

Optimizing GPU-Accelerated Similarity Joins: Addressing Data-Dependent Workloads

Benoît Gallet, Michael Gowanlock

benoit.gallet@nau.edu, michael.gowanlock@nau.edu

School of Informatics, Computing and Cyber Systems, Northern Arizona University



Introduction

Given D a dataset in n dimensions

- Distance similarity self-join \rightarrow Find pairs of objects in D whose distance is within ϵ
- Use a grid indexing to prune the search space
- Thread _{i} = Query _{i}

- Depending on data characteristics \rightarrow Workload between threads varies a lot
- GPU's architecture (SIMT) \rightarrow Idle periods for threads with less workload
 - Results in higher execution time

Solution

- Balance the workload between threads to reduce the response time
- Sort the points by workload, from most to least
 - Consecutive threads should have similar workload (Fig. 1)
- However \rightarrow GPU's hardware scheduler controls the execution order of the threads
- Use a work queue to force this execution's order
 - Atomic operation to retrieve the first element of the queue, i.e., the one with most workload (Fig. 2)

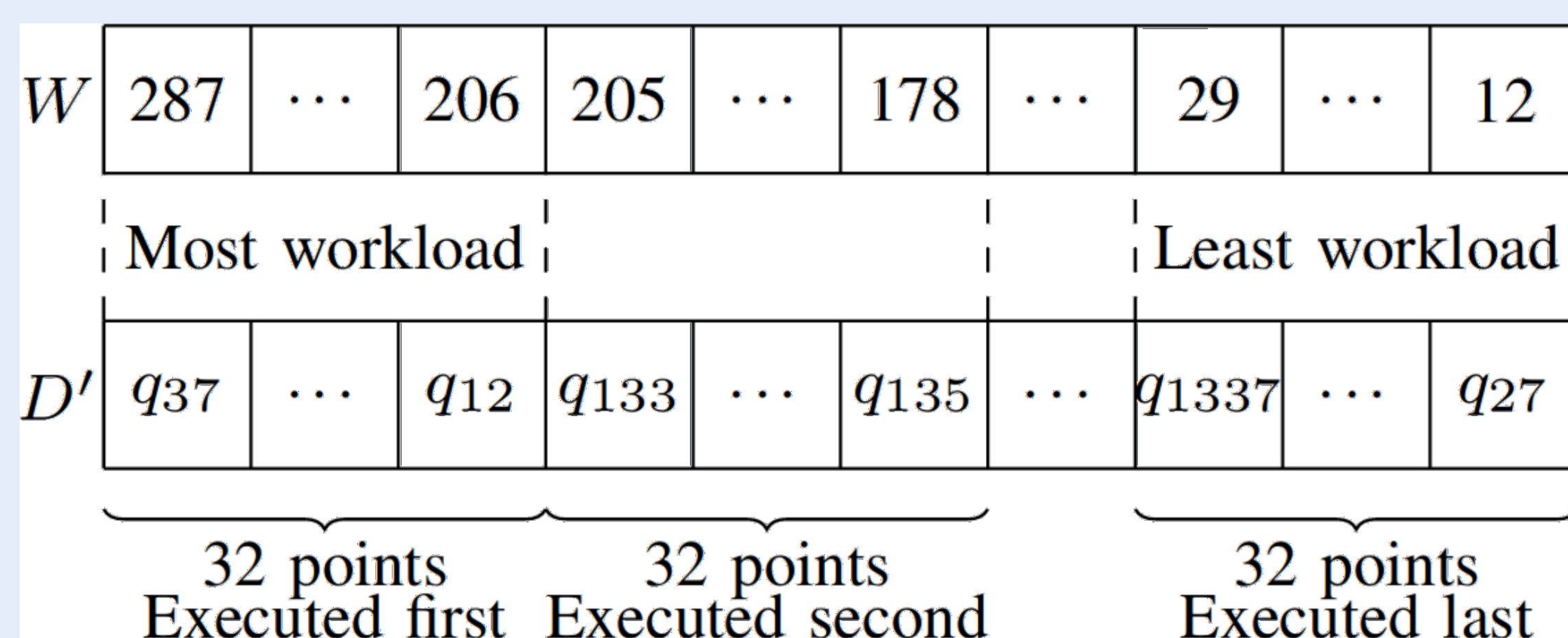


Figure 2

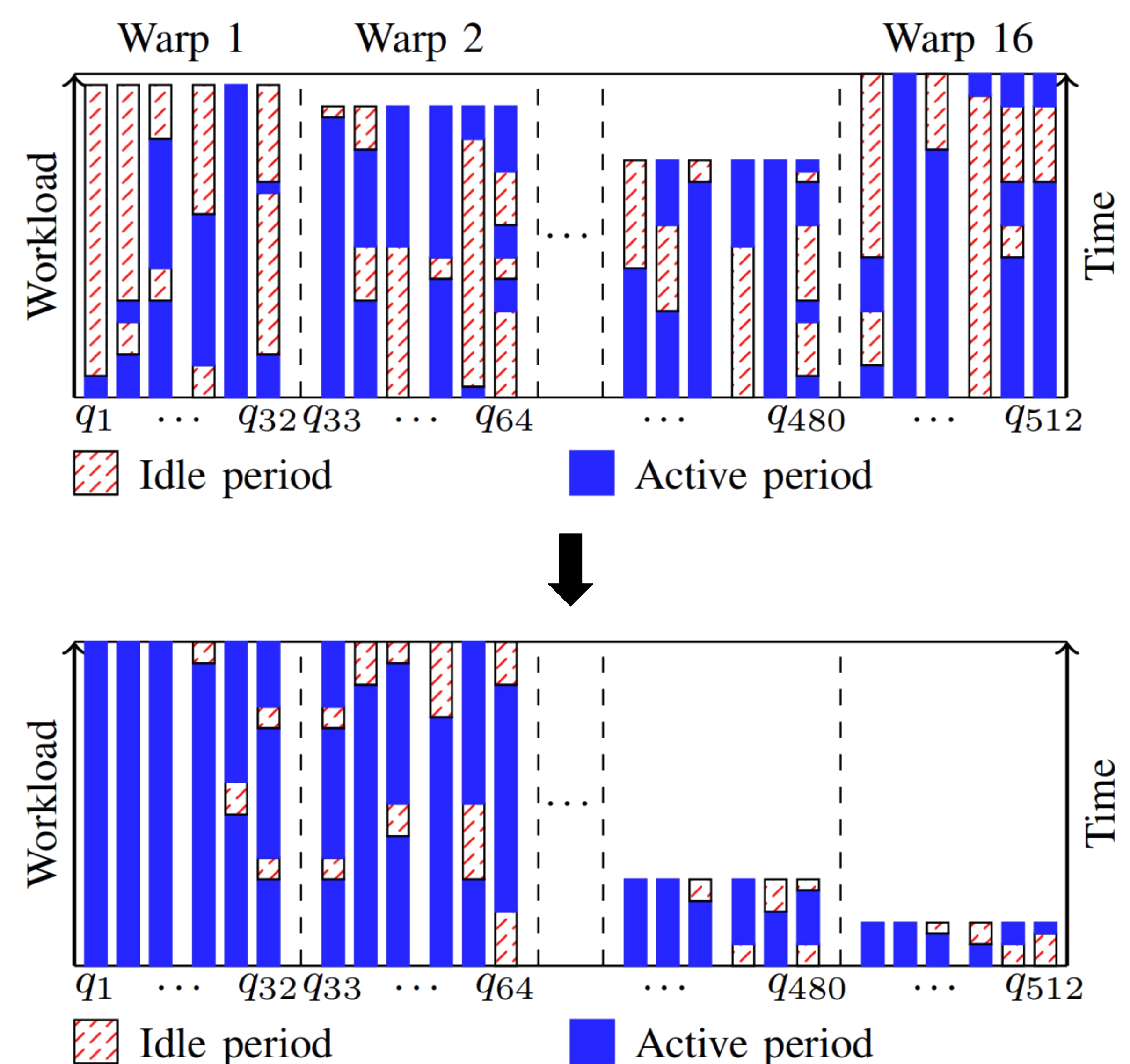


Figure 1

Results

- Compare GPUCalcGlobal [1], sorting by workload (SortByWL) and our work queue (WorkQueue) [2]
- Focus on the execution time and warp execution efficiency (WEE)
 - Percentage of active threads within a warp \rightarrow higher is better
- Uniformly (Fig. 3) and exponentially (Fig. 4) distributed synthetic datasets, $n = 6$, 2M points

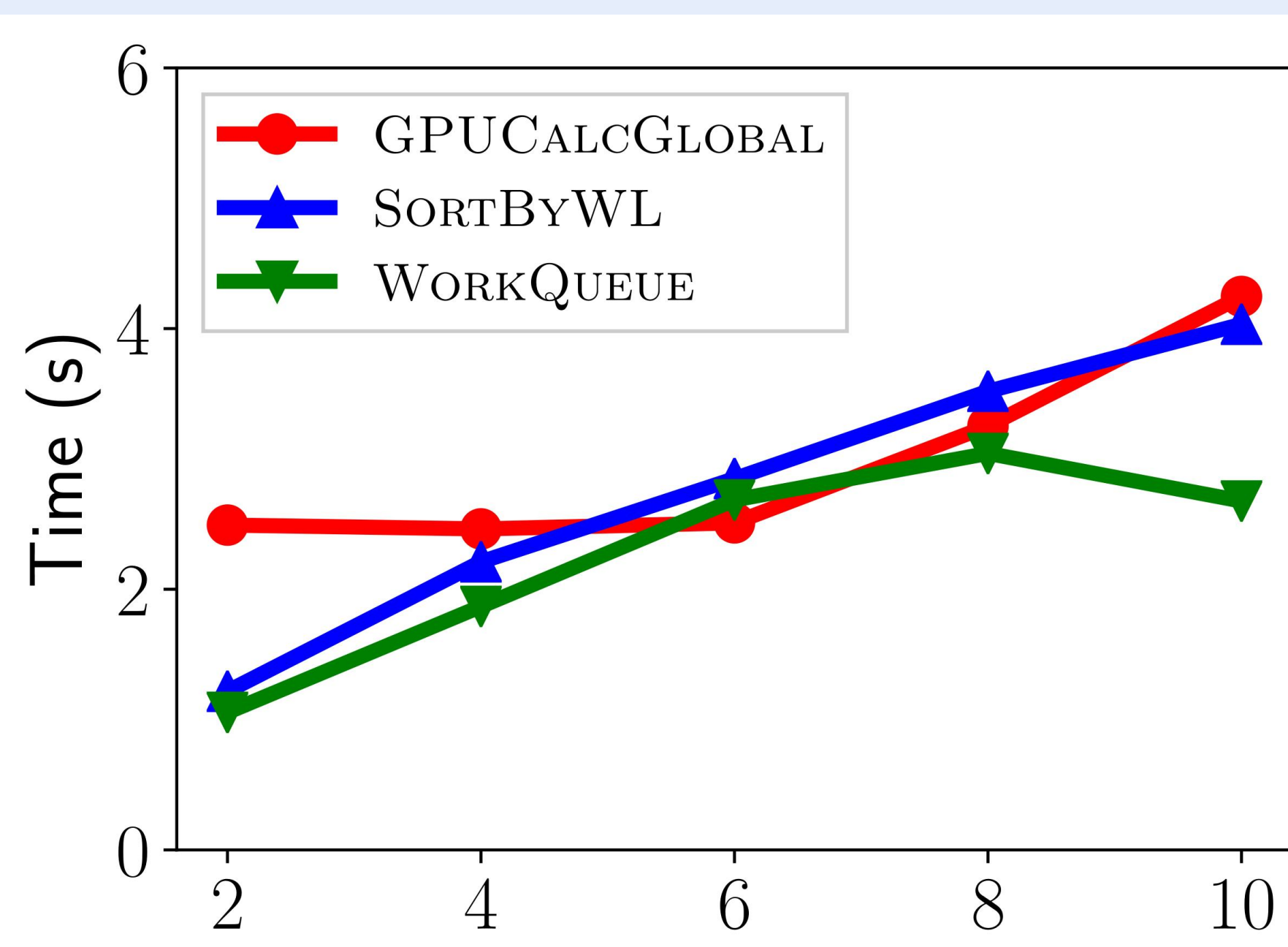


Figure 3: Uniform

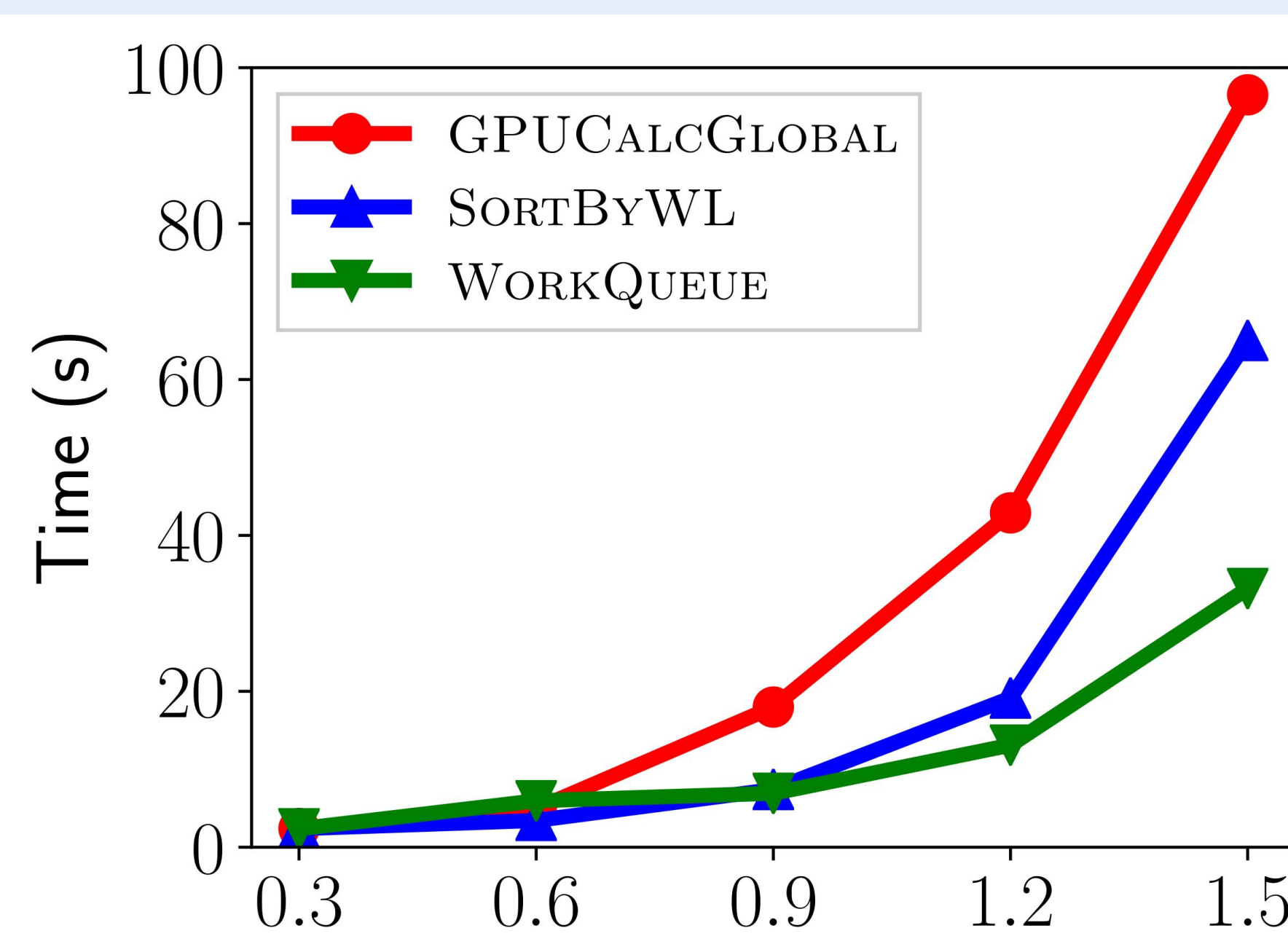
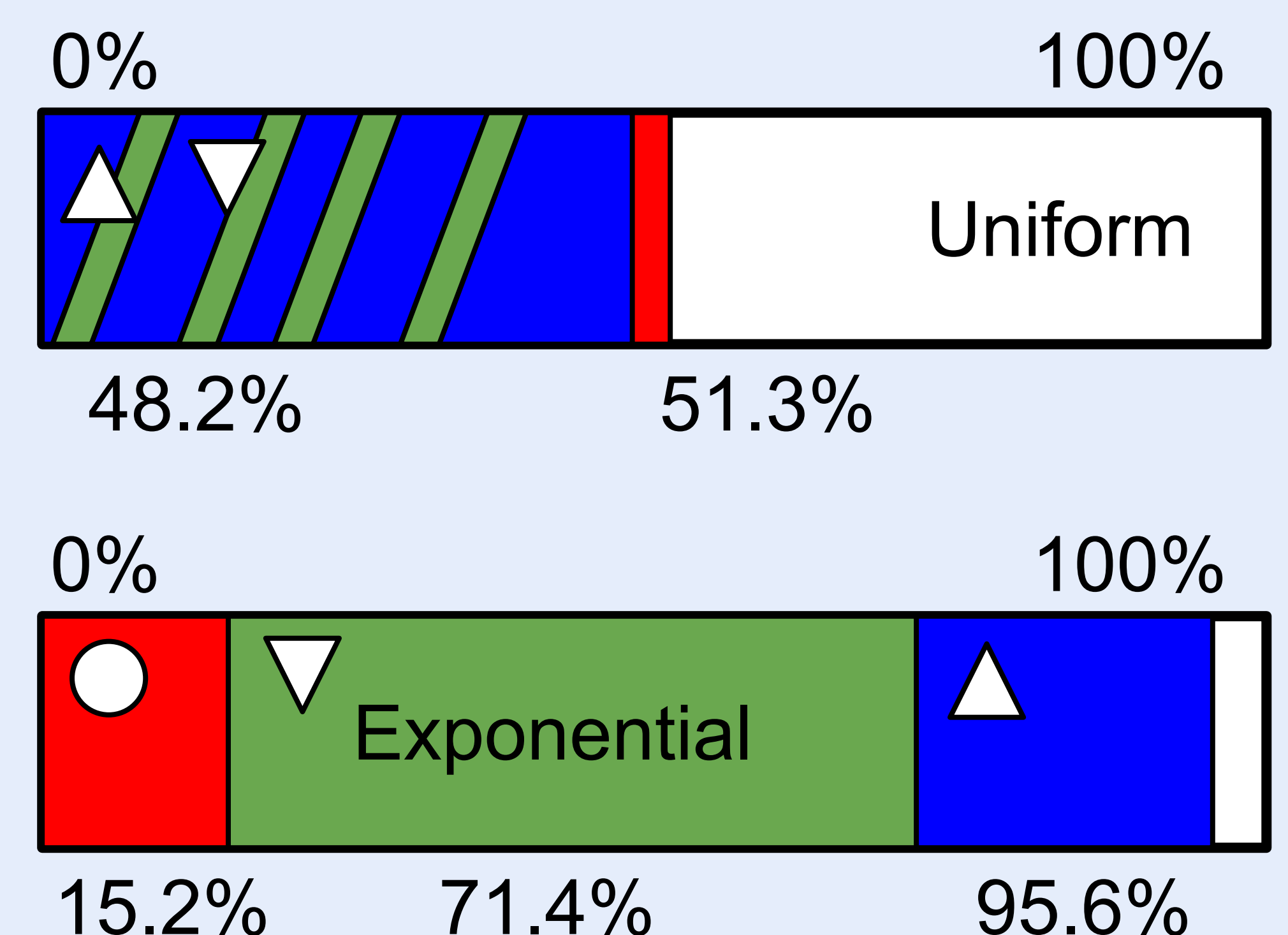


Figure 4: Exponential



- Uniformly distributed dataset \rightarrow uniform workload \rightarrow no need to balance, contrary to exponentially distributed

Conclusion

- Warp execution efficiency impacts response time
- 100% warp execution efficiency may indicate a computational bound
- Use the WorkQueue to improve other data dependent applications

References

- [1] M. Gowanlock and B. Karsin, "GPU Accelerated Self-join for the Distance Similarity Metric," Proc. of the 2018 IEEE Intl. Parallel and Distributed Processing Symposium Workshops, pp. 477–486, 2018.
- [2] B. Gallet and M. Gowanlock, "Load Imbalance Mitigation Optimizations for GPU-Accelerated Similarity Joins", Proc. of the 2018 IEEE Intl. Parallel and Distributed Processing Symposium Workshops, 2019