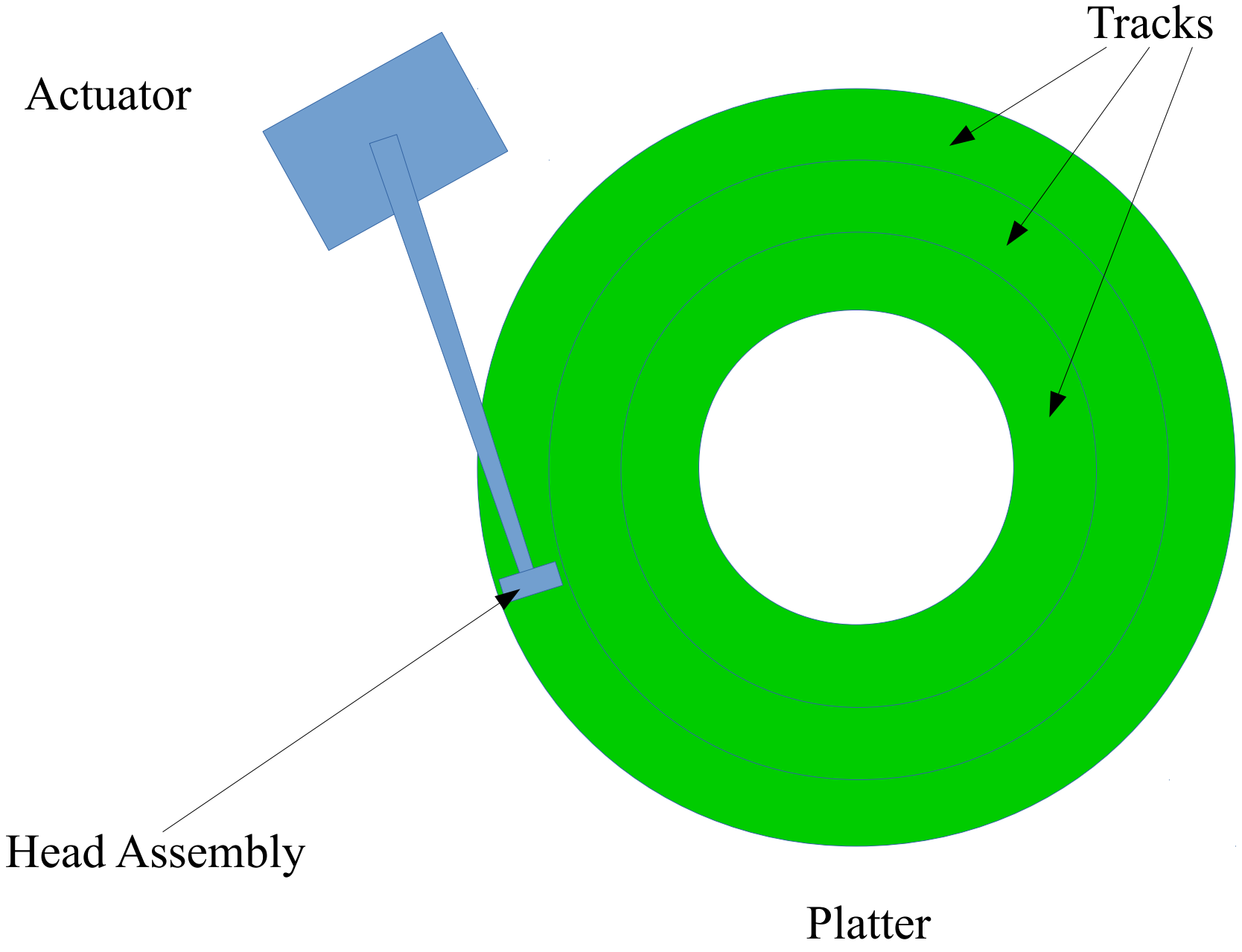
Skylight – A Window on Shingled Disk Operation

Abutalib Aghayev, Peter Desnoyers
Northeastern University

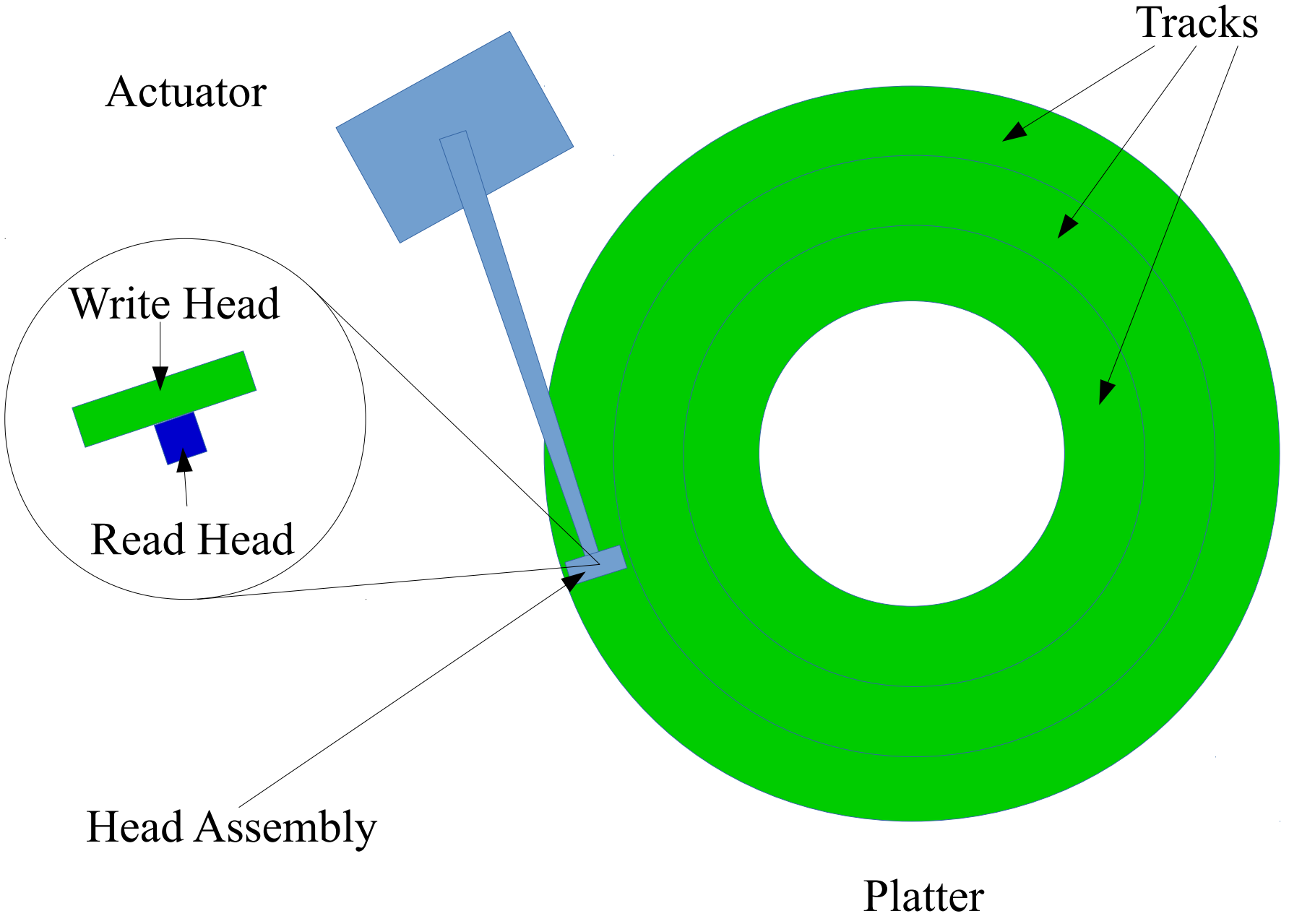
What is Shingled Magnetic Recording (SMR)?

- A new way of recording tracks on the disk platter.
- Evolutionary – uses existing infrastructure.
- Fits more tracks onto platter → increases capacity.
- Disallows random writes → increases complexity.

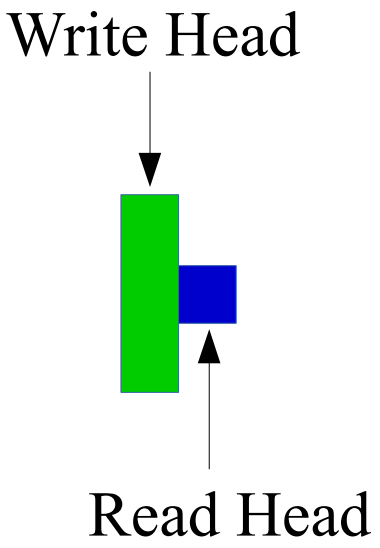
Disk Drive Internals

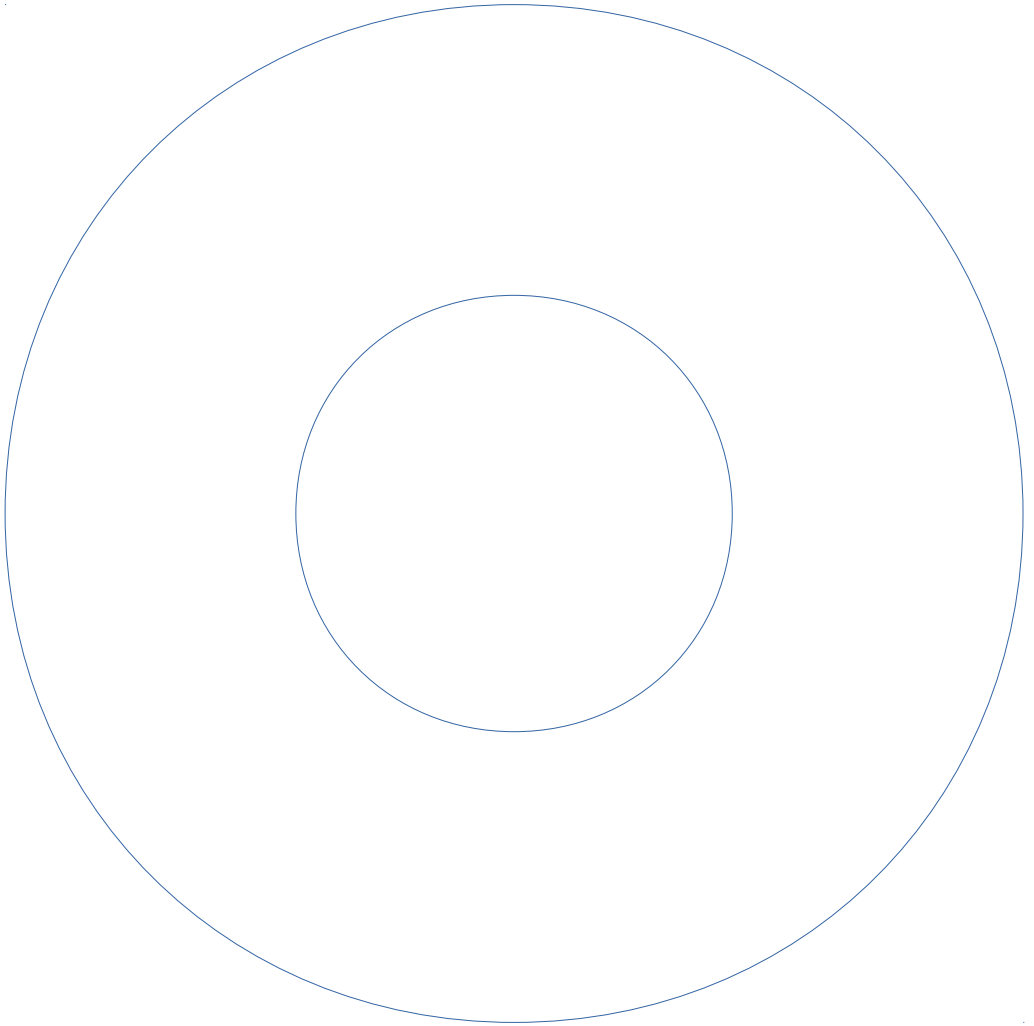


Disk Drive Internals



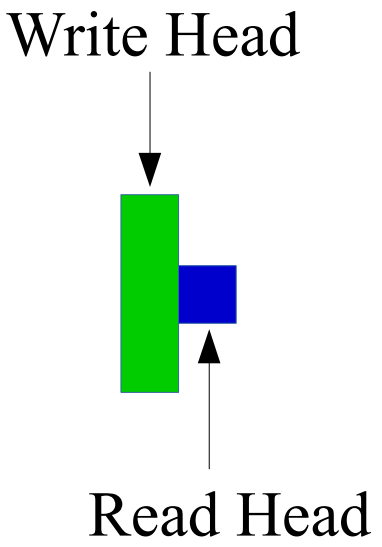
Conventional Magnetic Recording

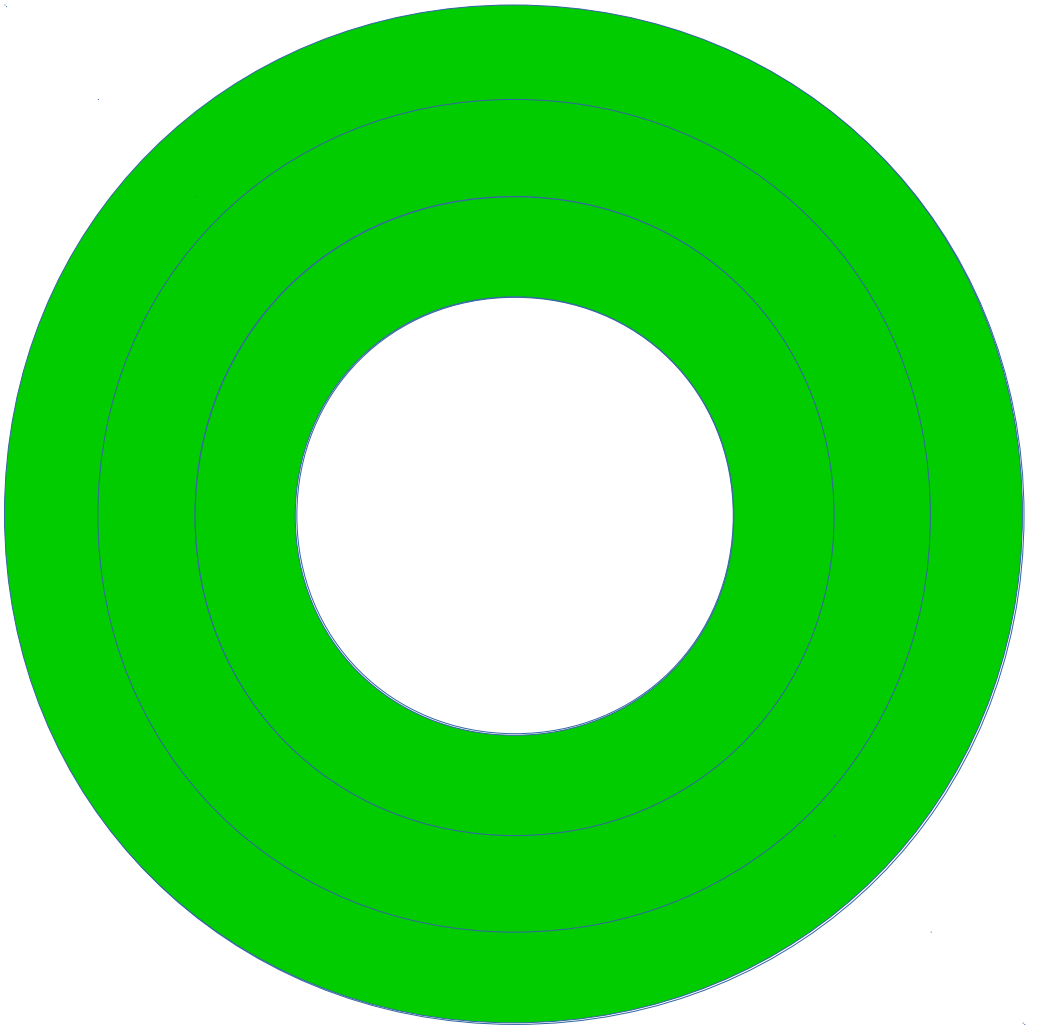




Platter

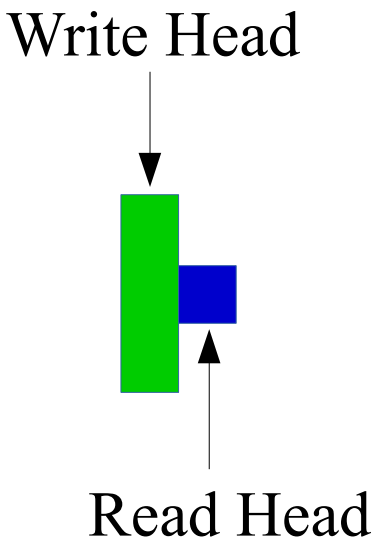
Conventional Magnetic Recording

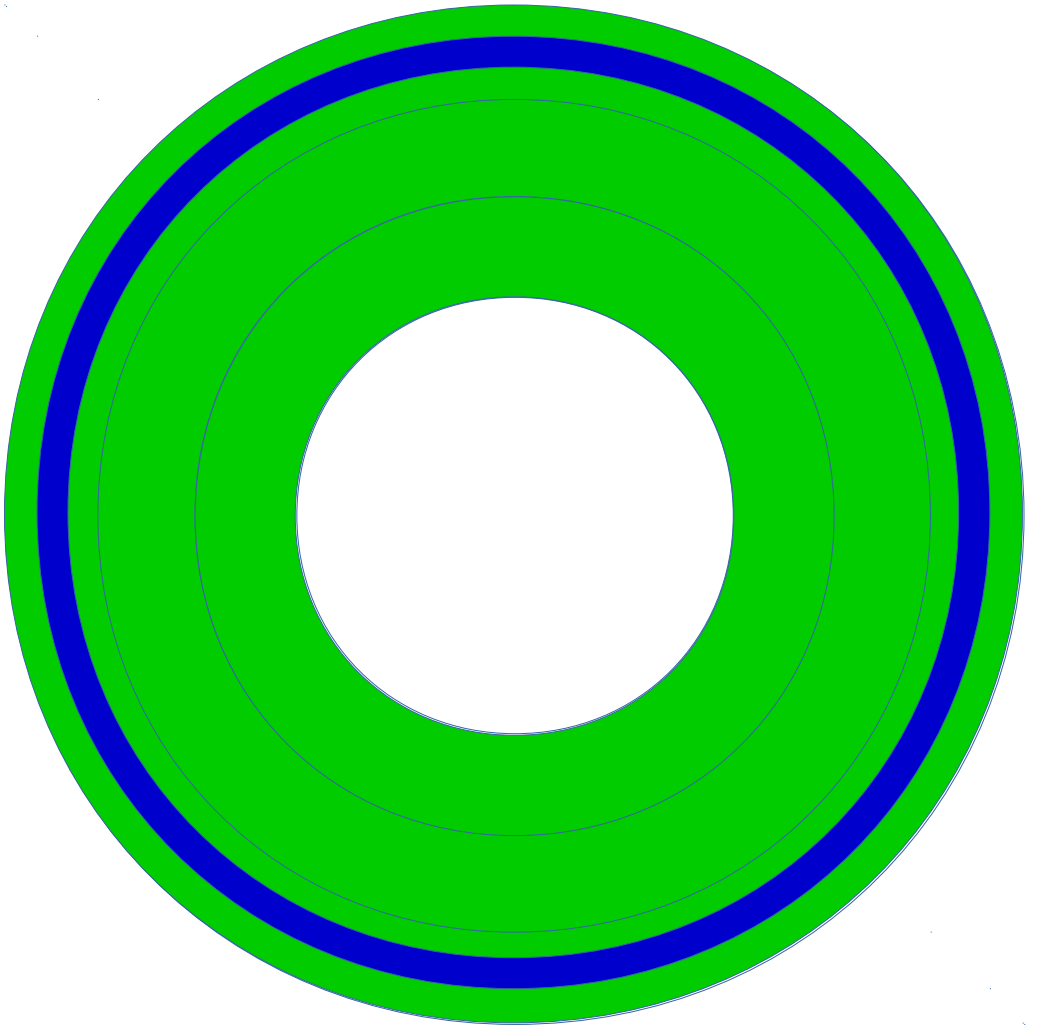




Platter

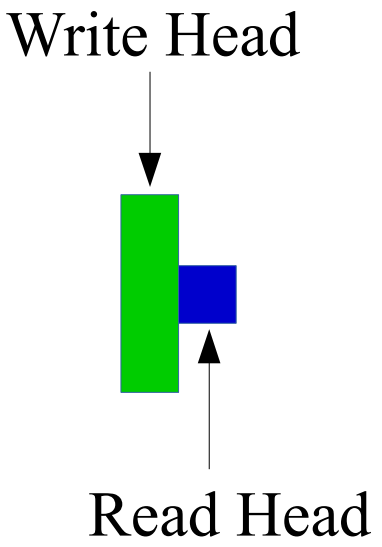
Conventional Magnetic Recording

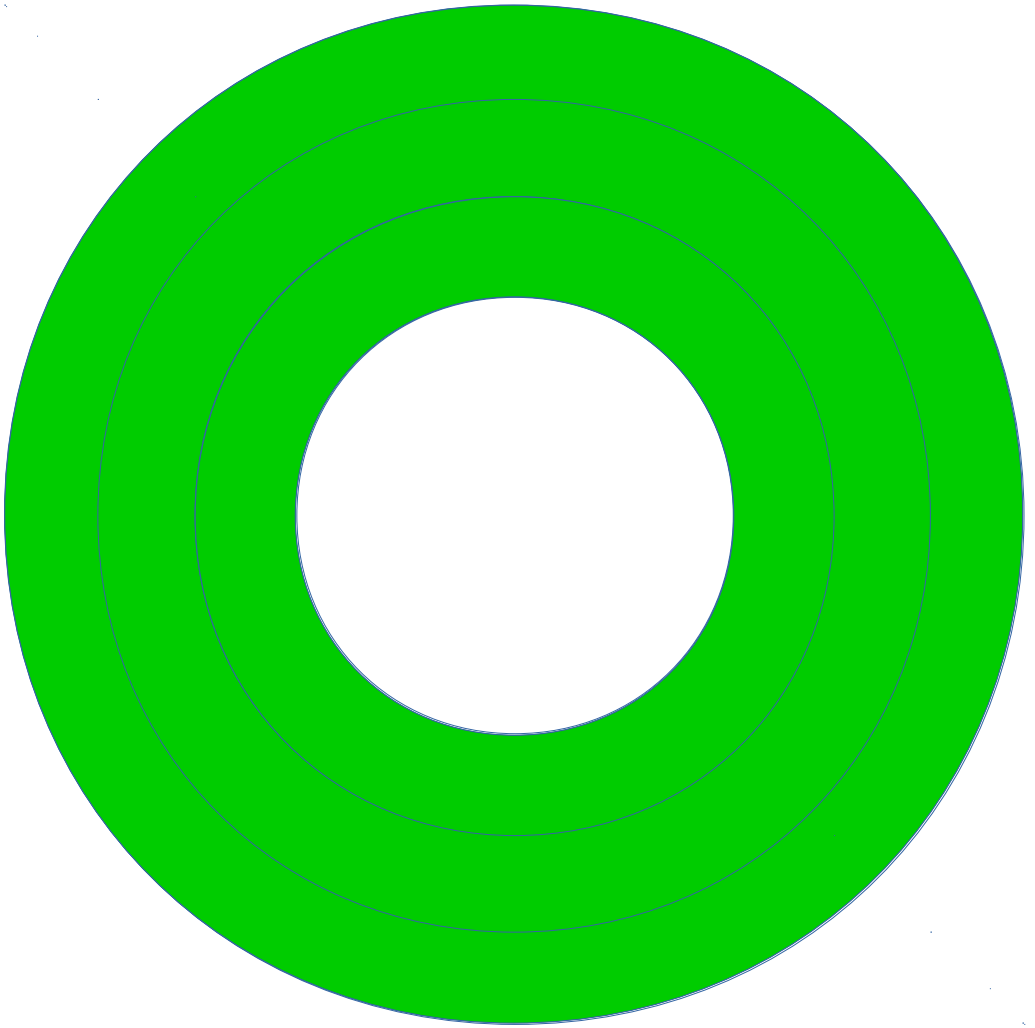




Platter

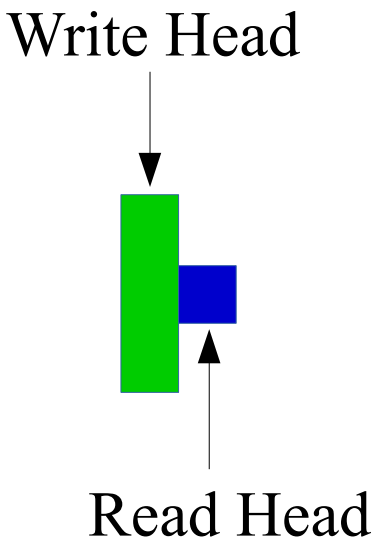
Conventional Magnetic Recording

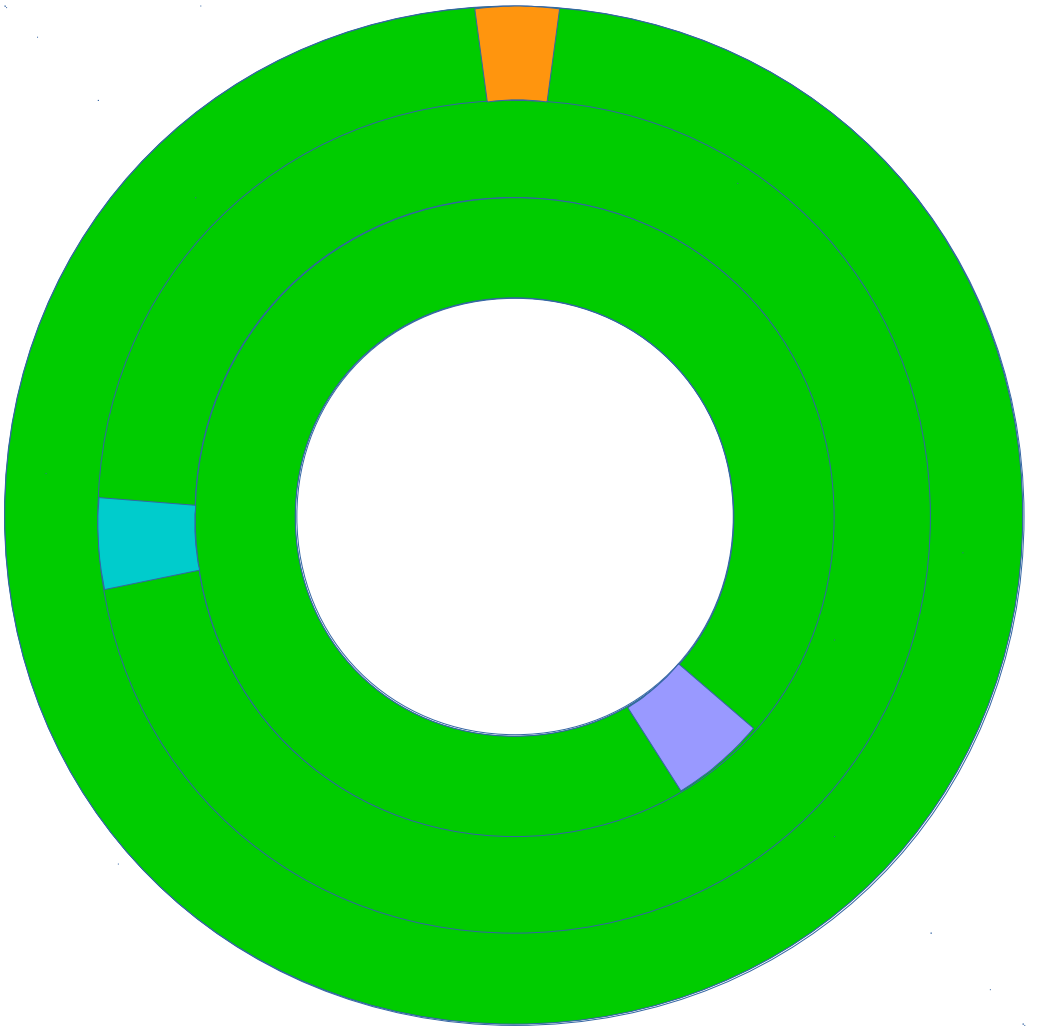




Platter

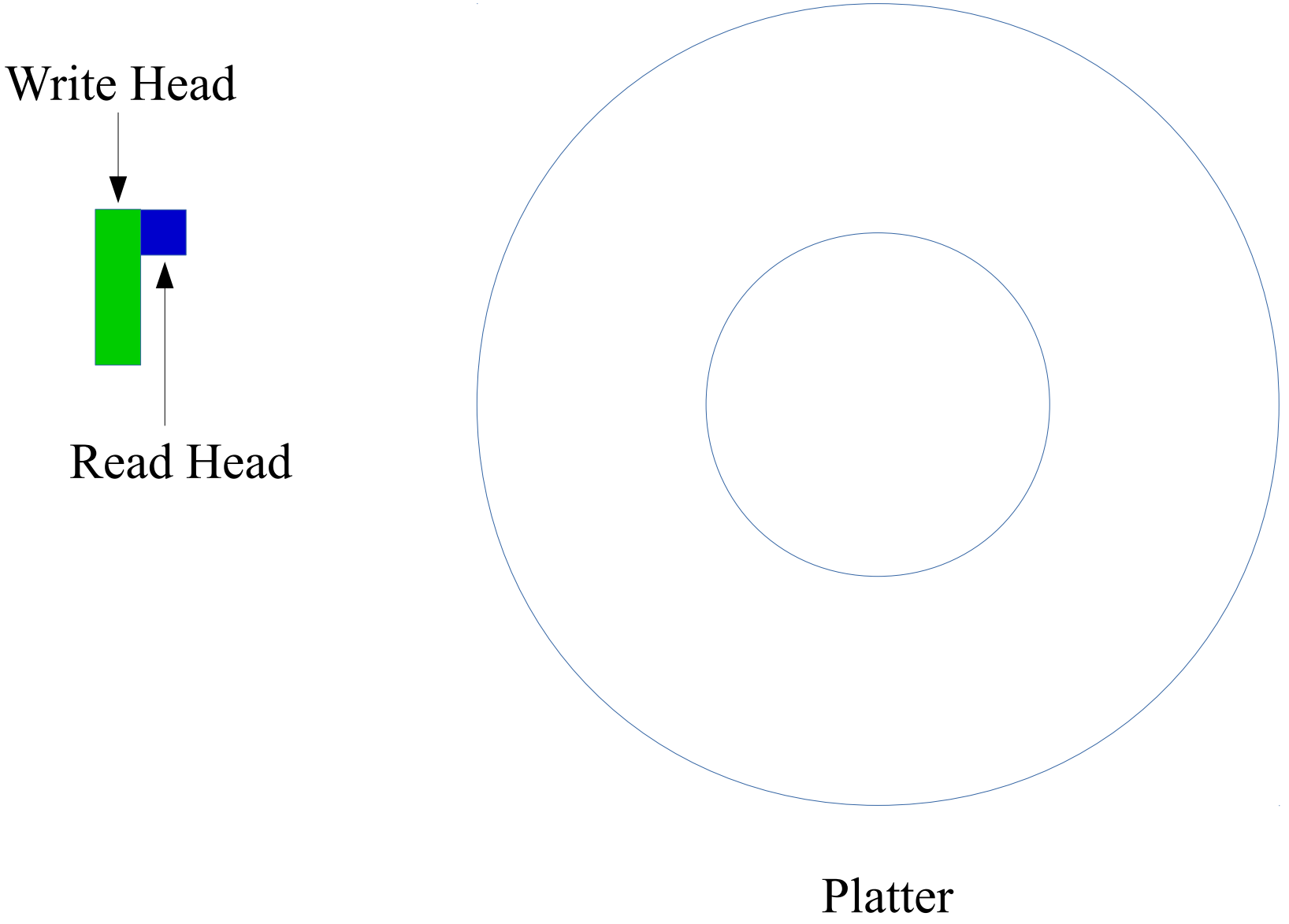
Conventional Magnetic Recording



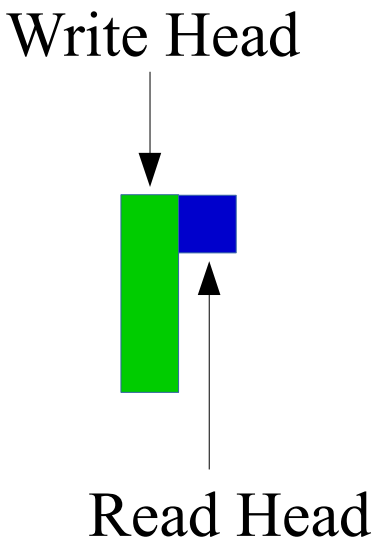


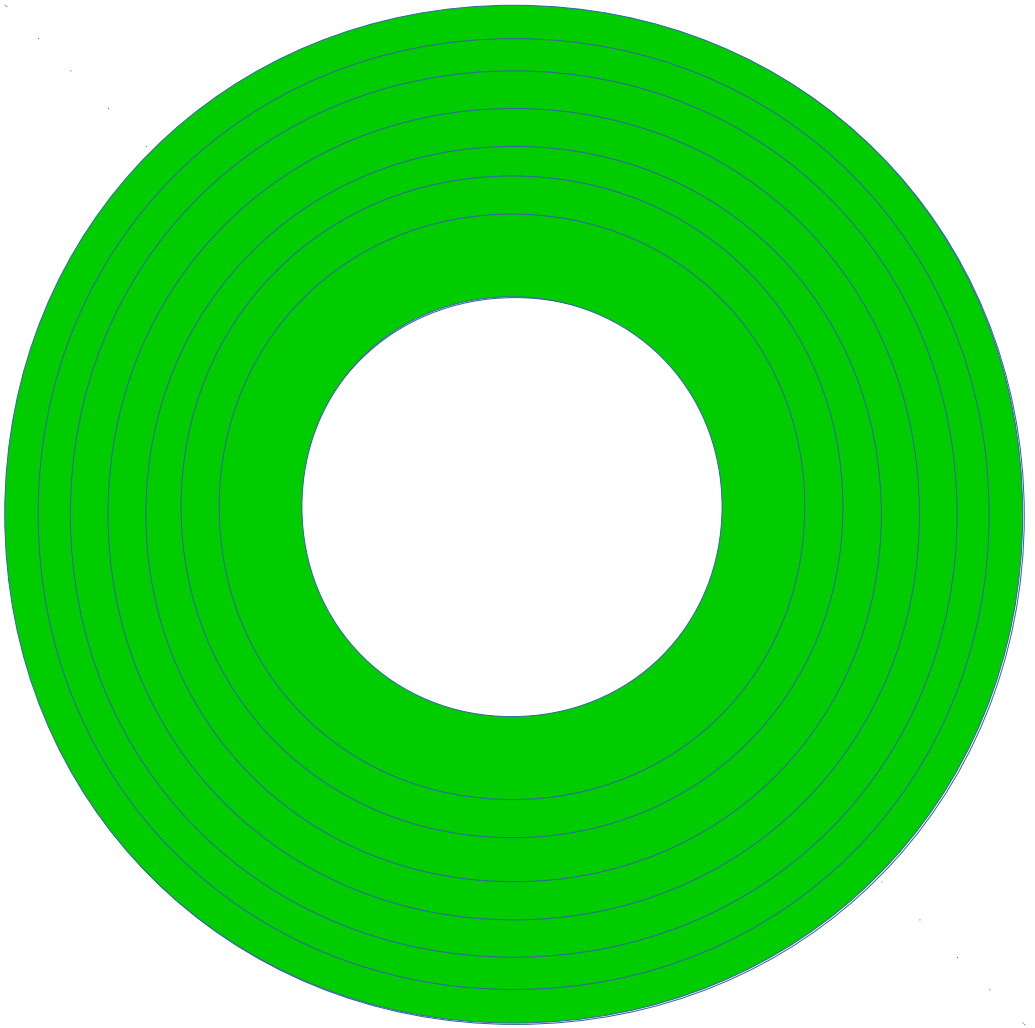
Platter

Shingled Magnetic Recording



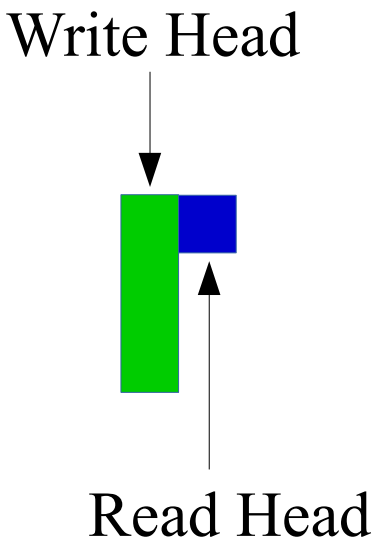
Shingled Magnetic Recording

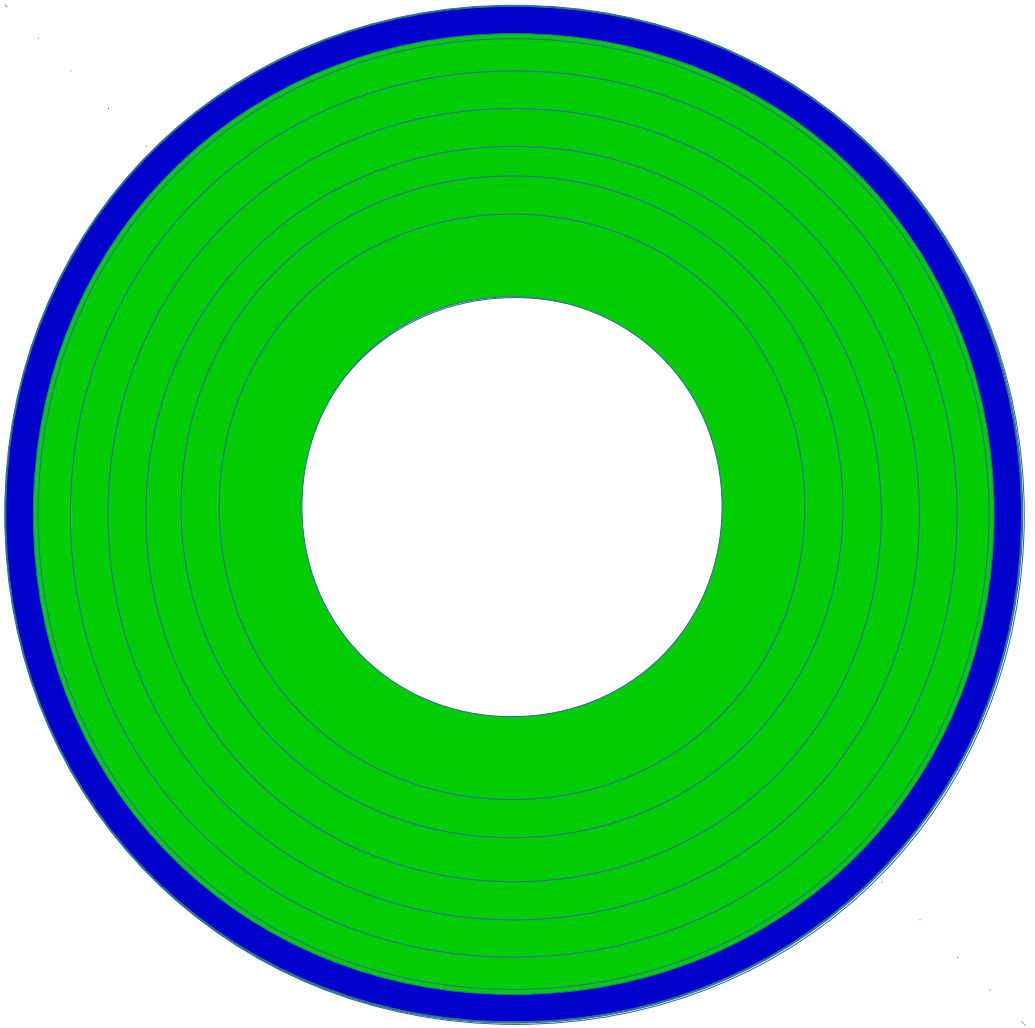




Platter

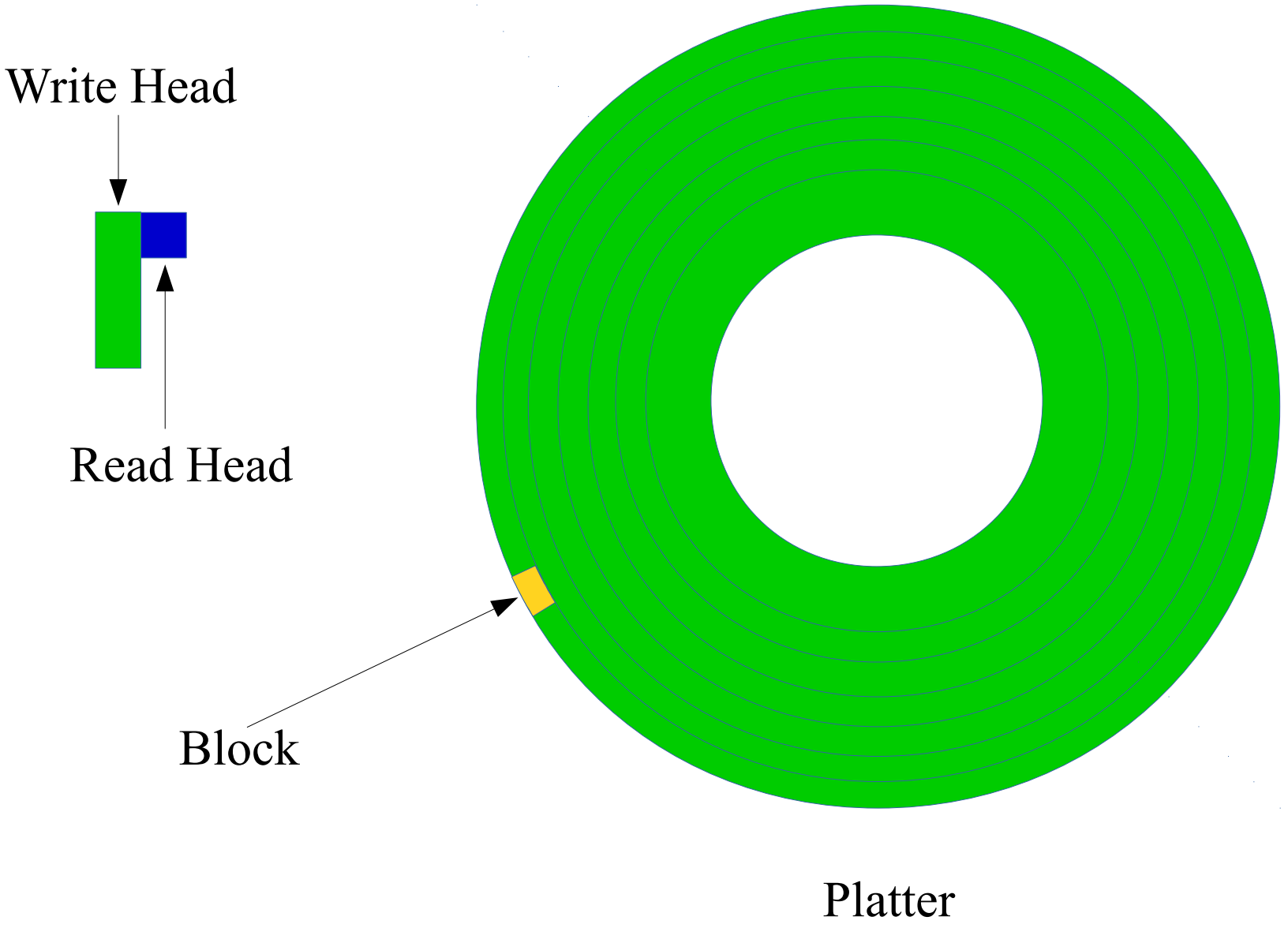
Shingled Magnetic Recording



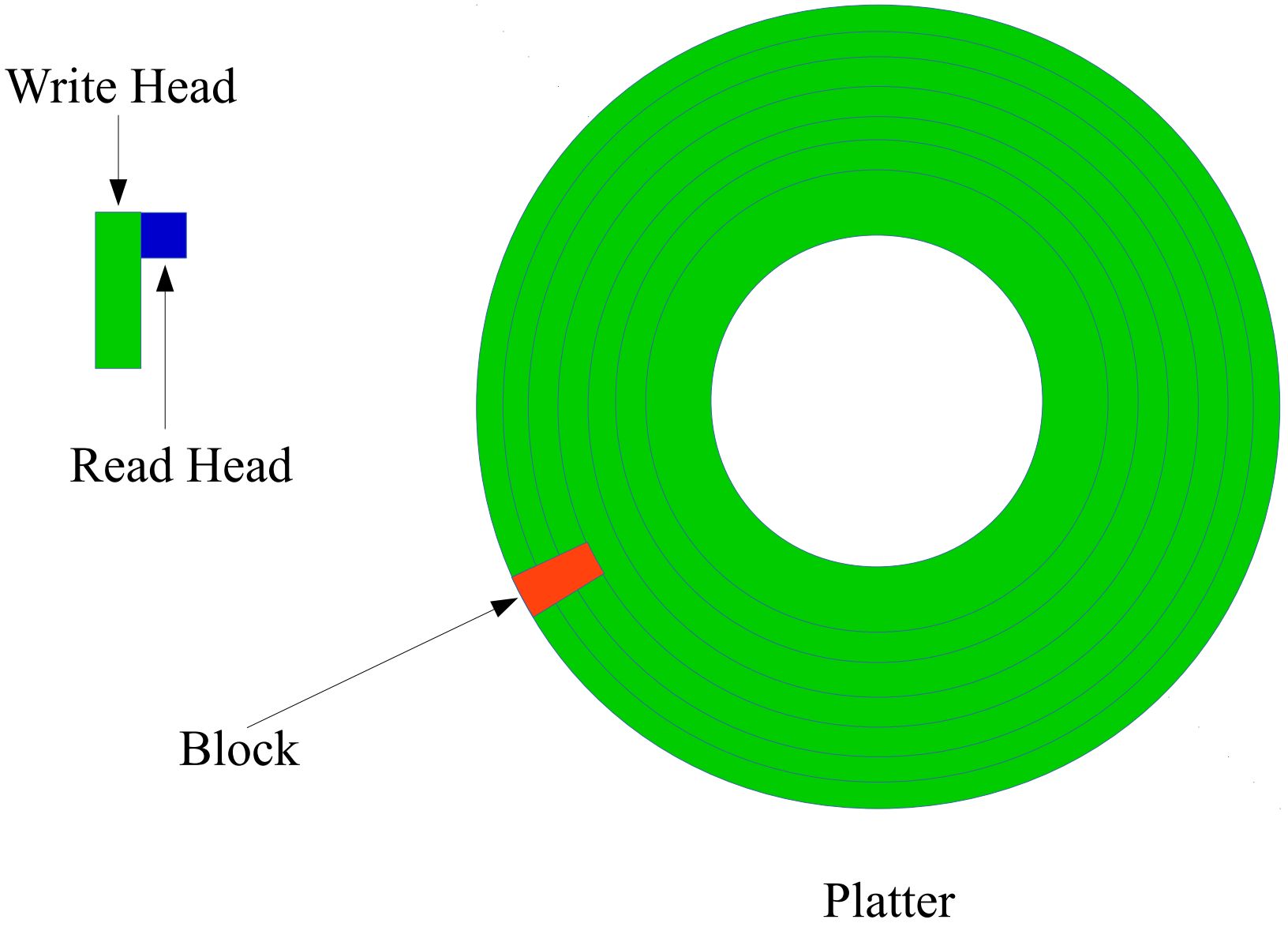


Platter

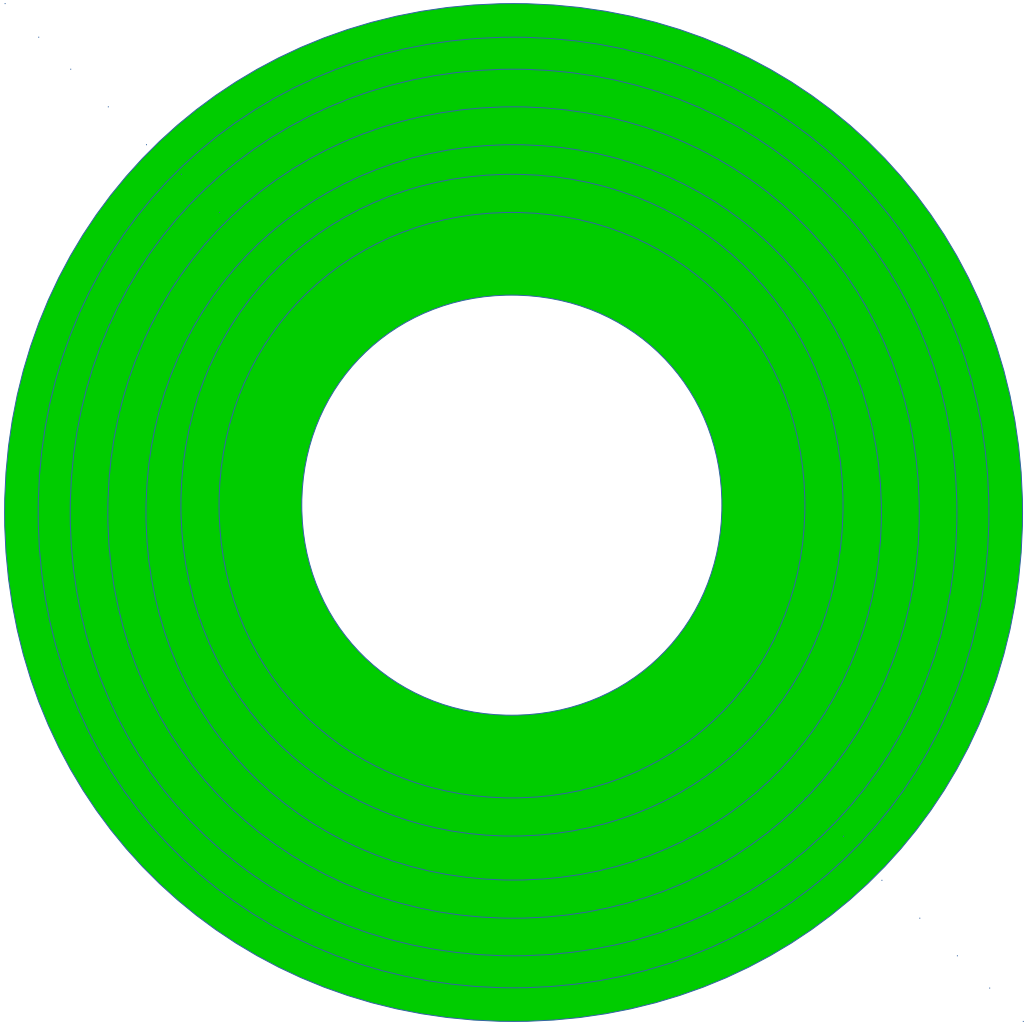
Shingled Magnetic Recording



Shingled Magnetic Recording



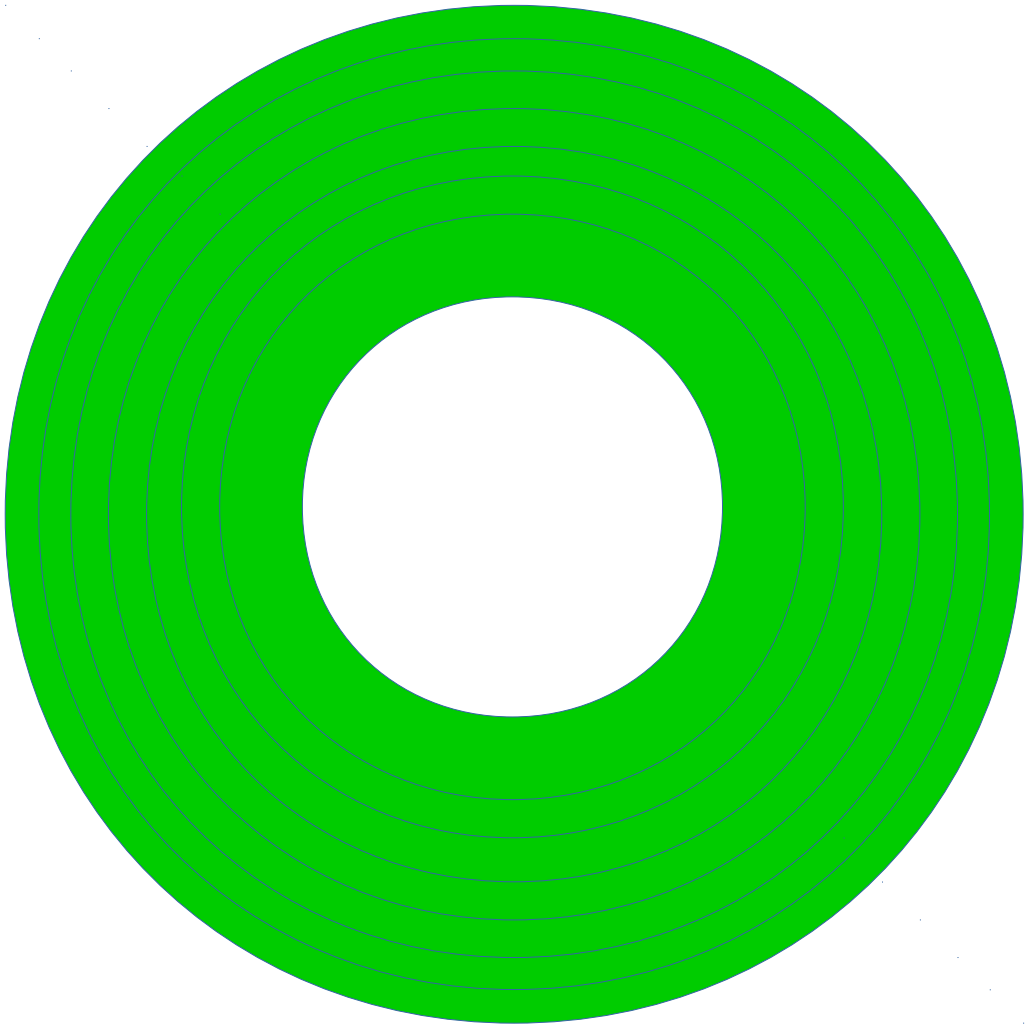
Shingled Magnetic Recording



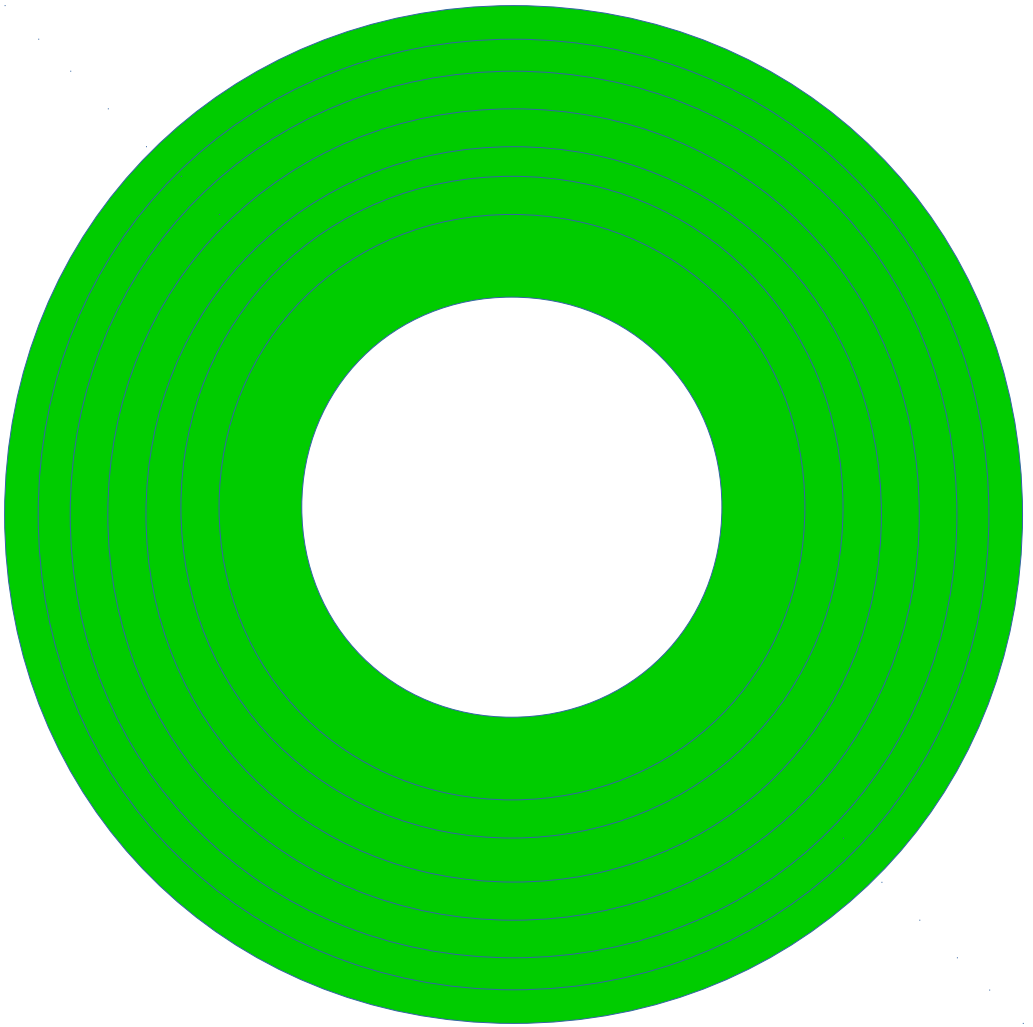
Memory

Platter

Shingled Magnetic Recording

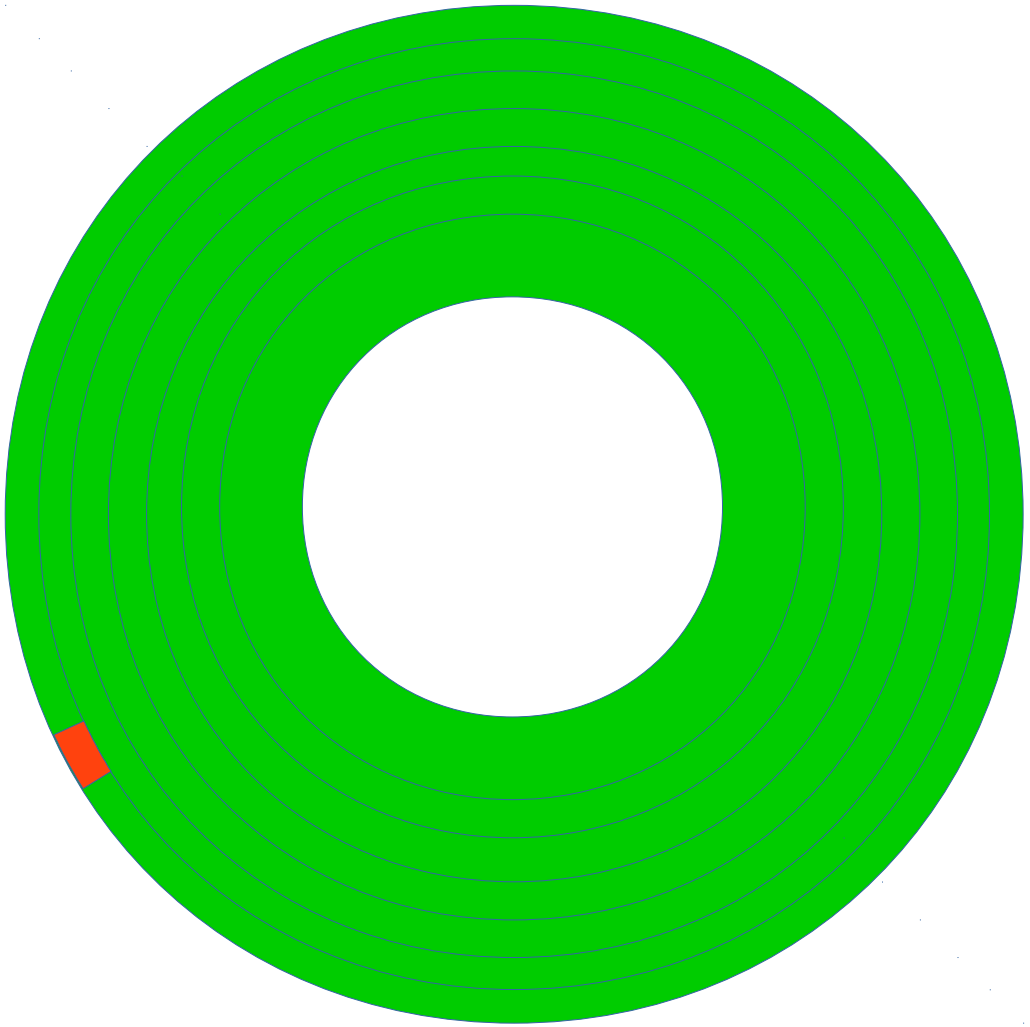


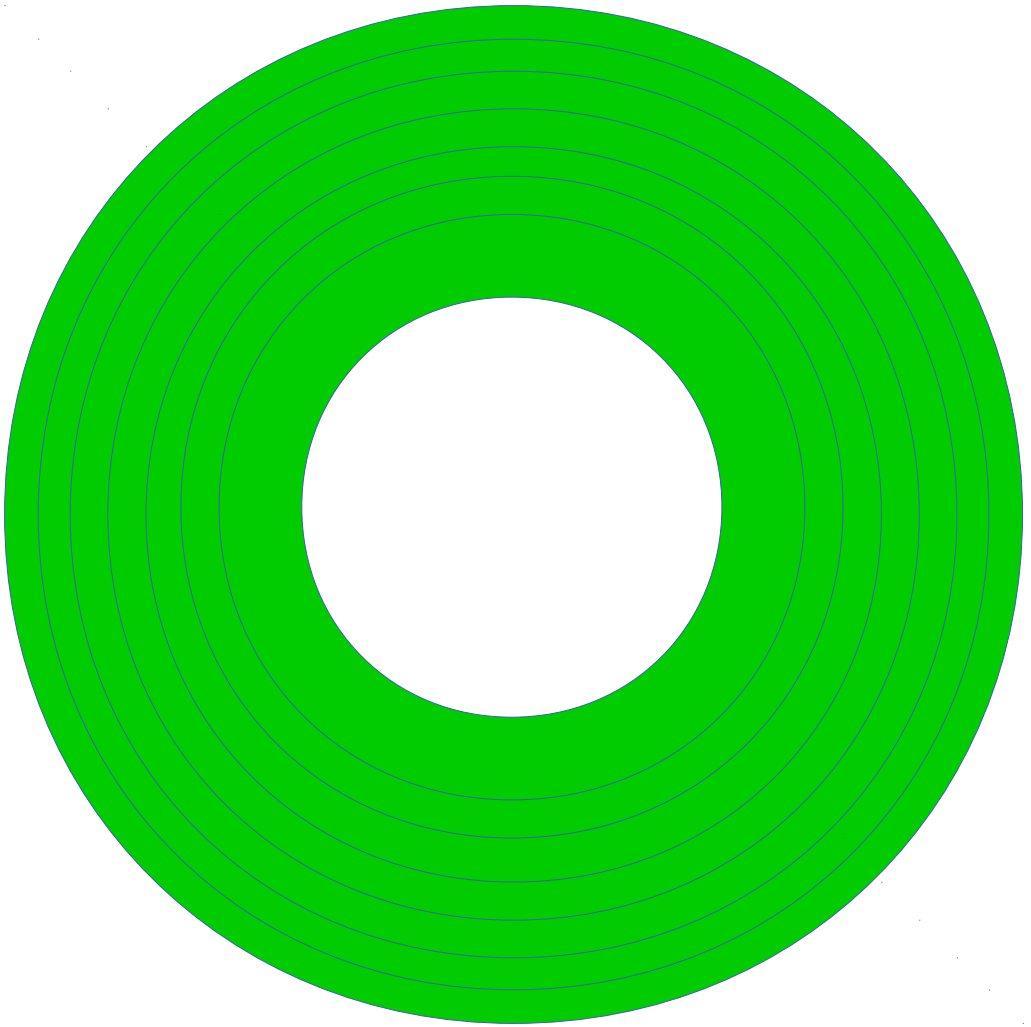
Memory



Platter

Shingled Magnetic Recording

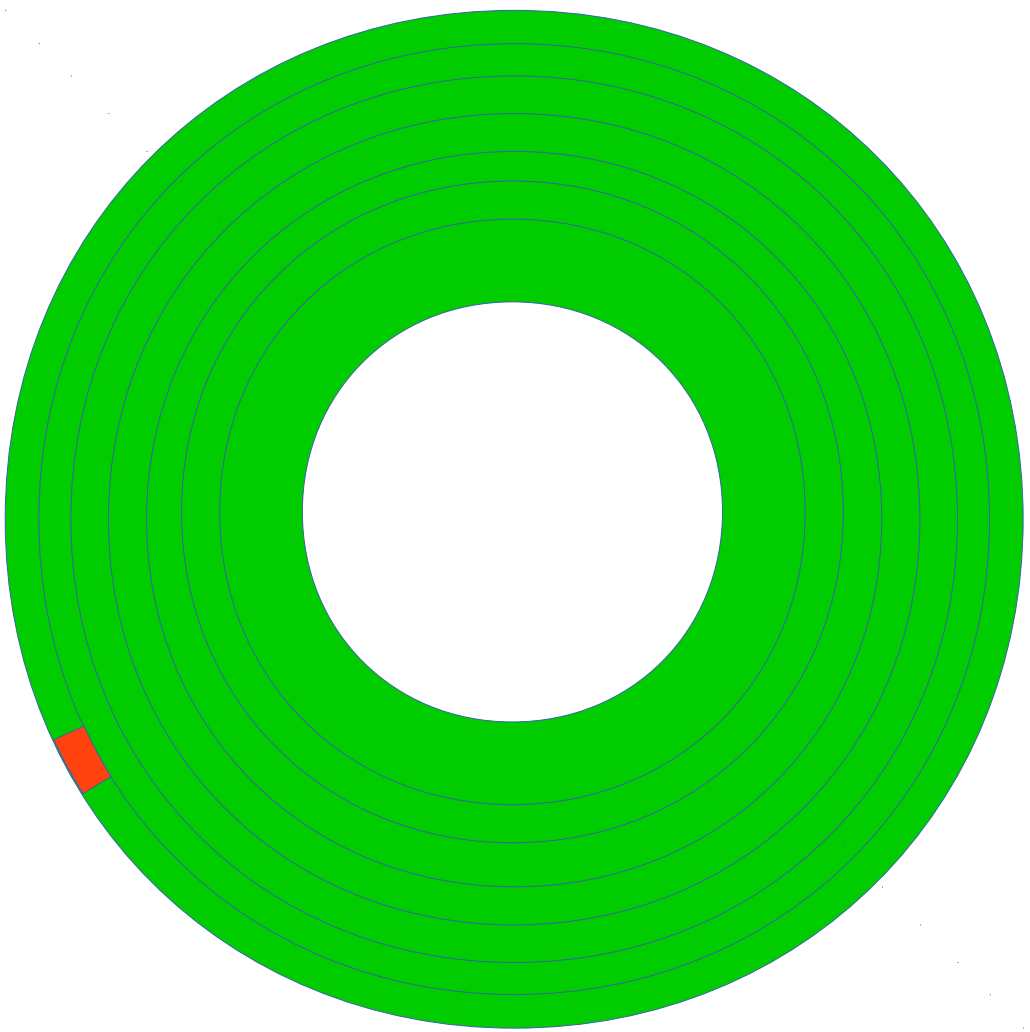




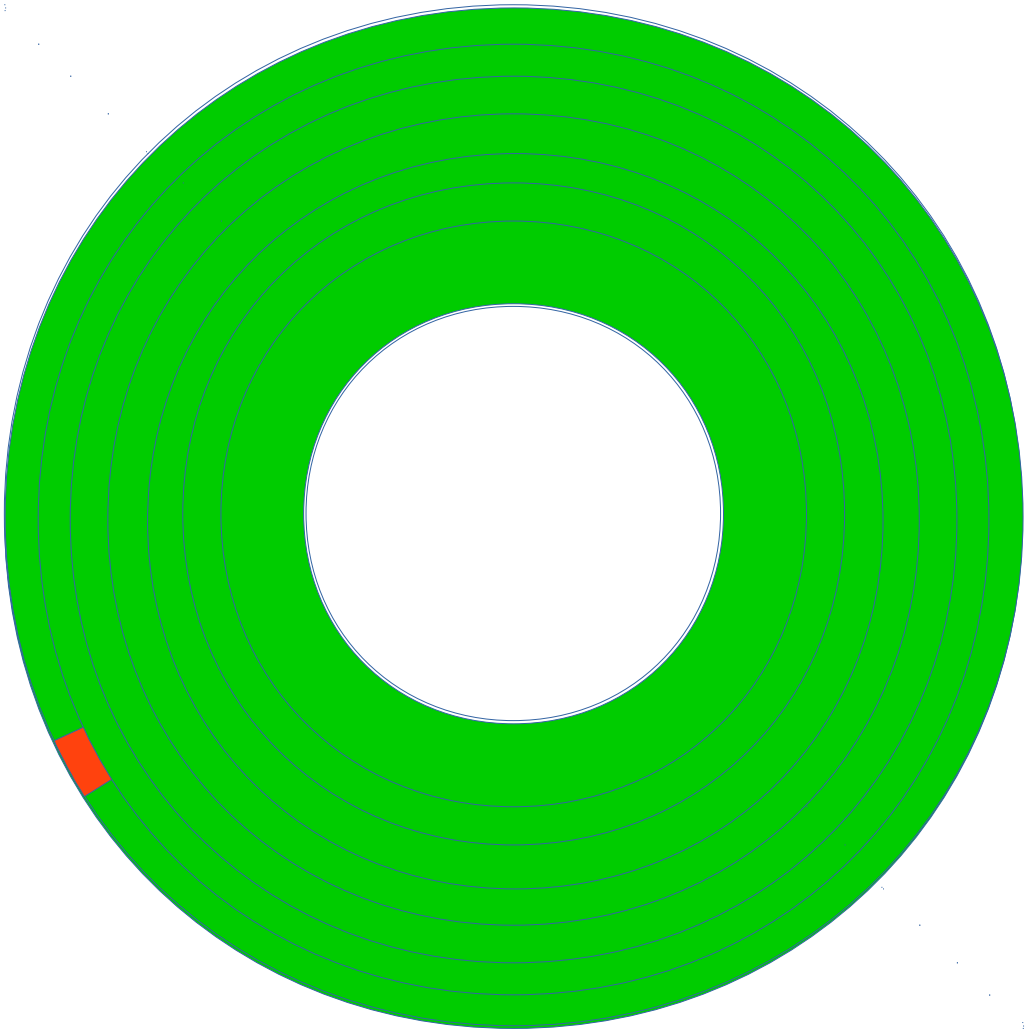
Memory

Platter

Shingled Magnetic Recording

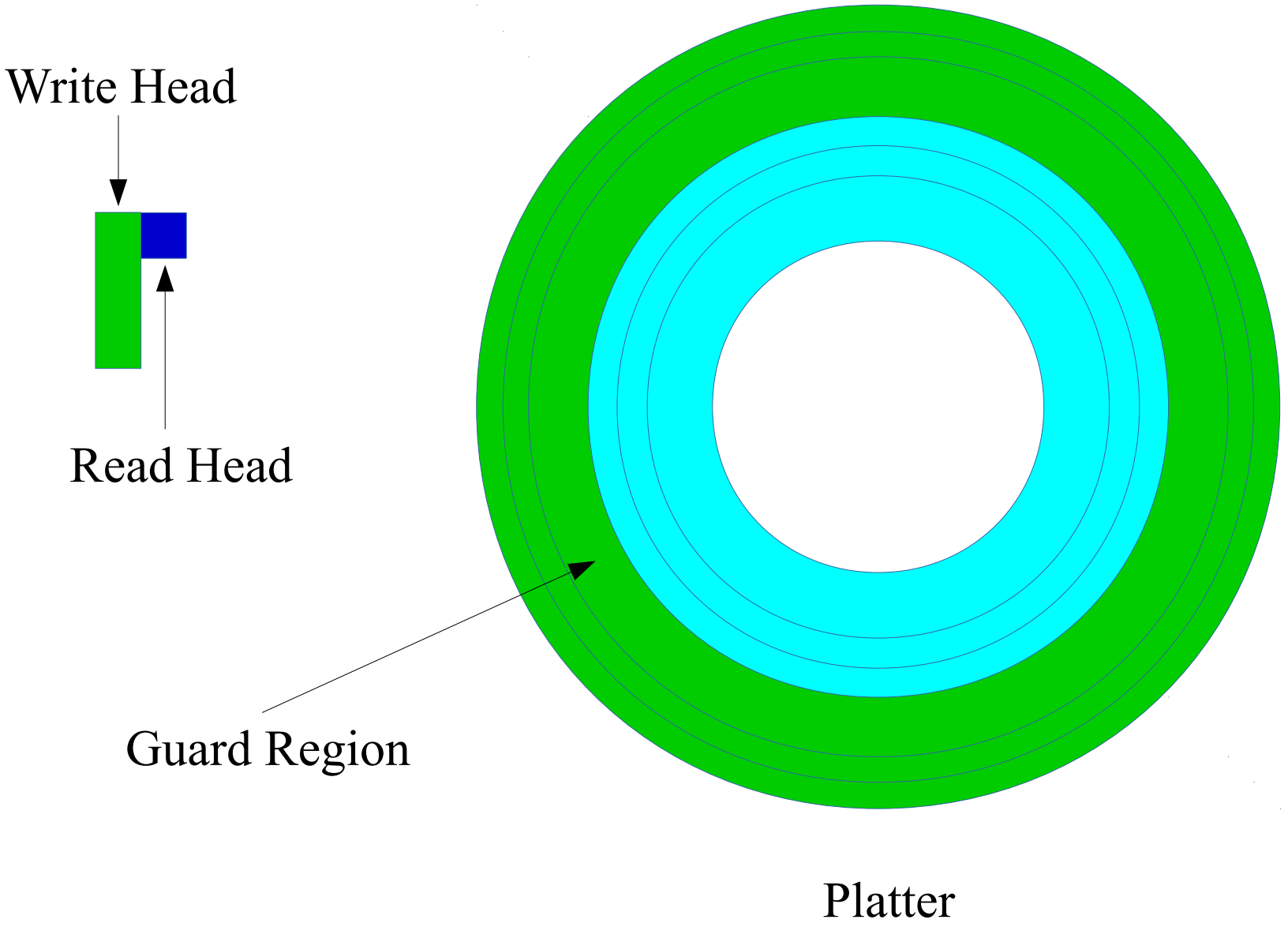


Memory



Platter

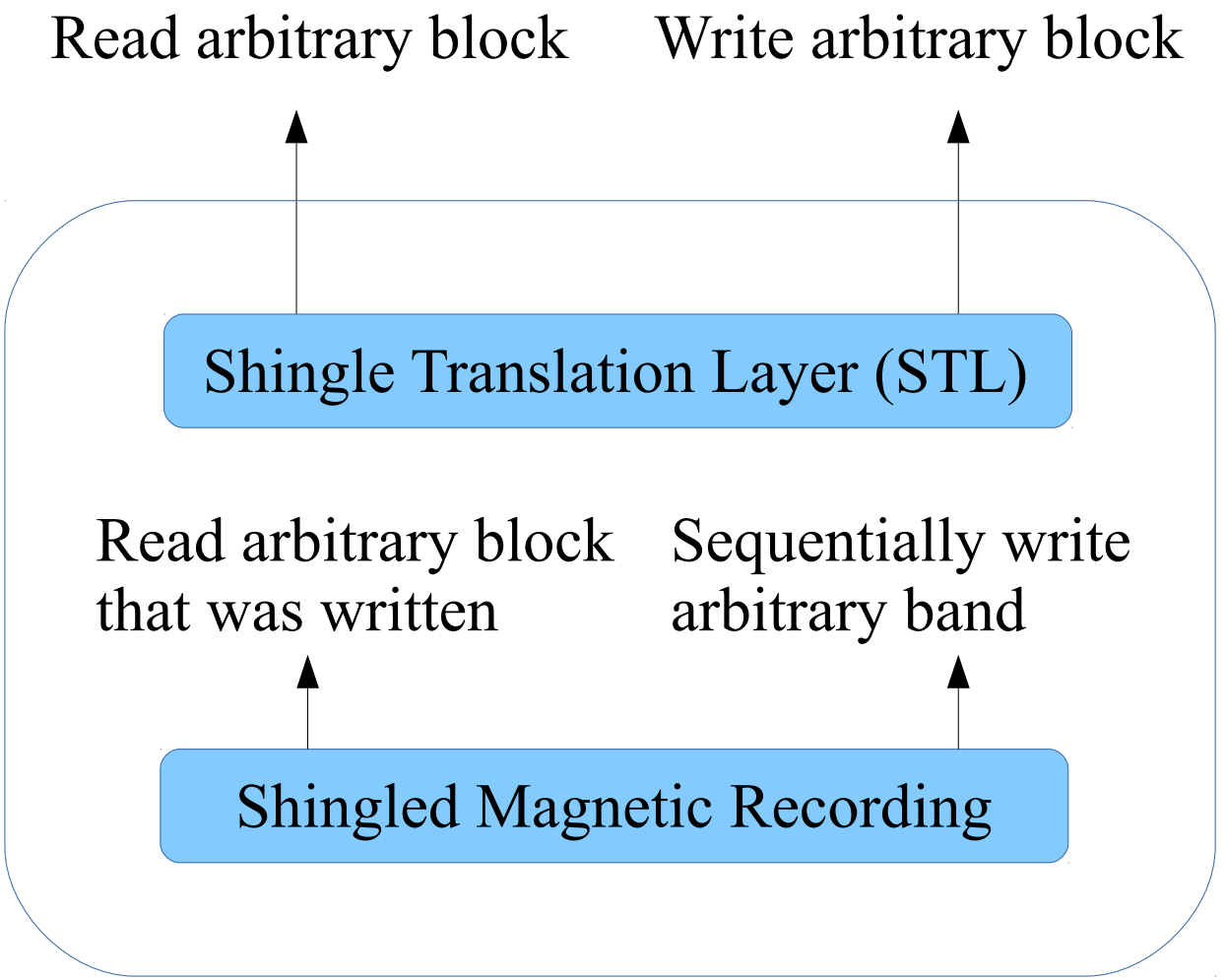
Shingled Magnetic Recording



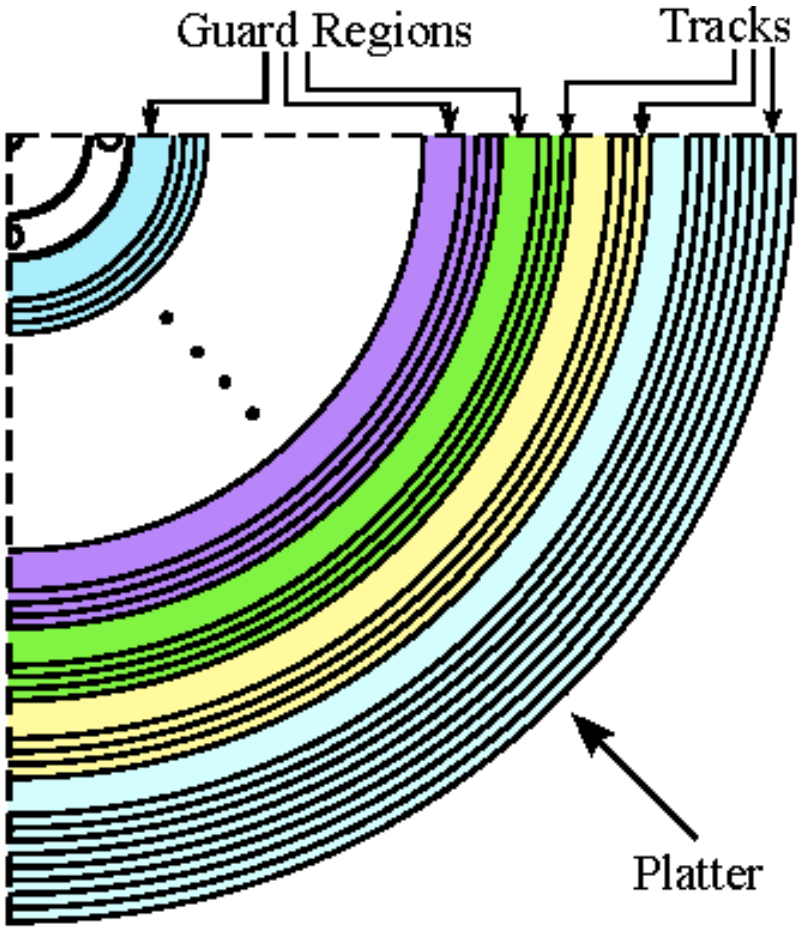
SMR Drive Implementations

- Host-Managed
  - Reports band to host.
  - Bands must be written sequentially.
  - Random writes or reads before writes will fail.
- Host-Aware
  - Reports band to host.
  - Also handles random writes – backward compatible.
- Drive-Managed
  - Hides SMR details.
  - Drop-in replacement for existing drives.

Drive-Managed SMR

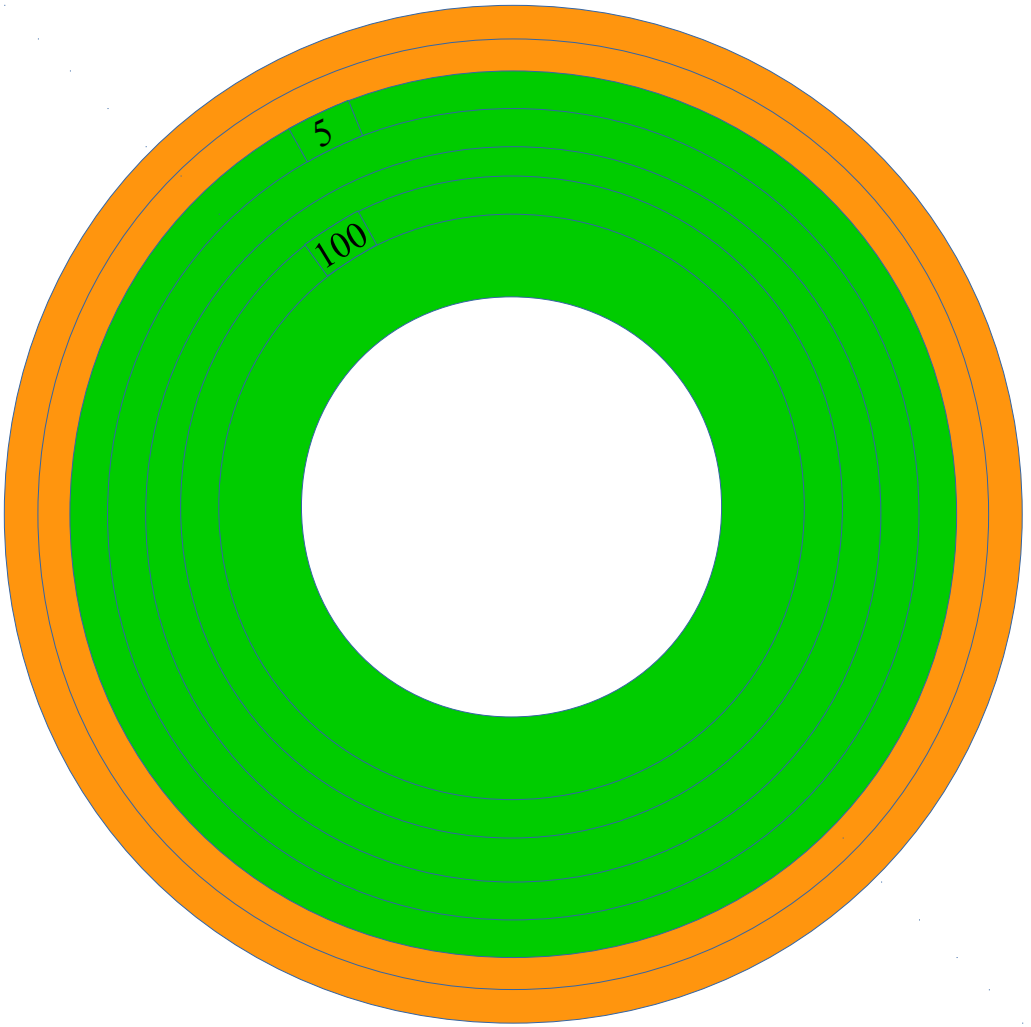


Drive-Managed SMR



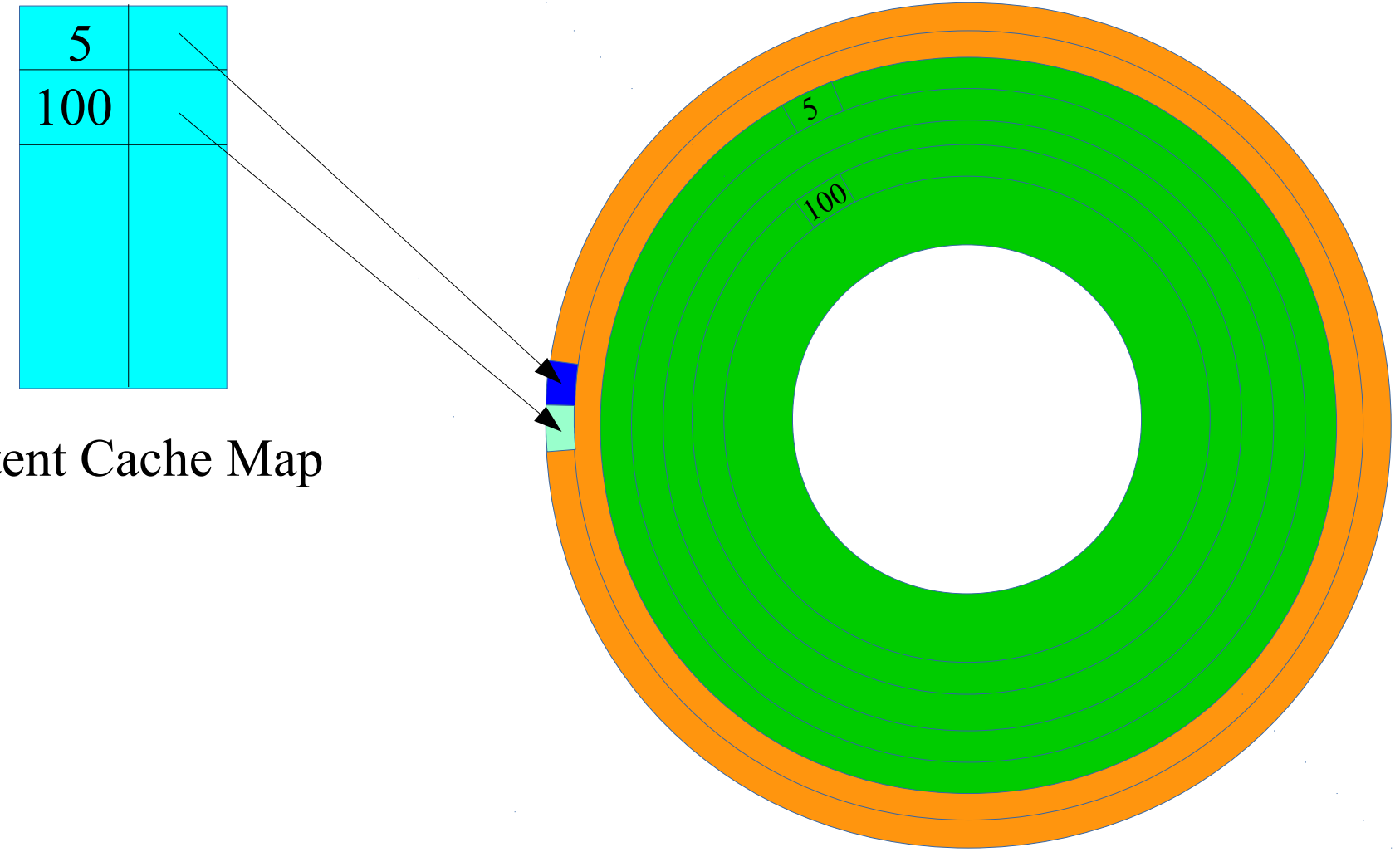
- Small region of disk, called persistent cache, used for staging random writes.
- Other non-volatile memory like flash can also be used for persistent cache.
- Disk is mapped at band granularity; persistent cache uses extent mapping.

Drive-Managed SMR



Bands are shown in green.
Persistent Cache is shown in orange.

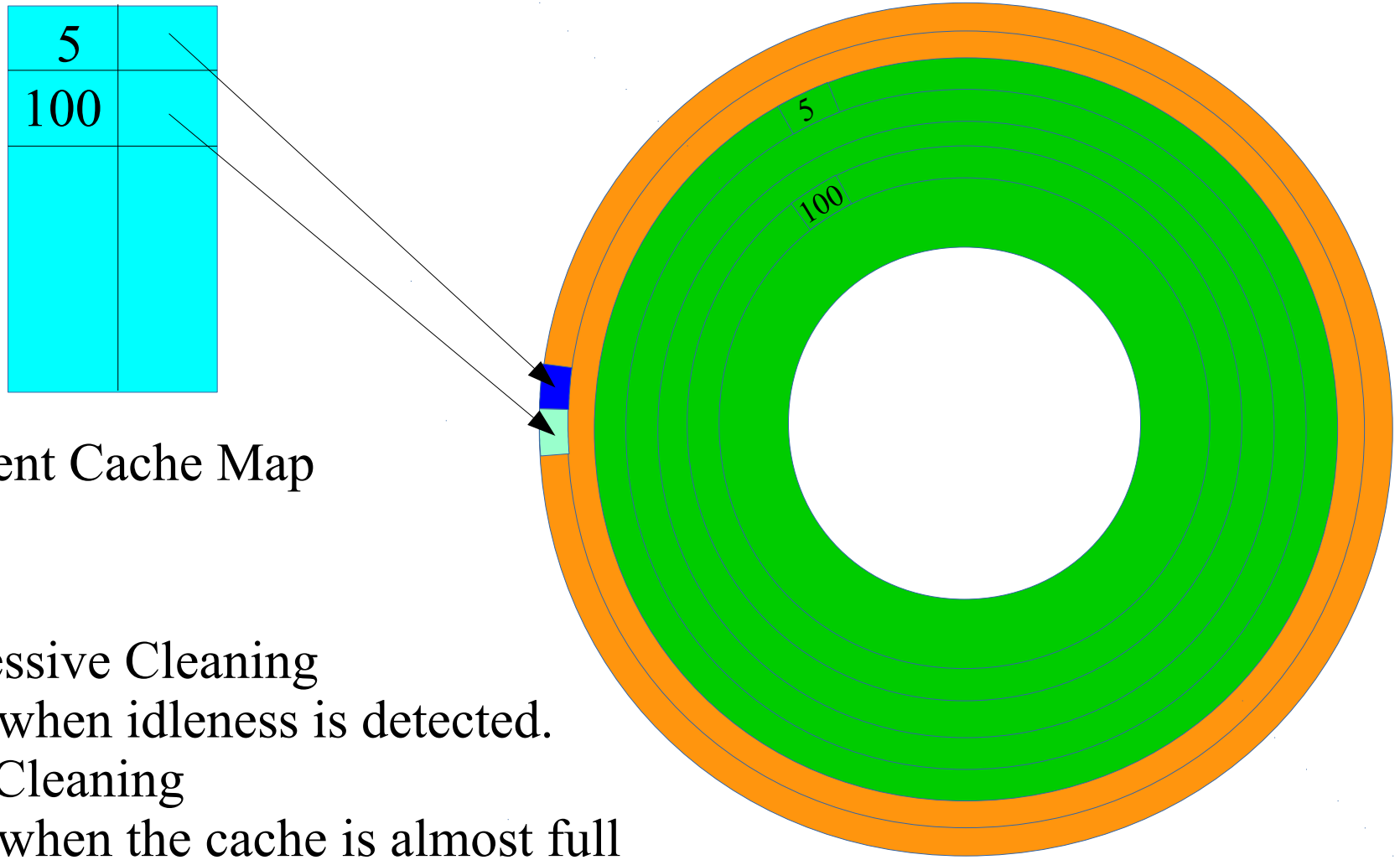
Drive-Managed SMR



Persistent Cache Map

Bands are shown in green.
Persistent Cache is shown in orange.

Drive-Managed SMR



Persistent Cache Map

- Aggressive Cleaning
  starts when idleness is detected.
- Lazy Cleaning
  starts when the cache is almost full

Bands are shown in green.
Persistent Cache is shown in orange.

Outline

- Introduction to SMR
- Characterization goals and test setup
- Test results

Characterization Goals

- Drive Type
- Persistent Cache Type
- Cache Location and Layout
- Cache Size
- Cache Map Size
- Band Size
- Block Mapping
- Cleaning Type
- Cleaning Algorithm
- Band Cleaning Time
- Zone Structure
- Shingling Direction

Skylight Components

- Software part:
  - Launch crafted I/O operations using **`fio`**.
  - Disable kernel read-ahead, drive look-ahead, on-board volatile cache.
  - Use latency to infer drive properties.
- Hardware part:
  - Install a transparent window on the drive.
  - Track the head movements using a high-speed camera.
  - Convert movements to head position graphs.

Emulation Strategy

- STLs from the literature implemented as Linux device-mapper targets.

Drive-Managed SMR with persistent disk cache

Drive-Managed SMR with persistent flash cache

Tested Drives

- Emulated Drives

Drive Name	Cache Type (Size)	Cache Location	Band Size	Capacity
Emulated-SMR-1	Disk (37.2 GB)	Single at ID	40 MiB	3.9 TB
Emulated-SMR-2	Flash (9.9 GB)	N/A	25 MiB	3.9 TB
Emulated-SMR-3	Disk (37.2 GB)	Multiple	20 MiB	3.9 TB

  - All were emulated using a 4TB conventional Seagate drive.
- Real Drives
  - 5TB and 8TB Seagate drive-managed SMR drives.
  - We only show 5TB results – labeled as Seagate-CMR.
- All disk drives are 5900RPM => ~10 ms rotation time.
- Write cache, read-ahead both disabled

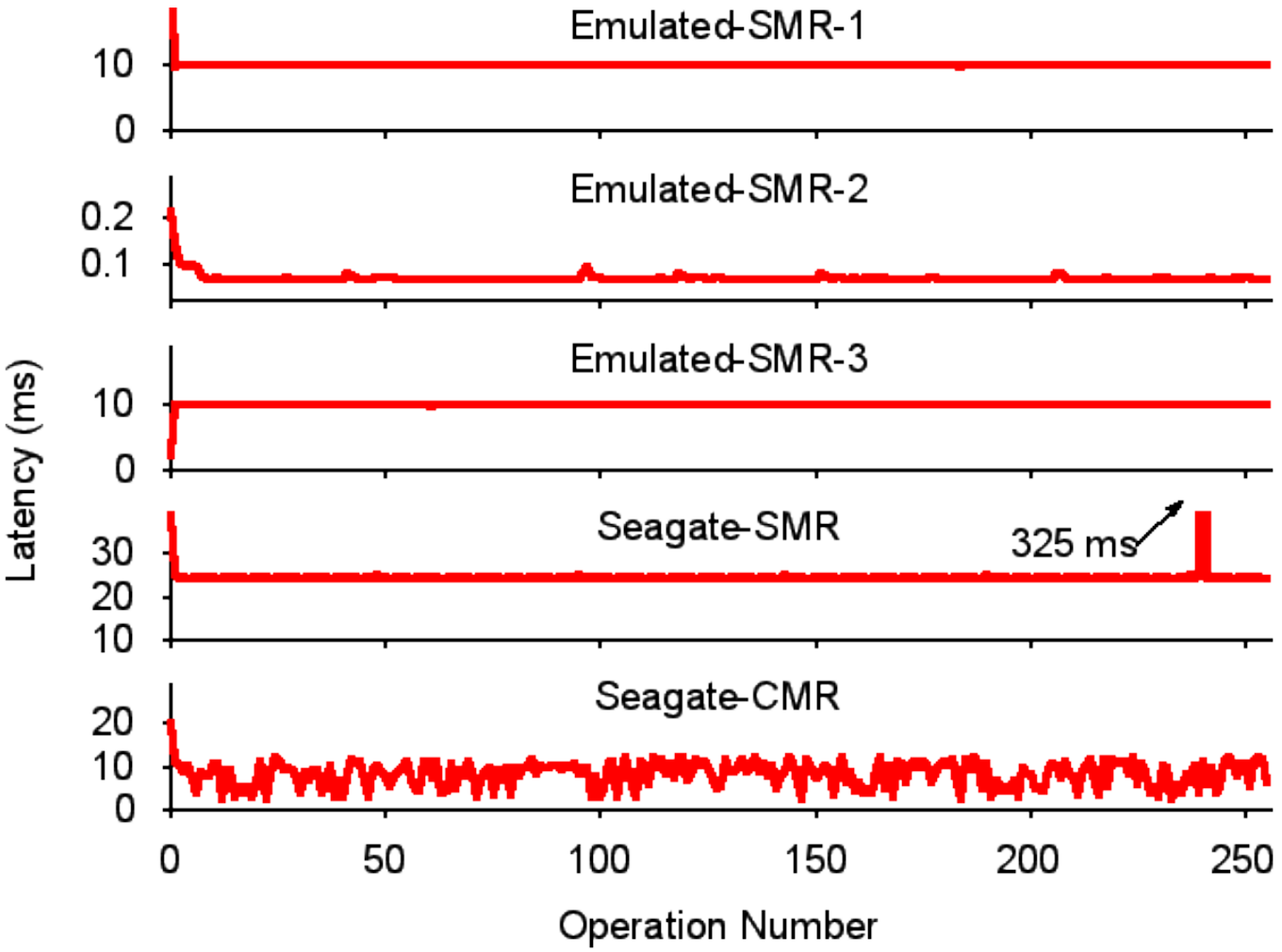
Outline

- Introduction to SMR
- Characterization goals and test setup
- Test results

Test 1: Discovering the drive type and the persistent cache type

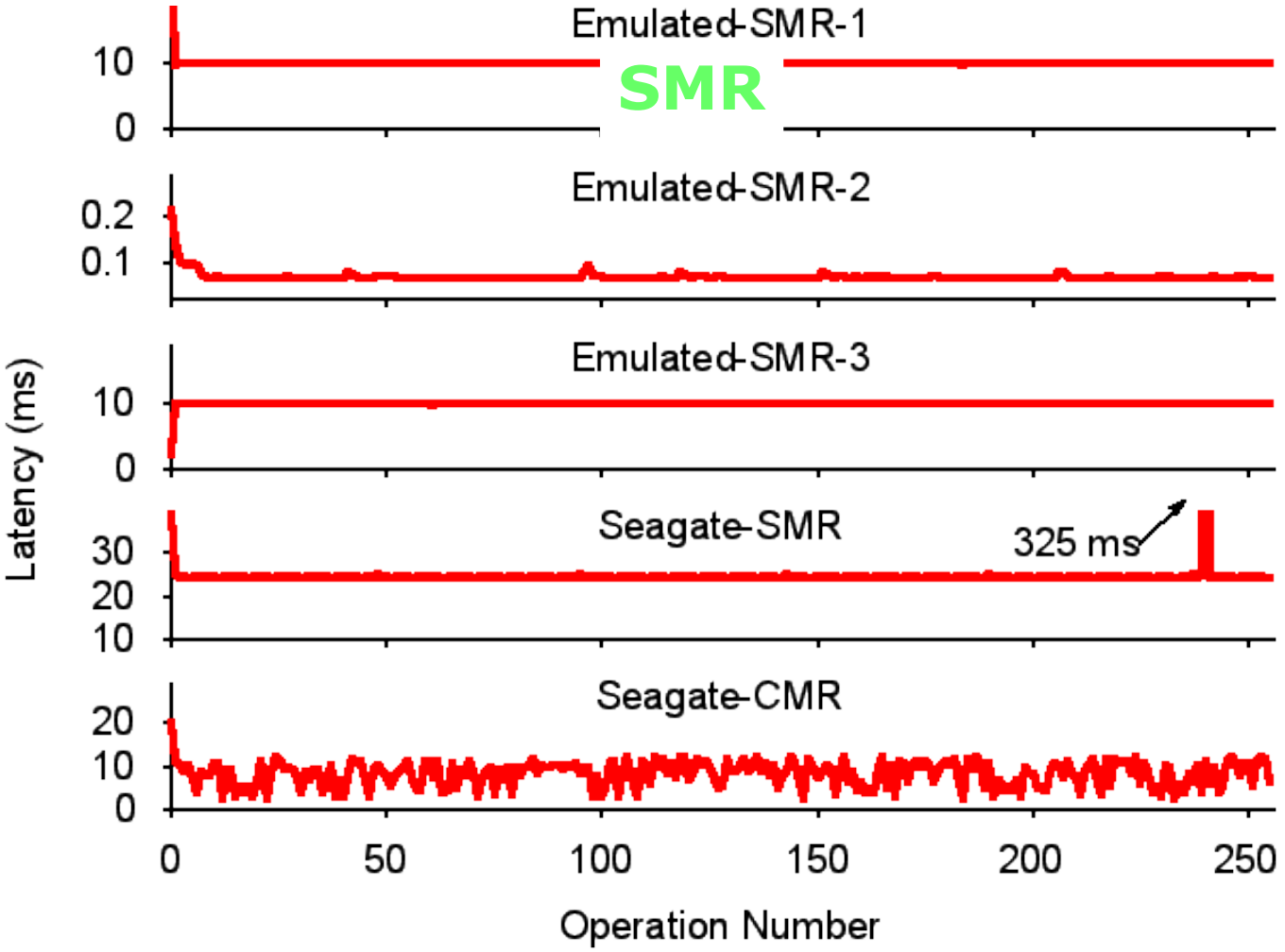
- Test exploits unusual random write behavior in SMR drives.
- Write blocks in the first 1GiB in random order.
- If latency is fixed then the drive is SMR, otherwise it is a conventional magnetic recording (CMR).
- Sub-millisecond latency indicates a drive with a persistent flash cache.

Random Write latency



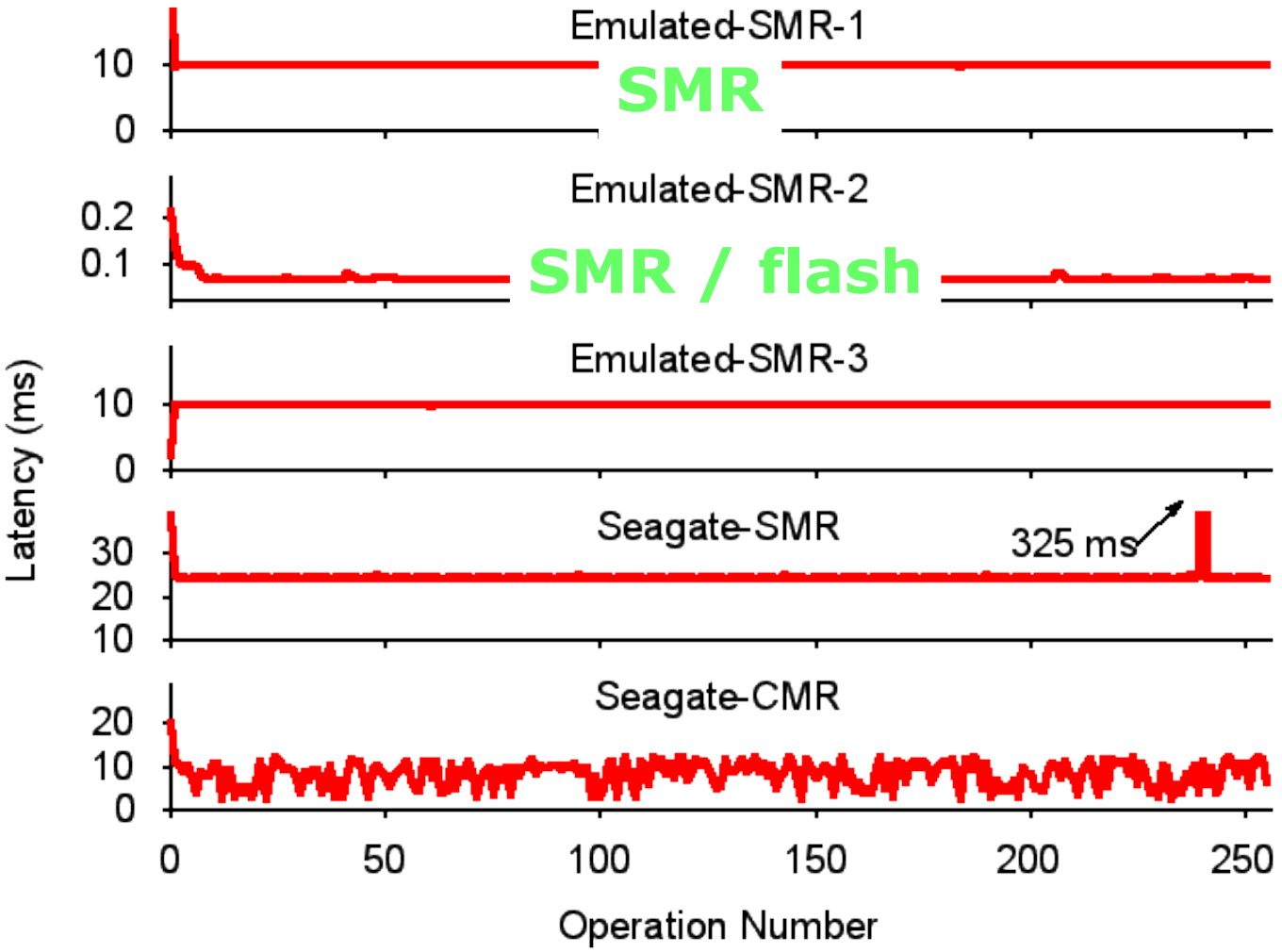
- Y-axis varies in each graph.
- Conventional drive (Seagate-CMR) stands out from the rest.
- Emulated drive with persistent flash cache has sub-ms latency.
- Latency is high for the real SMR drive.

Random Write latency



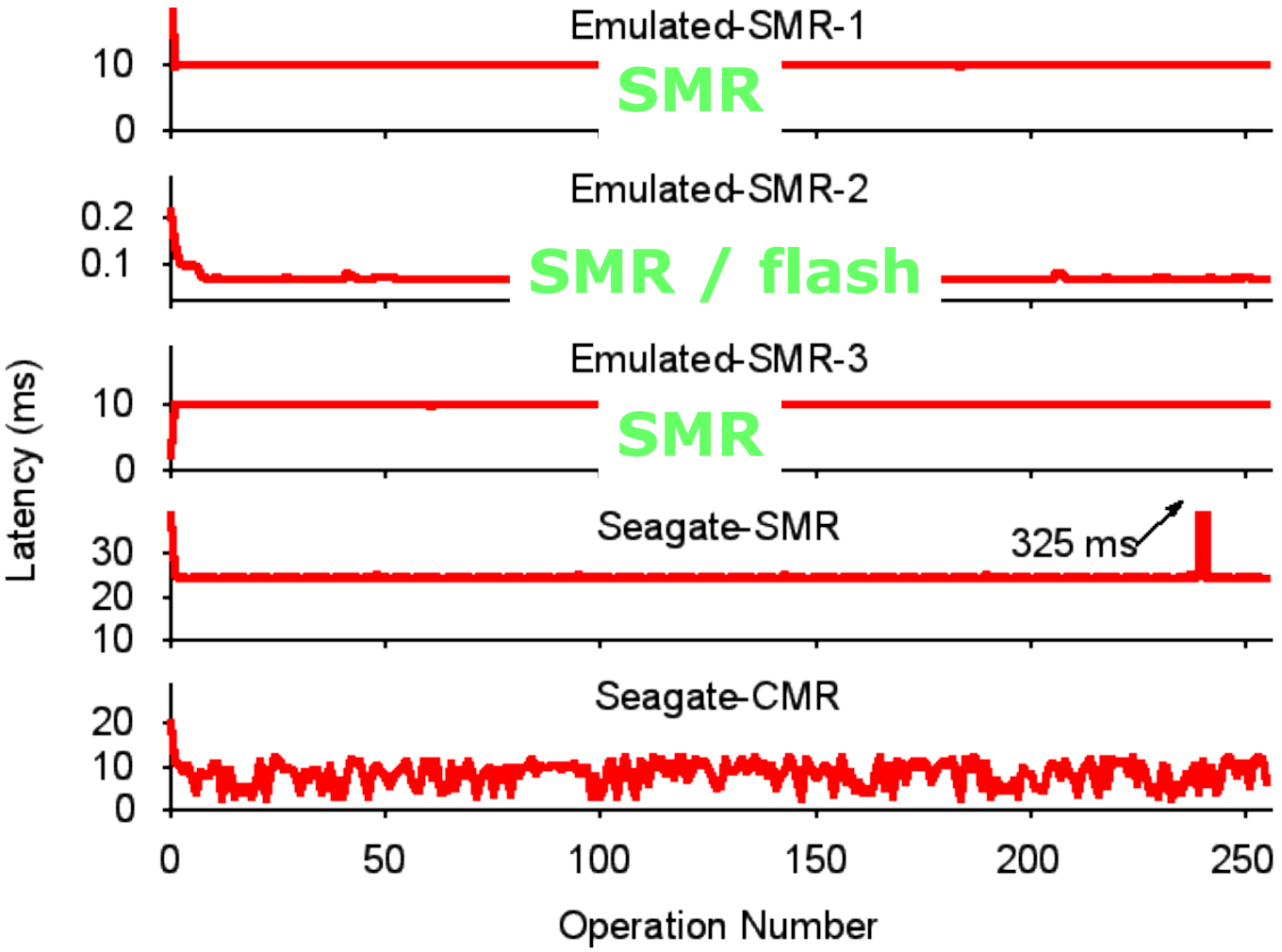
- Y-axis varies in each graph.
- Conventional drive (Seagate-CMR) stands out from the rest.
- Emulated drive with persistent flash cache has sub-ms latency.
- Latency is high for the real SMR drive.

Random Write latency



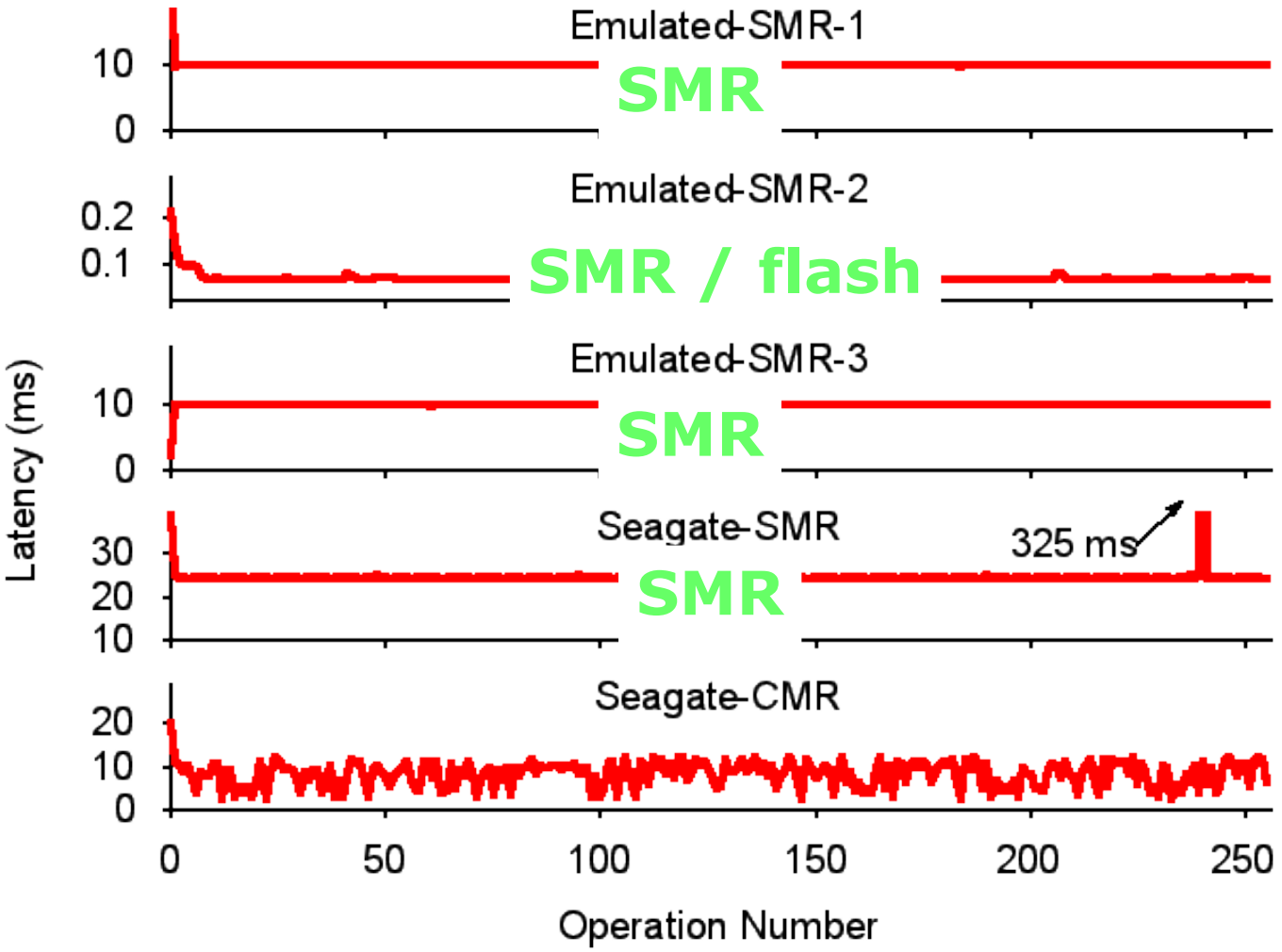
- Y-axis varies in each graph.
- Conventional drive (Seagate-CMR) stands out from the rest.
- Emulated drive with persistent flash cache has sub-ms latency.
- Latency is high for the real SMR drive.

Random Write latency



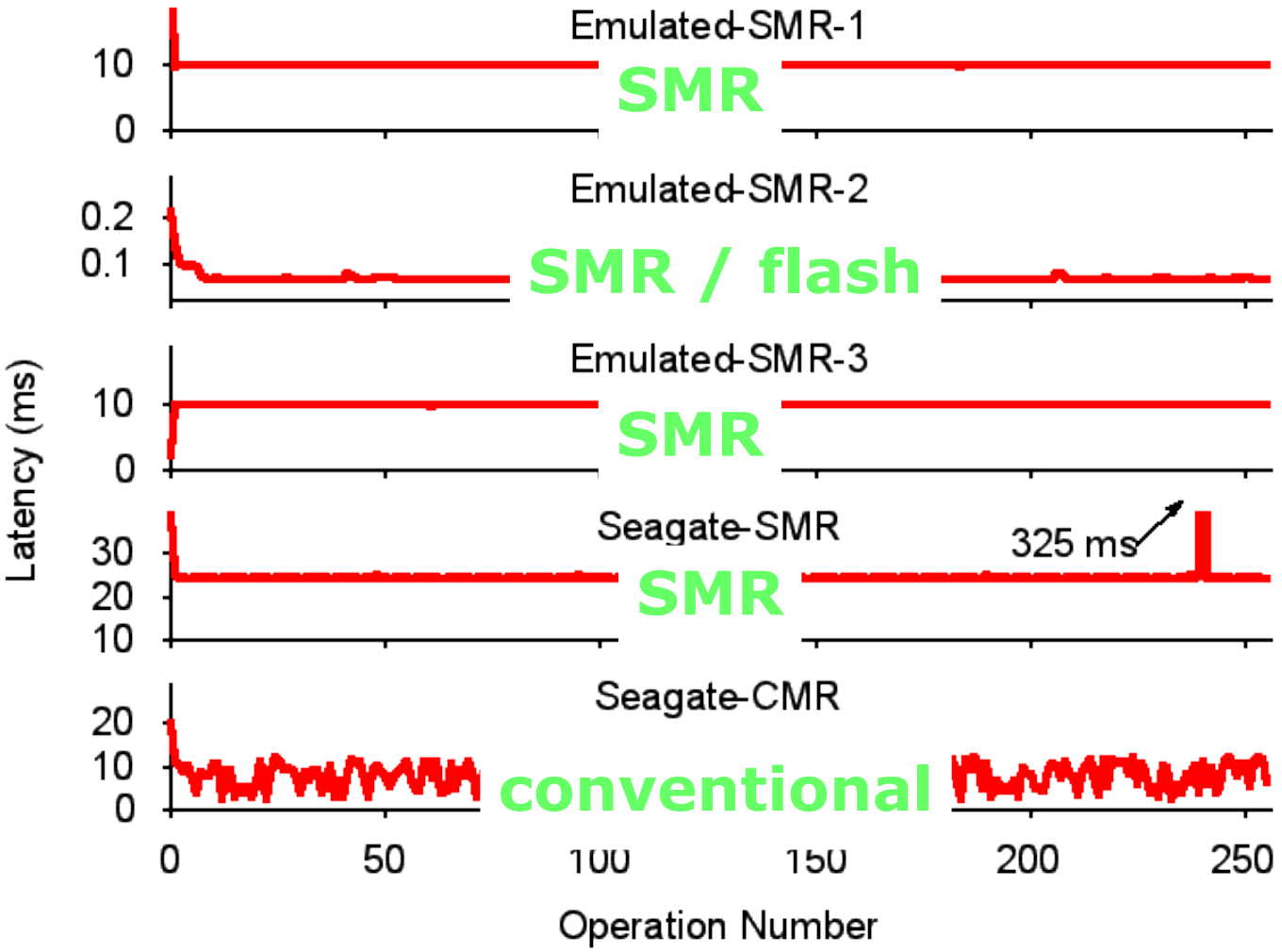
- Y-axis varies in each graph.
- Conventional drive (Seagate-CMR) stands out from the rest.
- Emulated drive with persistent flash cache has sub-ms latency.
- Latency is high for the real SMR drive.

Random Write latency



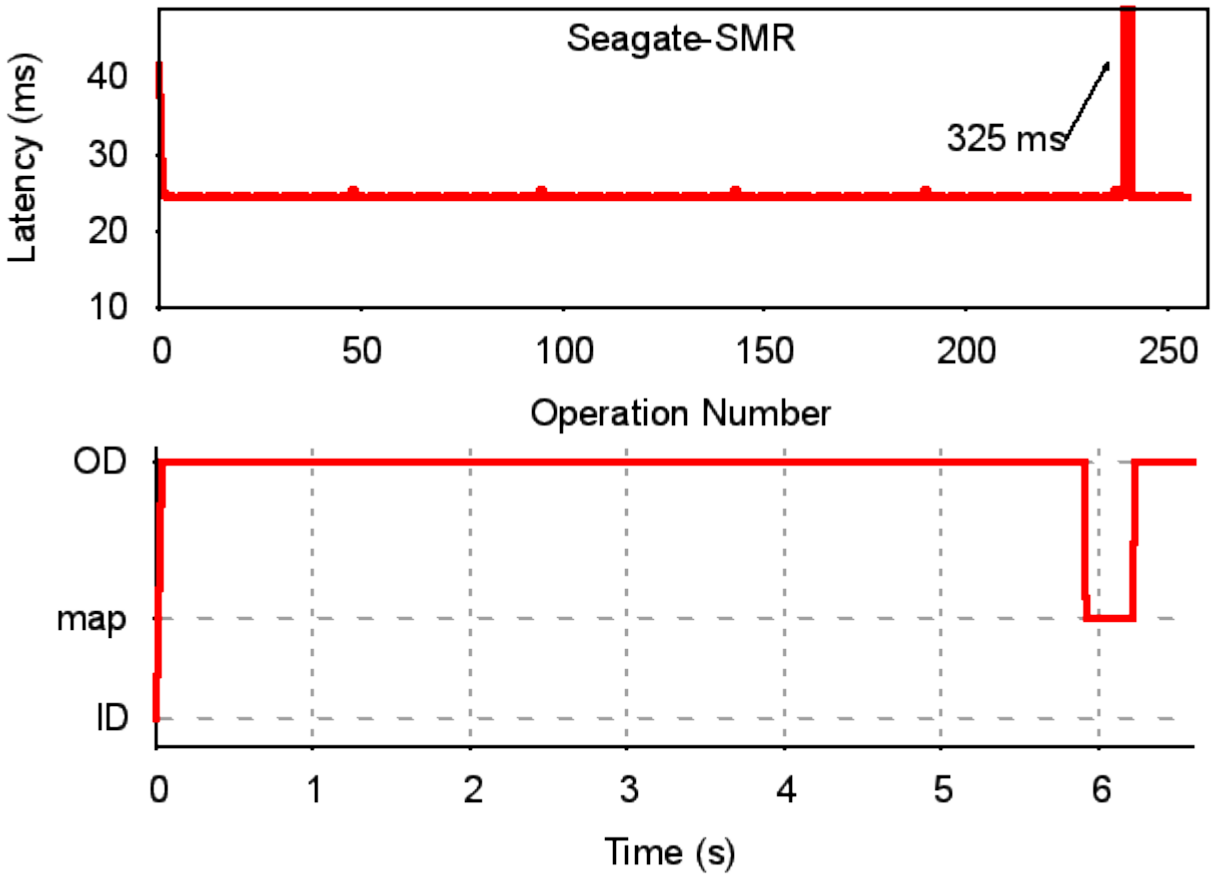
- Y-axis varies in each graph.
- Conventional drive (Seagate-CMR) stands out from the rest.
- Emulated drive with persistent flash cache has sub-ms latency.
- Latency is high for the real SMR drive.

Random Write latency



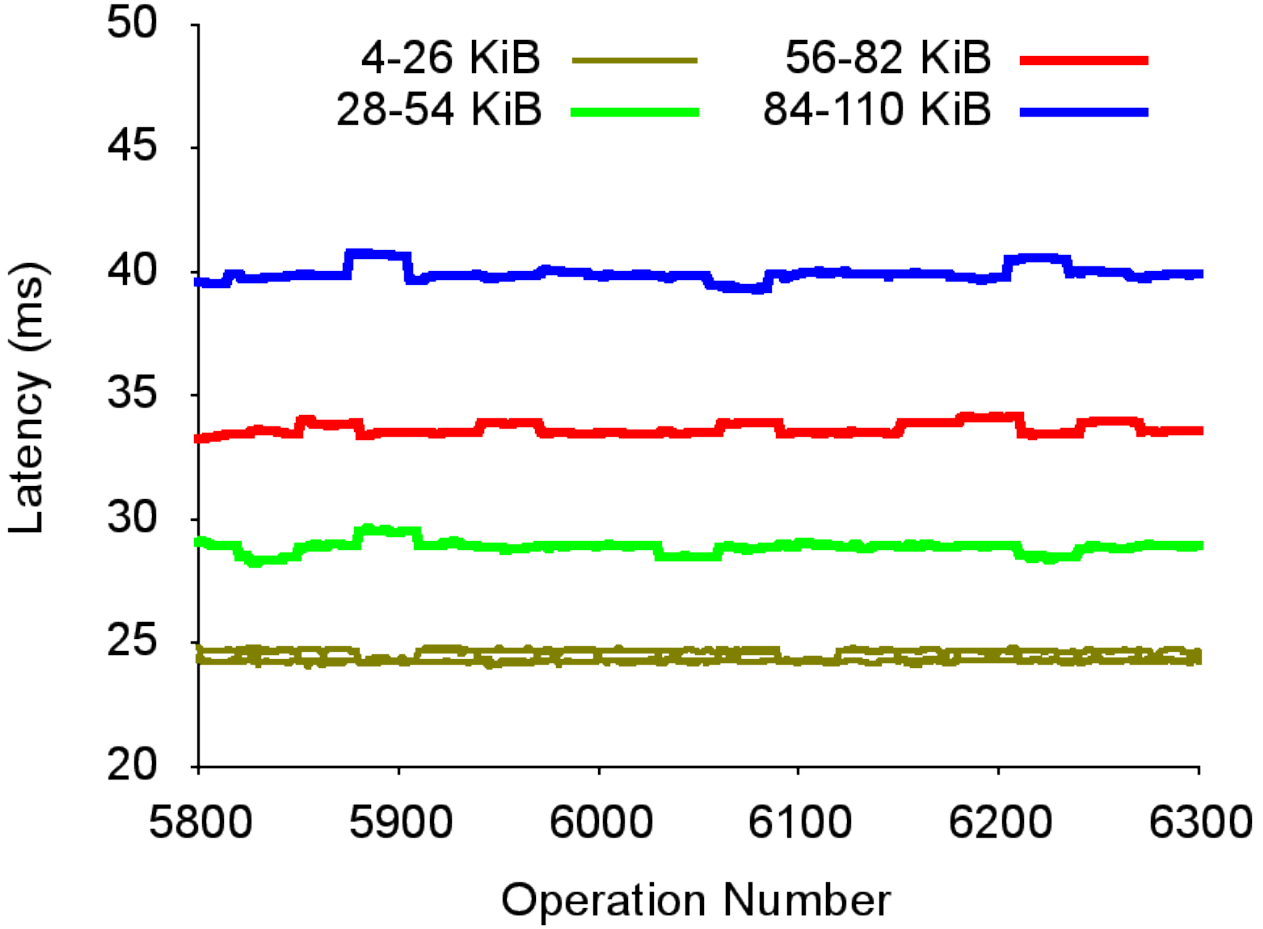
- Y-axis varies in each graph.
- Conventional drive (Seagate-CMR) stands out from the rest.
- Emulated drive with persistent flash cache has sub-ms latency.
- Latency is high for the real SMR drive.

Random Write Latency + Head Position



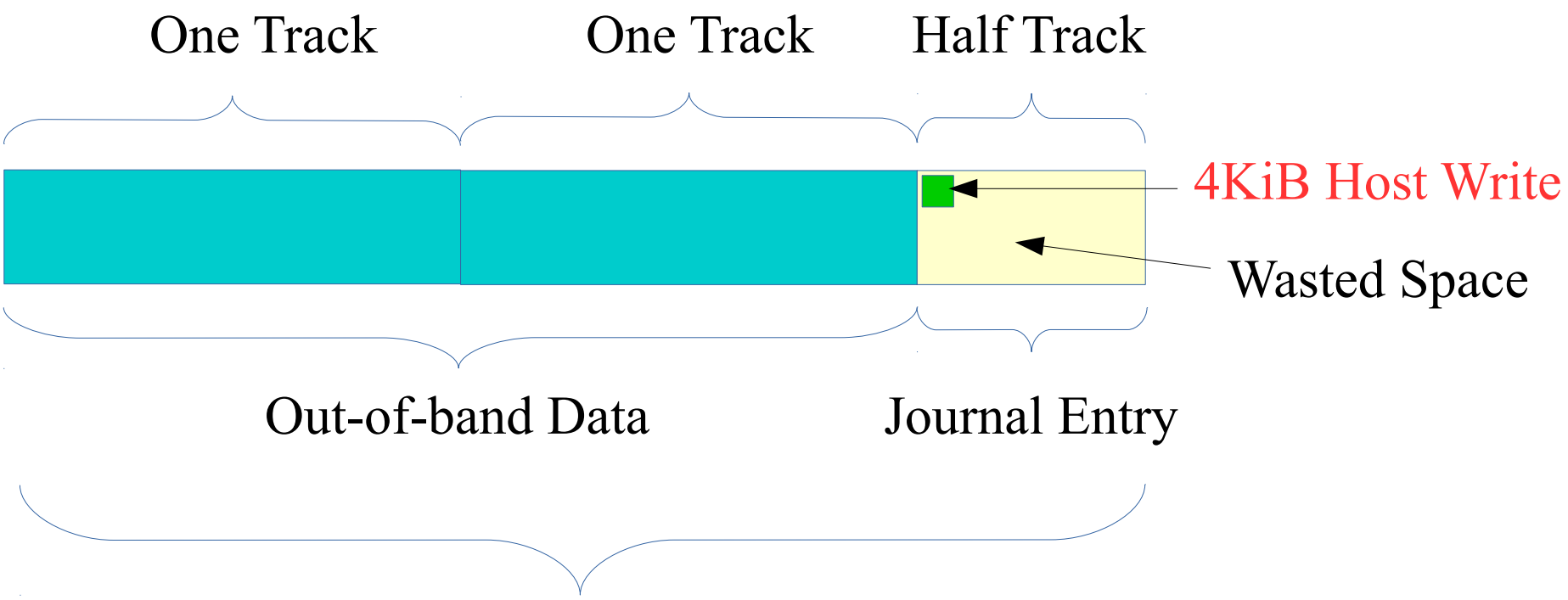
- There is a persistent cache at the outer diameter (OD).
- Writes are (likely) piggy backed with out-of-band data.
- There is (likely) a persistent map stored at the middle diameter.

Random Writes with Max Queue Depth



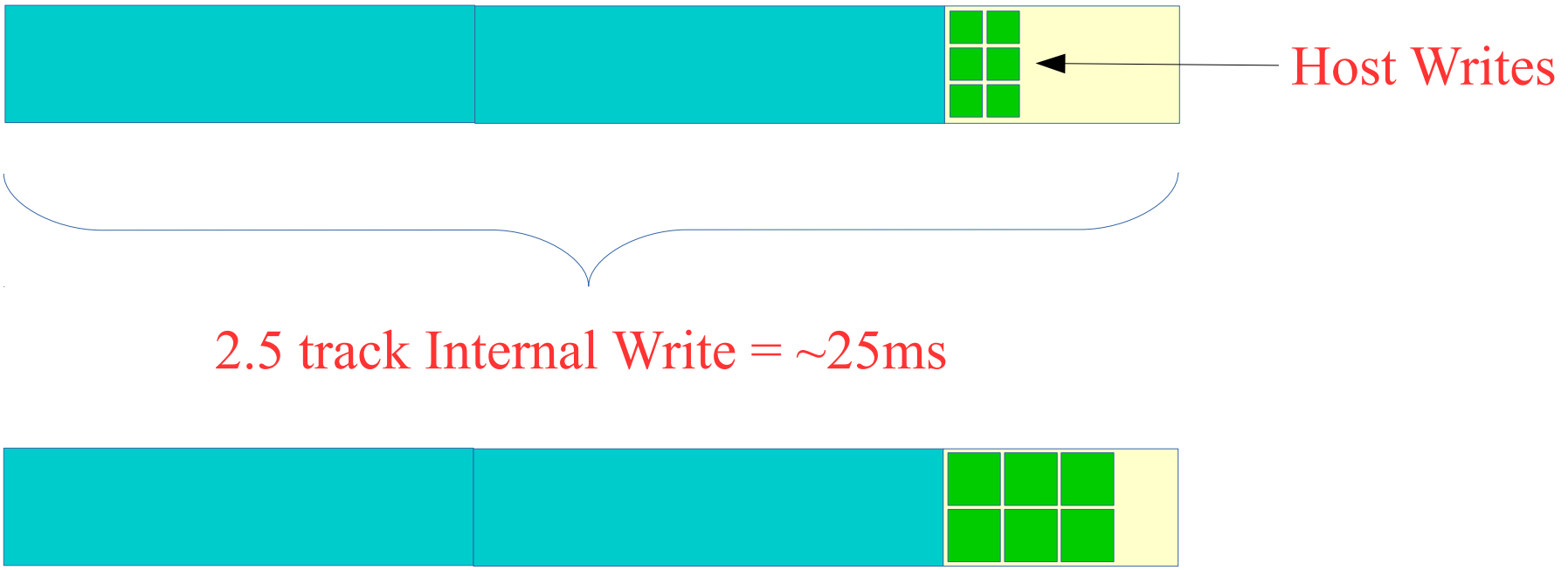
- Different write sizes produce equal latencies.
- Latency increases in ~5ms jumps.
- Given ~10ms rotation time, ~5ms is ~ half-track increase in write size.

Host Write vs Internal Write



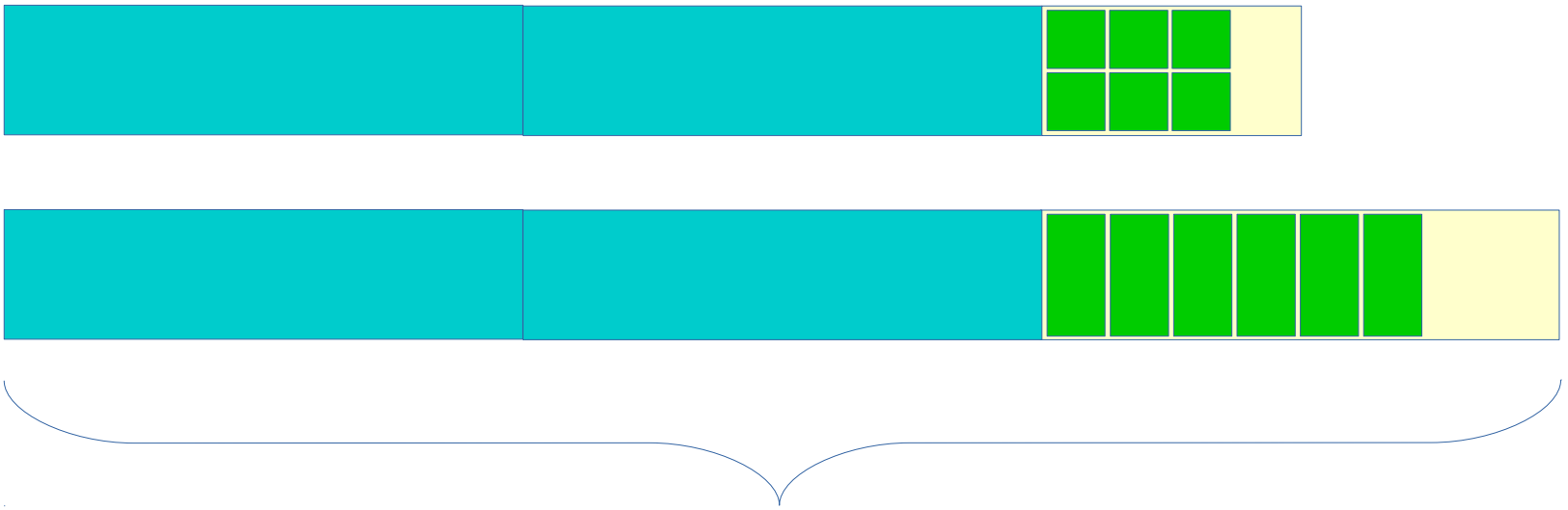
Journal Entries with Quantized Sizes

Journal Entries with Quantized Sizes



Journal Entries with Quantized Sizes

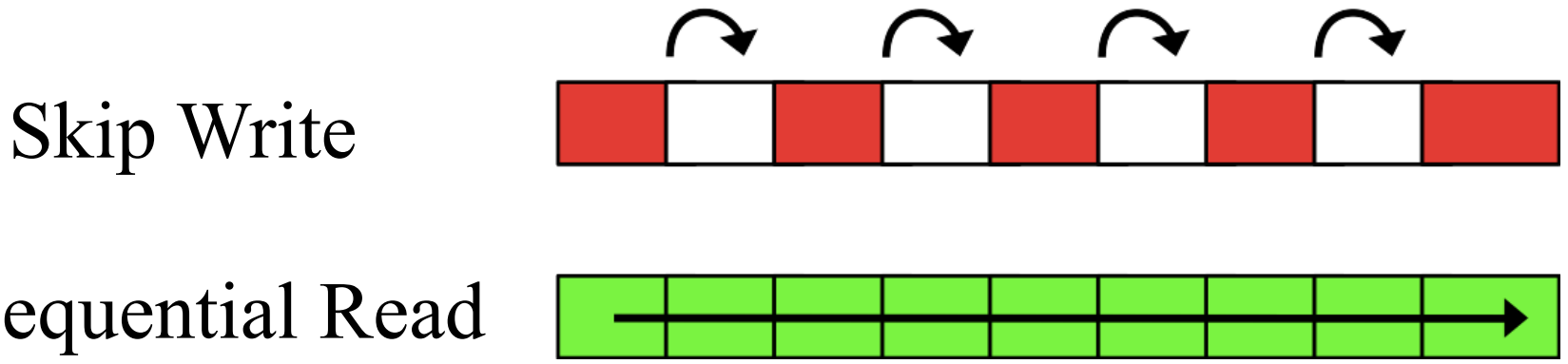
2.5 track Internal Write = ~25ms



3 track Internal Write = ~30ms

Test 2: Discovering Disk Cache Location and Structure

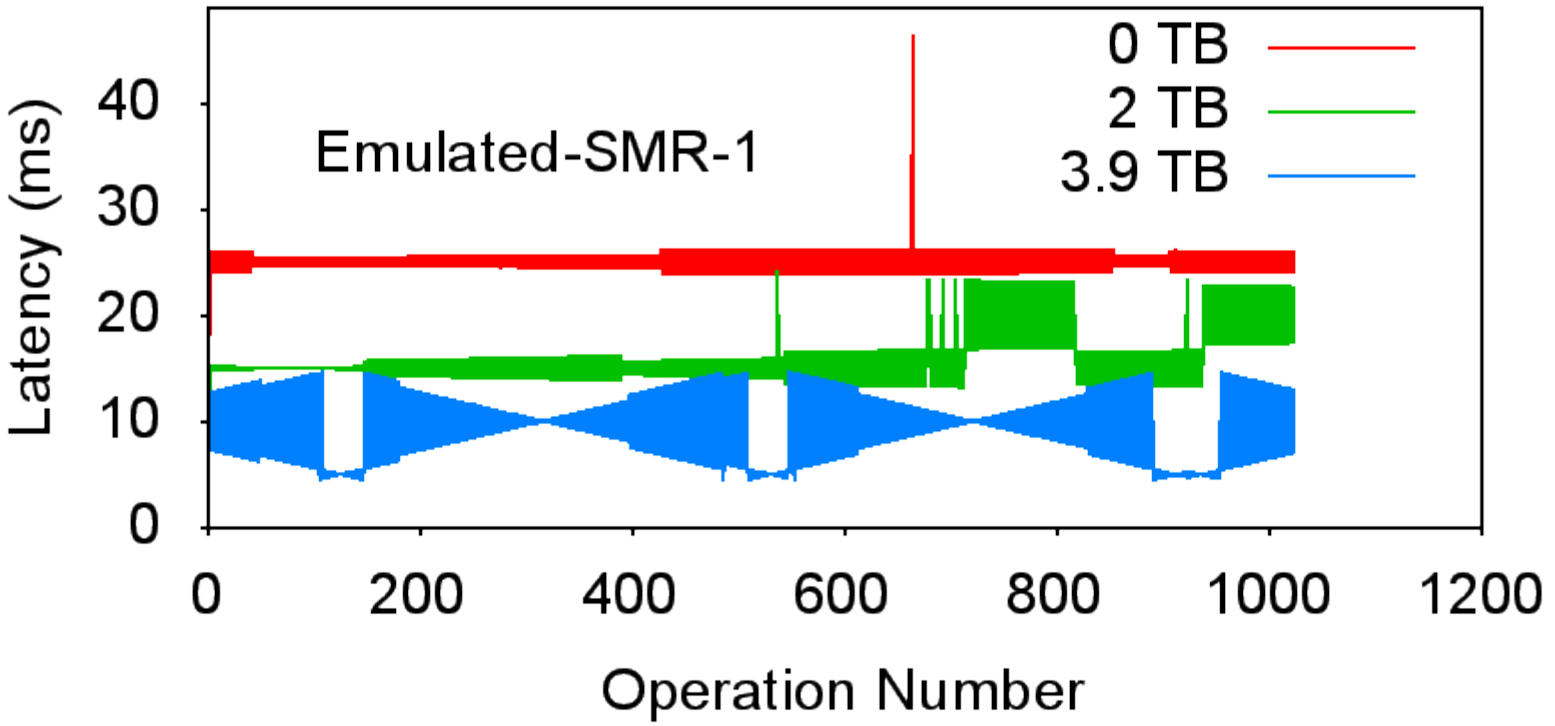
- Test exploits a phenomenon called “fragmented reads”.
- Fragmented read: during sequential read, seek to the persistent cache and back to read an updated block.
- Force fragmented reads at different offsets to infer persistent cache location based on seek time.



Fragmented Read at 5TB Offset

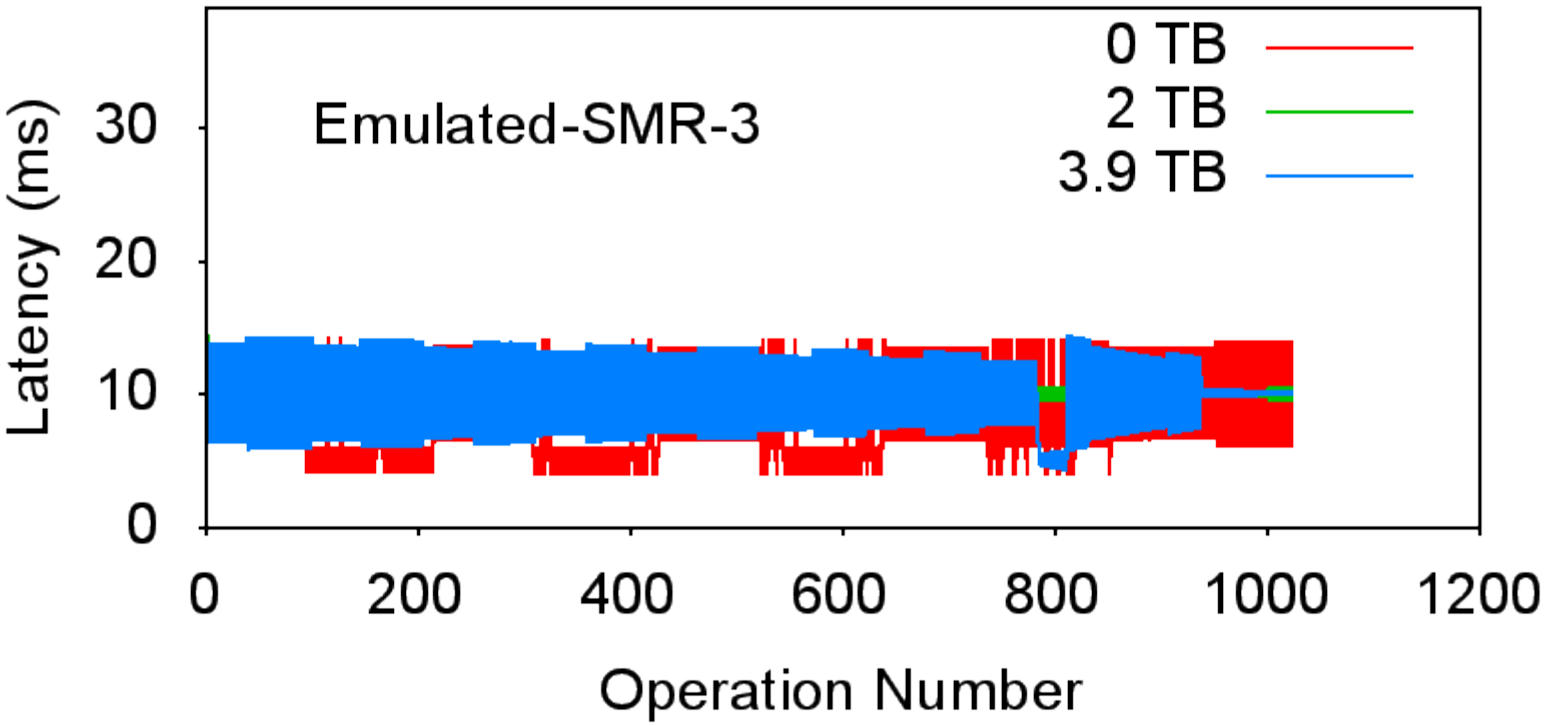
- Head seeks back and forth between a track and persistent cache.
- Persistent Cache is at OD, therefore, 5TB offset is at ID.
- Block numbering convention starts at OD proceeds towards ID.

Fragmented Read Latency at Low, Middle, and High Offsets



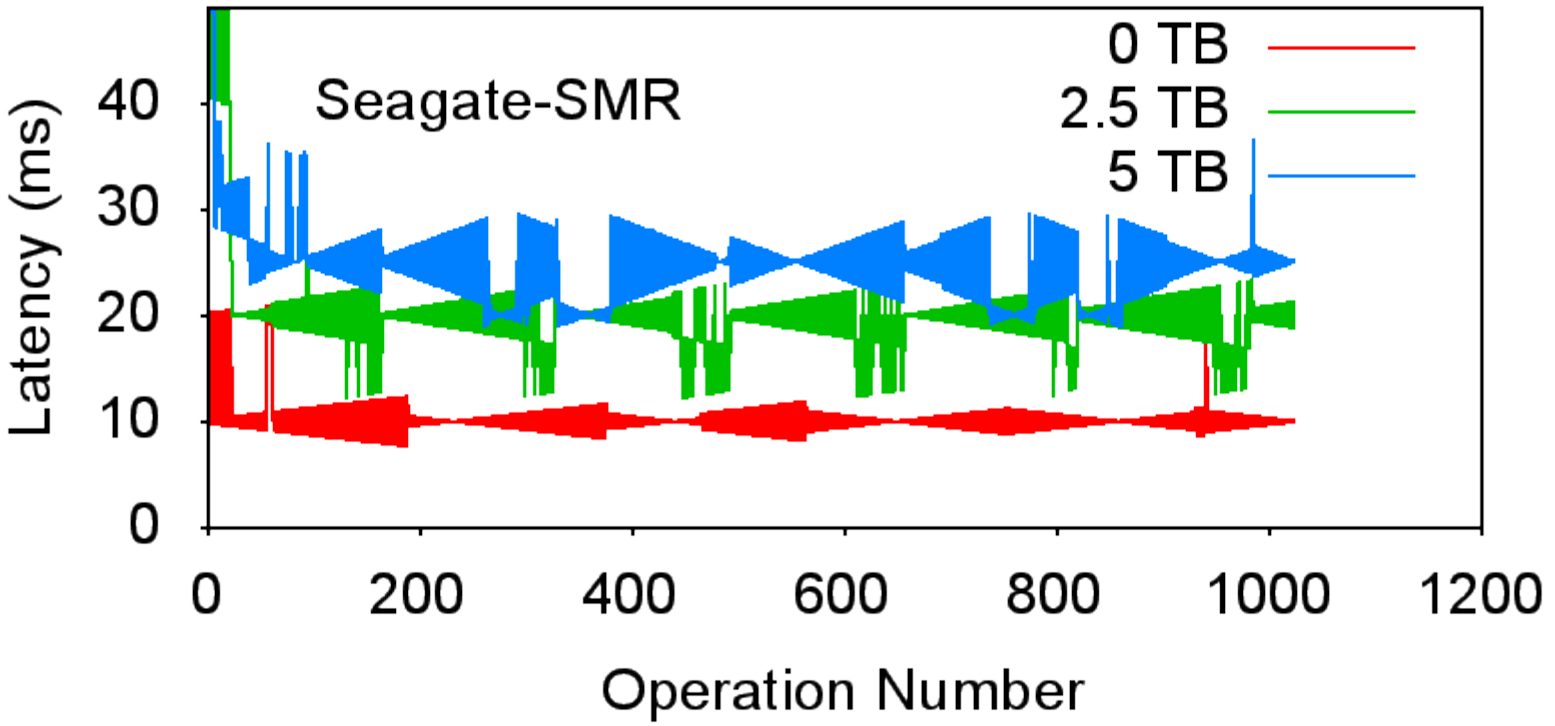
- Average latency high at low offset => cache at ID..

Fragmented Read Latency at Low, Middle, and High Offsets



- Average latency is roughly fixed => distributed cache.

Fragmented Read Latency at Low, Middle, and High Offsets



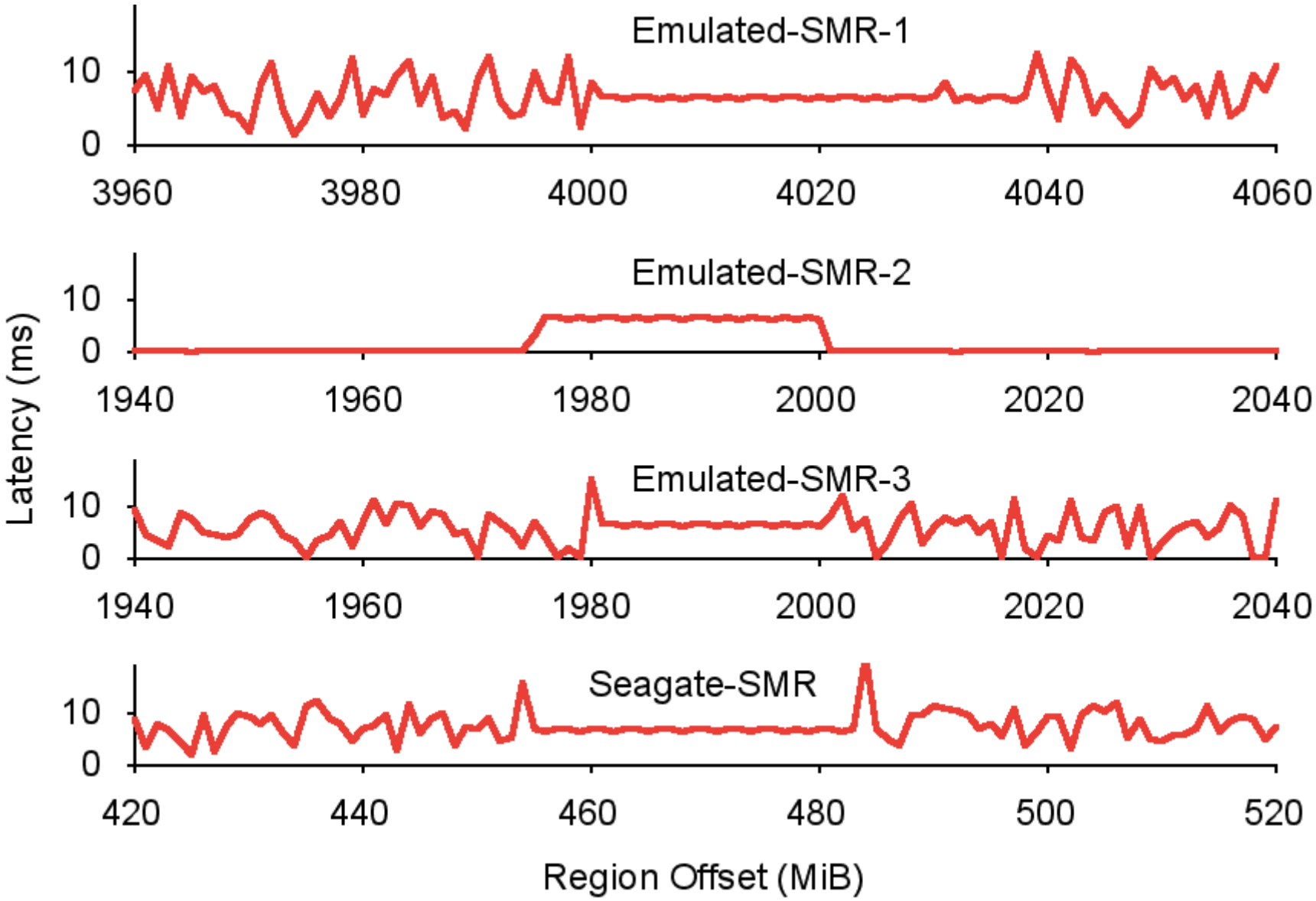
- Average latency is high at high offset => cache at OD.

Test 3: Discovering the Band Size

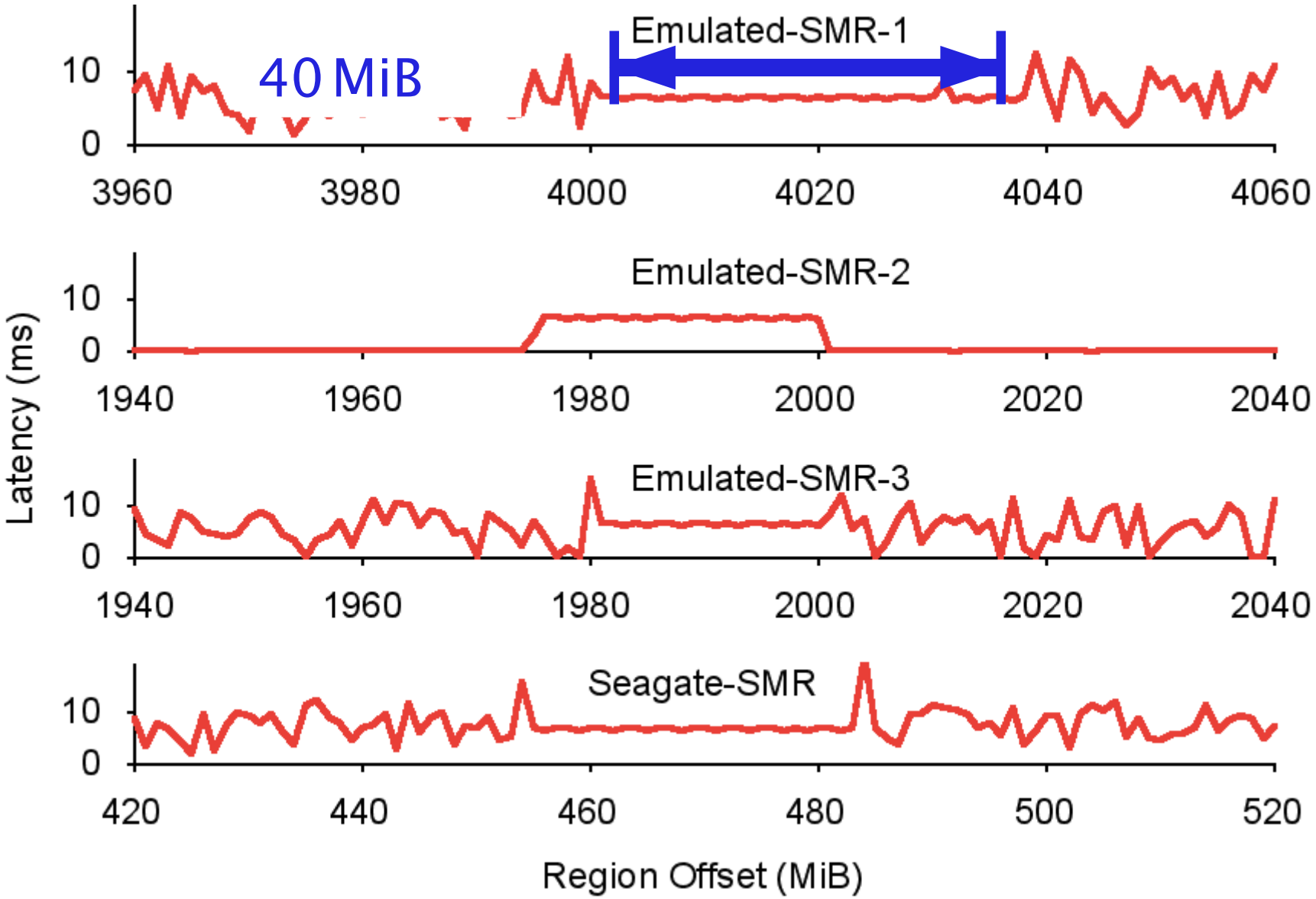
- Test relies on the fact that cleaning proceeds at a band granularity.
- Choose a small region (~1GiB) and write blocks in random order.
- Pause for a short (~3-5s) period, letting the cleaner to clean a few bands.
- Sequentially read the blocks in the region.
- Most latencies will be random – a streak of flat latencies will identify a band.

Band Size Detection

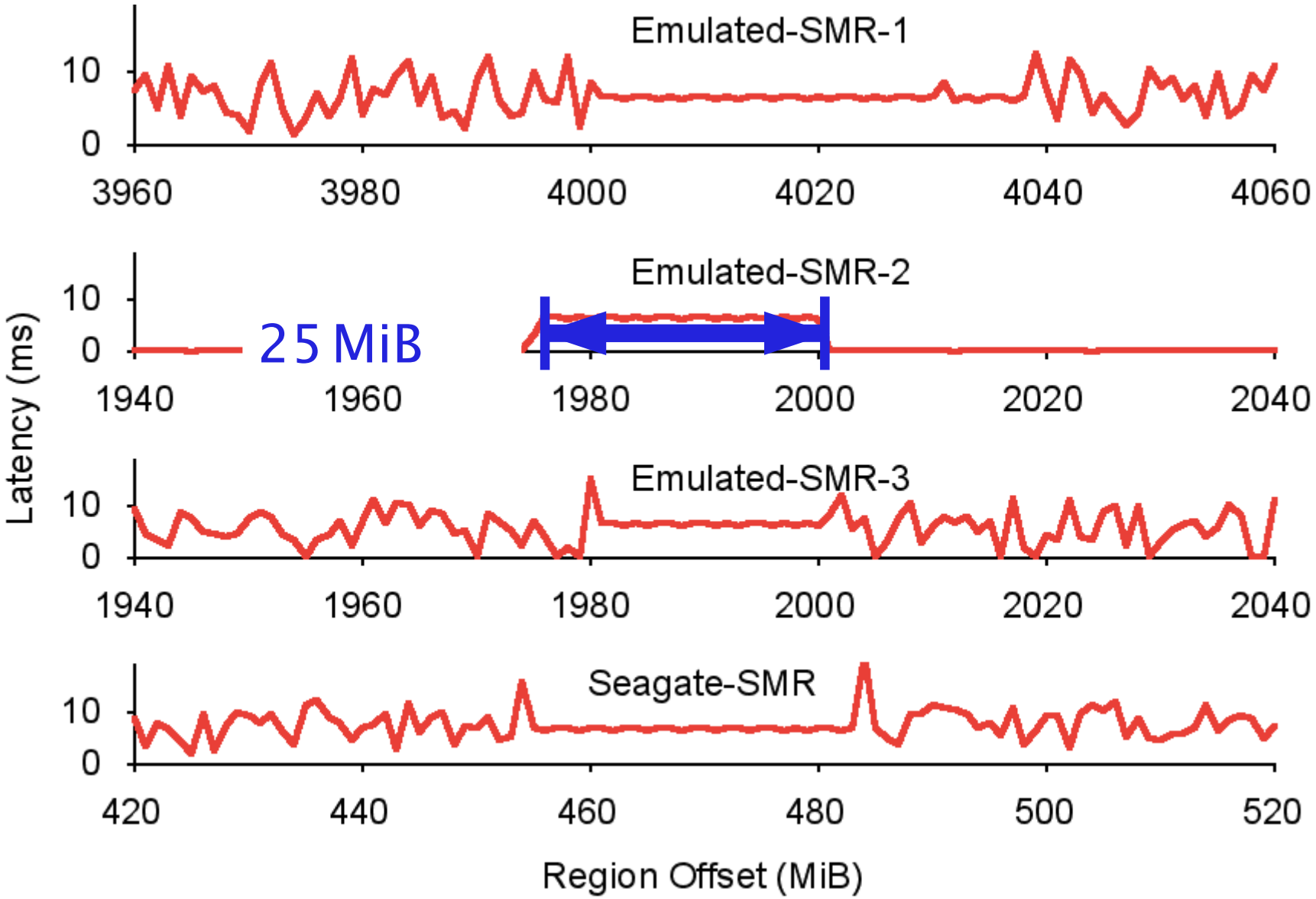
Sequential Read of Random Writes



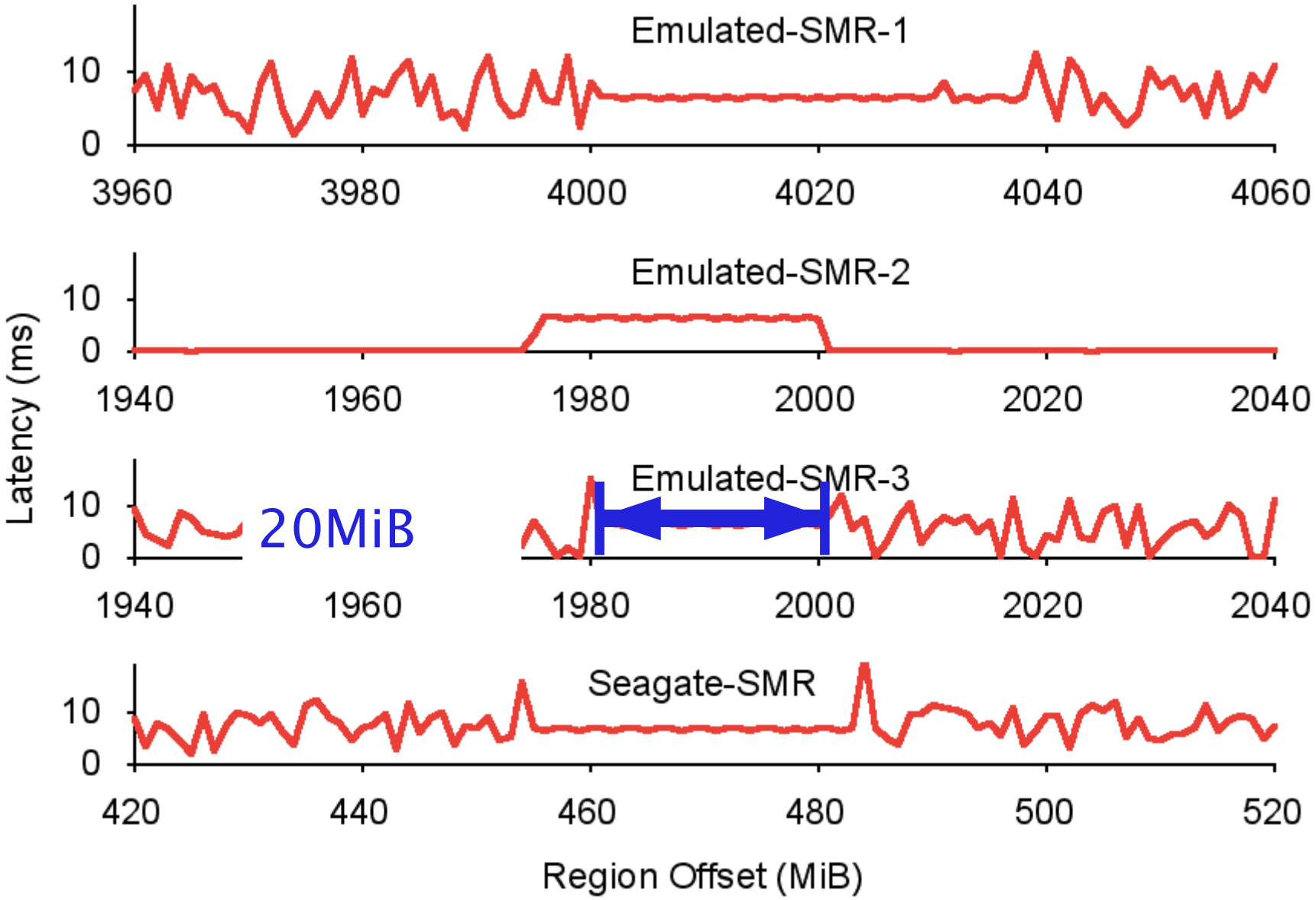
Sequential Read of Random Writes



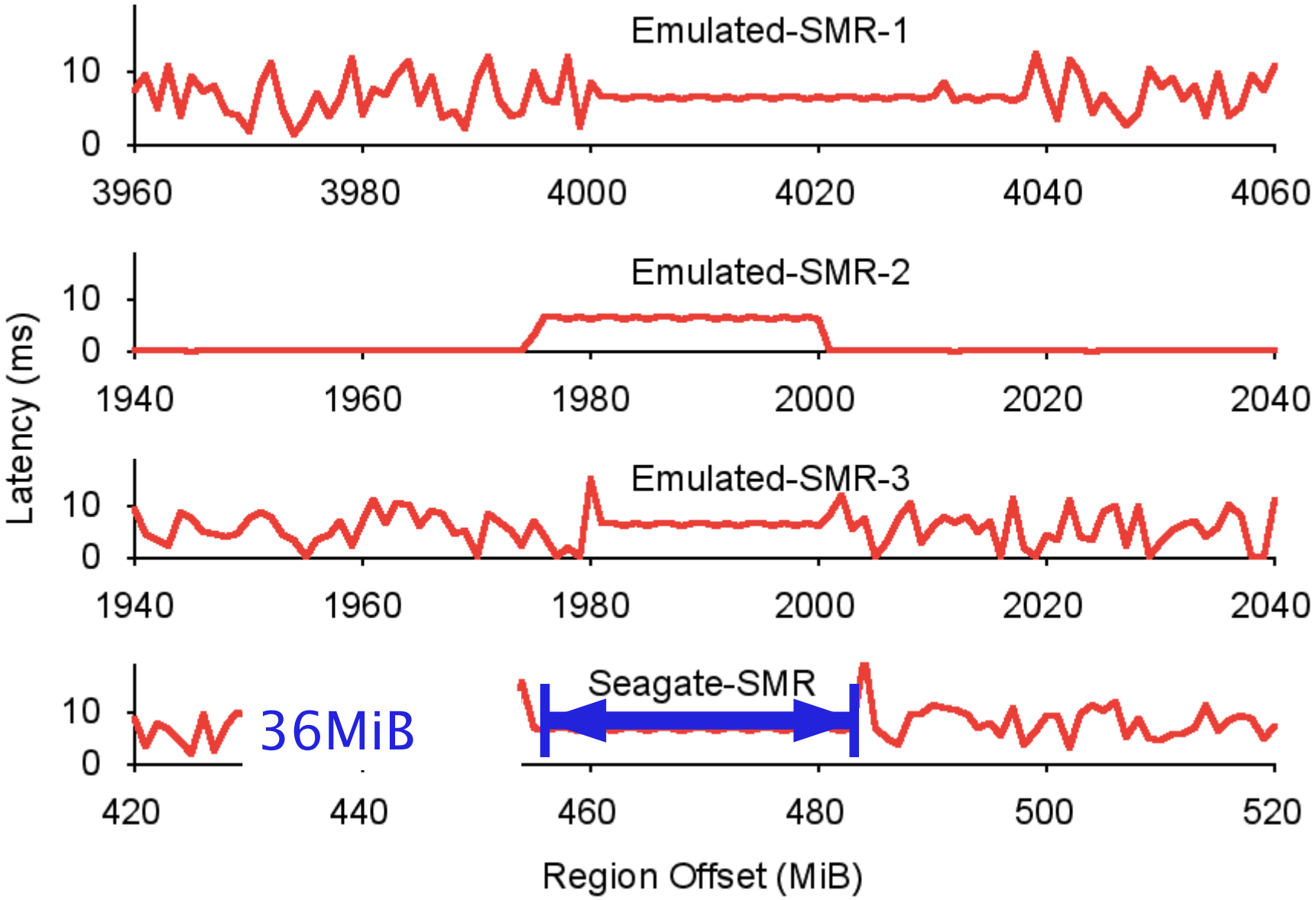
Sequential Read of Random Writes



Sequential Read of Random Writes



Sequential Read of Random Writes



Conclusion

- Drive-Managed SMR drives have different performance characteristics.
- Using them efficiently will require changes to software stack.
- Skylight aims to guide these changes.
- We aim for generality, more work may be needed.
- Tests, STL source code, video clips are available at

  http://sssl.ccs.neu.edu/skylight