Why Cache in DRAM if you have NVMe-oF?

Radu Stoica, Nikolas Ioannou, Kornilios Kourtis
IBM Research Zurich

Santa Clara, CA August 2019
The architecture of a web application

Benefits:

- Scalability & elasticity
- Tolerance to failures
- Develop, deploy & manage layers independently

DRAM caching drawbacks:

- X Cost
- X Underutilization
- X Limited size
- X Cold start

~10ms
~50µs
~25µs
~1ms
~25µs
~1ms
~25µs
~100µs
25µs
~100ms

Total response time: ~1s
Trends in high-performance storage

1. Advances in non-volatile storage:
   • NAND Flash – highest improvement in density, IOPS, and cost
   • New non-volatile technologies (3DXP, Z-NAND, XL-FLASH)

2. Advances in software and protocols:
   • NVMe & NVMe-oF enables fast access to NVM storage
   • OS bypass for data path (SPDK)

Idea: replace the DRAM cache with NVMe-oF storage
NVMe-oF caching

End users

Load balancers

Frontend (HTTP servers)

Application servers (backends)

NVMe cache (Memcached compatible)

Data persistency layer (object store, HA database)

Disaggregation benefits:
- Scale compute independently from storage
- Better elasticity
- Lower resource waste

Cost benefits:
- Replace DRAM with NVMe storage
- Service other workloads on the same HW
- Data reduction @ no performance loss

Management benefits:
- One way to provision storage
- Redundancy reduces failure rate
- Unlimited capacity
Existing caching systems

DRAM-based systems (Memcached, Redis) are not designed to support persistent storage

Storage-based caching solutions are too slow:

- Built for slower devices (e.g., use synchronous IO)
- Data structures with inherent IO amplification (LSM- or B-trees)
- Cache data in DRAM, limiting scalability
- Rich feature set (e.g., transactions, snapshots)

Faster HW not enough, software needs to change
uDepot* architecture

- Application (uDepot client)
- Application (memcache client)

- Application (Embedded use)
- uDepot server
- memcache server

- Low overhead
- Small memory footprint
- Persistency

- Drop-in replacement for Memcached

Local storage
- Block device
- NVMe

Remote storage
- FC-NVMe
- NVMe-oF
- iSCSI

- Scale up to multiple devices, multiple cores, and PBs of data
- Delivers the low latency of new classes of NVM (3DXP, Z-NAND)

*Reaping the performance of fast NVM storage with uDepot*, Kornilios Kourtis et al., FAST 2019
Proof-of-concept deployment

- Standard Memcached benchmark
- 95% reads, 5% writes
- 1KB entries

- uDepot using Linux AIO
- Default FC-NVMe driver

- Redundant network paths

IBM FlashSystem® 900
- 180-400TB in 2U
- Redundancy & availability
- Compression
- Encryption

<table>
<thead>
<tr>
<th></th>
<th>Read</th>
<th>Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency (µs)</td>
<td>155</td>
<td>95</td>
</tr>
<tr>
<td>IOPS (millions)</td>
<td>1.1</td>
<td>0.6</td>
</tr>
<tr>
<td>Bandwidth (GB/s)</td>
<td>10</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Santa Clara, CA August 2019
Throughput & latency comparison

Similar throughput
Up to 400TB cache size

uDepot has ~160µs higher latency

Latency increase due to overload

Elasticity > storage medium!
Fast NVMe storage offers opportunities to replace DRAM, but existing data store technologies fail to match their performance.

We demonstrate the DRAM-performance of a system composed of:

- uDepot: a Memcached drop-in replacement that delivers storage performance
- IBM FlashSystem® 900 (155us latency, 10GB/s throughput, NVMe-ready)

Benefits:

- Disaggregation
- Cost reduction
- Simplified management
Thank You!

Questions?