

Accelerating Data Deduplication with Content-Defined Skips and Vector Instructions

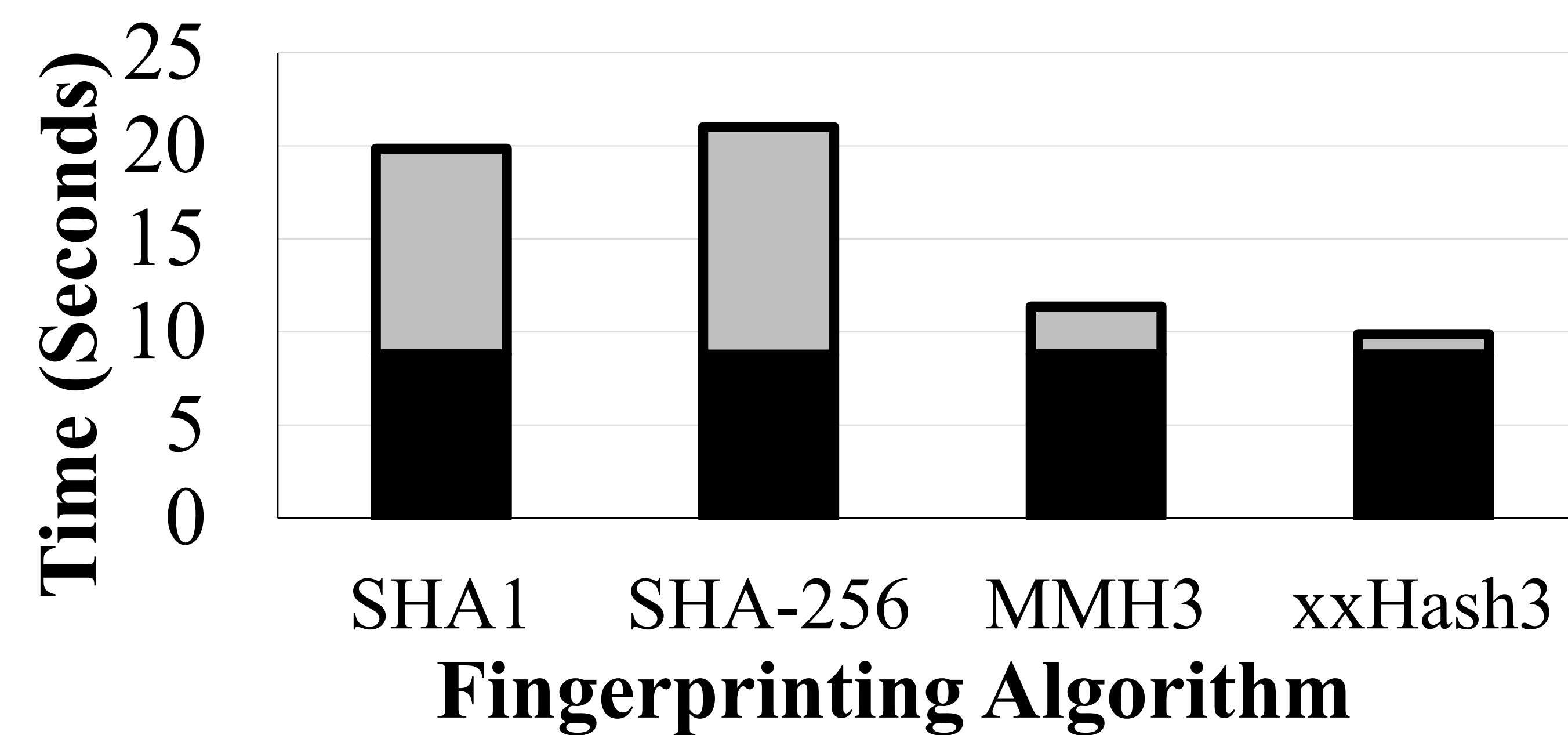
Sreeharsha Udayashankar and Samer Al-Kiswany

Introduction

1

- Data deduplication is used to conserve storage space
- Four phases: Data Chunking, Fingerprinting, Comparison and Storage
- Data chunking performed using CDC algorithms – **Performance bottleneck!**

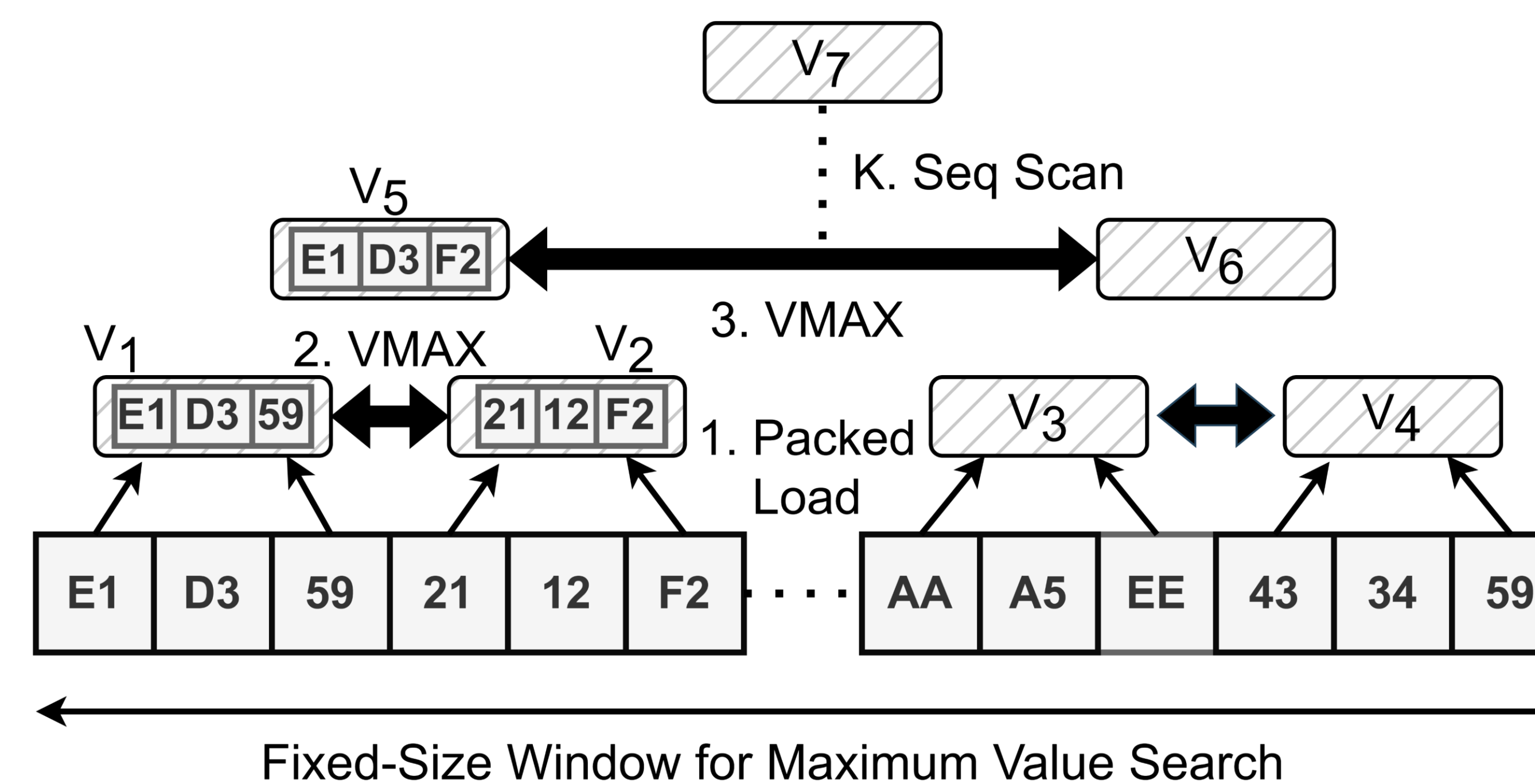
■ Data Chunking ■ Fingerprinting



VectorCDC: Accelerating Data Chunking with Vector Instructions [1, 2]

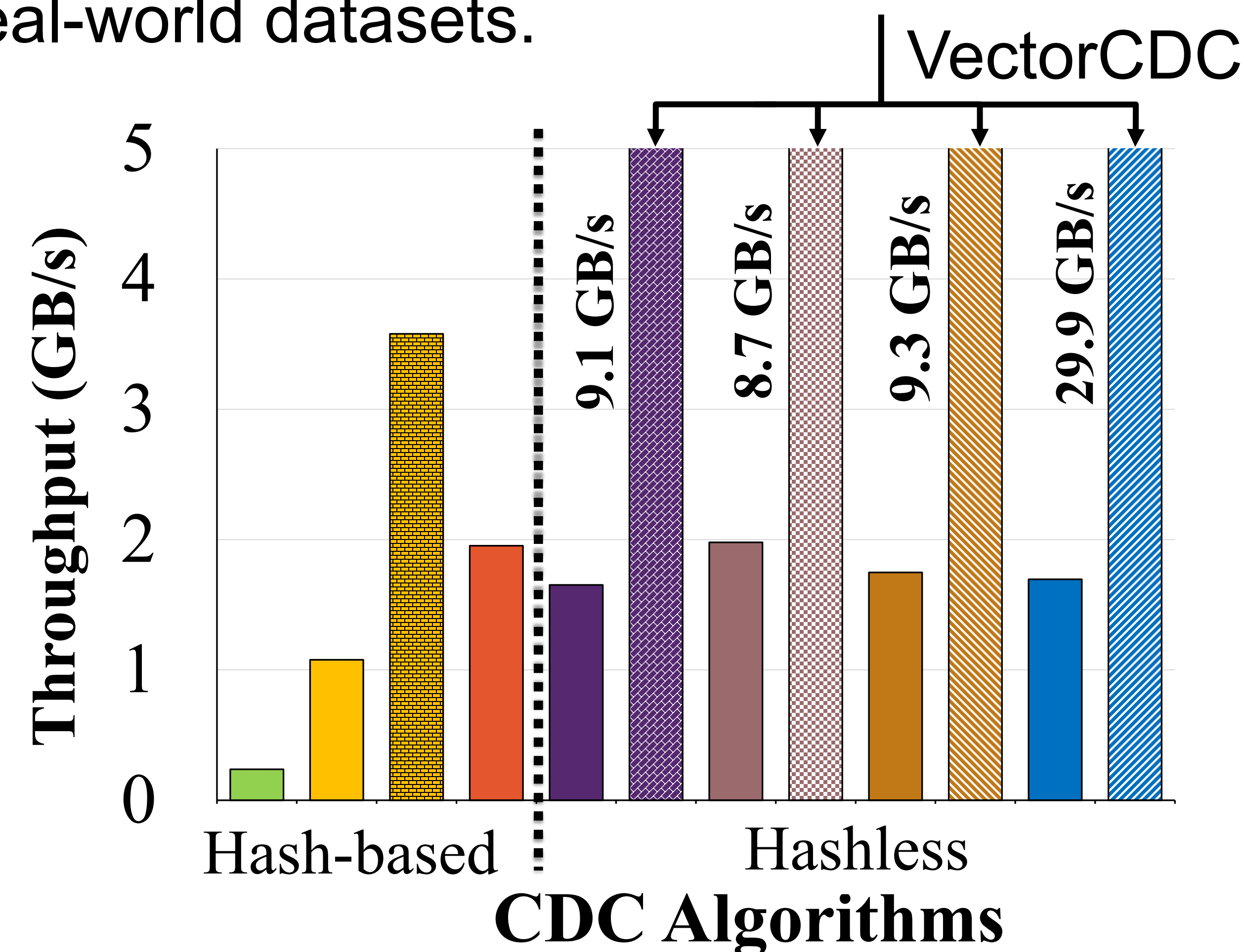
2

- Idea:** Use special *vector CPU instructions* (SSE/AVX) to accelerate data chunking.
- Insight:** Hashless chunking algorithms have two phases; Extreme Byte Search and Range Scan.



Tree-based Extreme Byte Search

- Accelerates throughput by 8x – 24.2x on real-world datasets.



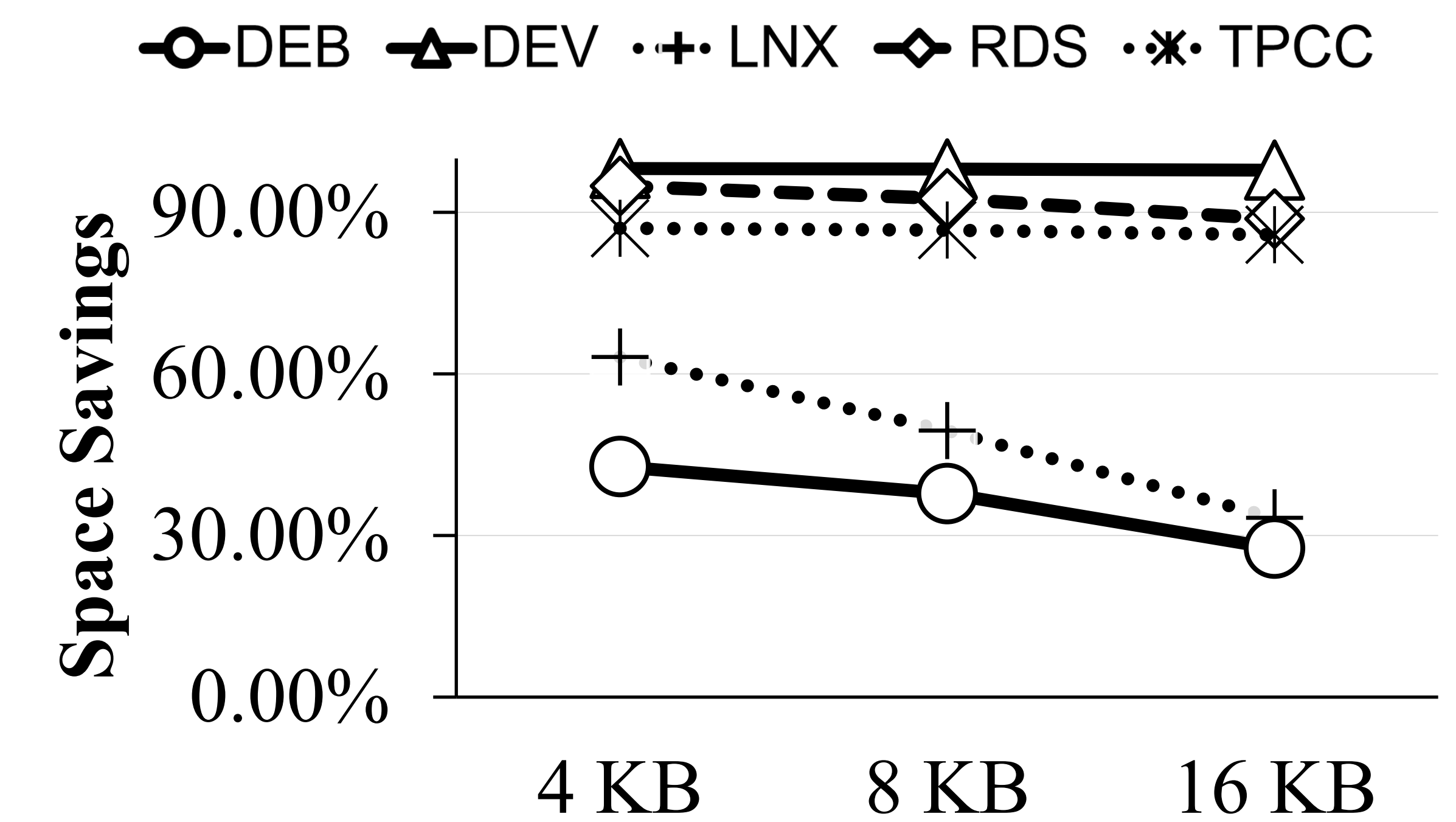
Legend for CDC Algorithms:

- TTTD, GEAR, SS-GEAR, FCDC, AE-Max, VAE-Max, AE-Min, VAE-Min, MAXP, VMAXP, RAM, VRAM

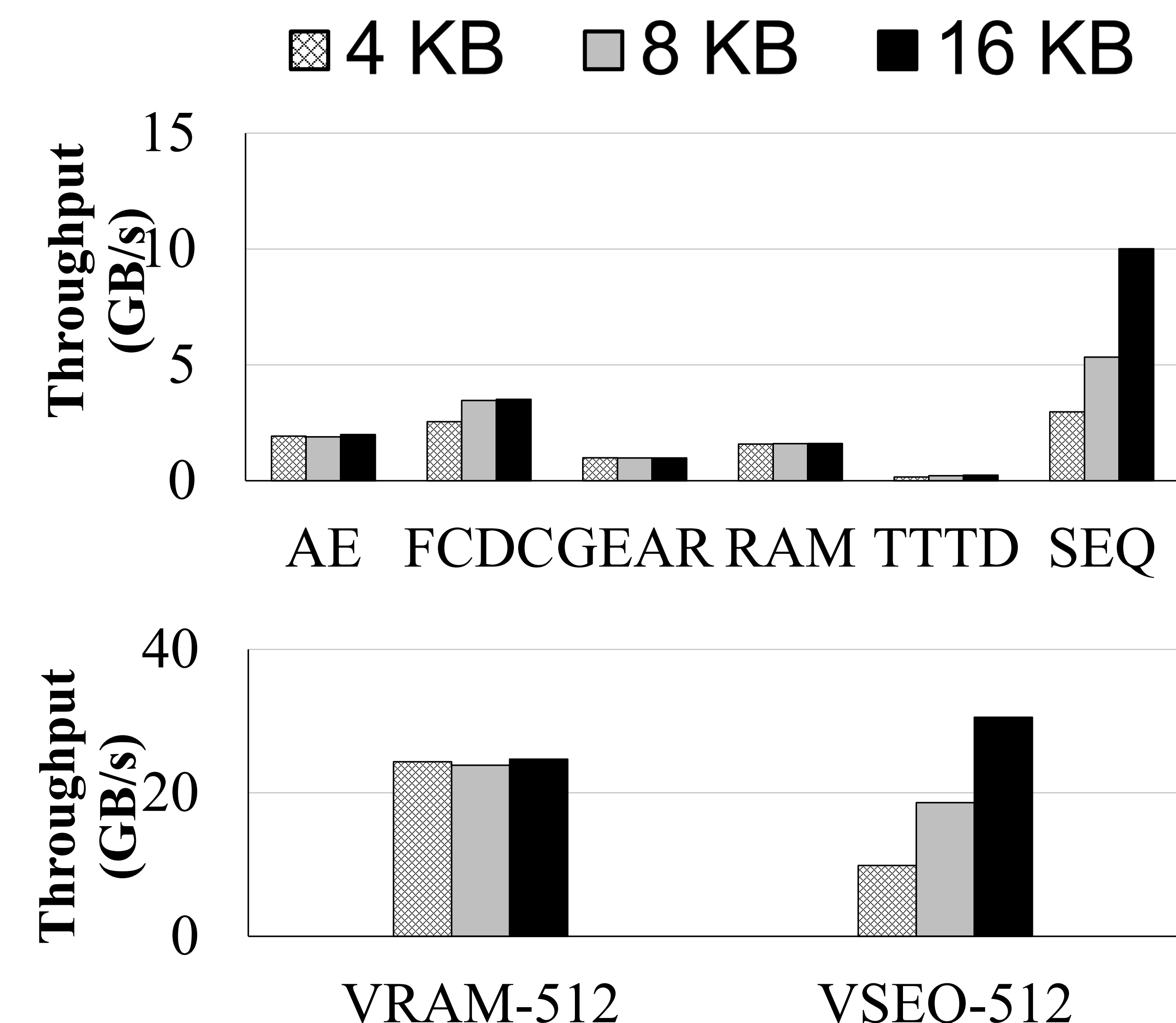
SeqCDC: Content-Defined Skips for Large Chunk Sizes [3, 4]

3

- Systems in production favor larger chunks for certain datasets. Existing algorithms designed for smaller chunks.



- Idea:** Skip *unfavorable regions* to increase throughput.
- Insight:** Randomly skipping data regions causes loss of space savings; we regulate it with content-defined heuristics.



Summary

4

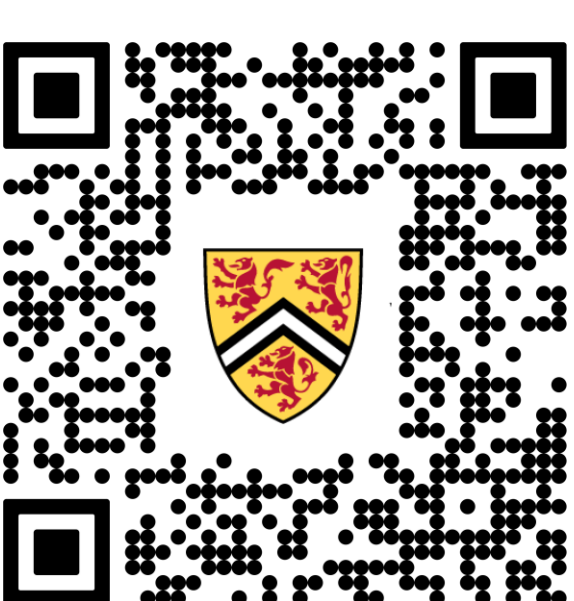
- VectorCDC and SeqCDC effectively alleviate the performance bottleneck in data deduplication.
- Patented in partnership with Acronis
- Published at top systems venues [1-4]

[1] VectorCDC: Accelerating Data Deduplication with Vector Instructions. *USENIX FAST '25*.

[2] Accelerating Data Chunking in Deduplication Systems using Vector Instructions, *ACM Transactions on Storage*.

[3] SeqCDC: Hashless Content-Defined Chunking for Data Deduplication. *ACM Middleware '24*.

[4] Vectorized Sequence-Based Chunking for Data Deduplication. *IEEE Transactions on Parallel and Distributed Systems*.



UNIVERSITY OF
WATERLOO

DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

Acronis