Carnegie Mellon University

Timestamp Ordering

Lecture #19

Database Systems
15-445/15-645
Fall 2017

Andy Pavlo
Computer Science Dept.
Carnegie Mellon Univ.

ADMINISTRIVIA

Homework #4: Monday November 13th @ 11:59pm

Project #3: Wednesday November 15th @ 11:59am

CONCURRENCY CONTROL APPROACHES

Two-Phase Locking (2PL)

→ Determine serializability order of conflicting operations at runtime while txns execute.

Timestamp Ordering (T/O)

→ Determine serializability order of txns before they execute.

CONCURRENCY CONTROL APPROACHES

Two-Phase Locking (2PL)
→ Determine serializability order of conflicting operations at runtime while txns execute.

Pessimistic

Timestamp Ordering (T/O)
→ Determine serializability order of txns before they execute.

Optimistic

T/O CONCURRENCY CONTROL

Use timestamps to determine the serializability order of txns.

If $TS(T_i) < TS(T_j)$, then the DBMS must ensure that the execution schedule is equivalent to a serial schedule where T_i appears before T_j.

TIMESTAMP ALLOCATION

Each txn T_i is assigned a unique fixed timestamp that is monotonically increasing.

→ Let $TS(T_i)$ be the timestamp allocated to txn T_i
→ Different schemes assign timestamps at different times during the txn.

Multiple implementation strategies:

→ System Clock.
→ Logical Counter.
→ Hybrid.

TODAY'S AGENDA

Basic Timestamp Ordering

Optimistic Concurrency Control

Partition-based Timestamp Ordering

BASIC T/O

Txns read and write objects without locks.

Every object `X` is tagged with timestamp of the last txn that successfully did read/write:

→ `W-TS(X)` – Write timestamp on `X`

→ `R-TS(X)` – Read timestamp on `X`

Check timestamps for every operation:

→ If txn tries to access an object "from the future", it aborts and restarts.

BASIC T/O – READS

If $TS(T_i)$ < $W\text{-}TS(X)$, this violates timestamp order of T_i with regard to the writer of X.

→ Abort T_i and restart it with same TS.

Else:

→ Allow T_i to read X.

→ Update $R\text{-}TS(X)$ to $\max(R\text{-}TS(X), TS(T_i))$

→ Have to make a local copy of X to ensure repeatable reads for T_i.

BASIC T/O – WRITES

If $TS(T_i) < R\text{-}TS(X)$ or $TS(T_i) < W\text{-}TS(X)$

→ Abort and restart T_i.

Else:

→ Allow T_i to write X and update $W\text{-}TS(X)$

→ Also have to make a local copy of X to ensure repeatable reads for T_i.

BASIC T/O – EXAMPLE #1

Schedule

TS(T_1)=1 TS(T_2)=2

TIME

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	COMMIT

Database

Object	R-TS	W-TS
A	0	0
B	0	0

BASIC T/O – EXAMPLE #1

Schedule

$TS(T_1)=1$ $TS(T_2)=2$

TIME

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	COMMIT

Database

Object	R-TS	W-TS
A	0	0
B	0	0

BASIC T/O – EXAMPLE #1

Schedule

TS(T_1)=1 TS(T_2)=2

TIME

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	COMMIT

Database

Object	R-TS	W-TS
A	0	0
B	1	0

BASIC T/O – EXAMPLE #1

Schedule

TIME

TS(T_1)=1

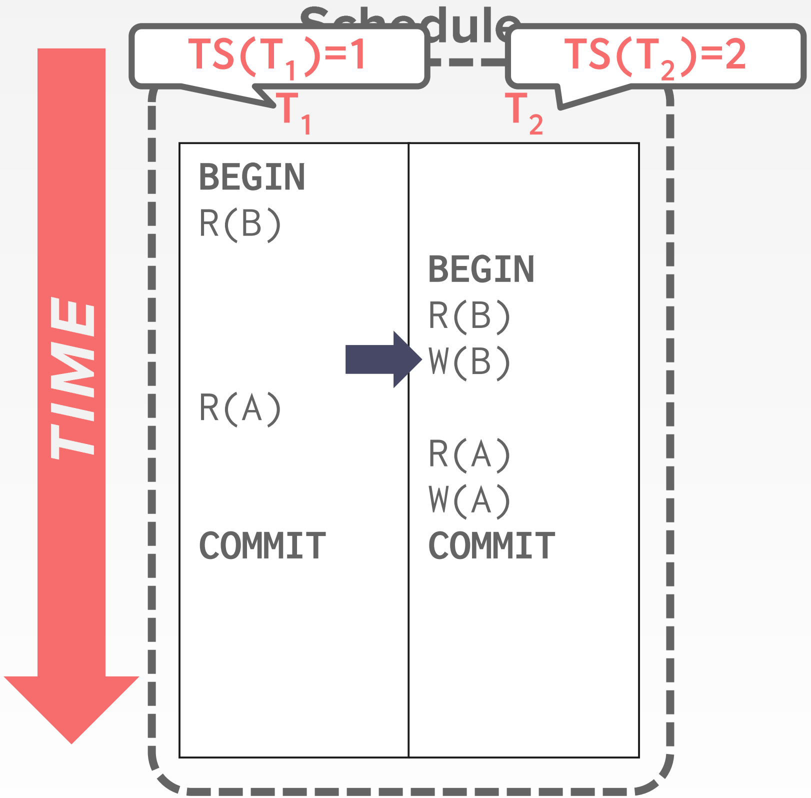
TS(T_2)=2

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	**COMMIT**

Database

Object	R-TS	W-TS
A	0	0
B	2	0

BASIC T/O – EXAMPLE #1



Database

Object	R-TS	W-TS
A	0	0
B	2	2

BASIC T/O – EXAMPLE #1

Schedule

$TS(T_1)=1$ $TS(T_2)=2$

TIME

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	COMMIT

Database

Object	R-TS	W-TS
A	1	0
B	2	2

BASIC T/O – EXAMPLE #1

Schedule

TIME

TS(T_1)=1 TS(T_2)=2

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	**COMMIT**

Database

Object	R-TS	W-TS
A	2	0
B	2	2

BASIC T/O – EXAMPLE #1

Schedule

$TS(T_1)=1$ $TS(T_2)=2$

TIME

T_1	T_2
BEGIN	
R(B)	
	BEGIN
	R(B)
	W(B)
R(A)	
	R(A)
	W(A)
COMMIT	COMMIT

Database

Object	R-TS	W-TS
A	2	2
B	2	2

No violations so both txns are safe to commit.

BASIC T/O – EXAMPLE #2

Schedule

TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

Database

Object	R-TS	W-TS
A	0	0
B	0	0

BASIC T/O – EXAMPLE #2

Schedule

TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

Database

Object	R-TS	W-TS
A	1	0
B	0	0

BASIC T/O – EXAMPLE #2

TIME

Schedule

T_1	T_2
BEGIN R(A)	
	BEGIN W(A) COMMIT
W(A) COMMIT	

Database

Object	R-TS	W-TS
A	1	2
B	0	0

BASIC T/O – EXAMPLE #2

Schedule

TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

T_1 cannot overwrite update by T_2, so the DBMS has to abort it!

Database

Object	R-TS	W-TS
A	1	2
B	0	0

Violation:
$TS(T_1) < W\text{-}TS(A)$

CARNEGIE MELLON DATABASE GROUP

THOMAS WRITE RULE

If $\texttt{TS}(\texttt{T}_i) < \texttt{R-TS}(\texttt{X})$:

→ Abort and restart $\texttt{T}_i$.

If $\texttt{TS}(\texttt{T}_i) < \texttt{W-TS}(\texttt{X})$:

→ **Thomas Write Rule**: Ignore the write and allow the txn to continue.

→ This violates timestamp order of $\texttt{T}_i$.

Else:

→ Allow $\texttt{T}_i$ to write $\texttt{X}$ and update $\texttt{W-TS}(\texttt{X})$

BASIC T/O – EXAMPLE #2

Schedule

TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

Database

Object	R-TS	W-TS
A	0	0
B	0	0

BASIC T/O – EXAMPLE #2

Schedule

TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

Database

Object	R-TS	W-TS
A	1	0
B	0	0

BASIC T/O – EXAMPLE #2

Schedule

TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

Database

Object	R-TS	W-TS
A	1	2
B	0	0

BASIC T/O – EXAMPLE #2

Schedule

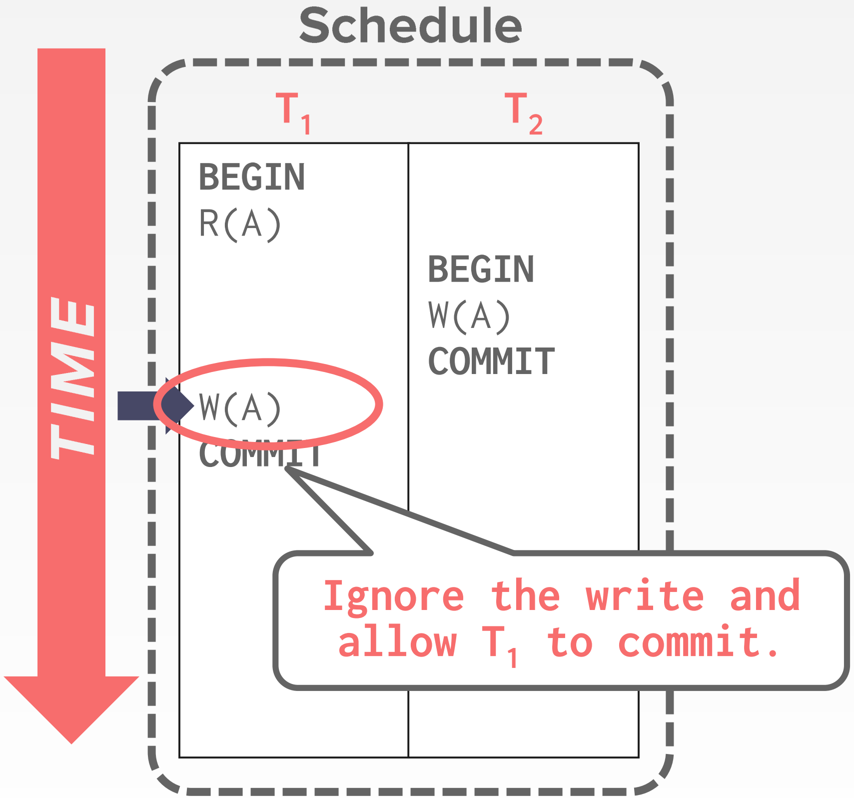
TIME

T_1	T_2
BEGIN	
R(A)	
	BEGIN
	W(A)
	COMMIT
W(A)	
COMMIT	

Database

Object	R-TS	W-TS
A	1	2
B	0	0

BASIC T/O – EXAMPLE #2



Object	R-TS	W-TS
A	1	2
B	0	0

BASIC T/O

Generates a schedule that is conflict serializable if you do **not** use the Thomas Write Rule.

→ No deadlocks because no txn ever waits.

→ Possibility of starvation for long txns if short txns keep causing conflicts.

Permits schedules that are not **recoverable**.

RECOVERABLE SCHEDULES

A schedule is **recoverable** if txns commit only after all txns whose changes they read, commit.

Otherwise, the DBMS cannot guarantee that txns read data that will be restored after recovering from a crash.

RECOVERABLE SCHEDULES

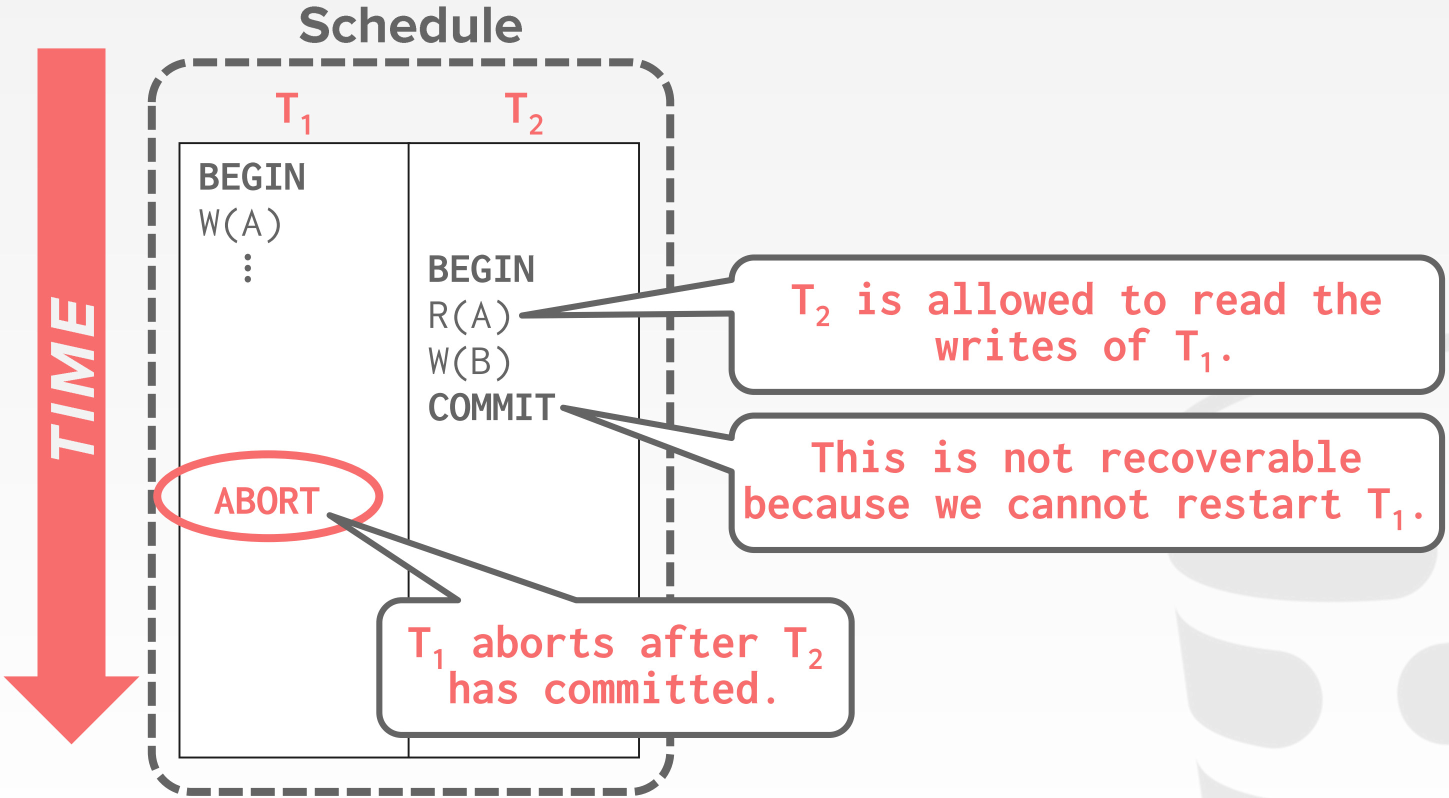
Schedule

TIME

T_1	T_2
BEGIN	
W(A)	
⋮	
	BEGIN
	R(A)
	W(B)
	COMMIT

T_2 is allowed to read the writes of T_1.

RECOVERABLE SCHEDULES



BASIC T/O – PERFORMANCE ISSUES

High overhead from copying data to txn's workspace and from updating timestamps.

Long running txns can get starved.

→ The likelihood that a txn will read something from a newer txn increases.

Suffers from timestamp bottleneck.

OBSERVATION

If you assume that conflicts between txns are **rare** and that most txns are **short-lived**, then forcing txns to wait to acquire locks adds a lot of overhead.

A better approach is to optimize for the no-conflict case.

OPTIMISTIC CONCURRENCY CONTROL

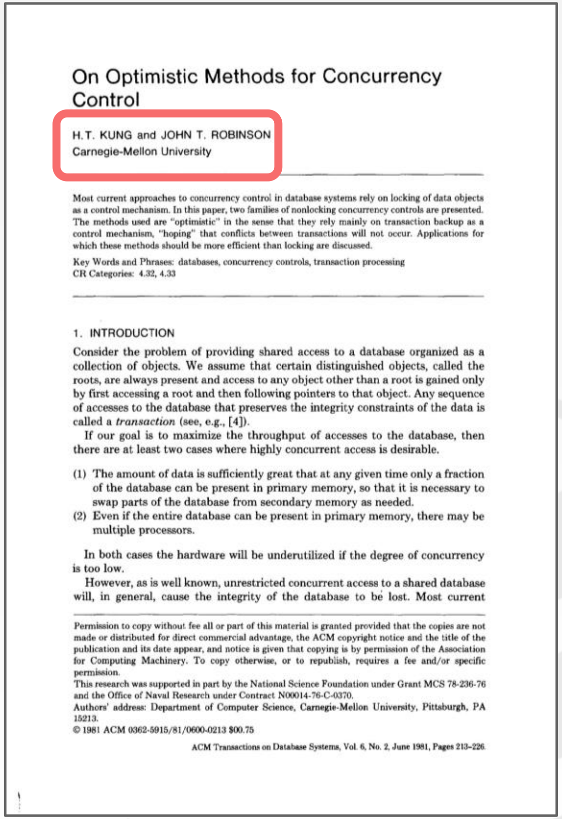
The DBMS creates a private workspace for each txn.

→ All modifications are applied to workspace.

→ Any object read is copied into workspace.

When a txn commits, the DBMS compares its workspace write set to see whether it conflicts with other txns.

If there are no conflicts, the write set is installed into the "global" database.



On Optimistic Methods for Concurrency Control

H.T. KUNG and JOHN T. ROBINSON
Carnegie-Mellon University

Most current approaches to concurrency control in database systems rely on locking of data objects as a control mechanism. In this paper, two families of nonlocking concurrency controls are presented. The methods used are "optimistic" in the sense that they rely mainly on transaction backup as a control mechanism, "hoping" that conflicts between transactions will not occur. Applications for which these methods should be more efficient than locking are discussed.

Key Words and Phrases: databases, concurrency controls, transaction processing
CR Categories: 4.32, 4.33

1. INTRODUCTION

Consider the problem of providing shared access to a database organized as a collection of objects. We assume that certain distinguished objects, called the roots, are always present and access to any object other than a root is gained only by first accessing a root and then following pointers to that object. Any sequence of accesses to the database that preserves the integrity constraints of the data is called a *transaction* (see, e.g., [4]).

If our goal is to maximize the throughput of accesses to the database, then there are at least two cases where highly concurrent access is desirable.

(1) The amount of data is sufficiently great that at any given time only a fraction of the database can be present in primary memory, so that it is necessary to swap parts of the database from secondary memory as needed.
(2) Even if the entire database can be present in primary memory, there may be multiple processors.

In both cases the hardware will be underutilized if the degree of concurrency is too low.

However, as is well known, unrestricted concurrent access to a shared database will, in general, cause the integrity of the database to be lost. Most current

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
This research was supported in part by the National Science Foundation under Grant MCS 78-236-76 and the Office of Naval Research under Contract N00014-76-C-0370.
Authors' address: Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213.
© 1981 ACM 0362-5915/81/0600-0213 $00.75

ACM Transactions on Database Systems, Vol. 6, No. 2, June 1981, Pages 213–226.

CARNEGIE MELLON DATABASE GROUP

OCC PHASES

#1 – Read Phase:
→ Track the read/write sets of txns and store their writes in a private workspace.

#2 – Validation Phase:
→ When a txn commits, check whether it conflicts with other txns.

#3 – Write Phase:
→ If validation succeeds, apply private changes to database. Otherwise abort and restart the txn.

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

Database

Object	Value	W-TS
A	123	0
–	–	–

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

Database

Object	Value	W-TS
A	123	0
-	-	-

OCC – EXAMPLE

TIME

Schedule

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	*READ*
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

Database

Object	Value	W-TS
A	123	0
–	–	–

OCC – EXAMPLE

Schedule

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

TIME

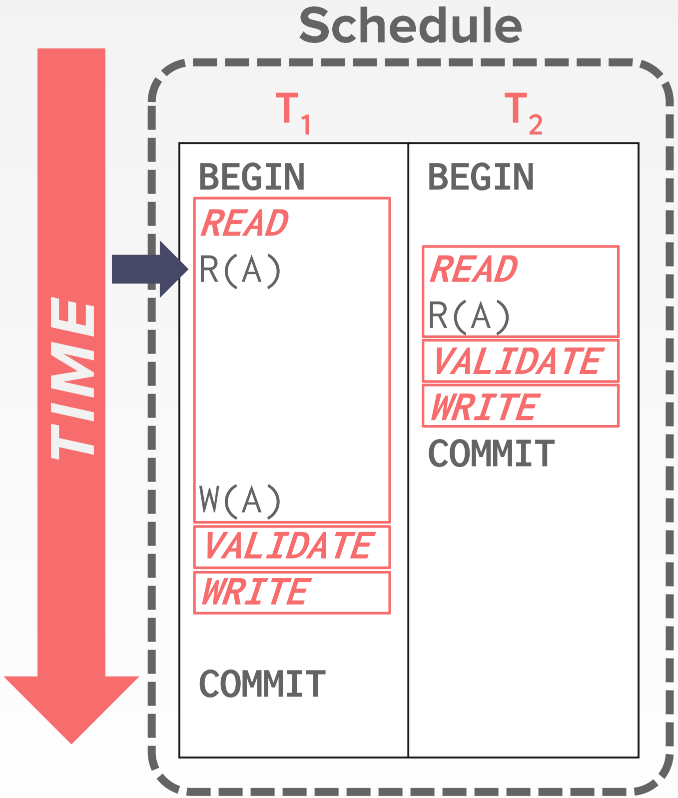
Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
–	–	–
–	–	–

OCC – EXAMPLE



Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

Database

Object	Value	W-TS
A	123	0
-	-	-

T_1 Workspace

Object	Value	W-TS
A	123	0
-	-	-

T_2 Workspace

Object	Value	W-TS
-	-	-
-	-	-

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

Database

Object	Value	W-TS
A	123	0
-	-	-

T_1 Workspace

Object	Value	W-TS
A	123	0
-	-	-

T_2 Workspace

Object	Value	W-TS
A	123	0
-	-	-

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

$TS(T_2)=1$

Database

Object	Value	W-TS
A	123	0
-	-	-

T_1 Workspace

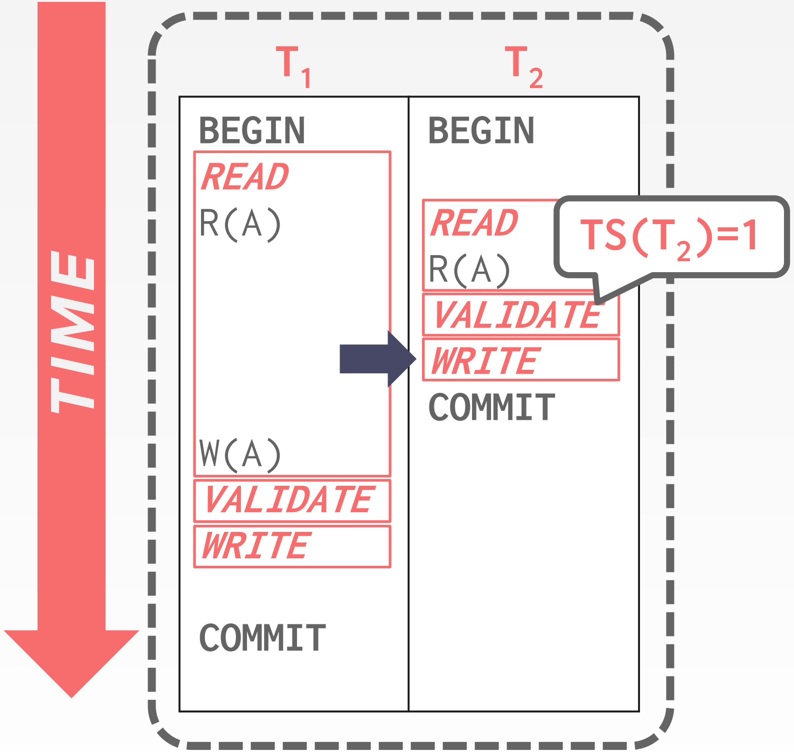
Object	Value	W-TS
A	123	0
-	-	-

T_2 Workspace

Object	Value	W-TS
A	123	0
-	-	-

OCC – EXAMPLE

Schedule



Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	123	0
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	*READ*
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

$TS(T_2)=1$

Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	456	∞
–	–	–

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ R(A)	
	READ R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

$TS(T_2)=1$

Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	456	∞
–	–	–

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
	READ
	R(A)
	VALIDATE
	WRITE
	COMMIT
W(A)	
VALIDATE	
WRITE	
COMMIT	

$TS(T_2)=1$

Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	456	∞
–	–	–

OCC – EXAMPLE

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ R(A)	
	READ R(A)
	VALIDATE
	WRITE
W(A)	
VALIDATE	
WRITE	
COMMIT	

$TS(T_2)=1$

$TS(T_1)=2$

Database

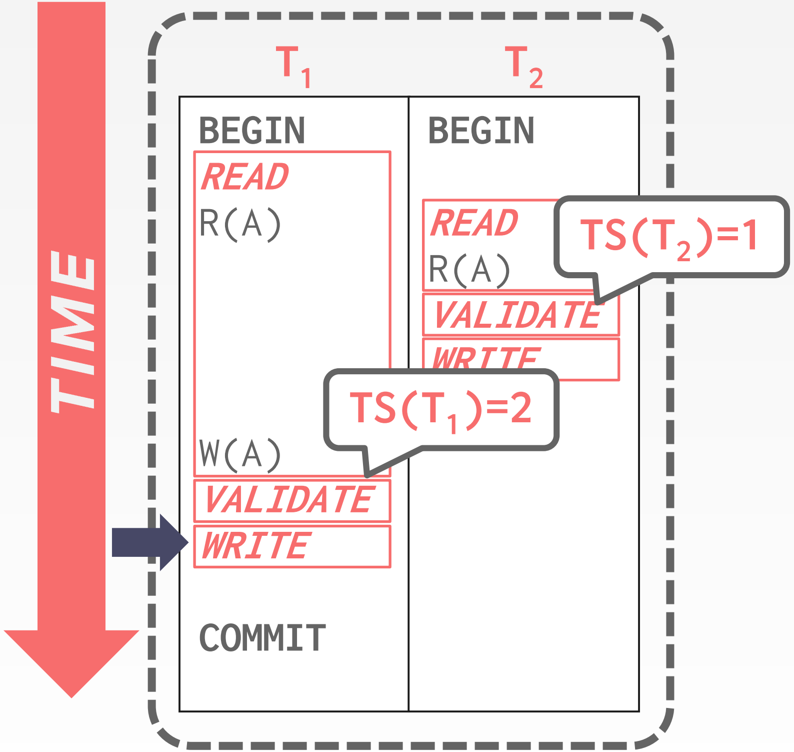
Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	456	∞
–	–	–

OCC – EXAMPLE

Schedule



Database

Object	Value	W-TS
A	456	2
–	–	–

T_1 Workspace

Object	Value	W-TS
A	456	∞
–	–	–

OCC – VALIDATION PHASE

The DBMS needs to guarantee only serializable schedules are permitted.

T_i checks other txns for RW and WW conflicts and makes sure that all conflicts go one way (from older txns to younger txns).

OCC – SERIAL VALIDATION

Maintain global view of all active txns.

Record read set and write set while txns are running and write into private workspace.

Execute **Validation** and **Write** phase inside a protected critical section.

OCC – VALIDATION PHASE

Each txn's timestamp is assigned at the beginning of the validation phase.

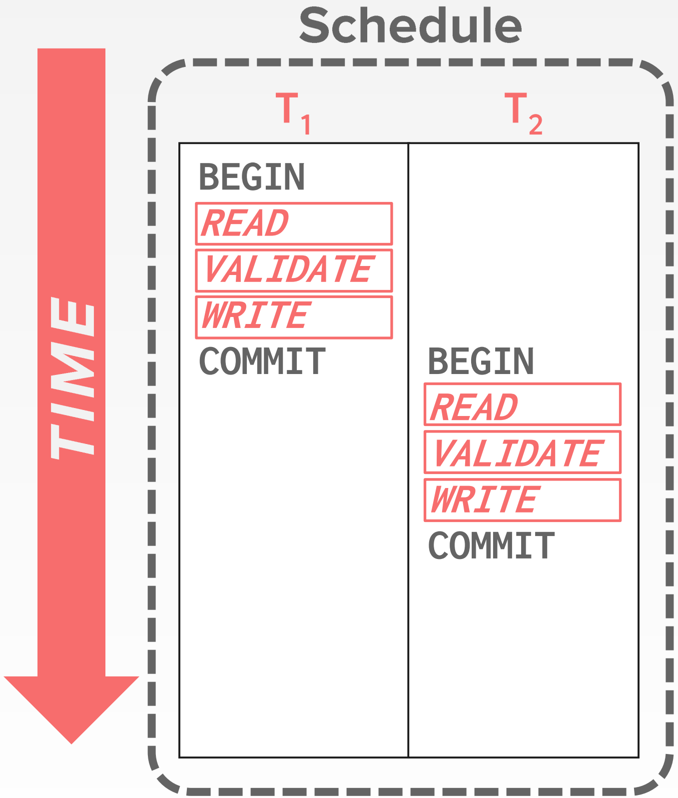
Check the timestamp ordering of the committing txn with all other running txns.

If $TS(T_i) < TS(T_j)$, then one of the following three conditions must hold…

OCC – VALIDATION STEP #1

T_i completes all three phases before T_j begins.

OCC – VALIDATION STEP #1



OCC – VALIDATION STEP #2

T_i completes before T_j starts its **Write** phase, and T_i does not write to any object read by T_j.

$\rightarrow$ $\texttt{WriteSet}(T_i) \cap \texttt{ReadSet}(T_j) = \varnothing$

OCC – VALIDATION STEP #2

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
W(A)	
	READ
	R(A)
VALIDATE	
	VALIDATE
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

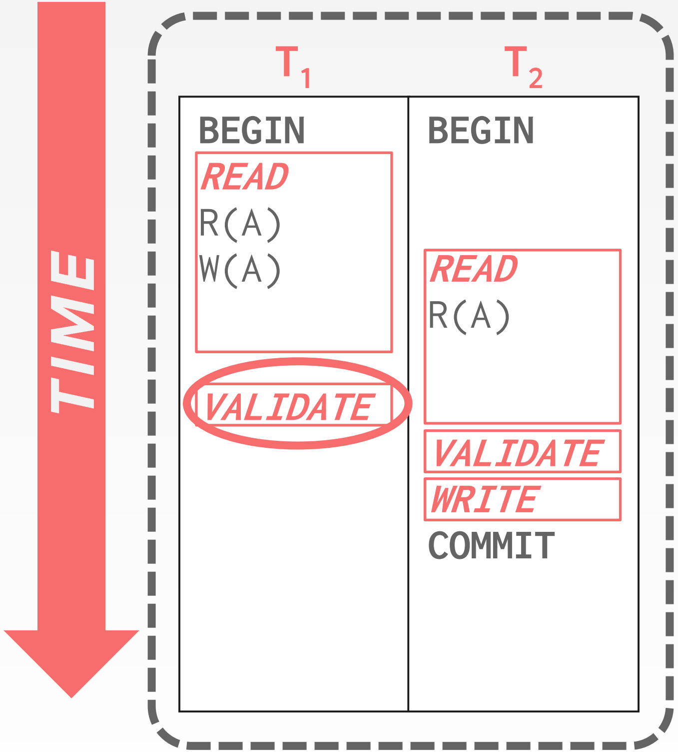
Object	Value	W-TS
A	123	∞
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – VALIDATION STEP #2

Schedule



Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	123	∞
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – VALIDATION STEP #2

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ R(A) W(A)	
	READ R(A)
VALIDATE	
	VALIDATE
	WRITE
	COMMIT

T_1 has to abort even though T_2 will never write to the database.

Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	123	∞
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – VALIDATION STEP #2

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
W(A)	
	READ
	R(A)
	VALIDATE
VALIDATE	
WRITE	
COMMIT	
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

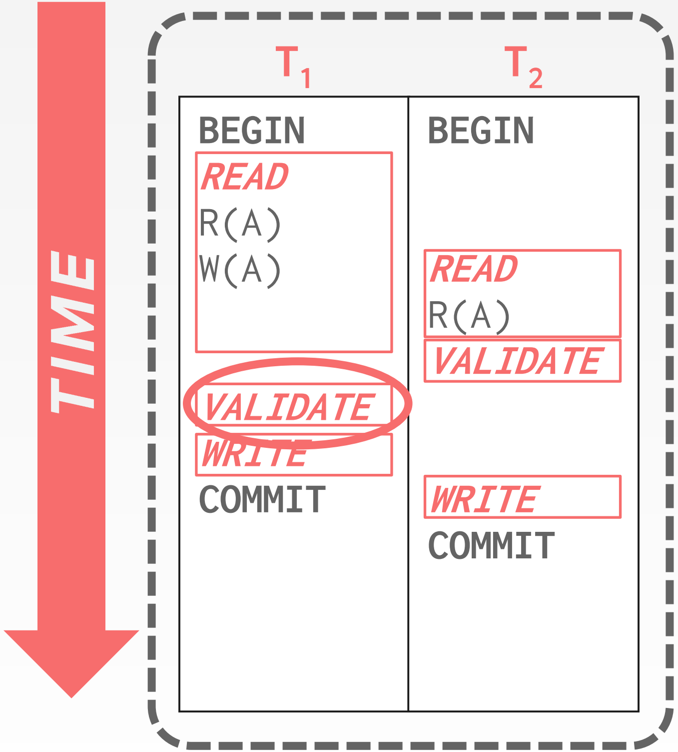
Object	Value	W-TS
A	456	∞
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – VALIDATION STEP #2

Schedule



Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

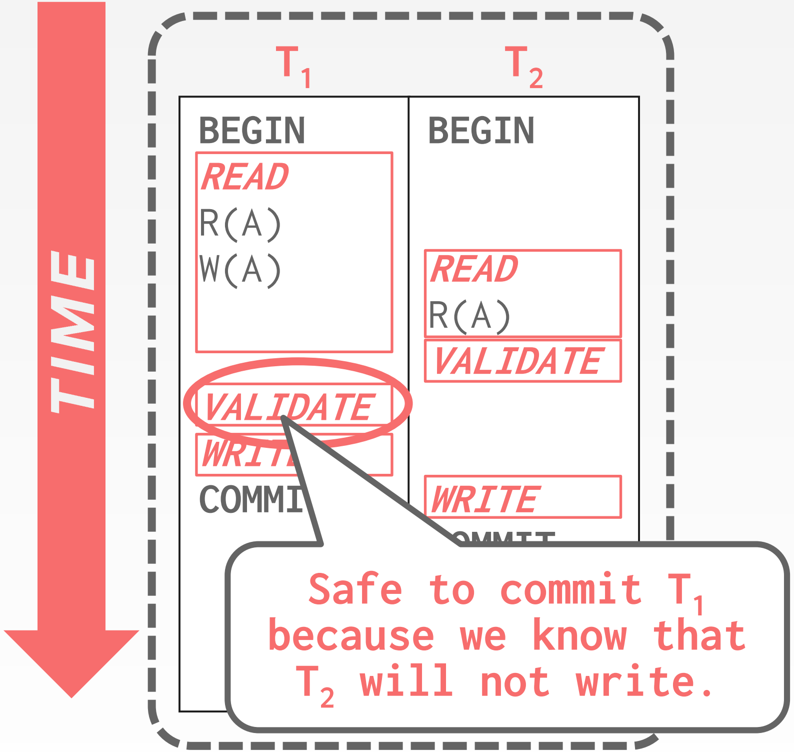
Object	Value	W-TS
A	456	∞
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

OCC – VALIDATION STEP #2

Schedule



Database

Object	Value	W-TS
A	123	0
–	–	–

T_1 Workspace

Object	Value	W-TS
A	456	∞
–	–	–

T_2 Workspace

Object	Value	W-TS
A	123	0
–	–	–

CARNEGIE MELLON
DATABASE GROUP

OCC – VALIDATION STEP #3

T_i completes its **Read** phase before T_j completes its **Read** phase

And T_i does not write to any object that is either read or written by T_j:

→ `WriteSet(`T_i`)` ∩ `ReadSet(`T_j`)` = Ø

→ `WriteSet(`T_i`)` ∩ `WriteSet(`T_j`)` = Ø

OCC – VALIDATION STEP #3

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
W(A)	
	READ
	R(B)
VALIDATE	
WRITE	
COMMIT	
	R(A)
	VALIDATE
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	123	0
B	XYZ	0

T_1 Workspace

Object	Value	W-TS
A	456	∞
-	-	-

T_2 Workspace

Object	Value	W-TS
B	XYZ	0
-	-	-

OCC – VALIDATION STEP #3

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
W(A)	*READ*
	R(B)
VALIDATE	
WRITE	
COMMIT	
	R(A)
	VALIDATE
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	456	1
B	XYZ	0

T_1 Workspace

Object	Value	W-TS
A	456	∞
-	-	-

T_2 Workspace

Object	Value	W-TS
B	XYZ	0
-	-	-

CARNEGIE MELLON DATABASE GROUP

OCC – VALIDATION STEP #3

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
W(A)	
	READ
	R(B)
VALIDATE	
WRITE	
COMMIT	
	R(A)
	VALIDATE
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	456	1
B	XYZ	0

T_1 Workspace

Object	Value	W-TS
–	–	–
–	–	–

T_2 Workspace

Object	Value	W-TS
B	XYZ	0
–	–	–

OCC – VALIDATION STEP #3

Schedule

TIME

T_1	T_2
BEGIN	BEGIN
READ	
R(A)	
W(A)	
	READ
	R(B)
VALIDATE	
WRITE	
COMMIT	
	R(A)
	VALIDATE
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	456	1
B	XYZ	0

T_1 Workspace

Object	Value	W-TS
-	-	-
-	-	-

T_2 Workspace

Object	Value	W-TS
B	XYZ	0
-	-	-

OCC – VALIDATION STEP #3

Schedule

TIME

T1	T2
BEGIN	BEGIN
READ	
R(A)	
W(A)	*READ*
	R(B)
VALIDATE	
WRITE	
COMMIT	
	R(A)
	VALIDATE
	WRITE
	COMMIT

Database

Object	Value	W-TS
A	1	1
B	XYZ	0

T_1 Workspace

Object	Value	W-TS
-	-	-
-	-	-

T_2 Workspace

Object	Value	W-TS
B	XYZ	0
A	456	1

CARNEGIE MELLON DATABASE GROUP

OCC – OBSERVATIONS

When does OCC work well?

When # of conflicts is low:

→ All txns are read-only (ideal).
→ Txns access disjoint subsets of data.

If the database is large and the workload is not skewed, then there is a low probability of conflict, so again locking is wasteful.

OCC – PERFORMANCE ISSUES

High overhead for copying data locally.

Validation/Write phase bottlenecks.

Aborts are more wasteful because they only occur after a txn has already executed.

Suffers from timestamp allocation bottleneck.

OBSERVATION

When a txn commits, all previous T/O schemes check to see whether there is a conflict with concurrent txns.

→ This requires latches.

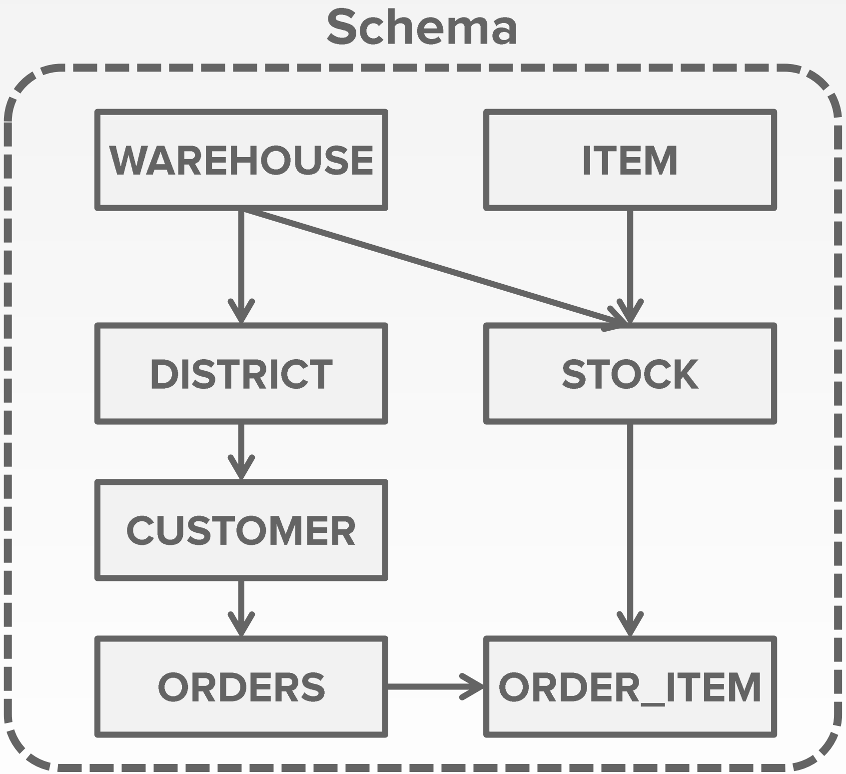
If you have a lot of concurrent txns, then this is slow even if the conflict rate is low.

PARTITION-BASED T/O

Split the database up in disjoint subsets called **partitions** (aka shards).

Only check for conflicts between txns that are running in the same partition.

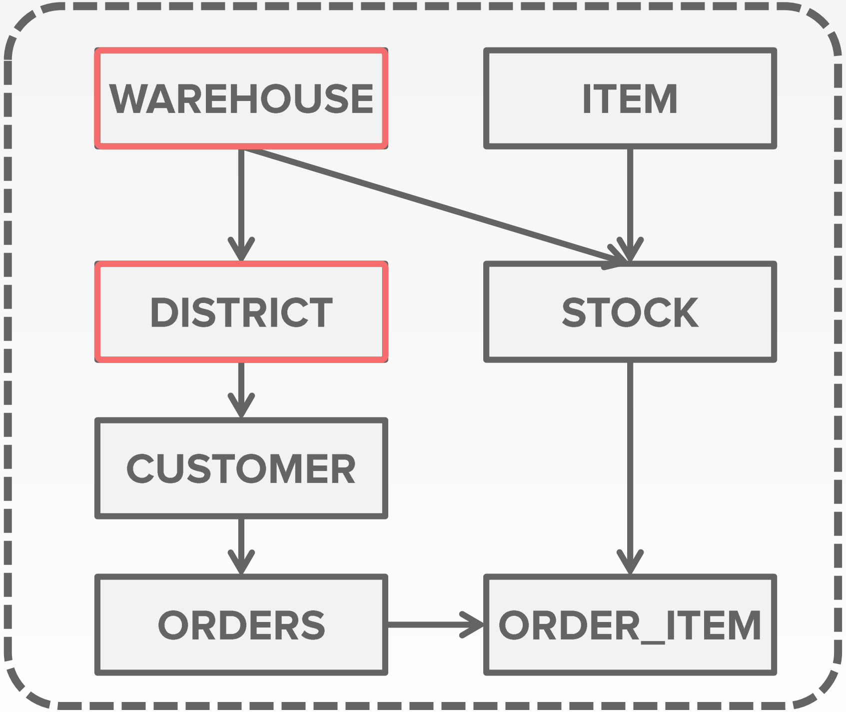
DATABASE PARTITIONING



CARNEGIE MELLON DATABASE GROUP

DATABASE PARTITIONING

Schema

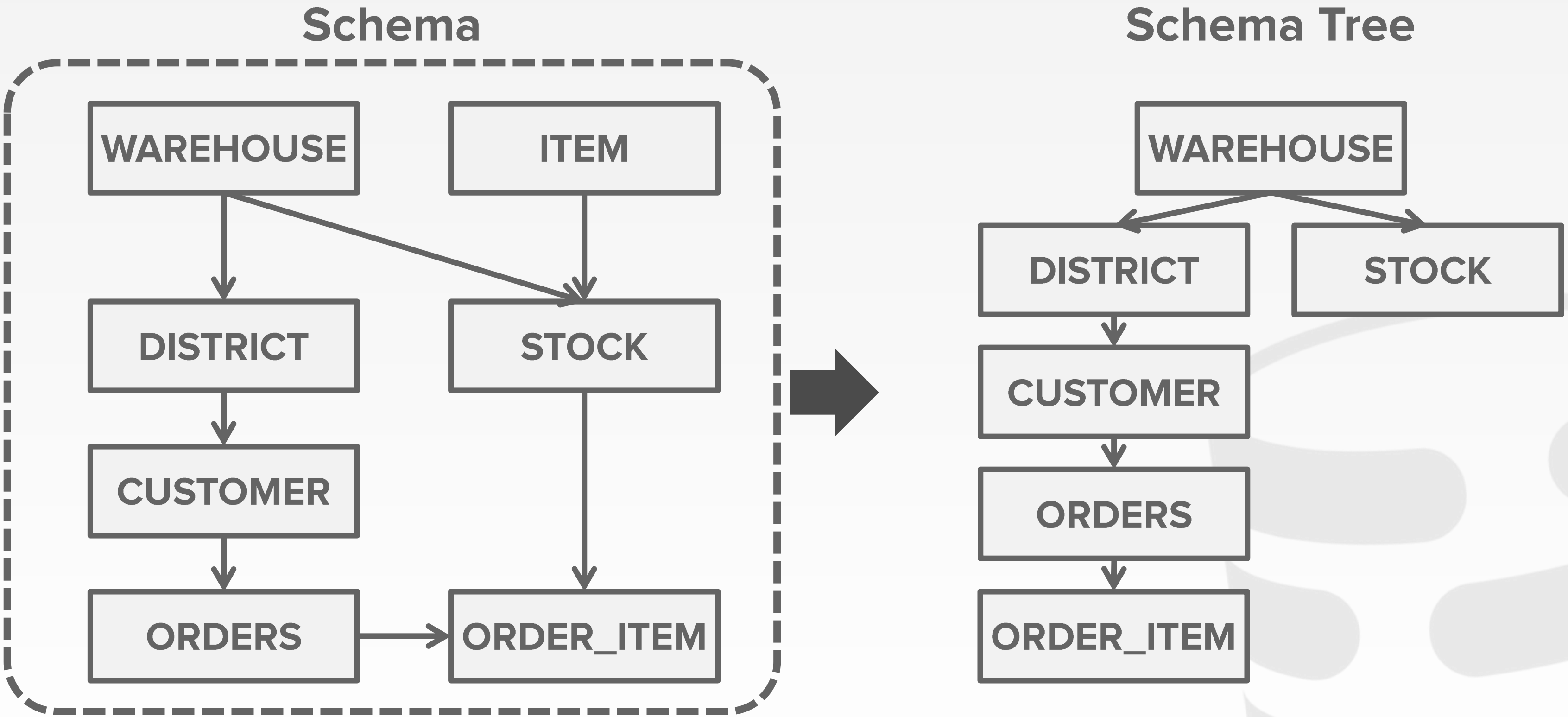



```
CREATE TABLE WAREHOUSE (
  w_id INT PRIMARY KEY,
  w_name VARCHAR UNIQUE,
  ⋮
);
```

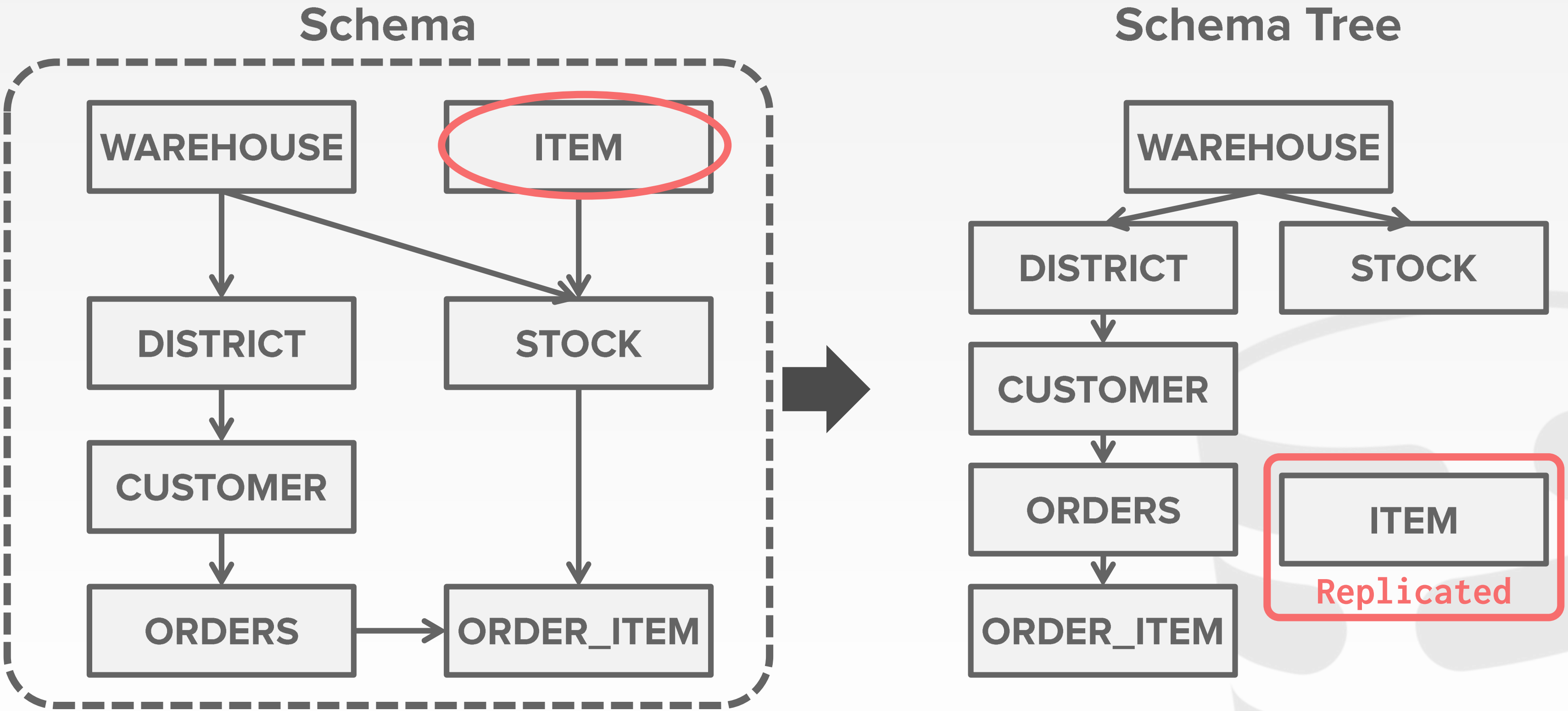
```
CREATE TABLE DISTRICT (
  d_id INT,
  d_w_id INT REFERENCES
             ↪ WAREHOUSE (w_id),
  ⋮
  PRIMARY KEY (d_w_id, d_id)
);
```


CARNEGIE MELLON
DATABASE GROUP

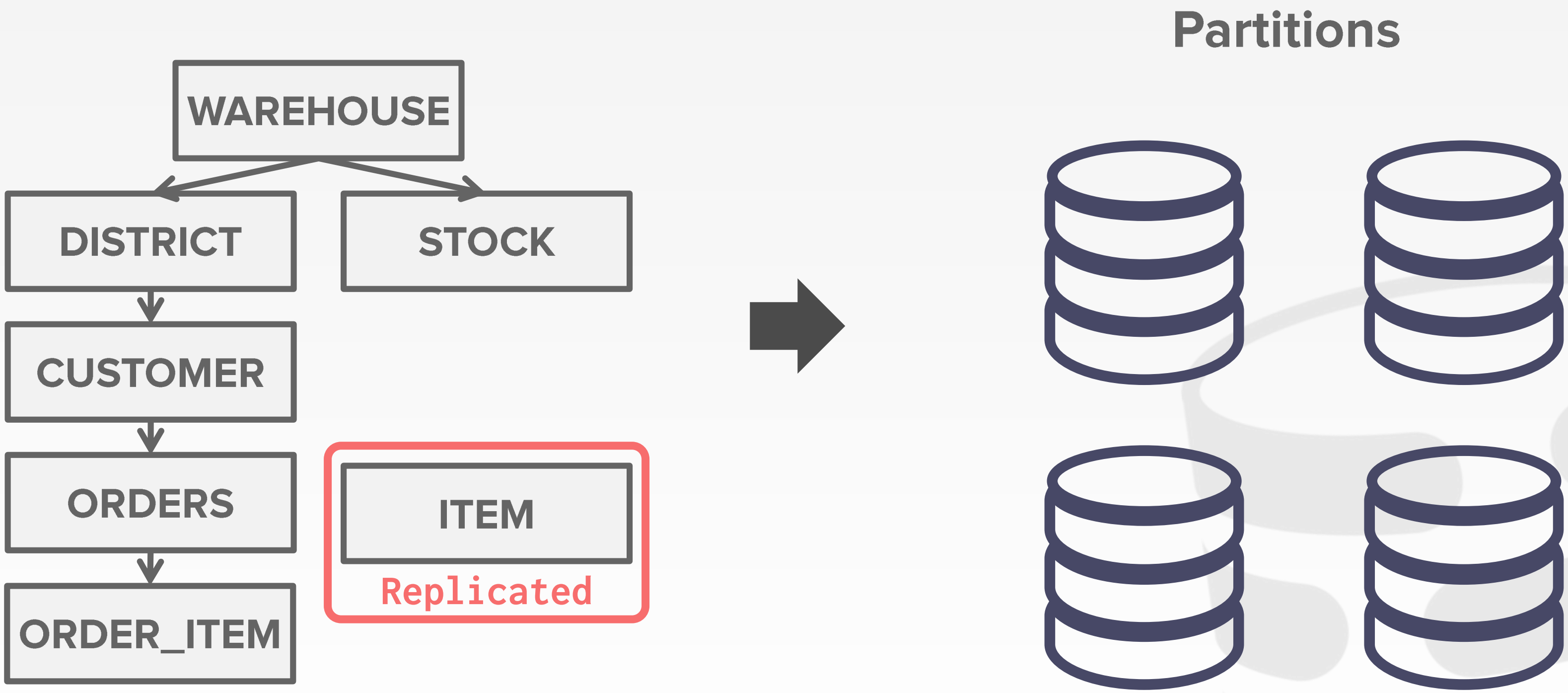
DATABASE PARTITIONING



DATABASE PARTITIONING

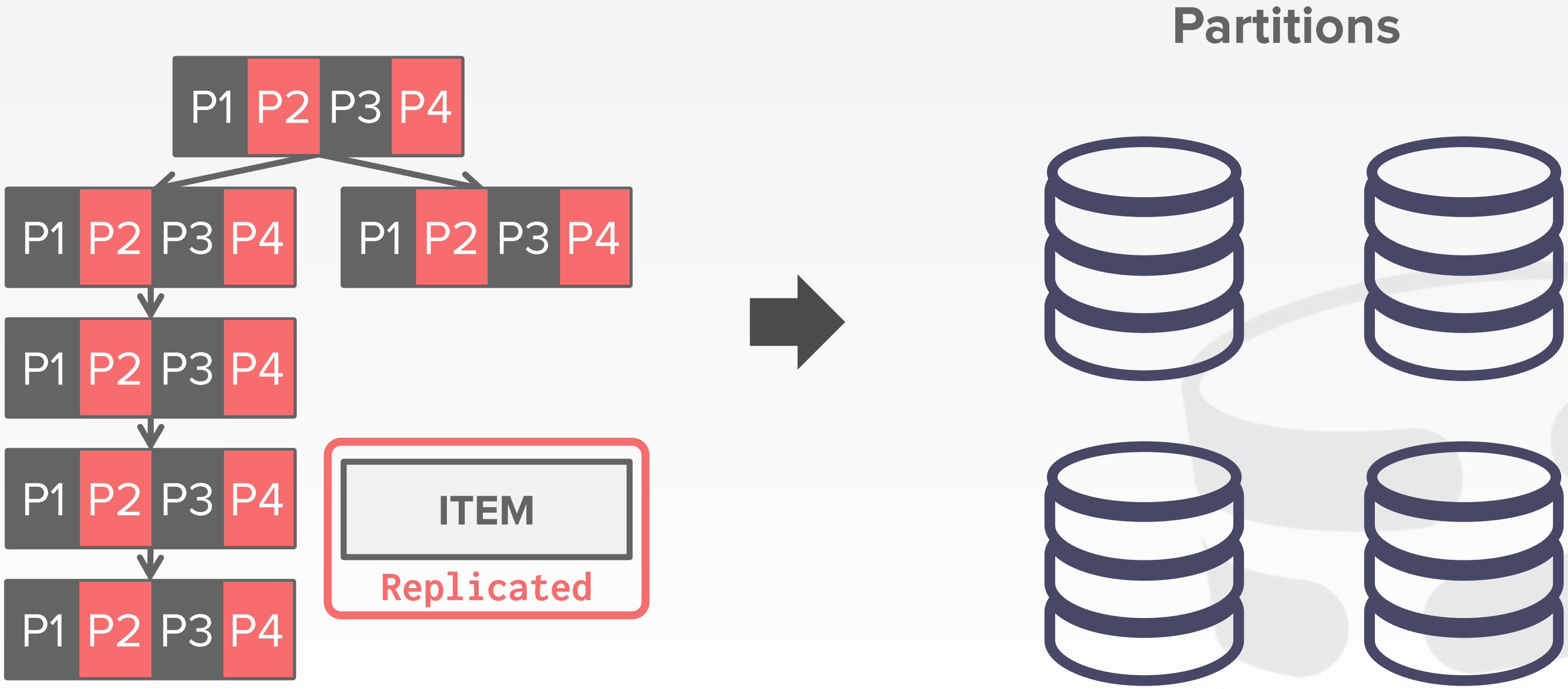


DATABASE PARTITIONING

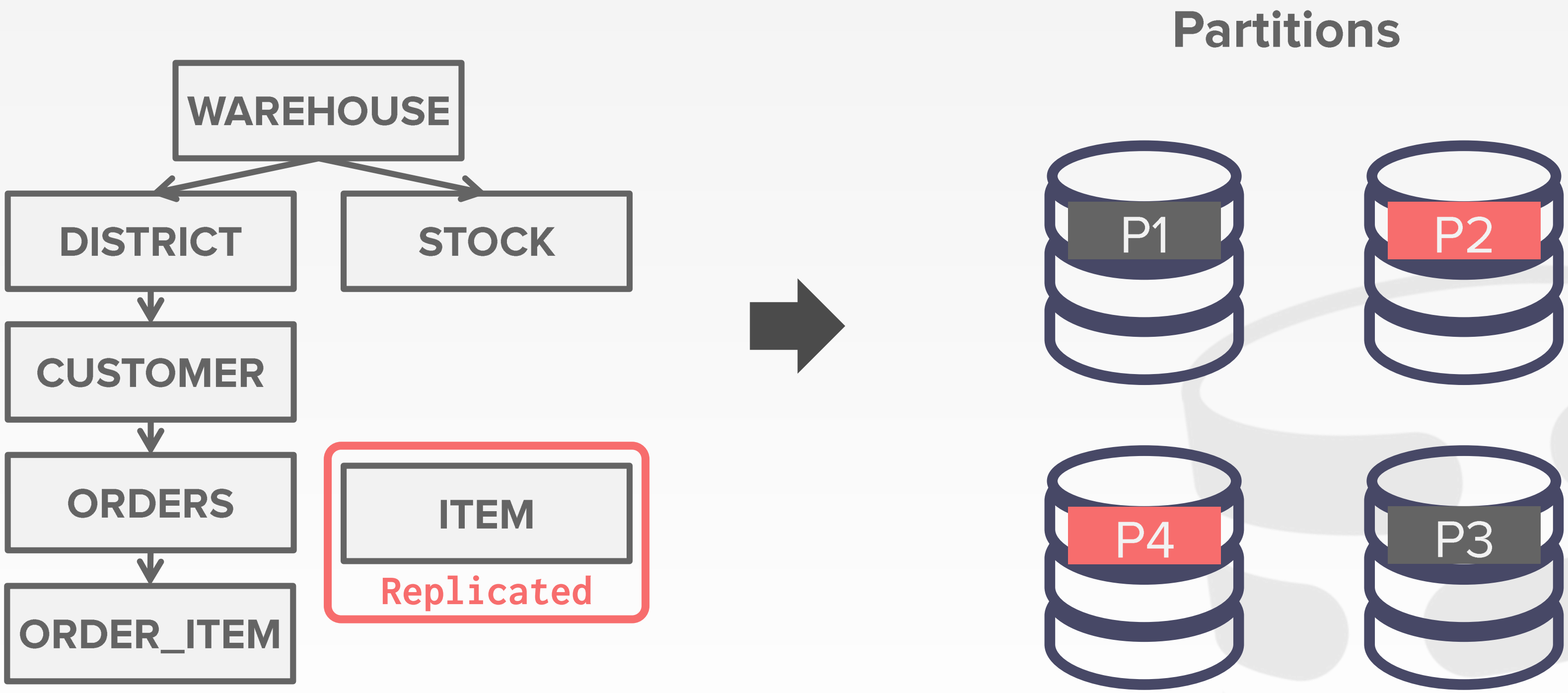


CARNEGIE MELLON
DATABASE GROUP

DATABASE PARTITIONING

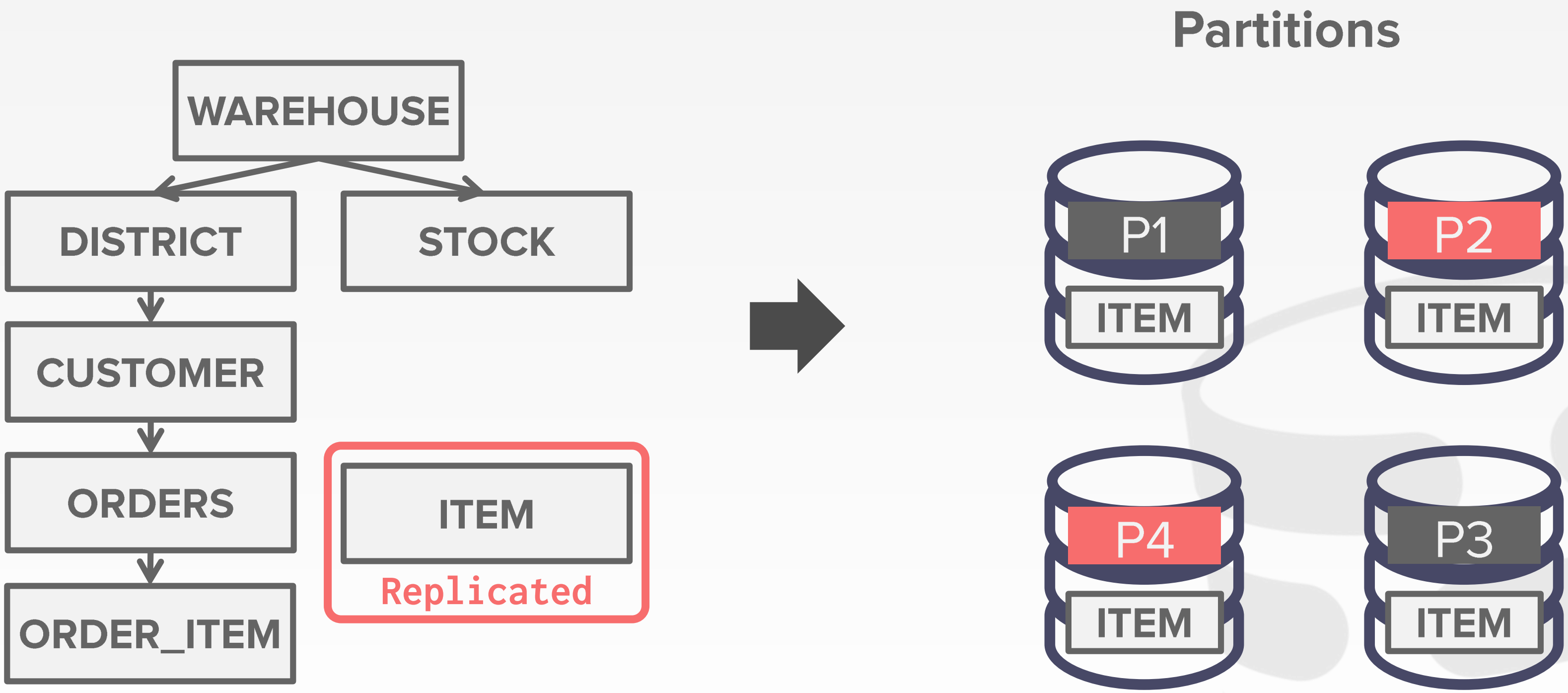


DATABASE PARTITIONING



CARNEGIE MELLON DATABASE GROUP

DATABASE PARTITIONING



PARTITION-BASED T/O

Txns are assigned timestamps based on when they arrive at the DBMS.

Partitions are protected by a single lock:

→ Each txn is queued at the partitions it needs.

→ The txn acquires a partition's lock if it has the lowest timestamp in that partition's queue.

→ The txn starts when it has all of the locks for all the partitions that it will read/write.

PARTITION-BASED T/O – READS

Txns can read anything that they want at the partitions that they have locked.

If a txn tries to access a partition that it does not have the lock, it is aborted + restarted.

PARTITION-BASED T/O – WRITES

All updates occur in place.

→ Maintain a separate in-memory buffer to undo changes if the txn aborts.

If a txn tries to write to a partition that it does not have the lock, it is aborted + restarted.

PARTITION-BASED T/O – PERFORMANCE ISSUES

Partition-based T/O protocol is very fast if:

→ The DBMS knows what partitions the txn needs before it starts.

→ Most (if not all) txns only need to access a single partition.

Multi-partition txns causes partitions to be idle while txn executes.

CONCLUSION

Every concurrency control can be broken down into the basic concepts that I've described in the last two lectures.

I'm not showing benchmark results because I don't want you to get the wrong idea.

NEXT CLASS

Multi-Version Concurrency Control