zadara

NEUTRAL AI FACTORY

WHAT IS A NEUTRAL AI FACTORY?

AI Factory is a multi-node computing engine, purpose-built for AI workloads. Its operation involves three major stages. The first is the data acquisition pipeline and its related algorithms. The second stage is data ingestion into AI model creation and training (reasoning, language, vision or other), and the third stage is large volumes of decision and intelligence inferences, based on trained models, producing a series of generated tokens.

Driven by rapid growth in AI demand, building an AI Factory represents a significant opportunity for telcos, data center builders and operators.

Now they can build an AI Factory that will host numerous AI application providers, letting enterprises as customers have a choice to decide with what AI application provider to engage, and what type of AI application to consume. This model is, in essence, a Neutral AI Factory. A key advantage of a Neutral AI Factory is its ability to support sovereign AI for local enterprises through strong data residency controls and adherence to digital sovereignty requirements. This is particularly true compared to hyperscalers that are leveraging their cloud dominance and economics of scale to entrench the AI cloud market, while lacking in guaranteeing digital sovereignty. However, AI factories require considerable capital investments in high-performance computing infrastructure. All AI workloads require heavy GPU machinery and other expensive hardware components and systems.

Therefore, the establishment of an AI factory requires careful planning to justify such major capital investments for high-profile and expensive hardware, for large and intensive workloads. For many AI Factory builders, the primary challenge is ensuring an attractive Return-On-Investment (ROI) on the significant capital investment, achieved by maintaining high utilization of the factory. While enterprise demand for AI services is strong, AI application requirements vary widely across verticals. Most AI Factory builders lack the domain expertise to develop solutions for multiple industries and are wary of the sales and marketing investment required to pursue enterprise customers directly.

Recent evolutions in AI customization techniques such as Retrieval Augmented Generative AI (RAG), and growing accessibility of high performant pre-trained open source models (e.g., Qwen, Claude, DeepSeek R1), opened a myriad of possibilities for a growing number of enterprise AI applications & use cases, that can be optimized for enterprises of any scale. Such use cases evolve across numerous areas, including copilots, support chatbots for customer service automation, data search and retrieval, meeting summarization, and more.

The majority of enterprises can not obtain the required investment to kick start an in-house AI application development, let alone invest in building an AI factory themselves. While hyperscaler AI solutions are available, they are costly, poorly aligned with enterprise specific requirements, and often fail to meet digital sovereignty needs. Although well suited for consumer use cases, they are not a strong fit for enterprises, particularly in regulated industries. While bare metal GPU access is not new, it falls short of meeting enterprise-grade requirements around performance isolation, cost predictability, governance, and data sovereignty.

Managed Service Providers (MSPs) see AI-as-a-Service as a strong growth opportunity and are experiencing significant demand from both existing enterprise customers and new prospects. Unlike most enterprises, MSPs have the skills and domain knowledge to develop or tailor AI services for the specific verticals they serve. However, MSPs lack the ability to obtain the capital investment required to build and operate an AI factory.

While MSPs can use hyperscalers' AI cloud services, hyperscalers' margins leave little room for MSP profitability. Moreover, hyperscalers' lack of support for sovereign AI is an issue for many MSPs customers. Therefore, MSPs require access to an AI factory to stay competitive and offer differentiated AI-as-a-Service. In practice, they seek to consume a fraction of such a factory. On top of this fraction of a factory they can run their home grown or tailored AI applications in a SaaS-like model for their enterprise customers. Unlike traditional SaaS, built on general-purpose cloud infrastructure, enterprise-grade AI applications as a service require a cloudified AI factory, consumed elastically and only to the extent required.

These shared challenges across AI factory builders, MSPs, and enterprises point to a new operating model: the Neutral AI Factory.

In this model, AI factory builders dynamically provision shared hardware capacity to multiple MSPs based on real-time demand. Each MSP draws AI capacity on demand to run the right AI applications for its customers.

For AI factory builders, the Neutral AI Factory accelerates ROI and drives high utilization by tapping into established MSP demand. It also creates a repeatable go-to-market model that avoids the need to sell directly to enterprise verticals or deeply understand each industry's specific requirements. As a result, builders can significantly reduce sales and marketing investment.

For MSPs, the model delivers competitive AI infrastructure on demand, enabling them to focus on packaging, customizing, and operating AI services for enterprise customers, including regulated industries.

Last but not least, this model is also well suited for sovereign AI by supporting data residency, digital sovereignty requirements, and locally controlled AI operations.

This is where multi-tenant cloud operations on top of a neutral AI factory, built on GPU clusters, play a critical role.

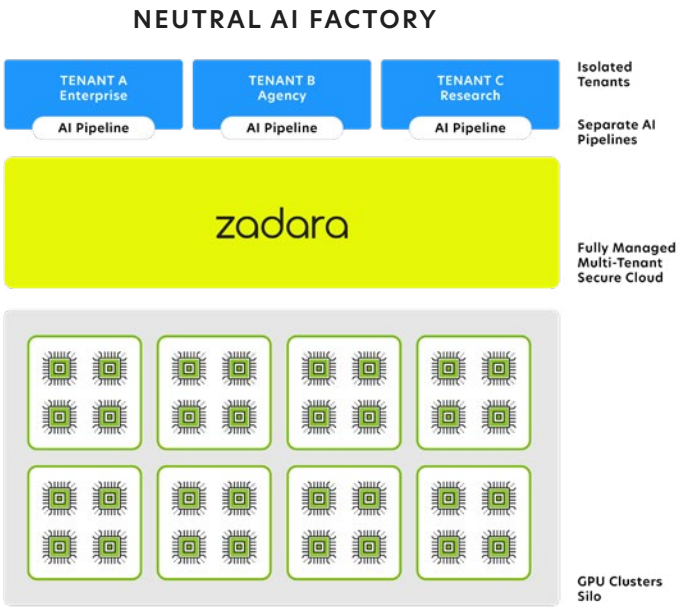

Fig. 1: By using Zadara cloud it is now possible to successfully address multi-tenancy challenges, and to operate a Neutral AI Factory. Zadara creates virtual sub-factories for each tenant, which in turn runs its own AI pipeline.

ZADARA ENABLEMENT OF NEUTRAL AI FACTORY

Zadara delivers a purpose-built cloud platform for operating a neutral AI factory on NVIDIA GPU clusters, enabling multiple independent MSPs or AI service providers to securely coexist, scale, and innovate on shared AI factory infrastructure. The three most critical challenges in making this model operational at scale are addressed by Zadara.

CHALLENGE ONE: SECURING AND SERVING MULTIPLE DISTINCT TENANTS ON A SHARED GPU CLUSTER

A neutral AI factory operates as a shared AI industrial zone, in which multiple operators of enterprise AI applications (e.g., co-pilot operator, or meeting summarization service provider), onboard as tenants to deploy and operate their AI workloads. Each tenant, in turn, serves its downstream customers, including enterprises, agencies, and research organizations.

These tenants are mutually independent and must remain isolated from one another. Their tenancy must be securely siloed, confined and protected within a clearly defined portion of the AI factory.

Addressing this challenge requires a robust, infrastructure-level multi-tenancy model that Zadara provides. By using Zadara cloud control plane, Factory operators are capable of carving out a virtual "sub-factories" from shared physical GPU clusters, granting each tenant its digital sovereignty to run its own AI frameworks, data pipelines and MLOps workflows, guaranteed against cross-tenant interference, and data leakage.

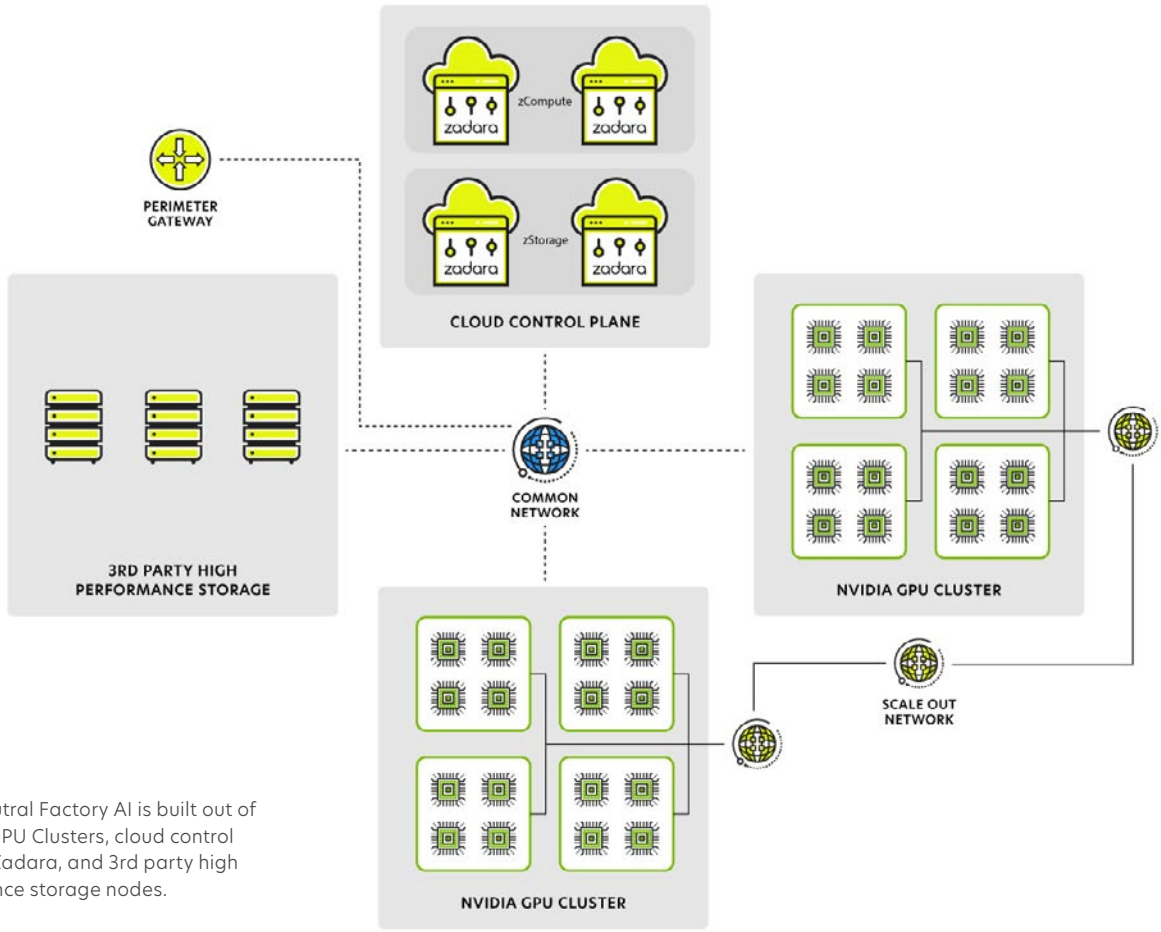

Fig. 2: Neutral Factory AI is built out of NVIDIA® GPU Clusters, cloud control plane by Zadara, and 3rd party high performance storage nodes.

CHALLENGE TWO: SUPPORTING HETEROGENEOUS WORKLOADS AT SCALE

AI workloads across tenants are inherently heterogeneous. Inference pipelines, agentic AI services, retrieval-augmented generation (RAG) workflows, and large model training jobs, each exhibit distinct execution patterns, memory footprints, and GPU utilization characteristics. Zadara enables GPU cloud infrastructure to dynamically, securely and concurrently support diverse workloads types and scales on shared physical resources. Each workload must execute within an independently governed virtual environment, with enforced isolation across compute, memory, storage, and network domains. This approach enables high cluster utilization while preserving deterministic performance, operational fairness, and compliance with regulatory and service-level requirements.

CHALLENGE THREE: MAXIMIZING DENSE AND EFFICIENT WORKLOAD PLACEMENT

AI factories cloud operators make substantial capital investments in GPU infrastructure and therefore must maximize cluster utilization by densely packing and efficiently tiling multiple paying tenants across shared GPU resources. Achieving high utilization is essential to sustaining economic viability while preserving predictable performance.

At the same time, tenants (MSPs, managed AI service providers or enterprises operating their own AI workloads) lease a defined portion of the AI factory as a virtual GPU sub-cluster, and each tenant seeks to fully leverage its allocated resources throughout the lease period, without exposure to contention, disruption, or security risk.

Precise, automated orchestration at the infrastructure layer becomes crucial.

This is where Zadara's automated, secure, and production-proven virtual cloud infrastructure orchestration delivers its full value. Zadara dynamically allocates and manages GPU resources across virtual AI sub-factories, enabling dense workload placement without compromising tenant isolation, regulatory compliance, or operational sovereignty.

AI FACTORY: PHYSICAL INFRASTRUCTURE BUILDING BLOCKS

AI Factory foundational building blocks include data center silo construction, cooling and electricity engineering, assembly of hardware infrastructure of several types, software suites with related services, and more.

The specifications herein provide abstract synopsis of the compute, networking and storage hardware infrastructure components, and give a high level overview of the software pieces.

From the hardware components perspective, a neutral AI Factory cloud consists of: (a) the NVIDIA® GPU cluster nodes; (b) Zadara zCompute and zStorage control plane nodes; (c) AI applications, MLOps, and Kubernetes masters control server nodes; (d) 3rd party high performance low latency storage nodes; (e) Network switches for three types of networks: GPU scale up fabric, common converged network for high performance storage, cloud management and MLOps control, and lastly cloud admin and user access network.

NVIDIA® GPU CLUSTER NODES

AI Factory consists of one or more of NVIDIA® defined cluster units. Two types of NVIDIA® GPU cluster architectures are supported by Zadara, as described below, in Table 1.

TABLE 1: NVIDIA® CLUSTER TYPES, BASIC UNIT

#		NVIDIA® NVL72 RACK-SCALE	SIZE	NVIDIA® CLOUD PARTNER (NCP) REFERENCE ARCHITECTURE, SCALABLE UNIT	SIZE
1	**GPU Configuration**	NVIDIA® Blackwell GB200 or Blackwell Ultra GB300 GPUs	72	NVIDIA® H100/H200/B200/B300/ L4/L40S/RTX Pro™ 6000	256
2	**CPU Configuration**	NVIDIA® Grace™ CPUs (72 ARM Neoverse V2 cores)	36	Intel® x86 on NVIDIA® DGX and HGX	64
3	**Memory Capacity**	Per GPU: 288 GB HBM3/HBM3e Per CPU: 512 GB DDR5		Per GPU: 80 GB HBM and up Per CPU: 1TB-2TB DDR	
4	**Interconnect (NVLink and NVSwitch)**	NVIDIA® 5th-generation NVLink C2C		NVIDIA® NVLink fabric	
5	**Scale Out Network**	NVIDIA® ConnectX-8 2 x 400GbE	72	NVIDIA® ConnectX-7 2 x 200GbE	256
		NVIDIA® Quantum-X800 InfiniBand or NVIDIA® Spectrum-X™ Ethernet	3	NVIDIA® Quantum-X800 InfiniBand or NVIDIA® Spectrum-X™ Ethernet	7
6	**Common Converged Network**	64 x 200G Fabric and Ethernet Switch; NVIDIA® SN4600 or other	4	64 x 200G Fabric and Ethernet Switch; NVIDIA® SN4600 or other	4

CONTROL PLANE:ZADARA CLOUD AND MLOPS PLATFORM

The GPU cloud control plane, which is a separate domain where management, control and applications run, consists of Zadara cloud control plane, as well as AI applications, MLOps, and Kubernetes masters control plane. The control plane of Zadara cloud consists of zCompute cloud management & control subsystem, and zStorage Elastic Block Storage boot volumes. Control plane utilities operate on dedicated standalone servers, separate from the GPU cluster.

Similarly, a list of control plane utilities, e.g., AI applications pipeline control utilities, including MLOps and Kubernetes masters, ML platform management, and more, operate on separate and dedicated standalone server nodes. For every 1,000 GPUs on the NVIDIA® cluster that are managed by Zadara cloud, additional standalone control plane server nodes shall be required, as described in Table 2 herein.

TABLE 2: CLOUD CONTROL PLANE INFRASTRUCTURE NVIDIA® CLUSTER TYPES, BASIC UNIT

#	PURPOSE	# OF NODES
1	zStorage: Zadara Elastic Block Storage (EBS) volumes	2
2	zCompute: Zadara compute cloud control	3
3	AI platforms and AI applications control planes (for, e.g., PaaS, Kubernetes masters, AI & data applications, MLOps).	3

Control plane server nodes are x86 compute platforms (applicable on both types of NVIDIA® Clusters including Scalable Units and Rack-Scale), running control and management specific workloads. The required specifications of the control plane platforms appear in Table 3 herein.

TABLE 3: CONTROL PLANE SERVER SPECIFICATIONS

	zCOMPUTE CONTROL NODE/ PAAS CONTROL NODE	zSTORAGE ELASTIC BLOCK VOLUME NODE
Processor*	Intel® Xeon® 6 Performance 67xxP "Granite Rapids" Processor 32/64/160 Cores	
Memory	512GB	256GB
Boot NVMe	1 x M.2 NVMe PCIe4 480GB	
Storage*	–	32 x 7.68 SATA 6Gb/s SSD, Samsung PM88
Network	1 x 1GbE RJ45 Dual port 25GbE SFP28 ConnectX6-LX	2 x 1GbE RJ45; 2 x 100GbE QSFP28 RDMA support; ConnectX6-Dx
BMC/IPMI	1G OOB Management w/full remote Keyboard/Video/Mouse/CD	

* Core count sizing, and number of SSD drives are subject to projected workload scale.

Note: Control plane servers are intended to run control and management workloads only.

GENERAL PURPOSE CLOUD NEXT TO AI GPU CLUSTER CLOUD

Zadara supports a configuration where general-purpose compute cloud (VMs, K8s, etc.) operates next to the AI GPU cluster cloud deployment, for the purpose of general compute workloads.

In such cases, additional dedicated x86 servers for general-purpose compute cloud workloads are deployed as general purpose cloud clusters. The Zadara control plane manages both the general-purpose compute clusters and the NVIDIA® GPU clusters. The control plane server specifications remain as described in Table 3, while the exact number of required control plane servers may differ from the configuration outlined in Table 2, and may increase. For further details and sizing guidance for such hybrid AI clusters and general purpose cloud setup, please contact Zadara sales.

3RD PARTY HIGH PERFORMANCE, LOW LATENCY STORAGE AND DATA SERVICES

High performance, low latency storage and data services are foundational requirements for AI operations and MLOps. Data movement (not compute) is the more common bottleneck across the AI pipelines and lifecycle. While in training, GPUs must continuously stream large volumes of data, and any storage latency or bandwidth shortfall directly results in idle GPU time. In inference and retrieval augmented generation (RAG), fast access to model artifacts, embeddings, and feature stores is fundamental to achieving best "time to first token" and to eliminate tail latency.

From the MLOps perspective, storage is the basis to the entire control loop. Versioned datasets, model checkpoints, metadata, CI/CD pipelines, models rollback and the rest of MLOps operations, all rely on fast, concurrent and consistent access to data.

AI platforms operate storage as a parallel service, with NVMe arrays and scale out architectures that are integrated with the AI fabric. This way high throughput is delivered with microsecond level latency.

High performance low latency storage is to be supplied by a third party vendor. However, as a guideline, for every 256 GPUs in the NVIDIA® GPU cluster, the following types of servers shall be required for storage scale out.

TABLE 4: HI PERFORMANCE LOW LATENCY SCALE OUT STORAGE NODES, GUIDELINE

	CORE HIGH PERFORMANCE DATA PLATFORM NODE	HIGH CAPACITY STORAGE ELEMENTS NODE
Processor*	Dual Intel® Xeon® 6 Performance 67xxP "Granite Rapids" Processor 32/64/160 Cores	Dual Intel® Xeon® 6 Performance 65xxP "Granite Rapids" Processor 12/16
Memory	256GB	64GB
Boot NVMe	1 x M.2 NVMe PCIe4 480GB	
Storage*	–	44 x SSD; 12 x SCM
Network	2x 100GbE/HDR100, 2x 200GbE/NDR200 (Dual NIC)	4 x 100GBE OR 4 x EDR IB
BMC/IPMI	1G OOB Management w/full remote Keyboard/Video/Mouse/CD	
# of Nodes	13	12

COMMON NETWORKING INFRASTRUCTURE

The common converged network in the NVIDIA® AI cluster is used for several types of network traffic, including high performance storage RDMA, GPU cloud operations and control, user access management, MLOps management, and more. The common network physical architecture is elaborate and should adhere to the technical requirements of each of the mentioned uses. However, as a guideline, the switch models that described herein in Table 5 should be considered for common network.

TABLE 5: GUIDELINE, COMMON NETWORK SWITCH FABRIC

#	SWITCH MODEL	I/O
1	NVIDIA® SN5600; NVIDIA® SN5610	64 x 400GbE/800GbE OSFP; 2 x 10GbE/25GbE SFP28
2	NVIDIA® SN5400	64 x 400GbEQSFP-DD; 2 x 10GbE/25GbE SFP28