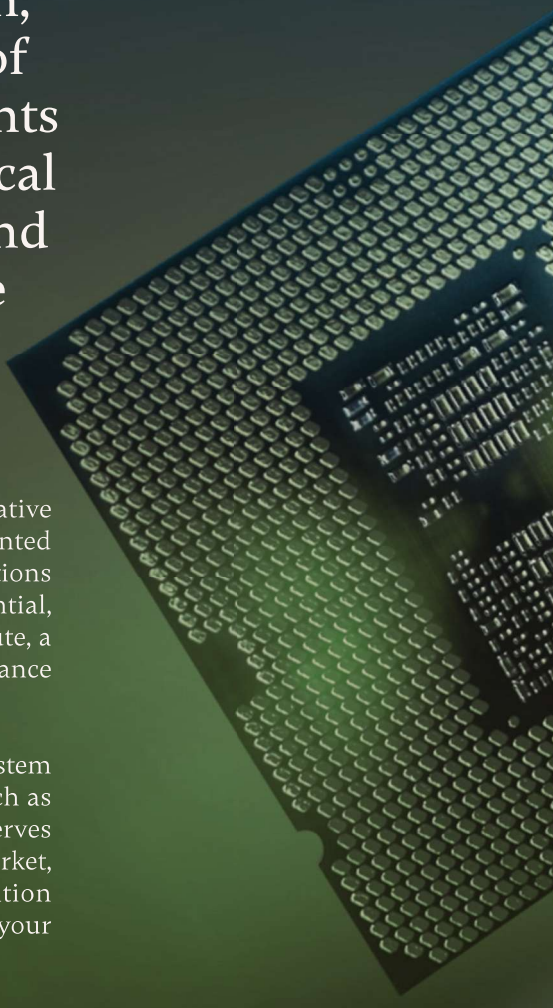
IREN

GPU Compute *Buyer's Guide*

While graphics processing units (GPUs) have long been synonymous with artificial intelligence (AI) / machine learning (ML) innovation, the evolving landscape of GPU infrastructure presents organizations with a critical decision: how to access and optimize GPU compute effectively.



The rapid rise of generative AI as a transformative force in modern business has placed unprecedented demands on computing infrastructure. Organizations across industries are racing to unlock AI's potential, and at the core of this endeavor lies GPU compute, a foundational element enabling the high-performance parallel processing that AI and ML requires.

The decision involves navigating a complex ecosystem of providers and solutions, balancing factors such as performance, cost, and scalability. This guide serves as a resource for navigating the GPU compute market, providing insights into evaluating next generation providers and choosing a partner aligned with your AI requirements.

Contents

01

The evolving GPU compute *ecosystem*

AI's insatiable appetite for computational power has driven the development of three primary models for GPU access.

1.1 Cloud-based GPU access via hyperscalers

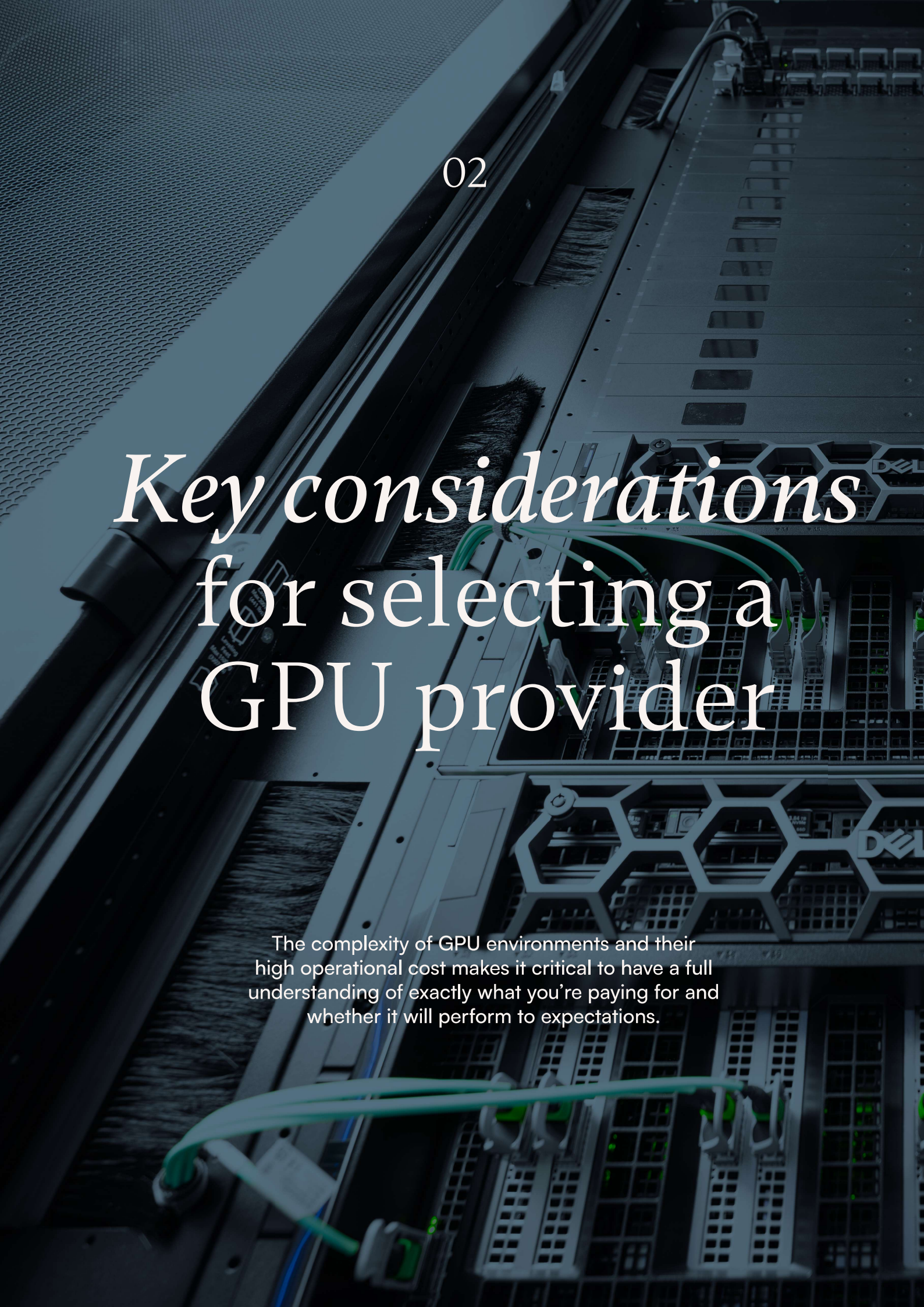
Hyperscale cloud providers offer a familiar approach to accessing GPU compute resources. While this model benefits from integration with existing cloud ecosystems and data, it could come with performance trade-offs due to multi-tenancy, limited customization, data entrapment and opaque pricing structures.

1.2 Specialized providers of cloud-based GPU access

Often referred to by various names such as high performance computing (HPC) cloud providers, neocloud, "GPU as a Service" (GPUaaS), GPU infrastructure providers, or GPU cloud, these service providers deliver tailored solutions designed specifically for AI workloads. As part of their core offering, they often offer specialist technical expertise, transparent pricing, and dedicated support, capabilities that are often considered add-ons or reserved for premium with other models.

1.3 On-premises GPU infrastructure

For organizations with highly specialized needs and established expertise in GPU technology, investing in on-premises infrastructure may be appealing. However, the significant capital expenditure, large power requirements, and operational complexity often make this option viable only for select enterprises.



02

Key considerations for selecting a GPU provider

The complexity of GPU environments and their high operational cost makes it critical to have a full understanding of exactly what you're paying for and whether it will perform to expectations.

GPU brokers often lack full control over their capacity, leading to potential availability issues or performance degradation during periods of peak demand.

2.1 Performance capabilities

Technical excellence is critical for GPU compute environments. Providers should demonstrate expertise in:

Best-in-class hardware

Providers using trusted NVIDIA GPUs such as the NVIDIA H100 and H200 Tensor Core GPUs, and NVIDIA's next-generation Blackwells, deliver exceptional computational power and efficiency. Paired with advanced interconnect technologies like NVIDIA Quantum-2 400Gb/s InfiniBand, these GPUs ensure optimal performance for even the most complex AI/ML tasks.

Infrastructure resilience

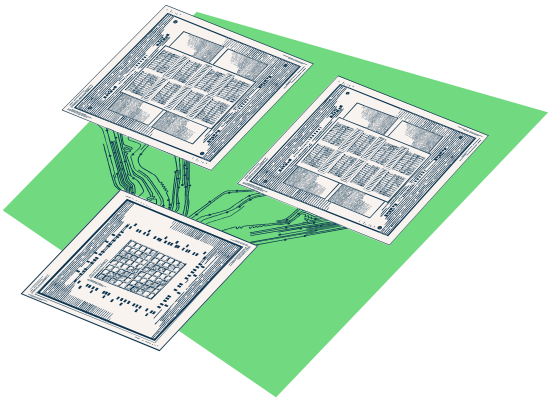
Reliable GPU operations depend on robust architecture, incorporating advanced cooling, high-speed interconnects, and fault-tolerant designs. Providers should ensure proactive hardware health monitoring and real-time diagnostics to identify and mitigate potential issues before they impact performance.

Dynamic scalability and capacity control

GPU brokers often lack full control over their capacity, leading to potential availability issues or performance degradation during periods of peak demand. In contrast, direct GPU infrastructure providers with ownership of their own data centers and full operational control can guarantee consistent access and performance.

Latency and throughput optimization

Evaluate how providers design their systems for low-latency, high-bandwidth workloads, including support for NVIDIA Quantum-2 400Gb/s InfiniBand networking and 3.2TBps NVIDIA NVLink interconnect technologies, which minimize bottlenecks in data movement. Providers that lack advanced interconnect systems may struggle to meet demanding AI/ML performance needs.



GPU provider offerings vary widely. To make an informed decision, organizations must evaluate providers based on three critical dimensions: performance capabilities, cost structure and support ecosystem.

2.2 Cost Structure

The cost of GPU compute is multifaceted and requires a nuanced understanding to optimize budgets and outcomes:

Pricing models and predictability

Assess whether the provider offers clear and flexible pricing models - such as reserved capacity, on-demand usage for both instances and clusters, or spot pricing - to match workload requirements.

Data transfer and storage costs

AI workloads often involve significant data movement and storage demands. Evaluate providers based on their ability to offer low-latency, high bandwidth data pipelines, and cost-efficient storage that scales with your workload. Consider hidden fees for ingress and egress of data.

Operational overhead

Hiring and maintaining a specialized DevOps teams for GPU infrastructure can be costly and challenging, as skilled professionals are in high demand and difficult to find. Look for providers that have optimized these processes, allowing organizations – particularly if you are a startup – to avoid unnecessary headcount expenses. This enables faster speed to market without the need for in-house expertise.

Transparency

The complexity of GPU environments and their high operational costs makes it critical for clients to fully understand what they are paying for and whether it is performing to expectations. Ideally, a service provider would have full access to the data center and be able to provide usage reporting across availability.

Energy optimization and sustainability

Energy is one of the largest cost drivers in GPU operations. Providers leveraging advanced energy efficient technologies, intelligent cooling systems and renewable energy sources not only reduce operating costs but also align with corporate sustainability goals.

Failure to work with an experienced support team at the service provider can lead to failed activities and significant cost overruns.

2.3 Support ecosystem

When selecting a GPU provider, the support ecosystem plays a pivotal role in ensuring the reliability and efficiency of operations. Organizations should prioritize providers that demonstrate the following capabilities:

Deep knowledge

Providers should have a team with deep subject matter knowledge. The rapid evolution of GPU technology means there is only a relatively small pool of specialist technicians. This knowledge can be critical for those organizations whose resources are stretched thin or who have yet to develop their own skill sets in AI and GPUs. Failure to work with an experienced support team at the service provider can lead to failed activities and significant cost overruns.

24/7 Technical support

Around-the-clock access to technical experts ensures quick resolution to any issues that arise, minimizing downtime and maintaining productivity efficiency.

Robust security and compliance

GPU service providers are trusted with both clients' data and the personal information that they hold. This makes it critical that any service provider is able to adhere to the highest level of cyber security and data privacy requirements, with all necessary safeguards.

In-house expertise and end-to-end control

The ability to manage infrastructure internally without relying on third-party vendors for maintenance or repairs enables faster response times and consistent service quality. This eliminates delays associated with logging tickets or waiting for external technicians, providing greater reliability and operational efficiency.

03

The *IREN Cloud™* difference

IREN Cloud™ is passionate about pushing the boundaries of GPU computing. By combining cost efficiency, top-tier performance, and 24/7 support, we're helping organizations unlock the full potential of AI.

Exceptional cost effectiveness

IREN Cloud™ delivers lower costs through vertical integration and strategically located data centers in the United States and Canada, with access to abundant, low-cost renewable energy. By owning and managing our data centers and GPUs directly, we ensure full transparency in pricing and reporting.

Our data centers are powered by 100% renewable energy (whether from clean or renewable energy sources or through the purchase of RECs). Advanced cooling systems further enhance efficiency, achieving extremely low PUE and WUE ratios. With no data ingress/egress charges and opensource storage solutions, we make high-speed data transfer and storage accessible to clients of all sizes.

Owner-operated data center control

At IREN Cloud™, we operate with precision, leveraging our extensive expertise in GPU technology to maintain peak performance. Direct ownership of our data centers provides exceptional operational transparency and reliability. Our 24/7 onsite team is always ready to address issues swiftly, ensuring reliable operations.

Whether clients require pre-configured GPU stacks for rapid deployment or fully customizable baremetal solutions, IREN Cloud™ offers flexible options tailored to diverse workloads. With the latest NVIDIA GPUs and advanced networking technologies like InfiniBand, we deliver the high performance required for today's demanding AI applications.

Comprehensive support

IREN Cloud™ offers a hands-on approach which ensures clients receive personalized guidance at every stage. Our team works closely with customers to scope, design, and optimize GPU environments for peak performance.

With proactive monitoring, real-time reporting, and robust security frameworks, we provide reliable and secure operations.

By partnering with IREN Cloud™, clients gain access to the expertise and tools needed to quickly and cost effectively achieve their AI goals, while benefiting from infrastructure that pushes the boundaries of GPU capabilities.

Get in touch with one of our experts today.

Whether you need more insight, have a few questions, or want to get started, we'd love to chat.

For more information, visit **irencloud.com**