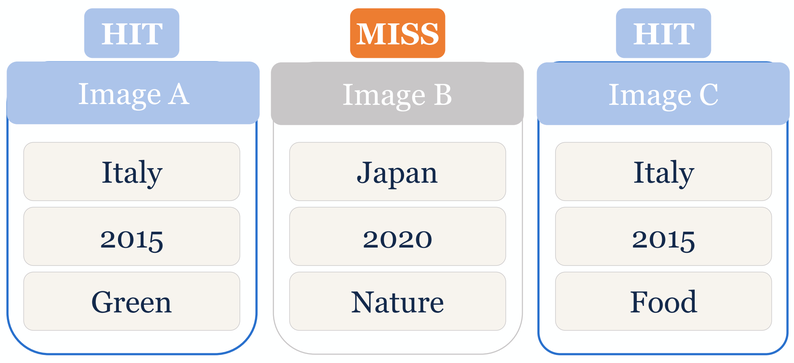
VecFlow: A High-Performance Vector Data Management System for Filtered-Search on GPUs

Jingyi Xi[1*], Chenghao Mo[1*], Ben Karsin[2], Artem Chirkin[2], Mingqin Li[3], Minjia Zhang[1]

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

Filtered Vector Search

Datasets and queries carry labels



Query: similar images AND labels ⊇ {Italy}

Return top-K similar items satisfying all filters.

Motivation

Wide applications

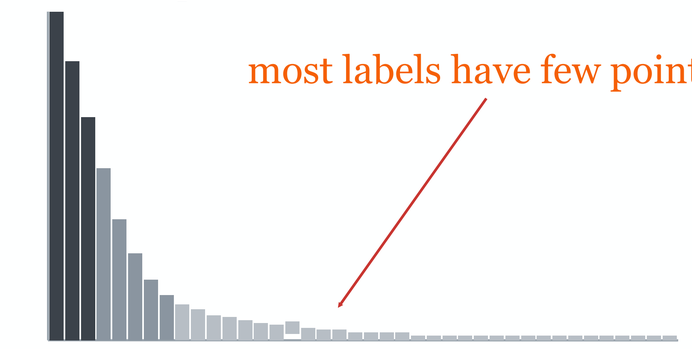
e.g., multisearch with text hints, region-relevant ads, enterprise search with permissions

GPUs give the highest throughput

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

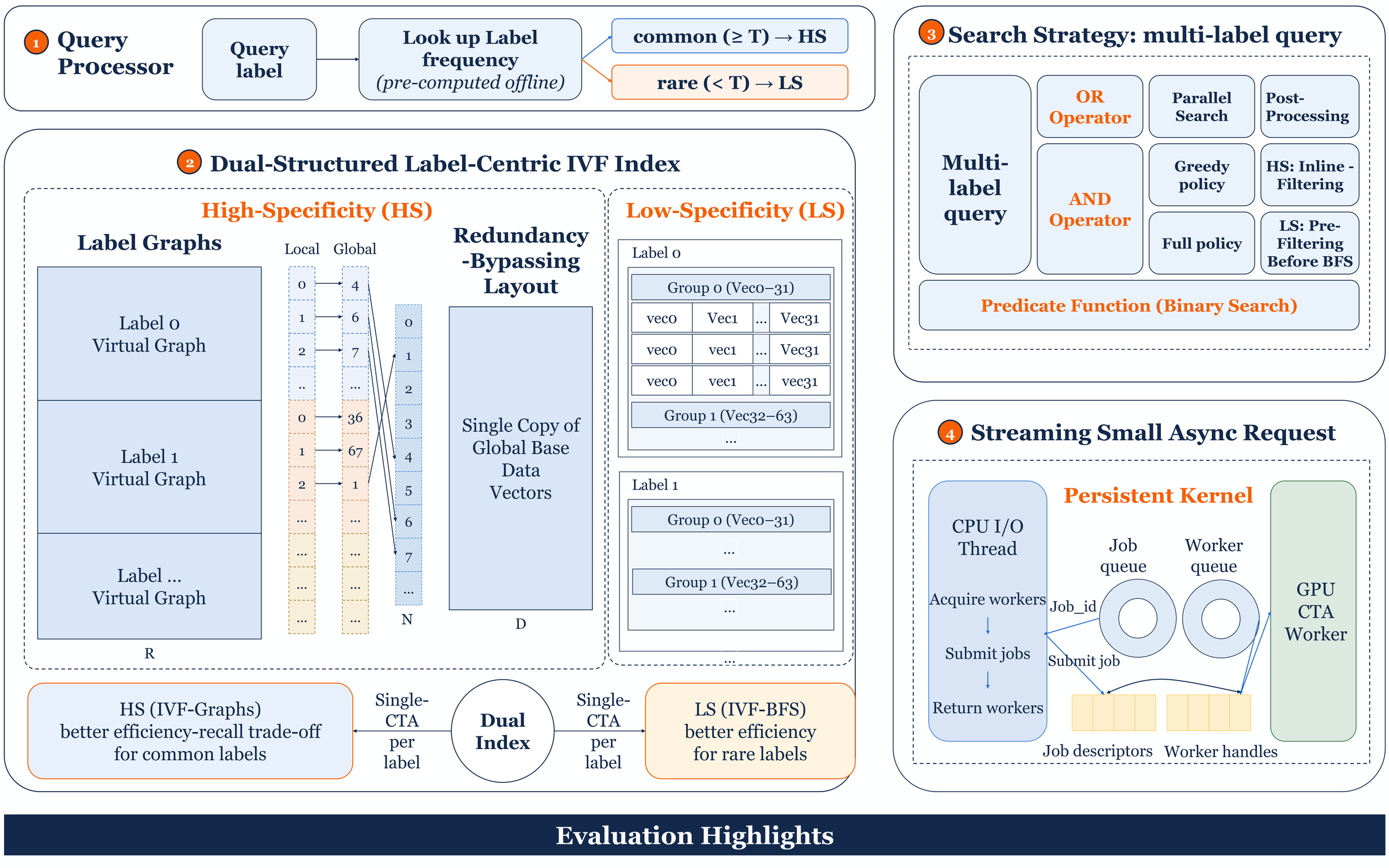
Label Selectivity Challenge

Label frequencies are long-tailed



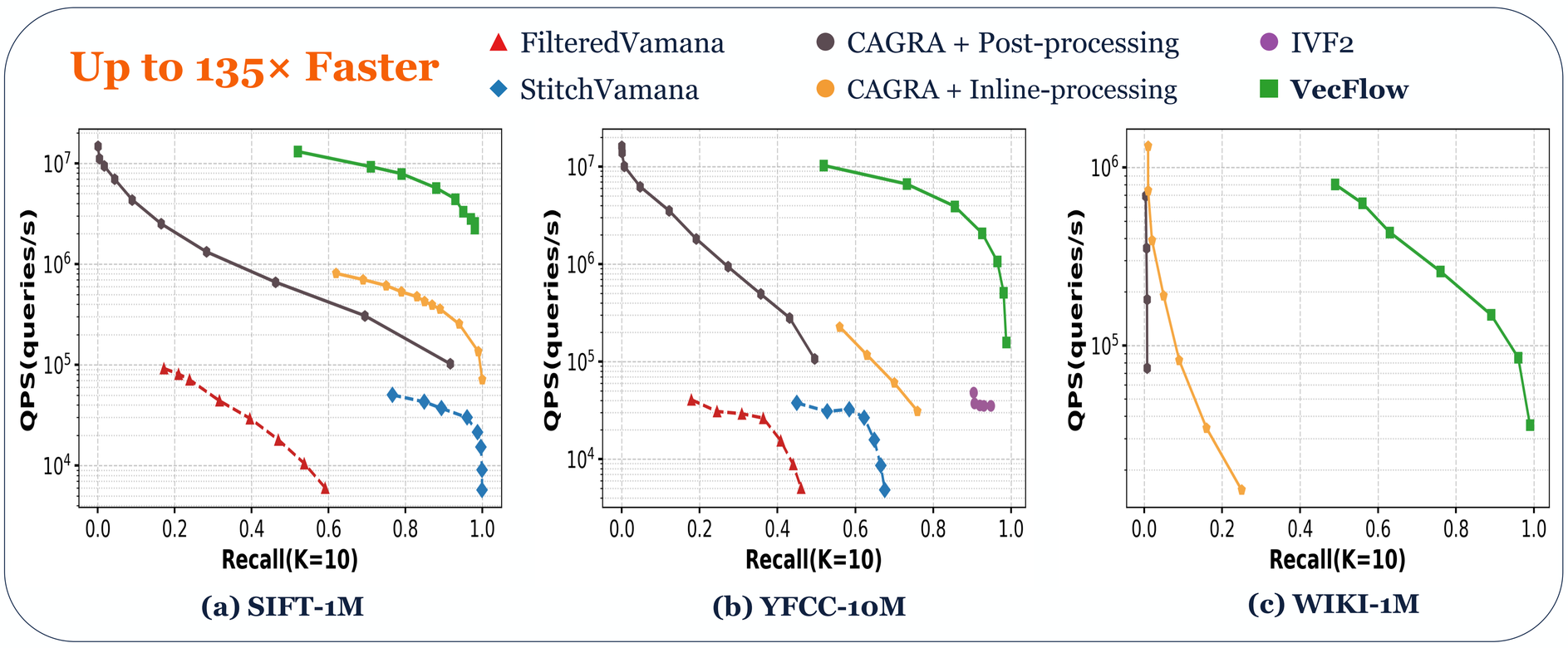
Queries can hit rare labels with few points.

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



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