Qeios

Peer Review

Review of: "Performant Automatic BLAS Offloading on Unified Memory Architecture with OpenMP First-Touch Style Data Movement"

Ruobing Han[1]

1. Georgia Institute of Technology, United States

This work introduces a software solution that automatically updates off-the-shelf CPU BLAS programs to leverage GPU computational resources. The author presents a lightweight approach that does not require manual code modifications and seamlessly improves runtime performance. To further exploit the unified memory features in advanced GPU architectures, the author proposes the “Device First-Use Policy,” which significantly improves runtime compared to two other memory management solutions.

The author evaluates the solution using two BLAS benchmarks and discusses the relationship between data size/shape and performance. These results are valuable for other researchers in selecting suitable memory management strategies.

The discussion on performance issues in Section 4.4 is particularly interesting. The author provides data to highlight potential GPU memory and/or software issues that can degrade performance and discusses methods to overcome these limitations (e.g., using different page sizes and alignments).

I hope the author can also provide artifacts that would be helpful for CPU BLAS users.

From a writing perspective, I believe the two paragraphs in Section 4.4.3 are repetitive.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/5TZMH8