

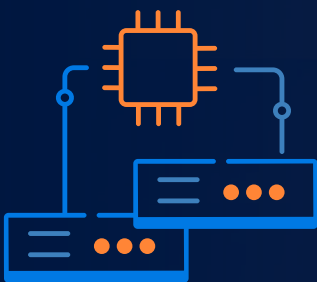
GPUaaS

THE ENGINE BEHIND AI TRANSFORMATION

Why GPU-as-a-Service matters? Let's break it down

Alright, let me ask you something. Have you ever wondered how your favourite apps – like those AI writing tools or even gaming platforms – seem to magically get smarter, faster, and more tailored to your needs every day? What's behind the scenes making that happen?

It's Graphical Processing Unit (GPUs). But not just any GPUs – we're talking about GPUs on demand, in the cloud, available when needed, and gone when they're not. It's called GPU-as-a-Service (GPUaaS). And it's the secret sauce transforming how AI applications get built and delivered to people like you and me.



GPU-as-a-Service (GPUaaS) has emerged as a critical enabler of high-performance computing. The demand for more advanced machine learning models, particularly **Large Language Models (LLMs)**, has skyrocketed in recent years. These models require an enormous amount of computational power to train and run. However, not all businesses or organisations have the resources to invest in the physical infrastructure necessary to support these power-hungry tasks.

Whether you're an enterprise training Large Language Models (LLMs), a startup building the next breakthrough foundation model, or a government driving your AI mission, GPU-as-a-Service (GPUaaS) is the core engine powering your ambitions.

Emerging trends in the GPU-as-a-Service market

The GPUaaS market is transitioning with several key trends corresponding to technological advancements and changes in industry requirements. These trends are altering how various sectors access and use GPU resources.



INCREASED AI AND ML WORKLOAD ADOPTION:

GPUaaS is driven by the increasing adoption of AI and ML applications across numerous industries. High-performance GPUs enable complex algorithms and big data processing, fuelling demand for GPUaaS solutions.

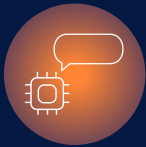


**GROWTH OF EDGE COMPUTING:**

The rise of edge computing requires GPUs to perform local processing tasks. This includes edge-based GPUaaS solutions to support real-time data processing and reduce latency in IoT and autonomous vehicles, among other applications.

**EXPANSION OF HYBRID CLOUD SOLUTIONS:**

Organisations are moving towards hybrid cloud models, taking advantage of GPUaaS for scalable and on-demand computing. This trend enables firms to combine public and private cloud assets, optimising cost-efficiency for different workloads.

**SUPPORT FOR LARGE LANGUAGE MODELS (LLMs):**

Growing demand for training and deploying LLMs and other generative AI models is boosting the adoption of GPUaaS.

**ENHANCED SECURITY FEATURES:**

Advanced encryption, access controls, and compliance with global data protection standards.











What is GPU-as-a-Service?

GPU-as-a-Service is a cloud-based model that offers users access to high-performance GPU resources via the Internet, eliminating the need for costly GPU hardware ownership and maintenance. This flexible, on-demand approach allows businesses to efficiently scale their computing capabilities, prioritising innovation while reducing the burden of managing infrastructure. As AI workloads scale, businesses must decide between managing GPUs in-house or opting for GPU-as-a-Service (GPUaaS). While on-premise GPUs offer complete control, they come with high upfront costs, maintenance demands, and scalability challenges. GPUaaS, on the other hand, provides flexible, cloud-based access to high-performance GPUs, eliminating infrastructure burdens while optimising costs and scalability.

But how do they compare? Let's find out!



Catagory	GPU (On-Premise)	GPU-as-a-service
 COST	High upfront investment (hardware, setup, and maintenance).	Pay-as-you-go or subscription-based, lower upfront cost.
 SCALABILITY	Limited by physical hardware and upgrade cycles.	Highly scalable with the ability to add resources on-demand.
 FLEXIBILITY	Fixed resources; difficult to adapt to fluctuating needs.	Dynamic allocation; ideal for varying workloads.
 MAINTENANCE	Requires in-house expertise for updates and troubleshooting.	Managed by the service provider, reducing operational burden.
 PERFORMANCE	High performance with direct control over hardware.	Performance depends on service tier and provider's infrastructure.
 ACCESSIBILITY	Restricted to local access or network configuration.	Cloud-based, accessible globally with internet connectivity.
 LATENCY	Low latency for on-site applications.	Potential latency due to cloud access and data transfer.
 DATA SECURITY	Greater control over sensitive data on-premises.	Depends on the provider's compliance and security protocols.

LLM and GPU's linkage

Large Language Models (LLMs) like GPT and others are revolutionising industries by processing and analysing vast amounts of data to generate insights and inform decision-making. GPUs (Graphics Processing Units) play a critical role in accelerating this process by providing the massive parallel computational power needed to train and run these models efficiently. Unlike CPUs, which handle tasks sequentially, GPUs can perform thousands of operations simultaneously making them ideal for handling the complex matrix multiplications and data-heavy computations that underpin LLMs. As data continues to grow exponentially, GPUs enable faster model training and real-time inference, allowing organisations to harness insights from large datasets almost instantly. This acceleration empowers businesses to make data-driven decisions quickly, improving outcomes in areas like personalised customer experiences, predictive analytics, and operational optimisation. In essence, GPUs serve as the backbone of LLMs, driving innovation in a data-rich world.

LLMs in action: Real-world examples for enterprises, startups, and governments

1 Enterprises: Revolutionising customer experience with LLMs

LLMs powered by GPUaaS can revolutionise customer support, sales, and marketing for large enterprises. Take an example of an enterprise that provides AI-driven CRM solutions. Using LLMs for customer interaction, they can offer smarter chatbots and virtual assistants that understand nuanced customer inquiries, generate personalised responses, and automate repetitive tasks. Running these models on GPUaaS allows enterprises to scale up during high-demand periods, like product launches or seasonal promotions, without worrying about investing in and maintaining expensive GPU infrastructure.

GPUaaS provides the flexibility to access cutting-edge GPUs when needed, allowing enterprises to test, iterate, and deploy their AI-powered solutions quickly and cost-effectively.

2 Startups: Accelerating innovation in foundation models

Startups working on foundation models, which form the core AI architecture for tasks like text generation, translation, and summarisation, are particularly dependent on massive computational power. A startup (at its early stages) or a new player aiming to build the next big breakthrough in AI can leverage GPUaaS to train LLMs at a fraction of the cost. Without GPUaaS, they would need to invest in expensive hardware – a significant hurdle for early-stage companies with limited funding. Additionally, staying competitive requires access to the latest GPUs as soon as they become available, ensuring faster time-to-value and accelerated innovation.

With GPUaaS, these startups can rent high-performance GPUs for as long as needed, dramatically reducing costs while maintaining the ability to scale as their models grow. It allows them to prototype, test, and refine their LLMs faster, getting their products to market faster than ever.

3 Governments: Enhancing policymaking with AI-powered insights

Governments can harness the power of LLMs and GPUaaS to analyse vast amounts of text data, such as legislative documents, research papers, and public sentiment. For example, a government agency looking to implement better healthcare policies could use LLMs to analyse trends in patient data, medical research, and public feedback. By running these models on GPUaaS, they can perform real-time data analysis, which helps draft policies responsive to the population's needs. Additionally, governments can use LLMs to automate administrative tasks, like processing legal documents or public inquiries, making public services more efficient and accessible.

GPUaaS allows governments to quickly deploy powerful AI models without the massive upfront investment in infrastructure. Furthermore, governments are launching multiple AI-driven initiatives, such as hackathons and programs focused on solving societal challenges, all contributing to nation-building. For these large-scale projects, GPUaaS becomes essential, providing the computational power needed to drive innovation and public impact.

How to evaluate the right GPU-as-a-Service provider

Choosing the right GPUaaS provider isn't a one-size-fits-all game—it's about aligning their offerings with your specific needs. Here are the key factors to consider:



Performance and flexibility

- **HARDWARE OPTIONS**

Check what GPUs they offer—are they the latest or outdated models? Your choice should align with the intensity of your AI/ML tasks.

- **SCALABILITY**

Look for providers that allow you to ramp up resources as your project demands grow. A sudden spike in workloads shouldn't slow you down.



Pricing structure

- **FLEXIBLE PRICING**

Understand how they charge—hourly rates, reserved plans, or discounted preemptible instances. Each pricing model caters to a different type of workload.

- **HIDDEN COSTS**

Look beyond the hourly rates—factor in data transfer fees, storage costs, and software licensing requirements.

- **TRIALS AND CREDITS**

Many platforms offer free credits for testing—make the most of these to evaluate performance without committing financially.



Compatibility and ease of use

- **FRAMEWORK SUPPORT**

Ensure the service is compatible with tools like TensorFlow, PyTorch, and other frameworks you rely on.

- **PRE-CONFIGURED ENVIRONMENTS**

Providers with pre-built setups save you time, especially for everyday AI tasks.

- **INTEGRATION**

Can it slot easily into your existing workflows, whether that's through APIs, Docker containers, or CI/CD pipelines?



Cost vs. long-term ROI

- **RESERVED INSTANCES FOR PREDICTABLE NEEDS**

Reserved plans can be more economical if you know you'll need GPUs in the long term.

- **COST TRACKING TOOLS**

Does the provider offer dashboards to help monitor and optimise your spending?



Data and security

- **STORAGE OPTIONS**
Does the provider offer fast-access storage options, like SSDs, and how scalable are they?
- **DATA TRANSFER COSTS**
Moving large datasets can get expensive. Understand their pricing structure for uploads and downloads.
- **SECURITY STANDARDS**
Check for compliance with standards like GDPR, HIPAA, or SOC 2, especially if you're handling sensitive data.
- **DATA SOVEREIGNTY AND COMPLIANCE**
Ensure that the provider allows data to be stored and processed in compliance with local regulations. Auditing capabilities should align with the laws of the land, preventing data from being governed by foreign jurisdictions.



Reliability and support

- **UPTIME GUARANTEES**
Review their Service Level Agreements (SLAs) to understand uptime commitments.
- **SUPPORT AVAILABILITY**
Reliable customer support can be a lifesaver during downtime. Does the provider offer 24X7 assistance or dedicated account managers?
- **FAIL-SAFE MECHANISMS**
Providers with redundant systems can ensure continuity in case of hardware or network failures.



Specialised features

- **AI-SPECIFIC OFFERINGS**
Look for value-added features like pre-trained models, AI accelerators, or ready-to-use SDKs.
- **TEAM COLLABORATION**
Services that enable shared workspaces or team collaboration can streamline your projects.



Environmental sustainability

- **GREEN DATA CENTERS**
Some providers prioritise renewable energy-powered operations, which is a big plus if sustainability is part of your company ethos.
- **ENERGY EFFICIENCY**
GPUs with better performance-per-watt ratios are more eco-friendly and cost-effective.

Tata Communications GPU-as-a-Service powers your business with on-demand, high-performance GPU resources, offering flexibility, reliability, and cost efficiency.

"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

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



Efficient

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

