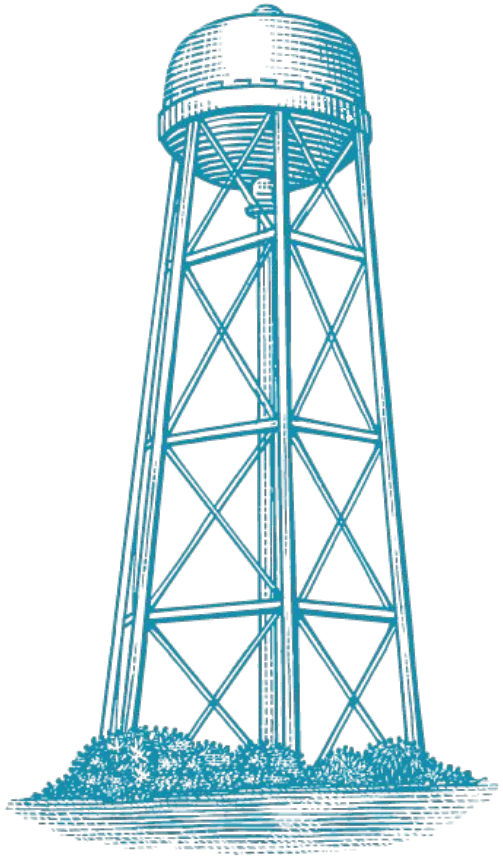
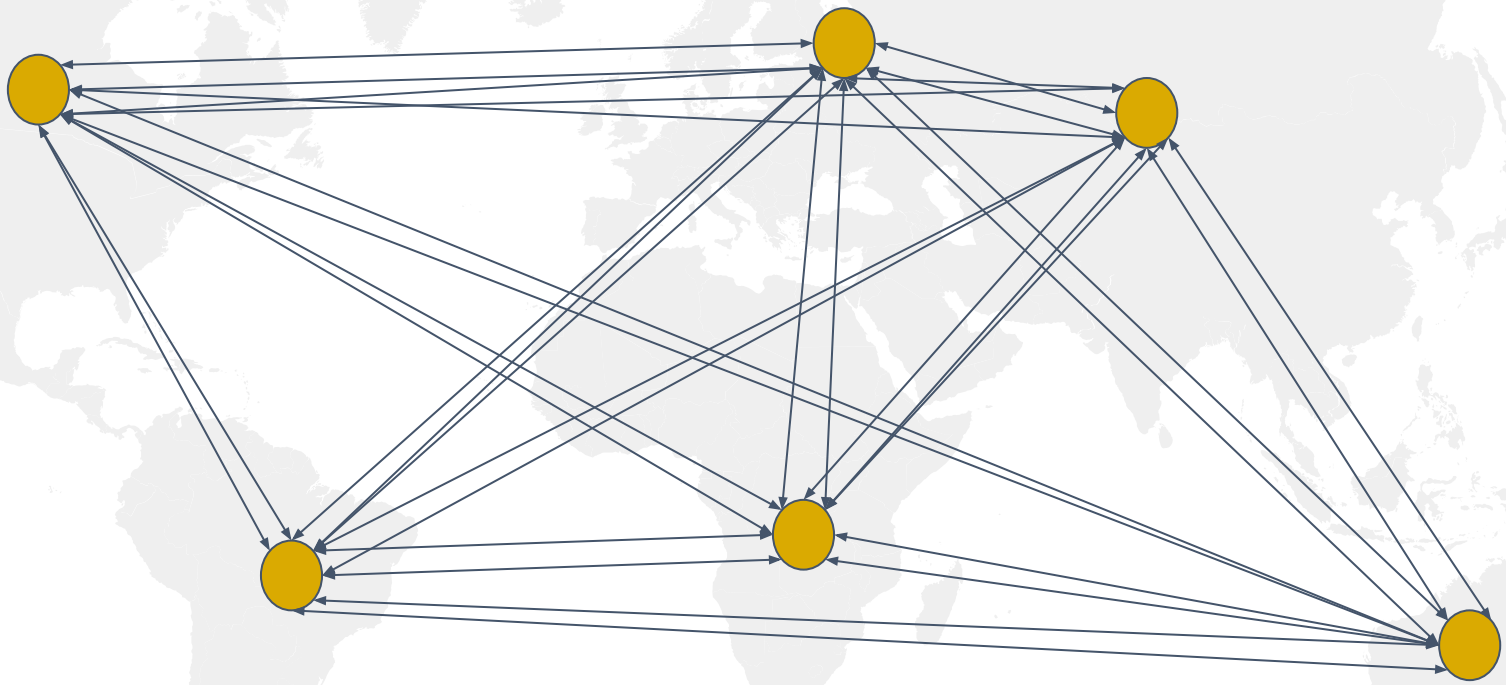
ResilientDB: Global Scale Resilient Blockchain Fabric.

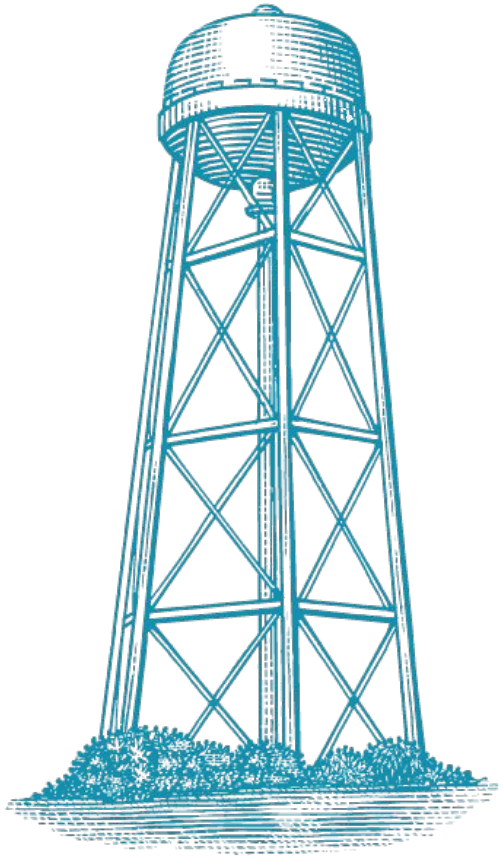
Rohan Sood, Utkarsh Drolia, Ashwitha Kassetty, Kunlin Tan, Alejandro Armas

UCDAVIS
COMPUTER SCIENCE





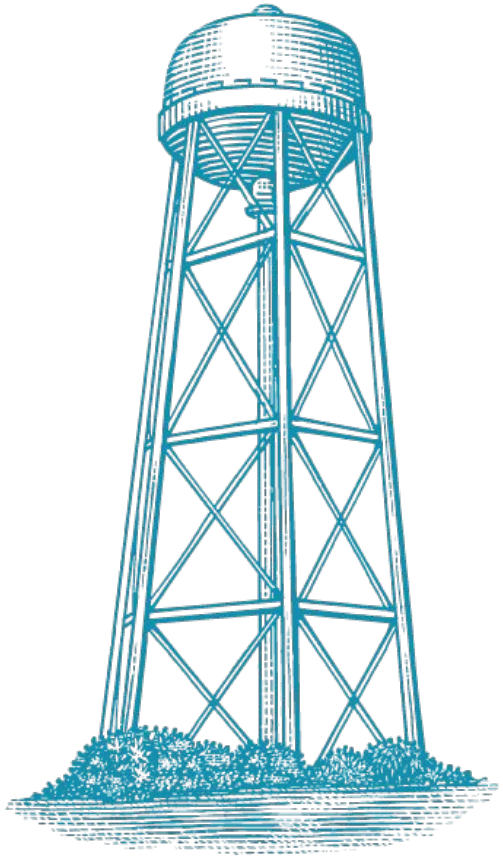
UCDAVIS
COMPUTER SCIENCE



What do you think is going wrong?

UCDAVIS
COMPUTER SCIENCE

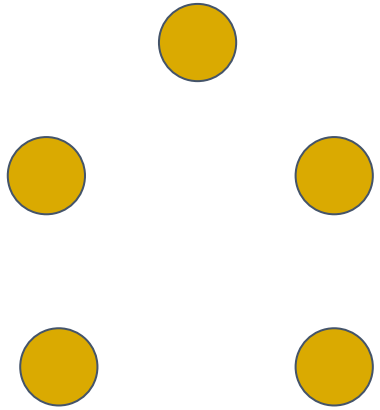
Latency speeds



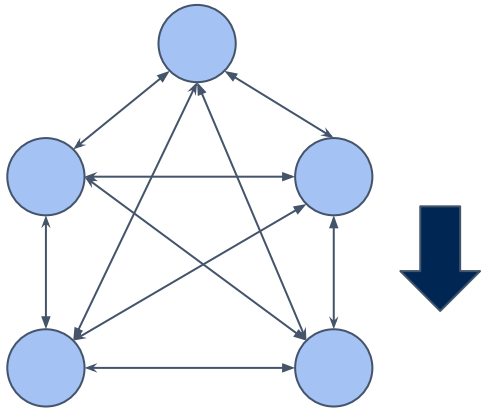
	Ping round-trip times (ms)						*Bandwidth* (Mbit/s)					
	O	I	M	B	T	S	O	I	M	B	T	S
Oregon (O)	≤ 1	38	65	136	118	161	7998	669	371	194	188	136
Iowa (I)		≤ 1	33	98	153	172		10004	752	243	144	120
Montreal (M)			≤ 1	82	186	202			7977	283	111	102
Belgium (B)				≤ 1	252	270				9728	79	66
Taiwan (T)					≤ 1	137					7998	160
Sydney (S)						≤ 1						7977

UCDAVIS
COMPUTER SCIENCE

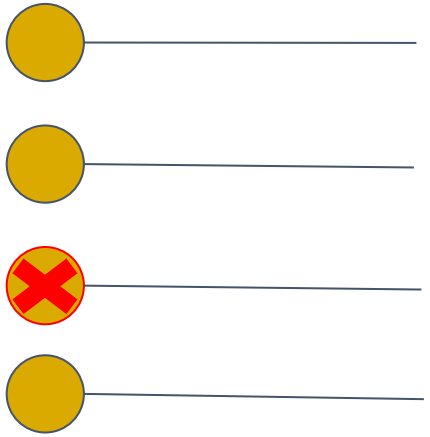
Creating a global scale protocol



Making a topologically aware grouping of replicas ie CLUSTERS



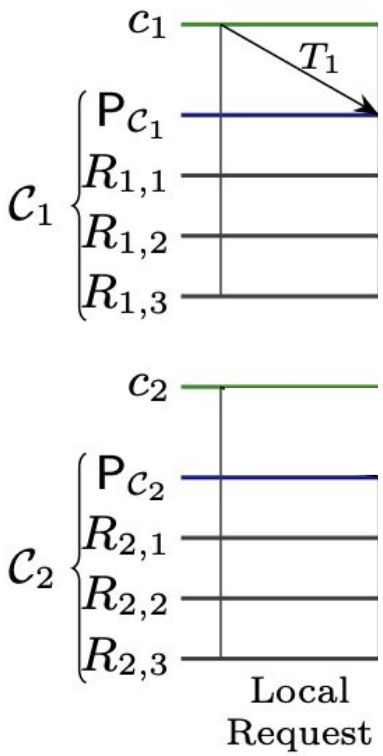
Optimistically reducing global communication between these clusters



Making the protocol live and safe by having novel remote view change that deals with maliciousness and failures

UCDAVIS
COMPUTER SCIENCE

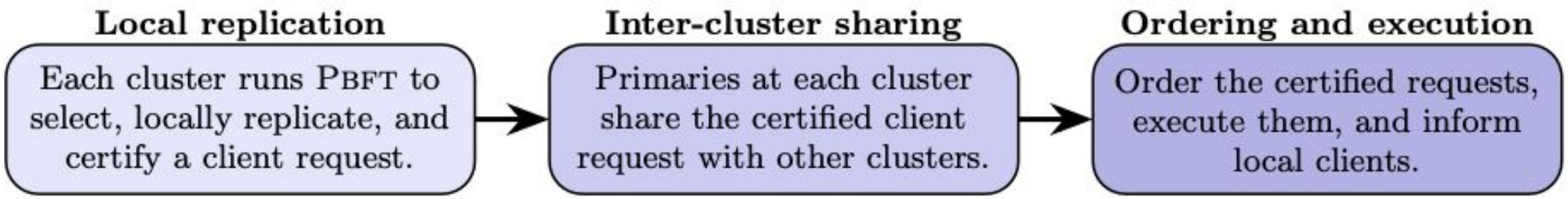
Introducing GeoBFT from a high level



Here we have 2 clusters having 4 nodes, GeoBFT operates in rounds, and in each round every cluster will be proposing a single client request for execution

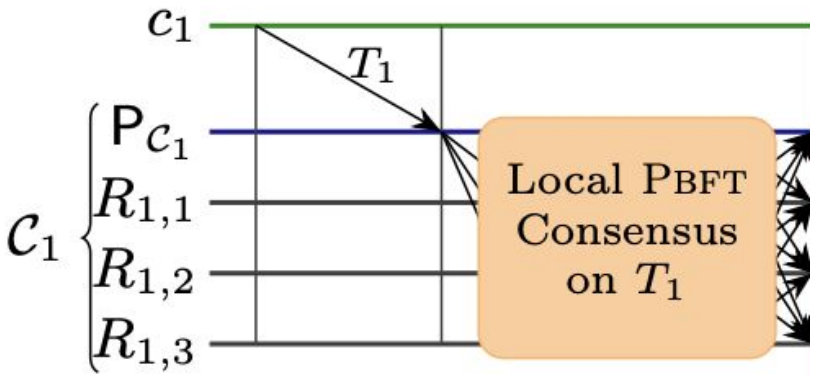
UCDAVIS
COMPUTER SCIENCE

Introducing GeoBFT from a high level



UCDAVIS
COMPUTER SCIENCE

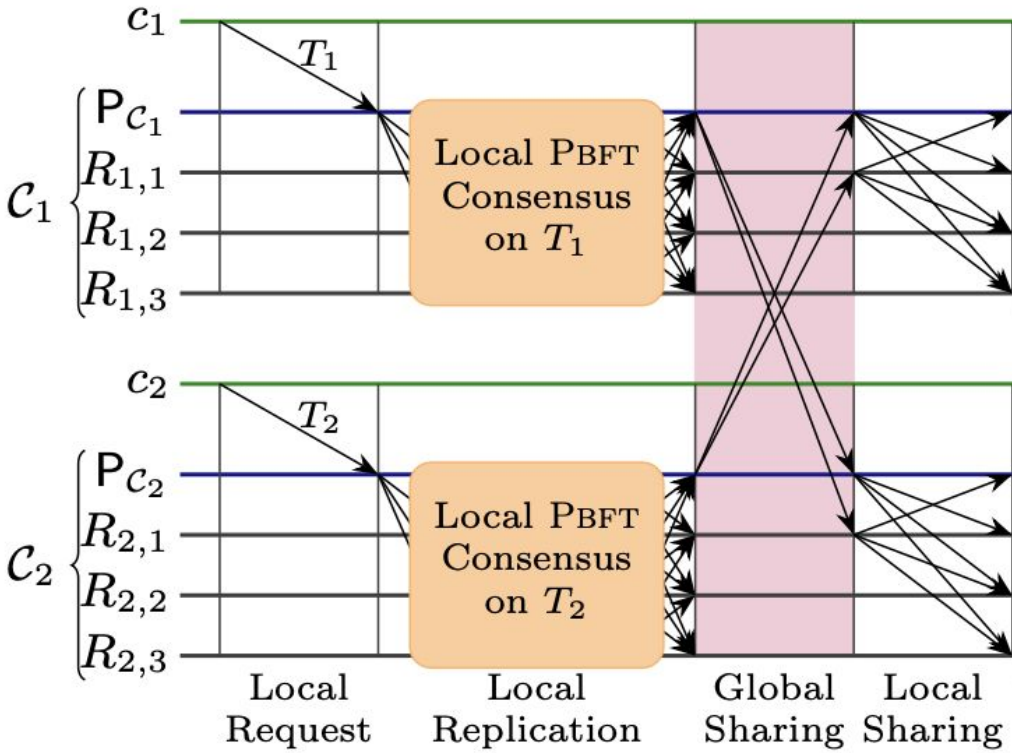
Introducing GeoBFT from a high level



The local client sends the transaction to the primary and performs PBFT in its local cluster

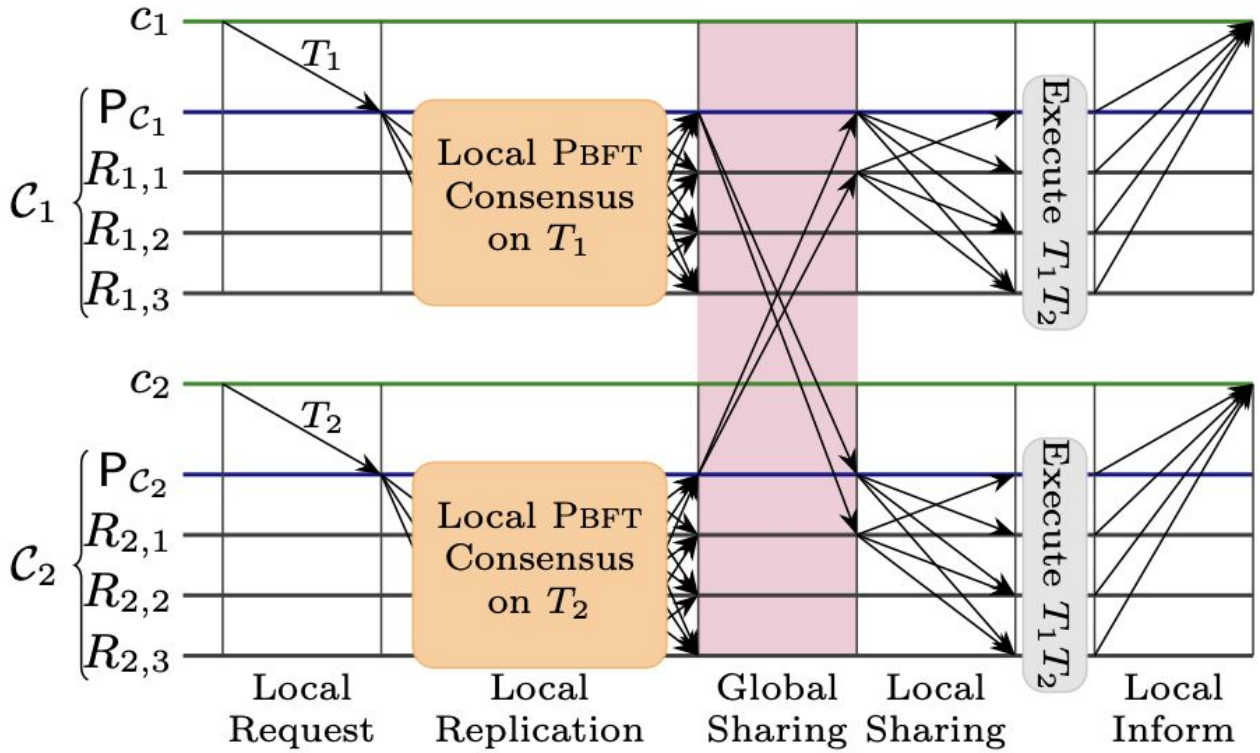
UCDAVIS COMPUTER SCIENCE

Introducing GeoBFT from a high level



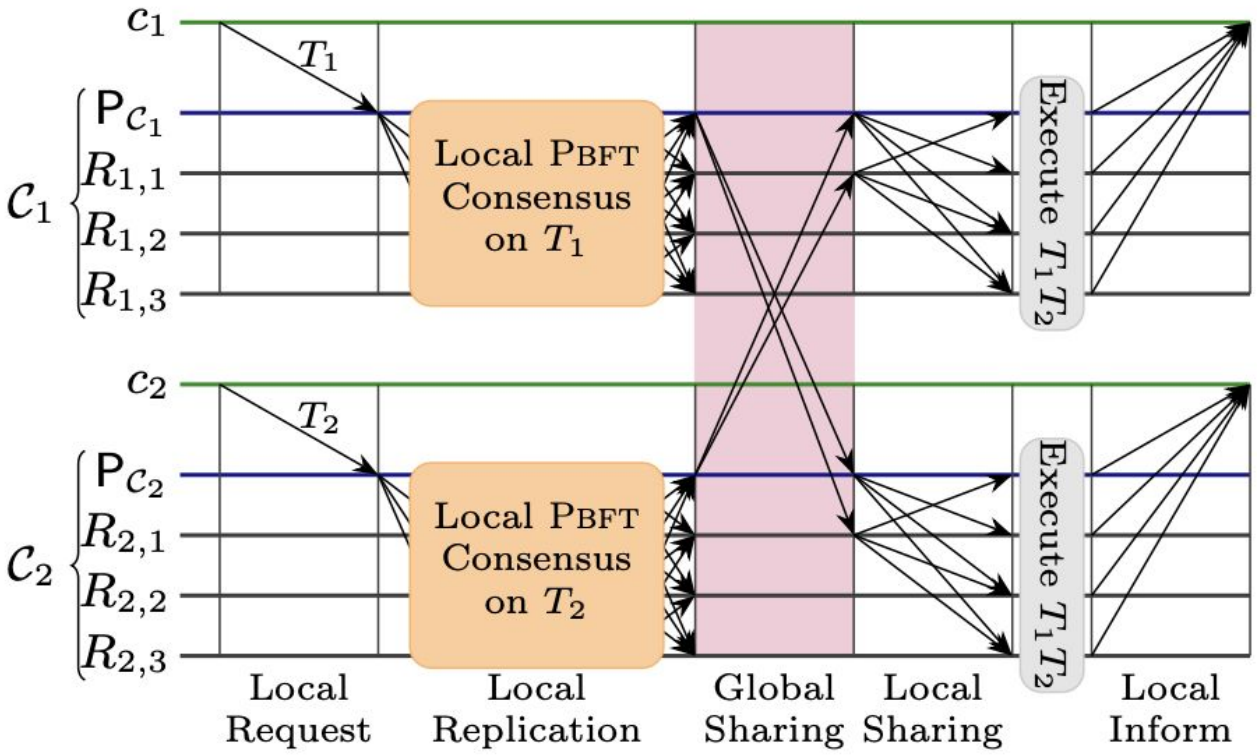
UCDAVIS COMPUTER SCIENCE

Introducing GeoBFT from a high level



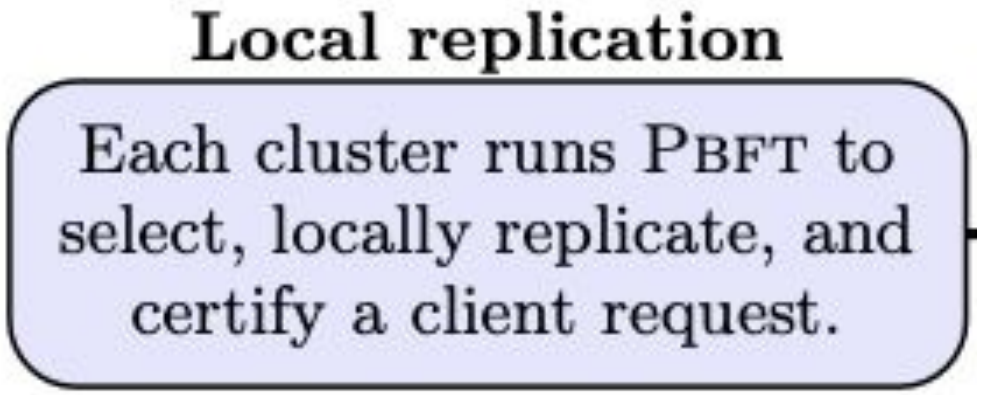
UCDAVIS COMPUTER SCIENCE

Introducing GeoBFT from a high level



UCDAVIS COMPUTER SCIENCE

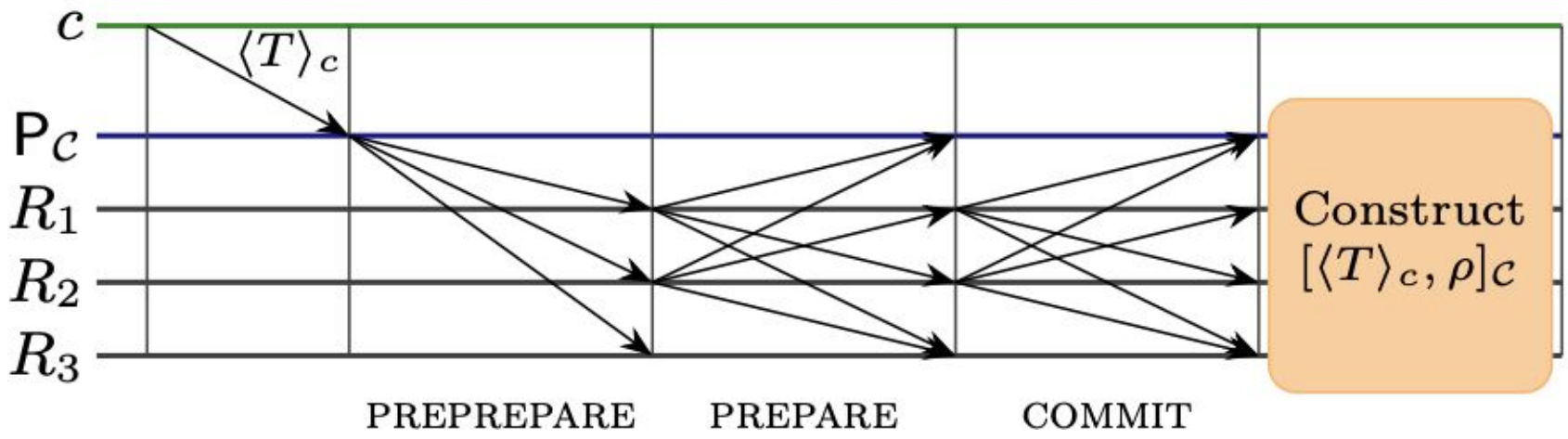
Local Replication (PBFT)



Local Replication (PBFT)

Local replication

Each cluster runs PBFT to select, locally replicate, and certify a client request.



Local Replication (PBFT) - commit certificate

UCDAVIS COMPUTER SCIENCE

Inter-Cluster Sharing

UCDAVIS
COMPUTER SCIENCE

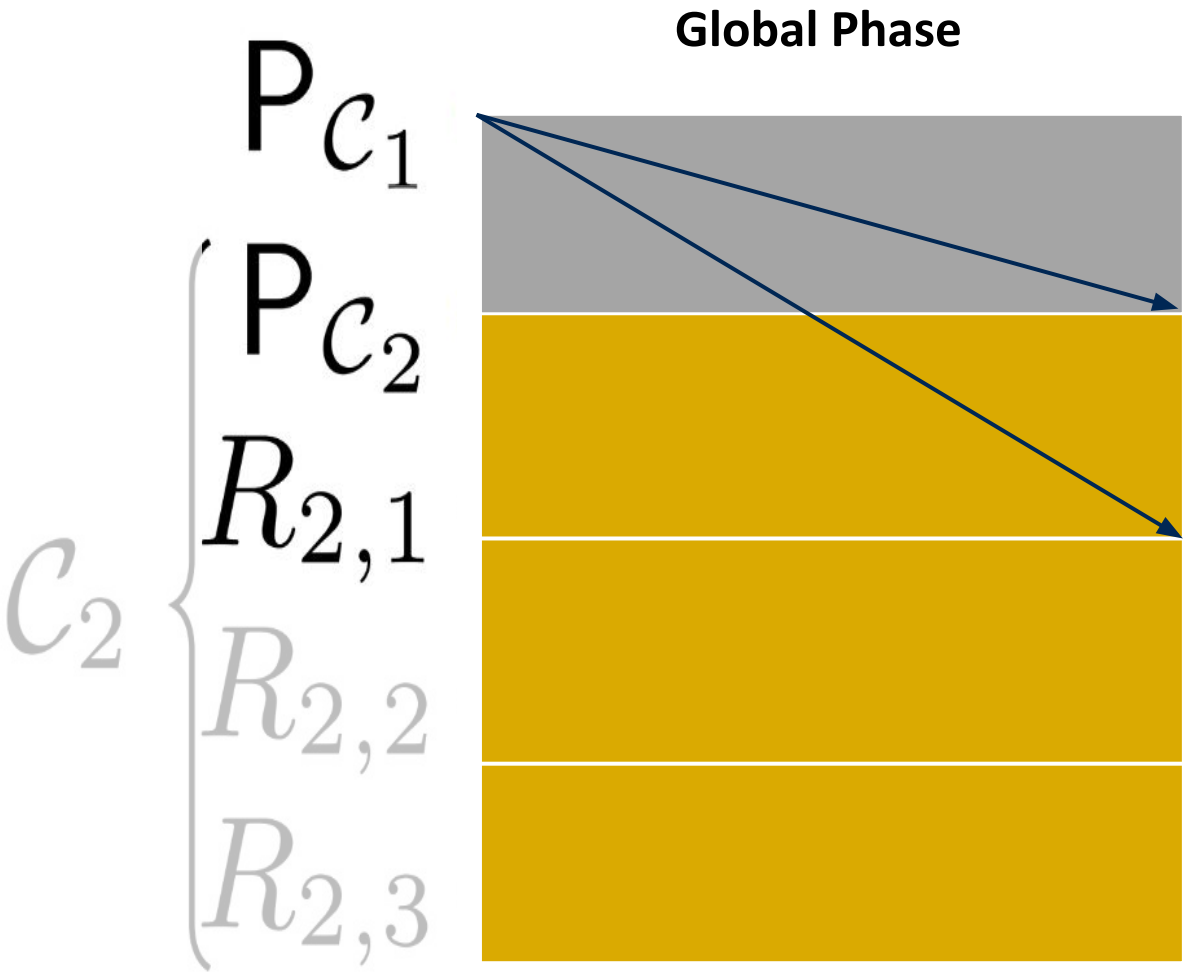
Message Forwarded (m)

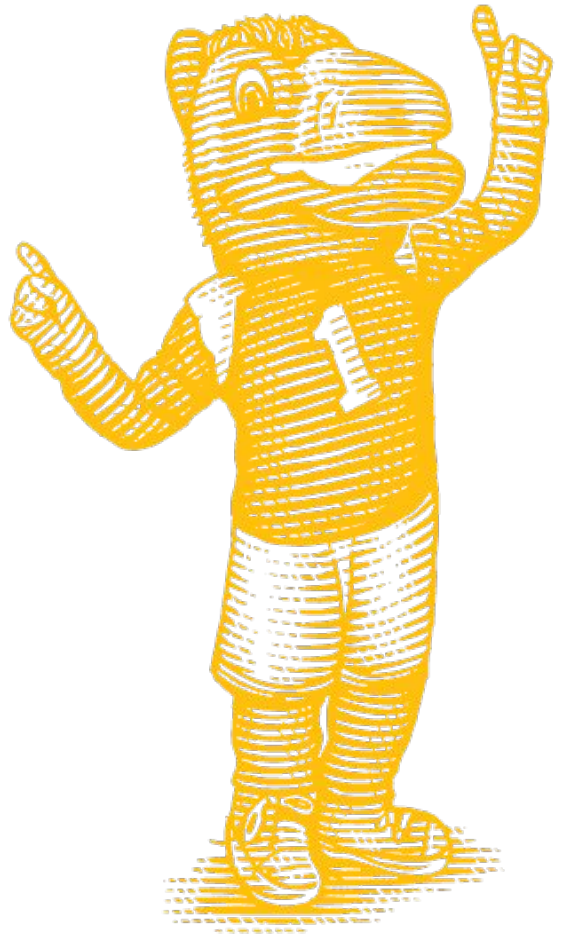
$$m = (\langle T \rangle_c, [\langle T \rangle_c, \rho] c_1)$$

$\langle T \rangle_c$ - Client Request of the cluster forwarding the message

$[\langle T \rangle_c, \rho] c_1$ - Commit Certificate of Cluster 1

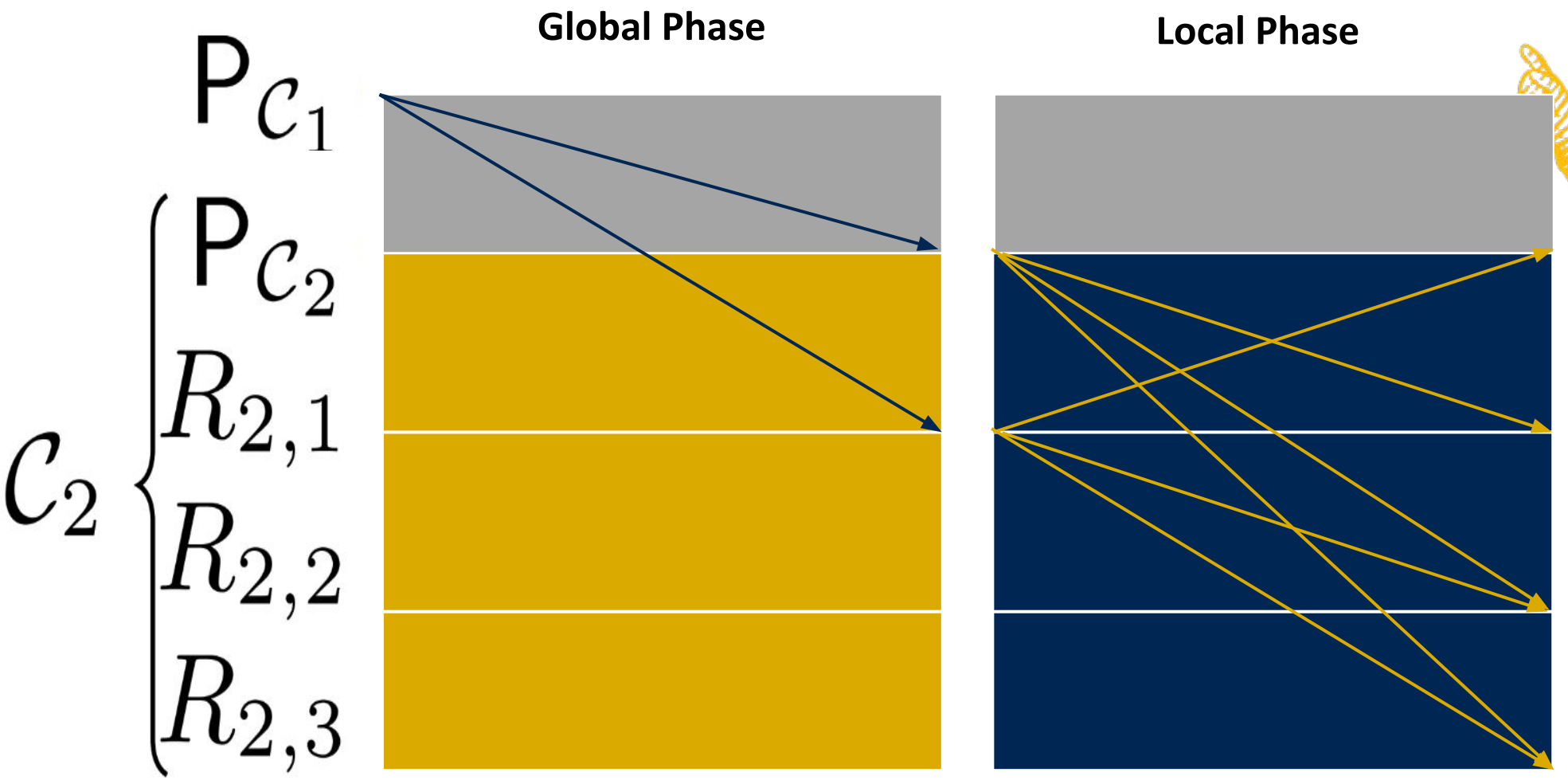
Optimistic Inter-cluster Sharing (ICS)

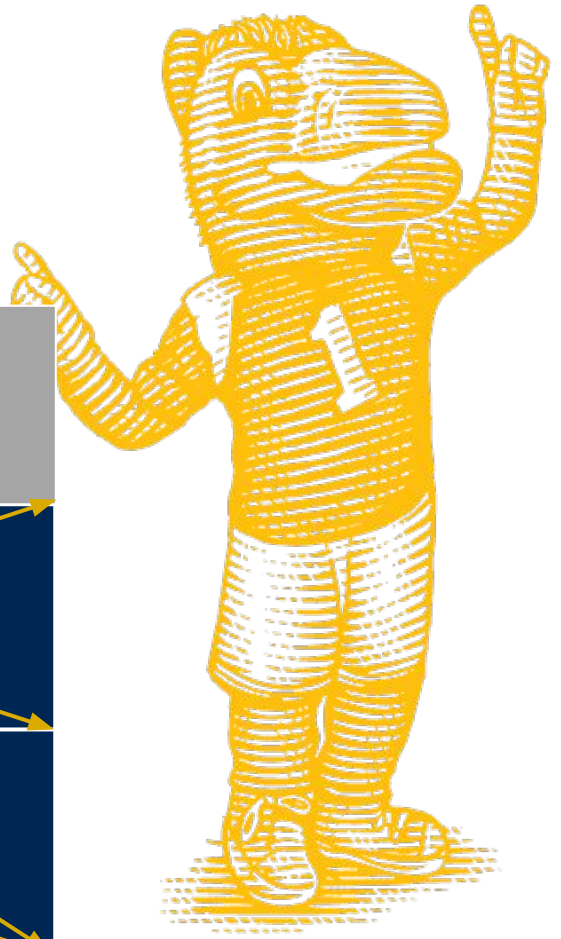




UCDAVIS
COMPUTER SCIENCE

Optimistic Inter-cluster Sharing (ICS)

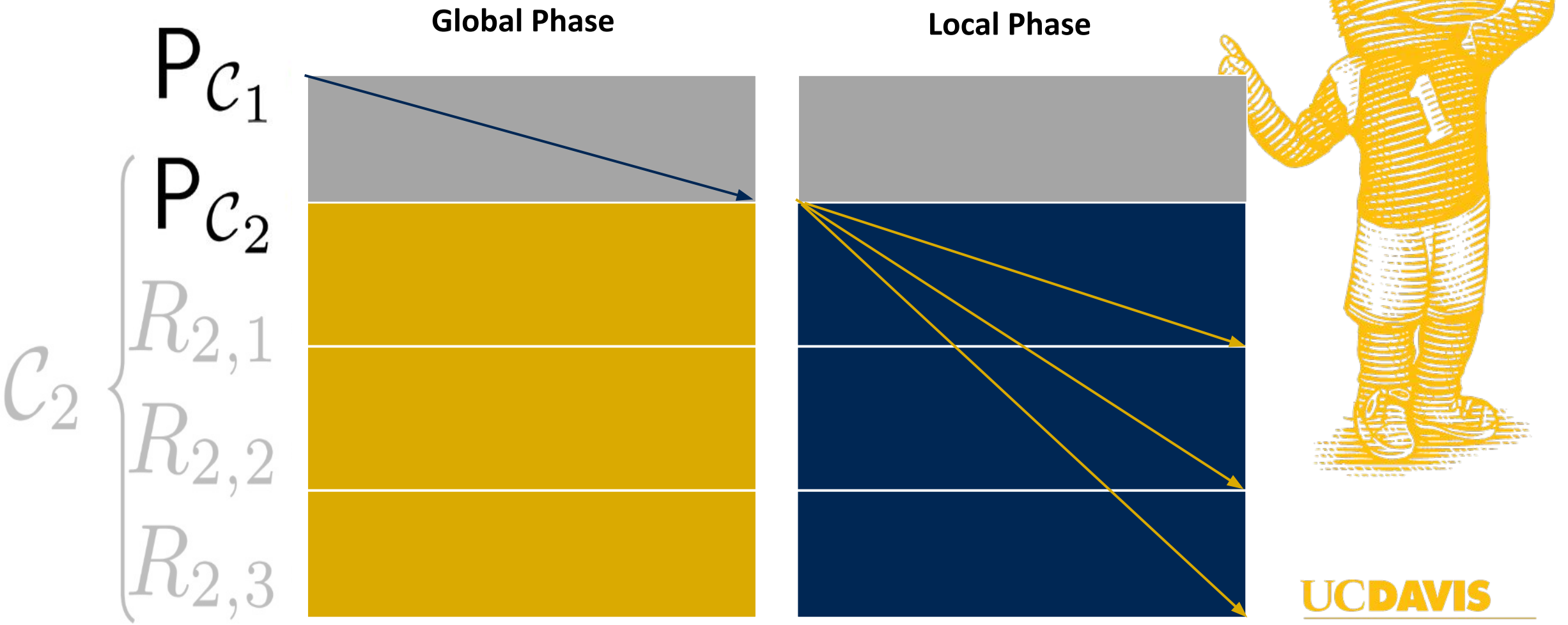




Forward m = ($\langle T \cdot \rangle_c$, $[\langle T \cdot \rangle_c, \varrho]_{C1}$)

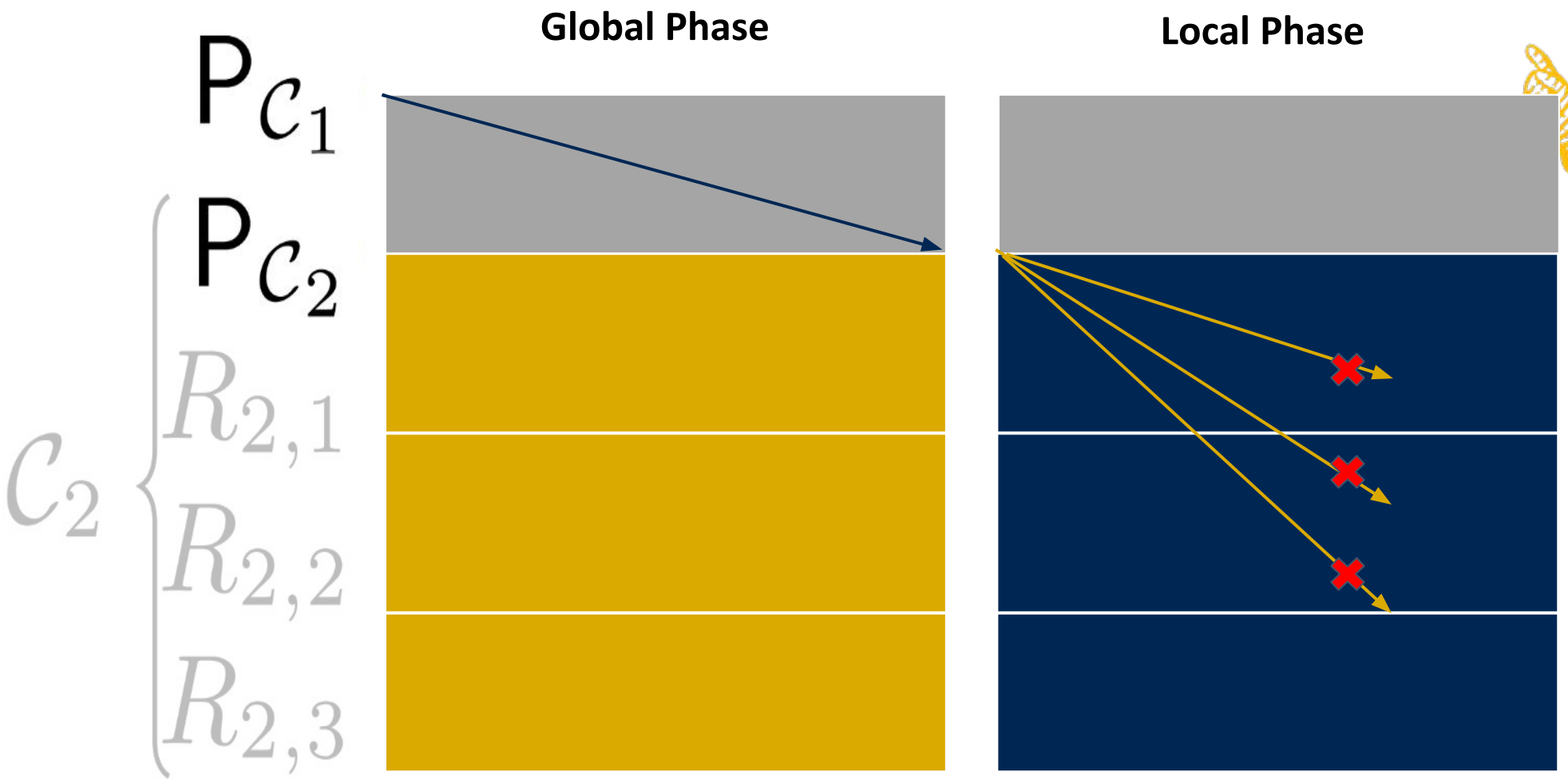
UCDAVIS
COMPUTER SCIENCE

Can we reduce communication?



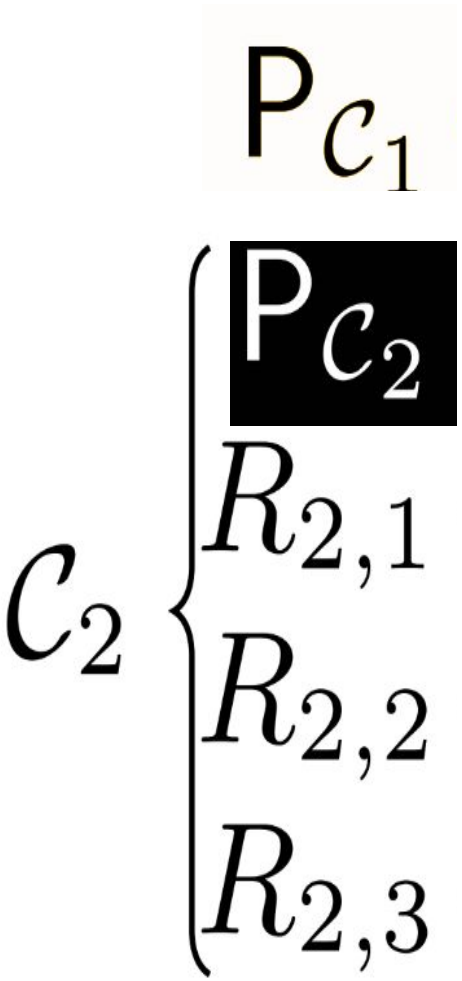
UCDAVIS
COMPUTER SCIENCE

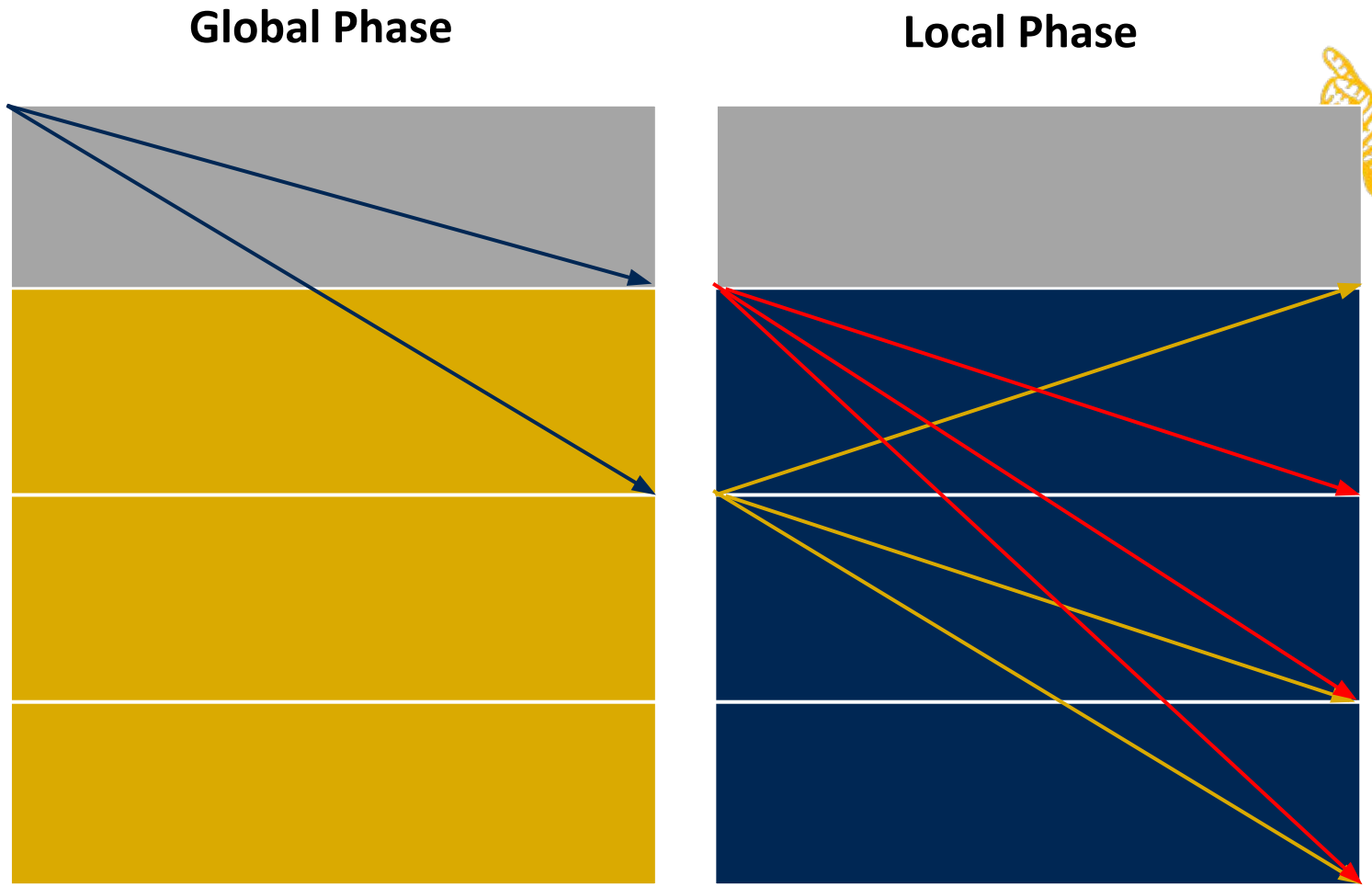
Inter-cluster Sharing (Scenario 1)

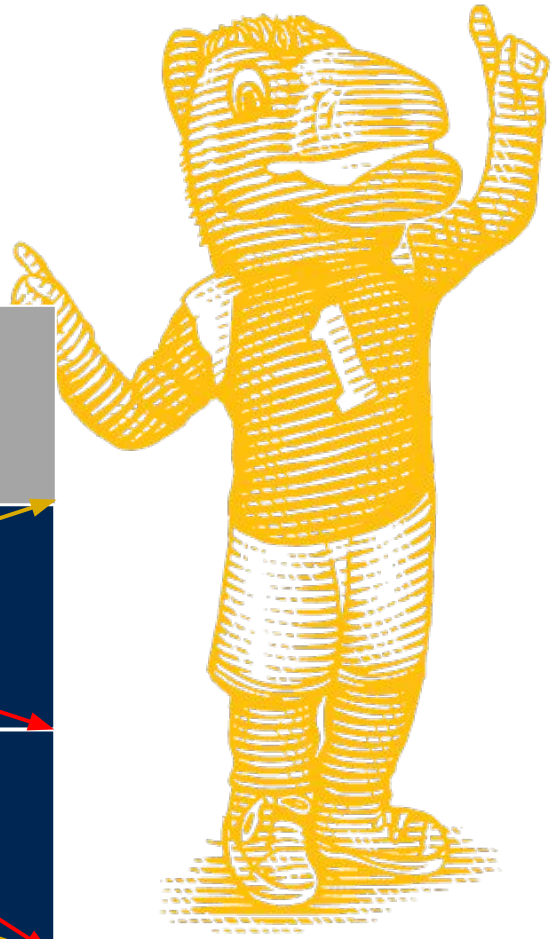


UCDAVIS
COMPUTER SCIENCE

How Optimistic ICS Resolves This





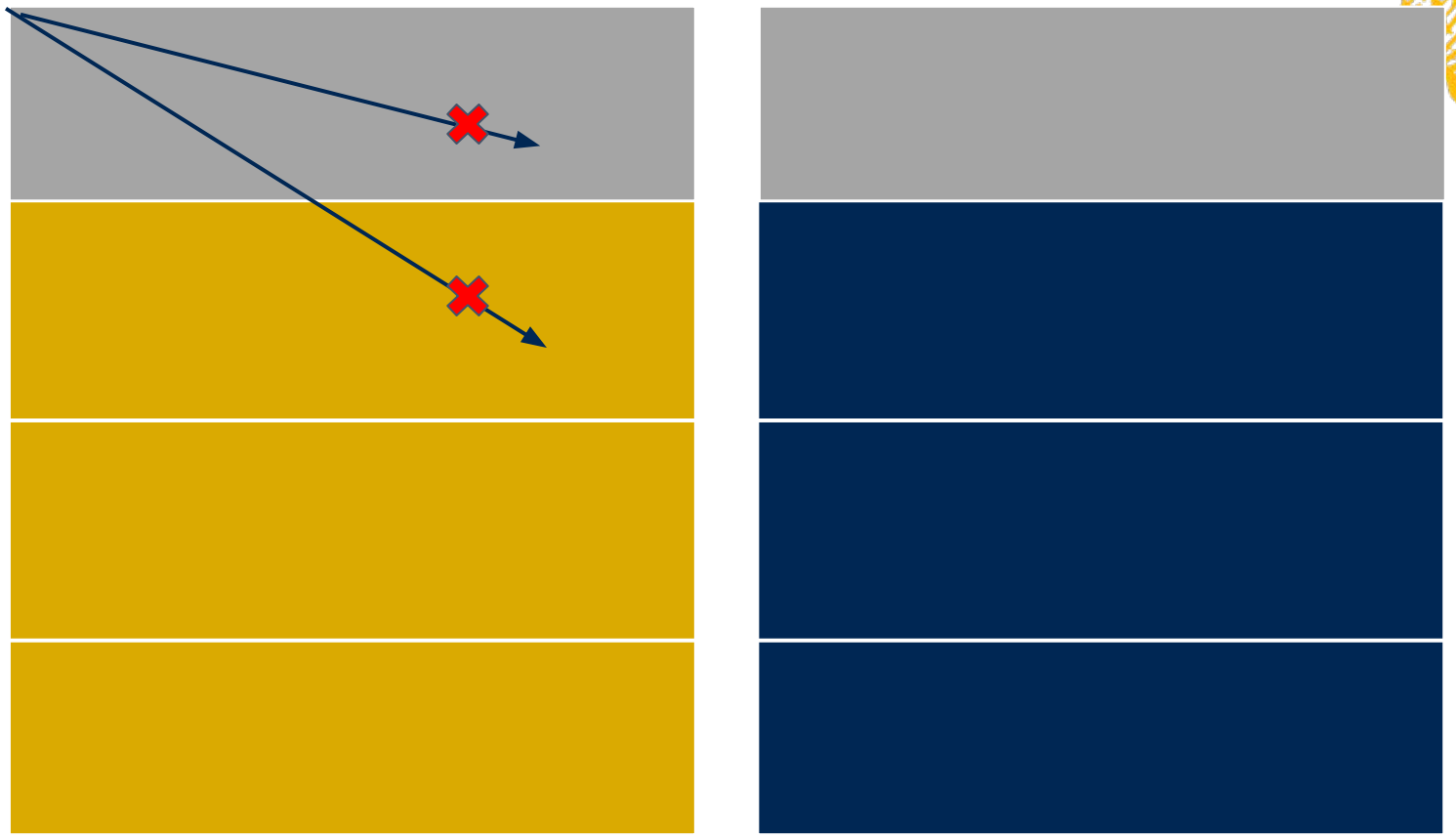


UCDAVIS
COMPUTER SCIENCE

ICS (Scenario: Byzantine P_{C1})

P_{C_1}

C_2 { P_{C_2}, $R_{2,1}$, $R_{2,2}$, $R_{2,3}$



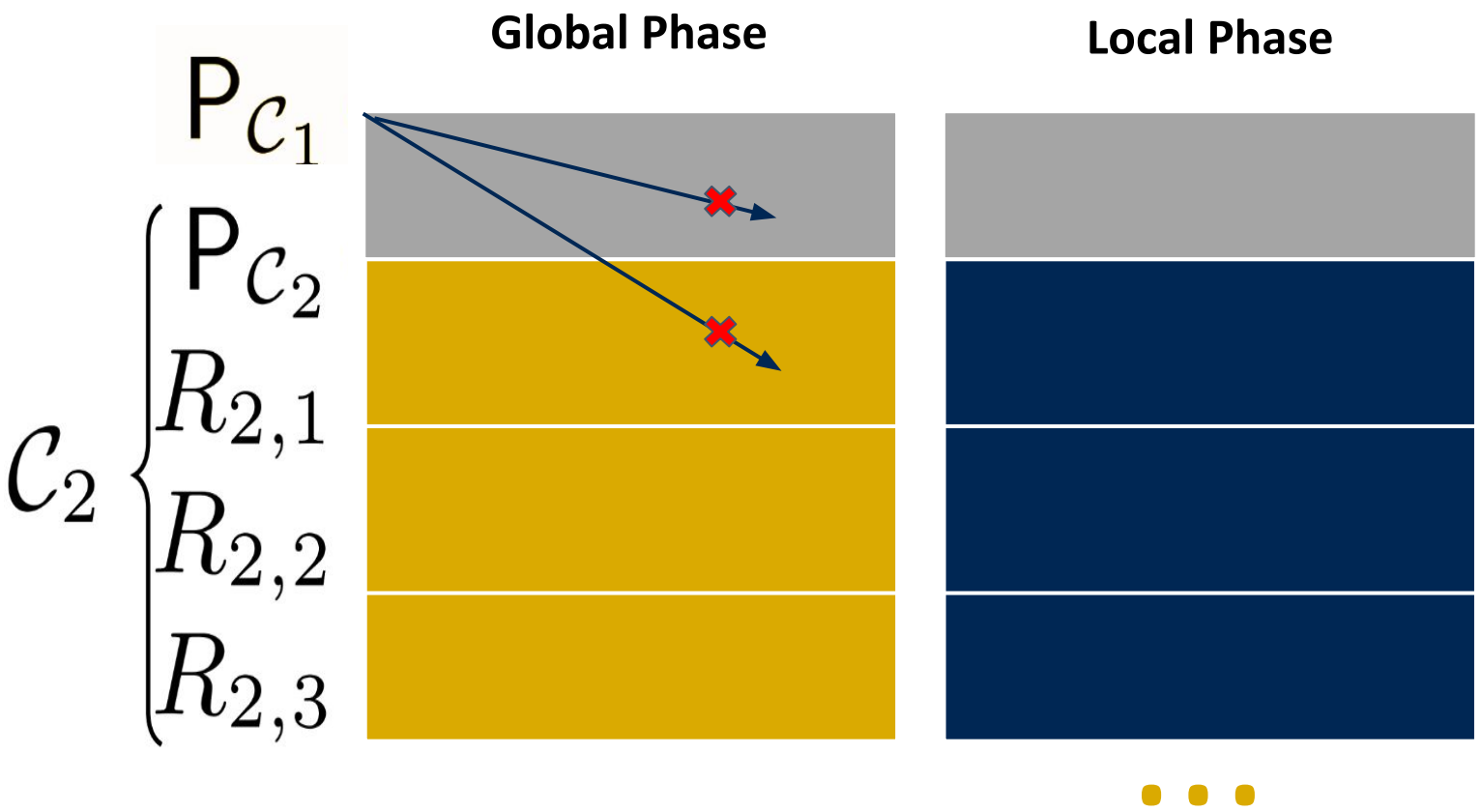
UCDAVIS
COMPUTER SCIENCE

Proposition 2.5. *Let $\mathfrak{S}$ be a system, let $\mathcal{C}_1, \mathcal{C}_2 \in \mathfrak{S}$ be two clusters, and let $m = (\langle T \rangle_c, [\langle T \rangle_c, \rho]_{\mathcal{C}_1})$ be the message $\mathcal{C}_1$ sends to $\mathcal{C}_2$ using the normal-case global sharing protocol of Figure 5. We have the following:*

Receipt *If the primary $\mathsf{P}_{\mathcal{C}_1}$ is non-faulty and communication is reliable, then every replica in $\mathcal{C}_2$ will eventually receive m.*

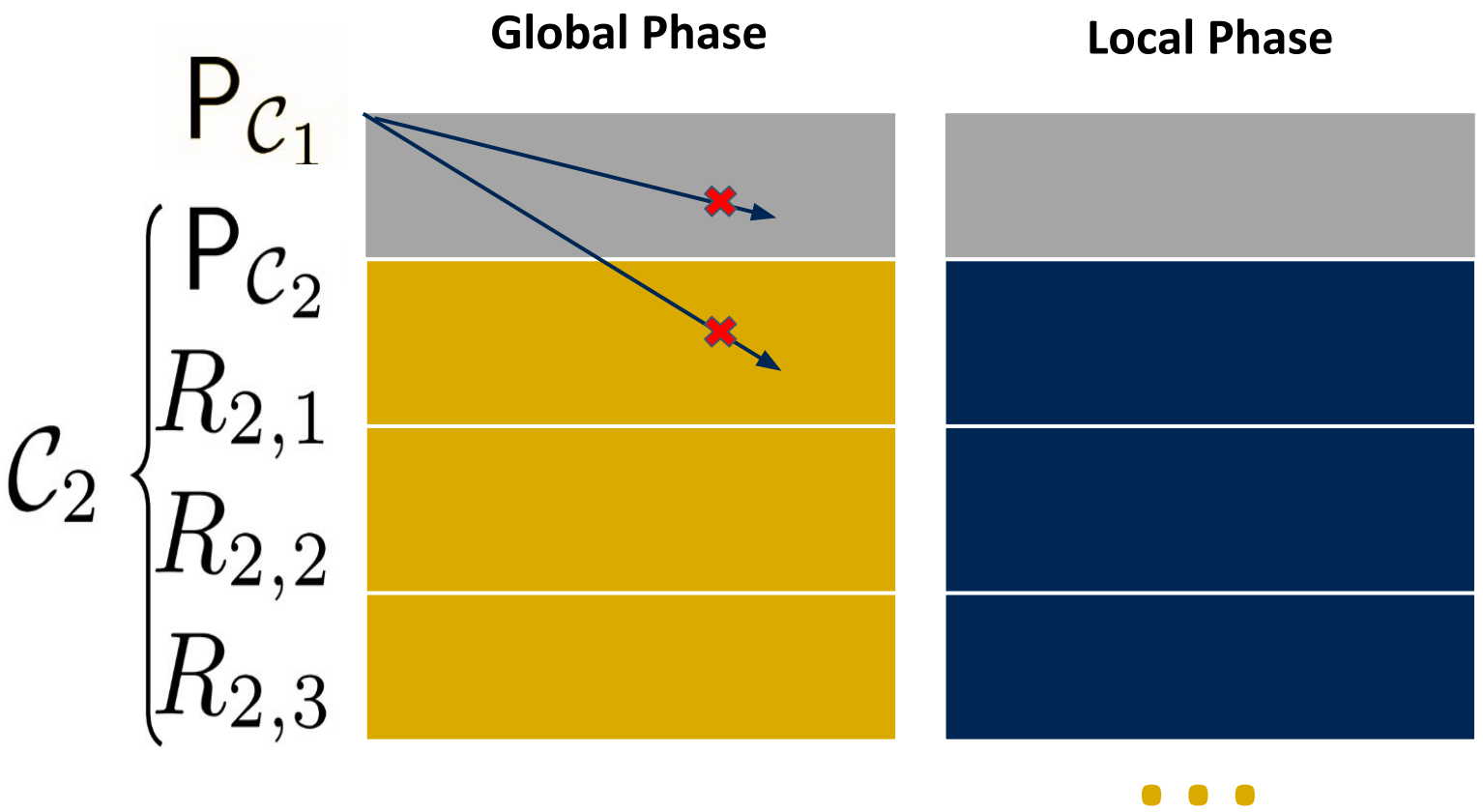
Agreement *Replicas in $\mathcal{C}_2$ will only accept client request $\langle T \rangle_c$ from $\mathcal{C}_1$ in round ρ.*

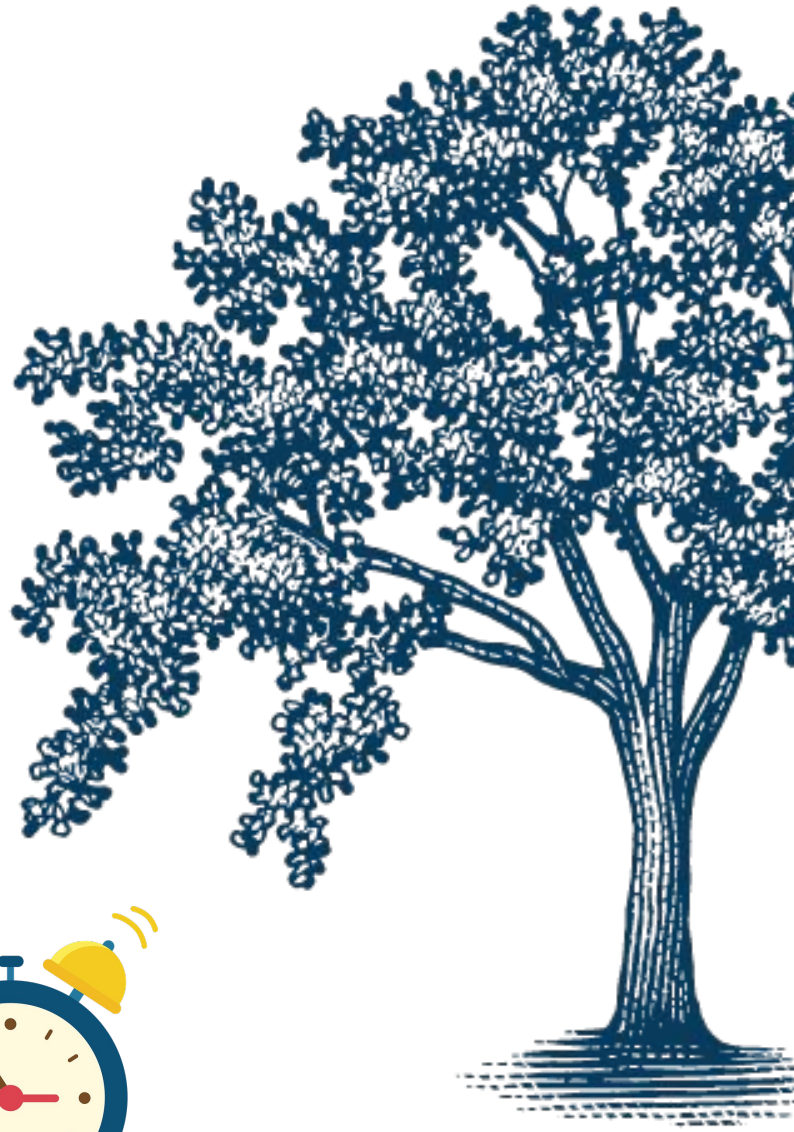
Remote View Change



UCDAVIS
COMPUTER SCIENCE

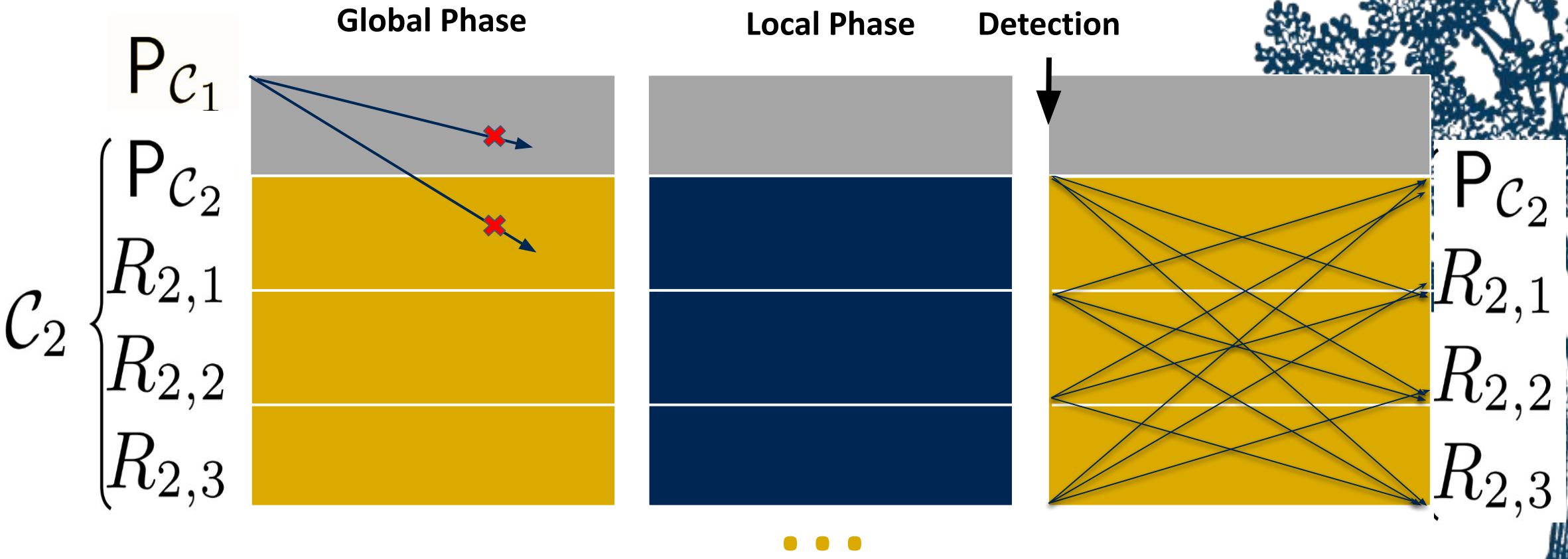
Remote View Change





UCDAVIS
COMPUTER SCIENCE

Remote View Change (Phase 1 of 4)

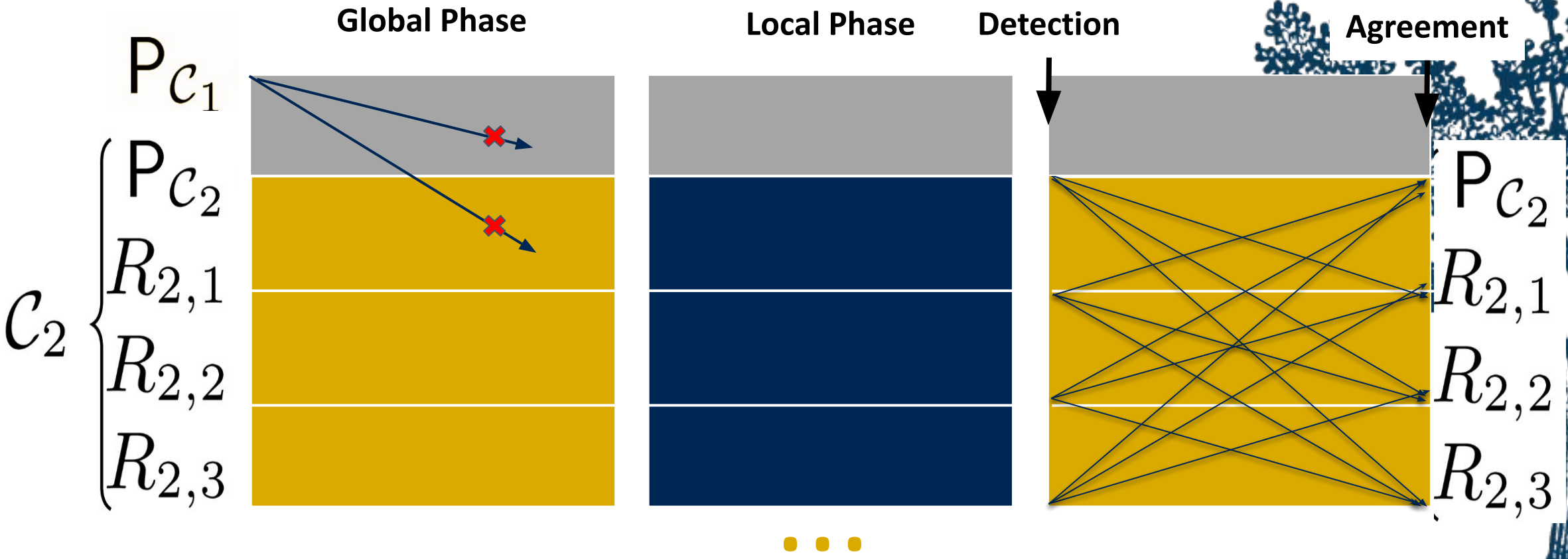


$R_{2,2}$ detects the failure of C_1

$R_{2,2}$ Broadcast $DR_{VC}(C_1, \varrho, v_1)$ to all replicas in C_2

UCDAVIS COMPUTER SCIENCE

Remote View Change (Phase 2 of 4)

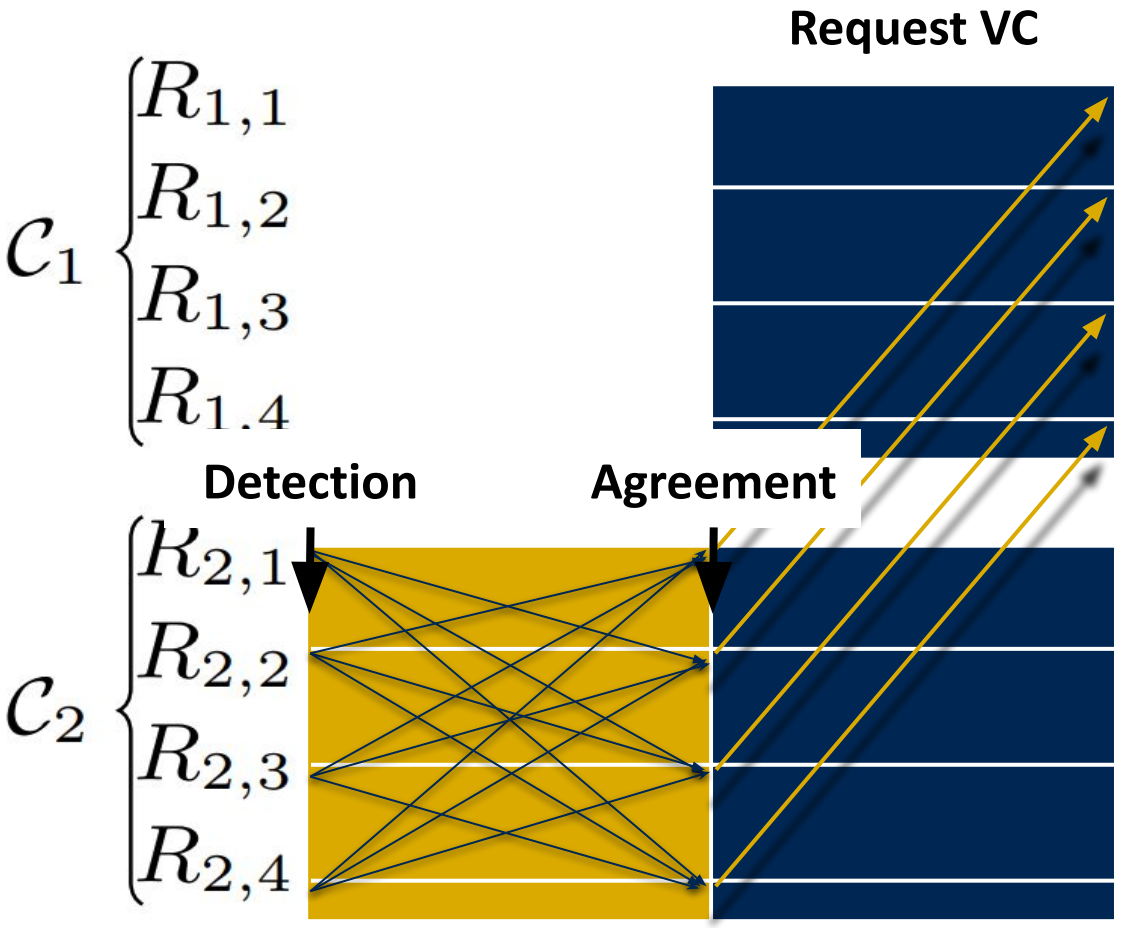


$R_{2,2}$ waits for n-f identical $DR_{VC}(C_1, \varrho, v_1)$

Non-fault replicates in C_2 initiate the agreement

UCDAVIS COMPUTER SCIENCE

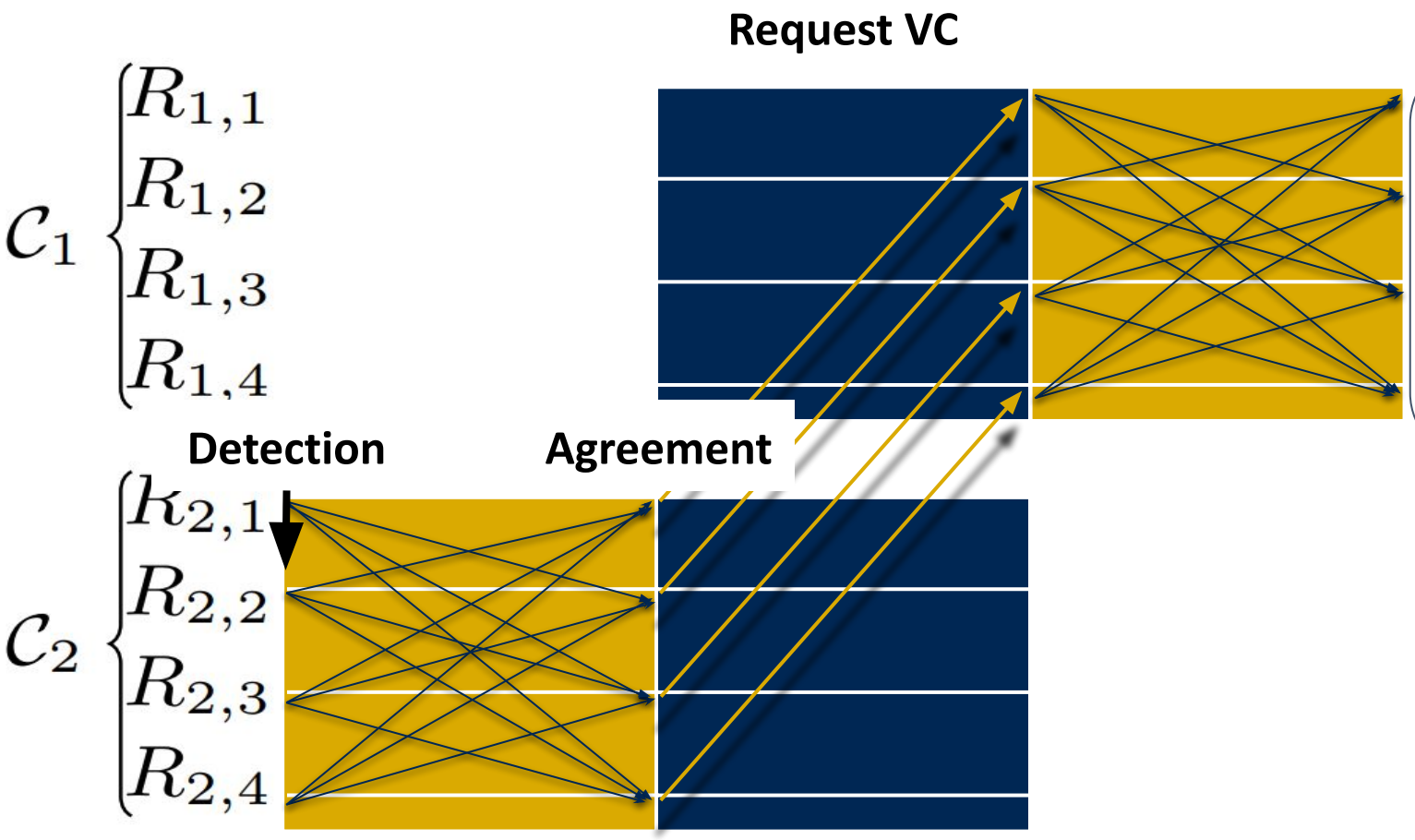
Remote View Change (Phase 3 of 4)



C_2 send request m_{RCV} for remote view-change in C_1

UCDAVIS
COMPUTER SCIENCE

Remote View Change (Phase 4 of 4)



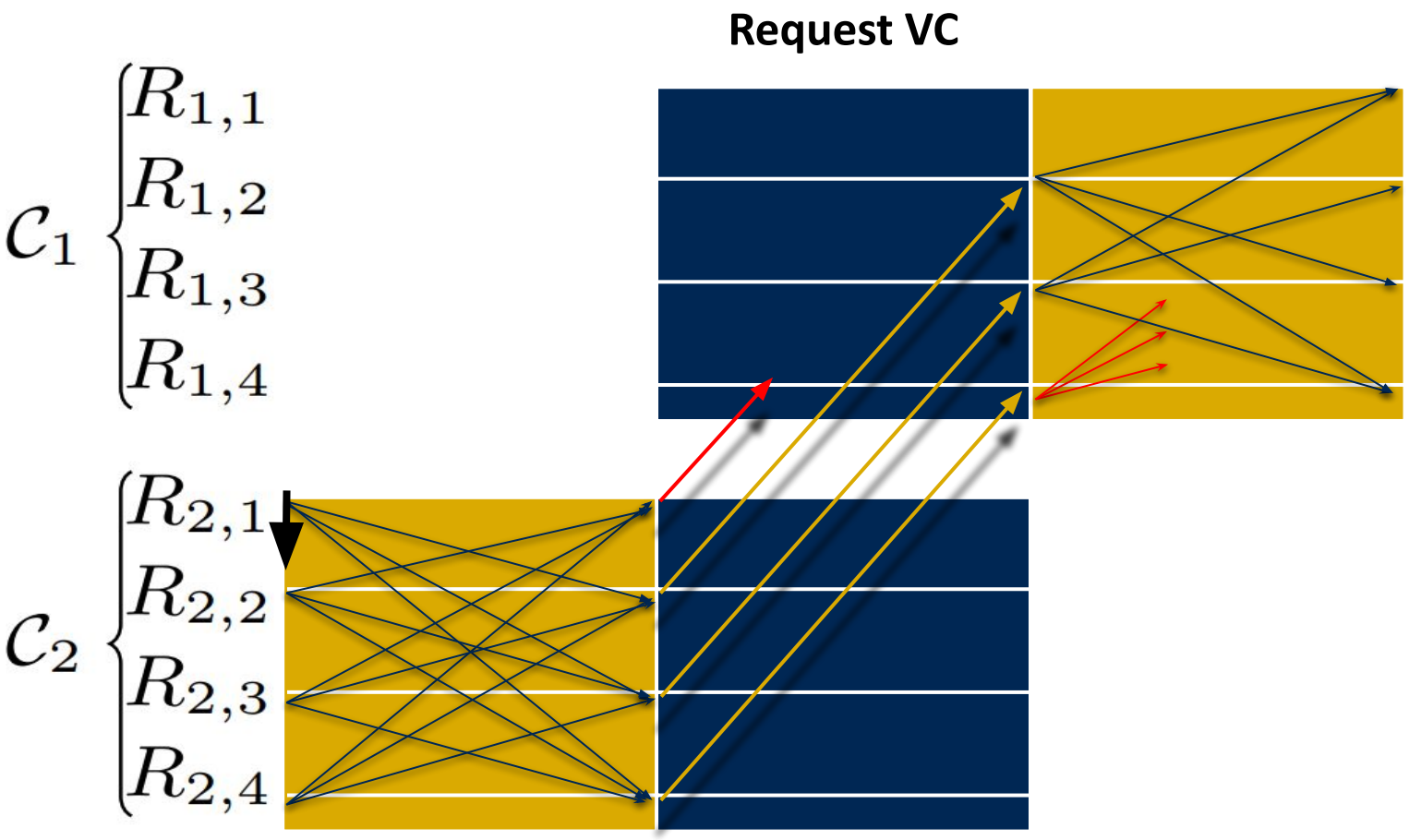
Induced view-change in C_1 (PBFT)

C_1 forward the request to all replicas in C_1

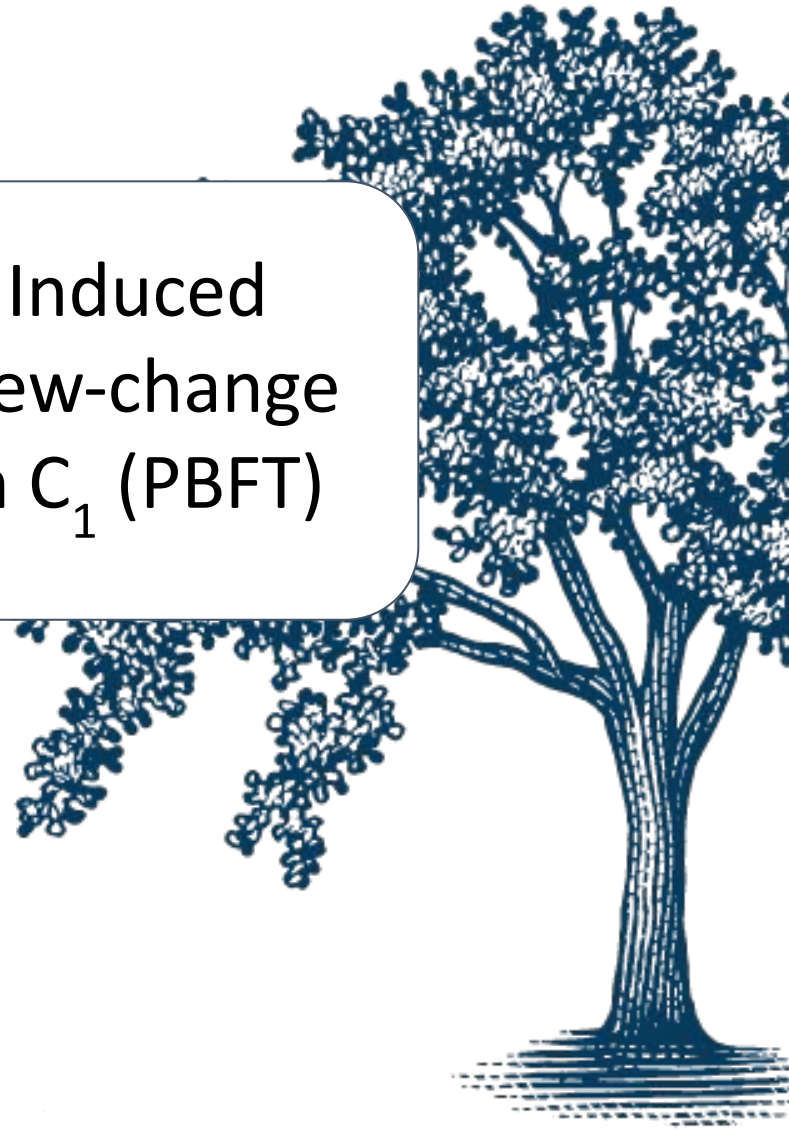
Waits for f+1 identical $R_{VC}(C, \varrho, v)$

UCDAVIS COMPUTER SCIENCE

Remote View Change (Scenario 1)

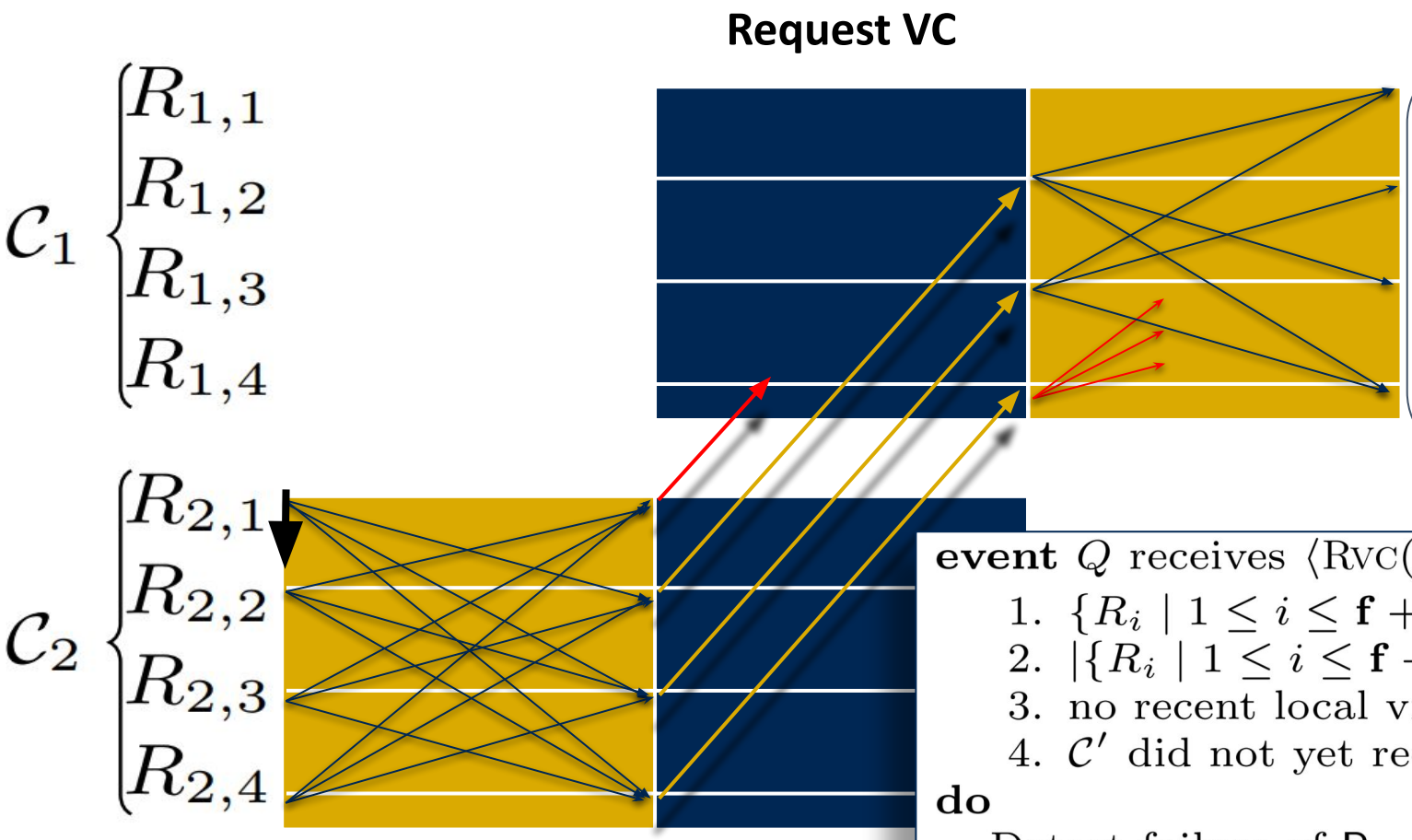


Induced view-change in C_1 (PBFT)



UCDAVIS
COMPUTER SCIENCE

Remote View Change (Scenario 1)



Induced view-change in C_1 (PBFT)

event Q receives $\langle \text{RVC}(\mathcal{C}_1, \rho, v) \rangle_{R_i}$, $1 \leq i \leq \mathbf{f}+1$, such that:

1. $\{R_i \mid 1 \leq i \leq \mathbf{f}+1\} \subset \mathcal{C}'$, $\mathcal{C}' \in \mathfrak{S}$;
2. $|\{R_i \mid 1 \leq i \leq \mathbf{f}+1\}| = \mathbf{f}+1$;
3. no recent local view-change was triggered; and
4. $\mathcal{C}'$ did not yet request a v-th remote view-change

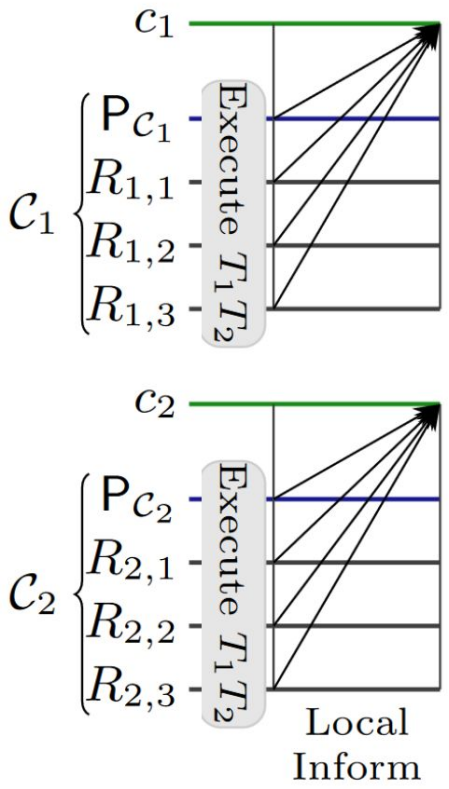
do

Detect failure of $\text{P}_{\mathcal{C}_1}$ (if not yet done so).

UCDAVIS COMPUTER SCIENCE

Ordering and Execution

Theorem 2.7. *Let* $\mathfrak{S} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{\mathbf{z}}\}$ *be a system over* $\mathfrak{R}$. *If communication is reliable and has bounded delay, then every replica* $R \in \mathfrak{R}$ *will, in round* ρ, *receive a set* $\{(\langle T_i \rangle_{c_i}, [\langle T_i \rangle_{c_i}, \rho] c_i) \mid (1 \leq i \leq \mathbf{z}) \wedge (c_i \in \text{clients}(\mathcal{C}_i))\}$ *of* $\mathbf{z}$ *messages. These sets all contain identical client requests.*



$$m = (\langle T \rangle_c, [\langle T \rangle_c, \rho] c_1)$$

Evaluation

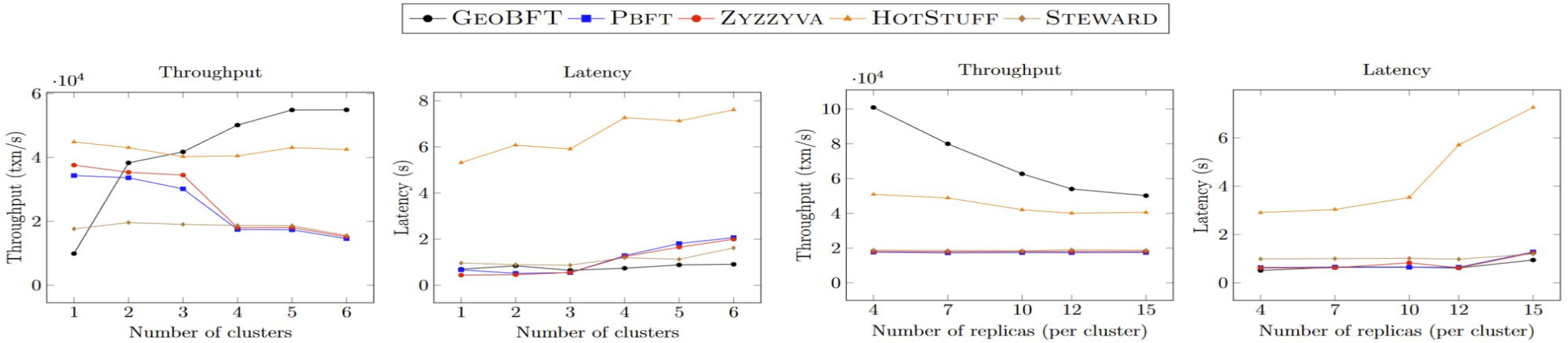


Figure 10: Throughput and latency as a function of the number of clusters; zn = 60 replicas.

Figure 11: Throughput and latency as a function of the number of replicas per cluster; z = 4.

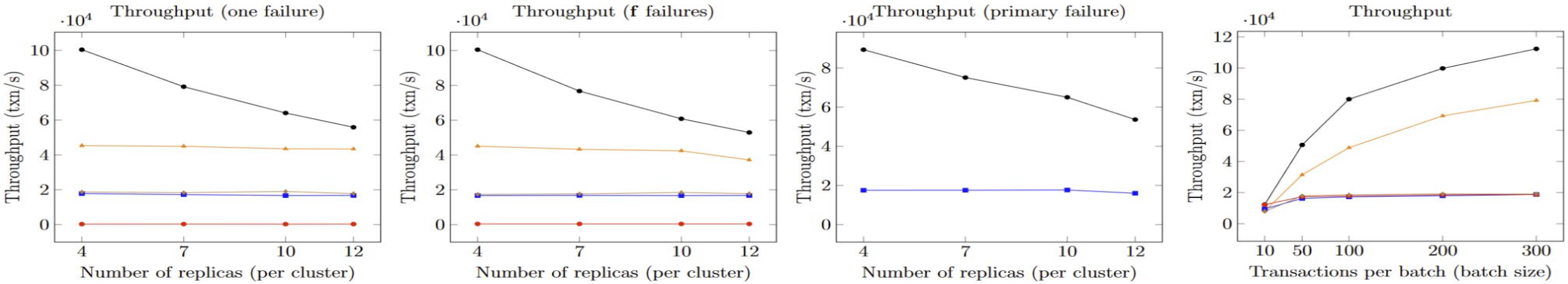
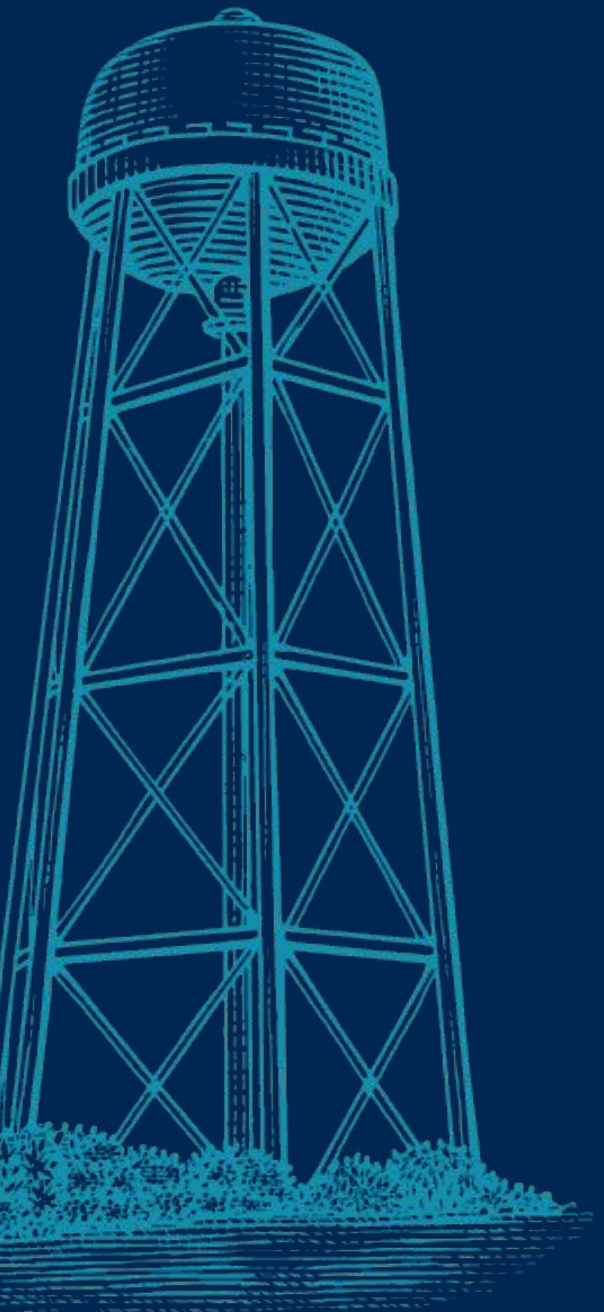


Figure 12: Throughput as a function of the number of replicas per cluster; z = 4. *Left*, throughput with one non-primary failure. *Middle*, throughput with f non-primary failures. *Right*, throughput with a single primary failure.

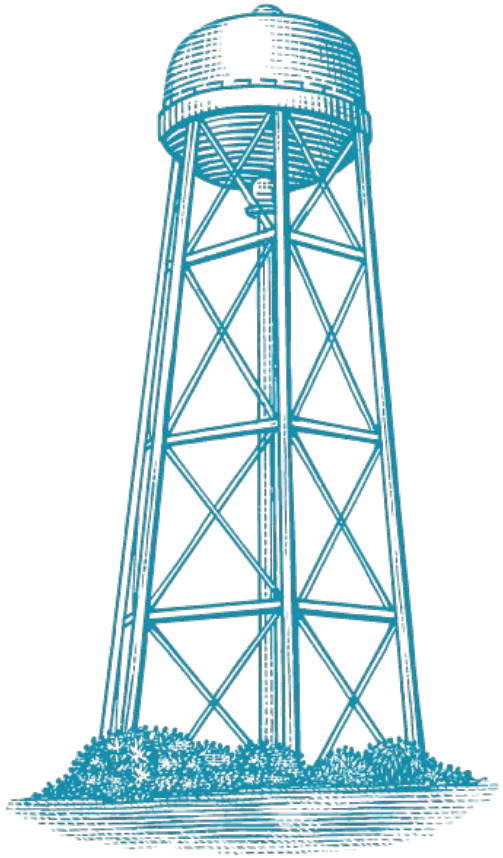
Figure 13: Throughput as a function of the batch size; z = 4 and n = 7.



Thank you!

UCDAVIS
COMPUTER SCIENCE

Comparing to other existing protocols



Protocol	Decisions	Communication		Centralized
		(Local)	*(Global)*	
GeoBFT (our paper)	$\mathbf{z}$	$\mathcal{O}(2\mathbf{zn}^2)$	$\mathcal{O}(\mathbf{fz}^2)$	No
↳ *single decision*	1	$\mathcal{O}(4\mathbf{n}^2)$	$\mathcal{O}(\mathbf{fz})$	No
Zyzzyva	1	$\mathcal{O}(\mathbf{zn})$		Yes
Pbft	1	$\mathcal{O}(2(\mathbf{zn})^2)$		Yes
PoE	1	$\mathcal{O}((\mathbf{zn})^2)$		Yes
HotStuff	1	$\mathcal{O}(8(\mathbf{zn}))$		Partly

UCDAVIS
COMPUTER SCIENCE