XDP-Rocks: KV Storage, Revolutionized
XDP-Rocks: KV Storage, Revolutionized

Edward Bortnikov, VP Technology, Pliops
Key-Value Storage Engines on the Rise

KV Abstractions

Put(K,V)
Get(K)
Delete(K)
...

KV Storage Engines

Used By

[Images and logos of various KV storage engines and tools]
Challenges and Tradeoffs

- Write/Read/Space Amplification
- Compute Overhead
- Index Footprint
Use Case - RocksDB

וסט-אוןprésentation-

**State-of-the-art Embedded KV Storage Engine**
- Optimized for writes - Log Structured Merge (LSM) Tree storage
- Optimized for reads – Binary Search index, Bloom filters, Caching, etc.

... and yet, facing multiple challenges, especially with flash storage
- Compaction → write amplification → low throughput, high jitter, SSD wear
- Compaction → data copy, sort/merge, compression → high CPU overhead
- Compaction → space amplification → storage overprovisioning
- Uncompressed index → low cache HR → read amplification → high tail latency
Data Accelerators Come to Help

Domain-Specific Processor (ASIC or FPGA)

KV Centric Architecture
A native design, not an afterthought
SSD space inherently managed as a KV collection

Storage Access Bottlenecks Offload
Search, Indexing, Compression, GC, etc.

Failure protection
Storage media and power failures
XDP™ - Extreme Data Processor

PCIe Data Access Accelerator for SSD storage

Technology  XDP 1.0 FPGA, XDP 2.0 ASIC (upcoming).
Low profile    HHHL, avg power <35W, max power 49W, OAM FF (long term)
High reliability  Temperature 10-52°C @ 250 LFM, MTBF 4.5M hours

Customer Value

Performance  3.2M IOPS RR, 1.2M IOPS RW, 45GB/s SR, 6.4 GB/s SW.
Capacity     Line rate compression - LZ4, ZSTD^{ASIC}.
Protection   Zero-impact RAID5, power failure reliability.
Endurance    7x SSD lifetime extension.
XDP™ - a Native KV Storage Platform

Core KV Abstractions in Silicon
    Get/Put/Delete, Multi-DB (up to 1K shards)

Direct KV Access
    Bypassing the heavy OS storage stack.

Performance
    Low Write Amp  log structured storage, native GC.
    Low Read Amp    fine grained, compressed (<3B/value) KV index.
                    zero false positives → single-I/O access.
    Low CPU Footprint  index access completely in HW.
XDP-Rocks™ – an Accelerated KV Engine

Applications expect rich storage engine API’s
RocksDB is a de-facto KV standard.

XDP Rocks - the best of all worlds

**RocksDB API**  **drop-in compatibility** for developers.

**XDP performance** 10x’s throughput↑, 100x tail latency↓.

Implementation
Based on RocksDB 7.0 OSS.
Extensive re-write of internals to work with XDP.
W Microbenchmark

Setting: 16 DB's (~7 TB data), 4 threads, 16B key, 1KB value
W Microbenchmark – Dynamics

Setting: 1 DB (~7TB data), 1 thread, 16B key, 1KB value

Average: 149.6 kops
Stdev: 3.6 kops (2.3%)

Average: 22.9 kops
Stdev: 7.3 kops (31.8%)
R & W Microbenchmarks - Scaling w/Value

Put

<table>
<thead>
<tr>
<th>Size</th>
<th>RocksDB</th>
<th>XDP-Rocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>10.7x</td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>20.3x</td>
<td></td>
</tr>
<tr>
<td>4096</td>
<td>27.1x</td>
<td></td>
</tr>
<tr>
<td>16384</td>
<td>36.3x</td>
<td></td>
</tr>
</tbody>
</table>

Get

<table>
<thead>
<tr>
<th>Size</th>
<th>RocksDB</th>
<th>XDP-Rocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>5.7x</td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>3.9x</td>
<td></td>
</tr>
<tr>
<td>4096</td>
<td>2.8x</td>
<td></td>
</tr>
<tr>
<td>16384</td>
<td>1.8x</td>
<td></td>
</tr>
</tbody>
</table>
R and Mixed Microbenchmarks

Setting: 32 DB’s (~5TB data), 16B key, 1KB value

100% Get
48-1024 threads, post massive writes

50% Get, 50% Put
48-1024 threads

6.8x throughput, 0.24x CPU
5.9x throughput, 0.22x CPU
KVRocks – W Benchmark

OSS distributed NoSQL database
Compatible with the Redis protocol
RocksDB as storage engine

Widely popular
Baidu, Ctrip, Meitu, Baishan Cloud

We ported it to use XDP-Rocks
… and compared the performance

Setting: 32 DB’s (~7TB data), 32 clients, 8B key, 4KB value

Throughput, kops

<table>
<thead>
<tr>
<th></th>
<th>Fill (SET)</th>
<th>Update (SET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XDP-Rocks</td>
<td>4.7x</td>
<td>10.7x</td>
</tr>
<tr>
<td>RocksDB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KVRocks – R and Mixed Benchmarks

100% Read (GET)

50% Read (GET), 50% Write (SET)
Meeting the OCP Tenets

Openness
- XDP HW – PCIe compliant, supports all NVMe drives, OCP friendly (OAM) long term
- XDP Rocks SW - committed to open sourcing, community engagement welcome!

Efficiency
- Performance translates to 3x-10x TCO savings (CapEx and OpEx).

Impact
- Applicable to a broad range of data management systems

Scale
- In production at Fortune 100 SaaS provider.
Wrapping Up

Data accelerators can revolutionize KV access to flash storage.

**XDP™**
- High-performance, reliable native KV storage engine.

**XDP-Rocks™**
- RocksDB API + XDP performance.

**Microbenchmark- and application-level gains**
- Up to 10x’s in throughput. Up to 100x in tail latency.
Let’s Engage!

Try out our accelerator cards demo@pliops.com

Sign up for notifications on XDP-Rocks xdp-rocks@pliops.com

Learn more https://pliops.com/
Thank you!

EMPOWERING OPEN.

OCTOBER 18-20, 2022
SAN JOSE, CA
The Write Path – KV Separation

**RocksDB**
Fat LSM, High WA

- Memtable
  - SST file
  - Flush
  - Compaction

**XDP-Rocks**
Slim LSM, Low WA

- Memtable
  - SST file
  - Flush
  - Compaction
  - Put/ Delete

Core KV Store

*Challenge: advanced read semantics (snapshot) support*
The Read Path – LSM Bypass

RocksDB
LSM read, Low/Medium RA

XDP-Rocks
LSM bypass, Very low RA

Challenge: advanced read semantics (snapshot) support