Differential RAID: Rethinking RAID for SSD Reliability

Mahesh Balakrishnan

Asim Kadav[1], Vijayan Prabhakaran, Dahlia Malkhi

Microsoft Research Silicon Valley

[1]The University of Wisconsin-Madison

Solid State Storage

- Flash storage is now mainstream

- Commodity SSDs
  – Thousands of IOPS
  – Low power consumption

- How do we protect data on SSDs?
  – Device-Level redundancy: RAID levels

Primer on SSDs

- Smallest unit of read/write is a **Page** (e.g., 4KB)
- Pages must be *erased* before they are overwritten

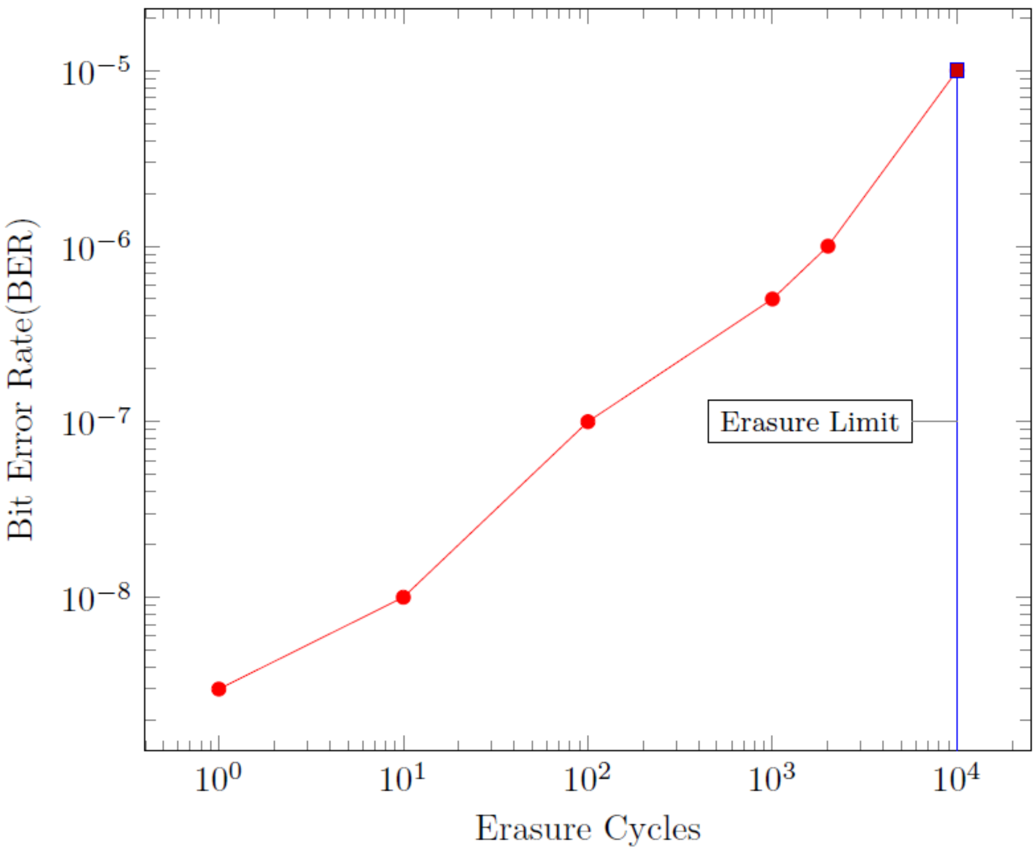
More writes → More erasures

- MTTF, Bit Error Rate (BER)

More erasures → Higher BER

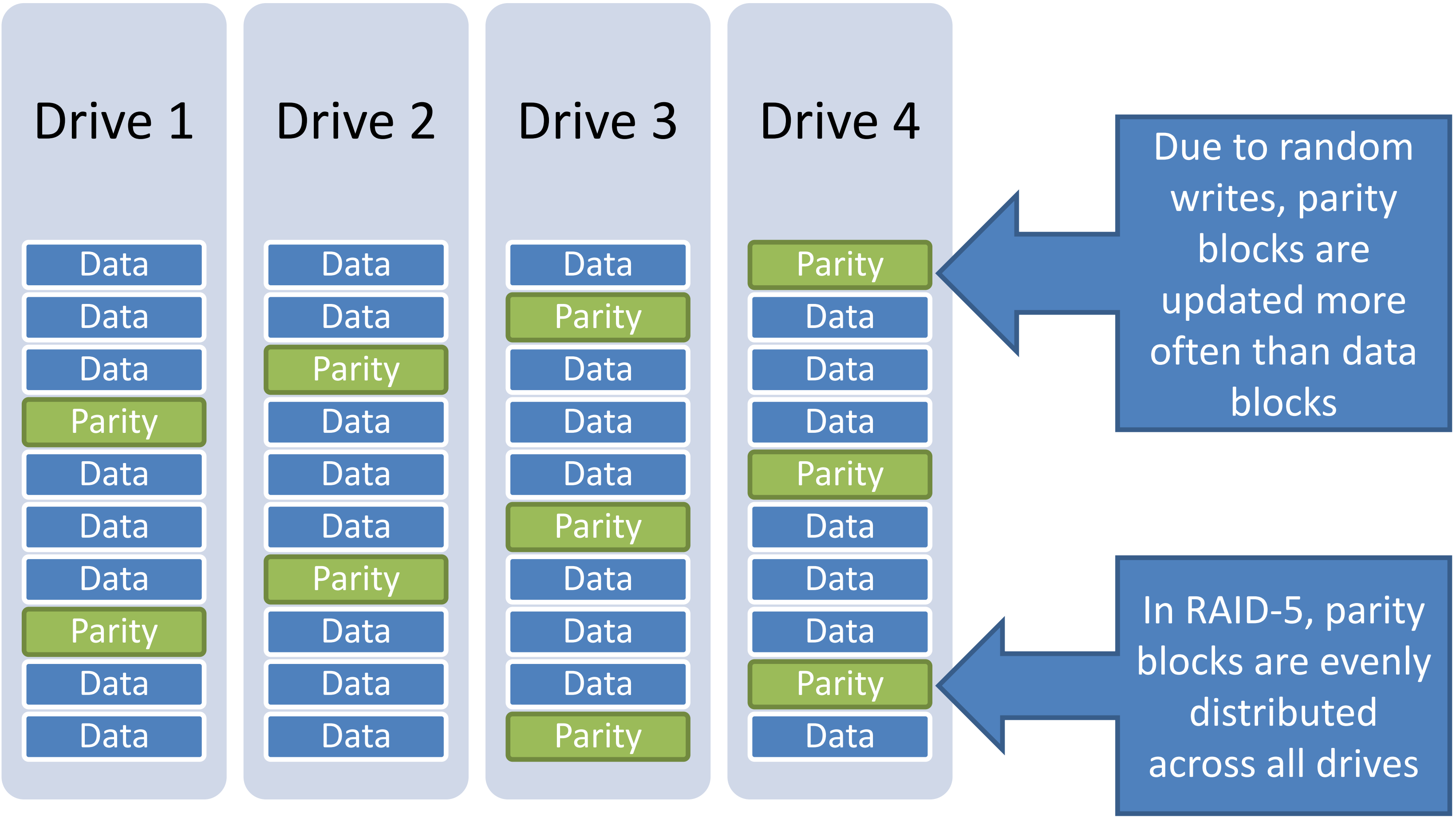
More writes → Higher BER

The BER Curve



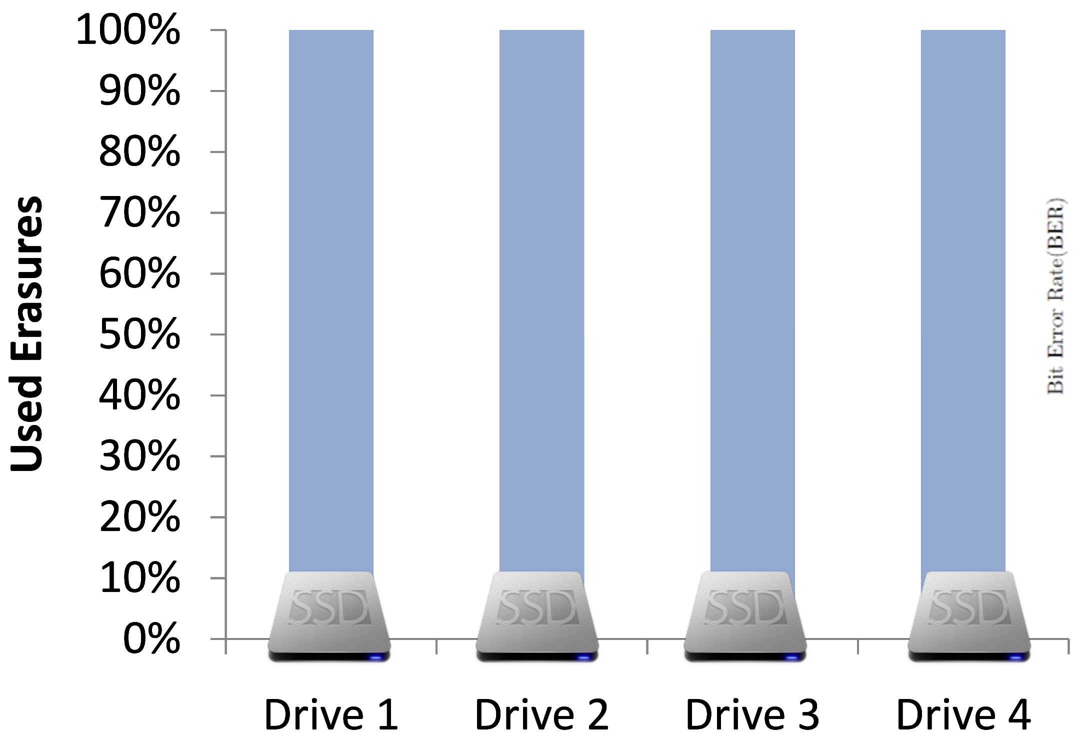
Can we use RAID to protect data on SSDs?

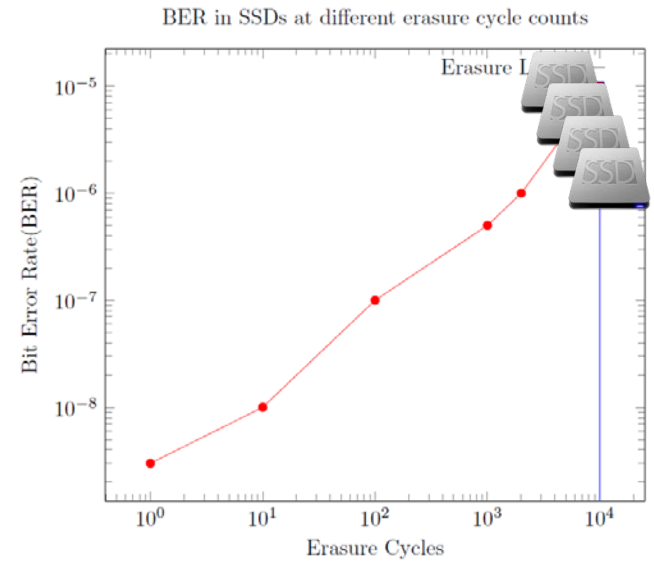
RAID-5 and SSDs



In RAID-5, all drives age at the same rate

Correlated Failures in RAID-5



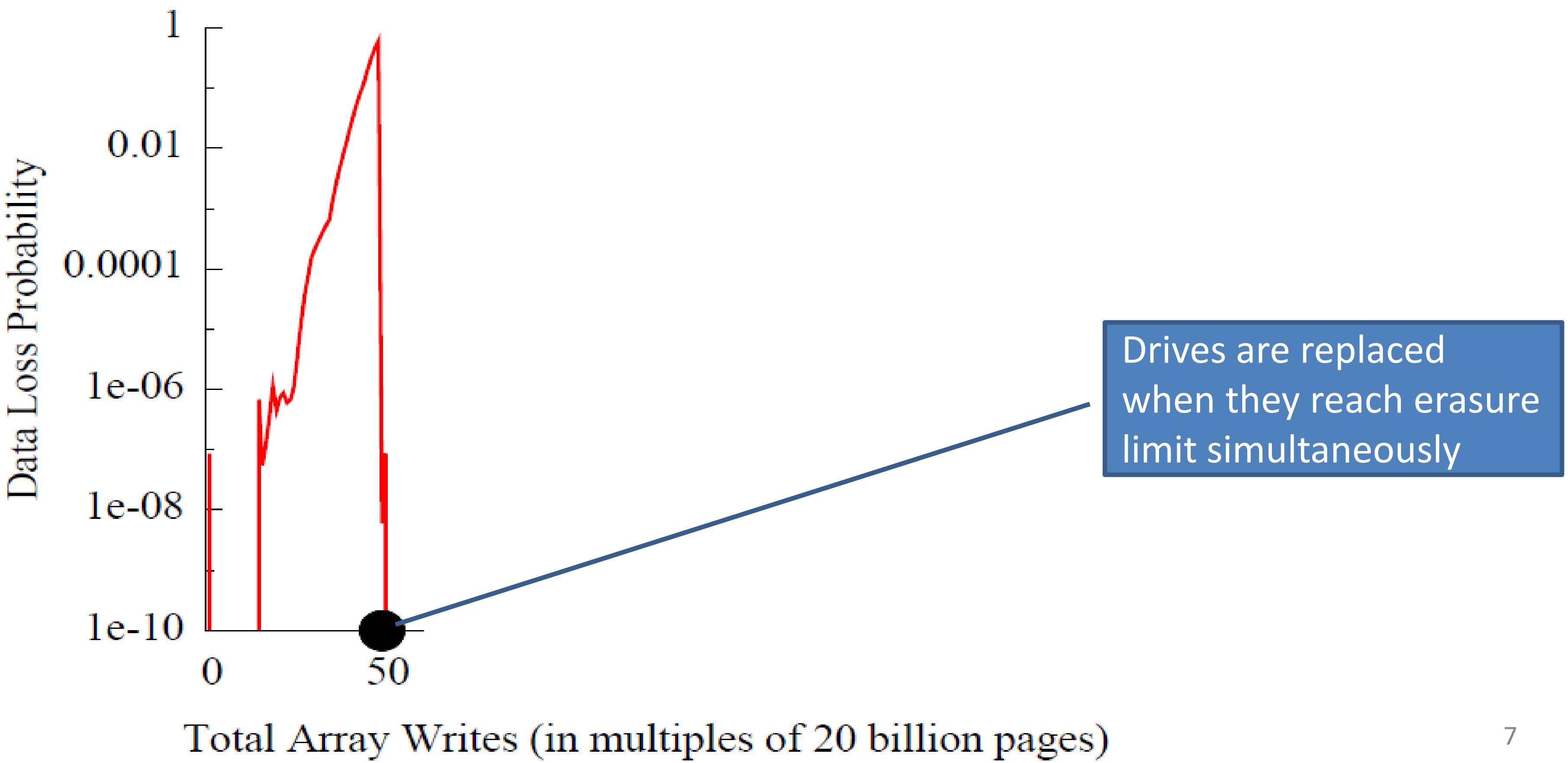


Parity Distribution:

Conventional RAID can induce correlated failures with SSDs

RAID-5 Reliability

Data Loss Probability: *When a drive fails, the probability that it cannot be completely reconstructed.*



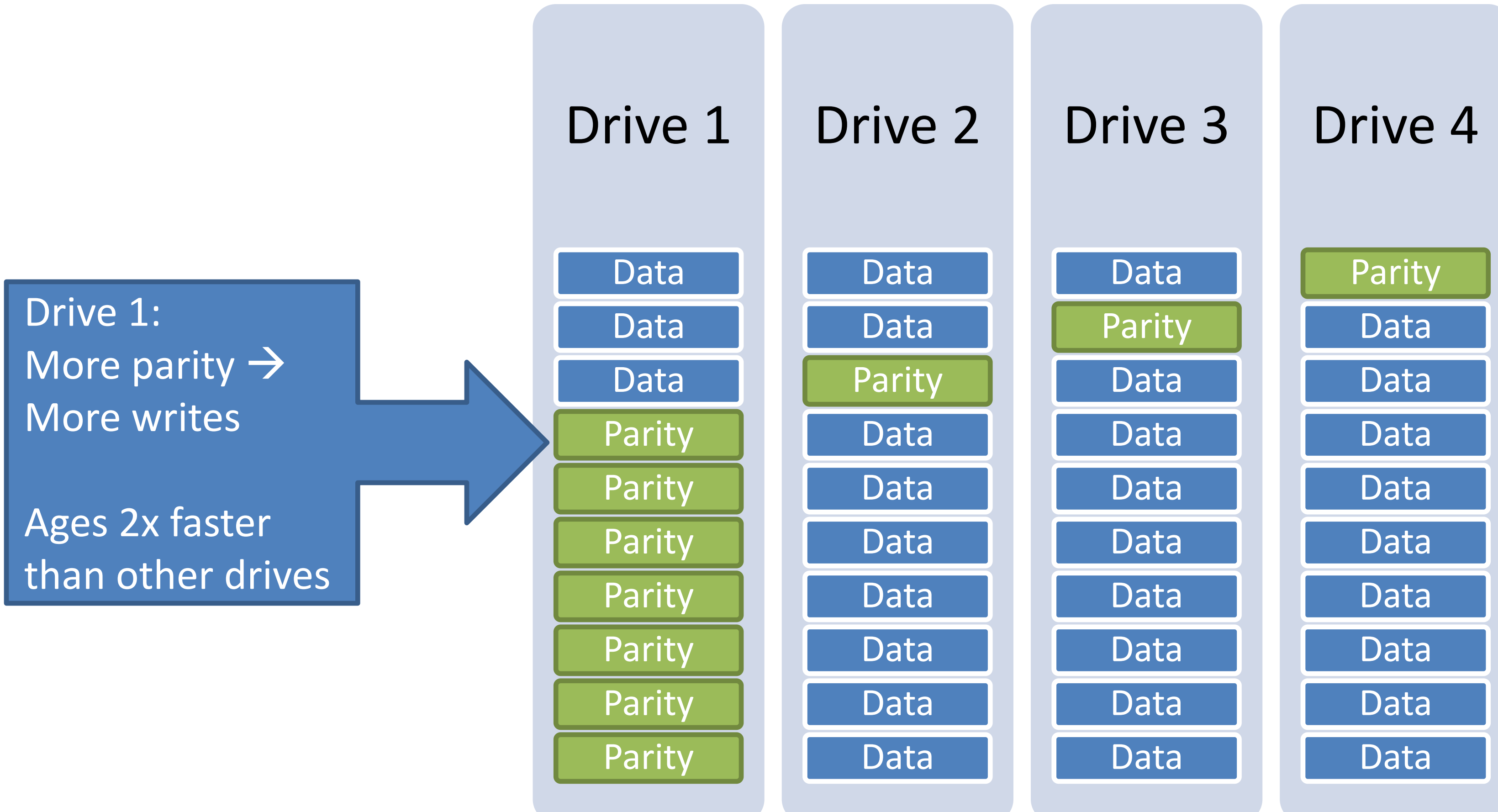
Solution: Differential RAID

- Goal: Age SSDs at different rates

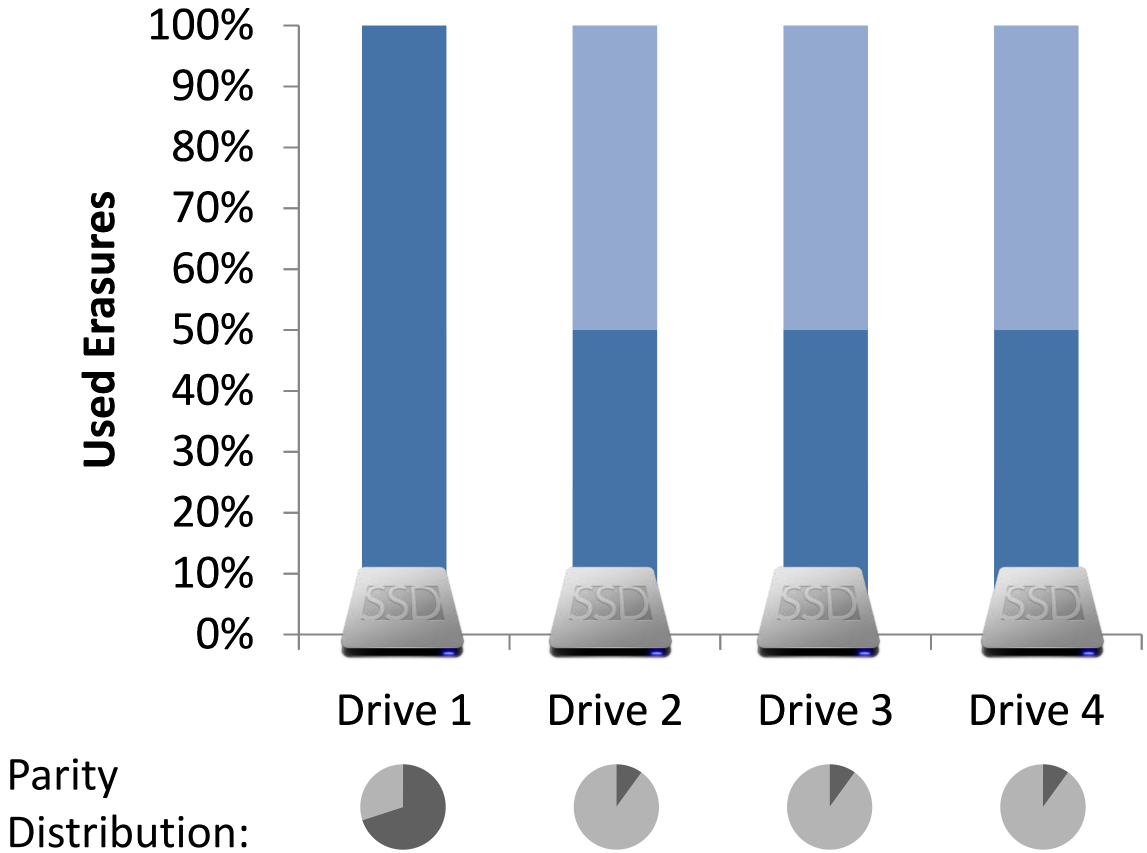
- Uneven Parity Assignment
    - Example: (70, 10, 10, 10): 70% of parity on Device 1
    - Any possible configuration between RAID-4 and RAID-5

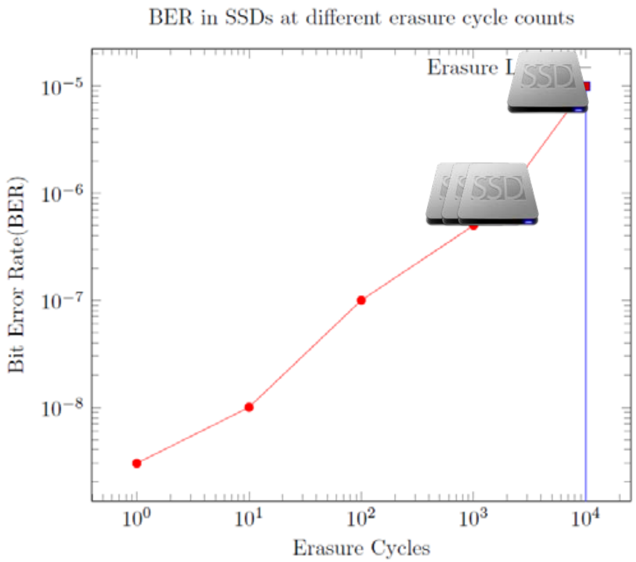
- Drive Replacement

Uneven Parity Distribution



Uneven Parity Distribution





Aging SSDs at different rates helps

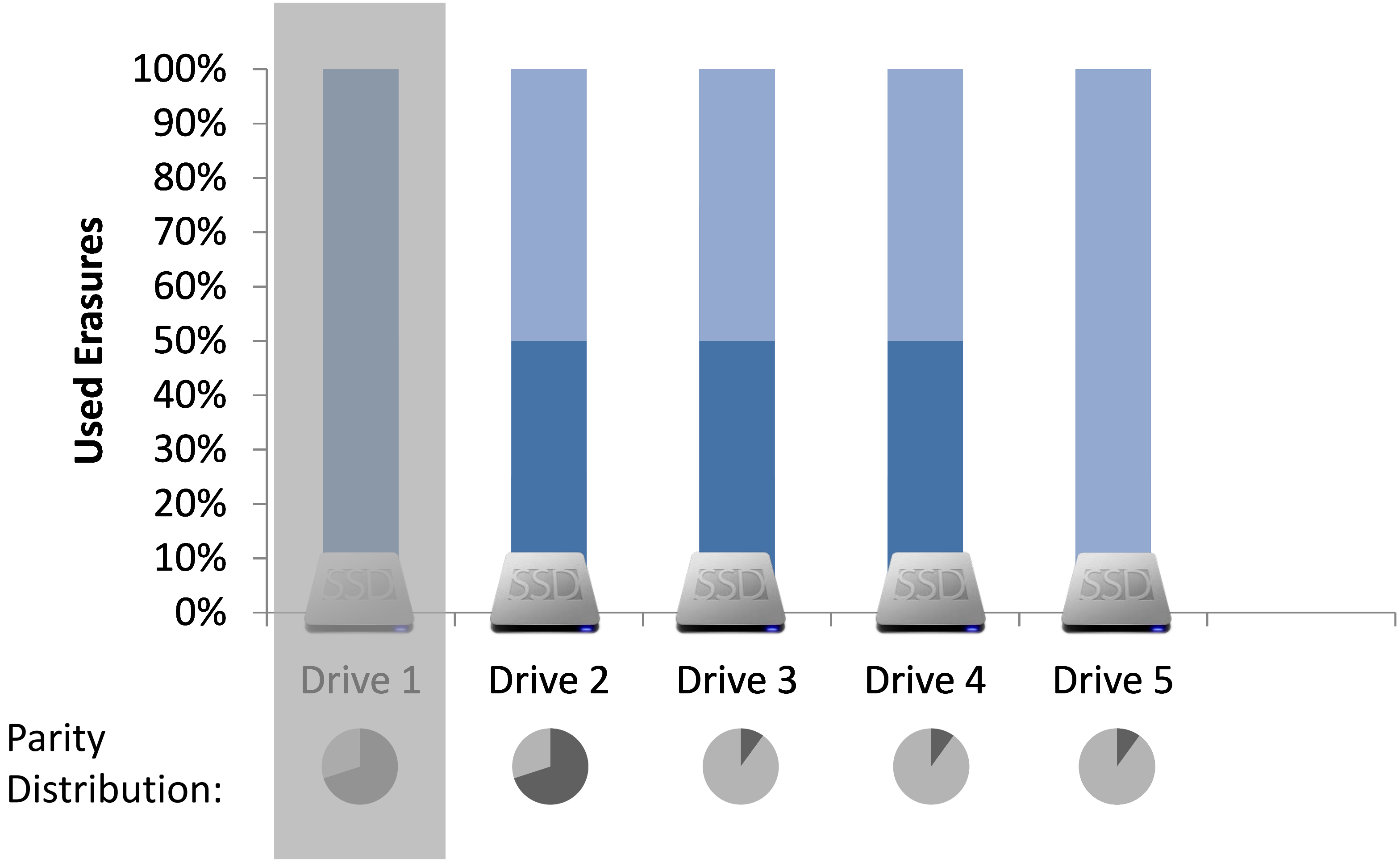
Solution: Differential RAID

- Goal: Age SSDs at different rates

- Uneven Parity Assignment
  - Example: (70, 10, 10, 10): 70% of parity on Device 1
  - Any possible configuration between RAID-4 and RAID-5

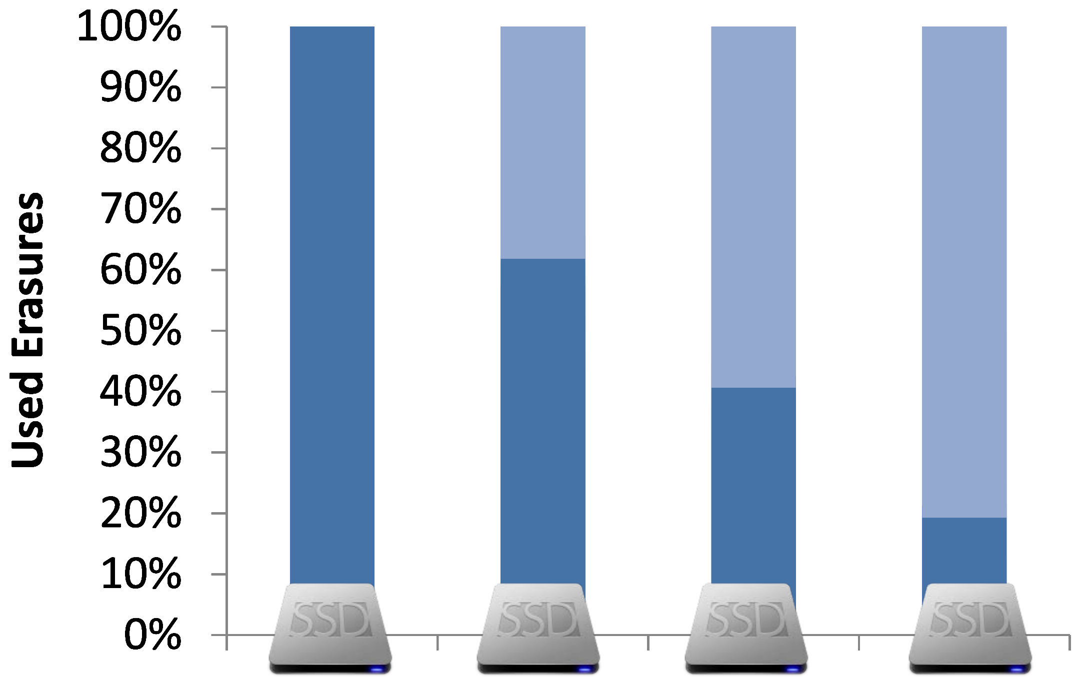
- Drive Replacement

Drive Replacement



Naïve Drive Replacement can still lead to correlated failures!

Convergence



Parity Distribution:

Age distribution provably converges for any parity assignment!

Evaluation

- Simulation for reliability
  - Real BER data for 12 chips from two studies[1]
  - 5-10K erase cycles
  - Assumed 4-bit ECC per 512-byte sector
  - Metric: Data Loss Probability (DLP)
- Implementation for performance

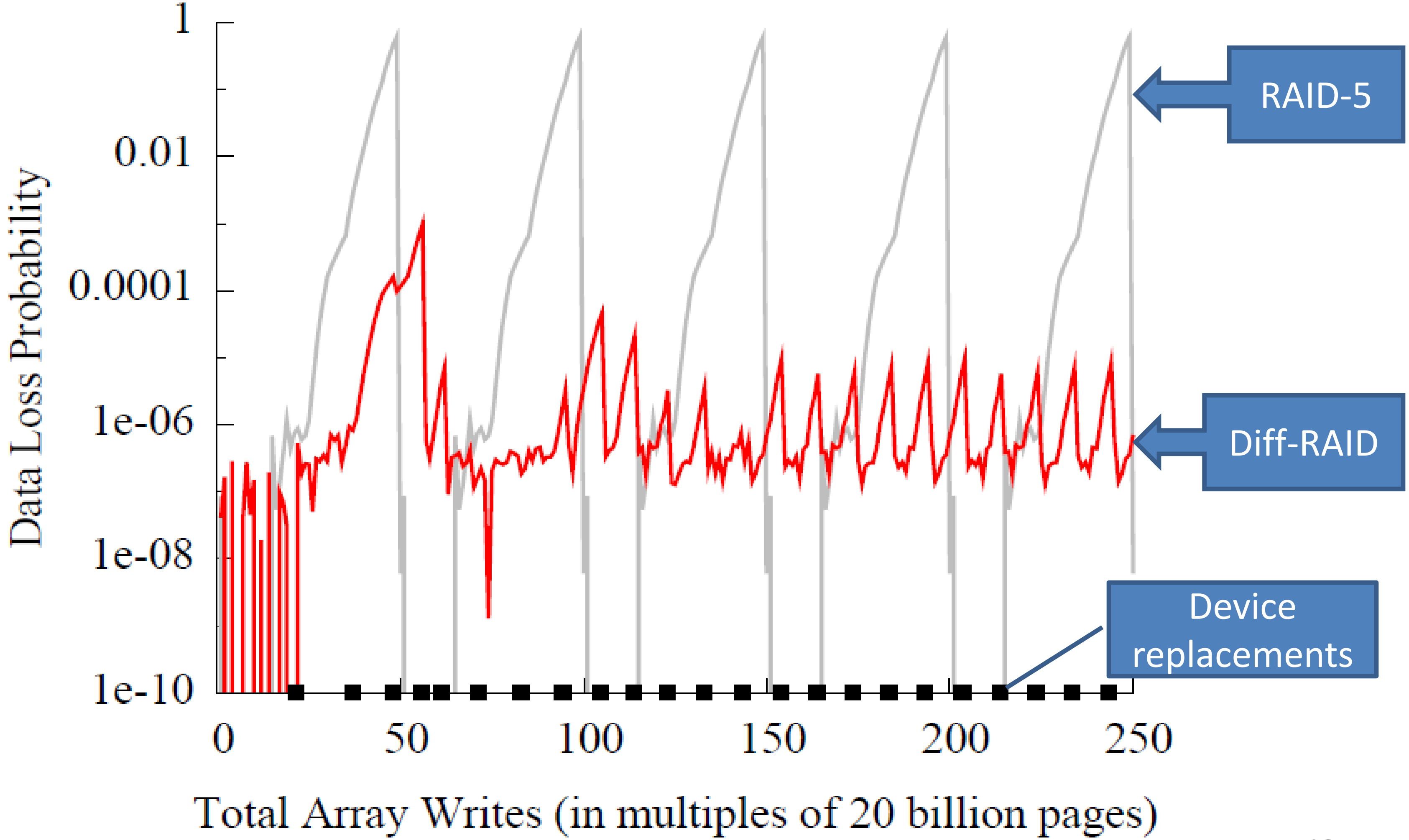
[1] N.Mielke et al., *Bit Error Rate in NAND Flash Memories.* International Reliability Physics Symposium, 2008.
L.M.Grupp et al., *Characterizing Flash Memory: Anomalies, Observations, and Applications.* Micro 2009.

Evaluation

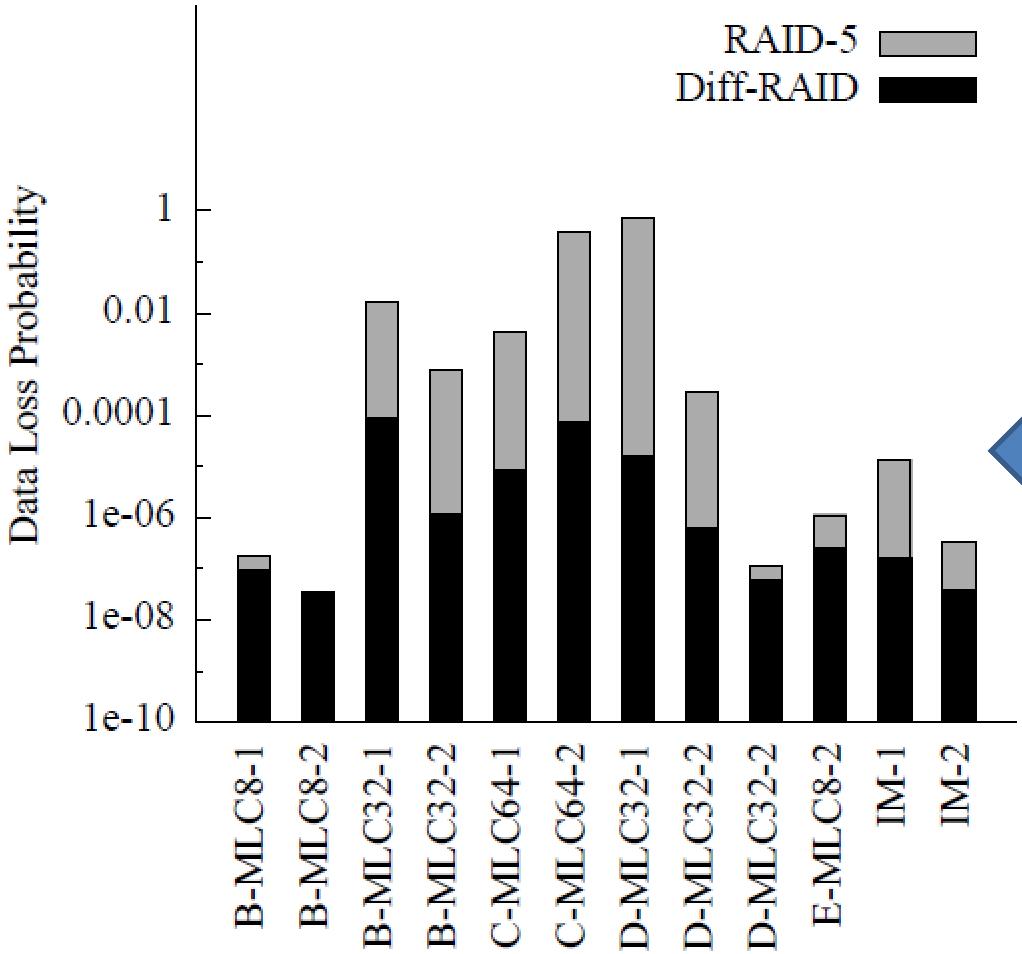
- Simulation for reliability
- Implementation for performance
  - 5 x Intel X25-M MLC SSDs
    - 80 GB each
    - Random write: 3.3K IOPS, Random read: 35K IOPS
    - Sequential write: 70 MB/s, Sequential read: 250 MB/s

[1] N.Mielke et al., *Bit Error Rate in NAND Flash Memories.* International Reliability Physics Symposium, 2008.
L.M.Grupp et al., *Characterizing Flash Memory: Anomalies, Observations, and Applications.* Micro 2009.

Diff-RAID Reliability

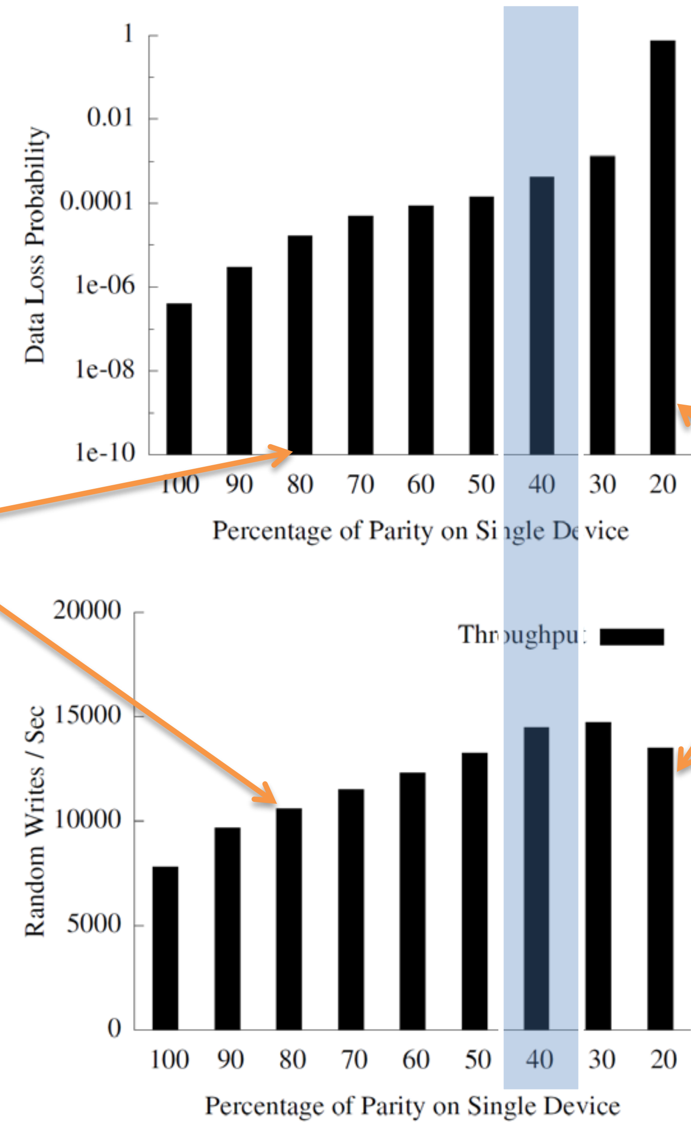


Diff-RAID Reliability



Diff-RAID is more reliable than RAID-5 by at least **10x for 7/12 chips**

Reliability vs Throughput



Reliability decreases as parity is less concentrated

(80,5,5,5,5)

(20,20,20,20,20)
RAID-5

Throughput increases as parity is less concentrated

Conclusion

- RAID can cause correlated failures with Flash
  – Not just RAID-5; not just SSDs

- Differential RAID:
  – Key Idea: Age SSDs at different rates
  – Same space overhead as RAID-5
  – Trade-off between reliability and throughput

Other Results

- Diff-RAID works on real workloads

- Diff-RAID can be used to extend SSD lifetime
  - Replace drives at 13K cycles → DLP < 0.001
  - Replace drives at 15K cycles → DLP < 0.01

- Diff-RAID can lower ECC requirements
  - Diff-RAID on 3-bit ECC == RAID-5 on 5-bit ECC

SSDs and RAID

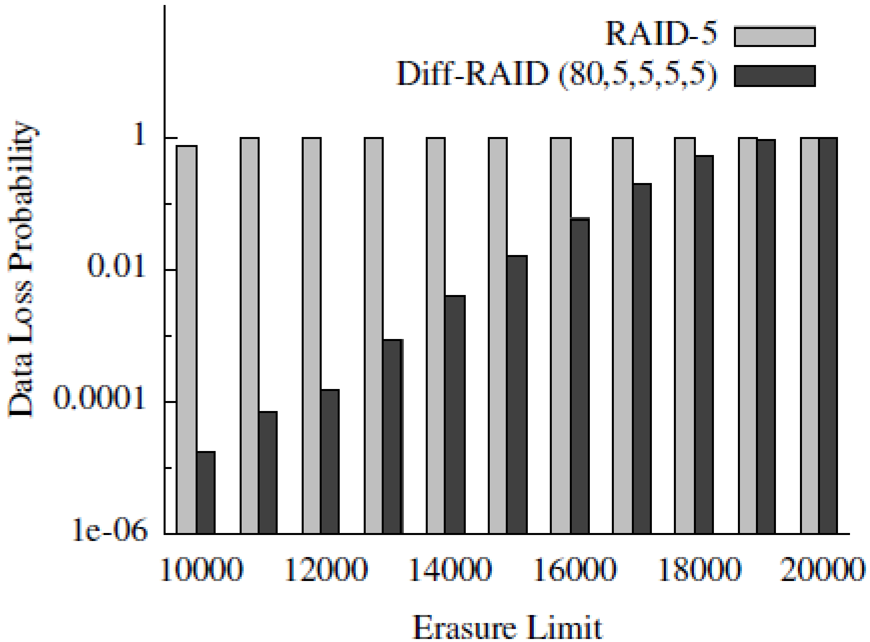
SSDs are *not* hard disks!

- RAID-5 is a bad idea for Hard Disks:
  - ✘ Performance: Slow random writes
  - ✘ Cost: Storage is cheap → Just use RAID-1/10
  - ✘ Reliability: High probability of data loss in large arrays
- RAID-5 is a great idea for SSDs!
  - ✓ Performance: Fast random writes (5 SSDs = 14K/sec)
  - ✓ Cost: Storage is expensive → Can't use RAID-1/10
  - ? Reliability: Correlated Failures!

  (Not just RAID-5: RAID-1, RAID-4, RAID-10, RAID-6…)

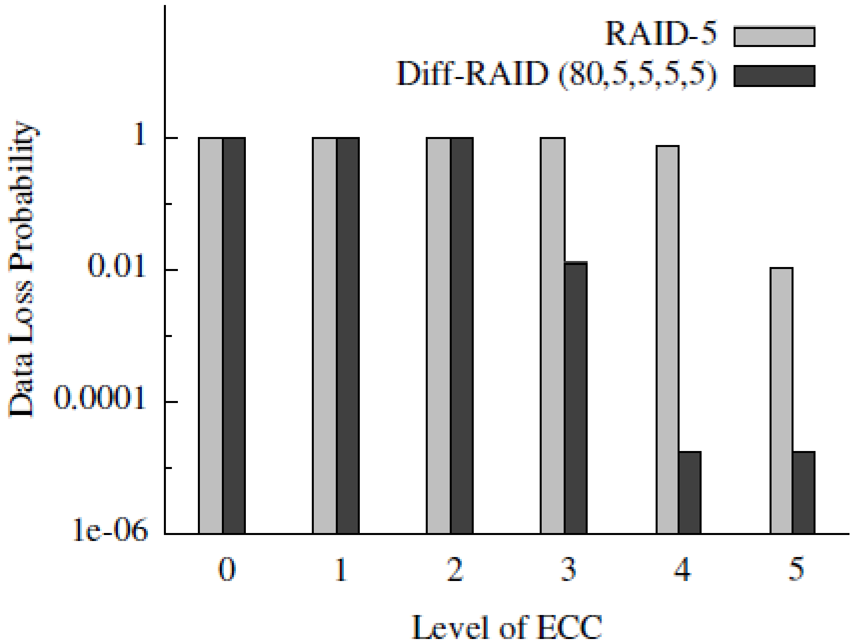
Other Benefits

Diff-RAID allows SSDs to be used past the erasure limit:



30% more lifetime with DLP 0.001
50% more lifetime with DLP 0.01

Diff-RAID allows SSDs to be used with less ECC:

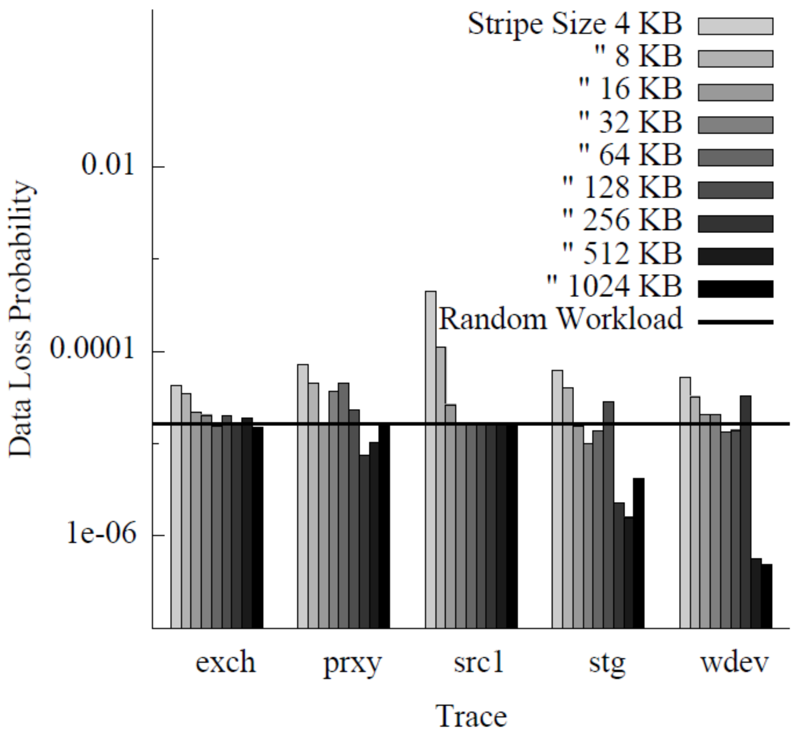


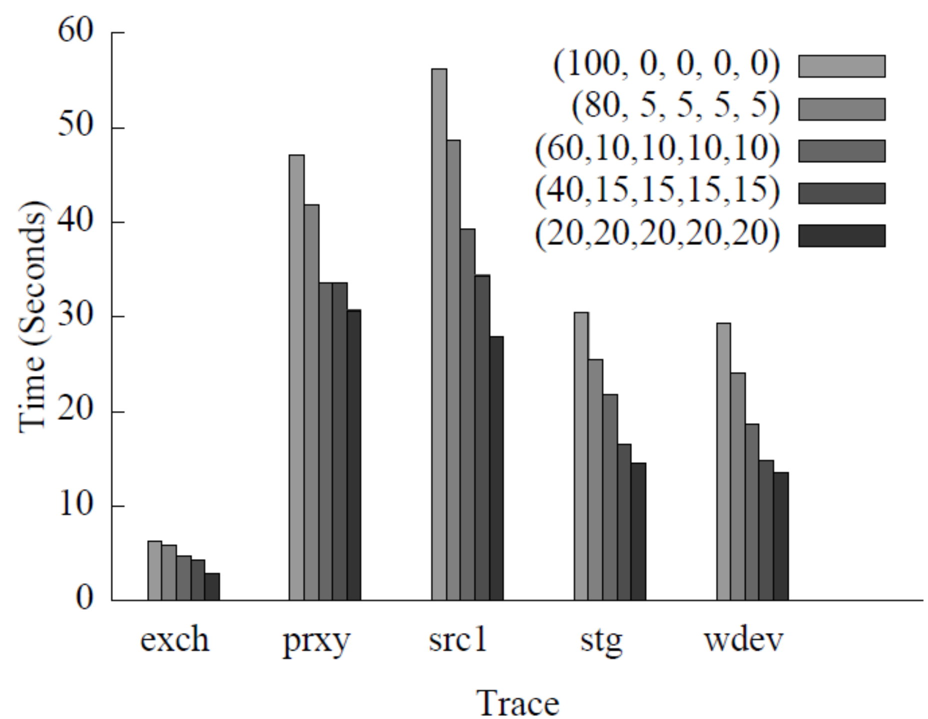
Diff-RAID with 3-bit ECC is equal to RAID-5 with 5-bit ECC

Real Workloads

RAID stripe size matters:
larger stripe size → more random writes

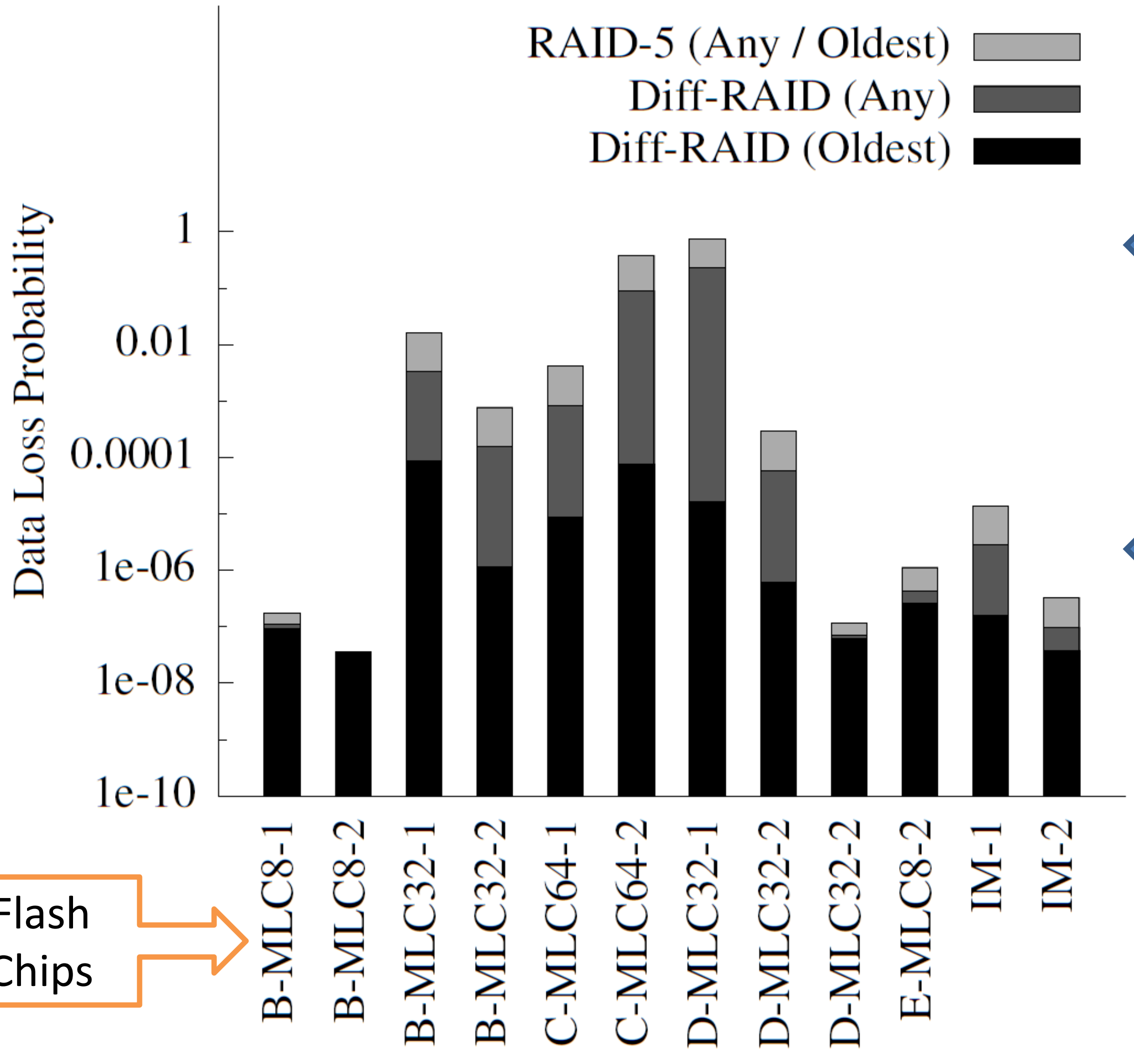
Reliability versus throughput
on real workloads





Larger stripe → more reliable

Reliability



Flash Chips

Random device fails: Diff-RAID is at least **2x** more reliable for 9/12 chips.

Oldest device fails: Diff-RAID is at least **10x** more reliable for 7/12 chips.