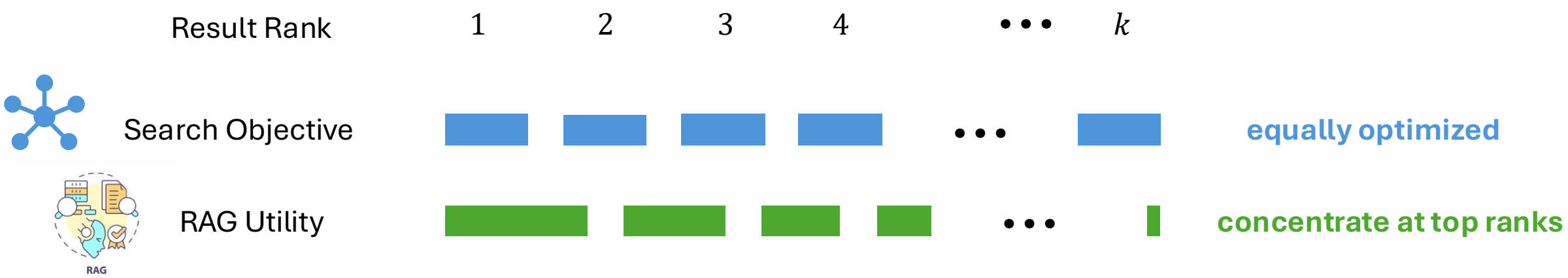
When Enough Is Enough: Rank-Aware Early Termination for Vector Search

Jianan Lu[1], Asaf Cidon[2], and Michael J. Freedman[1]

[1] PRINCETON UNIVERSITY

[2] COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Search Optimizes the Wrong Objective



Application utility is highly rank-sensitive, but current vector search systems are not.

We redesign vector search around rank-aware utility.

Vector Search: Empowering Everyday Intelligence

- A way of finding data based on their **semantic similarity** (e.g., Euclidean distance) in some high-dimensional **embedding space**, instead of using exact keyword matches.

- The AI + LLM Boom
  - Retrieval-augmented generation (RAG)
  - Semantic search and recommendation
  - Many more …

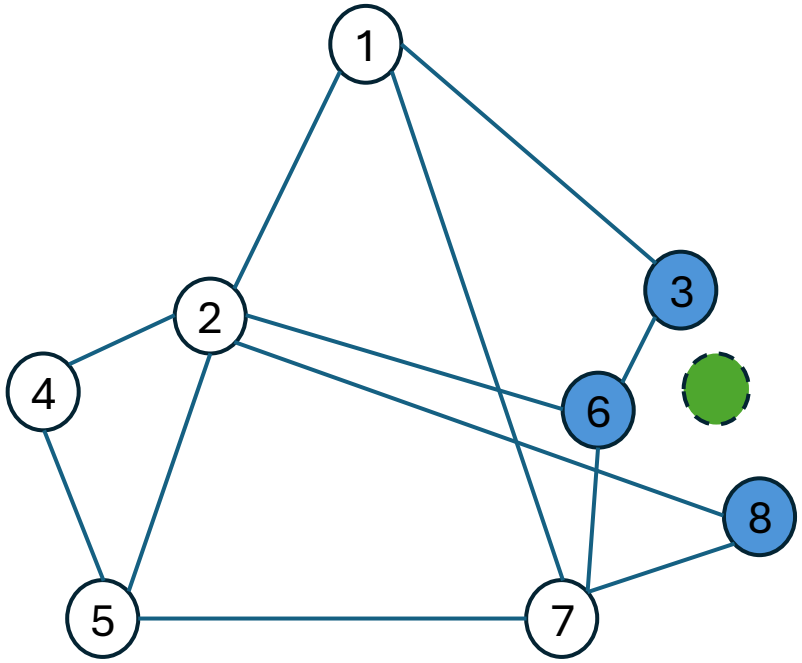
Recommend songs

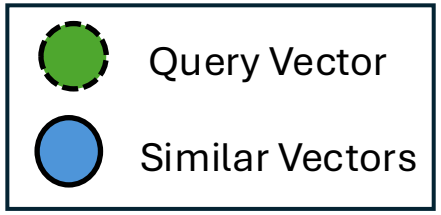
Show photos of family time

The user bought a sci-fi book. They may also enjoy …

Better Scalability with Graph ANN Indexes

- Approximate Nearest Neighbor (ANN) Search

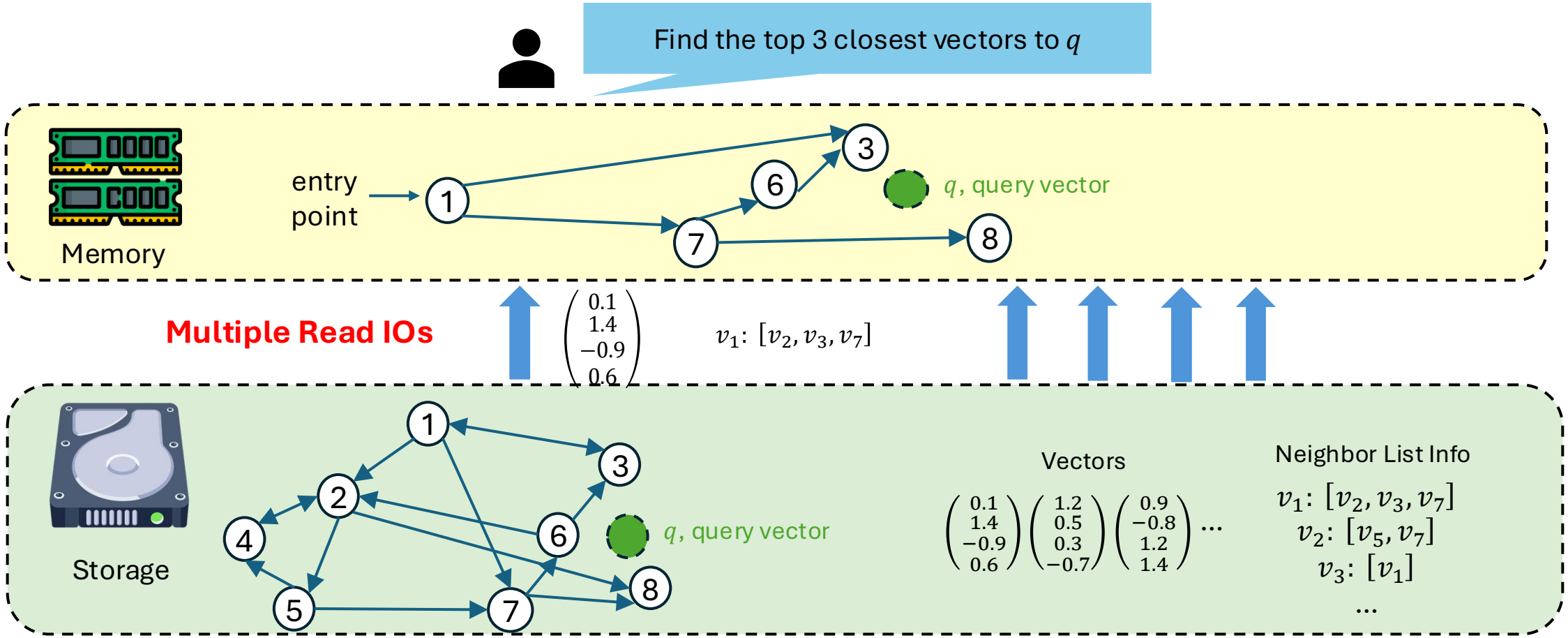




+ Better Scalability
+ Low Search Latency
+ High Search Accuracy

Note: In this talk, I will use vector search and graph-based ANN search interchangeably.

Read Amplification in Storage-Backed Graphs



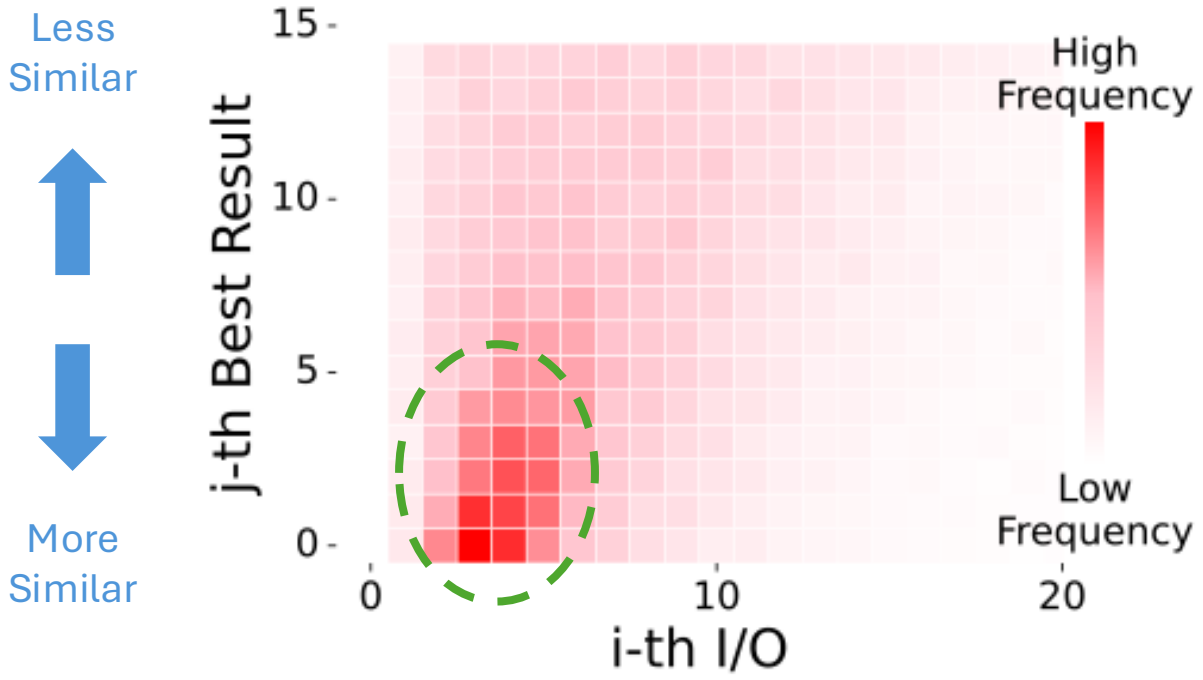
Disk IOPS Is Becoming the Bottleneck

- IOPS: number of I/O operations (reads or writes) a storage device can perform per second
- The issue of IOPS **amplification**: one ANN search query → 10s – 100s small disk reads
- **Multi-Tenancy** on a disk is VERY common in the cloud today

Storage Class	HDD	Flash SSD	NVMe SSD
Random IOPS (4KB)	50 – 200	10K – 100K	300K – 1M+
Upper Bound on ANN QPS	**0.5 – 2**	**10 – 100**	**3K – 10K**

Assume 100 I/Os per ANN search query

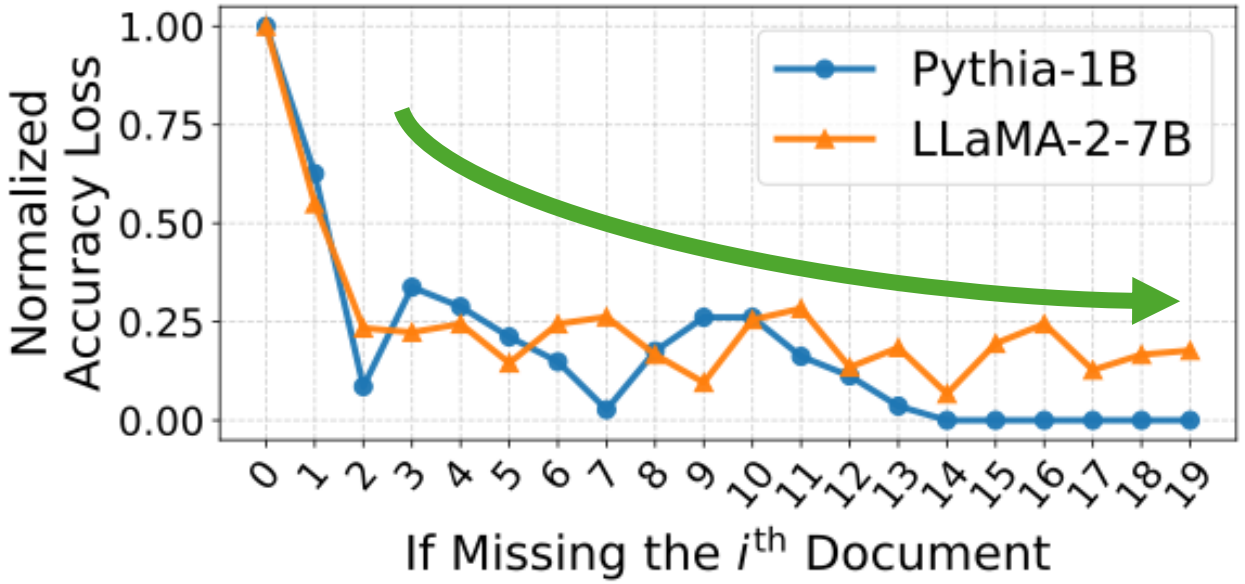
I/O Inefficiency in Search



- **The most similar results are found early in the search.**
- **More I/Os are wasted on searching less similar results.**

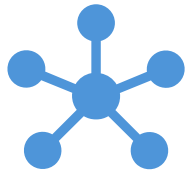
Top-Ranked Results Dominate Application Accuracy

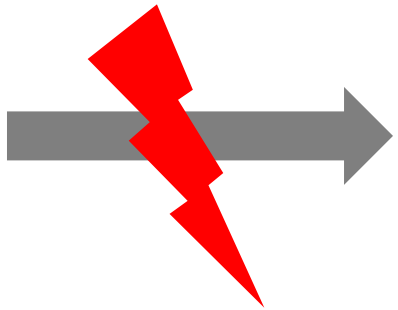
Task: Natural Questions (NQ)
Goal: Measure the relative impact of individual documents to the overall RAG accuracy
Y-axis: Report the normalized accuracy loss w.r.t. the oracle that uses the best 20 documents



The Trend

The Retrieval-Application Mismatch

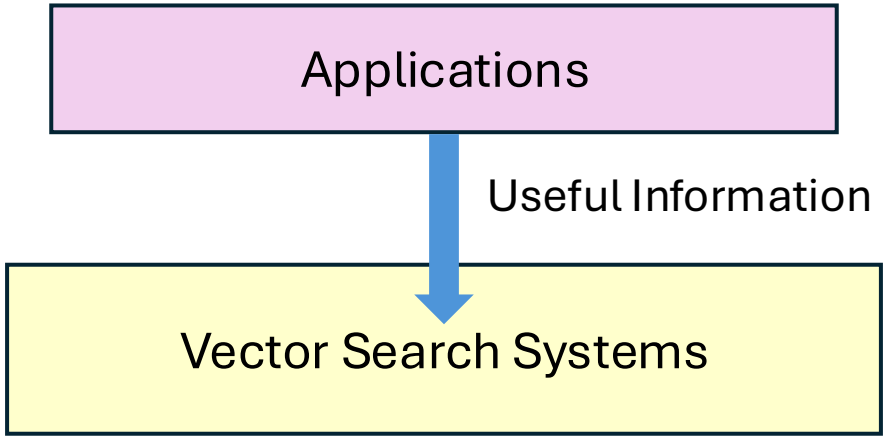




Vector search systems optimize retrieval of **all** top-k results.

RAG applications care about **top-ranked** results.

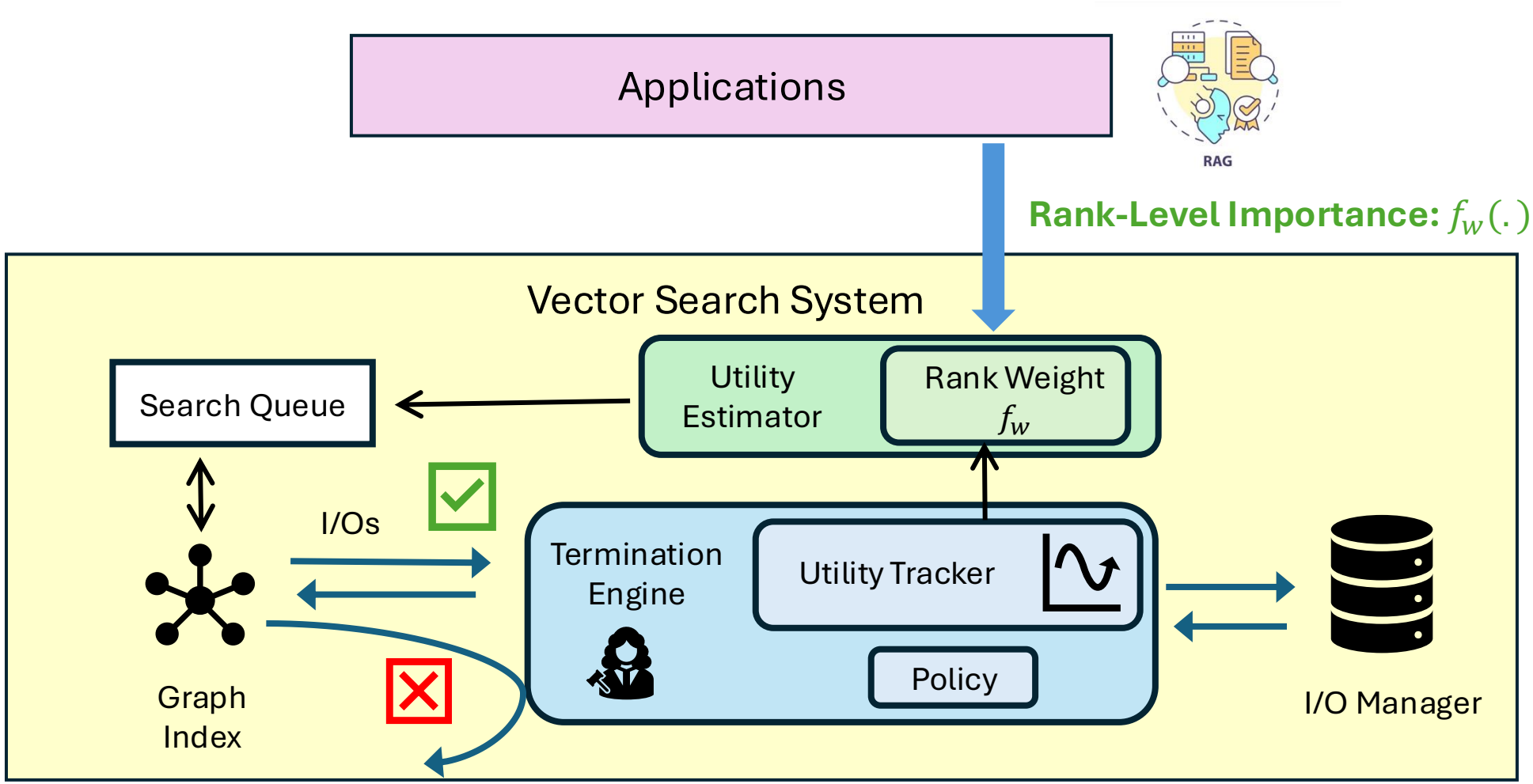
Key Insight: Align Vector Search with Application Utility



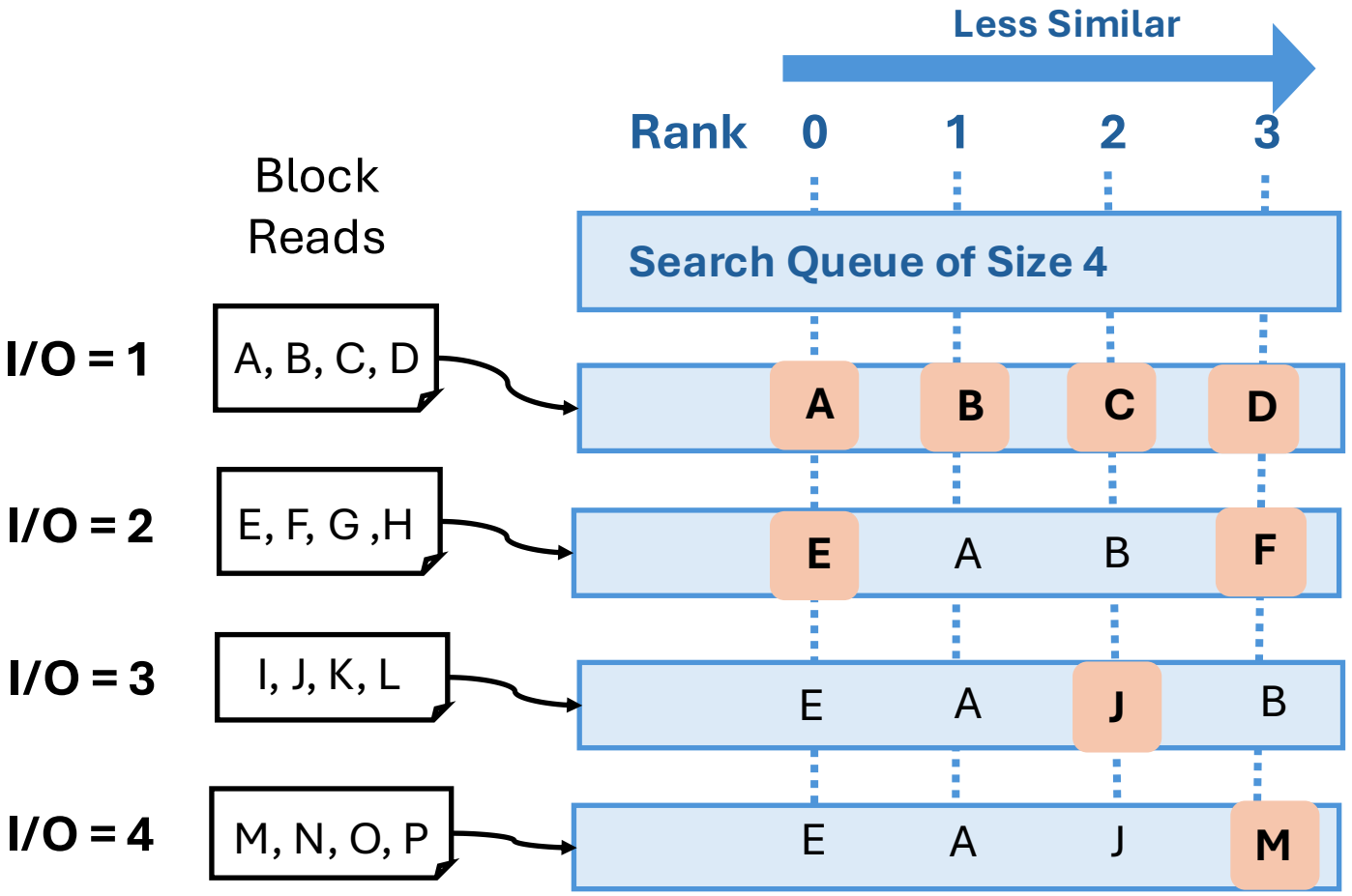
Our Design Goals

- Alleviate disk IOPS bottleneck.
- Preserve the accuracy of high-value results.
- Achieve a better performance-accuracy trade-off.

Terminus: **Rank-Aware** Early Termination for Vector Search



Rank-Aware I/O Utility Modeling



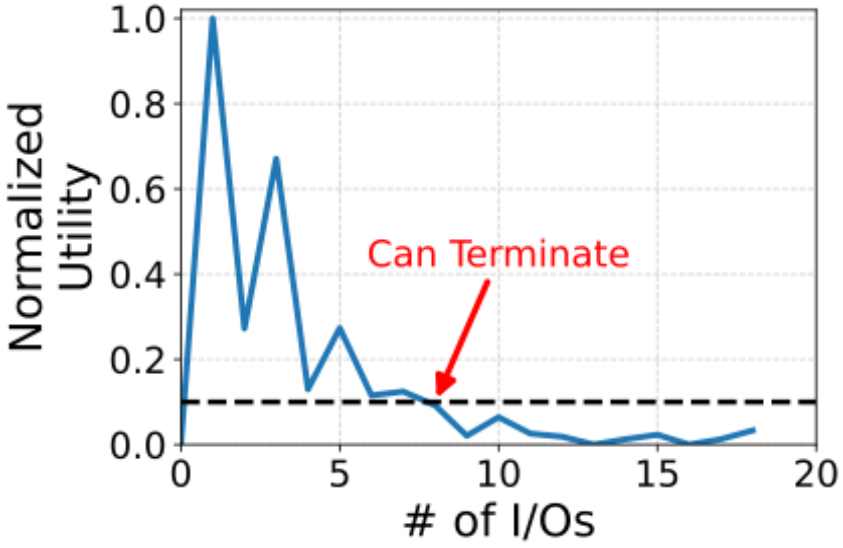
I/O Utility

$\mathbf{0}$
$f_w(0) + f_w(1) + f_w(2) + f_w(3)$
$f_w(0) + f_w(3)$
$f_w(2)$
$f_w(3)$

$f_w(.)$: Rank Weight Function

Dynamic Early Termination

- U_i: the rank-aware utility of the i^{th} I/O
- Terminate when utility of the most recent X I/Os becomes negligible
  - If $U_i < \varepsilon, \forall\, i \in [i+1-X,\ i]$ where X is the sliding window size and ε is the termination threshold



Evaluations

- How does Terminus compare to other early termination methods for vector search?
- What are the impacts of early termination?

See the paper

- What is the overall trade-off between vector search performance and RAG accuracy in Terminus?

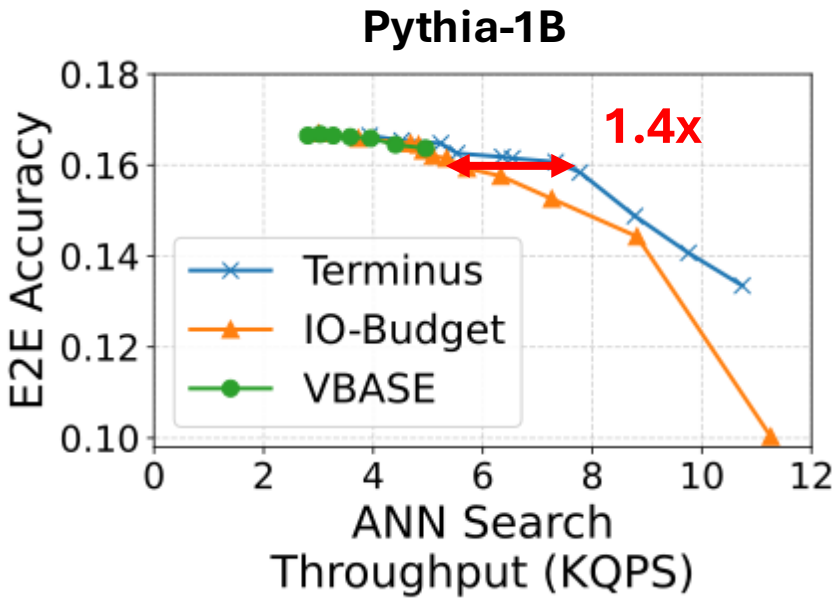
Same E2E Accuracy, Much Higher Throughput

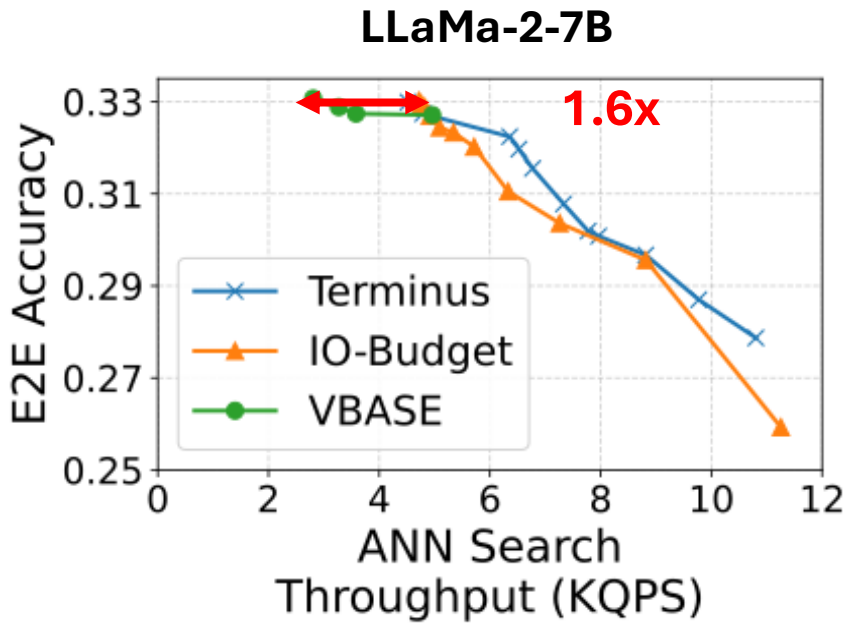
k = 20

Task: Natural Questions (NQ)

IO-Budget: an I/O-first rank-agnostic strawman that terminates when the total number of I/Os of a query exceeds some upper bound N

VBASE: a conservative method that terminates when the top-E results in the search queue stabilize, where $\mathrm{E} \geq k$





Conclusions

- Current vector search systems optimize retrieval quality **uniformly across all** top-k results.
- Retrieval utility in RAG **concentrates at top ranks**.
- This mismatch leads to unnecessary I/O consumption.
- Terminus is a **utility-driven early termination** system for graph ANN search.
- **Rank-aware** retrieval offers **a better trade-off** between retrieval performance and application accuracy.

You can reach out to us at amberljn@icloud.com.