Real-time distributed concurrency control algorithms with mixed time constraints

Citation for published version (APA):

Document status and date:
Published: 01/01/1996

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Real-Time Distributed Concurrency Control Algorithms
with mixed time constraints

by

P.D.V. van der Stok

ISSN 0926-4515

All rights reserved
editors: prof.dr. R.C. Backhouse
        prof.dr. J.C.M. Baeten

Reports are available at:
http://www.win.tue.nl/win/cs

Computing Science Report 96/18
Eindhoven, September 1996
Real-Time Distributed Concurrency Control Algorithms with mixed time constraints

P.D.V. van der Stok
Eindhoven University of Technology
Department of Computing Science
P.O. Box 513
5600 MB Eindhoven, Netherlands
August 7, 1996

Abstract

Two types of applications are considered: Hard Real-Time (HRT) and Soft Real-Time (SRT). HRT applications need to meet their deadlines under all circumstances; deadlines of SRT applications may occasionally be missed. HRT and SRT applications coexist on the same set of processors and need to exchange data in a consistent manner.

Several concurrency control algorithms exist to assure the consistency of the results of transactions in distributed database systems. Three types of concurrency control are generally recognized: locking, timestamp ordering and optimistic concurrency control. They all impose an acyclic order on the individual actions of the transactions. The addition of versions to the data can diminish the execution time of read-only transactions. Imposing two different orderings, one for read-only transactions and one for all other transactions, the wait-time of read-only transactions is considerably reduced. The specified RO-Read allows the exchange of data between HRT and SRT applications without perturbations to the HRT applications.

1 Introduction

Real-Time application software reacts to stimuli from its environment. The state of the application is determined by the environment, while the state of the environment in its turn is determined by the state of the application. The application software (controlling process) often interacts quite heavily with a physical process (process under control).

The results of Real-Time applications not only need to be functionally correct, they must also be delivered in time. A result that is delivered too late is as bad as, or even worse than, a result that is not delivered at all. The timeliness aspects of the Real-Time applications make Real-Time systems notoriously complex. In this paper, an application is divided into two parts: A Hard Real-Time (HRT) and a Soft Real-Time (SRT) part (application). Both applications have deadlines which have to be met. However, the missing of the deadline of a HRT application has catastrophic consequences for the process under control; the missing of a SRT application only constitutes a temporary degradation of the system performance. This division has been made to provide the system designer with a means to put some structure in its application to increase the number of designs that meet a given specification.
The object-oriented programming paradigm is followed for the development of the applications. An application consists of a program which is constructed from classes. A class defines a data-structure and the actions which can be executed on this structure. An instance of a class is called an object. Objects are distributed over a set of processors. Objects do not migrate between processors. The interleaving of actions to the same object may lead to inconsistent results. When during two concurrent accesses to the object, an item \( a \) with value \( A \) is read followed by the storage of \( A + 1 \) into \( a \), a possible final result \( A + 1 \) differs from the intended result \( A + 2 \). To prevent such unintended outcome, objects can be declared atomic. This means that the results of the actions on an object are the same as if those actions were executed at one unique atomic moment of time. Inconsistent results caused by the interleaving of the actions are then excluded. The atomicity concept allows the system designer to concentrate on the functionality of his actions and not to bother about the intricate interactions between the different concurrent executions of these actions. The concept of atomic objects is motivated by the well-known concept of Conversation Schemes [AL81] used for the construction of reliable Real-Time systems. The duality of atomic objects and conversation schemes has been shown in [SMB85]. Concurrency Control Algorithms (CCA) impose an order on the concurrent actions on the objects with the purpose to meet the atomicity requirements of the objects. The research area of Real-Time (RT)-databases is lively one [Gra92, HSRT91]. The objective is to specify serialization criteria or CCA's to create transactions which meet their deadlines. The here considered objects are only stored in memory. Recovery of objects from stable storage is a different topic treated separately. This restriction is the major difference with RT-databases.

Transactions in RT-databases can be classified in three categories [Ram93]:

- Write-Only (WO)-transactions which read data from the environment and store them (in the database),
- Update transactions which read stored data and derive new data to be stored,
- Read-Only (RO)-transactions which read stored data and send them to the environment.

The addition of versions [Wei87] to objects increases the number of possible interleavings thus diminishing the probability that transactions need to be restarted and increasing the probability that they meet their deadlines. In this paper an algorithm based on versions is presented which diminishes the time that (RO)-transactions wait for other transactions. Consequently, the execution time of RO-transactions decreases.

Typically a large part of the WO- and RO-transactions, which interact directly with the environment need to be HRT-transactions. The other transactions can for a larger part be SRT-transactions. Data between HRT- and SRT-transactions is exchanged via objects. It is shown that the requirement of non-perturbation of HRT-transactions by SRT transaction considerably reduces the number of possible interleavings of SRT- and HRT-transactions. When the HRT-transactions are designed such that the order in which the results of HRT-transaction are delivered is more important than the serialization order imposed by the CCA used for the HRT-transactions, the concept of transfer serializability can be introduced. Transfer serializability is a correct criterium iff the transfer of object-modifications by a given HRT(SRT)-transaction to an equivalent set of objects readable by SRT(HRT)-transactions does not lead to unwanted system behaviour.

Within the actions of an object, invocations to actions of other objects can be specified. Such nested actions can also be declared atomic. However, the actions invoked by the atomic
action should in their turn also be atomic. These actions are called nested atomic actions similar to the concept of nested transactions [Mos87].

In section 2 a historical view on database CCA's is presented, followed by a informal description of transaction properties in section 3. The first presented algorithm is an extension to Optimistic Concurrency Control with Time Intervals (OCC-TI) [LS93]. OCC-TI is a forward validation variation [HSRT91] on Optimistic Concurrency Control (OCC) [KR81]. The here presented Versioned OCC-TI (OCC-VTI) algorithm is an extension of OCC-TI with versions and modified for execution on a distributed computer platform. A specification and proof of OCC-VTI is presented in sections 5 and 6. The extension to HRT/SRT communication is made in section 7. The concept of nested transactions is discussed in section 8. The same additions as discussed for OCC-VTI are added to several CCA's developed for nested transactions, followed by their proof of correctness in section 9. The rest of the paper discusses implementation aspects of the different actions. A measure is calculated to decide the number of versions which are necessary to increase the performance of the RO-transactions.

2 Database historical

Actions on a database consist of a series of read or write actions on individual database elements. These actions are grouped in transactions, which can be executed in an interleaved fashion. An execution of a set of transactions is a serial execution iff for any pair of transactions $T_i$ and $T_j$ all actions of $T_i(T_j)$ are terminated before any action of $T_j(T_i)$ starts. Correct transactions executed in isolation, take the database from a given consistent state to a new consistent state. Consequently, starting from a consistent state, a serial execution of the transactions always leaves the database in a consistent state. In accordance, transactions executed in an interleaved fashion should only see values belonging to one and the same consistent database state and produce corresponding database values. The concept of serializability states that the result, delivered by a set of interleaved transactions, should be the same as a given serial execution of these transactions, called the equivalent serial execution. A lot of research has been done to specify the serializability criteria for databases. Simultaneously, research has been done on the algorithms which assure that the implementation of the transactions satisfies these serializability criteria [BHGS7, PapS6].

Apart from serializability, another important aspect of databases is the permanency of the results even when failures occur such as processor or disk crashes [BHGS7, GMA87]. The requirement that the database is always in a consistent state has as consequence that all actions of a transaction should be executed or none.

The concept of nested transactions was introduced to increase the parallelism inside a transaction and render recovery more efficient [Mos85, Mos87, BBG89]. The individual read and write actions no longer needed to be executed in a fixed serial order, but could be grouped and be executed in parallel. When a subtransaction of a transaction fails, only the subtransaction needs to be recovered.

The first concurrency control algorithms were based on locking. When two transactions simultaneously need to act on the same data-item and one of these actions is a write-action, the actions are said to conflict. When an action of one transaction conflicts with an action of another transaction, the transactions are said to be conflicting. When all conflicting actions are write-actions, the transactions are said to be WW-conflicting. When one of the conflicting actions is a read-action, the transactions are said to be RW-conflicting. The locking algorithm
should assure that for any two conflicting transactions the conflicting actions of the one are executed before those of the other. The algorithms based on this principle are known to realize conflict serializability [Pap86]. The concept of two phase locking (2PL) assures this even in the case of processor or disk failures. 2PL states that a transaction first acquires locks in a growing phase and acquires none once it has released a lock. The main problem is the possibility of dead-lock. Notoriously difficult dead-lock detection and repair algorithms have been specified [Kna87].

Deadlock prevention assigns to each transaction a unique priority, usually defined by its start time. When a conflicting action is recognized by a transaction, lower priority transactions wait for higher priority transactions. When a lower priority transaction arrives first, only one of the two transactions is allowed to continue and the other aborts and restarts the execution of the transaction from scratch. As the priorities are totally ordered, no cycles can occur in the wait-for graph and deadlock is prevented. Deadlock prevention leads to transaction aborts which are not always necessary and to continuously aborted transactions (starvation).

The timestamp ordering (TSO)-approach completely abandons the lock concept. It is assumed that events in the distributed systems are totally ordered by logical clocks or synchronized clocks. Every transaction is assigned a timestamp that consists of the local time of the processor on which the transaction started suffixed with the processor identifier. Assuming that the processor-identifiers are totally ordered, also the timestamps are totally ordered. The CCA imposes the order prescribed by the timestamps on the equivalent serial order of the transactions. The algorithm aborts transactions or puts them in a wait state such that read actions only read values written by transactions with a lower timestamp and no values are written to data-items which have been read by transactions with higher timestamps than the timestamp of the writing transaction [Ree83]. Every data-item is typically equipped with a write-timestamp, that is the timestamp of the writing transaction, and a read-timestamp, that is the maximum of the timestamps of the transactions that have read this particular value of the item.

The drawback of this method is that frequently updated values make it difficult for slow transactions to read a value corresponding with its usually too low timestamp. The solution to this problem is the addition of versions, such that also slow Read-Only (RO)-transactions can proceed independent of the frequent updates of the values by other transactions. This approach is called Multi Version TSO (MVTSO) [Wei87]. Each data-item consists of a set of versions and each version consists of a read-timestamp and a write-timestamp as specified for the TSO algorithm.

For many databases, especially the large distributed ones, the number of conflicts will be relatively low. A more performant database system can then be created by assigning local copies of the database items to the transactions and let them execute their actions on these local copies. At the end of the transaction a serial order of the transactions is calculated and the validity of the values in the copies is verified against actions executed by other transactions on the same data-items. When an equivalent serial execution exists, the local copies are installed in the database. In case of invalid results, the copies are destroyed and the transaction is restarted. This approach is called Optimistic Concurrency Control (OCC) [KR81, Sch81]. At validation time, a timestamp is allocated to the transaction as explained for TSO and MVTSO. The algorithm assures that the order of an equivalent serial execution of the transactions corresponds with the order imposed by these timestamps. Two approaches exist: Backward Validation (BV) and Forward Validation (FV). In the BV variant,
the validating transaction only commits when all commits in the past allow this, otherwise
the validating transactions aborts. In the FV variant, the validating transaction aborts still
executing transactions which will not be allowed to commit in the future. More recent OCC-
FV algorithms determine on the basis of additional criteria (e.g. meeting of deadlines) whether
the executing or the validating transaction should be aborted.

Most commercial database systems use the locking approach. The performance gain of
the database system does not seem to outweigh the disadvantages of a new unchartered
domain of concurrency control algorithms. However, Real-Time databases are gaining impor-
tance and the performance characteristics of the algorithms described above are investigated
[Gra92, HSRT91]. The performance issue is relatively important in Real-Time systems. Typi-
cally, values will be updated with a relatively high frequency determined by the Real-Time
environment. Relatively slow reads of the data-items follow a different time pattern. The
introduction of versions is ideal to decouple the write frequencies from the read frequencies.
However, in the inverse case, a frequent reading of the data at moments specified by the Real-
Time environment can be seriously hampered by slow updates of these items. To solve this
problem, an optimization of the RO-transactions supported by multiple versions is proposed
in this paper. The performance improvement can be realized by specifying a different order
for RO-transactions with respect to the other Writing(W)-transactions. RO-transactions will
only read those versions of the data-items which have been validated by the writing transac-
tions. This is enforced by allocating timestamps to RO-transactions equal to their start-time,
and allocating timestamps to the W-transactions equal to their commit time plus an eventual
constant.

3 Transactions

Concurrent actions on the atomic objects of a system meet the following two requirements:

- Concurrency atomicity: Actions behave as if they are executed in one unique point in
time.

- Exception atomicity: All modifications of an action are visible to other actions or none.

The more frequently used term failure atomicity implies exception atomicity and moreover
that all modifications of the database remain available in spite of processor and memory
failures (permanency of results [BH87, GMA87]). In this paper, permanency of results is
not considered.

The concept of concurrency atomicity is equivalent with the isolation criterion used for
transactions in databases. The serializability concept which characterizes transactions which
conform to the isolation criterion applies equally well to the concurrency atomicity concept.
Therefore, many results obtained by the research into RT-databases can be directly applied
to atomic objects.

A database consists of a set of database items $D$. A database is consistent iff a database
dependent predicate $P$ holds. Atomic Read and Write actions can be executed on the individu-
al items. Items are composed of versions which reflect the history of values atributed to
these items. A transaction, $T_i$, is composed of a set of partially ordered actions. An execution
of a set of transactions is called a serial execution iff for any two transactions $T_i$ and $T_j$ all
actions of $T_i(T_j)$ are finished before any action of $T_j(T_i)$ starts.
When \( T_i \) executes a write action on item \( X \) it creates a version \( z_i \) notated with \( W_i[z_i] \). When \( T_i \) reads an item, it reads the value of a version written by some transaction \( T_j \), notated like: \( R_i[x_j] \). Each element from \( I \) is read or written once by a given transaction. All read actions are executed before the write actions. This often justified simplification [Pap86] renders the algorithms and their accompanying proofs much simpler. An example of the execution of transaction \( T_i \) is \( T_i = R_i[x_j]R_i[y_k]W_i[z_i]W_i[z_i] \). The values of \( j \) and \( k \) depend on the history of the data-items. A transaction \( T_i \) has a read set \( RSi \) and a write set \( WSi \) defined by: \( RSi = \{ x_j \mid x_j \text{ is a version of } X \in D \text{ read by } T_i \} \) and \( WSi = \{ x_i \mid x_i \text{ is a version of } X \in D \text{ written by } T_i \} \). A transaction only composed of read(write) actions is called a RO(WO)-transaction. A transaction with one or more write-actions is called a \( W \)-transaction. Two transactions which have versions \( x_i, x_j \) of the same data-item \( X \) as argument in one of their actions are called related. Two related transactions of which at least one contains a write-action to a common data-item are called conflicting.

\[
T_1 \quad \bullet \quad \cdots \quad \cdots \\
T_2 \quad \bullet \quad \cdots \quad \cdots \\
T_3 \quad \bullet \quad \cdots \quad \cdots \\
\text{RO-transactions} \\
\text{W-transactions} \\
T_4 \quad \times \quad \cdots \quad \bullet \quad \cdots \\
T_5 \quad \times \quad \cdots \quad \bullet \quad \cdots \\
T_6 \quad \times \quad \cdots \quad \bullet \quad \cdots \\
T_7 \quad \times \quad \cdots \quad \bullet \quad \cdots \\
\]

Figure 1: transaction times and ordering

In Fig. 1, transactions \( T_i, 0 < i \leq 7 \), are shown which take a certain time as defined by the lines. A typical transaction is separated in a read-write phase denoted by the solid line and a commit or validation phase denoted with the broken lines. The transactions governed by the here presented algorithms behave with respect to RO-transactions as if they are executed in the timepoints denoted with large dots. \( W \)-transactions should behave with respect to \( W \)-transactions as if they are executed in the timepoints denoted with the crosses. The positions of the crosses and dots depend on the CCA. In current CCA's, the crosses and the dots coincide. With the presented RO variant of OCC-TI, the crosses are placed anywhere on the solid line. When a RO-transaction starts, the data for which no waits are needed.
are the data of transactions that are finished. Therefore, the dots of the W-transactions are preferably placed at the end of the broken lines and the dots of the RO-transactions are placed at the beginning of the solid lines. In fig. 1, the RO serial execution, determined by the dots, is \([T_1, T_4, T_5, T_6, T_2, T_7]\) and the W serial execution, determined by the crosses, is \([T_4, T_5, T_7, T_6]\). RO-transaction \(T_1\) will read the results of none of the W-transactions \(T_4 - T_7\) as the ordering prescribes that \(T_1\) happens before \(T_4 - T_7\). However, RO-transaction \(T_2\) may read results of \(T_4, T_5\) and \(T_6\).

Problems may arise in the case of transactions \(T_6\) and \(T_7\). According to the crosses, \(T_6\) happens before \(T_7\), but according to the dots \(T_7\) happens before \(T_6\). Consequently, the values of the database-items as seen by the RO-transactions can be different from the ones seen by the W-transactions. For example assume that the transactions \(T_6\) and \(T_7\) act on three items \([X,Y,Z]\) with consistency constraint \(Z = X + Y\). Suppose that \(T_6\) executes \(Y := Y + 2; Z := Z + 2\); and \(T_7\) executes \(X := X + 5; Z := Z + 5\). From the initial contents \([2,3,5]\), the contents of \([X,Y,Z]\) will either pass via \([2,5,7]\) or via \([7,3,10]\) to \([7,5,12]\). The two different orderings then impose that a given transaction can read the values \([7,3,10]\), while another can read \([2,5,7]\). However, only one of the two possibilities is physically executed. Therefore, the set of predecessors of a RO-transaction is determined in the following way. All W-transactions with dots placed after the dots of the RO-transactions succeed the RO-transaction. All W-transactions which have crosses placed before the crosses of all transactions belonging to the successor set are members of the predecessor set. In the example the dot of \(T_7\) is placed after the dot of \(T_2\) and \(T_7\) is part of the successor set of \(T_2\). Transactions \(T_4\) and \(T_5\) are in the predecessor set of transaction \(T_2\) because their dots are placed before the dot of \(T_2\) and their crosses are placed before the cross of \(T_7\). \(T_6\) does not belong to the predecessor set because its cross is placed after the cross of \(T_7\).

4 Informal description of OCC-VTI algorithms

The following rules apply for OCC-VTI. A transaction \(T_i\) precedes transactions \(T_j\) if \(T_i.C_f < T_j.C_f\). The following precedence relations are assured between a validating and X-writing transaction \(T_v\) and the currently X-writing transaction \(T_w\) and X-reading transaction \(T_r\); \(T_v\) precedes \(T_r\) precedes \(T_w\) [LS93]. No precedence relations are maintained between a X-reading \(T_v\) and \(T_r\) [LS93]. The precedence relation between a X-reading \(T_v\) and X-writing \(T_w\) is: \(T_v\) precedes \(T_w\).

The algorithm of Fig. 2 determines which transactions read the values last written by a preceding transaction. A validity interval \([C_f, C_i]\) is attributed to the transactions and a validity interval \([C_b, C_e]\) is attributed to the versions of all items \(X \in D\). The relation between the \([C_b, C_e]\) and the \([C_f, C_i]\) interval is the following. For a given version \(x_i\) written by \(T_i\): \(x_i.C_b = T_i.C_f\). The value of \(x_i.C_e\) is given by: \(x_i.C_e = \max\{T_j.C_f \mid T_j \text{ has read } x_i\}\). For a given version \(x_i\) holds that: \(x_i.C_b < x_i.C_e\). For two consecutively committed versions \(x_i\) and \(x_{i+1}\) holds: \(x_{i+1}.C_b > x_i.C_e\).

The transaction is activated with the specification, \(RS\) and \(WS\), of the actions it must execute and the data-items, \(I\), on which they are executed. After the creation of the transaction identifier, \(T_a\), the validity interval \([C_f, C_i]\) is initialised to \([0, \infty]\) and all read actions specified in \(RS\) are performed. The value read from each item \(X\) is stored in \(A[X]\). Then the transaction can calculate results and store these into \(A[X]\). The modified values are returned with the write actions specified in \(WS\). At the end of the transaction Validate is executed in which the final \(C_f\) and \(C_i\) values are calculated and the end-time-stamp \(T_a\).
PROCEDURE W-Transaction(RS: read items, WS: write items, I: data-items)
VAR A[data-items]: values
BEGIN
Create time-stamp Ts = time o pid
Create interval \([C_f, C_i] = [0, \infty]\)
\(\forall X \in RS: A\text{-read}(X, T_s, A[X])\)
execute calculations
\(\forall X \in WS: A\text{-write}(X, T_s, A[X])\)
Create time-stamp Te = 0
Validate(I, Ts, Te, G)
IF \((C_f < C_i)\) THEN \(\forall X \in I: A\text{-Commit}(X, T_s, T_e, C_f)\)
ELSE \(\forall X \in I: A\text{-Abort}(X, T_s)\)
END

PROCEDURE Validate(I: data-items, Ts, Te: Timestamp, C_f, C_i: Counter)
VAR C_b, C_e: Counter
BEGIN
\(C_f := 0; C_i := \infty\)
\(\forall X \in I\ DO\)
\(A\text{-Validate}(X, T_s, t, C_b, C_e)\)
\(C_f := \max(C_b, C_f); C_i := \min(C_i, C_e)\)
\(T_e = \max(T_s, t)\)
OD
END

Figure 2: W-transaction Algorithm

During the validation phase, a validating transaction, \(T_v\), determines its own validity interval by taking the intersection of its validity interval \([C_f, C_i]\) with the validity intervals \([C_b, C_e]\) as returned by the A-Validate invocations.

When the Validate has succeeded, i.e. a non-empty interval has been found, \(T_v\) commits and the versions written by \(T_v\) are installed with the modifications to the \(C_b\) and \(C_e\) values of other versions. The time-stamp \(T_e\) attributed to each installed version for later use by the RO-transactions, is set equal to \(T_e\) that represents the maximum of the values returned by the A-Validate invocations.

The composing actions A-Validate, A-Read, A-Write, A-Abort and A-Commit are explained in more detail in section 5.

PROCEDURE RO-Transaction(I: data-items, t: time-stamp)
VAR A[data-items]: values
BEGIN
Create time-stamp Ts = time o PID
IF \(\exists X \in I: (\exists x \in X: T_s > T_x) \lor \exists S \neq \text{Commit}\) THEN
\(C_m := (\min X \in I: (\min x \in X, x \in S \neq \text{Commit} \lor x.T_s > T_x \lor x.C_b))\)
ELSE \(C_m := (\min X \in I: (\max x \in X : x.C_b))\)
\(\forall X \in I: RO\text{-read}(X, C_m, A[X])\)
execute calculations
END

Figure 3: RO-Transaction Algorithm
In Fig. 3, the algorithm for a RO-transaction is shown. At the start the identification time-stamp $T_s$ is created. For all $X \in I$ with uncommitted versions or versions with a commit time-stamps larger than $T_s$, the minimum $C_b$ value is determined. When no such versions exist the maximal $C_b$ value is taken. This value is used to find the corresponding version as explained in section 3.

This distributed version of OCC-TI has a disadvantage with respect to the uniprocessor one discussed in [LS93]. In the uniprocessor case, validating transactions will succeed because the intervals of still executing transactions are adjusted possibly leading to their immediate abort. This is no longer possible because global distributed knowledge on all relevant $[C_b, C_e]$ intervals becomes available after the Validation phase.

5 Specification of OCC-VTI and RO-Read Actions

A given data-item $X$ consists of a set of versions $X.V$. Each version $x_i$ is a tuple:

$$< C_b, C_r, C_e, T_e, T_s, S, list, value >$$

where (1) $C_b$ is the begin counter of the validity interval, (2) $C_r$ the maximum $C_f$ of the transactions that read $X$, (3) $C_e$ is the end counter of the validity interval, (4) $T_e$ is the time-stamp of the commit action (5) $T_s$ is the time-stamp that identifies the writer of the version, (6) $S$ is the state of the version, (7) $list$ is a list of identifiers of transactions that read this version and (8) $value$ is the value of the item's version. The state $S$ of a version is (1) tentative: the version is created but not yet usable by other transactions than the creating one, (2) validating: the transaction that created this version has finished all write and read actions but has not yet asserted their validity and (3) commit: the version can be used by transactions different from the creating one.

In Fig. 4, the specifications for the reading, writing, validating, and committing of one data-item are shown.

**A-Read** returns the value of the last committed version, $v$, and adds the identifier, $T_i$, of the reading transaction to $v.list$.

**A-Write** adds a new version, $v$, to the set of versions, $X.V$. The specified value is stored in $v.value$ and $v.T_s$ is set to the identifier of the writing transaction.

**A-Validate** is executed when no other transaction is validating the specified item, $X$. A-Validate calculates the authorised $C_f, C_l$ value pairs for $X$. When the version, $v$, has been read, $C_l$ is set equal to $v.C_r - 1$, otherwise $C_l$ is set to $\infty$. When version $v$ has been written, $C_f$ is set equal to $b.C_r + 1$ where $b$ is the last committed version of $X$. When no version has been written, a version $v$ must have been read and $C_f$ is set equal to $v.C_b + 1$ with $v$ the read version. A local end timestamp is calculated and returned in $t$.

**A-Commit** sets the state of the written version, $v$, equal to Commit and $v.C_b$ equal to $C_f$. When a version $v$ is written, $b.C_e$ is set to $v.C_f$ where $b$ is the last committed version. When a version, $v$, is read, $v.C_r$ is set equal to the maximum of $C_f$ and the former value of $v.C_r$.

**A-Abort** removes all versions, $x$, owned by the aborting transaction, $x.T_o = T_i$.

The specification of the **RO-Read** states that a committed version with the largest $C_b$ value is selected that is smaller than the value, $C$, determined by the reading transaction (see Fig. 3).
ACTION A-Read( X: data-item, Ti: Time-stamp, var a: value)
POST: v.list = L ∪ {Ti} ∧ a = A
}

ACTION A-Write( X: data-item, Ti: Time-stamp, a: value)
{PRE: X.V = Q ∧ a = A
POST: v = newly created version ∧ v.S = tentative ∧ v.TC = Ti ∧ v.value = A ∧ v.list = ∅ ∧ X.V = Q ∪ {v}
}

ACTION A-Validate( X: data-item, Ti: Time-stamp, var Cf, Ci: Counter)
{PRE: ∀x ∈ X.V : x.S ≠ validating
POST: ∃(∀x ∈ X.V ∧ x.TC = Ti : x.S = validating) ∧ (∃u ∈ X.V : u.TC = Ti) ⇒ Cf = (max x ∈ X.V ∧ x.S = commit : x.Cr + 1)) ∧ (∀x ∈ X.V ∧ x.TC = Ti) ⇒ Ci = (max x ∈ X.V ∧ x.TC = Ti : x.Cr + 1)) ∧ (∃u ∈ X.V ∧ u.TC = Ti ∧ u.list) ⇒ Ci = Gc - 1)) ∧ (∃v ∈ X.V : v.TC = Ti ∧ v.list) ⇒ Ci = oo)
}

ACTION A-Commit( X: data-item, Ti, Tc: Time-stamp, Cf: Counter)
{PRE: ∀x ∈ X.V ∧ x.Cr = Cc ∧ ∀a ∈ X.V ∧ x.S = commit ∧ ∀v ∈ X.V ∧ x.S = commit : x.Cb ≤ x.Cb0
POST: (∃u ∈ X.V : x.TC = Ti) ⇒ h.Cc = Cf) ∧ (∃v ∈ X.V ∧ v.TC = Ti : v.S = commit ∧ v.Cc = ∞ ∧ v.Cb = Cc ∧ v.Cr = Cf ∧ v.TC = Tc) ∧ (∃v ∈ X.V ∧ v.TC = Ti ∧ v.list : v.Cc = max(Cc, Cf))
}

ACTION A-Abort( X: data-item, Ti: Time-stamp)
{PRE: X.V = Q ∧ C = {v | v ∈ X.V ∧ v.TC = v.Tp})
POST: X.V = Q.C
}

ACTION RO-Read( X: data-item, C: Counter, var a: value)
POST: a = A
}

Figure 4: OCC-VTI read, write, validate, commit and abort specifications

6 Proof of correctness

A serial history h = < T0, T1, T2, ..., Tn, Tf > of transactions is represented by an ordered list of transactions Ti, 0 ≤ i ≤ n, preceded by a dummy transaction T0 = W[D] and terminated by a dummy transaction Tf = R[D]. A transaction set H consists of the transactions Ti, 1 ≤ i ≤ n, joint with the set {T0, Tf}. The set H is composed of a set of RO-transactions Hro and a set of W-transactions Hw, with H = Hro ∪ Hw and Hro ∩ Hw = ∅. A write by a transaction Ti to a version xTi, that is not read by any transaction Tj is called a useless write, denoted by Tiwx. The set Hw is defined by { Tiwx | none of the transactions Tj ∈ H reads the value xTi; written by Ti}. The read-from relation rf(h) in history, h, is defined by: { (Ti, xTi, Tj) | Ti reads the value xTi; written by Tj }.

It must be verified that a history h of read and write actions, as ordered by the algorithms and by the partial order prescribed by the transactions, is equivalent with a serial history
of these same transactions. This is proven by making a directed graph \( G \) with vertices 
\( V = H \cup H^u \) and edges \( E \), which represent the causal order of the read and write actions 
on individual items of \( D \). For a given edge \( e \in E \), \( s(e) \) is the source vertex and \( d(e) \) is the 
destination vertex. In [Vid85] it is shown that the graph of view serializable histories has 
certain characteristics. Edges are added to the graph \( G \) to demonstrate that the graphs 
describing the histories enforced by the CCA's have these characteristics.

Before the correctness proof, a few hypotheses are made about the implementation of 
the actions described in section 4. Motivated by the distribution of the data-items over 
the processors, it is assumed that a monotonously increasing clock is associated with each 
data-item. These clocks represent the local times of the data-items. A generally accepted 
hypothesis on the specified actions needs to be stated.

**Hypothesis 1** The ACTIONs specified in section 4 are executed atomically.

It is important that A-Validate and A-Commit constitute one concurrency atomic action. 
The following example will show that when this hypothesis is not met, inconsistent states can 
be reached. Suppose two transactions \( T_1 \) and \( T_2 \) access two items \( \{X,Y\} \). The transactions 
are defined by \( T_1 = R_1[Y] W_1[X] \) and \( T_2 = R_2[X] W_2[Y] \) and they execute according to the 
unserializable schedule \( h = R_1[x_0] R_2[y_0] W_1[x_1] W_2[y_2] \). Both read the initial values \( x_0 \) and 
\( y_0 \) and want to install the values \( x_1 \) and \( y_2 \). Both transactions first validate. The validation 
result is TRUE for both transactions having read the last installed values \( x_0 \) and \( y_0 \). The 
ensuing commit of both transactions installs \( x_1 \) and \( y_2 \), which is an unwanted result as in a 
serial schedule either \( T_1 \) reads \( y_0 \) and \( T_2 \) reads \( x_1 \) or \( T_2 \) reads \( x_0 \) and \( T_1 \) reads \( y_2 \). Therefore, 
a hypothesis is made on the atomicity of the validate and commit actions.

**Hypothesis 2** For all \( X \in D \) and for any transaction \( T \), the action pair 
A-Validate( \( X, T_i, t, C_i, C_f \)) and A-Commit( \( X, T_i, T_c, C_f \)) constitutes one atomic action.

A hypothesis is needed about the commit timestamps of X-writing transactions with 
respect to X-reading RO-transactions. For a given RO-transaction, \( T_{ro} \) and a W-transaction, 
\( T_w \), which conflict in \( X \), the following hypothesis is made:

**Hypothesis 3** If \( T_{ro} \) reads \( X \) at local time \( t \) and \( T_w \) commits at local time \( t' > t \) then 
\( T_{ro}.T_s < T_w.T_c \).

The transaction-graph \( G \) is constructed as described in [Vid85]. The edge set \( E \) of \( G \) has 
the following characteristics:

- An edge labeled \( X \) from \( T_i \) to \( T_j \) for each \( (T_i, x_i, T_j) \in r_f(h) \), called useful \( X \)-edge.
- An edge labeled \( X \) from \( T_i \) to \( T_{ix} \) for each \( T_{ix} \in H^u \), called useless \( X \)-edge.
- An unlabeled edge for each vertex \( T_{ix} \in H^u \) to \( T_f \).
- An unlabeled edge from each RO-transaction other than \( T_f \) to \( T_f \).
- An unlabeled edge from \( T_0 \) to each WO-transaction.

**Definition 1** \( \text{TP}(h) \) An acyclic Transaction Precedence graph \( \text{TP}(h) \) of history \( h \) has the 
following characteristics:
• The vertex set of $TP(h)$ is the same as the vertex set $G$.

• The edge set of $TP(h)$ includes the edge set of $G$ joint with a (possibly empty) set of unlabelled edges.

• For any two edges $e, e'$, where $e$ is a useful $X$-edge and $e'$ another (useful or useless) $X$-edge with $s(e) \neq s(e')$, there exists a directed path in $TP(h)$ that contains them both.

**Theorem 1** [Vid85] A history $h$ is serializable iff there exists an acyclic $TP(h)$.

The graph $G$ is constructed as described above. Every transaction $T_i \in H$ has a start-timestamp $T_s$ and an end timestamp $T_e$ such that $T_i.T_s < T_i.T_e$. $T_s$ and $T_e$ of transactions $T_0$ and $T_f$ are defined such that:

$$\forall T_i \in H \setminus \{T_0, T_f\} : T_0.T_s < T_i.T_s < T_f.T_s \land T_0.T_e < T_i.T_e < T_f.T_e$$

(2)

RO-transactions read their data from one or more W-transactions or $T_0$. For a given RO-transaction, $T_{ro}$, a set of versions exists associated with the read set $RS_{ro}$. The maximum $M_{ro}$ is defined as the minimum of all $C_b$-values of uncommitted versions and the versions for which their commit-timestamp $T_c > T_{ro}.T_o$:

$$M_{ro} = \min X \in RS_{ro} : (\min x \in X, (x.T_o > T_{ro}.T_o \land x.S \neq commit) : x.C_b)$$

(3)

Otherwise: $M_{ro} = (\min X \in I : (\max x \in X : x.C_b))$. An $X$-order, $\prec$, can be defined for the transactions.

**Definition 2 (X-order)** For any two transactions $T_i$ and $T_j$, $T_i \prec T_j$ iff

• $T_i$ and $T_j$ are conflicting W-transactions with $X \in RS_i \cap WS_j \lor X \in WS_i \cap RS_j$ and $T_i.C_f < T_j.C_f$.

• $(i = 0 \land X \in RS_j)$ or $(j = f \land X \in WS_i)$.

• $T_i$ is a W-transaction and $T_j$ is a RO-transaction and $X \in WS_i \cap RS_j$ and $T_i.C_f \leq M_j$.

• $T_j$ is a W-transaction and $T_i$ is a RO-transaction and $X \in WS_j \cap RS_i$ and $M_i < T_j.C_f$.

The consistency of these relations is proven in the following lemma:

**Lemma 1** The $C_f$ values of any two conflicting transactions $T_i$ and $T_j$ differ.

**Proof.** Suppose the transactions conflict in $X$. Suppose that both transactions are writing transactions. Assume without loss of generality that $T_i$ has committed version, $x_i$, after $T_j$ validates and commits version $x_j$. In that case $x_j.T_o = T_j$ and the post condition of $A$-Validate states:

$$C_f = (\max x \in X.V, x.S = commit : x.C_r) + 1$$

(4)

together with the calculation of the maximum $C_f$ value in the Procedure Validate, the calculation of $C_r$ in the pre and post-conditions of $A$-Commit:

**PRE:** $(\forall v \in X.V : v.C_r = C_v)$

**POST:** $(\forall v \in X.V, T_i \in v.list : v.C_r = \max(C_v, C_f))$
and the postcondition in A-Commit:

\[(\forall v \in X.V, v.T_o = T_i : v.S = Commit \land v.C_e = C_f \land v.C_b = C_f)\]  (6)

assure that:

\[T_i.C_f = x_i.C_b \leq x_i.C_e < x_j.C_b = T_j.C_f\]  (7)

This holds for all \(X \in D\). Consequently, the \(C_f\) values of \(T_i\) and \(T_j\) differ when they write to the same item.

Again without loss of generality assume that \(T_i\) writes version \(x_i\) of \(X\) and \(T_j\) reads \(X\). When \(T_j\) reads \(x_i\): \(x_i.T_o \neq T_j\). The post-condition of A-Validate:

\[(\forall v \in X.V : v.T_o \neq T_j \Rightarrow C_f = (\max x \in X.V, T_j \in X.list : x.C_b + 1))\]  (8)

assures that \(T_j.C_f > x_i.C_b = T_i.C_f\).

\(T_j\) can read a version \(x_p\) with \(x_p \neq x_i\). According to the post-condition of A-Read: \(T_j \in x_p.list\). Suppose \(x_p.C_b > x_i.C_b\). Similar to the reading of \(x_i\): \(T_j.C_f > x_p.C_b > x_i.C_b = T_i.C_f\). Suppose that \(x_p.C_b < x_i.C_b\). In this case \(T_j\) reads \(x_p\) before \(T_i\) commits. If \(T_i\) had committed before \(T_j\)'s reading of \(X\), \(T_j\) would have read \(x_i\) or another after \(x_i\) committed version. Two possibilities exist: \(T_i\) commits after \(T_j\) or before. When \(T_i\) invokes A-Commit before \(T_j\), the pre- and post-conditions of A-Commit:

\[
\text{PRE: } b \in X.V \land b.S = commit \land (\forall x \in X.V, x.S = commit : x.C_b \leq b.C_b) \\
\text{POST: } (\exists x \in X.V : x.T_o = T_i) \Rightarrow b.C_e = C_f
\]  (9)

assure that \(x_p.C_e \leq T_i.C_f\). The post condition of A-Validate assures that \(T_j.C_f \leq T_j.C_i < x_p.C_e \leq T_i.C_f\).

When \(T_j\) invokes A-Commit before \(T_i\), the post- and pre-conditions of A-Commit:

\[
\text{PRE: } (\forall v \in X.V : v.C_r = C_v) \\
\text{POST: } (\forall v \in X.V, T_j \in v.list : v.C_r = \max(C_v, C_f))
\]  (10)

assures that for the entry \(T_j \in x_p.list: x_p.C_r \geq T_j.C_f\). The post condition of A-Validate assures that \(T_j.C_f \leq x_p.C_r < T_i.C_f\).

Consequently, the \(C_f\) values of any two conflicting \(T_i\) and \(T_j\) differ.  \(\square\)

**Lemma 2** For two related transactions \(T_i\) and \(T_j\), with \(\{X, Y\} \subseteq (RS_i \cap WS_j) \cup (RS_j \cap WS_i)\):

\(T_i \prec T_j \Rightarrow T_i \prec T_j\)

**Proof.** Assume both transactions to be conflicting W-transactions. According to lemma 1 the \(C_f\)-values differ and according to the order definition 2 the transactions are ordered. The \(C_f\) values are independent of the item in which they conflict. Consequently, the lemma is true for W-transactions.

When both transactions are RO-transactions, no order is defined and the lemma is trivially true.

Consider that \(T_j \in H^x\) and \(T_i \in H^y\). No distinction exists between \(X\) and \(Y\). Consequently, it is sufficient to prove that: \(T_i \prec T_j \Rightarrow T_i \prec T_j\). In the case \(T_i \prec T_j\), the third bullet in definition 2 applies. Because \(\{X, Y\} \subseteq WS_i \cap RS_j\), \(X\) can be replaced by \(Y\). This
Figure 5: impossible cycle in transaction order

is exactly the definition for $T_i \prec T_j$. The case that $T_j \in H^w$ and $T_i \in H^{ro}$ is treated in the same way. □

Lemma 3 The order of definition 2 is acyclic.

Proof. According to lemma 1 and the ordering of the numbers assigned to the $C_f$ values, the sets of conflicting W-transactions are ordered without cycles. No order is defined between RO-transactions, but the ordering of RO-transactions with respect to W-transactions is different from the ordering between W-transactions. Consequently, cycles can occur when a RO-transaction $T_{ro}$ is ordered between two W-transactions, $T_{w1}$ and $T_{w2}$, such that $T_{w1} \prec T_{w2} \wedge T_{w2} \prec T_{ro} \prec T_{w1}$, as shown in fig 5. From definition 2, the following relations can be derived:

\begin{align}
T_{w2} \prec T_{ro} \Rightarrow T_{w2}.C_f < M_{ro} \\
T_{ro} \prec T_{w1} \Rightarrow M_{ro} \leq T_{w1}.C_f
\end{align}

Combining both relations leads to: $T_{w2}.C_f > T_{w2}.C_f$. Because $(Z \in RS_{w1} \cap WS_{w2}) \vee (Z \in WS_{w1} \cap RS_{w2})$, it follows that $T_{w2} \prec T_{w1}$ and the cycle is impossible. □

The following abbreviation is used:

$$T_i \prec T_j \equiv \exists X \in D : T_i \prec X \prec T_j$$

In the relation $T_i \prec T_j$ both $T_i$ and $T_j$ can be freely exchanged with $s(e)$, $s(e')$, $d(e)$, and $d(e')$.

A transaction $T_j$ is the immediately related successor to a transaction $T_i$ iff $T_i \prec T_j$ and there is no other transaction $T_k$ such that $T_i \prec T_k \prec T_j$. The graph $G'$ is constructed from
G by adding edges between related transactions. From each vertex \( T_i \) to its immediately related successors \( T_{i+k} \) and to \( T_{iz} \) and from \( T_{iz} \) to its immediately related successor \( T_{i+k} \). It will be proven that the graph \( G' \) represents a serializable history for the transactions which constitute \( h \). In Fig. 6, an example of \( G' \) is drawn. A history \( h \) is assumed for the transactions \( T_0, T_{10} - T_{13}, T_f \), where each transaction consists of the following read and write actions:

\[
\begin{align*}
T_{10} &= R_{10}[x_0] R_{10}[y_0] W_{10}[z_{10}] \\
T_{11} &= R_{11}[x_0] R_{11}[z_{10}] W_{11}[z_{11}] \\
T_{12} &= R_{12}[x_0] W_{12}[y_{12}] \\
T_{13} &= R_{13}[y_{12}] W_{13}[z_{13}]
\end{align*}
\]  

(14)

The following history, \( h \), of actions is associated with the 4 transactions:

\[
R_{10}[x_0] R_{12}[x_0] R_{10}[y_0] R_{11}[x_0] W_{10}[z_{10}] R_{11}[z_{10}] W_{11}[y_{12}] R_{13}[y_{12}] W_{11}[z_{11}] W_{13}[z_{13}]
\]  

(15)

The graph \( G \) is represented in fig. 6 by the closed arrows. Remark that \( W_{11}[z_{11}] \) is a useless write. The graph \( G' \) is constructed by adding the dashed arrows to \( G \). An arrow is added from \( T_{10} \) to \( T_{12} \) because they are immediate successors via \( Y \) and from \( T_{11}, T_{12} \) to \( T_{13} \) because they are immediate successors via \( Z \).

Figure 6: Example of graph \( G' \)

**Lemma 4** For any labelled edge \( e \) from \( G \), with \( s(e) = T_i \) and \( d(e) = T_j \), \( T_i \) and \( T_j \) related via \( X \), it holds that \( T_i \prec X T_j \).
Proof. A RO-transaction $T_{ro}$ is the destination of a given edge $e$, or is the source of an edge with destination $T_{f}$. In the first case, the pre-condition in RO-Read and the assignment to $C$ in RO-Transaction:

$$v.C_b < C \land v.S = commit \land C = M_{ro}$$

assures that $s(e).C_f < M_{ro}$ and $s(e)$ precedes $d(e)$ in the defined order.

In the second case, $T_{f}$ follows all other transactions per definition. Consequently, the lemma is true for all edges involving a RO-transaction. The same reasoning is valid for any edge involving $T_{ix} \in H^u$.

Edges which only involve W-transactions are now considered. The version $x_{s(e)}$ written and committed by $s(e)$ is read by $d(e)$. The post-condition of A-Read assures that an entry $d(e)$ is added to $x_{s(e)}.list$. The POST condition of A-Validate:

$$(\forall v \in X.V : v.T_e \neq T_i) \Rightarrow C_f = (\max x \in X.V, T_i \in X.list : x.C_b + 1)$$

returns a $C_f$ values that is larger than $x_{s(e)}.C_b = s(e).C_f$. The maximum of the returned $C_f$ values is calculated by the Validate Procedure. It can be inferred that: $d(e).C_f > x_{s(e)}.C_b = s(e).C_f$.

From the order-definition 2 it follows that the source of a labelled edge is ordered before its destination.

Theorem 2 (View serializability) The graph $G'$ constructed from the ACTION specifications of section 4 is an acyclic $TP(h)$.

Proof. By construction, the vertex set of G is equal to the vertex set of $G'$ and the edge set of $G'$ is equal to the edge set of G with the addition of some unlabelled edges. An acyclic order is defined on all transactions $T_i$ and for all $X \in D$ there is a directed path in $G'$ that includes all X-edges. For two X-edges $e$ and $e'$ with different sources of which at least one is useful, their destinations are different, $d(e) \neq d(e')$, because a transaction reads a given value only from one transaction at the time. From lemma 4 it is known that $d(e) \succ s(e)$ and $d(e') \prec s(e')$.

Suppose that one of them, $e'$ is useless. The case $s(e) \prec s(e')$ is proven in the same way as the case that both edges are useful. When $s(e') \prec s(e)$, $s(e)$ is either the direct successor of $s(e')$ or ordered after the direct successor of $s(e')$. The destination of the useless write, $d(e')$, is defined to be ordered before the direct successor of the source: $d(e') \prec s(e')$.

Suppose both $e$ and $e'$ are useful edges. In that case $s(e)$ and $s(e')$ are both W-transactions because no labelled edges exist from a RO-transaction or a from a vertex $T_{ix} \in H^u$. Without loss of generality it is assumed that $s(e) \prec s(e')$. When $d(e) = s(e')$, the theorem is trivially true by taking the X-path from $T_0$ to $s(e)$ and via $d(e) = s(e')$ to $d(e')$ and finally to $T_f$. When $d(e) \neq s(e')$, the case that $d(e)$ is a W-transaction is treated first, followed by the case that $d(e)$ is a RO-transaction.

Assume $d(e)$ is a W-transaction. The proof is similar to the proof of lemma 1 when $d(e)$ reads a version $x_p$ with $x_p.C_b < x_i.C_f$ with $x_i$ written by $T_i = s(e')$. It was demonstrated that $s(e')$ commits after the reading of $X$ by $d(e)$. Two cases were considered: (1) $d(e)$ commits after $s(e')$ and (2) $d(e)$ commits before $s(e')$. In both cases it was demonstrated that $d(e).C_f < s(e').C_f$. 16
Consider the case that \( d(e) \) is a RO-transaction. When \( s(e') \) is committed before \( d(e) \) reads, then according to the PRE-condition of RO-Read:

\[
\forall x \in X. V : (x.C_b \leq v.C_b \lor x.C_b > C)
\]  

and the selection of \( C = M_{d(e)} \):

\[
s(e').C_f = x_{s(e')}C_b > M_{d(e)}
\]  

When \( s(e') \) commits after the reading by \( d(e) \), \( s(e') \) also commits after \( s(e) \) and according to the post-condition of A-Validate \( s(e').C_f = x_{s(e')}C_b > x_{s(e)}.C_f = s(e).C_f \). According to hypothesis 3: \( s(e').T_e > d(e).T_e \), which means that the value of \( M_{d(e)} \) before and after the commit of \( s(e') \) is the same and \( s(e').C_f > M_{d(e)} \). According to the order of definition 2: \( d(e) < s(e') \) in both cases.

From the ordering and the additional unlabelled edges added to \( G' \), an acyclic path can be constructed started at \( T_0 \) from \( s(e) \) to \( d(e) \), from \( d(e) \) to \( s(e') \) and from \( s(e') \) to \( d(e') \) to end at \( T_f \).

\[\square\]

7 HRT/SRT communication

A Real-Time environment puts constraints on the moments actions have to be executed. Starting from the requirement that no actions are better than too late actions, each transaction will have a deadline attribute. Transactions that do not finish before their deadlines loose their usefulness and may be aborted.

A careful scheduling of the HRT transactions should prevent that HRT-transactions interfere with each other in such a way that some deadlines are not met, or that preconditions cannot be met. For the purpose of this paper, it is assumed that such a scheduling of HRT transactions is actually possible albeit imperfectly from an efficiency point of view. The separation between HRT- and SRT-transactions, as introduced in section 1, necessitates the formulation of conditions under which objects can be accessed by both HRT- and SRT-transactions.

During the execution of the HRT schedule the meeting of the deadlines of the HRT-transactions is guaranteed as long as the schedule is not perturbed by the execution of the SRT-transactions. Consider two transactions \( T_i \) and \( T_j \) with \( T_i \) of a different type (HRT or SRT) from \( T_j \). In the histories not only the Read and Write actions but also the Commit actions are included. \( C[X] \) stands for the commit of data-item \( X \in D \) and \( C[Q] \) for the commit of the set of data-items \( Q \subseteq D \). The notation \( t(A_i[X]) \) stands for the time on the local clock of item \( X \) at the moment of excution of action \( (A = R, W \text{ or } C) \) by transaction \( T_i \).

**Lemma 5** The HRT-transaction, \( T_i \) (\( T_j \)), is not perturbed by the SRT-transaction, \( T_j \) (\( T_i \)), if transactions are ordered such that: \( t(R_i[X]) < t(C_i[X]) \Rightarrow T_j < T_i \).

**Proof.** Consider the case that \( T_i \) is a HRT-transaction and \( T_j \) a SRT transaction. At the commit of the HRT transaction \( T_i \), \( T_i \) must be aborted if \( T_i \) has written a version \( x_i \) to be read by an already committed \( T_j \). This is the case when \( t(R_i[X]) < t(C_i[X]) \land T_i < T_j \). The abort of a HRT transaction should be avoided and therefore \( t(R_i[X]) < t(C_i[X]) \) implies that \( T_j < T_i \).
Consider the case that \( T_j \) is a HRT transaction and \( T_i \) a SRT-transaction. Consider two transactions: \( T_i = \ldots W_i[x_i] C_i[X] \) and \( T_j = \ldots R_j[X] C_j[X] \). A possible history, with \( T_i \prec T_j \land t(R_j[X]) < t(C_i[X]) \), is given by: \( \mathcal{h} = \ldots W_i[x_i] R_j[x_i] C_j[X] C_i[X] \). In this case \( T_j \) should wait for the commit of \( T_i \) before it can read the value \( x_i \) written by \( T_i \). As HRT-transactions should not wait for SRT-transactions (with unknown durations), \( T_j \) should have read an earlier value of \( X \) and \( T_j \prec T_i \) should hold.

For the following take \( \mathcal{H}^h(\mathcal{H}^s) \) as the set of HRT(\text{srt})-transactions, \( Q \) is the set of items accessed by transactions from both \( \mathcal{H}^h \) and \( \mathcal{H}^s \). Take \( T_h \in \mathcal{H}^h, T_s \in \mathcal{H}^s \) and \( p,q \in Q \). From the above lemma the following theorem can be concluded.

**Theorem 3** For all \( T_h, T_s, q \) and \( p \), with \( T_h = \ldots R_h[q] \ldots C_h[q,p] \) and \( T_s = \ldots R_s[p] \ldots C_s[q,p] \), no interleaving is possible between the Read action and the Commit action.

**Proof.** Assuming the theorem is not valid, try to construct a valid interleaved history. Suppose the \( R_s[p] \) is interleaved between \( R_h[q] \) and \( C_h[q,p] \). The action \( C_s[q,p] \) can be scheduled before or after \( C_h[p,q] \). Independent of the position of \( C_s[q,p] \), an unwanted situation is created according to lemma 5:

\[
\begin{align*}
    t(R_h[q]) < t(C_s[p,q]) \Rightarrow & T_h \prec T_s \\
    t(R_s[p]) < t(C_h[p,q]) \Rightarrow & T_s \prec T_h
\end{align*}
\]

This constitutes a cycle and leads to a non-view-serializable history. The interleaving of \( R_h[q] \) between \( R_s[p] \) and \( C_h[q,p] \) leads to the same contradiction. Consequently, the only valid history is one without the discussed interleaving.

A simple and valid way to solve this problem is to execute HRT- and SRT transactions in a perfect serial order. SRT-transactions which are preempted by HRT transactions must be aborted. The scheduling of the SRT transactions such that they never are executed concurrently with conflicting HRT-transactions restricts the possible schedules enormously and leads to many SRT-transaction aborts when conflicting HRT transactions are started during the SRT-transaction execution. Another option is to relax the view-serializability requirements. When HRT-transactions deliver sets of values such that not the serial order chosen by the CCA is important for the SRT-transactions but the order in which these sets are committed, another serializability criterion can be formulated. The same is true for values delivered by SRT transactions to HRT transactions. The concept of transfer serializability is introduced:

**Definition 3 (Transfer-serializability)** A history of HRT- and SRT-transactions is transfer serializable iff in an equivalent view-serializable history the values written by a given HRT-transaction are transferred by a single transaction to reading SRT-transactions and vice-versa.

A TR-Read action must be defined which obeys lemma 5. The RO-Read action does not influence the other W-transactions. Consequently, a modified version of the RO-Read action, shown in Fig. 7 as TR-Read, can be used by the SRT-transactions when the HW-objects are read. However, when a HRT-transaction invokes the TR-Read action on a data-item containing uncommitted versions for which the commit is already started, the transaction
is obliged to wait until the commit is finished. The use of TR-Read by HRT-transactions to read SW-objects prevents such intolerable waits when the time to execute a commit of a SRT-transaction is bounded. With bounded commit execution times the $T_c$ values can be chosen such that no waiting occurs.

```
1 ACTION TR-Read( X: data-item, T: Timestamp, var a: value)
2 {PRE: v ∈ X.V ∧ u.v.Tc < T ∧ v.S = commit ∧ v.value = A
3 ∀x ∈ X.V - {v}, x.S = commit : (x.Ic < u.v.Tc ∨ x.Tc > T)}
4 POST: a = A
5 }
```

Figure 7: TR-Read specification

An additional constraint is generated on the use of data-items by this definition. When data-items can be read and written by both HRT- and SRT-transactions, the following scenario can be envisaged: the HRT-transaction $T_H$ writes to A and reads from A and the SRT-transaction $T_S$ also reads from and writes to A. HRT-transactions must follow a different CCA to read the versions written by $T_H$ from the one to read the version written by $T_S$. A labelling of A's versions is then necessary. This actually points to a separation of A into a HRT-part and a SRT-part but may lead to badly understood side-effects for the programmer.

Therefore, three sets of objects are defined: (1) SRT-objects only used by SRT-transactions, (2) HRT-objects only used by HRT-transactions and (3) HS-objects used by both types of transactions. The HS-objects are separated into two categories: Hard-Write (HW-)objects and Soft-Write (SW-)objects, where:

- **HW-object** can be written and read by HRT-transactions, but only read by SRT-transactions with the TR-Read action.
- **SW-object** can be written and read by SRT-transactions, but only read by HRT transactions with the TR-Read action.

### 7.1 Proof of HRT/SRT serializability

The *hrt-graph*, $G_{hrt}$, is constructed in the same way as the graph $G$, described in section 6, but includes both HRT and SRT transactions. All edges between HRT-transactions and between SRT-transactions are drawn as specified for the transaction-graph $G$.

For each HRT-transaction that writes to a set of HW-objects $HO$, a transaction $T_{HO}$ is imagined that copies all items from $HO$ to another set of objects $HO'$: $T_{HO} = R[HO] W[HO']$. The same is done for each SRT-transaction that writes to a set of SW-objects: $T_{SO} = R[SO] W[SO']$. $T_{SO}$ and $T_{HO}$ are called specification transactions.

The definition of immediately related successor is restricted to transactions of the same class, HRT or SRT. The graph $G_{hrt}$ is extended by adding vertices which represent the above mentioned $T_{SO}$ and $T_{HO}$ and their connecting transfer-edges. Edges $e$ labelled with $X ∈ HW$ with as destination a SRT-transaction $T_a$ are replaced by an edge labelled $X$ from $s(e) = T_h$ to a new HRT-vertex $T_{hd}$ and an edge labelled $X'$ from $T_{hd}$ to $d(e) = T_h$. Edges $e$ labelled with $X ∈ SW$ with as destination a HRT-transaction $T_h$ are replaced by an edge labelled $X$ from $s(e) = T_s$ to a new SRT-vertex $T_{sd}$ and an edge labelled $X'$ from $T_{sd}$ to $d(e) = T_h$. The $T_x$ value of a specification transaction, $T_{id}$ is defined as:
\[ T_{id} : T_e = T_i : T_e \] (20)

The graph \( G'_{hrt} \) is constructed from \( G_{hrt} \) by adding edges from each vertex \( T_i \) to its immediately related successor \( T_{i+k} \), to \( T_{ix} \) and to \( T_{id} \), from \( T_{ix} \) to \( T_i \)'s immediately related successor \( T_{i+k} \) and from \( T_{id} \) to \( T_i \)'s immediately related successor \( T_{i+k} \).

In Fig. 8 this is visualised for the transactions of \( T_H \) and \( T_S \), acting on \( hrt : \)

\[ (hrt :) T_H = R_H[H] R_H[S] W_H[H] \]
\[ (srt :) T_S = R_S[H] R_S[S] W_S[S] \] (21)

The order in which they can be executed by the CCA is:


The addition of the specification transaction \( T_{od} \) makes that \( T_3 \) is ordered with respect to \( T_{od}, T_H \) and \( T_S \) via \( H \) and \( S \), \( T_j \) is ordered with respect to \( T_H \) and \( T_S \) via \( H \) and \( S \), \( T_{od} \) is ordered with respect to \( T_H \) and \( T_S \) via \( S' \) and \( H' \), but \( T_H \) and \( T_S \) are not related and consequently no longer ordered with respect to each other.

![Diagram of the graph \( G_{hrt} \)](image)

Figure 8: An example of the graph \( G_{hrt} \)

Now the order can be defined:

**Definition 4 (hrt-order)** For any two transactions \( T_i \) and \( T_j \), \( T_i \prec T_j \) iff

- \( T_i \) and \( T_j \) are transactions of the same class and ordered according to definition 2.
- \( T_i \) writes \( X \in HO \cup SO \) and \( T_j = T_{id} \).
- \( T_j \) accesses \( X \in HO \cup SO \) and \( T_i \) is a specification transaction that writes \( X' \) and \( T_i, T_e \prec T_j, T_s \).
- $T_i$ accesses $X \in HO \cup SO$ and $T_j$ is a specification transaction that writes $X'$ and $T_i.T_s \leq T_j.T_s$.

It is important to note that the transfer-serializability concept is independent of the choice of CCA. The SRT-transactions may use a different CCA from the HRT-transactions. The order definition must then be adapted correspondingly.

**Lemma 6** The order of definition 4 is acyclic.

**Proof.** Consider the ordering without the specification transactions. Two mutually unordered sets of HRT and SRT transactions exist. According to lemma 3 each individual set is ordered. The addition of the specification transactions can lead to cycles in two ways: (1) when a pair of HRT(SRT)-transactions are differently ordered with respect to each other than with respect to a given specification transaction and (2) when a cyclic ordering can be found from a specification transaction $T_{id}$ via a set of HRT(SRT) transactions to another specification transaction $T_{jd} \neq T_{id}$ and from $T_{jd}$ via a set of SRT(HRT) transactions back to $T_{id}$.

ad 1) Consider that the following order is valid: $T_j<T_{id} \prec T_k$, with $T_j$ and $T_k$ HRT(SRT)-transactions of the same class. According to definition 4: $T_j.T_s \leq T_{id}.T_s < T_k.T_s$. $T_j$ and $T_k$ write to the same items and conflict. The condition $T_j.T_s < T_k.T_s$ implies that $T_k$ committed after $T_j$. As shown in the proof of lemma 1, this implies that $T_j.C_f < T_k.C_f$. For conflicting $T_j$ and $T_k$ this means that according to definition 2: $T_j < T_k$. Consequently, no cycle is possible.

ad 2) Without loss of generality, assume that a chain of HRT(SRT)-transactions is ordered after the specification transaction $T_{id}$ and before the specification transactions $T_{jd}$. According to the definition 4: $T_{id}.T_s < T_{jd}.T_s$. To create a cycle, a chain of SRT(HRT)-transactions is ordered before the specification transaction $T_{id}$ and after the specification transactions $T_{jd}$. According to the definition 4: $T_{id}.T_s > T_{jd}.T_s$. This constitutes a contradiction. Consequently, the order is acyclic. $\square$

For a given transaction, $T_r$, which uses TR-read and a W-transaction, $T_w$, of a different class which both access $X \in HO \cup SO$, the following hypothesis, similar to hypothesis 3, is made:

**Hypothesis 4** If $T_r$ reads $X$ at local time $t$ and $T_w$ commits at local time $t' > t$ then $T_r.T_s < T_w.T_s$.

It is proven that the graph $G'_{hrt}$ represents a transfer-serializable history for the leaf-transactions which constitute $h$.

**Lemma 7** For any $Y$-labelled edge $e$ from $G_{hrt}$, with $Y = X$ or $Y = X'$: $s(e) \prec d(e)$.

**Proof.** When $s(e)$ and $d(e)$ are both HRT- or SRT-transactions, lemma 4 applies and $s(e) \prec d(e)$.

One of the two can be a specification transaction. When $d(e)$ is a specification transaction, $s(e) \prec d(e)$ is determined by the hrt-order definition 4. When $s(e)$ is a specification transaction, there exists a $T_i$ such that $s(e) = T_id$. When $T_i$ is a HRT(SRT)-transactions then $d(e)$ is a SRT(HRT)-transactions.
The transaction, \( d(e) \), only reads \( X' \) after the commit of \( T_i \). From the precondition of TR-Read follows: \( d(e).T_s > T_i.T_e \). From the definition of \( T_e \)-value of the specification transaction: \( s(e).T_s < d(e).T_s \), which according to the definition 4 leads to the conclusion that \( s(e) \prec X' d(e) \). Combining all cases, it follows that \( s(e) \prec d(e) \).

**Theorem 4 (Transfer-serializability)** The graph \( G_{ht}^{\prime} \) constructed from the ACTION specifications of section 4 and transfer-edges is an acyclic TP(h).

**Proof.** For two X-edges \( e \) and \( e' \) with different sources of which at least one is useful, say \( e' \), their destinations are different, \( d(e) \neq d(e') \), because a transaction reads a given value only from one transaction at the time. Without loss of generality it is assumed that \( s(e) \prec s(e') \). From lemma 7 it is known that \( d(e) \succ s(e) \) and \( d(e') \succ s(e') \). By construction and by the definition of the transfer-edges, HRT transactions access other data-items than SRT-transactions and are not related. The identical labels of \( e \) and \( e' \) imply that when one of the four transactions is a HRT(SRT) transaction then the other transactions are either HRT(SRT) transactions or specification transactions.

When \( s(e), d(e), s(e') \) and \( d(e') \) are all HRT-transactions or SRT-transactions, theorem 2 applies.

Only two cases need to be considered: (1) \( d(e) \) is a specification transaction and (2) \( s(e') \) is a specification transaction.

**Ad (1):** According to the definition of transfer transactions, \( d(e) \) is ordered before the immediate successor of \( s(e) \). \( s(e') \) belongs to the same class as \( s(e) \) (HRT or SRT) and is the immediate successor of \( s(e) \) or \( s(e') \) is ordered after the immediate successor and therefore: \( d(e) \prec s(e') \).

**Ad (2):** When \( s(e') \) is a specification transaction, \( s(e) \) is necessarily also a specification transaction because both edges, \( e \) and \( e' \) are transfer-edges labeled with the same label \( X' \) and only specification transactions are the sources of transfer-edges. There exists a transaction \( T_i \) corresponding with the specification transaction \( T_{id} = s(e) \) and a transaction \( T_j \) corresponding with transaction \( T_{jd} = s(e') \). The precondition of TR-Read implies: \( T_i.T_s < d(e).T_s < T_j.T_s \) when \( T_j \) commits before \( d(e) \) reads. Hypothesis 4 assures that the same is true when \( T_j \) commits after the read-action of \( d(e) \). The definition of the \( T_e \)-value of the specification transaction \( T_{id} \) states: \( T_{id}.T_e = T_i.T_e \). Substitution of the definition into the above equation results in: \( s(e).T_e < d(e).T_s < s(e').T_e \). According to definition 4: \( d(e) \prec X' s(e') \).

A path can be constructed from \( T_0 \) via the acyclic order to \( s(e) \), from \( s(e) \) via \( d(e) \) and \( s(e') \) to \( d(e') \) and from \( d(e') \) via the acyclic order to \( T_f \). Which concludes the proof of the theorem.

**8 Nested atomic objects**

A simple object that does not invoke any other objects can be considered equivalent to a transaction. The execution of the methods of a set of simple objects should obey the same atomicity requirements as a set of distinct transactions. A nested object invokes other objects. Parts of the code of the class separated by invocations of other objects can be identified. A part of a nested object is defined by: (1) it is placed between two object invocations, (2) between
the PROCEDURE start and an object invocation and (3) between an object invocation and
the PROCEDURE end. The objects invoked by an atomic object should be atomic as well.
The concept of atomic nested objects is equivalent to the concept of nested transactions. In
Fig. 9, a possible class structure is shown. The PROCEDURE Trans is identified with one
transaction $T_0$. It is divided into three parts ($T_1, T_3, T_5$) delimited by two object invocations.
The graph in the same figure represents the corresponding nested transaction graph. The
invoked objects $O_1$ and $O_2$ are represented by $T_2$ and $T_4$.

Figure 9: relation between object and nested transaction

The concept of nested transactions was introduced to increase the parallelism inside a
transaction and render recovery more efficient [Mos85, Mos87, BBG89]. The individual read
and write actions no longer needed to be executed in a fixed serial order, but could be
grouped and be executed in parallel. When a subtransaction of a transaction fails, only the
subtransaction needs to be recovered.

Although the code in the class should be executed in a serial fashion determined by the
language semantics, it is not excluded that object invocations can be done in parallel. When
no parallel object invocation are possible, a simplified CCA can be found.

A nested transaction $T_i$ either consists of a sequence of read-actions $R_i$ and write-actions
$W_i$, acting on items $I \in D$ partially ordered in time, or consists of a set of transactions.
When the transaction consists of a sequence of actions, it is called a leaf-transaction. When
a transaction consists of a set of transactions and is part of another transaction, it is called
an intermediate-transaction. A transaction that is not part of any other transaction is called
a root-transaction. A root-transaction is defined to be its own root. The function $root(T)$
returns TRUE iff $T$ is a root-transaction. Fig. 10 illustrates this further. Not shown are leaf-
transactions which can be root-transactions at the same time. When a transaction $T_i$ is part
of a transaction $T_j$, $T_j$ is called a parent of $T_i$, denoted by $T_j = parent(T_i)$. The transaction
$T_i$ is an ancestor of $T_j$, denoted by $T_i \in ancestor(T_j)$, when $T_i$ is a parent of $T_j$ or when $T_i$ is
a parent of another transaction $T_k$ which is an ancestor of $T_j$. The function $\text{ancestor}$ has as domain the transactions and as range a set of transactions. The inverse functions are defined by:

$$T_j = \text{parent}(T_i) \Leftrightarrow T_i \in \text{child}(T_j)$$
$$T_j \in \text{ancestor}(T_i) \Leftrightarrow T_i \in \text{descendant}(T_j)$$

In the Fig. 10, $T_4^1$ is an ancestor of $[T_2, T_4]$ and $T_7^3$ is an ancestor of $T_6^2$, $T_4^1$, $T_6^1$ and $[T_1, T_7]$. The nested structure limits the possible orders in which the leaf-transactions can be executed. For example, the order $[T_2, T_4, T_12, T_1]$ in Fig. 10 can be changed to $[T_1, T_12, T_11, T_13]$ but not to $[T_10, T_11, T_12, T_13]$.

A nested transaction $T$ executes a set of transactions denoted by $N$. $N$ consists of 1 or more transactions $T_j$. All root transactions are executed such that their result is the same as the result of some serial execution of the same transactions. For a given nested transaction, the transactions $T_i \in N$ are executed such that their result is the same as the result of some serial execution of these same transactions. In Fig 10, a serial execution of root transactions $[T_7^3, T_9^1, T_13^2]$ can be realized with the serial execution of the leaf-transactions $[T_1, \ldots T_{13}]$.

The nesting of transactions has a consequence for the exception atomicity. The results of a leaf-transaction $T_i$ are only visible to another leaf-transaction $T_j$ iff:

- $T_i$ and $T_j$ have different roots and the root of $T_i$ has committed.
- $T_i$ and $T_j$ have a common ancestor $T_k$ which is composed of at least two transactions $T_i'$ and $T_j'$ with $T_i = T_i'$ or $T_i' \in \text{ancestor}(T_i)$, $T_j = T_j'$ or $T_j' \in \text{ancestor}(T_j)$ and $T_i'$ has
In Fig. 10, $T_5$ sees the results of $T_5$ when $T_5$ has committed and $T_5$ sees the results of $T_3$ when $T_4$ has committed.

8.1 Nested transaction algorithm

In Fig. 11, the nested transaction algorithm for OCC-NVTI adapted from Fig. 2 is shown. At the beginning of the nested transaction a timestamp, $T_s$, as function of the parent timestamp and the current time is created. $T_s$ uniquely identifies a given transaction. For all transactions of which this transaction is the parent, the leaf or nested procedure is invoked. (see Fig. 2 for leaf-transaction) For all data-items that are accessed by the transaction, a commit is executed.

\begin{verbatim}
1  PROCEDURE Nest-Trans(N: transactions, T: timestamp)
2      Create timestamp $T_s = f(T, time)$
3      \forall Ti \in N : IF Ti is nested THEN Nest-Trans(Ti, $T_s$)
4          ELSE W-Transaction(Ti.RS, Ti.WS, Ti.I, $T_s$)
5  FI
6  Create timestamp $T_c = 0$
7  Validate(T, $T_c$, $C_f$, $C_i$)
8      IF ($C_f < C_i$) THEN \forall X \in N.I: A-Commit( X, $T_c$, $C_f$)
9          ELSE \forall X \in N.I: A-Abort( X, $T_c$)
10 END
\end{verbatim}

Figure 11: Nested transaction algorithm for OCC-NVTI

The visibility of the results within a nested transactions has a consequence for the specification of the actions. Results of intermediate- and leaf-transactions are available to other transactions when they have an ancestor in common. At the commit of a leaf- or intermediate transaction, ownership of its version is passed to its parent; at the commit of the root, the version’s state is passed to commit. The Read and Commit actions are adapted accordingly. The parameters for Validate and Commit depend on the CCA under consideration. In case of OCC-BV, Validate is only invoked for leaf-transactions. In case of 2PL and MVTSO, Validate is not invoked at all. The actual implementation of the Validate and Commit procedures is discussed in section 10.3.

8.2 Specification of nested transaction actions

Four sets of actions based on 2PL, OCC-BV, OCC-NVTI and MVTSO are specified. All four incorporate versions to allow the ordering of the RO-transactions with respect to the W-transactions such that the values of data-items are available to RO-transactions with minimal waiting times. The read- and write-actions are specified with PRE- and POST-conditions.

The representation of the data-elements contains attributes for all four types of CCAs.

In case of 2PL, a given data-item $X$ is a tuple $<V, L>$, which consists of a set of versions $V$ and a lock $L$. Each version $x_i \in V$ is a tuple $<T_w, T_o, S, value>$.

In case of OCC-BV, a given data-item $X$ is a set of versions $V$. Each version $x_i \in V$ is a tuple $<T_w, T_o, S, list, value>$. 

25
In case of MVTSO, a given data-item \( X \) is a set of versions \( V \). Each version \( x_i \in V \) is a tuple \( < T_w, T_e, T_o, S, \text{value}> \).

In case of OCC-NVTI, a given data-item \( X \) is a set of versions \( V \). Each version \( x_i \in V \) is a tuple \( < C_b, C_r, C_e, T_e, T_o, S, \text{list}, \text{value}> \).

The elements of the tuples are: (1) \( T_w \) is the timestamp of the writing W-transaction, (2) \( T_r \) is the largest timestamp of all the transactions that have read this version, (3) \( T_o \) is the timestamp that identifies the owner of the version, (4) \( C_b \) is the begin counter of the validity interval defined by the \( C_f \) value of the writing transaction, (5) \( C_r \) is the maximum of the \( C_f \) values of the transaction which read this version, (6) \( C_e \) is the end counter of the validity interval, (8) \( S \) is the state of the version, (9) \( \text{list} \) is a list of identifiers of transactions that read this version and (10) \( \text{value} \) is the value of the item's version. The state \( S \) of a version is (1) tentative: the version is created but not yet usable by other transactions than the creating one, (2) validating: the transaction that created this version has finished all write and read actions but has not yet asserted their validity and (3) commit: the version can be used by transactions with roots different from the creating one. A lock \( L \) is a tuple \( < s, l > \), which consists of a state \( L.s \) and a list \( L.l \) of timestamps which uniquely identify the W-transactions using this lock. The lock-state \( L.s \) can be in three states: (1) free: no W-transaction uses this item, (2) write: one W-transaction uses this item for writing and (3) read: one or more W-transactions use this item for reading. A lock is only used by W-transactions; RO-transactions don’t use the lock but only use the time-stamp \( T_e \) which determines if a version can be read.

2 PROCEDURE RO-Transaction(RS: read actions, I: data-items, t: timestamp)
3 Create timestamp \( T_i := f(t, \text{time}) \)
4 \( \forall x \in \text{I} \in RS: \text{RO-read}(X, T_e, \lambda_X) \)
5 execute calculations
6 END

ACTION RO-Read( X: data-item, T: Timestamp, var a: value)
9 {PRE: \( v \in X.V \land v.T_e < T \land v.S = \text{commit} \land v.\text{value} = A \)}
10 \( \forall (v \in X.V - (v), x.S = \text{commit} : (x.T_e < v.T_e \land x.T_e > T)) \)
11 POST: \( a = A \)

Figure 12: RO-Read specification and RO-Transaction Algorithm for 2PL and OCC-BV

All the algorithms select a serial order of the transactions, which determines which transactions read the values last modified by a preceding transaction. MVTSO algorithms order with respect to the start time of the transactions, OCC algorithms order with respect to the validation time and 2PL algorithms order with respect to the first lock conflict between any two related transactions.

In Fig. 12, the algorithm for a RO-transaction is shown in case of OCC-BV and 2PL. The specification of the RO-Read states that a version with the largest write timestamp is selected which has a commit timestamp smaller than the timestamp, \( T \), of the reading transaction. For MVTSO (see Fig. 13), the version with the largest write-timestamp \( T_w \) is selected. For OCC-BV and 2PL (see Fig. 12), the version with the largest commit-timestamp \( T_c \) is selected. The value of the selected version is returned. In case of MVTSO, the \( T_c \) values of the versions \( x \) with \( x.T_w < T \) are set equal to the \( T_w \) value of their immediate successor \( x' \): \( x.T_c = x'T_w \).
PROCEDURE RO-Transaction(RS: read actions, I: data-items, t: timestamp)

Create timestamp $T_e = f(t, t_e)$

IF $\exists x \in I: x.T_e > T_e$ THEN $T_m := (\min x \in I, x.S \neq \text{Commit} \lor x.T_e > T_e : x.T_w)$

ELSE $T_m := T_e$

Vac[X] $\in$ RS: RO-read(X, T_m, A_x)

execute calculations

END

ACTION RO-Read( X: data-item, T: Timestamp, var a: value)

{PRE: $v \in X.V \land v.T_w < T \land v.S = \text{commit} \land v.value = A$

$(\forall x \in X.V : (x.T_w \leq v.T_w \lor x.T_o > T))$

POST: $a = A \land \text{Close}(v)$}

FUNCTION Close( v : version)

{PRE: $\exists X \in D : v \in X.V \land x_m.T_o = (\max x \in X.V, x.T_w < v.T_w : x.T_o) \land v.T_e = T$

POST: $x_m.T_r = T \land \text{Close}(x_m)$}

Figure 13: RO-Read specification and RO-Transaction Algorithm for MVTSO

The case for OCC-VTI is already explained in section 4

8.3 Two Phase locking

The locking algorithm for nested transactions differs from traditional 2PL in the treatment of the locks [Mos85]. Locks set by transactions are inherited by the parents after commit of these transactions. Once a write-lock has been set, transactions can read write-locked data-items as long as the locking transaction is an ancestor of the read-requesting transaction. This strategy assures that when an intermediate transaction aborts, removing intermediate results, all modifications done by one or more transactions that read those intermediate results are also removed. Exception atomicity is maintained inside the nested transaction, because reads on new versions can only take place when these new versions are owned by an ancestor of the reading transaction. A write lock can be set by a leaf-transaction, when all owners of the lock are ancestors of the leaf-transaction. The only owner of the write-lock is the leaf-transaction until its commit.

In Fig. 14, the specifications for the reading, writing and committing of one data-item are shown when the 2PL strategy is used. The last written version (identified by its commit timestamp $T_e$) is always selected for reading. The reading can start when no tentative version exists and the lock-state is read or free, or a tentative version is selected of which the owner is an ancestor. After execution of A-Read, the value of $v$ is returned, the reading transaction $T$ has been added to the lock list and the free lock-state is set to read.

The writing can only start when the lock-state is free or the lock-list is occupied by the invoking transaction or by its ancestors. After execution of A-Write, a version $v$ which contains the specified value is added to the data-item $X$. The lock-state is set to write with $T$ the only locking transaction specified in the lock-list and the version state is set to tentative. The writer and owner of the version identified by respectively $T_w$ and $T_o$ are set to $T$.

The commit is executed on data-items $X$ which are locked by the executing transaction ($T \in X.L.i$). A-Commit starts with the definition of the set $C$ of versions owned by the committing transaction $T$. After A-Commit, the executing transaction identifier $T$ is removed from the lock-list $L.i$ of $X$. When the root commits, the state of the lock $L.s$ is set to free.
ACTION A-Read( X: data-item, T: Timestamp, var a: value)
{PRE: v ∈ X.V ∧ X.L.l = L ∧ v.value = A ∧ X.L.s = LS
∧(X.L.a ≠ write ∨ (X.L.s = write ∧ v.To ∈ ancestor(T)))
∧(∀x ∈ X.V : v.Tx ≥ a.Tx)}
POST: (LS = free ⇒ X.L.s = read) ∧ a = A
∧X.L.l = (L ∪ {T})
}

ACTION A-Write( X: data-item, T: Timestamp, a: value)
{PRE: (X.L.s = free ∨ X.L.l = {T}) ∨ ∀T, ∈ X.L.l : T, ∈ ancestor(T))
∧ X.V = Q ∧ a=A
POST: v = newly created version ∧X.V = Q ∪ {v}
∧a.S = tentative ∧ v.value = A ∧ v.To = T ∧ v.To = T
∧X.L.s = write ∧ X.L.l = {T}
}

ACTION A-Commit( X: data-item, T: Timestamp, t: Timestamp)
{PRE: X.L.l = L ∧ X.V = Q ∧ T ∈ L
∧C = {v | v ∈ Q ∧ v.To = T} }
POST: (∀u ∈ C : (v.Tw = T ⇒ u.Tw = t))
∧(root(T) ⇒ (v.S = commit ∧ X.L.l = φ ∧ X.L.s = free))
∧(¬root(T) ⇒ v.Tw = parent(T) ∧ X.L.l = (L - {T}) ∪ {parent(T)})
}

ACTION A-Abort( X: data-item, T: Timestamp)
{PRE: X.V = Q ∧ C = {v | v ∈ X.V ∧ v.To = T}
POST: X.V = Q - C
}

Figure 14: 2PL read, write and commit specifications

When the transaction is not the root, T's parent is added to the lock-list and the owners of all versions of C are set to the parent of T. The Commit times are set to t for all leaf-transactions which wrote this version (Tw = T).

The A-Abort action specifies that the tentative versions owned by this transaction are removed.

8.4 Multi-Version Timestamp ordering

MVTSO assures that for a set of transactions, the order of the equivalent serialized set is equal to the order in which the transactions are started. For nested transactions, the order of the root-transactions is determined by the absolute start times of the root-transactions, but the order within a nested transaction is determined by the relative start times of the intermediate transactions. As for 2PL, transactions can only read values from versions committed by earlier root transactions or from versions which are owned by ancestors. Each version has a write timestamp of the writing transactions and a read timestamp that is the maximum of the timestamps of all transactions that read this version. A serializable schedule is enforced by the simple rule that for any two versions u and v with write-timestamps Tw < Tt, and v the immediate successor of u, the timestamps Tt of the transactions that read version u obey: Tw ≤ Tt < Tt.

In Fig. 15, the specifications for the reading, writing and committing of one data-item are shown when the MVTSO strategy is used. The reading can start when there is a version, committed by an ancestor (identified with v.To ∈ ancestor(T)) or by a root transaction, with
a write-timestamp smaller than the timestamp of the executing transaction. The version $v$ with the largest write-timestamp is selected. After execution of A-Read, the read-timestamp is updated with the timestamp of the executing transaction when this timestamp is larger than the original read-timestamp. The value of $v$ is returned in $a$.

The writing can start when there is no version with a write-timestamp smaller and a read-timestamp larger than the timestamp of the executing transaction. After execution of A-Write, a new version is created with the specified value and a write-timestamp, owner-timestamp, and read-timestamp equal to the timestamp of the executing transaction. The state of the version is set to tentative.

A-Commit is executed to make the versions accessible to other (root) transactions and their descendants. The set $C$ contains the versions owned by the committing transaction. After A-Commit, the owners of the versions are equal to the parent of the committing transaction. In case of a leaf-transaction ($v.T_w = T$), the commit timestamps of these versions are set equal to the specified commit timestamp. For root transactions, the states of $C$'s versions are set to commit.

The A-Abort action specifies that all versions owned by the executing transaction are removed.

### 8.5 OCC-BV

In Optimistic Concurrency Control with Backward Validation (OCC-BV), the results of all write actions of a given transaction are installed in one atomic step including a "validate" phase. Before, all write actions have taken place without any checks on their validity. The ordering of the serialized equivalent set of transactions is the same as the one dictated by the commit times. At validation time, each transaction verifies that no more recent version
is installed since its reading of the corresponding item. When new versions are installed, the transaction should have read this newer version. Aborts of transactions are then necessary.

**ACTION A-Read:** (X: data-item, T: Timestamp, v: value)

**PRE:**
\[ v \in X.V \land (x.S = \text{commit} \lor x.T_0 \in \text{ancestor}(T)) \land v.l = L \land v.a = A \]
\[ \land (\forall x \in X.V - \{v\} : (x.S = \text{commit} \lor x.T_0 \in \text{ancestor}(T)) \Rightarrow v.T_e > x.T_e) \]

**POST:**
\[ v.l = L \cup \{T\} \land v.a = A \]

**ACTION A-Write:** (X: data-item, T: Timestamp, a: value)

**PRE:**
\[ x.V = Q \land a = A \]

**POST:**
\[ a = \text{newly created version} \]
\[ \land v.S = \text{tentative} \land v.T_0 = T \land v.T_r = T \land v.a = A \land x.V = \emptyset \}

**ACTION A-Validate:** (X: data-item, T: Timestamp, t: Timestamp)

**PRE:**
\[ (\forall x \in X.V : x.S \neq \text{validating}) \land T_L = (\max \{x.S : x.T_0 \in \text{ancestor}(T)\} : x.T_e) \]
\[ \land (\forall x \in X.V : x.T_0 \neq T_L \Rightarrow T \notin x.l) \]

**POST:**
\[ v = (x.V, x.T_0 = T : x.S = \text{validating}) \land (\exists x \in X.V : v.T_0 = T) \Rightarrow t = \text{time opid} \]
\[ \land (\forall x \in X.V : x.T_0 \neq T_L) \Rightarrow t = T_L \circ \text{pid} \]

**ACTION A-Commit:** (X: data-item, T: Timestamp, t: Timestamp)

**PRE:**
\[ C = \{x \in X.V : x.T_0 = T\} \land \forall x \in C : x.S = ST_e \]
\[ \land \forall x \in X.V : (x.T_0 < (\min \{x = C : x.T_e\}) \lor (x.T_e > (\max \{x \in C : x.T_e\})) \land \forall x \in C : (ST_e = \text{validating} \Rightarrow x.T_0 = t) \]
\[ \land (\text{root}(T) \land x.S = \text{commit}) \lor (\neg \text{root}(T) \land x.S = \text{tentative} \land x.T_0 = \text{parent}(T)) \}

**ACTION A-Abort:** (X: data-item, T: Timestamp)

**PRE:**
\[ X.V = Q \land \forall x \in X.V : T = x.T_0 \]

**POST:**
\[ X.V = Q \lor C \}

---

**Figure 16:** OCC-BV read, write, validate and commit specifications

In Fig. 16, the specifications for the reading, writing, validating, and committing of one data-item are shown when the OCC-BV strategy is used. The reading can start by selecting a committed version or a tentative version owned by an ancestor of T (identified by \(v.T_0 \in \text{ancestor}(T)\)). The version v with the highest commit-timestamp is selected. After execution of A-Read, the value of v is returned and the timestamp of the executing transaction is added to the version’s list.

The writing can always start. A new version with a tentative state and the specified value is created. The owner is set to be the writing leaf-transaction.

Only leaf-transactions validate their results. Validation can only start when no other versions are being validated. After execution of A-Validate, the state of the version written by this transaction, \(v.T_0 = T\), is set to validating. The read-actions of the transaction are valid when at validation time there are no new versions created since the reading of the data-items. This is verified by looking at the lists of transactions that read a given version. Only the last validated version (defined by \(T_L\)) is allowed to be read by the validating transaction. A value for the local timestamp of \(T_e\) is returned in t.

Each item X is represented by a list of versions ordered by the commit times of the leaf-transactions. A nested transaction T can own one or more versions of an item. All versions
not belonging to $T$ must be ordered either before all versions of $T$ or after. Consequently, A-Commit can only start when all versions created by leaf-transactions of the committing transaction, $T$, constitute one contiguous sequence. A-Commit is executed on all versions which are created by the committing transaction. When leaf-transactions ($v.S = validating$) are committed, the commit timestamp is replaced by the specified one. In case of a root transaction, the version's state is set to commit; in the other case, the state is set to tentative and the owner-timestamp is set equal to the timestamp of the parent of the committing transaction. In case this data-item was read, the reading timestamp is set equal to the maximum of the committing transaction and former read-timestamp.

The A-Abort action specifies that all versions owned by the executing transaction are removed.

8.6 OCC-NVTI

ACTION A-Read( $X$: data-item, $T_i$: Timestamp, var $a$: value)
{PRE: $v \in X.V \land (v.S = commit \lor v.T_0 \in ancestor(T_i)) \land v.list = L \land v.value = A$
$\land (\forall x \in X.V, x.S = commit \lor x.T_0 \in ancestor(T_i)): x.T_0 \leq v.T_0)$

{POST: $v.list = L \cup (T_i) \land a = A$}

ACTION A-Write( $X$: data-item, $T_i$: Timestamp, var $C_f$: counter, var $a$: value)
{PRE: $X.V = Q \land a = A$

{POST: $v \in newly created version$
$\land n.S = tentative \land v.T_0 = T_i \land v.value = A \land v.list = \emptyset \land X.V = Q \cup \{v\}$}

ACTION A-Validate($X$: data-item, $T_i$, $T$: Time-stamp, var $C_f$, $C_t$: Counter)
{PRE: $\forall x \in X.V: x.S \neq validating$
$P(x) \equiv (\neg root(T_i) \Rightarrow x.T_0 = parent(T_i)) \land (root(T_i) \Rightarrow x.S = commit)$

{POST: $(\forall x \in X.V, x.T_0 = T_i \land x.S = validating)$
$\land (\exists v \in X.V: v.T_0 = T_i \Rightarrow C_f = (\max x \in X.V, P(x) \cdot x.C_r + 1))$
$\land (\exists v \in X.V: v.T_0 = T_i \Rightarrow C_f = (\max x \in X.V, P(x) \land T_i \in x.list : x.C_r + 1))$
$\land (\exists v \in X.V, P(x): T_i \in v.list) \Rightarrow C_t = v.C_t - 1$
$\land (\exists v \in X.V, T_i \notin v.list) \Rightarrow C_t = \infty$
$\land (\exists v \in X.V: v.T_0 = T) \Rightarrow t = time + pid$
$\land (\forall v \in X.V: v.T_0 \neq T) \Rightarrow t = (\max x \in X.V, P(x) \cdot x.T_0) \circ pid$}

ACTION A-Commit( $X$: data-item, $T_i$, $T_0$: Time-stamp, var $C_f$: Counter)
{PRE: $(\forall x \in X.V: x.C_r = C_r \land x.list = L_o \land v.T_0 = T_0)$

$P(x) \equiv (\neg root(T_i) \Rightarrow x.T_0 = parent(T_i)) \land (root(T_i) \Rightarrow x.S = commit)$
$\land b \in X.V \land P(b) \land (\forall x \in X.V, P(x): x.C_r \leq b.C_r)$

{POST: $(\exists x \in X.V: T_0 = T_i \Rightarrow b.C_r = C_f)$
$\land (\forall v \in X.V, T_0 = T_i \Rightarrow b.C_r = \infty \land v.C_r = C_f \land v.T_0 = T_0)$
$\land (\forall v \in X.V, T_0 = T_i \Rightarrow v.C_r = \max(C_r, C_f) \land (\neg P(v) \Rightarrow v.list = (L_o \land T_i) \cup parent(T_i)))$
$\land (\forall v \in X.V, T_0 = T_i : (root(T_i) \Rightarrow v.S = committing) \land (\neg root(T_i) \Rightarrow v.T_0 = parent(T_i)))$}

ACTION A-Abort( $X$: data-item, $T_i$: Timestamp)
{PRE: $X.V = Q \land C = \{v \mid \exists v \in X.V \land T_i = v.T_0\}$

{POST: $X.V = Q \land C$}

Figure 17: nested OCC-NVTI read, write, validate, commit and abort specifications
Optimistic Concurrency Control with Nested Versions and Time Intervals (OCC-NVTI) is based on OCC-VTI. Nesting is added by ordering all transactions with the same parent according to the $C_f$ values. Recursive application of the ordering until the roots are committed results in root transactions which are ordered like the unnested OCC-VTI transactions.

Readable versions are the committed versions and the versions which are owned by an ancestor of the reading transactions. Reading takes place by selecting from all readable versions the last written one which has the largest $T_e$ value. After reading, the value of the version is returned to the invoking transaction and the identifier of the reading transaction is added to the read-list of the selected version.

Writing can always take place. The version is added to the list of versions with state tentative.

When A-Validate starts, no other transactions can be validating the same item. For every item the valid interval is determined. Two cases are considered: (1) the transactions is a root-transaction and (2) the transaction is not a root-transaction. In case 2, the versions belonging to the parent transaction and in case 1, the committed versions are considered. Additionally the read versions are considered by looking at versions with a list containing the identifier of the validating transaction. The returned $C_f$ value is the maximum of all version's $C_b$; the returned $C_l$ value is the minimum of all $C_l$ values.

A-Commit adjusts the $C_b$, $C_l$, and $C_e$ values of the versions and the involved lists. When the Committing transaction is the root, the version's state is set to commit. When not a root, the owner is set to the parent of the committing transaction. The last readable version is identified and the $C_f$ values in the readers list are set equal to the $C_b$ value of the currently committing transaction. The $C_f$ values of the transactions in the readers list of the versions owned by the committing transaction are set equal to $C_b + 1$.

**ACTION A-Commit** (X: data-item, $T_i$, $T_e$: Timestamp, $C_f$: Counter)

\[
\text{PRE: } \quad X.V = Q \land Z = \{ v \mid v \in X.V \land T_i = v.T_e \} \land C_{max} = \{ \max v \in Z : v.C_b \} \\
R = \{ v \mid v \in Z \land v.C_b < C_{max} \} \\
\text{POST: } \quad X.V = Q-R
\]

Figure 18: specification of version removal in A-Commit for OCC-NVTI

Only the latest version of a set of versions owned by the same transaction can be read. All versions owned by the same transaction have the same $C_b$ value which leads to ambiguities. Therefore, only the version with the largest $C_b$ value will be retained; all others are removed (see Fig. 18).

8.7 Commit

The precedence relation together with view serializability make that when for a given data-item several versions are committed by one root-transaction, only one of those versions is actually read by other root-transactions or their descendants. To diminish the number of versions, all unused versions should be removed. The commit timestamp $T_e$ of the last version has been modified to the current time plus some constant $d$. The post conditions of commit
have been extended as shown:

\[
\begin{align*}
\text{OCC-BV, 2PL} & \\
PRE: & \quad v.T_e = (\max x \in C : x.T_e) \\
& \quad t = \text{timestamp}(\text{time}) + d \\
& \quad \land R = \{x \mid x \in C \land x \neq v\} \\
POST: & \quad \text{root}(T) \Rightarrow X.V = Q - R \\
& \quad v.T_e = t
\end{align*}
\]

\[
\begin{align*}
\text{MVTSO} & \\
PRE: & \quad v.T_w = (\max x \in C : x.T_w) \\
& \quad t = \text{timestamp}(\text{time}) + d \\
& \quad \land R = \{x \mid x \in C \land x \neq v\} \\
POST: & \quad \text{root}(T) \Rightarrow X.V = Q - R \\
& \quad v.T_e = t
\end{align*}
\]

9 Proof of correctness

For the proof of OCC-BV, OCC-NVTI and 2PL, the following hypothesis relates the commit timestamps of the transactions with the order in which the transactions execute the commit actions on the individual data-items.

**Hypothesis 5** In case of 2PL, OCC-BV and OCC-NVTI, two RW-conflicting transactions T and T', with commit-timestamps T_e and T'_e, execute their commit-actions at local times t and t' such that: t > t' \iff T_e > T'_e.

A last hypothesis on the implementation of timestamps must be stated for MVTSO.

**Hypothesis 6** \(T_i.T_s < \text{parent}(T_k).T_s < T_j.T_s \land T'_j \notin \text{descendant}(\text{parent}(T_k)) \land \text{parent}(T_k) \notin \text{descendant}(T_i) \Rightarrow T_i.T_s < T_k.T_s < T_j.T_s\)

RO-transactions read their data from one or more W-transactions or T_0. For a given RO-transaction, T_\text{ro}, a set of versions exists associated with the read set RS_\text{ro}. In case of MVTSO, the maximum write-timestamp M_\text{ro} is defined as the minimum of all T_w for which the versions have a commit-timestamp T_e when there are such versions:

\[
M_{\text{ro}} = (\min x \in RS_{\text{ro}}, (x.T_e > T_{\text{ro}}, T_s \land x.S \neq \text{commit}) : x.T_w)
\]

Otherwise: \(M_{\text{ro}} = T_s\). The X-order, \(\prec\), can be extended for MVTSO, OCC-BV and 2PL.

**Definition 5 (Nested X-order)** For any two leaf-transactions T_i and T_j, T_i \(\prec\) T_j iff

- T_i and T_j are conflicting W-transactions with X \in RS_i \cap WS_j \lor X \in WS_i \cap RS_j and T_i.T_y < T_j.T_y, where T_y = T_e for OCC-BV \& 2PL and T_y = T_s for MVTSO.

- \((i = 0 \land X \in WS_i)\) or \((j = f \land X \in WS_i)\).

- T_i is a W-transaction and T_j is a RO-transaction and X \in WS_i \cap RS_j and for OCC-BV \& 2PL: T_i.T_e < T_j.T_s and for MVTSO: T_i.T_s < M_j.

- T_j is a W-transaction and T_i is a RO-transaction and X \in WS_j \cap RS_i and for OCC-BV \& 2PL: T_i.T_s < T_j.T_e and for MVTSO: M_i \leq T_j.T_s.

A different order is defined in the case of OCC-NVTI. The root of transaction T_i is denoted with T_i*.
Definition 6 (Nested-X-order2) For any two transactions \( T_i \) and \( T_j \), conflicting in \( X \), \( T_i \prec X T_j \) iff

- \( T_i \) and \( T_j \) are root-transactions and ordered according to definition 2
- \( T_i \) is a root-transaction and \( T_j \) is not a root-transactions and \( T_i^* \prec C_f T_j \)
- \( T_j \) is a root-transaction and \( T_i \) is not a root-transactions and \( T_i^* \prec C_f T_j \)
- \( T_j \) and \( T_i \) are not root-transactions and \( T_i^* \prec C_f T_j \)
- \( T_i \) and \( T_j \) have the same root and there is a transaction \( T_k \) with \( T_i, T_j \in \text{descendant}(T_k) \) with children \( T_m \) and \( T_i \) such that \( T_i = T_m \lor T_i \in \text{descendant}(T_m) \) and \( T_j = T_i \lor T_j \in \text{descendant}(T_i) \) and \( T_m \prec C_f T_i \).

The consistency of these relations is proven in the following lemma:

Lemma 8 For two related transactions \( T_i \) and \( T_j \), with \( \{X, Y\} \subseteq (RS_i \cap WS_j) \cup (RS_j \cap WS_i) \):

\( T_i \prec X T_j \Rightarrow T_i \prec Y T_j \)

Proof. When both transaction are W-transactions, they are ordered by the same time-stamps independent of the variable with which they are related. Consequently, the lemma is true for W-transactions.

The rest of the proof is identical to the proof of lemma 2.

Lemma 9 The order of definition 5 is acyclic.

Proof. The W-transactions are ordered by their uniquely attributed timestamps. The total ordering of the timestamps assures that the set of W-transactions is ordered without cycles. In case of 2PL and OCC-BV the ordering criterium of RO-transactions and W-transactions is the same (defined by \( T_i \)) and the union of both sets is also ordered without cycles. The case of MVTSO must be considered separately. No order is defined between RO-transactions, but RO-transactions are ordered with respect to W-transactions in two ways. Consequently, cycles can occur when a RO-transaction \( T_{ro} \) is ordered between two W-transactions, \( T_{w1} \) and \( T_{w2} \), such that \( T_{w1} \prec X T_{w2} \land T_{w2} \prec X T_{ro} \prec X T_{w1} \), as shown in fig 5.

From definition 2, the following relations can be derived:

\[ T_{w2} \prec T_{ro} \Rightarrow T_{w2} \prec T_{ro} \prec Y T_{s} \]

\[ T_{ro} \prec Y T_{w1} \Rightarrow Y T_{ro} \leq T_{w1} \prec T_{s} \]

Combining both relations leads to: \( T_{w1} \prec T_{ro} \prec T_{w2} \prec T_{s} \). Because \( Z \in RS_{w1} \cap WS_{w2} \lor (Z \in WS_{w1} \cap RS_{w2}) \), it follows that \( T_{w2} \prec Z T_{w1} \) and the cycle is impossible.

The same needs to be proven for the order definition 6 of OCC-NVTI:

Lemma 10 The order of definition 6 is acyclic.
such a version is committed by either a root transaction or a leaf transaction. It follows that
$v$ is committed by the writing leaf-transactions and one or more ancestors when $d(e)$ reads
$v$. According to the local clock of $X$, a reading leaf-transaction commits after the commit of
the leaf-transactions that wrote the version. According to hypothesis 5: $s(e).T_e < d(e).T_e$.

**MVTSO** A-Read proceeds by reading a version, $v$, with its state *commit* or by reading the
tentative version, $v$, owned by an ancestor of the reading transaction.

$$v.S = \text{commit} \lor v.T_o \in \text{ancestor}(T)$$  \hspace{1cm} (31)

The reading MVTSO leaf-transaction reads versions such that $s(e).T_e < d(e).T_e$, as prescribed
by the precondition of the A-Read action:

$$v.T_w \leq T_e$$  \hspace{1cm} (32)

with $v.T_w = s(e).T_e$ and $T_e = d(e).T_e$.

**OCC-BV** A-Read proceeds by reading a version with its state *commit* or by reading the
tentative version owned by an ancestor of the reading transaction.

$$v.S = \text{commit} \lor v.T_o \in \text{ancestor}(T)$$  \hspace{1cm} (33)

According to the post condition of A-Commit

$$(\text{root}(T) \Rightarrow (v.S = \text{commit})) \land (\neg \text{root}(T) \Rightarrow (v.T_o = \text{parent}(T)))$$  \hspace{1cm} (34)

such a version is committed by either a root transaction or a leaf transaction. According to
the local clock of $X$, the reading leaf-transaction commits after the commit of the leaf-
transactions that wrote the version. According to hypothesis 5: $s(e).T_e < d(e).T_e$.

**OCC-NVTI** Assume that leaf-transaction $T_i$ has written a version $v$ read by leaf-
transaction $T_j$. According to the precondition of A-Read

$$v \in X.V \land (v.S = \text{commit} \lor v.T_o \in \text{ancestor}(T_j))$$  \hspace{1cm} (35)

$T_j$ can read either a committed version or a version owned by an ancestor.

Assume the version is owned by an ancestor. According to the postcondition of A-Commit

$$\neg \text{root}(T_j) \Rightarrow v.T_o = \text{parent}(T_j)$$  \hspace{1cm} (36)

the ownership of $v$ passes to the parent of $T_i$. At the commit of the parent or ancestor of $T_i$,
ownership passes to another ancestor of $T_i$. When $T_j$ reads $v$, the owner $T_k$ of $v$ is both an
ancestor of $T_j$ and $T_i$. The postcondition of A-Read specifies that after the reading of $v$ by
$T_j$, an entry $T_j$ is added to $v.list$. According to the postcondition of A-Commit

**PRE:**

$$\forall v \in X.V : v.list = L_v$$  \hspace{1cm} (37)

$$P(x) \equiv (\neg \text{root}(T_i) \Rightarrow x.T_o = \text{parent}(T_i))$$  \hspace{1cm} (38)

**POST:**

$$\neg P(v) \Rightarrow v.list = (L_v \setminus T_j) \cup \text{parent}(T_j)$$  \hspace{1cm} (39)
Proof. The ordering between RO-transactions and W-transactions is already proven in lemma 3. Consider the case of ordered nested W-transactions with $T_i \prec T_j \prec T_k$. Assume that $T_i$ and $T_k$ conflict in $Z$. When all transactions are root-transactions: $T_i.C_f < T_j.C_f < T_k.C_f$.

According to definition 6: $T_i \prec Z T_k$.

Assume $T_i$ and $T_j$ have different roots: $T_i \neq T_j$. According to the nested order definition: $T_i.C_f < T_j.C_f$ and $T_j.C_f \leq T_k.C_f$. Combining both equations leads to $T_i.C_f < T_k.C_f$.

According to the order definition it follows that $T_i \prec Z T_k$.

Assume $T_i$ and $T_j$ have a common ancestor $T_l$; $T_j$ and $T_k$ have a common ancestor $T_m$. Then there are two children $T_l_1$ and $T_l_2$ of $T_l$ with $T_l_1 = T_l_1 \lor T_i \in \text{descendant}(T_l_1)$ and $T_l_2 \lor T_j \in \text{descendant}(T_l_2)$ and $T_l_1.C_f < T_l_2.C_f$. There are also two children $T_m_1$ and $T_m_2$ of $T_m$ with $T_j = T_m_1 \lor T_j \in \text{descendant}(T_m_1)$ and $T_k = T_m_2 \lor T_k \in \text{descendant}(T_m_2)$ and $T_m_1.C_f < T_m_2.C_f$. If $T_m_1 = T_l_2$ then it follows immediately that $T_i \prec Z T_k$. If $T_m_1 \neq T_l_2$ then $T_m \in \text{ancestor}(T_l) \lor T_i \in \text{ancestor}(T_m)$. Both cases are treated in the same way.

Consider $T_m \in \text{ancestor}(T_l)$. Then either $T_m_1 = T_l$ or $T_m_1 \in \text{ancestor}(T_l)$. From the definition of $T_l$ it follows that $T_m_1 \prec T_m_1$. From the order definition, $T_k \in \text{ancestor}(T_m_2)$ and $T_k.C_f \prec T_m_2.C_f$ it follows that $T_i \prec Z T_k$.

Based on order definitions 6 and 5 the view-serializability of the nested CCA's can be proven. It is first demonstrated that the history of leaf-transactions is serializable.

Lemma 11 For any labelled edge $e$ from $G$, with $s(e) = T_i$ and $d(e) = T_j$, $T_i$ and $T_j$ both leaf-transactions related via $X$, it holds that $T_i \prec X T_j$.

Proof. A RO-transaction $T_{ro}$ is the destination of a given edge $e$, or is the source of an edge with destination $T_f$. In the first case, the pre-condition in RO-Read and the assignment to the parameters, $T$ or $C$, of RO-Read in RO-Transaction:

\[
\begin{align*}
2PL & \land OCC-BV & \forall T \prec T \land \forall S = commit \land T = T_f \\
MVTSO & \forall T_w \prec T \land \forall S = commit \land T = M_{ro} \\
OCC-NVTI & \forall C_b \prec C \land \forall S = commit \land C = M_{ro}
\end{align*}
\]

assure that $s(e).T_e < d(e).T_s$ in case of OCC-BV & 2PL, $s(e).T_s < M_{ro}$ for MVTSO and $s(e).C_f < M_{ro}$ for OCC-NVTI. For all four CCA’s, $s(e)$ precedes $d(e)$ in the defined order.

In the second case, $T_f$ follows all other transactions per definition. Consequently, the lemma is true for all edges involving a RO-transaction. The same reasoning is valid for any edge involving $T_{ir} \in H^u$.

Edges which only involve W-transactions are considered for all four types of CCA separately.

\[2PL\ A-Read\ proceeds\ by\ reading\ the\ item\ that\ is\ free,\ read-locked\ or\ write-locked\ but\ owned\ by\ a\ parent:\\n\]

\[X.L.s \neq write \lor (X.L.s = write \land v.T_o \in \text{ancestor}(T))\]

According to the post condition of Commit

\[(\text{root}(T) \Rightarrow (v.S = commit)) \land (\neg \text{root}(T) \Rightarrow (v.T_o = \text{parent}(T))\]

35
ancestors may have locked the item when A-Write starts.

\[
X.L.s = free \lor X.L.l = \{T\} \lor \forall T_i \in X.L.l : T_i \in ancestor(T)
\]  

Consequently, no new versions of an item \(X\) can be created with A-Write as long as leaf-
transactions reading the same item have not yet committed. According to the postcondition
of A-Commit:

\[
\text{root}(T) \Rightarrow (X.L.l = \emptyset \land X.L.s = free)
\]

\[
\land \lnot \text{root}(T) \Rightarrow (X.L.l = (L - \{T\}) \cup \{parent(T)\})
\]  

The state of \(X\) is set to \(free\) or the reading leaf-transaction identifier is replaced by its
parent in the lock list. A new version of an item \(X\) can only be created by \(s(e')\) when the
transaction \(d(e)\) that reads an earlier version has committed. At the local clock of \(X\), \(d(e)\)
commits at time \(t\) before \(s(e')\) writes, after which \(s(e')\) commits at time \(t' > t\). After the
writing and commit by \(s(e')\), the precondition of A-Read assures that no transactions can
read the version written by \(s(e)\).

\[
\forall x \in X.V : v.T_c \geq x.T_c
\]  

Together with hypothesis 5: \(d(e) \prec s(e')\).

**OCC-BV** Suppose that \(d(e)\) is a W-transaction. The precondition of A-Validate and hy-
pothesis 2 prevent that a new version is written and committed by \(s(e')\) before the transaction
\(d(e)\) that reads an earlier version has committed.

\[
T_L = (\max x \in X.V, (x.S = commit \lor x.T_o \in ancestor(T)) : x.T_c)
\land (\forall x \in X.V : x.T_c \neq T_L \Rightarrow T \notin x.list)
\]  

According to the local clock of \(X\), \(s(e')\) writes after the validate/commit of \(d(e)\). From
hypothesis 5 follows: \(s(e').T_c > d(e).T_c\). After the commit by \(s(e')\) no transaction will read
the version written by \(s(e)\) as specified in the precondition of A-Read.

\[
\forall x \in X.V - \{v\} : (x.S = commit \lor x.T_o \in ancestor(T)) \Rightarrow v.T_c > x.T_c
\]  

Consequently, before and after the commit moment of \(s(e')\) and according to the order def-
nition: \(d(e) \prec s(e')\).

**OCC-BV & 2PL** Suppose \(d(e)\) is a RO-transaction. The RO-Read precondition

\[
v.T_c < T \land v.S = commit \land (\forall x \in X.V - \{v\}, x.S = commit : x.T_c < v.T_c \lor x.T_c > T)
\]  

assures that the committed version of \(X\) with the largest timestamp \(T_c\) is chosen. Assume
that this version was read at time \(t\) on the local clock of \(X\). Suppose the transaction \(s(e')\)
commits at time \(t' < t\) on the local clock of \(X\). The precondition of RO-Read assures that
\(s(e').T_c < d(e).T_c < s(e').T_c\). Suppose the transaction \(s(e')\) commits at time \(t' > t\) on the
local clock of \(X\). According to hypothesis 3, \(s(e')\) commits with timestamp \(s(e').T_c > d(e).T_c\).
According to the order definition 2 for OCC-BV and 2PL: \(d(e) \prec s(e')\).
the entry is changed to the parent of $T_j$. Define $T_i, T_m$ as children of the owning transaction $T_k$ with $T_j = T_i$ or $T_i$ is an ancestor of $T_j$ and $T_m = T_i$ or $T_m$ is an ancestor of $T_i$. The pre- and post-conditions A-Validate

**PRE:** \[ P(x) \equiv (\neg \text{root}(T_i) \Rightarrow x.T_o = \text{parent}(T_i)) \]  
**POST:** \[ (\forall v \in X.V : v.T_o \neq T_i) \Rightarrow C_f = (\max x \in X.V, P(x) \land T_i \in x.l) : x.C_b + 1) \] together with the $C_f$ calculation in the Validate PROCEDURE assure that $T_i.C_f > v.C_b$. The postcondition of A-Commit assures that the value of $v.C_b = T_m.C_f$. From the order definition 6, $T_m.C_f < T_i.C_f$, $T_k$ is the parent of both $T_m$ and $T_i$ and $T_j \in \text{ancestor}(T_i) \lor T_j = T_i$ and $T_i \in \text{ancestor}(T_m) \lor T_i = T_m$ follows: $T_i < T_j$

Assume that $T_j$ reads a committed version. The root of $T_i$ has committed because the version has the state commit. $T_j$ has not committed yet and has a different root: $T_j^* = T_j$. The same reasoning is valid as above. When the root of $T_j$ commits the $C_f$ value calculated for $T_j^*$ in the Validate PROCEDURE is larger than the $C_b$ value of $v$; $T_j^*.C_f = v.C_b > T_j^*.C_f$.

common From the order-definitions 6 and 5 it follows that the source of a labelled edge is ordered before its destination for all four CCA’s.

**Theorem 5 (Leaf serializability)** The graph $G'$ constructed from the ACTION specifications of section 8 is an acyclic $\text{TP}(h)$.

**Proof.** By construction, the vertex set of $G$ is equal to the vertex set of $G'$ and the edge set of $G'$ is equal to the edge set of $G$ with the addition of some unlabelled edges. An acyclic order is defined on all leaf-transactions $T_i$ and for all $X \in D$ there is a directed path in $G'$ that includes all $X$-edges. For two $X$-edges $e$ and $e'$ with different sources of which at least one is useful, their destinations are different, $d(e) \neq d(e')$, because a transaction reads a given value only from one transaction at the time. From lemma 11 it is known that $d(e) \succ s(e)$ and $d(e') \succ s(e')$.

Suppose that one of them, $e'$ is useless. The case $s(e) \prec s(e')$ is proven in the same way as the case that both edges are useful. When $s(e') \prec s(e)$, the destination of the useless write, $d(e')$, is defined to be ordered before the direct successor of the source: $d(e') \succ s(e)$.

Suppose both $e$ and $e'$ are useful edges. In that case $s(e)$ and $s(e')$ are both W-transactions because no labelled edges exist from a RO-transaction or a from a vertex $T_{ix} \in H^u$. Without loss of generality it is assumed that $s(e) \prec s(e')$. When $d(e) = s(e')$, the theorem is trivially true by taking the X-path from $T_0$ to $s(e)$ and via $d(e) = s(e')$ to $d(e')$ and finally to $T_j$. When $d(e) \neq s(e')$, the case that $d(e)$ is a W-transaction is treated first, followed by the case that $d(e)$ is a RO-transaction.

The four CCA’s are considered separately

**2PL** Suppose that $d(e)$ is a W-transaction. The postcondition of A-Read:

\[ X.L.l = (L \cup \{T\}) \land (LS = \text{free} \Rightarrow X.L.s = \text{read}) \] states that the lock-list, $X.L.l$, contains the identifier of the reading leaf-transaction and the state is unequal to free. The precondition of A-Write states that apart from itself only
Suppose that \(d(e)\) is a W-transaction. When \(s(e')\) has already written a version before \(d(e)\) reads, it follows for the precondition of A-Read that: \(s(e') \succ d(e)\). When \(s(e')\) writes a version after the reading by \(d(e)\) the precondition in A-Write states:

\[
\forall x \in X.V : (x.T_w < T_s \land x.T_r < T_s) \lor x.T_w > T_s
\]

Consequently, after the reading by \(d(e)\) only versions can be created by transaction \(s(e')\) if \(s(e').T_s < s(e).T_s \lor s(e').T_s > d(e).T_s > s(e).T_s\). According to \(s(e) \prec s(e')\) and the ordering definition: \(d(e) < s(e')\).

Suppose \(d(e)\) is a RO-transaction. The precondition of RO-Read

\[
v.T_w < T \land v.S = \text{commit} \land (\forall x \in X.V : (x.T_w \leq v.T_w \lor x.T_w > T_s))
\]

assures that \(s(e).T_s < M_{ro}\) and \(s(e').T_s > M_{ro}\) when \(s(e)\) and \(s(e')\) are committed before \(d(e)\) reads at time \(t\) on the local clock of \(X\). The postcondition of RO-Read

\[\text{Close}(v)\]

assures that transactions writing a version, \(x\), after the invocation of RO-Read cannot create this version with \(x.T_w < d(e).T_s\). Consequently, the value of \(M_{ro}\) is not modified by transactions which commit after the execution of RO-read. When \(s(e')\) commits at a time \(t' > t\) on the local clock of \(X\), hypothesis 3 states: \(s(e').T_s > d(e).T_s\). Together with \(s(e').T_s > M_{ro}\) it follows that: \(d(e) < s(e')\).

OCC-NVTI The case RO-Read has been proven in theorem 2

Suppose that \(d(e)\) is a W-transaction. From the assumption that \(s(e) \prec s(e')\) it follows that \(s(e).C_f < s(e').C_f\) and \(v.C_b < v'.C_b\). Suppose \(s(e')\) writes version \(v'\) which is an immediate successor of version \(v\) written by \(s(e)\).

Assume that \(s(e)\) and \(s(e')\) have a common ancestor \(T_k\), then \(d(e)\) must have the same common ancestor. If this is not the case \(d(e)\) has a root different from \(T_k\) or \(d(e)\) and \(T_k\) share a common ancestor. In both cases \(d(e)\) can only read the version written by \(s(e)\) after the commit of the ancestor of \(T_k\) or its root. According to the precondition of A-Read, \(d(e)\) would have read the version with the largest \(T_c\) value which is \(v'\). This contradicts the assumption that \(d(e)\) reads \(v\). Consequently, \(d(e)\) has the same ancestor \(T_k\) as \(s(e)\) and \(s(e')\).

Introduce three transactions \(T_l, T_m, T_n \in \text{child}(T_k)\), with \(T_l = s(e) \lor T_l \in \text{ancestor}(s(e))\) and \(T_m = d(e) \lor T_m \in \text{ancestor}(d(e))\) and \(T_n = s(e') \lor T_n \in \text{ancestor}(s(e'))\). After the reading of \(v\) by \(d(e)\), an entry \(d(e)\) is entered in \(v.list\). When \(T_m\) commits before \(T_n\) the pre- and post-conditions of A-Commit:

\[
\begin{align*}
\text{PRE:} & \quad \forall v \in X.V : v.C_r = C_v \land v.list = L_v \land P(v) \equiv (X.T_o = \text{parent}(T_i)) \\
\text{POST:} & \quad \forall v \in X.V, T_i \in L_v : P(v) \Rightarrow v.C_r = \max(C_v, C_f)
\end{align*}
\]

assures that \(v.C_r \geq T_m.C_f\). The postcondition of A-Validate and the calculation of the \(C_f\) value in Validate assures that after invocation of A-Validate by \(T_n\): \(T_n.C_f > v.C_r \geq T_m.C_f\). From the order definition it follows that: \(d(e) \prec s(e')\).

When \(T_n\) commits before \(T_m\), \(d(e)\) must already have read \(v\), otherwise \(d(e)\) would have read \(v'\). Because \(v'\) is an immediate successor of \(v\), the pre- and post-conditions of A-Commit:

\[
\begin{align*}
\text{PRE:} & \quad b \in X.V \land P(b) \land (\forall x \in X.V, P(x) : x.C_b \leq b.C_b) \\
& \quad \forall v \in X.V : v.T_o = T_{ov} \\
\text{POST:} & \quad (\exists x \in X.V : T_{ox} = T_i) \Rightarrow b.C_e = C_f
\end{align*}
\]
assure that $T_l \cdot C_f = b \cdot C_b < b \cdot C_e \leq T_n \cdot C_f$. The post-condition of $A$-Validate and the calculation of $C_f$ assures that $T_m \cdot C_f < T_m \cdot C_l \leq b \cdot C_e = T_n \cdot C_f$. From the order definition it follows that $d(e) < s(e')$.

When $s(e)$ and $s(e')$ have no common ancestors, their roots are different. Exactly the same reasoning applies as for the case that both have the same ancestor.

When $s(e')$ is not an immediate successor to $s(e)$ an immediate successor $T_x \prec s(e')$ can be found. Above it is proven that $d(e) \prec T_x$. Combining both equations leads to $d(e) \prec s(e')$.

From the order definition it follows that $d(e) \prec s(e')$. When $s(e')$ and $s(e')$ have no common ancestors, their roots are different. Exactly the same reasoning applies as for the case that both have the same ancestor.

When $s(e')$ is not an immediate successor to $s(e)$ an immediate successor $T_x \prec s(e')$ can be found. Above it is proven that $d(e) \prec T_x$. Combining both equations leads to $d(e) \prec s(e')$.

common From the ordering and the additional unlabelled edges added to $G'$, an acyclic path can be constructed started at $T_0$ from $s(e)$ to $d(e)$, from $d(e)$ to $s(e')$ and from $s(e')$ to $d(e')$ to end at $T_f$. Consequently the graph $G'$ constitutes an acyclic TP($h$).

It remains to prove that the intermediate transactions and the root transactions constitute a serializable history. In the graph $G$, the set $N$ of vertices representing transactions which are children of a given transaction can be collapsed to form a single vertex. The vertex set $N$ is replaced with one vertex $T_n^1$. All edges with both their source and their destination in $N$ are removed. All edges with their source in $N$ and their destination outside $N$ have their source replaced with $T_n^1$; all edges with their destination in $N$ and their sources outside $N$ have their destination replaced with $T_n^1$. The thus created graph is called $G^{(1)}$. All edges in the graph $G^{(1)}$ with the same source, destination and label are replaced by one single edge with an identical label. In the graph $G^{(1)}$, unlabelled edges are drawn which connect the nested transactions and the leaf-transaction as specified for the construction of $G'$ from $G$. This new graph is called $G^{(1)'}$. It then needs to be proven that this graph also represents a TP($h$). This is a called $L$-serializability in [Vid91].

One of the requirements on the nested transactions are pointed out in section 4: when a nested transaction $T_i$ precedes nested transaction $T_j$, then all descendants of $T_i$ also precede $T_j$ and all its descendants. The order number $n$ of $T_n^1$ is chosen such that $n = \max i : T_i \in \text{child}(T_n^1) : i$. This order is reflected by the value of the timestamps attributed by the OCC and MVTSO algorithms.

**Definition 7 (nested timestamps)** The start timestamp $T_s^x$ of an intermediate/root transaction at level $x$ is defined to be the minimum of the timestamps of all leaf-transactions which are descendants of this transaction. The end timestamp $T_e^x$ of an intermediate/root transaction at level $x$ is defined to be the maximum of the timestamps of all leaf-transactions which are descendants of this transaction.

\[
\begin{align*}
T_s^x &= \min T \in \text{descendant}(T_n^x) : T \cdot T_s \\
T_e^x &= \max T \in \text{descendant}(T_n^x) : T \cdot T_e
\end{align*}
\]

The transactions $T_{10}$ and $T_{11}$ of the fig. 11 are collapsed to $T_{11}'$. Graph $G'$ shown in fig. 6 is modified to graph $G^{(1)'}$ shown in fig. 19. For the remainder of this paper: $T_i^0 \equiv T_i$. The following theorem is used:

**Theorem 6 (L-serializability)** [Vid91] A history $h$ is $[L]$-serializable iff both $G^{(1)'}$ and the history of the collapsed transaction are an acyclic TP($h$).

It is easily demonstrated that the history of the collapsed transaction is serializable. Remove in $G$ all vertices and edges which have neither source nor destination in the collapsed...
transaction. Replace the source of the ingoing edges by $T_0$ and the destinations of the outgoing edges by $T_f$. View serializability (theorem 5) has already been proven for all edges with both vertices representing leaf-transactions of the collapsed transaction. For the other edges, the lemma is trivially true from the definition of the ordering of $T_0$ and $T_f$. The serializability of the history of the collapsed transaction follows. Below it is proven that $G^{(1)'}$ is an acyclic TP(h).

**Lemma 12** For any labelled edge $e$ from $G^{(1)}$, with $s(e) = T^x_i$ and $d(e) = T^y_j$, where one is a collapsed parent transaction and the other a leaf-transaction not a descendant of the first ($(x = 0 \land y = 1) \lor (x = 1 \land y = 0)$), it holds that $T^x_i < T^y_j$.

**Proof.** Consider the first alternative that $T^x_i = d(e)$ is a collapsed parent with children $T_k$. For every labelled edge $e_k$ with $s(e_k) = T_i$ and $d(e_k) = T_k$, lemmas 4 and 11 state that

---

**Figure 19:** Example of graph $G^{(1)'}$ with collapsed $T_{10}$ and $T_{11}$

**Figure 20:** Collapsing of either $s(e)$ or $d(e)$
$T_i \prec T_k$. All these edges, $e_k$ with $s(e_k) = T_i$ and $d(e_k) = T_k$, are replaced by the edges $e$ with $s(e) = T_i$ and $d(e) = T_j$. Edges with the same label and source are collapsed to one edge (see fig. 20). It is then required that for all $e_k$ with the label $X$, every $X$-writing $s(e)$ is ordered before all $X$-reading $T_k$.

**2PL** The collapsed transaction can belong to a different root from $T_i$ or to the same one. If the roots are different, the version of $T_i$ can only be read when the root has committed. The reading transactions, $T_k$, commits after the commit of $T_i$ and from hypothesis 5 it follows that the above condition is satisfied. When the collapsed transaction belongs to the same root, $T_i$ is not a descendant of $T_j$. All $T_k$ can only start reading when the version written by $T_i$ is owned by an ancestor of $T_k$. This has as consequence that an ancestor of $T_i$ has committed. $T_k$ can only commit after this ancestor and this one in turn can only commit after $T_i$. Consequently, $T_i$ is ordered before all $T_k$.

**OCC-BV** Similar to 2PL proof.

**MVTSO** $T_j \notin descendant(T_i)$ and $T_i.T_s < T_j.T_s = parent(T_k).T_s$ implies according to hypothesis 6 that: $T_i \prec T_k$.

**OCC-NVTI** Suppose that $s(e)$ and $d(e)$ have different roots $s(e)^* = T_i$ and $d(e)^* = T_j$. According to definition 6: $s(e)^*.C_f < d(e)^*.C_f$. All children $T_k$ of $d(e)$ have the same root. According to the order definition: $\forall T_k : s(e) \prec T_k$. When $s(e)$ and $d(e)$ have the same root, they have a common ancestor with two children $T_m$ and $T_i$ with $T_m.C_f < T_i.C_f$ and $s(e) = T_m \lor s(e) \in descendant(T_m)$ and $d(e) = T_i \lor d(e) \in descendant(T_i)$. The same reasoning as developed for $s(e)^*$ and $d(e)^*$ applies to $T_m$ and $T_i$, which leads to the same conclusion: $\forall T_k : s(e) \prec T_k$.

**common** Consequently, the lemma is satisfied for all four CCA's when the destination is the collapsed transaction.

Consider the second alternative that $T_i = s(e)$ is a collapsed parent with children $T_k$. For every labelled edge $e_k$ with $d(e_k) = T_j$ and $s(e_k) = T_k$, lemma 4 states that $T_k \prec T_j$. Again, edges with the same label $X$ and the same destination $d(e)$ are collapsed. Consequently, for all $e_k$ with the same label $X$, all $T_k$ must be ordered before every $d(e)$.

**2PL** When $T_i$ and $T_j$ have different roots, all leaf transactions $T_k$ must have committed before $T_j$ can read them. $T_j$ commits after $T_i$ and (hypothesis 5) each $T_j$ is ordered after all $T_k$.

Suppose the collapsed transaction and the reading transaction $T_j$ have the same root. The version written by one of the descendants of $s(e)$ has the state write when it is read by $d(e)$. According to the precondition of A-Read:

\[ X.L.s = write \land v.T_o \in ancestor(T) \quad (58) \]

the leaf-transaction $T_j$ can only read the versions of the children of $T_i$ when an ancestor of $T_j$ is owner of these versions. This happens after the commit of $T_i$ which only occurs after the last commit of all children $T_k$ of $T_i$. Consequently, $T_j$ can only commit after the last commit of $T_k$.

42
When $T_i^l$ and $T_j$ have different roots, all leaf transactions $T_k$ must have committed before $T_j$ can read them. $T_j$ commits after $T_i^l$ and (hypothesis 5) each $T_j$ is ordered after all $T_k$.

Suppose the collapsed transaction and the reading transaction $T_j$ have the same root. A leaf-transaction can only read a version that is committed or a tentative version that has a common ancestor. The latter implies that $T_i^l$ has committed and consequently all its children $T_k$ have committed before the versions can be read. The reading transactions $T_j$ commits after $T_i^l$ and it follows that $T_k < T_j$ for all $k$.

The graph $G^{(1)'}$ constructed from $G^{(1)}$ where a set of leaf-transactions is replaced by their parent, is an acyclic TP(h).

**Proof.** The part of the proof that concerns RO-transactions and useless writes, is identical to the proof of theorem 2. Consider two X-edges $e$ and $e'$ where both sources and destinations are W-transactions. When all edges are simple leaf-transactions, theorem 2 applies. The cases that one of the four vertices are collapsed parents is of interest. Without loss of generality take $s(e) < s(e')$. When $s(e)$ or $d(e')$ is a collapsed parent the theorem is trivially true by applying theorem 2 and lemma 12. Two cases are left: (1) $s(e')$ is a collapsed parent and $d(e)$ is a leaf-transaction and (2) $d(e)$ is a collapsed parent and $s(e')$ is a leaf-transaction.

**Ad 1)** When $d(e)$ is a child of $s(e')$, $d(e) = s(e')$ after the collapse of $d(e)$. Applying the lemma 12 twice proves the lemma. Take the case that $d(e)$ is not a child of $s(e')$. It must be proven that $d(e) < s(e')$. This is equivalent with proving that all children $T_k$ of $s(e')$ obey $d(e) < T_k$ for every $d(e)$.

2PL $d(e)$ reads before any descendant of $s(e')$ commits, otherwise $d(e)$ would have read a version written by a descendant of $s(e')$. A child of $s(e')$ can only write a version when the lock is held by its ancestors or is free. This means that either $d(e)$ has committed and the lock is passed to a common ancestor of $d(e)$ and $s(e')$, or that the lock is released and $d(e)$ is the descendant of a root different from the root of $s(e')$. In both cases $d(e)$ has committed before all writing children of $s(e')$.

The precondition on A-Commit

$$\forall x \in X.V - C : (x.T_c < (\min v \in C : v.T_c)) \lor (x.T_c > (\max v \in C : v.T_c))$$  (59)

assures that all commit by children of $s(e')$ occur either before or after the commit of $s(e)$. From the assumption that $s(e) < s(e')$ it follows that all children $T_k$ of $s(e')$ commit after $s(e)$. The pre-condition of A-Validate

$$T_L = (\max x \in X.V, (x.S = commit \lor x.T_c \in ancestor(T)) : x.T_c)$$

\[
\land (\forall z \in X.V : z.T_c \neq T_L \Rightarrow T \notin x.list)
\]

assures that all commit by children of $s(e')$ occur either before or after the commit of $s(e)$.

From the assumption that $s(e) < s(e')$ it follows that all children $T_k$ of $s(e')$ commit after $s(e)$. The pre-condition of A-Validate

$$T_L = (\max x \in X.V, (x.S = commit \lor x.T_c \in ancestor(T)) : x.T_c)$$

\[
\land (\forall z \in X.V : z.T_c \neq T_L \Rightarrow T \notin x.list)
\]
assures that all children of \( s(e') \) commit after the the commit of \( d(e) \). From hypothesis 5 it follows that:

\[
\forall T_k, T_k \in \text{child}(s(e')) : d(e).T_e < T_k.T_e
\]

Consequently, \( s(e) < d(e) < s(e') \).

**MVTSO** The application of hypothesis 6 validates the lemma.

**OCC-NVTI** The same reasoning as used in lemma 12 applies.

**Ad 2)** \( d(e) \) is a collapsed parent.

**2PL** The precondition of A-Write states that all locking transactions are ancestors of \( s(e') \).

\[
\forall T_i \in X.L.1 : T_i \in \text{ancestor}(T)
\]

Consequently, all leaf-transactions \( T_k \) of \( d(e) \) must have committed before the write can start. Because \( d(e) \) commits later than the committed leaf-transactions, it follows that for all \( T_k \), \( T_k < s(e') \).

**OCC-BV** The precondition on A-Validate

\[
T_L = (\max x \in X.V, (x.S = \text{commit} \lor x.T_e \in \text{ancestor}(T)) : x.T_e) \\
\land (\forall x \in X.V : x.T_e \neq T_L \Rightarrow T \notin x.\text{list})
\]

assures that no new version can be added between the commit of the reading transactions and the commit of the writing transaction \( s(e) \). Consequently, all children of \( d(e) \) have committed before the new version of \( s(e') \) can be installed. Together with hypothesis 5, this implies that \( s(e') > d(e) \).

**MVTSO** Hypothesis 6 assures that \( d(e) < s(e') \).

**OCC-NVTI** The same reasoning as applied in lemma 12 applies.

**common** In all cases it is true that \( s(e) < d(e) \) and \( s(e') < d(e') \) and \( d(e) < s(e') \lor d(e) = s(e') \), which proves the lemma.

Transactions can be collapsed to their parents and in their turn these parents can be collapsed to their parents. Root-transactions cannot further be collapsed. A completely collapsed history is the history of the root-transactions. The set which consists of history \( h \) of leaf-transactions and all collapsed histories derived from \( h \) is called the collapsed history set, \( CH(h) \).

**Theorem 7** [Nested Serializability] All collapsed histories \( ch \in CH(h) \) with \( h \) a history governed by the ACTION specifications of section 4 are serializable histories.
Proof. For theorem 5 it is proven that \( h \) is a serializable history. In lemma 13, it is proven that a singly collapsed history \( h^{(1)} \) is a serializable history because the derived graph \( G^{(1)} \) is a acyclic TP(h). Remove in \( G \) for all \( X \in D \) all write and read actions on \( X \) where W[X] precedes R[X] in the collapsed parent and replace them with equivalent write and read actions on internal variables of the collapsed parent. The modified collapsed parent has now all the characteristics of a leaf-transaction. \( G^{(1)} \) can be considered a graph of leaf-transactions.

Consider a given graph \( G^{(k+1)} \) recursively constructed by collapsing a parent of the graph constructed in \( G^{(k)} \). Assume that \( G^{(k)} \) consists of leaf-transactions and that a graph \( G^{(k)}_1 \) can be constructed which is a TP(h). Collapsing a parent in \( G^{(k)} \) creates the graph \( G^{(k+1)} \). The graph \( G^{(k+1)}_1 \) can then be proven to be a TP(h) as shown by lemma 13. In the graph \( G^{(k+1)} \), the single collapsed transaction can be replaced by an equivalent leaf-transaction as shown for \( G^{(1)} \).

Theorem 8 (Exception atomicity) The versions created by a transaction are visible to all following transactions or to none.

Proof. A version’s state can be committed or tentative. When the version’s state is committed, the version cannot be removed and is consequently visible to all transactions following the committing transaction. According to all A-Read preconditions

\[
v.S = \text{commit} \lor v.T \in \text{ancestor}(T)
\]

versions with the state tentative are only visible to transactions which are a descendant of the owning transaction. Other transactions see the version only after its commit. According to the A-Abort pre- and post-conditions,

\[
\begin{align*}
\text{PRE:} & \quad X.V = Q \land C = \{v \mid v \in X.V \land v.T = T\} \\
\text{POST:} & \quad X.V = Q - C
\end{align*}
\]

versions are removed when the owning transaction is aborted. The owning transaction owns the versions written by all descendants as specified by the postcondition of A-Commit:

\[
\neg \text{root}(T) \Rightarrow v.T = \text{parent}(T)
\]

When the owning transaction aborts, the results of all its descendants are removed as well. The removed version was only read by aborted descendants of the owning transaction. Consequently, no transaction has had access to the version under consideration.

9.1 Proof of HRT/SRT serializability

The hrt-order is already defined for OCC-VTI in section 7. The hrt order can be defined for 2PL, OCC-BV and MVTSO:

Definition 8 (hrt-order extension) For any two transactions \( T_i \) and \( T_j \), \( T_i \preceq T_j \) iff

- \( T_i \) and \( T_j \) are leaf-transactions of the same class and ordered according to definitions 5 and 2.

45
• $T_i$ writes $X \in HO \cup SO$ and $T_j = T_{id}$.

• $T_j$ accesses $X \in HO \cup SO$ and $T_i$ is a specification transaction and $T_i.T_y < T_j.T_y$, with $T_y = T_s$ for MVTSO and $T_y = T_e$ for 2PL & OCC-BV.

• $T_i$ accesses $X \in HO \cup SO$ and $T_j$ is a specification transaction and $T_i.T_y < T_j.T_y$, with $T_y = T_s$ for MVTSO and $T_y = T_e$ for 2PL & OCC-BV.

It is important to note that the transfer-serializability concept is independent of the choice of CCA. The SRT-transactions may use a different CCA from the HRT-transactions.

Lemma 14 The order of definition 8 is acyclic.

Proof. Consider the ordering without the specification transactions. Two mutually unordered sets of HRT and SRT transactions exist. According to theorem 7 each individual set is ordered. The addition of the specification transactions can lead to cycles in two ways: (1) when a pair of HRT(SRT)-transactions are differently ordered with respect to each other than with respect to a given specification transaction and (2) when a cyclic ordering can be found from a specification transaction $T_{id}$ via a set of HRT(SRT) transactions to another specification transaction $T_{jd} \neq T_{id}$ and from $T_{jd}$ via a set of SRT(HRT) transactions back to $T_{id}$.

ad 1) Consider that the following order is valid: $T_j.X' < T_{id} < T_k$, with $T_j$ and $T_k$ HRT(SRT)-transactions of the same class. In case of MVTSO, $T_j.T_s < T_{id}.T_s < T_k.T_s$ implies that for conflicting $T_j$ and $T_k$: $T_j < T_k$. In case of OCC-BV and 2PL, $T_j.T_e < T_{id}.T_e < T_k.T_e$ implies that for conflicting $T_j$ and $T_k$: $T_j < T_k$. Consequently, no cycle is possible.

ad 2) Without loss of generality, assume that a chain of HRT(SRT)-transactions is ordered after the specification transaction $T_{id}$ and before the specification transactions $T_{jd}$. According to the definition 4: $T_{id}.T_y < T_{jd}.T_y$ with $T_y = T_s$ for MVTSO and $T_y = T_e$ for OCC-BV and 2PL. To create a cycle, a chain of SRT(HRT)-transactions is ordered before the specification transaction $T_{id}$ and after the specification transactions $T_{jd}$. According to the definition 4: $T_{id}.T_y > T_{jd}.T_y$ with $T_y = T_s$ for MVTSO and $T_y = T_e$ for OCC-BV and 2PL. This constitutes a contradiction. Consequently, the order is acyclic. □

It is proven that the graph $G'_{hrt}$ represents a transfer-serializable history for the leaf-transactions which constitute $h$.

Lemma 15 For any $Y$-labelled edge $e$ from $G_{hrt}$, with $Y = X$ or $Y = X'$, $s(e)$ and $d(e)$ both leaf-transactions, it holds that $s(e) \prec d(e)$.

Proof. When $s(e)$ and $d(e)$ are both HRT- or SRT-transactions, lemma 12 applies and $s(e) \prec d(e)$.

One of the two can be a specification transaction. When $d(e)$ is a specification transaction, $s(e) \prec d(e)$ is determined by the hrt-order extension definition. When $s(e)$ is a specification transaction, there exists a $T_i$ such that $s(e) = T_{id}$. When $T_i$ is a HRT(SRT)-transactions then $d(e)$ is a SRT(HRT) transactions. Consequently, they cannot have a common ancestor.

In case of 2PL and OCC-BV, $d(e)$ only reads $X'$ after the commit of $T_i$ at local time $t$. According to the precondition of TR-Read, the commit of $d(e)$ at local time $t' > t$ and hypothesis 5: $d(e).T_e < T_i.T_e$. According to the definition of specification transaction timestamps: $s(e).T_e = T_i.T_e$. Combination of both conditions leads to the conclusion that $s(e) X' \prec d(e)$. 

46
In case of MVTSO, \(d(e)\) reads \(X'\) after the commit of \(T_i\). From the precondition of TR-Read follows: \(d(e).T_s > T_i.T_e\) and from the definition of the specification transaction follows: \(s(e).T_s = T_i.T_e\). Their combination leads to the conclusion that \(s(e) \prec d(e)\). Consequently, combining all cases, it follows that \(s(e) \prec d(e)\).

Theorem 9 (Transfer-serializability) The graph \(G_{\text{ht}}\) constructed from the ACTION specifications of section 4 and transfer-edges is an acyclic TP(h).

Proof. For two X-edges \(e\) and \(e'\) with different sources of which at least one is useful, say \(e'\), their destinations are different, \(d(e) \neq d(e')\), because a transaction reads a given value only from one transaction at the time. Without loss of generality it is assumed that \(s(e) \prec s(e')\). From lemma 7 it is known that \(d(e) \succ s(e)\) and \(d(e') \succ s(e')\). By construction and by the definition of the transfer-edges, HRT transactions access other data-items than SRT-transactions and are not related. The identical labels of \(e\) and \(e'\) imply that when one of the four transactions is a HRT(SRT) transaction then the other transactions are either HRT(SRT) transactions or specification transactions.

When \(s(e), d(e), s(e')\) and \(d(e')\) are all HRT-transactions or SRT-transactions, theorem 7 applies.

When \(s(e)\) is a specification transaction, \(s(e')\) is necessarily also a specification transaction because both edges, \(e\) and \(e'\) are transfer-edges labeled with the same label \(X'\) and only specification transactions are the sources of transfer-edges. Consequently, only two cases need to be considered: (1) \(d(e)\) is a specification transaction and (2) \(s(e')\) is a specification transaction.

Ad (1): According to the definition of transfer transactions, \(d(e)\) is ordered before the immediate successor of \(s(e)\). \(s(e')\) belongs to the same class as \(s(e)\) (HRT or SRT) and is the immediate successor of \(s(e)\) or \(s(e')\) is ordered after the immediate successor and therefore: \(d(e) \prec s(e')\).

Ad (2): There exists a transaction \(T_i\) corresponding with the specification transaction \(T_{id} = s(e)\) and a transaction \(T_j\) corresponding with transaction \(T_{jd} = s(e')\). The SRT(HRT)-transaction \(d(e)\) has a different root from the HRT(SRT)-transactions \(T_i\) and \(T_j\). \(T_i\) has a different root from \(T_j\) as \(d(e)\) considers them as two separate transactions.

In case of OCC-BV and 2PL, the ordering \(s(e) \prec s(e')\) implies that \(T_i.T_e < T_j.T_e\). Assume that TR-Read is invoked at local time \(t\). \(T_i\) has committed before TR-Read is invoked, otherwise \(d(e)\) could not have read the value written by \(T_i\). Two cases must be discerned: (1) \(T_j\) commits before \(t\) and (2) \(T_j\) commits after \(t\). In case 1 the precondition of TR-Read stipulates that \(T_i.T_e < d(e).T_s < T_j.T_e\). In case 2 hypothesis 4 assures also: \(T_i.T_e < d(e).T_s < T_j.T_e\). Using the definition of specification transaction timestamps, the substitution \(s(e').T_s = T_j.T_e\) yields: \(d(e).T_s < s(e').T_e\). Which leads to: \(s(e) < d(e) \prec s(e') \prec d(e')\).

In case of MVTSO, the precondition of TR-Read implies: \(s(e).T_s = T_i.T_e < d(e).T_s < T_j.T_e\) when \(T_j\) commits before \(d(e)\) reads. Hypothesis 4 assures that the same is true when \(T_j\) commits after the read-action of \(d(e)\). Further, the definition of the specification transaction timestamp implies: \(s(e').T_s = T_j.T_e\). Combining the two equations above yields: \(s(e).T_s < d(e).T_s < s(e').T_s\).

A path can be constructed from \(T_0\) via the total order to \(s(e)\), from \(s(e)\) via \(d(e)\) and \(s(e')\) to \(d(e)\) and from \(d(e)\) via the acyclic order to \(T_j\). Which concludes the proof of the theorem.  

\[\square\]
10 Implementation

10.1 Timestamp creation

During run-time, transactions are identified with a timestamp.

\[
\text{PROCEDURE TimeStamp( } T : \text{ Timestamp) : TimeStamp} \\
\text{BEGIN} \\
\quad \text{IF } T = 0 \text{ THEN TimeStamp := time } \circ \text{ PID} \\
\quad \text{ELSE TimeStamp := } T \circ \text{ OID} \\
\text{END}
\]

Figure 21: Creation of Timestamps

In Fig. 21, the creation of a timestamp is shown. A root-transaction invokes TimeStamp with \( T = 0 \) as parameter. The timestamp is returned as the local time suffixed with the Processor IDentifier (PID). This establishes unique timestamps. Hypothesis 6 requires that when root transaction \( T_1 \) precedes \( T_2 \), all descendants of \( T_1 \) precede \( T_2 \) and precede all descendants of \( T_2 \). When the timestamp is generated for an intermediate transaction, the timestamp of the parent-transaction suffixed with the Object identifier OID is returned. It is assumed that (possibly in time) unique OID can be generated. These timestamps allow to check if a given timestamp \( T_i \) is the ancestor of another transaction represented with timestamp \( T_j \). This relation is recursively represented with:

\[
T_i \in \text{ancestor}(T_j) \iff \begin{cases} 
T_j = T_i \circ \text{ OID} \\
T_j = T_k \circ \text{ OID} \land T_k \in \text{ancestor}(T_i)
\end{cases} \quad (67)
\]

A relation is defined for any two timestamps \( T_i \) and \( T_j \) where \( K_i \) and \( K_j \) are integer values:

\[
T_i < T_j \iff \begin{cases} 
T_j = T_i \circ T' \\
T_i = K \circ T'_i \land T_j = K \circ T'_j \land T'_i < T'_j \\
T_i = K_i \circ T \land T_j = K_j \circ T' \land K_i < K_j
\end{cases} \quad (a,b,c) \quad (68)
\]

The above relation for the timestamps satisfies the hypothesis 6 for MVTSO discussed in section 6.

**Lemma 16** \( (T_i,T_s < \text{parent}(T_k).T_s < T_j.T_s \land T_j \notin \text{descendant(parent(T_k)))} \land \text{parent(T_k)} \notin \text{descendant(T_i)}) \Rightarrow T_i.T_s < T_k.T_s < T_j.T_s \)

**Proof.** From \( T_i.T_s < \text{parent}(T_k).T_s \) and equations 67 and 68 follows that

\[
T_i \notin \text{descendant(parent}(T_k)) \quad (69)
\]

Similarly it follows that \( \text{parent}(T_k) \notin \text{descendant}(T_j) \). The line (a) of equation 68 which defines the order for descendants does not apply:

\[
T_j.T_s \neq \text{parent}(T_k).T_s \circ T \\
\text{parent}(T_k).T_s \neq T_i.T_s \circ T \\
\text{parent}(T_k).T_s \neq T_i.T_s \circ T \
\]

The lines (b) and (c) of equation 68 apply. This means that the timestamps of \( T_i \) and \( \text{parent}(T_k) \) consist of an equal number \( (n > 0) \) of identical integers followed by integer values.
\[ \text{parent}(T_k).T_s = T \circ K_k \circ T_s < T \circ K_j \circ T_j = T_j.T_s \]
\[ T_i.T_s = T' \circ K_i \circ T_i < T' \circ K'_k \circ T'_k = \text{parent}(T_k).T_s \]

where \( T, T', T_i, T_j, T'_k \) and \( T_k \) are possibly empty sequences of integers and \( K_i \) and \( K_j \) are integer values. According to the construction of the timestamp shown in fig. 21:

\[ (\text{parent}(T_k).T_s) \circ A_k = T_k.T_s \] (72)

Inserting in equation 71 the result from eq. :68, \( T_i < T_i \circ A_k \) yields:

\[ \text{parent}(T_k).T_s = T \circ K_k \circ T_k < T \circ K_j \circ T_j < T \circ K'_j \circ T'_j \circ A_k = T_k.T_s \] (73)

Substituting equations 72 and 71 the final result is obtained:

\[ \text{parent}(T_k).T_s < T_k.T_s < T_j.T_s \]
\[ T_i.T_s < \text{parent}(T_k).T_s < T_k.T_s \] (74)

Their combination and removal of \( \text{parent}(T_k).T_s \) leads to the wanted result.

### 10.2 System model

In this section the underlying assumptions and services are discussed. It is assumed that the underlying system is composed of processors interconnected via links. Processors are fail-silent: they behave according to specification until the moment of failure after which they perform no detectable actions. The following services are available.

**Multicast service** delivers messages to a set of processes with the following characteristics:

- **Integrity** A message received by a receiver is sent by a sender.
- **Order** Two HRT commit messages \( m \) and \( m' \) sent at local times \( t_s \) and \( t'_s \), with \( t_s < t'_s \), will be received at local times \( t_r < t'_r \).
- **Validity** Messages from a correct sender are received by all correct receivers.
- **Unanimity** A message is received by all correct receivers or by none.
- **Timeliness** A message sent by a HRT sender arrives within bounded time at all correct HRT receivers.

**Clock synchronization service** assures that correct processors have local clocks which are approximately synchronized with a known maximum deviation \( \epsilon \). Clock values read at different instants are monotonically increasing. \( C_p(t) \) denotes for each processor \( p \) its clock value at real time \( t \). For any two processors \( p \) and \( q \), it holds that \( |C_p(t) - C_q(t)| < \epsilon \).

**Online HRT scheduler** maintains a list of executions, processors, start-times and deadlines, that constitutes the HRT schedule. The online scheduler starts and stops executions to assure that the executions obey the online schedule. Processors are replicated and a replicated group is assumed not to fail such that all scheduled actions can be successfully completed.
Communication delay

The communication duration $\Delta$ of a multicast and the communication duration $\delta$ of a point to point message is related to the clock deviations. When a message is sent (multicast) by sender, $p$, at local clock time $C_p(t)$ and received by receiver, $q$, at local clock time $C_q(t')$ then the following relation holds:

$0 < C_q(t') - C_p(t) \leq \Delta(\delta)$.

Hypothesis 5 can be restated as lemma 17.

Lemma 17 In case of 2PL, OCC-NVT and OCC-BV, two RW-conflicting transactions $T$ and $T'$, with commit-timestamps $T_e$ and $T'_e$ execute their commit-actions at local times $C_q(t)$ and $C_q(t')$ such that: $C_q(t) > C_q(t') \Leftrightarrow T_e > T'_e$.

Proof. Suppose two RW-conflicting transaction $T$ and $T'$ exist.

2PL Without loss of generality, $T'$ has locked the conflicting item on processor $q$ before $T$. The precondition of $T$ is only fulfilled after the liberation of the data-item by $T'$. The execution by $T'$ of A-Commit on processor $q$ at local time $C_q(t')$ liberates the item. The commits of $T$ and $T'$ will be initiated by the coordinator $c'$ at a local time $C_{c'}(t'_c)$ and at coordinator $c$ at a local time $C_c(t_c)$. Due to the clock synchronization and the communication delay, $c'$ sent this message at local time $C_{c'}(t'_c) \geq C_q(t)$. The value of the data-item will be read by $T$ at a local time $C_q(t_s) > C_q(t')$. Due to the clock synchronization and the communication delay, the message will be received by $c$ at a local time $C_c(t_c) > C_q(t_s)$. This means that

$$C_q(t) > C_q(t') \Rightarrow C_{c'}(t'_c) < C_c(t_c).$$

Inversely, when $C_{c'}(t'_c) < C_c(t_c)$ the liberation of the item occurs before the the start of the commit: $C_{c'}(t'_c) > C_q(t)$. Using $C_q(t) > C_c(t_c)$ leads to:

$$C_q(t) > C_q(t') \Leftrightarrow C_{c'}(t'_c) < C_c(t_c).$$

According to the construction of the timestamps at local times $C_{c'}(t'_c)$ and $C_c(t_c)$, the accompanying commit timestamps will obey: $C_{c'}(t'_c) < C_c(t_c) \Leftrightarrow TS(C_{c'}(t'_c)) < TS(C_c(t_c))$ and $TS(C_{c'}(t'_c)) < TS(C_c(t_c)) \Leftrightarrow T'_e < T_e$. Combining the two last equations with equations 75 and 76 yields the wanted result.

OCC-BV & OCC-NVTI The validate/commit action pair of two transactions that write to the same item $X$ is done in a strictly serial order because only one transaction at the time can add a new version with state validating to the set of versions of $X$. The minimum difference between the timestamps, $T_e, T'_e$ of two consecutively created versions is determined by the maximum deviations between the local clocks and the minimum communication delay $\delta$. The difference between local time $t$ of the creation of the timestamp $T_e$ and the local time $t'$ of the creation of the next timestamp $T'_e$ is determined by the sending of a message to the coordinator and the sending of the commit message by the coordinator: $t' - t > 2.\delta$. The chosen timestamp $T_e$ is taken from the maxima of the returned timestamps such that $T_e = timestamp(t_e) \wedge t_e - t < \epsilon < \delta$. For a minimum difference between $T_e$ and $T'_e$: $T_e = timestamp(t')$. This has as consequence that for any two consecutively created version with $T_e = timestamp(t)$ and $T'_e = timestamp(t')$ that $t' - t > \delta$. 

50
Two cases must be proven to be impossible: (1) a writing transaction commits after the commit of a reading transaction with a larger commit-timestamp and (2) a reading transaction commits after the commit of a writing transaction with a larger commit-timestamp.

Ad 1) The timestamps of the reading transaction $T_r$ and the writing transaction $T_w$, with $T_w > T_r$, are generated from local times $t_r$ and $t_w$ respectively. In this case the reading transaction has read an earlier version with timestamp $T_o$ generated at local time $t_e$. Due to the precondition of A-Validate the writing transaction with the larger commit timestamp invokes A-Validate after the invocation of A-Validate by the reading transaction. In this case $t_r - t_c < \delta$ and $t_w - t_c > \delta$ as stated by the lemma.

Ad 2) A reading transaction that commits with timestamp $T_r$ after another writing transaction with commit timestamp $T_w$ has either read the version written by this transaction or has invoked A-Validate before the writing transaction. In the first case the creation of the commit timestamp assures by construction that $T_r > T_w$. The second case is similar to the one treated in ad 1).

10.3 Commit algorithms

The correctness of the specified transactions strongly depends on a timely commit implementation. During a commit all versions created by the transaction under consideration are either installed and committed or removed. It is important that all related versions reach the same state after the commit. In addition, the commit should be executed in bounded time. The here presented commit algorithms depend on the above presented services. It is assumed that there is one coordinator process which initiates the transaction and at the end of the transaction also initiates the commit. By construction only one coordinator exists and when no faults occur, a coordinator will be present. When needed, the coordinator waits a fixed time on answers from the participants in the commit protocol. When this time has elapsed without response from some participant, the coordinator will assume that a failure occurred. Participants will also wait for the coordinator's response on their answer to the coordinator. When this time has elapsed, participants will assume a failure occurred and will send an "abort" message to all participants. This way the non-blocking property of the commit protocol is acquired.

The requirements on the Commit are:

**Unanimity** All participants that decide reach the same decision.

**Integrity** Only when all (correct) participants agree to commit, can a participant commit.

**Validity** When all participants agree to commit and no failures occur, then all participants commit.

**Ordering** Two transactions $T$ and $T'$ that execute A-Commit at local times $C_p(t)$ and $C_p(t')$ with commit timestamps, $T_c$ and $T'c$ such that $C_p(t) < C_p(t') \Rightarrow T_c < T'_c$.

**Timeliness** When the coordinator starts the commit at time $C_c(t)$ at its local clock, then there is a constant $d$ such that all correct participants have either committed or aborted at a local time $C_p(t') = C_c(t) + d$.

In addition, hypotheses 2, 3 and 4 must be met.
Timely behavior is solved in three ways: (1) the HRT transactions commit according to a schedule and their commit times are known, (2) the commit of a finite number of SW-objects is done in finite time by reserving timeslots in the HRT-schedule and (3) the commit of SRT-objects is scheduled in the time slots left by the HRT transactions and no guarantee can be given for the timeliness requirement of the Commit. After the execution of the Commit procedure, all participants will have set the same value of $T_e$ in the versions associated with the transaction.

10.3.1 HRT Commit

The example of OCC-NVTI is taken. In Fig. 22 pseudo code for the implementation of a HRT-Commit procedure is indicated. The Commit procedure for the other CCA’s is similar but for the specified A-Commit parameters. At every processor one process executes the Procedure Wait-Mess. This process permanently listens for Commit messages sent by coordinators. When a Commit message is received, the Commit is executed for all data-items owned by the committing transaction.

```
1 PROCEDURE HRT-Commit( N: set transactions, $T_s,T_e$: Timestamp, $C_f$: Counter)
2 BEGIN
3 $T_e :=$ Timestamp(time + delay)
4 Multicast($N \oplus T_e \oplus C_f$, participants)
5 END

6 PROCEDURE Wait-Mess()
7 VAR
8 $D_p$: data-items on processor $p$
9 BEGIN
10 WHILE TRUE DO
11 ON reception of "N \oplus T_e \oplus C_f"
12 $\forall t \in N : \forall X \in T,I \cap D_p$: A-Commit(X, T, T_e, T_s, C_f)
13 OD
14 END
```

Figure 22: HRT-Commit algorithm

It is assumed that HRT transactions are scheduled such that their interleaving always generates view-serializable schedules. When a component of a HRT system fails, often the schedule cannot be met and whole system shuts down or fails. By rendering the components very reliable the probability of such an event is made acceptably small. A hypothesis must be formulated about the consequences of a component failure.

**Hypothesis 7** When a HRT component fails, the whole system fails.

The correctness of the protocol of Fig. 22 is rather trivially proven.

**Theorem 10** The procedures HRT-Commit and Wait-Mess together satisfy the five commit requirements and hypotheses 2 and 3.

**Proof.** Unanimity, Integrity, Validity, Ordering and Timeliness: they are based on Multicast properties and availability of coordinator and participants.

52
Hypothesis 2: this is assured by the HRT schedule. The schedule should be constructed such that A-Validate always succeeds. In that case, the Validate Procedure is empty and the atomicity of A-Validate, A-Commit action pairs is assured.

Hypothesis 3: from the timeliness property it is known that the commit protocol is bounded. Equate the constant \( \Delta_e \) with the time from the determination of the timestamp to the termination of the last Commit action. Take the "delay" in the setting at line 2 of \( T_e \) equal to \( \Delta_e \). The timestamp, \( T_e = C_p(t_e) \circ p \) of transaction \( T_{ro} \) is generated at local time \( C_p(t_e) \) on processor \( p \). The commit timestamp, \( T_e = (C_q(t_e) + \Delta_e) \circ q \) of transaction \( T_w \) is generated at processor \( q \) at local time \( C_q(t_e) \). \( T_{ro} \) reads a data-item, \( X \), at local processor \( d \), at local time \( C_d(t_r) \). Due to the communication delay \( C_p(t_e) < C_d(t_r) \). A version of data-item, \( X \), is committed by \( T_w \) at time \( C_d(t_e) \). From the timeliness property of the multicast and the choice of the value of "delta" \( C_q(t_e) + \Delta_e \geq C_d(t_e) \). The monotonicity of the clock function implies \( t_e > t_r \ L_2 \ C_d(t_e) > C_d(t_r) \). It follows that for a RO-transaction \( T_{ro} \) reading \( X \) at time \( t_r \) and a transaction \( T_w \) committing at time \( t_e > t_r : C_q(t_e) + \Delta_e \geq C_d(t_e) > C_d(t_r) > C_p(t_e) \Rightarrow T_{ro}.T_e \leq T_w.T_e \), which concludes the proof.

10.3.2 HRT/SRT Commit

When a SRT transaction reads a value produced by a HRT transaction, the TR-Read procedure is used. The commit of the HW-objects by the HRT transactions can be done with above described HRT-commit protocol (see Fig. 22). The proof of hypothesis 4 for a SRT-transaction reading a data-item from the set HW-objects is similar to the proof of hypothesis 3 in the proof of theorem 10. The commit of the HW-objects by the HRT transactions can be done with above described HRT-commit protocol.

Data-items from the set of SW-objects can be read and written by SRT transactions and read by HRT transactions. When a root SRT-transaction writes an item to be read by a HRT-transaction, another commit protocol is needed than the one used for HRT because the timeliness of the underlying SRT multicast can no longer be guaranteed. The timeliness of the commit of the SW-objects is essential. Without special measures the commit protocol can be preempted. In such a case, part of the involved objects has the state commit and another part is left with the state tentative. The verification of these unwanted consequences by the HRT transactions takes a lot of time from the HRT-schedule which can be spent differently. The timeliness of the Commit can be guaranteed by executing the Commit protocol for one or more SRT-transactions in sufficiently large and frequent gaps in the HRT schedule.

To guarantee the timeliness of the commits involving SW-objects, one process per processor executes the COMPROC Procedure in a repetitive gap of the HRT-schedule (see Figs 23 and 24). This process has access to all data-items connected to the processor. The coordinators of completed SRT-transactions fill a queue, \( RQ \), at the coordinator’s processor with their commit requests. A time-out is associated with each request. When the gap is met in the schedule, all processors execute the COMPROC procedure.

**2PL & MVTSO** The requests which exceed their deadline are removed from \( RQ \) (line 15). A limited amount of requests is read from the queue and stored in \( nr \) (line 16).

All processors multicast their requests to all correct processors, where they are received in \( nr \) (lines 17-19). The commits of the transactions are executed in the same order (in this case the time-out order) for the data-items associated with the executing processor (line
VAR

RS = \{ pid | pid is identifier of correct processor \}

RQ = \{ <T, t, ws, rs> | T is identifier of currently active transaction with timeout value t and write(read)-set wo(rs) \}

D_p = \{ q | q is data-item resident on processor p \}

PROCEDURE COMPROC()

VAR

nr : set of requests;

rqst : request;

cntr : Integer; T_c : timestamp

BEGIN

WHILE ingap DO

    \forall rq \in RQ \text{ DO IF } rq.t < \text{endgap THEN } RQ := RQ \backslash rq \text{ FI OD;}

    nr := \{ rq | rq \in RQ \land rq.t > \text{endgap} \};

    Multicast( nr, RS);

    \forall p \in RS \text{ DO } nr := nr \cup \text{receive}(p) \text{ OD}

    T_c := \text{timestamp}(\text{endgap});

    WHILE nr \neq \emptyset \text{ DO}

        rqst := (\min \{ rq | rq.t < \text{nr} \}) \land \text{nr} \land \text{receive};

        cntr := cntr + 1; \forall d \in (\text{rqst.ws} \cup \text{rqst.rs}) \cap D_p \text{ DO A-Commit(d, rqst.T, T_c \circ cntr) OD}

    OD; OD

END

Figure 23: SW-Commit process COMPROC for 2PL & MVTSO

22-23). All SRT-transactions verify that their requests have been executed in time. When the timeout expires, they are aborted following the SRT-commit protocol rules described in section 10.3.3.

OCC-BV The A-Commit action needs some modifications with respect to the original specification. In contrast to former requirements, a transaction that is committed in gap \( g \) can afterwards still be aborted in the same gap \( g \). A transaction that is committed in gap \( g \) cannot be aborted in gap \( g + k \) with \( k > 0 \).

When the gap is met in the schedule, all processors execute the COMPROC procedure (see Fig. 24). The requests which exceed their deadline are removed from \( RQ \) (line 16). A limited amount of requests is read from the queue and stored in \( nr \) (line 17).

All processors multicast their requests to all correct processors, where they are received in \( lr \) (lines 18-19). A-Validate is executed for all requests on the local data-items. When no data exist locally, A-Validate returns TRUE. When A-Validate succeeds, A-Commit is executed and the requests is added to \( nr \) (line 25-26). The set of committed requests is broadcast to all participants. All participants receive the same request sets from all participants (line 28-29). For requests that are removed at at-least one processor, a local A-Abort is executed (line 30).

The later Abort removes versions that are not yet read by any other transactions. Transactions read versions that are installed in one of the earlier terminated gaps.

Theorem 11 COMPROC satisfies the five commit requirements and hypotheses 2, 3 and 4.

Proof. Unanimity: the Unanimity requirement is met by the unanimity property of the supporting multicast. All requests are sent to all participants. When a participant receives a
1 VAR
2 3 RS = \{ pid | pid is identifier of correct processor \}
3 RQ = \{ < T_t, t, ws, rs, T_e > | T_t is identifier of currently active transaction with time-out value t and write(read)-set ws(rs) \}
4 D_p = \{ q | q is data-item resident on processor p \}
5 PROCEDURE COMPROC()
6 VAR
7 8 tr, nr, lr : set of requests;
9 rq : request;
10 wal : set of data-items;
11 b: boolean, cntT: integer;
12 len : Timestamp;
13 BEGIN
14 WHILE ingap DO
15   \forall rq \in RQ DO IF rq.t \leq endgap THEN RQ := RQ \setminus rq FI OD;
16   nr := \{ rq | rq \in RQ \setminus \forall r < K \wedge rq.t > endgap \};
17   Multicast( nr, RS)
18   lr := \emptyset; \forall p \in RS DO lr := lr \cup receive(p) OD
19   lr := lr; nr := \emptyset; cntT := 1
20   T_e := timestamp(endgap)
21   WHILE lr \neq \emptyset DO
22     \forall q \in RQ DO IF rq.t \leq endgap THEN RQ := RQ \setminus rq FI OD;
23     \forall d \in ((\forall ws \cup \forall rs) \cap D_p) DO b := b \wedge A-Validate(d, rq.T_t, rq.T_e) OD
24     IF b THEN cntT := cntT + 1; \forall d \in ((\forall ws \cup \forall rs) \cap D_p) DO A-Commit(d, rq.T_t, rq.T_e \circ cntT)
25       nr := nr \cup \{ rq \}; FI
26   OD; OD
27   Multicast( nr, RS)
28   lr := \emptyset; \forall p \in RS DO lr := lr \cup receive(p) OD
29   \forall rq \in lr - lr, \forall d \in (\forall ws \cap D_p) DO A-Abort(d, rq.T_t) OD
30   OD
31 END

Figure 24: SW-Commit process COMPROC for OCC-BV

multicast message, all participants receive it. At line 19 of both COMPROC procedures, all processors have the same commit requests in the variable lr. In case of OCC-BV, Validate requests are executed in the order determined by the time-out values at line 24 and all processors execute the Validate requests in the same order. When a transaction cannot be validated, the request is not added to nr. All processors send the contents of nr to all processors. From the unanimity of the multicast it follows that all processors receive the same values and end up with the same requests in lr at line 29. At all processors the same transactions are aborted at line 30. At line 25 all remaining data-items of a given transaction are committed with the same Te value. This concludes the unanimity part.

Integrity: in case of 2PL & MVTSO, all participants agree to commit all received requests due to hypothesis 7 and the requirement is met for all requests. In case of OCC-BV, participants only agree when A-Validate returns TRUE at line 22. Refused requests are not inserted into nr. All participants send each other their nr lists and abort all transactions that are refused by any processor. Agreement depends on the former commits. The removal of a committed version in a given gap g does not change the agreement of later transactions committed in the same gap g.

Validity: in case of 2PL & MVTSO all participants agree to participate in the Commit due to hypothesis 7 and their requests to commit are stored in nr. Once stored in nr, the Commit actions are executed at line 23. In case of OCC-BV, the requests validated by all
participants are maintained in \( nr \). All participants receive the same lists of validated requests. When all participants agree, their requests are maintained in \( Ir \) at line 28. For those requests no Aborts are executed.

**Ordering:** the Ordering is trivially met by choosing the commit time at the end of the gap and postfixing the value of counter \( cntr \).

**Timeliness:** the Timeliness property of the underlying multicast assures that messages are sent within bounded time. From the moment that the request is transmitted, all participants will receive this request in bounded time. They only receive a bounded number of requests and will send a bounded number of commit requests to the participants. Within bounded time the participants receive the message and will execute the commit. Consequently, the timeliness requirement is met.

**Hypothesis 2:** the proof of hypothesis 2 is separated in two parts.

Exception atomicity: for each transaction either an Abort or a Commit action is executed. After the execution of COMPROC all versions belonging to a given transactions have the state commit or have been removed.

Concurrency atomicity: an order is established by the time-out values. The A-Validate and A-Commit actions are sequentially executed in this same order on all processors.

**Hypotheses 4 and 3:** new HRT or SRT transactions will start after the gap and will find the appropriate versions ready and committed. Tentative versions will be committed in one of the following gaps with a timestamp larger than the starting times of the active transactions.

The COMPROC procedure for OCC-BV leads to unwanted Aborts. A transaction \( T_1 \) that aborts on site \( p \) and but writes an item on site \( q \) may lead to the abort of a transaction \( T_2 \) that read \( q \). The later global abort of \( T_1 \) permits the Commit of \( T_2 \) that is not performed in COMPROC. Remedies to this situation can be envisaged by sending information about all accessed data by all committable transactions to all sites. This additional information can be used to construct a commit order that allows the minimum number of aborts. However, such a strategy leads to many messages.

In case of OCC-NVTI, the determination of a globally correct \( C_f \) value leads to the exchange of even more information. In this case it is probably wiser to execute the Validate and Commit pairs one after the other and not group them as is the case in COMPROC.

### 10.3.3 SRT Commit

The example of OCC-NVTI is taken. Other CCA's can be treated similarly. The SRT-Commit (shown in figs 26 and 25) depends on the clock-synchronization and the Multicast facilities. The major difference with the former two protocols is the invalidity of the timeliness property of the underlying multicast protocol. Therefore, extra measures must be taken to validate hypothesis 3. A "validate" message is sent to all participants. Each participant calculates an appropriate commit timestamp based on its local clock. The transaction identifier with the calculated timestamp is stored in the history \( H : identifier \rightarrow Timestamp \). Each participant initiates a time-out \( > Time + delay \) and returns the calculated Commit timestamp. When all results are returned and validated, a "commit" together with the maximum of the received timestamps is sent by the coordinator. Commits that take too long are aborted. The timelines of the "Commit" is verified by each participant. When a participant sees that the time-out has passed, it sends an "abort" message to all participants including itself. Because all multicast
messages are ordered, all participants will receive the same messages in the same order and will execute either A-Abort or A-Commit. After arrival of the "commit" message, the entry in $H$ is removed and the commit is executed on all data-items belonging to $T$. On arrival of an "abort" message the corresponding item is removed from $H$ and the abort is executed on the with $T$ associated data-items. When a sequence of messages arrives for one transaction, the first request specified in the message is executed and the transaction identifier is removed from $H$, after which all messages for this transaction are ignored.

\[
\text{VAR} \\
H : \text{identifier} \rightarrow \text{Timestamp} \\
D_p = \{ q \mid q \text{ is data-item resident on processor } p \}
\]

\[
\text{PROCEDURE Wait-Mess()} \\
\text{BEGIN} \\
H := \emptyset \\
\text{WHILE TRUE DO} \\
\quad \text{ON reception of } I \oplus T_i \oplus C_f \oplus C_i \oplus \text{"validate"} \\
\quad \text{set Time-Out;} \\
\quad \forall v \in I \cup D_p \text{ DO } A-\text{Validate}(v, T_i, C_f, C_i) \text{ OD} \\
\quad H := H \cup \{ T_i \rightarrow TS(T_i + \text{delay}) \} \\
\quad \text{send}(T_i \oplus TS(T_i + \text{delay})) \\
\quad \text{ON reception of } I \oplus T_i \oplus T_e \oplus C_f \oplus \text{"commit"} \\
\quad \text{IF } T_i \in \text{dom}(H) \text{ THEN} \\
\quad \quad \text{dom}(H) := \text{dom}(H) \setminus \{ T_i \} \\
\quad \forall v \in I \cap D_p \text{ DO } A-\text{Commit}(v, T_i, T_e, C_f) \text{ OD} \\
\quad \text{FI} \\
\quad \text{ON reception of } I \oplus T_i \oplus \text{"abort"} \\
\quad \text{IF } T_i \in \text{dom}(H) \text{ THEN} \\
\quad \quad \forall v \in I \cap D_p \text{ DO } A-\text{Abort}(v, T_i) \text{ OD} \\
\quad \text{dom}(H) := \text{dom}(H) \setminus \{ T_i \} \\
\quad \text{FI} \\
\quad \text{ON occurrence of } \text{"time-out"} \\
\quad \text{Multicast}(I \oplus T_i \oplus \text{"abort"}, \text{participants}) \\
\text{OD} \\
\text{END}
\]

Figure 25: Wait-Mess for SRT-Commit

The precondition of the RO-Read action must be extended with:

\[
T < (\min s \in \text{dom}(H) : H(s)) \tag{77}
\]

to assure that hypothesis 3 is met. The RO-Read action may only be started when no transaction waits to be committed with a $T_r$ smaller than the $T_s$ of the RO-Read invoking RO-transactions.

**Theorem 12** The procedures SRT-Commit, SRT-Validate and Wait-mess satisfy the first four commit requirements and hypotheses 2 and 3.

**Proof.** **Unanimity:** the Unanimity requirement is met by the unanimity and ordering properties of the underlying multicast. All correct participants will receive the same multicast messages in the same order. The first decision on a given transaction is received by all participants and consecutively executed. All participants will take the same decision for any
PROCEDURE SRT-Commit( I: data-items, T_s, T_e: Timestamp, C_f : Counter)
BEGIN
    Multicast( I \oplus T_s \oplus T_e \oplus C_f \oplus "commit", participants)
END

PROCEDURE SRT-Validate(I: data-items, T_s, T_e : Timestamp, C_f, C_t : Counter)
VAR
    C_{3n}, C_{4n} : Counter
    T : Timestamp
BEGIN
    Multicast( I \oplus T_s \oplus C_f \oplus C_t \oplus "validate", participants)
    T_e := 0
    \forall participants DO receive(T); T_e := max(T_e, T_e) OD
END

Figure 26: SRT-Commit algorithm

given transaction. Once a A-Commit or A-Abort has been executed for a transaction \( T \), this transaction is removed from \( H \). Later messages in connection with transaction \( T \) are ignored.

**Integrity:** participants only receive "commit" requests from the coordinator when all participants have returned an answer, implying their agreement.

**Validity:** when no failures or time-outs occur, all participants will send their answers to the coordinator. The "commit" message of the coordinator arrives at all destinations after which A-Commit is executed.

**Ordering:** See proof of lemma 17.

**Hypothesis 2:** follows directly from proof of lemma 17.

**Hypothesis 3:** the RO-Read precondition assures that when RO-Read is executed at local time \( t \) with a given \( T_s \), all committing transactions will have a \( T_e < T_s \). After \( t \), commit requests are added to \( H \) with a \( T_e \) that is larger than \( T_s \). Consequently, all transactions committing at local time \( t' > t \) have a \( T_e > T_s \).

Transactions can be aborted when the precondition of A-Validate is not met because another transaction has put the state of the specified item \( X \) equal to validating, or the "validate" request can be put in a wait-queue till the state of the data-item has changed. In the latter case, the total ordering of all multicast messages prevents deadlocks. When a transaction \( T_1 \) waits for the commit of another transaction \( T_2 \), The validate request of \( T_1 \) was sent before the request of \( T_2 \). The total ordering of the "validate" requests prevents cycles in the wait-for graph.

11 Implementation restrictions

11.1 finite number of versions

In the above it has been assumed that once created versions remain permanently available. This is not realistic and a fixed number of versions per object is assumed.

The conditions under which a RO-transaction based on A-Read outperforms the RO-transaction based on the RO-Read action need to be investigated.

RO-transactions can be aborted for two reasons: (1) missing of the deadline and (2) un-
availability of the required version. In Fig. 27, the different behaviours are shown schematically. Time progresses from left to right. In the upper part the A-Read action behaviour and in the lower part the RO-Read action behaviour is shown. The evolution of the versions are shown by the sequence of rectangles and the evolution of the transactions by the full and dotted lines. At time $T_w$ version $n$ of a given item is created. At time $T_c$ the version is committed and from the following time $T'_w$ the versions $n + 1$ and $n + 2$ are created. For this example it is assumed that only two versions are kept per item. Therefore, version $n + 2$ is drawn with a dotted line because at its creation time version $n$ is removed. When A-Read is used (upper part of figure), transactions which are started between the creation times, $T_w < T_s < T'_w$, of version $n$ and $n + 1$ will select version $n$. When version $n$ is selected at the beginning of the transaction, the version can only be accessed after time $T_e$. In the Figure this is shown with the dotted line. The transaction can be so delayed that it meets its deadline before it terminates and is aborted. The transaction started with $T_c < T_s < T'_w$, will not wait and terminate before its deadline (shown by the dotted line between $T_c$ and $T_d$).

In the lower part of the figure the behaviour of the RO-Read actions is visualized. A transaction starting between the commit-times, $T_c < T_s < T'_c$, of version $n$ and version $n + 1$ selects version $n$. Transactions will not wait and terminate correctly assuming that they are not preempted. This is shown by the uninterrupted line at the bottom of the figure. When this transaction accesses version $n$ at the end of the transaction, the version has been replaced by version $n + 2$ and the transaction needs to be aborted.

Performance in the case of the A-Read action is limited by the longer duration of the transaction and in the case of the RO-Read actions is limited by the shorter availability of a given version for the transactions. A criterion can be established when one method will
outperform the other.

Assume that \( k \) versions of all items are regularly updated with a fixed update interval \( \theta \):

\[ \theta = T'_w - T_w = T'_c - T_c \]  

(78)

It is assumed that the distribution, \( P(T_s) \), of start times of transactions which need version \( n \) is flat.

\[
P(T_s) = \begin{cases} 
1/\theta & \text{if } T_w < T_s < T'_w, \ T_c < T_s < T'_c \\
0 & \text{otherwise}
\end{cases}
\]

(79)

The probability, \( P(T_s < X) \), that \( T_s \) has a value lower than a given value, \( X \), is given by:

\[
P(T_s < X) = \int_{-\infty}^{X} P(T_s) dT_s = \begin{cases} 
\frac{X - T_w}{\theta} & \text{if } T_w \leq X < T'_w \\
\frac{X - T_c}{\theta} & \text{if } T_c \leq X < T'_c \\
1 & \text{if } X \geq T'_w \\
1 & \text{if } X \geq T'_c
\end{cases}
\]

(80)

When an item is read with A-Read, the selected version may be committed or still tentative. The probability, \( P(w_1 > 0) \), that the transaction needs to wait for time, \( w_1 \), on the first item is given by the probability that \( T_s < T_c \):

\[ P(w_1 > 0) = P(T_s < T_c) = \frac{T_c - T_w}{\theta} \text{ if } T_w \leq T_c < T'_c \]  

(81)

Assume that the reading of a data-item takes a fixed time, \( \delta \) (including eventual preemption times), and that each transaction reads a fixed number, \( n \), of data-items. The transaction execution time, \( \Delta T \), is then given by:

\[ \Delta T = n.\delta \]  

(82)

It is possible that the reading of a data-item is shorter than the time between the start and commit of a transaction;

\[ \delta < T_c - T_w \]  

(83)

Assume that the creation times of versions of different data-items are not related. In that case after the access of the first data-item the transaction may wait for the commit of the second data-item and after the access of the \( n^{th} \) for the commit of the \((n + 1)^{th}\). \( P_1(w) \) is the probability that a transaction waits an interval of length \( w_1 \) for the commit of item 1. \( P_1(w) \) is equal to the probability \( P(T_c - T_s = w_1) \) if \( T_s < T_c \) and is equal to the probability \( P(w_1 = 0) = P(T_s \geq T_c) \):

\[ P_1(w) = P(w_1 = T_c - T_s \mid T_s < T_c) + P_1(w_1 = 0 \mid T_s \geq T_c) \]

(84)

This can be rewritten as:

\[ P_1(w) = P(w_1) + P_1(0) \]

(85)
where each term is given by:

\[
\begin{align*}
\text{P}_1(0) = P(T_s > T_c) &= \int_{T_c}^{T_c + w_2} \frac{1}{\theta} d T_s = \frac{T_c - T_s}{\theta} & \text{if } w_1 = 0 \\
P(w_1) &= \frac{1}{\theta} & \text{if } 0 < w_1 \leq T_c - T_w
\end{align*}
\]

(86)

(87)

\(P_2(w)\) is the probability that a transaction waits a time interval \(w_2\) for the commit of item 2. \(P_2(w)\) can be calculated with the aid of \(P_1(w)\). The probability, \(P_2(w)\), must be separated in two parts: (1) when \(T_s + \delta + w_1 > T_c\), then \(P_2(w)\) is determined by the probability distribution \(P_1(w)\) and (2) when \(T_s + \delta + w_1 \leq T_c\), then \(P_2(w)\) is determined by the probability that \(T_c - T_s - \delta = w_2\) for item 2:

\[
P_2(w) = P_2(w \mid T_s + w_1 + \delta > T_c) + P_2(w \mid T_s + w_1 + \delta \leq T_c)
\]

(88)

Each of these terms must be considered for the case that \(w_1 = 0\) and the case \(w_1 > 0\).

\[
P_2(w \mid T_s + w_1 + \delta > T_c) = P_1(0). P(T_s + \delta > T_c) = \\
\int_{T_c}^{T_c + w_2} \frac{1}{\theta} d T_s = \frac{(T_c - T_s)(T_c - T_s + \delta)}{\theta^2} & \text{if } w = 0
\]

(89)

\[
P_2(w \mid T_s + w_1 + \delta \leq T_c) = P_1(0). \int_{T_s + \delta - w_1}^{T_c + \delta} \frac{1}{\theta} d T_s = \frac{T_c - T_s + \delta + w}{\theta^2} & \text{if } 0 < w < T_c - T_w \\
P(w_1). \int_{T_c + \delta - w_1}^{T_c + w_2} \frac{1}{\theta} d T_s = \frac{(T_c - T_s + \delta + w)}{\theta^2} & \text{if } 0 < w < T_c - T_w
\]

(90)

Each of these terms can be calculated:

\[
P_2(w \mid T_s + w_1 + \delta > T_c \land w_1 > 0) = P(w_1). P(T_s + \delta > T_c) = \\
\int_{T_c}^{T_c + w_2} \frac{1}{\theta} d T_s = \frac{(T_c - T_s)(T_c - T_s + \delta + w)}{\theta^2} & \text{if } w = 0
\]

(91)

\[
P_2(w \mid T_s + w_1 + \delta \leq T_c \land w_1 > 0) = P(w_1). P(T_s + \delta \leq T_c) = \\
\int_{T_c + \delta - w_1}^{T_c + w_2} \frac{1}{\theta} d T_s = \frac{(T_c - T_s + \delta + w)}{\theta^2} & \text{if } 0 < w < T_c - T_w
\]

(92)

The wait distribution after two items, under the condition: \((T_w + \delta < T_c)\), is given by:

\[
P_2(w) = (2(w + T_w' - T_c + \delta))/\theta^2 & \text{if } 0 < w < T_c - T_w - \delta
\]

(93)

\[
P_2(w) = (T_w' - T_c + \delta)/\theta^2 & \text{if } w = 0
\]

(94)

The wait distribution after \(i\) items, under the condition: \((T_w + (i - 1)\delta < T_c)\), is given by:
\[ P_i(w) = \begin{cases} 
\frac{\prod_{m=0}^{i-1}(T_w - T_c + m.\delta)}{\sum_{k=0}^{i-1} A_{i,k} w^k} & \text{if } w = 0 \\
\frac{\sum_{k=0}^{i-1} A_{i,k} w^k}{\sum_{k=0}^{i-1} A_{i,k,i-1} w^k} & \text{if } 0 < w < T_c - T_w - (i-1)\delta \\
\frac{\sum_{k=0}^{i-1} A_{i,k} w^k}{T_c - T_w - (i-1)\delta \leq w < T_c - T_w - (i-2)\delta} \\
. & \text{if } T_c - T_w - \delta \leq w < T_c - T_w 
\end{cases} \] (96)

The coefficients \( A_{i,k,j} \) will be recursively defined. Again separate the probability of waiting on the \( i \)th item in two parts:

\[ P_i(w) = P(w \mid T_s + w_{i-1} + (i-1)\delta > T_c) + P_i(w \mid T_s + w_{i-1} + (i-1)\delta \leq T_c) \] (97)

Each of these terms must be considered for the case that \( w_{i-1} = 0 \) and the case \( w_{i-1} > 0 \).

\[ P_i(w \mid T_s + w_{i-1} + (i-1)\delta > T_c) = \] (98)

\[ P_i(w \mid T_s + (i-1)\delta > T_c \land w_{i-1} = 0) + P_i(w \mid T_s + w_{i-1} + (i-1)\delta > T_c \land w_{i-1} > 0) \]

\[ P_i(w \mid T_s + w_{i-1} + (i-1)\delta \leq T_c) = \] (99)

\[ P_i(w \mid T_s + (i-1)\delta \leq T_c \land w_{i-1} = 0) + P_i(w \mid T_s + w_{i-1} + (i-1)\delta \leq T_c \land w_{i-1} > 0) \]

Each of these terms can be calculated:

\[ P_i(w \mid T_s + (i-1)\delta > T_c \land w_{i-1} = 0) = P_{i-1}(0).P(T_s + (i-1)\delta > T_c) = \]

\[ P_{i-1}(0).\int_{T_c-(i-1)\delta}^{T_s} 1/\theta \, dT_s = P_{i-1}(0).\frac{(T_c - T_s + (i-1)\delta)/\theta}{1/\theta^i.\prod_{m=1}^{i-1}(T_w - T_c + (m-1)\delta)} \] if \( w = 0 \)

\[ P_i(w \mid T_s + w_{i-1} + (i-1)\delta > T_c \land w_{i-1} > 0) = P(w_{i-1}).P(T_s + w_{i-1} + (i-1)\delta > T_c) = \]

\[ P(w_{i-1}).\int_{T_c-(i-1)\delta-w_{i-1}}^{T_s} 1/\theta \, dT_s = \]

\[ P_{i-1}(w).\frac{(T_c - T_w - (i-1)\delta + w)/\theta}{1/\theta^i.\prod_{m=1}^{i-1}(T_w - T_c + (m-1)\delta)} \] if \( 0 < w < T_c - T_w \)

\[ P_i(w \mid T_s + w_{i-1} + (i-1)\delta \leq T_c \land w_{i-1} = 0) = \]

\[ P_{i-1}(0).P(w_{i} = T_c - T_s - (i-1)\delta) = P_{i-1}(0)/\theta = \] (102)

\[ (1/\theta^i.\prod_{m=1}^{i-1}(T_c - T_s + (m-1)\delta)) \] if \( 0 < w < T_c - T_w - (i-1)\delta \)

\[ P_i(w \mid T_s + w_{i-1} + (i-1)\delta \leq T_c \land w_{i-1} > 0) = \]

\[ \int_0^{w_{i-1}} P(w_{i-1}) \, dw_{i-1}.P(w_{i} = T_c - (i-1)\delta - T_s) = \]

\[ 1/\theta.\int_0^{w_{i-1}} P(w_{i-1}) \, dw_{i-1} \] if \( 0 < w < T_c - T_w - (i-1)\delta \)

(103)

From the above equations the values of \( A_{i,k,j} \) can be determined:
if $1 \leq j < i \land i < k < 0$

$A_{i,k,j} = A_{i-1,k,j}(T_w - T_c + (i-1)\delta)/\theta + A_{i-1,k-1,j}/\theta$ \hspace{1cm} (104)

if $1 \leq j < i \land k = 0$

$A_{i,0,j} = A_{i-1,0,j}(T_w - T_c + (i-1)\delta)/\theta$ \hspace{1cm} (105)

if $j = i \land i < k < 0$

$A_{i,k,i} = A_{i-1,k,i}(T_w - T_c + (i-1)\delta)/\theta + A_{i-1,k-1,i}/\theta + A_{i-1,k-1,i}/(k\theta)$ \hspace{1cm} (106)

if $j = i \land k = 0$

$A_{i,0,i} = A_{i-1,0,i}(T_w - T_c + (i-1)\delta)/\theta + (1/\theta^i) \cdot \prod_{m=1}^{i-1} (T_w' - T_c + (m-1)\delta)$ \hspace{1cm} (107)

The mean duration, $\Delta T'$, of a transaction which uses A-Read and waits for the versions to commit is determined by:

$$\Delta T' = \Delta T + \int_{0}^{T_c - T_w} w P_n(w) \, dw$$

Next the probability must be calculated that a transaction, using A-Read, is aborted because its waiting time is unacceptably long. Transactions can only be aborted when the worst case wait-time, $T_c - T_w$, plus the total execution time, $\Delta T$, is larger than the for execution allocated time period $T_d - T_s$:

$$T_d - T_s < \Delta T + T_c - T_w$$

The maximum time, $W$, that a transaction, using A-Read, may wait is defined by:

$$W = T_d - T_s - \Delta T > 0$$

Negative values of $W$ imply that all transactions will be aborted. The probability that a transaction which uses A-Read, executes beyond its deadline is given by: $P_n(w > W)$:

$$P_n(w > W) = \int_{W}^{\infty} P_n(w) \, dw \quad \text{if } 0 \leq W < T_c - T_w$$
$$= 0 \quad \text{otherwise} \hspace{1cm} (111)$$

The chances $P_f(P_{f'})$ that a transaction using A-Read(RO-Read) cannot find the appropriate version of the last, $k^i h$, item is calculated now. The probability, $P_{f'}$, of abort of a transaction using Ro-Read is equal to the probability $P(T_s + \Delta T > T_w + k\theta)$:

$$P_{f'} = \int_{T_w + T_c - T_s}^{\infty} P(T_s) \, dT_s = \frac{T_c - T_w - k\theta + n\delta}{\theta}$$

$$= 1$$

if $T_c < T_w + k\theta - n\delta \leq T_c'$ \hspace{1cm} (112)

if $T_c' \geq T_w + k\theta - n\delta$.

The probability, $P_f$, that under the same circumstances a transaction does not find the appropriate version using A-Read is equal to the probability $P(T_s + w + \Delta T > T_w + k\theta)$. This is equivalent to the probability $P_n(w > T_w + k\theta - T_s - n\delta)$ for all possible values of $T_w$. 

63
\[ P_f = \int_{T_w}^{T_s} P(T_s) dT_s \int_{T_w+k-T_s-n\delta}^{\infty} P_n(w) dw \] (113)

Assume that the aborts due to the unavailability of versions is independent of the aborts due to missing of deadlines. The RO-Read actions will result in less aborts than the Read actions when

\[ (1 - P_n(w > W))(1 - P_f) < 1 - Pf' \] (114)

11.2 Implementation of RO-Read Action

The RO-Read implementation for SRT transactions is straightforward in the case of OCC-BV and 2PL. Care must be taken in the case of MVTSO and OCC-NVTI, as the value of \( M_{ro} \) must be determined in an efficient way. Three possibilities exist: (1) requesting the values from the provider when needed in two phases, (2) requesting the values from the provider in one phase and (3) regularly providing the values to requester at the moment they are available. The third option is a valid one for Real-Time systems which have a periodic behaviour and often transactions require values on a repetitive basis.

11.2.1 2 phase communication

The RO-transaction as shown in figures 3, 12 and 13 need to be refined to calculate the \( M_{ro} \) value. The algorithm for requesting a value in two phases is shown in fig. 28. Two functions \( loc \) and \( ident \) are introduced. The function \( loc \) returns the identifier of the processor where data-item \( X \) is located, and \( ident \) returns the identifier of data-item \( X \).

MVTSO The function Close, with version parameter \( v \), assures that for all versions, \( x \), with \( x.T_w < v.T_w \): the read timestamp, \( T_r \) is equal to the write Timestamp \( T_w \) of its immediate successor. Consequently, no more versions can be added between the versions created before version \( v \).

In the RO-Transaction, for every required data-item, \( X \), the minimum, \( T_m \), is determined of the not yet committed versions and the maximum of the \( T_w \) values of the committed versions, \( x \), with \( x.T_w < T_m \) is returned. At line 36 these \( T_m \) values are stored in the set \( ST \). For all returned data-item, \( X \), no new version, \( x \), can be created with a \( x.T_w < T_m \) because all \( T_r \) values are made equal to all the \( T_w \) values of their immediate successors.

At line 38, for all data-items, \( X \), RO-Read is invoked with a parameter, \( M_{ro} \), which is equal to the minimum of the set \( ST \). The returned values are stored in the local variables \( A[X] \) at line 39. At every processor the Procedure Wait-Mess is permanently waiting to receive the messages of the RO-transactions.

OCC-NVTI In the RO-Transaction, for every required data-item, \( X \), the minimum, \( C_m \), is determined of the not yet committed versions and the versions with a \( T_r \) value larger than \( T_s \). The minimum of \( C_m \) and the maximum of the \( C_b \) values of the committed versions, \( x \), with \( x.C_b < C_m \) is returned. At line 30 of Fig. 29 these \( C_m \) values are stored in the set \( SC \). The calculated \( C_m \) value stays correct because new versions are created by A-Write with \( C_b \) values larger than all existing ones.
FUNCTION Close(v: version)
{PRE: (\exists X \in D: v \in X.V) \land \pi_m.T_w = (\max x \in X.V, x.T_w < v.T_w \land v.T_w = T}
{POST: \pi_m.T_r = T \land \operatorname{Close}(\pi_m)}

PROCEDURE ROtime(X: data-item): Timestamp
{PRE: T_m = (\min x \in X.V, x.S \neq \text{commit} \land x.T_w)
\land v \in X.V \land v.T_w = (\max x \in X.V, x.T_w < T_m, x.S = \text{Commit} \land x.T_w)
{POST: \operatorname{Close}(v) \land \operatorname{RET}: T_m}

PROCEDURE RO-Read(X: data-item, T: Timestamp, var a: value)
{PRE: v \in X.V \land v.T_w = (\max x \in X.V, x.S = \text{commit} \land x.T_w \leq T \land v.T_w)
\land v.value = A
{POST: a = A}

PROCEDURE Wait-Mess()
BEGIN
WHILE TRUE DO
ON reception of ident(X)
T := ROtime(X)
send(T)
ON reception of ident(X) \oplus T
RO-Read(X, T, a)
send(a)
END

PROCEDURE RO-Transaction(RS: read actions, I: data-items, t: timestamp)
VAR
ST: set of Timestamps, T_m, \mathcal{M}_r: Timestamp
A[\text{data-items}]: values
BEGIN
T_s = f(t, times); ST := \emptyset
\forall X \in RS: \text{send}(\text{loc}(X), \text{ident}(X))
\forall X \in RS: \text{receive}(T_m); ST := ST \cup T_m
\mathcal{M}_r := \text{min}(ST)
\forall X \in RS: \text{send}(\text{loc}(X), \text{ident}(X) \oplus \mathcal{M}_r)
\forall X \in RS: \text{receive}(A), A[X] := A
execute calculations
END

Figure 28: RO-Read requests for MVTSO, two communication phases

At line 32 of Fig. 29, RO-Read is invoked with a parameter, \mathcal{M}_r, which is equal to the minimum of the set \mathcal{S}. The returned values are stored in the local variables \text{A}[X].

In these algorithms the communication is done in two phases. In the first phase, requests for the minima are sent to all participating processors. Based on the returned values, \mathcal{M}_r is calculated. In the second phase the RO-Read Procedure is invoked at all participants with the communicated \mathcal{M}_r value. The returned values are used. This method is quite expensive in messages.

11.2.2 1 phase communication

MVTSO In Fig. 30, a one phase algorithm is shown for MVTSO. It is assumed that a lower limit on the age of the acquired values can be specified. Older values are assumed to be unusable for the application. This age parameter is sent over to the participants which
PROCEDURE ROtime(X: data-item, Ts: Timestamp): Counter

{PRE: \( C_m = \{ \text{min} x \in X.V, (x.S \neq \text{commit} \lor x.T > x.C_b) \} \)}

{POST: \( \text{RET: min}(C_m, \{ \text{max} x \in X.V, x.S = \text{commit} : x.C_b \}) \)}

PROCEDURE RO-Read(X: data-item, C: Counter, var a: value)

{PRE: \( v \in X.V \land (\forall x \in X.V, x.S = \text{commit}, x.C_b \leq C : x.C_b \leq v.C_b) \land v.value = A \land v.C_b \leq C \land v.S = \text{commit} \)}

{POST: \( a = A \)}

PROCEDURE Wait-Mess()

WHILE TRUE DO

ON reception of ident(X)

\( C_m := \text{ROtime}(X) \)

send(\( C_m \))

ON reception of ident(X) \( \oplus C \)

RO-Read(X, C, v)

send(v)

OD

PROCEDURE RO-Transaction(RS: read actions, I: data-items, t: timestamp)

VAR SC: set of Timestamps, \( T_m, M_{ro} : \) Timestamp

A[data-items]: values

BEGIN

\( T_m = f(t, \text{time}); SC := \emptyset \)

\( \forall x \in RS: \text{send}(\text{loc}(X), \text{ident}(X)) \)

\( \forall x \in RS: \text{receive}(\text{cm}); SC := SC \cup C_m \)

\( M_{ro} := \text{min}(SC) \)

\( \forall x \in RS: \text{send}(\text{loc}(X), \text{ident}(X) \oplus M_{ro}) \)

\( \forall x \in RS: \text{receive}(A); A[X] := A \)

execute calculations

END

Figure 29: RO-Read requests for OCC-NVTI, two communication phases

return committed versions which are younger than the specified age-limit and which are older than the oldest not yet committed version. On the base of the returned sets of versions the \( M_{ro} \) value can be calculated and the required values can be selected locally because they are already available at the requesting processor.

In the Procedure VClose, the minimum, \( T_m \), of all \( T_w \) values of the uncommitted versions is determined. The set of versions is determined which have a \( T_w \) values which lies between this minimum and the specified age-limit \( T \). This set is returned and the \( T_r \) values of the committed versions are set equal to the \( T_w \) values of their immediate committed successor versions. The Close function in the POST condition of VClose is the same as specified in Fig. 28. The Procedure RO-Read has the same specification as the RO-Read of the former algorithm with the exception that the data-item parameter is replaced by a set of versions.

Again at every processor the Wait-Mess procedure is started which waits for messages. On reception of a message, the Function VClose is invoked and the result set is sent back to the requester.

The RO-Transaction sends a request for the set of versions to all participants and waits for the results which consecutively are stored in the set \( X.S \). For every data-item the maximum of all \( T_w \) values of the versions is calculated. The minimum of all maxima is stored in \( T_m \) and
PROCEDURE VClose( X: data-item, T: Timestamp) : set of versions

PRE: \[ T_m = (\min x \in X.V x.S \neq \text{commit} : x.T_w) \land A \subseteq X.V \land (\forall x \in A : T \leq x.T_w \leq T_m) \]

\[ \land \forall x \in A : x.T_w \leq (\forall x \in X.V - A. x.S = \text{commit} : x.T_w < T) \]

POST: Close(v) \land \text{RET : A}

PROCEDURE RO-Read( V: set of versions, T: Timestamp, var a: value)

PRE: \[ v \in V \land v.T_w = (\max x \in X.V, x.S = \text{commit}, x.T_w \leq T : x.T_w) \land v.value = A \]

POST: a = A

PROCEDURE Wait-Mess()

BEGIN
WHILE TRUE DO
ON reception of ident(X) \oplus T
V := VClose(X, T)
send(V)
END

PROCEDURE RO-Transaction(RS: read actions, I: data-items, t: timestamp)

VAR
XS: set of set of versions, Ts, M_{ro} : Timestamp
A[data-items]: values
BEGIN
Ts = f(t, time - interval); XS := \emptyset
\forall x \in X.S: send( loc(x), ident(x) \oplus T_x)
\forall x \in RS: receive( V); X.S := X.S \cup V
M_{ro} := (\min : V \in X.S: (\max : x \in V : x.T_w))
\forall V \in X.S: RO-Read(V, M_{ro}, A[V])
execute calculations
END

Figure 30: RO-Read requests for MVTSO, one communication phase

constitutes the \( M_{ro} \) value. The RO-Read procedure is invoked with this \( M_{ro} \) values for each set of versions present in \( X.S \).

OCC-NVTI In Fig. 31, a one phase algorithm for OCC-NVTI is shown. It is assumed that a lower limit, \( T_a \), on the age of the acquired values can be specified. Older values are assumed to be unusable for the application. This age parameter is sent over to the participants which return committed versions which are younger than the specified age-limit and which are older than the oldest not yet committed version. On the base of the returned sets of versions, the \( M_{ro} \) value can be calculated and the required values can be selected locally because they are already available at the requesting processor.

The Procedure RO-Read has the same specification as the RO-Read of the former algorithm with the exception that the data-item parameter is replaced by a set of versions.

Again at every processor the Wait-Mess procedure is started which waits for messages. On reception of a message, the relevant versions are selected in \( V \) and the result set is sent back to the requester.

The RO-Transaction sends a request for the set of versions to all participants and waits for the results which consecutively are stored in the set \( X.S \). For every data-item the maximum of all \( C_b \) values of the versions is calculated. The minimum of all maxima is stored in \( C_m \) and constitutes the \( M_{ro} \) value. The RO-Read procedure is invoked with this \( M_{ro} \) values for each...
1. PROCEDURE RO-Read( V: set of versions, C: Counter, var a: value)
2. {PRE: v ∈ V ∧ v.S = commit ∧ v.Cb ≤ C ∧ (∃x ∈ X.V, x.S = commit : x.Cb ≤ v.Cb) ∧ v.value = a
3. POST: a = A
4. }
5. }
6. PROCEDURE Wait-Mess()
7. VAR
8. V: set of Version
9. BEGIN
10. WHILE TRUE DO
11. ON reception of ident(X) ⊕ T₀ ⊕ Tₛ
13. send(V)
14. OD
15. END
16. }
17. PROCEDURE RO-Transaction(RS: read actions, I: data-items, t: timestamp)
18. VAR
19. XS: set of set of versions, Tₛ, T₀: Timestamp, Mᵀᵥ : Counter
20. A[data-items]: values
21. BEGIN
22. Tₛ := time o pid; T₀ = (time - interval) o pid; XS := φ
23. ∀act[X] ∈ RS: send( loc(X), ident(X) ⊕ Tₛ ⊕ T₀)
24. ∀act[X] ∈ RS: receive(V); XS := XS ∪ V
25. Mᵀᵥ := (min V ∈ XS: (max x ∈ V: x.Cb))
26. ∀V ∈ XS : RO-Read(V, Mᵀᵥ, A[V])
27. execute calculations
28. END

Figure 31: RO-Read requests for OCC-NVTI, one communication phase

set of versions present in XS.

11.2.3 Repetitive update

MVTSO In Fig. 32 the last alternative for MVTSO is shown. Every time a Commit action is executed on a data-item X, relevant versions are sent to all destinations which want to acquire X on a regular basis. For that reason the data-item structure is extended with another field, req, in which all requesting processors are stored. In all processors the identifier to data-item mapping, XS, is stored. The Procedure Wait-Mess, executing on every processor, waits for messages and adds new versions to the version lists of the specified data-item replicated in XS.

In RO-transaction the maximum Tᵥ values of the versions of a given identifier locally stored in XS are determined for all required data-items. The minimum of these values represents Mᵀᵥ. RO-Read is invoked on the versions of the data-items replicated in XS. The specification of RO-Read is the same as for the RO-Read in Fig. 30.

Every time a version of a requested data-item (X.req ≠ φ) is committed, the Procedure PCommit is invoked with the Commit timestamp. Pcommit invokes XClose and sends the returned versions to the requesting processors. The Function XClose returns the last committed versions which have not been sent to the requesters. At any given moment a series of already communicated versions exists. They can be identified from their Tᵥ values which are equal to the Tᵥ values of their immediate successors.

When a version commits, it can be a version which is the immediate successor of an
PROCEDURE XClose( X: set of versions, T: Timestamp, var V: set of versions): boolean

{PRE: v.T_r = (max x ∈ X.V : x.T_w) ∧ T = T_r ∧ T_r = v.T_r ∧ (v.T_r ≠ T ⇒ V = W ∪ {v}) ∧ RET: TRUE}

{POST: (v.S = commit ∧ XClose(X \ v, v.T_w, W))}

PROCEDURE RO-Read( V: set of versions, T: Timestamp, var a: value)

{PRE: v ∈ V ∧ v.T_w = (max x ∈ V, x.T_w ≤ T : x.T_w) ∧ v.value = A}

{POST: a = A}

VAR XS: ident → data-item

PROCEDURE PCommit( X: data-item, T: Timestamp)

VAR W: set of versions

BEGIN

IF XClose(X, T, W) THEN Send( dest, W) FI

END

PROCEDURE Wait-Mess()

BEGIN

WHILE TRUE DO

ON reception of ident(X) ⊆ W

XS(ident(X)).V := XS(ident(X)).V ∪ W

OD

END

PROCEDURE RO-Transaction(RS: read actions, l: data-items, t: timestamp)

VAR

M_ro : Timestamp

A[data-items]: Values

BEGIN

M_ro := (min : act[X] ∈ RS : (max : x ∈ XS(X).V : x.T_w))

∀act[X] ∈ RS : RO-Read( XS(X).V, M_ro, A[X])

execute calculations

END

Figure 32: RO-Read updates for MVTSo

uncommitted version. In that case, XClose returns the empty set. When the committed version is the immediate successor to an already communicated version, then XClose returns this version and the committed immediate successors. The read timestamp of the returned versions are adjusted such that the \( T_r \) values of their immediate predecessors are equal to their \( T_w \) value.

OCC-NVTI In Fig. 33 the last alternative is shown. Every time A-Commit is executed on a data-item \( X \), relevant versions are sent to all destinations that want to acquire \( X \) on a regular basis. For that reason the data-item structure is extended with another field, req, in which all requesting processors are stored. In all processors the identifier to data-item mapping, XS, is stored. The Procedure Wait-Mess, executing on every processor, waits for messages and adds new versions to the version lists of the specified data-item replicated in XS.

In RO-transaction the maximum \( C_b \) values of the versions of a given identifier locally stored in XS are determined for all required data-item. The minimum of these values rep-
Figure 33: RO-Read updates for OCC-NVTI

respects $M_{ro}$. RO-Read is invoked on the versions of the data-items replicated in $XS$. The specification of RO-Read is the same as for the RO-Read in Fig. 31.

Every time a version of a requested data-item ($X$.req $\neq \emptyset$) is committed, the Procedure PCommit is invoked with the Commit timestamp. Pcommit sends the returned versions to the requesting processors.

12 Conclusions

An application consisting of transactions acting on a database is divided in two parts: a HRT and a SRT part. HRT- and SRT-transactions communicate via database items. The presented extensions to existing algorithms allow RO-transactions to proceed without waiting or with considerably reduced wait-times. The concept of transfer serializability is formulated to increase the number of interleavings between HRT and SRT transactions.

13 Acknowledgements

K. Vidyanshar has pointed out a mistake in the order definition proposed in an earlier version of this document. This work has benefited from discussions with P.T.A. Thijssen. A.Engel
and A.Glim have contributed to the specifications of the ACTIONS presented here.

References


<table>
<thead>
<tr>
<th>Paper Number</th>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>93/01</td>
<td>R. van Geldrop</td>
<td>Deriving the Aho-Corasick algorithms: a case study into the synergy of programming methods, p. 36.</td>
</tr>
<tr>
<td>93/02</td>
<td>T. Verhoeff</td>
<td>A continuous version of the Prisoner's Dilemma, p. 17</td>
</tr>
<tr>
<td>93/03</td>
<td>T. Verhoeff</td>
<td>Quicksort for linked lists, p. 8</td>
</tr>
<tr>
<td>93/04</td>
<td>E.H.L. Aarts, J.H.M. Korst, P.J. Wieting</td>
<td>Deterministic and randomized local search, p. 78.</td>
</tr>
<tr>
<td>93/05</td>
<td>J.C.M. Baeten, C. Verhoef</td>
<td>A congruence theorem for structured operational semantics with predicates, p. 18.</td>
</tr>
<tr>
<td>93/06</td>
<td>J.P. Volkamp</td>
<td>On the unavoidability of metastable behaviour, p. 29</td>
</tr>
<tr>
<td>93/07</td>
<td>P.D. Moerland</td>
<td>Exercises in Multiprogramming, p. 97</td>
</tr>
<tr>
<td>93/08</td>
<td>J. Verhoosel</td>
<td>A Formal Deterministic Scheduling Model for Hard Real-Time Executions in DEDOS, p. 32.</td>
</tr>
<tr>
<td>93/10</td>
<td>K.M. van Hee</td>
<td>Systems Engineering: a Formal Approach Part II: Frameworks, p. 44.</td>
</tr>
<tr>
<td>93/16</td>
<td>H. Schepers, J. Hoornan</td>
<td>A Trace-Based Compositional Proof Theory for Fault Tolerant Distributed Systems, p. 27</td>
</tr>
<tr>
<td>93/17</td>
<td>D. Alstein, P. van der Stok</td>
<td>Hard Real-Time Reliable Multicast in the DEDOS system, p. 19.</td>
</tr>
<tr>
<td>93/18</td>
<td>C. Verhoef</td>
<td>A congruence theorem for structured operational semantics with predicates and negative premises, p. 22.</td>
</tr>
<tr>
<td>93/19</td>
<td>G.J. Houben</td>
<td>The Design of an Online Help Facility for ExSpect, p.21.</td>
</tr>
<tr>
<td>93/21</td>
<td>M. Codish, D. Dams, G. Filé, M. Bruynooghe</td>
<td>Freeness Analysis for Logic Programs - And Correctness, p. 24</td>
</tr>
<tr>
<td>93/22</td>
<td>E. Poll</td>
<td>A Typechecker for Bijective Pure Type Systems, p. 28.</td>
</tr>
<tr>
<td>93/23</td>
<td>E. de Kogel</td>
<td>Relational Algebra and Equational Proofs, p. 23.</td>
</tr>
<tr>
<td>93/24</td>
<td>E. Poll and Paula Severi</td>
<td>Pure Type Systems with Definitions, p. 38.</td>
</tr>
<tr>
<td>93/26</td>
<td>W.M.P. van der Aalst</td>
<td>Multi-dimensional Petri nets, p. 25</td>
</tr>
<tr>
<td>93/27</td>
<td>T. Kloks and D. Kratsch</td>
<td>Finding all minimal separators of a graph, p. 11.</td>
</tr>
<tr>
<td>93/28</td>
<td>F. Kamareddine and R. Nederpelt</td>
<td>A Semantics for a fine λ-calculus with de Bruijn indices, p. 49.</td>
</tr>
<tr>
<td>93/29</td>
<td>R. Post and P. De Bra</td>
<td>GOLD, a Graph Oriented Language for Databases, p. 42.</td>
</tr>
<tr>
<td>93/30</td>
<td>J. Deogun, T. Kloks, D. Kratsch, H. Müller</td>
<td>On Vertex Ranking for Permutation and Other Graphs, p. 11.</td>
</tr>
</tbody>
</table>
Derivation of delay insensitive and speed independent CMOS circuits, using directed commands and production rule sets, p. 40.


ILIAS, a sequential language for parallel matrix computations, p. 20.

Real Time Process Algebra with Infinitesimals, p. 39.

Abstract Reduction and Topology, p. 28.

Non Interleaving Process Algebra, p. 17.

Design and Analysis of Dynamic Leader Election Protocols in Broadcast Networks, p. 73.

A general conservative extension theorem in process algebra, p. 17.

Job Shop Scheduling by Constraint Satisfaction, p. 22.


Temporal operators viewed as predicate transformers, p. 11.

Automatic Verification of Regular Protocols in P/T Nets, p. 23.

A taxonomy of finite automata construction algorithms, p. 87.

A taxonomy of finite automata minimization algorithms, p. 23.

A precise clock synchronization protocol, p. 4.


Verifying Sequentially Consistent Memory using Interface Refinement, p. 20.

The object-oriented paradigm, p. 28.

Canonical typing and II-conversion, p. 51.


Graph Isomorphism Models for Non Interleaving Process Algebra, p. 18.


Time and the Order of Abstract Events in Distributed Computations, p. 29.


A Hierarchical Diagrammatic Representation of Class Structure, p. 22.

Process Algebra with Partial Choice, p. 16.
The testing Paradigm Applied to Network Structure. p. 31.


A New Method for Integrity Constraint checking in Deductive Databases. p. 34.

Ups and Downs of Type Theory. p. 9.

Job Shop Scheduling by Local Search. p. 21.

Mathematical Induction Made Calculational. p. 36.

An Algebraic Semantics of Basic Message Sequence Charts. p. 9.

Refining Reduction in the Lambda Calculus. p. 15.

The performance of single-keyword and multiple-keyword pattern matching algorithms. p. 46.

Beyond β-Reduction in Church's λ→. p. 22.

An introduction to the FIRE engine: A C++ toolkit for Finite automata and Regular Expressions.

The design and implementation of the FIRE engine: A C++ toolkit for Finite automata and regular Expressions.

An algebraic semantics of Message Sequence Charts. p. 43.


On the foundations of functional programming: a programmer's point of view. p. 54.


Correctness of Real Time Systems by Construction. p. 22.

Process Algebra with Feedback. p. 22.

A Boyer-Moore type algorithm for regular expression pattern matching. p. 22.


A formalization of the Ramified Type Theory. p.40.

The Barendregt Cube with Definitions and Generalised Reduction. p. 37.

Delayed choice: an operator for joining Message Sequence Charts. p. 15.

Canonical typing and Π-conversion in the Barendregt Cube. p. 19.

Simulating and Analyzing Railway Interlockings in ExSpect, p. 30.

Point-free substitution, p. 10.
94/39  A. Blokhuis  T. Koks  On the equivalence covering number of splitgraphs, p. 4.
94/40  D. Alstein  Distributed Consensus and Hard Real-Time Systems, p. 34.
94/41  T. Koks  D. Kratsch  Computing a perfect edge without vertex elimination ordering of a chordal bipartite graph, p. 6.
94/43  R.C. Backhouse  M. Bijsterbosch  Category Theory as Coherently Constructive Lattice Theory: An Illustration, p. 35.
94/45  G.J. Houben  Tutorial voor de ExSpect-bibliotheek voor "Administratieve Logistiek", p. 43.
94/46  R. Bloo  F. Kamareddine  R. Nederpelt  The λ-cube with classes of terms modulo conversion, p. 16.
94/48  Mathematics of Program Construction Group  Fixed-Point Calculus, p. 11.
94/50  H. Geuvers  A short and flexible proof of Strong Normalization for the Calculus of Constructions, p. 27.
94/51  T. Koks  D. Kratsch  H. Müller  Listing simplicial vertices and recognizing diamond-free graphs, p. 4.
94/52  W. Penczek  R. Kuiper  Traces and Logic, p. 81
95/01  J.J. Lukkien  The Construction of a small Communication Library, p. 16.
95/02  M. Bezem  R. Bel  J.F. Groote  Formalizing Process Algebraic Verifications in the Calculus of Constructions, p. 49.
95/03  J.C.M. Baeten  C. Verhoef  Concrete process algebra, p. 134.
95/04  J. Hidders  An Isotopic Invariant for Planar Drawings of Connected Planar Graphs, p. 9.
95/05  P. Severi  A Type Inference Algorithm for Pure Type Systems, p. 20.
95/07  G.A.M. de Bruyn  O.S. van Roostwijk  Drawing Execution Graphs by Parsing, p. 10.
95/08  R. Bloo  Preservation of Strong Normalisation for Explicit Substitution, p. 12.
95/09  J.C.M. Baeten  J.A. Bergstra  Discrete Time Process Algebra, p. 20
95/10  R.C. Backhouse  R. Verhoeven  Q. Weber  Mathpad: A System for On-Line Preparation of Mathematical Documents, p. 15
95/11  R. Seljée  Deductive Database Systems and integrity constraint checking, p. 36.
95/12  S. Mauw and M. Reniers  Empty Interworkings and Refinement Semantics of Interworkings Revised, p. 19.
95/14  A. Posse, C. Verhoef, S.E.M. Vlijmen (eds.)  De proceedings: ACP'95, p.
95/16  D. Dams, O. Grumberg, R. Gerth  Abstract Interpretation of Reactive Systems: Preservation of CTL*, p. 27.
95/17  S. Mauw and E.A. van der Meulen  Specification of tools for Message Sequence Charts, p. 36.
95/19  J.C.M. Baeten and J.A. Bergstra  Discrete Time Process Algebra with Abstraction, p. 15.
95/20  F. van Raamsdonk and P. Severi  On Normalisation, p. 33.
95/21  A. van Deursen  Axiomatizing Early and Late Input by Variable Elimination, p. 44.
95/23  W.M.P. van der Aalst  Petri net based scheduling, p. 20.
95/25  L. Feijs  Synchronous Sequence Charts In Action, p. 36.
95/26  W.M.P. van der Aalst  A Class of Petri nets for modeling and analyzing business processes, p. 24.
95/29  H. Jurjus  On Nesting of a Nonmonotonic Conditional, p. 14
95/32  W.M.P. van der Aalst  Handboek simulatie, p. 51.
95/33  J. Engelfriet and JJ. Vereijken  Context-Free Graph Grammars and Concatenation of Graphs, p. 35.
95/34  J. Zwanenburg  Record concatenation with intersection types, p. 46.
95/35  T. Basten and M. Voorhoeve  An algebraic semantics for hierarchical P/T Nets, p. 32.
96/01  M. Voorhoeve and T. Basten  Process Algebra with Autonomous Actions, p. 12.
96/02  P. de Bra and A. Aerts  Multi-User Publishing in the Web: Dress, A Document Repository Service Station, p. 12
96/04  S. Mauw  Example specifications in phi-SDL.
96/06  W.M.P. van der Aalst and T. Basten  Life-Cycle Inheritance A Petri-Net-Based Approach, p. 18.
96/07  M. Voorhoeve  Structural Petri Net Equivalence, p. 16.
96/11  T. Laan  AUTOMATH and Pure Type Systems, p. 30.
96/12  F. Kamareddine and T. Laan  A Correspondence between Nuprl and the Ramified Theory of Types, p. 12.
96/13  T. Borghuis  Priorean Tense Logics in Modal Pure Type Systems, p. 61.
96/14  S.H.J. Bos and M.A. Reniers  The $I^2$ C-bus in Discrete-Time Process Algebra, p. 25.
96/15  M.A. Reniers and J.J. Vereijken  Completeness in Discrete-Time Process Algebra, p. 139.
96/16  P. Hoogendijk and O. de Moor  What is a data type?, p. 29.
96/17  E. Boiten and P. Hoogendijk  Nested collections and polytypism, p. 11.