Abstract

GPU parallelism for real applications can achieve enormous performance gain. CPU-GPU communication is one of the major bottlenecks that limit this performance gain. Among several libraries developed so far to optimize this communication, DyManD (Dynamically Managed Data) provides better communication optimization strategies and achieves better performance on a single GPU. Smith-Waterman is a well known algorithm in the field of computational biology for finding functional similarities in a protein database. CUDA implementation of this algorithm speeds up the process of sequence matching in the protein database. When input databases are large, multi-GPU implementation gives better performance than single GPU implementation. Since this algorithm acts upon large databases, there is need for optimizing CPU-GPU communication. DyManD implementation provides efficient data management and communication optimization only for single GPU. For providing communication optimization on multiple GPUs, an approach of combining DyManD with a multi-threaded framework called GPUWorker was proposed. Our contribution in this work is to propose an optimized CUDA implementation of this algorithm on multiple GPUs i.e., GPUWorker-DyManD which reduces the communication overhead between CPU and multiple GPUs. This implementation combines DyManD functionality with GPUWorker for optimizing communication. The performance gain obtained for the GPUWorker-DyManD implementation of this algorithm over default multi-GPU
implementation is 3.5x.

References

- SARJAT SAHNI JUNJIE LI & SANJAY RANKA. Pairwise sequence alignment for very long sequences on gpu.

Index Terms

Computer Science Algorithms

Keywords

DyManD GPUWorker Data Management Communication Optimization