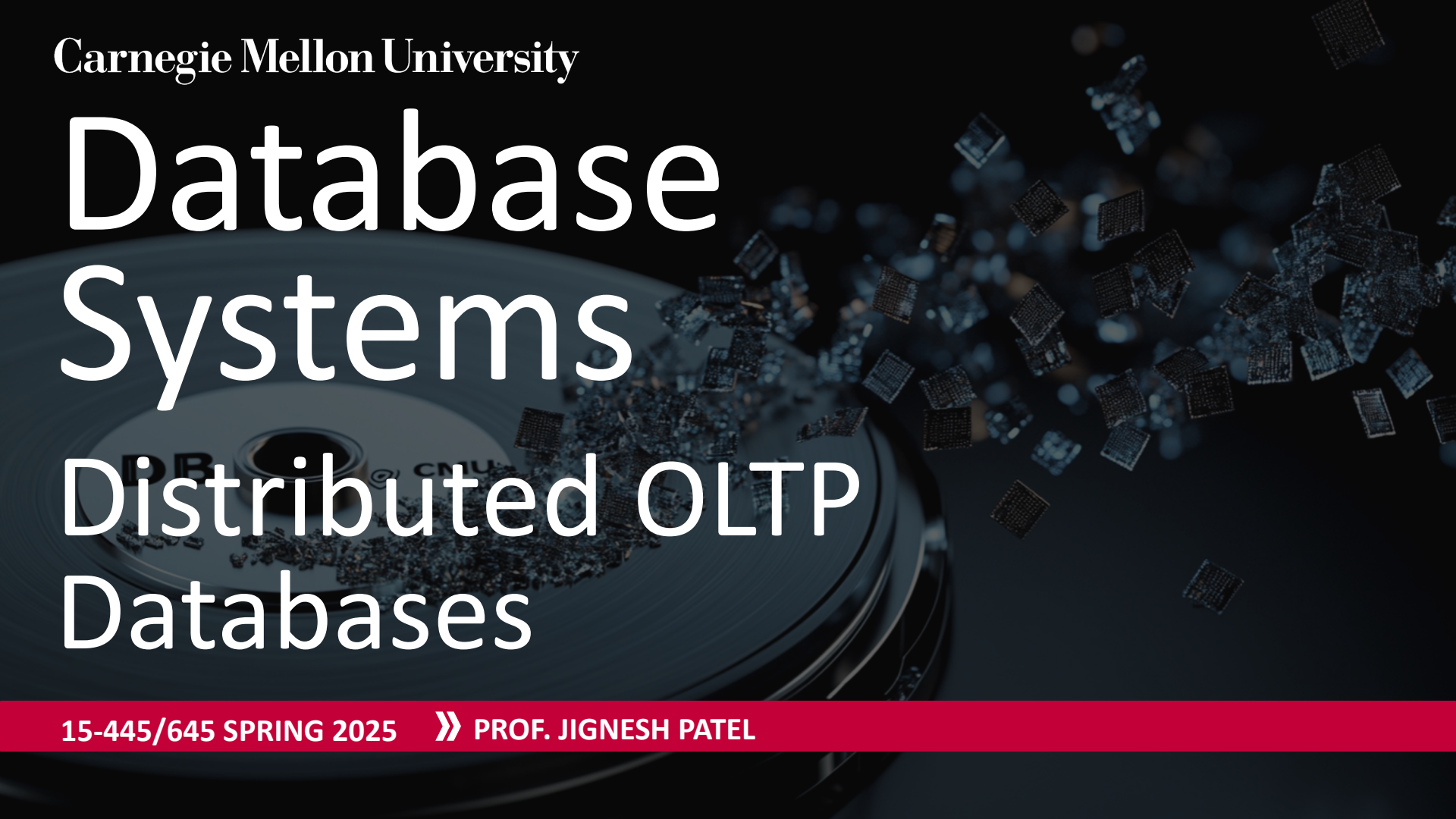
Carnegie Mellon University

Database Systems

Distributed OLTP Databases



15-445/645 SPRING 2025 » PROF. JIGNESH PATEL

ADMINISTRIVIA

Project #4 is due Sunday April 20th @ 11:59pm
→ Recitation: Friday, April 11th in GHC 4303 from 3:00 - 4:00 PM

HW6 is due Sunday, April 20, 2025 @ 11:59pm

Final Exam is on Monday, April 28, 2025, from 05:30pm - 08:30pm
→ Early exam will not be offered. Do not make travel plans.

This course is recruiting TAs for the next semester
→ Apply at: https://www.ugrad.cs.cmu.edu/ta/F25/

ADMINISTRIVIA

Class on Monday, April 21: **Review Session**

→ Come to class prepared with your questions. What material do you want me to go over again?

Class on Wednesday, April 23: **Guest Lecture**

→ Real-world applications of Gen AI and Databases

→ Speaker: Sailesh Krishnamurthy, Google

UPCOMING DATABASE TALKS

MariaDB (DB Seminar)

→ Monday, April 14 @ 4:30pm
→ MariaDB's New Query Optimizer
→ Speaker: Michael Widenius
→ https://cmu.zoom.us/j/93441451665

Gel (DB Seminar)

→ Monday, April 21 @ 4:30pm
→ EdgeQL with Gel
→ Speaker: Michael Sullivan
→ https://cmu.zoom.us/j/93441451665

LAST CLASS

System Architectures
→ Shared-Everything, Shared-Disk, Shared-Nothing

Partitioning/Sharding
→ Hash, Range, Round Robin

Transaction Coordination
→ Centralized vs. Decentralized

OLTP VS. OLAP

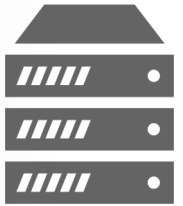
On-line Transaction Processing (OLTP):

→ Short-lived read/write txns.
→ Small footprint.
→ Repetitive operations.

On-line Analytical Processing (OLAP):

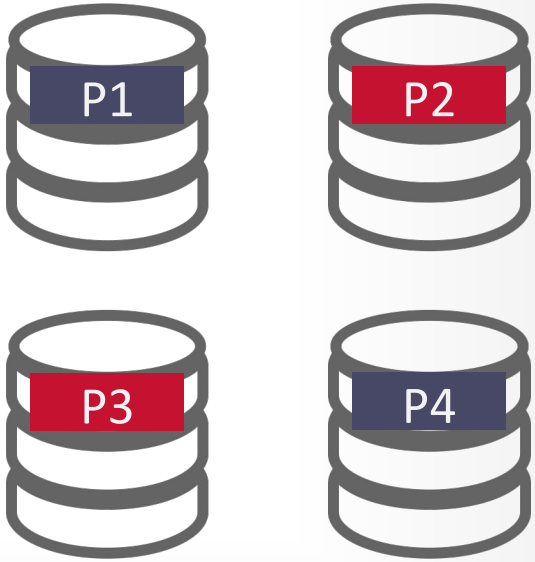
→ Long-running, read-only queries.
→ Complex joins.
→ Exploratory queries.

DECENTRALIZED COORDINATOR



Application Server

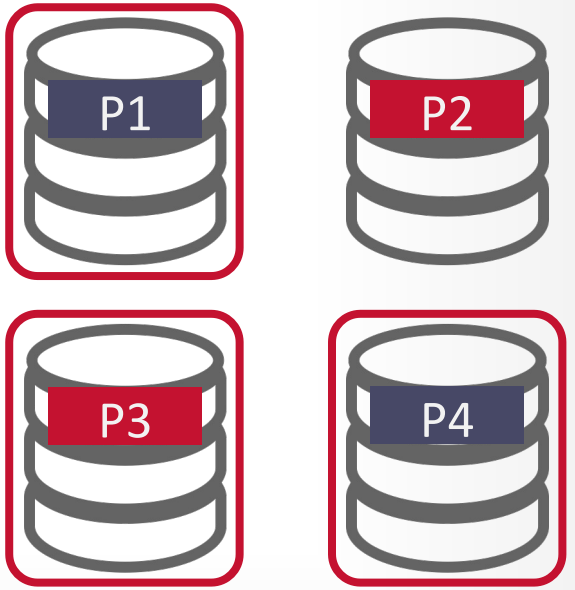
Partitions



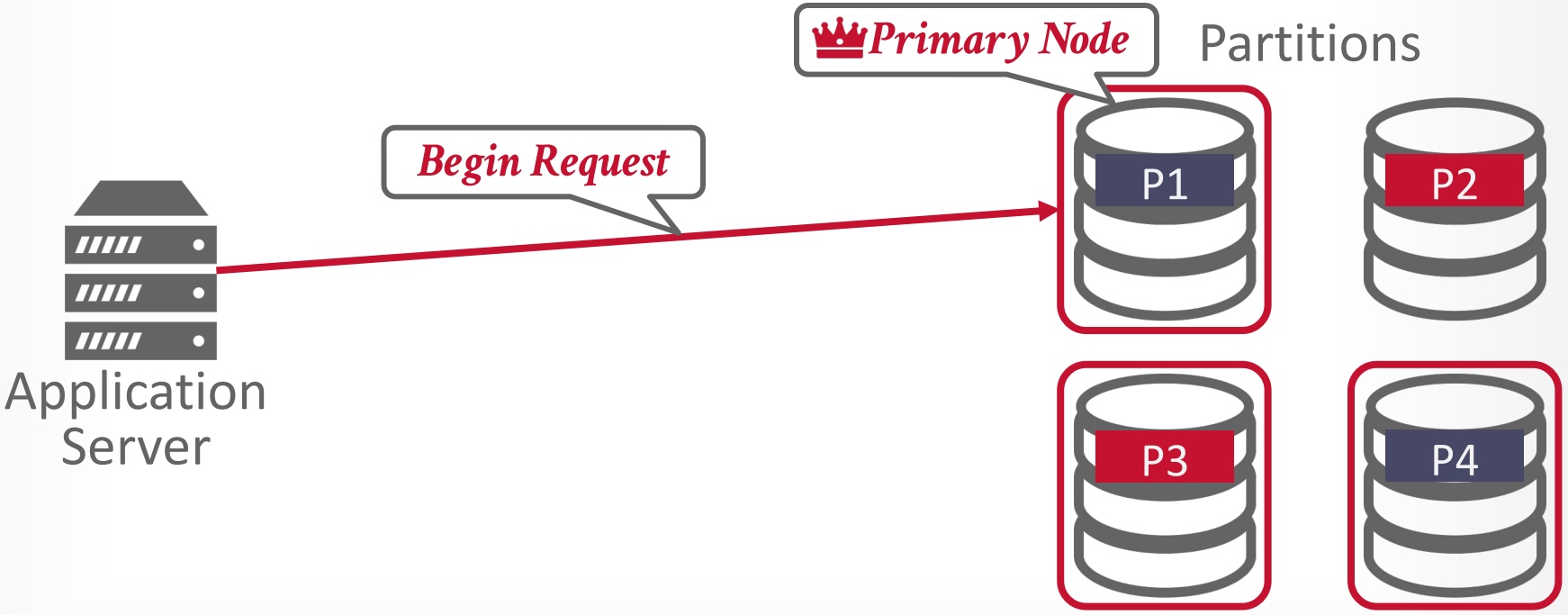
DECENTRALIZED COORDINATOR

Application Server

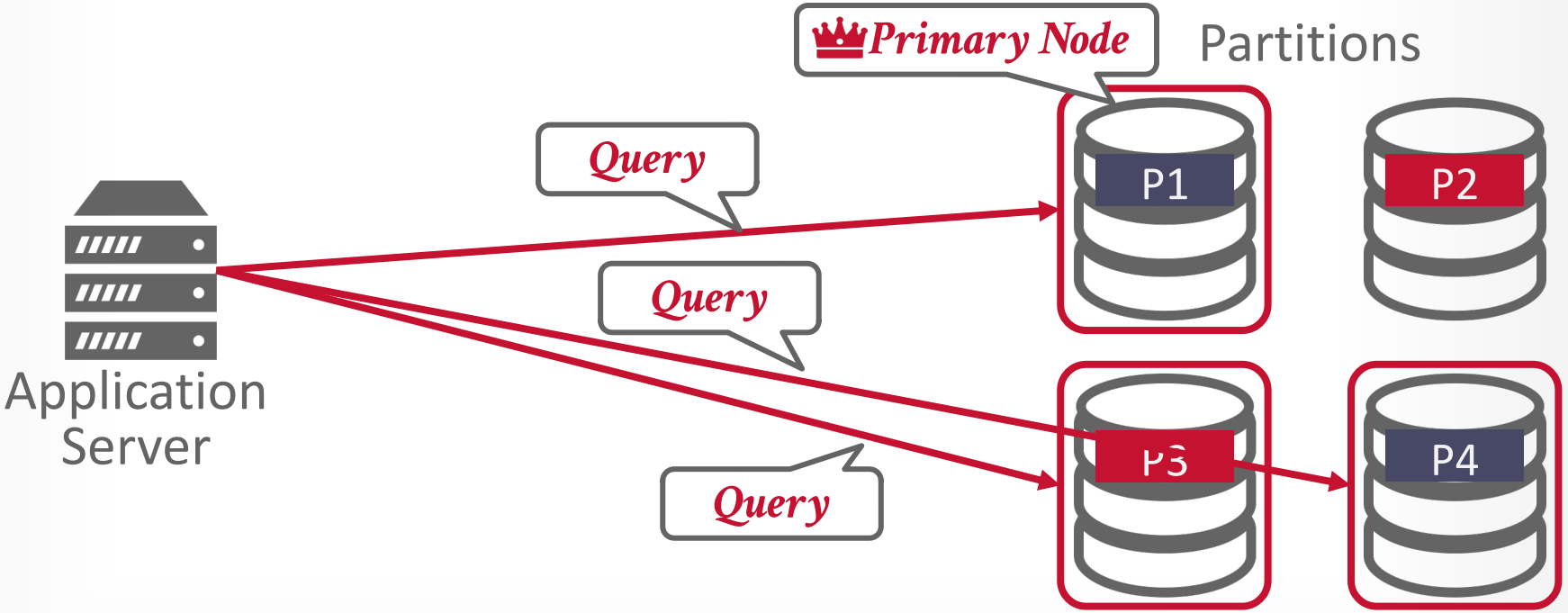
Partitions



DECENTRALIZED COORDINATOR

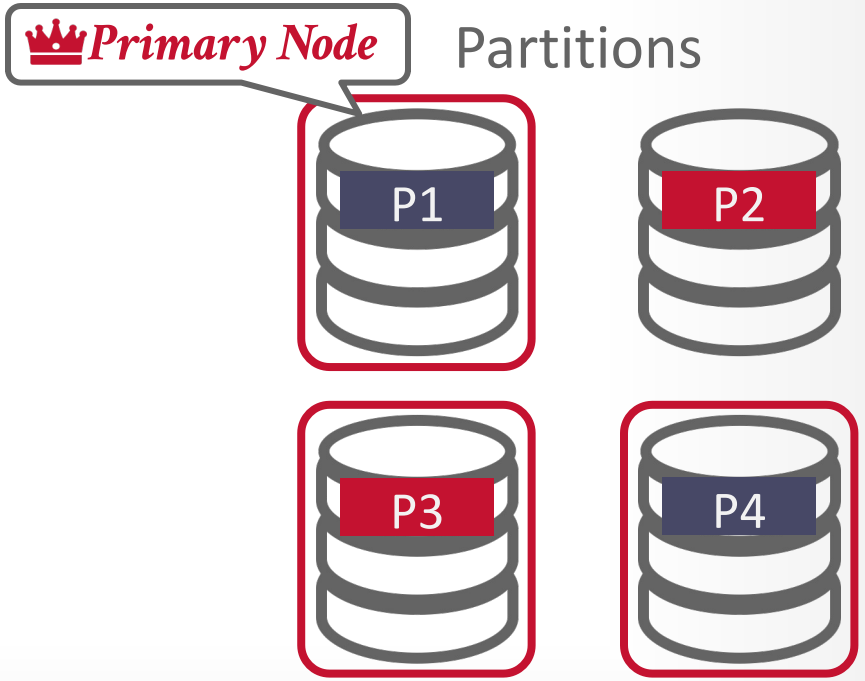


DECENTRALIZED COORDINATOR

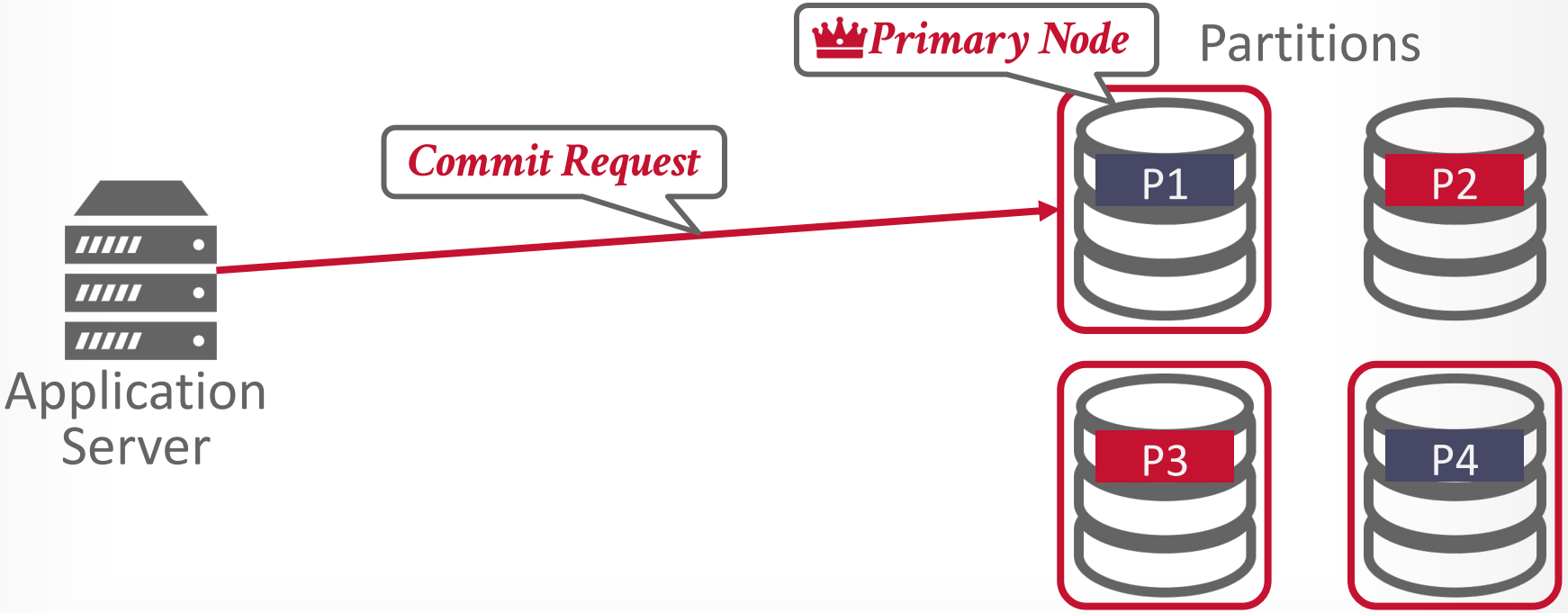


DECENTRALIZED COORDINATOR

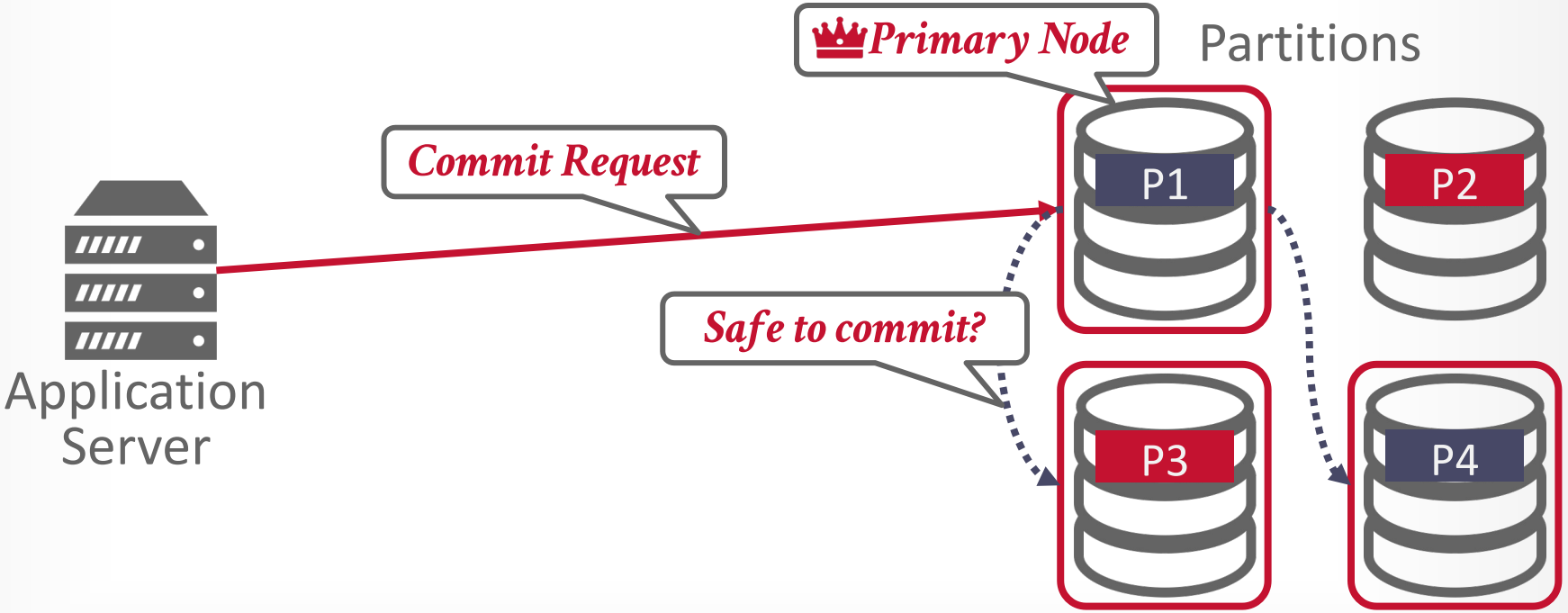
Application Server



DECENTRALIZED COORDINATOR



DECENTRALIZED COORDINATOR



OBSERVATION

Recall that our goal is to have multiple physical nodes appear as a single logical DBMS.

We have not discussed how to ensure that all nodes agree to commit a txn and then to make sure it does commit if the DBMS decides it should.

→ What happens if a node fails?
→ What happens if messages show up late?
→ What happens if the system does not wait for every node to agree to commit?

IMPORTANT ASSUMPTION

We will assume that all nodes in a distributed DBMS are well-behaved and under the same administrative domain.

→ If we tell a node to commit a txn, then it will commit the txn (if there is not a failure).

If you do not trust the other nodes in a distributed DBMS, then you need to use a Byzantine Fault Tolerant protocol for txns (blockchain).

→ Blockchains are **not** good for high-throughput workloads.

IMPORTANT ASSUMPTION

We will assume that all nodes in a distributed DBMS are well-behaved and under the same administrative domain.

→ If we tell a node to commit a txn, then it will commit the txn (if there is not a failure).

If you do not trust the other nodes in a distributed DBMS, then you need to use a Byzantine Fault Tolerant protocol for txns (blockchain).

→ Blockchains are **not** good for high-throughput workloads.

TODAY'S AGENDA

Replication

Atomic Commit Protocols

Consistency Issues (CAP / PACELC)

REPLICATION

The DBMS can replicate a database across redundant nodes to increase availability.

→ Partitioned vs. Non-Partitioned
→ Shared-Nothing vs. Shared-Disk

Design Decisions:

→ Replica Configuration
→ Propagation Scheme
→ Propagation Timing
→ Update Method

REPLICA CONFIGURATIONS

Approach #1: Primary-Replica

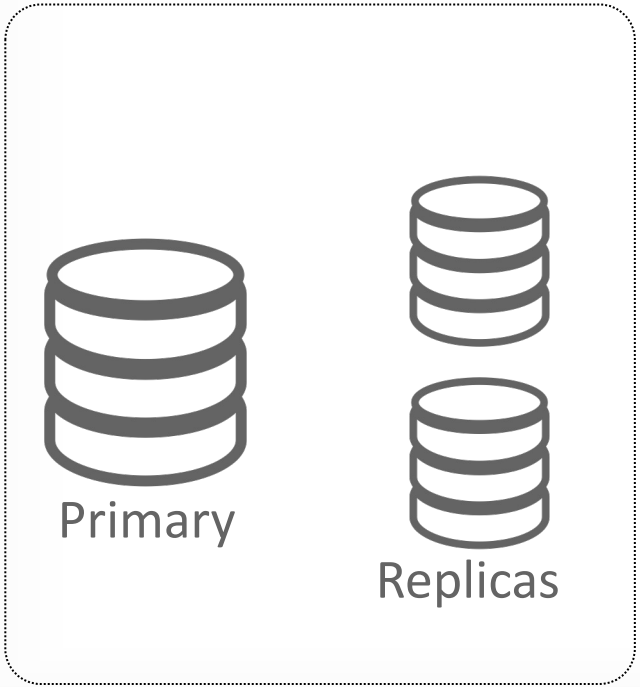
→ All updates go to a designated primary for each object.
→ The primary propagates updates to its replicas by shipping logs.
→ Read-only txns may be allowed to access replicas.
→ If the primary goes down, then hold an election to select a new primary.

Approach #2: Multi-Primary

→ Txns can update data objects at any replica.
→ Replicas must synchronize with each other using an atomic commit protocol.

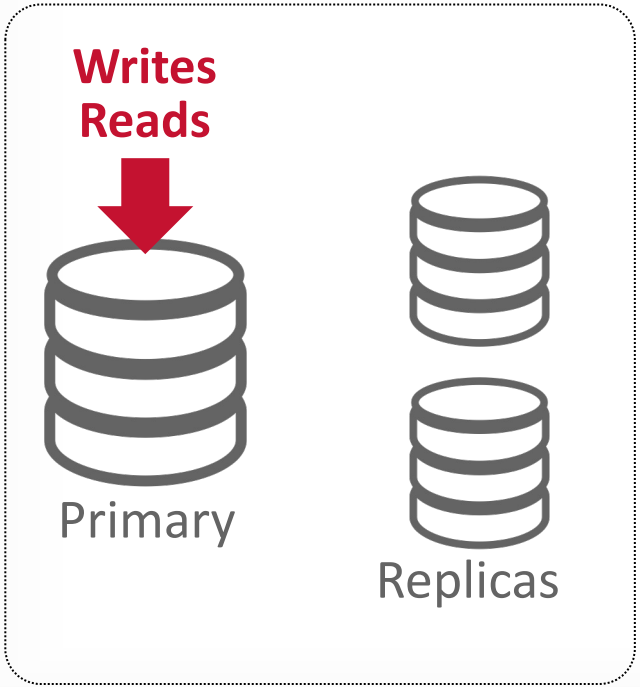
REPLICA CONFIGURATIONS

Primary-Replica



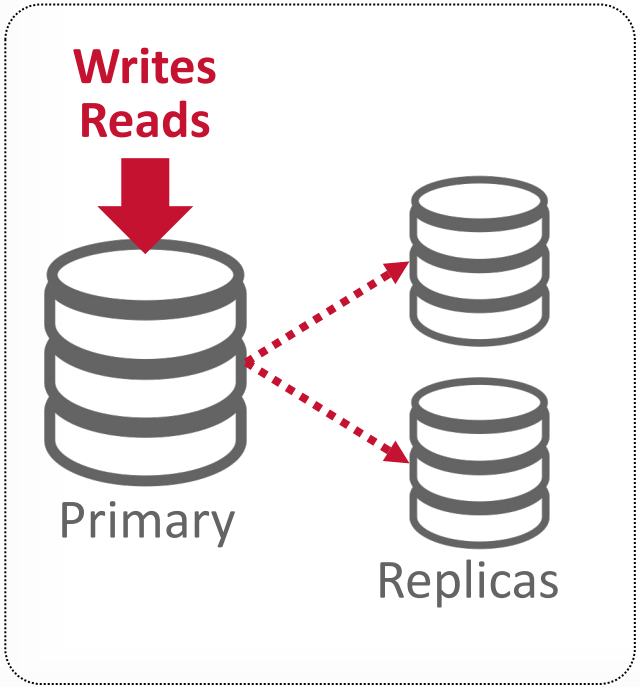
REPLICA CONFIGURATIONS

Primary-Replica



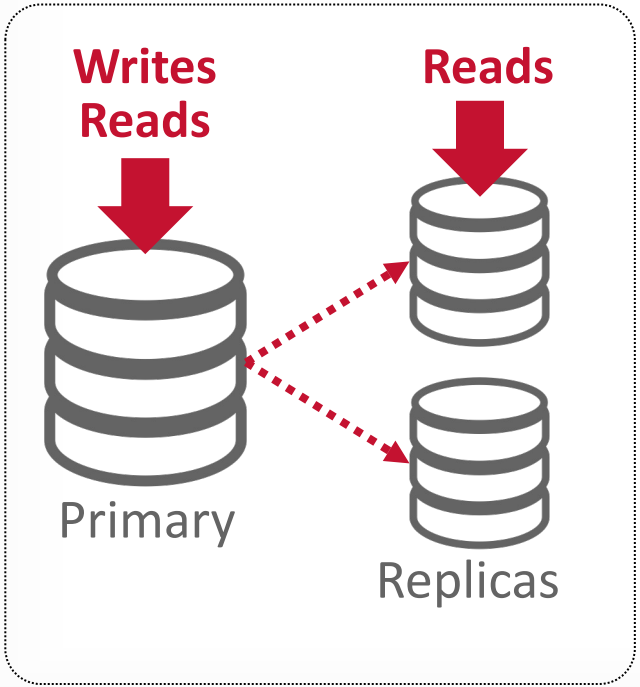
REPLICA CONFIGURATIONS

Primary-Replica



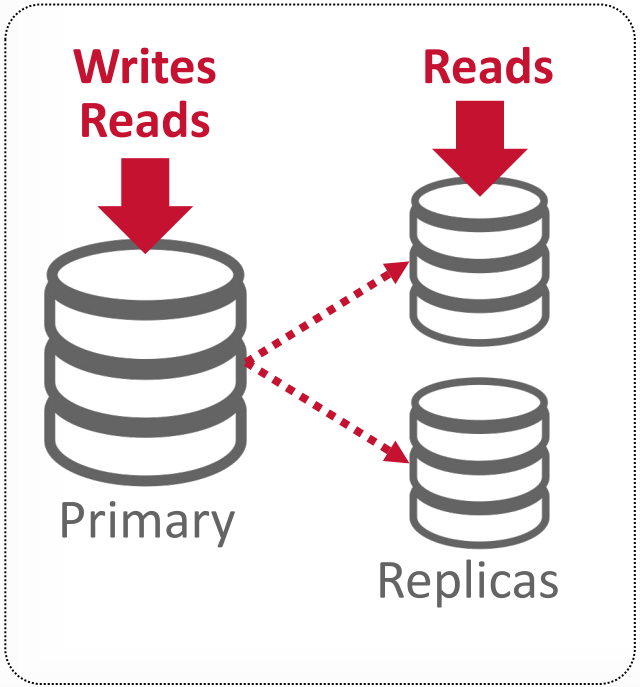
REPLICA CONFIGURATIONS

Primary-Replica

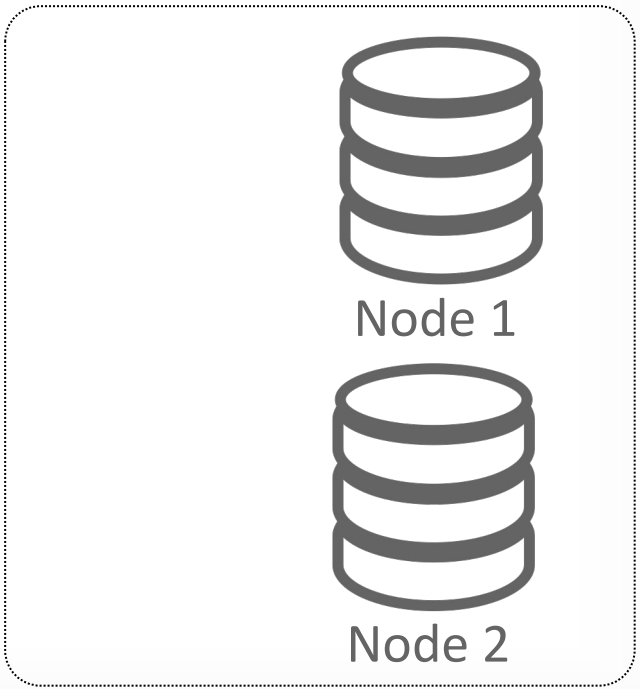


REPLICA CONFIGURATIONS

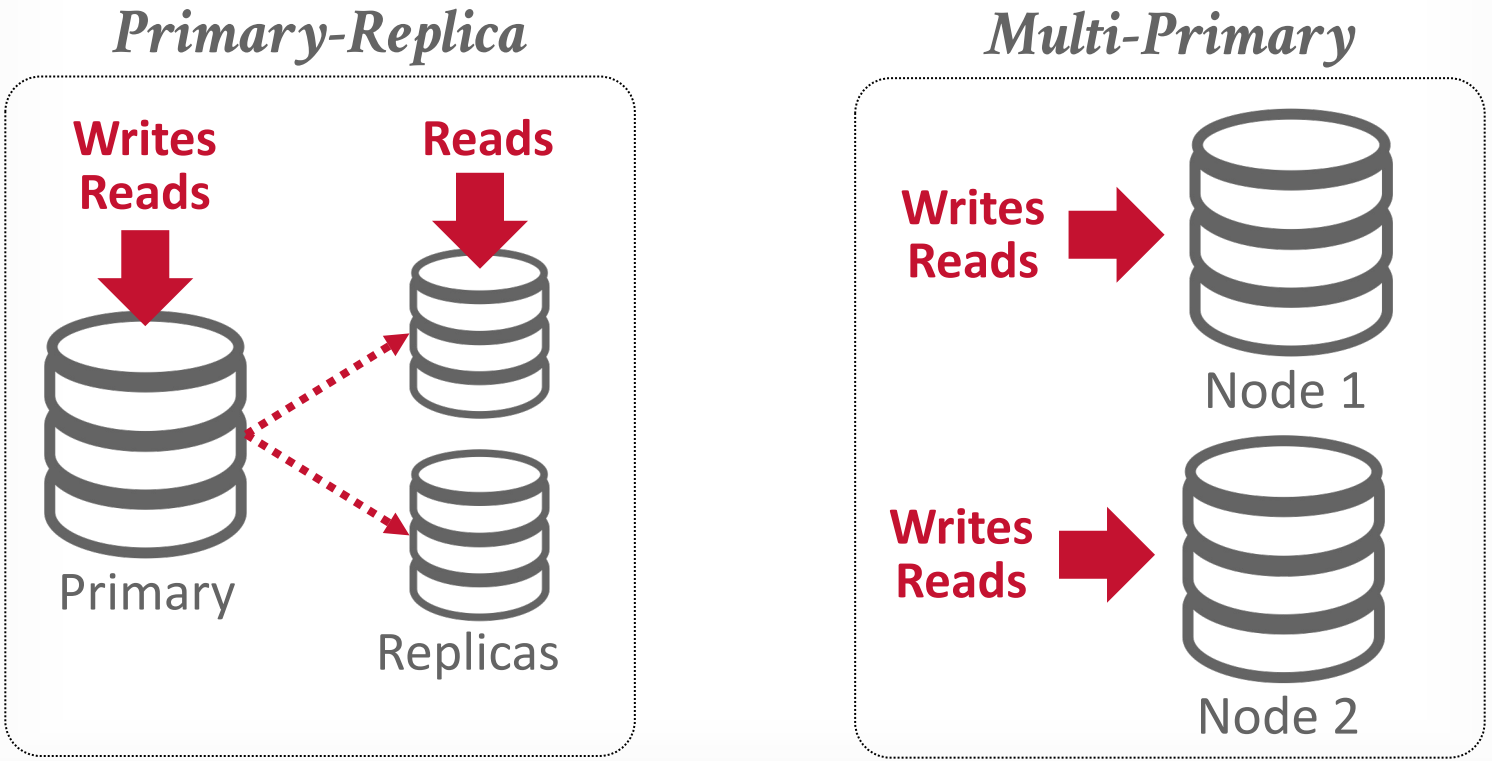
Primary-Replica



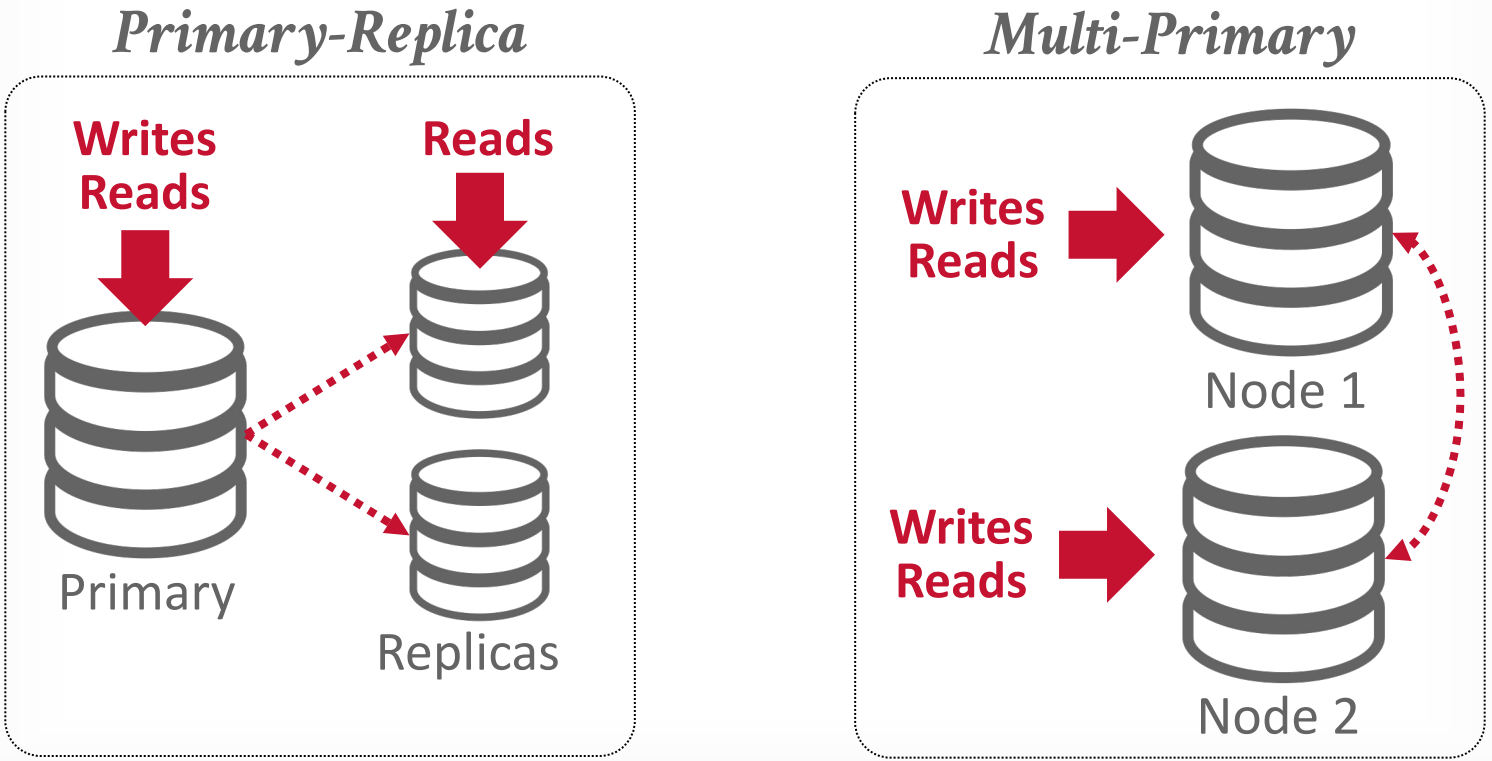
Multi-Primary



REPLICA CONFIGURATIONS



REPLICA CONFIGURATIONS



K-SAFETY

***K*-safety is a threshold for determining the fault tolerance of the replicated database.**

The value *K* represents the number of replicas per data object that must always be available.

If the number of replicas goes below this threshold, then the DBMS halts execution and takes itself offline.

PROPAGATION SCHEME

When a txn commits on a replicated database, the DBMS decides whether it must wait for that txn's changes to propagate to other nodes before it can send the acknowledgement to application.

Propagation levels:

→ Synchronous (*Strong Consistency*)

→ Asynchronous (*Eventual Consistency*)

PROPAGATION SCHEME

Approach #1: Synchronous

→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.

PROPAGATION SCHEME

Approach #1: Synchronous

→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.

PROPAGATION SCHEME

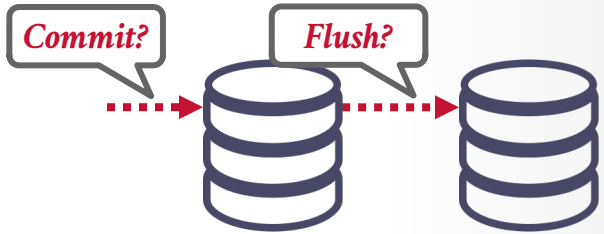
Approach #1: Synchronous

→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.

PROPAGATION SCHEME

Approach #1: Synchronous

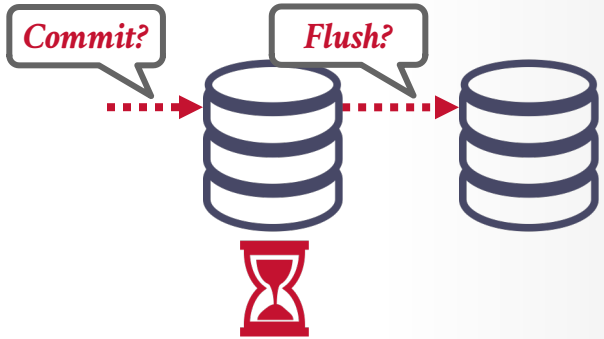
→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



PROPAGATION SCHEME

Approach #1: Synchronous

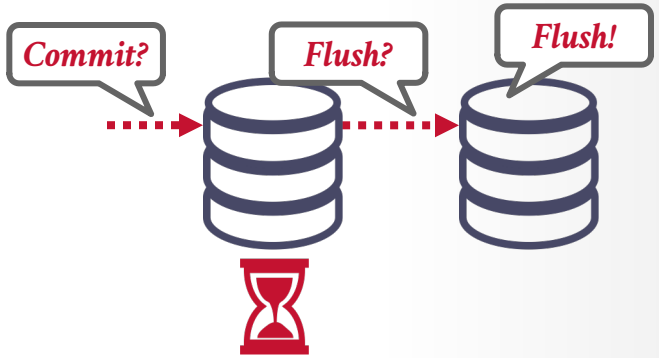
→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



PROPAGATION SCHEME

Approach #1: Synchronous

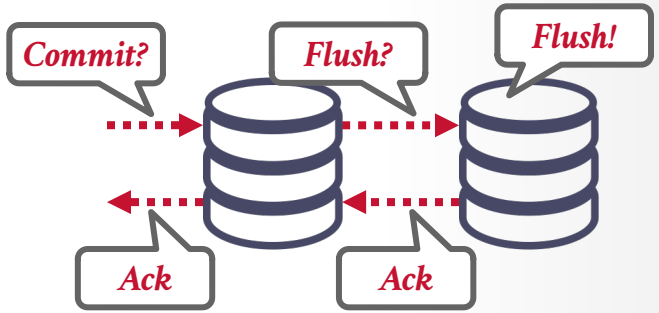
→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



PROPAGATION SCHEME

Approach #1: Synchronous

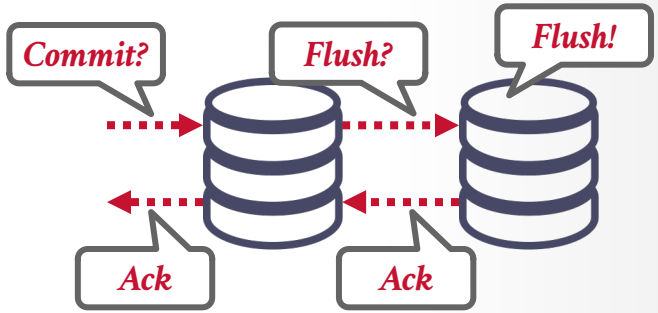
→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



PROPAGATION SCHEME

Approach #1: Synchronous

→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



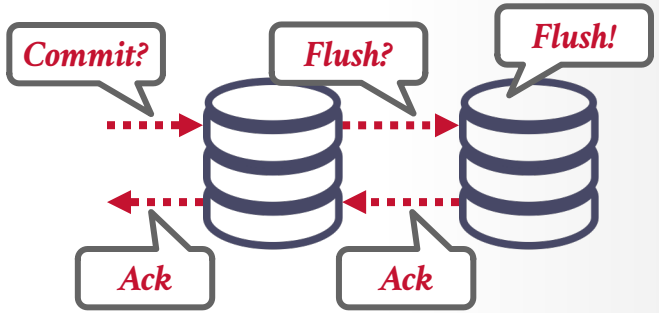
Approach #2: Asynchronous

→ The primary immediately returns the acknowledgement to the client without waiting for replicas to apply the changes.

PROPAGATION SCHEME

Approach #1: Synchronous

→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



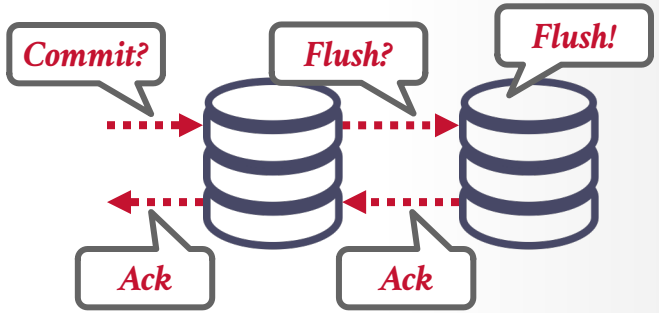
Approach #2: Asynchronous

→ The primary immediately returns the acknowledgement to the client without waiting for replicas to apply the changes.

PROPAGATION SCHEME

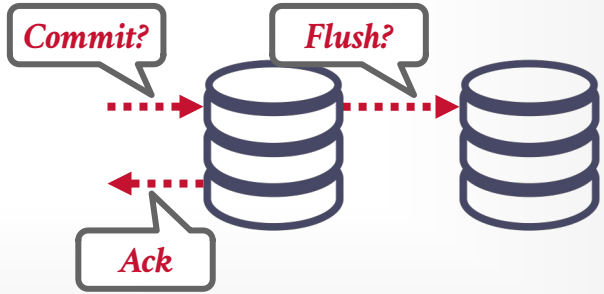
Approach #1: Synchronous

→ The primary sends updates to replicas and then waits for them to acknowledge that they fully applied (i.e., logged) the changes.



Approach #2: Asynchronous

→ The primary immediately returns the acknowledgement to the client without waiting for replicas to apply the changes.



PROPAGATION TIMING

Approach #1: Continuous

→ The DBMS sends log messages immediately as it generates them.
→ Also need to send a commit/abort message.

Approach #2: On Commit

→ The DBMS only sends the log messages for a txn to the replicas once the txn is commits.
→ Do not waste time sending log records for aborted txns.

ACTIVE VS. PASSIVE

Approach #1: Active-Active

→ A txn executes at each replica independently.
→ Need to check at the end whether the txn ends up with the same result at each replica.

Approach #2: Active-Passive

→ Each txn executes at a single location and propagates the changes to the replica.
→ Can either do physical or logical replication.
→ Not the same as Primary-Replica vs. Multi-Primary

OBSERVATION

If only one node decides whether a txn is allowed to commit, then making that decision is easy.

Life is much harder when multiple nodes are allowed to decide:

→ What if multiple nodes need to agree a txn is allowed to commit?

→ What if a primary node goes down and the system needs to choose a new primary?

ATOMIC COMMIT PROTOCOL

Coordinating the commit order of txns across nodes in a distributed DBMS.

→ Commit Order = State Machine

→ It does not matter whether the database's contents are replicated or partitioned.

Examples:

→ Two-Phase Commit (1970s)

→ Three-Phase Commit (1983)

→ Viewstamped Replication (1988)

→ Paxos (1989)

→ ZAB (2008?)

→ Raft (2013)

ATOMIC COMMIT PROTOCOL

Coordinating the commit order of txns across nodes in a distributed DBMS.

→ Commit Order = State Machine

→ It does not matter whether the database's contents are replicated or partitioned.

Examples:

→ Two-Phase Commit (1970s)

→ Three-Phase Commit (1983)

→ Viewstamped Replication (1988)

→ Paxos (1989)

→ ZAB (2008?)

→ Raft (2013)

ATOMIC COMMIT PROTOCOL

Resource Managers (RMs)

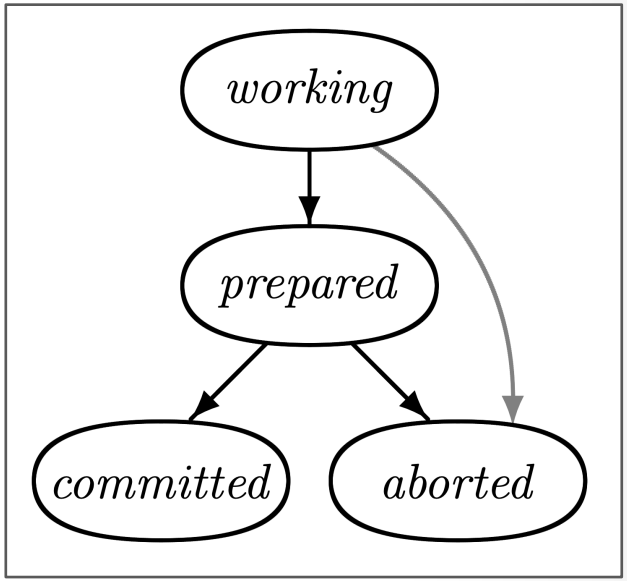
→ Execute on different nodes
→ Coordinate to decide fate of a txn.

Properties of the Commit Protocol

→ **Stability**: Once the fate is decided, it cannot be changed.
→ **Consistency**: All RMs end up in the same state.

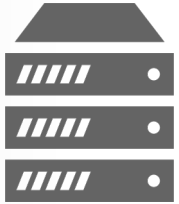
Assumes Liveness:

→ There is some way of progressing forward.
→ Enough nodes are alive and connected for the duration of the protocol.



https://www.microsoft.com/en-us/research/publication/consensus-on-transaction-commit/

TWO-PHASE COMMIT (SUCCESS)



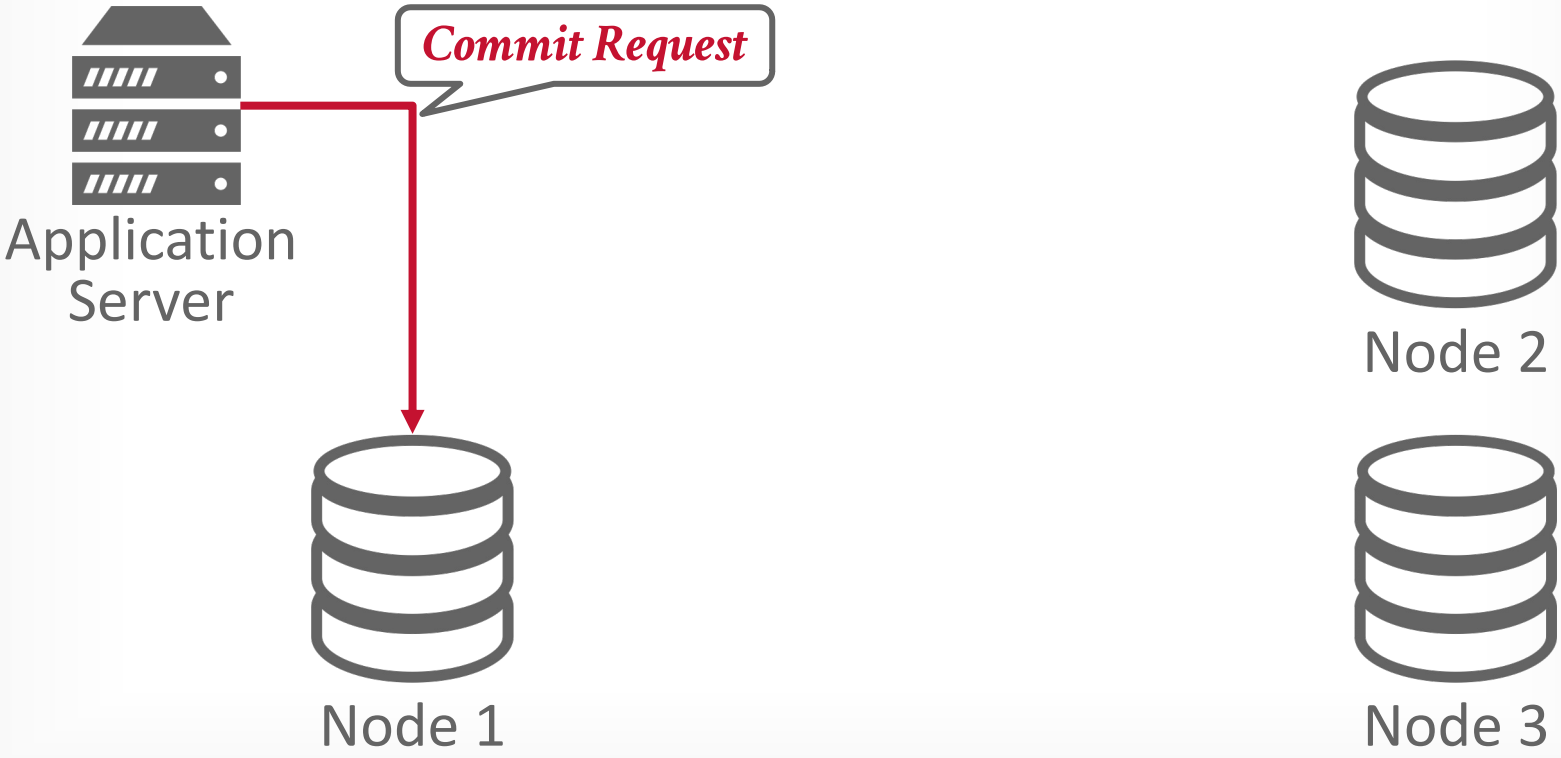
Application Server

Node 2

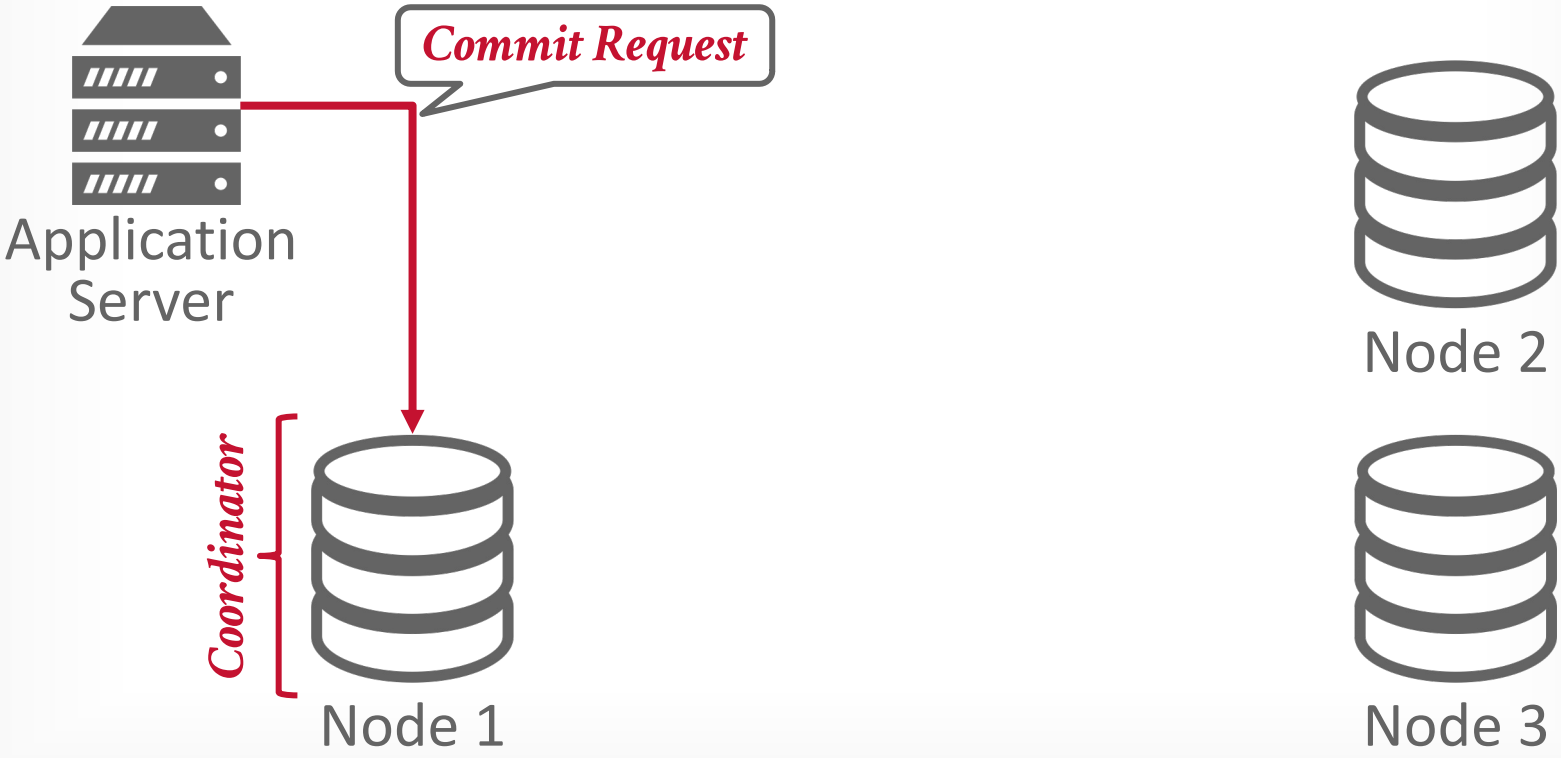
Node 1

Node 3

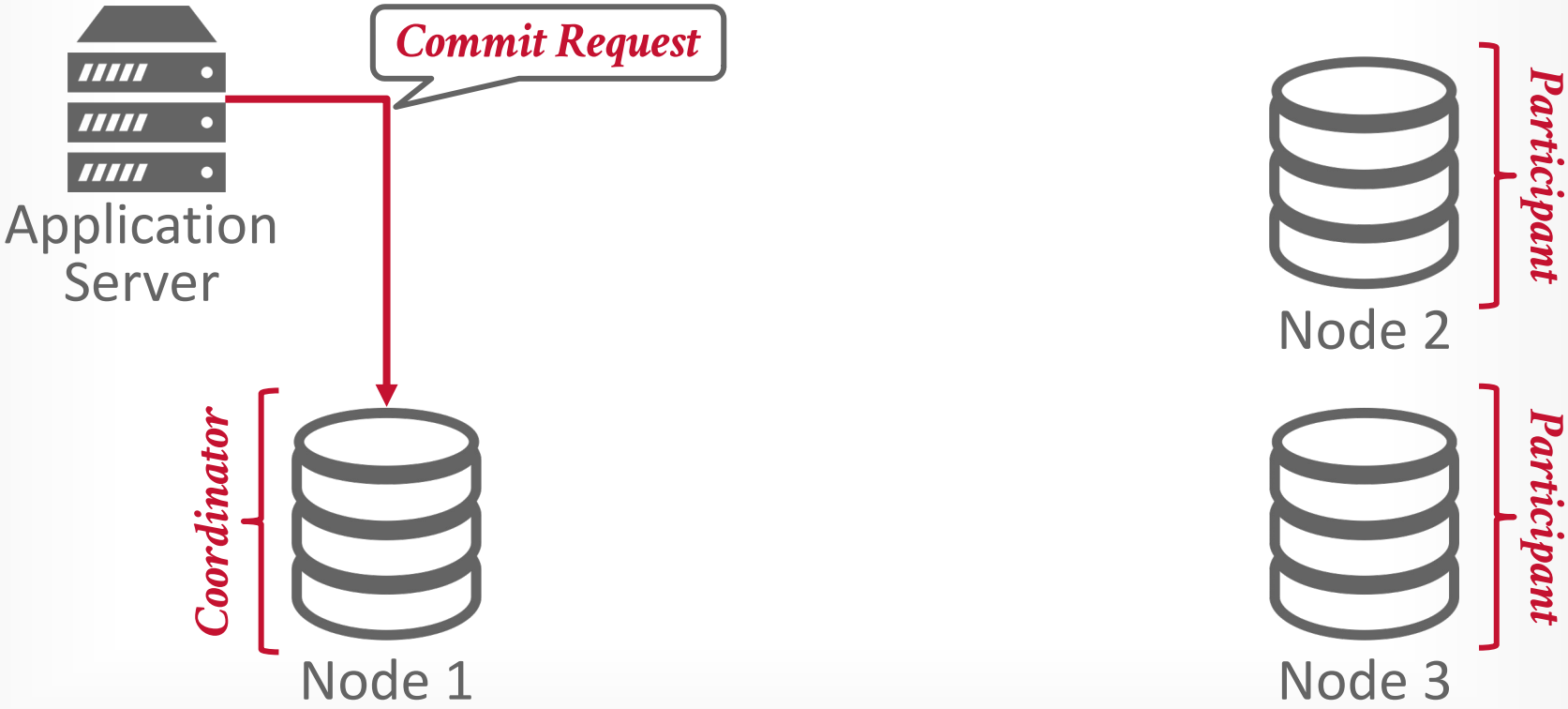
TWO-PHASE COMMIT (SUCCESS)



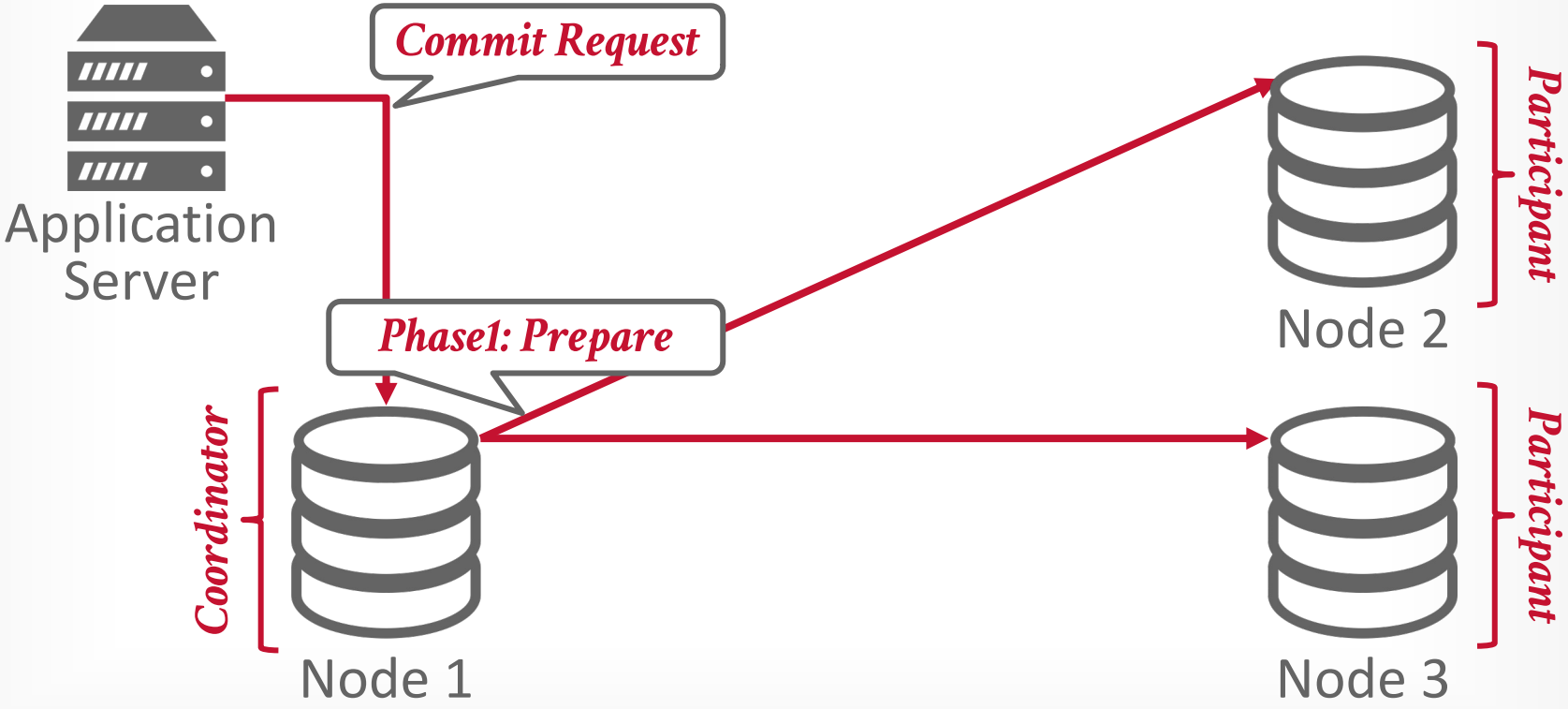
TWO-PHASE COMMIT (SUCCESS)



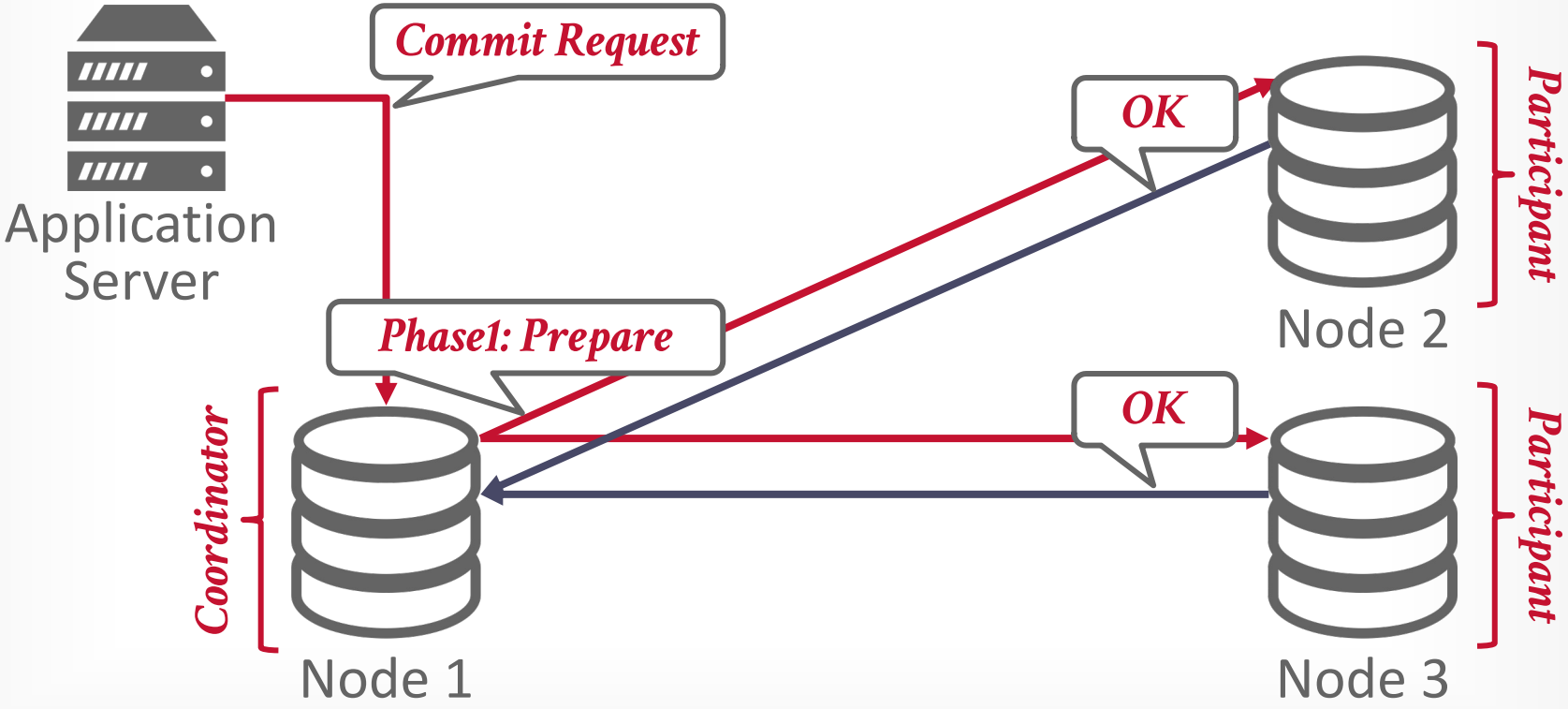
TWO-PHASE COMMIT (SUCCESS)



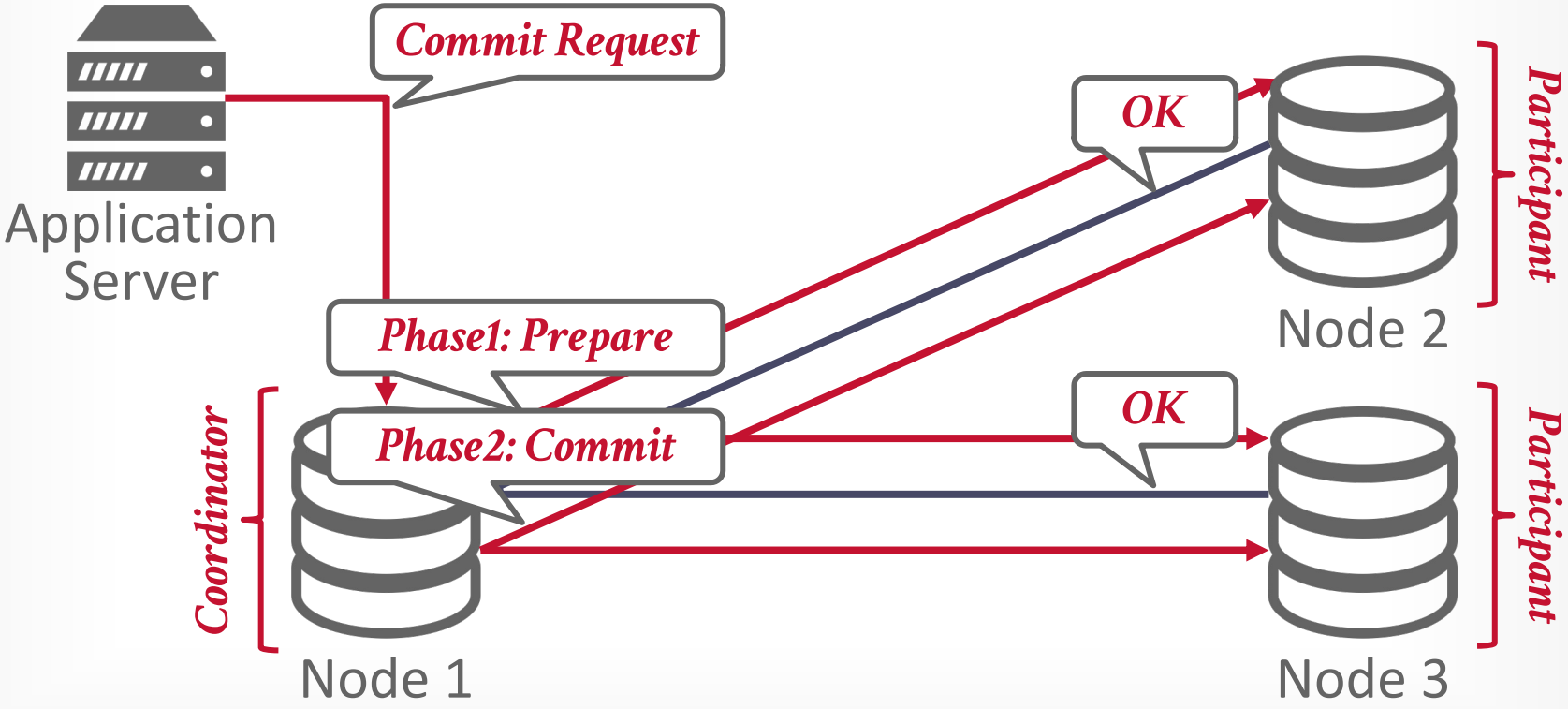
TWO-PHASE COMMIT (SUCCESS)



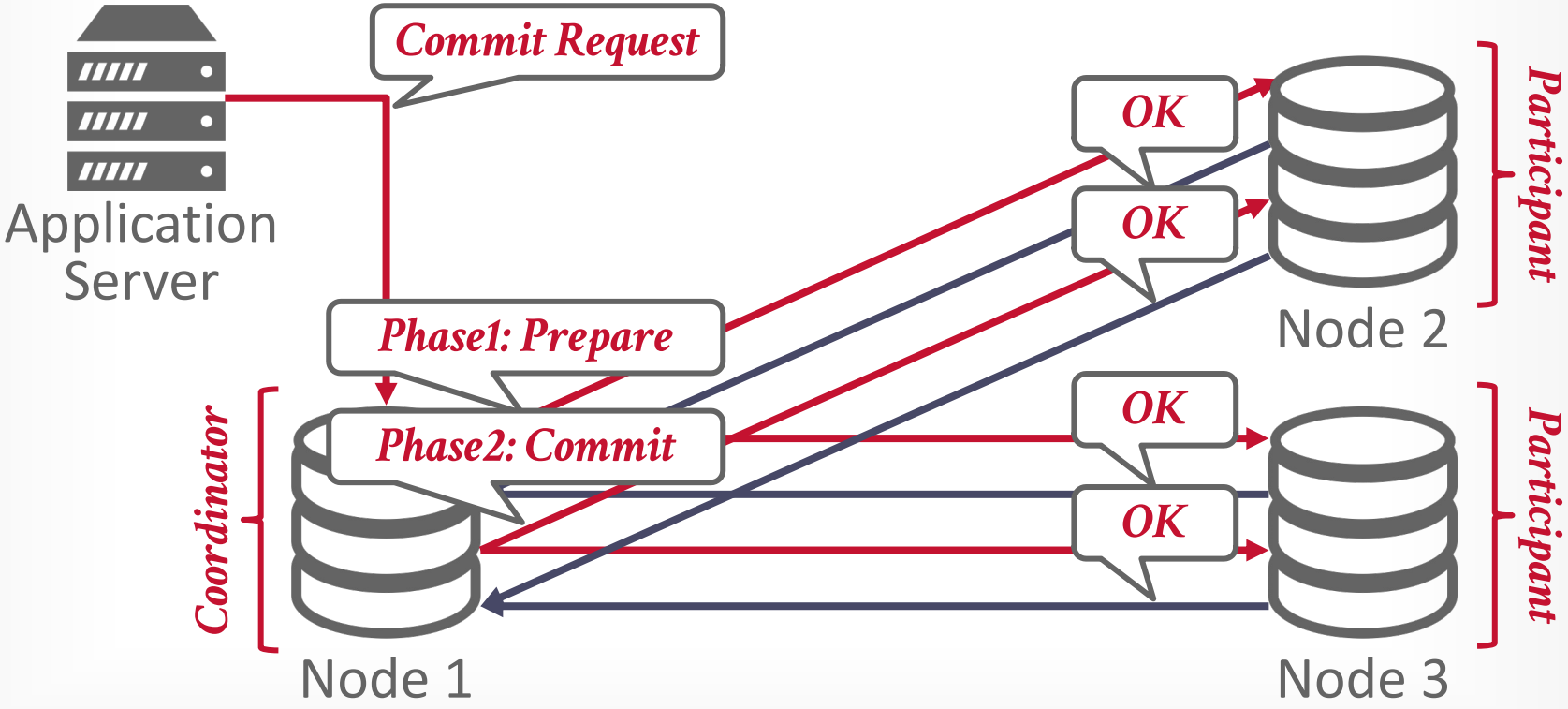
TWO-PHASE COMMIT (SUCCESS)



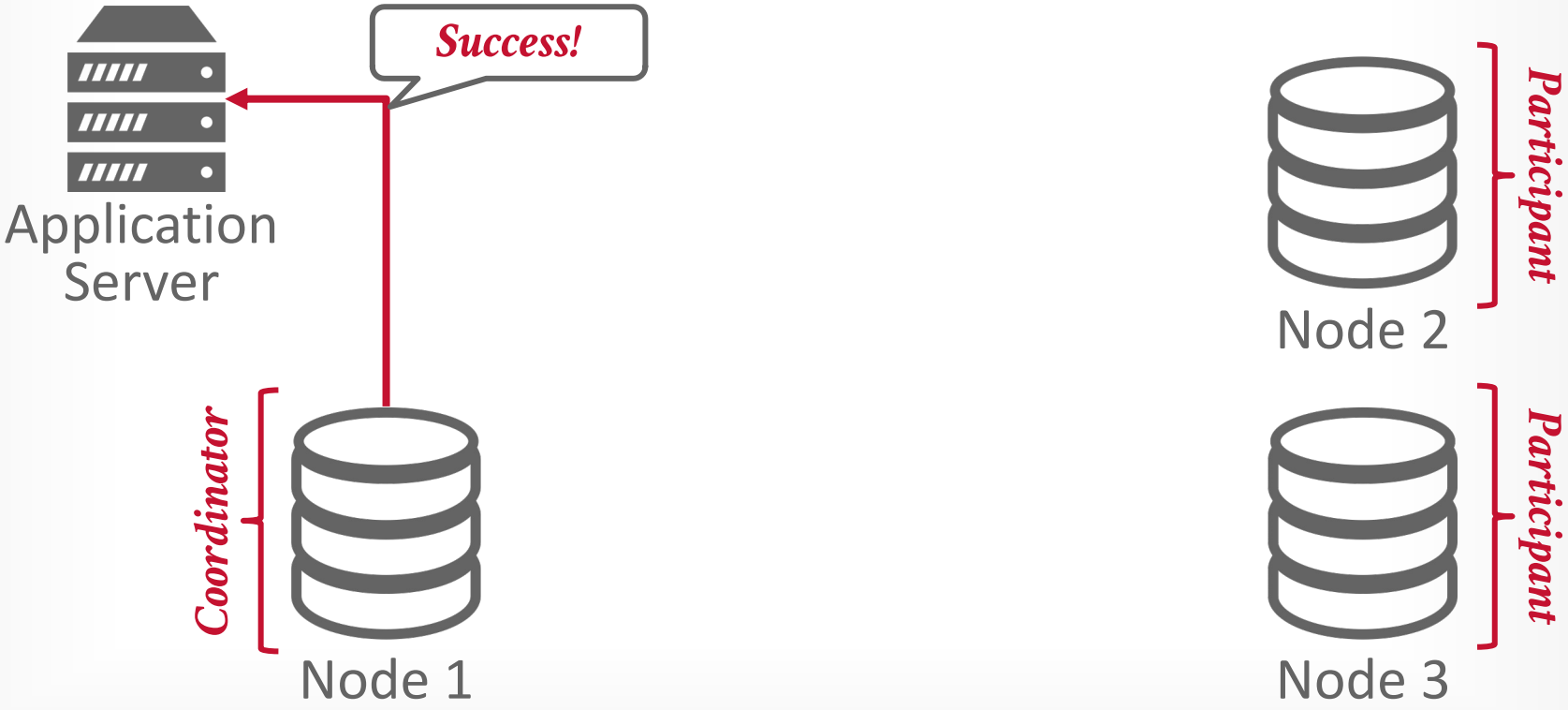
TWO-PHASE COMMIT (SUCCESS)



TWO-PHASE COMMIT (SUCCESS)



TWO-PHASE COMMIT (SUCCESS)



TWO-PHASE COMMIT (ABORT)

Application Server

Participant

Node 2

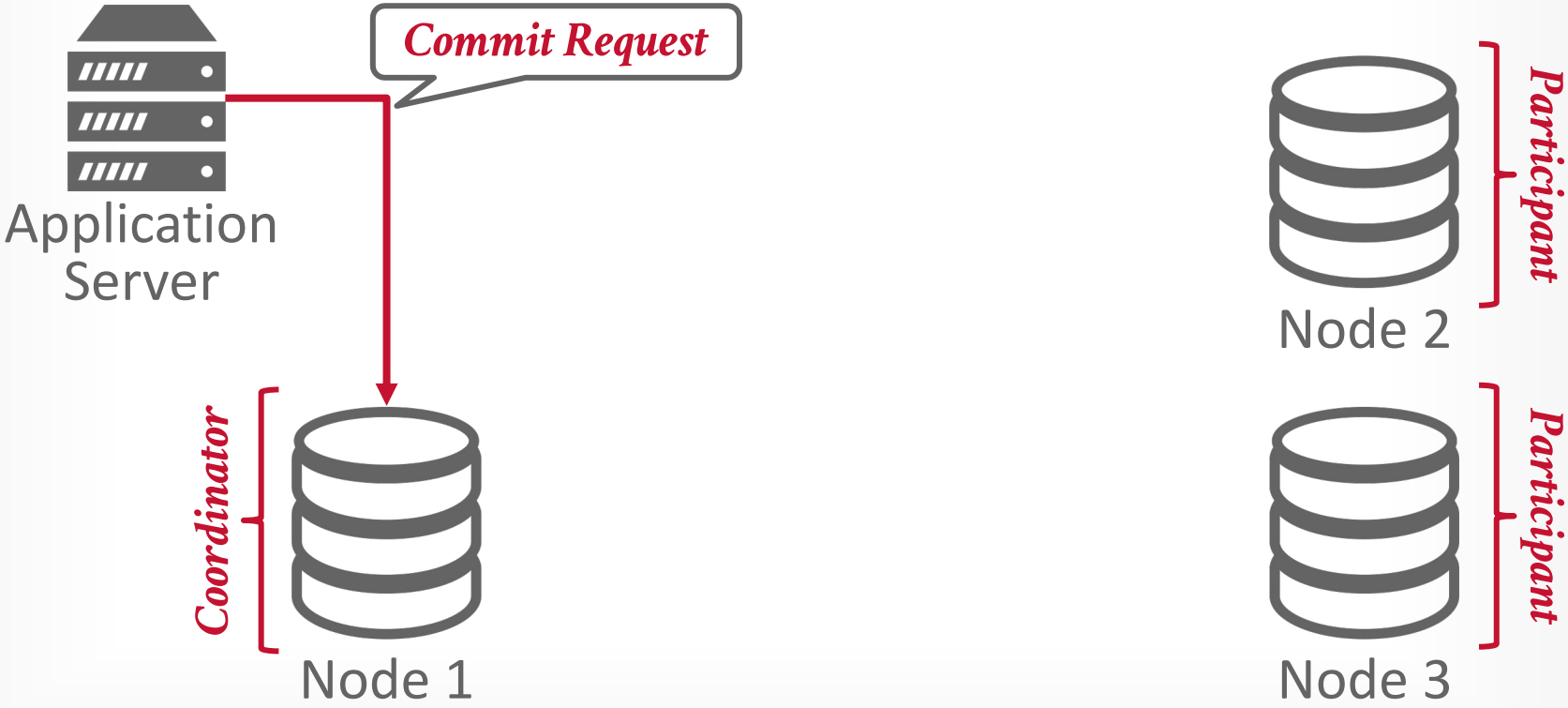
Coordinator

Node 1

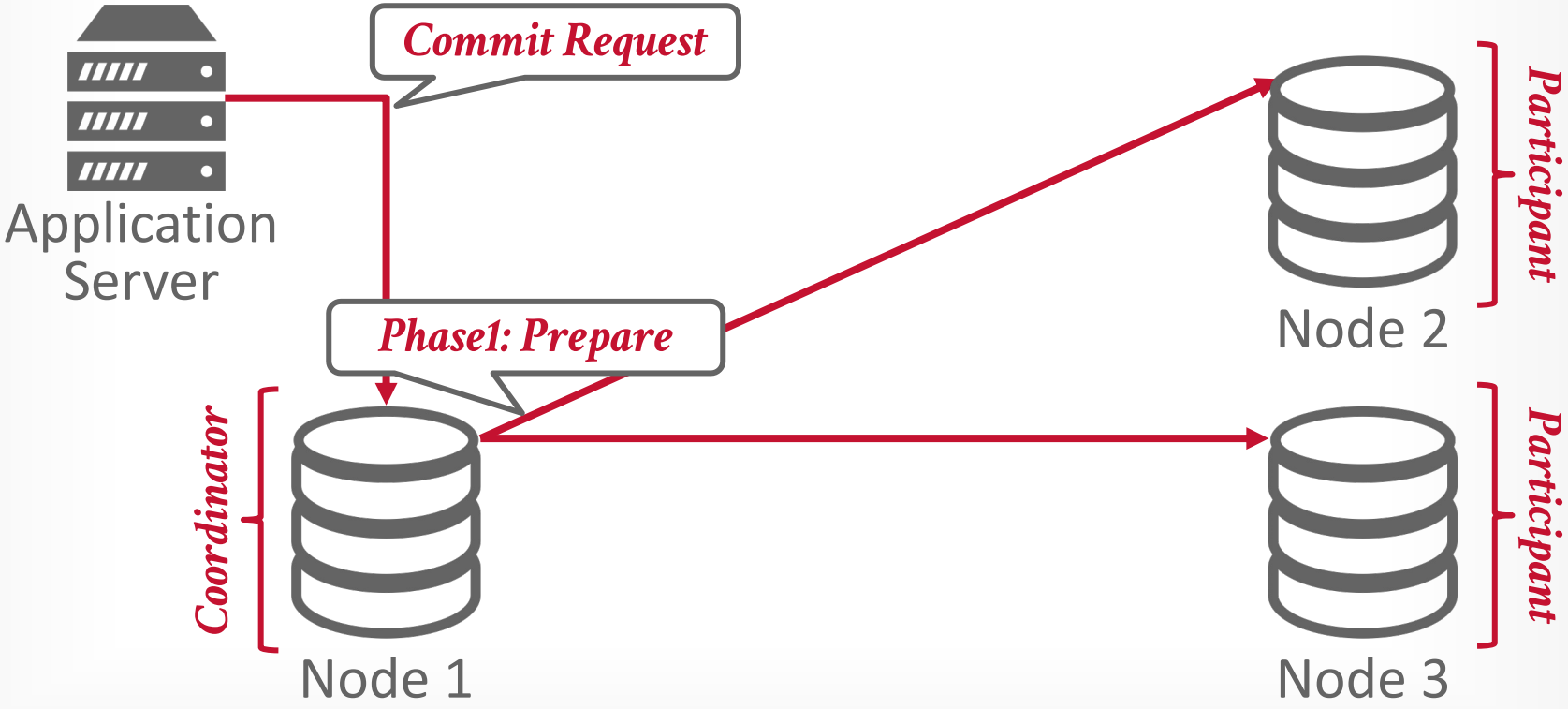
Participant

Node 3

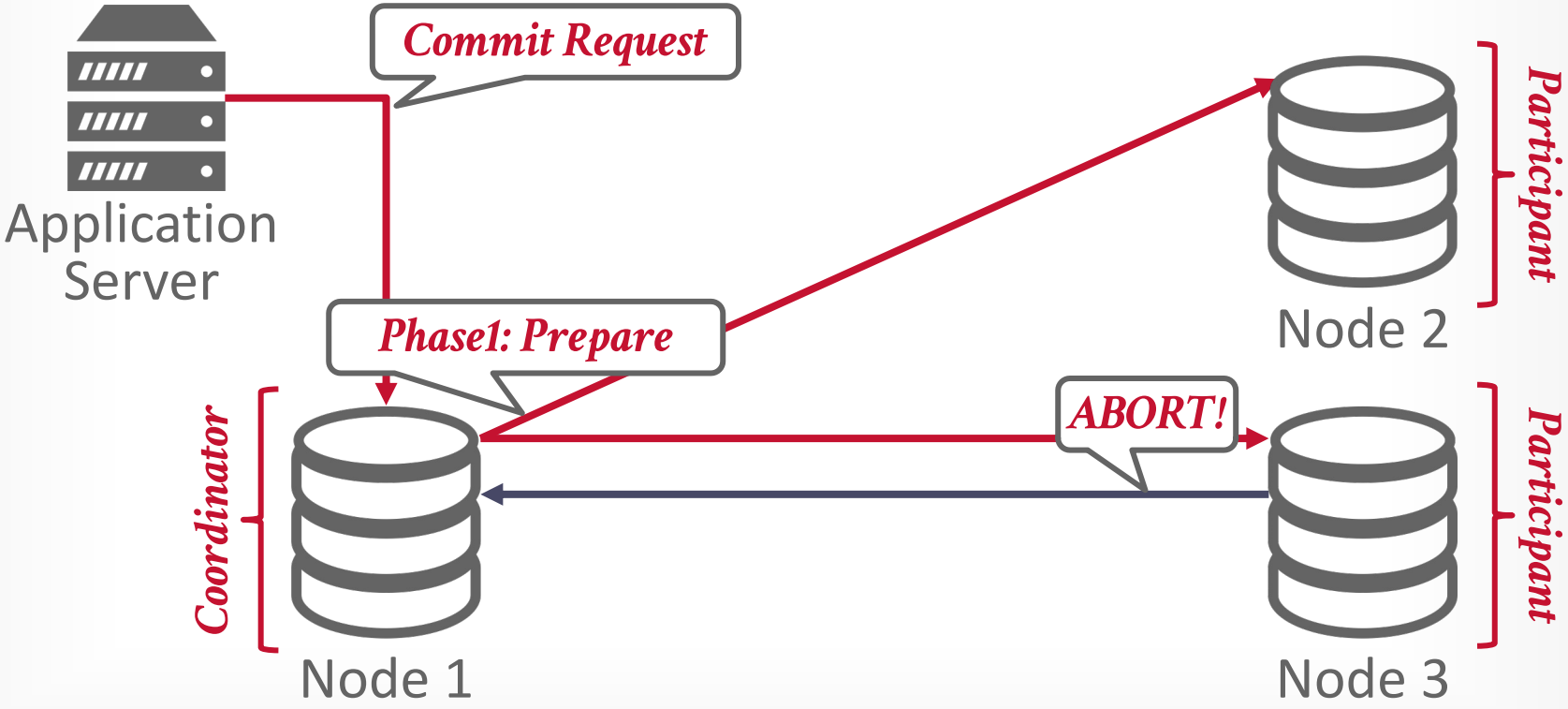
TWO-PHASE COMMIT (ABORT)



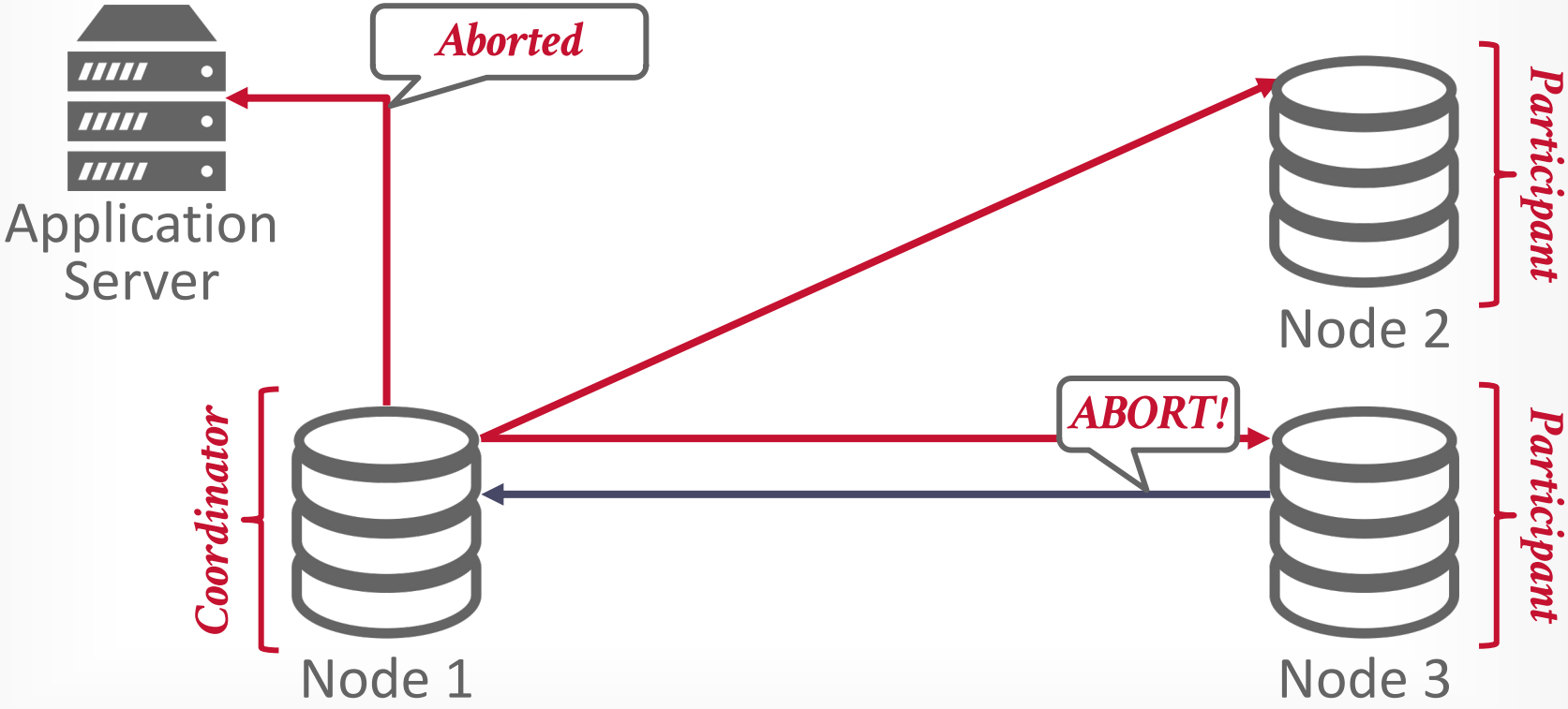
TWO-PHASE COMMIT (ABORT)



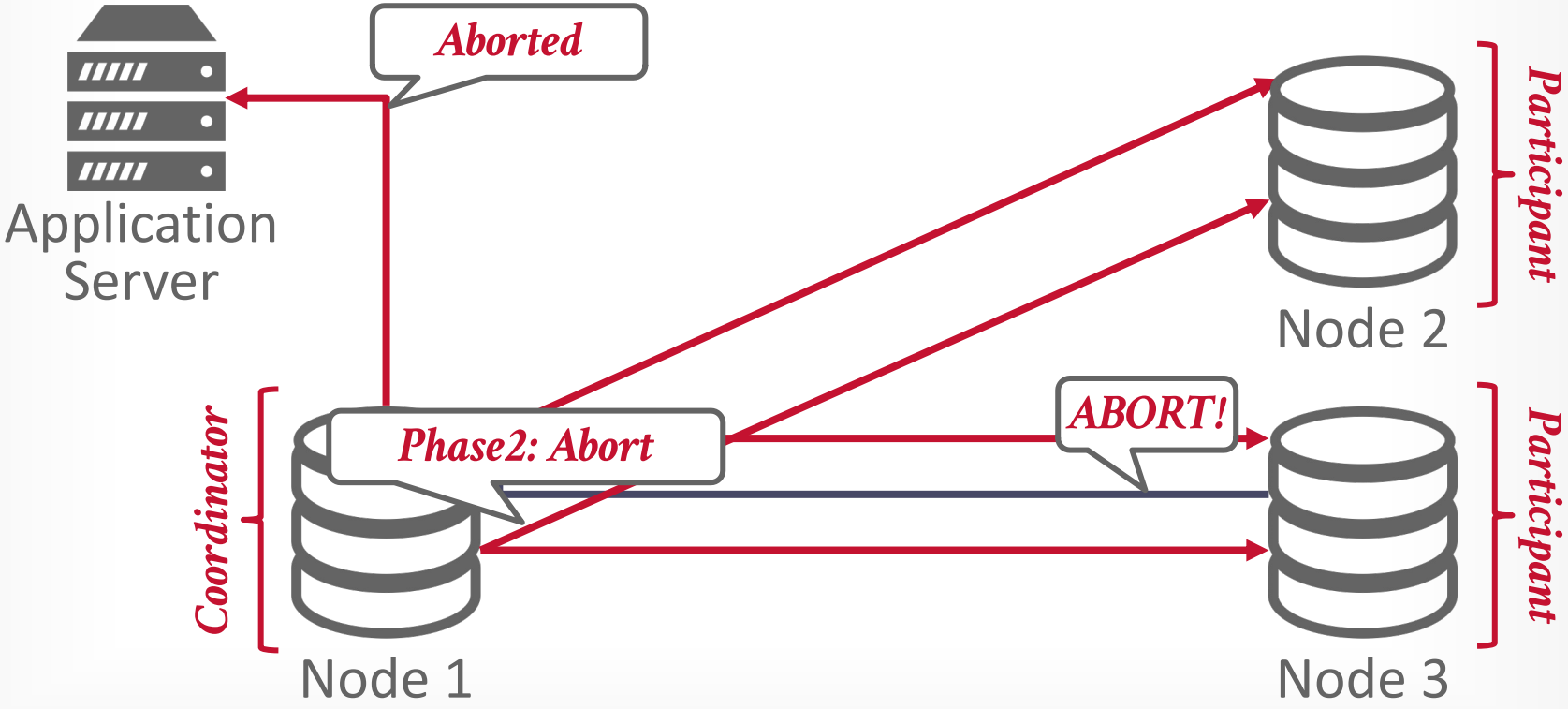
TWO-PHASE COMMIT (ABORT)



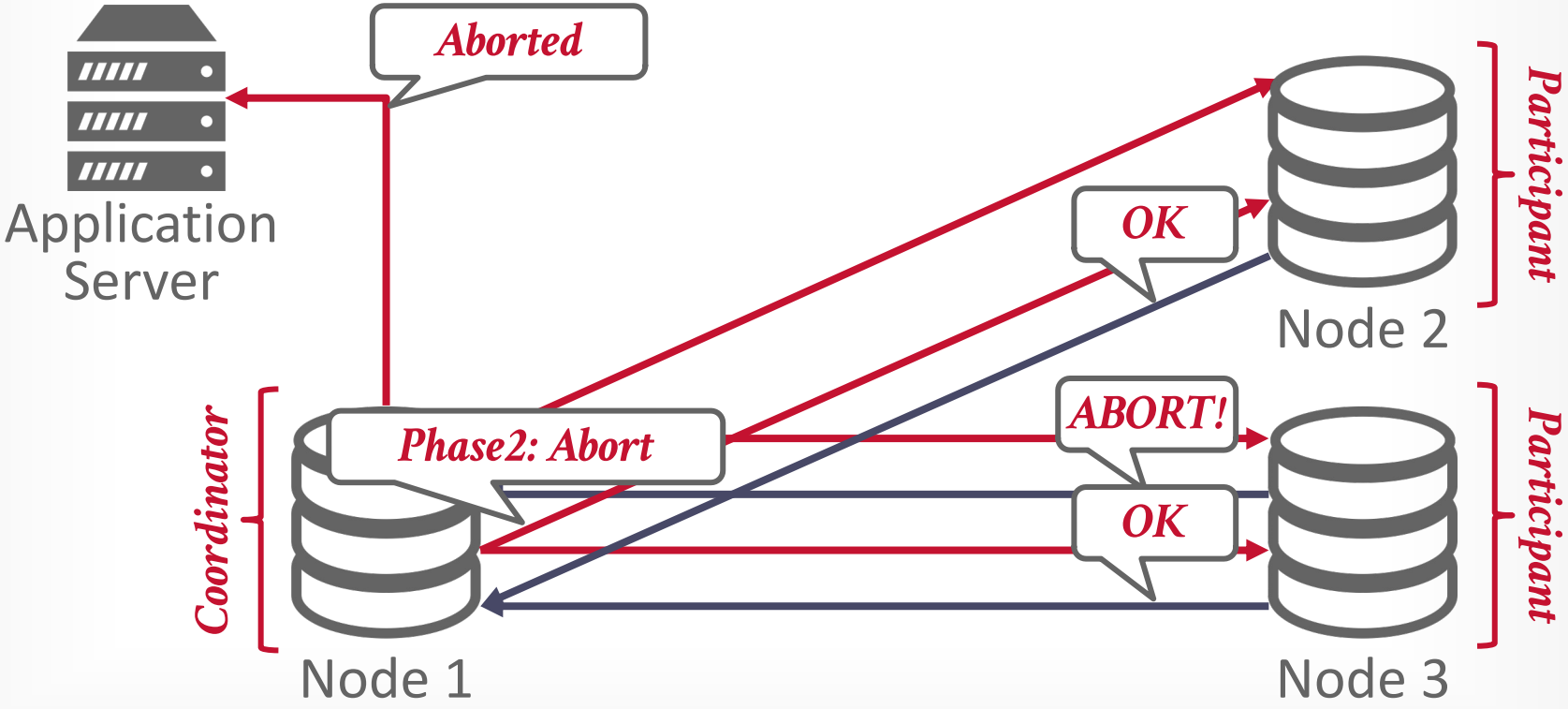
TWO-PHASE COMMIT (ABORT)



TWO-PHASE COMMIT (ABORT)



TWO-PHASE COMMIT (ABORT)



TWO-PHASE COMMIT

Each node records the inbound/outbound messages and outcome of each phase in a non-volatile storage log.

On recovery, examine the log for 2PC messages:

→ If local txn in prepared state, contact coordinator.
→ If local txn not in prepared, abort it.
→ If local txn was committing and node is the coordinator, send `COMMIT` message to nodes.

TWO-PHASE COMMIT FAILURES

What happens if the coordinator crashes?

What happens if the participant crashes?

TWO-PHASE COMMIT FAILURES

What happens if the coordinator crashes?

→ Participants must decide what to do after a timeout *(this only applies if the participants know of all other participants)*.

→ System is not available during this time.

What happens if the participant crashes?

TWO-PHASE COMMIT FAILURES

What happens if the coordinator crashes?

→ Participants must decide what to do after a timeout *(this only applies if the participants know of all other participants)*.

→ System is not available during this time.

What happens if the participant crashes?

→ Coordinator assumes that it responded with an abort if it has not sent an acknowledgement yet.

→ Again, nodes use a timeout to determine whether a participant is dead.

2PC OPTIMIZATIONS

Early Prepare Voting *(Rare)*

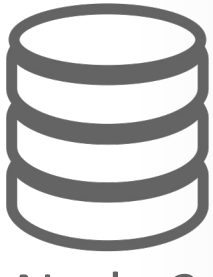
→ If you send a query/request to a remote node that you know will be the last one to execute in this txn, then that node will also return their vote for the prepare phase with the query result.

Early Ack After Prepare *(Common)*

→ If all nodes vote to commit a txn, the coordinator can send the client an acknowledgement that their txn was successful before the commit phase finishes.

EARLY ACKNOWLEDGEMENT

Application Server



Participant

Node 2

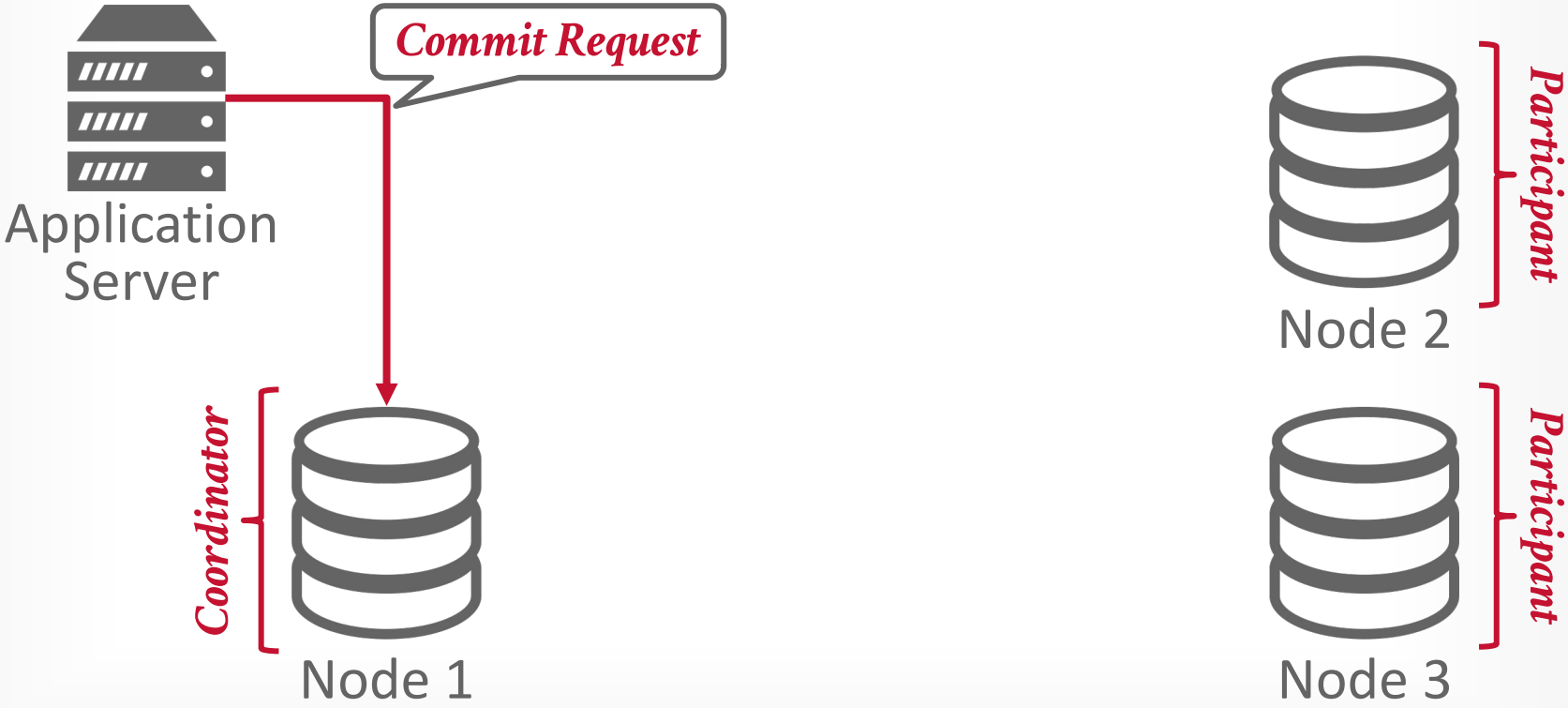
Coordinator

Node 1

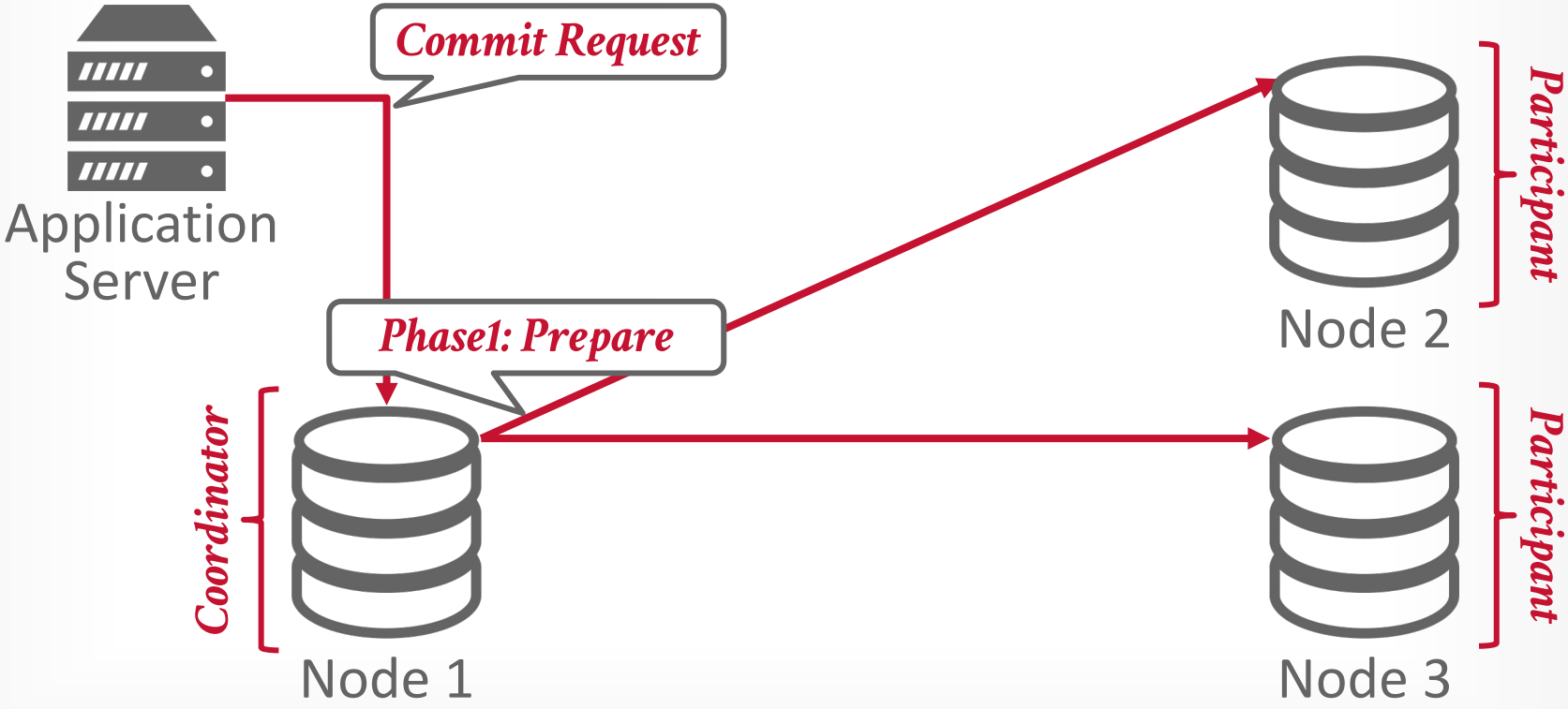
Participant

Node 3

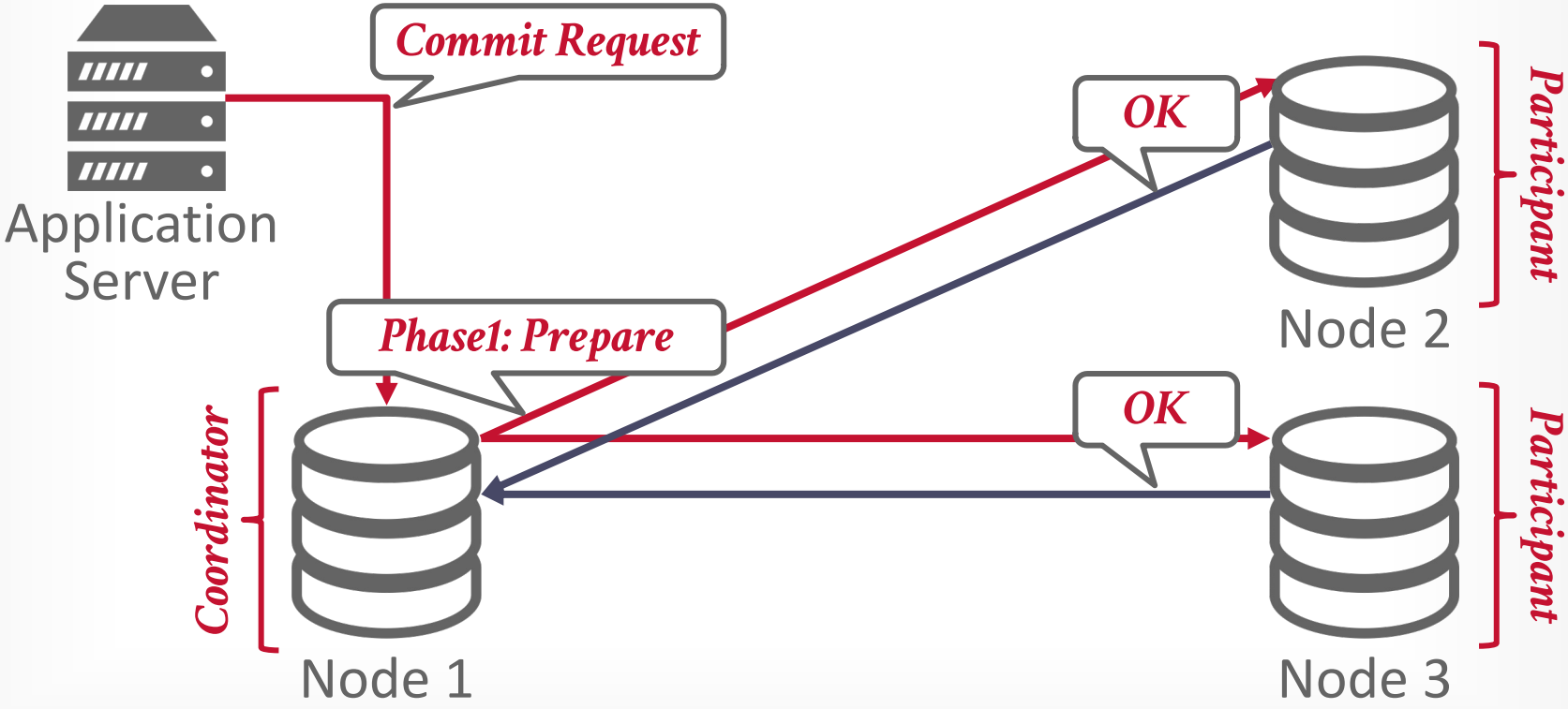
EARLY ACKNOWLEDGEMENT



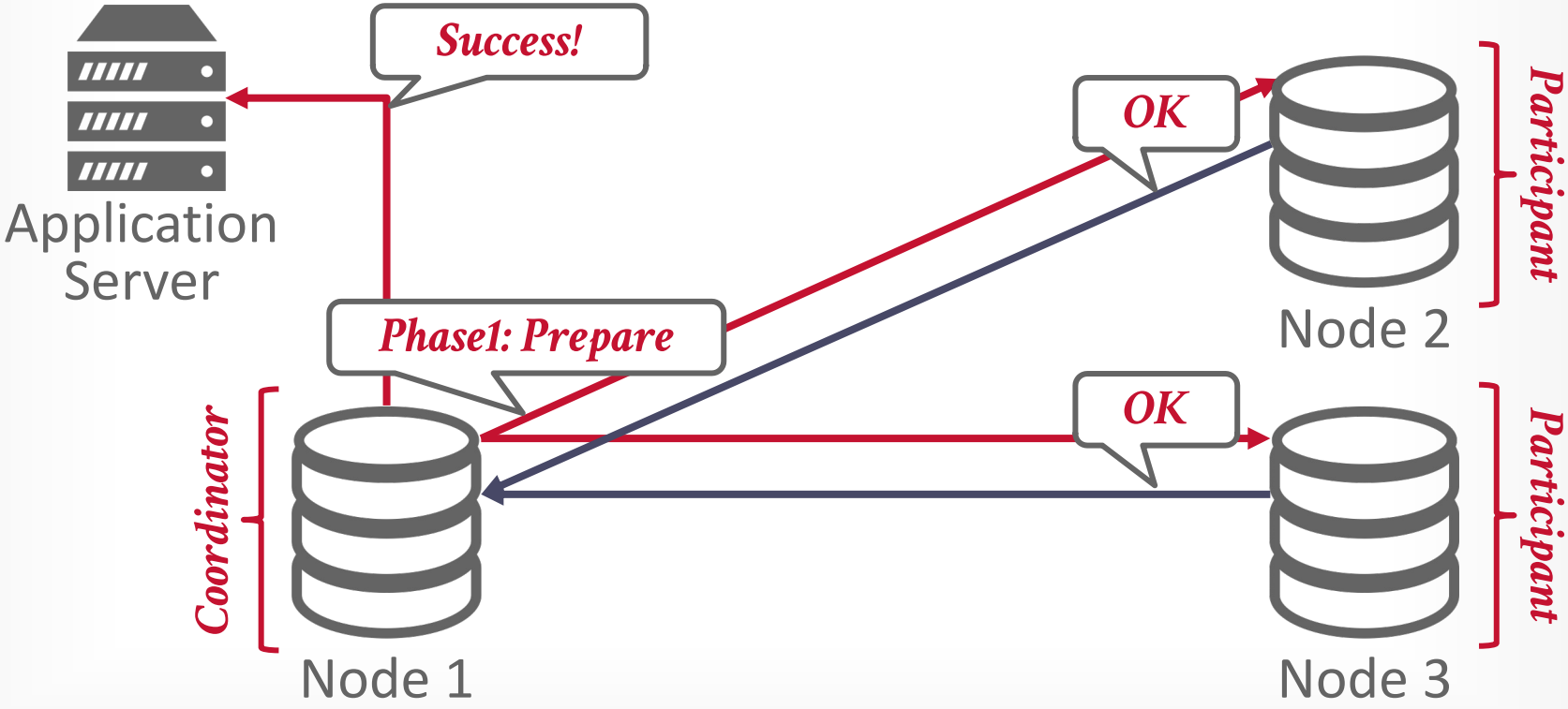
EARLY ACKNOWLEDGEMENT



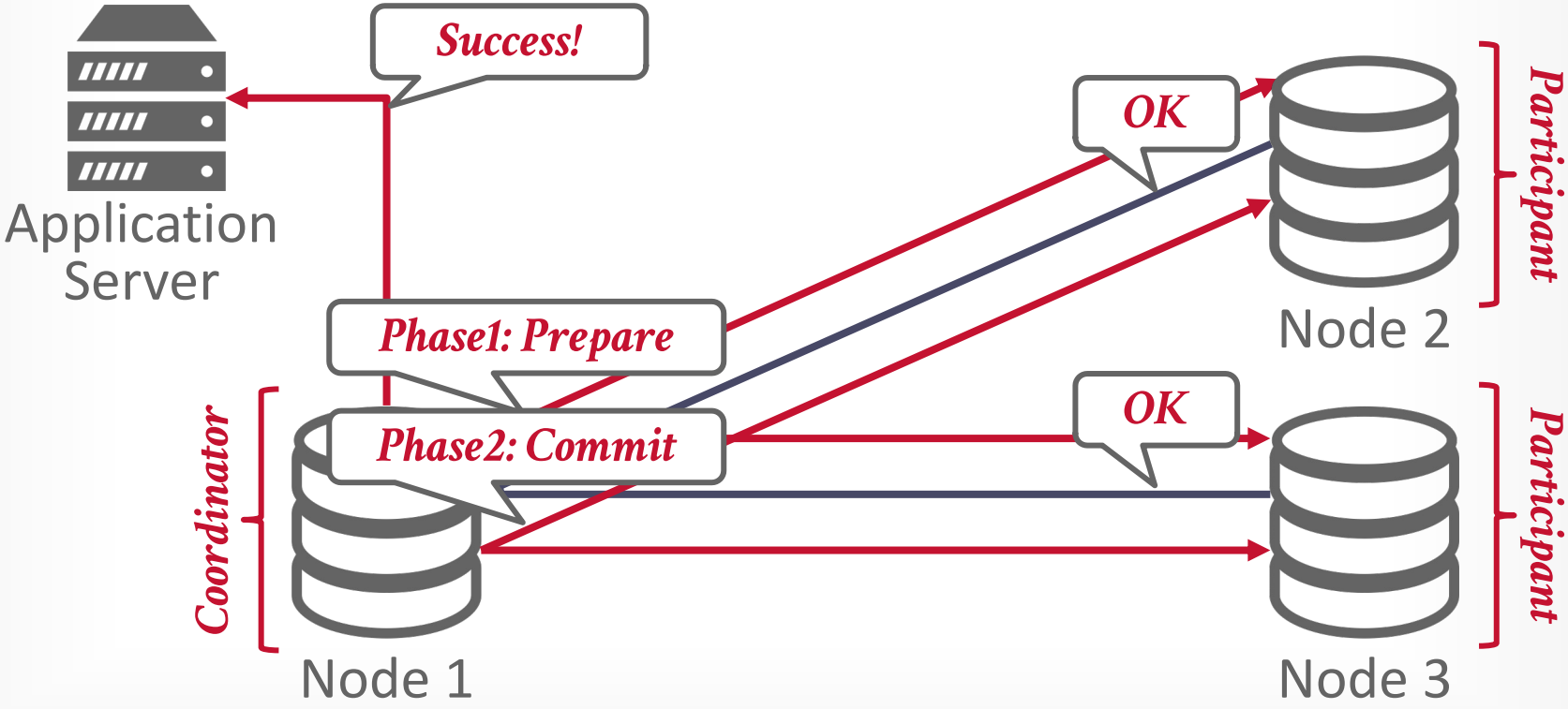
EARLY ACKNOWLEDGEMENT



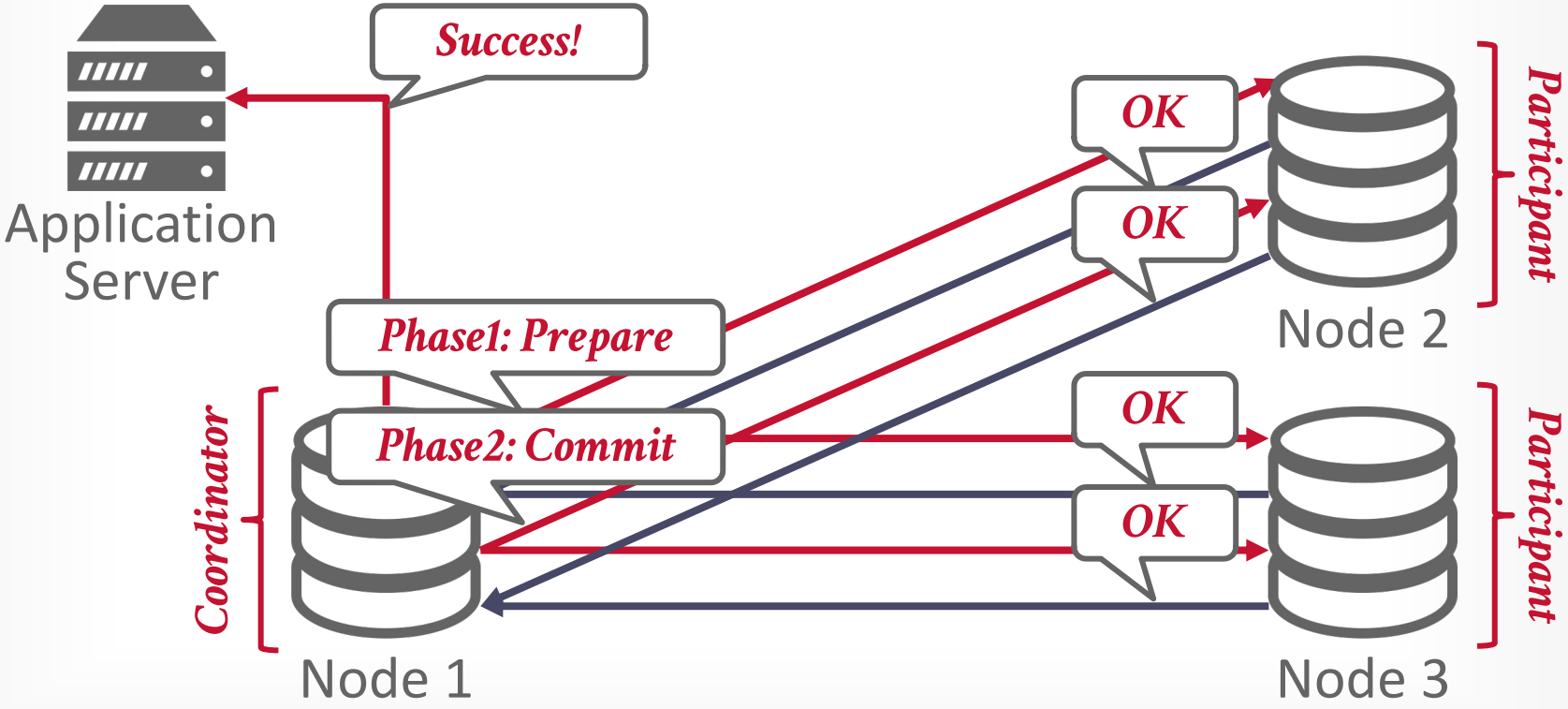
EARLY ACKNOWLEDGEMENT



EARLY ACKNOWLEDGEMENT



EARLY ACKNOWLEDGEMENT



PAXOS

Consensus protocol where a coordinator proposes an outcome (e.g., commit or abort) and then the participants vote on whether that outcome should succeed.

Does not block if a majority of participants are available and has provably minimal message delays in the best case.

The Part-Time Parliament

LESLIE LAMPORT
Digital Equipment Corporation

Recent archaeological discoveries on the island of Paxos reveal that the parliament functioned despite the peripatetic propensity of its part-time legislators. The legislators maintained consistent copies of the parliamentary record, despite their frequent forays from the chamber and the forgetfulness of their messengers. The Paxon parliament's protocol provides a new way of implementing the state-machine approach to the design of distributed systems.

Categories and Subject Descriptors: C2.4 [**Computer-Communications Networks**]: Distributed Systems—*Network operating systems*; D4.5 [**Operating Systems**]: Reliability—*Fault-tolerance*; J.1 [**Administrative Data Processing**]: Government

General Terms: Design, Reliability

Additional Key Words and Phrases: State machines, three-phase commit, voting

This submission was recently discovered behind a filing cabinet in the *TOCS* editorial office. Despite its age, the editor-in-chief felt that it was worth publishing. Because the author is currently doing field work in the Greek isles and cannot be reached, I was asked to prepare it for publication.

The author appears to be an archeologist with only a passing interest in computer science. This is unfortunate; even though the obscure ancient Paxon civilization he describes is of little interest to most computer scientists, its legislative system is an excellent model for how to implement a distributed computer system in an asynchronous environment. Indeed, some of the refinements the Paxons made to their protocol appear to be unknown in the systems literature.

The author does give a brief discussion of the Paxon Parliament's relevance to distributed computing in Section 4. Computer scientists will probably want to read that section first. Even before that, they might want to read the explanation of the algorithm for computer scientists by Lampson [1996]. The algorithm is also described more formally by De Prisco et al. [1997]. I have added further comments on the relation between the ancient protocols and more recent work at the end of Section 4.

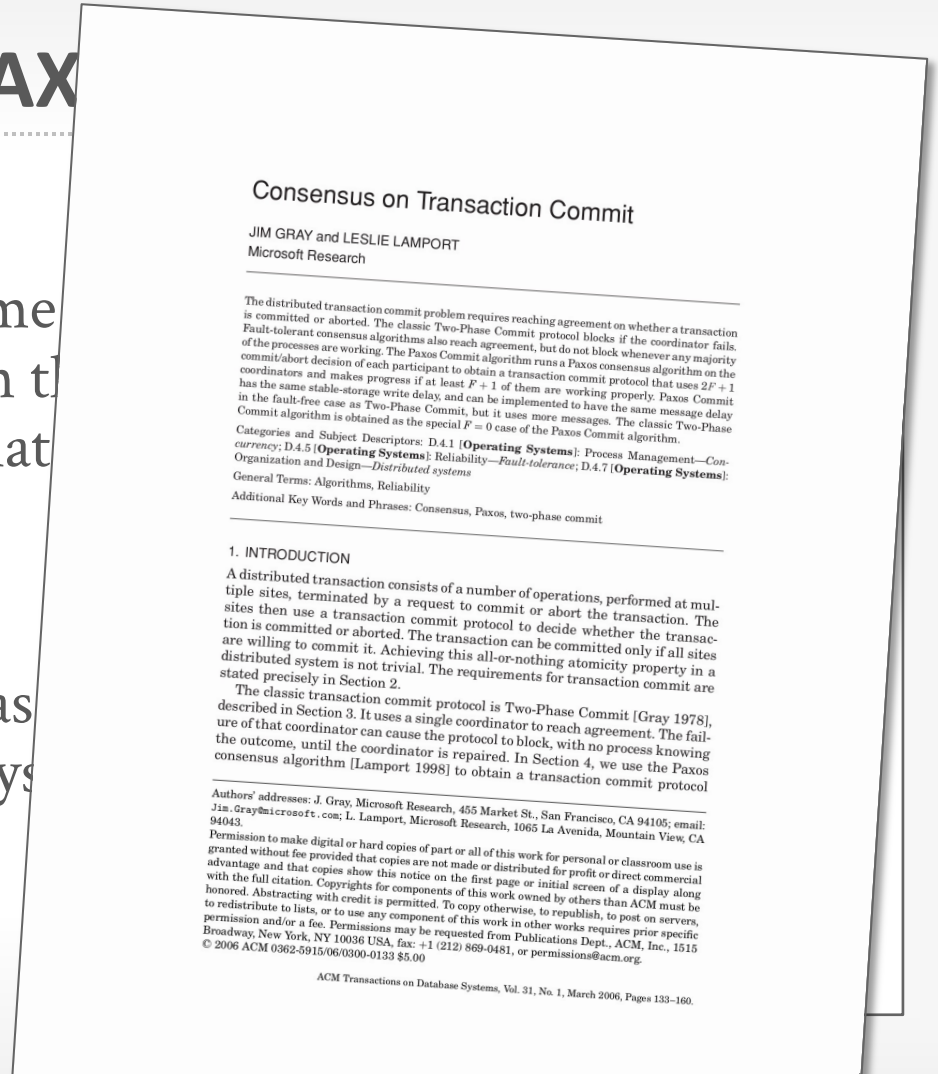
Keith Marzullo
University of California, San Diego

Authors' address: Systems Research Center, Digital Equipment Corporation, 130 Lytton Avenue, Palo Alto, CA 94301.

PAX

Consensus protocol where a coordinator proposes an outcome (e.g., commit or abort) and then t participants vote on whether that outcome should succeed.

Does not block if a majority of participants are available and has provably minimal message delays the best case.



Consensus on Transaction Commit

JIM GRAY and LESLIE LAMPORT
Microsoft Research

The distributed transaction commit problem requires reaching agreement on whether a transaction is committed or aborted. The classic Two-Phase Commit protocol blocks if the coordinator fails. Fault-tolerant consensus algorithms also reach agreement, but do not block whenever any majority of the processes are working. The Paxos Commit algorithm runs a Paxos consensus algorithm on the commit/abort decision of each participant to obtain a transaction commit protocol that uses $2F + 1$ coordinators and makes progress if at least $F + 1$ of them are working properly. Paxos Commit has the same stable-storage write delay, and can be implemented to have the same message delay in the fault-free case as Two-Phase Commit, but it uses more messages. The classic Two-Phase Commit algorithm is obtained as the special $F = 0$ case of the Paxos Commit algorithm.

Categories and Subject Descriptors: D.4.1 [**Operating Systems**]: Process Management—*Concurrency*; D.4.5 [**Operating Systems**]: Reliability—*Fault-tolerance*; D.4.7 [**Operating Systems**]: Organization and Design—*Distributed systems*

General Terms: Algorithms, Reliability

Additional Key Words and Phrases: Consensus, Paxos, two-phase commit

1. INTRODUCTION

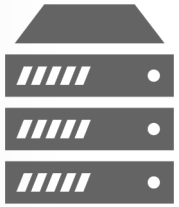
A distributed transaction consists of a number of operations, performed at multiple sites, terminated by a request to commit or abort the transaction. The sites then use a transaction commit protocol to decide whether the transaction is committed or aborted. The transaction can be committed only if all sites are willing to commit it. Achieving this all-or-nothing atomicity property in a distributed system is not trivial. The requirements for transaction commit are stated precisely in Section 2.

The classic transaction commit protocol is Two-Phase Commit [Gray 1978], described in Section 3. It uses a single coordinator to reach agreement. The failure of that coordinator can cause the protocol to block, with no process knowing the outcome, until the coordinator is repaired. In Section 4, we use the Paxos consensus algorithm [Lamport 1998] to obtain a transaction commit protocol

Authors' addresses: J. Gray, Microsoft Research, 455 Market St., San Francisco, CA 94105; email: Jim.Gray@microsoft.com; L. Lamport, Microsoft Research, 1065 La Avenida, Mountain View, CA 94043.

ACM Transactions on Database Systems, Vol. 31, No. 1, March 2006, Pages 133–160.

PAXOS



Application Server

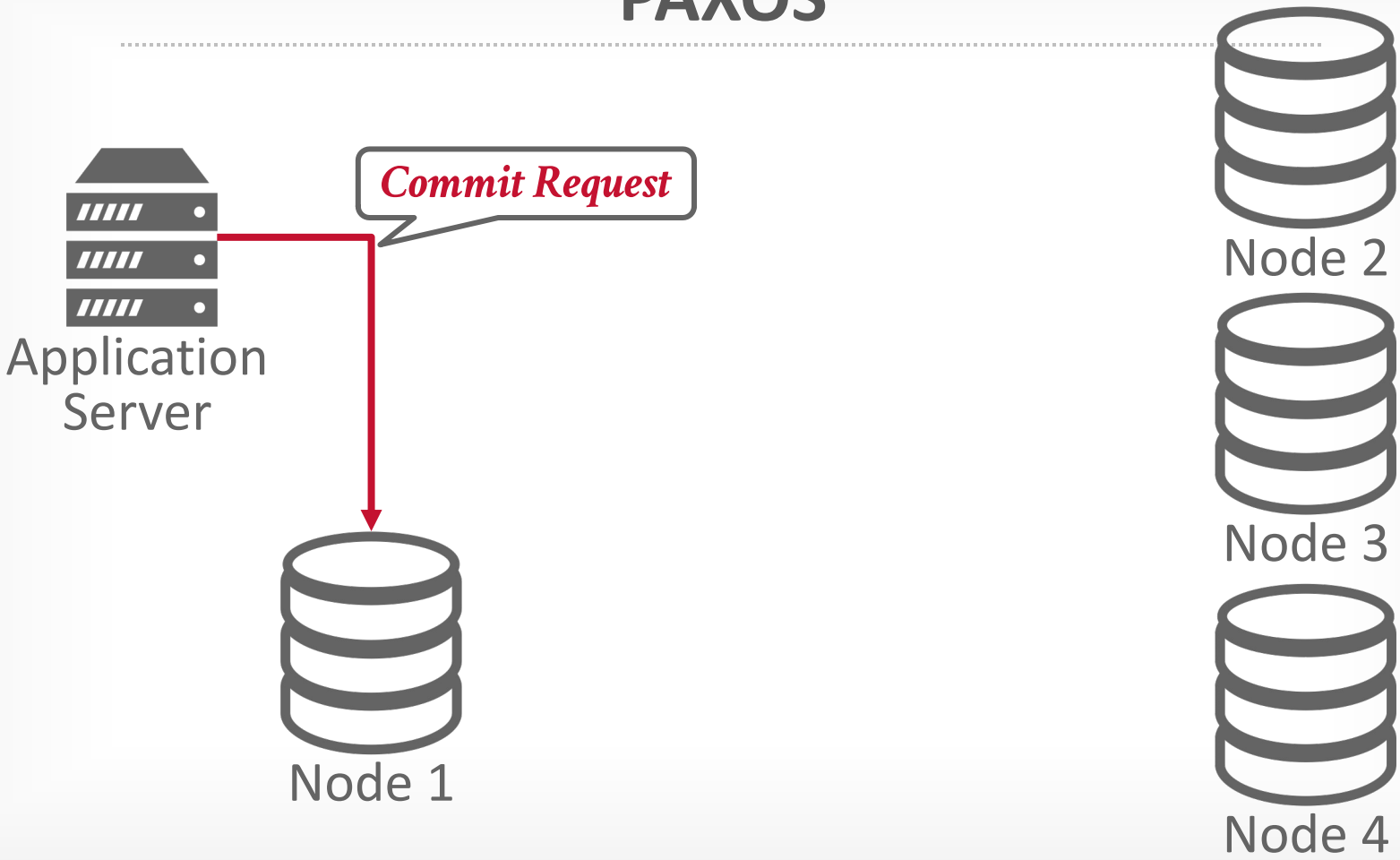
Node 1

Node 2

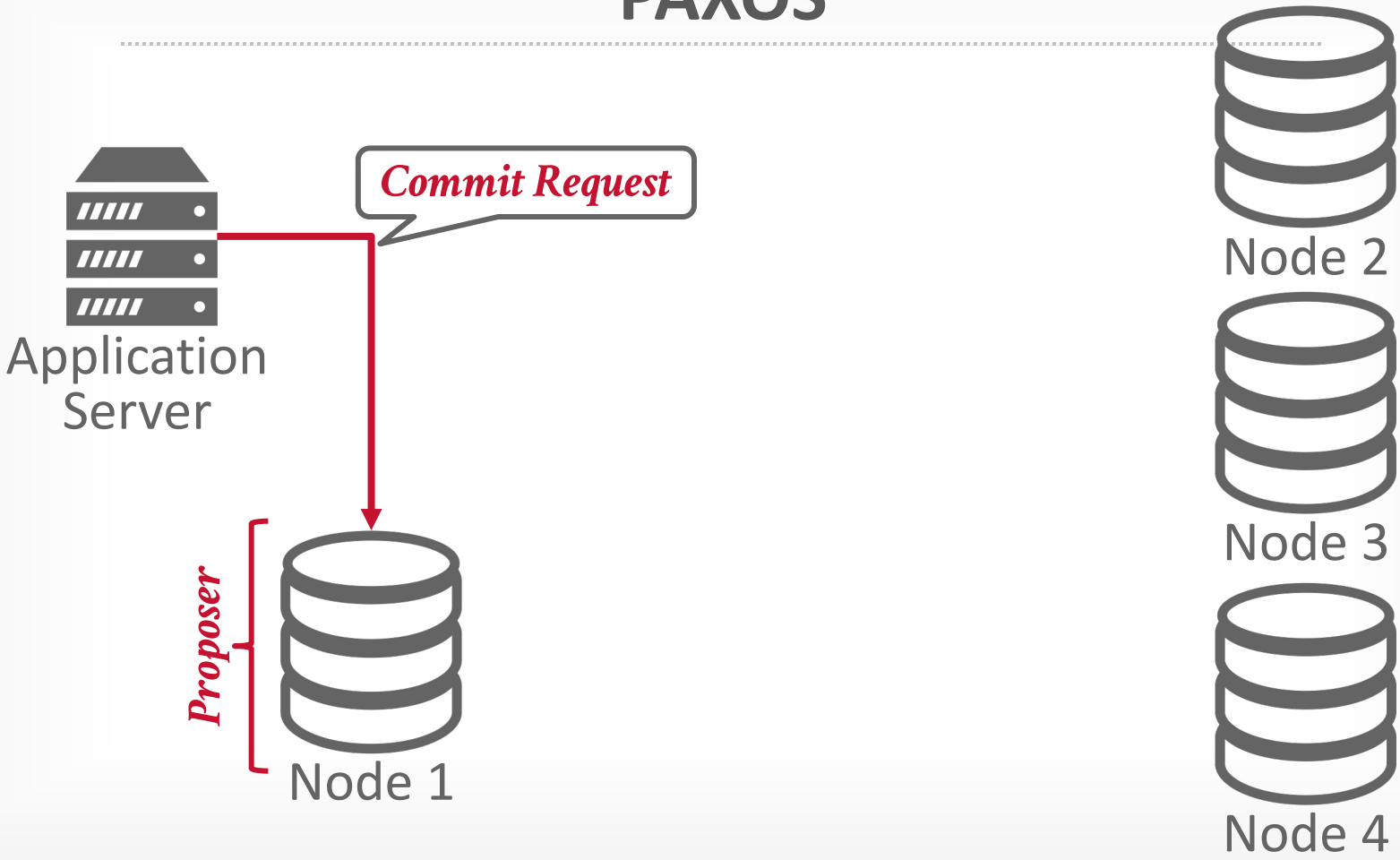
Node 3

Node 4

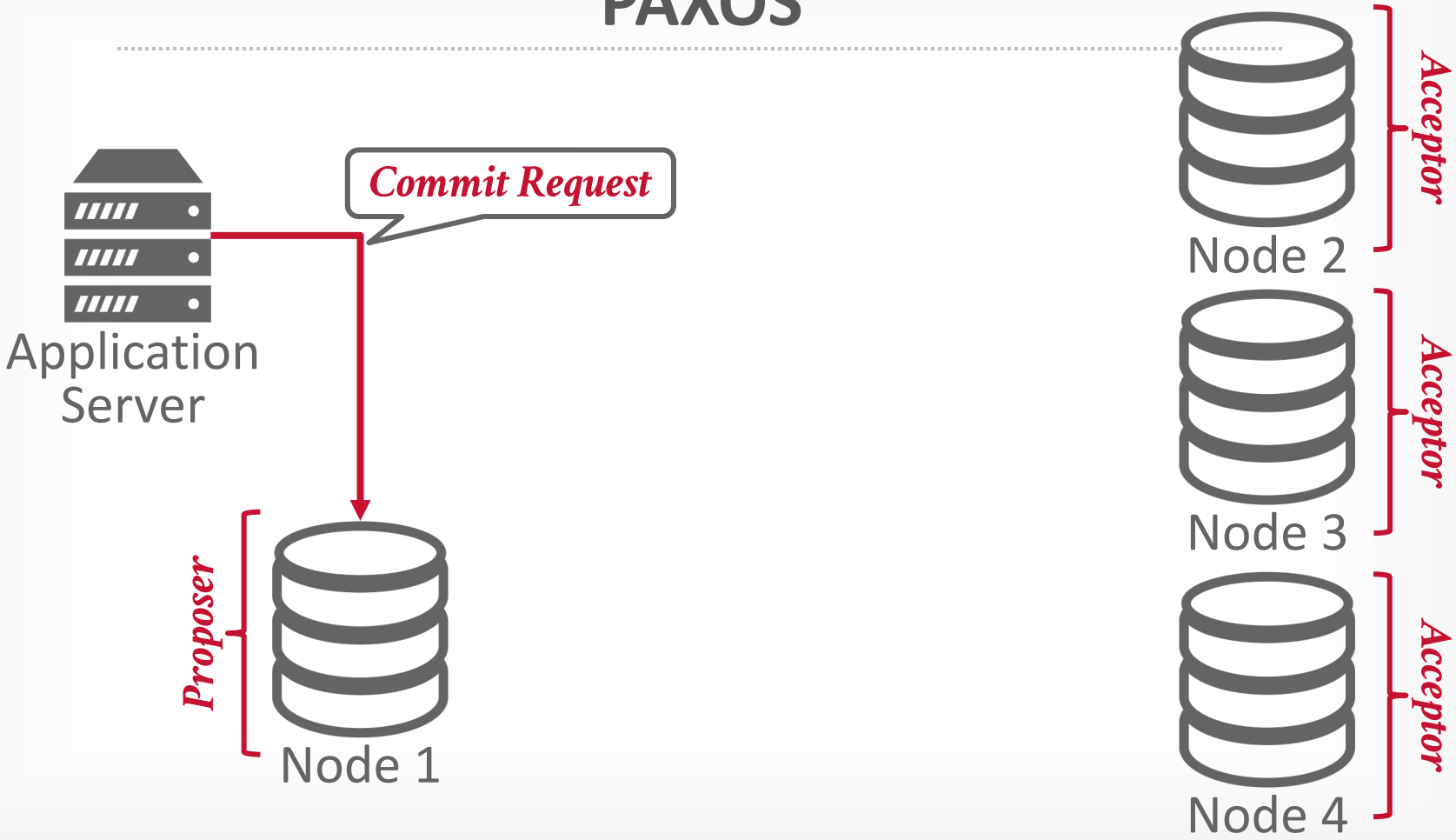
PAXOS


Commit Request
Application Server
Node 1
Node 2
Node 3
Node 4

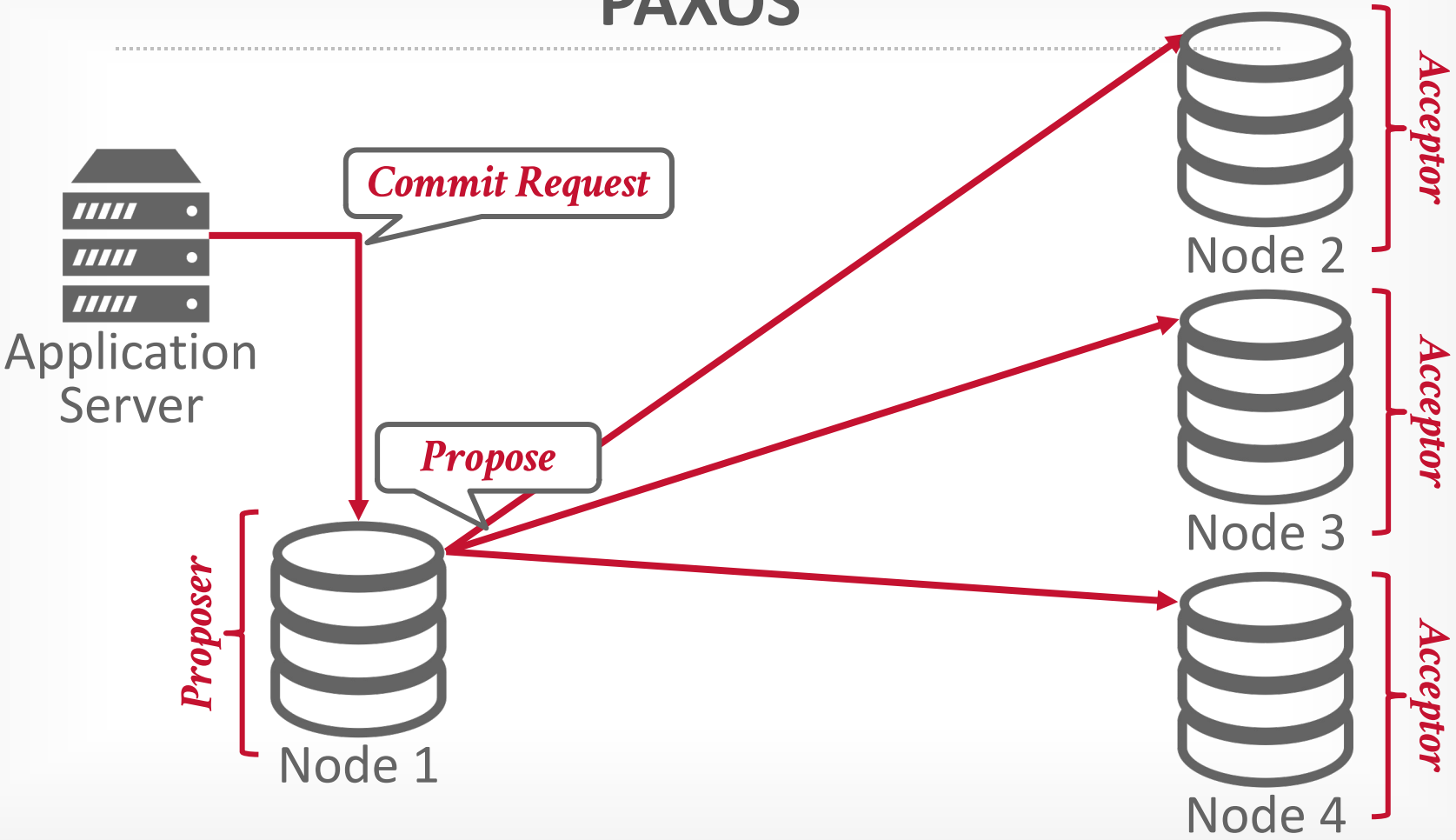
PAXOS


Commit Request
Application Server
Proposer
Node 1
Node 2
Node 3
Node 4

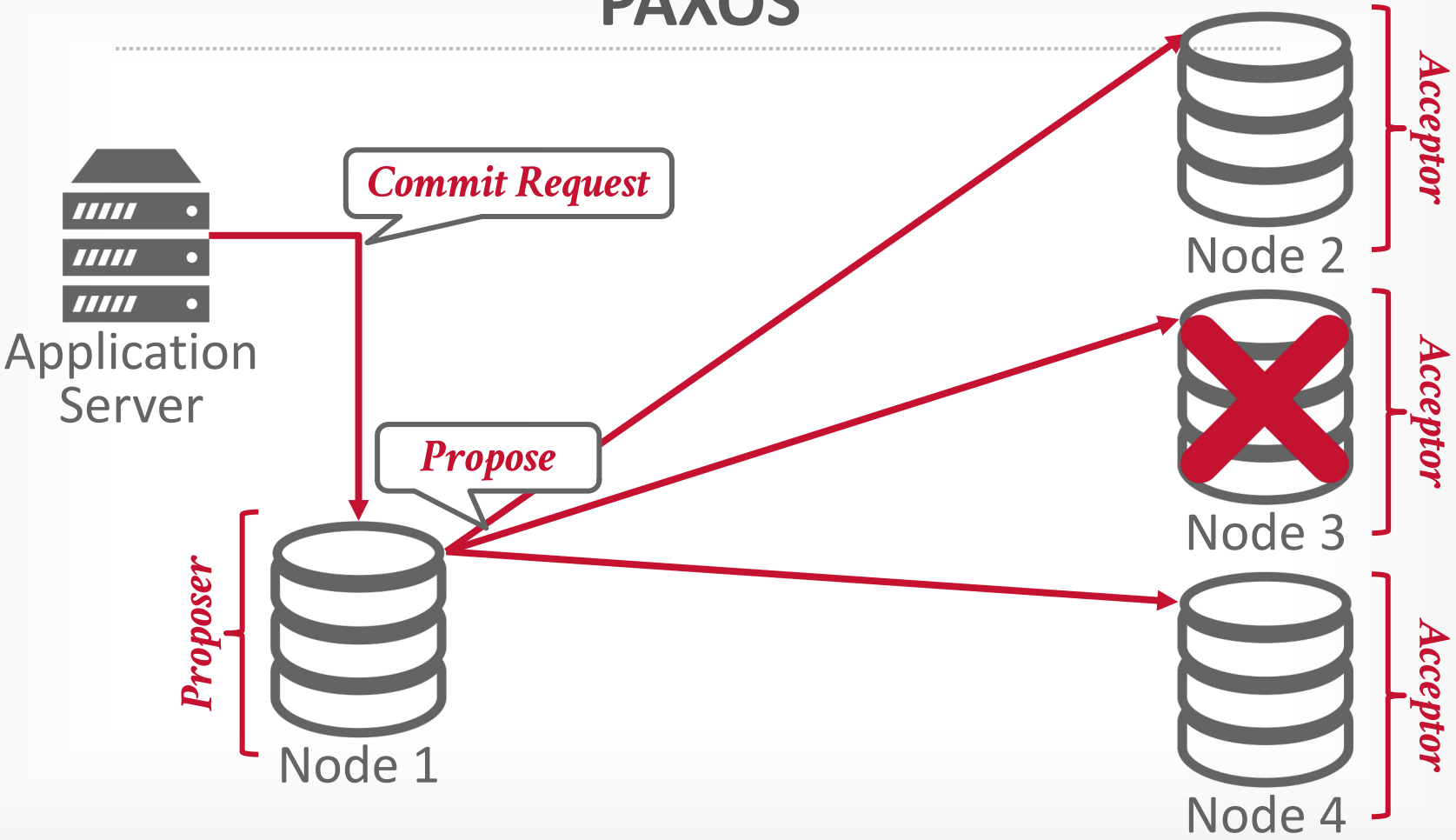
PAXOS


Commit Request
Application Server
Proposer
Node 1
Acceptor
Node 2
Acceptor
Node 3
Acceptor
Node 4

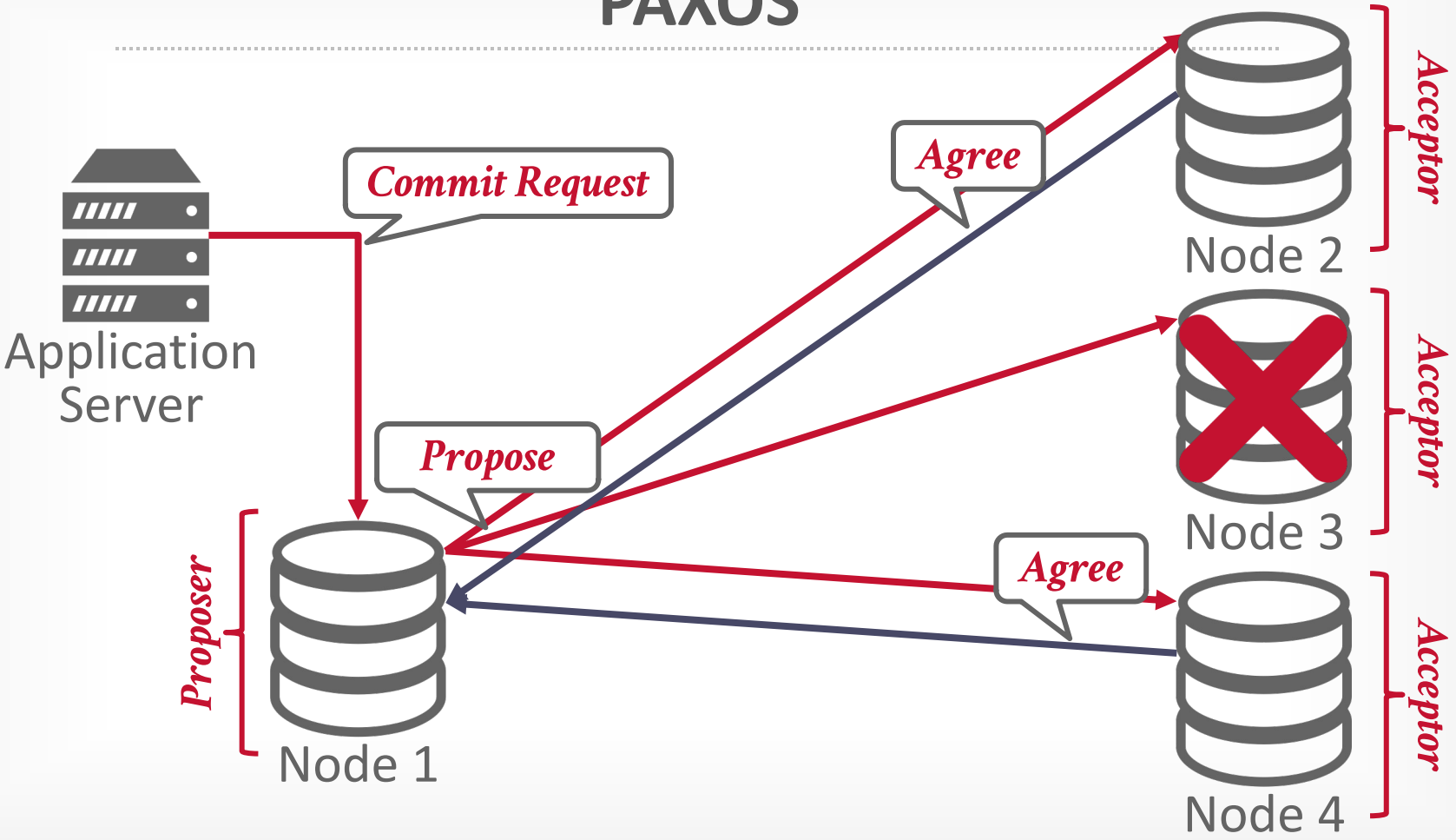
PAXOS


Commit Request
Application Server
Propose
Proposer
Node 1
Node 2
Node 3
Node 4
Acceptor
Acceptor
Acceptor

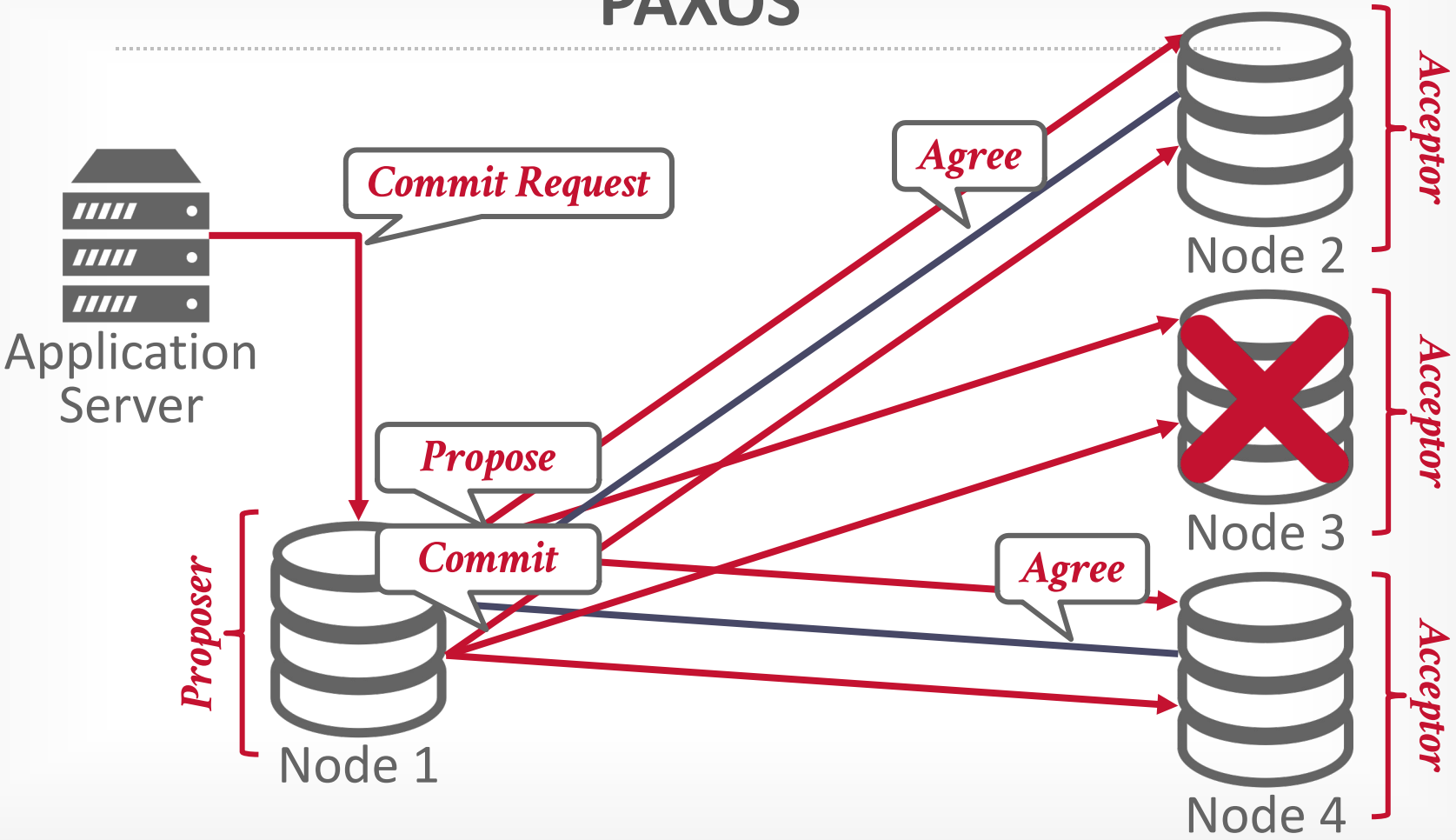
PAXOS


Commit Request
Application Server
Propose
Proposer
Node 1
Node 2
Node 3
Node 4
Acceptor
Acceptor
Acceptor

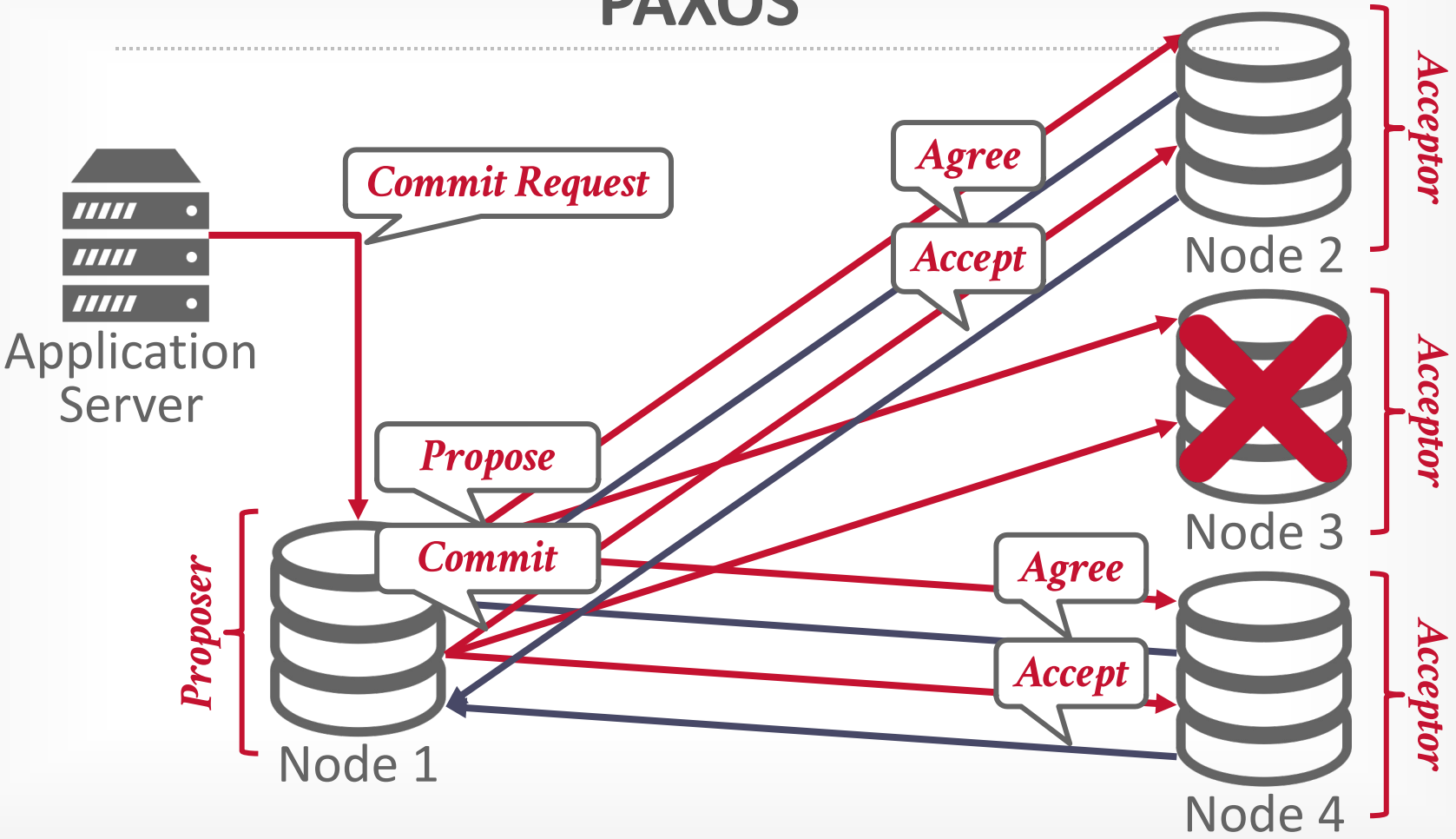
PAXOS


Application Server
Commit Request
Propose
Agree
Agree
Proposer
Node 1
Node 2
Node 3
Node 4
Acceptor
Acceptor
Acceptor

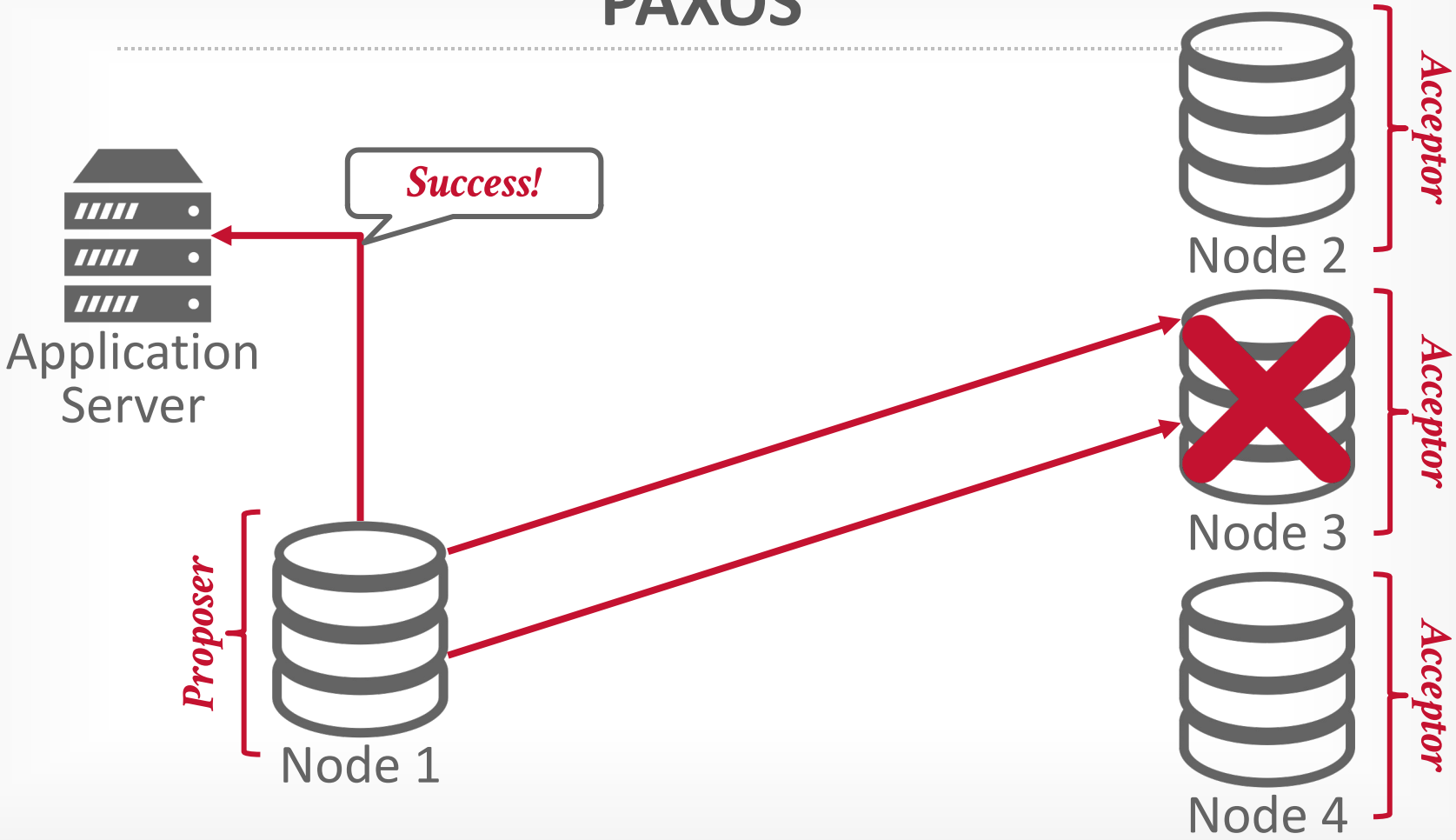
PAXOS


Application Server
Commit Request
Propose
Commit
Agree
Agree
Proposer
Node 1
Node 2
Node 3
Node 4
Acceptor
Acceptor
Acceptor

PAXOS


Application Server
Commit Request
Propose
Commit
Proposer
Node 1
Agree
Accept
Node 2
Acceptor
Node 3
Acceptor
Agree
Accept
Node 4
Acceptor

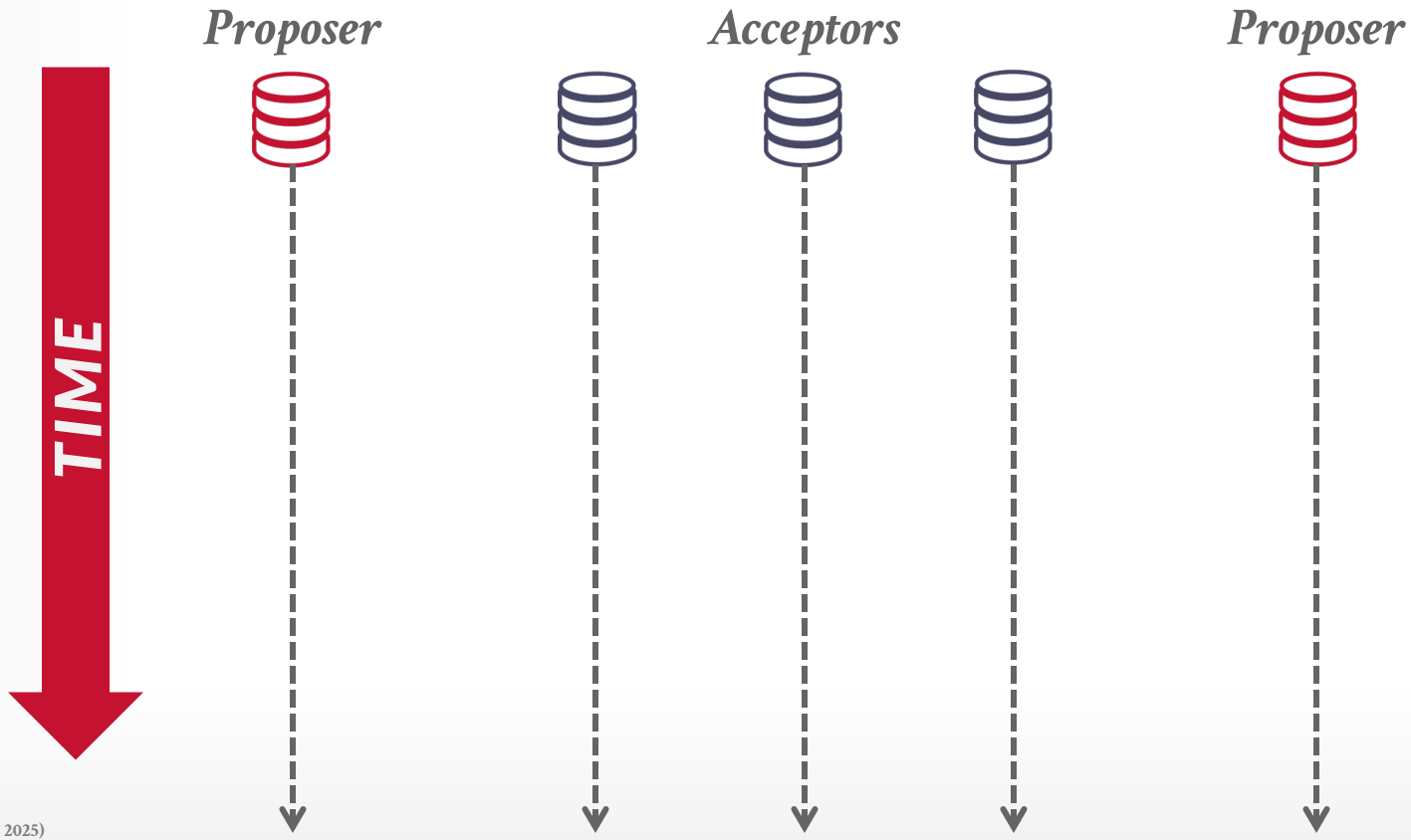
PAXOS


Success!
Application Server
Proposer
Node 1
Node 2
Node 3
Node 4
Acceptor
Acceptor
Acceptor

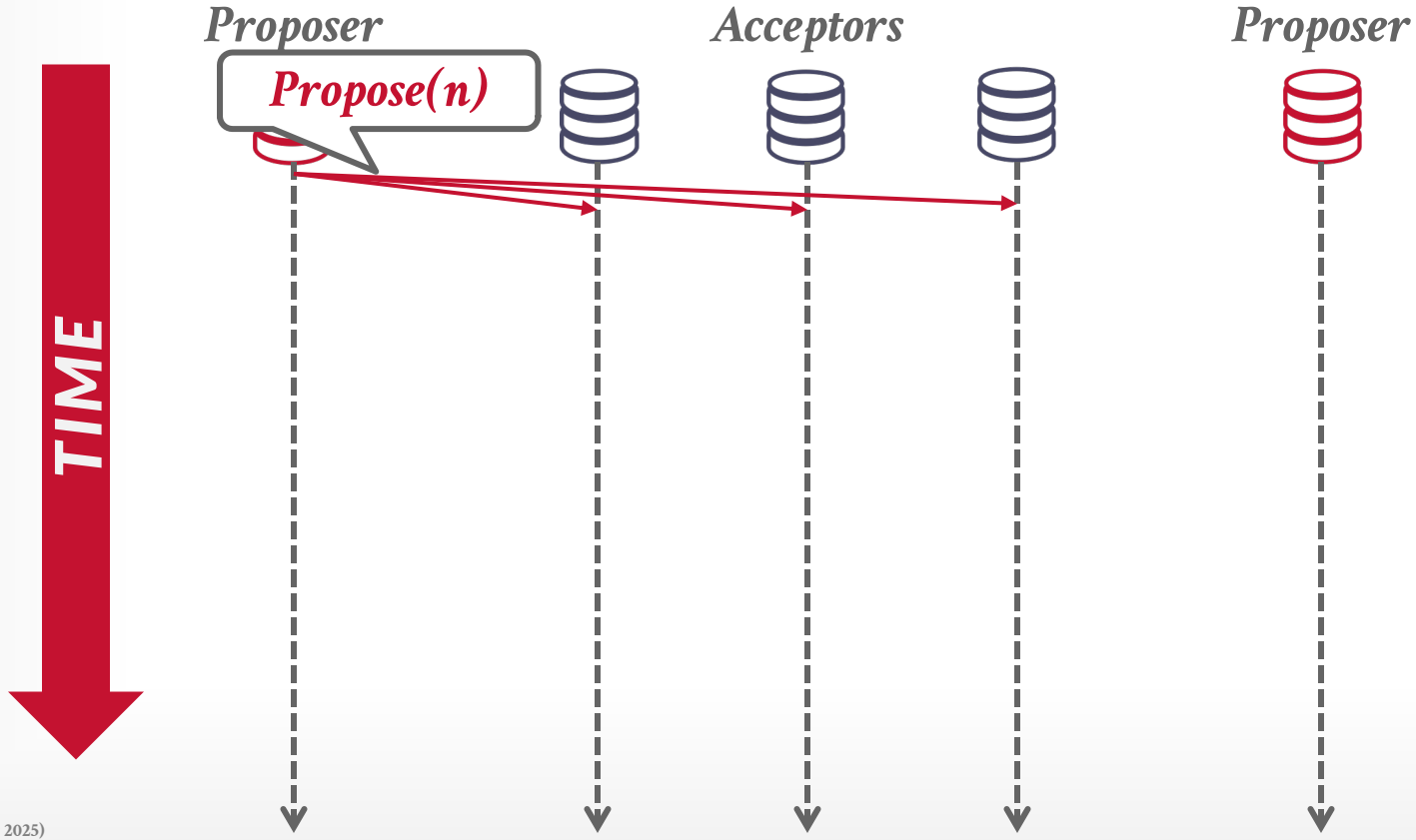
PAXOS


Success!
Application Server
Proposer
Node 1
Node 2
Node 3
Node 4
Acceptor
Acceptor
Acceptor

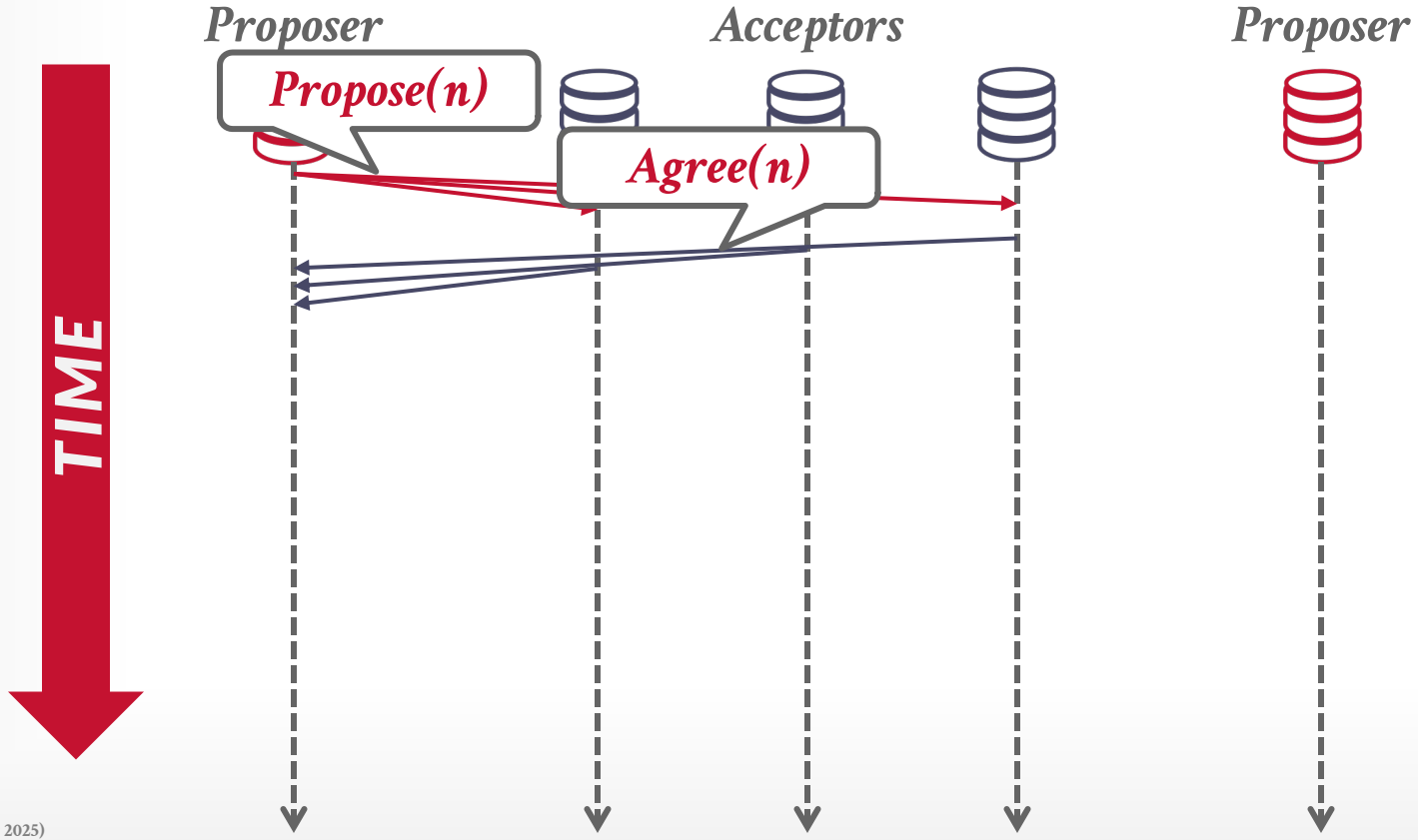
PAXOS



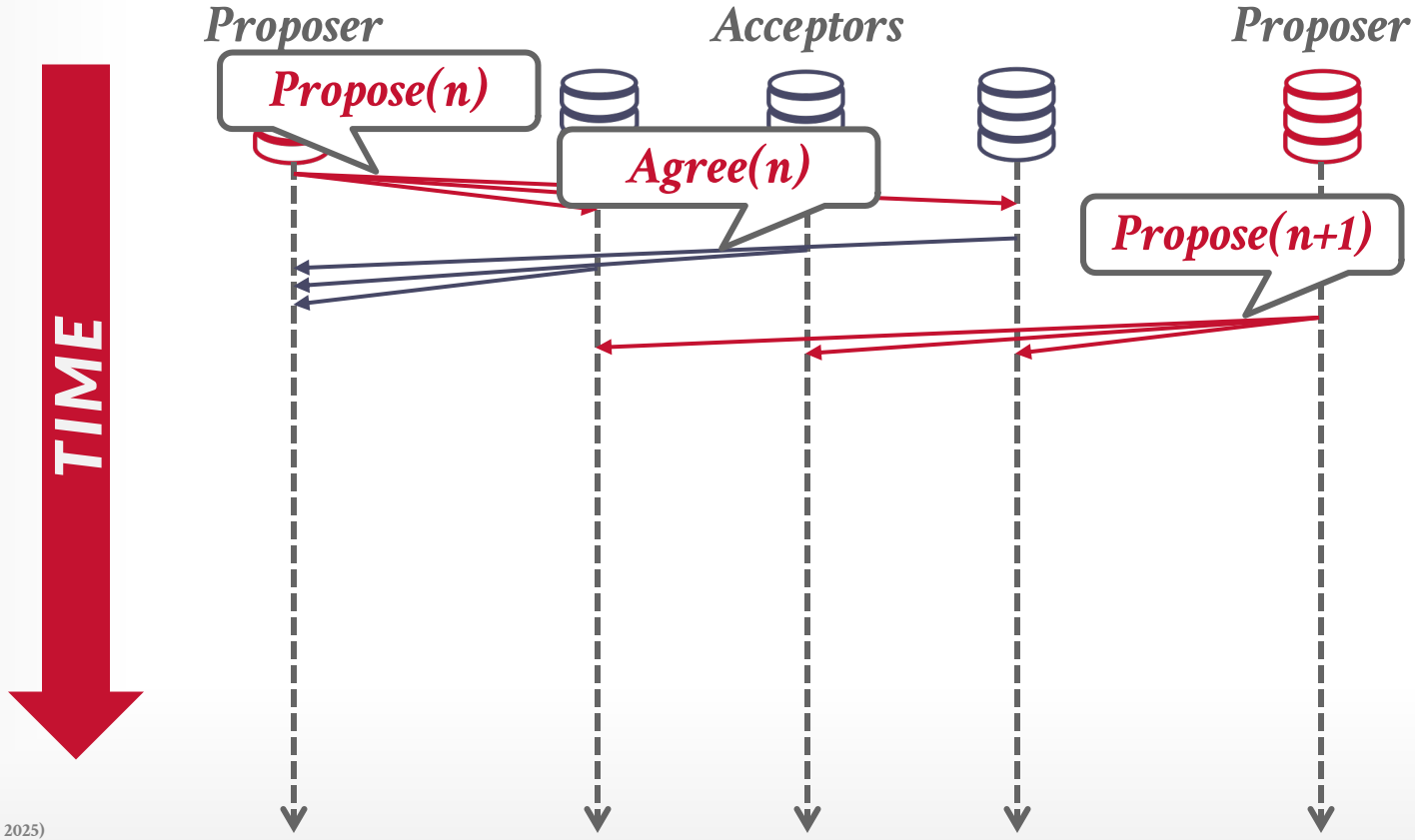
PAXOS



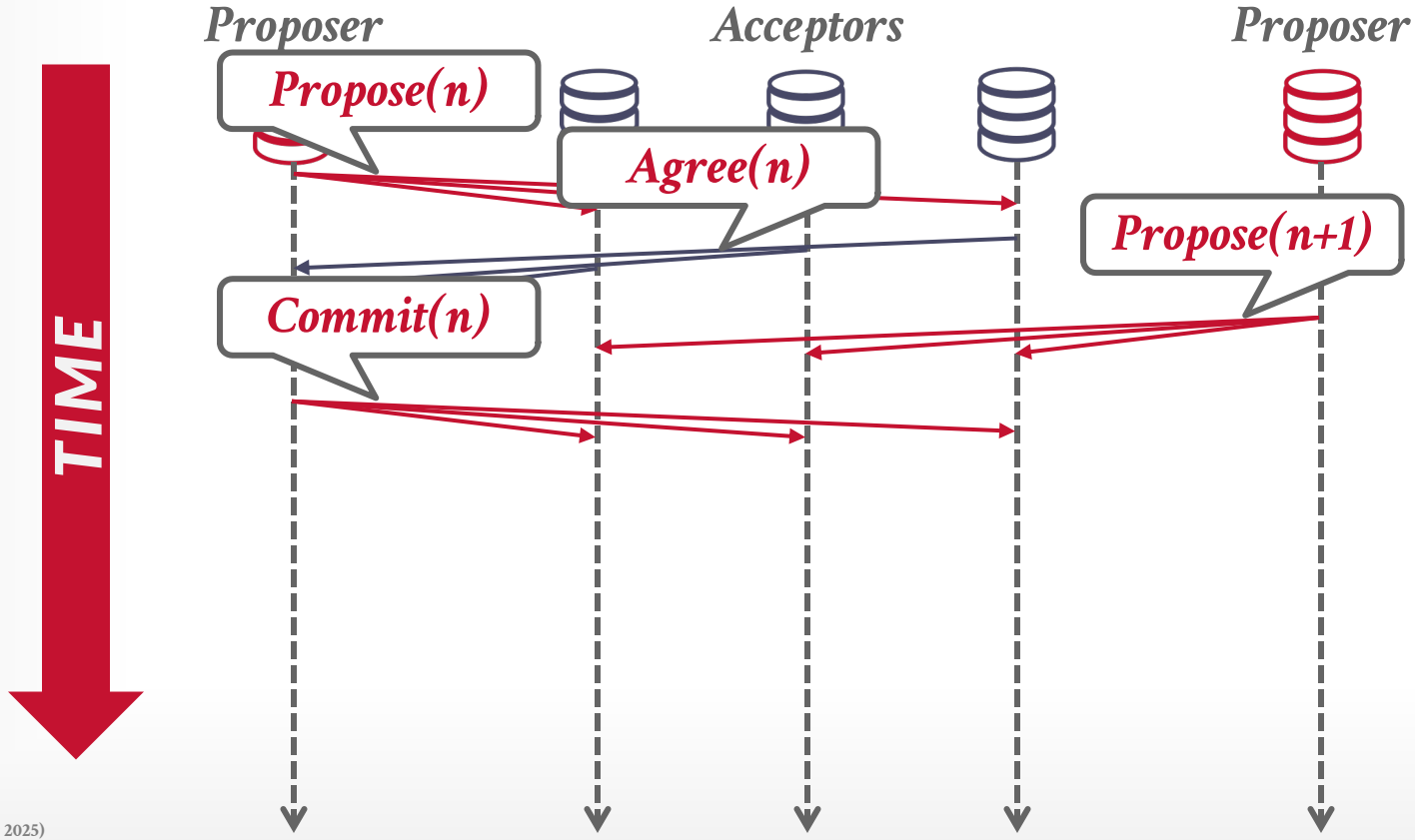
PAXOS



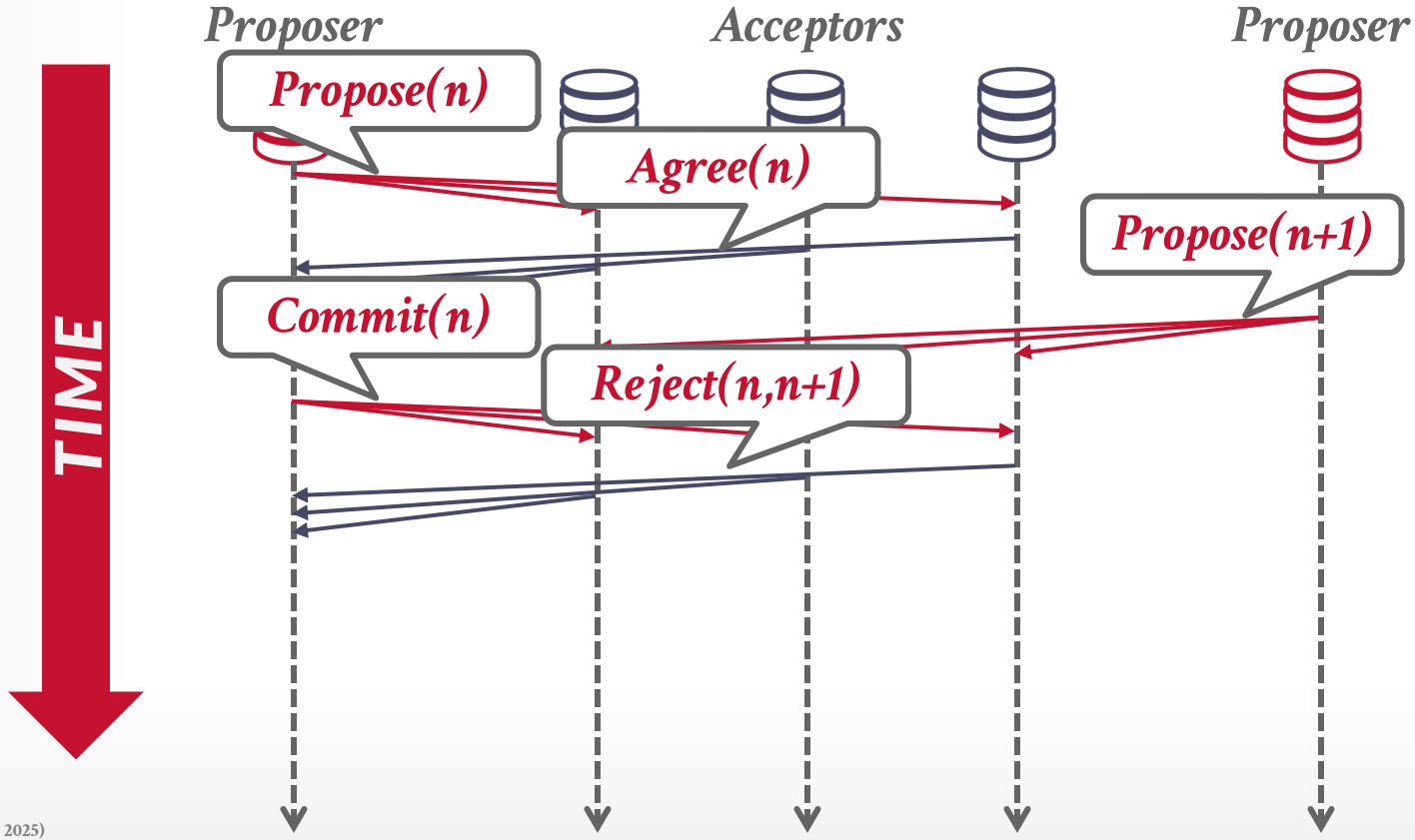
PAXOS


Proposer
Acceptors
Proposer
Propose(n)
Agree(n)
Propose(n+1)
TIME

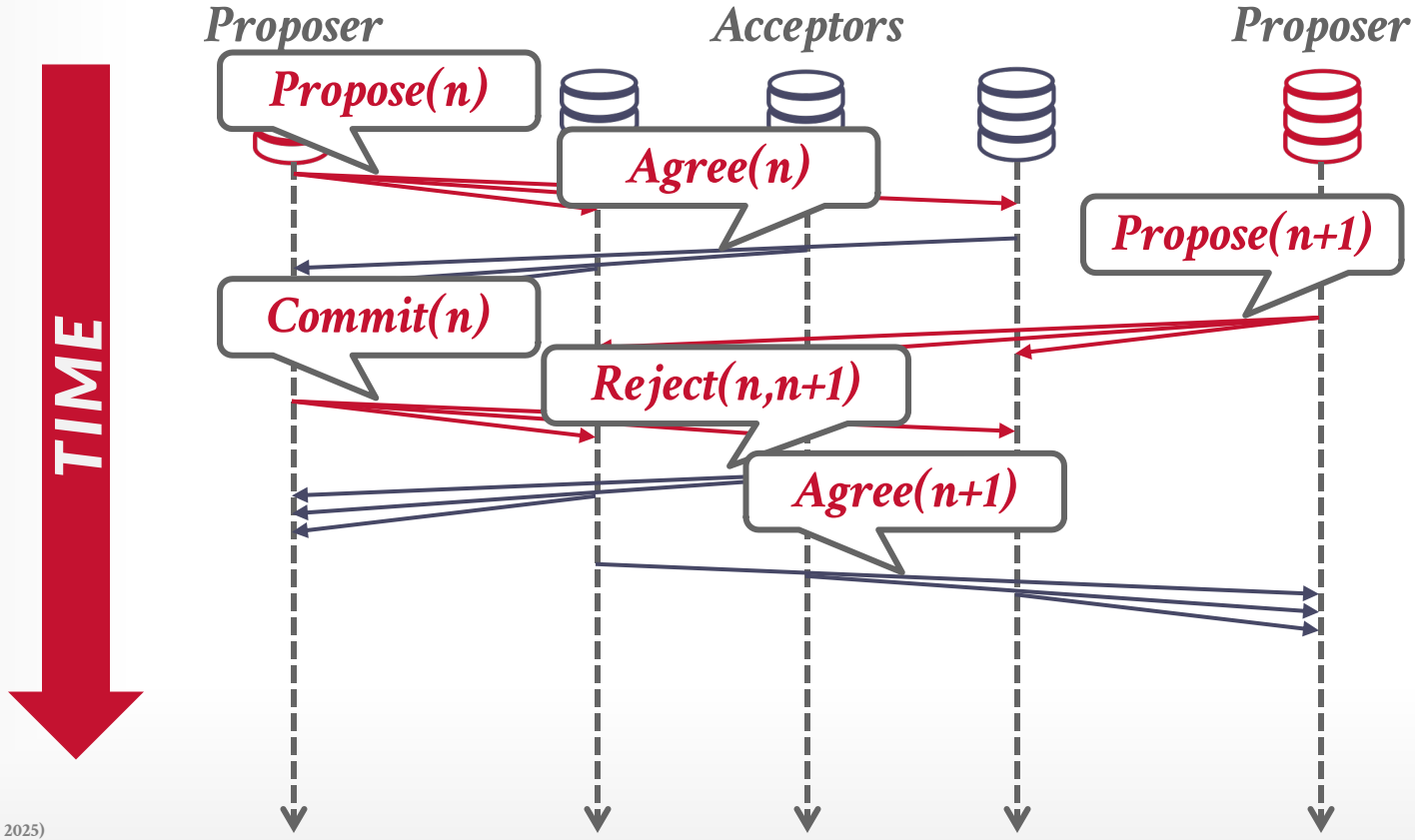
PAXOS


Proposer
Acceptors
Proposer
Propose(n)
Agree(n)
Propose(n+1)
Commit(n)
TIME

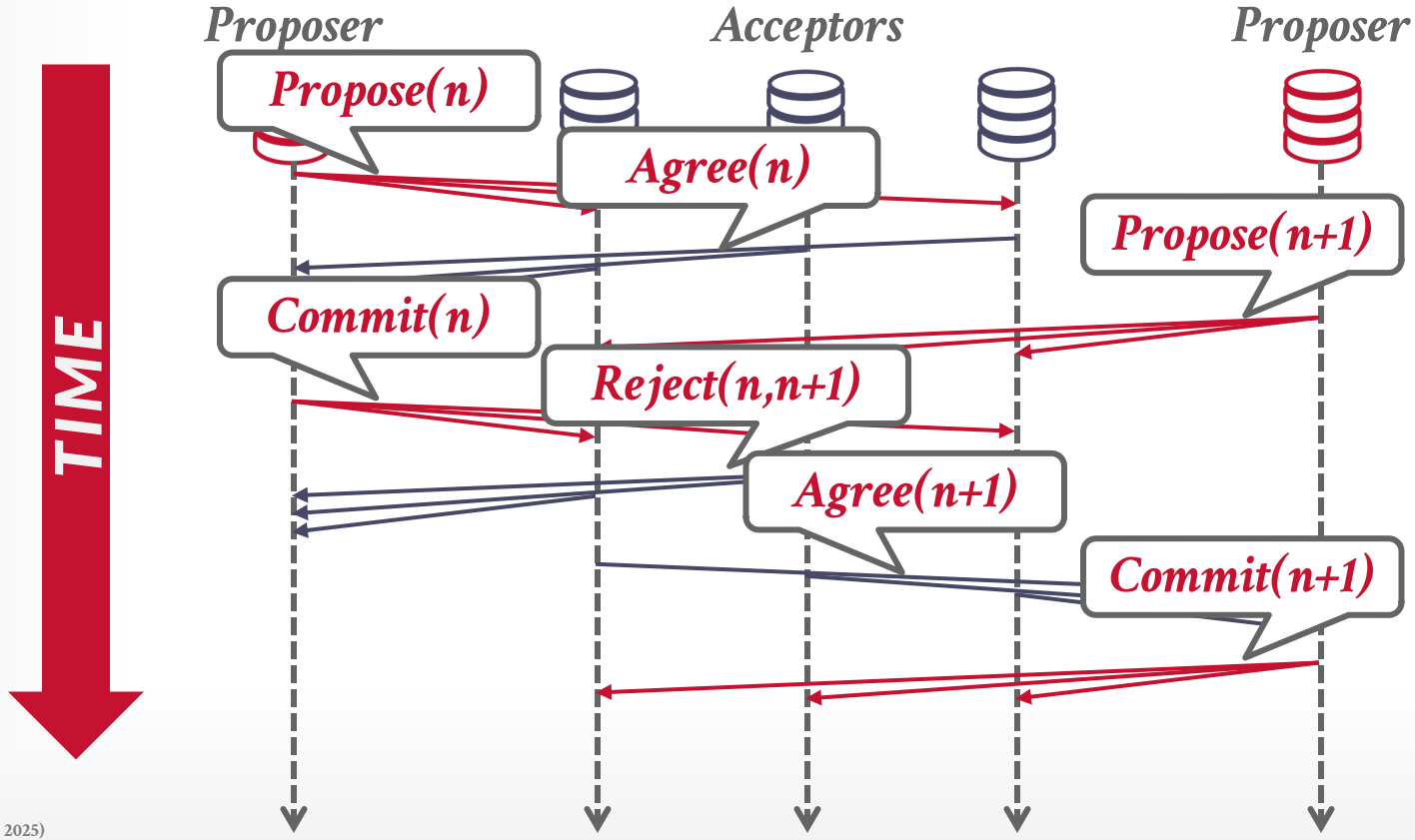
PAXOS


Proposer
Acceptors
Proposer
Propose(n)
Agree(n)
Propose(n+1)
Commit(n)
Reject(n,n+1)
TIME

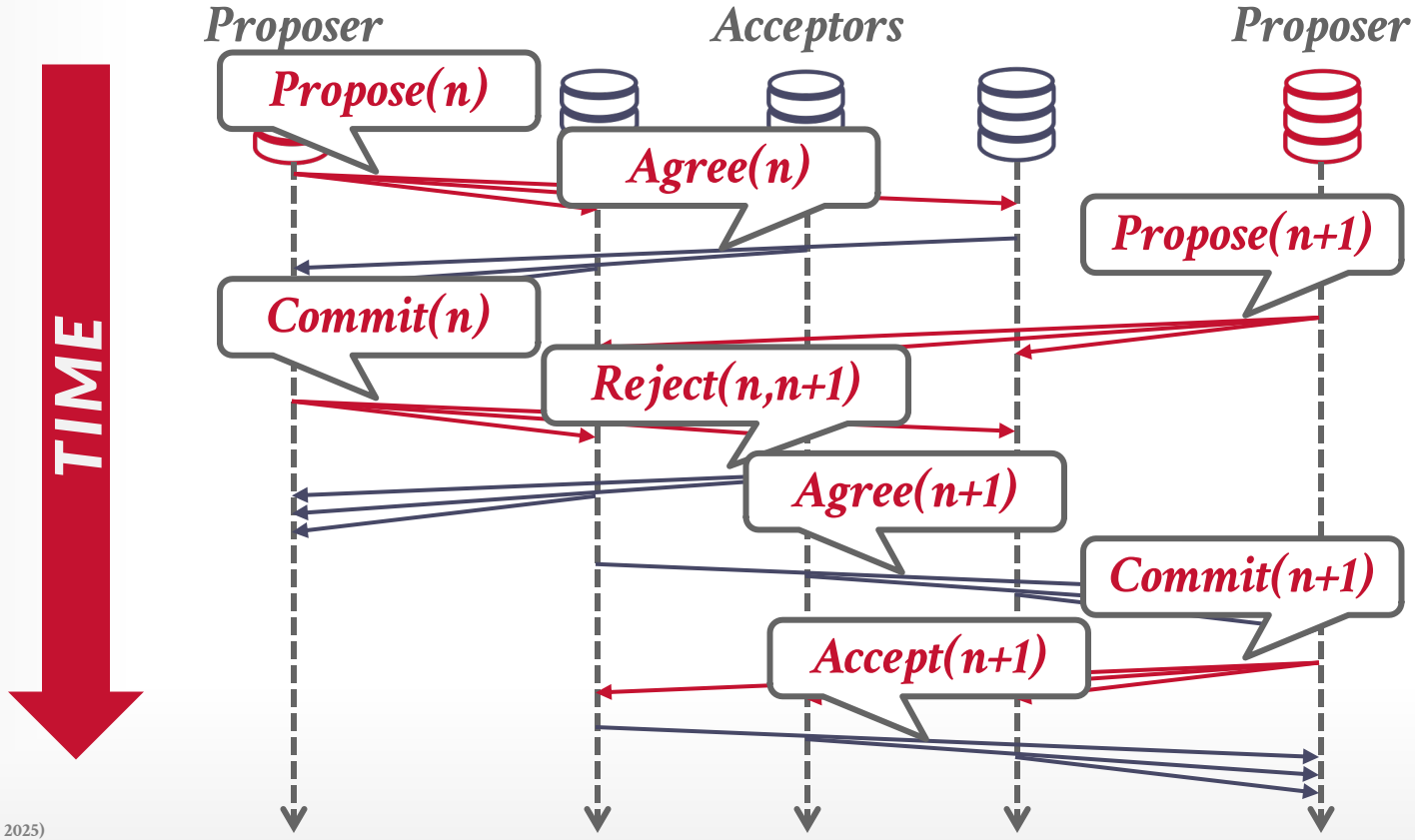
PAXOS


Proposer
Acceptors
Proposer
Propose(n)
Agree(n)
Propose(n+1)
Commit(n)
Reject(n,n+1)
Agree(n+1)
TIME

PAXOS


Proposer
Acceptors
Proposer
Propose(n)
Agree(n)
Propose(n+1)
Commit(n)
Reject(n,n+1)
Agree(n+1)
Commit(n+1)
TIME

PAXOS


Proposer
Acceptors
Proposer
TIME
Propose(n)
Agree(n)
Propose(n+1)
Commit(n)
Reject(n,n+1)
Agree(n+1)
Commit(n+1)
Accept(n+1)

MULTI-PAXOS

If the system elects a single leader that oversees proposing changes for some period, then it can skip the **Propose** phase.

→ Fall back to full Paxos whenever there is a failure.

The system periodically renews the leader (known as a *lease*) using another Paxos round.

→ Nodes must exchange log entries during leader election to make sure that everyone is up-to-date.

2PC VS. PAXOS VS. RAFT

Two-Phase Commit

→ Blocks if coordinator fails after the prepare message is sent, until coordinator recovers.

Paxos

→ Non-blocking if a majority participants are alive, provided there is a sufficiently long period without further failures.

Raft:

→ Similar to Paxos but with fewer node types.
→ Only nodes with most up-to-date log can become leaders.

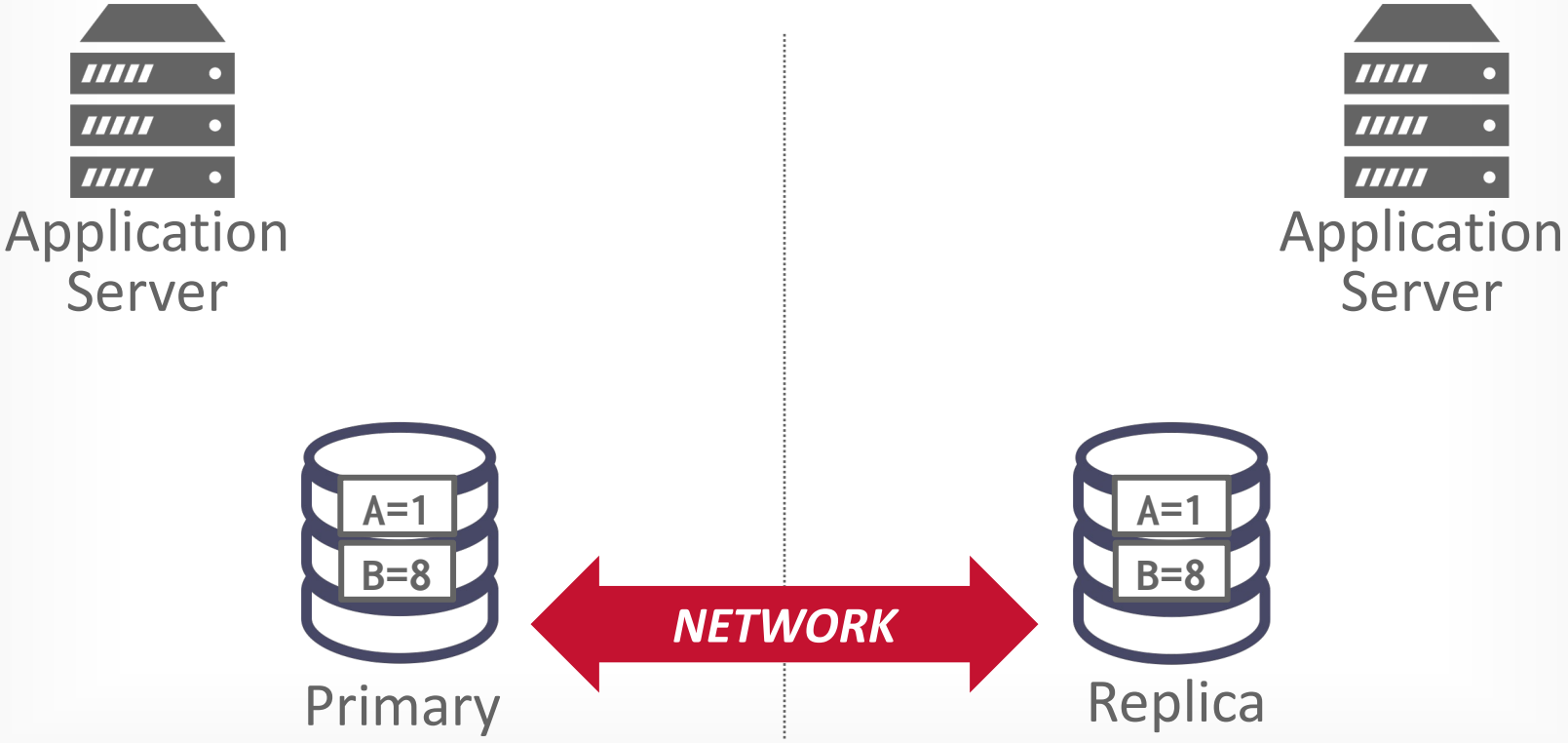
CAP THEOREM

Proposed in the late 1990s that is impossible for a distributed database to always be:

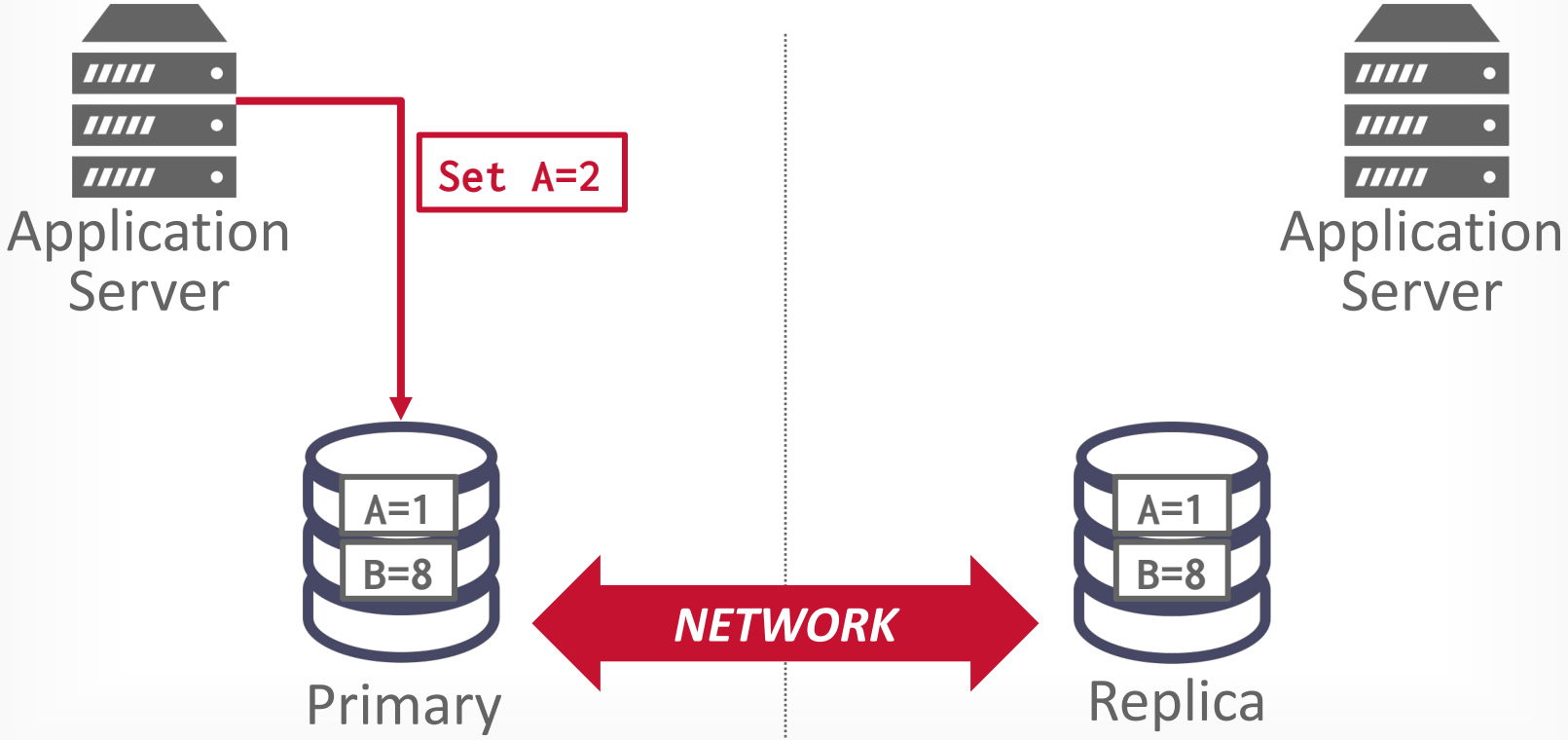
→ **C**onsistent
→ **A**lways Available
→ **N**etwork Partition Tolerant

Whether a DBMS provides **C**onsistency or **A**vailability during a **N**etwork partition.

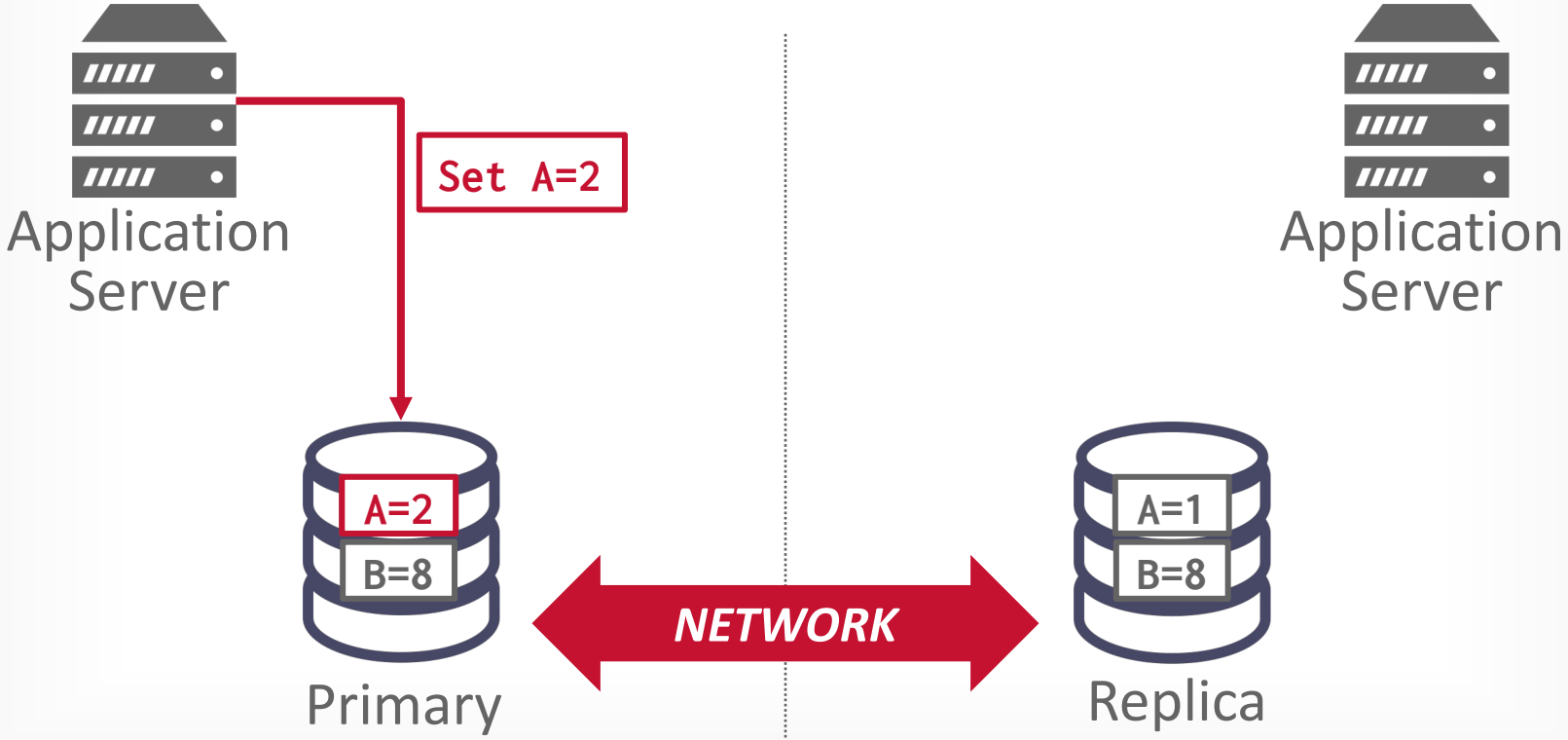
CONSISTENCY



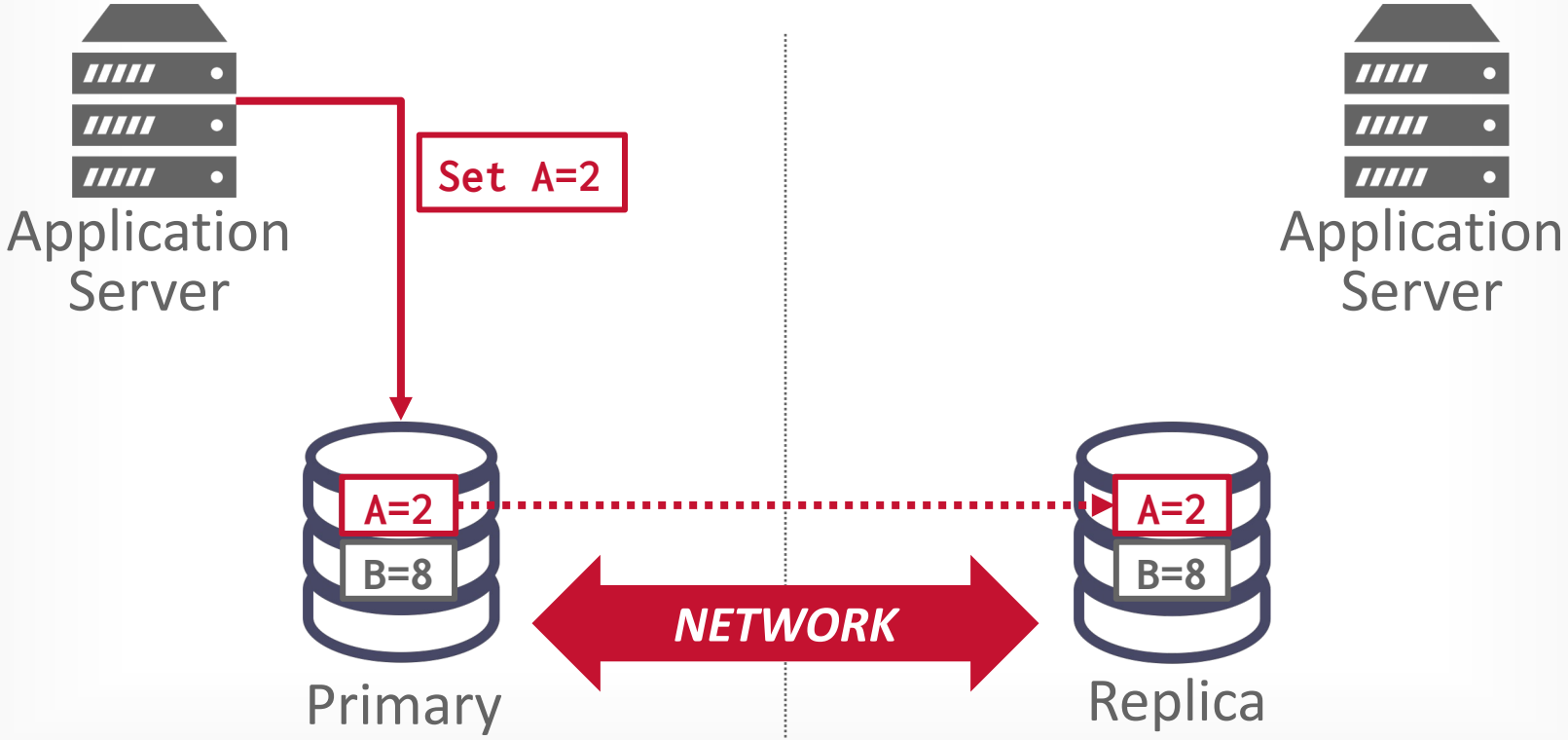
CONSISTENCY



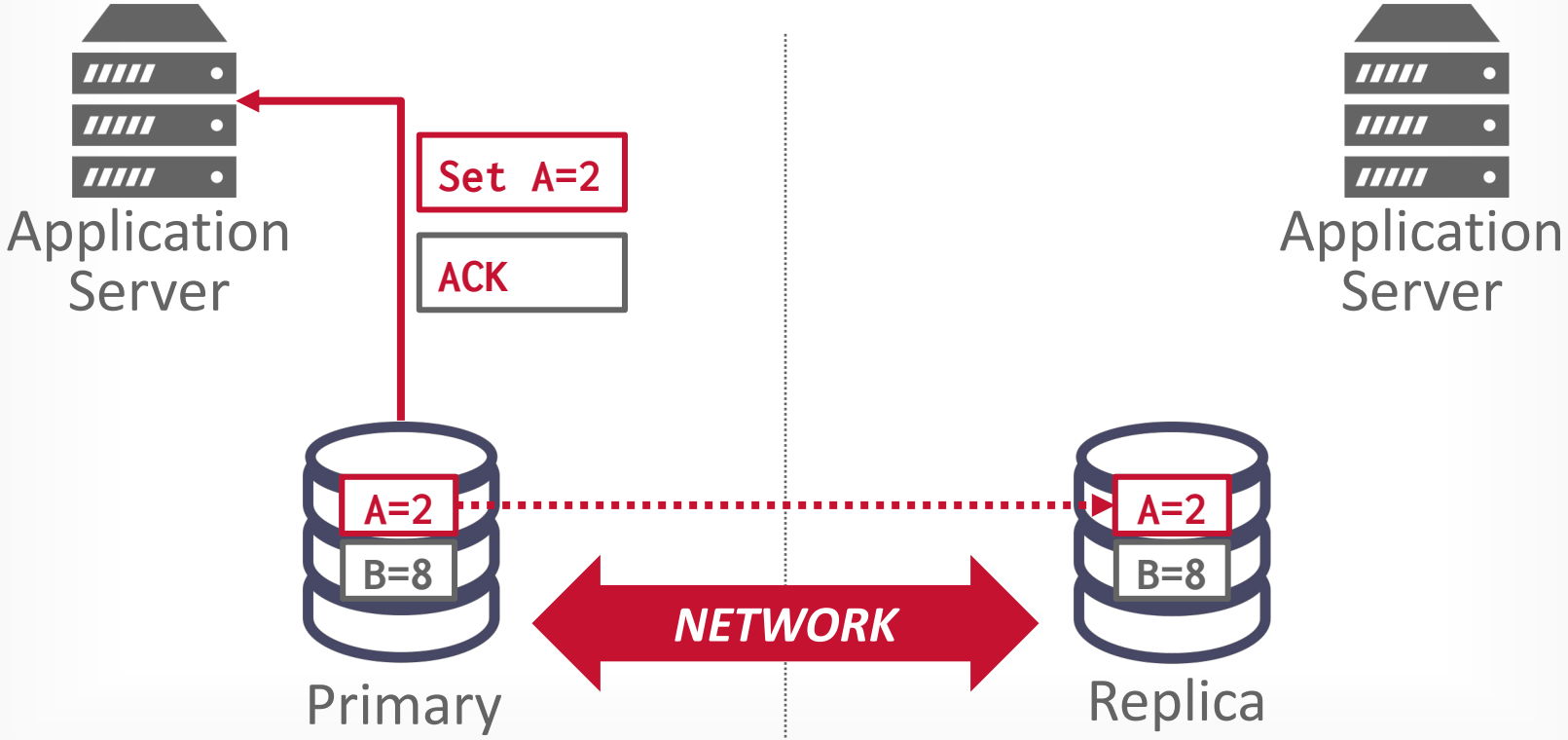
CONSISTENCY



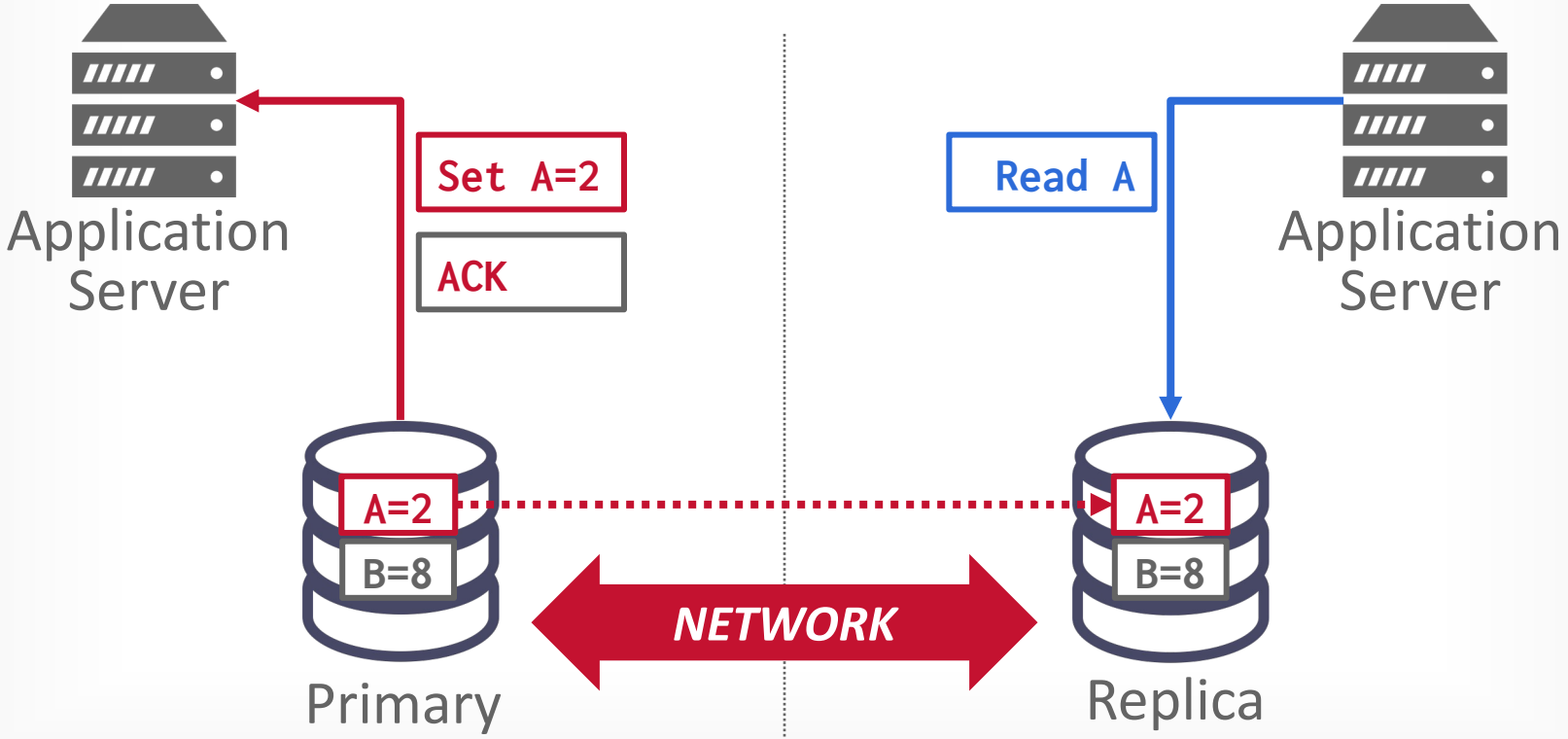
CONSISTENCY



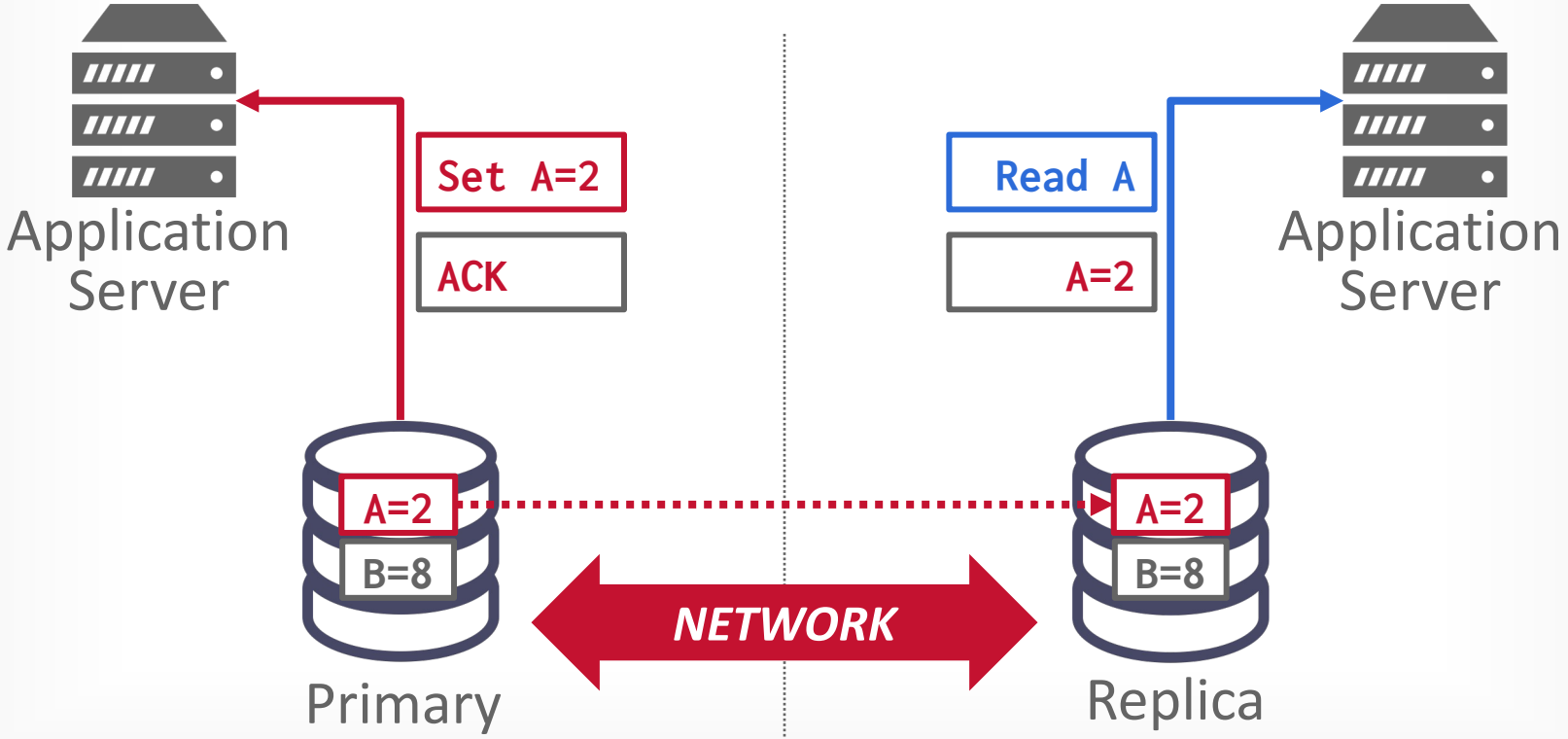
CONSISTENCY



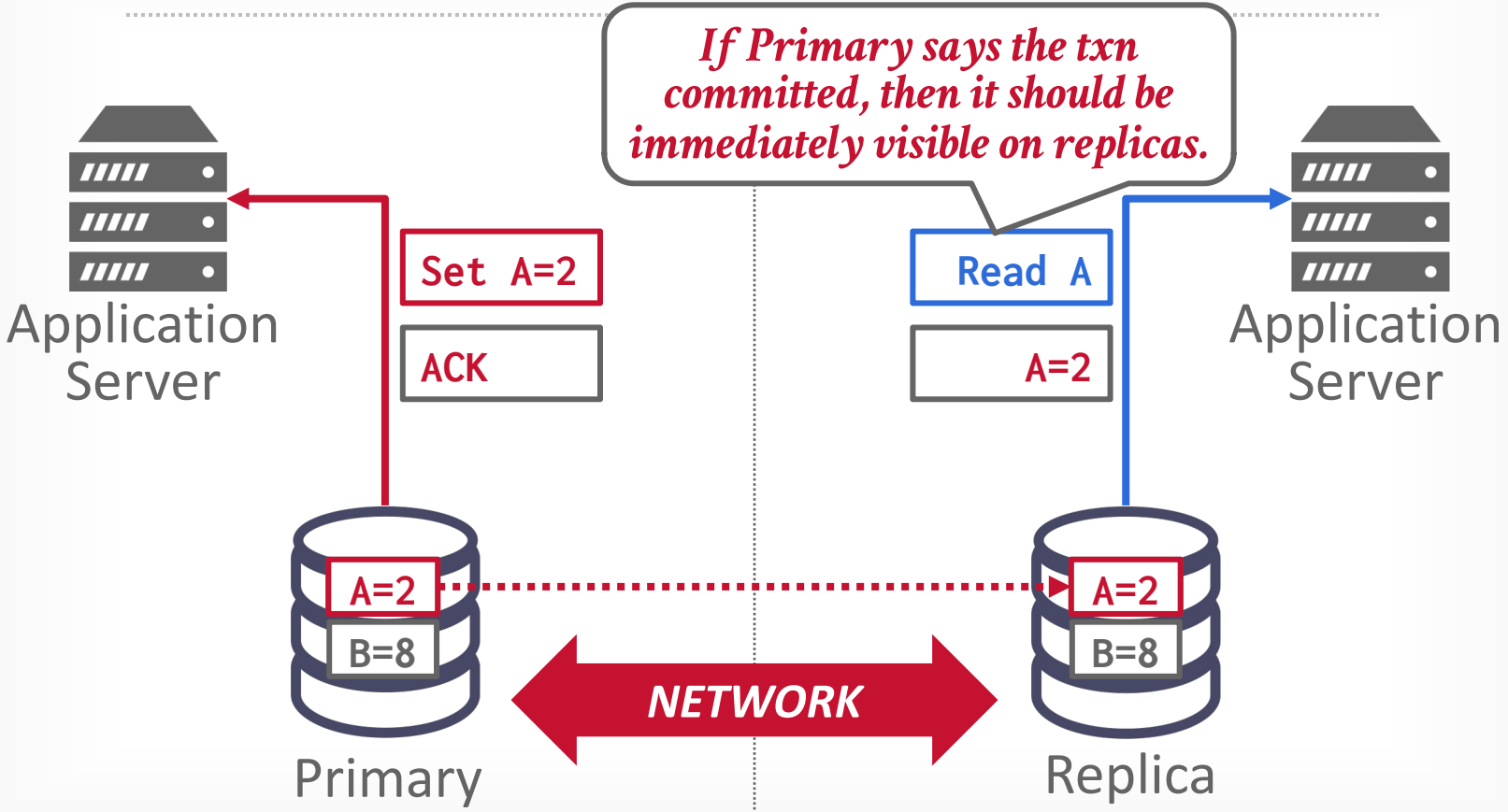
CONSISTENCY



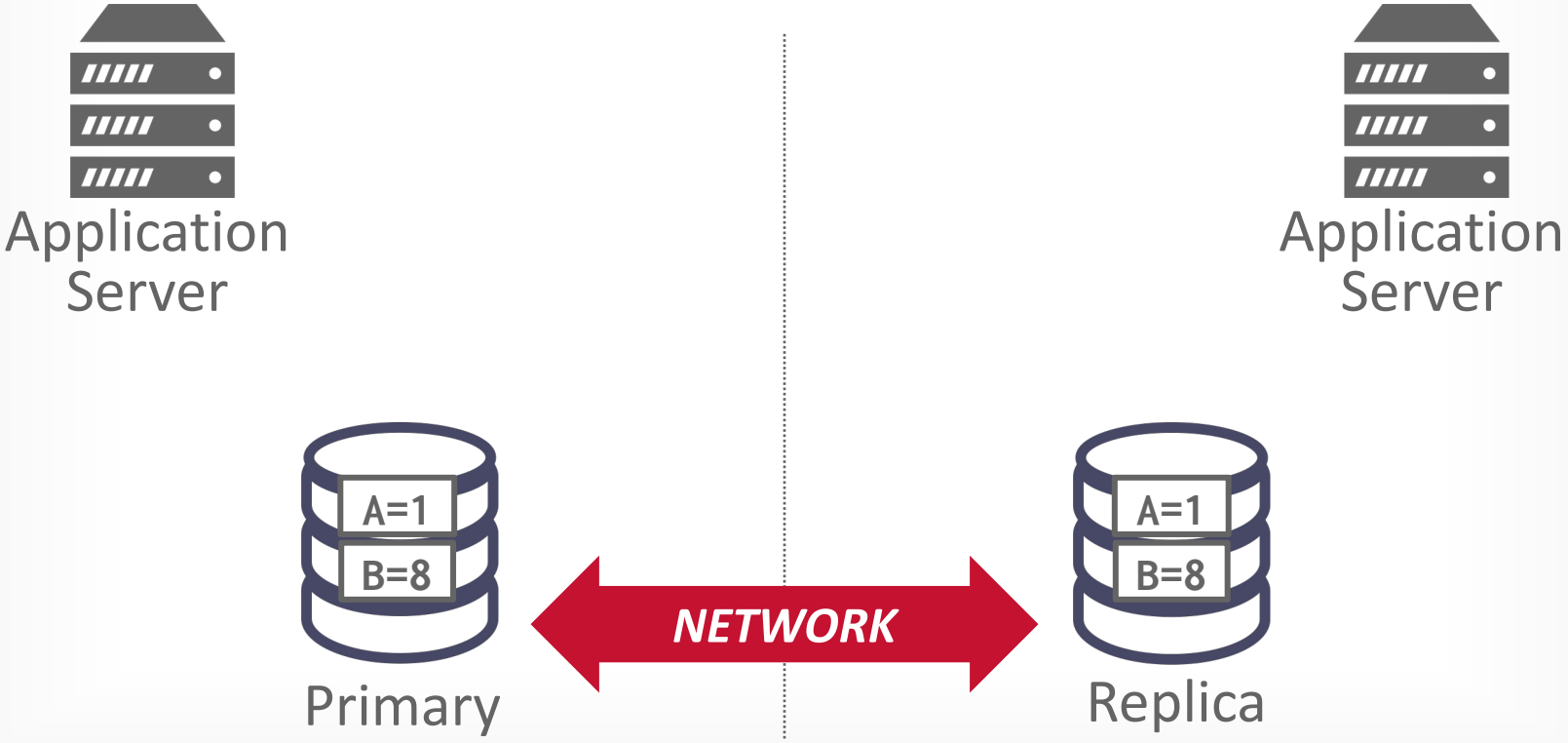
CONSISTENCY



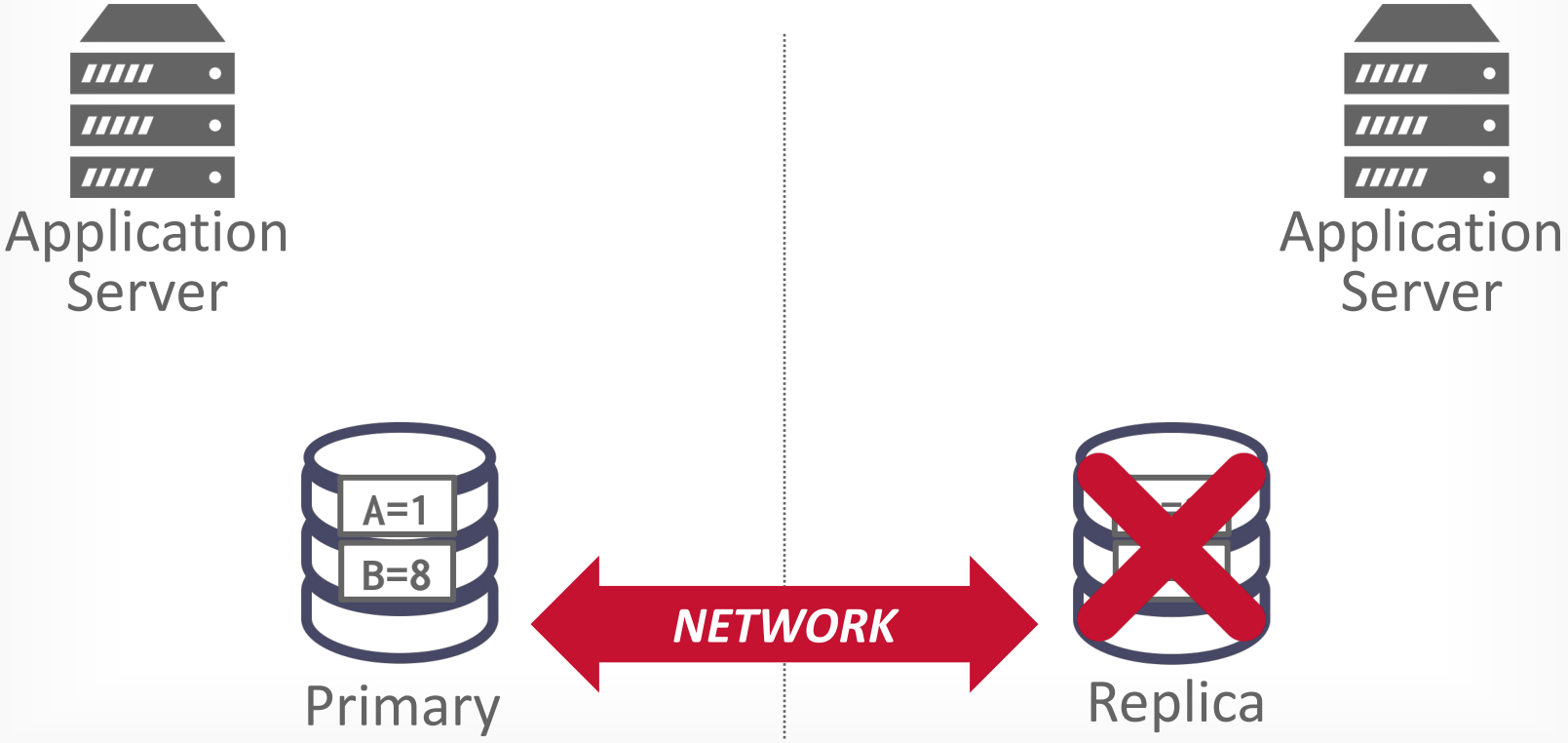
CONSISTENCY



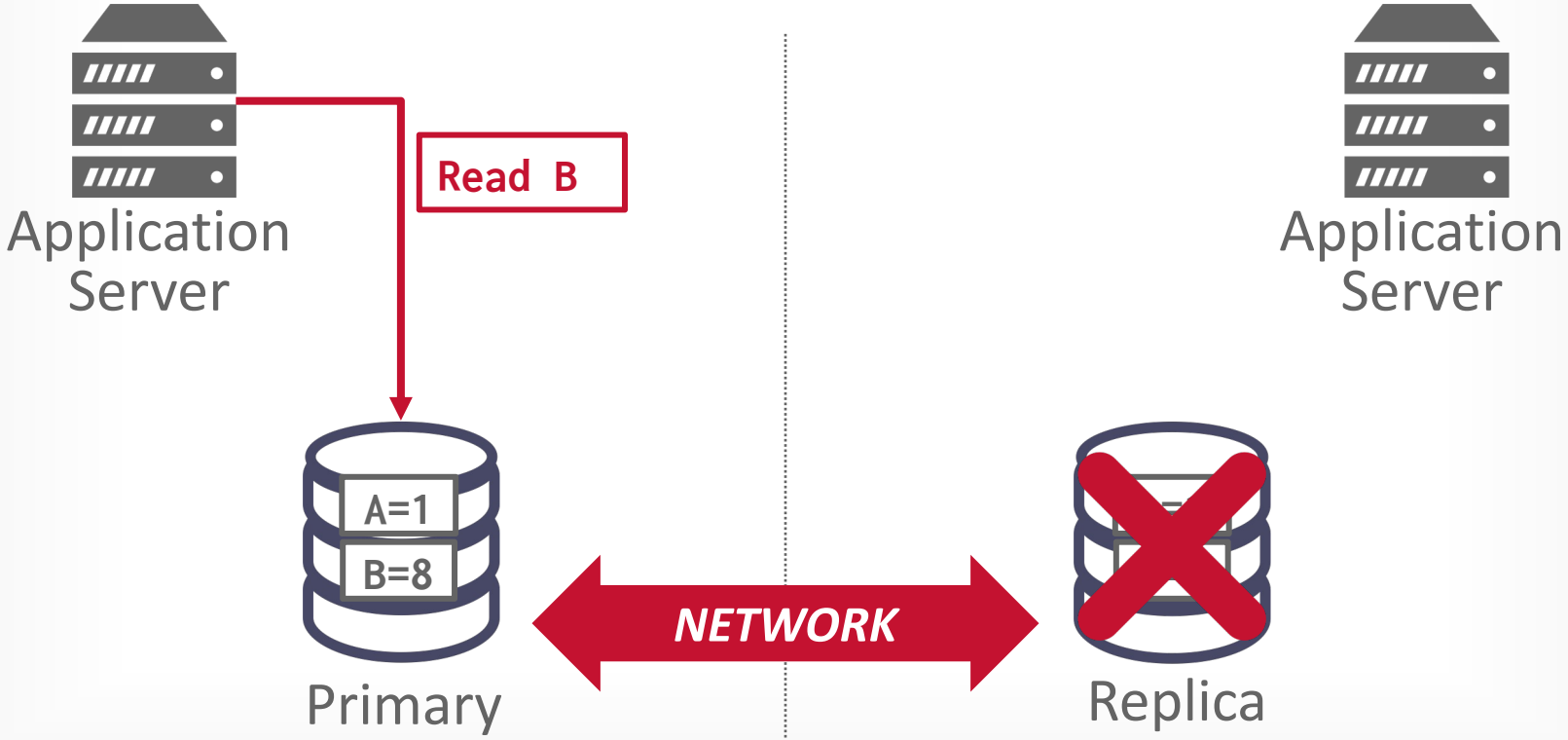
AVAILABILITY



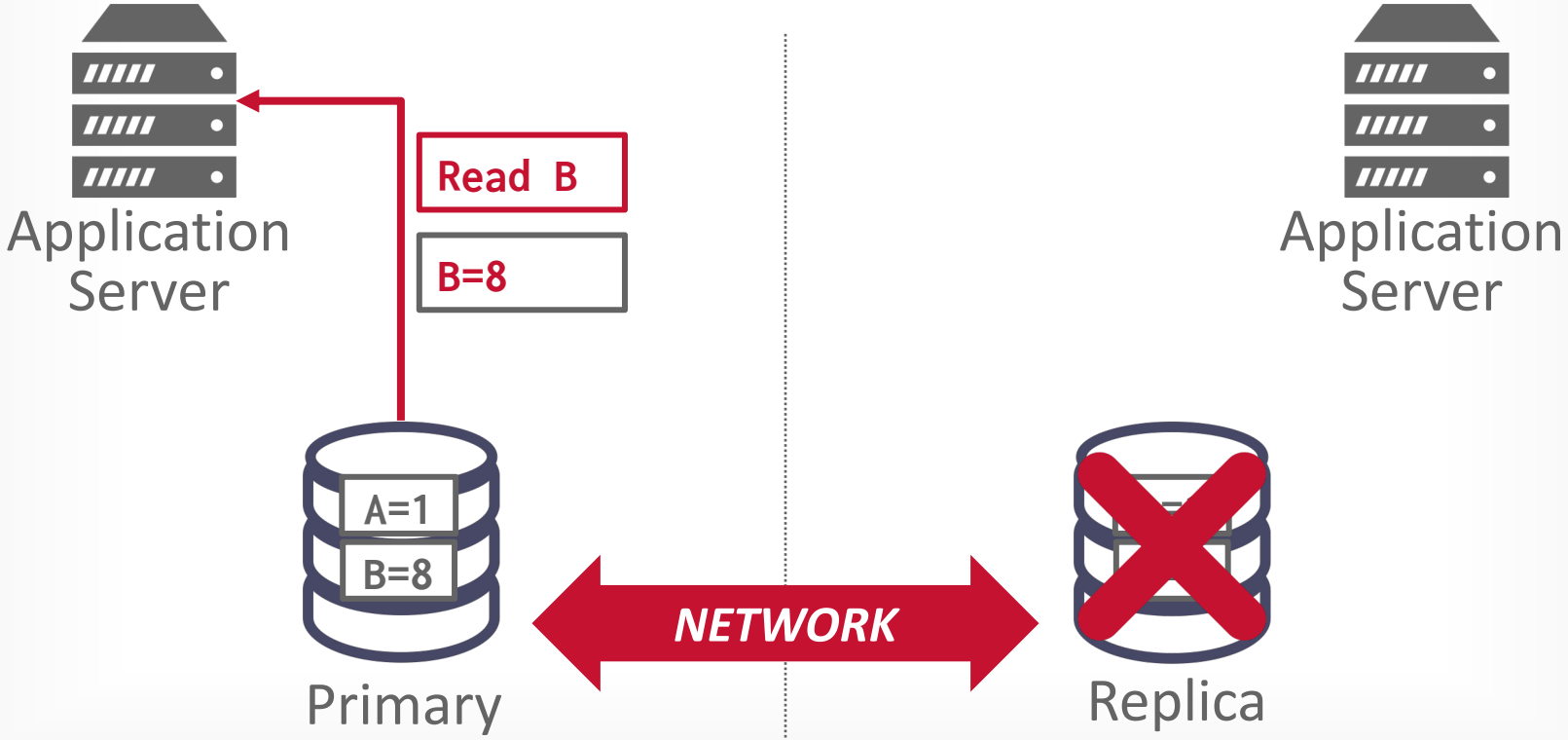
AVAILABILITY



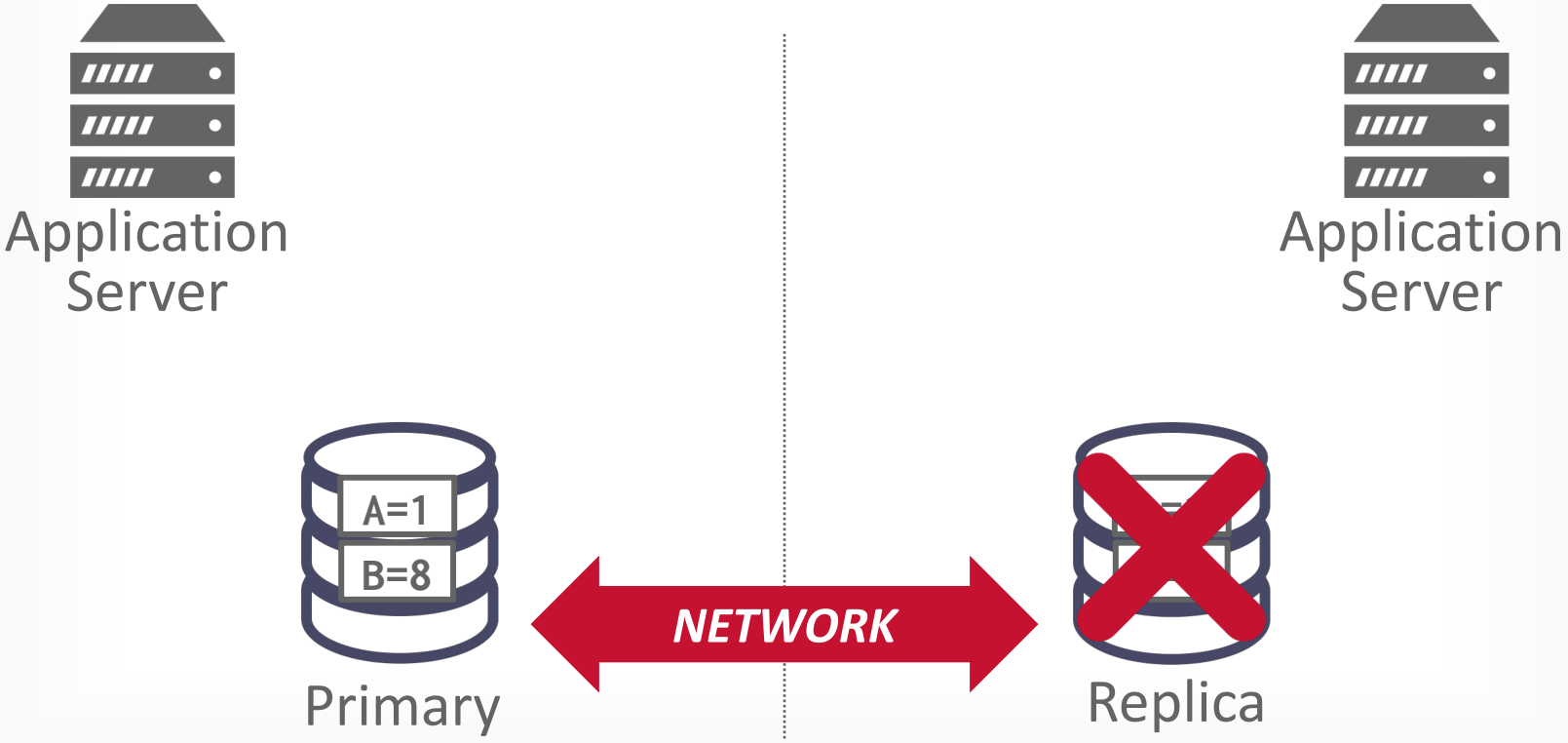
AVAILABILITY



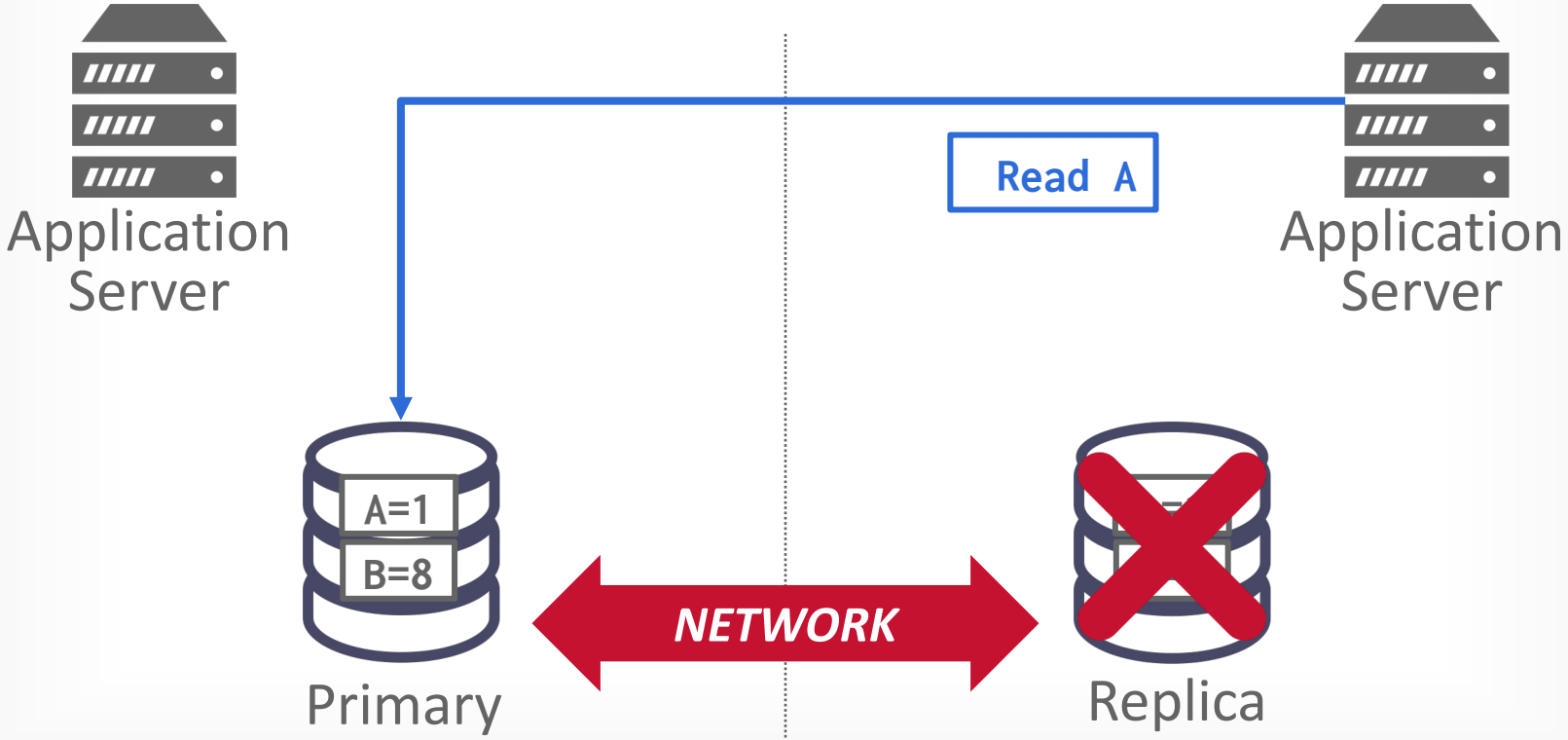
AVAILABILITY



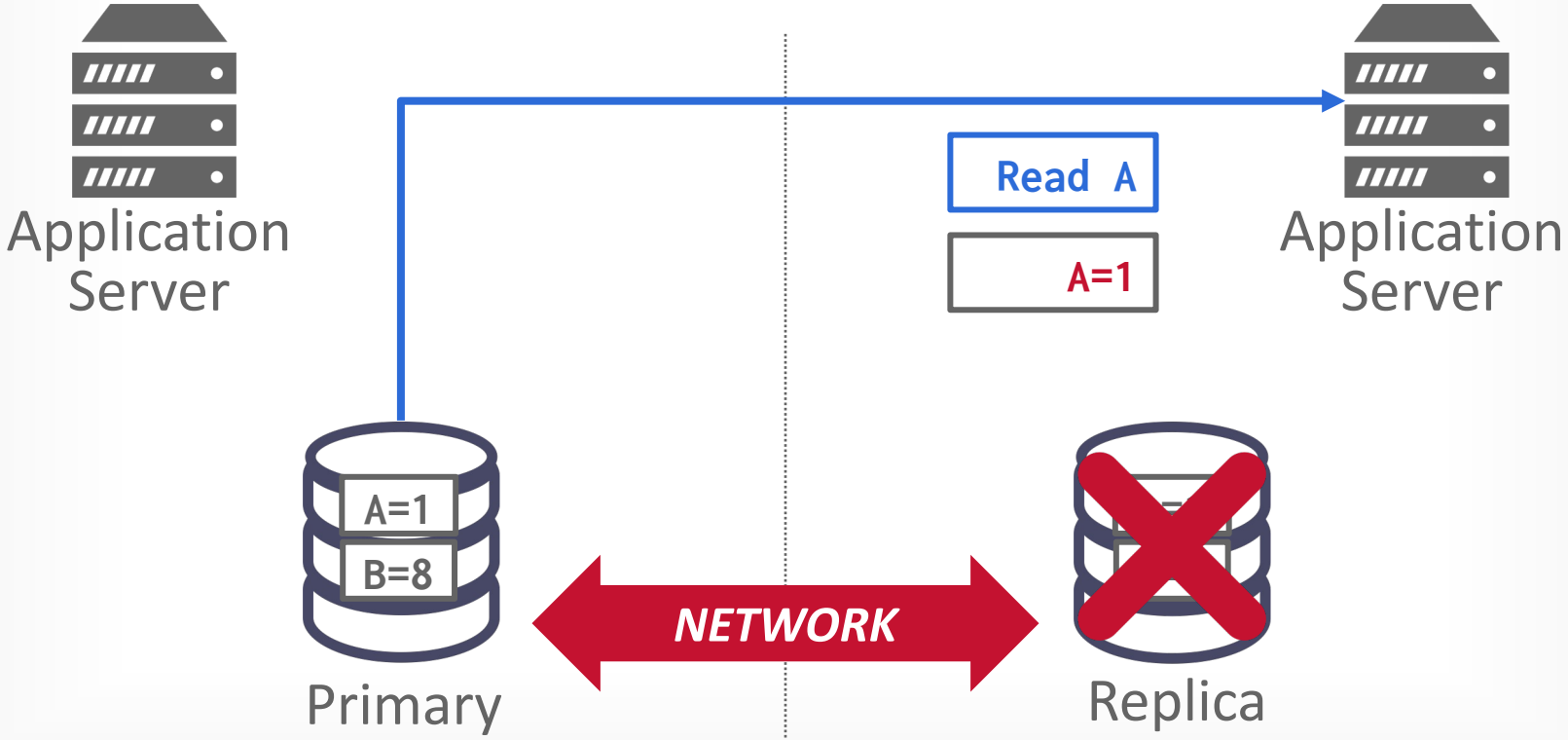
AVAILABILITY



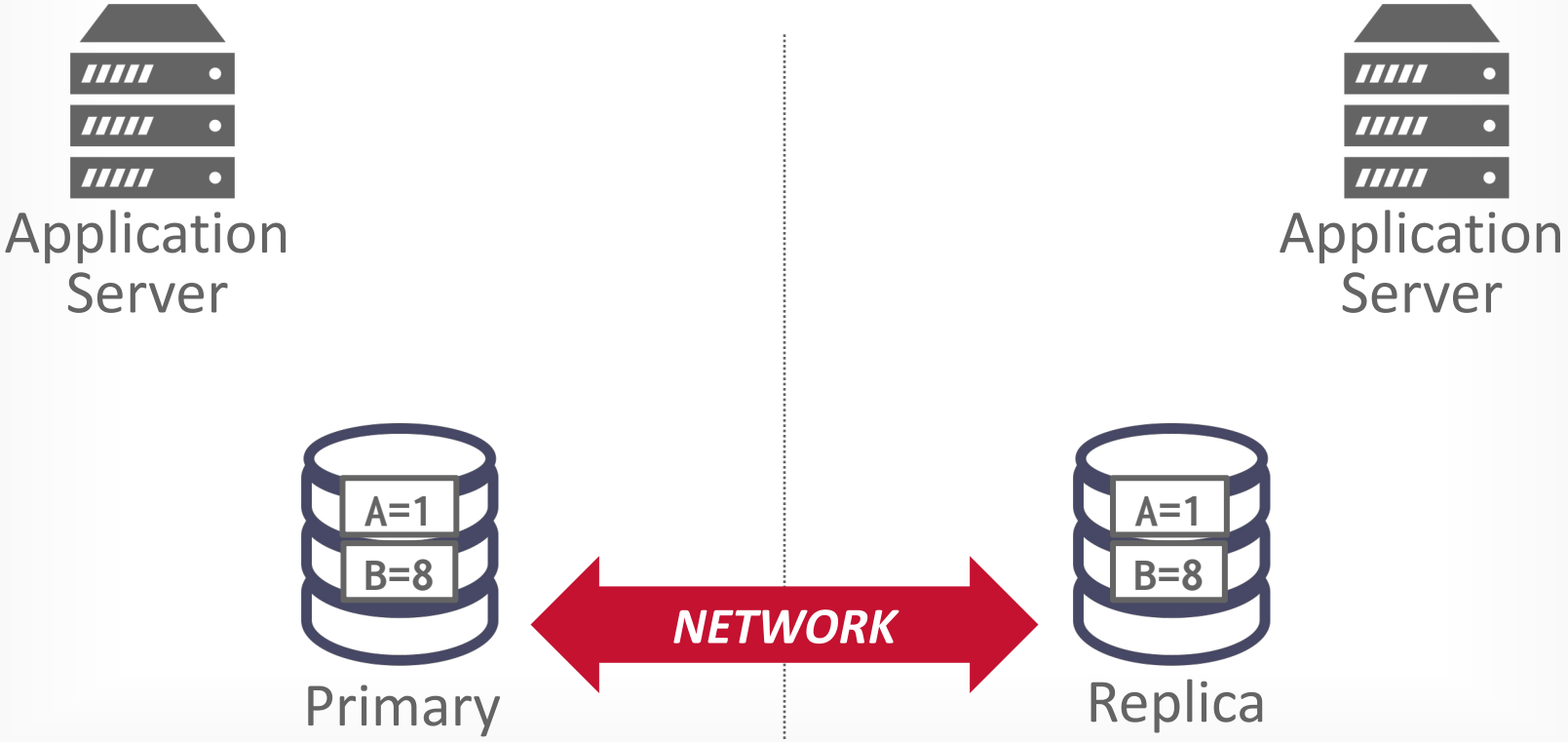
AVAILABILITY



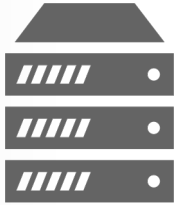
AVAILABILITY



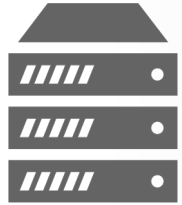
PARTITION TOLERANCE



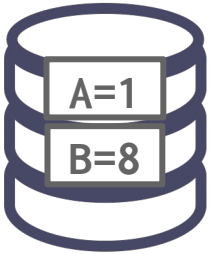
PARTITION TOLERANCE



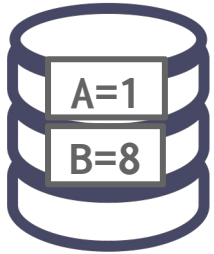
Application Server



Application Server

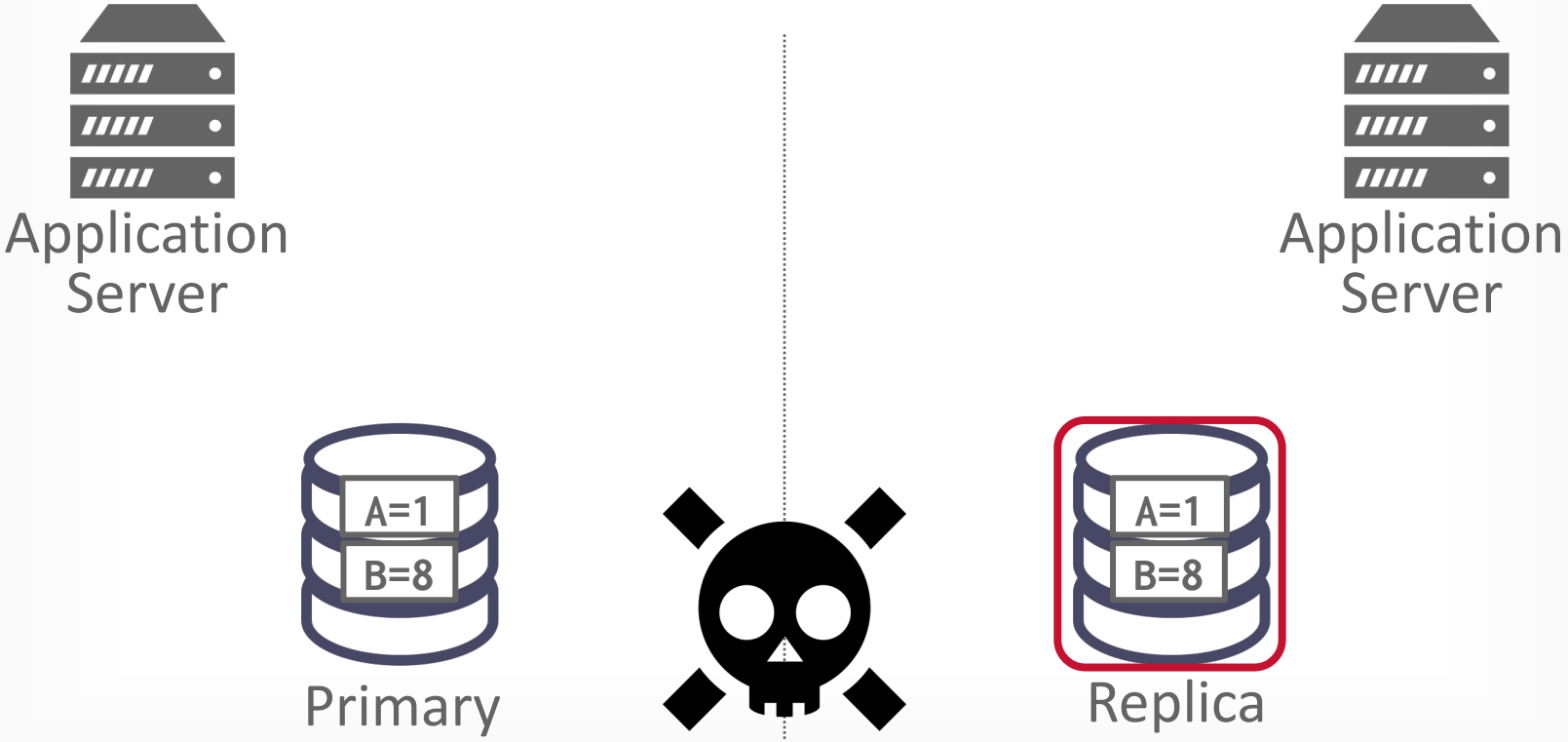


Primary

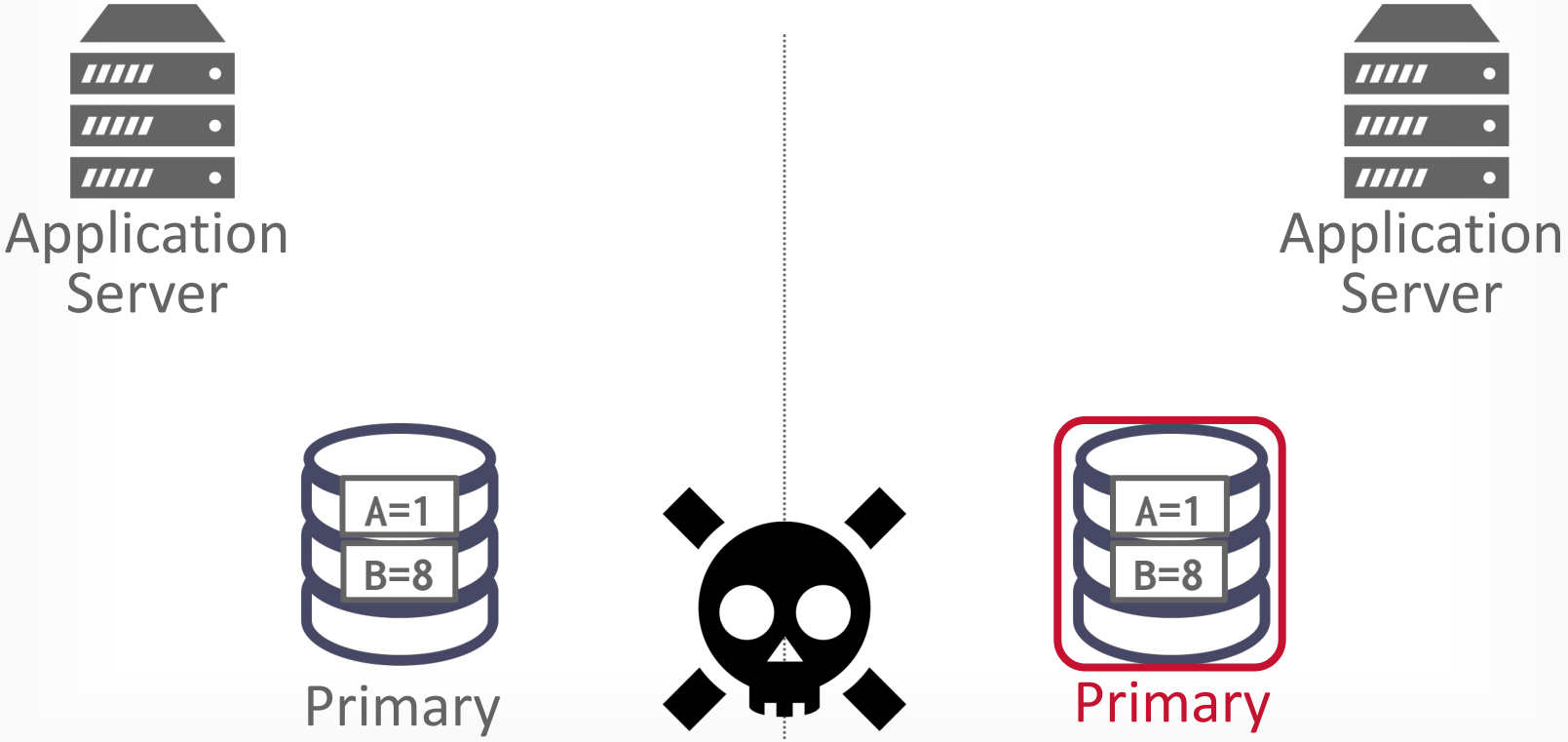


Replica

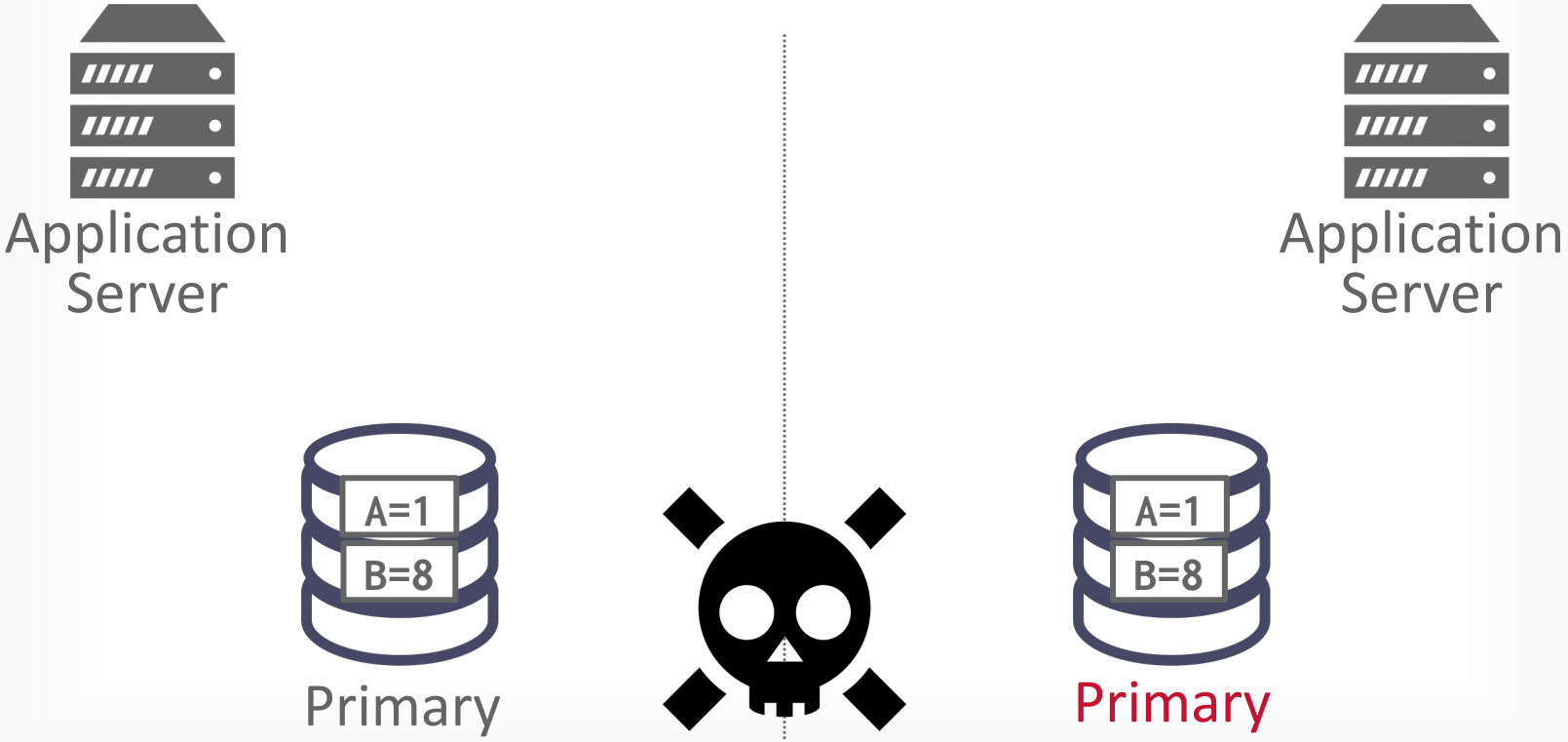
PARTITION TOLERANCE



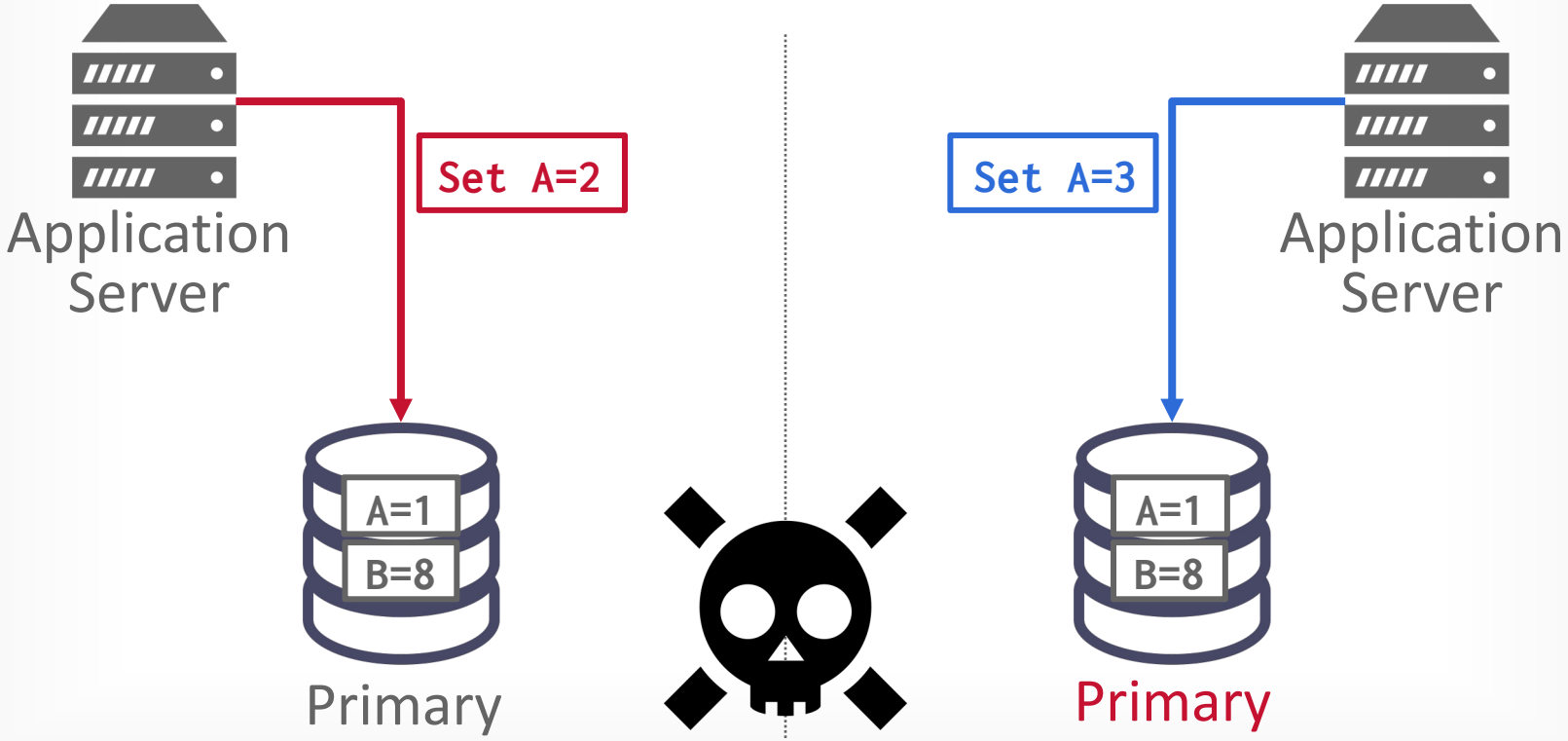
PARTITION TOLERANCE



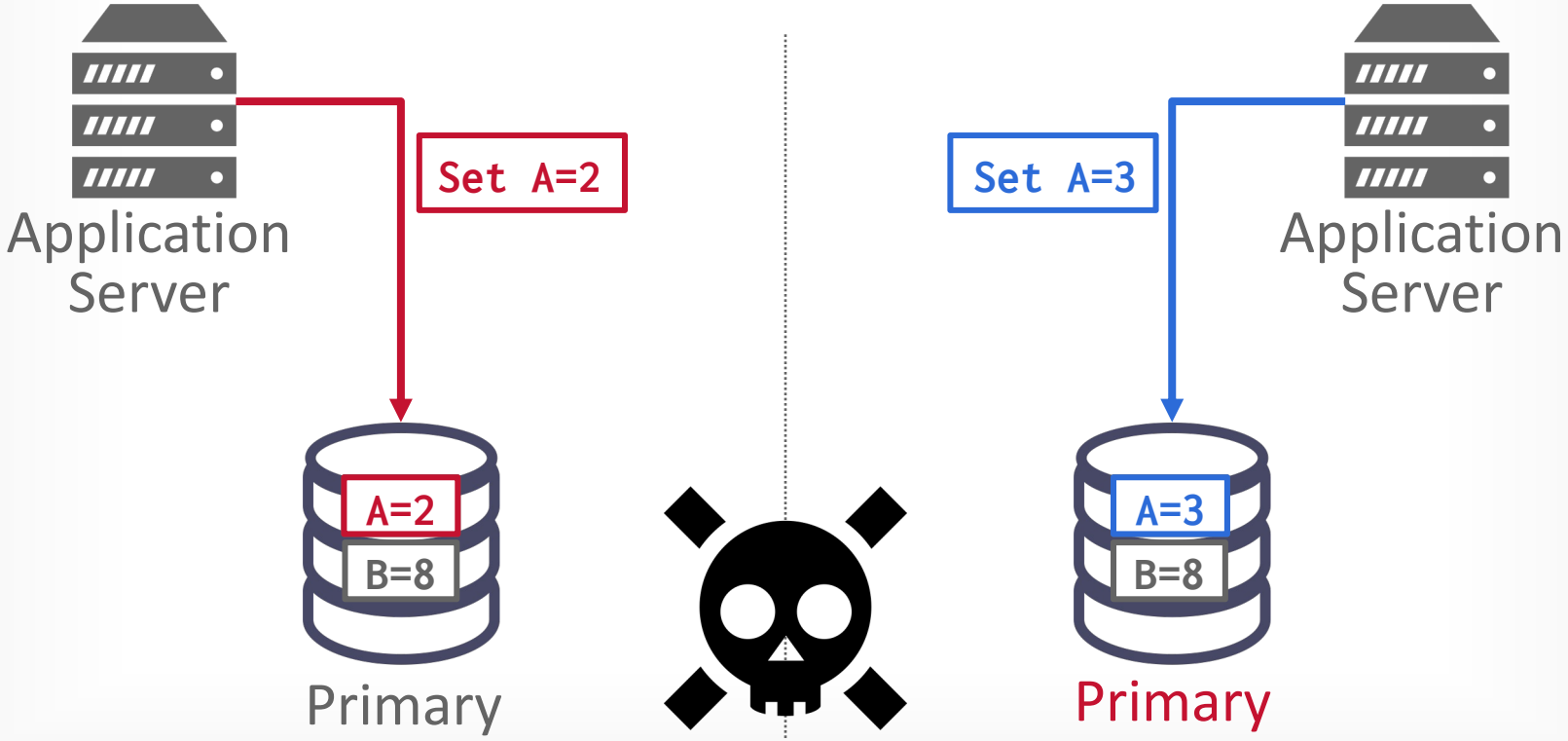
PARTITION TOLERANCE



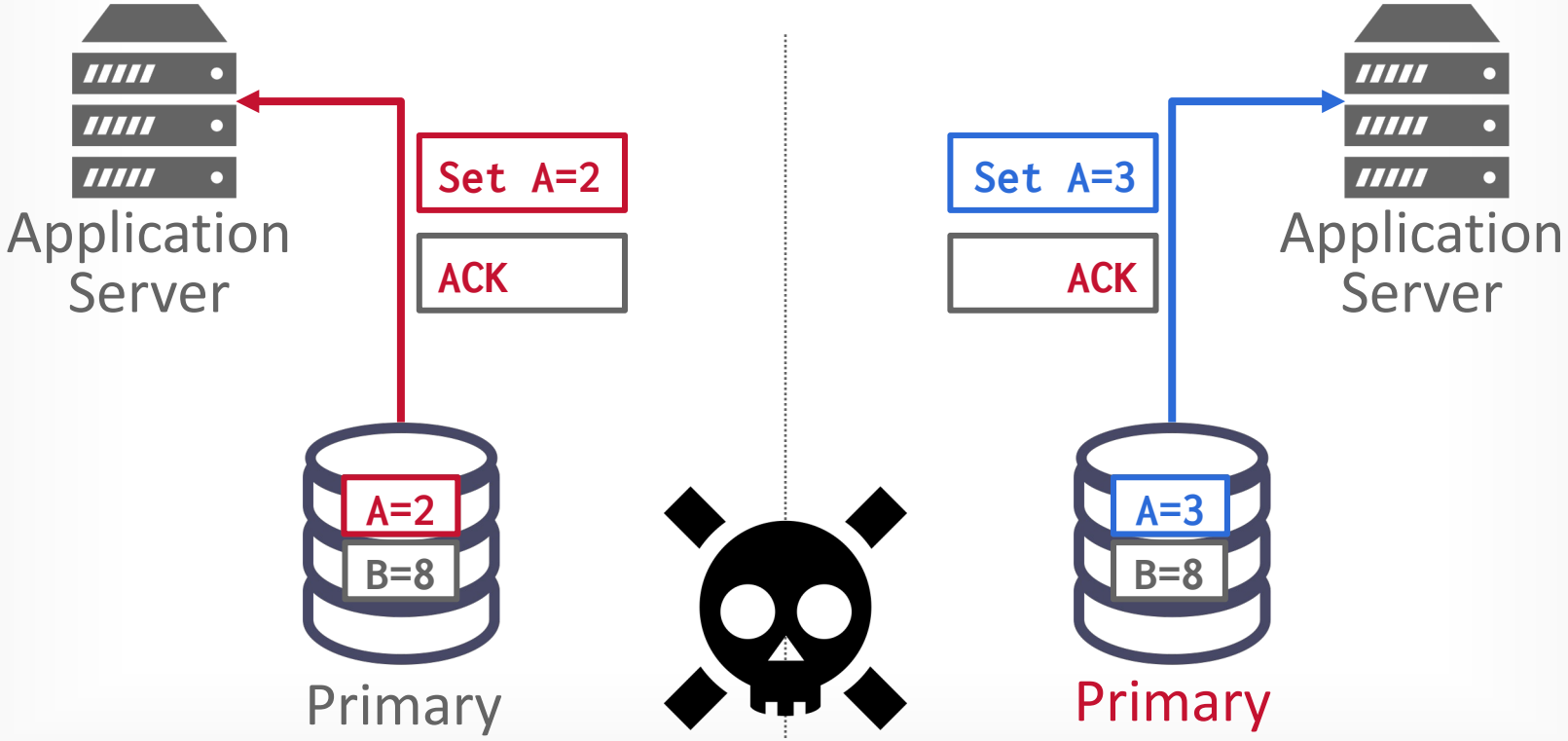
PARTITION TOLERANCE



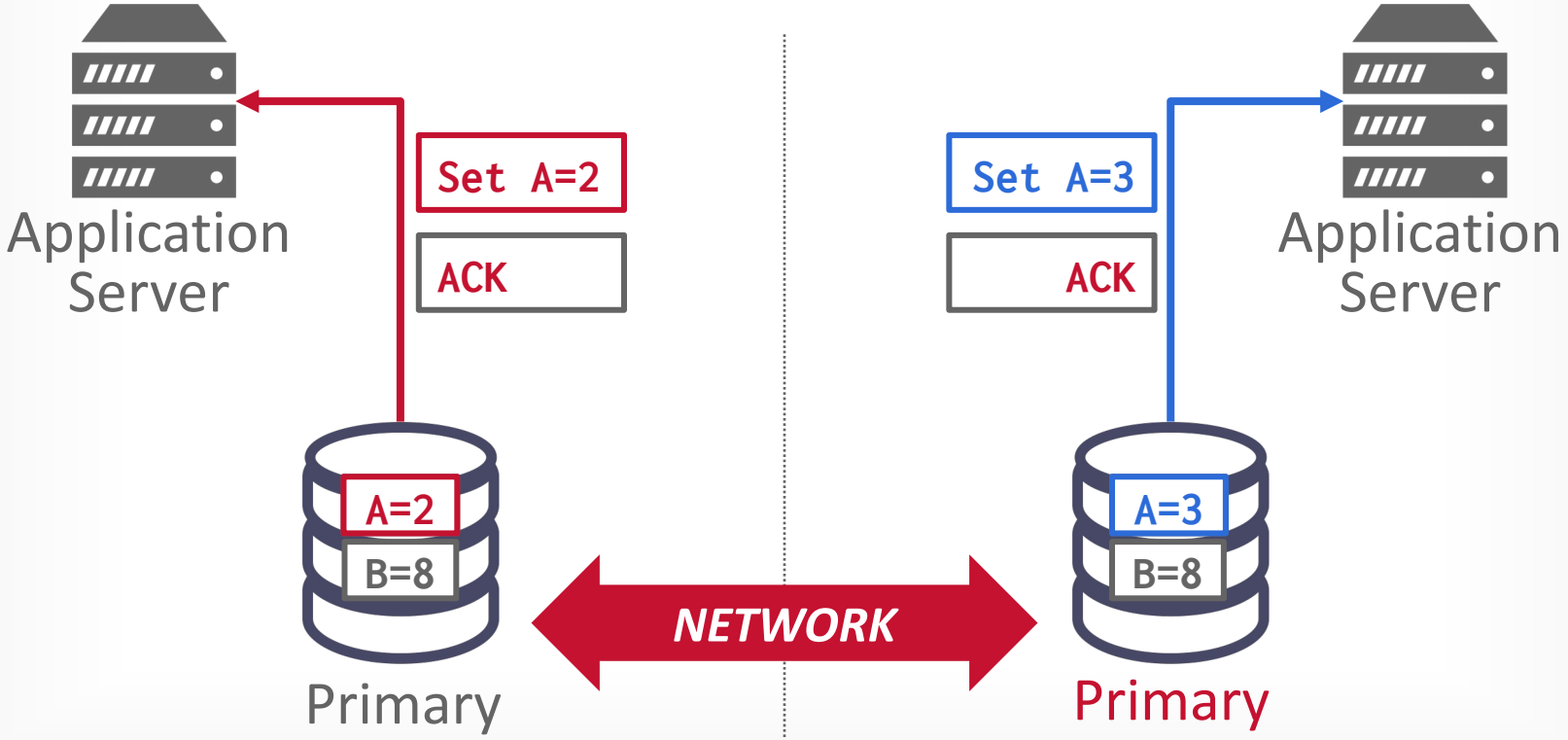
PARTITION TOLERANCE



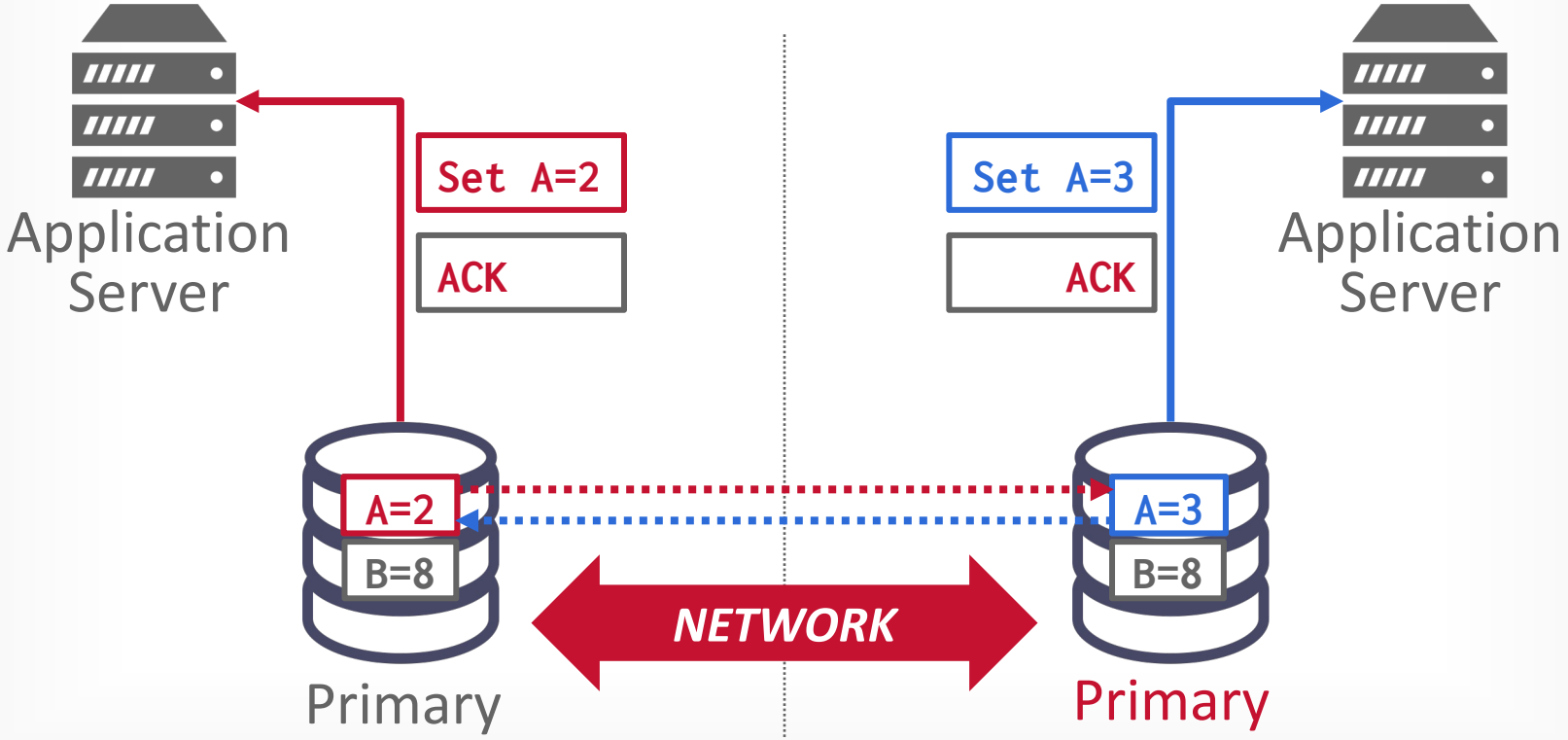
PARTITION TOLERANCE



PARTITION TOLERANCE



PARTITION TOLERANCE



PARTITION TOLERANCE

Choice #1: Halt the System

→ Stop accepting updates in any partition that does not have a majority of the nodes.

Choice #2: Allow Split, Reconcile Changes

→ Allow each side of partition to keep accepting updates.
→ Upon reconnection, perform reconciliation to determine the "correct" version of any updated record
→ Server-side: Last Update Wins
→ Client-side: Vector Clocks

PARTITION TOLERANCE

Choice #1: Halt the System

→ Stop accepting updates in any partition that does not have a majority of the nodes.

Choice #2: Allow Split, Reconcile Changes

→ Allow each side of partition to keep accepting updates.
→ Upon reconnection, perform reconciliation to determine the "correct" version of any updated record
→ Server-side: Last Update Wins
→ Client-side: Vector Clocks

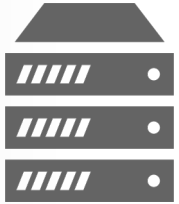
Don't Do This!

PACELC THEOREM

Extension to CAP proposed in 2010 to include consistency vs. latency trade-offs:

→ **P**artition Tolerant
→ **A**lways Available
→ **C**onsistent
→ **E**lse, choose during normal operations
→ **L**atency
→ **C**onsistency

LATENCY VS. CONSISTENCY



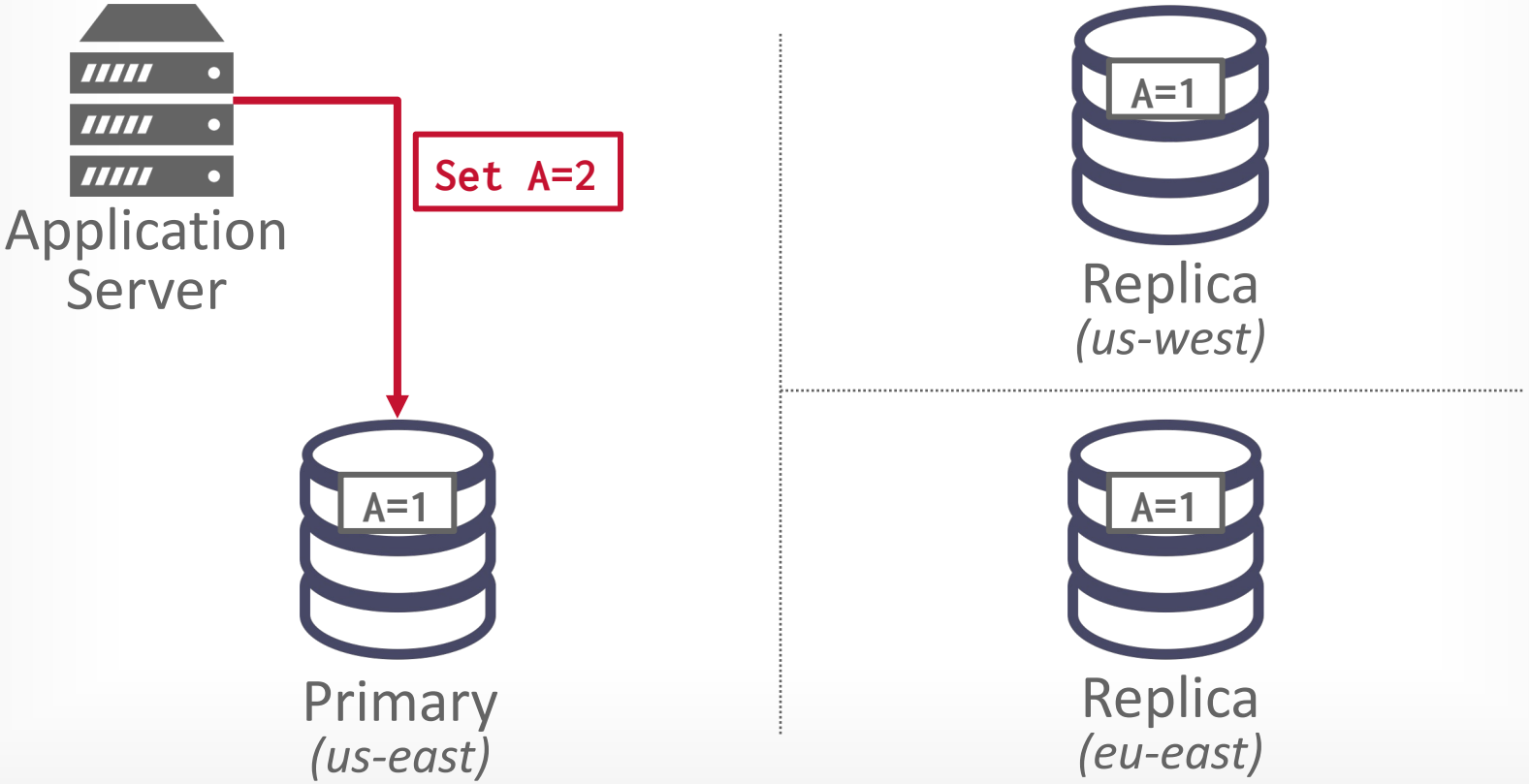
Application Server

Primary
(us-east)

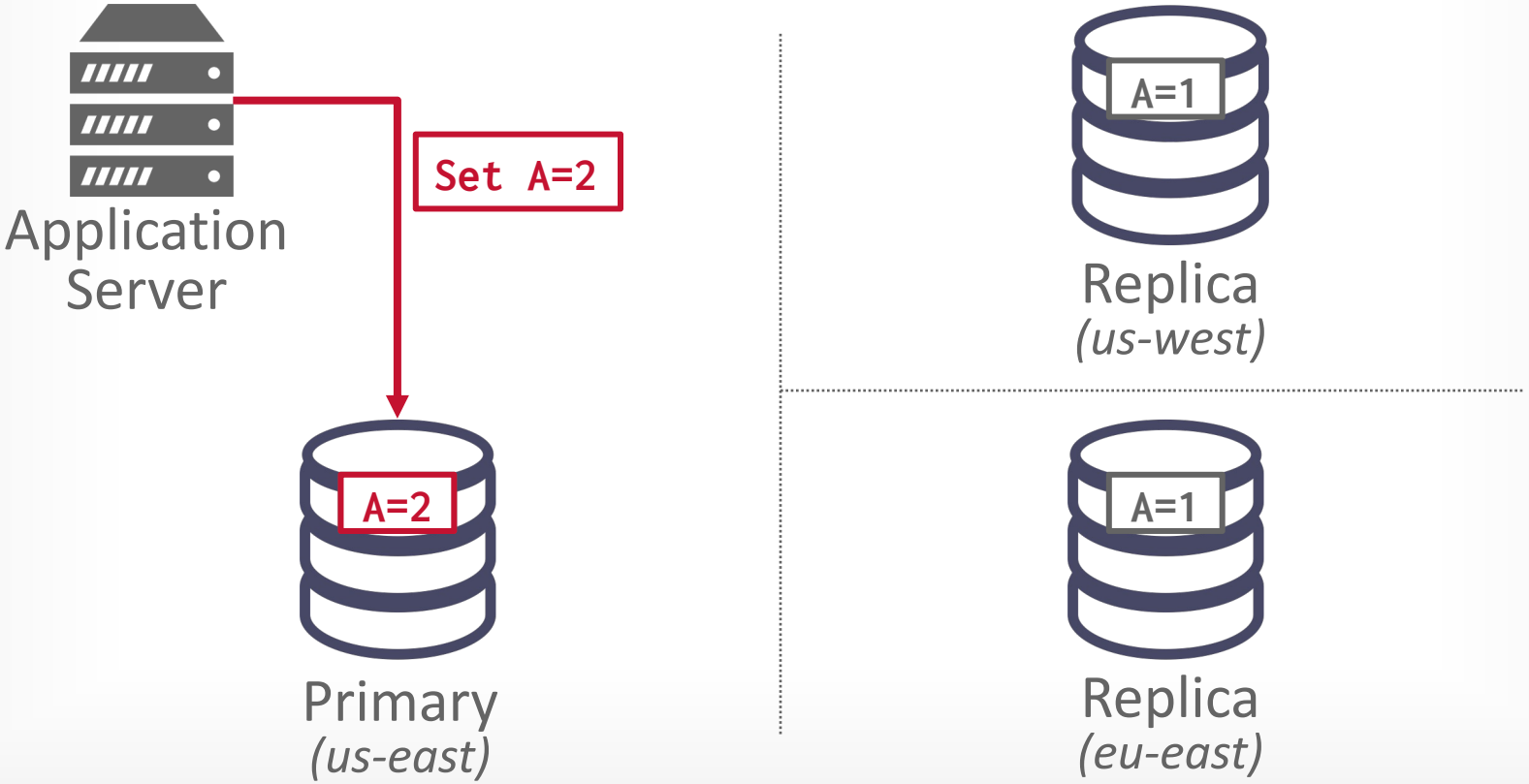
Replica
(us-west)

Replica
(eu-east)

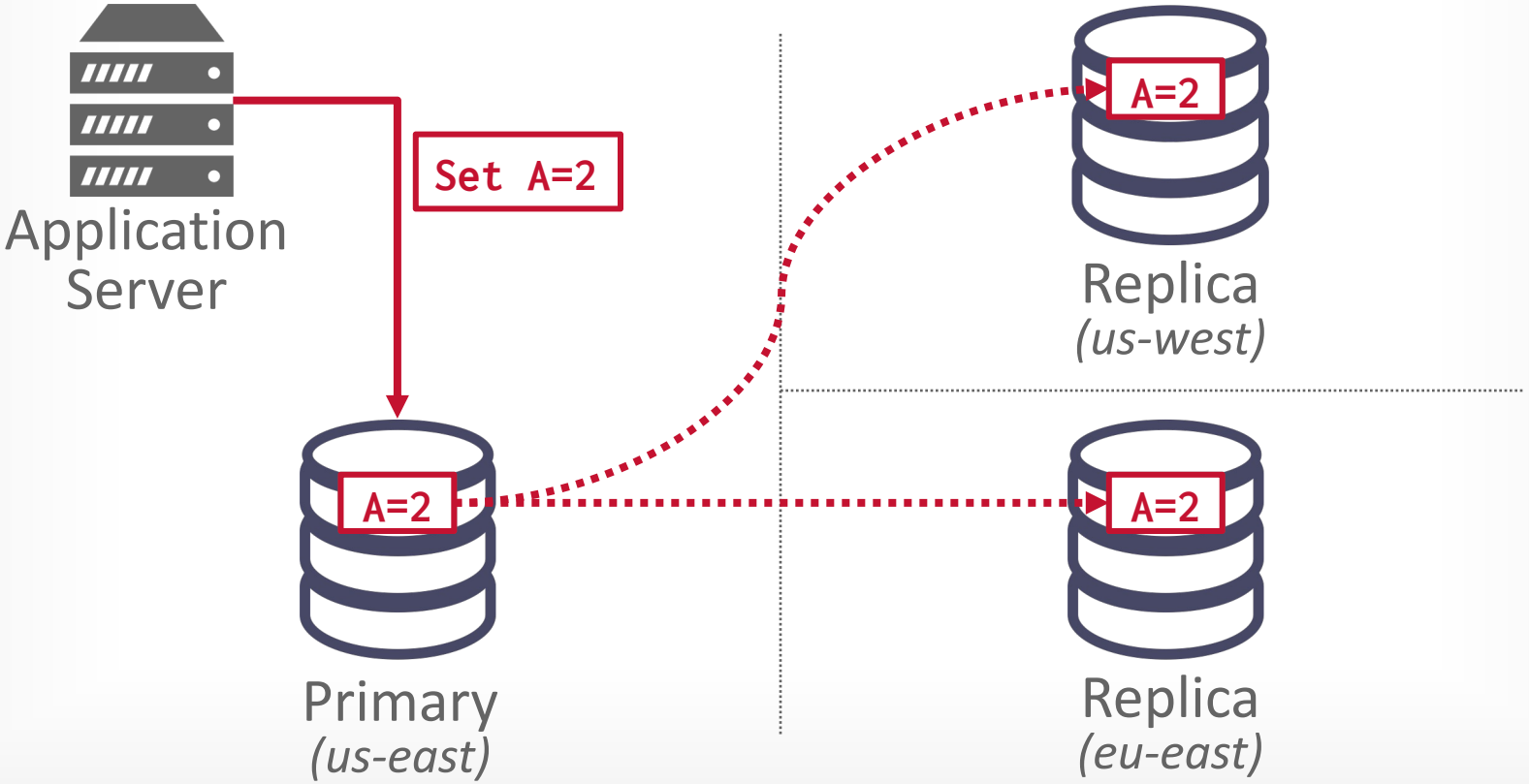
LATENCY VS. CONSISTENCY



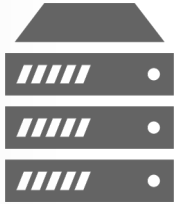
LATENCY VS. CONSISTENCY



LATENCY VS. CONSISTENCY



LATENCY VS. CONSISTENCY



Application Server

Primary
(us-east)

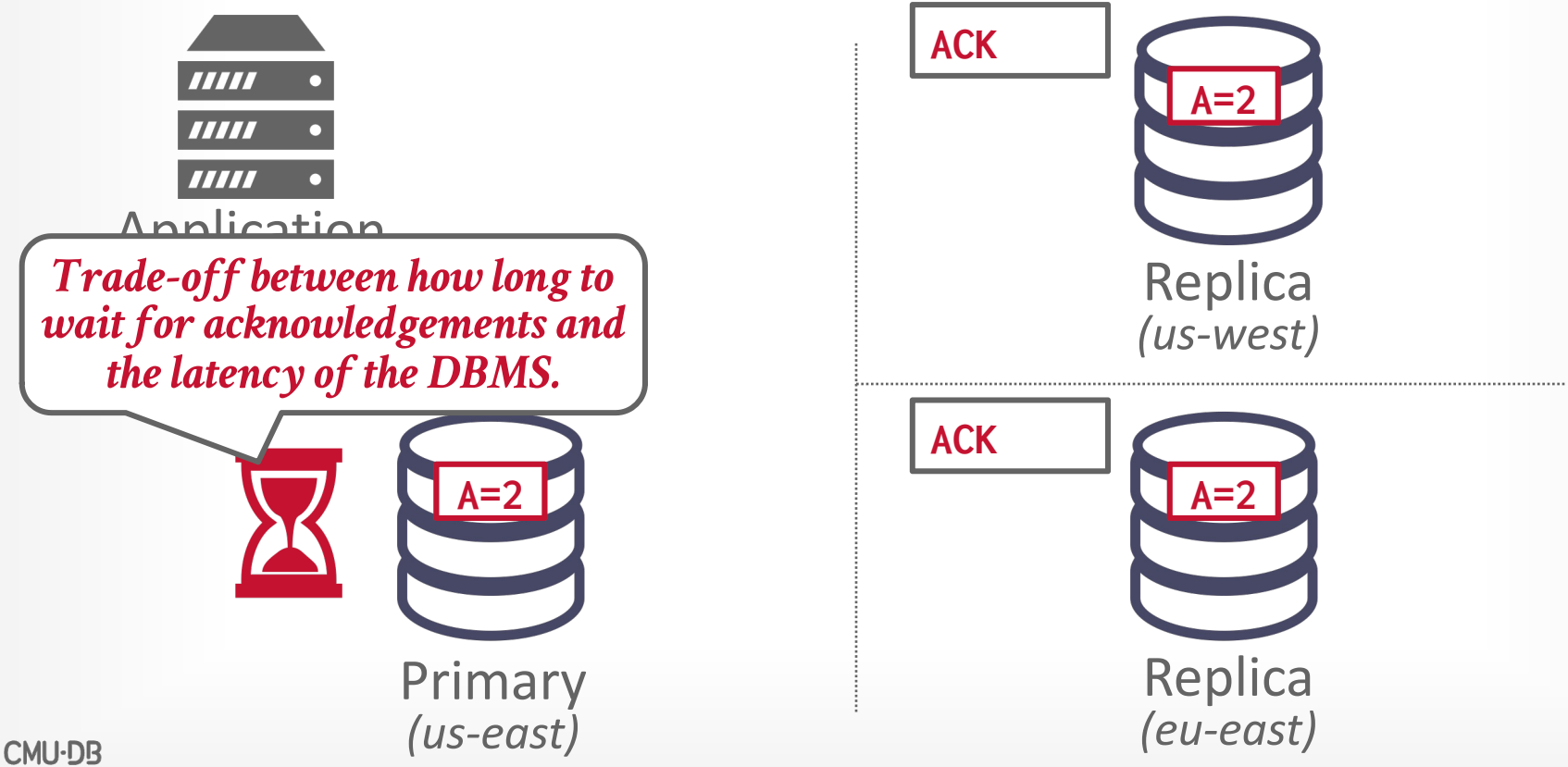
ACK

Replica
(us-west)

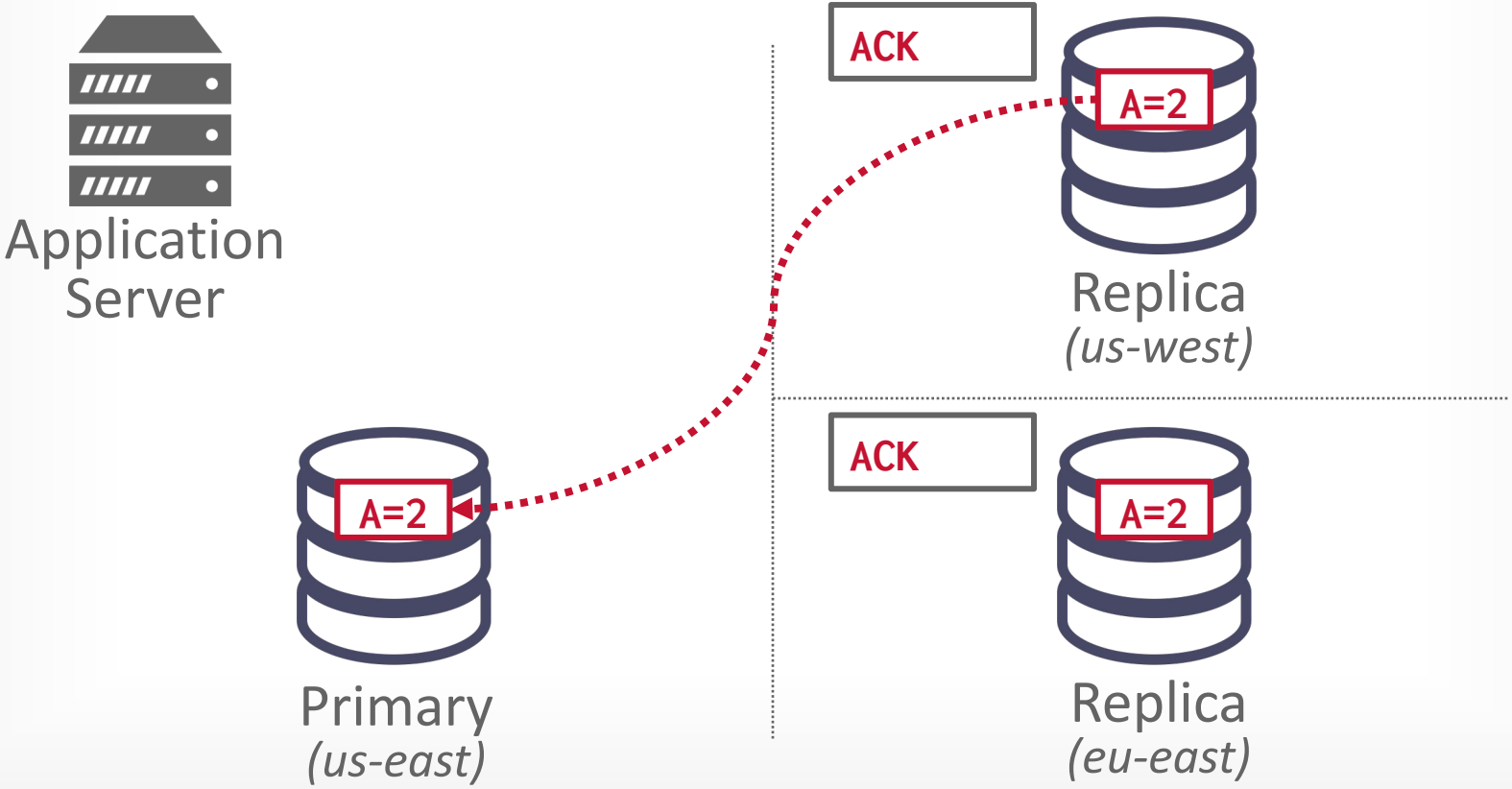
ACK

Replica
(eu-east)

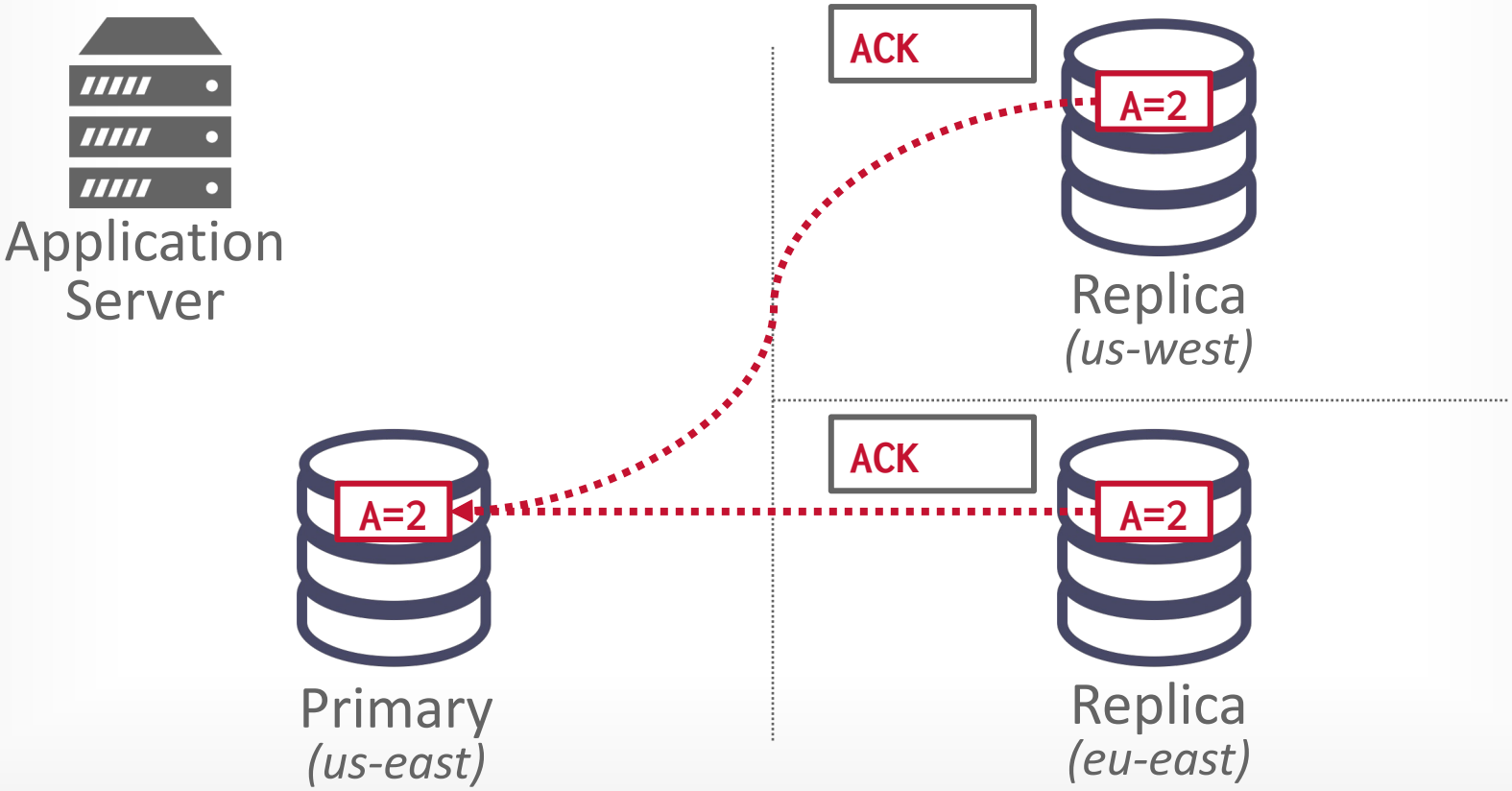
LATENCY VS. CONSISTENCY



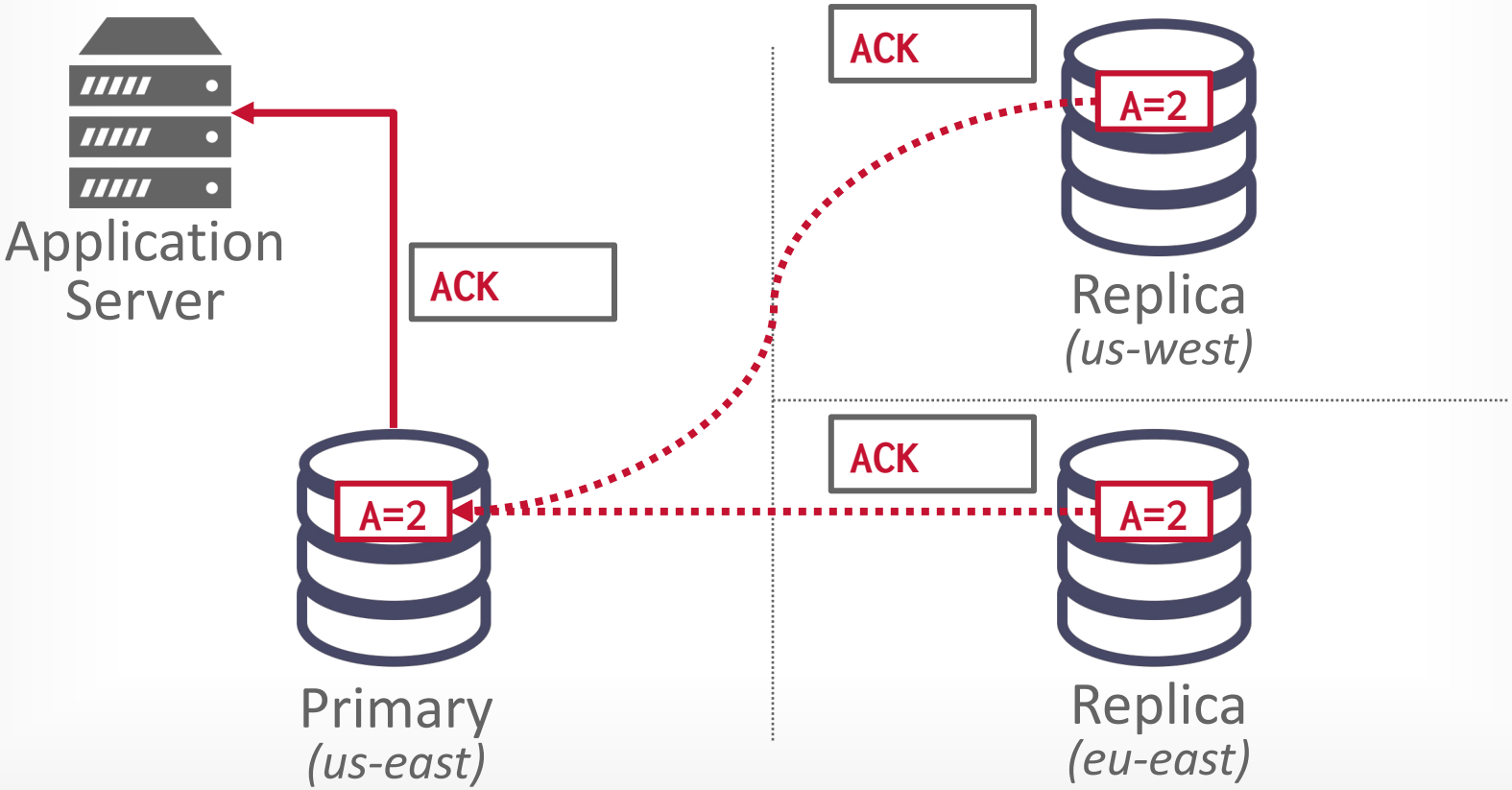
LATENCY VS. CONSISTENCY



LATENCY VS. CONSISTENCY



LATENCY VS. CONSISTENCY



CONCLUSION

Maintaining transactional consistency across multiple nodes is hard. Bad things will happen.
→ Don't let the "unwashed masses" go without txns!

2PC / Paxos / Raft are the most common protocols to ensure correctness in a distributed DBMS.

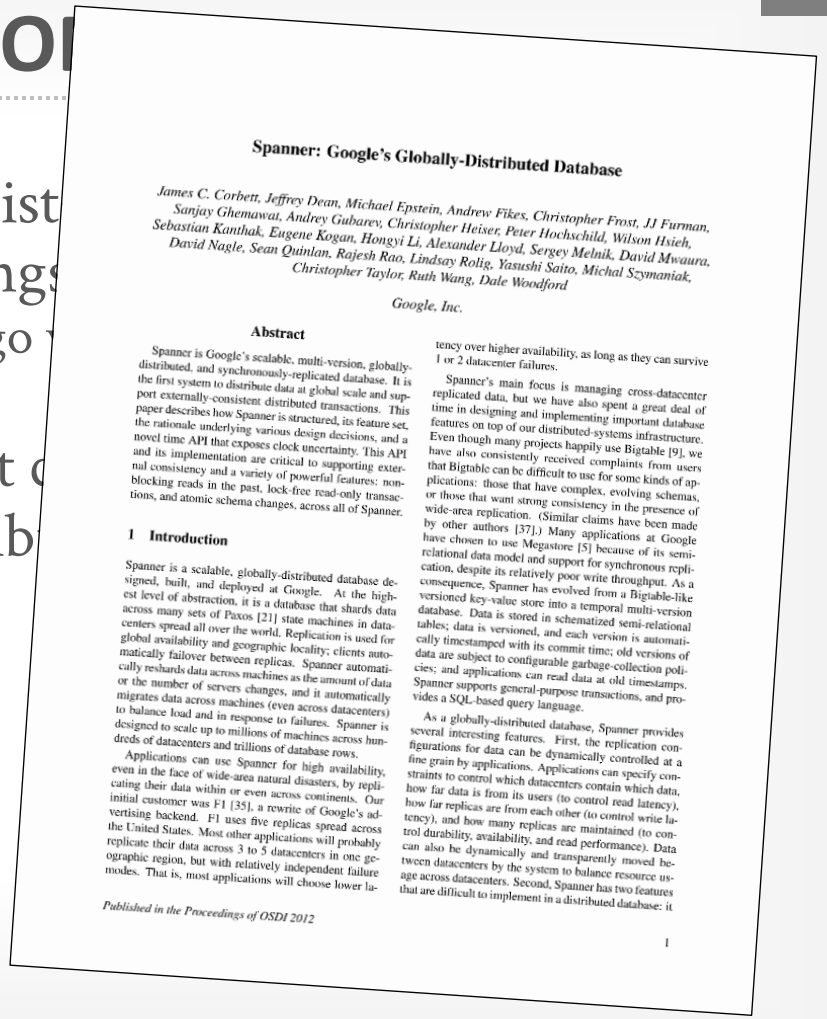
More info (and humiliation):
→ Kyle Kingsbury's Jepsen Project

CONCLUSIO

Maintaining transactional consist
multiple nodes is hard. Bad things
→ Don't let the "unwashed masses" go

2PC / Paxos / Raft are the most
to ensure correctness in a distrib

More info (and humiliation):
→ Kyle Kingsbury's Jepsen Project



Spanner: Google's Globally-Distributed Database

James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, Dale Woodford

Google, Inc.

Abstract

Spanner is Google's scalable, multi-version, globally-distributed, and synchronously-replicated database. It is the first system to distribute data at global scale and support externally-consistent distributed transactions. This paper describes how Spanner is structured, its feature set, the rationale underlying various design decisions, and a novel time API that exposes clock uncertainty. This API and its implementation are critical to supporting external consistency and a variety of powerful features: non-blocking reads in the past, lock-free read-only transactions, and atomic schema changes, across all of Spanner.

1 Introduction

Spanner is a scalable, globally-distributed database designed, built, and deployed at Google. At the highest level of abstraction, it is a database that shards data across many sets of Paxos [21] state machines in datacenters spread all over the world. Replication is used for global availability and geographic locality; clients automatically failover between replicas. Spanner automatically reshards data across machines as the amount of data or the number of servers changes, and it automatically migrates data across machines (even across datacenters) to balance load and in response to failures. Spanner is designed to scale up to millions of machines across hundreds of datacenters and trillions of database rows.

Applications can use Spanner for high availability, even in the face of wide-area natural disasters, by replicating their data within or even across continents. Our initial customer was F1 [35], a rewrite of Google's advertising backend. F1 uses five replicas spread across the United States. Most other applications will probably replicate their data across 3 to 5 datacenters in one geographic region, but with relatively independent failure modes. That is, most applications will choose lower latency over higher availability, as long as they can survive 1 or 2 datacenter failures.

Spanner's main focus is managing cross-datacenter replicated data, but we have also spent a great deal of time in designing and implementing important database features on top of our distributed-systems infrastructure. Even though many projects happily use Bigtable [9], we have also consistently received complaints from users that Bigtable can be difficult to use for some kinds of applications: those that have complex, evolving schemas, or those that want strong consistency in the presence of wide-area replication. (Similar claims have been made by other authors [37].) Many applications at Google have chosen to use Megastore [5] because of its semi-relational data model and support for synchronous replication, despite its relatively poor write throughput. As a consequence, Spanner has evolved from a Bigtable-like versioned key-value store into a temporal multi-version database. Data is stored in schematized semi-relational tables; data is versioned, and each version is automatically timestamped with its commit time; old versions of data are subject to configurable garbage-collection policies; and applications can read data at old timestamps. Spanner supports general-purpose transactions, and provides a SQL-based query language.

As a globally-distributed database, Spanner provides several interesting features. First, the replication configurations for data can be dynamically controlled at a fine grain by applications. Applications can specify constraints to control which datacenters contain which data, how far data is from its users (to control read latency), how far replicas are from each other (to control write latency), and how many replicas are maintained (to control durability, availability, and read performance). Data can also be dynamically and transparently moved between datacenters by the system to balance resource usage across datacenters. Second, Spanner has two features that are difficult to implement in a distributed database: it

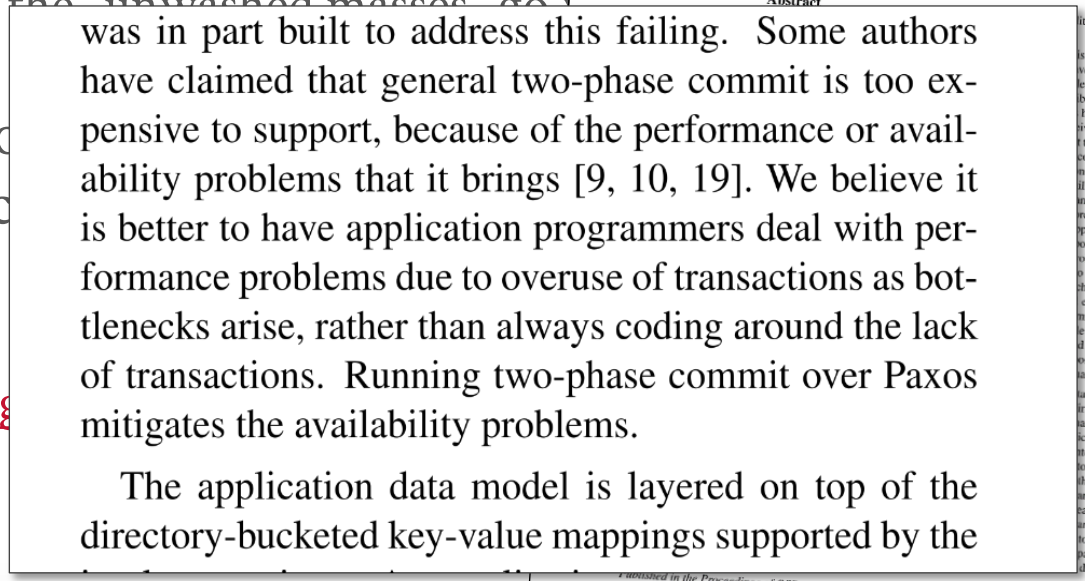
Published in the Proceedings of OSDI 2012

CONCLUSIO

Maintaining transactional consist
multiple nodes is hard. Bad things
→ Don't let the "unwashed masses" go

2PC / Paxo
to ensure c

More info
→ Kyle King



was in part built to address this failing. Some authors have claimed that general two-phase commit is too expensive to support, because of the performance or availability problems that it brings [9, 10, 19]. We believe it is better to have application programmers deal with performance problems due to overuse of transactions as bottlenecks arise, rather than always coding around the lack of transactions. Running two-phase commit over Paxos mitigates the availability problems.

The application data model is layered on top of the directory-bucketed key-value mappings supported by the

Spanner: Google's Globally-Distributed Database

James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, Dale Woodford

Google, Inc.

Abstract

Published in the Proceedings of OSDI 2012

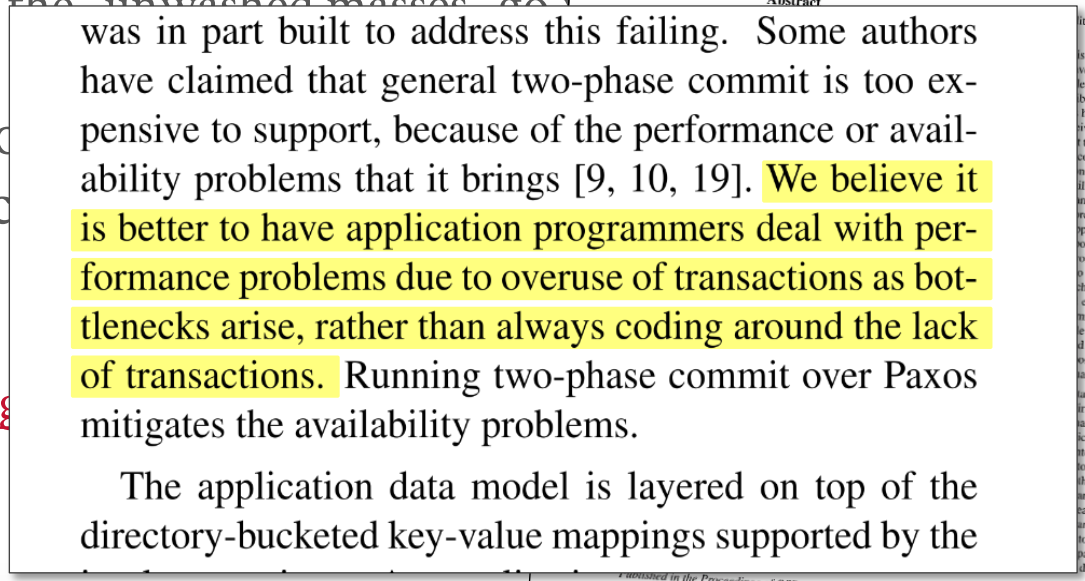
1

CONCLUSIO

Maintaining transactional consist
multiple nodes is hard. Bad things
→ Don't let the "unwashed masses" go

2PC / Paxo
to ensure c

More info
→ Kyle King



was in part built to address this failing. Some authors have claimed that general two-phase commit is too expensive to support, because of the performance or availability problems that it brings [9, 10, 19]. We believe it is better to have application programmers deal with performance problems due to overuse of transactions as bottlenecks arise, rather than always coding around the lack of transactions. Running two-phase commit over Paxos mitigates the availability problems.

The application data model is layered on top of the directory-bucketed key-value mappings supported by the

Spanner: Google's Globally-Distributed Database

James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, Dale Woodford

Google, Inc.

Abstract

Published in the Proceedings of OSDI 2012

CONCLUSION

Maintaining transactional consistency across multiple nodes is hard. Bad things will happen.
→ Don't let the "unwashed masses" go without txns!

2PC / Paxos / Raft are the most common protocols to ensure correctness in a distributed DBMS.

More info (and humiliation):
→ Kyle Kingsbury's Jepsen Project

NEXT CLASS

Distributed OLAP Systems