Distributed Software
Transactional Memories
Foundations, Algorithms and Tools

Maria Couceiro
(with Paolo Romano and Luís Rodrigues)
IST/ INESC-ID
Contents

- Part I: (Non-Distributed) STMs
- Part II: Distributed STMs
- Part III: Case-studies
- Part IV: Conclusions
Contents

- Part I: (Non-Distributed) STMs
- Part II: Distributed STMs
- Part III: Case-studies
- Part IV: Conclusions
(Non-Distributed) STMs

- Basic Concepts
- Example Algorithms
Basic Concepts

- Concurrent programming has always been a challenge.
- One needs to control the concurrent access to shared data by multiple threads.
- This is hard for most programmers.
- Concurrent programming has been a “niche”.
Basic Concepts

- In the past:
  - More performance via faster CPUs

- Now:
  - More performance via more CPUs
  - Concurrent programming has to become mainstream
Basic Concepts

- Ideally
  - Performance would scale linearly with the number of cores
  - (with 8 cores we would have a program 8 times faster)

- Reality:
  - Speed up limited by % serial code
  - Small % can kill performance (Amdahl’s Law)
  - Say 25% of the program is serial
  - 8 cores = 2.9 speedup.
Basic Concepts

- Ideally
  - Performance would scale linearly with the number of cores
  - (with 8 cores we would have a program 8 times faster)

- Reality:
  - Small % of serial code can kill performance (Amdahl’s Law)
  - Say 25% of the program is serial
  - 32 cores = 3.7 speedup.
Basic Concepts

- Ideally
  - Performance would scale linearly with the number of cores
  - (with 8 cores we would have a program 8 times faster)

- Reality:
  - Small % of serial code can kill performance (Amdahl’s Law)
  - Say 25% of the program is serial
  - 128 cores = 3.9 speedup.
Basic Concepts

- It is hard or impossible to structure a program in a set of parallel independent tasks.
- We need efficient and simple mechanisms to manage concurrency.
Explicit synchronization

One of the most fundamental and simple synchronization primitive is the \texttt{lock}

non-synchronized code;
\texttt{lock ()};
do stuff on shared data;
\texttt{unlock ()};
more non-synchronized code;
Many problems with locks

- Deadlock:
  - locks acquired in “wrong” order.

- Races:
  - due to forgotten locks

- Error recovery tricky:
  - need to restore invariants and release locks in exception handlers
Fine Grained Parallelism?

- Very complex:
  - Need to reason about deadlocks, livelocks, priority inversions.
  - Verification nightmare as bugs may be hard to reproduce.
- Make parallel programming accessible to the masses!!!
Concurrent Programming Without Locks

- Lock-free algorithms.
- Hard to design and prove correct.
- Only for very specialized applications.
- Designed and implemented by top experts.
Abstractions for simplifying concurrent programming...
Atomic transactions

atomic {
  access object 1;
  access object 2;
}

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Transactional Memories

- Hide away synchronization issues from the programmer.
- Advantages:
  - avoid deadlocks, priority inversions, convoying;
  - simpler to reason about, verify, compose.
TMIs: where we are, challenges, trends

- Theoretical Aspects
  - Formalization of adequate consistency guarantees, performance bounds.

- Hardware support
  - Very promising simulation-based results, but no support in commercial processors.
TMs: where we are, challenges, trends

- Software-based implementations (STM)
  - Performance/scalability improving, but overhead still not satisfactory.
- Language integration
  - Advanced supports (parallel nesting, conditional synchronization) are appearing...
  - ...but lack of standard APIs & tools hampers industrial penetration.
TMs: where we are, challenges, trends

- Operating system support
  - Still in its infancy, but badly needed (conflict aware scheduling, transactional I/O).

- Recent trends:
  - Shift towards *distributed* environments to enhance scalability & dependability.
Run-time

- How does it work?
  - The run time implements concurrency control in an automated manner.

- Two main approaches:
  - Pessimistic concurrency control (locking).
  - Optimistic concurrency control.

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Example of pessimistic concurrency control

- Each item has a read/write lock.
- When an object is read, get the read lock.
  - Block if write lock is taken.
- When an object is written, get the write lock.
  - Block if read or write lock is taken.
- Upon commit/abort:
  - Release all locks.
Example of optimistic concurrency control

- Each item has a version number.
- Read items and store read version.
- Write local copy of items.
- Upon commit do atomically:
  - If all read items still have the read version (no other concurrent transaction updated the items)
    - then apply all writes (increasing the version number of written items).
  - Else,
    - abort.
Many, many, variants exist

- For instance, assume that two phase locking is used and a deadlock is detected. It is possible:
  - Abort both transactions.
  - Abort the oldest transaction.
  - Abort the newest transaction.
  - Abort the transaction that did less work.
Many, many variants exist

- For instance, assume that two phase locking is used and a deadlock is detected. It is possible:
  - Abort both transactions
  - Abort the oldest transaction
  - Abort the newest transaction
  - Abort the transaction that did less work

Each alternative offers different performance with different workloads.
How to choose?

- What is a correct behavior?
- Which safety properties should be preserved?
- Which liveness properties should be preserved?
How to choose?

- What is a correct behavior?
- Which safety properties should be preserved?
- Which liveness properties should be preserved?

To answer these questions we need a bit of theory.
Theoretical Foundations

- Safety:
  - What schedules are acceptable by an STM?
  - Is classic atomicity property appropriate?

- Liveness:
  - What progress guarantees can we expect from an STM?
Theoretical Foundations

Safety:
- What schedules are acceptable by an STM?
- Is classic atomicity property appropriate?

Liveness:
- What progress guarantees can we expect from an STM?
Classic atomicity property

- A transaction is a sequence of read/write operations on variables:
  - sequence unknown a priori (otherwise called static transactions).
  - asynchronous (we do not know a priori how long it takes to execute each operation).
- Every operation is expected to complete.
- Every transaction is expected to abort or commit.
Histories

- The execution of a set of transactions on a set of objects is modeled by a **history**

- A history is a total order of operation, commit and abort events
Histories

Two transactions are **sequential** (in a history) if one invokes its first operation after the other one commits or aborts; they are **concurrent** otherwise.
Two transactions are **sequential** (in a history) if one invokes its first operation after the other one commits or aborts; they are **concurrent** otherwise.

Non-sequential:
Two transactions are **sequential** (in a history) if one invokes its first operation after the other one commits or aborts; they are **concurrent** otherwise.

**Sequential:**

```
OP  C  OP  OP  OP  C
```
Histories

- A history is sequential if it has only sequential transactions; it is concurrent otherwise
Histories

- A history is sequential if it has only sequential transactions; it is concurrent otherwise.

- Sequential:

```
OP  C  OP  OP  C  OP  C
```
Histories

- A history is sequential if it has only sequential transactions; it is concurrent otherwise.

- Non-sequential:
Histories

- Two histories are **equivalent** if they have the same transactions.
Histories

Two histories are equivalent if they have the same transactions.

Equivalent:

\[ \text{OP} \quad \text{C} \quad \text{OP} \quad \text{OP} \quad \text{C} \quad \text{OP} \quad \text{C} \]

\[ \text{OP} \quad \text{C} \quad \text{OP} \quad \text{OP} \quad \text{OP} \quad \text{C} \quad \text{C} \]
Histories

- Two histories are **equivalent** if they have the same transactions.
- **Non-equivalent:**

```
OP  C  OP  OP  C  OP  C
OP  C  OP  OP  OP  C  C
```
What the programmer wants?

- Programmer does not want to be concerned about concurrency issues.
- Execute transactions “as if” they were serial
- No need to be “serially executed” as long as results are the same
A history $H$ of committed transactions is **serializable** if there is a history $S(H)$ that is:

- equivalent to $H$
- sequential
- every read returns the last value written
Serializable?

WO2(1) → RO2(1) → RO1(0) → C → WO1(1) → C

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
 Serializable!
Serializable?

WO2(1) → RO2(0) → RO1(0) → C → WO1(1) → C
Serializability

Non-serializable!

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Opacity

 Serializable (blue aborts)?

WO2(1) → RO2(1) → RO1(0) → A → WO1(1) → C
Opacity

- Serializable: only committed transactions matter!
In a database environment, transactions run SQL:
- no harm if inconsistent values are read as long as the transaction aborts.

This is not the same in a general programming language:
- observing inconsistent values may crash or hang an otherwise correct program!
Opacity: example

Initially: \( x:=1; \ y:=2 \)

- \( T1: x := x+1; \ y := y+1 \)
- \( T2: z:= 1 / (y-x); \)

If \( T1 \) and \( T2 \) are atomic, the program is correct.
Opacity: example

Initially: \( x := 1; \ y := 2 \)

- T1: \( x := x + 1; \ y := y + 1 \)
- T2: \( z := 1 / (y - x) \);

Otherwise...
Opacity: example

Initially: \( x := 1; \ y := 2 \)

- T1: \( x := x + 1; \ y := y + 1 \)
- T2: \( z := 1 / (y - x) \);

Otherwise...
Opacity: example

Initially: x:=1; y:=2

- T1: x := x+1; y := y+1
- T2: z:= 1 / (2-x);

Otherwise...
Opacity: example

Initially: \( x := 1; \ y := 2 \)

- T1: \( x := x + 1; \ y := y + 1 \)

- T2: \( z := 1 / (2 - x) \)

Otherwise...
Opacity: example

Initially: $x:=1; \ y:=2$

After $T1$: $x:=2; \ y:=3$

- $T1: x := x+1; \ y := y+1$

- $T2: z:= 1 / (2-x)$;

Otherwise...
Opacity: example

Initially: \( x := 1; \ y := 2 \)

After T1: \( x := 2; \ y := 3 \)

- T1: \( x := x + 1; \ y := y + 1 \)
- T2: \( z := \frac{1}{2-2} \);

Otherwise... divide by zero!
Intuitive definition:

- every operation sees a consistent state
  (even if the transaction ends up aborting)
Opacity
[GK08]

- Intuitive definition:
  - every operation sees a consistent state
    (even if the transaction ends up aborting)

- Following history is serializable but violates opacity!
Does classic optimistic concurrency control guarantee opacity?

- Writes are buffered to private workspace and applied atomically at commit time.
- Reads are optimistic and the transaction is validated at commit time.
- Opacity is not guaranteed!
Theoretical Foundations

- **Safety:**
  - What schedules are acceptable by an STM?
  - Is classic atomicity property appropriate?

- **Liveness:**
  - What progress guarantees can we expect from an STM?
Progress

- STMs can abort transactions or block operations...
- But we want to avoid implementations that abort all transactions!
- We want operations to return and transactions to commit!
Requirements

- **Correct transactions:**
  - *commit* is invoked after a finite number of operations
  - either *commit* or perform an infinite number of (low-level) steps

- **Well-formed histories:**
  - every transaction that aborts is immediately repeated until it commits
Conditional progress: obstruction freedom

- A correct transaction that eventually does not encounter contention eventually commits

- ...but what to do upon contention?
Contention-managers

- Abort is unavoidable
- But want to maximize the number of commits
- Obstruction freedom property: progress and correctness are addressed by different modules.

Contention-managers encapsulate policies for dealing with contention scenarios.
Contention-managers

Let $TA$ be executing and $TB$ a new transaction that arrives and creates a conflict with $TA$.
CM: Aggressive

Let TA be executing and TB a new transaction that arrives and creates a conflict with TA.

- **Aggressive contention manager:**
  - always aborts TA
CM: Backoff

Let TA be executing and TB a new transaction that arrives and creates a conflict with TA.

- **Backoff contention manager:**
  - TB waits an exponential backoff time
  - If conflict persists, abort TA
CM: Karma

Let TA be executing and TB a new transaction that arrives and creates a conflict with TA.

Karma contention manager:
- Assign priority to TA and TB
  - Priority proportional to work already performed
- Let Ba be how many times TB has been aborted
- Abort TA if $Ba > (TA - TB)$
CM: Greedy

Let TA be executing and TB a new transaction that arrives and creates a conflict with TA.

- **Greedy contention manager:**
  - Assign priority to TA and TB based on start time
  - If TB<TA and TA not blocked then wait
  - Otherwise abort TA
(Non-Distributed) STMs

- Basic Concepts
- Example Algorithms
  - DSTM
  - JVSTM
(Non-Distributed) STMs

- Basic Concepts
- Example Algorithms
  - DSTM
  - JVSTM
DSTM

- Software transactional memory for dynamic-sized data structures.
- Prior designs: static transactions.
- DSTM: dynamic creation of transactional objects.

DSTM

- Killer write:
  - Ownership.

- Careful read:
  - Validation.
DSTM - Writes

- To write \( T \) requires a write-lock on \( o \).
- \( T \) aborts \( T' \) if some \( T' \) acquired a write-lock on \( o \):
  - Locks implemented via Compare & Swap.
- Contention manager can be used to reduce aborts.
DSTM – Reads and Validation

- Concurrent reads do not conflict.

- To read $o$, $T$ checks if all objects read remain valid;
  - else abort $T$.

- Before committing, $T$ checks if all objects read remain valid and releases all its locks.
  - Make sure that the transaction observes a consistent state.
  - If the validation fails, transaction is restarted.
DSTM - Why is careful read needed?

- No lock is acquired upon a read:
  - invisible reads
  - visible read invalidate cache lines
  - bad performance with read-dominate workloads due to high bus contention

- What if we validated only at commit time?
  Serializability? Opacity?

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
No lock is acquired upon a read:
- invisible reads
- visible read invalidate cache lines
- bad performance with read-dominate workloads due to high bus contention

What if we validated only at commit time?

Serializability? Y
Opacity? N
(Non-Distributed) STMs

- Basic Concepts
- Example Algorithms
  - DSTM
  - JVSTM
JVSTM

- Java Versioned Software Transactional Memory.
- Cachopo and Rito-Silva. 2006.
- Versioned boxes as the basis for memory transactions.
JVSTM

- Optimized for read-only transactions:
  - Never aborted or blocked;
  - No overhead associated with readset tracking.

- How?
  - Multi-version concurrency control.
  - Local writes (no locking, optimistic approach)
  - Commit phase in global mutual exclusion.
  - Recently introduced a parallel commit version [FC09].
  - Global version number (GVN)
JVSTM - Versioned boxes

- Versioned boxes
- Each transactional location uses a versioned box to hold the history of values for that location.
JVSTM - Algorithm

Upon begin $T$, read GVN and assigned it to $T$ snapshot ID (sID).

Upon read on object $o$:
- If $o$ is in $T$’s writeset, return last value written,
- else return the version of the data item whose sID is “the largest sID to be smaller than the $T$’ sID”.
- If $T$ is not read-only, add $o$ to readset.
 Upon write, just add to the writeset.
  - No early conflict detection.

 Upon commit:
  - Validate readset:
    - Abort if any object read has changed.
  - Acquire new sID (atomic increase of GVN).
  - Apply writeset: add new version in each written VBox.
JVSTM - Execution

T1 2

T2 2 3

T3 2 5

T4 3 4

Time
Contents

- Part I: (Non-Distributed) STMs
- **Part II: Distributed STMs**
- Part III: Case-studies
- Part IV: Conclusions
Distributed STMs

- Origins
- Goals
- Distribution Strategies
- Programming Models
- Toolbox
Origins

- Convergence of two main areas:
  - Distributed Shared Memory
  - Database Replication
Distributed Shared Memory

- DSM aims at providing a single system image
  - Fault-tolerance via checkpointing

- Strong consistency performs poorly
  - Myriad of weak-consistency models
  - Programming more complex

- Explicit synchronization
  - Locks, barriers, etc
DSTMs vs DSM

- DSTMs are simpler to program
- Transactions introduce boundaries where synchronization is required
- By avoiding to keep memory consistency at every (page) access or at the level of fine-grain locks, it may be possible to achieve more efficient implementations

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Database Replication

- Databases use transactions
  - Constrained programming model
  - Durability is typically a must

- Database replication was considered too slow

- In the last 10 years new database replication schemes have emerged
  - Based on atomic broadcast and on a single coordination phase at the beginning/end of the transaction.
DSTMs vs DBMS

- Transactions are often much shorter in the STM world
  - This makes coordination comparatively more costly
- Durability is often not an issue
  - This makes coordination comparatively more costly
- Database replication techniques can be used as a source of inspiration to build fault-tolerant DSTMs
Distributed STMs

- Origins
- Goals
- Distribution Strategies
- Programming Models
- Toolbox
Goals

- Better performance:
  - Doing reads in parallel on different nodes.
  - Computing writes in parallel on different items.

- Fault-tolerance:
  - Replication the memory state so that it survives the failure of a subset of nodes.
Distributed STMs

- Origins
- Goals
- Distribution Strategies
- Potential Problems
- Toolbox

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Distribution Strategies

- Single System Image
  - Distribution is hidden
  - Easier when full replication is implemented
  - No control of the data locality

- Partitioned Global Address Space
  - Different nodes have different data
  - Distribution is visible to the programmer
  - Programmer has fine control of data locality
  - Complex programming model
Distribution Strategies

- Partitioned non-replicated
  - Max capacity
  - No fault-tolerance
  - No load balancing for reads on multiple nodes

- Full replication
  - No extra capacity
  - Max fault-tolerance
  - Max potential load balancing for reads
Distributed STMs

- Origins
- Goals
- Distribution Strategies
- Programming Models
- Toolbox
Dataflow Model

- Transactions are immobile and objects move through the network.
- Write: processor locates the object and acquires ownership.
- Read: processor locates the object and acquires a read-only copy.
- Avoids distributed coordination.
- Locating objects can be very expensive.
Control Flow Model

- Data is statically assigned to a home node and does not change over time.

- Manipulating objects:
  - In the node (via RPC);
  - First data is copied from the node then the are changes written back.

- Relies on fast data location mechanism.

- Static data placement may lead to poor data locality.
Distributed STMs

- Origins
- Goals
- Distribution Strategies
- Programming Models
- Toolbox
Toolbox

- Atomic Commitment
- Uniform Reliable Broadcast (URB)
- Atomic Broadcast (AB)
- Replication Strategies
Atomic Commitment

- Atomicity: all nodes either commit or abort the entire transaction.

- Set of nodes, each node has input:
  - CanCommit
  - MustAbort

- All nodes output same value
  - Commit
  - Abort

- Commit is only output if all nodes CanCommit

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
2-phase commit

- prepare msg
- validate/ acquire locks
- vote msg (Yes or No)
- decision msg (Commit or Abort)
- apply decision
2PC is blocking

Coordinator

Participant

Participant

prepare msg

vote msg (Yes or No)

decision msg (Commit or Abort)

validate/ acquire locks

validate/ acquire locks

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
3PC

Pre-decision msg (Pre-Commit)

Decision msg (Commit)

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Toolbox

- Atomic Commitment
- **Uniform Reliable Broadcast (URB)**
- Atomic Broadcast (AB)
- Replication Strategies
Uniform Reliable Broadcast

- Allows to broadcast a message \( m \) to all replicas
- If a node delivers \( m \), every correct node will deliver \( m \)
- Useful to propagate updates

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Toolbox

- Atomic Commitment
- Uniform Reliable Broadcast (URB)
- Atomic Broadcast (AB)
- Replication Strategies
Atomic Broadcast

- Reliable broadcast with total order
- If replica R1 receives \( m1 \) before \( m2 \), any other correct replica Ri also receives \( m1 \) before \( m2 \)
- Can be used to allow different nodes to obtain locks in the same order.

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Sequencer-based ABcast

R1: sequencer
Sends message
Assigns SN
Receive Msg + order
Receive Msg + order
Commit order
Commit order
Commit order
final uniform order

R2

R3
Abcast with optimistic delivery

- Total order with optimistic delivery.
- Unless the sequencer node crashes, final uniform total order is the same as regular total order.
- Application may start certificating the transaction locally based on optimistic total order delivery.
ABcast with optimistic delivery

R1: sequencer
Assigns SN
Sends message
R2
Receive Msg + order
Receive Msg + order
R3
Commit order
Commit order
Commit order
Optimistic delivery
Final delivery

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
ABcast with optimistic delivery

R1: sequencer
Assigns SN
Sends message
Spontaneous order delivery

R2
Spontaneous order
Receive SN
Commit order

R3
Spontaneous order
Receive SN
Commit order

final uniform order

Final delivery

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Toolbox

- Atomic Commitment
- Uniform Reliable Broadcast (URB)
- Atomic Broadcast (AB)
- Replication Strategies
In absence of replication, there’s no chance to fall into deadlocks with a single lock... what if we add replication?
Replicating a single lock

T1
- lock()
- Update R1
- Waiting for R2

T2
- lock()
- Update R2
- Waiting for R1
Replicating a single lock

T1
lock() Updates R1 unlock()

T2
lock() Updates R1

T1
lock() Updates R2 unlock()

T2
lock() Updates R1

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Coordination is slow

- Drawback of previous approach:
  - Coordination among replicas needs to be executed at every lock operation.
  - Atomic broadcast is an expensive primitive.
  - The system becomes too slow.

- Solution:
  - Limit the coordination among replicas to a single phase, at the beginning of the transaction or commit time.
Single-phase schemes

- State machine replication
- Single master (primary-backup)
- Multiple master (certification)
  - Non-voting
  - Voting
State-machine replication

- All replicas execute the same set of transactions, in the same order.
- Transactions are shipped to all replicas using atomic broadcast.
- Replicas receive transactions in the same order.
- Replicas execute transaction by that order.
  - Transactions need to be deterministic!
State-machine replication

R1
AB of T1’s input params
T1 pre-acquires its locks
T1 execs
T1 commits
T2 is blocked due to T1
T2 execs
T2 commits

R2
AB of T2’s input params
T1 pre-acquires its locks
T1 execs
T1 commits
T2 is blocked due to T1
T2 execs
T2 commits
Single-phase schemes

- State machine replication
- Single master (primary-backup)
- Multiple master (certification)
  - Non-voting
  - Voting
Primary-backup

- Write transactions are executed entirely in a single replica (the primary).
- If the transaction aborts, no coordination is required.
- If the transaction is ready to commit, coordination is required to update all the other replicas (backups).
  - Reliable broadcast primitive.
- Read transactions may be executed on backup replicas.
  - Works fine for workloads with very few update transactions.
  - Otherwise the primary becomes a bottleneck.
Primary-backup

- **Synchronous updates:**
  - Updates are propagated during the commit phase:
    - Data is replicated immediately
    - Read transactions observe up to date data in backup replicas
    - Commit must wait for reliable broadcast to finish

- **Asynchronous updates:**
  - The propagation of updates happens in the background:
    - Multiple updates may be batched
    - Commit is faster
    - There is a window where a single failure may cause data to be lost
    - Read transactions may read stale data

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Single-phase schemes

- State machine replication
- Single master (primary-backup)
- Multiple master (certification)
  - Non-voting
  - Voting
Multi-master

- A transaction is executed entirely in a single replica.
- Different transactions may be executed on different replicas.
- If the transaction aborts, no coordination is required.
- If the transaction is ready to commit, coordination is required:
  - To ensure serializability
  - To propagate the updates
Multi-master

- Two transactions may update concurrently the same data in different replicas.
- Coordination must detect this situation and abort at least one of the transactions.
- Two main alternatives:
  - Non-voting algorithm
  - Voting algorithm
Single-phase schemes

- State machine replication
- Single master (primary-backup)
- Multiple master (certification)
  - Non-voting
  - Voting
Non-voting

- The transaction executes locally.
- When the transaction is ready to commit, the read and write set are sent to all replicas using atomic broadcast.
- Transactions are certified in total order.
- A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set).
Non-voting

Execution Transaction T1

Execution Transaction T2

AB of T1’s read & writeset

AB of T2’s read & writeset

Validation & Commit T1

Validation & Commit T1

Validation & Commit T1

Validation & Abort T2

Validation & Abort T2

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Single-phase schemes

- State machine replication
- Single master (primary-backup)
- Multiple master (certification)
  - Non-voting
  - Voting
Voting

- The transaction executes locally at replica R
- When the transaction is ready to commit, **only the write set** is sent to all replicas using atomic broadcast
- Transactions’ commit requests are processed in total order
- A transaction may commit if its read set is still valid (i.e., no other transaction has updated the read set):
  - **Only R can certify the transaction!**
- R send the outcome of the transaction to all replicas:
  - Reliable broadcast
Voting

Transaction T1

T1’s AB
(write set)

R1

T1’s validation

wait for R1’s vote

T1’s RB
(vote)

commit

R2

commit
Contents

- Part I: (Non-Distributed) STMs
- Part II: Distributed STMs
- Part III: Case-studies
- Part IV: Conclusions
Part III: Case-Studies

- Partitioned Non-Replicated
  - STM for clusters (Cluster-STM)

- Partitioned (Replicated)
  - Static Transactions (Sinfonia)

- Replicated Non-Partitioned
  - Certification-based with Bloom Filters (D²STM)
  - Certification with Leases (ALC)
  - Active Replication with Speculation (AGGRO)
Part III: Case-Studies

- Partitioned Non-Replicated
  - STM for clusters (Cluster-STM)

- Partitioned (Replicated)
  - Static Transactions (Sinfonia)

- Replicated Non-Partitioned
  - Certification-based with Bloom Filters (D²STM)
  - Certification with Leases (ALC)
  - Active Replication with Speculation (AGGRO)
Cluster-STM

- Software Transactional Memory for Large Scale Clusters
- Bocchino, Adve, and Chamberlain. 2008
- Partitioned (word-based) address space
- No persistency, no replication, no caching
- Supports only single thread per node
- Various lock acquisition schemes + 2PC
Cluster-STM

- Various methods for dealing with partitioned space
- Data movement (Dataflow model):
  - \texttt{stm get}(src proc, dest, work proc, src, size, open)
  - \texttt{stm put}(src proc, work proc, dest, src, size, open)
- Remote execution (Control flow model):
  - \texttt{stm on}(src proc, work proc, function, arg buf, arg buf size, result buf, result buf size)

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Cluster-STM

```c
increment(proc_t proc, int *addr) {
    atomic {
        on(proc) {
            ++*addr
        }
    }
}
```
Cluster-STM

\[
\text{increment(proc_t proc, int* addr)} \{ \\
\quad \text{stm_start(MY_ID)} \\
\quad \text{stm_on(MY_ID, proc, increment_local, addr, sizeof(int*), 0, 0)} \\
\quad \text{stm_commit(MY_ID)} \\
\}
\]

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Cluster-STM

```
increment_local(proc_t src_proc,
    void* arg,
    size_t arg_size,
    void *result,
    size_t result_size) {
    int *addr = *((int*) arg);
    int tmp;
    stm_open_read(src_proc, addr, sizeof(int))
    stm_read(src_proc, &tmp, addr, sizeof(int))
    ++tmp;
    stm_open_write(src_proc, addr, sizeof(int))
    stm_write(src_proc, addr, &tmp, sizeof(int))
}
```
Cluster-STM

- Read locks (RL) vs. read validation (RV)

  - **RL:**
    - immediately acquire a lock as a read (local or remote) is issued
    - abort upon contention (avoid deadlock)
    - as coordinator ends transaction, it can be committed w/o 2PC

- Note: distributed model w/o caching:
  - each access to non local data implies remote access:
  - eager locking is for free
  - with caching only RV could be employable

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Cluster-STM

- Read locks (RL) vs. read validation (RV)

- RV:
  - commit time validation (not opaque)
  - validity check requires 2PC
Cluster-STM

- Write buffering schemes

- UL undo log:
  - write is applied and an undo log is maintained
  - forced sync upon each write

- WB write buffering:
  - writes applied in local buffer
  - avoid communications for writes during exec phase
  - requires additional communication at commit time
Cluster-STM

- Write buffering: two lock acquisition schemes
  - LA: Late acquire
    - at commit time.
    - may allow for more concurrency
  - EA: Early acquire
    - as the write is issued
    - may avoid wasted work by doomed transactions
Cluster-STM: Graph Analysis (SSCA2)

![Bar chart showing the ratio of locks to STM for different numbers of processors. The chart compares various STM implementations: RL-EA-UL, RV-EA-UL, RL-EA-WB, and RL-LA-WB.](image)

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Cluster-STM: Graph Analysis (SSCA2)
Part III: Case-Studies

- Partitioned Non-Replicated
  - STM for clusters (Cluster-STM)

- Partitioned (Replicated)
  - Static Transactions (Sinfonia)

- Replicated Non-Partitioned
  - Certification-based with Bloom Filters (D²STM)
  - Certification with Leases (ALC)
  - Active Replication with Speculation (AGGRO)
Sinfonia

- Sinfonia: A new paradigm for building scalable distributed systems.
- Partitioned global (linear) address space
- Optimized for static transactions
Sinfonia

- Mini-transactions:
  - A-priori knowledge on the data to be accessed

- Two types of nodes:
  - Application nodes
  - Memory nodes

- Fault-tolerance via:
  - In-memory replication
  - Sync (log) + async checkpoint for persistency on memory nodes
Sinfonia

In a nutshell, Sinfonia is a service that allows hosts to share application data in a fault-tolerant, scalable, and consistent manner. Our approach targets particular data center applications, such as metadata, tables, and configuration and status information. Protocols for handling distributed state include complex protocols for handling distributed state. Distributed state refers to data that application hosts need to manipulate and share. Existing services that allow hosts to share data include database systems, which all tend to increase coupling. Thus, application hosts must be fault-tolerant and scalable, and must provide functionality than needed, resulting in performance overheads. For instance, attempts to build file systems using a database system resulted in an unusable system due to poor performance. Existing DSM systems lack the scalability or fault tolerance required for infrastructure applications. Section 8 discusses some of the DSM functionalities that applications use to atomically access and conditionally modify that applications use to atomically access and conditionally modify.

Sinfonia seeks to provide a balance between functionality and performance overheads. For instance, attempts to build file systems using a database system resulted in an unusable system due to poor performance. Existing DSM systems lack the scalability or fault tolerance required for infrastructure applications. Section 8 discusses some of the DSM functionalities that applications use to atomically access and conditionally modify that applications use to atomically access and conditionally modify.
Sinfonia

Minitransactions have compare items, read items, and write items. A minitransaction with one compare and one write item on different memory nodes.

Minitransaction

- **compare items**
  - mem-id addr len data
  - mem-id addr len data
  - ...

- **read items**
  - mem-id addr len
  - mem-id addr len
  - ...

- **write items**
  - mem-id addr len data
  - mem-id addr len data
  - ...

In Sinfonia, coordinators are application nodes and participants are memory nodes.

In the first phase, the coordinator asks all participants if they are ready to commit. If they all vote yes, in the second phase the coordinator commits. If the transaction's last action does not affect the coordinator's decision, then we can also piggyback the action.

A frequent minitransaction idiom is to use compare items to validate data, without read or write items. Such a minitransaction validates data, without read or write items.

In Section 4 we explain how minitransactions are executed and controlled caching has three clear advantages:

1. There is greater support for applications to do their own caching.
2. Managing caches in Sinfonia is always current (not stale).
3. Sinfonia becomes a simpler service to use because data accessed know their data better than what Sinfonia can infer.

Another minitransaction idiom is to have only compare items to modify data indicated by a set of compare-and-swap operations. A compare item compares the current value against a constant; if equal, a write item replaces it.

A compare item checks that a lease is held and, if so, write items update data.

A read item returns the old value and a write item replaces it.

A compare item replaces it.

Create multiple increment items, possibly at different memory nodes, to increment data and, if data is valid, use write items to apply some changes.

If an increment item atomically changes data, the increment is guaranteed to be performed in the commit phase of the two-phase commit protocol. The increment is then committed.
Sinfonia

**API**

class Minitransaction {
    public:
        void cmp(memid, addr, len, data); // add cmp item
        void read(memid, addr, len, buf); // add read item
        void write(memid, addr, len, data); // add write item
        int exec_and_commit(); // execute and commit
};

**Example**

...  
t = new Minitransaction;
t->cmp(memid, addr, len, data);
t->write(memid, addr, len, newdata);
status = t->exec_and_commit();
...
Sinfonia

- Global space is partitioned
  - Transaction may need to access different memory nodes
  - It can only commit if it can commit at all memory nodes
  - 2-phase commit
4.1 Basic architecture

Phase 1, the coordinator (application node) generates a new minitransaction, while phase 2 commits it. More precisely, in the two-phase protocol in Figure 4. Phase 1 executes and prepares to block anyway. Furthermore, Sinfonia can optionally replicate by all its replicas. We accomplish this by blocking on participant crashes instead of logging to improve performance.

Our two-phase commit protocol also reflects new system fail-compromises: a participant tries to acquire locks without blocking; if it finds that locks are busy, it discards the write items. The coordinator never logs any information, unlike in some predefined order, but with that scheme, the coordinator in phase 1 has to contact participants in series (to ensure lock order).

Our minitransaction protocol integrates execution of the minitransaction with memory nodes. Each participant then (a) tries to lock the addresses of all compare items succeeded, the vote is for committing, otherwise it discards them. In all compare items failed, the vote is for aborting.

In the compare phase, the coordinator sends the compare items to the participants. Each participant then (a) compares the data of its items in the minitransaction (without blocking), (b) executes the compare items and, if all comparisons succeed, (c) votes for committing, otherwise it discards them. In all compare items failed, the vote is for aborting.

In the compare phase, the coordinator sends the compare items to the participants. Each participant then (a) compares the data of its items in the minitransaction (without blocking), (b) executes the compare items and, if all comparisons succeed, (c) votes for committing, otherwise it discards them. In all compare items failed, the vote is for aborting.

In the compare phase, the coordinator sends the compare items to the participants. Each participant then (a) compares the data of its items in the minitransaction (without blocking), (b) executes the compare items and, if all comparisons succeed, (c) votes for committing, otherwise it discards them. In all compare items failed, the vote is for aborting.

In the compare phase, the coordinator sends the compare items to the participants. Each participant then (a) compares the data of its items in the minitransaction (without blocking), (b) executes the compare items and, if all comparisons succeed, (c) votes for committing, otherwise it discards them. In all compare items failed, the vote is for aborting.

In the compare phase, the coordinator sends the compare items to the participants. Each participant then (a) compares the data of its items in the minitransaction (without blocking), (b) executes the compare items and, if all comparisons succeed, (c) votes for committing, otherwise it discards them. In all compare items failed, the vote is for aborting.
Sinfonia

- No support for caching:
  - delegated to application level
  - same applies for load balancing

- Replication:
  - aimed at fault-tolerance, not enhancing performance
  - fixed number of replicas per memory node
  - primary-backup scheme ran within first phase of 2PC
Part III: Case-Studies

- Partitioned Non-Replicated
  - STM for clusters (Cluster-STM)
- Partitioned (Replicated)
  - Static Transactions (Sinfonia)
- Replicated Non-Partitioned
  - Certification-based with Bloom Filters (D²STM)
  - Certification with Leases (ALC)
  - Active Replication with Speculation (AGGRO)
D$^2$STM

- D$^2$STM: Dependable Distributed STM
- Couceiro, Romano, Rodrigues, Carvalho, 2009
- Single-image system
  - Full replication
  - Strong consistency
- Certification-based replication scheme
  - Based on Atomic Broadcast
  - Built on top of JVSTM
**D\textsuperscript{2}STM**

- Non-voting replication scheme
- Transactions execute in a single replica
- No communication during the execution
- Writeset **and** readset AB at commit time
- Deterministic certification executed in total order by all replicas
- No distributed deadlocks

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
D²STM

Execution of Transaction T1

AB of T1’s read & writeset

Validation & Commit T1

Validation & Commit T1
D$^2$STM

AB of both T1’s readset & writeset

Problem:
(very) big message size

R1 Execution

R2

Validation & Commit

Validation & Commit

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
In STMs, transaction’s execution time is often 10-100 times short than in DBs:
- the cost of AB is correspondingly amplified

Bloom Filter Certification:
- space-efficient encoding (via Bloom Filter) to reduce message size
**D\textsuperscript{2}STM**

- Application
- D2STM API
- JVSTM
- Replication Manager
- Generic Group Communication Service
- Network

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
**Bloom filters**

- A set of \( n \) items is encoded through a vector of \( m \) bits
- Each item is associated with \( k \) bits through \( k \) hash functions having as image \( \{1..m\} \):
  - insert: set \( k \) bits to 1
  - query: check if all \( k \) bits set to 1
D\(^2\)STM

- **False Positives:**
  - An item is wrongly identified as belonging to a given set
  - Depend on the number of bits used per item \((m/n)\) and the number of hash functions \((k)\)

- **D2STM** computes the size of the Bloom filter based on:
  - User-defined false positive rate
  - Number of items in the read set (known)
  - Number of BF queries, estimated via moving average over recently committed transactions

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
D$^2$STM

- Read-only transactions:
  - local execution and commit
Write transaction T:
- Local validation (read set)
- If the transaction is not locally aborted, the read set is encoded in a Bloom filter
- Atomic broadcast of a message containing:
  - the Bloom filter encoding of tx readset
  - the tx write set
  - the snapshot ID of the tx
- Upon message delivery: validate tx using Bloom filter’s information
D²STM

for each committed $T'$ s.t. $T'.\text{snapshotID} > T.\text{snapshotID}$
for each data item $d$ in the writeset of $T'$
if $d$ is in Bloom filter associated with $T'$s readset
abort $T$

// otherwise...

commit $T$
D²STM

STM Bench 7: Results

![Graph showing % Execution Time Reduction of Write Transactions]
Part III: Case-Studies

- Partitioned Non-Replicated
  - STM for clusters (Cluster-STM)

- Partitioned (Replicated)
  - Static Transactions (Sinfonia)

- Replicated Non-Partitioned
  - Certification-based with Bloom Filters ($D^2$STM)
  - Certification with Leases (ALC)
  - Active Replication with Speculation (AGGRO)
Asynchronous Lease Certification Replication of Software Transactional Memory

Carvalho, Romano, Rodrigues, 2010

Exploit data access locality by letting replicas dynamically establish ownership of memory regions:

- replace AB with faster coordination primitives:
- no need to establish serialization order among non-conflicting transactions
- shelter transactions from remote conflicts
Data ownership established by acquiring an Asynchronous Lease

- mutual exclusion abstraction, as in classic leases...
- ...but detached from the notion of time:
- implementable in a partially synchronous system
- Lease requests disseminated via AB to avoid distributed deadlocks.
Transactions are locally processed

At commit time check for leases:
- An Asynchronous Lease may need to be established

Proceed with local validation
If local validation succeeds, its writeset is propagated using Uniform Reliable Broadcast (URB):
- No ordering guarantee, 30-60% faster than AB

If validation fails, upon re-execution the node holds the lease:
- Transaction cannot be aborted due to a remote conflict!
ALC

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
ALC

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
ALC
ALC

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
ALC

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
If applications exhibit some access locality:
- avoid, or reduce frequency of AB
- locality improved via conflict-aware load balancing

Ensure transactions are aborted at most once due to remote conflicts:
- essential to ensure liveness of long running transactions
- benefic at high contention rate even with small running transactions
ALC

Application

Distributed STM API Wrapper

JVSTM

Lease Manager

Replication Manager

Group Communication Service
ALC

- Synthetic “Best case” scenario
- Replicas accessing distinct memory regions
ALC

- Synthetic “Worst case” scenario
- All replicas accessing the same memory region
ALC

Lee Benchmark
Part III: Case-Studies

- Partitioned Non-Replicated
  - STM for clusters (Cluster-STM)

- Partitioned (Replicated)
  - Static Transactions (Sinfonia)

- Replicated Non-Partitioned
  - Certification-based with Bloom Filters ($D^2$STM)
  - Certification with Leases (ALC)
  - Active Replication with Speculation (AGGRO)

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
AGGRO

AGGRO: Boosting STM Replication via Aggressively Optimistic Transaction Processing

R. Palmieri, Paolo Romano and F. Quaglia, 2010

Active Replication for STMs

- Multiple replicas
- All replicas execute update transactions
- Read-only transactions can execute in any replica
- Data survives failures of replicas
Basic Active Replication
With Optimistic Delivery
Improve improvement

Atomic Broadcast

Opt

Tx Exec

Hold Locks

C

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
But...

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Speculative

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Transactions are started in speculative order immediately after the optimistic delivery.

- Writes kill all transactions that have read stale data.
- Items touched by speculative transactions Tspec are marked as “work-in-progress (WIP)” while Tspec executes.
  - When Tspec terminates (but not yet committed) items are unmarked as WIP.
- Transaction only read values from terminated transactions.
AGGRO Algorithm

upon $\text{opt-Deliver}(T_i)$

start transaction $T_i$ in a speculative fashion
AGGRO Algorithm

upon write(Ti, X, v)

if (X not already in Ti.WS)

add X to Ti.WS

mark X as WIP // C&S

for each Tj that follows Ti in OAB order:

if (Tj read X from Tk preceding Ti) abort Tj

else

update X in Ti.WS
AGGRO Algorithm

upon read(Ti, X)

if (X in Ti.WS) return X.value from Ti.WS

if (X in Ti.RS) return X.value from Ti.RS

wait while (X is marked WIP)

let Tj be tx preceding Ti in OAB order that wrote X

Ti.readFrom.add(Tj)
AGGRO Algorithm

upon completed (Ti)
   atomically {
      for each X in Ti.WS: unmark X as WIP by Ti
   }

upon commit(Ti)
   atomically {
      for each X in Ti.WS: mark X as committed
   }

M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
AGGRO Algorithm

upon abort(Ti)

abort any transaction that read from Ti

restart Ti

upon TO-Deliver(Ti)

append Ti to TO-order

wait until all xacts preceding Ti in TO-order committed

if (validation of Ti’s readset fails) abort (Ti)

else commit(Ti)
AGGRO Performance

Performance speed-up
(20% reordering, only one SO explored)

![Graph showing performance speed-up](image)
Contents

- Part I: (Non-Distributed) STMs
- Part II: Distributed STMs
- Part III: Case-studies
- Part IV: Conclusions
Conclusions

- Replication helps in read-dominanted workloads or when writes have low conflicts
- Replication provides fault-tolerance
- Some techniques have promising results
Conclusions

- No technique outperforms the others for all workloads, networks, number of machines, etc
- Autonomic management of the distributed consistency and replication protocols
  - Change the protocols in runtime, in face of changing workloads
A bit of publicity

- Time for the commercials
DTMs: a programming paradigm for the Cloud?

Stay tuned on www.cloudtm.eu
Euro-TM
Euro-TM Cost Action

- Research network bringing together leading European experts in the area of TMs
- Contact us if you are interested in joining it:
  - romano@inesc-id.pt
  - ler@inesc-id.pt
- www.euromt.org
Bibliography


Bibliography


Bibliography


[FC10] Sérgio Fernandes and João Cachopo, A scalable and efficient commit algorithm for the JVSTM, 5th ACM SIGPLAN Workshop on Transactional Computing

Bibliography

[GHP05] Rachid Guerraoui, Maurice Herlihy, Bastian Pochon: Toward a theory of transactional contention managers. PODC 2005: 258-264

[GU09] Rachid Guerraoui, Tutorial on transactional memory, CAV 2009


M. Couceiro, P. Romano, L. Rodrigues, HPCS 2011
Bibliography


Bibliography

[KAJLKW08] Christos Kotselidis, Mohammad Ansari, Kim Jarvis, Mikel Luján, Chris Kirkham and Ian Watson, DiSTM: A Software Transactional Memory Framework for Clusters, In the 37th International Conference on Parallel Processing (ICPP'08), September 2008


Bibliography


The End

Thank you.