CoreWeave

COMPARISON GUIDE

Choose the Right Cloud for Your AI

Get the evaluation criteria you need to make the right infrastructure choice.

INTRODUCTION

Discover which variables matter most and which provider is right for you.

AI is moving faster than ever, and your cloud strategy needs to be purpose-built to keep up. Choosing the wrong cloud partner doesn't just slow down your team and impact your budget; it puts you at risk for missing product launch windows, losing investor confidence, and giving competitors a head start.

The truth is, there's no one-size-fits-all cloud for AI. The right partner depends on your use case: whether you're training a model, serving inference at scale, or balancing hybrid workloads. Every AI workload has unique requirements, which map to different strengths across providers, from networking and orchestration to cost predictability and speed of innovation.

This guide breaks down the key criteria that matter for AI workloads and compares how hyperscalers, neoclouds, and CoreWeave measure up—so you can move forward with clarity and confidence.

Table of contents

All cloud types are not created equal — 3

Which cloud is best for your AI workloads? — 5

Evaluating clouds for your AI use case — 8

Checklist for your cloud partner — 13

Find your partner in the cloud — 14

All cloud types are not created equal.

Hyperscalers, smaller neoclouds, and CoreWeave all offer their own forms of functional AI infrastructure, but their design choices and trade-offs directly affect your AI workloads. That's why it's so important to evaluate providers, ensuring that you get the right mix of performance, cost, and support for your AI workloads.

Category	CoreWeave AI Cloud	Hyperscalers	Neoclouds
General information	Datacenter footprint: Multiple enterprise-scale facilities; largely U.S.-based with locations across Europe Focus: Large-scale AI and HPC workloads	Datacenter footprint: Global scale, multi-region Focus: General-purpose workloads with a recent growing focus on AI	Datacenter footprint: Limited regional presence Focus: Often hyper-specialized for a niche workload
Tech stack	• Bare-metal infrastructure purpose-built for AI (no hypervisor overhead) • Industry-leading managed environments, like Slurm and Kubernetes orchestration • SHARP-enabled InfiniBand networking and topology-aware scheduling	• Virtualized, hypervisor-based environments • Broad AI product suites, but largely built for general-purpose workloads • Largely ethernet-based networking, limited SHARP support	• Virtualized, hypervisor-based environments • Competitive GPU pricing, but limited orchestration depth • Networking and AI infrastructure capabilities vary significantly
Strengths	• Rated #1 cloud for AI and awarded Platinum-rated provider in ClusterMAX™ by SemiAnalysis • Best-in-class price/performance for AI workloads • Strong SLAs, proactive support, and rapid failover • Fast adoption of next-gen GPUs (e.g., H200, GB200)	• Trusted brand recognition with boards and investors • Large-scale global datacenter presence • Strong baseline security and compliance	• Lower costs and flexible contracts • Fast adoption of new hardware generations • Developer-first user experience
Weaknesses	• Smaller brand footprint compared to hyperscalers • Hyper-optimized to only support AI workloads • Performance-adjusted cost benefits are not as evident at small-scale clusters	• High costs, especially for GPUs • Rigid contracts and high egress fees • Delayed adoption of bleeding-edge GPUs and networking innovations	• Limited enterprise-grade security and compliance • Support often less formalized, with fewer enterprise-grade SLAs • Perceived risk due to shorter track record

Every cloud is designed with a purpose

Hyperscalers were built for general-purpose computing, not the specialized demands of AI. Smaller clouds can help with quick bursts of compute, but they rarely scale to meet enterprise needs. The CoreWeave AI Cloud breaks that mold with built-for-purpose AI infrastructure delivered at scale.

CoreWeave AI Cloud	Hyperscalers	Neoclouds
CoreWeave combines bare-metal performance with enterprise-grade security, delivering hardware and software solutions that set a new standard for AI. The result is industry-leading price per performance and faster time-to-value.	These large, legacy providers offer global reach and strong brand recognition. However, their virtualized environments and rigid contracts often mean higher costs, vendor lock-in, and slower adoption of the latest AI innovations.	New, smaller cloud players deliver affordability and agility, making them initially appealing for some startups and short-term projects. However, their maturity, enterprise-grade compliance, and support can vary widely, as well as their varying degrees of ability to support at scale.



Which cloud is best for your AI workloads?

Here's how the different providers stack up on the factors that matter most for AI workloads, based on the independent ClusterMAX™ benchmark from SemiAnalysis.[1]

To make these categories easier to digest, we've divided them into three main groups: infrastructure operations, performance innovation, and business trust.

Infrastructure operations

This is the backbone of AI cloud performance. From cluster lifecycle management to networking, these factors determine whether workloads run smoothly, scale reliably, and recover quickly when things go wrong.

Category	CoreWeave AI Cloud	Hyperscalers	Neoclouds
Node validation and lifecycle management	Full lifecycle management; includes rigorous cluster burn-in, proactive maintenance, 90-second node failover, and automated health checks	Standard lifecycle management; slower failover and less transparency in burn-in processes	Basic lifecycle management; less automation and validation, which may impact the slower failover
Managed orchestration	Best-in-class managed Slurm and Kubernetes	Mature Kubernetes services	Containerized support; orchestration flexibility varies
Storage	Highest-performing object storage in market; provides other storage offerings (Weka, VAST Data); proven to offer the best performance and cost for AI workloads	Wide range of storage options; solid performance but highly proprietary and most costly	Strong storage performance (e.g., Weka, VAST Data); options vary widely by provider
Networking	Enabled NVIDIA Quantum InfiniBand networking at scale; NVIDIA SHARP-enabled and topology-aware scheduling	Mostly proprietary Ethernet networking with some NVIDIA Quantum InfiniBand networking; limited NVIDIA SHARP support	Some NCCL and limited NVIDIA SHARP support; varies in maturity

Performance innovation

Performance innovation means rethinking what's possible. By combining cutting-edge GPUs, high-speed networking, and intelligent orchestration, teams can achieve breakthroughs in speed and efficiency—accelerating time-to-market for new AI applications.

Category	CoreWeave AI Cloud	Hyperscalers	Neoclouds
Observability	Full cluster and infrastructure transparency with out-of-the-box Grafana dashboards, NCCL profiler plugins, and silent data corruption detection	Limited workload visibility; monitoring is fragmented across services	Basic dashboards; limited depth for AI-specific workloads
Ease of use	Quick time-to-value with pre-configured environments; UI is improving	Polished dashboards; enterprise-focused UX	Developer-first UX; fast spin-up but less enterprise polish
Technical partnership	Deep NVIDIA collaboration (e.g. NCCL tuning) and first to implement SHARP in production; strong open-source integrations and enterprise Slurm support	Strong vendor ties; co-develop features at scale	Partnerships vary; some lack enterprise-grade security
Innovation speed	Rapid adoption of next-gen hardware; first to market with many leading accelerators; fast rollouts of new networking, orchestration, and observability innovations	Bureaucratic, slower adoption of new hardware and networking innovations	Quick to adopt new hardware, but slower to integrate advanced networking and orchestration features

Business trust

Every enterprise is focused on cost, flexibility, security, and support. These are the factors that shape whether a cloud can be trusted for mission-critical AI workloads at scale and whether the economics reliably hold up over time.

Category	CoreWeave AI Cloud	Hyperscalers	Neoclouds
Price snapshot	Best performance-adjusted cost for AI workloads, factoring in metrics like MFU, Goodput, and mean time to failure (MTTF)	Higher performance-adjusted cost; designed for enterprise budgets, sometimes offset by signing discounts	Lower sticker price but higher performance-adjusted cost; often the cheapest for bursts
Contract flexibility	Flexible terms; no egress/ingress fees, long-term pricing stability	Rigid terms; lock-in by design via long-term commitments and high egress fees	Improved flexibility; some charge egres fees, unpredictable ong-term view
Support and SLAs	Expert solutions architecture support during onboarding, POCs, and contracts; high uptime backed by strong SLAs, with proactive monitoring and rapid failover support	Reliable support, but outages and slower recovery are more common; ticket-driven response	Responsive for smaller teams, but with limited SLAs and lower overall uptime
Security	Full enterprise-grade certifications and physical tenant isolation	Full enterprise certifications (SOC2, ISO27001, etc.)	Mixed maturity; some lack certifications, others are struggling to meet requirements
Proven performance	Customers (like OpenAI and IBM) are live with clusters featuring the latest hardware; multiple industry-leading MLPerf benchmarks submissions in the last few years	Quick to announce support for the latest NVIDIA hardware; but customer access can be inconsistent; limited participation in MLPerf benchmarks	Slower to hand over ready-for-production clusters that feature the latest hardware; some participation in recent MLPerf benchmark submissions

Evaluating clouds for your AI use case

When it comes to AI, legacy factors like ecosystem breadth or bundled discounts are no longer measures of success. What matters today is whether your cloud can deliver the right mix of speed, cost, scalability, orchestration, simplicity, and reliability for your specific AI workload.

No two AI projects look alike. Every project demands its own mix of speed, scale, and control. Choosing the right cloud is all about aligning the right infrastructure to your specific needs.

That's why it helps to evaluate clouds through what really matters—and what helps get you to market faster. However, you also need to consider the broader context and company-specific goals of your AI project.

Find your partner in the cloud

In short, the best AI cloud partner is the one that enables you to move fast without compromising on cost efficiency or long-term defensibility.

Let's take a look at some use cases from different industries to see their unique challenges and requirements. Then, we'll compare how traditional hyperscalers, neoclouds, and CoreWeave answer the call.

Questions to consider that help define your unique AI requirements

1. How large are your research and engineering teams?
2. What's your budget and runway?
3. How urgent is your timeline to train, fine-tune, or deploy?
4. How critical is data privacy and security to your IP?
5. Do latency and throughput directly impact your end users or revenue?
6. Do you need a specific GPU spec to run your workload efficiently?
7. How important is predictable, transparent pricing for your CFO and board?

Model training for high-frequency trading

A global investment firm wants to train proprietary AI models to analyze market signals and execute trades faster than competitors. They need flexible, full-stack compute for scalability, efficiency, and securit

Challenges

- Large-scale training runs must be completed within tight market-driven timelines
- GPU capacity shortages or delays can stall time-sensitive launches
- Data privacy and regulatory compliance are non-negotiable

What's required

- Dedicated direct-connect options to integrate with on-prem data centers
- Access to the latest GPU hardware, delivering maximum performance
- Ultra-low latency networking for distributed training across nodes
- Enterprise-grade security and tenant isolation trusted by finance firms

Cloud provider non-negotiables

- SOC2 and ISO 27001 compliance
- Predictable cost curves that a CFO can defend
- SHARP-enabled InfiniBand networking
- Rigorous lifecycle management to prevent downtime

Solutions by provider type

CoreWeave AI Cloud	Hyperscalers	Neoclouds
Enterprise-grade security, predictable pricing, and industry-best networking, though with less brand familiarity compared to hyperscalers	Strong compliance and brand recognition, but expensive and slower networking performance	Lower cost, but limited track record and security maturity for sensitive financial data



INDUSTRY STORY | GENERATIVE AI

Fine-tuning and inference for text-to-image models

Generative AI startups need to fine-tune open-source models on proprietary datasets and and then serve inference for millions of user requests. They need infrastructure that can scale elastically in real time, with transparent tracking of utilization and billing.

Challenges

- Workloads swing between fine-tuning bursts and real-time inference demand
- Budget pressures make efficiency as important as scale
- High-performance storage and observability are necessary to sustain growth

What's required

- High-performance storage with automated archival
- Flexible contracts with no egress surprises
- Rapid scaling supported by robust orchestration (Slurm, Kubernetes) within the same compute cluster
- Observability tools to monitor utilization and costs across workloads

Cloud provider non-negotiables

- Transparent performance/$
- Reliable observability
- Fast onboarding with pre-configured environments
- Storage tuned for checkpointing and large datasets

Solutions by provider type

CoreWeave AI Cloud	Hyperscalers	Neoclouds
Flexible and cost-efficient at scale; fastest storage solution in the market at 7GB/s/GPU with automated archival that saves customers up to 65% on inactive data. SUNK (Slurm on Kubernetes) enables training and inference runs on the same cluster	Stable and enterprise-ready, but with high costs, limited observability, and slower GPU refresh cycles; storage solutions and orchestration tools are less performant	Affordable and fast adopters of new GPUs, but with less reliable SLAs; storage solutions are not as performant or cost-effective, but some don't charge egress fees



INDUSTRY STORY | HEALTHCARE

Training and inference for audio transcription

A healthcare technology company wants to run real-time transcription models to support clinicians during patient visits, ensuring both speed and compliance. Data privacy is of the utmost importance, especially as the move to inference involves sensitive data.

Challenges

- Sensitive patient data must remain secure and compliant
- Partitioning the data between both training and inference
- Real-time transcription requires reliable low-latency inference
- Downtime can directly disrupt patient care

What's required

- Enterprise-grade security, including certifications and network isolation
- SLA-backed uptime with rapid failover (seconds, not hours)
- Deep observability for auditability and inference performance monitoring

Cloud provider non-negotiables

- HIPAA compliance and SOC2/ISO-certified infrastructure
- Proactive SLAs for fast support
- Ability to run training and inference workloads in the same cluster
- NCCL profiler plugins for visibility into training workloads

Solutions by provider type

CoreWeave AI Cloud	Hyperscalers	Neoclouds
Compliance-ready, highly observable, and cost-efficient, though with less global datacenter reach than hyperscalers; provides unique capabilities to partition data between training and inference	Trusted on compliance, but costly and less transparent for workload-level observability; limited ability to trian and serve inference on the same cluster	Lower cost and agile, but lack compliance maturity needed for healthcare workloads; limited ability to trian and serve inference on the same cluster



INDUSTRY STORY | ACADEMIA

Burst compute for pilot project

A university research lab has a 3–month grant to train a mid-sized language model for a publication. They need GPU access quickly, but don't expect to run workloads beyond the grant period.

Challenges

- Strict budget cap makes cost the top priority
- Project length is fixed, with no need for ongoing infrastructure
- Reliability and lifecycle management are less critical since the workload is short-lived

What's required

- Low absolute GPU pricing
- Flexible contracts with no long-term commitment
- Minimal setup time for fast onboarding

Cloud provider non-negotiables

- Simple UI for easy onboarding
- Can support burst workloads without lock-in
- Affordable compute

Solutions by provider type

CoreWeave AI Cloud	Hyperscalers	Neoclouds
Delivers strong performance/$ and orchestration at scale, but may be overkill for small, time-limited projects where lifecycle management and SLAs are a lower priority,	Reliable and compliant, but may be too expensive for budget-limited, short-term work unless there's a partner program for researchers.	Lowest cost and maximum flexibility; ideal for short, bursty projects, but downsides include weaker SLAs and limited enterprise security.



Checklist for choosing your cloud partner

Here's a list of items to help you evaluate if an AI cloud is the right fit for your use case and requirements. Keep in mind: not every item will be critical to your project, and the priority of each will depend on your goals, timeline, and team.

Performance-adjusted cost

- [ ] Benchmarked performance/$ aligned to your workloads
- [ ] Transparent, predictable pricing with no hidden egress fees

Flexibility and scalability

- [ ] Ability to scale up and down quickly
- [ ] Fast access to the latest GPUs without lock-in

Reliability

- [ ] Ability to scale up and down quickly
- [ ] Fast access to the latest GPUs without lock-in
- [ ] Rapid failover and GPU replacement measured in minutes, not hours

Security

- [ ] Enterprise certifications (SOC2, ISO27001)
- [ ] Strong tenant isolation and secure handling of sensitive data

Observability and support

- [ ] Out-of-the-box tools for monitoring utilization, costs, and cluster health
- [ ] Proactive support backed by strong SLAs

Want to go deeper?

For a closer look at how to evaluate AI infrastructure partners, download the ebook: Smarter Compute, Smarter Spend: 5 Critical Considerations When Choosing an AI Infrastructure Partner.

CONCLUSION

Find your partner in the cloud

The cloud you choose for AI workloads is one of the most consequential business decisions you can make. It determines how fast you can launch, how much you'll spend over time, and how confidently you can deliver results to your board, investors, and customers. And perhaps most importantly, how quickly you can get your paradigm-shifting AI innovations to market.

Hyperscalers bring scale and brand recognition, but often at the cost of flexibility and speed. Neoclouds offer affordability and agility, but lack the maturity enterprises need. CoreWeave is purpose-built for AI, delivering industry-leading price/performance, enterprise-grade security, and lifecycle management designed to help you move fast without compromise.

The best cloud is the one that fits your use case and priorities. Your AI roadmap depends on making the right infrastructure choice today. Let's make sure you have the clarity, confidence, and partnerships to deliver and win.

Ready to take the next step?

Here are three ways we can help you make an informed decision:

Personalized TCO for your AI

Explore the technical and operational decisions that influence cost, value, performance, and scale.

Schedule a consultation



Download the ebook: Smarter Compute, Smarter Spend

Explore five critical considerations when choosing an AI infrastructure partner.

Download the Ebook



Read the SemiAnalysis Whitepaper

Dive into the world's first independent GPU cloud benchmark and see how providers measure up.

Get the whitepaper



Sources

[1] The GPU Cloud ClusterMAX™ Rating System | How to Rent GPUs, SemiAnalysis, March 26, 2025.