SCIENCE
PASSION
TECHNOLOGY

Architecture of DB Systems
11 Modern Concurrency Control

Matthias Boehm

Graz University of Technology, Austria
Computer Science and Biomedical Engineering
Institute of Interactive Systems and Data Science
BMK endowed chair for Data Management

Last update: Jan 20, 2021

Announcements/Org

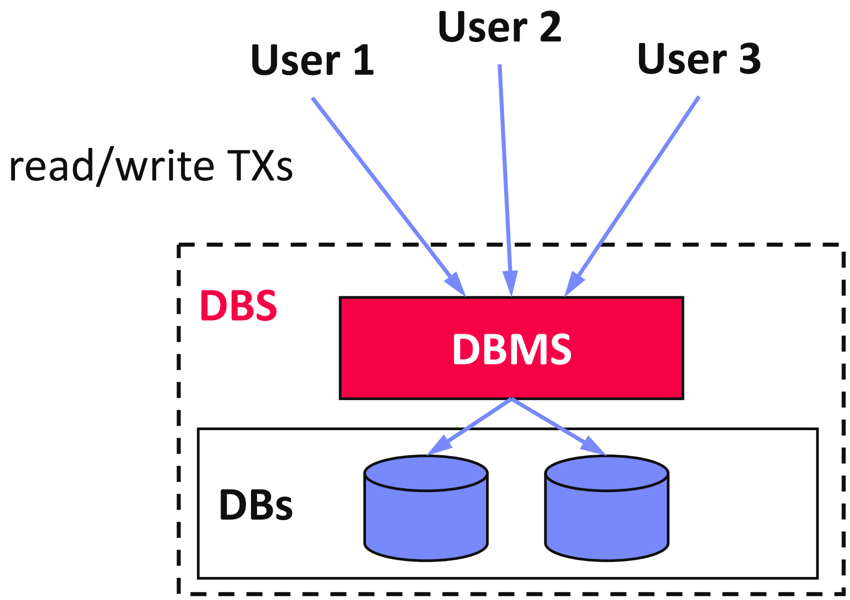
- **#1 Video Recording**
  - Link in TeachCenter & TUbe (lectures will be public)
  - Optional attendance (independent of COVID)
- **#2 COVID-19 Restrictions** (HS i5)
  - Corona Traffic Light: **RED**
  - Temporarily webex lectures until end of semester
- **#3 Course Evaluation and Exam**
  - Evaluation period: **Dec 15 – Jan 31**
  - Exam date: **Feb 19** (virtual webex oral exams, 45min each)

Agenda

- **TX Processing Background**
- **Pessimistic and Optimistic Concurrency Control**
- **Multi-Version Concurrency Control**
- **Excursus: Coordination Avoidance**

TX Processing Background

Transaction (TX) Processing



#1 Multiple users
➔ **Correctness?**

#2 Various failures
(TX, system, media)
➔ **Reliability?**

Deadlocks

Constraint violations

Network failure

Crash/power failure

Disk failure

- **Goal:** Transaction Processing
  - **#1** Locking and concurrency control to ensure **#1 correctness**
  - **#2** Logging and recovery to ensure **#2 reliability**

Terminology of Transactions

- **Database Transaction**
  - A transaction (TX) is a series of steps that brings a database from a consistent state into another (not necessarily different) consistent state
  - ACID properties (atomicity, consistency, isolation, durability)

- **Terminology by Example**

#2 Start/begin of TX (BOT/BT)

#1 Isolation level (defined by addressed anomalies)

```
START TRANSACTION ISOLATION LEVEL SERIALIZABLE;
  UPDATE Account SET Balance=Balance-100
     WHERE AID = 107;
  UPDATE Account SET Balance=Balance+100
     WHERE AID = 999;

  SELECT Balance INTO lbalance
     FROM Account WHERE AID=107;
  IF lbalance < 0 THEN
     ROLLBACK TRANSACTION;
  END IF
COMMIT TRANSACTION;
```

#3 Reads and writes of data objects

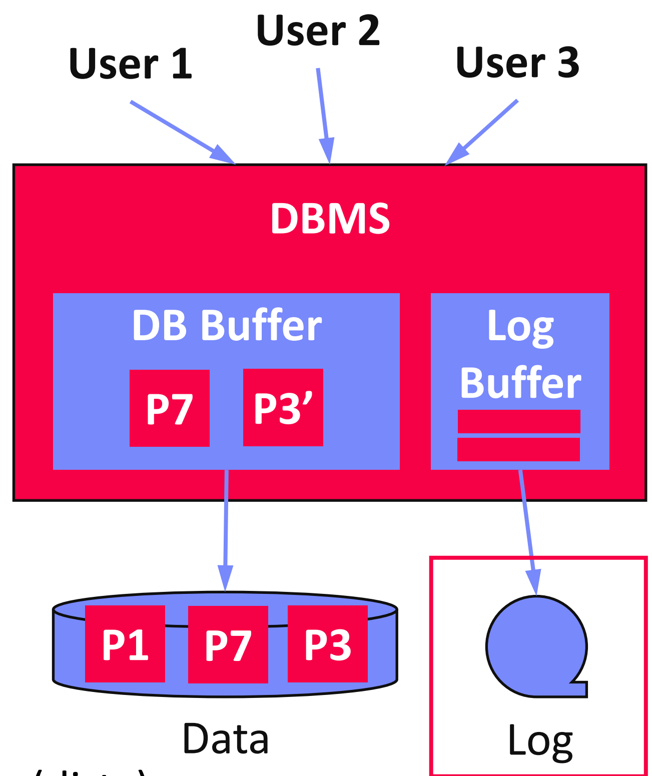
#6 Savepoints (checkpoint for partial rollback)

#4 Abort/rollback TX (unsuccessful end of transaction, EOT/ET)

#5 Commit TX (successful end of transaction, EOT/ET)

Database (Transaction) Log

- **Database Architecture**
  - Page-oriented storage on disk and in memory (DB buffer)
  - Dedicated **eviction algorithms**
  - Modified in-memory pages marked as dirty, flushed by cleaner thread
  - Log: append-only TX changes
  - Data/log often placed on different devices and periodically archived (backup + truncate)



- **Write-Ahead Logging (WAL)**
  - The log records representing changes to some (dirty) data page must be on **stable storage before the data page** (UNDO - atomicity)
  - **Force-log on commit** or full buffer (REDO - durability)
  - **Recovery:** forward (REDO) and backward (UNDO) processing
  - Log sequence number (LSN)

[C. Mohan, Donald J. Haderle, Bruce G. Lindsay, Hamid Pirahesh, Peter M. Schwarz: ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging. **TODS 1992**]

Isolation Levels

- **Different Isolation Levels**
  - **Tradeoff Isolation vs performance** per session/TX
  - SQL standard requires **guarantee against lost updates** for all

```
SET TRANSACTION
  ISOLATION LEVEL
  READ COMMITTED
```

- **SQL Standard Isolation Levels**

Isolation Level	Lost Update	Dirty Read (P1)	Unrepeatable Read (P2)	Phantom Read (P3)
READ UNCOMMITTED	No*	Yes	Yes	Yes
READ COMMITTED	No*	No	Yes	Yes
REPEATABLE READ	No*	No	No	Yes
[SERIALIZABLE]	No*	No	No	No

* Lost update potentially w/ different semantics in standard

  - Serializable w/ highest guarantees (**pseudo-serial execution**)

- **How can we enforce these isolation levels?**
  - **User:** set default/transaction isolation level (mixed TX workloads possible)
  - **System:** dedicated concurrency control strategies + scheduler

Excursus: A Critique of SQL Isolation Levels

[Hal Berenson, Philip A. Bernstein, Jim Gray, Jim Melton, Elizabeth J. O'Neil, Patrick E. O'Neil: A Critique of ANSI SQL Isolation Levels. **SIGMOD 1995**]

- **Summary**
  - **Criticism:** SQL standard isolation levels are ambiguous (strict/broad interpretations)
  - Additional anomalies: dirty write, cursor lost update, fuzzy read, read skew, write skew
  - Additional isolation levels: **cursor stability** and **snapshot isolation**
- **Snapshot Isolation (< Serializable)**
  - **Type of optimistic concurrency control** via multi-version concurrency control
  - TXs reads data from a snapshot of committed data when TX started
  - **TXs never blocked on reads**, other TXs data invisible
  - TX **T1 only commits if no other TX wrote the same data items** in the time interval of T1
- **Current Status?**

  [http://dbmsmusings.blogspot.com/2019/05/introduction-to-transaction-isolation.html]

  - "SQL standard that **fails to accurately define database isolation levels** and database vendors that attach liberal and non-standard semantics"

Excursus: Isolation Levels in Practice

- **Default and Maximum Isolation Levels for "ACID" and "NewSQL" DBs** [as of 2013]
  - 3/18 `SERIALIZABLE` by default
  - 8/18 did not provide `SERIALIZABLE` at all

[Peter Bailis, Alan Fekete, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica: HAT, Not CAP: Towards Highly Available Transactions. **HotOS 2013**]

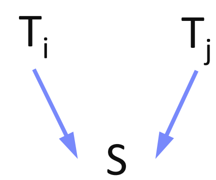
Beware of defaults, even though the SQL standard says **`SERIALIZABLE`** is the default

Database	Default	Maximum
Actian Ingres 10.0/10S [1]	S	S
Aerospike [2]	RC	RC
Akiban Persistit [3]	SI	SI
Clustrix CLX 4100 [4]	RR	RR
Greenplum 4.1 [8]	RC	S
IBM DB2 10 for z/OS [5]	CS	S
IBM Informix 11.50 [9]	Depends	S
MySQL 5.6 [12]	RR	S
MemSQL 1b [10]	RC	RC
MS SQL Server 2012 [11]	RC	S
NuoDB [13]	CR	CR
Oracle 11g [14]	RC	SI
Oracle Berkeley DB [7]	S	S
Oracle Berkeley DB JE [6]	RR	S
Postgres 9.2.2 [15]	RC	S
SAP HANA [16]	RC	SI
ScaleDB 1.02 [17]	RC	RC
VoltDB [18]	S	S

RC: read committed, RR: repeatable read, SI: snapshot isolation, S: serializability, CS: cursor stability, CR: consistent read

Serializability Theory

- **Operations of Transaction T_j**
  - Read and write operations of A by T_j: **$r_j(A)$ $w_j(A)$**
  - Abort of transaction T_j: **a_j** (unsuccessful termination of T_j)
  - Commit of transaction T_j: **c_j** (successful termination of T_j)
- **Schedule S**
  - Operations of a transaction **T_j are executed in order**
  - Multiple transactions may be executed concurrently
  - ➔ **Schedule describes the total ordering of operations**



- **Equivalence of Schedules S1 and S2**
  - Read-write, write-read, and write-write dependencies on data object A executed in same order:

$$r_i(A) <_{S1} w_j(A) \Leftrightarrow r_i(A) <_{S2} w_j(A)$$
$$w_i(A) <_{S1} r_j(A) \Leftrightarrow w_i(A) <_{S2} r_j(A)$$
$$w_i(A) <_{S1} w_j(A) \Leftrightarrow w_i(A) <_{S2} w_j(A)$$

Serializability Theory, cont.

- **Example Serializable Schedules**
  - Input TXs
    T1: BOT $r_1(A)$ $w_1(A)$ $r_1(B)$ $w_1(B)$ c_1
    T2: BOT $r_2(C)$ $w_2(C)$ $r_2(A)$ $w_2(A)$ c_2
  - Serial execution
    $r_1(A)$ $w_1(A)$ $r_1(B)$ $w_1(B)$ c_1 $r_2(C)$ $w_2(C)$ $r_2(A)$ $w_2(A)$ c_2
  - Equivalent schedules
    $r_1(A)$ $r_2(C)$ $w_1(A)$ $w_2(C)$ $r_1(B)$ $r_2(A)$ $w_1(B)$ $w_2(A)$ c_1 c_2
    $r_1(A)$ $w_1(A)$ $r_2(C)$ $w_2(C)$ $r_1(B)$ $w_1(B)$ $r_2(A)$ $w_2(A)$ c_1 c_2
  - Wrong schedule
    $r_1(A)$ $r_2(C)$ $w_2(C)$ $r_2(A)$ $w_1(A)$ $r_1(B)$ $w_1(B)$ $w_2(A)$ c_1 c_2

- **Serializability Graph (conflict graph)**
  - Operation dependencies (read-write, write-read, write-write) aggregated
  - **Nodes:** transactions; **edges:** transaction dependencies
  - **Transactions are serializable** (via topological sort) **if the graph is acyclic**
  - **Beware:** Serializability Theory considers only successful transactions, which disregards anomalies like dirty read that might happen in practice

Pessimistic and Optimistic Concurrency Control

Overview Concurrency Control

- **Terminology**
  - **Lock:** logical synchronization of TXs access to database objects (row, table, etc)
  - **Latch:** physical synchronization of access to shared data structures
- **#1 Pessimistic Concurrency Control**
  - Locking schemes (lock-based database scheduler)
  - Full serialization of transactions
- **#2 Optimistic Concurrency Control (OCC)**
  - Optimistic execution of operations, check of conflicts (validation)
  - Optimistic and timestamp-based database schedulers
- **#3 Mixed Concurrency Control (e.g., PostgreSQL)**
  - Combines locking and OCC
  - Might return **synchronization errors**

```
ERROR: could not serialize access
       due to concurrent update
ERROR: deadlock detected
```

Locking Schemes

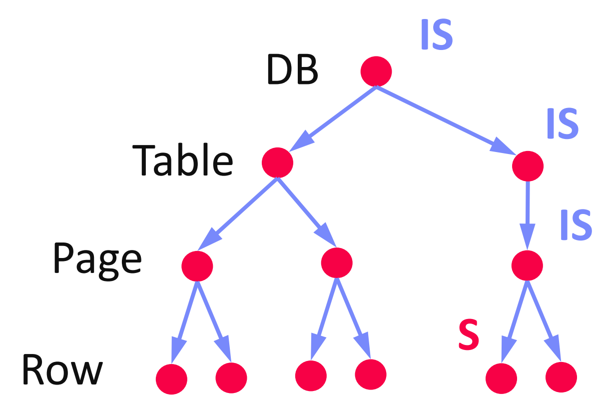
- **Compatibility of Locks**
  - X-Lock (exclusive/write lock)
  - S-Lock (shared/read lock)

Existing Lock / Requested Lock

	None	S	X
S	Yes	Yes	No
X	Yes	No	No

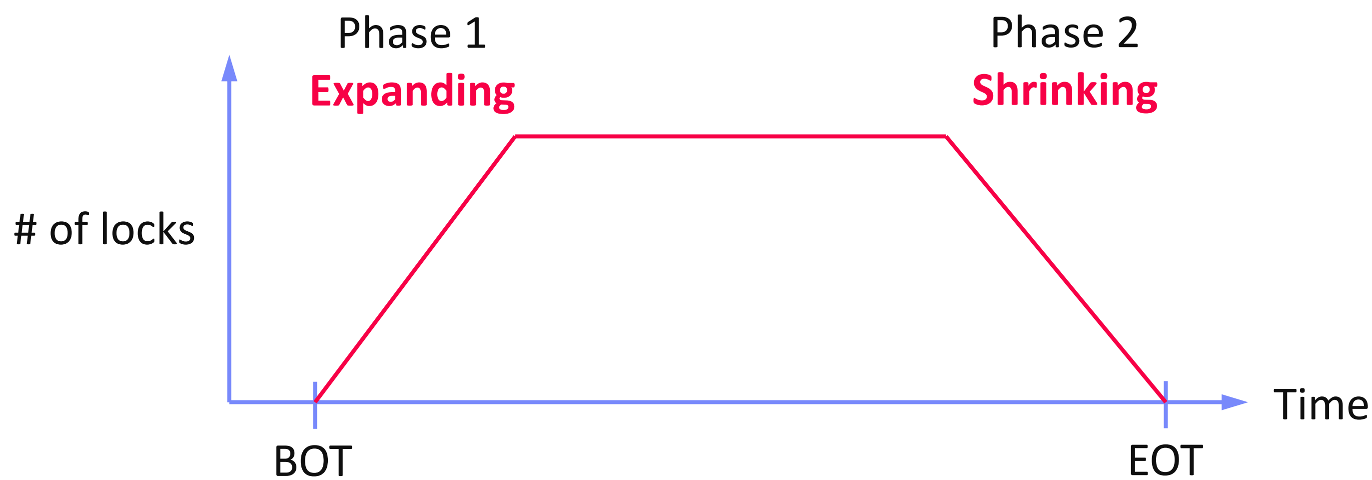
- **Multi-Granularity Locking**
  - Hierarchy of DB objects
  - Additional intentional IX and IS locks

	None	S	X	IS	IX
S	Yes	Yes	No	Yes	No
X	Yes	No	No	No	No
IS	Yes	Yes	No	Yes	Yes
IX	Yes	No	No	Yes	Yes



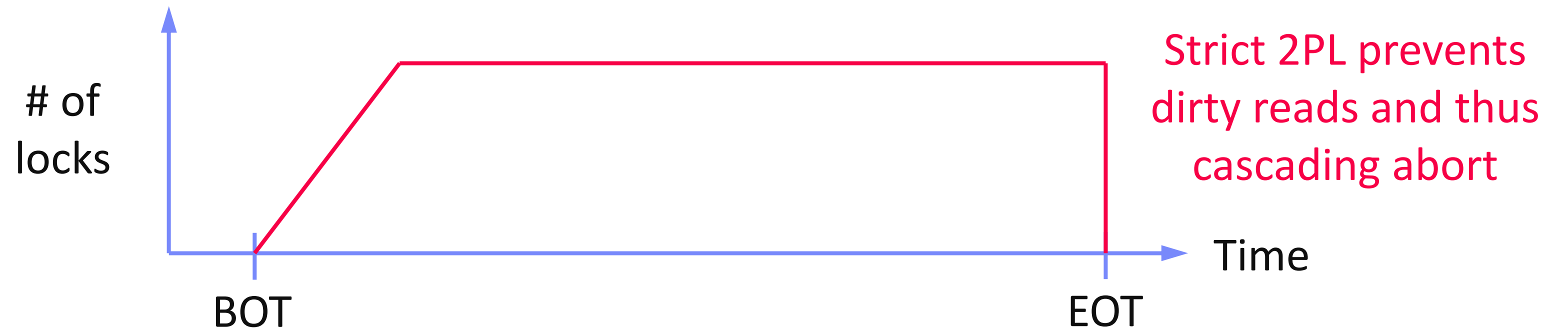
Two-Phase Locking (2PL)

- **Overview**
  - 2PL is a concurrency protocol that guarantees **SERIALIZABLE**
  - **Expanding phase** (growing): acquire locks needed by the TX
  - **Shrinking phase**: release locks acquired by the TX
    (can only start if all needed locks acquired)

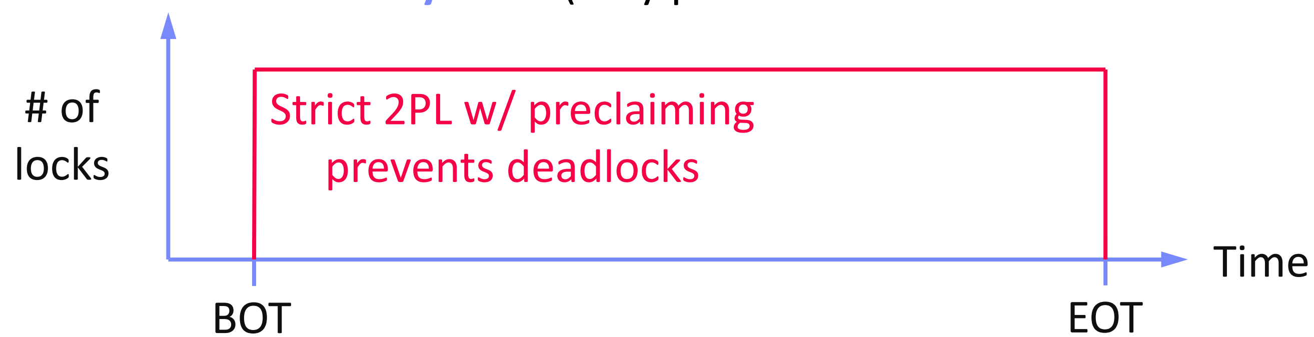


Two-Phase Locking, cont.

- **Strict 2PL (S2PL) and Strong Strict 2PL (SS2PL)**
  - **Problem:** Transaction rollback can cause (**Dirty Read**)
  - Release all X-locks (S2PL) or X/S-locks (SSPL) **at end of transaction (EOT)**

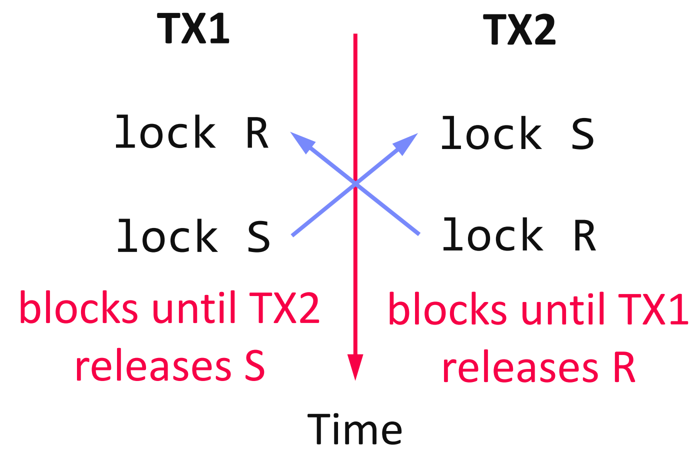


- **Strict 2PL w/ pre-claiming (aka conservative 2PL)**
  - Problem: incremental expanding can cause deadlocks for interleaved TXs
  - **Pre-claim all necessary locks** (only possible if entire TX known + **latches**)



2PL – Deadlocks

- **Deadlock Scenario**
  - Deadlocks of concurrent transactions
  - Deadlocks happen due to **cyclic dependencies without pre-claiming** (wait for exclusive locks)



DEADLOCK, as this will never happen

- **#1 Deadlock Prevention**
  - Pre-claiming (guarantee if TX known upfront)

[Philip A. Bernstein, Nathan Goodman: Concurrency Control in Distributed Database Systems. **ACM Comput. Surv. 1981**]

- **#2 Deadlock Avoidance**
  - Preemptive vs non-preemptive strategies
  - NO_WAIT (if deadlock suspected wrt timestamp TS, abort lock-requesting TX)
  - WOUND-WAIT (T1 locks something held by T2 → if T1<T2, restart T2)
  - WAIT-DIE (T1 locks something held by T2 → if T1>T2, abort T1 but keep TS)
- **#3 Deadlock Detection (DL_DETECT)**
  - Maintain a wait-for graph of blocked TX (similar to serializability graph)
  - Detection of cycles in graph (on timeout) → abort one or many TXs

Basic Timestamp Ordering (BTO)

[Philip A. Bernstein, Nathan Goodman: Concurrency Control in Distributed Database Systems. **ACM Comput. Surv. 1981**]

- **Synchronization Scheme**
  - Transactions get timestamp (or version) **$TS(T_j)$** at BOT
  - Each data object A has **readTS(A)** and **writeTS(A)**
  - Use timestamp comparison to validate access → serialized schedule
- **Read Protocol $T_j(A)$**
  - If $TS(T_j)$ >= writeTS(A): **allow read**, set readTS(A) = max($TS(T_j)$, readTS(A))
  - If $TS(T_j)$ < writeTS(A): **abort T_j** (older than last modifying TX)
- **Write Protocol $T_j(A)$**
  - If $TS(T_j)$ >= readTS(A) & $TS(T_j)$ >= writeTS(A): **allow write**, set writeTS(A)=$TS(T_j)$
  - If $TS(T_j)$ < readTS(A): **abort T_j** (older than last reading TX)
  - If $TS(T_j)$ < writeTS(A): **abort T_j** (older than last modifying TX)
- **BEWARE:** BTO requires handling of dirty reads, recoverability in general (e.g., via abort or versions)
  - Strict Timestamp Ordering (dirty bit) w/ deadlock avoidance techniques

[Stephan Wolf et al: An Evaluation of Strict Timestamp Ordering Concurrency Control for Main-Memory Database Systems. **IMDM@ VLDB 2013** (Revised Selected Papers)]

Excursus: BTO in Project Reference Impl

- **Overview TX Processing**
  - Implements variant of **basic timestamp ordering** (w/ handling of dirty reads)
  - **TX log for UNDO** of aborted transactions
  - **TIDs:** `__sync_fetch_and_add(&VAR,1)`

```
./speed_test 1468 0 0 0 0 \
             4000 160000 100
```

- **#1 Basic TO**
  - isReadable: TID >= WTS
  - IsWriteable: TID >= max(WTS, RTS)

```
NUM_TXN_FAIL: 0
NUM_TXN_COMP: 16,000,000
Time to run:  15.223s.
```

- **#2 Basic TO w/ Read Committed**
  - Basic TO w/ isReadable: TID >= WTS && !(TID != WTS && scanTXTable(ix, WTS))

```
NUM_TXN_FAIL: 0
NUM_TXN_COMP: 16,000,000
Time to run:  15.394s.
```

- **#3 Basic TO w/ Serializable**
  - Basic TO w/ read committed
  - Deleted bit, forced cleanup in epochs (∄ TS < max(RTS,WTS))

NotImplementedException

Optimistic Concurrency Control (OCC)

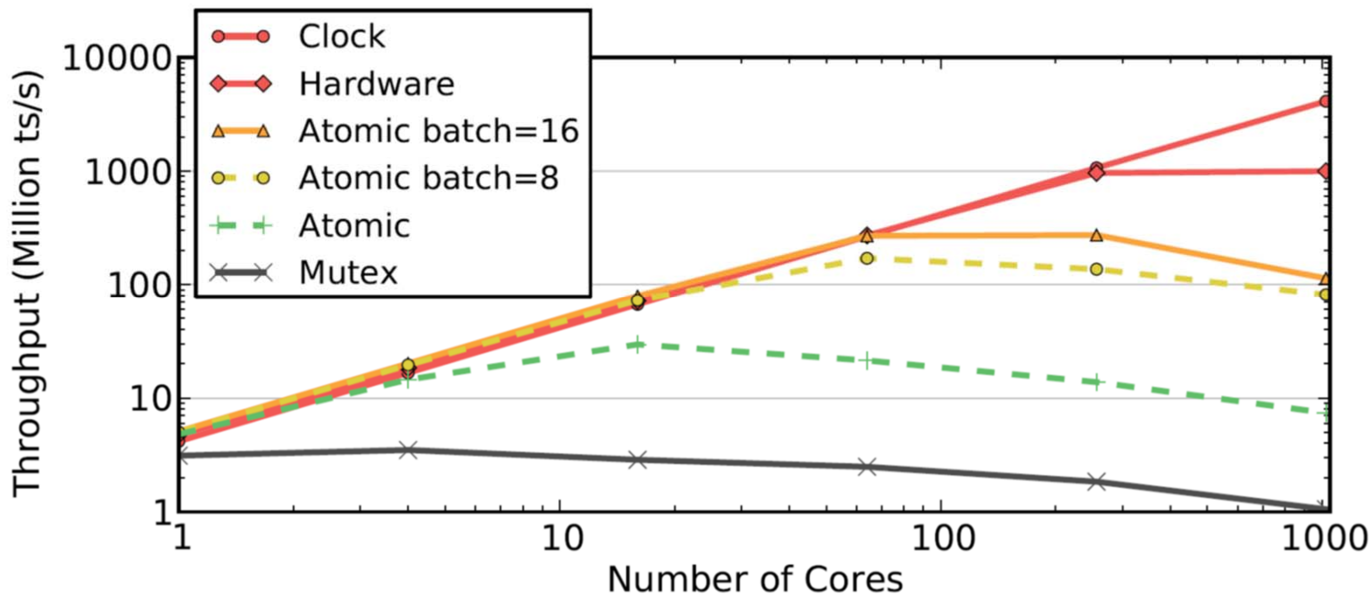
- **#1 Read Phase**
  - Initial reads from DB, **repeated reads and writes into TX-local buffer**
  - Maintain **ReadSet(T_j)** and **WriteSet(T_j)** per transaction T_j
  - TX seen as read-only transaction on database
- **#2 Validation Phase**
  - Check read/write and write/write conflicts, **abort on conflicts**
  - BOCC (Backward-oriented concurrency control) – check all older TXs T_i that finished (EOT) while T_j was running ($EOT(T_i) \geq BOT(T_j)$)
    - **Serializable:** if $EOT(T_i) < BOT(T_j)$ or $WSet(T_i) \cap RSet(T_j) = \emptyset$
    - **Snapshot isolation:** $EOT(T_i) < BOT(T_j)$ or $WSet(T_i) \cap WSet(T_j) = \emptyset$
  - FOCC (Forward-oriented concurrency control) – check running TXs
- **#3 Write Phase**
  - Successful TXs with write operations propagate their local buffer into the database and log

Timestamp Allocation

[Xiangyao Yu, George Bezerra, Andrew Pavlo, Srinivas Devadas, Michael Stonebraker: Staring into the Abyss: An Evaluation of Concurrency Control with One Thousand Cores. **PVLDB 8(3) 2014**]

[Stephen Tu, Wenting Zheng, Eddie Kohler, Barbara Liskov, Samuel Madden: Speedy transactions in multicore in-memory databases. **SOSP 2013**]

- **#1 Mutex**
- **#2 Atomic add / Batched Atomics**
- **#3 Decentralized / CPU Clock**
- **#4 Hardware** (CPU HW counter)

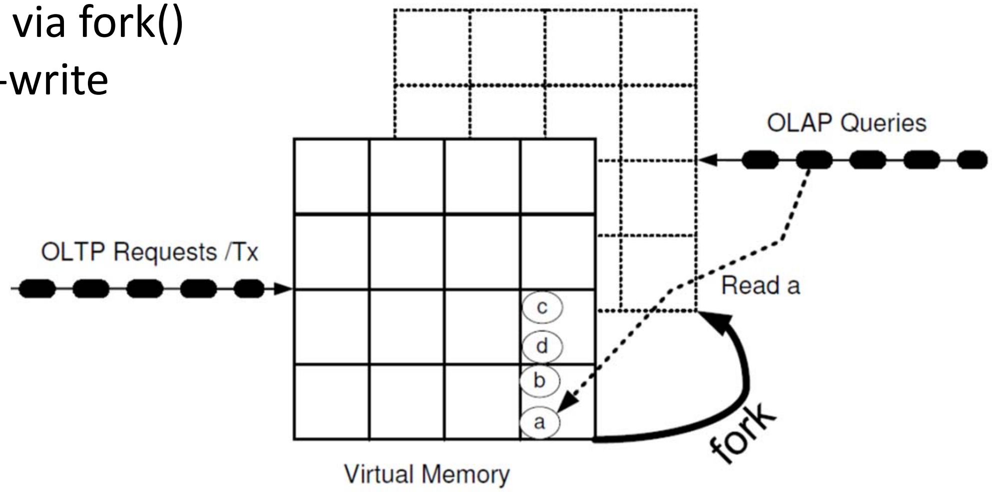


Multi-Version Concurrency Control (MVCC)

Snapshot Isolation w/ Snapshots

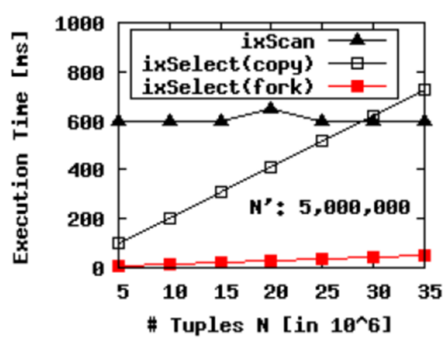
- **#1 Shadow Storage**
- **#2 Snapshots via Fork**
  - Partitioned, single-threaded OLTP ops
  - Snapshots via fork()
    + copy-on-write

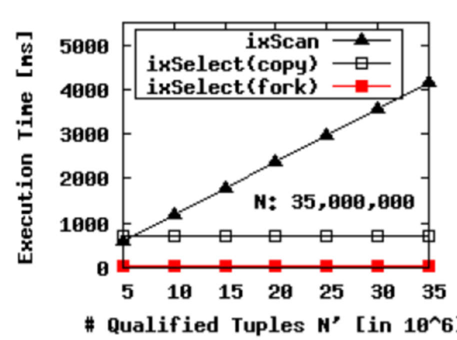
[Alfons Kemper, Thomas Neumann: HyPer: A hybrid OLTP&OLAP main memory database system based on virtual memory snapshots. **ICDE 2011**]



- **Excursus: Query Processing on Prefix Trees (via fork)**

[Matthias Boehm Patrick Lehmann Peter Benjamin Volk Wolfgang Lehner: Query Processing on Prefix Trees, **HPI Future SOC Lab 2011**]



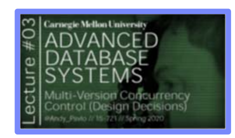


MVCC Overview

- **MVCC Motivation**
  - Read TXs without need for locks, read sets, or copies (fine-grained management of individual versions)
  - Copy-on-write (readers never block writers), garbage collection when safe
  - Additional benefits: time travel, clear semantics, snapshot isolation
  - Mixed HTAP workloads → focus of many recent systems

- **Design Decisions**
  - #1 Concurrency Control Protocol
  - #2 Version Storage
    - Append-only, time-travel, delta
    - Oldest-to-newest/newest-to-oldest
  - #3 Garbage Collection
    - Tuple (background, coop), TX-level
  - #4 Index Management
    - Logical, physical pointers

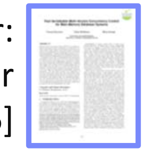
[Andy Pavlo: Advanced Database Systems – Multi-Version Concurrency Control (Design Decisions), **CMU 2020**]



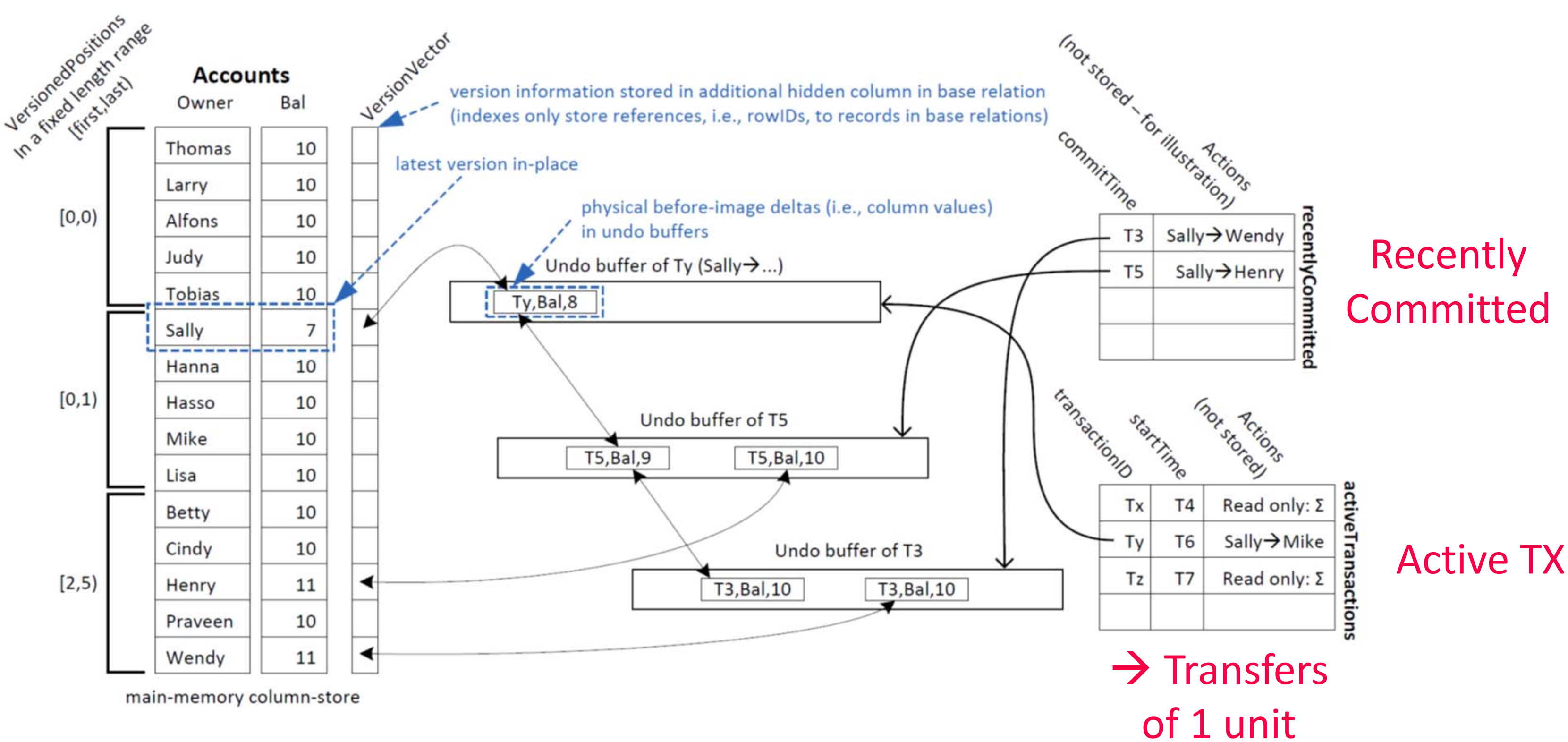
[Yingjun Wu, Joy Arulraj, Jiexi Lin, Ran Xian, Andrew Pavlo: An Empirical Evaluation of In-Memory Multi-Version Concurrency Control. **PVLDB 10(7) 2017**]

Version Storage

[Thomas Neumann, Tobias Mühlbauer, Alfons Kemper: Fast Serializable Multi-Version Concurrency Control for Main-Memory Database Systems. **SIGMOD 2015**]



- **Example Hyper**
  - In-place update, backward delta in UNDO buffer
  - Almost no storage overhead (VersionVector), TX-local commit processing
  - Newest-to-oldest (preference for fast analytical queries)



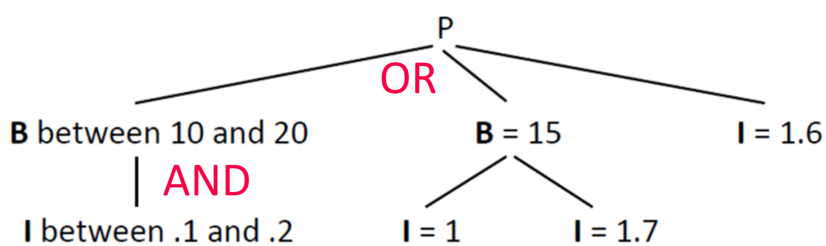
Abort TX write-write conflicts on uncommitted changes

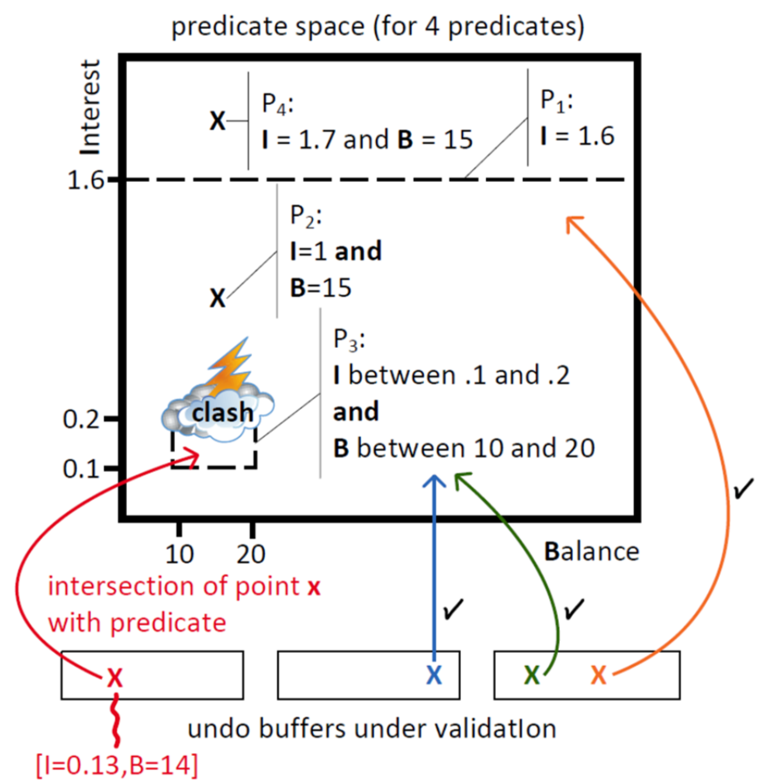
Serializability Validation

[Thomas Neumann, Tobias Mühlbauer, Alfons Kemper: Fast Serializable Multi-Version Concurrency Control for Main-Memory Database Systems. **SIGMOD 2015**]

- **(Extended) Precision Locking**
  - **Predicate logging:** Instead of maintaining read-set, store read predicates of index and table scan of validated T_i in **predicate tree** (PT)
  - Recap: **Serializable:** if $EOT(T_i) < BOT(T_j)$ or $WSet(T_i) \cap RSet(T_j) = \emptyset$
  - Probe UNDO buffers (write set) of all T_j against predicate tree

Predicate Tree of T_i



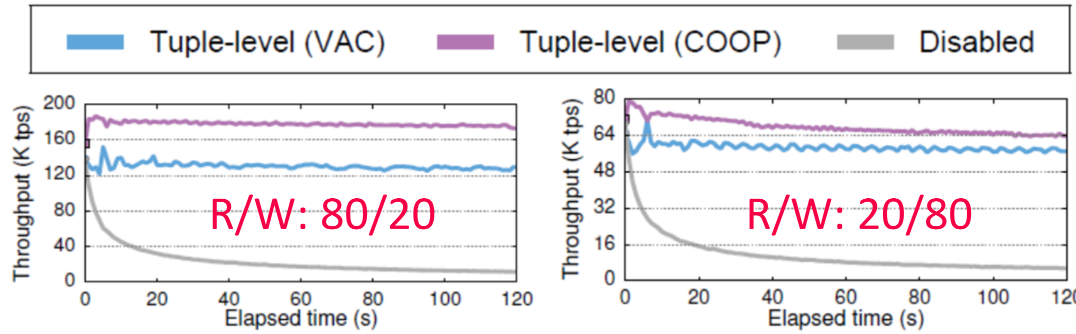


Abort Ti if a single UNDO buffer's data point matches

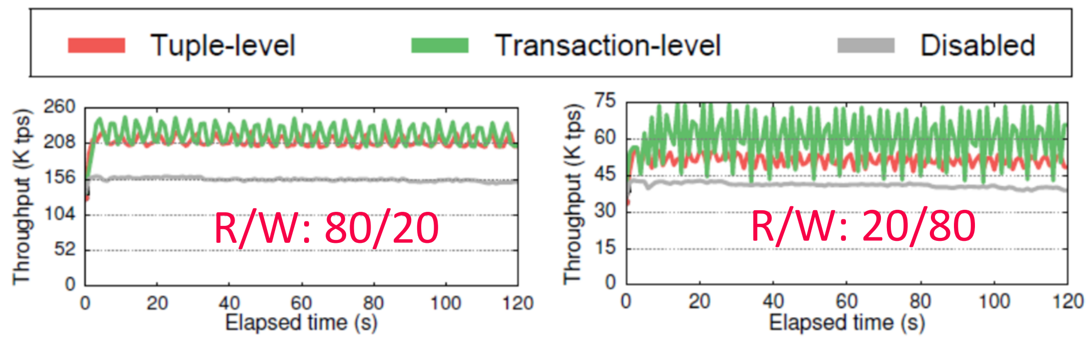
Garbage Collection

[Yingjun Wu, Joy Arulraj, Jiexi Lin, Ran Xian, Andrew Pavlo: An Empirical Evaluation of In-Memory Multi-Version Concurrency Control. **PVLDB 10(7) 2017**]

- **#1 Tuple-level Garbage Collection**
  - Background vacuuming
  - Cooperative cleaning on traversal)



- **#2 Transaction-level**
  - E.g., by epoch



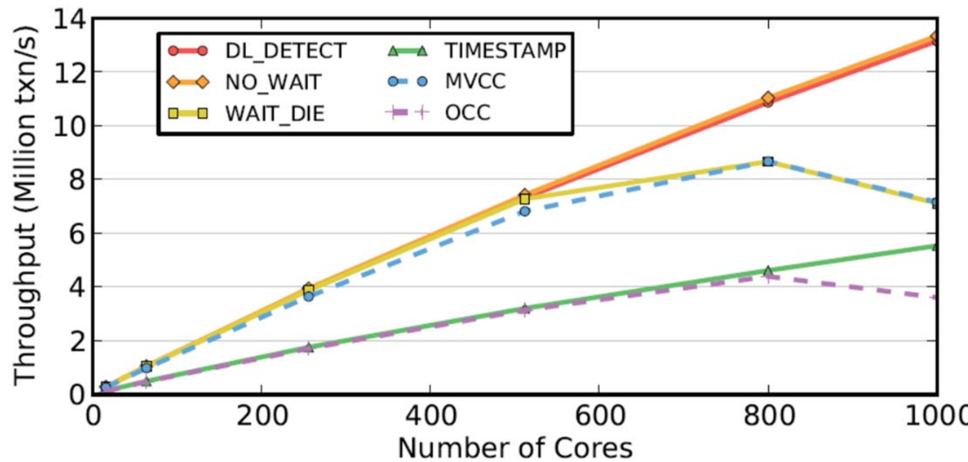
- **Deferred Action Framework (DAF)**
  - Maintenance tasks for GC, plan cache invalidation, data transformation

[Ling Zhang et al: Everything is a Transaction: Unifying Logical Concurrency Control and Physical Data Structure Maintenance in Database Management Systems, **CIDR 2021**]

Comparison (simulated)

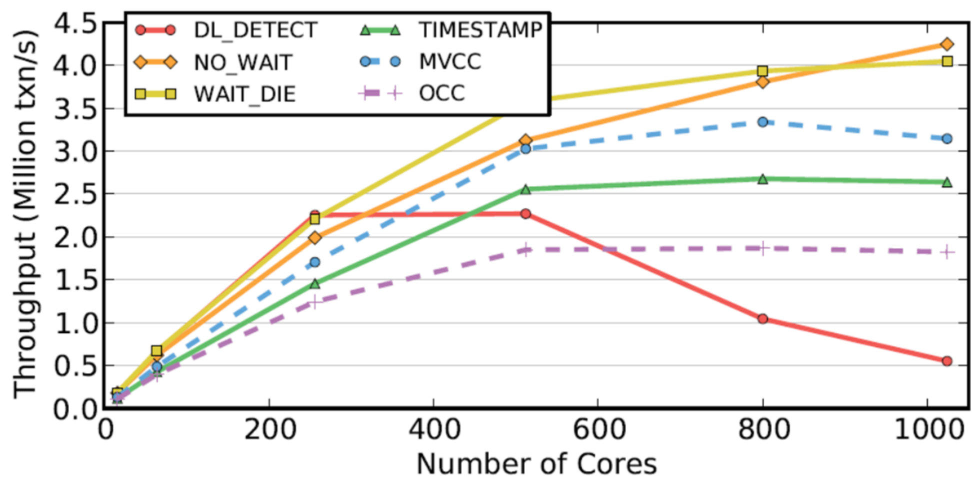
[Xiangyao Yu, George Bezerra, Andrew Pavlo, Srinivas Devadas, Michael Stonebraker: Staring into the Abyss: An Evaluation of Concurrency Control with One Thousand Cores. **PVLDB 8(3) 2014**]

- **Read-only Workload**



Timestamp Allocation

- **Write-intensive Workload** (medium contention)



Abort Rates

Lock Thrashing

Excursus: Coordination Avoidance

Overview Coordination Avoidance

- **Overview**
  - Ensure application-level invariants and convergence instead of (serializability vs weaker) with as little coordination as possible (different approaches)

With Transactions

[Peter Bailis, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica: Bolt-on causal consistency. **SIGMOD 2013**]

[Peter Bailis et al.: Coordination Avoidance in Database Systems. **PVLDB 8(3) 2014**]

[Peter Bailis: Coordination Avoidance in Distributed Databases. **PhD UC Berkeley 2015**]

Without Transactions

[Peter Alvaro, Neil Conway, Joseph M. Hellerstein, William R. Marczak: Consistency Analysis in Bloom: a CALM and Collected Approach. **CIDR 2011**]

[Peter Alvaro: Data-centric Programming for Distributed Systems. **PHD UC Berkeley 2015**]

[Chenggang Wu, Jose M. Faleiro, Yihan Lin, Joseph M. Hellerstein: Anna: A KVS for Any Scale. **ICDE 2018**]

[Chenggang Wu, Vikram Sreekanti, Joseph M. Hellerstein: Autoscaling Tiered Cloud Storage in Anna. **PVLDB 12(6) 2019**]

Summary and Q&A

- **TX Processing Background**
- **Pessimistic and Optimistic Concurrency Control**
- **Multi-Version Concurrency Control**
- **Excursus: Coordination Avoidance**

- **Next Lectures** (Part C)
  - **12 Modern Storage and HW Accelerators** [Jan 27]