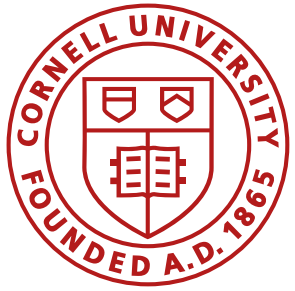


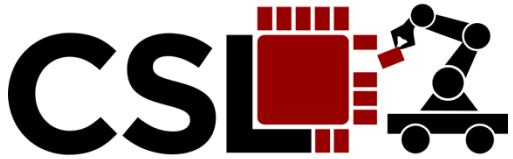
DReX: Accurate & Scalable Dense Retrieval Acceleration via Algorithmic-Hardware Codesign

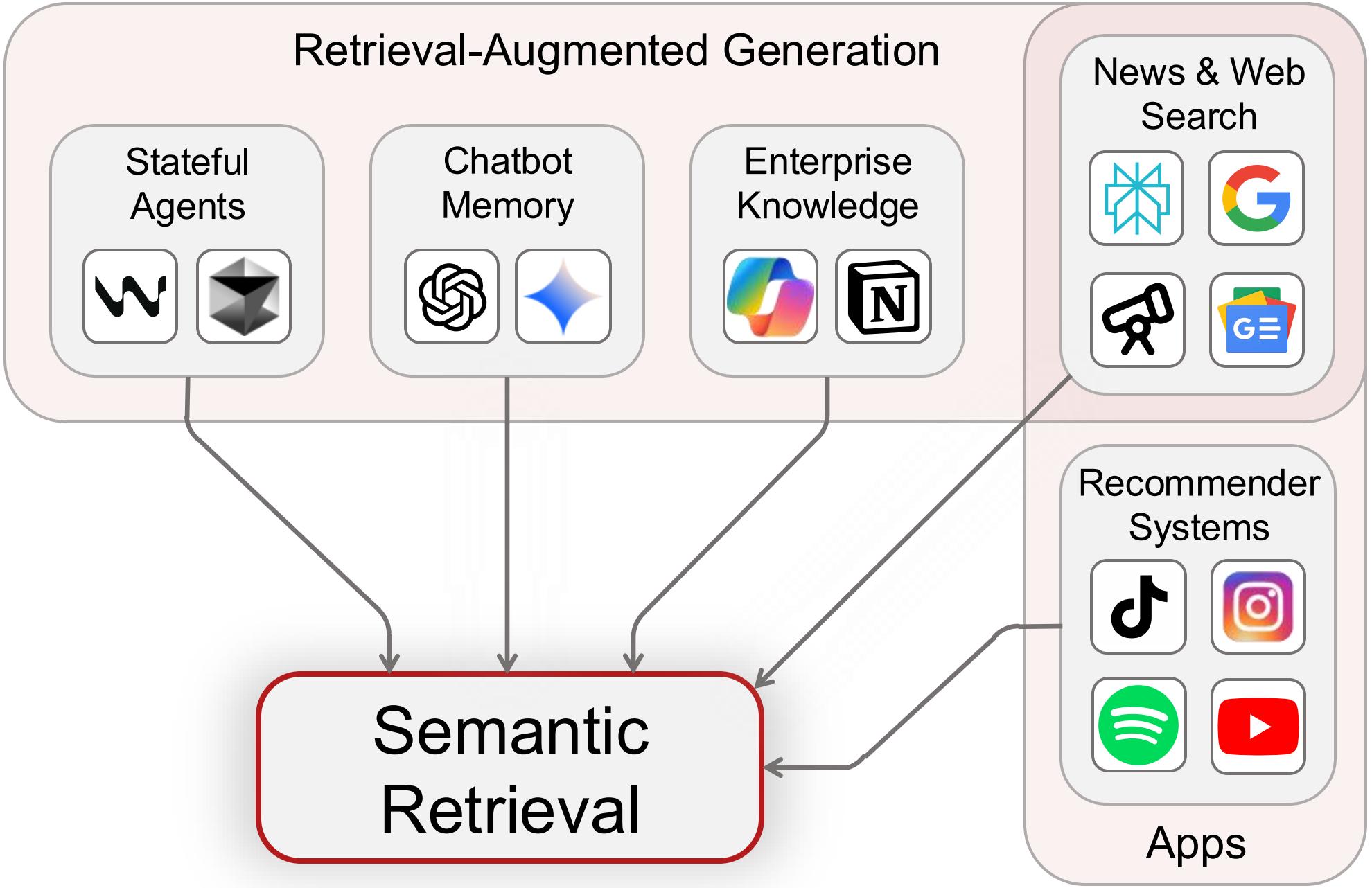
Derrick Quinn, E. Ezgi Yücel, Martin Prammer[1], Zhenxing Fan[2], Kevin Skadron[2], Jignesh M. Patel[1], José F. Martínez, Mohammad Alian

[1]: Carnegie Mellon University

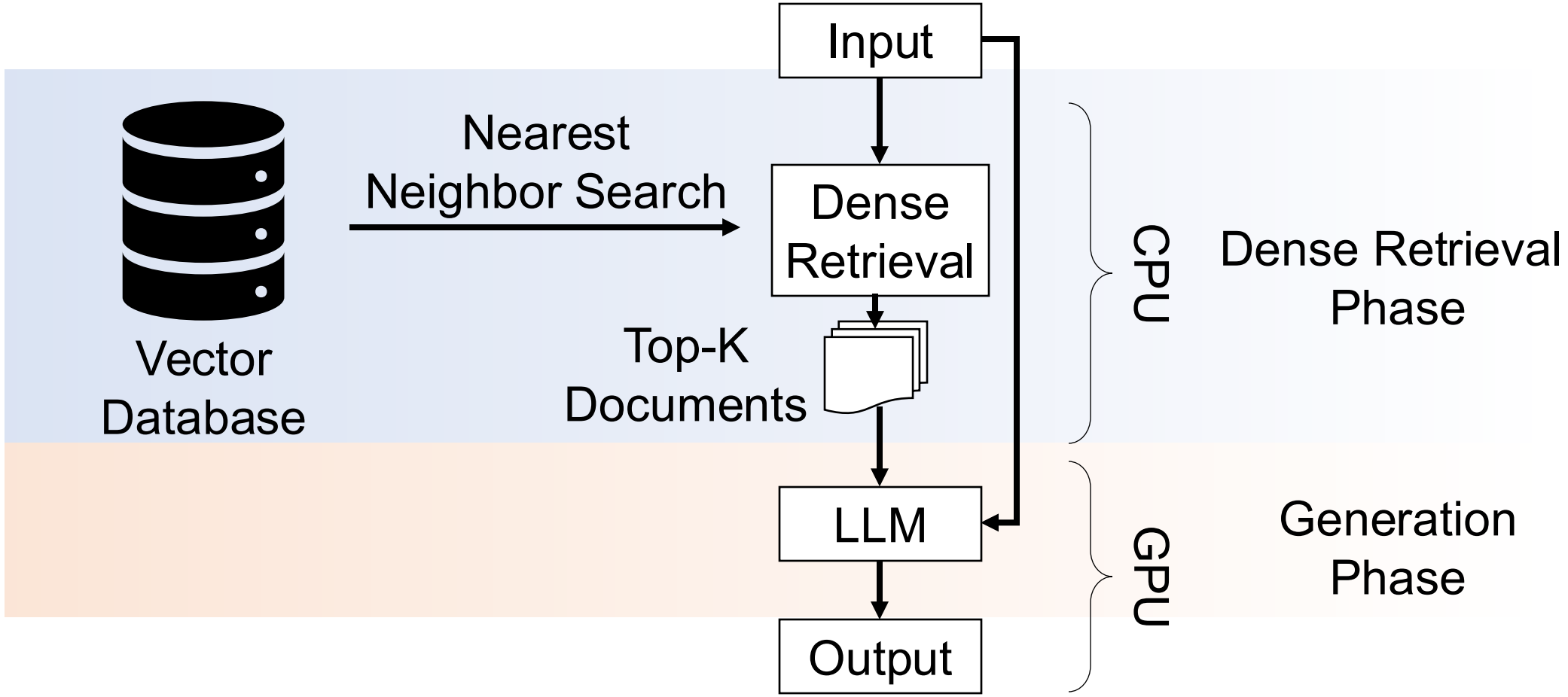
[2]: University of Virginia

CornellEngineering
Electrical and Computer Engineering

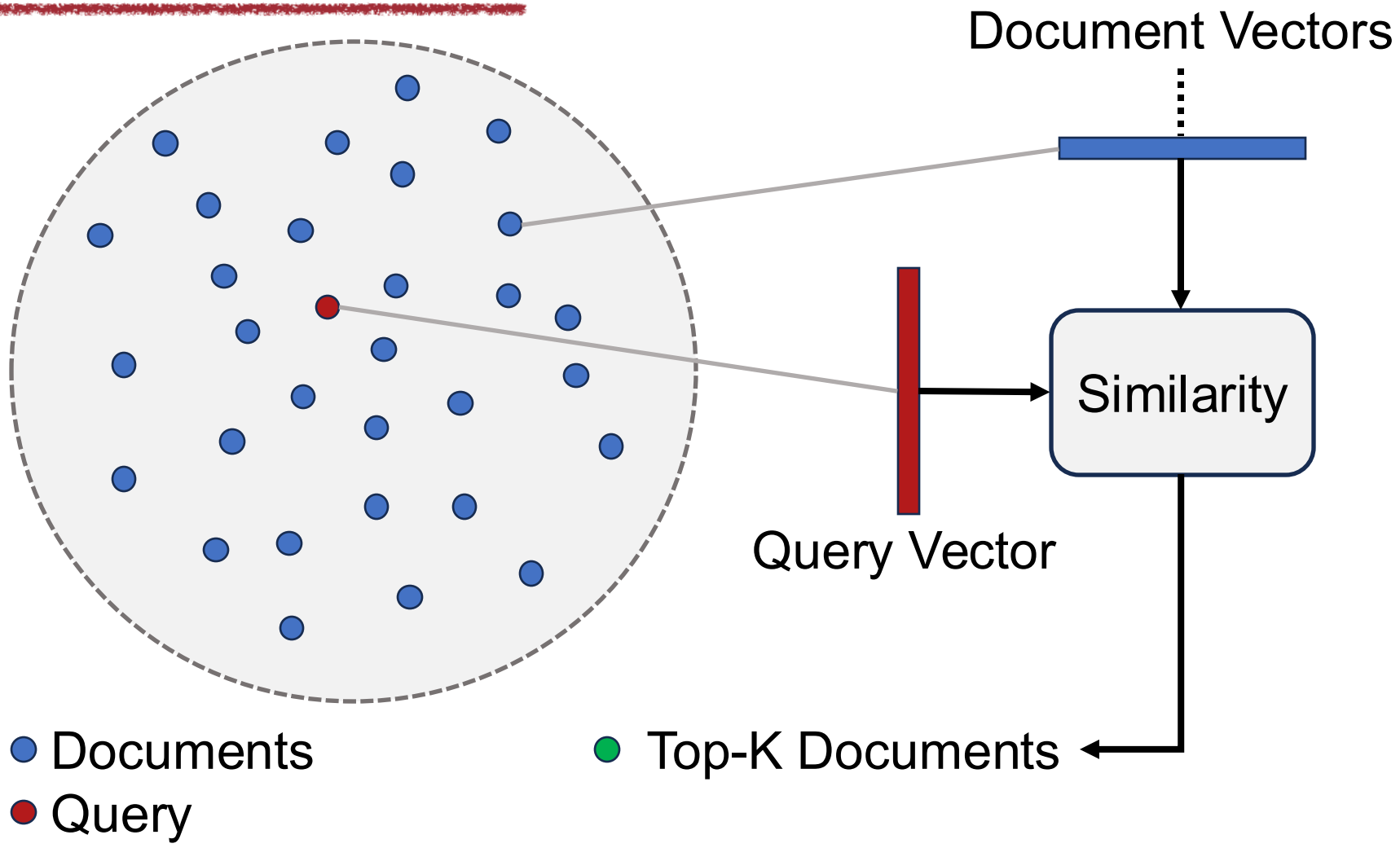



Retrieval-Augmented Generation
Stateful Agents
Chatbot Memory
Enterprise Knowledge
News & Web Search
Recommender Systems
Apps
Semantic Retrieval

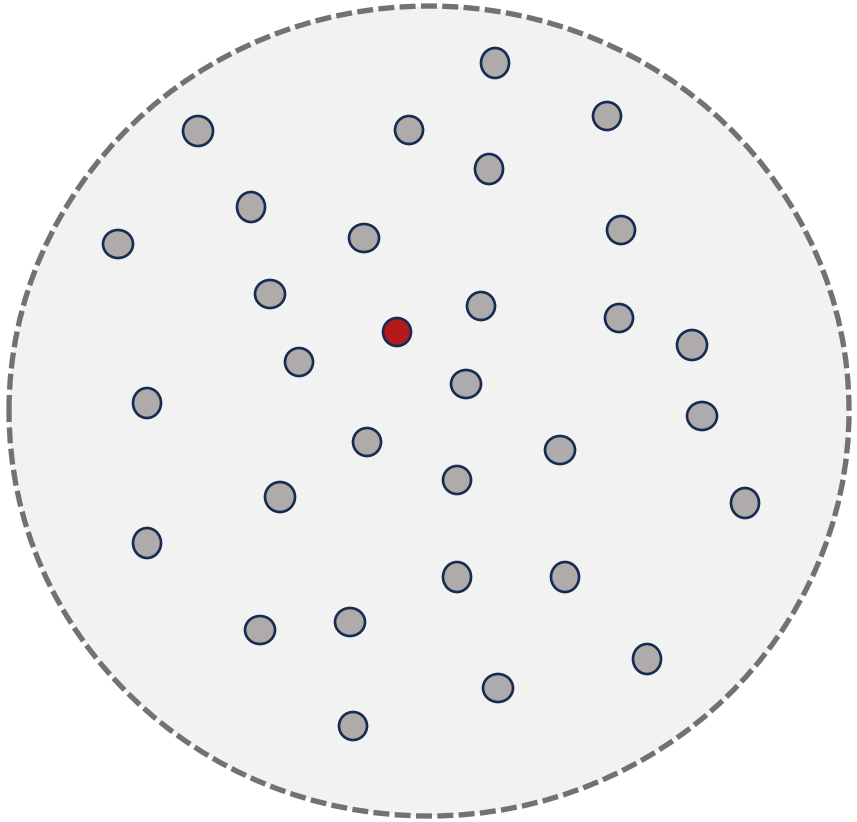
Retrieval-Augmented Generative AI



Dense Retrieval

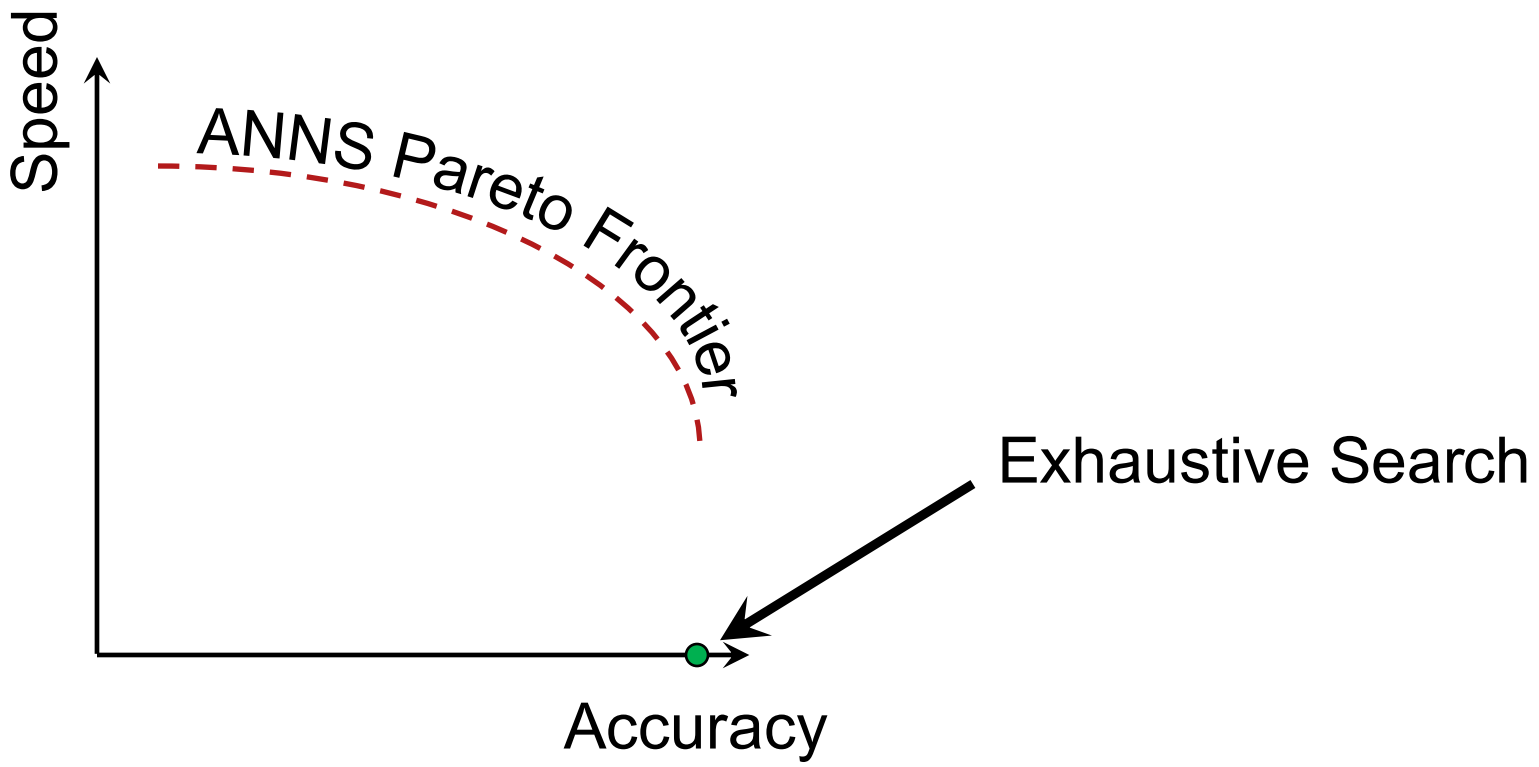


Exhaustive Search

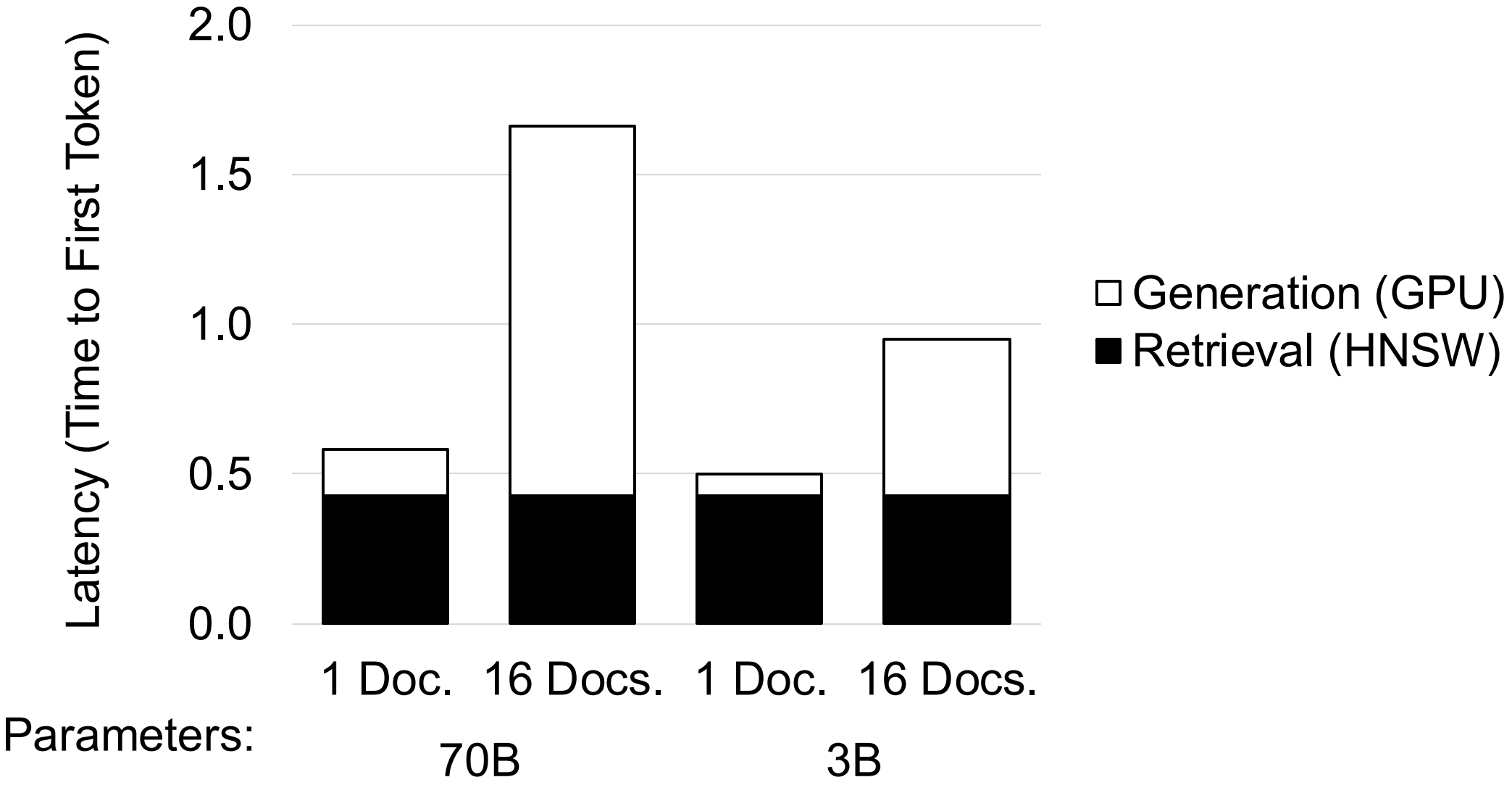


- Documents
- Query
- Top-K Documents

Approximate Nearest-Neighbor Search (ANNS)



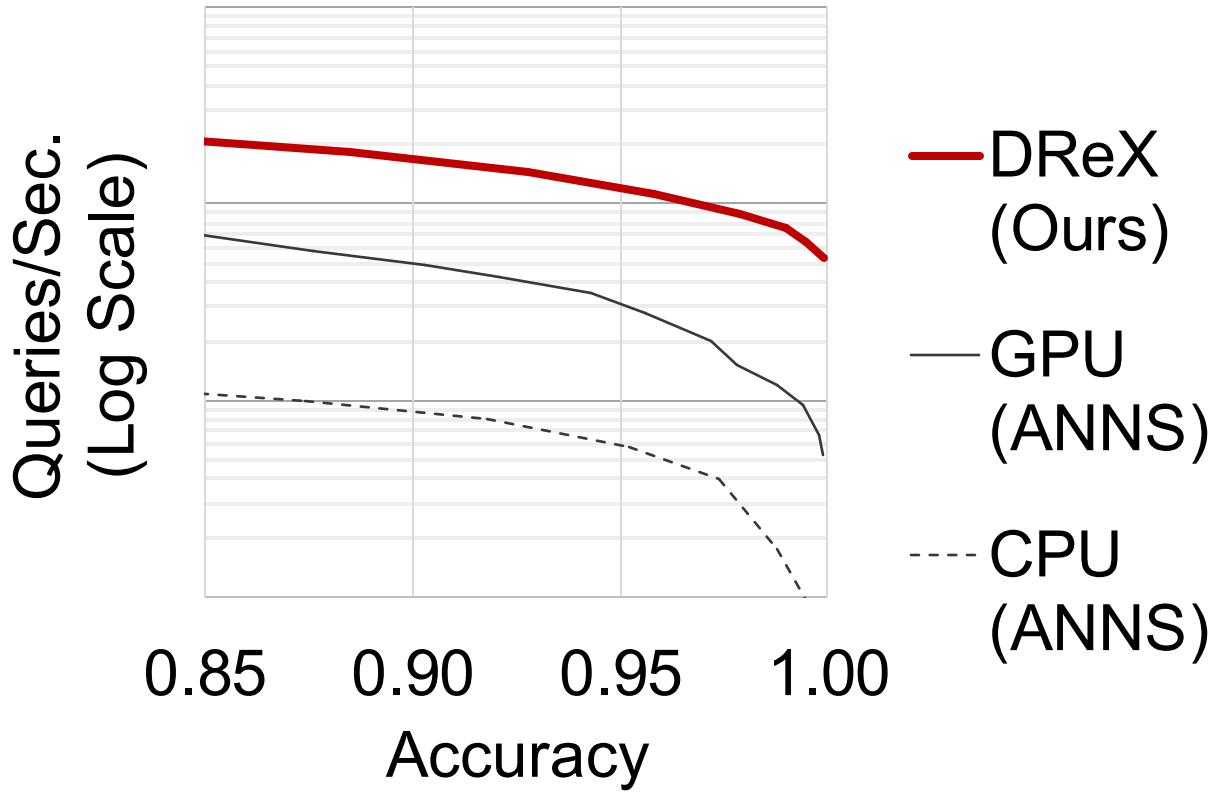
What's Wrong With Today's ANNS?



Our Contribution: Algorithmic-Hardware Co-Design for Dense Retrieval

Sign-Concordance Filtering: ANNS algorithm *Designed for Near-Data Processing*

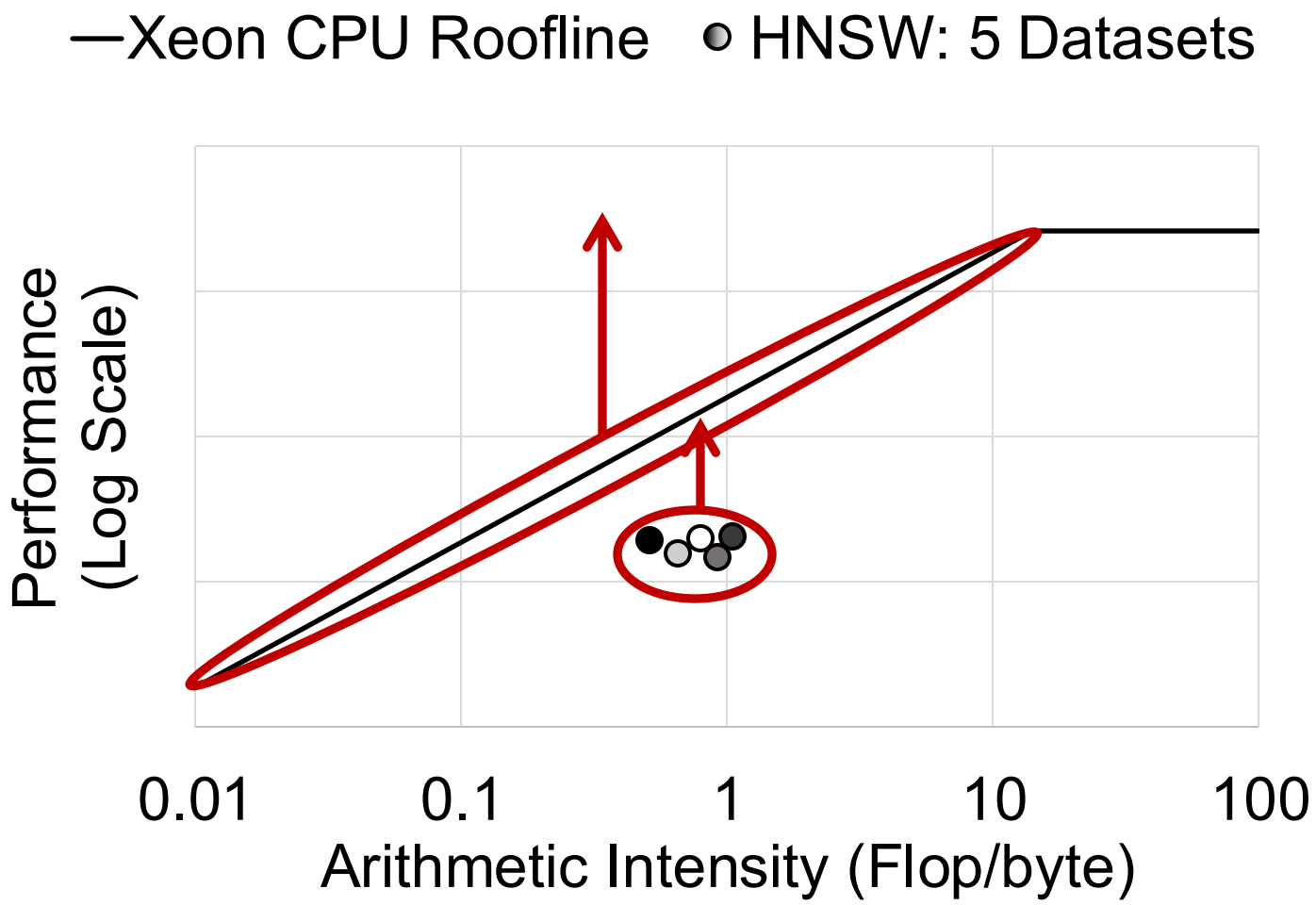
DReX: In- & Near-Memory Dense Retrieval Accelerator



Contents

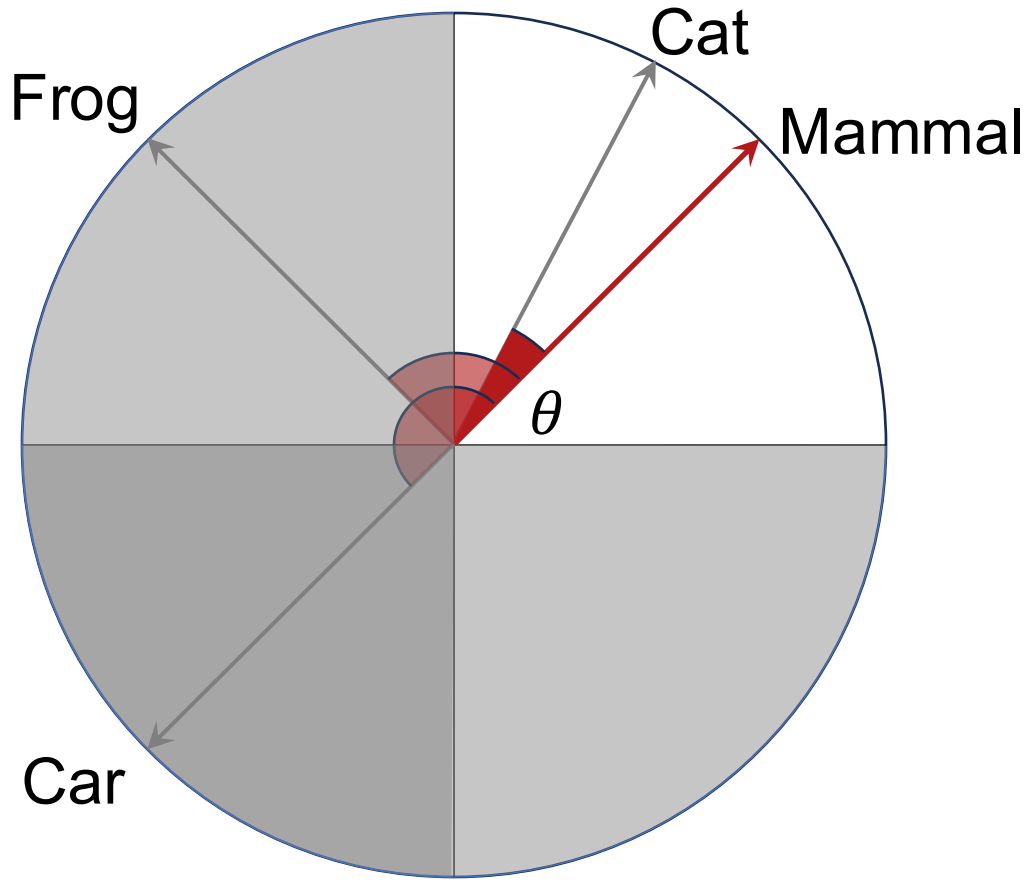
- Background
- Sign Concordance Filtering
- Architecture
- Methodology
- Evaluation

Bottleneck in Current ANNS?



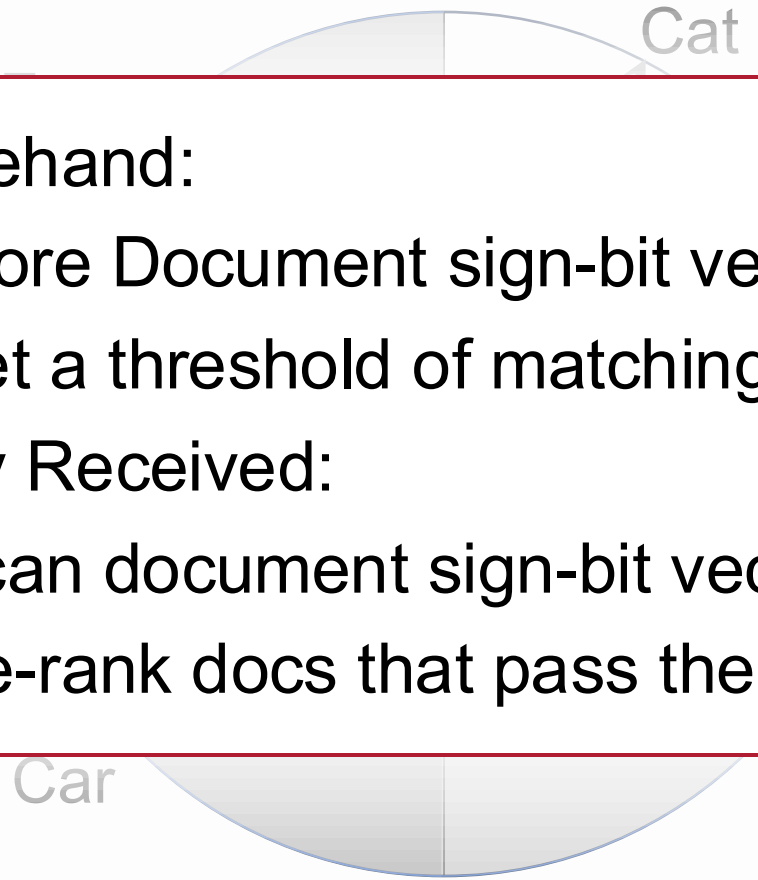
Sign Concordance Filtering

Background: Cosine Similarity



$$\cos(\theta) = D \cdot Q$$

Sign Concordance Filtering



Beforehand:

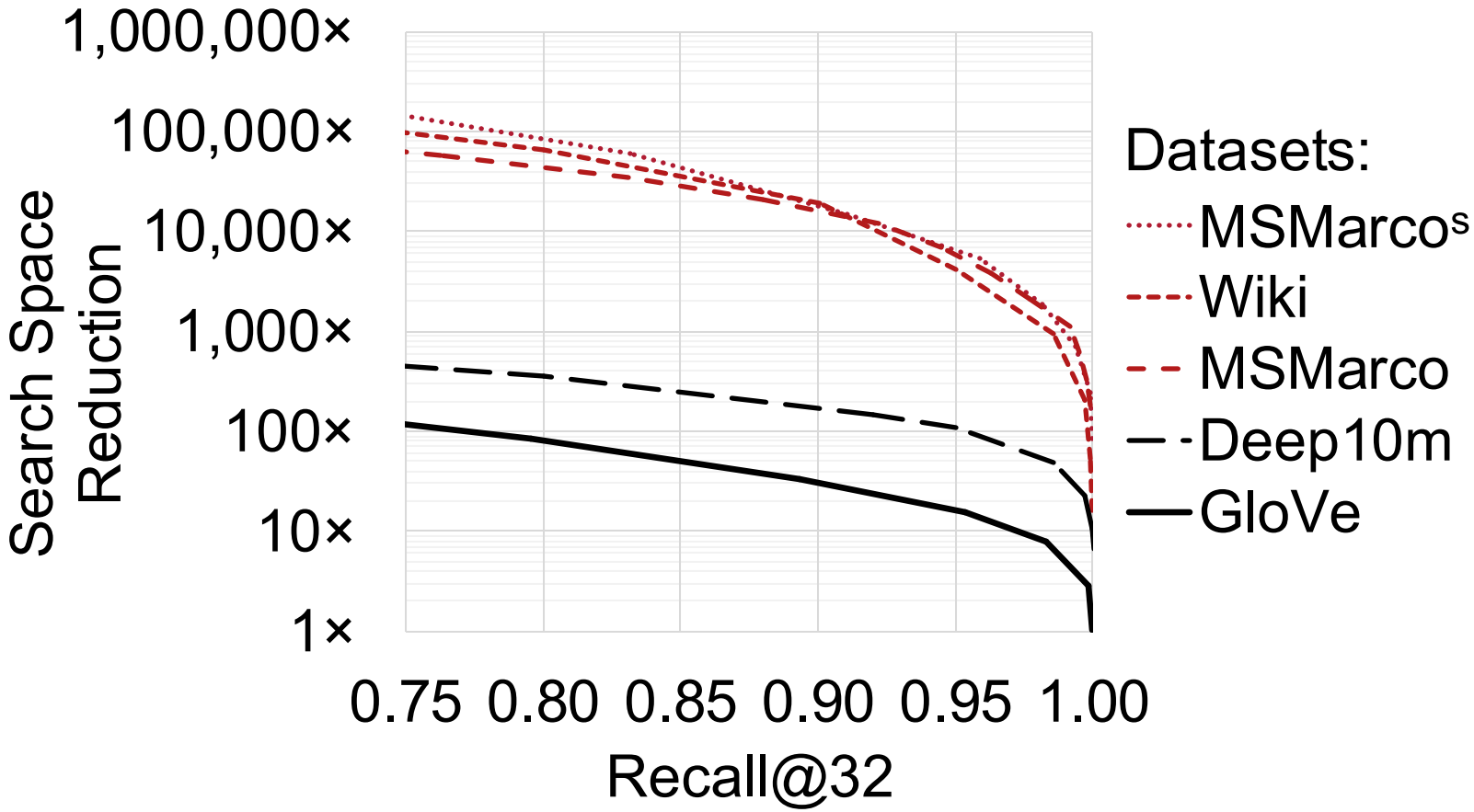
1. Store Document sign-bit vectors
2. Set a threshold of matching signs

Query Received:

1. Scan document sign-bit vectors
2. Re-rank docs that pass the threshold

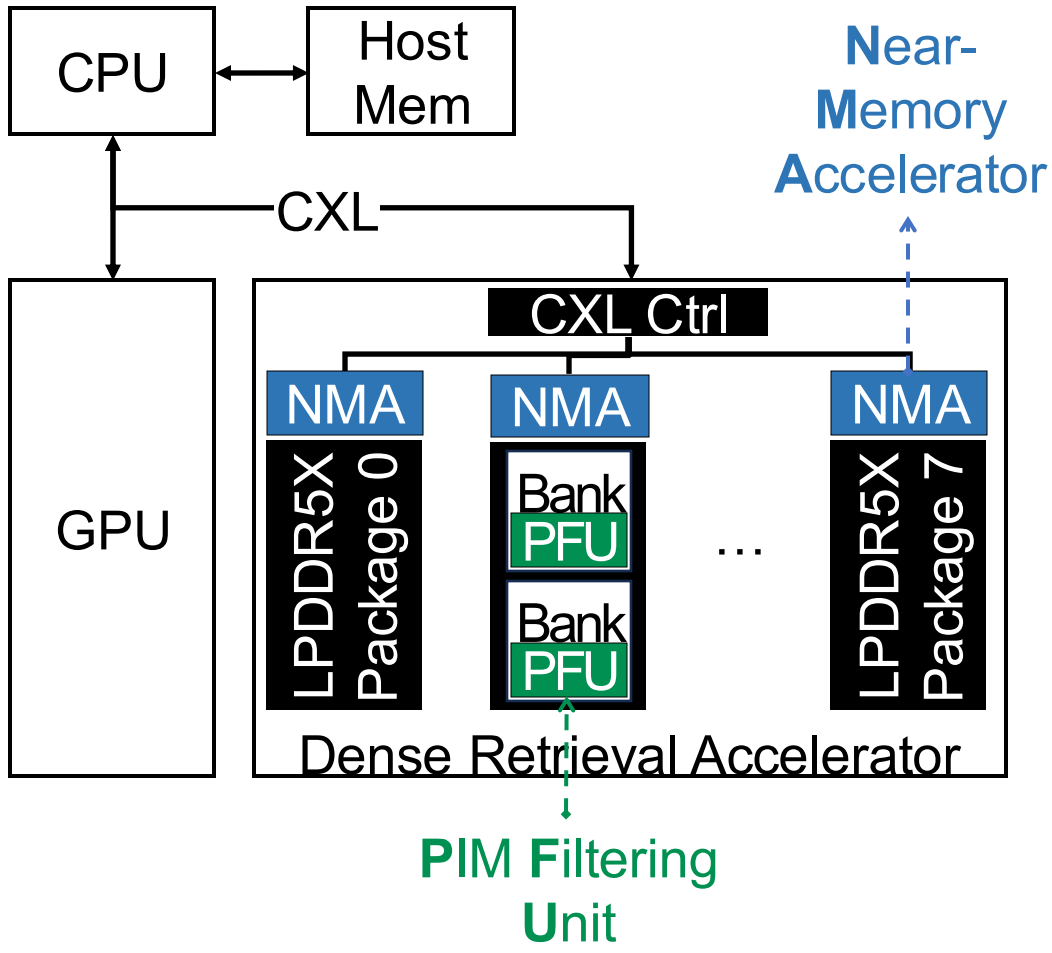
$= D \cdot Q$

Sign Concordance Filtering

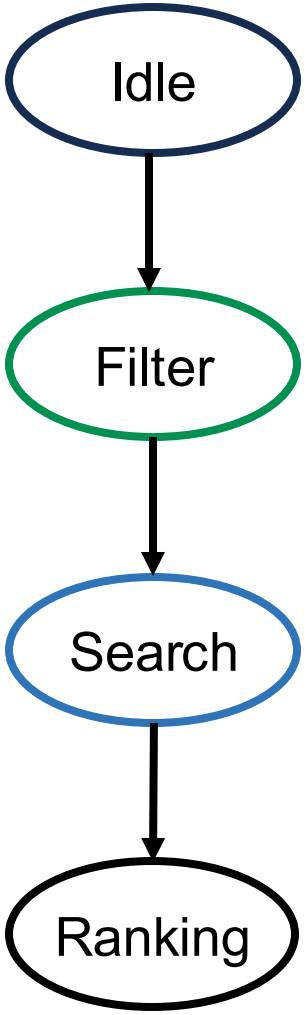


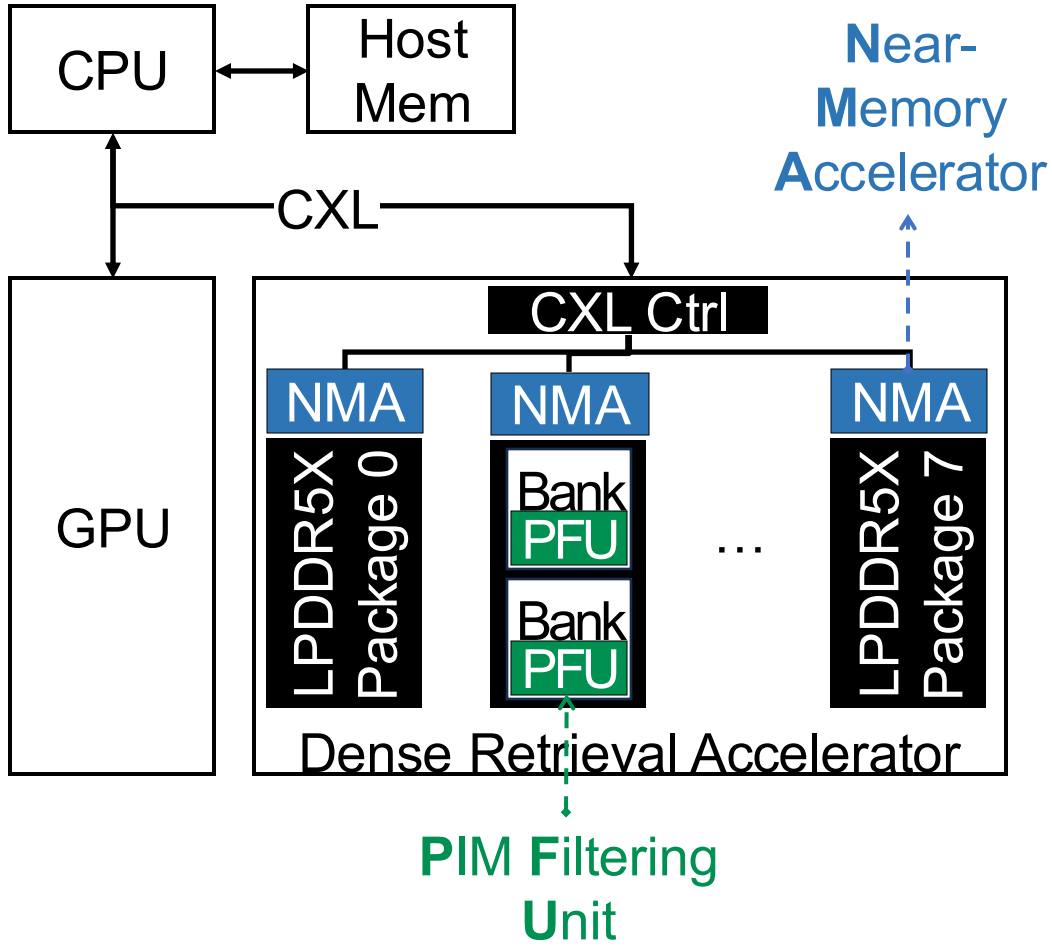
DReX Organization

- 512 GB LPDDR5X DRAM
- Byte addressable through CXL.mem
- Control plane through CXL.cache

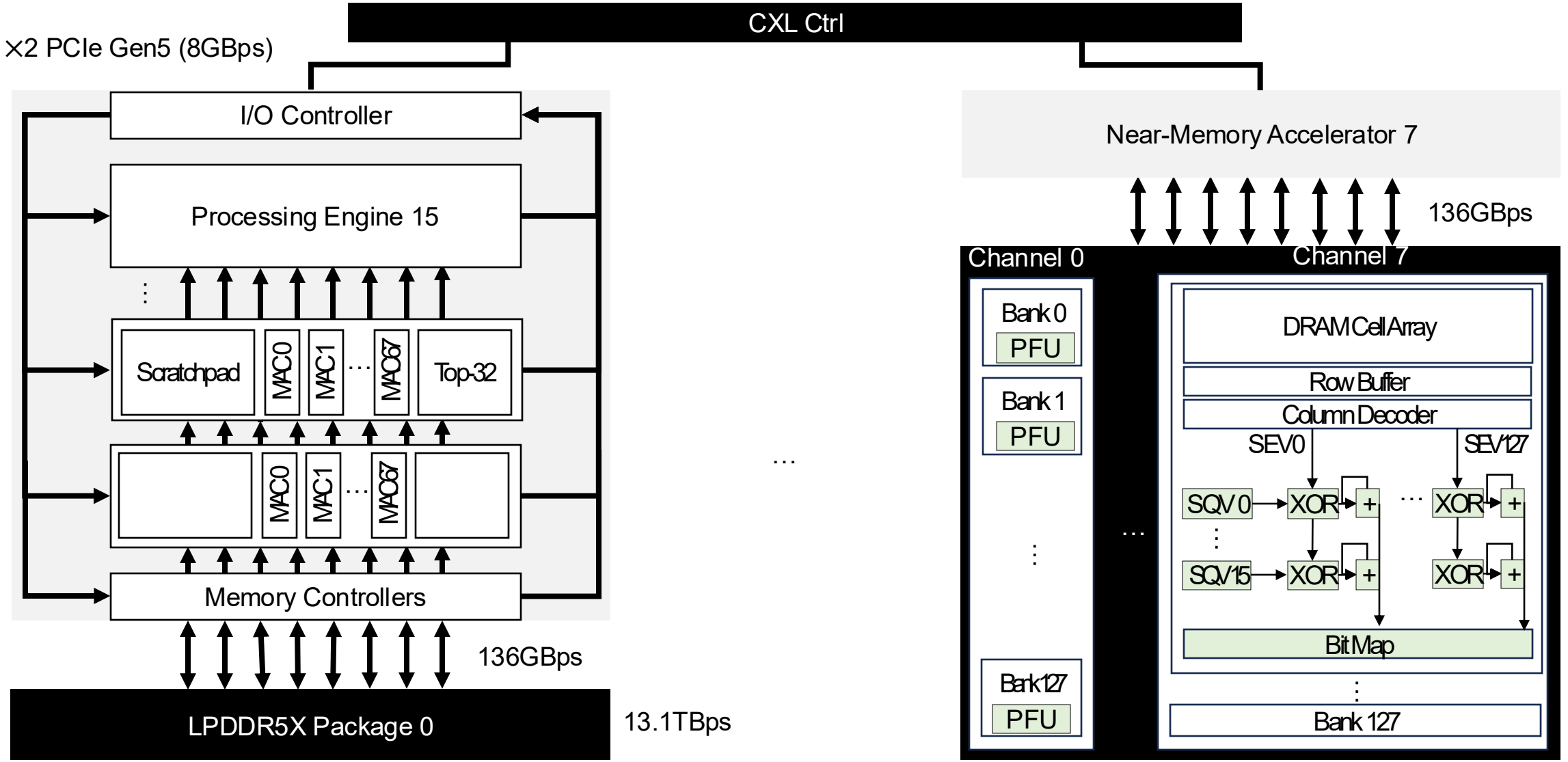


Algorithmic-Hardware Co-Design: DReX





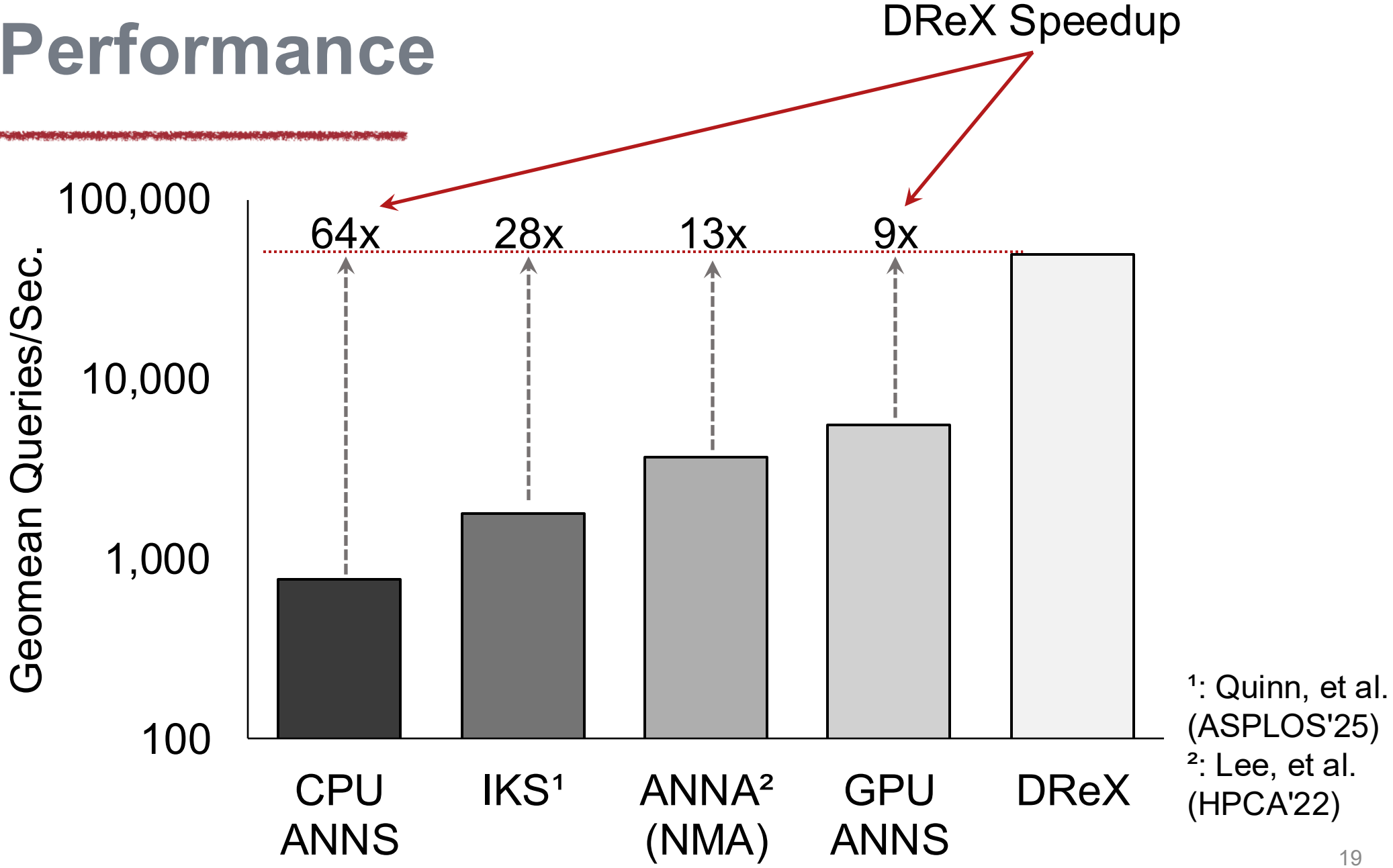
PFU & NMA Internal Structure



Experimental Methodology

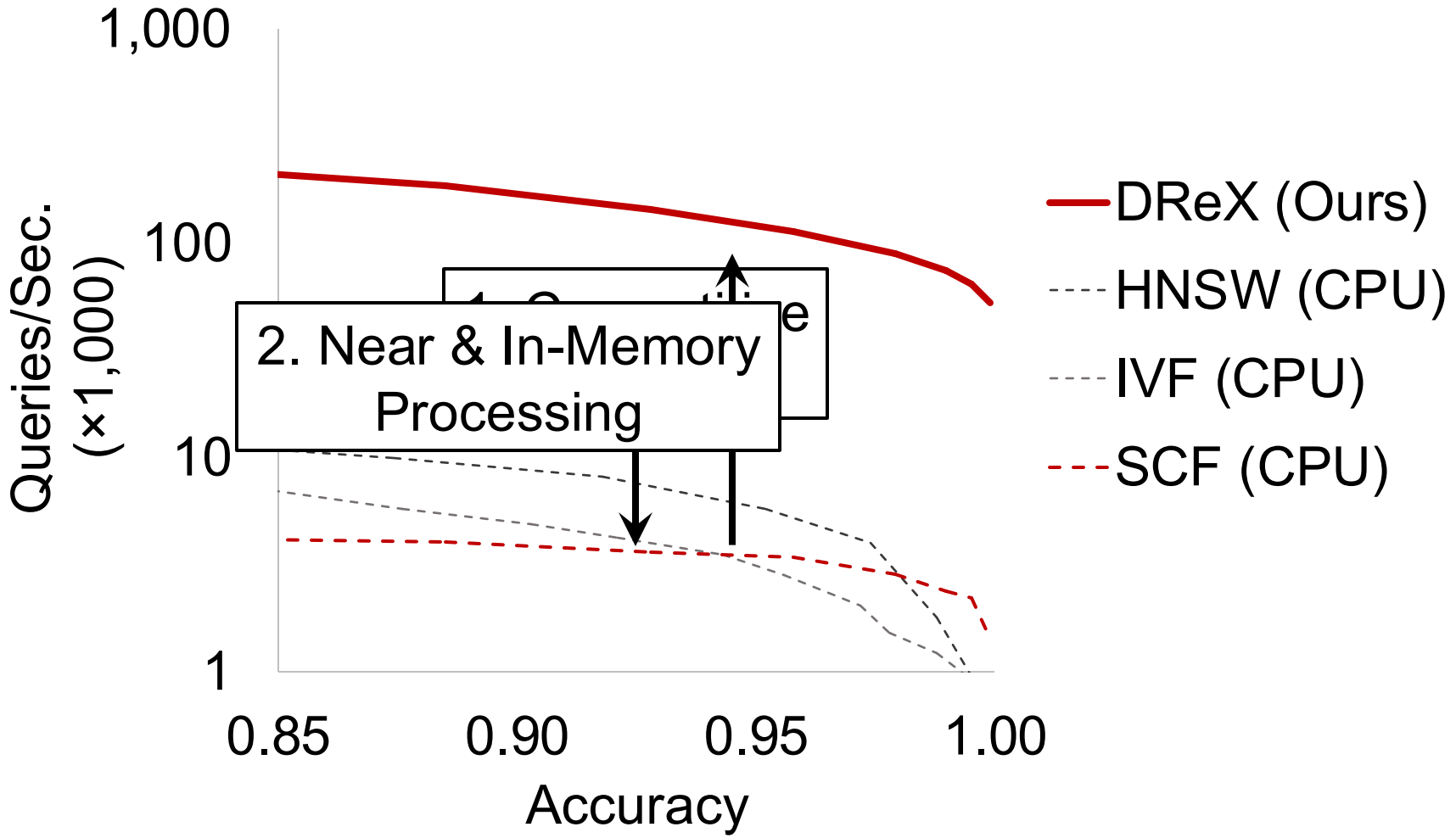
- Software: Meta FAISS (CPU) & NVIDIA CUVS (GPU)
- Experimental Setup:
  - CPU: Intel Xeon Max 9462
  - GPU: NVIDIA H100
- DReX Simulator:
  - DRAMSim3
  - RTL Synthesis
  - Timing Simulator
- 5 representative datasets

DReX Performance



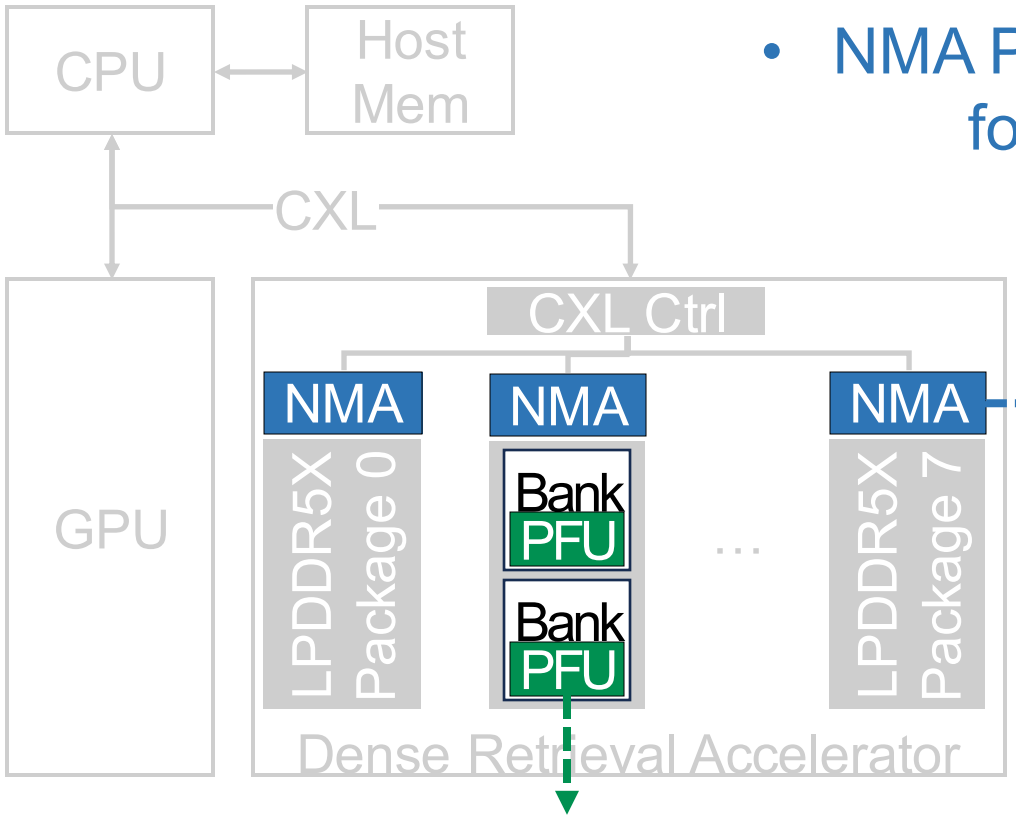
¹: Quinn, et al. (ASPLOS'25)
²: Lee, et al. (HPCA'22)

Why Does DReX Work?



Power and Area of PFU & NMA

- NMA Area: 15.07 mm^2
- NMA Peak Power: 1.07 W for batch size 16



- PFU Area: 0.1 mm^2 (6.7% area overhead)
- PFU Peak Power: 14 mW for batch size 16 (18.4W per package)

Conclusions

- **Contributions**
  - **Sign-Concordance Filtering**: ANNS algorithm *Designed for Near-Data Processing*
  - **DReX**: In- & Near-Memory Dense Retrieval Accelerator
- **Key results**
  - **64**× speedup at Recall@32 = 0.95 compared to HNSW on CPU
  - **7**× reduction in time-to-first-token for a representative RAG application