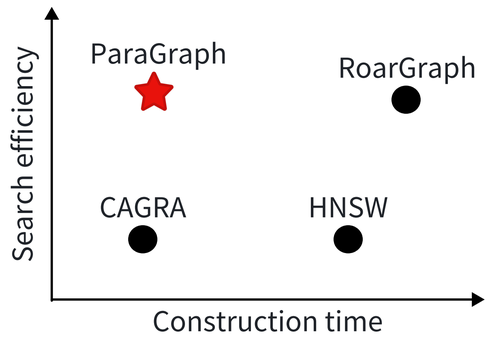
ParaGraph: Accelerating Graph Indexing through GPU-CPU Parallel Processing for Efficient Cross-modal ANNS

Yuxiang Yang, Shiwen Chen, Yangshen Deng, **Bo Tang**
Southern University of Sciecne and Technology; AlayaDB AI
research@alayadb.ai

Cross-model ANNS

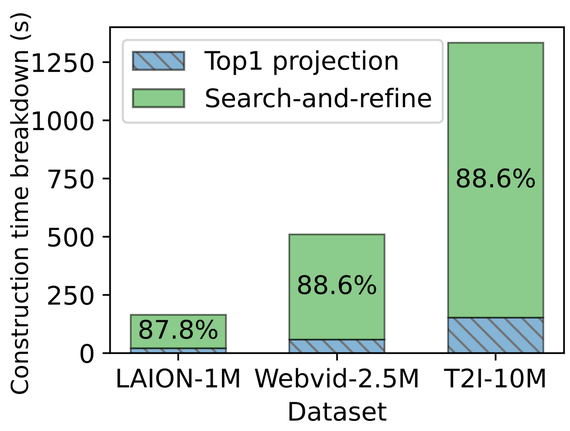
Existing methods struggle with either **slow index construction** or **poor search efficiency**.

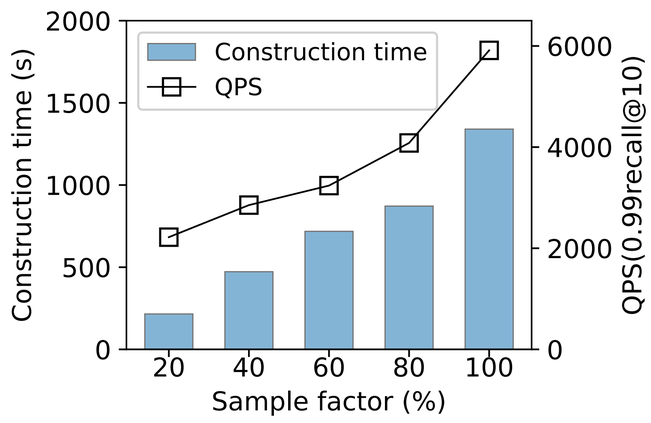


*How to bridge the gap between **construction** and **search**?*

The Bottleneck: Iterative Search-and-refine

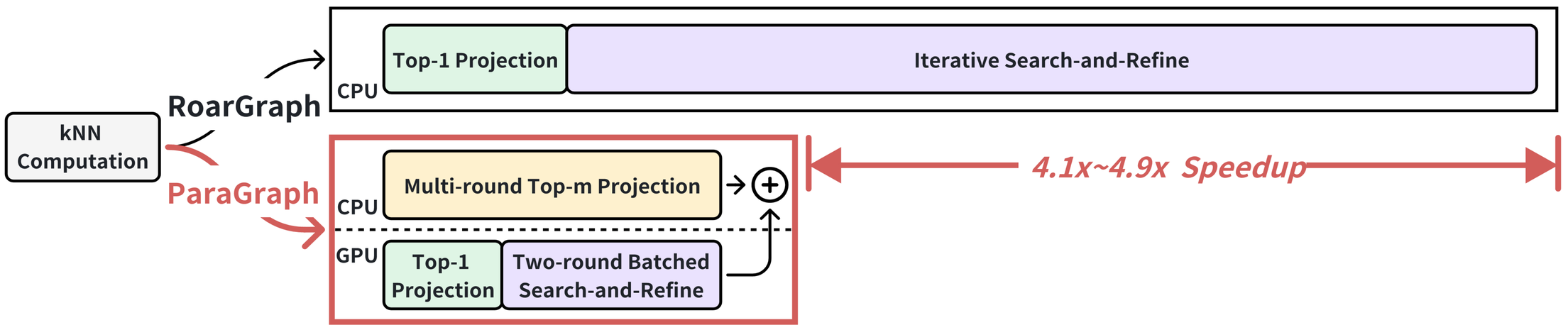
- For each vector, *RoarGraph* **iteratively searches** its nearest neighbors and **refines** the index by updating its neighbor list.
- Simply reducing #iteration will harm the search efficiency.





The key to our solution:

The *PROJECTION* is not only *fast*, but also as *powerful* as the search-and-refine!



Algorithm

New highly-parallelizable algorithms

Multi-round top-m projection + Batched search-and-refine

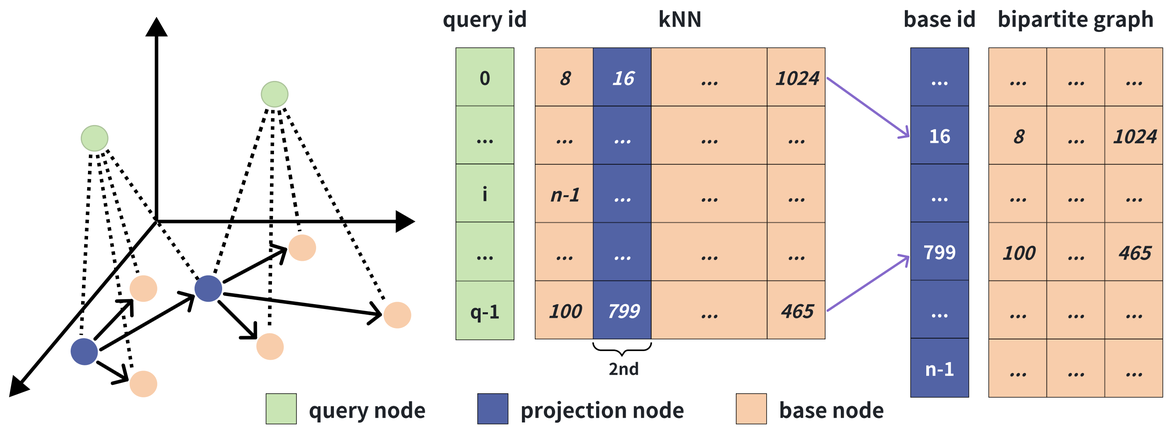


Figure: An example of top-m projection ($m = 2$)

System

Lock-free pruning and reversing in one kernel

Kernel fusion with prune-reverse array

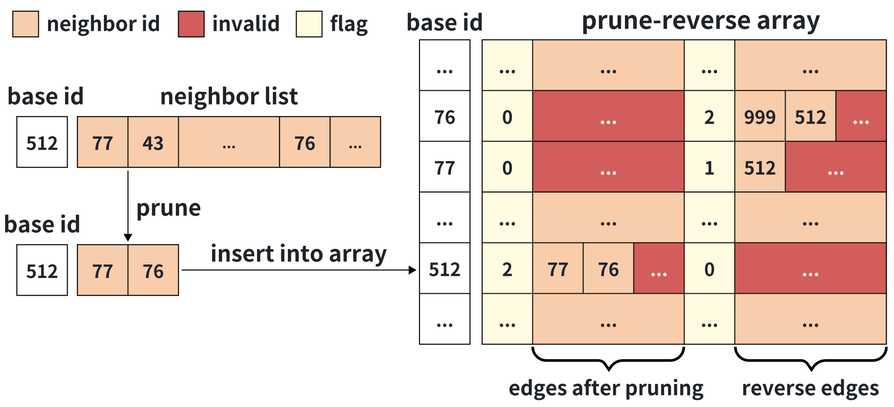
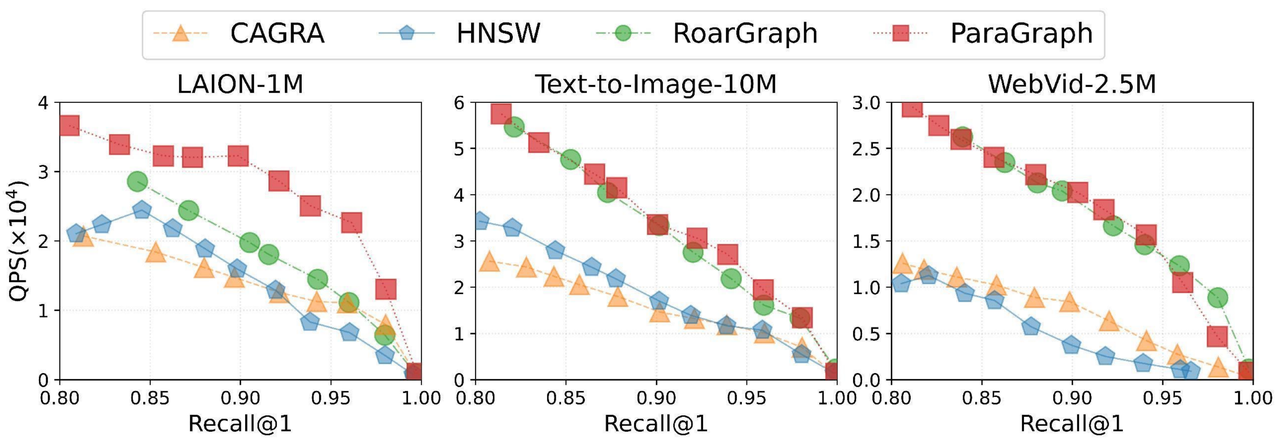


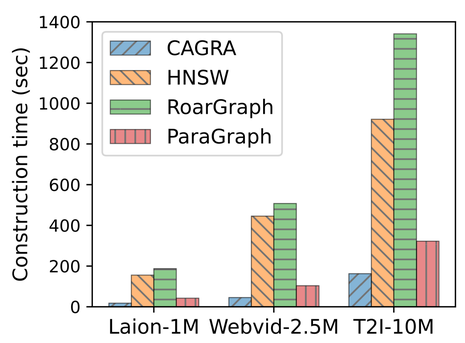
Figure: An example of our new index format prune-reverse array

Experiments Results

Comparable or even better search efficiency than SOTA



Up to 4.9x faster construction



2x reduction in memory footprint