Separating Data via Block Invalidation Time Inference for Write Amplification Reduction in Log-Structured Storage

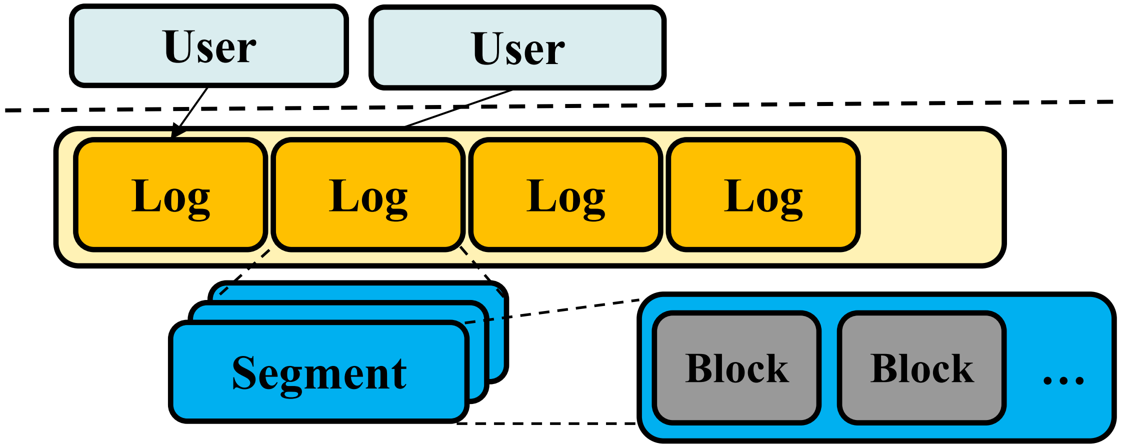
Qiuping Wang[1,2], Jinhong Li[1], Patrick P. C. Lee[1],
Tao Ouyang[2], Chao Shi[2], Lilong Huang[2]

[1]The Chinese University of Hong Kong (CUHK)
[2]Alibaba Group

Log-Structured Storage

- Alibaba Cloud ESSDs
  - Log-structured block storage atop Alibaba Cloud Pangu
  - Backed by flash-based storage, with ~100μs latency and up to 1M IOPS

- Abstraction
  - Each ESSD is a block-level **volume** as an append-only log
  - Each log contains **segments** (hundreds of MiB) composed of **blocks** (several KiB), each identified by logical block addresses (**LBAs**)



Garbage Collection

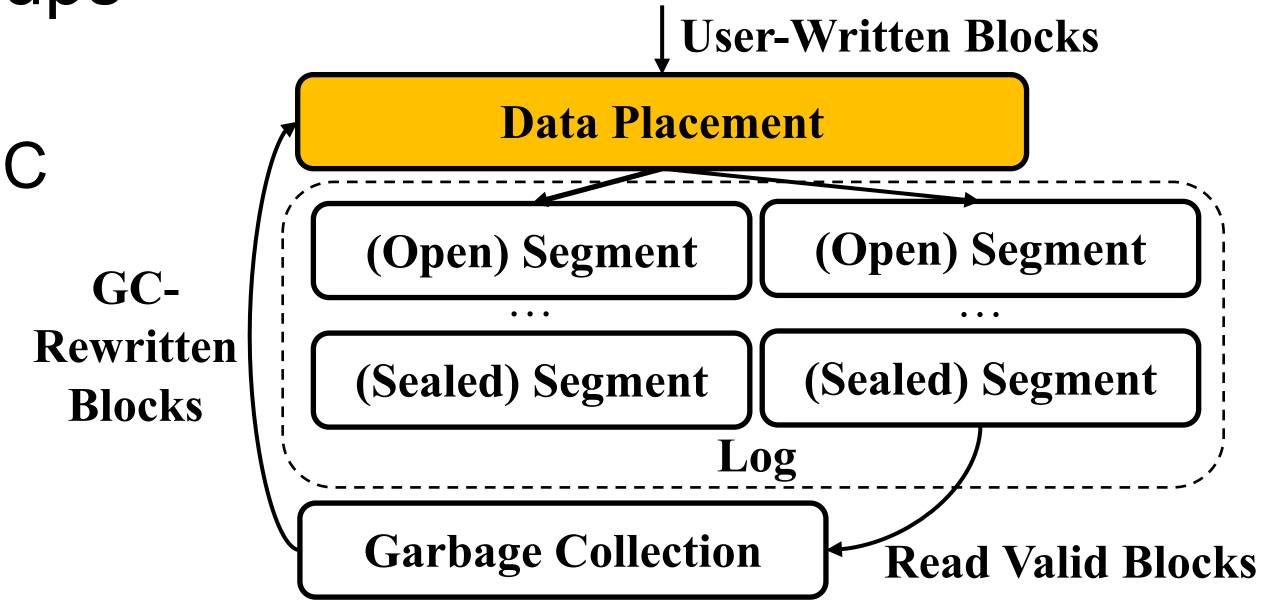
- Space reclamation of invalid (stale) blocks
  - **Triggering**: garbage proportion (GP) over a threshold, e.g., 15%
    - Garbage proportion = #invalid blocks / (#invalid blocks + #valid blocks)
  - **Selection**: select segments according to some algorithm, e.g., Greedy
  - **Rewriting**: rewrite valid blocks, delete old segments and reuse space

- **Write amplification (WA)**: repeated rewrites of valid blocks
  - Reduced flash lifespan and bandwidth at Alibaba Cloud ESSDs

Data Placement

- Each write/update to an LBA gives one **user-written block** and zero or multiple **GC-rewritten blocks**
  - WA = 1 + #GC-rewritten blocks / #user-written blocks

- Goal: separates blocks into groups by properties
  - Produce high-GP segments for GC



Contribution

- **SepBIT**: Separates blocks via **block invalidation time (BIT)** inference [He, EuroSys'17]
  - Effective BIT inference based on mathematical and trace analyses
    - Separates each set of user-written blocks and GC-rewritten blocks by BIT inference
  - Deployed at Alibaba Cloud ESSDs
- Extensive trace analysis and prototype evaluation to validate SepBIT effectiveness

Ideal Data Placement and Limitations

- ➢ Ideal data placement can achieve the minimum WA of 1
  - Strictly places all written blocks based on their invalidation orders
  - Selects segments whose blocks are earliest invalidated for GC
    - No rewrites of valid blocks
- ➢ Impractical assumptions:
  - Future knowledge: the invalidation order (time) of each written block
  - Space reservation: memory/storage reservation for all written blocks
- ➢ Practical solution:
  - Infers accurately the BIT for each written block
  - Groups blocks of close BITs

Observations

- Derive three observations based on **lifespan** analysis
  - Lifespan: number of bytes written from when a block is written until it is invalidated (or until the end of the trace)
- Focus on Alibaba Cloud block traces:
  - 186 out of 1000 volumes over one full month in January 2020
    - Large write working set size (WSS) (# unique LBAs written x block size): 10GiB - 1TiB
    - Large write traffic size: 43GiB - 36.2TiB.
- Also validated on Tencent Cloud block traces

Observations

- O1: User-written blocks generally have short lifespans
  - e.g., 10% of write WSS
  - GC-rewritten blocks have long lifespans
- O2: Frequently updated blocks have highly varying lifespans
  - Large variations in different groups of frequently updated blocks
- O3: Rarely updated blocks dominate with highly varying lifespans
  - Spanning both long and short ranges
- *Temperature-based placement (e.g., via access frequencies) are ineffective in BIT inference*

SepBIT Design

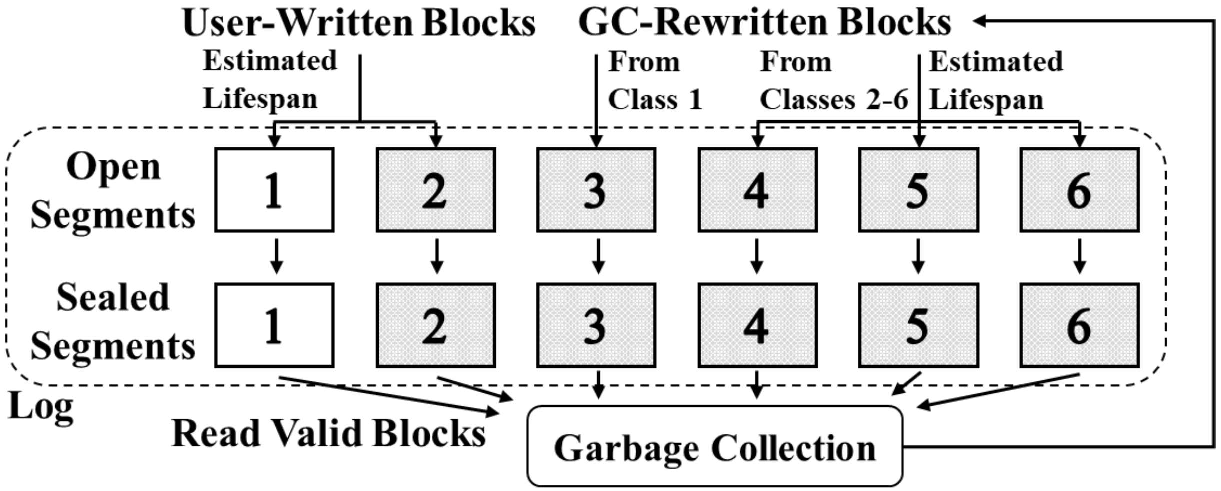
➢ User-written blocks

- **Short-lived** blocks (Class 1) written near the same time have similar BITs
- Remaining **long-lived blocks** (Class 2) span large BIT ranges

➢ GC-rewritten blocks

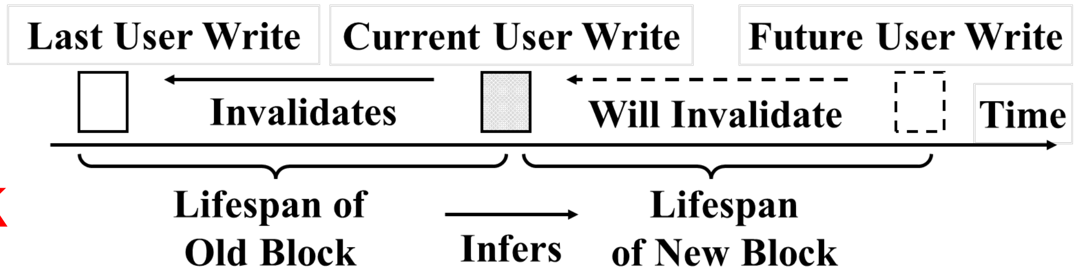
- Short-lived blocks (Class 3) identified in user-written blocks
- Blocks with similar BITs inferred are grouped to Classes 4-6

Note: WA reduction of SepBIT is less sensitive to # of classes



User-Written Block Separation

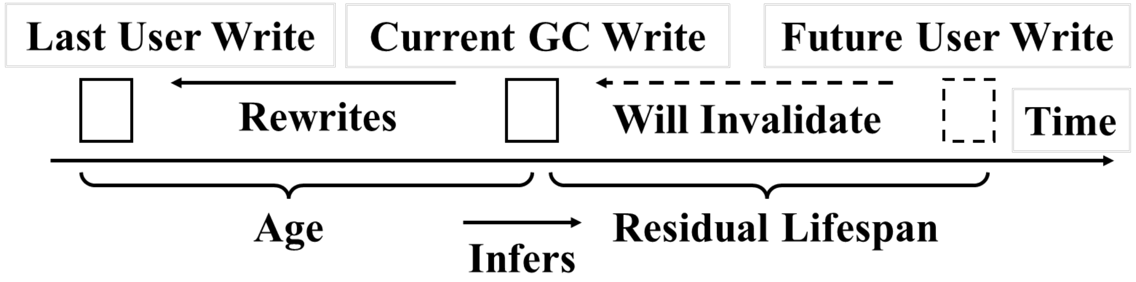
- Intuition: Any user-written block that invalidates a short-lived block is also likely to be a short-lived block



- Analysis
  - Probability analysis: $\Pr(u \leq u_0 | v \leq v_0)$
  - u, v: lifespans of user-written block and invalidated block, respectively
- High conditional probabilities in Alibaba Cloud block traces
  - e.g., 77.8-90.9% for v_0= 40% and u_0 in [2.5%,40%] of write WSS

GC-rewritten Block Separation

- Intuition: Any GC-rewritten block with a smaller age is likely to have a short residual lifespan



- Analysis
  - Probability analysis: $\Pr(u \leq g_0 + r_0 | u \geq g_0)$
  - g_0, r_0: thresholds of age and residual lifespans, respectively
- Conditional probabilities drop significantly when g_0 increases in Alibaba Cloud block traces
  - From 90% to 14.5% when g_0 increases from 0.8X to 6.4X write WSS

SepBIT Implementation

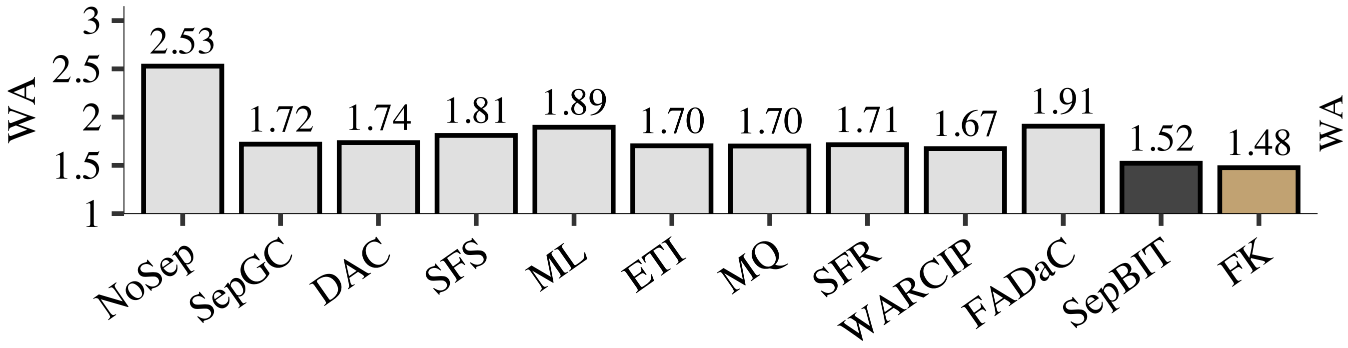
- ℓ: average segment lifespan of collected segments in Class 1
  - Segment lifespan: #bytes written since first append until collection
  - Compute ℓ for each fixed number (e.g., 16) of collected segments
- Threshold selection
  - Classes 1 and 2: Use ℓ as lifespan threshold for user-written blocks based on the lifespans of their invalidated blocks
    - Track (in memory) blocks whose ages are less than ℓ
  - Classes 4-6: Use 4ℓ and 16ℓ as age thresholds for GC-rewritten blocks according to their ages
    - Maintain age information for each block in flash page spare space

Evaluation

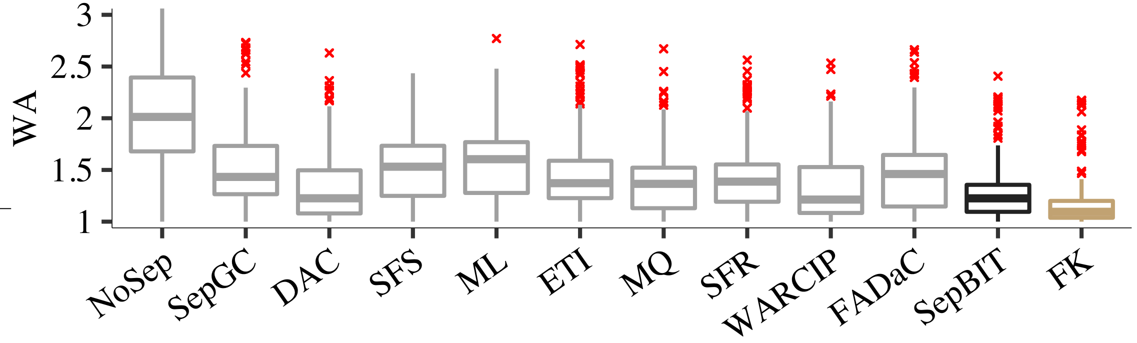
- Trace analysis for per-volume WA
  - 186 volumes from Alibaba block traces
    - Varying segment selection algorithms: Greedy and Cost-Benefit
    - Varying segment sizes and GP thresholds for GC
  - Schemes: 12 schemes, including 8 state-of-the-art placement schemes and FK (oracle scheme)
- Prototype evaluation for throughput
  - Build C++ prototype implementing SepBIT
  - Use emulated zoned storage devices using Optane PM
    - For reproducibility and best match with production storage at Alibaba

Trace Analysis on WA

- SepBIT reduces overall WA of existing schemes by **9.1-20.2%**
  - Only 3.1% higher overall WA than Future Knowledge (FK)
- SepBIT has lowest 75th percentile in per-volume WA



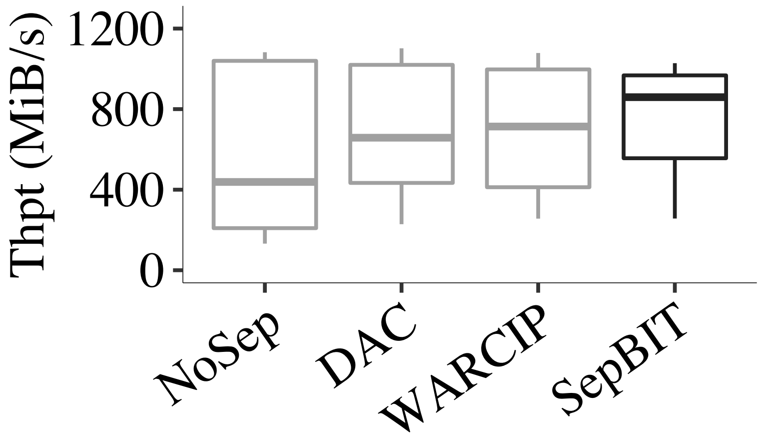
Overall WA of Cost-Benefit



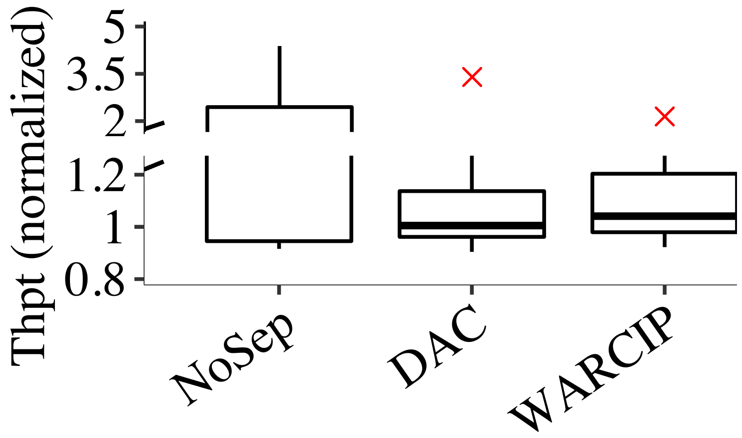
Per-volume WA of Cost-Benefit

Prototype Throughput

- Throughput on 20 write-heavy volumes
- SepBIT achieves the highest throughput
  - 25th and 50th percentiles: **28.3% and 20.4%** higher than the second best, respectively



Absolute throughput



Normalized throughput

Conclusion

- **SepBIT** is a novel data placement scheme that mitigates WA in log-structured storage via BIT inference
  - Infers BIT patterns based on trace analysis
  - Deployed at Alibaba Cloud ESSDs

- See paper, technical report, and source code for more details

- Source code: http://adslab.cse.cuhk.edu.hk/software/sepbit

Thank You!
Q & A