ResilientDB: Building Global-Scale Privacy-Preserving Blockchain Fabric

Mohammad Sadoghi

REIMAGINE 2020

Global Blockchain Conference

May 18, 2020

Mohammad Sadoghi

Exploratory Systems Lab

Department of Computer Science

UC DAVIS
UNIVERSITY OF CALIFORNIA

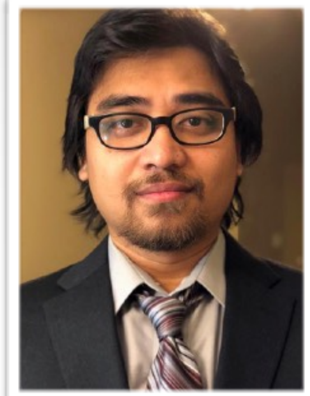
ExpoLab
Creativity Unfolded

ExpoLab Team

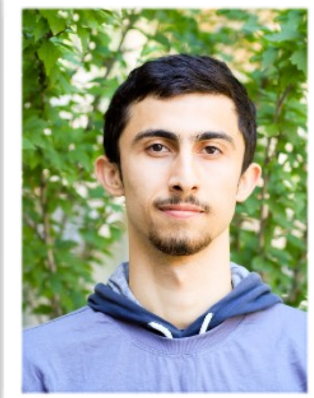
Jelle Hellings, PostDoc
(Fault-tolerant Complexity Analysis)

Suyash Gupta, PhD
(Scalable Consensus Meta-Protocols)

Mohammad Sadoghi
(Principal Investigator)



Thamir Qadah, PhD
(Distributed & Coordination-free Concurrency)



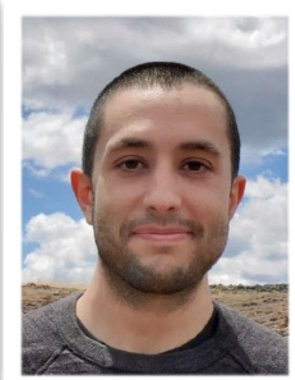
Sajjad Rahnama, PhD
(Global Scale Consensus)

Dhruv Krishnan, MSc
(Scaling Fabric via Sharding)

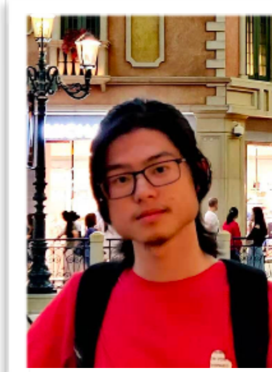
Priya Holani, MSc
(Scaling Fabric via Sharding)

Shubham Pandey, MSc
(Scaling Fabric via RDMA)

Rohan Sogani, MSc
(Scaling Fabric via Sharding)



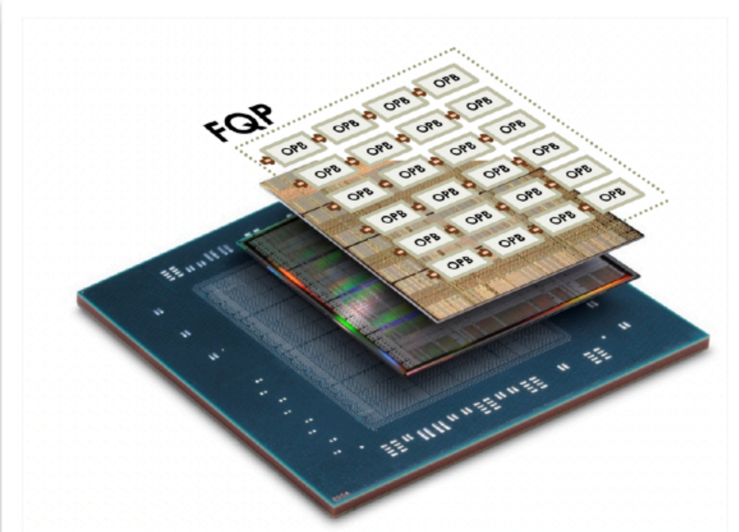
Erik Linsenmayer, BSc
(Scaling Fabric via RDMA)



Xinyuan Sun, BSc
(Scaling Fabric via RDMA)

Resilient Journey…

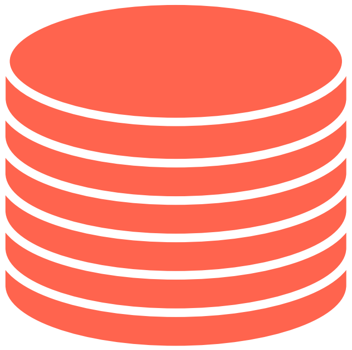
SQL Analytics



Flexible Query Processing (FQP) on a Reconfigurable Computing Fabric

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]

Resilient Journey…

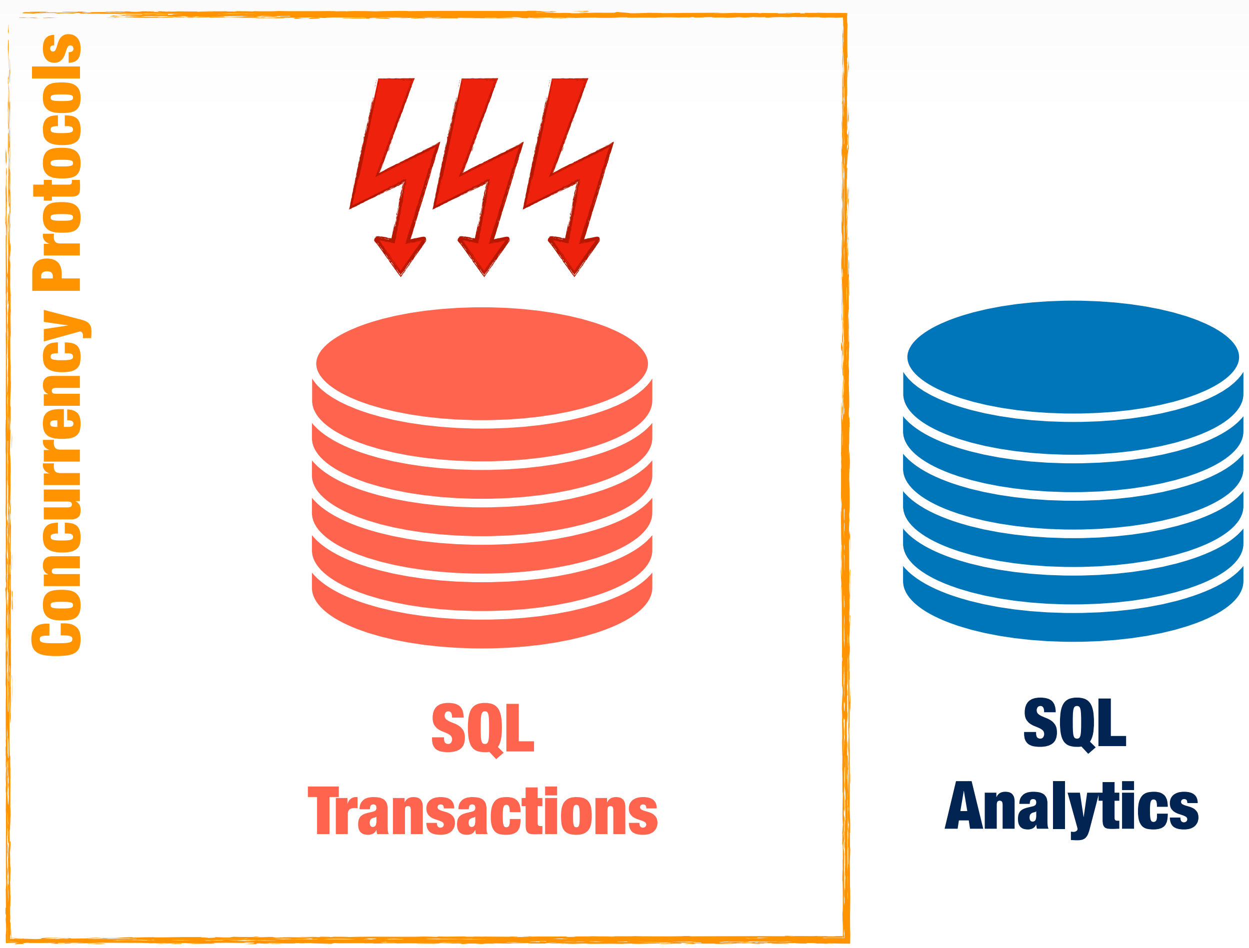


SQL
Transactions

SQL
Analytics

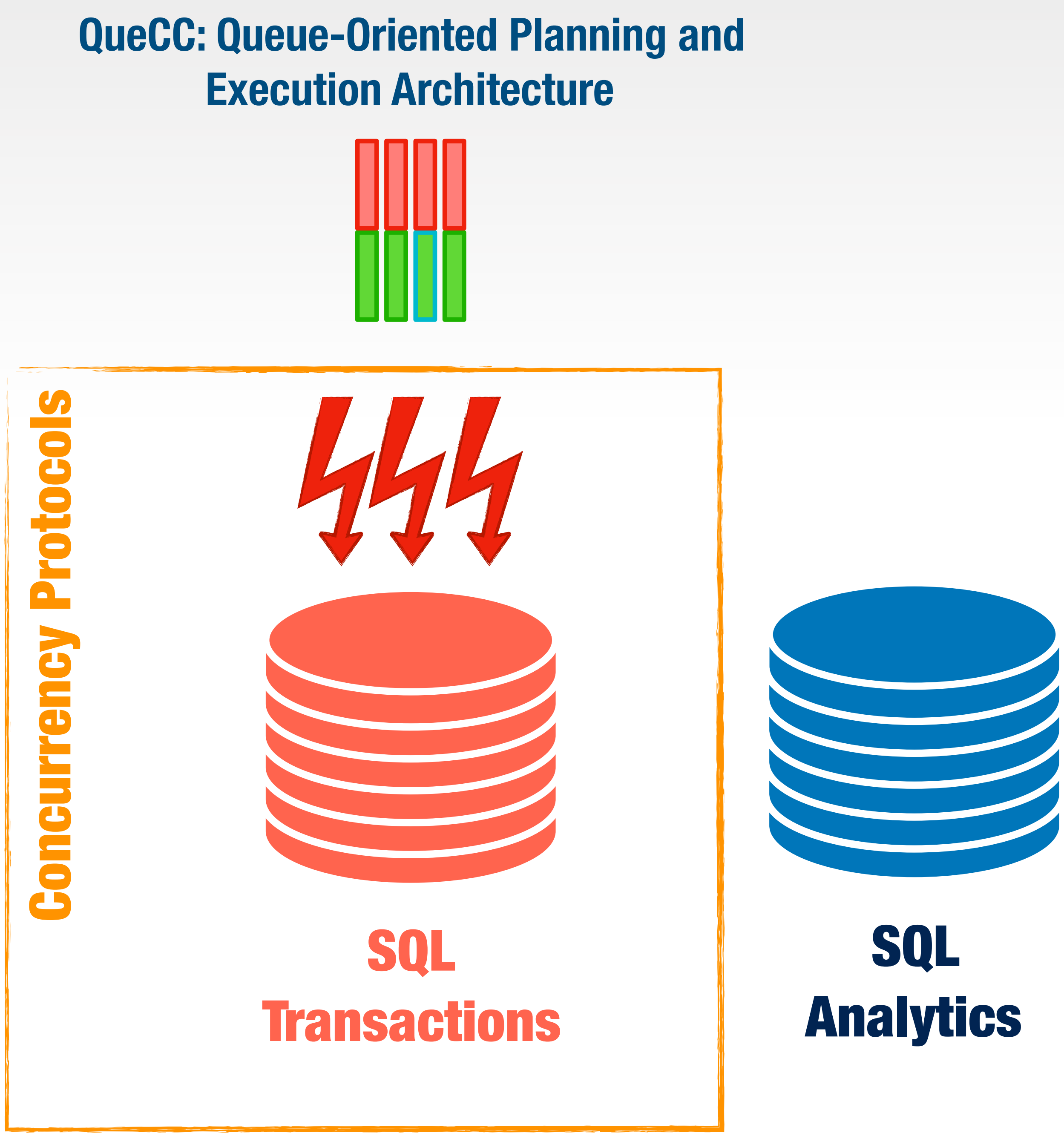
High-dimensional Indexing: (e.g., BE-Tree, BE-topK)
[SIGMOD’11, ICDE’12, TODS’13, ICDCS’13, ICDE’14, ICDCS’17, Middleware'17]

Resilient Journey…



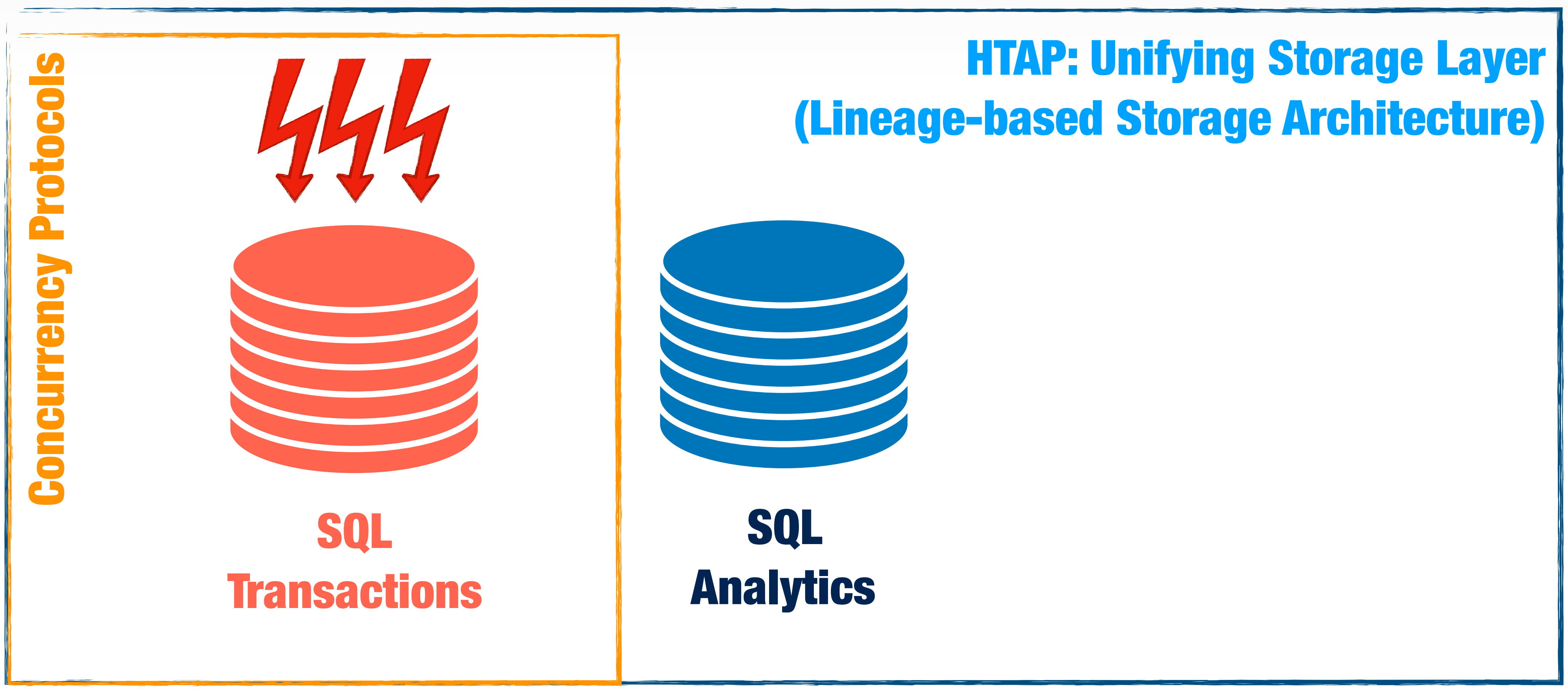
Concurrency Control Protocols: (e.g., 2VCC, QueCC - Best Paper Award)
[VLDB’13, VLDB’14, VLDBJ’16, Middleware’16, TDKE’15, SIGMOD’15, ICDE’16, Middleware'18]

Resilient Journey…



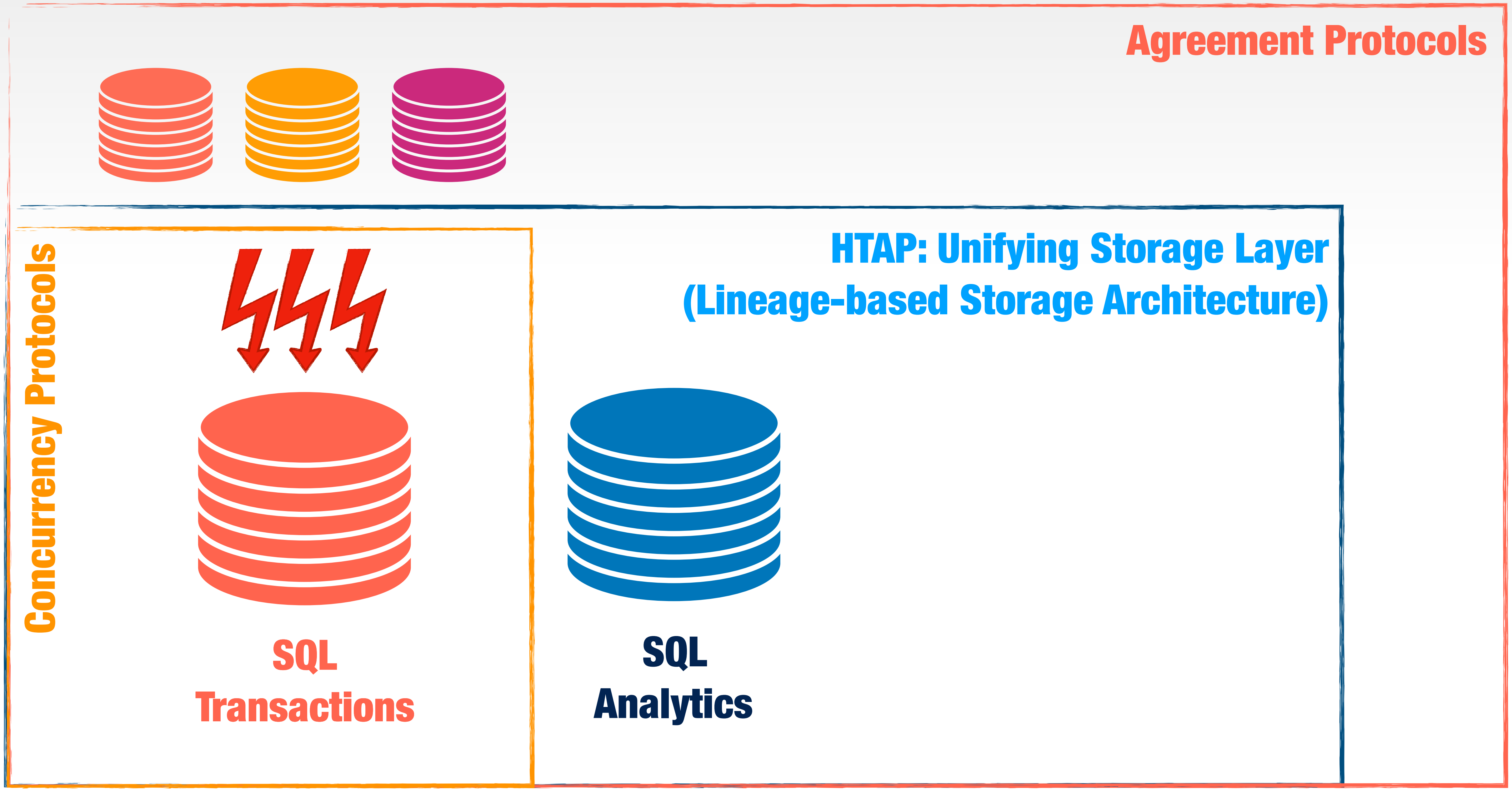
Concurrency Control Protocols: (e.g., 2VCC, QueCC - Best Paper Award)
[VLDB’13, VLDB’14, VLDBJ’16, Middleware’16, TDKE’15, SIGMOD’15, ICDE’16, Middleware'18]

Resilient Journey…



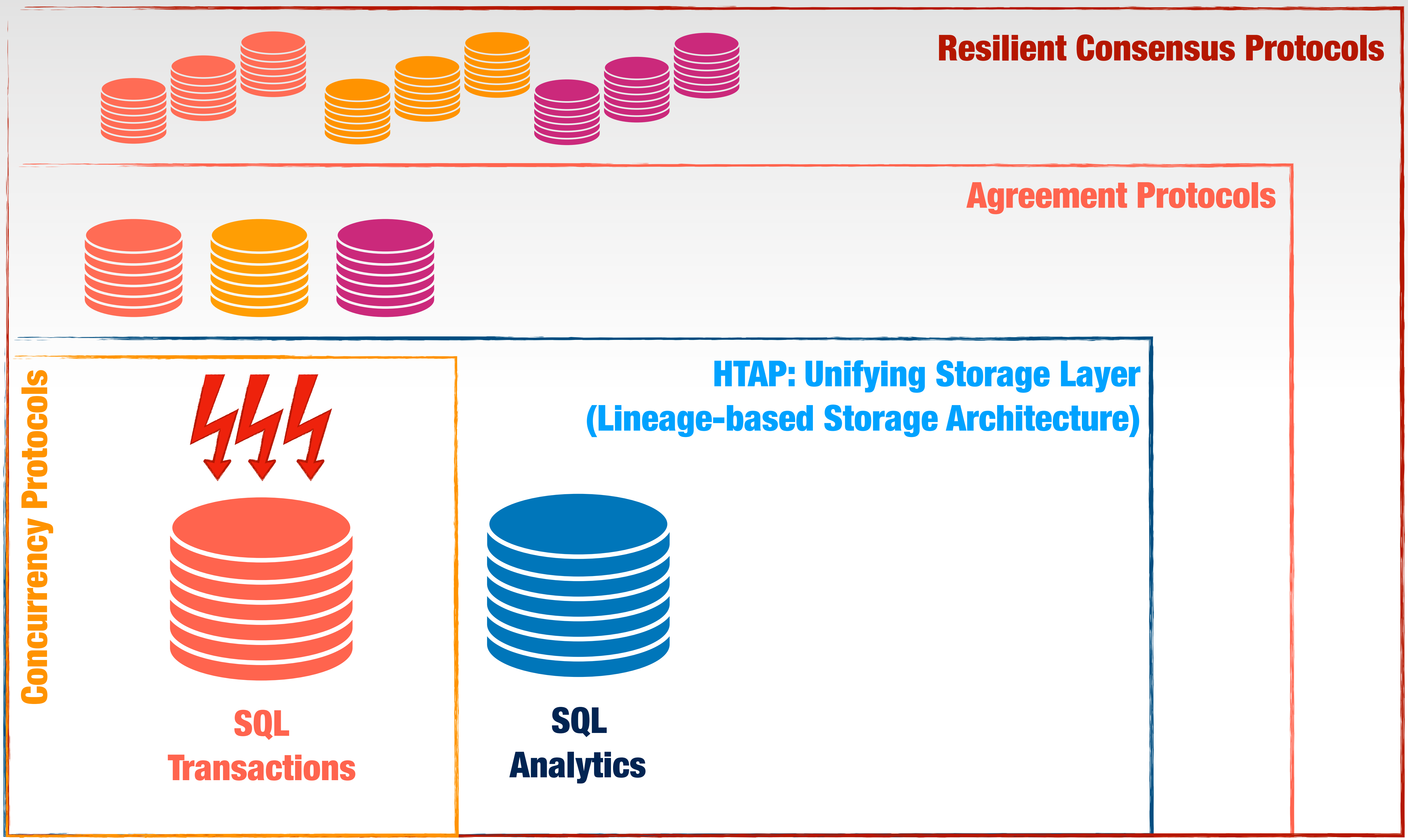
HTAP Column-store: L-Store (Lineage-based Data Store)
[VLDB'12, ICDE'14, ICDCS'16, EDBT'18, 34 filed US patents]
Graphs on SQL: (e.g., GRFusion) [SIGMOD'18, EDBT'18]

Resilient Journey…



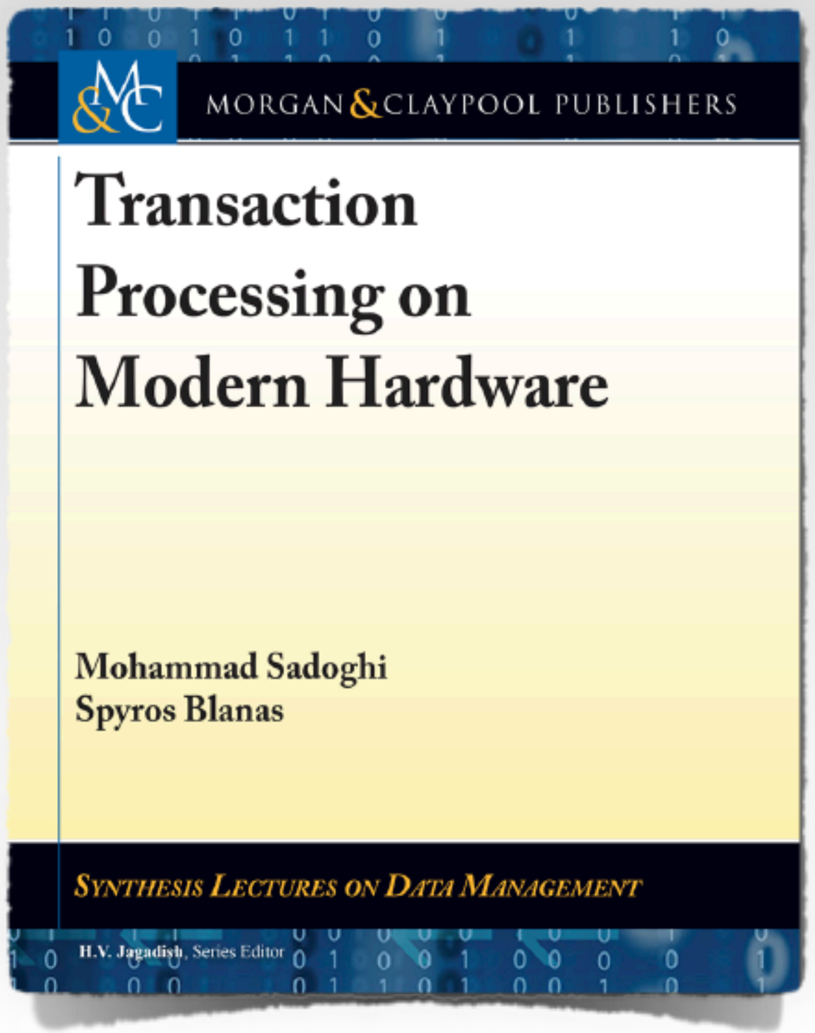
Agreement Protocols: (e.g., EasyCommit, Q-Store)
[EDBT'20, EDBT'18, DAPD'19]

Resilient Journey…



Consensus Protocols: (e.g., ResilientDB, GeoBFT, PoE, MBFT, Delayed Replication, CSP, Blockplane)

[VLDB’20, ICDCS’20, ICDT’20, DISC'19 (2x), SC’19, ICDE'19, 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, CityAM, Medium, Yahoo! Finance, Market Insider, CoinDesk, Crypto Media, Davis Enterprise, Times Union, WBOC TV/Radio

Books

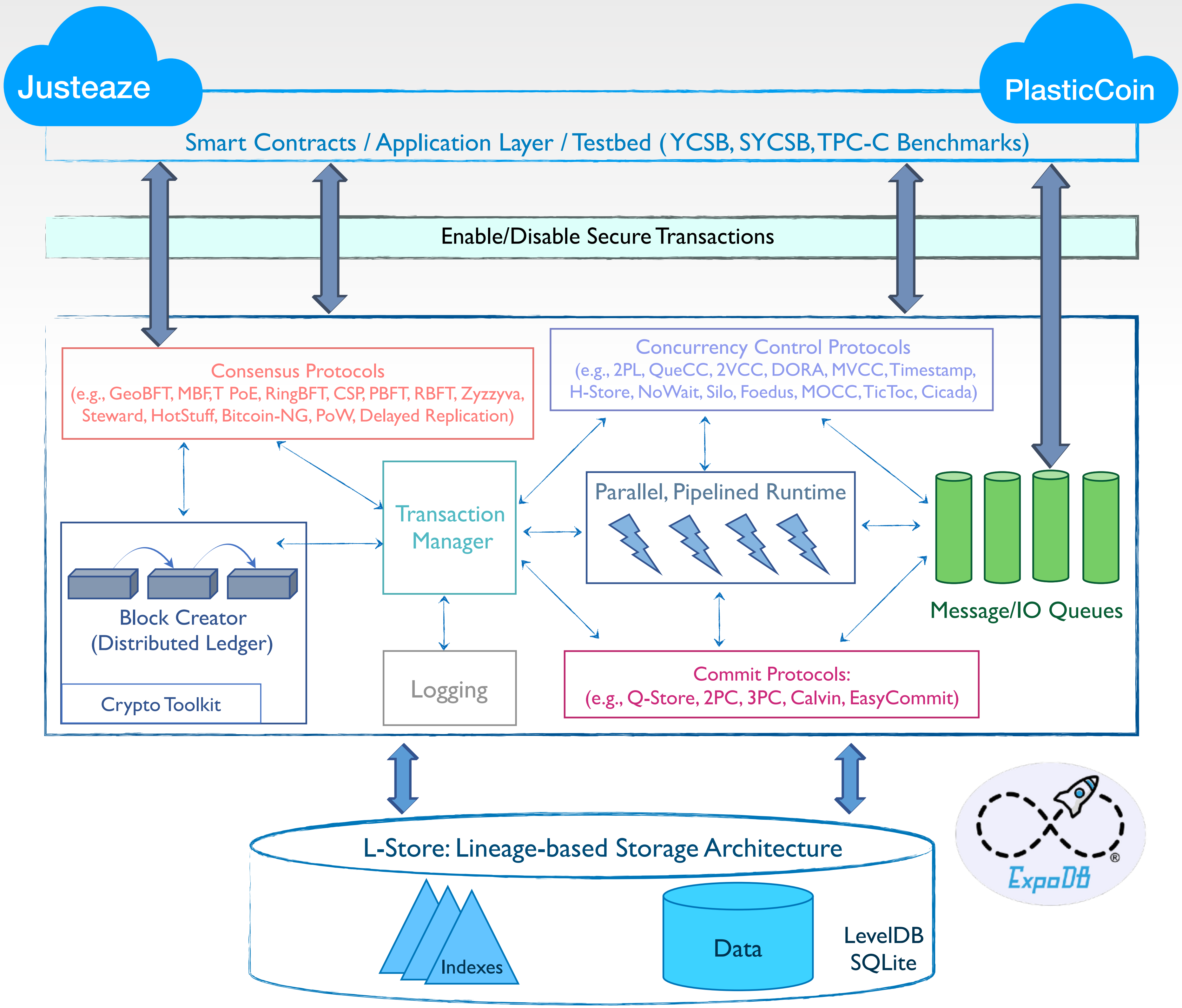
Transaction Processing on Modern Hardware.
Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2019

Fault-Tolerant Distributed Transactions on Blockchain.
Synthesis Lectures on Data Management, Morgan & Claypool Publishers, *to appear 2020*

ExpoLab
Creativity Unfolded

UCDAVIS

ExpoDB Architecture



ResilientDB

Security, Privacy Reloaded

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

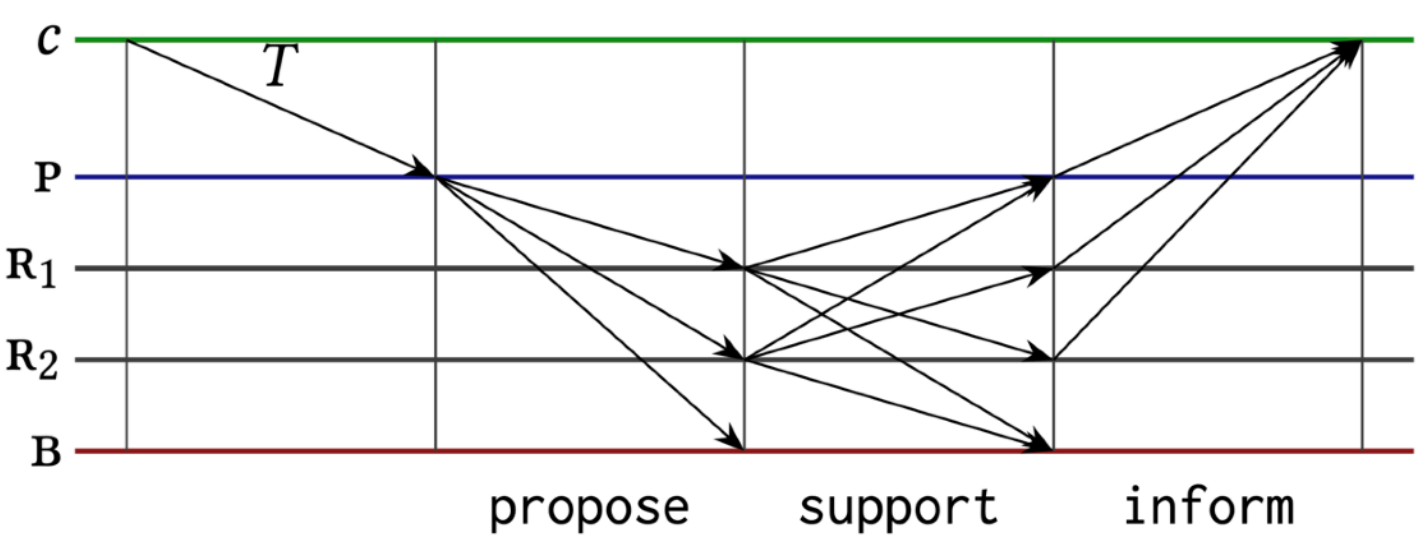
1,579 Feet Elevation Gain

Non-Quantifiable Resiliency

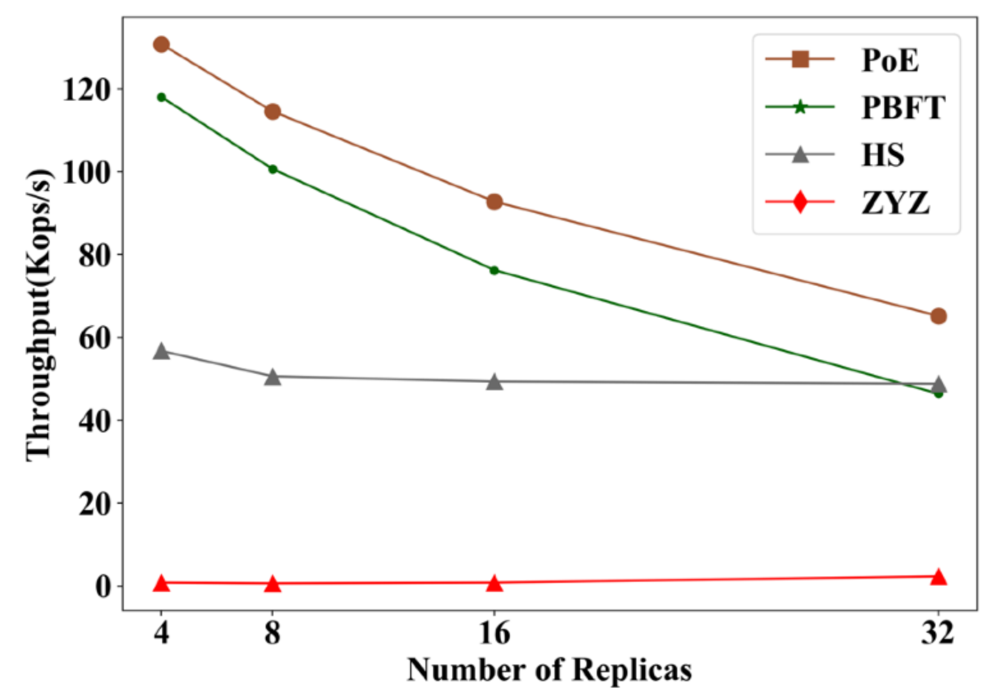
Proof-of-Execution: Reaching Consensus Through Fault-Tolerant Speculation [arXiv'19]

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 <u>32 replicas</u>, in presence of failures, outperforms <u>PBFT up to 40%</u>

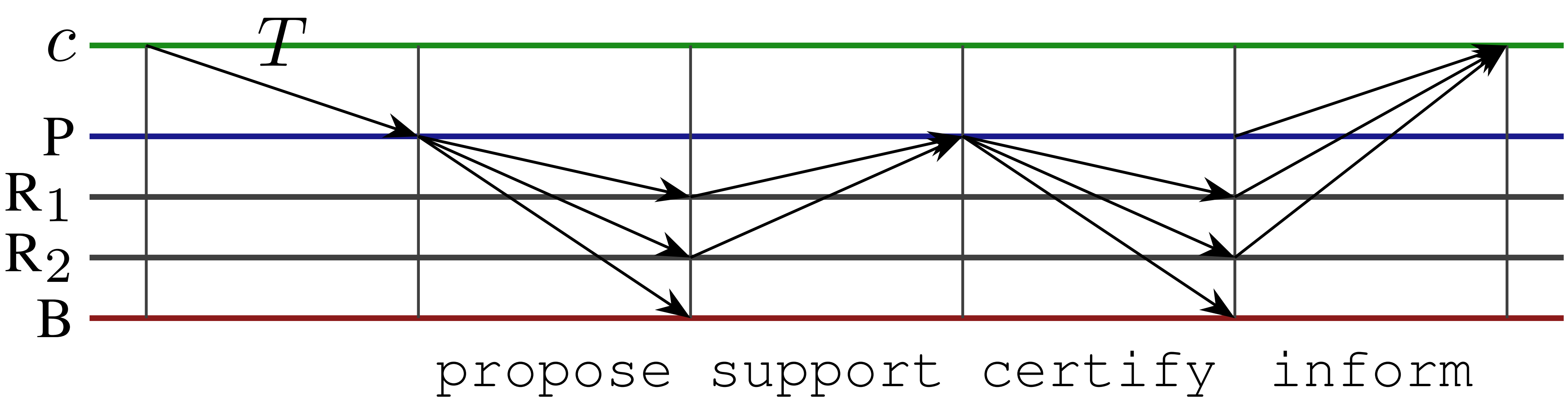
Proof-of-Execution: Reaching Consensus Through Fault-Tolerant Speculation [arXiv'19]

Out-of-Order message processing to reduce replica idleness

Speculative Execution with revertible/divergent replicas &

eager/irrevertible client commit

introducing linear message complexity



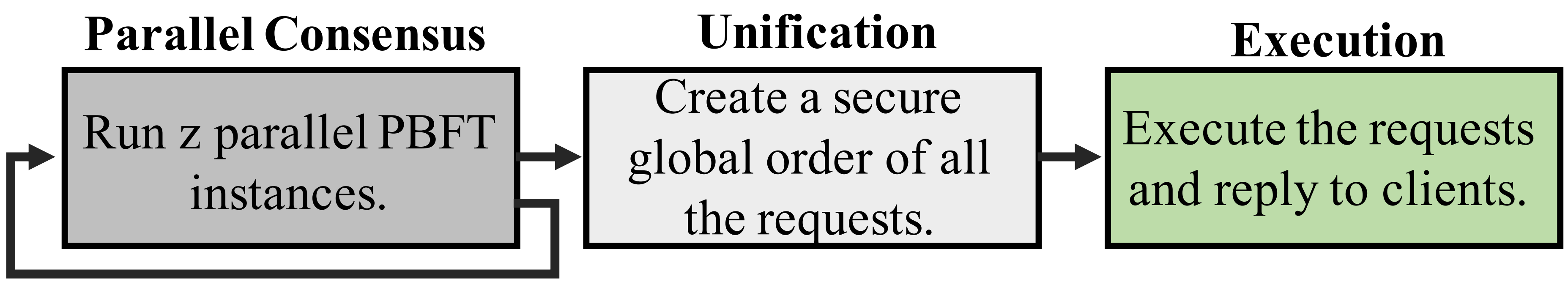
Linearized Proof-of-Execution Protocol

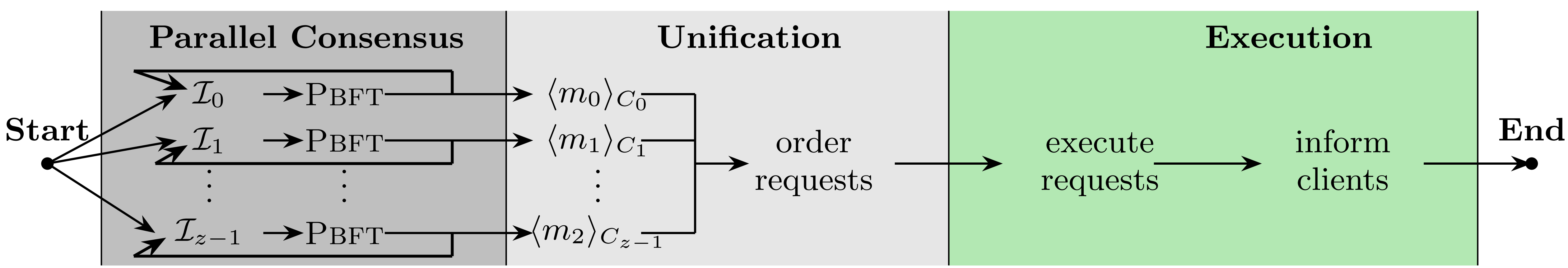
MultiBFT: Scaling Blockchain Databases Through Parallel Resilient Consensus Paradigm [arXiv'19]

A wait-free meta-protocol…

Designate multiple replicas as Primaries!

Run multiple parallel consensuses on each replica independently

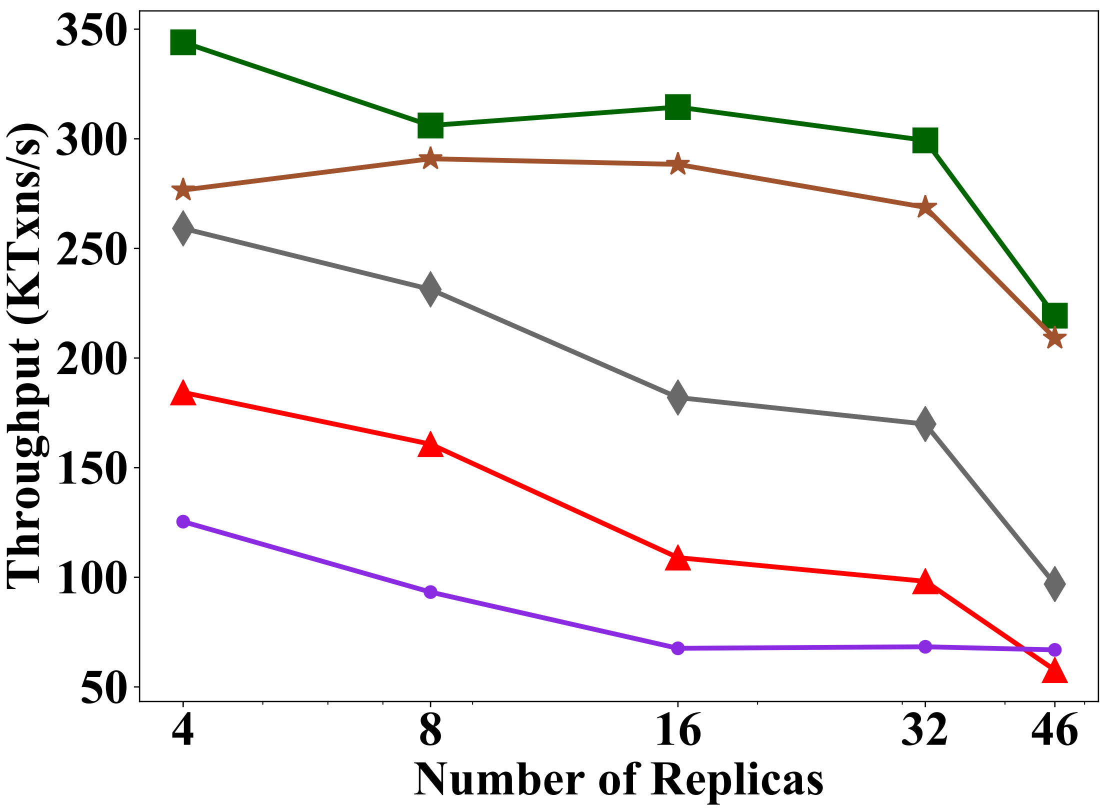


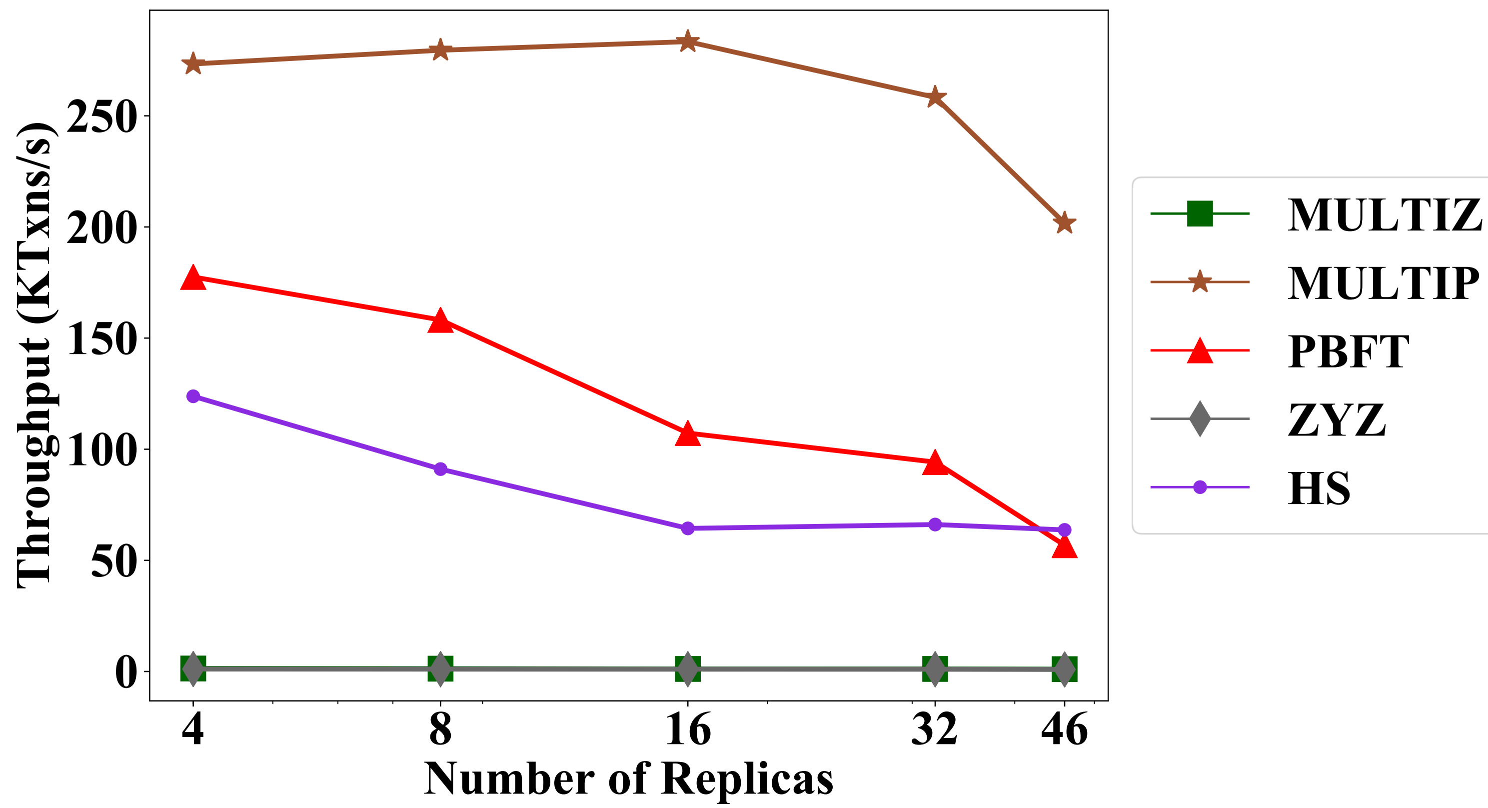


Fault-tolerant MultiBFT Protocol

Brief Announcement: Revisiting Consensus Protocols through Wait-free Parallelization. DISC 2019

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)

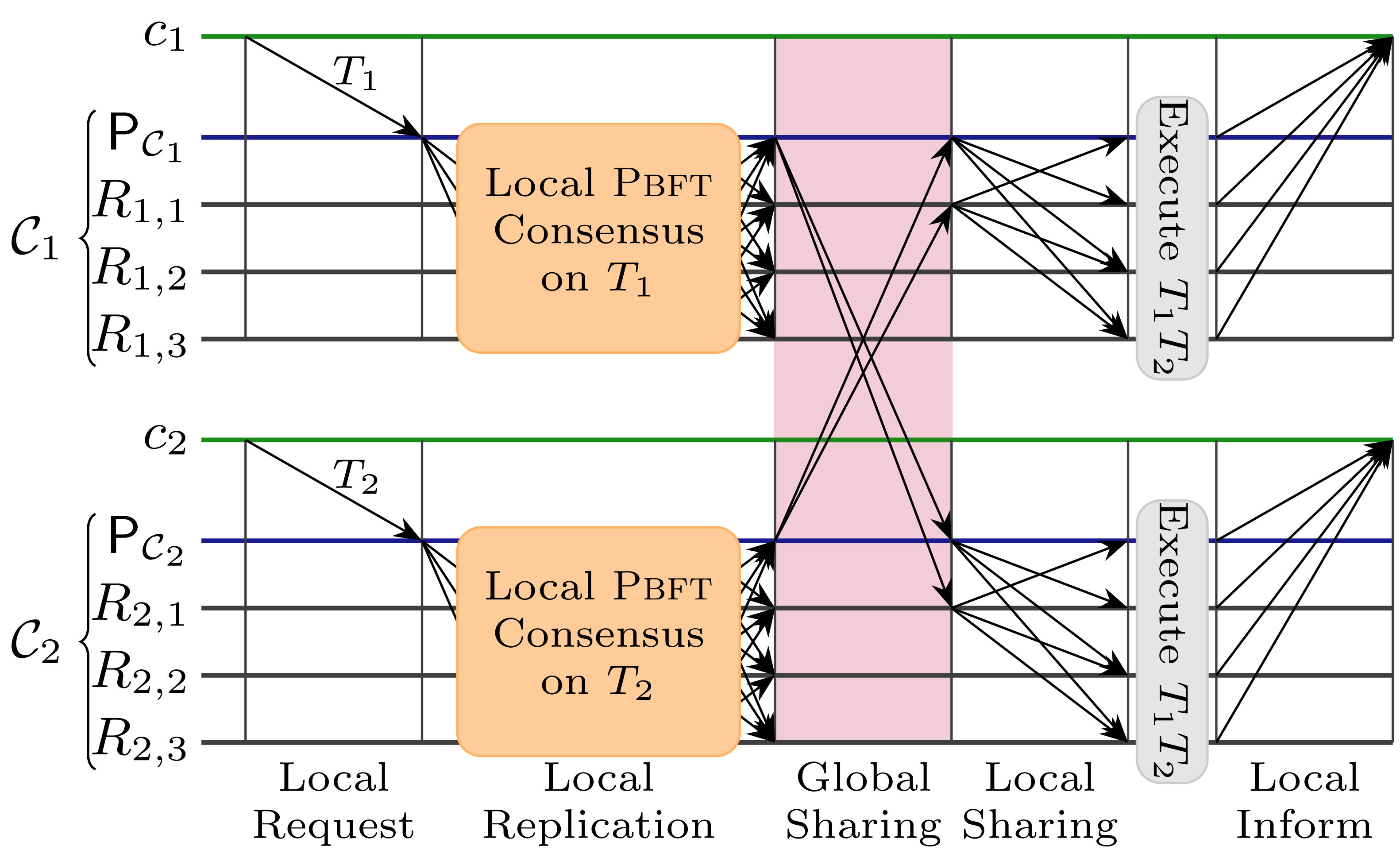
Brief Announcement: Revisiting Consensus Protocols through Wait-free Parallelization. DISC 2019

GeoBFT: Global Scale Resilient Blockchain Fabric [VLDB'20]

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

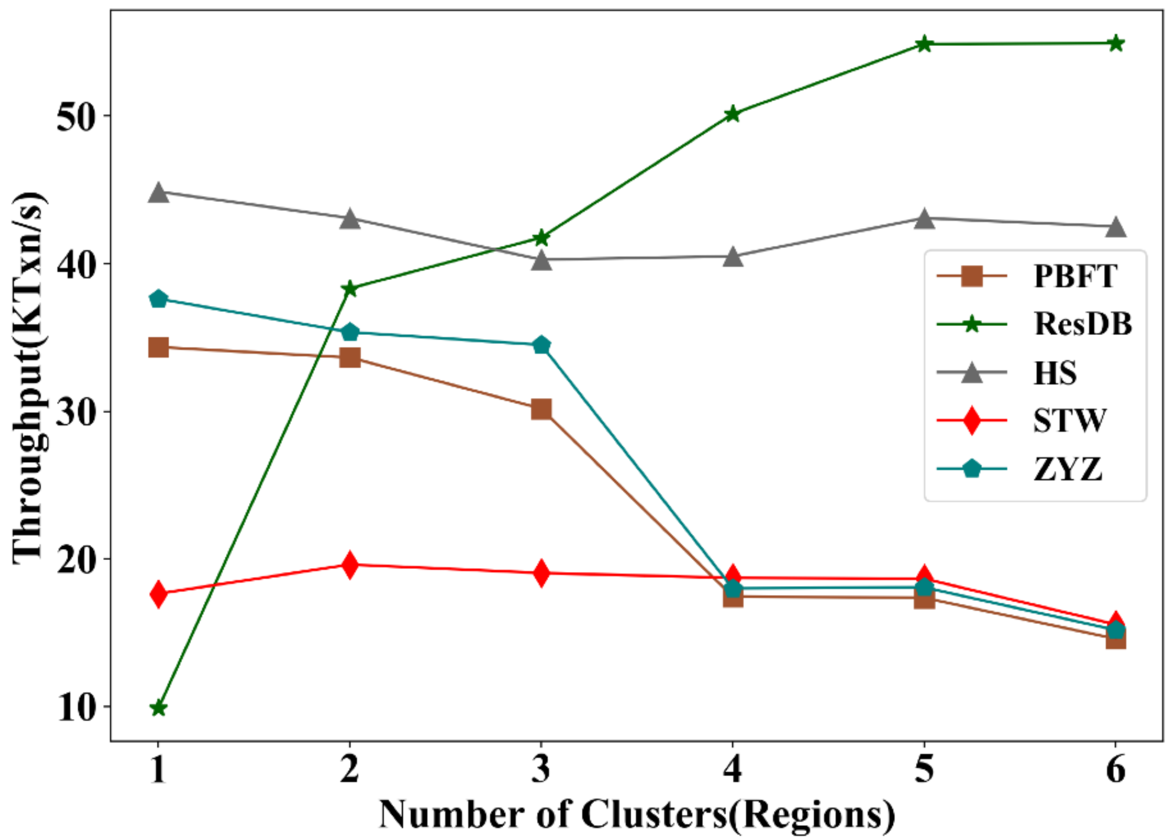
Global ordering provably requires only linear communication

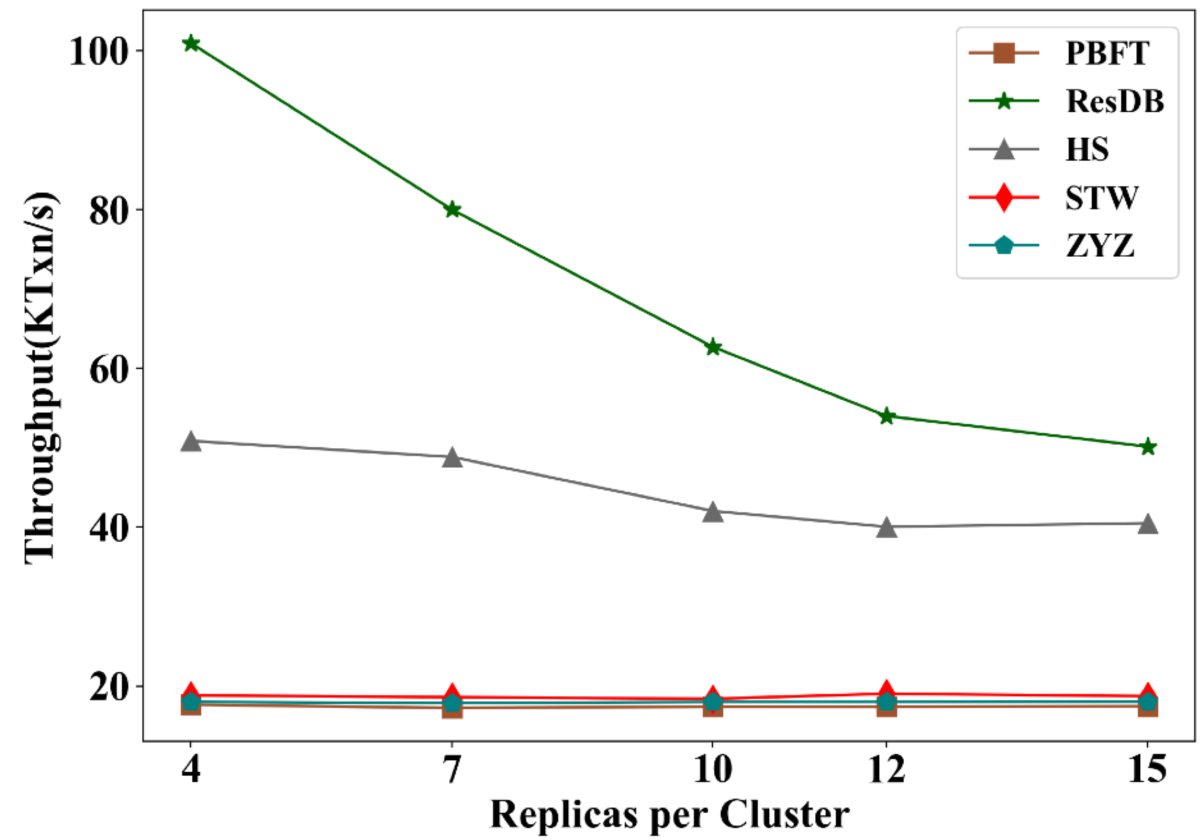
Provably sufficient for primary to send a certificate to at most f+1 replicas, malicious primary is detectable and replaceable



Fault-tolerant GeoBFT Protocol

GeoBFT: Global Scale Resilient Blockchain Fabric [VLDB'20]





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

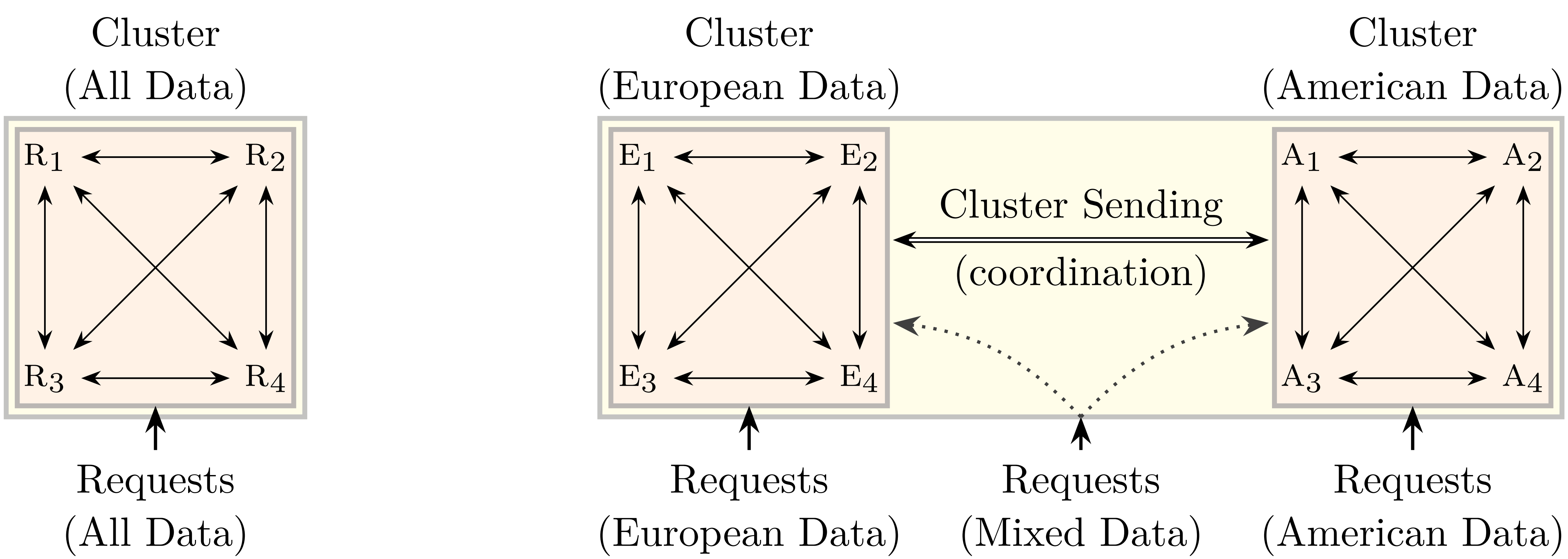
GeoBFT scales a permissioned blockchain up to 60 replicas globally.

The Fault-Tolerant Cluster-Sending Problem [DISC'19]

formalizing 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)



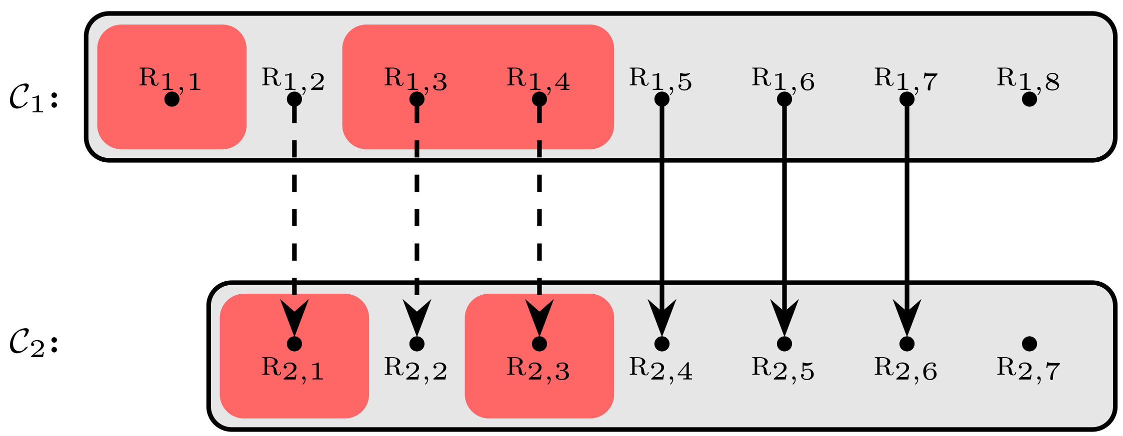
Brief Announcement: The Fault-Tolerant Cluster-sending Problem. DISC 2019

The Fault-Tolerant Cluster-Sending Problem [DISC'19]

formalizing 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	$n_{\mathcal{C}_1} > 2f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > f_{\mathcal{C}_2}$	$(f_{\mathcal{C}_1}+1)\cdot(f_{\mathcal{C}_2}+1)$	$\mathcal{O}(\Vert v\Vert)$
	RB-brs	Byzantine, RS	$n_{\mathcal{C}_1} > 2f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > f_{\mathcal{C}_2}$	$(2f_{\mathcal{C}_1}+1)\cdot(f_{\mathcal{C}_2}+1)$	$\mathcal{O}(\Vert v\Vert)$
	RB-bcs	Byzantine, RS	$n_{\mathcal{C}_1} > 2f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > f_{\mathcal{C}_2}$	$(f_{\mathcal{C}_1}+1)\cdot(f_{\mathcal{C}_2}+1)$	$\mathcal{O}(\Vert v\Vert + f_{\mathcal{C}_1})$
	RB-bcs	Byzantine, CS	$n_{\mathcal{C}_1} > 2f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > f_{\mathcal{C}_2}$	$(f_{\mathcal{C}_1}+1)\cdot(f_{\mathcal{C}_2}+1)$	$\mathcal{O}(\Vert v\Vert)$
linear	PBS-bcs	Omit	$n_{\mathcal{C}_1} > 3f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > 3f_{\mathcal{C}_2}$	$\mathcal{O}(\max(n_{\mathcal{C}_1}, n_{\mathcal{C}_2}))$ (optimal)	$\mathcal{O}(\Vert v\Vert)$
	PBS-brs	Byzantine, RS	$n_{\mathcal{C}_1} > 4f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > 4f_{\mathcal{C}_2}$	$\mathcal{O}(\max(n_{\mathcal{C}_1}, n_{\mathcal{C}_2}))$ (optimal)	$\mathcal{O}(\Vert v\Vert)$
	PBS-bcs	Byzantine, RS	$n_{\mathcal{C}_1} > 3f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > 3f_{\mathcal{C}_2}$	$\mathcal{O}(\max(n_{\mathcal{C}_1}, n_{\mathcal{C}_2}))$	$\mathcal{O}(\Vert v\Vert + f_{\mathcal{C}_1})$
	PBS-bcs	Byzantine, CS	$n_{\mathcal{C}_1} > 3f_{\mathcal{C}_1}$, $n_{\mathcal{C}_2} > 3f_{\mathcal{C}_2}$	$\mathcal{O}(\max(n_{\mathcal{C}_1}, n_{\mathcal{C}_2}))$ (optimal)	$\mathcal{O}(\Vert v\Vert)$

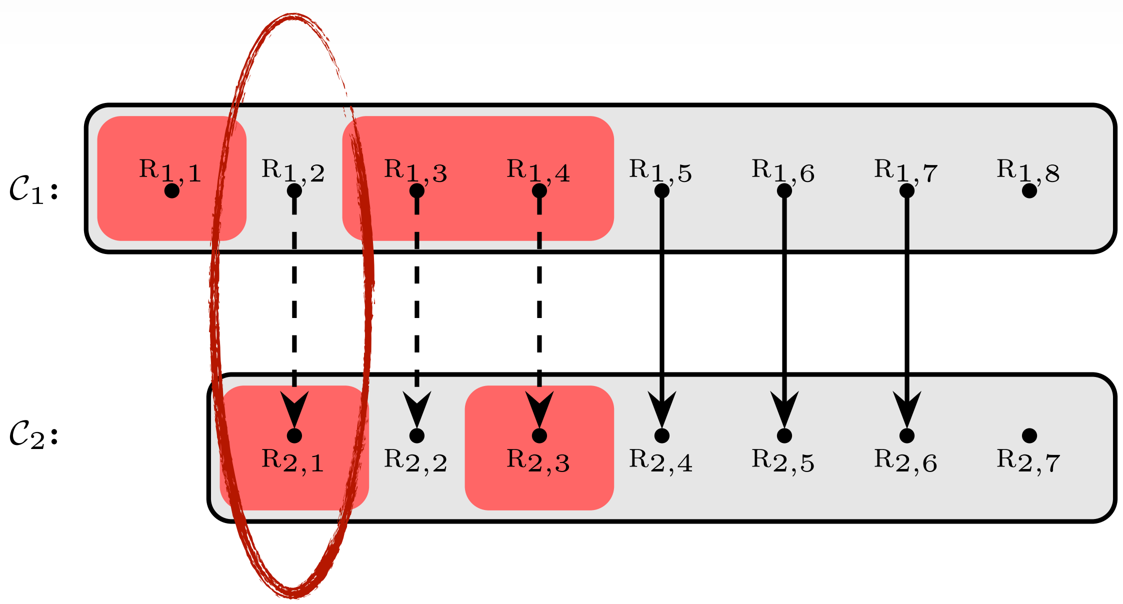
Brief Announcement: The Fault-Tolerant Cluster-sending Problem. DISC 2019

Byzantine Cluster-Sending in Expected Constant Communication [arXiv'20]

formalizing the problem of probabilistically 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

(expected constant message complexity)



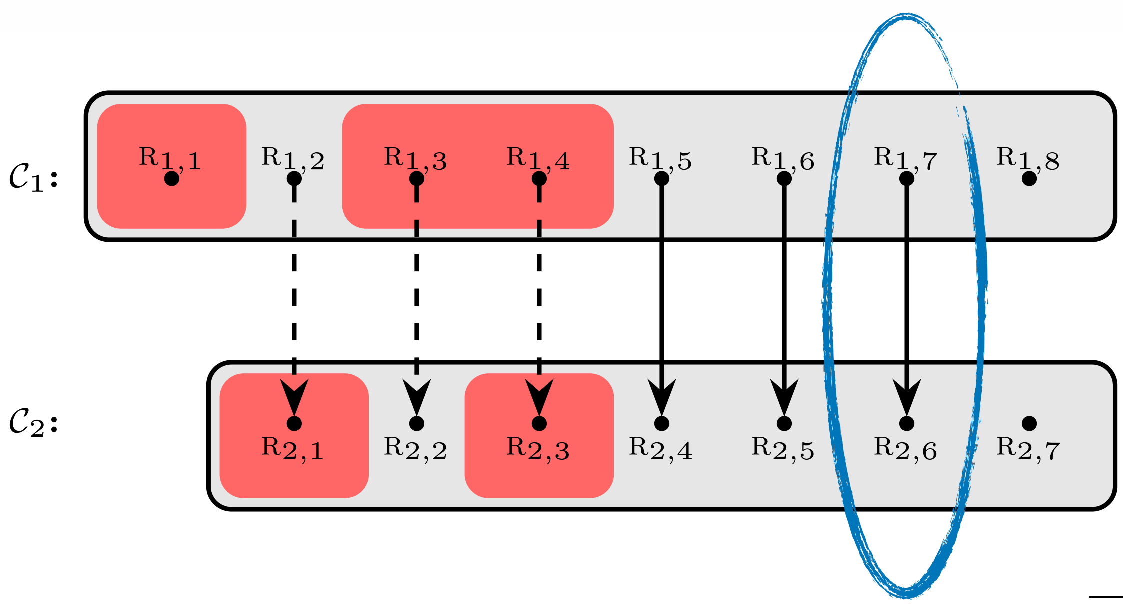
	Protocol	Robustness	Message Steps (expected)	(worst)	O.	U.
	PBS-cs [13]	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > \mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2}$	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$		✔	✘
	PBS-cs [13]	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$\max(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2})$		✔	✘
	GeoBFT [12]	$\mathbf{n}_{\mathcal{C}_1} = \mathbf{n}_{\mathcal{C}_2} > 3\max(\mathbf{f}_{\mathcal{C}_1}, \mathbf{f}_{\mathcal{C}_2})$	$\mathbf{f}_{\mathcal{C}_2} + 1^{\ddagger}$	$\Omega(\mathbf{f}_{\mathcal{C}_1}\mathbf{n}_{\mathcal{C}_2})$	✘	✔
This Paper	PPCS	$\mathbf{n}_{\mathcal{C}_1} > 2\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 2\mathbf{f}_{\mathcal{C}_2}$	4	$(\mathbf{f}_{\mathcal{C}_1} + 1)(\mathbf{f}_{\mathcal{C}_2} + 1)$	✘	✔
	PPCS	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$2\frac{1}{4}$	$(\mathbf{f}_{\mathcal{C}_1} + 1)(\mathbf{f}_{\mathcal{C}_2} + 1)$	✘	✔
	PLCS	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > \mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2}$	4	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$	✔	✔
	PLCS	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > 2(\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2})$	$2\frac{1}{4}$	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$	✔	✔
	PLCS	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	3	$\max(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2})$	✔	✔

Byzantine Cluster-Sending in Expected Constant Communication [arXiv'20]

formalizing the problem of probabilistically 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

(expected constant message complexity)

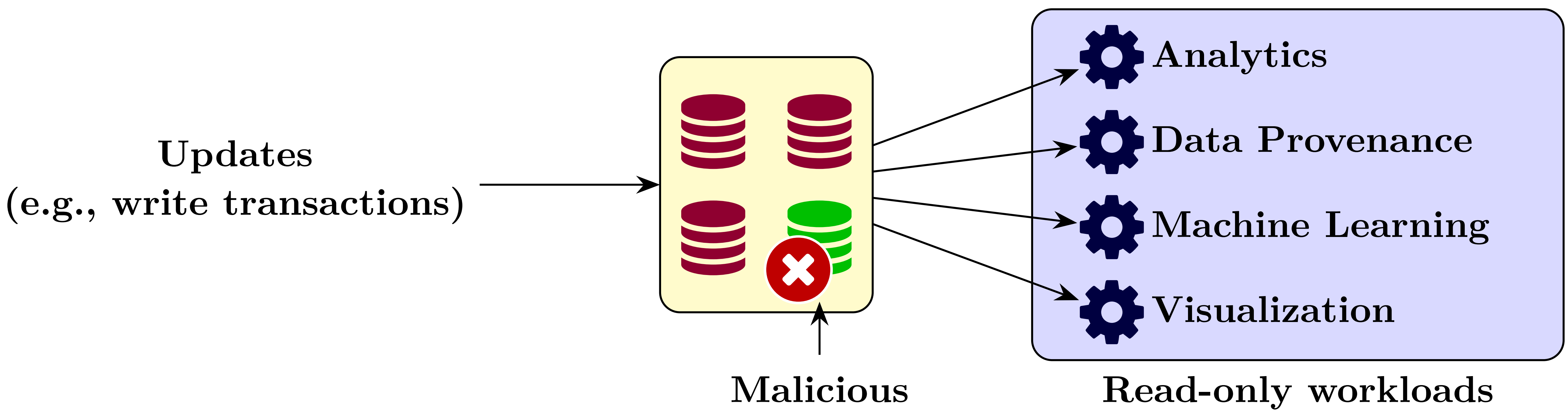


	Protocol	Robustness	Message Steps (expected)	(worst)	O.	U.
	PBS-cs [13]	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > \mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2}$	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$		✔	✖
	PBS-cs [13]	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$\max(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2})$		✔	✖
	GeoBFT [12]	$\mathbf{n}_{\mathcal{C}_1} = \mathbf{n}_{\mathcal{C}_2} > 3\max(\mathbf{f}_{\mathcal{C}_1}, \mathbf{f}_{\mathcal{C}_2})$	$\mathbf{f}_{\mathcal{C}_2} + 1^{\ddagger}$	$\Omega(\mathbf{f}_{\mathcal{C}_1}\mathbf{n}_{\mathcal{C}_2})$	✖	✔
This Paper	PPCS	$\mathbf{n}_{\mathcal{C}_1} > 2\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 2\mathbf{f}_{\mathcal{C}_2}$	4	$(\mathbf{f}_{\mathcal{C}_1} + 1)(\mathbf{f}_{\mathcal{C}_2} + 1)$	✖	✔
	PPCS	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$2\frac{1}{4}$	$(\mathbf{f}_{\mathcal{C}_1} + 1)(\mathbf{f}_{\mathcal{C}_2} + 1)$	✖	✔
	PLCS	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > \mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2}$	4	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$	✔	✔
	PLCS	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > 2(\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2})$	$2\frac{1}{4}$	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$	✔	✔
	PLCS	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	3	$\max(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2})$	✔	✔

Coordination-Free Byzantine Replication With Minimal Communication Costs [ICDT'20]

formalizing the Byzantine learner problem to support efficient analytics for blockchain applications

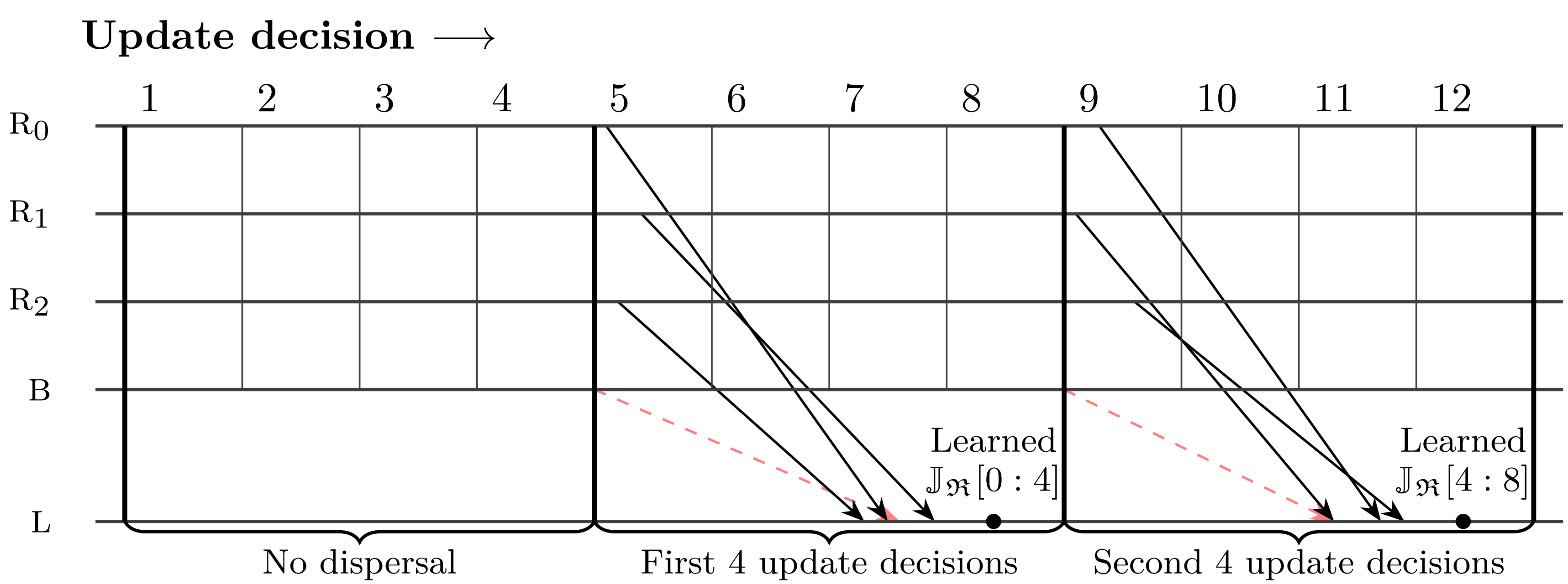
introducing the delayed-replication algorithm, utilizing information dispersal techniques, giving rise to a coordination-free, push-based, minimal communication protocol



Coordination-Free Byzantine Replication With Minimal Communication Costs [ICDT'20]

formalizing the Byzantine learner problem to support efficient analytics for blockchain applications

introducing the delayed-replication algorithm, utilizing information dispersal techniques, giving rise to a coordination-free, push-based, minimal communication protocol



System	Checksum	Complexity for the learner		
		Data sent per replica	*Data received*	*Decode steps*
$\mathbf{b} = 0$	None	$\mathcal{O}(s/\mathbf{g})$	$\mathcal{O}(s(\mathbf{n}/\mathbf{g}))$	$u/\mathbf{n}$
$\mathbf{b} < \mathbf{g}$	Simple	$\mathcal{O}(s/\mathbf{g})$	$\mathcal{O}(s(\mathbf{n}/\mathbf{g}))$	$\binom{\mathbf{g}+\mathbf{b}}{\mathbf{g}}(u/\mathbf{n})$
$\mathbf{b} < \mathbf{g}$	Tree	$\mathcal{O}(s/\mathbf{g} + (u/\mathbf{n})\log(\mathbf{n}))$	$\mathcal{O}(s(\mathbf{n}/\mathbf{g}) + u\log(\mathbf{n}))$	$u/\mathbf{n}$

Permissioned Blockchain Through the Looking Glass: Architectural and Implementation Lessons Learned [ICDCS'20]

Single-threaded Monolithic Design

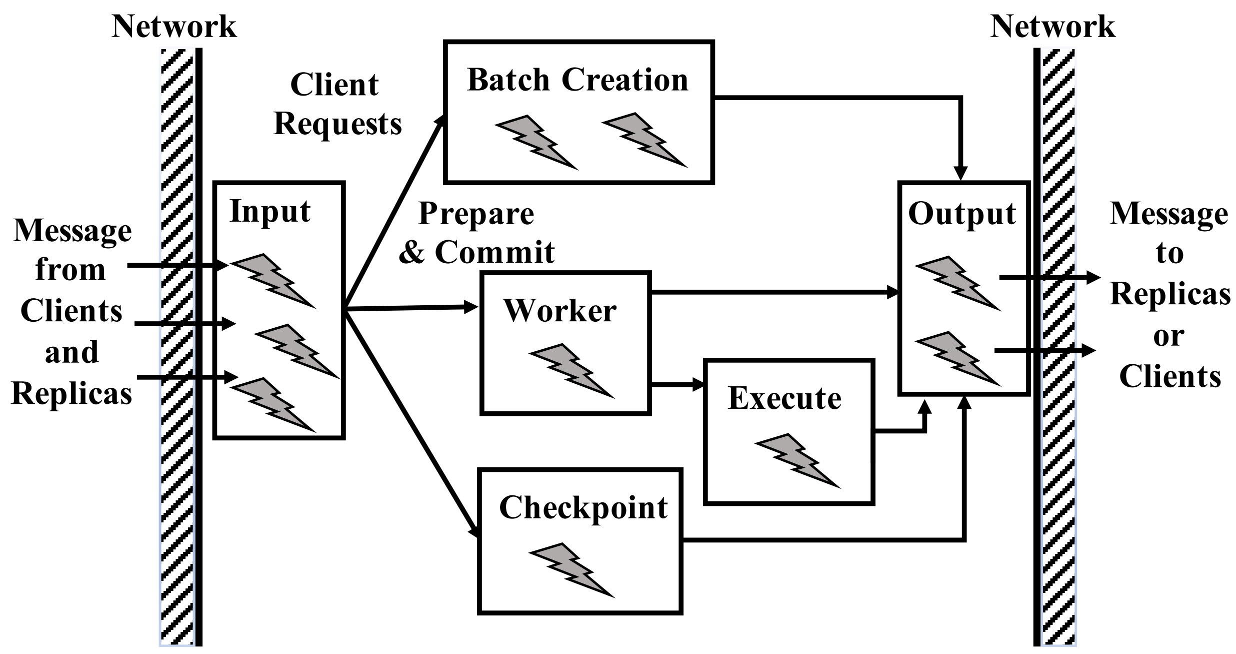
Out-of-ordering Consensus Communication

De-coupled Ordering and Execution

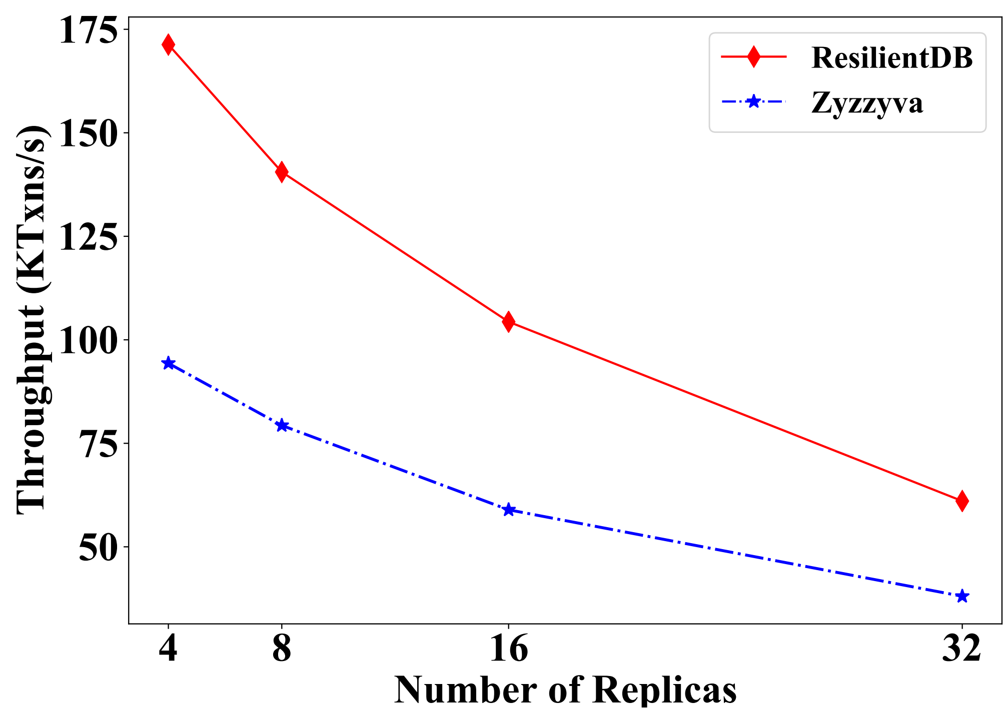
Off-Chain Memory Management

Expensive Cryptographic Practices (DS vs. MAC)

Smart Contracts Code Generation (Pre-compilation)



Multi-Threaded Deep Pipeline



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

Revisit Resiliency

(Graduate Student Experiment Continues)

Mount Tallac, Lake Tahoe

12.1 Miles Long

3,931 Feet Elevation Gain

(9,738 Feet at Summit)

Fostering Resiliency

(Offering Stress Management and Well-Being Courses at UC Davis)

THE CALIFORNIA AGGIE

Seminar spotlight: "Becoming an Extraordinary Human"

The California Aggie, April 6, 2020

M.T.O. Tamarkoz®
BE BALANCED

Thank You

For Complete References