IBM Cloud GPU Solutions for AI and HPC Workloads

Solve complex challenges and increase your business value.

IBM Cloud

Introduction

Advancements in GPU technology and the growing demand for artificial intelligence and high-performance computing have changed the way a growing number of industries extract value from their data and push IT performance to new heights.

Healthcare institutions now use chatbots to provide better customer service, petroleum enterprises use reservoir simulators to reduce environmental risks and costs of exploration, and automotive companies teach autonomous cars to adhere the rules of the road. The one thing these data-intensive workloads have in common is a fast, secure and cost-effective cloud infrastructure with the added power of GPU computing.

At IBM Cloud, we know that proper foundations for higher performing infrastructures help improve your bottom line, raise your output potential, and give your IT team the essentials they need without tacking on management hassles. Helping you design and provision cutting-edge NVIDIA GPU architectures within your specific bare metal and virtual server environments is what we do best.

From analytics and graphics enhancement to energy exploration and machine learning – the added power of GPUs is undeniable. The simplicity and cost-savings of building your GPU solutions with IBM Cloud is where you'll find the most value.

"With the powerful GPUs in IBM Cloud Bare metal servers, we knew we would able to dramatically reduce the time taken to optimize and validate our machine learning algorithms."

— Jeffrey Thatcher, Chief Scientist, Spectral MD

Table of Contents

Why IBM Cloud for GPU Solutions

The value of choosing IBM Cloud for your GPU requirements rests within our enterprise infrastructure, platform, and services. You get direct access to one of the most flexible server selection processes in the industry thanks to a broad spectrum of bare metal and virtual server environments. Those cost-savings, control and visibility can help your team overcome GPU adoption barriers. The NVIDIA GPUs we offer were hand-picked to seamlessly integrate into your IBM Cloud architecture, APIs, applications, and workloads. We understand your data location needs, and with IBM Cloud it's a light lift. You get a globally distributed network of modern data centers at your fingertips that improve essentials like high-availability and security for today's "always on" and "always secure" demands.

Ease of use	Highly distributed	Advanced essentials
Customizable bare metal and virtual servers	Worldwide data centers	Hybrid ecosystem expertise
On-demand provisioning and rapid scaling	High-speed private network options	Network application variety
Pre-configured workload options	Maximum control and visibility for high-availability demands	Powerful AI software and hardware data solutions
No vendor lock-ins	Fast data ingest offerings	White-glove service and support

"IBM gave us the flexibility to choose the right GPU for our needs and to pay hourly or monthly as our requirements change. We're now looking at how we can scale our GPU compute needs elastically on the IBM Cloud."

— Danny Tomsett, Founder and CEO, FaceMe

GPUs. Not Just Graphics

GPUs are becoming the modern general processor

Let's quickly go over why GPUs in the first place. By operating in a cloud environment, GPUs work in conjunction with a server's CPU to accelerate application and processing performance. The CPU offloads compute-intensive portions of the application to the GPU, which processes large blocks of data at one time rather than sequentially to boost overall performance in a server environment. By running smaller tasks concurrently, GPUs harness the processing power needed for performance-hungry workloads.

Industry workloads for GPU-enhanced servers have transcended far beyond graphic design. With advancements in technology, lower costs, and a rise in deep learning and AI have turned a once high-priced add-on into a much more accessible, must-have component. With the power of GPUs, deep learning algorithms can perform sophisticated mathematical and statistical computations for a wide range of applications inside your IT architecture.

NVIDIA GPUs by Industry

GPUs are benefiting a wide range of industries

Industries	Businesses	Opportunities for GPU solutions
Financial	Investment banks, brokerage firms, commercial banks, retails	• Risk adverse applications • Faster complex transactions
Healthcare	Hospitals, research centers	• Security application concerns • Large medical image support
Manufacturing	Automobile, Aerospace, consume electronics	• Accelerating time-to-market • Deliver innovations faster
Government	Defense, intelligence, state and local governments	• Security and data breaches • Cyber threats • Evolving regulations
Architecture, Engineering, and Construction (AEC)	Architectural firms, engineering and design firms	• Processing power for applications • Fast remote access when working onsite
Tier One Research	In-house or outside agency scientific research business units and marketing agencies	• Complex molecular modeling calculations • Demographic data mining
Media and Entertainment	Digital studios and agencies	• AI-based toolsets for applying deep learning to their content creation • Customer insight from media assets • Video inferencing, transcoding, and media archiving

Bare Metal vs. Virtual

Raw power of bare metal and the scale and
rapid provisioning of virtual

IBM Cloud Bare Metal Servers:

- High level of customization to fit any workload requirement

- Superior security

- No hypervisor overhead

- All resources dedicated to a single user; no "noisy neighbors"

IBM Cloud Virtual Servers:

- Low provisioning time; pre-configured virtual servers

- Flexible instance options per storage type, memory, and number of GPUs

- Efficient and scalable for burst workloads with hourly and monthly options

*"In our business, there are so many computations that occur.
A process that would normally take us 10 days has shrunk to a
couple of hours by running NVIDIA GPUs on IBM Cloud."*

— Jeffrey Thatcher, Chief Scientist, Spectral MD

NVIDIA GPUs on IBM Cloud

GPUs are benefiting a wide range of industries

Selection	NVIDIA Tesla M60	NVIDIA Tesla K80	NVIDIA Tesla P100	NVIDIA Tesla V100
Ideal environment	Fundamental enterprise performance for virtualization and professional graphics	Reliable enterprise performance for introductory AI computing	Essential performance for growing advanced AI and HPC capabilities	Maximum performance for progressive deep learning workloads
Availability	Monthly bare metal servers	Monthly bare metal servers	Monthly and hourly bare metal servers Monthly and hourly virtual servers	Monthly bare metal servers Monthly and hourly virtual servers
Enabled data centers	North Amerca, Europe, Asia, Australia and South America data centers	North America, Europe, Asia, Australia and South America data centers	(Bare Metal) Dallas, TX San Jose, CA Washington D.C. Amsterdam Seoul Tokyo (Virtual) Dallas, TX Washington D.C. London	Dallas, TX Washington D.C.
Specifications	1TB SATA-3.8TB SSD, 12-28 Cores, 100 MBps – 10 GBps, 64 GB – 1.5 TB			Up to 16 GB of memory capacity with up to 900 GB/s memory bandwidth
Starting at	$650	monthly	$500	monthly	$1.95	hourly $500	monthly	$3.06	hourly $900	monthly

Nuts and Bolts

Get to know the ins and outs of your infrastructure

Add GPUs to handle your most complex workloads, including machine learning, deep learning, and AI. Configure bare metal servers or virtual servers with your choice of the latest NVIDIA Tesla GPUs, designed for high-performance acceleration of scientific computation, data analytics, and large-scale calculations, or NVIDIA GRID, engineered for professional-grade virtualized graphics.

- **Bare metal servers** | https://ibm.com/cloud/bare-metal-servers
  High-performance cloud servers configurable in hourly and monthly options to meet any workload need

- **Virtual Servers** | https://ibm.com/cloud/virtual-servers
  Compute resources that allow you to provision and scale on-demand, meeting workload requirements in minutes

- **Storage** | https://www.ibm.com/cloud/storage
  A flexible storage approach for rapid storage growth driven by new data sources and evolving technologies

- **Network** | https://www.ibm.com/cloud/network
  Solutions designed to enhance your traffic quickly and consistently around the world

- **Watson Studio Deep Learning** | https://www.ibm.com/cloud/deep-learning
  Design complex neural networks then experiment at scale to deploy optimized deep learning models

- **Veeam on IBM Cloud** | https://ibm.co/veeam
  Automated backup, recovery, and replication solutions designed to keep your company running and your data protected with 24x7x365 availability

- **Zerto on IBM Cloud** | https://ibm.co/zerto
  Protect, expand, or migrate your existing VMware vSphere and other hypervisor workloads onto IBM Cloud's global high-performance infrastructure

- **IBM Cloud Object Storage** | https://ibm.co/object-storage
  A highly scalable cloud storage service, designed for high durability, resiliency, and security

IBM **Cloud**

Accelerate AI and HPC workloads with NVIDIA GPUs on IBM Cloud.

To learn more or to order your NVIDIA GPUs on IBM Cloud solutions, visit:

https://www.ibm.biz/nvidia-gpus

Get started →