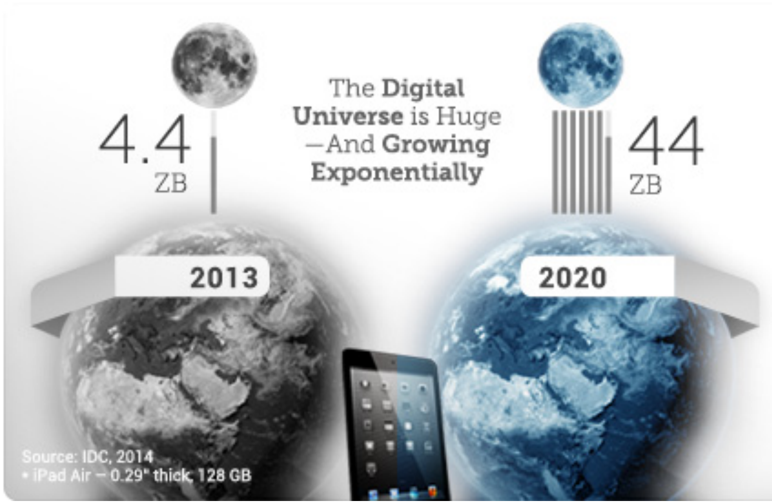
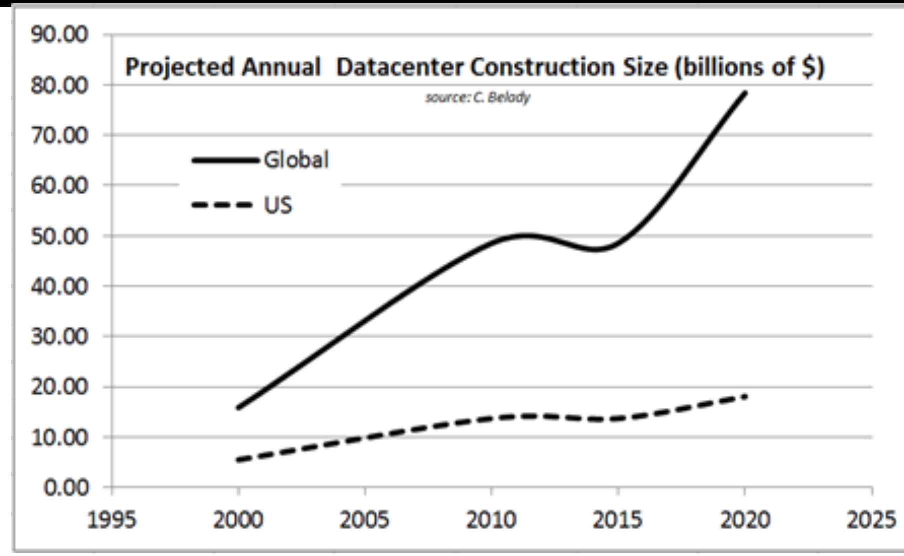
Facilitating Magnetic Recording Technology Scaling for Data Center Hard Disk Drives through Filesystem-level Transparent Local Erasure Coding

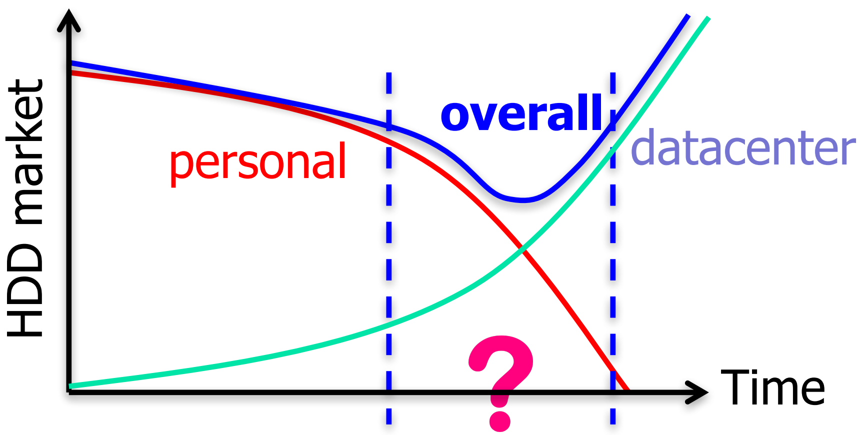
Yin Li, Hao Wang, Xuebin Zhang, Ning Zheng, Shafa Dahandeh, and Tong Zhang

Data Center HDDs: Rationale





➡ Data center: The main driver for future HDD market growth



? Minimize HDD $/GB

⬇

Data Centers

[1] E. Brewer et al., "Disks for data centers," Technical report, Google, 2016.

Data Center HDDs: Rationale

Exploit the characteristics of datacenter infrastructure & workloads

Relax the per-HDD reliability spec ➡ Lower manufacturing cost

Read retry rate: $<10^{-5}\sim10^{-6}$

Hard sector failure rate: $<10^{-12}\sim10^{-14}$

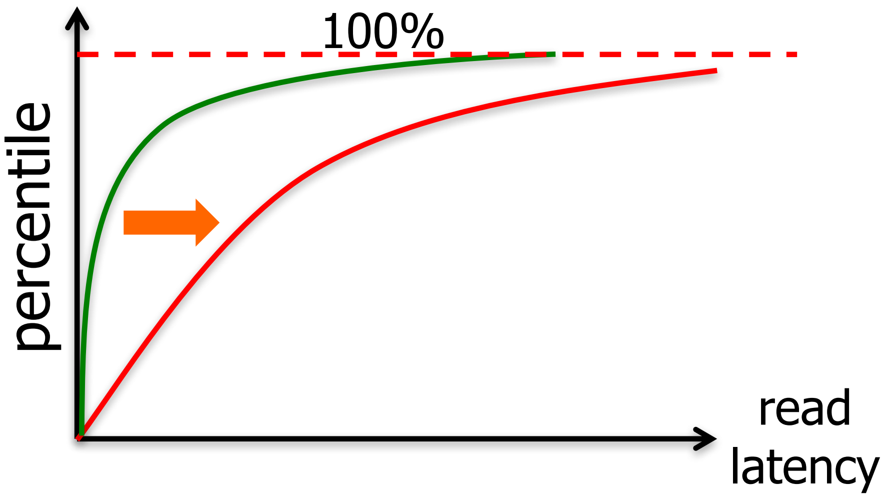
- The pervasive use of replication and distributed erasure coding to ensure system-level reliability in datacenters
- Dominantly coarse-grained HDD data access in datacenters

Data Center HDDs: Our First Step

? How datacenters can embrace HDD with relaxed read retry rate

Higher HDD read retry rate

Longer per-HDD tail read latency

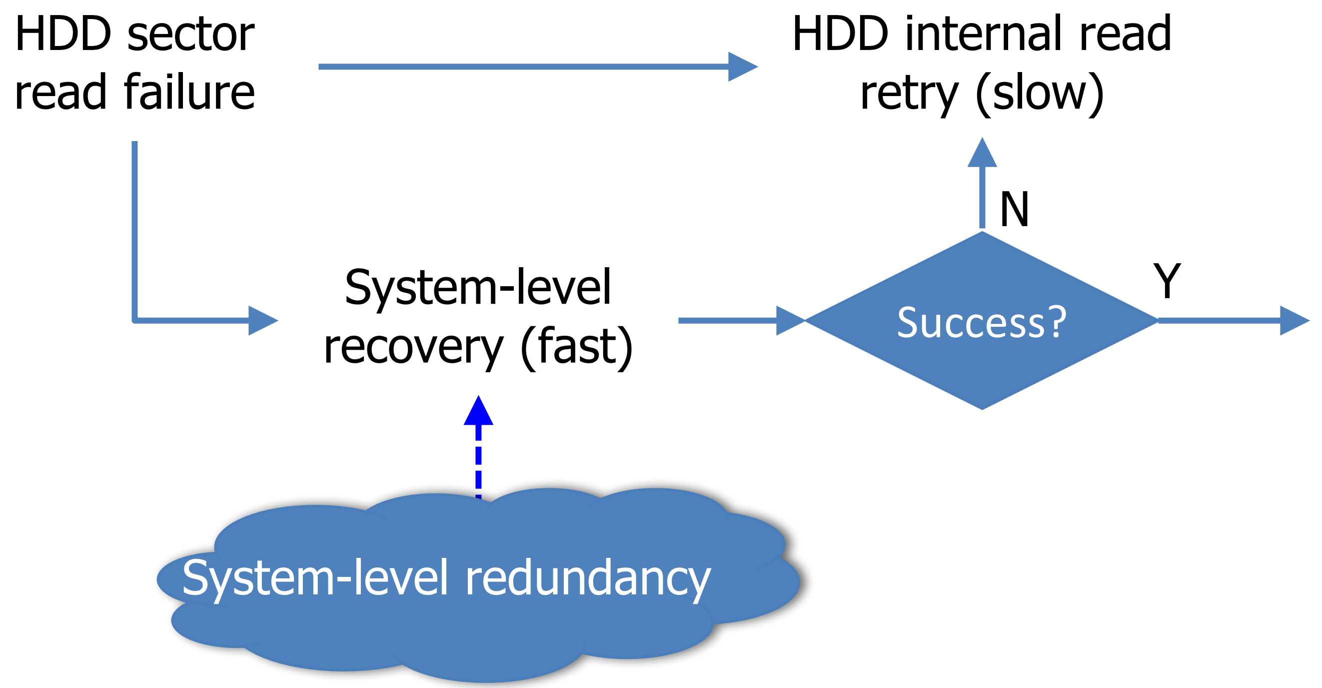


Effect will be amplified in large-scale systems (e.g., datacenters)

[1] J. Dean and L. A. Barroso, "The tail at scale," Communications of the ACM, 56:74–80, 2013.

Data Center HDDs: A Simple First Step

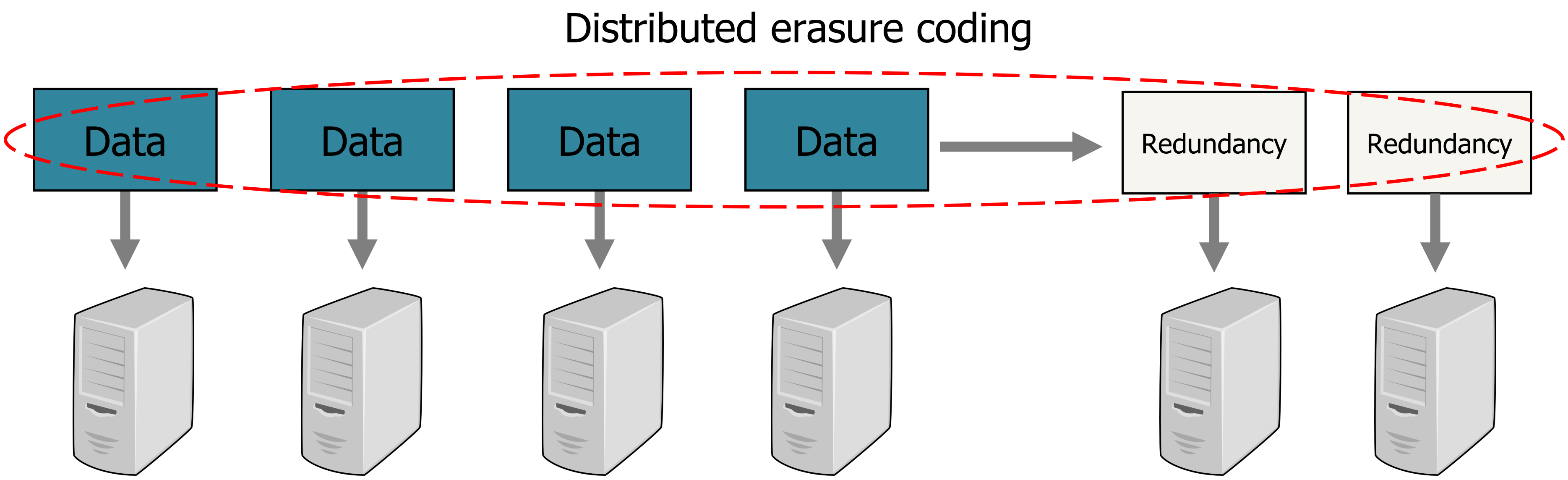
? How datacenters can embrace HDD with relaxed read retry rate



RAID 5/6 → Distributed erasure coding ← High network traffic

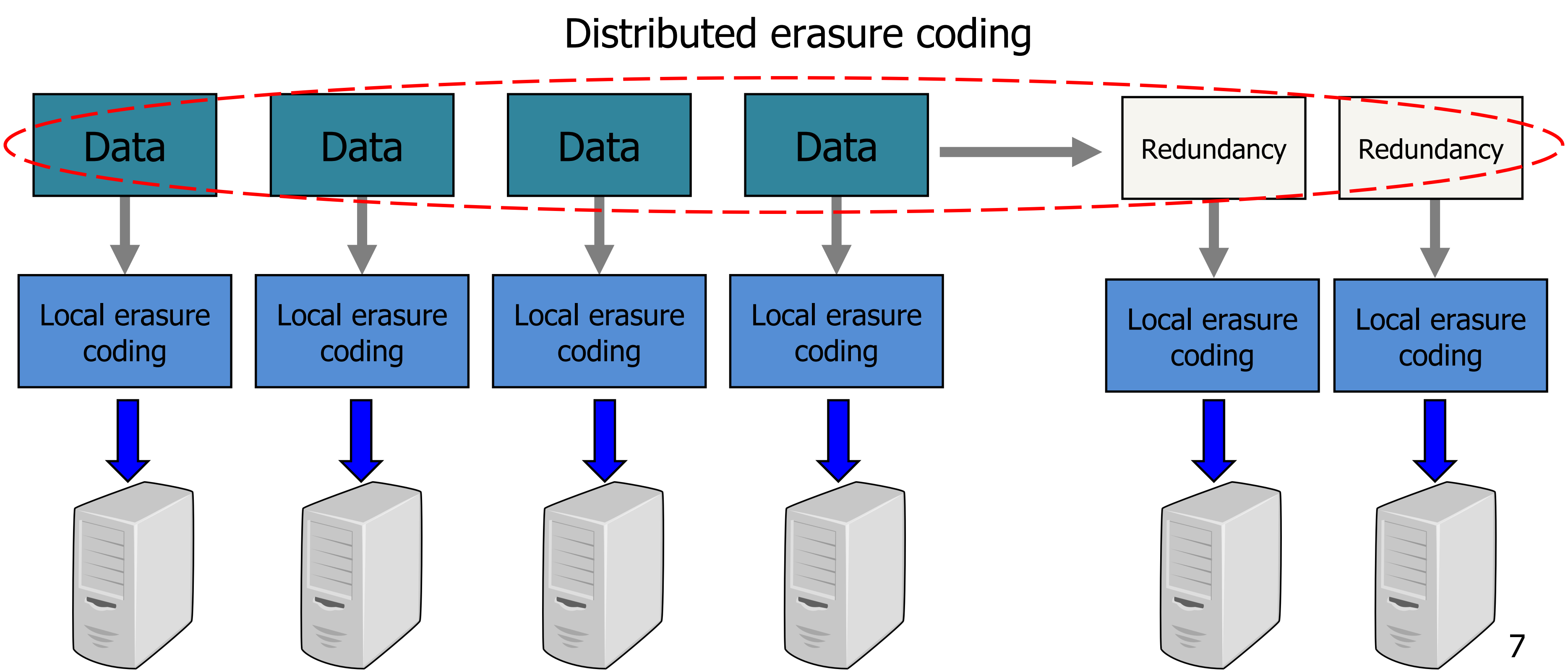
Hybrid Erasure Coding for Data Centers

- Distributed erasure coding: Mitigate catastrophic HDD failures & server unavailability at high coding redundancy (e.g., 25%~50%)



Hybrid Erasure Coding for Data Centers

- Distributed erasure coding: Mitigate catastrophic HDD failures & server unavailability at high coding redundancy (e.g., 25%~50%)
- Local erasure coding: Mitigate HDD sector read failures at low coding redundancy (e.g., 3% and below)

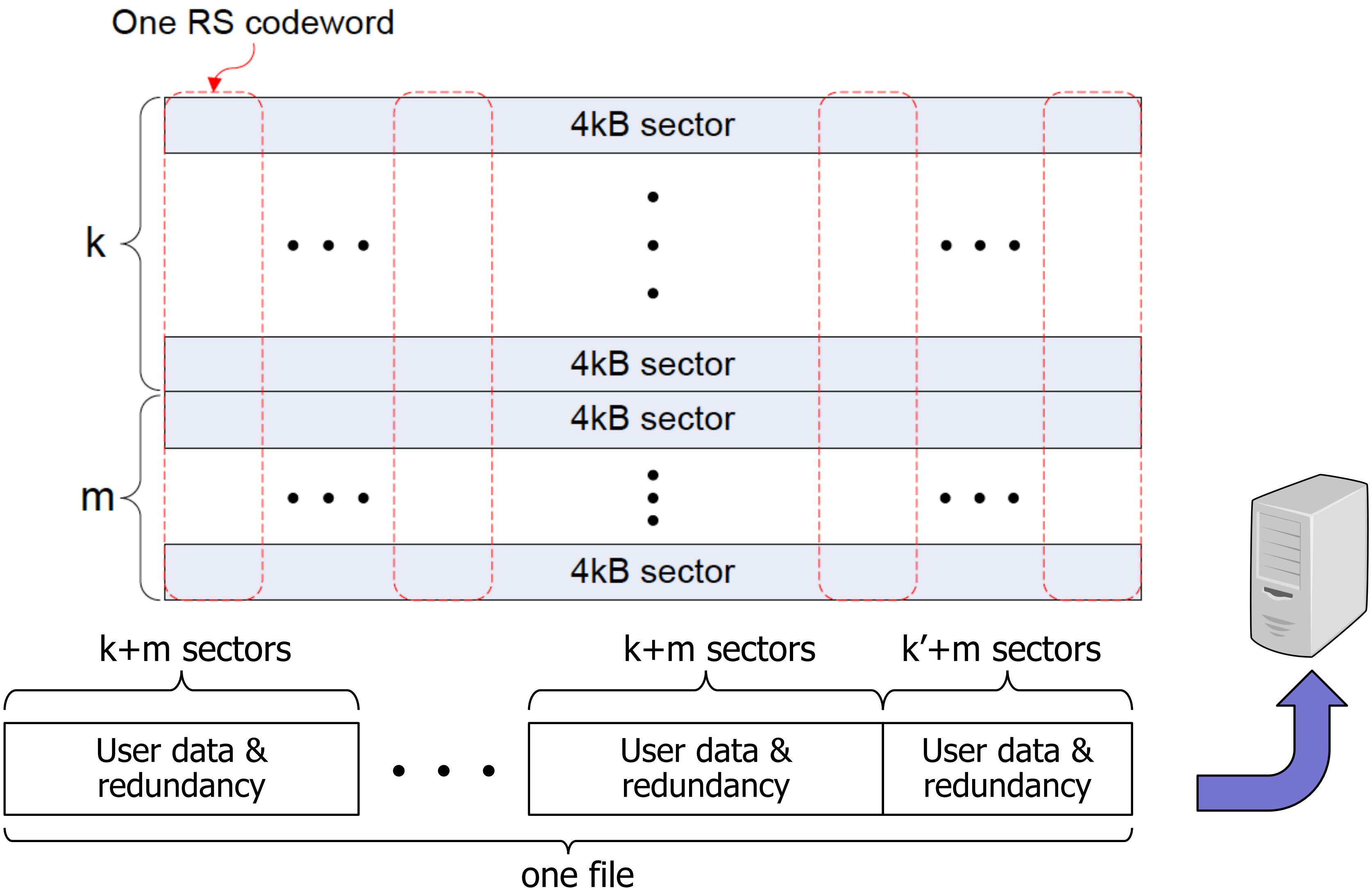


Simple Basic Concept

- **Local erasure coding**: data + coding redundancy on the same HDD

Filesystem-level Local Erasure Coding

- Per-file erasure coding for data & per-sector replication for metadata



Some Non-trivial Issues

? Mathematically formulate its effect on HDD read tail latency

? How to deal with unaligned HDD write and data update

? Impact of encoding/decoding on system speed performance

Tail Latency

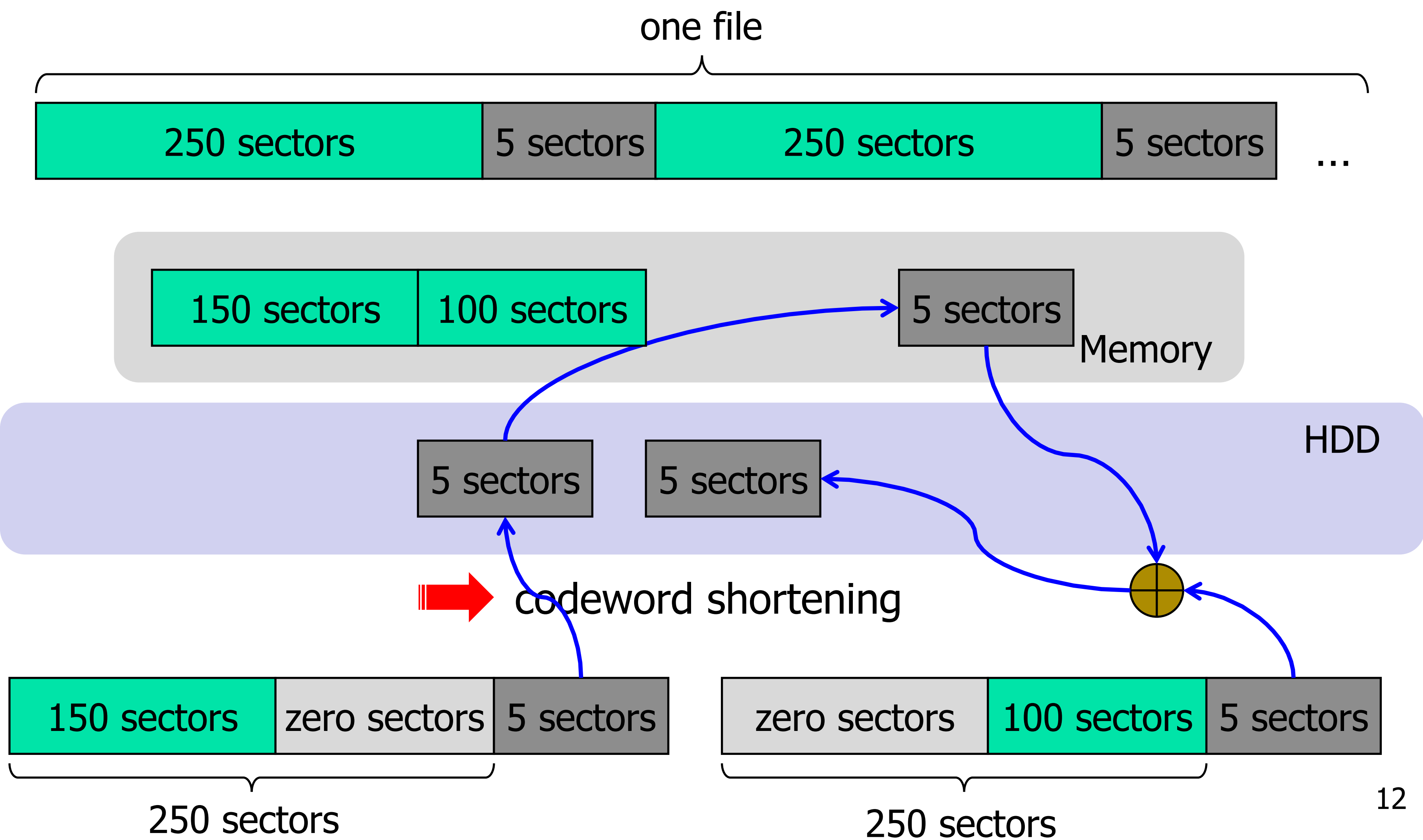
- Let T denote the latency to read N consecutive sectors from HDD
- Model T as a discrete variable and let f(T) denote its probability mass function
- Given the latency percentile P_{tail} (e.g., 99%), we search for the tail latency T_{tail} subject to

$$\sum_{T=0}^{T_{tail}} f(T) \geq P_{tail}$$

➡ Derived a set of mathematical formulations to estimate the data read tail latency when using local erasure coding (see the paper for details)

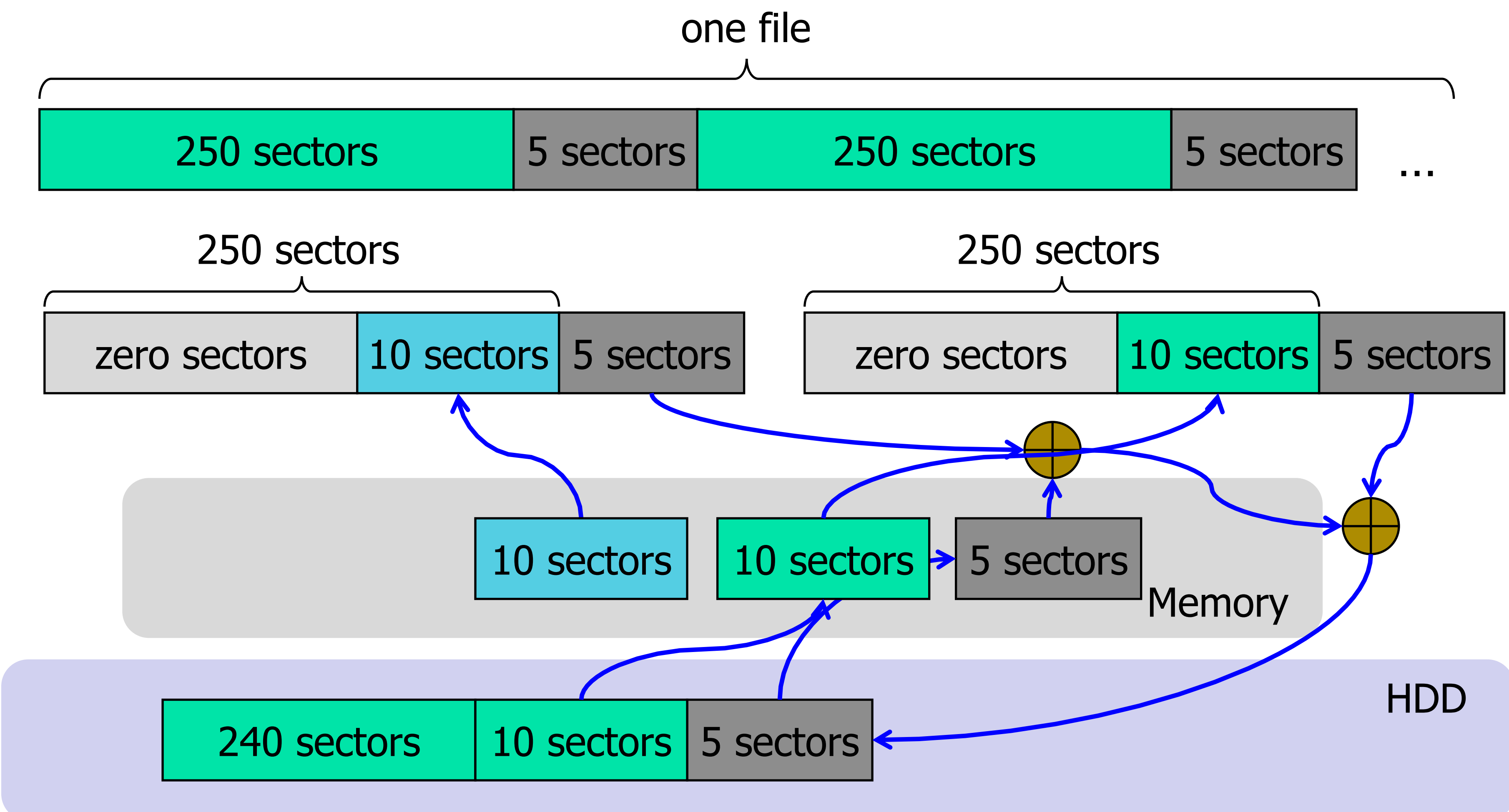
Unaligned HDD Write

- Use of (250, 5) local erasure code



Data Update

- Use of (250, 5) local erasure code



Analysis and Experimental Results

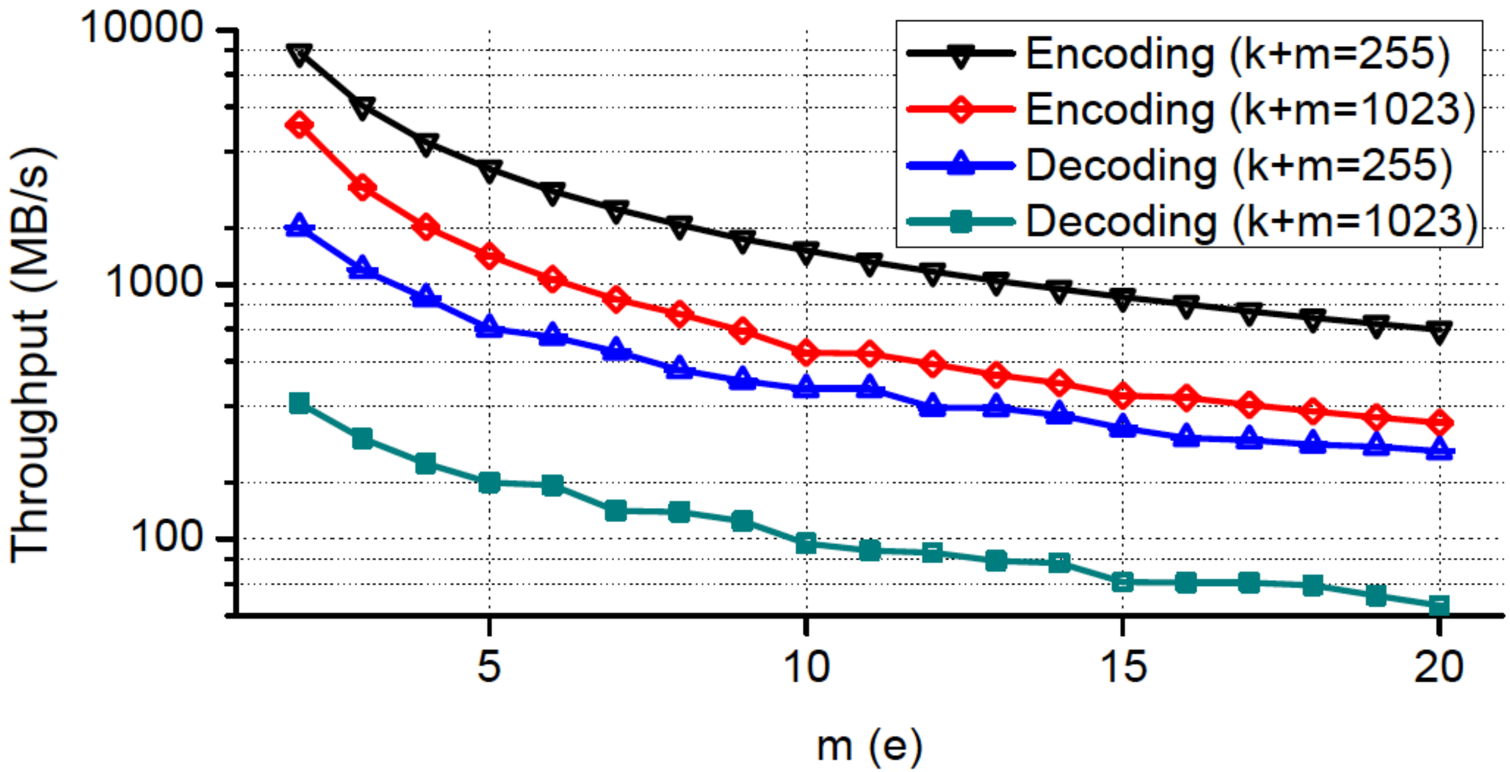
- RS-based local erasure codes with codeword length of 255 and 1023
- Relaxed HDD sector read failure probability: 1×10^{-4}, 5×10^{-4}, 1×10^{-3}, 5×10^{-3}
- Target local erasure code decoding failure probability: 1×10^{-8}

	255		**1023**	
	k	m	k	m
1×10^{-4}	252	3	1019	4
5×10^{-4}	251	4	1016	7
1×10^{-3}	250	5	1014	9
5×10^{-3}	246	9	1004	19

Encoding/Decoding Engine

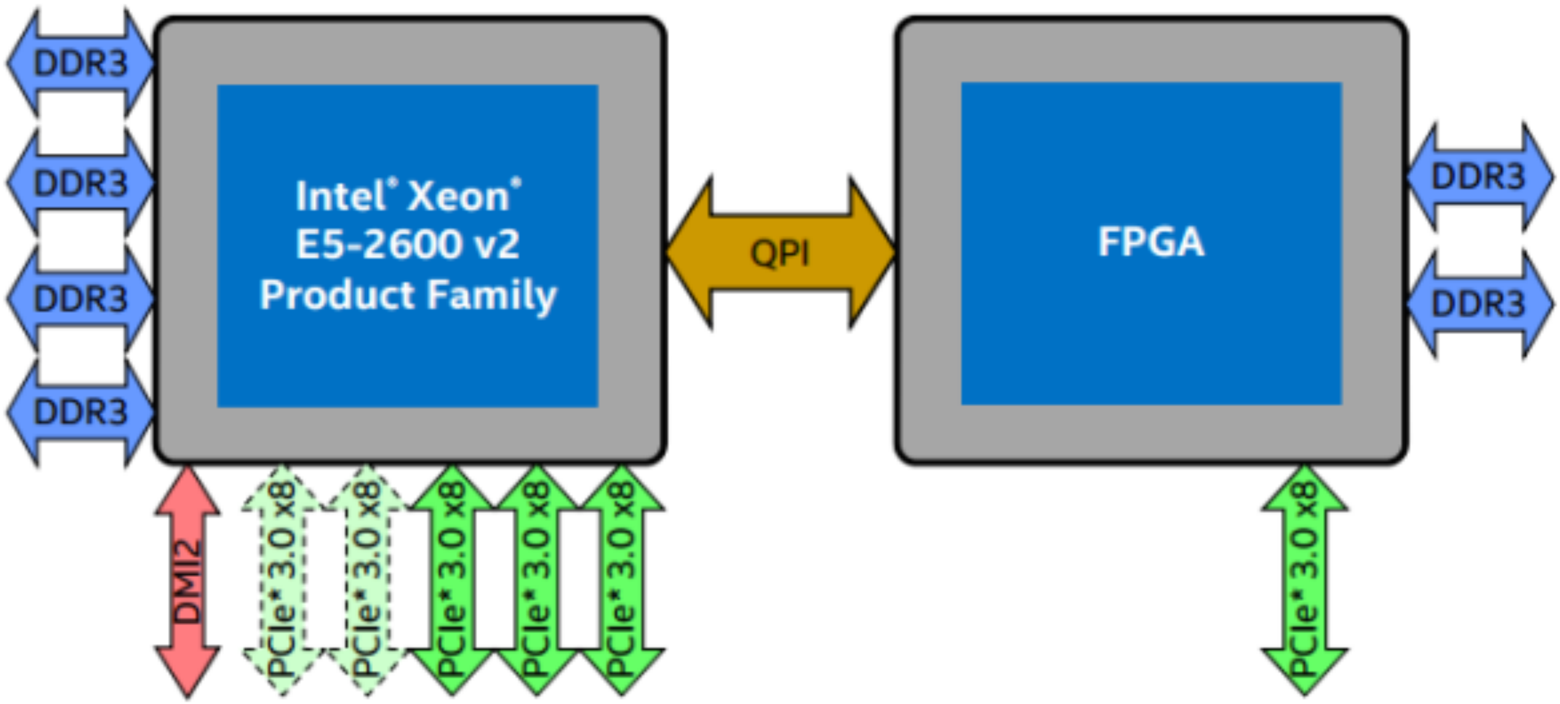
Software-based implementation

- Intel CPU 3.3GHz with 64kB L1, 256kB L2, and 6MB L3
- Matrix-based encoding and decoding
- Utilization of x86 SSE (Streaming SIMD Extensions) instructions



Encoding/Decoding Engine

Emerging CPU chip with built-in FPGA



Encoding/Decoding Engine

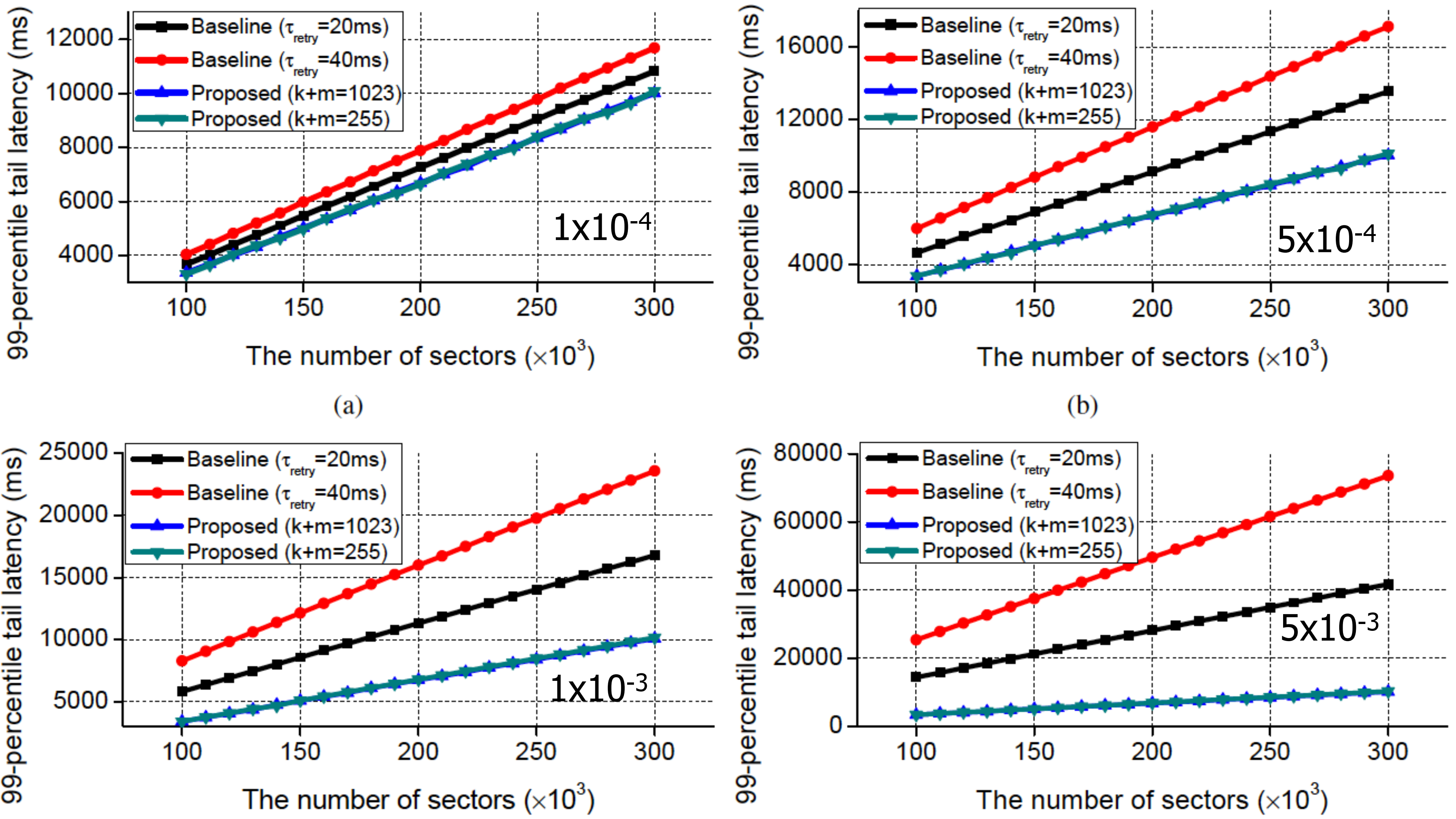
Hardware-based Implementation

- Parallel polynomial-based encoder and parallel Berlekamp-Massey decoder
- Verilog-based HDL design entry with target throughput of 4GB/s

Coding Parameters		**Equivalent XOR Gate Number**	
k	m	Encoder	Decoder
252	3	11k	156k
251	4	11k	161k
250	5	17k	185k
246	9	28k	232k
1019	4	16k	634k
1016	7	31k	699k
1014	9	39k	732k
1004	19	78k	894k

Tail Latency Analysis

- 7200rpm, P_{tail} = 99%, average per-sector retry latency: 20ms and 40ms



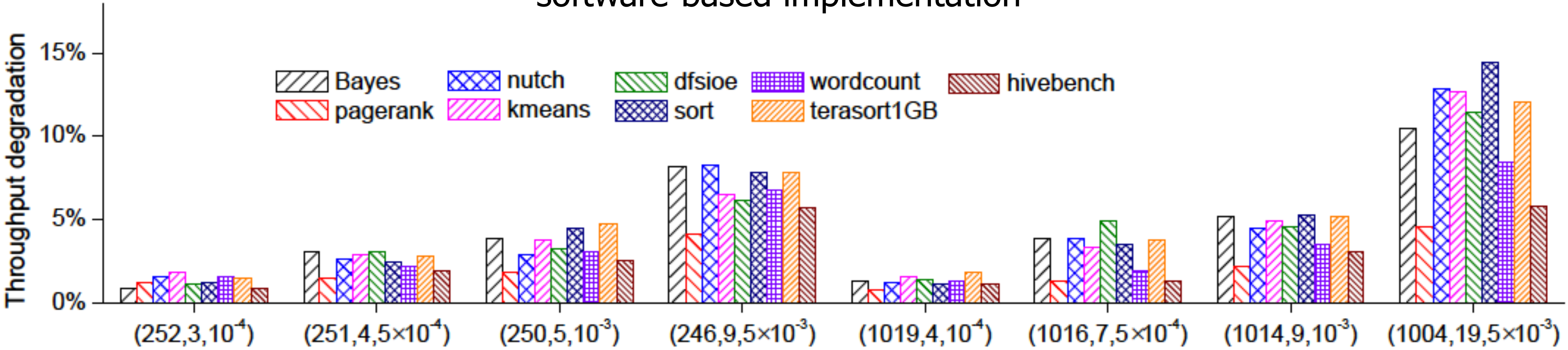
(a) (b)

Impact on System Speed Performance

- Integrate the local erasure coding into Kernel 3.10.102 (ext4 filesystem)
- Big data benchmark suite HiBench 3.0
  1. Job based micro benchmarks *sort* and *wordcount*
  2. SQL benchmark *hivebench*
  3. Web search/indexing benchmarks *pagerank* and *nutch*
  4. Machine learning benchmarks *bayes* and *kmeans*
  5. HDFS benchmark *dfsioe*
  6. Big data sorting benchmark *terasort*

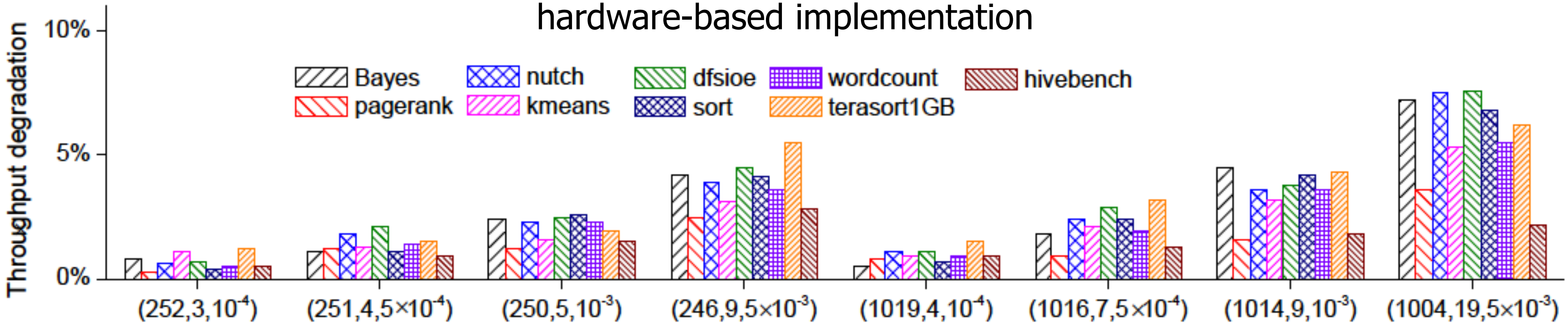
Impact on System Speed Performance

software-based implementation



(a)

hardware-based implementation



(b)

Conclusion and Future Work

- ✓ A first step exploring datacenter HDDs: local erasure coding
- ? Minimize CPU workload for distributed & local erasure coding
- ? Cross-layer system/HDD design
  - ? Software-defined datacenter HDD with configurable read channel
  - ? Iterative read channel and system-level erasure code decoding
  - ? Use of >4kB HDD sector size
- ? Modeling of soft and hard sector failures in future HDDs, and development of corresponding system-level coding design techniques
- ? Implication to overall HDD design (read channel, servo, head, …)
- ? …