CASE STUDY | MSOE

RESHAPING THE FUTURE OF LEARNING WITH LEADERSHIP-CLASS AI INFRASTRUCTURE

NVIDIA

The Milwaukee School of Engineering Deployed NVIDIA DGX Systems and NVIDIA GPU-based Servers to Deliver AI Computing to Every Student

Founded in 1904, the Milwaukee School of Engineering (MSOE) is one of the top schools for engineering in the country, ranking ninth among the best undergraduate engineering programs and fourth among the best computer engineering programs in the U.S.[1] This small university of 2,800 students is dedicated to achieving lofty goals, being known for its low student-to-faculty-ratio, real-world application through hands-on coursework, and deep industry integrations.

CHALLENGE

MSOE established a computer science degree offering in 2018, which is augmented with specialized courses in AI, such as computational science, data science, and deep learning. Since that time, the university has seen a huge influx of students pursuing majors involving computing. Providing realistic AI projects for students to work on requires a tremendous amount of computational resources along with a software stack that's optimized to run on GPU clusters.

The university originally had students use their own laptops, but this created problems, as most laptops have limited GPU acceleration. Additionally, it was challenging to deploy a consistently performing software stack on all student laptops. For instance, inconsistent AI or high-performance computing (HPC) libraries can lead to software working on some machines and not others.

CUSTOMER PROFILE

Organization:	**Industry:**	**Location:**	**Founded:**	**Size:**	**Website:**
Milwaukee School of Engineering	Education	Milwaukee, WI	1904	2,820	msoe.edu

[1] *MSOE. "Excellent Reputation and Rankings: U.S. News & World Report (2021)." www.msoe.edu/about-msoe/who-we-are/rankings-2/*

SUMMARY

- MSOE wanted to integrate AI courses in their computer science curriculum to meet growing industry demand.
- To achieve this, they needed a tremendous amount of computational resources and an optimized software stack to enable a wide variety of AI workloads for a large number of students.
- Cloud instances were limiting experimentation, while personal laptops were limited in performance.
- They leveraged a combination of NVIDIA DGX™ systems and NVIDIA T4 Tensor Core GPU-based servers to support their AI, teaching, and research workloads.
- NVIDIA GPU cluster management tools allow a single administrator to easily support the cluster. Open OnDemand with Slurm and Jupyter notebooks simplifies use, driving high utilization and integration early in the curriculum.
- Faculty use NVIDIA NGC™ software containers and pre-trained models for course laboratory projects. This saves them hours of course development time and lets them better replicate environments to improve debugging.

MSOE initially used cloud instances, but experienced limitations. As Derek Riley, MSOE professor and program director of computer science, explained, "There was a steep learning curve to being able to get instances set up and running with the right libraries." Professor Riley also mentioned the cloud wasn't able to provide the latest and greatest technology, and the students were experiencing throttled performance.

One of Professor Riley's biggest concerns was how cloud computing was constraining student learning. "With the cloud, you have a certain number of credits, and when you run out of credits, you have to pay more," he said. "We really didn't want our students to feel like they were going to have to spend more money if they were struggling with an assignment or if they wanted to explore something further."

Beyond interactive use cases, they had to support batch use cases from both faculty and students, who sometimes needed to run larger- scale workloads with bigger datasets on systems with multi-GPU capabilities. This stemmed from undergraduate research projects, side projects, faculty collaborations, and industry consulting.

Professor Riley considered a traditional Secure Shell (SSH)-based HPC cluster interface but dismissed the idea due to the steep learning curve required for undergraduate students. The environment also made it difficult to compile programs and link libraries.

SOLUTION

In 2017, Dr. Dwight Diercks, senior vice president of software engineering at NVIDIA and a graduate of MSOE, provided a $34 million gift to MSOE to fund the Dwight and Dian Diercks Computational Science Hall. Diercks Hall opened in September of 2018 and includes laboratories, classrooms, and a data center.

Professor Riley worked with NVIDIA solutions architects to design the architecture and development of a computing cluster that could easily support the growing demand for interactive and batch use cases. "We wanted a system that maximized training performance but also was flexible enough for a wide variety of use cases," Professor Riley commented. NVIDIA solutions architects and Microway, an NVIDIA Partner Network (NPN) provider[2], collaborated with MSOE on the design.

INFRASTRUCTURE

- **AI sub-cluster:** DGX systems
- **Teaching sub-cluster:** Microway servers with NVIDIA T4 GPUs
- **Storage:** NetApp FAS8200

SOFTWARE

- **GPU cluster management:** DeepOps
- **Cluster monitoring:** NVIDIA Data Center GPU Manager (DCGM) and Ganglia
- **Containers and Models:** NVIDIA NGC Catalog
- **User access:** Open OnDemand with Slurm and Jupyter notebooks

Microway delivered all the hardware, deployed the cluster at the university, and performed the initial software configuration.[2]

The AI sub-cluster is composed of three NVIDIA DGX systems, while the teaching and research sub-cluster is composed of 20 servers, each with four NVIDIA T4 GPUs. The nodes are joined together by NVIDIA networking fabric and share 200 terabytes (TB) of network-attached storage. The cluster storage uses NetApp FAS8200, providing great performance and flexibility for expansion. Lastly, the login and management nodes contain four servers that enable redundancy and load balancing. This supercomputer is called "Rosie", whose name was inspired by the women who programmed one of the earliest computers - the Electronic Numerical Integrator and Computer (ENIAC) - and were captured in the documentary *Top Secret Rosies: The Female Computers of WWII*.[3]

The system administrator can easily manage the cluster using DeepOps, an NVIDIA GPU cluster management toolkit. NVIDIA Data Center GPU Manager (DGCM) is also being utilized to monitor health and utilization of the GPUs along with Ganglia, which provides a web UI that gives visibility into individual systems and GPUs.

Professor Riley also wanted to ensure ease of use for students, faculty, and staff. There's a separate software stack for each use case: The teaching stack uses Singularity containers and Jupyter notebooks and is built on an Open OnDemand web-based interface for students; the research stack has Singularity containers, provides an SSH-based interface, and supports bare-metal programming. Both stacks leverage the NGC catalog — NVIDIA's GPU-optimized hub of HPC and AI applications, including containers, pre-trained models, and SDKs — to simplify software and library management.

MSOE professors use the NGC catalog as a basis for class projects and assignments. With NGC containers, students can simply run their HPC or AI applications on the cluster, without needing to build complex environments. Students can fast-track their AI projects by fine-tuning

[2] ***Microway*** *delivers HPC & AI solutions to the marketplace with expertise in design and manufacture of high quality hardware and software. They are part of the* ***NVIDIA Partner Network (NPN)*** *program, which gives partners the opportunity to offer full-stack accelerated computing solutions from NVIDIA, provided they meet specific competency requirements.*

[3] *MSOE. "Meet Rosie." Sept 14, 2019. www.msoe.edu/about-msoe/news/details/meet-rosie/*

"With DGX systems, we can now work on larger problems, and because we can work on larger problems, we can achieve different learning outcomes than if we were working on toy problems."

Derek Riley, PhD
Professor, Program Director of Computer Science
Milwaukee School of Engineering

the pre-trained models from the catalog, saving them hours of work and exposing them to common workflows that exist to solve industrial challenges.

RESULTS

"We've seen a great increase in demand for the system due to ease of use. Most students access Rosie through our Open OnDemand portal web UI," said Professor Riley. "They can easily request the resources they need, and the base images already come with a special set of pre-configured libraries. Students can easily run single GPU jobs on this partition or run jobs across multiple GPUs or DGX systems."

Professor Riley commented, "With DGX systems, we can now work on larger problems, and because we can work on larger problems, we can achieve different learning outcomes than if we were working on toy problems."

In its first year, only two distinct courses used Rosie. Today, one year later, nine courses are using Rosie. And next year, over 15 courses, with multiple sections, are expected to use Rosie. Beyond teaching, Rosie is being heavily used in research — from undergraduate to faculty projects in computer science, as well as faculty across the university, from engineering to business administration.

"Prior to Rosie's deployment, we used AWS, Google Colab, and personal laptops. As a professor, I want to help the students debug and find answers to the problem, as opposed to spending a lot of time addressing their configuration challenges," said Professor Riley. "With NVIDIA DGX systems, our students have access to the best-in-class AI infrastructure and no longer have to worry about the cloud 'odometer' always running

> "Eighty percent of undergraduate computer science students leverage Rosie today, and faculty GPU usage has increased 10X in just one year. Additionally, the system has had over 20,000 Slurm jobs so far."

Derek Riley, PhD
Professor, Program Director of Computer Science
Milwaukee School of Engineering

and limiting experimentation. Eighty percent of undergraduate computer science students leverage Rosie today, and faculty GPU usage has increased 10X in just one year. Additionally, the system has had over 20,000 Slurm jobs so far."

Because of the recent COVID-19 pandemic, most classes at MSOE are conducted online, and they've seen very little interruption of service. All students are able to access Rosie remotely from wherever they are via VPN. Having a weak connection, whether they're in rural areas or sharing it with others, is not a limiting factor when accessing Rosie and using Jupyter notebooks.

LOOKING AHEAD

As MSOE's motto is to enable their students to learn differently and provide practical applications through hands-on opportunities, they plan to leverage Rosie to boost university-AI research collaborations even further. MSOE already has active collaborations with several local businesses, in addition to research grant proposal opportunities with several organizations. Additional security hardware is being planned to ensure data privacy and security. "We plan to support industrial projects in a large way going forward, as we feel this drives a lot of learning for our students," Professor Riley mentioned.

To learn more about NVIDIA DGX systems, visit: **www.nvidia.com/dgx**

www.nvidia.com