SIAS-Chains: Snapshot Isolation Append Storage Chains

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

**5k HDD**
- 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:**

- Parallelism
- Write Sequentialization
- Direct overwrite
- 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

**Intel X25-ESLC 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

01.09.2017 | Dr. Robert Gottstein |
Introduction
Version Organization & Invalidation

Relation R | A
---|---
Version X₀ | 9
Version X₁ | 10
Version X₂ | 11

Visibility
- **Timestamps**
  - creation: \( t_{\text{create}} \)
  - invalidation: \( t_{\text{inval}} \)

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

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

Version Organization & Invalidation

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

Visibility
- Timestamps
  - creation: $ts_{\text{create}}$
  - invalidation: $ts_{\text{inval}}$

<table>
<thead>
<tr>
<th>Relation R</th>
<th>...</th>
<th>A</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version $X_0$</td>
<td>9</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Version $X_1$</td>
<td>10</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Version $X_2$</td>
<td>11</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

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

Version Organization & Invalidation
Small Random Updates
SIAS: Snapshot Isolation Append Storage

SIAS in a nutshell: redesign architecture and algorithms

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

- **Version Organization**
  - Backward *Chaining* of versions
  - Chain identified by virtual ID (VID)
  - Store the entrypoint in datastructure: \( \text{VID}_{\text{map}} \)

- **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

Variant above is widely spread in multi version databases!

Variant below allows to address Flash storage properties

- Append tuple versions to a new page
- Write page when filled or on a threshold

01.09.2017 | Dr. Robert Gottstein |
Multi Version DBMS Example

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

Item X

Invalidation

Creation

Transaction

DB Page

Device Block
Multi Version DBMS Example

$W_1[X_0=9];C_1; \quad W_2[X_1=10];C_2; \quad W_3[X_2=11];C_3;$
Multi Version DBMS Example

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

- Random Writes
- In-Place Updates
- Mixed Load
SIAS Principle Example

Tuple Append Storage Management

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

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

DBMS specific
- Write reduction
- Simplified Buffer Management

**Invalidation**

**Creation**

**Transaction**

**DB Page**

**Device Block**

**Write Order**
SIAS Principle Example

Tuple Append Storage Management

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

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

\[ X_0=9 \]
\[ X_1=10 \]
\[ X_2=11 \]

\[ T_1 \]
\[ T_2 \]
\[ T_3 \]

\[ P_n \]

\[ B_k-3 \quad B_k-2 \quad B_k-1 \quad B_k \quad \ldots \]

Write Order

Invalidation

Creation

Transaction

DB Page

VID Map

Item X

SIAS: TPC-C - Stock

01.09.2017 | Dr. Robert Gottstein |
SIAS Principle Example

Tuple Append Storage Management

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

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

Invalidation
Creation
Transaction

Write Order

Significant Write Reduction (5 pages vs. 1 page) Sequentialization
SIAS Principle: $\text{VID}_{\text{map}}$

- 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
  - $\text{BucketNr} = \left\lfloor \frac{\text{VID}}{1024} \right\rfloor$

Hash Bucket $B_0$

<table>
<thead>
<tr>
<th>VID_0</th>
<th>VID_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Hash Bucket $B_1$

<table>
<thead>
<tr>
<th>VID_{1024}</th>
<th>VID_{1023}</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Hash Bucket $B_N$

<table>
<thead>
<tr>
<th>VID_{N x 1024 + 0}</th>
<th>VID_{N x 1024 + 1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\text{Virtual ID}$</th>
<th>$\text{TupleID}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0x{123}$</td>
<td>$0x{021}$</td>
</tr>
<tr>
<td>$0x{002}$</td>
<td>$0x{133}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\text{DB - Relation}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_0$</td>
</tr>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_0$</td>
</tr>
</tbody>
</table>

01.09.2017 | Dr. Robert Gottstein |
Evaluation

Transaction Processing Council Benchmark C

TPC-C OLTP Benchmark
Write reduction

TPC-C OLTP benchmark – Stock relation

- Significant reduction of host writes
  - SI: TPC-C - Stock
    - Random & In Place
    - 2566 Page Writes!
  - 53 Page Writes!

- 97% write reduction
  - SSD Endurance?
  - SIAS: TPC-C - Stock
    - Sequentialization

01.09.2017 | Dr. Robert Gottstein |
TPC-C: Throughput on 2x SSD RAID
New Order Transactions per Minute (NOTPM)

<table>
<thead>
<tr>
<th>NOTPM</th>
<th>350</th>
<th>400</th>
<th>450</th>
<th>500</th>
<th>520</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIAS-Chains</td>
<td>4480</td>
<td>5094</td>
<td>5676</td>
<td>6123</td>
<td>6164</td>
</tr>
<tr>
<td>SI</td>
<td>4468</td>
<td>4858</td>
<td>4862</td>
<td>4799</td>
<td>4716</td>
</tr>
</tbody>
</table>

SI: Saturated with <450 WHs
SIAS scales further >540WHs
+30%
TPC-C: Response Time on 2x SSD RAID

Average response time: new order transaction

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

SI: saturated
3x lower

Peak Load tolerance: SIAS scales with ~540WHs
TPC-C: Throughput on 6x SSD RAID (Sylt)

New Order Transactions per Minute (NOTPM)

TPC-C on Sylt: Throughput (NOTPM)

Saturation

<table>
<thead>
<tr>
<th>NOTPM</th>
<th>500</th>
<th>800</th>
<th>1000</th>
<th>1200</th>
<th>1300</th>
<th>1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIAS-Chains</td>
<td>6424</td>
<td>10254</td>
<td>12693</td>
<td>13482</td>
<td>13375</td>
<td>13054</td>
</tr>
<tr>
<td>SI</td>
<td>6422</td>
<td>10113</td>
<td>10964</td>
<td>10553</td>
<td>10485</td>
<td>9294</td>
</tr>
</tbody>
</table>

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

- New Version Organization
- VID Optimized Index
- Selective Scan over VID_{map}
- Write Retention
- One Place invalidation
- Chaining
- Tuple Granularity Append Storage Manager
- Implementation in PostgreSQL

01.09.2017  |  Dr. Robert Gottstein  |
Thank You

Publications


- **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.