Consistency Analysis of Replication-Based Probabilistic Key-Value Stores

Ramy E. Ali

Was with  PennState

Now with USC Viterbi
School of Engineering

IEEE International Conference on Communications, June 2021

Outline

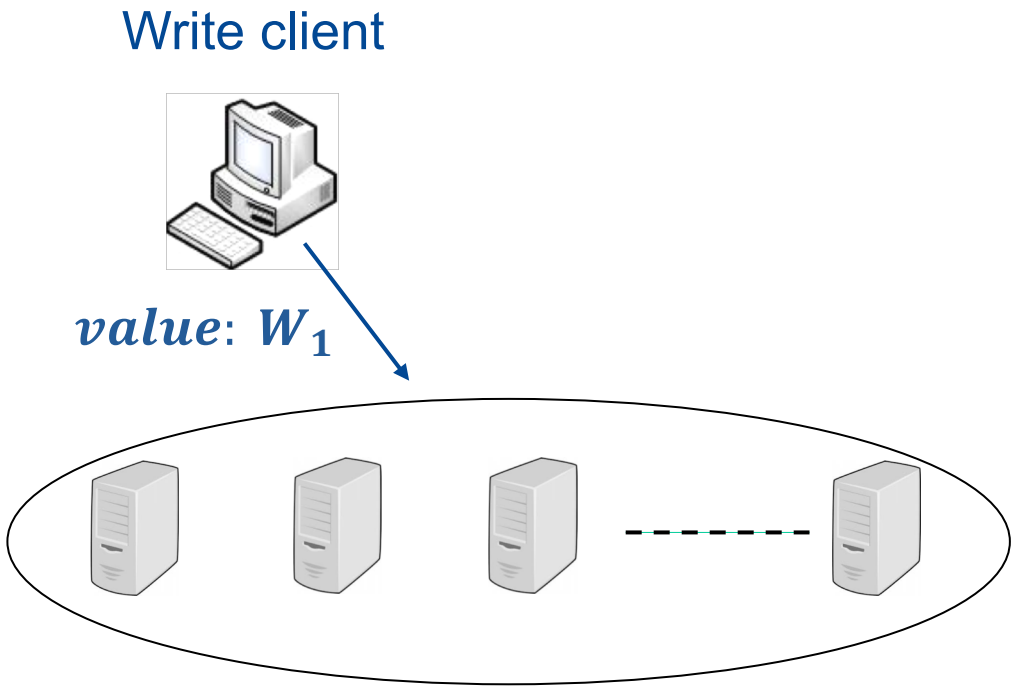
- Overview of Key-value Stores
- Strong Consistency
- Probabilistic Consistency
- Discussion & Future Work

Key-value Stores

- Applications: reservation systems, financial transactions, distributed computing, …
- Numerous key-value stores: Amazon Dynamo, Apache Cassandra, and CouchDB

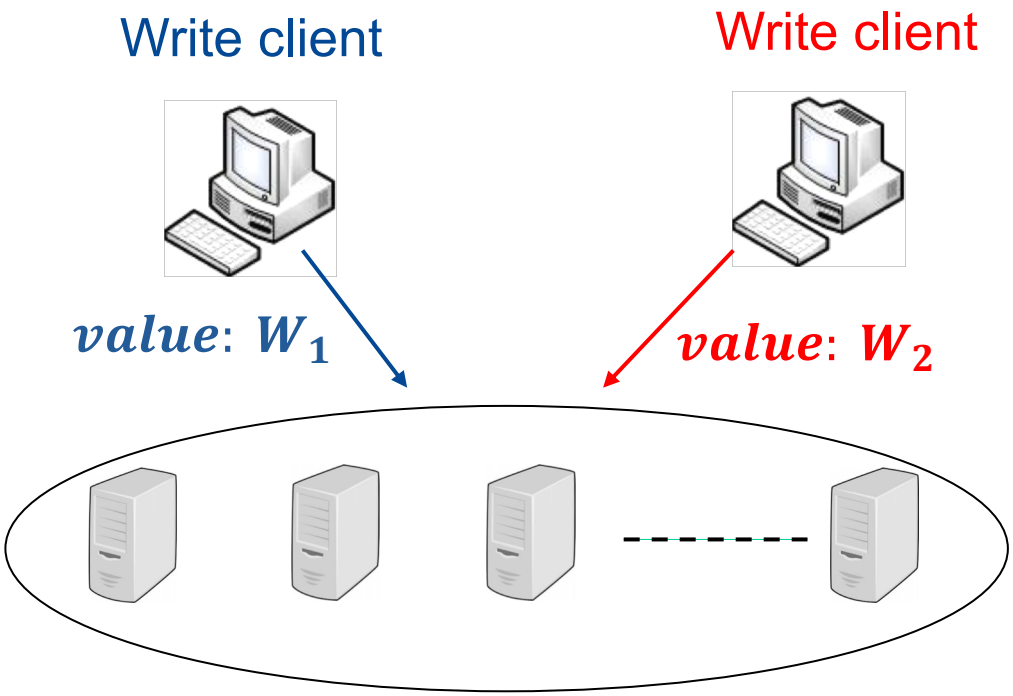
Distributed Key-value Stores

- Data is stored over multiple nodes.
- Data is asynchronously updated.



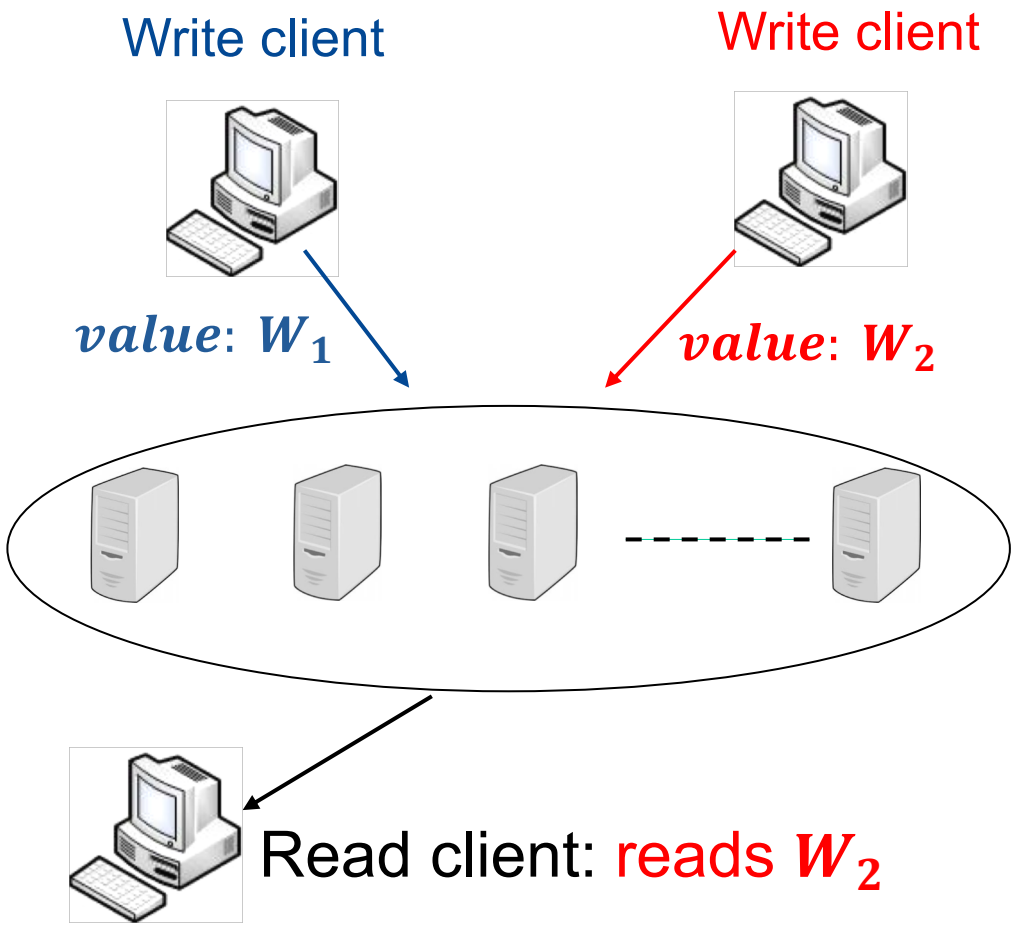
Distributed Key-value Stores

- Data is stored over multiple nodes.
- Data is asynchronously updated.



Distributed Key-value Stores

- Data is stored over multiple nodes.
- Data is asynchronously updated.
- Client must get the *latest possible version* of the data [Lamport 1979, ABD 1995].



Distributed Key-value Stores

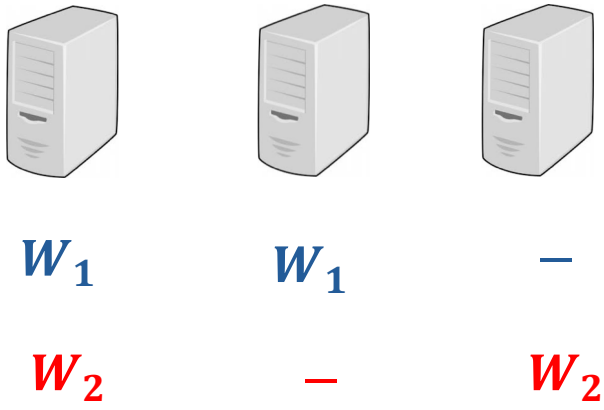
1. *Asynchrony*

   Data updates may not arrive at all servers simultaneously.

Distributed Key-value Stores

1. *Asynchrony*

   Data updates may not arrive at all servers simultaneously.



Distributed Key-value Stores

1. *Asynchrony*

   Data updates may not arrive at all servers simultaneously.

2. *Decentralized Nature*

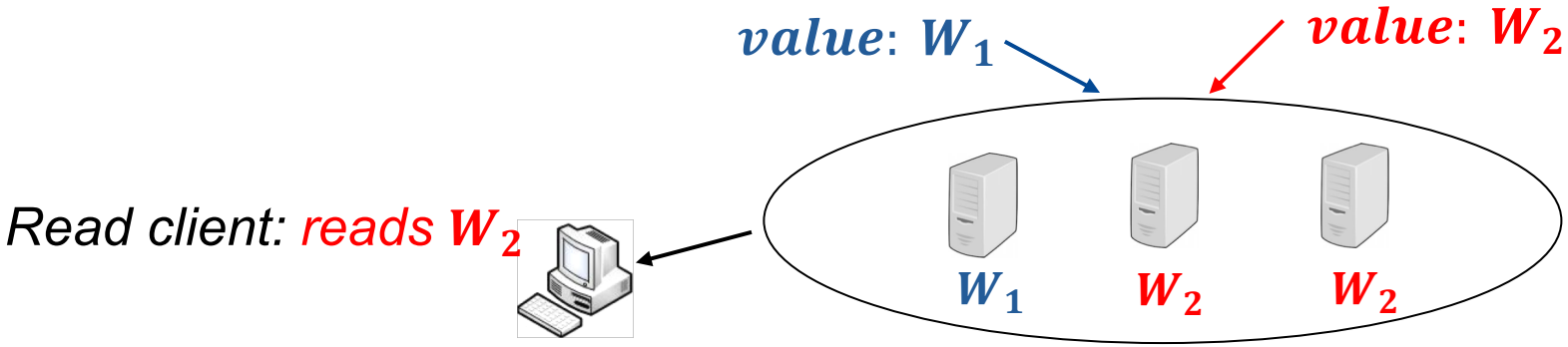
   A server may not be aware of which updates received by others.

Distributed Key-value Stores

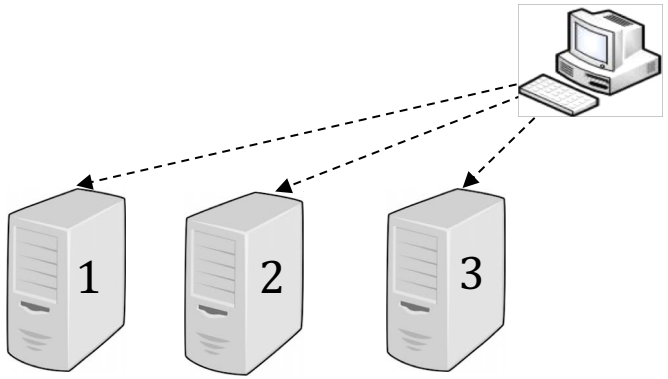
1. *Asynchrony*
   Data updates may not arrive at all servers simultaneously.

2. *Decentralized Nature*
   A server may not be aware of which updates received by others.

3. *Consistency*
   A client must retrieve the latest possible update.



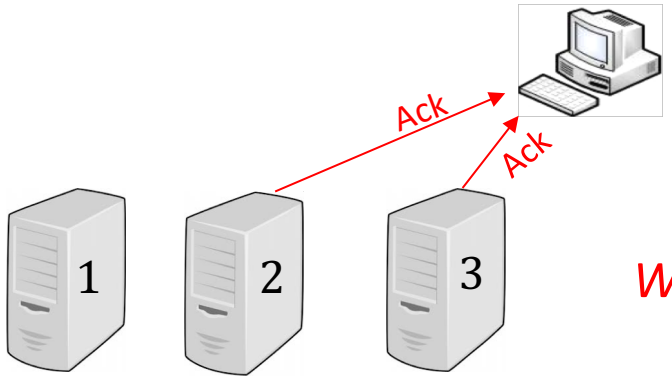
Strong Consistency: Strict Quorums



Write client

Send the request to all, wait for W to respond

Strong Consistency: Strict Quorums

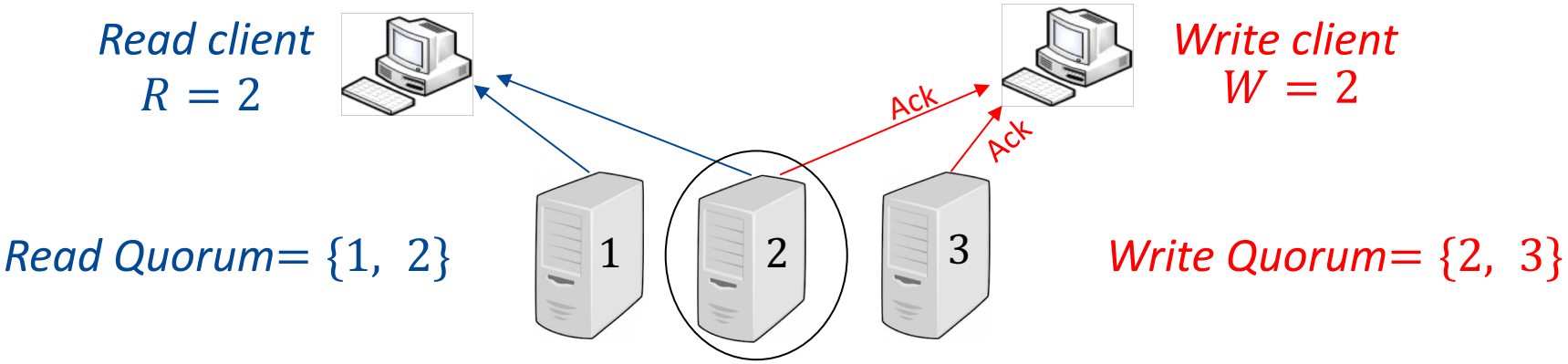


Write client
$W = 2$

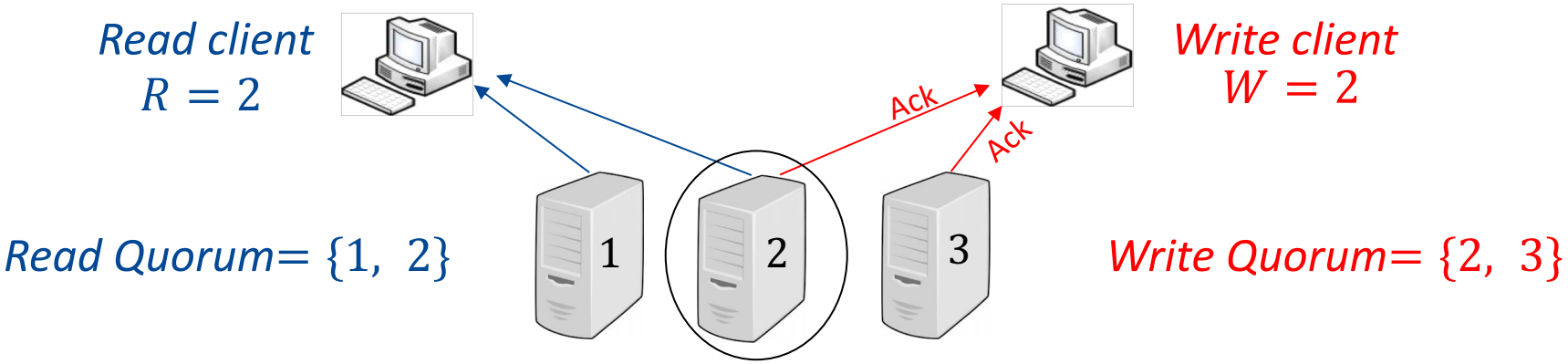
Write Quorum$= \{2,\ 3\}$

Send the request to all, wait for W to respond

Strong Consistency: Strict Quorums



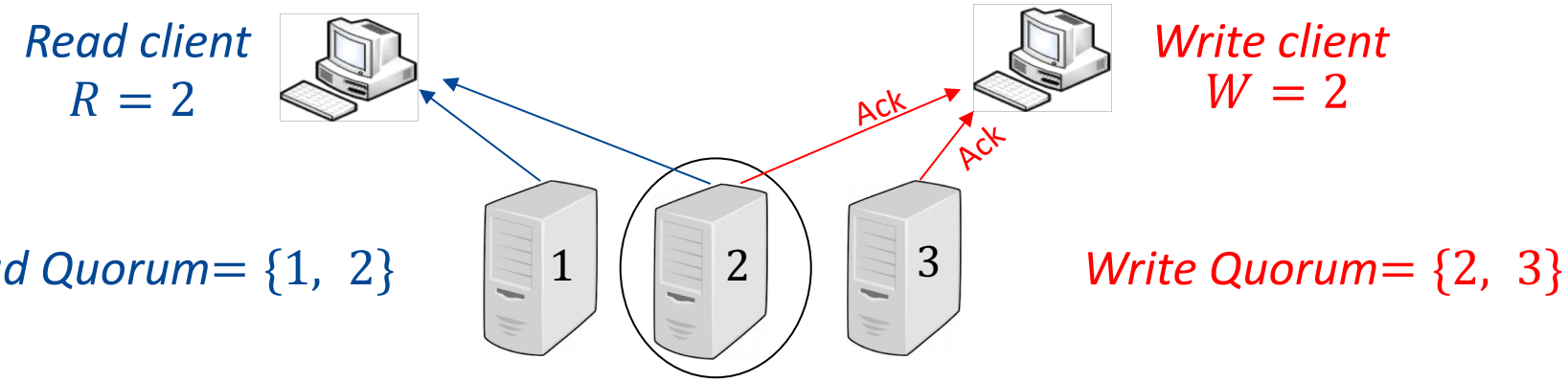
Strong Consistency: Strict Quorums



Strong Consistency: $W + R > N$

$$\{1, 2\} \cap \{2,3\} = \{2\}$$

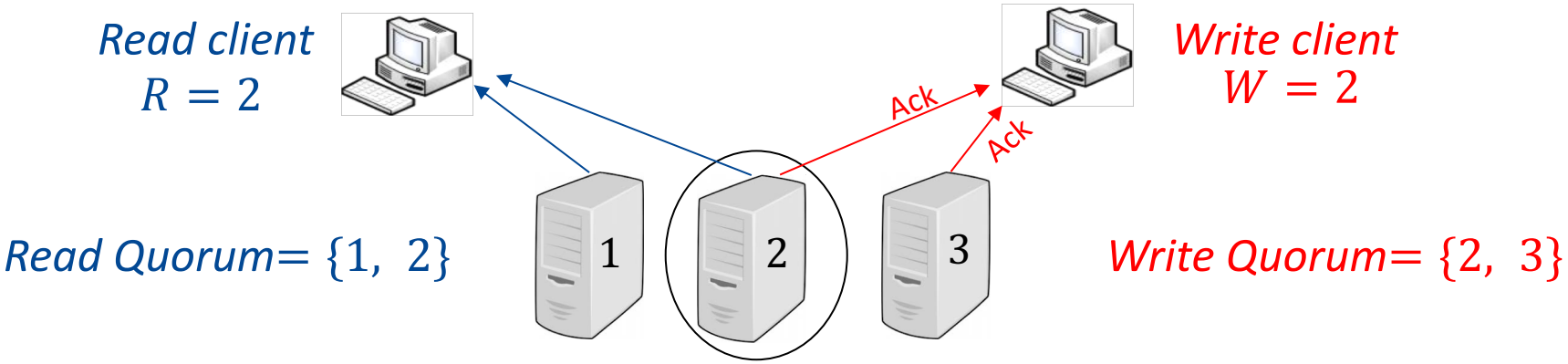
Strong Consistency: Strict Quorums



Strong Consistency: $W + R > N$

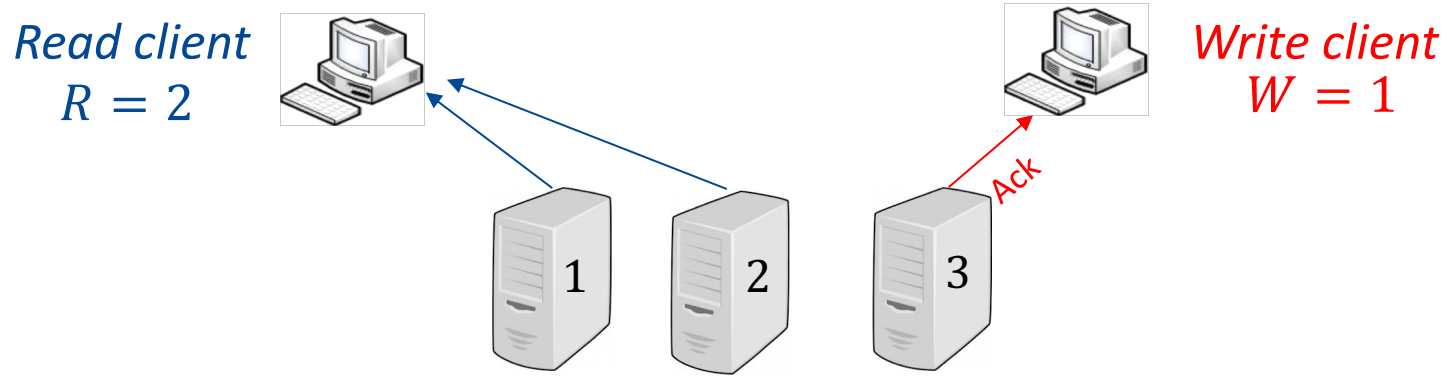
High Write and Read Latency

Strong Consistency: Strict Quorums



Strong Consistency: $W + R > N$

High Write and Read Latency

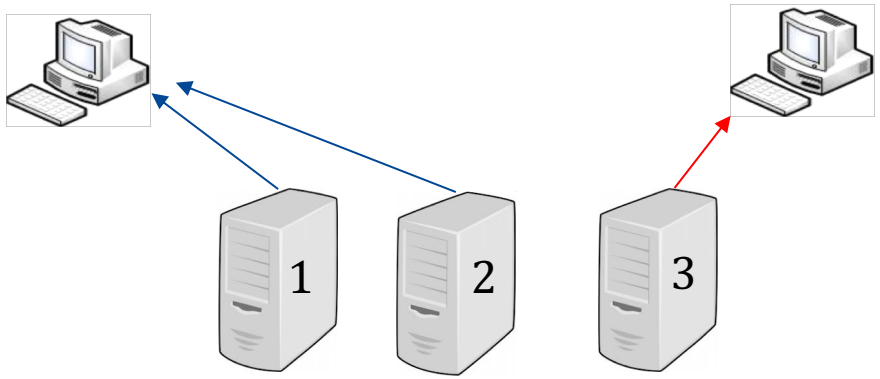


Probabilistic Consistency: $W + R \leq N$

Quorums may not intersect!

Probabilistic Consistency: Probabilistic Quorums

Read client
$R = 2$



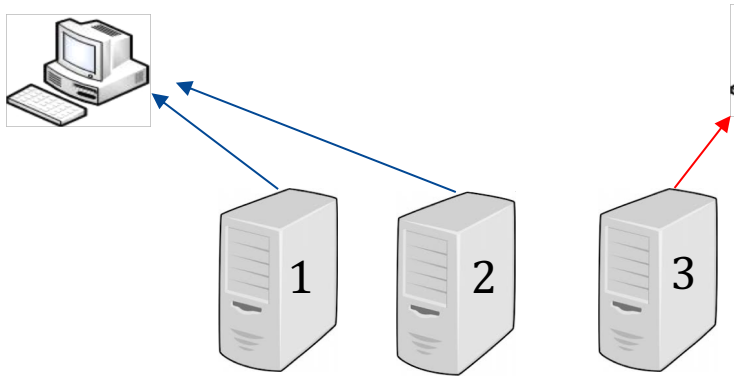
Write client
$W = 1$

$W + R \leq N$

Quorums may not intersect!

Probabilistic Consistency: Probabilistic Quorums

Read client
$R = 2$



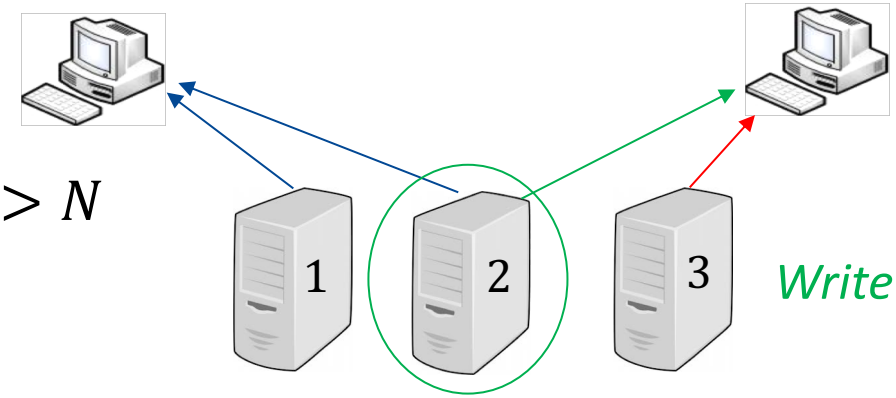
Write client
$W = 1$

$W + R \leq N$

Quorums may not intersect!

Expanding Quorums: More severs receive the write request

Read client
$R = 2$



Write client
$W = 1$

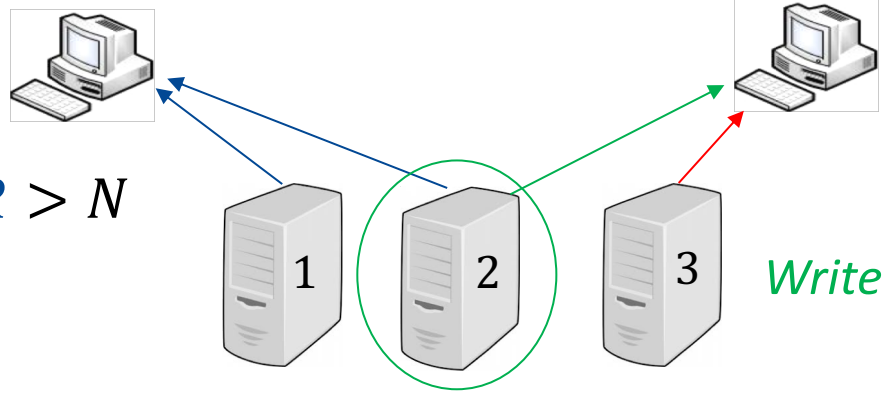
Eventually $S(t) + R > N$

$S(t) = 2$

Write quorum size at time t

Probabilistic Consistency: Probabilistic Quorums

Read client
$R = 2$



Write client
$W = 1$

Eventually $S(t) + R > N$

$S(t) = 2$

Write quorum size at time t

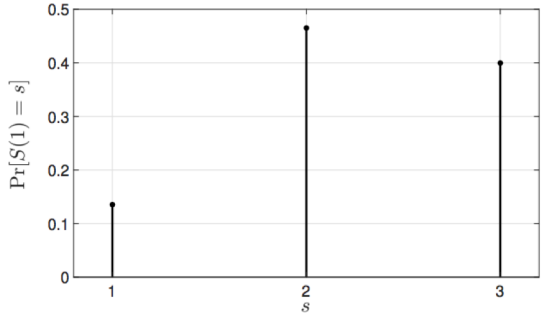


Fig. 1: The probability mass function of $S(1)$ for the case where $N = 3, W = 1$ and $\lambda = 1$.

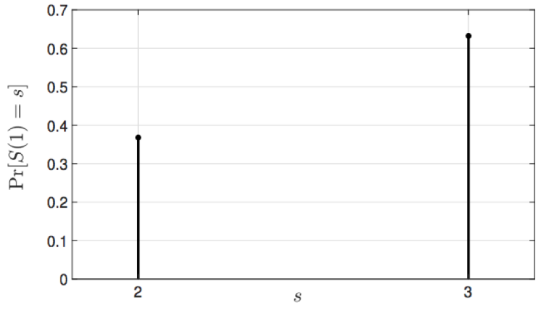
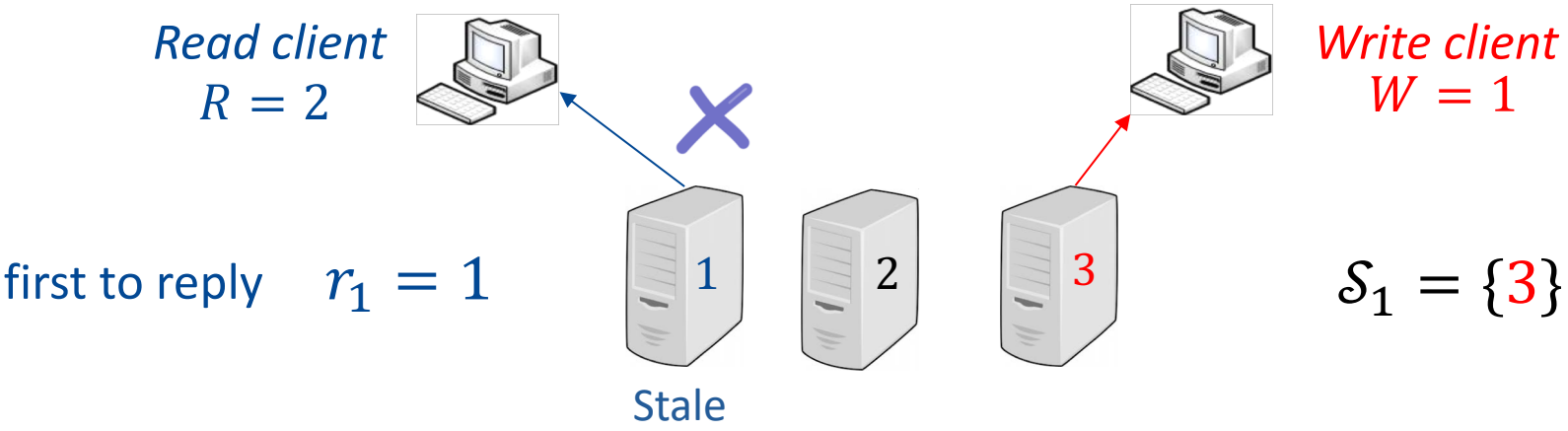


Fig. 2: The probability mass function of $S(1)$ for the case where $N = 3, W = 2$ and $\lambda = 1$.

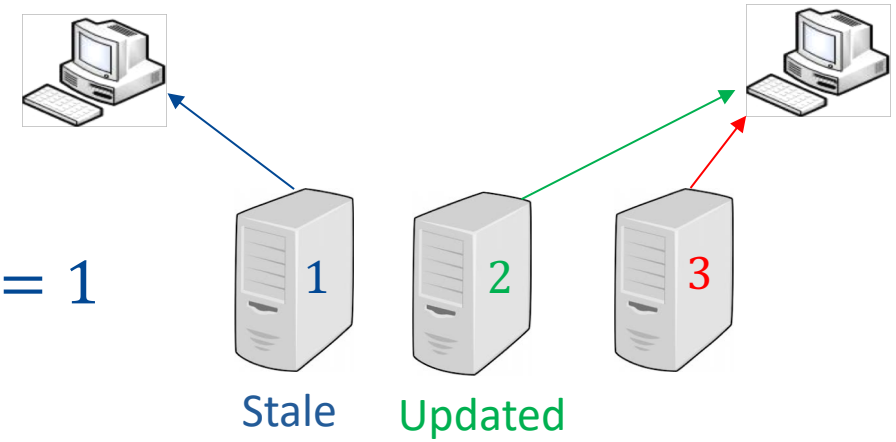
Probabilistic Consistency: Probabilistic Quorums



$$\Pr[\text{Inconsistency at time } t] = ?$$

Stale: replies to the read request before receiving the write request

Probabilistic Consistency: Probabilistic Quorums



first to reply $r_1 = 1$

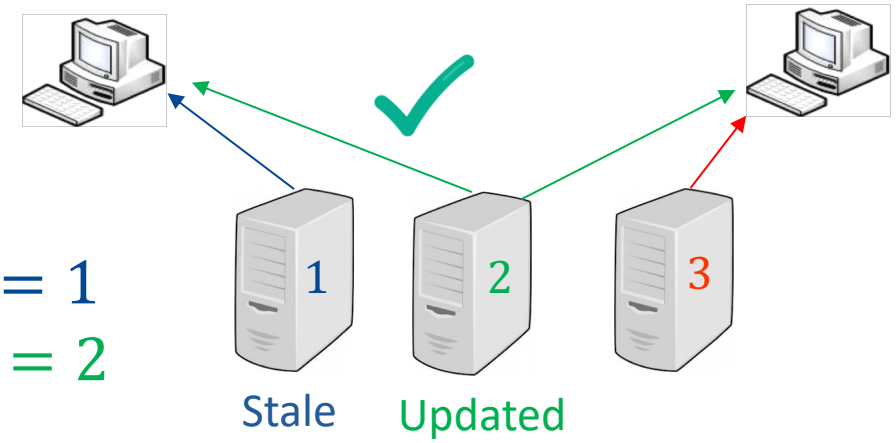
$\mathcal{S}_1 = \{3\}$

$\mathcal{S}_2 = \{2, 3\}$

$\Pr[\text{Inconsistency at time } t] = ?$

Stale: replies to the read request before receiving the write request

Probabilistic Consistency: Probabilistic Quorums



first to reply $r_1 = 1$

second to reply $r_2 = 2$

$\mathcal{S}_1 = \{3\}$

$\mathcal{S}_2 = \{2, 3\}$

$$\Pr[\text{Inconsistency at time } t] = ?$$

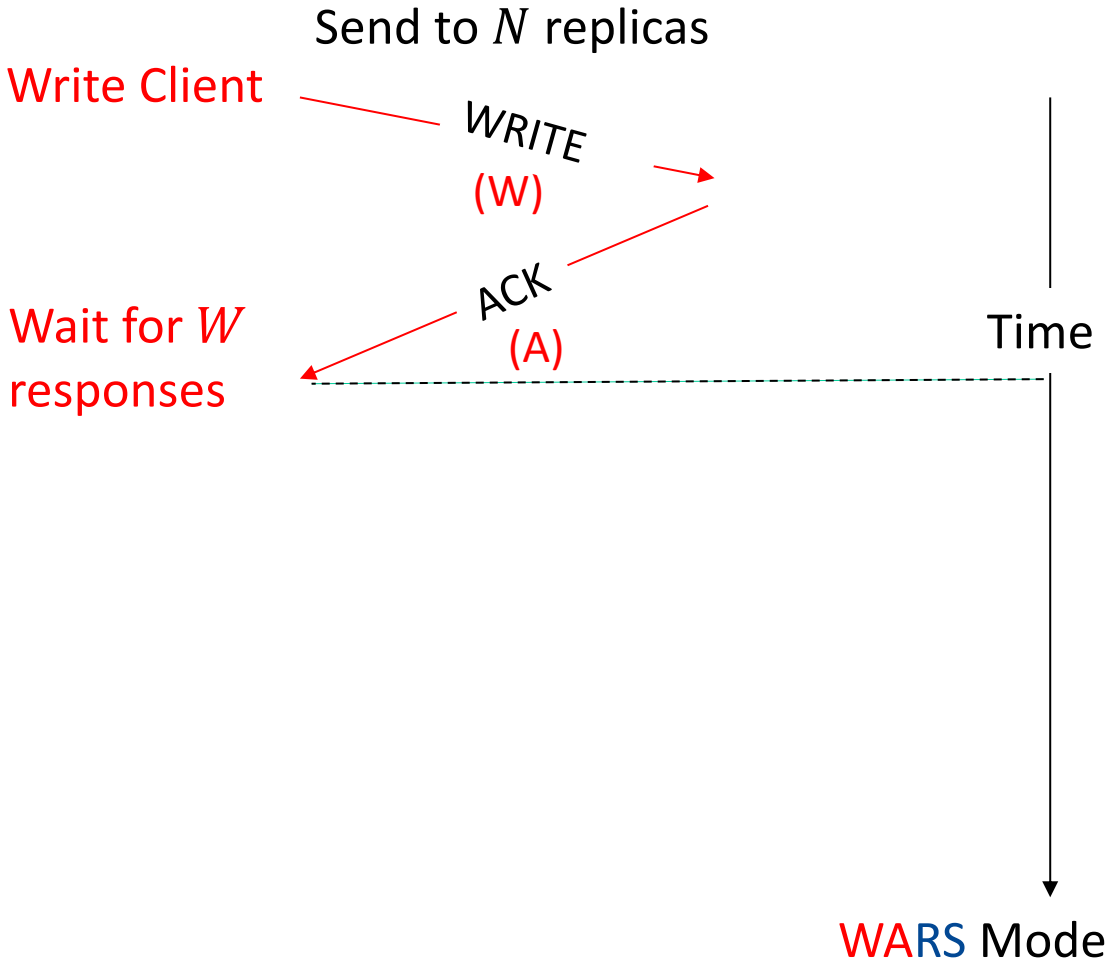
Stale: replies to the read request before receiving the write request

$$\Pr[\text{Inconsistency at time } t] = \Pr[r_1\ stale, r_2\ stale]$$

Probabilistic Consistency: 3-way Replication

$$\Pr[\text{Inconsistency at time } t] = ?$$

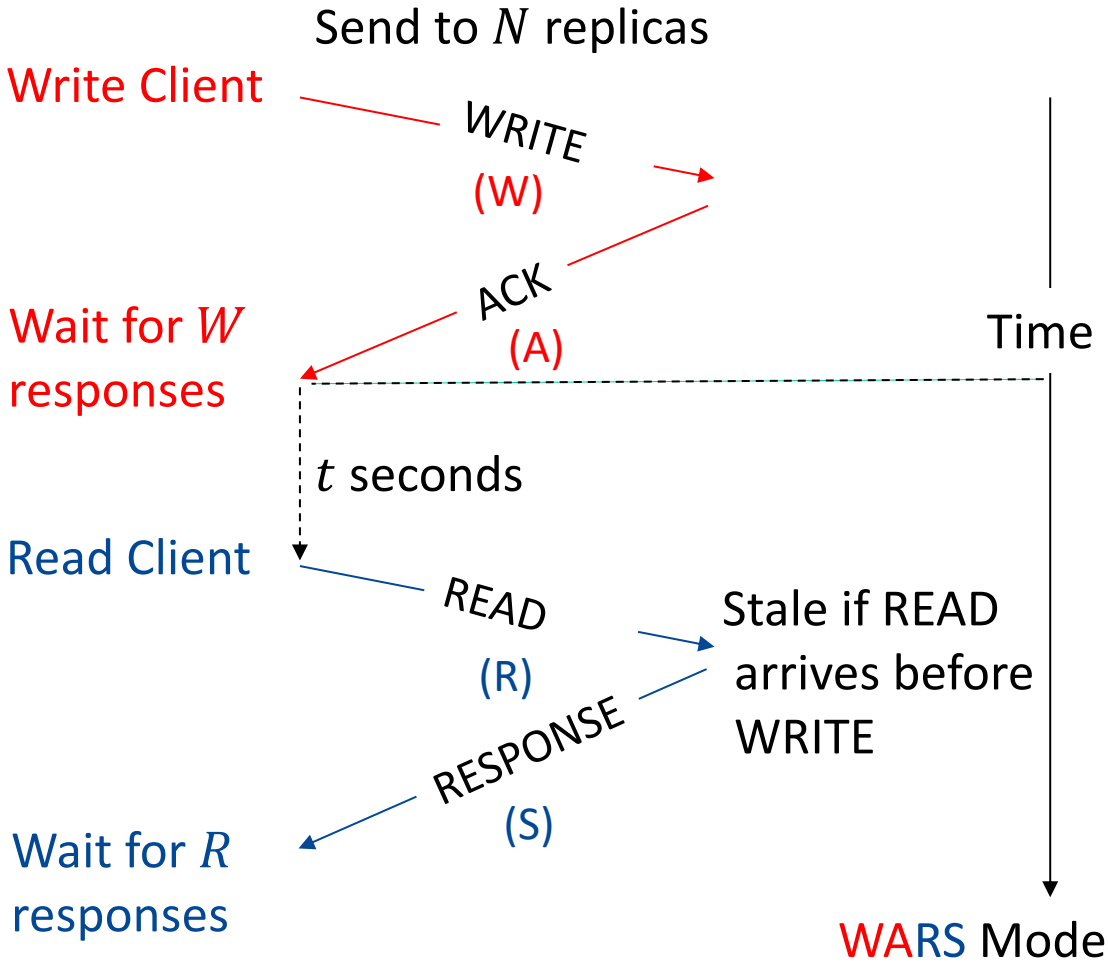
[Bailis et al. 2014]



Probabilistic Consistency: 3-way Replication

$$\Pr[\text{Inconsistency at time } t] = ?$$

[Bailis et al. 2014]

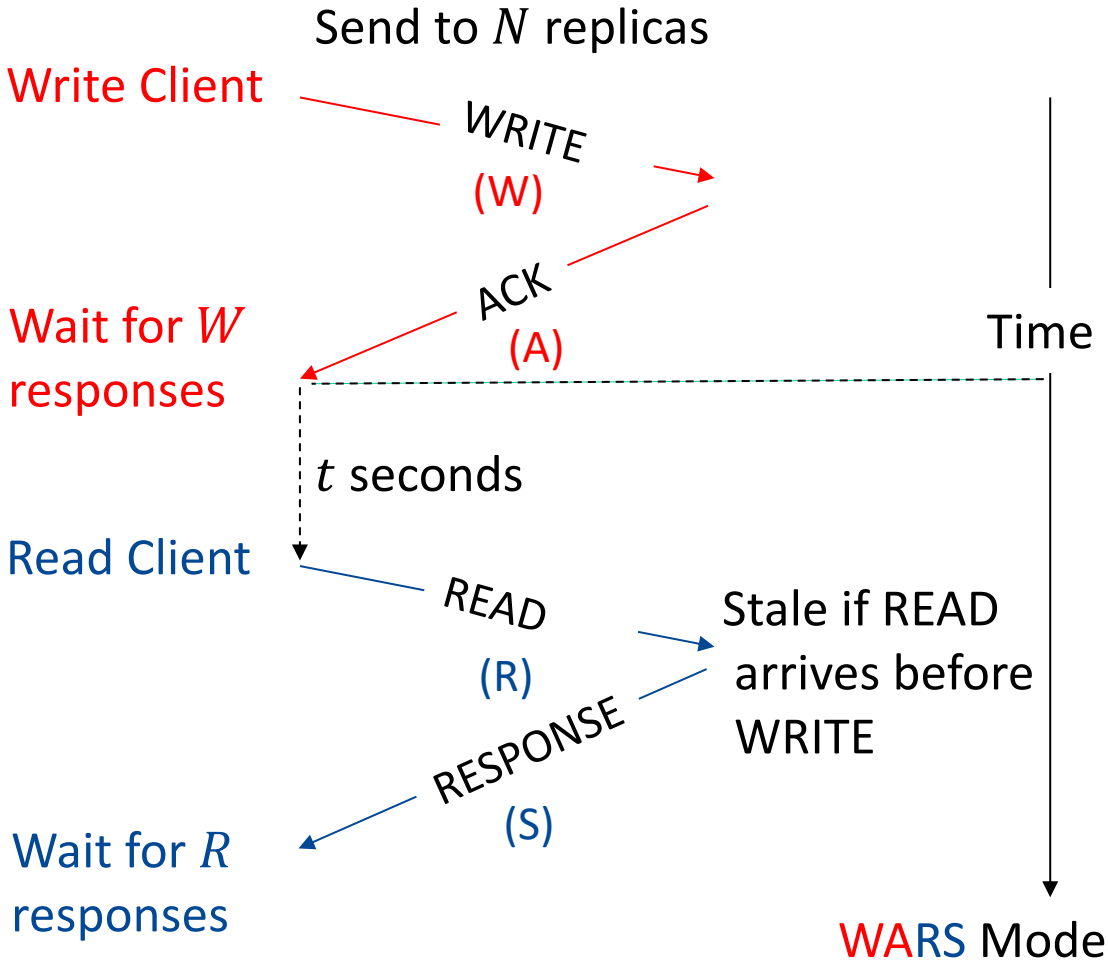


Probabilistic Consistency: 3-way Replication

$$\Pr[\text{Inconsistency at time } t] = ?$$

[Bailis et al. 2014]

Monte-Carlo Simulations



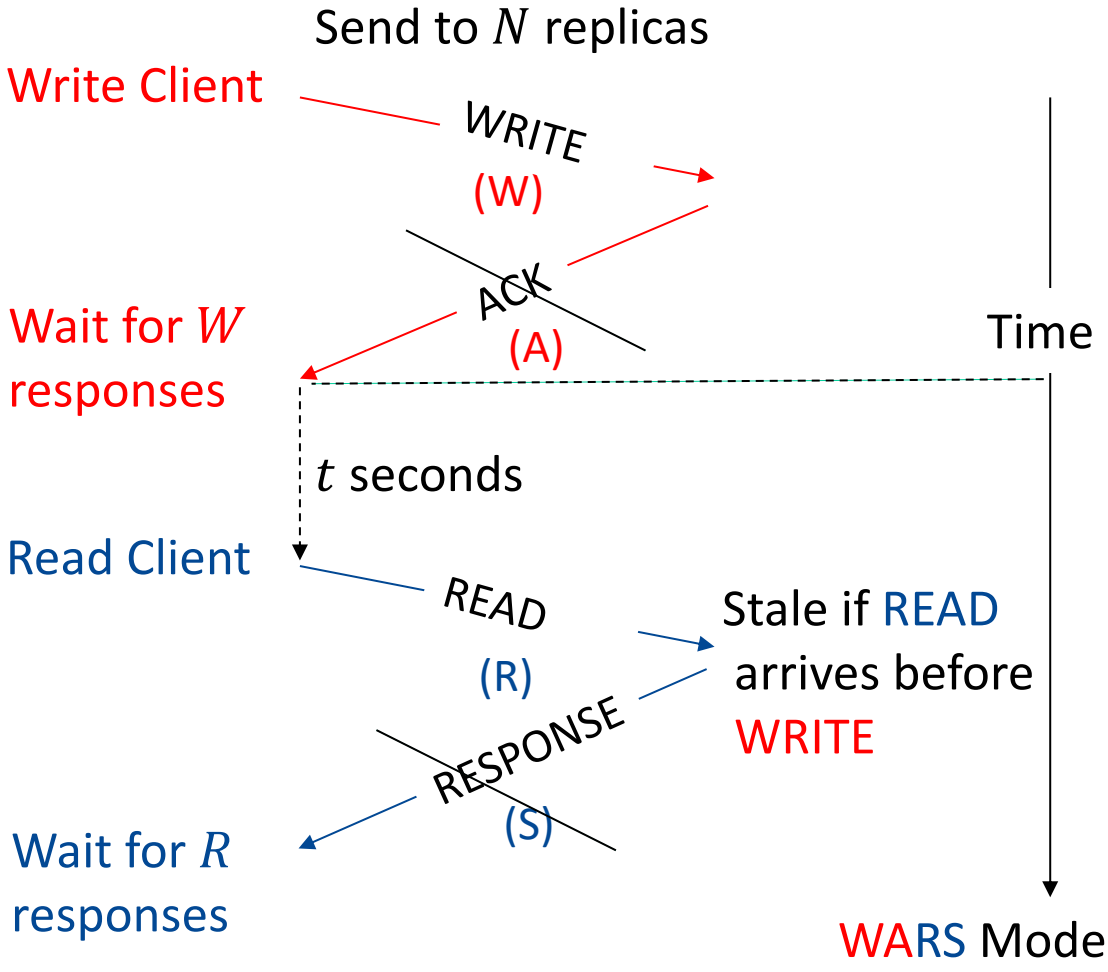
Probabilistic Consistency: 3-way Replication

$$\Pr[\text{Inconsistency at time } t] = ?$$

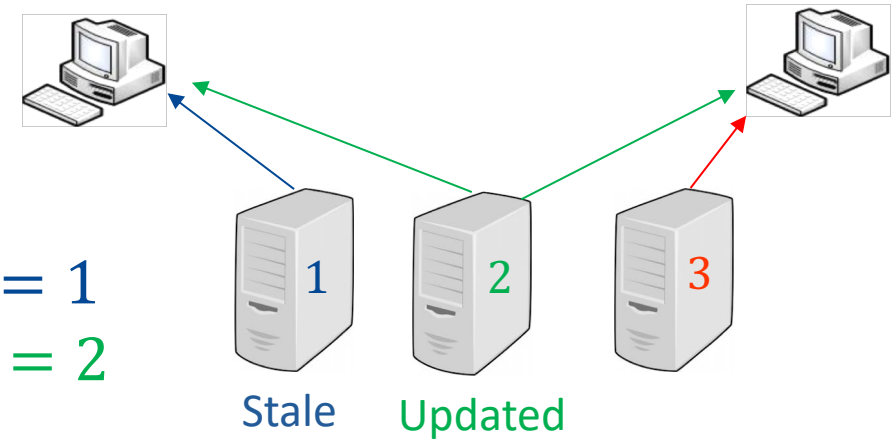
[Bailis et al. 2014]

Monte-Carlo Simulations

This Work: Solves the problem analytically



Probabilistic Consistency: 3-way Replication



first to reply $r_1 = 1$

second to reply $r_2 = 2$

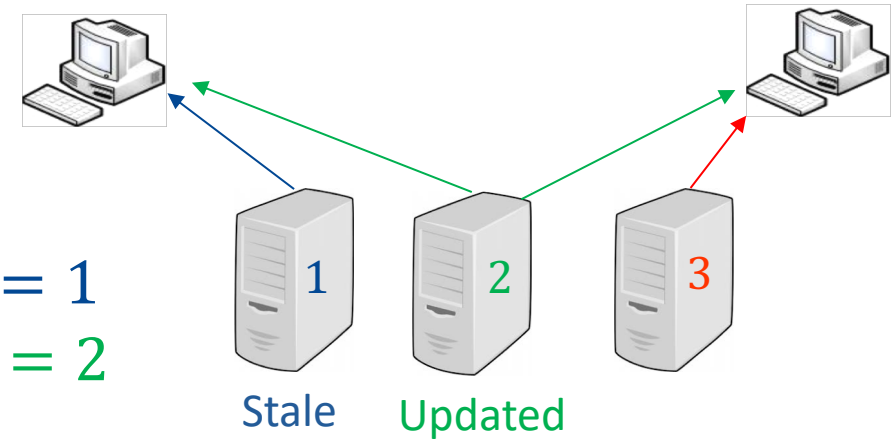
$\mathcal{S}_1 = \{3\}$

$\mathcal{S}_2 = \{2, 3\}$

$$\Pr[\text{Inconsistency at time } t] = \Pr[r_1 \notin \mathcal{S}_1, r_2 \notin \mathcal{S}_2]$$

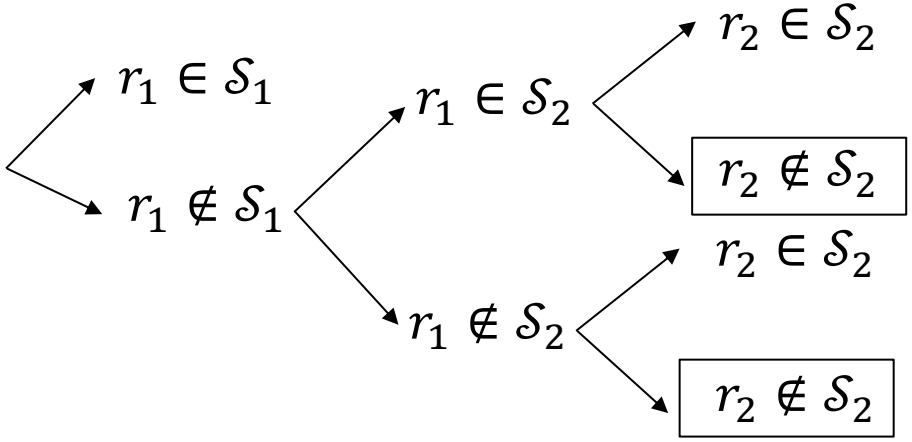
Dependent Events

Probabilistic Consistency: 3-way Replication

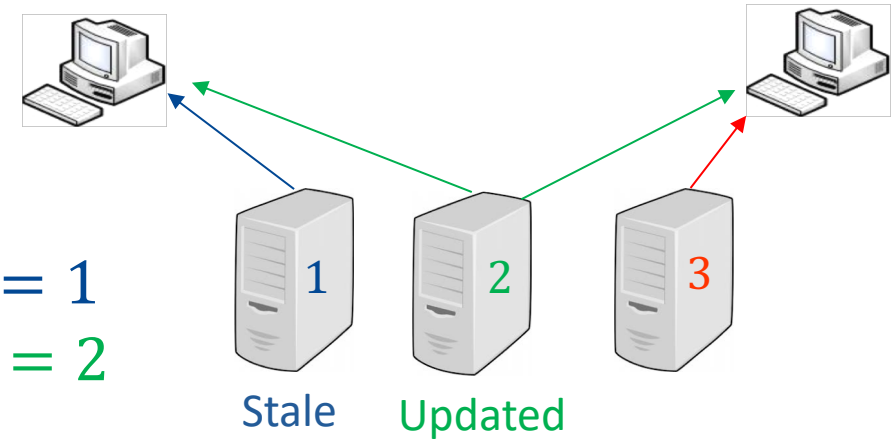


$$\Pr[\text{Inconsistency at time } t] = \Pr[r_1 \notin \mathcal{S}_1, r_2 \notin \mathcal{S}_2]$$

Dependent Events



Probabilistic Consistency: 3-way Replication



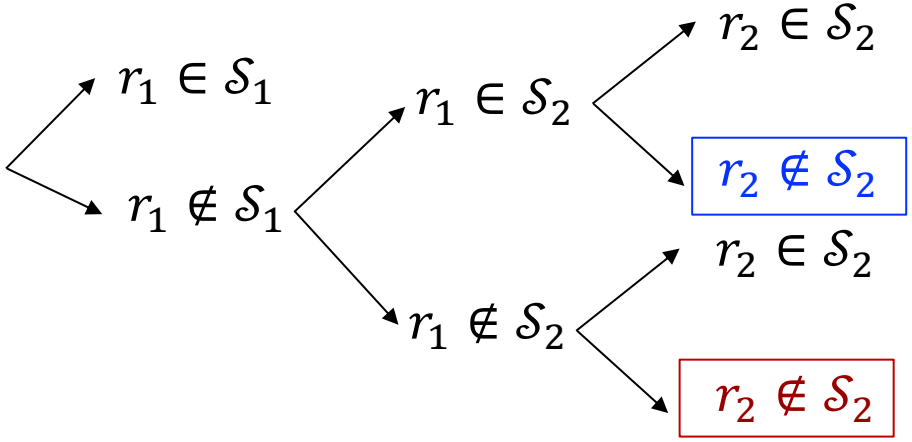
first to reply $r_1 = 1$

second to reply $r_2 = 2$

$\mathcal{S}_1 = \{3\}$

$\mathcal{S}_2 = \{2, 3\}$

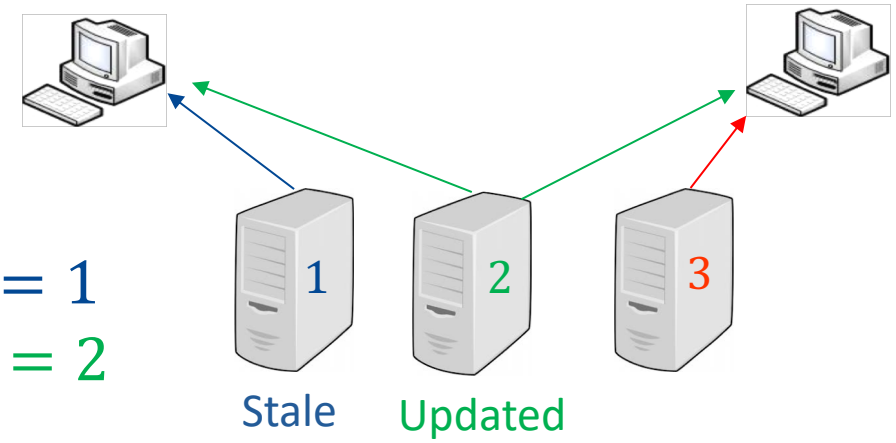
$$\Pr[\text{Inconsistency at time } t] = \Pr[r_1 \notin \mathcal{S}_1, r_2 \notin \mathcal{S}_2]$$



$$= \Pr[r_1 \in \mathcal{S}_2 - \mathcal{S}_1, r_2 \notin \mathcal{S}_2]$$

$$+ \Pr[r_1 \notin \mathcal{S}_2, r_2 \notin \mathcal{S}_2]$$

Probabilistic Consistency: 3-way Replication



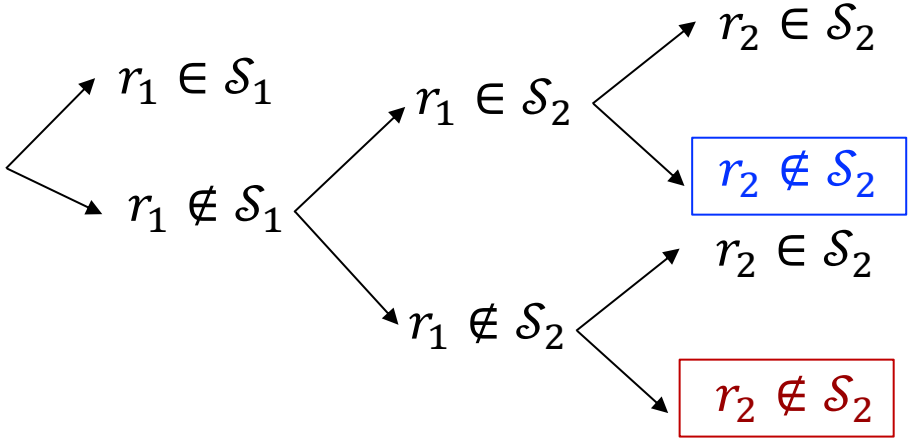
first to reply $r_1 = 1$

second to reply $r_2 = 2$

$\mathcal{S}_1 = \{3\}$

$\mathcal{S}_2 = \{2, 3\}$

$$\Pr[\text{Inconsistency at time } t] = \Pr[r_1 \notin \mathcal{S}_1, r_2 \notin \mathcal{S}_2]$$



$$= \Pr[r_1 \in \mathcal{S}_2 - \mathcal{S}_1, r_2 \notin \mathcal{S}_2]$$

$$+ \Pr[r_1 \notin \mathcal{S}_2, r_2 \notin \mathcal{S}_2]$$

⇒ Closed-form

Probabilistic Consistency: 3-way Replication

$W = 1, R = 1$

$$\Pr[\text{Inconsistency at time } t] = \frac{2\xi}{\lambda + 3\xi} e^{-\lambda t}$$

$W = 2, R = 1$

$$\Pr[\text{Inconsistency at time } t] = \frac{\xi}{\lambda + 3\xi} e^{-\lambda t}$$

$W = 1, R = 2$

$$\Pr[\text{Inconsistency at time } t] = \frac{6\xi^3 e^{-2\lambda t}}{(\lambda + 2\xi)(\lambda + 3\xi)} \left(\frac{2\lambda}{(\lambda + 2\xi)(\lambda + 3\xi)} - \frac{(\lambda - \xi)e^{-\lambda t}}{(\lambda + \xi)(2\lambda + 3\xi)} \right)$$

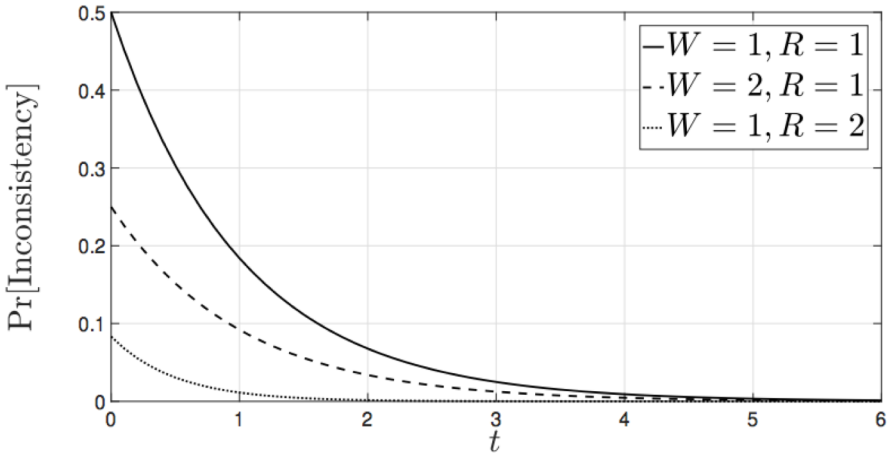


Fig. 3: The probability of inconsistency for the case where $N = 3, \lambda = 1$ and $\xi = 1$.

Write Delay ~ exp(λ)

Read Delay ~ exp(ξ)

Probabilistic Consistency: General N

In general: $R!$ cases

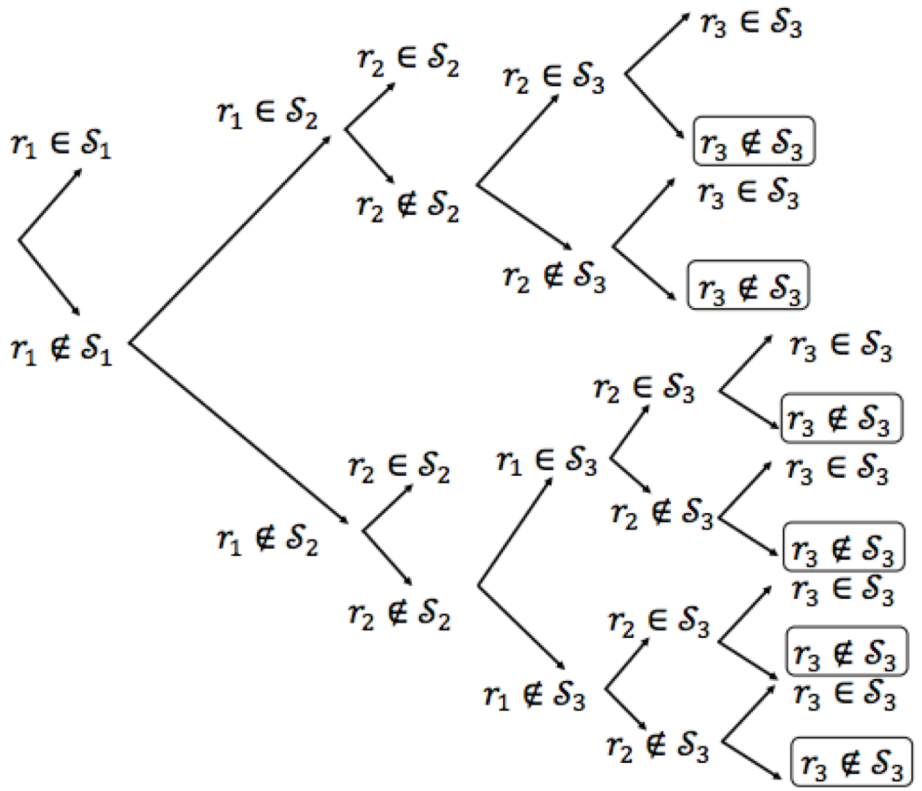


Fig. 4: Inconsistency cases for the case where $R = 3$.

Discussion & Future Work

- A closed-form expression for 3-way replication is derived assuming exponential delays.
- The inconsistency probability can be derived for any delay distributions.
- Extending this work to erasure-coded key-value stores is an interesting future direction.

Questions?

E-mail: reali@usc.edu

Thank You

References

- P. Bailis et al. **"Probabilistically bounded staleness for practical partial quorums."** Proceedings of the VLDB Endowment 5.8 (2012): 776-787.
- P. Bailis et al. **"Quantifying eventual consistency with PBS."** The VLDB Journal 23.2 (2014): 279-302.