Generative AI

Generative AI

The Cirrascale AI Innovation Cloud is specifically designed to handle Generative AI models so your customers can create breathtaking content.

Purpose Built for Generative AI Innovation
Cirrascale is at the forefront of cloud innovation, providing a platform purpose-built for generative AI. Designed to meet the unique demands of generative AI models, Cirrascale enables organizations to confidently scale and deploy their AI applications with unmatched efficiency. Our infrastructure is tailored to support every phase of your AI journey, from pre-training and training to real-time inference, ensuring that your generative AI solutions are always production ready.

Optimized Performance for Real-Time AI Modes
Generative AI models require significant computing power and specialized hardware to function effectively. The models used for Generative AI primarily come from foundational models, designed for more general use cases. The most recognized Generative AI models today are Large Language Models (LLMs) that generate text, but models are also designed for generating images and code. These models are pre-trained with extensive data sets to cover more general use cases. These models can however be enhanced for a specific enterprise, industry or use cases by retraining with augmented or synthetic data, tuning with labeled examples to update weights or using retrieval-augmented-generation (RAG) in LLMs to utilize authoritative data to improve responses.

Managed Services and Cost-Effective Scalability
At Cirrascale we understand that managing complex AI infrastructure can be a challenge. That’s why we offer managed services that take the burden off your team, allowing you to focus on innovating with generative AI. Our hands-free approach includes setting up scheduling services, managing user access, and ensuring your infrastructure is secure and optimized. With transparent, flat-rate billing, you can scale your AI efforts without worrying about unexpected costs. Trust Cirrascale to provide the secure, scalable platform your generative AI initiatives need to thrive.. Cirrascale’s AI Innovation Cloud platform is a much better choice for those companies needing cloud-based, dedicated servers for Generative AI. You'll get charged one flat rate per month and that's it. No ingress or egress fees, no overage charges, no surprises... and we provide a level of support that goes beyond other providers.

Autonomous Systems and Robotics

Autonomous Systems and Robotics

Cirrascale can help you identify and eliminate aspects of your autonomous system or robotics workflows that hinder your ability to quickly process collected data for ground-truth generation and system performance improvement.

Autonomous system and robotics developers use various workflows to create their perception, planning, and control models. These workflows typically involve collecting, processing, and storing massive amounts of data so that data scientists can build datasets for training and validating their models. As these workflows are utilized, customers encounter bottlenecks that prevent them from using their resources. These bottlenecks usually manifest as slow data processing, leading to backlogs that force teams to select subsets of data for processing, potentially missing essential edge cases that help improve system performance.

Larger cloud providers excel at one thing: hyperscale. However, they struggle to handle the large volumes of data generated by autonomous systems and robotics. Multiple cameras, LiDAR's, Radars, bus-collected data, and other sensor outputs create massive data challenges if the necessary compute acceleration devices cannot manage the volume of data. A more custom-tailored approach is required to eliminate bottlenecks and achieve the full performance needed to process data in a timely manner. This is where Cirrascale comes in. With in-house expertise in data collection and analysis for autonomous systems and robotics, we understand where the risks lie and how to mitigate them. Additionally, our cloud is structured differently, so you never experience hidden fees. Large providers often squeeze you with extra fees and charges to get your data out of their storage clouds, leading to a sense of lock-in because the cost of extracting data becomes prohibitively expensive. With us, you won't face that issue, as we don't charge any ingress or egress fees.

Computer Vision

Computer Vision

Image and Object Classification, Detection, and Tracking

Computers can be trained to accurately identify objects in images and videos faster than ever before, but they need massive amounts of data to do so. Even for pre-trained models, re-training or tuning is essential to ensure high accuracy and effectiveness of a model. What was once possible with on-premises workstations or servers, using initial models with limited pattern matching capabilities, has now shifted to deep learning and machine learning models that add AI capabilities, providing a deeper understanding of visual content. As the needs for acceleration of these models increases, companies often attempt to scale their operations for computer vision needs on cloud services such as AWS or GCP, only to realize that prolonged usage of these services can be prohibitively expensive.

With powerful models at their disposal, specific market segments- such as autonomous and robotic systems, healthcare, manufacturing, agriculture and security applications- can take data from sensors and other imaging devices and analyze it in a timely manner to meet the needs of specific workflows.

Cirrascale has been leveraging GPUs for computer vision use cases, enabling groundbreaking improvements across image and object classification, detection and tracking applications. Today, with a wider variety of acceleration options available for computer vision, Cirrascale's AI Innovation Cloud platform is a superior choice for companies needing cloud-based, dedicated servers. You'll be charged one flat rate per month -no ingress or egress fees, no overage charges, no surprises. Plus, we provide a level of support that goes beyond other providers.

High Performance Computing

High Performance Computing

High Performance Computing (HPC) is a cornerstone in driving forward the frontiers of scientific discovery and innovation

From climate modeling to advanced simulations in aerospace and biomedical research, Cirrascale empowers researchers by seamlessly integrating traditional computational methods with the latest advancements in AI, machine learning, big data analytics, and edge computing. Together, we're unlocking the secrets of our world, one calculation at a time.

Empowering Demanding Workloads
Whether it's complex simulations, the fusion of HPC and AI, or data-intensive visualizations, our HPC solutions are designed to accelerate breakthroughs across diverse industries—from cutting-edge scientific research to intricate financial modeling. Cirrascale's cloud-based HPC platform, powered by top-tier GPUs, enables professionals to achieve their most ambitious goals and transform their visions into reality.

With a focus on scalability, reliability, and unmatched performance, Cirrascale's infrastructure adapts to the evolving needs of your projects. Our dedicated support team and extensive resources ensure that you have the tools and expertise to tackle the most challenging computational tasks. By offering flexible configurations and on-demand resources, we help you optimize costs while maximizing efficiency, enabling you to focus on what truly matters—innovating and pushing the boundaries of what's possible.

Today, with a wider variety of acceleration options available for high-performance computing, Cirrascale's AI Innovation Cloud platform is a superior choice for companies needing cloud-based, dedicated servers. You'll be charged one flat rate per month -no ingress or egress fees, no overage charges, no surprises. Plus, we provide a level of support that goes beyond other providers.

Audio Processing

Audio Processing

NLP / Audio Processing

Audio Processing has been used for years to detect, decipher, and understand complex language. Unlike Large Language Models (LLMs), which focus on text input, NLP / Audio Processing applications are literally everywhere in our daily lives, as most communication occurs through spoken language. Whether it’s a simple audio-based search, a digital assistant request, language translation, or more, a vast amount of deep learning models power these applications. The resources required to run these models continue to grow, demanding large-scale, multi-GPU servers for processing. Companies often attempt to scale their operations for natural language and audio processing on cloud services such as AWS or GCP, only to find that prolonged usage can be prohibitively expensive.

Cirrascale Cloud Services leverages NVIDIA GPUs for Audio Processing to make groundbreaking improvements across various language and audio applications. Given the immense GPU compute power needed, the Cirrascale Cloud Services innovation cloud is a superior choice for companies needing dedicated, long-term cloud-based servers. You'll be charged one flat rate per month-no ingress or egress fees, no overage charges, no surprises. Plus, we provide a level of support that goes beyond other providers.