Researchers tackle new machine learning algorithm designs with powerful AI platform

University of Wisconsin-Madison | Industry: Education/research | Location: Madison, Wisconsin

Founded in 1848 and known as the flagship campus of the University of Wisconsin system, the University of Wisconsin-Madison (UW-Madison) is a public land-grant university that boasts $1.2 billion in research expenditures annually across its approximately 100 research centers and programs, spanning the spectrum of agriculture, arts, education, and engineering. UW-Madison has 20 schools and colleges, with an enrollment of 30,361 undergraduate and 14,052 graduate students in 2018. Academic programs include 136 undergraduate majors, as well as 148 master's degree programs and 120 doctoral programs.

Objectives

- Develop better systems for large-scale machine learning
- Build a multi-tenant training setup to support elastic machine learning training-as-a-service
- Achieve ultra-low latency inference
- Evaluate emerging hardware to support a longer shelf life for solutions being designed and tested

Solution

- Participated in Early Availability Program (EAP) for Cisco UCS® C480 ML Server with NVIDIA V100 Tensor Core GPUs

Findings

- Trends in the latest AI platform (i.e. balance between compute and communication) drove the design of different types of algorithms within the research
- Research leveraged massive parallel compute capability to perform many iterations on a large number of permutations and to accelerate design cycles to converge on a solution
- There is potential to improve the algorithms to enhance speed and accuracy of machine learning capabilities (e.g. personalized recommendations, sensors, digital assistants)

For more information:

Cisco UCS C480 ML Server

Objective: Develop better systems for large-scale machine learning

University of Wisconsin Assistant Professor Shivaram Venkataraman and Professor and H. I. Romnes Faculty Fellow Aditya Akella, have a clear mandate. They, along with their student researchers Surya Teja Chavali, Adarsh Kumar, and Kshiteej Mahajan, are pursuing the goal of developing better systems for large-scale machine learning by designing and testing a variety of solutions for specific use cases.

The use cases include:

- Build a multi-tenant training setup to enable the proper apportioning of NVIDIA GPU and interconnect resources to support machine learning training-as-a-service.
- Make algorithm level changes to support elasticity in machine learning training jobs.
- Improve the accuracy of inference (enables the predictive capability used in most web-facing applications) while accelerating response times.

"Within our research, we are considering how different large-scale machine learning models behave with different data sets and parameters," says Venkataraman. "It is critical that we have the AI compute platform that supports the effective testing of our solutions on these models."

The research team is committed to designing algorithms with a long shelf life. As a result, there is a focus on experimenting with emerging technologies that will have the longevity to support the large models and data sets now becoming popular in machine learning. In addition, Venkataraman, Akella, and their students are keen to explore the trends in the underlying computing platform and how their designs would work best given those trends.

"We develop software solutions. While we want the latest compute platform to support emerging machine learning capabilities, we also want the assurance that eventually the platform will be used widely. That way, our software can benefit the greatest number of people," says Akella.

With these objectives in mind, the research team chose to explore systems with the NVIDIA NVLink high-speed GPU interconnect and multiple high-end GPUs with ample memory, connected in scalable topologies.

"We ran several different benchmarks, including compute and communication timeframes. The Cisco UCS C480 ML Server did the best job of balancing compute and communication, which is important because communication typically creates a bottleneck in the process."

Shivaram Venkataraman
Assistant Professor, University of Wisconsin-Madison

Revelations from testing the Cisco technology

In the technology exploration phase, the research team leveraged the university's cloud lab to benchmark its work, as well as investigate NVIDIA GPU servers from other providers including Amazon EC2.

However, participating in the Early Availability Program for the new Cisco UCS C480 ML Server with NVIDIA V100 Tensor Core GPUs enabled the research team to meet the majority of its research requirements.

Venkataraman really appreciated the amount of parallel computation available in the Cisco UCS C480 ML platform. "The Cisco hardware gave us the opportunity to use up to eight NVIDIA GPUs, with each GPU having a lot of parallelism within it. With that capability, we were able to prototype a number of different training methodologies and test out a number of different models for inference," he says.

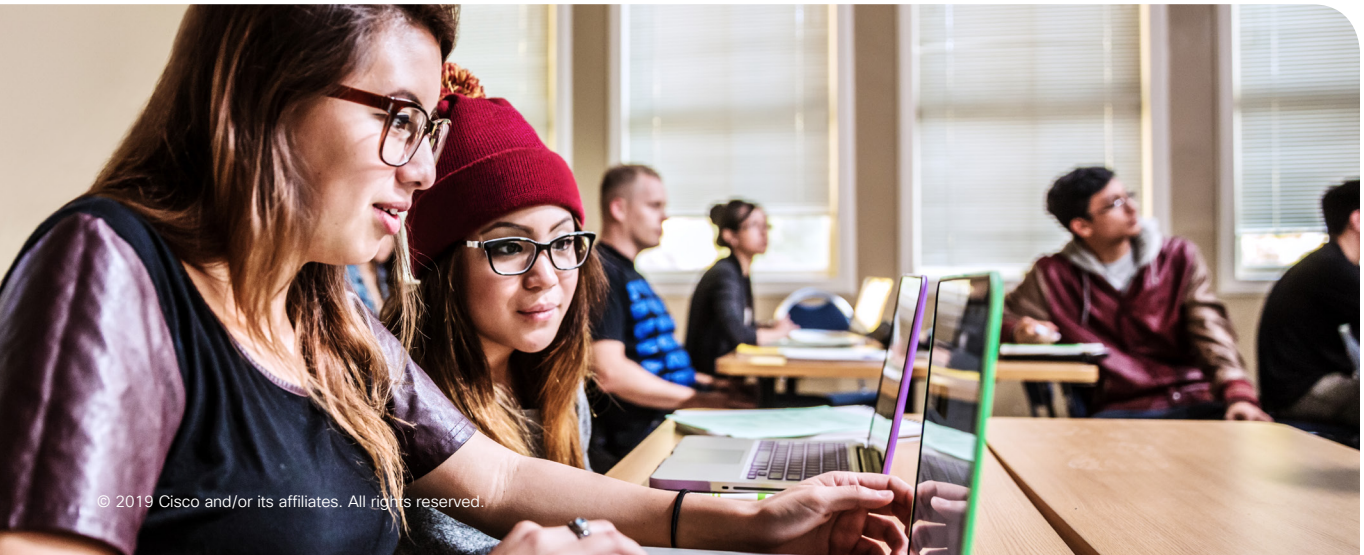
The communication speeds enabled by the NVIDIA NVLink interconnect between the NVIDIA GPUs in the Cisco server were also critical for the research team's work. The university's cloud lab machines did not offer such a high interconnect speed. The research team leveraged interconnect speed specifically when testing distributed training situations to understand quickly how well distributed training works with a change in the degree of parallelism.

"We could have gotten a multi-GPU or machine instance from Amazon," notes Akella. "The NVLink capabilities within the Cisco server really appealed to us because we could explore what it means to configure different topologies and their impact on training algorithms."

Having access to the latest generation CPUs was useful in enabling the research team to model and understand the trade-offs between using CPU and GPU, especially for inference tasks. The NVME storage provided high-speed access at large capacity, which was especially important when the team ran training on larger-sized image net models. The faster storage accelerated the overall performance of the actual training algorithm. Most data in machine learning training applications is read only. In terms of memory, having a large buffer cache enabled a much faster IO access pattern for all machine learning training workloads.

The flexibility of connectivity and storage configuration was also of great interest to the research team and is something that it plans to explore more deeply in the future. Having the connectivity option to big data can enable the team to do machine learning right where the data lives. The server's 24 drives can enable the choice between local or shared storage, or integration directly to the data lake. That level of flexibility is vitally important for distributed machine learning training.

"We ran several different benchmarks, including compute and communication timeframes," says Venkataraman. "The Cisco UCS C480 ML Server did the best job of balancing compute and communication, which is important because communication typically creates a bottleneck in the process."



> "Having underlying hardware like the Cisco UCS C480 ML Server with NVIDIA GPUs makes us very excited for how we can develop faster, more accurate algorithms that will propel personalized recommendations, sensors, digital assistants, and IoT to the next level."
>
> **Aditya Akella**
> Professor, University of Wisconsin-Madison

Inspiring new algorithmic ideas to drive better speed and accuracy

After completing its participation in the EAP for the Cisco UCS C480 ML Server, the research team felt encouraged by the potential impacts of the technology on its work.

Akella felt that the trends inherent in the technology led to new design ideas and approaches. For example, the compute/communication balance within the server facilitated the design of different types of algorithms for providing elasticity.

Venkataraman also found the vast amount of parallel resources hugely valuable to accelerating a research process that relies heavily on ongoing solution experimentation and numerous iterations. With more powerful hardware resources, the research team can test more ideas and shorten the time to final design.

"Researchers typically make assumptions about the underlying hardware, and those notions often get baked into the algorithm design to a certain extent," says Venkataraman. "The Cisco UCS C480 ML Server goes against conventional assumptions with respect to expected bottlenecks, which I'm confident will result in a newer class of algorithms that will benefit the machine learning community at large."

Based on their experience and learnings, specifically around communication and computation, both Venkataraman and Akella believe that algorithms can change based on the power and potential of underlying hardware like the Cisco UCS C480 ML Server.

What does that mean for machine learning and associated research breakthroughs? Researchers measure advances in algorithms based on accuracy. Pushing the frontier of algorithm design has a direct, positive impact on accuracy and the speed of achieving it.

"If we improve algorithms, we improve speed and accuracy, which in turn can impact all sorts of capabilities that rely on machine learning," notes Akella. "Having underlying hardware like the Cisco UCS C480 ML Server with NVIDIA GPUs makes us very excited for how we can develop faster, more accurate algorithms that will propel personalized recommendations, sensors, digital assistants, and IoT to the next level."

Products

- Cisco UCS C480 ML Server complete with 8 NVIDIA V100 Tensor Core GPU modules with NVLink interconnect