Scaling Retrieval With Sharding

1.2B Chunks at <100ms

https://github.com/skeptrunedev/scaling-search

Nicholas Khami

Founder/CEO

trieve

Twitter: @skeptrune
Mastodon: @skeptrune
LinkedIn: - nicholas-khami
GitHub: github.com/skeptrunedev

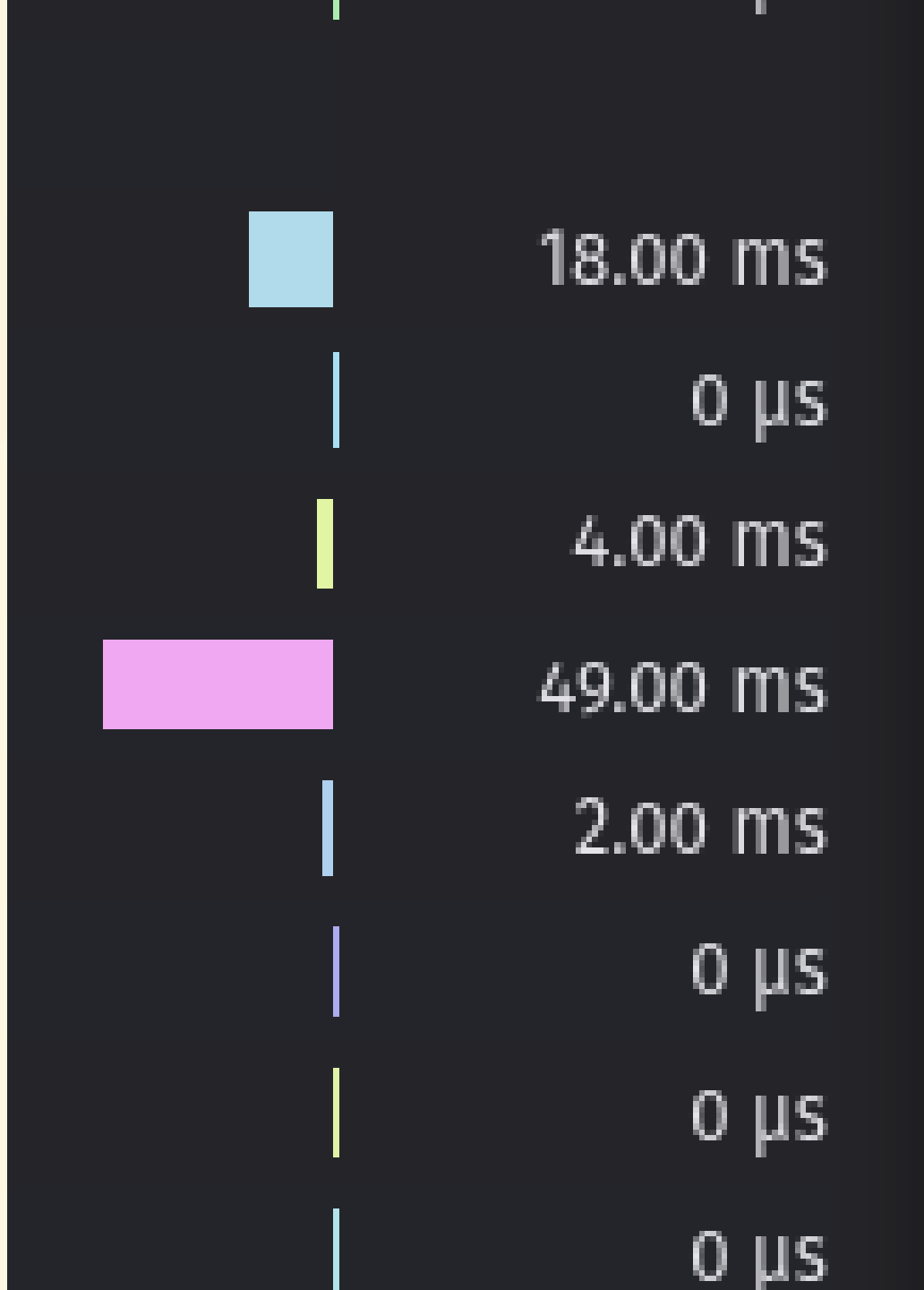
Agenda

- Latency by layer in retrieval
- Fast Typo Detection
- Optimizing vector inference
- Scaling ingest
- Sharding your search index

@skeptrune - nick.k@trieve.ai

Latency by layer

1. Inference vectors for query ~15ms
2. Typo detection+correction ~1ms
3. Score docs in search index ~10ms
4. Re-rank (LTR/cross-encode) ~15ms
5. First LLM token (TTFT) ~250ms



@skeptrune - nick.k@trieve.ai

Fast Typo Correction

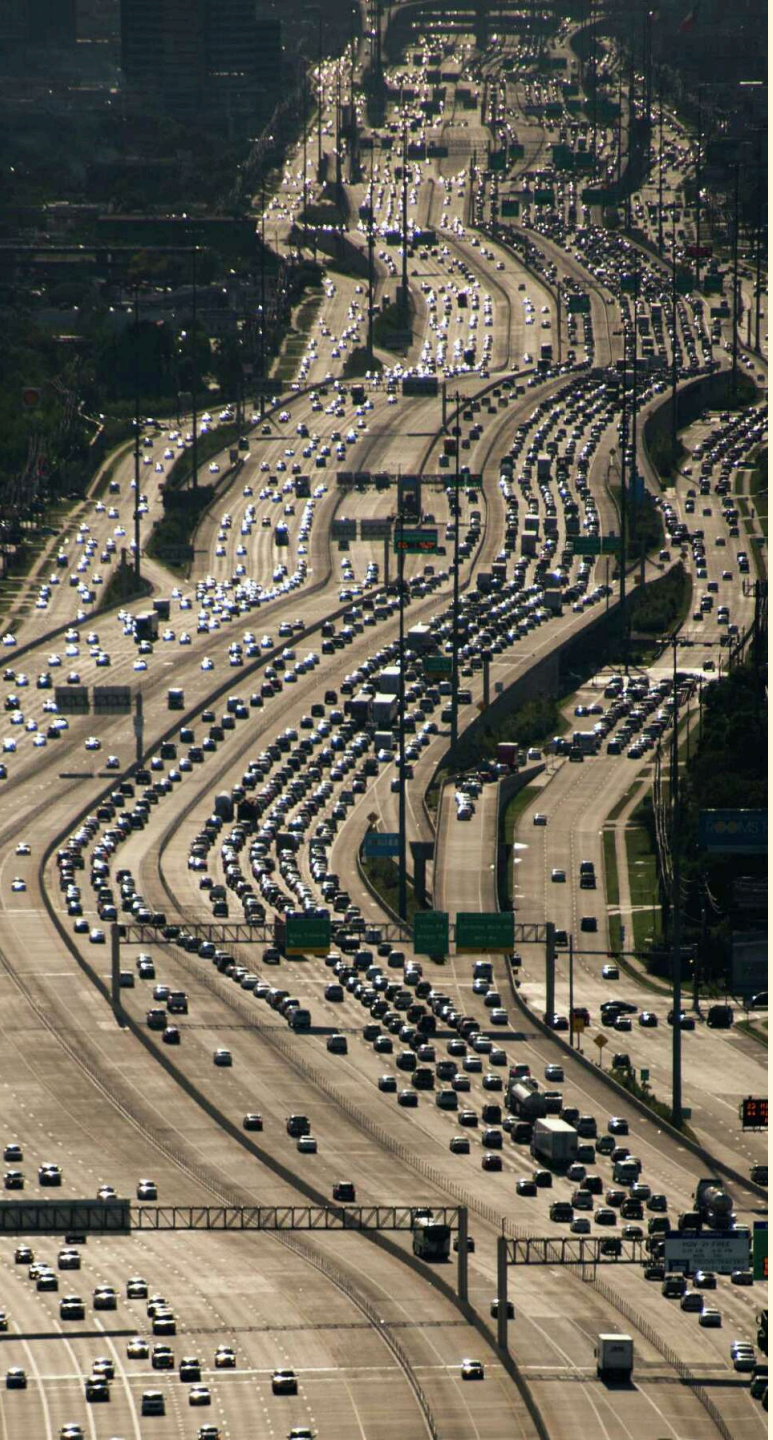
- Many vector db's don't correct typos
  - Harder when multi-tenant
- BK-Tree vs. SymSpell Data Structures
  - SymSpell is 100x faster

trieve.ai/building-blazingly-fast-typo-correction-in-rust

PSA: Vector/Re-ranker Latency

- Cloud API's take ~150ms, but 15ms possible w/ Candle
- Ingesting 1,000,000 chunks (33k batches)
  - ~80mins using OpenAI
  - ~8mins on Candle
- LTR plugins usually <10ms

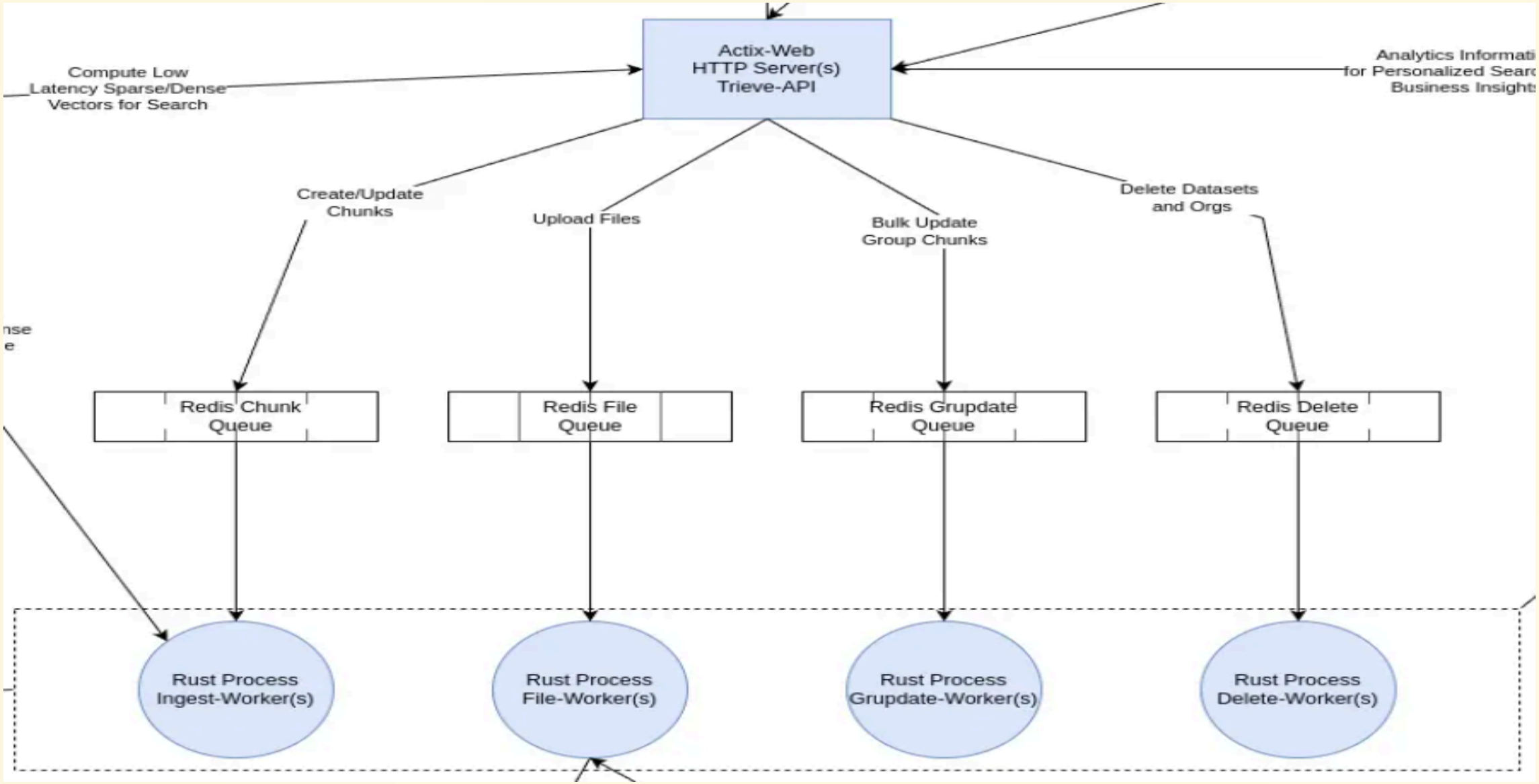
docs.trieve.ai/vector-inference



Scaling Ingest

- You almost certainly need a queue+worker architecture
- Using a worker for updates and deletions enables chunking experiments at scale
- Consider an id system which allows you to upsert easily

github.com/trieve/ingestion-worker



@skeptrune - nick.k@trieve.ai

Sharding your search index Pt. 1

TLDR: think about it

What is a shard?

- A shard is an instance of the search lib (Lucene, FAISS, etc.)
- Shards are made up of segments
- Segments are indices (IDF or HNSW)
- Small segments should be merged off-peak hours

@skeptrune - nick.k@trieve.ai

Sharding your search index Pt. 2

How to optimize?

- Usually best to think about sharding before HNSW params
- Always have `shards > 2*nodes` such that you can horizontally scale CPU if needed
- Replicate shards for more read throughput
- Different db's map their thread pools to shards differently and you should work with your vendor to address it

@skeptrune - nick.k@trieve.ai