

Peer Review

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

Antonio J. Peña¹

1. Computer Sciences, Barcelona Supercomputing Center, Barcelona, Spain

This paper proposes a very nice abstraction to GPU BLAS for Grace-Hopper (and other unified-memory designs) that (1) respects the original BLAS API and (2) does smart data movement based on a proposed "device first-use data movement policy". This policy proves to be effective on a couple of relevant use cases in HPC.

The topic is timely, and the storyline is nice. The writing could still be improved a little, but the shortcomings are minor and don't distract from reading.

IMHO, Figure 1 doesn't really add much :)

I like the proposal. I believe the author should be clearer when comparing that the baseline `cudaMemcpy` doesn't have to be as bold as in Listing 1, since one could (should) easily keep track of when the memory has to be moved and not move it back and forth every single time - in fact, that's precisely why `cublasXXX` doesn't include automatic data transfers, and it's the main benefit of explicit data movements vs. UM. The author is very much aware of this and is not dishonest, but I believe it would benefit readers to clarify this ahead and also to add some experimentation to clarify how far performance is from manually-tuned data movements and that what's provided is a very convenient way of avoiding the manual movements by implementing a heuristic that proves to work fairly well in common scenarios - the author has some discussion toward this at the end, focusing this mechanism for proof-of-concept, but I believe adding a little bit more on this would add more clarity to what the contribution is. While MuST has some native support, this paper proves it's not as optimized as it should be, since hand-crafted memory management can always be at least as good as a heuristic.

Minor: in “InfiniBand,” the “B” should be capitalized.

Declarations

Potential competing interests: No potential competing interests to declare.