EXPODB: EFFICIENT TRANSACTION PROCESSING IN BYZANTINE FAULTTOLERANT ENVIRONMENTS

Suyash Gupta, Jelle Hellings, Thamir Qadah, Sajjad Rahnama, Mohammad Sadoghi

18th International Workshop on
Transaction High Performance Transaction Systems (HPTS)
November 3-6, 2019

Mohammad Sadoghi
Exploratory Systems Lab
Department of Computer Science
UCDAVIS
UNIVERSITY OF CALIFORNIA

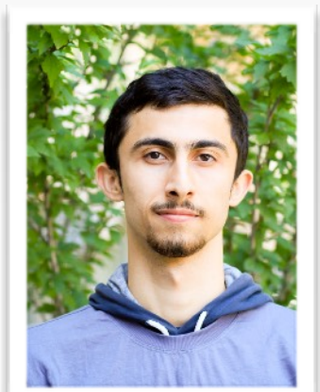
ExpoLab
Creativity Unfolded

ExpoLab
Creativity Unfolded

Mohammad Sadoghi
(Principal Investigator)

Jelle Hellings, PostDoc
(Blockchain)

Suyash Gupta, PhD
(Blockchain)

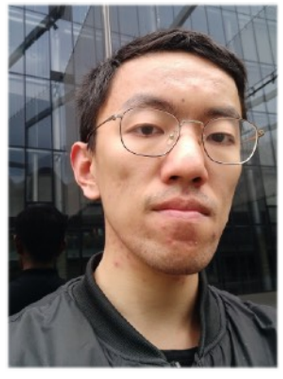


Sajjad Rahnama, PhD
(Blockchain)

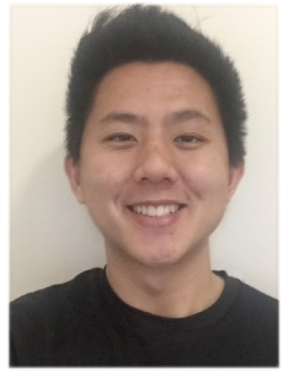
Thamir Qadah, PhD
(Coordination-free Concurrency)

Masoud Hemmatpour, PhD
(RDMA KV-Stores)

Domenic Cianfichi, MSc
(Blockchain)



Robert He, MSc
(Coordination-free Concurrency)

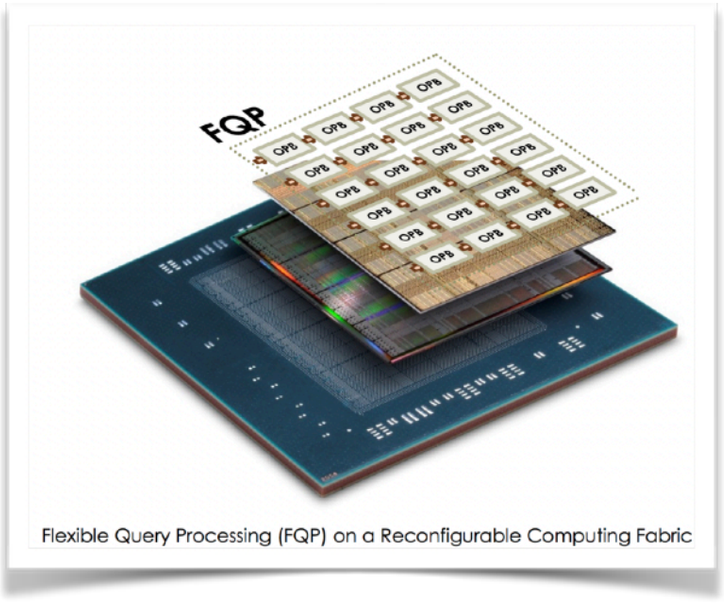


Patrick Liao, BSc
(Blockchain)

Shreenath Iyer, MSc
(Blockchain)

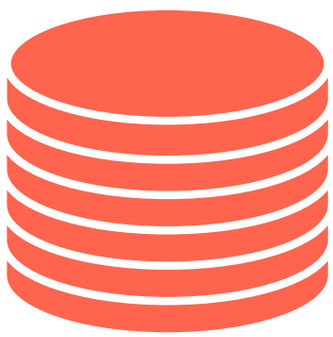
Journey…

SQL
Analytics



FPGA Acceleration: FQP (Flexible Query Processor)
[VLDB'10, ICDE'12, VLDB'13, ICDE'15, SIGMOD Record'15, ICDE'16, USENIX ATC'16, ICDCS'17, ICDE'18, TKDE'19]

Journey…

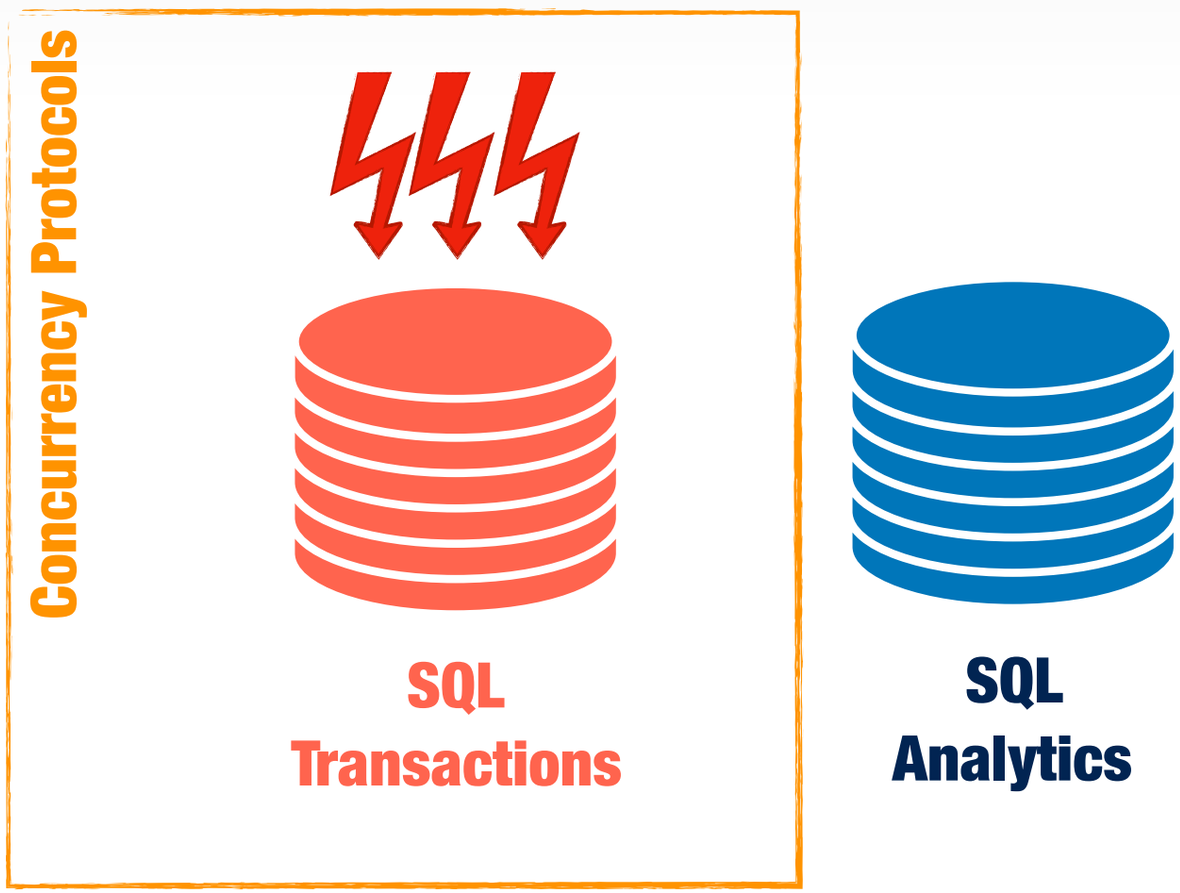


SQL
Transactions

SQL
Analytics

High-dimensional Indexing: (e.g., BE-Tree, BE-topK)
[SIGMOD'11, ICDE'12, TODS'13, ICDCS'13, ICDE'14, ICDCS'17, Middleware'17]

Journey…

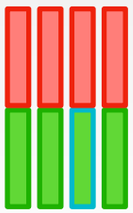


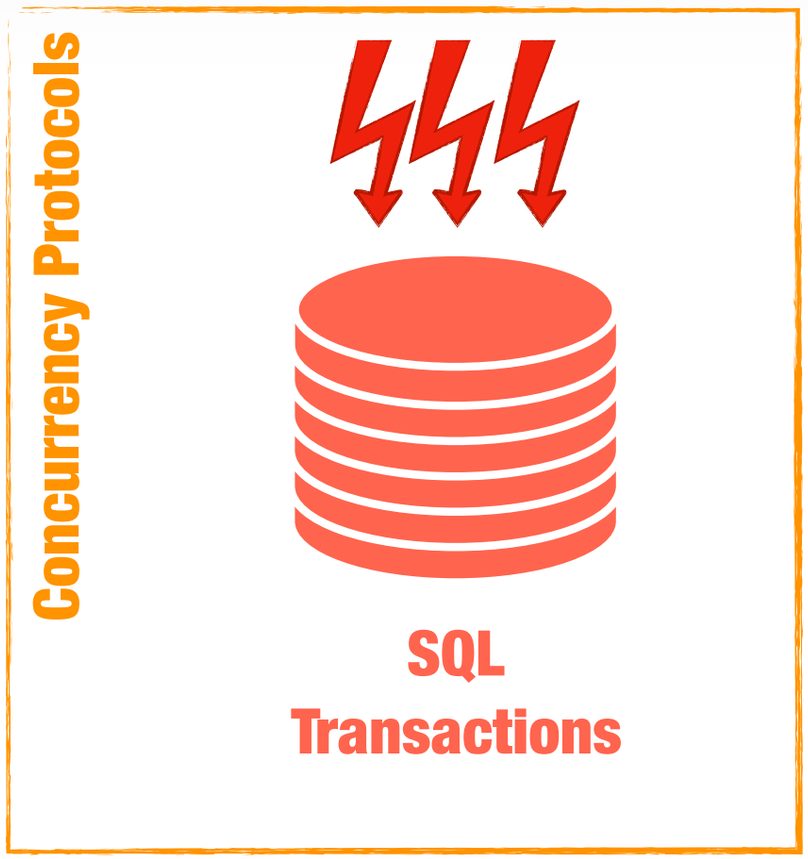
Concurrency Control Protocols: (e.g., 2VCC, QueCC - Best Paper Award)

[VLDB'13, VLDB'14, VLDBJ'16, Middleware'16, TDKE'15, SIGMOD'15, ICDE'16, Middleware'18]

Journey…

QueCC: Queue-Oriented Planning and Execution Architecture



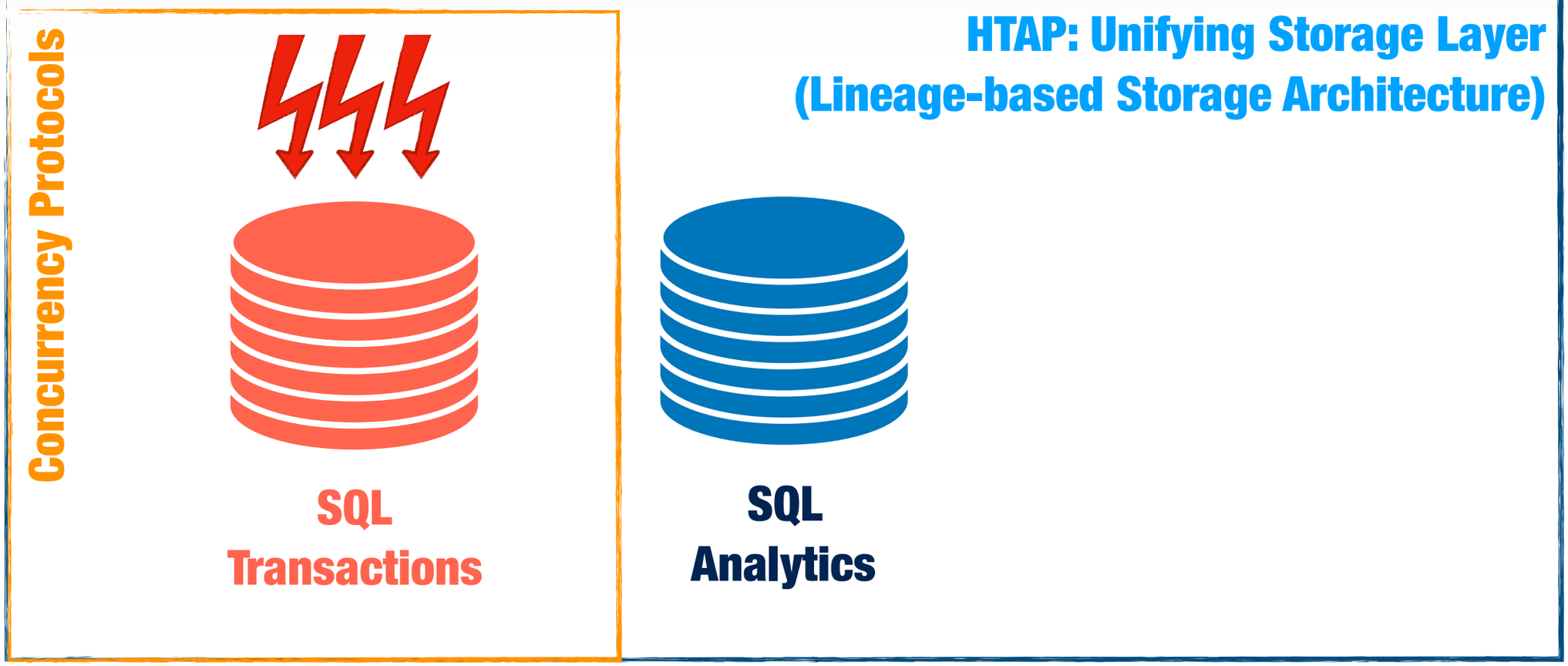


SQL Analytics

Concurrency Control Protocols: (e.g., 2VCC, QueCC - Best Paper Award)

[VLDB'13, VLDB'14, VLDBJ'16, Middleware'16, TDKE'15, SIGMOD'15, ICDE'16, Middleware'18]

Journey…



HTAP Column-store: L-Store (Lineage-based Data Store)

[VLDB'12, ICDE'14, ICDCS'16, EDBT'18, 34 filed US patents]

Graphs on SQL: (e.g., GRFusion) [SIGMOD'18, EDBT'18]

Journey…

Agreement Protocols: (e.g., EasyCommit, Q-Store)
[EDBT'18, DAPD'19]

Journey…

Resilient Consensus Protocols

Agreement Protocols

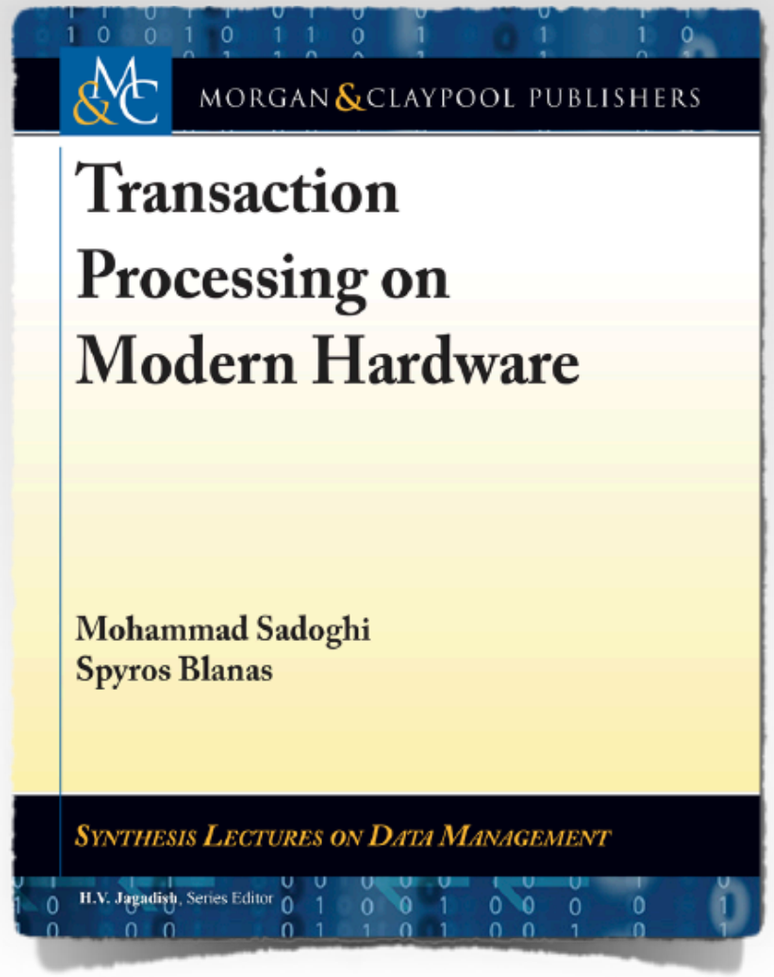
Concurrency Protocols

SQL Transactions

HTAP: Unifying Storage Layer (Lineage-based Storage Architecture)

SQL Analytics

Consensus Protocols: (e.g., ResilientDB, Blockplane, Blocklite)
[SC'19, ICDE'19, DISC'19 (2x), arXiv'19 (6x)]



Books

Transaction Processing on Modern Hardware.

Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2019

Fault-Tolerant Distributed Transactions on Blockchain.

Synthesis Lectures on Data Management, Morgan & Claypool Publishers, *to appear 2020*

ExpoLab
Creativity Unfolded

UCDAVIS

Press

Advancements TV With Ted Danson - CNBC, Yahoo! Finance, Market Insider, CoinDesk, Crypto Media, Davis Enterprise, Times Union, WBOC TV/Radio

Books

Transaction Processing on Modern Hardware.

Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2019

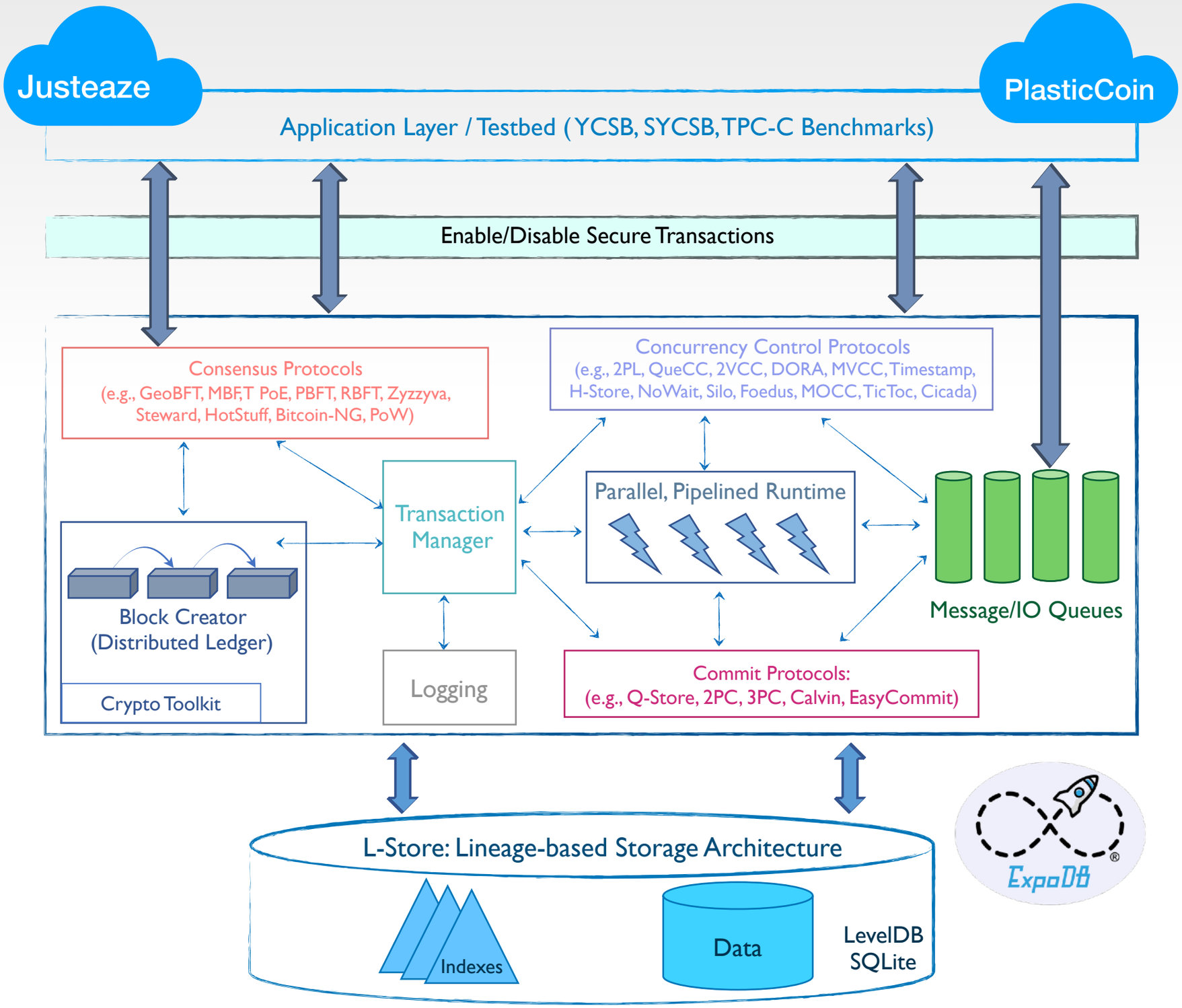
Fault-Tolerant Distributed Transactions on Blockchain.

Synthesis Lectures on Data Management, Morgan & Claypool Publishers, *to appear 2020*

ExpoLab
Creativity Unfolded

UCDAVIS

ExpoDB Architecture



ResilientDB

Coming Soon…

Quantifiable Resiliency (Graduate Student Experiments)

Aloha Lake, Desolation Wilderness

15 Miles Long

2,500 Feet Elevation Gain

(8,700 Feet at Summit)

Tomales Point Trail, Point Reyes National Seashore

9.4 Miles Long

1,579 Feet Elevation Gain

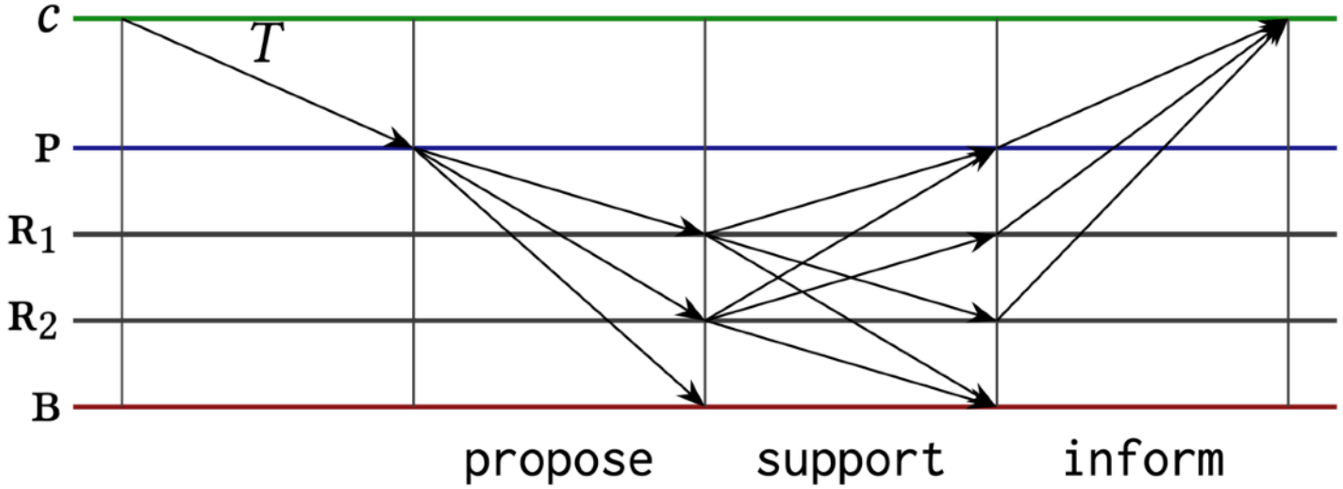
Non-Quantifiable Resiliency

Proof-of-Execution: Reaching Consensus Through Fault-Tolerant Speculation [arXiv'19]

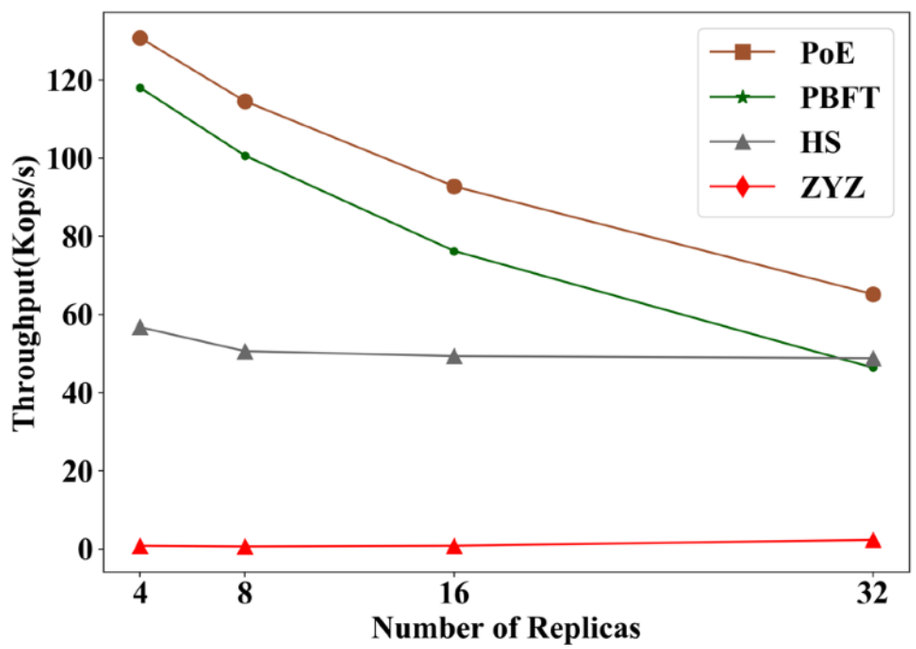
Out-of-Order message processing to reduce replica idleness

Speculative Execution with revertible/divergent replicas &

eager/irrevertible client commit



Fault-tolerant Proof-of-Execution Protocol



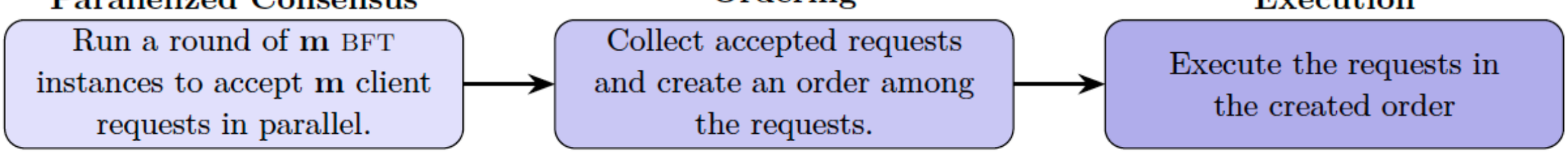
PoE scales beyond 32 replicas, in presence of failures, outperforms PBFT up to 40%

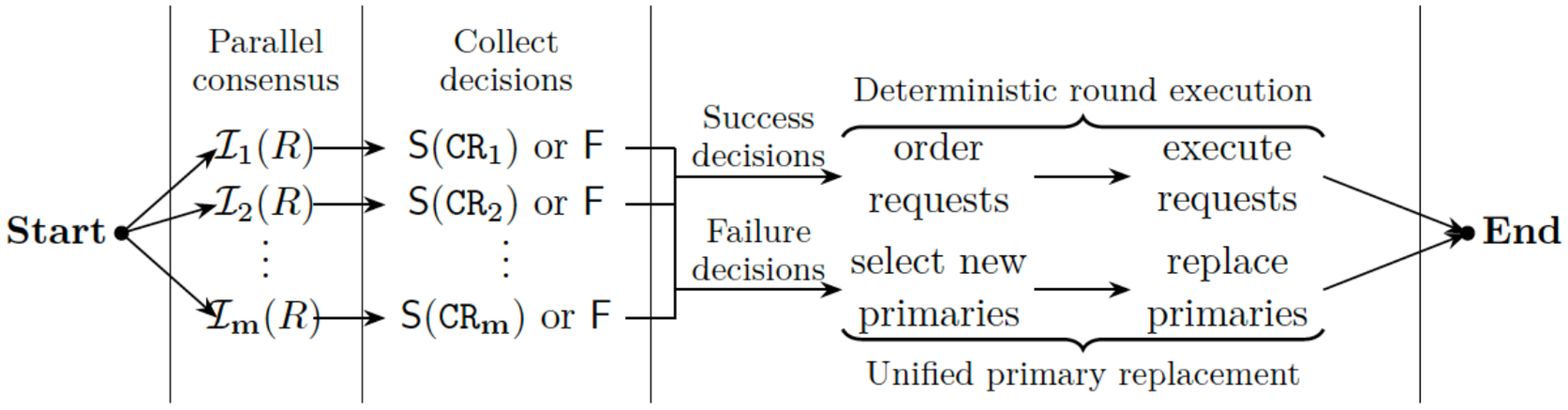
MultiBFT: Scaling Blockchain Databases Through Parallel Resilient Consensus Paradigm [arXiv'19]

A wait-free meta-protocol…

Designate multiple replicas as Primaries!

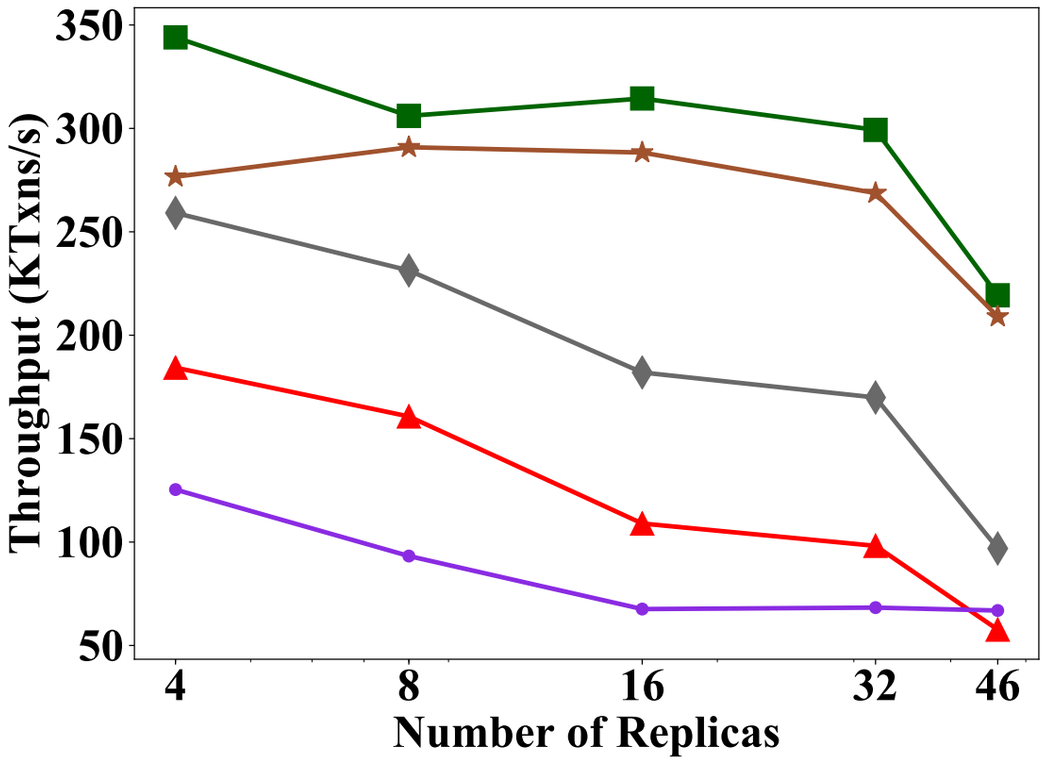
Run multiple parallel consensuses on each replica independently



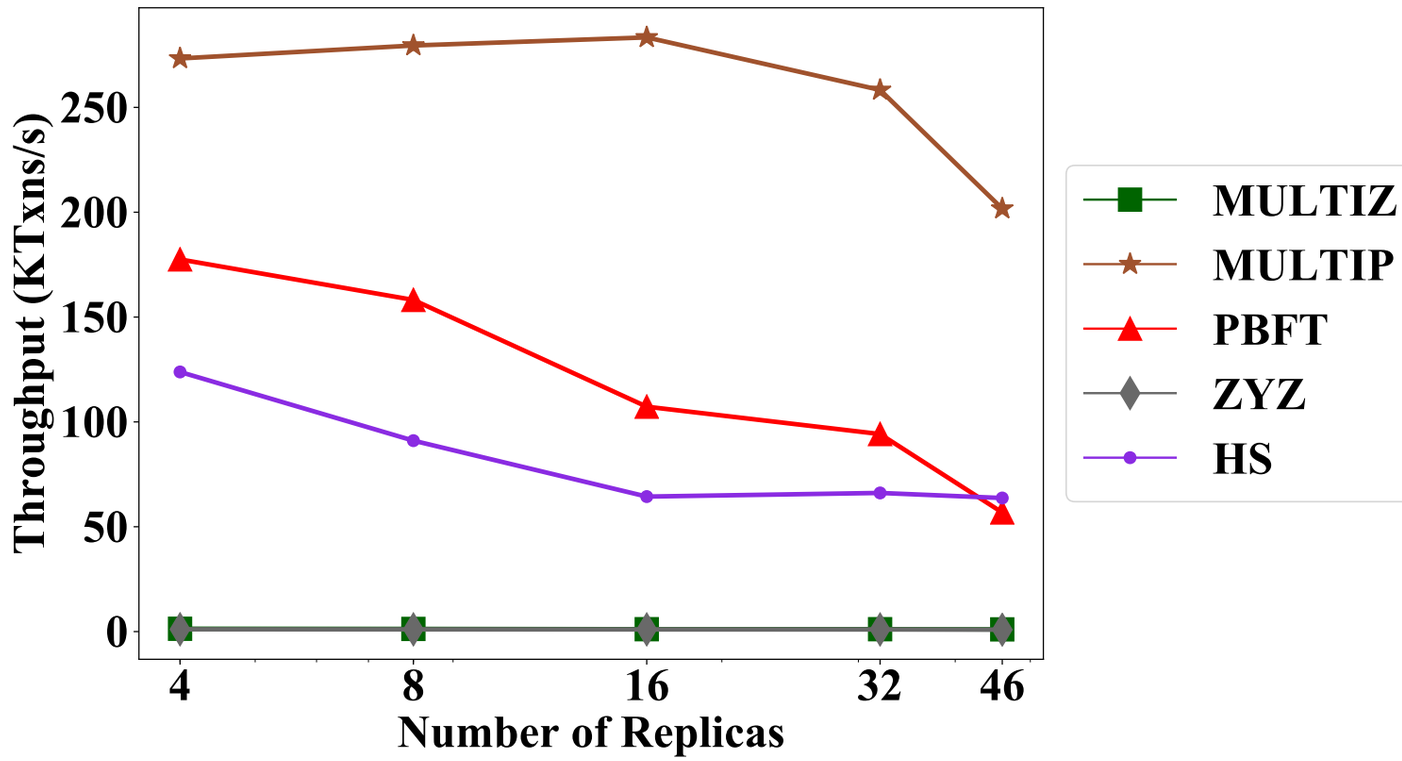


Fault-tolerant MultiBFT Protocol

MultiBFT: Scaling Blockchain Databases Through Parallel Resilient Consensus Paradigm



Throughput up to 350,000 txns/s (without failures)



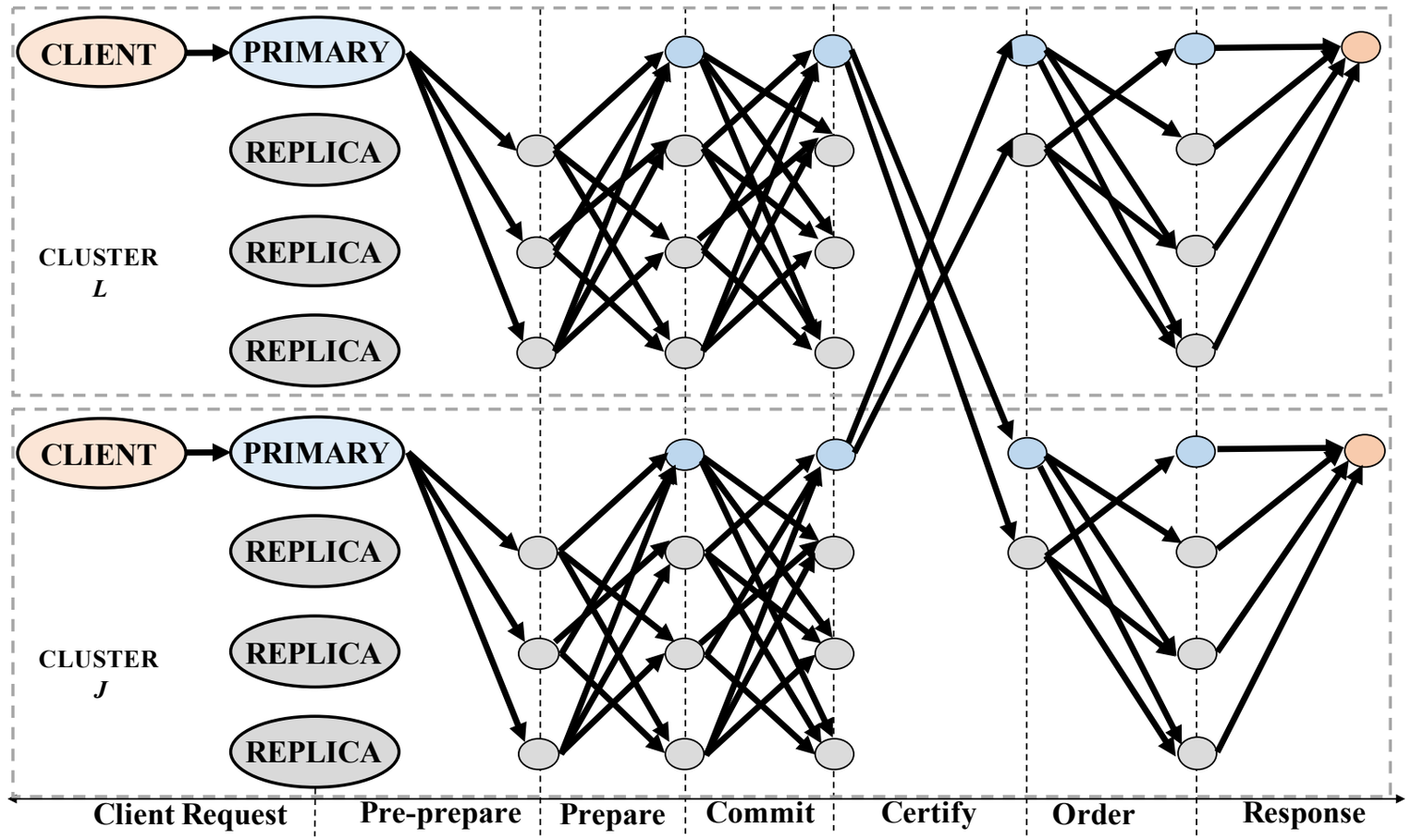
Throughput up to 300,000 txns/s (with failures)

GeoBFT: Global Scale Resilient Blockchain Fabric [arXiv'19]

A meta-protocol, locally running any BFT in parallel and independently

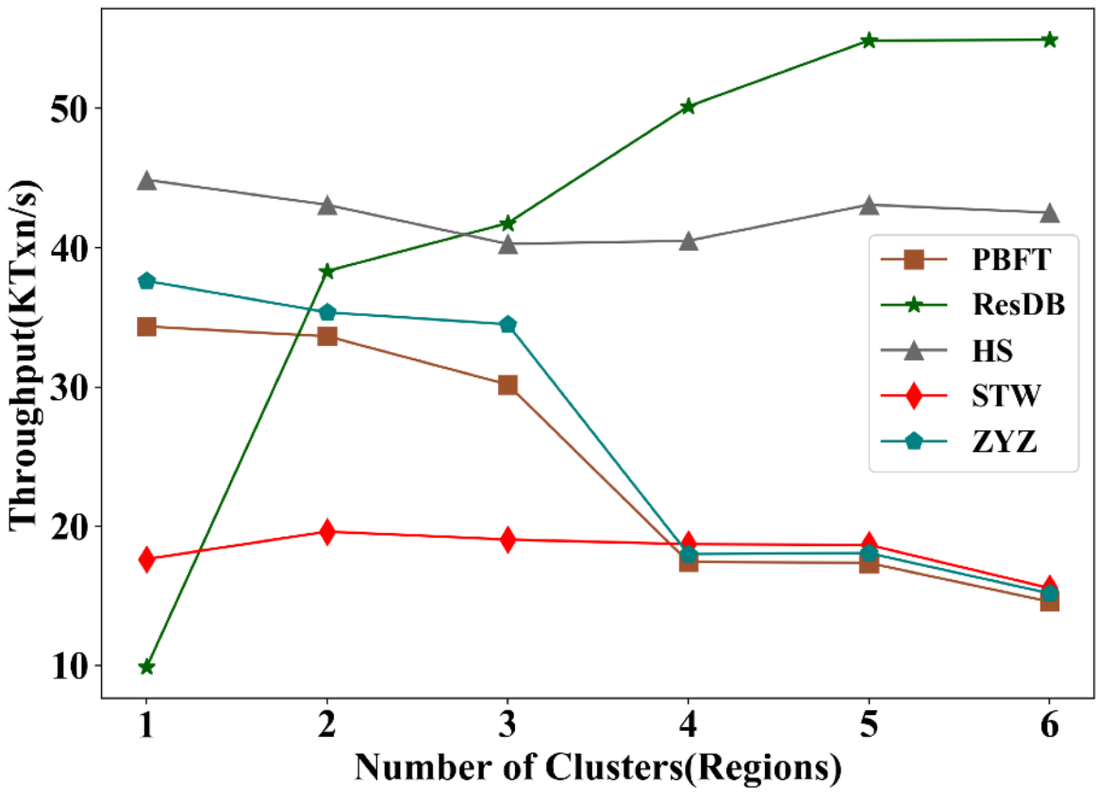
Global ordering provably requires only linear communication

Provably sufficient for primary to send a certificate to at most f+1 replicas, malicious primary is detectable and replaceable

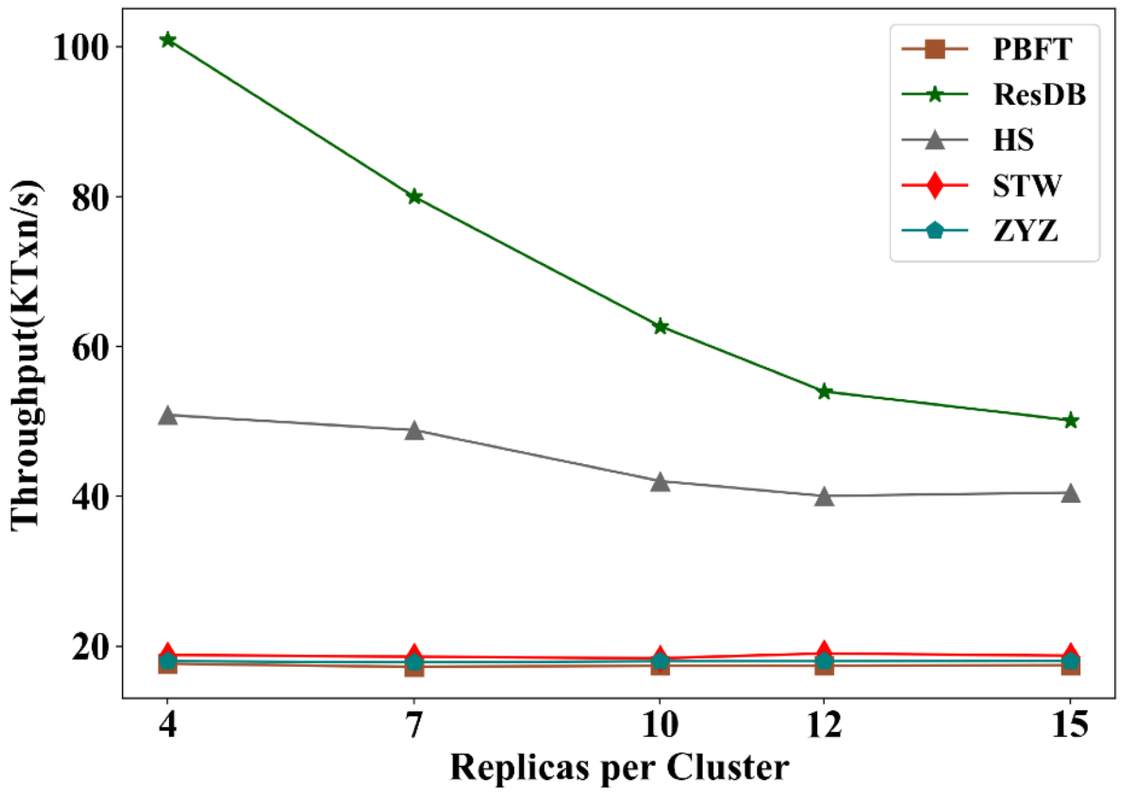


Fault-tolerant GeoBFT Protocol

GeoBFT: Global Scale Resilient Blockchain Fabric [arXiv'19]



ResilientDB easily scales across 6 countries in 4 continents due to GeoBFT protocol.



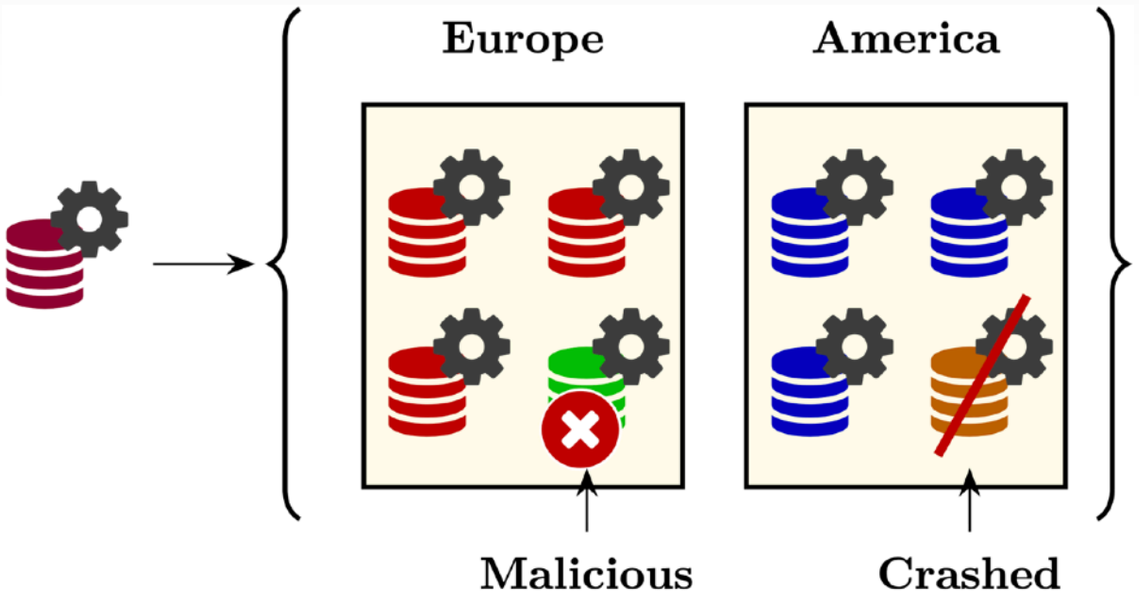
GeoBFT scales a permissioned blockchain up to 60 replicas globally.

The Fault-Tolerant Cluster-Sending Problem

the problem of sending a message from one Byzantine cluster to another Byzantine cluster in a reliable manner,

establishing lower bounds on the complexity of this problem under crash failures and Byzantine failures

(linear in the size of clusters)



	Protocol	System	Robustness	Messages	Message size
non-linear	RB-bcs	Omit	$\mathbf{n}_{\mathcal{C}_1} > 2\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > \mathbf{f}_{\mathcal{C}_2}$	$(\mathbf{f}_{\mathcal{C}_1}+1)\cdot(\mathbf{f}_{\mathcal{C}_2}+1)$	$\mathcal{O}(\|\|v\|\|)$
	RB-brs	Byzantine, RS	$\mathbf{n}_{\mathcal{C}_1} > 2\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > \mathbf{f}_{\mathcal{C}_2}$	$(2\mathbf{f}_{\mathcal{C}_1}+1)\cdot(\mathbf{f}_{\mathcal{C}_2}+1)$	$\mathcal{O}(\|\|v\|\|)$
	RB-bcs	Byzantine, RS	$\mathbf{n}_{\mathcal{C}_1} > 2\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > \mathbf{f}_{\mathcal{C}_2}$	$(\mathbf{f}_{\mathcal{C}_1}+1)\cdot(\mathbf{f}_{\mathcal{C}_2}+1)$	$\mathcal{O}(\|\|v\|\|+\mathbf{f}_{\mathcal{C}_1})$
	RB-bcs	Byzantine, CS	$\mathbf{n}_{\mathcal{C}_1} > 2\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > \mathbf{f}_{\mathcal{C}_2}$	$(\mathbf{f}_{\mathcal{C}_1}+1)\cdot(\mathbf{f}_{\mathcal{C}_2}+1)$	$\mathcal{O}(\|\|v\|\|)$
linear	PBS-bcs	Omit	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$\mathcal{O}(\max(\mathbf{n}_{\mathcal{C}_1},\mathbf{n}_{\mathcal{C}_2}))$ (optimal)	$\mathcal{O}(\|\|v\|\|)$
	PBS-brs	Byzantine, RS	$\mathbf{n}_{\mathcal{C}_1} > 4\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 4\mathbf{f}_{\mathcal{C}_2}$	$\mathcal{O}(\max(\mathbf{n}_{\mathcal{C}_1},\mathbf{n}_{\mathcal{C}_2}))$ (optimal)	$\mathcal{O}(\|\|v\|\|)$
	PBS-bcs	Byzantine, RS	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$\mathcal{O}(\max(\mathbf{n}_{\mathcal{C}_1},\mathbf{n}_{\mathcal{C}_2}))$	$\mathcal{O}(\|\|v\|\|+\mathbf{f}_{\mathcal{C}_1})$
	PBS-bcs	Byzantine, CS	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$\mathcal{O}(\max(\mathbf{n}_{\mathcal{C}_1},\mathbf{n}_{\mathcal{C}_2}))$ (optimal)	$\mathcal{O}(\|\|v\|\|)$

Permissioned Blockchain Through the Looking Glass: Architectural and Implementation Lessons Learned [arXiv'19]

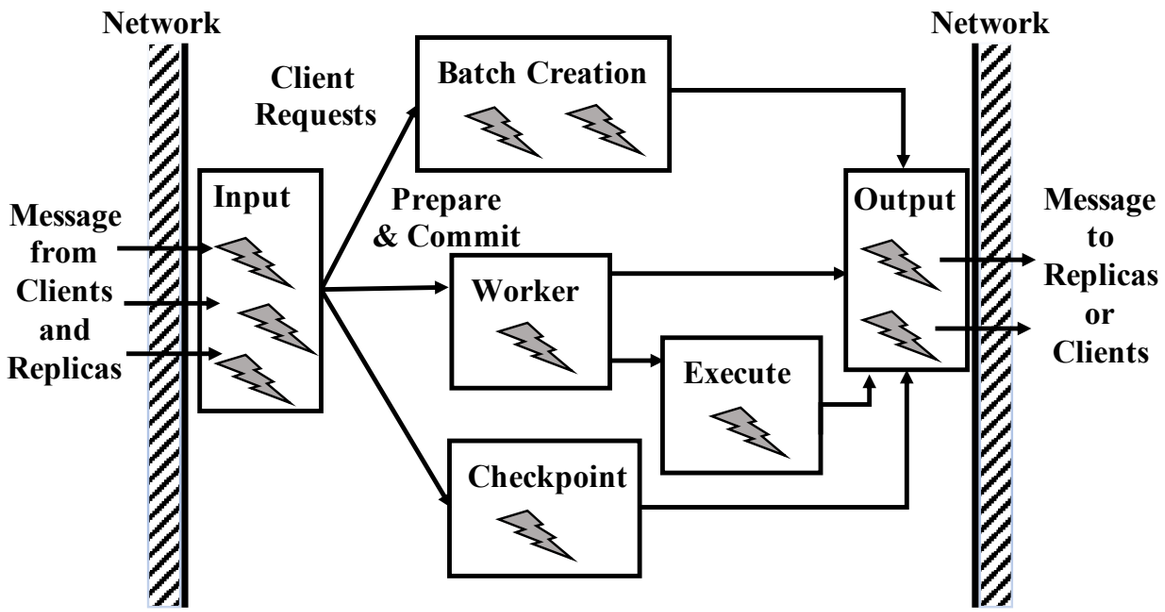
Single-threaded Monolithic Design

Out-of-ordering Consensus Communication

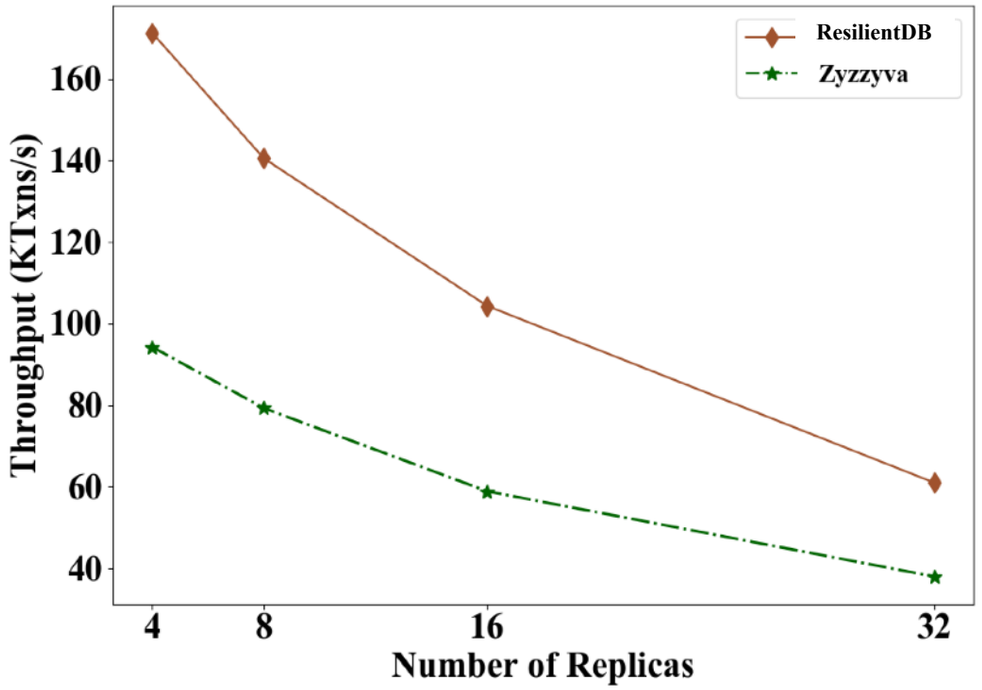
De-coupled Ordering and Execution

Off-Chain Memory Management

Expensive Cryptographic Practices (DS vs. MAC)



Multi-Threaded Deep Pipeline



Can a well-crafted system based on a classical BFT protocol outperform a modern protocol?

Mount Tallac, Lake Tahoe
12.1 Miles Long
3,931 Feet Elevation Gain
(9,738 Feet at Summit)

Thank You

ACM MIDDLEWARE 2019

2019.middleware-conference.org

Coming to UC Davis in December 2019

For Complete References