

Peer Review

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

Basilio B. Fraguera¹

1. Universidad de La Coruña, Spain

The manuscript presents a novel tool to offload BLAS operations to a GPU in an effortless fashion for users, as it is automatically performed through the interception of BLAS symbols thanks to dynamic binary instrumentation. The most novel part of the proposal is the mechanism proposed for data management, which is always more complicated when we have to deal with accelerators, as it relies on a first-touch policy similar to that of OpenMP in unified memory architectures such as Grace-Hopper. The tests performed achieve very good results, and they identify a critical improvement point for NVIDIA, which, if solved, would further improve the performance of the proposal.

While at the scientific level the contribution is fine, the writing could certainly be improved. There are not only sentences but even full paragraphs that are a repetition of a preceding one. In addition, there is a large number of problems with the language, mostly in the form of missing words. I provide publicly a PDF where the repetitions and language issues are marked for the benefit of the authors so that this helps them enhance these aspects of the paper. The document is located at <https://rb.gy/sk8yoo>

Other improvements that can be made in the paper involve proposing some kind of future work and trying to identify scenarios where the first-touch policy may not be the best option, and then discussing which other possible automatic policies would be more suitable in those situations. This discussion is proposed to be performed only as a theoretical analysis, not requiring implementations and tests, which could be part of a new different contribution from the authors.

Declarations

Potential competing interests: No potential competing interests to declare.