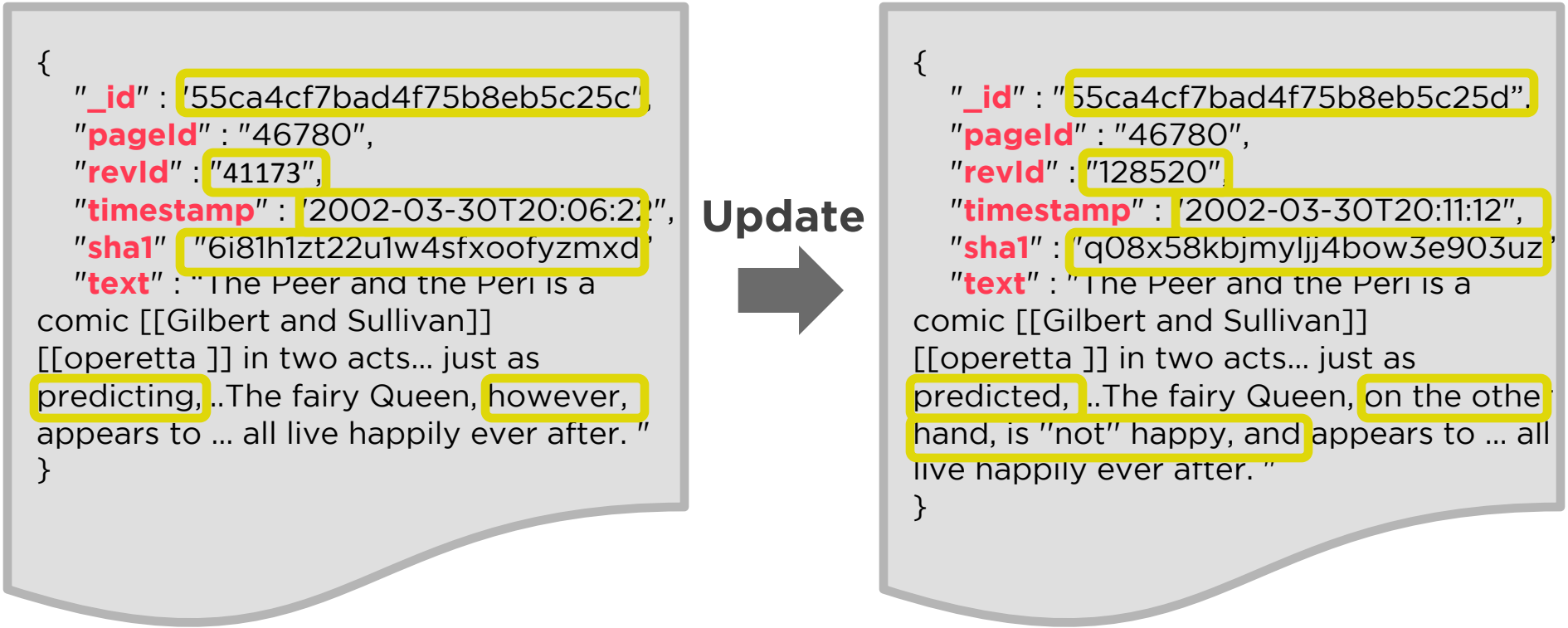
Reducing Replication Bandwidth for Distributed Document Databases

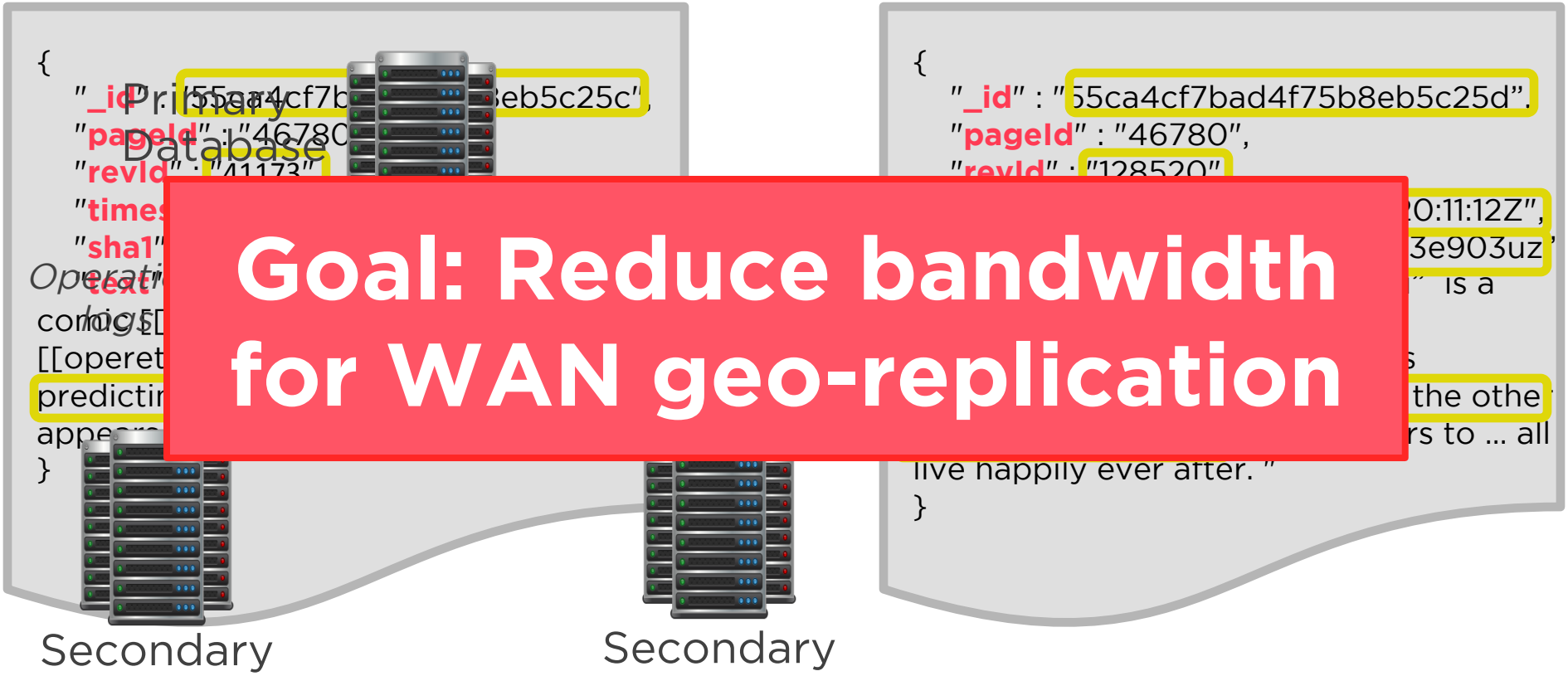
Lianghong Xu[1], Andy Pavlo[1], Sudipta Sengupta[2]
Jin Li[2], Greg Ganger[1]
Carnegie Mellon University[1], Microsoft Research[2]

Document-oriented Databases



Update: Reading a recent doc and writing back a similar one

Replication Bandwidth

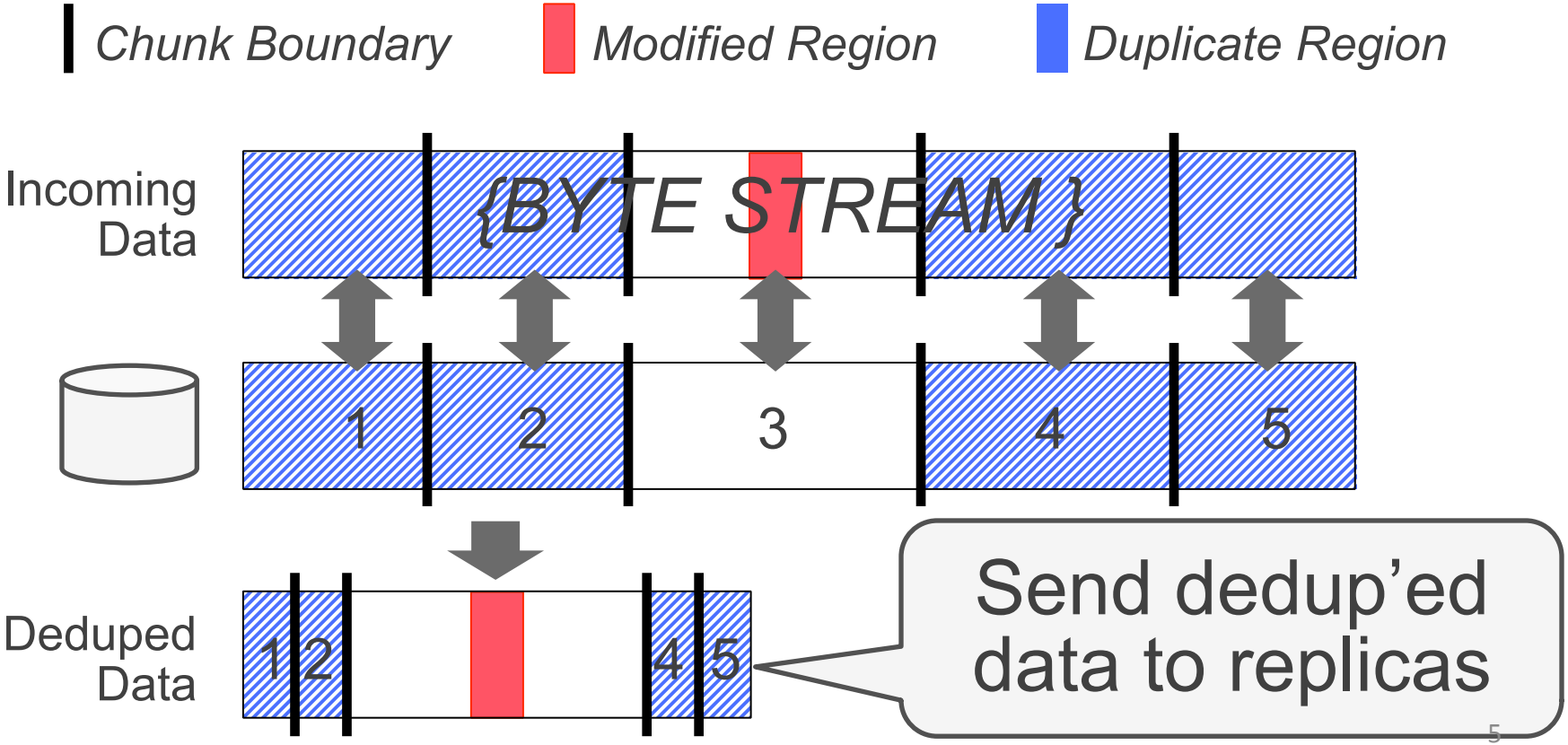


Goal: Reduce bandwidth for WAN geo-replication

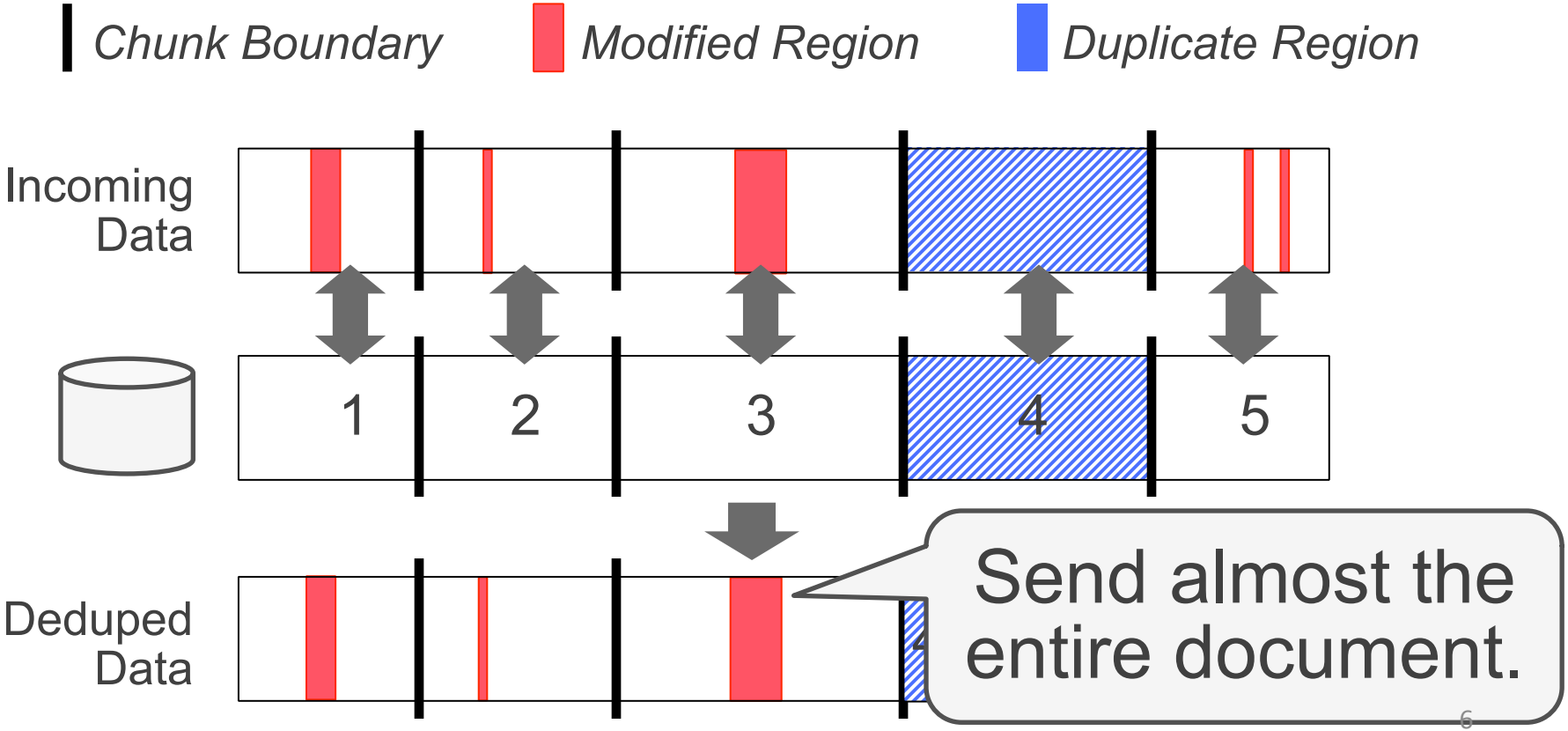
Why Deduplication?

- Why not just **compress**?
  - *Oplog batches are small and not enough overlap*
- Why not just use **diff**?
  - *Need application guidance to identify source*
- **Dedup** finds and removes redundancies
  - *In the entire data corpus*

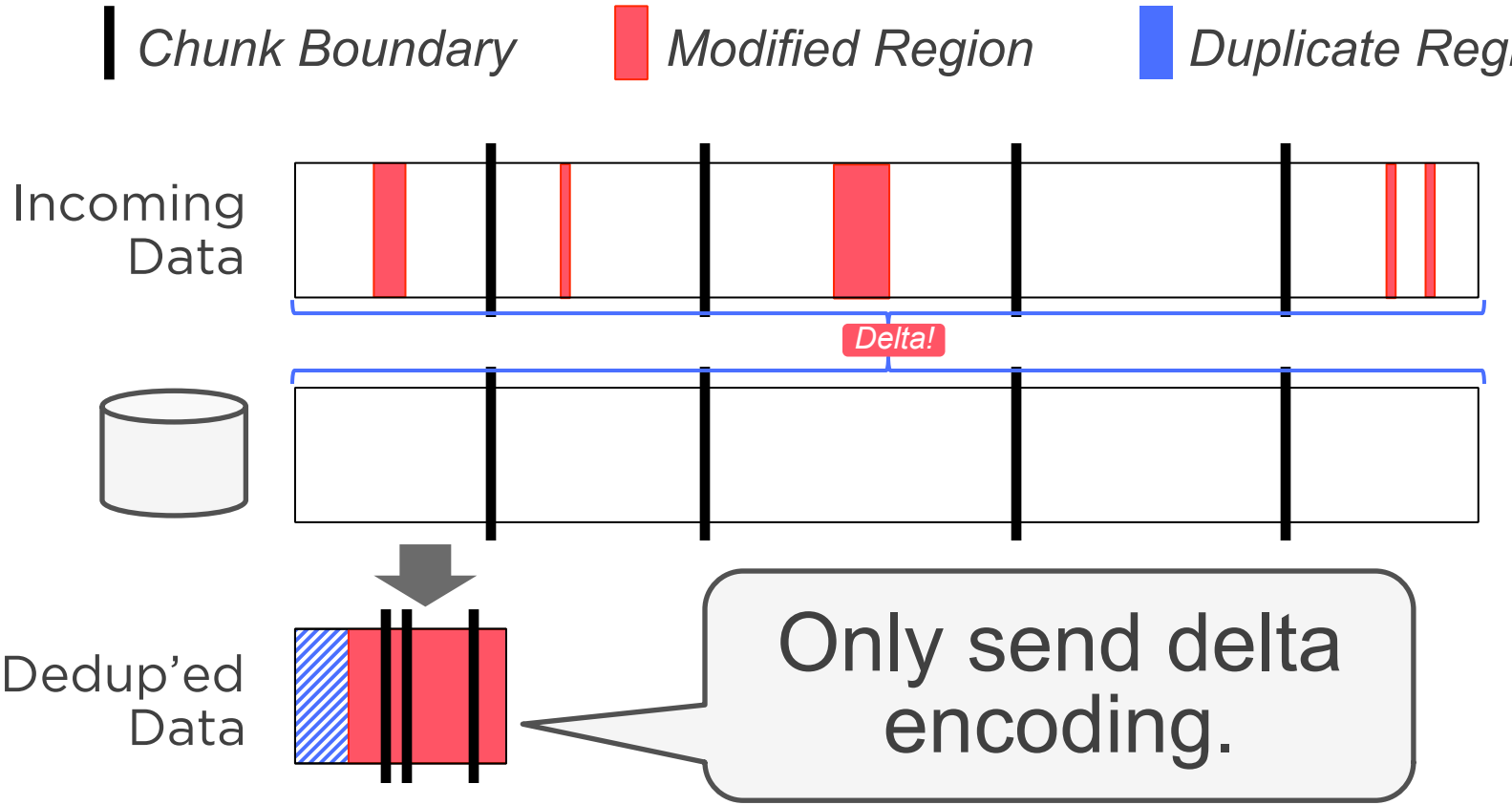
Traditional Dedup: Ideal



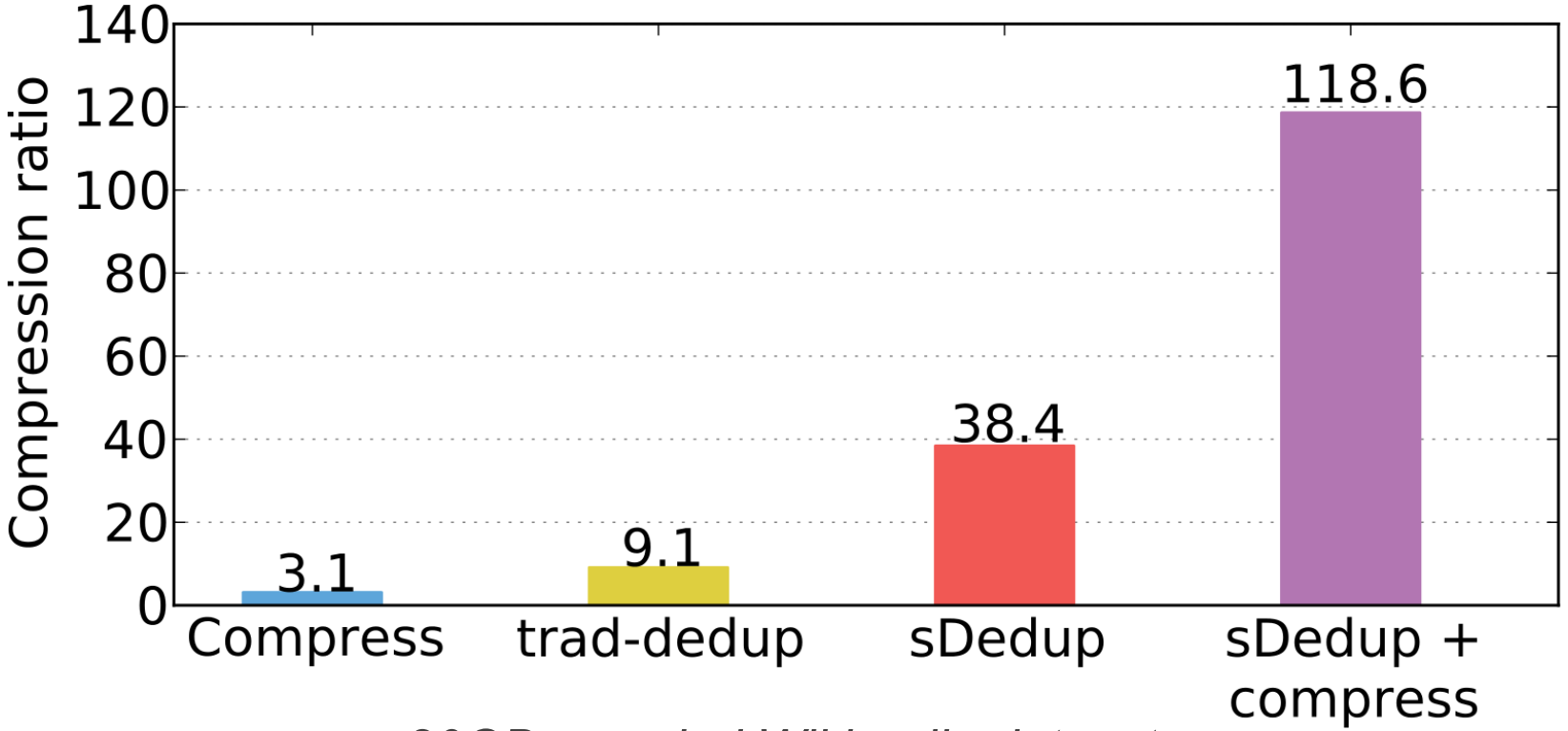
Traditional Dedup: Reality



Similarity Dedup

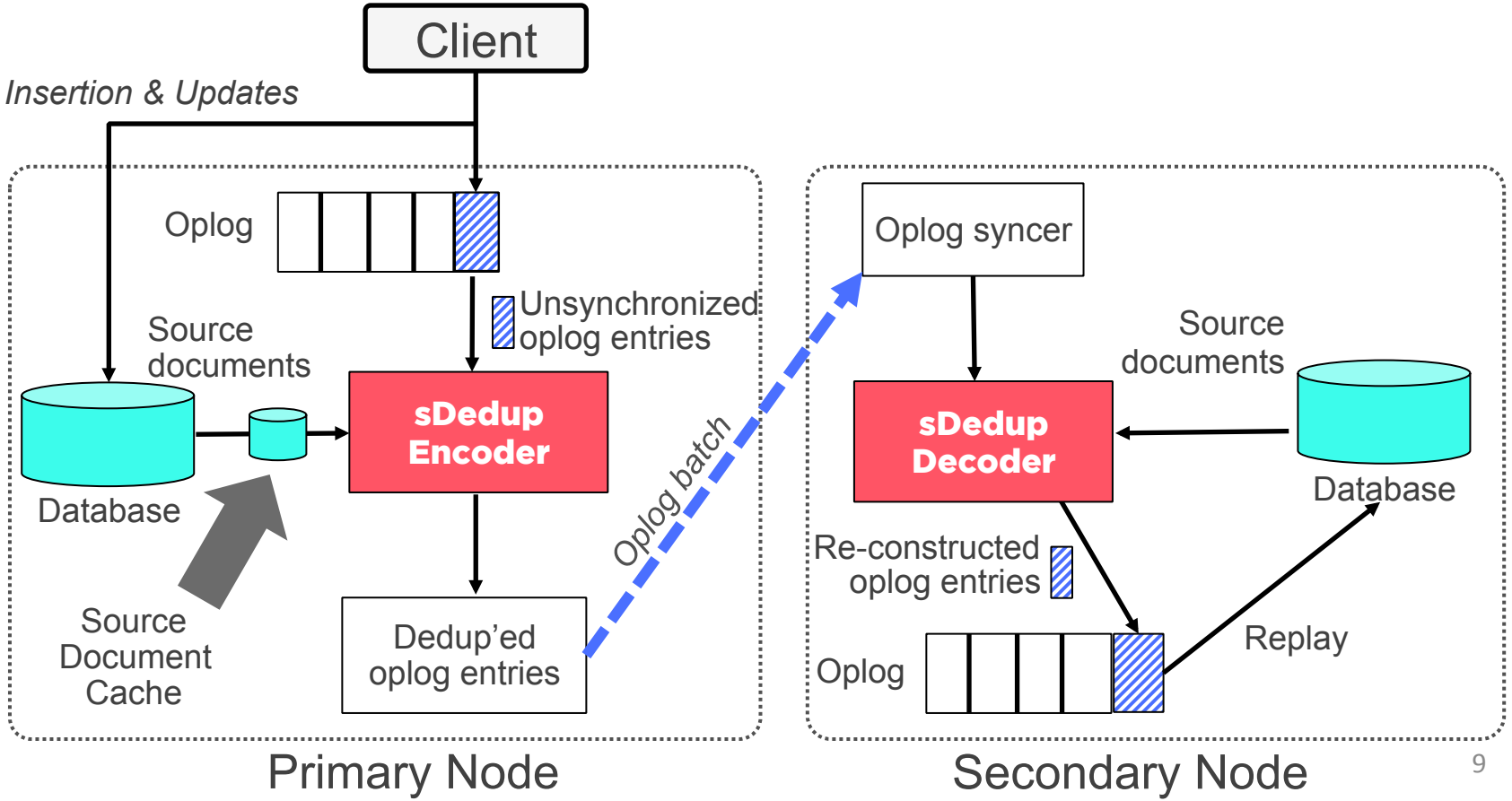


Compress vs. Dedup



20GB sampled Wikipedia dataset
MongoDB v2.7 // 4MB Oplog batches

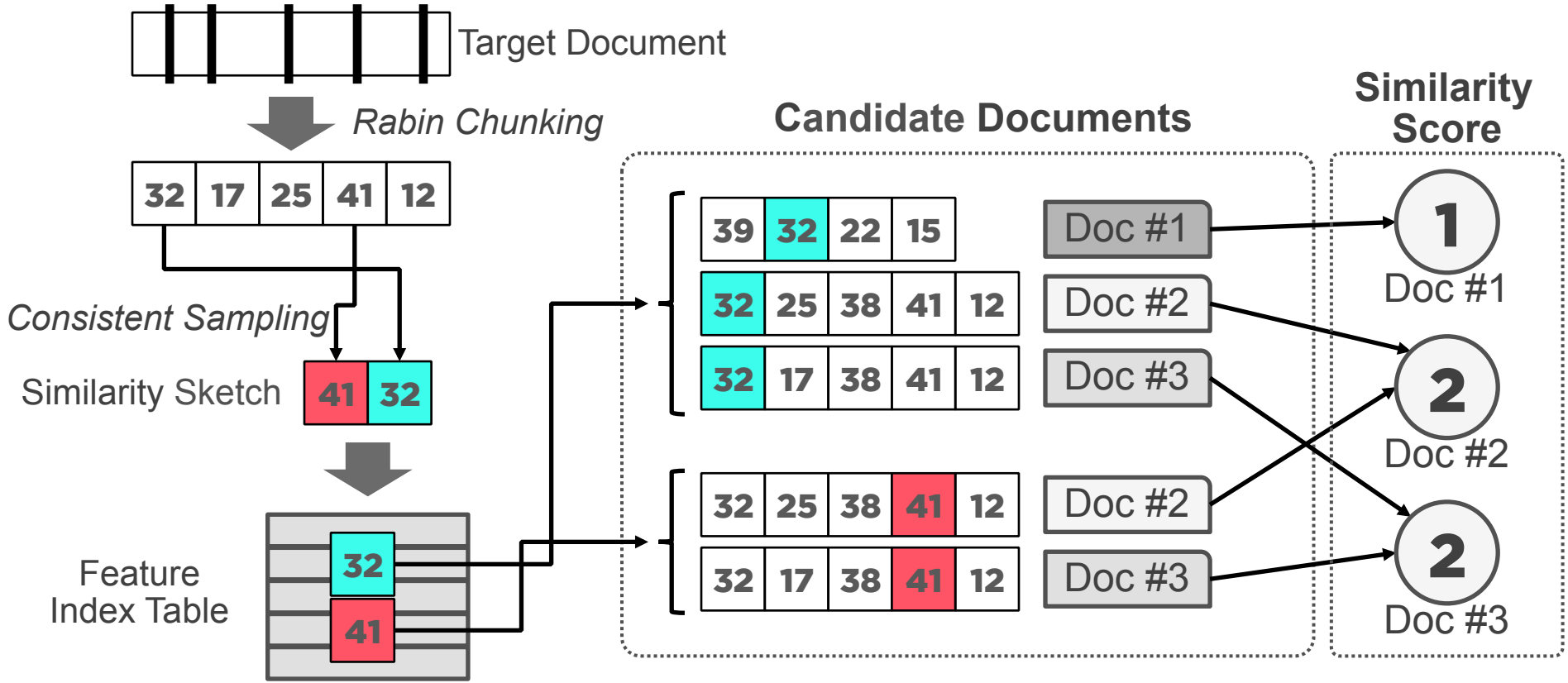
sDedup: Similarity Dedup



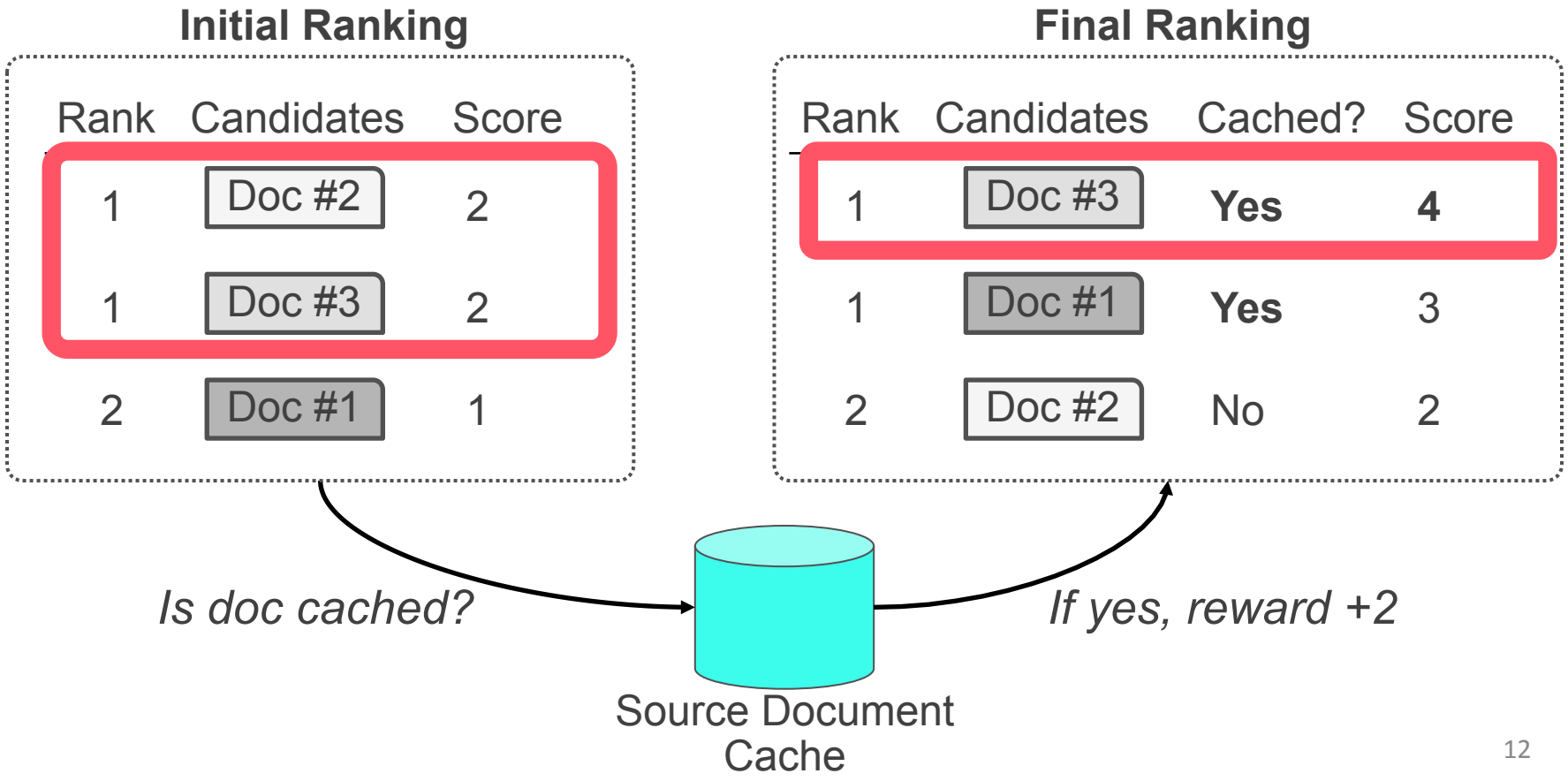
sDedup Encoding Steps

- Identify Similar Documents
- Select the Best Match
- Delta Compression

Identify Similar Documents



Select the Best Match

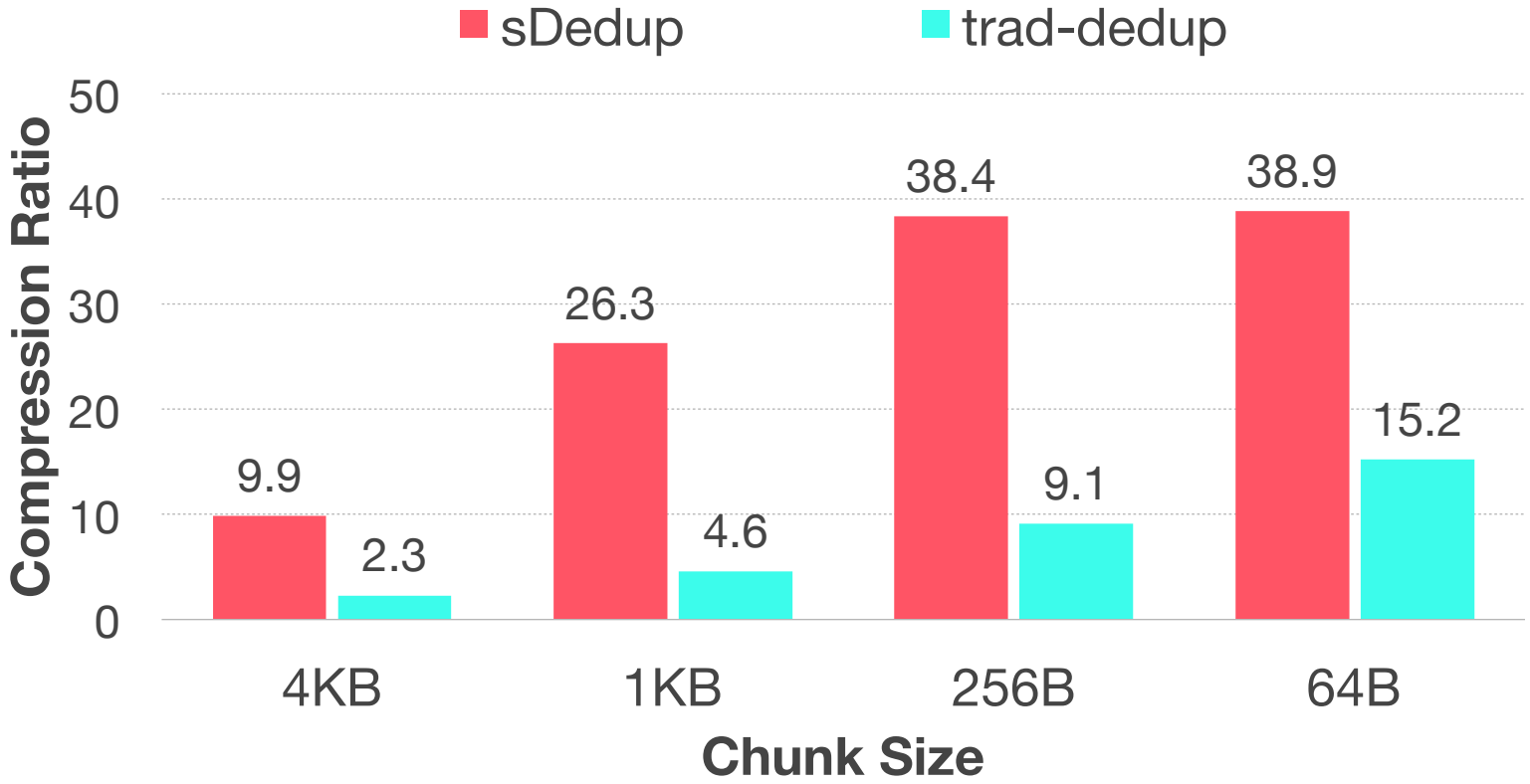


Evaluation

- MongoDB setup (v2.7)
  - *1 primary, 1 secondary node, 1 client*
  - *Node Config: 4 cores, 8GB RAM, 100GB HDD storage*

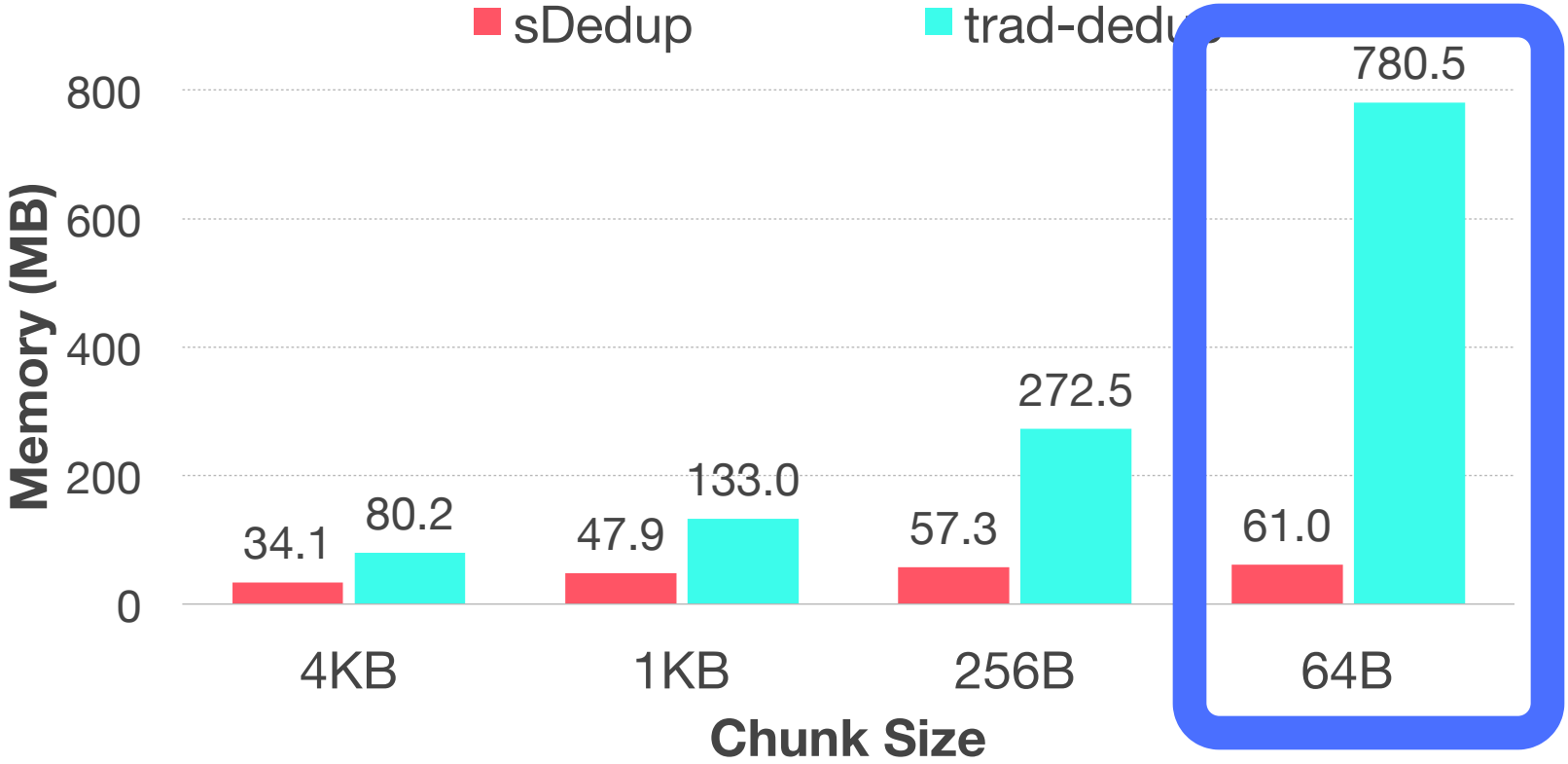
- Datasets:
  - *Wikipedia dump (20GB out of ~12TB)*
  - *Additional datasets evaluated in the paper*

Compression



20GB sampled Wikipedia dataset

Memory



20GB sampled Wikipedia dataset

Other Results (See Paper)

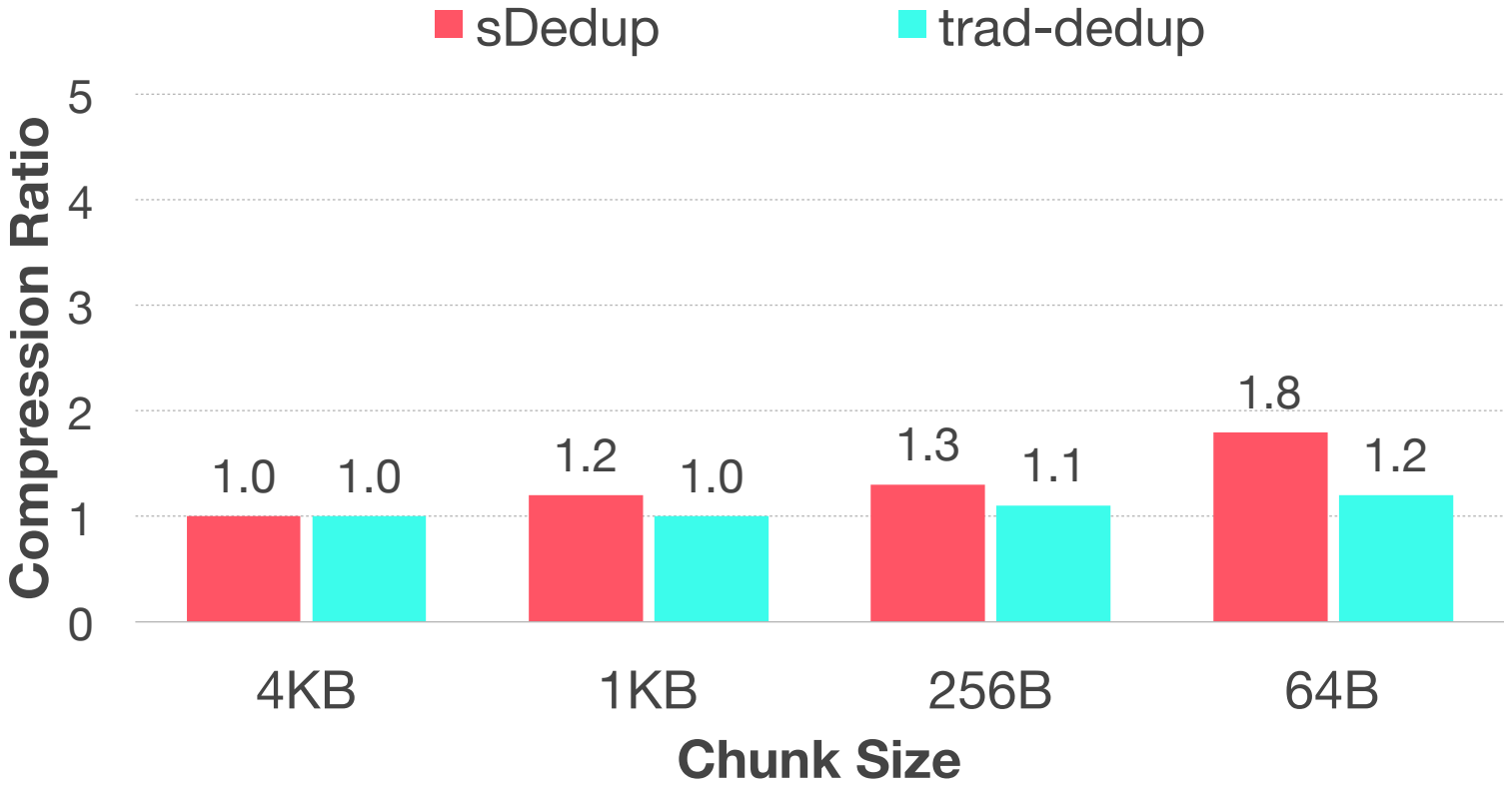
- **Negligible client performance overhead**
- **Failure recovery is quick and easy**
- **Sharding does not hurt compression rate**
- **More datasets**
  - *Microsoft Exchange, Stack Exchange*

Conclusion & Future Work

- **sDedup**: Similarity-based deduplication for replicated document databases.
  - *Much greater data reduction than traditional dedup*
  - *Up to 38x compression ratio for Wikipedia*
  - *Resource-efficient design with negligible overhead*
- **Future work**
  - *More diverse datasets*
  - *Dedup for local database storage*
  - *Different similarity search schemes (e.g., super-fingerprints)*

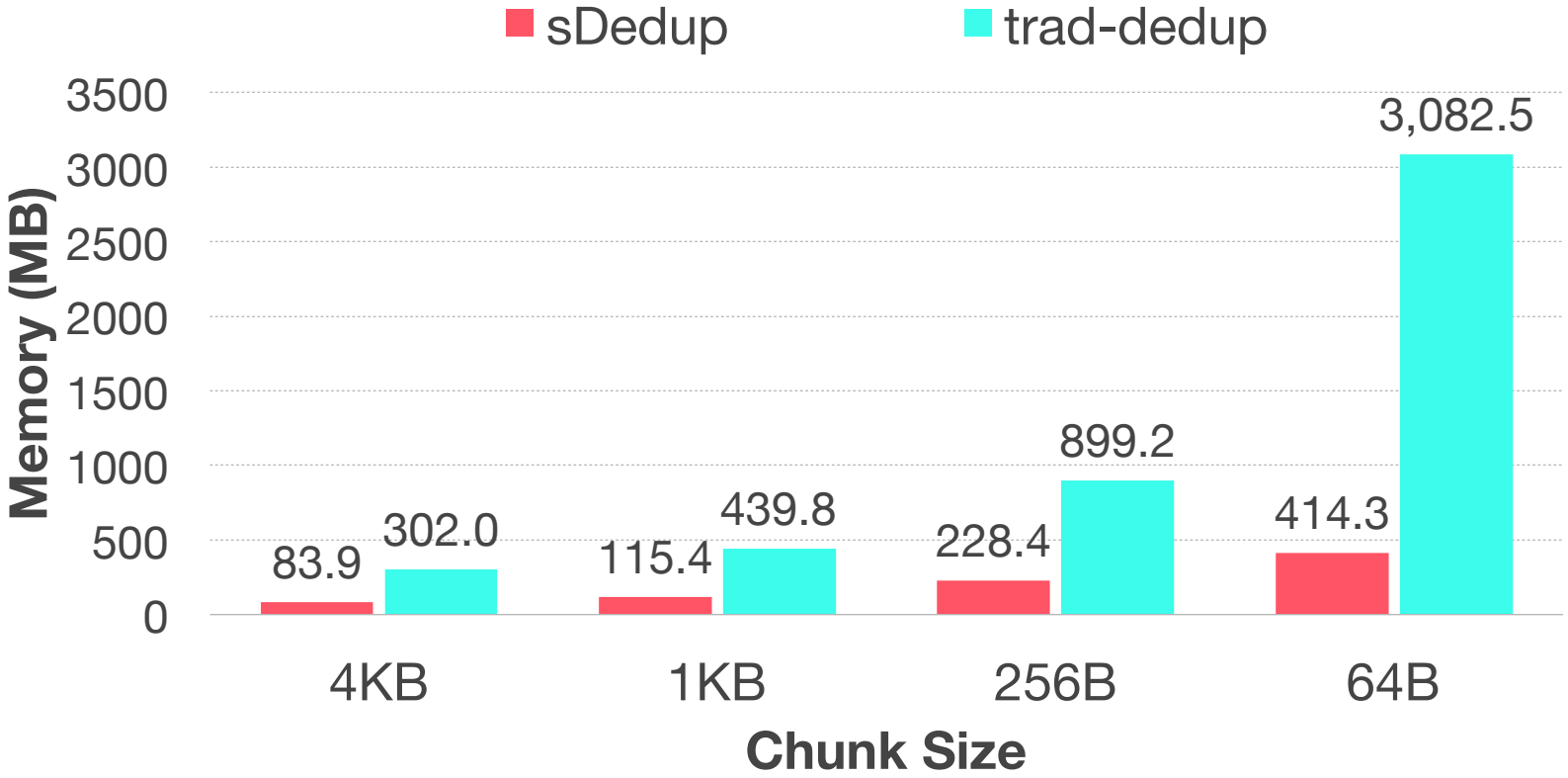
Backup Slides

Compression: StackExchange



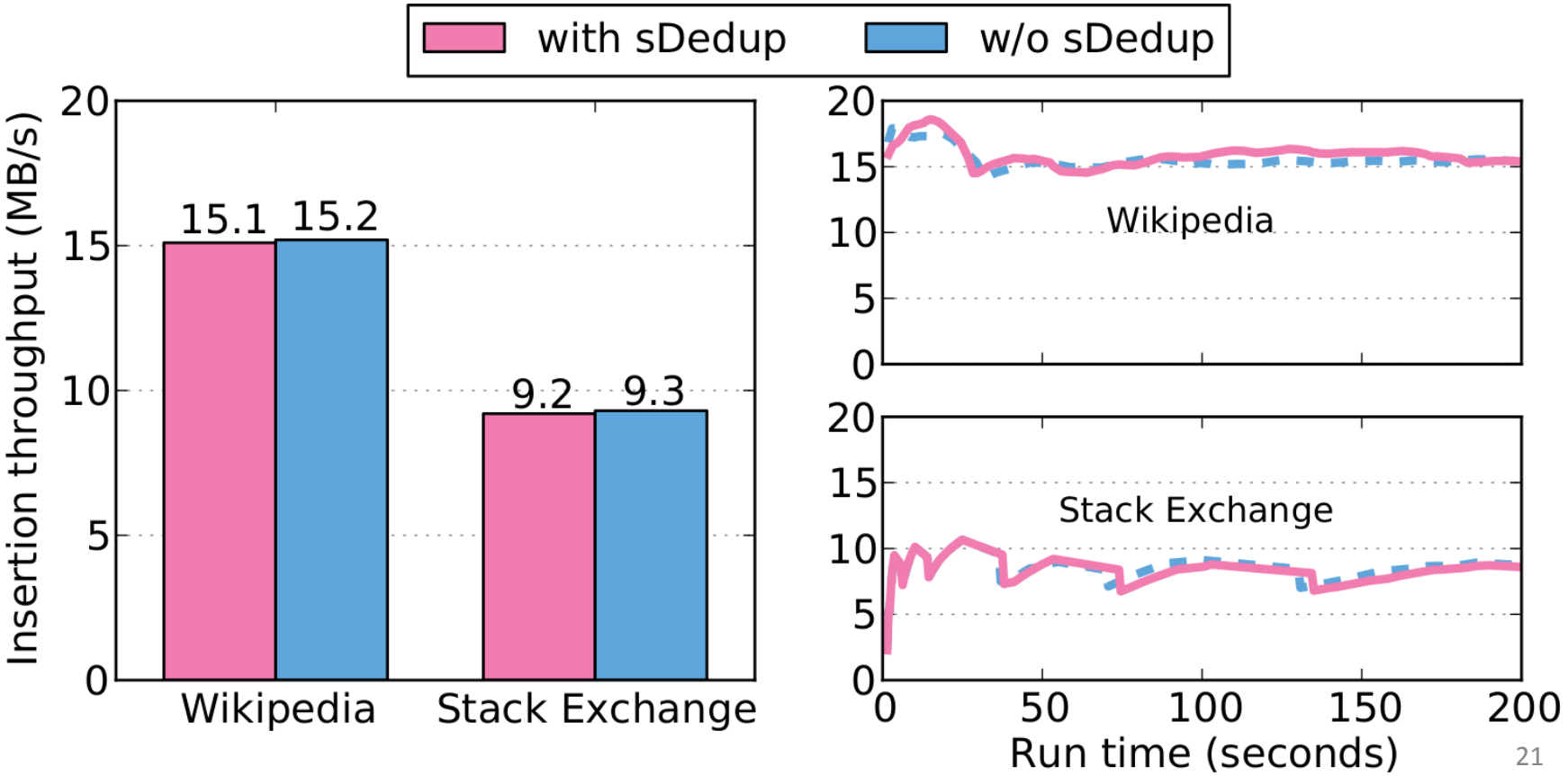
10GB sampled StackExchange dataset

Memory: StackExchange

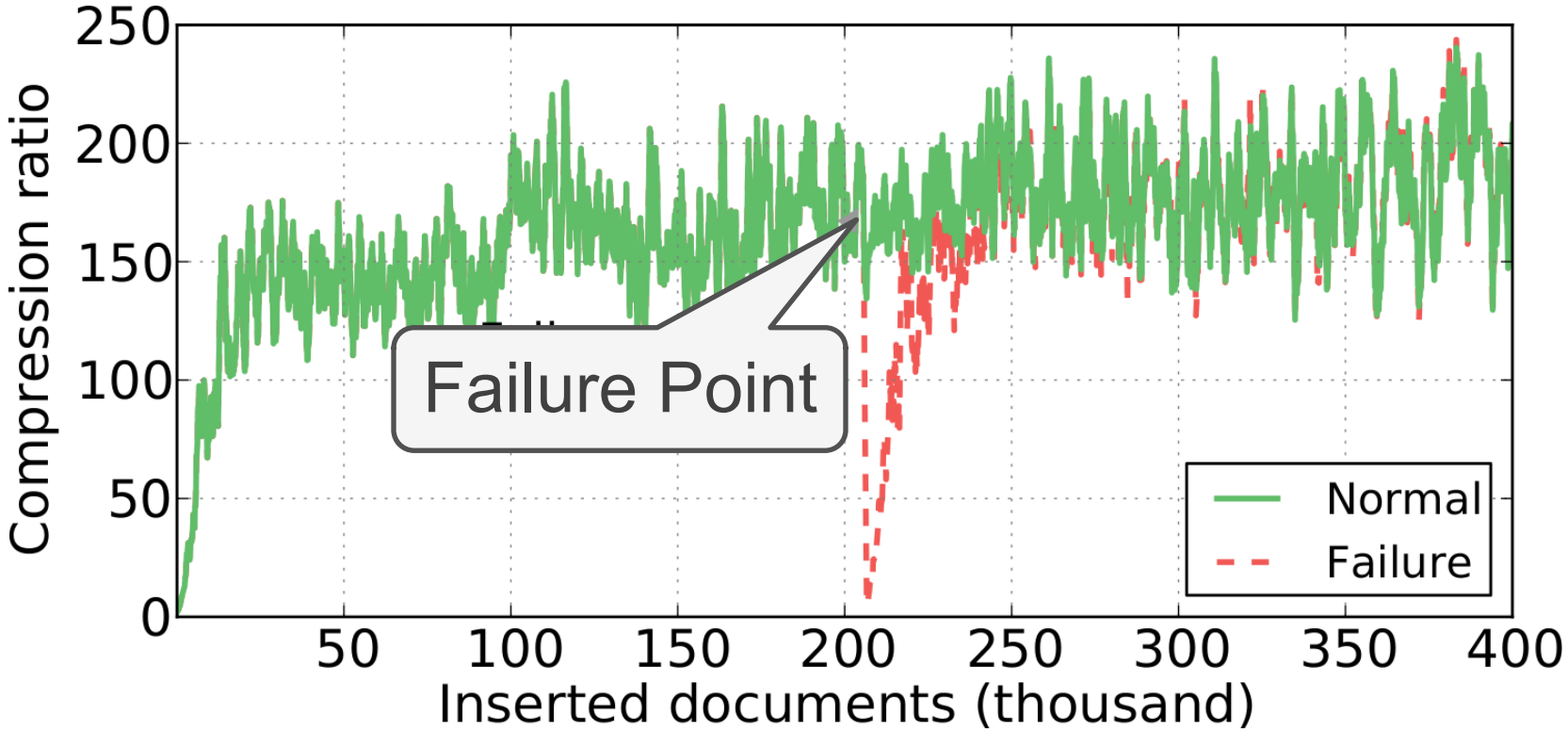


10GB sampled StackExchange dataset

Throughput Overhead

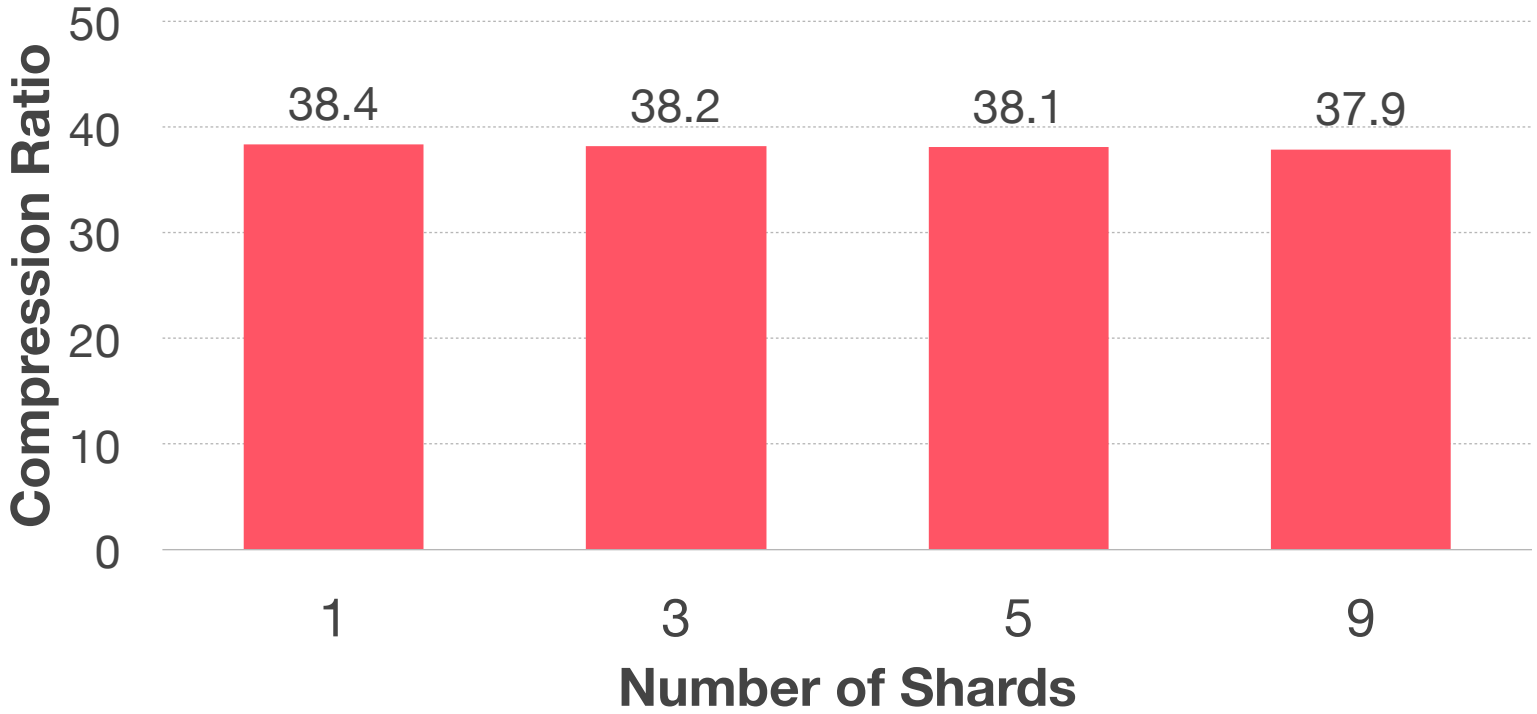


Failure Recovery



20GB sampled Wikipedia dataset.

Dedup + Sharding



20GB sampled Wikipedia dataset

Delta Compression

- Byte-level diff between source and target docs:
  - *Based on the xDelta algorithm*
  - *Improved speed with minimal loss of compression*

- **Encoding**:
  - *Descriptors about duplicate/unique regions + unique bytes*
- **Decoding**:
  - *Use source doc + encoded output*
  - *Concatenate byte regions in order*