Computational Storage
Make Relational Databases Efficiently
Support Data Analytics

Yang Liu
Chief Architect, ScaleFlux Inc.
Agenda

- Computational Storage
- Analytical Workloads with Computational Storage
The Rise of Computational Storage

Computational Storage

- FPGA/GPU/TPU
- SmartNICs
- End of Moore’s Law
- 10 → 100-400Gb/s
- Fast & Big Data Growth
- Tremendous Value-Add to Applications & Infrastructure

- Scale performance with increasing workload capacities
- Optimize total infrastructure footprint
- Highly adaptable to evolving modern applications
Benefits of Computational Storage

- By putting computational engines near storage
  - More efficient heterogeneous computing
  - Better Performance (lower latency and higher throughput)
  - Reduced IO traffic
  - Offload CPU
  - Scalable performance with increasing workload capacities
  - Optimize total infrastructure footprint (TCO)
  - A lot more!
Computational Storage Unifies Fast & Big Data

- Low-Latency FAST Data Transactions
- Scale BIG Data analytics workloads with capacity
- Commodity infrastructure

Intensive Compute
Extract/Transform
Record Filtering
1. Computational storage with **in-line transparent compression**
Computational Storage for Database

2. Computational storage with **in-line data filtering**

![Diagram of computational storage with in-line data filtering]

- **Truly intelligent SQL processing**
- **Data filtering**
- **Decompression**
- **Flash Control**
- **NAND Flash**

The diagram illustrates the flow of data and processing components within a computational storage system, emphasizing inline data filtering and truly intelligent SQL processing.
Accelerating Analytical Workloads with Computational Storage
OLTP Vs. OLAP

- Database Workload
  - OLTP (Online Transactional Processing)
  - OLAP (Online Analytical Processing)

- NOT possible to do analytical processing on TP database, even for moderate level analytical query
  - Cache pollution / CPU contention / High system utilization
  - Unacceptable Quality of Service (QoS)!
ETL Complications

- ETL (Extraction, Transformation, Load)
  - Data formats changing over time
    - row based Vs. column based
  - Complicated to learn/maintain/integration
  - Cannot query on real time data
  - Concurrency (Isolation control)
  - Etc…
Database Pushdown

- Common database query
  - SELECT col1, col2, ... FROM table WHERE conditions
    - Projection: SELECT
    - Predicate (Filter): WHERE conditions

- Database pushdown:
  - Pushdown projection, predication or other query operations from SW query engine to SW storage engine
    - Storage Node in distributed database system
    - Storage controller for computational storage system
MyRocks Pushdown

- MySQL built-in support of **Engine Condition Pushdown** & **Index Condition Pushdown**
- Elegant RocksDB data structure/format ➔ Simplify hardware implementation

![Diagram showing the integration of MySQL, RocksDB, and hardware components]
HW Database Pushdown

- Implementation: MySQL 5.6 & RocksDB 5.18
- Support in-line Snappy decompression, filter conditions: =, !=, >, <, >=, <=, !Null, Null
- Engine condition pushdown (ECP): Direct comparison between non-indexed columns and constants inside computational storage devices
- Index condition pushdown (ICP): Condition evaluation on the index inside computational storage devices
HW Database Pushdown

Middle-range Xilinx KU15P 16nm FPGA

TBs 3D TLC NAND Flash

Flash Control (soft LDPC)

16 channels

2.4GB/s

Proc. Engine #1
Proc. Engine #2
Proc. Engine #3
Proc. Engine #4

3.2GB/s, 750K IOPS

PCIe

Performance benefit

Hardware utilization

Parallel & batch processing

2019 Storage Developer Conference. © ScaleFlux Inc. All Rights Reserved.
Implementation Considerations

- Block-by-block iterative processing
  ➔ Look-ahead batch processing
- Sync call
  ➔ Mixed sync/async call
- Error tolerance
  ➔ Recovery from computational errors and no impacts on normal storage IO
Data Block Format

- Streaming process any size of data blocks without DDRs
OLAP

CPU

MyRocks CPU Util %
Lower is better

CPU UTILS %

0 20 40 60 80 100

0 250 500 750 1000 1250 1500 1750 2000

Time (Seconds)

SFXDB CPU Util %
Lower is better

CPU UTILS %

0 20 40 60 80 100

0 250 500 750 1000 1250 1500 1750 2000

Time (Seconds)

I/O

MyRocks IO Util (MB/s)
Lower is better

IO UTILS MB/s

0 100 200 300 400 500 600 700 800 900

0 250 500 750 1000 1250 1500 1750 2000

Time (Seconds)

SFXDB IO Util (MB/s)
Lower is better

IO UTILS MB/s

0 100 200 300 400 500 600 700 800 900

0 250 500 750 1000 1250 1500 1750 2000

Time (Seconds)
Scalability

- **MyRocks**
  - CPU
  - 768GB DRAM
  - 8 host CPU cores, 100GB, 1 drive per DB instance
  - TPC-H Benchmark Q6

- Faster FPGA engine vs. CPU
  - 60% Faster!
  - 75% Faster!

- Scale Performance with Capacity
HTAP: Sysbench (TP) + TPC-H (AP)

<table>
<thead>
<tr>
<th># TPC-H Analytical Processing Threads</th>
<th>MyRocks (Sysbench TPS, TPS %)</th>
<th>SFXDB (Sysbench TPS, TPS %)</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sysbench TPS</td>
<td>TPS %</td>
<td>Sysbench TPS</td>
</tr>
<tr>
<td>0</td>
<td>16890.9</td>
<td>100.00%</td>
<td>16890.9</td>
</tr>
<tr>
<td>4</td>
<td>8932.0</td>
<td>52.88%</td>
<td>15038.81</td>
</tr>
<tr>
<td>8</td>
<td>6778.63</td>
<td>40.13%</td>
<td>14929.94</td>
</tr>
<tr>
<td>16</td>
<td>5687.97</td>
<td>33.67%</td>
<td>13619.02</td>
</tr>
<tr>
<td>24</td>
<td>4605.07</td>
<td>27.26%</td>
<td>12675.46</td>
</tr>
<tr>
<td>32</td>
<td>3637.12</td>
<td>21.53%</td>
<td>13300.13</td>
</tr>
</tbody>
</table>

- When there is no TPC-H AP workload on the system, the TPS performance is the same with or without computational storage.
- As the TPC-H AP workload starts to increase, the Sysbench TPS benchmark is negatively impacted in the case where analytical TPC-H queries are executed on the host x86 CPU.
HTAP | CPU | Computational Storage

**Compute**

**I/O**
Summary & Call to Action

- Computational Storage is in production!
- Database: One ideal target of computational storage
  - Successful demonstration with MyRocks on computational storage
- Very large design space to be explored
  - Cross-layer innovation across software and hardware
  - Close collaboration across industry sectors
Thank You!

Deploying Computational Storage at Scale
• Bullet one
  • Bullet two
    • Bullet 3
      • Bullet 4