Mixed Precision: A Strategy for New Science Opportunities

Jonathan Hines
Oak Ridge National Laboratory

Editors:
James J. Hack
Oak Ridge National Laboratory
jhack@ornl.gov

Michael E. Papka
Argonne National Laboratory
papka@anl.gov

Since the days of vector supercomputers, computational scientists have relied on high-precision arithmetic to accurately solve problems. But changes to hardware, spurred by the demand for more computing capability and growth in machine learning, have researchers considering lower precision alternatives.

Since the days of vector supercomputers, computational scientists have relied on high-precision arithmetic to accurately solve a wide range of problems, from modeling nuclear reactors to predicting supernova physics to measuring the forces within an atomic nucleus.

But changes to hardware, spurred by the demand for more computing capability and growth in machine learning, have researchers rethinking the balance between the number of digits needed to perform a given calculation and computational efficiency. For portions of a calculation that do not require 64-b double-precision arithmetic—the longtime floating-point number standard in high-performance computing (HPC)—lower precision alternatives may provide enough accuracy. The tradeoff could lead to scientific discoveries that would otherwise remain years away.

A glimpse into this brave new mixed-precision world was previewed at the launch of the Summit supercomputer at the U.S. Department of Energy's (DOE's) Oak Ridge National Laboratory (ORNL) in June 2018. Named the fastest system in the world by the TOP500 list upon its debut, Summit derives 95% of its computing power from its more than 27 000 NVIDIA Tesla V100 GPUs. For standard double-precision problems, Summit's peak performance maxes out at 200 petaflops, or 200 million billion double-precision calculations per second. Application developers who can utilize low-precision arithmetic, however, will find that the IBM AC922 system has an extra gear—one that peaks at more than 3 exaops, or 3 billion billion mixed-precision calculations per second.

 Published by the IEEE Computer Society

This capability stems from a specialized NVIDIA integrated circuit called a tensor core designed to boost deep-learning research by executing a simple matrix operation quickly. By building the feature directly into hardware, NVIDIA gifted deep-learning researchers with a technology that can train and run neural networks several times faster than the speed that would otherwise be expected.

A part of the reason that tensor cores operate so quickly—around 16 times faster than standard computation—is because they support 16-b, half-precision arithmetic, a floating-point format that accommodates only a fraction of the digits compared with double precision. A secondary step of tensor cores' matrix operation runs at 32-b single precision. On Summit, researchers have already demonstrated the value of tensor cores by obtaining speeds surpassing 1 mixed-precision exaop for distributed neural networks.

Although the usefulness of tensor cores for supercharging low-precision deep learning is obvious, its relevance for flavors of scientific computing that require more accuracy remains less so. However, that has not stopped some computational scientists from experimenting with this new technology.

SEARCH FOR ACCELERATION

As one of the few researchers with early access to Summit, ORNL computational scientist Wayne Joubert has gotten a head start in this respect.

Since NVIDIA announced the tensor core GPU architecture in May 2017, Joubert has wondered if the feature could be useful for more than training neural networks. "Whenever hardware has some new feature, scientists are going to ask whether it is useful for their science," Joubert said.

Working as the lead methods developer for a comparative genomics project called Combinatorial Metrics (CoMet) in 2018, Joubert entertained the idea of accelerating the code to use Summit's low-precision capabilities to boost analysis of genomic datasets. Just a year earlier, he had ported CoMet to the Titan supercomputer, fine-tuning the application to take advantage of the Cray XK7's GPU acceleration. The resulting speedup contributed to research led by ORNL computational systems biologist Dan Jacobson on regulatory genes of plant cell walls that can be manipulated to enhance biofuels and bioproducts.

One CoMet algorithm that Joubert thought might be particularly well suited for tensor cores was the custom correlation coefficient (CCC) method, which specializes in comparing variations of the same genes, known as alleles, present in a given population. Tensor core-enhanced performance of CCC could potentially allow researchers to analyze datasets composed of millions of genomes—an impossible task for current leadership-class systems—and study variations among all possible combinations of two or three alleles at a time (see Figure 1). Scientists could then use this information to uncover hidden networks of genes in plants and animals that contribute to observable traits, such as biomarkers for drought resistance in plants or disease in humans.

In conversations with colleagues on the systems biology team and in HPC circles, Joubert discussed the CCC algorithm in depth. He studied the algorithm's description in old research papers and considered possible reformulations that could map to low-precision hardware. "It's going back to mathematics and asking if there is some way to rearrange the pieces and get the same result," Joubert said.

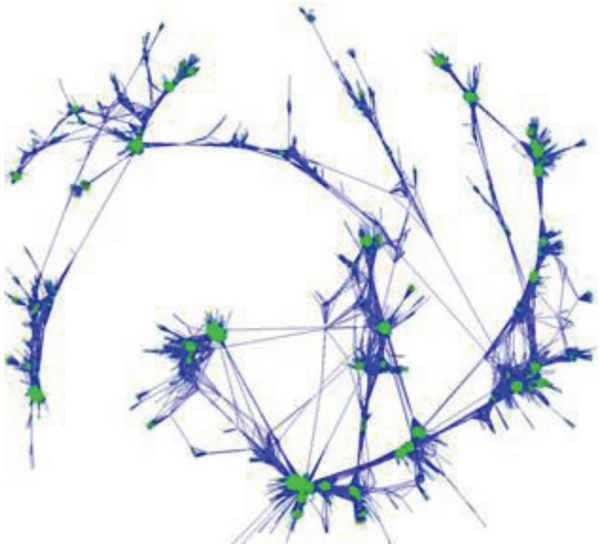


Figure 1. One component of a correlation network mapping variations in single nucleotides that occur at the same location in the genome across a population. These correlations can be used to identify genetic markers linked to complex observable traits. Image: Dan Jacobson, Oak Ridge National Laboratory.

ACHIEVING EXAOPS

Mixed-precision arithmetic is not new to HPC. Scientists have been using it regularly to selectively boost application performance when it makes sense. For example, in 2008, an ORNL team achieved the first sustained petaflop simulation using a materials science application called DCA++ that utilized a combination of single and double precision. In that instance, the team employed single precision for a portion of its code concerning a quantum Monte Carlo calculation that solved embedded material clusters and did not require double precision. The arrangement resulted in a twofold speedup for that portion of the code—the typical performance gain expected in the switch from double to single precision.

New types of accelerators such as tensor cores, however, raise the reward substantially for application developers who can successfully recast their problem. Instead of factors of two, the potential exists for acceleration approaching factors of 10. For Joubert, a deep dive into CCC's mathematics—examining how the genomic data translates into bits, or sequences of numbers, and how these sequences are compared across a population—proved to be the key.

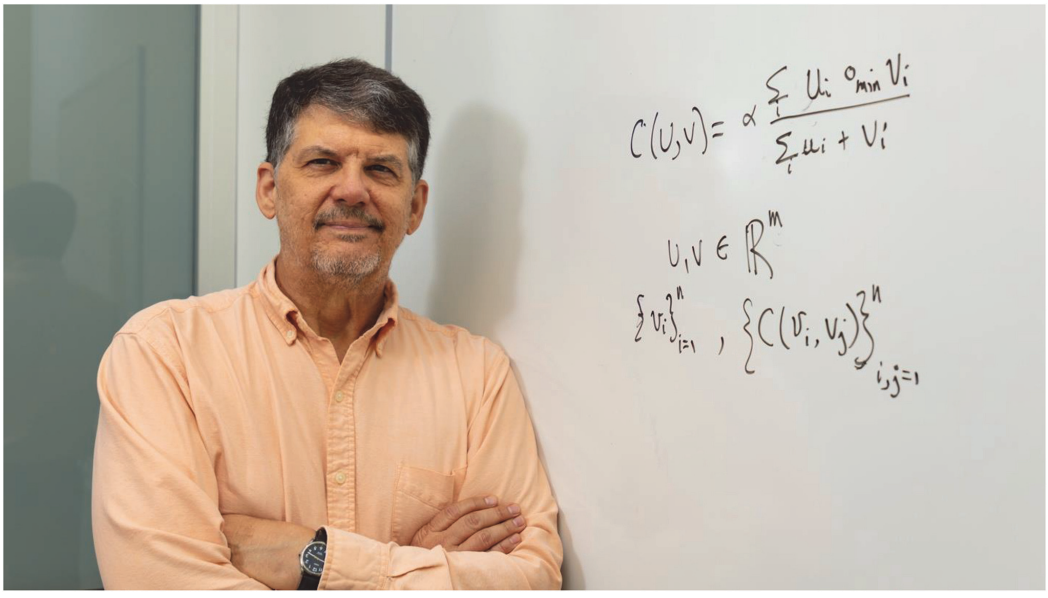
"The heart of the method is to count the number of occurrences of certain combinations of these 2-b pairs, which represent different alleles within a gene," Joubert said. "You compare the bits, get counts of the number of occurrences of these combinations across your population, and calculate the result. It occurred to me that we can actually take these 2-b values and insert them into the half-precision format."

Normally, mapping 2 b into a 16-b floating-point number would be inefficient—like ferrying boat passengers two at a time on a vessel that could otherwise hold the entire group. Joubert, however, intuited that the speed of the tensor cores would more than make up for this unconventional data packaging.

From the moment inspiration struck him, Joubert spent 2 weeks writing his idea into code. He spent another month optimizing it. The performance gains on Summit became apparent immediately—delivering a machine-to-machine 37-fold speedup compared with the Titan implementation. Testing the algorithm across 4000 nodes using a representative dataset, the genomics application team achieved a peak throughput of 1.88 mixed-precision exaops—faster than any previously reported

science application. With some additional fine-tuning, the algorithm leaped to 2.36 mixed-precision exaops running on 99% of Summit (4560 nodes). The final result was a rate of science output four to five orders of magnitude beyond the current state of the art. The journey that started with a low-precision puzzle had opened a new frontier in a comparative genomics analysis.

Joubert said close collaboration with Jacobson and his systems biology colleagues played a large role in the team's technical triumph. "We speak different languages," he said. "At the start of the project I didn't know anything about their science problem and they knew very little about GPUs. I think you need to have multidisciplinary interaction to find these new opportunities."



Wayne Joubert. Image: Carlos Jones, Oak Ridge National Laboratory.

HOW LOW CAN HPC GO?

With machine learning continuing to drive processor architectures, opportunities for researchers to push into lower precisions will likely expand in the future. Low-precision calculations have the potential to benefit not only deep learning and data science applications but also modeling and simulation, where linear algebra dominates computation.

Developing HPC libraries that allow scientists to automatically exploit lower precisions with minimal or no additional work is an ongoing area of research that could dramatically expand use of mixed precision. The Innovative Computing Laboratory (ICL) at the University of Tennessee, Knoxville (UTK) is playing a leading role in creating such a tool.

In 2007, ICL, which was founded by UTK–ORNL distinguished researcher Jack Dongarra, added an algorithm to its Linear Algebra PACKage (LAPACK) that could automatically switch between double and single precision when advantageous. The technique involves the generation of a fast low-precision solver, followed by an iterative process to derive high-precision accuracy.

"What you get is a solver that is as fast as single precision with the accuracy of double precision," said Stan Tomov, an ICL research director and a UTK research assistant professor.

Extending this capability to a tensor core-like accelerator that operates at even lower precisions adds an additional series of mathematical challenges. ICL has made progress nonetheless, developing new mixed-precision solvers that can quadruple the speed of conventional double-precision solvers. An ICL-led paper on this topic, presented at the 2017 International Conference for High Performance Computing, Networking, Storage and Analysis (SC17), and a poster shared at the 2018 ISC

High Performance Conference have been downloaded more than 12 000 times, an indication of the high interest surrounding low-precision solvers.

“The bottom line is that there are a number of algorithms already that have proven to be successful,” Tomov said. “As more people get interested, it becomes more likely that there will be even further progress.”

In the meantime, computational scientists like Joubert continue to experiment with the latest architectures that technology vendors put on the market, searching for promising strategies to harness hardware in the name of science’s most pressing questions.

“To run our science faster and move forward, I think we need to be open to all kinds of ideas,” Joubert said.

ACKNOWLEDGMENTS

The Oak Ridge Leadership Computing Facility is a DOE Office of Science User Facility supported under contract DE-AC05-00OR22725. This manuscript has been authored by UT-Battelle, LLC, under contract number DE-AC05-00OR22725 with the U.S. Department of Energy. The U.S. government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

ABOUT THE AUTHOR

Jonathan Hines is a science writer at Oak Ridge National Laboratory. He received his bachelor’s degree in journalism from Indiana University. Contact him at hinesjd@ornl.gov.

This article originally appeared in
Computing in Science & Engineering, *vol. 20, no. 6, 2018.*