Qeios

Peer Review

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

Dominik Ernst[1]

1. NHR@FAU, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

Some sentences appear twice after each other:

“then the resultant matrix is copied back after cuBLAS execution.”

“This strategy works on all GPUs, including the conventional PCIe-based cards, without needing unified memory capability,”

As far as I am aware, access counters are a feature introduced with Volta already.

If I understand correctly: Buffers touched by an intercepted BLAS call are automatically and permanently moved to the GPU, instead of being copied back and forth. I assume this works well for applications that can generally profit from BLAS call interception because all work is done in BLAS routines. In that case, what the library does feels like the only sensible choice.

The anti-case would be an application that has a single BLAS call but then does a lot more computation outside of BLAS routines. That computation would happen on the CPU, which now has to access the slower device RAM.

I like the concept; I dislike the analogy to OpenMP's first touch. The CPU might have touched the page plenty before the first BLAS call, but the automatic BLAS offload wrappers only see BLAS calls.

doi.org/10.32388/AoNLYE

Declarations

Potential competing interests: No potential competing interests to declare.