A Formal Characterization of Epsilon Serializability

Krithi Ramamritham[1]
Dept. of Computer Science
University of Massachusetts
Amherst MA 01003

Calton Pu[2]
Dept. Computer of Science and Eng.
Oregon Grad. Inst. of Sci. & Tech
Beaverton, OR 97006

Abstract

Epsilon Serializability (ESR) is a generalization of classic serializability (SR). In this paper, we provide a precise characterization of ESR when queries that may view inconsistent data run concurrently with *consistent* update transactions.

Our first goal is to understand the behavior of queries in the presence of conflicts and to show how ESR in fact is a generalization of SR. So, using the ACTA framework, we formally express the inter-transaction conflicts that are recognized by ESR and through that define ESR, analogous to the manner in which conflict-based serializability is defined. Secondly, expressions are derived for the amount of inconsistency (in a data item) *viewed* by a query and its effects on the *results* of a query. These inconsistencies arise from concurrent updates allowed by ESR. Thirdly, in order to maintain the inconsistencies within bounds associated with each query, the expressions are used to determine the *preconditions* that operations have to satisfy. The results of a query, and the errors in it, depend on what a query does with the, possibly inconsistent, data viewed by it. One of the important byproducts of this work is the identification of different types of queries which lend themselves to an analysis of the effects of data inconsistency on the results of the query.

[1] partially supported by the National Science Foundation under grants IRI-9109210 and IRI-9314376.

[2] partially supported by NSF, IBM, DEC, AT&T, Oki Electric Ind. and Texas Instruments.

Contents

Keywords: Epsilon Serializability, Concurrency Control, Transaction Processing, Formal Techniques, ACTA.

1 Introduction

Epsilon Serializability (ESR) [21, 29], a generalization of classic serializability (SR), explicitly allows some limited amount of inconsistency in transaction processing (TP). ESR enhances concurrency since some non-SR execution schedules are permitted. For example, epsilon-transactions (ETs) that just perform queries may execute in spite of ongoing concurrent updates to the database. Thus, the query ETs may view uncommitted, i.e., possibly inconsistent, data. Concretely, an update transaction may *export* some inconsistency when it updates a data item while query ETs are in progress. Conversely, a query ET may *import* some inconsistency when it reads a data item while uncommitted updates on that data item exist. The correctness notion in ESR is based on bounding the amount of imported and exported inconsistency for each ET. The benefits of ESR have been discussed in the papers cited above. For instance, ESR may increase system availability and autonomy [22] in distributed TP systems, since asynchronous execution is allowed. But in this paper we restrict our attention to ESR in a centralized TP system.

In its full generality, update ETs may view inconsistent data the same way query ETs may. However, in this paper we focus on the situation where query-only ETs run concurrently with *consistent* update transactions. That is, the update transactions are *not* allowed to view uncommitted data and hence *will* produce consistent database states.

Our first goal is to understand the behavior of queries in the presence of conflicts and to show how ESR in fact is a generalization of SR. So, in section 2, using the ACTA framework [5, 6, 4] we formally express the inter-transaction conflicts that are recognized by ESR and, through that, define ESR, analogous to the manner in which conflict-based serializability is defined.

Our second goal is to quantify the amount of inconsistency experienced by queries. To this end, in section 3, expressions are derived for the amount of inconsistency (in a data item) *viewed* by a query. These inconsistencies arise from concurrent updates allowed by ESR. This section also considers how transaction aborts affect the inconsistency of data.

ESR imposes limits on the amount of inconsistency that can be viewed by a query. So, our third goal is to find ways by which these bounds are maintained. Using the expressions quantifying the inconsistency, we derive *preconditions* that operations have to satisfy. Derivation of these preconditions is the subject of Section 4. These preconditions point to possible mechanisms that can be used to realize ESR and show that more flexible implementations than those presented in [21, 29] are possible.

The effects of the inconsistent view on the results of a query depend on what a query does with the viewed data. In general, a small data inconsistency can translate into an arbitrarily large result inconsistency. So our fourth goal is to derive the effect of the inconsistency of the data read by a query on the *results* produced by the query. This derivation is done in Section 5 which also shows some of the restrictions that need to be imposed on the queries and updates so as to be able to bound the inconsistency in the result of the query to lie within reasonable limits. This helps characterize the situations in which ESR is applicable. Thus, one of the important byproducts of this work is the identification of different types of queries which lend themselves to an analysis of the effects of data inconsistency on the results of the query.

Related work is discussed in Section 6 while section 7 concludes the paper and offers suggestions for further work.

In the rest of this introduction, we provide an informal introduction to ESR and define the terms used.

1.1 ESR and ETs

A database is a set of data items. Each data item contains a value. A database state is the set of all data values. A database state space is the set of all possible database states. A database state space S_{DB} is a *metric space* if it has the following properties:

- A distance function $distance(u, v)$ is defined over every pair of states $u, v \in S_{DB}$ on real numbers.

  The distance function can be defined as the absolute value of the difference between two states of an account data item. For instance, the distance between \$50 and \$120 is \$70. Thus, if the current account balance is \$50 and \$70 is credited, the distance between the new state and the old state is \$70.

- Symmetry. For every $u, v \in S_{DB}$, $distance(u, v) = distance(v, u)$.

  Continuing with the example, suppose, the current account balance is \$120 and \$70 is debited. The distance between the new state and the old state is still \$70.

- Triangle inequality. For every $u, v, w \in S_{DB}$, $distance(u, v) + distance(v, w) \geq distance(u, w)$.

  The account data clearly satisfies triangle inequality. For example, suppose the current account balance is \$50 and \$70 is credited. The distance between the new state and

> the old state, as we saw before is \$70. Suppose \$40 is now debited. The distance between the state after the credit and the state after the debit is \$40. The distance between the initial state of the account (\$50) and the one after both updates (\$80) is \$30. Since \$70 + \$40 $\geq$ \$30, triangle inequality is satisfied.

Many database state spaces have such a regular geometry. As we just saw, in banking databases, dollar amounts possess these properties. Similarly, airplane seats in airline reservation systems also form a metric space.

Usually the term "database state space" refers to the state on disk (implicitly, only the committed values). We are not restricted to the database state on disk, however, since we also consider the intermediate states of the database, including the contents in the main memory. We will use the shorter term "data state" to include the intermediate states. Note that the magnitude of an update can be measured by the distance between the old data item state and the new data item state.

ESR defines correctness for both consistent states and inconsistent states. In the case of consistent states, ESR reduces to classic serializability. In addition, ESR associates an *amount* of inconsistency with each inconsistent state, defined by its distance from a consistent state. Informally, *inconsistency* in a data item x with respect to a query q is defined as the difference between the current value of x and the value of x if no updates on x were allowed to execute concurrently with q. A query imports inconsistency when it views, i.e., reads, an inconsistent data item. Conversely, an update transaction exports inconsistency when it updates, i.e., writes to, a data item while query ETs that read the data item are in progress.

ESR has meaning for any state space that possesses a distance function. In general, serializable executions produce answers that have zero inconsistency, but if a (non-serializable) query returns an answer that differs from a serializable result by at most \$10,000 we say that the amount of inconsistency produced by the query is \$10,000. In addition, the triangle inequality and symmetry properties help us design efficient algorithms. In this paper, we will confine our attention to state spaces that are metric spaces.

To an application designer and transaction programmer, an ET is a classic transaction with the addition of inconsistency limits. A query ET has an *import-limit*, which specifies the maximum amount of inconsistency that can be imported by it. Similarly, an update ET has an *export-limit* that specifies the maximum amount of inconsistency that can be exported by it. Since our focus is on queries, and for simplicity of presentation, we examine in detail ETs when *import-limit*s are placed on individual data items (a single attribute in the relational model). The algorithms can be extended to handle an *import-limit* that spans several attributes (e.g., checking accounts and savings accounts).

An application designer specifies the limit for each ET and the TP system ensures that these limits are not exceeded during the execution of the ET. For example, a bank may wish to know how many millions of dollars there are in the checking accounts. If this query were executed directly on the checking accounts during the banking hours, serious interference would arise because of updates. Most of the interference is irrelevant, however, since typical updates refer to small amounts compared to the query output unit, which is in millions of dollars. Hence we must be able to execute the query during banking hours. Specifically, under ESR, if we specify an *import-limit* for the query ET, for example, of $100,000, for this query, the result also would be guaranteed to be within $100,000 of a consistent value (produced by a serial execution of the same transactions). For example, if the ET returns the value $357,215,000 (before round-off) then at least one of the serial transaction executions would have yielded a serializable query result in the $325,215,000±$100,000 interval.

The inconsistency accumulated by a query that reads multiple data items, such as in the example above, depends on how the values read are used within the query. The percolation of inconsistency from the data items read by the query to the results of the query is an interesting issue and is discussed in Section 5.

Sections 3 and 4 focus on individual data items. Let us assume that limits are imposed on the amount of inconsistency an ET can import or export with respect to a particular data item. Let $import\_limit_{t,x}$ stand for the import-limit that has been set for ET t with respect to data x. Let $import\_inconsistency_{t,x}$ stand for the amount of inconsistency that has already been imported by ET t on data item x. The system that supports queries reading inconsistent data must ensure the following for every ET t (that accesses data item x):

$$\begin{aligned} import\_inconsistency_{t,x} &\leq import\_limit_{t,x} && (1)\\ export\_inconsistency_{t,x} &\leq export\_limit_{t,x}. && (2)\end{aligned}$$

We call the invariants (1) and (2) $Safe(t,x)$ for brevity. For query ET q reading x, $Safe(q,x)$ reduces to:

$$\begin{aligned} import\_inconsistency_{q,x} &\leq import\_limit_{q,x} && (3)\\ export\_inconsistency_{q,x} &= 0. && (4)\end{aligned}$$

$Safe(q,x)$ states that a query q cannot exceed its *import-limit* and that q cannot export inconsistency.

Thus, during the execution of each ET, the system needs to maintain the amount of inconsistency the ET has imported so far. Note that the amount of inconsistency is given by the distance function and the incremental accumulation of inconsistency depends on the triangle inequality property of metric spaces. Without triangle inequality, we would have to recompute the distance function for the entire history each time a change occurs. In Section 3 we derive the algorithms necessary to maintain the specified limit on the inconsistency imported from individual data items.

Before we end this section we would like to point out that throughout the paper, it is assumed that the read set of a query, i.e., the set of data items read by a query is not affected by the inconsistency in the data read by a query.

2 A Formal Definition of ESR

We use the ACTA framework [5, 4, 6] to introduce the notion of conflicts between operations and discuss the dependencies induced between transactions when they invoke conflicting transactions.

For a given state s of a data item, we use $return(s, a)$ to denote the output produced by operation a, and $state(s, a)$ to denote the state produced after the execution of a. $value(s, P)$ denotes the value of predicate P in state s.

Given a history H, $H^{(x)}$ is the projection of the history containing the operation invocations on a data item x. $H^{(x)} = a_1 \circ a_2 \circ \ldots \circ a_n$, indicates both the order of execution of the operations, (a_i precedes a_{i+1}), as well as the functional composition of operations. Thus, a state s of a data item produced by a sequence of operations equals the state produced by applying the history $H^{(x)}$ corresponding to the sequence of operations on the data item's initial state s_0 ($s = state(s_0, H^{(x)})$). For brevity, we will use $H^{(x)}$ to denote the state of a data item produced by $H^{(x)}$, implicitly assuming initial state s_0. Note that $H^{(x)}$ may depend on values read in H from data items other than x.

Definition 1 *Two operations a and b* conflict *in a state produced by $H^{(x)}$, denoted by* conflict$(H^{(x)}, a, b)$*, iff*

$$\begin{aligned}
(state(H^{(x)} \circ a,\ b) &\neq state(H^{(x)} \circ b,\ a)) \quad \vee \\
(return(H^{(x)},\ b) &\neq return(H^{(x)} \circ a,\ b)) \vee \\
(return(H^{(x)},\ a) &\neq return(H^{(x)} \circ b,\ a)).
\end{aligned}$$

Thus, two operations conflict if their effects on the state of a data item or their return values are not independent of their execution order.

Let $a_{t_i}[x]$ denote operation a invoked by t_i on data item x. $(a_{t_i}[x] \rightarrow b_{t_j}[x])$ implies that $a_{t_i}[x]$ appears *before* $b_{t_j}[x]$ in H.

Let us first define the classic serializability correctness criterion.

Definition 2 *Let t_i and t_j be transactions $\in T$. Given a history H of events relating to transactions in T, $\mathcal{C}_{SR}$, a binary relation on T, is defined as follows:*

$$(t_i\ \mathcal{C}_{SR}\ t_j),\ t_i \neq t_j \textit{ iff}$$
$$\exists x\ \exists a, b\ (\text{conflict}(H^{(x)}, a_{t_i}[x], b_{t_j}[x]) \wedge (a_{t_i}[x] \rightarrow b_{t_j}[x])).$$

Let $\mathcal{C}_{SR}^*$ be the transitive-closure of $\mathcal{C}_{SR}$; i.e.,

$$(t_i\ \mathcal{C}_{SR}^*\ t_j) \textit{ if } [(t_i\ \mathcal{C}_{SR}\ t_j) \vee \exists t_k\ (t_i\ \mathcal{C}_{SR}\ t_k \wedge t_k\ \mathcal{C}_{SR}^*\ t_j)].$$

H is (conflict preserving) serializable *iff*

$$\forall t \in T\ \neg(t\ \mathcal{C}_{SR}^*\ t).$$

To illustrate the practical implications of this definition, let us consider the case where all operations perform in-place updates. In this case, if transactions t_i and t_j have a $\mathcal{C}_{SR}$ relationship, i.e., t_j has invoked an operations which conflicts with a previous operation by t_i, as long as t_i is serlialized before t_j, the conflict can be tolerated. Consider the (serialization) graph corresponding to the $\mathcal{C}_{SR}$ relation induced by a history. The above definition states that for the history to be serializable, there should be no cycles in the graph. That is, the serialization order must be acyclic.

The following three definitions constitute the definition of ESR.

Definition 3 *Let t_i and t_j be transactions $\in T$ whose events are recorded in history H. $\mathcal{C}_{ESR}$, a binary relation on transactions in T, is defined as follows:*

$$(t_i\ \mathcal{C}_{ESR}\ t_j), t_i \neq t_j \textit{ iff}$$
$$\exists x\ \exists a, b\ (\text{conflict}(H^{(x)}, a_{t_i}[x], b_{t_j}[x]) \wedge (a_{t_i}[x] \rightarrow b_{t_j}[x])$$
$$\wedge\ value(state(H^{(x)} \circ a,\ b), \neg Safe(t_j, x))).$$

In other words, t_i and t_j are related by $\mathcal{C}_{ESR}$ if and only if they are related by $\mathcal{C}_{SR}$ *and* they violate one of the invariants that constitute the predicate $Safe$. Note that the last term in the definition of $\mathcal{C}_{ESR}$ makes $\mathcal{C}_{ESR}$ strictly *weaker* than $\mathcal{C}_{SR}$; if $(t_i\ \mathcal{C}_{ESR}\ t_j)$ then $(t_i\ \mathcal{C}_{SR}\ t_j)$. Just as $\mathcal{C}_{SR}$ denotes ordering requirements due to conflicts under serializability, $\mathcal{C}_{ESR}$ denotes the ordering requirements imposed by conflicts under epsilon serializability. Since $\mathcal{C}_{ESR}$ is a subset of the $\mathcal{C}_{SR}$ relationship, a smaller number of orderings are imposed under ESR than under classic serializability.

Consider the graph corresponding to the $\mathcal{C}_{SR}$ and $\mathcal{C}_{ESR}$ relations induced by a history.

Definition 4 *A cycle formed by transactions t_0, t_1, t_2, ..., t_{n-1}, has a $\mathcal{C}_{ESR}$ edge iff*

$$\exists i,\ 0 \leq i < n,\ (t_i\ \mathcal{C}_{ESR}\ t_{(i+1\ mod\ n)}).$$

As the next definition shows, (unlike SR) ESR can tolerate cycles formed by the $\mathcal{C}_{SR}$ relation. However, if the graph has a cycle consisting of a $\mathcal{C}_{ESR}$ edge, then the history is not ESR.

Definition 5 *A history H is* (conflict-preserving) epsilon serializable *iff, in the graph which corresponds to the $\mathcal{C}_{SR}$ and $\mathcal{C}_{ESR}$ relations induced by the history, there is no cycle that has a $\mathcal{C}_{ESR}$ edge.*

Before we examine the practical meaning of the above definitions, let us summarize the properties of ESR compared to serializability:

- When all *import-limit* and *export-limit* are zero, $\mathcal{C}_{ESR}$ reduces to $\mathcal{C}_{SR}$. $\mathcal{C}_{ESR}$ is then just $\mathcal{C}_{SR}$ and ESR reduces to serializability.
- A set of transactions may not satisfy serializability because of cycles in the $\mathcal{C}_{SR}$ relation, but may satisfy ESR.
- When some *import-limit*s and *export-limit*s are greater than zero, $\mathcal{C}_{ESR} \subseteq \mathcal{C}_{SR}$ (given the additional term in definition 3). That is, ESR may allow more operations to execute concurrently than serializability.

To understand the practical meaning of the definitions, let us focus on a query q executing concurrently with an update transaction t. Suppose q reads x and this is followed by t's write to x. Assume that t's write does not violate $safe(t,x)$. Thus ($q\ \mathcal{C}_{SR}\ t$) but ($q\ \mathcal{C}_{ESR}\ t$) is not true. Assume that now q does another read of x. Let us consider two scenarios:

1. Assume that q's second read does not violate $safe(q,x)$ and so ($t\ \mathcal{C}_{SR}\ q$) but not ($t\ \mathcal{C}_{ESR}$ q). So we have a cyclic $\mathcal{C}_{SR}$ relationship and yet the read is permitted by ESR. The reason for this is that, under ESR, the values of x read by q are considered acceptable, i.e., within the limits of inconsistency specified. More precisely, the value of x read by q

when concurrently executed with t is within the inconsistency limits considering either of the serialization orderings: (q, t) or (t, q). That is why no orderings are imposed by ESR, since according to ESR, both orderings are acceptable.

2. Assume that q's second read violates $safe(q,x)$. So (t $\mathcal{C}_{ESR}$ q). This imposes an ordering requirement such that it is as though q read x serially after t. Thus (t, q) is the only serialization order acceptable – in order to conform to the inconsistency limits. This implies that we cannot have (q ${\mathcal{C}_{SR}}^*$ t) since that corresponds to the opposite serialization ordering. Hence it is required that there be no cycles consisting of $\mathcal{C}_{SR}$ and $\mathcal{C}_{ESR}$ edges.

Given the above characterization of ESR, one of the first tasks is to quantify the inconsistency experienced by a query so that we can check if the *safe* predicates hold. This is done in Section 3. Then in Section 4 we examine how to ensure that only epsilon serializable histories are produced. One way is to allow no $\mathcal{C}_{ESR}$ to form, i.,e., to disallow an operation if it violates *safe*. The question of how the inconsistency in the data read by a query percolates to the the results of the query is studied in Section 5. Different types of queries are identified with a view to determining the amount of data inconsistency they can tolerate in order to maintain specified limits on result inconsistency.

3 Inconsistency Imported by a Query ET

We focus on the inconsistency of a single data item x read by a query q. Informally, *inconsistency* in x with respect to a query q is defined as the difference between the current value of x and the value of x if no updates on x were allowed to execute concurrently with q.

Consider update transactions $t_1 \ldots t_n$ where each of the t_i's updates x. We allow a query q to read x multiple times and each of the updating t_i's to write x multiple times. Let us define a transaction t_i's *write interval* with respect to x to be the interval of time between its first write and the last write. A *read interval* is defined similarly.

Every query q has a set of *Concurrent Update Transactions* (denoted by $\textsc{cut}(q)$). Update ET $t_i \in \textsc{cut}(q)$ if its write interval intersects with q's read interval. Note that lock-based realizations of serializability ensure that $\textsc{cut}(q) = \emptyset$.

The question we are attempting to answer here is the following: What can one say about the value of x read by q given the $\textsc{cut}(q)$? Our main objective is to bound the inconsistency in the value of x read by q. But first we establish that the write intervals of transactions in $\textsc{cut}(q)$ are totally ordered, since *consistent* update ETs are serializable.

Theorem 1 *The serialization order of the transactions $t_i \in \textsc{cut}(q)$, w.r.t. x, is the same as the order in which each t_i enters its write interval which in turn is the same as the order in which they commit.*

Now we name the values of x at different points in time:

- $x_{current}$ is the current value of x.
- $x^{t_i}_{final}$ is the value of x committed by transaction t_i.
- $x^{t_i}_{initial}$ is the value of x when transaction t_i in $\textsc{cut}(q)$ begins, i.e., $x^{t_i}_{initial} = x^{t_{i-1}}_{final}$.
- $x^{q}_{initial}$ is defined to be the value of x before any of the transactions in $\textsc{cut}(q)$ begin execution. That is, if $\textsc{cut}(q) \neq \emptyset$, $x^{q}_{initial} = x^{t_1}_{initial}$, else, $x^{q}_{initial} = x_{current}$.

From these values of x we can derive:

$$current\_change_{t_i,x} = distance(x_{current}, x^{t_i}_{initial})$$

$$max\_change_{t_i,x} = \max_{during\ t_i} \{current\_change_{t_i,x}\}$$

$$final\_change_{t_i,x} = distance(x^{t_i}_{initial}, x^{t_i}_{final})$$

Clearly, $final\_change_{t_i,x} \leq max\_change_{t_i,x}$ and $current\_change_{t_i,x} \leq max\_change_{t_i,x}$.

We are in a position to define inconsistency formally.

$$(x^{q}_{initial} - inconsistency_{q,x}) \leq x_{current} \leq (x^{q}_{initial} + inconsistency_{q,x})$$

That is, $inconsistency_{q,x}$ denotes the *distance* between $x^{q}_{initial}$ and $x_{current}$. So, inconsistency in the value of x for a query q while t_i is in progress and update ETs $t_1 \ldots t_{i-1}$ have already committed is given by

$$\begin{aligned} inconsistency_{q,x} &= distance(x_{current}, x^{q}_{initial}) = distance(x_{current}, x^{t_1}_{initial}) \\ &\leq distance(x_{current}, x^{t_i}_{initial}) + distance(x^{t_i}_{initial}, x^{t_1}_{initial}) \\ &\leq distance(x_{current}, x^{t_i}_{initial}) + \sum_{j=1}^{i-1} distance(x^{t_j}_{final}, x^{t_j}_{initial}) \end{aligned}$$

$$= current\_change_{t_i,x} + \sum_{j=1}^{i-1} final\_change_{t_j,x}.$$

Let $committed\_\textsc{cut}(q)$ denote the subset of $\textsc{cut}(q)$ containing the ETs that have committed. Let $t_{current} \in \textsc{cut}(q)$ denote the update transaction whose write interval has begun but has not ended yet. If no such $t_{current}$ exists, it has a "null" value and $current\_change_{null,x}$ is defined to be 0.

From these discussions we can state the following theorem which expresses (bounds on) the inconsistency of a data item read by a query q when its read interval intersects with the write intervals of ETs in $\textsc{cut}(q)$.

Theorem 2

$$inconsistency_{q,x} = distance(x_{current}, x^q_{initial}) \tag{5}$$

$$\leq \sum_{t_j \in committed\_\textsc{cut}(q)} final\_change_{t_j,x} + current\_change_{t_{current},x} \tag{6}$$

$$\leq \sum_{t_j \in committed\_\textsc{cut}(q)} final\_change_{t_j,x} + max\_change_{t_{current},x} \tag{7}$$

$$\leq \sum_{t_j \in committed\_\textsc{cut}(q)} max\_change_{t_j,x} + max\_change_{t_{current},x} \tag{8}$$

Whereas expression (5) is an exact expression of the inconsistency, expressions (6) through (8) can be viewed as different *bounds* on $inconsistency_{q,x}$.

We are now in a position to relate the inconsistency bound with the conflict-based definition of ESR given in Section 2. Recall the definitions of $\mathcal{C}_{SR}$ and $\mathcal{C}_{ESR}$:

> A pair of transactions have a $\mathcal{C}_{SR}$ relationship but not a $\mathcal{C}_{ESR}$ relationship iff one of them is a query and the other is an update *and* the import limits are not violated. Let us focus on $\mathcal{C}_{SR}$ relationships induced by operations on x. Given (8), each of the update transactions t_i that appears in the pairs that belongs to $\mathcal{C}_{SR}$ but not to $\mathcal{C}_{ESR}$ contributes an inconsistency of at most $max\_change_{t_i,x}$ to the value of x read by q.

So far we have considered the case when all transactions commit. As stated by the following theorem, abortion of update transactions has the effect of increasing the inconsistency imported by a query without changing the value of x.

Theorem 3 *The maximum* increase *in imported inconsistency caused by aborted transactions is given by*

$$\max_{t \in CUT(q)\ aborted} \{max\_change_{t,x}\}.$$

Proof: Suppose transactions t_1 to t_{i-1} have committed and then t_i begins but subsequently aborts. In addition to the inconsistency due to t_1 to t_{i-1}, derived earlier, if q reads x any time during t_i's execution, it will experience an additional inconsistency of $max\_change_{t_i,x}$. Assume t_i aborts whereby changes made by t_i are obliterated and thus subsequent updates will increase the value of x only with respect to that resulting from t_1 to t_{i-1}.

Suppose all the transactions in $CUT(q)$ that follow t_i commit. Then $max\_change_{t_i,x}$ is the only increase to the inconsistency due to aborted transactions and hence the theorem holds.

Suppose instead that t_{i+1} to t_{j-1} commit and t_j aborts. When q reads x after t_j begins, x will only reflect the changes done by (1) transactions t_1 to t_{i-1}, (2) transactions t_{i+1} to t_{j-1}, and (3) transaction t_j. (3) is bounded by $max\_change_{t_j,x}$. If this is larger than $max\_change_{t_i,x}$, then $max\_change_{t_j,x}$ is the increase in inconsistency due to the aborted transactions t_i and t_j and hence the theorem follows for two transaction aborts. If this is smaller, $max\_change_{t_i,x}$ remains the upper bound on the increase. That is, the maximum of the two is the effective increase in inconsistency due to two transaction aborts. This proof extends easily if further transactions abort.

4 Ensuring Epsilon Serializability: Pre-Conditions for ET Operations

To make sure that all histories are ESR (as per Definition 4) we should ensure that no cycles are formed with $\mathcal{C}_{ESR}$ edges in them. But what if we do not even allow $\mathcal{C}_{ESR}$ relations to form? Just like SR can be realized by preventing the formation of serialization orderings (i.e., $\mathcal{C}_{SR}$ relations), ESR can be realized by preventing the formation of $\mathcal{C}_{ESR}$ relations). Thus, if we ensure that a query is always *safe*, i.e, ($import\_inconsistency_{q,x} \leq import\_limit_{q,x}$) is an invariant, then ESR is guaranteed. Specifically, the inequality must hold (before and) after every read and write operation as well as every transaction transaction management event. We derive the preconditions for performing the operations. These are sufficient to ensure that import limits of transactions are not exceeded. The preconditions will in turn be used to show how the transaction executions should be managed.

Let $begin\_write_{t,x}$ denote the attempt by ET t to begin its write interval with respect to x. $begin\_read_{t,x}$ is invoked by t to begin its read interval with respect to x. Let $end\_write_{t,x}$ denote that t has completed its writes on x. We will now consider the semantics of *begin_write*, *begin_read*, *end_write*, *end_read*, *read* and *write*. There are two situations to consider. The first is if a query ET q is already in progress (initially with $committed\_\textsc{cut}(q) = \emptyset$) when an update transaction's write interval begins. This may be followed by other update ETs before q commits. The second is if an update ET is in progress when the query begins.
Recall that our attention is confined to a centralized database with a single transaction manager.
Let q be a query and t be an update ET. ← stands for assignment.
If query q is in progress,

$$\begin{aligned} begin\_write_{t,x} &\equiv (t_{current} \leftarrow t) \wedge (\textsc{cut}(q) \leftarrow \textsc{cut}(q) \cup t) \\ end\_write_{t,x} &\equiv (t_{current} \leftarrow null) \wedge (committed\_\textsc{cut}(q) \leftarrow committed\_\textsc{cut}(q) \cup t) \end{aligned}$$

Otherwise, $begin\_write_{t,x} \equiv ()$ and $end\_write_{t,x} \equiv ()$.
If an update transaction t is in progress, $begin\_read_{q,x} \equiv (t_{current} \leftarrow \mathrm{t}) \wedge (\textsc{cut}(q) \leftarrow t)$.
Otherwise, $begin\_read_{q,x} \equiv (t_{current} = null)$.
Here are the semantics of the other operations.

$$\begin{aligned} end\_read_{q,x} &\equiv (q \leftarrow null) \\ read_{t,x} &\equiv () \\ read_{q,x} &\equiv (import\_inconsistency_{q,x} \leftarrow inconsistency_{q,x}) \\ write_{t,x}(\Delta) &\equiv (x_{current} \leftarrow x_{current} + \Delta) \end{aligned}$$

Δ is a parameter to the *write* operation that denotes the amount by which x is modified when the write occurs.

It is important to note from the above semantics that a query imports inconsistency only if it performs a read operation. That is, the inconsistency in the value of x due to updates translates to imported inconsistency only when read operations occur.

We will now establish the preconditions necessary to maintain (3), i.e.,

$$(import\_inconsistency_{q,x} \leq import\_limit_{q,x}) \tag{9}$$

Case 1: Preconditions only on $read_{q,x}$ Operations.

Given that inconsistency is imported by q only when it performs a *read*, the following precondition is all we need to maintain (9):

$$inconsistency_{q,x} \leq import\_limit_{q,x}.$$

From (5), this implies the precondition

$$distance(x_{current}, x^{q}_{initial}) \leq import\_limit_{q,x}.$$

Every *read* operations must be intercepted by the transaction management mechanism to ensure that the above precondition holds. If the predicate does not hold, the read by the query will have to be aborted or delayed. If q is a long query, this has performance implications. This is the motivation for examining other possible ways to maintain (9).

Case 2: Preconditions on $write_{t,x}$ Operations and $begin\_read_{q,x}$ Operations

Suppose we satisfy the following invariant:

$$inconsistency_{q,x} \leq import\_limit_{q,x},$$

i.e.,

$$distance(x_{current}, x^{q}_{initial}) \leq import\_limit_{q,x}$$

Note that this is a stronger invariant than (9), i.e, if this is maintained, then (9) will be maintained. (This has a negative side-effect: If the query does not read x at all, then the allowable inconsistency on x has been restricted unnecessarily.) Given the semantics of the various operations, and the expression (5) for inconsistency, the following precondition on *write* results.

$$distance(x_{current} + \Delta, x^{q}_{initial}) \leq import\_limit_{q,x}$$

and given that x is in metric space, this implies the precondition

$$|\Delta| + distance(x_{current}, x^{q}_{initial}) \leq import\_limit_{q,x}$$

where $|\Delta|$ denotes the *absolute value* of Δ. (We also use $\mid S \mid$ to denote the cardinality of set S. The meaning should be obvious from the context.) This says that a write should be allowed only if the increase in inconsistency caused by the intended increment will not violate the limit imposed on the inconsistency imported by q.

Even though no precondition is necessary for a *read*, the following precondition is required for $begin\_read_{q,x}$ when it is invoked while an update transaction t is already in progress:

$$distance(x_{current}, x^{t}_{initial}) \leq import\_limit_{q,x}.$$

Note that $x^{q}_{initial} = x^{t}_{initial}$ when q begins its read interval while t's writes are in progress. This says that if the changes that have already been done by the update transaction exceed the import limit imposed on q then the query must not be allowed to begin its read on x.

The above preconditions imply that for each query q, we should maintain $x^{q}_{initial}$. This can be avoided by maintaining an even stronger invariant, corresponding to the inconsistency bound (6), i.e., by maintaining

$$\sum_{t_j \in committed\_CUT(q)} final\_change_{t_j,x} + current\_change_{t_{current},x} \leq import\_limit_{q,x}.$$

This imposes the following precondition on $write_{t,x}$:

$$\sum_{t_j \in committed\_CUT(q)} final\_change_{t_j,x} + current\_change_{t_{current},x} + |\Delta| \leq import\_limit_{q,x}$$

and the following precondition on $begin\_read_{q,x}$:

$$current\_change_{t_{current},x} \leq import\_limit_{q,x}.$$

This implies that write operations by update ETs and requests by query ETs to begin their reading have to be monitored to ensure that they are allowed only when the above preconditions hold.

Both these invariants require maintenance of the most recent committed state of x. This is available anyway. However, the need to check every *write* by an update ET implies increased overheads and may also result in aborts or delays of update ETs in progress. Both can be avoided as shown below if an even stronger invariant is maintained.

Case 3: Preconditions on $begin\_read_{q,x}$ and $begin\_write_{t,x}$ Operations

Consider the following invariant corresponding to inconsistency bound (7):

$$\sum_{t_j \in committed\_CUT(q)} final\_change_{t_j,x} + max\_change_{t_{current},x} \leq import\_limit_{q,x}.$$

This inequality turns out to be the precondition for $begin\_write_{t,x}$. $begin\_read_{q,x}$ has the following precondition:

$$max\_change_{t,x} \leq import\_limit_{q,x}. \tag{10}$$

This implies that unlike the previous case, no preconditions are associated with *individual* writes by update transactions. While this reduces transaction management overheads, it does introduce some pessimism into the decision making since worst case changes to x by t are assumed.

The precondition for $begin\_write_{t,x}$ requires knowledge about $final\_change$ of transactions. This can be avoided if the following invariant, corresponding to inconsistency bound (8), is maintained:

$$\sum_{t_j \in committed\_\textsc{cut}(q)} max\_change_{t_j,x} + max\_change_{t_{current},x} \leq import\_limit_{q,x} \tag{11}$$

(11) is also the precondition for $begin\_write_{t,x}$. (10) stays as the precondition for $begin\_read_{q,x}$.

Suppose $max\_change_{t_i,x}$ is the same for all update ETs t_i. Then, a given $import\_limit_{q,x}$ for a query q translates into a limit on the *cardinality* of $\textsc{cut}(q)$.

In terms of the impact of the above derivation on an implementation of ESR, note that we progressed from preconditions on individual read and write operations to preconditions for read and write intervals to begin. The latter introduce more pessimism, because of the the assumptions that have to be made about the amount of changes done by a given update transaction.

Modeling query and transaction executions in terms of their read and write intervals allows us to capture different types of concurrency control techniques. For instance, if the *begin* events correspond to the acquisition of locks and the *end* events correspond to the release of locks, we get lock based protocols. Assume we use the preconditions on these events to ensure bounds. This is the basis for the lock-based implementation in [29] wherein precondition (11) for *begin_write* corresponds to LOK-2 and precondition (10) for *begin_read* corresponds to LOK-1.

However, the above derivation is not restricted to lock-based implementations. In optimistic concurrency control, writes are done after the validation phase. In this case, precondition checking for writes will be part of the validation phase of an update transaction.

5 Inconsistency in the Results of a Query

Since a query, by definition, does not update data, it does not affect the permanent state of the database. Furthermore, we have assumed that updates do not import inconsistency, i.e., they operate on consistent database states. Thus, assuming that each update ET maintains database consistency, updates also do not affect the consistency of the database. The only effect of the updates is on the inconsistency of the data read by queries. In Section 3 we derived expressions for the amount of inconsistency imported by a query. Given this inconsistency, the only *observable* effect of a query ET is on the results produced by a query. In other words, the inconsistency imported by a query can *percolate* to the results of a query, in ways that are obviously dependent on the manner in which the query utilizes the values read.

This section is devoted to determining the effect of the inconsistency of data read by a query on its results. In general, a small input inconsistency can translate into an arbitrarily large result inconsistency. Therefore, we study the properties of a query that make the result inconsistency more predictable.

First we establish some terminology. Consider the situation where a query q reads data items $x_1, x_2, \ldots, x_n$ and produces a result based on the values read. In general, the results of such a query can be stated as a function of the form:

$$g(f_1(x_1), f_2(x_2), \ldots, f_n(x_n)) \tag{12}$$

where g denotes a query ET and f_i's are functions such that $f_i : S_{DB} \rightarrow R_f$, where R_f is the range of f_i. We assume that R_f is also a metric space. In practice, typically R_f is a subset of S_{DB}. For example, aggregate functions and queries on the database usually return a value in S_{DB}.

Focusing on *monotonic* queries, in Section 5.1 we derive the inconsistency in the result of a query and show that even though the inconsistency can be bound, the bound may not be tight. Suppose, similar to *import_limit* and *export_limit*, a limit is placed on the inconsistency in the result of a query. In Section 5.2, we derive the preconditions on ET operations imposed by such a limit. In Section 5.3 a class of queries called *bounded* queries is considered. Section 5.4 examines *steady* queries and discusses how queries can be designed to have tighter inconsistency bounds thereby requiring less restrictive preconditions.

5.1 Monotonic Queries

The first important class of queries consists of *monotonic* functions. A function f is *monotonically increasing* if $x \leq y \Rightarrow f(x) \leq f(y)$. A function g is *monotonically decreasing* if $x \leq y \Rightarrow f(x) \geq f(y)$. A function is called *monotonic* if it is either monotonically increasing or decreasing. Without loss of generality in the rest of this section we describe only monotonically increasing functions.

The result returned by a monotonic ET q assuming that the value of x_i read by q is given by $x_{i,read}$ is

$$g(f_1(x_{1,read}), f_2(x_{2,read}), \ldots, f_n(x_{n,read}))$$

where, if $max\_inconsistency_{x_i}$ is the maximum inconsistency in the value of x_i read by q (given by Theorem 2 of Section 3), $x_{i,initial}$ is the value of x_i when the first update ET in $\textsc{cut}(q)$ begins, and $x_{min} = x_{i,initial} - max\_inconsistency_{x_i}$ and $x_{max} = x_{i,initial} + max\_inconsistency_{x_i}$, then

$$x_{i,min} \leq x_{i,read} \leq x_{i,max}. \tag{13}$$

Thus, since g and the f_i's are monotonic, the result of the query can lie between

$$min\_result_q = g(f_1(x_{1,min}), \ldots, f_n(x_{n,min})) \tag{14}$$

and

$$max\_result_q = g(f_1(x_{1,max}), \ldots, f_n(x_{n,max})) \tag{15}$$

Note that if f_i is not monotonic, the smallest (largest) value of f_i need not correspond to the smallest (largest) value of x_i.

Thus, by our definition of inconsistency,

$$result\_inconsistency_q = \frac{(max\_result_q - min\_result_q)}{2}. \tag{16}$$

Let us look at some examples:

Example 1: n=1; $g = f_i =$ the identity function. This corresponds to the single data element case and hence the inconsistency in the result of q can be seen to be given by (13).

Example 2: n=20; $g = \sum_{i=0}^{20}$; $f_i =$ the identity function. In this case, as one would expect, the result of the query, according to (14) and (15), will lie between $\sum_{i=0}^{20}(x_{i,initial} - max\_inconsistency_{x_i})$ and $\sum_{i=0}^{20}(x_{i,initial} + max\_inconsistency_{x_i})$.

Example 3: n=20; $g = \sum_{i=0}^{20}$; $f_i = ((x_i > 5000) \times x_i)$. (A predicate has a value 1 if it is true, otherwise 0.) In this case, the result of the query, according to (14) and (15), will lie between

$$\sum_{i=0}^{20}(((x_{i,initial} - max\_inconsistency_{x_i}) > 5000) \times (x_{i,initial} - max\_inconsistency_{x_i}))$$

and

$$\sum_{i=0}^{20}(((x_{i,initial} + max\_inconsistency_{x_i}) > 5000) \times (x_{i,initial} + max\_inconsistency_{x_i})).$$

Example 4: This is a concrete case of Example 3. Consider a bank database with 20 accounts, numbered 1-20. Each account with an odd number happens to have \$5,001 and even-numbered accounts have \$4,999. The only update transaction in the system is: Transfer(Acc_i, Acc_j, 2), which transfers \$2 from Acc_i into Acc_j. The query ET sums up all the deposits that are greater than \$5,000. Suppose that the first set of transactions executed by the system are: Transfer(Acc_{2i-1}, Acc_{2i}, 2), for i=1, ... , 10. When these finish, the following are executed: Transfer(Acc_{2i}, Acc_{2i-1}, 2), for i=1, ... , 10.

These update transactions maintain the total of money in the database, and it is easy to see that a serializable execution of the query ET should return \$50,010, since at any given time, exactly 10 accounts have more than \$5,000.

This query will produce a result between \$0 and \$100,080 since it is exactly Example 3, where,

$$\forall i = 1, \ldots, 10, x_{(i*2)-1,initial} = \$5,001.$$
$$\forall i = 1, \ldots, 10, x_{(i*2),initial} = \$4,999.$$
$$\forall i = 1, \ldots, 20, max\_inconsistency_{x_i} = 4.$$

The range of the result does include the serializable result of \$50,010. However, given that the range is not very "tight", it is too pessimistic. This occurs because the inconsistency caused by the updates percolate, in a rather drastic manner, to the results of the query. In Section 5.4, we identify a class of queries for which tight bounds on the results of a query exist.

One other point to note here is that even this bound requires knowledge of $x_{i,initial}$, the value of x_i when the first ET in $\textsc{cut}(q)$ begins. This has practical implications. Specifically, before an update is begun, the data values may have to be logged in order to derive the inconsistency for the queries that may subsequently begin. This is the case of systems that require UNDO capability (using the STEAL buffering policy [12]).

Given that the lower bound on the result of the above query is 0, one may be tempted to take the following solution: Assume that $x_{i,initial}$ is the smallest value x_i can take, i.e., 0. It is not too difficult to see why this will not produce the correct range for the above query's result.

5.2 Pre-Conditions for Monotonic Queries

Suppose $result\_inconsistency\_limit_q$ denotes the maximum inconsistency that an application can withstand in the result of a query q. Then

$$result\_inconsistency_q \leq result\_inconsistency\_limit_q$$

is an invariant. Just as we derived preconditions to maintain $import\_limit_{q,x}$ and $export\_limit_{q,x}$, we can derive preconditions to maintain the above invariant.

For instance, consider the expression (8) for $max\_inconsistency_x$. From this, given (16) and the semantics of ET operations (see Section 3), we have the following precondition for $begin\_write_{t,x_i}$:

$$\frac{1}{2}\left(g(\ldots, f_i(x_{i,initial} + (\sum_{t_j \in committed\_\mathrm{CUT}(q)} max\_change_{t_j,x_i} + max\_change_{t,x_i})), \ldots)\right) -$$
$$\frac{1}{2}\left(g(\ldots, f_i(x_{i,initial} - (\sum_{t_j \in committed\_\mathrm{CUT}(q)} max\_change_{t_j,x_i} + max\_change_{t,x_i})), \ldots)\right) \leq$$
$$result\_inconsistency\_limit_q$$

and the following precondition for $begin\_read_{q,x_i}$:

$$\frac{1}{2}\left(g(\ldots, f_i(x_{i,initial} + max\_change_{t,x_i}), \ldots) - g(\ldots, f_i(x_{i,initial} - max\_change_{t,x_i}), \ldots)\right) \leq$$
$$result\_inconsistency\_limit_q$$

In a similar manner, preconditions can be derived in case the other expressions for inconsistency are used.

5.3 Bounded Queries

We say that a function f is *bounded* if there is a maximum bound in the result of f. It is easy to see that we can calculate bounds on the inconsistency in the results of a query composed from bounded functions.

Example 5: Consider the following variation of Example 4. The query ET sums up all the deposits that are *not* greater than \$5,000. For this query, n=20; $g = \sum_{i=0}^{20}$; $f_i = ((x_i \leq 5000) \times x_i)$. The f_i's are not monotonic because when x_i increases from \$4999 to \$5001, f_i decreases from \$4999 to \$0. So the expressions derived for $result\_inconsistency$ in Section 5.2 do not apply.

It is easy to see that a serializable execution of the query ET should return \$49,990, since at any given time, exactly 10 accounts have balance $\leq$ \$5,000. It is also not difficult to see that for the above ET query, the smallest possible result is \$0 and the largest possible result is \$99,980.

Even though the the f_i's are not monotonic, we now show that it is possible to obtain bounds on the query results. Let min_f_i denote the smallest value of f_i for any value of x_i in $(x_{i,min},\ x_{i,max})$ and let max_f_i denote the largest value of f_i for any value of x_i in $(x_{i,min},\ x_{i,max})$. Then as long as g is monotonic, the result of the query can lie between $g(min\_f_1, \ldots, min\_f_n)$ and $g(max\_f_1, \ldots, max\_f_n)$.

Let us return to Example 5. In this case,

$$\begin{aligned} \forall i = 1, \ldots, 10, x_{(i*2)-1,min} &= \$4,997. \\ \forall i = 1, \ldots, 10, x_{(i*2)-1,max} &= \$5,005. \\ \forall i = 1, \ldots, 10, x_{(i*2),min} &= \$4,995. \\ \forall i = 1, \ldots, 10, x_{(i*2),max} &= \$5,003. \end{aligned}$$

min_f_i = 0 and max_f_i = \$5,000 and hence, the result of the query can lie between \$0 and \$100,000. Since the actual result of the query lies between \$0 and \$99,980, using the maximum and minimum possible f_i values leads to an overestimate of the inconsistency in the query results.

A generalization of bounded functions and monotonic functions is the class of functions *of bounded variation*. To avoid confusion for readers familiar with mathematical analysis, we follow closely the usual definition of these functions in compact metric spaces.

Definition 6 *If $[a, b]$ is a finite interval in a metric space, then a set of points*

$$P = \{x_0, x_1, \ldots, x_n\}$$

satisfying the inequalities $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ is called a partition *of $[a, b]$. The interval $[x_{k-1}, x_k]$ is called the k^{th} subinterval of P and we write $\Delta x_k = x_k - x_{k-1}$, so that $\sum_{k=1}^{n} \Delta x_k = b - a$.*

Definition 7 *Let f be defined on $[a, b]$. If $P = \{x_0, x_1, \ldots, x_n\}$ is a partition of $[a, b]$, write $\Delta f_k = f(x_k) - f(x_{k-1}), k = 1, 2, \ldots, n$. If there exists a positive number M such that*

$$\sum_{k=1}^{n} |\Delta f_k| \leq M$$

for all partitions of $[a, b]$, then f is said to be of bounded variation *on $[a, b]$.*

It is clear that all bounded functions are of bounded variation. In Example 5, $M = 5000$. Furthermore, all monotonic functions are also of bounded variation. This happens because for a monotonically increasing function f we have $\Delta f_k \geq 0$ and therefore:

$$\sum_{k=1}^{n} |\Delta f_k| = \sum_{k=1}^{n} \Delta f_k = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] = f(b) - f(a) = M.$$

In general, for a function of bounded variation, the M bound can be used as an (over)estimate of result inconsistency given the interval $[a, b]$ caused by input inconsistency. However, the examples above show that what we need is to restrict the forms of ET queries such that tighter bounds on result inconsistency can be found without overly restricting the type of queries allowed.

5.4 Steady Queries

Let DS denote the set of distances defined by S_{DB} and DR the set of distances defined by R_f. We say that f is *steady* if for every $\epsilon \in DR, \epsilon > \epsilon_0 \geq 0$ we can find a $\delta \in DS, \delta > 0$ such that $|f(x) - f(x + \delta)| \leq \epsilon$. Steady functions on discrete metric spaces are analogous to continuous functions on compact sets. The definition is similar, except that we exclude a fixed number of small ϵ due to the discrete nature of S_{DB}. Informally, if $\epsilon < \epsilon_0$ we allow δ to be zero.

The importance of steady functions is that the application designer may specify a limit on the result inconsistency, *result_inconsistency_limit* (ϵ), and the TP system can calculate the limit on the imported inconsistency, *max_inconsistency* (δ), that guarantees the specified limit on the result inconsistency. Section 5.2 shows how this calculation can be done for monotonic functions. Note that every monotonic function can be steady with a convenient choice of ϵ_0. However, the smaller is the ϵ_0 the tighter is the bound on δ. In the following example, the bound is tight because $\epsilon_0 = 0$.

Example 6: Consider a query ET that returns the balance of a bank account. If an update is executing, say transferring some money into the account, then the query result inconsistency is equal to imported inconsistency and $\delta = \epsilon$.

For an example where ϵ_0 is large, consider Example 4. When an account balance is actually 5000, an input inconsistency of 1 may change the result by 5000. Therefore we have $\epsilon_0 = 5000$, since a smaller ϵ requires $\delta = 0$.

One way to handle such a situation is to reduce or eliminate the imported inconsistency in the data item that causes a large ϵ_0. For instance, suppose that $q = g(f_1(x_1), f_2(x_2))$ and that a large ϵ_0 is due to x_1. We should tighten the *import_limit* for x_1 and allow inconsistency only for x_2. Consider the following example which is a simple variation of Example 4.

Example 7: The query ET returns the checking account balance of customers that have savings accounts with balance greater than \$5,000. Note that in this example, x_1 refers to the savings account and x_2 to the checking account. In this case, we may specify *import_limit* = 0 for the savings account balance and *import_limit* = \$100 for the checking account balance. This way, we avoid the large ϵ_0 with respect to x_1 but maintain the tight control over result inconsistency since the function that returns the checking account balance is a steady function with $\epsilon_0 = 0$ (from Example 6).

Being able to calculate ϵ from δ and vice-versa are properties of ET queries that allow the system to maintain tight bounds on result inconsistency. Functions of bounded variation and steady functions are abstract classes of functions that have these properties. Clearly, more elaborate characterization of these functions defined on discrete metric spaces will be useful.

6 Related Work

6.1 General Weak Consistency Criteria

Several notions of correctness weaker than SR have been proposed previously. A taxonomy of these correctness criteria is given in [23]. Here we contrast those that are closely related to ESR with ESR.

Gray's different degrees of consistency [11] is an example of a coarse spectrum of consistency. Of specific interest to us is degree 2 consistency which trades off reduced consistency for higher concurrency for queries. Since degree 2 allows unbounded inconsistency, degree 2 queries become less accurate as a system grows larger and faster. In general, ESR offers a much finer granularity control than the degrees of consistency.

Garcia-Molina and Wiederhold [10] have introduced the *weak consistency* class of read-only transactions. In contrast to their WLCA algorithm, ESR is supported by many divergence control methods [29]. Similarly, Du and Elmagarmid [7] proposed quasi-serializability (QSR). QSR has limited applicability because of the local SR requirements despite unbounded inconsistency. Korth and Speegle [16] introduced a formal model that include transaction pre-conditions and post-conditions. In contrast, ESR refers specifically to the amount of inconsistency in state space.

Sheth and Rusinkiewicz [26] have proposed *eventual consistency*, similar to identity connections introduced by Wiederhold and Qian [28], and *lagging consistency*, similar to asynchronously updated copies like quasi-copies [1]. They discuss implementation issues in [24, 25]. In comparison, ESR achieves similar goals but has a general approach based on state space properties and functional properties. Barbara and Garcia-Molina [2] proposed *controlled inconsistency*, which extends their work on quasi-copies [1]. Their demarcation protocol [3] can be used for implementing ESR in distributed TP systems. ESR is applicable to arithmetic and other kinds of consistency constraints.

6.2 Asynchronous Transaction Processing

Garcia-Molina et al. [9] proposed *sagas* that use semantic atomicity [8] defined on transaction semantics. Sagas differ from ESR because an unlimited amount of inconsistency (revealed before a compensation) may propagate and persist in the database. Levy et al [19] defined *relaxed atomicity* and its implementation by the Polarized Protocol. ESR is defined over state space properties and less dependent on application semantics.

An important problem in asynchronous TP is to guarantee uniform outcome of distributed transactions in the absence of a commit protocol. Unilateral Commit [13] is a protocol that uses reliable message transmission to ensure that a uniform decision is carried out asynchronously. Optimistic Commit [18] is a protocol that uses Compensating Transactions [15] to compensate for the effects of inconsistent partial results, ensuring a uniform decision. Unilateral Commit and Optimistic Commit can be seen as implementation techniques for ESR-based systems.

Another way to increase TP concurrency is Escrow Method [20]. Like the escrow method, ESR also uses properties of data state space, but ESR does not rely on operation semantics to preserve consistency. Similarly, *data-value partitioning* [27] increases distributed TP system availability and autonomy. ESR can be used in the modeling and management of escrow and partitioned data-values.

7 Conclusions

Previous ESR papers discussed ESR in informal terms by motivating it via specific applications [21, 22] and by presenting implementation-oriented considerations [29]. An evaluation of the performance improvement due to ESR is reported in [14].

In this paper, we have examined epsilon serializability (ESR) from first principles. We showed precisely how ESR is related to SR, for example, which conflicts considered by SR are ignored by ESR. A conflict based specification of ESR using the ACTA formalism was employed to bring out the differences between SR and ESR.

We began our formalization of query behavior by deriving the formulae that express the inconsistency in the data values read by a query. From these expressions we derived the preconditions, that depend on the data values and the import limits, for the read and write operations invoked by transactions and for transaction management events. In other words, from a precise definition of ETs and ESR, we have been able to derive the behavioral specifications for the necessary transaction management mechanisms. These form the second contribution of this paper. Results showed that more flexible transaction management techniques, than the ones discussed previously, are possible.

Another important aspect of this paper is the derivation of expressions for the inconsistency of the *results of queries.* We showed that since arbitrary queries may produce results with large inconsistency, it is important to restrict ET queries to have certain properties that permit tight inconsistency bounds. Towards this end, we came up with different types of queries that allow us to bound the result inconsistency, and in some cases, to find tight bounds as well. Clearly, more work is needed in this area since generality of the queries has to be traded off against the tightness of the result inconsistency.

Among the other active topics of research is the formal treatment of general ETs that both import and export inconsistency. Also, the effect of relaxing some of the assumptions, for instance, that read set of a query is unaffected by the inconsistency, needs to be studied.

Acknowledgements

The authors thank P. Chrysanthis, H.V. Jagadish, V. Wolfe, and the referees for their comments on previous versions of this paper.

References

[1] R. Alonso, D. Barbara, and H. Garcia-Molina. Data caching issues in an information retrieval systems. *ACM Transactions on Database Systems*, 15(3):359–384, September 1990.

[2] D. Barbara and H. Garcia-Molina. The case for controlled inconsistency in replicated data. In *Proceedings of the Workshop on Management of Replicated Data*, pages 35–42, Houston, November 1990.

[3] D. Barbara, H. Garcia-Molina, The Demarcation Protocol: A Technique for Maintaining Arithmetic Constraints in Distributed Database Systems, Extending Database Technology Conference, Vienna, March 1992, in Lecture Notes in Computer Science #580, Springer Verlag, pp. 373-397.

[4] P. Chrysanthis and K. Ramamritham. A formalism for extended transaction models. In *Proceedings of the Seventeenth International Conference on Very Large Data Bases*, September 1991.

[5] P.K. Chrysanthis and K. Ramamritham. ACTA: A framework for specifying and reasoning about transaction structure and behavior. In *Proceedings of SIGMOD Conference on Management of Data*, pages 194–203, June 1990.

[6] P.K. Chrysanthis and K. Ramamritham. ACTA: The Saga continues. In Ahmed Elmagarmid, editor, *Transaction Models for Advanced Applications.* Morgan Kaufmann, 1991.

[7] W. Du and A. Elmagarmid. Quasi serializability: a correctness criterion for global concurrency control in InterBase. In *Proceedings of the International Conference on Very Large Data Bases*, pages 347–355, Amsterdam, The Netherlands, August 1989.

[8] H. Garcia-Molina. Using semantic knowledge for transactions processing in a distributed database. *ACM Transactions on Database Systems*, 8(2):186–213, June 1983.

[9] H. Garcia-Molina and K. Salem. Sagas. In *Proceedings of ACM SIGMOD Conference on Management of Data*, pages 249–259, May 1987.

[10] H. Garcia-Molina and G. Wiederhold. Read-only transactions in a distributed database. *ACM Transactions on Database Systems*, 7(2):209–234, June 1982.

[11] J.N. Gray, R.A. Lorie, G.R. Putzolu, and I.L. Traiger. Granularity of locks and degrees of consistency in a shared data base. In *Proceedings of the IFIP Working Conference on Modeling of Data Base Management Systems*, pages 1–29, 1979.

[12] T. Haerder and A. Reuter. Principles of transaction-oriented database recovery. *ACM Computing Surveys*, 15(4):287–317, December 1983.

[13] M. Hsu and A. Silberschatz. Unilateral commit: A new paradigm for reliable distributed transaction processing. In *Proceedings of the Seventh International Conference on Data Engineering*, Kobe, Japan, February 1990.

[14] M. Kamath and K. Ramamritham, "Performance Characteristics of Epsilon Serializability with Hierarchical Inconsistency Bounds", *International Conference on Data Engineering*, April 1993.

[15] H. Korth, E. Levy, and A. Silberschatz. A formal approach to recovery by compensating transactions. In *Proceedings of the 16th International Conference on Very Large Data Bases*, Brisbane, Australia, August 1990.

[16] H.F. Korth and G.D. Speegle. Formal model of correctness without serializability. In *Proceedings of 1988 ACM SIGMOD Conference on Management of Data*, pages 379–386, May 1988.

[17] N. Krishnakumar and A.J. Bernstein. Bounded Ignorance in Replicated Systems. In *Proceedings of the 1991 ACM Symposium on principles of Database Systems*, pages 63-74, May 1991,

[18] E. Levy, H. Korth, and A. Silberschatz. An optimistic commit protocol for distributed transaction management. In *Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data*, Denver, Colorado, May 1991.

[19] E. Levy, H. Korth, and A. Silberschatz. A theory of relaxed atomicity. In *Proceedings of the 1991 ACM Symposium on Principles of Distributed Computing*, August 1991.

[20] P. E. O'Neil. The escrow transactional method. *ACM Transactions on Database Systems*, 11(4):405–430, December 1986.

[21] C. Pu and A. Leff. Replica control in distributed systems: An asynchronous approach. In *Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data*, pages 377–386, Denver, May 1991.

[22] C. Pu and A. Leff. Autonomous transaction execution with epsilon-serializability. In *Proceedings of 1992 RIDE Workshop on Transaction and Query Processing*, Phoenix, February 1992. IEEE/Computer Society.

[23] Ramamritham, K. and P. Chrysanthis. "In Search of Acceptability Criteria: Database Consistency Requirements and Transaction Correctness Properties" *Distributed Object Management*, Ozsu, Dayal, and Valduriez, Ed., Morgan Kaufmann Publishers, 1993.

[24] A. Sheth and P. Krishnamurthy. Redundant data management in Bellcore and BCC databases. Technical Report TM-STS-015011/1, Bell Communications Research, December 1989.

[25] A. Sheth, Yungho Leu, and Ahmed Elmagarmid. Maintaining consistency of interdependent data in multidatabase systems. Technical Report CSD-TR-91-016, Computer Science Department, Purdue University, March 1991.

[26] A. Sheth and M. Rusinkiewicz. Management of interdependent data: Specifying dependency and consistency requirements. In *Proceedings of the Workshop on Management of Replicated Data*, pages 133–136, Houston, November 1990.

[27] N. Soparkar and A. Silberschatz. Data-value partitioning and virtual messages. In *Proceedings of the Ninth ACM Symposium on Principles of Database Systems*, Nashville, Tennessee, April 1990.

[28] G. Wiederhold and X. Qian. Modeling asynchrony in distributed databases. In *Proceedings of the Third International Conference on Data Engineering*, pages 246–250, February 1987.

[29] K.L. Wu, P. S. Yu, and C. Pu. Divergence control for epsilon-serializability. In *Proceedings of Eighth International Conference on Data Engineering*, Phoenix, February 1992. IEEE/Computer Society.

Biographical Information on Krithi Ramamritham

Krithi Ramamritham received the Ph.D. in Computer Science from the University of Utah in 1981. Since then he has been with the Department of Computer Science at the University of Massachusetts where he is currently a Professor. During 1987-88, he was a Science and Engineering Research Council (U.K.) visiting fellow at the University of Newcastle upon Tyne, U.K. and a visiting professor at the Technical University of Vienna, Austria.

His current research activities deal with enhancing performance of applications that require transaction support through the use of semantic information about the objects, operations, transaction model, and the application. He is also a director of the Spring project whose goal is to develop scheduling algorithms, operating system support, architectural support, and design strategies for distributed real-time applications.

Dr. Ramamritham has served on numerous program committees of conferences and workshops devoted to databases as well as real-time systems and will serve as Program Chair for the Real-Time Systems Symposium in 1994. He is an editor of the Real-Time Systems Journal and the Distributed Systems Engineering Journal and has co-authored two IEEE tutorial texts on hard real-time systems.

Biographical Information on Calton Pu

Calton Pu received the B.S. degrees in Physics and Computer Science from the University of Sao Paulo (1979 and 1980, respectively), and the M.S. and Ph.D. degrees in Computer Science from the University of Washington (1983 and 1986, respectively).

He is presently an Associate Professor in the Department of Computer Science and Engineering at the Oregon Graduate Institute of Science and Technology. He has been doing research in transaction processing (epsilon serializability), heterogeneous databases (the Superdatabase), operating systems (the Synthesis kernel), and scientific data management using object-oriented databases and programming languages.

Dr. Pu is a member of IEEE, ACM, and AAAS.