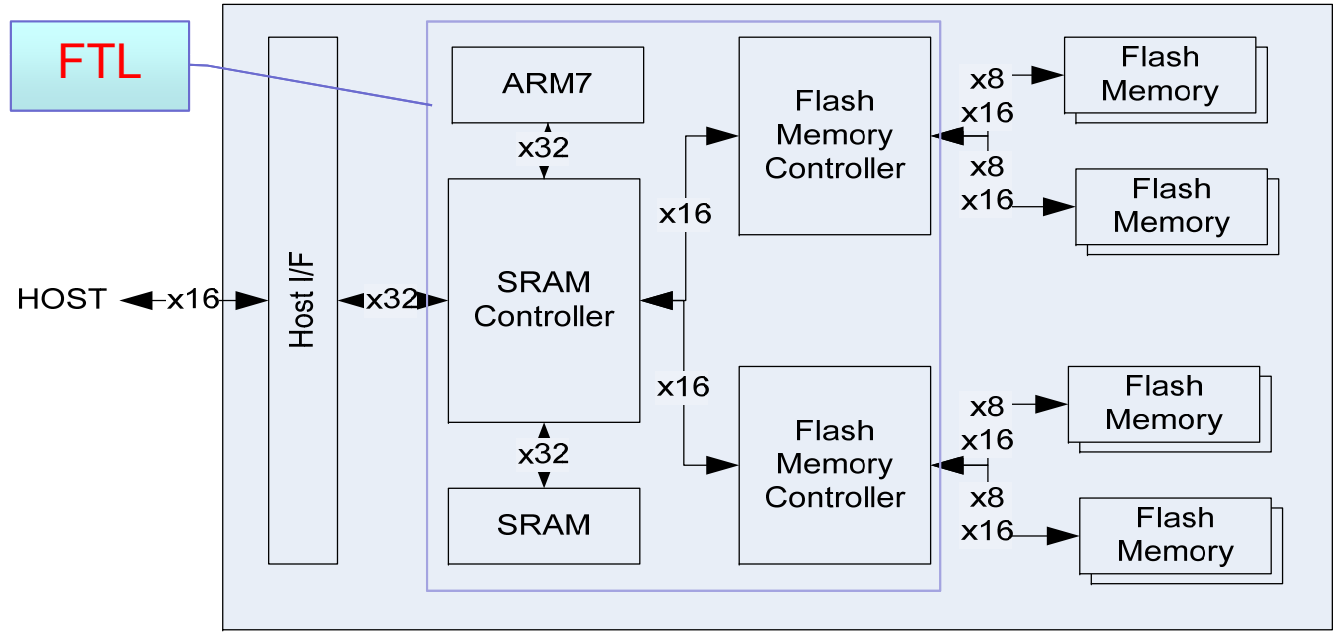
Using Flash SSDs as Primary Database Storage

TECHNISCHE UNIVERSITÄT DARMSTADT

Robert Gottstein,
Ilia Petrov,
Guillermo G. Almeida,
Todor Ivanov,
Alex Buchmann
{lastname}@dvs.tu-darmstadt.de

flashyDB DVS

Flash SSDs, X25-E, ioDriveDuo



Specification

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Specification – Intel X25-E 64GB, SLC**
- Seq. Read/Write: 250 / 170 MB/s
- Read/Write IOPS (4K): 35 000 / 3 300
- Latency Read/Write (4K): 0.075/0.085 ms
- Price: € 650

- **Specification:** Savvio 146GB,15k
- Seq. Read / Write: 160 MB/s
- Read/Write IOPS: 350 / 300
- Latency Read/Write: 3.2 / 3.5 ms
- Price: € 180

10x

20x

Flash vs Magnetic Storage

TECHNISCHE UNIVERSITÄT DARMSTADT

10x ...
20x

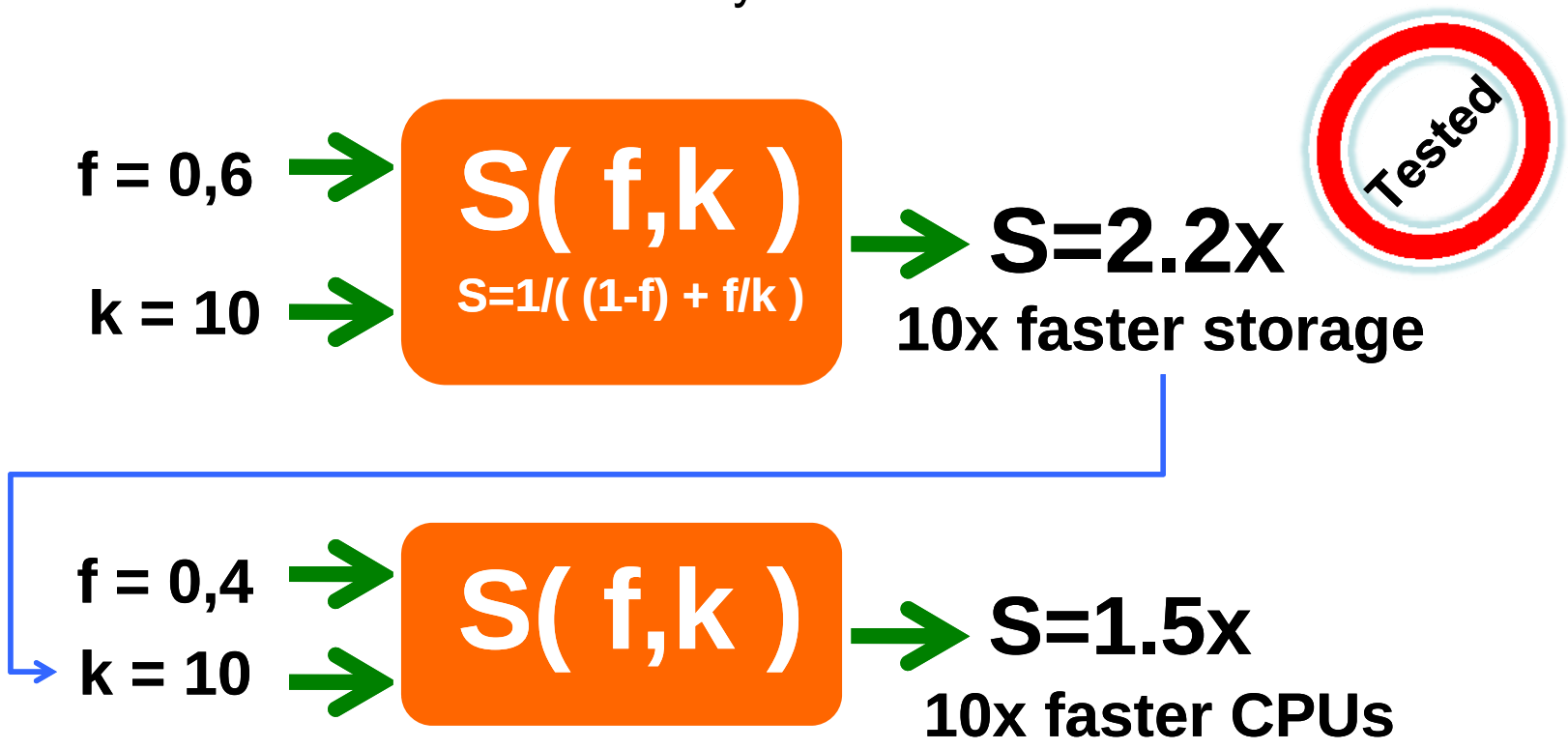
> 1000x

- **IoFusion ioDrive Duo**
- Seq. Read/Write: 1.5 / 1.4 GB/s
- Read/Write IOPS (4K): 130 000 / 80 000
- Latency Read/Write (4K): 0.025/0.035 ms
- Price: approx. € 6000

Amdahl's Law – Speedup [1]

TECHNISCHE UNIVERSITÄT DARMSTADT

- An OLTP database performs IO approx. 60% of the time [Patterson]
- 10x faster CPUs or 10x faster IO-Subsystem?

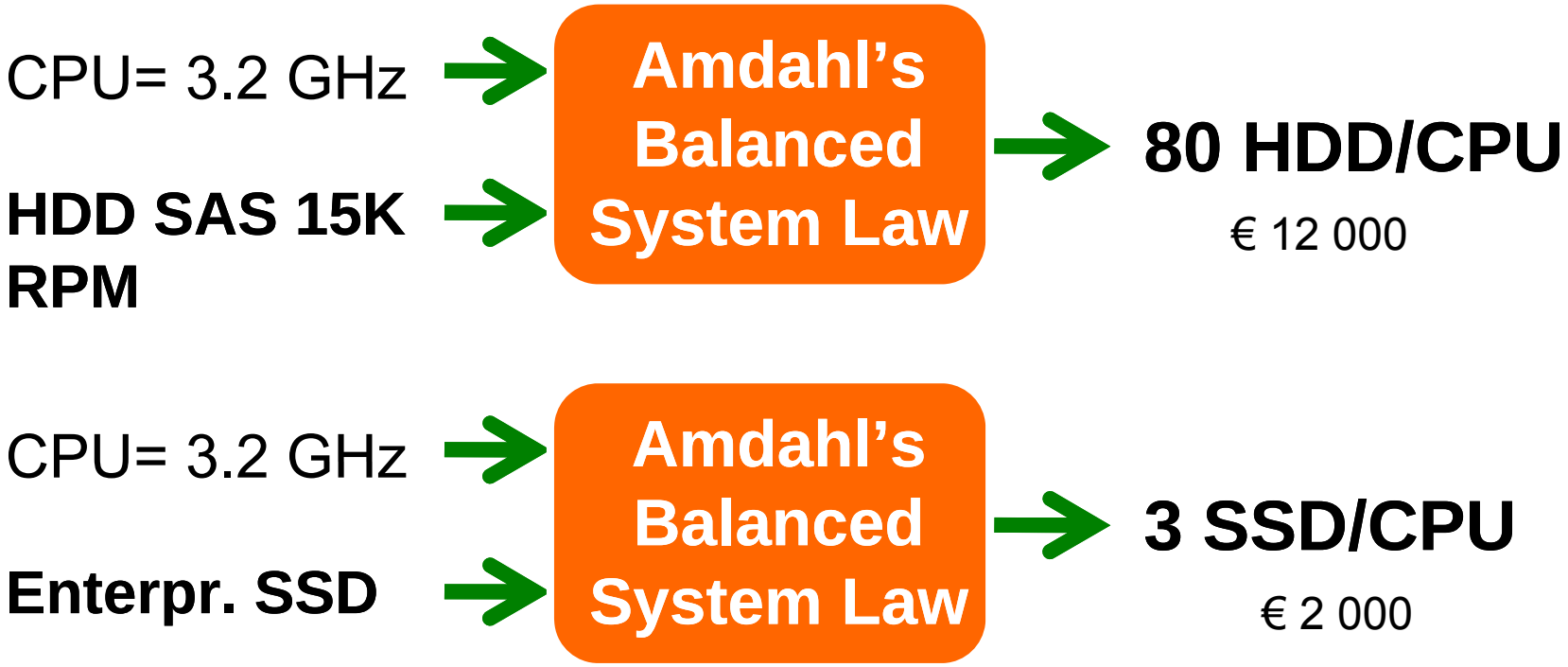


[1] Amdahl, Gene. "Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities". In Proc. AFIPS Conference pp.483–485. 1967

Amdahl's Revised Balanced System Law [2]:

TECHNISCHE UNIVERSITÄT DARMSTADT

- A system needs 8 MIPS/MB/s IO
- The instruction rate and IO rate workload dependent → OLTP, CPI=2.1
  - Assume 75% random write, 25% random read, 8KB page size, 3.2 GHz CPU



[2] Jim Gray, Prashant Shenoy, "Rules of Thumb in Data Engineering," In Proc. , ICDE 2000

flashyDB DVS

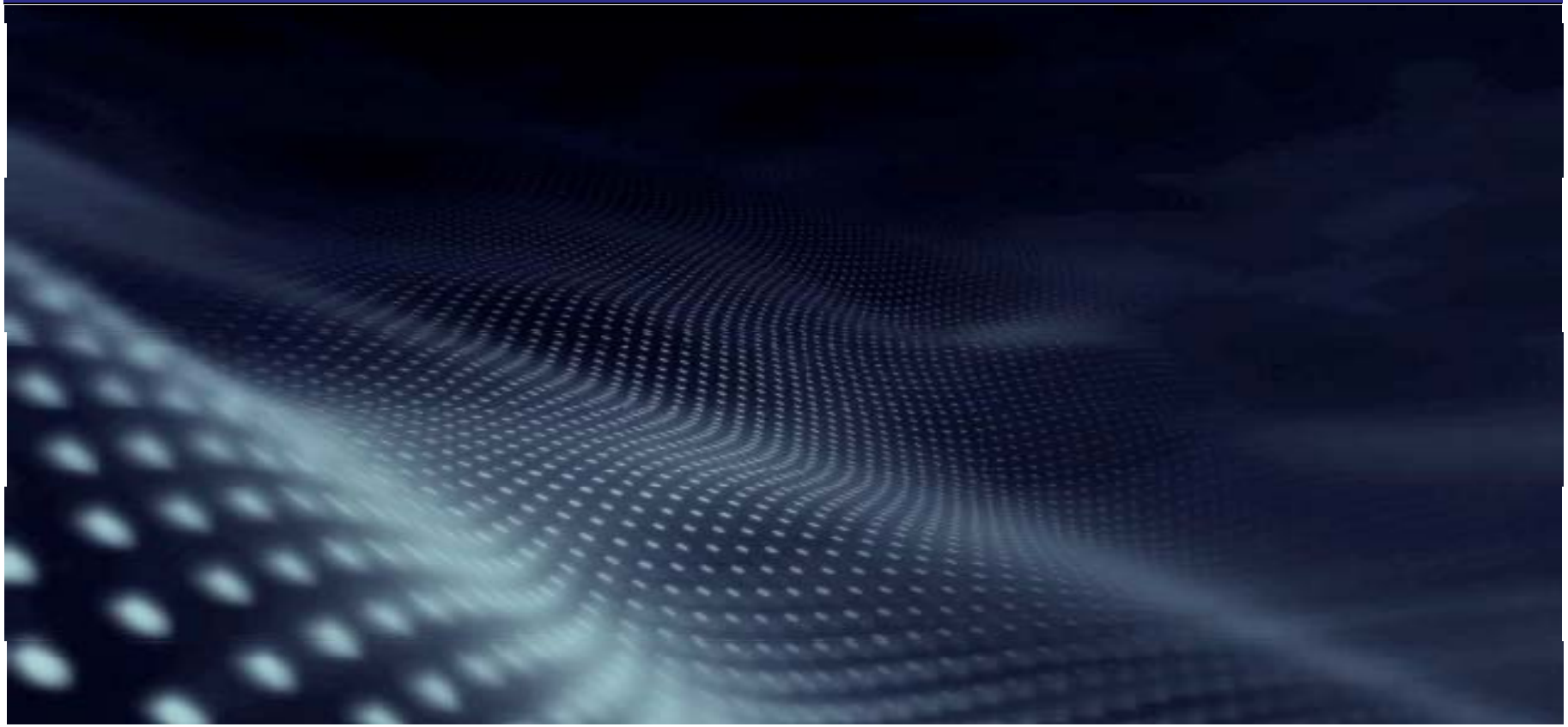
In summary

TECHNISCHE UNIVERSITÄT DARMSTADT

- HDD have reached physical limits
  - Fighting low access density with thousands of HDDs is unreasonable
  - Outdated storage technology
- Data-Intensive Systems are IO-Bound
- Data-Intensive systems built around HDD properties
  - Access Gap / Access Density
  - Larger Buffer Sizes
  - Larger Page Sizes
  - Algorithms optimized for streaming access rather than random access
- SSDs come at the right moment

Flash SSD Characteristics

TECHNISCHE UNIVERSITÄT DARMSTADT



Characteristics

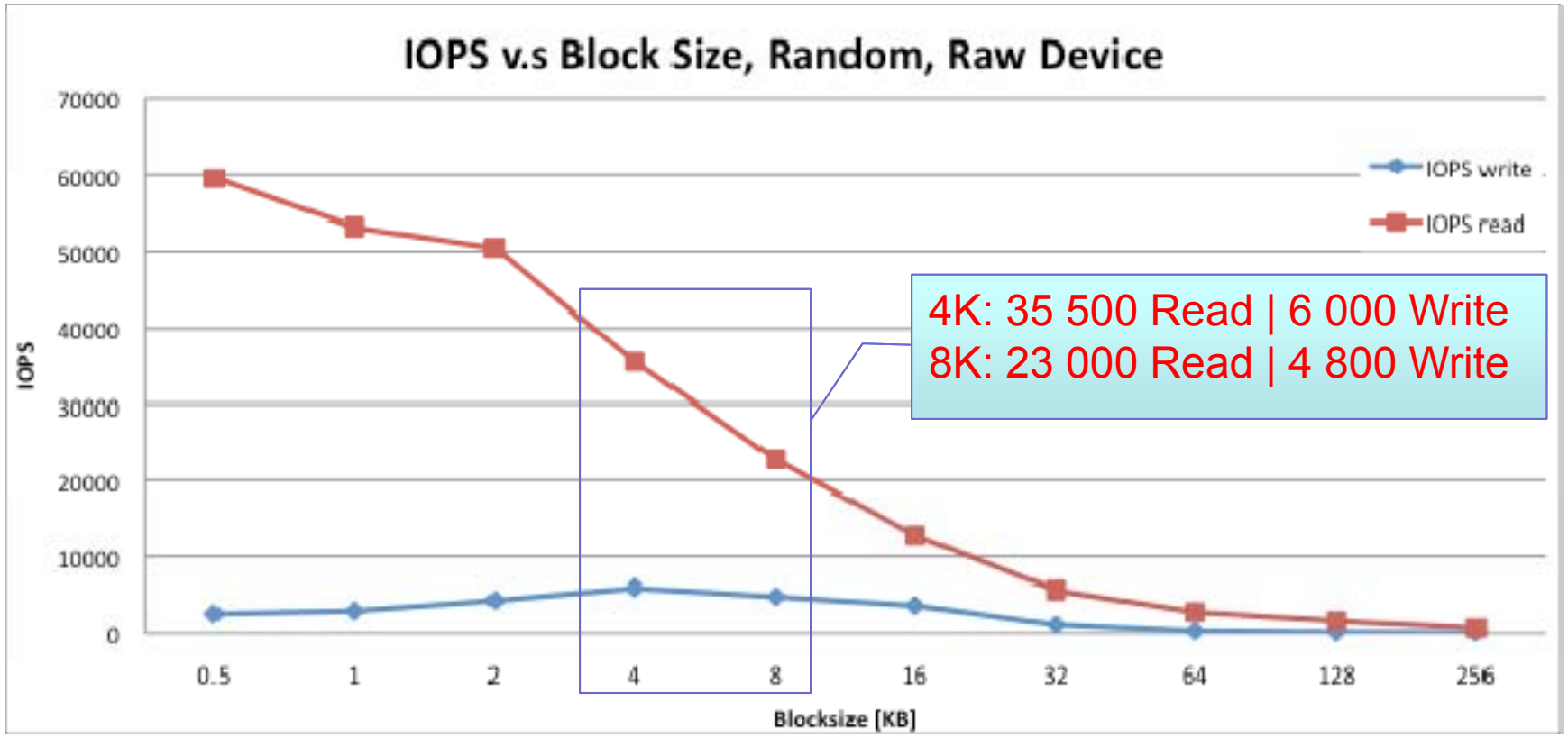
TECHNISCHE UNIVERSITÄT DARMSTADT

- Throughput asymmetry
- Random Throughput
  - Better for small block-sizes
  - Random Writes are an issue
- Very good sequential throughput
  - Still asymmetric
  - Caching
- Very low latency
- Command Queuing and internal parallelism

Random Throughput

TECHNISCHE UNIVERSITÄT DARMSTADT

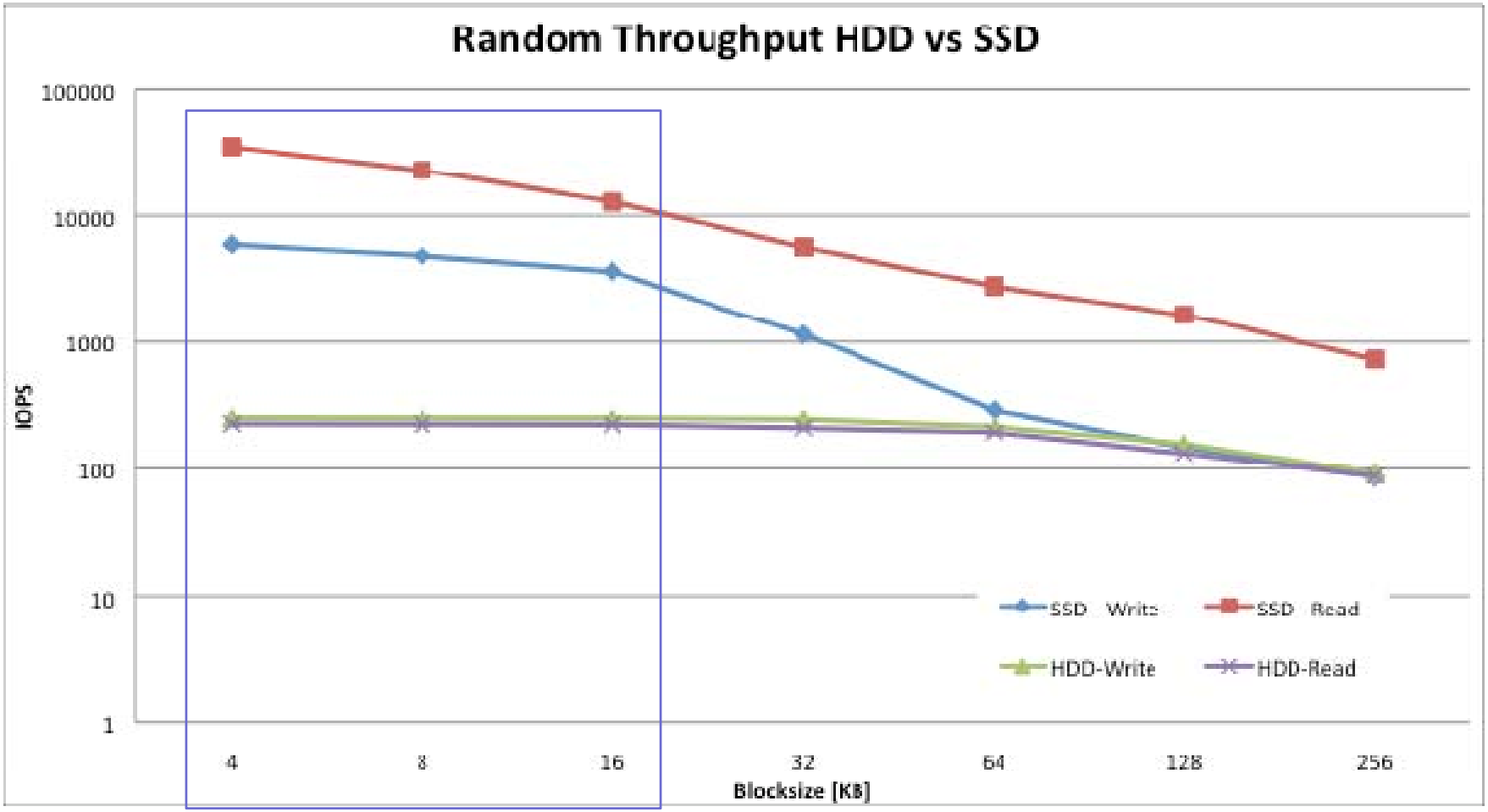
- Random Throughput-Very High
  - Asymmetric: Read vs. Write
  - Up to 10x difference
- Better for small blocksizes:
  - Major weakness of HDD



flashyDB DVS

Random Throughput – SSD and HDD

TECHNISCHE UNIVERSITÄT DARMSTADT

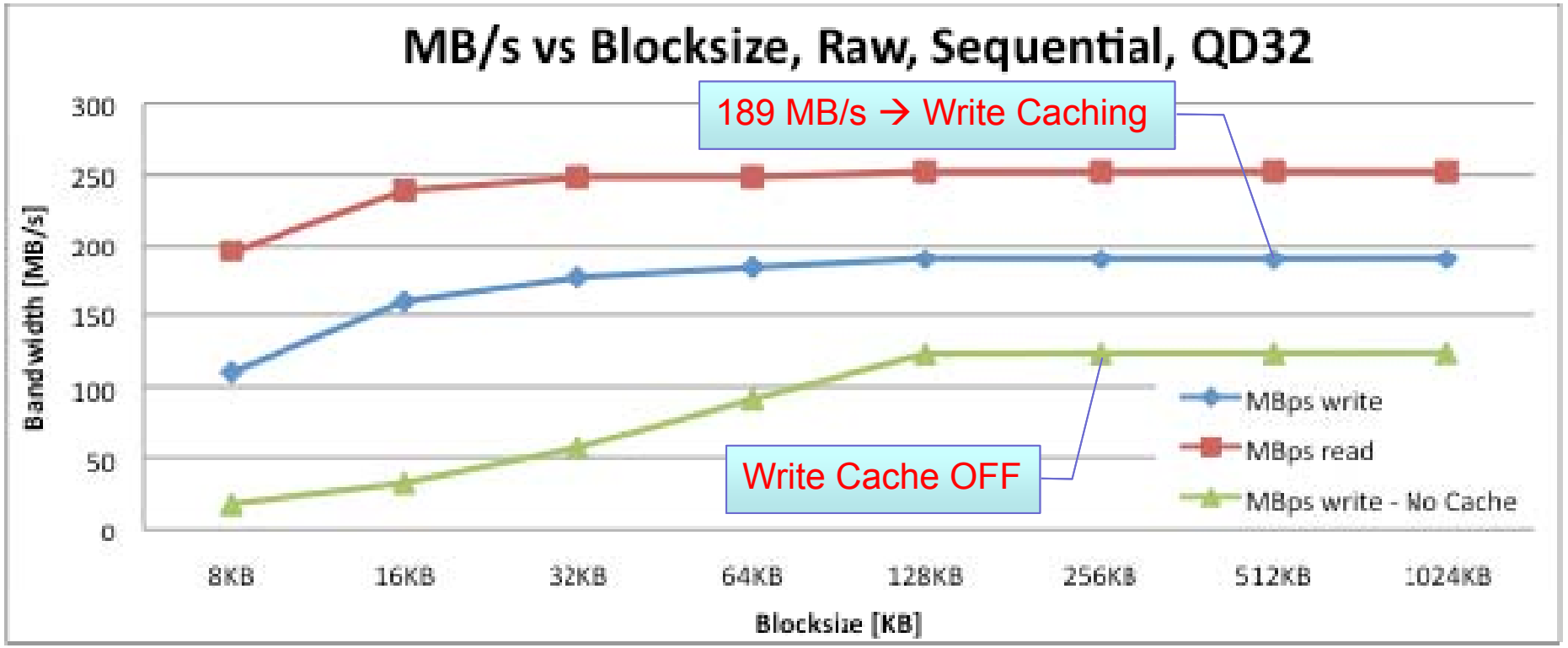


flashyDB DVS

Sequential Throughput

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sequential Bandwidth MB/s
  - Asymmetric
  - >= HDD
- Caching
- Command Queuing



flashyDB DVS

Average Access Time / Latency (AVG)

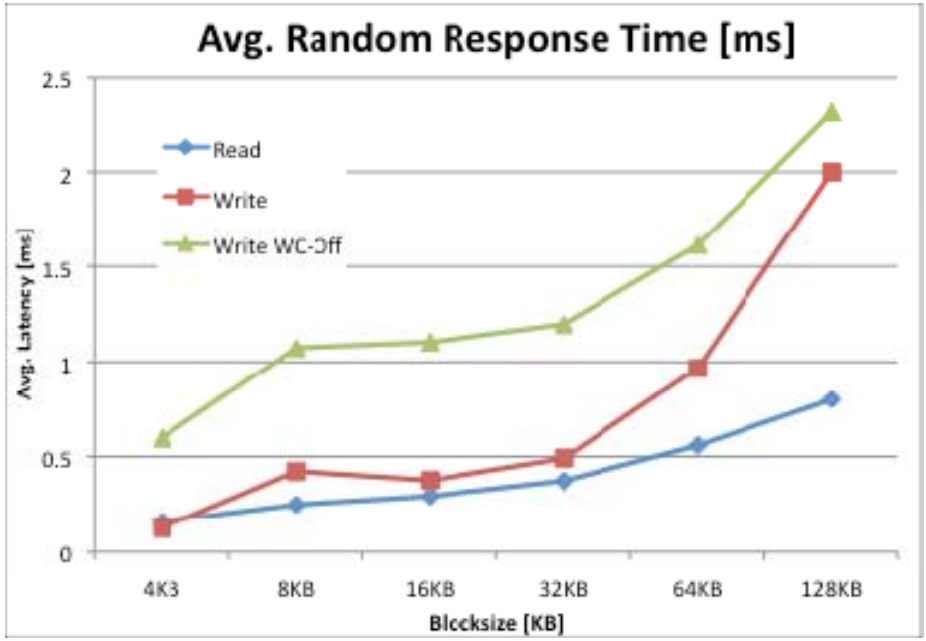
TECHNISCHE UNIVERSITÄT DARMSTADT

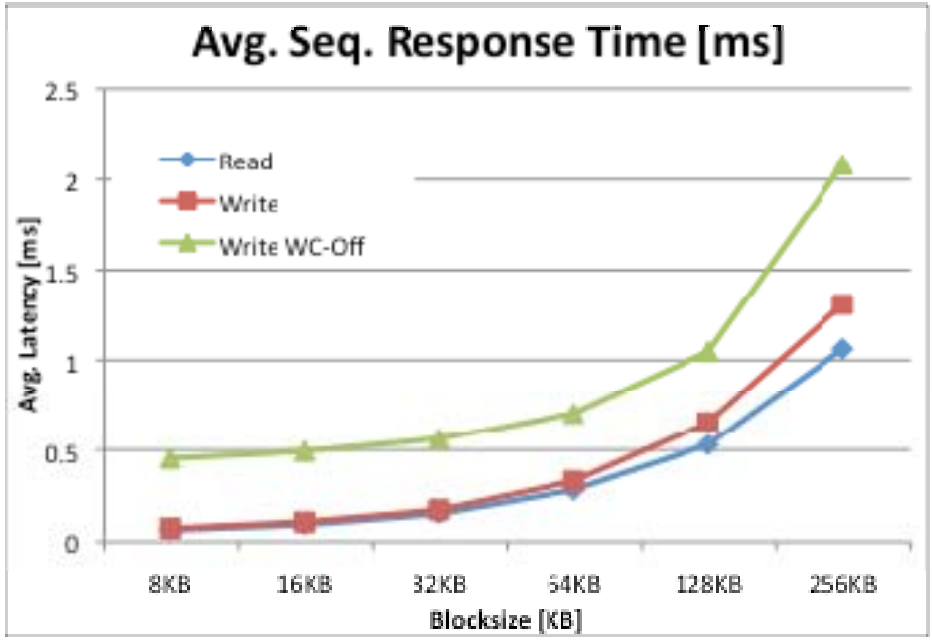
AVG. Latency [ms]

	WC On	WC Off
Seq. Read	0.053	--
Seq.Write	0.059	0.455
Rand. Read	0.167	--
Rand. Write	0.113	0.435

Max Latency[ms]

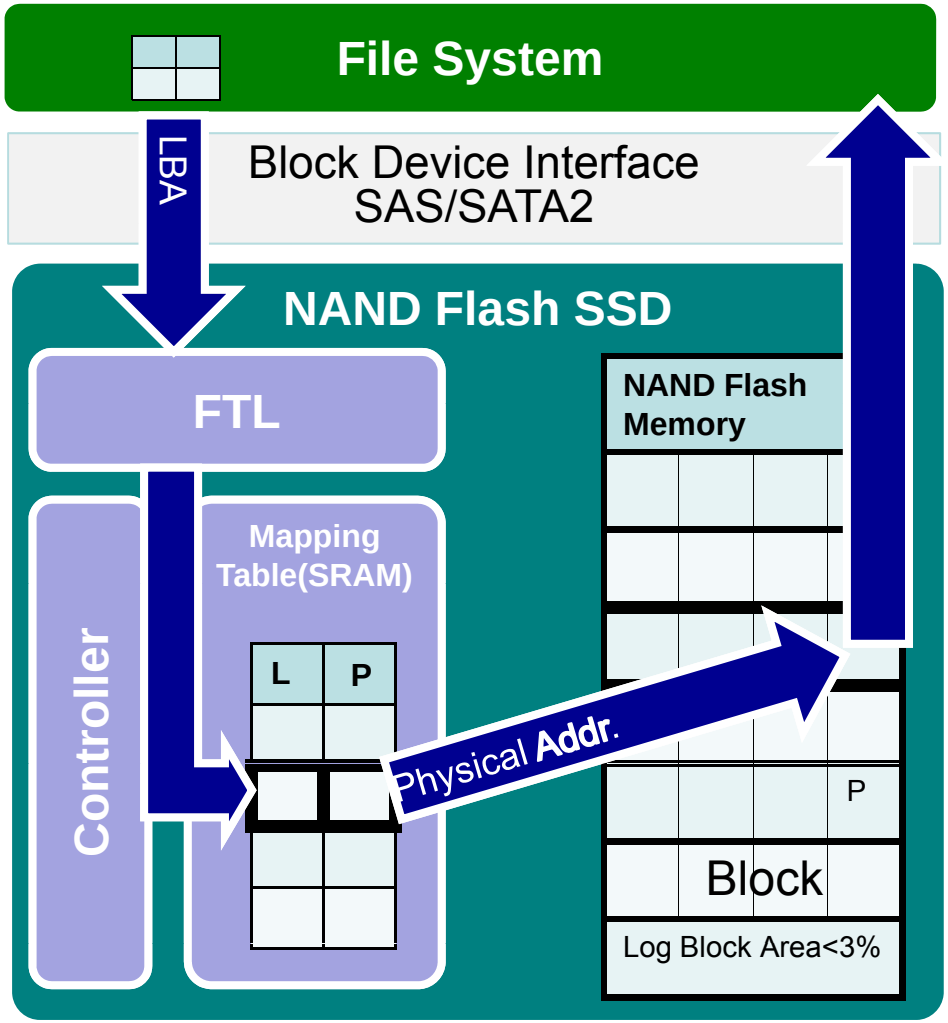
	WC On	WC Off
Seq. Read	12.29	--
Seq.Write	94.82	100.26
Rand. Read	12.41	--
Rand. Write	175.27	100.68





flashyDB DVS

FTL, Address-Mapping



- Block device interface
- Logical Blocks, LBA
- Pages, (Erase)Blocks, Log records
- FTL- Flash Translation Layer

- Background Processes
  - Wear-leveling
  - Garbage collection
  - Metadata synch
  - Log-block merging

- SATA2/SAS – TRIM
  - RAID

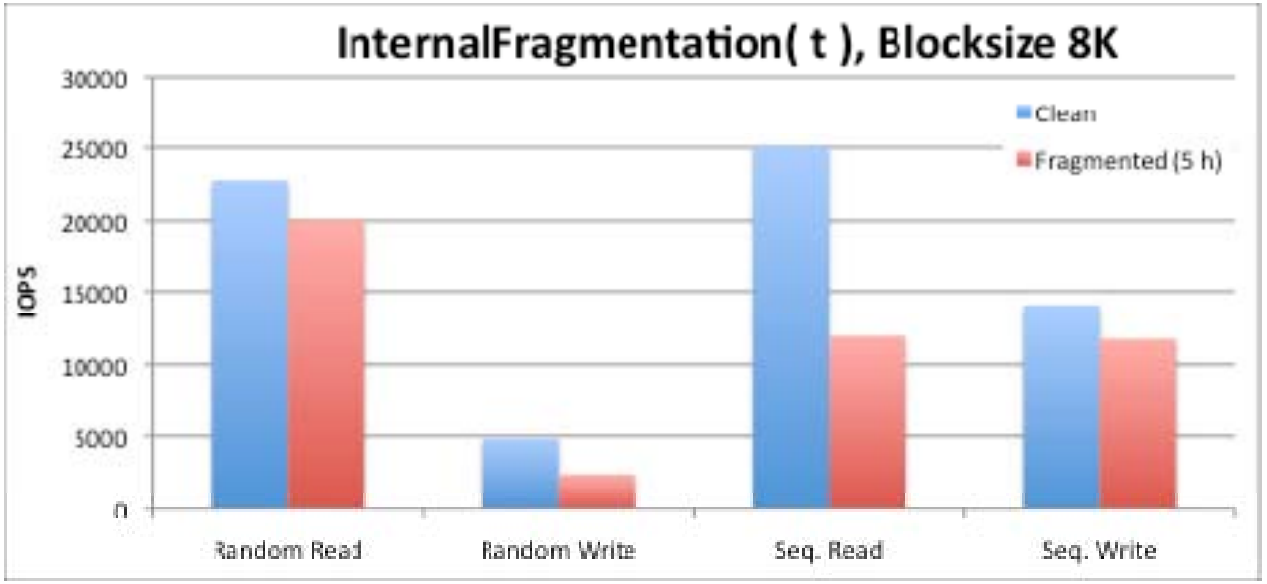
Fragmentation and Background Processes

TECHNISCHE UNIVERSITÄT DARMSTADT

Single Drive Fragmentation – max. 70% full

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Fragment: 5h write (rand., seq.)**
- Random reads less affected - 11%
- Seq. writes – 18% slower
  - Reason: (+) write cache/write back for small block sizes,(-) garbage collection
  - ➔ Worse for larger block sizes
- **Most affected Seq.Read, Rnd.Write**
- Sequential reads – 52% slower !
  - Read ahead not possible
  - Better for larger block sizes
- Random writes – 50% slower !
  - Reason: excessive garbage collection



flashyDB DVS

Single Drive Fragmentation – over 90% full

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Reads less affected**
  - Random reads not affected
  - Sequential reads approx 30% slower
- **Writes affected significantly**
  - Random writes 75% slower
  - Sequential writes 79% slower

SEQUENTIAL, 64K

	Read		Write	
	Fragmented	Non-Fragment.	Fragmented	Non-Fragment.
Bandw. [MB/s]	177	255	38	185
Avg. Latency [ms]	9	8	52	11

RANDOM, 4K

	Read		Write	
	Fragmented	Non-Fragment.	Fragmented	Non-Fragment.
IOPS	38900	39810	828	3358
Avg. Latency [ms]	0.8	0.8	39	10

Flash Trends [A. v. Bechtolsheim HPTS 2009]

TECHNISCHE UNIVERSITÄT DARMSTADT

- Density doubling each year → 1TB in 4 years
  - Costs falling by 50% per year
- Access times falling by 50% per year → 5µs in 4 years
- Throughput doubling every year
- Interface moving from SATA to PCI Express
- Very large-scale I/O looks feasible

SSD RAID Storage

TECHNISCHE UNIVERSITÄT DARMSTADT

How do we build large SSD storage?

Single SSD vs. SSD RAID

TECHNISCHE UNIVERSITÄT DARMSTADT

Device	Seq. Read [MB/s]	Seq. Write [MB/s]	Rnd. Read [ms]	Rnd. Write [ms]	Rnd. Read IOPS	Rnd. Write IOPS	Price [€/GB]	Price Read IOPS/€	Price Write IOPS/€
E.SSD	250	170	0.075	0.085	35 000	3 300	10	56	5.3
RAID0 2xSSD	422	631	0.375	0.458	24 371	2 035	19	13	1.1

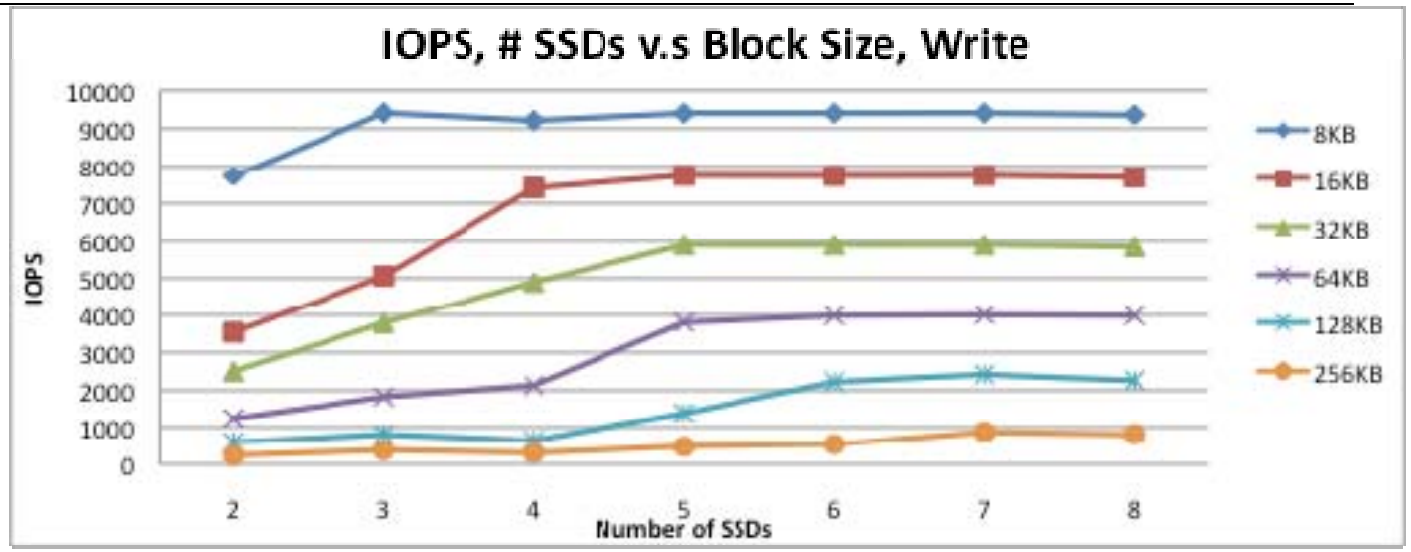
What did go wrong?

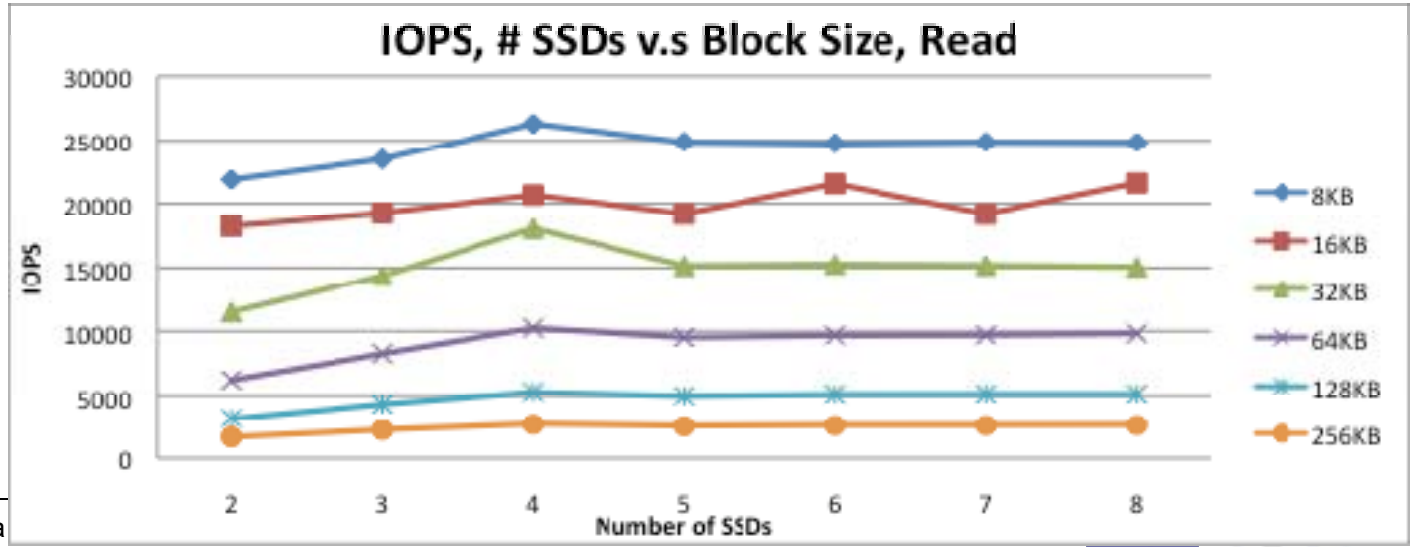
- RAID benefits come at a high cost in SSD configurations
- Random throughput (IOPS) → approx. 30% lower
- Sequential read throughput (MB/s) → better than that of a single SSD
- Sequential write throughput good
  - Entirely due to write caching

Scalability Tests – Random Load

TECHNISCHE UNIVERSITÄT DARMSTADT

- RAID 0
- Controller saturated with:
- SMALL Block size
  → 2 SSDs!!!
  (even 1)
- Larger block sizes
  → more SSDs
  → less than 4

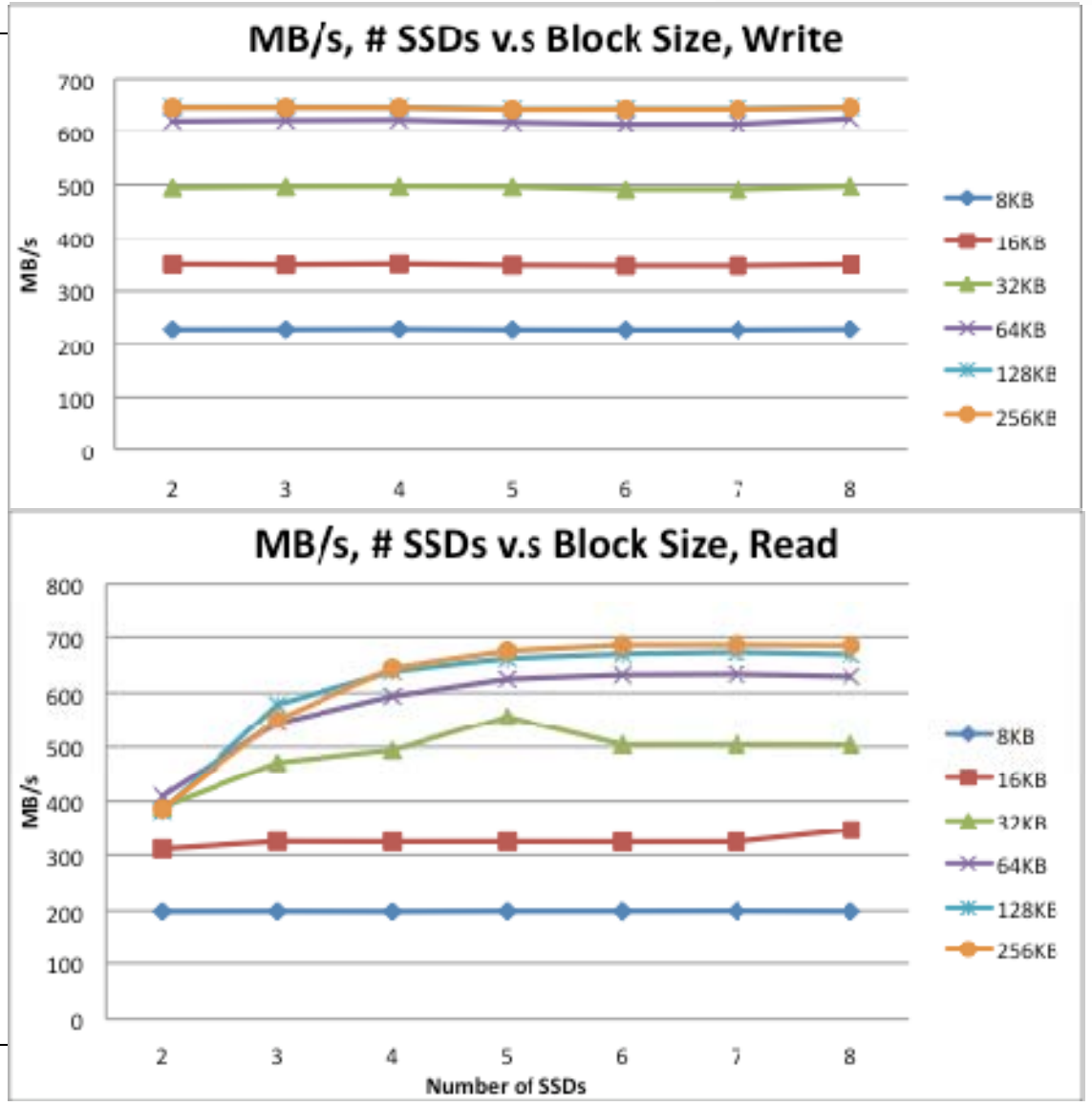




Scalability Tests – Sequential Tests

TECHNISCHE UNIVERSITÄT DARMSTADT

- RAID 0
- Write saturated from start
  - Controller Cache → **Seq**.!
  - Scales with writing threads
- Read
  - Contrl. Cache - ineffective
- File System = Raw Dev.



Hardware/Software RAID Configurations

TECHNISCHE UNIVERSITÄT DARMSTADT

Hardware/Software RAID

TECHNISCHE UNIVERSITÄT DARMSTADT

4 SSD Total		1 Controller (4 SSDs)				2 Controllers (2 SSD per Controller)			
		RAID0 HW		RAID0 SW SimpleVolumes		RAID0 SW 2 SSD per Controller RAID0 HW		RAID0 SW 2 SSD per Controller RAID0 SimpleVolumes	
Quantity	blockSize	Read	Write	Read	Write	Read	Write	Read	Write
Sequential Throughput	256KB	672	397	671	462	1033	762	1031	684
	512KB	670	398	674	468	1039	760	1030	687
Sequential Latency	256KB	0.743	0.743	0.688	0.711	0.772	0.512	0.531	0.461
	512KB	1.168	1.382	1.152	1.254	1.303	0.913	0.877	0.791
Random Throughput	4KB	24787	10193	27675	11704	44537	19529	49054	22512
	8KB	20987	6289	25417	10575	41091	13657	44129	13765
Random Latency	4KB	0.353	0.204	0.277	0.120	0.282	0.114	0.277	0.109
	8KB	0.429	0.220	0.365	0.196	0.334	0.161	0.332	0.138

Two Controllers double the Performance!
Simple Volumes better random throughput!
HW RAID0 better sequential throughput!
Host Based Storage!

Thank You!

TECHNISCHE UNIVERSITÄT DARMSTADT

http://www.dvs.tu-darmstadt.de/research/flashydb/