SIVF: Streaming Inverted File Indexing on GPUs

Dongfang Zhao
dzhao@uw.edu
HPDIC Lab
University of Washington, USA

Abstract

1 Introduction

1.1 Motivation

Real-world vector search applications, such as real-time recommendation and fraud detection, increasingly operate on streaming data where timeliness is critical. These systems require a sliding window model where expired vectors must be evicted as new vectors arrive to maintain bounded memory usage. However, existing GPU-accelerated approximate nearest neighbors (ANN) indices are predominantly optimized for write-once-read-many workloads.

To quantify the gap between insertion and eviction performance, we conducted a benchmark using Faiss [1] on an NVIDIA RTX 6000 GPU with the SIFT1M dataset [2] on the Chameleon testbed [3]. As illustrated in Figure 1, we observe a severe performance asymmetry. While inserting a batch of 10,000 vectors takes only 28.2 ms due to efficient GPU parallelism, evicting the same number of vectors incurs a latency of 212.7 ms. This constitutes a 7.6 times slowdown.

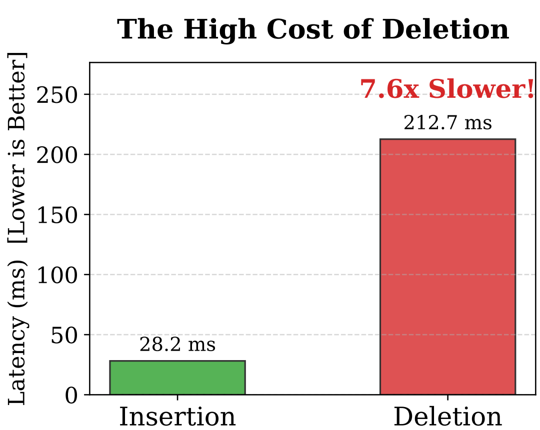

Figure 1: The High Cost of Deletion. While GPU parallelism accelerates vector insertion (28.2 ms), the lack of in-place eviction support causes deletion latency to increase to 212.7 ms. This represents a 7.6 times slowdown. This asymmetry makes data eviction the primary bottleneck in streaming scenarios.

This bottleneck stems from the architectural mismatch in current library designs. Since standard IVF (Inverted File) indices lack support for efficient in-place deletion on the GPU, evicting data necessitates a costly CPU-GPU roundtrip. The entire index structure must be copied back to the host, compacted on the CPU, and re-uploaded to the device. This IO-bound operation prevents the system from fully utilizing the GPU compute throughput and limits the maximum sustainable ingestion rate of the system.

Acknowledgment

Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation.

References

[1] Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data 7*, 3 (2021), 535–547.

[2] Jégou, H., Douze, M., and Schmid, C. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 1 (2011), 117–128.

[3] Keahey, K., Anderson, J., Zhen, Z., Riteau, P., Ruth, P., Stanzione, D., Cevik, M., Colleran, J., Gunawi, H. S., Hammock, C., Mambretti, J., Barnes, A., Halbach, F., Rocha, A., and Stubbs, J. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association, July 2020.