Mojim: A Reliable and Highly-Available Persistent Memory System

Yiying Zhang, Jian Yang, Amirsaman Memaripour, Steven Swanson

Problem

Traditional data-replication schemes are designed for disk-based data

Too slow for next-generation non-volatile main memory (NVMM)

- Heavy protocol and software
- I/O based instead of memory access

Consequence

Expensive NVMMs produce little performance improvement

Need to reconsider data replication for NVMM!

Mojim Solution

Memory-to-memory replication

Flexible Modes

- Provide different levels of reliability, availability, consistency, and $ cost

Atomic Support

Results

Mojim replication can even be faster than no replication!

29% - 72% latency

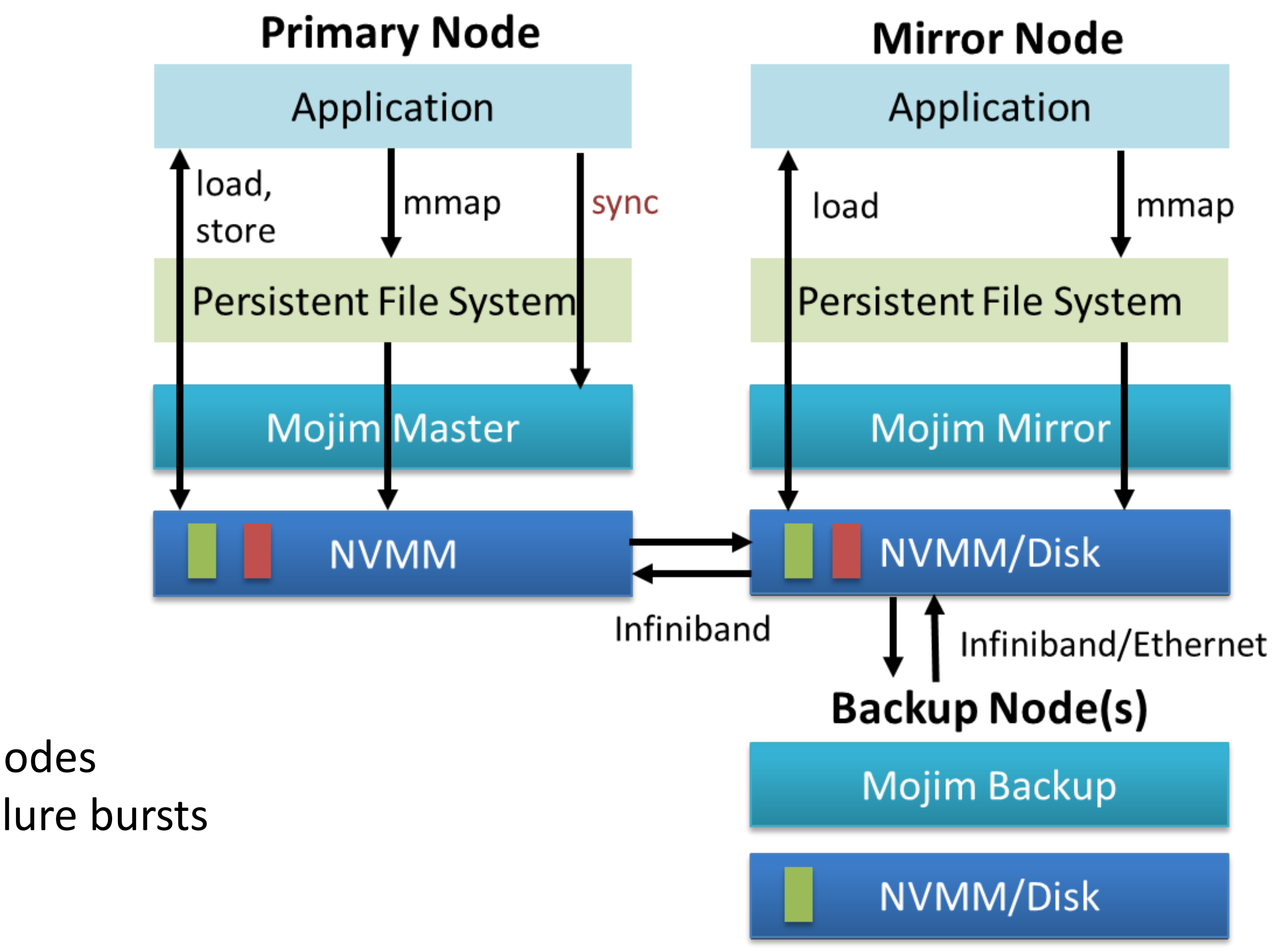
0.5 – 3.5x throughput

Mojim improves current replication schemes by up to 42x

Mojim Architecture

Primary Tier

- Pair of mirrored nodes
- Ensures good performance
- RDMA protocol with minimal software overhead
- Fast fail-over to mirror node

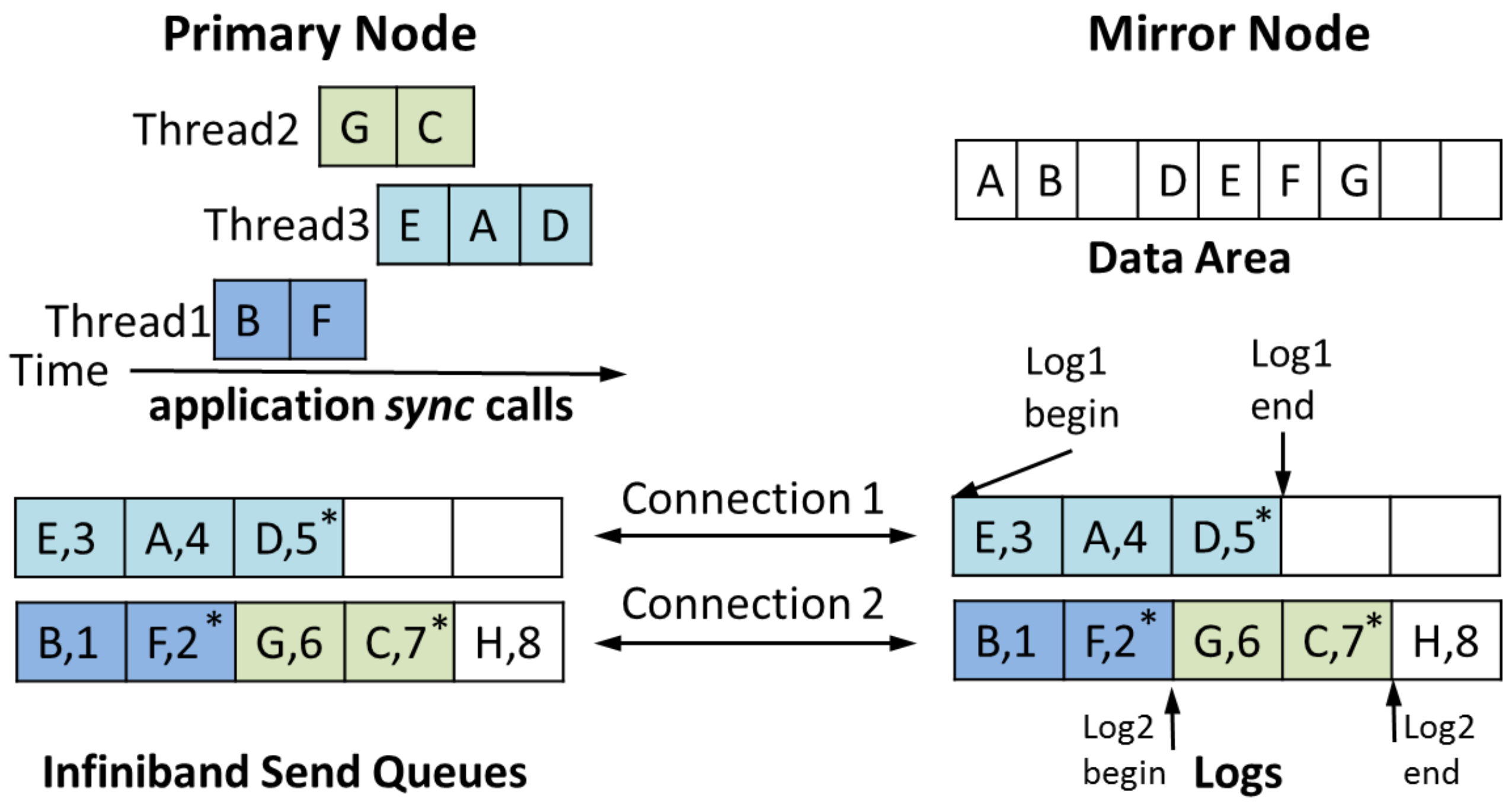


Secondary Tier

- Optional one or more backup nodes
- More redundancy to sustain failure bursts
- Replicate data in background
- Low $ cost option

Implementation

- Implemented in the Linux kernel
- In-kernel RDMA protocol
- Use logs and page tables to support atomic operations
- Fast recovery



Mojim Modes

	Scheme	Performance	Reliability	Availability	Consistency	$ Cost
	Un-replicated	Good	0	Worst	N/A	Low
Mojim	Async	Good	1	Good	Weak	Fair
	Sync	Good	1	Good	Strong	Fair
	Sync-disk	Good	1	OK	Strong	Low
	Sync-two-tier	Good	N-1	Best	Strong+Weak	High
	Sync-twotier-ETH	Bad	N-1	Good	Strong+Weak	Fair
Existing	Write-all	Bad	N-1	Best	Strong	High
	Chain-rep	OK	N-1	Best	Strong	High
	Broadcast-rep	OK	N-1	Best	Strong	High

Existing replication schemes:
Write-all: allow write to all nodes with strong consistency
Chain-rep: write to primary, serialize replication to secondaries
Braodcast-rep: write to primary, broadcast replication to secondaries

Evaluation

Research Questions

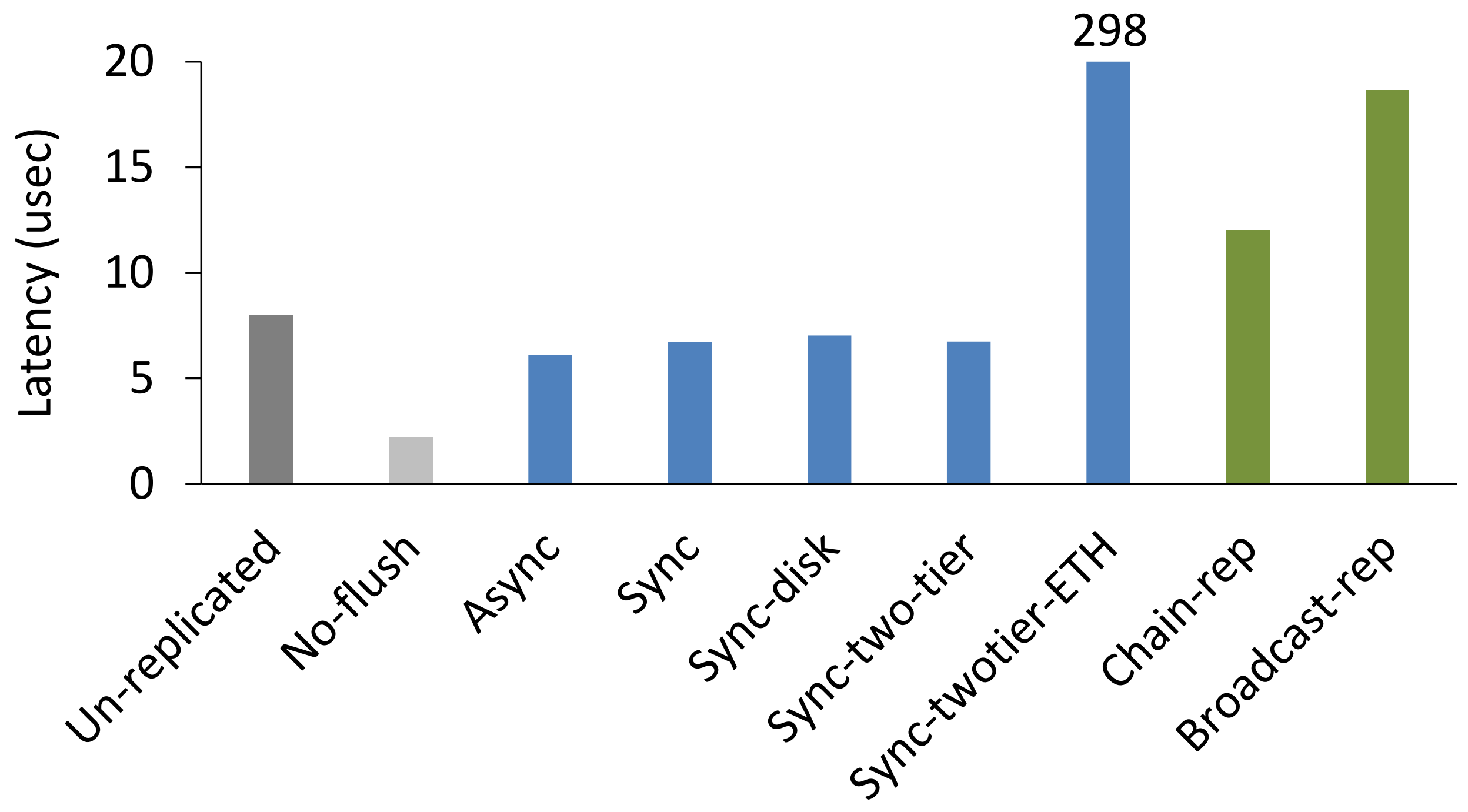
- What is the performance of different Mojim modes?
- How does Mojim compare with other replication methods?
- What is the performance of Mojim with real applications?

Environment

- 40 Gbps Infiniband and 1 Gbps Ethernet
- DRAM as a stand-in for NVMM

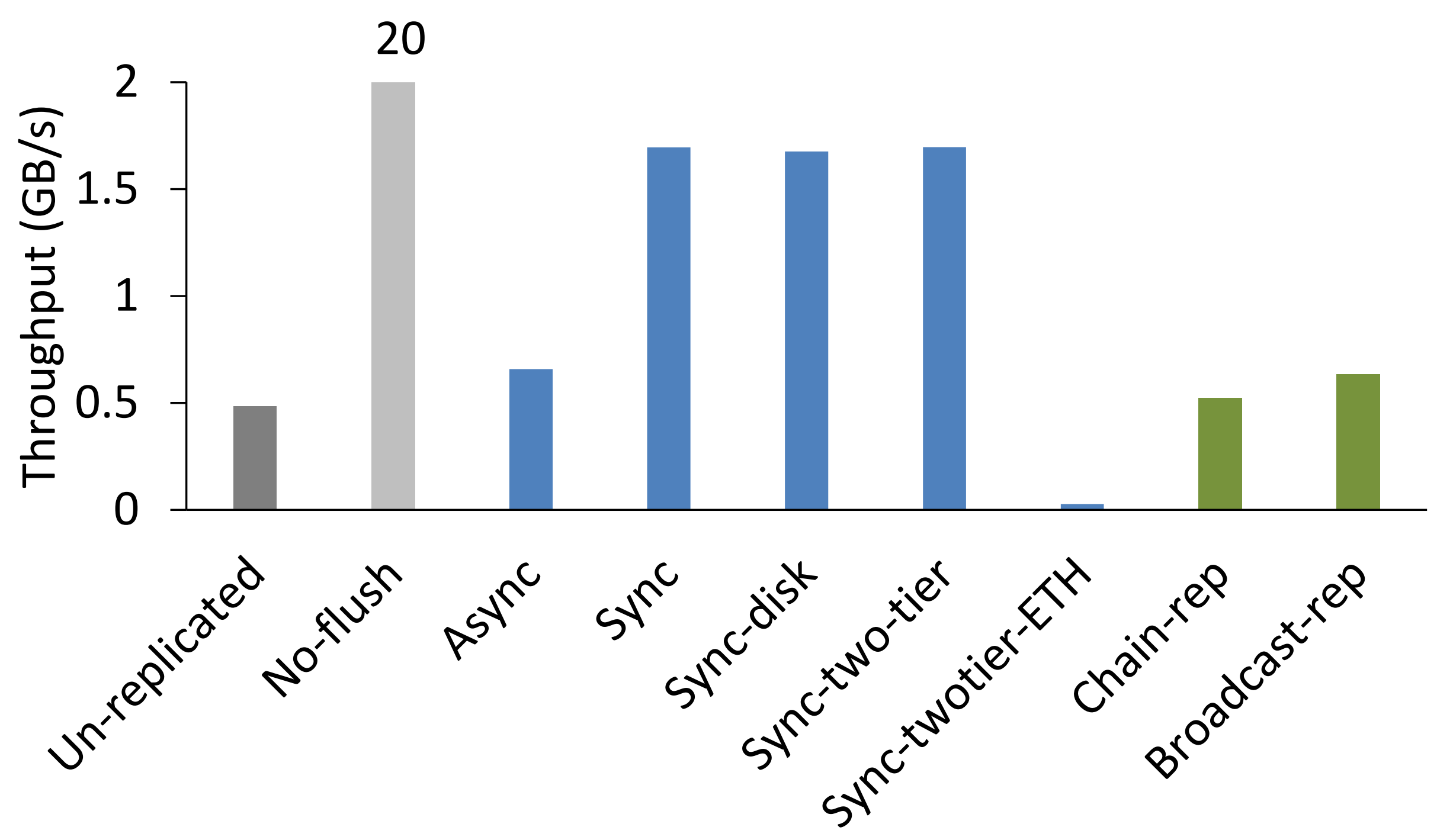
Avg msync Latency

- Random 4KB *msync* calls in a 4GB *mmap*'d file
- no-flush: un-replicated without CPU cache flushes

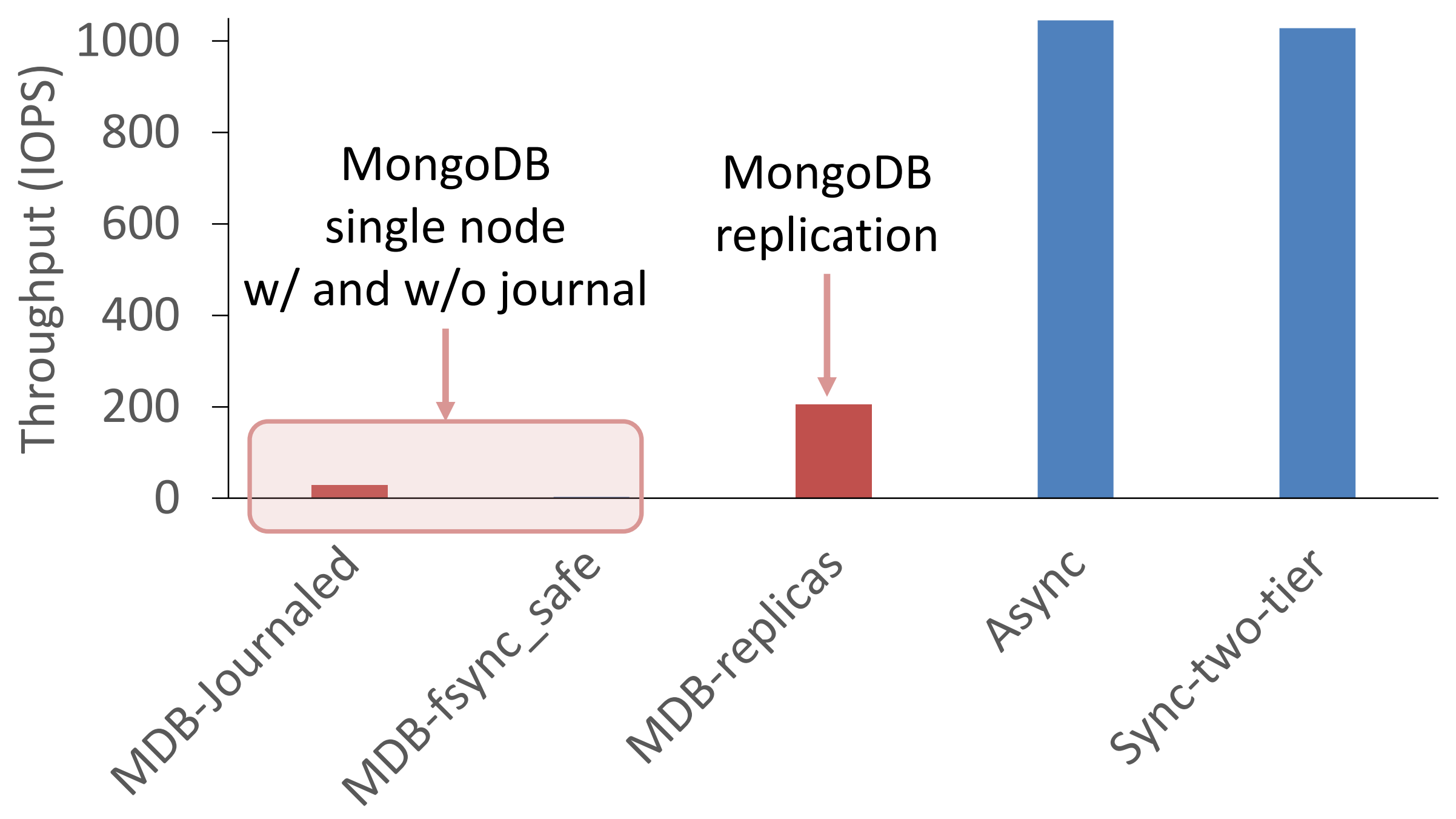


msync Throughput

- Random 12KB *msync* calls in a 4GB *mmap*'d file



MongoDB Key-Value Pair Load


Throughput (IOPS)
MongoDB single node w/ and w/o journal
MongoDB replication
MDB-Journaled
MDB-fsync_safe
MDB-replicas
Async
Sync-two-tier