

Peer Review

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

Daniel Abdi¹

1. National Oceanic and Atmospheric Administration, Washington, D.C., United States

The paper presents *SCILIB-Accel*, a novel automatic BLAS offloading tool designed for unified memory architectures like NVIDIA Grace-Hopper. The key contribution is the Device First-Use policy, an efficient data management strategy inspired by OpenMP's First-Touch approach, aimed at minimizing CPU-GPU data transfers. The tool uses dynamic binary instrumentation (DBI) to intercept and replace BLAS calls at runtime, requiring no modifications to existing code. The performance evaluation on scientific applications like MuST and PARSEC demonstrates a substantial 3x speedup.

I believe the manuscript is a great contribution to the existing literature on the automatic offloading of BLAS and LAPACK type libraries. The device-first use policy is a significant improvement over the traditional mem-copy based method, which may not yield any performance improvements due to the memory transfer bottleneck on every BLAS call. Having said that, the onus is on the developer of a program to use BLAS routines exclusively to benefit from this approach. Some legacy codes (e.g., weather models) that do use BLAS extensively also have compute-intensive code that does not use BLAS. In this case, the performance gain from accelerating the BLAS kernels alone may be wiped out by code that is running on the CPU due to Amdahl's law and the additional need to transfer data between the CPU and GPU.

As other reviewers have suggested, the writing could certainly be improved. There are grammatical errors that still need to be fixed, and some statements and sections are repetitive, particularly the discussion of unified memory. Some of the tables, such as Table 4, are better presented with a figure, in my opinion. I also find the overuse of the first person "I" distracting.

Also, please explicitly state whether this work is limited to Grace-Hopper or generalizable, and how performant it would be for non-BLAS heavy legacy codes. Detailed discussion on the downsides of the device first-use policy in terms of memory overhead, memory fragmentation when migrating large arrays to GPUs, NUMA effects, and how this method scales beyond current tests would be helpful too.

Declarations

Potential competing interests: No potential competing interests to declare.