Contention-Aware Lock Scheduling for Transactional Databases

Boyu Tian
University of Michigan
Ann Arbor, MI, USA
bytian@umich.edu

Barzan Mozafari
University of Michigan
Ann Arbor, MI, USA
mozafari@umich.edu

Grant Schoenebeck
University of Michigan
Ann Arbor, MI, USA
schoeneb@umich.edu

ABSTRACT

Lock managers are among the most well-studied components in concurrency control and transactional systems. However, one question seems to have been generally overlooked: "when there are multiple lock requests on the same object, which one(s) should be granted first?"

Nearly all existing systems rely on a FIFO (first in, first out) strategy to decide which transaction(s) to grant the lock to. However, in this paper, we show that the choice of lock scheduling has significant ramifications on the overall performance of a transactional system. Despite the large body of research on job scheduling outside the database context, lock scheduling presents subtle but challenging requirements that render existing results on scheduling inapt for a transactional database. By carefully studying this problem, we present the concept of contention-aware scheduling, show the hardness of the problem, and propose novel lock scheduling algorithms (LDSF and bLDSF), which guarantee a constant factor approximation of the best scheduling. We conduct extensive experiments using a popular database on both TPC-C and a microbenchmark. Compared to FIFO—the default scheduler in most database systems—our bLDSF algorithm yields up to 300x speedup in overall transaction latency. On the other hand, our LDSF algorithm, which is simpler and achieves comparable performance to bLDSF, has already been adopted by open-source community, and chosen as default scheduling strategy in MySQL 8.0.3+.

1. INTRODUCTION

Lock management forms the backbone of concurrency control in modern software, including many distributed systems and transactional databases. A lock manager guarantees both correctness and efficiency of a concurrent application by solving the data contention problem. For example, before a transaction accesses a database object, it has to acquire the corresponding lock; if the transaction fails to get a lock immediately, it is blocked until the system grants it the lock. This poses a fundamental question: when multiple transactions are waiting for a lock on the same object, which should be granted first when the object becomes available? This question, which we call *lock scheduling*, has received surprisingly little attention, despite the large body of work on concurrency control and locking protocols [15, 45, 8, 65, 18, 40, 50, 54, 23]. In fact, almost all existing DBMSs rely on variants of a basic first-in-first-out (FIFO) strategy, which grants (all) compatible lock requests based on their arrival time in the queue [1, 3, 4, 5, 6]. In this paper, we carefully study the problem of lock scheduling and show that it has significant ramifications on overall performance of a DBMS.

Related Work — There is a long history of research on scheduling problems in a general context [25, 42, 66, 67, 61, 41, 35, 59], whereby a set of jobs is to be scheduled on a set of processors such that a goal function is minimized, e.g., the sum of (weighted) completion times [61, 41, 39] or the variance of the completion or wait times [14, 17, 72, 49, 28]. There is also work on scheduling in a real-time database context [71, 40, 8, 36, 70], where the goal is to minimize the total tardiness or the number of transactions missing their deadlines.

In this paper, we address the problem of lock scheduling in a transactional context, where jobs are transactions and processors are locks, and the scheduling decision is about which locks to grant to which transactions. However, our transactional context makes this problem quite different than the well-studied variants of the scheduling problem. First, unlike generic scheduling problems, where at most one job can be scheduled on each processor, a lock may be held in either exclusive or inclusive modes. The fact that transactions can sometimes share the same resources (i.e., shared locks) significantly complicates the problem (see Section 3.4). Moreover, once a lock is granted to a transaction, the same transaction may later request another lock (as opposed to jobs requesting all of their needed resources upfront). Finally, in the scheduling literature, the execution time of each job is assumed to be known upon its arrival [52, 67, 14, 72], whereas the execution time of a transaction is often unknown *a priori*.

Although there are scheduling algorithms designed for real-time databases [55, 68, 73, 13], they are not applicable in general DBMS context. For example, real-time settings assume that each transaction comes with a deadline, whereas most database workloads do not have explicit deadlines. Instead, most workloads wish to minimize latency or maximize throughput.

Challenges — Several aspects of lock scheduling make it a uniquely challenging problem, particularly under the performance considerations of a real-world DBMS.

1. **An online problem.** At the time of granting a lock to a transaction we do not know when the lock will be released, since the transaction's execution time will only be known once it is finished.
2. **Dependencies.** In a DBMS, there are dependencies among concurrent transactions when one is waiting for a lock held by another. In practice, these dependencies

can be quite complex, as each transaction can hold locks on several objects and several transactions can hold inclusive locks on the same object.

3. **Non-uniform access patterns.** Not all objects in the database are equally popular. Also, different transaction types might each have a different access pattern.
4. **Multiple locking modes.** The possibility of granting a lock to one writer exclusively or to multiple readers inclusively is a source of great complexity (see Section 3.4).

Contributions — In this paper, to the best of our knowledge, we present the first formal study of lock scheduling problem with a goal of minimizing transaction latencies in a DBMS context. Furthermore, we propose a contention-aware transaction scheduling algorithm, which captures the contention and the dependencies among concurrent transactions. The key insight is that a transaction blocking many others should be scheduled earlier. We carefully study the difficulty and optimality of our algorithm. Most importantly, we show that our results are not merely theoretical, but lead to dramatic speedups in a real-world DBMS. Our ultimate hope is that our results will bring attention to the significant ramifications of the choice in lock scheduling algorithm on the overall performance of a transactional system.[1] In summary, we make the following contributions:

1. We propose a contention-aware lock scheduling algorithm, called Largest-Dependency-Set-First (LDSF). We prove that, in the absence of inclusive locks, LDSF is optimal in terms of the expected mean latency (Theorem 2). With inclusive locks, we prove that LDSF is a constant factor approximation of the optimal scheduling under certain regularity constraints (Theorem 3).
2. We propose the idea of granting only *some* of the inclusive lock requests on an object (as opposed to granting them *all*). We study the difficulty of the scheduling problem under this setting (Theorem 5), and propose another algorithm, called bLDSF (batched Largest-Dependency-Set-First), which improves upon LDSF in this setting. We prove that bLDSF is also a constant factor approximation of the optimal scheduling (Theorem 6).
3. In addition to our theoretical analysis, we use a real-world DBMS and extensive experiments to empirically evaluate our algorithms on the TPC-C benchmark, as well as a microbenchmark. Our results confirm that, compared to the commonly used FIFO strategy, LDSF and bLDSF reduce mean transaction latencies by up to 300x and 290x, respectively. They also increase throughput by up to 6.5x and 5.5x. As a result, LDSF (which is simpler than bLDSF) has already been adopted as the default scheduling algorithm in all MySQL (8.0.3+) distributions.

2. RELATED WORK

In short, the large body of work on traditional job scheduling is unsuitable in a database context due to the unique requirements of locking protocols deployed in databases. Although there is some work on lock scheduling for real-time databases, they aim at supporting explicit deadlines rather than minimizing the mean latency of transactions.

Job scheduling — Outside the database community, there has been extensive research on scheduling problems in general. Here, the duration (and sometimes the weight and arrival time) of each task is known *a priori*, and a typical goal is to minimize (i) the sum of (weighted) completion times (SCT) [61, 41, 39], (ii) the latest completion time [20, 34, 64], (iii) the completion time variance (CTV) [14, 17, 72, 49], or even (iv) the waiting time variance (WTV) [28].

None of these results are applicable to our setting, mainly because of their assumption that each processor/worker can be used by only one job at a time, whereas in a database locks can be held both inclusively and exclusively. Moreover, they assume the execution time of each job is known, which is not the case in a database (i.e., the database does not know when the application/user will commit and release its locks). Finally, with the exception of [61, 41], prior work on scheduling either assumes that all tasks are available at the beginning, or that their arrival time is known. In a database, however, such information is unavailable.

Dependency-based scheduling — Scheduling tasks with dependencies among them has been studied for both single machines [66, 42] and multiprocessors [25, 26, 30, 58]. Here, each job only needs one processor and once scheduled, it will not be blocked again. However, in a database, a transaction can request many locks, and thus, can be blocked even after it is granted some locks.

Real-time databases (RTDB) — There is some work on lock scheduling in the context of RTDBs, where transactions are scheduled to meet a set of user-given deadlines [71, 8, 55, 68, 73, 13, 36, 70, 69, 38, 33, 9, 53, 63, 19, 40]. It is shown that the First-In-First-Out (FIFO) policy performs poorly in this setting [33, 8, 9, 53], compared to the Earliest-Deadline-First policy [55, 68, 73], which is also used in practice [13].

Unfortunately, the work in this area is not applicable to general-purpose database systems. First, in an RTDB, each transaction comes with a pre-specified deadline, while in a general-purpose database such deadlines are not provided. Second, a key assumption in this line of work is that the execution time of each transaction is known in advance, whereas in a general database the execution time of a transaction is only known once it is finished. Finally, the scheduling goal in a RTDB is to minimize the total tardiness or the number of missed deadlines. In other words, as long as a transaction meets its deadline, they do not care whether it finishes right before the deadline or much earlier. In contrast, general databases aim to execute transactions as fast as possible.

Scheduling in existing DBMS — For simplicity and fairness [12], the First-In-First-Out (FIFO) policy and its variants are the most widely adopted scheduling policies in many of today's databases [10], operating systems [16], and communication networks [51]. For example, FIFO is the default lock scheduling policy in MySQL [3], MS SQL Server [6], Postgres [5], Teradata [4], and DB2 [1]. Despite its popularity, FIFO does not provide any guarantees in terms of average or percentile latencies. Huang et al. [44] propose a scheduling algorithm, called Variance-Aware Transaction Scheduling (VATS), which aims at minimizing the variance of transaction latencies, and its optimality holds only when there are no shared locks in the system. In contrast, we focus

[1] Despite decades of research on all aspects of transaction processing, the importance of lock scheduling seems to have gone unnoticed, to the extent that all DBMSs still use FIFO.

on minimizing mean latency, and allow for both shared and exclusive locks.[2] In short, designing optimal lock scheduling algorithms for databases has remained an open problem.

Deadlock resolution — The problem of *deadlock resolution* is about deciding which transaction(s) to abort (a.k.a. victims) in order to resolve a deadlock [62, 37, 56, 43, 32]. Typically, transactions with lower priority are chosen as victims in order to reduce the abortion cost or the amount of work wasted. Here, a transaction's priority can be based on its age [32, 11], its deadline [47], or the number of locks it holds [57]. Franaszek et al. [32] empirically show that an age-based priority improves concurrency, and reduces the amount of work wasted. Agrawal et al. [11] argue that choosing victims based on their age and the number of currently held locks leads to fewer rollbacks, compared to (i) choosing a transaction randomly, or (ii) aborting the most recently blocked transaction.

These proposals take contention into consideration, but only for deadlock resolution. In this paper, we focus on lock scheduling and show that contention-aware scheduling can yield significant performance improvements in practice.

3. PROBLEM STATEMENT

In this section, we first describe our problem setting and define dependency graphs. We then formally state the lock scheduling problem.

3.1 Background: Locking Protocols

Locks are the most commonly used mechanism for ensuring consistency when a set of shared objects are concurrently accessed by multiple transactions (or applications). In a locking system, there are two main types of locks: inclusive and exclusive. Before a transaction can read an object (e.g., a row), it must first acquire an inclusive lock (a.k.a. shared or read lock) on that object. Likewise, before a transaction can write to or update an object, it must acquire an exclusive lock (a.k.a. write lock) on that object. An inclusive lock can be granted on an object as long as no exclusive locks are currently held on that object. However, an exclusive lock on an object can be granted only if there are no other locks currently held on that object. We focus on the strict 2-phase locking (strict 2PL) protocol: once a lock is granted to a transaction, it is held until that transaction ends. Once a transaction finishes execution (i.e., it commits or gets aborted), it releases all its locks.

3.2 Dependency Graph

Given the set T of transactions currently in the system, and the set O of objects in the database, we define the dependency graph of the system as an edge-labeled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$. The vertices of the graph $\mathcal{V} = T \cup O$ consist of the current transactions and objects. The edges of the graph $\mathcal{E} \subseteq T \times O \cup O \times T$ describe the locking relationships among the objects and transactions. Specifically, for transaction $t \in T$ and object $o \in O$,

- $(t, o) \in \mathcal{E}$ if t is waiting for a lock on o;
- $(o, t) \in \mathcal{E}$ if t already holds a lock on o.

[2]For naming consistency, our bLDSF algorithm is called *VATS version 3.0* in MySQL's codebase [2], overriding the original algorithm proposed by Huang et al. [44], called *VATS version 1.0 [7]*.

The label $\mathcal{L} : \mathcal{E} \to \{S, X\}$ indicates the lock type:

- $\mathcal{L}(t, o) = X$ if t is waiting for an exclusive lock on o;
- $\mathcal{L}(t, o) = S$ if t is waiting for an inclusive lock on o;
- $\mathcal{L}(o, t) = X$ if t already holds an exclusive lock on o;
- $\mathcal{L}(o, t) = S$ if t already holds an inclusive lock on o.

We assume that deadlocks are rare and are handled by an external process (e.g., a deadlock detection and resolution module). Thus, for simplicity, we assume that the dependency graph $\mathcal{G}$ is always a directed acyclic graph (DAG).

3.3 Lock Scheduling

A lock scheduler makes decisions about which transactions are granted the locks upon one or both of the following events: (i) when a transaction requests a lock, and (ii) when a lock is released by a transaction.[3] Let $\mathbb{G}$ be the set of all possible dependency graphs of the system. A scheduling algorithm $\mathcal{A} = (\mathcal{A}_{req}, \mathcal{A}_{rel})$ is a pair of functions $\mathcal{A}_{req}, \mathcal{A}_{rel} : \mathbb{G} \times O \times T \times \{S, X\} \to 2^T$. For example, when transaction t requests an exclusive lock on object o, $\mathcal{A}_{req}(\mathcal{G}, o, t, X)$ determines which of the transactions currently waiting for a lock on o (including t itself) should be granted their requested lock on o, given the dependency graph $\mathcal{G}$ of the system. (Note that the types of locks requested by transactions other than t are captured in $\mathcal{G}$.) Likewise, when transaction t releases a shared lock on object o, $\mathcal{A}_{rel}(\mathcal{G}, o, t, S)$ determines which of the transactions currently waiting for a lock on o should be granted their requested lock, given the dependency graph $\mathcal{G}$. When all transactions holding a lock on an object o release the lock, we say that o has *become available*. When the lock request of a transaction t is granted, we say that t is scheduled.

Since the execution time of each transaction is typically unknown in advance, we model their execution time using a random variable with expectation R. Given a particular scheduling algorithm $\mathcal{A}$, we define the latency of a transaction t, denoted by $l_{\mathcal{A}}(t)$, as its execution time plus the total time it has been blocked waiting for various locks. Since $l_{\mathcal{A}}(t)$ is a random variable, we denote its expectation as $\bar{l}_{\mathcal{A}}(t)$. We use $\bar{l}(\mathcal{A})$ to denote the expected transaction latency under algorithm $\mathcal{A}$, which is defined as the average of the expected latencies of all transactions in the system, i.e., $\bar{l}(\mathcal{A}) = \frac{1}{|T|} \sum_{t \in T} \bar{l}_{\mathcal{A}}(t)$.

Our goal is to find a lock scheduling algorithm under which the expected transaction latency is minimized. To ensure consistency and isolation, in most database systems $\mathcal{A}_{req}$ simply grants a lock to the requesting transaction only when (i) no lock is held on the object, or (ii) the currently held lock and the requested lock are compatible and no transaction in the queue has an incompatible lock request. This choice of $\mathcal{A}_{req}$ also ensures that transactions requesting exclusive locks are not starved. The key challenge in lock scheduling, then, is choosing an $\mathcal{A}_{rel}$ such that the expected transaction latency is minimized.

3.4 NP-Hardness

Minimizing the expected transaction latency under the scheduling algorithm is, in general, an NP-hard problem.

[3]These are the only occasions in which the dependency graph changes. If a scheduler grants locks at other times, the same decision could have been made upon the previous event, i.e., a transaction was unnecessarily blocked. A lock scheduler is thus an event-driven scheduler.

Notation	Description
T	the set of transactions in the system
O	the set of objects in the database
$\mathcal{G}$	the dependency graph of the system
$\mathcal{V}$	vertices in the dependency graph
$\mathcal{E}$	edges in the dependency graph
$\mathcal{L}$	labels of the edges indicating the lock type
$\mathcal{A}$	a scheduling algorithm
$l_{\mathcal{A}}(t)$	the latency of transaction t under $\mathcal{A}$
$\bar{l}_{\mathcal{A}}(t)$	the expectation of $l_{\mathcal{A}}(t)$
$\bar{l}(\mathcal{A})$	the expected transaction latency under $\mathcal{A}$

Table 1: Table of Notations.

Intuitively, the hardness is due to the presence of inclusive locks, which cause the system's dependency graph to be a DAG, but not necessarily a tree.

THEOREM 1. *Given a dependency graph $\mathcal{G}$, when a transaction t releases a lock (S or X) on object o, it is NP-hard to determine which pending lock requests to grant, in order to minimize the expected transaction latency. The result holds even if all transactions have the same execution time, and no transaction requests additional locks in the future.*

The proof of the theorem is deferred to Appendix A

Given the NP-hardness of the problem in general, in the rest of this paper, we propose algorithms that guarantee a constant-factor approximation of the optimal scheduling in terms of the expected transaction latency.

4. CONTENTION-AWARE SCHEDULING

We define contention-aware scheduling as any algorithm that prioritizes transactions based on their impact on the overall contention of the system. In this section, we first study several heuristics for comparing the contribution of different transactions to the overall contention, and illustrate their shortcomings through intuitive examples. We then propose a particular contention-aware scheduling that formally quantifies this contribution, and guarantees a constant-factor approximation of the optimal scheduling when inclusive locks are not held by too many transactions. (Later, in Section 5, we generalize this algorithm for situations where this assumption does not hold.)

4.1 Capturing Contention

The degree of contention in a database system is directly related to the number of transactions concurrently requesting conflicting locks on the same objects.

For example, a transaction holding an exclusive lock on a popular object will naturally block many other transactions requesting a lock on that same object. If such a transaction is itself blocked (e.g., waiting for a lock on a different object), it will negatively affect the latency of many transactions, increasing overall contention in the system. Thus, our goal in contention-aware scheduling is to determine which transactions have a more important role in reducing the overall contention in the system, so that they can be given higher priority when granting a lock. Next, we discuss heuristics for measuring the priority of a transaction in reducing the overall contention.

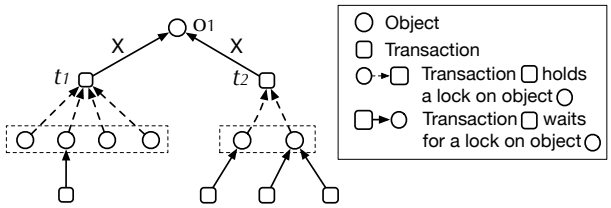

Figure 1: Transaction t_1 holds the most number of locks, but many of them on unpopular object.

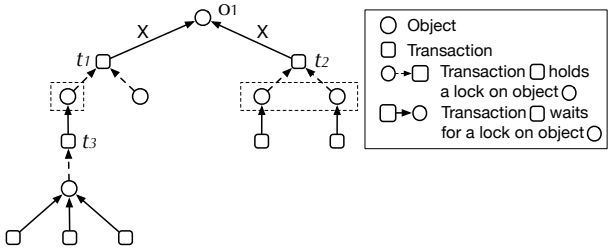

Figure 2: Transaction t_2 holds two locks that are waited on by other transactions. Although only one of t_1's locks is blocking other transactions, the blocked transaction (i.e., t_3) is itself blocking three others.

Number of locks held — The simplest criterion for prioritizing transactions is the number of locks they currently hold. We refer to this heuristic as Most Locks First (MLF). The intuition is that a transaction with more locks is more likely to block other transactions in the system. However, this approach does not account for the popularity of objects in the system. In other words, a transaction might be holding many locks but on an unpopular objects, which are unlikely to be requested by other transactions. Prioritizing such a transaction will not necessarily reduce contention in the system. Figure 1 demonstrates an example, where transaction t_1 holds the most number of locks but on unpopular objects. It is therefore better to keep t_1 waiting and instead schedule t_2 first, which holds fewer locks but on more popular objects.

Number of locks that block other transactions — An improvement over the previous criterion is to only count those locks that have at least one transaction waiting on them. This approach disregards transactions that hold many locks, but on these locks no other transactions are waiting. We call this heuristic Most Blocking Locks First (MBLF). The issue with this criterion is that it treats all blocked transactions as the same, even if they contribute unequally to the overall contention. Figure 2 shows an example in which the scheduler must decide between transactions t_1 and t_2 when the object o_1 becomes available. Here, this criterion would choose t_2, which currently holds two locks, each at least blocking one other transaction. However, although t_1 holds only one blocking lock, it is blocking t_3 which itself is blocking three other transactions. Thus, by scheduling t_2 first, t_3 and its three subsequent transactions will remain blocked in the system for a longer period of time than if t_1 had been scheduled first.

Depth of the dependency subgraph — A more sophisticated criterion is the depth of a transaction's dependency subgraph. For a transaction t, this is defined as the subgraph of the dependency graph comprised of all vertices that can reach t (and all edges between such vertices). The depth of

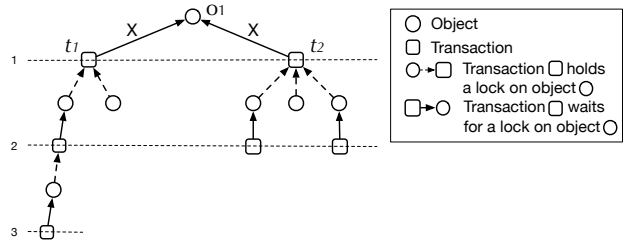

Figure 3: Transaction t_1 has a deeper dependency subgraph, but granting the lock to t_2 will unblock more transactions which can run concurrently.

t's dependency subgraph is characterized by the number of transactions on the longest path in the subgraph that ends in t. We refer to this heuristic as Deepest Dependency First (DDF). Figure 3 shows an example, where the depth of the dependency subgraph of transaction t_1 is 3 while that of transaction t_2 is only 2. Thus, when deciding between t_1 and t_2 based on this criterion, the exclusive lock on object o_1 should be granted to t_1. The idea behind this heuristic is that a longer path indicates a larger number of transactions that are sequentially blocked. Thus, to unblock such transactions sooner, the scheduling algorithm must start with a transaction whose dependency graph is deeper. However, considering only the depth of this subgraph can limit the overall degree of concurrency in the system. For example, in Figure 3, if the exclusive lock on o_1 is granted to t_1 instead of t_2, upon the completion of t_1 only one transaction in its dependency subgraph will be unblocked and can resume execution. On the other hand, if the lock is granted to t_2, upon its completion two other transactions in its dependency subgraph will be unblocked, which can run concurrently.

Later, in Section 7.4, we empirically evaluate these heuristics. While none of these heuristics alone is able to guarantee an optimal lock scheduling strategy, they offer valuable insight in understanding the relationship between scheduling and overall contention. In particular, the first two heuristics focus on what we call *horizontal contention*, whereby a transaction holds locks on many objects directly needed by other transactions. In contrast, the third heuristic focuses on reducing *vertical contention*, whereby a chain of dependencies causes a series of transactions to block each other. Next, we present an algorithm which is capable of resolving both horizontal and vertical aspects of contention.

4.2 Largest-Dependency-Set-First

In this section, we propose an algorithm, called Largest-Dependency-Set-First (LDSF), which provides formal guarantees on the expected mean latency.

Consider two transactions t_1 and t_2 in the system. If there is a path from t_1 to t_2 in the dependency graph, we say that t_1 is dependent on t_2 (i.e., t_1 depends on t_2's completion/abortion for at least one of its required locks). We define the dependency set of t, denoted by $g(t)$, as the set of all transactions that are dependent on t (i.e., the set of transactions in t's dependency subgraph). Our LDSF algorithm uses the size of the dependency sets of different transactions to decide which one(s) to schedule first. For example, in Figure 4, there are five transactions in the dependency set of transaction t_1 (including t_1 itself) while there are four transactions in t_2's dependency set. Thus, in a situation where both t_1 and t_2 have requested an exclusive lock on object o_1, LDSF grants the lock to t_1 (instead of t_2) as soon as o_1 becomes available.

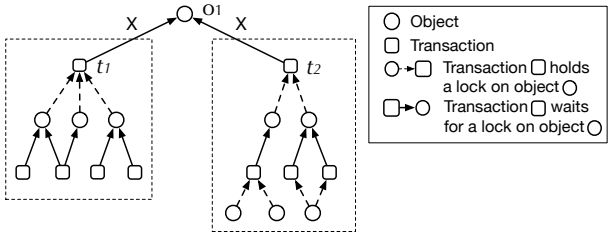

Figure 4: Lock scheduling based on the size of the dependency sets.

Input : The dependency graph of the system $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$, transaction t, object o, label $L \in \{X, S\}$
// meaning t has just released a lock of type L on o
Output: The set of transactions whose requested lock on o should be granted

if *there are other transactions still holding a lock on* o **then**
  return $\emptyset$;
Obtain the set of transactions waiting for an inclusive lock on o, $T^i \leftarrow \{t^i \in \mathcal{V} : (t^i, o) \in \mathcal{E} \text{ and } \mathcal{L}(t^i, o) = S\} = \{t^i_1, t^i_2, \cdots, t^i_m\}$;
Obtain the set of transactions waiting for an exclusive lock on o, $T^x \leftarrow \{t^x \in \mathcal{V} : (t^x, o) \in \mathcal{E} \text{ and } \mathcal{L}(t^x, o) = X\} = \{t^x_1, t^x_2, \cdots, t^x_n\}$;
Let $\tau(T^i) = \left|\bigcup_{i=1}^{n} g(t^i_i)\right|$;
Find a transaction $t^{\hat{x}} \in W$ s.t. $|g(t^{\hat{x}})| = \max_{t^x_i \in T^x} \left|g(t^x_i)\right|$;
if $\tau(T^i) < \left|g(t^{\hat{x}})\right|$ **then**
  return T^i;
else
  return $\{t^{\hat{x}}\}$;

Algorithm 1: *Largest-Dependency-Set-First Algorithm*

Now, we can formally present our LDSF algorithm. Suppose an object o becomes available (i.e., all previous locks on o are released), and there are $m + n$ transactions currently waiting for a lock on o: m transactions $t^i_1, t^i_2, \cdots, t^i_m$ are requesting an inclusive lock o, and n transactions $t^x_1, t^x_2, \cdots, t^x_n$ are requesting an exclusive lock on object o. Our LDSF algorithm defines the priority of each transaction t^x_i requesting an exclusive lock as the size of its dependency set, $|g(t^x_i)|$. However, LDSF treats all transactions requesting an inclusive lock on o, namely $t^i_1, t^i_2, \cdots, t^i_m$, as a single transaction—if LDSF decides to grant an inclusive lock, it will be granted to all of them. The priority of the inclusive requests is thus defined as the size of the union of their dependency sets, $\left|\bigcup_{i=1}^{m} g(t^i_i)\right|$. LDSF then finds the transaction $t^{\hat{x}}$ with the highest priority among $t^x_1, t^x_2, \cdots, t^x_n$. If $t^{\hat{x}}$'s priority is higher than the collective priority of the transactions requesting an inclusive lock, LDSF grants the exclusive lock to $t^{\hat{x}}$. Otherwise, an inclusive lock is granted to all transactions $t^i_1, t^i_2, \cdots, t^i_m$. The pseudo-code of the LDSF algorithm is provided in Algorithm 1.

Analysis — We do not make any assumptions about the future behavior of a transaction, as they may request various locks throughout their lifetime. Furthermore, since we cannot predict new transactions arriving in the future, in our analysis, we only consider the transactions that are already in the system. Since the system does not know the execution time of a transaction *a priori*, we model the execution time of each transaction as a memoryless random variable. That is, the time a transaction has already spent in execution does not necessarily reveal any information about the trans-

action's remaining execution time. We denote the remaining execution time as a random variable R with expectation $\bar{R}$. We also assume that the execution time of a transaction is not affected by the scheduling.[4] Transactions that are sensitive to the actual time (e.g., stop if run before a certain time 2pm, otherwise run for a long time) are also excluded from our discussion.

We first study a simplified scenario, in which there are only exclusive locks in the system (we relax this assumption in Theorem 2). The following theorem states that LDSF minimizes the expected latency in this scenario.

THEOREM 2. *When there are only exclusive locks in the system, the LDSF algorithm is the optimal scheduling algorithm in terms of the expected latency.*

PROOF. Let $w_{\mathcal{A}}(t)$ be a random variable representing the total time transaction t will eventually spend in the system waiting for various locks, with expectation $\bar{w}_{\mathcal{A}}(t)$. In other words, the latency of transaction t under scheduling algorithm $\mathcal{A}$ can be modeled as $l_{\mathcal{A}}(t) = w_{\mathcal{A}}(t) + R$. Thus, minimizing the expected transaction latency under $\mathcal{A}$ is equivalent to minimizing the expected wait time under $\mathcal{A}$, defined as $\bar{w}(\mathcal{A}) = \frac{1}{|T|}\sum_{t\in T}\bar{w}_{\mathcal{A}}(t)$. Apparently, we have $\bar{l}(\mathcal{A}) = \bar{w}(\mathcal{A}) + \bar{R}$.

For each transaction $t \in T$ which is currently waiting for a lock, let o be an object reachable from t in the dependency graph that is locked by a running transaction (i.e, a critical object of t as defined in Section 4.2). Then we say that t is delayed by object o. Suppose all locks in the system are exclusive, that is, each lock can be held by at most one transaction. Then, the dependency graph of the system $\mathcal{C}$, which is a DAG, becomes a forest (a set of disjoint trees). Therefore, each transaction is delayed by at most one object.

Given a scheduling algorithm $\mathcal{A}$, let $d_{\mathcal{A}}(t, o)$ be the expected time transaction t is delayed by object o. Then, the expected wait time of transaction t is given by

$$\bar{w}_{\mathcal{A}}(t) = \sum_{o\in O} d_{\mathcal{A}}(t, o). \tag{1}$$

Therefore, the mean wait time by algorithm $\mathcal{A}$ is given by

$$\begin{aligned}\bar{w}(\mathcal{A}) &= \frac{1}{|T|}\sum_{t\in T}\bar{w}_{\mathcal{A}}(t)\\ &= \frac{1}{|T|}\sum_{t\in T}\sum_{o\in O} d_{\mathcal{A}}(t, o)\\ &= \frac{1}{|T|}\sum_{o\in O}\sum_{t\in g(o)} d_{\mathcal{A}}(t, o),\end{aligned}$$

where $g(o)$ is defined as the set of transactions that can reach object o in the dependency graph $\mathcal{G}$.

Let $d_{\mathcal{A}}(o) = \sum_{t\in g(o)} d_{\mathcal{A}}(t, o)$. Then, the mean expected wait time is

$$\bar{w}(\mathcal{A}) = \frac{1}{|T|}\sum_{o\in O} d_{\mathcal{A}}(o). \tag{2}$$

Therefore, minimizing $\bar{d}_{\mathcal{A}}$ is equivalent to minimizing $d_{\mathcal{A}}(o)$ for each $o \in O$.

For a given object o that has become available, suppose there are n transactions waiting on it, denoted by $t_1, t_2, \cdots, t_n$. We also assume that the transactions are sorted in the decreasing order of their dependency set sizes, that is:

$$|g(t_1)| \geq |g(t_2)| \geq \cdots \geq |g(t_n)|.$$

Let $a_1, a_2, \cdots, a_n$ be the expected time that $t_1, t_2, \cdots, t_n$, respectively, are granted the lock. Assume that the current time is considered as 0. Since the execution time of a transaction is characterized by a memoryless random variable R, the remaining time of the transaction R^{rem} has the same distribution as R. Therefore, its expectation is given by $\bar{R}$ and $\{a_1, a_2, \cdots, a_n\} = \{0, \bar{R}, \cdots, (n-1)\bar{R}\}$.

For transaction t_i, the time that t_i is delayed by object o is a_i. The time that the other transactions in $g(t_i)$ are delayed by o is given by

$$(|g(t_i)| - 1)(a_i + \bar{R}),$$

since these transactions are also delayed by o during the time t_i is being executed.

Thus, the expected time for all transactions to be delayed by object o is

$$d_{\mathcal{A}}(o) = \sum_{i=1}^{n} a_i + (|g(t_i)| - 1)(a_i + \bar{R}).$$

By the rearrangement inequality, $d_{\mathcal{A}}(o)$ is minimized only when $a_1 \leq a_2 \leq \cdots \leq a_n$, which corresponds to our LDSF algorithm. In this situation, $\bar{w}(\mathcal{A})$ and, thus, $\bar{l}(\mathcal{A})$ are minimized as well.

Therefore, the LDSF algorithm results in the minimum expected latency $\bar{d}_{\mathcal{A}}$ of the transactions in the system. □

The intuition here is that if a transaction t_1 is dependent on t_2, any progress in the execution of t_2 can also be considered as t_1's progress since t_1 cannot receive its lock unless t_2 finishes execution. Thus, by granting the lock to the transaction with the largest dependency set, LDSF allows the most transactions to make progress toward completion.

However, this does not necessarily hold true with the existence of inclusive locks. Even if transaction t_1 is dependent on t_2, the execution of t_2 does not necessarily contribute to t_1's progress. Specifically, consider the set of all objects that are reachable from t_1 in the dependency graph, but are locked (inclusively or exclusively) by *currently running* transactions. We call these objects the *critical objects* of t_1, and denote them as $C(t_1)$.[5] For example, in Figure 5, we have $C(t_1) = \{o_1, o_2\}$. Note that not all transactions that hold a lock on a critical object of t_1 contribute to t_1's progress. Rather, only the transaction that releases the last lock on that critical object allows for the progress of t_1. In the example of Figure 5, the execution of t_2 does not necessarily contribute to the progress of t_1, unless t_3 releases the lock earlier than t_2.

Nonetheless, if the number of transactions waiting for each shared lock is bounded, then LDSF is a constant-factor approximation of the optimal scheduler in terms of the expected latency.

THEOREM 3. *Let the maximum number of critical objects for any transaction in the system be c. Assume that the number of transactions waiting for shared locks on the same object is bounded by u. The LDSF algorithm is a $(c \cdot u)$-approximation of the optimal scheduling (among strategies*

[4]For example, scheduling causes context switches, which may affect performance. For simplicity, in our formal analysis, we assume that their overall effect is not significant.

[5]Note that the critical objects of a transaction may change throughout its lifetime.

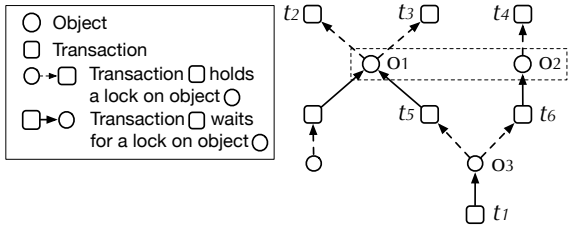

Figure 5: The critical objects of t_1 are o_1 and o_2, as they are locked by transactions t_2 and t_3. Note that, although o_3 is reachable from t_1, it is not a critical object of t_1 since it is locked by transactions that are not currently running, i.e., t_5 and t_6 which themselves are waiting for other locks.

that grant all shared locks simultaneously) in terms of the expected latency.

PROOF. Since a lock can be held by multiple transactions, a transaction can have multiple critical objects. This means that Equation (1) does not hold anymore.

Assume that a transaction has at most c critical objects, i.e., each transaction t is delayed by at most c objects. Instead of Equation (1), we have

$$\bar{w}_{\mathcal{A}}(t) \leq \sum_{o \in O} d_{\mathcal{A}}(t, o) \leq c \cdot \bar{w}_{\mathcal{A}}(t). \tag{3}$$

Let $\hat{w}(\mathcal{A}) = \frac{1}{|T|} \sum_{o \in O} d_{\mathcal{A}}(o)$. We have the following:

$$\bar{w}(\mathcal{A}) \leq \hat{w}(\mathcal{A}) \leq c \cdot \bar{w}(\mathcal{A}). \tag{4}$$

Suppose that $f(k)$ be the delay factor defined in Section 5.1. Let $\tilde{w}(\mathcal{A})$ be the value of $\hat{w}(\mathcal{A})$ when all transactions take $f(u)$ time to finish. Then,

$$\bar{w}(\mathcal{A}) \leq \tilde{w}(\mathcal{A} \leq f(u)\hat{w}(\mathcal{A}) \leq u \cdot \hat{w}(\mathcal{A}). \tag{5}$$

By the same argument in the proof of Theorem 2, we can prove that LDSF minimizes $\tilde{w}(\mathcal{A})$. Let $\hat{\mathcal{A}}$ be our LDSF algorithm, and $\mathcal{A}_{OPT}$ be the optimal scheduling algorithm. Therefore, by Equation (4) and (5),

$$\begin{aligned} \bar{w}(\hat{\mathcal{A}}) &\leq \tilde{w}(\hat{\mathcal{A}}) \\ &\leq \tilde{w}(\mathcal{A}_{OPT}) \\ &\leq u \cdot \hat{w}(\mathcal{A}_{OPT}) \\ &\leq u \cdot c \cdot \bar{w}(\mathcal{A}_{OPT}). \end{aligned}$$

Therefore, $\bar{l}(\hat{\mathcal{A}}) = \bar{w}(\hat{\mathcal{A}}) + \bar{R} \leq c \cdot u \cdot (\bar{w}(\mathcal{A}_{OPT}) + \bar{R}) = c \cdot u \cdot \bar{l}(\mathcal{A}_{OPT})$, which means that LDSF is a $(c \cdot u)$-approximation of the optimal algorithm in terms of expected latency. □

5. SPLITTING SHARED LOCKS

In the LDSF algorithm, when a shared lock is granted, it is granted to all transactions waiting for it. In Section 5.1, we show why this may not be the best strategy. Then, in Section 5.2, we propose a modification to our LDSF algorithm, called bLDSF, which improves upon LDSF by exploiting the idea of not granting all shared locks simultaneously.

5.1 The Benefits and Challenges

As noted earlier, when the LDSF algorithm grants an inclusive lock, it grants the lock to all transactions waiting for it. However, this may not be the optimal strategy. In general, granting a larger number of shared locks on the same object increases the probability that at least one of them will take a long time before releasing the lock. Until the last transaction completes and releases its lock, no exclusive locks can be granted on that object. In other words, the expected duration that the slowest transaction holds a shared lock grows with the number of transactions sharing the lock. This is the well-known problem of *stragglers* [21, 27, 31, 60, 74], which is exacerbated as the number of independent processes grows.

To illustrate this more formally, consider the following example. Suppose that a set of m transactions, $t_1, \cdots, t_m$, are sharing an inclusive lock. Let $R_1^{rem}, R_2^{rem}, \cdots, R_m^{rem}$ be a set of random variables representing the remaining times of these transactions. Then, the time needed before an exclusive lock can be granted on the same object is the remaining time of the the slowest transaction, denoted as $R_{\max,m}^{rem} = \max\{R_1^{rem}, \cdots, R_m^{rem}\}$, which itself is a random variable. Let $\bar{R}_{\max,m}^{rem}$ be the expectation of $R_{\max,m}^{rem}$. As long as the R_i^{rem}'s have non-zero variance[6] (i.e., $\sigma_i^2 > 0$), $\bar{R}_{\max,m}^{rem}$ strictly increases with m, as stated next.

LEMMA 4. *Suppose that $R_1^{rem}, R_2^{rem}, \cdots$ are random variables with the same range of values. If $\sigma_{k+1}^2 > 0$, then $\bar{R}_{\max,k}^{rem} < \bar{R}_{\max,k+1}^{rem}$ for $1 \leq k < m$.*

PROOF. Let $F_i(x)$ be the cumulative distribution function (CDF) of R_i, and $F_{\max,k}(x)$ be the CDF of $R_{\max,k}^{rem}$. Then,

$$F_{\max,k}(x) = \Pi_{i=1}^{k} F_i(x). \tag{6}$$

Thus,

$$\begin{aligned} \bar{R}_{\max,k+1}^{rem} &= \int_0^\infty (1 - F_{\max,k+1}(x))\mathrm{d}x \\ &= \int_0^\infty (1 - \Pi_{i=1}^{k+1} F_i(x))\mathrm{d}x \\ &\geq \int_0^\infty (1 - \Pi_{i=1}^{k} F_i(x))\mathrm{d}x \\ &= \int_0^\infty (1 - F_{\max,k}(x))\mathrm{d}x \\ &= \bar{R}_{\max,k}^{rem} \end{aligned} \tag{7}$$

as $F(x) \in [0, 1]$. However, since $\sigma_{k+1}^2 > 0$, there exists $a, b \geq 0$ such that for all $x \in (a, b)$, we have $F_{k+1}(x) \in (0, 1)$ and $F_i(x) > 0$ for $i = 1, 2, \cdots, k$. Thus, the equality in Equation (7) does not hold. Therefore, $\bar{R}_{\max,k+1}^{rem} > \bar{R}_{\max,k}^{rem}$. □

We define the delay factor as $f(m) = \frac{\bar{R}_{\max,m}^{rem}}{\bar{R}^{rem}}$. According to Lemma 4, $f(m)$ is strictly monotonically increasing with respect to m. The exact formula for $f(m)$ depends on the specific distribution of R_i's. For example, if all the R_i's are exponentially distributed (which is a memoryless distribution) with the same mean $\bar{R}$, then their CDF is given by

$$F(x) = 1 - e^{-x/\bar{R}^{rem}}, x > 0. \tag{8}$$

Then, $f(m)$ can be computed as:

$$f(m) = \sum_{i=1}^{m} \frac{1}{i}. \tag{9}$$

[6]This assumption holds unless all instances of a transaction type take exactly the same time, which is unlikely.

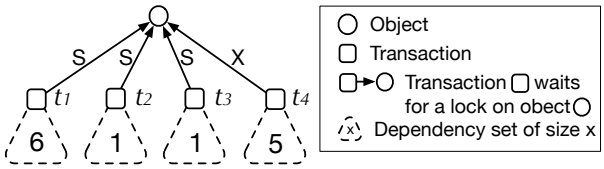

Figure 6: Assume that $f(2) = 1.5$ and $f(3) = 2$. If we first grant a shared lock to all of t_1, t_2, and t_3, all transactions in t_4's dependency set will wait for at least $2\bar{R}$. The total wait time will be $10\bar{R}$. However, if we only grant t_1's lock, then t_4's lock, and then grant t_2's and t_3's locks together, the transactions in t_4's dependency set will only wait $\bar{R}$, while those in t_2's and t_3's dependency sets will wait $2\bar{R}$. Thus, the total wait time in this case will be only $9\bar{R}$.

For other distributions, the exact form of $f(m)$ will be different. However, regardless of the distribution of the latencies, $f(m)$ is guaranteed to satisfy the following properties:

C1. $f(1) = 1$;

C2. $f(m) < f(m+1)$;

C3. $f(m) \leq m$.

The first property is trivial: granting the lock to only one transaction at a time does not incur any delays. The second property is based on Lemma 4. The third is based on the fact that shared a lock between a group of m transactions cannot be slower than granting the lock to them one after another and sequentially.

Since granting a shared lock to more transactions can delay the exclusive lock requests, it is conceivable that granting a shared lock to only a subset of the transactions waiting for it might reduce the overall latency in the system. Intuitively, when many transactions are waiting for the same shared lock, it would be better to grant the shared lock only to a few that have a higher priority (i.e., a larger dependency set), and leave the rest until the next time. This strategy can therefore reduce the time that other transactions have to wait for an exclusive lock, as illustrated in Figure 6.

However, lock scheduling in this situation becomes extremely difficult. When the time the last shared lock is released depends on (and grows with) the number of transactions, there is no constant-factor approximation strategy, if we have no prior knowledge of the expected latency of a set of transactions. We have the following negative result.

THEOREM 5. *Let $\mathbb{A}_{\neg f}$ be the set of scheduling algorithms that do not use the knowledge of the delay factor $f(k)$ in their decisions. For any algorithm $\mathcal{A}_{\neg f} \in \mathbb{A}_{\neg f}$, there exists an algorithm $\mathcal{A}$, such that $\frac{\bar{w}(\mathcal{A}_{\neg f})}{\bar{w}(\mathcal{A})} = \omega(1)$ for some delay factor $f(k)$.*

The proof of the theorem is deferred to Appendix B.

According to this theorem, any algorithm that does not rely on knowing the delay factor is not competitive: it performs arbitrarily poor, compared to the optimal scheduling. Thus, in the next section, we take the delay factor $f(k)$ as an input, and propose an algorithm that adopts the idea of granting shared locks only to a subset of the transactions requesting it. We also discuss the criteria for choosing delay factors that can yield good performance in practice.

5.2 The bLDSF Algorithm

In this section, we present a simple algorithm, called bLDSF, which inherits the intuition behind the LDSF algorithm, but also exploits the idea that a shared lock does not have to be granted to all transactions waiting for it.

While LDSF measures the progress enabled by different scheduling decisions, our bLDSF algorithm measures the *speed of progress*. If a transaction t^x waiting for an exclusive lock is scheduled, $|g(t^x)|$ transactions will make progress over the next $\bar{R}$ (expected) units of time. Thus, the speed of progress can be measured as $\frac{|g(t^x)|}{\bar{R}}$. On the other hand, by scheduling a batch of transactions $t_1^i, t_2^i, \cdots, t_k^i$ waiting for a shared lock together, $|\bigcup_{i=1}^{k} g(t_i^i)|$ transactions will make progress over the next $f(k) \cdot \bar{R}$ units of time. The speed of progress can then be measured as $\frac{|\bigcup_{i=1}^{k} g(t_i^i)|}{f(k)\bar{R}}$.

The bLDSF algorithm works as follows. First, it finds the transaction waiting for an exclusive lock with the largest dependency set, denoted as $\hat{t^x}$. Denote the size of its dependency set as $p = \left|g(\hat{t^x})\right|$. Then, bLDSF finds the batch of transactions, $\hat{t_1^i}, \hat{t_2^i}, \cdots, \hat{t_k^i}$, waiting for a shared lock such that $q = \frac{|\bigcup_{i=1}^{k} g(\hat{t_i^i})|}{f(k)}$ is maximized. If $q < p$, the system will make faster progress if $\hat{t^x}$ is scheduled first, in which case bLDSF will grant an exclusive lock to $\hat{t^x}$. Conversely, if $q > p$, the system will make faster progress if the batch of $\hat{t_1^i}, \hat{t_2^i}, \cdots, \hat{t_k^i}$ is scheduled first, in which case bLDSF will grant shared locks to $\hat{t_1^i}, \hat{t_2^i}, \cdots, \hat{t_k^i}$ simultaneously. When $q = p$, the speed of progress in the system will be the same under both scheduling decisions. In this case, bLDSF grants shared locks to the batch, in order to increase the overall degree of concurrency in the system. The pseudocode for bLDSF is provided in Algorithm 2.

We show that, when the number of transactions waiting for shared locks on the same object is bounded, the bLDSF algorithm is a constant factor approximation of the optimal scheduling algorithm in terms of the expected wait time.

THEOREM 6. *Let the maximum number of critical objects for any transaction in the system be c. Assume that the number of transactions waiting for shared locks on the same object is bounded by v. Then, given a delay factor of $f(k)$, the bLDSF algorithm is an h-approximation of the optimal scheduling algorithm in terms of the expected wait time, where $h = cv^2 \cdot f(v)$.*

The proof of the theorem is deferred to Appendix C

Unlike the LDSF algorithm, bLDSF requires a delay factor for its analysis. However, since the remaining times of transactions can be modeled as random variables, the exact form of the delay factor $f(k)$ will also depend on the distribution of these random variables. For example, the delay factor for exponential random variables is $f(k) = O(\log k)$ [22], for geometric random variables is $f(k) = O(\log k)$ [29], for Gaussian random variables is $f(k) = O(\sqrt{\log k})$ [46], and for power law random variables with exponent 3 is $f(k) = \sqrt{k}$ (See Appendix D). In Section 7.7, we empirically show that bLDSF's performance is not sensitive to the specific choice of the delay factor, as long as it is a sub-linear function that grows monotonically with k (conditions C1, C2, and C3 from Section 5.1). This is because when the batch size is small, the difference between all sub-linear functions is

Input : The dependency graph of the system $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$, transaction t, object o, label $L \in \{X, S\}$
// meaning t has just released a lock of type L on o
Output: The set of transactions whose requested lock on o should be granted

if *there are other transactions still holding a lock on* o **then**
  return $\emptyset$;
Obtain the set of transactions waiting for an inclusive lock on o, $T^i \leftarrow \{t^i \in \mathcal{V} : (t^i, o) \in \mathcal{E} \text{ and } \mathcal{L}(t^i, o) = S\} = \{t^i_1, t^i_2, \cdots, t^i_m\}$;
Obtain the set of transactions waiting for an exclusive lock on o, $T^x \leftarrow \{t^x \in \mathcal{V} : (t^x, o) \in \mathcal{E} \text{ and } \mathcal{L}(t^x, o) = X\} = \{t^x_1, t^x_2, \cdots, t^x_n\}$;
Let $t^{\hat{i}}_1, t^{\hat{i}}_2, \cdots, t^{\hat{i}}_k$ be the set of transactions in T^i such that $\frac{|\bigcup_{i=1}^{k} g(t^{\hat{i}}_i)|}{f(k)}$ is maximized ;
Let $t^{\hat{x}}$ be the transaction in T^x with the largest dependency set;
if $\left|g(t^{\hat{x}})\right| \cdot f(k) \leq \left|\bigcup_{i=1}^{k} g(t^{\hat{i}}_i)\right|$ **then**
  return $\{t^{\hat{i}}_1, t^{\hat{i}}_2, \cdots, t^{\hat{i}}_k\}$;
else
  return $t^{\hat{x}}$;

Algorithm 2: *The bLDSF Algorithm*

also small. For example, when $b = 10$, $\sqrt{b} \approx 3.16$ and $\log_2(1 + b) \approx 3.46$, leading to similar scheduling decisions. Even though $\sqrt{\log_2(1 + b)} \approx 1.86$ is smaller than the other two, it can still capture condition C2 quite well.

6. IMPLEMENTATION

We implement our scheduling algorithm in MySQL. Similar to all major DBMSs, the default lock scheduling policy in MySQL was FIFO.[7] Specifically, all pending lock requests on an object are placed in a queue. A lock request is granted immediately upon its arrival only if one of these two conditions holds: (i) there are no other locks currently held on the object, or (ii) the requested lock type is compatible with all the locks currently held on the object, and there are no incompatible requests ahead of it waiting in the queue. Whenever any transaction releases a lock on an object, MySQL's scheduler scans the entire queue from the beginning to the end: it grants any waiting requests if (i) there are no other locks held on the object, or (ii) the request is compatible with the remaining locks held on the object. However, as soon as the scheduler encounters the first lock request that cannot be granted, it stops scanning the rest of the queue.

One issue in implementing LDSF and bLDSF is that keeping track of the sizes of the dependency sets can be difficult. Exact calculation would require either (i) searching down the reverse edges in the dependency graph whenever a scheduling decision is to be made, or (ii) storing the dependency sets for all transactions and updating them each time any transaction is blocked or a lock is granted. Both of these options can cause large overheads. Specifically, option (i) can be slow: in the worse case, finding the size of the dependency set of a single transaction takes $O(|\mathcal{E}| + |\mathcal{V}|)$ time. Option (ii) incurs a large space overhead: in the worst case, storing the dependency set for each transaction requires $O(\mathcal{V})$ space. Therefore, in our implementation, we rely on an approximation of the sizes of the dependency sets, rather than computing their exact values. When a transaction t holds no locks that block other transactions, $|g(t)| = 1$. Otherwise, let T_t be the set of transactions waiting for an object currently held by transaction t. Then, $|g(t)| \approx \sum_{t' \in T_t} |g(t')| + 1$. The reason this method is only an approximation of $|g(t)|$, is that the dependency graph is a DAG (but not necessarily a tree), which means the dependency sets of different transactions may overlap. Figure 7 illustrates an example, where the dependency set of t_1 is $\{t_1, t_2, t_3, t_4\}$ and is therefore of size 4. However, its effective size is calculated as one plus the sum of the effective sizes of t_2 and t_3's dependency sets, resulting in 5.

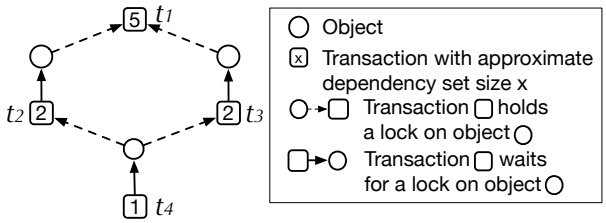

Figure 7: The effective size of t_1's dependency set is 5. But its exact size is only 4.

Another implementation issue lies in the difficulty of finding the desired batch of transactions in bLDSF. Calculating the size of the union of several dependency sets requires detailed information about the elements in each dependency set, since the dependency sets may not be disjoint (due to the presence of shared locks). Therefore, we rely on an approximation in our implementation. Specifically, we first sort all transactions waiting for a shared lock in the decreasing order of their dependency set sizes. Then, for $k = 1, 2, \cdots$, we calculate the q value (see Section 5.2) for the first k transactions. Here, we approximate the size of the union of the dependency sets as the sum of their individual sizes. Let k^* be the k value that maximizes q. We then take the first k^* transactions as our batch, which we consider for granting a shared lock to.

We show in Section 7 that, despite using these approximations in our implementation, our algorithms remain quite effective in practice.

Starvation Avoidance — In MySQL's implementation of FIFO, when there is an exclusive lock request in the queue, it serves as a conceptual barrier: later requests for shared locks cannot be granted, even if they are compatible with the currently held locks on the object. This mechanism prevents starvation when using FIFO. In our algorithms, we prevent starvation of transactions using a similar mechanism: we also place a barrier at the end of the current wait queue. Lock requests that arrive later are all placed behind this barrier, and are not considered for scheduling. In other words, the only requests that are considered are those that are ahead of the barrier. Once all such requests are granted, this barrier is lifted, and a new barrier is added to the end of the current queue, so that those requests previously behind the barrier are now ahead of it. This mechanism prevents a transaction with a small dependency set from waiting indefinitely behind an infinite stream of newly arrived transactions with larger dependency sets.

Space Complexity — Given the approximation methods mentioned above, both LDSF and bLDSF only require maintaining the approximate size of dependency set of each trans-

[7]Now, our LDSF algorithm is the default (MySQL 8.0.3+).

action. Therefore, the overall space overhead of our algorithms is only $O(|T|)$.

Time Complexity — In MySQL, all lock requests on an object (either granted or not) are stored in a linked list. Whenever a transaction releases a lock on the object, the scheduler scans this list for requests that are not granted yet. For each of these requests the scheduler scans the list again for requests that have been granted and checks for compatibility. If the request is found compatible with all existing locks, the request is granted, and the scheduler moves on to check the compatibility of the next request. Otherwise, the request is not granted, and the scheduler stops granting any further requests. Let N is the number of lock requests on an object (either granted or not). Then, FIFO takes $O(N^2)$ time in the worst case. In both LDSF and bLDSF, we use the same procedure as FIFO to find compatible requests that are not granted yet, which takes $O(N^2)$ time. For bLDSF, we also sort all transactions waiting for a shared lock by the size of their dependency sets, which takes $O(N \log N)$ time. Thus, time complexity of both LDSF and bLDSF is still $O(N^2)$.

7. EXPERIMENTS

Our experiments aim to answer several key questions:

- How do our scheduling algorithms (LDSF and bLDSF) affect the overall throughput of the system?
- How do our algorithms compare against FIFO—the default policy in almost every major database—in terms of reducing transaction latencies?
- How do our algorithms compare against various heuristics?
- How much overhead do our algorithms incur? Are they significant or negligible compared to the latency of a transaction?
- How does the effectiveness of our algorithms vary with different levels of contention?
- What is the impact of the choice of delay factor on the effectiveness of bLDSF?

In summary, our experiments show the following:

1. By resolving contention much more effectively than FIFO, bLDSF improves throughput by up to 6.5x (by 4.5x on average). (Section 7.2)
2. Compared to FIFO, bLDSF can reduce transaction latencies by up to 300x (30x on average). (Section 7.3)
3. Both bLDSF and LDSF outperform various heuristics by 2.5x in terms of throughput, and by up to 100x (8x on avg.) in terms of transaction latency. (Section 7.4)
4. Our algorithms reduce queue length by reducing contention, and thus incur much less overhead than FIFO. (Section 7.5)
5. As the degree of contention increases in the system, bLDSF's improvement over FIFO widens. (Section 7.6)
6. bLDSF is not sensitive to the specific choice of delay factor, as long as it is chosen to be an increasing and sub-linear function. (Section 7.7)

7.1 Experimental Setup

Hardware & Software — All experiments were performed using a 5 GB buffer pool on a Linux server with 16 Intel(R) Xeon(R) CPU E5-2450 processors and 2.10GHz cores. The clients were run on a separate machine, submitting transactions to MySQL 5.7 running on the server.

Methodology — We used the OLTP-Bench tool [24] to run the TPC-C workload. We also modified this tool to run a microbenchmark (explained below). OLTP-Bench generated transactions at a specified rate, and client threads issued these transactions to MySQL. The latency of each transaction was calculated as the time from when it was issued until it finished. In all experiments, we controlled the number of transactions issued per second within a safe range to prevent MySQL from falling into a thrashing regime. We also noticed that the number of deadlocks was negligible compared to the total number of transactions, across all experiments and algorithms.

TPC-C workload — We used a 32-warehouse configuration for the TPC-C benchmark. To simulate a system with different levels of contention, we relied on changing the following two parameters: (i) number of clients, and (ii) number of submitted transactions per second (a.k.a. throughput). Each of our client threads issued a new transaction as soon as its previous transaction finished. Thus, by creating a specified number of client threads, we effectively controlled the number of in-flight transactions. To control the system throughput, we created client threads that issued transactions at a specific rate.

Microbenchmark — We create a microbenchmark for a more thorough evaluation of our algorithm under different degrees of contention. Specifically, we created a database with only one table that had 20,000 records in it. The clients would send transactions to the server, each comprised of 5 queries. Each query was randomly chosen to be either a "SELECT" query (acquiring a shared lock) or an "UPDATE" query (acquiring an exclusive lock). The records in the table were accessed by the queries according to a Zipfian distribution. To generate different levels of contention, we varied the following two parameters in our microbenchmark:

1. the skew of the access pattern (the parameter θ of the Zipfian distribution)
2. the fraction of exclusive locks (the probability of "UPDATE" queries).

Baselines — We compared the performance of our bLDSF algorithm (with $f(k)$=$\log_2(1 + k)$ as default) against the following baselines:

1. **First In First Out (FIFO).** FIFO is the default scheduler in MySQL and nearly all other DBMSs. When an object becomes available, FIFO grants the lock to the transaction that has waited the longest.
2. **Variance-Aware Transaction Scheduling (VATS).** This is the strategy proposed by Huang et al. [44]. When an object becomes available, VATS grants the lock to the eldest transaction in the queue.
3. **Largest Dependency Set First (LDSF).** This is the strategy described in Algorithm 1, which is equivalent to bLDSF with $b = \inf$, and $f(k) = 1$.
4. **Most Locks First (MLF).** When an object becomes available, grant a lock on it to the transaction that holds the most locks. This heuristic is introduced in Section 4.1.

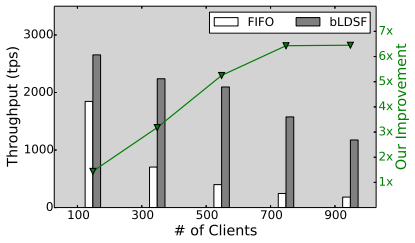

Figure 8: Throughput improvement under bLDSF (TPC-C).

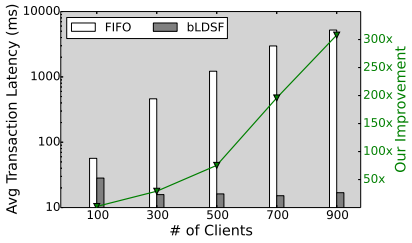

Figure 9: Transaction latency improvement under bLDSF, when running the same number of transactions per second (TPC-C).

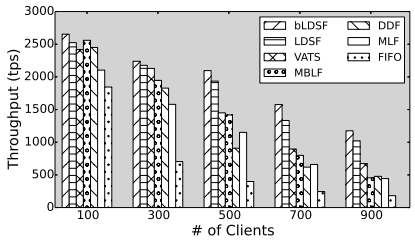

Figure 10: Maximum throughput under various algorithms (TPC-C).

5. **Most Blocking Locks First (MBLF).** When an object becomes available, grant a lock on the object to the transaction that holds the most locks which block at least one other transaction. This heuristic is introduced in Section 4.1.
6. **Deepest Dependency First (DDF).** When an object becomes available, grant a lock on the object to the transaction that has the deepest dependency subgraph. This heuristic is introduced in Section 4.1.

7.2 Throughput

We compared the system throughput when using FIFO versus bLDSF, given an equal number of clients (i.e., in-flight transactions). We varied the number of clients from 100 to 900. The results of this experiment for TPC-C are presented in Figure 8.

In both cases, the throughput dropped as the number of clients increased. This is expected, as more transactions in the system lead to more objects being locked. Thus, when a transaction requests a lock, it is more likely to be blocked. In other words, the number of transactions that can make progress decreases, which leads to a decrease in throughput.

However, the throughput decreased more rapidly when using FIFO than bLDSF. For example, when there were only 100 clients, bLDSF outperformed FIFO by only 1.4x. However, with 900 clients, bLDSF achieved 6.5x higher throughput than FIFO. As discussed in Section 5.2, bLDSF always schedules transactions that maximize the speed of progress in the system. This is why it allows for more transactions to be processed in a certain amount of time.

7.3 Transaction Latency

We also compared transaction latencies between FIFO and bLDSF with an equal number of transactions per second (i.e, throughput). We varied the number of clients (and hence, the number of in-flight transactions) from 100 to 900 for FIFO, and then ran bLDSF at the same throughput as FIFO. The result is shown in Figure 9. Our bLDSF algorithm dramatically outperformed FIFO by a factor of up to 300x. This outstanding improvement confirms our Theorems 3 and 6, as our algorithm is designed to minimize average transaction latencies.

7.4 Comparison with Other Heuristics

In this section, we report our comparison of both bLDSF and LDSF algorithms against the heuristic methods introduced in Section 4, i.e., MLF, MBLF, and DDF. Moreover, we compare our algorithms with VATS too.

First, we compared their throughput given an equal number of clients. We varied the number of clients from 100 to 900. The results are shown in Figure 10. LDSF and bLDSF achieve up to 2x and 2.5x improvement over the other heuristics in terms of throughput, respectively.

We also measured transaction latencies under an equal number of transactions per second (i.e, throughput). We varied the number of clients from 100 to 900 for the heuristics, and then ran bLDSF and LDSF at the maximum throughput achieved by any of the heuristics. For those heuristics which were not able to achieve this throughput, we compared our algorithms at a higher throughput than they achieved. The results are shown in Figure 11, indicating that MLF, MBLF, and DDF outperformed FIFO by about 2.5x in terms of average latency, while our algorithms achieved up to 100x improvement over the best heuristics (MBLF with 900 transactions). Furthermore, bLDSF was better than LDSF by a small margin.

7.5 Scheduling Overhead

We also compared the overhead of our algorithms (LDSF and bLDSF) against FIFO: the overhead of a scheduling algorithm is the time needed by the algorithm to *decide* which lock(s) to grant.

In this experiment, we fixed the number of clients to 100 while varying throughput from 200 to 1000. The result is shown in Figure 12. We can see that, although all three algorithms have the same time complexity in terms of the queue length (Section 6), ours resulted in much less overhead than FIFO because they led to much shorter queues for the same throughput. This is because our algorithms effectively resolve contention, and thus, reduce the number of waiting transactions in the queue. To illustrate this, we also measured the average number of waiting transactions whenever an object becomes available. As shown in Figure 13, this number was much smaller for LDSF and bLDSF.

7.6 Studying Different Levels of Contention

In this section, we study the impact of different levels of contention on the effectiveness of our bLDSF algorithm. Contention in a workload is a result of two factors: (i) skew in the data access pattern (e.g., popular tuples), and (ii) a large number of exclusive locks. There is more contention when the pattern is more skewed, as transactions will request a lock on the same records more often. Likewise, exclusive lock requests cause more contention, as they cannot be granted together and result in blocking more transactions. We studied the effectiveness of our algorithm under different degrees of contention, by varying these two factors using our microbenchmark:

1. We fixed the fraction of exclusive locks to be 60% of all lock requests, and varied the θ parameter of the Zipfian

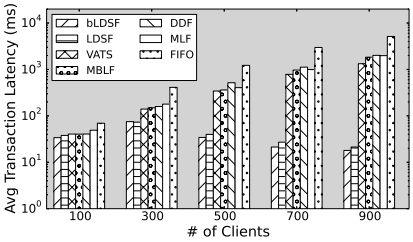

Figure 11: Transaction latency under various algorithms (TPC-C).

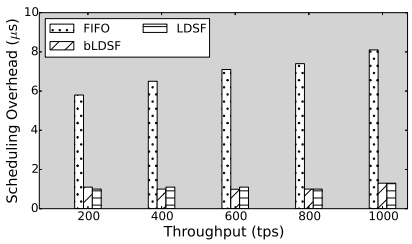

Figure 12: Scheduling overhead of our algorithms versus FIFO (TPC-C).

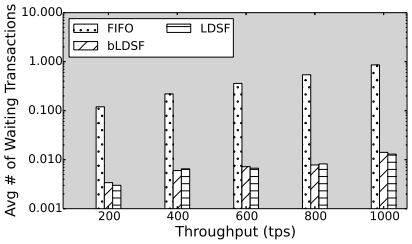

Figure 13: Average number of waiting transactions in the queue under our algorithms versus FIFO (TPC-C).

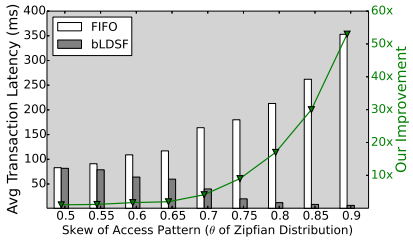

Figure 14: Average transaction latency for different degrees of skewness (microbenchmark).

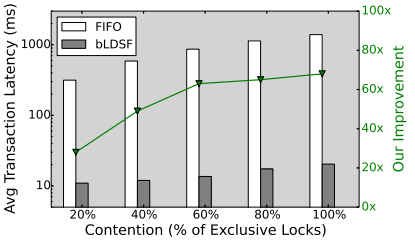

Figure 15: Average latency for different number of exclusive locks (microbenchmark).

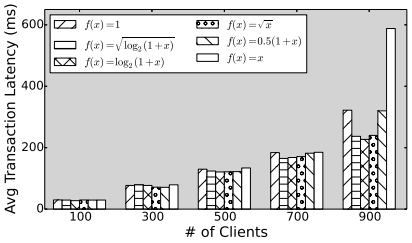

Figure 16: The impact of delay factor on average latency.

distribution of our access distribution between 0.5 and 0.9. A Zipfian distribution becomes more skewed as θ increases.

2. We fixed the θ parameter to be 0.8, and varied the probability of an "UPDATE" query in our microbenchmark between 20% and 100%. The larger this probability, the larger the fraction of exclusive locks.

First, we ran FIFO using 300 clients, and then ran bLDSF at the same throughput as FIFO. The results of these experiments are shown in Figures 14 and 15.

Figure 14 shows that when there is no skew, there is no contention, and thus most queues are either empty or only have a single transaction waiting. Since there is no scheduling decision to be made in this situation, FIFO and bLDSF become equivalent and exhibit a similar performance. However, the gap between the two algorithms widens, as skew (and thereby contention) increases. For example, when the data access is highly skewed ($\theta = 0.9$), bLDSF outperforms FIFO by more than 50x. Figure 15 reveals a similar trend: as more exclusive locks are requested, bLDSF achieves greater improvement. Specifically, when 20% of the lock requests are exclusive, bLDSF outperforms FIFO by 20x. However, when all the locks are exclusive, the improvement is even more dramatic, i.e., 70x. In summary, when there is no contention in the system, there are no scheduling decisions to be made, and all scheduling algorithms are equivalent. However, as contention rises, so does the need for better scheduling decisions, and so does the gap between bLDSF and other algorithms.

7.7 Choice of Delay Factor

To better understand the impact of delay factors on bLDSF, we experimented with several functions of different growth rates, ranging from the lower bound of all functions that satisfy conditions C1, C2, and C3 (i.e., $f(k) = 1$) to their upper bound (i.e., $f(k) = k$). Specifically, we used each of the following delay factors in our bLDSF algorithm, and measured the average transaction latency:

- $f_1(k) = 1$;
- $f_2(k) = \sqrt{\log_2(1+k)}$;
- $f_3(k) = \log_2(1+k)$;
- $f_4(k) = \sqrt{k}$;
- $f_5(k) = 0.5(1+k)$;
- $f_6(k) = k$.

The results are shown in Figure 16. We can see that all sublinear functions (i.e., f_2, f_3, and f_4) performed comparably, and that they performed better than the other functions. Understandably, f_1 did not perform well, as it did not satisfy condition C2 from Section 5.1. Functions f_5 and f_6 did not perform well either, since linear functions over-estimate the delay. For example, two transactions running concurrently take less time than if they ran one after another.

8. CONCLUSION

We study a fundamental (yet, surprisingly overlooked) problem: *lock scheduling* in a database system. Despite the massive body of work on transactional databases, the astonishing impact of lock scheduling on overall performance of a transactional system seems to have been largely underrecognized—to the extent that every DBMS to date has simply relied on FIFO. To our knowledge, we are the first to propose the idea of contention-aware lock scheduling, and present efficient algorithms that are guaranteed to reduce mean transaction latencies down to a constant-factor-approximation of the optimal scheduling. We also empirically confirm our theoretical analysis by modifying a real-world DBMS. Our extensive experiments show that our algorithms reduce transaction latencies by up to two orders of magnitude, while delivering 6.5x higher throughput. More importantly, our algorithm has already been adopted by MySQL, and has started to show impact on real world applications.

9. REFERENCES

[1] Db2 documentation. https://www.ibm.com/support/knowledgecenter/en/SSEPEK_12.0.0/perf/src/tpc/db2z_lockcontention.html.

[2] LDSF pull request. https://github.com/mysql/mysql-server/pull/115.

[3] Mysql source code. https://github.com/mysql/mysql-server/blob/5.7/storage/innobase/lock/lock0lock.cc.

[4] Overview of teradata database locking. http://info.teradata.com/HTMLPubs/DB_TTU_16_00/index.html#page/General_Reference%2FB035-1091-160K%2Fmtg1472241438567.html.

[5] Postgres documentation. https://github.com/postgres/postgres/blob/master/src/backend/storage/lmgr/README.

[6] Sql server, lock manager, and relaxed fifo. https://blogs.msdn.microsoft.com/psssql/2009/06/02/sql-server-lock-manager-and-relaxed-fifo/.

[7] VATS pull request. https://github.com/mysql/mysql-server/pull/106.

[8] R. Abbott and H. Garcia-Molina. Scheduling real-time transactions. *ACM SIGMOD Record*, 1988.

[9] R. K. Abbott and H. Garcia-Molina. Scheduling real-time transactions: A performance evaluation. *TODS*, 1992.

[10] B. Adelberg, B. Kao, and H. Garcia-Molina. Database support for efficiently maintaining derived data. In *International Conference on Extending Database Technology*, 1996.

[11] R. Agrawal, M. J. Carey, and L. W. McVoy. The performance of alternative strategies for dealing with deadlocks in database management systems. *IEEE Transactions on Software Engineering*, 1987.

[12] S. Altmeyer, S. M. Sundharam, and N. Navet. The case for fifo real-time scheduling. Technical report, 2016.

[13] R. F. Aranha, V. Ganti, S. Narayanan, C. Muthukrishnan, S. Prasad, and K. Ramamritham. Implementation of a real-time database system. *Information Systems*, 1996.

[14] C. Bector, Y. P. Gupta, and M. C. Gupta. V-shape property of optimal sequence of jobs about a common due date on a single machine. *Computers & operations research*, 1989.

[15] P. A. Bernstein and N. Goodman. Concurrency control in distributed database systems. *ACM Computing Surveys*, 1981.

[16] D. P. Bovet and M. Cesati. *Understanding the Linux kernel*. 2005.

[17] X. Cai. V-shape property for job sequences that minimize the expected completion time variance. *European Journal of Operational Research*, 1996.

[18] M. J. Carey and M. Stonebraker. The performance of concurrency control algorithms for database management systems. In *VLDB*, 1984.

[19] S. Chakravarthy, D.-K. Hong, and T. Johnson. Real-time transaction scheduling: A framework for synthesizing static and dynamic factors. *Real-Time Systems*, 1998.

[20] B.-C. Choi, S.-H. Yoon, and S.-J. Chung. Minimizing maximum completion time in a proportionate flow shop with one machine of different speed. *European Journal of Operational Research*, 2007.

[21] J. Cipar, Q. Ho, J. K. Kim, S. Lee, G. R. Ganger, G. Gibson, K. Keeton, and E. P. Xing. Solving the straggler problem with bounded staleness. In *HotOS*, 2013.

[22] A. DasGupta. Finite sample theory of order statistics and extremes. In *Probability for Statistics and Machine Learning*. 2011.

[23] L. Daynès, O. Gruber, and P. Valduriez. Locking in oodbms client supporting nested transactions. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, 1995.

[24] D. E. Difallah, A. Pavlo, C. Curino, and P. Cudre-Mauroux. Oltp-bench: An extensible testbed for benchmarking relational databases. *PVLDB*, 2013.

[25] J. Du and J. Y.-T. Leung. Scheduling tree-structured tasks with restricted execution times. *Information processing letters*, 1988.

[26] J. Du and J. Y.-T. Leung. Scheduling tree-structured tasks on two processors to minimize schedule length. *SIAM journal on discrete mathematics*, 1989.

[27] A. C. Dusseau, R. H. Arpaci, and D. E. Culler. Effective distributed scheduling of parallel workloads. *ACM SIGMETRICS Performance Evaluation Review*, 1996.

[28] S. Eilon and I. Chowdhury. Minimising waiting time variance in the single machine problem. *Management Science*, 1977.

[29] B. Eisenberg. On the expectation of the maximum of iid geometric random variables. *Statistics & Probability Letters*, 2008.

[30] A. Feldmann, M.-Y. Kao, J. Sgall, and S.-H. Teng. Optimal online scheduling of parallel jobs with dependencies. In *STOC*, 1993.

[31] K. B. Ferreira, P. G. Bridges, R. Brightwell, and K. T. Pedretti. The impact of system design parameters on application noise sensitivity. *Cluster computing*, 2013.

[32] P. Franaszek and J. T. Robinson. Limitations of concurrency in transaction processing. *TODS*, 1985.

[33] L. George and P. Minet. A fifo worst case analysis for a hard real-time distributed problem with consistency constraints. In *ICDCS*, 1997.

[34] A. Guinet and M. Solomon. Scheduling hybrid flowshops to minimize maximum tardiness or maximum completion time. *International Journal of Production Research*, 1996.

[35] L. A. Hall, D. B. Shmoys, and J. Wein. Scheduling to minimize average completion time: Off-line and on-line algorithms. In *SODA*, 1996.

[36] J. R. Haritsa, M. J. Canrey, and M. Livny. Value-based scheduling in real-time database systems. *VLDBJ*, 1993.

[37] J. R. Haritsa, M. J. Carey, and M. Livny. Data access scheduling in firm real-time database systems. *Real-Time Systems*, 1992.

[38] J. R. Haritsa, M. Livny, and M. J. Carey. Earliest deadline scheduling for real-time database systems. In *RTSS*, 1991.

[39] C. He, J. Y.-T. Leung, K. Lee, and M. L. Pinedo.

Improved algorithms for single machine scheduling with release dates and rejections. *4OR*, 2016.

[40] D. Hong, T. Johnson, and S. Chakravarthy. *Real-time transaction scheduling: a cost conscious approach.* 1993.

[41] J. A. Hoogeveen and A. P. Vestjens. Optimal on-line algorithms for single-machine scheduling. In *International Conference on Integer Programming and Combinatorial Optimization*, 1996.

[42] W. Horn. Single-machine job sequencing with treelike precedence ordering and linear delay penalties. *SIAM Journal on Applied Mathematics*, 1972.

[43] J. Huang. *Real-time transaction processing: design, implementation, and performance evaluation.* PhD thesis, University of Massachusetts, 1991.

[44] J. Huang, B. Mozafari, G. Schoenebeck, and T. Wenisch. A top-down approach to achieving performance predictability in database systems. In *SIGMOD*, 2017.

[45] T. Ibaraki, T. Kameda, and N. Katoh. Cautious transaction schedulers for database concurrency control. *IEEE transactions on software engineering*, 1988.

[46] G. C. Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www.gautamkamath.com/writings/gaussian_max.pdf*, 2015.

[47] B. Kao and H. Garcia-Molina. An overview of real-time database systems. In *Real Time Computing*. 1994.

[48] R. Kohli, R. Krishnamurti, and P. Mirchandani. The minimum satisfiability problem. *SIAM Journal on Discrete Mathematics*, 1994.

[49] A. M. Krieger and M. Raghavachari. V-shape property for optimal schedules with monotone penalty functions. *Computers & operations research*, 1992.

[50] J. Lee and S. H. Son. Performance of concurrency control algorithms for real-time database systems., 1996.

[51] J. P. Lehoczky. Scheduling communication networks carrying real-time traffic. In *RTSS*, 1998.

[52] J. K. Lenstra, A. R. Kan, and P. Brucker. Complexity of machine scheduling problems. *Annals of discrete mathematics*, 1977.

[53] H. Leontyev and J. H. Anderson. Tardiness bounds for fifo scheduling on multiprocessors. In *Euromicro Conference on Real-Time Systems*, 2007.

[54] Q. Lin, P. Chang, G. Chen, B. C. Ooi, K.-L. Tan, and Z. Wang. Towards a non-2pc transaction management in distributed database systems. In *Proceedings of the 2016 International Conference on Management of Data*, 2016.

[55] C. L. Liu and J. W. Layland. Scheduling algorithms for multiprogramming in a hard-real-time environment. *JACM*, 1973.

[56] P. P. Macri. Deadlock detection and resolution in a codasyl based data management system. In *SIGMOD*, 1976.

[57] D. P. Mitchell and M. J. Merritt. A distributed algorithm for deadlock detection and resolution. In *Proceedings of the third annual ACM symposium on Principles of distributed computing*, 1984.

[58] R. R. Muntz and E. G. Coffman Jr. Preemptive scheduling of real-time tasks on multiprocessor systems. *JACM*, 1970.

[59] J. Pei, X. Liu, P. M. Pardalos, W. Fan, and S. Yang. Scheduling deteriorating jobs on a single serial-batching machine with multiple job types and sequence-dependent setup times. *Annals of Operations Research*, 2017.

[60] F. Petrini, D. J. Kerbyson, and S. Pakin. The case of the missing supercomputer performance: Achieving optimal performance on the 8,192 processors of asci q. In *Supercomputing, 2003 ACM/IEEE Conference*, 2003.

[61] C. Phillips, C. Stein, and J. Wein. Scheduling jobs that arrive over time. In *Workshop on Algorithms and Data Structures*, 1995.

[62] K. Ramamritham. Real-time databases. *Distributed and parallel databases*, 1993.

[63] R. Rastogi, S. Seshadri, P. Bohannon, D. Leinbaugh, A. Silberschatz, and S. Sudarshan. Improving predictability of transaction execution times in real-time databases. *Real-Time Systems*, 2000.

[64] A. J. Ruiz-Torres and G. Centeno. Scheduling with flexible resources in parallel workcenters to minimize maximum completion time. *Computers & operations research*, 2007.

[65] L. Sha, R. Rajkumar, and J. P. Lehooczky. Concurrency control for distributed real-time databases. *SIGMOD Rec.*, 1988.

[66] J. B. Sidney. Decomposition algorithms for single-machine sequencing with precedence relations and deferral costs. *Operations Research*, 1975.

[67] W. E. Smith. Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 1956.

[68] J. A. Stankovic, M. Spuri, K. Ramamritham, and G. C. Buttazzo. *Deadline scheduling for real-time systems: EDF and related algorithms.* 2012.

[69] D. Towsley and S. Panwar. On the optimality of minimum laxity and earliest deadline scheduling for real-time multiprocessors. In *Workshop on Real Time*, 1990.

[70] S.-M. Tseng, Y.-H. Chin, and W.-P. Yang. Scheduling real-time transactions with dynamic values: a performance evaluation. In *Workshop on Real-Time Computing Systems and Applications*, 1995.

[71] Ö. Ulusoy and G. G. Belford. Real-time transaction scheduling in database systems. *Information Systems*, 1993.

[72] V. Vani and M. Raghavachari. Deterministic and random single machine sequencing with variance minimization. *Operations Research*, 1987.

[73] M. Xiong, Q. Wang, and K. Ramamritham. On earliest deadline first scheduling for temporal consistency maintenance. *Real-Time Systems*, 2008.

[74] R. Zajcew, P. Roy, D. L. Black, C. Peak, P. Guedes, B. Kemp, J. LoVerso, M. Leibensperger, M. Barnett, F. Rabii, et al. An osf/1 unix for massively parallel multicomputers. In *USENIX Winter*, 1993.

APPENDIX

A. PROOF OF THEOREM 1

PROOF. In order to show that this problem is NP-hard, we reduce the problem from MAX-2-DNF. In this problem, we are given a disjunctive normal formula φ where each clause has at most 2 literals, and the goal is to output an assignment which satisfies the maximum number of clauses. (So each clause will be the AND of two literals). Max-DNF, which is the same problem as Min-SAT, is known to be NP-complete [48] even when the number of literals in each clause is bounded by 2, which is called Max-2-DNF.

Given a disjunctive normal formula φ with variable set V and clause set C, we create a dependency graph $\mathcal{G}$ as follows:

- Transactions:
  - For each variable $v \in V$, create transactions t_v, and $t_{\neg v}$
  - For each clause $c \in C$, create one transaction t_c.
- Objects:
  - For each variable $v \in V$, create an object o_v: t_v, and $t_{\neg v}$ are waiting for an exclusive lock on o_v.
  - For each clause $c \in C$, create an object o_c: for each literal ℓ in c, t_ℓ currently holds an inclusive lock on o_c, and t_c is waiting for an exclusive lock on o_c.

Suppose that each transaction takes 1 unit time to finish. We now prove that an assignment for the DNF problem that satisfies at least m of the clauses is equivalent to an algorithm under which the total latency is $3|V| + 3|C| - m$.

Given an assignment that satisfies m clauses, we first grant the lock to t_x such that x is true in the assignment at time 0, for $x \in V$ or $\neg x \in V$. There are $|V|$ such transactions and they will finish at time 1, so the latency for these transactions is $|V|$. At time 1, we grant the lock to t_x such that x is false in the assignment, for $x \in V$ or $\neg x \in V$. There are $|V|$ such transactions and they will finish at time 2, so the latency for these transactions is $2|V|$. Moreover, for the clauses c such that all its literals are true, we grant the lock to t_c. There are m such transactions and they will finish at time 2 as well, so the latency for them is $2m$. At time 2, we grant the lock to all other transactions. There are $|C| - m$ such transactions and they will finish at time 3, so the latency for them is $3(|C| - m)$. Therefore, the total latency is $3|V| + 3|C| - m$.

On the other hand, if we have a scheduling algorithm whose total latency is $3|V| + 3|C| - m$, the only decisions made when constructing the schedule are whether to grant the lock to t_v or $t_{\neg v}$ for each $v \in V$. Thus, we can construct an assignment from the schedule. Because the schedule total latency $3|V| + 3|C| - m$, we know that m of the t_c complete at time 2. But this only happens if the assignment satisfies the clause c. Therefore, the assignment satisfies m clauses in φ. □

B. PROOF OF THEOREM 5

PROOF. Given a scheduling algorithm $\mathcal{A}_{\neg f}$, consider the following situation. Suppose that n transactions are waiting for an exclusive lock on object o, each holds another lock that blocks $n^{1.6} - 1$ other transactions. Meanwhile, $m = n^3$ transactions are waiting for a shared lock on o, none of which blocks another transaction. Let $t_{m/2}$ be the transaction waiting for a shared lock such that half of the transactions waiting for the shared lock is scheduled before or together with it, while the other half is scheduled after or together with it by $\mathcal{A}_{\neg f}$. Let $t_{n/2}$ be the $\frac{n}{2}$-th transaction waiting for an exclusive lock scheduled by $\mathcal{A}_{\neg f}$. Consider the following two scenarios.

Scenario 1. Suppose that $t_{m/2}$ is scheduled after $t_{n/2}$. Then,

$$\begin{aligned} d_{\mathcal{A}_{\neg f}}(o) &\geq \frac{n^3}{2} \cdot \frac{n}{2} \bar{R} \\ &= \frac{n^4}{4} \bar{R}. \end{aligned}$$

Let algorithm $\mathcal{A}_1$ be an algorithm that first schedules all the transactions waiting for the shared lock together. Then,

$$d_{\mathcal{A}_1}(o) \leq n^{2.6} \cdot (n + f(n^3))\bar{R}.$$

Suppose that $f(k) = 1$ for all k. Then,

$$\begin{aligned} \frac{d_{\mathcal{A}_{\neg f}}(o)}{d_{\mathcal{A}_1}(o)} &= \frac{n^4}{4(n^{3.6} + n^{2.6})} \\ &= \omega(1). \end{aligned}$$

Therefore, $\frac{\bar{w}(\mathcal{A}_{\neg f})}{\bar{w}(\mathcal{A}_1)} = \omega(1)$.

Scenario 2. Suppose that $t_{m/2}$ is scheduled before $t_{n/2}$. Then,

$$d_{\mathcal{A}_{\neg f}}(o) \geq \frac{n^{2.6}}{2} \cdot f\left(\frac{n^3}{2}\right) \bar{R}.$$

Let $\mathcal{A}_2$ be an algorithm that schedules all the transactions waiting for the shared lock together after all transactions waiting for an exclusive lock. Then,

$$d_{\mathcal{A}_2}(o) \leq (n^3 + n^{2.6}) \cdot n\bar{R}.$$

Suppose that $f(k) = \sqrt{k}$. Then,

$$\begin{aligned} \frac{d_{\mathcal{A}_{\neg f}}(o)}{d_{\mathcal{A}_2}(o)} &= \frac{n^{4.1}}{2\sqrt{2}(n^4 + n^{3.6})} \\ &= \omega(1). \end{aligned}$$

Therefore, $\frac{\bar{w}(\mathcal{A}_{\neg f})}{\bar{w}(\mathcal{A}_2)} = \omega(1)$. □

C. PROOF OF THEOREM 6

We define $\hat{w}(\mathcal{A})$ for scheduling algorithm $\mathcal{A}$ as in the proof of Theorem 3. Also, we consider a batch of transactions to be scheduled together as a single transaction. We use the same notation t for the batch as for a transaction for simplicity. Let $a_{\mathcal{A}}(t)$ be the expected time transaction is granted the lock by scheduling algorithm $\mathcal{A}$, and T_o be the set of transactions waiting for a lock on object o. Then,

$$d_{\mathcal{A}}(o) = \sum_{t \in T_o} a_{\mathcal{A}}(t) + (|g(t)| - 1)(a_{\mathcal{A}}(t) + \bar{R}). \tag{10}$$

Let $\tilde{\mathcal{A}}$ be the algorithm performs the same way as bLDSF, except that it schedules all the transactions waiting for a shared lock together (like LDSF).

To prove Theorem 6, we first prove the following lemma.

LEMMA 7. *$\tilde{\mathcal{A}}$ is minimizes $\hat{w}(\mathcal{A})$ among all algorithms $\mathcal{A}$ that schedule all the transactions waiting for a shared lock together.*

PROOF. Let $\mathcal{A}_1$ be a scheduling algorithm that schedules all the transactions waiting for a shared lock together. Suppose t_1 is scheduled before t_2 by $\mathcal{A}_1$, i.e., $a_{\mathcal{A}_1}(t_1) < a_{\mathcal{A}_1}(t_2)$. Consider the following scenarios.

Scenario 1. Suppose that t_1 and t_2 are both transactions waiting for an exclusive lock on object o and $|g(t_1)| \leq |g(t_2)|$. Consider algorithm $\mathcal{A}_2$ such that $\mathcal{A}_2$ schedules t_1 when $\mathcal{A}_1$ schedules t_2, and vice versa; $\mathcal{A}_2$ schedules all the other transactions in the same order as $\mathcal{A}_1$. Then,

$$\begin{aligned}
&d_{\mathcal{A}_1}(o) - d_{\mathcal{A}_2}(o) \\
=&|g(t_1)|(a_{\mathcal{A}_1}(t_1) - a_{\mathcal{A}_2}(t_1)) + |g(t_2)|(a_{\mathcal{A}_1}(t_2) - a_{\mathcal{A}_2}(t_2)) \\
=&|g(t_1)|(a_{\mathcal{A}_1}(t_1) - a_{\mathcal{A}_1}(t_2)) + |g(t_2)|(a_{\mathcal{A}_1}(t_2) - a_{\mathcal{A}_1}(t_1)) \\
=&(|g(t_1)| - |g(t_2)|)(a_{\mathcal{A}_1}(t_1) - a_{\mathcal{A}_1}(t_2)) \geq 0
\end{aligned}$$

Therefore, $\hat{w}(\mathcal{A}_1) \geq \hat{w}(\mathcal{A}_2)$.

Scenario 2. Suppose that t_1 is the batch of k transactions waiting for a shared lock on o, t_2 is a transaction waiting for an exclusive lock on o, and $|g(t_1)| \leq |g(t_2)|f(k)$. Moreover, suppose that $a_{\mathcal{A}_1}(t_2) = a_{\mathcal{A}_1}(t_1) + f(k)\bar{R}$, that is, t_2 is scheduled right after t_1. Consider algorithm $\mathcal{A}_3$ such that $\mathcal{A}_3$ schedules t_1 right after t_2 and schedules every other transaction the same way as $\mathcal{A}_1$. Then, $a_{\mathcal{A}_3}(t_1) = a_{\mathcal{A}_3}(t_2) + \bar{R}$, $a_{\mathcal{A}_3}(t_2) = a_{\mathcal{A}_1}(t_1)$. Thus,

$$\begin{aligned}
&d_{\mathcal{A}_1}(o) - d_{\mathcal{A}_3}(o) \\
=&|g(t_1)|(a_{\mathcal{A}_1}(t_1) - a_{\mathcal{A}_3}(t_1)) + |g(t_2)|(a_{\mathcal{A}_1}(t_2) - a_{\mathcal{A}_3}(t_2)) \\
=&|g(t_1)|(a_{\mathcal{A}_3}(t_2) - a_{\mathcal{A}_3}(t_1)) + |g(t_2)|(a_{\mathcal{A}_1}(t_2) - a_{\mathcal{A}_1}(t_1)) \\
=&(|g(t_2)|f(k) - |g(t_1)|)\bar{R} \geq 0
\end{aligned}$$

Therefore, $\hat{w}(\mathcal{A}_1) \geq \hat{w}(\mathcal{A}_3)$.

Scenario 3. Suppose that t_1 is a transaction waiting for an exclusive lock on o, t_2 is the batch of k transactions waiting for a shared lock on o, and $|g(t_1)|f(k) \leq |g(t_2)|$. Moreover, suppose that $a_{\mathcal{A}_1}(t_2) = a_{\mathcal{A}_1}(t_1) + \bar{R}$, that is, t_2 is scheduled right after t_1. Consider algorithm $\mathcal{A}_4$ such that $\mathcal{A}_4$ schedules t_1 right after t_2 and schedules every other transaction the same way as $\mathcal{A}_1$. Then, $a_{\mathcal{A}_4}(t_1) = a_{\mathcal{A}_3}(t_2) + f(k)\bar{R}$, $a_{\mathcal{A}_4}(t_2) = a_{\mathcal{A}_1}(t_1)$. Thus,

$$\begin{aligned}
&d_{\mathcal{A}_1}(o) - d_{\mathcal{A}_4}(o) \\
=&|g(t_1)|(a_{\mathcal{A}_1}(t_1) - a_{\mathcal{A}_3}(t_1)) + |g(t_2)|(a_{\mathcal{A}_1}(t_2) - a_{\mathcal{A}_3}(t_2)) \\
=&|g(t_1)|(a_{\mathcal{A}_3}(t_2) - a_{\mathcal{A}_3}(t_1)) + |g(t_2)|(a_{\mathcal{A}_1}(t_2) - a_{\mathcal{A}_1}(t_1)) \\
=&|g(t_2)| - |g(t_1)|f(k) \geq 0
\end{aligned}$$

Therefore, $\hat{w}(\mathcal{A}_1) \geq \hat{w}(\mathcal{A}_4)$.

Let $\hat{\mathcal{A}}$ be the algorithm that minimizes $\hat{w}(\mathcal{A})$ among all algorithms that schedule all the transactions waiting for a shared lock together. $\hat{\mathcal{A}}$ can be modified according to the three scenarios mentioned above to bLDSF without increasing $\hat{w}(\mathcal{A})$. Therefore, bLDSF also minimizes $\hat{w}(\mathcal{A})$ among all algorithm $\mathcal{A} \in \mathbb{A}$. □

With the lemma above, we present the proof of Theorem 6.

Let $\mathcal{A}_1$ be the bLDSF algorithm, and $\mathcal{A}_2$ be an algorithm described below.

- $\mathcal{A}_2$ schedules all transactions waiting for the exclusive lock in the same order as bLDSF.
- Upon the first time bLDSF schedules a batch of transactions waiting for a shared lock, $\mathcal{A}_2$ schedules all transactions waiting for the shared lock.

For all transactions t scheduled before the transactions waiting for a shared lock by $\mathcal{A}_3$,

$$a_{\mathcal{A}_1}(t) = a_{\mathcal{A}_2}(t).$$

For all transactions t scheduled after the transactions waiting for a shared lock by $\mathcal{A}_3$,

$$a_{\mathcal{A}_1}(t) \leq v a_{\mathcal{A}_2}(t).$$

Thus,

$$d_{\mathcal{A}_1}(o) \leq v d_{\mathcal{A}_2}(o).$$

And therefore,

$$\hat{w}(\mathcal{A}_1) \leq v\hat{w}(\mathcal{A}_2). \tag{11}$$

Let $\mathcal{A}_3$ be Algorithm $\tilde{\mathcal{A}}$ described above. Then, $\mathcal{A}_2$ and $\mathcal{A}_3$ schedules the transactions waiting for an exclusive lock in the same order. Thus, for a transaction t waiting for an exclusive lock,

$$\begin{aligned}
\bar{R} + a_{\mathcal{A}_2}(t) &\leq \bar{R} + a_{\mathcal{A}_3}(t) + f(v)\bar{R} \\
&\leq f(v) a_{\mathcal{A}_3}(t) \\
&\leq v \cdot a_{\mathcal{A}_3}(t);
\end{aligned}$$

for a batch of transaction t waiting for a shared lock,

$$\begin{aligned}
\bar{R} + a_{\mathcal{A}_2}(t) &\leq \bar{R} + a_{\mathcal{A}_3}(t) + sb \cdot \bar{R} \\
&\leq v \cdot a_{\mathcal{A}_3}(t).
\end{aligned}$$

Therefore, $d_{\mathcal{A}_2}(o) \leq v \cdot d_{\mathcal{A}_3}(o)$, resulting in

$$\hat{w}(\mathcal{A}_2) \leq v \cdot \hat{w}(\mathcal{A}_3) \tag{12}$$

Let $\hat{\mathcal{A}}$ be the algorithm that minimizes $\hat{w}(\mathcal{A})$, and $\mathcal{A}_4$ be an algorithm described below.

- $\mathcal{A}_4$ schedules all transactions waiting for the exclusive lock in the same order as $\hat{\mathcal{A}}$.
- Upon the first time $\hat{\mathcal{A}}$ schedules a batch of transactions waiting for a shared lock, $\mathcal{A}_4$ schedules all the batches consecutively in the same order as $\hat{\mathcal{A}}$.

Then, for a transaction t waiting for an exclusive lock that is scheduled before the transactions waiting for a shared lock by $\mathcal{A}_4$,

$$a_{\mathcal{A}_4}(t) = a_{\hat{\mathcal{A}}}(t);$$

for any other transaction t,

$$\begin{aligned}
\bar{R} + a_{\mathcal{A}_4}(t) &\leq \bar{R} + a_{\hat{\mathcal{A}}}(t) + f(v)\bar{R} \\
&\leq f(v) a_{\mathcal{A}_4}(t).
\end{aligned}$$

Therefore, $d_{\mathcal{A}_4}(o) \leq f(v) \cdot d_{\hat{\mathcal{A}}}(o)$, resulting in

$$\hat{w}(\mathcal{A}_4) \leq f(v) \cdot \hat{w}(\hat{\mathcal{A}}). \tag{13}$$

Let $\mathcal{A}_5$ be an algorithm described below.

- $\mathcal{A}_5$ schedules all transactions waiting for the exclusive lock in the same order as $\mathcal{A}_4$.
- Upon the first time $\mathcal{A}_4$ schedules a batch of transactions waiting for a shared lock, $\mathcal{A}_5$ schedules all transactions waiting for the shared lock.

Since the time used for multiple batches of transactions to finish is more than the time used if they are put in a single batch, $a_{\mathcal{A}_5}(t) \leq a_{\mathcal{A}_4}(t)$. Therefore, $d_{\mathcal{A}_5}(o) \leq d_{\mathcal{A}_4}(o)$, resulting in

$$\hat{w}(\mathcal{A}_5) \leq \hat{w}(\mathcal{A}_4). \tag{14}$$

According to Lemma 7,

$$\hat{w}(\mathcal{A}_3) \leq \hat{w}(\mathcal{A}_5). \tag{15}$$

Let $\mathcal{A}_{OPT}$ be the optimal scheduling algorithm. Then,

$$\hat{w}(\hat{\mathcal{A}}) \leq \hat{w}(\mathcal{A}_{OPT}). \tag{16}$$

According to Equation (11), (12), (13), (14), (15), and (16),

$$\hat{w}(\mathcal{A}_1) \leq v^2 \cdot f(v) \cdot \hat{w}(\mathcal{A}_{OPT}). \tag{17}$$

Furthermore, according to Equation (4),

$$\begin{aligned}\bar{w}(\mathcal{A}_1) &\leq \hat{w}(\mathcal{A}_1) \\ &\leq v^2 \cdot f(v) \cdot \hat{w}(\mathcal{A}_{OPT}) \\ &\leq cv^2 \cdot f(v) \cdot \bar{w}(\mathcal{A}_{OPT}).\end{aligned}$$

Therefore, $\bar{l}(\mathcal{A}_1) \leq cv^2 \cdot f(v) \cdot \bar{l}(\mathcal{A}_{OPT})$.

D. EXPECTATION OF MAXIMUM OF POWER LAW RANDOM VARIABLES

For $i = 1, 2, \cdots k$, let X_i be independent power law random variables with exponent 3 on the domain $[1, \infty)$. Their CDF is given by

$$F(x) = \begin{cases} 0 & x < 1, \\ 1 - x^{-2} & x \geq 1. \end{cases} \tag{18}$$

Let $X_{\max}$ be the maximum of these random variables. Then, the expectation of $X_{\max}$ is given by

$$\begin{aligned}\mathbb{E}[X_{\max}] &= \int_0^\infty 1 - F(x)^k \mathrm{d}x \\ &= 1 + \int_1^\infty 1 - (1 - x^{-2})^k \mathrm{d}x \\ &= 1 + \int_1^\infty 1 - \sum_{i=0}^{k} \binom{k}{i} (-1)^i x^{-2i} \mathrm{d}x \\ &= 1 + \sum_{i=1}^{k} \binom{k}{i} (-1)^{i+1} \frac{1}{2i-1} \\ &= \sum_{i=0}^{k} \binom{k}{i} (-1)^{i+1} \frac{1}{2i-1} \\ &= \frac{\sqrt{\pi} \cdot \Gamma(n+1)}{\Gamma(n+\frac{1}{2})}. \end{aligned} \tag{19}$$

The series expansion of Equation (19) at $k = \infty$ is given by

$$\mathbb{E}[X_{\max}] = \sqrt{\pi k} + \frac{1}{8}\sqrt{\frac{\pi}{k}} + o(\frac{1}{k}).$$

Therefore, $\mathbb{E}[X_{\max}] = O(\sqrt{k})$.