Researchers develop open-source mixed-precision benchmark tool for supercomputers

September 25 2023, by Coury Z Turczyn

Members of the Analytics and AI Methods at Scale group in the National Center for Computational Sciences at ORNL developed the mixed-precision performance benchmarking tool OpenMxP. From left are group leader Feiyi Wang, technical lead Mike Matheson and research scientist Hao Lu. Credit: Carlos Jones/ORNL, U.S. Dept. of Energy

As Frontier, the world's first exascale supercomputer, was being assembled at the Oak Ridge Leadership Computing Facility in 2021, understanding its performance on mixed-precision calculations remained a difficult prospect. That gap in understanding wasn't an oversight but rather a sign of just how novel supercomputer systems that excel at mixed precision remain in computational science, which has been dominated by double precision–focused systems for almost its entire history.

Double-precision—or 64-bit—arithmetic is the primary standard for computational accuracy in simulations. Mixed-precision arithmetic—16 or 32 bits—often calculated by GPUs, can potentially offer required levels of accuracy at much faster speeds, especially for data science and AI. But no widely available, open-source software has existed for testing lower-precision performance at extreme scale on GPU-accelerated supercomputers, which first saw large-scale deployment in 2012 with the OLCF's Titan system.

Consequently, researchers at the Department of Energy's Oak Ridge National Laboratory developed a new cross-platform benchmarking software package just in time for Frontier's launch in May 2022: OpenMxP. They've also made it available to other computing facilities as an open-source code.

"The supercomputer is a vital foundation for maintaining U.S. computing technological leadership, and we're in the business of pushing the frontier—pun intended—of supercomputing. But you cannot improve it if you cannot measure it, which highlights the importance of benchmarking. This reference implementation of OpenMxP as a capability benchmark will benefit all the other leadership computing systems," said Feiyi Wang, leader of the Analytics and AI Methods at Scale, or AAIMS, group at the National Center for Computational Sciences at ORNL.

For its utility in supercomputer assessment and operation, OpenMxP was recently recognized as a 2023 R&D 100 Awards finalist in the Software/Services category.

Running the numbers

OpenMxP implements the HPL-MxP benchmarking task, which was introduced in 2019 and is the industry standard for measuring mixed-precision performance of supercomputing systems. HPL-MxP presents a problem to solve—a dense system of linear equations—but not the software to solve it. That's up to the benchmarkers. Previously, for the OLCF's Summit supercomputer, its GPU chip vendors developed and ran proprietary codes that evaluated their speed at making mixed-precision calculations.

"In the past, this kind of benchmark has always been run by the vendors or integrators. They develop their own code—it is their secret sauce to differentiate themselves and provide their own unique competitive advantage. They run it, and we take their result as is," Wang said.

This was not an option for Frontier because it is powered by AMD CPUs and GPUs, so new benchmarking codes would need to be developed to run properly on the next-generation chips. When NCCS Director Georgia Tourassi raised the question about whether ORNL could run the benchmark itself, Wang proposed forming a team to do just that. They began in April 2021 by studying the benchmark problem itself and consulting scientists who had worked on similar problems.

"We didn't have experience doing these kinds of problems, so we ran into issues with software stacks that we didn't expect. You just expect the Message Passing Interface is going to work like you think it's going to work in your head. But it didn't work that way," said Mike Matheson, the OpenMxP project's technical lead in the AAIMS group. "So, we

would try things, and then it wouldn't work, and then we'd talk to other people, and then we'd try something else. We were kind of just probing ahead, trying to figure out what actually worked. That was a learning curve—we just had to do it."

Fortunately, Frontier was still many months away from being completed. Unfortunately, that also meant they would be developing the code for a machine they couldn't test it on yet. But once OpenMxP was ready for its initial runs in mid-2021, the team did have another very fast—if not exascale fast—system nearby to serve as a test bed.

"Our target was the Frontier system, but Frontier at that point did not really exist. So, we leveraged what we had, which is Summit," Wang said. "It was actually a good thing, meaning that once we have a stable system to start with, got our code up and running, and tuned on Summit at this scale, our code was battle-tested. We knew it would scale up, and the rest was adapting it or preparing it for Frontier."

In May 2022, Frontier was ready to run. So was OpenMxP. Frontier's initial mixed-precision benchmark of 6.86 exaflops—or 6.86 billion billion floating point operations per second—put it at the top of the 2022 HPL-MxP list. One year later, it hit 9.95 exaflops for another first-place finish. The European High-Performance Computing Joint Undertaking's LUMI supercomputer also used OpenMxP to make its HPL-MxP submission and came in second place to Frontier in that June 2023 ranking.

A multipurpose tool

Putting a number to contest entries is not OpenMxP's true strength. The software package ultimately provides insights into how well GPU/CPU supercomputers are operating, which helps improve their performance by revealing how small changes in programming can lead to leaps in

computational speed. With the powerfully fast results it unveils, OpenMxP can also demonstrate to computational scientists the advantages of using GPU-equipped systems capable of mixed-precision calculations.

"Lots of simulations are solving large systems of equations, and traditionally it's all just double precision—researchers take the hammer out and go down the path they know. It made sense in the past because there was no special-purpose hardware," Matheson said. "But with the advent of all these GPUs that do low-precision calculations faster than CPUs, it has enabled this new solution process to be attractive."

Furthermore, OpenMxP can serve as a tool itself for solving certain problems in science and engineering at speeds and energy efficiency never before possible. The HPL-MxP benchmark problem that OpenMxP solves consists of large linear systems of equations, which are also the building blocks for science and engineering applications.

In 2022, an ORNL team—including NCCS research scientists Wang, Matheson, Hao Lu and Jens Glaser—used OpenMxP as a solver for TwoFold, a software stack that predicts how strongly a given drug molecule will bind to a pathogen and that predicts the 3D structure of how it will attach to the target. TwoFold was named a finalist for the 2022 Gordon Bell Special Prize for HPC-Based COVID-19 Research by the Association for Computing Machinery.

"The real thing we're trying to do is push science forward by giving scientists an open-source piece of software that they can build upon and modify to solve their science problems. Since we're at extreme scale and extreme size, OpenMxP can help tackle the biggest science problems in a faster way," Matheson said.

Provided by Oak Ridge National Laboratory

Citation: Researchers develop open-source mixed-precision benchmark tool for supercomputers (2023, September 25) retrieved 28 April 2024 from https://techxplore.com/news/2023-09-open-source-mixed-precision-benchmark-tool-supercomputers.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.