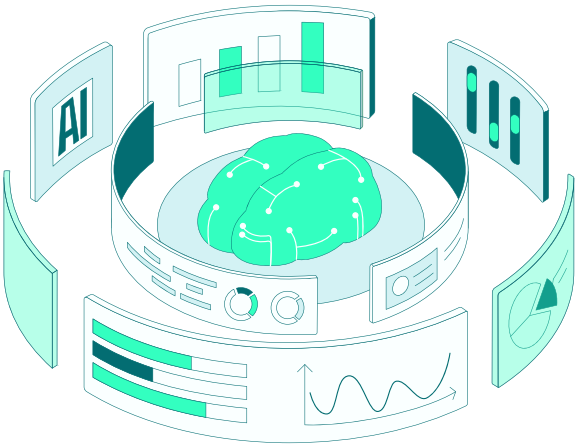
Cirrascale Cloud Services

Cirrascale Inference Platform

A New Era of Smarter Inferencing for Enterprise

Analyze models, deploy on optimal accelerators, and balance workloads across regions automatically



Go Beyond Traditional Hyperscalers

Your AI journey may have begun with hyperscaler AI services to evaluate models, test specific use-cases and start initial deployments. As you move to full-scale deployment, you need predictable costs, resource control, and experienced AI services tailored to your requirements. Cirrascale's Inference Platform enables you to scale up beyond the limitations of hyperscalers with key enterprise-ready features.

Burns Through Workloads, Not Your Budget

The Cirrascale Inference Platform is designed to allow for easier deployment, use, and scale-up of Generative AI models, including Large Language Models (LLM's), image, audio, and video models. The platform provisions AI model pipelines that can be used with existing enterprise, SaaS or proprietary workflows that may be run on-premises or with a hyperscaler. Additionally, it dynamically balances workloads across regions, helping to smooth out peak demands, enhance operational efficiencies, and reduce costs.

Agile and Scaleable Performance

A key advantage of the platform is its ability to deploy your own tuned or customized models based on token performance needs and user demand. Cirrascale's Inference Platform intelligently selects the most ideal AI accelerators to balance cost and performance for any given AI inference workflow. Our model pipelines dynamically scale the necessary resources to maintain required token throughput and time-to-answer, ensuring seamless support for both low-latency, real-time applications and batch processing workloads.

Platform Features

- Premiering with NVIDIA Blackwell B200, RTX PRO 6000 Blackwell Server Edition, and Hopper H100 and H200 GPUs.
- Dynamically balances workloads across regions so you don't have to.
- Serverless deployments that provide an instant AI Model Pipeline.
- Pre-compiled foundational models optimized for underlying accelerators (such as Llama 3.3 Instruct and Deepseek R1).
- Web console-based deployment and configuration – no SSH access required, includes monitoring tools for throughput and performance analysis.
- API Access to the AI model utilizing OpenAI API's and other common interfaces via public access endpoints.
- Supports fine-tuned models, allowing for custom weights or parameters to be applied on an AI model pipeline.
- Priced by token volume for the given AI Model used in the pipeline.

Optional Features

- Customized weights, tuning, and Retrieval-Augmented-Generation (RAG) capabilities for an AI Model Pipeline.
- Interconnect of the AI model pipeline to an on-premises or hyperscaler environment to run directly aside existing workflows.
- Proprietary model support.
- Reserved capacity for time-based burst requirements.

Cirrascale Cloud Services 5775 Kearny Villa Road, San Diego, CA 92123 USA Phone 888-942-3800 Web www.cirrascale.com

Cirrascale

CCSIPDS - REV B - 3/2025