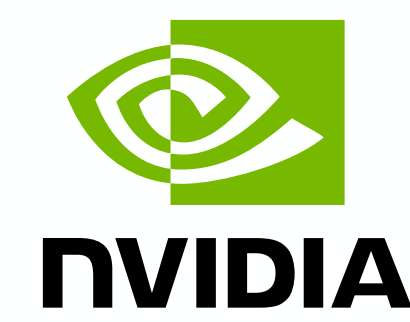




VecFlow: A High-Performance Vector Data Management System for Filtered-Search on GPUs



Jingyi Xi^{1*}, Chenghao Mo^{1*}, Ben Karsin², Artem Chirkin², Mingqin Li³, Minjia Zhang¹

¹: SSAIL Lab, UIUC ²: Nvidia ³: Microsoft

Filtered Vector Search

Datasets and queries carry **labels**

HIT	MISS	HIT
Image A	Image B	Image C
Italy	Japan	Italy
2015	2020	2015
Green	Nature	Food

Query: similar images AND labels \supseteq {Italy}

Return top-K similar items satisfying **all filters**.

Motivation

Wide applications

e.g., multisearch with text hints, region-relevant ads, enterprise search with permissions

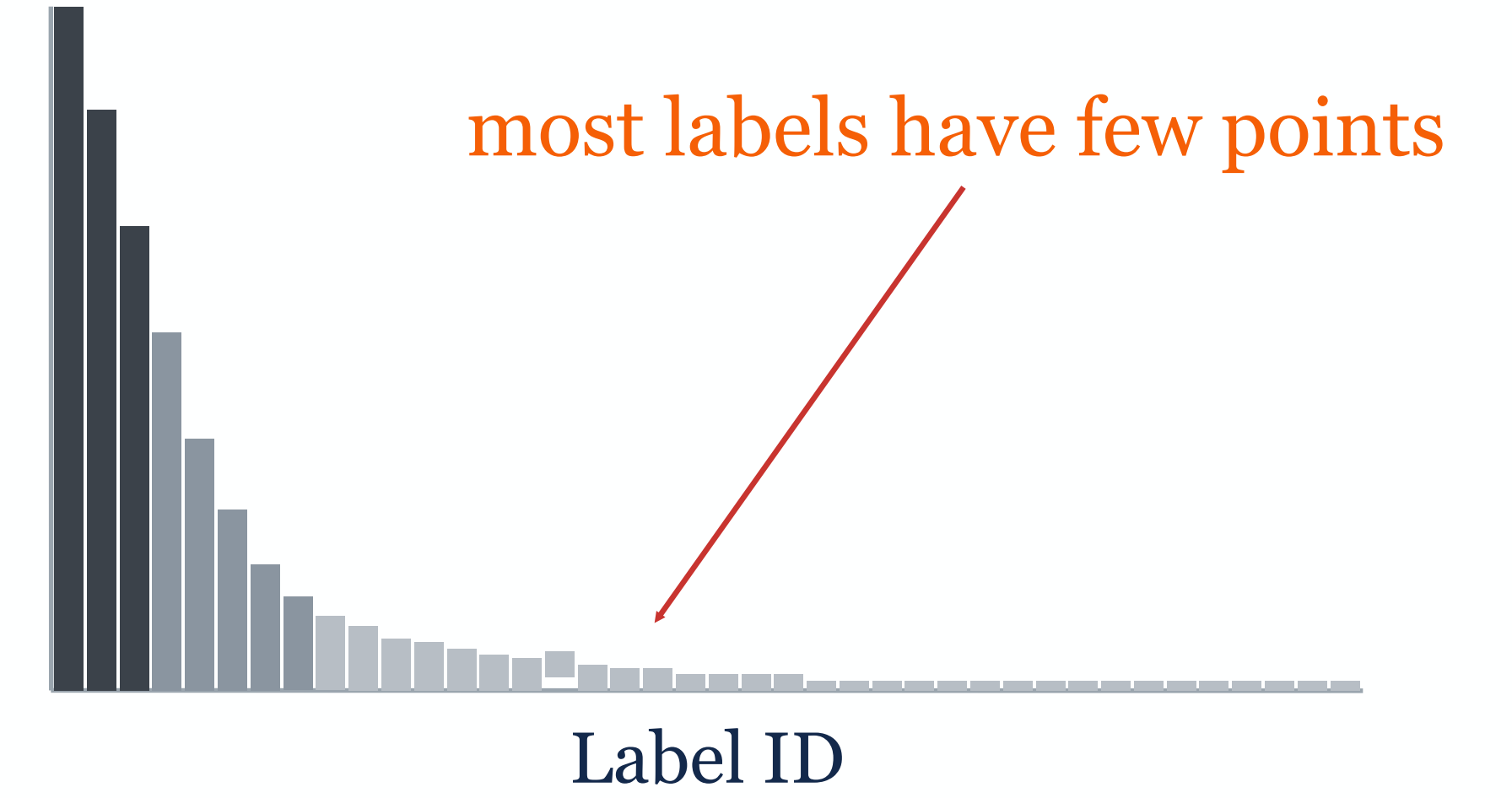
GPUs give the highest throughput

high memory bandwidth (>2 TB/s, ~10x a CPU) and massive parallelism

Label Selectivity Challenge

Label frequencies are **long-tailed**

Number of points



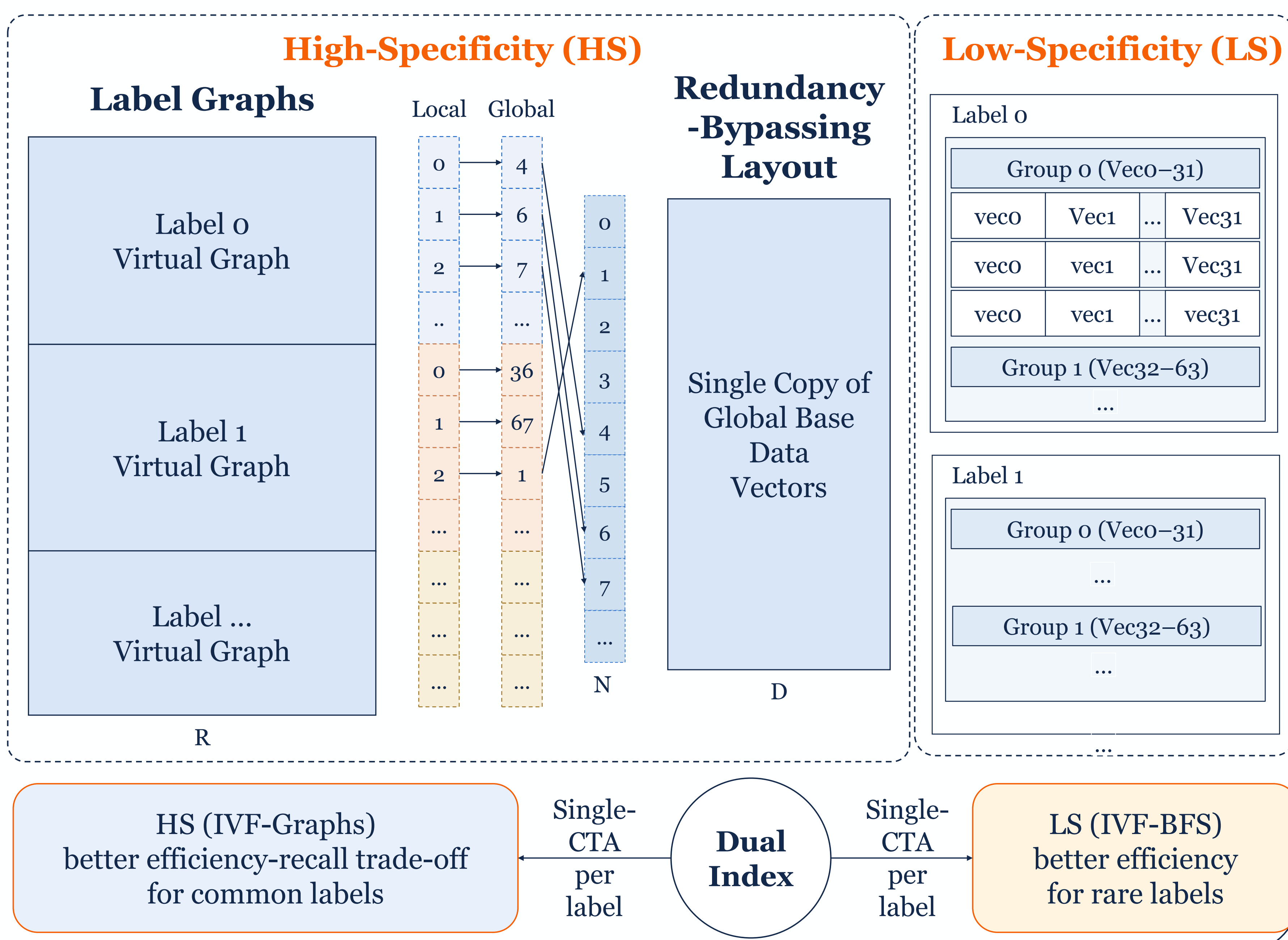
Queries can hit **rare labels** with few points.

VecFlow Architecture Overview: Query Processing → Dual Indexing → GPU Search

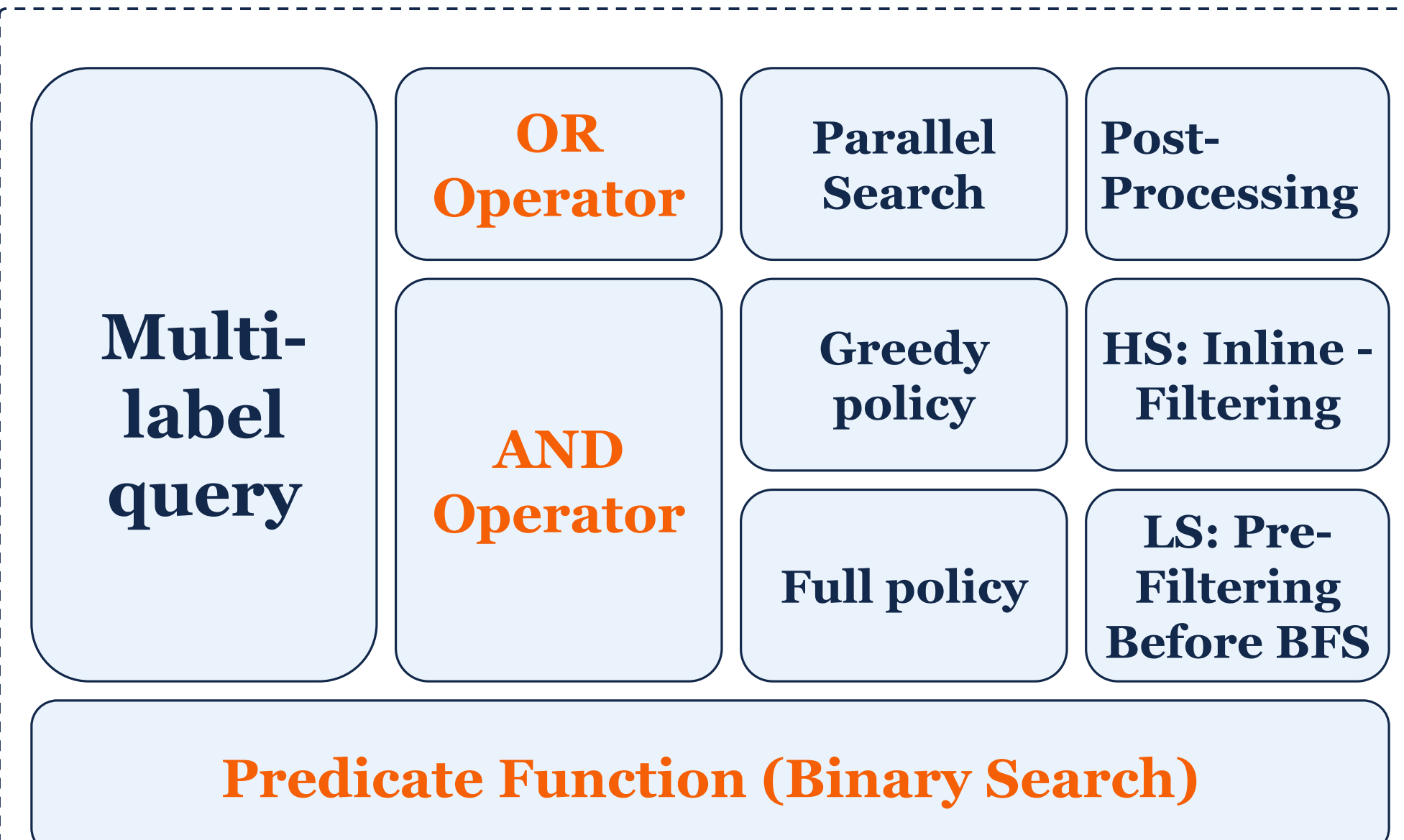
1 Query Processor



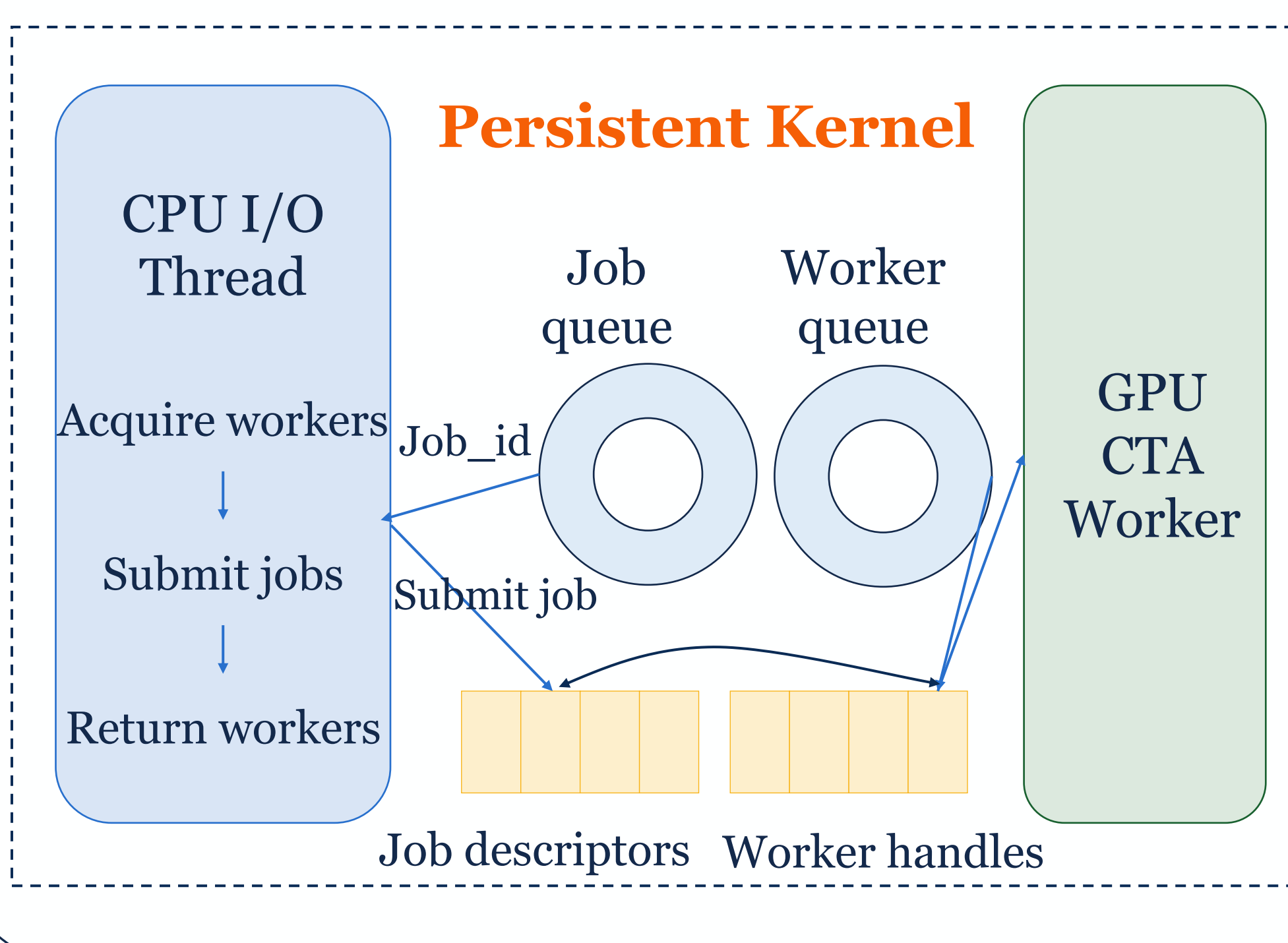
2 Dual-Structured Label-Centric IVF Index



3 Search Strategy: multi-label query

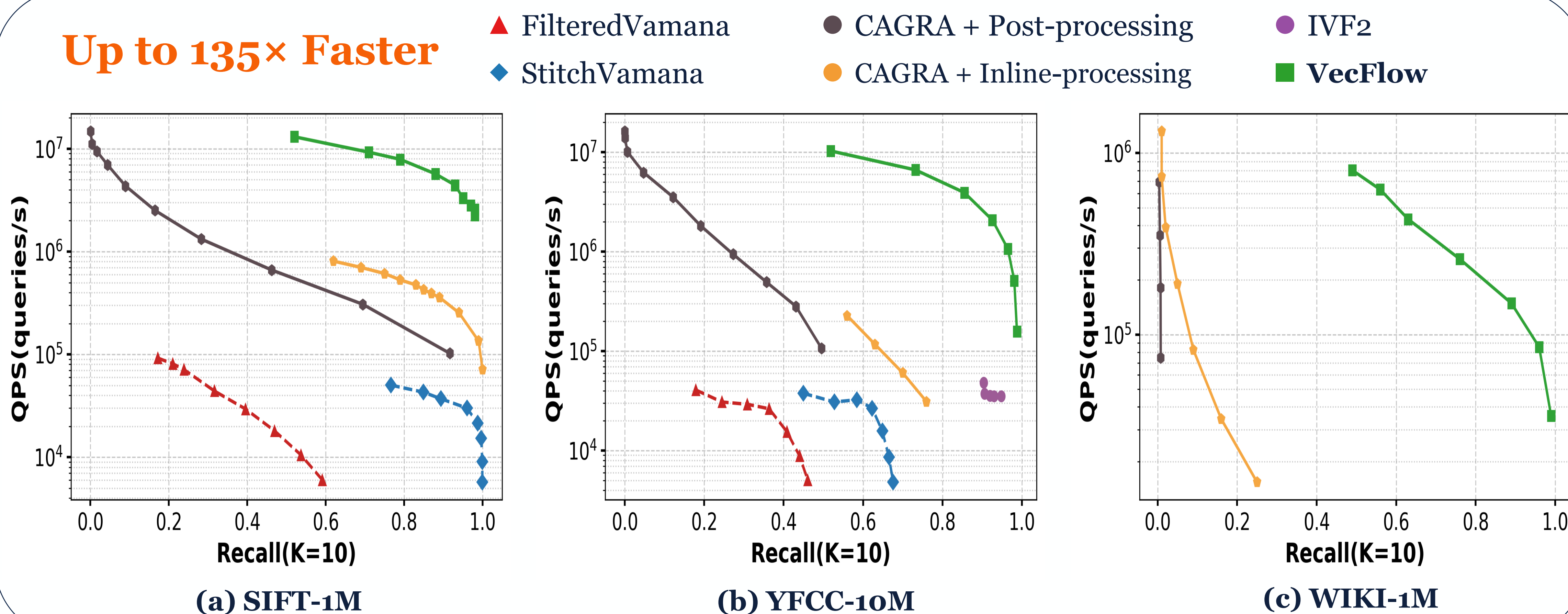


4 Streaming Small Async Request



Evaluation Highlights

Up to 135× Faster



Scan to Share



Code:
github.com/
Supercomputing-
System-AI-Lab/
VecFlow



Paper:
https://dl.acm.org/
doi/abs/10.1145/3749189