

NVIDIA simplifies GPU-accelerated computing



Known for popularizing the graphical processing unit (GPU), NVIDIA is now helping enterprise customers adopt GPU-accelerated computing for artificial intelligence (AI) and high-performance computing (HPC) applications. Through its partnership with Red Hat, NVIDIA helps customers run GPU-accelerated computing on Red Hat OpenShift. A GPU operator, developed initially by Red Hat and now owned by NVIDIA, simplifies GPU-accelerated computing on an enterprise-level container platform.

Partner resources

Red Hat ISV Program

Software

Red Hat® OpenShift®

Hardware

Graphical Processing Unit (GPU)

Data Processing Unit (DPU)



Technology

19,000 employees

Benefits

- ▶ Offered crucial insight for ongoing development of the operator
- ▶ Ensured optimal compute efficiency for customers' AI and HPC workloads
- ▶ Saved customers time and avoided manual errors because of automation
- ▶ Provided customers with support and expertise from the right partner

“Red Hat OpenShift is very important to NVIDIA as it allows our customers to develop, deploy, and deliver new apps faster and easier.”

Matt Akins

Business Development Manager, NVIDIA

“When we adapted
CUDA for Kubernetes,
Red Hat OpenShift was
top of mind.”

Matt Akins
Business Development Manager,
NVIDIA

Facilitating processing-intensive computing in the enterprise

NVIDIA popularized the GPU, a specialized processor that can process many pieces of data simultaneously, and helps enterprise customers adopt the processor for running processing-intensive HPC, AI, and cloud operations. GPU-accelerated computing—where the compute-intensive portion of a workload runs on GPUs—is reshaping transportation, healthcare, manufacturing, and many other industries.

NVIDIA developed Compute Unified Device Architecture (CUDA)—a parallel computing platform and programming model for general computing on GPUs—to simplify the development of GPU-accelerated applications. The framework includes libraries, a toolkit, runtime, and plugins that communicate with the GPU. “We made CUDA available on Kubernetes as pretty much all of today’s AI software is container-first,” said Akins.

Customers who initially wanted to take advantage of running Kubernetes on top of GPUs had to manually write containers for CUDA and all the software needed to run GPU-accelerated applications on Kubernetes. Developers also had to write additional code to tell Kubernetes which nodes contained GPUs. The process was time-consuming and prone to errors, but is now greatly simplified using Red Hat OpenShift.

Partnering for an optimal solution

While NVIDIA caters to all Kubernetes distributions, Red Hat OpenShift is seen as a priority. “Red Hat OpenShift is very important to NVIDIA as it allows our customers to develop, deploy, and deliver new apps faster and easier,” said Akins. “When we adapted CUDA for Kubernetes, Red Hat OpenShift was top of mind.”

NVIDIA produced a series of Red Hat OpenShift techniques for CUDA and the software needed by GPU-accelerated applications with guidance from Red Hat. “Working with the Red Hat team, we wrapped all these different plugins up into a single operator, to provide a better way to communicate with the container platform,” said Akins.

Making AI and HPC accessible

When a customer deploys Red Hat OpenShift on top of a server with GPUs, the GPU operator automatically containerizes CUDA and all the software needed before deploying to Red Hat OpenShift.

More than 100 customers are currently using the GPU operator to help them implement and run GPU-accelerated workloads across a wide range of application types, including AI, machine learning, model training, and inferencing.

Red Hat provided NVIDIA with the initial code for the GPU operator, and NVIDIA now maintains it. “The Red Hat engineering team played an integral role in helping us develop the GPU operator,” said Akins.

Building on each other's strengths

Offered crucial insight for ongoing development of the operator

NVIDIA uses Red Hat's expertise and influence regarding Kubernetes, helping NVIDIA understand the container platform's future direction so they can build critical evolutionary advancements into the GPU operator.

Ensured optimal compute efficiency for customers' AI and HPC workloads

The GPU operator allows NVIDIA to optimize compute efficiency for its customers. A process orchestrated by Red Hat OpenShift uses node-labeling techniques, so workloads can automatically find the specific type of GPU they need.

Saved customers time and avoided manual errors because of automation

The GPU operator makes it easier for customers to use CUDA to take advantage of GPU technology for running HPC and AI workloads on Red Hat OpenShift. Automation saves customers time and helps them avoid errors.

Provided customers with support and expertise from the right partner

With the NVIDIA and Red Hat teams aligned, any customer facing an issue with the GPU operator can submit a ticket to Red Hat. The partners then triage the ticket together and have an escalation path before escalating it to either NVIDIA or Red Hat experts, ensuring customers have access to the best support.

Building the next generation of computing

NVIDIA has a GPU network operator development team that regularly collaborates with Red Hat. "We work with Red Hat on the development of the GPU networking operator," said Akins. "It's very much a Red Hat-influenced roadmap."

That team is currently creating an operator for NVIDIA DOCA software framework. An analog to the NVIDIA GPU operator, the network operator simplifies scale-out network design for Kubernetes by automating aspects of network deployment and configuration that would otherwise require manual work. It loads the required drivers, libraries, device plugins, and CNIs on any cluster node with an NVIDIA network interface. DPUs are a new class of programmable processors for cloud-native computing platforms for modern cloud-scale computing. The DPU operator provides security enhancements, future SDKs and frameworks, and other features critical to DOCA.

The partners are working together to help Red Hat OpenShift customers take advantage of GPUs to run AI and HPC workloads and help NVIDIA customers use Red Hat OpenShift. "Our partnership is a great fit," said Akins. "Red Hat is helping more enterprise customers deploy NVIDIA GPUs and NVIDIA is helping Red Hat maximize effectiveness with their presence in the AI and HPC space."

About NVIDIA

NVIDIA's popularization of the GPU sparked the PC gaming market. The company's pioneering work in accelerated computing—a supercharged form of computing at the intersection of computer graphics, high performance computing and AI—is reshaping trillion-dollar industries, such as transportation, healthcare, manufacturing, and fueling the growth of many others.




About Red Hat Innovators in the Open

Innovation is the core of open source. Red Hat customers use open source technologies to change not only their own organizations, but also entire industries and markets. Red Hat Innovators in the Open proudly showcases how our customers use enterprise open source solutions to solve their toughest business challenges. Want to share your story? [Learn more.](#)



About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. [A trusted adviser to the Fortune 500](#), Red Hat provides [award-winning](#) support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.

 facebook.com/redhatinc
 @RedHat
 linkedin.com/company/red-hat

North America
 1 888 REDHAT1
 www.redhat.com

**Europe, Middle East,
and Africa**
 00800 7334 2835
 europe@redhat.com

Asia Pacific
 +65 6490 4200
 apac@redhat.com

Latin America
 +54 11 4329 7300
 info-latam@redhat.com