Mixing precision for model acceleration

July 27 2021

The international team of researchers drew upon the power of three of the world's most powerful supercomputers: HAWK in Germany, Shaheen-II in KAUST and Summit (the second most powerful computer in the world). Credit: KAUST

A mixed-precision approach for modeling large geospatial datasets can achieve benchmark accuracy with a fraction of the computational run time.

By applying high-precision calculations only where they're needed most, a KAUST-led research team has been able to significantly speed up

modeling of large geospatial datasets without overall precision loss. The approach, implemented on a high-performance computing system based on highly parallelized graphics processor units (GPUs), will allow larger datasets to be analyzed in shorter time.

While computers have the capacity to perform very large calculations very quickly, the result can sometimes be less precise than hand calculations because of the limitation of how numbers are stored in digital systems. Standard or "single" precision numbers effectively have 6–9 actual decimal digits of precision, meaning that any calculation resulting in a longer sequence of digits will be truncated, thereby losing information.

While double-precision numbers can be used, this doubles the memory and calculation intensity. For geospatial datasets where the accumulation of such precision errors can lead to erroneous modeling results, this imparts a major limitation on the size of dataset that can be calculated precisely.

Sameh Abdulah and colleagues including Hatem Ltaief, Marc Genton, Ying Sun and David Keyes from KAUST, in collaboration with researchers from the University of Tennessee, Knoxville (UTK) in the U.S., have now developed an elegant solution to this problem by mixing precision as required.

"For decades, modeling of environmental data relied on double-precision arithmetic to predict missing data," says Abdulah. "Today, there is high-performance computing hardware that can run single- and half-precision arithmetic with a speedup of 16 and 32 times compared with double-precision arithmetic. To take advantage of this, we propose a three-precision framework that can exploit the acceleration of lower precision while maintaining accuracy by using double-precision arithmetic for vital information."

Using the PaRSEC runtime system developed by UTK, which allows for on-demand precision and the orchestration of tasks and data movement across multiple parallel GPUs, the researchers exploited the statistical relationships in the data to reduce precision for weakly correlated spatial locations to single- or half-precision based on distance.

Double-precision calculations are only applied for the strongly correlated locations that have the most influence on model accuracy.

"The main goal of this project is to leverage the recent parallel linear algebra algorithms developed by KAUST's Extreme Computing Research Center to scale up geospatial statistics applications on leading-edge parallel architectures," says Abdulah.

"We have shown that we can achieve significant speedup compared to full double-precision arithmetic modeling while preserving the parameter estimations and prediction accuracy to meet the application requirements," he explains. "Next, we intend to integrate approximations with mixed precision to further reduce memory footprint and shorten calculation time."