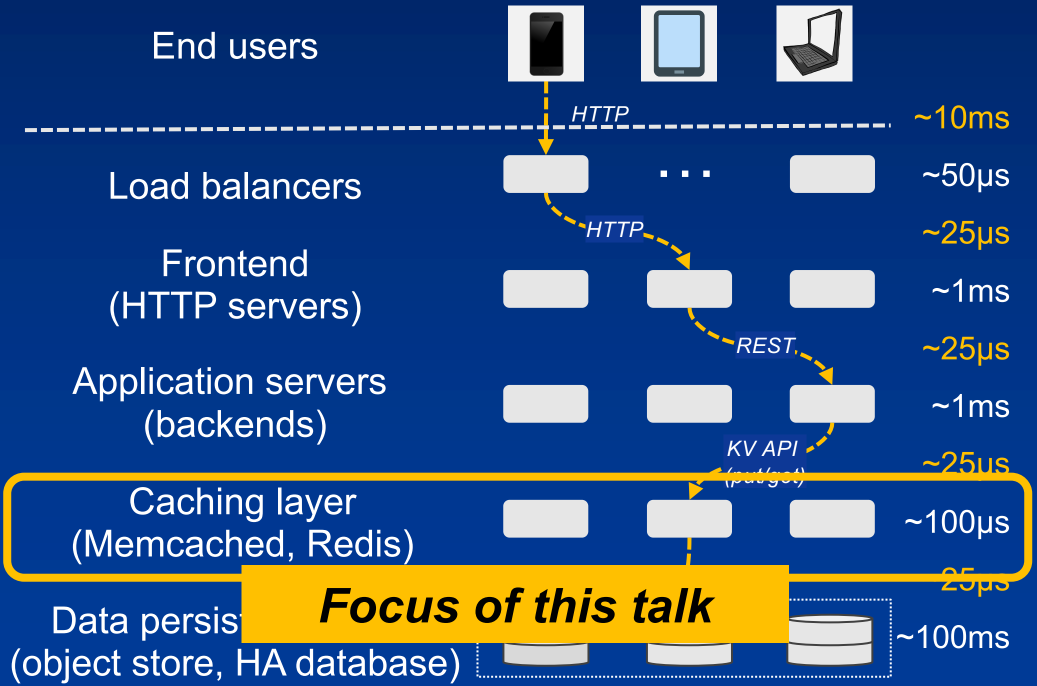
Why Cache in DRAM if you have NVMe-oF?

Radu Stoica, Nikolas Ioannou, Kornilios Kourtis
IBM Research Zurich

The architecture of a web application



Total response time: ~1s

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

DRAM caching drawbacks:
- ✗ Cost
- ✗ Underutilization
- ✗ Limited size
- ✗ Cold start

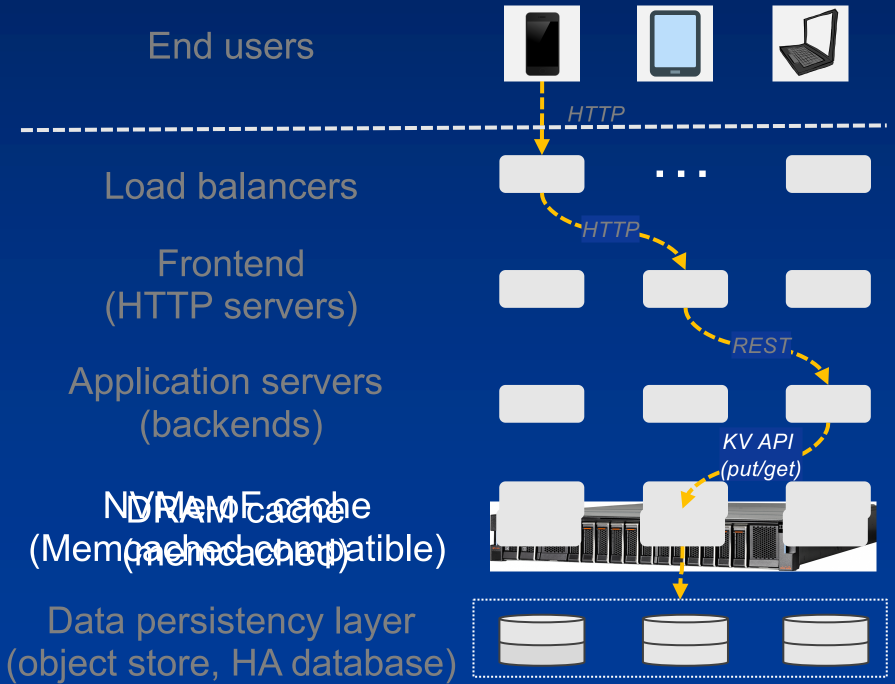
Trends in high-performance storage

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

Flash Memory Summit

NVMe-oF caching



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

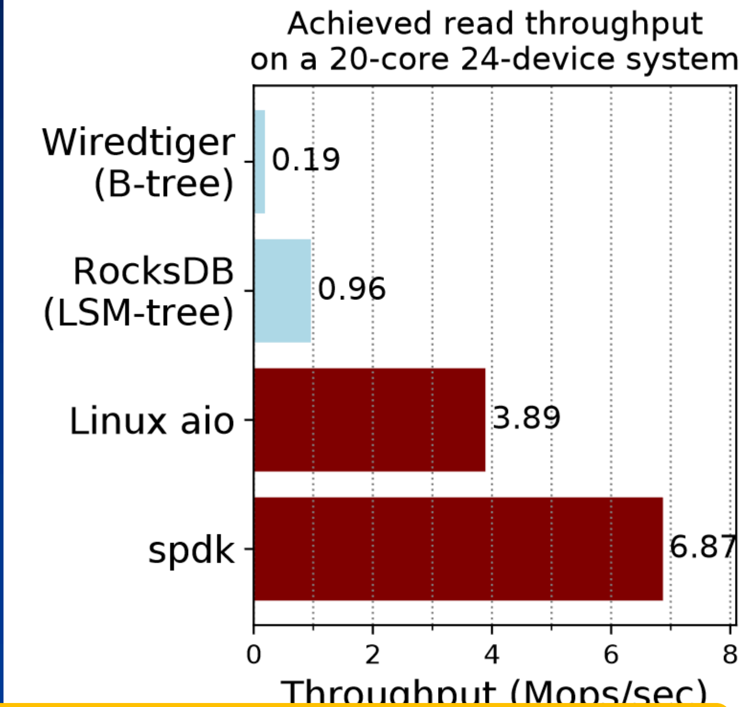
Flash Memory Summit

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

uDepot KV store

- ✓ *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

Flash Memory Summit

Proof-of-concept deployment

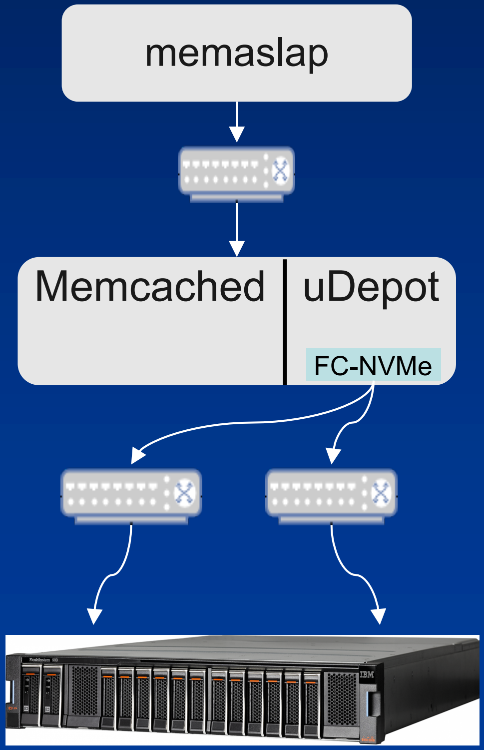
Caching client

10 GbE network

Caching server

16 Gbps FC network

FC-NVMe storage



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

- uDepot using Linux AIO
- Default FC-NVMe driver

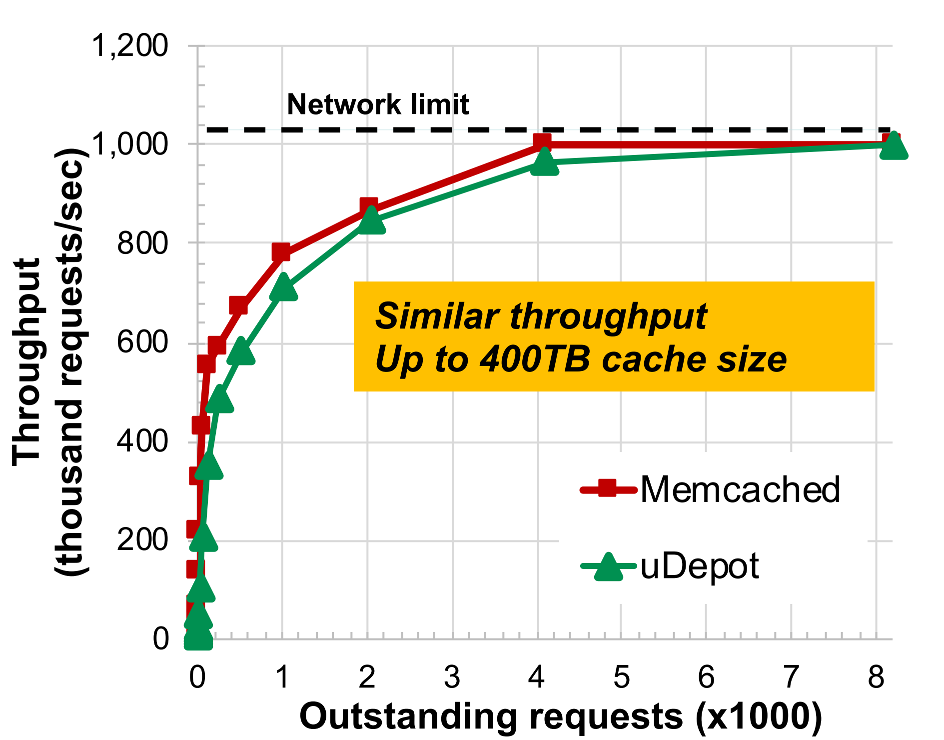
- Redundant network paths

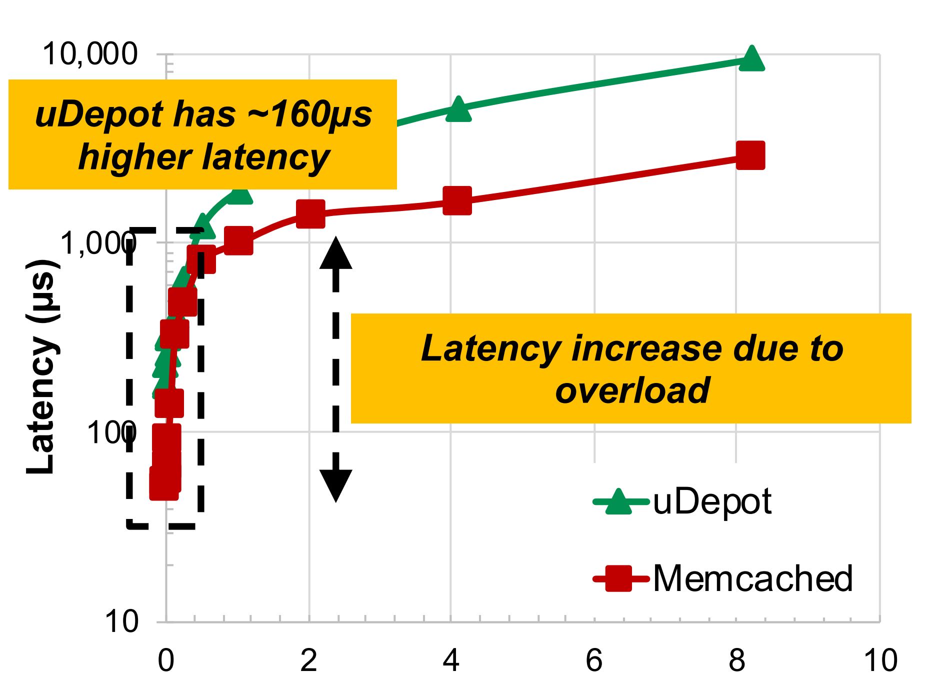
IBM FlashSystem® 900

- *180-400TB in 2U*
- *Redundancy & availability*
- *Compression*
- *Encryption*

	Read	**Write**
Latency (μs)	155	95
IOPS (millions)	1.1	0.6
Bandwidth (GB/s)	10	4.5

Throughput & latency comparison





Elasticity > storage medium!

Flash Memory Summit

Summary

- 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 ?

www.research.ibm.com/labs/zurich/cci/