Multiple Pairwise Sequence Alignments with the Needleman-Wunsch Algorithm on GPU

Da Li, Michela Becchi

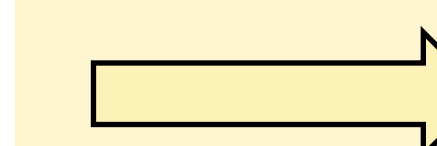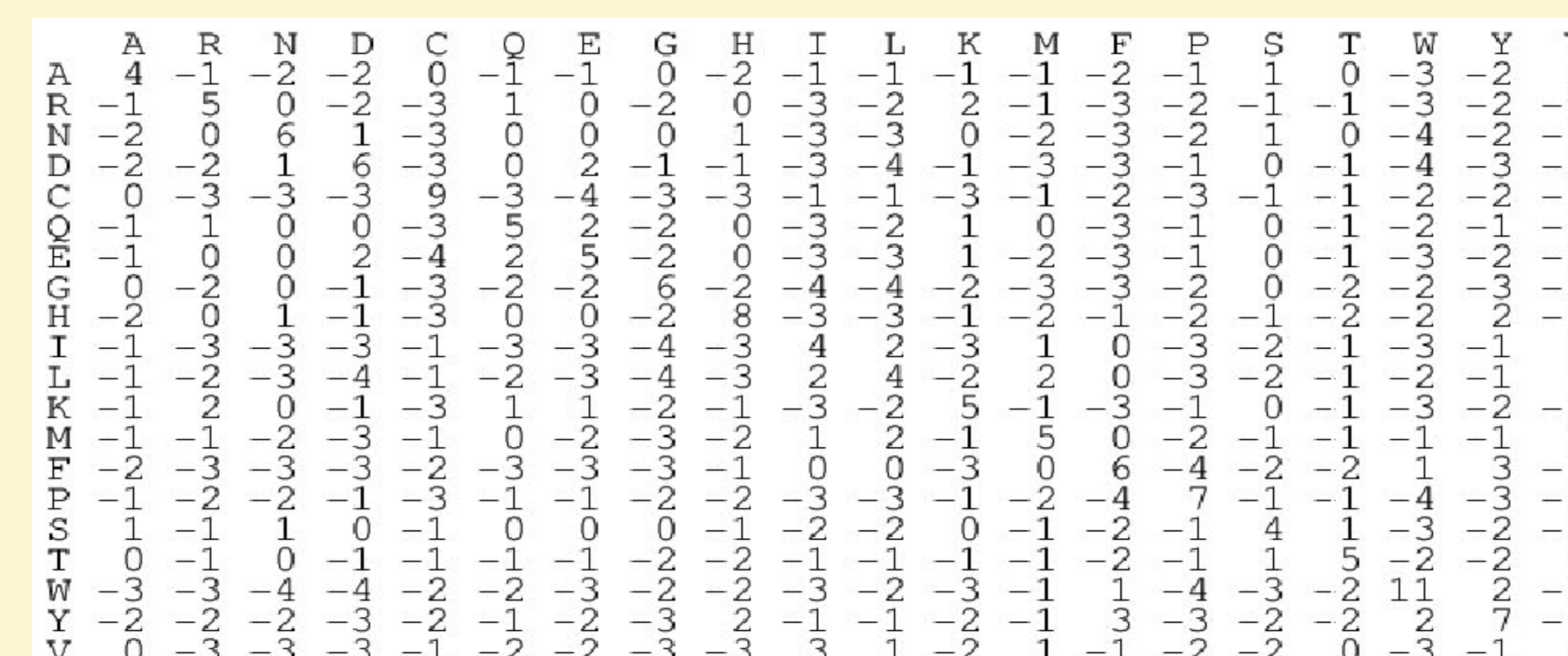**University of Missouri - Columbia**

Abstract

Pairwise sequence alignment is a method used in bioinformatics to determine the similarity between DNA, RNA and protein sequences. The Needleman-Wunsch algorithm is typically used to perform global alignment, and has been accelerated on Graphics Processing Units (GPUs) on single pairs of sequences. Many applications require multiple pairwise comparisons over sets of sequences. The large size of modern bioinformatics datasets leads to a need for efficient tools that allow a large number of pairwise comparisons. Because of their massive parallelism, GPUs are an appealing choice for accelerating these computations. In this paper, we propose an efficient GPU implementation of multiple pairwise sequence alignments based on the Needleman-Wunsch algorithm. Compared to a well-known existing solution, our implementation improves the memory transfer time by a factor 2X, and achieves a ~3x speedup in kernel execution time.

Introduction

- Global alignment (Needleman-Wunsch)
- Local alignment (Smith-Waterman)

Example: GCGCATGGATTGAGCGA and TGCGCCATTGATGACCA

Global	Local
-GCGC-ATGGATTGAGCGA	**GCGC**ATGGATTGAGCGA
TGCGCCATTGAT-GACC-A	T**GCGC**CATTGATGACCA

Needleman-Wunsch

$$M(i,j) = \max \begin{cases} M(i-1, j-1) + S(x_i, y_j) \\ M(i-1, j) + G \\ M(i, j-1) + G \end{cases} +$$
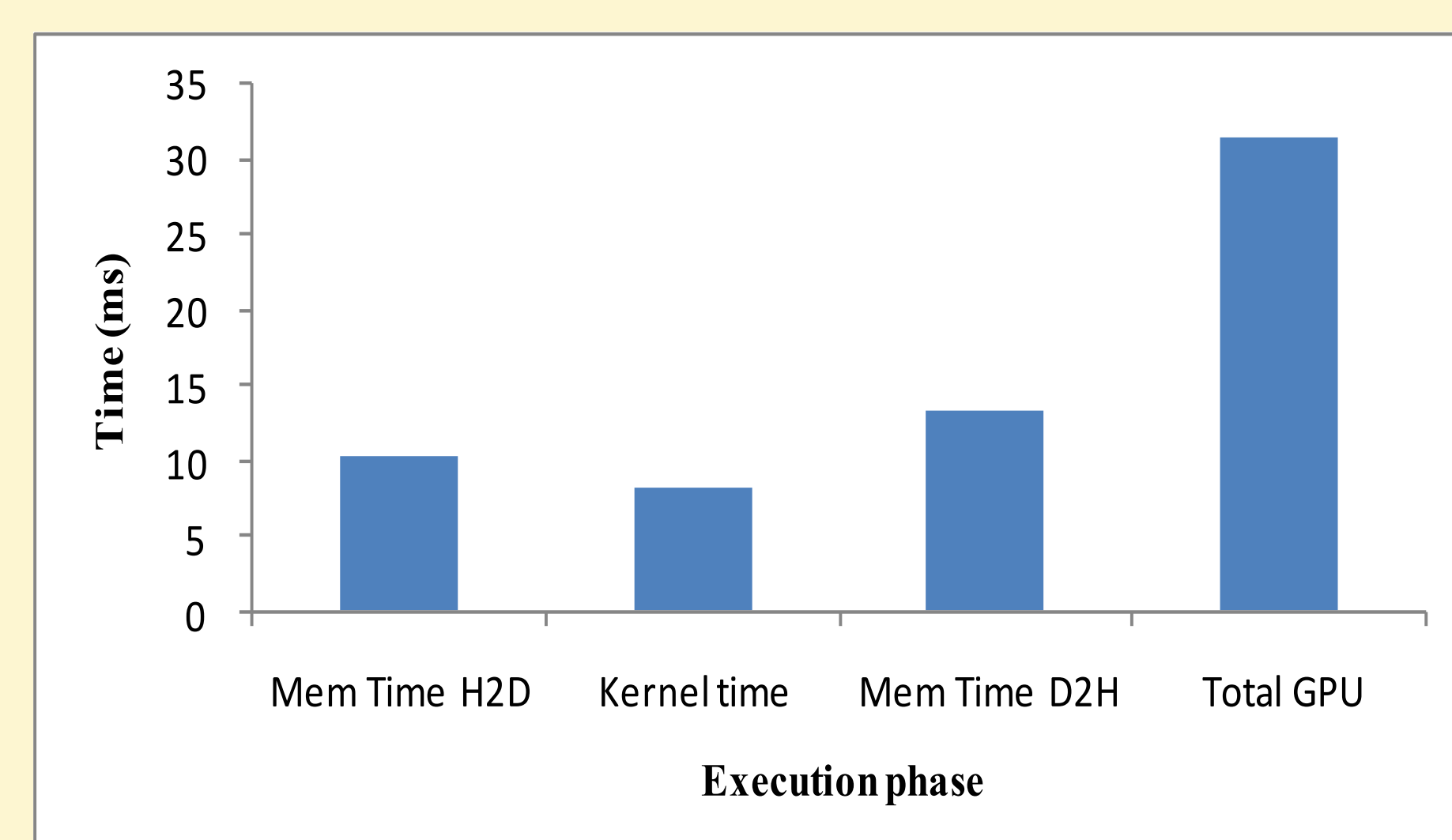

Motivation & Method

Previous work (Rodinia's NW) has following limitations:

- Designed for single (as opposed to multiple) pair-wise alignment
- Communication overhead
- Kernel launch overhead
- GPU underutilization
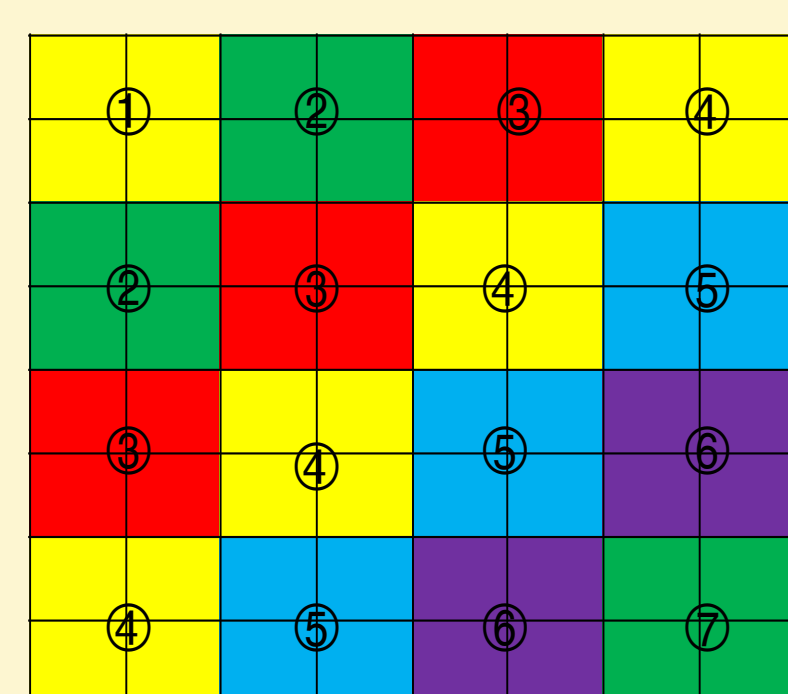
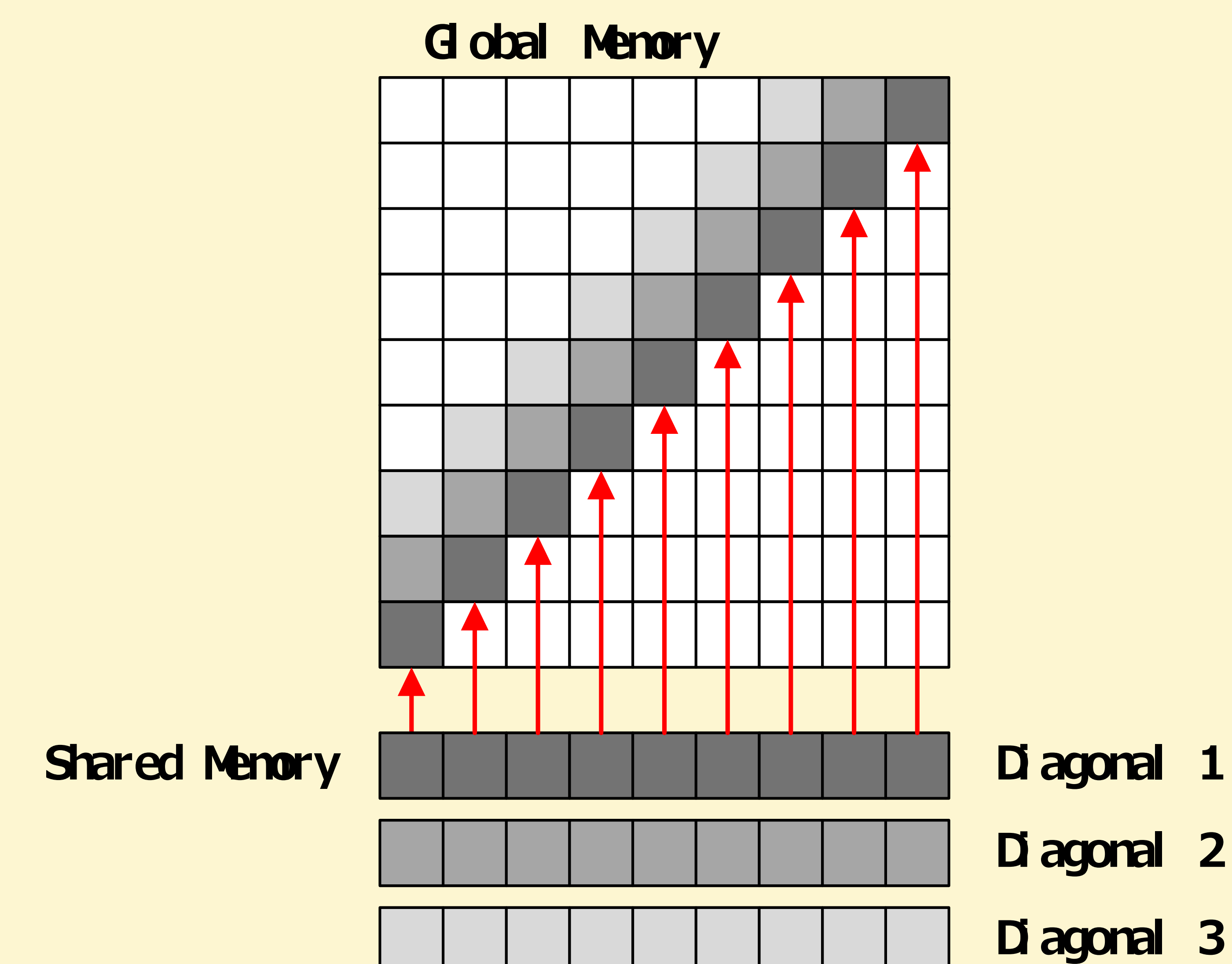**Break-down of execution time of Rodinia's NW**

Experiments performed on Nvidia GeForce GTX 480 GPU, 15 x 32 CUDA cores, 1,536 Mbytes global memory and 2,048 base sequences

Two levels of parallelism

- Each pair of sequences maps to one thread-block
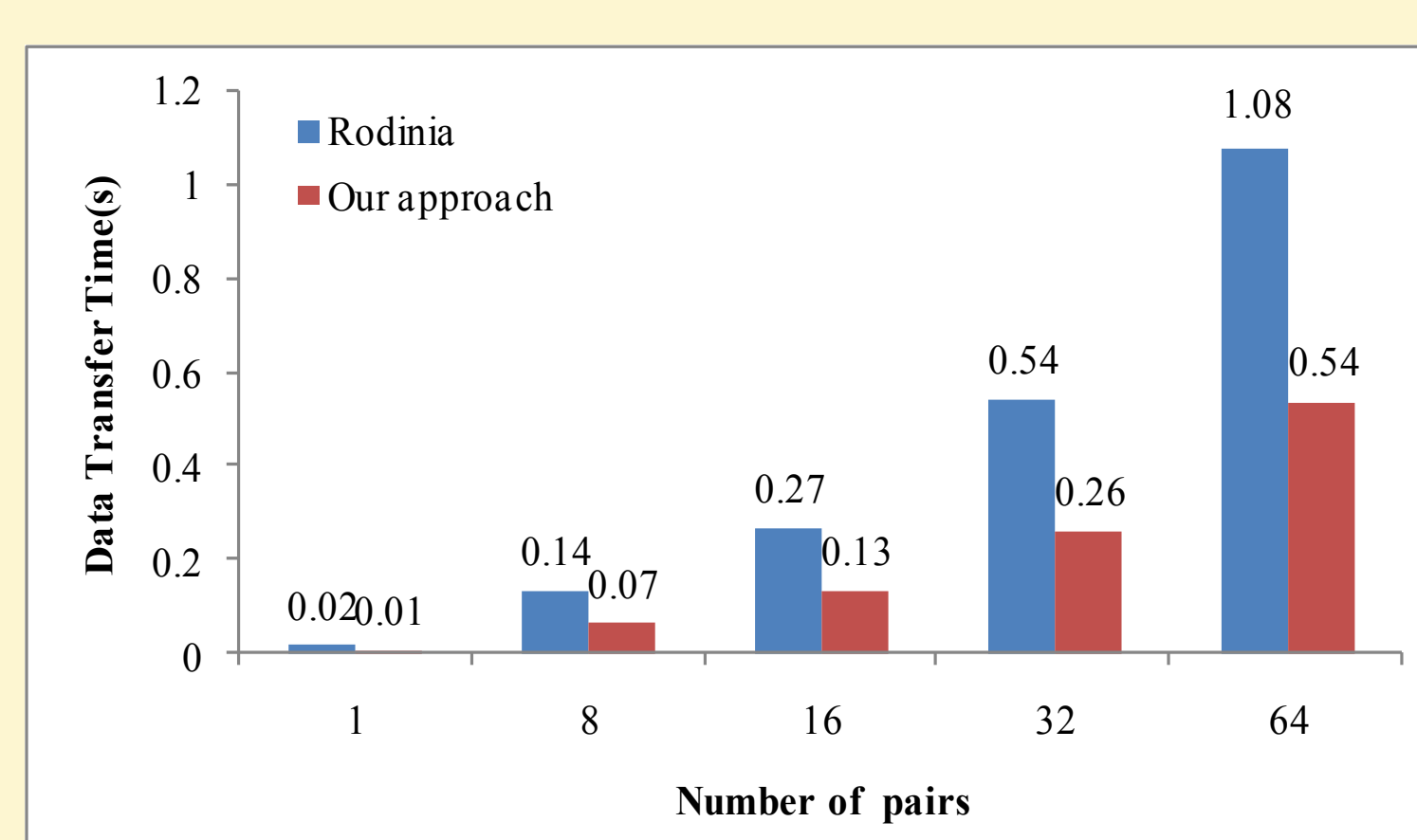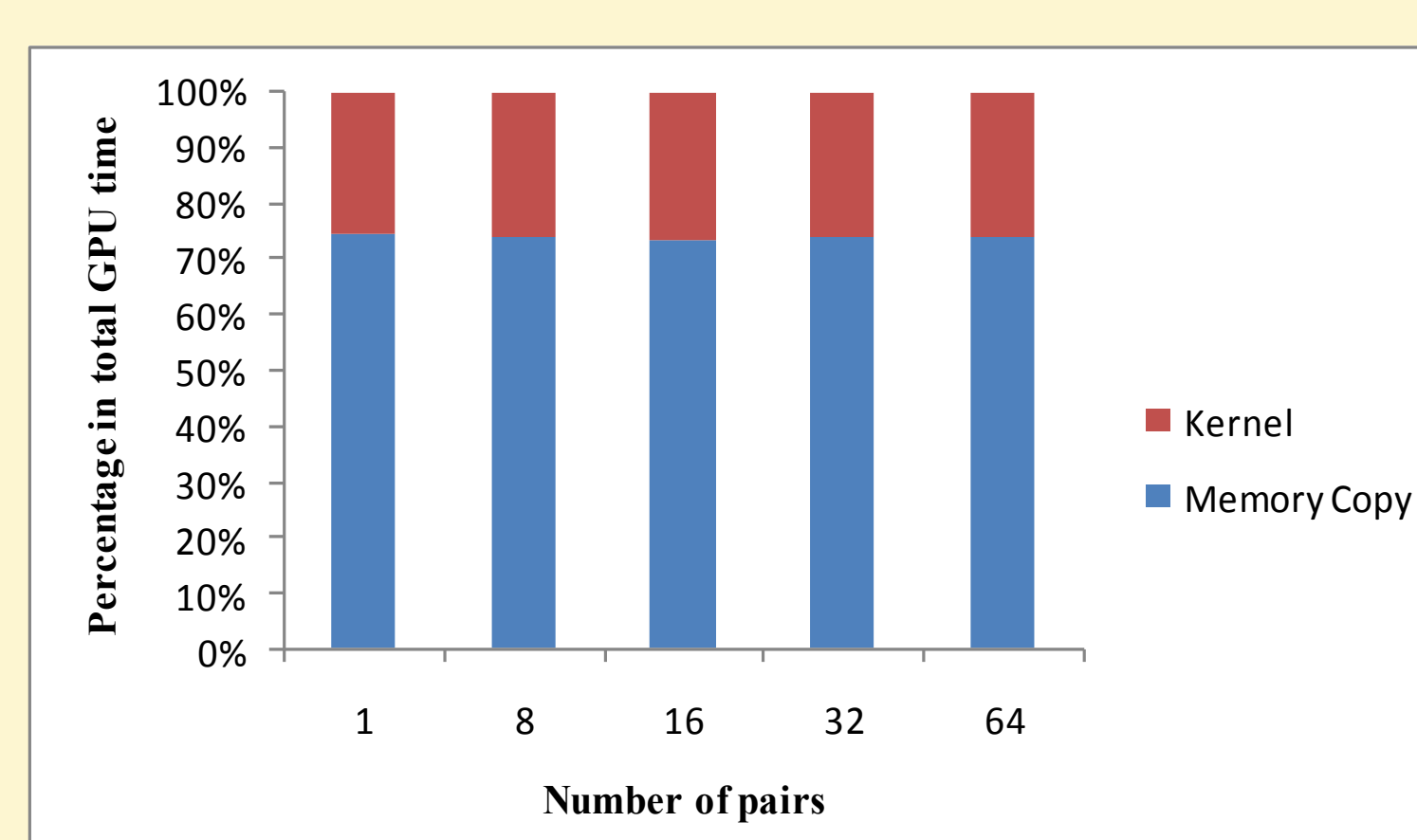- Multiple threads cooperate in a minor diagonal manner

Block 0 | Pair 0 Block 1 | Pair 1
Block 2 | Pair 2 Block 3 | Pair 3
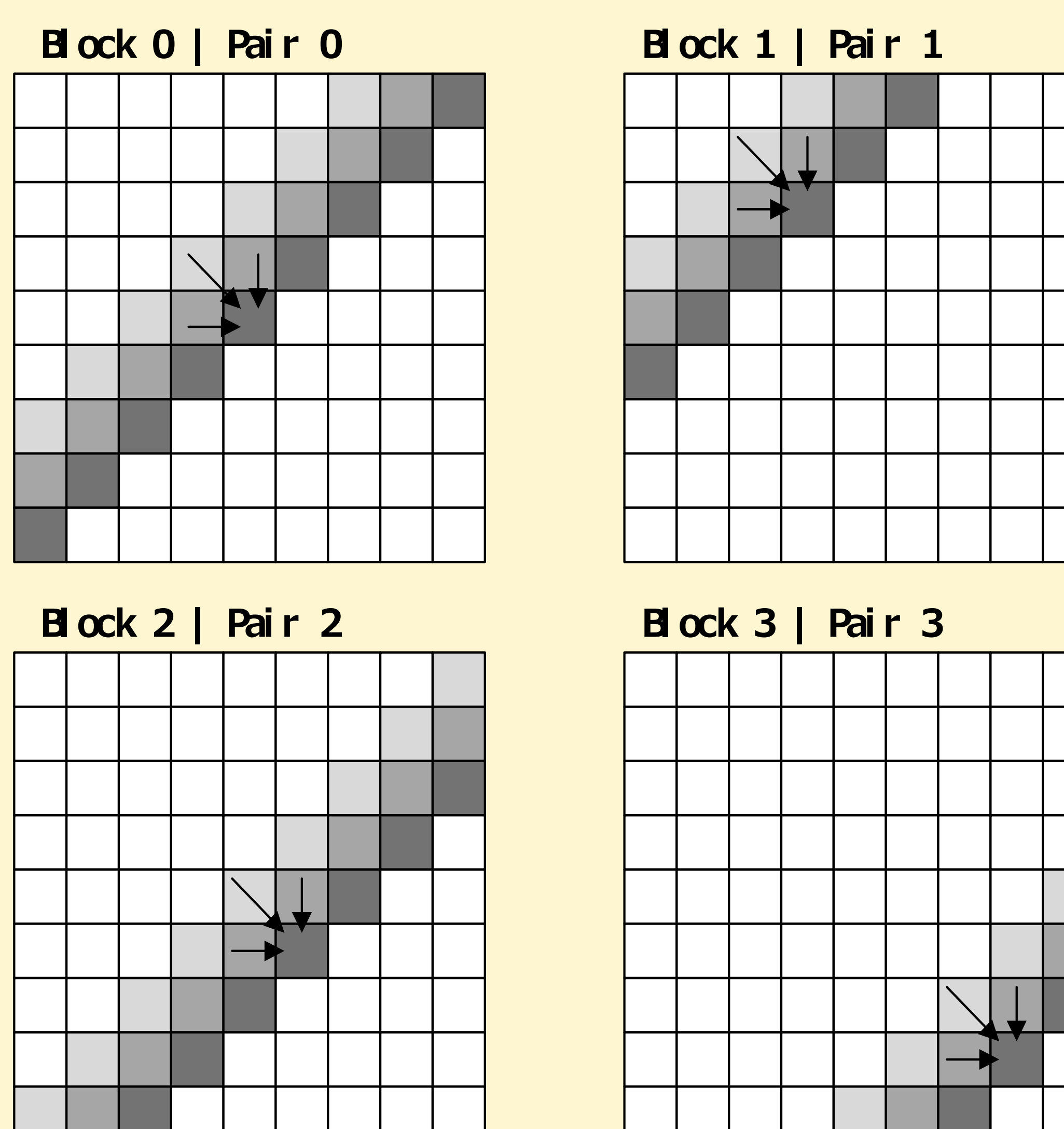
Data mapping to shared memory

Global Memory

Shared Memory   Diagonal 1
Diagonal 2
Diagonal 3

Experiments

Comparison of data transfer time

Comparison of kernel time

GPU usage of Rodinia's NW
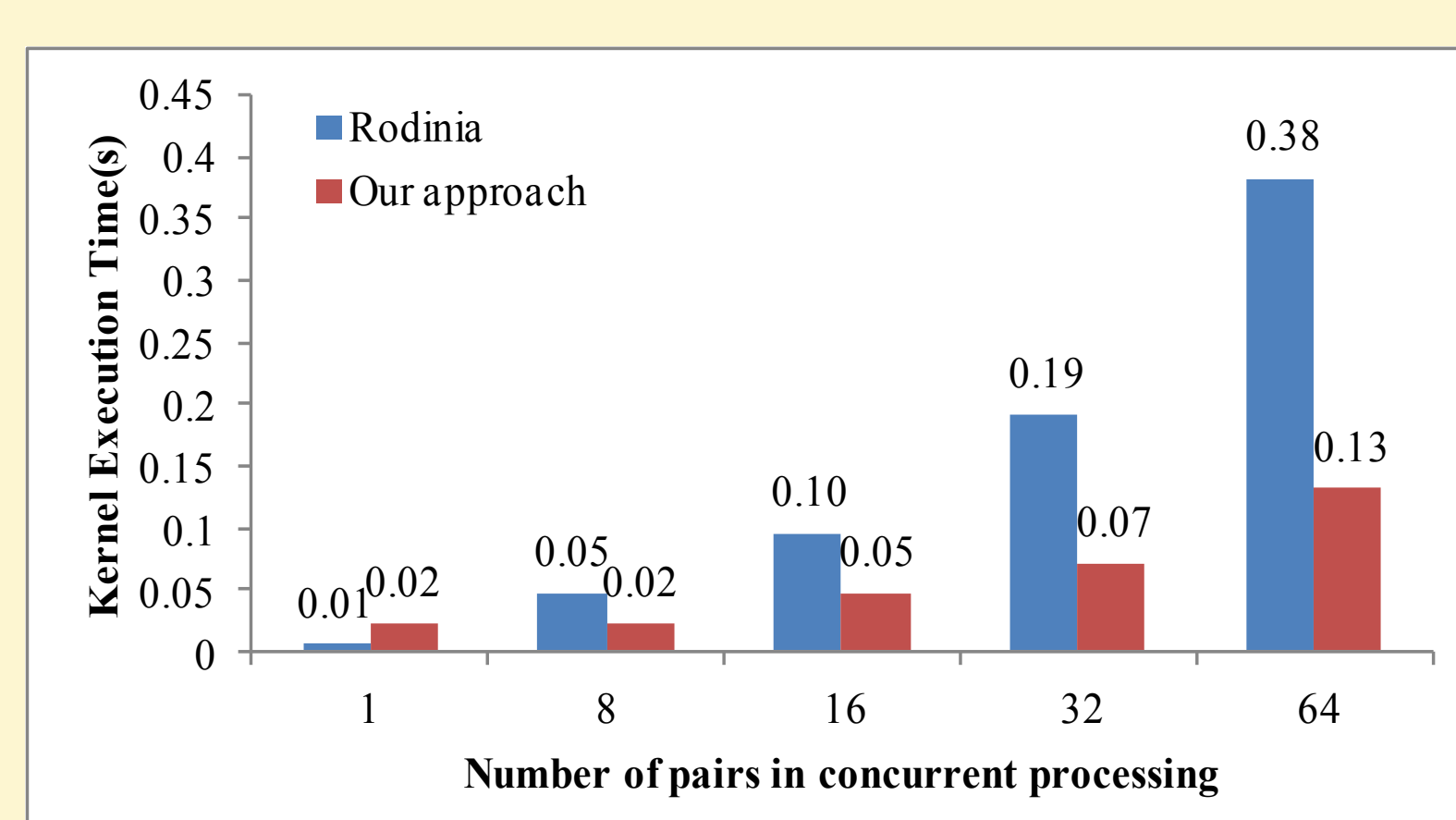
GPU usage of our NW implementation

Comparison between GPU and multi-threaded CPU implementation
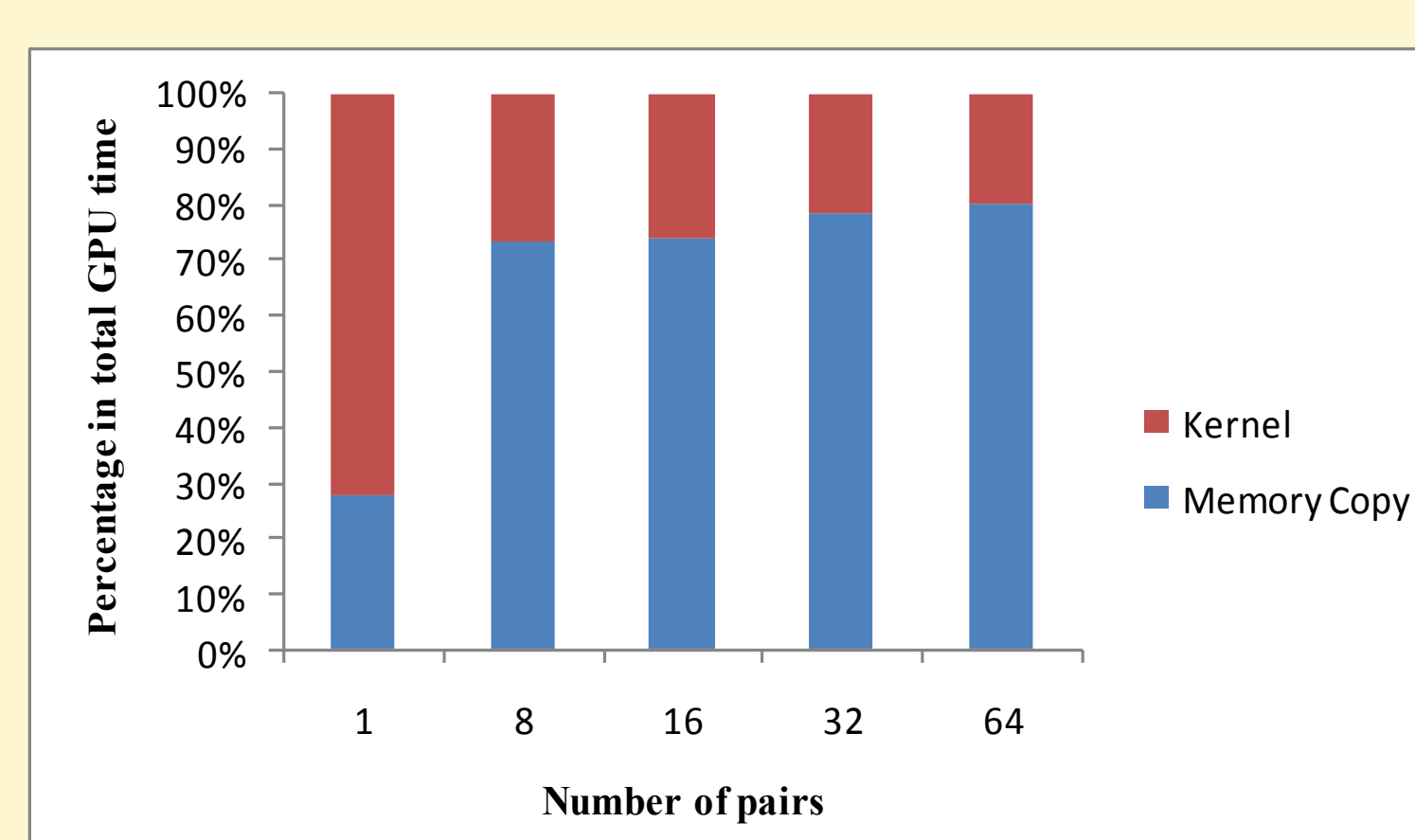
- GPU platform: Nvidia GeForce GTX 480 GPU, 15 x 32 CUDA cores, 1,536 Mbytes global memory, 16KB L1 cache and 2,048 base sequences
- CPU platform: Intel Xeon E5620, 12MB L1 Cache
- 8 threads OpenMP implementation
- CUDA kernel configuration of our implementation: 512 threads per block, #block = #pair

References

[1] Saul B. Needleman, Christian D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48 (3): 443-53, 1970.
[2] Temple F. Smith, Michael S. Waterman, "Identification of Common molecular Subsequences", Journal of Molecular Biology 147: 195-197, 1981.
[3] Tahir Naveed, Imitaz Saeed Siddiqui, Shaftab Ahmed, "Parallel Needleman-Wunsch Algorithm for Grid", Proceedings of the PAK-US International Symposium on High Capacity Optical Networks and Enabling Technologies(HONET 2005), Islamabad, Pakistan, 2005.
[4] Shuai Che, Jie Li, Jeremy W. Sheaffer, Kevin Skadron, John Lach, "Accelerating Compute-Intensive Applications with GPUs and FPGAs", SASP, 2008.
[5] Shuai Che, Michael Boyer et al, "Rodinia: A Benchmark Suite for Heterogeneous computing", In Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), pp. 44-54. 2009.

Contact us:
dlx7f@mail.missouri.edu