SIAS-Chains: Snapshot Isolation Append Storage Chains

TECHNISCHE UNIVERSITÄT DARMSTADT

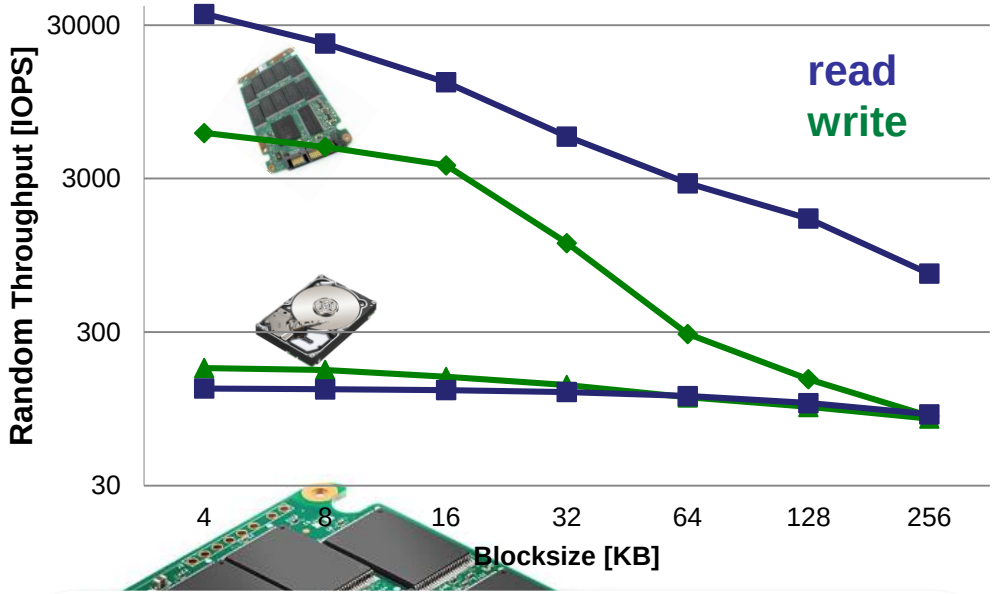
Hochschule Reutlingen
Reutlingen University

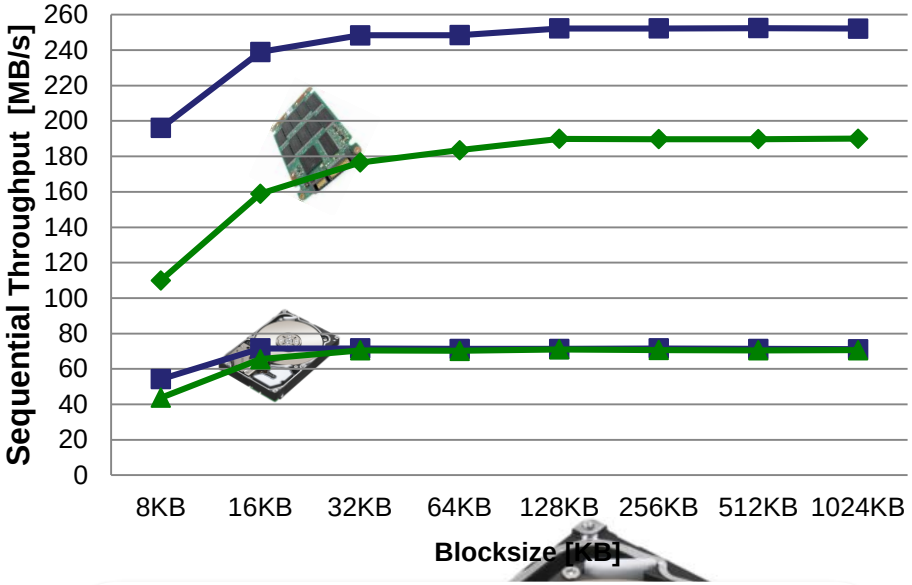
Dr. Robert Gottstein
Prof. Ilia Petrov
M.Sc. Sergej Hardock
Prof. Alejandro Buchmann

DVS

Motivation: Storage Technology Evolution

Significant impact of storage technology evolution



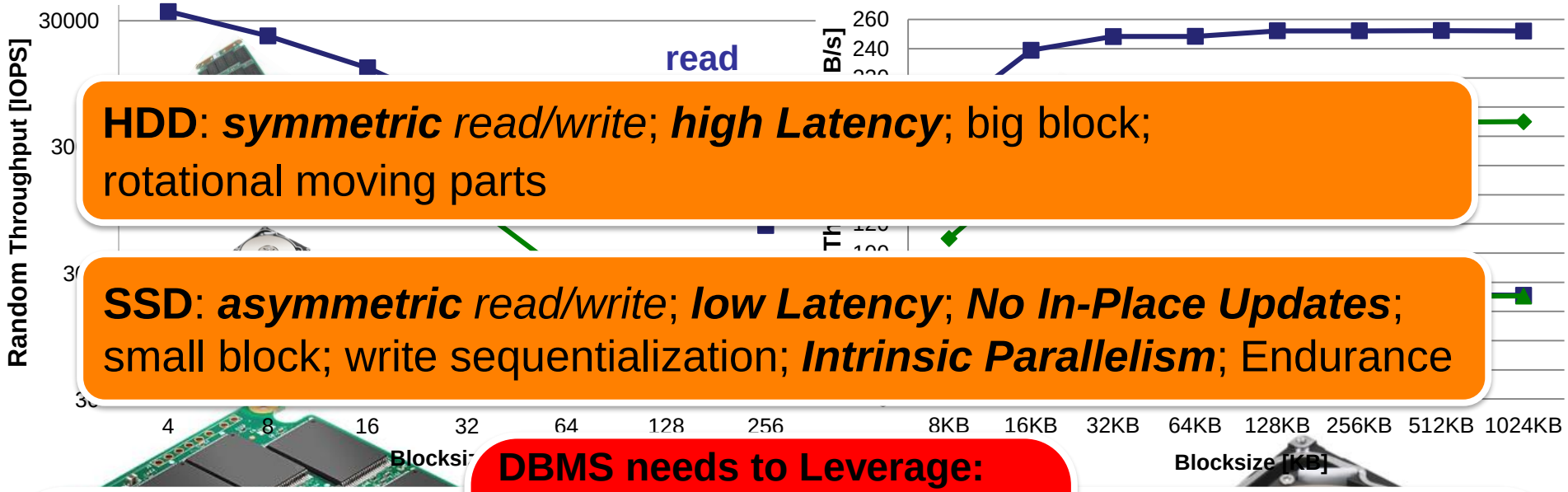


- **Intel X25-E SLC SSD**
- Seq. Read/Write: 250 / 170 MB/s
- Read/Write IOPS (4K): 35 000 / 3 300
- Latency Read/Write (4K): 0.075/0.085 ms
- Erase before overwrite
  - slow & large granularity

- **Savvio 15k HDD**
- Seq. Read / Write: 160 MB/s
- Read/Write IOPS: 350 / 300
- Latency Read/Write: 3.2 / 3.5 ms
- Direct overwrite

Motivation: Storage Technology Evolution

Significant impact of storage technology evolution



HDD: ***symmetric*** *read/write*; ***high Latency***; big block; rotational moving parts

SSD: ***asymmetric*** *read/write*; ***low Latency***; ***No In-Place Updates***; small block; write sequentialization; ***Intrinsic Parallelism***; Endurance

DBMS needs to Leverage:

- **Fast Reads**
- **Low Latencies**
- **Asymmetry**
- **Parallelism**
- **Write Sequentialization**

- **Intel X25-E SLC SSD**
- Seq. Read/Write:
- Read/Write IOPS (4K):
- Latency Read/Write (4
- Erase before overwrite
  - slow & large granularity

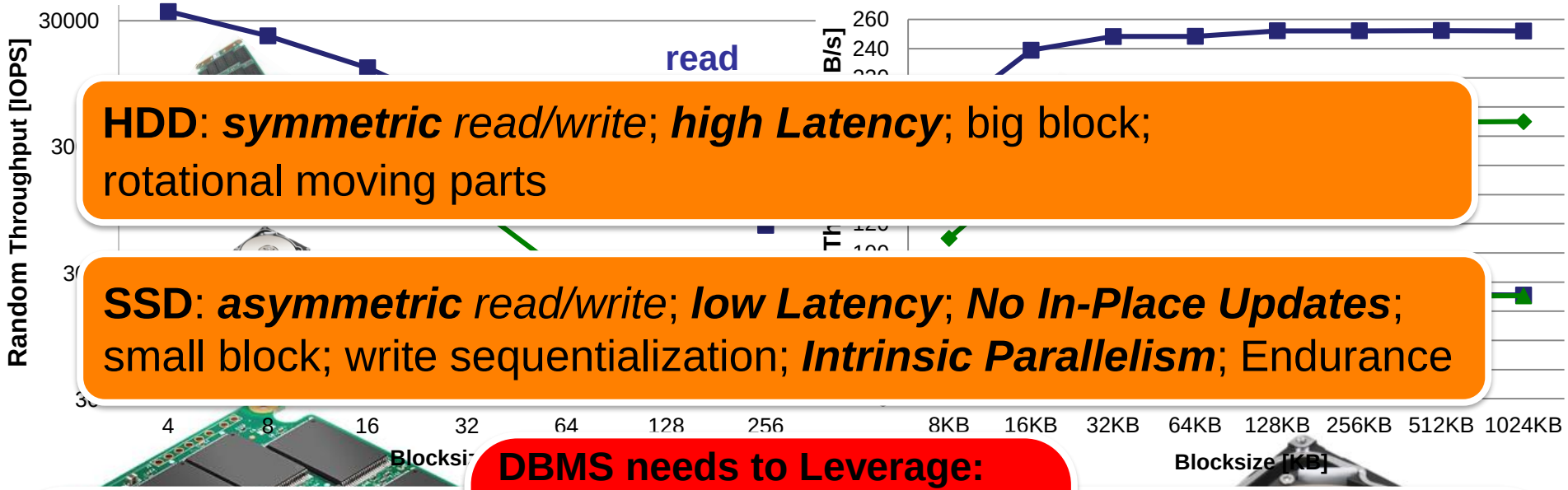
.5k HDD

ad / Write:	160 MB/s
ite IOPS:	350 / 300
Read/Write:	3.2 / 3.5 ms

- Direct overwrite

Motivation: Storage Technology Evolution

Significant impact of storage technology evolution



HDD: ***symmetric*** *read/write*; ***high Latency***; big block; rotational moving parts

SSD: ***asymmetric*** *read/write*; ***low Latency***; ***No In-Place Updates***; small block; write sequentialization; ***Intrinsic Parallelism***; Endurance

DBMS needs to Leverage:

- Parallelism
- Write Sequentialization

Multi Version DBMS: In principle suitable for asymmetric storage. Parallelism. Out-of place updates. Sequentialization....

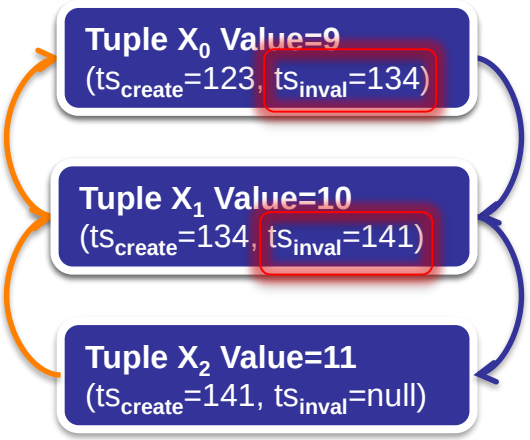
- Read/Write IOPS (4K): ite IOPS: 350 / 300
- Latency Read/Write (4K): Read/Write: 3.2 / 3.5 ms
- Erase before overwrite
  - slow & large granularity
- Direct overwrite

Introduction

Version Organization & Invalidation

Relation R	…	A	….
Item X	Version X_0 …	9	…
	Version X_1 …	10	…
	Version X_2 …	11	…

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



Visibility

- **Timestamps**
  creation: ts_{create}
  invalidation: ts_{inval}

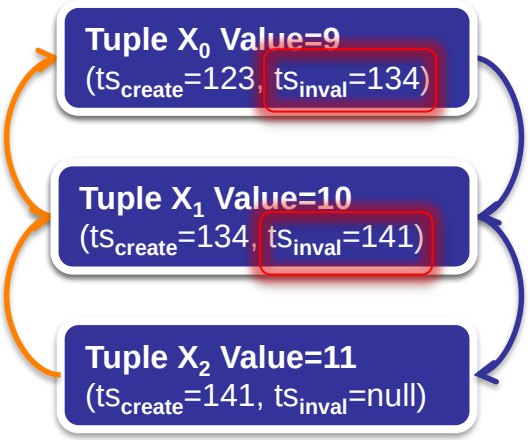
- ***Asymmetric***: Fast Reads & Slow Writes
- ***Low Latency***: no moving parts
- ***No In-Place Updates***: Need to erase first (slow)
- ***Intrinsic Parallelism***: Read in parallel

Introduction

Version Organization & Invalidation

Relation R	…	A	….
Item X	Version X_0 …	9	…
	Version X_1 …	10	…
	Version X_2 …	11	…

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



Visibility

- **Timestamps**
  creation: ts_{create}
  invalidation: ts_{inval}

- ***Asymmetric***: Fast Reads & Slow Writes
- ***Low Latenc***
- ***No In-Place***
- ***Intrinsic Pa***

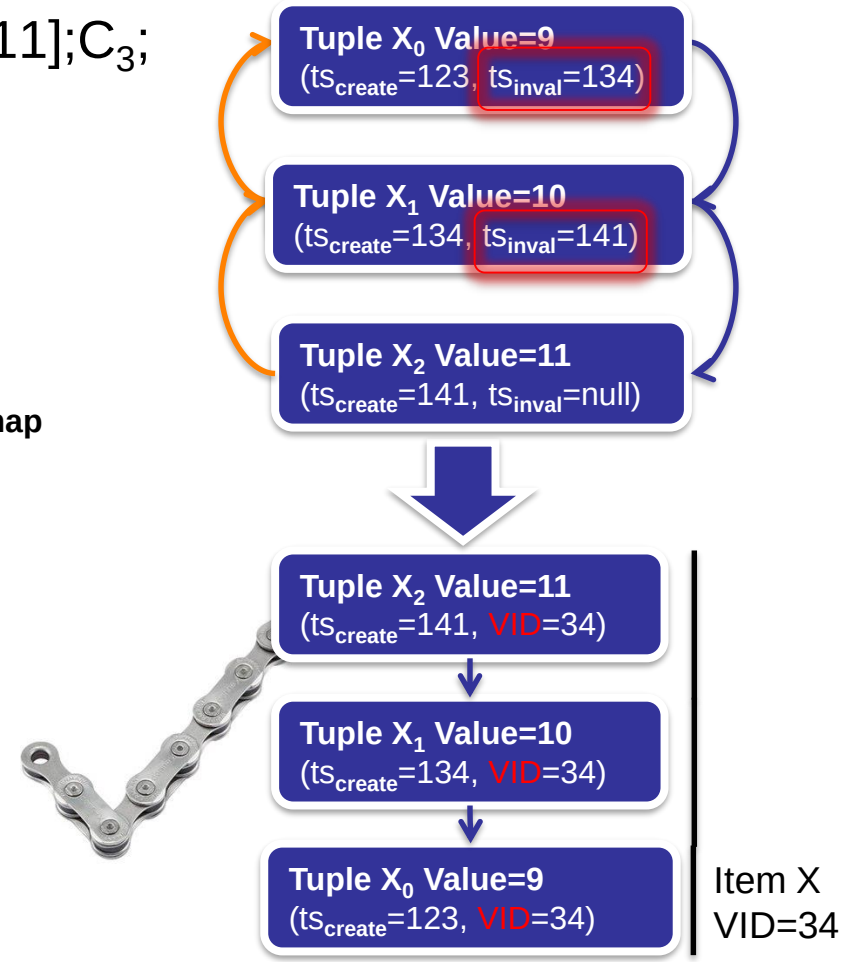
Version Organization & Invalidation
Small Random Updates

SIAS: Snapshot Isolation Append Storage

SIAS in a nutshell: redesign architecture and algorithms

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$

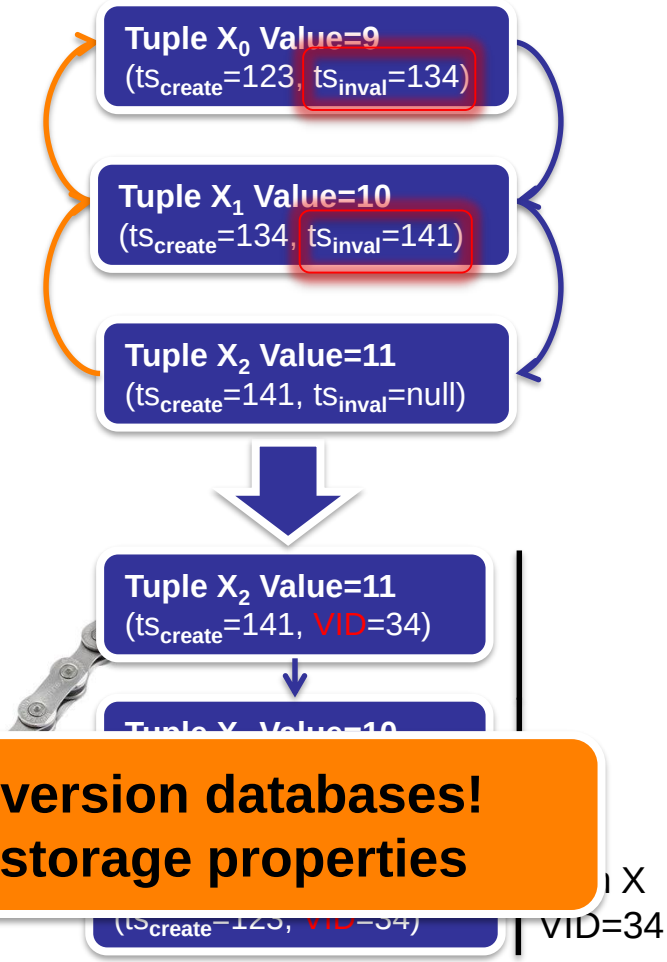
- **Version Organization**
  - Backward ***Chaining*** of versions
  - Chain identified by virtual ID (**VID**)
  - Store the entrypoint in datastructure: $\mathbf{VID_{map}}$
- **New Invalidation**
  - Invalidation coded within the chain
  - „One-place“ Invalidation
- **Append Storage**
  - Append tuple versions to a new page
  - Write page when filled or on a threshold



SIAS: Snapshot Isolation Append Storage

SIAS in a nutshell: redesign architecture and algorithms

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$

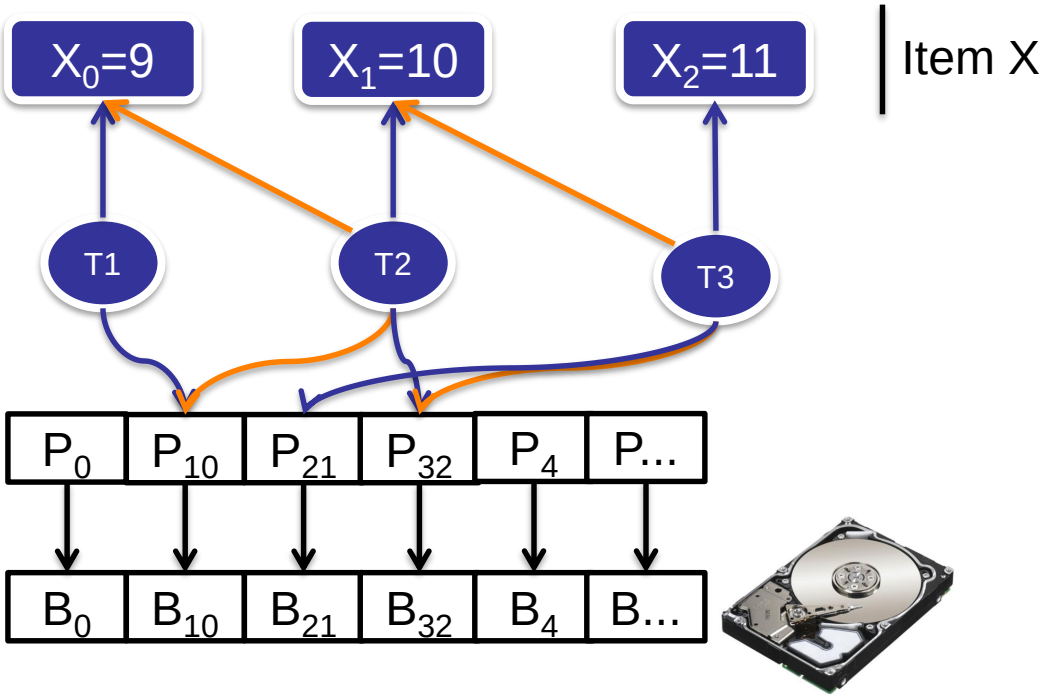


- **Version Organization**
  - Backward ***Chaining*** of versions
  - Chain identified by virtual ID (**VID**)
  - Store the entrypoint in datastructure: **VID_{map}**
- **New Invalidation**
  - Invalidation coded within the chain
  - „One-place“ Invalidation
-
  - Append tuple versions to a new page
  - Write page when filled or on a threshold

Variant above is widely spread in multi version databases!
Variant below allows to address Flash storage properties

Multi Version DBMS Example

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



Invalidation

Creation

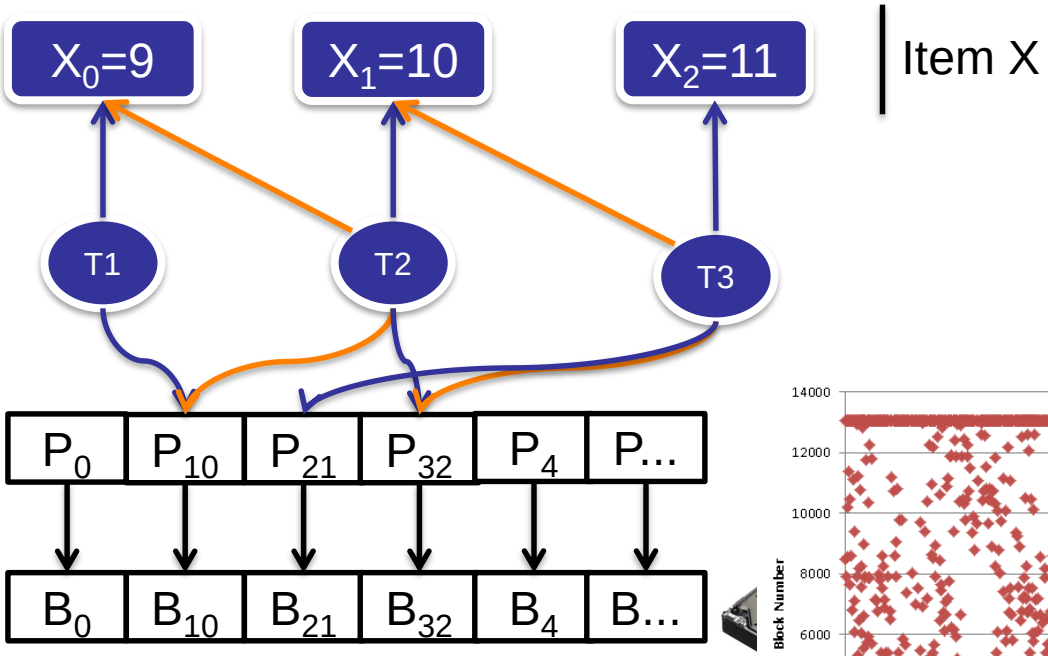
Ti Transaction

P_n DB Page

B_n Device Block

Multi Version DBMS Example

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



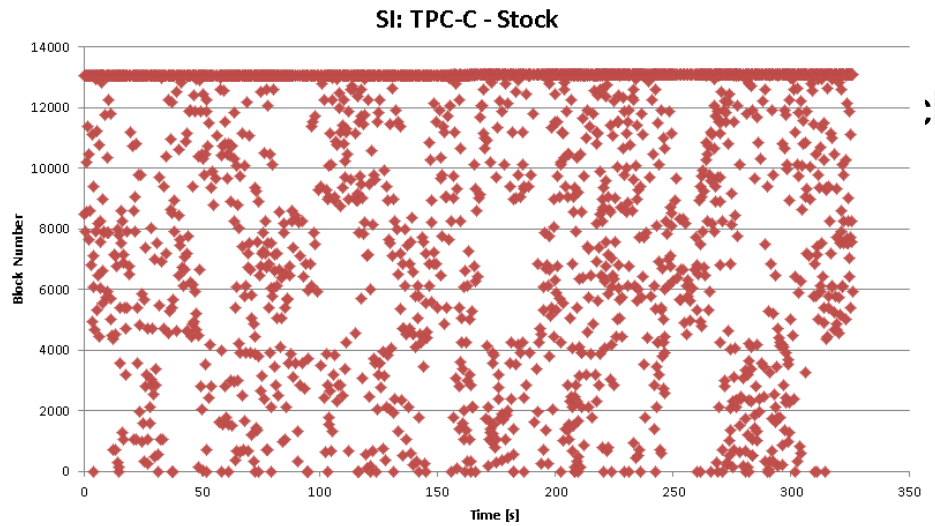
Item X

Invalidation

Creation

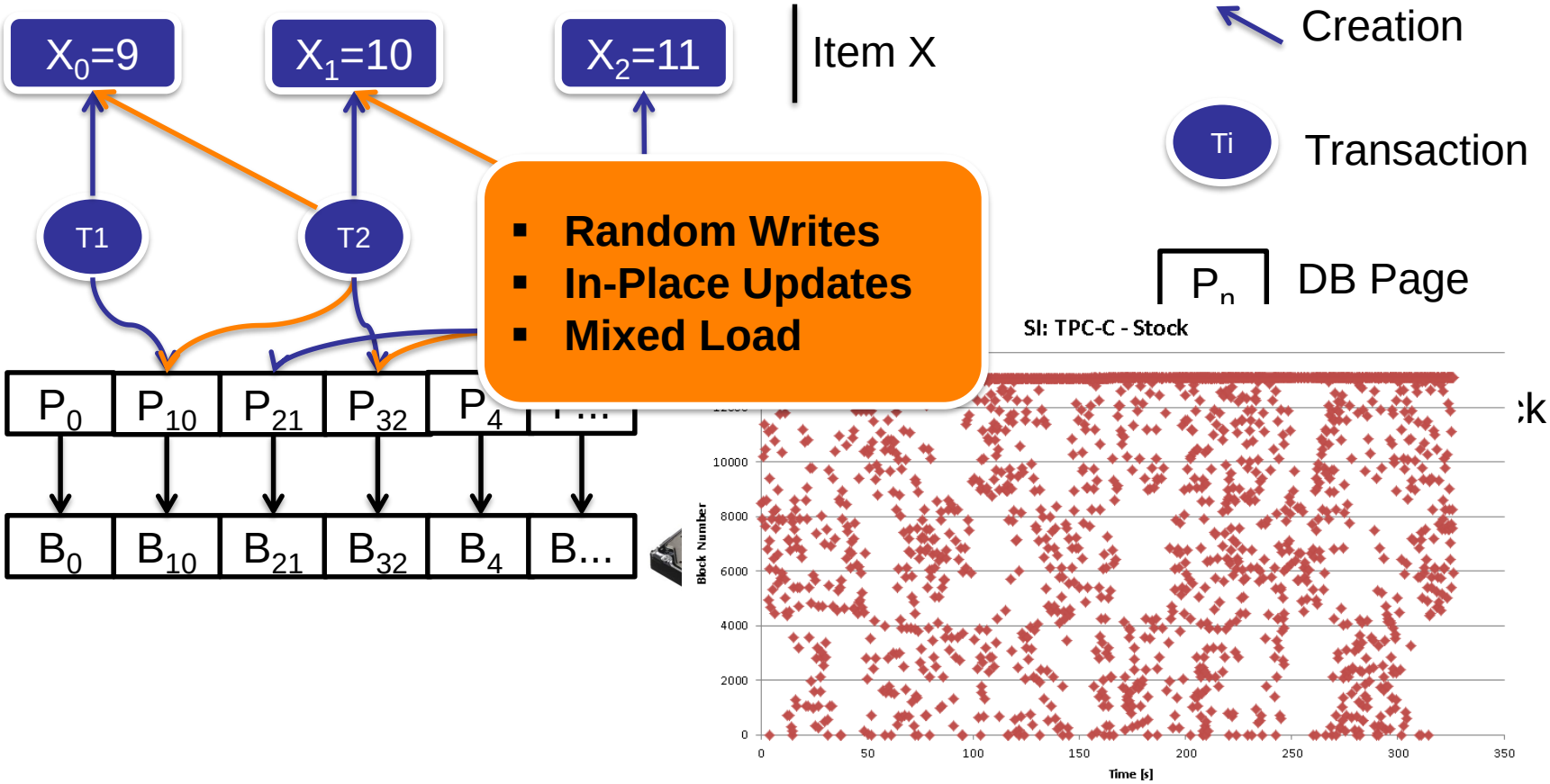
Ti Transaction

P_n DB Page



Multi Version DBMS Example

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$

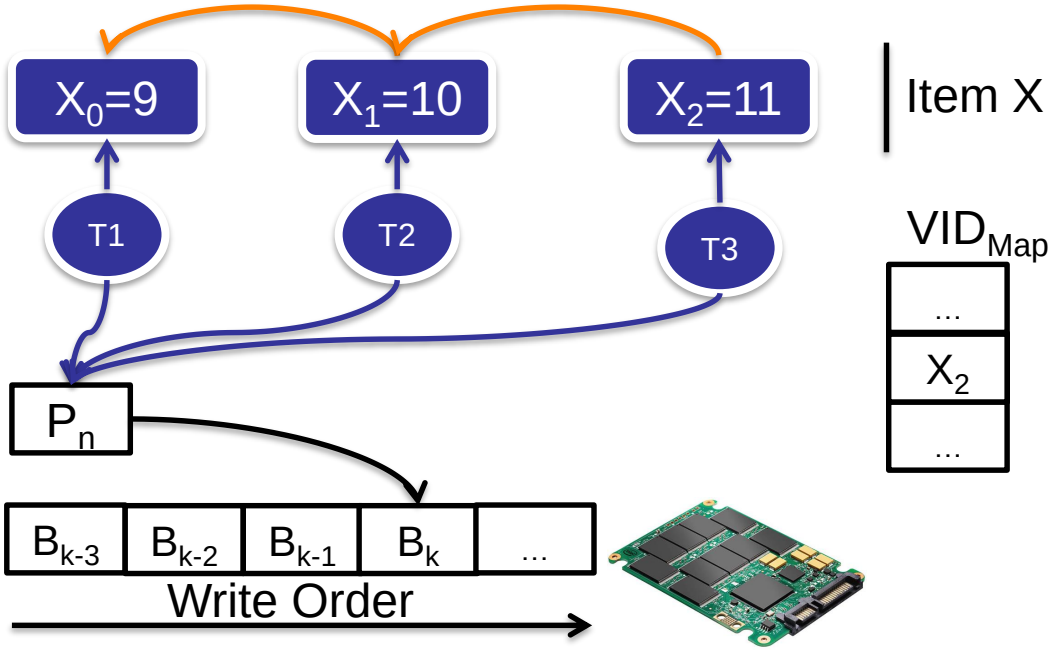


- **Random Writes**
- **In-Place Updates**
- **Mixed Load**

SIAS Principle Example

Tuple Append Storage Management

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



Invalidation

Creation

Ti Transaction

P_n DB Page

B_n Device Block

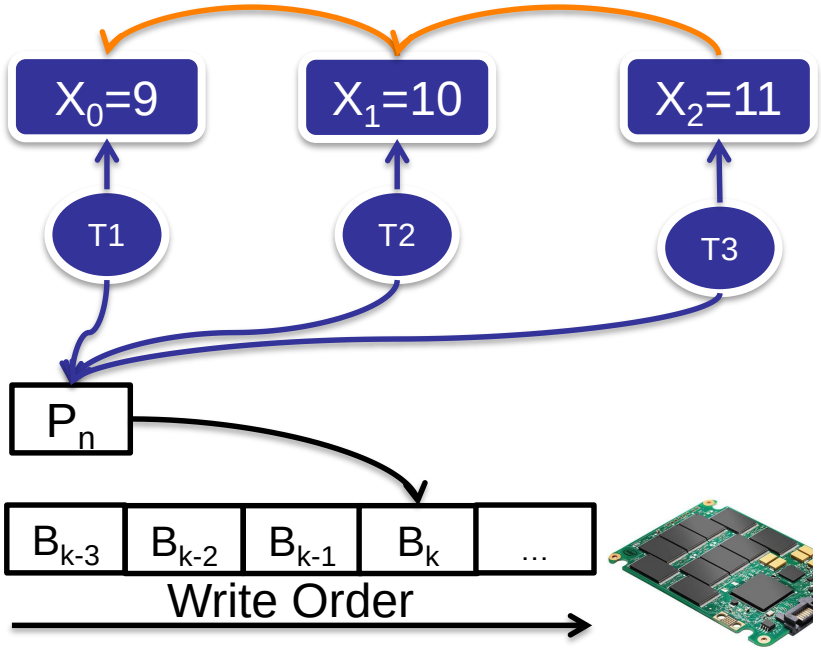
- No in-place invalidation
- Append versions instead of pages
- Write filled pages

- DBMS specific
- Write reduction
- Simplyfied Buffer Management

SIAS Principle Example

Tuple Append Storage Management

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



Item X

VID_{Map}

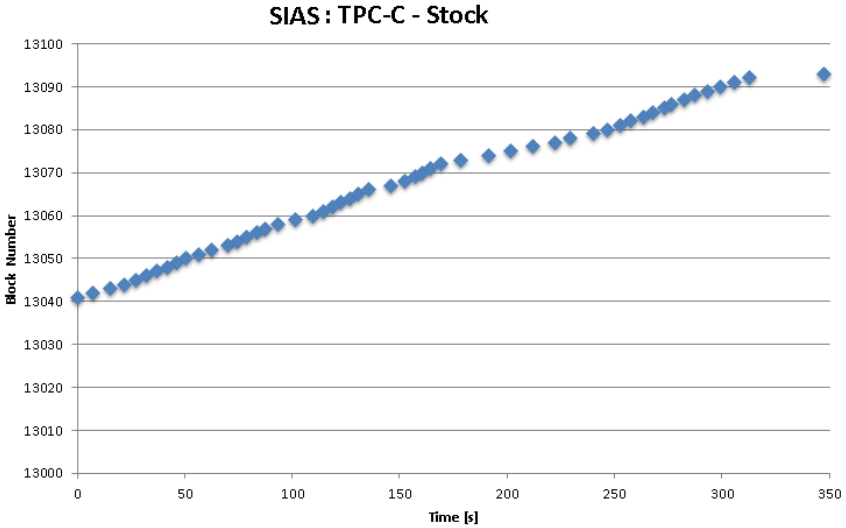
...

Invalidation

Creation

Ti Transaction

P_n DB Page



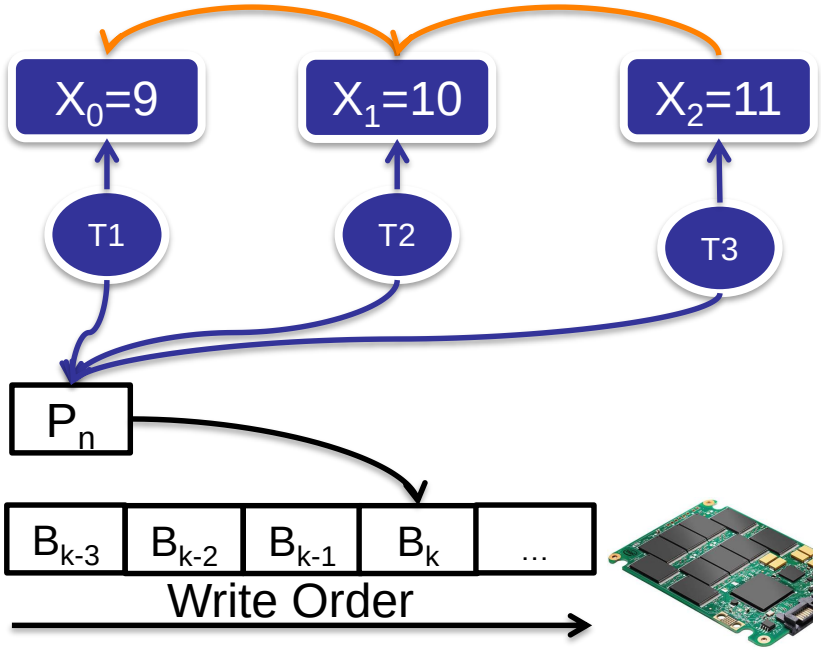
ock

- No in-place invalidation
- Append versions instead of pages
- Write filled pages

SIAS Principle Example

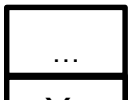
Tuple Append Storage Management

$W_1[X_0=9];C_1;$ $W_2[X_1=10];C_2;$ $W_3[X_2=11];C_3;$



Item X

VID_{Map}

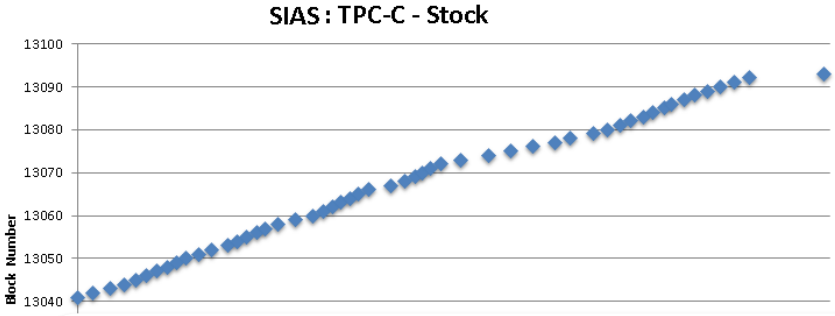


Invalidation

Creation

Ti Transaction

P_n DB Page

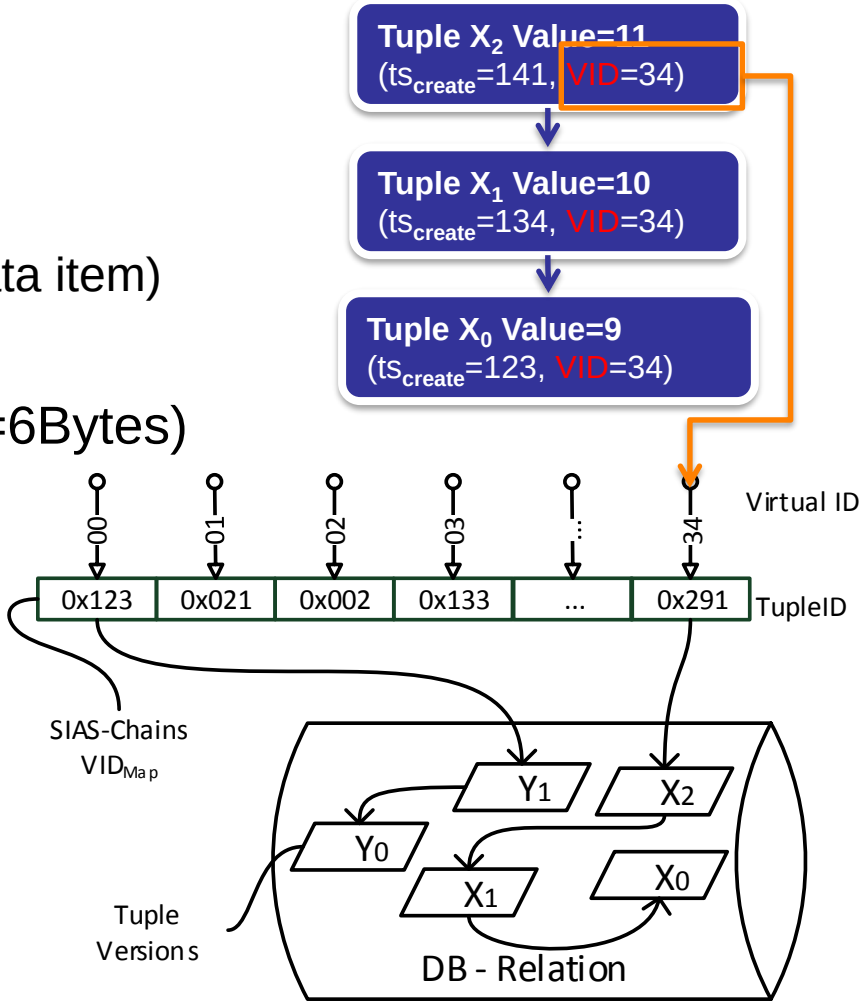


- No in-place invalidation
- Append versions instead of pages
- Write filled pages

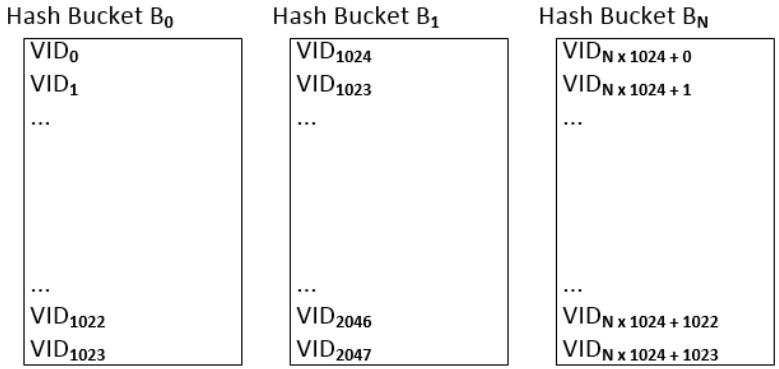
Significant Write Reduction (5 pages vs. 1 page) Sequentialization

SIAS Principle: VID_{map}

- VID not explicitly stored
  → Index to Hash Bucket
- No Overflow buckets
  → VIDs are unique: one VID per tuple (data item)



- Storing one TID per VID (sizeof(TID)=6Bytes)
  - TIDpos=VID mod 1024
  - $BucketNr = \lfloor \frac{VID}{1024} \rfloor$



Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

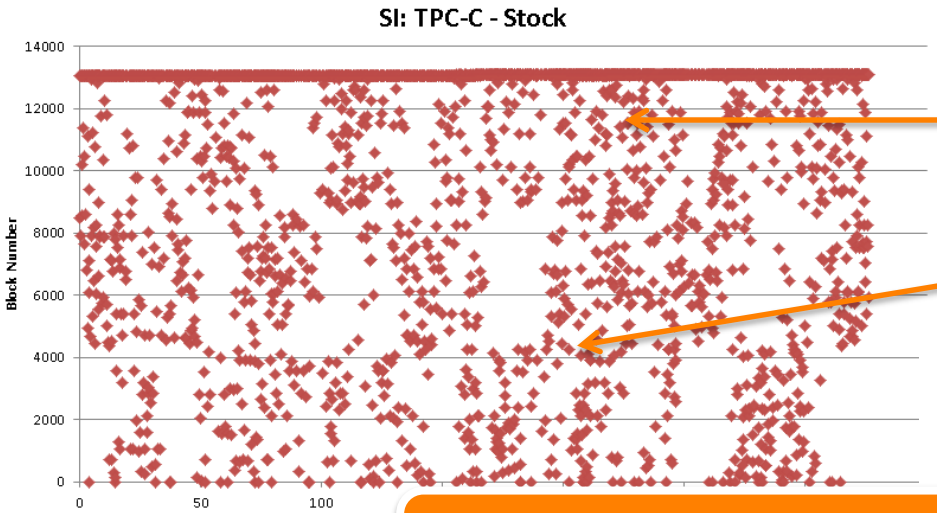
Transaction Processing Council Benchmark C

TPC-C OLTP Benchmark

Write reduction

TPC-C OLTP benchmark – Stock relation

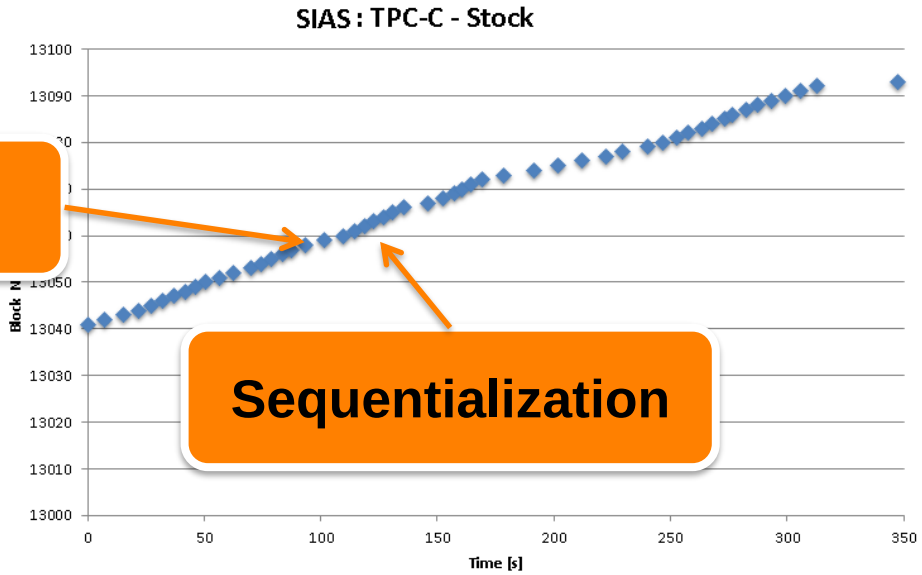
- **Significant reduction of host writes**



Random & In Place

2566 Page Writes!

53 Page Writes!



Sequentialization

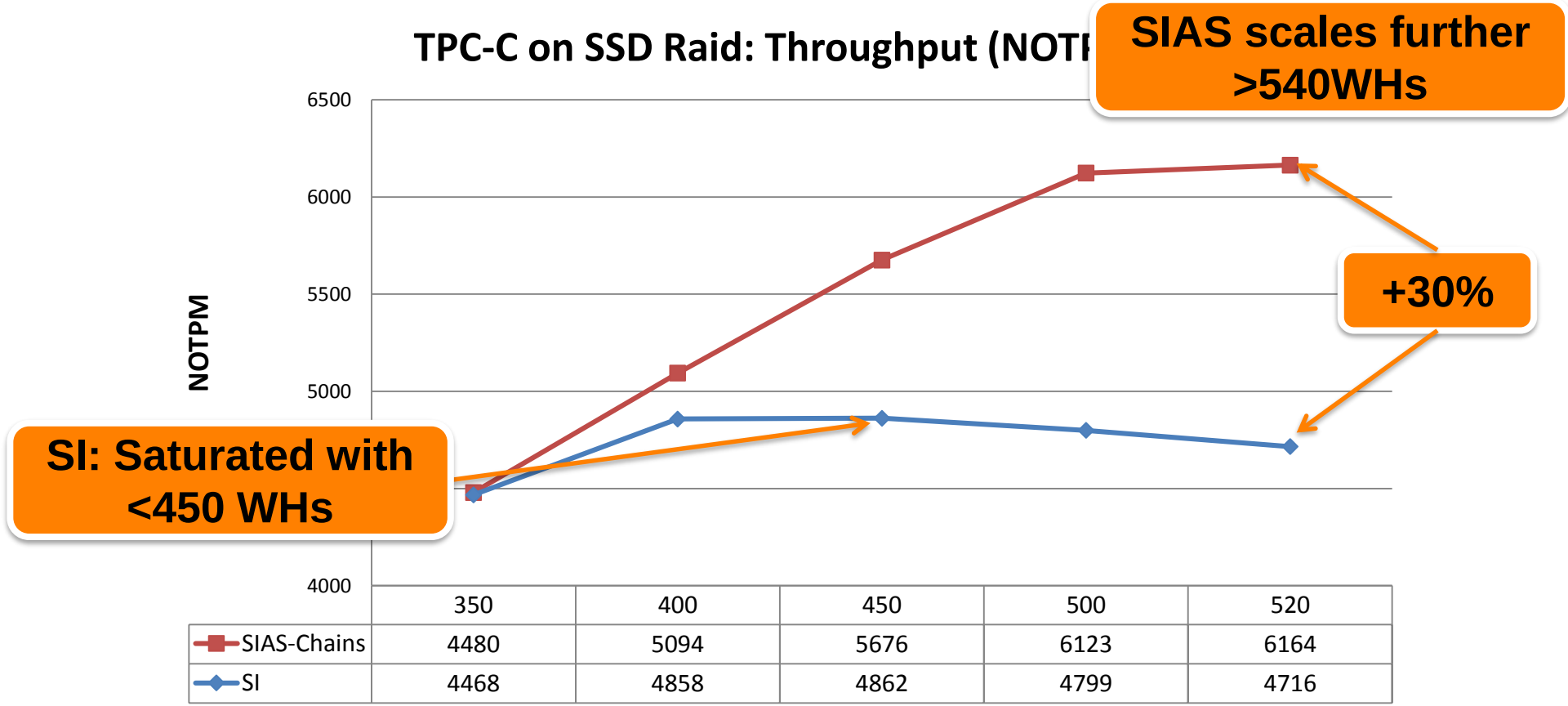
- 97% write reduction

SSD Endurance?

TPC-C: Throughput on 2x SSD RAID

New Order Transactions per Minute (NOTPM)

TECHNISCHE UNIVERSITÄT DARMSTADT



	350	400	450	500	520
SIAS-Chains	4480	5094	5676	6123	6164
SI	4468	4858	4862	4799	4716

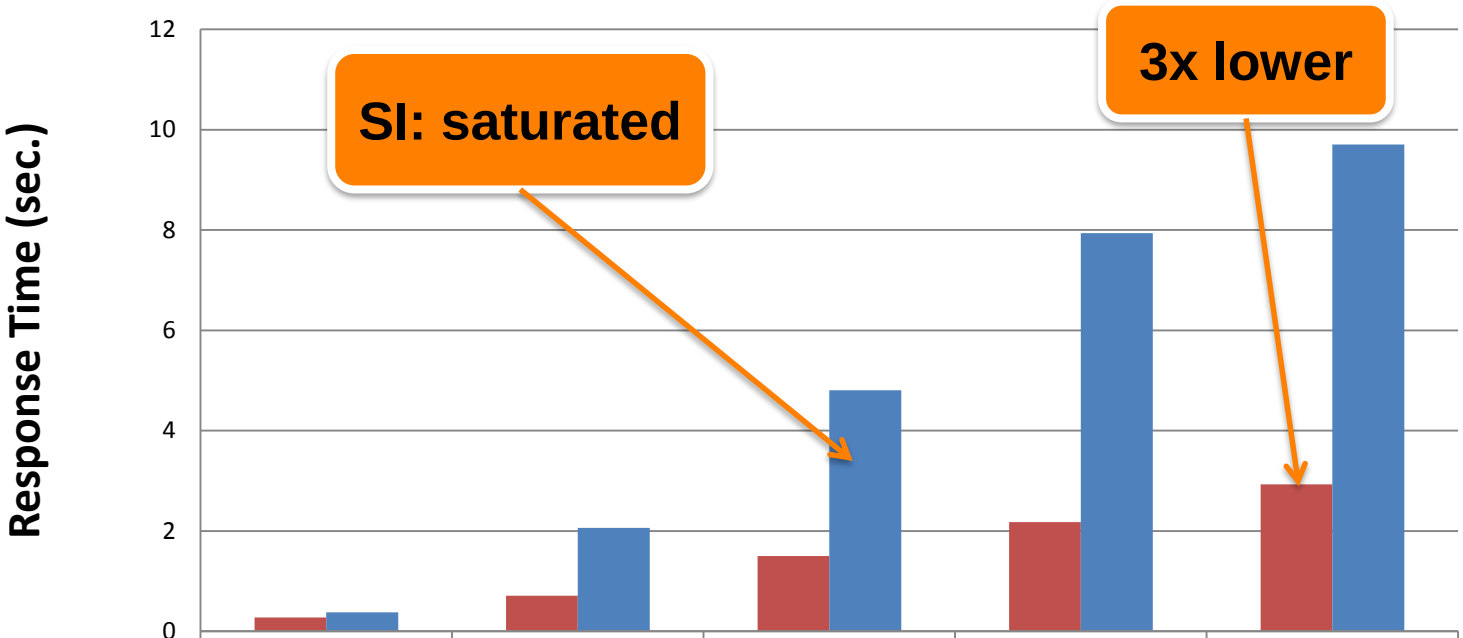
DVS

TPC-C: Response Time on 2x SSD RAID

Average response time: new order transaction

TECHNISCHE UNIVERSITÄT DARMSTADT

TPC-C on SSD Raid: Response Time (sec.)



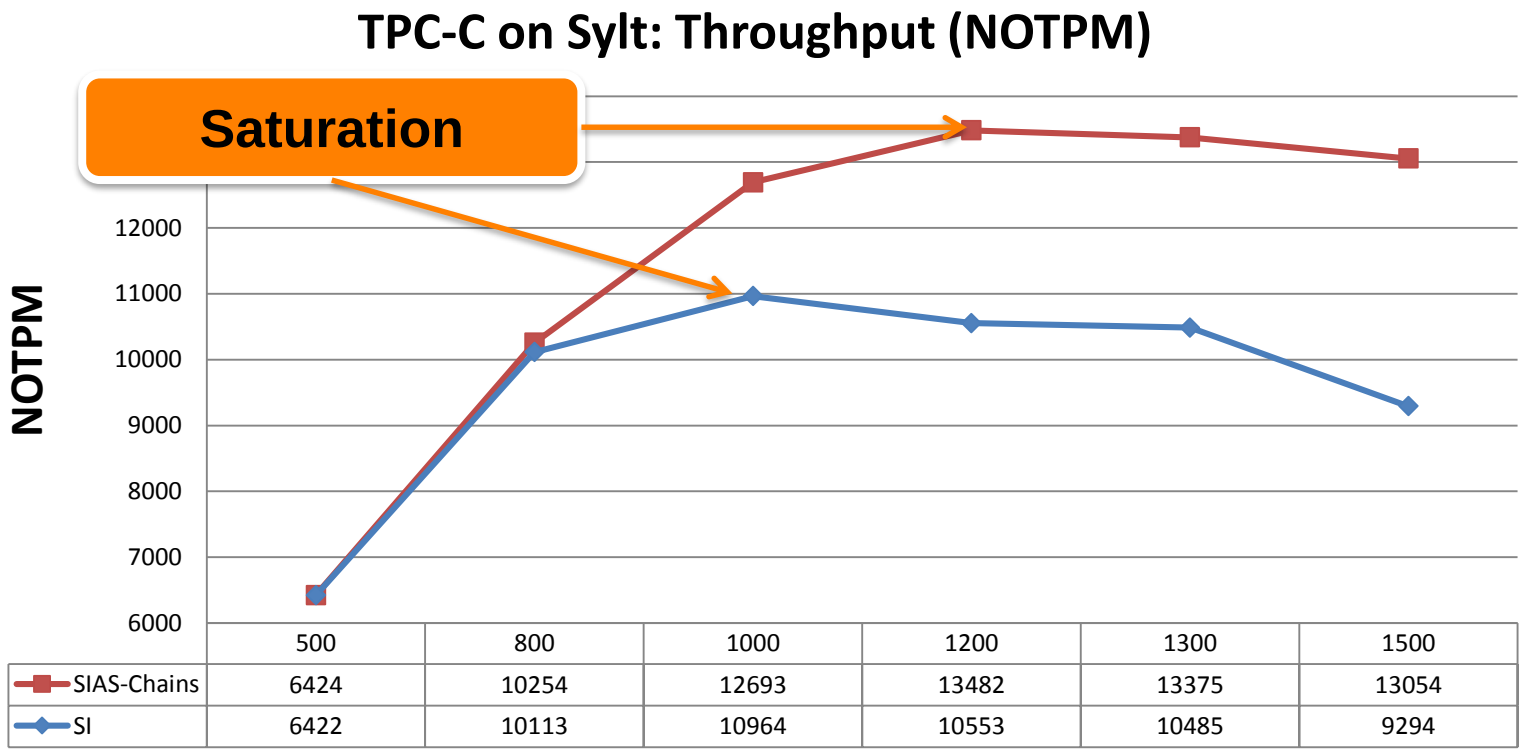
	350	400	450	500	520
SIAS-Chains	0,274	0,711	1,5	2,179	2,931
SI	0,376	2,063	4,806	7,937	9,707

Peak Load tolerance: SIAS scales with ~540WHs

DVS

TPC-C: Throughput on 6x SSD RAID (Sylt)

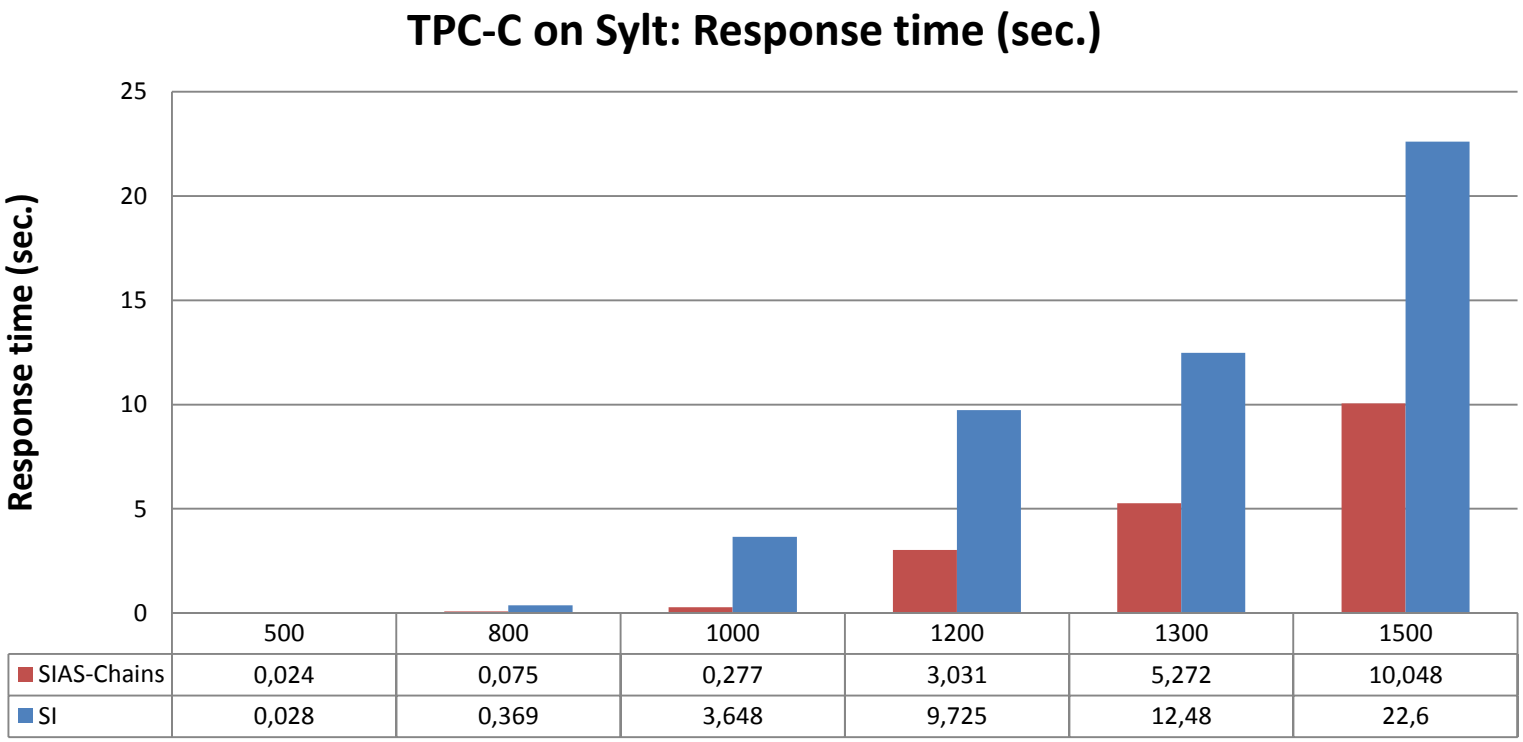
New Order Transactions per Minute (NOTPM)



SIAS removes I/O Bottleneck

TPC-C: Response Time on 6x SSD RAID (Sylt)

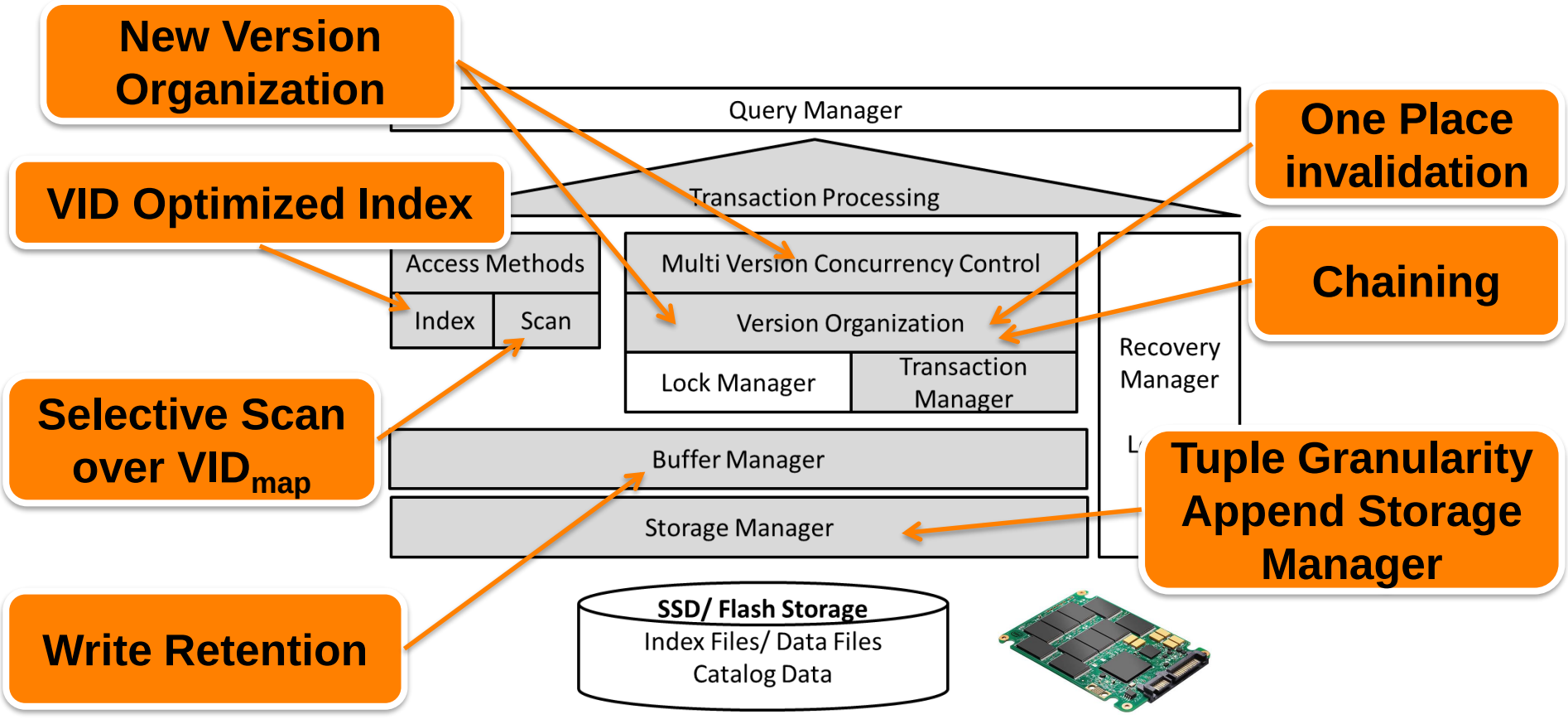
Average response time of new order table



SI Saturation <1200WH
SIAS Saturation ~1300WH

Contributions

SIAS Architectural Changes



Implementation in PostgreSQL

Thank You

Publications



TECHNISCHE UNIVERSITÄT DARMSTADT

- **Gottstein, Robert. Impact of New Storage Technologies on an OLTP DBMS, Its Architecture and Algorithms.** Doctoral dissertation, Technische Universität Darmstadt. 2016.
- **DBMS on Modern Storage Hardware (Tutorial).** I. Petrov, R. Gottstein, S. Hardock. ICDE 2015.
- **SIAS-V in Action: Snapshot Isolation Append Storage - Vectors on Flash.** Robert Gottstein, Thorsten Peter, Ilia Petrov and Alejandro Buchmann. In 17th International Conference on Extending Database Technology (EDBT) 2014.
- **Multi-Version Databases on Flash:** Append Storage and Access Paths. Robert Gottstein, Ilia Petrov, Alejandro Buchmann. International Journal On Advances in Software, Vol. 6, Number 3 and 4 2013.
- **Read Optimisations for Append Storage on Flash.** Robert Gottstein, Ilia Petrov, Alejandro Buchmann. 17th International Database Engineering and Applications Symposium, Barcelona, Spain, ACM, 2013. (IDEAS 2013).
- **FBARC: I/O Asymmetry-Aware Buffer Replacement Strategy.** Paul Dubs, Ilia Petrov, Robert Gottstein, Alejandro Buchmann . Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures" (ADMS 2013), in conjunction with VLDB 2013, Riva del Garda, Trento, Italy in August 2013.
- **Aspects of Append-Based Database Storage Management on Flash Memories.** Robert Gottstein, Ilia Petrov, Alejandro Buchmann. DBKDA 2013. Best Paper Award.
- **Append Storage in Multi-Version Databases on Flash.** Robert Gottstein, Ilia Petrov, Alejandro Buchmann. 29th British National Conference on Databases, BNCOD 2013, University of Oxford, United Kingdom, 2013.
- **SI-CV: Snapshot Isolation with Co-located Versions.** Robert Gottstein, Ilia Petrov, Alejandro Buchmann. In Raghunath Nambiar, Meikel Poess: Topics in Performance Evaluation, Measurement and Characterization, Lecture Notes in Computer Science 7144, ISBN 978-3-642-32626-4, Springer Berlin / Heidelberg, 2012 10.1007/978-3-642-32627-1_9
- **Data-Intensive Systems on Evolving Memory Hierarchies.** I. Petrov, D. Bausch, R. Gottstein, A. Buchmann. EEbS 2012.
- **Revisiting DBMS Space Management for Native Flash**. S. Hardock, I. Petrov, R. Gottstein, A. Buchmann. EDBT 2016.
- **NoFTL for Real:** Databases on Real Native Flash Storage. S. Hardock, I. Petrov, R. Gottstein, A. Buchmann. EDBT 2015.
- **NoFTL: Database Systems on FTL-less Flash Storage.** Sergej Hardock, Ilia Petrov, Robert Gottstein, Alejandro Buchmann. 39th International Conference on Very Large Databases (VLDB), Riva del Garda, Italy, 2013.