Inference Cloud

Inference Cloud powered by Qualcomm

Efficient and Scalable AI - No Complex Infrastructure Management Required

Experience seamless one-click AI deployment. Effortlessly use generative AI models to build custom applications and agents using popular frameworks.

Inference Cloud powered by Qualcomm and leveraging the Qualcomm Cloud AI 100 Ultra.

Ease AI Deployments
The web-based platform for deployment, configuration, and monitoring simplifies access to leading AI models as well as pre-built applications and agents. API endpoints enable rapid integration with your existing applications and workflows. You pay only for what you use, with pricing based on tokens that vary for selected AI models.

Run with Confidence
Enjoy high availability and strict data privacy with no storage of model inputs or outputs. Our solution is designed and stress-tested for enterprise environments.

Top Performance, Future-Proofed
Maximize performance and cost efficiency with Qualcomm Cloud AI 100 Ultra inference accelerators, embedded optimization techniques, and state-of-the-art models available in the Qualcomm AI Inference Suite for Cloud.

Customized Options Available
For specialized needs or enhanced scalability, Cirrascale offers the Qualcomm Cloud AI 100 Ultra in a bare-metal solution that enables deep integration of custom DevOps workforces with your inference requirements. We work with you to develop the solution you need.

Inference Cloud Powered by Qualcomm

The Qualcomm AI Inference Suite with the Qualcomm Cloud AI 100 Ultra

- Play with generative AI applications with easy-to-use endpoints.
- Build your applications with Qualcomm Cloud AI tutorials and documentation.
- Start with free access to test, then seamlessly transition to production using the Inference Cloud powered by Qualcomm.

Ready-To-Use Applications and Agents

- Chatbot

- Summarization

- AI agents
- Image generation
- Retrieval-augmented text generation (RAG)
- Real-time translation
- Code development
- Real-time transcription
- Your next use case

NVIDIA Inception

We Help NVIDIA® Inception Members

Cirrascale Cloud Services helps NVIDIA Inception members get the most out of their program benefits. Whether you're a community member or a premier member, we can help you.

What is the NVIDIA Inception Program?
NVIDIA Inception is designed to help startups build their products and grow faster. Members get exclusive technology discounts, free technical training, marketing opportunities, and increased exposure to the venture capital community. The program is free and available for tech startups of all stages

NVIDIA works closely with members to provide the best technical tools, latest resources, and opportunities to connect with investors. As your startup matures, your program benefits also evolve to further your growth. Premier members receive increased NVIDIA marketing support, access to Premier-only member events, and a dedicated NVIDIA relationship manager.

What Can Cirrascale Do To Help?
Cirrascale has been helping NVIDIA Inception program members such as AssemblyAI, MosaicML, Metaphor, and more, evolve faster by providing them with unique benefits. By partnering with Cirrascale, approved Inception members can purchase hardware at preferred pricing levels, receive discounts on cloud and managed services provided by Cirrascale, and partner with us to receive referral discounts for other startups you refer.

Cirrascale Cloud Services works closely with NVIDIA and your assigned dedicated relationship manager to provide you with the best experience as an Inception member.

NVIDIA® Inception Benefits with Cirrascale

Full Cloud Services
We offer fully-managed GPU clusters at a fraction of the cost of traditional cloud service providers. These bare-metal servers are completely dedicated to you with no contention and no performance issues due to virtualization overhead.

Our flat-rate, no surprises billing model means we can provide you with a price that is up to 30% lower than the other cloud service providers. We also don't nickel-and-dime you by charging to get your data into or out of our cloud. Instead, we charge no ingress or egress fees so you never receive a supplemental bill.

Hardware Rebates
We're one of NVIDIA's earliest NPN CSP Elite Compute partner's which allows us to sell hardware to you while providing you with the maximum program member benefits for NVIDIA GPU discounts on eligible accelerator cards.

These discounts are provided to you up-front and show on your invoice so that you receive the benefit immediately. We then work with NVIDIA to receive the rebates on the back end. We can work with you to ensure you get the absolute most out of your NVIDIA Inception membership.

Managed Services
Our managed services offering is designed to give you the edge without worrying about server management or maintenance. You can focus on what matters: solving problems, speeding up your workflow and taking advantage of cutting-edge computing power.

Once you purchase your hardware, Cirrascale can provide full managed services with specialized discounts so that you don't have to deal with long-term data center contracts or managing your hardware. We even include break/fix and RMA service to our customers. It's truly worry free with Zero DevOps.

Referral Discounts
Inception members are eligible to take part in our referral program enabling them to save even more on their monthly service costs. Once part of our referral program, Inception members can refer other companies and receive discounts of up to 10% off their monthly service costs.

Leasing Options
Cirrascale Cloud Services works with several financing companies to help provide leasing options for those Inception program members that need assistance with long-term solutions. We can work with your team to help manage the overall process, making it easy on you.

Co-Marketing Opportunities
If desired, we can help to provide additional marketing opportunities for Inception program members that work directly with Cirrascale. We promote your company within our own blog posts, social media postings, case studies, customer references, presentations and more.

Multi-GPU Compute

Unleash Your Deep Learning Frameworks

Whether you're just starting your GPU accelerated application development or ready to take your production and training applications to the next level, we provide you with the features you need in a cloud hosted environment that's unmatched.

Largest Variety of GPUs and Other Accelerators
Other providers make you train your models on two generation old GPUs. Not us, we use the latest NVIDIA GPU accelerators, as well as other accelerator partner architecture, to help you advance the AI revolution and enable HPC breakthroughs. Giving you accelerator options like these is what makes us different.

Physical, Dedicated GPUs and Servers
We don’t virtualize GPUs and give you a sliver of the power, you get all of them and all of the resources that are included in our bare metal servers. No one shares these servers with you. Rest assured that when you use our service, you're getting a dedicated resource for you, and you alone.

Perfect for AI Applications
Building the ultimate generative AI, speech recognition or natural language processing application takes different tools along the way. Our variety of configuration offerings are set up to allow you to start small and scale up.

Real Transparent Pricing with No Surprises
Others may claim to have transparent pricing, but once you start to look at the bottom line it can be a big surprise. From our start, we have always delivered a no surprises billing model with flat rates, so what you see as our price is what you'll pay.

Always the Latest Technology
Our offerings are constantly being updated to offer our customers the highest level of service and technology available. Discover multi-GPU cloud servers with NVMe storage, the latest CPUs, and GPUs at the same price as what our competition offers older, more outdated technology. Why pay more?

Need Something Different? Customize It
Can't find a solution that meets your needs? Over the years, we've gotten a reputation for helping companies find the right solution that fits what they need. We work with start-ups and enterprise customers to build solutions that tackle today’s toughest challenges. We're always happy to discuss something special.

Cloud Storage Solutions

Cloud Storage Solutions for Generative AI, Computer Vision, and NLP Workflows

For unmatched performance in feeding your deep learning training models, you can't rely on the unknown storage performance that you're handed at the other cloud service providers. Instead, you can count on the Cirrascale Cloud Services platform to deliver a wide variety of storage options that will meet your needs.

Local NVMe Storage
All of our systems support high-speed NVMe drives for your local storage. By supporting NVMe drives in our cloud servers, customers gain the true advantages of state-of-the-art technology.

With reduced I/O overhead and various performance improvements in comparison to previous logical-device interfaces, including multiple, long command queues, and reduced latency, NVMe drives are superior for Generative AI, Autonomous Vehicle, Computer Vision, and Natural Language Processing (NLP) workflows.

Additionally, local bulk tier storage is also available to each server if customer need access to low-cost options.

NVMe Hot-Tier Storage Solutions
Cirrascale uses NVMe flash-based solutions to deliver the overall best storage experience to its customers. Whether it's application workflows for generative AI, high-performance computing, medical imaging, financial risk simulations, or natural language processing, we have the right solutions to meet your needs.

Cirrascale has selected the WEKA Data Platform to utilize within its cloud for removing storage bottlenecks faced by customers who use inference and training datasets consisting of millions of files. Both data and metadata are distributed across the entire storage infrastructure to ensure massively parallel access to NVMe drives. The WEKA® Data Platform is certified as a high-performance data store solution for NVIDIA Cloud Partners like Cirrascale and has demonstrated the ability to easily saturate any accelerator cluster and deliver ultra-fast performance per node across an NVIDIA Quantum InfiniBand network.

Object Storage
Cost-effective, S3-compatible Object Storage that solves your biggest storage challenges while boosting interoperability, data durability, and operational efficiency making it possible to store practically limitless amounts of data, simply and cost effectively.

Cirrascale offers object storage solutions that are perfect for the retention of massive amounts of unstructured data. Unlike other cloud solutions, Cirrascale provides the ability to store data in our cloud and not get hit with variable ingress/egress fees when trying to move data either from your office or from other cloud providers.

When part of a multi-tiered storage solution, and linked with a Cirrascale hot-tier storage offering, customers are able to seamlessly move data to be worked on between storage tiers. Overall, customers can reduce latency, improve throughput, and eliminate access charges.

Networking

Flat, Stable Network Bandwidth

Cirrascale utilizes the next-generation of networking for our high-performance, scalable, and secure AI-driven data centers. Our cloud servers come standard with bonded ethernet network connectivity to each physical node within our cloud. Unlike other providers, we provide the same network bandwidth speeds to our high-speed or object storage solutions too, so you don't experience any issues when accessing your data. For 90% of our customers, these solutions work incredibly well for their AI heavy workflows; however, if higher bandwidth and decreased latency are needed, we have high-speed solutions to meet those needs.

Cirrascale Networking Benefits

Rail-Optimized NVIDIA Quantum InfiniBand Networking

Cirrascale Cloud Services provides its customers with the ability to go beyond and experience connectivity up to 3200Gb per server with NVIDIA Quantum InfiniBand networking.

Cirrascale utilizes rail-optimized NVIDIA Quantum InfiniBand networking that is designed to enhance AI workload performance by providing a high-bandwidth, low-latency interconnect that is particularly well-suited for dense, multi-node cloud server configurations. In this setup, the "rail" refers to optimized data pathways that connect servers within the same or adjacent racks, making the most of NVIDIA Quantum InfiniBand's ability to deliver exceptionally fast data transfer speeds while minimizing latency. By focusing on rack-level or rail-specific networking, these systems reduce the distance data needs to travel and decrease the chances of congestion in the network, enabling smoother and faster data processing.

This optimization is especially valuable for customers that have compute-intensive workloads, such as those in high-performance computing (HPC), artificial intelligence, and deep learning environments, where rapid communication between nodes is critical. Rail-Optimized NVIDIA Quantum InfiniBand networks support direct communication channels between servers, allowing data centers to achieve consistent performance and scalability as they grow. Our engineers are skilled in optimizing our networks so you get the absolute best performance.

Fast Connectivity to Storage Resources

Cirrascale provides options for its customers to connect their dedicated multi-accelerator cloud servers to the industry's fastest, ultra high speed NVMe Flash storage from WEKA for the fastest available feeding of large deep learning and AI training, fine-tuning, and inference workloads requiring exceptionally high IOPS per server and centralized access. The WEKA Data Platform is certified as a high-performance data store solution for NVIDIA Cloud Partners like Cirrascale.

Private Networking

Private Networking enables our multi-accelerator cloud servers to communicate with other cloud servers in the same data center. Systems can be clustered together for replication, larger job

analysis, or more. Additionally, private networked servers can connect to the same storage resources making the delivery and sharing of data across nodes easier than ever.

Guaranteed Uptime
Cirrascale Cloud Services has built its entire high-availability network around top-of-the-line providers in our state-of-the-art West, Central, and East data centers. We provide an uptime SLA around network, power and server availability. If we fail to hit our target, just call us and we'll credit you based on the amount of time your servers were unavailable.