The Journey of Building Global-Scale ResilientDB Blockchain Fabric

Mohammad Sadoghi

URCS Seminar Series
University of Rochester
November 5, 2021

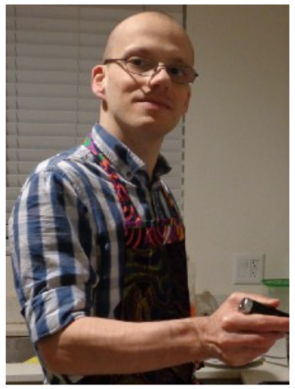
Mohammad Sadoghi
Exploratory Systems Lab
Department of Computer Science
UC DAVIS
UNIVERSITY OF CALIFORNIA

ExpoLab
Creativity Unfolded

ExpoLab Team

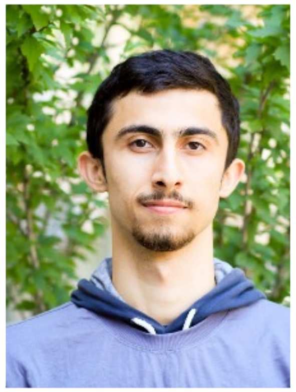
Mohammad Sadoghi
(Principal Investigator)

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



Jelle Hellings, PostDoc
(Fault-tolerant Complexity Analysis)

Suyash Gupta, PhD
(Scalable Consensus Meta-Protocols)



Sajjad Rahnama, PhD
(Global Scale Consensus)

Haojun (Howard) Zhu, MSc
(Re-Configurable Consensus Protocols)

Alejandro Armas, BSc
(Re-engineering ResilientDB Toolkits)

Dakai Kang, BSc
(View-change-less Protocols)

Dhruv Krishnan, MSc
(Scaling Fabric via Sharding)

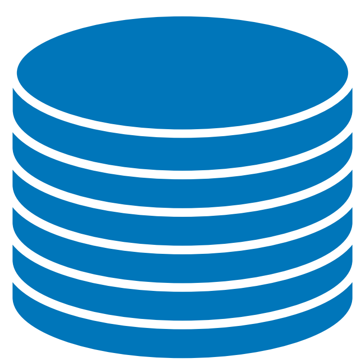
Priya Holani, MSc
(Scaling Fabric via Sharding)

Shubham Pandey, MSc
(Scaling Fabric via RDMA)

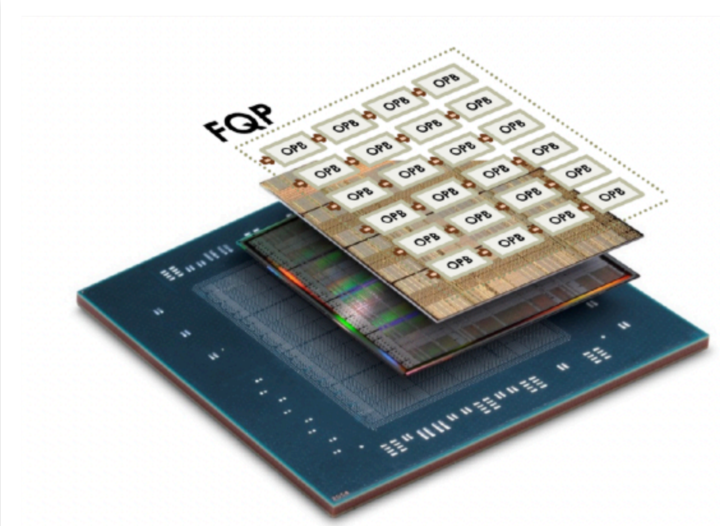
Rohan Sogani, MSc
(Scaling Fabric via Sharding)

Xinyuan Sun, MSc
(Scaling Fabric via RDMA)

Resilient Journey…



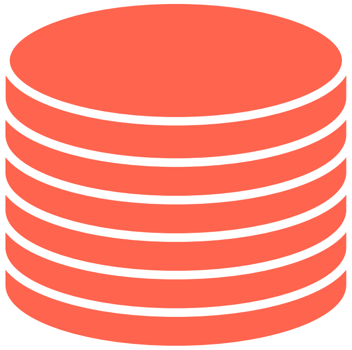
SQL Analytics



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…

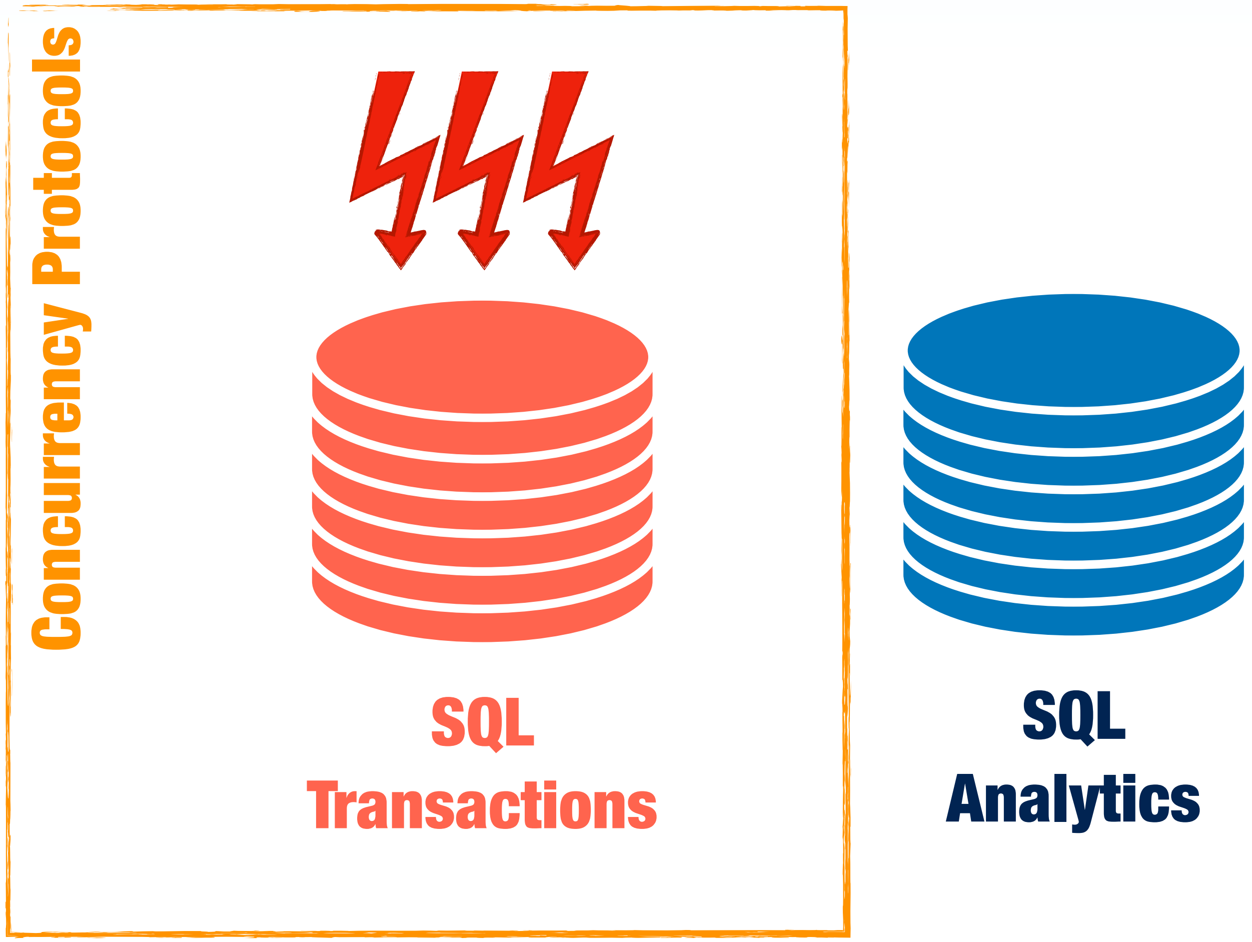


SQL Transactions

SQL Analytics

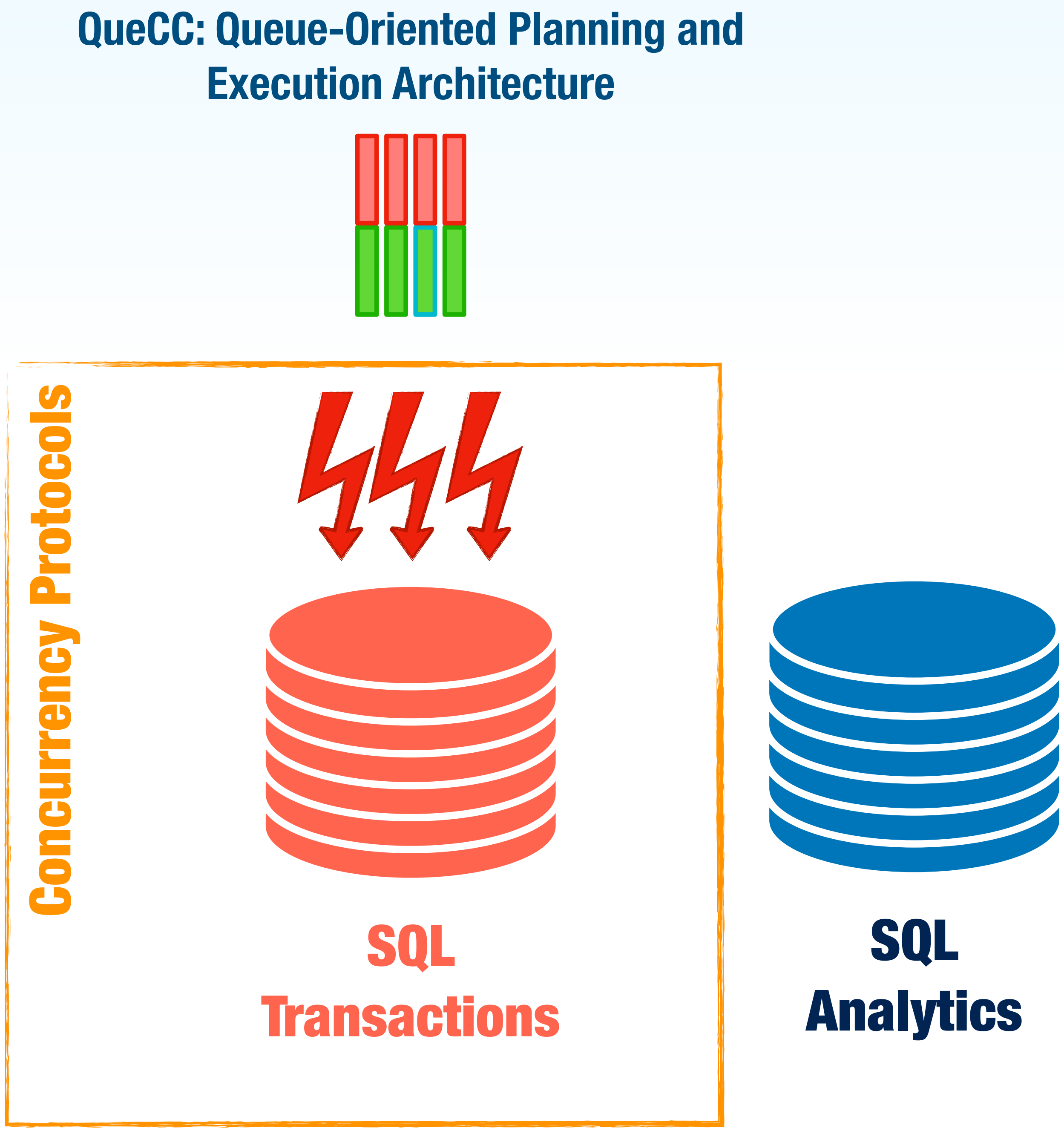
High-dimensional Indexing: (e.g., BE-Tree, BE-topK)
[SIGMOD’11, ICDE’12, TODS’13, ICDCS’13, ICDE’14, ICDCS’17, Middleware'17]

Resilient Journey…



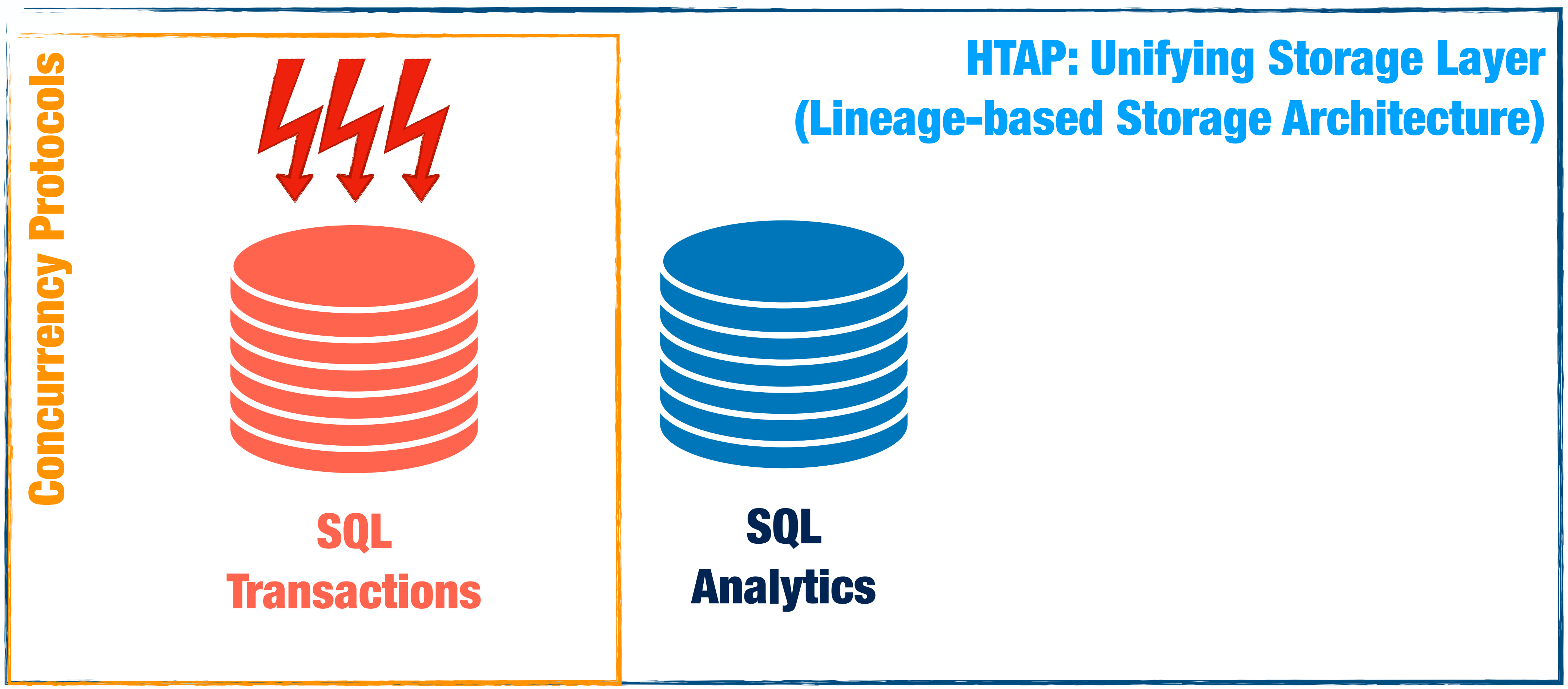
Concurrency Control Protocols: (e.g., 2VCC, QueCC - Best Paper Award)
[VLDB'13, VLDB'14, VLDBJ'16, Middleware'16, TDKE'15, SIGMOD'15, ICDE'16, Middleware'18]

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…

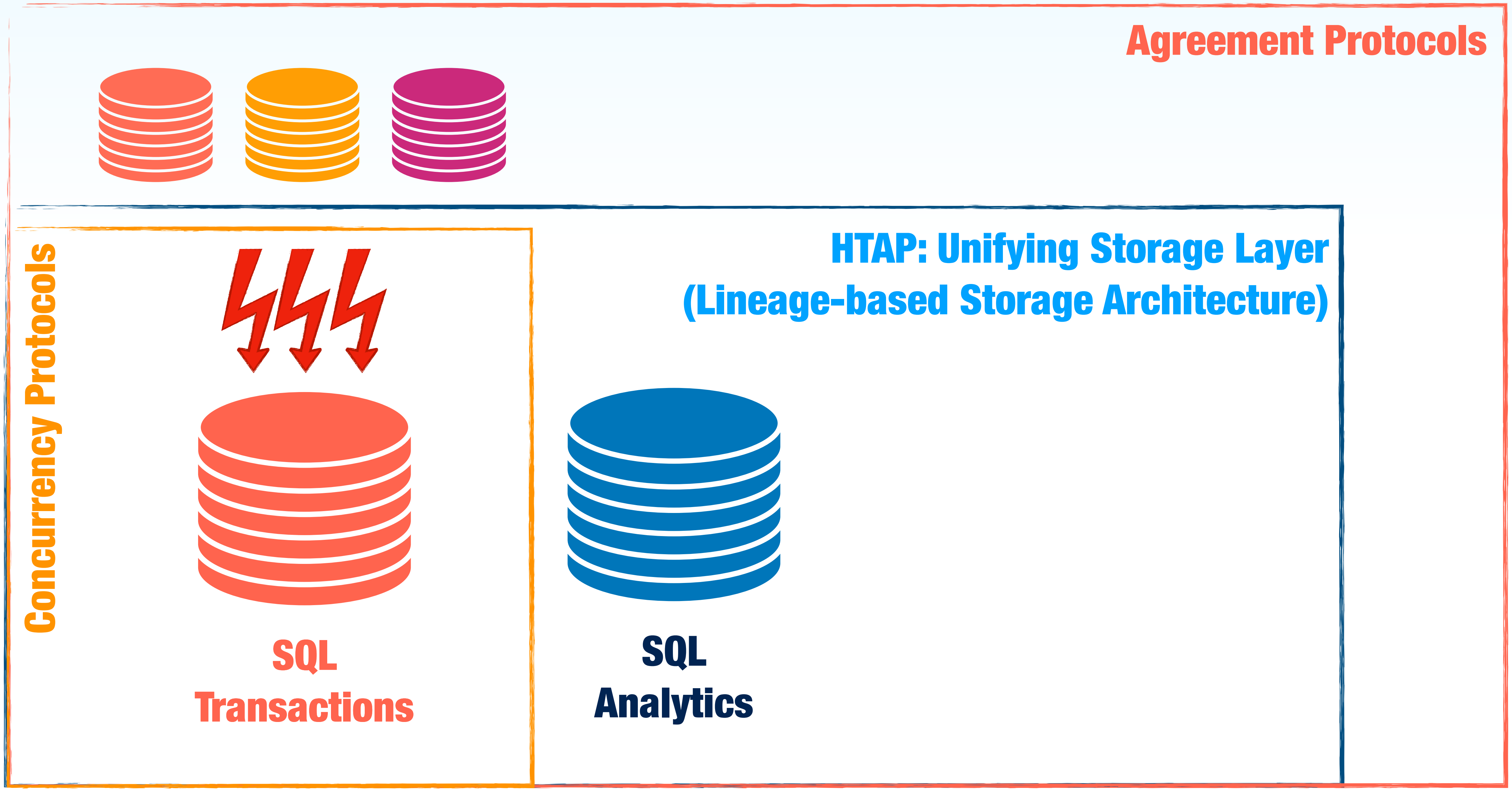


HTAP Column-store: L-Store (Lineage-based Data Store)

[VLDB’12, ICDE’14, ICDCS’16, EDBT’18, TKDE’20 (2x) 34 filed US patents]

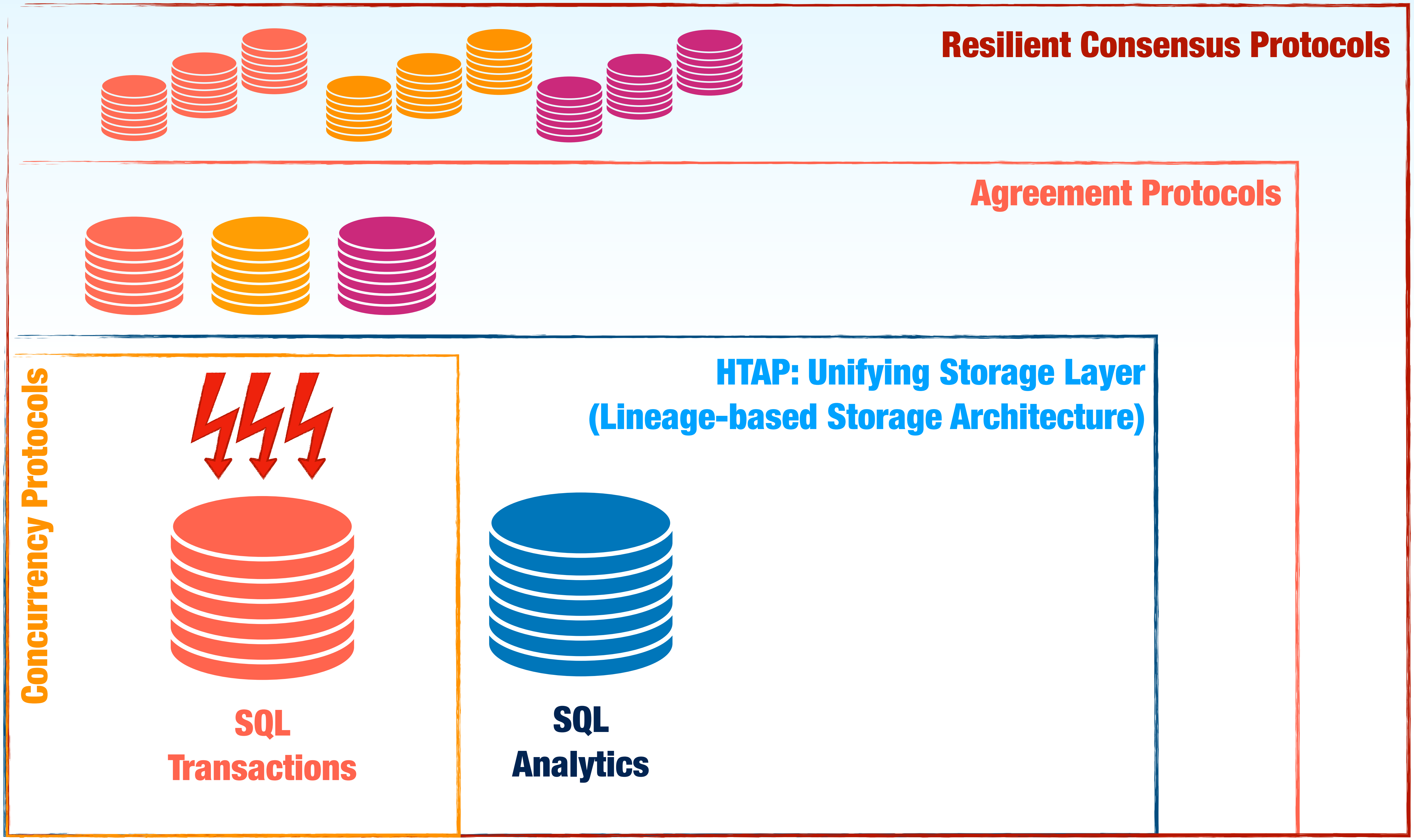
Graphs on SQL: (e.g., GRFusion) [SIGMOD’18, EDBT’18]

Resilient Journey…

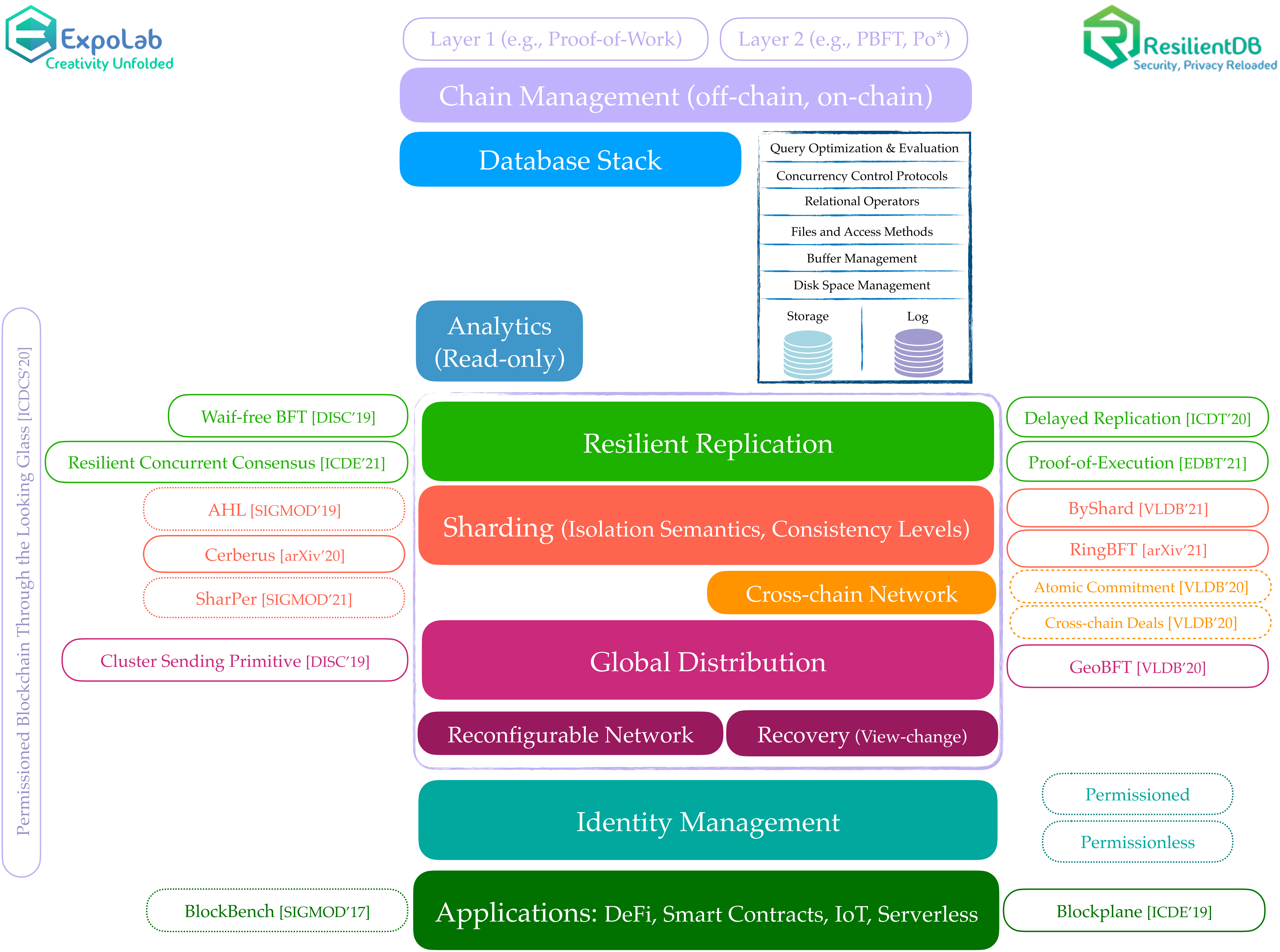


Agreement Protocols: (e.g., EasyCommit, Q-Store)
[EDBT'20, EDBT'18, DAPD'19]

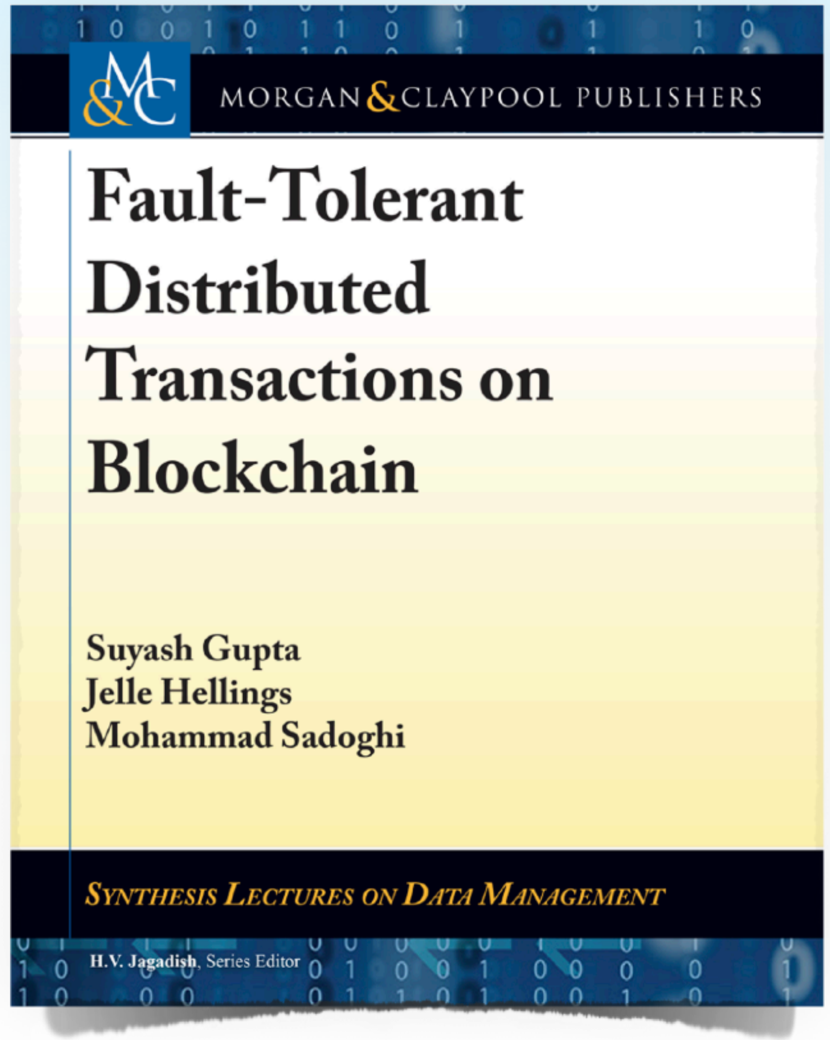
Resilient Journey…

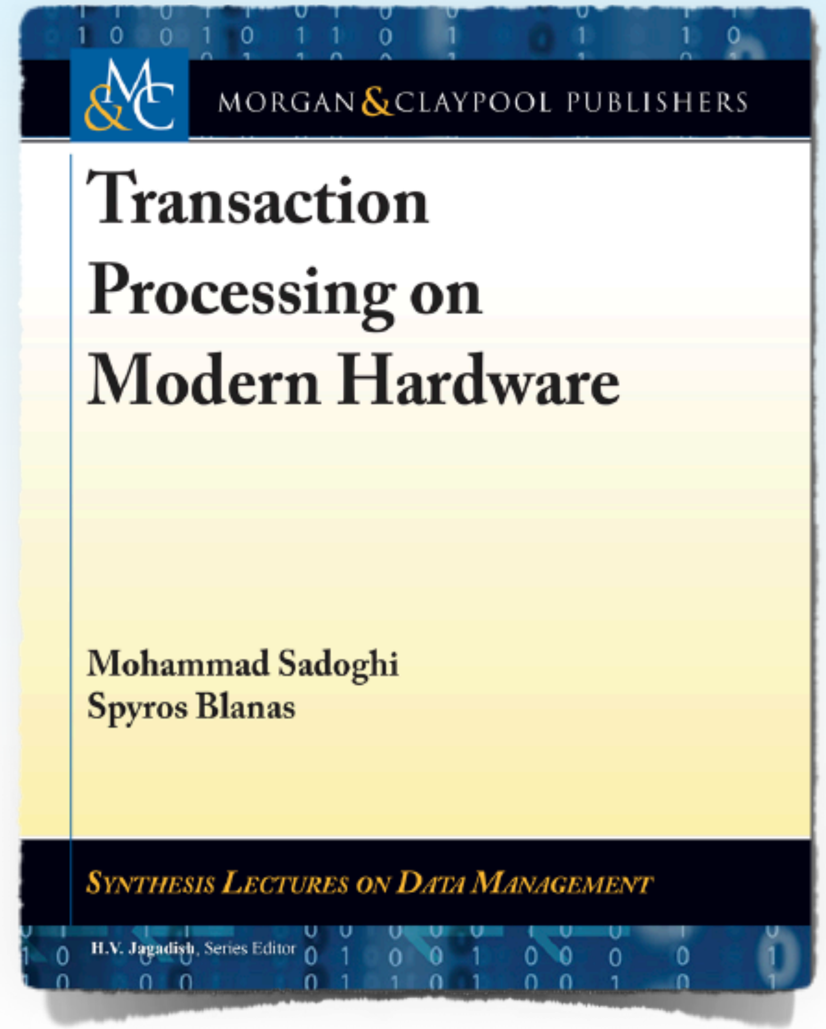


Consensus Protocols: (e.g., GeoBFT, PoE, RCC, ByShard, RingBFT, Delayed Replication, CSP, Blockplane)
[VLDB'21, ICDE'21, EDTB'21, VLDB'20, ICDCS'20, ICDT'20, DISC'19 (2x), SC'19, ICDE'19, arXiv'19 (8x)]



ExpoLab
Creativity Unfolded
ResilientDB
Security, Privacy Reloaded
Layer 1 (e.g., Proof-of-Work)
Layer 2 (e.g., PBFT, Po*)
Chain Management (off-chain, on-chain)
Database Stack
Query Optimization & Evaluation
Concurrency Control Protocols
Relational Operators
Files and Access Methods
Buffer Management
Disk Space Management
Storage
Log
Analytics (Read-only)
Permissioned Blockchain Through the Looking Glass [ICDCS'20]
Waif-free BFT [DISC'19]
Resilient Concurrent Consensus [ICDE'21]
AHL [SIGMOD'19]
Cerberus [arXiv'20]
SharPer [SIGMOD'21]
Cluster Sending Primitive [DISC'19]
Resilient Replication
Sharding (Isolation Semantics, Consistency Levels)
Cross-chain Network
Global Distribution
Reconfigurable Network
Recovery (View-change)
Identity Management
Applications: DeFi, Smart Contracts, IoT, Serverless
Delayed Replication [ICDT'20]
Proof-of-Execution [EDBT'21]
ByShard [VLDB'21]
RingBFT [arXiv'21]
Atomic Commitment [VLDB'20]
Cross-chain Deals [VLDB'20]
GeoBFT [VLDB'20]
Permissioned
Permissionless
BlockBench [SIGMOD'17]
Blockplane [ICDE'19]





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 *2021*

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 *2021*

ExpoLab

Creativity Unfolded

UCDAVIS

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

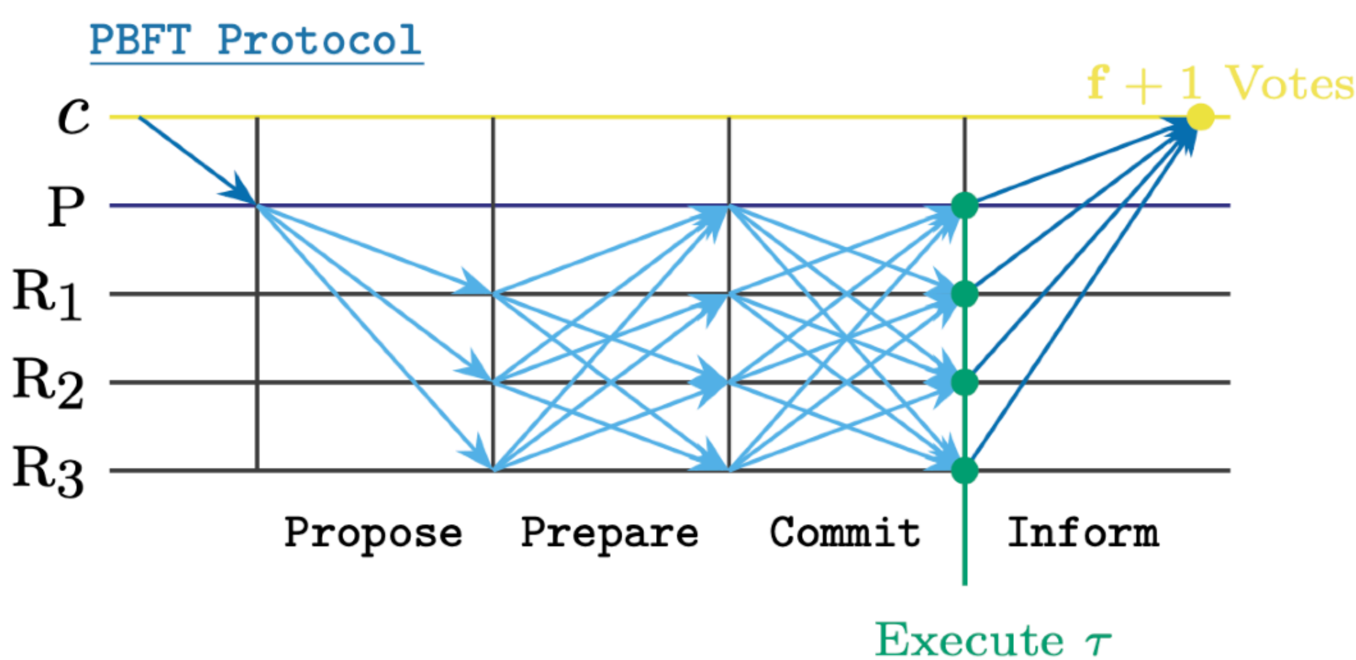
Non-Quantifiable Resiliency

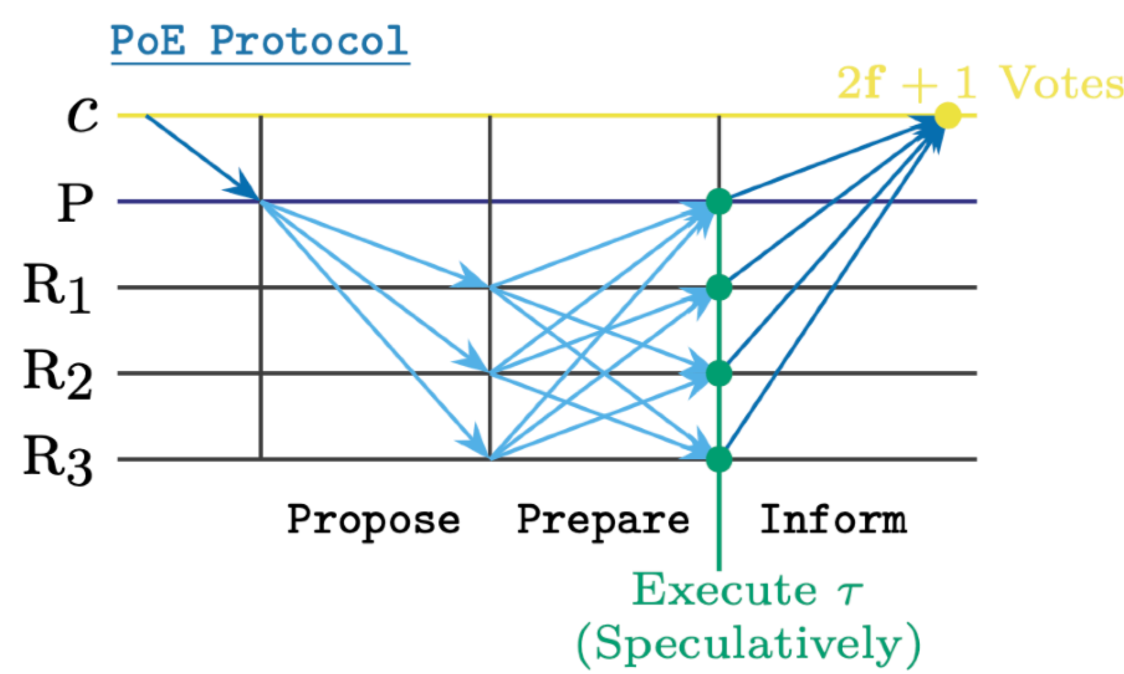
Proof-of-Execution: Reaching Consensus Through Fault-Tolerant Speculation [EDBT'21]

Out-of-Order message processing to reduce replica idleness

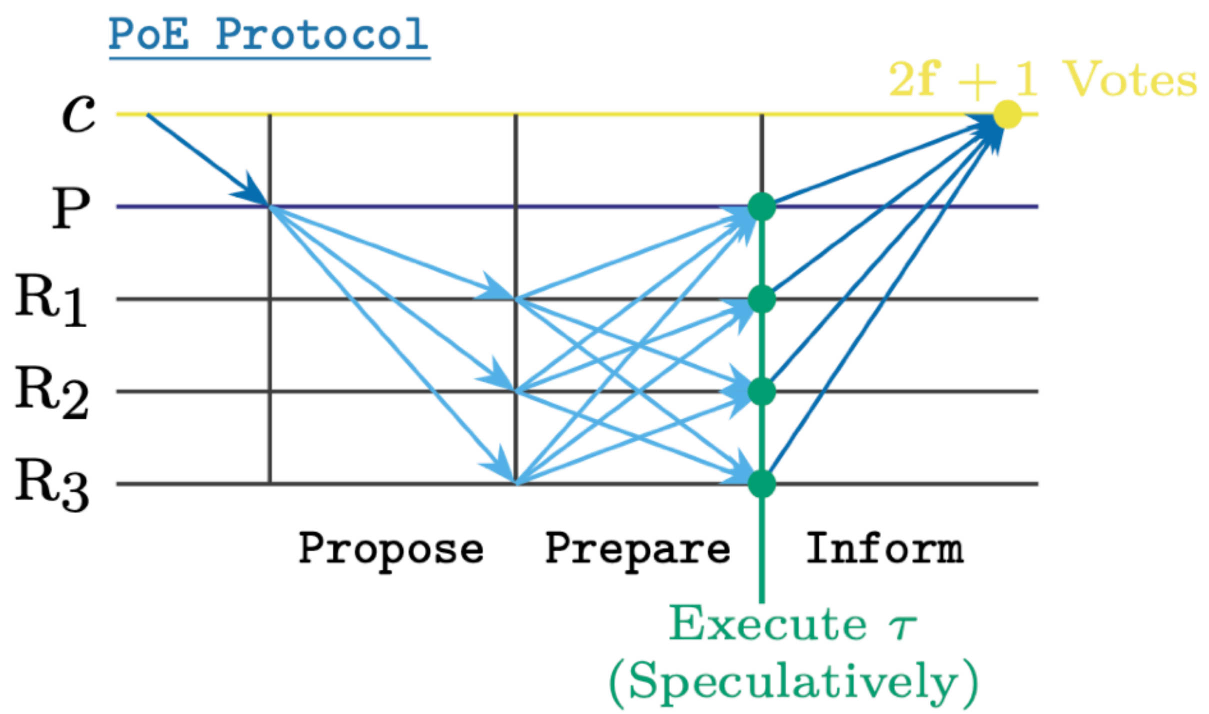
Speculative Execution with revertible/divergent replicas & eager/irrevertible client commit

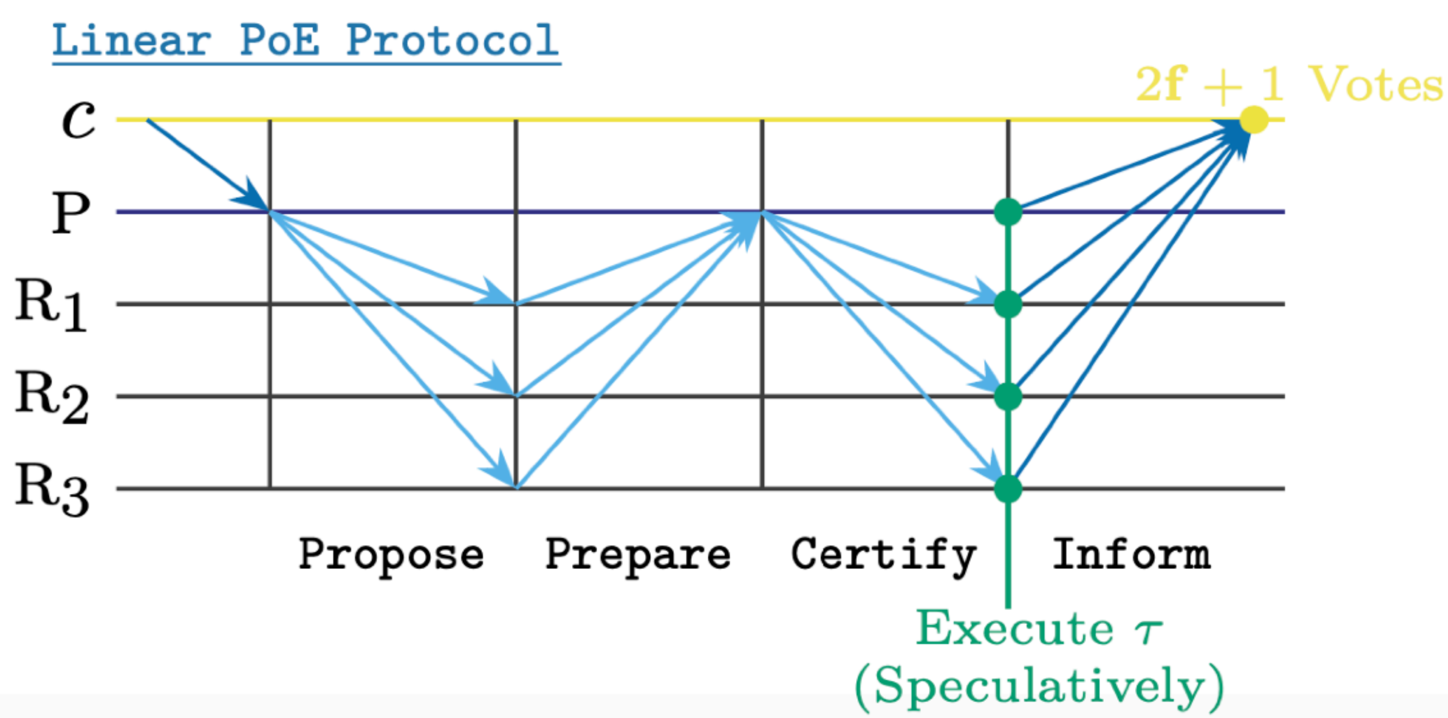
introducing linear message complexity



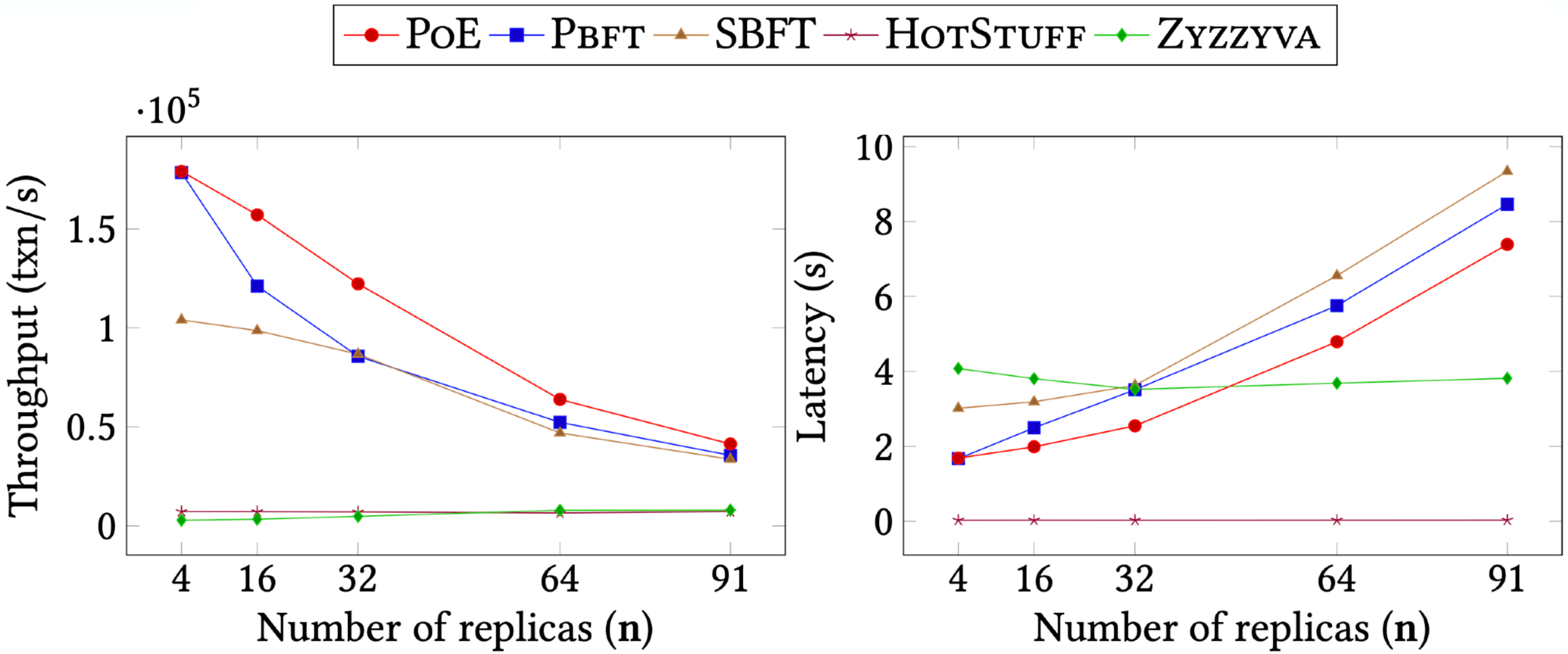


Proof-of-Execution: Reaching Consensus Through Fault-Tolerant Speculation [EDBT'21]





Proof-of-Execution: Reaching Consensus Through Fault-Tolerant Speculation [EDBT'21]



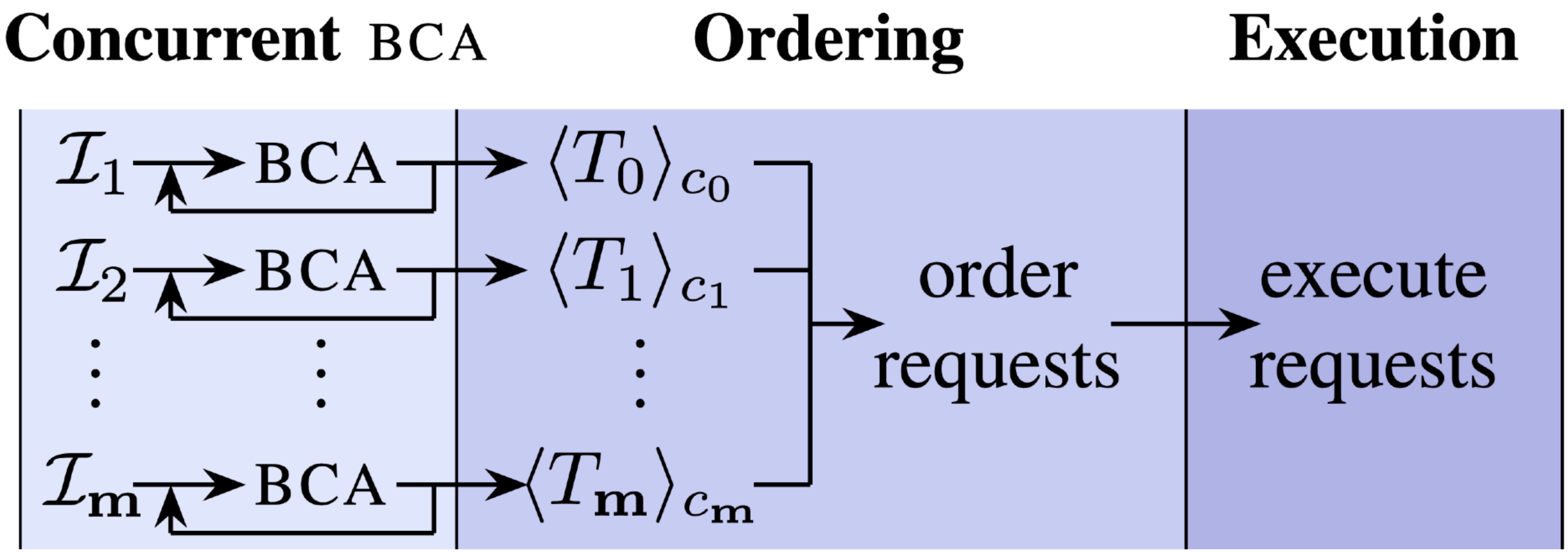
PoE scales beyond 91 replicas, in presence of failures, outperforms PBFT up to 43%

RCC: Resilient Concurrent Consensus Paradigm [ICDE'21]

A wait-free meta-protocol…

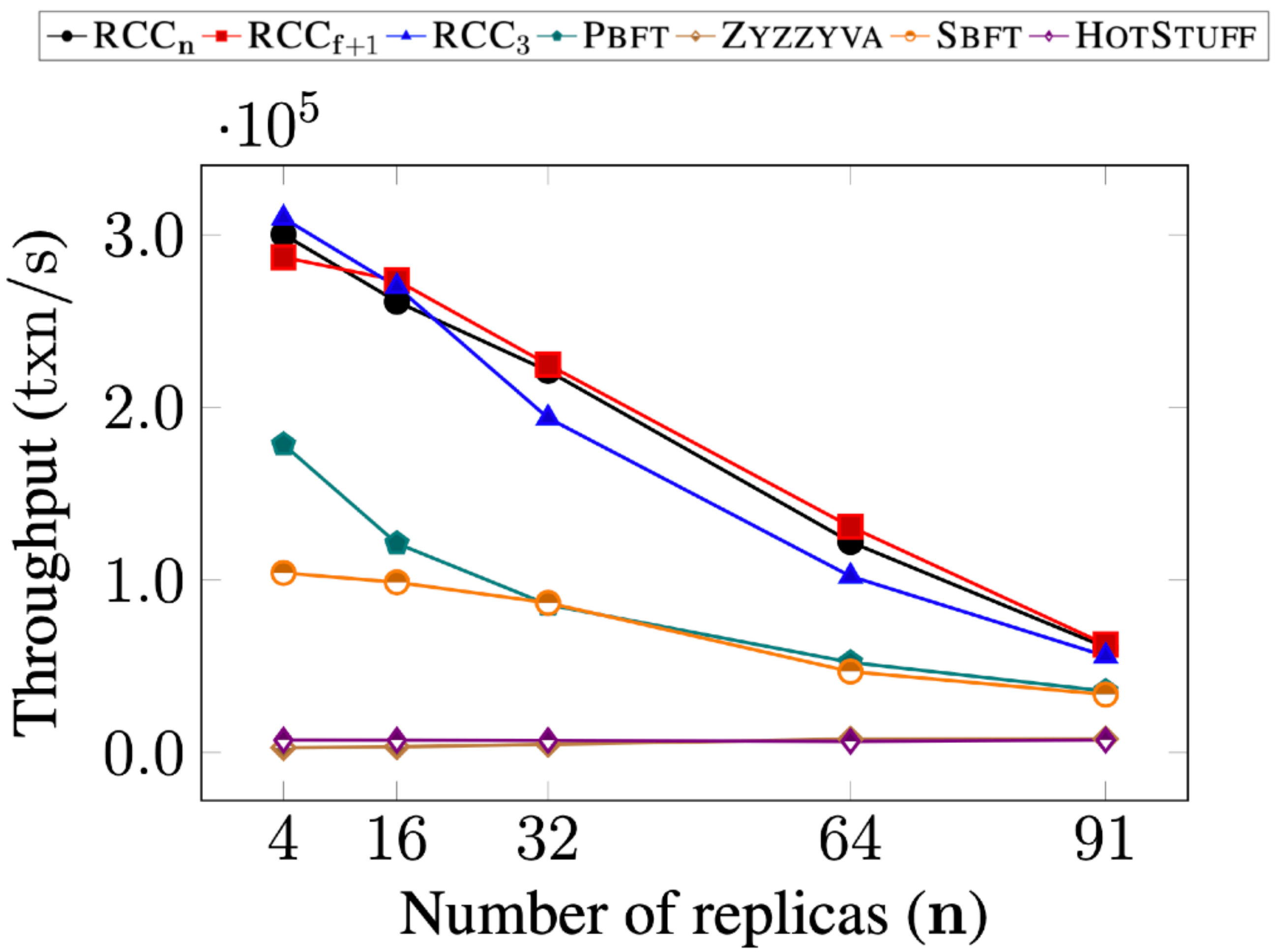
Designate multiple replicas as primaries!

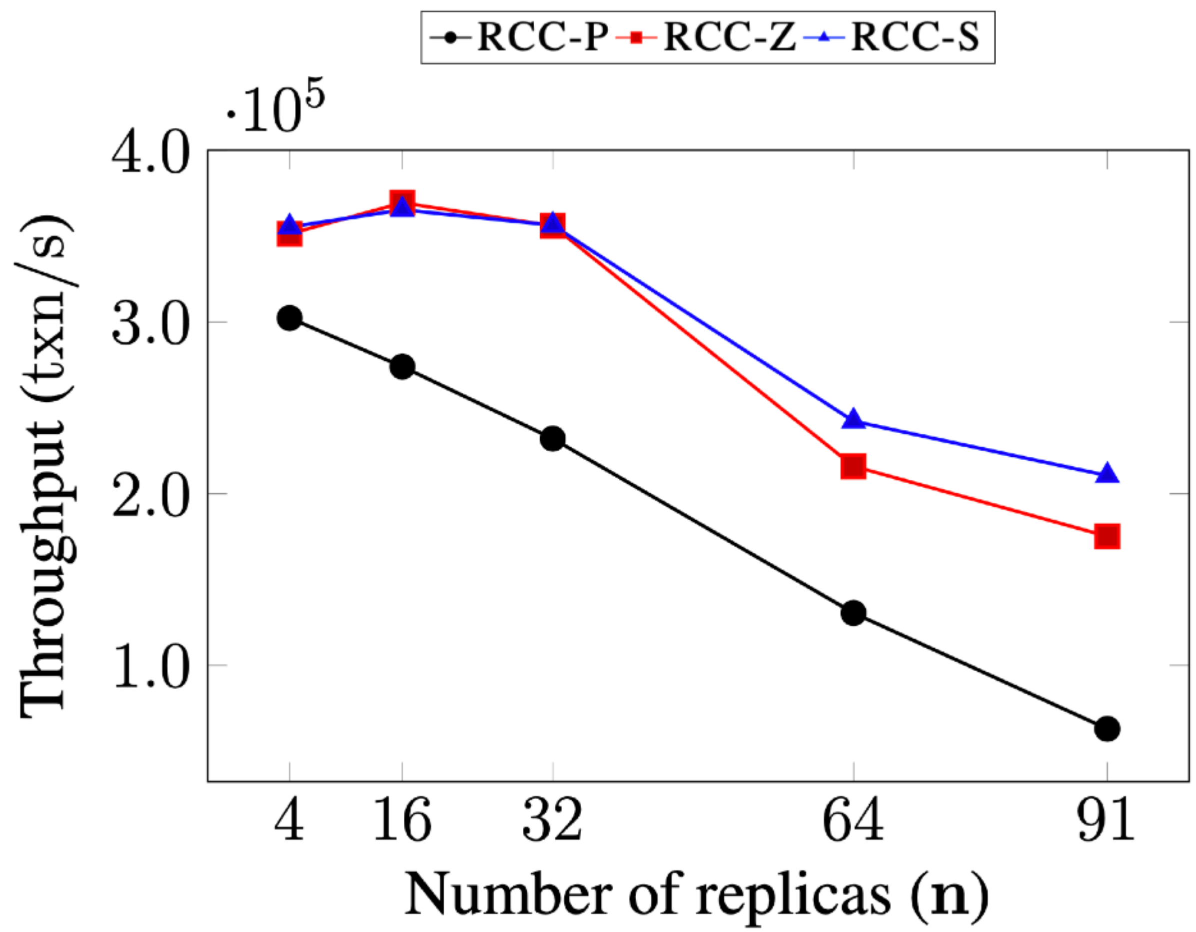
Run multiple parallel consensuses on each replica independently



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

RCC: Resilient Concurrent Consensus Paradigm [ICDE'21]





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

Throughput up to <u>400,000 txns/s</u>
(without failures)

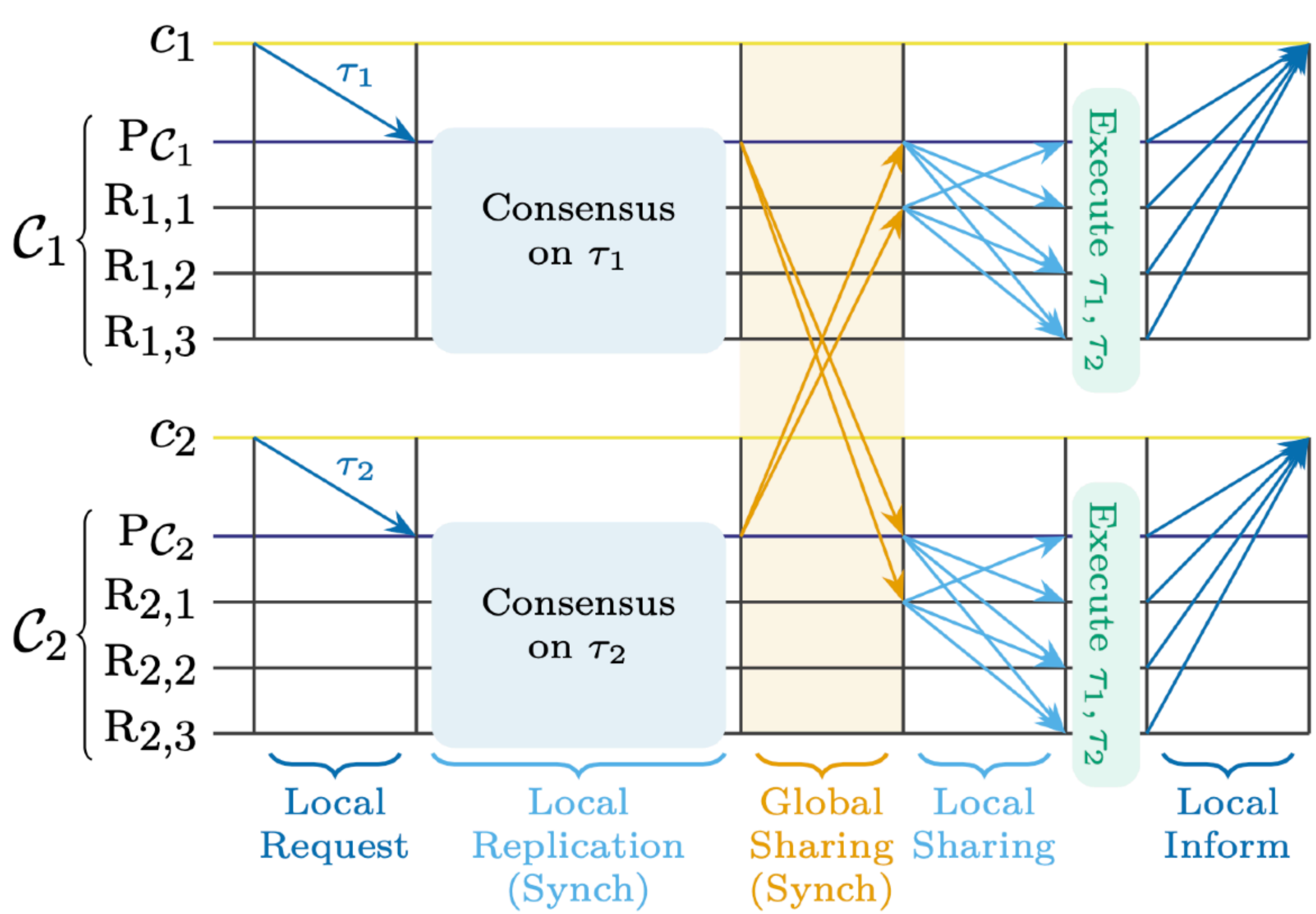
Brief Announcement: Revisiting Consensus Protocols through Wait-free Parallelization. DISC 2019

GeoBFT: Global Scale Resilient Consensus [VLDB'20]

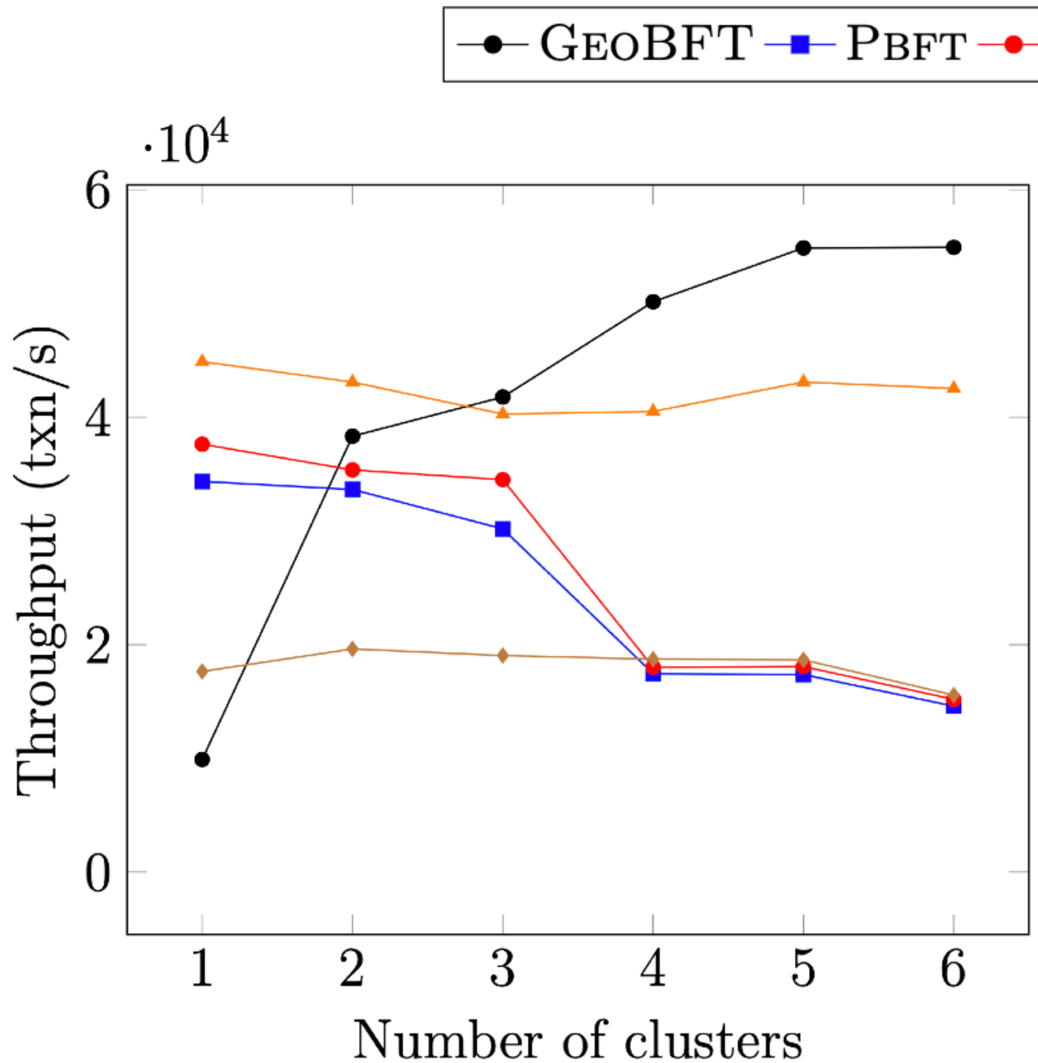
A meta-protocol, locally running any BFT in parallel and independently

Global ordering provably requires only linear communication

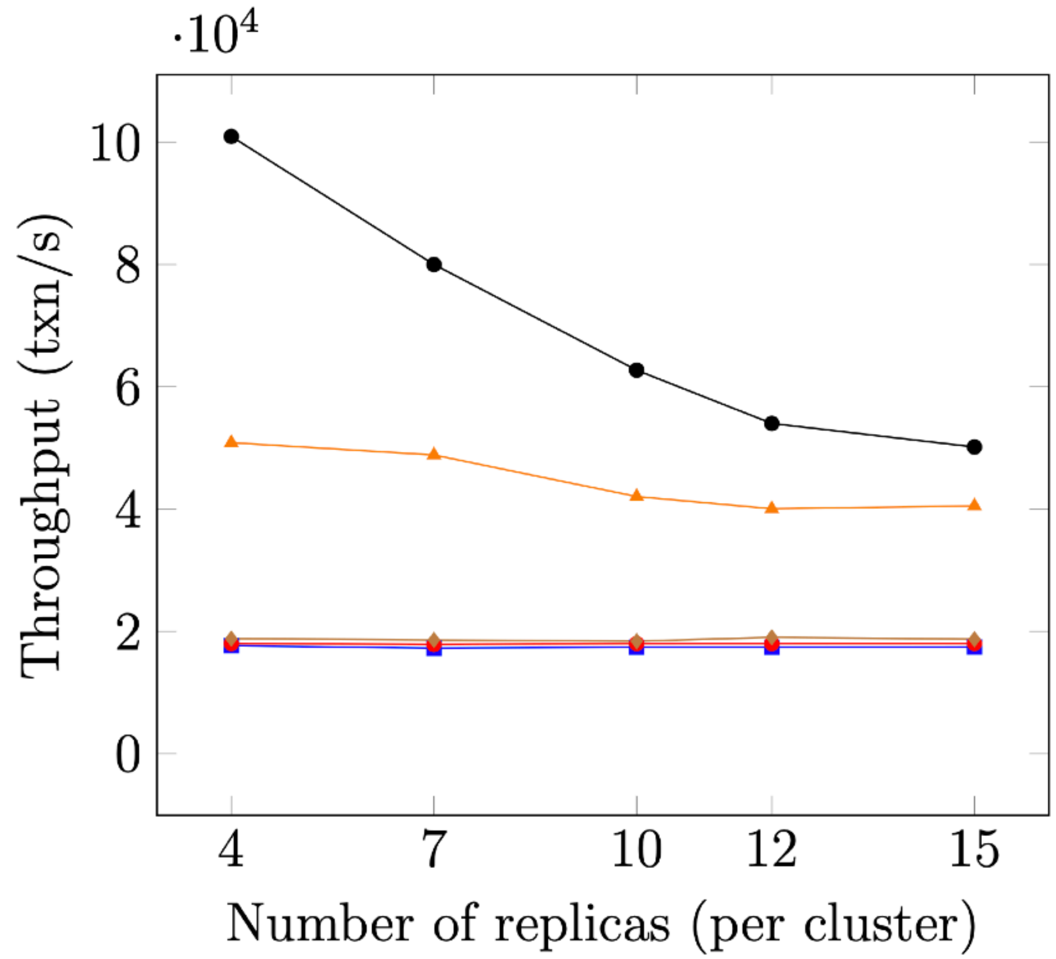
Provably sufficient for primary to send a certificate to at most f+1 replicas, malicious primary is detectable and replaceable



GeoBFT: Global Scale Resilient Consensus [VLDB'20]



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



GeoBFT scales a permissioned blockchain up to 60 replicas globally.

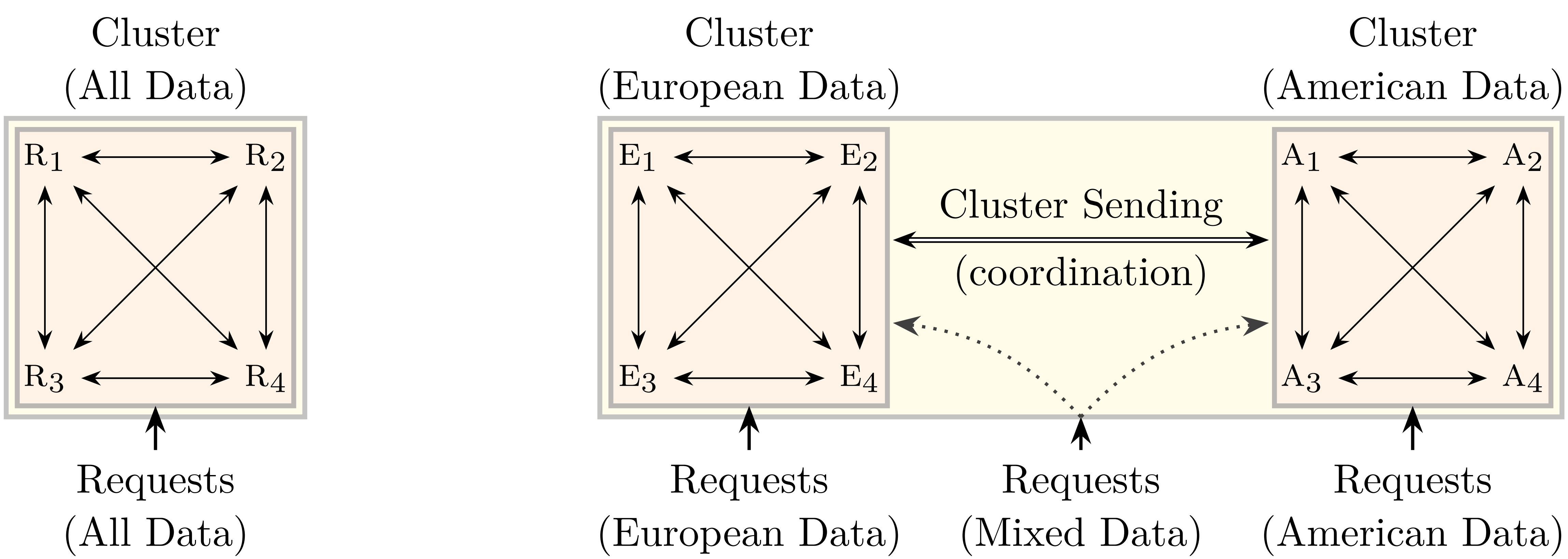
Brief Announcement: The Fault-Tolerant Cluster-sending Problem. DISC 2019

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)



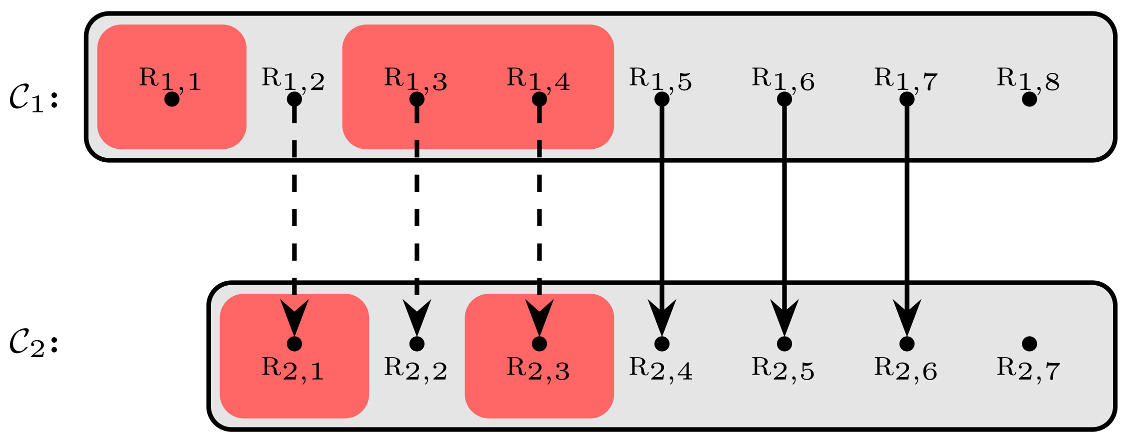
Brief Announcement: The Fault-Tolerant Cluster-sending Problem. DISC 2019

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

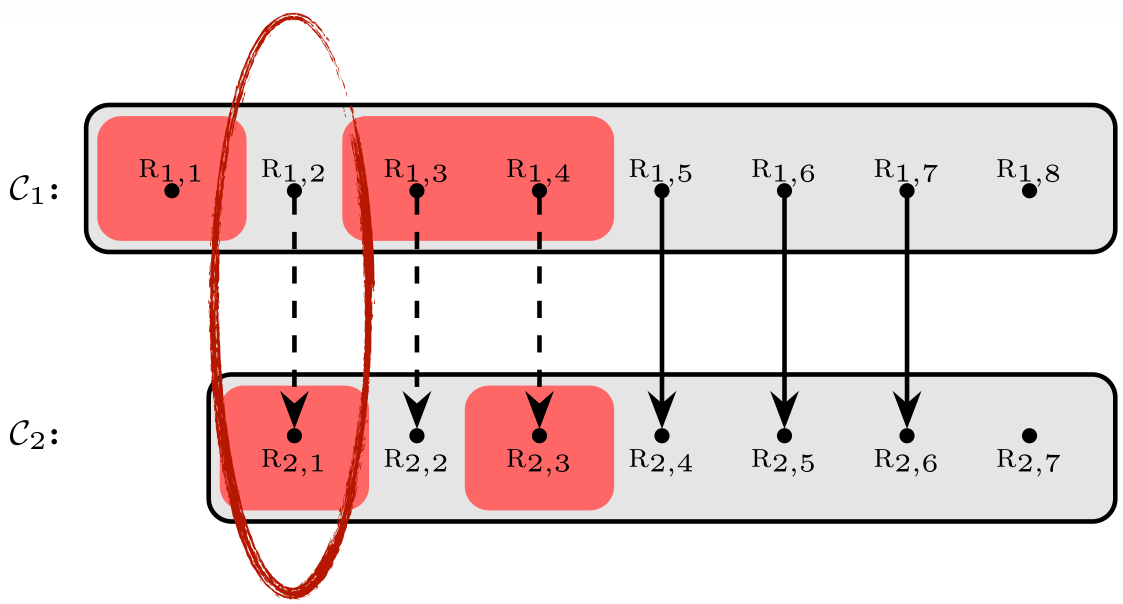
Brief Announcement: The Fault-Tolerant Cluster-sending Problem. DISC 2019

Byzantine Cluster-Sending in Expected Constant Communication [arXiv'21]

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)



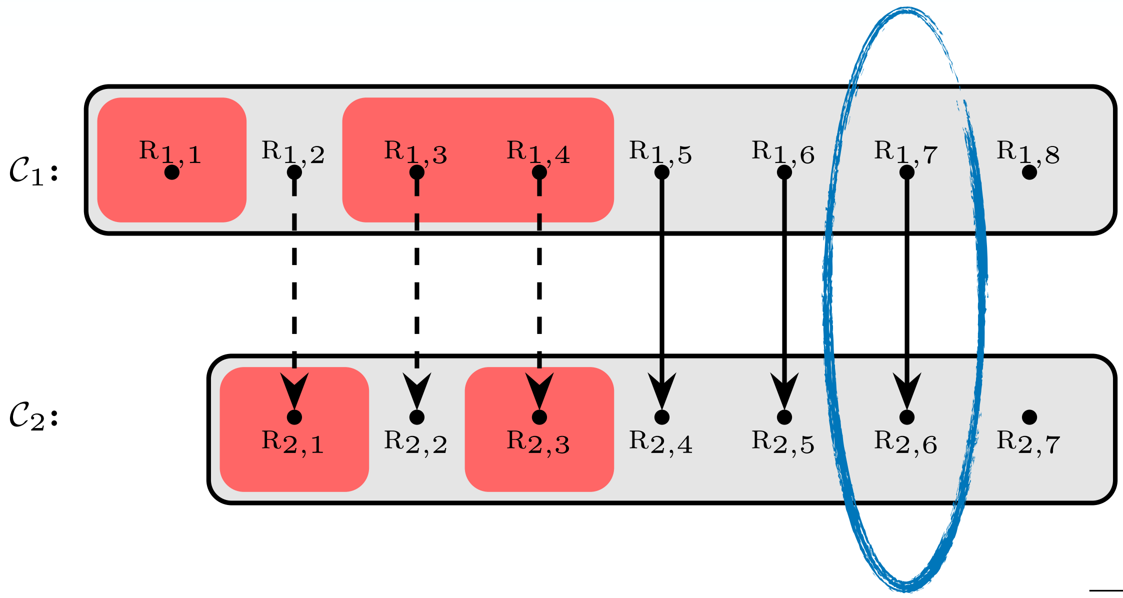
	Protocol	Robustness	Message Steps (expected)	Message Steps (worst)	O.	U.
	PBS-cs [13]	$\min(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2}) > \mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2}$	$\mathbf{f}_{\mathcal{C}_1} + \mathbf{f}_{\mathcal{C}_2} + 1$		✔	✘
	PBS-cs [13]	$\mathbf{n}_{\mathcal{C}_1} > 3\mathbf{f}_{\mathcal{C}_1}$, $\mathbf{n}_{\mathcal{C}_2} > 3\mathbf{f}_{\mathcal{C}_2}$	$\max(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2})$		✔	✘
	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'21]

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)	Message Steps (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	$\max(\mathbf{n}_{\mathcal{C}_1}, \mathbf{n}_{\mathcal{C}_2})$	✔	✔

RingBFT: Resilient Consensus Over Sharded Ring Topology [arXiv'21]

A meta-protocol adhering to the ring order, and follow the principle of

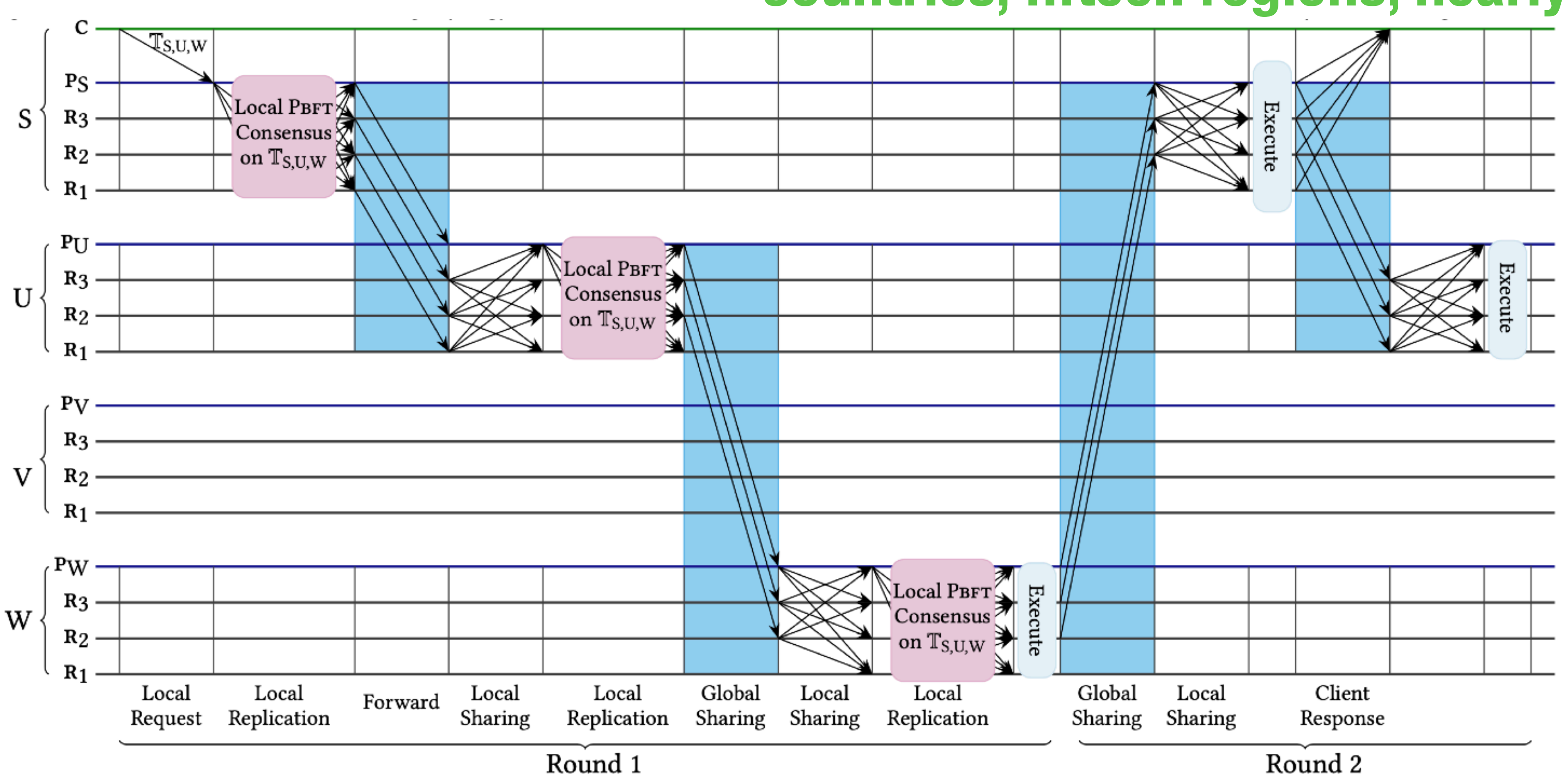
process, forward, and re-transmit

Guarantees consensus for each cross-shard transaction in

at most two rotations around the ring

Sustaining over 1,200,000 transactions per second when deployed globally spanning ten

countries, fifteen regions, nearly 500 replicas.



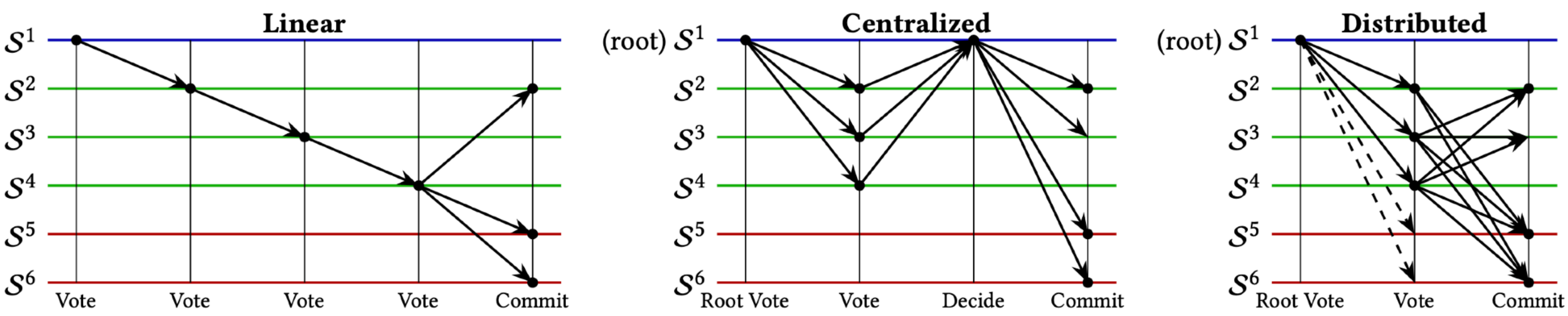
ByShard: Sharding in a Byzantine Environment [VLDB'21]

Processing multi-shard transaction via the orchestrate-execute model

Processing is broken down into three types of shard-steps: vote, commit, and abort

Each shard-step is performed via one consensus step

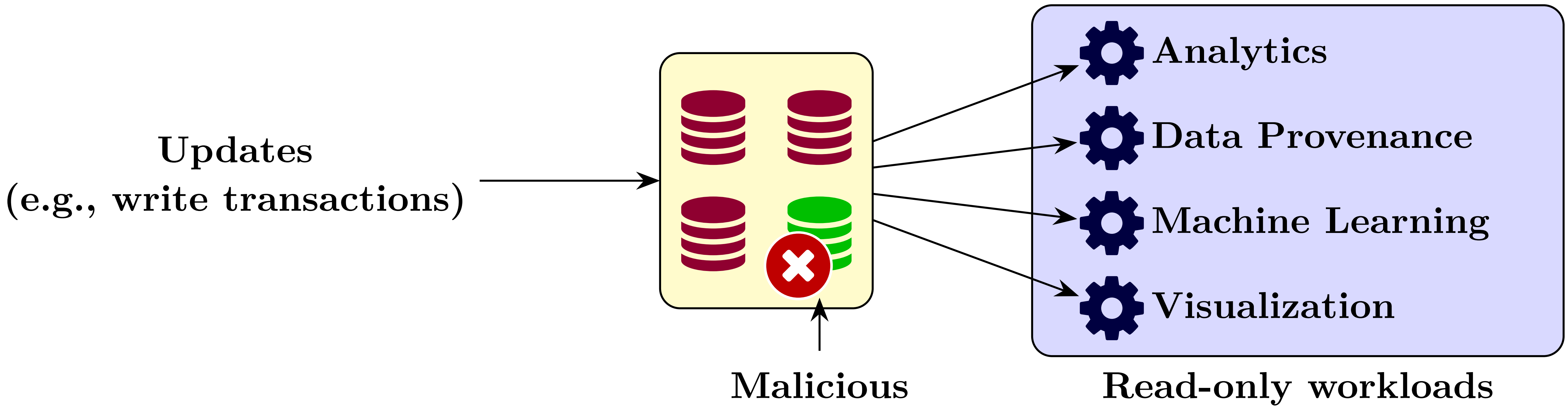
Steps are communicated via cluster-sending



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

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

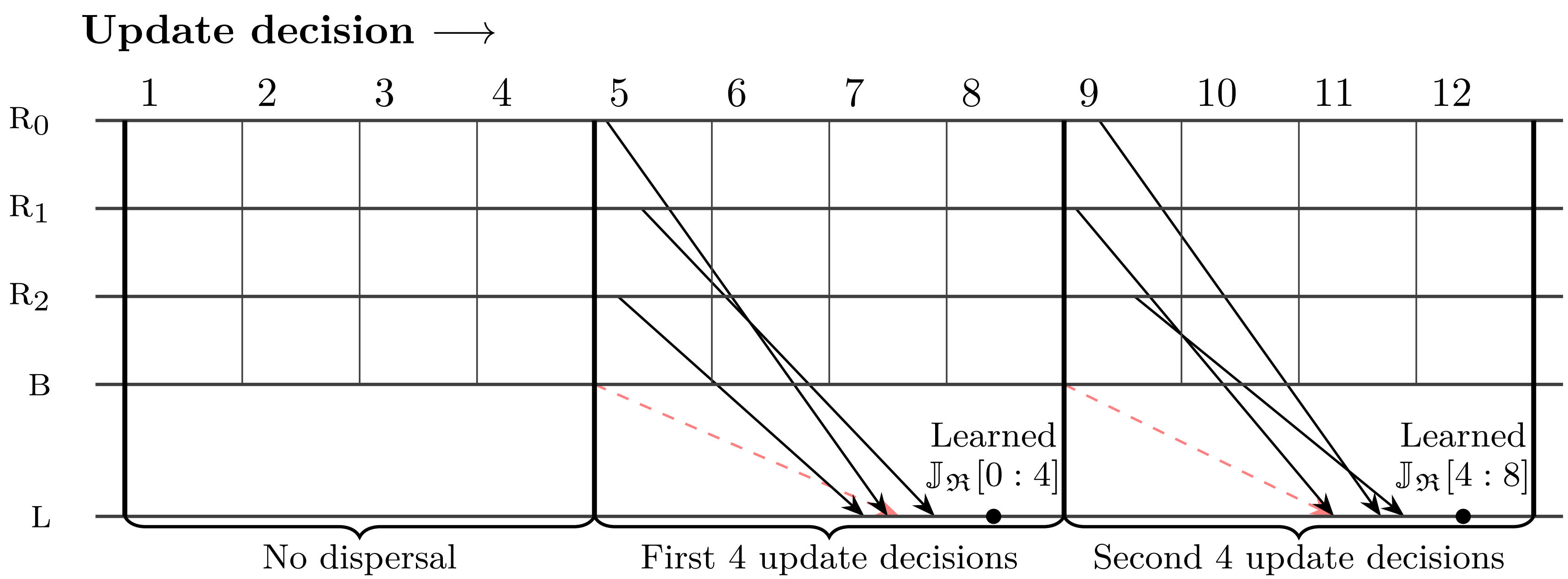
introducing the delayed-replication algorithm, utilizing information dispersal techniques, giving rise to a coordination-free, push-based, minimal communication protocol



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

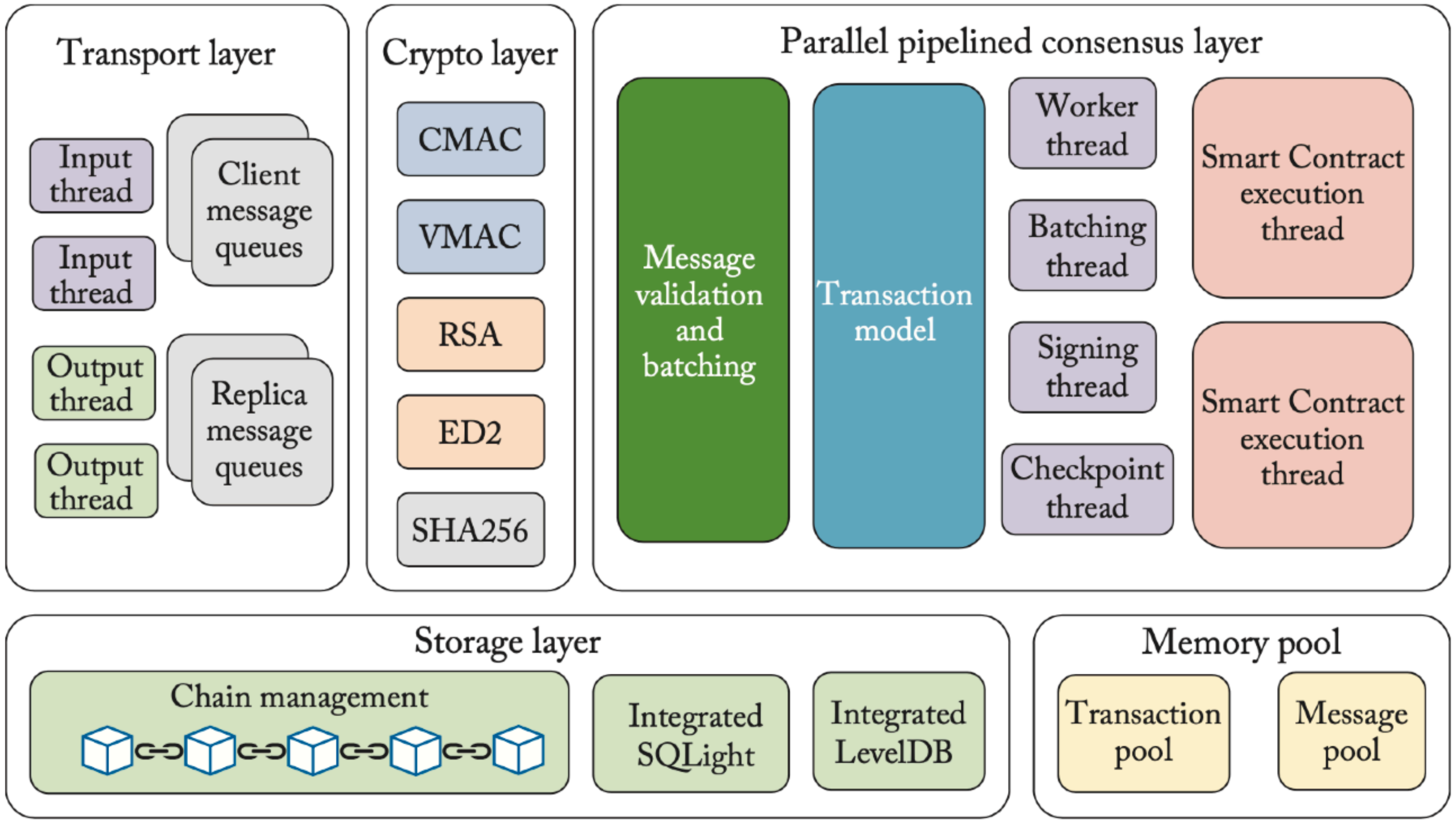
Out-of-ordering Consensus Communication

De-coupled Ordering and Execution

Off-Chain Memory Management

Expensive Cryptographic Practices (DS vs. MAC)

Smart Contracts Code Generation (Pre-compilation)



ResilientDB
Security, Privacy Reloaded

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

Single-threaded Monolithic Design

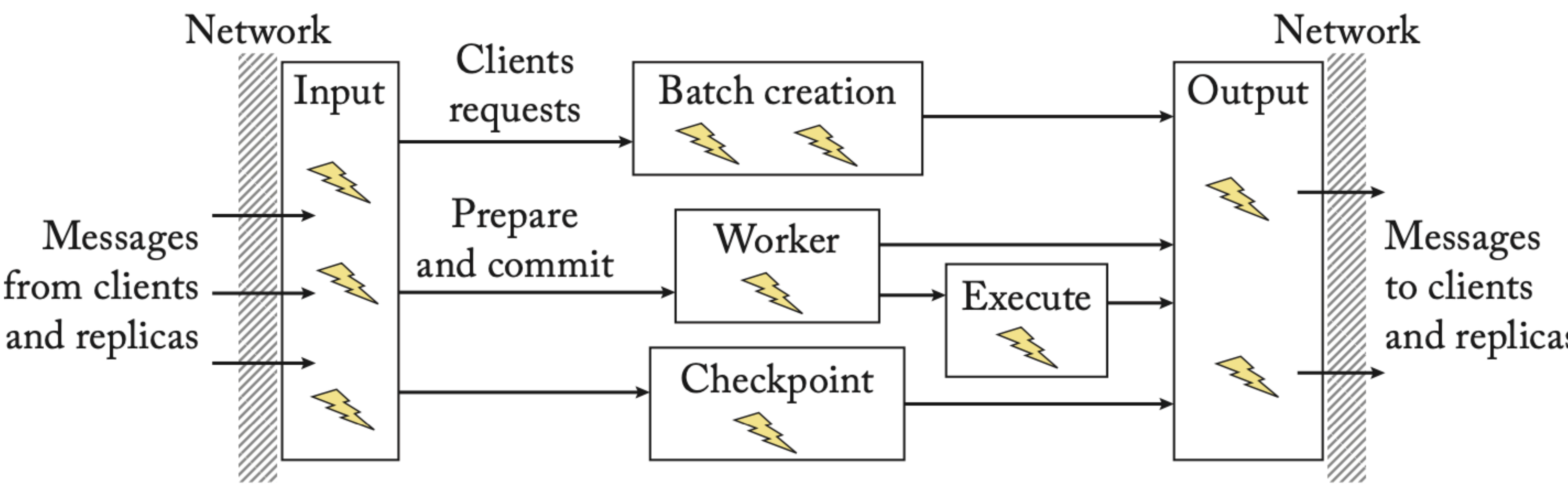
Out-of-ordering Consensus Communication

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

ResilientDB
Security, Privacy Reloaded

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

Single-threaded Monolithic Design

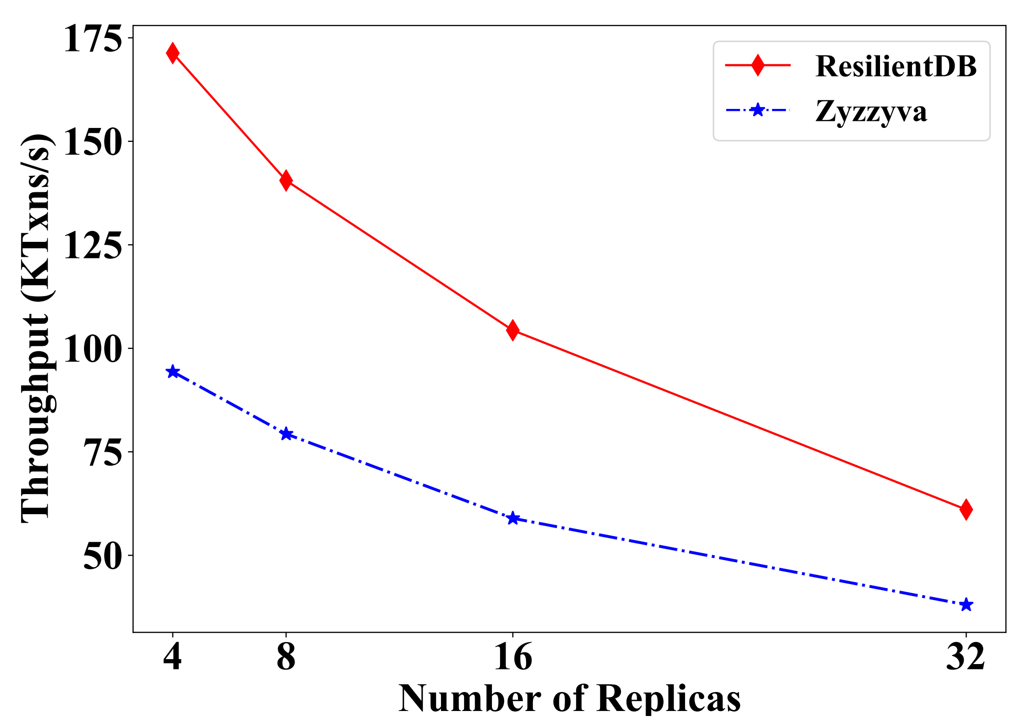
Out-of-ordering Consensus Communication

De-coupled Ordering and Execution

Off-Chain Memory Management

Expensive Cryptographic Practices (DS vs. MAC)

Smart Contracts Code Generation (Pre-compilation)



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

ResilientDB
Security, Privacy Reloaded

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"

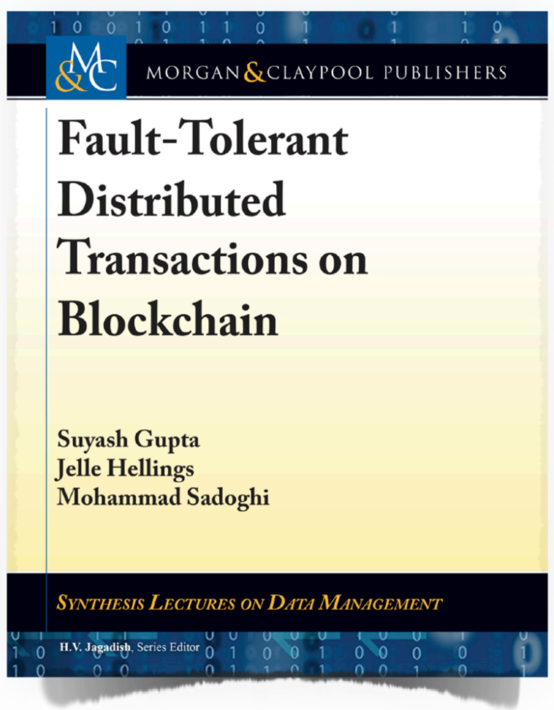
The California Aggie, April 6, 2020

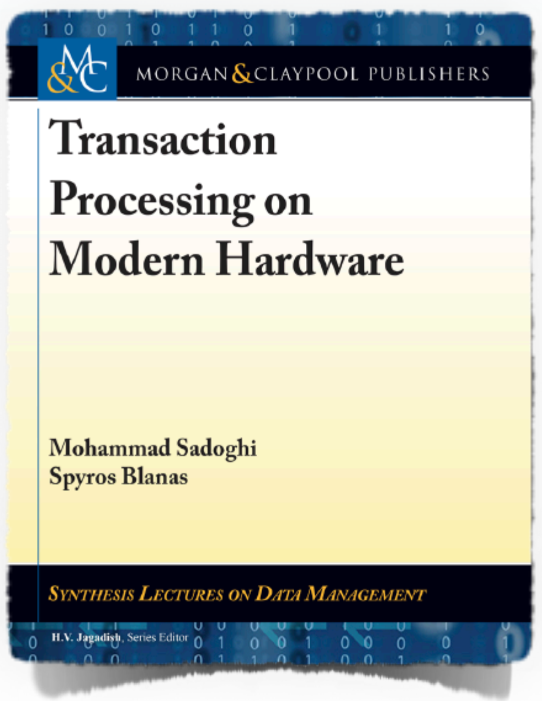
M.T.O. Tamarkoz®
BE BALANCED

ExpoLab
Creativity Unfolded

ResilientDB
Security, Privacy Reloaded

Thank You





For Complete References

dblp computer science bibliography

Google Scholar

RG