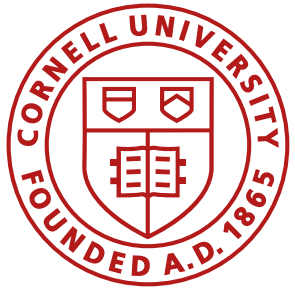
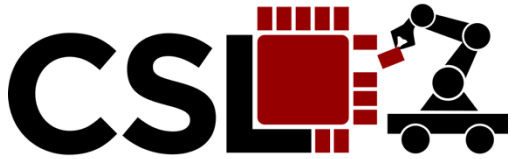


LongSight: Compute-Enabled Memory to Accelerate Large-Context LLMs via Sparse Attention

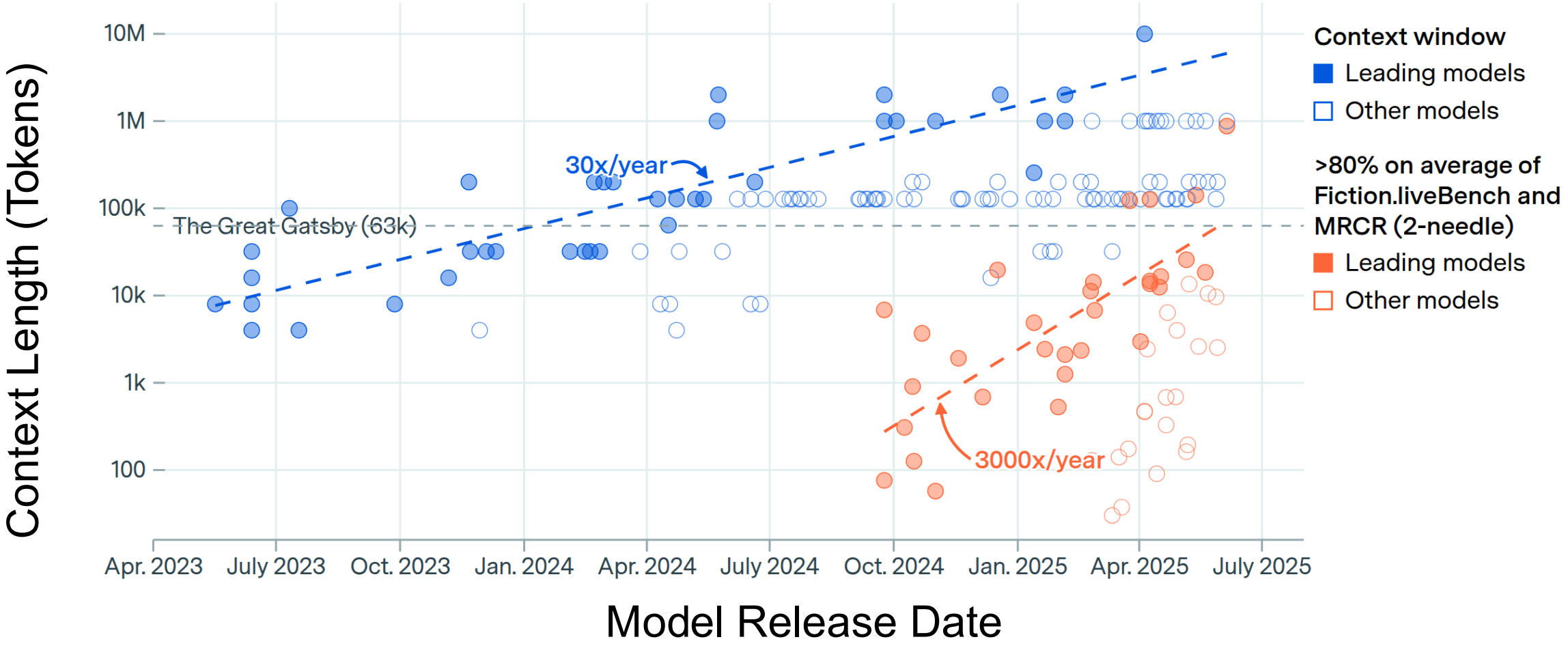
Derrick Quinn, E. Ezgi Yücel, Jinkwon Kim, José F. Martínez, Mohammad Alian

Cornell University

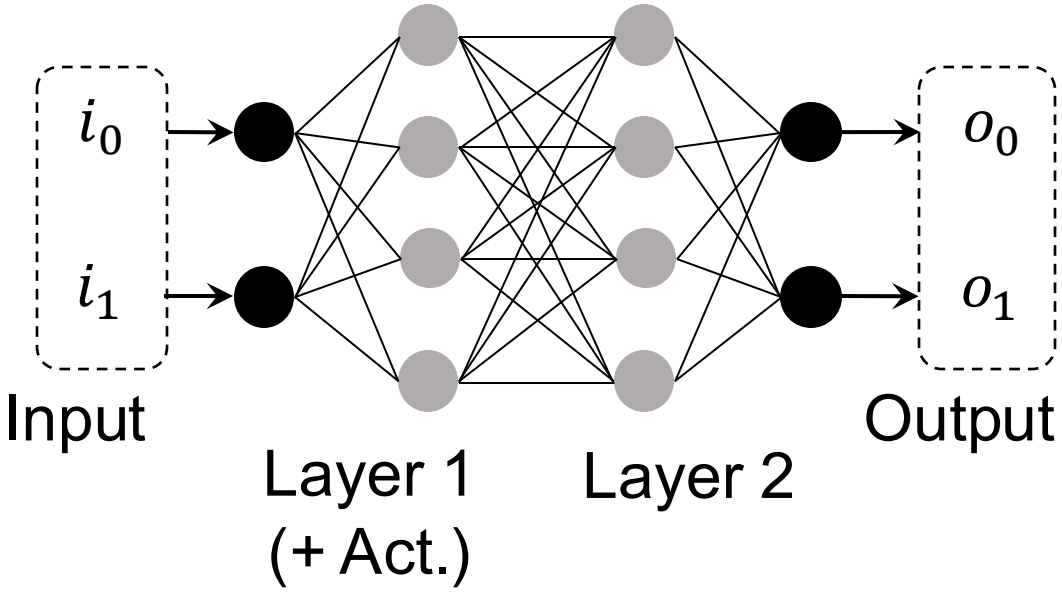
CornellEngineering
Electrical and Computer Engineering



Context Lengths Over Time (Epoch AI)



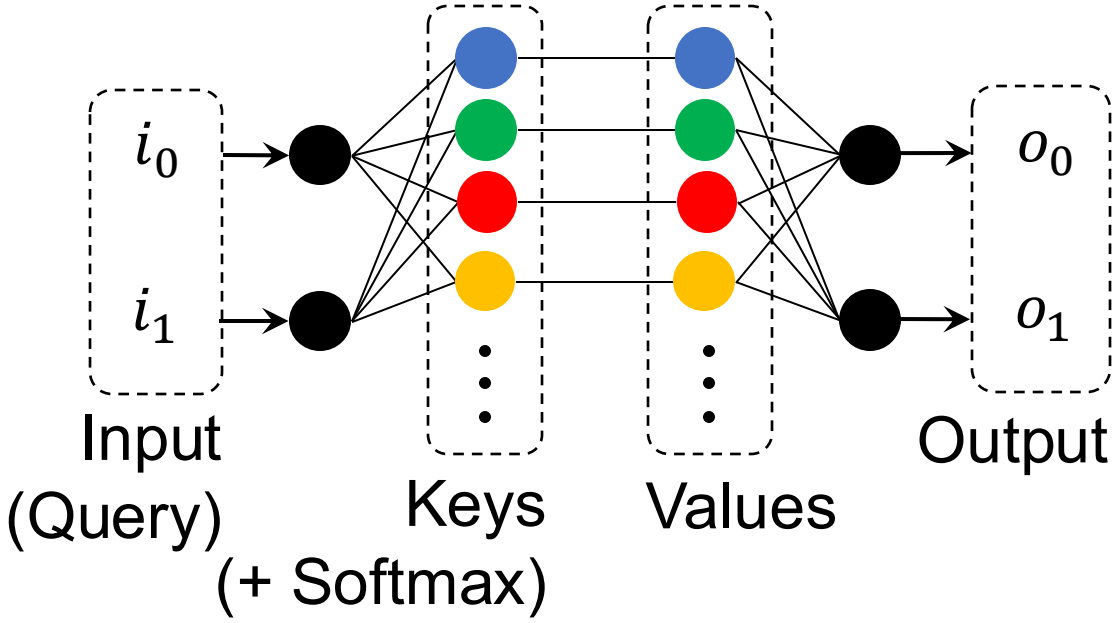
Feed-Forward Network (FFN)



Parameters: $O(D^2)$

Weights: Shared across users

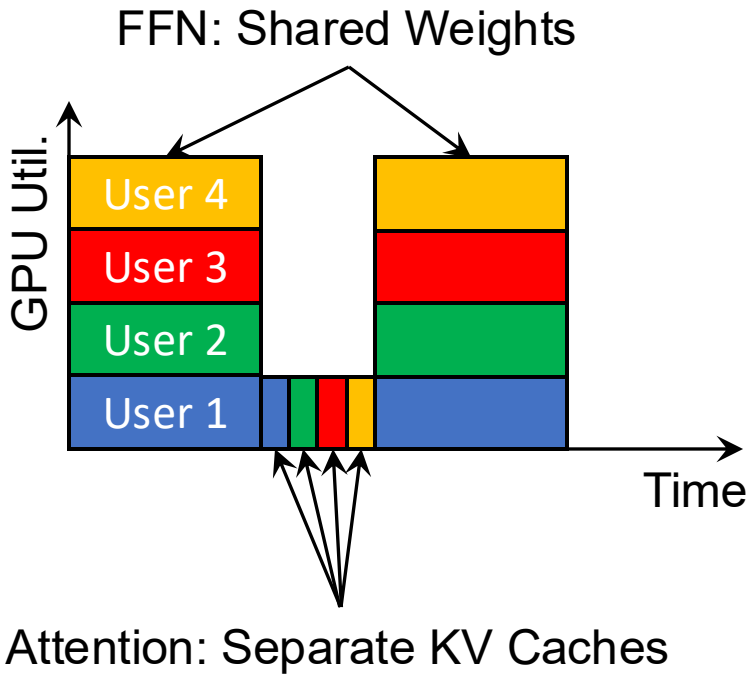
Attention



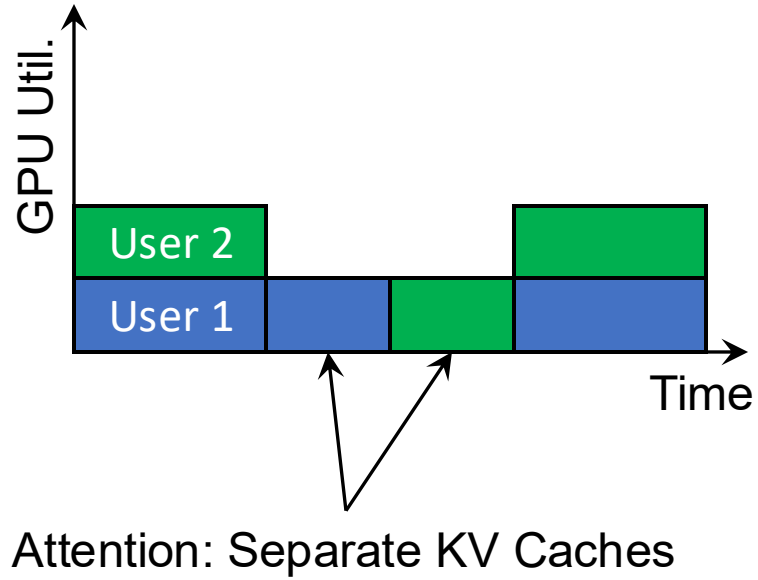
KV Cache "Parameters": $O(D \times \text{Ctx.})$

Weights: Separate per user

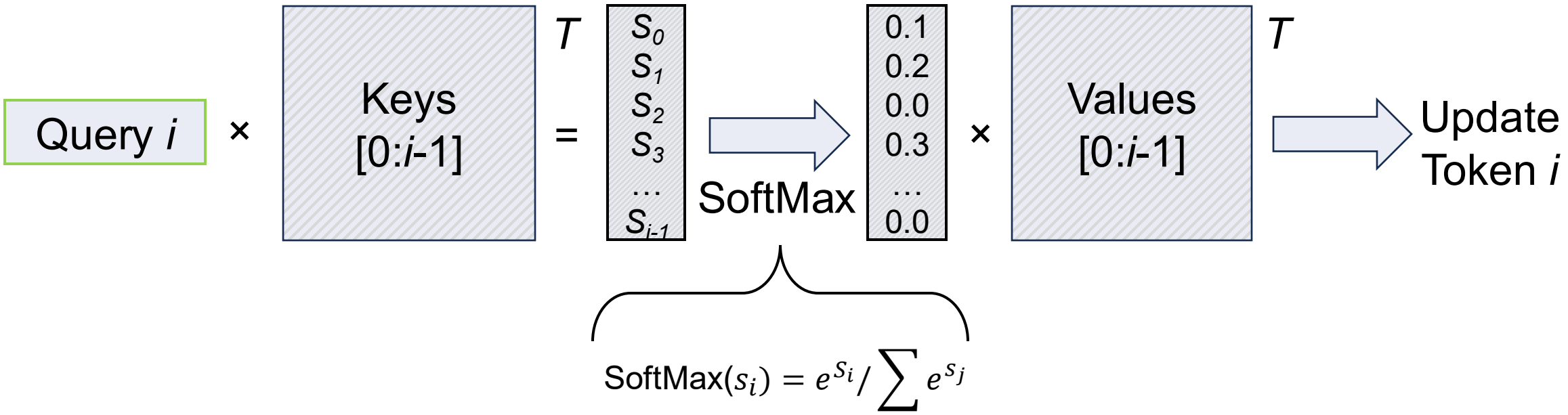
Batched Inference: Short Context

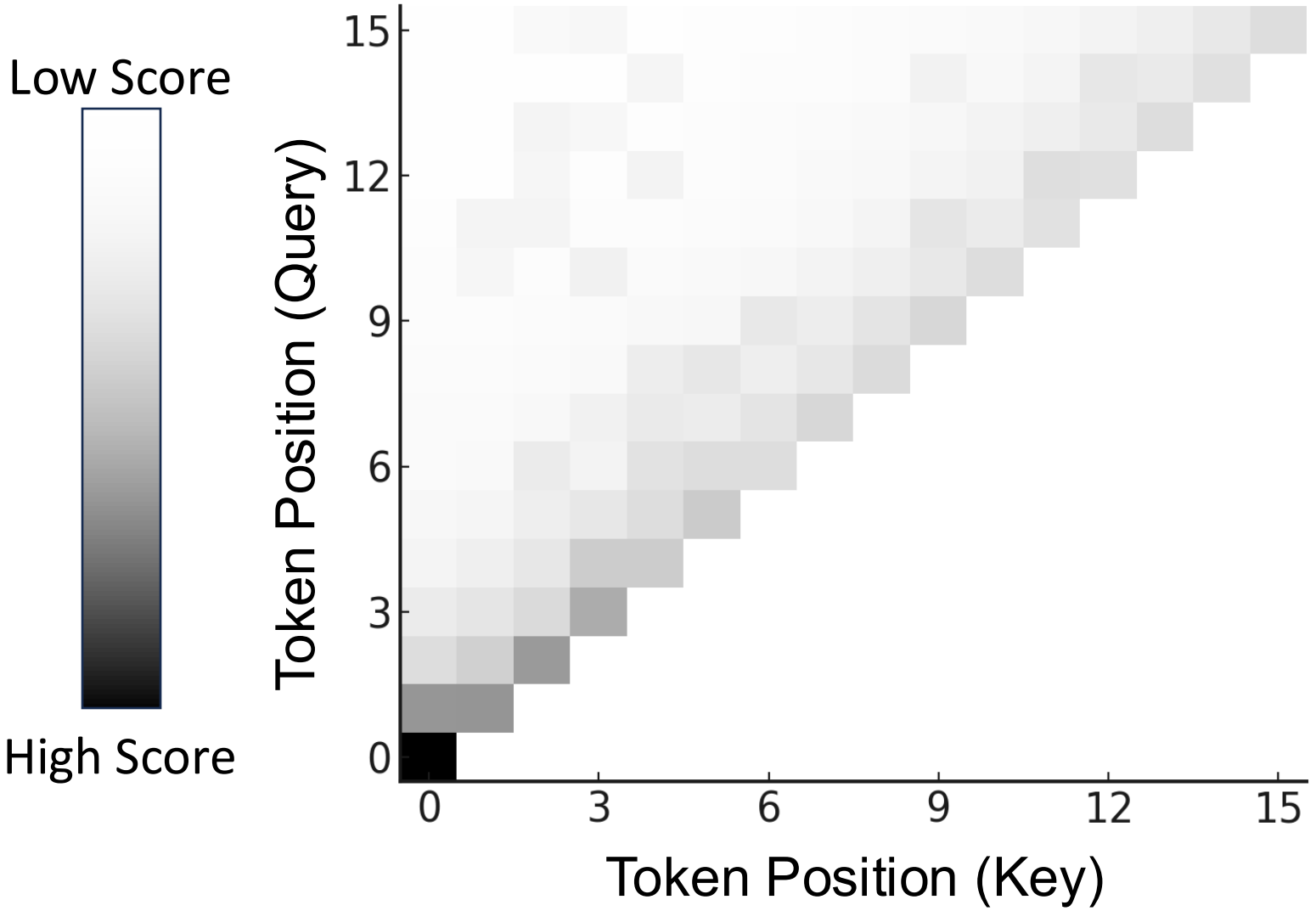


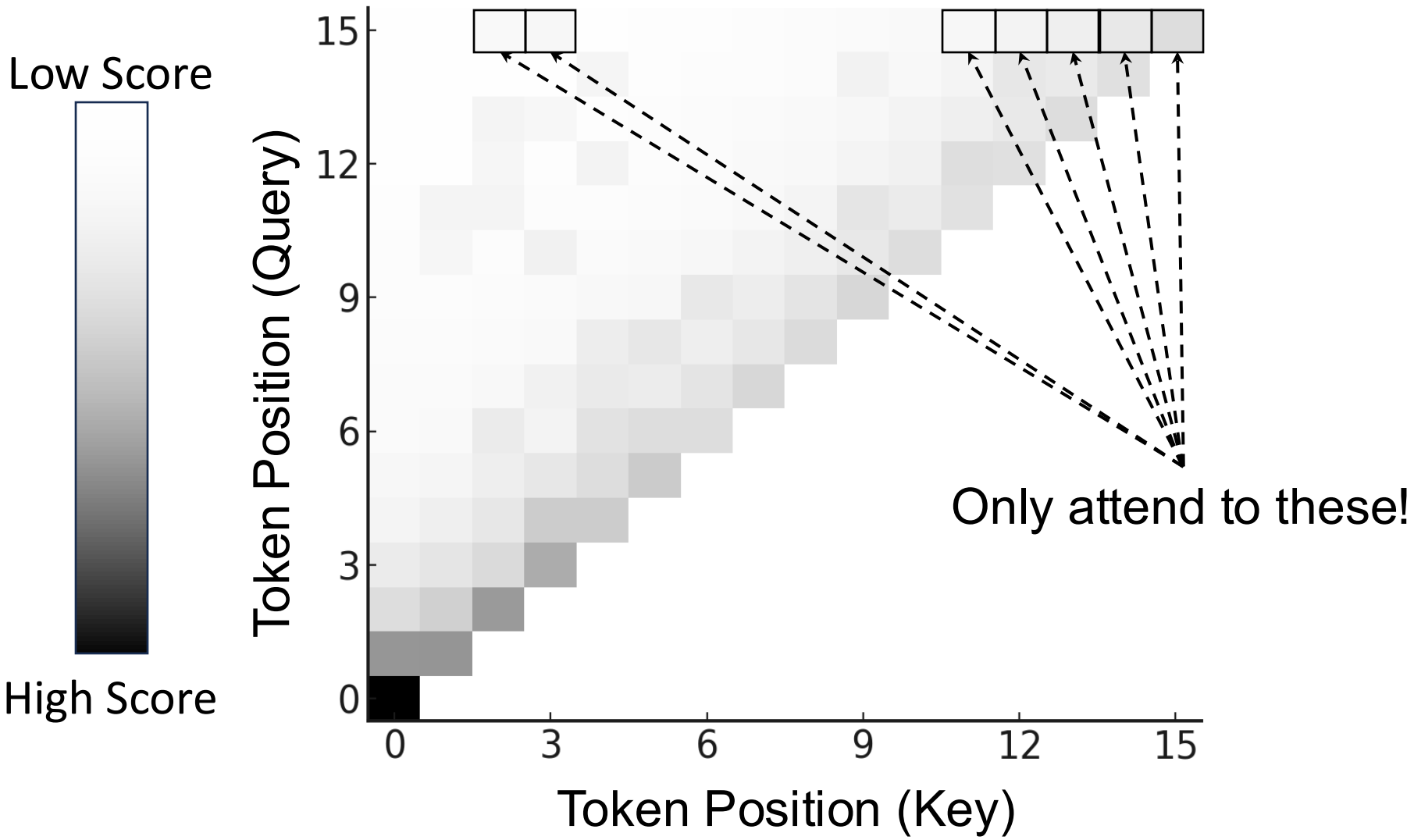
Batched Inference: Long Context



Opportunity: Sparse Attention




Low Score
High Score
Token Position (Query)
Token Position (Key)
0
3
6
9
12
15

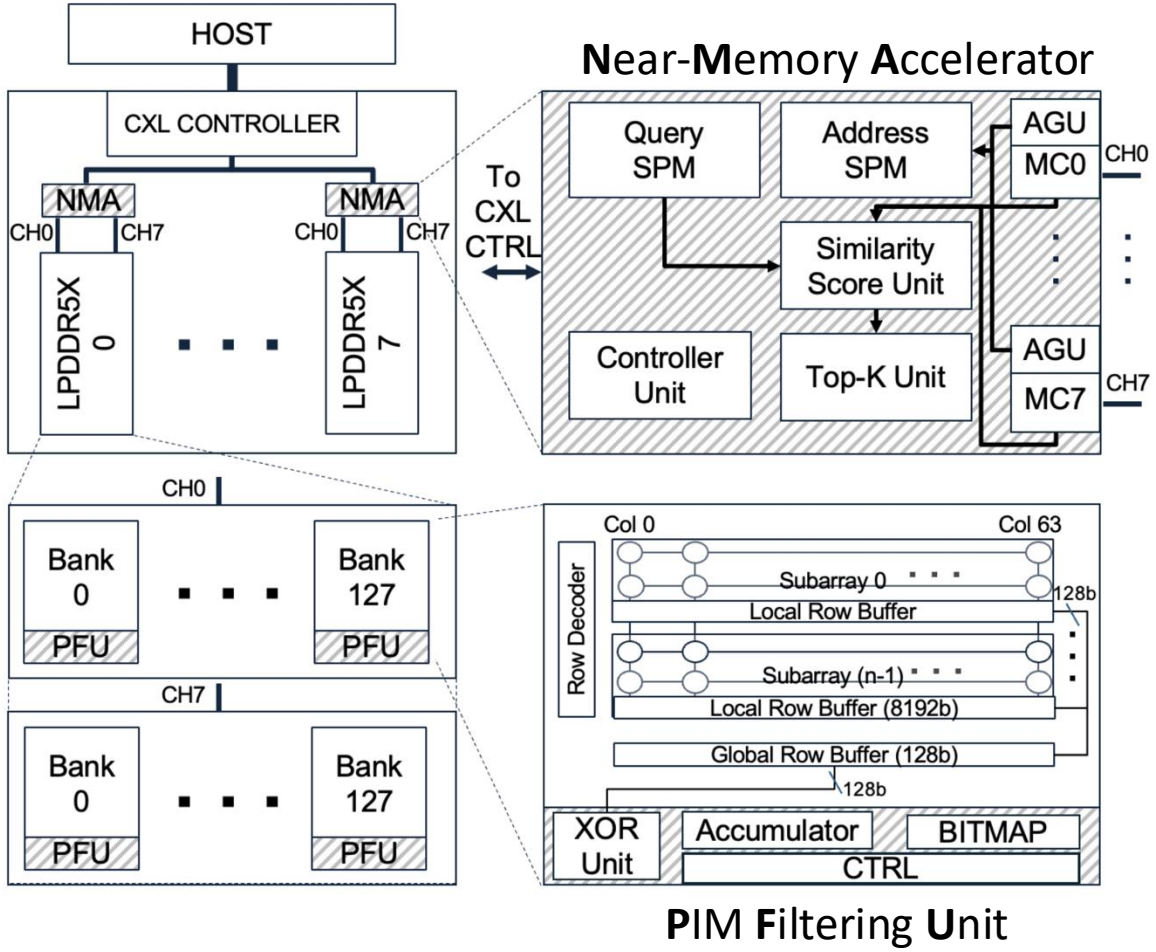

Low Score
High Score
Token Position (Query)
Token Position (Key)
0
3
6
9
12
15
Only attend to these!

Proposal: Top-K Sparse Attention

1. Store KV cache in a high-capacity, compute-enabled CXL memory
2. Filter out less relevant keys **inside DRAM**
3. Find Top-K Keys with the highest attention score **near DRAM**
4. Ship the **Top-K Keys and Values** to GPU for Attention

DReX (ISCA 2025*)

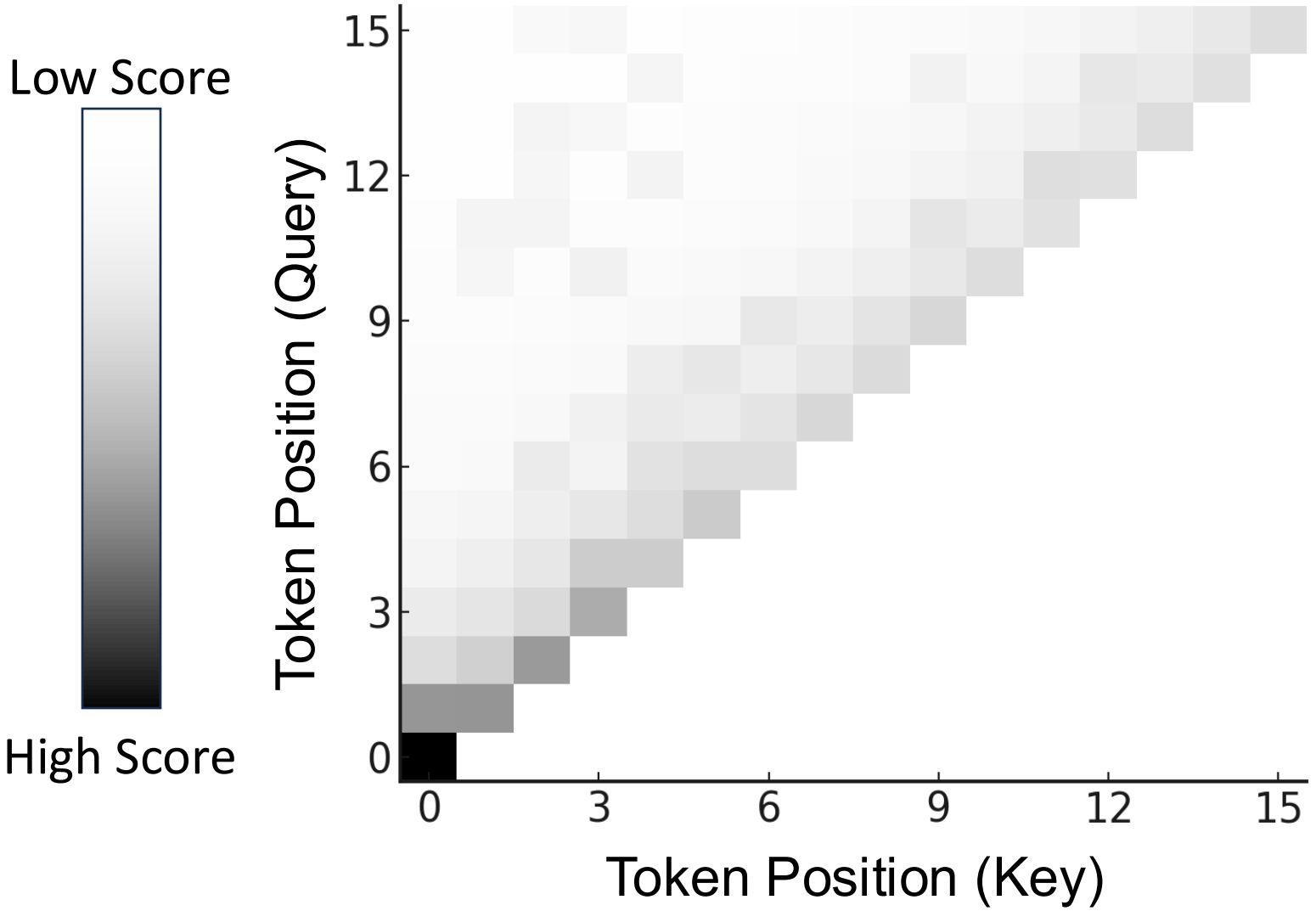
- **D**ense **R**etrieval **Acc**elerator (DReX)
  - Compute-enabled CXL-memory expander
- 512 GB Capacity
- Sign Concordance Filtering (**SCF**)
  - Filter vectors based on sign bits
- PIM Filtering
  - Run SCF inside DRAM
- Near Memory Ranking -> Top-k retrieval
  - Full-precision similarity score near DRAM

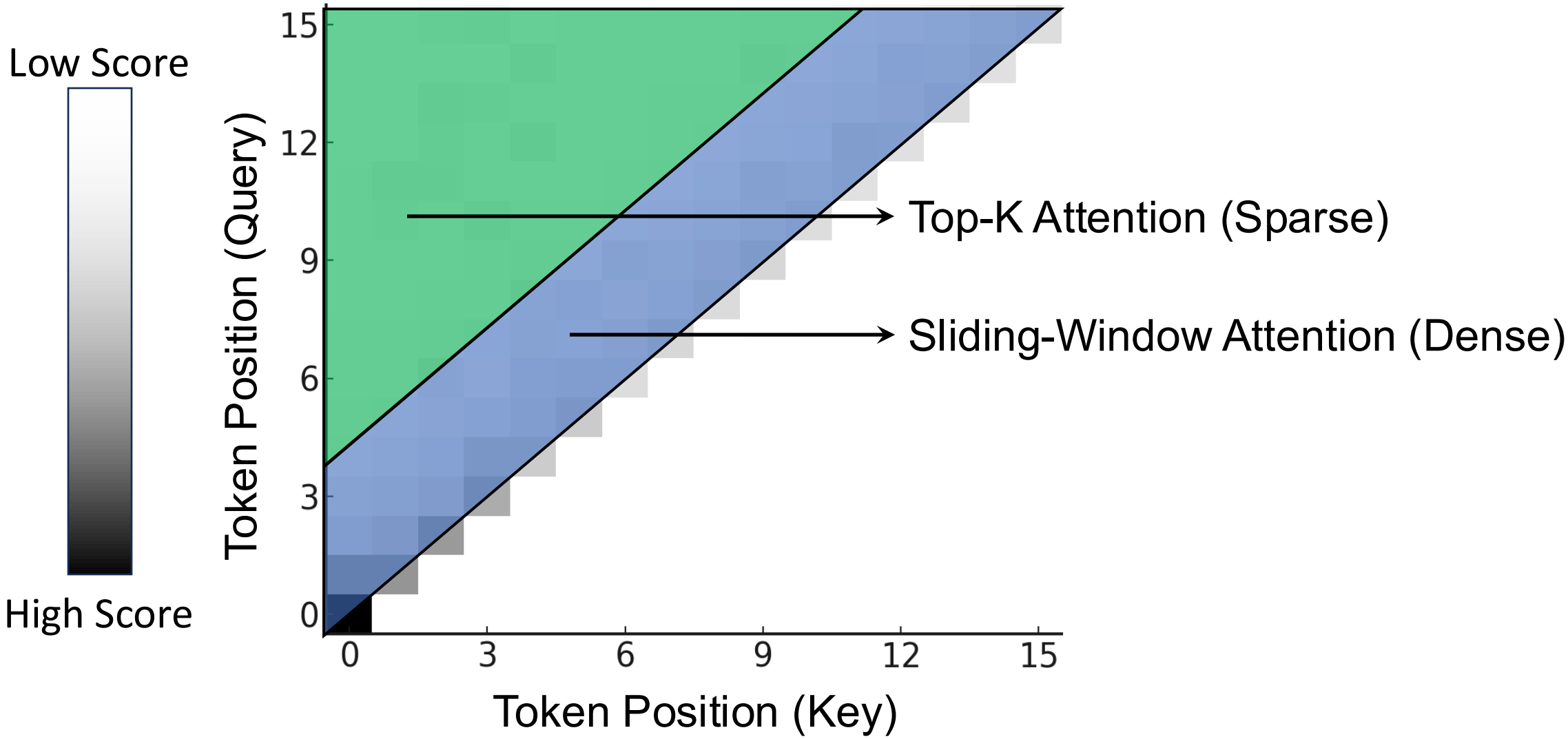


* ***D. Quinn***, *E. Yucel, M. Prammer, Z. Fan, K. Skadron, J. Patel, J. Martínez, and M. Alian, "DReX: Accurate and Scalable Dense Retrieval Acceleration,"* ISCA 2025

Reimagining Attention for DReX

1. Combine SCF with sliding-window attention: **balances load with GPU** and **improves SCF sparsity**


Low Score
High Score
Token Position (Query)
Token Position (Key)
0
3
6
9
12
15

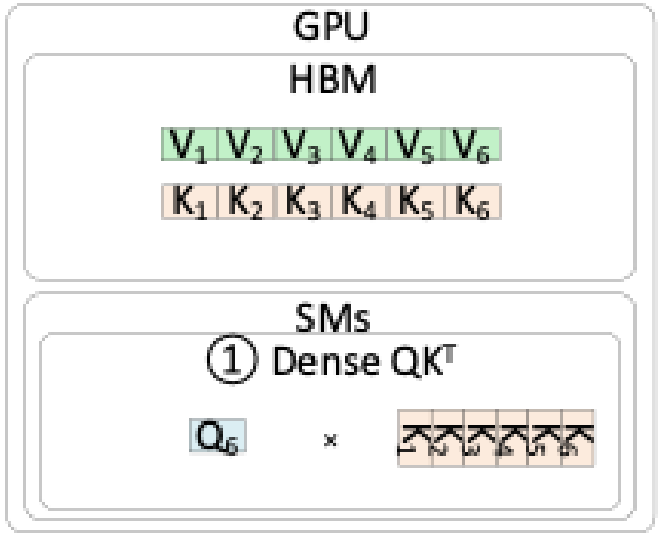

Low Score
High Score
Token Position (Query)
Token Position (Key)
0
3
6
9
12
15
Top-K Attention (Sparse)
Sliding-Window Attention (Dense)

Reimagining Attention for DReX

1. Combine SCF with sliding-window attention: **balances load with GPU** and **improves SCF sparsity**

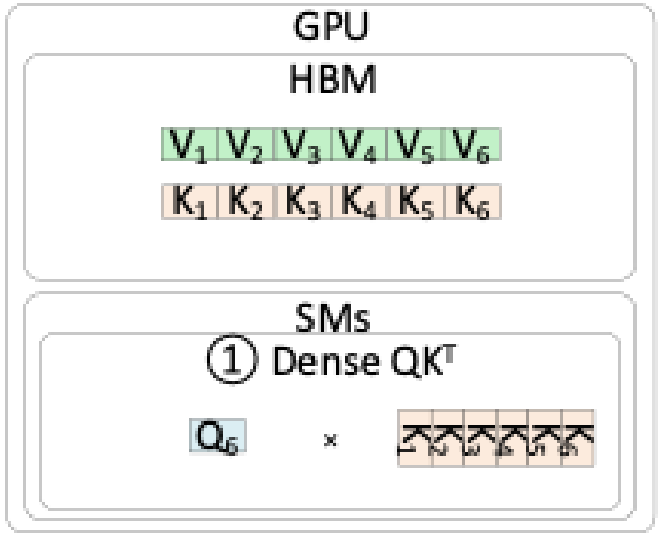
2. Optimize Queries and Keys via ITQ: **further improves SCF sparsity** by rebalancing vectors

Conventional GPU-Only Attention

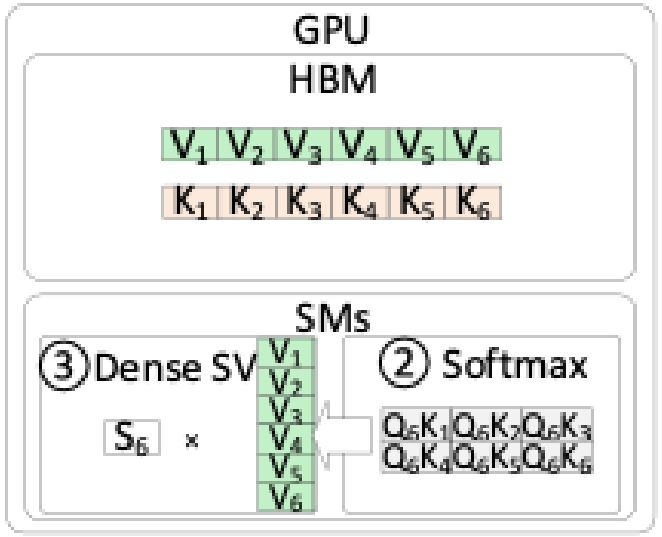


Step 1. Dense QK^T for all tokens

Conventional GPU-Only Attention

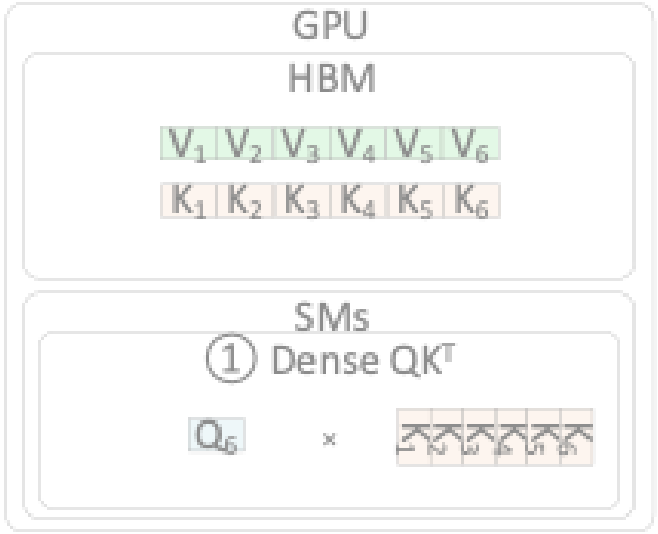


Step 1. Dense QK^T for all tokens

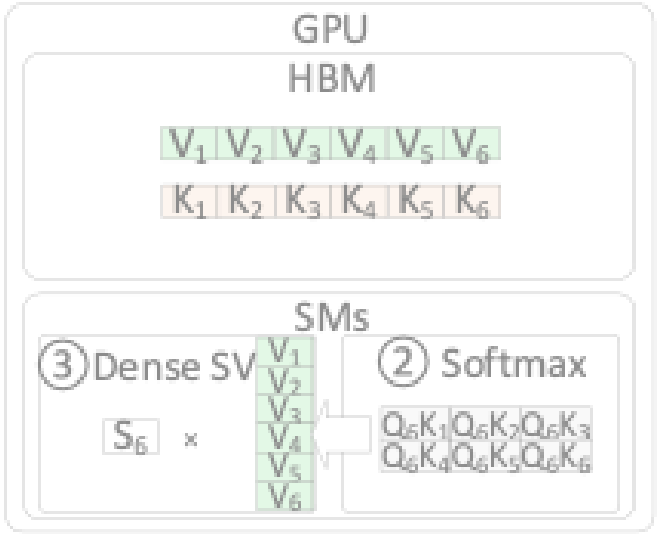


Step 2. Softmax & Dense SV

Conventional GPU-Only Attention

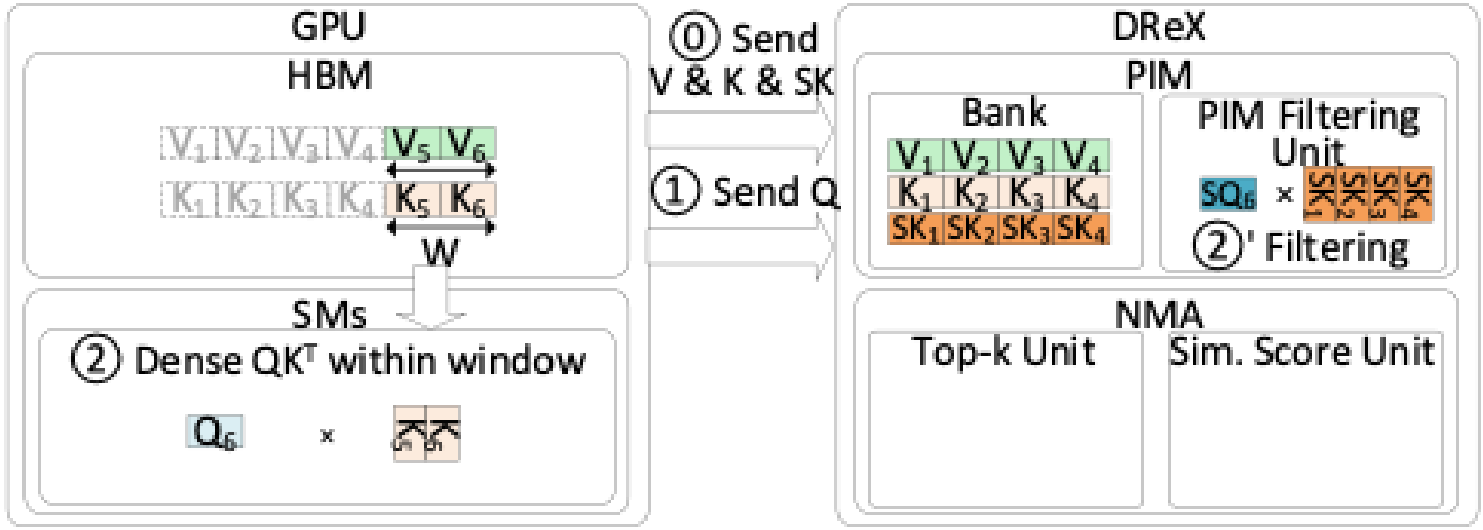


Step 1. Dense QK^T for all tokens



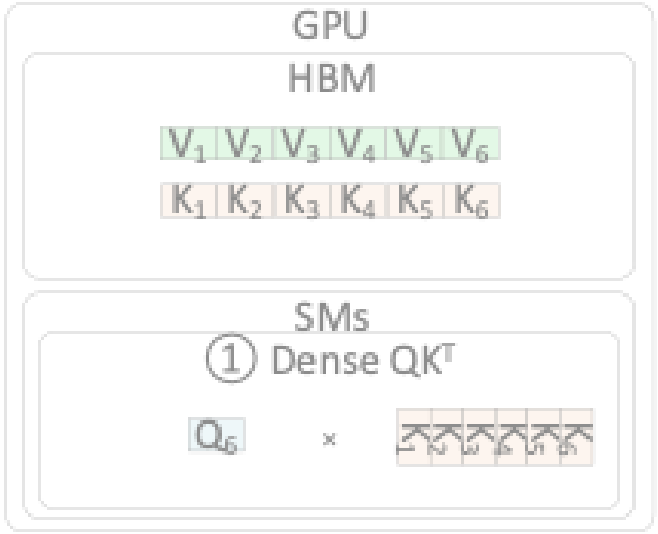
Step 2. Softmax & Dense SV

Our Approach (LongSight Attention)

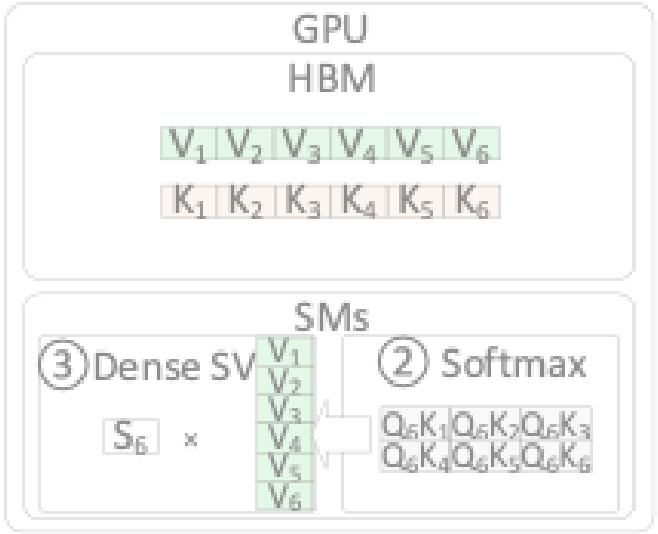


Step 1. GPU : Dense QK^T within Window / DReX : Filtering

Conventional GPU-Only Attention

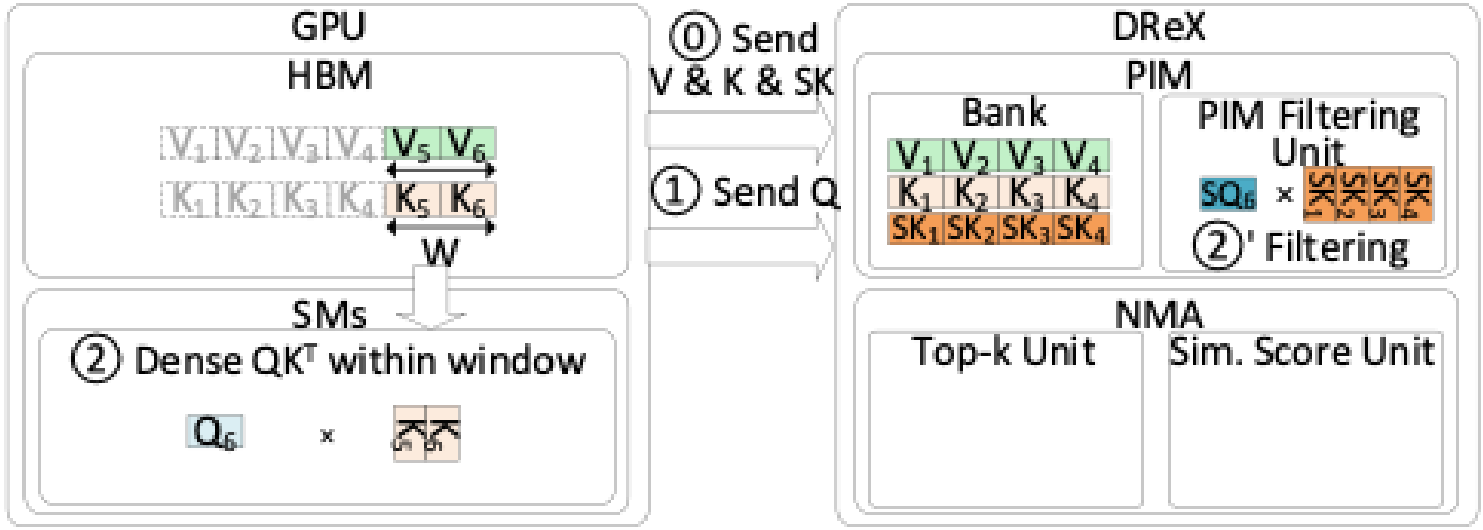


Step 1. Dense QK^T for all tokens

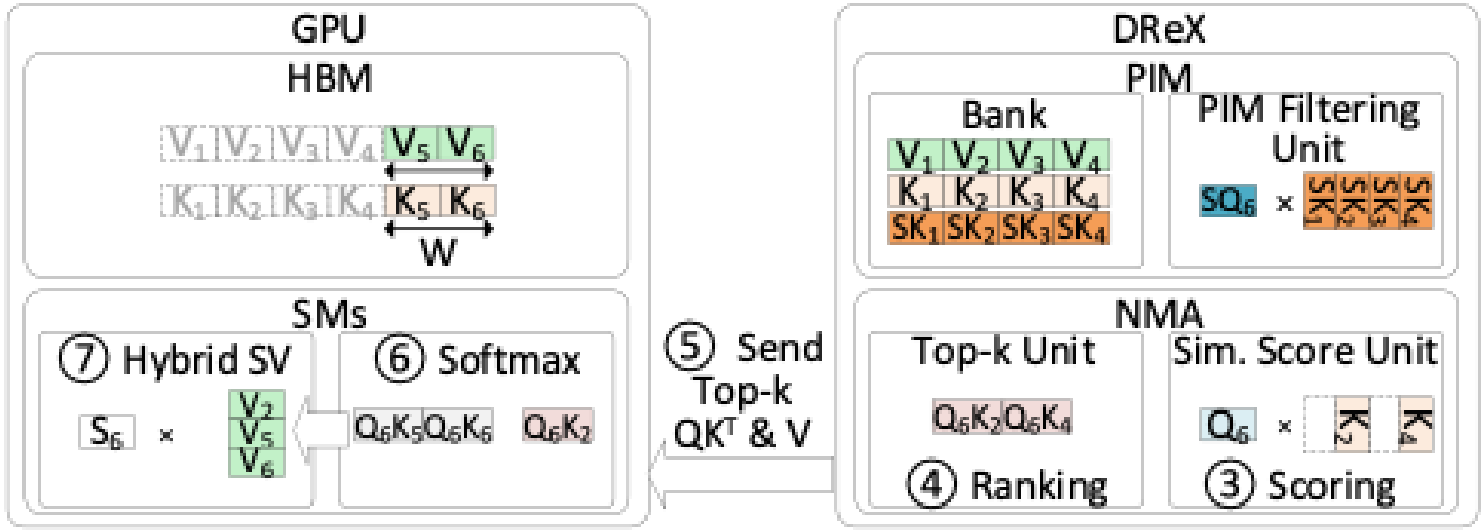


Step 2. Softmax & Dense SV

Our Approach (LongSight Attention)



Step 1. GPU : Dense QK^T within Window / DReX : Filtering



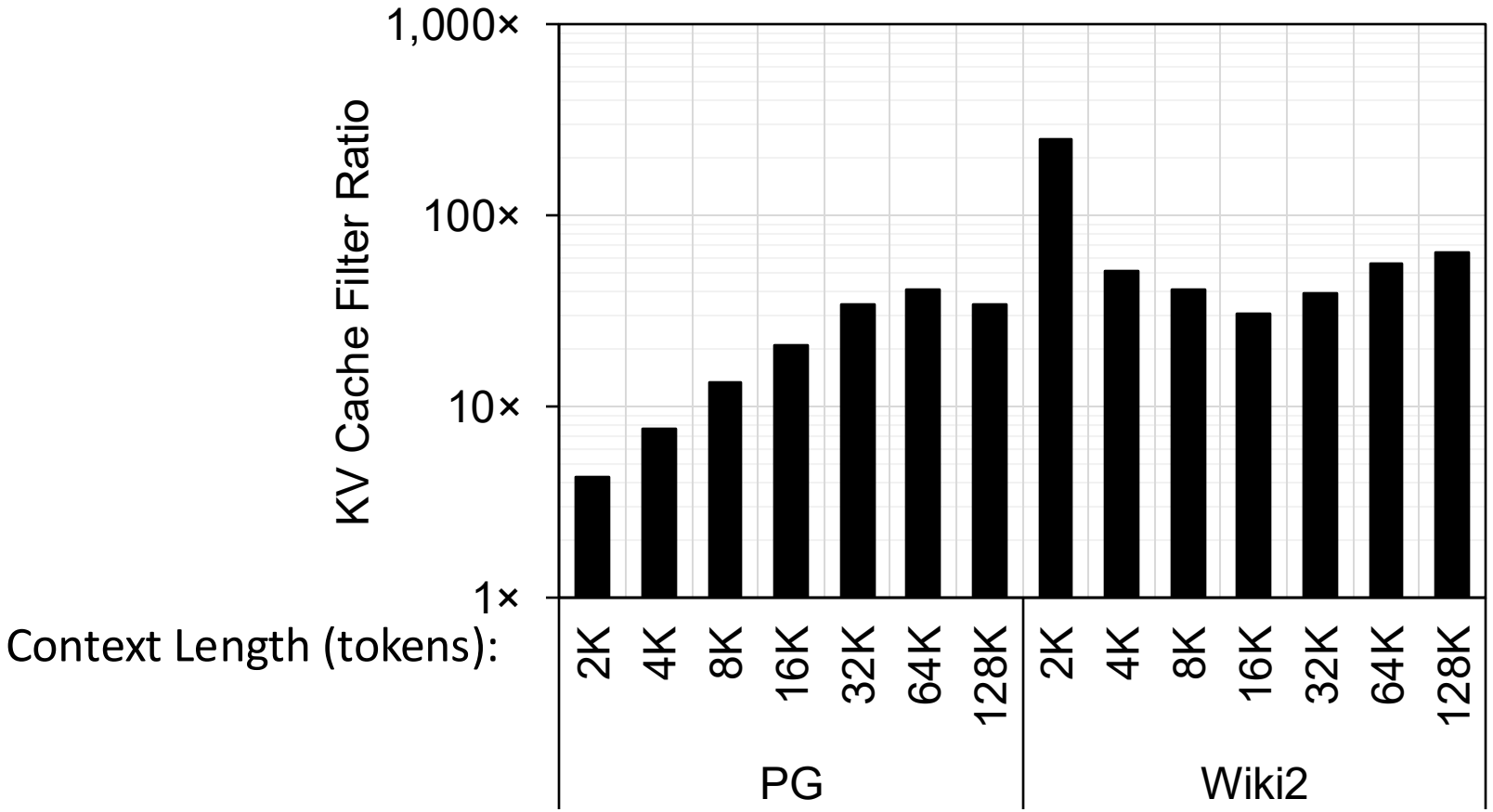
Step 2. GPU : Softmax & Hybrid SV / DReX : Scoring & Ranking

Methodology

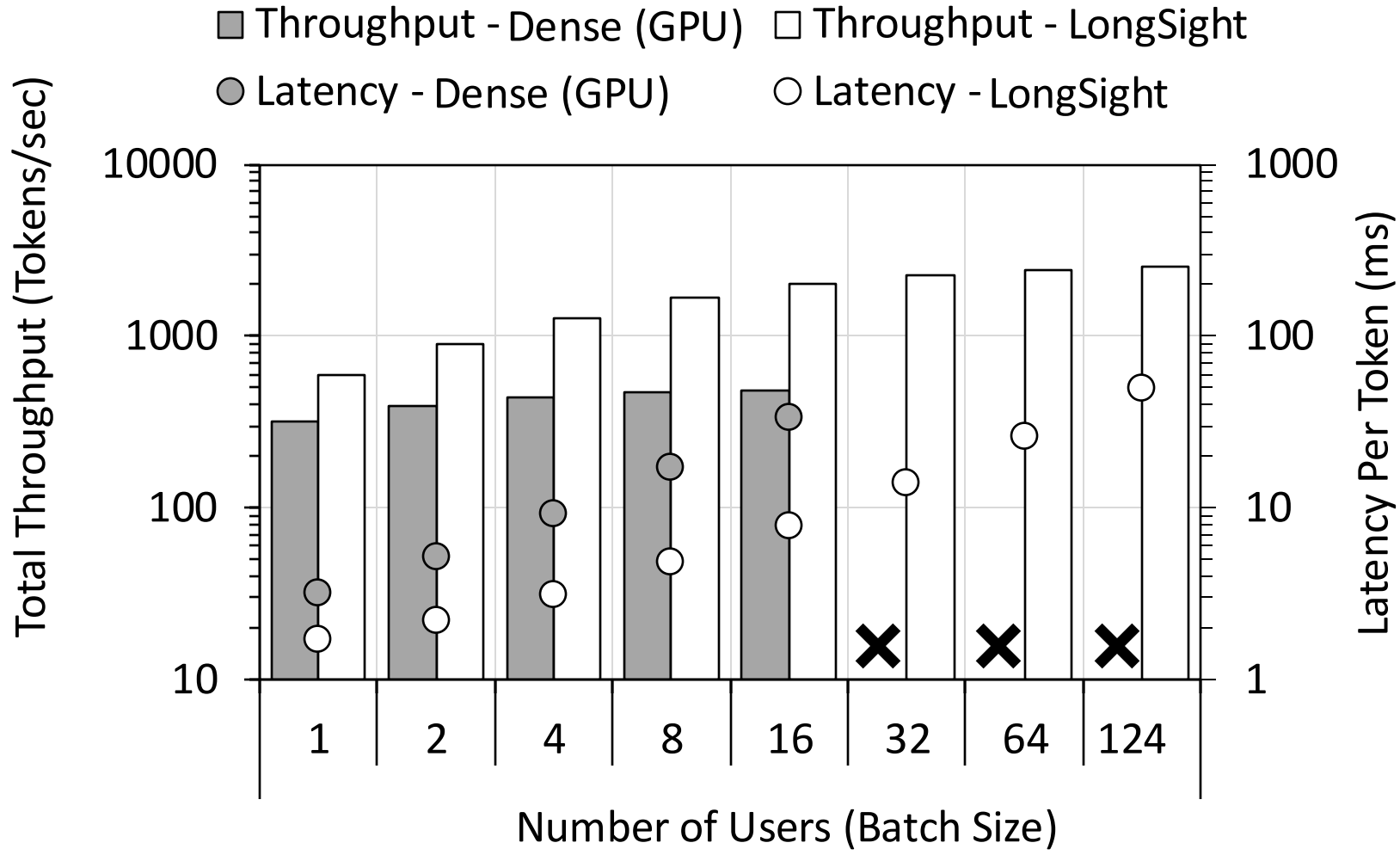
- Models: Llama-3-1B
- Datasets: Project Gutenberg (PG) and Wikipedia (Wiki2)
- Context lengths: Up to 128K (evaluated), 1M (projected)
- We evaluate decode phase (no prefill)
- End-to-end timing simulator:
  - Non-attention components (FFN, etc.): GPU-calibrated
  - Dense Attention Baseline: GPU-calibrated
  - Sparse Attention: RTL-validated analytical model for DReX
  - CXL interface: Remote accesses in two-socket Xeon server

Algorithmic Results: Hybrid Attention

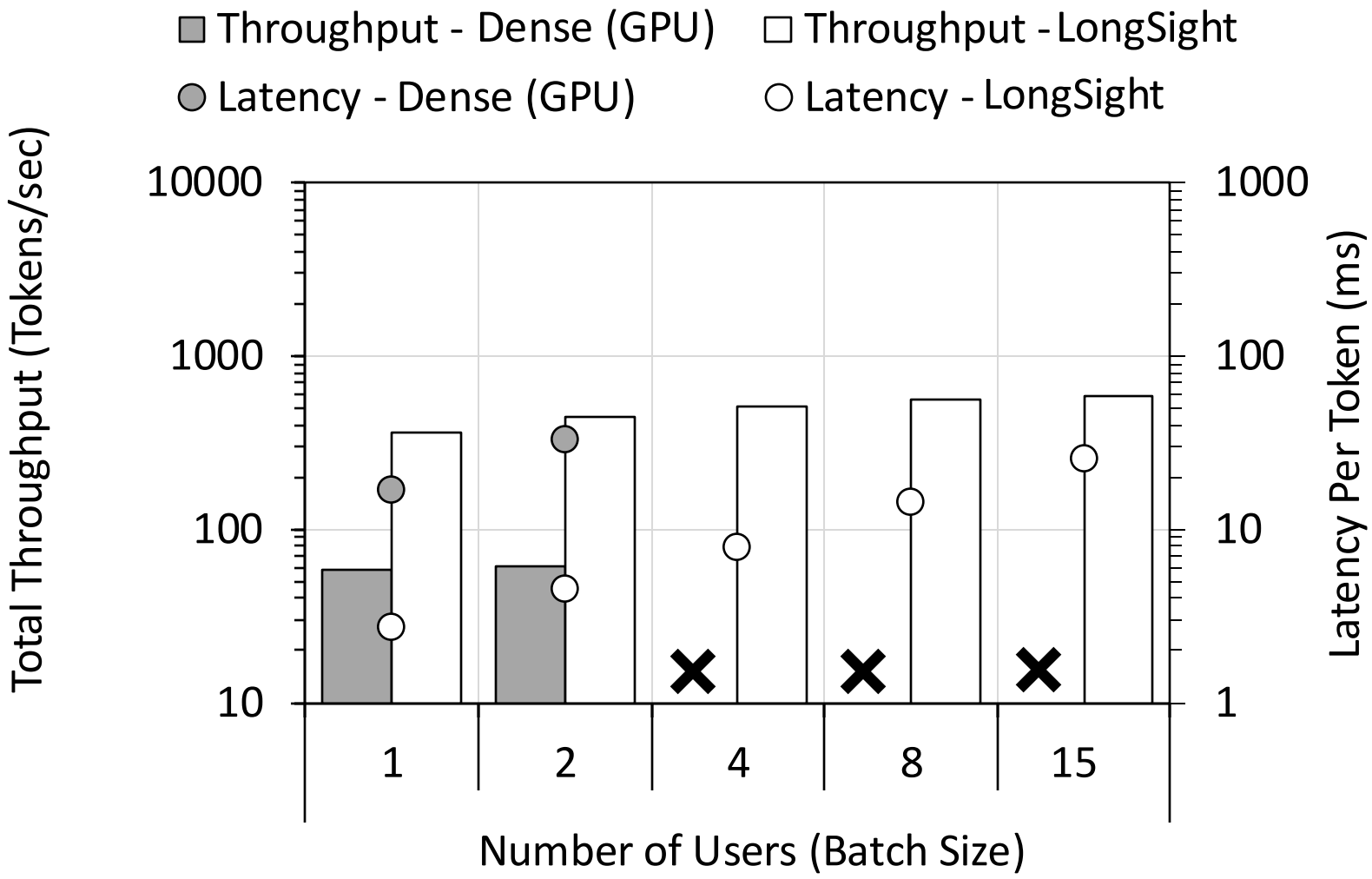
- Perplexity is within 5% of full dense attention.
- $K \leq 1024$



Results: Llama-3-1B – 128K Token Context



Results: Llama-3-1B – 1M Token Context



Contributions

1. LongSight leverages DReX to sparsify long-context attention

2. LongSight's hybrid attention balances load between DReX and GPU and improves robustness

3. LongSight efficiently disaggregates long-context attention, expanding system capacity and curtailing underutilization