MySQL and SSD

Vadim Tkachenko
Percona Inc, co-founder, CTO
www.percona.com
www.SSDPerformanceBlog.com
vadim@percona.com

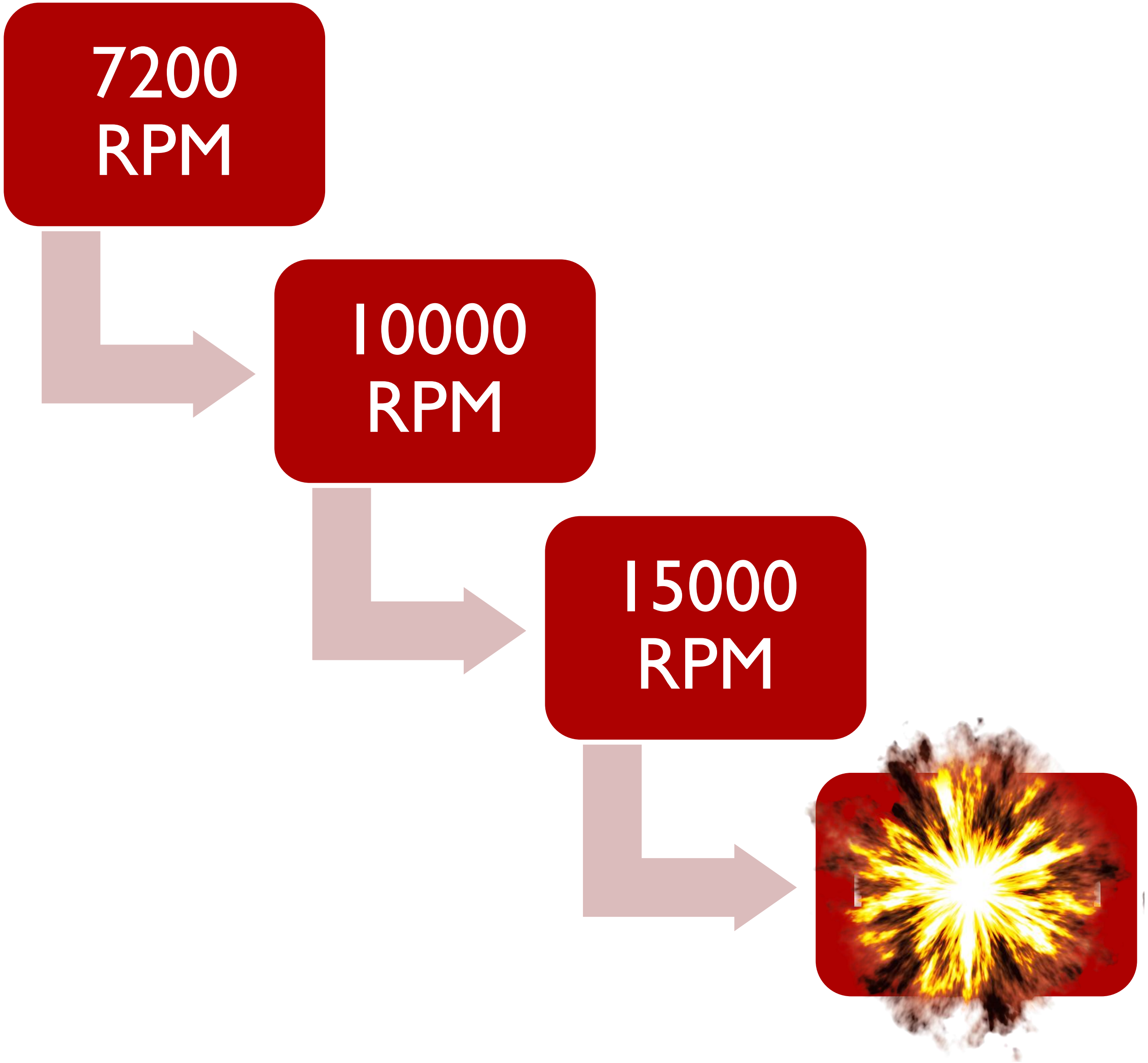
This talk online

- PowerPoint
  - http://bit.ly/MySQL-SSD-2012
- PDF
  - http://bit.ly/MySQL-SSD-2012-PDF
- Contacts
  - vadim@percona.com
  - Twitter @VadimTk

World is spinning

Physical limits

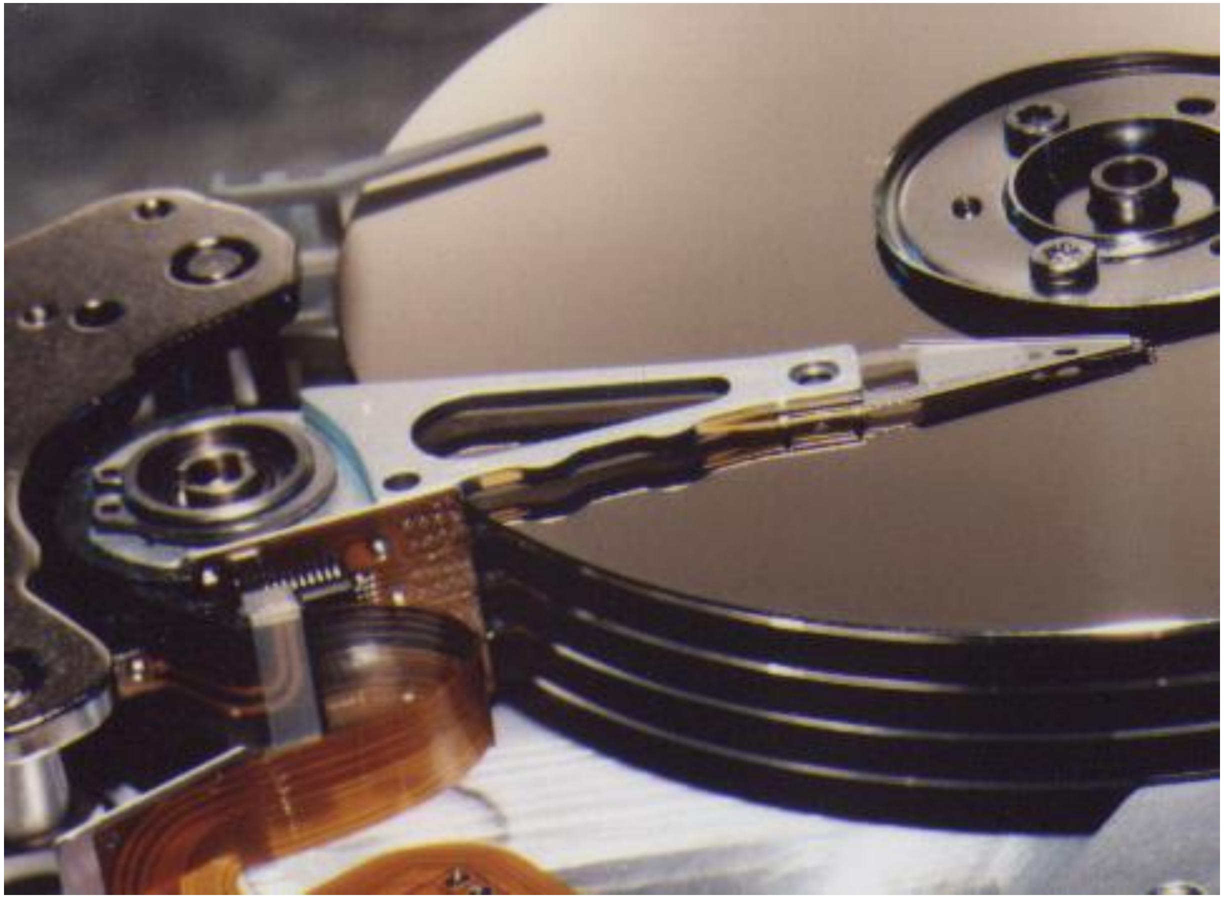
Rotate faster



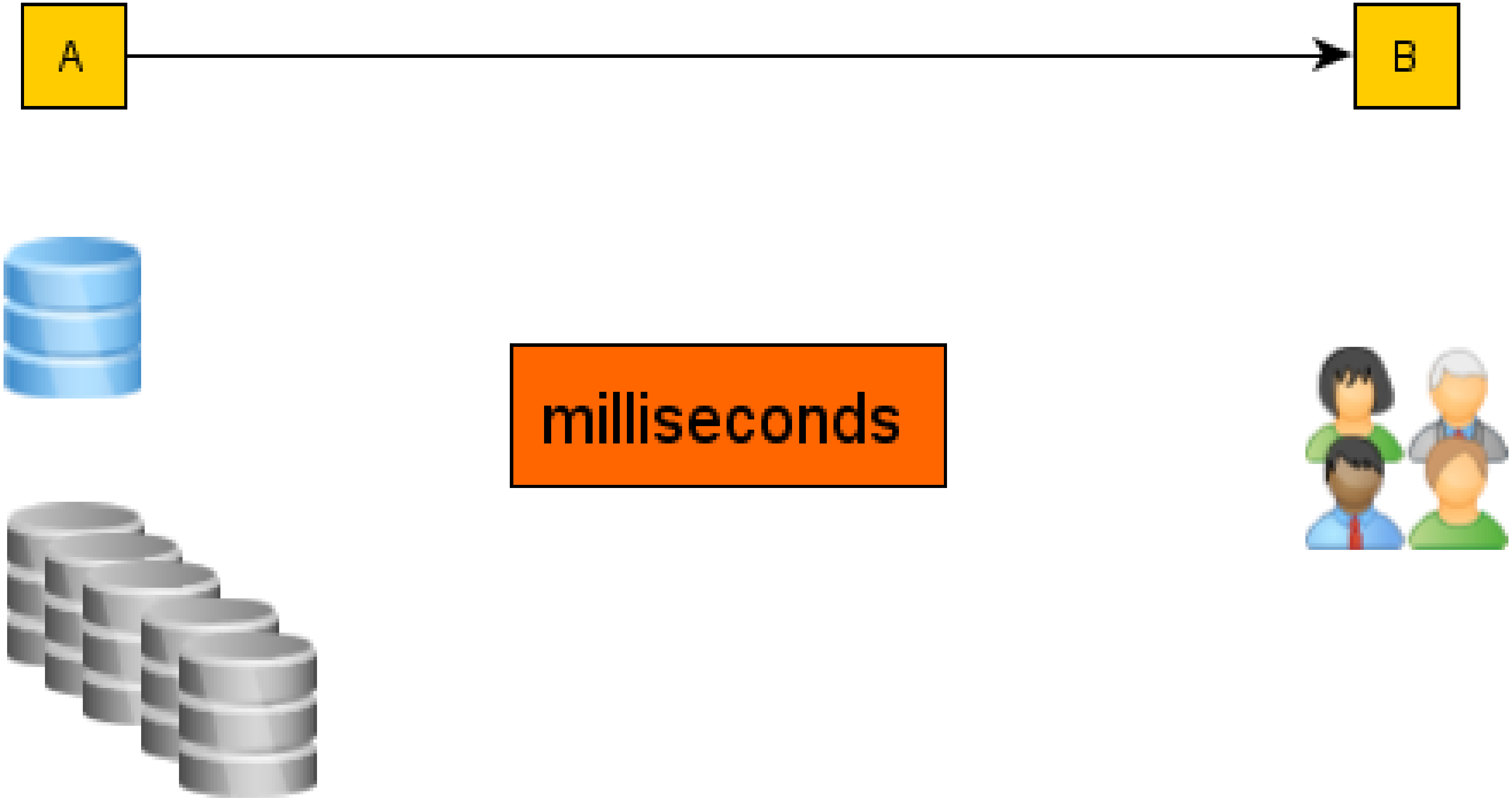
milliseconds

Access time

More spindles



Still milliseconds

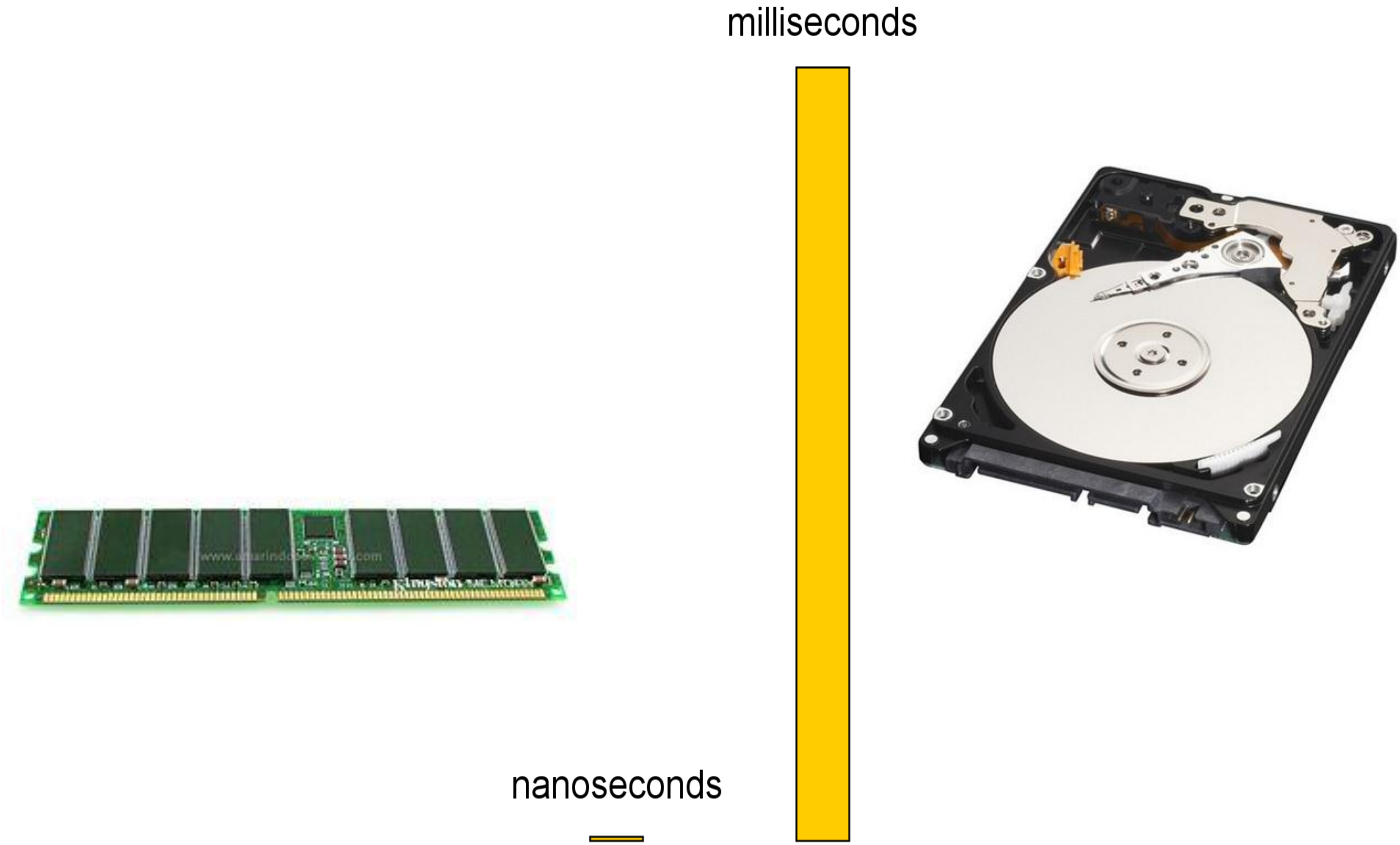


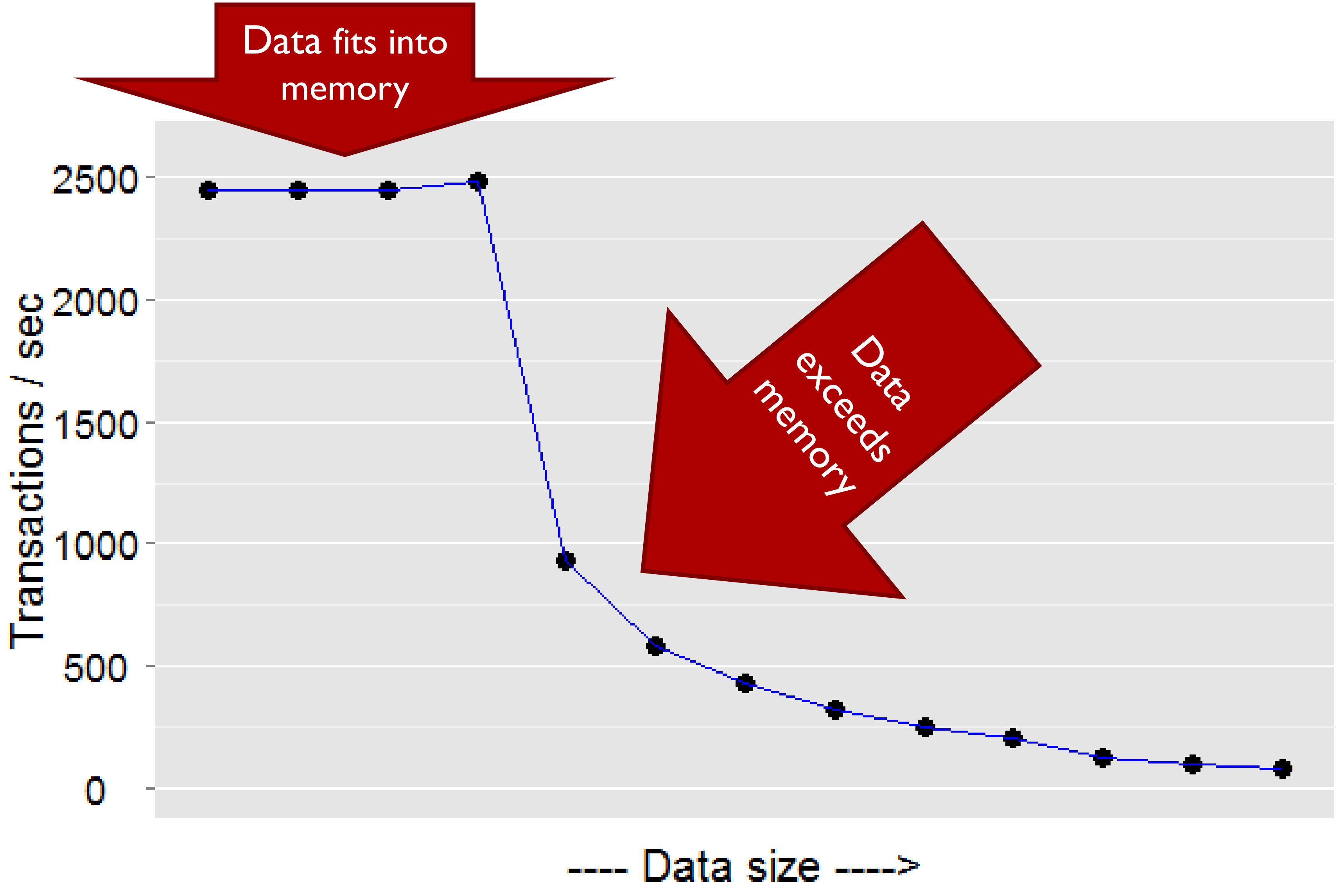
A

B

milliseconds

Access time

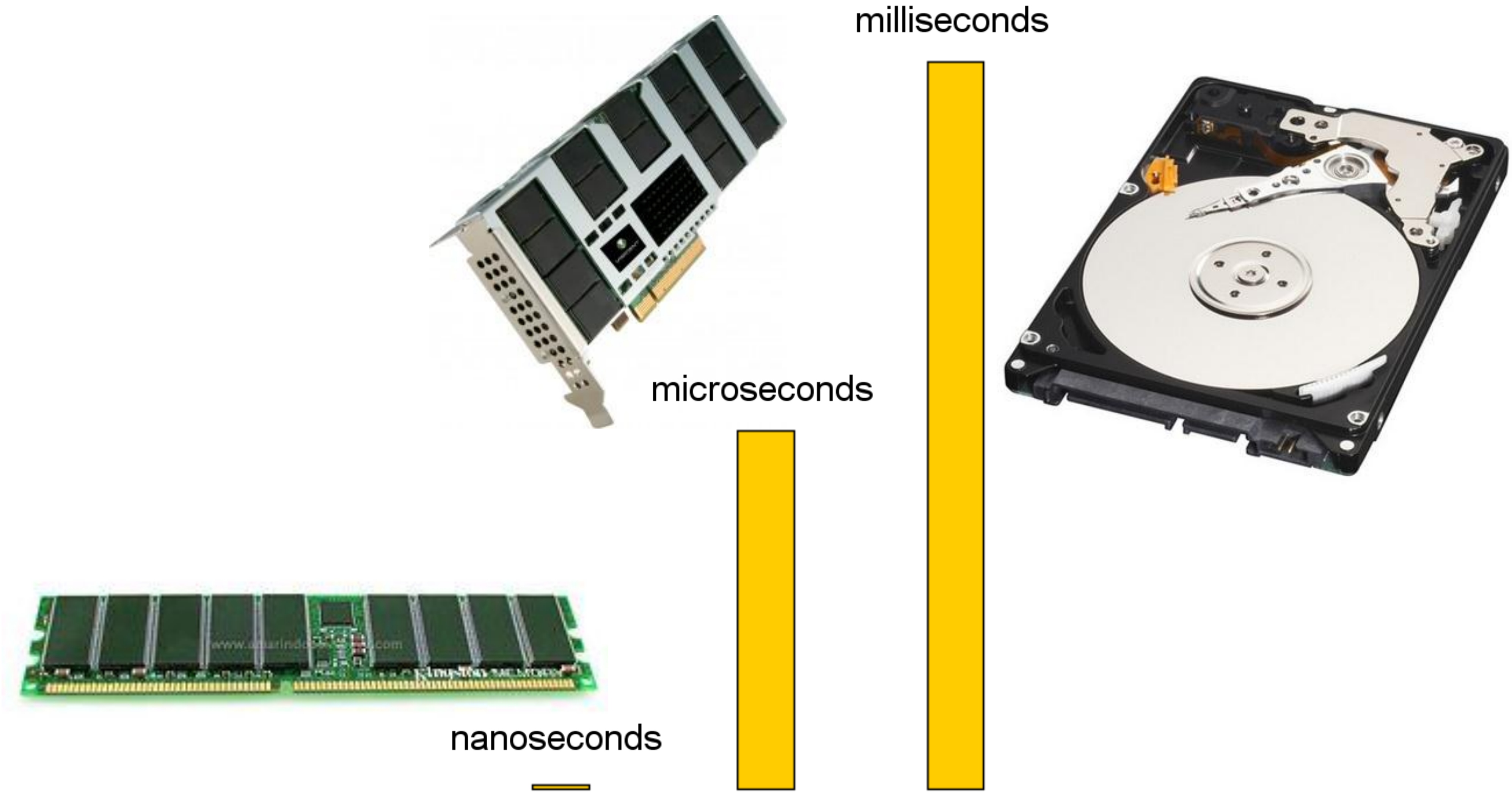



Data fits into memory
Data exceeds memory
2500
2000
1500
1000
500
0
Transactions / sec
---- Data size ---->

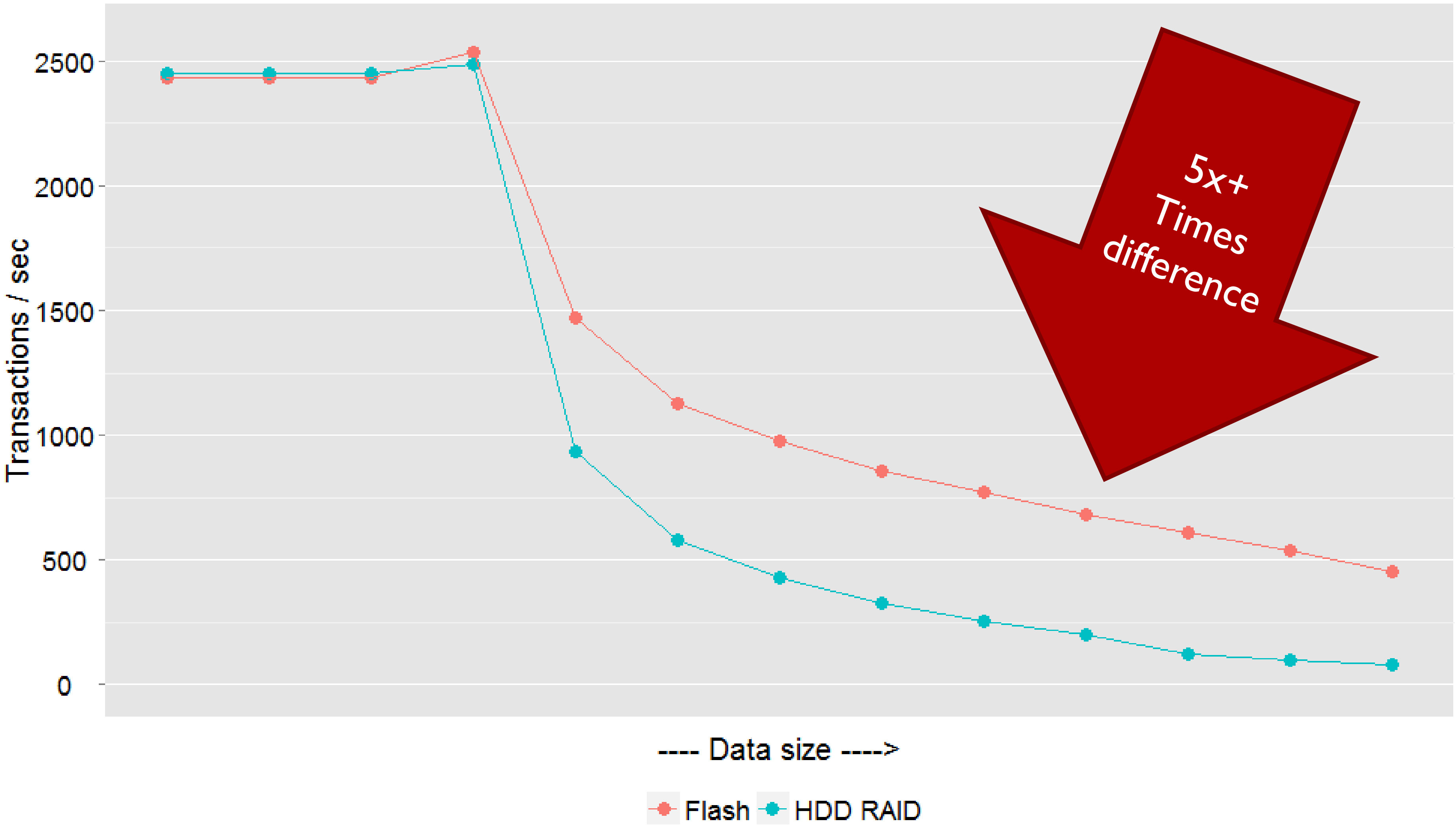
10% data growth

70% throughput drop

Flash Access time

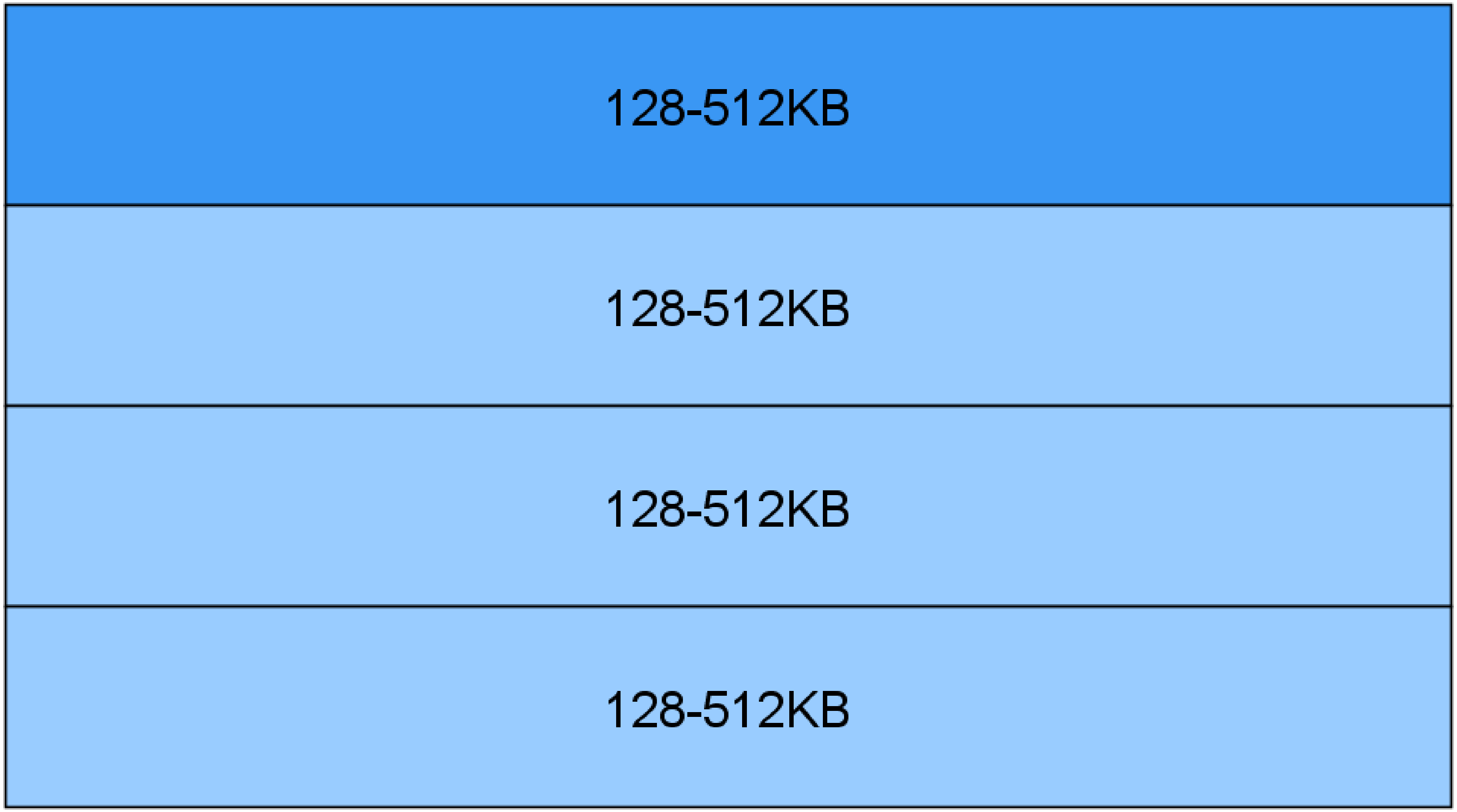


MySQL throughput with Flash



Flash

Erase size



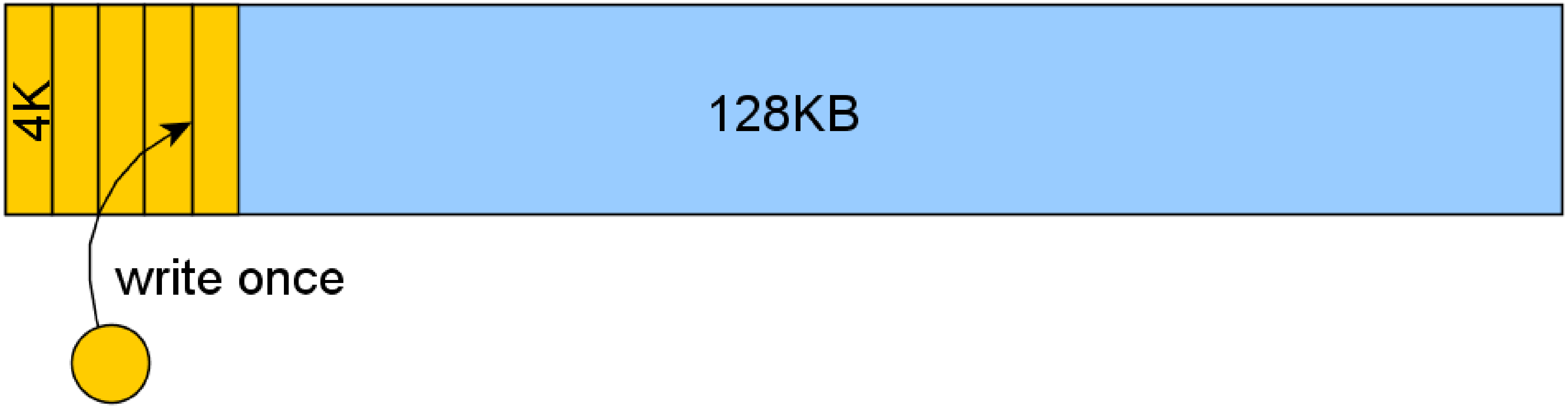
128-512KB

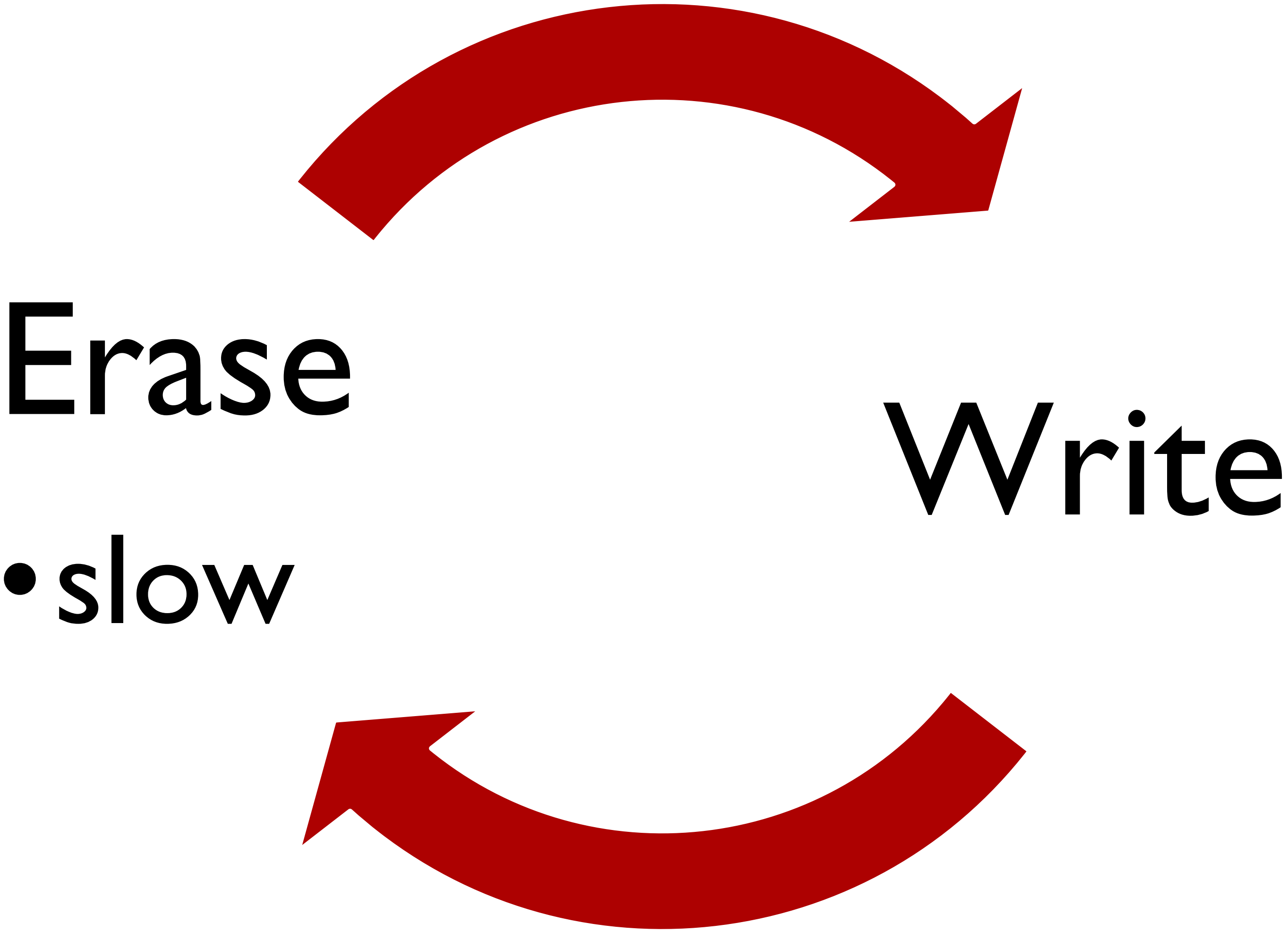
128-512KB

128-512KB

128-512KB

Write once



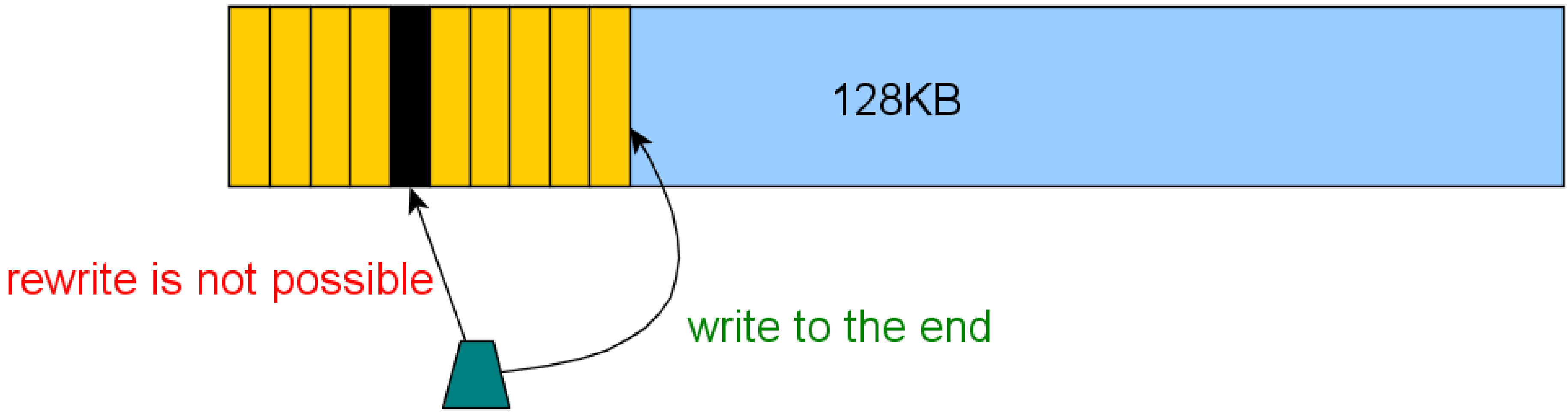


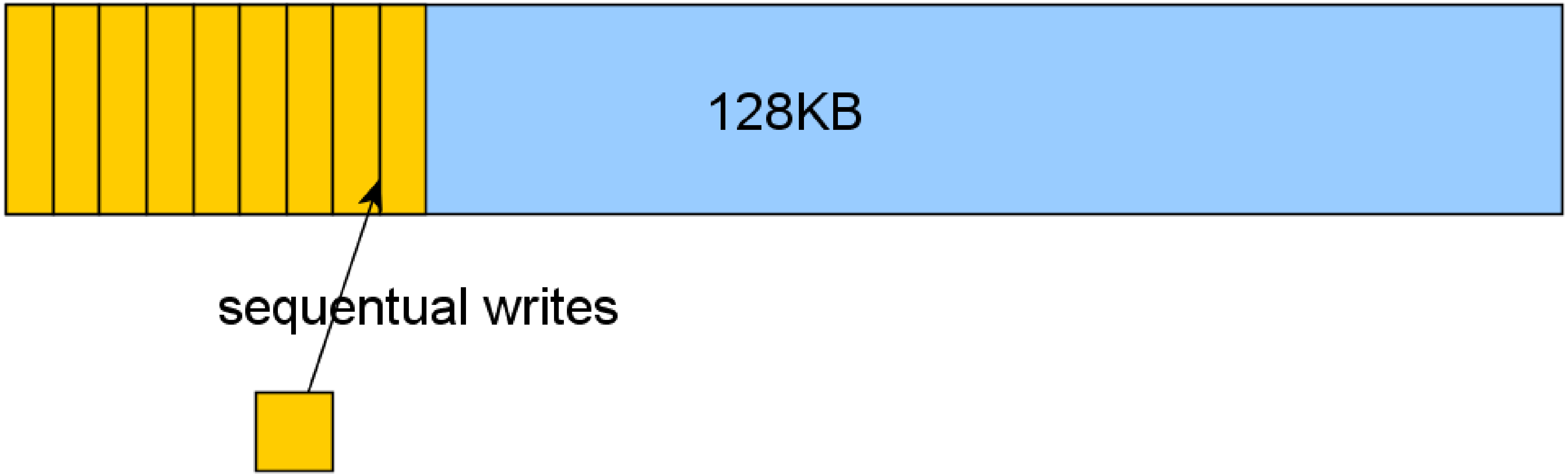
Erase

- slow

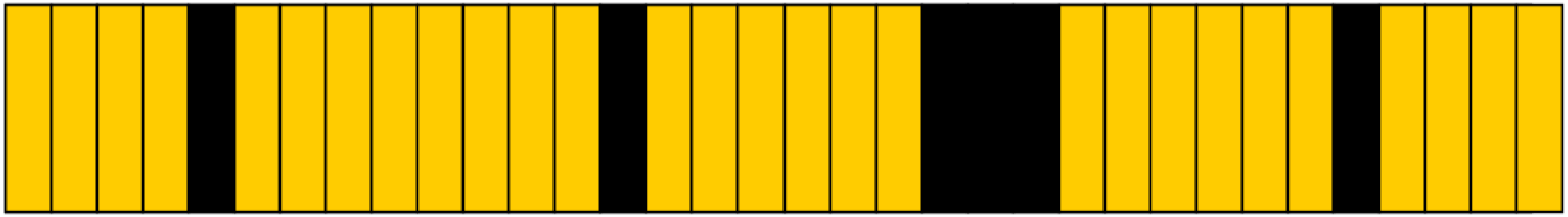
Write

No rewrites




128KB
sequentual writes

Garbage collector

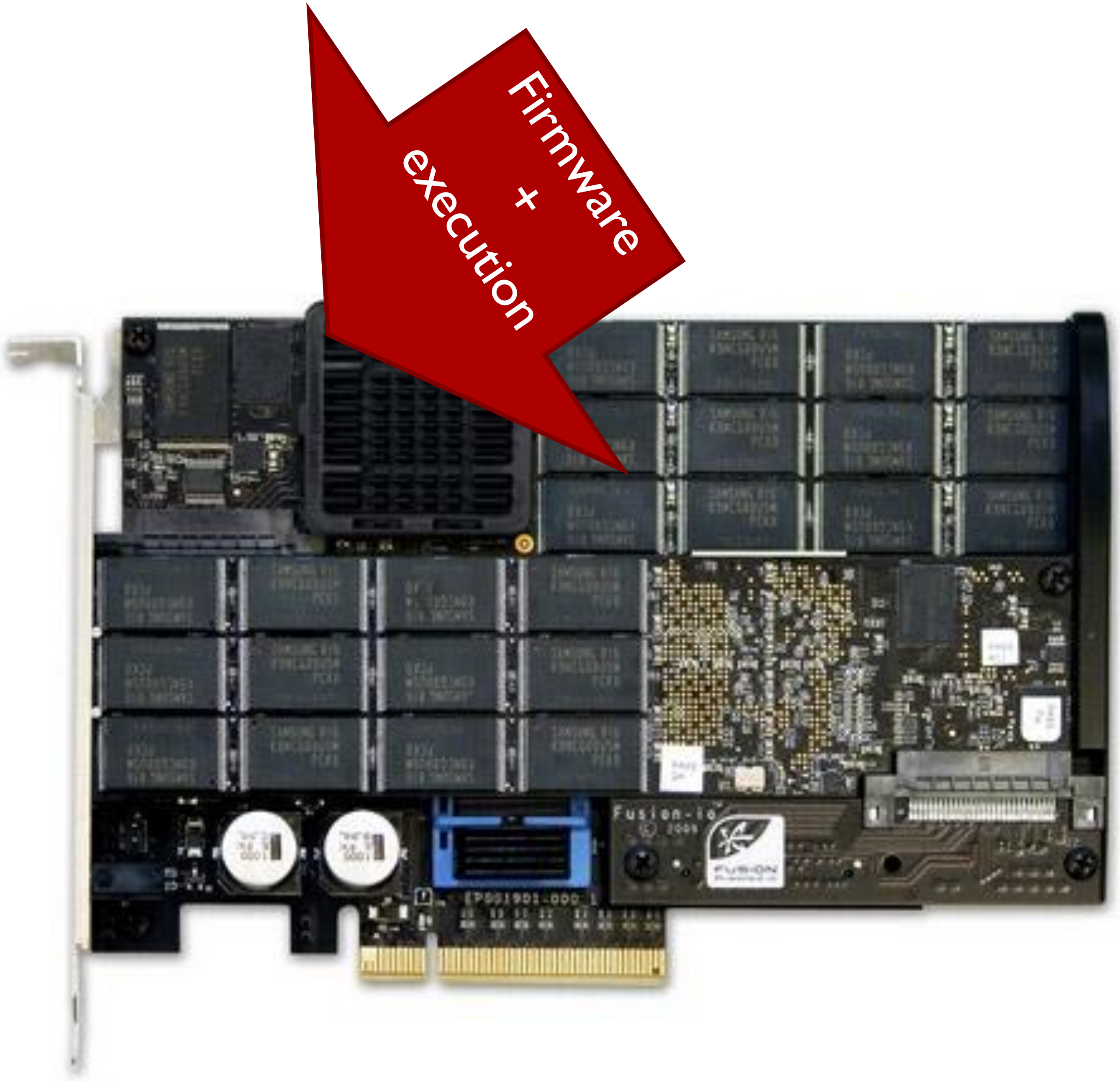


Write amplification

Flash writes more than application

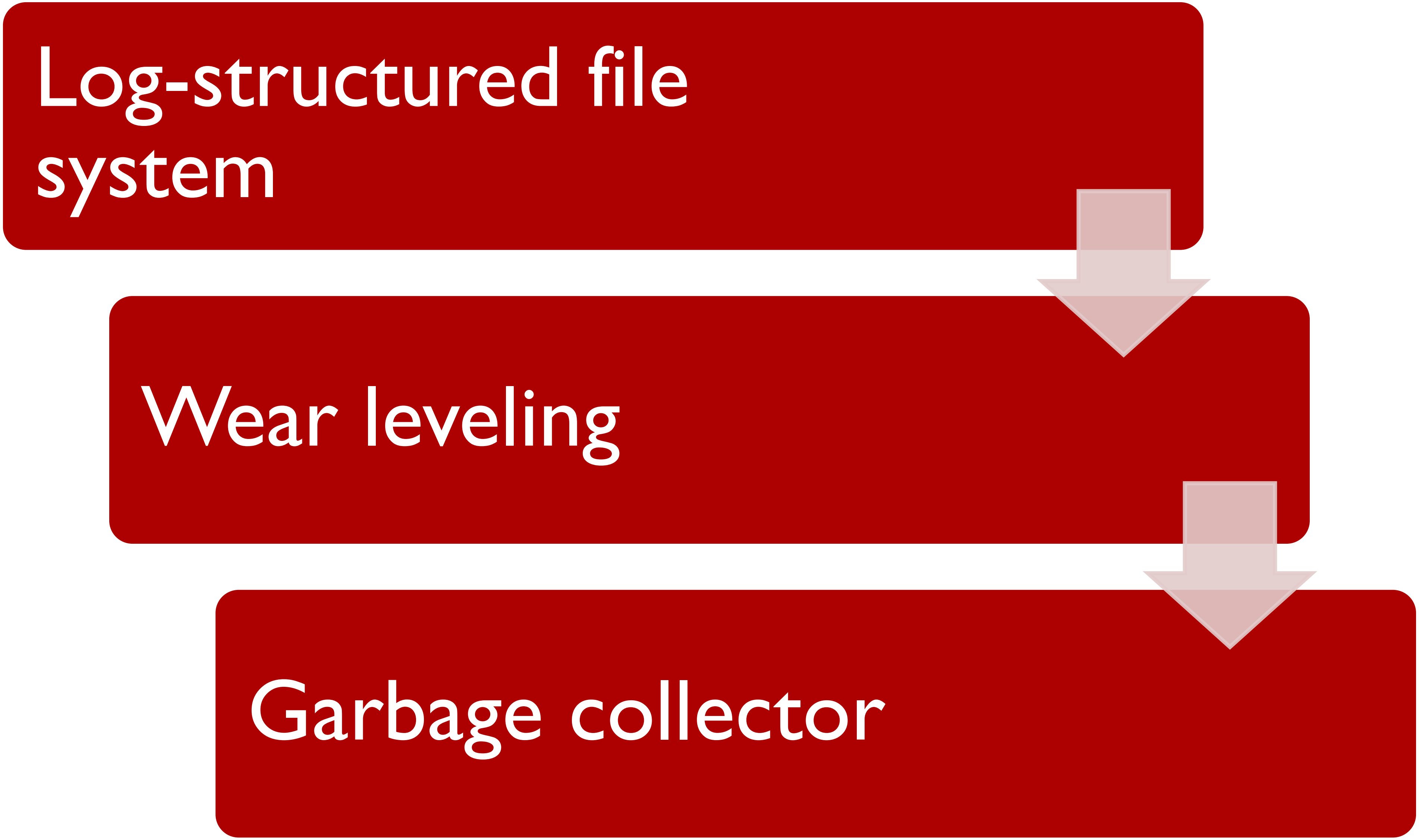
Software matters

Hardware is less important



Firmware
+
execution

Flash quality is defined by software



Log-structured file system

Wear leveling

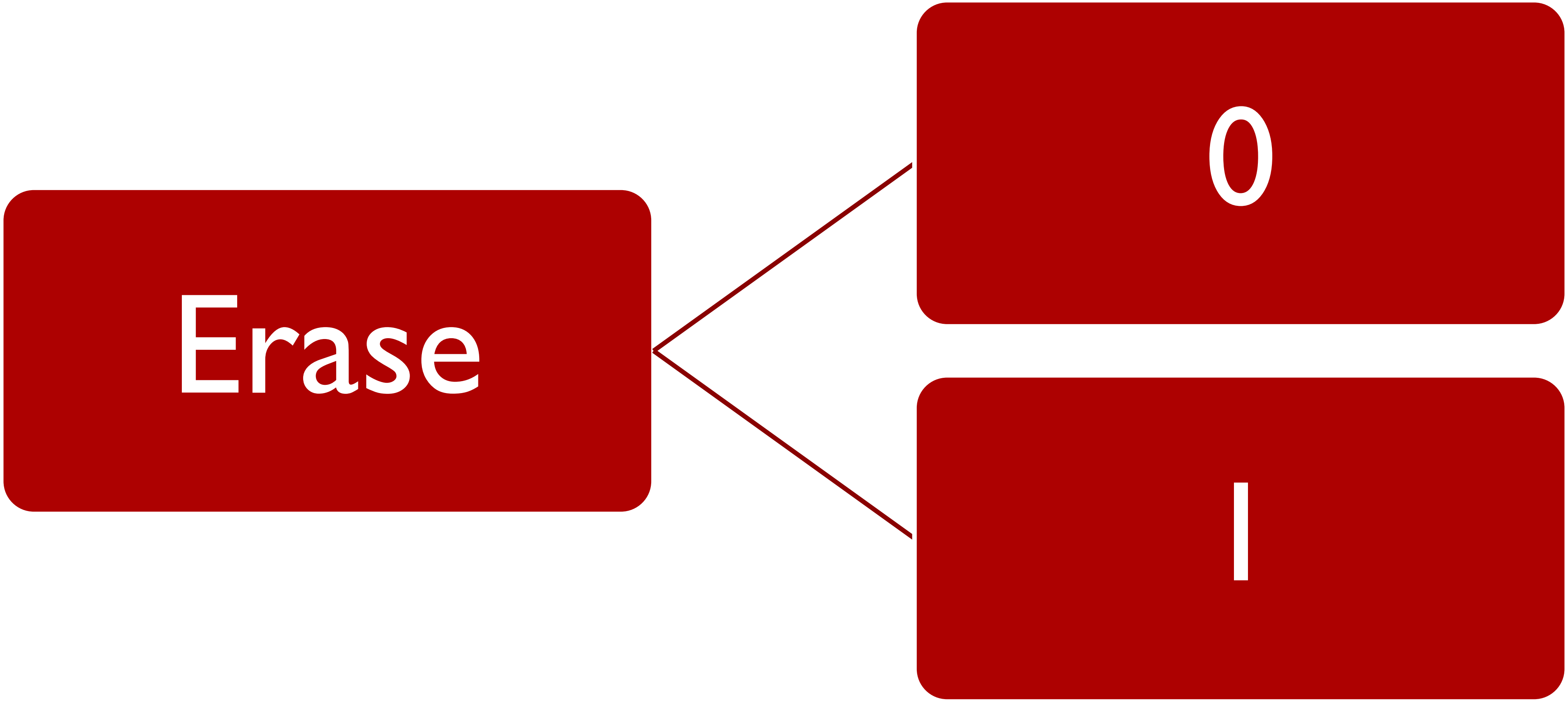
Garbage collector

Flash types

SLC

MLC

Single Level Cell – 1 bit

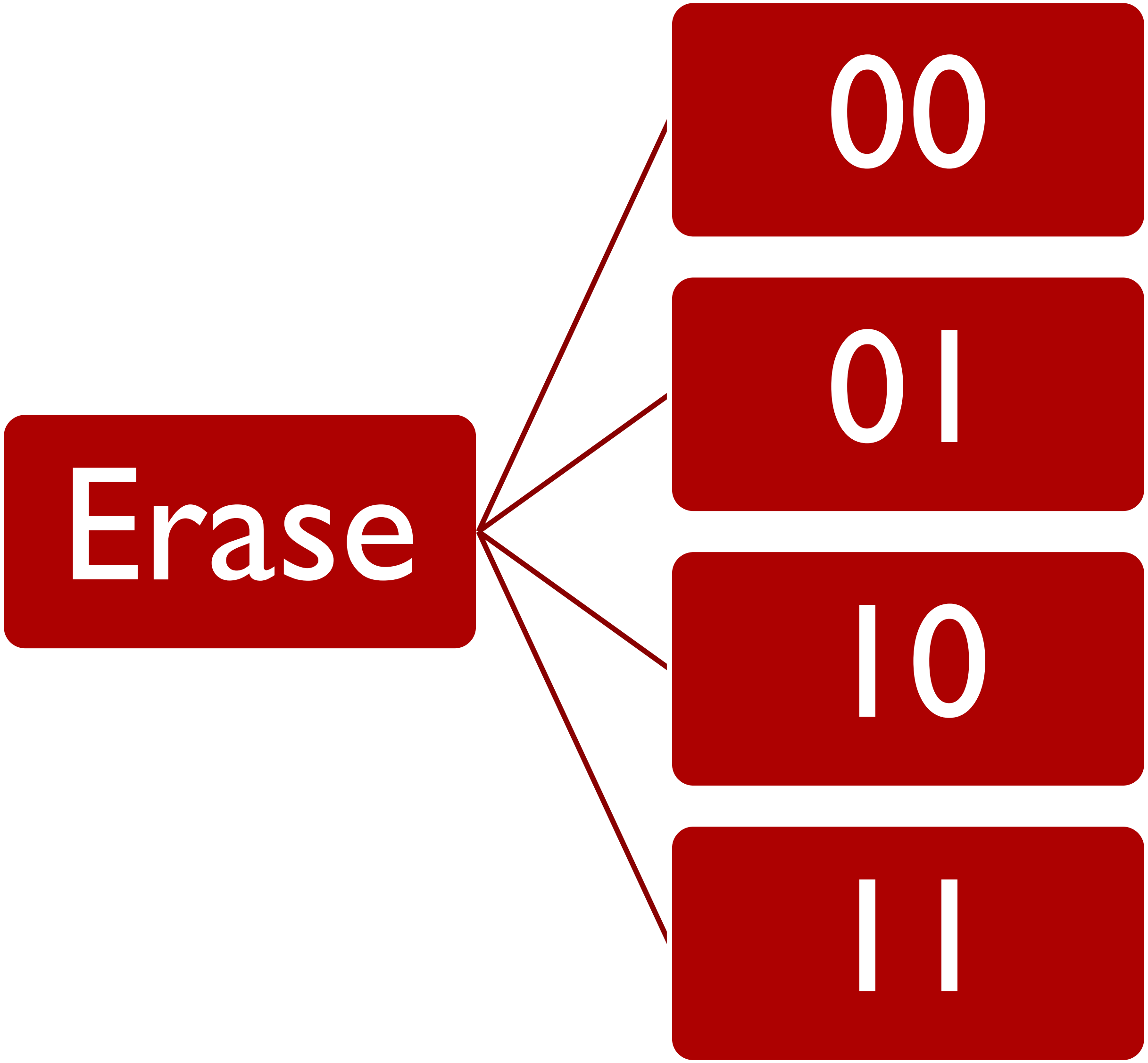


Erase

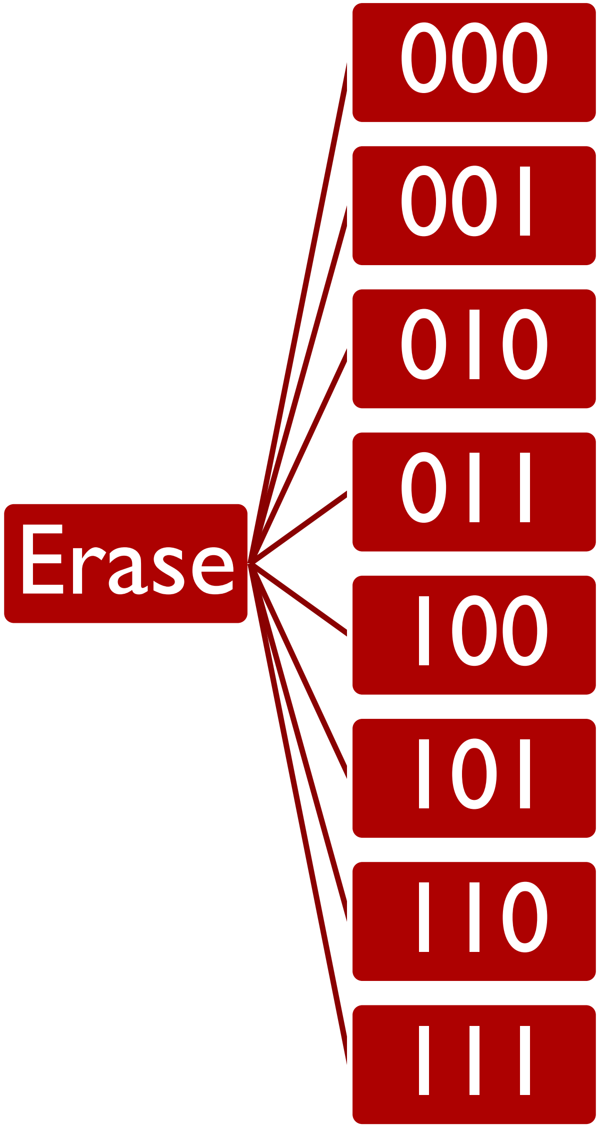
0

1

Multi Level Cell – 2 bit



Multi Level Cell – 3 bit



Erase

000

001

010

011

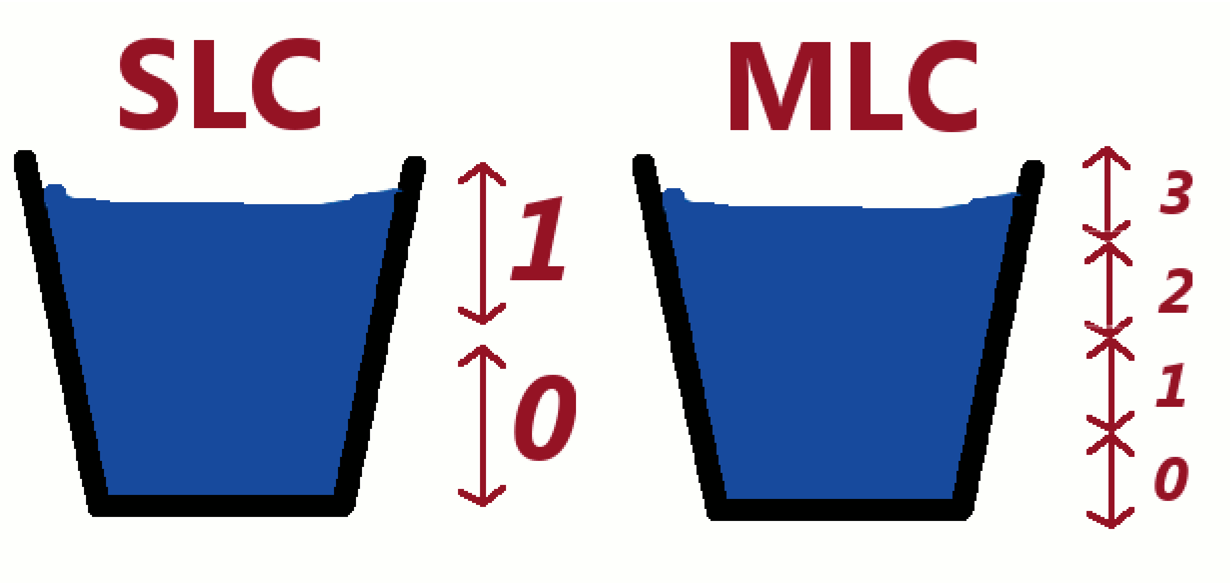
100

101

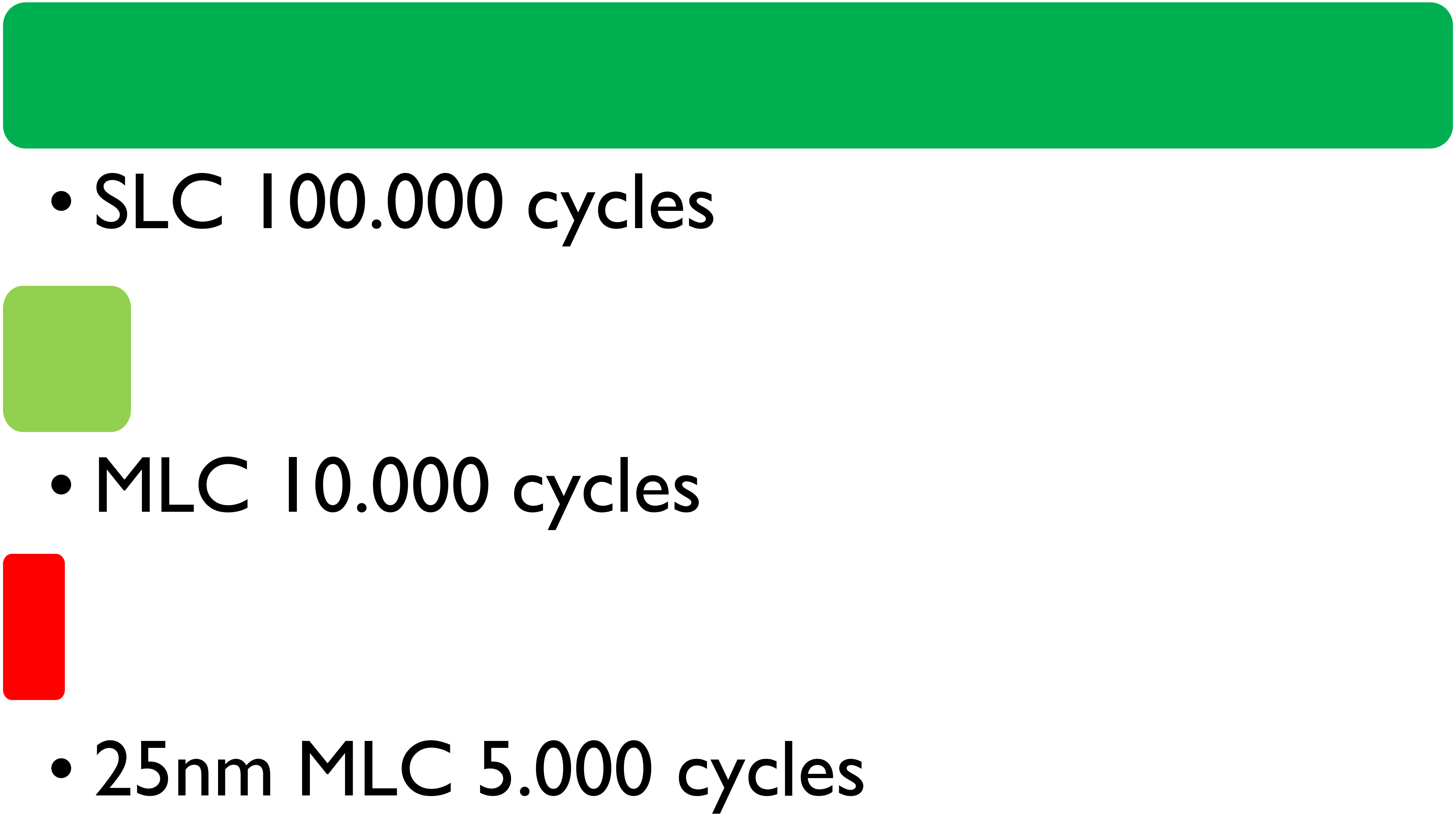
110

111

SLC vs MLC



Erase cycles



- SLC 100.000 cycles
- MLC 10.000 cycles
- 25nm MLC 5.000 cycles

SLC

Benefits

- Reliability
- Performance

Drawbacks

- Up to 800GB
- Expensive
- 30-50$/GB

MLC

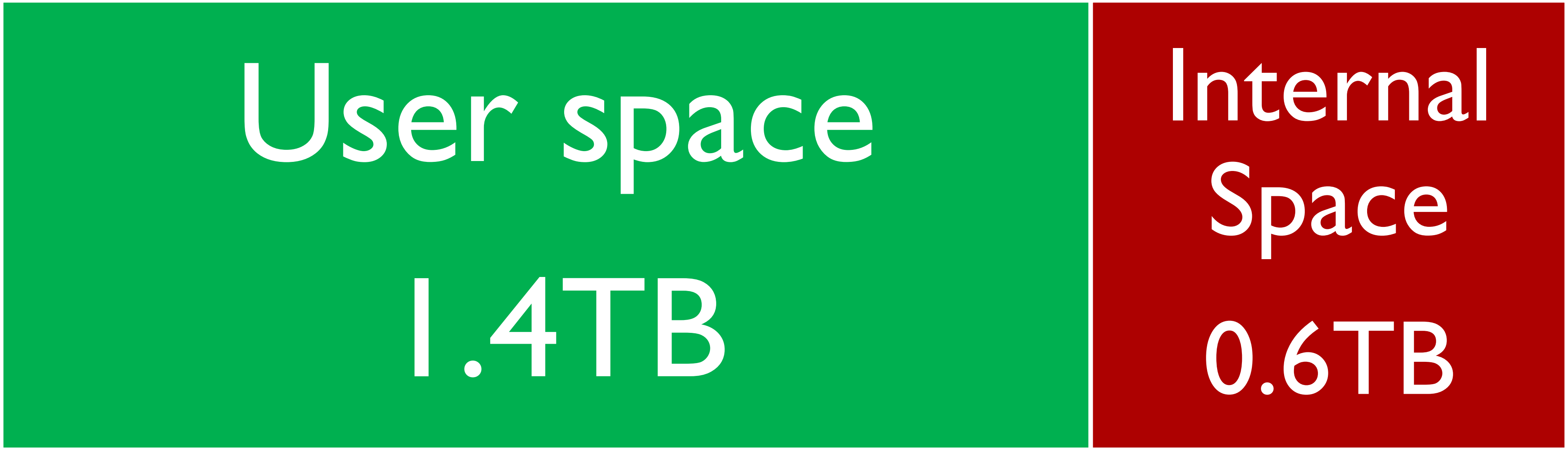
+

- Over 1TB
- 10-15$/GB

−

- Life time
- Reliability

Space provisioning –Virident FlashMax 1400



User space

1.4TB

Internal Space

0.6TB

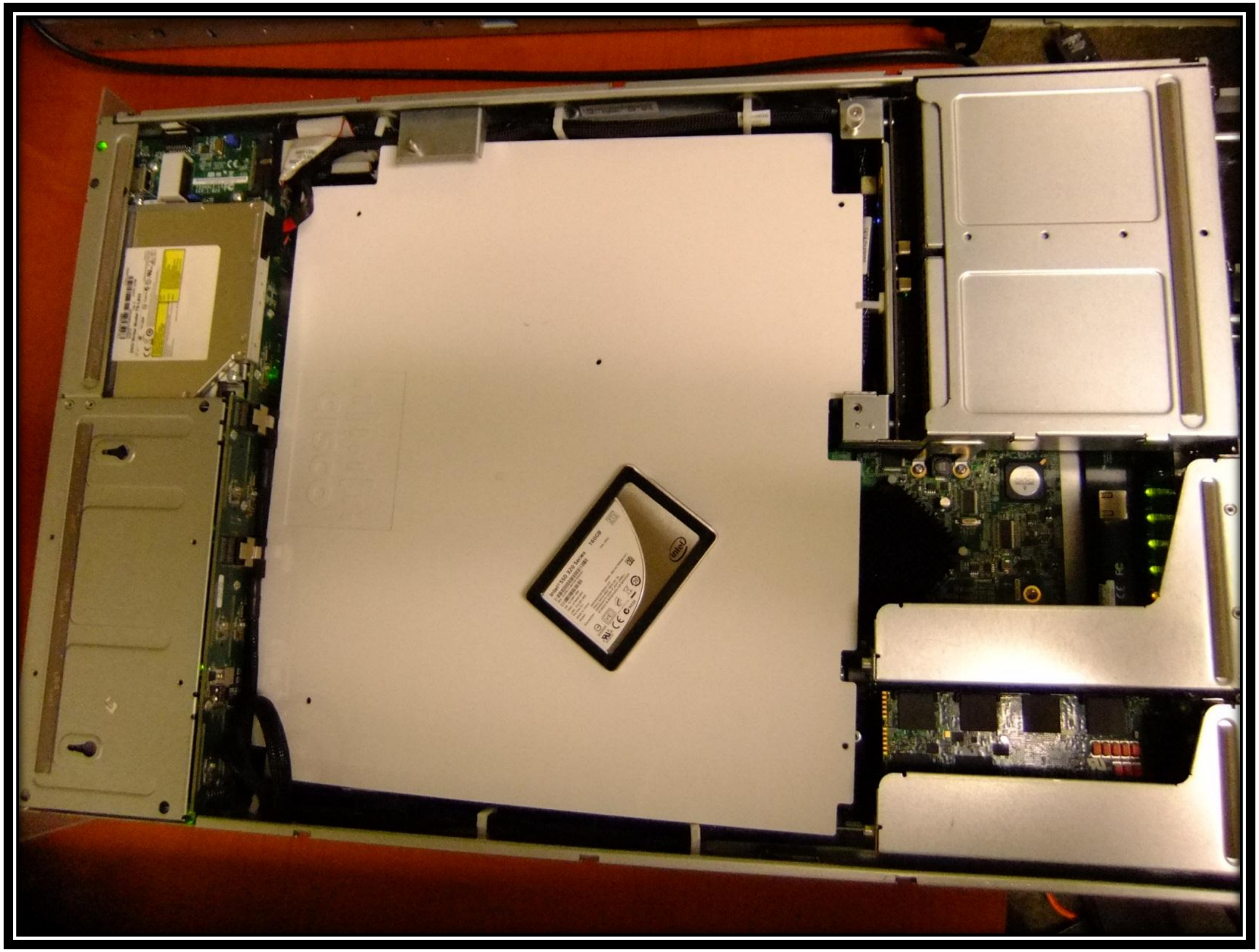
SATA vs PCI Express

SATA SSD

My benchmarks story

2 Intel 320 SSD cards

How do I install it?



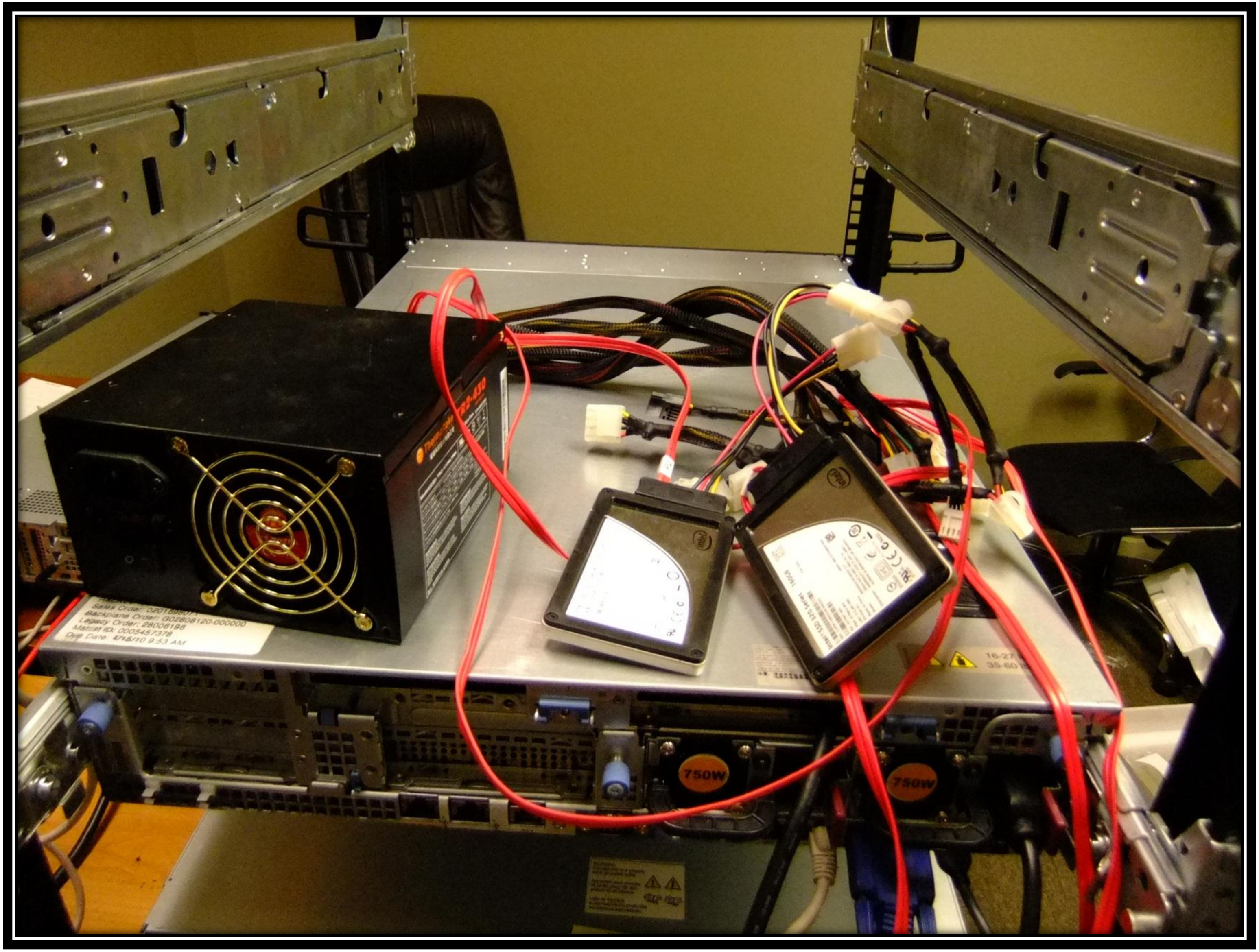
I need

Space

Power

Controller +cables

Initial setup



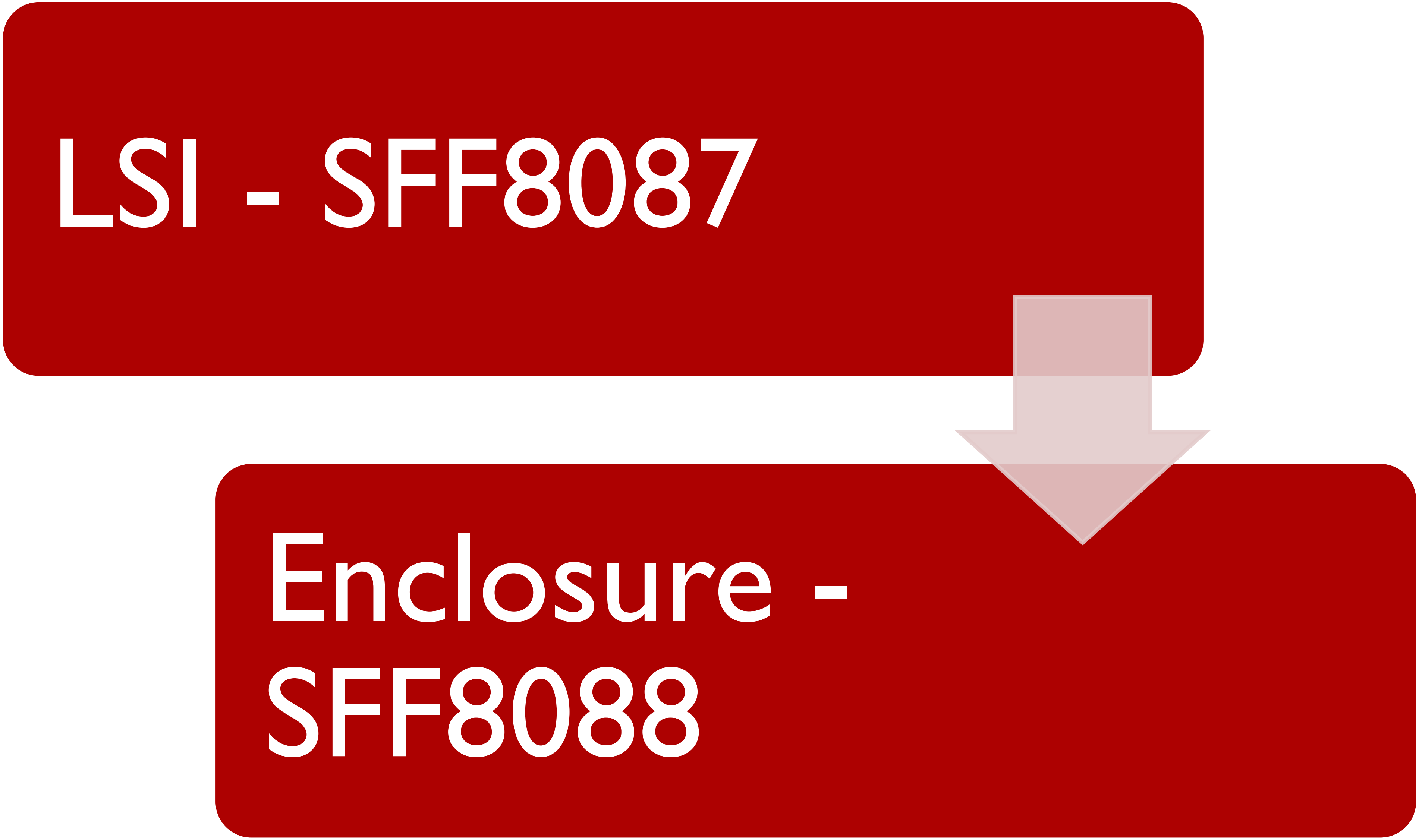
Polished setup - AccuSTOR AS108X

RAID controllers

LSI 9260

LSI 9211

Last component - cable



LSI - SFF8087

Enclosure - SFF8088

PCIe is different

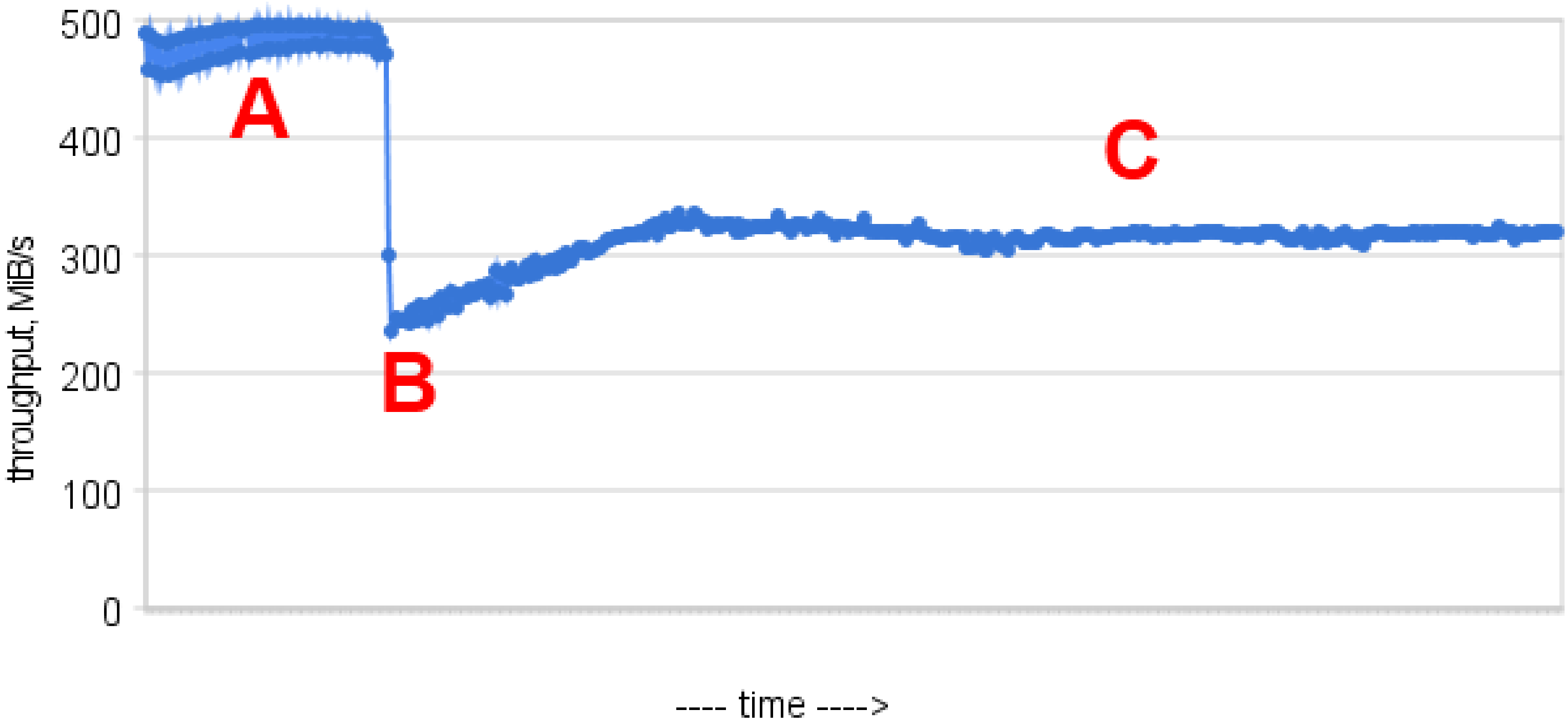
PCIe

Just plug into a PCIe slot

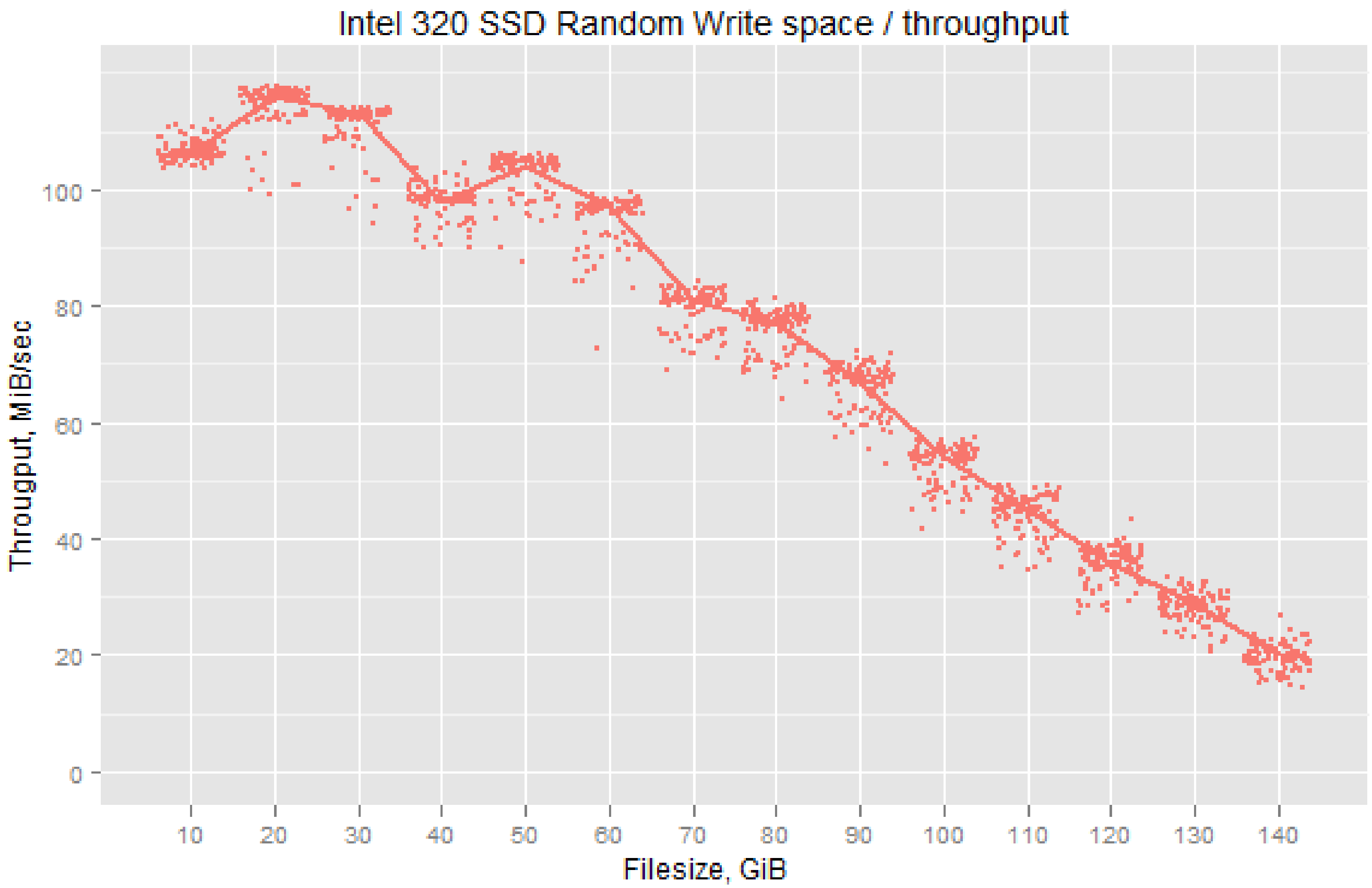
But SATA is hot-swap

Benchmarks lie

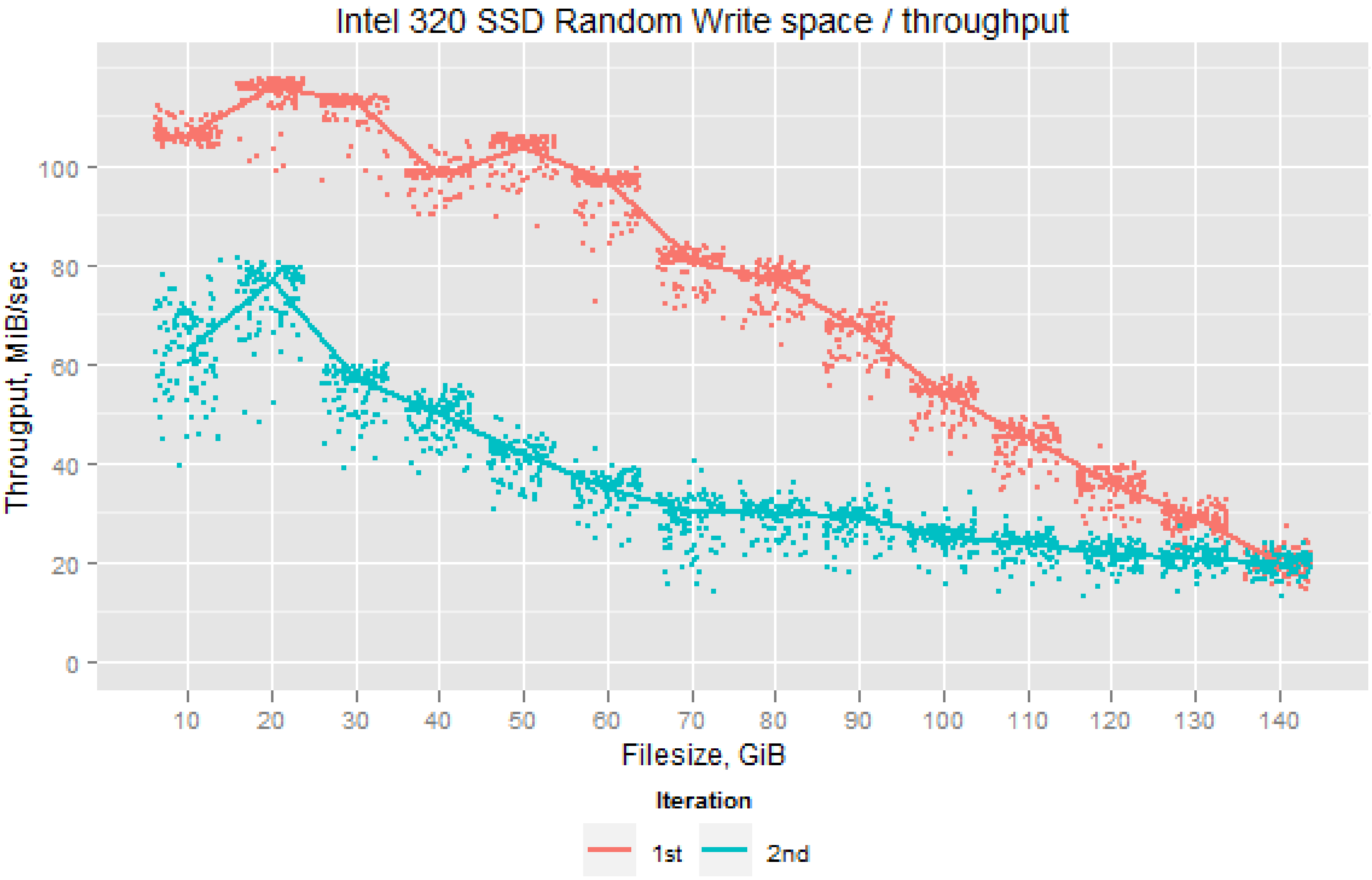
Benchmark challenge: internal state



Benchmark challenge: capacity



Unrepeatable results

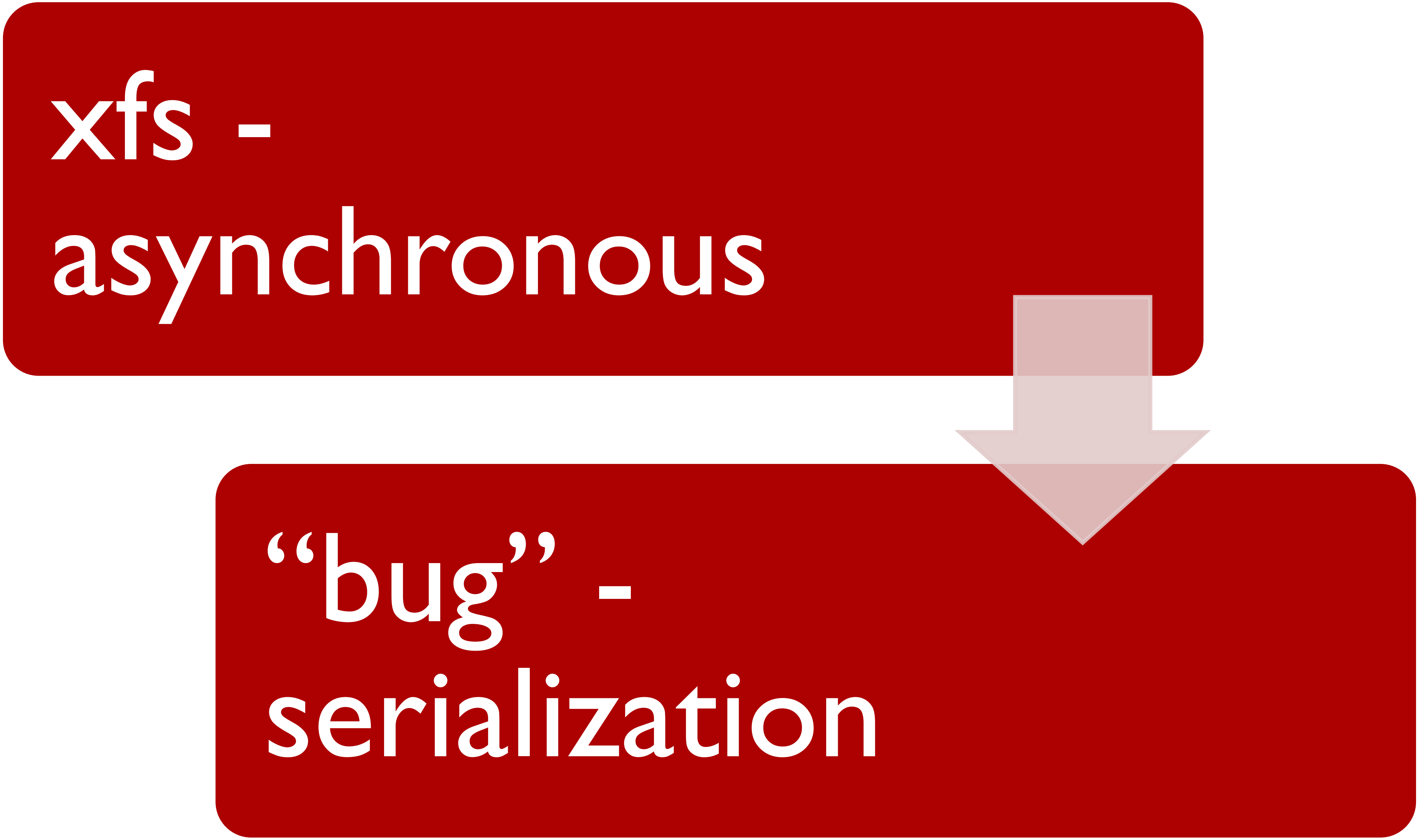


Benchmark challenge: filesystems

Ext3/4 – synchronous IO

“bug” in O_DIRECT

Benchmark challenge: filesystems



xfs - asynchronous

“bug” - serialization

Xfs already fixed bug in source code

4x improvement

Benchmark challenge: filesystems

btrfs – not ready yet

To add to confusion: in MySQL

Reads - sync

Writes – async

Readahead reads - async

Ext4 vs xfs – your choice

Comparing apples

8xHDD RAID10

- 2.5" 15K RPM HP Smart Array

STEC MACH16 200GB

- SATA SLC

4xSTEC MACH16

- RAID10 – LSI 9211-4i

Intel 320 SSD 160GB

- SATA MLC

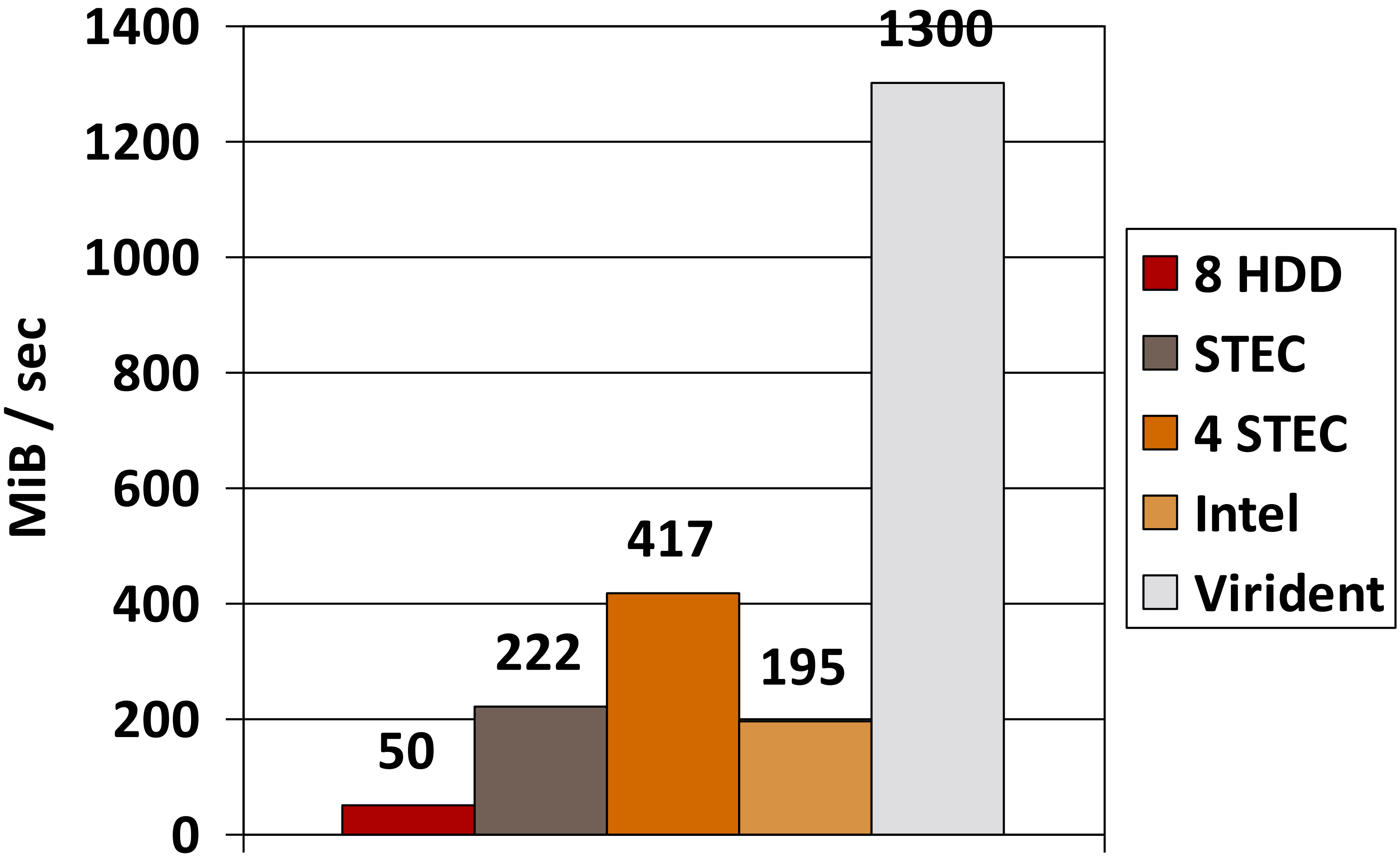
Virident FlashMax 1400

- PCIe MLC

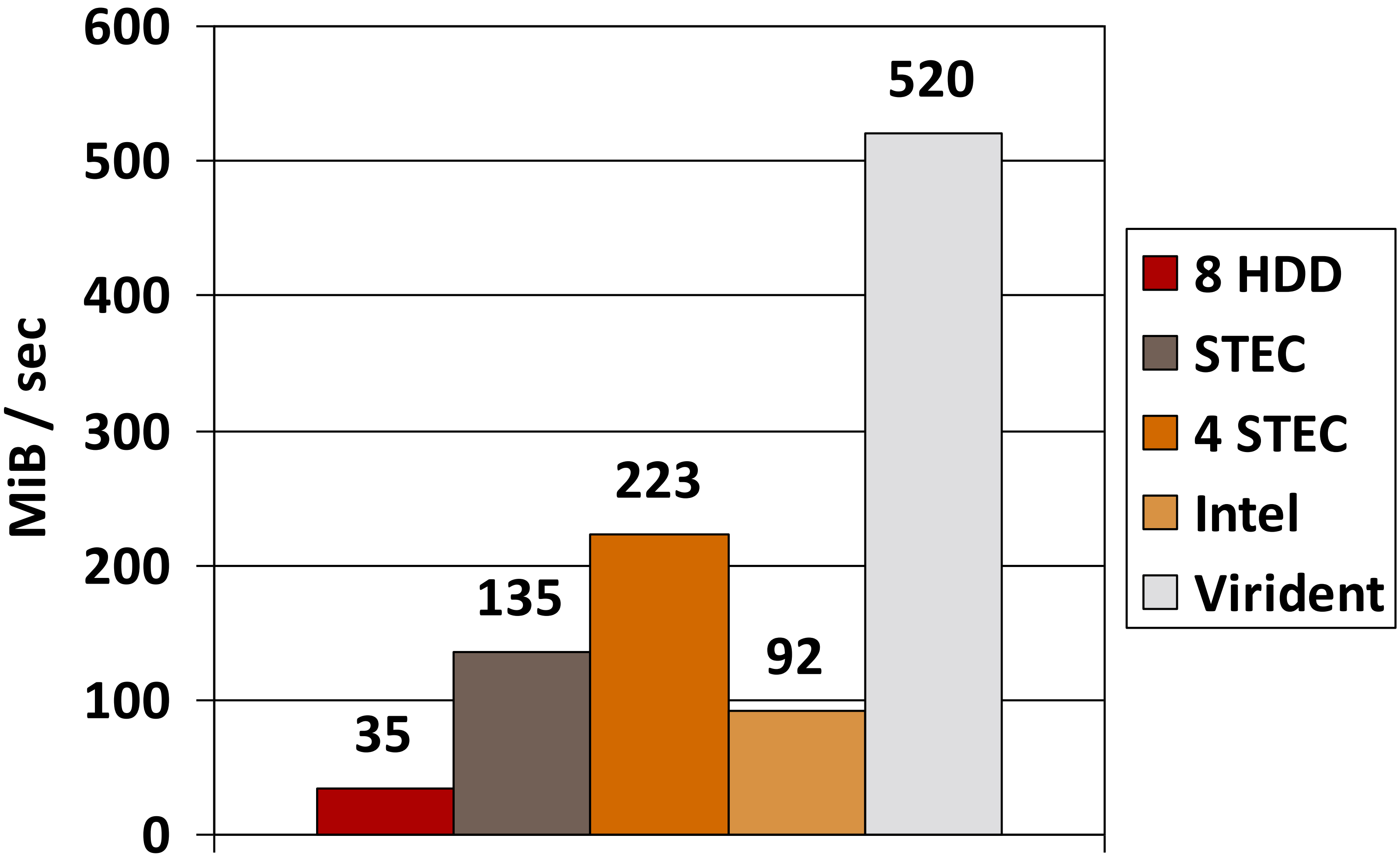
Flash vendors

~50 on market

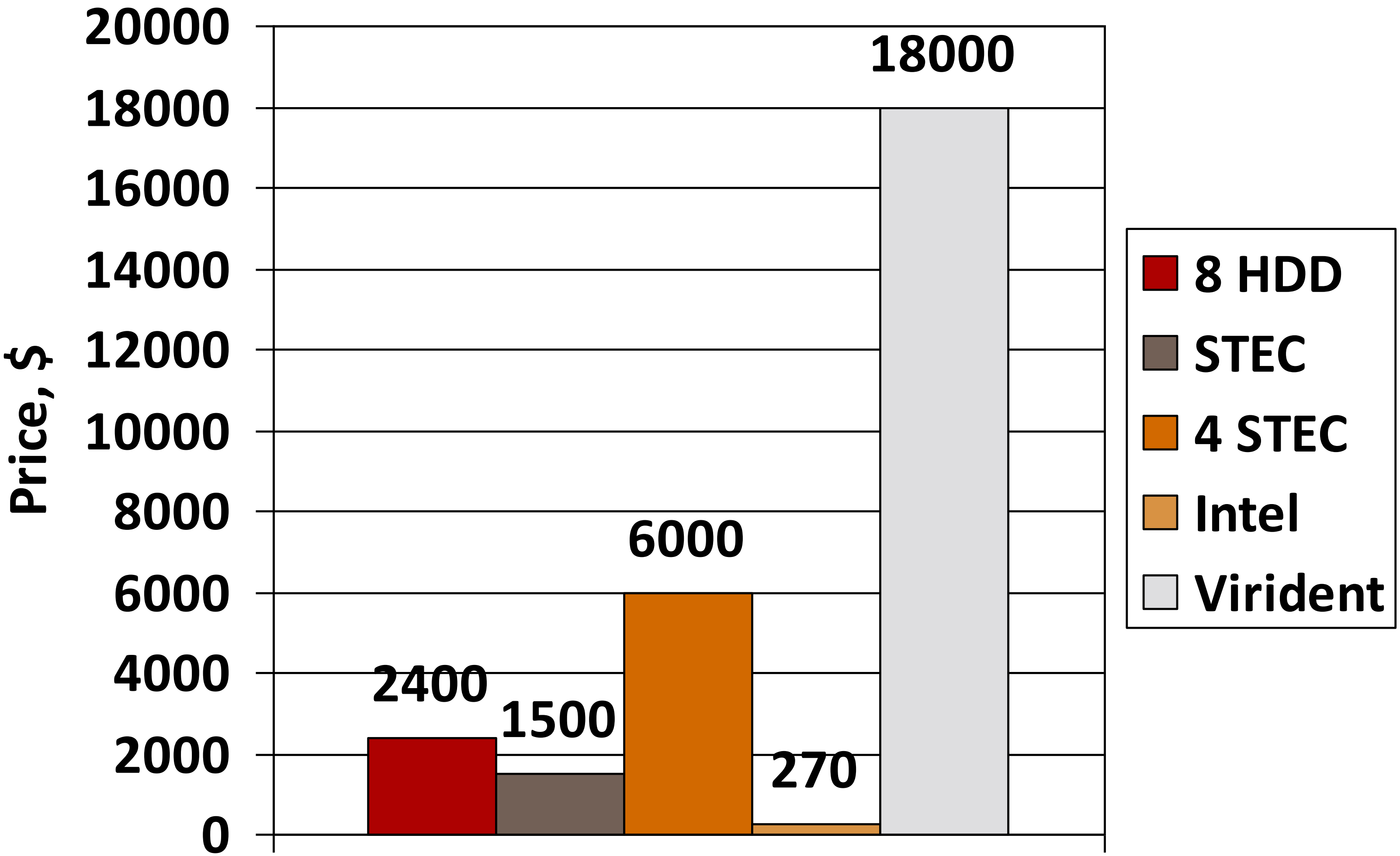
Random 16KiB reads



Random 16KiB writes



Price


Price, $
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
2400
1500
6000
270
18000
8 HDD
STEC
4 STEC
Intel
Virident

PCIe vs SATA

Which one to choose?

PCIe for absolute performance

I use it because I have free samples

SATA for performance per $

I would use it if I had to buy…

When should I use flash?

Very good for random reads

Both SLC and MLC

Random Writes

Maybe challenge for MLC

SLC lifetime

20 years?

MLC lifetime

8PB

15PB

Write amplification

Flash writes more than application

Experiment – tpcc-mysql

Virident FlashMAX 1400

Write amplification: 1.143

1125.65 GiB writes per hour

Lifetime: **1.52** years

Flash for MySQL

When Flash helps

Low-latency requirement

Joins, large tables, mixed workloads, replication

High throughput workloads

High concurrency workloads

Most important decision

MySQL version

MySQL 5.1 with builtin InnoDB

Not good

You need

Multiple IO threads

Async

Choices

Percona Server 5.5

MySQL 5.5

Percona Server 5.1

Benchmarks again

Percona Server 5.5

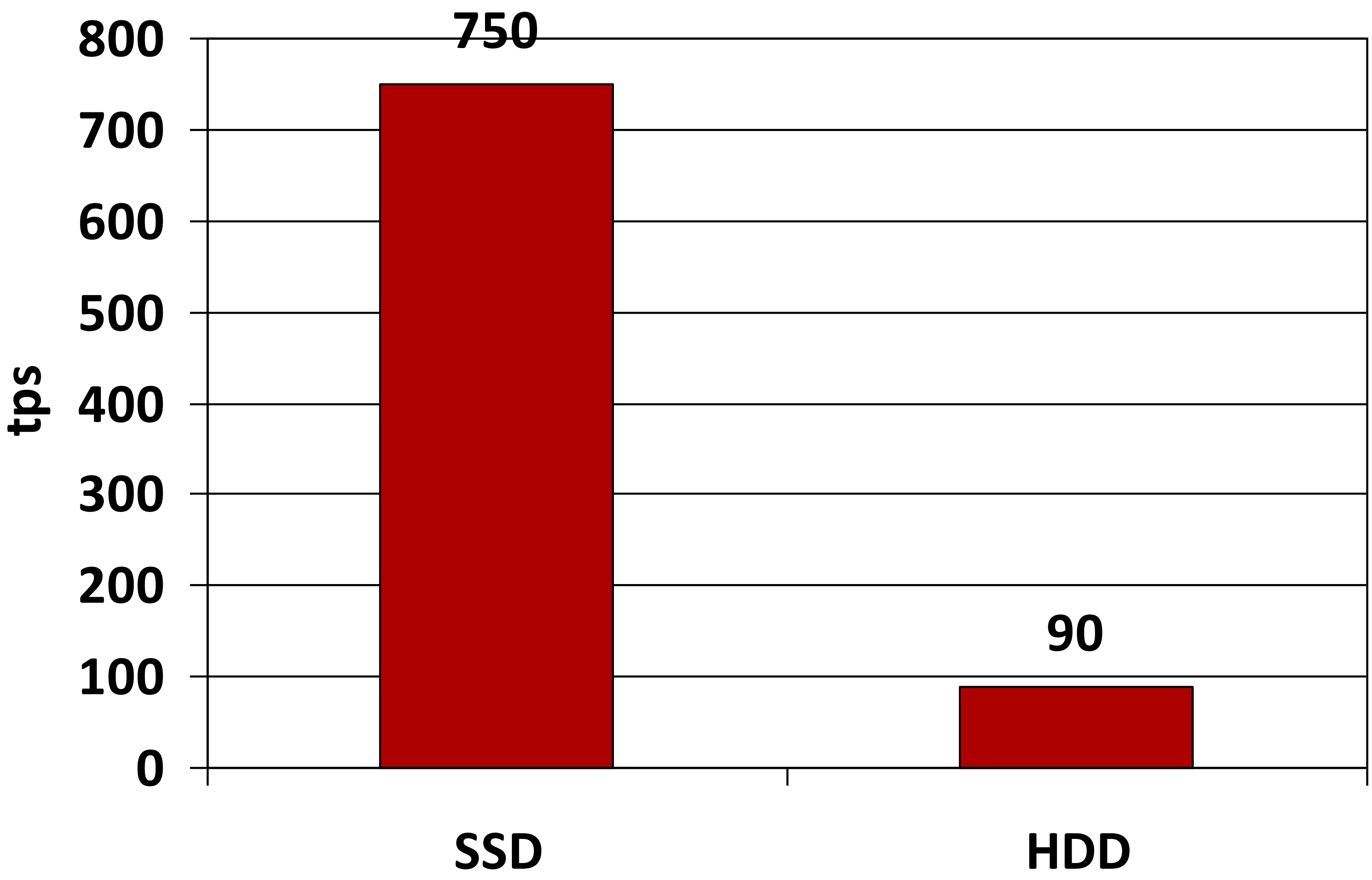
4x STEC MACH16 RAID10

- LSI 9211
- LSI 9260 – with cache

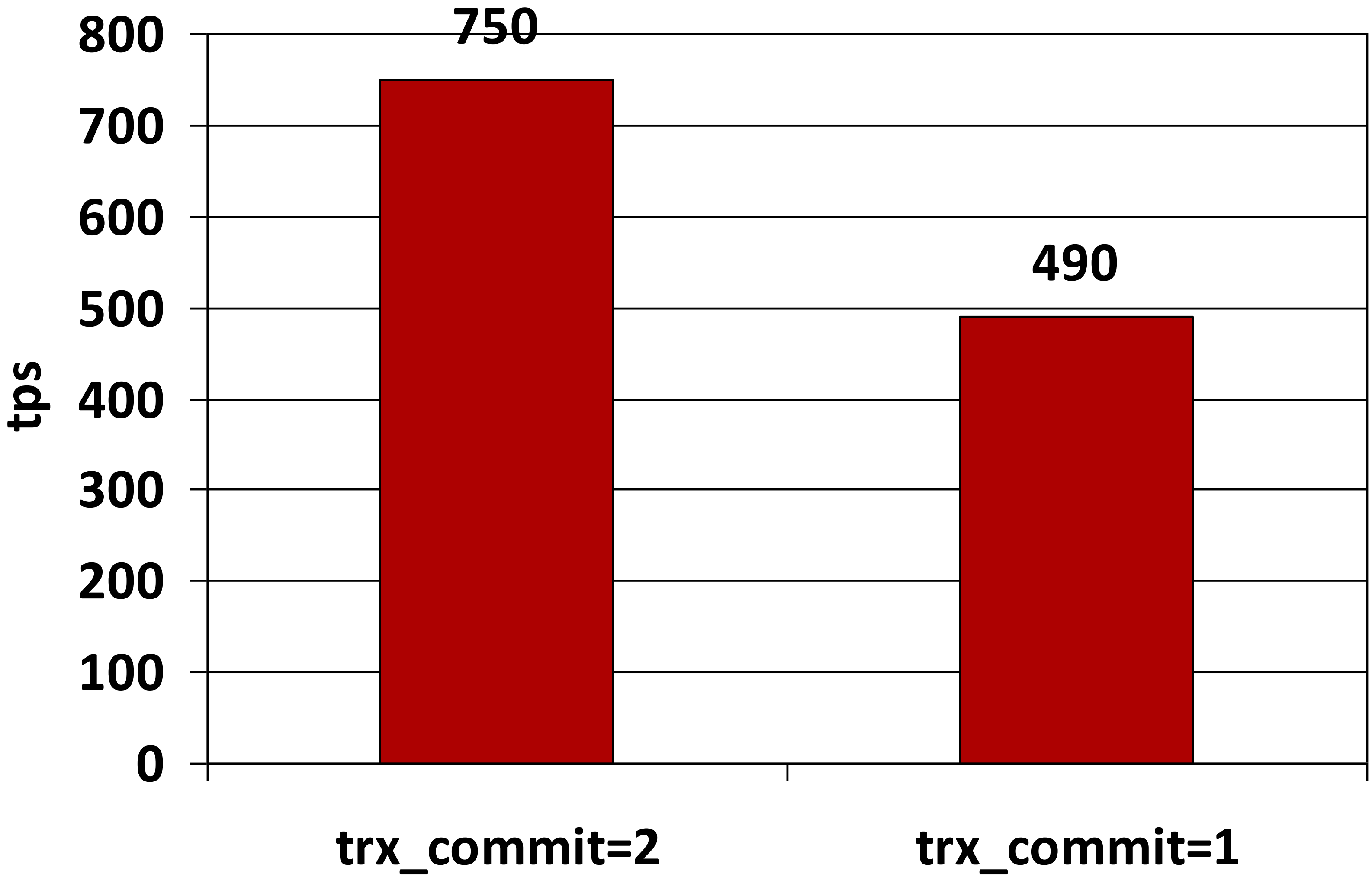
Sysbench oltp

- 100GB database, 50GB memory

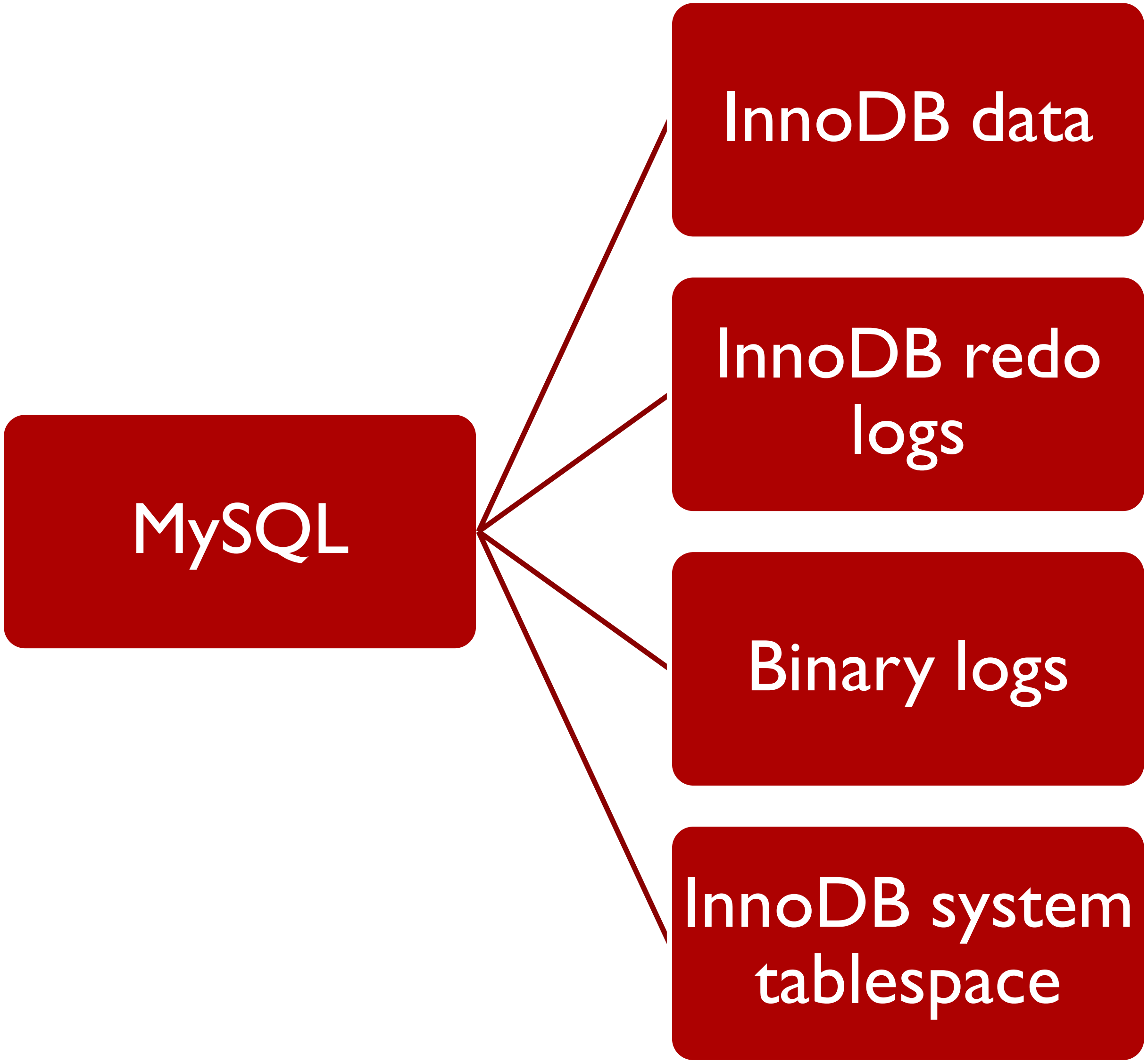
STEC SSD vs HDD: 8x gain



STEC innodb_flush_log_at_trx_commit



MySQL IO workloads



MySQL

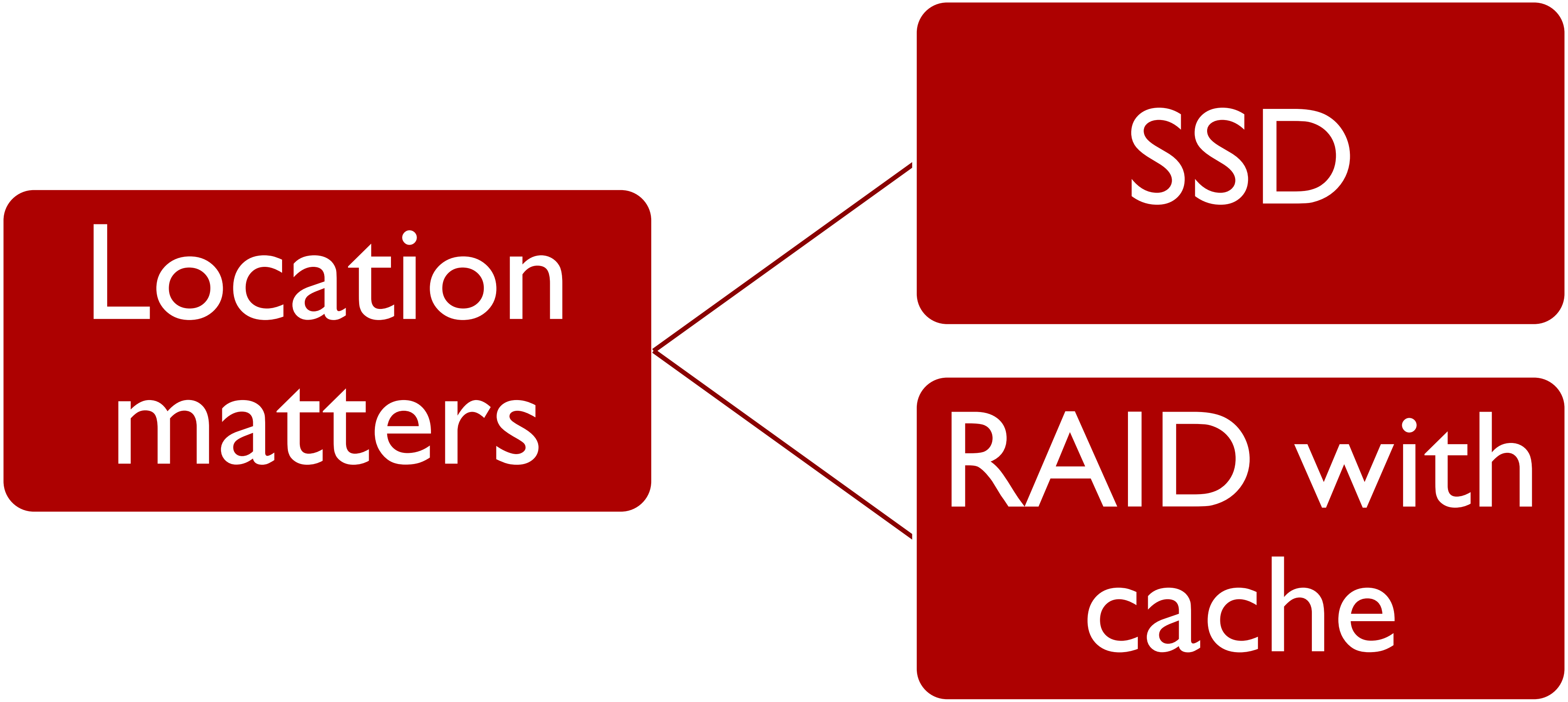
InnoDB data

InnoDB redo logs

Binary logs

InnoDB system tablespace

innodb_flush_log_at_trx_commit=1

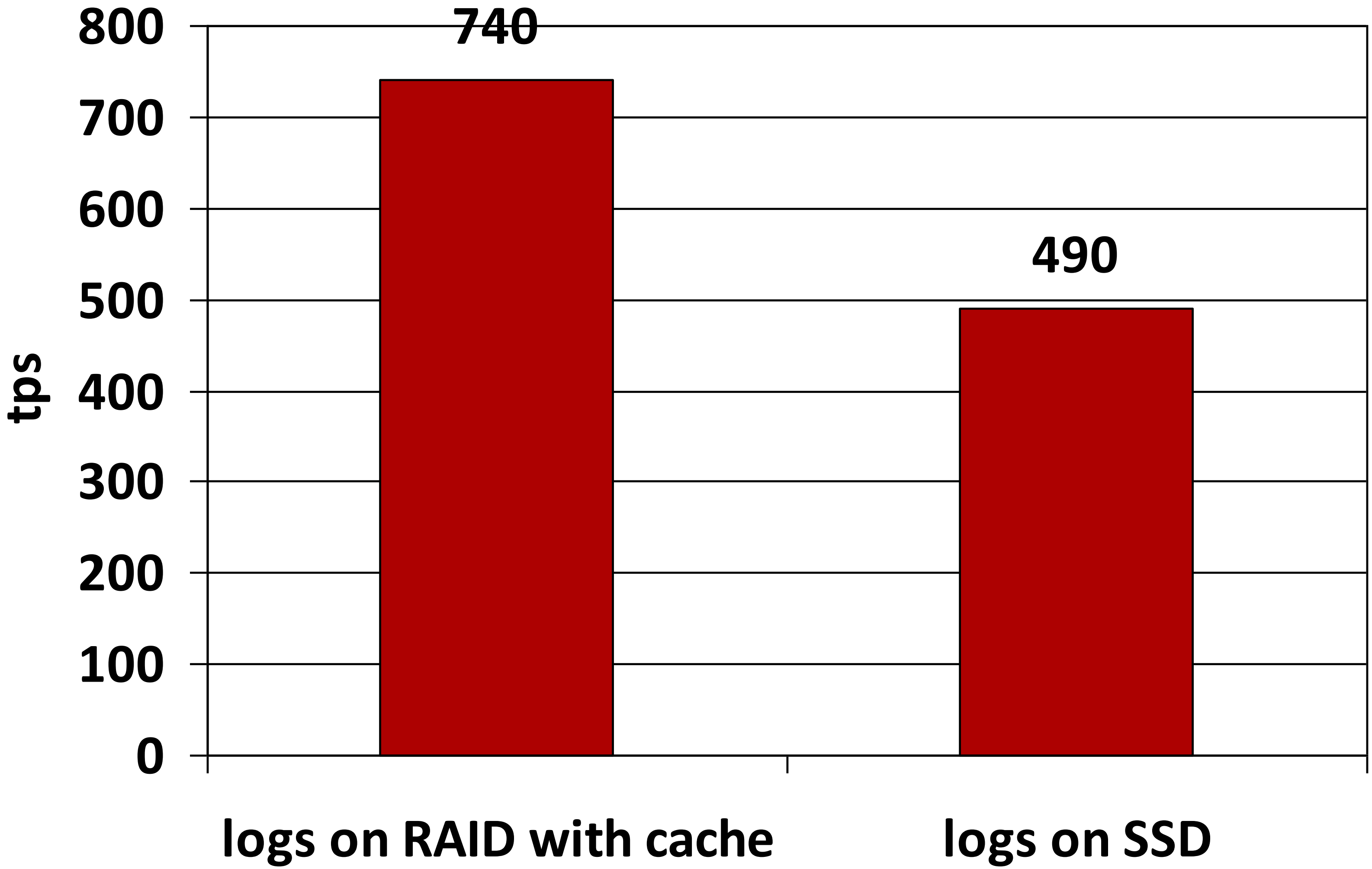


Location matters

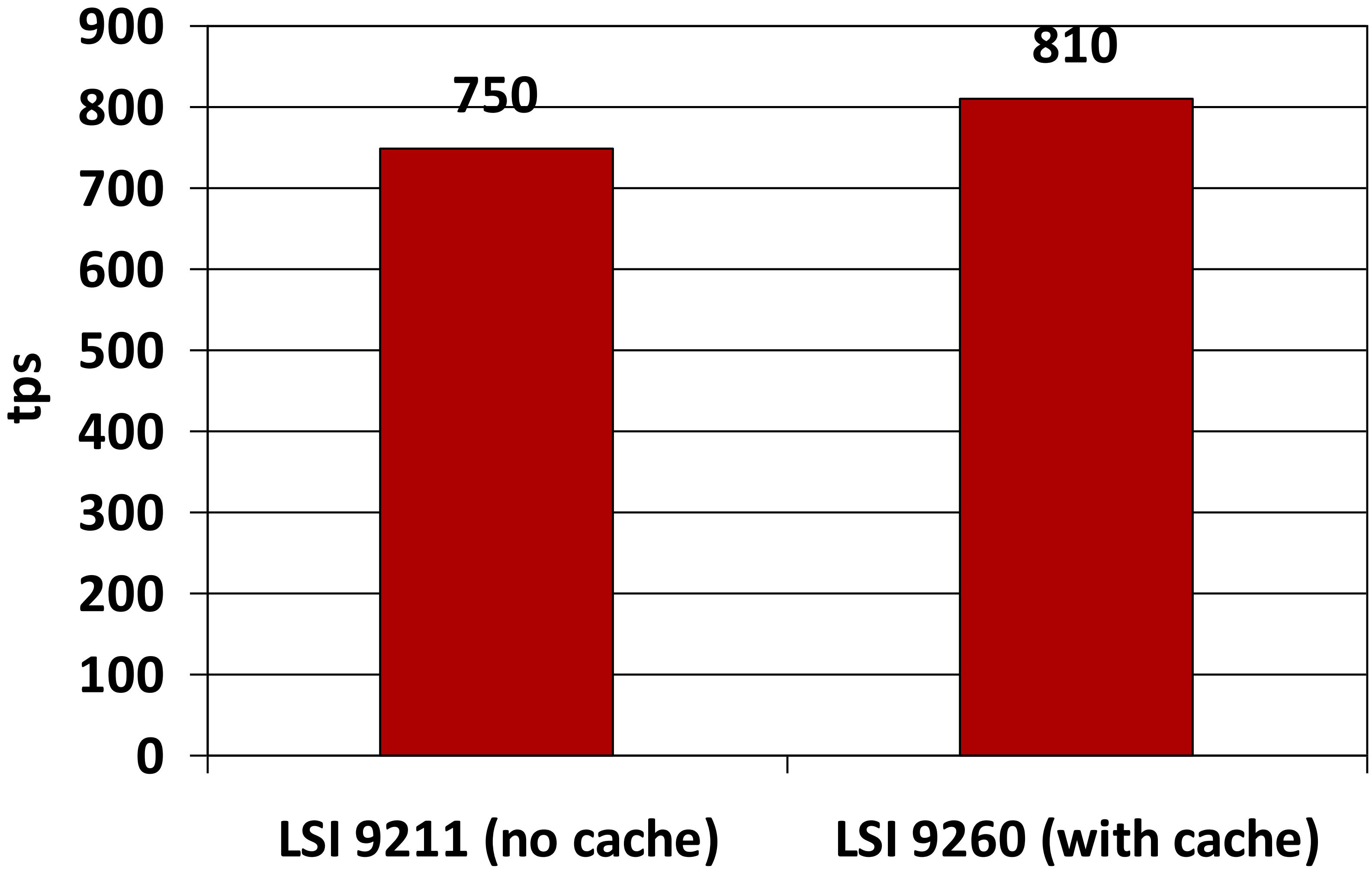
SSD

RAID with cache

STEC: InnoDB log location

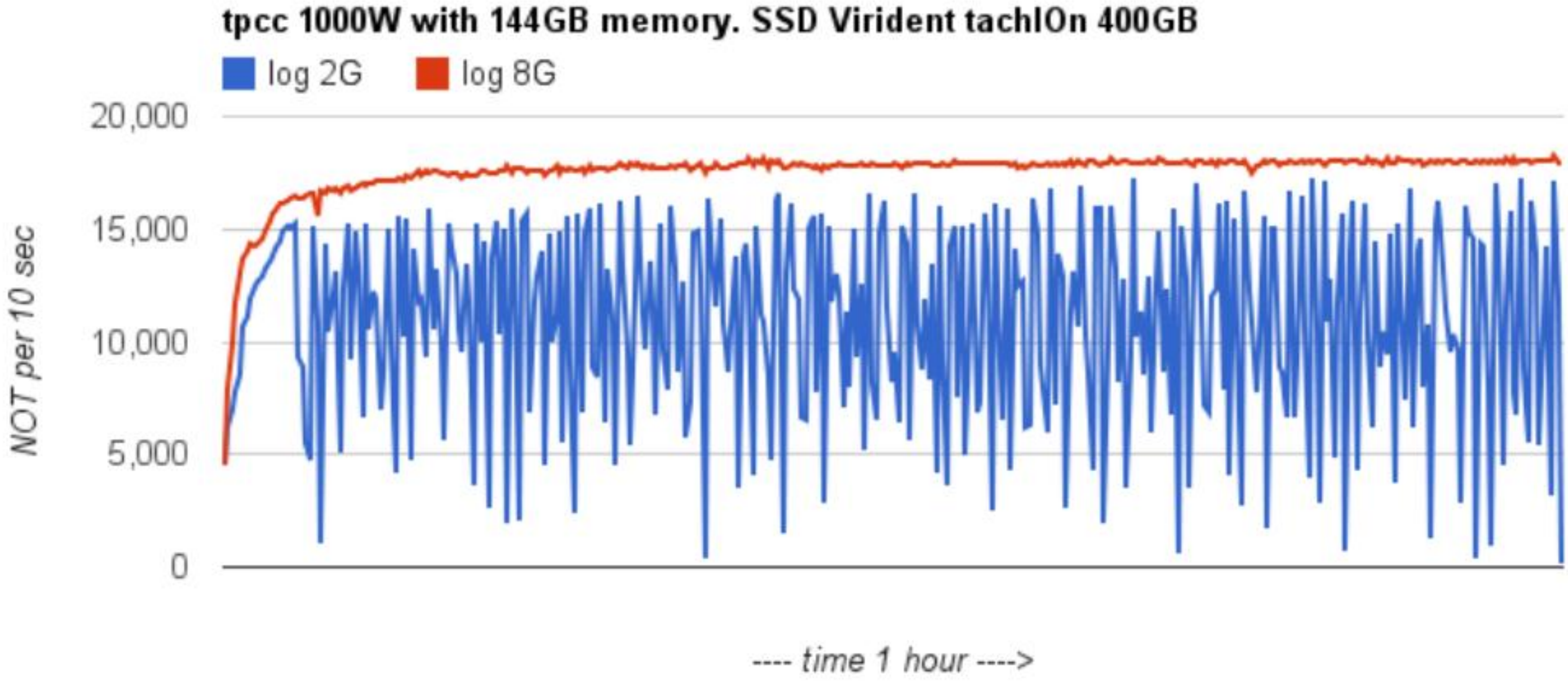


STEC – RAID card



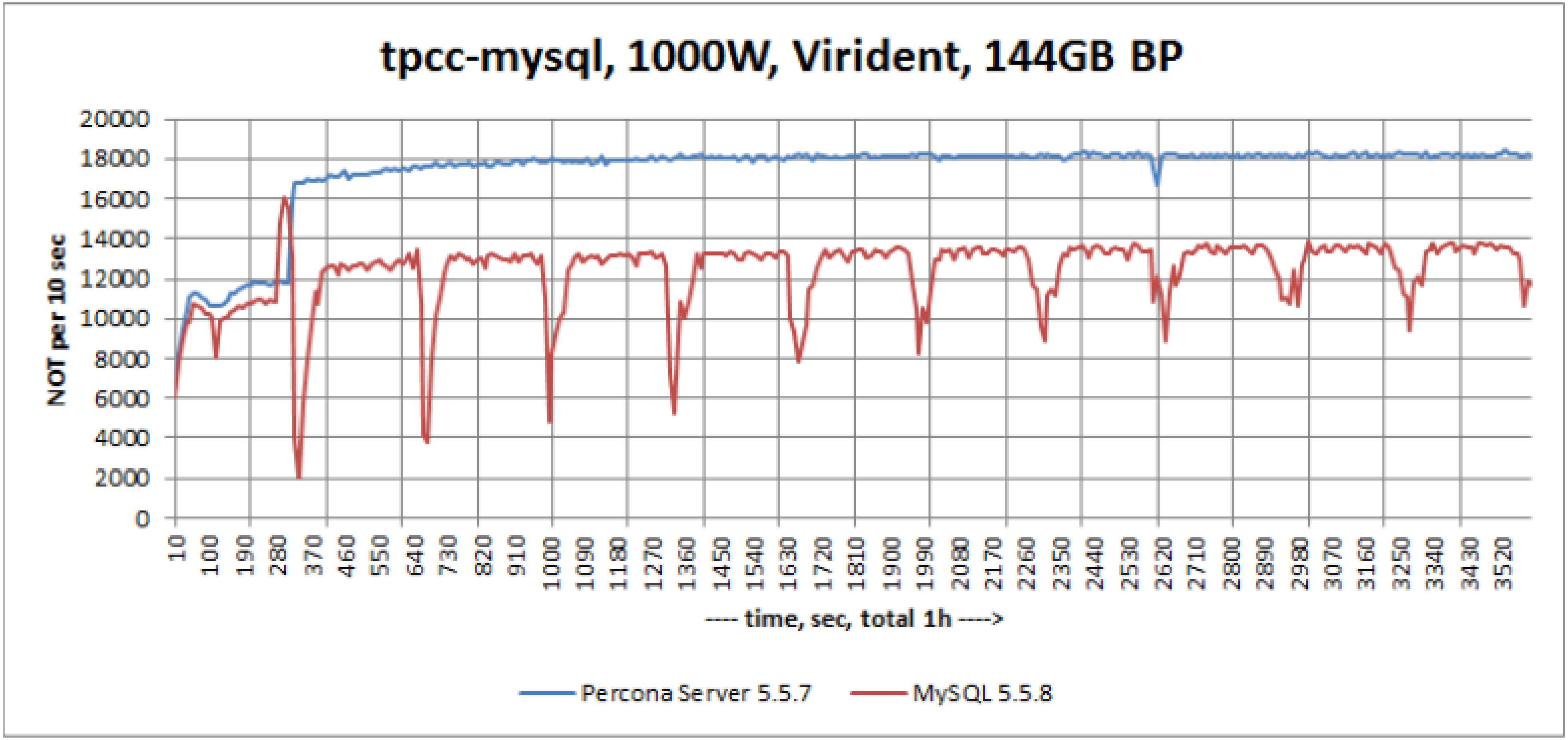
Log size matters

Big log file 8GB (Percona Server)

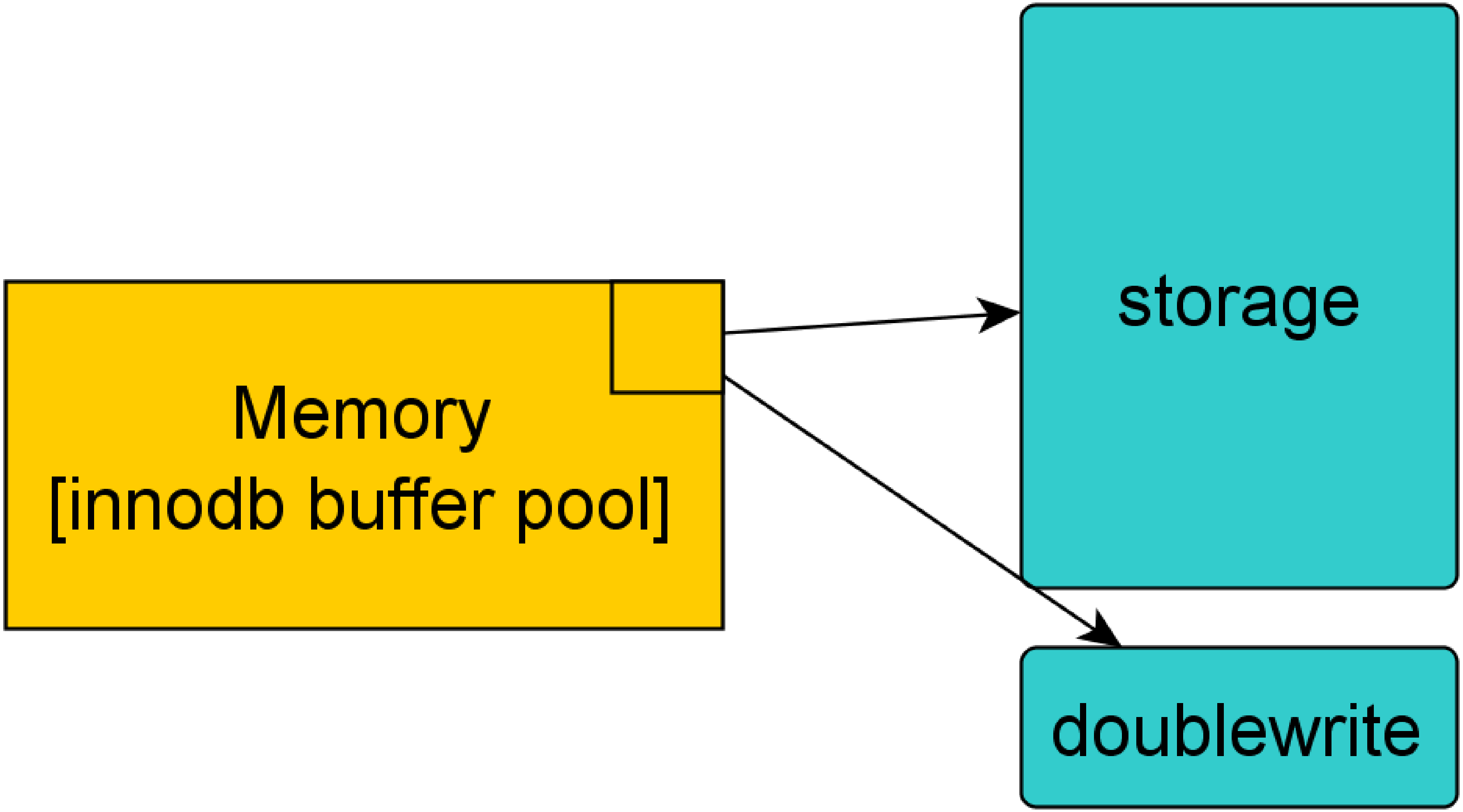


Flushing algorithm is important

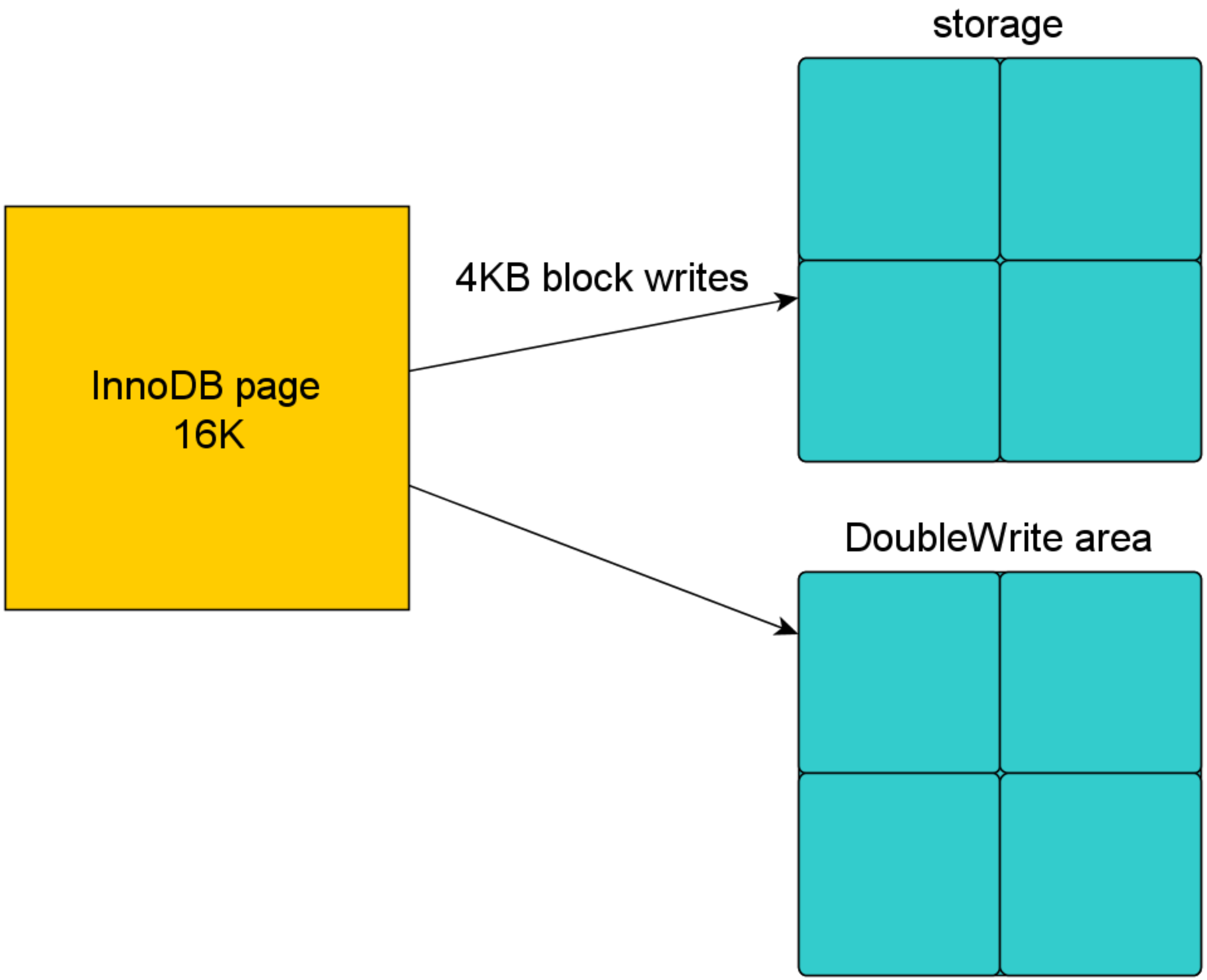
Innodb_adaptive_checkpoint=keep_average (Percona Server)



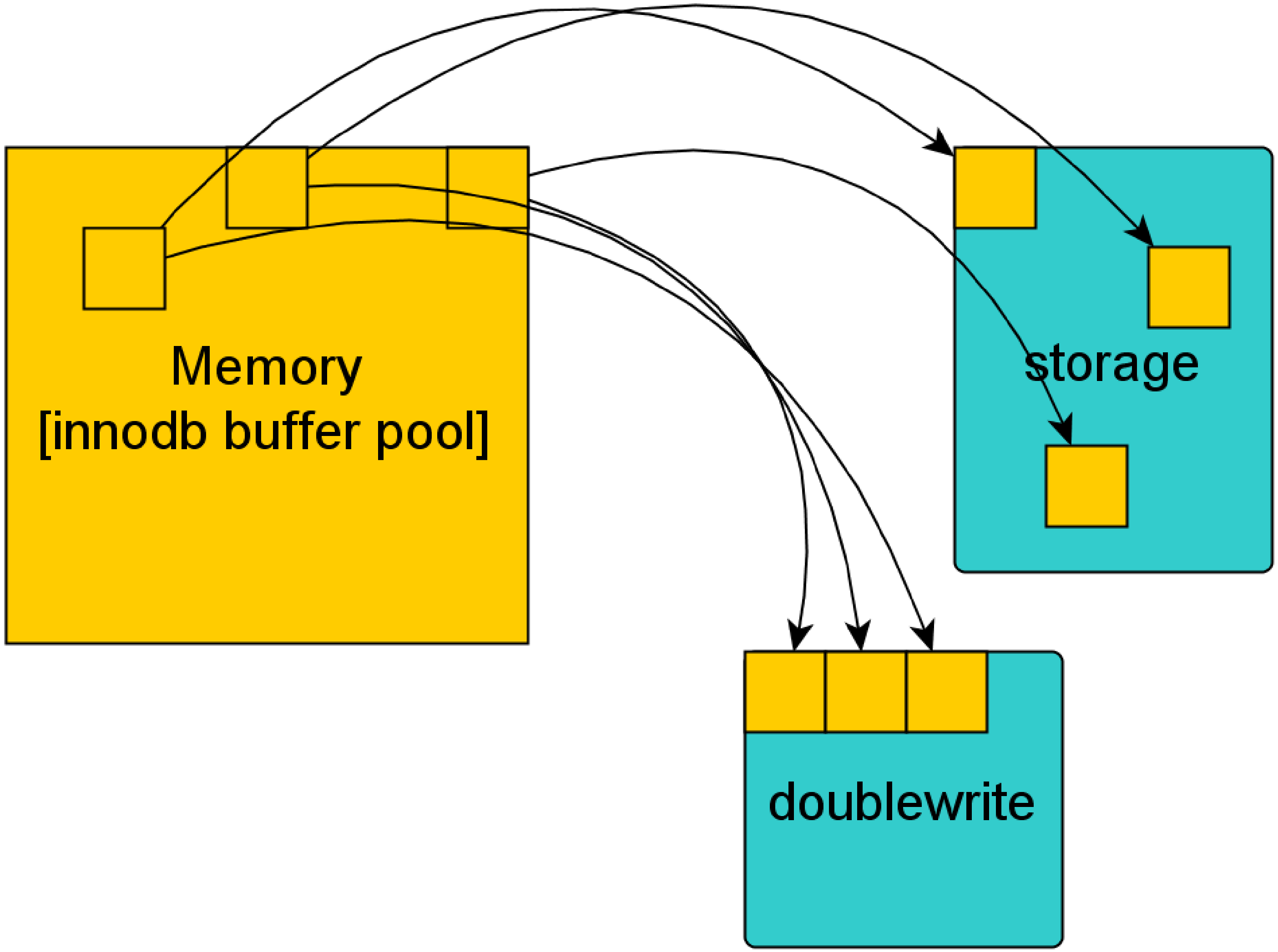
DoubleWrite area is important



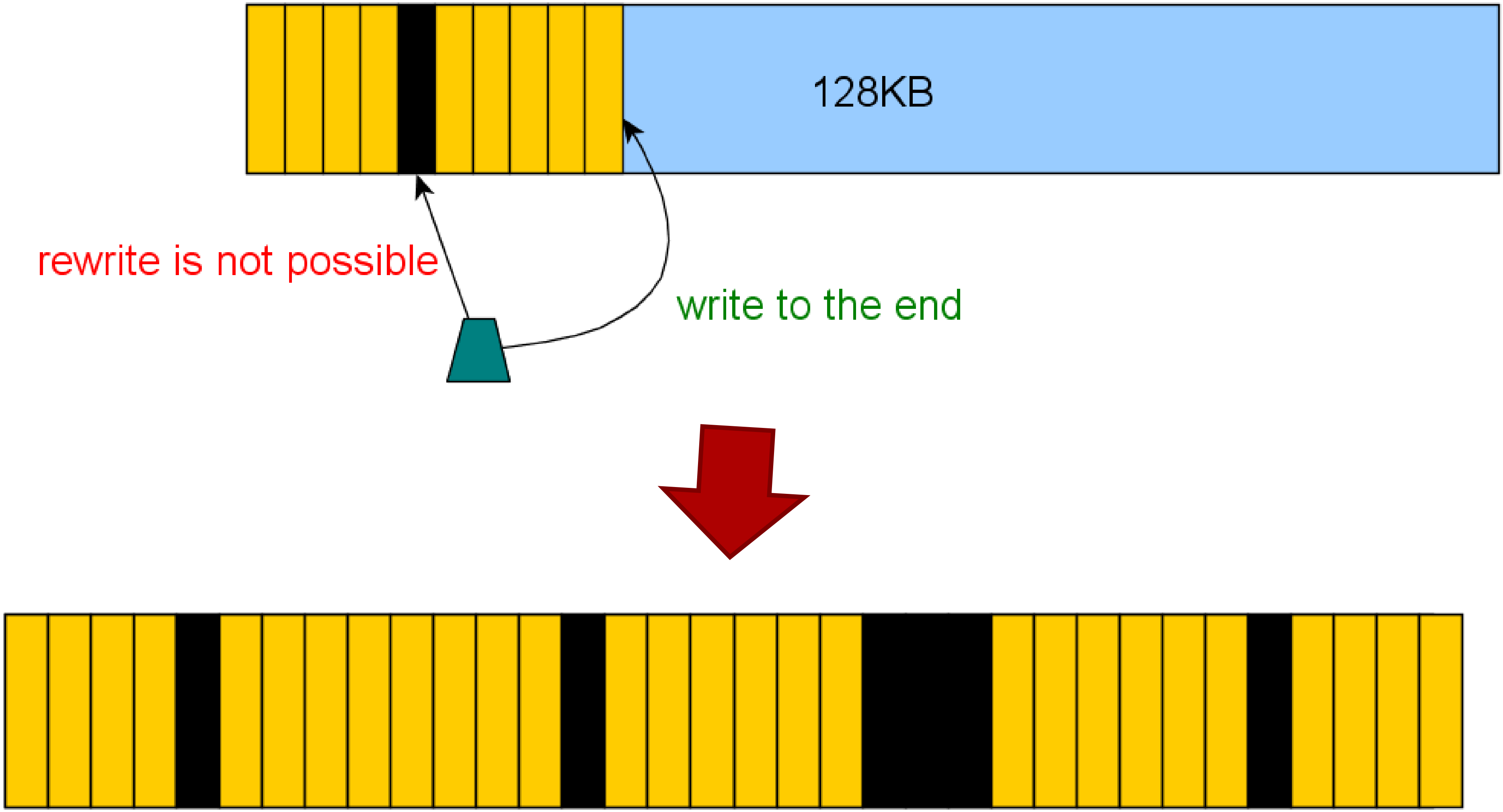
DoubleWrite is a protection



DoubleWrite is rewriting the same area



Rewrites of the same area



Consider moving doublewrite

- Innodb_doublewrite_file =
  - Percona Server
- ibdata1
  - For general MySQL

Fusion-io to support atomic writes

1.5x performance improvement

Misc

innodb_flush_neighbor_pages= ON | OFF

innodb_log_block_size = 512 | 4096

Misc filesystem

```
Mkfs.xfs –s size=4096
```

```
Mount –o nobarrier
```

Is Flash expensive?

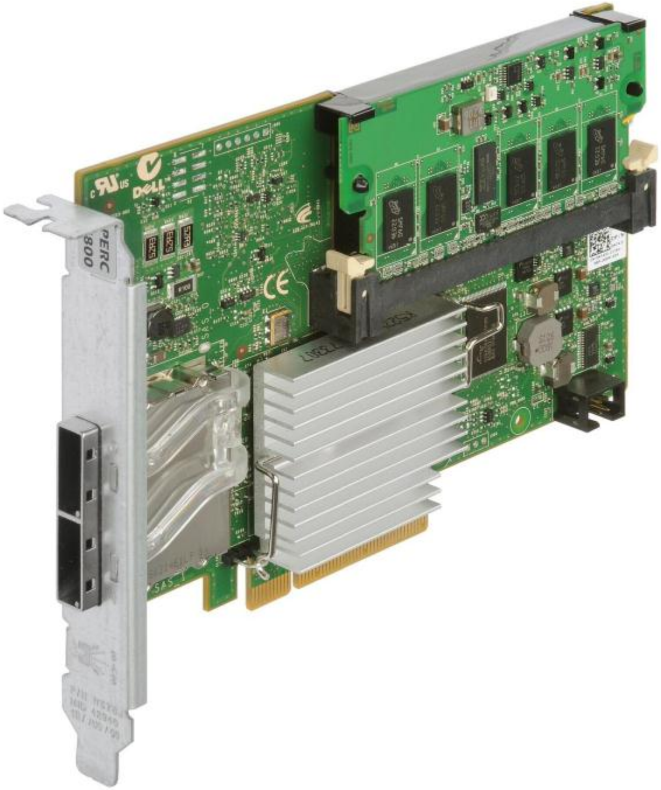
Consolidation

Power savings

New Relic

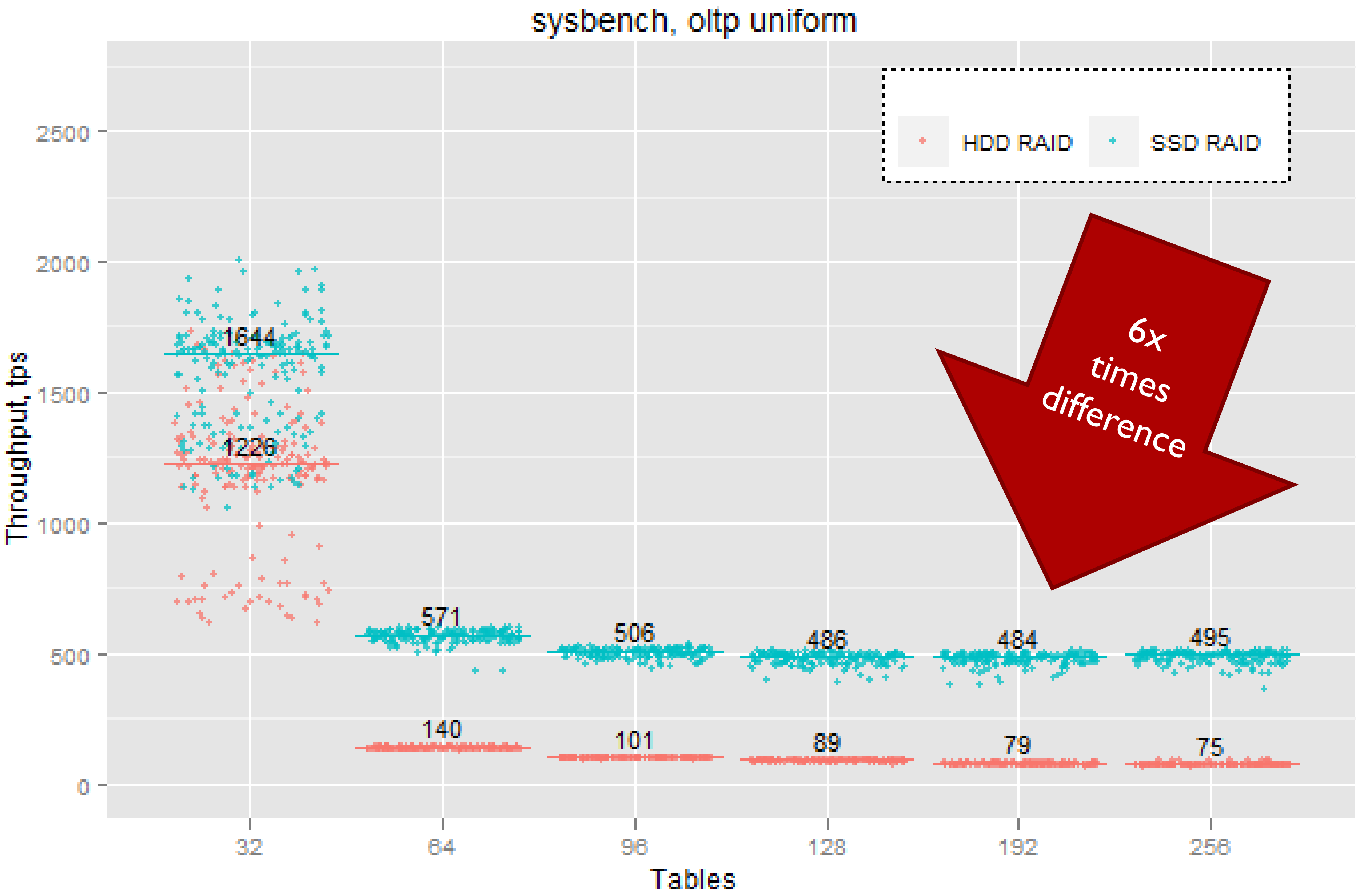
Dell PowerEdge R610

Dell PowerVault MD1220



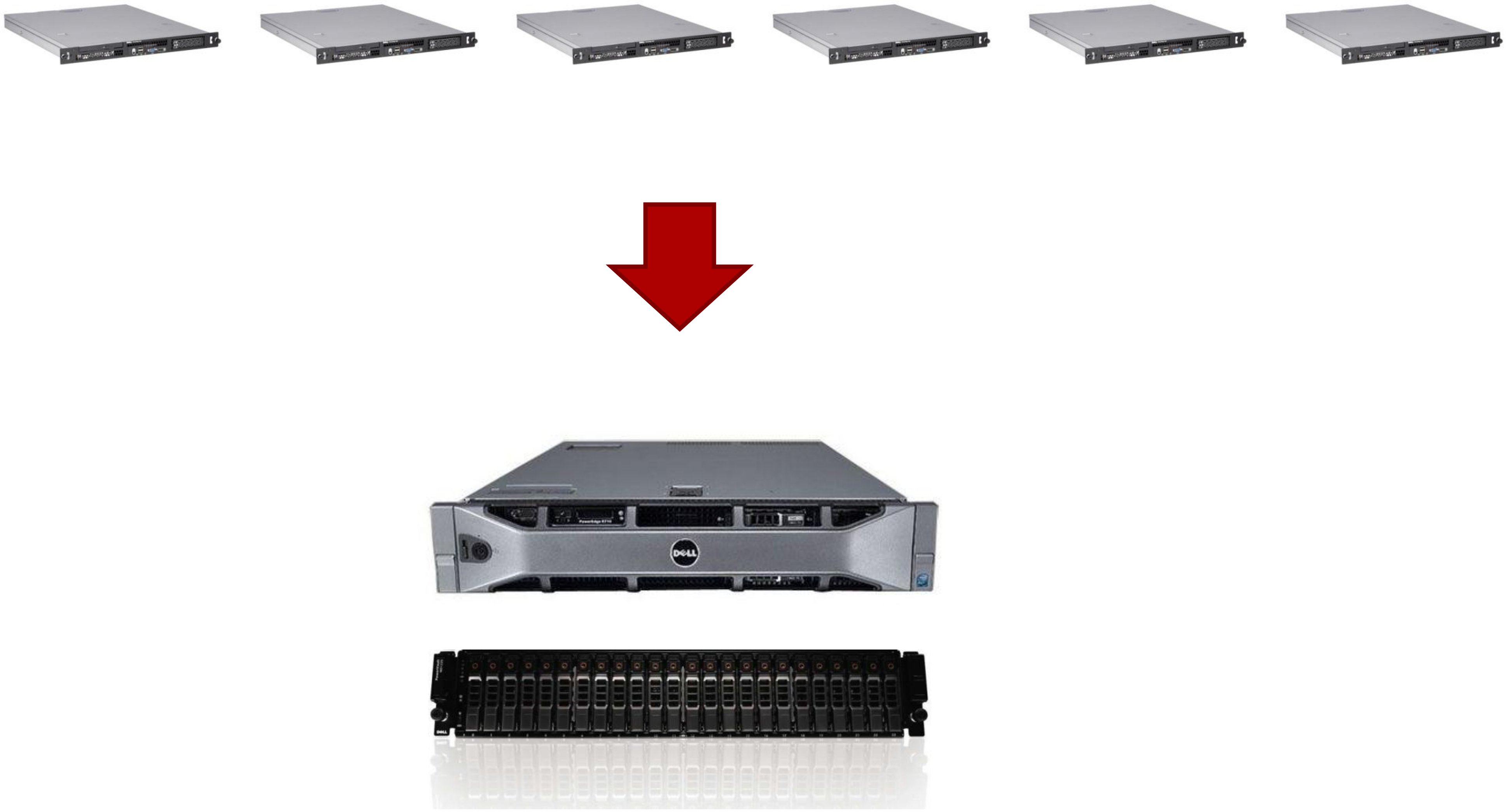
Perc H800

RAID5
11 Intel 320 SSD 600GB



6x times difference

Scale Up, not Scale Out



“Flash made everything faster, but more confusing”

Pictures credits

- http://www.sunrainet.com/hdd-vs-ssd-speed-test-video-windows-7-boot-up.html
- http://blog.familytreemagazine.com/insider/content/binary/datacenter-2.jpg

Thank you!

Questions ?

Flash is exciting!

vadim@percona.com