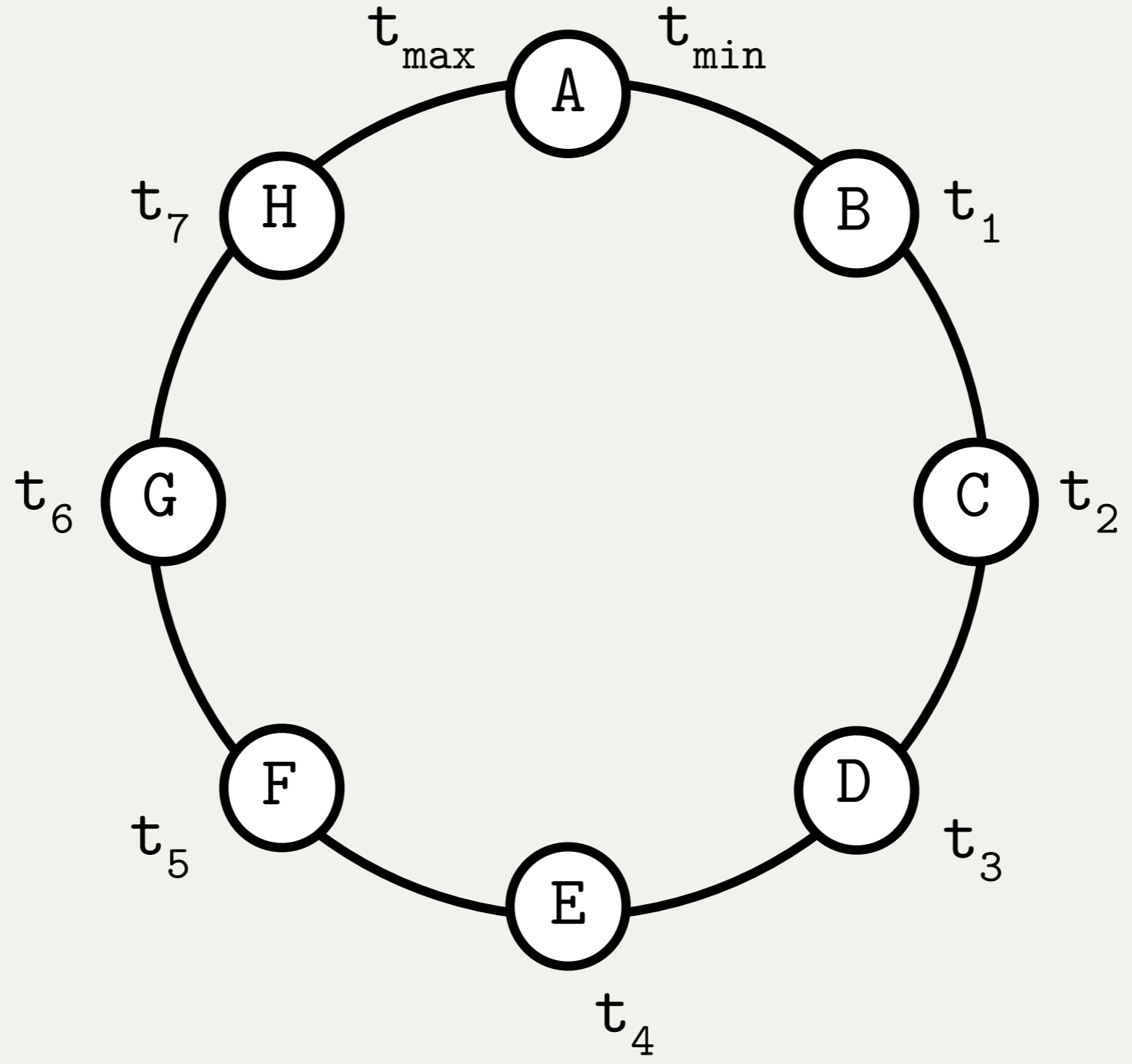


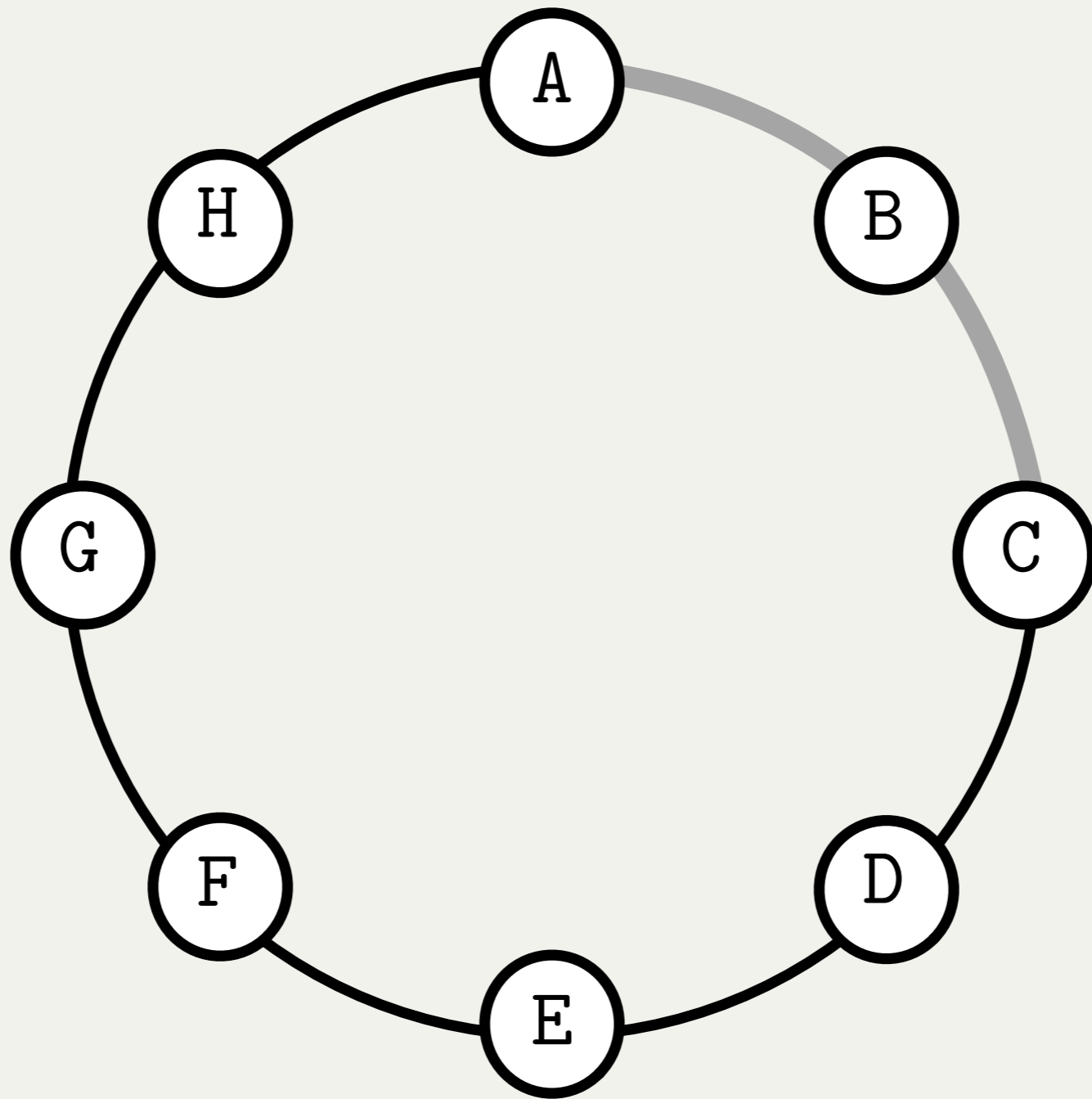
# Transient Replication

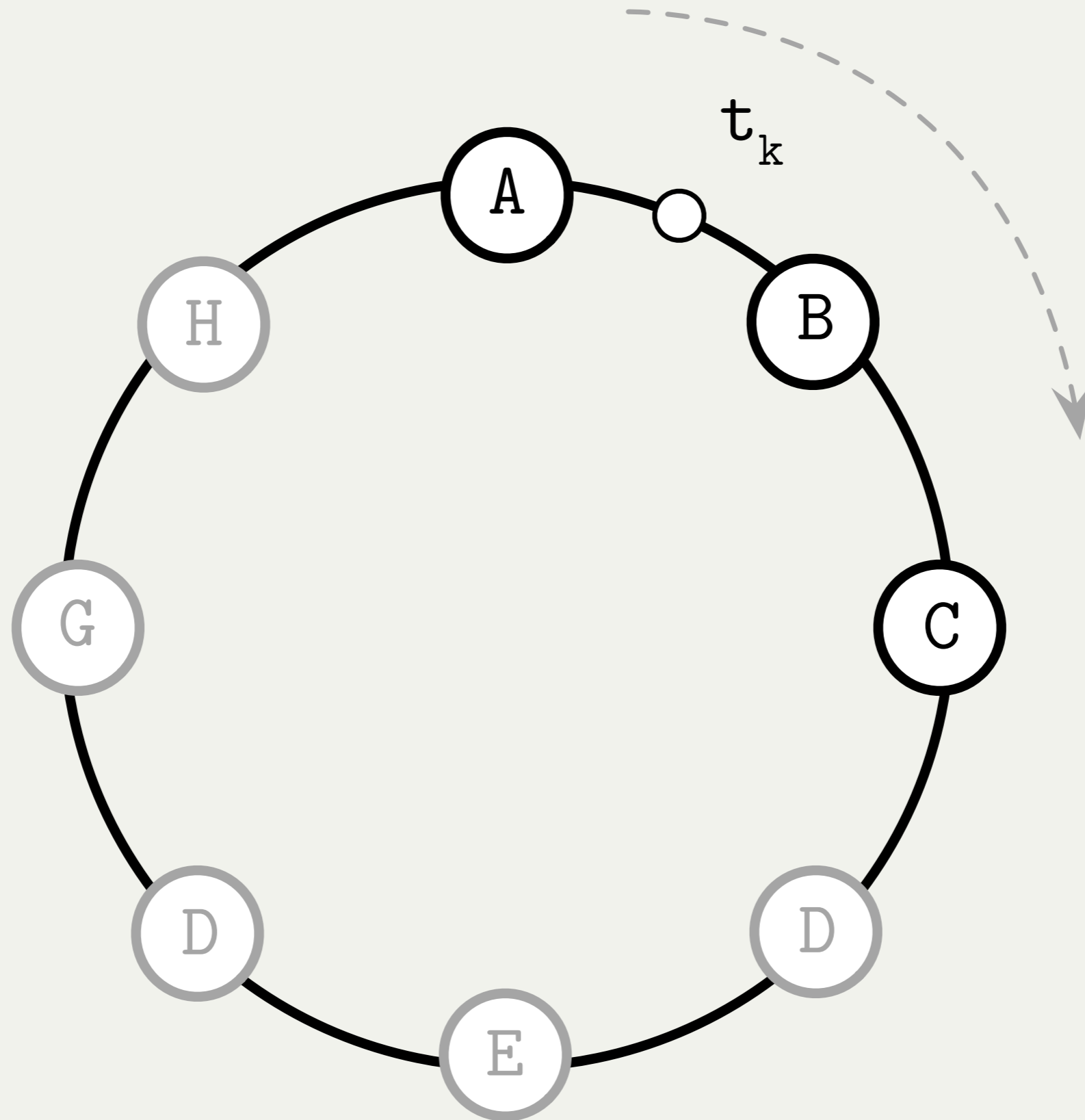
and Cheap Quorums

# Prerequisites

- Eventual Consistency
- Quorums
- Consistent Hashing
- Anti-Entropy



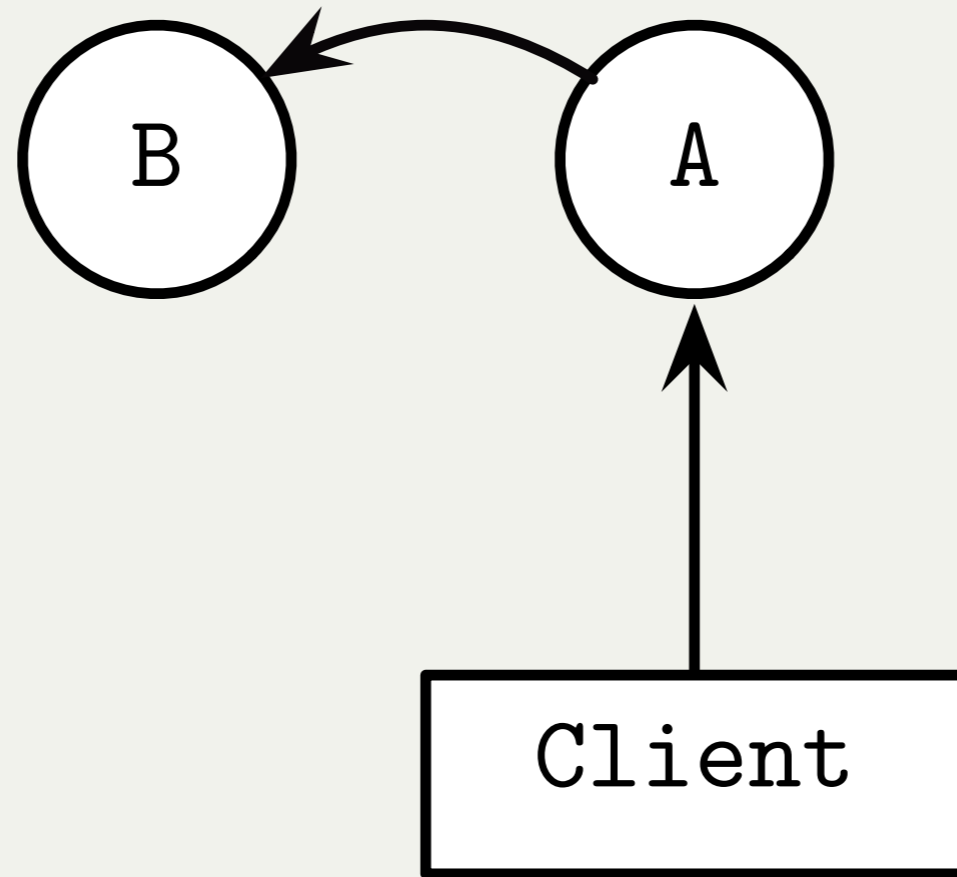


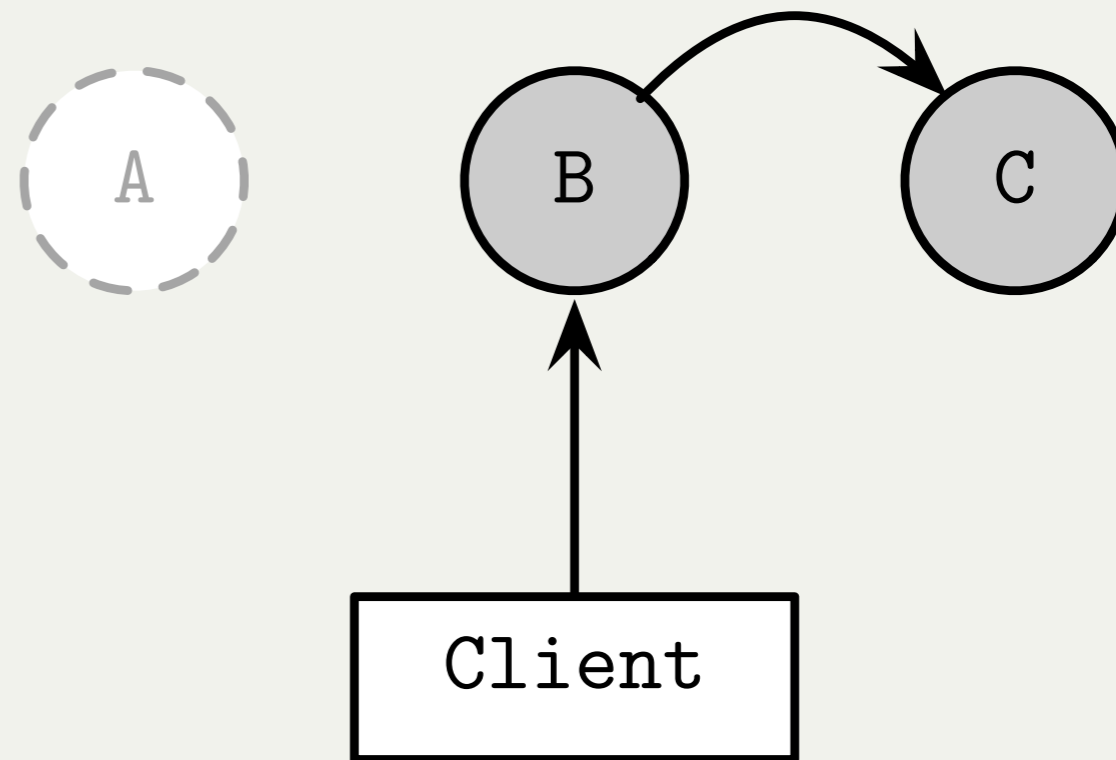
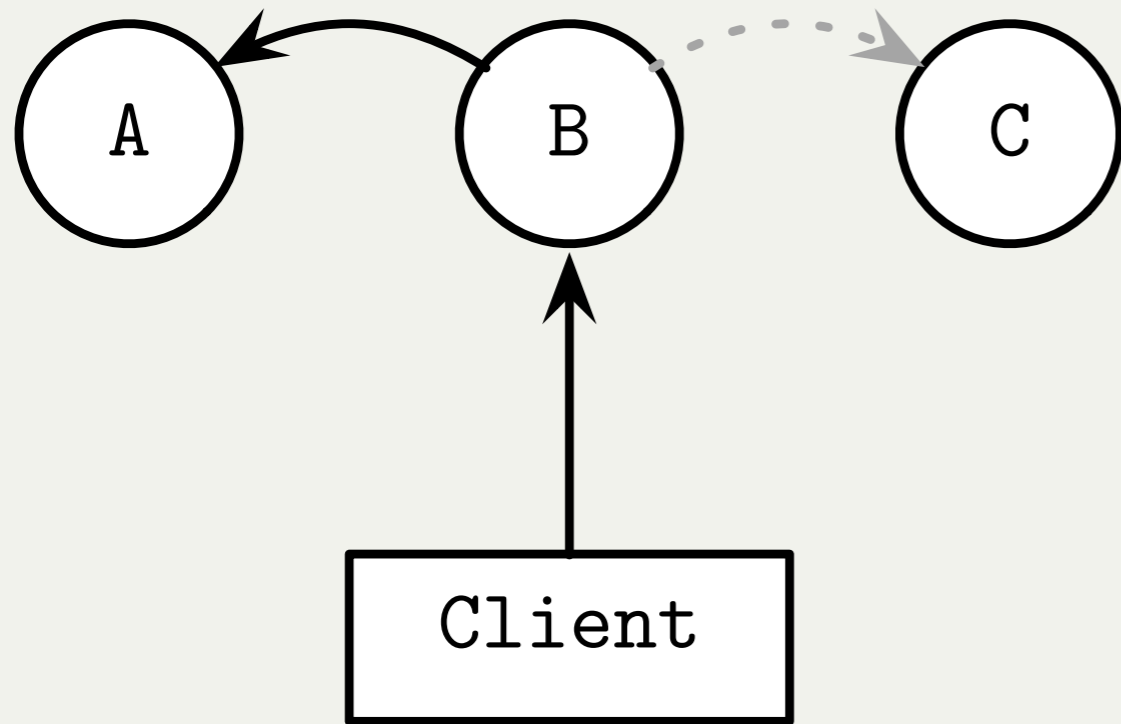


# Eventual Consistency

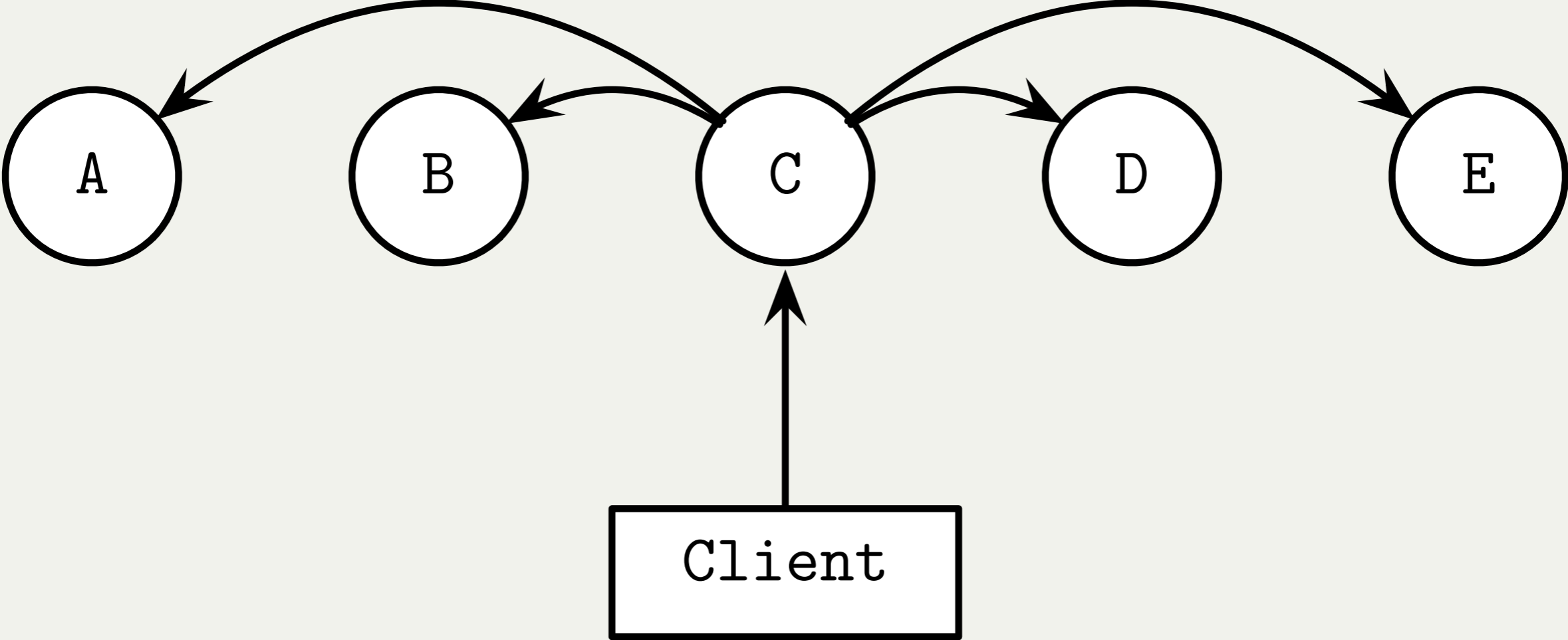
“The storage system guarantees that if no new updates are made to the object, eventually all accesses will return the last updated value.”

- Werner Vogels. 2008. Eventually Consistent.

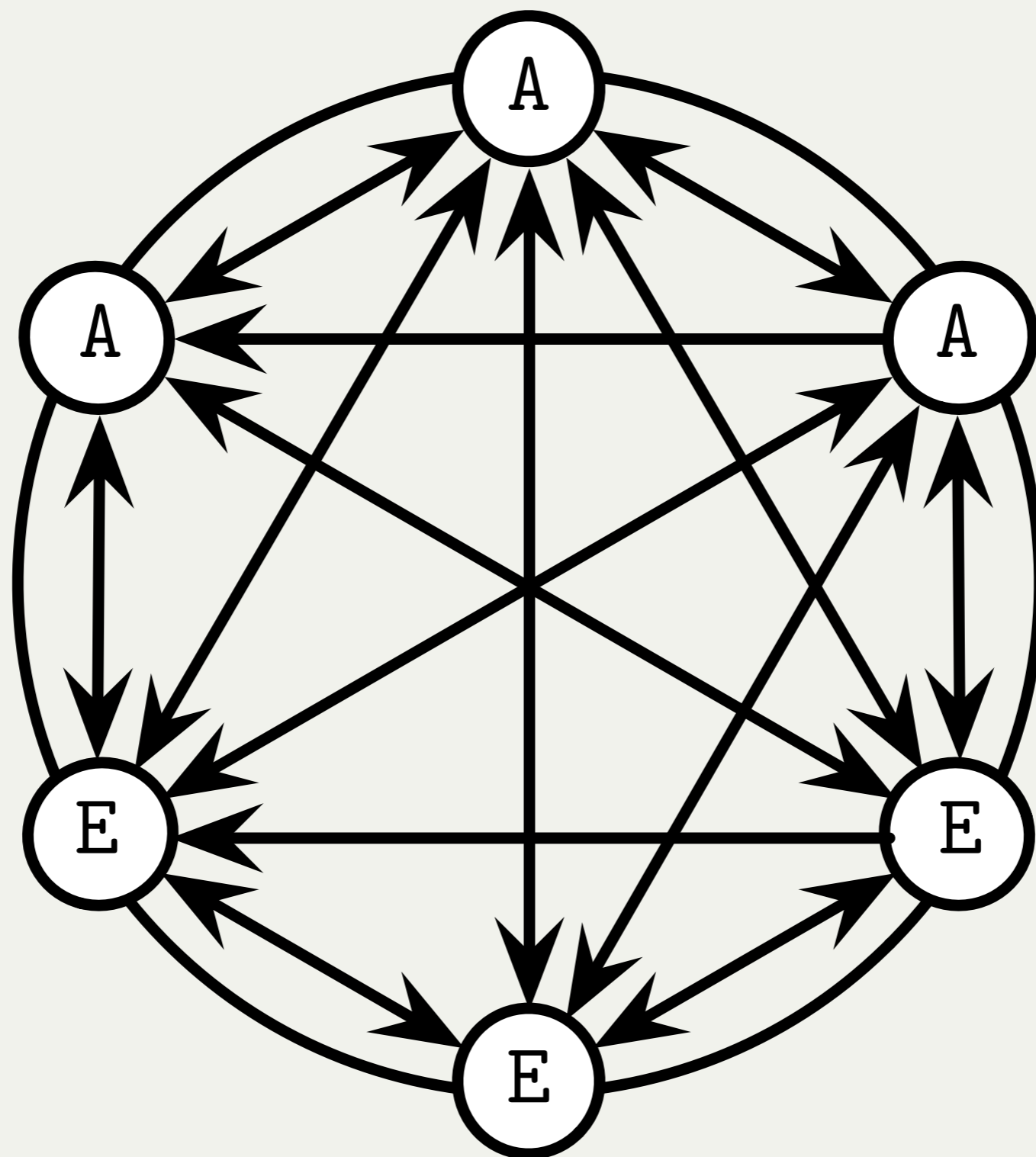




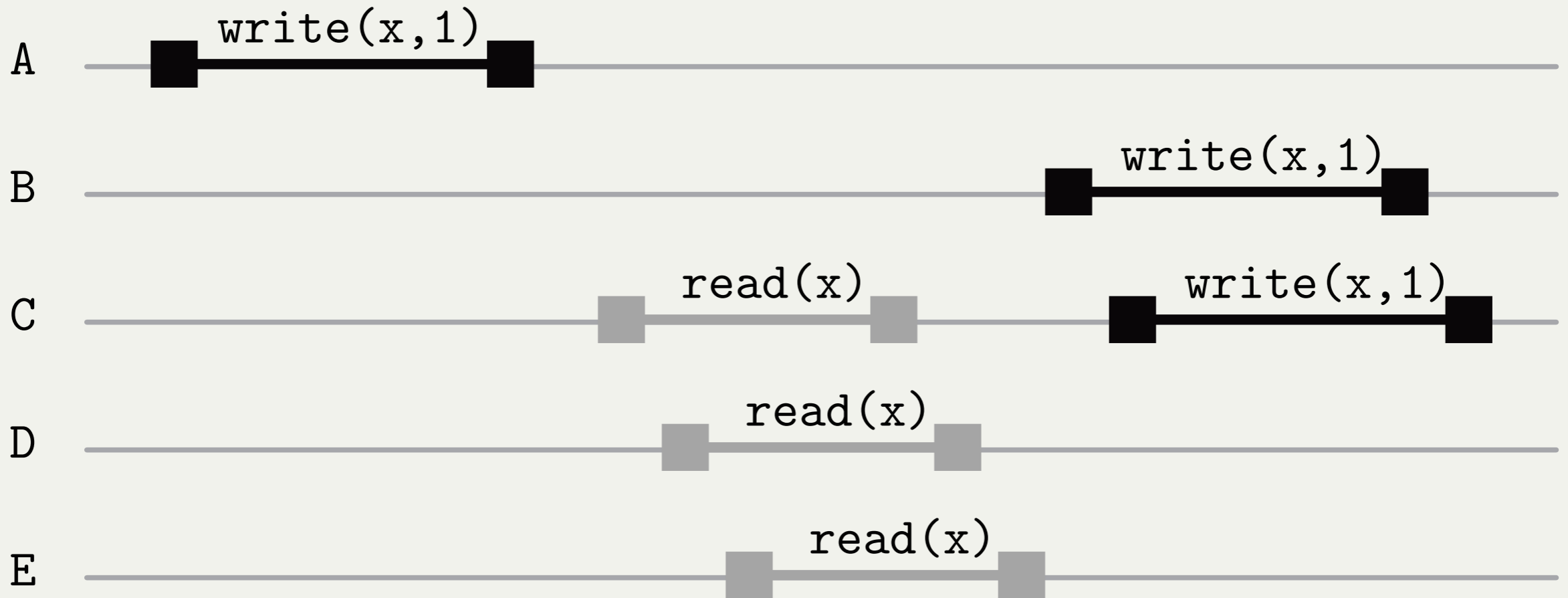




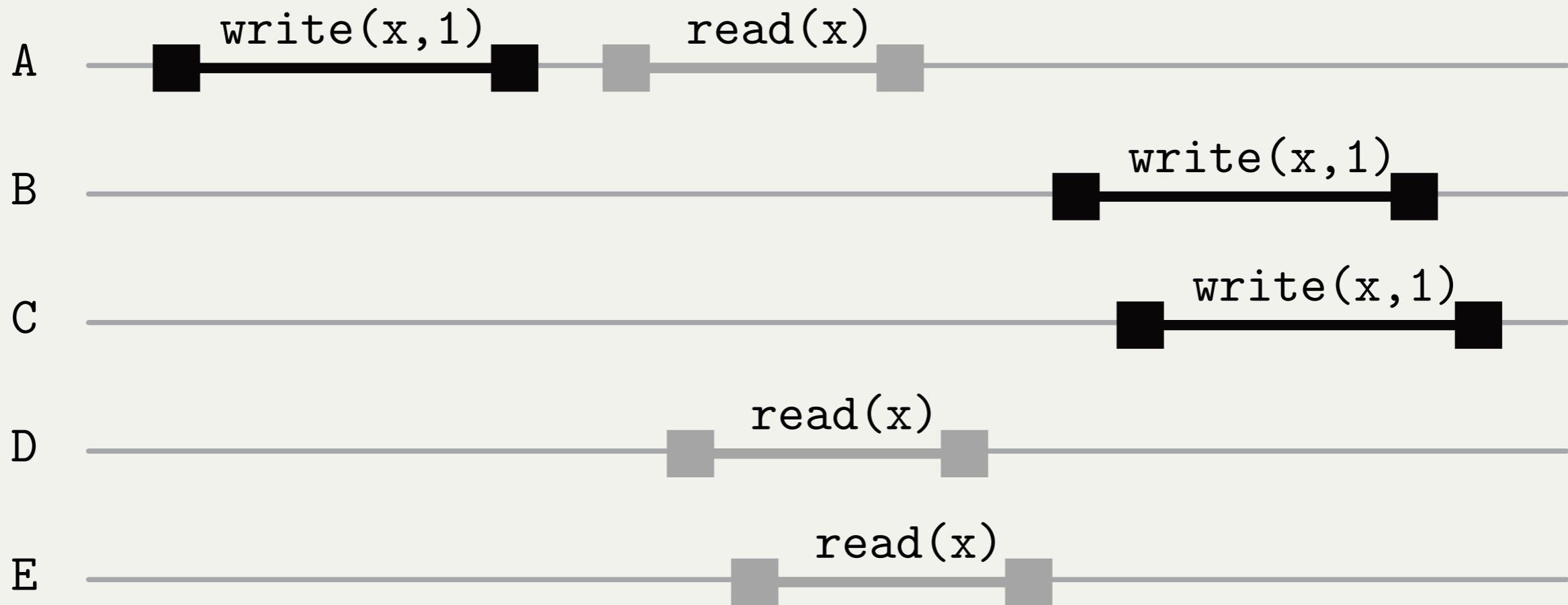
# Anti-Entropy



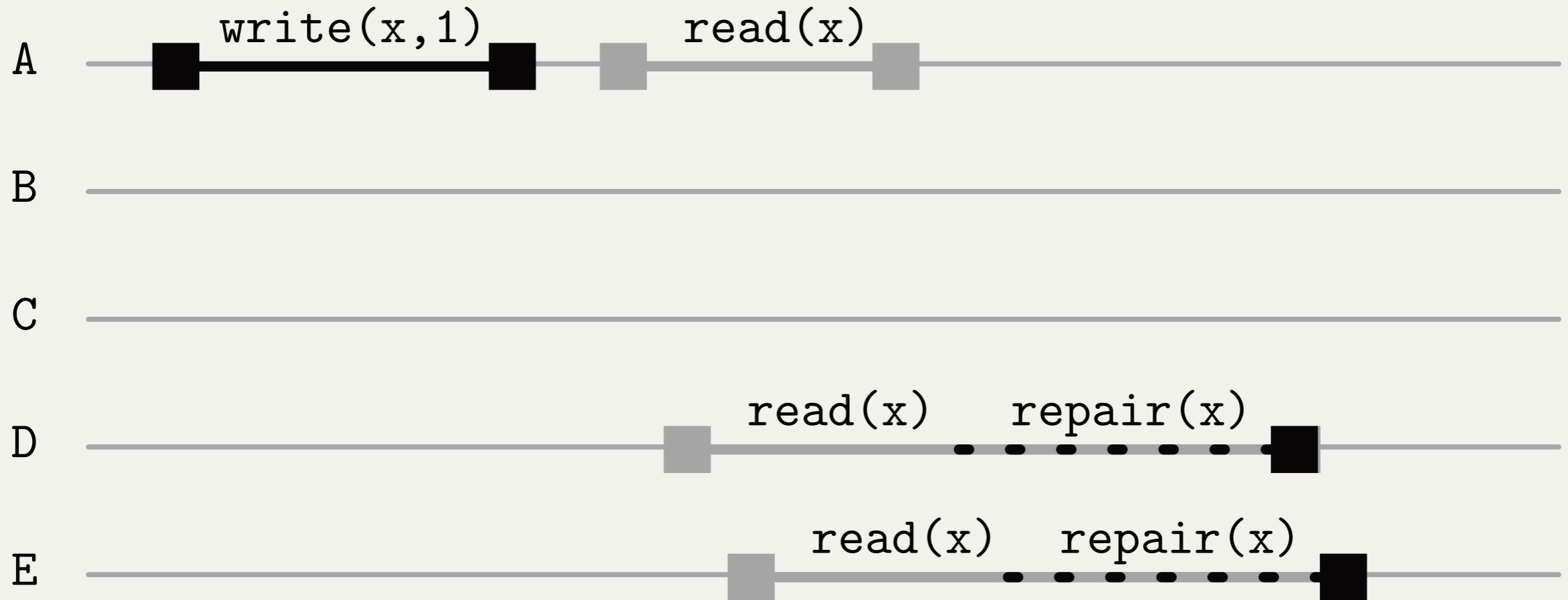
`read(x) → ∅`



# read(x) → 1



# Read Monotonicity



# Tunable Consistency

- N: the number of nodes that store replicas of the data
- W: the number of replicas that need to acknowledge the receipt of the update before the update completes
- R: the number of replicas that are contacted when a data object is accessed through a read operation
  - Werner Vogels. 2008. Eventually Consistent.

$$W + R > N$$

# Replication

$$W + R > N$$



# Voting

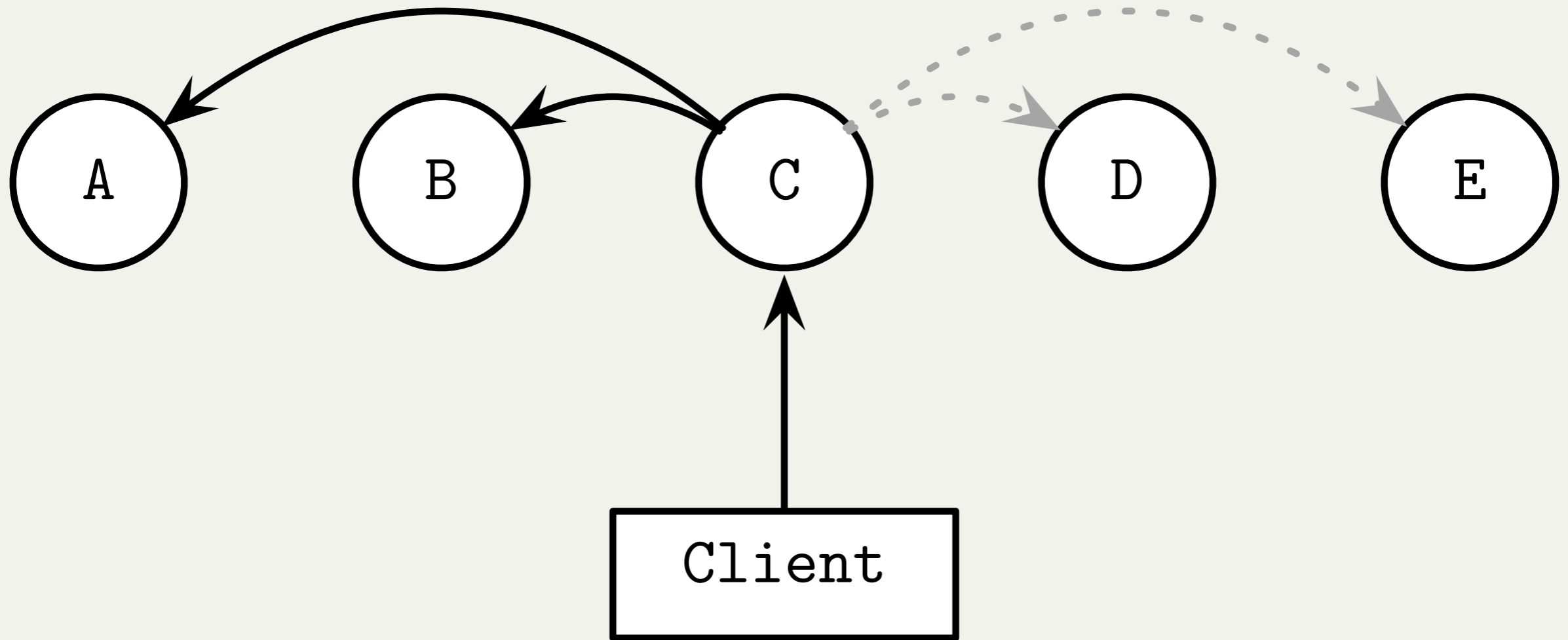
$$W + R > N$$

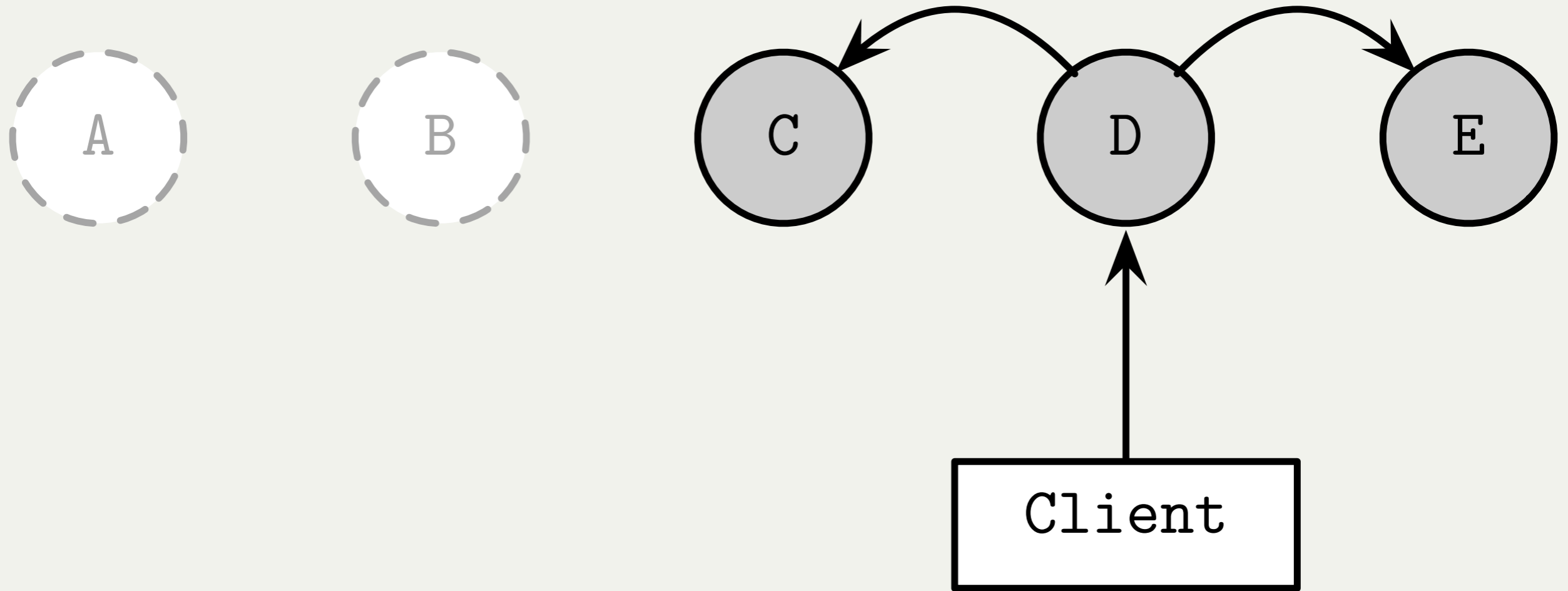
# Quorum

$$W = R = \lfloor N/2 \rfloor + 1$$

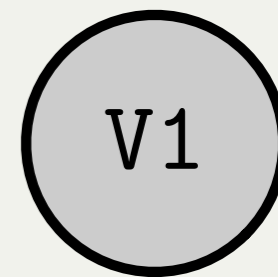
# Fault Tolerance

$$N = 2F + 1$$

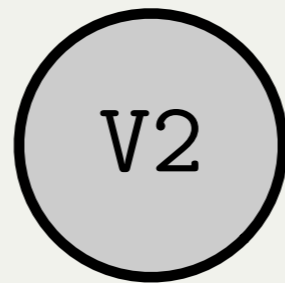
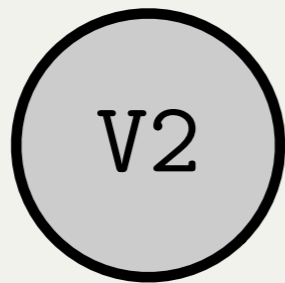
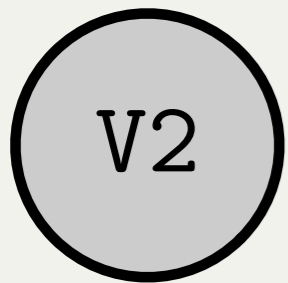




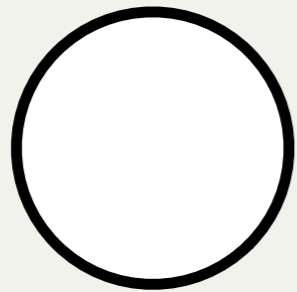
# Cost of maintaining a Quorum



# Cost of a Quorum

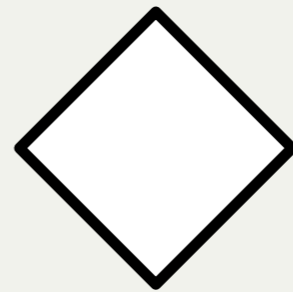


# Witness Replicas



Copy

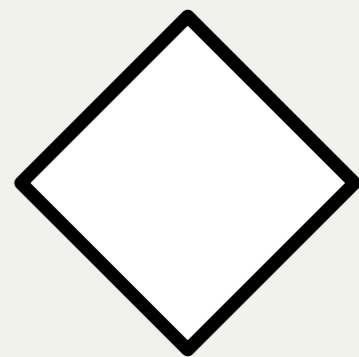
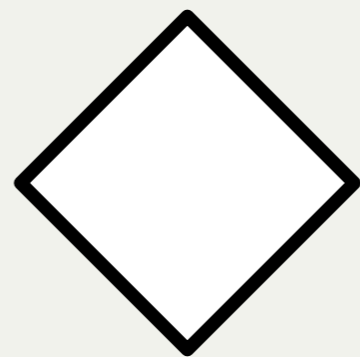
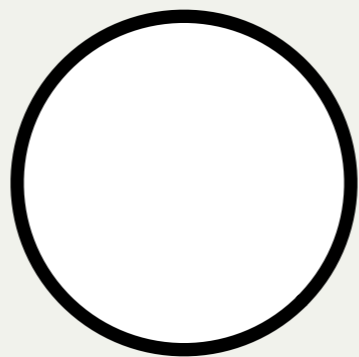
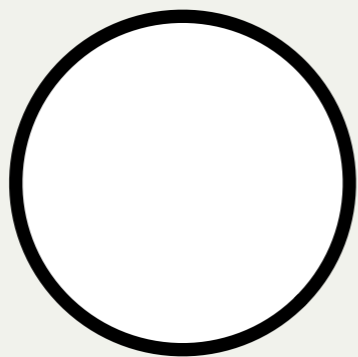
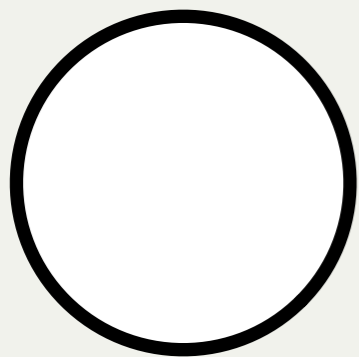
Version ID  
+ Data Record

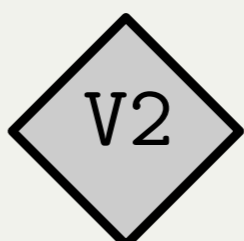
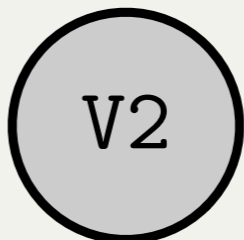
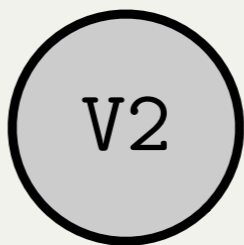


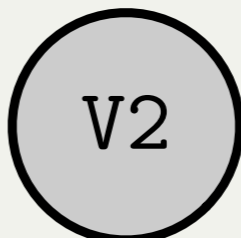
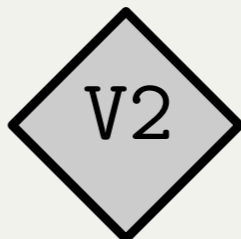
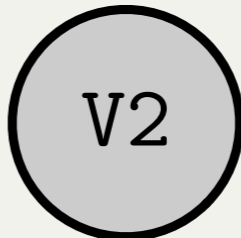
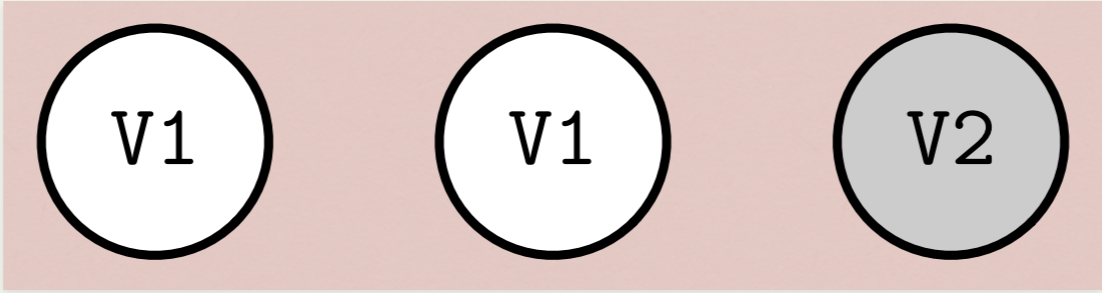
Witness

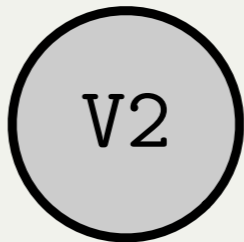
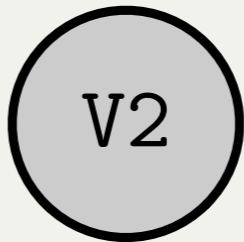
Version ID only

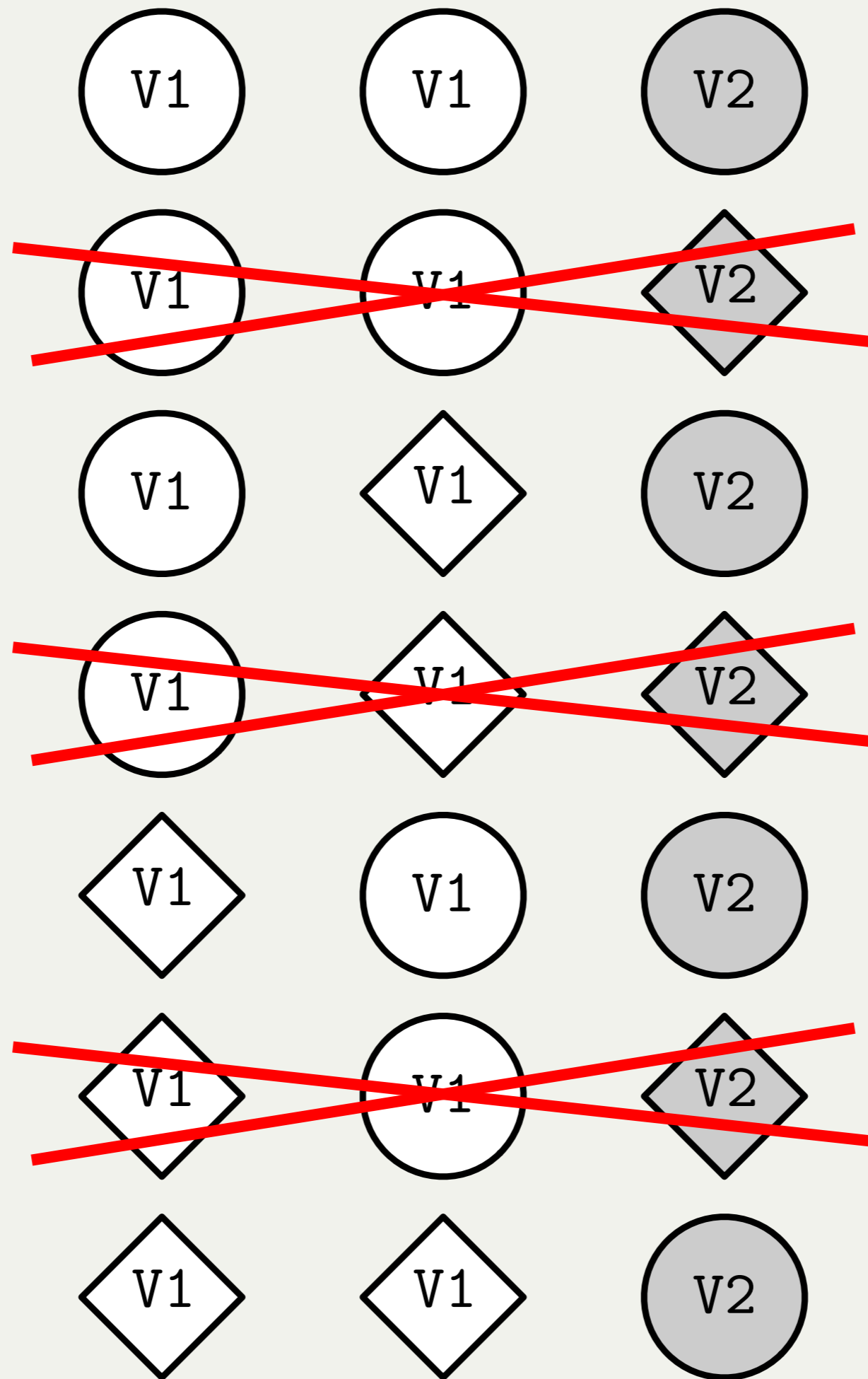


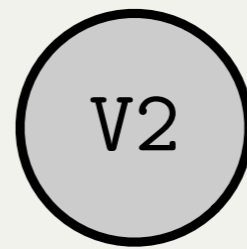
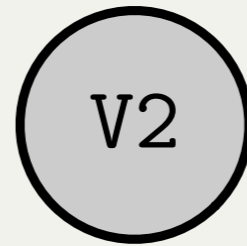
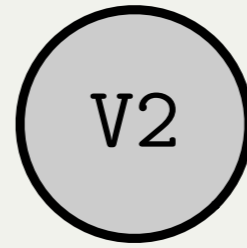
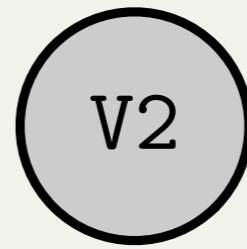




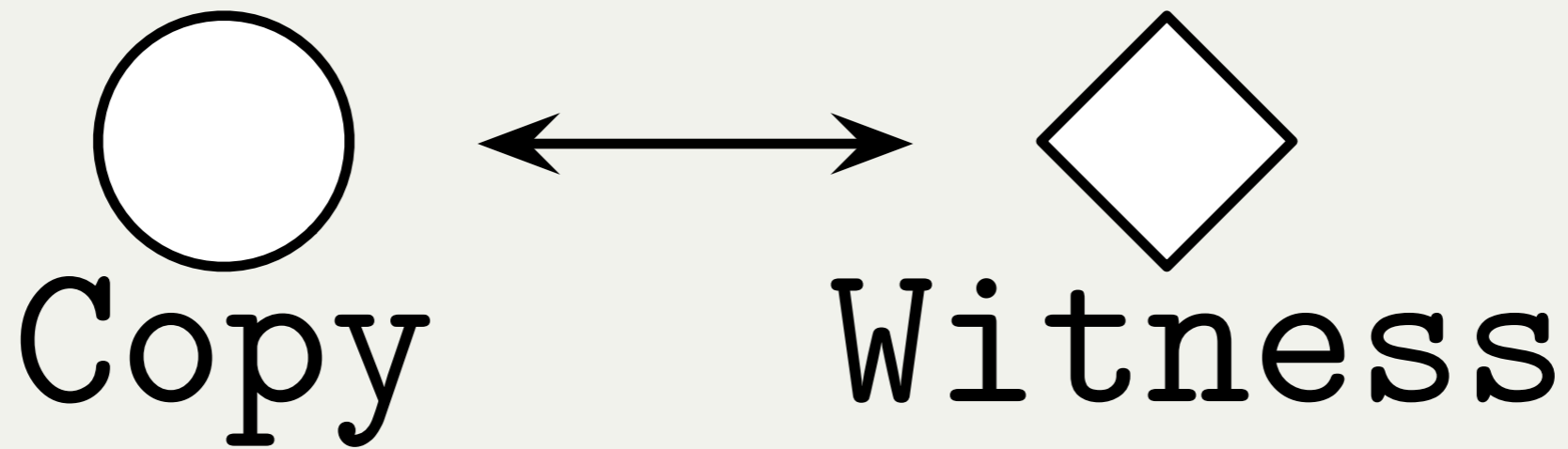


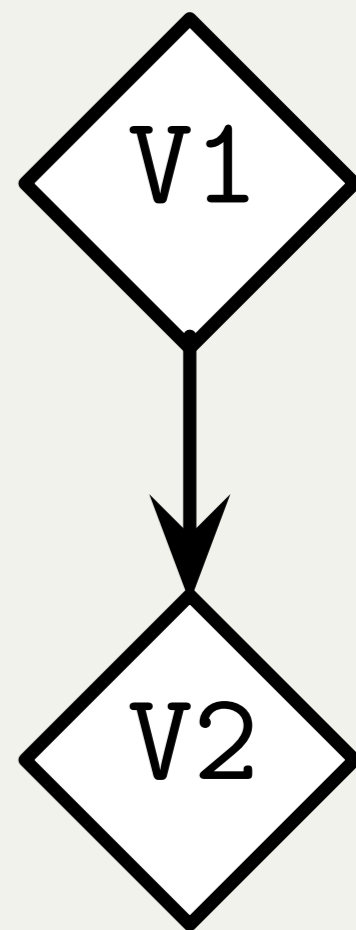
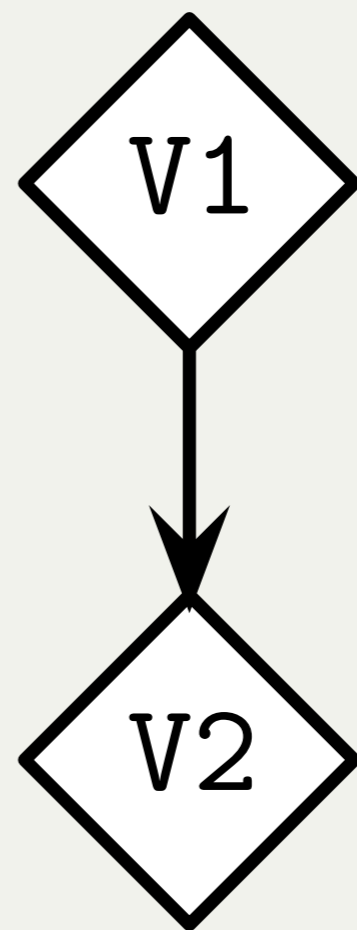
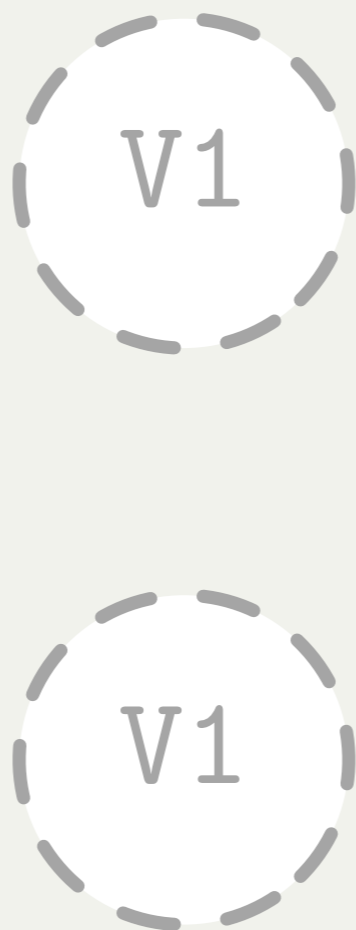
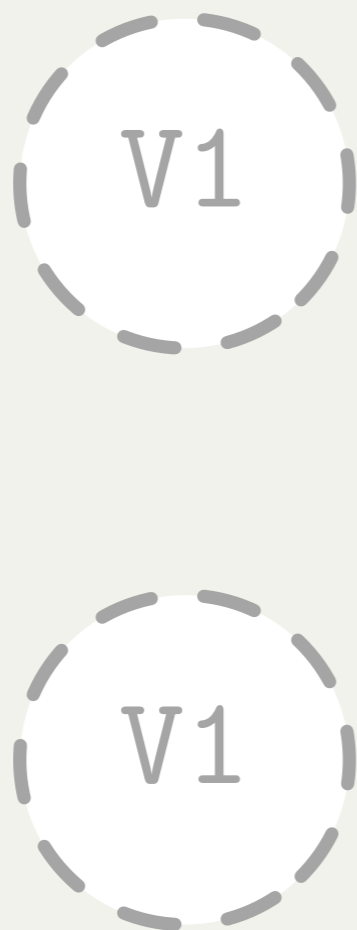
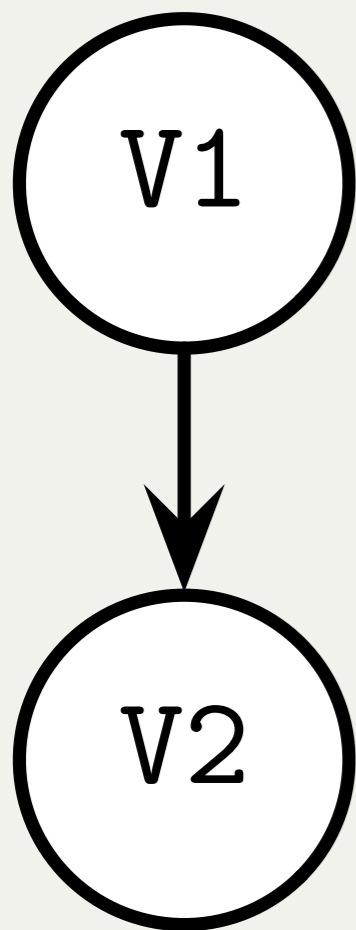




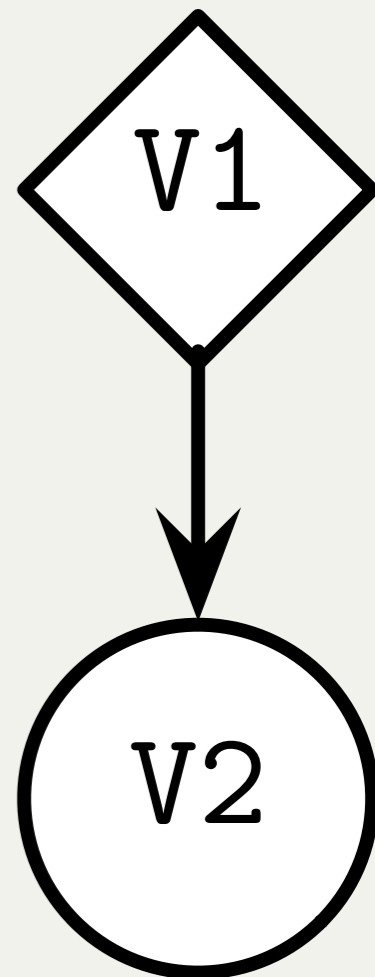
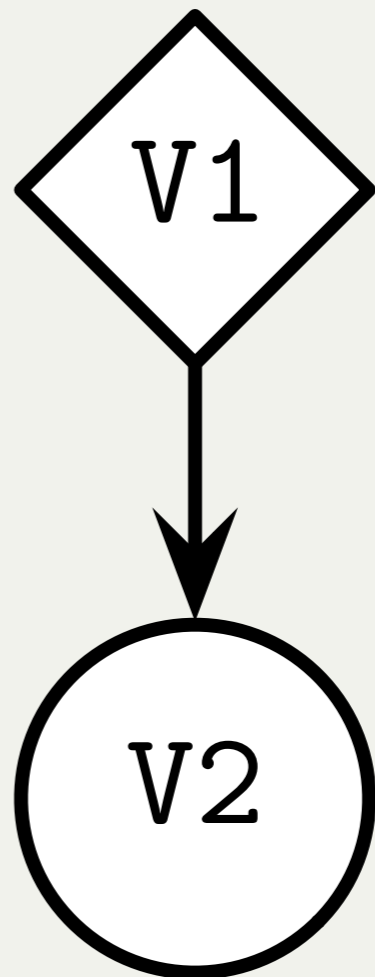
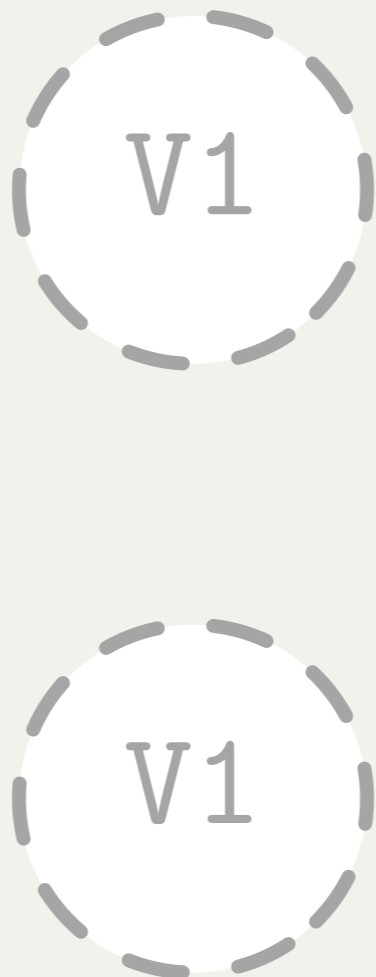
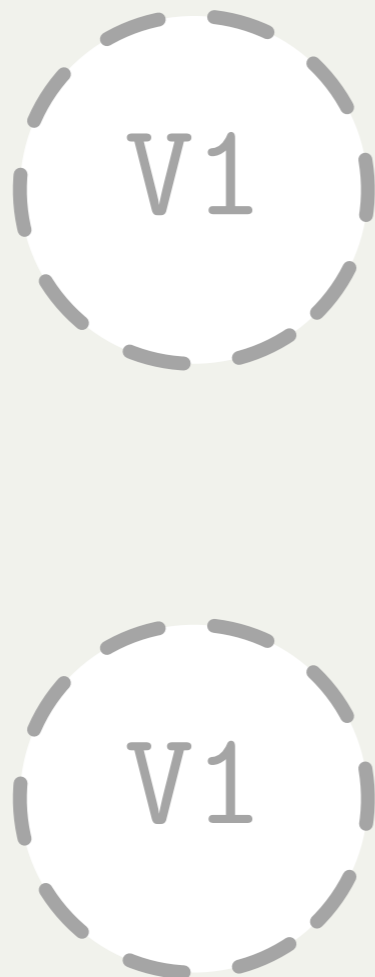
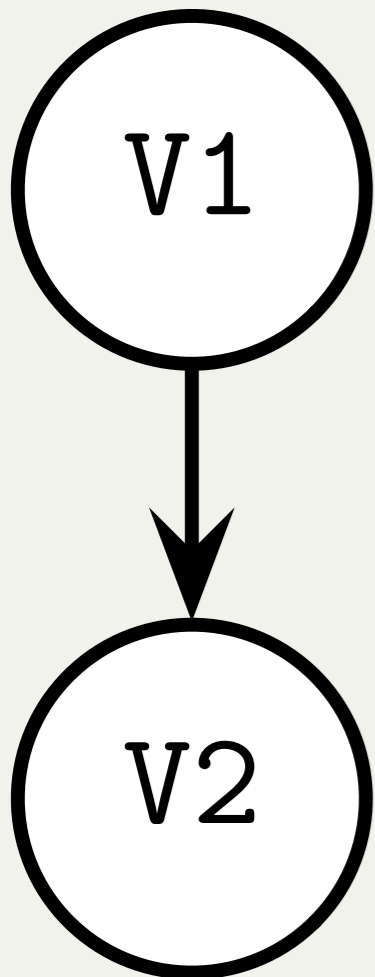


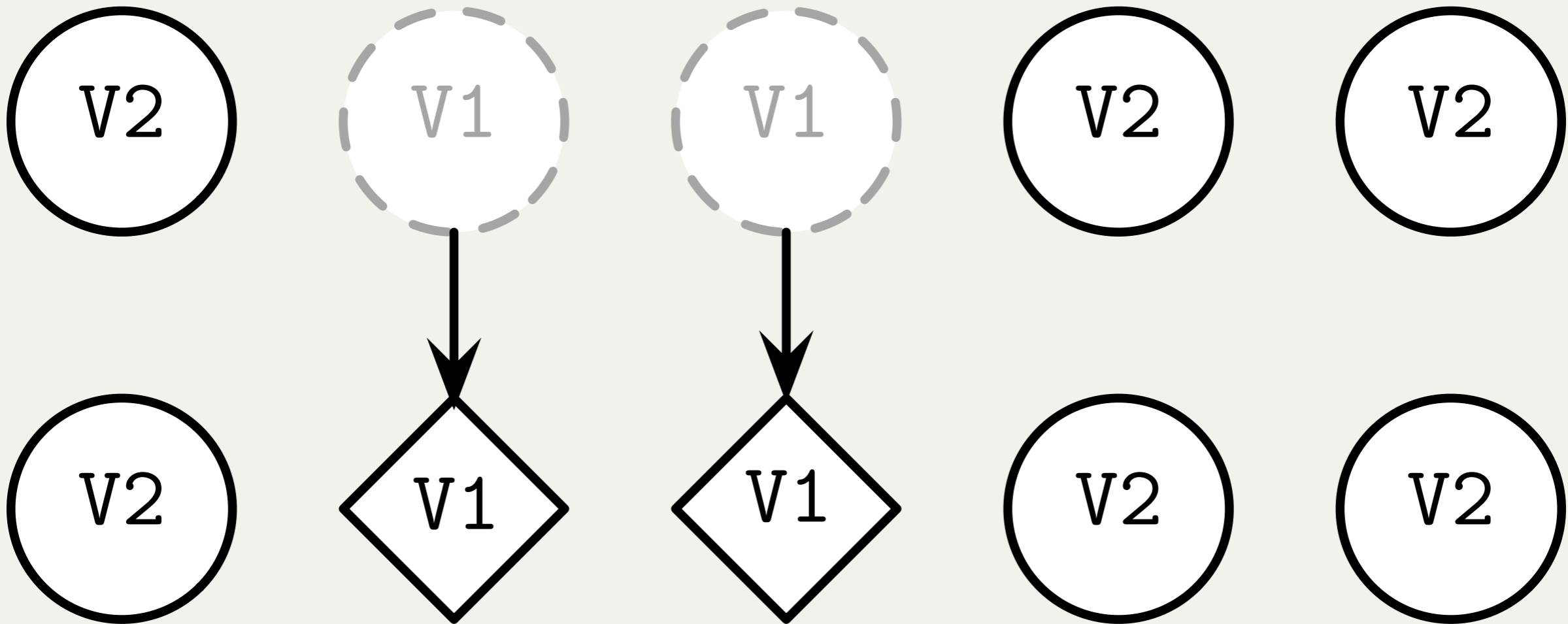
# Upgradable Witnesses





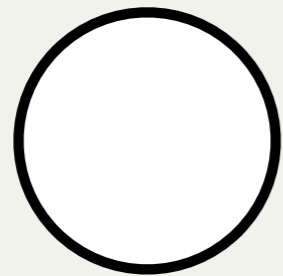




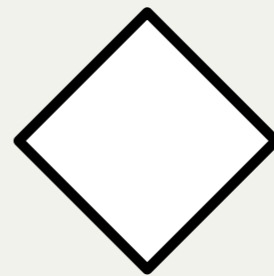


Cost of topology change

# Transient Replication

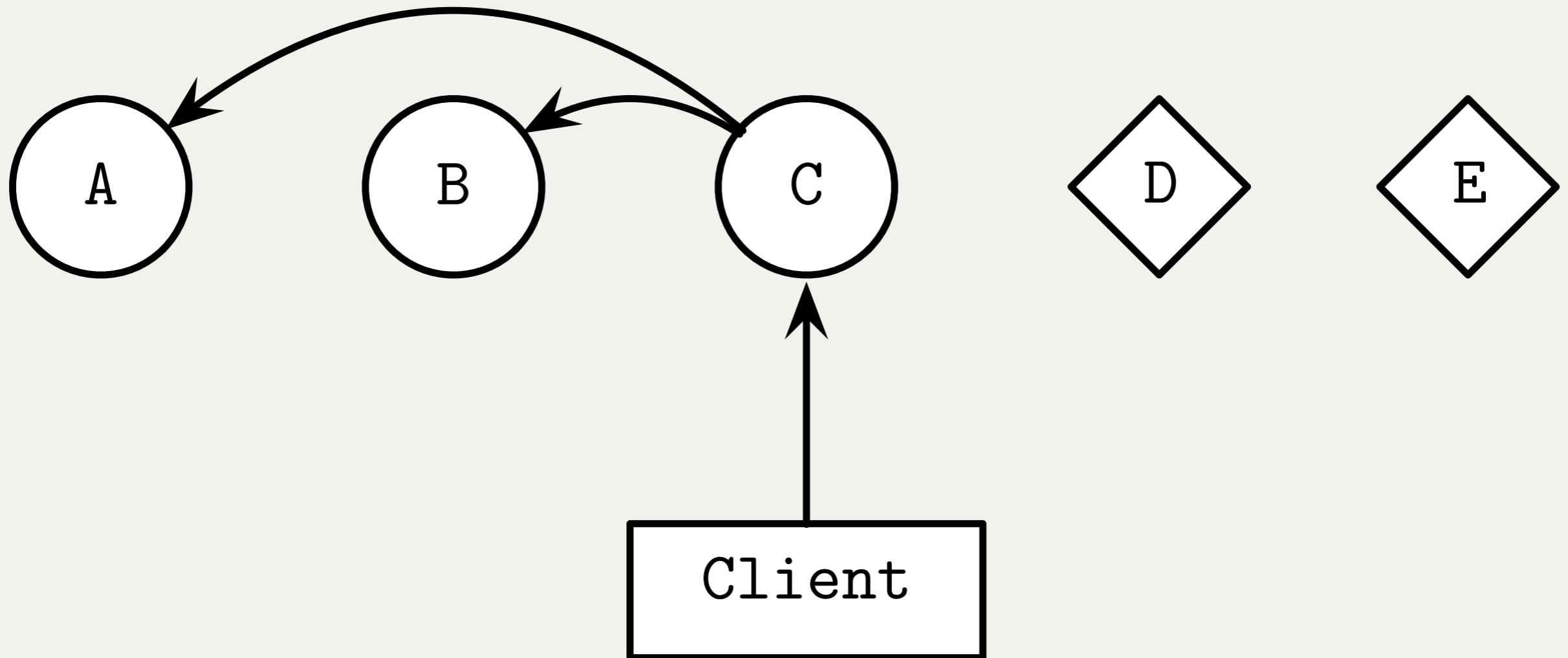


Full

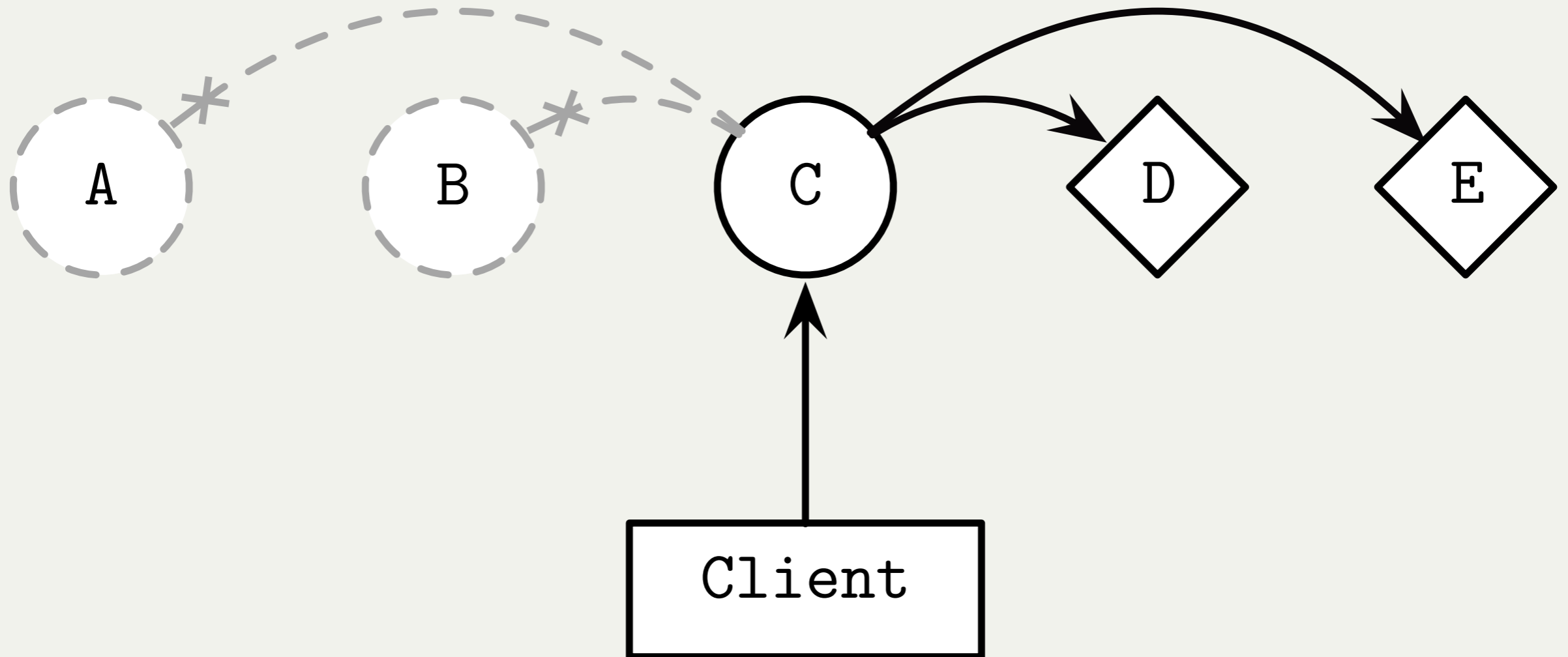


Transient

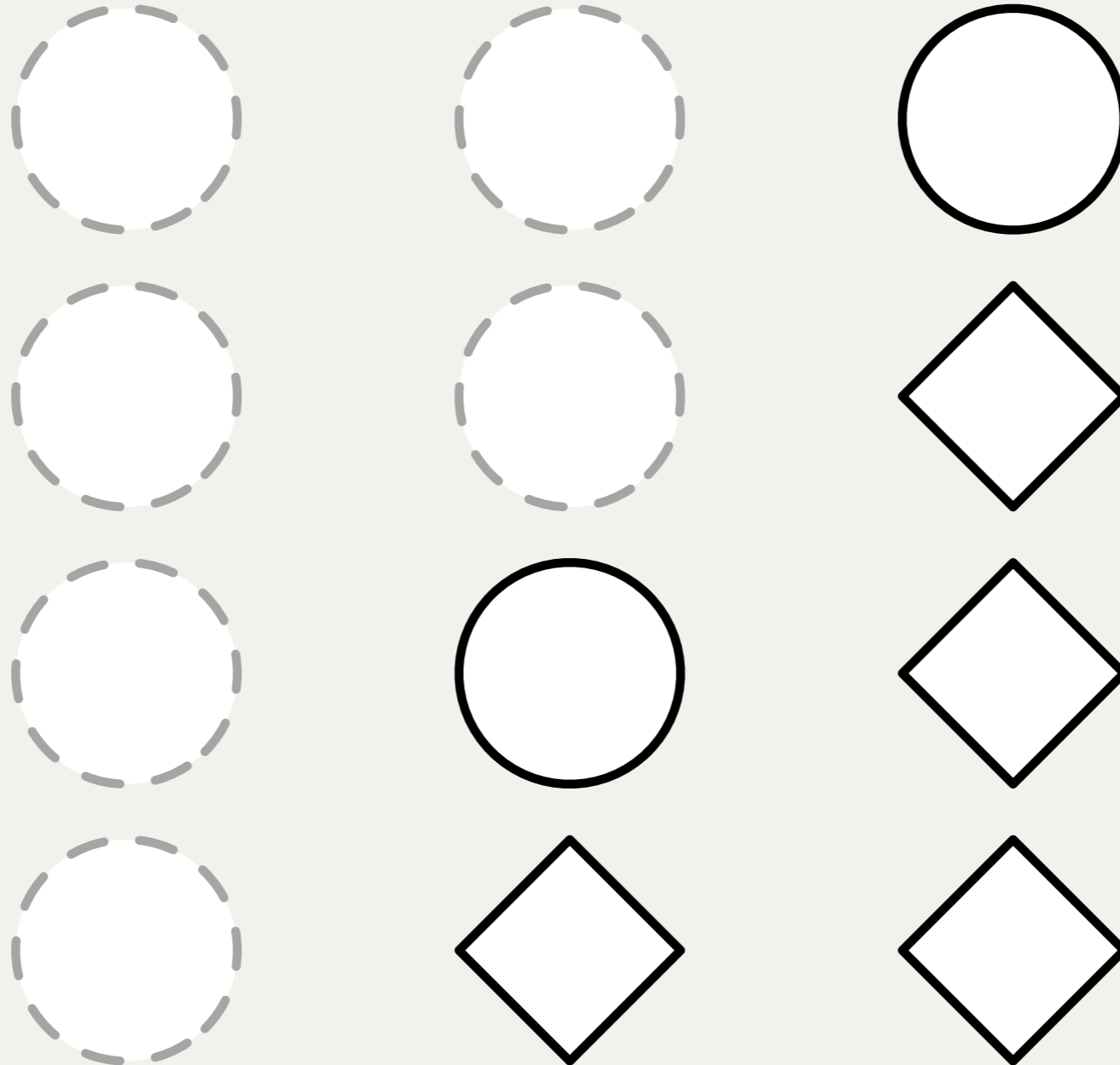
# Cheap Quorums: Write



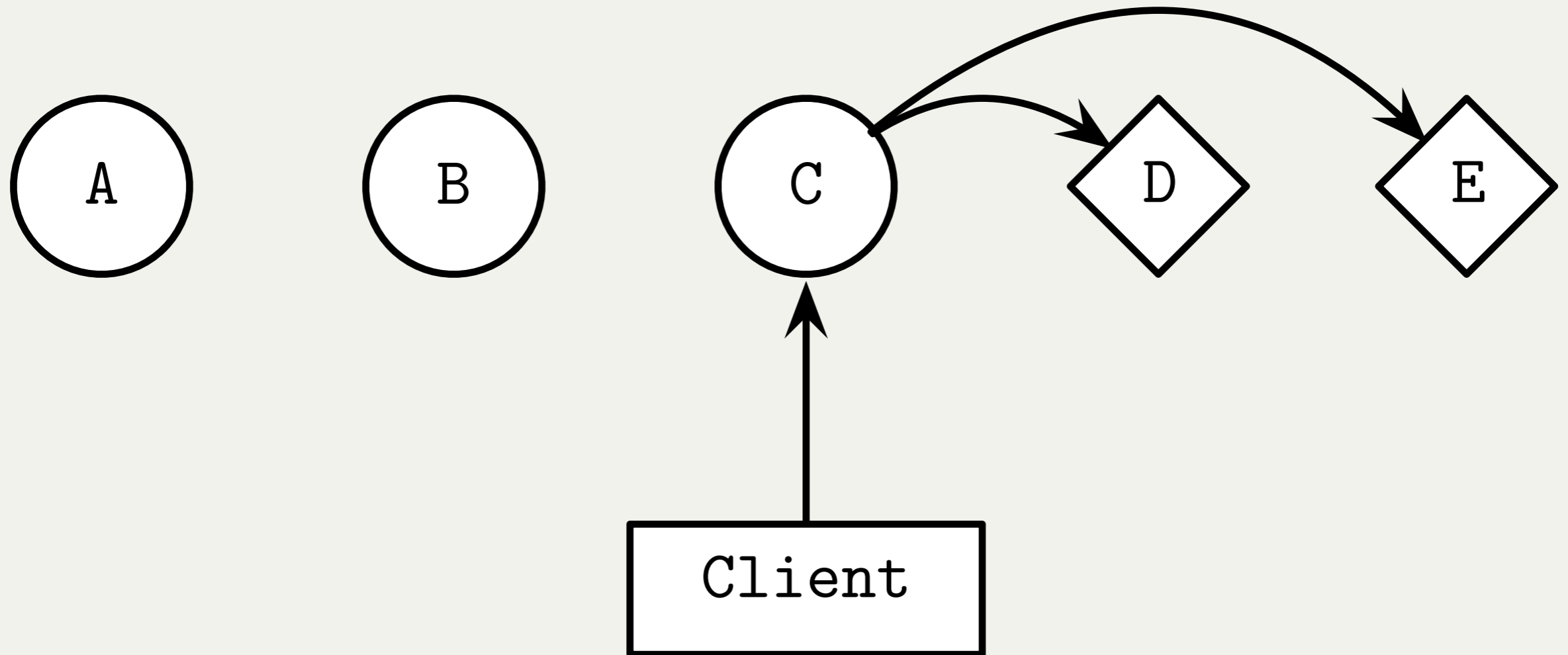
# Cheap Quorums: Write



# Cheap Quorums: Read

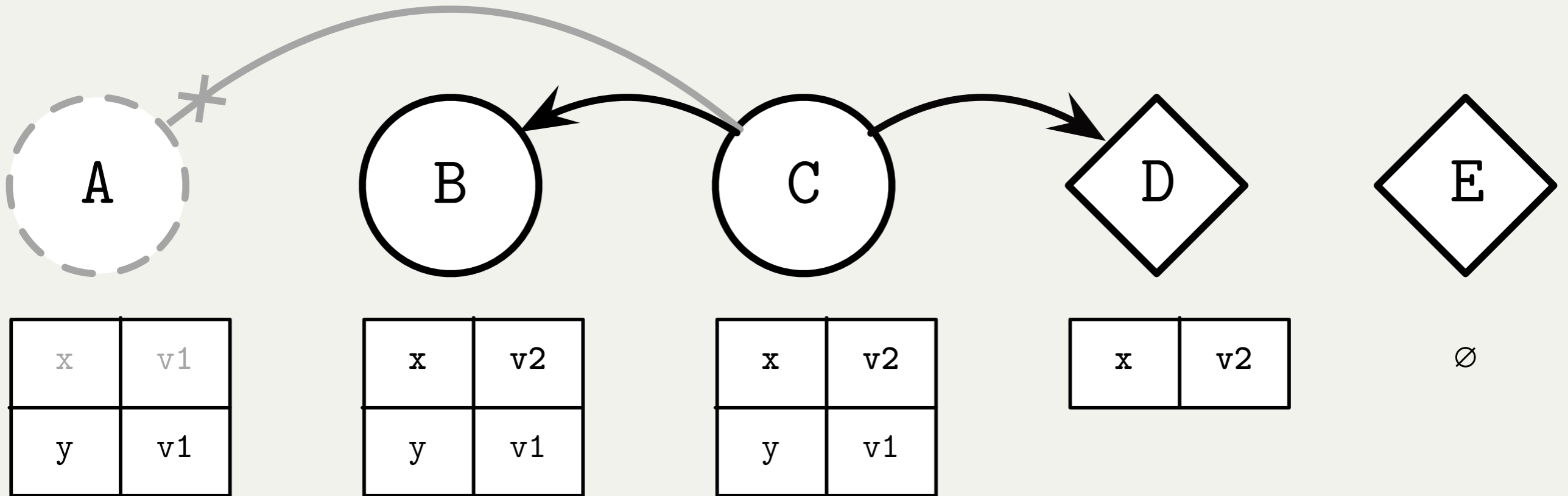


# Cheap Quorums: Read

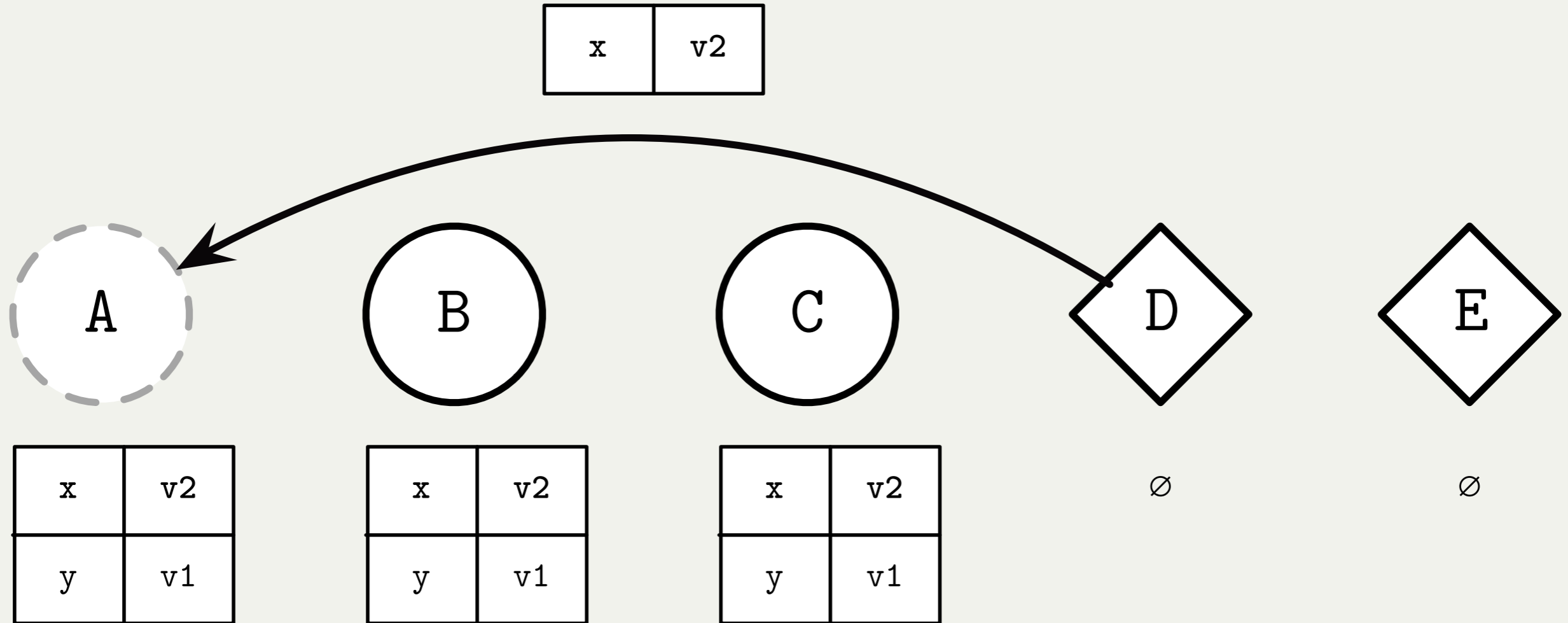




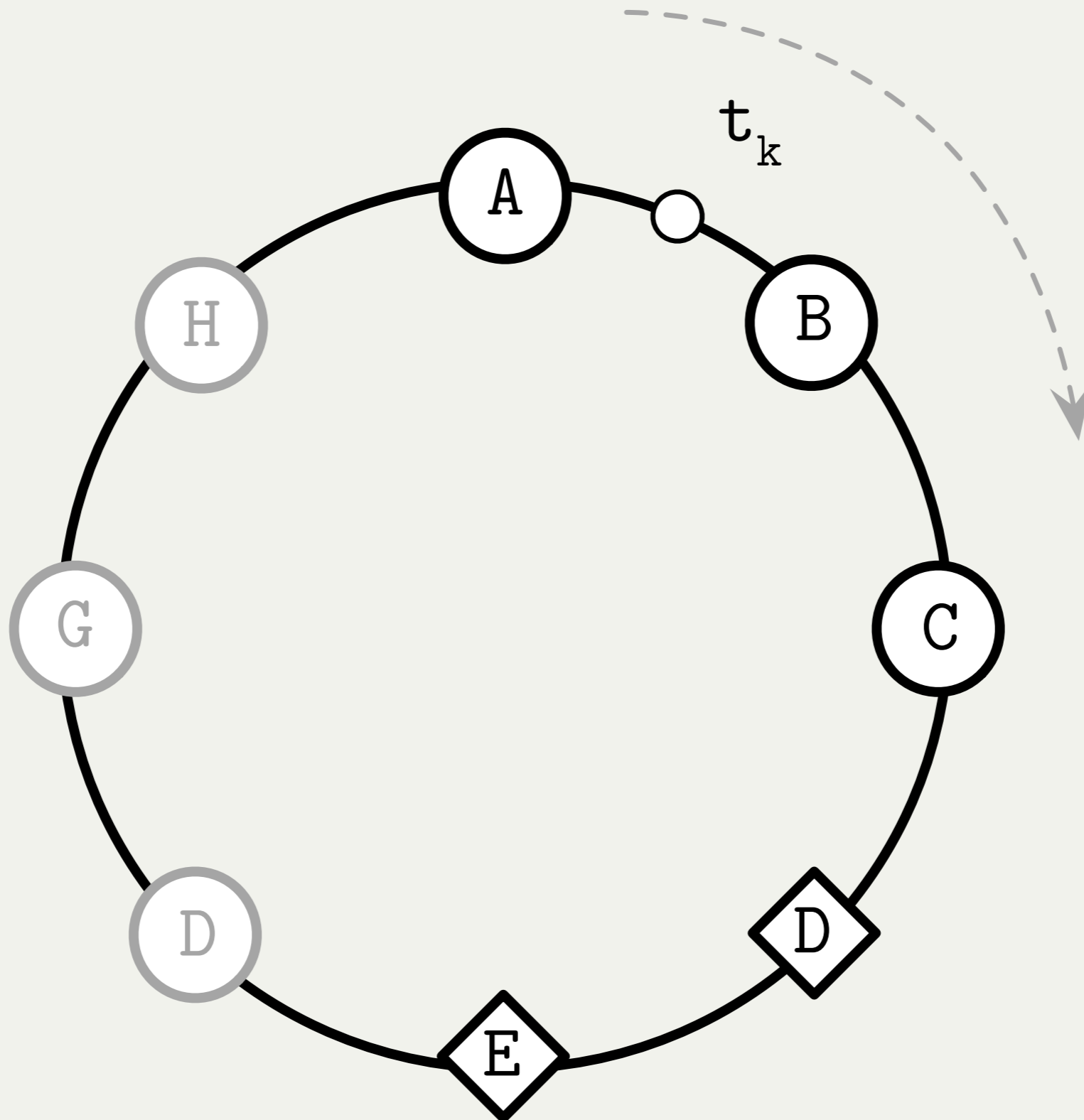
# Repair



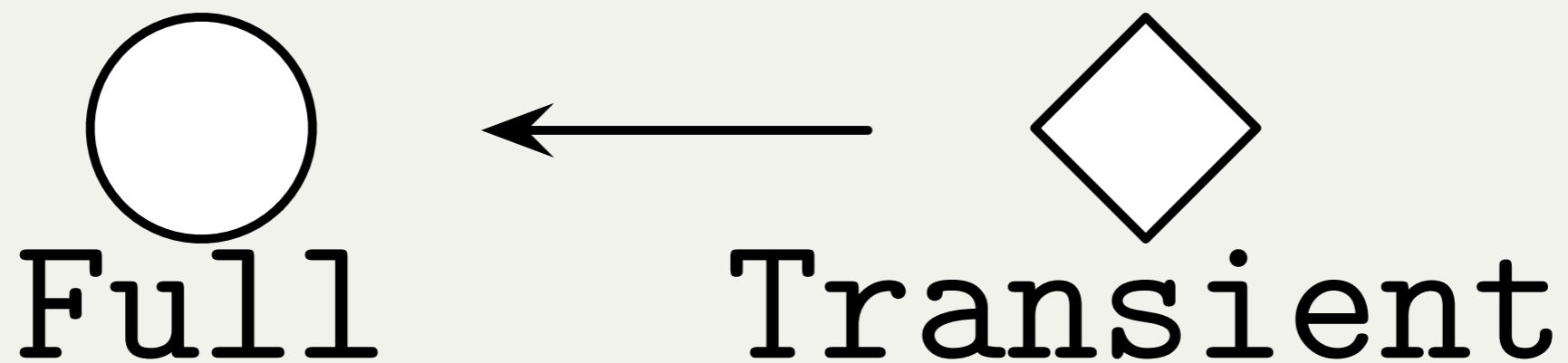
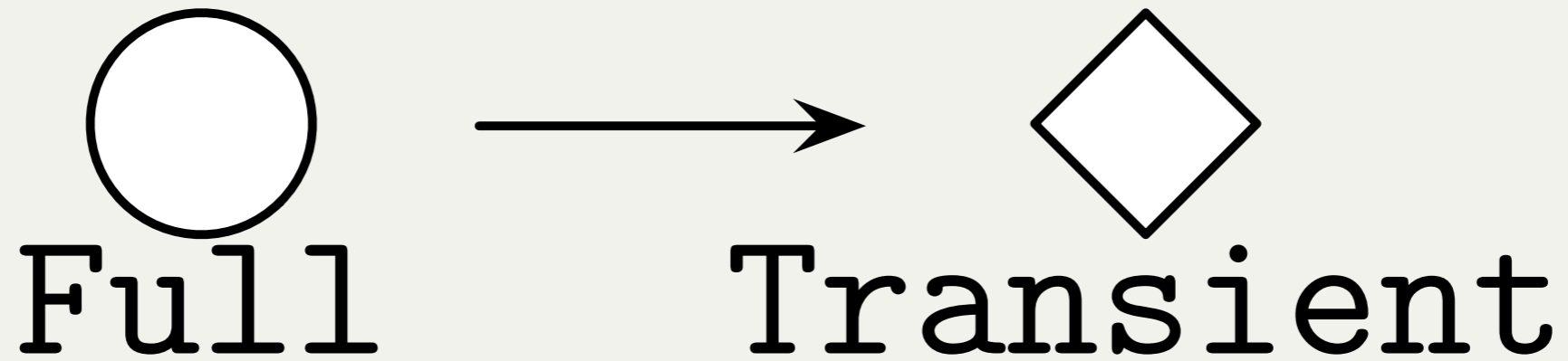
# Repair



# Consistent Hashing



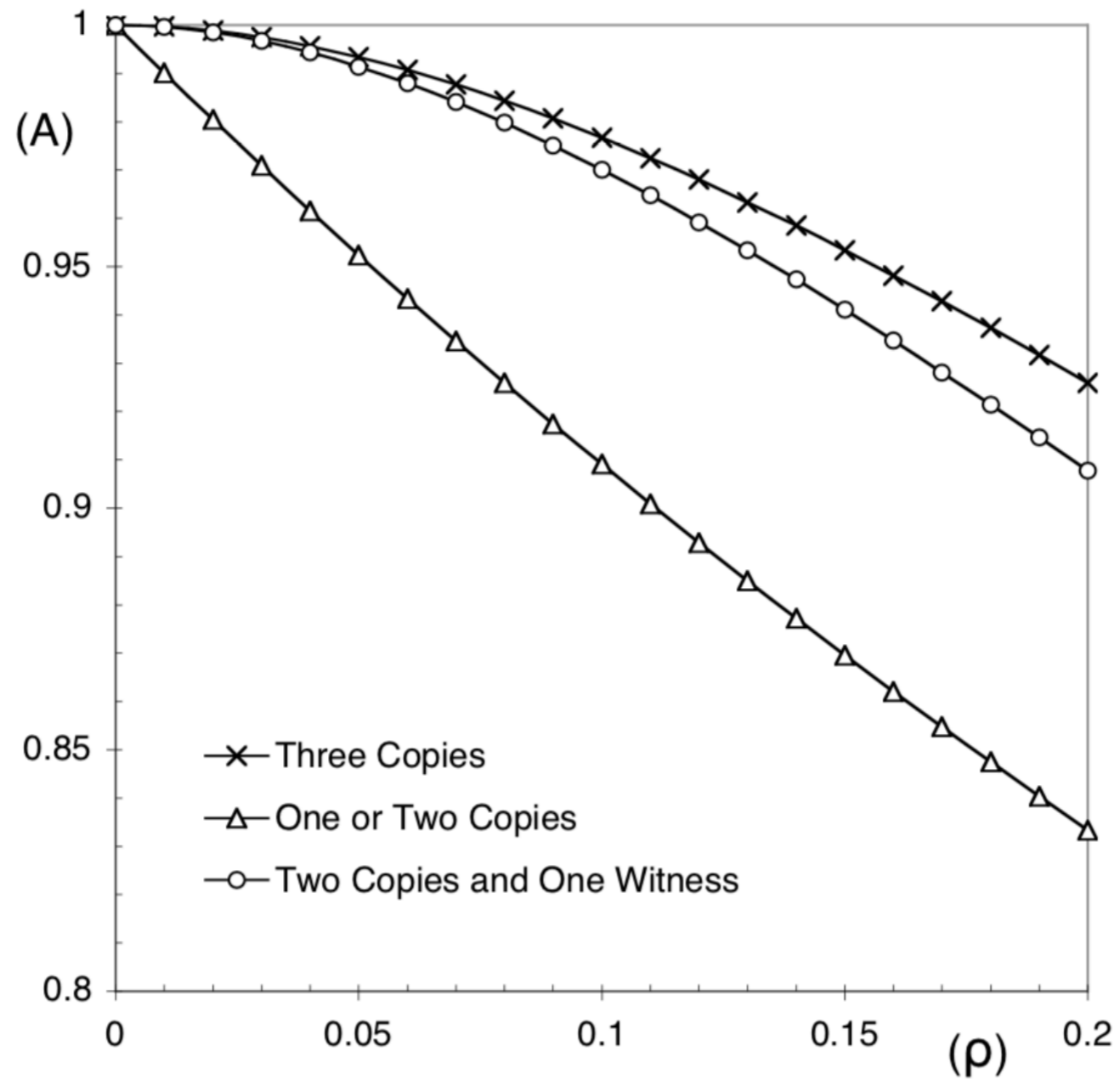
# Ring



# Availability

$$A(n_f + m_t) \approx A(n_f + m_f)$$

# Availability



# What TR is not

- Sloppy Quorums
- Hinted Handoff
- A way to reduce a number of replicas
- A way to reduce a quorum size

# Summary

- No decrease in availability
- No impact on durability
- Up to 50% storage savings
- Lower message overhead
- Smaller write quorum
- Less overhead for reads



## Transient Replication and Cheap Quorums

[Browse files](#)

Patch by Blake Eggleston, Benedict Elliott Smith, Marcus Eriksson, Alex Petrov, Ariel Weisberg; Reviewed by Blake Eggleston, Marcus Eriksson, Benedict Elliott Smith, Alex Petrov, Ariel Weisberg for CASSANDRA-14404


Co-authored-by: Blake Eggleston <bdeggleston@gmail.com>

Co-authored-by: Benedict Elliott Smith <benedict@apache.org>

Co-authored-by: Marcus Eriksson <marcuse@apache.org>

Co-authored-by: Alex Petrov <oleksandr.petrov@gmail.com>

---

 trunk (#3)



5 people committed on Jul 6, 2018

1 parent [5b645de](#)

commit [f7431b432875e334170ccdb19934d05545d2cebd](#)

# References

- Gifford D. K. Weighted Voting for Replicated Data. 1979.
- J.-F. Pâris. Voting with Witnesses: A Consistency Scheme for Replicated Files. 1986.
- Divyakant Agrawal and Amr El Abbadi. Reducing Storage for Quorum Consensus Algorithms. 1988.



O'REILLY®

# Database Internals

A Deep-Dive into How Distributed Data Systems Work



**Early  
Release**

RAW &  
UNEDITED

Alex Petrov

**@ifesdjeen**

**Images attributed to their respective owners.**