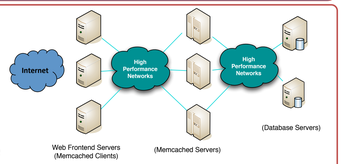
Designing High-Performance, Resilient and Heterogeneity-Aware Key-Value Storage for Modern HPC Clusters

Dipti Shankar (shankar.50@osu.edu)

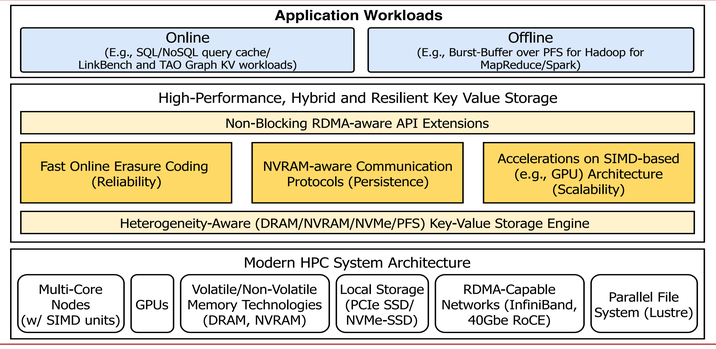
Advisor: Dhabaleswar K. (DK) Panda (panda.2@osu.edu), Co-Advisor: Dr. Xiaoyi Lu (lu.932@osu.edu)

Introduction

- Key-Value Stores (e.g., Memcached) serve as the heart of many production-scale distributed systems and databases
- Accelerating Online and Offline Analytics in High-Performance Computing (HPC) environments
- Our Basis: High-performance and hybrid key-value storage
  - Remote Direct Memory Access (RDMA) over high-performance network interconnects interconnects (e.g., InfiniBand, RoCE)
  - 'DRAM+NVMe/NVRAM' hybrid memory designs
- Research Focus: Designing a high-performance key-value storage system that can leverage: (1) RDMA-capable networks (2) heterogeneous I/O and (3) compute capabilities on HPC clusters
- Goals: (1) End-to-end performance (2) Scalability (3) Resilience / High Availability

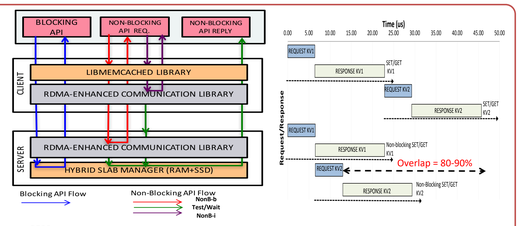


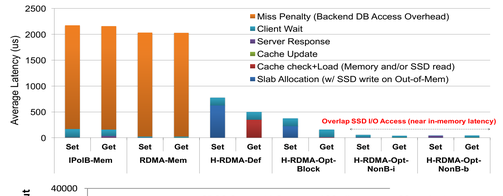
Research Framework

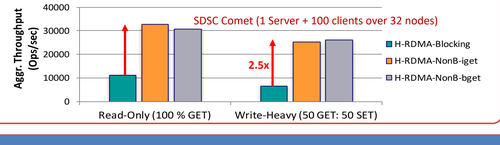


High-Performance Non-Blocking API Semantics

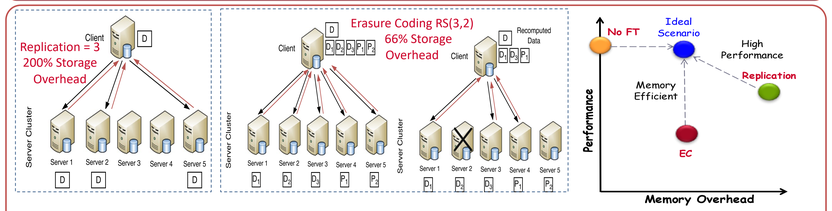
- Motivation: Hybrid `DRAM+PCIe/ NVMe-SSD' Key-Value Stores
  - Higher data retention; fast random reads
  - Performance limited by Blocking API semantics
- Goals: Achieve near in-memory speeds while being able to exploit hybrid memory
- Approach: Novel Non-blocking API Semantics
  - Extensions for RDMA-based Libmemcached library
  - memcached_(iset/bset/bget) for SET/GET operations
  - memcached_(test/wait) for progressing communication
  - Ability to overlap request and response phases; hide SSD I/O overheads
  - Up to 8x gain in overall latency vs. blocking API semantics



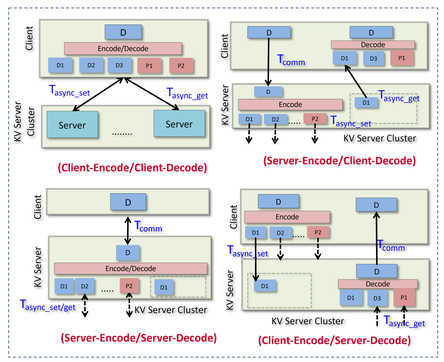


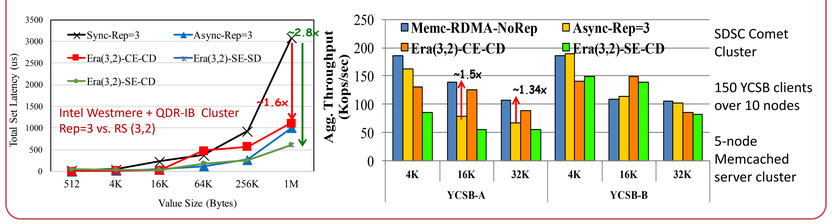


Fast Online Erasure Coding with RDMA

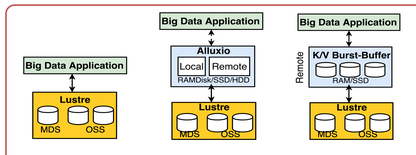


- Erasure Coding (EC): Storage-efficient alternative to Replication for Resilience
- Goal: Making Online EC viable for key-value stores
- Bottlenecks: (1) Encode/decode computation (2) Scattering/gathering the data/parity chunks
- Approach: Non-blocking RDMA-aware semantics to enable compute/communication overlap
- Encode/Decode offloading integrated into Memcached client (CE/CD) and server (SE/SD)
- Experiments with Yahoo! Cloud Serving Benchmark (YCSB) for Online EC vs. Async. Rep: (1) Update-heavy: CE-CD outperforms; SE-CD on-par (2) Read-heavy: CE-CD/SE-CD on-par

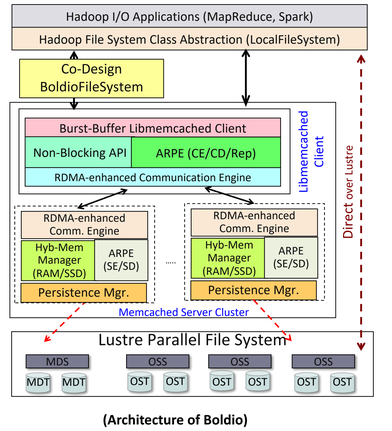


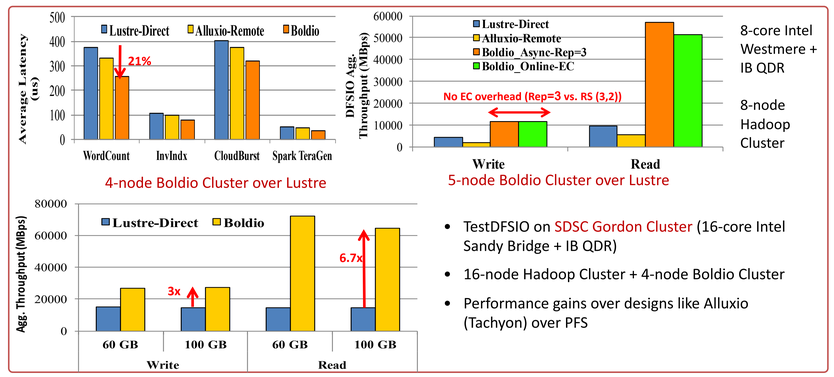


Co-Designing Key-Value Store-based Burst Buffer over PFS



- Big Data I/O infrastructure (e.g., HDFS, Alluxio) vs. HPC Storage (e.g., GPFS, Lustre); limited 'data locality' in HPC
- Bottleneck: Heavy reliance on PFS (limited I/O bandwidth); affects data-intensive Big Data applications
- Approach: High-Performance and Hybrid and Resilient Key-Value Store-Based Burst-Buffer Over Lustre (Boldio)
  - Hadoop I/O over Lustre; transparent FileSystem plugin for Hadoop MapReduce/Spark
  - No dependence on local storage at compute nodes
  - Resilience via Async. Replication or Online EC
  - RDMA-Memcached as Burst-Buffer servers + Non-blocking client APIs for efficient I/O pipelines

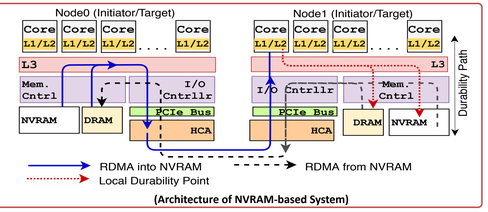




- TestDFSIO on SDSC Gordon Cluster (16-core Intel Sandy Bridge + IB QDR)
- 16-node Hadoop Cluster + 4-node Boldio Cluster
- Performance gains over designs like Alluxio (Tachyon) over PFS

Exploring Opportunities with NVRAM and RDMA

- Emerging non-volatile memory technologies (NVRAM)
  - Potential: Byte-addressable and persistent; capable of RDMA
  - Observations: RDMA writes into NVRAM needs to guarantee remote durability
  - Opportunities: RDMA-based Persistence Protocols for NVRAM Systems



Conclusion and Future Work

- The proposed framework enables key-value storage systems to exploit the capabilities of HPC clusters for maximizing performance and scalability, while ensuring data resilience/availability.
- Provides efficient non-blocking API semantics to design efficient read/write pipelines with resilience via RDMA-aware asynchronous replication and fast online EC
- Future work for this thesis: Works-in-Progress
  - Explore opportunities for exploiting the SIMD compute capabilities (e.g., GPU, AVX); End-to-end SIMD-aware key-value storage system designs
  - Work on co-designing memory-centric data-intensive applications over key-value stores: (1) Read-Intensive Graph-based Workloads (e.g., LinkBench, RedisGraph) (2) Key-value store engine for Parameter Server frameworks for ML workloads

References

[1] **D. Shankar ,** X. Lu , and D. K. Panda**,** "High-Performance and Resilient Key-Value Store with Online Erasure Coding for Big Data Workloads", 37th International Conference on Distributed Computing Systems (**ICDCS 2017**)

[2] **D. Shankar** , X. Lu , and D. K. Panda, "Boldio: A Hybrid and Resilient Burst-Buffer Over Lustre for Accelerating Big Data I/O", 2016 IEEE International Conference on Big Data (**IEEE BigData 2016**) [Short Paper]

[3] **D. Shankar**, X. Lu , N. Islam , M. W. Rahman , and D. K. Panda, "High-Performance Hybrid Key-Value Store on Modern Clusters with RDMA Interconnects and SSDs: Non-blocking Extensions, Designs, and Benefits", 30th IEEE International Parallel & Distributed Processing Symposium (**IPDPS 2016**)

[4] **D. Shankar**, X. Lu , M. W. Rahman , N. Islam , and D. K. Panda, "Benchmarking Key-Value Stores on High-Performance Storage and Interconnects for Web-Scale Workloads", 2015 IEEE International Conference on Big Data (**IEEE BigData '15**) [Short Paper]

[5] **D. Shankar**, X. Lu , J. Jose , M. W. Rahman , N. Islam , and D. K. Panda, "Can RDMA Benefit On-Line Data Processing Workloads with Memcached and MySQL", 2015 IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS 2015**) [Poster]

Software Distribution

- The RDMA-based Memcached and Non-Blocking API designs (RDMA-Memcached) proposed in this research are available for the community as a part of the HiBD project http://hibd.cse.ohio-state.edu/#memcached
- Micro-benchmarks and YCSB plugin for RDMA-Memcached available in as a part of of the OSU HiBD Micro-benchmark Suite (OHB) http://hibd.cse.ohio-state.edu/#microbenchmarks

Acknowledgements

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu

HiBD High-Performance Big Data (HiBD)
High-Performance Big Data
http://hibd.cse.ohio-state.edu

This research is supported in part by National Science Foundation grants #CNS-1513120, #IIS-1636846, and #CCF-1822987