zadara

ENABLING MULTI-TENANT AI INFRASTRUCTURE

Deploying NVIDIA Reference Architectures on Zadara IaaS

EXECUTIVE SUMMARY

NVIDIA is leading the evolution of AI infrastructure with reference architectures that support a new class of scalable, software-defined AI factories. These environments demand performance, security, and agility across multi-tenant, sovereign, and private cloud footprints.

Zadara meets this need with a cloud-native Infrastructure-as-a-Service system that combines enterprise-grade orchestration, GPU-performance-aware scheduling, and native integration with NVIDIA networking and DPU technologies. This enables service providers, telcos, and enterprises to deploy NVIDIA-powered infrastructure for multi-tenant environments quickly, securely, and with full control over cost and compliance.

This white paper explores Zadara's central role in enabling NVIDIA AI Factory architectures, from the underlying infrastructure to containerized AI stack execution. It explains how Zadara supports secure GPU networking, VM-level allocation of 1, 2, 4, or 8 GPUs per VM as aligned with NVIDIA virtualization guidance, and DPU-accelerated network offload using the NVIDIA Cloud Partner reference architecture.

Additionally, this white paper highlights how Zadara delivers real-world deployments that simplify operational complexity and allow precise control of performance. Its infrastructure is designed to help organizations build and scale sovereign, multi-tenant AI clouds with confidence.

ZADARA ADVANTAGE AT A GLANCE

Zadara provides a unified managed cloud system that runs GPU and non-GPU infrastructure side by side. Customers can consolidate AI, compute, storage, and network workloads within a single operational model. This enables hybrid deployments, seamless workload migration, and a consistent user experience across environments.

Each tenant environment is instantiated with a secure, fully wired GPU and networking access, allowing data scientists, ML engineers, or DevOps deploying inference environments to begin building immediately with no manual intervention required by the cloud provider.

- Natively supports NVIDIA BlueField DPUs and Spectrum-X networking
- Fully aligned with NCP (NVIDIA Cloud Partners) reference architectures
- Global Sovereign AI Cloud footprint with full regulatory compliance
- Transparent GPU-to-GPU networking with Zadara GPU-Net
- VM-based GPU node allocation (1, 2, 4, or 8 GPUs per VM) with GPU and SuperNIC passthrough
- DPU-offloaded network stack with SR-IOV interface and enabling DOCA-accelerated services (firewall, telemetry, microsegmentation)

Zadara delivers these capabilities as a fully managed service, providing 24/7 monitoring, infrastructure lifecycle management, and second-day operational support. This allows cloud providers, telcos, and enterprise IT teams to offer AI services with confidence, while relying on Zadara to ensure performance, availability, and continuous compliance.

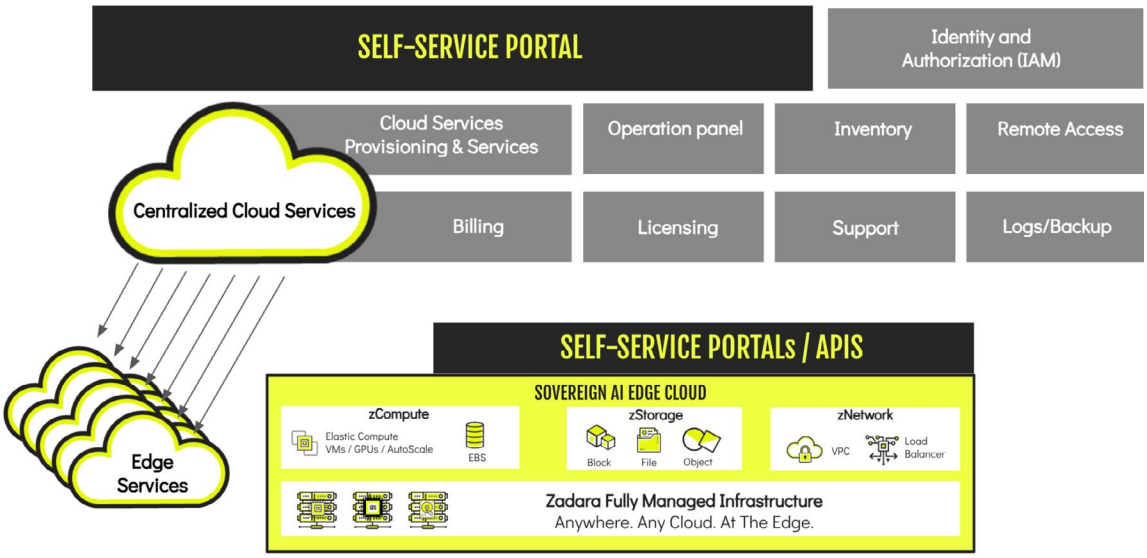

Figure 1: Zadara Cloud Structure Distributed, self-managed edge locations with optional centralized control

INTRODUCTION:

ENABLING THE SOVEREIGN AI EDGE CLOUD

As AI workloads scale, infrastructure providers must deliver secure, high-performance environments that support multitenancy, dynamic provisioning, and hardware-level isolation. These capabilities are no longer exclusive to hyperscalers, they are essential for regional clouds, managed services, telcos, and private enterprise systems.

Zadara was built to meet this demand from the ground up. Zadara's cloud-native Infrastructure-as-a-Service system supports GPU and non-GPU workloads, integrates with NVIDIA reference architectures, and exposes tightly controlled orchestration, policy, and network segmentation per tenant. From GPU allocation to network offload using DPUs, Zadara abstracts complexity while preserving performance and compliance.

The NVIDIA Cloud Partner reference architecture defines the principles required to deploy a scalable, sovereign, multi-tenant GPU infrastructure. This includes full-stack orchestration, GPU allocation policies, PCI passthrough, tenant-isolated networking, and observability. The goal is to simplify the deployment of AI-ready clouds that can support training, inference, and agentic workloads across varied environments; from edge to multi-region factories.

Zadara is uniquely aligned with these principles. The Zadara IaaS system provides the necessary orchestration layer, GPU and SuperNIC passthrough, DPU offload, and tenant policy enforcement to deliver this architecture as a product—not a framework. Rather than assembling discrete hardware and software components, Zadara offers a full-stack managed cloud infrastructure that includes everything needed to run the NVIDIA software stack securely and efficiently.

Zadara enables consistent multi-tenant GPU cloud deployments using GPU-Net fabric orchestration, dynamic provisioning, and integrated DPU-based policy enforcement. The architectural and operational details of Zadara's implementation are explored in the following sections.

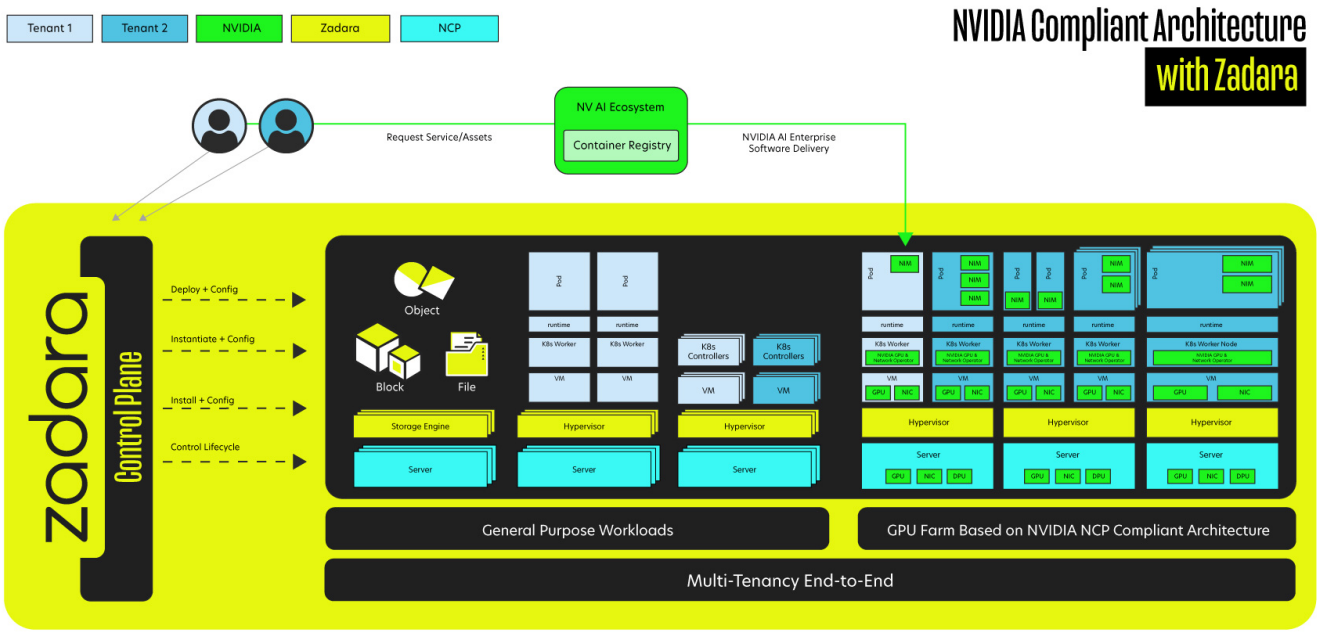

Figure 2: Zadara-NVIDIA Reference Architecture Alignment

KEY DESIGN PRINCIPLES FROM THE NVIDIA IAAS SOFTWARE REFERENCE ARCHITECTURE:

- Full-stack virtualization using a KVM-based hypervisor model for VM-based GPU tenancy
- Identity services and IAM
- GPU passthrough architecture with full physical GPU allocation (1, 2, 4, or 8 GPUs per VM)
- SR-IOV virtual NIC interfaces assigned directly to tenant VMs
- Tenant K8s clusters running on virtual machines with no direct access to physical hosts
- Containerized AI workloads deployed inside tenant-controlled Kubernetes clusters
- DOCA-enabled networking stack using BlueField DPUs for isolation and telemetry
- Dynamic overlay fabric using VRF, VXLAN, and EVPN segmentation per tenant
- Support for multiple orchestration models (e.g., Run:ai, NVCF, Lepton)
- NVIDIA AI software components (NIM, NeMo) run as microservices within tenant K8s clusters

This white paper details how Zadara implements these principles through its managed IaaS system, with native support for NVIDIA SuperNIC, BlueField DPU, and GPU-Net fabric policies orchestrated dynamically at provisioning time.

TECHNICAL FOUNDATIONS:

SPECTRUM-X GPU FABRIC AND GPU-NET

Each HGX GPU server in the architecture contains BlueField devices functioning exclusively as SuperNICs for GPU-to-GPU communication (east-west traffic), and one or two additional BlueField DPUs used for other networks and storage offload (common networking in NVIDIA's reference architecture terminology). This separation enables efficient traffic handling and role-specific hardware acceleration.

Zadara implements Spectrum-X as a dedicated GPU-to-GPU networking domain, optimized for AI training and inference pipelines. Zadara calls this GPU-Net, its dedicated east-west GPU fabric. Note that GPU-Net is not a virtual overlay fabric, but an orchestrated configuration of VRF, VXLAN, and Spectrum-X policy that is provisioned automatically when tenant VMs are deployed.

Each HGX GPU server includes dedicated NICs for east-west GPU communication. These NICS are either ConnectX-based NICs or BlueField card models NVIDIA refers to as SuperNICs. These NICs are installed directly in the GPU nodes and managed as part of the GPU fabric. When configured under Spectrum-X, they provide high-throughput connectivity between GPUs across servers. Zadara also uses additional BlueField DPUs per server for north-south (common networking) traffic and infrastructure offload.

SPECTRUM-X FABRIC DESIGN HIGHLIGHTS:

- Each GPU node includes multiple ConnectX-based NICs (SuperNICs) connected to separate Spectrum-X switches
- These connections are "rail groups" aligned: Parallel, non-blocking paths to separate what NVIDIA defines as leaf switches, ensuring deterministic latency and symmetric east-west traffic
- Rail groups enable line-rate communication between GPU nodes and improve fault isolation
- Multiple rail groups can be orchestrated into a larger scale-out unit while preserving performance isolation per GPU workload
- Spectrum policies orchestrate deterministic paths and full-bandwidth delivery across the rail groups, ensuring uniform GPU-to-GPU performance for model-parallel and inference workloads

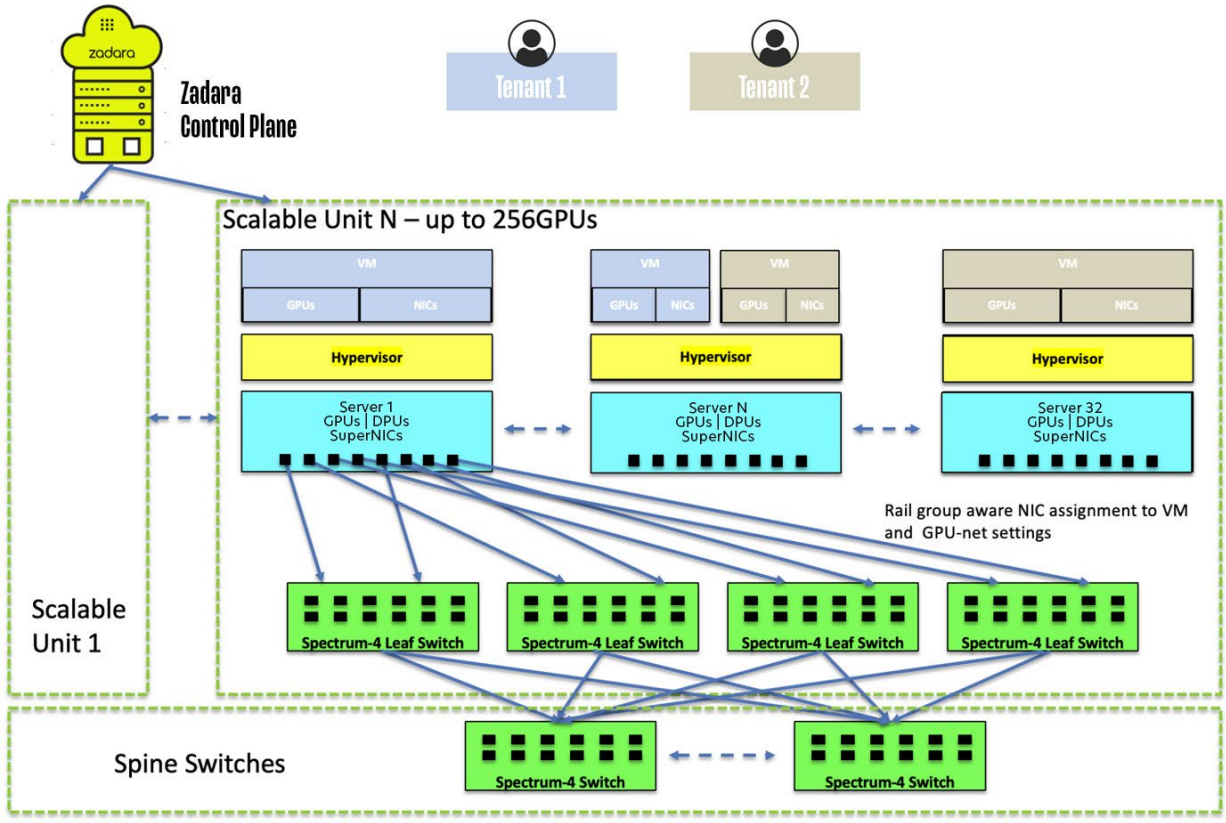

Figure 3: Spectrum X reference architecture implemented by Zadara

ZADARA'S GPU-NET ABSTRACTION PROVIDES:

- Dynamic, per-tenant L3 segments on top of VXLAN overlays over Ethernet fabrics
- Integrated EVPN control plane per tenant VRF, enabling distributed virtual routing and policy enforcement
- High-throughput interconnect that supports NVLink and Ethernet-based GPU communication, including NVIDIA's optimized transport with or without RDMA
- Transparent inter-node GPU communication within and across VMs with isolation by VRF and virtual network segment, all dynamically and automatically provisioned during VM creation. Each VPC is backed by a dedicated VRF and VXLAN overlay network, ensuring complete traffic isolation between tenants. This enables a seamless cloud experience, where any VM within a tenant's VPC is automatically and securely connected to the GPU fabric without user intervention

OPTIMIZED VM CONFIGURATION WITH GPU AND SUPERNIC PASSTHROUGH

EACH VM IN THE ZADARA AI CLOUD IS CONFIGURED WITH:

- Full PCI passthrough of NVIDIA GPUs, enabling full bare-metal performance with no hypervisor scheduling interference
- Full PCI passthrough of SuperNICs, bypassing virtual switches or emulated devices entirely
- Physical GPU allocation (1, 2, 4, or 8 GPUs per VM) in alignment with NVIDIA virtualization guidelines
- NUMA alignment and CPU pinning, ensuring local memory access paths to the assigned GPUs
- Awareness of underlying PCI topology, so GPU and NIC passthrough devices are mapped cleanly to avoid cross-CPU bus traffic or I/O bottlenecks
- VM-level awareness of NVLink switch domains, ensuring multi-GPU workloads can fully leverage NVLink bandwidth within the same node when possible

These settings result in performance that approaches bare-metal operation while preserving the isolation, policy enforcement, and lifecycle flexibility of virtualized infrastructure. These properties are kept even when a GPU node is split between tenants. Zadara orchestrates this configuration automatically for each tenant deployment, eliminating manual tuning and maximizing workload efficiency.

When a VM is allocated multiple GPUs, Zadara makes use of NVIDIA Fabric Manager to ensure those GPUs are initialized with full NVLink bandwidth and topology awareness. This ensures that GPU groupings reflect optimal placement within the node's NVLink switch fabric. As a result, multi-GPU workloads benefit from low-latency peer access and NVLink interconnects without requiring explicit tenant configuration or tuning.

BLUEFIELD DPU INTEGRATION

Network Offload and Security Isolation BlueField DPUs serve as programmable Data Processing Units (DPUs) that offload key infrastructure functions from the host:

- SR-IOV virtual interfaces assigned directly to tenant VMs
- Virtual Switch, traffic filtering (Security-groups), and VXLAN acceleration performed in the NIC, reducing host CPU load and improving throughput

Each VM is assigned with one or more virtual functions (VF) for direct access to SuperNIC (not via hypervisor) with isolated L2 segments managed via BlueField. This architecture enables fully isolated, high-throughput AI tenant environments with DPU-enforced policy and metrics collection.

DOCA-enabled services (firewall, telemetry, microsegmentation) offload to the DPU will soon follow.

AI SYSTEM ENABLEMENT AND NVIDIA SOFTWARE STACK INTEGRATION

Zadara provides the orchestration, identity, and tenant-level policy control that could support the NVIDIA AI Enterprise software stack. This includes containerized environments for running NVIDIA NIM microservices, NeMo frameworks, Triton inference servers, and RAPIDS pipelines within secure and performance-optimized GPU VMs.

Zadara zCompute system abstracts and simplifies infrastructure complexities. It exposes APIs, enforces quotas, and isolates GPU tenancy. This could be used to deploy production grade AI workloads. Identity management, multi-tenant control policies, and resource quotas are all in place in alignment with NVIDIA's Agentic AI Factory architecture guidance.

Zadara's system could also serve as the infrastructure foundation for NVIDIA Mission Control and Dynamo. While Zadara manages the orchestration and lifecycle of GPU VM instances, NVIDIA Mission Control could be layered into the deployment to provide cross-site observability and DPU-level policy enforcement. In this model, Zadara provisions and manages the GPU cloud infrastructure, and NVIDIA Mission Control offers centralized insight and lifecycle tracking across sites and fabrics. NVIDIA Mission Control does not orchestrate containerized workloads across tenant Kubernetes clusters, but it can integrate with Zadara via agent and API-level extensions to monitor infrastructure state, DPU health, and policy status over time. Zadara provides the secure multi-tenant substrate and orchestration layer required to operationalize these control planes in both sovereign and service-provider-hosted environments.

To demonstrate the breadth of integration possibilities, the following software components from NVIDIA's AI Enterprise stack and AI Factory design could be deployed atop Zadara:

- **NVIDIA NIM:** Containerized microservices offering optimized AI model inference endpoints
- **NVIDIA Triton Inference Server:** Scalable, multi-framework inference serving inside VM- or container-managed clusters
- **NVIDIA NeMo:** Training and fine-tuning pipelines for LLMs and foundation models
- **NVIDIA RAPIDS:** GPU-accelerated data analytics and preprocessing libraries
- **Agentic orchestration layers:** Integration points for emerging AI agent workflows coordinated through services like NVIDIA Dynamo

Together, these integrations enable Zadara to support modular, scalable deployments of AI Factory architectures, merging infrastructure readiness with full-stack execution of NVIDIA AI workloads.

SUMMARY

Zadara addresses modern IaaS critical requirements such as compliance, GPU density, operational control, cost efficiency, and ecosystem alignment. It enables NVIDIA Solution Architects and partners to deliver value faster by removing complexity and surfacing a fully integrated GPU infrastructure that can be deployed into any data center, telco, enterprise, or regional cloud environment.

Zadara's deployment of NVIDIA's IaaS reference architecture enables real-world sovereign GPU cloud environments with superior performance, security, and compliance. This implementation spans GPU networking, VM tuning, and full-stack isolation with zero operational compromise.

To learn more about deploying secure, AI-optimized GPU infrastructure using Zadara's validated approach of NVIDIA-powered sovereign AI infrastructure contact us.