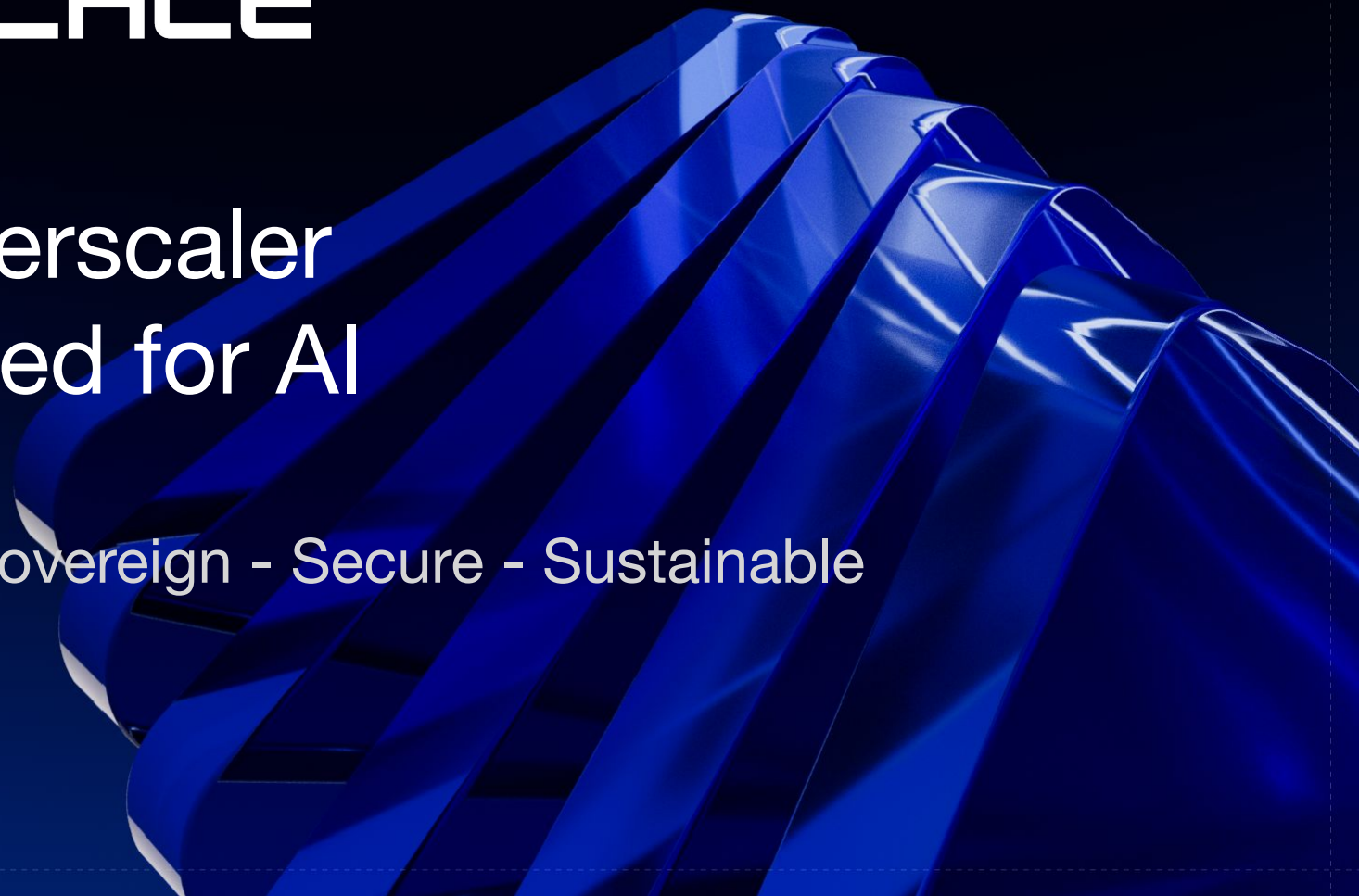
NSCALE

The Hyperscaler engineered for AI

Scaleable - Sovereign - Secure - Sustainable



NSCALE

Storage scaling and performance for massive-scale GPU infrastructure

Elise Jennings
Head of HPC

Nscale - The hyperscaler engineered for AI

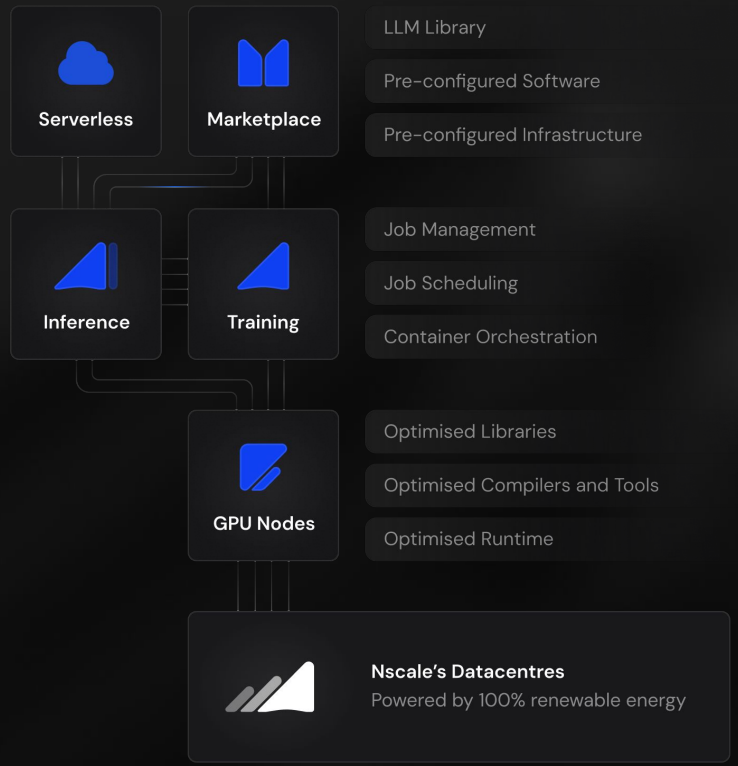
Nscale designs, builds, and operates **sustainable modular AI-ready data centres** and **delivers large-scale GPU clusters** to accelerate AI and advanced computing globally.

Nscale provides a **fully integrated suite of AI services** and compute, reducing costs and running AI workloads efficiently on a unified platform. Designed to simplify the journey from development to production, whether using Nscale's built-in AI/ML tools or your own.

Nscale accelerates training and inference workloads with best-in-class GPUs, high-performance storage, and advanced networking to offer unmatched performance and scalability.

Key Benefits

- Advanced optimisation for AI workloads
- Transparent pricing and cost-effective compute resources
- Expert support from AI specialists
- 100% renewable energy-powered data centres



Sovereign AI with Nscale

Infrastructure Built for Sovereign AI
Nscale delivers dense, high-performance GPU infrastructure designed to meet the needs of national AI programmes. Our scalable, sustainable data centres enable governments to retain control over data, models, and infrastructure—critical for digital sovereignty.

White-Label National AI Cloud
Nscale's full-stack AI platform supports the entire generative AI lifecycle—from training to deployment—and can be white-labeled to launch sovereign AI cloud services under a national brand.

We Provide

- Bare metal and virtualised GPU clusters
- Fully managed Kubernetes (NKS) and Slurm
- Serverless and dedicated inference endpoints
- Multimodal playground for fine-tuning and RAG workflows

AI Factory-as-a-Service
Our modular data centre model accelerates time-to-launch, with infrastructure deployed where it's needed, powered by renewable energy, and designed for long-term scalability.

NSCALE

Modular AI Factories

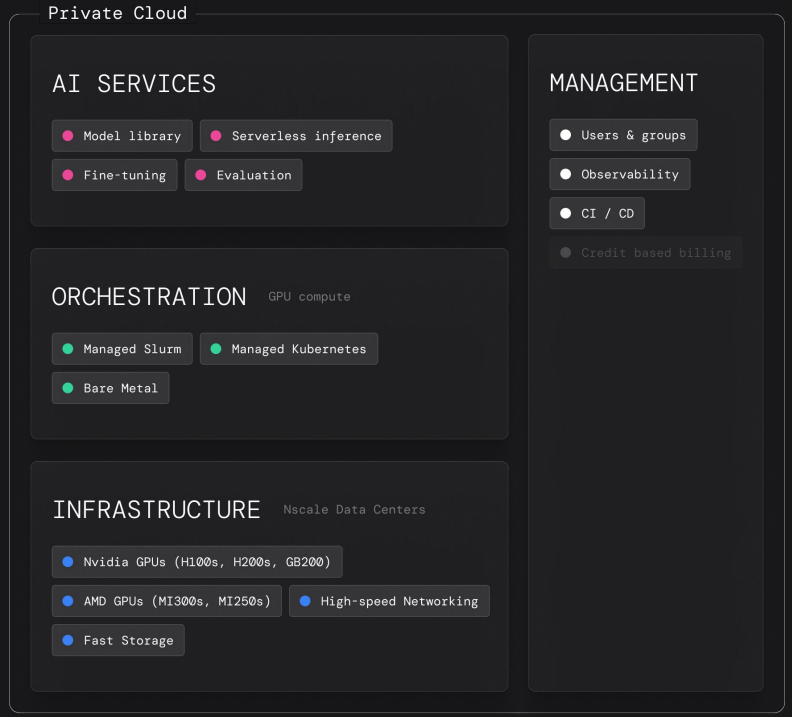
Nscale Private Cloud

Overview
A self-serve private cloud solution for enterprises, model makers, and hyperscalers, offering sophisticated cloud and orchestration capabilities such as Slurm, Kubernetes, and bare metal, with flexible contractual agreements.

Target
Hyperscalers, model makers, enterprises

Status
Full managed ✓
Self-serve
Whitelabel

Commercials
Take or pay long term contracts, 3+ years



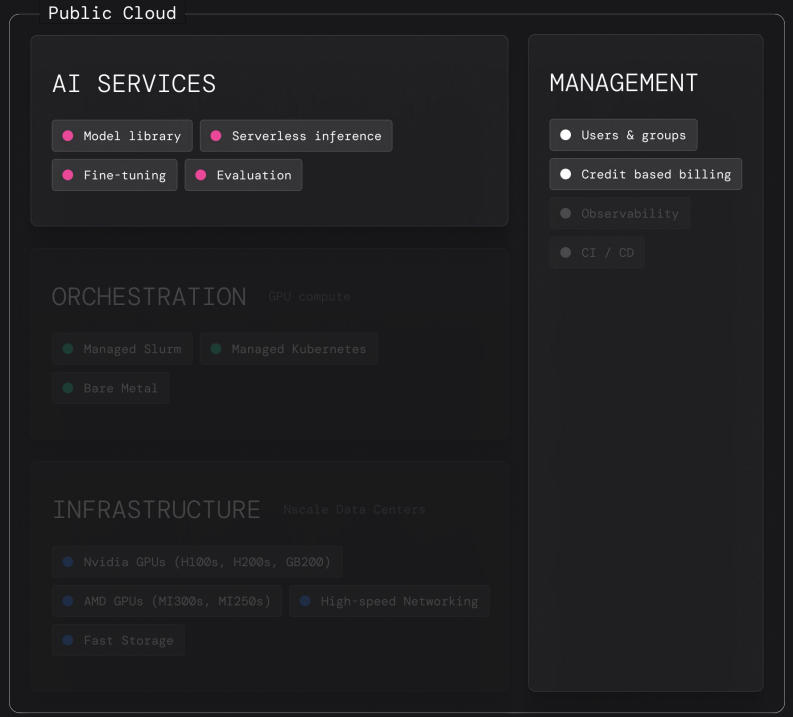
Nscale AI Services

Overview
An on-demand cloud platform designed for all businesses, providing seamless AI tools for fine-tuning, scalable inference, evaluation, and a curated model library

Target
Anyone, likely SMEs

Status
- Serverless Inference ✓
- Credit-based billing ✓
- Fine-tuning

Commercials
Credit based on demand, pre-paid.



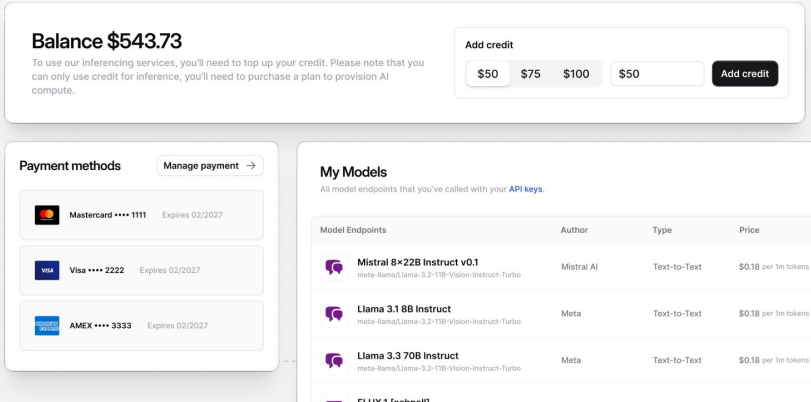
AI Services

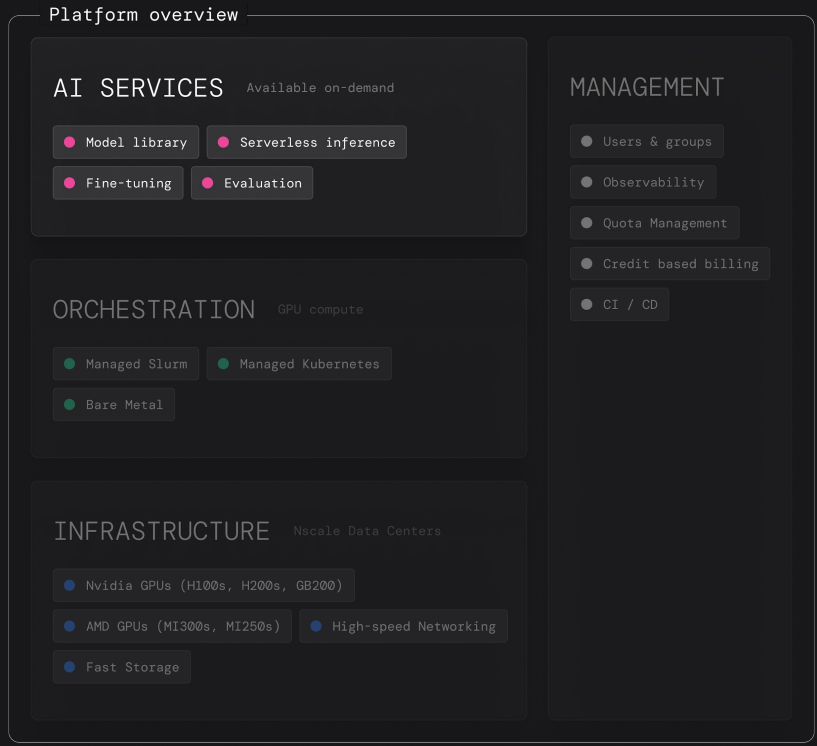
Nscale's on-demand AI services are designed for businesses of all sizes, providing tooling for fine-tuning, evaluating, deploying, and inferencing with AI models. These services leverage Nscale's underlying technology but are available in a self-serve and prepaid capacity to appeal to a wider audience.

Technology

Nscale's AI services are built using a combination of in-house development and open-source technologies. Rigorous benchmarking and GPU-specific testing is conducted to select technologies such as:

PyTorch





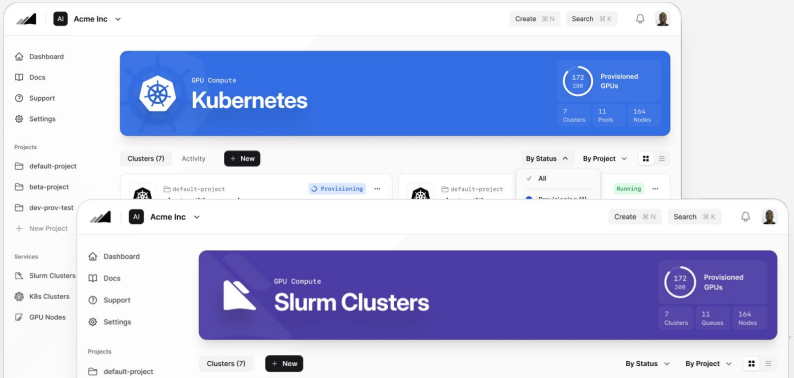
Orchestration

Nscale provides a proprietary orchestration and scheduling environment for AI workloads. This is offered through three configuration options—Bare metal, Managed Slurm, and Managed Kubernetes—giving Nscale coverage across all enterprise needs and allowing customers to orchestrate infrastructure at scale.

Technology

Nscale's orchestration capabilities leverage popular technologies to create an abstracted service purpose-built for AI. Open-source technologies we leverage as part of our proprietary solution include:

slurm workload manager · kubernetes · vcluster · argo · cilium



Platform overview

AI SERVICES — Available on-demand
- Model library
- Serverless inference
- Fine-tuning
- Evaluation

ORCHESTRATION — GPU compute
- Managed Slurm
- Managed Kubernetes
- Bare Metal

INFRASTRUCTURE — Nscale Data Centers
- Nvidia GPUs (H100s, H200s, GB200)
- AMD GPUs (MI300s, MI250s)
- High-speed Networking
- Fast Storage

MANAGEMENT
- Users & groups
- Observability
- Quota Management
- Credit based billing
- CI / CD

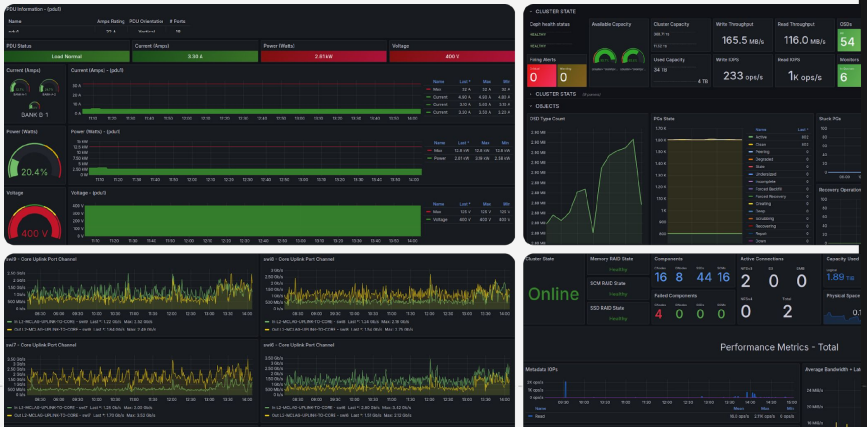
Management & Observability

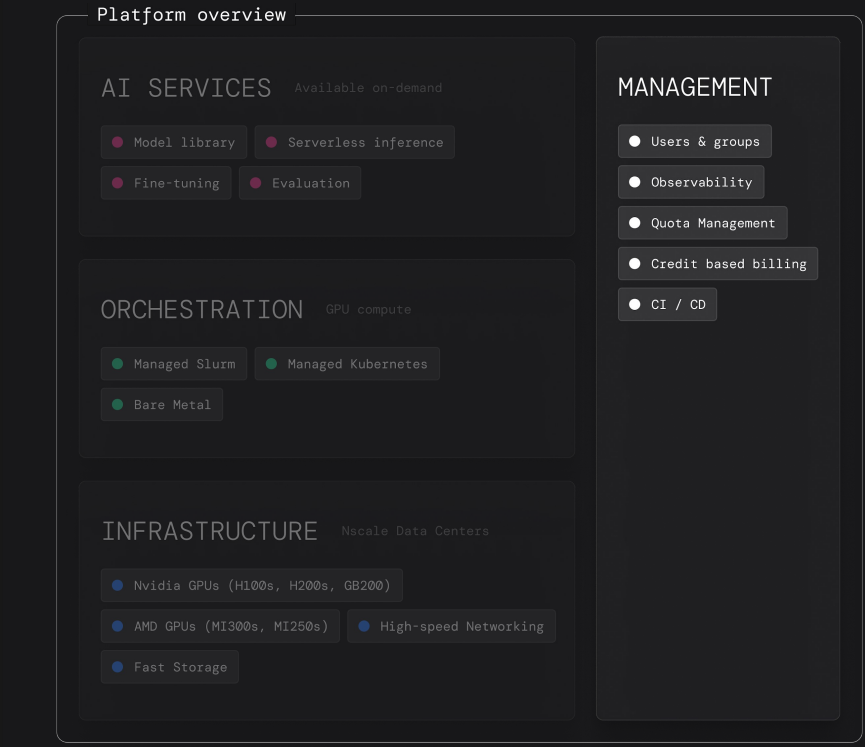
Nscale provides sophisticated management and observability into the infrastructure being consumed, allowing customers to continually monitor and optimise their cloud environments. These capabilities are flexible and built to work alongside any combination of Nscale services, allowing the platform to adapt to all enterprise use cases.

Technology

Nscale leverages open-source technologies to create a cloud environment optimised for GPU-based infrastructure. Open-source technologies we leverage as part of our proprietary solution include:

Grafana Prometheus HARBOR authentik





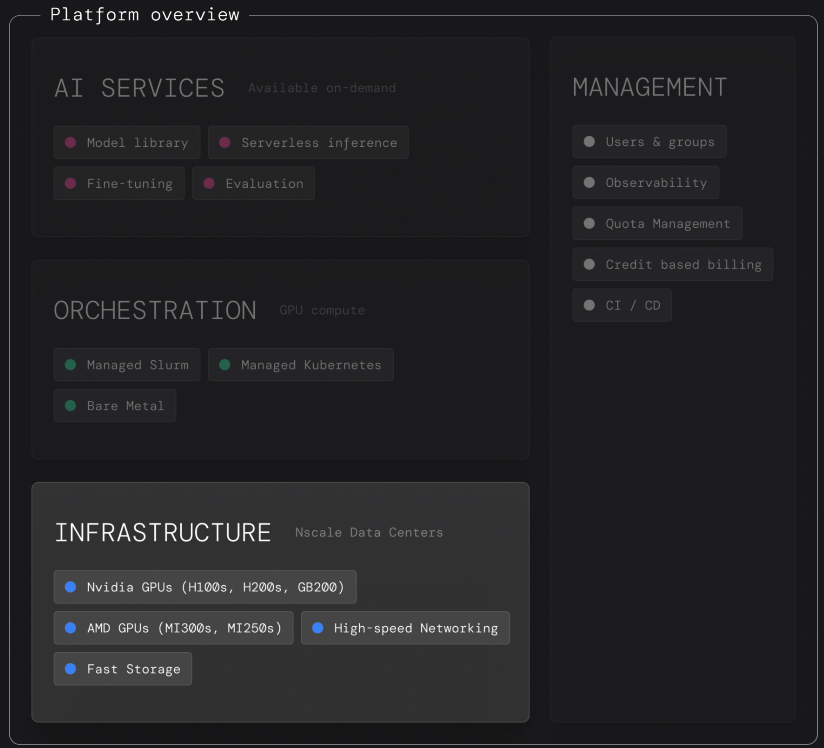
Infrastructure

Nscale, unlike public cloud providers with fixed reference architectures, can design and deliver a bespoke GPU cluster to specific customer requirements, maximising performance while minimising cost on computationally intensive AI workloads.

Technology

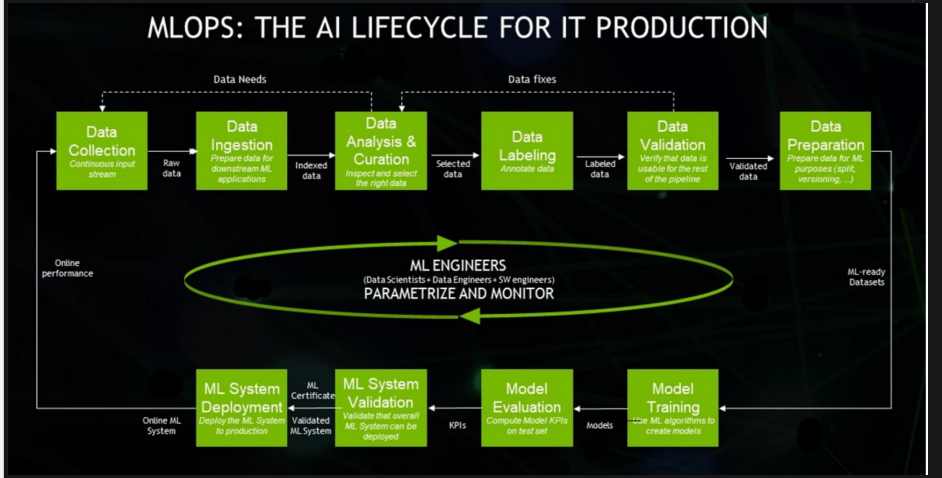
The Nscale Cloud Stack is technology agnostic. Nscale's unique approach enables the deployment of the latest technologies across the stack from GPU accelerators, to high-performance networking fabrics, to fast file storage.

NVIDIA. Lenovo



Infrastructure –Storage

- AI workloads at scale demand intelligent management of both infrastructure and storage.
- Storage systems play a critical role in AI training and inference performance
- Access to stored data can become the bottleneck in the entire system.
- DC design and orchestration requires mapping entire storage data path to correctly assess the delivered performance e.g, DAS and NAS
- Understanding PCIe topology and determining:
  - which IO devices and GPUs are on the same PCIe switch or root complex
  - communication paths which might traverse multiple PCIe ports or cross CPU socket boundaries



https://blogs.nvidia.com/blog/what-is-mlops/

Key storage concerns for AI hyperscalers

- Scalability
    - Adding more nodes and accelerators to serve ever-larger training datasets increases the throughput demands
- Resiliency, fault tolerance, availability
- Data diversity: distributed, parallel file systems crucial for multi-GPU, multi-node training, object and container access
- I/O characteristics of different models/workloads
    - Low latency, high IOPS for short bursty workloads
    - High throughput training/data staging
- Data security and lineage

Usefulness of industry standard benchmarks for hyperscalers

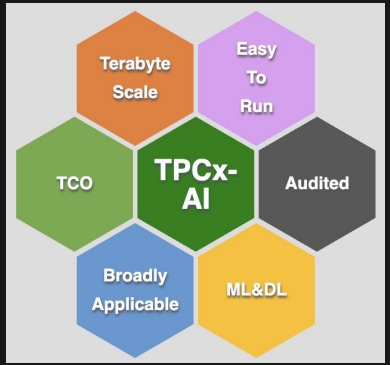
- beyond fio for disk/storage performance needs
- effective tool for purchasing, configuring, and optimizing storage
- Effective reference for designing next-generation systems

Nvidia provide specific benchmarking suites for e.g. NVIDIA® GPUDirect® Storage (GDS)

Benchmarking

TCPx-AI

- benchmark focused on emulating the behavior of representative industry AI solutions that are relevant in current production datacenters and cloud environments.
- diverse representative dataset up to Terabytes
- data management stages, and data science pipeline:
    - Data diversity: different sources and formats
    - data management stages: cleansing, exploration, preprocessing
    - mimic modern commercial pipelines from production environments
    - training, serving and scoring phases using production datasets available in the datacenter



Benchmarking

MLPerf Storage

- V1.0 suite May 2023
- Goal is to model the I/O patterns posed by AI workloads
- Measure sustained perf for PyTorch and Tensorflow for BERT and 3D UNet
- Emulation method to simulate variety of GPUs (A100, H100)

- Highlights hardware specifics
  H100 per-batch computation time for the 3D-UNet workload reduced by 76% compared to V100: converts a typically a bandwidth-sensitive workload into one which is latency-sensitive.

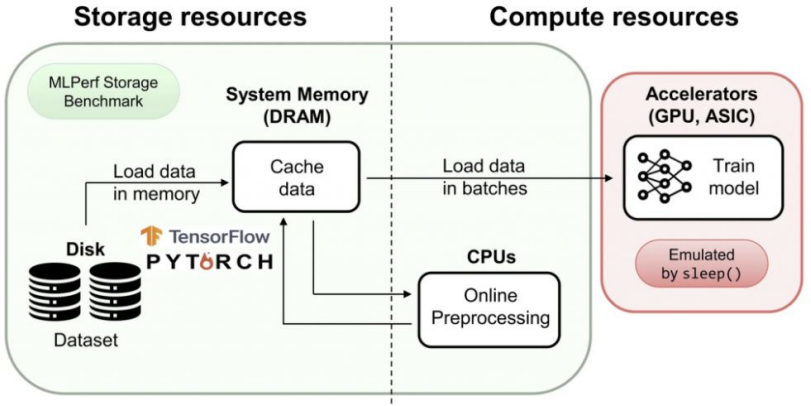
- Distributed training presents specific challenges for a storage system not only in delivering higher throughput but also in serving multiple training nodes simultaneously.

"The MLPerf Storage v1.0 results demonstrate a renewal in storage technology design," said Oana Balmau, MLPerf Storage working group co-chair. "At the moment, there doesn't appear to be a consensus 'best of breed' technical architecture for storage in ML systems: the submissions we received for the v1.0 benchmark took a wide range of unique and creative approaches to providing high-speed, high-scale storage."

BENCHMARK SUITE RESULTS

MLPerf Storage

Area	Task	Model	Nominal Dataset
Vision	Medical image segmentation	3D U-Net	KITS 2019 (602x512x512)
Vision	Image classification	ResNet50	ImageNet
Scientific	Cosmology parameter prediction	CosmoFlow	CosmoFlow N-body simulation
Language	Language processing	BERT-large	Wikipedia (2.5KB/sample)



https://mlcommons.org/working-groups/benchmarks/storage/

NSCALE

Thank You

elise.jennings@nscale.com

May 2025