Leopard: A Black-Box Approach for Efficiently Verifying Various Isolation Levels

Keqiang Li[1], Siyang Weng[1], Peiyuan Liu[1], Lyu Ni[1], Chengcheng Yang[1], Rong Zhang[1], Xuan Zhou[1], Jianghang Lou[2], Gui Huang[2], Weining Qian[1], Aoying Zhou[1]
East China Normal Unversity[1], Alibaba Group[2]
{kqli, syweng, pyliu}@stu.ecnu.edu.cn, {lni, ccyang, rzhang, xzhou, wnqian, ayzhou}@dase.ecnu.edu.cn,
{jianghang.loujh, qushan}@alibaba-inc.com

***Abstract*—Isolation Levels (IL) act as correct contracts between applications and database management systems (DBMSs). The complex code logic and concurrent interactions among transactions make it a hard problem to expose violations of various ILs stated by DBMSs. With the recent proliferation of new DBMSs, especially the cloud ones, there is an urgent demand for a general way to verify various ILs. The core challenges come from the requirements of: (a) lightweight (verifying without modifying the application logic in workloads and the source code of DBMSs), (b) generality (verifying various ILs), and (c) efficiency (performing efficient verification on a long running workload). For lightweight, we propose to deduce transaction dependencies based on time intervals of operations collected from client-sides without touching the source code of DBMSs. For generality, based on a thorough analysis of existing concurrency control protocols, we summarize and abstract four mechanisms which can implement ILs in all commercial DBMSs we have investigated. For efficiency, we design a *two-level pipeline* to organize and sort massive time intervals in a time and memory conservative way; we propose a *mechanism-mirrored verification* to simulate the concurrency control protocols implemented in DBMSs for high throughputs. *Leopard* outperforms existing methods by up to $114\times$ in verification time with a relative small memory usage. In practice, *Leopard* has a superpower to verify various ILs on any workload running on all commercial DBMSs. Moreover, it has successfully discovered 23 bugs that can not be found by other existing methods.**

I. Introduction

The concept of isolation level (IL) was first introduced in [1] with the name "degrees of consistency". IL serves as a correctness contract between applications and DBMSs. The strongest IL is *serializable*, but it usually exhibits a relative poor performance. A weak IL, on the other hand, offers better performance but sacrifices the guarantees of a perfect isolation. Thus, commercial DBMSs support various ILs to provide applications with a trade-off between consistency and performance [2].

Although theoretical correctness proofs have been provided for almost all ILs, the practical implementations might not strictly follow the definitions [3], [4]. Moreover, to provide both serializability and scalability in distributed systems, the distributed DBMS usually uses a consensus protocol to interact with the atomic commit protocol [5]–[7], which would lead to an extraordinarily complex protocol combination over multiple remote machines [8]–[10]. This complexity also results in a number of subtle bugs [11]–[15]. Thus, it is significantly important to perform a thorough verification of ILs when deploying critical applications on a DBMS. However, verifying various ILs has always been a tough work [11], [16]–[23] and the key challenges are summarized as follows:

Black-box Verification (C1). Most concurrency control protocols are non-deterministic, that is the final database state might be different even given the same input [24]. Thus, the traditional differential testing method [25] which checks the final database state is impractical for IL verification. Existing studies can be broadly classified into kernel-oriented and workload-oriented methods. The first type usually instruments the kernels to catch the internal execution state of DBMSs [23], [26]–[31]. However, modifying the kernels is laborious or even impossible, especially for cloud services provided by the third party. The second type relies on specific workloads to make transaction dependencies easy to obtain [11], [16]–[19]. However, these methods have a limited scope of applications, and might neglect some subtle bugs in other application scenarios. Thus, it is more attractive to design a black-box verifying method which is independent of kernels and workloads. However, verifying ILs in a black-box mode has been proven to be an NP-complete problem [32], [33].

Various Isolation Levels (C2). There are various ILs in DBMSs [34]–[38]. Even for the same IL, there might exist some subtle differences between different DBMSs. The main reason is that different DBMSs might combine different concurrency control protocols to implement these ILs. Take *repeatable read* as an example, InnoDB [39] allows *lost update* anomalies in this level of isolation, while PostgreSQL [40] and Oracle [41] do not since they use the *first updater wins* mechanism in their implementations [42]. Considering such a variety of ILs, it is rather laborious to verify ILs one by one. The differences between ILs increase the complexity of designing a general approach for IL verifications. Most of the existing work [11], [43] has been designed to only verify the strongest IL, i.e., *serializable*. Although *Elle* [19] has tried to find different IL anomalies, it could not distinguish the *repeatable read* and *serializable* in PostgreSQL [15].

Efficient Verification (C3). Many applications are suffering from isolation-related bugs that might cause data corruptions, and they might be exploited by determined adversaries to make

some mischief and profits [3], [4]. Therefore, it is necessary to timely verify each transaction executed on a DBMS, such that bugs can be reported and fixed as soon as possible. However, the online transaction processing always continuously runs with a high throughput, then massive transactions pose a significant challenge on the efficiency of IL verification. Previous studies often fail to preform an efficient IL verification on a long running workload. For example, the verification time of *Cobra* [11] grows superlinearly with the number of transactions. *Elle* can only verify ILs in an offline way.

We propose *Leopard* to address above challenges. It exhibits excellent properties of (a) *lightweight* (doing verification in a black-box mode), (b) *generality* (verifying various ILs), and (c) *efficiency* (achieving efficient verification even for a long running workload). To address **C1**, we propose to collect *interval-based traces* from client-sides without touching any source code of a DBMS or changing application logic. This method is workload-insensitive and can be applied to any DBMS. Specifically, the traces contain the execution time interval of each database operation and can be leveraged by our verification method to deduce the operation orders and transaction dependencies. To address **C2**, we summarize and abstract the implementation of various ILs into four classic mechanisms, including *consistent read*, *first updater wins*, *serialization certifier* and *mutual exclusion*. These four mechanisms constitute all ILs in commercial DBMSs we have investigated. Thus our method can be generally applied to various IL verification. To address **C3**, we first design a *two-level pipeline* to sort massive streaming traces produced by a running workload. Based on the sorted traces, we propose *mechanism-mirrored verification*, which directly simulates the internal processing of a DBMS to verify the four mechanisms. In this way, the verification process can catch up with the performance of DBMSs. Additionally, we also design some garbage collection methods to remove unnecessary structures to reduce the memory usage. It's worth noting that, *Leopard* does not guarantee the correctness of ILs since the exact execution time point of each database operation is not available in the black-box mode. Thus, we position it as a bug finding tool which verifies the test cases with best efforts. In summary, we make the following contributions.

1) We are the first work to design a lightweight IL verification framework for verifying ILs in a black-box mode.
2) We summarize and abstract the implementation of various ILs in commercial DBMSs into four mechanisms.
3) We design *two-level pipeline* and *mechanism-mirrored verification* methods to provide efficient IL verification.
4) The experiments show that *Leopard* outperforms existing methods by up to $114\times$ in verification time with a relative small memory usage. Moreover, we have successfully found 23 bugs (13 fixed, 15 confirmed and 8 open reported) which existing methods could not find.

II. Background

In this section, we first introduce the notations used in our paper. Then, we abstract and summarize four mechanisms to implement isolation levels (ILs) of commercial DBMSs.

A. Transaction Dependencies

A database $\mathbb{D}$ has a set of records. A transaction t consists of several operations typed read or write, ended with either commit or abort as a terminal operation. A write creates a new version for a record while a read queries a specific database snapshot. A database snapshot is consistent with versions created at the time of the snapshot creation. A commit installs all versions created by a transaction while an abort discards them. For a transaction t, we denote $r_t(rs)$ as a read in t with its read set rs, and denote $w_t(ws)$ as a write in t with its write set ws. Each element in rs (resp. ws) is an accessed version by r (resp. w). We denote x^i as the i^{th} version of record x, and x^{i+1} as its direct successor version.

ILs define the degree to which a transaction must be isolated from the data modifications made by any other transaction. As an isolation anomaly can be indicated by a specific transaction dependency pattern [36], the first step of IL verification is to get dependencies between all transactions. In general, there are three kinds of transaction dependencies between any two committed transactions (denoted as t_m and t_n): 1) If t_m installs a version x^i and t_n installs x^i's direct successor x^{i+1}, t_n has a **direct write-dependency** (ww) on t_m. 2) If t_m installs a version x^i and t_n reads x^i, t_n has a **direct read-dependency** (wr) on t_m. 3) If t_m reads a version x^i and t_n installs x^i's direct successor version x^{i+1}, t_n has a **direct anti-dependency** (rw) on t_m.

B. Isolation Level Implementations

DBMSs often define different ILs, e.g., *read committed* (*RC*), *repeatable read* (*RR*), *snapshot isolation* (*SI*), and *serializable* (*SR*), to prevent different isolation anomalies [34]–[38]. Specifically, there exist no unified definitions of ILs and each DBMS has its specific definition (more details can be found in [44]). To implementing these ILs, the community has devised many concurrency control protocols (*CCP*), including *optimistic concurrency control* (*OCC*) and *two-phase locking* (*2PL*), *multi-version concurrency control* (*MVCC*), *serializable snapshot isolation* (*SSI*) and *timestamp ordering* (*TO*).

After carefully investigating 18 popular DBMSs, we discover that all (*CCP*)s in these DBMSs eliminate isolation anomalies through the following four classic mechanisms: *consistent read* (*CR*), *mutual exclusion* (*ME*), *first updater wins* (*FUW*), and *serialization certifier* (*SC*). In table I, we summarize the implementations of ILs in popular DBMSs. For example, *serializable* in PostgreSQL takes all the above four mechanisms to eliminate isolation anomalies while *snapshot isolation* may suffer from *write skew* that is a kind of isolation anomaly prohibited by *SC* [40]. Note that there still exist some other mechanisms [24], [45]–[49], almost all of which stay only in the academic papers instead of in practical products.

Consistent Read (*CR*) provides a consistent view of the database at a specific time point. To eliminate the *read skew* anomaly which might lead to a transaction see an inconsistent state of the database, MVCC takes *CR* to see a snapshot of the database by maintaining multiple physical versions of

TABLE I
THE IMPLEMENTATIONS FOR ISOLATION LEVELS IN POPULAR DBMSS

DBMS	CCP	IL	CR	ME	FUW	SC
PostgreSQL [40], OpenGauss [50], yugabyteDB [51]	2PL+MVCC +SSI [40]	SR SI RC	✓ ✓ ✓	✓ ✓ ✓	✓ ✓	✓
InnoDB [39], SQL server [52], Aurora [53], PolarDB [54],	2PL+MVCC	SR,RR, RC	✓	✓		
TiDB [7]	2PL+MVCC Percolator [55]	RR,RC SI	✓ ✓	✓		✓
RocksDB [56]	2PL+MVCC OCC+MVCC	SI SI	✓ ✓	✓	✓	✓
SQLite [57]	2PL	SR		✓		
FoundationDB [58]	OCC+MVCC	SR	✓			✓
SingleStore [59]	2PL+MVCC	RC	✓	✓		
CockroachDB [6]	TO+MVCC	SR	✓			✓
Spanner [5]	2PL+MVCC	SR	✓	✓		
Oracle [41], SAP HANA [60], Oceanbase [61], NuoDB [62]	2PL+MVCC	SI RC	✓ ✓	✓ ✓	✓	

each record. Specifically, a read in MVCC sees the changes made by earlier operations within the same transaction and the changes made by other transactions committed before a specific time point. To provide different isolation levels, there exist transaction-level *CR* and statement-level *CR*. Transaction-level *CR* provides a consistent view of the database at the beginning of a transaction, while statement-level *CR* at the beginning of an operation. **Mutual Exclusion** (*ME*) uses the locking strategy to provide a kind of exclusive access a shared resource. 2PL takes the mechanism of *ME* to generate serializable histories by detecting conflicts and delaying the conflicting transactions. Specifically, for every transaction following 2PL, a phase during which locks are acquired is distinguished from and strictly followed by a phase during which locks are released. **First Updater Wins** (*FUW*) prevents transactions which modify the same record from concurrent executions. *Lost update* indicates a transaction does not see the update of another transaction when updating the same record. To eliminate the *lost update* anomaly, most DBMSs take *FUW* to ensure that transactions modifying the same record should execute in a serial order. DBMSs often combine *CR* and *FUW* mechanisms to guarantee *snapshot isolation* [40], [60]–[62]. **Serialization Certifier** (*SC*) guarantees the transaction execution is conflict serializable [63]. Many *CCP*s takes *SC* to eleminate the serialization anomaly. The anomaly might lead to a transaction schedule could not be transformed into a serial schedule by swapping non-conflicting operations. However, each *CCP* has its own "certifier". Specifically, SSI uses the detection of *write skew* as its certifier, TO uses the timestamp ordering as its certifier, and OCC uses the conflicts checking as its certifier. Due to space constraints, more details of the IL implementations can be found in [44].

III. LEOPARD FRAMEWORK

The framework of *Leopard* is illustrated in Fig. 1, which contains two components, i.e., *Tracer* and *Verifier*.

Tracer. *Tracer* continuously collects traces from each client connected to the DBMS in a black-box mode. It collects the client-side invocation and completion timestamps of each operation and without modifying any application logic. Thus,

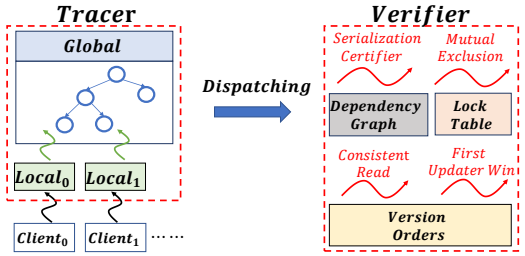
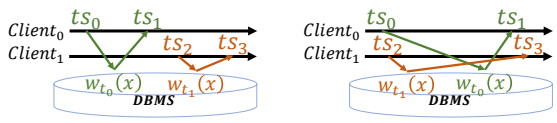


Fig. 1. Leopard Framework

it is workload-insensitive and generally applied to any DBMS. To help *Verifier* deduce the operation orders and transaction dependencies, *Tracer* needs to sort the traces according to their timestamps. As each client continuously generates its own traces individually, one naive idea is to use a min-heap to store exact one trace for each client. To dispatch a trace to *Verifier*, it pops the heap and gets the trace with globally smallest timestamp. In addition, it also fetches one new trace from the client which generates the popped trace previously. However, this would lead to high synchronization and communication costs between the clients and min-heap. *Tracer* proposes to use a *two-level pipeline* which consists of local and global buffers to address this issue. As Fig. 1 shows, the local buffers asynchronously buffer traces from each client and slice the traces into batches. The global buffer then batch fetches traces from local buffers into its min-heap round by round. Moreover, it uses a watermark to coordinate the order of trace fetching between local buffers. The watermark can also help control the size of min-heap and reduce the heap maintaining cost.

Verifier. DBMSs usually exhibit a high throughput with massive operations executed in a short period of time. As a result, it poses a significant challenge on efficient IL verification. To improve the efficiency of IL verification, previous studies usually focus on optimizing the cycle searching process on the dependency graph, such as splitting the graph into isolated segments [11] and sampling representative subsets from the graph [21]. Unfortunately, due to the inherent high complexity of cycle searching on a large graph, these methods fail to verify each operation in an efficient fashion. To this end, *Verifier* proposes to directly simulate the workflow of concurrency control protocols inside the DBMS. The main reason is that the time spent in concurrency control is much less than other components, such as the query execution and disk access. *Verifier* abstracts the implementation of various ILs into four mechanisms. In this way, verifying various ILs can be decomposed into verifying the four mechanisms. Specifically, *Verifier* tries to mirror the internal states of the DBMS, such as the version orders, lock table and dependency graph. To this end, it processes traces the same as the operation processing of a DBMS and executes each dispatched trace on these internal states. Then, *Verifier* uses these internal states to check whether there exists a violation of the four mechanisms. It's worth noting that there still exist some manual efforts for our *Leopard* framework. That is, the *Verifier* should be aware of the IL definitions of the tested DBMS, such that it can find out how to combine the four mechanisms. Fortunately,



(a) No Overlap (b) Overlap

Fig. 2. All Cases between Traces of Two Conflicting Operations

these definitions are usually described in the DBMS's official web documents. In table I, we summarize the IL definitions and the combination of the four mechanisms of popular DBMSs. In addition, the structures of these internal states would accumulate as time goes on. To save the memory usage, *Verifier* periodically prunes the obsolete structures that are not involved in the active transactions. Note that, *Leopard* can be also deployed offline. Specifically, the client only logs the time interval of each database operation and store these *interval-based traces* in the persistent storage. Then, our Leopard can sort (i.e. dispatch) and verify these traces in an offline fashion.

IV. Trace Management

In this section, we first introduce the concept of *interval-based trace* (Section IV-A). Then we discuss the opportunity of deducing transaction dependencies in a black-box mode (Section IV-B). Finally, we describe how to efficiently sort and dispatch massive traces to *Verifier* (Section IV-C).

A. Interval-based Trace

To perform IL verification in a black-box mode, we log the client-side time interval of each operation in all clients connected to the DBMS. The trace of an operation consists of 1) the timestamp before the operation executed ts_{bef}; 2) the timestamp after the operation executed ts_{aft}; 3) operation type and the data touched by the operation. Specifically, for a read (resp. write) operation, we log its belonging transaction t and its read set rs (resp. write set ws). For a commit/abort operation, we only log its belonging transaction t. Thus, the trace logging process does not need to modify the application logic and the DBMS kernel. We formalize the trace for an operation by $\mathcal{T} = \{ts_{bef}, ts_{aft}, r_t(rs)/w_t(ws)/a_t/c_t\}$.

Note that, the timestamps appearing in traces require clock synchronizations. If the test clients are deployed on a single machine (resp. multiple machines), we use its hardware (resp.software) time for clock synchronizations. Specifically, we propose to use the centralized logical timestamp generation method since it is widely used in TiDB [7] and OceanBase [61] to synchronize clocks in software mode. The detailed description and evaluation of *Leopard's* distributed deployments can be found in Appendix IV and V.B of our technical report [64].

B. Deducing Transaction Dependencies

As each IL has a specific restriction on the allowed dependency patterns [36], it is critical to determine the dependencies between transactions. The interval-based traces provide an opportunity to do this because each trace represents a specific operation issued to the DBMS. For example, consider the two conflicting operations $w_{t_0}(x)$ and $w_{t_1}(x)$ in Fig. 2. Both

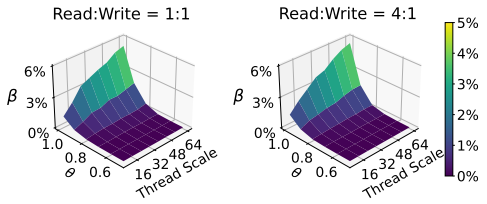


Fig. 3. Overlapping Ratio β in YCSB-A

of them create a new version on the record x. Suppose the traces of the two operations are $\mathcal{T}_0 = \{ts_0, ts_1, w_{t_0}(x)\}$ and $\mathcal{T}_1 = \{ts_2, ts_3, w_{t_1}(x)\}$. If the time intervals of the two traces do not overlap (see Fig. 2(a)), we can deduce that t_1 has a ww dependency on t_0. Otherwise, we can not determine the dependencies between t_0 and t_1 since the exact time points of the two write operations are not available. Therefore, the overlapped traces would lead to uncertain dependencies.

However, the main operations in a transactional workload are item reads/writes and the time intervals of their traces are usually very short. This indicates that the time intervals are not likely to be overlapped. To explore it practically, we run a standard benchmark *YCSB-A* [65] on PostgreSQL with a single table and 1 million records. We vary the skew parameter θ, the thread scale, and the read/write ratio to simulate different contentions and different ratios of the three dependencies. Here, we define $\beta = n_{uncertain}/n_{total}$ to represent the ratio of uncertain dependencies, where n_{total} is the total number of actual dependencies and $n_{uncertain}$ is the number of uncertain dependencies due to overlapped trace time intervals. As shown in Fig. 3, the skew parameter and thread scale are the major factors which affect the ratio β. This indicates that a high contention between operations would lead to a high ratio of uncertain dependencies. Moreover, we also observe that the value of β is kept relative small (below 6%) in all cases. This indicates that most of the dependencies can be directly deduced by the interval-based traces. In Section V, the designs in *Leopard* can further eliminate uncertain dependencies.

C. Two-Level Pipeline

The trace sorting procedure is critical for determining the operation orders and transaction dependencies. As sorting either by before timestamp or after timestamp can help check whether two time intervals overlap, we sort all traces by the before timestamp in this paper. To sort the massive traces continuously generated by multiple clients in an online fashion, we design a *two-level pipeline* which consists of several local buffers and a global buffer. The local buffers cache the streaming traces from each client asynchronously. In the meanwhile, the global buffer fetches and sorts all traces from each local buffer. The global buffer is organized as a min-heap whose time complexity increases logarithmically with the number of traces. It also uses a watermark to coordinate the order of the traces fetching from the local buffers and control the size of global buffer. Specifically, the watermark is set as the smallest before timestamp among all traces in local buffers.

Based on the *two-level pipeline*, we design a round-by-round algorithm to dispatch traces. As shown in Algorithm 1, each

Algorithm 1 Dispatching Trace

Input: n_{local} : the number of local buffers; $local_i$: the i^{th} local buffer.
Output: a trace dispatched to *Verifier*.

1: **procedure** DISPATCH()
2: $\mathcal{T}$ = $global$.top()
3: **while** $\mathcal{T}$ == null or $\mathcal{T}.ts_{bef} > watermark$ **do**
4: **for** i=0 to $n_{local} - 1$ **do**
5: fetch all of traces in $local_i$ and sort them in global buffer;
6: push the traces produced by client i into $local_i$;
7: $watermark = \min_{0 \le i < n_{local}} local_i[0].ts_{bef}$
8: **return** $\mathcal{T}$

round consists of four stages: (a) the global buffer dispatches the traces whose before timestamps are less than watermark to *Verifier* (lines 2, 8); (b) the global buffer fetches traces from the local buffers (lines 4 $\sim$ 5); (c) the global buffer updates the watermark (line 7). Note that, the traces in each local buffer are naturally sorted since the traces are generated in increasing order of before timestamps in each client; (d) each client pushes traces into its corresponding local buffer (*line* 6). Theorem 1 proves that the *two-level pipeline* is guaranteed to dispatch ordered traces to *Verifier*. Due to space constraints, we leave the formal proof of all theorems to Appendix I of our technical report [64].

Theorem 1: Algorithm 1 dispatches traces in monotonically increasing order of before timestamps.

Fig. 4 shows a running example based on two clients. In $Round_0$, the global buffer and watermark are initialized as $\varnothing$ and 1. Then, the two clients push their collected traces into the local buffers. In $Round_1$, as there exist no traces in the global buffer, it does not dispatch any trace to *Verifier*. Then, the global buffer fetches traces from each local buffer and sorts them with the min-heap. After that, the clients fill the empty local buffers with their newly collected traces. Finally, the watermark is set as the smallest before timestamp of the two local buffers, which is 3. In $Round_2$, the global buffer first dispatches all traces whose before timestamps are smaller than the watermark to *Verifier*, which are trace 1 and 2. Then, it repeats the three steps in $Round_1$, i.e., fetching traces 3, 4, 7 and 8 into the global buffer, pushing traces 9, 10, 11 and 12 into the local buffers and setting watermark as 9. Similarly, in $Round_3$, it repeats the four steps as described in $Round_2$.

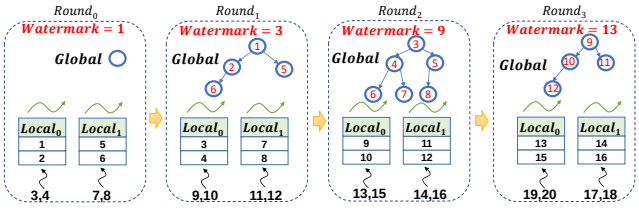


Fig. 4. Example for Dispatching Trace

Optimizations. Our trace sorting procedure also takes advantage of three optimizations, which are multi-thread processing, batch dispatching, and adaptive fetching. Specifically, to fully utilize the parallelism of modern CPUs, we propose to partition the traces, and then parallelly sort and verify the traces with multi-threads. Moreover, as we observe that there usually exist some consecutive traces in a local buffer whose before timestamps do not interleave with any trace in other local buffers, we propose a batch processing optimization which aims to directly dispatch these consecutive traces to the verifier at once. Further, as the watermark has a significant impact on the size of global buffer, we propose an adaptive method which fetches traces and controls the watermark according to the distribution of timestamps in each local buffer. The detailed descriptions of these optimizations are put in Appendix II of our technical report [64].

Algorithm 2 Mechanism-mirrored Verification

Input: *op*: operation; $\mathcal{T}$: the trace of *op*; *traces*: the trace set dispatched.
Output: *bug descriptor*

1: **procedure** CONSISTENTREAD
2: $\mathcal{S}^{\mathcal{T}} \leftarrow$ the visible snapshot time interval of *op*;
3: $OV \leftarrow$ use *traces* to construct ordered versions of each record;
4: $CV^{\mathcal{T}} \leftarrow candidate\_version\_set(OV, \mathcal{S}^{\mathcal{T}})$
5: **for** each version x^i in the read set of *op* **do**
6: **if** $x^i \notin CV^{\mathcal{T}}$ **then**
7: **return** a *CR* violation in *bug descriptor*;
8: **if** there exists only one version in $CV^{\mathcal{T}}$ matches x^i **then**
9: Deduce a wr dependency;
10: **procedure** MUTUALEXCLUSION
11: $LT \leftarrow$ use *traces* to construct a lock table for each record;
12: **for** each lock l_i released by *op* **do**
13: **for** each conflicting lock l_j in LT **do**
14: **if** each of possible orders indicates incompatible locks **then**
15: **return** an *ME* violation in *bug descriptor*;
16: **else**
17: Deduce a ww dependency;
18: **procedure** FIRSTUPDATERWINS
19: $\mathcal{S}^{\mathcal{T}} \leftarrow$ the visible snapshot time interval of *op*;
20: $OV \leftarrow$ use *traces* to construct ordered versions of each record;
21: **for** each version x^i in the write set of *op* **do**
22: **for** each version x^j in OV **do**
23: **if** each of possible orders indicates concurrent versions **then**
24: **return** an *FUW* violation in *bug descriptor*
25: **else**
26: Deduce a ww dependency;
27: **procedure** SERIALIZATIONCERTIFIER
28: **for** each dependency d deduced from $\mathcal{T}$ **do**
29: $DG \leftarrow DG \cup \{d\}$
30: **if** d causes a prohibited dependency pattern **then**
31: **return** an *SC* violation in *bug descriptor*

Complexity Analysis. The space complexity is $O(n_{local} \cdot (s_{local} + s_{client}))$, where n_{local} is the number of local buffers, s_{local} is the size of each local buffer and s_{client} is the size of traces temporally stored in each client. Note, our adaptive fetching optimization can bound the global buffer size as $O(n_{local} \cdot s_{local})$. As each trace dispatch consists of a heap push and a heap pop, the time complexity of each dispatched trace is $O(log(n_{local} \cdot s_{local}))$. Moreover, the batch processing can further reduce the time complexity to $O(\frac{1}{B} \cdot log(\frac{n_{local} \cdot s_{local}}{B}))$, where B is the average batch size. The detailed analysis and evaluation can be found in Appendix II and V-A of our technical report [64].

V. ISOLATION LEVEL VERIFICATION

In this section, based on the dispatched interval-based traces, we introduce our *mechanism-mirrored verification* to verify the four classic implementation mechanisms.

A. Verifying Consistent Read

Consistent read (*CR*) provides a consistent view of the database at a specific time point. In general, there exist two cases for the *CR* verification. In the first case, an operation sees the changes made by earlier operations within the same transaction. In the second case, an operation sees the visible changes made by other transactions. Specifically, the second case can be further classified into the transaction-level and statement-level consistent reads. The transaction-level consistent read sees the snapshot of a database as of the beginning of a transaction, while the statement-level consistent read sees the snapshot as of the beginning of an operation.

As we could not obtain the exact time point of each operation under the black-box mode, we propose a time interval based verification approach, which leverages the visible snapshot time interval of each operation and potential version evolution of each record to guide the *CR* verification. We first formally define the version installation and visible snapshot time intervals as follows.

Definition 1: **Version Installation Time Interval.** Suppose a write operation creates a new version and $\mathcal{T}$ is the operation trace. Then, the version installation time interval indicated by $\mathcal{T}$ is defined as $\mathcal{V}^{\mathcal{T}} = (\mathcal{T}.ts_{bef}, \mathcal{T}.ts_{aft})$.

Definition 2: **Visible Snapshot Time Interval.** Suppose an operation *op* sees a snapshot of the database as of an operation op_s. Then, the visible time interval of the snapshot saw by *op* is defined as $\mathcal{S}^{\mathcal{T}} = (\mathcal{T}_s.ts_{bef}, \mathcal{T}_s.ts_{aft})$, where $\mathcal{T}$ and $\mathcal{T}_s$ are traces of *op* and op_s, respectively.

The *version installation time interval* contains the exact time point when a version is created. For example, consider the trace $\mathcal{T} = \{ts_0, ts_1, w_{t_0}\}$. It indicates that a write operation in transaction t_0 creates a new version between ts_0 and ts_1, then we have $\mathcal{V}^{\mathcal{T}} = (ts_0, ts_1)$. Additionally, the *visible snapshot time interval* contains the exact time point when a specific snapshot is visible. For example, suppose there exist two consecutive reads r_t and r'_t in a transaction t. $\mathcal{T} = \{ts_0, ts_1, r_t\}$ and $\mathcal{T}' = \{ts_2, ts_3, r'_t\}$ are their corresponding traces. In transaction-level consistent read, an operation sees the snapshot as of the beginning of a transaction. Thus, both $\mathcal{S}^{\mathcal{T}}$ and $\mathcal{S}^{\mathcal{T}'}$ are set as (ts_0, ts_1). However, in statement-level consistent read, an operation sees the snapshot as of the beginning of each operation. Thus, $\mathcal{S}^{\mathcal{T}}$ and $\mathcal{S}^{\mathcal{T}'}$ are set as (ts_0, ts_1) and (ts_2, ts_3), respectively.

Next we discuss how to find *CR* violations based on the two kinds of time intervals. If all the *version installation* and *visible snapshot time intervals* do not overlap with each other, then we can directly determine the visibility of each version to a given read operation. Let's consider the first case in *CR* verification. There exists no time interval overlaps in the same transaction. Then, the *CR* mechanism can be verified by checking if a read operation sees a version which should be invisible to it. However, the read and write operations are often executed parallelly inside a DBMS, which would lead to a certain number of overlapped time intervals (see Fig.3).

To address this issue, we propose to leverage the visible snapshot time interval of a read operation to find a "candidate version set" to help with the *CR* verification process. Specifically, the candidate version set contains all the versions that are possibly visible to the read operation. If there exists no version in the candidate version set matching the read set of a read operation, we can confirm that a *CR* violation occurs. Intuitively, the size of candidate version set has a critical impact on the effectiveness of our *CR* verification. The smaller the size is, the stricter checks it poses for the *CR* verification, and the higher probability a *CR* violation can be detected.

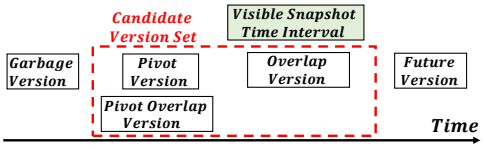


Fig. 5. Minimizing Candidate Versions

Lastly, we show how to find a minimal candidate version set. In order to efficiently prune the versions that must be invisible to the read operation, we first classify the versions into five categories according to the relationships between version installation and visible snapshot time intervals.

1) *Future version*. The version whose installation time interval appears after the visible snapshot time interval.
2) *Overlap version*. The version whose installation time interval overlaps with the visible snapshot time interval.
3) *Pivot version*. The version whose installation time interval appears before the visible snapshot time interval, and its before timestamp is closest to the before timestamp of visible snapshot time interval.
4) *Pivot overlap version*. The version whose installation time interval overlaps with that of the pivot version.
5) *Garbage version*. The version whose installation time interval appears before that of the pivot version.

Fig. 5 demonstrates which versions should be included in the candidate version set. We can see that it only consists of the *overlap versions*, *pivot versions* and *pivot overlap versions* whose installation time intervals are close to the visible snapshot time interval. Specifically, the *future version* arrives after the visible snapshot time interval, so it is invisible to the read operation. In contrast, the *garbage version* arrives before the visible snapshot time interval. However, it will be overwritten by the versions in the candidate version set. Thus, it must be invisible to the read operation. For the rest three versions, as their exact arrive times are not available in hand, either one of the three versions might be seen by the read operation. For example, if the *overlap version* arrives before the read operation, but after the *pivot version* and *pivot overlap version*, then the read operation sees the *overlap version*. As another example, the read operation sees the *pivot overlap version* (resp. *pivot version*) if that version arrives after the *pivot version* (resp. *pivot overlap version*) and the *overlap version* arrives after the read operation. We proceed to show the effectiveness of our *CR* verification approach. Theorem 2 proves our approach finds a minimal set of candidate versions.

Theorem 2: The candidate version set contains a minimum number of versions that are possibly visible to a given read

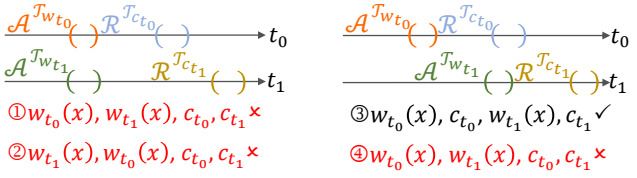
operation.

Additionally, if the candidate version set has only one version matching the read set of a read, then we can deduce that the read must happen after the write creating this matched version, even though the time intervals of the two operations are overlapped. That is, we can still confirm that the read transaction has a wr dependency on the write transaction creating this matching version. The dependencies deduced in a specific mechanism can be used by other ones to improve the effectiveness of their *IL* verification. For example, the ww dependencies deduced in the *mutual exclusive* or *first update wins* mechanism (see Sections V-B and V-C) can help the *CR* verification determine the installation order of versions with overlapped time intervals. With ordered versions, only the *overlap versions* and the last version among the *pivot overlap versions* and *pivot versions* need to be added to the candidate version set. Take the *serialization certifier* as another example, it first uses the dependencies deduced in other mechanisms to build a dependency graph, and then detects *SC* violation by checking whether an invalid dependency pattern exists.

For the reasons above, in our implementation, we verify the four mechanisms in parallel and continuously transfer the deduced dependencies between them. Note that, the dependencies deduced in each mechanism is based on the assumption that there exist no bugs inside the DBMS. However, with the cooperation of the four verification mechanisms, the potential bugs hidden inside the DBMS are likely to deduce several contradictory dependencies. Then, they would be identified as a violation of isolation levels in the bug descriptor.

The pseudo-code of the *CR* verification method is shown in Algorithm 2. Given a read operation *op*, it first uses *op*'s trace $\mathcal{T}$ to get the visible snapshot time interval (line 2). Then, it uses the traces dispatched from *Tracer* to construct the ordered versions of each record, which are sorted according to the after timestamp of their corresponding version installation time intervals (line 3). With the help of ordered versions, it generates a minimal candidate version set $CV^{\mathcal{T}}$ that contains all the versions possibly visible to *op* (line 4). Next, for each version x^i in the read set of *op*, it checks whether there exists a version in $CV^{\mathcal{T}}$ matches x^i. If not, it reports a *CR* violation in the bug descriptor (lines $6 \sim 7$). Otherwise, if there exists only one match, it can deduce that *op* has a *wr* dependency on the write which creates that version (lines $8 \sim 9$).

Complexity Analysis. The time complexity of a read operation in our *CR* verification is $O(n_r \cdot n_v)$, where n_r is the average number of versions in the read set of an operation and n_v is the average number of record versions. Note that, the construction of ordered versions for each record is carried out by the write operations. Specifically, the ordered versions of each record are stored as a sorted linked list, and each record version created by a write operation is added to the list with the insertion sort method. Thus, the time complexity of a write operation is $O(n_w \cdot n_v)$, where n_w is the average number of versions in the write set of an operation. The space complexity is $O(n_x \cdot n_v)$, where n_x is the total number of recently accessed records. As a long-running workload might change its working set and continuously create new versions for each record, we observe that n_v and n_x would increase as time goes on. To alleviate this issue, we propose to asynchronously prune garbage versions and records that do not conflict with current active transactions.



(a) Triggering a Violation (b) Deducing a ww Dependency

Fig. 6. Mutual Exclusion Verification Examples

B. Verifying Mutual Exclusion

Mutual exclusion (*ME*) uses the locking strategy to ensure exclusive accesses to the shared resources. An *ME* violation happens when a transaction acquires an incompatible lock holding by another transaction. Similar to the *CR* verification case, it is impractical to get the exact lock acquiring and releasing time in the black-box mode. We propose a time interval based approach to address this issue.

Definition 3: **Lock Acquiring and Releasing Time Intervals.** Suppose $\mathcal{T}$ (*resp.* $\mathcal{T}'$) is the trace of a lock acquiring (*resp.* releasing) operation. Then, the lock acquiring (*resp.* releasing) time interval indicated by $\mathcal{T}$ (*resp.* $\mathcal{T}'$) is defined as $\mathcal{A}^{\mathcal{T}} = (\mathcal{T}.ts_{bef}, \mathcal{T}.ts_{aft})$ (*resp.* $\mathcal{R}^{\mathcal{T}'} = (\mathcal{T}'.ts_{bef}, \mathcal{T}'.ts_{aft})$).

The lock acquiring and releasing time intervals contain the exact time points when a lock is acquired and released. On the one hand, if all the time intervals do not overlap with each other, then we can directly determine the order of lock acquiring and releasing on each record. On the other hand, for overlapped time intervals that are generally caused by parallel data accesses, we propose to leverage the mutual exclusion between locks to guide the verifying process.

Next, we discuss how to use time intervals to identify incompatible locks and deduce dependencies between transactions. Specifically, consider two transactions t_0 and t_1. Suppose they acquire two locks on the record x with two write operations $w_{t_0}(x)$ and $w_{t_1}(x)$. When t_0 and t_1 commit, their commit operations c_{t_0} and c_{t_1} would release the locks posed on x. $\mathcal{T}_{w_{t_0}}$, $\mathcal{T}_{w_{t_1}}$, $\mathcal{T}_{c_{t_0}}$ and $\mathcal{T}_{c_{t_1}}$ are the traces of these four operations. As the exact lock acquiring and releasing time points are not available in the black box mode, there might exist multiple possible orders of lock operations for any given operation traces. We broadly classify the orders into two cases.

(1) If each of the possible orders of lock operations is identified to have incompatible locks, then we can infer that there must exist an *ME* violation inside the DBMS. For example, consider the four lock operations in Fig. 6(a). There exist two possible orders of lock operations (① and ②). However, both of them are incompatible locks since a lock is acquired by two write operations simultaneously.

(2) Otherwise, from Theorem 3, we observe that we can deduce exact one ww dependency between t_0 and t_1. For example, consider the four lock operations with two possible

orders in Fig. 6(b), there exists only one order (③) in which a ww dependency can be deduced. Thus, we deduce a ww dependency (Note we assume there are no bugs inside the DBMS when deducing dependencies), and take this dependency with the dependencies deduced from other verifying mechanisms to check whether contradictory dependencies exist.

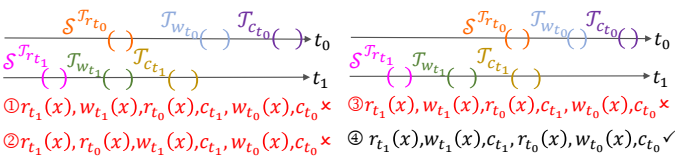
Theorem 3: Given two transactions t_0 and t_1, for any overlapped time intervals of two conflicting locks, there exists at most one possible order in which a ww dependency can be deduced. Specifically, each of the other possible orders is identified to have incompatible locks.

The pseudo-code of the *ME* verification method is shown in Algorithm 2 (lines 10 $\sim$ 17). Based on the traces dispatched from the *Tracer*, it first constructs a lock table to organize the locks on each record (line 11). The lock table contains the time intervals of lock acquiring and releasing operations. When an operation releases its previously acquired locks, the verification process first refers to the lock table and finds all locks that conflict with the released locks (lines 12 $\sim$ 13). Next, for each released lock l_i and its conflicted lock l_j, it enumerates all possible orders of the lock operations based on their lock acquiring and releasing time intervals. If each of the possible orders indicates incompatible locks, then it reports an *ME* violation in bug descriptor (lines 14 $\sim$ 15). Otherwise, it deduces a ww dependency from the possible orders (line 17).

Complexity Analysis. The time complexity of a lock releasing (or acquiring) operation in our *ME* verification is $O(n_l \cdot n_t)$, where n_l is the average number of locks released (or acquired) by an operation and n_t is the average number of conflicted locks on each record in the lock table. Specifically, the lock acquiring and releasing time intervals of each record are stored as a sorted linked list in the lock table, and each time interval is added to or removed from the list with the insertion sort method. Thus, the time complexity of maintaining the lock table for each operation is $O(n_l \cdot n_t)$. Further, for any two conflicted locks, the cost of enumerating all possible orders of the lock operations is a constant value. This is because there are at most 4 possible orders if we enforce the lock acquiring operation must happen before lock releasing operation. The space complexity is $O(n_x^l \cdot n_t)$, where n_x^l is the total number of recently locked records. Similar to the *CR* process, to reduce the size of n_x^l and n_t, we propose to asynchronously prune the locks of committed or aborted transactions that do not conflict with the locks of active transactions.

C. Verifying First Updater Wins

First updater wins (*FUW*) addresses the issue of *lost update* which might happen in concurrent transactions. The lost update occurs when the update of a transaction is overwritten by the update of another concurrent transaction. For example, suppose transaction t_0 and t_1 write independently using their own previously read values. If neither t_0 nor t_1 sees the update made from each other, then the first update on the record will be overwritten by the second one from the other transaction. *FUW* ensures that the updates of concurrent transactions are serializable. At first, *FUW* only permits the first updating



(a) Triggering a Violation (b) Deducing a ww Dependency

Fig. 7. First Updater Wins Verification Examples

transaction to commit or abort, and the other transaction has to wait for the execution result of the first updating transaction. If the first updating transaction commits successfully, then the other transaction would be forced to abort. If the first updating transaction aborts and rollbacks its update, then the other transaction would attempt to proceed with its update.

To verify *FUW* in the black-box mode, we propose to use the time intervals of visible snapshot, version installation and transaction commit/abort to identify whether there exist other concurrent versions during the period of a transaction execution. Specifically, the visible snapshot time interval contains the exact time point when a transaction get the snapshot of all its read records. The version installation and transaction commit/abort time intervals contain the exact time points when the transaction updates the record and then ends with a committed/aborted status.

The process of *FUW* verification is somewhat similar to the *ME* verification. As the aborted transactions always rollback their updates and would not lead to the case of lost update, we only consider the committed transactions in *FUW* verification. More specifically, suppose there exist two concurrent transactions operating on the same record. For the three kinds of time intervals mentioned above, if each possible order of their corresponding operations is identified to have concurrent record versions, then there must exist an *FUW* violation inside the DBMS. For example, consider the two transactions t_0 and t_1 which update the same record x in Fig. 7(a). The visible snapshot time interval of t_0 (i.e., $\mathcal{S}^{\mathcal{T}_{r_{t_0}}}$) lies between the time intervals of visible snapshot and transaction commit of t_1 (i.e., $\mathcal{S}^{\mathcal{T}_{r_{t_1}}}$ and $\mathcal{T}_{c_{t_1}}$). Then we can infer that, for both of the two possible orders ① and ②, there must exist two concurrent versions. This is because the snapshot seen by t_0 does not contain the uncommitted update of t_1. As a result, the case of lost update would happen in t_1. Otherwise, there must exist exact one possible order in which a ww dependency can be deduced (see Theorem 4). Takes Fig. 7(b) as another example, for the possible orders ③ and ④, only ④ can deduce a ww dependency. Thus, we use ④ to deduce a ww dependency and transfer it to other verifying mechanisms.

Theorem 4: Given two committed transactions t_0 and t_1, for any overlapped time intervals of the two transactions, there exists at most one possible order in which a ww dependency can be deduced. Specifically, each of the other possible orders is identified to have concurrent versions.

The pseudo-code of the *FUW* verification method is shown in Algorithm 2 (lines 18 $\sim$ 26). Given a write operation *op*, it first gets the visible time interval of the snapshot observed by *op* (line 19). Then, based on the trace set dispatched from the

Tracer, it constructs the ordered versions of each record (line 20). With the help of ordered versions, it checks whether there exists a concurrent version regarding each record updated by *op*. Specifically, for each version x^i in the write set of *op* and its conflicted version x^j (lines 21 $\sim$ 22), it enumerates all possible orders of their operations of visible snapshot, version installation and transaction commit. If each of the possible orders indicates concurrent record versions, then it reports a *FUW* violation in *bug descriptor* (lines 23 $\sim$ 24). Otherwise, it deduces a ww dependency (line 26).

Complexity Analysis. The time complexity of our *FUW* verification is $O(n_w \cdot n_v)$, where n_w is the average number of versions in the write set of an operation and n_v is the average number of record versions. For the space complexity, it uses the ordered version lists maintained by the *CR* verification mechanism and does not incur extra space cost.

D. Verifying Serialization Certifier

Serialization certifier (*SC*) applies some certifier-based approaches to guarantee that the transactions executed inside a DBMS are conflict serializable. Conflict serializability means that the dependencies between parallel transactions is equivalent to the dependencies between serial transactions.

A general approach of verifying conflict serializability is to build a dependency graph (*DG*) and performs cycle searches on the graph [11], [20]–[22]. Specifically, each node in *DG* corresponds to a committed transaction and each directed edge corresponds to a dependency between two transactions. If a cycle is found in *DG*, then it indicates that a violation of conflict serializability occurs. However, the complexity of cycle searching increases super-linearly with the scale of *DG*. To avoid the high cost of cycle searching, commercial DBMSs usually employ a lightweight certifier-based approach.

Specifically, the concurrency control protocols inside the DBMSs often take advantage of *SC* to guarantee the conflict serializability, and each protocol has its specific "certifier". For example, the SSI protocol of PostgreSQL uses two consecutive rw dependencies as its certifier. Specifically, the certifier achieves this goal by avoiding *write skew* anomalies for *snapshot isolation* [40]. *Snapshot isolation* can be implemented by the lightweight *CR* and *FUW* mechanisms mentioned before. *Write skew* happens in the situation where each transaction writes to the individual version it sees, while the final result is not equivalent to that of any serial transactions. Fortunately, it can be efficiently detected by checking whether there exist two consecutive rw dependencies [40]. Thus, the certifier would abort one of the three transactions if there exist two rw dependencies between them. Takes CockroachDB as another example, its TO protocol uses the timestamp ordering as its certifier. Specifically, the certifier does not allow a transaction with an older timestamp to have a dependency on the transaction with a newer timestamp. Thus, the cycles would never appear in *DG*. More details of the certifier can be found in Appendix III of our technical report [64].

To efficiently detect the violation of conflict serializability, we propose to follow the idea of certifier-based approach

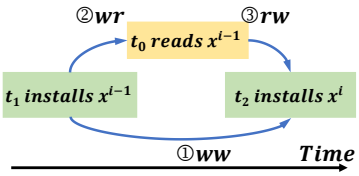


Fig. 8. Deducing Dependencies

inside the DBMS and directly use the *DG* to check whether there exists a violation of *SC*. Now the main challenge is how to build the *DG* in black-box mode. Fortunately, the wr dependencies can be obtained in the *CR* verification method and the ww dependencies can be obtained in the *ME* and *FUW* verification methods. Additionally, the rw dependencies can be further deduced from wr and ww dependencies. Fig. 8 illustrates the deducing method. Each rectangle represents a version installation time interval (colored green) or a visible snapshot time interval (colored yellow). If there exist two dependencies which indicate that transaction t_2 has a ww dependency on t_1 and t_1 has a rw dependency on t_0, then we can deduce that t_1 has a rw dependency on t_0.

The pseudo-code of the *SC* verification method is shown in Algorithm 2 (lines 27 $\sim$ 31). It first uses the dependencies deduced from the traces (including dependencies obtained from the other verification methods and the dependencies deduced by itself) to build a *DG* (lines 28 $\sim$ 29). Then, it checks whether there exists a dependency causing a dependency pattern that should be prohibited by the certifier inside the DBMS (line 30). If so, it reports a *SC* violation (line 31). For example in PostgreSQL, it checks whether a dependency leads to two consecutive rw dependencies.

As the dependencies are continuously deduced by the four verification methods, the size of *DG* would keep growing and lead to a large amount of memory usage. To address this issue, we propose to asynchronously prune the garbage transactions (see Definition 4) and their dependencies to reclaim the memory space. Theorem 5 guarantees that the garbage transactions can be pruned without affecting *SC* verification.

Definition 4: **Garbage Transaction.** A transaction t is a *garbage transaction* if t satisfies that: (*C1*) the in-degree of t is zero and (*C2*) $ts_{aft} \leq \mathcal{S}_e$. Here, ts_{aft} is the after timestamp of t's commit or abort operation, and $\mathcal{S}_e$ is earliest visible snapshot timestamp of any trace that has not been verified.

Theorem 5: A *garbage transaction* t is not a part of any future cycle on *DG*.

Complexity Analysis. The time complexity of verifying a trace $\mathcal{T}$ in *SC* is related to the implementation of the certifier inside the DBMS. For example, the time complexities of PostgreSQL and CockroachDB are $O(d)$, where d is the average degree of each node in *DG*. This is because they only need to track two consecutive rw dependencies or check whether a transaction with an older timestamp has a dependency on the transaction with a newer timestamp. The space complexity of *DG* is $O(n_t^2)$ where n_t is the number of transactions in *DG*.

VI. Experimental Evaluation

In this section, we launch sufficient experiments to answer the following questions:

(1) How efficient are *two-level pipeline* and *mechanism-mirrored verification*? (Section VI-A and VI-B) Can the throughput of *Leopard* surpass that of a DBMS? (Section VI-C)
(2) How effective is *Leopard* to deduce transactions dependencies even with overlapped time intervals? (Section VI-D)
(3) Can *Leopard* outperform state-of-the-art work, including Cobra (Section VI-E) and Elle (Section VI-F)?

Environment & Settings. *Leopard* is implemented by Java (v.1.8). Our experiments are conducted on four servers connected using 1 Gigabit Ethernet. Each server is equipped with 2 Intel Xeon Silver 4110 @ 2.1 GHz CPUs, 120 GB memory, and 4 TB HDD disk. We deploy one centalized DBMS, i.e., PostgreSQL (v12.7) and one distributed DBMS, i.e., OceanBase [61] (v3.1), to explore technical designs in *Leopard* (For space contraints, the experiment results on more DBMSs are put in Appendix V-D and V-E of our technical report [64]). Note that, OceanBase are deployed on three machines with three replicas and a client is deployed on one node. In default, *Leopard* switches on all optimizations and garbage collection, and we take PostgreSQL to demonstrate its performance.

Comparison Work. *Cobra* [11] and *Elle* [19] are the state-of-the-art work for verifying ILs. *Cobra* only verifies serializable key-value stores. It enumerates all possible dependency graphs on which it verifies ILs by cycle searching. *Elle* requires its workload to make all dependencies manifest, based on which it then builds a dependency graph and does the cycle search.

Workload. TPC-C [66] and SmallBank [67] are used to check the ability of *Leopard* in verifying workload with complex application logic. TPC-C dose not suffer from *write skew* anomalies [68], so it is insufficient for serializable verification. However, SmallBank is to benchmark the strategies to achieve serializable by eliminating *write skew* anomalies. Thus, we take both of them in our experiments. By default, they populate database with *scale factor=1*. Specifically, SmallBank populates the database with $1,000$ accounts, while TPC-C populates the database with 1 warehouse. *Cobra* [11] requires each update operation to a record must write a unique value. So *BlindW* designed by *Cobra* is used and extended to evaluate the designs in *Leopard*. In default, it creates a table sized $2K$ (*scale factor=1*) with values of 140 fixed-length strings; each transaction has 8 operations and keys are accessed uniformly under *serializable*. For different evaluation purposes, we generate three workload variants of *BlindW*:

(1) *BlindW-W* contains 100% *blind-write* transactions with uniquely written values, i.e., a write not preceded by a read to the same key. Because *blind-write* can not access the record version before creating a new version, it is a tough scenario for tracking ww dependencies (Section VI-D).
(2) *BlindW-RW* evenly contains *item-read* and *blind-write* transactions. *BlindW-RW* can build three types of dependencies, used to evaluate the usefulness of the *mechanism-mirrored verification* on deducing dependencies (Section VI-D).

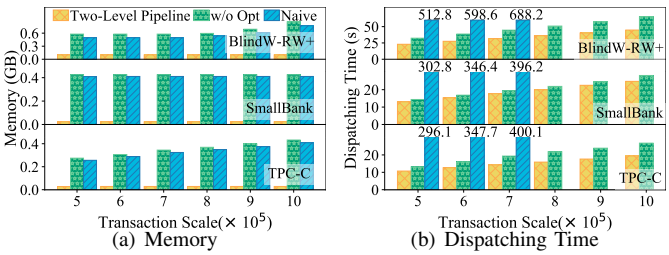


Fig. 9. Two-Level Pipeline Performance

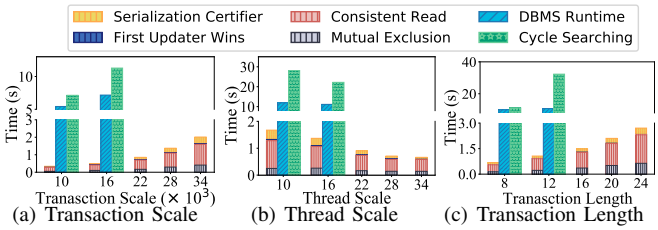


Fig. 10. Verification Time on Blind-RW+ Workload

(3) *BlindW-RW+* replaces 50% *item-read* in *BlindW-RW* with 10-keys range-read, which challenges the performance of *Leopard* with more dependencies (Section VI-B).

A. Two-level Pipeline

We design a *two-level pipeline* to sort and dispatch traces. To show its efficiency, we compare it with the naive approach, which collects traces from multiple clients and sorts them in global buffer. To demonstrate the effectiveness of optimizations to *two-level pipeline*, we also compare *Leopard* with the one without optimization, i.e., *w/o Opt*. As the timestamp distribution in traces affects the performance of *two-level pipeline* greatly, we run TPC-C, SmallBank and *BlindW-RW+* on PostgreSQL, which have different timestamp distributions.

As varying transaction scales, we collect the memory usage and the dispatching time of *two-level pipeline* in Fig. 9. Our approach is consistent far better than the other two approaches on memory usage and dispatching time. In Fig. 9(a), the naive approach has similar maximal memory consumption with *Leopard w/o Opt* due to the accumulation of a huge amount of traces in the global buffer when the distribution of timestamps in each client is extremely uneven. The naive approach has the worst dispatching time as in Fig. 9(b). The reason is that the naive approach sorts traces synchronously, while *Leopard* sorts a batch of traces asynchronously. Specifically, running *BlindW-RW+* on PostgreSQL, it has the maximum traces, which then has the longest dispatching time. When transaction scale $>7\times10^5$, the naive approach takes too much time to dispatch traces and we do not plot its time any more. In short, *two-level pipeline* can dispatch traces efficiently.

B. Mechanism-mirrored Verification

Verifying ILs can be decomposed into verify the execution of the four mechanisms. The naive approach is to build a *dependency graph* and do cycle searching, whose efficiency is significantly affected by the number of traces. We then propose an *mechanism-mirrored verification* by simulating the workflow of concurrency control protocols inside a DBMS, which only occupies a small part of the DBMS runtime

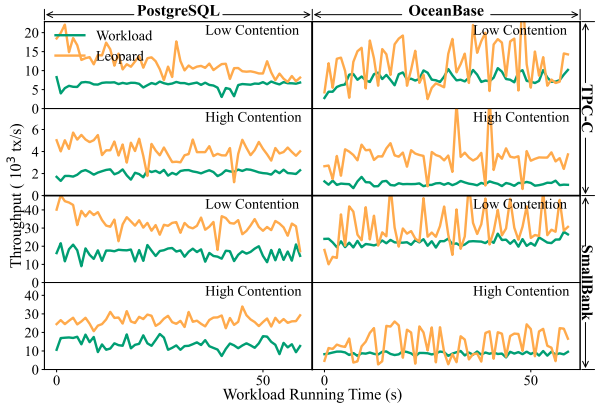


Fig. 11. DBMS Throughput vs. Leopard Throughput

theoretically. We vary the critical factors affecting verification time, including 1) transaction scale to control the number of transactions; 2) thread scale to control the data contentions; 3) transaction length to control the number of operations in a transaction. Since *BlindW-RW+* can control the three factors quantitatively, we compare the verification time of our approach with the DBMS runtime and the naive cycle searching approach in Fig. 10.

The verification time of our four mechanisms is linear with transaction scale as shown in Fig. 10(a), which benefits from timely garbage pruning. Increasing thread scale aggravates transaction contentions, leading to a higher abort rate. Since the aborted transactions are not involved in verification, the verification time for our approach decreases accordingly, as shown in Fig. 10(b). Increasing transaction length expands the read or write set in a transaction. The verification time then linearly increases with transaction length, as shown in Fig. 10(c). This is because the complexity of our verification method is linearly with the size of read or write set. Note, ours can significantly outperform the two comparison work, so we do not plot figures for thread scale>16, transaction scale>16K and length>12. Thus, our approach can efficiently verify ILs.

C. Comparison with DBMS Throughput

To show the verification efficiency of *Leopard*, we compare the performance of DBMSs with that of *Leopard* by running TPC-C and SmallBank on both PostgreSQL and OceanBase. Since contentions greatly affect DBMS throughputs as well as operation overlappings, we launch the experiments under both high and low contentions. Specifically, on PostgreSQL, for a high contention, we run TPC-C and SmallBank with 24 threads and 1 scale factor; for a low contention, we run the workloads with 24 threads, each of which is coupled with a scale factor. Since OceanBase is deployed on three machines, we expand the workloads in proportion, that is 72 threads and 3 scale factors (resp. 72 threads and 72 scale factors) for a high (resp. low) contention. We run each workload for 300s and send the traces to *Leopard* every 0.5s.

Fig. 11 shows the results. Except running TPC-C on OceanBase under a low contention, for all the other workloads, the throughput of *Leopard* dominates that of DBMSs. Even though SmallBank on OceanBase has the highest throughput under a low contention, it has simpler application logic than TPC-C, i.e., short transactions with only item-read/write and no range-read, which greatly decreases the burden on verification. So *Leopard* is fast enough to accomplish verification for SmallBank. But when running TPC-C on OceanBase under a low contention, the high throughput with complex application logic burdens *Leopard* in verification. As verifying ILs usually focuses on the conflicting operations, operations which are not conflict can be verified in parallel. Then we can scale *Leopard* to multiple instances by partitioning traces. Traces belonging to a data partition can be mapped to a *Leopard* instance, and multiple *Leopard* instances can work in parallel (More details for distributed deployment of *Leopard* are presented in Appendix IV-B of our technical report [64]). So when running TPC-C on OceanBase under a low contention, 3 *Leopard* instances are depolyed to do verification. In short, *Leopard* can be scaled out for the high throughput DBMS.

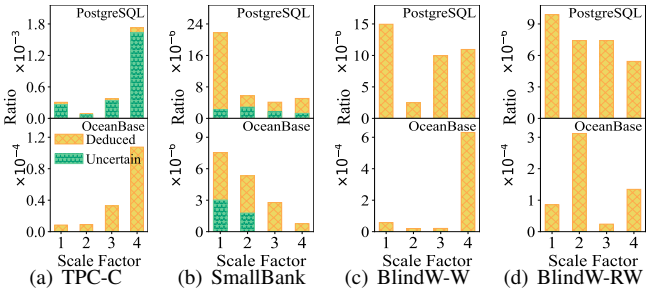


Fig. 12. Deducing Dependencies

D. Effectiveness of Deducing Dependencies

Uncertain dependencies exist due to client-side trace overlappings. We have proposed to deduce wr, ww and rw dependencies during IL verification, which can expose the order between conflict operations. To demonstrate its effectiveness, we run TPC-C, SmallBank, *BlindW-W* and *BlindW-RW* on PostgreSQL and OceanBase for 20 minutes to cover trace overlappings as much as possible. As shown in Fig. 12, we plot the ratio of the deduced dependencies (yellow bar).

The ratio of the uncertain dependencies is generally a small number ($< 10^{-3}$). For the complex application logic in TPC-C and SmallBank, there is still a part of uncertain dependencies that can not be deduced (green bar). Specifically, in TPC-C, many transactions read/write a part of attributes instead of the whole record, which makes it impossible to deduce the dependencies of two operations if they operate on different attributes of the same record. In SmallBank, transaction *amalgamate* always writes the same values, and duplicate values can not be distinguished in its candidate version set. In *BlindW-W*, all uncertain dependencies in traces are ww dependencies; the *blind-write* in *BlindW-RW* writes distinct values. *Leopard* can then expose all dependencies as in Fig. 12(c)-12(d). In short, *Leopard* can expose more uncertain dependencies effectively.

E. Comparison with Cobra on Efficiency

Cobra is designed only to verify serializable key-value stores [11]. So we set the IL of PostgreSQL as serializable.

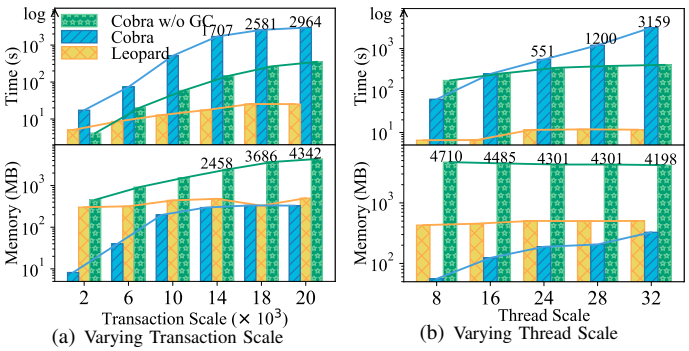


Fig. 13. Comparison with Cobra on Blind-RW Workload

Moreover, it enables garbage collection if meeting fence transactions which are inserted every 20 transactions in our experiment; *Cobra w/o GC* disables fence transactions.We report the verification time and memory usage in Fig. 13. Because of the long verification time of *Cobra* which will be exacerbated by range-read, we only run *BlindW-RW* here.

As increasing the transaction scale, the verification time of *Leopard* and *Cobra w/o GC* increases linearly and super-linearly, respectively; but *Cobra* is the worst, which spends much time identifying garbage dependencies on its polygraph as in Fig. 13(a). In memory usage, *Cobra w/o GC* cost the most to store all dependencies. *Leopard* is stable in memory usage which is almost the same as that of *Cobra* for a large transaction scale, but *Cobra* has the lowest verification efficiency. Specifically, for $20K$ transactions, *Leopard* outperforms *Cobra* by $114\times$ in verification efficiency with almost the same memory usage. In Fig. 13(b), we vary the number of workload threads to generate $20K$ transactions. *Cobra w/o GC* has the worst memory usage. Though *Cobra* has a lower memory consumption, it has uncontrollable verification time on the high concurrent workload, which is $271\times$ slower than *Leopard* when thread scale=32. Thus, *Leopard* has a much better scalability $w.r.t$ the scales of transactions and threads.

F. Comparison with Elle on Bug Cases

Elle [19] has been deployed to test TiDB. We explain one bug detected by *Leopard* but can not be located by *Elle*. (More bugs are put in Appendix V-G of our technical report [69]).

Inconsistent Read. Transaction $TID = 914$ reads the record written by the first update $TID = 904$, but does not read the latest one written by the second update $TID = 907$, which violates consistent read.

```
CREATE TABLE t(a INT PRIMARY KEY, b FLOAT);
INSERT INTO t(3873, -1.123);
UPDATE t SET b=-0.386 WHERE a=3873;--TID:904
UPDATE t SET b=0.484 WHERE a=3873;--TID:907
SELECT b FROM t WHERE a=3873;
--TID:914, Result:{-0.386}✖
```

Elle depends on the cycle in dependency graph to verify ILs. But the dependency graph of the above bug case has no cycle, so it fails to find it. *Leopard* is more general in verifying ILs without specifical requirement on workload and can expose more subtle bugs. We also compare the performance with Elle by running the workload designed for it [70] and *Leopard* is more efficient than Elle (Details are in Appendix V-F of our technical report [69]).

Summaries and Practical Applications. *Leopard* can be easily integrated with any workload generator, e.g., *OLTP-Bench* [71], without modifying application logic and is general to verify ILs. After running *Leopard* on several popular DBMSs, we have discovered 23 bugs (13 fixed, 15 confirmed and 8 open reported), demonstrated in [72].

VII. Related Work

Verifying isolation levels (ILs) is usually achieved by elaborating workloads or instrumenting source codes of DBMSs [4], [11], [19], [73]. However, none of current work is general to verify various ILs in a black-box mode with arbitrary workloads. *Cobra* [11] can only exposes serializability violations of transactional key-value stores based on a workload following a specific application logic. By an expensive graph traverse, it prunes garbage transactions. *Elle* [19] specifies a short workload to expose version orders in history and its verification is only based on a graph cycle detection. Both *Cobra* and *Elle* can not verify arbitrary workloads, e.g., TPC-C. Mai et.al [4] diagnose violations of the ACID properties in a white-box method by injecting power faults while replaying its workloads. Yu et.al [73] design a DBMS that can provide verifiable proofs of transaction correctness and semantic properties. But it is only verifiable for the serializable IL. In contrast, *Leopard* can verify the workload with any application logic and is not limited to a specific IL.

Some work proposes to detect anomalies in application workloads [3], [17], [20]–[22] instead of DBMSs. *IsoDiff* [20] takes an analysis of application codes to debug anomalies (caused by weak isolations) on the representative subsets of dependency graphs. *Rushmon* [21] monitors real-time anomalies caused by the asynchronous algorithm for the inconsistency-tolerant applications on weak isolation systems. It takes the idea of serializability to achieve detection by sampling the dependency graph, which is not a comprehensive checking method. *ConsAD* [22] quantifies isolation anomalies by detecting cycles in the dependency graph. But it costs a lot to analyze the application logic to track dependencies. Todd et.al [3] aim to detect potential isolation anomalies in web applications. They reason the possible concurrent interleavings among clients to generate workloads to violate the ACID principle. In contrast, *Leopard* aims to verify various ILs in DBMSs without modifying application logic.

VIII. Conclusion and Future Work

Leopard abstracts four general implementation mechanisms for various ILs. It proposes to verify ILs in a black-box way based on client-side execution traces. *Two-level pipeline* and *mechanism-mirrored verification* are designed to accomplish efficient and effective verification. Compared with existing studies, *Leopard* has order-of-magnitudes improvement on performance and better ability in bug detection. However, the side-effect of time interval overlapping of traces prevents us from detecting all dependencies, and digging up all bugs is still impossible. We leave it as our future work.

REFERENCES

[1] R. L. J. Gray, G. Putzolu, and I. Traiger, "Granularity of locks and degrees of consistency," *Modeling in Data Base Management Systems, GM Nijssen ed., North Holland Pub*, 1976.

[2] A. Pavlo, "What are we doing with our lives? nobody cares about our concurrency control research," in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017, pp. 3–3.

[3] T. Warszawski and P. Bailis, "Acidrain: Concurrency-related attacks on database-backed web applications," in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017, pp. 5–20.

[4] M. Zheng, J. Tucek, D. Huang, F. Qin, M. Lillibridge, E. S. Yang, B. W. Zhao, and S. Singh, "Torturing databases for fun and profit," in *11th USENIX Symposium on Operating Systems Design and Implementation*, 2014, pp. 449–464.

[5] J. C. C. r and et al., "Spanner: Google's globally-distributed database," in *OSDI*, 2012, pp. 251–264.

[6] R. Taft, I. Sharif, A. Matei, N. VanBenschoten, J. Lewis, T. Grieger, K. Niemi, A. Woods, A. Birzin, R. Poss *et al.*, "CockroachDB: The resilient geo-distributed SQL database," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 1493–1509.

[7] D. Huang, Q. Liu, Q. Cui, Z. Fang, X. Ma, F. Xu, L. Shen, L. Tang, Y. Zhou, M. Huang *et al.*, "TiDB: a Raft-based HTAP database," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3072–3084, 2020.

[8] C. A. Stuardo, T. Leesatapornwongsa, R. O. Suminto, H. Ke, J. F. Lukman, W.-C. Chuang, S. Lu, and H. S. Gunawi, "Scalecheck: A single-machine approach for discovering scalability bugs in large distributed systems," in *17th USENIX Conference on File and Storage Technologies (FAST 19)*, 2019, pp. 359–373.

[9] H. S. Gunawi, M. Hao, T. Leesatapornwongsa, T. Patana-anake, T. Do, J. Adityatama, K. J. Eliazar, A. Laksono, J. F. Lukman, V. Martin *et al.*, "What bugs live in the cloud? a study of 3000+ issues in cloud systems," in *Proceedings of the ACM symposium on cloud computing*, 2014, pp. 1–14.

[10] H. S. Gunawi, M. Hao, R. O. Suminto, A. Laksono, A. D. Satria, J. Adityatama, and K. J. Eliazar, "Why does the cloud stop computing? lessons from hundreds of service outages," in *Proceedings of the Seventh ACM Symposium on Cloud Computing*, 2016, pp. 1–16.

[11] C. Tan, C. Zhao, S. Mu, and M. Walfish, "Cobra: Making transactional key-value stores verifiably serializable," in *OSDI*, 2020, pp. 63–80.

[12] K. P. Gaffney, R. Claus, and J. M. Patel, "Database isolation by scheduling," *Proceedings of the VLDB Endowment*, vol. 14, no. 9, pp. 1467–1480, 2021.

[13] "Cockroachdb bugs," https://github.com/cockroachdb/cockroach/issues.

[14] "Yugabyte bugs," https://github.com/yugabyte/yugabyte-db/issues.

[15] "Jepsen: Postgresql 12.3." https://jepsen.io/analyses/postgresql-12.3.

[16] A. Fekete, S. N. Goldrei, and J. P. Asenjo, "Quantifying isolation anomalies," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 467–478, 2009.

[17] S. Jorwekar, A. Fekete, K. Ramamritham, and S. Sudarshan, "Automating the detection of snapshot isolation anomalies," *Proceedings of the VLDB Endowment*, 2007.

[18] A. Dey, A. Fekete, R. Nambiar, and U. Röhm, "Ycsb+ t: Benchmarking web-scale transactional databases," in *2014 IEEE 30th International Conference on Data Engineering Workshops*, 2014, pp. 223–230.

[19] P. Alvaro and K. Kingsbury, "Elle: Inferring isolation anomalies from experimental observations," *Proceedings of the VLDB Endowment*, vol. 14, no. 3, pp. 268–280, 2020.

[20] Y. Gan, X. Ren, D. Ripberger, S. Blanas, and Y. Wang, "Isodiff: debugging anomalies caused by weak isolation," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 2773–2786, 2020.

[21] Z. Shang, J. X. Yu, and A. J. Elmore, "Rushmon: Real-time isolation anomalies monitoring," in *Proceedings of the 2018 International Conference on Management of Data*, 2018, pp. 647–662.

[22] K. Zellag and B. Kemme, "Real-time quantification and classification of consistency anomalies in multi-tier architectures," in *2011 IEEE 27th International Conference on Data Engineering*, 2011, pp. 613–624.

[23] K. Nagar and S. Jagannathan, "Automated detection of serializability violations under weak consistency," *arXiv preprint arXiv:1806.08416*, 2018.

[24] Y. Lu, X. Yu, L. Cao, and S. Madden, "Aria: a fast and practical deterministic oltp database," *Proceedings of the VLDB Endowment*, 2020.

[25] W. M. McKeeman, "Differential testing for software," *Digital Technical Journal*, vol. 10, no. 1, pp. 100–107, 1998.

[26] L. Brutschy, D. Dimitrov, P. Müller, and M. Vechev, "Serializability for eventual consistency: criterion, analysis, and applications," in *SIGPLAN*, 2017, pp. 458–472.

[27] C. Hammer, J. Dolby, M. Vaziri, and F. Tip, "Dynamic detection of atomic-set-serializability violations," in *Proceedings of the 30th international conference on Software engineering*, 2008, pp. 231–240.

[28] A. Sinha and S. Malik, "Runtime checking of serializability in software transactional memory," in *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)*, 2010, pp. 1–12.

[29] W. N. Sumner, C. Hammer, and J. Dolby, "Marathon: Detecting atomic-set serializability violations with conflict graphs," in *International Conference on Runtime Verification*, 2011, pp. 161–176.

[30] M. Xu, R. Bodík, and M. D. Hill, "A serializability violation detector for shared-memory server programs," *ACM Sigplan Notices*, vol. 40, no. 6, pp. 1–14, 2005.

[31] K. Zellag and B. Kemme, "Consistency anomalies in multi-tier architectures: automatic detection and prevention," *The VLDB Journal*, vol. 23, no. 1, pp. 147–172, 2014.

[32] P. A. Bernstein and N. Goodman, "Multiversion concurrency control—theory and algorithms," *TODS*, vol. 8, no. 4, pp. 465–483, 1983.

[33] C. H. Papadimitriou, "The serializability of concurrent database updates," *Journal of the ACM (JACM)*, vol. 26, no. 4, pp. 631–653, 1979.

[34] A. X3, "American national standard for information systems-database language-sql," 1992.

[35] H. Berenson, P. Bernstein, J. Gray, J. Melton, E. O'Neil, and P. O'Neil, "A critique of ansi sql isolation levels," *ACM SIGMOD Record*, vol. 24, no. 2, pp. 1–10, 1995.

[36] A. Adya and B. H. Liskov, "Weak consistency: a generalized theory and optimistic implementations for distributed transactions," Ph.D. dissertation, Massachusetts Institute of Technology, 1999.

[37] N. Crooks, Y. Pu, L. Alvisi, and A. Clement, "Seeing is believing: A client-centric specification of database isolation," in *Proceedings of the ACM Symposium on Principles of Distributed Computing*, 2017, pp. 73–82.

[38] A. Szekeres and I. Zhang, "Making consistency more consistent: A unified model for coherence, consistency and isolation," in *Proceedings of the 5th Workshop on the Principles and Practice of Consistency for Distributed Data*, 2018, pp. 1–8.

[39] "Innodb," https://dev.mysql.com/doc/refman/8.0/en/innodb-storage-engine.html.

[40] D. R. K. Ports and K. Grittner, "Serializable snapshot isolation in postgresql," *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 1850–1861, 2012.

[41] "Oracle database," https://www.oracle.com/hk/database/technologies/.

[42] A. Fekete, E. O'Neil, and P. O'Neil, "A read-only transaction anomaly under snapshot isolation," *ACM SIGMOD Record*, vol. 33, no. 3, pp. 12–14, 2004.

[43] R. Biswas and C. Enea, "On the complexity of checking transactional consistency," *Proceedings of the ACM on Programming Languages*, vol. 3, no. OOPSLA, pp. 1–28, 2019.

[44] "Leopard IL_Description," https://github.com/Coconut-DB1024/Leopard/blob/main/IL_Description.pdf.

[45] S. Tu, W. Zheng, E. Kohler, B. Liskov, and S. Madden, "Speedy transactions in multicore in-memory databases," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, 2013, pp. 18–32.

[46] X. Yu, A. Pavlo, D. Sanchez, and S. Devadas, "Tictoc: Time traveling optimistic concurrency control," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 1629–1642.

[47] A. Thomson, T. Diamond, S.-C. Weng, K. Ren, P. Shao, and D. J. Abadi, "Calvin: fast distributed transactions for partitioned database systems," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 1–12.

[48] J. Wang, D. Ding, H. Wang, C. Christensen, Z. Wang, H. Chen, and J. Li, "Polyjuice: High-performance transactions via learned concurrency control," in *OSDI*, 2021, pp. 198–216.

[49] D. Tang and A. J. Elmore, "Toward coordination-free and reconfigurable mixed concurrency control," in *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, 2018, pp. 809–822.

[50] G. Li, X. Zhou, J. Sun, X. Yu, Y. Han, L. Jin, W. Li, T. Wang, and S. Li, "opengauss: An autonomous database system," *Proceedings of the VLDB Endowment*, vol. 14, no. 12, pp. 3028–3042, 2021.

[51] "yugabyteDB," https://www.yugabyte.com/.

[52] P.-Å. Larson, A. Birka, E. N. Hanson, W. Huang, M. Nowakiewicz, and V. Papadimos, "Real-time analytical processing with sql server," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1740–1751, 2015.

[53] A. Verbitski, A. Gupta, D. Saha, M. Brahmadesam, K. Gupta, R. Mittal, S. Krishnamurthy, S. Maurice, T. Kharatishvili, and X. Bao, "Amazon aurora: Design considerations for high throughput cloud-native relational databases," in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017, pp. 1041–1052.

[54] F. Li, "Cloud-native database systems at alibaba: Opportunities and challenges," *Proceedings of the VLDB Endowment*, vol. 12, no. 12, pp. 2263–2272, 2019.

[55] P. Bhatotia, A. Wieder, İ. E. Akkuş, R. Rodrigues, and U. A. Acar, "Large-scale incremental data processing with change propagation," in *3rd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 11)*, 2011.

[56] "Rocksdb," https://www.rocksdb.org/.

[57] "SQLite," https://www.sqlite.org/index.html/.

[58] J. Zhou, M. Xu, A. Shraer, B. Namasivayam, A. Miller, E. Tschannen, S. Atherton, A. J. Beamon, R. Sears, J. Leach *et al.*, "Foundationdb: A distributed unbundled transactional key value store," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 2653–2666.

[59] "Singlestore," https://www.singlestore.com/.

[60] V. Sikka, F. Färber, W. Lehner, S. K. Cha, T. Peh, and C. Bornhövd, "Efficient transaction processing in sap hana database: the end of a column store myth," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 731–742.

[61] Z. Yang, C. Yang, F. Han, M. Zhuang, B. Yang, Z. Yang, X. Cheng, Y. Zhao, W. Shi, H. Xi *et al.*, "OceanBase: a 707 million tpmC distributed relational database system," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3385–3397, 2022.

[62] "NuoDB," https://nuodb.com/.

[63] G. Weikum and G. Vossen, *Transactional information systems: theory, algorithms, and the practice of concurrency control and recovery*. Elsevier, 2001.

[64] "Leopard Technical Report," https://github.com/Coconut-DB1024/Leopard/blob/main/Technical%20Report.pdf.

[65] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking cloud serving systems with ycsb," in *Proceedings of the 1st ACM symposium on Cloud computing*, 2010, pp. 143–154.

[66] "TPC-C benchmark," http://www.tpc.org/tpcc/.

[67] M. Alomari, M. Cahill, A. Fekete, and U. Rohm, "The cost of serializability on platforms that use snapshot isolation," in *2008 IEEE 24th International Conference on Data Engineering*, 2008, pp. 576–585.

[68] A. Fekete, D. Liarokapis, E. O'Neil, P. O'Neil, and D. Shasha, "Making snapshot isolation serializable," *TODS*, vol. 30, no. 2, pp. 492–528, 2005.

[69] "Leopard," https://github.com/Coconut-DB1024/Leopard.

[70] "Elle github," https://github.com/pingcap/tipocket/tree/master/testcase/rw-register.

[71] D. E. Difallah, A. Pavlo, C. Curino, and P. Cudre-Mauroux, "Oltp-bench: An extensible testbed for benchmarking relational databases," *Proceedings of the VLDB Endowment*, vol. 7, no. 4, pp. 277–288, 2013.

[72] "Leopard Bug List," https://github.com/Coconut-DB1024/Leopard/blob/main/Bug%20List.pdf.

[73] Y. Xia, X. Yu, M. Butrovich, A. Pavlo, and S. Devadas, "Litmus: Towards a practical database management system with verifiable acid properties and transaction correctness."

[74] D. Peng and F. Dabek, "Large-scale incremental processing using distributed transactions and notifications," in *9th USENIX Symposium on Operating Systems Design and Implementation (OSDI 10)*, 2010.

[75] "Robert Tarjan," Depth-first Search and Linear Graph Algorithms (SWAT), 1971.

APPENDIX

I. Theorem Proof

Theorem 1: Algorithm 1 dispatches traces in monotonically increasing order of before timestamps.

Proof 1: Supposing we have n_{local} local buffers and W is the watermark that is the smallest before timestamp among all traces in local buffers. According to Algorithm 1, the dispatched trace $\mathcal{T}$ satisfies:

$$\mathcal{T}.ts_{bef} \leq \min_{\mathcal{T}_i \in global} \mathcal{T}_i.ts_{bef} \quad (1)$$

$$\mathcal{T}.ts_{bef} \leq W = \min_{0 \leq i \leq n_{local}-1} local_i[0].ts_{bef} \quad (2)$$

Since the traces in each client C_i are generated in an increasing order of before timestamps, then we have:

$$\begin{aligned} local_i[0].ts_{bef} &\leq \min_{\mathcal{T}_j \in local_i} \mathcal{T}_j.ts_{bef} \\ &\leq \min_{\mathcal{T}_j \in C_i} \mathcal{T}_j.ts_{bef}, 0 \leq i \leq n_{local}-1 \quad (3) \end{aligned}$$

From Equations (1)-(3), we can infer that the dispatched trace has the minimum before timestamp among all traces in the global buffer, local buffers and clients. Thus theorem is proven.

Theorem 2: The candidate version set contains a minimum number of versions that are possibly visible to a given read operation.

Proof 2: Suppose a version x^i falls into the candidate version set but is impossibly visible to a given read operation. There exist two cases if x^i must be invisible to the read operation. In the first case, the version x^i appears after the read operation. Then, we can infer that the version installation time interval of x^i must not overlap with the visible snapshot time interval. Otherwise, x^i is possibly visible to the read operation since we could not determine the chronological order of the exact read operation time point and version installation time point. This implies that x^i is a future version*.*

In the second case, the version x^i appears before the read operation but has been overwritten by another version. As discussed above, we could not determine the chronological order of the pivot overlap version*,* pivot version *and* overlap version *since their exact arrive times are not available. That is, any version among them is possibly visible to the read operation. Thus, the version x^i must not be one of the three versions, which implies that x^i is a* garbage version*.*

Since our approach excludes all the future versions *and* garbage versions *from the candidate version set, this is contradicted with the initial assumption. The theorem is proven.*

Theorem 3: Given two transactions t_0 and t_1, for any overlapped time intervals of two conflicting locks, there exists at most one possible order in which a ww dependency can be deduced. Specifically, each of the other possible orders is identified to have incompatible locks.

Proof 3: Suppose there exist two possible orders in which two ww dependencies can be deduced. On the one hand, if t_0 has a ww dependency on t_1, then we can infer that the exact lock acquiring time of t_0 must happen before that of t_1. On the other hand, if t_1 has a ww dependency on t_0, then the exact lock acquiring time of t_1 must happen after the lock releasing time of t_0. To make the two ww dependencies possibly deduced, the lock acquiring time interval of t_0 (i.e., $\mathcal{A}^{\mathcal{T}_{w_{t_0}}}$) must overlap with the lock acquiring and releasing time intervals of t_1 (i.e., $\mathcal{A}^{\mathcal{T}_{w_{t_1}}}$ and $\mathcal{R}^{\mathcal{T}_{c_{t_1}}}$). Similarly, $\mathcal{R}^{\mathcal{T}_{c_{t_0}}}$ must also overlap with $\mathcal{A}^{\mathcal{T}_{w_{t_1}}}$ and $\mathcal{R}^{\mathcal{T}_{c_{t_1}}}$.

From $\mathcal{A}^{\mathcal{T}_{w_{t_0}}}$ overlaps with $\mathcal{A}^{\mathcal{T}_{w_{t_1}}}$ and $\mathcal{R}^{\mathcal{T}_{c_{t_1}}}$, we have $\mathcal{A}^{\mathcal{T}_{w_{t_1}}}.ts_{aft} < \mathcal{A}^{\mathcal{T}_{w_{t_0}}}.ts_{aft}$. Additionally, the lock releasing time interval must happen after the lock acquiring time interval, then we have $\mathcal{A}^{\mathcal{T}_{w_{t_0}}}.ts_{aft} < \mathcal{R}^{\mathcal{T}_{c_{t_0}}}.ts_{bef}$. Taken together, we have $\mathcal{A}^{\mathcal{T}_{w_{t_1}}}.ts_{aft} < \mathcal{R}^{\mathcal{T}_{c_{t_0}}}.ts_{bef}$, which indicates that $\mathcal{R}^{\mathcal{T}_{c_{t_0}}}$ would not overlap with $\mathcal{A}^{\mathcal{T}_{w_{t_1}}}$ and $\mathcal{R}^{\mathcal{T}_{c_{t_1}}}$ simultaneously. This is contradicted with the initial assumption. The theorem is proven.

Theorem 4: Given two committed transactions t_0 and t_1, for any overlapped time intervals of the two transactions, there exists at most one possible order in which a ww dependency can be deduced. Specifically, each of the other possible orders is identified to have concurrent versions.

Proof 4: The proof is similar to that of Theorem.3.

Theorem 5: A *garbage transaction* t is not a part of any future cycle on *DG*.

Proof 5: From C1 *in Definition 4, we can infer that the in-degree of t would be zero unless a future transaction creates a new dependency on t. Let $\mathcal{T}_k$ be the trace of the first operation of any committed transaction in future. From* C2 *in Definition 4, we can deduce that $\mathcal{T}_t.ts_{aft} \leq \mathcal{S}_e \leq \mathcal{S}^{\mathcal{T}_k}.ts_{bef}$. This indicates that any future transaction would not have dependencies on t. Taken together, the in-degree of t will keep as zero. However, the in-degree of* garbage transaction t *must be large than zero if t is contained in a cycle. Thus, t is not a part of any future cycle on* DG*. The theorem is proven.*

II. Batch-based Two-level Pipeline

In our implementation, our watermark based two-level pipeline also takes advantage of three optimizations, which are multi-thread processing, batch dispatching, and adaptive fetching. Next, we discuss the two optimizations in detail.

Trace Partitioning for Multi-Thread Processing. As the IL verifying process usually focuses on the conflicting operations which accesses the same records, we observe that the operations which are not in conflict can be verified in parallel. To fully utilize the parallelism of modern CPUs, we propose to parallelly sort and verify the traces with the help of trace partitioning. That is, each partition runs its sorting and verifying process individually. Specifically, the number of partitions is jointly determined by the CPU cores and the number of required threads for the sorting and verifying process. In our implementation, for each partition, we use 1 thread for sorting and 4 threads for parallelly verifying the 4 mechanisms. Moreover, the rule of trace partitioning can borrow ideas from the database sharding. Note that, the trace

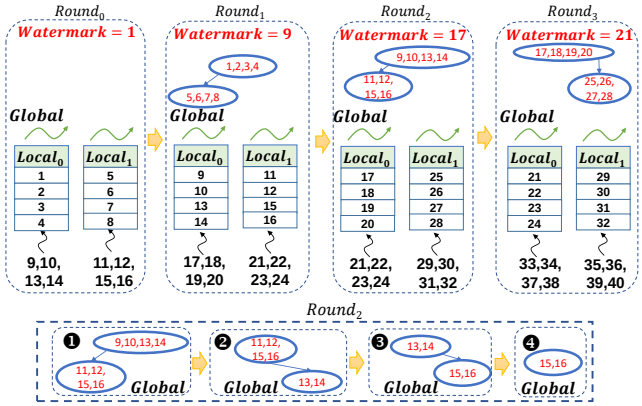


Fig. 14. Batch-based Two-level Pipeline Example

partitioning method also helps increase the batch size for the following batch process method.

Amortizing the Sorting Cost with Batch Processing. As a transactional workload often has the scenario in which a transaction continuously executes multiple operations while blocking other conflicting transactions, we observe that there usually exist some consecutive traces in a local buffer whose before timestamps' range does not cover the before timestamp of any trace in other local buffers. If we can directly dispatch these consecutive traces to the verifier in a batch, the sorting cost can be efficiently amortized. For this purpose, instead of sorting traces one by one in the global buffer, we propose to sort traces with the granularity of batches. Specifically, as shown in Fig. 14, each heap node in the global buffer consists of a batch traces, and the min-heap is organized according to the first (i.e. smallest) timestamp in each batch. For the trace dispatching process, it first pops a node from the heap and checks if its range of timestamps overlaps with that of the current heap root or overlaps with the watermark. If not, it can infers that the batch of popped traces must have smaller timestamps than all the other traces in the global buffer. Then, it directly dispatches the batch of traces to the verifier. Otherwise, it only dispatches traces whose timestamps are smaller than the watermark and the first timestamp in the heap root, and then inserts the rest traces into the heap as a batch. Note that, the trace partitioning method can help reduce trace interleavings among different partitions, then the batch size can be effectively increased.

Fig. 14 gives a running example of our batch sorting optimization. In this example, each trace is represented by its before timestamp, and we set batch size as 4 traces. In $Round_0$, the global buffer and watermark are initialized as $\varnothing$ and 1. Then, the two clients push their collected traces into the local buffers. In $Round_1$, the global buffer first fetches a batch of traces from each local buffer and sorts them with the min-heap. The watermark is set as the smallest before timestamp of the two local buffers, which is 9. Then, it pops the heap and get a trace batch $< 1, 2, 3, 4 >$. As all the traces in the trace batch are smaller than the watermark and the first trace 5 in the heap root, so it directly dispatches them as a batch to the verifier. Similarly, $< 5, 6, 7, 8 >$ can be also dispatched as a batch to the verifier.

In $Round_2$, it repeats the steps in $Round_1$, that is, fetches a batch of traces from each local buffer into the global buffer, sorts them with the min-heap, and the watermark is set as 17. Specifically, it takes the following steps to dispatch the traces in batches. ❶ It pops the heap and get traces $< 9, 10, 13, 14 >$. However, as $< 13, 14 >$ are large than the first trace 11 in current heap root. Then, it split the popped trace batch and dispatches $< 9, 10 >$ as a batch to the verifier. Next, it inserts $< 14, 15 >$ into the global buffer as a batch. ❷ Similarly, it pops traces $< 11, 12, 15, 16 >$, and dispatches traces $< 11, 12 >$. Then, it inserts $15, 16$ into the global buffer as a batch. ❸ It directly dispatches traces $< 13, 14 >$ which is smaller than the watermark and the first trace 15 in current heap root. ❹ Finally, it directly dispatches traces $< 15, 16 >$ which is smaller than the watermark. Similarly, in $Round_3$, it repeats the steps as described in $Round_2$.

Controlling the Watermark with Adaptive Fetching. The performance of trace sorting depends on the number of traces in the global buffer. We observe that the distribution of timestamps in each local buffer has a significant impact on the size of global buffer. For example, if the operations in a client are executed much slower than other clients, then the timestamps in the corresponding log buffer are much smaller than that of others. As a result, it would hinder the increase of watermark, and the size of global buffer will keep increasing due to the accumulation of traces which are fetched from other local buffers. To address this issue, we prefer to fetch traces from the local buffer with the smallest timestamp. Further, to keep the size of global buffer stable, we always restrict that the speed of trace dispatching is equivalent to the speed of trace fetching.

Complexity Analysis. We denote n_{local} and s_{local} as the number of local buffers and the size of traces in each local buffer, respectively. Intuitively, the space complexity of local buffers is $O(n_{local} \cdot s_{local})$. For initializing global buffer, we fetch all traces of local buffers into global buffer. After that, taking advantage of adaptive fetching, we guarantee that the size of traces that are transferred into global buffer equals the size of traces that are dispatched out of global buffer. In such a way, the space complexity of global buffer equals that of local buffers. Additionally, there may be some traces temporally stored in each test client. Then, the space complexity of traces stored in test clients is $O(n_{local} \cdot s_{client})$ where s_{client} is the size of traces in each test client. In summary, the space complexity of two-level pipeline is $O(n_{local} \cdot (s_{local} + s_{client}))$. Suppose each trace is treated as a node in global buffer. Then, the time complexity of each dispatched trace is $O(log(n_{local} \cdot s_{local}))$. Taking advantage of batch processing, we treat several consecutive traces as a batch in global buffer. Here, we denote B as the average batch size. Then, the number of nodes in global buffer can be reduced to $\frac{n_{local} \cdot s_{local}}{B}$. Meanwhile, since we can dispatch a batch of trace at once, we can further reduce the time complexity to $O(\frac{1}{B} \cdot log(\frac{n_{local} \cdot s_{local}}{B}))$.

III. More details about Serialization Certifier

The concurrency control protocols (CCP) in popular DBMSs often take advantage of the *serialization certifier* (SC) mechanism to guarantee that the executed transactions are conflict serializability [63]. However, each concurrency control protocol has its specific "certifier". Table II lists the the certifiers used by popular DBMSs. Next, we discuss how each certifier is used.

The *serializable snapshot isolation* (*SSI*) protocol implements the *SC* mechanism based on *snapshot isolation*, and uses the detection of the *write skew* anomalies as its certifier. Specifically, the *write skew* anomaly can be efficiently detected by checking whether there exist two consecutive rw dependencies [40]. Thus, the certifier would abort one of the three transactions if there exist two rw dependencies between them.

The *timestamp ordering* (*TO*) protocol uses the transaction timestamp ordering as its certifier when implementing its *SC* mechanism. Specifically, the certifier checks whether a transaction with an older timestamp has a dependency on another transaction with a newer timestamp. If so, it would abort the old transaction.

The *optimistic concurrency control* (*OCC*) protocol specifies that a transaction in the DBMS is executed in three phases: *read*, *validation*, and *write*. In the *read* phase, the transaction performs read and write operations on records without blocking. When the transaction prepares to commit, the *validation* phase uses the conflicts checking as a certifier to guarantee the conflict serializability. If there exist conflicts, the certifier would abort the transaction; otherwise, it commits the transaction in *write* phase.

The *percolator* [74] protocol[1] also implements the idea of *serialization certifier* mechanism. Similar to the three concurrency control protocols mentioned above, *percolator* use its certifier to eliminate some specific anomalies when the transaction starts to commit. However, the certifier of *percolator* only guarantees snapshot isolation, and could not guarantee the conflict serializability. Specifically, *percolator* uses the check of concurrent versions as a certifier to eliminate *lost update* anomaly. That is, when a transaction starts to commit, the certifier checks whether a newer version has been installed since the transaction begins. If there exist concurrent versions, the certifier would abort the transaction; otherwise, it commits the transaction. We consider *percolator* as a special case of implementing the *serialization certifier* mechanism.

TABLE II
Serialization Certifier

CCP	Certifier
SSI	detecting write skew anomalies based on snapshot isolation
TO	ordering transactions as a monotonically increasing timestamp
OCC	checking the transaction conflicting with other active ones
Percolator	detecting lost update anomalies before committing a transaction

IV. Leopard Implementation

A. Single Machine Deployment

Fig. 15 depicts our implementation details on a single machine. A workload consists of multiple clients. ❶ Each client issues the requested operations to a DBMS. ❷ Each client also receive results responded by a DBMS. Each client encodes an responded operations as an interval-based trace (trace for short). Note that, as described in Appendix II, we shard a database into several data partitions, and each partition is coupled with a parallelism unit. ❸ According to the data partition accessed by the operation, we branch off traces into the corresponding parallelism unit. Parallelism unit retrieves traces from clients and attempts to verify whether the workload satisfies the definition of isolation levels.

Specifically, the *local buffer*s cache the streaming traces from each client asynchronously. ❹ In the meanwhile, the *global buffer* fetches and sorts all traces from each *local buffer*. Based on sorted traces from *global buffer*, *Leopard* launches the verification. The verification includes three phases, including context preparation, forking and garbage collection. ❺ Context preparation installs sorted traces into three contexts, i.e.., *version orders*, *dependency graph* and *lock table*. Specifically, *version orders* maintains the version evolution of each record for verifying *consistent read* and *first updater wins*, *dependency graph* captures the transaction dependencies for verifying *serialization certifier*, and *lock table* temporarily saves the locks acquired by each operation for verifying *mutual exclusion*. ❻ Then, based on the prepared context, it forks four threads to verify the four implementation mechanisms in parallel. ❼ Finally, for catching up with a long running workload, garbage collection aggressively cleans up the contexts that have nothing to do with the successive verification.

B. Multiple Machine Deployment

Fig. 16 depicts the distributed deployments of our *Leopard* framework. Not only the tested DBMS can be deployed in a distributed environment, but also the test clients, the trace sorting and verifying components can be scaled to multiple machines. Next, we discuss the key issues of *Leopard*'s distributed deployments.

Clock Synchronizations among Test Clients. As we need to log the time interval of each database operation, the clock synchronization of timestamps is a critical issue when the test clients are deployed on multiple machines. We propose to make use of the clock synchronizing method inside the DBMSs. Specifically, if the data nodes are deployed in a single cluster with low network latencies, a centralized logical timestamp generation method is more appropriate. For example, both TiDB [7] and OceanBase [61] use a centralized

[1]Strictly speaking, *percolator* is a transaction processing framework, not only a concurrency control protocol.

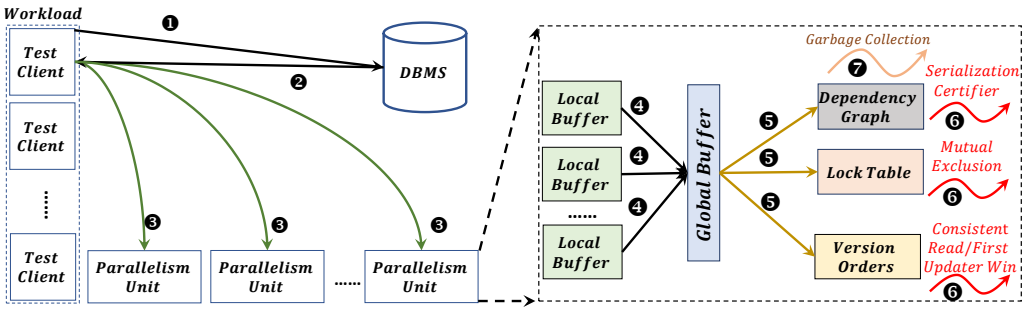


Fig. 15. Implementation Details of Leopard on A Single Machine

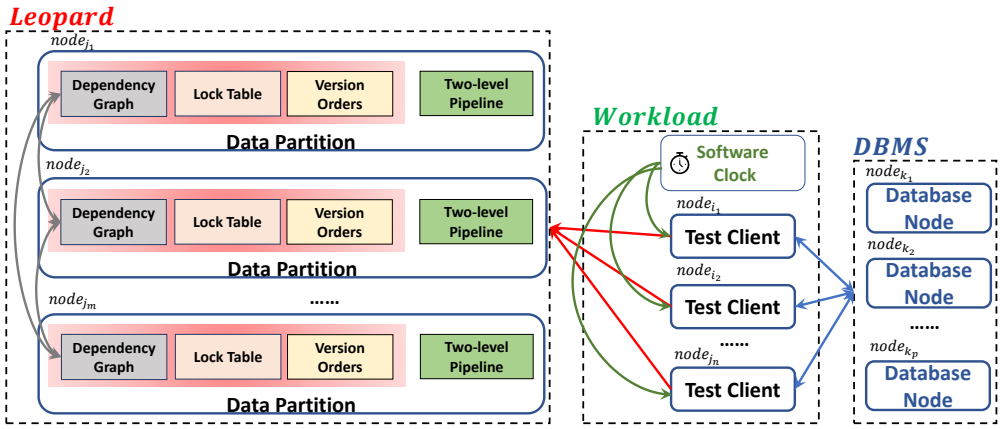


Fig. 16. Distributed Deployments of Leopard

timestamp oracle to allocate logical timestamps for all transactions. Moreover, OceanBase suggests that applications should avoid expensive acquisitions of cross-region timestamps. Our experimental results show that the latency of each timestamp acquisition is much smaller (at least $45\times$) than the latency of database operation.

Scale Leopard to Multiple Instances. As the verifying process of isolation levels usually focuses on the conflicting operations which accesses the same records, we can deploy multiple Leopard instances to do verification in parallel. That is, we can make use of the database sharding information provided by the DBMSs (or borrow ideas from popular database sharding algorithms), and divide all workload traces into partitions accordingly. Then, we can deploy each Leopard instance on a single node and use it to verify the traces belonging to one (or more) specific partitions. Note that, as the *CR*, *ME* and *FUW* mechanisms only concern the data conflicts, they can be served by each node individually. For the *SC* mechanism, as we we need construct the dependency graph whose nodes are transactions and edges are dependencies between transactions, it might need to verify the transaction dependencies which across different nodes. Nevertheless, it needs a low communication cost if the data nodes are deployed in a single cluster with low network latencies. This is because the verification of each cross-node dependency only needs one network round trip. Takes verifying *serialization certifier* of PostgreSQL as an example. It detects wheter there exist two consecutive rw dependencies. In the worst case, verifying *serialization certifier* only requires two network round trips.

V. More Experiments

In default, our experimental environment is the same as the one declared in our paper. Specifically, TiDB and OceanBase are deployed on three servers with three replicas and a client is deployed on one server. MySQL and PostgreSQL are deployed on one server individually.

A. Evaluation on Two Optimizations of Two-level Pipeline

An online transaction processing always continuously produces a massive amount of traces, which pose a significant challenge on the high-performance trace sorting. External merge sort is a traditional approach to sort massive amounts of data. We launch an experiment to compare it with our approach. We run the standard benchmark TPC-C with 24 threads and 1 warehouse on PostgreSQL and MySQL. In Fig. 17, as varying transaction scale, we collect the dispatching time of our approach and the external merge sort. From results, the dispatching time of our method outperforms the external merge sort method by up to $3.6\times$. Thus, taking advantage of two optimizations based on multi-thread and batch processing, our sorting method is more efficient than the external merge sort.

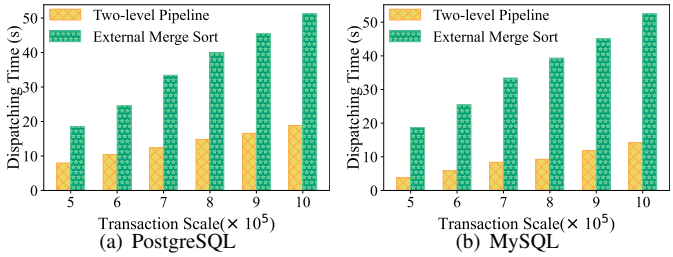


Fig. 17. Comparing Two-level Pipeline Sort with External Merge Sort

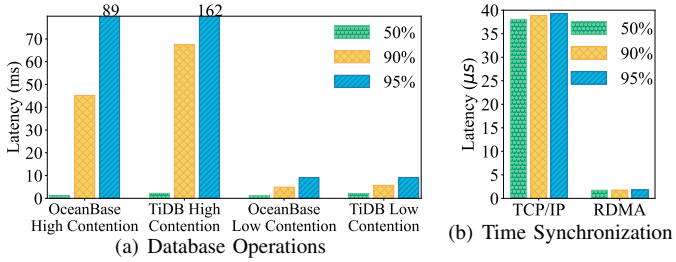


Fig. 18. Comparing Latencies of Time Synchronization and Database Operations

B. Evaluation on Clock Synchronization Cost

We conduct experiments to compare the latency of clock synchronization with the latency of operations in distributed DBMSs. Specifically, the clock synchronization is implemented on three servers. One server works as the centralized timestamp oracle, and the other two servers work as the clients which issue 100 threads to acquire timestamps concurrently. For the distributed DBMSs, they are deployed on three servers with three replicas.

We run standard benchmark TPC-C on two distributed DBMSs, i.e., TiDB and OceanBase, and report the tail latency of each operation. Specifically, as the workload contention heavily affects the operation latency, we simulate both the high and low contention scenarios. For a high contention, we run TPC-C with 72 threads and 3 warehouses (denoted as high contention), while for a high contention, we run TPC-C with 72 thread and 72 warehouses (denoted as low contention). For comparison, we report the tail latencies of clock synchronization service on the cluster which uses different network protocols, including TCP/IP and RDMA. Fig. 18 shows the results. We can see that,

- Under a high contention, the 50%/90%/95% latencies of TiDB (resp. OceanBase) are 60/1749/4125× (resp. 60/1158/2266×) higher than that of the clock synchronization service based on the TCP/IP protocol; Based on the RDMA protocol, the 50%/90%/95% latencies of TiDB (resp. OceanBase) are 1322/38590/90884× (resp. 1150/25538/49930×) higher than that of the clock synchronization service.
- Under a low contention, the 50%/90%/95% latencies of TiDB (resp. OceanBase) are 47/154/229× (resp. 45/116/229×) higher than that of the clock synchronization service based on the TCP/IP protocol. For the RDMA protocol, the 50%/90%/95% latencies of TiDB (resp. OceanBase) are 1035/3405/5049× (resp. 978/2554/5046×) higher than that of the clock synchronization service.

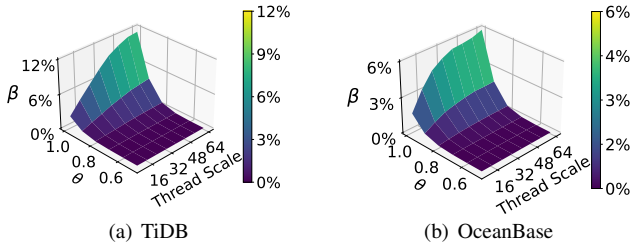


Fig. 19. Overlapping Ratio β

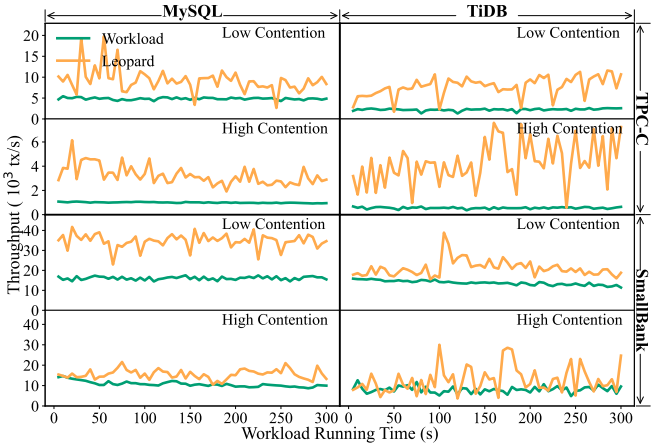


Fig. 20. Workload Throughput vs. Leopard Throughput on MySQL and TiDB

Usually, a distributed DBMS uses an extraordinarily complex protocol combination (e.g., the consensus protocol and the atomic commit protocol) to achieve transactional consistency over multiple remote machines, which would lead to a certain number of synchronous network round trips and disk data accesses. However, a logical timestamp acquisition only needs one network round trip. For DBMSs with particularly high throughputs, we suggest to use the RDMA network protocol to reduce the clock synchronization cost.

C. Overlapping Ratio on Distributed DBMSs

Workload contention has an obvious impact on overlapping ratio β. The standard benchmark YSCB-A can easily control workload contention by varying skew parameter θ and thread scale. Thus, we run YCSB-A on a table with 1 million records and the read/write ratio set as 4:1 to demonstrate the change of β on two distributed DBMSs, i.e., TiDB and OceanBase, as shown in Fig. 19(a) and Fig. 19(b). From the results, only when the skew parameter θ and thread scale are increased at the same time, β increases significantly. For the maximum value of β, TiDB's β is larger than the one of OceanBase. When increasing θ to 1 and thread scale to 64 on TiDB, β is still less than 10%. OceanBase has a smaller β than TiDB. This is because TiDB treats each transaction as a distributed transaction, while OceanBase tries to avoid the distributed transactions through the database sharding. The latency of DBMS processing operations increases the ratio of overlapping. We can see that although the overlapping ratio β

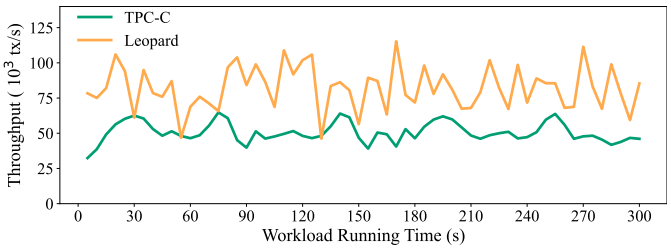


Fig. 21. Multiple-Leopard Throughput vs. High Throughput of OceanBase

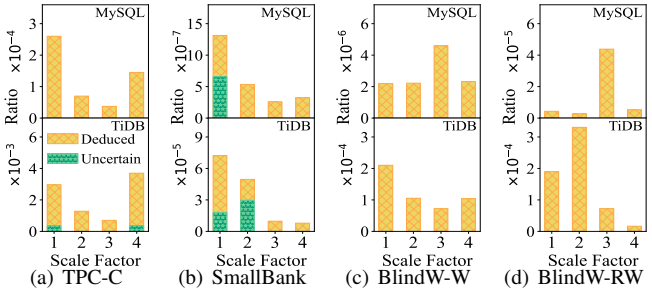


Fig. 22. Deducing Dependencies on MySQL and TiDB

in the distributed DBMS might be higher than the centralized DBMS, the value of β is still relatively small in all cases.

D. More Comparison with DBMS Throughput

To demonstrate the ability to keep up with DBMS throughputs, we launch the same experiments on two more popular DBMSs, i.e., MySQL ($v5.7$) and TiDB ($v5.0$). Fig. 20 shows the results. *Leopard* can be easily scaled out to keep up with the high throughput DBMS service.

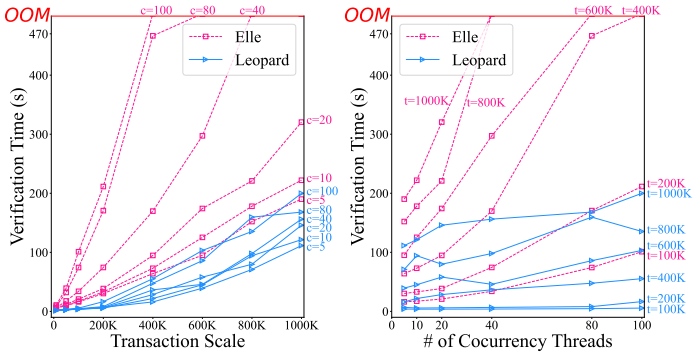
In addition, to further illustrate the scalability of *Leopard*, we have also conducted experiments by running TPC-C on OceanBase which is deployed on 20 cloud nodes connected using 1 Gigabit Ethernet. Each cloud node is equipped with 32 cores, 64 GB memory, and 100 GB HDD disk. Specifically, we deploy our Leopard only on 10 nodes. As shown in Fig. 21, we can see that the throughput of OceanBase can be as high as $65K$ transactions per second. In TPC-C, each transaction consists of $15 \sim 17$ operations on average. The peak throughput of *Leopard* can be as high as $115K$ transactions per second, which is much higher than the throughput of OceanBase. This indicates that *Leopard* scales well in the distributed setting.

E. Deducing Dependencies on More DBMSs

To demonstrate the generality of our approach on deducing dependencies, besides PostgreSQL, we also run the same experiments with four benchmarks, i.e., TPC-C, SmallBank, BlindW-W and BlindW-RW, on MySQL and TiDB to demonstrate the generality of our approach. Fig. 22 show the experiment results, which are similar with the results of PostgreSQL and OceanBase. In summary, *Leopard* can expose more uncertain dependencies effectively.

F. Comparison with Elle on Efficiency

Elle [19] carefully designs some specific workloads that can manifest dependencies for IL verification. Specifically, it takes the two properties to deduce dependencies, including



(a) Time vs. # of Transactions (b) Time vs. # of Concurrency Threads

Fig. 23. Comparison with Elle on Efficiency

recoverability and *traceability*. *Recoverability* means that every version of a read operation can be mapped to a specific write operation. *Traceability* means that the complete version order can be obtain by a read operation. Based on these two properties, Elle expects to deduce dependencies easily.

To find isolation anomalies, Elle puts these deduced dependencies into a graph, and search for cycles. Specifically, Elle applies Tarjan's algorithm to identify a cycle [75]. However, Tarjan's algorithm is superlinear with the number of transactions in the graph. To make matters worse, the number of transactions is massive for a long running workload. Elle then has poor scalability to a long running workload. To demonstrate it practically, we launch an experiment to compare the performance between *Leopard* and Elle.

We take the workload from the github repository of Elle [70], which can be verified by Elle. Specifically, the workload is composed of randomly generated transactions each of which covers one to five item-read/write operations and runs on a table of 1000 records. We run the workload on MySQL as varying the number of transactions (denoted as t) and the number of concurrency threads (denoted as c). Fig. 23 shows the verification efficiency of *Leopard* and Elle.

From results, Elle's verification time rises dramatically with the number of transactions and the number of concurrency threads, which is caused by the superlinear time complexity of Tarjan's algorithm. By contrast, *Leopard*'s verification time is linear with the number of transactions and the number of concurrency threads. Elle lacks the garbage collection for the dependency graph, so it is not suitable for verifying a long running workload that contains massive transactions and dependencies. Moreover, the graph expands significantly as the number of concurrency threads increases, which has a large number of conflicting operations (i.e., many dependencies on the graph). So, when the number of transactions or concurrency threads is up to a specific threshold, Elle terminates the verification and throws the out of memory exception (OOM). For example, when the number of concurrency threads is 100, Elle throws OOM at the number of transactions $400K$.

G. More Bug Case Demonstration

```
1 CREATE TABLE t(a INT PRIMARY KEY, b INT);
2 INSERT INTO t(676, -5012153);
3 BEGIN TRANSACTION;--TID:739
```

```
4   UPDATE t SET b=-5012153 WHERE a=676;--TID:739
5   UPDATE t SET b=-852150 WHERE a=676;--TID:723✖
6   COMMIT;--TID:739
```

Bug 1: Dirty Write on TiDB. Transaction $TID = 739$ writes a record, i.e., a=676, and then another transaction $TID = 723$ also writes this record before 739 commits, which results in a dirty write [35]. We find that the first update does not modify the record, leading to *TiDB* acquiring no lock. We report this bug to *TiDB*, which is confirmed and fixed.

```
1   CREATE TABLE t(a INT PRIMARY KEY, b INT);
2   CREATE TABLE s(a INT PRIMARY KEY, b INT);
3   ALTER TABLE s ADD FOREIGN KEY(b) REFERENCES t(a));
4   INSERT INTO t(1, 2);
5   INSERT INTO s(2, 1);
6   BEGIN TRANSACTION;--TID:211
7   UPDATE t SET b=3 WHERE a=1;--TID:211
8   SELECT * FROM t, s WHERE t.a=s.b AND s.a>1
9   FOR UPDATE; --TID:324, Result:{2,1,2}✖
10  COMMIT;--TID:211
```

Bug 2: Violating Mutual Exclusion on TiDB. Transaction $TID = 211$ acquires a long write lock on record 1 in table t, and another concurrent transaction $TID = 324$ successfully reads record 1 by *FOR UPDATE* statement, which violates the mutual exclusion between write locks. We report this bug to *TiDB*, which is confirmed and fixed.

```
1   CREATE TABLE t(a INT PRIMARY KEY, b INT);
2   CREATE TABLE s(a INT PRIMARY KEY, b INT);
3   ALTER TABLE s ADD FOREIGN KEY(b) REFERENCES t(a));
4   INSERT INTO t(1, 2);
5   INSERT INTO s(2, 1);
6   DELETE FROM s WHERE a=2;--TID:213
7   BEGIN TRANSACTION;--TID:412
8   INSERT INTO s VALUES(2,3);--TID:412
9   SELECT * FROM t WHERE a=2;
10  --TID:412, Result:{2,1},{2,3}✖
```

Bug 3: A Query Returning Two Versions of a Record on TiDB. Transaction $TID = 412$ returns two versions for a record. One is the version written by 412 itself, and the other is the deleted version, which should not be available. We report this bug to *TiDB* and confirmed that it is a known bug.