Customer Case Study | Brad Gauthier

Governed Autonomous AI on the HP ZGX Nano AI Station

Air-Gapped Organizational Intelligence with Post-Quantum Cryptographic Audit

The Challenge

Organizations in regulated and high-trust environments face a structural barrier to AI adoption that goes beyond model selection: sensitive internal data (including policy documents, financial records, operational logs, and institutional knowledge) cannot leave controlled environments. Cloud-based AI services, regardless of their capability, introduce external dependencies that conflict with data sovereignty requirements, create unpredictable cost exposure, and produce governance gaps where AI actions occur outside organizational oversight.

For industries like defense, federal government, healthcare, and financial services, these are not theoretical concerns. Compliance frameworks such as FedRAMP, ITAR, HIPAA, and CMMC impose strict controls on where data is processed and who has access. When AI execution happens in third-party infrastructure, meeting these requirements adds significant overhead, and in some cases makes deployment impractical entirely.

Most existing AI platforms treat security as a configuration layer applied after deployment rather than an architectural constraint enforced at execution time.

A separate challenge emerges with autonomous AI agents specifically. Unlike static model inference, agents plan and execute multi-step workflows that may generate and run code, query internal systems, and produce artifacts. This creates a larger attack surface than traditional AI deployments. If an agent produces malicious or faulty code during autonomous operation, the consequences extend beyond incorrect output to potential system compromise. Most existing AI platforms treat security as a configuration layer applied after deployment rather than an architectural constraint enforced at execution time.

"The rise of open-source local AI tools has demonstrated that capable AI can run entirely on user owned hardware. This has validated the core premise: people want AI that lives on infrastructure they own and control."

Bradley Gauthier,
Founder & Chief Architect,
Quantum Pipes (Sitecast.com / RecSystems.com / Others)

The rise of open-source local AI tools has demonstrated that capable AI can run entirely on user owned hardware. This has validated the core premise: people want AI that lives on infrastructure they own and control.

But capability without governance is a liability in regulated environments. These tools are designed for personal productivity, not enterprise accountability. There is no audit trail, no approval workflow, no cryptographic proof of what the AI did or why. For a developer experimenting at home, this is fine. For an organization handling classified, financial, or patient data, it is disqualifying.

The missing layer is not intelligence. It is trust infrastructure: the ability to prove what an AI system did, why it did it, and who authorized it, with cryptographic certainty.

The Solution

Quantum Pipes is an autonomous AI platform designed to run entirely on-premises, co-locating AI execution with organizational data and systems rather than routing either through external services. Deployed on the HP ZGX Nano AI Station, the platform provides a self-contained environment for AI agent workflows where governance and security are enforced at the architecture level.

The HP ZGX Nano, powered by the NVIDIA GB10 Grace Blackwell Superchip[1] with 128GB unified LPDDR5x memory, provides the compute foundation. Models run directly on the hardware rather than being accessed through external APIs, eliminating the data-transit concerns that complicate regulated deployments.[2]

If the first wave of local AI was personal, fast, and capable, then Quantum Pipes on the HP ZGX Nano represents the next: enterprise-grade, governed, and provably trustworthy. Same fundamental premise, AI that lives on your hardware, but built for the organizations that cannot afford to get trust wrong. The platform's two flagship capabilities make this concrete:

The Vault is a local intelligence layer that converts organizational documents, data, and institutional knowledge into queryable context. Documents go in; cited, grounded answers come out. Everything stays on the device.

The Capsule system is an open-source, cryptographically sealed audit architecture that captures the complete decision record for every AI action, including the AI's reasoning *before* execution occurs. This is not logging. It is contemporaneous evidence, sealed with both classical and postquantum cryptographic signatures, and linked in a tamper-evident hash chain.

Beneath these, the platform provides serverless execution environments, built on over a decade of production infrastructure technology, that enable Quantum Pipes to take the intelligence and insights from the Vault and Capsule system and render them into rich, interactive interfaces that run entirely locally and air-gapped. This is what transforms governed AI from a backend process into an operational capability that people actually use.

Quantum Pipes is an autonomous AI platform designed to run entirely on-premises, co-locating AI execution with organizational data and systems rather than routing either through external services.

Quantum Pipes includes a complete document intelligence pipeline for converting organizational data into queryable knowledge. The system processes

50+

document formats

including PDF, DOCX, XLSX, PPTX, HTML, and images with OCR support, converting them to a normalized representation.

Network Architecture: Air-Gapped by Design

The deployment architecture enforces network-level data sovereignty through physical isolation rather than policy-based access controls. The HP ZGX Nano runs Quantum Pipes Core and connects to the local network via a 10 Gbps switch. Client devices (laptops, desktops, mobile) access the system over this internal network. A firewall sits between the GPU server and the external internet with egress fully blocked. No data leaves the network boundary because there is no outbound path.

NAS storage on the local network holds models and organizational data, making them available to Quantum Pipes without any external transfer. An organization loads open-weight models and internal documents onto the HP ZGX Nano, and the entire AI stack operates within the existing security perimeter.

Air-gap integrity is verifiable: Quantum Pipes includes a built-in verification command (`make verify-airgap`) that runs an automated 5-minute packet capture on all network interfaces, confirming zero outbound traffic. This provides auditable evidence that the deployment is genuinely air-gapped, not merely configured to be.

Document Ingestion

Quantum Pipes includes a complete document intelligence pipeline for converting organizational data into queryable knowledge. The system processes 50+ document formats, including PDF, DOCX, XLSX, PPTX, HTML, and images with OCR support, converting them to a normalized representation. Documents are then semantically chunked and embedded locally into high dimensional vectors for retrieval.

At query time, Quantum Pipes executes hybrid retrieval combining semantic similarity search with traditional keyword matching, along with a relevancy scoring system that learns from user trust signals. Responses include inline citations linking back to source material, providing verifiable grounding for every answer.

The entire ingestion and retrieval pipeline runs locally. No document content, embeddings, or queries leave the device at any stage.

Performance Characteristics

Metric	Value
Cold start (boot + load + execute)	~150ms
Warm execution (pre-warmed VM pools)	<50ms
Memory overhead per isolated execution	<5MB
MicroVM creation rate	150+ per second per host

This delivers VM-level security isolation with container-like speed. Each execution is as isolated as a separate physical machine, but boots in milliseconds rather than minutes.

Execution Security: Two Concentric Cages

Autonomous AI agents create a security challenge that static inference does not: agents generate and execute code as part of their workflow. Quantum Pipes addresses this with a dual-layer isolation model that contains all AI-generated code execution within two concentric security boundaries. If one layer is compromised, the other contains the threat.

Cage 1: Hardware Isolation

The outer cage provides hardware-level containment for the execution environment. Each code execution receives its own dedicated guest kernel, running inside a hardware-backed microVM isolated from the host operating system via hardware virtualization. This is the same class of isolation technology trusted by major cloud providers for serverless workloads.

- **Dedicated Guest Kernel:** Each execution gets its own kernel with no shared-kernel vulnerabilities.
- **Hardware-Enforced Memory:** Memory boundaries are enforced at the hardware level, preventing code inside the cage from accessing memory outside its allocated region.
- **Network Disabled by Default:** MicroVMs have no network interface. Communication with host services occurs through a hypervisor-level channel that bypasses the network stack entirely.
- **Read-Only Root Filesystem:** Persistent modifications to the execution environment between runs are prevented.

Cage 2: Language Sandbox

The inner cage constrains what AI-generated code can do at the language runtime level, even within the hardware-isolated environment.

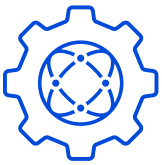
- **No Direct System Calls:** Generated code cannot interact with the operating system kernel directly.
- **Capability-Based Permissions:** The sandbox exposes only explicitly granted operations. All external access (reading from the Vault, writing logs, querying databases) requires a specific capability grant from a strict allowlist.
- **Memory Bounds Checking:** The runtime prevents buffer overflows and out-of-bounds access.
- **Execution Limits Enforced:** Timeouts and resource caps prevent runaway processes from consuming resources.

Every capability check, whether granted or denied, is logged with full context for audit purposes.

Defense in Depth

The result is defense in depth: even if AI-generated code contains a vulnerability or behaves unexpectedly, the blast radius is contained to an isolated execution environment with no network access, no persistent storage, and no path to the host system.

Threat	Cage 1 (Hardware)	Cage 2 (Sandbox)
Kernel exploit	Isolated guest kernel	N/A
Memory corruption	Hardware-enforced isolation	Bounds checking
Syscall abuse	Minimal syscall allowist	No direct syscalls
File system access	Read-only rootfs	No FS access
Network exfilrtation	No network interface	No network access
Resource exhaustion	Resource limits	Execution limits



At the center of Quantum Pipes is the Capsule system, an open-source, cryptographically sealed audit architecture that makes every AI action provably accountable.

The Capsule System: Cryptographic Trust for AI

At the center of Quantum Pipes is the Capsule system, an open-source, cryptographically sealed audit architecture that makes every AI action provably accountable.

A Capsule is not a log entry. It is a six-section immutable record that captures the complete decision context of an AI operation:

1. **Trigger:** What initiated the action: user request, scheduled event, or autonomous decision. Includes microsecond-precision timestamp and correlation ID.
2. **Context:** System state at the time: agent identity, session, environment configuration, model version.
3. **Reasoning:** How the AI decided: analysis, options considered with pros/cons, the selected option with rationale, confidence score, and *mandatory rejection reasons for every non-selected option*. Captured *before* execution, not reconstructed after the fact.
4. **Authority:** Who authorized: autonomous, human-approved, or policy-approved, with a cryptographically signed authorization chain.
5. **Execution:** What happened: tool calls with per-call duration, resource utilization (tokens, memory, CPU).
6. **Outcome:** What resulted: the output, a PII-free summary, side effects, error details if any.

Most AI audit systems log inputs and outputs. Capsules capture the reasoning between them, and they capture it before the action occurs, not after. This means the record of why an AI made a decision exists independently of whether the decision succeeded, failed, or was interrupted.

Pre-Execution Reasoning Capture

Most AI audit systems log inputs and outputs. Capsules capture the *reasoning between them*, and they capture it before the action occurs, not after. This means the record of why an AI made a decision exists independently of whether the decision succeeded, failed, or was interrupted.

This has two implications that matter for regulated environments:

1. **Legal Admissibility:** Contemporaneous evidence of decision-making is more defensible than post-hoc reconstruction. The reasoning section is captured before execution begins.
2. **Computational Efficiency:** Pre-execution capture reduces overhead by approximately 50% compared to post-hoc explainability approaches, because the reasoning is recorded as it naturally occurs in the agent's cognitive loop rather than being regenerated from outputs.

Dual-Signature Quantum-Ready Sealing

Each Capsule is sealed with three cryptographic layers:

1. **SHA3-256 Hash:** (FIPS 202) of canonical JSON (RFC 8785), providing deterministic, reproducible content addressing.
2. **Ed25519 Signature:** FIPS 186-5), classical cryptographic signature, 128-bit security.
3. **ML-DSA-65 Signature:** (FIPS 204), post-quantum cryptographic signature, NIST Level 3 security

Hash Chain Integrity

Each Capsule includes the SHA3-256 hash of the previous Capsule, creating an append-only chain where tampering with any record invalidates all subsequent records. Chain integrity can be verified on-demand through the web interface, providing auditable proof that the historical record has not been altered.

Mandatory Capture Architecture

AI operations in Quantum Pipes cannot complete without creating a sealed Capsule. This is enforced at the architecture level, not through policy or configuration, but through runtime wrappers and type-system constraints that make record creation an atomic part of every AI operation. There is no code path that bypasses the audit trail.

Deployment: From Unboxing to Operational in 30 Minutes

In observed deployments, Quantum Pipes was installed and operational on the HP ZGX Nano in approximately 30 minutes. The following walkthrough describes the deployment sequence.

Step 1: Hardware Provisioning (~5 min)
Connect HP ZGX Nano to power and internal network. The system runs NVIDIA DGX™ OS out of the box.

Step 2: Offline Package Creation (~10 min)
On a staging machine with internet access, run `make offline-package ARCH=arm64` to bundle all container images, dependencies, and platform code into a single transferable package with SHA-256 checksums. Run `make offline-models` to bundle AI models for local inference. Transfer to the ZGX Nano via USB drive or SCP over isolated LAN.

Step 3: Installation (~5 min)
On the ZGX Nano, run `make offline-install` followed by make go. All services start automatically. No cloud accounts, API keys, or external services are required.

Step 4: Document Ingestion (~5 min)
Access the web interface at `https://hub.local`. The Vault page provides a drag-and-drop file manager for document uploads. Uploaded documents are automatically processed (converted, chunked, embedded, and indexed) with progress tracking in the UI. For bulk ingestion, existing document directories can be mounted as volumes.

Step 5: Operational Verification (~5 min)
Upload documents to the Vault, switch to Chat, and ask a question. The system retrieves relevant context, generates a cited response using the local LLM, and seals the interaction as a Capsule. Run `make verify-airgap` to produce auditable evidence of air-gapped operation.

No cloud accounts, API keys, or external services are required at any step. The entire deployment occurs within the organization's existing network boundary.

In observed deployments, Quantum Pipes was installed and operational on the HP ZGX Nano in approximately

30

minutes.

How It Works: End-to-End Query with Governance

The following sequence describes what happens when a user submits a natural-language query to Quantum Pipes running on the HP ZGX Nano, from input through governed execution to auditable output.

Example scenario: A compliance officer asks, *"What are our obligations under HIPAA for retaining patient imaging data?"*

1. **User submits query** through the web interface at https://hub.local. The Chat interface supports streaming responses, image attachments, and model selection.
2. **Local Retrieval:** The query is embedded locally and matched against organizational documents in the Vault using hybrid search, combining semantic similarity with keyword matching. Relevant document passages are retrieved with source references.
3. **Local Inference:** The local LLM (e.g., Llama 3.3 70B running entirely in the ZGX Nano's 128GB unified memory) generates a response grounded in the retrieved context.
4. **Capsule Creation:** Before the response is returned, a Capsule is created capturing the full decision record (trigger, context, pre-execution reasoning, authority, execution details, and outcome) as described above.
5. **Cryptographic Sealing:** The Capsule is sealed with SHA3-256 hash, Ed25519 signature, and ML-DSA-65 post-quantum signature, then hash-chain-linked to the previous Capsule.
6. **Response Delivery:** The response streams to the user with inline source citations. Each citation links to the specific document and passage in the Vault.

If the query triggers an agent workflow requiring multi-step execution (e.g., *"Draft a summary comparing our retention policies across all departments"*), each step runs inside the dual-cage isolation boundary, and human approval checkpoints fire at defined control points based on the organization's configured autonomy mode. Every step, including approvals and denials, is sealed as a Capsule.

Governance and Auditability

Audit Dashboard

The web dashboard provides:

- Filtering by event type, date range, user, and agent
- On-demand chain integrity verification across all Capsule records
- Export to JSON and CSV formats
- A dedicated Auditor Portal providing read-only access for external auditors

Privacy-Aware Retention

Capsule records support three-tier differential retention:

- **Tier 1: Permanent.** Header, seal, summary, reference hashes. Retained indefinitely. Accessible to all authorized users.
- **Tier 2: Operational.** Full reasoning, authority chain, tool parameters. Configurable retention (default 90 days). Accessible to operators.
- **Tier 3: Full Context.** Full prompts, full results, PII. Configurable retention (default 7 days).

This architecture supports GDPR Article 17 (Right to Erasure): sensitive data can be deleted while preserving hash chain integrity through reference hashes, maintaining the audit trail's cryptographic validity even after personal data is removed.

Compliance Frameworks

Quantum Pipes includes an enterprise compliance engine that helps organizations track and demonstrate their compliance posture across ten frameworks:

SOC 2 | HIPAA | GDPR | PCI DSS | FedRAMP | ISO 27001 | NIST CSF | FINRA | ITAR | CMMC

The compliance dashboard maps Capsule audit records to framework-specific controls, providing scorecards, control mapping, gap analysis, compliance calendars, and evidence export (JSON, PDF, ZIP with signed manifests). Organizations use these tools to document and manage their own compliance obligations. The open-source Capsule system provides the underlying evidence foundation; the compliance engine structures that evidence for auditors and regulators.

Quantum Pipes includes an enterprise compliance engine that helps organizations track and demonstrate their compliance posture across

10

frameworks.

Human-in-the-Loop Controls

Quantum Pipes supports configurable human-in-the-loop approval points within agent workflows. Organizations define control points where agent actions require human review before proceeding.

Autonomy Modes

Mode	Behavior
Conservative	Most tool executions require explicit human review
Balanced	Routine operations proceed autonomously; high-risk or novel actions trigger approval
Progressive	Only operations exceeding defined risk thresholds trigger approval

Approval Experience

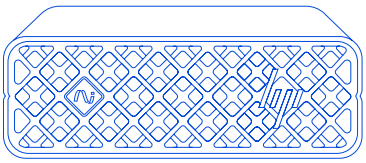
When a checkpoint fires, the approver sees a visual approval card displaying the requesting agent, the specific tool and parameters being requested, a risk-level indicator, and the agent's reasoning. The approver can approve or deny with a mandatory reason. All decisions, approvals and denials, are sealed as Capsule records.

Emergency Controls

A Guardian process runs as a dedicated service, providing a kill switch that cannot be overridden by any other process. It supports graceful shutdown (allowing final Capsule records) and immediate termination. Every kill event is itself recorded as a sealed Capsule.

Observed Results

Install to operational	~30 minutes
External connectivity required	None
Cloud services required	None
Compliance frameworks with built-in tooling	10 (SOC 2, HIPAA, GDPR, PCI DSS, FedRAMP, ISO 27001, NIST CSF, FINRA, ITAR, CMMC)
Document formats supported	50+ (PDF, DOCX, XLSX, PPTX, HTML, images with OCR)
MicroVM cold start	~150ms
MicroVM warm execution	<50ms
Air-gap verification	Automated packet capture, auditable evidence



Why HP ZGX Nano

The HP ZGX Nano AI Station provides specific hardware characteristics that align with the requirements of on-premises autonomous AI workloads:

- **GPU:** NVIDIA GB10 Grace Blackwell Superchip[1]. Capable of local model inference without external API calls.
- **Memory:** 128GB unified LPDDR5x. Supports large language models entirely in local memory.
- **AI Performance:** 1000 TOPS. Sufficient throughput for concurrent agent workloads.
- **Operating System:** NVIDIA DGX™ OS (Ubuntu 24.04). Pre-configured AI development environment.
- **Form Factor:** 150mm L × 150mm W × 51mm H[3] desktop. Deployable in offices, SCIFs, field locations without rack infrastructure.
- **Network:** Capable of operating fully offline. Supports air-gapped deployment architecture.[2]
- **Replication:** Standardized configuration. Scale by deploying additional units with identical setup.

The compact form factor is particularly relevant for deployments where traditional rack-mounted server infrastructure is impractical: secure facilities, branch offices, or field-deployed operations. Organizations scale AI capacity by deploying additional ZGX Nano units rather than expanding centralized infrastructure.

Open Source

Quantum Pipes Core, the foundational trust layer including the Capsule audit system, cryptographic sealing, cognitive loop, and kill switch, is open source. The Capsule system's six-section record structure, pre-execution reasoning capture, and dual-signature quantum-ready sealing are freely available for inspection, extension, and integration.

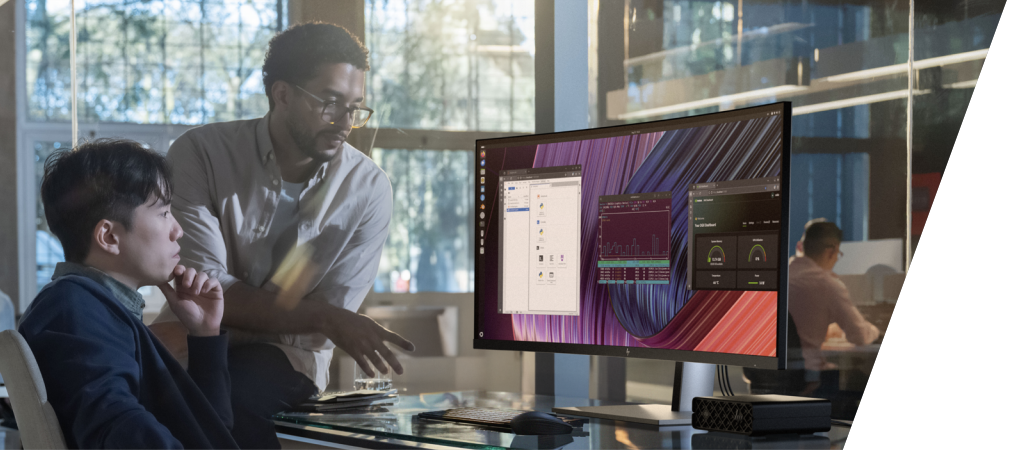
Open source means you can audit the code, verify there are no hidden connections, examine the security implementation, and build on it. For regulated environments, this is not a feature. It is a requirement for trust.

The full Quantum Pipes platform, including the Vault, governance dashboards, compliance frameworks, deployment tooling, and dual-cage isolation infrastructure, is available commercially, with additional capabilities released as intellectual property protections are established.

Get Started

Quantum Pipes Core (Open Source):
https://www.quantumpipes.com/platform/

Quantum Pipes Platform:
https://www.quantumpipes.com



About the HP ZGX Nano AI Station

The HP ZGX Nano G1n AI Station is a compact desktop system designed for AI development and evaluative inference workloads. Built on the NVIDIA GB10 Grace Blackwell Superchip[1] with 128GB unified memory, it runs NVIDIA DGX™ OS and, when paired with the HP ZGX Toolkit[4], provides a ready-to-use environment for PyTorch, llama.cpp, vLLM, and other AI frameworks. The system delivers powerful AI compute in a 150mm L × 150mm W × 51mm H[3] form factor, enabling researchers and developers to iterate on complex models without cloud dependencies or infrastructure complexity.

The HP ZGX Nano is a variant of the NVIDIA DGX Spark™, developed in partnership between HP and NVIDIA.

HP ZGX Nano
https://www.hp.com/us-en/workstations/zgx-nano-ai-station.html

HP ZGX Toolkit
https://www.hp.com/us-en/workstations/zgx-onboard.html

NVIDIA DGX Spark™
https://www.nvidia.com/en-us/products/workstations/dgx-spark/

HP provided the customer with an HP ZGX Nano free of charge for evaluation, and their experience is shared here.

1 Multi-core is designed to improve performance of certain software products. Not all customers or software applications will necessarily benefit from use of this technology. Performance and clock frequency will vary depending on application workload and your hardware and software configurations.

2 Models can be executed locally on the HP ZGX Nano without reliance on external APIs or cloud-hosted services. HP offers a version of the HP ZGX Nano with Wi-Fi and Bluetooth components physically removed, eliminating wireless data transmission capabilities. Network connectivity, if present, is limited to customer-controlled wired interfaces.

3 Height excludes feet.

4 The HP ZGX Toolkit is provided free of charge. Use requires a client device running Windows 11 or Ubuntu 24.04 (or later) with Visual Studio Code installed and host device ZGX Nano. The client device must be x86-based, but otherwise there are no restrictions regarding hardware specifications or device manufacturer. Availability may vary by region and is subject to applicable local laws, regulations, and restrictions.

c09271400 May 2026.