident response (SIEM triage, threat hunting, IR). Managed-service tiers (Part 3.40) determine whether TQA, the client, or a shared model staffs each function. AIOps assists triage and capacity forecasting but does not replace human IR.

3.20 Disaster recovery and business continuity

Per hub cell: replica=3 Ceph + cross-zone snapshot copy; DB cluster with automatic failover and point-in-time recovery; backup jobs to an isolated target (VLAN 70). Cross-hub: the control plane treats hubs as DR regions where residency permits; **RPO/RTO are set per workload tier** (e.g., Tier-1 regulated: RPO ≤ 15 min, RTO ≤ 1 hr). **DR is tested on a schedule** (game-days), not assumed. Air-gapped enclaves use offline backup rotation.

3.21 Compliance and auditability

Every action is policy-gated and immutably logged (who, what, when, which keys, which residency decision). Evidence maps to client frameworks: government procurement rules, financial-services regulation, ISO 27001, SOC 2, and emerging LatAm sovereignty law (Bolivia/Colombia localization — flagged for primary-source legal validation). Target a **LatAm sovereign-cloud accreditation** early (the OVHcloud SecNumCloud playbook) as both a moat and a procurement entry ticket. EU AI Act full enforcement (Aug 2026) is relevant for clients operating into the EU.

3.22 Multi-tenancy model

Three isolation tiers: (1) **soft** — shared cluster, namespace/project isolation, ABAC quotas (dev/non-sensitive); (2) **hard** — per-tenant vCluster/CloudStack project, dedicated GPU/MIG, network microsegmentation (enterprise default); (3) **dedicated/sovereign** — dedicated hub cell or in-client deployment, client-held keys, optional air-gap (government/banking). Tenancy identity, quota and billing are federated at the control plane so a tenant is the **same entity** across compute substrates and sites.

3.23 API model

A **signed REST/JSON API** at the gateway (Kong/APISIX) with OIDC tokens, rate limiting, and WAF; resource APIs per plane (provisioning, data, model, quantum job, key/PQC, billing, observability); an

AWS-compatible layer (EC2/S3) via CloudStack for lift-and-shift; webhooks/events via Kafka. Everything the Control Center UI does is an API call, so the platform is automatable and partner-embeddable.

3.24 Billing / metering model

A **unified metering service** reconciles four meters into one chargeback: CloudStack Usage Server (VM/laaS), Kubecost/OpenCost (GPU-seconds, per-MIG), quantum job-seconds, and PQC/AI seats. Supports setup fees, recurring subscriptions, usage-based consumption, and managed-service tiers (Part 3.40). FinOps dashboards surface cost per tenant, per workload, per hub. This service is **TQA-built IP** — it is what turns federated open components into a sellable platform.

3.25 Hardware / software options

Software: the open stack in Part 3.5, plus the commercial options (ShapeBlue support, OpenShift on demand, non-US HSM, IonQ access, optional NIM). **Hardware:** NVIDIA GPUs (H100/H200/B200 class; A100 acceptable for inference/MIG); InfiniBand/RoCE fabric for multi-node training; BlueField DPUs for tenant isolation; liquid cooling mandatory at Blackwell density (a partner-site qualification bar); Ceph on NVMe+HDD tiers. Procurement is partner/colo-led to keep depreciation off TQA's book.

3.26 Vendor comparison (decision matrices)

Orchestration substrate (●●● strong / ●●○ adequate / ●○○ weak, for TQA's use case).

Platform	Maturity	GPU/AI	Multi-tenant	Cost/Lock-in	LatAm talent	Fit	Verdict
CloudStack+KVM+Ceph	●●●	●○○	●●●	●●●	●●○	●●●	Keep — VM/laaS plane
Bare-metal K8s+GPU Operator	●●●	●●●	●●○	●●●	●●●	●●●	Adopt — AI plane
OpenStack+Ceph	●●●	●●○	●●●	●●○	●●○	●●○	Only if client mandates
OpenShift	●●●	●●●	●●○	●●○	●●●	●●○	Paid SLA tier on demand
VMware Cloud Foundation	●●●	●●●	●●●	●○○	●●○	●○○	Reject (Broadcom)
Nutanix	●●●	●●○	●●●	●○○	●●○	●●○	Skip unless mandated
Proxmox	●●●	●●○	●○○	●●●	●●○	●○○	Edge only

Platform	Maturity	GPU/AI	Multi-tenant	Cost/Lock-in	LatAm talent	Fit	Verdict
SUSE Harvester	●●○	●●○	●●○	●●○	●●○	●●○	PoC, revisit 12 mo

Data platform. Open lakehouse (Iceberg+Trino+Polaris on Ceph) = the sovereign default; Databricks/Snowflake can't air-gap; Palantir Foundry can but Ontology + per-install pricing fight "client-owned." **Storage.** Ceph (sovereign default) over MinIO (archived), NetApp/Pure/Dell (proprietary), cloud object (non-sovereign). **Serving.** vLLM+KServe (default) over NIM/Seldon (taxes) and TGI (maintenance). **Vector DB.** Qdrant/pgvector (default) over Weaviate (outbound vectorizers — air-gap liability) and Pinecone (SaaS). **HSM/PQC.** Non-US Utimaco/Thales root of trust; US AWS/Google/IBM PQC referenced as proof-of-readiness, never as custodian. **Quantum.** IonQ anchor + optional Braket/Azure backends.

3.27 Build / buy / partner recommendations

Capability	Decision	Rationale
Control/billing federation plane	Build	Defensible IP; nothing open does cross-substrate cross-site metering
QuantOS RAG/GraphRAG + citation pipeline	Build	The real AI differentiator; traceability is the moat
Crypto-agility abstraction + CBOM workflow	Build	NCCoE says own this; PQC-readiness lock-in
SLM lifecycle (QLoRA/distillation/eval)	Build (on open tools)	Glue over MLflow/vLLM; client-ownership story
Lakehouse, serving engines, vector/graph DBs	Buy/Integrate (open)	Don't author engines or formats
GPU silicon + NVIDIA stack	Buy	Unavoidable for competitive AI
HSM root of trust	Buy (non-US)	Sovereignty constraint
Colo + energy capacity	Partner	LatAm time-to-power moat; asset-light
Quantum hardware + QKD	Partner (IonQ)	Exclusive regional anchor
Frontier models	Partner (open-weights: Mistral/Granite/Qwen)	Don't build a frontier model

3.28 What to TAKE from the Cloud Orchestrator proposal

Ceph distributed storage (block/object/file, replica=3, S3 via RadosGW); 8-VLAN segmentation with default-deny; firewall HA active-active (<1s failover); OOB isolation; CloudStack HA management + DB cluster + API gateway; KVM live migration and NUMA-aware placement; the Zero-Trust posture (MFA, SSO, RBAC, JWT, WAF, secrets vault, IDS/IPS).

3.29 What to CHANGE from the Cloud Orchestrator proposal

Add the **bare-metal Kubernetes GPU plane** (don't run AI on KVM pods); add **MIG/DPU tenant isolation**; add the entire **data, AI, PQC, observability** planes (absent); add the **federation/control/billing plane** spanning hubs (it's single-DC today); upgrade **ABAC** over plain RBAC; specify **storage tiering + erasure coding + tested DR**; make the **PQC plane real** (HYOK + non-US HSM + hybrid PQC + crypto-agility), not just named; size GPU/network for **multi-node training** (InfiniBand/RoCE), not just passthrough.

3.30 Alternative architecture options (considered and rejected/deferred)

(1) **Single-CloudStack everything** (the proposal as-is) — rejected: CloudStack 4.21 GPU is preview-grade; the NVIDIA AI ecosystem is Kubernetes-native. (2) **OpenStack-only** — deferred: too operationally heavy for distributed client-DC/colo. (3) **Kubernetes-only with KubeVirt for VMs** — viable single-pane, but service-provider multi-tenancy/billing/networking less mature than CloudStack at scale; revisit as Harvester matures. (4) **VMware** — rejected (Broadcom). (5) **Frontier-API-first AI** — rejected for sovereign clients (data leaves, no ownership). (6) **Vendor-managed PQC SaaS at the trust root** — rejected (undercuts sovereignty).

3.31 Risks and mitigations (register)

#	Risk	Sev	Mitigation
R1	Two-plane operational complexity	High	Shared KVM/Ceph skills; GitOps/IaC; unified observability; staff K8s first
R2	GPU multi-tenant isolation in client DC/colo	High	MIG + BlueField DPU + Confidential Computing; dedicated GPUs for high-assurance
R3	LatAm talent depth for niche platforms	Med	Lead with K8s (deepest pool); ShapeBlue/Proxmox training; nearshore bench
R4	Quantum overclaiming → credibility loss	High	Binding honesty rules (3.17); strong baselines
R5	Single-counterparty concentration	High	Diversify anchor demand; cap any tenant <50%
R6	CloudStack support concentration (ShapeBlue)	Med	In-house expertise; ASF open-source insurance; SLA as backstop
R7	SAP integration restriction (ODP-RFC)	Med	SLT+CDS/OData or certified connector; budget OData penalty
R8	GPU supply/economics; NVIDIA dependence	Med	Partner-financed iron; AMD MI300X hedge for inference; reserved

#	Risk	Sev	Mitigation
			capacity vs contracts
R9	Unit-economics claim teardown (11B tokens)	Med	Restate per 3.10 before any external use
R10	Sovereignty-washing exposure	High	Prove with certification + HYOK, not branding; in-jurisdiction operators
R11	Energy/site availability per hub	Med	LOIs with colo/energy partners before commit; time-to-power as selection criterion
R12	DR untested / availability gaps	Med	Scheduled DR game-days; per-tier RPO/RTO; SLAs

3.32 Open decisions

(1) Confirm two-plane compute as controlling. (2) Vanilla-K8s-default vs OpenShift policy. (3) HSM vendor selection (non-US). (4) Approve quantum/PQC honesty line as binding. (5) Restate SLM breakeven. (6) MVP beachhead (Neuquén + PQC anchor recommended). (7) Certification target. (8) AMD-hedge scope for inference. (9) Harvester PoC go/no-go. (10) Air-gapped enclave offering scope. (Resolution recommendations in Part 9.)

3.33–3.40 Roadmap · Build plan · Evolution · Team · Costs · Appendix · Glossary

- **3.33 Implementation roadmap / 3.34 30-60-90 / 3.35 6-12-24 months** — see Part 8.
- **3.36 Team required.** Core: Head of AI/ML (funded by the round), Chief Architect, Platform/SRE leads (K8s + CloudStack), Data Engineering lead, Security/PQC lead, Integration engineers (SAP/OT/banking), Quantum solutions lead, Delivery PM, SOC analysts. Leverage LatAm talent arbitrage (5–10× lower cost) and nearshore pods; lead hiring with the deepest-talent Kubernetes/data roles.
- **3.37 Cost drivers.** GPU capex (partner-financed), power/cooling (PUE + \$/kWh — LatAm advantage), network fabric, HSM, IonQ access amortization, talent, support contracts. The asset-light model externalizes depreciation; TQA's book carries software, contracts, IP.
- **3.38 Technical appendix.** Reference component versions; hub-cell IaC layout (OpenTofu modules + Helm charts + Argo CD apps); network VLAN table (from S4, extended with K8s pod/service CIDRs and DPU fabric); RPO/RTO tier table; eval-gate thresholds; CBOM schema (CycloneDX).
- **3.39 Glossary.** ABAC, AQ (Algorithmic Qubits), CBOM, CDC, CNSA 2.0, Crypto-agility, DPU, FIPS 203/204/205, GraphRAG, HNDL, HYOK/BYOK, IaC, MIG, ML-KEM/ML-DSA/SLH-DSA, NISQ, PDP/PEP, PQC, QKD, QLoRA, RAG, SLM, vLLM, Zero Trust. (Full definitions in the Word edition appendix.)
- **3.40 Commercial layer (offering).** Setup (assessment, deployment, integration, model build) · Recurring (platform subscription, managed service, support tiers) · Usage (GPU-seconds, tokens, quantum job-seconds, storage) · Professional services (PQC readiness,

integration, vertical solutions) · Quantum access · IP licensing (Year 4+). Packaging detail in Part 4 and the client deck.

PART 4 — CLIENT-FACING OFFERING (summary; full deck = Deliverable 2)

The client narrative — *TQA Sovereign AI, Data & Quantum Infrastructure* — is delivered in full as the PPTX deck (Phase 2). Its spine, in plain C-level language:

The problem. You rent your intelligence and borrow your security. Your models train on someone else's cloud, your sensitive data leaves the country to do it, and when the contract ends the intelligence leaves with the vendor. Your encrypted data is being harvested today to be decrypted later. And choosing AWS or Huawei is choosing *which foreign government* can compel your data — not whether one can.

Why now. PQC deadlines are legally fixed (2030/2035). AI shifted to owned models on owned compute. Data is being forced home by law. The first eighteen months define the regional standard.

What TQA is. A sovereign advanced-computing layer for Latin America: one control plane over sovereign micro-data-centers, GPU clusters and quantum access; client-owned AI built inside your environment; quantum-safe by default. We don't build the steel — we make it sovereign, intelligent and secure.

What you can buy (the ten modules) — whole or in parts: Sovereign Data Foundation · Secure Private/Hybrid Compute · Sovereign AI (QuantOS) · Orchestration & Control · Security & Post-Quantum Readiness · Quantum Access · Enterprise Integration · Industry Solutions · Operations/Managed Service · Marketplace.

What's in a hub; what installs in you; what we manage. A hub gives shared sovereign capacity in-country; we can install the same hub cell inside your data center; or operate it for you as a managed service in a neutral colo — same product, your choice of placement and key custody.

How sovereignty is real, not a label. Your data stays in your jurisdiction; your model and its weights are yours; *you hold the keys in your own HSM so we are technically unable to read your data*; everything is quantum-safe and audit-logged. By jurisdiction (non-US operator), by architecture (you hold the keys), and by algorithm (post-quantum) — a triple no hyperscaler can match.

Use cases by industry. Banking (PQC migration, ISO-20022 AI, sovereign copilots) · Government (sovereign data + citizen services + net-zero/carbon accounting) · Oil & Gas (wells-engineering copilots, the 4–7-day → 5-minute workflow, quantum-inspired optimization) · Energy (grid dispatch optimization, net-zero reporting) · Logistics (routing optimization) · Mining/Agribusiness (forecasting,

optimization) · Healthcare/Pharma (private RAG on clinical/regulatory corpora) · Smart cities (mobility/energy) · Defense/Security (air-gapped sovereign AI + quantum-safe comms).

ROI/TCO and comparisons. Versus hyperscalers (disqualified from sovereign procurement by foreign law; data leaves); versus consultancies (front-end someone else's cloud, own nothing); versus telcos (real estate without orchestration or a quantum-safe layer); versus pure-play quantum (frontier hardware, no near-term security or AI delivery). TQA owns the empty top-right quadrant: sovereign + full-stack hybrid compute.

Next steps for a client. A PQC-readiness assessment + a sovereign-data quick-start is the low-risk entry (Crawl); a client-owned copilot is the first AI win (Walk).

The deck renders this with the dark-premium mint-green brand, one idea per slide, no jargon without architecture behind it.

PART 5 — DEMO / PROTOTYPE SPEC (summary; full demo = Deliverable 3)

TQA Control Center — a navigable, dark-premium prototype (HTML + Figma spec, Phase 3) that makes sovereignty *visible*. Information architecture and the 25 screens:

35. Secure login / SSO · 2. Client home dashboard · 3. Hub & infrastructure overview · 4. **Data sovereignty dashboard** (residency map, classification, who-can-touch-this) · 5. Data ingestion flows · 6. Connected enterprise systems · 7. AI model registry (client-owned) · 8. Local SLM dashboard · 9. RAG/knowledge-base + **citation/traceability viewer** · 10. Agentic workflows · 11. GPU cluster usage (per-MIG) · 12. Compute allocation · 13. Cost/usage/FinOps · 14. Security posture · 15. **PQC readiness score** + CBOM · 16. Encryption/key-management status (HYOK) · 17. Quantum access module · 18. Hybrid quantum-classical workflow · 19. Observability · 20. Alerts/incident response · 21. Compliance & audit logs · 22. Marketplace/modules · 23. Client support · 24. Executive ROI dashboard · 25. Industry-specific workflow (O&G wells-engineering reference).

Roles: Executive (ROI, posture), Platform admin (compute, tenancy), Data steward (residency, lineage), Security/CISO (PQC, keys, audit), AI engineer (models, RAG, eval), Quantum analyst (jobs).

Demo scripts (delivered with the prototype): a tight **5-minute** path (login → sovereignty dashboard → client-owned SLM answer with citation → PQC score → quantum-inspired optimization result → ROI) and a **15-minute** technical path (adds ingestion, model registry/eval, GPU/MIG allocation, key custody, hybrid quantum workflow, audit log). Each screen carries: what the salesperson says, what the technical buyer should understand, what the investor should understand.

Visual style: black canvas, mint-green (#A0F0D0) accents and the orbital-globe mark, monospaced data, generous negative space, "advanced-computing" aesthetic. Realistic dummy data (Neuquén net-zero, a bank PQC migration, an O&G copilot).

PART 6 — COMPARABLES & RESEARCH (cited dossier)

Condensed from four live research squads (full citations in the research appendix). Labels: **FACT** (cited) vs **ASSESSMENT**.

6.1 The market has bifurcated (ASSESSMENT)

Capital-heavy GPU **owners** (CoreWeave, Nscale, Crusoe, Nebius, OCI) vs asset-light **orchestrators** (RunPod, Together AI, NVIDIA DGX Cloud Lepton, Lightning AI). **TQA's lane is the asset-light one NVIDIA itself retreated to** after trying capital-heavy DGX Cloud. Margin is migrating from bare-metal rental to inference/orchestration (DigitalOcean: ~70% of AI ARR from inference); raw \$/GPU-hr is collapsing (H100 ~\$8 → ~\$2 in two years).

6.2 Neocloud comparables (FACT, 2025–26)

CoreWeave: 2025 rev **\$5.13B (+170%)**, backlog ~\$99B, ~250k GPUs, >1 GW; Microsoft ~67% of revenue (concentration warning). Nebius: Q1-2026 rev **\$399M (+684%)**, ARR guide **\$7–9B**, sold out. Nscale: **\$14.6B** valuation, ~\$14B Microsoft GB300 deal — *closest sovereign-narrative comp*. Together AI: ~\$1B ARR at ~\$7.5B on far less capital — *closest asset-light template*. RunPod: explicitly asset-light, two-tier supply (Secure/sovereign + Community), \$1B valuation on ~\$22M raised — *closest model analog; the gap it doesn't fill (sovereignty/PQC) is TQA's*. **Public H100 on-demand June 2026:** ~\$1.99 (Voltage Park) to ~\$6.16 (CoreWeave) — a 3–4× spread that *is the orchestrator's arbitrage surface*.

6.3 Sovereign cloud (FACT + ASSESSMENT)

Residency ≠ sovereignty ≠ immunity from foreign law (the 2024–25 Ontario-vs-OVH ruling proves commercial presence can create jurisdiction). The credible non-capital-intensive winners are **Tier-2 operators — T-Systems Sovereign Cloud, Delos Cloud** — local legal entities operating partner infra under local law **with local key custody**. **That is exactly TQA's model**. Oracle EU Sovereign Cloud's "**isolated decision-making entity designed to refuse foreign requests**" is the governance pattern to copy. **PQC precedent:** NxtGen + Thales brought PQC to India's sovereign cloud (May 2025) — proof buyers pay for "sovereign *and* quantum-safe." **AI-delivery pattern:** Cohere North runs air-gapped on as few as 2 GPUs; Mistral supplies open-weights sovereign reference architectures — adopt, don't build a frontier model.

6.4 GPU economics (FACT)

H100 capex ~\$25–40k/card, ~\$200–400k/8-GPU server; per-GPU capex is **rising** generation-over-generation while rental prices **fall** — a widening squeeze that is the core argument for asset-light. Power is the binding constraint (IEA: DC electricity ~doubles to ~945 TWh by 2030; US interconnection 4–10 yrs). **LatAm time-to-power and cheap clean baseload (Patagonian hydro/wind, Chilean curtailed solar ~6 TWh wasted in 2024, Vaca Muerta gas) is TQA's strongest non-GPU differentiator** — but state honestly: Stargate Argentina is hydro/wind-positioned, not "Vaca Muerta gas powers OpenAI." The GPU-depreciation debate (Burry, Nov 2025) favors the non-owner: depreciation risk sits with whoever owns the iron.

6.5 LatAm context (FACT)

Brazil leads on sovereign-AI policy (PBI 2024–28 ~\$4.2B; Santos Dumont supercomputer; REDATA renewable-DC tax regime) — likely first beachhead, and its state cloud still runs on foreign hardware (a citable gap TQA fills). **CAF and IDB are the connective tissue behind nearly every multilateral effort — the highest-leverage relationships.** The CLOUD-Act "symmetry" argument (US-or-China is still a foreign overlord) and the June-2025 Microsoft-France sworn admission are the most weaponizable points in any government procurement debate.

6.6 Data/AI stack (FACT)

A 100% client-owned, air-gappable open stack is now credible: Iceberg has won the table-format war; the components that turned commercial (MinIO archived, Seldon BSL, TGI maintenance, NIM licensing) all have clean open replacements — and being the alternative to those taxes **is** TQA's pitch. The graph-backed citation/traceability pipeline is the genuine differentiator.

6.7 PQC / Quantum (FACT)

NIST FIPS 203/204/205 final (Aug 2024); deadlines 2030/2035; hybrid PQC live at scale (Cloudflare: >2/3 of human TLS traffic now hybrid ML-KEM). HSMs PQC-ready (Entrust nShield 5 CAVP-validated; Utimaco Quantum Protect; Thales Luna). IonQ: real trapped-ion fidelity leader, Tempo #AQ 64, but NISQ/pre-fault-tolerant and behind Quantinuum on logical qubits; pivoting to quantum networking/QKD. ****No peer-reviewed quantum *advantage* exists for commercial optimization today**** — near-term value is quantum-**inspired** on classical hardware. TQA's honesty here is a moat.

PART 7 — RED TEAM CHALLENGES (seven rounds)

Each round lists the critiques, the adjustment made to the architecture, the residual trade-off, and the resulting improvement. The architecture in Parts 2–3 already incorporates these changes; this part is the audit trail.

Round 1 — Architecture Challenge

Critiques. (1) The original single-CloudStack design is a single point of architectural failure for AI — CloudStack 4.21 GPU support is a Technical Preview validated on one GPU model. (2) Running two substrates risks overengineering and a doubled operational surface. (3) "Federation" was hand-waved — unclear interfaces between hubs. (4) The hub cell could become a snowflake per deployment. (5) Live migration and N+1 don't help a GPU training job that spans nodes.

Adjustments. Split compute into CloudStack-IaaS + bare-metal-K8s-AI with explicit, documented interfaces (identity, quota, billing federated at the control plane only — Part 3.22–3.24). Made the hub cell **reproducible via one IaC codebase + placement profiles** (Part 3.6) to prevent snowflakes. Specified InfiniBand/RoCE fabric and Kueue gang-scheduling for multi-node training (Part 3.11). Defined federation as a thin control-plane membership, not data movement (Part 2.6).

Residual trade-off. Two substrates is more surface than one; accepted because the alternative (single platform) sacrifices either AI competitiveness (CloudStack-only) or service-provider multi-tenancy/billing maturity (K8s-only). Revisit single-pane (Harvester) in 12 months.

Improvement. AI competitiveness without losing IaaS maturity; no architectural SPOF for AI; reproducible deployments.

Round 2 — Security / PQC Challenge

Critiques. (1) "PQC plane" was a label with no algorithms, key model, or HSM. (2) BYOK does not defeat the CLOUD Act — the provider still holds technical access. (3) RBAC alone is too coarse for dynamic, residency-aware policy. (4) Tenant isolation on shared GPUs is unproven against side-channels. (5) Hosting a US vendor's PQC SaaS at the trust root would undercut the entire sovereignty claim. (6) Secrets/lateral-movement not addressed.

Adjustments. Engineered the PQC plane (Part 3.16): hybrid PQC (X25519MLKEM768) on OpenSSL 3.5/liboqs, **HYOK with a non-US customer-held HSM** (making compelled decryption technically impossible), crypto-agility abstraction, CBOM workflow. Upgraded to **ABAC** (Part 3.15). Mandated MIG + DPU + Confidential Computing for shared-GPU tenants (Part 3.11). Forbade US-vendor SaaS at the trust root (Part 3.16, 3.26). Added Vault + short-lived credentials + microsegmentation against lateral movement.

Residual trade-off. HYOK shifts key-management burden to the client and complicates support (TQA literally cannot recover client data) — acceptable and indeed the selling point for sovereign buyers; document operational runbooks for key loss.

Improvement. Sovereignty becomes *technically true and demonstrable*, not branding — TQA's single strongest argument.

Round 3 — Enterprise Buyer Challenge (CIO/CTO/CISO/CFO/Procurement)

Critiques. *CISO:* "Prove you can't read my data." *CIO:* "I have SAP and a 20-year-old core — your AI is useless if it can't reach them." *CTO:* "Who operates this at 3am?" *CFO:* "Where's the TCO vs my current cloud, and what's the lock-in?" *Procurement:* "What certifications, what SLAs, what exit clause?"

Adjustments. CISO answered by HYOK + audit + air-gap option. CIO answered by the integration architecture (Part 3.14) — and honestly flagging the SAP ODP-RFC restriction with a compliant path. CTO answered by the SRE/NOC/SOC model + managed-service tiers (Part 3.19, 3.40). CFO answered by FinOps/metering transparency + open-format exit paths (no dead-ends) + asset-light TCO (no capex). Procurement answered by the certification target (Part 3.21), per-tier SLAs, and Iceberg/standard-API exit clauses.

Residual trade-off. Certification takes time and money before it's a moat; in the interim, lead with HYOK + reference architectures as proof.

Improvement. Every buyer persona has a concrete, non-hand-wavy answer; objections become discovery questions.

Round 4 — Hyperscaler Challenge ("Why not AWS/Azure/GCP/Oracle?")

Critiques. Hyperscalers have sovereign SKUs (EU Data Boundary, Oracle's EU entities), vastly more scale, mature managed AI, and deep pockets. Why would a client pick a startup?

Honest answer (not slogans). (1) **Jurisdiction:** a US-parented provider remains subject to US process regardless of where data sits — the CLOUD Act is a control test, not a location test; a Microsoft France executive admitted under oath he couldn't guarantee otherwise. Hyperscaler "sovereign" SKUs reduce residency risk but do not remove foreign jurisdiction. TQA, as a LatAm-incorporated operator with HYOK, can make a claim they structurally cannot. (2) **Ownership:** hyperscaler AI rents you intelligence; QuantOS gives you the model and weights. (3) **PQC + quantum bundle:** unclaimed regionally. (4) **Where they can't go fast:** in-client-DC and energy-advantaged colo, asset-light, without a multi-year region build. **Where we concede:** raw scale, breadth of managed services, and price on commodity compute — so we don't compete there; we compete on sovereignty, ownership, and security, and we *interoperate* (hybrid burst) where it's non-sensitive.

Residual trade-off. For non-sovereign, price-sensitive, commodity workloads, hyperscalers win — and TQA should not pretend otherwise; the hybrid model routes those to public cloud.

Improvement. A defensible, honest "why TQA" that survives a sophisticated buyer and a hyperscaler's counter-pitch.

Round 5 — Implementation Challenge

Critiques. (1) The full seven-plane stack is too much to build at once. (2) Two substrates need scarce talent. (3) QuantOS "5-minute workflow" needs proof. (4) The control/billing federation plane is real software, not config. (5) Quantum is not a near-term revenue engine.

Adjustments. Defined a phased MVP (Part 8): **PQC-readiness + Sovereign Data Foundation + one QuantOS SLM use case** first; bare-metal K8s plane before CloudStack expansion (talent depth); the federation/billing plane built incrementally (single-hub metering → multi-hub). Reframed quantum as access + quantum-inspired now, advantage later. Flagged the O&G benchmark for reproducible substantiation (Part 0.6).

Residual trade-off. A phased MVP means the full "platform" story outruns the built reality for a few quarters; manage by selling what's real (PQC, data, one copilot) and roadmapping the rest transparently.

Improvement. A buildable sequence with the hardest-talent and highest-IP pieces staged sensibly; MVP is sellable in months, not years.

Round 6 — Commercial Challenge

Critiques. (1) Pricing on GPU-hours is a losing game (prices collapsing). (2) The "11B tokens/month" breakeven is over-precise and invites teardown. (3) Layer revenue figures (\$138M/\$478M/\$252M/\$27M) are an "architectural grouping" that may not reconcile to the P&L. (4) Long government sales cycles vs startup runway. (5) Managed services don't scale margin like software.

Adjustments. Repriced on outcomes/capacity/seats, not commodity GPU-hours (Part 3.40); **restated the SLM breakeven** (Part 3.10) to ~0.5–1.5B vs frontier / high-single-digit-billions vs cheapest APIs. Kept the layer figures explicitly labeled as architectural groupings reconciled to the \$1,097M P&L (Part 0.2). Addressed sales-cycle/runway with the land-and-expand motion (PQC-readiness as a fast, low-risk first sale) and milestone-gated financing (Part 8). Positioned managed services as the *bridge* to platform/IP margin (the four-layer compounding logic), not the end state.

Residual trade-off. Land-and-expand means lower initial contract values than a single mega-deal; accepted for faster cycles and reference accumulation.

Improvement. Pricing that doesn't erode, claims that survive technical scrutiny, and a commercial motion matched to government/enterprise reality.

Round 7 — Investor / Board Challenge

Critiques. (1) What is defensible IP vs services? (2) What scales and at what margin? (3) Isn't this just systems integration with a quantum sticker? (4) Single-counterparty and partner-dependency risk (IonQ, NVIDIA, colo). (5) Can a LatAm team execute a public-company-grade platform?

Adjustments / answers. Defensible IP: the control/billing federation plane, the QuantOS RAG/GraphRAG + traceability pipeline, the crypto-agility/CBOM workflow, and the sovereignty governance model — not the open components. **Scaling/margin:** the four-layer compounding model (infrastructure → solutions → platforms → IP), each layer structurally higher-margin, with software economics by Year 5; asset-light keeps depreciation off the book. **Not just SI:** the difference is *ownership and IP* — clients own models/keys/data, TQA owns the orchestration/security/delivery IP and recurring platform; pure SIs front-end someone else's cloud and own nothing. **Concentration/dependency:** diversify anchors (<50% any counterparty), AMD hedge for inference, multiple colo/energy partners, IonQ exclusivity as asset but quantum sold honestly so the thesis doesn't depend on its roadmap. **Execution:** operators who have run scale (ex-Accenture/Petrobras/Tenaris/Deutsche Bahn), LatAm talent arbitrage, and milestone-gated capital that de-risks each step.

Residual trade-off. Partner dependencies (NVIDIA silicon, IonQ quantum) are real and largely unavoidable; the mitigation is structural (hedges, honest positioning), not elimination.

Improvement. A board-grade articulation of what's defensible, what scales, what's de-risked, and where the honest residual risks sit — the basis for the \$30M-entry, milestone-stepped-to-\$250M+ story.

7.8 Net effect of the seven rounds

The architecture moved from a single-DC private-cloud proposal with a quantum/PQC label to a **defensible seven-plane, two-substrate, sovereign-by-construction platform** with: a corrected compute design, a real PQC plane (HYOK + non-US HSM + hybrid PQC + crypto-agility), engineered data/AI/integration/ops planes, an honest quantum posture, restated unit economics, a phased buildable MVP, outcome-based pricing, and a board-grade IP/defensibility story. The residual risks (operational complexity, talent, concentration, partner dependency) are named and mitigated, not hidden.

PART 8 — ROADMAP

Sequencing principle: build the **highest-IP, hardest-talent, lowest-technical-risk-to-revenue** pieces first; never build compute ahead of signed demand; lead with what is real today (PQC, sovereign data, one QuantOS use case).

8.1 30 / 60 / 90 days

Days 0–30 — Foundations & first sellable product.

- Stand up the bare-metal Kubernetes + GPU Operator reference cell (deepest talent first) on partner/colo hardware.
- Package the **PQC-readiness service** (discovery → CBOM template → roadmap deliverable). This is sellable with near-zero technical build.
- Stand up the Sovereign Data Foundation reference (Ceph + Iceberg + Polaris + Trino + OpenMetadata).
- Select the HSM vendor (non-US); execute the quantum/PQC honesty line; restate the SLM breakeven.
- Hire the Head of AI/ML (funded by the round).

Days 31–60 — First AI use case & control-plane seed.

- Deliver one **QuantOS SLM + GraphRAG** use case end-to-end (Neuquén net-zero or an O&G copilot) in-client or hub; prove the citation/traceability pipeline.
- Build the single-hub unified metering MVP (CloudStack Usage + Kubecost reconciliation).
- Stand up Keycloak/ABAC + Vault + hybrid-PQC TLS on the reference cell.
- Substantiate the 4–7-day → 5-minute O&G benchmark reproducibly.

Days 61–90 — Productionize & first managed service.

- Wire observability (OTel/Prometheus/Loki/Grafana) + SIEM + first SOC runbooks.
- Convert the first use case to a managed-service contract with SLAs.
- Run the SUSE Harvester PoC (single-pane future option, decide go/no-go).
- Deliver the Control Center demo to the first two pipeline accounts.

8.2 6 / 12 / 24 months

6 months — MVP platform, 1–3 hubs. PQC + Sovereign Data + QuantOS productized; Buenos Aires anchor delivery underway; Neuquén net-zero proposal converted; 1–3 hub cells live; unified billing across them; first quantum-inspired optimization in production for an energy/logistics client; certification process initiated.

12 months — Federation & vertical platforms. Multi-hub federation (control plane spans 3–6 hubs); first vertical SaaS platform productized from delivery IP (energy or finance); IonQ Edge live and quantum-access pilots running; SOC at 24/7 for managed clients; SAP/OT/banking connectors hardened; first debt drawn against contracts; Series-A-ready evidence base.

24 months — Platform & ecosystem. 6-hub footprint; platform revenue attaches with near-zero incremental hardware; IP licensing begins (Year-4 logic pulled forward where possible); marketplace live; quantum hybrid pilots maturing; LatAm sovereign-cloud certification achieved; priced on contracted revenue toward the \$250M+ Series A.

8.3 Scope discipline (MVP / Phase 2 / Phase 3)

- **MVP (real now):** PQC-readiness, Sovereign Data Foundation, one QuantOS use case, single-hub metering, bare-metal K8s GPU cell, the demo.
- **Phase 2 (6–12 mo):** multi-hub federation, vertical platform #1, hardened enterprise connectors, 24/7 SOC, quantum-inspired in production, certification underway.
- **Phase 3 (12–24 mo):** marketplace, IP licensing, additional verticals, quantum hybrid pilots, air-gapped enclave offering, federated capacity pooling.
- **Too ambitious to sell today (roadmap only):** "quantum advantage," fully autonomous agentic operations, a TQA frontier model. Keep these out of contracts.

8.4 RACI (abbreviated)

Workstream	Responsible	Accountable	Consulted	Informed
Compute planes	Platform/SRE lead	Chief Architect	NVIDIA/ShapeBlue	Exec
Data plane	Data Eng lead	Chief Architect	Client data owners	Exec
QuantOS/AI	Head of AI/ML	President	Clients	Board
Security/PQC	Security lead	CISO-advisor	HSM vendor, NIST guidance	Board
Quantum	Quantum lead	President	IonQ	Board
Commercial	Commercial lead	President	Clients/CAF-IDB	Board
Certification	Security lead	COO	Auditors	Board

PART 9 — FINAL RECOMMENDATIONS

9.1 Recommended decisions (resolve these)

36. **Adopt the two-substrate compute plane** (CloudStack-IaaS + bare-metal-K8s-AI) as the controlling architecture; retire the single-CloudStack proposal. *(Recommend: approve.)*
37. **Vanilla Kubernetes by default; OpenShift only as a paid client SLA tier.** *(Approve.)*
38. **Select a non-US HSM** (shortlist Utimaco, Thales) as the sovereign root of trust. *(Decide.)*
39. **Make the quantum/PQC honesty rules (Part 3.17) binding** on all external material. *(Approve.)*
40. **Restate the SLM breakeven** (~0.5–1.5B vs frontier; high-single-digit-billions vs cheapest APIs) before any external use. *(Approve.)*
41. **Choose the MVP beachhead:** Neuquén Net-Zero + a PQC-readiness anchor, with Buenos Aires as the infrastructure reference unit. *(Recommend.)*

42. **Initiate a LatAm sovereign-cloud certification** (OVHcloud SecNumCloud playbook).
(Approve.)
43. **Lead enterprise hiring with Kubernetes/data roles** (deepest LatAm talent) and the funded Head of AI/ML. *(Approve.)*

9.2 Open decisions (owner + by when)

AMD MI300X hedge scope for inference (Architect, 60d); Harvester single-pane go/no-go (Architect, 90d); air-gapped enclave offering scope (Security+Commercial, 90d); per-vertical platform #1 selection (Commercial, 60d); colo/energy partner LOIs per hub (COO, ongoing).

9.3 What to do this week

- Approve decisions 1, 2, 4, 5 (no external cost; unblock everything).
- Issue the HSM RFP to Utimaco and Thales.
- Package the PQC-readiness service one-pager and take it to the two most PQC-urgent pipeline accounts (banking/government).
- Stand up the bare-metal K8s + GPU Operator reference cell on existing partner hardware.
- Brief the team on the binding honesty rules.

9.4 What to ask vendors

- **IonQ:** exact Edge Appliance deliverable, SLAs, queue/latency, QKD-pivot roadmap availability for LatAm, co-marketing rights.
- **HSM (Utimaco/Thales):** CNSA 2.0 algorithm coverage, FIPS 140-3 + PQC validation status, field-activation, HYOK/XKS integration, LatAm support.
- **ShapeBlue:** CloudStack LATAM support SLA, GPU roadmap beyond preview, training for TQA staff.
- **Colo/energy partners:** power availability and \$/kWh, PUE, liquid-cooling readiness, time-to-power, jurisdiction.

9.5 What to show clients

The Control Center demo (sovereignty made visible: residency map, HYOK key custody, PQC score, client-owned SLM answer with citation, quantum-inspired optimization result, ROI) + the client offering deck + a PQC-readiness assessment proposal as the low-risk first step.

9.6 What to build first

PQC-readiness workflow + Sovereign Data Foundation + one QuantOS use case + single-hub unified metering, all on the bare-metal K8s reference cell. These are the lowest-technical-risk-to-revenue, highest-IP starting points.

9.7 What to avoid

Owning depreciating hardware on TQA's balance sheet; pricing on commodity GPU-hours; any quantum "advantage" claim; hosting a US vendor's PQC SaaS at the trust root; SAP BDC for sovereign clients; building a frontier model; single-counterparty concentration; shipping any AI answer without a citation object.

9.8 The one-paragraph thesis (for alignment)

TQA makes Latin America's compute sovereign, intelligent and secure without owning the steel: one control plane over two compute substrates, an open sovereign data foundation, client-owned AI (QuantOS) with graph-traceable answers, a real post-quantum security plane where the client holds the keys, and honest quantum access — sold from the urgent, low-risk security and data layers upward into differentiated AI and quantum. The defensible IP is the orchestration/billing federation, the QuantOS delivery and traceability pipeline, the crypto-agility/PQC workflow, and the sovereignty governance model. Everything else is best-of-breed open integration. The market is empty, the deadlines are legal, the talent is cheap, and the first mover defines the category.

End of Internal Technical Manual v1.0. Companion deliverables: (2) Client Offering deck (PPTX); (3) TQA Control Center demo (HTML + Figma spec); (4) Reference architecture in Figma. Research citations are maintained in the research appendix and inline in Part 6.