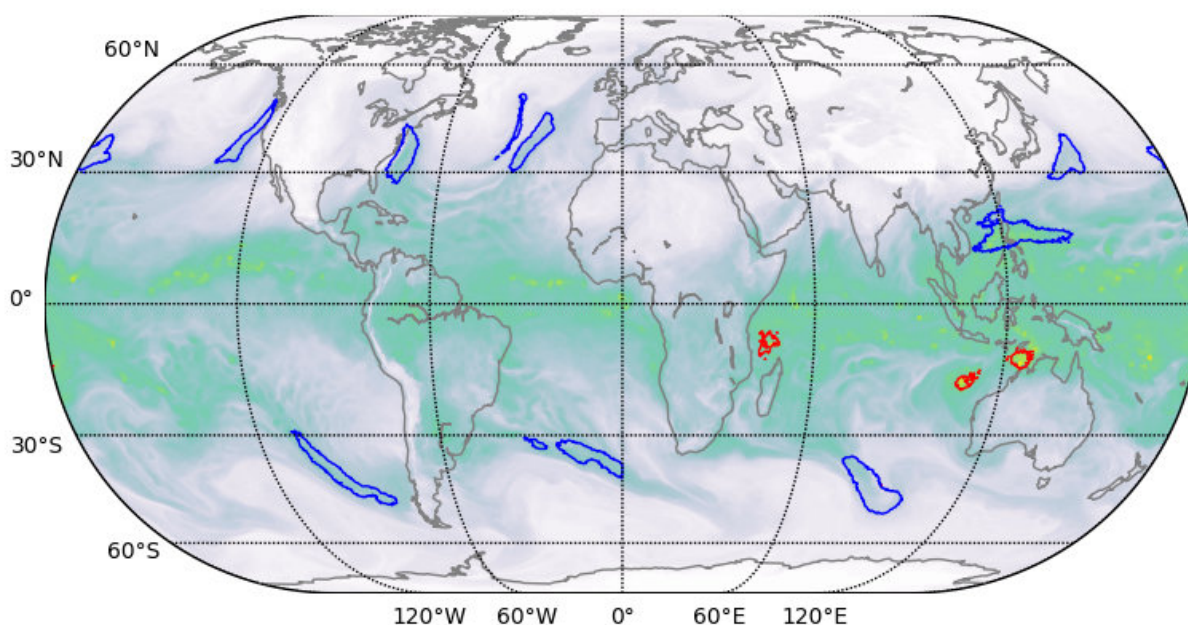# Team breaks exaop barrier with deep learning application

October 9 2018, by Kathy Kincade

High-quality segmentation results produced by deep learning on climate datasets. Credit: Berkeley Lab

A team of computational scientists from Lawrence Berkeley National Laboratory (Berkeley Lab) and Oak Ridge National Laboratory (ORNL) and engineers from NVIDIA has, for the first time, demonstrated an exascale-class deep learning application that has broken the exaop barrier.

Using a climate dataset from Berkeley Lab on ORNL's Summit system at the Oak Ridge Leadership Computing Facility (OLCF), they trained a deep neural network to identify extreme weather patterns from high-resolution climate simulations. Summit is an IBM Power Systems AC922 supercomputer powered by more than 9,000 IBM POWER9 CPUs and 27,000 NVIDIA Tesla V100 Tensor Core GPUs. By tapping into the specialized NVIDIA Tensor Cores built into the GPUs at scale, the researchers achieved a peak performance of 1.13 exaops and a sustained performance of 0.999—the fastest deep learning algorithm reported to date and an achievement that earned them a spot on this year's list of finalists for the Gordon Bell Prize.

"This collaboration has produced a number of unique accomplishments," said Prabhat, who leads the Data & Analytics Services team at Berkeley Lab's National Energy Research Scientific Computing Center and is a co-author on the Gordon Bell submission. "It is the first example of deep learning architecture that has been able to solve segmentation problems in climate science, and in the field of deep learning, it is the first example of a real application that has broken the exascale barrier."

These achievements were made possible through an innovative blend of hardware and software capabilities. On the hardware side, Summit has been designed to deliver 200 petaflops of high-precision computing performance and was recently named the fastest computer in the world, capable of performing more than three exaops (3 billion billion calculations) per second. The system features a hybrid architecture; each of its 4,608 compute nodes contains two IBM POWER9 CPUs and six NVIDIA Volta Tensor Core GPUs, all connected via the NVIDIA NVLink high-speed interconnect.The NVIDIA GPUs are a key factor in Summit's performance, enabling up to 12 times higher peak teraflops for training and 6 times higher peak teraflops for inference in deep learning applications compared to its predecessor, the Tesla P100.

"Our partnering with Berkeley Lab and Oak Ridge National Laboratory showed the true potential of NVIDIA Tensor Core GPUs for AI and HPC applications," said Michael Houston, senior distinguished engineer of deep learning at NVIDIA. "To make exascale a reality, our team tapped into the multi-precision capabilities packed into the thousands of NVIDIA Volta Tensor Core GPUs on Summit to achieve peak performance in training and inference in deep learning applications."

Improved Scalability and Communication

On the software side, in addition to providing the climate dataset, the Berkeley Lab team developed pattern-recognition algorithms for training the DeepLabv3+ neural network to extract pixel-level classifications of extreme weather patterns, which could aid in the prediction of how extreme weather events are changing as the climate warms. According to Thorsten Kurth, an application performance specialist at NERSC who led this project, the team made modifications to DeepLabv3+ that improved the network's scalability and communications capabilities and made the exaops achievement possible. This included tweaking the network to train it to extract pixel-level features and per-pixel classification and improve node-to-node communication.

"What is impressive about this effort is that we could scale a high-productivity framework like TensorFlow, which is technically designed for rapid prototyping on small to medium scales, to 4,560 nodes on Summit," he said. "With a number of performance enhancements, we were able to get the framework to run on nearly the entire supercomputer and achieve exaop-level performance, which to my knowledge is the best achieved so far in a tightly coupled application."

Other innovations included high-speed parallel data staging, an optimized data ingestion pipeline and multi-channel segmentation. Traditional image segmentation tasks work on three-channel

red/blue/green images. But scientific datasets often comprise many channels; in climate, for example, these can include temperature, wind speeds, pressure values and humidity. By running the optimized neural network on Summit, the additional computational capabilities allowed the use of all 16 available channels, which dramatically improved the accuracy of the models.

"We have shown that we can apply deep-learning methods for pixel-level segmentation on climate data, and potentially on other scientific domains," said Prabhat. "More generally, our project has laid the groundwork for exascale deep learning for science, as well as commercial applications."

Provided by Lawrence Berkeley National Laboratory