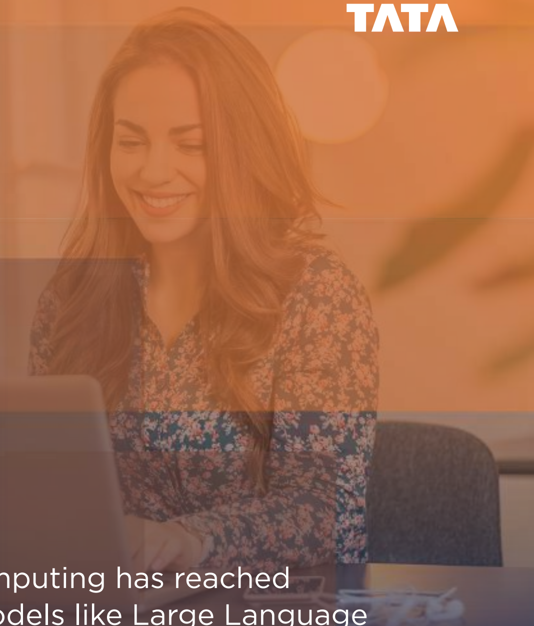
TATA COMMUNICATIONS

TATA



Tata Communications Vayu AI Cloud

GPU-as-a-Service

In today's data-driven world, the demand for high-performance computing has reached unprecedented levels. From training advanced machine learning models like Large Language Models (LLMs) to powering real-time edge computing applications, the need for scalable, efficient, and cost-effective solutions is critical.

GPU-as-a-Service (GPUaaS) bridges the gap, offering unparalleled computational power without the need for hefty upfront infrastructure investments. Whether you're an enterprise optimising AI workflows, a startup scaling innovative applications, or a government agency tackling complex simulations, GPUaaS delivers the flexibility, scalability, and security to meet your needs—all while driving cost efficiency and sustainability.

Tata Communications Vayu AI Cloud GPU-as-a-Service

Power your business with on-demand, high-performance GPU resources, offering flexibility, reliability, and cost efficiency.

Bare metal-as-a-Service

- Dedicated, physical host with GPUs, without any virtualisation.
- Full control over the hardware.
- Bring-your-own orchestration.

Cluster-as-a-Service

- Multiple dedicated GPU instances.
- Horizontal auto-scaling, clustered with single cluster engine (either as Kubernetes or as SLURM) along with required set of drivers for creating virtual GPU (Multi-Instance GPU).
- Resource pooling and IDEs installed in cluster for deployment of AI workloads.

Key features

GPU configuration and flexibility

- **Variety of GPUs:** H100, L40S.
- **Flexible pricing:** Pay-per-use and reserved options.

Easy provisioning and setup

- **User-friendly TC^{x} portal:** Simplified GPU provisioning.
- **Pre-installed AI frameworks:** CUDA, cuDNN, NCCL, NVIDIA AI Suite.

High-performance infrastructure

- **Scalable storage solutions:** 105 GB/s read, 75 GB/s write, 3M IOPS.
- **Superior data access:** Accelerate AI workloads.

Secure and efficient networking

- **BYON connectivity:** Leverage existing network infrastructure (ILL, MPLS, P2P).
- **Private and secure connections:** Site-to-site and client-to-site setups.
- **Multi-cloud connect:** Seamless connectivity to other clouds and on-prem data centers.

Scalable AI workloads

- **Support for orchestration engines:** Kubernetes, SLURM for scalable training and inference.
- **Fine-grained data control:** Ingress and egress control for enhanced security.

Configuration details

Compute

S.No	SKU	AI compute unit model	Configuration	GPU memory (GB)	Performance		Peer-to-peer bandwidth (Gbps)	Network bandwidth (Gbps)
					FP32	FP16		
1	AI.H100.IB.8X	NVIDIA H100 SXM	8* H100 GPU, 224 vCPU, 1024 GB RAM	640	67	1979	3200	3200
2	AI. L40S.4X	NVIDIA L40S	4* L40S GPU,128 vCPU, 512 GB RAM	192	91.6	733	400	180

Storage

S.No	Service name	Description
1	Object storage	Object storage service with S3 protocol support for storing datasets. Offered as per GB per month pricing model.
2	Parallel file system	High speed parallel filesystem with Lustre protocol for storing training data and intermediate model checkpoints across multiple GPUs. Provides performance throughput of 105 GB/s Read speed and 75 GB/s write speed and 3 million IOPS. Offered as per GB per month pricing model.

Network

S.No	Service name	Description
1	Internet data connectivity	Opt for bandwidth from 1 Mbps up to 10 Gbps. There is no usage-based ingress and egress cost.
2	Firewall-aaS	Detailed traffic rules, FQDN filtering, stateful firewalling, and NAT gateway, all in a single IDC-compliant SKU.
3	Loadbalancer-aaS	Seamless traffic distribution and high availability across multiple AI computes.
4	Virtual Private Network-aaS	Provides two services, one is IPSec Tunnel which provides site-to-site private connectivity, second is Client-to-site connectivity.
5	Multi-Cloud Connect	Seamless connectivity to multiple cloud providers and on-prem data centers.
6	Bring Your Own Network	Bring any third-party connectivity (ILL, P2P, MPLS) to Tata Communications Vayu AI Cloud to leverage existing investment on connectivity.

Pricing models

Pay-per-use	Pay only for GPU resources based on actual usage, without any long-term commitments.
Reserved-Instance	Commit to a fixed amount of GPU resources for a specified period (e.g. 6 months, 12 months, 1 year, 3 years).

TATA COMMUNICATIONS

TATA

"POWER" of our solution

Predictable

Maximise ROI with predictable costs and reduce egress costs by up to 40% through seamless multi-cloud connectivity options.

Optimised

Optimise large-scale AI training, fine-tuning, and on-demand inferencing—ensuring top performance, security, and compliance. Streamline data management with robust capabilities that reduce data noise and leverage enhanced Retrieval-Augmented Generation (RAG) for accurate, context-aware responses.

Well integrated

Modular architecture that allows GPU provisioning with pre-installed frameworks, APIs, and SDKs for seamless integration, supported by managed services with SLAs.

Efficient

Direct liquid cooling-enabled data-center deployments and efficient connectivity between GPUs and high-speed storage systems, such as parallel file systems, to enable distributed and latency-sensitive workloads.

Reliable

NVIDIA-certified GPUs for reliable performance and an end-to-end managed platform for scalable inference across all leading model frameworks.

vayu | AI Cloud

For more information, visit us at www.tatacommunications.com

CONTACT