BOSTON UNIVERSITY

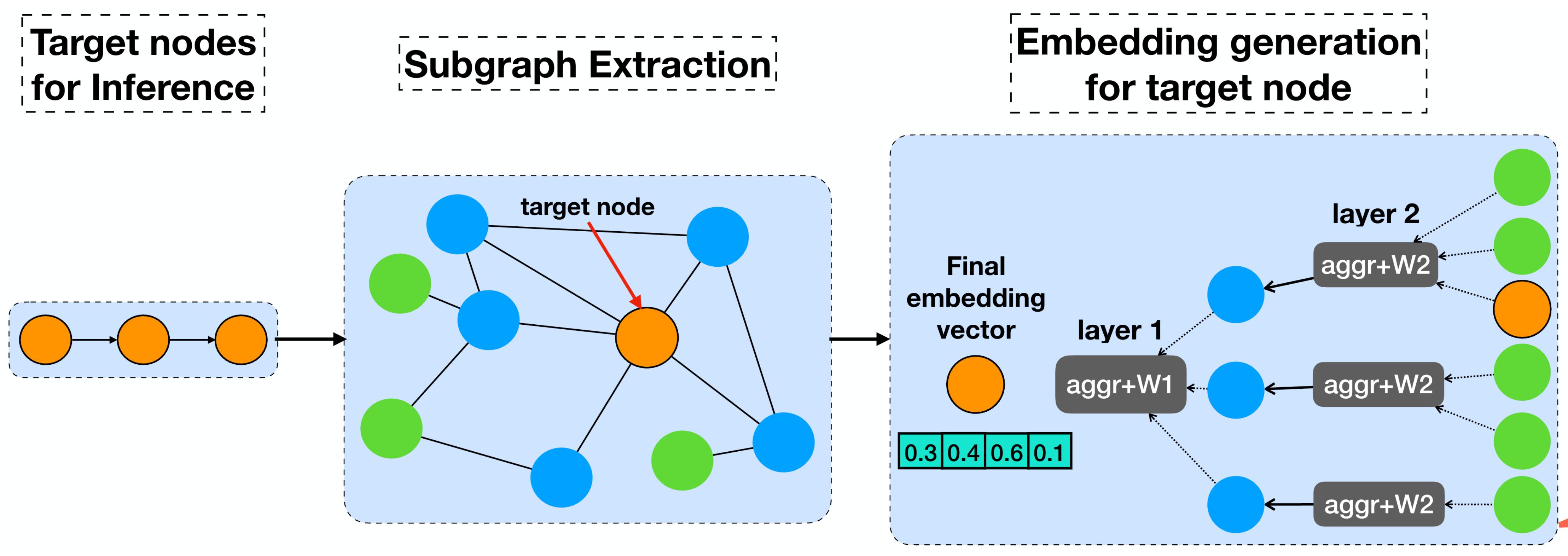
Bridging GNN Inference and Dataflow Stream Processing: Challenges and Opportunities

CASP Systems

Naima Abrar
anaima@bu.edu

Vasiliki Kalavri
vkalavri@bu.edu

Graph Representation Learning and GNN Inference

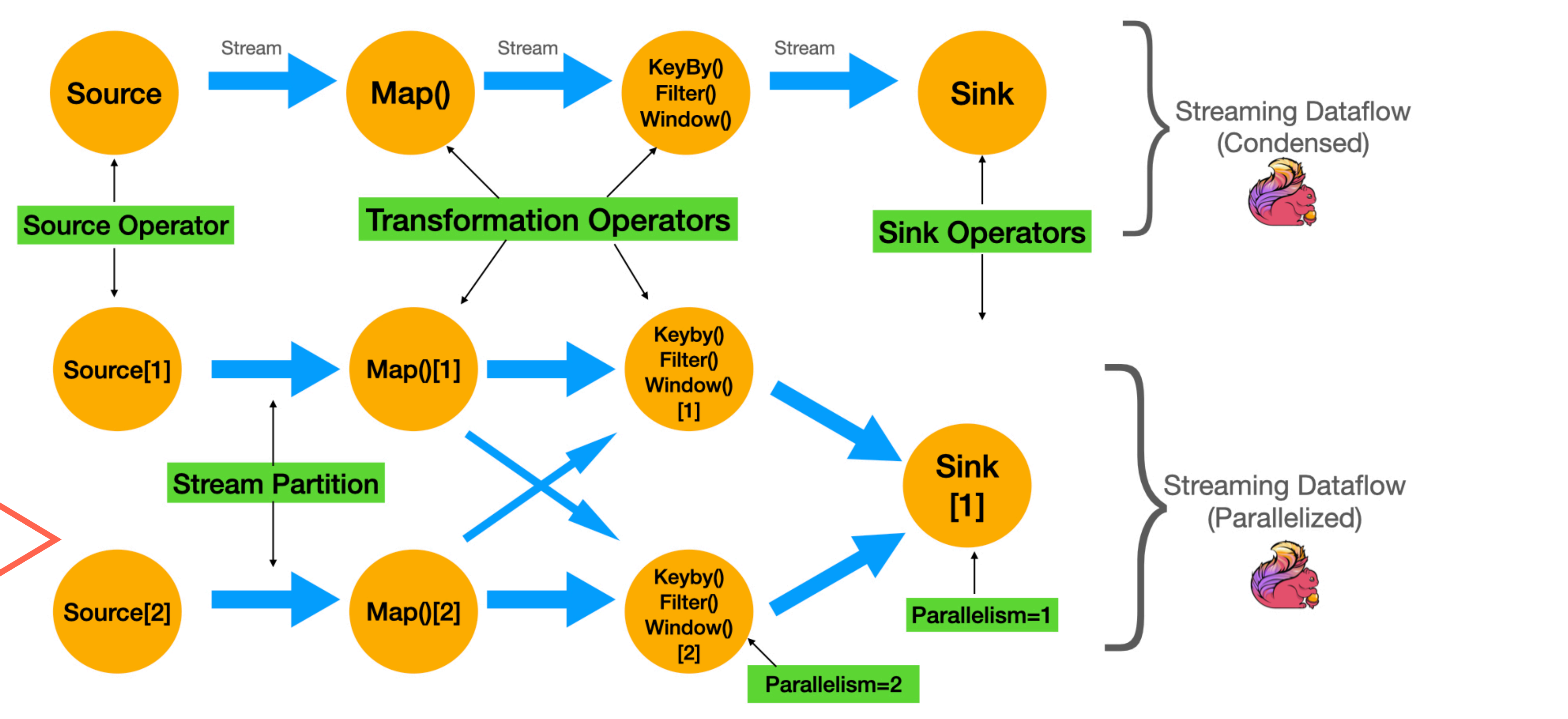


- GNN inference is triggered for a target node.
- A k-hop neighborhood is constructed around the target node.
- Features from the subgraph are retrieved.
- A GNN model applies aggregation and transformation layers to produce a final embedding.

- Dataflow systems like Apache Flink execute streaming programs as DAGs of operators (map, filter, keyBy).
- Operators are parallelized and maintain keyed state (e.g per node) using embedded stores like RocksDB.

GNN inference maps cleanly to this dataflow

- Each request involves stateful access (e.g. neighbors, features),
- Graph data is stored as KV pairs
- And node IDs act as keys for partitioned, per-node subgraph expansion.



Challenges of GNN Inference with Current Dataflow Systems

Lack of iterative support

No cycles in current dataflow systems

Keyed state prevents k-hop neighborhood access

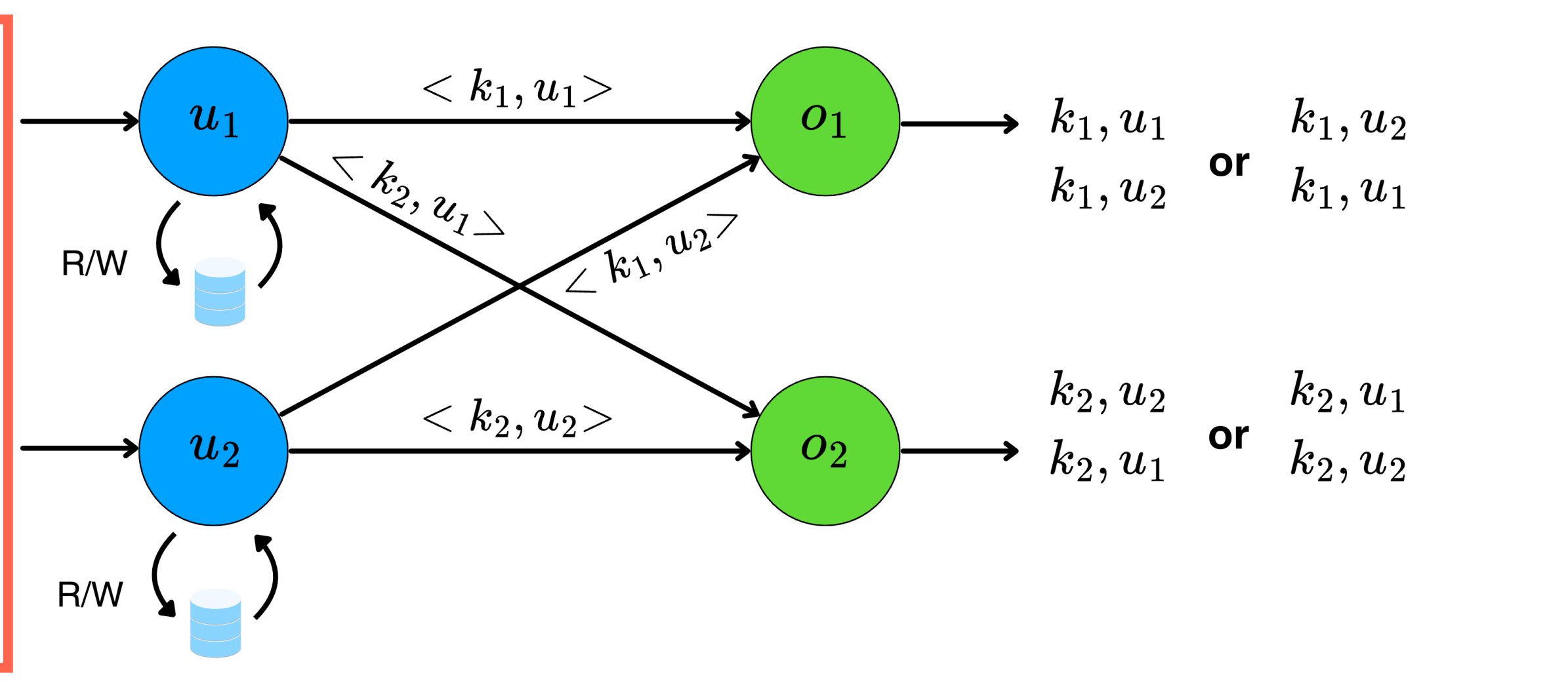
State access restricted per key

Lack of subgraph completion signal

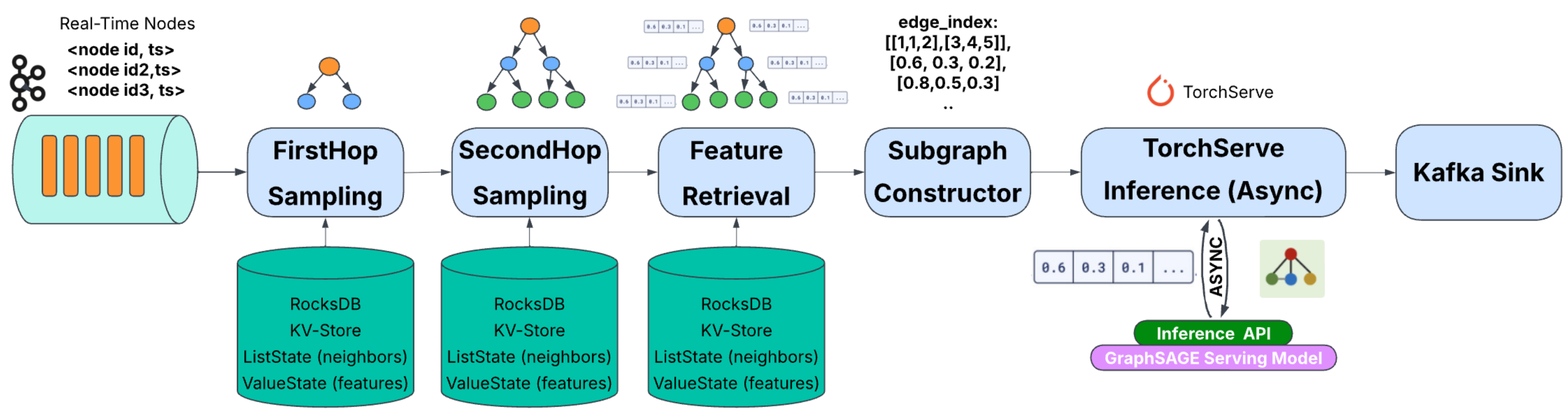
Due to pipelined and parallel execution, no built-in notion of when a subgraph is complete

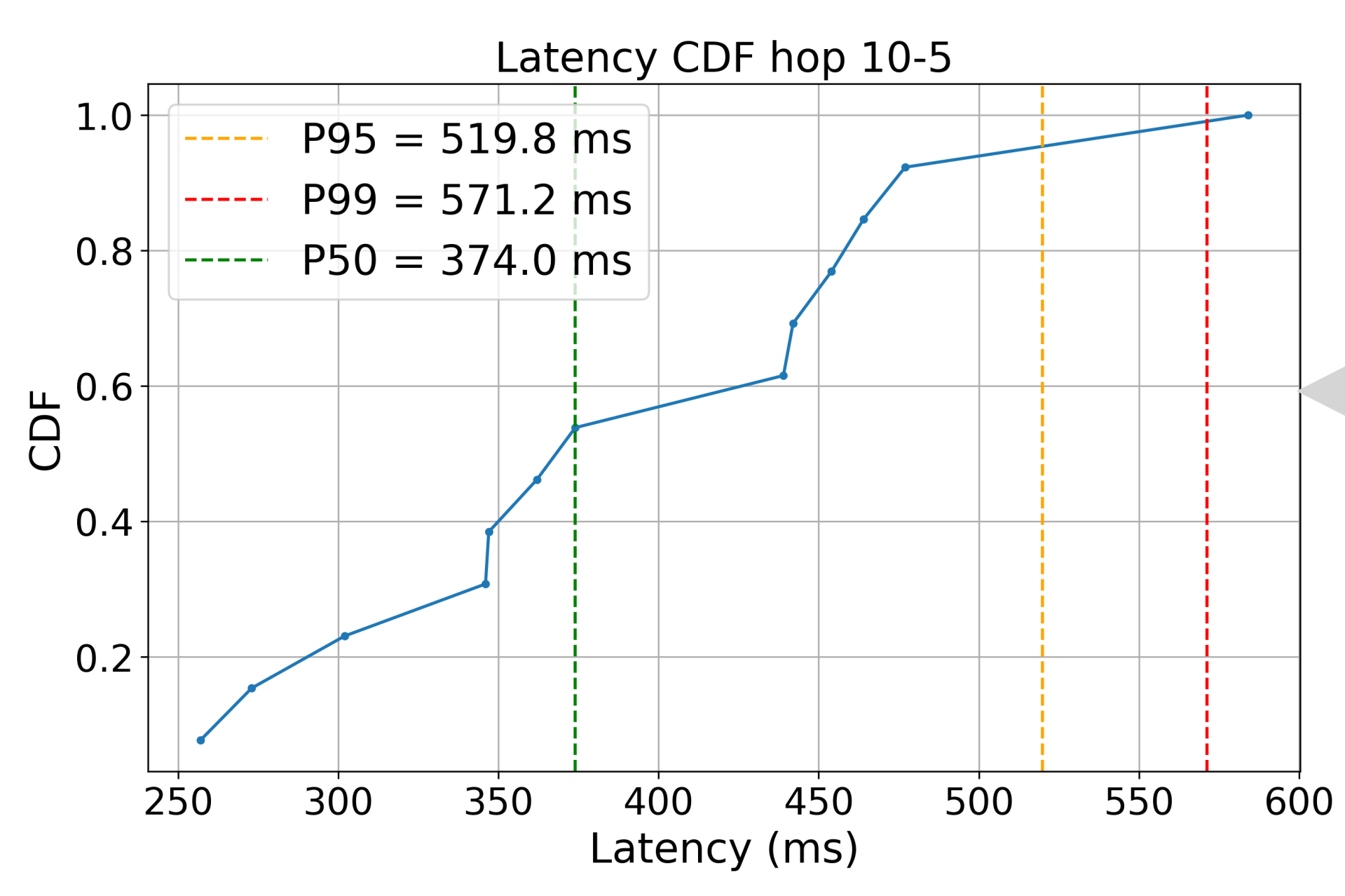
Graph state loading

Graph must be preloaded via an initialization job


u_1
u_2
o_1
o_2
$< k_1, u_1 >$
$< k_2, u_1 >$
$< k_1, u_2 >$
$< k_2, u_2 >$
R/W
R/W
k_1, u_1 k_1, u_2 or k_1, u_2 k_1, u_1
k_2, u_2 k_2, u_1 or k_2, u_1 k_2, u_2

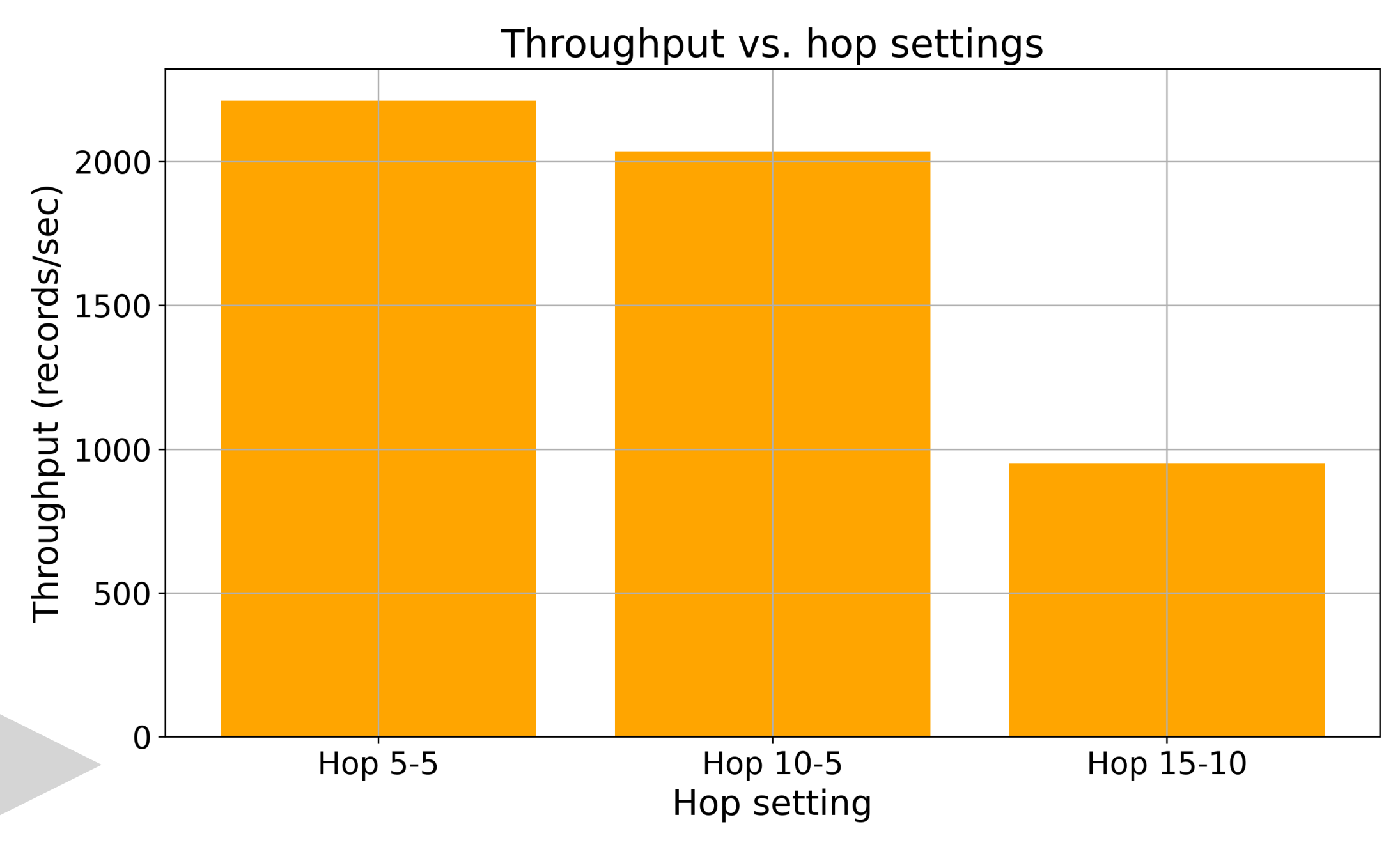
Pipeline for Streaming GNN Inference





50% of the requests complete in under 374 ms, 95% complete within 519 ms, and 99% within 571 ms.

Larger neighborhoods result in more data transfer, state access, and computation per request, which can directly impact overall performance



Future Work

- Incorporate dynamic graph updates (node/edge addition) directly into the real-time inference process
- Enhance neighborhood sampling with temporal awareness (e.g retrieving latest or Top-K neighbors based on timestamps)
- Integrate with external graph databases for enhanced state management flexibility and query capabilities.
- Implement advanced graph partitioning strategies within the dataflow system to improve locality and reduce communication.

Naima Abrar Shami and Vasiliki Kalavri. Bridging GNN Inference and Dataflow Stream Processing: Challenges and Opportunities. In Proceedings of the 8th Joint Workshop on Graph Data Management Experiences & Systems and Network Data Analytics (GRADES-NDA '25). ACM, 2025. (to appear).