SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

SMR and HAMR Advancing HDD Areal Density

Track: Data Architecture / Storage Architecture

Speaker: Babar Khan, Software/Hardware engineer/PhD candidate

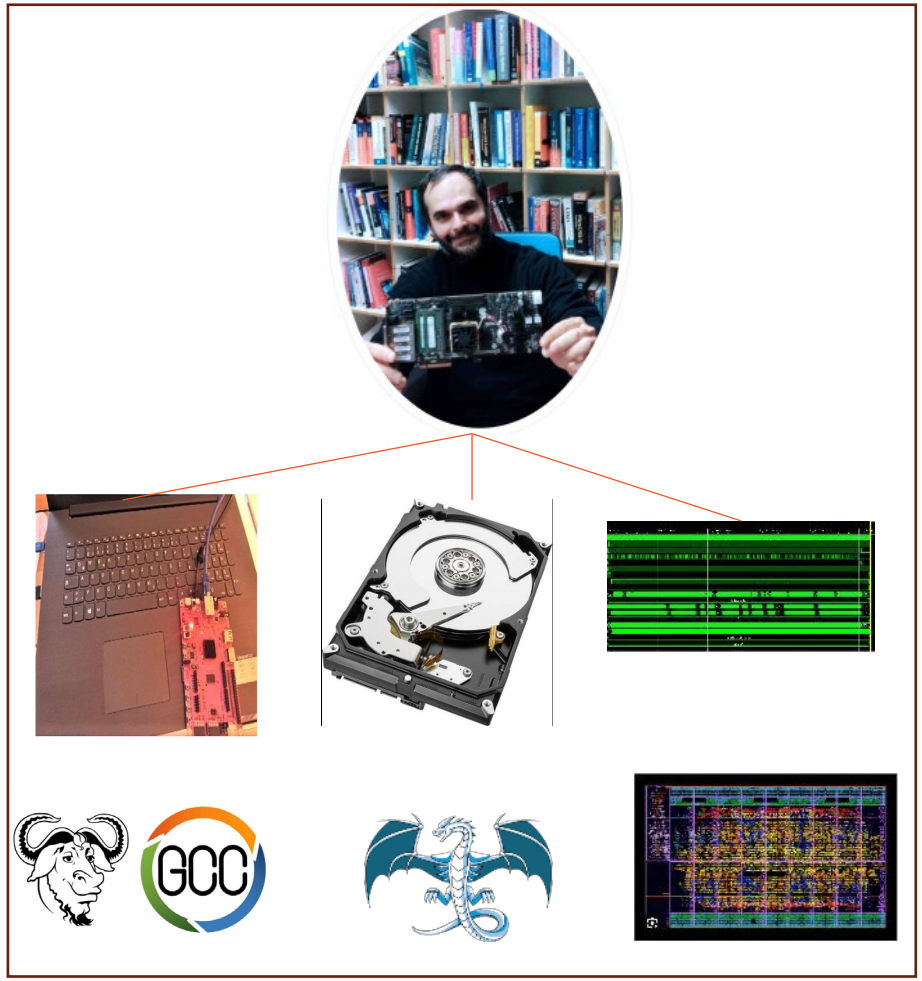
$ whoami

Academic:

- B.Sc./M.Sc. Electrical Engineering
- PhD candidate Computer Science (research: **distributed storage**)

Industry:

- Software/Hardware engineer AI Stealth startup (currently)
- Hands-on:
  - FPGA/ASIC/GPUs
  - Compilers, and DSLs
  - OS, kernel drivers, networks, and storage
  - C, C++, and Python
- Worked as broadcast engineer (past)



Codebase: BabarZKhan, and bk-TurbaAI (Babar Khan)
Stackoverflow: User BZKN - Stack Overflow (121K people reached through 52 answers)
Publications: Babar Khan - Google Scholar (mainly **distributed storage** and networks)

Overview (outline)

Mainly **3** topics:

1. **Background and Motivation**
2. **SMRs**: Reflecting 17 years of "*shingles* on roof or *clapboards* on a wall"
3. **SMR and HAMR**: 2025 and onwards

Takeaways of talk? Mainly **2:**

1. **SMRs as of 2025**? Alive and kicking
2. **HAMR 2025**? Physics DONE ✓. Now is the time for **systems** engineering/research in HAMR

Disclaimer ☺:

- I promise NOT to sell you an AI panacea for every problem.
- Strictly disk related talk.

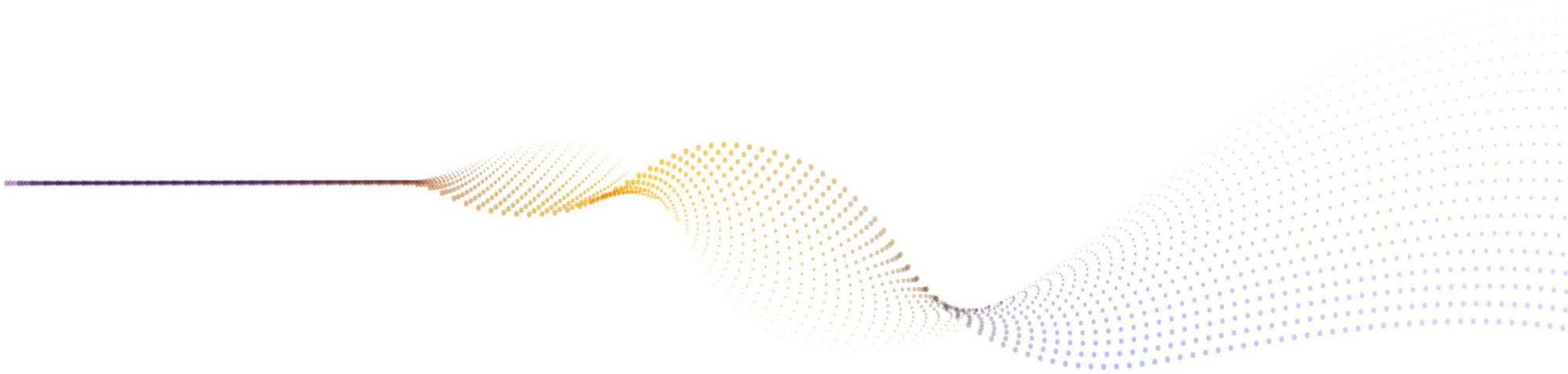
Setting the stage…..

The "spinning rust is dead" crowd can join the same never right "tape is dead" crowd.

"HDDs to AI is as important as Batteries are to EVs."

"When I entered the industry in the early 1990s I was told that tape is dead.."

"Forget about HDD's going away, tape was supposed to be extinct by this time....."



1. Background and Motivation

Background and Motivation

Background:

I have been doing academic research on following **two** topics:

1. I/O in data center and HPC
2. Limitations of traditional I/Os

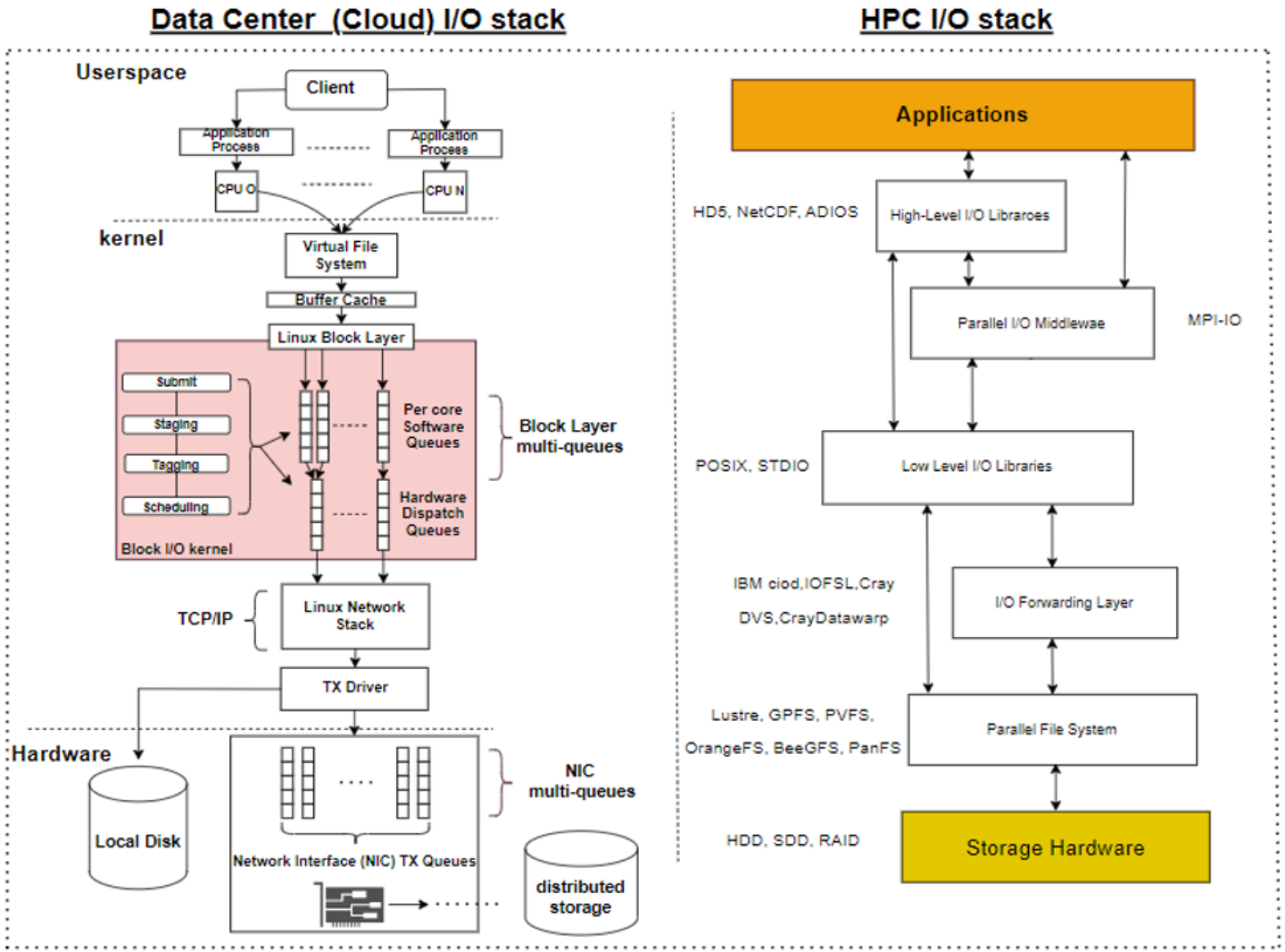
Motivation:

While working on above topics, I realized the following **two** things:

- Knowledge base for **SMR** (Shingled Magnetic Recording) is scattered
- Can we somehow help storage practitioners/researchers/community with lessons learnt from **SMRs**?

Next **2** slides will visually sum up **background** and **motivation**:

Background: I/O stack (layer) in **Data Center** and **HPC**



Note: Understandably, the end hardware remains same.

HPC I/O stack fig: I/O Access Patterns in HPC Applications: A 360-Degree Survey*
JEAN LUCA BEZ and SUREN BYNA, Lawrence Berkeley National Laboratory, USA, SHADI IBRAHIM, Inria, University of Rennes, CNRS, IRISA, Rennes, France

Background

Motivation based on a research problem:

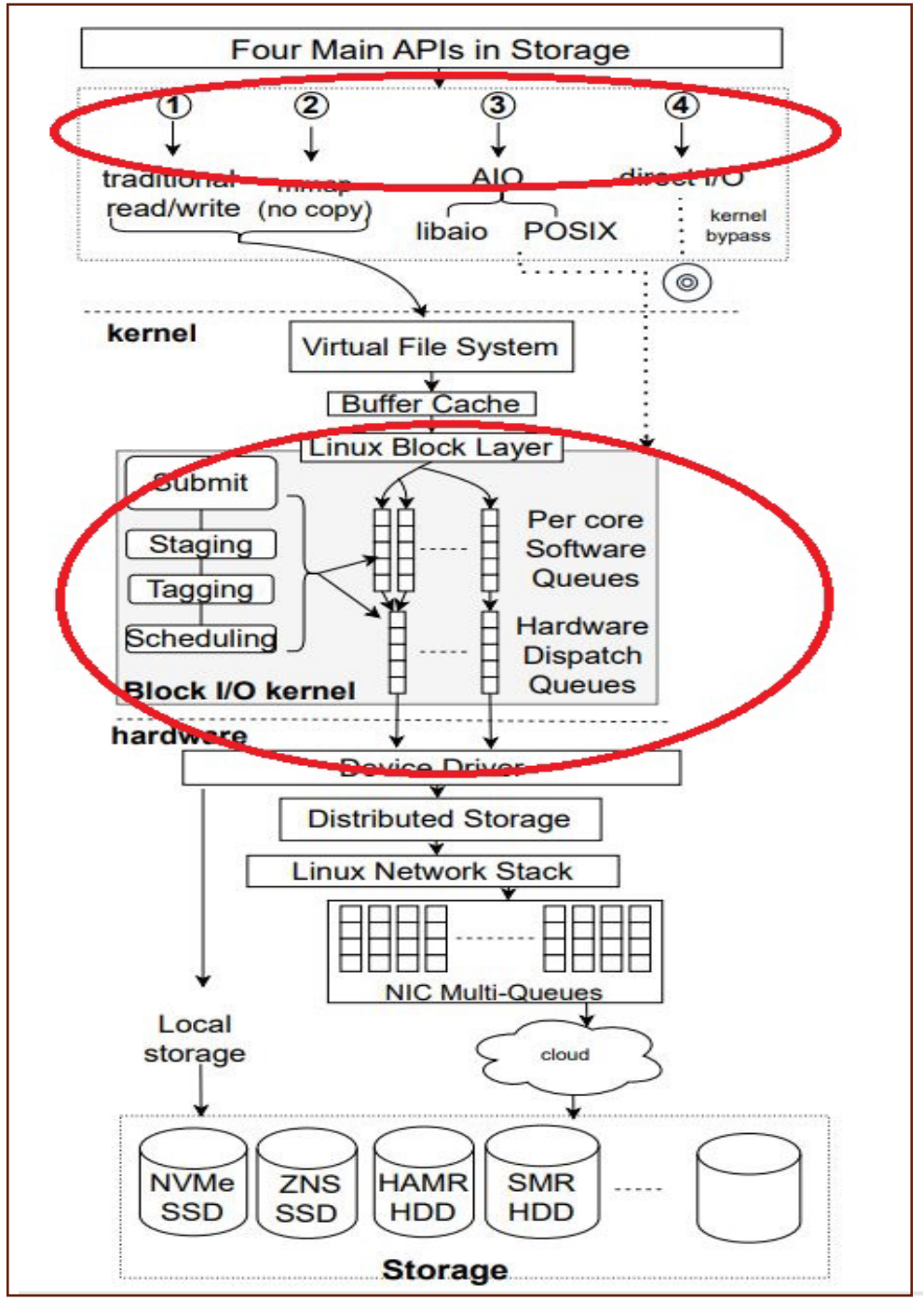
- **Problem 1** (first circle):
  Programming APIs: **intent** mismatches **end-result**

- **Problem 2** (second circle):
  Performance: **60%-90%** of total execution time in kernel

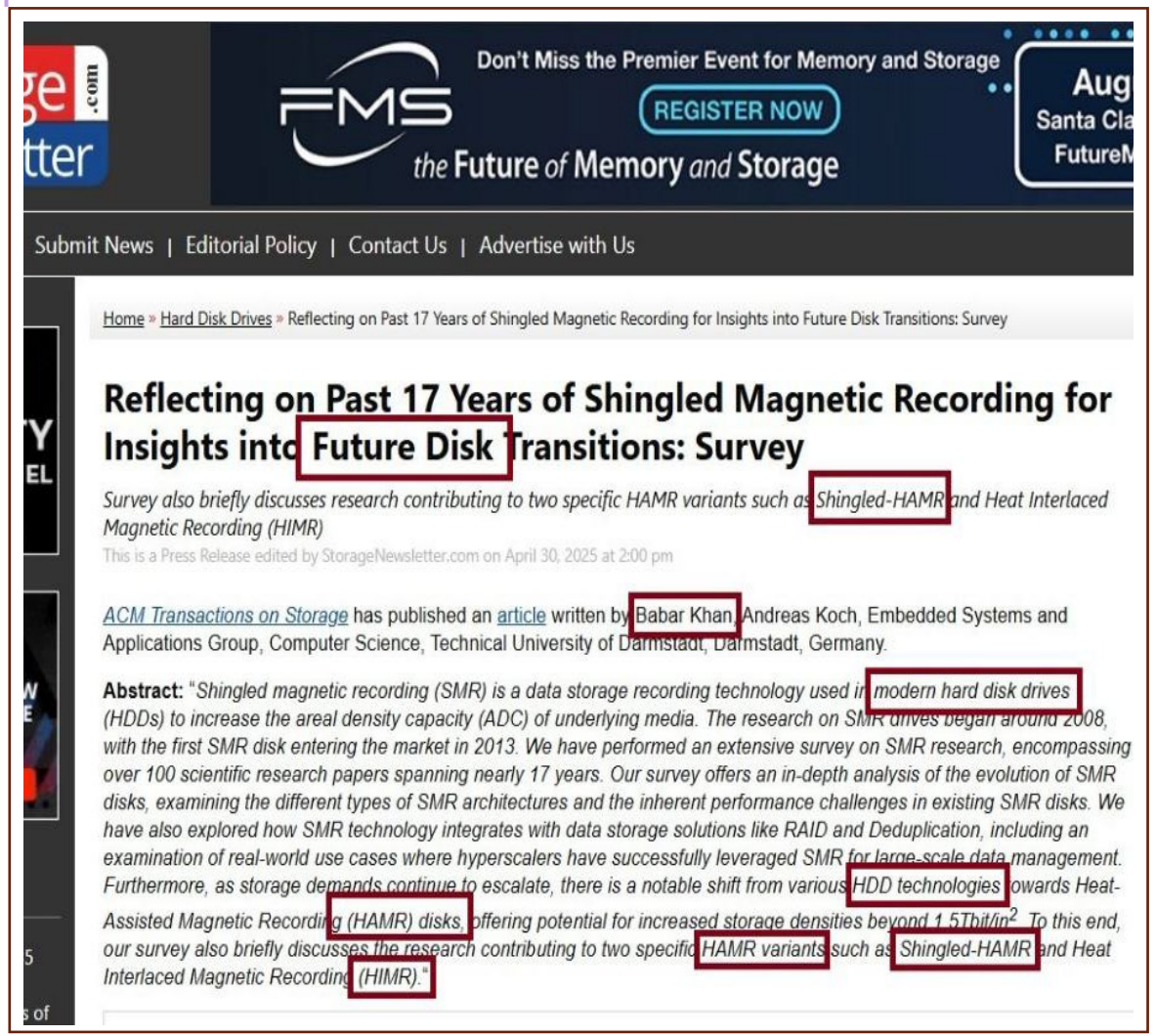
So, while working on these **two** problems, the question was:

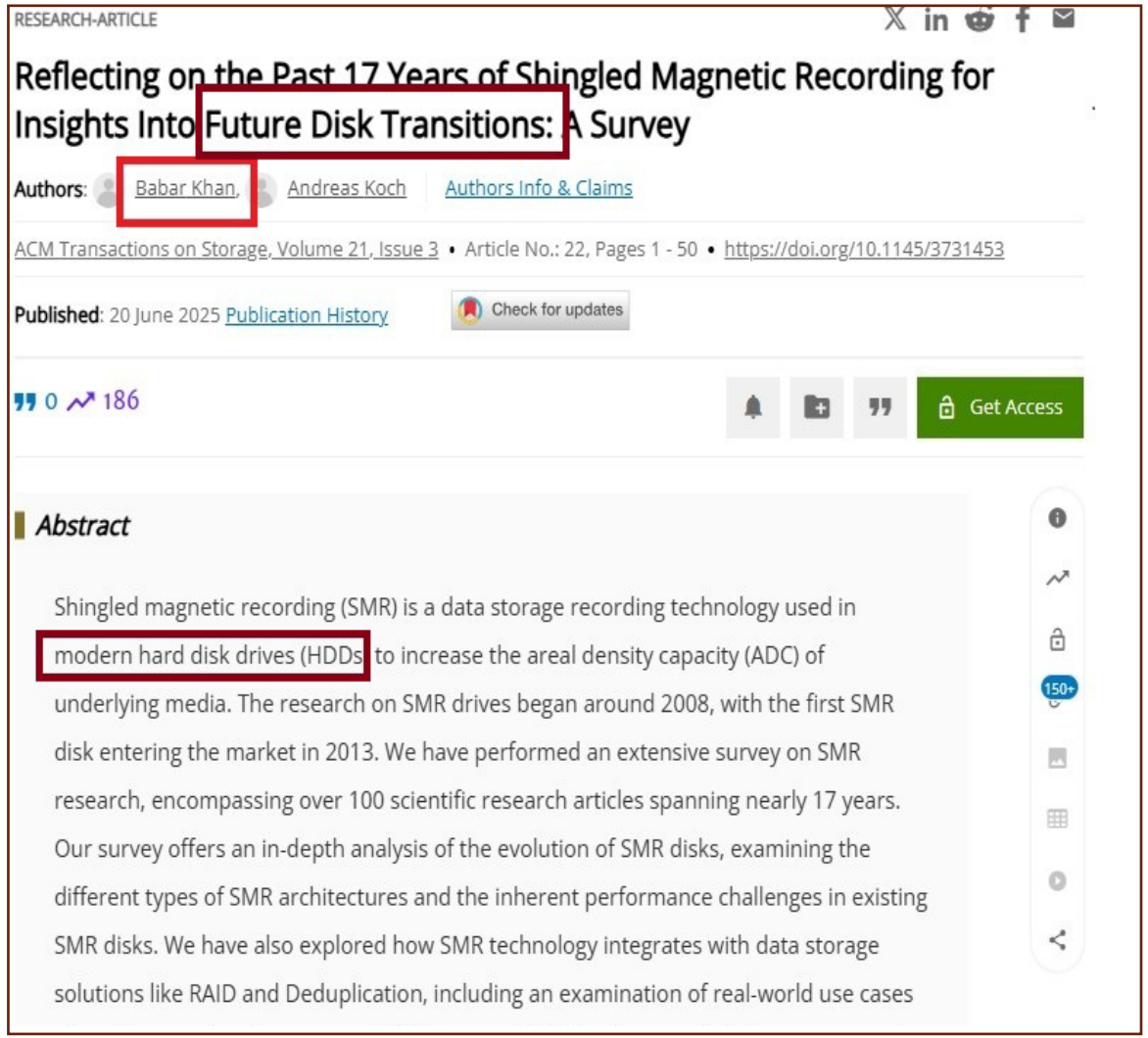
How are HDDs performing?

1. how is **SMR** performing?
2. how will **future** drives perform?
3. how will **HAMR** perform?



Motivation





rigorous peer review: took 1.5 year to finally publish (thanks to reviewers and editors)

2. **SMRs**: Reflecting 17 years of SMRs
(*shingles* on roof/clapboards on a wall)

SMRs: Reflecting 17 years of SMRs

Q: Why **17** years?
A: Most SMR practical and academic breakthroughs emerged around **2008/2009**

Following **5 key-takeaways** in next slides:

1. Lessons learnt from **Drive-Managed** Shingled Magnetic Recording (DM-SMR)
2. Lessons learnt from **Host-Managed** Shingled Magnetic Recording (HM-SMR)
3. Lessons learnt from **Host-Aware** Shingled Magnetic Recording (HA-SMR)
4. What is a **Hybrid**-SMR?
5. Has SMR paved way to **HAMR**?

HDD technology

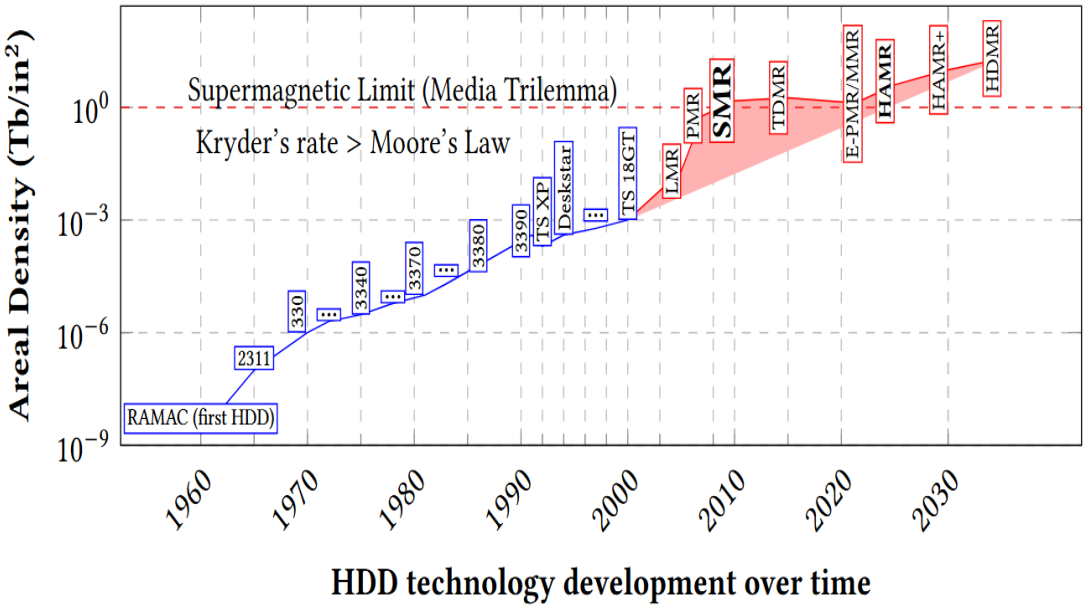
Past and Present:

- **1878**: Concept of magnetic recording (Oberlin Smith)
- **1956–57**: First HDD – IBM RAMAC, ~2 kbit/in²
- **1960s–2000s**: Rapid progress in CMR
- **2003**: ~60 Gbit/in² (30M× increase over RAMAC)
- **Kryder's Law**: Areal density doubled ~every 18 months
- **2000s**: Superparamagnetic effect → scaling limits ("***media trilemma***")
- **2013**: First consumer-grade SMR
- **Present**: Transition to SMR + HAMR (>1.5 Tbit/in² potential)

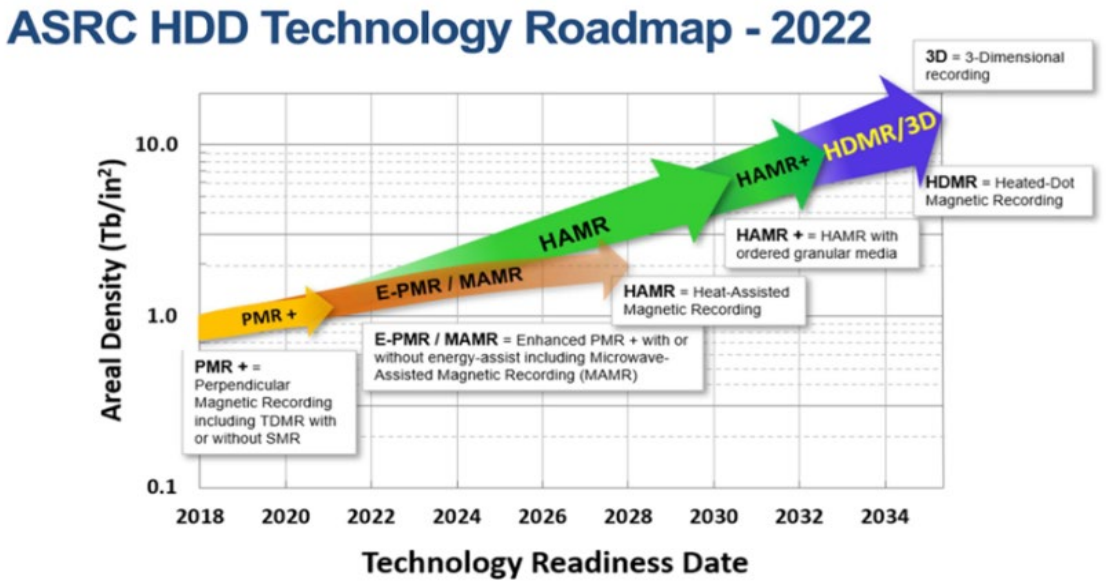
Future:

- **2028**: > ≈ 4 Tbit/in²
- **2032**: > ≈ 6 Tbit/in²
- **2036**: > ≈ 8 Tbit/in²
- **2040**: > ≈ 10 Tbit/in²

HAMR, HAMR+, HDMR/3D etc under AI/ML workload

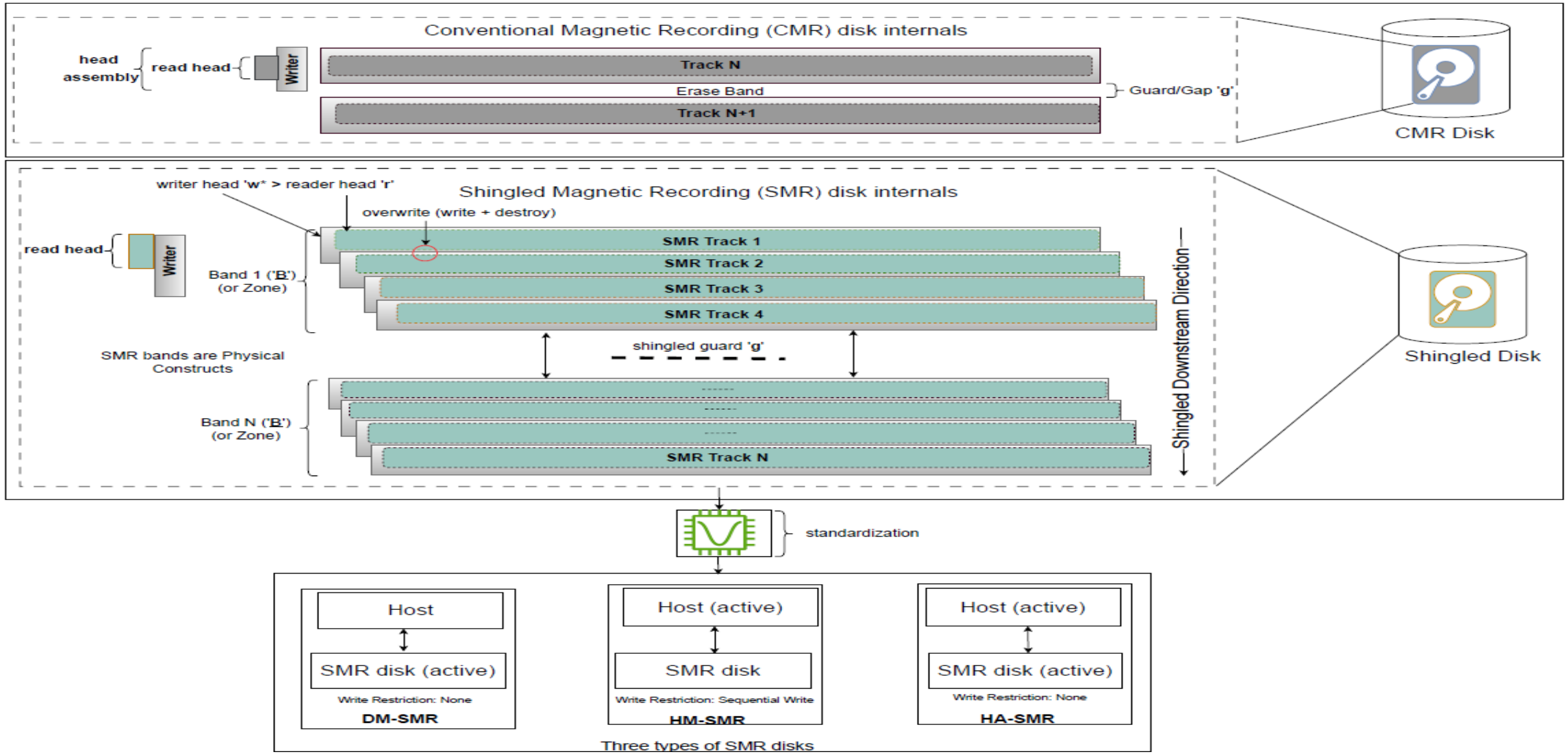


HDD technology development over time



*Magnetic recording has been a quantum marvel, with areal density improving by **600 million times** since the first hard drives. From the quantum effects behind GMR/TMR sensors to HAMR's laser-enabled breakthrough….***Xiaodong Che**

Shingled Magnetic Recording (SMR) Architecture

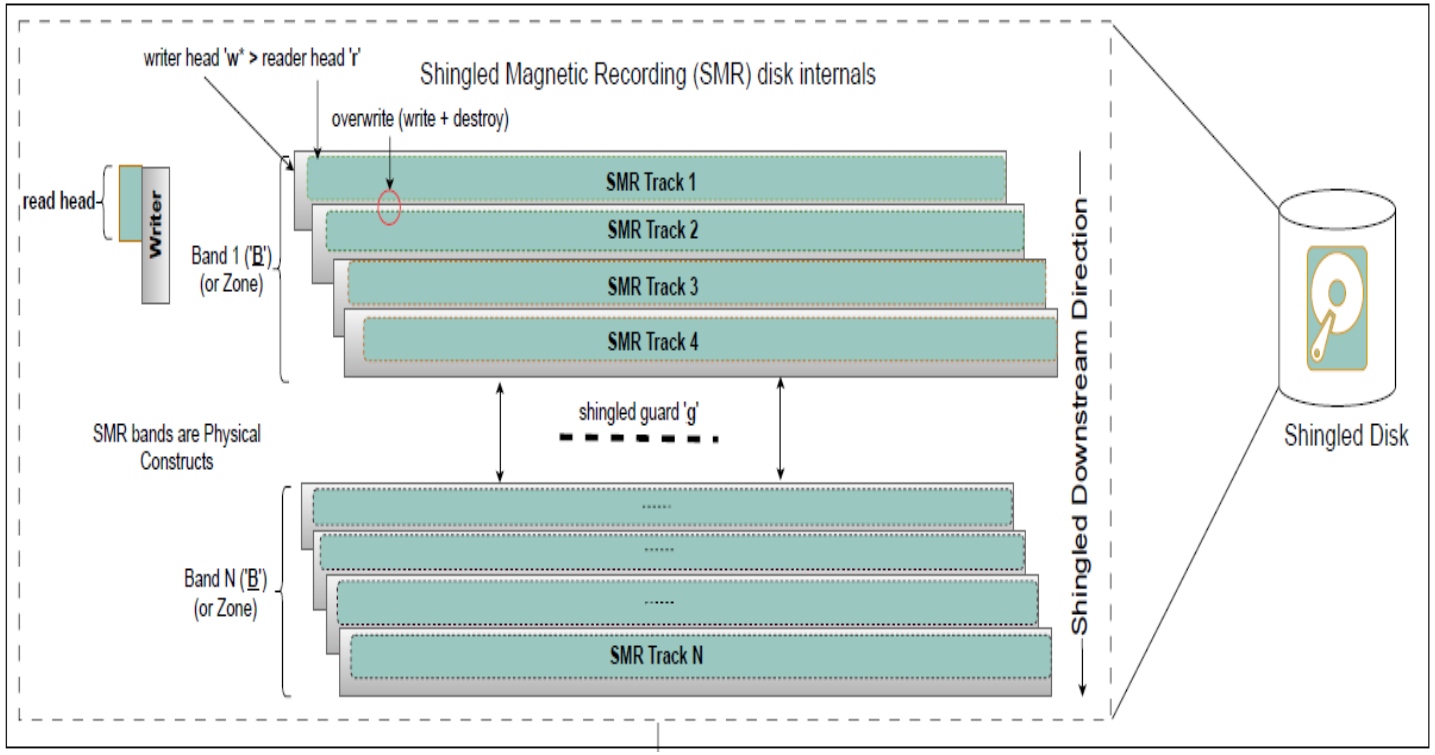


Next: Let's bisect briefly

Areal Density of SMR

Basically following **7** variables

1. Bands/Zones
2. Tracks per inch (TPI)
3. Common band sizes
4. Inside each band?
5. Individual track size?
6. Number of tracks shingled together?
7. Shingled guards?

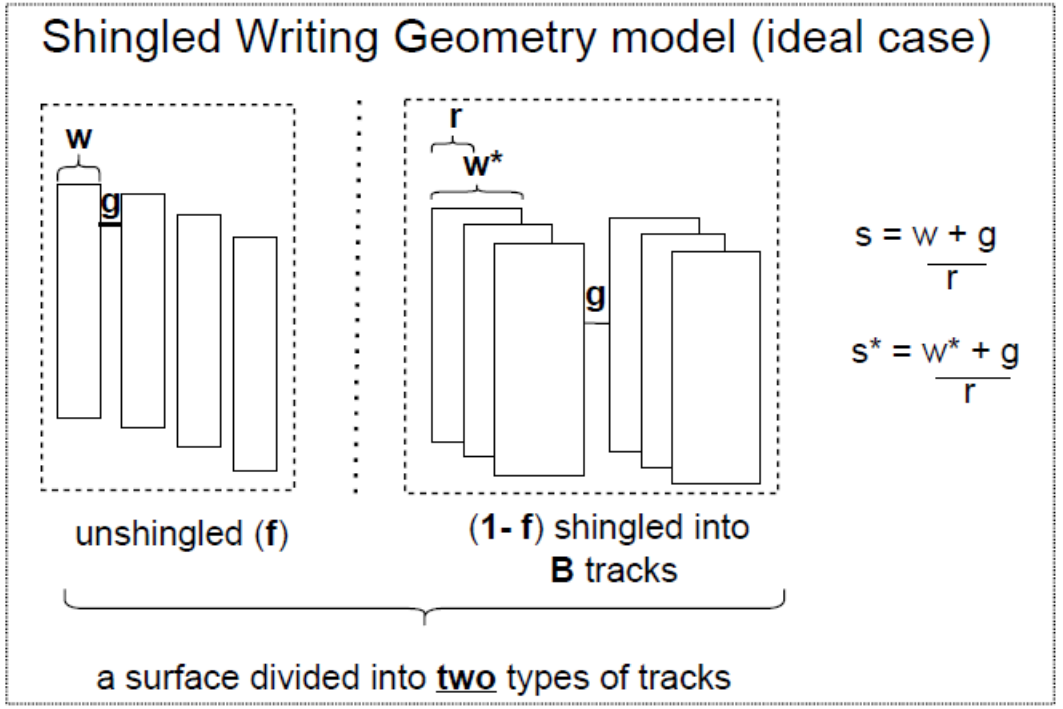


- **Bands/Zones**:Contiguous overlapping tracks, written sequentially
- **Tracks per Inch (TPI)**:TPI of disk radius → higher TPI = higher density
- **Common Band Sizes**:256 MiB (8 TB disks); also 13–36 MiB / 15–40 MiB
- **Inside each Band**:tracks overlap; random reads allowed, writes restrictive
- **Track Size**:100 kB – 1 MB
- **Tracks per Band**:~100 (limits overhead ≤10%)
- **Guard Regions**:Gaps between bands; prevent overwrite across zones

- Implications: balance of **increased areal Density** (**AD**) and **write amplification** (WA)
- Did the industry and research community **quantify** AD and WA? Yes, next slide

Increase Factor Areal Density: Shingling Writing Geometrical Model*

- **Conventional Recording**:
  - Track width **25 nm**, guard gap **5 nm**
- **Shingled Writing**:
  - Wider tracks (**70 nm**)
  - Sharper edges, down-band spacing reduced to **10–20 nm**
- **Geometrical Model Insights**:
  - Disk surface divided into **unshingled (f)** and **shingled (1–f)** regions
  - Areal density increase factor **A ≈ 2.25× – 2.5×**
  - Potential gains up to **3–5×** (industry estimates)



Quantification of Write amplification

Write Amplification Issue:

- Small updates trigger large sequential writes
- Example: updating 4 KB → 64 MB rewrite

Impact:

- Write amplification significantly HIGH for 4 KB update in 64 MB band
- Larger band sizes = more severe amplification

Key-takeaway: SMR geometry enables up to **threefold density gains**, surpassing the conventional **1 Tbit/in² superparamagnetic limit**.

**Storage Systems for Shingled Disks*
Garth Gibson Carnegie Mellon University and Panasas Inc
Anand Suresh, Jainam Shah, Xu Zhang, Swapnil Patil, Greg Ganger

Road to standardization:

SHINGLED MAGNETIC RECORDING FOUNDATIONAL RESEARCH: **BANDS** AND **GARBAGE COLLECTION** METHODOLOGIES

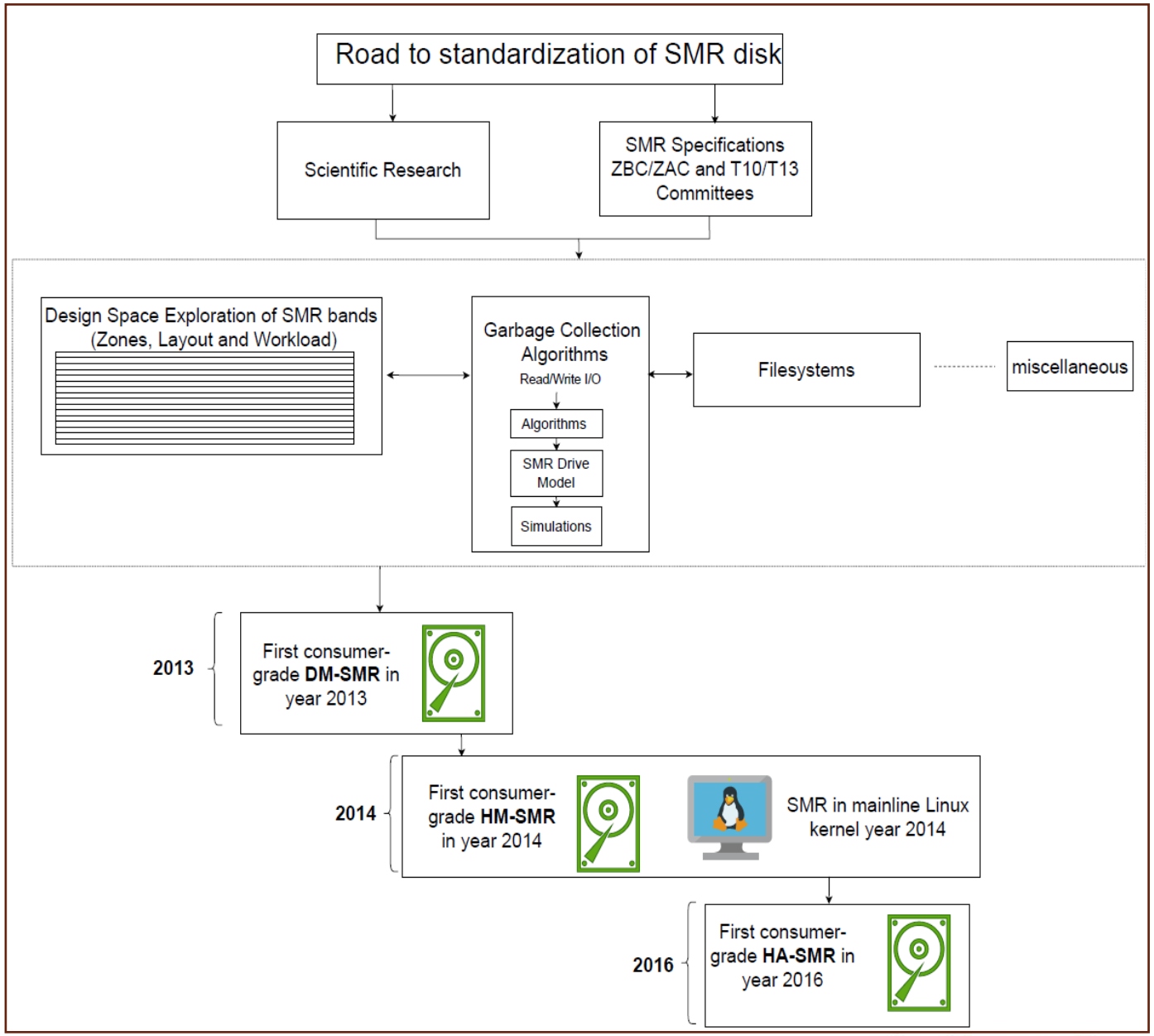
Classification of Shingled Magnetic Recording

Ref	DM-SMR			HM-SMR	HA-SMR	H-SMR	Source		Evaluation		Code
	O-SMR	I-SMR	hybrid				Peer	Misc	Shingled	Sim/CMR	
[32, 66, 82, 114, 129, 140, 146, 164, 189, 235] [25, 33, 34, 101, 102, 131, 145, 207, 236] [26, 37, 42, 117, 133, 138, 143, 144, 148, 149] [15, 16, 22, 78, 125, 147, 170, 181, 196, 210, 212]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
[89, 116]	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗
[88] [187]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
[3–5, 188, 192]	✓	✗	✗	✗	✗	✗	✓		✓	✗	✓
[193]	✓	✗	✗	✓	✗	✗	✓		✓		✓
[211, 227, 246]	✓						✓		✓		✗
[77, 80]	✗	✗	✗	✓	✓		✓		✓	✓	✓
[220, 224, 230, 233, 234, 240] [79, 228, 229, 232, 242]	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗	✗
[30, 126, 132, 134–136]	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗
[86, 99, 118, 130, 171] [28, 104, 139, 217, 218, 237, 238]	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✓	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✓	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✓	✗
[31, 150, 151, 156]	✗	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗
[239, 245]	✗	✗	✗	✓	✗	✗	✓	✗	✓	✗	✓
[65, 168]	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗
[6, 51]	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
[194]	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	✗
[108]	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗
[221, 222]	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗
[184]	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗

- **Extensive research (study) of <u>250</u> references that include:**
  - Scientific peer reviewed papers
  - Selective patents of SMRs
  - Specifications doc
  - Blogs/whitepapers/technical reports
  - SDC content

- **Classification Basis**:
  - Research categorized by SMR drive types: DM-SMR, HM-SMR, HA-SMR, and the newer H-SMR.

- **Code Contributions**:
  - Explicitly highlights open-sourced works



Road to Standardization

Period of Intensive Collaboration (2009-2016):

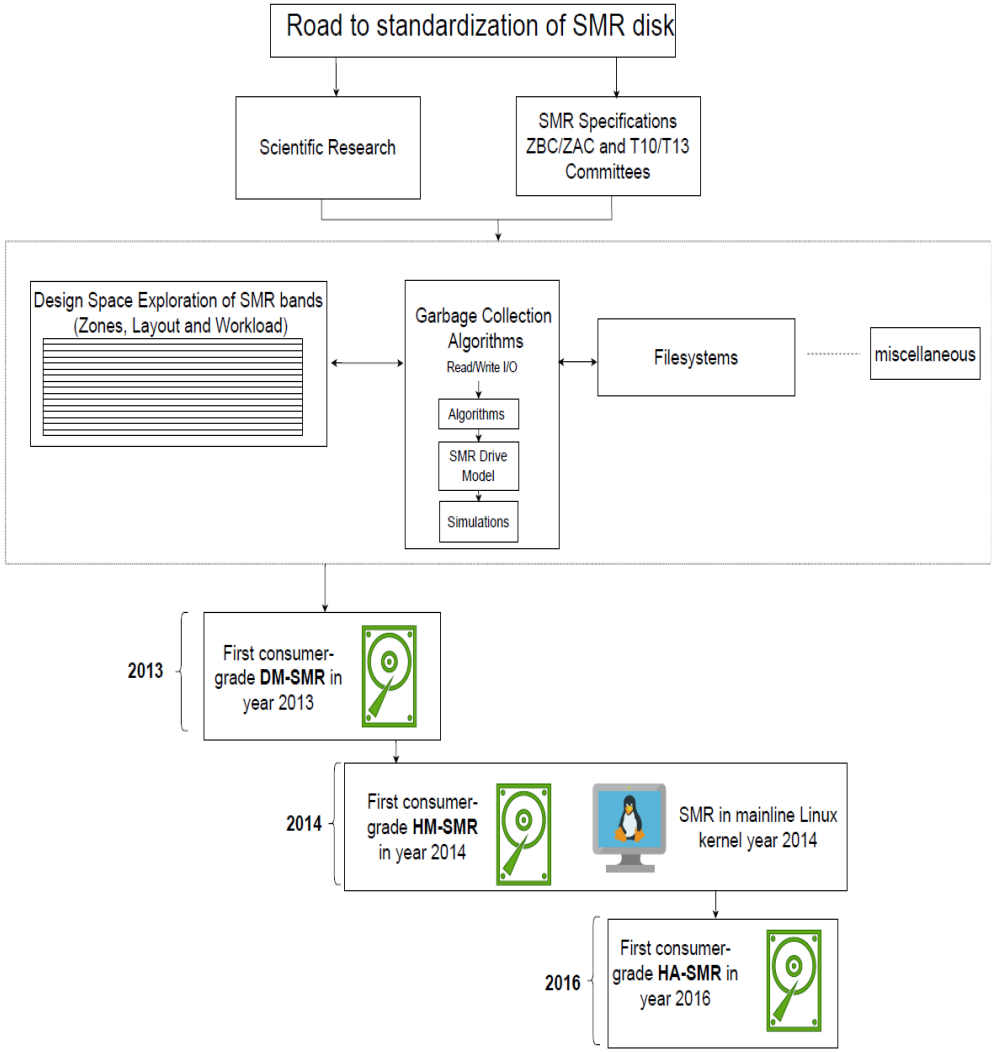
- Academic and industry practitioner shaped trajectory of SMR
- Focused on backward compatibility with legacy CMR

Three Pillars of Early SMR Research:

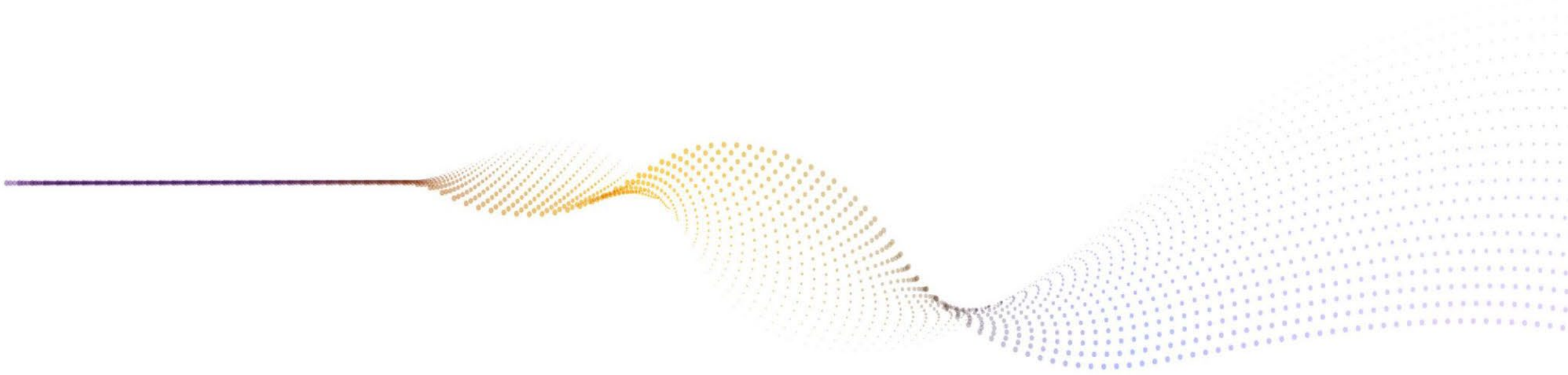
- Garbage Collection Algorithms
- Design Space Exploration of SMR bands (zones, layout, workload)
- Filesystem and workload

Path to Standardization:

- DM-SMR
- HM-SMR
- HA-SMR



Year	Garbage Collection Frameworks in SMR
2010	*set-associative* and *S-blocks* [22]
2012	*intraband GC*, *normal GC*, and *forced GC* [129]
2012	*short block (E-region and I-region)* [82, 83][2]
2014	*Round Robin mapping schemes:* R(4123), 124R(3), 14R(23) [89]
2015	*empty weight (greedy)* and *cold weight* [102]



Drive-Managed Shingled Magnetic Recording (DM-SMR)

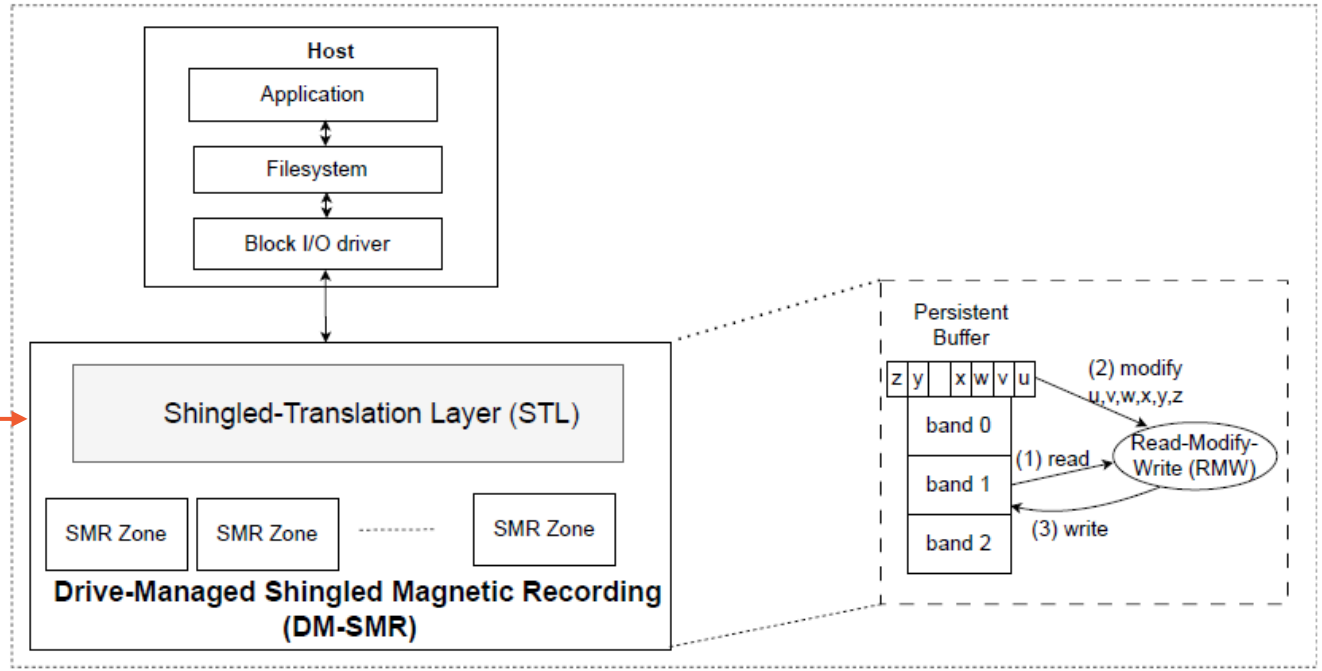
Drive-Managed Shingled Magnetic Recording (DM-SMR)

DM-SMR:

- First consumer-grade SMR (2013)
- Firmware-managed shingling & I/O
- Also known as *autonomous disks* or *type-1 SMR*
- Backward compatible, no standard command set

STL (SMR Translation Layer):

- SMR's version of FTL
- Maps LBA → PBA across shingled tracks
- Handles Read-Merge-Write (RMW) internally
- Performs garbage collection & data relocation
- Key factor in DM-SMR performance



Drive-Managed Shingled Magnetic Recording (DM-SMR) main architecture

Research focus of DM-SMR:

- Reverse-engineering DM-SMR internals
- On-disk persistent cache studies
- Hybrid drives (SSD + DM-SMR)

Types of DM-SMRs

Three* types of DM-SMRs:

- Out-of-place update SMR (O-SMR)
- In-place update SMR (I-SMR)
- Hybrid (O-SMR + I-SMR)

O-SMR:

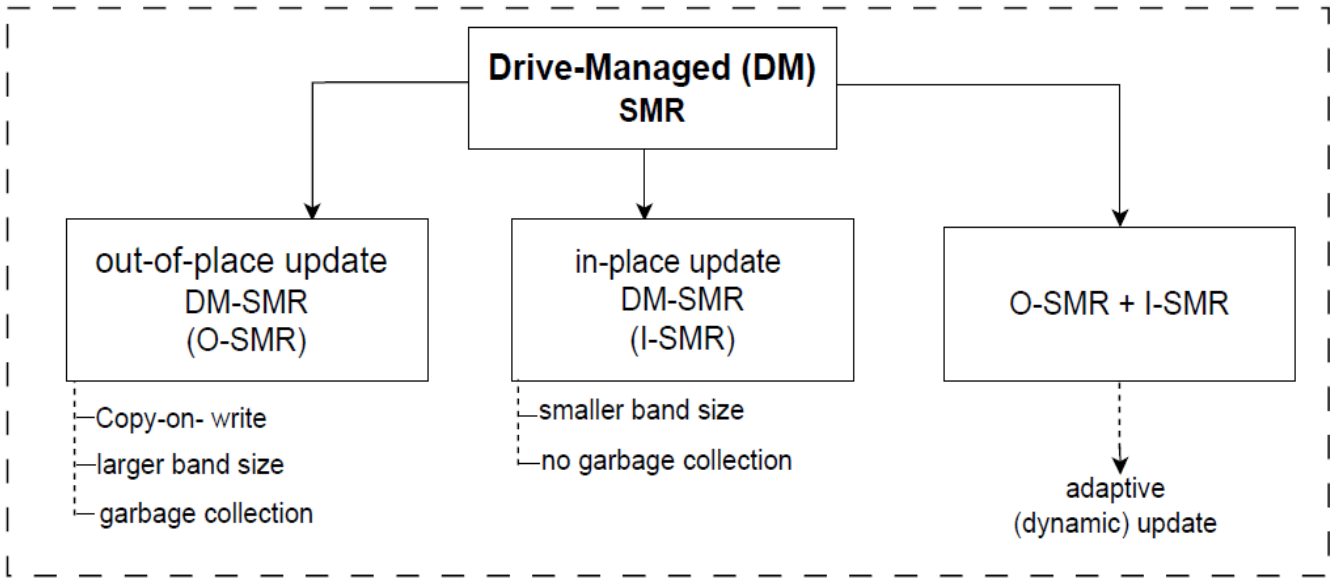
- updates managed by writing updated data to new location rather directly overwriting existing data.
- COW (copy-on-write)

I-SMR:

- Static LBA-to-PBA mappings (no mapping tables)
- No garbage collection required

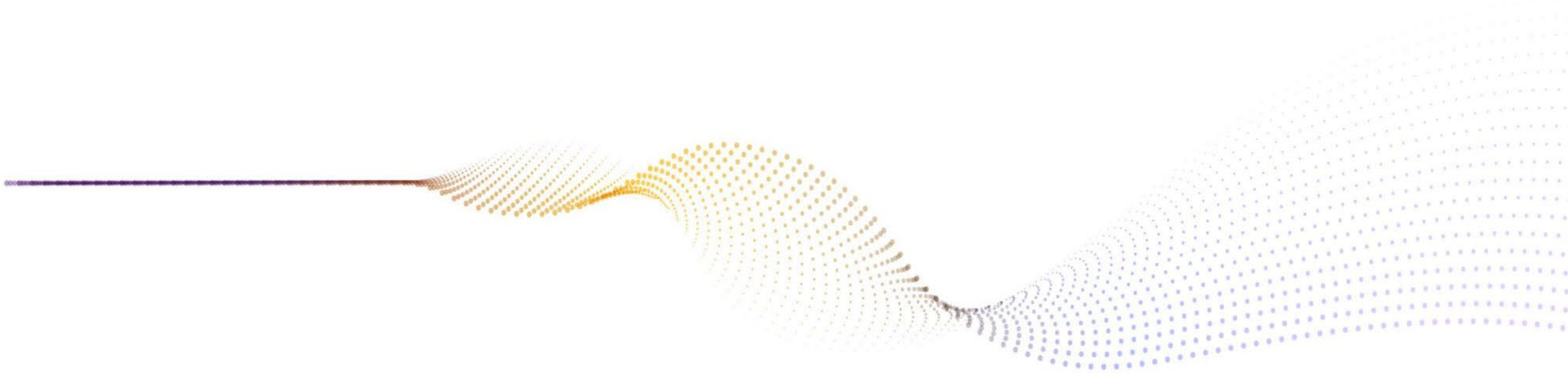
O-SMR + _I-SMR = Hybrid:

- Adaptive (dynamic) update
- Integrates **track-level mapping**
  novel space management scheme



**These three terminologies were coined by Lin et al. and He et al*

** These three types do not mean three types of consumer-grade drives*



1.Lessons learnt from DM-SMR

Lessons Learnt from DM-SMRs

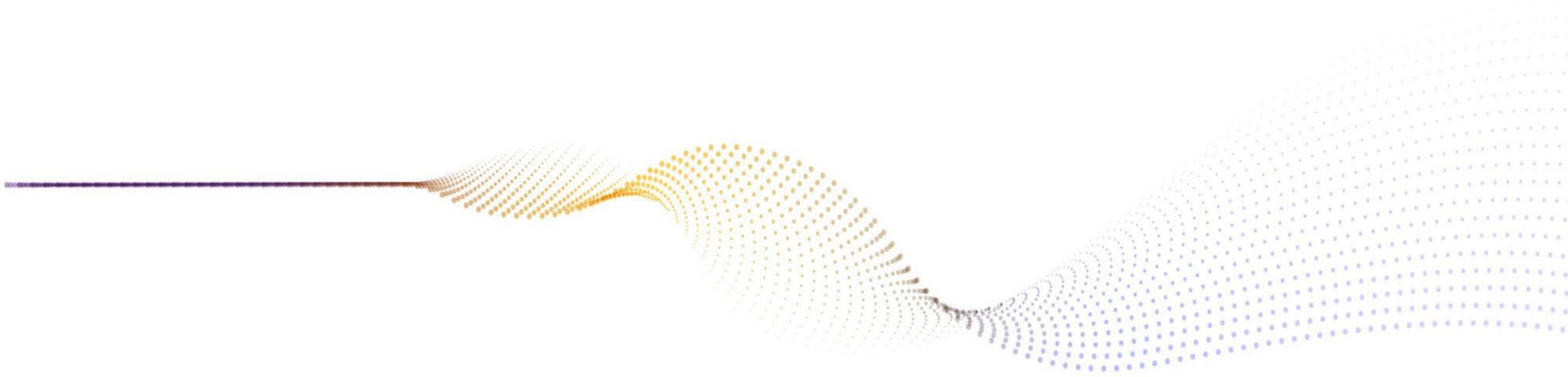
Two main lessons: (1) Unpredictability and (2) Black Box Model

Unpredictability:

- **STL vs. FTL**:
  - STL (SMR Translation Layer) in DM-SMR differs significantly from FTL (Flash Translation Layer) in NAND flash.
  - Flash memory avoids seek latency; SMR drives incur significant seek latency, impacting STL performance.
- **Write Constraints**:
  - SMR → sequential-write constraint to prevent overwrites.
  - Flash → erase-before-write with asymmetric program/erase sizes.
- **Cleaning Overheads**:
  - SMR cleaning operates on large zones (256 MB vs. 2–16 MB in flash).
  - Cleaning may re-read/re-write zones multiple times, causing **1–2 second delays**.

Black Box Model:

- **Lack of Standardization**: No standardized command set compared to other SMR drives.
- **Metadata vs. User Data**: Drives struggle to differentiate → risk of write amplification.
- **Fixed Metadata Handling**: Unlike SSDs, disks limited to 4 KB sectors with firmware-only metadata access.
- **Disk Heterogeneity**:
  - Even same-model drives show variable performance due to **adaptive zoning techniques**.
  - Results in non-uniform performance across multi-disk systems.
- **System Implication**: Storage optimization becomes harder due to unpredictable, heterogeneous behavior.



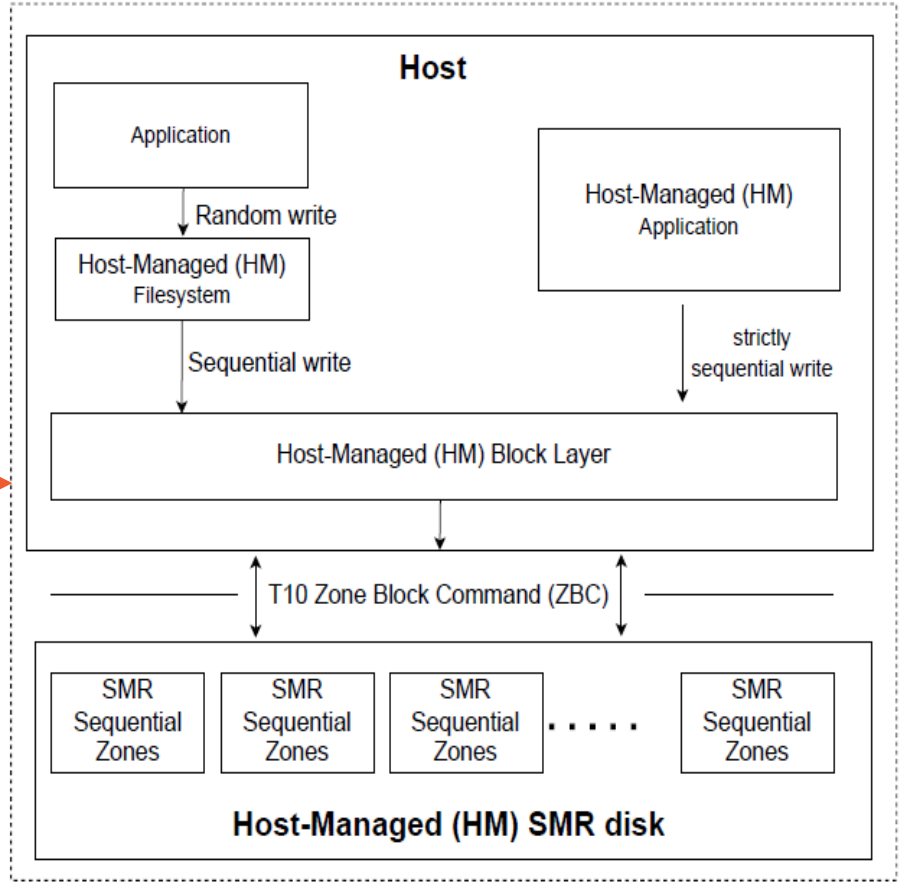
Host-Managed Shingled Magnetic Recording (HM-SMR)

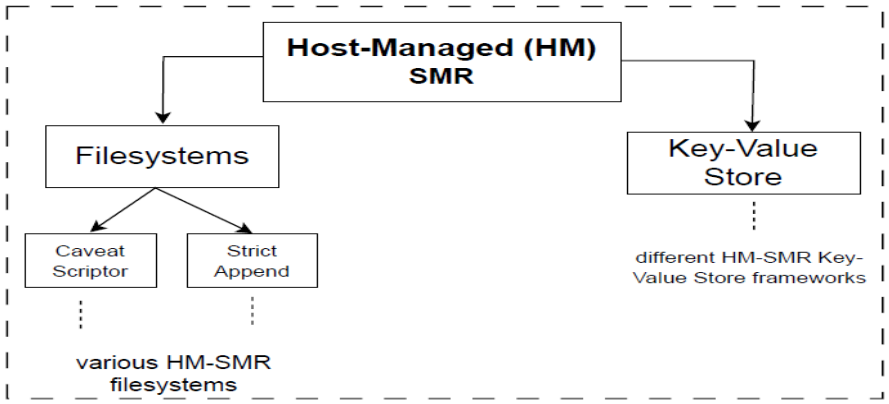
Pushing intelligence up: Host-Managed (HM) SMR

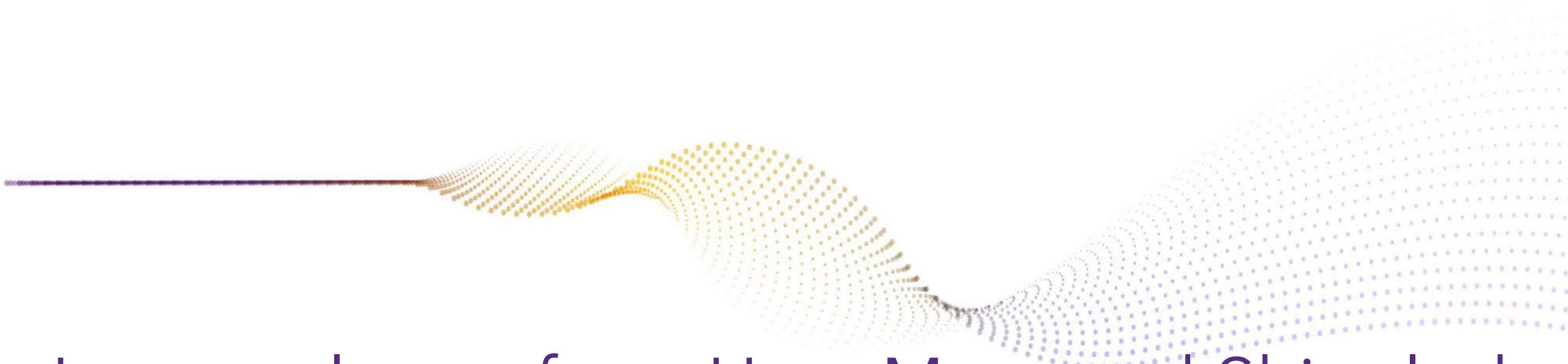
Host-Managed SMR (HM-SMR):
- Research began ~2012; first disk released in 2014.
- Alternative to DM-SMR due to GC-related bottlenecks.
- Also known as *Type-II SMR*.

Key Characteristics:
- Intelligence shifted to the host (not firmware-managed).
- Provides predictable performance vs. DM-SMR's
- Requires filesystem modifications or new filesystem designs.
- Uses standardized interfaces:
  ZBC (Zoned Block Commands, ANSI INCITS 536)
  ZAC (Zoned ATA Commands, ANSI INCITS 537)

Research focus of HM-SMR:
Understandably filesystems







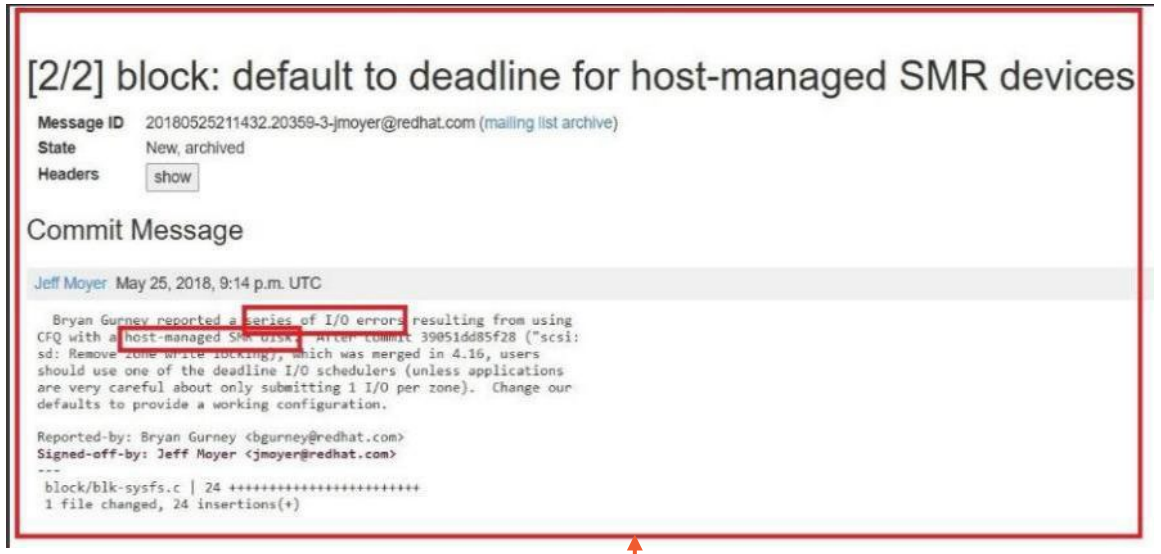
Lessons learnt from Host-Managed Shingled Magnetic Recording (HM-SMR)

Lessons learnt from HM-SMR DMRs

- **High Software Overhead:**
  - effort required to revamp I/O stack & filesystems.
  - use cases lean toward **backup & archival** workloads.

- **Sequential Write Challenges**
  - most I/O stacks reorder accesses → out-of-order writes.
  - lead to I/O errors in HM-SMR drives.

- **Linux Kernel Issues**
  - CFQ scheduler caused frequent write errors.
  - Deadline scheduler adopted as temporary SMR-aware fix.
  - Long-term goal → universal dispatch layer for all I/O schedulers.


[2/2] block: default to deadline for host-managed SMR devices

Message ID 20180525211432.20359-3-jmoyer@redhat.com (mailing list archive)
State New, archived
Headers show

Commit Message

Jeff Moyer May 25, 2018, 9:14 p.m. UTC

Bryan Gurney reported a series of I/O errors resulting from using CFQ with a host-managed SMR disk. After commit 39051dd85f28 ("scsi: sd: Remove zone write locking"), which was merged in 4.16, users should use one of the deadline I/O schedulers (unless applications are very careful about only submitting 1 I/O per zone). Change our defaults to provide a working configuration.

Reported-by: Bryan Gurney <bgurney@redhat.com>
Signed-off-by: Jeff Moyer <jmoyer@redhat.com>
---
block/blk-sysfs.c | 24 ++++++++++++++++++++++++
1 file changed, 24 insertions(+)

Temporary uphill battle with Linux mainline kernel

HM-SMR remains workhorse of hyperscalers

- *Dropbox*
- *Alibaba*
- *Google's advocacy for Caveat Scriptor*
- *Huawei*
- *……*

details mentioned in paper

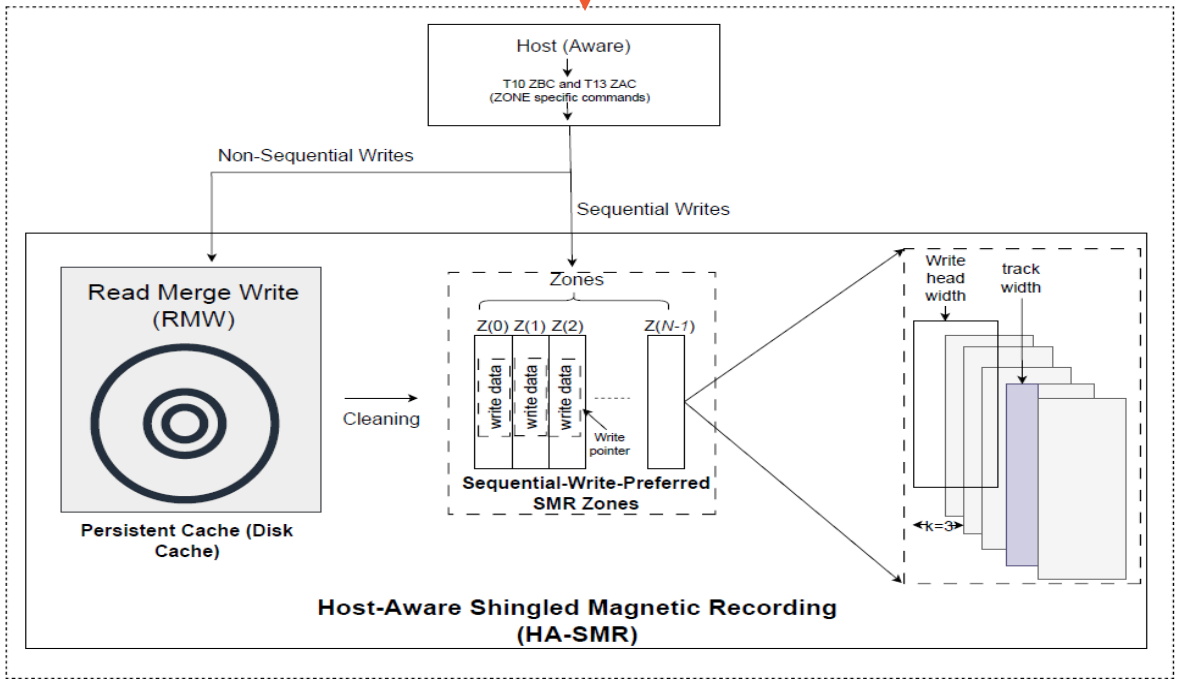
Host-Aware Shingled Magnetic Recording (HA-SMR)

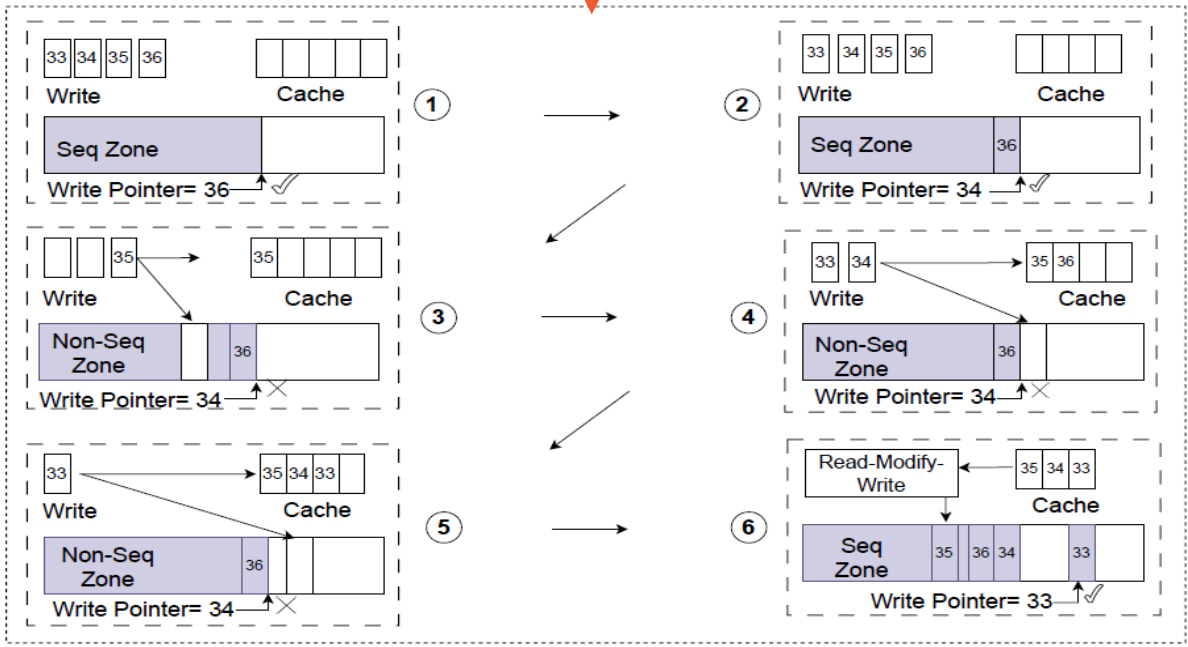
Host-Aware (HA) SMR

Host-Aware (HA)-SMR:
- Research began 2013 → first consumer grade 2016.
- Academic studies appeared 2017 onwards.
- Represents a superset of DM-SMR & HM-SMR.

Core Idea:
- data retention model of **DM-SMR** + performance traits of **HM-SMR**.
- Firmware manages mix of random + sequential writes
- Tolerates occasional deviations from strict sequential patterns.
- Integrates **T10 ZBC & T13 ZAC APIs** with disk firmware.

Key targets:
- Intelligent host policies via APIs for better control.
- Backward compatibility with fewer **Linux** changes vs HM-SMR.





Lessons learnt from Host-Aware Shingled Magnetic Recording (HA-SMR)

Lessons learnt from HA-SMR

By and large, HA-SMR was (or is) to mitigate and improve following, respectively:

- **DM-SMR**: Tail latency, characterized by prolonged response times due to blocking cleaning in STL
- **HM-SMR**: better filesystems

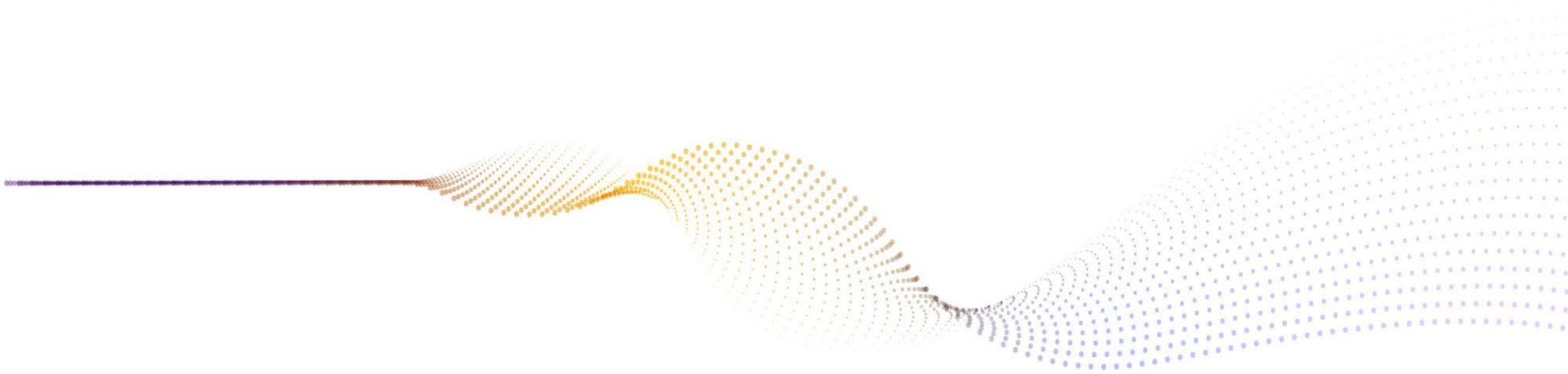
Some of the HA-SMR work cited in paper:

6.3.1 *Wu* et al. *on Evaluation of real HA-SMR drive.* Yang *et al.* [220, 224] evaluated the performance of HA-SMR drives, focusing on features like the open zone issue and media cache cleaning efficiency. This was the first work that has evaluated HA-SMR comprehensively. It proposes a host-controlled indirection buffer to enhance the drive's I/O performance. The open zone issue refers to performance degradation when the recommended maximum number of open zones in HA-SMR drives is exceeded. HA-SMR drives have specific open zones, areas where data can be written without affecting adjacent tracks. When the number of open zones exceeds the recommended limit, the performance of sequential writes can significantly decrease. The work creates sustained non-sequential write workloads to evaluate the performance of HA-SMR drives in handling a large number of zones. The testing program varies the update ratio, IO request number, and IO request size to analyze their impact on the average band cleaning time of HA-SMR drives.

6.3.2 *HA-SMR mitigating the Long Latency.* The work introduced in [233, 234] presents a comprehensive Virtual Persistent Cache design aimed at addressing the prolonged latency challenges observed in HA-SMR drives. While the work focuses on HA-SMR drives, its implications suggest that it could contribute to mitigating the prolonged latency bottleneck prevalent in DM-SMR drives as well. Similarly, the work in [242] also targets to mitigate the long latency in HA-SMR drives. Inspired by the varying performance impacts of idle and blocking cleanings, the work explores a potential strategy for mitigating the substantial tail response time resulting from blocking cleanings through Artificially Triggered Idle Cleanings (AT-IC). AT-IC aims to alleviate the performance degradation caused by blocking cleanings by simulating idle periods during workload execution. During these artificially induced idle durations, I/O requests are delayed and processed in subsequent execution duration. This allows for the invocation of idle cleanings, ensuring that cleaning operations proceed uninterrupted for a predefined idle duration.

6.3.3 *HA-SMR filesystem CosaFS.* To the best of our knowledge, the work [240] termed as *CosaFS* (cooperative shingle-aware file system) is the first work that has presented an HA-SMR based filesystem. The basic idea of CosaFS is to classify objects as hot or cold, where hot objects are served by SSDs and cold objects are stored on SMR drives. The authors conducted evaluations of CosaFS in comparison to filesystems based on DM-SMR and HM-SMR technologies. The findings of this study demonstrate improvements in throughput and reductions in latency when compared to DM-SMR and HM-SMR baseline performances. While the authors highlight the separation of metadata from file data as a primary distinguishing feature of CosaFS, aiming to optimize SMR drive bandwidth utilization, our research suggests that similar decoupling of metadata is also feasible in various HM-SMR-based file systems.

Question: Why lower adaptability?
Enterprises/Users/Customers want a **clear contract** and **control**



Hybrid Shingled Magnetic Recording (H-SMR)

Dynamic Hybrid – Shingled Magnetic Recording (H-SMR)

H-SMR Overview:

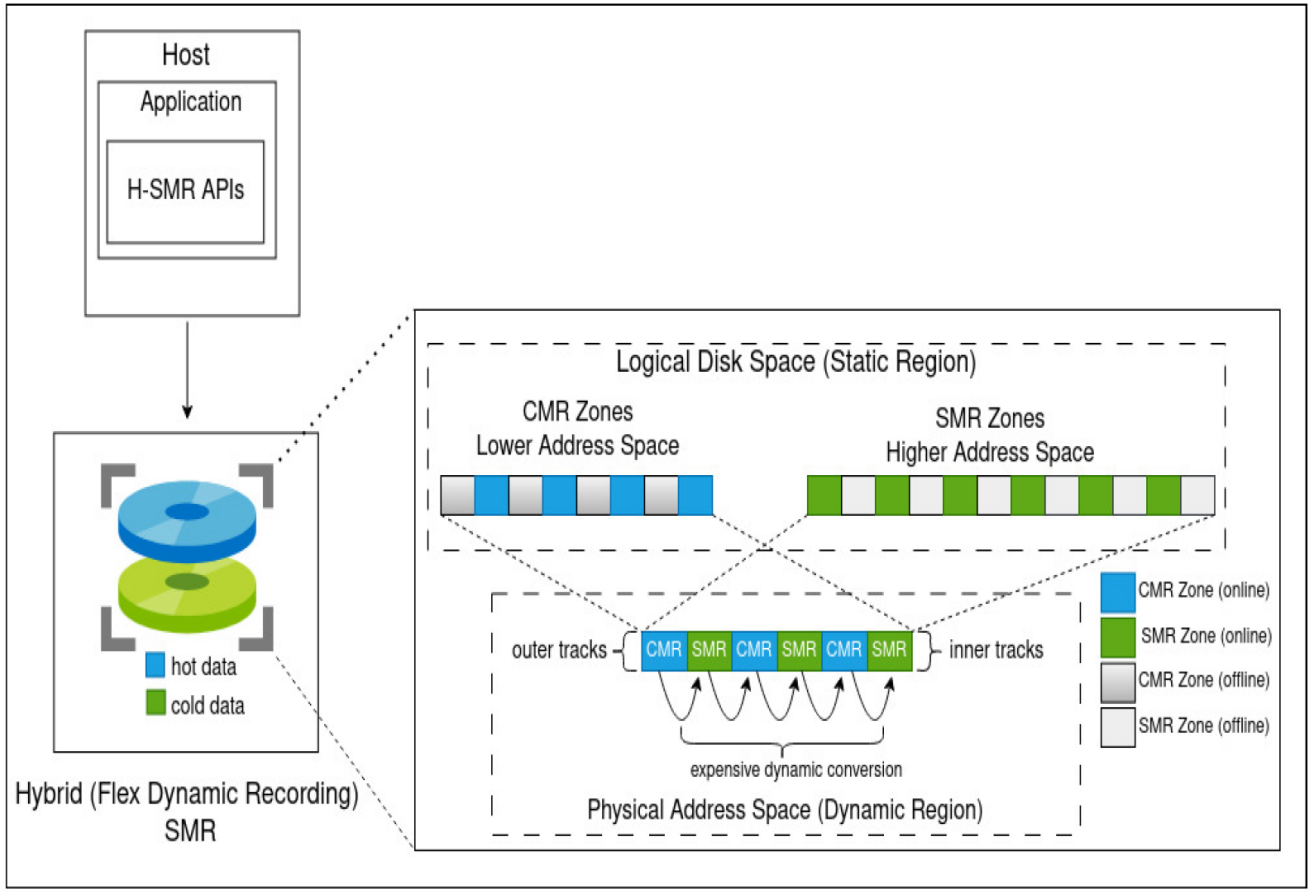
- Combines CMR zones (hot data) + SMR zones (cold data).
- Host can dynamically reconfigure zones via H-SMR APIs.
- Hot data → CMR: in-place updates, avoids SMR overhead.
- Cold data → SMR: sequential writes, high density.

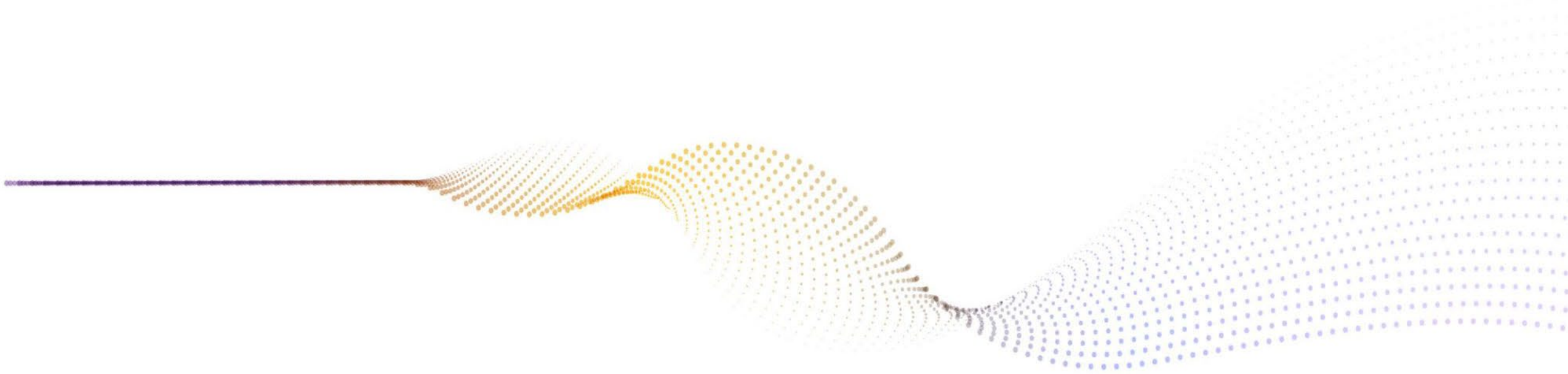
Architecture:

- **CMR zones**: lower LBA, short-stroked → low latency, high IOPS
- **SMR zones**: higher LBA → high capacity, cold data storage.
- **Zone conversion:** (online/offline) changes usable zone count due to areal density differences.Key factor in DM-SMR performance

Research Themes:

- Limited: only 3 key studies





3. SMR and HAMR 2025 onwards

Overall lessons learn and extending SMR to Future Disk Technologies

Overall lessons learnt:

- SMRs remain highly relevant.
- SMRs still untapped in non-hyperscalers
- Good with WORM
- **Most important**: paves way for future disks

In my paper, following **four** disk technologies were discussed:

- **TDMR** (Two-Dimensional Magnetic Recording)
- **MAMR** (Microwave-Assisted Magnetic Recording)
- **IMR** (Interlaced Magnetic Recording)
- **HAMR** (Heat Assisted Magnetic Recording)

Main focus?

- Obviously **HAMR**
  - 400+ patents, strong research focus
  - System-level study still underexplored

Two-Dimensional Magnetic Recording (TDMR) [24] and Microwave-Assisted Magnetic Recording (MAMR) [248] represent two key advancements in magnetic recording technologies. TDMR improves areal density by using multiple read heads to enhance signal quality and reduce interference between tracks, but it does not completely eliminate the inherent limitations of SMR. On the other hand, MAMR is an energy-assisted magnetic recording technology that uses microwave energy to assist in the recording process. This allows for higher data density without compromising write performance and is viewed as a promising step toward more efficient and scalable disk storage solutions.

Figure 12. Furthermore, as sketched in Figure 12, we will discuss another variant of HAMR media known as Interlaced Magnetic Recording (IMR). Both of these technologies are currently the subject of active investigation within the storage research domain.

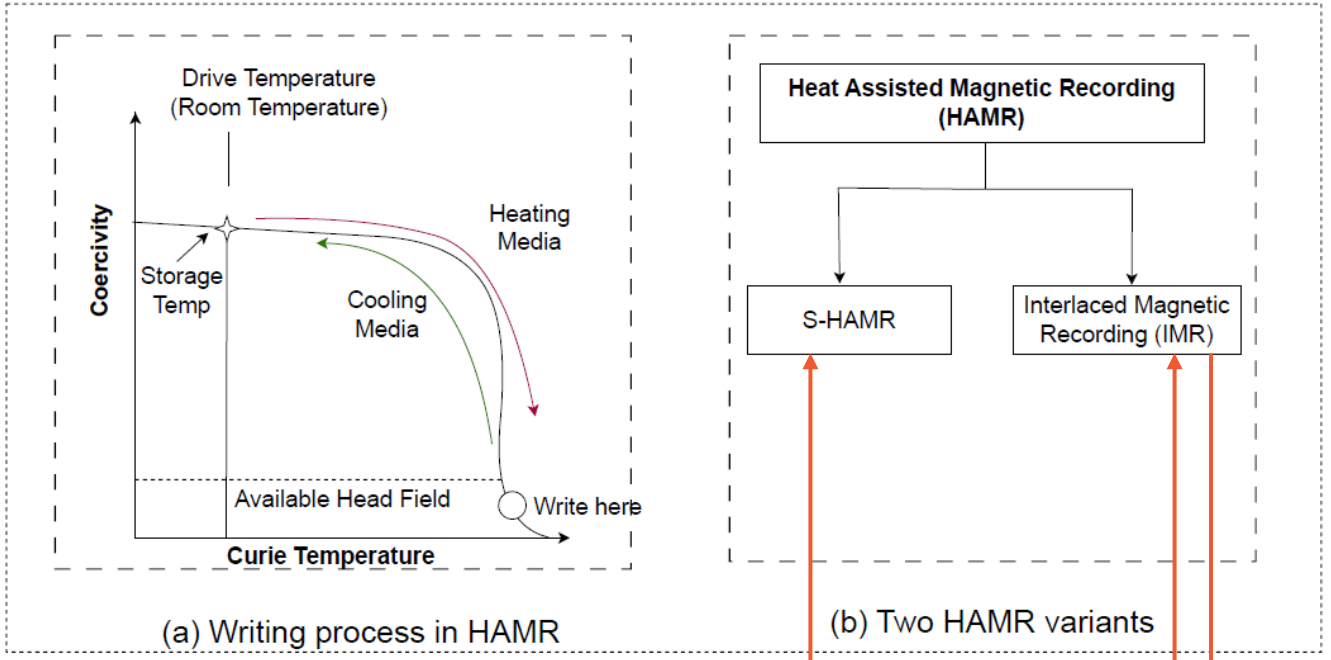
Additionally, technologies such as bit patterned media (briefly discussed in Section 10) are expected to be implemented, which define separate islands of magnetic media. One significant approach in energy-assisted magnetic recording is Heat Assisted Magnetic Recording (hereafter HAMR).

HAMR has emerged as the leading candidate for the next generation of magnetic recording technology, anticipated to surpass densities of 1.5 $Tbit/in^2$ and potentially reaching densities of up to 5 $Tbit/in^2$. Moreover, substantial research in HAMR is evident from the filing of over 400 patents for HAMR drives [40]. Although research on HAMR spans several decades, for brevity, the pivotal

Heat Assisted Magnetic Recording (HAMR)

HAMR principle:

- A **laser heats** the writing spot temporarily
- Heat reduces coercivity → magnetic field can write to **smaller region**.
- HAMR write head = **magnetic writer + laser diode**
- Enables high areal-density
- Boosts storage capacity

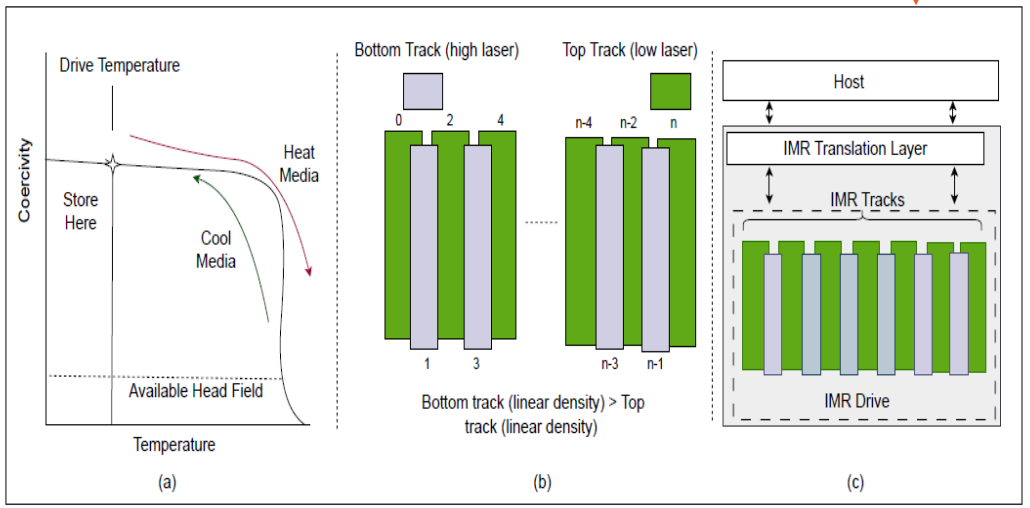


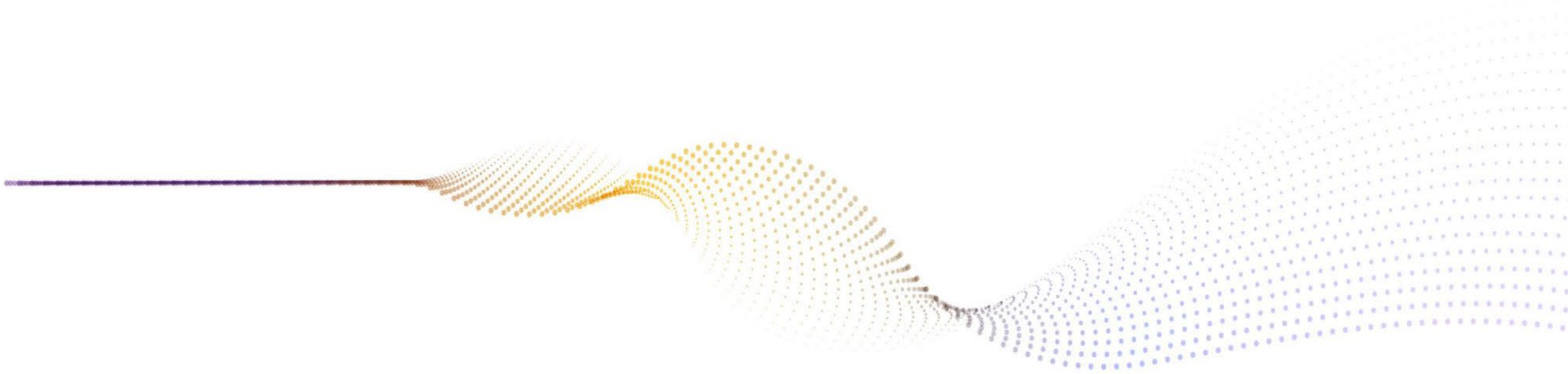
Two variants were discussed in our paper:

- Shingled-HAMR (S-HAMR)
- Heat Interlaced Magnetic Recording (HIMR)

Heat Interlaced Magnetic Recording (HIMR):

- Proposed in **2016**
- Combines **HAMR's heat-assisted writing** with **interlaced track layout**



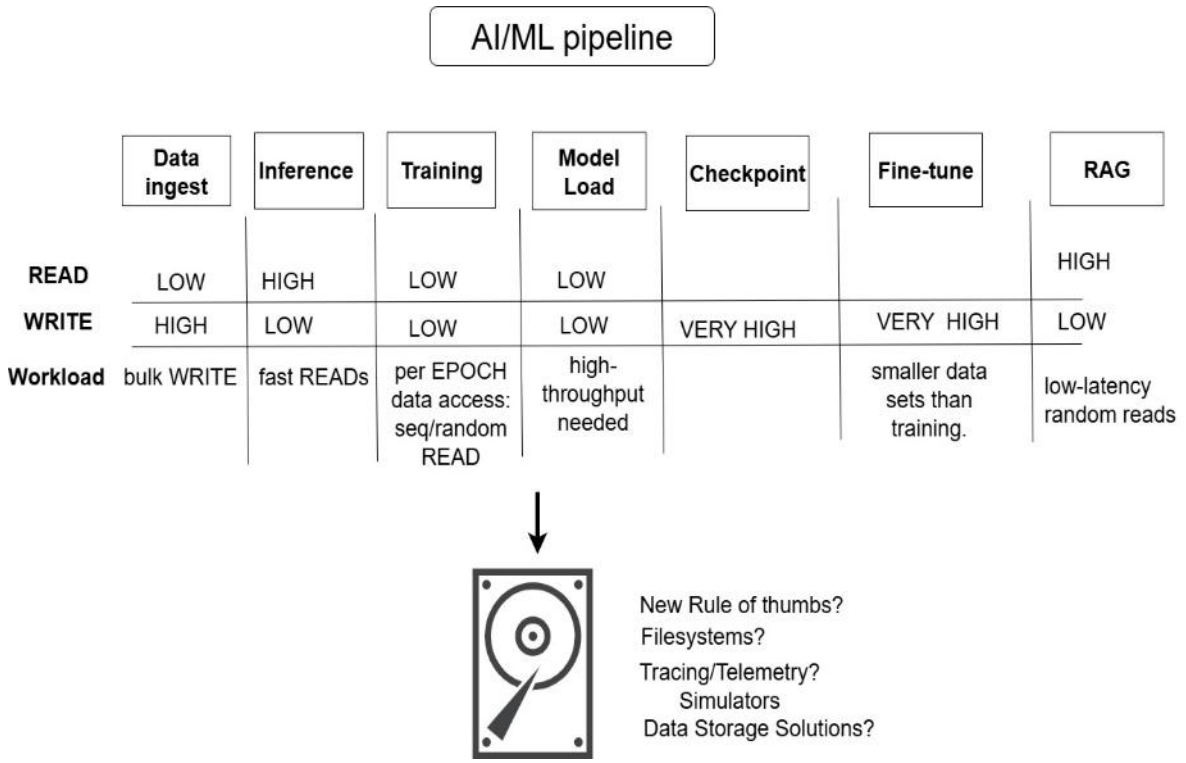


SMR+ HAMR = Systems

SMR+ HAMR = Systems and Application

What's next?

From materials to machines: the next leap is **systems** and **application**



What in Systems?

1. New Rule of thumbs?
2. Filesystems?
3. Simulators?
4. Data storage solutions: Deduplication and RAID

Larger disks and existing rules of thumbs

Citing two* papers (academia + industry) that has compelled to revisit rules of thumbs (not our papers, we are merely citing):

Two rule of thumbs in CMRs:

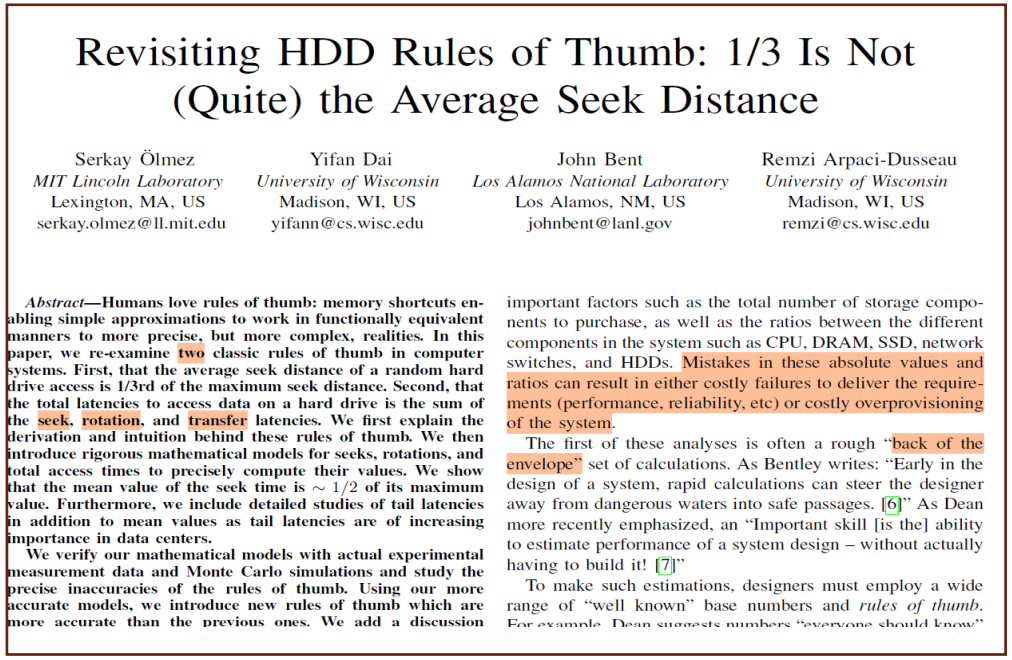
1. **Rule of thumb 1:** Average seek distance ≈ 1/3 of the full seek distance
2. **Rule of thumb 2**: seek latency + rotational latency + transfer latency

Question: Do these rule of thumbs are relevant to SMRs and HAMR?

Answer: It would be interesting to revisit. Why?

Future larger disks may break the rule:

- 1/3rd rule remains useful approximation for current disks (~4% error)
- As platter sizes increase (larger radius, more tracks), error will grow.
- For high-precision modeling, 1/3rd shortcut may no longer be acceptable.



Revisiting HDD Rules of Thumb: 1/3 Is Not (Quite) the Average Seek Distance

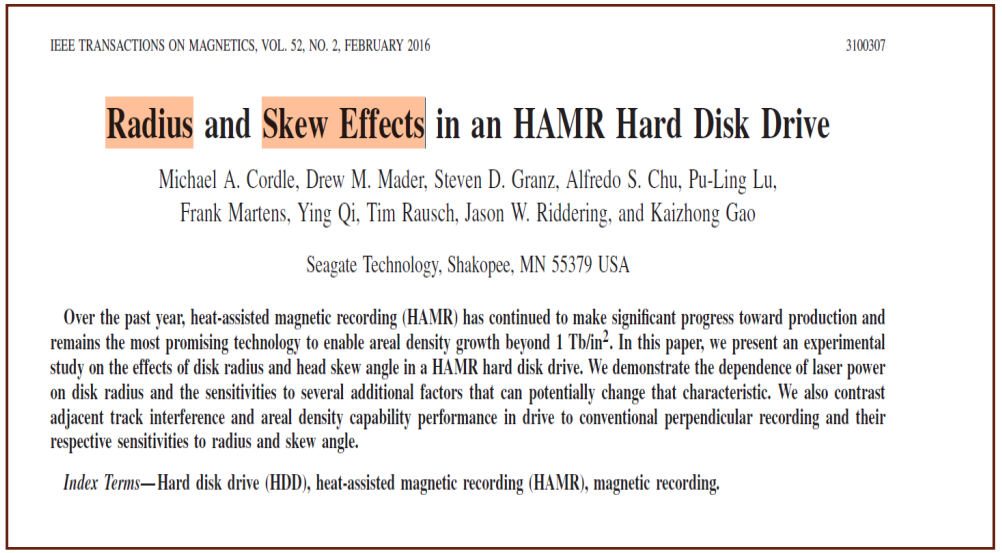
Serkay Ölmez, *MIT Lincoln Laboratory*, Lexington, MA, US, serkay.olmez@ll.mit.edu
Yifan Dai, *University of Wisconsin*, Madison, WI, US, yifann@cs.wisc.edu
John Bent, *Los Alamos National Laboratory*, Los Alamos, NM, US, johnbent@lanl.gov
Remzi Arpaci-Dusseau, *University of Wisconsin*, Madison, WI, US, remzi@cs.wisc.edu

***Abstract*—Humans love rules of thumb: memory shortcuts enabling simple approximations to work in functionally equivalent manners to more precise, but more complex, realities. In this paper, we re-examine two classic rules of thumb in computer systems. First, that the average seek distance of a random hard drive access is 1/3rd of the maximum seek distance. Second, that the total latencies to access data on a hard drive is the sum of the seek, rotation, and transfer latencies. We first explain the derivation and intuition behind these rules of thumb. We then introduce rigorous mathematical models for seeks, rotations, and total access times to precisely compute their values. We show that the mean value of the seek time is ∼ 1/2 of its maximum value. Furthermore, we include detailed studies of tail latencies in addition to mean values as tail latencies are of increasing importance in data centers.**

We verify our mathematical models with actual experimental measurement data and Monte Carlo simulations and study the precise inaccuracies of the rules of thumb. Using our more accurate models, we introduce new rules of thumb which are more accurate than the previous ones. We add a discussion important factors such as the total number of storage components to purchase, as well as the ratios between the different components in the system such as CPU, DRAM, SSD, network switches, and HDDs. Mistakes in these absolute values and ratios can result in either costly failures to deliver the requirements (performance, reliability, etc) or costly overprovisioning of the system.

The first of these analyses is often a rough "back of the envelope" set of calculations. As Bentley writes: "Early in the design of a system, rapid calculations can steer the designer away from dangerous waters into safe passages. [6]" As Dean more recently emphasized, an "Important skill [is the] ability to estimate performance of a system design – without actually having to build it! [7]"

To make such estimations, designers must employ a wide range of "well known" base numbers and *rules of thumb*. For example, Dean suggests numbers "everyone should know"



IEEE TRANSACTIONS ON MAGNETICS, VOL. 52, NO. 2, FEBRUARY 2016 3100307

Radius and Skew Effects in an HAMR Hard Disk Drive

Michael A. Cordle, Drew M. Mader, Steven D. Granz, Alfredo S. Chu, Pu-Ling Lu, Frank Martens, Ying Qi, Tim Rausch, Jason W. Riddering, and Kaizhong Gao

Seagate Technology, Shakopee, MN 55379 USA

Over the past year, heat-assisted magnetic recording (HAMR) has continued to make significant progress toward production and remains the most promising technology to enable areal density growth beyond 1 Tb/in^2. In this paper, we present an experimental study on the effects of disk radius and head skew angle in a HAMR hard disk drive. We demonstrate the dependence of laser power on disk radius and the sensitivities to several additional factors that can potentially change that characteristic. We also contrast adjacent track interference and areal density capability performance in drive to conventional perpendicular recording and their respective sensitivities to radius and skew angle.

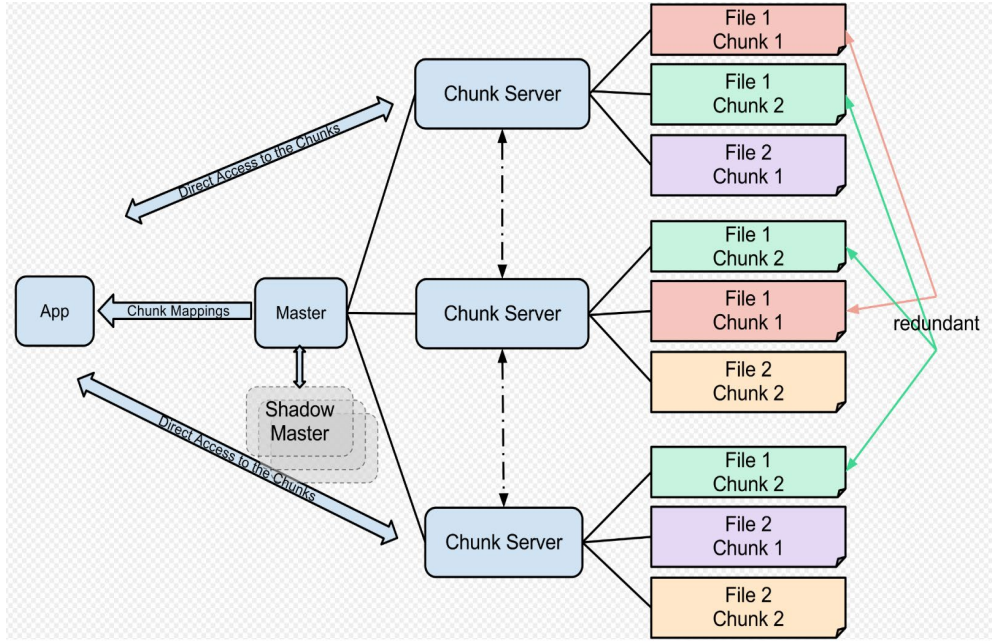
Index Terms—Hard disk drive (HDD), heat-assisted magnetic recording (HAMR), magnetic recording.

** Revisiting HDD Rules of Thumb: 1/3 Is Not (Quite) the Average Seek Distance*
Serkay Ölmez MIT Lincoln Laboratory,Yifan Dai University of Wisconsin,John Bent Los Alamos National Laboratory,Remzi Arpaci-Dusseau

** Radius and Skew Effects in an HAMR Hard Disk Drive*
Michael A. Cordle, Drew M. Mader, Steven D. Granz, Alfredo S. Chu, Pu-Ling Lu, Frank Martens, Ying Qi, Tim Rausch, Jason W. Riddering, and Kaizhong Gao

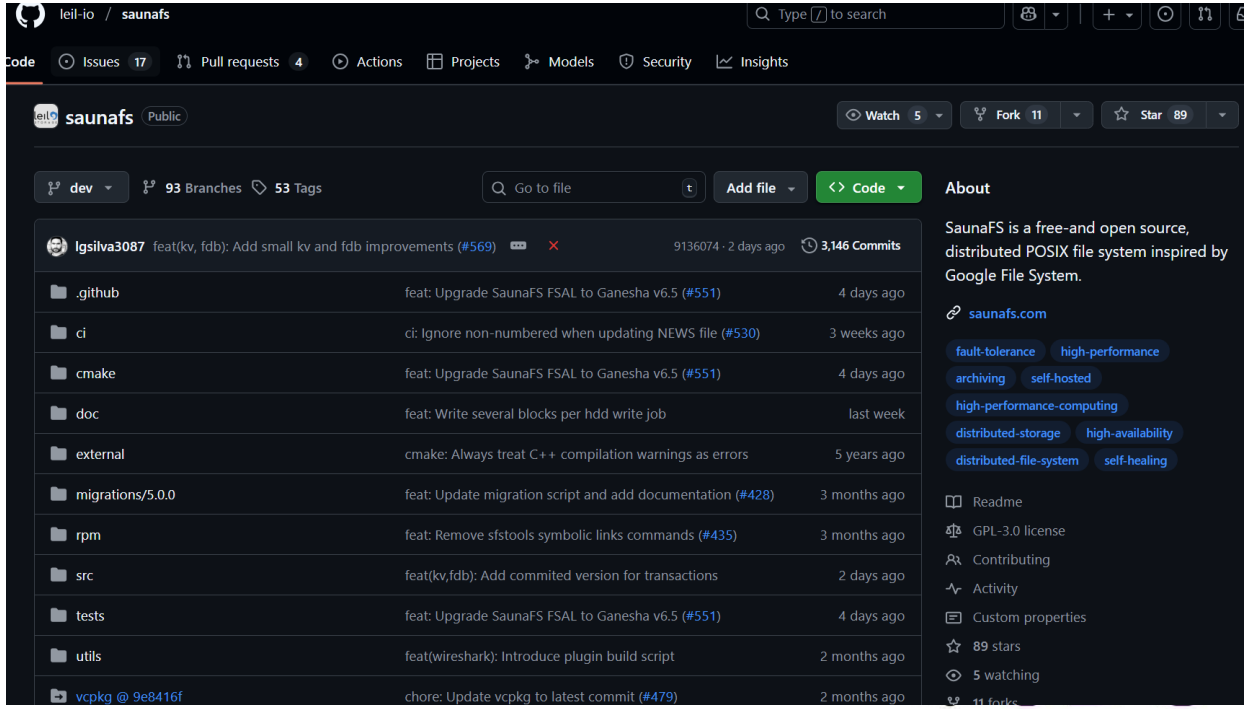
SaunaFS (Leil Storage):

What is SaunaFS?
- a distributed, exabyte-scale file system from Leil Storage
- has roots in Google Filesystem
- primarily supports Host-Managed SMR (HM-SMR)



Where does HAMR fits in SaunaFS?
- SaunaFS 5.0 for HAMR
- SaunaFS 5.0 for HAMR + SMR

SaunaFS codebase?
- open-sourced, C++ (of course), and active



Simulation of HAMR drives

Why simulation?

A significant part of SMR academic and industrial research is built on the foundation of simulators.

HAMR simulators and granularity?

In my research, I found following two* academic references (fine work – moved the needles):

- **Work:** *"Simulation of Heat Assisted Magnetic Recording System" et.al Jao*
  Granularity: work centers on simulation methodologies for HAMR (media generation, recording, playback, and state detection)

- **Work**: *"System and Media Optimizations for improved HAMR Performance" et.al Natekar*
  Granularity: work targets system and media optimization techniques for improved HAMR performance.

** Simulation of Heat Assisted Magnetic Recording System, Yipeng Jiao, 2018*
** System and Media Optimizations for improved HAMR Performance, RANDALL H. VICTORA, 2020*

RAID (black-art)

RAID 0 – Data striping, no fault tolerance.
RAID 1 – Disk mirroring
RAID 2 – Bit-level striping for error correction (rarely used).
RAID 3 – Byte-level striping (obsolete).
RAID 4 – Block-level striping with dedicated parity disk
RAID 5 – Block-level striping with distributed parity (most common)
RAID 6 – Like RAID 5 but with dual parity, survives two disk failures.

8.1.1 HiSMRfs: HiSMRfs [99] is a file system designed specifically for SMR drives, featuring a standard POSIX interface. HiSMRfs incorporated a file/object-based RAID 5 module tailored for SMR/hard disk drive (HDD) arrays, which computed parity for individual files or objects, ensuring that both data and parity writing occurred sequentially and fully within a stripe. This file system was also compatible with hybrid storage systems that integrated conventional HDDs and SSDs.

8.1.2 RAID 4S:. Le *et. al* [114, 116] proposed a novel application of SMR disks in conjunction with conventional disks, termed RAID 4S, which had the potential to enhance the performance of storage systems. Their evaluation demonstrated the feasibility of utilizing SMR disks within a RAID 4 array, outperforming the use of SMR disks in standard RAID 4 configurations with in-place updates.

8.1.3 DVS:. The authors [140] proposed a dynamic variable-width striping RAID specifically designed for shingled write disks (SWDs) to reduce the costs associated with parity updating. Their contributions were summarized as follows: First, they introduced DVS-RAID for SWDs, which dynamically generated new full or partial stripes and appended new data to the tail of the existing data. Second, they designed a write cache management system that took into account the unique properties of SWDs for DVS-RAID. Finally, they implemented a DVS-RAID simulator and evaluated its performance under various workloads. The results indicated that DVS-RAID outperformed traditional HDD-based RAID systems, particularly under read-dominant and sequential write-dominant workloads.

8.1.4 HSMR-RAID:. Lin *et. al* [130] proposed an SMR-friendly RAID 5 storage system, named HSMR-RAID, which was developed within a pure SMR environment without the use of additional storage drives. HSMR-RAID was based on a host-managed SMR architecture. The fundamental

"***The chain is only as strong as its weakest link***"
mixing SMRs and CMRs in RAID was not a preferable choice.
Performance of a RAID array that mixes SMR and CMR drives would be similar to an SMR-only RAID array.

↳ in-short: **RAID's fixed** data distribution pattern (striping + parity) assumes disk can handle random updates

Deduplication

1. SMR-aware Fingerprint Store (2018, Wu et al.):
2. SMRTS (2022):
3. LaDy (Recent):

SMR-aware Fingerprint Store (2018):

- Joint dedup + HM-SMR integration
- Used *chunk lifetime* to reduce reclamation delay
- Evaluated with Skylight emulator + fio tool

SMRTS (2022):

- SSD + SMR tiered storage
- SSD-tier: metadata, index, hot files
- SMR-tier: deduplicated cold files stored sequentially

LaDy (2023):

- Locality-Aware Deduplication
- Selectively writes duplicates to preserve locality

8.2.1 SMR-aware Fingerprint Store. The first work that applied deduplication techniques to SMR was in 2018.. Wu *et al.* discussed that naively applying deduplication techniques on SMR drives may downgrade the runtime performance of data storage services, because of the time-consuming SMR space reclamation processes. Therefore, the authors advocated a vertical integration solution

8.2.2 SMRTS. [21] The authors designed and implemented SMRTS, a performance optimized file system for SSD **SMR** **T**iered **S**torage. In SMRTS, SSDs and SMR drives were used as two storage tiers, called the SSD-tier and SMR-tier, respectively. First, metadata, the deduplication index, and frequently accessed files (hot files) were stored in the SSD-tier. Periodically, the data of rarely accessed files (cold files) were migrated to the SMR-tier to reclaim space in the SSD-tier. During the migration to the SMR-tier, files were deduplicated, and only unique data chunks were accumulatively stored in a container buffer, which was written to the SMR-tier sequentially once full. When these deduplicated files were requested, they were restored to the SSD-tier as the original files. Additionally, a fine-grained read/write logic was designed.

8.2.3 LaDy. The researchers in a recent publication [25] proposed a framework, **L**ocality **A**ware **D**eduplication technolog**y**, named LaDy, to address both the overhead of writing duplicate data and the impact on data locality. The framework determined the necessity of writing duplicate data to optimize performance. The researchers integrated LaDy with DiskSim, an open-source simulation tool, and modified it to model an SMR-based drive. Experimental results demonstrated that the framework significantly reduced response times in the best-case scenario compared with the well-known deduplication method CAFTL on an SMR drive. LaDy achieved this by selectively writing duplicate data, thereby preserving data locality and improving read performance.

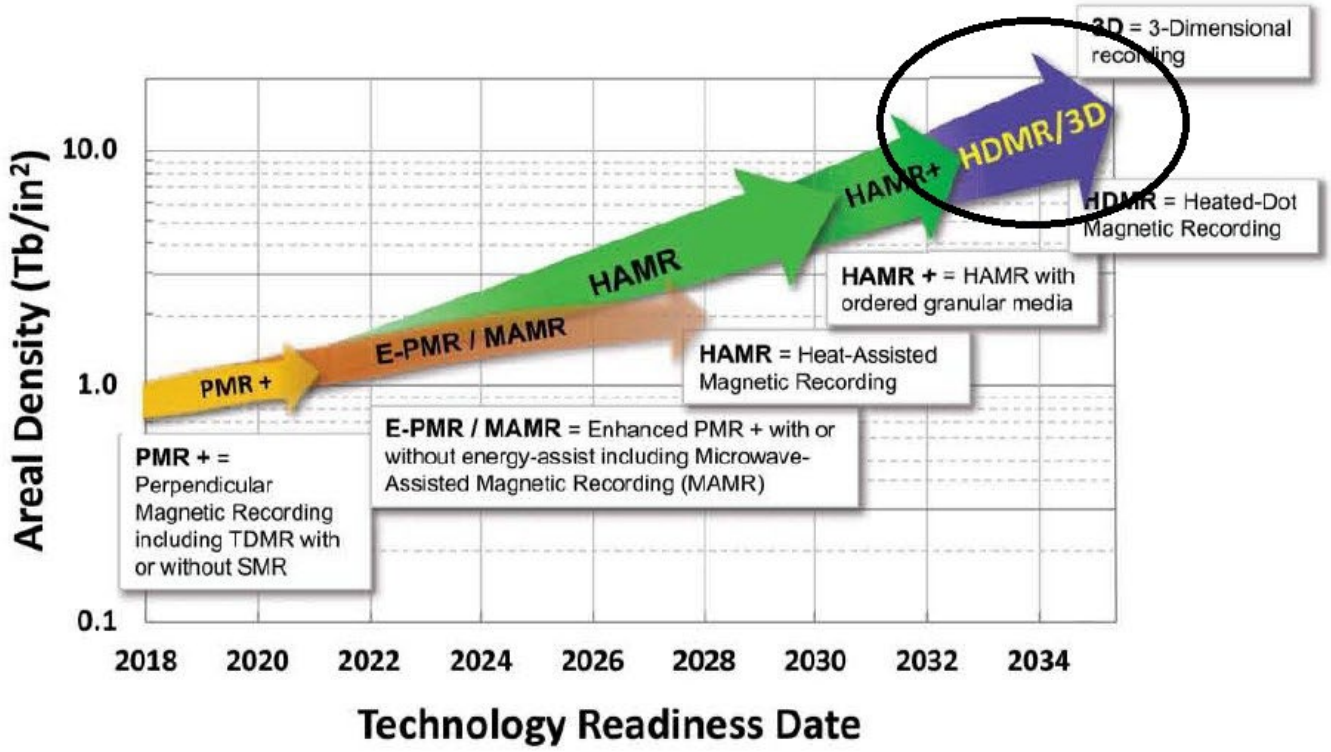
HAMR ≠ SMR:

- Characterize HAMR's write/latency/error profile under realistic dedup-induced write patterns
- HAMR-specific optimizations that leverage its higher density for deduplication benefits - like storing larger fingerprint indexes or enabling more aggressive dedup ratios due to increased capacity.

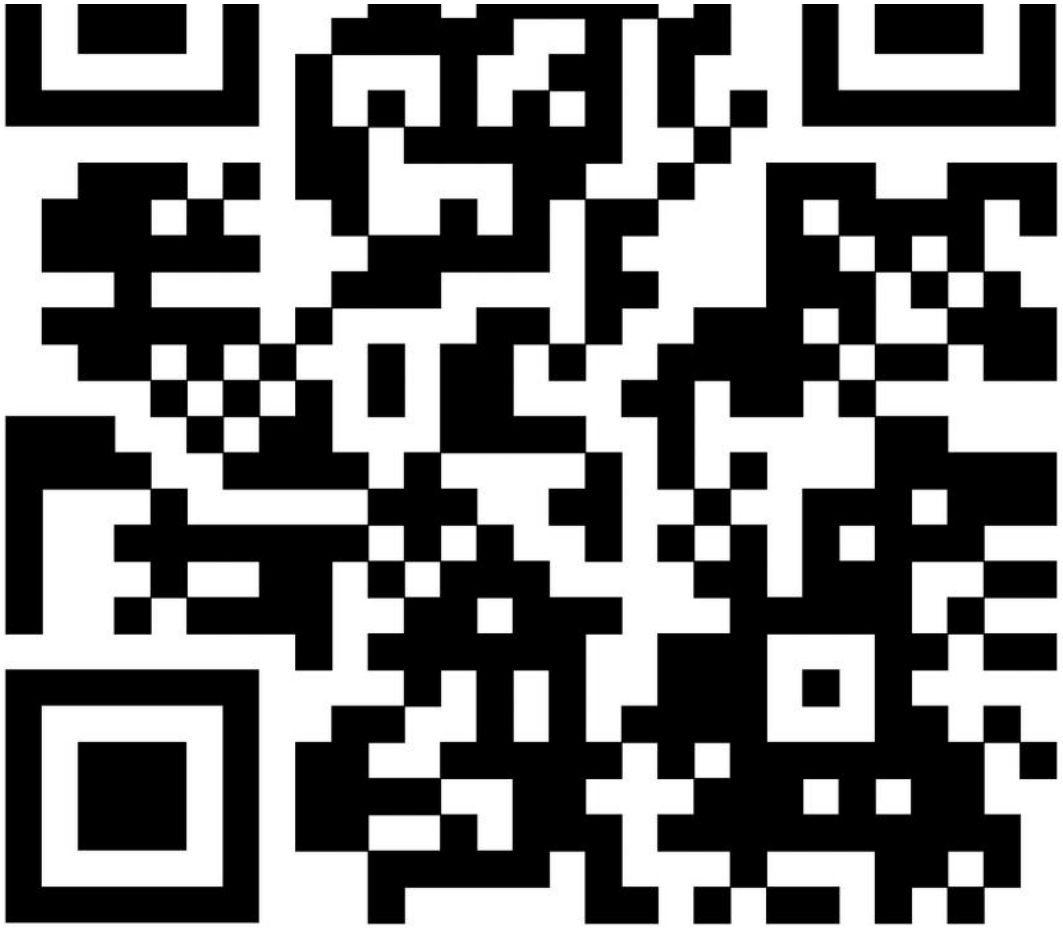
Heated **D**ot **M**agnetic **R**ecording (HDMR)

- **Definition:**
  HDMR = *Heat-Assisted Magnetic Recording (HAMR) + Bit Patterned Media (BPM)*.

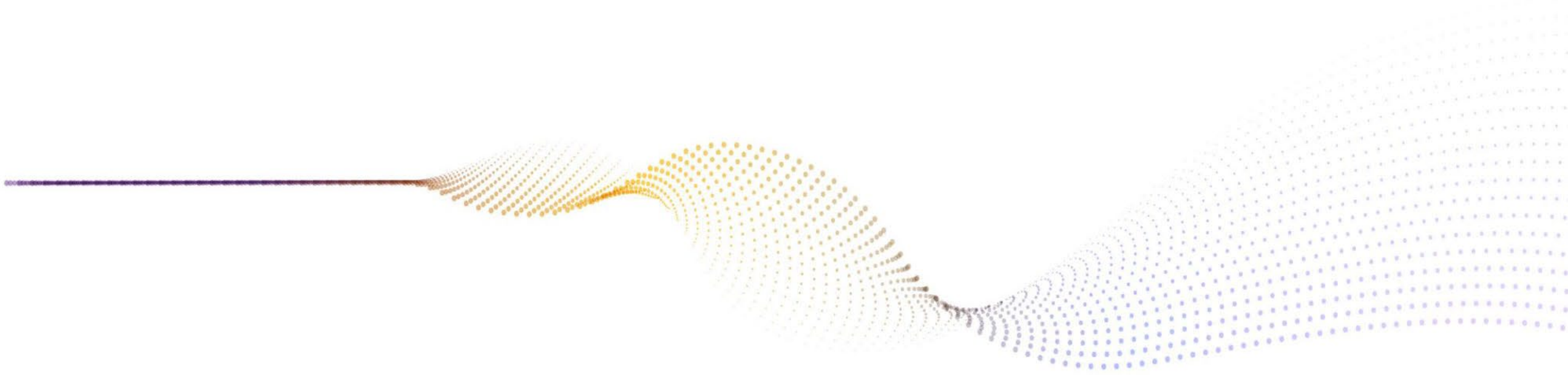
- **Target:**
  - ultra-high areal densities (potentially **>10 Tb/in²**).



Repository of **Reflecting 17 years of SMRs** publication



- BabarZKhan/shingled-magnetic-recording



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

References

- *Storage Systems for Shingled Disks*
  Garth Gibson Carnegie Mellon University and Panasas Inc, Anand Suresh, Jainam Shah, Xu Zhang, Swapnil Patil, Greg Ganger

- *A new Advanced Storage Research Consortium HDD Technology Roadmap*
  Eric Roddick, Mark Kief, and Akihiko Takeo

- *Revisiting HDD Rules of Thumb: 1/3 Is Not (Quite) the Average Seek Distance*
  Serkay Olmez, MIT Lincoln LaboratoryLexington, MA, US, Yifan Dai University of Wisconsin Madison, WI, US, John Bent Los Alamos National Laboratory Los Alamos, NM, US, Remzi Arpaci-Dusseau University of Wisconsin Madison, WI, US

- I/O Access Patterns in HPC Applications: A 360-Degree Survey
JEAN LUCA BEZ and SUREN BYNA, Lawrence Berkeley National Laboratory, USA SHADI IBRAHIM, Inria, University of Rennes, CNRS, IRISA, Rennes, France

- IEEE Roadmap Outlines Development of Mass Digital Storage Technology
Tom Coughlin , Coughlin Associates, Inc. Roger Hoyt, Consultant

- *System and Media Optimizations for improved HAMR Performance, RANDALL H. VICTORA, 2020*
- *Simulation of Heat Assisted Magnetic Recording System, Yipeng Jiao, 2018*

