



NUTANIX



Tactical Edge Supercomputing – It's Real, It's Today

Deploying AI workloads at the tactical edge is often limited by severe physical constraints, harsh environments, and unreliable cloud connectivity. Further, AI is evolving quickly, and hardware is having trouble keeping up when we look at the transition from generative AI to inferencing and the use of agents (Agentic AI). Our warfighters and their operations teams need datacenter-class compute power that can ensure operational continuity when cloud connectivity is intermittent and can operate securely in a completely disconnected or air-gapped environment.

This white paper details a validated, full-stack design combining the [Gigalo Gryf Carry-on AI Supercomputer](#), [Nutanix Kubernetes Platform \(NKP\)](#), and [Nutanix Enterprise AI \(NAI\)](#). This integrated solution delivers an agile, secure-by-default, and completely autonomous Agentic AI platform engineered to modernize our military's tactical and operational outcomes in the thick of operations.

01

Solution Component Overview

Nutanix Enterprise AI (NAI) - Inference, Large Language Models, AI Gateway

Nutanix Kubernetes Platform (NKP) - Fleet Management, GitOps, CNCF-Native Container PaaS

Gigalo Gryf AI Memory Fabric (FabreX) – The enabling technology for hardware component disaggregation and composability with dynamic resource allocation

Gigalo Gryf: The Hardware Foundation

The Gigalo Gryf is the industry's first suitcase-sized, field-ruggedized AI supercomputer. Built to fit into standard commercial airline overhead bins, it features a modular architecture that completely disaggregates hardware resources:

Composability

Accommodates 6 interchangeable sleds. Sleds include 64-core CPUs, high-end accelerators (GPUs/ASICs), and up to 5 PB of storage.

FabreX AI Fabric

Interconnects components at PCIe/CXL memory speeds, allowing resources to be dynamically partitioned via software.

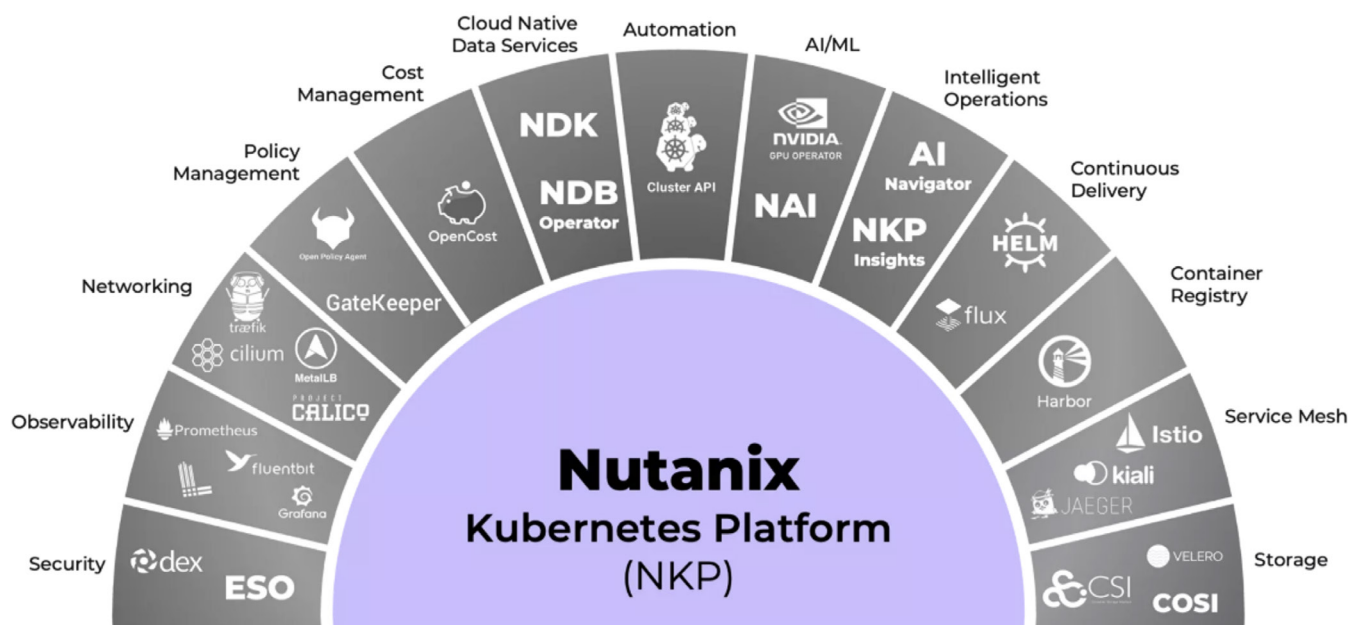
Extreme Resilience

Engineered with a 2,500W power capacity and isolated ventilation to handle power-hungry workloads in remote environments without cloud access.

Nutanix Kubernetes Platform (NKP): The Orchestration Layer

NKP acts as a production-ready, cloud-native container layer. It simplifies cluster operations through centralized GitOps and management tools:

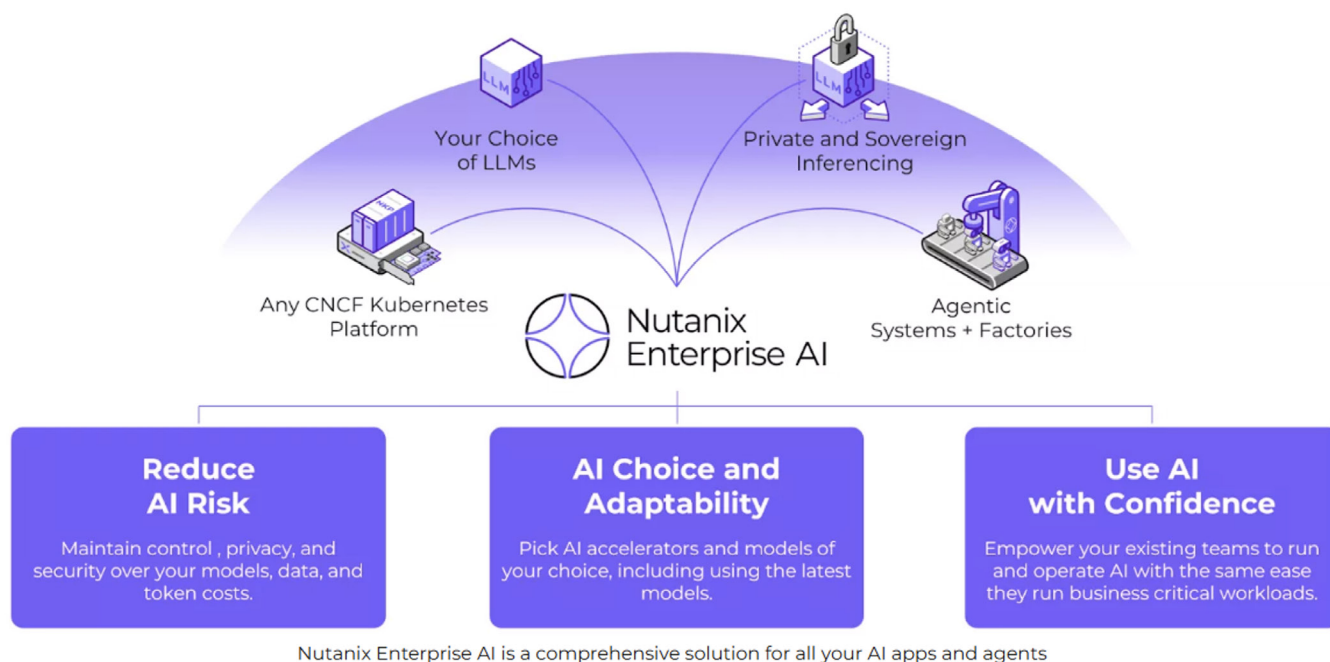
- **Security & Compliance:** Hardened out-of-the-box with FIPS-compliant cryptographic modules, Cilium for zero-trust networking, and pre-configured CIS benchmarks.
- **Air-Gapped Deployment:** Native capabilities to push container images and provision local cluster registries in completely disconnected environments.



Nutanix Enterprise AI (NAI): The Intelligence Layer

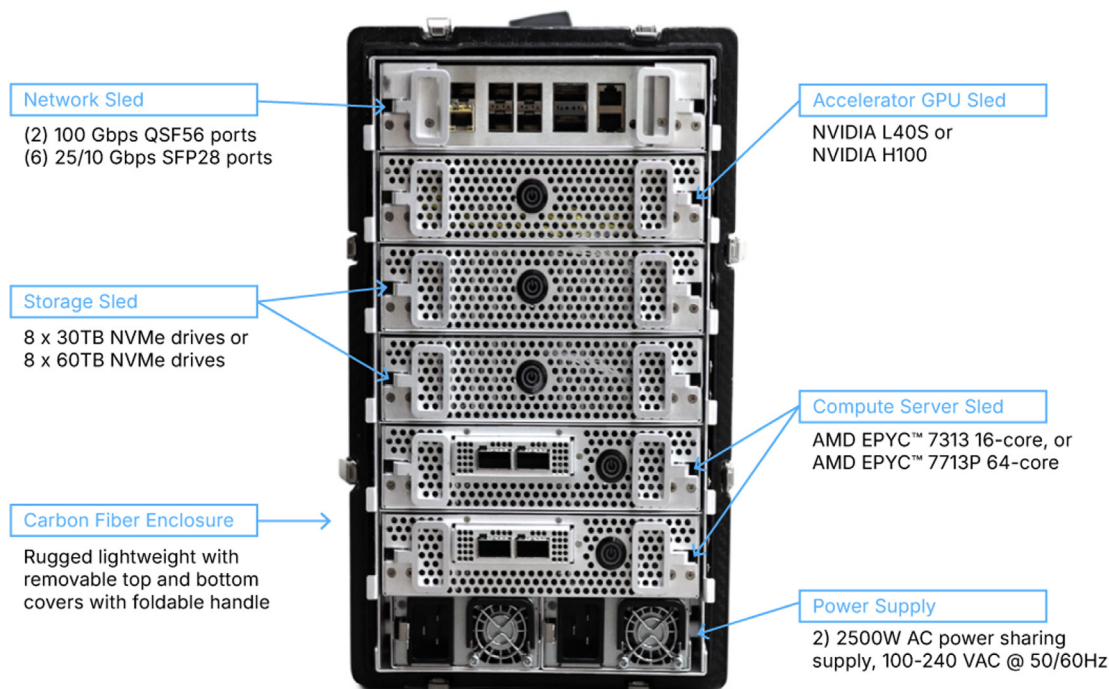
NAI provides a developer-friendly, cloud-native runtime environment for large language models (LLMs) and agentic applications. Integrated directly into the NKP dashboard catalog, NAI simplifies model deployment:

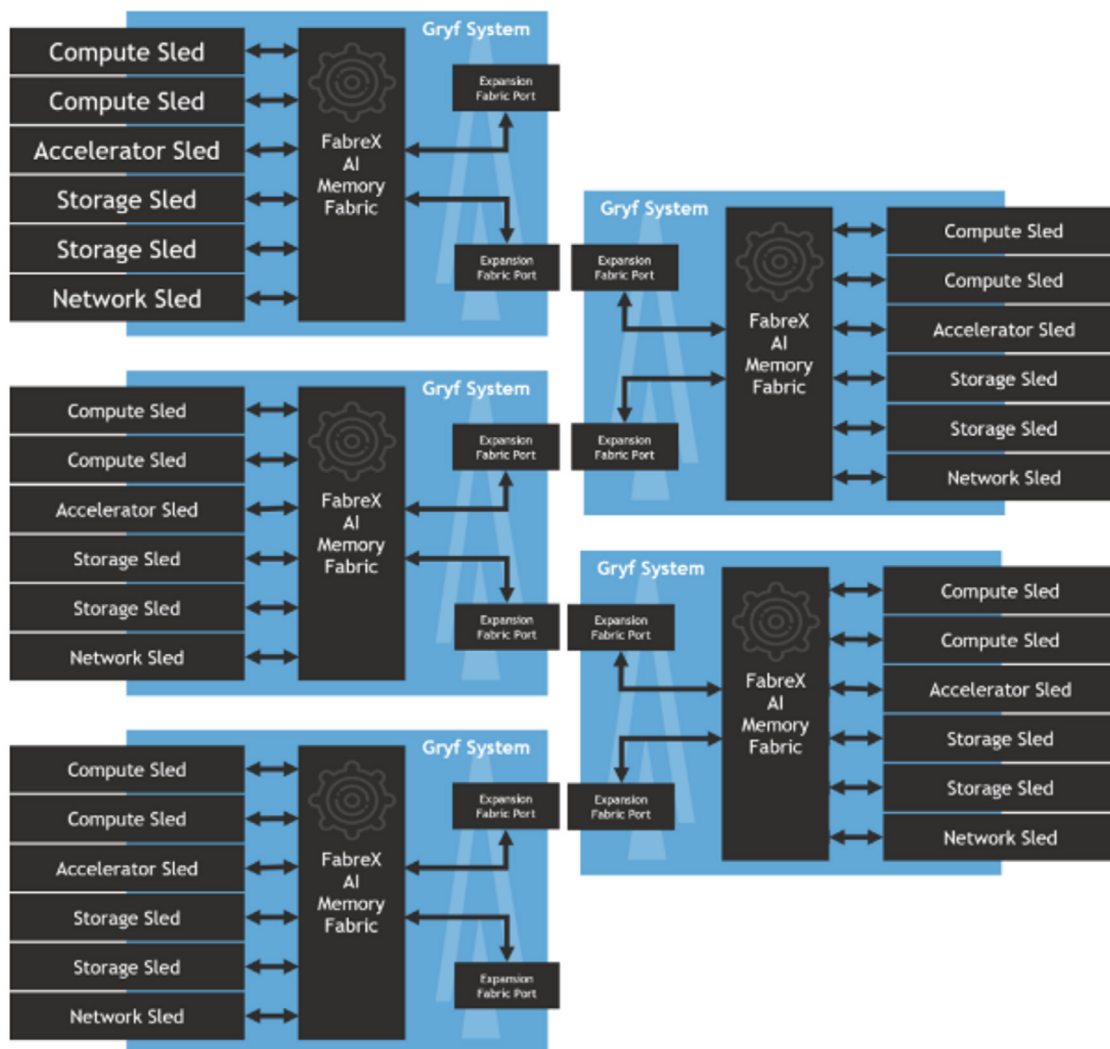
- **AI Gateway & Model-as-a-Service:** Exposes standard APIs for local LLM engines (such as NVIDIA NIM and Hugging Face). "Nutanix agent gateway provides centralized control layer that governs, routes, and secures communication between AI agents, LLMs, and enterprise tools. Organizations gain unified governance, observability, and cost control across the enterprise, from data center to edge."
- **Data Sovereignty:** Eliminates data leakage risks by processing all sensitive information within the localized Gryf hardware boundary.



02 Integrated Architecture & Resource Composability

The key advantage of this design is the integration between GigaIO's hardware-level composability and Nutanix's software-defined automation.





Dynamic Resource Allocation: Through [Gigalo FabreX](#), storage or GPUs are shared or reallocated across compute sleds or virtual environments dynamically, preventing resource lock-in.

Kubernetes-Driven Scheduling: NKP detects the underlying disaggregated compute and GPU topology mapped by the fabric. Pods requesting hardware acceleration are automatically scheduled onto nodes with optimal, low-latency access to the hardware fabric.

Optimized Token Economics: Because hardware resources can be precisely provisioned and models run completely locally, enterprises avoid the variable costs and unexpected bills associated with cloud-hosted AI providers.

03 Deployment Guide: Establishing the Air-Gapped AI Stack

Follow this sequential procedure to initialize the architecture in a disconnected, tactical edge scenario.

Step 1 Initialize the Hardened NKP Edge Cluster

1. **Prepare the Registry:** Push the NKP container images and dependencies to a localized, pre-provisioned registry or USB storage media within the perimeter.
2. **Bootstrap the Cluster:** Use the NKP CLI to build a secure bootstrap workspace using Ubuntu-hardened images running upstream Kubernetes.
3. **Apply Licenses:** Elevate the cluster to an NKP Pro or Ultimate tier to unlock advanced multi-cluster management and integrated data services.

Step 2 Provision Local Storage for AI Models

1. **Deploy Nutanix CSI:** Install the Nutanix CSI Driver directly to direct containerized requests to the Gryf's local high-speed NVMe storage sleds.
2. **Configure Nutanix Files:** Create a localized file server with sufficient volume allocation to cache heavy LLM weight structures.
3. **Export NFS v4 Share:** Configure a Network File System (NFS v4) export path to allow manual model importing and vector database configurations.

Step 3 Spin Up Nutanix Enterprise AI (NAI)

1. **Access the Catalog:** Open the NKP Central UI and navigate to the integrated applications catalog.
2. **Deploy NAI:** With a single click, activate the NAI application. The underlying platform automatically starts dependencies like KServe and registers the NAI pods within the NAI-system namespace.
3. **Expose the Gateway:** Configure local load balancing using MetalLB. Route your internal Fully Qualified Domain Name (FQDN) to the NAI API Gateway to begin the tactical edge solution.

Conclusion

The combination of [GigalO Gryf](#), [Nutanix NKP](#), and [Nutanix NAI](#) solves the last mile of enterprise generative AI at the tactical edge. By packaging datacenter-class hardware into an airline-friendly, composable suitcase, organizations can deploy secure, cloud-native Agentic AI applications anywhere on Earth.

About Flywheel Data

Flywheel Data provides elite solution design, system integration, software development, and product resale for data-driven-organizations.

Based on our experiences with the US Government and top commercial companies; Flywheel Data recognizes that data and people are at the center of a successful, data-driven organization.

Our goal is to arm our clients with the right tools, platforms, and culture to accelerate data-driven insights.

