GRIFFIN

Extending SSD Lifetimes with Disk-Based Write Caches

Gokul Soundararajan
Vijayan Prabhakaran
Mahesh Balakrishnan
Ted Wobber

FAST 2010

Flash Drives

Intel X25-M

Fast Reads

Sequential read rate	**250 MB/s**
Random 4 KB reads	**35000**

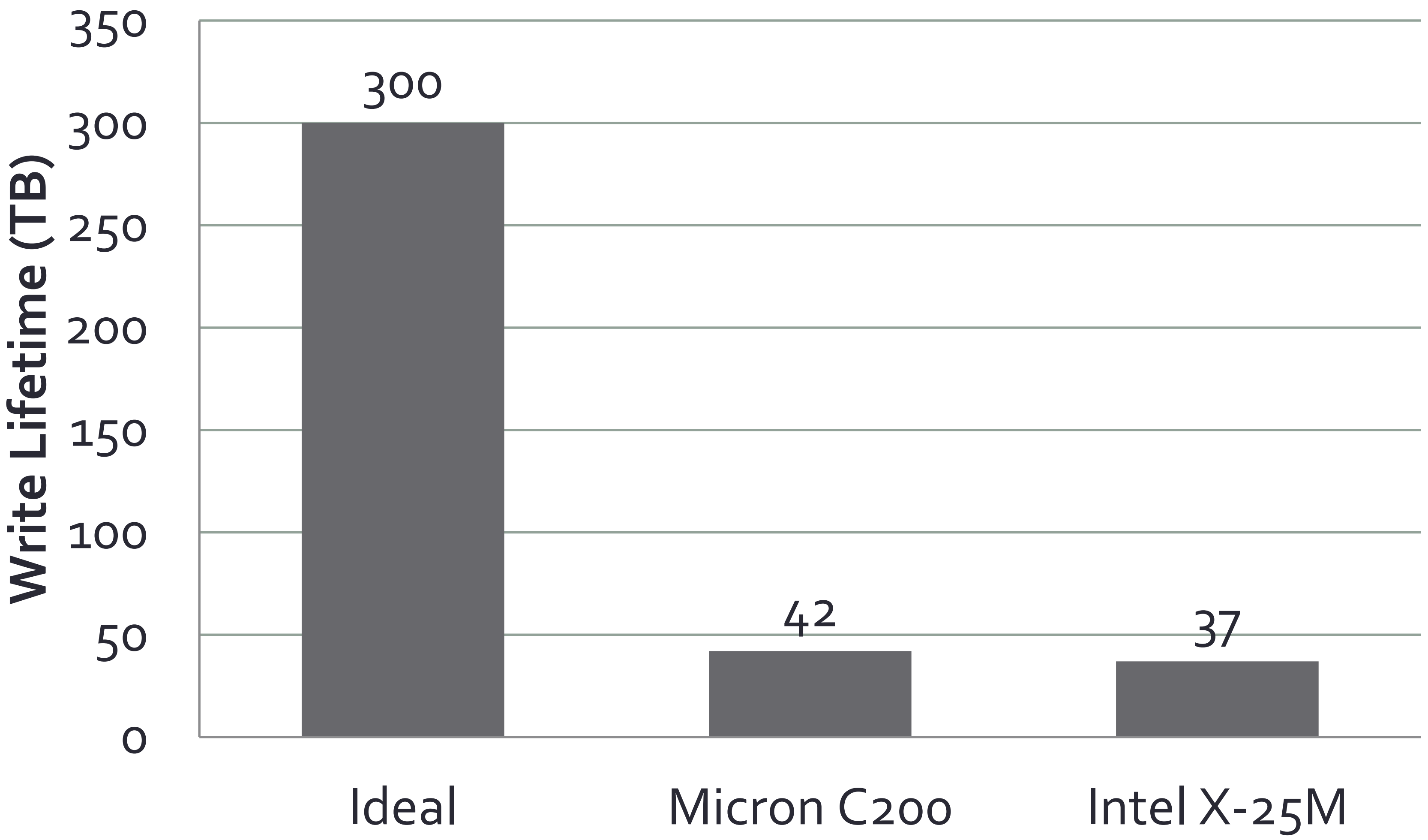
Slow Random Writes

Sequential write rate	**70 MB/s**
Random 4KB writes	**3300**

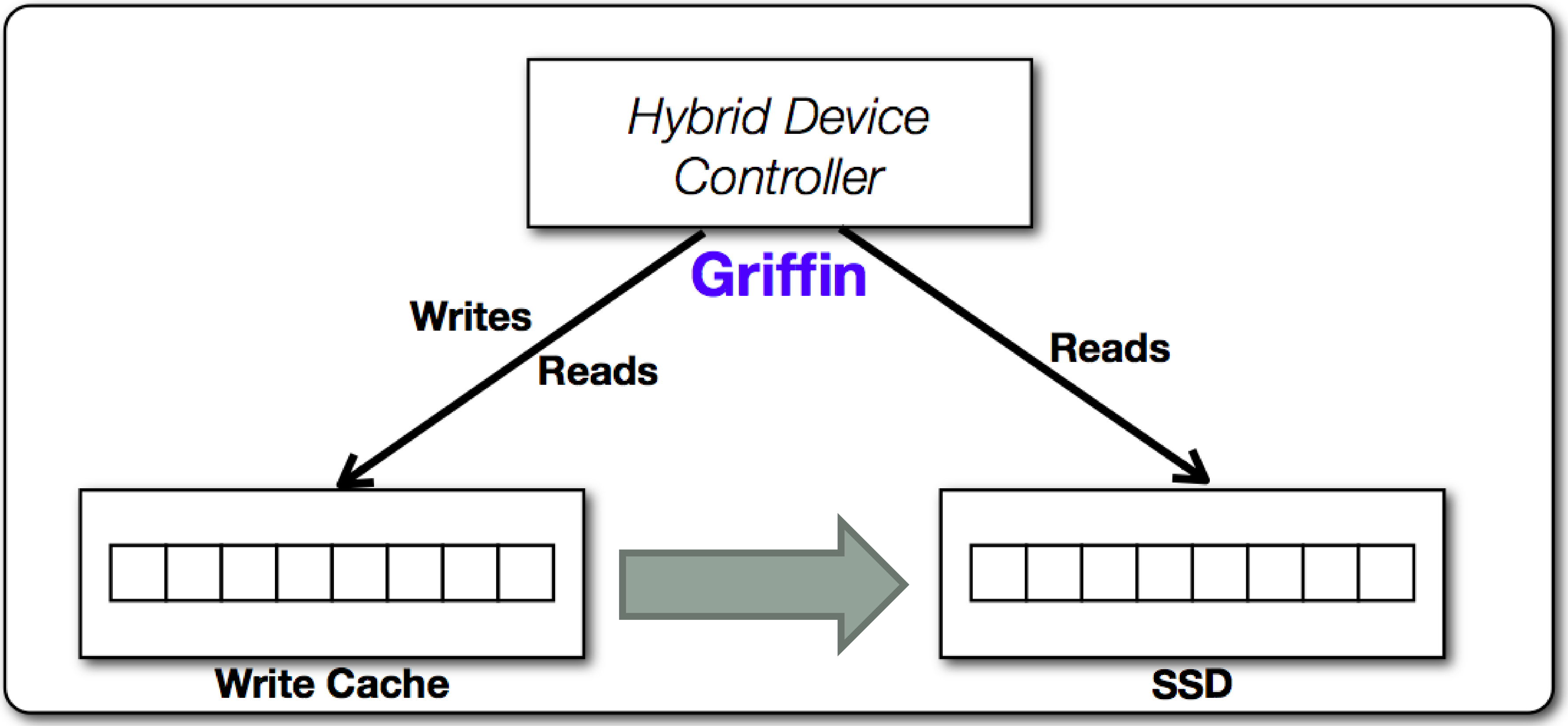
MLC vs SLC

	Single Level Cell	Multi Level Cell
Density	16Mbit	32Mbit / 64Mbit
Endurance	100,000 cycles	10,000 cycles
Cost (128 GB)	$1200	$300

SSD Write Lifetime

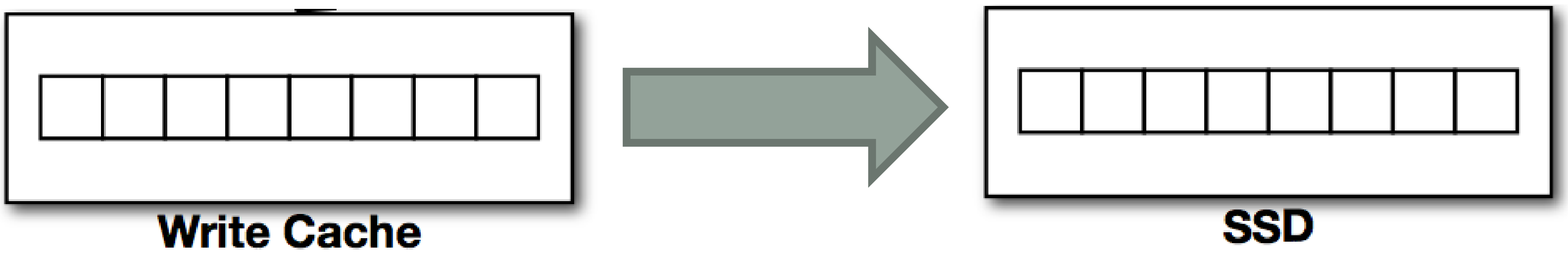


Write Cache



Benefit - Coalesce Overwrites to Increase Lifetime

Hybrid design Options

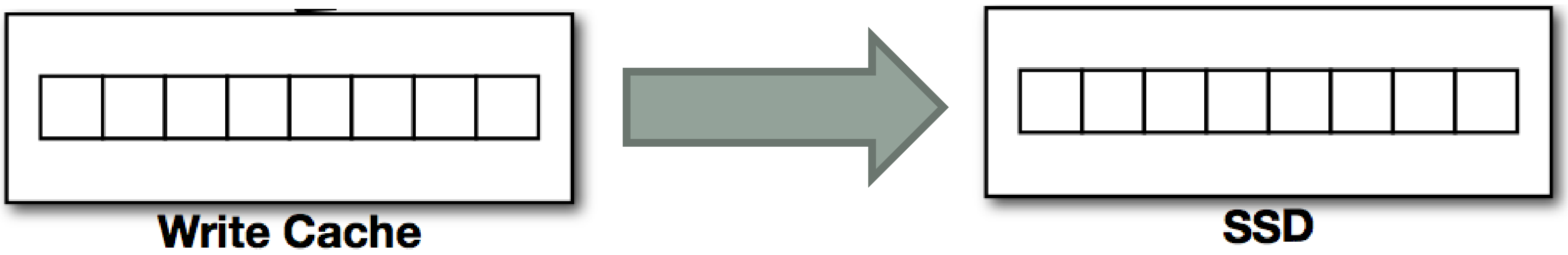


RAM
- Fast
- Not persistent

HDD
- Good Sequential Write Speed
- Cheap, Large Capacity

Flash
- Lower Power
- Higher Cost

Hybrid design Options

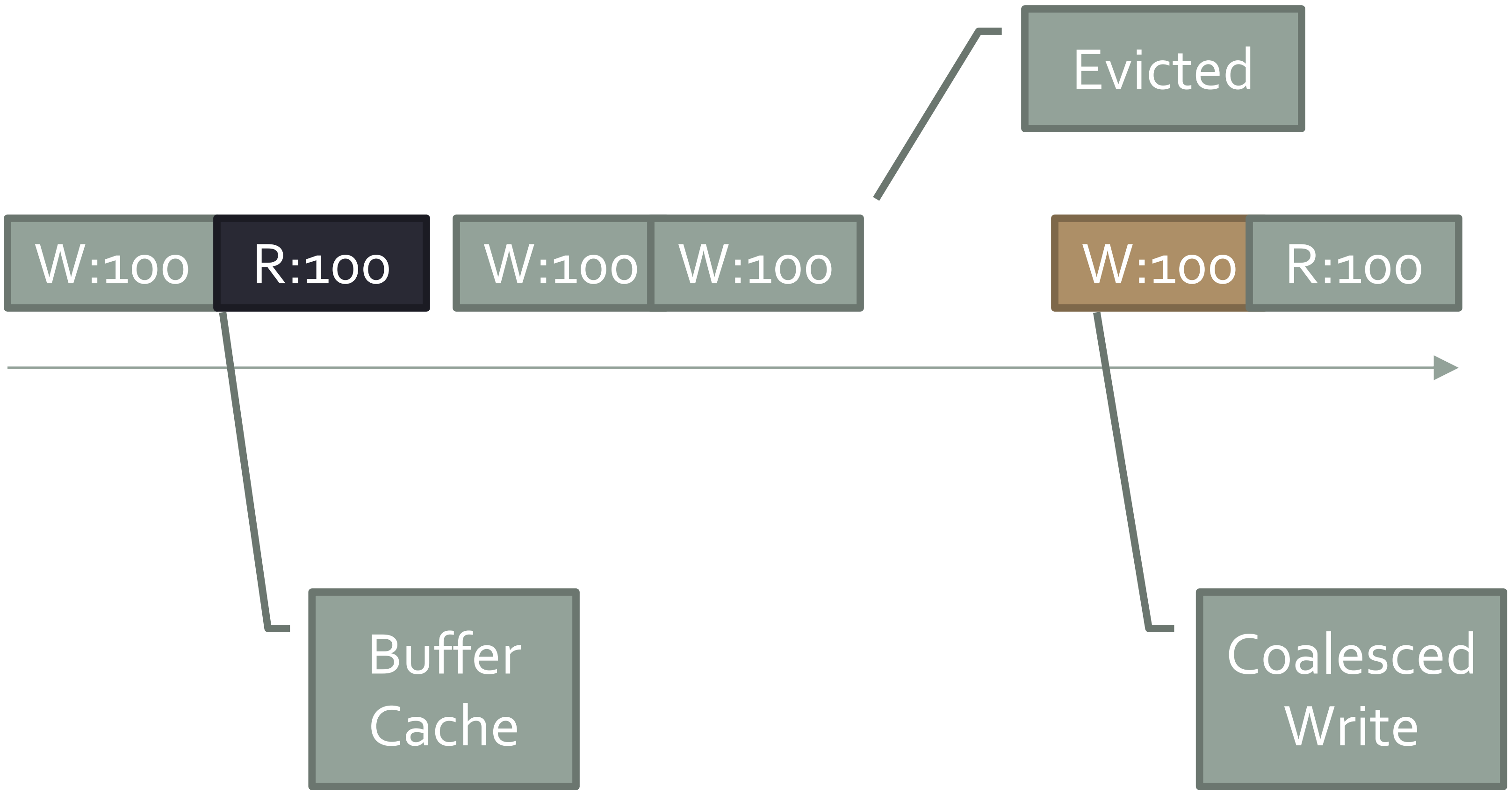


RAM
- Fast
- Not persistent

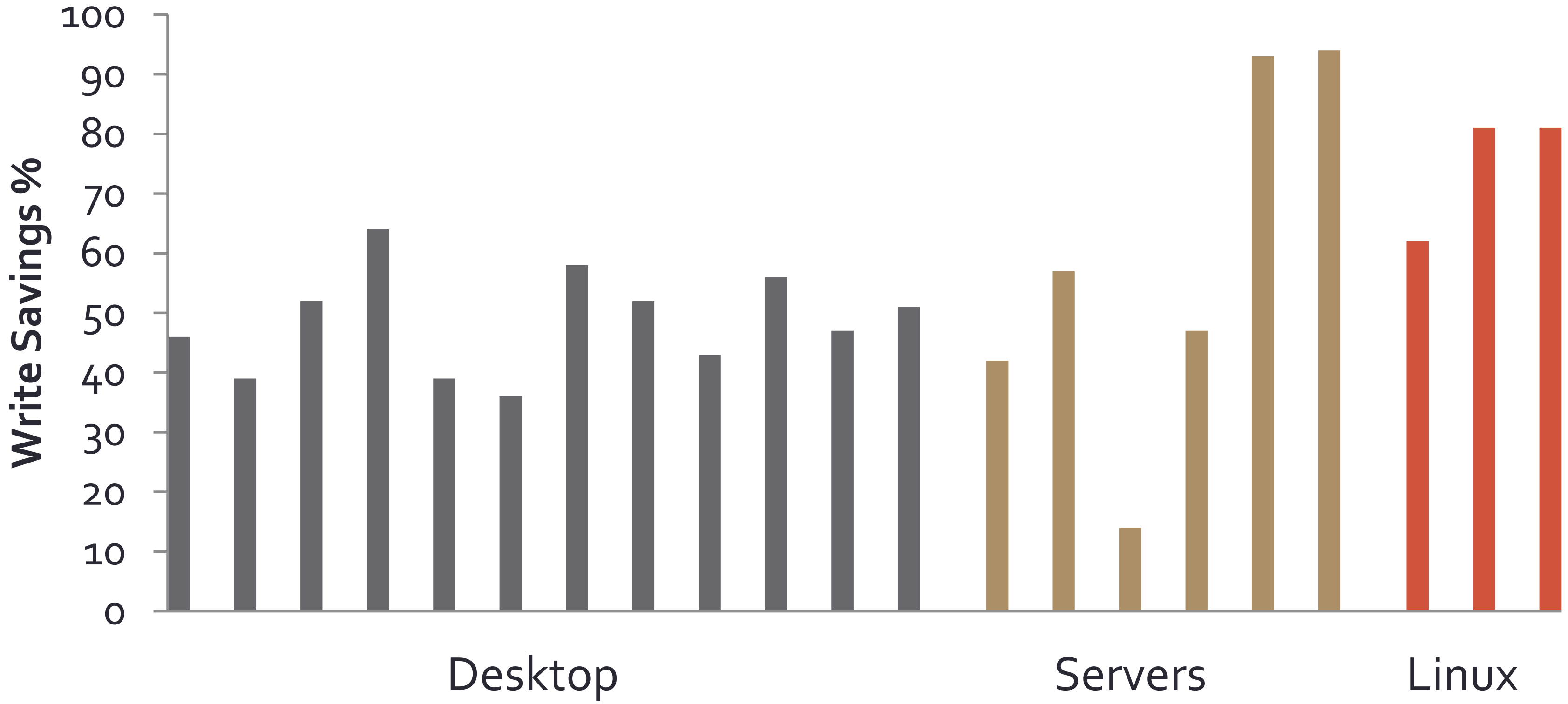
HDD
- Good Sequential Write Speed
- Cheap, Large Capacity

Flash
- Lower Power
- Higher Cost

Coalesced Writes



Ideal Savings



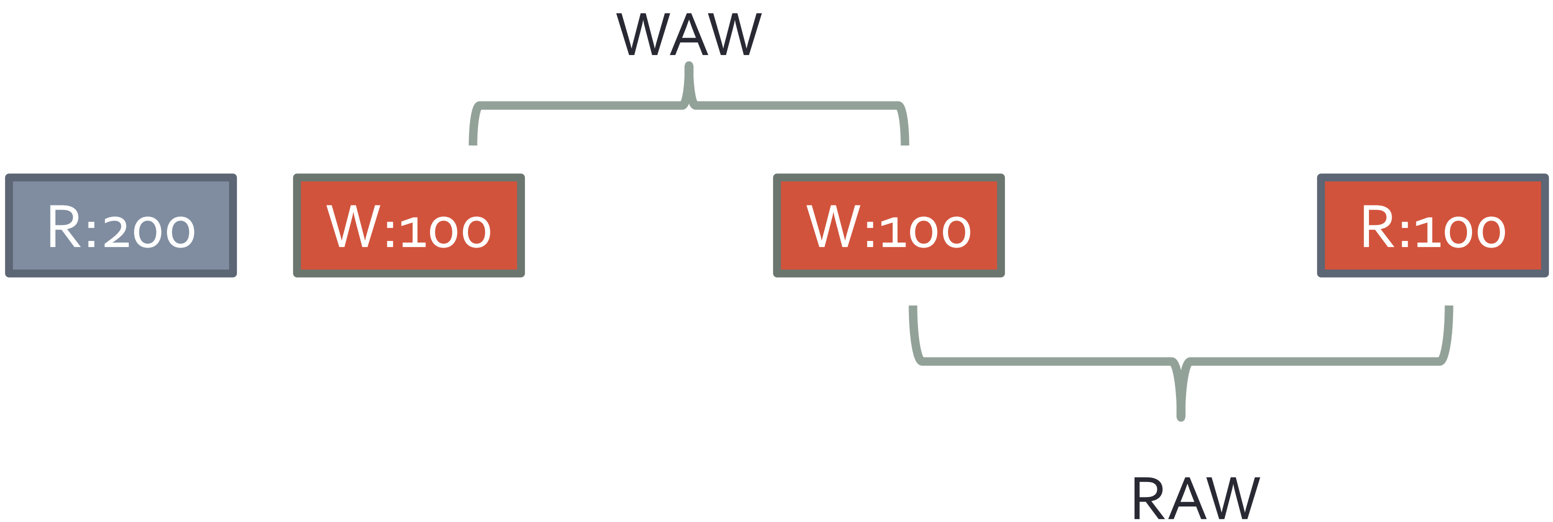
Potential Reduction:

36-64% - Desktops 94% - Web server 62% - Linux Desktop

WAW and RAW

- Write After Write (WAW)

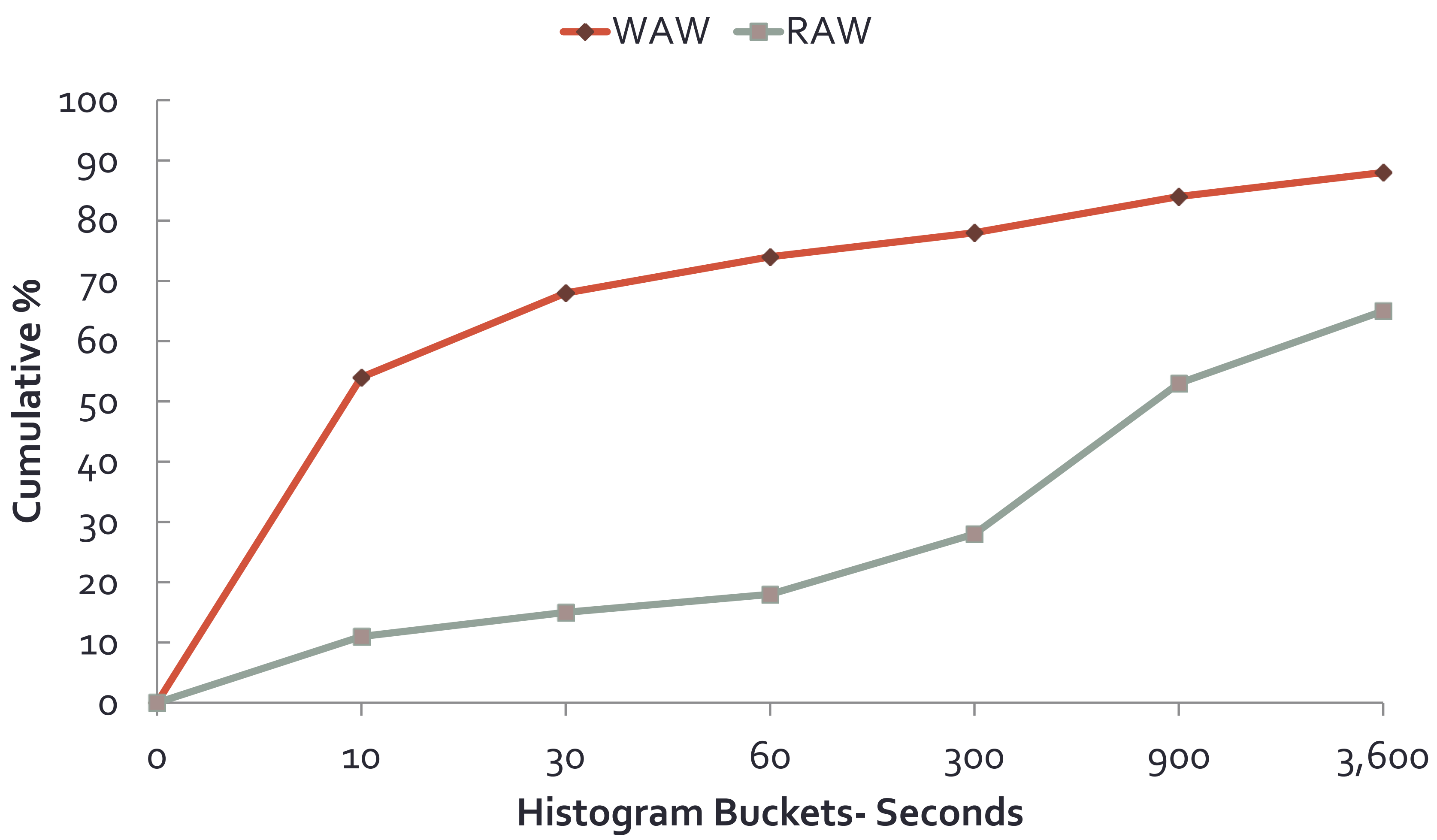
Time between two consecutive writes to the same block



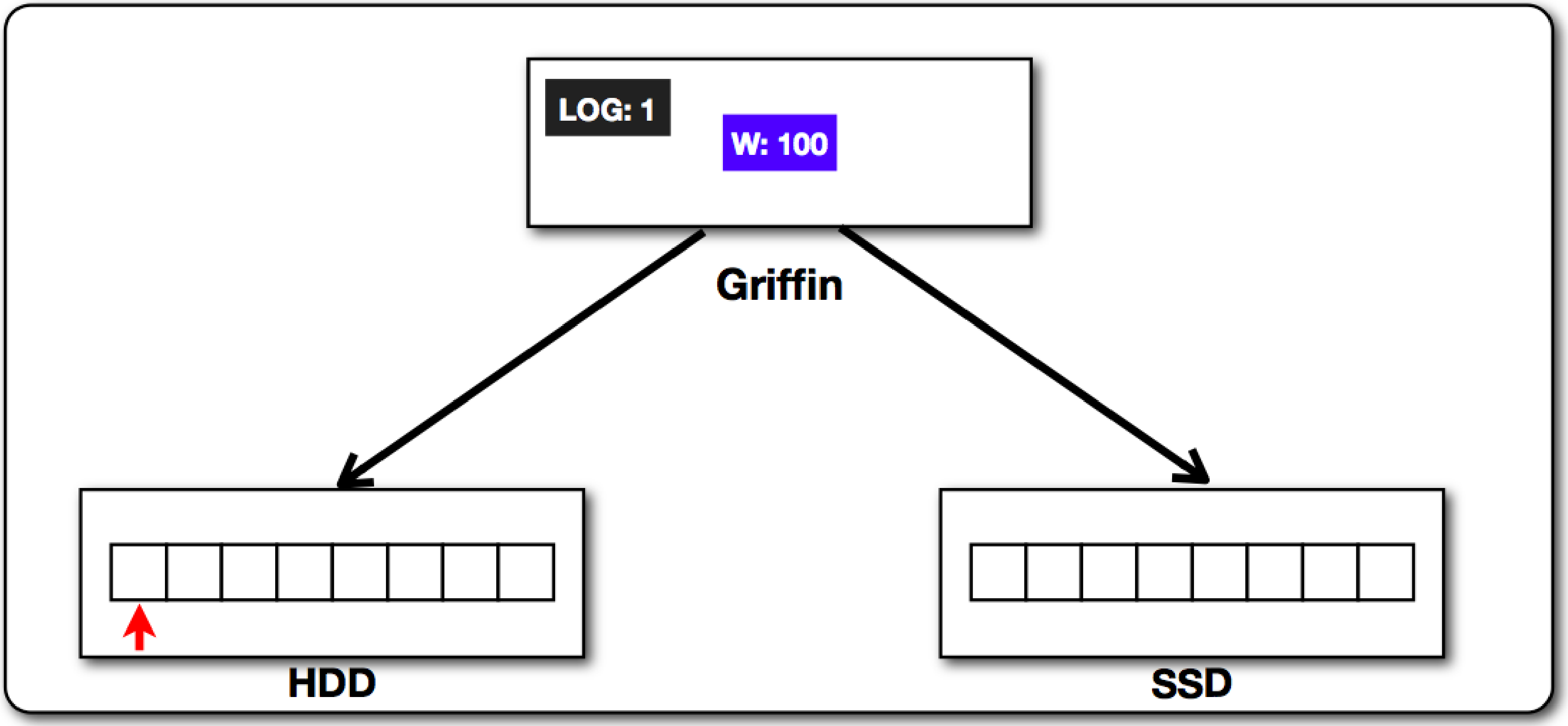
- Read After Write (RAW)

Time between a write and a subsequent read to the same block

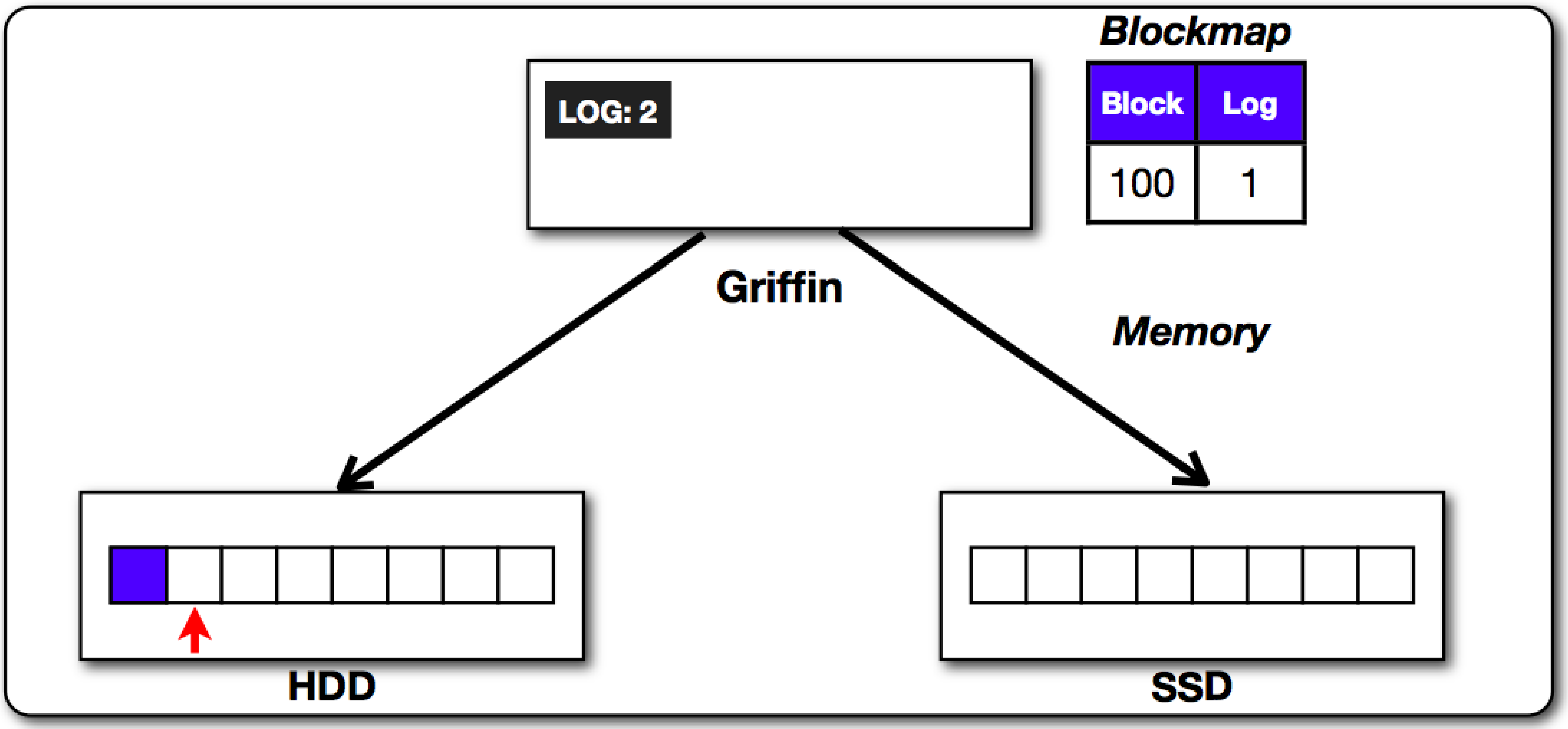
WAW/RAW Intervals



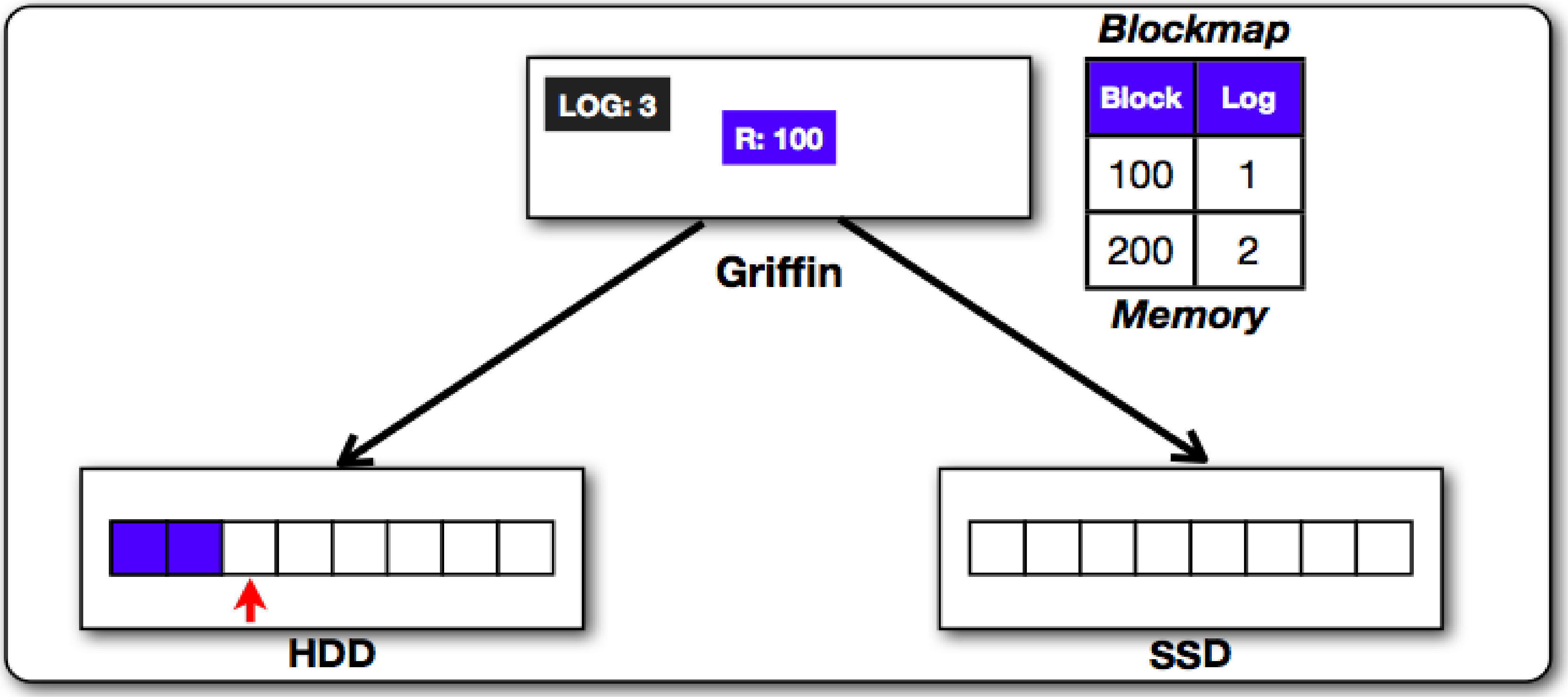
Griffin Write



Griffin Write



Griffin Read



What to cache and How long to Cache

- Metrics:
- Write Savings - Percentage of writes that don’t reach the SSD
- Read Penalty - Percentage of reads serviced by HDD
- Fault Tolerance - HDD failures

What to Cache – Selective Caching

Blockmap

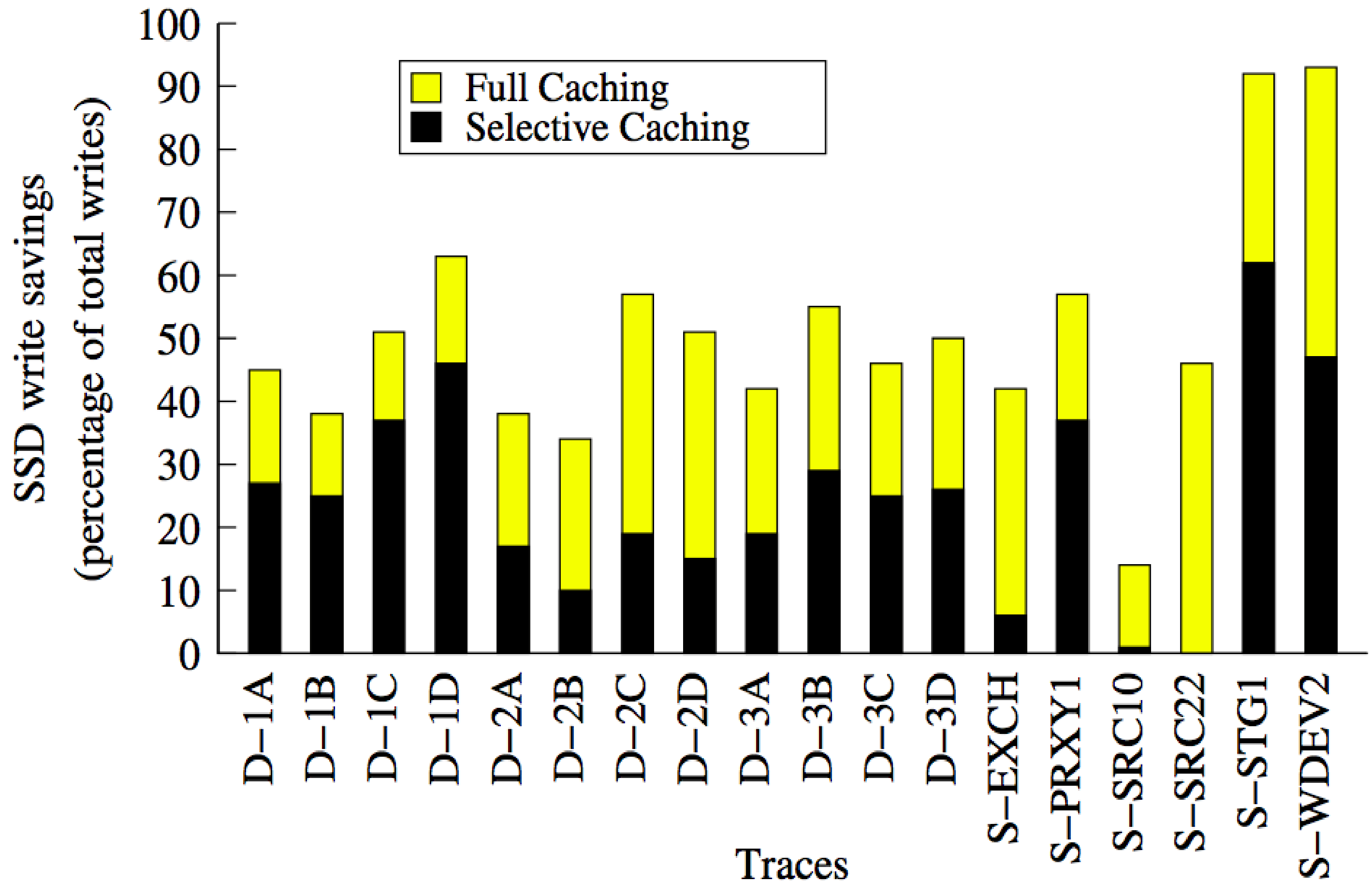
Block	Log
100	1

Blockmap

Block	Overwritten?	Log
100	✔	1

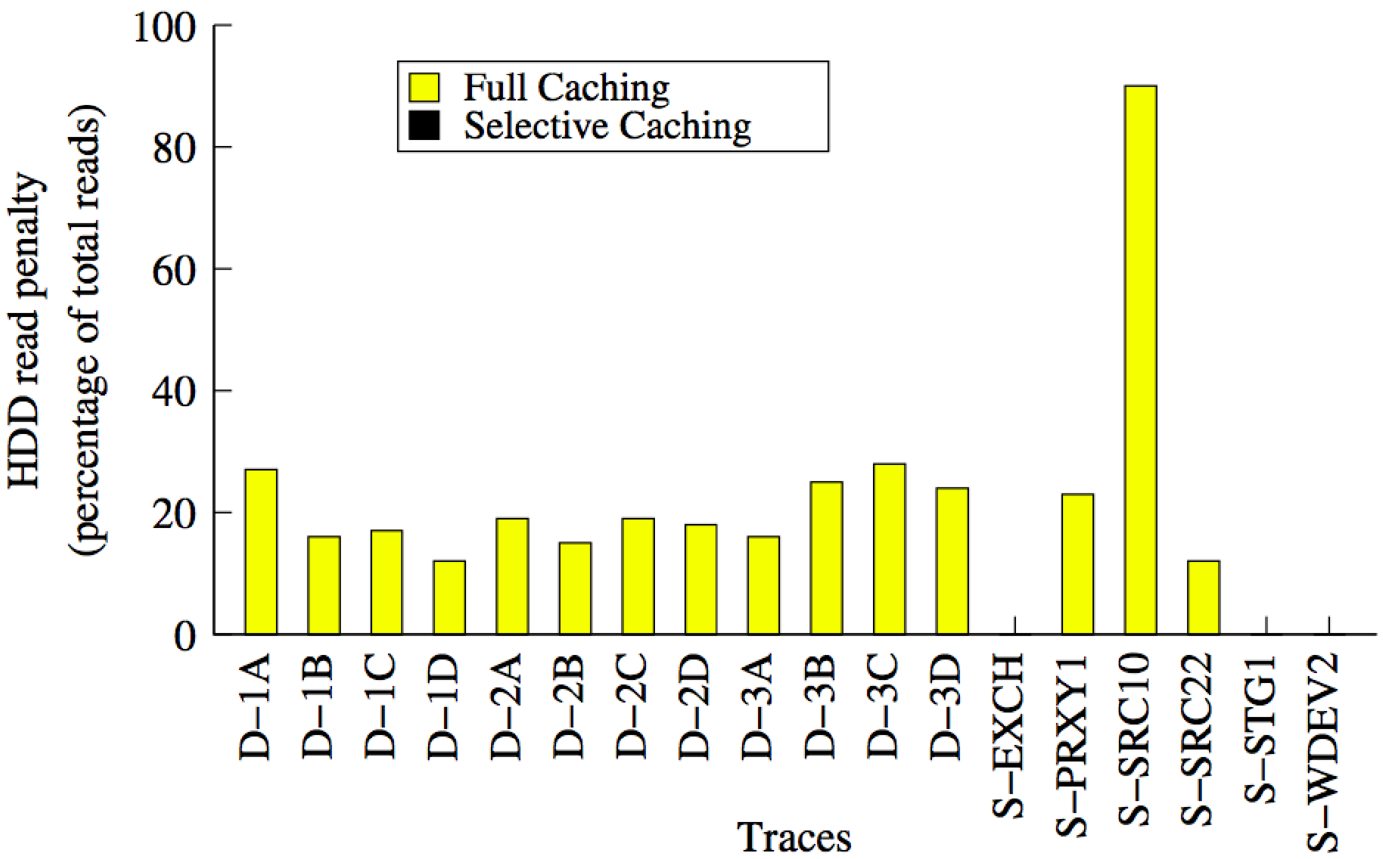
- Maintain overwrite ratio for each block
- Higher overhead but lower read penalty

What to Cache – Write Savings



Selective Caching (overwrite threshold = 0.25)	21 % average
Full Caching	51 % average

What to Cache – Read Penalty

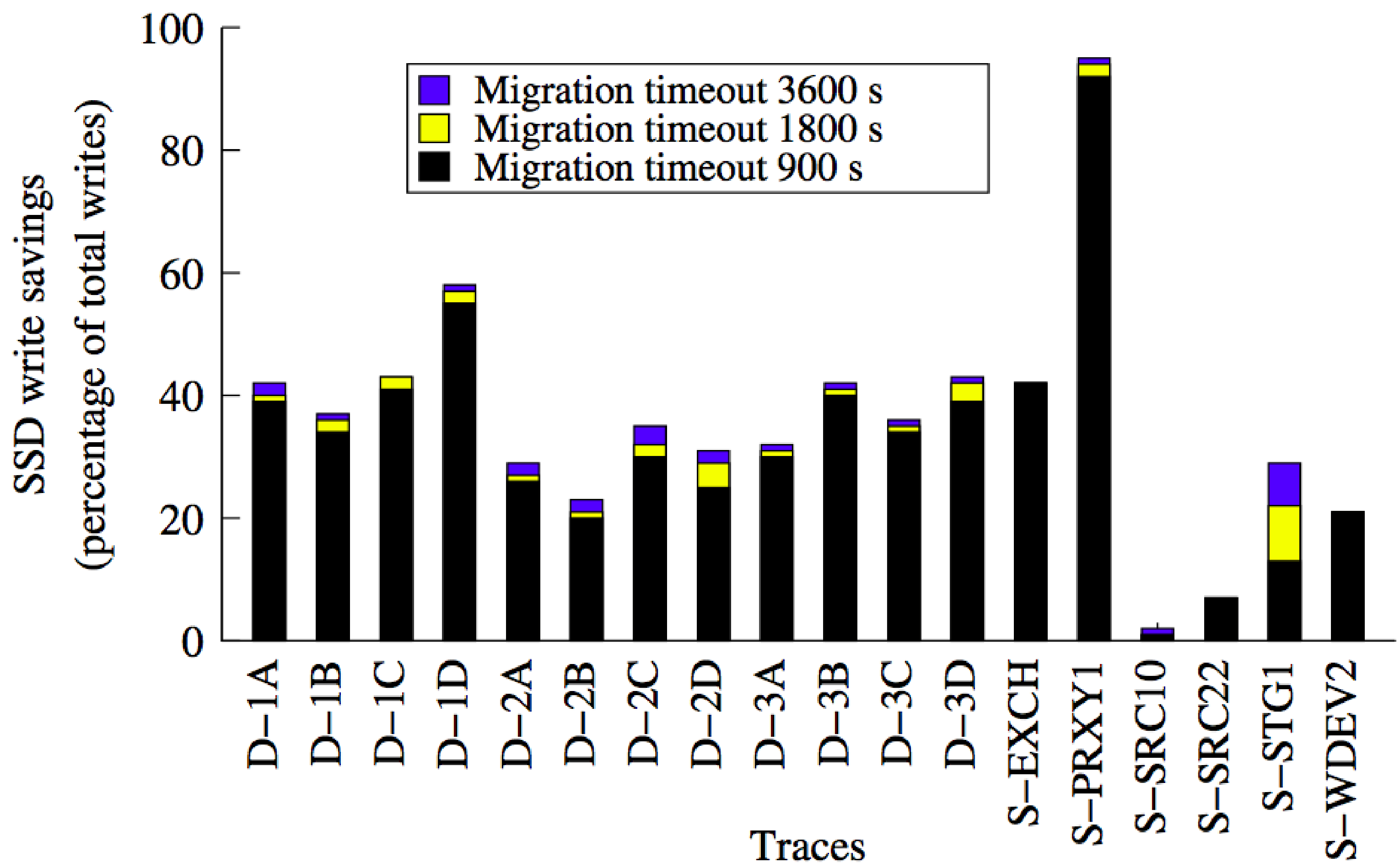


Average Read Penalty – 20%

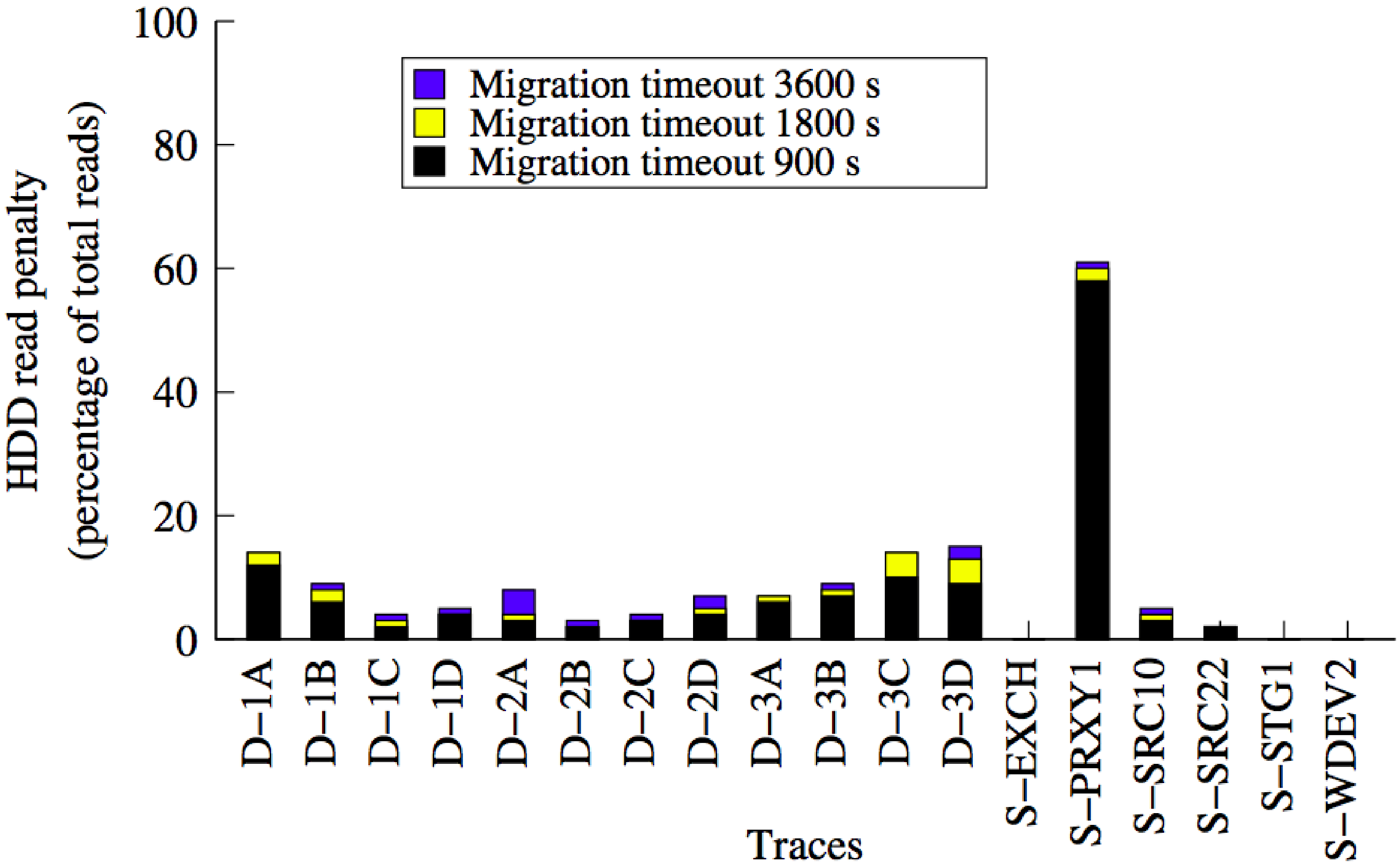
How long to Cache

- Timeout Trigger
  - Ex: Migrate every 5 minutes
  - Bounds data lost due to failure
- Read-Threshold Trigger:
  - Ex: Read Penalty should not exceed 5%
  - Ensures performance is reasonable
- Migration Size Trigger:
  - Ex: Migrate if cache size is > 100 MB.
- Hybrid scheme – Combine all triggers

How long to cache – Write Savings



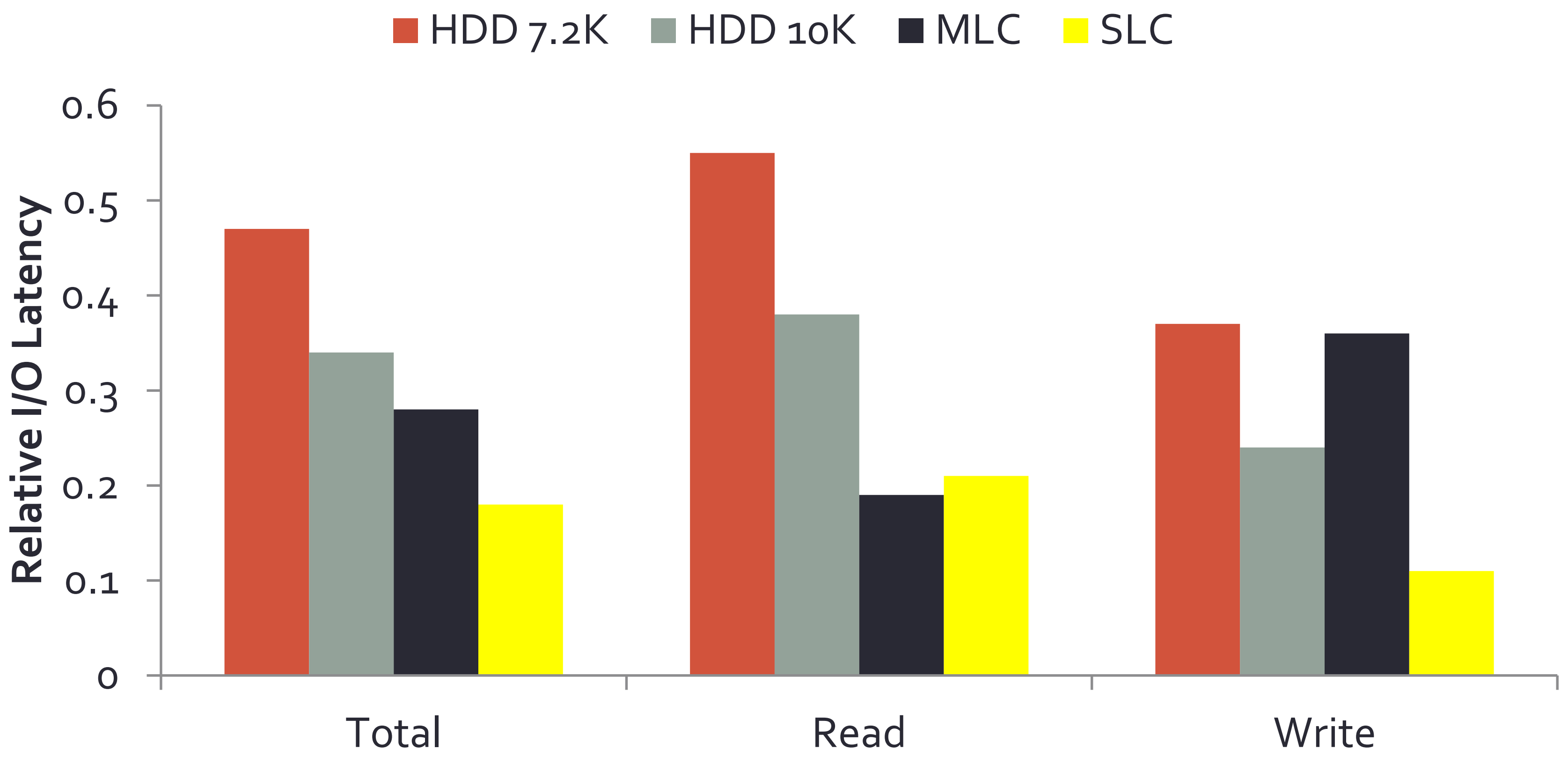
How long to cache – Read Penalty



Handling Failure

- Power Failures
  - Recovery similar to log-structured and journaling file systems
- Device Failures
  - HDD: Additional point of failure in storage stack
  - Full Caching - Most recent writes are lost
  - Selective Caching - More complex data recovery

Latency measurements



Relative to MLC-based SSD without Write Cache.

Discussion

- File system level vs Block level
  - File systems: More aware of what files to cache
  - Block Device: No modifications to software stack
- Power Consumption due to HDD
- Failure handling without caching all writes ?
- Adoption of Griffin given price and technology changes
- Phase Change Memory vs Flash

Comparison

	DRAM	Phase Change Memory	MLC NAND	HDD
Granularity	Bit	Bit	Block	Sector
Power	~W/GB	100-500mW /die	~100 mv/die	~10W
Write Bandwidth	~GB/s	1-100+ MB/s/die	~10 MB/s/die	200-400 MB/s
Write Latency	20-50 ns	~1 µs	~800 µs	~10 ms
Read Latency	50 ns	50-100 ns	25-50 µs	~10 ms
Endurance	∞	10^8	10^4	∞