# Qualcomm Cloud AI

## Qualcomm Cloud AI 100

Optimized inference for leading AI models, up to 5x performance of competing solutions.

Qualcomm Cloud AI, as part of the Cirrascale AI Innovation Cloud, utilizing the Qualcomm Cloud AI 100 Ultra, delivers the performance and power efficiency necessary to deploy and accelerate AI inference at scale.

Now providing bare-metal access and the new Inference Cloud powered by Qualcomm.

### Qualcomm Cloud AI 100 Ultra Solutions

Providing industry leading performance-per-TCO$ spanning GenAI, including Large Language Models, as well as Natural Language Processing and Computer Vision. Unlocking new possibilities for AI applications in the cloud for model developers, AI inference solution providers, and enterprises.

### Purpose-Built Inferencing

Generative AI models being developed today, require robust, high-performance acceleration during development and training. However, when deploying a pre-built model for a service or enterprise offering, the main requirement is cost-effective inference, avoiding the high costs from devices that are optimized for training.

The Qualcomm Cloud AI Platform includes devices like the Cloud AI 100 Ultra, purpose-built for Generative AI. It accelerates inference for Large Language Models (LLMs), Natural Language Processing (NLP), and Computer Vision (CV).

### Inference Cloud powered by Qualcomm

Accelerate generative AI development with ready-to-use applications and agents that enable production deployment with the Qualcomm AI Inference Suite.

### 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.

# AMD Instinct Series Cloud

## AMD Instinct™ Series

AMD Instinct Series accelerators, delivered in the Cirrascale AI Innovation Cloud, enable performance leadership that is uniquely well-suited to power even the most demanding AI and HPC workloads.

We've partnered with AMD to offer their AMD Instinct Series accelerators in the cloud for customers to test, utilize and fully deploy. These accelerators provide exceptional compute performance, large memory density, high bandwidth memory, and support for specialized data formats. AMD Instinct accelerators are built on AMD CDNA™ architecture, which features Matrix Core Technologies and supports a broad range of precision capabilities.

### AMD Benefits

### AMD ROCm Software

AMD ROCm™ is an open software stack including drivers, development tools, and APIs that enable GPU programming from low-level kernel to end-user applications. ROCm is optimized for Generative AI and HPC applications, and is easy to migrate existing code into.

ROCm enables AI and HPC application development across a broad range of demanding workloads.

### AMD Infinity Fabric Technology

Cirrascale-hosted AMD Instinct series accelerators with advanced peer- to-peer I/O connectivity through a maximum of eight AMD Infinity Fabric™ links deliver up to 800 GB/s I/O bandwidth performance. With a cache-coherent solution using optimized AMD EPYC™ CPUs and Instinct accelerators, Infinity Fabric unlocks the promise of unified computing, enabling a quick and simple on-ramp for CPU code to accelerated platforms.

### Discover the Benefits of AMD Instinct hosted by Cirracale

### Optimal Performance at the Right Price

- High performance with larger memory than other acceleration offerings
- Optimized for leading Generative AI models, including LLMs

### Ease of Use

- AMD drivers pre-installed and configured by Cirrascale
- ROCm software included for easy access to frameworks and tools
- Hugging Face transformers supported out of the box
- Highly scalable for the most demanding training, tuning and inference workloads

### Simple & Secure Cloud Operations

- Simple onboarding – No DevOps required
- SDKs, storage and network are configured and ready to go

## AMD Products

### AMD Instinct MI300X

AMD Instinct™ MI300X accelerators are uniquely well-suited to power even the most demanding AI and HPC workloads, offering exceptional compute performance, large memory density, high bandwidth memory, and support for specialized data formats.

AMD Instinct MI300X accelerators are built on AMD CDNA™ 3 architecture, which offers Matrix Core Technologies and support for a broad range of precision capabilities—from the highly efficient INT8 and FP8 (including sparsity support for AI) to the most demanding FP64 for HPC.

### AMD Instinct MI250

The AMD Instinct MI250 accelerator brings customers the compute engine selected for the first U.S. Exascale supercomputer.

AMD Instinct MI250 accelerators are built on AMD CDNA™ architecture, which offers Matrix Core Technologies and support for a broad range of precision capabilities—from the highly efficient INT8 and FP8 to the most demanding FP64 for HPC.

# Cerebras Cloud

## The Cerebras AI Model Studio

### Train GPT-Style Models 8x Faster Than Traditional Clouds at a Fraction of the Cost

Hosted on the Cerebras Cloud @ Cirrascale, the Cerebras AI Model Studio is a purpose-built platform, optimized for training and fine-tuning large language models on dedicated clusters. It provides deterministic performance, requires no distributed computing headaches, and is push-button simple to start.

### The Cerebras AI Model Studio

The Cerebras AI Model Studio is a simple pay by the model computing service powered by dedicated clusters of Cerebras CS-2's and hosted by Cirrascale Cloud Services. It is a purpose-built platform, optimized for training large language models on dedicated clusters of millions of cores. It provides deterministic performance, requires no distributed computing headaches, and is push-button simple to start.

### The Problem

Training large Transformer models such as GPT and T5 on traditional cloud platforms can be painful, expensive, and time consuming. Gaining access to large instances typically offered in the cloud can often takes weeks just to get access. Networking, storage, and compute can cost extra, and setting up the environment is no joke. Models with tens of billions of parameters end up taking weeks to get going and months to train.

If you want to train in less time, you can attempt to reserve additional instances – but unpredictable inter-instance latency, makes distributing AI work difficult, and achieving high performance across multiple instances challenging .

### Our Solution

The Cerebras AI Model Studio makes training large Transformer models fast, easy, and affordable. With Cerebras, you have millions of cores, predictable performance, no parallel distribution headaches – all of this enables you to quickly and easily run existing models on your data or to build new models from scratch optimized for your business.

A dedicated cloud-based cluster powered by Cerebras CS-2 systems with millions of AI cores for large language models and generative AI:

- Train 1-175 billion parameter models quickly and easily
- No parallel distribution pain: single-keystroke scaling over millions of cores
- Zero DevOps or firewall pain: simply SSH in and go
- Push-button performance: models in standard PyTorch or TensorFlow
- Flexibility: pre-train or fine-tune models with your data
- Train in a known amount of time, for a fixed fee

### Discover the Benefits of the Cerebras AI Model Studio

**Train Large Models in Less Time**

- Train 1–175 billion parameter models 8x faster than the largest publicly available AWS GPU instance
- Enable higher performing models with our longer sequence lengths (up to 50,000!)

**Ease of Use**

- Easy access: simply SSH in and go
- Simple programming: range of large language models in standard PyTorch and TensorFlow
- Push-button performance: the power of millions of AI cores dedicated to your work with no distributed programming required
- Even the largest GPT models run without a single minute spent on parallelizing work

**Price**

- Models trained at half the price of AWS
- Predictable fixed price cost for production model training

**Flexibility**

- Train your models from scratch or fine-tune open-source models with your data

**Ownership**

- Dependency free – keep the trained weights for the models you build

**Simple & Secure Cloud Operations**

- Simple onboarding: no DevOps required
- Software environment, libraries, secure storage, networking configured and ready to go

**Should You Fine-Tune or Train from Scratch?**

Ultimately the decision to fine-tune a pre-trained model or to train a model from scratch depends on various factors such as the size of the dataset, the similarity between the pre-trained model's task and the new task, the availability of necessary computational resources (we got you covered there), and overall time constraints. To make it as easy as possible, Cerebras developed the flow chart to the right to help guide you.

If you have a small dataset, fine-tuning a pre-trained model can be a good option. In fine-tuning, you take a pre-trained model and retrain it on a new dataset specific to your task. This approach can save you time since the pre-trained model has already learned general features from a large dataset. Fine-tuning can also help to avoid overfitting on small datasets.

However, if you have a large dataset, training a model from scratch may be a better option. Training a model from scratch allows you to have more control over the architecture, hyperparameters, and optimization strategy, which can lead to better performance on the specific task. Additionally, if the pre-trained model's task is significantly different from your task, fine-tuning may not be as effective.

### Fine-Tuning

### Standard Offering

The Fine-Tuning Standard Offering is a self-service process, similar to the Training from Scratch Standard Offering. Pricing is based per 1,000 tokens so there's no surprises. Minimum spend is $10,000.

### White-Glove Support with Cerebras Experts

With White-Glove Support, Cerebras thought leaders will fine-tune a model on the Cerebras Wafer-Scale Cluster on your behalf and will deliver you trained weights. Contact us directly for pricing.

### Train From Scratch

Train your own state-of-the-art GPT model for your application on your data. The process is simple:

- Pick a large model from the list below (or contact us for custom projects)
- See the price, time to train: no surprises
    - SSH in and get going
    - Enjoy secure, dedicated access to programming environment for the training period
    - Cerebras model implementation for the chosen model appear
    - Systems, code examples, documentation are at your fingertips
    - Scripts allow the user to vary training parameters, e.g. batch, learning rate, training steps, checkpointing frequency
    - Use Cerebras-curated Pile dataset to train upon, if desired
- Save and export trained weights and training log data from your work to use as you see fit

### Additional Services Available

Cirrascale and Cerebras provides additional services as needed, such as:

- Bigger dedicated clusters to are available to reduce time to accuracy and work on larger models
- Additional cluster time for hyperparameter tuning, pre-production training runs, post-production continuous pre-training or fine-tuning is available by the hour
- CPU hours from Cirrascale for dataset preparation
- CPU or GPU support from Cirrascale for production model inference

# NVIDIA GPU Cloud

## NVIDIA GPU Cloud

Unmatched End-to-End Accelerated Computing Platform

NVIDIA AI acceleration devices, hosted by Cirrascale, provide multiple GPUs with extremely fast interconnections and a fully accelerated software stack, creating the most optimal platform for HPC and AI training, tuning and inference.

### Expand Horizons with NVIDIA in the Cloud

### Purpose-Built for AI and HPC

AI, complex simulations, and massive datasets require multiple GPUs with extremely fast interconnections and a fully accelerated software stack. The NVIDIA HGX™ AI supercomputing platform brings together the full power of NVIDIA GPUs, NVIDIA NVLink™, NVIDIA networking, and fully optimized AI and high-performance computing (HPC) software stacks to provide the highest application performance and drive the fastest time to insights.

This fully connected topology from NVSwitch enables any GPU to talk to any other GPU concurrently. Notably, this communication runs at the NVLink bidirectional speed of 900 gigabytes per second (GB/s), which is more than 14x the bandwidth of the current PCIe Gen4 x16 bus.

### Accelerating HGX With NVIDIA Networking

The data center is the new unit of computing, and networking plays an integral role in scaling application performance across it. Paired with NVIDIA Quantum InfiniBand, HGX delivers world-class performance and efficiency, which ensures the full utilization of computing resources.

At Cirrascale, our NVIDIA HGX B200 and H200 clusters are built using NVIDIA InfiniBand NDR networking so you receive the most performant cluster for your training and inference needs. Our infrastructure is setup to be optimized for your specific configuration to make sure your training experiments maximize your compute per dollar.

### Discover the Benefits of NVIDIA AI Hosted by Cirrascale

### Flexibility for Training, Fine Tuning, and Inference

- Supports model development and tuning
- Provides tuned performance for all leading AI models
- The most recognized platform for developing and deploying AI and HPC solutions

### Ease of Use

- Direct support for leading frameworks
- The reference platform for model libraries such as Hugging Face, enabling easy model deployment and use

**Extensive Software Support**

- Native CUDA support for the most extensive compatibility with existing compute GPU software, frameworks and tools
- Predictable pricing model deployed on Cirrascale

**Simple & Secure Cloud Operations**

- Simple onboarding – No DevOps required
- SDKs, storage and network pre-configured and ready to go

## Popular NVIDIA Offerings on the Cirrascale AI Innovation Cloud

### NVIDIA HGX B200: The New Era of Accelerated Computing is Here

The NVIDIA Blackwell architecture introduces groundbreaking advancements for generative AI and accelerated computing. The incorporation of the second generation Transformer Engine, alongside the faster and wider NVIDIA NVLink interconnect, propels the data center into a new era, with orders of magnitude more performance compared to the previous architecture generation.

Cirrascale offers the HGX B200 in its AI Innovation Cloud as an 8-GPU configuration giving you full GPU-to-GPU bandwidth through NVIDIA NVLink™ Switch. As a premier accelerated scaleup x86 platform with up to 15X faster real-time inference performance, 12X lower cost, and 12X less energy use, HGX B200 is designed for the most demanding AI, data analytics, and high-performance computing (HPC) workloads.

### NVIDIA HGX H200: The World's Leading AI Computing Platform

As workloads explode in complexity, there's a need for multiple GPUs to work together with extremely fast communication between them. NVIDIA HGX H200 combines multiple H200 GPUs with a high-speed interconnect powered by NVIDIA NVLink and NVSwitch™ to enable the creation of the world's most powerful scale-up servers.

Cirrascale offers the HGX H200 as a dedicated, bare-metal offering in an eight H200 GPU configuration. The eight-GPU configuration offers full GPU-to-GPU bandwidth through NVIDIA NVSwitch. Leveraging the power of H200 multi-precision Tensor Cores, an eight-way HGX H200 provides over 32 petaFLOPS of FP8 deep learning compute and over 1.1TB of aggregate HBM memory for the highest performance in generative AI and HPC applications.

HGX H200 enables standardized servers that provide the highest performance on various application workloads, including LLM training and inference for the largest models beyond 175 billion parameters, while accelerating time to market for NVIDIA's ecosystem of partner server makers.

### NVIDIA HGX H100 in the Cloud with Cirrascale Cloud Services

The NVIDIA HGX H100 brings together the full power of NVIDIA H100 Tensor Core GPUs, NVIDIA® NVLink®, NVSwitch technology, and NVIDIA Quantum-2 InfiniBand networking. As a specialized cloud services provider, Cirrascale delivers all of this to you via the cloud. We offer fully-managed NVIDIA GPU-based 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 in to or out of our cloud. Instead, we charge no ingress or egress fees, so you never receive a supplemental bill.