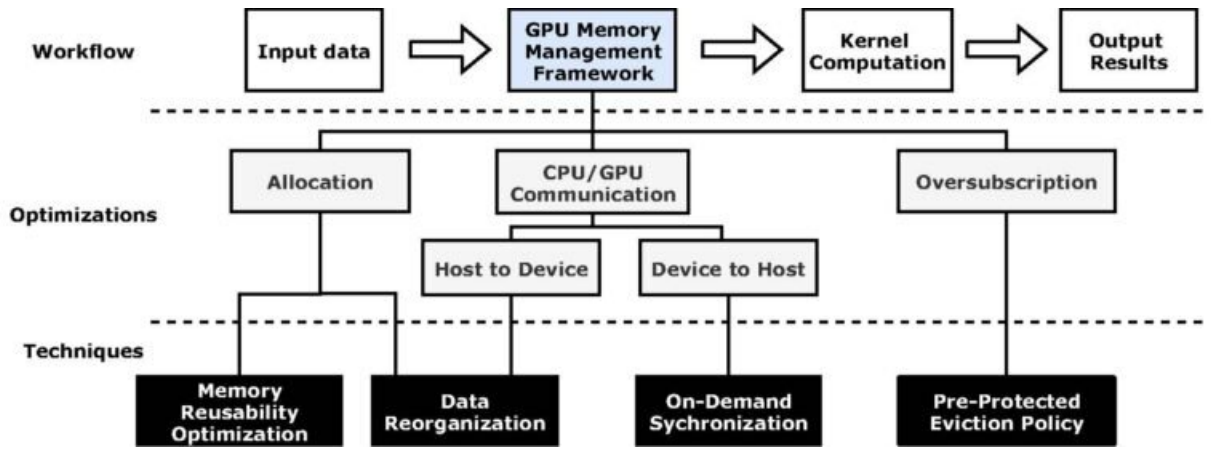
MemHC: An optimized GPU memory management framework for accelerating many-body correlation functions

November 15 2022, by Chris Patrick



MemHC incorporates three optimized memory management methods to eliminate redundant memory operations and enhance data reusability. Credit: Bin Ren, William & Mary

Computers help physicists solve complicated calculations. But some of these calculations are so complex that a regular computer is not enough. In fact, some advanced calculations tax even the largest supercomputers. Now, scientists at the U.S. Department of Energy's Thomas Jefferson National Accelerator Facility and the College of William & Mary have developed a new tool to ensure they get the most bang for their buck out of precious supercomputer time. They recently published details of the

work in the journal *ACM Transactions on Architecture and Code Optimization*.

The calculations in question are related to quantum chromodynamics (QCD). QCD is the theory that describes the structures of neutrons and protons—particles that make up atomic nuclei—and other hadrons. An approach used to calculate QCD, called lattice QCD, involves a type of calculation known as many-body correlation functions. These calculations are so complicated that they require the enhanced power of a supercomputer to solve them.

"If you used your desktop, it would take years to compute them," confirmed Bin Ren, associate professor of computer science at the College of William & Mary. "We cannot wait for that."

Even with supercomputers, many-body correlation functions are still time intensive. That's why Ren worked with researchers at Jefferson Lab to make these calculations more efficient.

Together, the collaboration of Jefferson Lab and William & Mary researchers developed a memory-management framework called MemHC. This framework organizes the memory of a graphics processing unit (GPU). In a gamer's desktop computer or a dedicated gaming system, GPUs quicken the rendering of graphics for smoother play. It turns out that GPUs can also accelerate complicated calculations in supercomputers.

One challenge in using GPUs in this way is that GPUs have limited memory. While a GPU might only have tens of gigabytes of memory, many-body correlation functions may require hundreds of gigabytes—or even terabytes. This means information must be frequently passed on and off the GPU to and from its host, the central processing unit (CPU), leading to inefficiencies.

MemHC addresses these inefficiencies to allow a GPU to calculate many-body correlation functions more quickly.

Three solutions

The main reason many-body correlation functions are so complicated is because each of these calculations contain an enormous number of complex operations known as tensor contractions. While previous work has optimized the computation of a single tensor contraction, lattice QCD involves a lot of them.

"Our problem is, if we want to calculate thousands of tensor contractions, how should we do that efficiently?" Ren said.

He worked with Jie Chen, a senior computer scientist at Jefferson Lab, and two Ph.D. students at William & Mary, Qihan Wang and Zhen Peng, to answer that question. They first identified the issues affecting a GPU's performance during these operations.

"Through rigorous discussions, performance evaluations, and literature studies, we came up with optimization ideas addressing the issues," Chen said.

They developed three memory management methods that reduce redundancy in the system to accelerate the calculation of tensor contractions.

Tensor contractions cannot be calculated at the same time to save power because they have a dependent relationship with each other: the answer of the first tensor contraction is used to calculate the second, and so on. The relationship of these tensor contractions forms over 100,000 graphs during calculation—that's 100,000 separate drawings—and that number can reach to nearly a million.

The researchers assessed the edges of these graphs describing tensor contraction calculations and noticed many of them overlapped. This meant the GPU was essentially calculating the same thing multiple times. These are known as redundant memory allocation operations, because they eat up more memory on the GPU than necessary.

"By realizing there are lots of passes overlapped with each other, we eliminated such kind of redundant computation," Ren said.

So instead of calculating the same quantity over and over, which would require the GPU to allocate that memory, then deallocate and reallocate it many times, the scientists coded the framework in such a way that a memory will persist on the GPU in a manner better suited to the calculations.

"And the GPU will reuse that memory again and again," he explained.

After reducing the number of input and output tasks of the GPU, the researchers focused on communication between the GPU and its host CPU.

The CPU only needs the result from the GPU, not the many intermediate computational steps leading up to it. The researchers eliminated the GPU's unnecessary communication with the CPU, which reduced redundant data transfer.

And finally, since the limited memory of a GPU does not meet the requirement for these types of computations, the team programmed MemHC to move data from the CPU onto the GPU only as needed, and back to the CPU when it's no longer needed.

"If we know which data are required and when we require this data, then we can optimize this data movement procedure," Ren said.

These three memory optimization methods allow a GPU to calculate many-body correlation functions ten times faster.

Moving forward with MemHC

"MemHC proposes advanced and novel solutions to efficiently reduce the significant time and memory expense of these calculations, which cannot be solved by prior related work," Wang said.

The team used MemHC through software called Redstar, which was developed by Chen and Robert Edwards, a senior staff scientist at Jefferson Lab and the principal investigator of the MemHC project, to evaluate many-body correlation functions for lattice QCD. Redstar is part of the DOE's Exascale Computing Project, which seeks to develop exascale supercomputers capable of quintillion calculations per second.

"MemHC is a very successful collaborative effort between William & Mary and Jefferson Lab," Chen said. "William & Mary provides much-needed expert knowledge in computer science to improve the performance of physics code, and as the result of the collaboration, Jefferson Lab has more lattice QCD physics calculations done and expects more physics results published. In addition, Jefferson Lab offers valuable software and hardware platforms for William & Mary computer science researchers to tackle real-world problems and to deliver world class research results."

In this work, MemHC was outfitted for a single GPU and CPU. More recently, the authors have been extending this framework to multiple GPUs.

Soon, they'll further extend it to clusters of CPUs and GPUs. The more clusters there are using MemHC, the faster physicists can compute calculations related to lattice QCD to better understand the properties of

protons and neutrons.

More information: Qihan Wang et al, MemHC: An Optimized GPU Memory Management Framework for Accelerating Many-body Correlation, *ACM Transactions on Architecture and Code Optimization* (2022). DOI: 10.1145/3506705