Qeios

Peer Review

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

Yuri Torres[1]

1. Universidad de Valladolid, Spain

General Evaluation

The paper **"Performant Automatic BLAS Offloading on Unified Memory Architecture with OpenMP First-Touch Style Data Movement"** presents an optimization for automatic BLAS offloading in unified memory architectures using the OpenMP First-Touch approach. The study introduces SCILIB-Accel, a tool that improves data movement management between CPUs and GPUs. This research is relevant to the high-performance computing (HPC) community and addresses a critical challenge in GPU offloading.

However, there are several areas that require **major revision** before the paper can be considered for publication. Below, I outline the strengths and the key areas that need improvement.

1. Contribution and Originality

Strengths:

- The **Device First-Use Policy** is novel and well-explained.
- The focus on **data movement optimization in Grace-Hopper architectures** is well-motivated and addresses a clear need in HPC.
- A detailed comparison of different BLAS offloading strategies is provided.

Areas for Improvement:

- The originality of SCILIB-Accel compared to existing tools like NVBLAS and Cray LIBSCI_ACC could be discussed in more depth. What are the fundamental differences in structure and

 doi.org/10.32388/FE5LAC

implementation?

- The paper does not discuss **potential limitations** of the proposed methodology in large-scale production environments beyond the testbed.

2. Clarity and Organization

Strengths:

- The paper follows a clear structure with well-defined sections (introduction, related work, methodology, experiments, conclusions).
- The different memory management strategies are introduced progressively and understandably.

Areas for Improvement:

- Some sections are **redundant**. For example, the comparison of data movement strategies is repeated across multiple sections with little new information.
- A **Limitations and Future Work** section is missing. This section should address:
  - The applicability of the approach to other CPU-GPU architectures.
  - Scenarios where the strategy may not perform well, such as extremely large matrices or workloads with high latency.

3. Methodology and Validity of Results

Strengths:

- The paper evaluates performance using **two real-world scientific applications (MuST and PARSEC)**, which strengthens its credibility.
- Performance metrics such as execution times and data transfer analysis are well presented.

Areas for Improvement:

- There is **no statistical analysis** of the results. It is recommended to include:
  - Minimum, maximum, and standard deviation of execution times.
  - Statistical significance tests if applicable.
- The paper does not mention **how many times each experiment was repeated** or whether results were verified for stability.

4. Comparison with Previous Work

Strengths:

- The paper references several previous tools (Cray LIBSCI_ACC, NVBLAS, IBM ESSL) and compares them to SCILIB-Accel.

Areas for Improvement:

- A **comparative table** summarizing key differences between SCILIB-Accel and other tools would improve clarity.
- The discussion of previous approaches is **more narrative than analytical**. A more technical comparison would strengthen the argument.

5. Writing Style and Readability

Strengths:

- The writing is clear and free of major grammatical errors.
- The use of technical terminology is appropriate.

Areas for Improvement:

- Some sentences are **too long and complex**. It would be helpful to break them into shorter sentences to improve readability.
- Certain key terms, such as *automatic offloading* and *data movement*, are **overused**. Varying terminology would prevent redundancy.

6. References and Citations

Strengths:

- The paper includes relevant and up-to-date references, including official NVIDIA documentation and recent research articles.

Areas for Improvement:

- The paper lacks citations from **high-impact journals or HPC conference proceedings**. It is recommended to add recent studies on BLAS offloading efficiency in modern GPUs.
- Some references appear isolated in the text without sufficient context.

Final Recommendation: Major Revision

Since the paper presents a relevant contribution but has weaknesses in **comparison with previous work, statistical analysis, and clarity of organization**, I recommend a **major revision** before publication.

Required Actions:

1. **Expand the comparison with previous tools** to highlight key technical differences and advantages.
2. **Include a statistical analysis** of the experiments to reinforce the validity of the results.
3. **Eliminate redundant sections** and improve the organization of the discussion.
4. **Add a Limitations and Future Work section** to discuss potential drawbacks and future research directions.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/FE5LAC