TechXplore

Math primes high-performance computing for the age of AI

May 24 2023, by John Roach

To overcome high-performance computing bottlenecks, the PNNL research team proposed using graph theory, a mathematical field that explores relationships and connections between a number, or cluster, of points in a space. Credit: Shannon Colson, Pacific Northwest National Laboratory

Increasing traffic congestion in the Seattle area is a good analogy for a similar increase in congestion on high-performance computing (HPC) systems, according to scientists at Pacific Northwest National Laboratory (PNNL).

More complex workloads, such as training artificial intelligence (AI) models, are to blame for the HPC bottlenecks, the scientists say in a paper published in *The Next Wave*, the National Security Agency's review of emerging technologies.

"We can solve the congestion through how we create the network," said Sinan Aksoy, a senior data scientist and team leader at PNNL who specializes in the mathematical field of graph theory and complex networks.

In HPC systems, hundreds of individual computer servers, known as nodes, work as a single supercomputer. The arrangement of the nodes and links between them is the network topology.

HPC congestion occurs when the exchange of data between nodes funnels onto the same link, creating a bottleneck.

HPC system bottlenecks are more common today than they were when the systems were designed, as Aksoy and his colleagues Roberto Gioiosa, a computer scientist in the HPC group at PNNL, and Stephen Young, a mathematician in the math group at PNNL, explain in *The Next Wave*.

That's because the way people use HPC systems today is different than the way they did when the systems were developed.

"This is an artifact of life changing," said Gioiosa. "We didn't have Facebook 20 years ago, we didn't have this big data, we didn't have big AI models, we didn't have ChatGPT."

Big tech expands

Starting in the 1990s, the computer technology industry began to blossom. New companies disrupted the Seattle area's economy and where people live and work. The resulting traffic patterns became less predictable, less structured, and more congested, especially along the east-west axis that constrains traffic to two bridges across Lake Washington.

Traditional HPC network topologies resemble the Seattle area road network, according to the researchers at PNNL. The topologies are optimized for physics simulations of things such as the interactions between molecules or regional climate systems, not modern AI workloads.

In physics simulations, the calculations on one server inform the calculations on adjacent servers. As a result, network topologies optimize the exchange of data among neighboring servers.

For example, in a physics simulation of a regional climate system, one server might simulate the climate over Seattle and another the climate over the waters of the Puget Sound to the west of Seattle.

"The Puget Sound climate model is not going to affect what's going on in New York City–I mean, it is eventually–but really it needs to talk to the Seattle model, so I might as well hook the Puget Sound computer and the Seattle computer right next to each other," said Young, a mathematician in PNNL's computational math group.

The communication patterns in data analytics and AI applications are irregular and unpredictable. Calculations on one server may inform calculations on a computer across the room. Running those workloads on traditional HPC networks is akin to driving around the greater Seattle region today on a scavenger hunt at rush hour, according to Gioiosa.

Network expansion

To overcome HPC bottlenecks, the research team at PNNL proposed using graph theory, a mathematical field that explores relationships and connections between a number, or cluster, of points in a space.

Young and Aksoy are experts in expanders, a class of graphs that can spread network traffic so that "there's always going to be a lot of options to get from point A to point B," explained Aksoy.

Their network, called SpectralFly, exhibits perfect mathematical symmetry—every node is connected to the same number of other nodes, and the connections from each node look the same throughout the network.

The options to get from one node to another—with each option identical to any node in the network—also mean it's easier for computer programmers to route information through the network, added Aksoy.

"It's the same roadmap no matter where you are, so it's a lot less computationally expensive to figure out how to route information on top of this network," he said, noting that this feature is like being in a city where the directions from any neighborhood to all destination neighborhoods are the same for any starting point.

Simulation results

The research team at PNNL ran simulations of their SpectralFly network across workloads from traditional physics-based simulations to training AI models and compared the results to those from other types of HPC network topologies.

They found that SpectralFly outperformed other network topologies on modern AI workloads and achieved comparable performance on traditional workloads, indicating that it could serve as a hybrid topology for people looking to do traditional science and AI on the same HPC system.

"We are trying to merge the two worlds, the traditional and the emerging ones in a way that we can still do science and we can also do AI and big data," said Gioiosa.

Provided by Pacific Northwest National Laboratory

Citation: Math primes high-performance computing for the age of AI (2023, May 24) retrieved 26 April 2024 from
https://techxplore.com/news/2023-05-math-primes-high-performance-age-ai.html