NVIDIA HGX™ H100 Cloud Service

The NVIDIA Accelerated Server Platform for AI and High Performance Computing is Now Part of the Cirrascale AI Innovation Cloud

Unprecedented performance, scalability, and security for Enterprise AI and High-Performance Computing

An Accelerated Server Platform for Any Workload

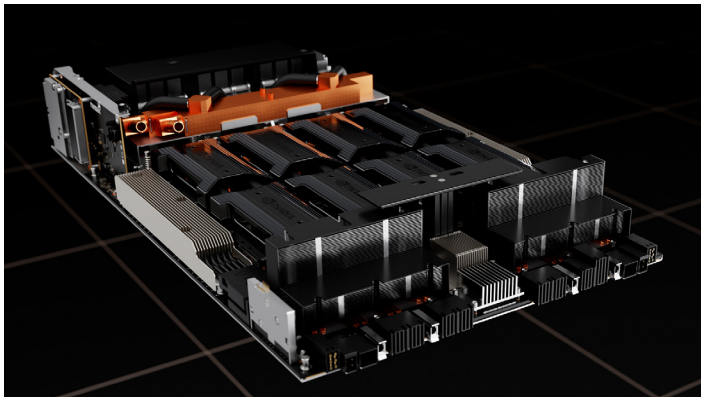
AI solves a wide array of business challenges, using an equally wide array of neural networks. A great AI accelerator has to not only deliver the highest performance but also the versatility to accelerate these networks. The NVIDIA HGX H100 server platform, as offered by Cirrascale, combines eight NVIDIA H100 GPUs with a high-speed interconnect powered by NVLink and NVSwitch technology to enable the creation of the world's most powerful scale-up servers. Leveraging the power of multi-precision Tensor Cores in H100, an eight-way HGX H100 provides over 32 petaFLOPS of FP8 deep learning compute performance. Additionally, Cirrascale offers large-scale NVIDIA HGX H100 clusters built using NVIDIA Quantum-2 InfiniBand networking platform, so users can experience unmatched application performance across multiple servers.

H100 further extends NVIDIA's market-leading inference leadership with several advancements that accelerate inference by up to 30X and deliver the lowest latency. Fourth-generation Tensor Cores speed up all precisions, including FP64, TF32, FP32, FP16, INT8, and now FP8 to reduce memory usage and increase performance while still maintaining accuracy for large language models.

Impressive Performance and Scalability

The HGX H100 8-GPU represents the key building block of the new Hopper generation GPU server. It hosts eight H100 Tensor Core GPUs and four third-generation NVSwitch. Each H100 GPU has multiple fourth generation NVLink ports and connects to all four NVSwitches. Each NVSwitch is a fully non-blocking switch that fully connects all eight H100 Tensor Core GPU. This fully connected topology from NVSwitch enables any H100 to talk to any other H100 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.

The third-generation NVSwitch also provides new hardware acceleration for collective operations with multicast and NVIDIA SHARP in-network reductions. Combining with the faster NVLink speed, the effective bandwidth for common AI collective operations like all-reduce go up by 3x compared to the HGX A100. The NVSwitch acceleration of collectives also significantly reduces the load on the GPU.



Exascale High-Performance Computing

The NVIDIA data center platform consistently delivers performance gains beyond Moore's Law. And H100's new breakthrough AI capabilities further amplify the power of HPC+AI to accelerate time to discovery for scientists and researchers working on solving the world's most important challenges.

H100 triples the floating-point operations per second (FLOPS) of double-precision Tensor Cores, delivering 60 teraFLOPS of FP64 computing for HPC. AI-fused HPC applications can leverage H100's TF32 precision to achieve one petaFLOP of throughput for single-precision, matrix-multiply operations, with zero code changes.

NVIDIA H100 SXM Specifications

Spec	H100 SXM
FP64	34 TFLOPS
FP64 Tensor Core	67 TFLOPS
FP32	67 TFLOPS
TF32 Tensor Core	989 TFLOPS*
BFLOAT 16 Tensor Core	1,979 TFLOPS*
FP16 Tensor Core	1,979 TFLOPS*
FP8 Tensor Core	3,958 TFLOPS*
INT8 Tensor Core	3,958 TOPS*
GPU Memory	80GB
GPU Memory Bandwidth	3.35TB/s
Interconnect	NVLink: 900GB/s PCIe Gen 5: 128GB/s

** Shown with sparsity. Specifications 1/2 lower without sparsity.*

NVIDIA HGX H100 GPU Instances are AVAILABLE NOW at Cirrascale Cloud Services

https://www.cirrascale.com/h100

Exascale High-Performance Computing (continued)

H100 also features DPX instructions that deliver 7X higher performance over NVIDIA A100 Tensor Core GPUs and 40X speedups over traditional dual-socket CPU-only servers on dynamic programming algorithms, such as Smith-Waterman for DNA sequence alignment.

NVIDIA Quantum-2 InfiniBand Networking

Cirrascale offers large-scale NVIDIA HGX H100 clusters built using NVIDIA Quantum-2 InfiniBand networking platform, so users can experience unmatched application performance across multiple servers. Smart adapters and switches reduce latency, increase efficiency, enhance security, and simplify data center automation to accelerate end-to-end application performance.

The data center is the new unit of computing, and HPC networking plays an integral role in scaling application performance across the entire data center. NVIDIA InfiniBand is paving the way with software-defined networking, In-Network Computing acceleration, remote direct-memory access (RDMA), and the fastest speeds and feeds.

NVIDIA HGX H100 in the Cirrascale AI Innovation Cloud is Different...

Flat-Rate Billing Model

Cirrascale offers the NVIDIA HGX H100 cloud service to its customers as a flat-rate billing model ensuring no hidden fees. You pay one price without the worry of fluctuating bills like other cloud providers.

No Ingress / Egress Fees on Data

Cirrascale doesn't charge any ingress or egress fees on transferring data in and out of our cloud service. We respect the fact that using your server means using big data.

Lightning Fast Storage Options

Cirrascale has partnered with the industry's top storage vendors to supply our customers with the absolute fastest storage options available. Specialized NVMe hot-tier storage offerings such as WekaIO Matrix or IBM Spectrum Scale ECE remove storage bottlenecks faced by customers. Options to saturate a GPU cluster and deliver more than 10GBytes/second per node across an InfiniBand network are something we specialize in implementing.to other cloud providers for a multi-cloud experience.

Kubernetes and Cluster Management

Cirrascale can help you with your various cluster management needs. We even have experts to help with Kubernetes -- a container scheduling, management, and orchestration platform. If you're running your containerized application in a single cloud or infrastructure, you're good with single Kubernetes pod/cluster. However, what if you want to build a hybrid application which will run in a multi-cloud environment? We can help answer all those questions, and get you on the right track.

Cross Connections with Third-Party Providers

You're just a cross connect away from a faster connection to cloud providers like Amazon AWS, Microsoft Azure, and Google Cloud through our Megaport connected data centers. It can also help to reduce data egress and port fees that are charged by hyperscale data centers.

NVIDIA GPU Instance Pricing

GPU	Processor	RAM	Local Storage	Hourly Equiv. Pricing*	Monthly Pricing	6-Month Pricing	Annual Pricing
8X NVIDIA H100	Dual 48-core	2TB	(4) 3.84TB NVMe	$34.25	$24,999	$22,499	$19,999

Above pricing is based on Cirrascale's No Surprises billing model with no hidden fees. Pricing shown for servers are per server per month. Hourly equivilent pricing is shown as a courtesy to customers for comparison against other clouds. Cirrascale does not offer hourly service.

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

CSM251 - REV A - 03/2023