A New LSM-style Garbage Collection Scheme for ZNS SSDs

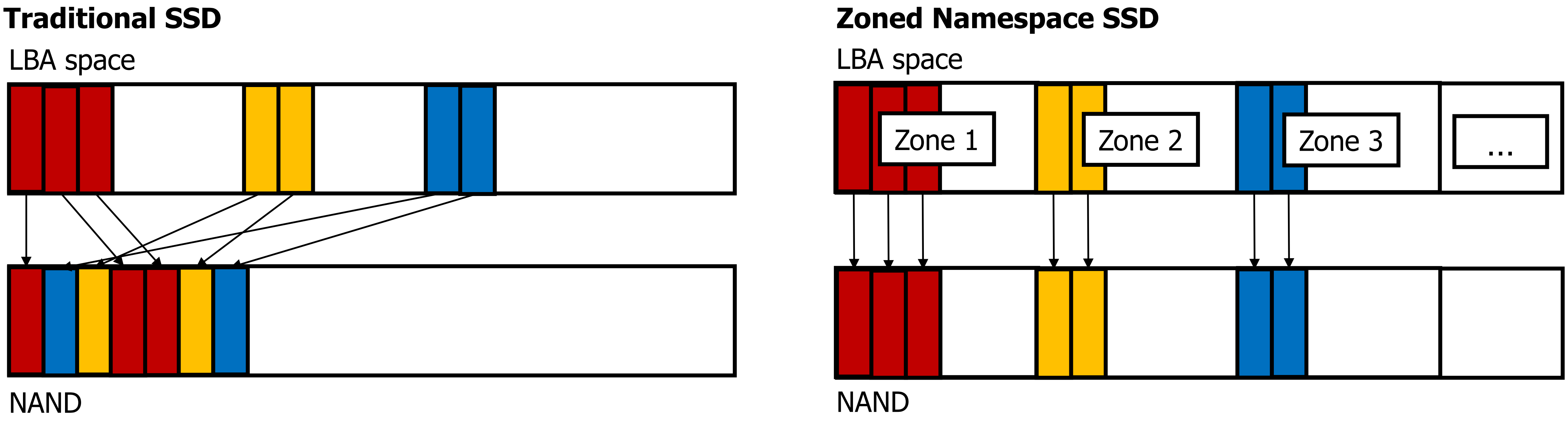
12th USENIX Workshop on Hot Topic in Storage and File System (HotStorage 20), 2020

2022.12.15

张茂荣
华中科技大学

Content

- **What is ZNS SSD?**

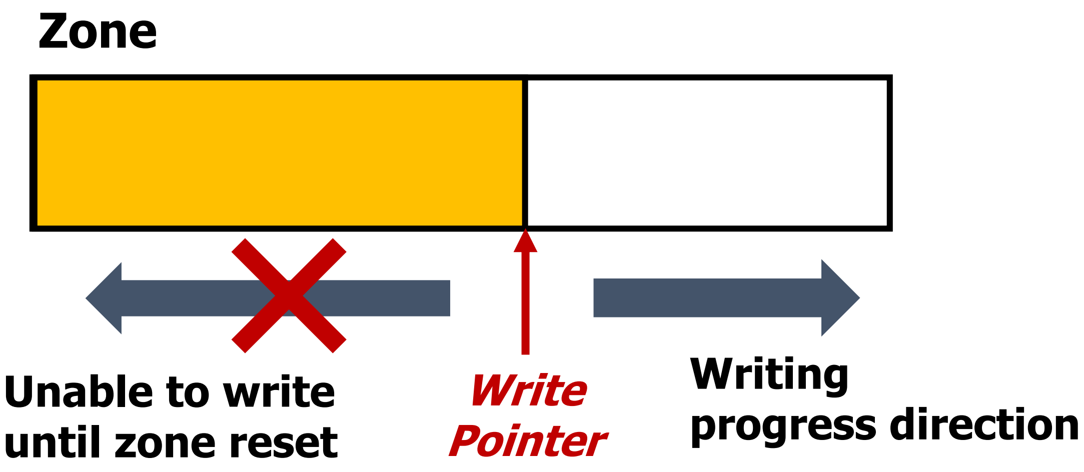


✓ Benefits

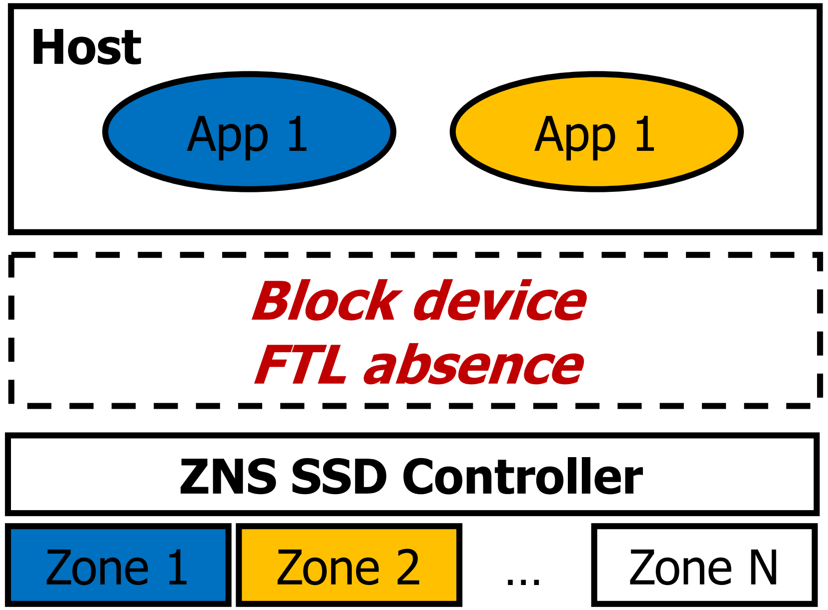
- Better performance and WAF by distributing different workloads into different zones
- Better isolation (IO Determinism)
- Reduce DRAM usage and Over-provisioning area in SSDs

- **What are the issues of ZNS SSD?**

  - ✓ Sequential write constraint: writes need to be conducted in a sequential manner, like the SMR drives.
  - ✓ Host needs to control zones directly such as zone open, close, reset and zone garbage collection.

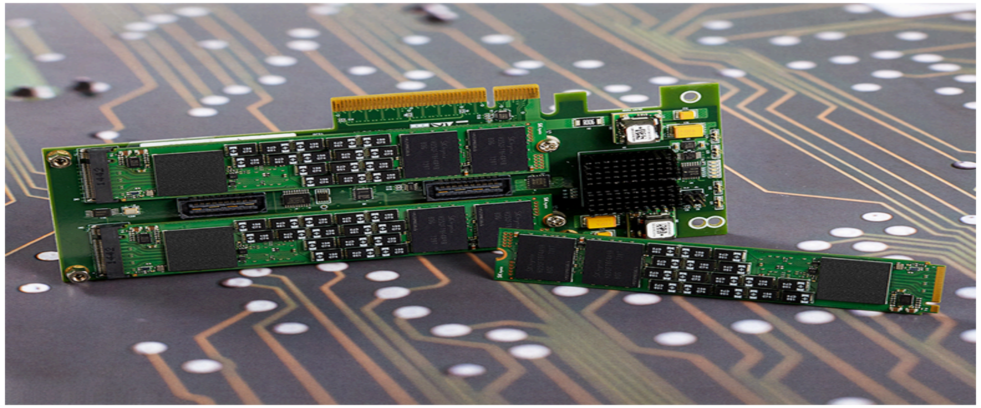


Sequential write constraint



Host Needs to handle zone controls

- **How much is the Zone Garbage Collection (hereafter ZGC) overhead?**
  - ✓ Using real ZNS SSD prototype
  - ✓ Zone size: 1GB (note that the typical segment size in LFS is 2MB)

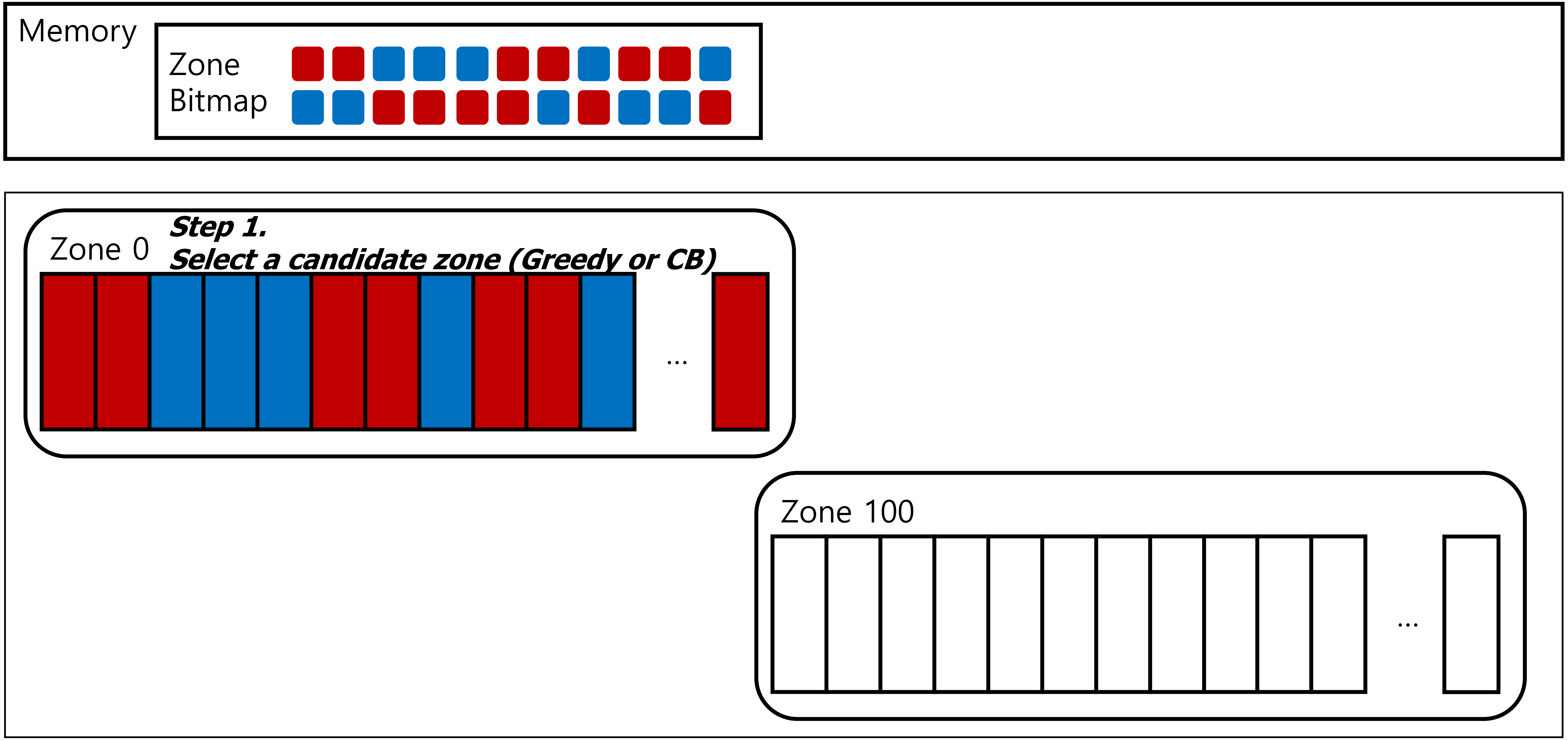


SK Hynix Prototype ZNS SSD

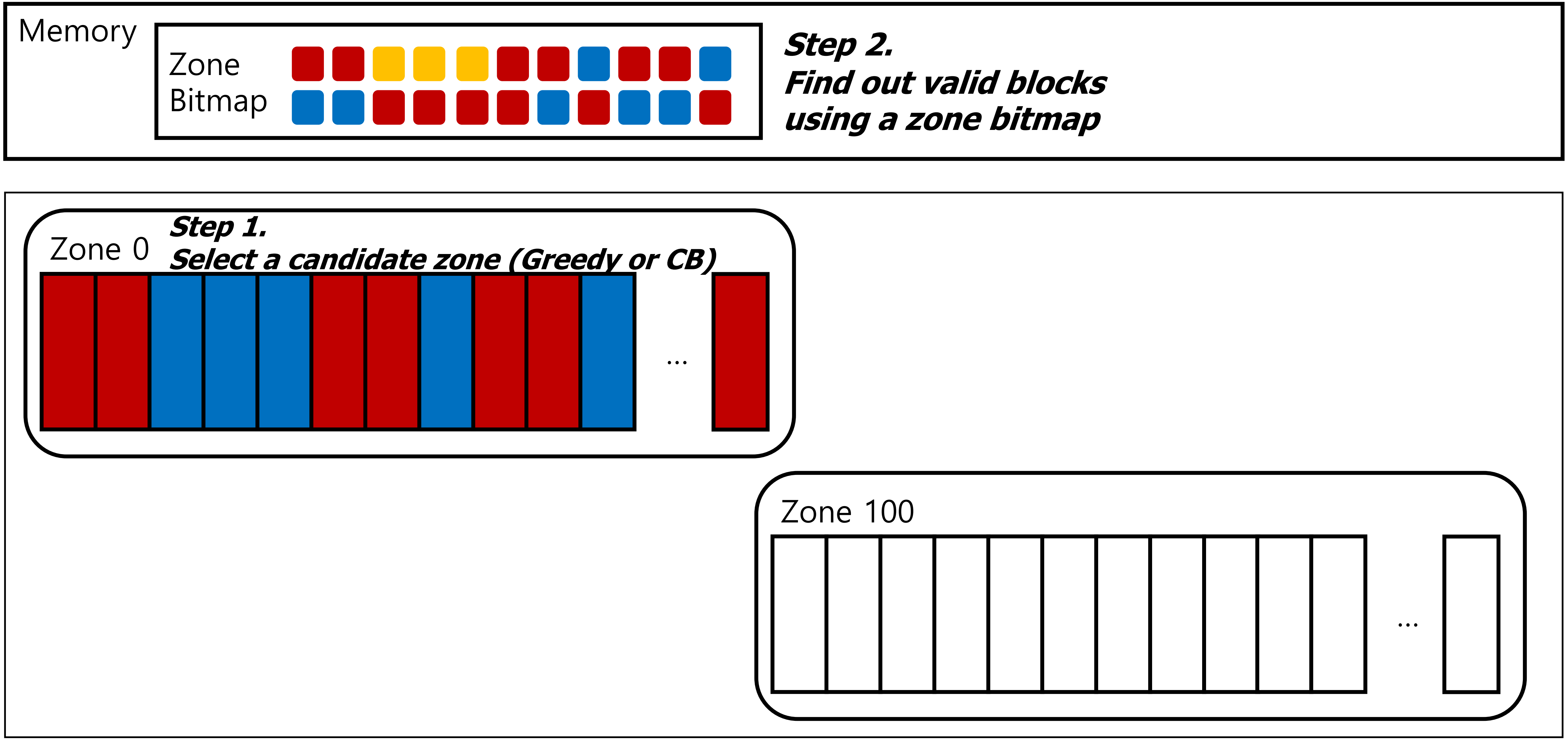
Table 1: ZNS SSD prototype information

Item	Specification
SSD Capacity	1TB
Size of a Zone	1GB
Number of Zones	1024
Interface	PCIe Gen3
Protocol	NVMe 1.2.1

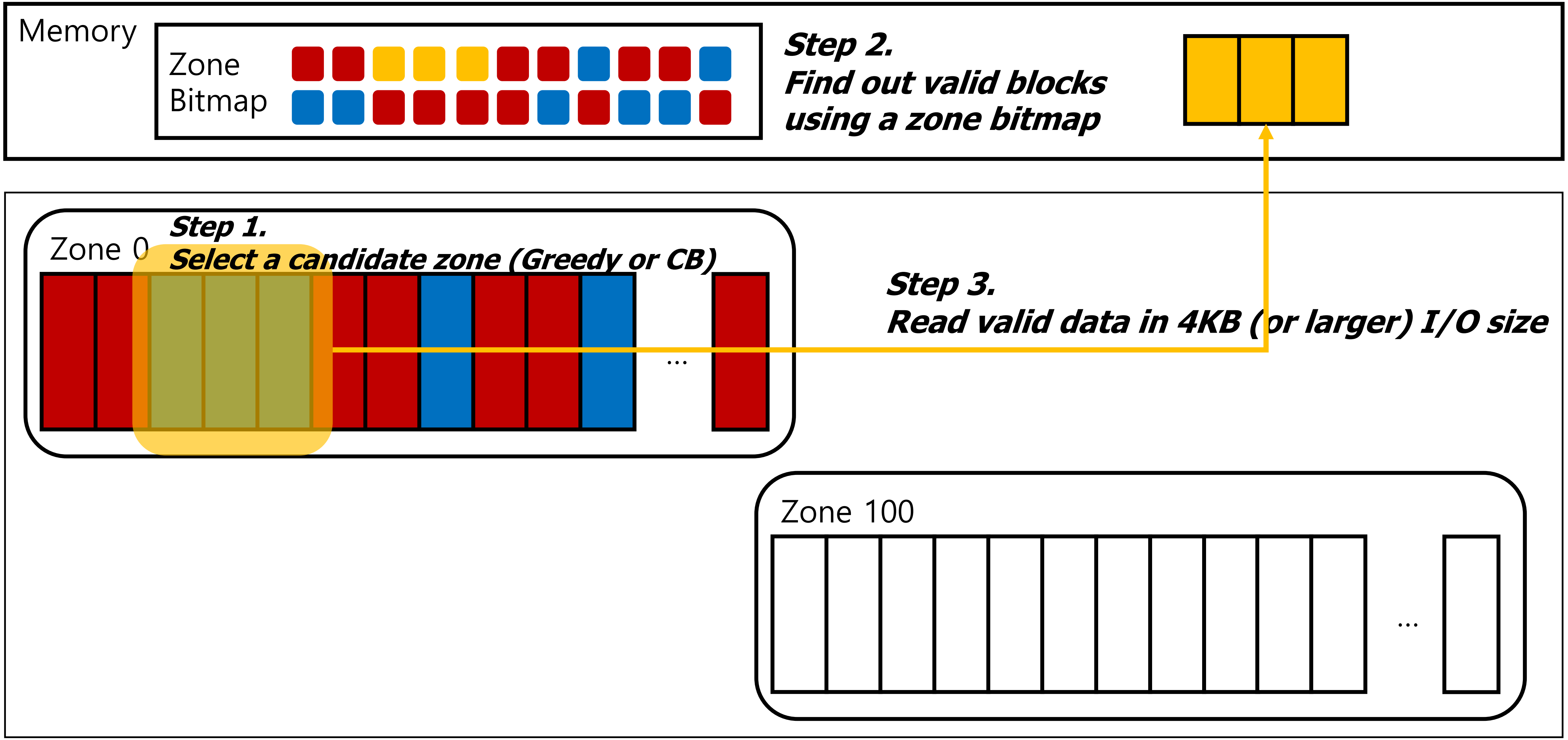
Basic Zone Garbage Collection (Basic_ZGC)



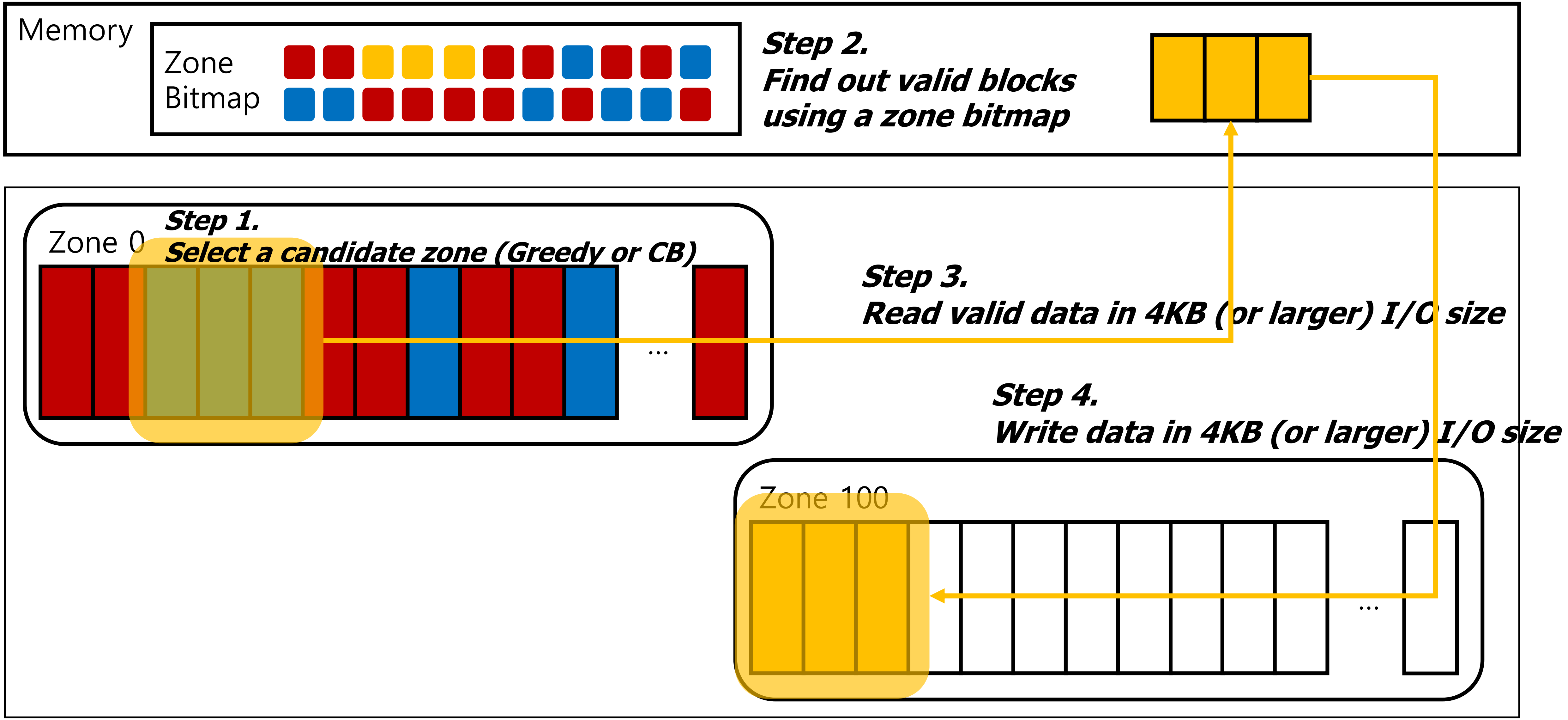
Basic Zone Garbage Collection



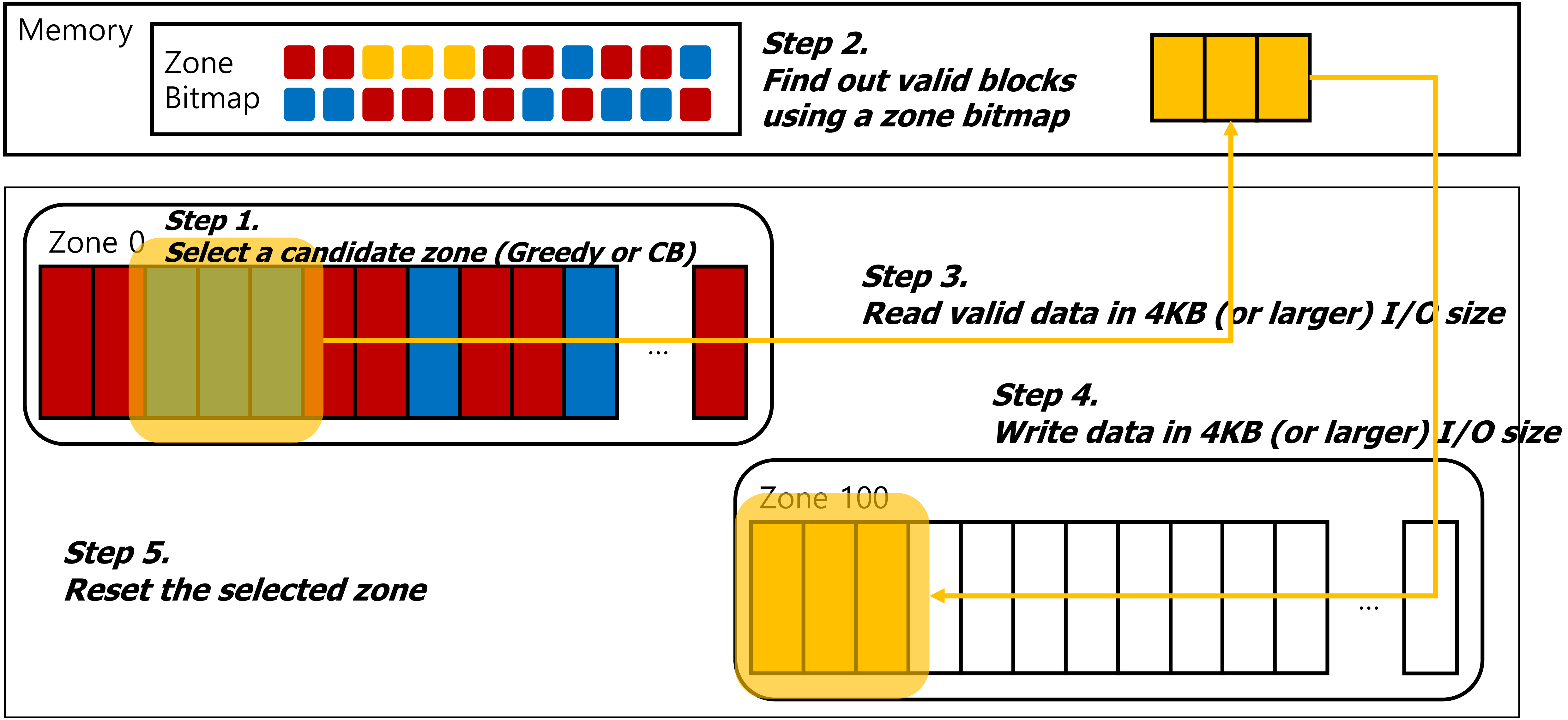
Basic Zone Garbage Collection



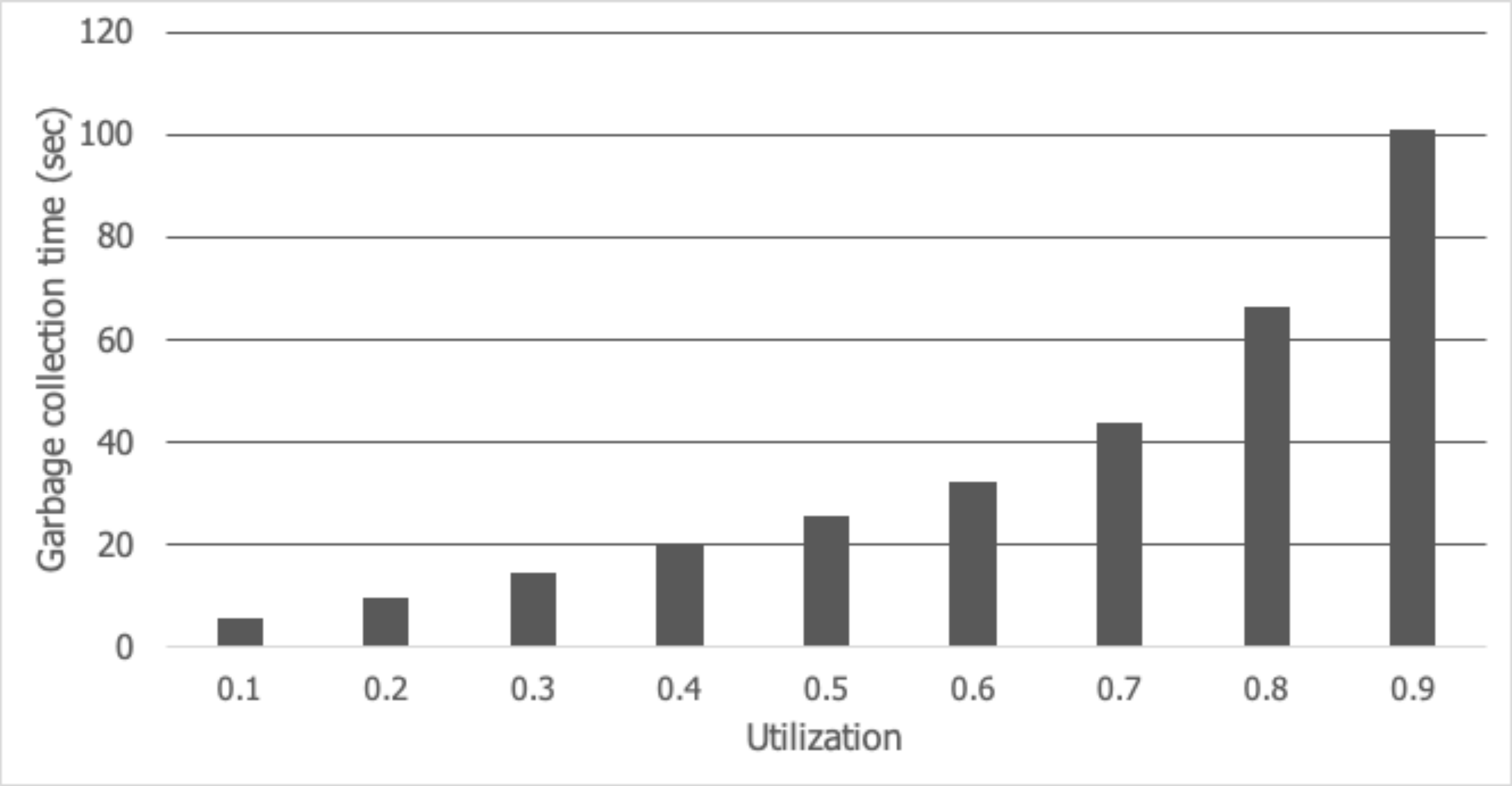
Basic Zone Garbage Collection



Basic Zone Garbage Collection

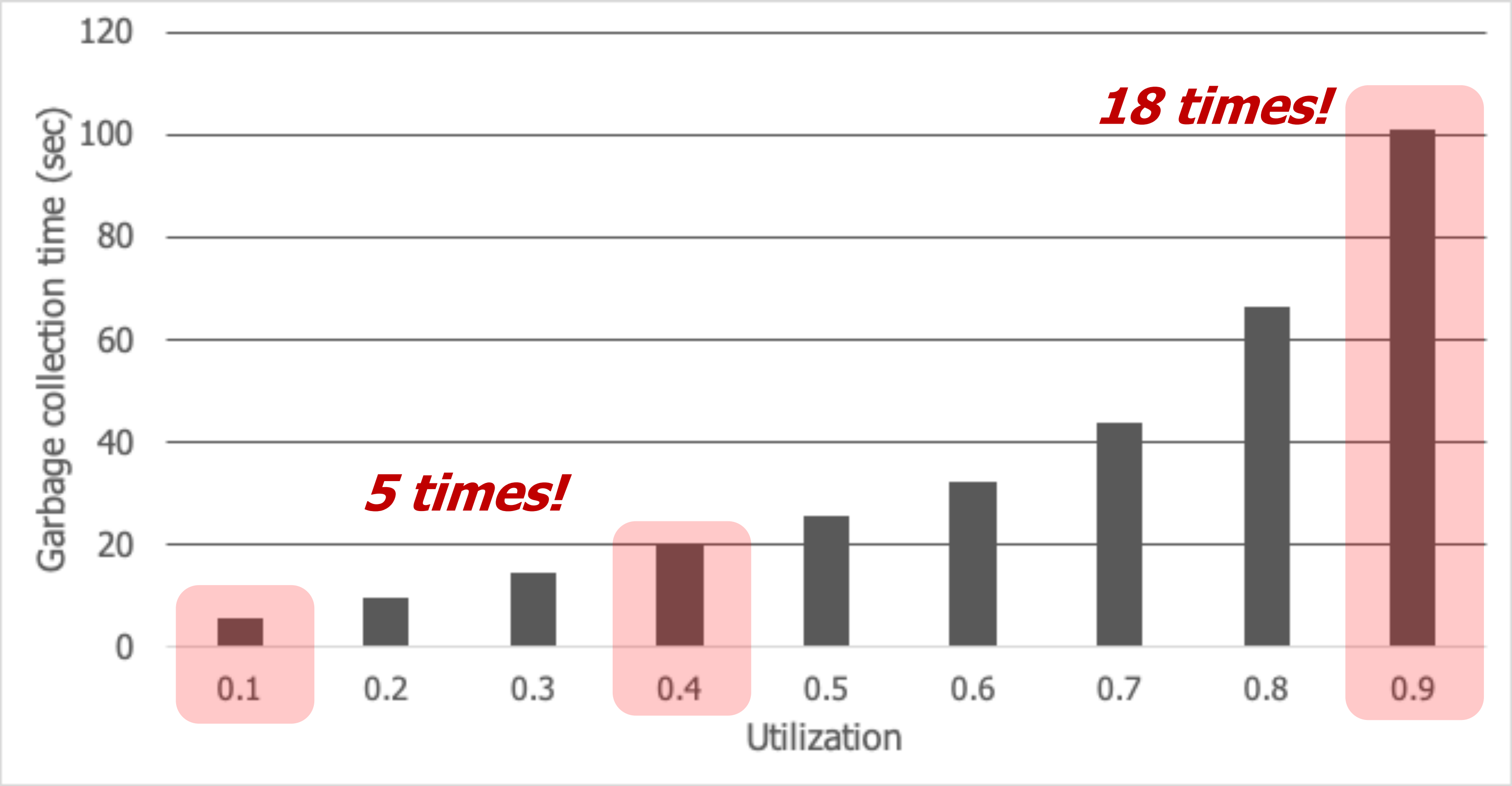


Observation 1: Zone garbage collection overhead



- Zone : 1GB
- Block : 4KB

Observation 1: Zone garbage collection overhead



- Zone : 1GB
- Block : 4KB

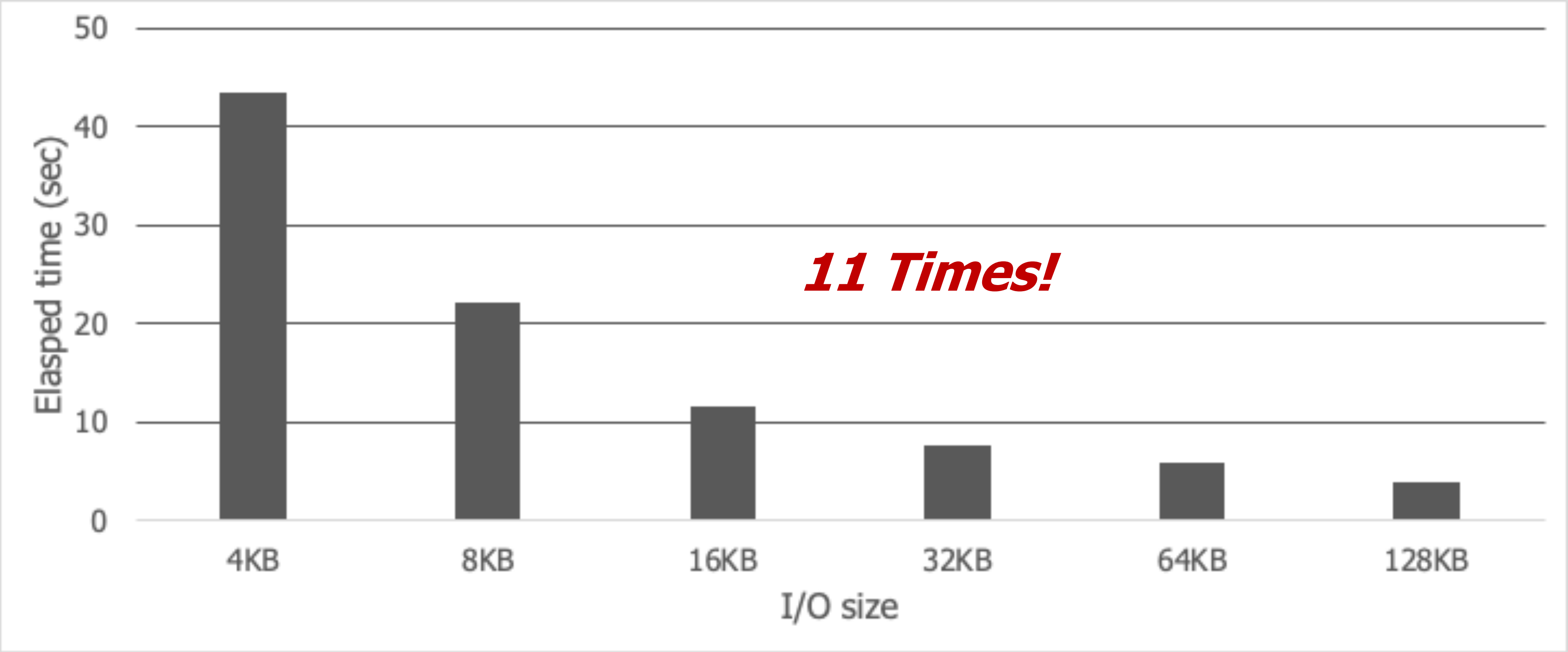
☞ ***Motivation 1: reducing utilization of a candidate zone is indispensable***

Observation 2: I/O size for Read/Write

- **Another feature of ZNS SSD**
  - ✓ **A zone is, in general, mapped into multiple channels/ways.**

- **Then, how about read/write data in a larger I/O size (e.g. 128KB)?**

Observation 2: I/O size for Read/Write



☞ ***Motivation 2: accessing in a larger I/O size is beneficial in ZNS SSDs***

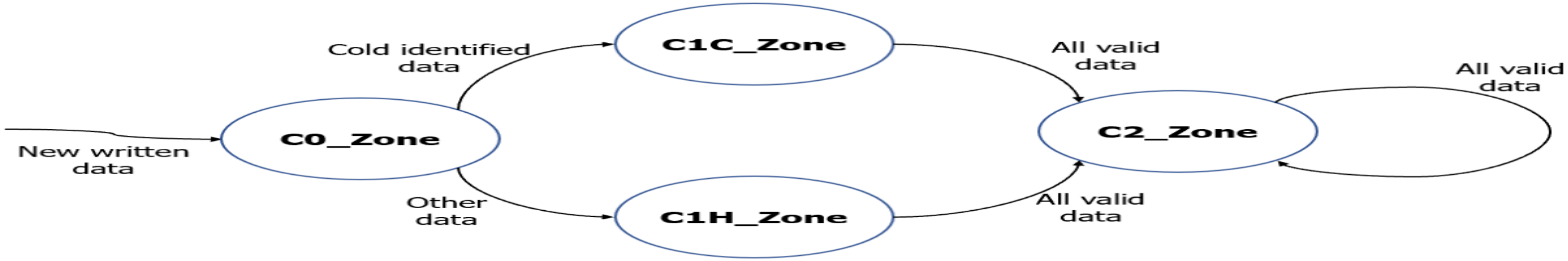
So, Our ideas are

1) Make the utilization of a candidate zone low
2) Access data in a larger I/O size

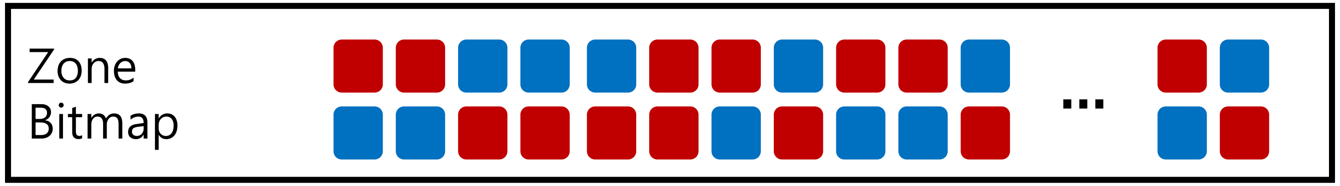
- **How to access data in a larger I/O size?**
  - ✓ **The coexistence of valid and invalid data makes it difficult**
  - ✓ **Read not only valid but also invalid data in a larger I/O size**

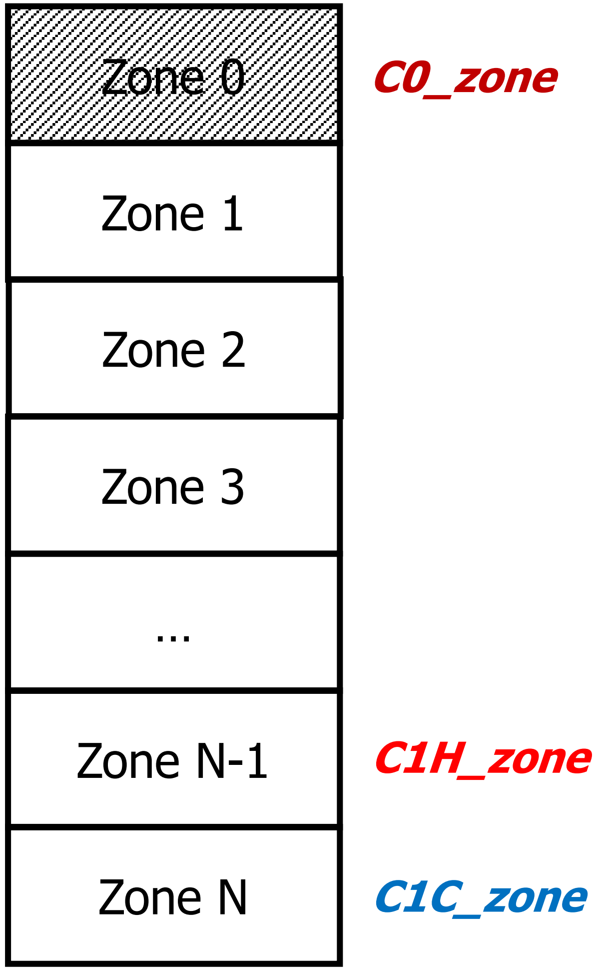
- **How to make the utilization of a candidate zone low?**
  - ✓ **Traditional hot/cold separation is not applicable in ZNS SSDs since zone is quite big**
  - ✓ **Employ the segment concept for finer-grained hot/cold separation**

- **Two management units**
  - ✓ **Zone: for garbage collection vs. Segment: for hot/cold separation**
  - ✓ **A zone is divided into multiple segments (1GB vs. 2MB in this study)**
- **Segment state and transition rule (refer to our paper for details)**
  - ✓ **New data ➔ C0**
  - ✓ **During ZGC, survived data from C0**
    - **Data in a high utilized segment (> threshold_{cold}): cold ➔ C1C**
    - **Others: hot (or unknown) ➔ C1H**
    - **Reasoning: spatial locality, also observed in previous studies such as F2FS (FAST'15), Multi-stream (FAST'19), Key-range locality (FAST'20)**
  - ✓ **During ZGC, survived data from C1C or C1H (second survived data)➔ C2**



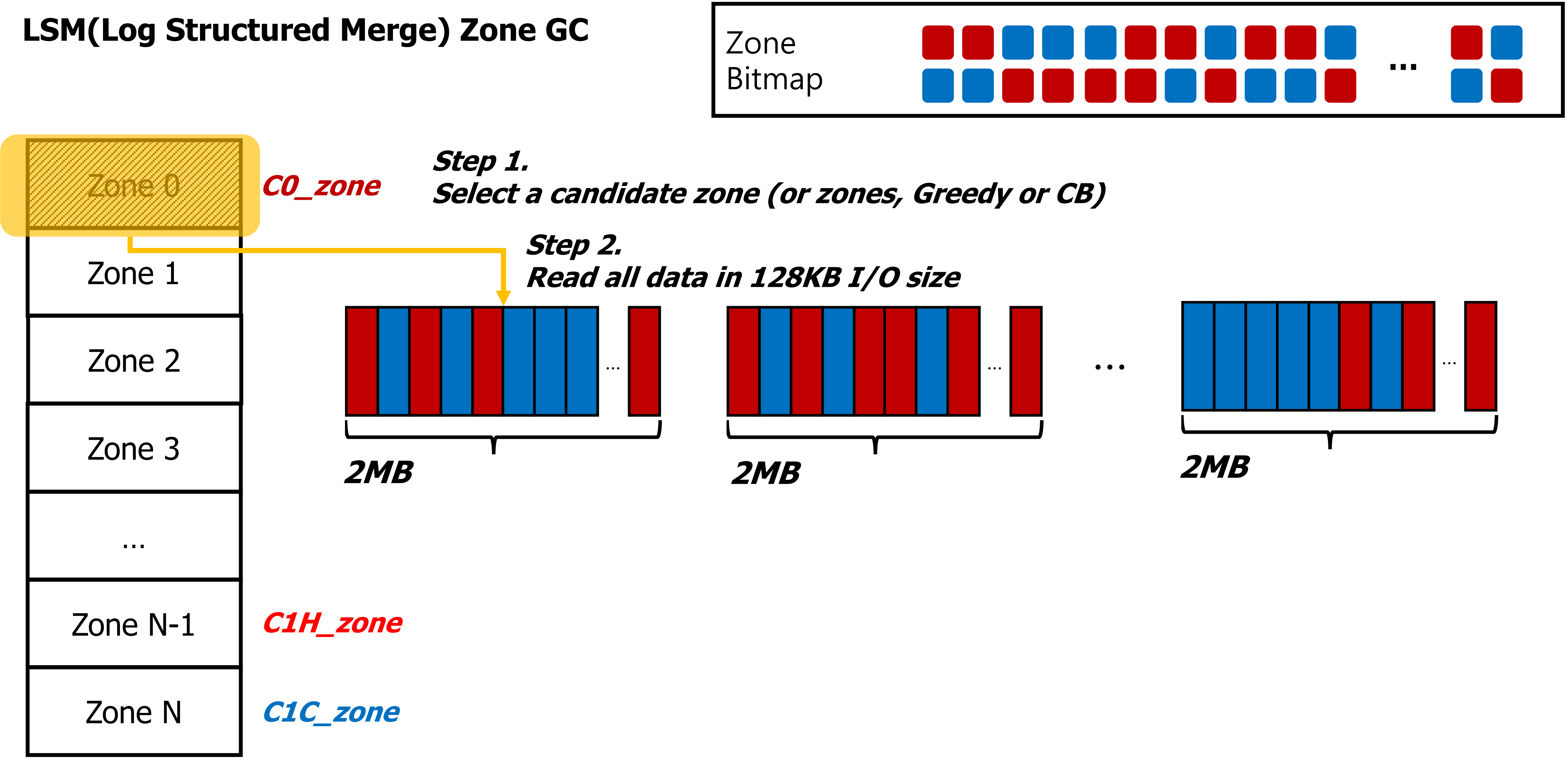
LSM(Log Structured Merge) Zone GC





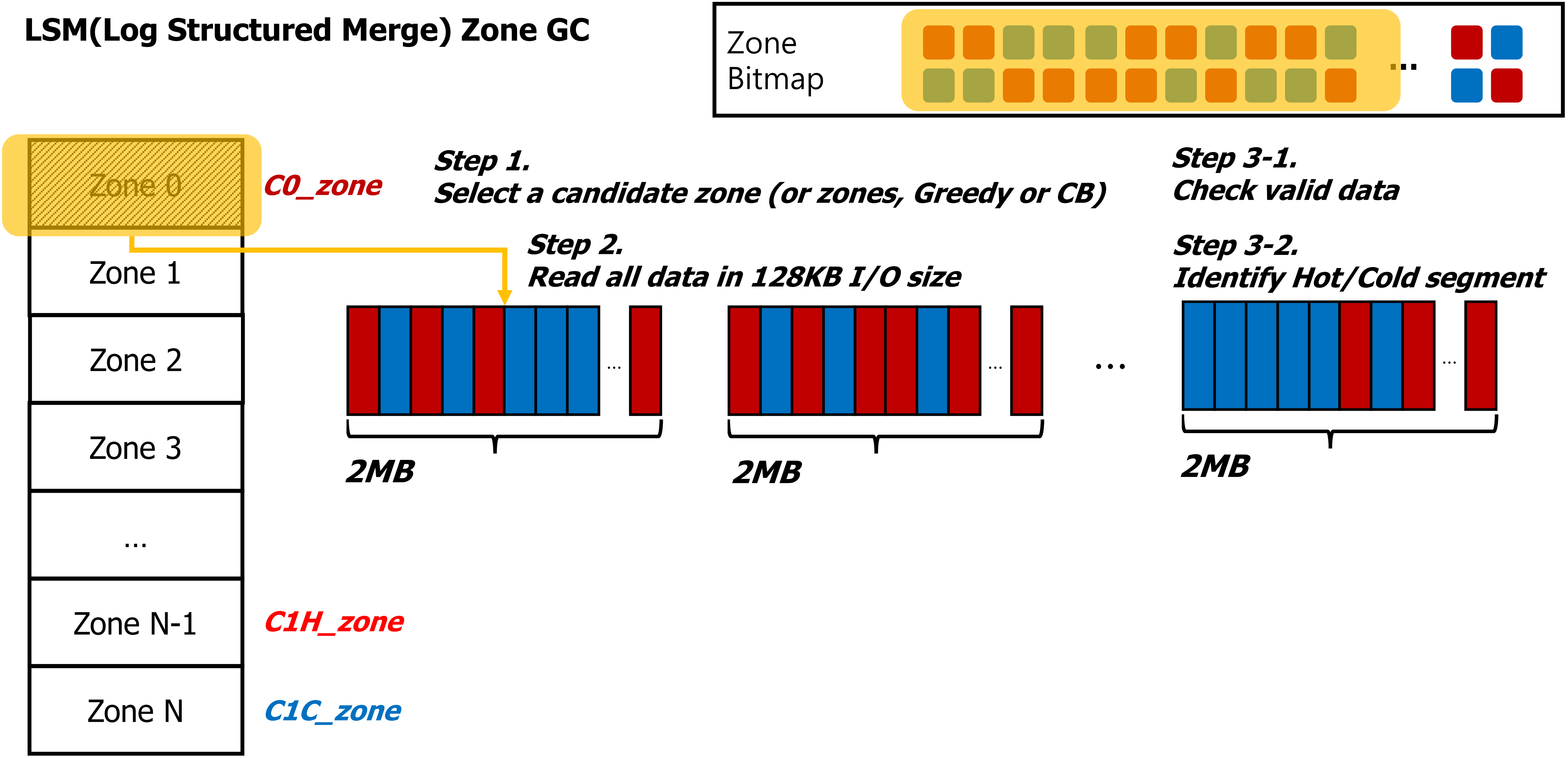
Step 1.
Select a candidate zone (or zones, Greedy or CB)

LSM(Log Structured Merge) Zone GC

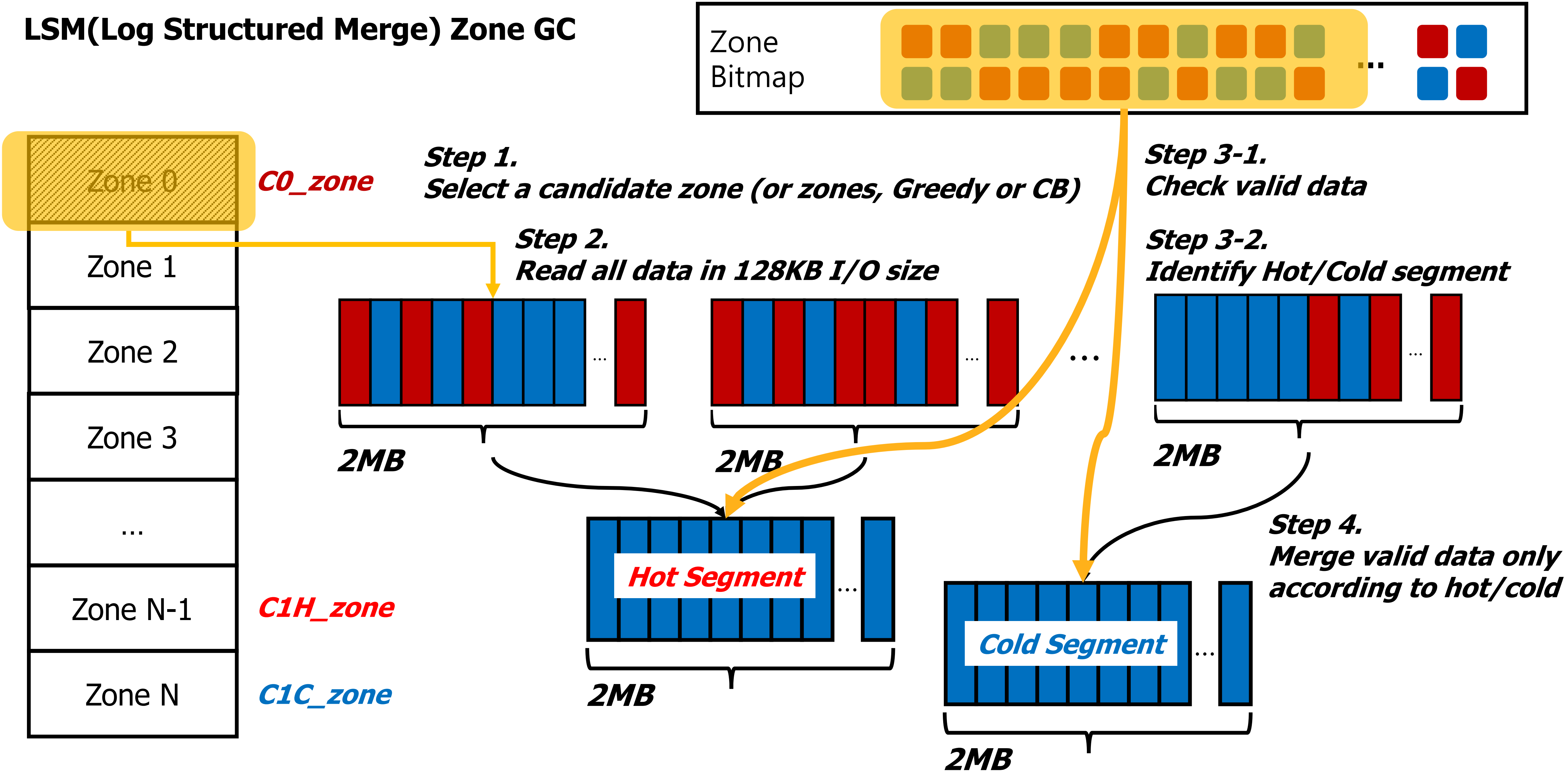


3. LSM-ZGC Design

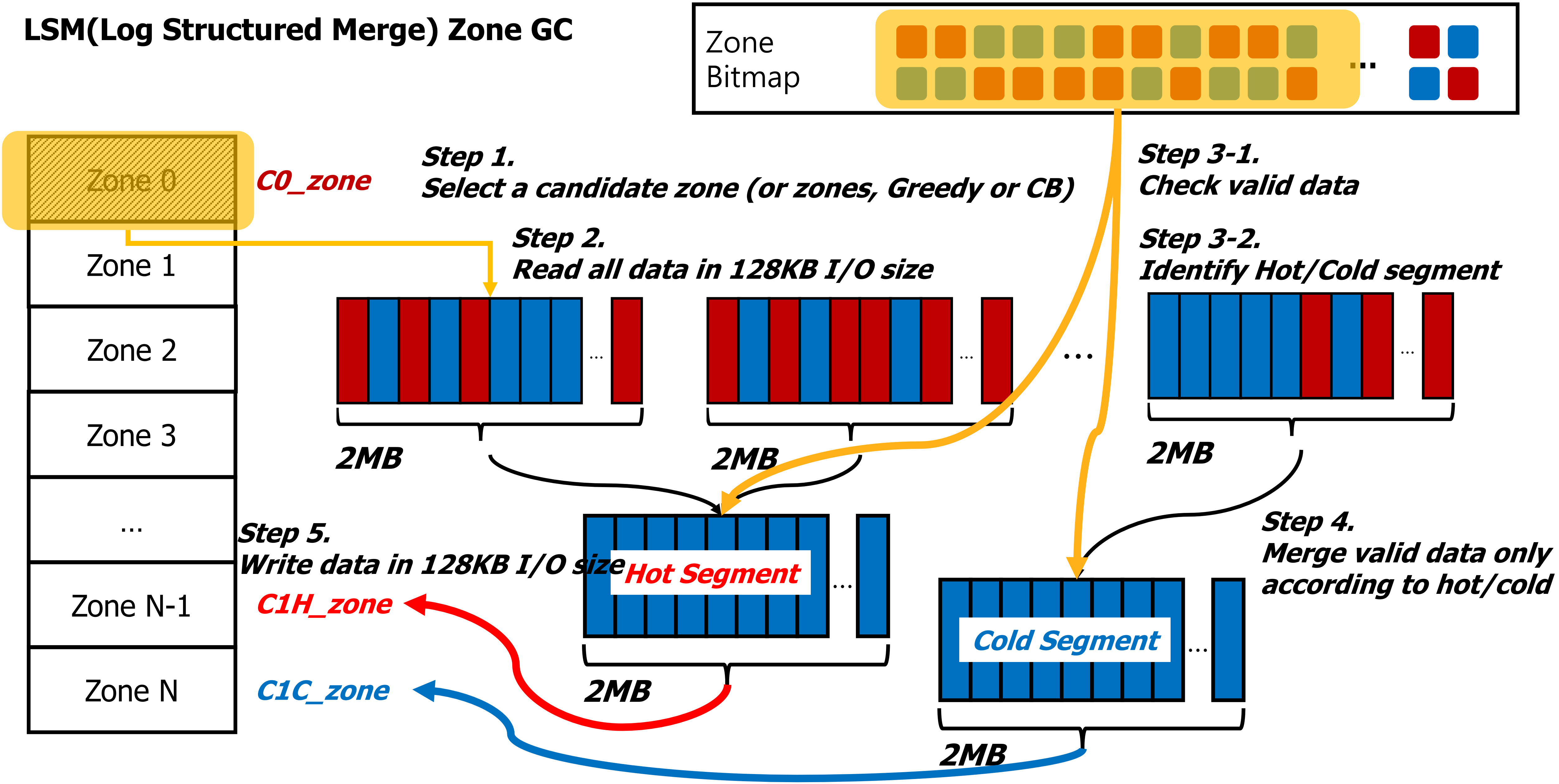
LSM(Log Structured Merge) Zone GC



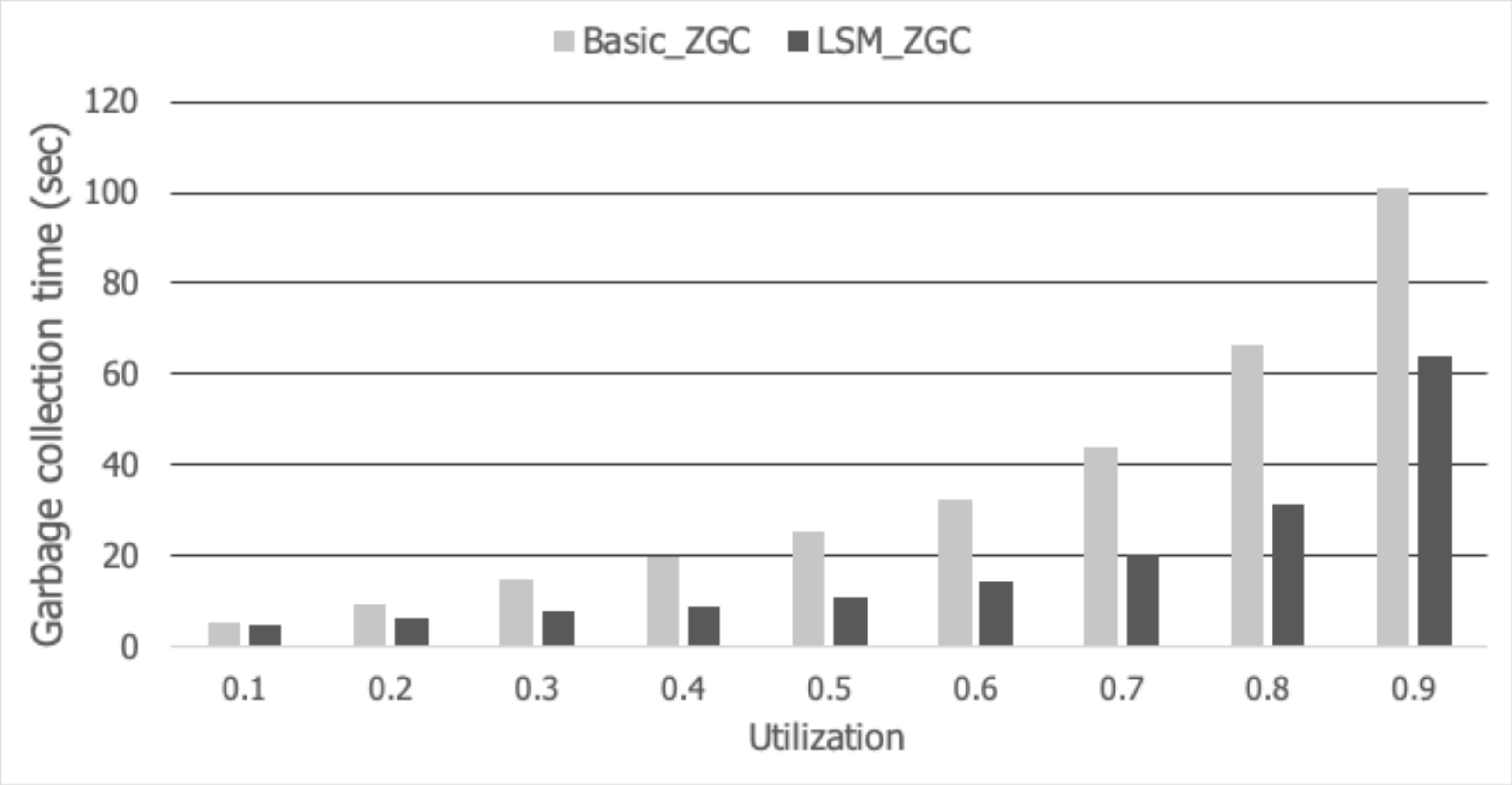
LSM(Log Structured Merge) Zone GC



LSM(Log Structured Merge) Zone GC



Garbage collection overhead: uniform update pattern



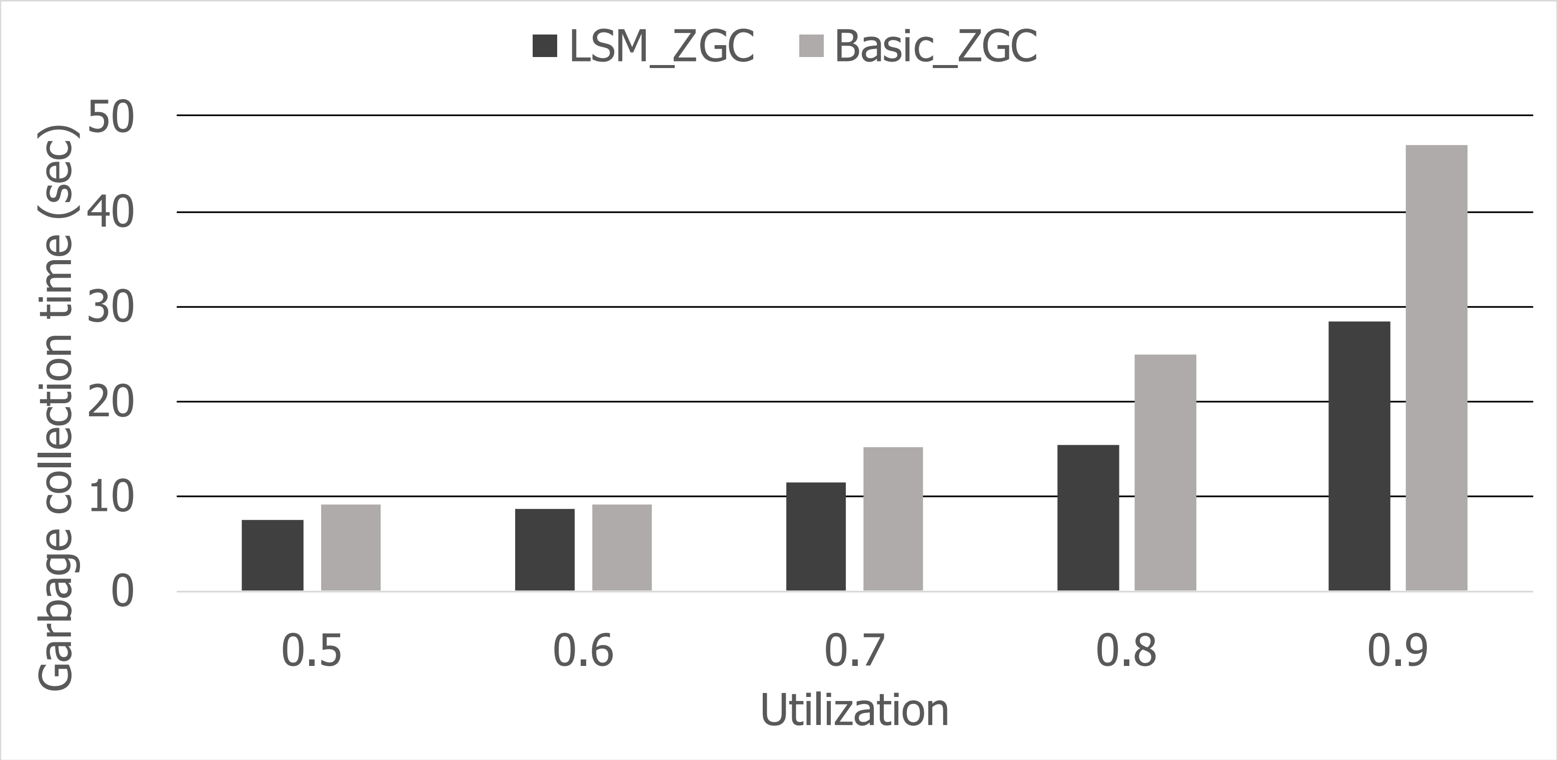
Average of 1.9 times

Max of 2.3 times

Experimental environment

- Intel Core i7 (8 core)
- 16GB DRAM
- 1TB ZNS SSD
- Size of Zone : 1GB

Garbage collection overhead: skewed update pattern



Average of 1.4 times

Max of 1.6 times

Parameters

- Workload: 70/30 hot/cold ratio
- Threahold$_{cold}$: 0.8
- average utilization: x-axis

Experimental environment

- Intel Core i7 (8 core)
- 16GB DRAM
- 1TB ZNS SSD
- Size of Zone : 1GB

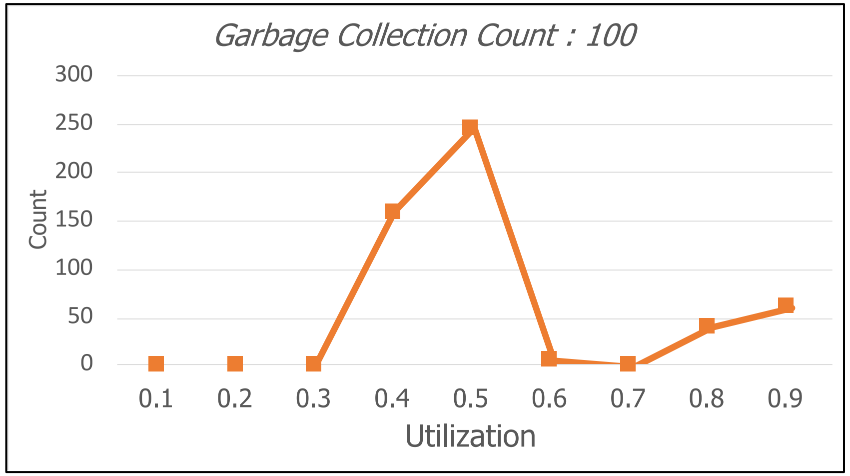
4. Evaluation

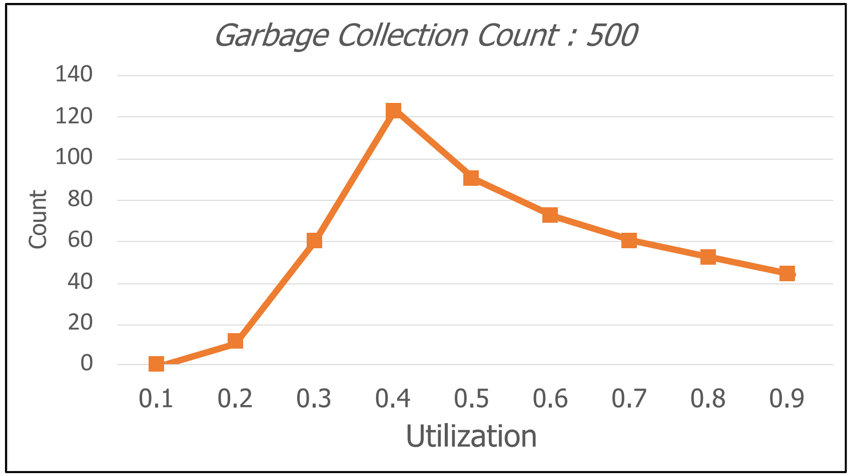
Parameters

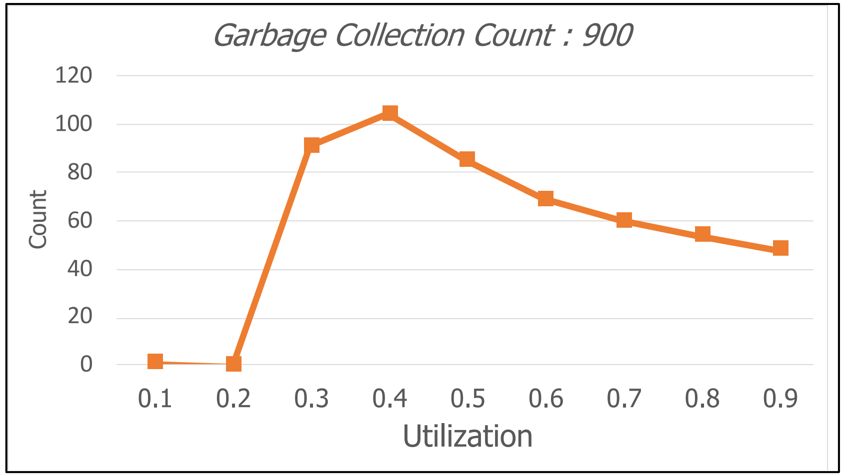
- Workload: 70/30 hot/cold ratio
- Threahold$_{cold}$: 0.8
- Average utilization: 0.6

Hot/Cold Separation

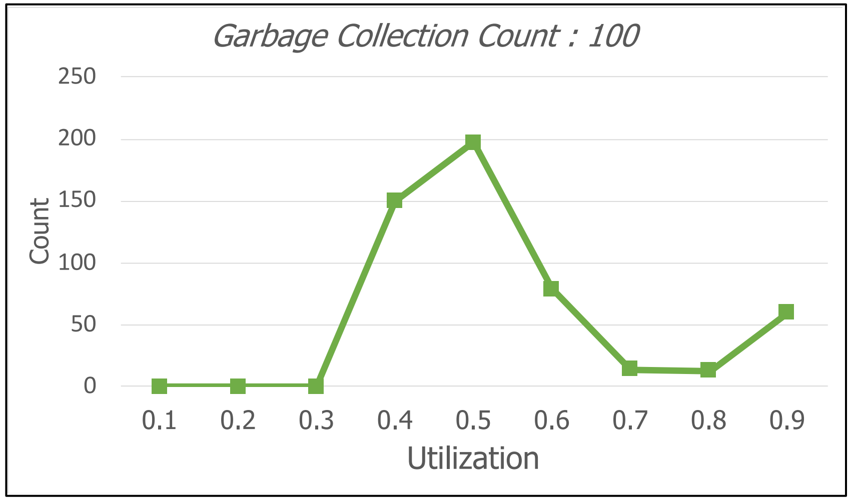
✓ Without hot/cold separation

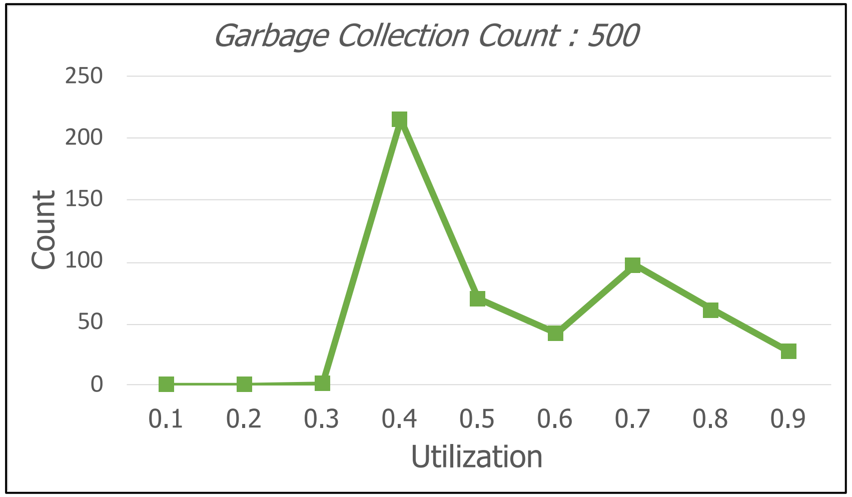


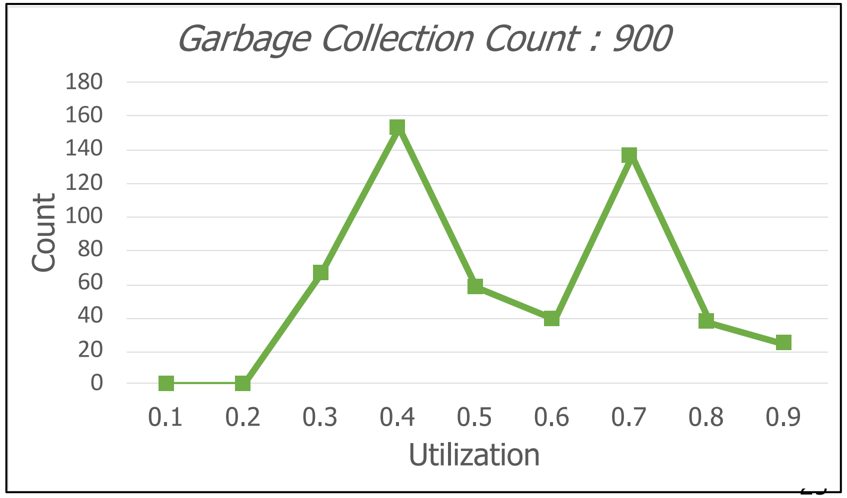




✓ With hot/cold separation







- **Our contributions**
  - *Observation: a zone garbage collection really matters*
  - *Proposal: a new LSM-style zone garbage collection scheme*
  - *Evaluation: real implementation based results*

- **Future work**
  - *We are currently extending F2FS on our ZNS SSD prototype*
  - *Also, evaluating LSM ZGC under diverse workloads with different hot /cold ratio, data size, initial placement and classification policies*

A New LSM-style Garbage Collection Scheme for ZNS SSDs

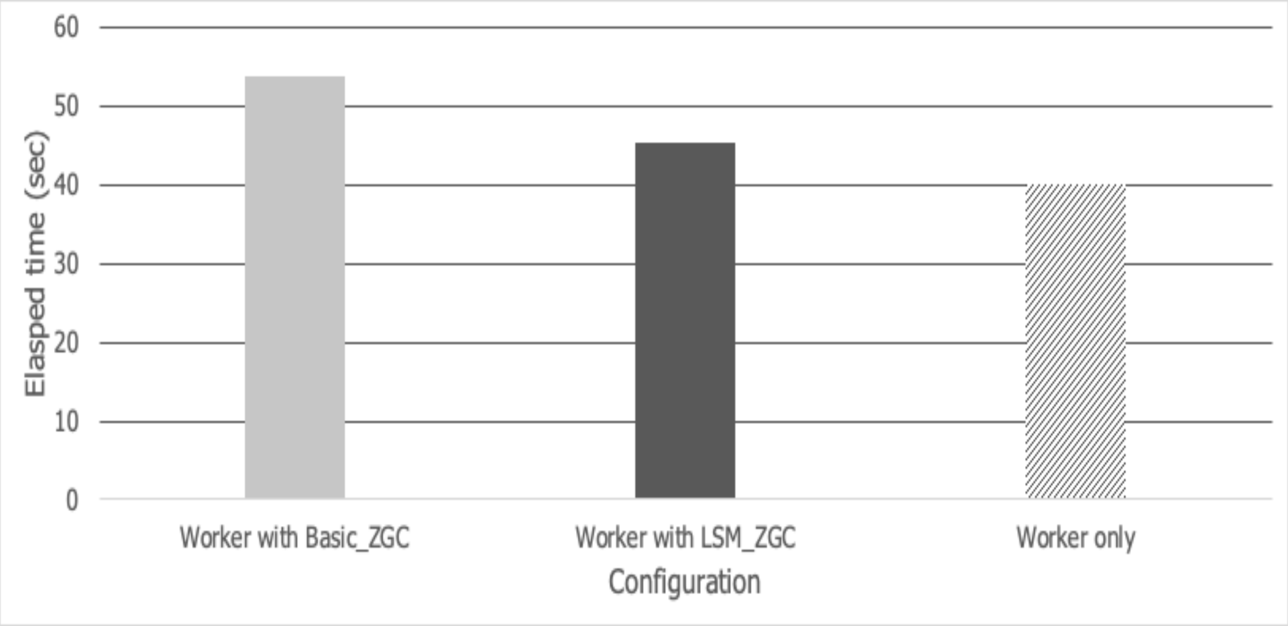
12th USENIX Workshop on Hot Topic in Storage and File System (HotStorage 20), 2020

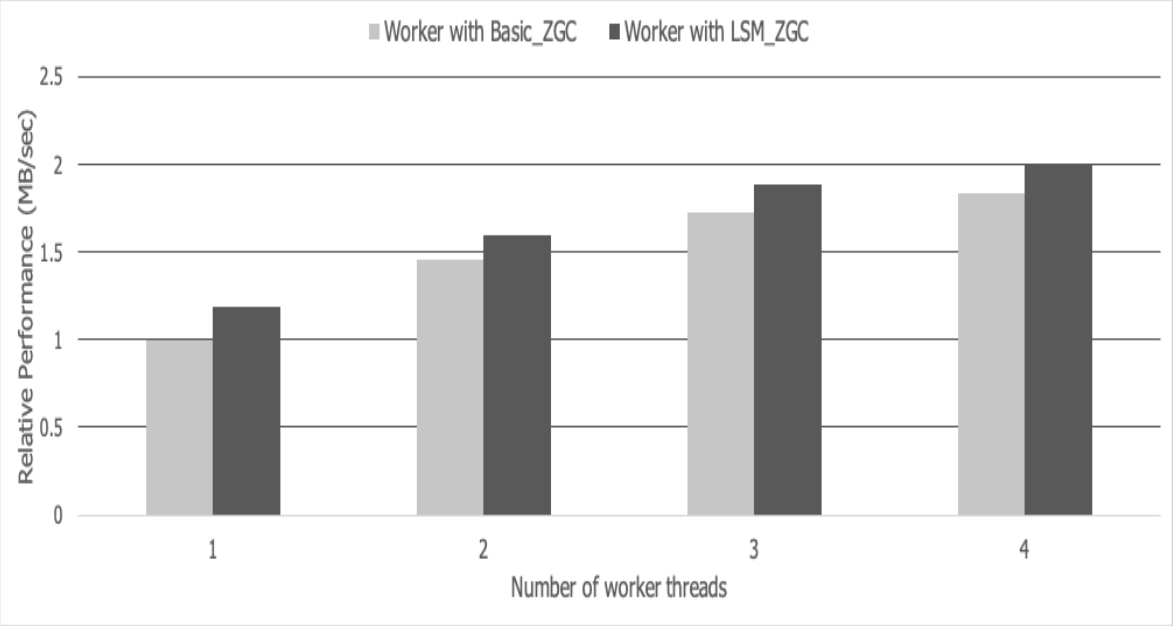
Thank You!

Questions?

2022.12.15
张茂荣
华中科技大学

Performance comparison using multi-thread & Scalability





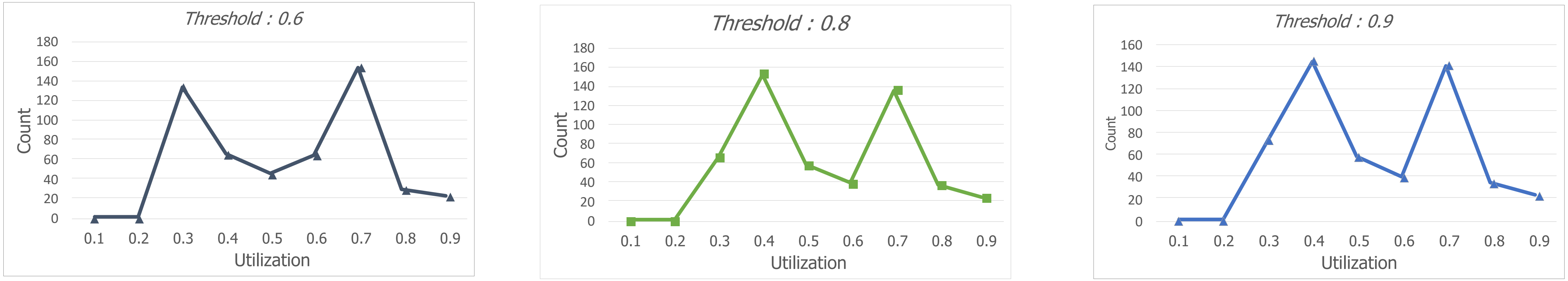
Worker only: 40
With LSM_ZGC : 45
With Basic_ZGC : 53

Non-linear Scalability

6. Appendix

Sensitive Analysis: various parameters

✓ Effect of threshold$_{cold}$ (initial utilization: 0.6)



✓ Effect of initial utilization of a zone (threshold$_{cold}$: 0.8)