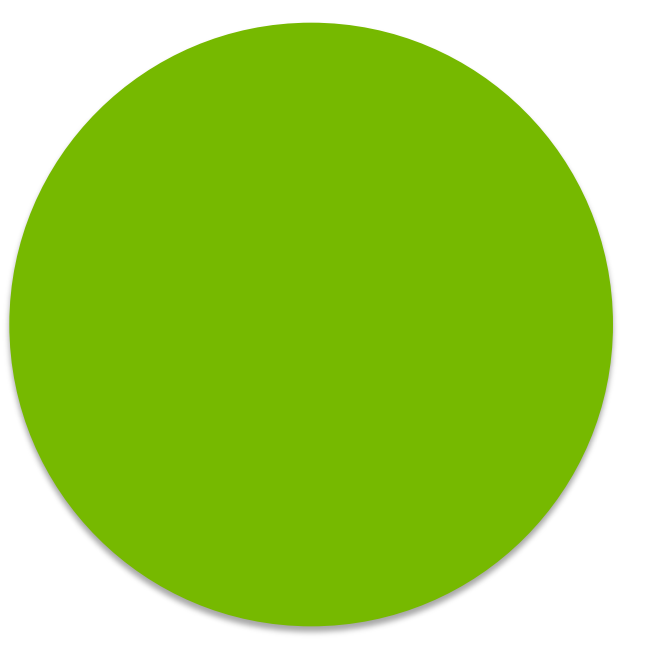


CUDA Implementation of Software for Identifying Post-Translational Modifications (PTMs)



Yantang Zhai, Qiang Tu, Xianyu Lang. Supercomputing Center, Chinese Academy of Sciences, Beijing, 100190

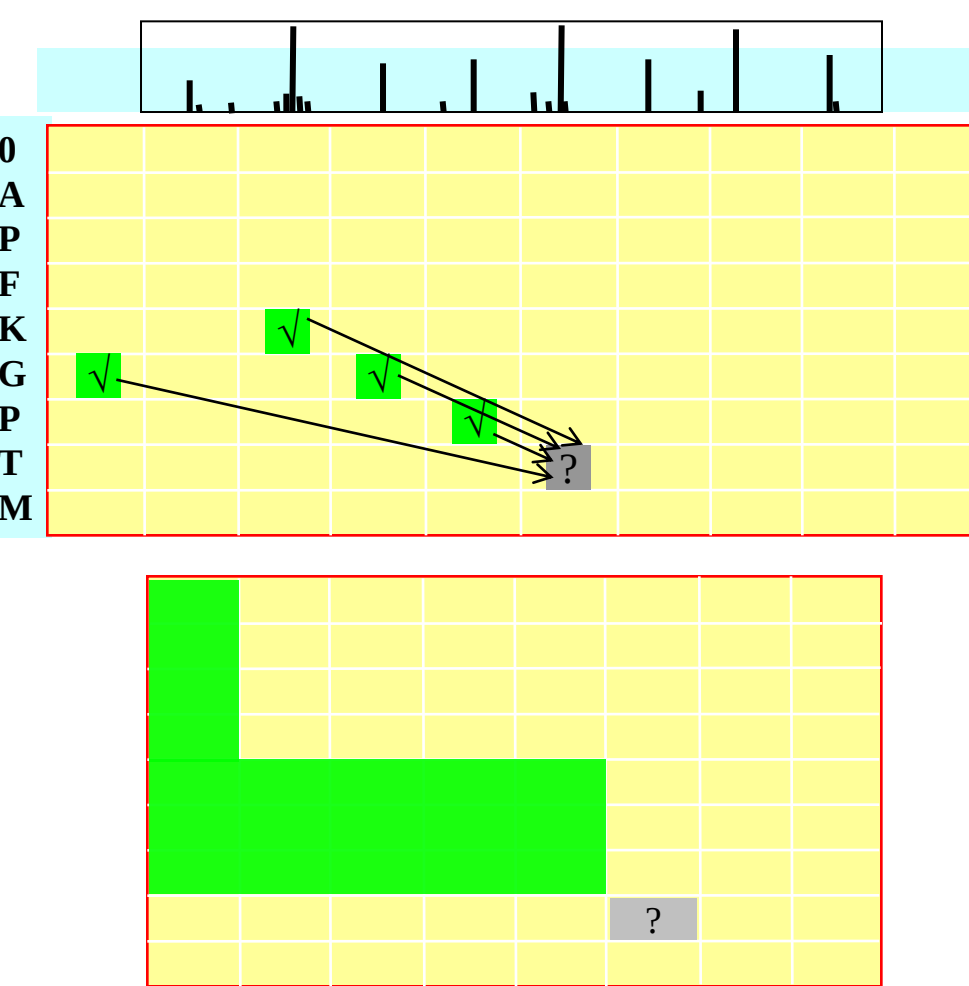
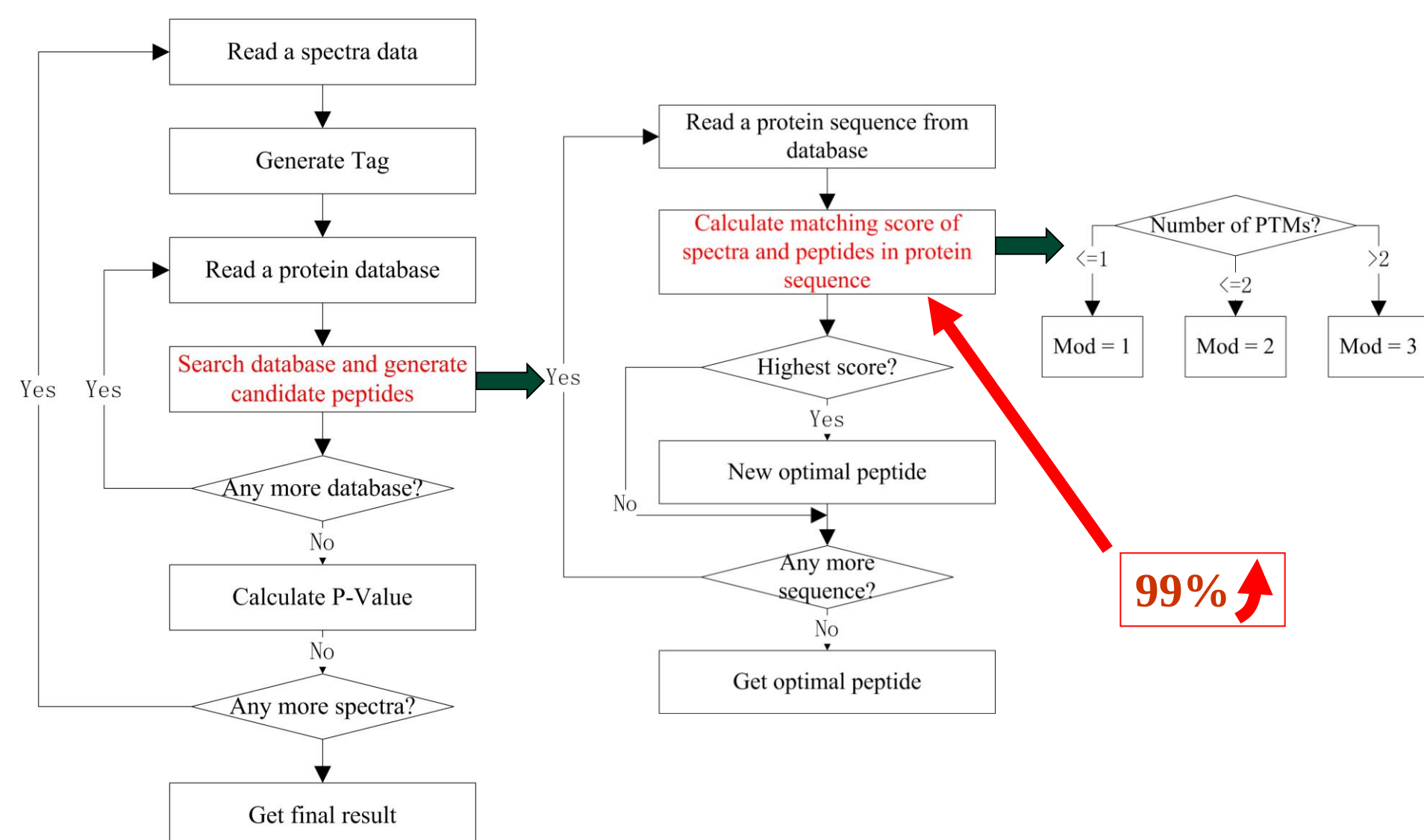
Introduction

InsPecT is a software for identifying post-translational modifications (PTMs) of protein. Because of involving the MS-Alignment algorithm, InsPecT can search PTMs in unrestrictive mode, even reveal unknown types of modifications. Because InsPecT has a tremendous time complexity, we accelerated it using GPUs.

The section we accelerated

MS-Alignment takes more than 99% computing complexity of InsPecT and uses 3 different algorithm models according to the maximum number of PTMs to identify. As we know that identifying no more than 2 PTMs in unrestrictive mode is popular and identifying 2 PTMs (mod=2) is much more complicated, so accelerating and optimizing “mod=2” model is the focused work we describe here.

MS-Alignment is a dynamic programming based algorithm. The complicated data dependence makes the implementation on GPUs more difficult.



$$DT(p_i, n_k) = score(mass(n_k)) + \max \{$$

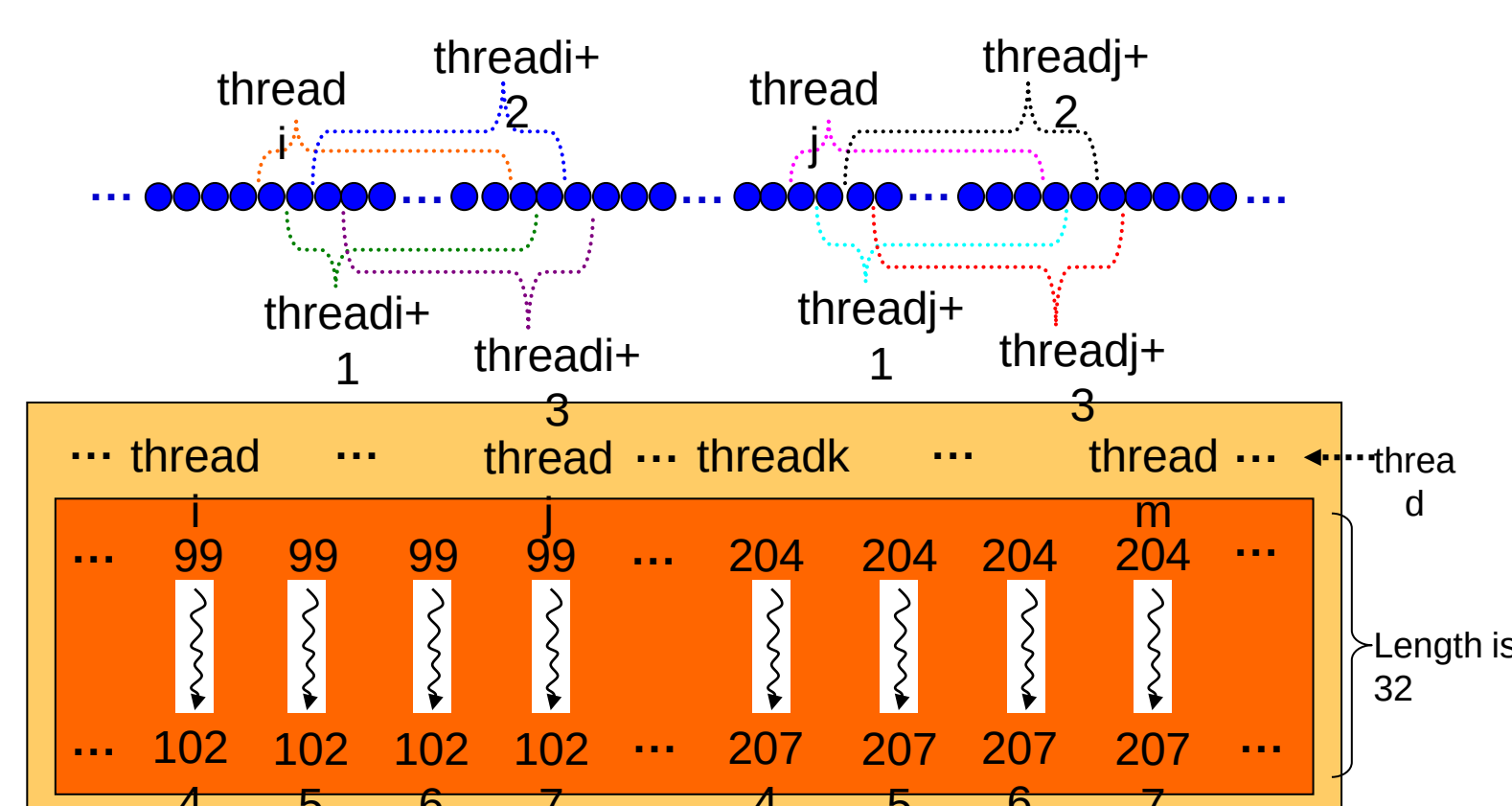
- ① Initial Value : $startscore(p_i)$;
- ② $\max \{ DT(p_{i-1}, n_i) + score(edge) \mid edge \in \{from\ n_i\ to\ n_k\}, i < k \}$;
- ③ $\max \{ DT(p_{i-2}, n_i) + score(doubleedge) \mid edge \in \{from\ n_i\ to\ n_k\}, i < k \}$;
- ④ $\max \{ DT(p_{i-3}, n_i) + score(tripleedge) \mid edge \in \{from\ n_i\ to\ n_k\}, i < k \}$;
- ⑤ $\max \{ DT(p_i, n_0) + prefixscore(p_i, p_i) \mid 0 < i \leq i \}$;
- ⑥ $\max \{ DT(p_i, n_0) + prefixscore(p_i, p_i) + ptmscore(\Delta, p_i) \mid 0 < i \leq i \}$;

$$\}$$

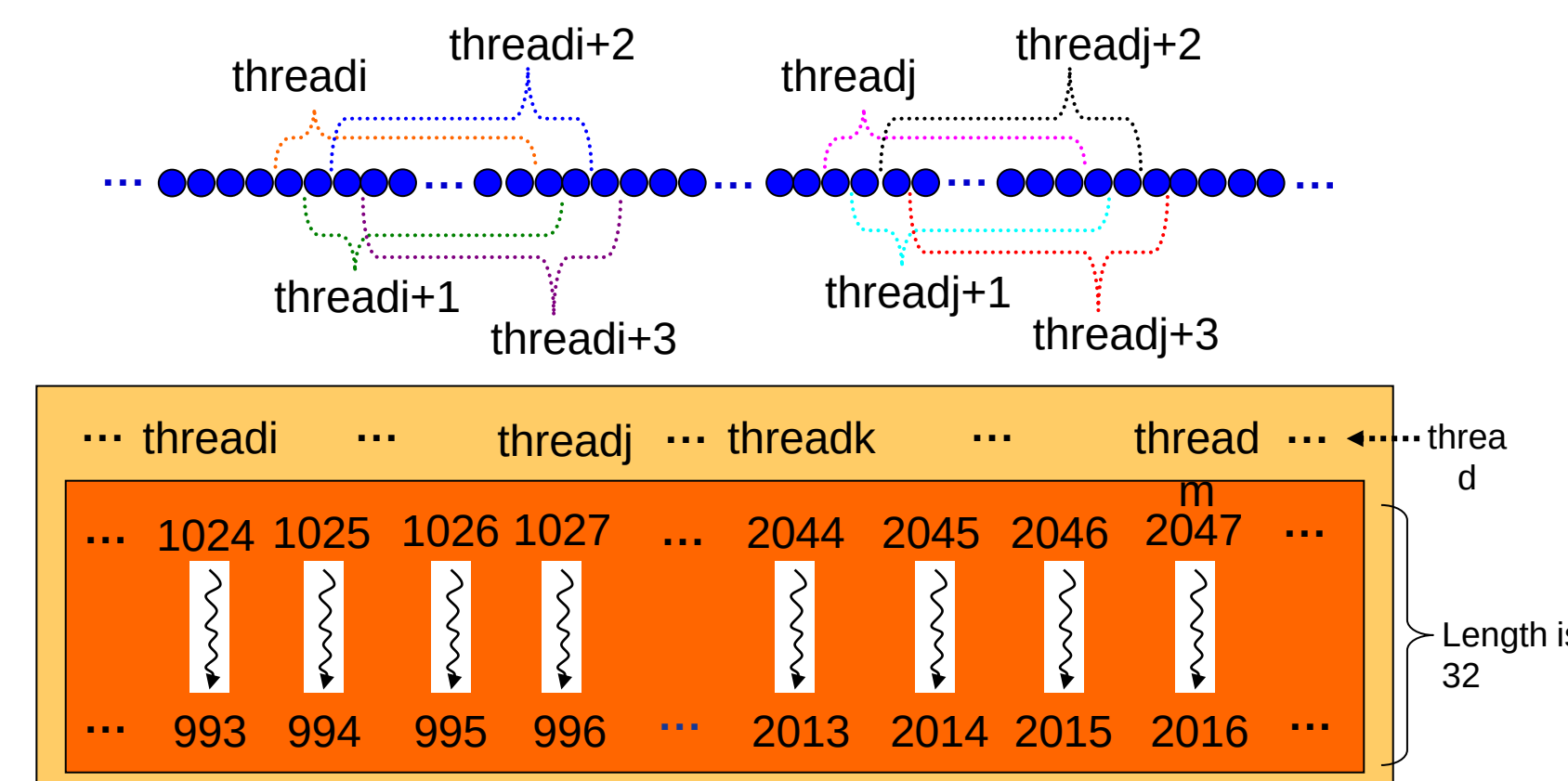
Acceleration on GPU

To identifying one peptide, InsPecT computes prefix table and suffix table of the peptide firstly, then gets candidate result according to those two tables, and select top 10 results from candidates finally.

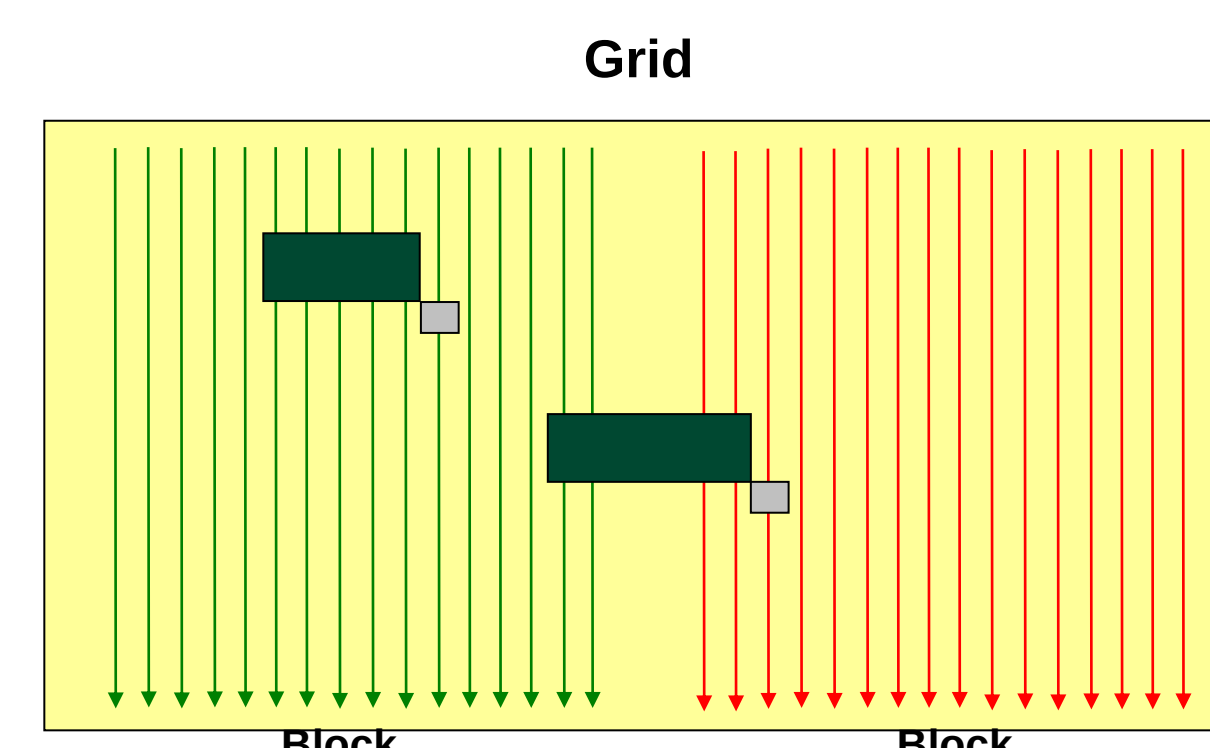
We used 4 CUDA kernels to realize the “mod=2” model on GPU. Kernel 1 and kernel 2 fill the prefix table and suffix table, and kernel 3 gets the candidates, then kernel 4 selects top 10 results from candidates. During the process of these four kernels, the data is stored in the global memory.



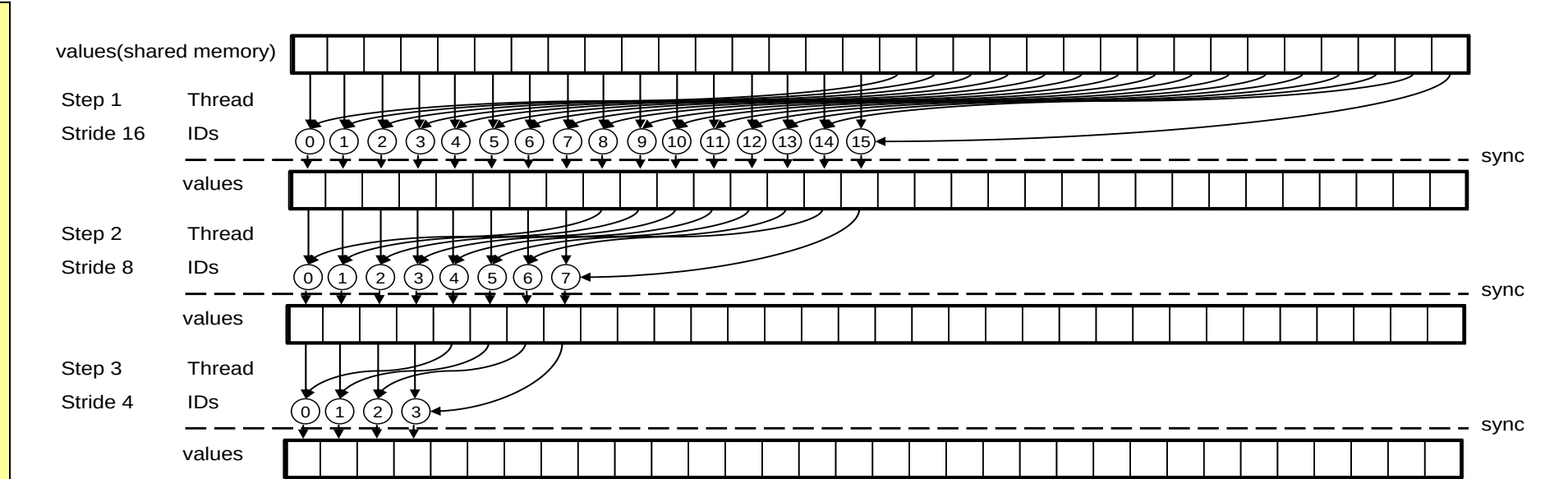
Kernel 1



Kernel 2

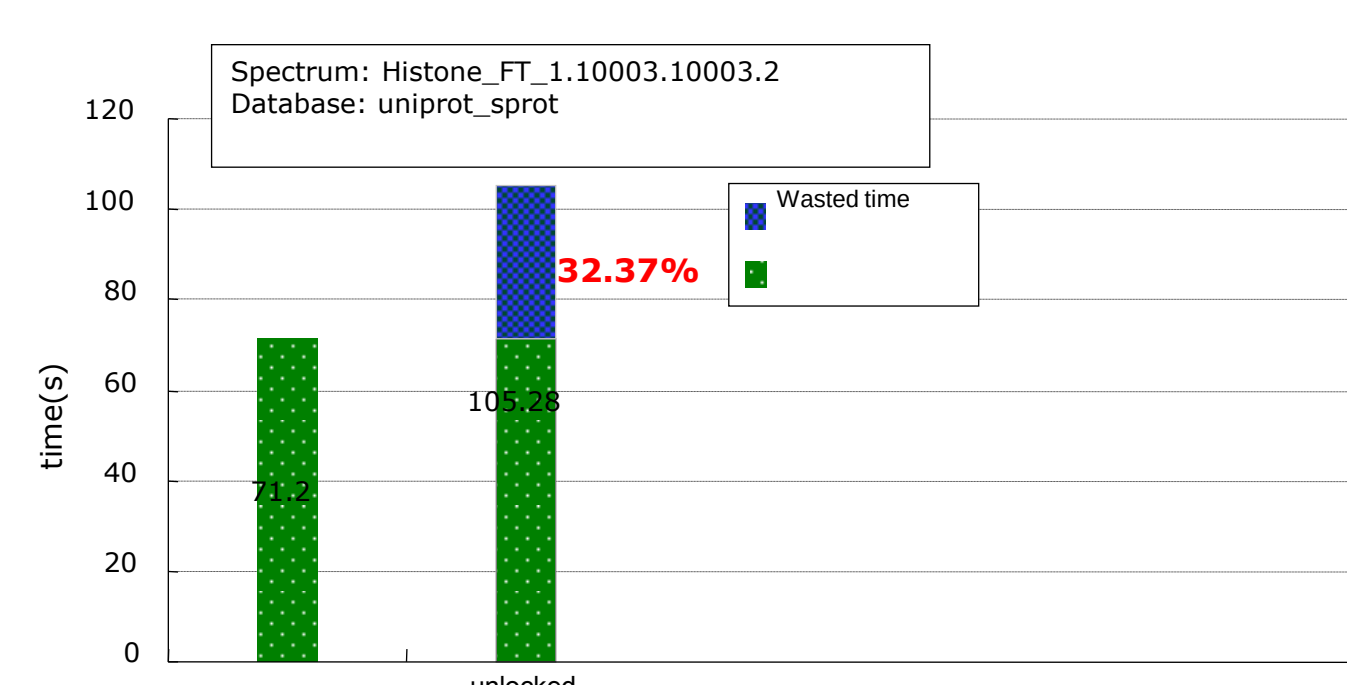


Kernel 3

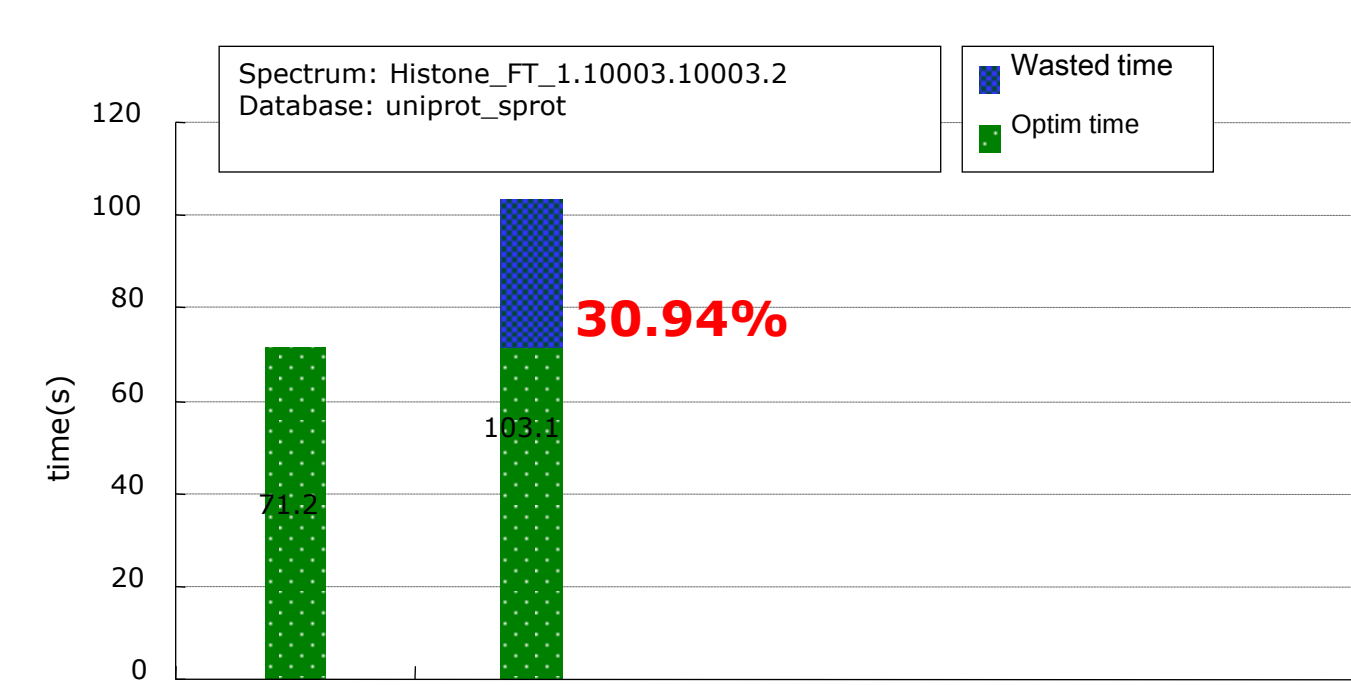


Kernel 4

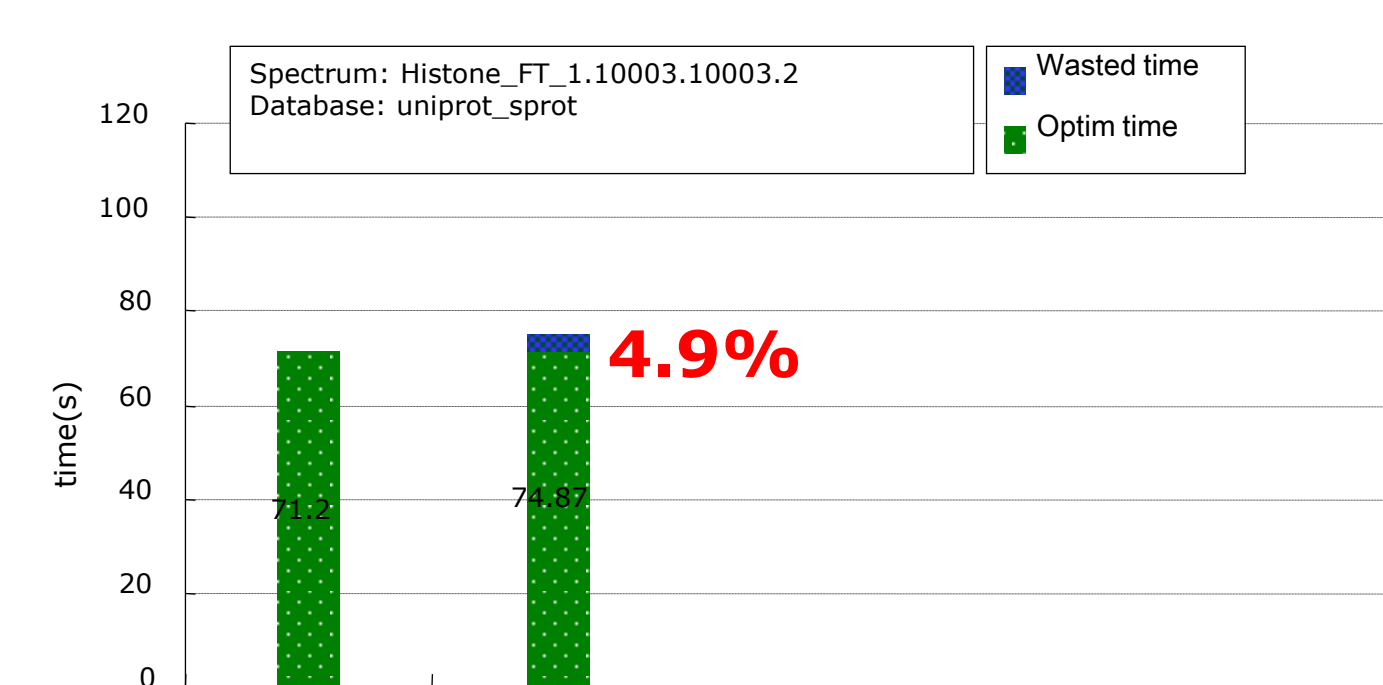
We use several optimization method to achieve higher speedup.



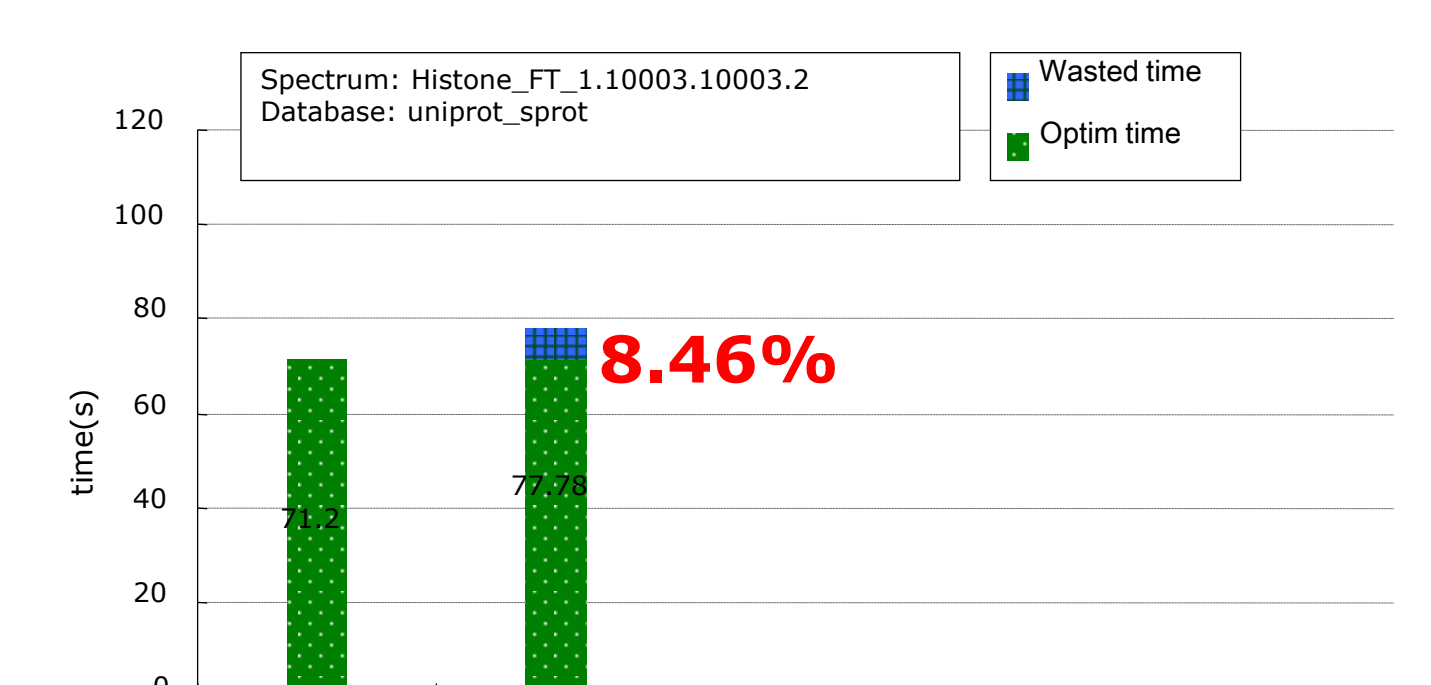
use page-locked memory



use texture memory

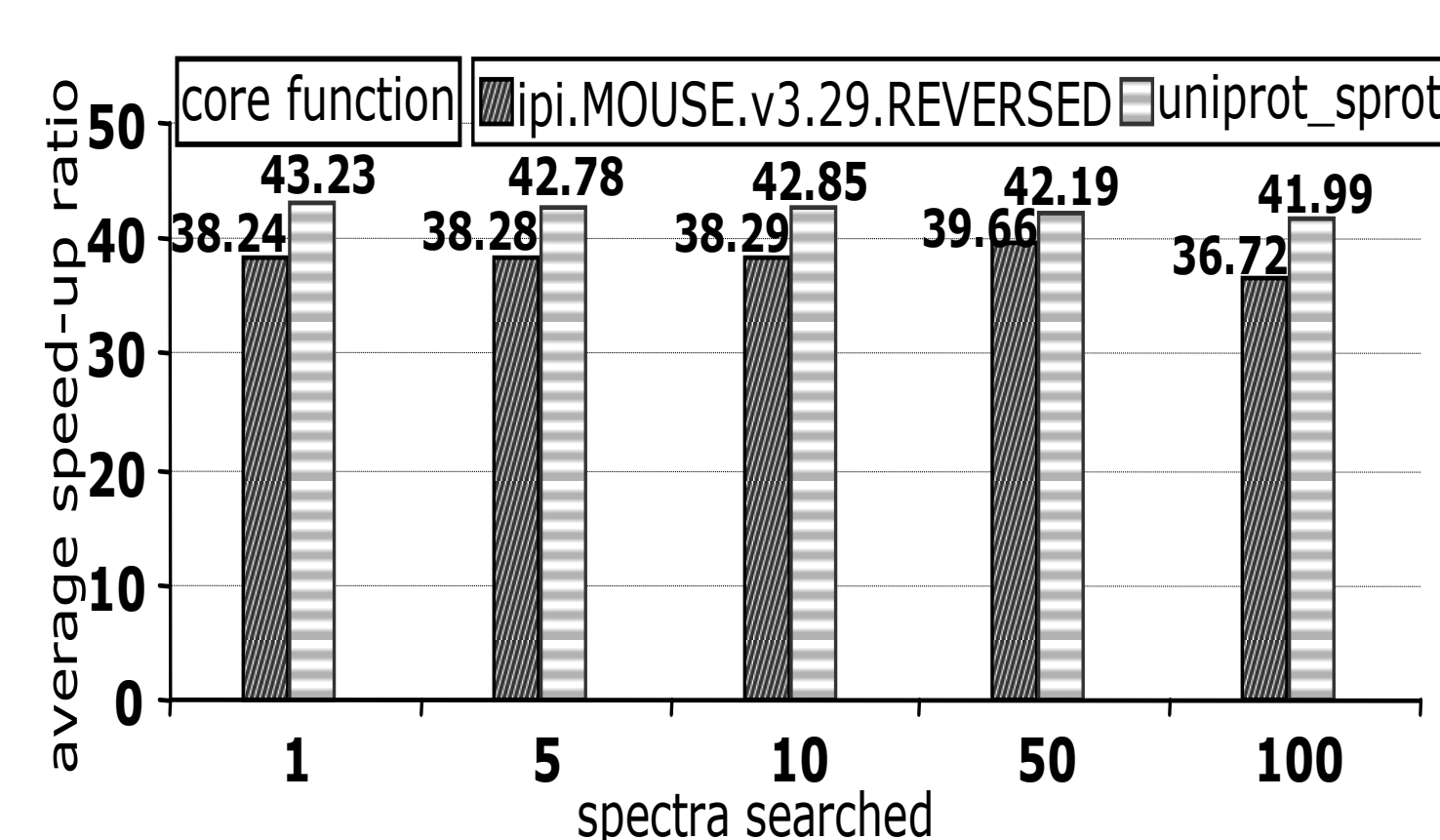


avoid bank conflict

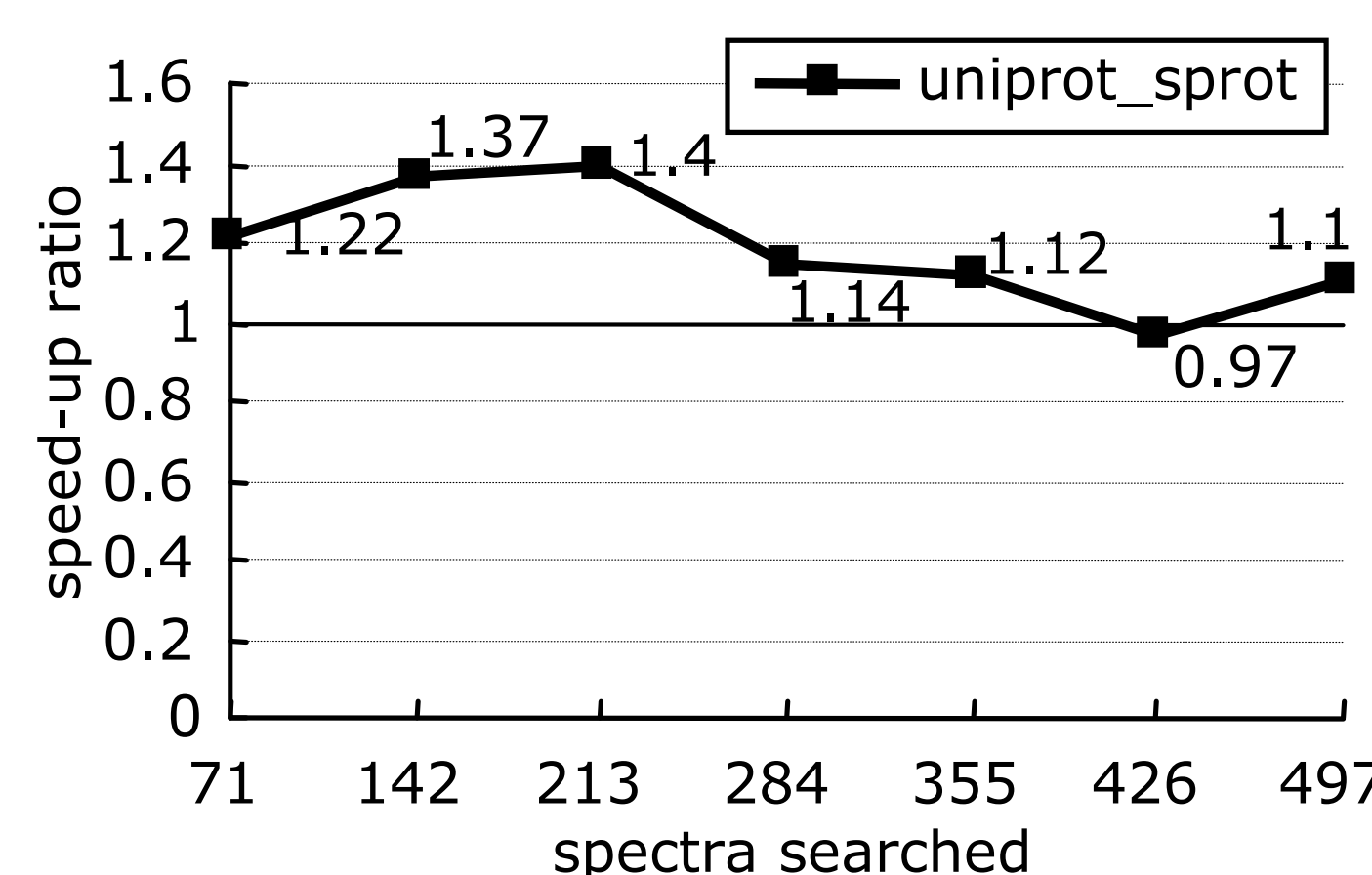


use global synchronization

We also parallelized InsPecT with MPI to make it running on multi-GPU environment. cuda-InsPecT is the MPI+CUDA version based on InsPecT, which will be open source.



1 GPU vs. 1 CPU



1 GPU vs. GPU cluster

Software : cuda-InsPecT		
Database : 62346 mass spectrometric; 107962 protein sequences		
Environment : Dawn 6000A(Nebulae)		
	1 Core in Intel X5650 (estimate)	677 Fermi C2050
Time	6 years	2.034 h

Performance on Dawn 6000A

Reference

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. *InsPecT: identification of post-translationally modified peptides from tandem mass spectra*. Anal Chem. 2005, 15;77(14): 4626-39

Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. *Identification of post-translational modifications by blind search of mass spectra*. Nat Biotechnol. 2005, 23(12): 1562-7