THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

TenGraph: A Tensor-Based Graph Query Engine

Guanghua Li
HKUST (Guangzhou)
Guangzhou, China
gli945@connect.hkust-gz.edu.cn

Hao Zhang
Huawei Cloud Database
Innovation Lab
Beijing, China
zhanghaowuda12@gmail.com

Xibo Sun
HKUST
Hong Kong, China
xsunax@connect.ust.hk

Qiong Luo
HKUST and HKUST (Guangzhou)
Guangzhou and Hong Kong, China
luo@cse.ust.hk

Yuanyuan Zhu
Wuhan University
Wuhan, China
yyzhu@whu.edu.cn

Presenter: Guanghua Li

Overview

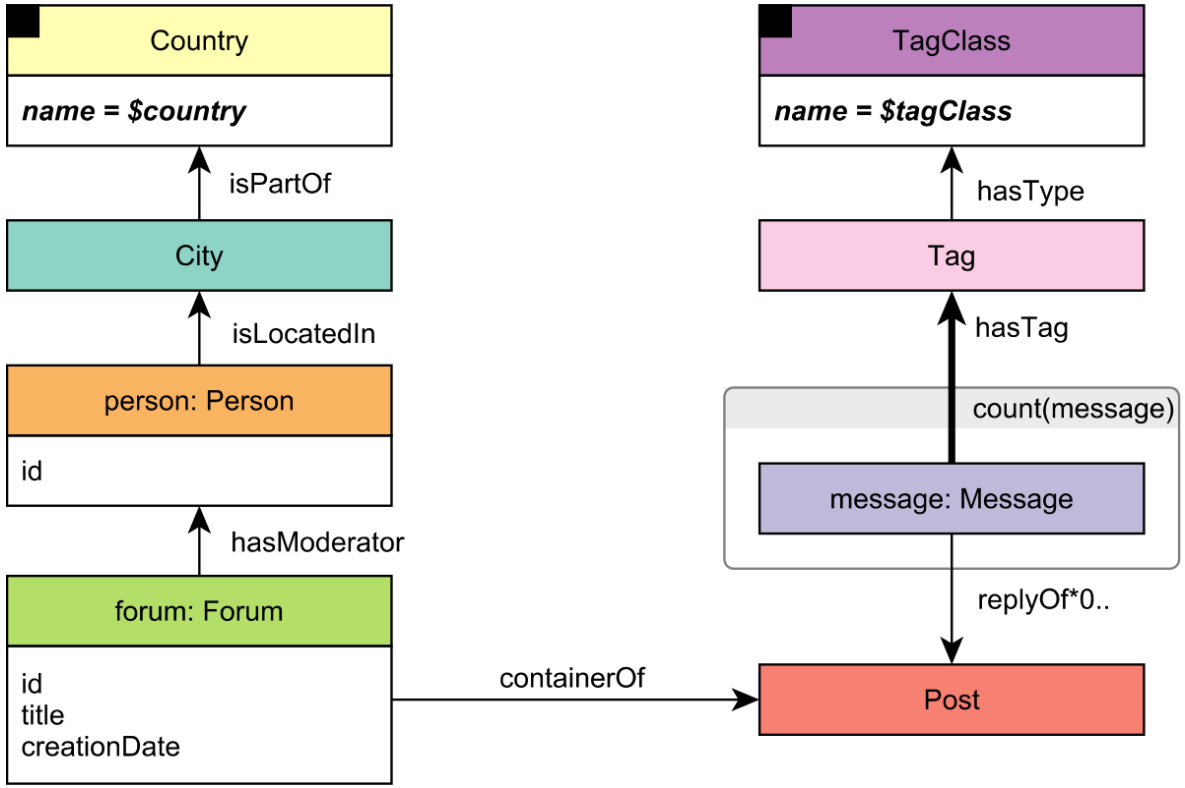
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Graph queries are common in analytical workloads.
- It is challenging to improve graph query performance on different hardware platforms.
- Existing deep learning models run on GPUs through tensor-based frameworks such as PyTorch.
- We explore the use of such tensor-based frameworks to process graph queries on CPUs and GPUs.

Graph Queries

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Graph queries are complex, consisting of pattern matching, filters and aggregates.



Query 3 of the LDBC SNB Businesss Intelligence Benchmark.

The figure is from *Renzo Angles et. al. 2023. The LDBC Social Network Benchmark. https://doi.org/10.48550/arXiv.2001.02299 arXiv:2001.02299*

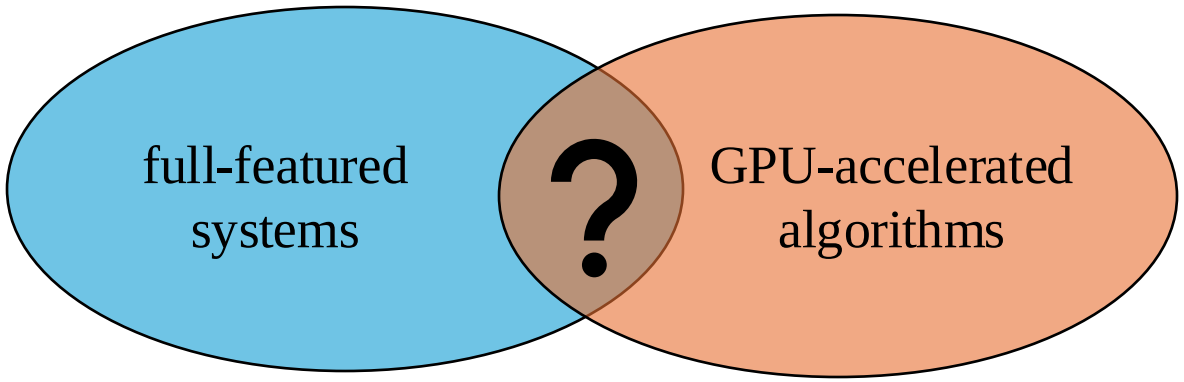
The Problem

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Existing work on graph query processing

either based on the CPU

or limited to accelerating a subset of graph queries on the GPU, through CUDA-based implementation.

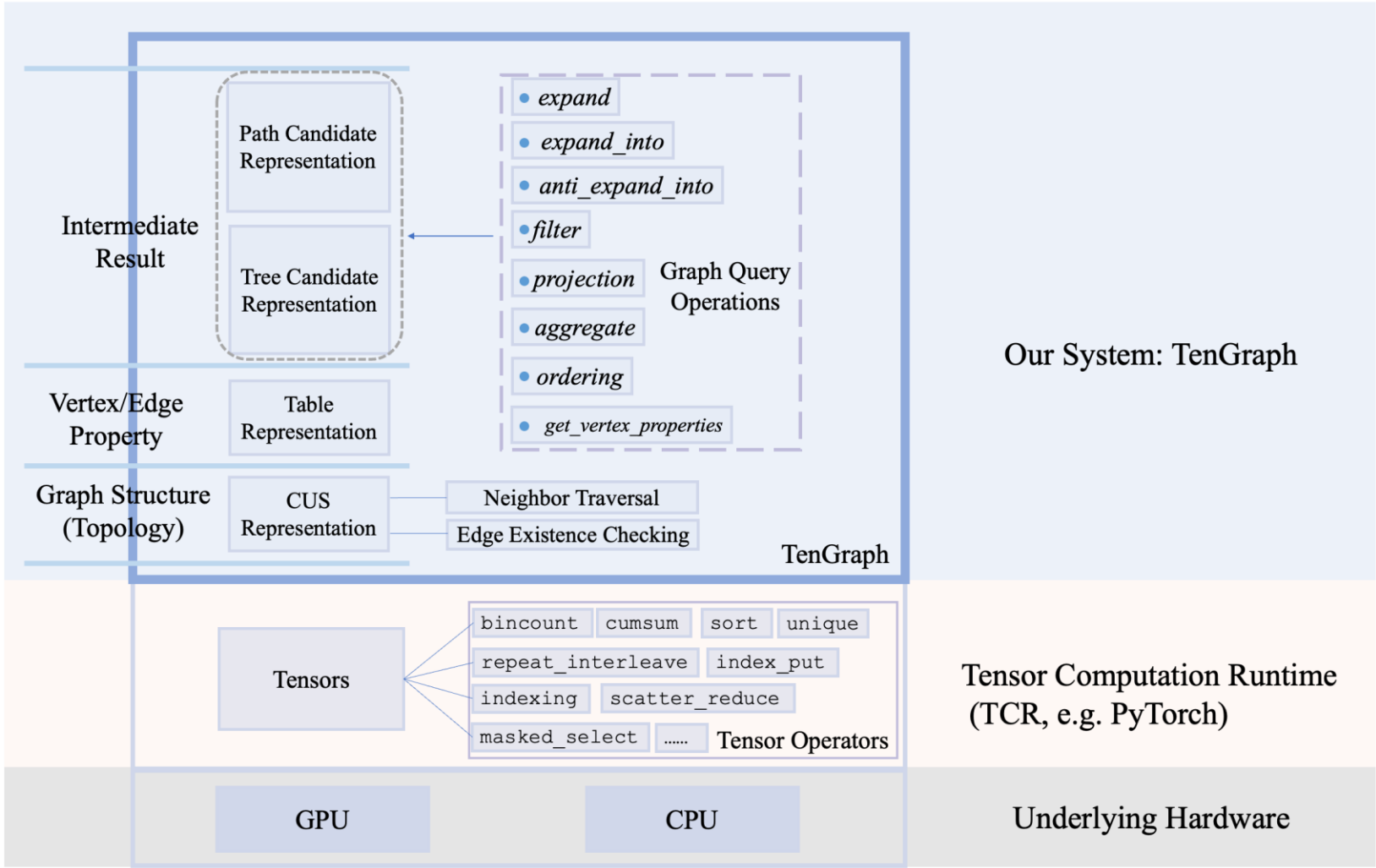


How to process full graph queries on GPUs automatically ?

Our Solution

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

To build our system upon Tensor Computation Runtime (TCR).



Tensors

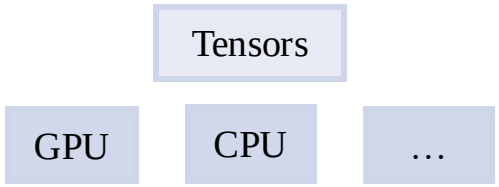
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

A tensor is a multi-dimensional array containing elements of a single data type.

Widely used data abstraction in deep learning frameworks such as PyTorch and TensorFlow.

PyTorch TensorFlow

Tensor-based frameworks run on various devices.



Tensor-based operators are highly optimized.

`cumsum` `index_select` `repeat_interleave`
`masked_select` `scatter_reduce` `……`

TenGraph has the potential to integrate deep learning methods.

TenGraph is cross-platform.

TenGraph is easy to implement and achieves automatic performance improvement.

Contributions

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- We show that graph queries can be expressed by a small set of tensor operators and we build the first tensor-based graph query engine, TenGraph.
- We implemented 16 of the 20 Business Intelligence queries of the LDBC SNB benchmark on top of TenGraph and evaluated its performance in comparison with the state-of-the-art.
- Our experiments show that TenGraph-GPU is up to two orders of magnitude faster than leading CPU-based graph databases. It is also up to five times faster than our own implementation of TQP, a tensor-based relational query engine, on the GPU.

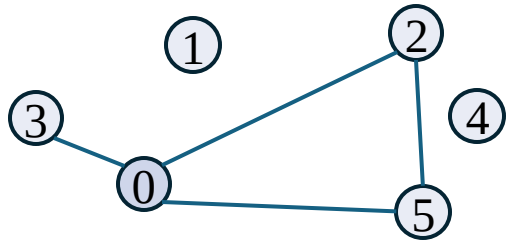
Design and Implementation - Neighbor Traversal as An Example

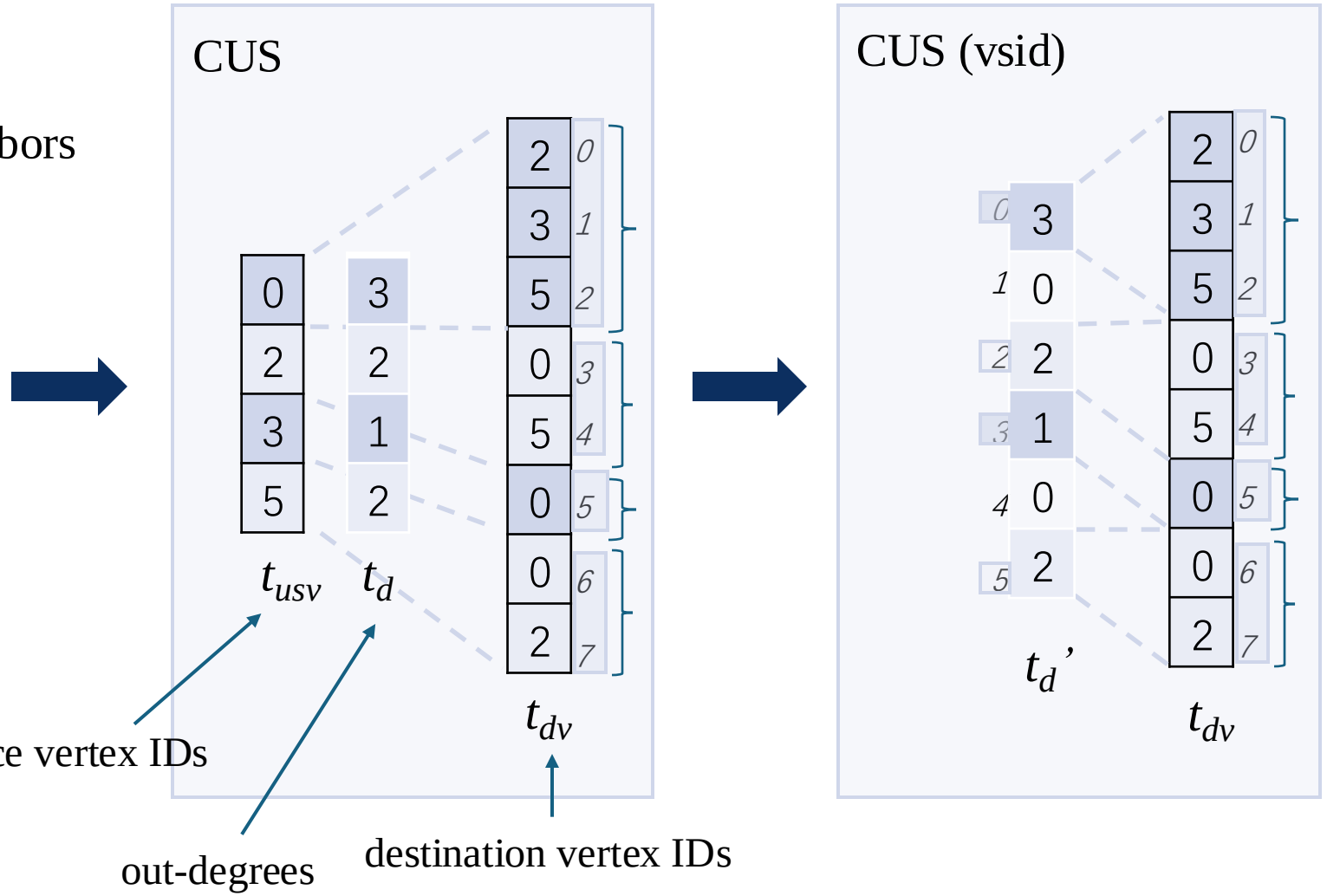
Graph Structure Representation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

CUS (Compressed Unique Source) Representation

Vertices	Out-degrees	Out-neighbors
0	3	[2, 3, 5]
2	2	[0, 5]
3	1	[0]
5	2	[0, 2]





Neighbor Traversal

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Input

inputIds: a tensor of input vertex IDs;

Output

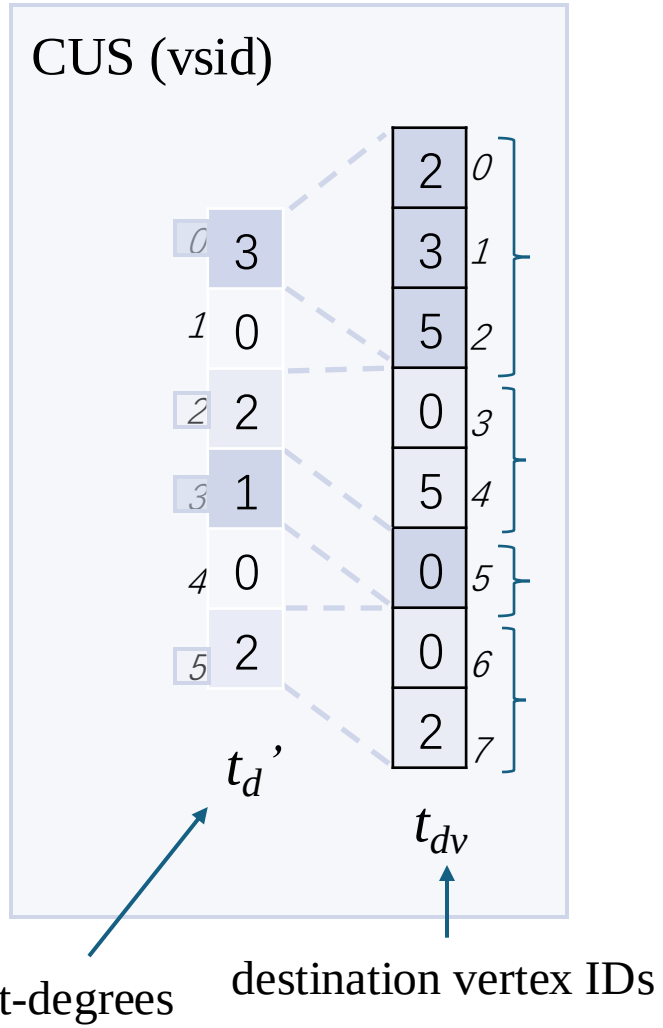
1. *neighborCounts*
2. *neighbors*.

Example:

Input: tensor([3, 2, 2, 4, 0])

Output: tensor([1, 2, 2, 0, 3]), tensor([0, 0, 5, 0, 5, 2, 3, 5])

1 *neighborCounts* 2 *neighbors*



Neighbor Traversal

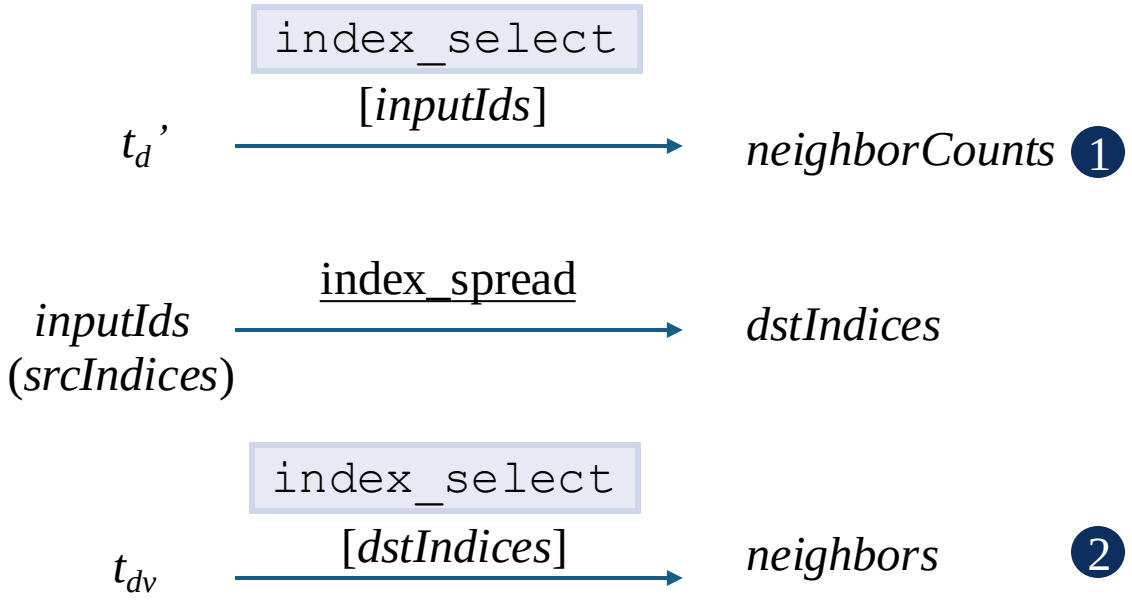
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

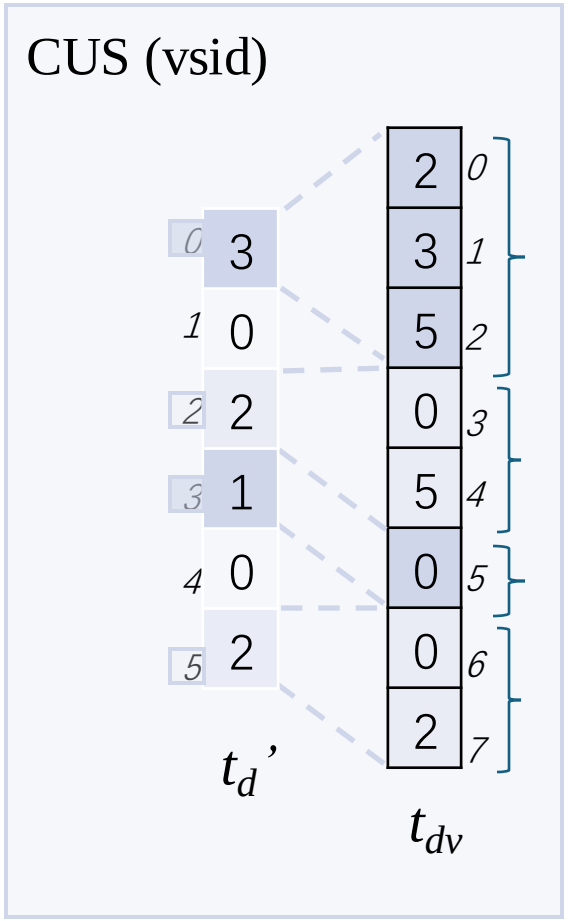
Input

inputIds: a tensor of input vertex IDs;

Output

1. *neighborCounts*
2. *neighbors*.





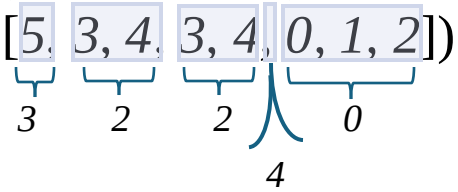
Index Spread

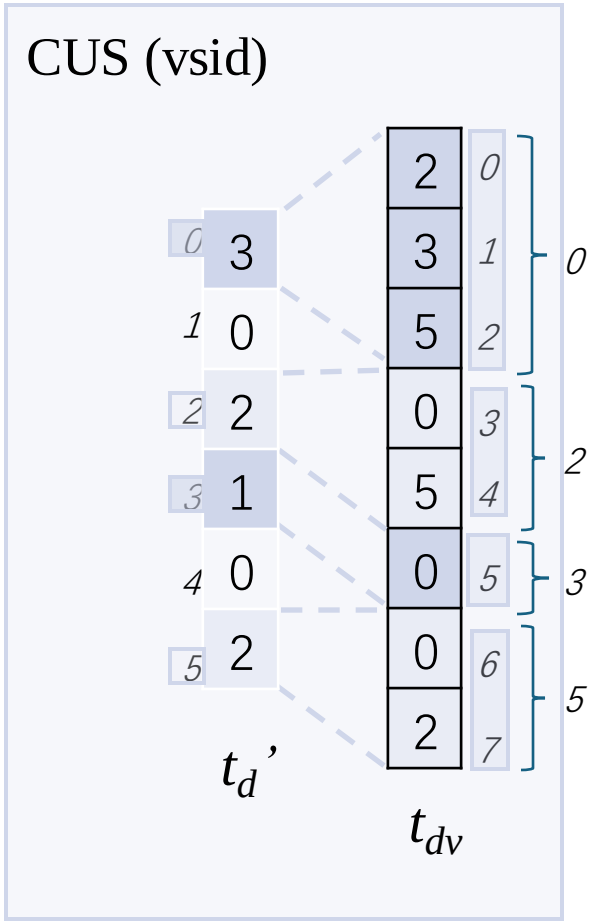
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Input
srcIndices: indices for source vertices;

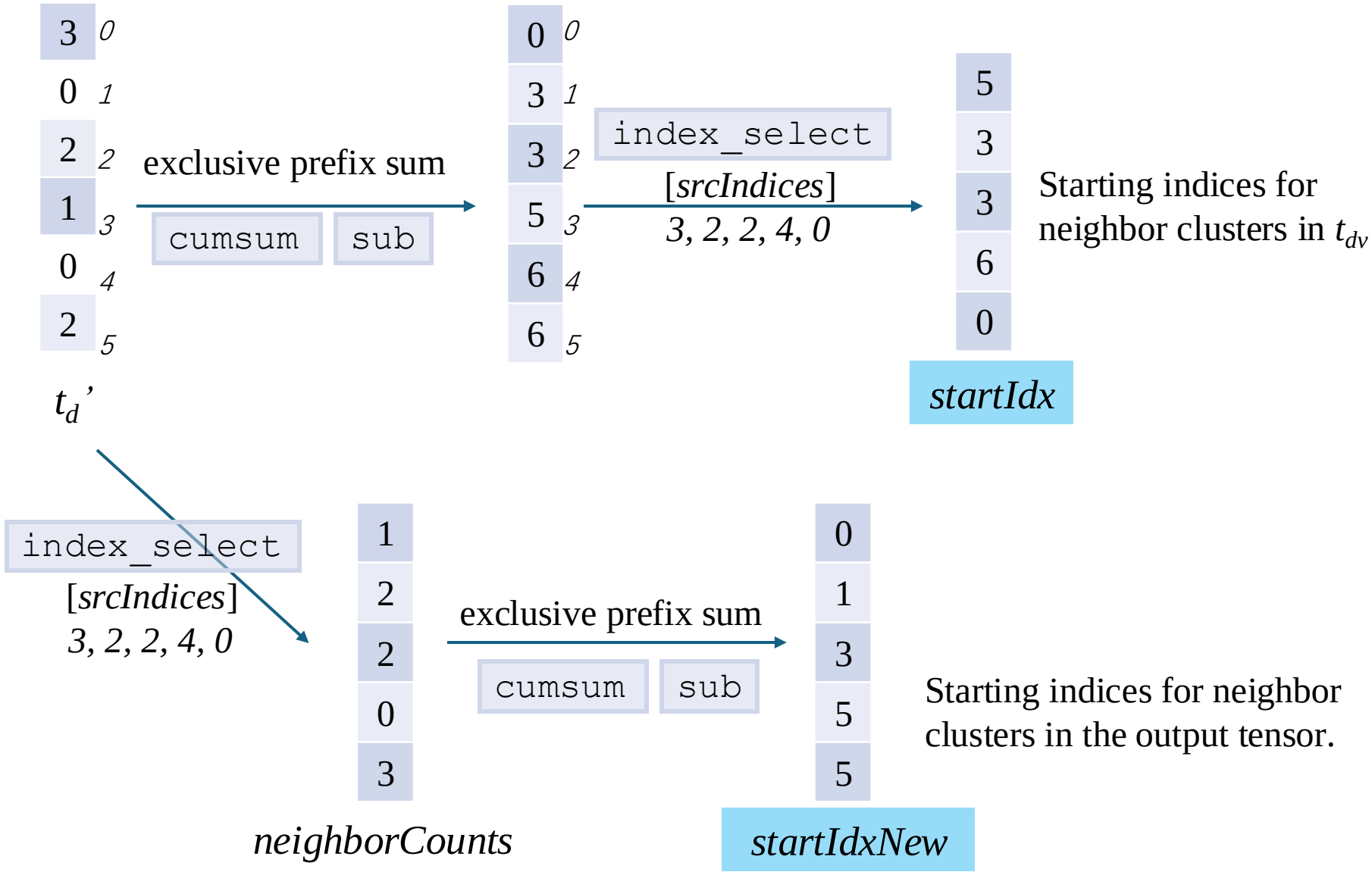
Output
dstIndices: indices for destination vertices.

Example:
Input: tensor([*3, 2, 2, 4, 0*])
Output: tensor([*5, 3, 4, 3, 4, 0, 1, 2*])

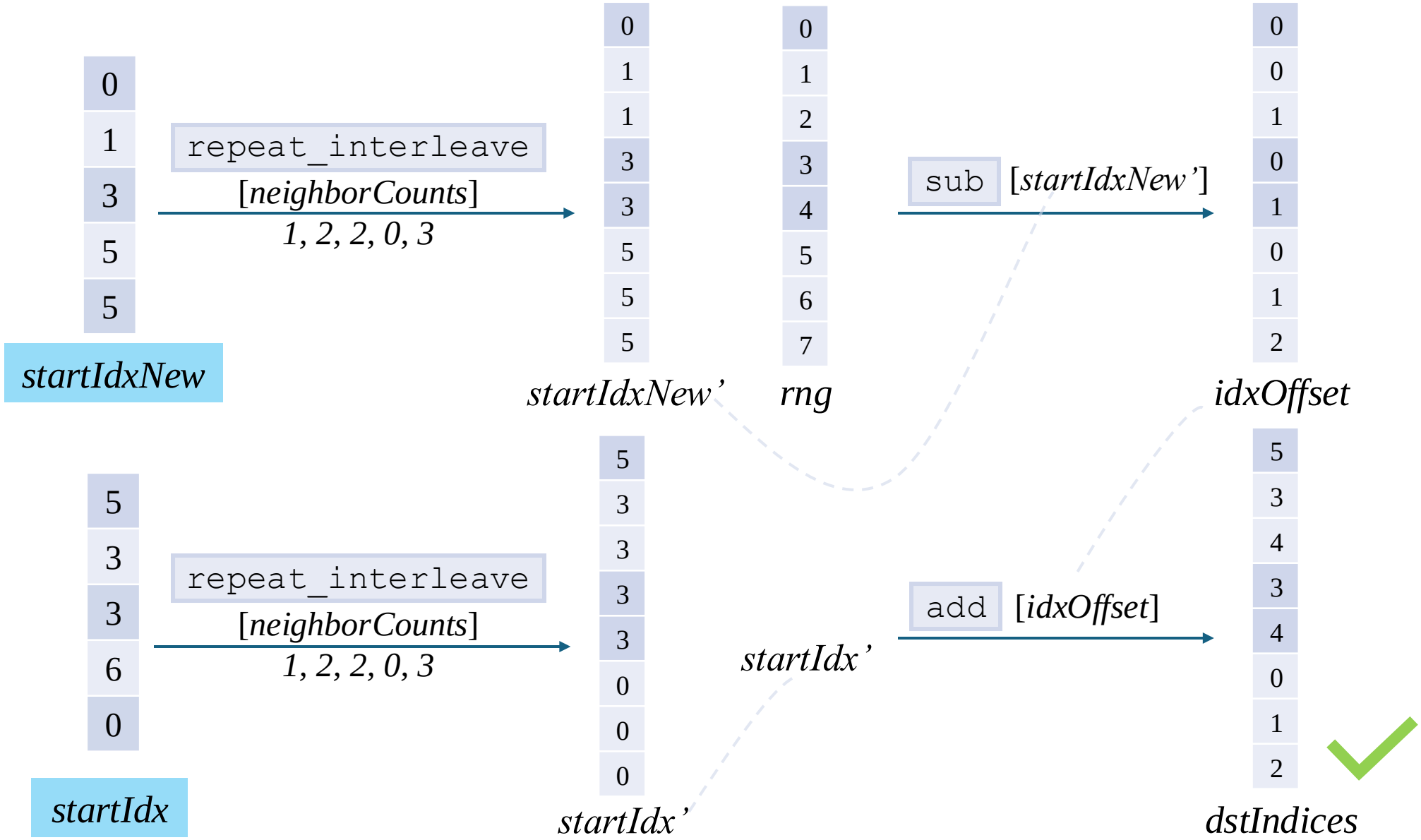




Index Spread – Getting startIdx and startIdxNew



Index Spread – Computing dstIndices

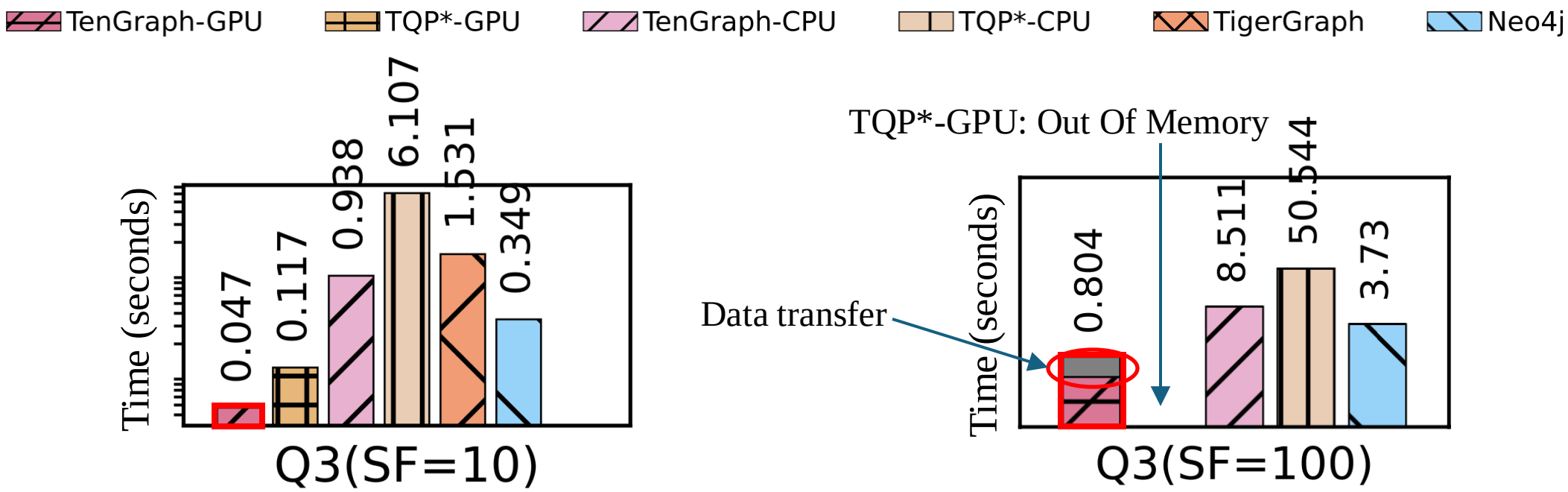


Experimental Results

LDBC Query 3 Performance

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

TenGraph-GPU is significantly faster than CPU-based systems.



Query time of LDBC SNB BI Query 3

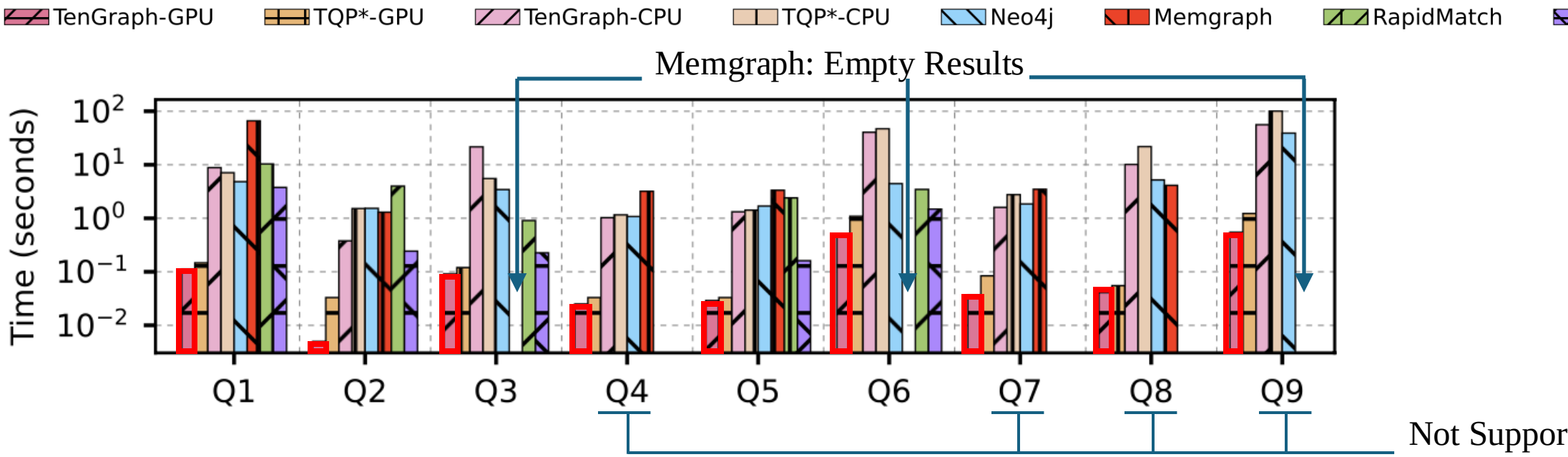
CPU: two AMD EPYC 7302 CPUs (64 virtual cores in total), with 125 GiB of RAM
GPU: one NVIDIA RTX 3090 GPU, with 24 GiB of device memory.

Subgraph Matching Performance

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

TenGraph supports all 9 queries.

TenGraph-GPU is faster than all other methods by up to two orders of magnitude.



Query time (in seconds) on LSQB. (SF=1, #vertices: 3.9M, #edges: 19.5M)

(LSQB* is a micro-benchmark for subgraph matching, derived from the LDBC benchmark.)

** Amine Mhedhbi et. al. 2021. LSQB: a large-scale subgraph query benchmark. (GRADES- NDA '21). https://doi.org/10.1145/3461837.3464516*

Conclusion

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- We build the first tensor-based graph query engine, TenGraph, to automatically take advantage of modern hardware such as multi-core CPUs and GPUs.
- We design the data structure to represent graph topology with tensors and show how to perform various operations on the data structure with tensor operators.
- The performance of TenGraph-CPU is comparable to leading graph databases and TenGraph-GPU is significantly faster than them.
- On the subgraph matching micro-benchmak, TenGraph-GPU also outperforms state-of-the-art CPU- and GPU-based methods.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Thanks

Email: gli945@connect.hkust-gz.edu.cn

Presenter: Guanghua Li