<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>A unified concurrency control algorithm for distributed database systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Wang, CP; Li, Victor OK</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>The 4th International Conference on Data Engineering (ICDE 1988), Los Angeles, CA., 1-5 February 1988. In Conference Proceedings, 1988, p. 410-417</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>1988</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/158031">http://hdl.handle.net/10722/158031</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
A UNIFIED CONCURRENCY CONTROL ALGORITHM
FOR DISTRIBUTED DATABASE SYSTEMS

C. P. Wang and Victor O. K. Li
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089-0272

ABSTRACT

We present a unified concurrency control algorithm for distributed database systems in which each transaction may choose its own concurrency control protocol. Specifically, we integrate Two-Phase Locking, Timestamp Ordering, and Precedence Agreement into one unified concurrency control scheme. We show the correctness of the new scheme and study the problem of selecting the best protocol for each transaction in order to optimize system performance.

1 INTRODUCTION

Many concurrency control algorithms have been proposed for distributed database systems[3]. Most of them fall into two categories, namely, Two-Phase Locking (2PL)[8,19,22] and Timestamp Ordering (T/O)[2,4,5,9,17]. 2PL, which was first designed for centralized systems and later applied to distributed ones, has the merit of simplicity but suffers from the distributed deadlock problem[1,6,11]. On the other hand, T/O avoids deadlocks by enforcing a special execution order (timestamp order) among transactions, which may cause transaction restarts and degradation of system performance. In [24], the Precedence Agreement (PA) algorithm is proposed. This concurrency control algorithm is free from deadlocks and restarts. However, communication cost increases as the system load increases.

The best concurrency control algorithm for a particular application depends on the system parameters[10,14,16,20]. For example, in an environment where each transaction only accesses one data item through a write operation, 2PL outperforms T/O since no deadlocks may occur. On the other hand, when system load is heavy and transaction size (in terms of the number of data items accessed) is small (but bigger than one), T/O is superior to 2PL. Many parameters may affect the performance of a concurrency control algorithm. Some of them are (1) transaction arrival rate, (2) the number of read operations versus the number of write operations, (3) transmission delay, (4) the number of data items requested by each transaction, (5) the cost of restarts, (6) deadlock detection time and cost, etc.

To implement a distributed database system, the system designers may estimate the system parameters beforehand and pick the best concurrency control algorithm for that system. From then on, every transaction in the database must follow that algorithm. We call this type of concurrency control static concurrency control. Most existing concurrency control methods are static. The major drawback is inflexibility. The originally chosen algorithm may not always be the best as the system parameters change. In addition, the term 'the best concurrency control algorithm' is really transaction dependent. Static concurrency control can only capture the average behavior but fails to reflect the individual differences among transactions.

The alternative is dynamic concurrency control. In this case, depending on the system states and the transaction types, different concurrency control methods are used for different transactions in an attempt to improve performance. There are several design problems associated with dynamic concurrency control. These include: (1) What are the candidate concurrency control algorithms? (2) How does one integrate these algorithms together and still preserve their correctness? (3) What are the system parameters relevant to choosing the best concurrency control algorithm and how does one determine them? (4) Given the relevant system parameters, how does one choose the best concurrency control method?

This paper presents the design of a dynamic concurrency control system which integrates 2PL, T/O, and PA. Our design consists of two steps:

1. Algorithm Integration: The goal is to design an integrated system such that each transaction can choose one of the above three algorithms. Our approach includes defining a unified model for distributed concurrency control algorithms, called the Precedence Assignment Model (PAM)[24]. Based on this model, the problem of unifying 2PL, T/O and PA is reduced to the distributed computation of two functions. Then we solve the distributed computation problems and prove the correctness of our unified algorithm.

2. Algorithm Selection: A cost function is defined to measure the loss of system throughput due to the requests issued by a transaction under its concurrency control method. The concurrency control algorithm will be chosen to minimize this cost function. We will discuss the evaluation of the cost function.

This paper is organized as follows. We describe the distributed database model and the precedence assignment model in sections 2 and 3, respectively. In section 4, the integration...
of concurrency control algorithms is presented. The selection of the best concurrency control algorithm is discussed in section 5. In section 6, we conclude with a list of future research tasks.

2 SYSTEM MODEL

A database consists of a finite set of logical data items \( D = \{ D_1, D_2, ...\} \). A transaction is a sequence of logical read/write operations requested by the user. We assume that a legal transaction consists of three phases, namely, the read phase, the local computing phase and the write phase. The read phase must be executed first and consists of operations which read from the database to the user's local memory. The local computing phase, consisting of all local computations, is then executed. Finally, during the write phase, data is copied from the user's memory back to the database. For database correctness, a transaction must either be completely executed or not executed at all.

Each logical data item \( D_i \) can be stored redundantly at different computer sites in the network. The physical copies of \( D_i \) are denoted \( D_{ij} \). Operations which access physical data item \( D_{ij} \) are physical reads \( r(D_{ij}) \), or physical writes \( w(D_{ij}) \). To execute a transaction, the system first translates the logical operations into their corresponding physical operations. The physical operations are sent to the sites where the physical data items are located and implemented there.

Let \( T = \{ t_1, t_2, ... \} \) be a set of transactions. An execution of \( T \) is the implementation of all operations associated with all the transactions in \( T \) on the database. An execution \( E \) is serial if no two transactions are executed concurrently at any time. An execution is said to be serializable [12] if the effect of the execution is equivalent to some serial execution.

Two logical (physical) operations conflict if they access the same logical (physical) data item and one of them is a write operation. We model the execution of transactions by a set of logs. There is one log associated with each physical data item. The log indicates the order in which physical operations are implemented on that data item. The properties of serializable execution have been extensively studied [3,12]. An important result is:

**Theorem 1** [12,13,18] Let \( T \) be a set of transactions. An execution \( E \) of \( T \) is serializable if there exists a total order on \( T \) such that if \( O_1 \) and \( O_2 \) are conflicting operations of distinct transactions \( t_1 \) and \( t_2 \), \( O_1 \) is implemented before \( O_2 \) in any log \( L_{t_1}, L_{t_2}, ... \) if and only if \( t_1 < t_2 \) in the total order.

An execution is conflict serializable if it satisfies the condition stated in Theorem 1. The total order defined in Theorem 1 is called the serialization order. All existing concurrency control algorithms use conflict serializability as their correctness criterion.

3 THE PRECEDENCE-ASSIGNMENT MODEL

PAM was developed to model concurrency control algorithms in distributed databases. It is shown that 2PL, T/O and many other existing concurrency control algorithms may be modeled by PAM. In this section, we will describe PAM.

3.1 THE CONCURRENCY CONTROL SUBSYSTEM

The concurrency control subsystem is modeled as follows:

1. **Request Issuer (RI):** Located at each user site is a request issuer. RI takes user transactions as its input and sends requests to data sites.

2. **Data Queue (Queue):** There is a data queue for each physical data item. Each entry in the queue is a user request. The user request at the head of the queue has the right to access the data.

3. **Data Queue Manager (QM):** There is a data queue manager for each data queue. The data queue manager communicates with the RI's and enforces the order of execution of requests on the data queue.

Each transaction \( t_i \) is sent to one of the RI's, say \( \tau_i \). Let \( O_1, O_2, ..., O_k \) be the read or write operations of \( t_i \). For each \( O_k \), \( \tau_i \) sends a read or write request to the corresponding data queue manager accessed by \( O_k \). \( O_k \) cannot be implemented until its request is granted by that data queue manager.

3.2 PRECEDENCE ASSIGNMENT AND ENFORCEMENT FUNCTIONS

The concurrency control subsystem must serve two major functions:

1. **Precedence Assignment:** Assigns a precedence to each request accessing the same data. That is, for each data \( D_j \), there is a non-empty set \( SP_j \) together with a total ordering \( <j \) on \( SP_j \). (\( SP_j \), \( <j \)) is called the precedence space for \( D_j \). Let \( O \) be the set of all operations accessing \( D_j \). There is a one-to-one function \( AS_j : O_j \rightarrow SP_j \) computed by the concurrency control subsystem which assigns a precedence (an element of \( SP_j \)) to each operation accessing \( D_j \).

2. **Precedence Enforcement:** Controls the implementation of operations on each data item to satisfy the following two conditions:

**E1:** Suppose \( O_1 \) and \( O_2 \) are conflicting operations accessing \( D_j \). If \( AS_j(O_1) \leq AS_j(O_2) \) then \( O_1 \) is implemented before \( O_2 \); else \( O_2 \) is implemented before \( O_1 \).

**E2:** There exists a serialization order on the set of data items such that the precedence order of two conflicting operations follows the corresponding transactions' serialization order.

3.3 MODELING 2PL AND T/O

In this section, we illustrate how to use PAM to model 2PL and T/O.

2PL

Static 2PL is considered in this paper. For 2PL, the requests are handled in a first-come-first-served (FCFS) manner as each data queue. The precedences are assigned to operations according to their order of arrival. (E1) is satisfied by the assignment function itself as long as the data queue is FCFS and conflicting operations are not allowed to be implemented concurrently. (E2) is satisfied by the following locking protocol:

1. A request is granted a lock if all conflicting operations with lower precedences have been implemented (locks have been released).
2. The operations in a transaction can be implemented only if all the requests for that transaction are granted.

Some transactions may be blocked by the locking protocol forever, i.e., there is a deadlock. Ignoring deadlocks, the locking protocol together with (E1) implies (E2).

To sum up, 2PL can be specified by
1. Precedence Assignment: the arrival order at the data queue.
2. Precedence Enforcement: (E1) is implied by the assignment function; (E2) is satisfied by the locking protocol.

T/O

There are several versions of the T/O algorithm. Here we consider the Basic T/O algorithm [3] only. In Basic T/O, all the data items have the same precedence space, which is the set of timestamps. Operations from the same transaction are assigned the same precedence (or timestamp).

This assignment function satisfies (E2) automatically as the serialization order can be defined to be the timestamp order of the transactions. (E1) is enforced by rejecting requests which arrive out of timestamp order.

In summary, T/O can be specified as:
1. Precedence Assignment: Assign the transaction's timestamp to each operation.
2. Precedence Enforcement: (E2) is automatically satisfied, (E1) is enforced by transaction restarts.

3.4 THE PRECEDENCE AGREEMENT ALGORITHM BASED ON TIMESTAMPS

PA decides the precedence for each operation through negotiation with the participating RI's and QM's, and is free from deadlocks and restarts. The following is a description of one version of PA based on timestamps.

Each transaction ti is assigned a timestamp tuple (TSi, INTi), where TSi is the timestamp for ti and INTi is the backoff interval associated with ti. This algorithm works just like T/O except when a request arrives at the data queue and finds some conflicting operation with a later timestamp has already been granted. Instead of rejecting the newly arrived request, the data queue for data item j will backoff the new request's timestamp, i.e., find the minimum TSji which is acceptable by the T/O protocol, with TSji = TSi + k·INTi, k ∈ N (the natural numbers). Then TSji is sent back to the request issuer ri. Each request will either receive grants from all the data queues accessed by ti (corresponding to no rejection case in T/O) or receive some backoff timestamps, TS'ij's. In the former case, the transaction will backoff its timestamp to maxj TS'ij. This new timestamp will be sent to all the data queues accessed by ti to update each request's timestamp. Now ti can wait for the grants and begin execution.

We state the algorithm formally:

1. For each ti, ri performs the following:
   (a) Generate Qj = (TS'ij, INTj).
   (b) Send requests together with Qj to each data site Dj accessed by the current transaction.
   (c) Wait until either a grant or TS'ij is received from each QM(j).
   (d) If grants are received from each QM(j), jump to (g).
   (e) Let TS'1j = maxj TS'ij. Send TS'1j to each QM(j).
   (f) Wait until lock grants arrive.
   (g) Perform local computing (the data read are attached to the corresponding lock grant).
   (h) Send lock release to each QM(j).

2. The data queue manager QM(j) for Dj performs the following:
   (a) Keep track of two variables, R.TS(j) and W.TS(j), which are the biggest timestamp of granted read requests and write requests, respectively.
   (b) Receive request qi and timestamp tuple Qi from ri.
   (c) Depending on whether qi is a read or write request do:
      read: If TSi > W.TS(j), let TSi be qi's timestamp and mark qi 'accepted'; insert qi in QUEUE(j). Otherwise, calculate TS'ij = TSi + k·INTi, k ∈ N, such that TS'ij > W.TS(j); let TS'ij be qi's timestamp and mark qi 'blocked'; insert qi in QUEUE(j). Otherwise, calculate TS'ij = TSi + k·INTi, k ∈ N, such that TS'ij > max (W.TS(j), R.TS(j)); let TS'ij be qi's timestamp and mark qi 'blocked'; insert qi in QUEUE(j); send TS'ij to ri.
      write: If TSi > W.TS(j) and TSi > R.TS(j), let TSi be qi's timestamp and mark qi 'accepted'; insert qi in QUEUE(j). Otherwise, calculate TS'ij = TSi + k·INTi, k ∈ N, such that TS'ij > max (W.TS(j), R.TS(j)); let TS'ij be qi's timestamp and mark qi 'blocked'; insert qi in QUEUE(j); send TS'ij to ri.
   (d) When TS'ij is received from ri, update the timestamp of qi and mark qi 'accepted'; re-insert qi into the proper position in QUEUE(j).
   (e) Queue(j) works as follows:
      i. All entries are sorted in increasing timestamp order.
      ii. Define HD(j) to be the request in QUEUE(j) such that all requests in QUEUE(j) with smaller timestamps have already been granted.
         A. If HD(j) is marked 'blocked', wait.
         B. If HD(j) is a read request marked 'accepted', and all previously granted write requests have been released, then set R.TS(j) to the timestamp of HD(j). Send grant to the transaction issuing HD(j).
         C. If HD(j) is a write request marked 'accepted', and all previously granted requests have been released, then set W.TS(j) to the timestamp of HD(j). Send grant to the transaction issuing HD(j).
         D. Otherwise, wait.

PA can be modeled by PAM as follows:
1. Precedence Assignment: Through the PA protocol, a single precedence is assigned to each request of a transaction.
2. Precedence Enforcement: The precedence assignment function guarantees both (E1) and (E2).

The correctness of PA will be shown in the next section.
4 THE UNIFIED CONCURRENCY CONTROL SYSTEM

We have already demonstrated the precedence assignment and enforcement functions for 2PL, T/O, and PA. To unify these algorithms, we have to unify their assignment functions and then their precedence enforcement functions.

4.1 UNIFIED PRECEDENCE ASSIGNMENT FUNCTION

The assignment function for 2PL resides at each data site, and assigns the next available precedence to an incoming request. The assignment function for T/O is at the request issuers, and assigns the timestamp of the current transaction to all of its requests. The assignment function for PA is more complicated: The initial timestamp is decided at the request issuer; then each data site decides a feasible timestamp; finally, collecting all the feasible timestamps, the request issuer decides the timestamp for all requests of the current transaction. To unify these three algorithms, we first define the unified precedence space (UPS) at each data queue. UPS is the timestamp space and the precedence is assigned as follows: The precedence for T/O and PA requests are their respective timestamps. The precedence for a 2PL request \( q \) accessing data item \( j \) is the biggest timestamp which has ever appeared in data queue \( j \) before \( q \)'s arrival. This ensures the new request is inserted at the tail of the queue and all 2PL requests follow the FCFS rule. The precedence order is defined as follows:

1. Compare the value of the timestamps first.
2. If it is a tie, compare the site id's of the transactions. A 2PL controlled transaction is regarded as having the biggest site id.
3. If still tied, then either both requests are 2PL or both are not. If both are 2PL transactions, compare their arrival order at the data queue; otherwise, compare their transaction id's.

4.2 UNIFIED PRECEDENCE ENFORCEMENT FUNCTION

In the unified system, many types of requests may appear in each data queue. The enforcement function must enforce both (E1) and (E2) under all possible situations. Some requests which may be granted by individual enforcement functions cannot be granted by the unified enforcement function. For example, in pure T/O, a granted read request never prevents write requests with bigger timestamps from accessing the data. But the following example shows that if T/O is used together with 2PL, then sometimes read requests must lock the data to preserve (E2).

Example: Consider three data items \( x, y, z \), and three transactions \( t_1, t_2, t_3 \). \( t_1 \) and \( t_2 \) are controlled by T/O and \( t_3 \) is controlled by 2PL. The operations for the transactions are:

- \( t_1 \): \( r_1(x), w_1(y) \)
- \( t_2 \): \( r_2(y), w_3(s) \)
- \( t_3 \): \( r_3(z), w_3(x) \)

The precedence order of these requests in the data queue may be:

- \( \text{Queue}(x): r_1 < w_3 \)
- \( \text{Queue}(y): r_2 < w_1 \)
- \( \text{Queue}(z): r_3 < w_2 \)

Both \( r_2 \) and \( w_4 \) are T/O type requests. Therefore, \( w_4 \) will be implemented before \( t_2 \) is executed. But then all three transactions will be executed and the execution is not serializable.

One solution to the unified enforcement function is to use locking for all requests, i.e., before a transaction leaves the system, lock releases must be sent to each data queue accessed by it and a request cannot be granted until all previously granted conflicting requests have been released. This method, while sacrificing the degree of concurrency for T/O transactions, guarantees (E2).

We now introduce a more sophisticated approach, called semi-locks. This type of locks preserve (E2) without reducing the degree of concurrency for T/O. Intuitively, a data is semi-locked if it is considered unlocked by the T/O protocol but must be treated as locked by 2PL and PA. The following semi-lock protocol decides, before a data item is unlocked, when the lock can become a semi-lock, i.e., ready for T/O type requests:

1. There are four possible locks: read (RL), write (WL), semi-read (SRL), and semi-write (SWL).
2. Two locks conflict if they lock the same data item and at least one is a WL or SWL. A lock is pre-scheduled if at least one conflicting lock is granted earlier but not yet released. All other locks are called normal locks. The following rules describe when HDU may be granted and what kind of locks are granted:
   - If HD(j) is a read request issued by a 2PL or PA transaction, then it can be granted a RL if all previously granted WL's and SWL's on the data j have been released.
   - If HD(j) is a write request issued by a 2PL or PA transaction, then it can be granted a WL if all previously granted locks on data j have been released.
   - If HD(j) is a read request issued by a T/O transaction, then it can be granted a SRL if all previously granted WL's on data j have been released.
   - If HD(j) is a write request issued by a T/O transaction, then it can be granted a WL if all previously granted RL's and WL's on data j have been released.
   - If HD(j) was once granted a pre-scheduled lock, which later became normal, then a normal lock grant will be issued for HD(j).
3. Before execution, 2PL and PA transactions must hold proper RL's or WL's. After execution, lock releases are sent to the locked data item.
4. Before execution, a T/O transaction must hold a SRL on each data item read by it, and a WL on each data item written by it. After execution, it sends lock releases to the locked data if it did not get any pre-scheduled locks. Otherwise, it transforms all its locks into semi-locks (RL → SRL, WL → SWL). At this point, the transaction is considered executed. But the request issuer will continue to collect lock grants for this transaction until there is one normal lock grant received from each data item accessed by this transaction. Then lock releases are sent to the data locked by that transaction.

The semi-lock protocol preserves the enforcement functions for 2PL and PA in the following sense: If the system runs 2PL (PA) transactions only, then the unified enforcement function
works in the same way as the enforcement function for 2PL (PA).
On the other hand, when the system runs T/O only, the unified enforcement function results in lock release messages which are not generated by the T/O enforcement function. But these messages are irrelevant in the sense that no T/O transactions need these messages for synchronization.

4.3 CORRECTNESS

We will show the correctness of the unified concurrency control system. By doing so, the correctness of PA is also proved since it can be regarded as a special instance of the unified scheme where all transactions use PA as their concurrency control algorithm.

First define when an operation is implemented: Operations of 2PL and PA type transactions are implemented when the corresponding locks are released; a T/O operation is implemented when the operation changes a lock to a semi-lock on the data or, if the lock is released.

Let <s be a binary relation on the set of executed transactions defined as follows: Let t1 and t2 be two executed transactions, t1 <s t2 iff (1) t1 = t2 or (2) there is a pair of conflicting operations o1 and q2 issued by t1 and t2 respectively, and o1 is implemented before q2. An execution is conflict serializable iff the conflict graph induced by <s is acyclic [12,18], i.e., <s is a partial order.

Theorem 2: Let T be a set of transactions executed under the unified concurrency control algorithm. <s must be a partial order on T. In other words, the execution is conflict serializable.

Proof: Suppose <s is not a partial order on T. Then there is a cycle t1 <s t2 <s ... <s tn <s t1. In addition, t1, ..., tn cannot all be T/O type transactions since the semi-lock protocol implements T/O type operations according to their timestamp order. Without loss of generality, assume t0 is a 2PL or PA type transaction. The semi-lock protocol requires t0 to get all normal locks on all data items accessed by it. This implies t0 has to receive all normal locks and to release all of these locks, either semi or normal, before t0 can release any locks. By induction, t1 has to release all the locks before t2 releases any lock. But t0 is of 2PL or PA type and t0 <s t1 implies t0 releases some locks before t1. This is a contradiction.

Q.E.D.

Theorem 2 shows that if the unified system allows a set of operations to be executed, the execution must be conflict serializable. Now we want to consider the situation when the system is blocked. It will be shown that if the system is blocked, there must be at least one 2PL type transaction in the wait-for cycle, that is, T/O and PA never block the system by themselves.

Definition 1: At time u, if there are requests not implemented yet, psmall(u) is the smallest precedence of such requests; else, psmall(u) is undefined.

Let P be the set of all possible precedences, i.e., extended timestamp space, and let T be the set of all transactions. Let PPT be the set of precedences assigned to requests associated with the transactions in T. Our proof starts with a special case of PPT and will be generalized.

Lemma 1: If PPT is isomorphic to the order of N, then the system is blocked if and only if there such that ∀x > u, psmall(x) = C where C is a constant.

Proof: In any data queue, different requests have different precedences. Therefore, the number of requests whose precedence is C is finite. At least one of these requests, say rq, cannot be implemented in finite time, else psmall() will be changed to a value different from C. Therefore, the transaction issuing rq is blocked.

(a) If no such C exists, there are two cases: (a) ∃u such that ∀x > u, psmall(x) is undefined or (b) ∀u, ∃x > u such that psmall(x) ≠ psmall(x).

Case (a) implies the system becomes empty after finite time. Therefore, no transaction is blocked. Case (b) implies that psmall() is changed for infinitely many times. Since the number of requests with the same precedence are finite, psmall() cannot be changed to the same value for infinitely many times. This implies the range of psmall() is infinite. Therefore, if at time u, a request rq with precedence p is not implemented yet, there is time x > u and psmall(x) > p. This implies at time x, (1) rq has been implemented, or (2) rq is of T/O type and has been rejected, or (3) rq is of PA type and it has been assigned a new backoff precedence. It is clear that if case (1) or (2) occurs, the transaction issuing rq is either executed or restarted. If (3) occurs, rq's precedence is backed off to a bigger one. But eventually psmall() will be greater than this precedence because PA can backoff the precedence of rq at most once. Then rq will be implemented and the transaction issuing rq will be executed.

Thus we conclude that all transactions will be either executed or restarted after finite time, i.e., no blocking occurs.

Q.E.D.

Theorem 3: Assume the precedence order on PPT is isomorphic to the order of N, then the system is blocked only if the blocked request with the smallest precedence is from a 2PL type transaction.

Proof: From Lemma 1, the system is blocked if and only if ∃u, such that ∀x > u, psmall(x) = C. There are three cases:

1. psmall() is owned by a request issued by a T/O type transaction t. If t is rejected, then psmall() will be updated to another value. If t is not rejected, all of its requests have precedence psmall() which is the smallest precedences not yet implemented. In short, t can be executed and psmall() is updated to another value. We reach contradictions in both cases, thus psmall() cannot be owned by T/O type requests.

2. psmall() is owned by a request issued by a PA type transaction t. There are two cases:

(a) psmall() is the original timestamp of t: t can receive either lock grants or back-off precedences from all the data accessed by it. Then either t can be executed or t will back off its precedence to the biggest of the back-off precedences received.

(b) psmall() is the back-off precedence: t can receive back-off precedences from all the data accessed by it. Similarly, t can be executed or t will back off its precedence to the biggest of the back-off precedences received.

In either case, psmall() will be changed. Thus psmall() cannot be owned by a request issued by PA type transactions.

3. psmall() is owned by a request issued by a 2PL type transaction t.

We conclude that case (3) must be true.

Q.E.D.
Corollary 1: PA is correct and free from deadlocks and rejections.
Proof: Theorem 2 shows PA will output a conflict serializable execution. Theorem 3 shows every transaction will be executed eventually. In addition, from the definition of PA, no transactions will be restarted
Q. E. D.

Corollary 2: Any deadlock cycle must contain at least one 2PL type transaction.
Proof: This is also a consequence of Theorem 3. The rationale is as follows. Our unified precedence assignment functions have the property that the removal of 2PL transactions will not affect the precedences assigned to T/O and PA type requests. Thus the removal of 2PL transactions will not destroy any deadlock cycles which contain T/O and PA transactions only. But the removal of 2PL transactions does guarantee freedom from blocked transactions and hence deadlock freedom. Thus we conclude there must be at least one 2PL transaction in each deadlock cycle.
Q. E. D.

The assumption that PT is isomorphic to N can be relaxed if, during any finite period of time, only finitely many transactions may arrive. Ignoring 2PL type transactions, the timestamp space is isomorphic to N since the values of timestamps are from N and only finitely many T/O or PA type transactions can have the same timestamp value. The key problem which prevents PT from being isomorphic to N is that there are potentially infinitely many 2PL requests assigned the same timestamp. But this problem can be fixed. Conceptually, we can modify the precedence assignment function of 2PL to prevent infinitely many requests assigned the same timestamp and still preserve the precedence order of requests in the data queue. This modification will make PT isomorphic to N without changing any implementation order of operations.

The 2PL precedence assignment function ASj for data j is modified as follows: If there are infinitely many T/O or PA type requests entering Queue(j), ASj does not need to be changed since the arrival of these requests will stop ASj from assigning the same precedence. If, on the other hand, there are only finitely many T/O or PA type requests entering Queue(j), ASj remains unchanged until the T/O (or PA) request omax with the largest timestamp TSMAX arrives. Then ASj changes and assigns TSMAX+m to the mth arriving 2PL requests after the arrival of omax. The new ASj will only assign a timestamp, if ever, finitely many times.

5 DYNAMIC CONCURRENCY CONTROL
In the previous section, the unified concurrency control system is presented. We now discuss how to select the most profitable concurrency control algorithm for each transaction.
Before developing the selection method, we define our performance measure to be the average transaction system time S and study how the transaction arrival rate, A, and the number of data items accessed by each transaction t, affect S. Our simulation results [23] show that 2PL performs well when is low. When is high, although the number of transactions directly involved in deadlocks does not increase very much, S goes up dramatically since more transactions are blocked by deadlocked transactions. For T/O, S grows steadily as is increases. It outperforms 2PL when is high. However, as is also shown in [10], T/O becomes worse than 2PL and PA as is increases. Apparently, this is due to the significant increase of restart probability. PA is a compromise between 2PL and T/O. It performs like 2PL when is low and like T/O while is is high. When is is moderate, it outperforms both 2PL and T/O.

5.1 THE SYSTEM THROUGHPUT LOSS FUNCTION
One possible method for selecting concurrency control algorithms is as follows: Each request issuer keeps a table of average system time versus transaction types and concurrency control method chosen. For each newly arrived transaction, the request issuer picks the concurrency control algorithm which minimizes the average system time. However, we believe this method does not perform well for the following reasons:
1. Minimizing the system time for individual transactions is not equivalent to optimizing S.
2. This method is biased towards 2PL since once deadlocks occur, not only is the system time of deadlocked transactions prolonged, but also the system time of all blocked transactions. Thus 2PL transactions tend to shorten their own system time by degrading other transactions’ performance.

The method we have adopted, instead of considering average system time, is to estimate the system throughput loss, STL. Our method picks the concurrency control algorithm for each transaction to minimize STL. Before discussing STL any further, let us introduce the following notations and definitions:
1. The current transaction t has m(t) read requests, r1, r2, ..., rm(t), and n(t) write requests, q1, q2, ..., qn(t).
2. Let D(rq) denote the data item accessed by request rq.
3. The write throughput of a data queue, which is the average number of write locks granted per unit time, is denoted λw(j). Similarly, the read throughput is denoted λr(j). Let λw(J) and λr(J) denote the throughput averaged over all λw(j)'s and λr(j)'s respectively.
4. λA, the system throughput, is the summation of all λw(j)'s and λr(j)'s.
5. The fraction of read requests among all requests is Qr.
6. K is the average number of requests per transaction.

We also make the following assumptions and approximations:
1. In computing STL, we assume a transaction gets all its locks at the same time, and all its locks are released at the same time.
2. The output process at each data queue (the process of lock releases) is Poisson.

Now we consider STL(t), the STL due to t, under our assumptions. Assume a request rq locks a data for U time units, what is the throughput loss on that data? This depends on what kind of lock rq has. If it is a read lock, then during this period of time, no write locks can be granted. Potentially, the throughput
loss is $\lambda_a D(rq) \cdot U$. On the other hand, if it is a write lock, the loss is $(\lambda_a D(rq) + \lambda_w D(rq)) \cdot U$. But there are other losses due to $rq$: Suppose that after $rq$ gets its lock, another request $rq'$ gets a lock on data $k$, but the transaction issuing $rq'$ also issues a request blocked by $rq$, or blocked by some request blocked by $rq$, then $rq'$ is blocked until $rq$ releases the locks. Therefore, $STL(t)$ is difficult to calculate or measure in general. Thus we use the following approximation.

Suppose transaction $t$ starts to hold all $(m(t) + n(t))$ locks at time $0$, the throughput loss, $\lambda_t$, is given by

$$\lambda_t = \sum_{i=1}^{n(t)} \lambda_d(r_i) + \sum_{i=1}^{m(t)} \lambda_u D(q_i) + \lambda_t D(q_t).$$

Now, the rate at which requests get locks is $\lambda_A - \lambda_t$. A request to obtain a lock will be blocked if the transaction issuing this request also issues a blocked request. A random transaction, on the average, issues $K$ requests. Each request is blocked with probability $\frac{\lambda}{\lambda_A}$. Assuming that the probability of blocking at different data site are independent, the rate at which requests obtaining locks is blocked is

$$(\lambda_A - \lambda_t) \cdot (1 - (1 - \frac{\lambda}{\lambda_A})^{K-1}).$$

If a request gets a lock and blocks the data, the new throughput loss depends on whether it gets a read lock or a write lock. Again, we approximate it by taking the average, that is, the new throughput loss is $\lambda_A - \lambda_t$. Now we are ready to solve $STL(t)$ recursively by defining the following functions. Let $STL\left(\lambda_{loss}, u\right)$ represents the $STL$ during period of $u$ time units where the initial throughput loss is $\lambda_{loss}$. There are two cases:

1. During this period, no other new lock grants block any data queue; then the $STL$ is $\lambda_{loss} \cdot u$.

2. At time $z$, a new lock grant blocks a data queue; then the $STL$ is $\lambda_{loss} \cdot z + STL\left(\lambda_{loss} + \lambda_u + (1 - Q_z) \cdot \lambda_t, u - z\right)$. Thus $STL'$ can be defined recursively as follows:

$$STL'\left(\lambda_{loss}, u\right) = \lambda_A \cdot u;$$

else

$$STL'\left(\lambda_{loss}, u\right) = (1 - \exp(-\lambda_{lock} \cdot u)) \cdot \lambda_{loss} \cdot u + \int_{0}^{\lambda_{lock}} \exp(-\lambda_{lock} \cdot z) \cdot (\lambda_{lock} \cdot z + STL\left(\lambda_{loss} + \lambda_{lock}, u - z\right))dz$$

where $\lambda_{lock} = (\lambda_A - \lambda_{loss}) \cdot (1 - (1 - \frac{\lambda}{\lambda_A})^{K-1})$ and $\lambda_{loss} = \lambda_A + \lambda_u + (1 - Q_z) \cdot \lambda_t$.

Note that $STL'$ can be evaluated efficiently through Dynamic Programming techniques [7].

### 5.2 SELECTION BASED ON $STL$

Now we describe the selection method based on $STL$. Our selection is based on the following parameters which can either be collected periodically or estimated through analytical methods[14,15,21,25]:

1. For 2PL transactions:
   - $(a) U'_{2PL}$: average lock time if a request is not aborted.
   - $(b) U'_{2PL}$: average lock time of an aborted request.
   - $(c) P_A$: probability of abortion of a transaction due to deadlocks.

2. For T/O transactions:
   - $(a) U'_{T/O}$: average lock time if a request is not aborted.
   - $(b) U'_{T/O}$: average lock time if a request is aborted.
   - $(c) P_r$: the probability of rejection of a read request.
   - $(d) P_w$: the probability of rejection of a write request.

3. For PA transactions:
   - $(a) U'_{PA}$: average lock time if a request is not backed off later.
   - $(b) U'_{PA}$: average lock time if a request is backed off later.
   - $(c) P_b$: the probability a read request is backed off.
   - $(d) P_w$: the probability a write request is backed off.

Let $t$ be a transaction as described in the last section, we will estimate $STL$ due to $t$ for 2PL, T/O and PA, by the $STL'$ function just defined.

### 2PL

The $STL$ due to $t$ under 2PL, denoted $STL_{2PL}(t)$ can be solved by the following equation:

$$STL_{2PL}(t) = (1 - P_A) \cdot STL_{2PL}(t \mid t \text{ is not aborted}) + P_A \cdot STL_{2PL}(t \mid t \text{ is aborted}),$$

or

$$STL_{2PL}(t) = (1 - P_A) \cdot STL(\lambda_t, U_{2PL}) + P_A \cdot (STL_{2PL(t)} + STL(\lambda_t, U_{2PL}))$$

### T/O

Similarly, $STL_{T/O}(t)$ can be solved by the following equation:

$$STL_{T/O}(t) = (1 - P_r) m(t) \cdot (1 - P_w) m(t) \cdot STL(\lambda_t, U_{T/O}) + (1 - (1 - P_r) m(t)) \cdot (1 - P_w) m(t) \cdot STL(\lambda_t, U_{T/O})$$

Here $\lambda^*_i$ is the average system throughput loss due to blocking by $t$'s requests given that at least one of $t$'s requests is rejected. $\lambda^*_i$ can be solved by the following equation:

$$(1 - P_r) \sum_{i=1}^{n(t)} \lambda_a D(r_i) + (1 - P') \sum_{i=1}^{n(t)} \lambda_u D(q_i) + \lambda_t D(q_t) = (1 - P_r) m(t) \cdot (1 - P') m(t) \cdot \lambda^*_i + (1 - P_r) m(t) \cdot (1 - P') m(t) \cdot \lambda_i$$

In the above equation, the left hand side is the sum of the expected throughput loss due to each request of $t$. The right hand side calculates the average throughput loss by conditioning on whether $t$ is rejected or not. Thus $\lambda^*_i$ is given by

$$\lambda^*_i = \frac{1}{1 - (1 - P_r) m(t) \cdot (1 - P') m(t)} \cdot \sum_{i=1}^{n(t)} \lambda_a D(r_i) + (1 - P') \cdot \sum_{i=1}^{n(t)} \lambda_u D(q_i) + \lambda_t D(q_t)$$

### PA

Finally, $STL_{PA}(t)$ can be derived as follows
where $\lambda_t^*$ is

$$
\lambda_t^* = \frac{1}{1 - (1 - P_B)^{\eta(t)}(1 - P_B)^{\eta(t)}} \sum_{i=1}^{m(t)} \lambda_i D(r_i) + (1 - P_B) \sum_{i=1}^{m(t)} \lambda_i D(q_i) + \lambda D(q_i) - \lambda(1 - P_B)^{\eta(t)}(1 - P_B)^{\eta(t)}
$$

In our selection method, for a new transaction $t$, $STL_{PAM}(t)$, $STL_{T/O}(t)$, and $STL_{PA}(t)$ are calculated separately. The algorithm with the smallest STL value is chosen to be the concurrency control algorithm. To speed up the performance of the selection process, transactions may be categorized into different classes and the STL for each class may be calculated in advance.

6 CONCLUSIONS AND FUTURE RESEARCH

This paper describes a dynamic concurrency control method for distributed database systems. Our approach is to first define a unified model, PAM, and to show that 2PL, T/O, and PA fall integrated to form a unified concurrency control system. Specifically, we unify the precedence space for these three algorithms and develop the semi-lock protocol, which serves to unify the precedence enforcement function.

To complete the design of the dynamic system, we also study the problem of algorithm selection. For this purpose, the STL criterion is proposed. We find efficient ways to estimate the STL for 2PL, T/O, and PA.

A number of problems remain to be investigated. Some of them are: (1) a detailed simulation of the proposed methodology, (2) integration of other concurrency control algorithms into our unified system, (3) development of more accurate criteria than STL, (4) allowing transactions to change their concurrency control methods, and (5) investigation of other applications of PAM.

References