Customer Success Story | Clemson University

Optimizing HPC Resources for Community Autonomous Vehicle Research

Image courtesy of Clemson University

Virtual GPU Technology Reduces Infrastructure Operating Costs While Meeting Researchers' Unique Needs on Computing Platforms.

Figure 1. The Palmetto HPC infrastructure. Credit: Clemson University.

Introduction

Clemson University is a public, land-grant research university located in Clemson, South Carolina. The university has seven colleges with economic development hubs and research facilities throughout the State of South Carolina. Annually, the university generates approximately $1.9 billion economic impact on the state. Clemson's hub of computational research, known as the Palmetto cluster (Figure 1), is one of the country's largest public academic supercomputers, ranking ninth among U.S. academic systems on the Top 500 list.

The Palmetto cluster is a homegrown, community condo model based HPC infrastructure. It is a heterogeneous system built incrementally, phase-by-phase, over a decade, and has more than 2,000 compute nodes with more than 1,000 NVIDIA GPUs.

To improve Palmetto's accessibility for researchers, Clemson's IT administrators are continuously seeking innovations and better ways to enhance the Palmetto high performance computing (HPC) infrastructure and to increase its resource utilization. Recently, while building a new cluster for an autonomous vehicle (AV) research project, Clemson discovered the value of virtual GPU (vGPU) technology across its infrastructure.

Customer Profile

Company	Industry	Founded	Location	Size	Website
Clemson University	Education	1889	Clemson, South Carolina	5,564 faculty and staff; 24,324 students	clemson.edu

Image courtesy of Clemson University

Summary

- Clemson's researchers needed to run Metamoto software on-prem for their OpenCAV Project.
- Metamoto requires GPU acceleration to provide optimal user experience and performance.
- The IT team deployed a virtual Kubernetes cluster, using one head node and one compute node with two NVIDIA V100 GPUs.
- A physical compute node provides eight VMs with V100 GPUs using NVIDIA® Virtual GPU Manager and NVIDIA RTX® Virtual Workstation (RTX vWS) software.
- Today, researchers can complete their work faster on vGPU-enabled VMs with 1/4 of the physical hardware.

Software

- **Hypervisor:** VMware vSphere Enterprise Plus 6.7
- **Graphics Acceleration:** NVIDIA RTX Virtual Workstation

Hardware

- **Server:** Dell PowerEdge R740
- **GPU:** NVIDIA V100 (32GB PCIe)

NVIDIA Virtual GPU Benefits

- Reduce computing hardware through CPU and GPU virtualization
- Accelerate simulations that need graphical UI and computing power

Challenge

Providing an advanced computing infrastructure to a diversified research community of creative and ambitious researchers requires customer-oriented thinking and creative problem-solving. "We run a cluster with zero red tape," said Corey Ferrier, Clemson University's Director of Advanced Computing Infrastructure (ACI). "All Clemson students and faculty researchers have free access to Palmetto and can ask for whatever job resources they want," Ferrier added.

While a barrier-free HPC infrastructure is the ideal environment for ambitious researchers, quickly delivering the resources to meet their needs isn't always easy. Not only is it difficult to maximize overall system resource utilization for jobs with a wide range of workload characteristics and resource requirements, but also it can be a challenge to run some novel types of applications on a general-purpose infrastructure due to their specialized requirements.

Recently, the ACI group faced one such scenario with the launch of a new autonomous vehicle research project known as Open Connected and Automated Vehicle (OpenCAV). The project, directed by Dr. Venkat Krovi at the Clemson University International Center for Automotive Research, essentially aims to provide an open research instrument that combines virtual reality-based simulation and hardware-based validation to support a wide range of connected automated vehicle research activities. More specifically, the OpenCAV team is working to develop a modular, open-architecture, open interface, and open-source-software research instrument that combines augmente reality (AR) based simulation and physical real-time hardware-in-the-loop validation, on a full-scale vehicle retrofitted with advanced sensing, drive-by-wire, perception connectivity, computation, and control modules. They've selected Metamoto, a Silicon Valley–based startup, as the simulation software provider to build the AR-based simulation component in the research instrument.

> **“NVIDIA vGPUs can significantly drive down IT costs. In the OpenCAV project, virtualization enabled us to reduce our total physical GPUs—and buy fewer servers—while still supporting the same workloads.”**
>
> **Corey Ferrier**
> Director of HPC
> Clemson University Advanced Computing Infrastructure

Unlike conventional HPC applications, Metamoto is provided as a simulation-as-a-service that runs on the Amazon Web Services (AWS) cloud platform. By subscribing to the Metamoto web service, researchers can train, test, and validate the autonomous system software that controls an AV’s sensors, driving, and interactions in a virtual but realistic environment. Enhanced with deep learning algorithms, Metamoto helps researchers design optimal AV sensor-suites and algorithms before integrating them into a full-scale vehicle for physical real-time hardware-in-the-loop validation. Soon, the OpenCAV project will be able to launch an AV that travels safely on public roads.

However, as a research project with a limited budget, OpenCAV could not afford to run Metamoto on AWS 24/7. Dr. Krovi, the principle investigator (PI) of the OpenCAV project, approached Clemson’s IT team about running the Metamoto software on Palmetto. “That was tricky because Palmetto currently doesn’t support Kubernetes-based workflows—and Metamoto requires Kubernetes to schedule its containerized workloads,” said Ferrier. “So, we thought we would handle this project’s needs by standing up a separate Kubernetes infrastructure with one head node and several compute nodes, each with two NVIDIA V100 GPUs, to provide the computational horsepower required by the simulation.”

Solution

During an internal discussion, Dr. Xizhou Feng, a senior HPC researcher and system scientist at Clemson, brought about the idea of using GPU-enabled virtual machines instead of physical servers to build the computing infrastructure for OpenCAV. This idea was inspired by an accidental question asked by a faculty member, “In an earlier Metamoto email they said in their current demo, they count each K80 as two GPU nodes, citing K80 being a ‘dual-GPU’ design. How do we count each V100?” Intrigued by the question, Feng researched NVIDIA’s virtual GPU technology website and replied, “The answer depends on the vGPU configuration. With proper hypervisor software support, one V100 PCIe 32G can support up to eight V100-4GB vGPU.” After consulting with the Metamoto team, the Clemson team adopted the vGPU design.

In July 2019, the ACI team began implementing an on-premise solution for OpenCAV. The system is comprised of two Dell R740 servers and the cluster consists of eight worker VMs, with one NVIDIA V100-8Q virtual GPU per VM, provided by two physical V100 PCIe 32GB cards. That is, the solution uses one single compute node with two V100 PCIe 32G GPUs to create eight virtual machines (VMs). Per Metamoto’s request, each VM has an 8Q profile vGPU, runs Ubuntu 16.04.6 LTS Linux server, and serves as a Kubernetes node that runs the Metamoto simulations. A separate server, which does not come with any GPU, provides another set of VMs running Kubernetes master, VMware vCenter, and license server.

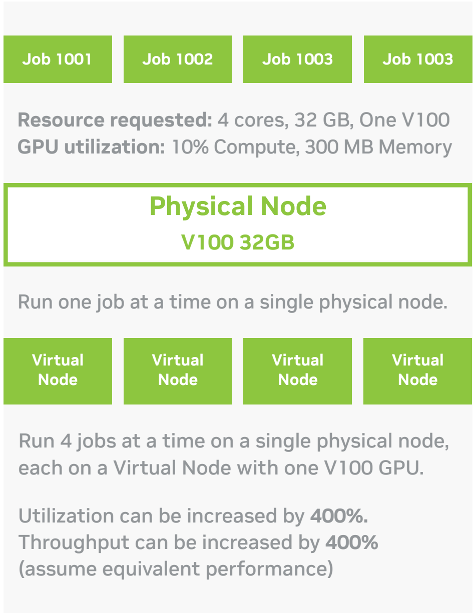

Figure 2. Illustration depicting how GPU virtualization increases GPU resource utilization and overall system throughput. Credit: Clemson University.

The team determined that NVIDIA RTX vWS was the right solution because Metamoto has design components which require visualization (OpenGL) support and RTX vWS provided researchers with the ability to design and visualize the simulations and analyses.

In late October, the Metamoto technical team came on site and installed the Metamoto software platform on the virtual cluster. Then, the cluster was handed over to the OpenCAV team, who had Metamoto up and running in the same way they would on AWS.

Result

Today, Metamoto performs smoothly on the dedicated OpenCAV infrastructure. After the Metamoto team installed their software in the infrastructure, they tested and compared it with AWS. The Metamoto team found that the VMs in the OpenCAV system outperformed those on AWS (Figure 2). Those tests validated that on-prem GPU virtualization is a feasible deployment option for Metamoto without performance compromise.

The IT team wasn't surprised by the improved performance as the OpenCAV environment runs on the NVIDIA V100 GPUs, which are based on a newer generation NVIDIA Volta™ architecture. In contrast, the AWS instances that were used included the NVIDIA K80 GPUs, which are based on a previous generation NVIDIA Kepler architecture. NVIDIA's newest GPUs are some of the most powerful on the market today, delivering exponentially faster performance for numerous deep learning and simulation applications, including Metamoto.

In addition to stellar performance, GPU virtualization enabled Clemson to reduce the number of physical nodes in a ratio of 4:1, while still meeting user requirements. Rough cost calculations show that over a three-year period it would have been more than twice as expensive to run Metamoto on AWS than on the new OpenCAV mini-cluster.

Looking Ahead

The successful completion of the OpenCAV Metamoto project provided the ACI team with two important things: a proof of concept that NVIDIA vGPU is capable of running actual compute-intensive simulations and a virtualization infrastructure on which the team can test new ideas to enhance HPC infrastructure services. Shortly after handing over the OpenCAV cluster to the researchers, the team launched the Palmetto vGPU Testbed project to study the feasibility and practicality of applying GPU virtualization technology to improve GPU resource utilization on the HPC cluster.

“It is undisputed that GPU computing has been successful in speeding up numerous compute-intensive computational problems. Many institutions face a common problem of GPU resource under-utilization.”

Xizhou Feng, PhD
HPC System Scientist

“It is undisputed that GPU computing has been successful in speeding up numerous compute-intensive computational problems, such as simulations and deep learning. However, many institutions face a common problem of GPU resource under-utilization,” said Dr. Feng, who has conducted extensive research on optimizing the performance and efficiency of HPC systems and applications. By virtualizing the GPU resources, it is possible for HPC infrastructure providers to develop promising solutions with better user and job isolation, higher system throughput, and improved system flexibility and scalability.

One of the major benefits they’ve realized is improved resource utilization. Conceptually, increasing GPU resource utilization by 400% is feasible through GPU virtualization and judicially job scheduling. This increased resource utilization not only reduces operational costs for the HPC infrastructure but also creates more available resources for researchers to run their applications.

GPU virtualization is particularly promising for research institutions like Clemson, which provides centralized computing infrastructures like Palmetto as a shared resource by a large community of researchers. Because GPU resources are limited, compute-intensive simulations and deep learning workloads are typically put into long queues before they can be started. To speed up the start time of these jobs, organizations often purchase additional GPU resources. However, a significant amount of these resources have historically been wasted due to low utilization, as many GPU-accelerated applications only use a small fraction of the memory and computing capacity of the GPUs allocated to run them.

While the ACI team at Clemson has consistently worked to fix the GPU resource underutilization problem, the team neither has control over the researcher’s code nor can direct how the researchers run their jobs. NVIDIA vGPU technology is capable of providing a workable solution: virtualizing a part of the GPU resources and scheduling jobs with low GPU utilization to the virtual nodes.

“Not only are NVIDIA vGPUs capable of supporting common deep learning workloads that can fit on the VM, they also improve throughput by 1.6X compared to physical GPUs.”

Xizhou Feng, PhD
HPC System Scientist

For the vGPU Testbed project, the ACI team used two GPU nodes to build a group of 16 virtual GPU nodes, each of them featuring a V100-8Q profile. These virtual nodes run the same operating system (Oracle Linux 7.6) and software stack as those run by a typical Palmetto physical node. These virtual nodes are provisioned with xcat and managed by the PBSPro scheduler. To a normal user, a virtual GPU node is no different from a physical node, except for a lower number of CPU cores and a smaller memory footprint.

In the Testbed pilot phase, two groups of graduate students advised by Prof. Rong Ge in the School of Computing conducted a benchmark study using a set of deep-learning benchmarks on the vGPU Testbed. Their study found that for deep learning problems that can fit on one vGPU VM, not only are NVIDIA vGPUs capable of supporting common deep learning workloads, but they also improve performance and throughput by approximately 1.6X compared to physical GPUs.

In the test example (Figure 3), it should be noted that 1 GPU means one V100 32G physical GPU. 1/2/4 vGPU means configuring one V100 physical GPU as 1 V100-32Q, 2 V100-16Q, and 4 V100-8Q virtual GPU-enabled VMs. The subfigure on the left compares the relative speedup, and the subfigure on the right compares the GPU utilization. As noted previously, these results show that running the same workload using four V100-8Q enabled VMs can speed up the benchmark by 1.6 times and increase utilization by four times.

Figure 3. Performance comparisons between physical GPUs and virtual GPUs. Credit: Rong Ge and her students, Deep Learning on vGPUs: A Case for High GPU Utilization and Throughput.

In January, the 16 vGPU nodes were deployed on the Palmetto system and made available to all users. Within the first month, 14 different users ran about 160 jobs on the nodes. No issues were reported by users, an indication that the vGPU nodes are capable of running jobs just like physical nodes, and thus, can potentially increase GPU resource utilization significantly when deployed on a larger scale.

However, there's still work to be done. The ACI team observed that not enough jobs have been allocated to the vGPU nodes to keep them 100% busy. The underlying cause for this is the mismatch between the overall resources requested by the user jobs (such as CPU cores) and those provided by the virtual GPU nodes. The team believes that because there are no technical problems related to GPU virtualization, the issue can be remedied through education.

One of Clemson's main takeaways derived from the OpenCAV and the Testbed projects is that GPU virtualization offers a practical solution for improving GPU resource utilization without compromising application performance on community HPC infrastructure. In addition, it allows the infrastructure provider to create compute systems that meet researchers' special needs without over-investing on physical hardware. In other words, GPU virtualization can potentially help the ACI team reduce the Palmetto cluster's need for physical GPUs—and therefore invest in fewer servers—while still supporting the same workloads.

To learn more about NVIDIA virtual GPU solutions, visit:
www.nvidia.com/virtualgpu