Helenos: A Realistic Benchmark for Distributed Transactional Memory

Paweł Kobyliński[a,*], Konrad Siek[a], Jan Baranowski[a], Paweł T. Wojciechowski[a]

[a]*Institute of Computing Science*
Poznań University of Technology
Piotrowo 2, 60–965 Poznań, Poland

Abstract

Transactional Memory (TM) is an approach to concurrency control that aims to make writing parallel programs both effective and simple. The approach has been initially proposed for non-distributed multiprocessor systems, but is gaining popularity in distributed systems to synchronize tasks at large scales. Efficiency and scalability are often the key issues in TM research, so performance benchmarks are an important part of it. However, while standard TM benchmarks like the STAMP suite and STMBench7 are available and widely accepted, they do not translate well into distributed systems. Hence, the set of benchmarks usable with distributed TM systems is very limited, and must be padded with microbenchmarks, whose simplicity and artificial nature often makes them uninformative or misleading. Therefore, this paper introduces Helenos, a realistic, complex, and comprehensive distributed TM benchmark based on the problem of the Facebook inbox, an application of the Cassandra distributed store.

Keywords: Transactional memory, distributed systems, performance testing, benchmark, heterogenous distributed systems

1. Introduction

Transactional Memory (TM) [1] is an approach to concurrency control that aims to make writing parallel programs both effective and simple. The programmer applies the *transaction* abstraction to denote sections of code whose atomicity must be preserved. The TM system is then responsible for enacting the required guarantees, and doing so efficiently. Efficiency is often the key question in TM research, and various TM systems employ different concurrency control algorithms that can achieve quite divergent levels of performance depending on the workload. Hence, there is a need for empirical evaluation of their performance and the trade-offs between efficiency and features.

Initially, TMs were evaluated using microbenchmarks, but these test specific features in isolation and use data structures that are too trivial to draw general conclusions about a TM. Alternatively, there are HPC benchmark suites, but these are difficult to translate meaningfully into the transactional idiom. That is, benchmarks from SPEComp [2] or SPLASH-2 [3] are already expertly optimized to minimize synchronization, so any incorporated transactions are used rarely and have little effect on overall performance. Hence, a set of TM-specific benchmarks was needed, whose transactional characteristics and contention for shared resources were both varied and controllable. Thus, benchmarks and benchmark suites like STMBench7 [4], LeeTM [5], and STAMP [6] were developed.

*Corresponding author
Email addresses: pawel.kobylinski@cs.put.edu.pl (Paweł Kobyliński), konrad.siek@cs.put.edu.pl (Konrad Siek), jan.baranowski@cs.put.edu.pl (Jan Baranowski), pawel.t.wojciechowski@cs.put.edu.pl (Paweł T. Wojciechowski)
URL: http://www.cs.put.poznan.pl/ksiek (Konrad Siek), http://www.cs.put.poznan.pl/pawelw (Paweł T. Wojciechowski)
[1]*Telephone numbers:* (+48) 61 665 3060 (Konrad Siek), (+48) 61 665 2943 (Paweł Kobyliński), (+48) 61 665 3021 (Paweł T. Wojciechowski)

Given that distributed systems face the same synchronization problems as their multiprocessor counterparts, distributed transactions are successfully used wherever requirements for strong consistency meet wide-area distribution, e.g., in Google's Percolator [7] and Spanner [8]. Distributed TM adds additional flexibility on top of distributed transactions and allows to lay the groundwork for well-performing middleware with concurrency control. However, in such environments, unlike its non-distributed analogue, distributed TM must also deal with a flood of additional problems like fault tolerance and scalability in the face of geo-distribution or heterogeneity. Distributed TMs were developed by using transactions on top of replication [9, 10, 11, 12] or by allowing clients to atomically access shared resources or to atomically execute remote code [13, 14, 15, 16]. The latter types are of particular interest and the main focus of this paper, since they incorporate the distributed elements directly into the transaction abstraction and apply to a wide variety of distributed system models from cluster computing to clouds and web services.

As with non-distributed TMs, the variety of differently-featured distributed TMs require empirical evaluation to find how their features, the workloads, and the configuration of distributed systems influences their performance. Therefore, as with non-distributed TMs, they must be evaluated empirically. However, the existing TM benchmarks are not appropriate for distributed TMs. This is primarily the case since the structures they use are not easy to distribute. Distributing non-distributed TM benchmarks often leads to arbitrary sharding of the structure that has no purpose for the application itself (e.g., the clients still must access the entire domain). Hence distributing STMBench7, LeeTM, or *labyrinth* or *k-means* from STAMP creates applications that do not reflect realistic use cases for distributed systems. On the other hand, even if a benchmark has valid distributed variants, the conversion is often non-trivial and should not be expected to be done *ad hoc*, if it is to be uniformly applied by various research teams. As a result systems like HyFlow [13], HyFlow2 [14], and Atomic RMI [15, 16] are all evaluated using a few microbenchmarks supplemented by a distributed version of the *vacation* benchmark from STAMP, which originally mimics a distributed database use case. In effect, the results of the evaluation are not always satisfactory because of the simplicity of the testing procedure. There is also little research showing comparisons between the performance of distributed TMs.

Hence, in this paper we introduce *Helenos*, a new complex benchmark dedicated for distributed TMs, that would help to remedy the lack of tools for evaluating such systems. The benchmark proposed here is also the first step towards creation of a comprehensive framework for evaluating all facets of distributed TMs. Such a complete system for an evaluation of distributed TMs would be an indispensable tool and is much awaited in the community of distributed TM researchers. Although Helenos is only the first part of such an extensive framework, it is certainly a great move forward and a substantial improvement in the quality of distributed TM evaluation, as it is created with this particular type of test subject in mind from the very beginning. The source code of Helenos is available for download [17].

The usefulness of the Helenos benchmark is demonstrated by an example evaluation of four different distributed concurrency control mechanisms and an analysis of the produced results. The aforementioned absence of other benchmarks that can fit a similar role makes conducting a straightforward comparison between Helenos and its competitors impossible. Thus, the only way for us to showcase Helenos is to perform an example evaluation presented in this paper and show that the obtained results are understandable, predictable and explicable, as well as meaningful.

This paper is structured as follows. Section 2 discusses the existing TM benchmarks in more detail and describes how they can be used for evaluating distributed TMs or what prevents them from being practical for that aim. Section 3 introduces the Helenos benchmark and provides detail as to its data model, executed transactions, and defined metrics. Then, in Section 4, we evaluate two distributed TMs and two lock-based distributed concurrency control mechanisms. Finally, we conclude in Section 5.

2. Related Work

There are a number of TM benchmarks for non-distributed TM systems that are noteworthy. We concentrate on the ones used more often (STAMP and STMBench7), but we also present some less known benchmarks further below. We also give our attention to the benchmarks bundled with HyFlow.

2.1. STAMP

The most prevalent suite of benchmarks in use with non-distributed TM is the STAMP benchmark introduced in [6] and retrofitted to work with more modern TM systems in [18]. It consists of eight applications, each presenting a different algorithm, as a whole providing a wide range of transaction characteristics. Out of the eight benchmarks within STAMP, only some can be used as distributed applications. The best candidate for a distributed benchmark is *vacation*. It simulates a travel agency application with a database of hotels, cars, and flights, where a number of clients attempt to book one of each, while offers are sporadically modified. The database is homogeneous in nature and can easily be distributed without incurring major modifications on the benchmark. Hence, the benchmark is often used to evaluate distributed TMs [15, 16, 19]. However, the benchmark has a limited range of transaction types, so it does not constitute a comprehensive evaluation tool.

Other benchmarks which lend themselves to distribution are *genome*, *bayes*, *ssca2*, *yada*, and *intruder*. The applications are based respectively on algorithms for gene sequencing, learning structures of Bayesian networks, creating efficient graph representations, Delaunay mesh refinement, and detecting intrusions in a network. The processing in each of these applications follows a pipeline where a complex data structure is processed into another form in a series of steps by multiple simultaneous threads. Such processing can be distributed among several network nodes (e.g. in high performance clusters) in order to provide more system resources (processing power, memory) to the algorithms. The issue with producing such distributed versions of these algorithms is that it cannot be done *ad hoc*, and often requires that an expert-prepared variant exists with at least a reference distributed implementation that could be simply instrumented by TM researchers. In any case, from the perspective of distributed system architecture *genome*, *bayes*, *ssca2*, *yada*, and *intruder* are all an example of a single use-case, and therefore even though these applications provide breadth for a concurrent benchmark, they provide little breadth in the distributed context.

The *kmeans* benchmark represents a K-means clusterer known from data mining. The algorithm executes in rounds: in every round a thread reads the values of some partition of objects and designates one of them the new center of the cluster, then all the data objects are reassigned to the closest center. While the data for the clusterer can be distributed onto multiple nodes, the resulting variant does not present a genuine distributed use case, because execution in rounds induces too high a level of coordination required among client threads. We find the remaining *labyrinth* benchmark to be a similar case. There, transactions operate on a central data structure representing a 3D maze and attempt to route a path from one point to another using Lee's algorithm. Since a path cannot intersect other paths, conflicting writes must be avoided. Such an application can be employed for circuit-board design. However, even if the maze is distributed, as with *kmeans*, the application has no reflection in distributed system use cases.

2.2. STMBench7

The other popular STM benchmark, STMBench7 [4], is based on an object-oriented database benchmark. In STMBench7 clients perform a wide range of transactions on a shared tree-based database. The tree, called a module, contains three levels of nodes: *a) complex assemblies (CAs)*, whose children are either other complex assemblies or base assemblies, *b) base assemblies (BAs)*, which link to several composite parts, or *c) composite parts (CPs)*, the leafs of the tree whose payload is a document and a graph. Each element contains links to its parent as well as children, allowing bottom-up or top-down traversal. Transactions in STMBench7 are either *traversals* (accessing a path from root to leaf), *queries* (accessing sets of random nodes), or *structure modifications* (adding or removing parts of a tree). They can be either long or short, and either read-only or update. The benchmark simulates a CAD/CAM/CASE application or a multi-user server.

The benchmark can be used for distributed TMs if the tree structure is spread among several servers in a network. This can be done in one of three ways: The benchmark can be distributed per CP, so that each CP is located on the same network node as its children, but it can be located on a different server than its parent base assembly, which, in turn, can be located somewhere different than its own parent complex assembly, etc. From the perspective of the client this leads to a system with a flat structure, where the tree structure from the original application becomes a simple collection of distributed objects with some objects referring to other ones. From the distributed TM's point of view the difference between CPs, BAs,

and CAs largely disappears, since their internal workings will be largely obscured. It is another practical consequence of this, that the difference between queries and traversals becomes blurred, since the only distinguishing feature between them now is how the access sets of transactions are selected (from other object in traversals rather than from an index in queries). In effect, this variant creates a very simple benchmark, which resembles the bank microbenchmark used with distributed TMs, only with a new transaction type—structure modifications. In addition, the benchmark can be distributed per CA or BA, i.e., a BA and its children are located on the same node, but any two BAs or CAs can be located in different parts of the network. This variant allows groups of objects to be treated as a single structure by the clients, and, rather than accessing each object individually, transactions can treat a BA with all its children jointly. However, from the point of view of a distributed TM, this again becomes a flat distributed collection of remote objects. In addition to the former variant, however, the benchmark is further simplified, because the number of interdependencies between objects is smaller. This is not offset by the increased complexity of remote objects, which simply means that processing time of an access increases, but does not have any other effect on transactional processing. Finally, the benchmark can be distributed by adding new modules and distributing per module. However, as the authors themselves note in [4], increasing the number of modules (or distributing by module) isolates transactions from one another, and is therefore not useful for evaluating TMs. Neither of the three variants provide a satisfactory tool for evaluating distributed TMs, since distribution leads to simplification and divorcement from the original realistic CAM/CAD/CASE use case.

2.3. HyFlow Benchmarks

The microbenchmarks and benchmarks included by the authors of HyFlow [13] in their implementation represent what can be considered to be the best available set of benchmarks for evaluating distributed TM systems. The suite consists of three microbenchmarks, bank, loan, and distributed hashtable (DHT), as well as a distributed version of the vacation benchmark from STAMP (described above). DHT is a micro-benchmark where each server node acts as a shard of a distributed key-value store. Transactions atomically perform a number of writes or a number of reads on some subset of nodes. The bank benchmark simulates a straightforward distributed application using the bank metaphor. Each node hosts a number of bank accounts that can be accessed remotely by clients who perform transfers between accounts or atomic reads of several accounts. Finally, the loan benchmark presents a more complex application where the execution of transactional code is also distributed among several nodes. Each server hosts a number of remote objects that allow write and read operations. Each client transaction atomically executes two reads or two writes on two objects. When a read or write method is invoked on a remote object, then it also executes two reads or writes (respectively) on two other remote objects. This recursion continues until it reaches a specified depth. Hence, the benchmark is characterized by long transactions and high contention, as well as relatively high network congestion, and is unique in focusing on the control flow transactional model.

However, while the benchmark suite is able to shed light on the performance of distributed TMs and includes a distributed TM-specific applications, it lacks complex benchmarks and therefore does not comprise a comprehensive evaluation tool.

2.4. Other TM Benchmarks

A number of other applications were used to test TM systems apart from the ones mentioned above. Notably, LeeTM [5] is an independent implementation of Lee's algorithm, analogous to the one used in STAMP's labyrinth. The benchmark has limited use in distributed TM evaluation for the same reasons as its STAMP counterpart. EigenBench [20] is a comprehensive and highly configurable microbenchmark for multiprocessor TM evaluation. One of its interesting features is that it allows induction of problematic executions (e.g. convoying) to observe a TM's behavior in their presence. EigenBench uses a number of flat arrays as its data structure, so it is easy to distribute for use with distributed TM. However, EigenBench is designed to test individual characteristics of a TM, and as such does not generate complex workloads containing heterogenous transactions. Another interesting application is Atomic Quake [21], a transactional

implementation of a multithreaded game server, which was used to compare the performance of TM with locking in a complex realistic use case. The benchmark uses a central server to coordinate remote clients, whose nature prevents the application from being employed as a truly distributed benchmark. Finally, in [11] the authors present a custom benchmark application for replicated transactional memory called Twitter Clone which simulates the operation of a social networking service. The benchmark is not useful for non-replicated TM, but the overall idea influenced the benchmark presented in this paper.

2.5. Distributed database benchmarks

There is another group of benchmarks that are being used to test distributed TMs, which were derived from distributed database benchmarks (such as a series of popular TPC benchmarks or others). As shown in [22], they can be altered to work as an evaluation tool for distributed TMs. Another example is Yahoo Cloud Serving Benchmark (YCSB) [23] which was proposed to compare performance of NoSQL database management systems. As they are designed to work in distributed environment and test the performance of data storage systems, they can be adapted to work with distributed TMs in a relatively straightforward way. However, databases differ from TMs in a completely fundamental way. They are used in different business scenarios , play a different role in larger systems, the underlying theory (such as consistency constraints) is different as well as many characteristics used to describe both of them. Additionally, in the distributed TM research the problem of the distributed data storage is only a part of a much larger picture. For instance, it also comprises the execution of arbitrary code within transactions. Distributed databases on the other hand are only responsible for correctly storing the data in a distributed environment. Therefore, benchmarks created to evaluate them assess characteristics useful for their descripton, but only partially useful for benchmarking distributed TMs – excluding such general metrics as total execution time, parallel transactions count or general throughput. So despite being a fairly good substitute benchmarks for distributed TMs, they are still clearly inferior to the benchmarks especially designed for this kind of test subjects.

3. The Helenos Benchmark

We introduce the Helenos benchmark which provides a platform for comprehensively evaluating a distributed TMs for use in non-uniform large scale distributed systems. Such a benchmark must contain a wide variety of transaction types and provide a high level of control over the workload that a TM can be subjected to. This allows thoroughly to test all aspects of a distributed TM, and observe its performance in diverse environments. However, it is equally important that a benchmark be evaluated in a realistic setting, i.e., such where there exists a need for both transactional memory and for distributed systems. The benchmark is implemented in Java, which allows us to interface with some of the existing distributed TM systems either also implemented in Java [13, 15, 16] or other languages on the Java Virtual Machine [14]. Places where transactions start and end must be manually marked for each of the tested libraries allowing great flexibility of measurement.

In order to achieve our goal of providing a realistic use case for a distributed TM, we base our benchmark on Cassandra [24], a distributed storage system for large volumes of structured data that emphasizes decentralization, scalability, and reliability. Cassandra was created to serve as a storage system for the user-to-user message inbox in Facebook, where it must be able to withstand large write-heavy throughputs. The authors specify a data model that is used for Cassandra's intended application, as well as two typical search procedures: *term search* (find a message by keywords) and *interaction search* (find all messages exchanged between two users). Cassandra allows to control consistency of its operations by managing which replicas reply to certain requests. However, the consistency guarantees can be further extended by introducing the transaction abstraction. Thus, the Cassandra implementation of the inbox is both an inherently distributed application, as well as one which can benefit from using the TM.

Hence, we use Cassandra's application as the Facebook inbox as the basis for our benchmark for distributed TM. We do this by implementing the data model specified in [24] and supplement the original search procedures with a comprehensive spectrum of pertinent transactions. On the other hand, we remove

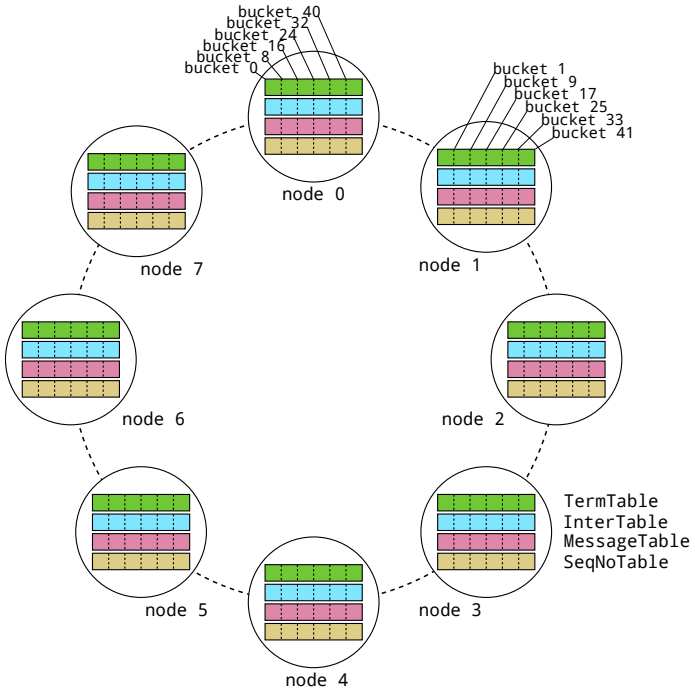


Figure 1: Database architecture.

features that are not directly *á propos*, including fault tolerance, local persistence, and partial replication. This makes any evaluation results simpler to predict and analyze, but does so without loosing the nuance of the original benchmark.

3.1. Data Model

We follow Cassandra's data model and a logical data model from the inbox application in [24]. First of all, a Cassandra cluster is composed of several nodes forming a logical topology of a ring. Each server node being assigned an arbitrary (random) position in the ring. As described in [24], the dataset is distributed among these nodes using consistent hashing. This means that for each data item a position in the ring is derived from a hash of it's key, and then that item is assigned to the first node with a greater position. In this way the data items are "wound around" the ring (multiple times) and one node becomes responsible for multiple ranges of keys. We refer to each such range as a *bucket* and use this level of granularity for synchronization. That is, two transactions conflict if they try to access the same bucket. We show this model in Fig. 1.

The original inbox application data model consists of two tables distributed in a Cassandra cluster, one for term search and one for interaction search. Each table row is a separate data item containing data from all of the table's columns in that row (as the object's fields). The users of the application all have unique identifiers (`UserID`), as do all messages in the system (`MsgID`). Note that the data resolution changes between the original application and the benchmark, since bucket-level granularity gives us a greater level of control over contention.

The term search table (`TermTable`) is used to find messages in a user's inbox by keywords. The logical structure of the table consists of one column containing a `UserID` and one column for each possible keyword. Each keyword column contains lists of message identifiers (`MsgID`s) of messages that include this keyword. More formally:

$$\texttt{TermTable} : \texttt{UserID} \mapsto \texttt{Keyword} \times (\texttt{MsgID list})$$

This schema assumes that the number of users is orders of magnitude greater than the size of one user's inbox. However, this is problematic for the purposes of a benchmark, since it makes contention both lower and more difficult to adjust. Hence, we decided to modify this table to increase the number of keys by moving the keyword data from separate columns into the key. In effect, the number of objects per user is increased and contention is no longer strictly dependent on the number of users but can be controlled by adjusting the population of keywords. Hence our schema looks as follows:

$$\texttt{TermTable} : \texttt{UserID} \times \texttt{Keyword} \mapsto \texttt{MsgID}\ \texttt{list}$$

Note that, since the benchmark only stores hashes of keys, the hashed keyword information has to be kept elsewhere in addition to the key column, but we omit this detail for clarity.

The interaction search table (`InterTable`) stores conversations between users. The table's key is the identifier of the user sending the message. The table also contains one column per receiving user which holds the IDs of the messages exchanged between the two users. Thus:

$$\texttt{InterTable} : \texttt{UserID} \mapsto \texttt{UserID} \times (\texttt{MsgID}\ \texttt{list})$$

This schema encounters the same problems as `TermTable`, so we perform a similar modification in order to increase the key domain: we use an ordered pair indicating the sender and the receiver of the message as the key, rather than just the sender. Thus, in the benchmark the table is defined as:

$$\texttt{InterTable} : \texttt{UserID} \times \texttt{UserID} \mapsto \texttt{MsgID}\ \texttt{list}$$

We extend the data model by adding additional tables which can be used to introduce new functionality into the application, and in this way increase the benchmark's depth—widen the range of transactions that can be performed. The message table (`MessageTable`) stores the contents of the messages used in the system (note that `TermTable` and `InterTable` both operate on message identifiers alone). The table's key is the identifier of the recipient of the message, and the value is the message, which can be expressed as five columns: the identifier of the message, the user identifiers of the sender and the recipient, the text of the message, and the timestamp of when it was sent. (The timestamps are used to sort messages, allowing transactions to return a group of most recent ones). In effect, we have the following definition:

$$\texttt{MessageTable} : \texttt{UserID} \mapsto (\texttt{MsgID} \times \texttt{UserID} \times \texttt{UserID} \times \texttt{Text} \times \texttt{Time})\ \texttt{list}$$

Finally, we employ a sequential number table (`SeqNoTable`) which is used to generate new message identifiers for a given user's inbox and to specify a cut off point for deleting old messages. Thus, the table's key is the identifier of an owner of an inbox and the data columns include the sequence number (`SequenceNo`) of the most recently created message in the inbox and the sequence number of the last deleted message there. This is defined as:

$$\texttt{SeqNoTable} : \texttt{UserID} \mapsto \texttt{SequenceNo} \times \texttt{SequenceNo}$$

The actual data to be used by the benchmark during operation is generated randomly while creating a scenario. The messages exchanged by the users are created using words from a dictionary file. This implies that the number of words in the dictionary is inversely proportional to the overall contention for the keyword buckets.

3.2. Tasks and Transactions

Clients in the benchmark can run 8 types of high-level tasks on the distributed database. Each of these tasks performs some non-transactional operations (primarily locating nodes in the ring from hashes and performing local processing) and employs one or more atomic transactions. The first two tasks: term search and interaction search are a part of the original example in [24], and the remainder are domain-appropriate additions that ensure the configurability of the benchmark. We describe the tasks below and provide detail about the transactions they use. All transactions are further detailed in Fig. 2. We also provide a summary in Fig. 3.

```
atomic getByKeyword(myUserID, myKeywords) {
  for (myKeyword : myKeywords)
    for ((userID, keyword) : TermTable)
     if (userID == myUserId and keyword == myKeyword) {
       msgIDs = TermTable[(userId, keyword)];
       result.addAll(msgIDs);
     }
}

atomic getMessages(msgIDs) {
  for (msgID : msgIDs)
    for (recipientID : MessageTable)
      if (recipientID == msgID.recipientID) {
        messages = MessageTable[recipientID];
        for (msg : messages)
          if (msgID == msg.msgID)
            result.add(msg);
      }
}

atomic getConversation(senderID, recipientID) {
  for ((sendID, recpID) : InterTable)
    if (sendID == senderID and recpID == recipientID)
      result.addAll(InterTable[(senderID, recipientID)]);
}

atomic sendMsg(senderID, recipientID, content, keywords) {
  SeqNoTable[recipientID] += 1;
  (currentSeq, deletedSeq) = SeqNoTable[recipientID];
  msgID = new MsgID(recipientID, currentSeq);
  timestamp = currentTimestamp();
  msg = new Message(msgID, senderID, recipientID,
                    content, timestamp);
  MessageTable[recipientID].add(msg);
  InterTable[(senderID, recipientID)].add(msgID);
  InterTable[(recipientID, senderID)].add(msgID);
  for (keyword : keywords)
    TermTable[(recipientID, keyword)].add(msgID);
}

atomic importMessages(messages) {
  for (msg : messages) {
    (currentSeq, deletedSeq) = SeqNoTable[msg.senderID];
    if (msg.sequenceNo < deletedSeq)
      continue
    if (MessageTable[msg.recipientID].exists())
      continue
    MessageTable[msg.recipientID].add(msg);
    InterTable[(msg.senderID, msg.recipientID)].add(msgID);
    InterTable[(msg.recipientID, msg.senderID)].add(msgID);
    for (keyword : msg.keywords)
      TermTable[(msg.recipientID, keyword)].add(msg.msgID);
  }
}

atomic resetCutoff(userID) {
  (currentSeq, deletedSeq) = SeqNoTable[userID];
  result.add((currentSeq, deletedSeq))
  SeqNoTable[userID] = (currentSeq, currentSeq);
}

atomic removeMessages(messages) {
  for (msg : messages) {
    for (keyword : msg.keywords)
      TermTable[(msg.userID, keyword)].remove(msg.msgID);
    InterTable[(msg.senderID, msg.recipientID)].remove(msg.msgID);
    InterTable[(msg.recipientID, msg.senderID)].remove(msg.msgID);
    MessageTable[(msg.recipientID)].remove(msg.msgID);
  }
}

atomic getAssociation(userID1, userID2) {
  result[MSG_ID].addAll(InterTable[(userID1, userID2)]);
  result[MSG_ID].addAll(InterTable[(userID2, userID1)]);
  result[SEQ].add(SeqNoTable[userID1]);
  result[SEQ].add(SeqNoTable[userID2]);
}

atomic indexMessages(queries) {
  for ((userID, keywords) : queries)
    for (keyword : keywords) {
      msgIDs = TermTable[(userID, keyword)];
      for (msgID : msgIDs) {
        messages = MessageTable[msgID.recipientID];
        for (msg : messages)
          if (msgID == msg.msgID)
            result[MSG].add(msg);
      }
     }
   result[MSG].sortByTimestamp();
   for (userID : queries)
     result[SEQ].add(SeqNoTable[userID]);
}
```

Figure 2: Benchmark transactions pseudocode.

Term Search. A user (identified by a specific `UserID`) specifies a set of keywords. The user's inbox is then searched to find all messages which contain one of the keywords by using transaction `getByKeyword`. This transaction traverses `TermTable` and returns all message identifiers whose key both matches the user's ID and at least one of the specified keywords. This section must be atomic to return a consistent state of the inbox. Finally, the contents of the matching messages are retrieved with transaction `getMessages`, which retrieves the contents of the messages from `MessageTable` on the basis of the list of message identifiers. Again, this is done atomically, in order to return a consistent snapshot of the inbox. Overall, the term search task reads from two tables: `TermTable` and `MessageTable` and we estimate it to be medium in length. The length of the task will depend on the size of the keyword domain. Depending on the distribution of buckets among nodes and the length of the query this task may involve from one to all nodes.

Interaction Search. A user specifies a `UserID` of another user in order to find all conversations between the two of them. In order to do this, the task runs transaction `getConversation`, which gets all message identifiers from `InterTable` whose keys fit the user identifiers of both users in question. Then, the contents of the messages are retrieved using `getMessages`. Thus, interactions search reads from two tables: `InterTable`

and `MessageTable` and is of small length, depending on the number of users in the system and the number of messages in both the users' inboxes. Depending on the distribution of buckets among nodes this task may involve from one to four nodes.

Send Unicast. A user, the sender, sends a message consisting of a text content to another user, the recipient. This consists of executing transaction `sendMsg` which is preceded by extracting the set of keywords from the contents of the message. The transaction then increments and retrieves the next sequence number from `SeqNoTable` and uses it to create a new `MsgID`. Then, a message object is created and written into `MessageTable`. Subsequently, the identifier of the message is inserted into `InterTable` twice: once to indicate the message sent out from the senders inbox, and once to indicate the message received in the inbox of the recipient. Finally, the transaction inserts an entry into `TermTable` for each keyword extracted from the contents of the message. All these actions must be performed atomically, in order to prevent another transaction from interfering and causing an inconsistent state. E.g., if a message were removed by another transaction from the `MessageTable` while this task was yet to add all the keywords to `TermTable`, the contents of `TermTable` would point to messages which no longer existed. Overall, this task updates `SeqNoTable` and writes to all the other tables. On the whole, this is a short read/write task with a relatively limited read/write set (R/W set). Depending on the distribution of buckets among nodes and the content of the message this task may involve from one to all nodes.

Send Multicast. A user sends a single message to several recipients. This task is an extension of the send unicast task (defined above), where the `sendMsg` transaction is executed in series, but the entire series is executed atomically. This task is a medium- to large-sized write task that touches the same tables as the send unicast task, but it strongly depends on the number of recipients. Similarly to the send unicast task this one may involve from one to all nodes.

Batch Import. The system is given a set of complete messages to store in the database. This represents a use case where the database is replicated and two of the replicas synchronize, by one sending a state update to the other. Another pertinent scenario is one where the database tries to recover from a crash. The task involves extracting keywords from each message's contents and executing transaction `importMessages`. The transaction filters the set of imported messages against `SeqNoTable` to remove all those that have a sequence number lower or equal to the sequence number of the last deleted message. If a message has a higher sequence number than the current highest sequence number in `SeqNoTable`, then the database is also updated. Then, the transaction filters out those messages which are already in `MessageTable`. All the remaining messages are added to `MessageTable`, `InterTable`, and `TermTable` by analogy to `sendMsg`. The task is a long write task (depending on the number of imported messages) which updates `SeqNoTable` and MessageTable, and writes to the two remaining tables. Similarly to the send unicast task this one may involve from one to all nodes.

Clear Inbox. A user requests that all messages from her inbox be removed. This task sequentially executes three separate transactions. First, the task runs transaction `resetCutoff` which retrieves the sequence numbers of the most recent message in the user's inbox as well as the sequence number of the most recent deleted message from `SeqNoTable`. In addition, the transaction sets the deleted sequence number to the current sequence number, signifying that all messages in the inbox are now below the cutoff for deletion. Next, the task uses transaction `getMessages` to retrieve all the existing messages in the inbox. Afterward, the task executes transaction `removeMessages` to remove all the messages from the three tables in the database that hold message information. Specifically, for each keyword of each message to be removed, the transaction removes each message's identifier from `TermTable`. Then, on the basis of the recipient and sender of each message, the transaction removes both occurrences of its message identifier from `InterTable`. Finally, the transaction searches each message in `MessageTable` by their recipient and remove them from there. The task is a medium-sized read/write tasks (although the execution can be shorter depending on the number of messages in the inbox) and touches all tables. The task is not executed atomically as a whole because any potential inconsistencies stemming from the lack of atomicity between transactions can be

Task	R/W set	Length	Used transactions
association level	read-only	short	`getAssociation`
term search	read-only	medium	`getByKeyword`, `getMessages`
interaction search	read-only	short	`getConversation`, `getMessages`
indexing	read-only	long	`indexMessages`
send unicast	read/write	short	`sendMsg`
send multicast	read/write	medium/long	`sendMsg`
batch import	read/write	long	`importMessages`
clear inbox	read/write	medium	`resetCutoff`, `getMessages`, `removeMessages`
Transaction	**R/W set**	**Length**	**Touched tables**
`getByKeyword`	read-only	short/medium	`TermTable`
`getConversation`	read-only	short	`InterTable`
`getAssociation`	read-only	short	`InterTable`
`getMessages`	read-only	medium	`MessageTable`
`indexMessages`	read-only	long	`TermTable`, `SeqNoTable`
`resetCutoff`	read/write	short	`SeqNoTable`
`sendMsg`	read/write	medium	`TermTable`, `InterTable`, `MessageTable`, `SeqNoTable`
`importMessages`	read/write	long	`TermTable`, `InterTable`, `MessageTable`, `SeqNoTable`
`removeMessages`	read/write	medium	`TermTable`, `InterTable`, `MessageTable`

Figure 3: Overview of tasks and transactions in Helenos.

fully and easily resolved by the application. Thus this decomposition makes for a more realistic workload. Depending on the distribution of buckets among nodes and contents of the removed messages this task may involve from one to all nodes.

Association Level. Given two users, the system checks what is the level of interaction between them, by counting the number of exchanged messages and normalizing the number against the number of messages in the inbox. This involves running transaction `getAssociation`, which retrieves the identifiers of messages involved in conversations between both users (from `InterTable`). It also retrieves the sequence numbers of the most recent existing and the most recent deleted message in both users' inboxes (from `SeqNoTable`). The results of these executions are then processed non-transactionally by the task. This produces a short read-only task operating on a single table—`InterTable`. We add this task to have a finer control over contention, since the existing short read-only transactions in interaction search and term search are always followed by a medium-sized read-only transaction. Depending on the distribution of buckets among nodes this task may involve from one to four nodes.

Indexing. Creates a cache for the most common keyword searches by the most active users. Given a list of users and a list of keyword search queries per user the task runs transaction `indexMessages`. For each user and for every keyword, the transaction searches through `TermTable` and collects message identifiers of all pertinent messages. Then, unique message identifiers are extracted from the list. The results are then sorted and cropped to a prescribed length. Then, the transaction retrieves the body of the three example messages for each user from `MessageTable`. Finally, the transaction retrieves sequence number information from `SeqNoTable` for each of the investigated users. Indexing is a work-intensive read-only task that contains a long transaction with a large readset. Depending on the distribution of buckets among nodes and the contents of the queries this task may involve from one to all nodes.

Fig. 3 contains the most important information about the tasks described above in a concise form, i.e. their length and whether they are read-only or not, as well as which transaction types are used in each of them. It also contains similar information about the transaction types used in the tasks, including the tables that are touched by each of them.

3.3. Metrics

In recognition of the fact that complex systems require comprehensive metrics, the benchmark implementation includes several metrics useful for monitoring distributed TM performance.

Classically, we provide the commonly employed *throughput* metric, defined as the number of transactions executed per second by the system. We also measure *latency* (or *mean flow time*), which is an alternative performance goal to throughput. A single transaction's latency (*flow time*) is the time between a transaction commences and finally completes. This time includes any time spent re-executing the transaction due to forced aborts. The system's metric is the mean of all transactions' latency. We measure latency on the system level as well as per transaction. Furthermore, we allow measuring transactions' *abort rate*, which indicates how often transactions are forced to re-execute. These metrics allow to measure to what extent a TM's overall performance goals were reached. They also constitute the most basic tool for comparing different TMs.

In addition, we provide more diagnostic metrics that allow to reason about the workload itself. First of all, we measure *retry rate*, *the total execution time of all transactions*, *the total execution time of all retries*, *sum of all startup times for all transactions*, i.e. the amount of time between a start of a transaction and a start of its first retry, as well as the *the total number of operations on all buckets*. Finally, we supplement these measures with metrics showing *the total execution time* of the whole benchmark run, *the total parallel execution time*, i.e. the time between the start of the first client thread and the end of the last client thread, and *transactional execution ratio*, the percentage of time the system spent within transactions.

3.4. Parameters

The benchmark includes a number of parameters that can be used to evaluate a distributed TM in a range of workload types. The most important aspect controlled by the parameters is contention, which can be controlled by adjusting the number of shared objects in relation to the number of transactions that use these objects on average, and to the number of objects used by each transaction. Larger R/W sets or more clients in relation to the same number of objects means increased contention.

We allow fine control over the domain of shared objects by changing the configuration of the system: the number of network nodes, and the number of remote objects (buckets) per table. We also allow changing the domain of words allowed in messages, which increases the number of keys in `TermTable`. The number of transactions is controlled by adjusting the number of simultaneous clients in the system, as well as the number of tasks executed by each of them. Finally, the R/W sets of transactions can be controlled directly by adjusting the cap on the number of keywords allowed in a query (`getByKeyword`, `indexMessages`) and setting the minimum and maximum number of messages for `importMessages`.

The benchmarks also allow to control the composition of the workload, by specifying what percentage of which types of transactions will be executed. This allows to test various features of distributed TMs that depend on e.g., a high ratio of read operation or read-only transactions. In addition, since moving data around a network can have a significant cost, we allow adjusting the size of data by setting the maximum length of the content of messages in `MessageTable`. Hence caching and storage strategies used by various distributed TMs can be evaluated in various environments. Lastly, we introduced a configurable delay in milliseconds that is applied to every operation on every bucket in the system. It can be used to mimic higher network latency or slower nodes.

4. Evaluation

In this section we present an example of empirical evaluation of several distributed concurrency control mechanisms using Helenos, including two distributed TMs that represent two different approaches to synchronization. We must stress, however, that the purpose of the paper is not to provide a comprehensive evaluation of the TM systems in question (both of which were thoroughly evaluated in [16, 14]). Instead, the main objective of our evaluation is to demonstrate how particular settings of the benchmark impact the system's workload and the performance of disparate concurrency control schemes. In effect, this shows that Helenos' parameters have strong influence on its behavior and they can be used to evaluate different aspects of the test subject. It is also demonstrated that the benchmark can be used to mimic complex, real-life distributed systems. In addition, we showcase how an evaluation can be used to draw meaningful conclusions that can help diagnose problems within the evaluated system and provide a reference benchmark for future research in distributed TM.

4.1. Frameworks

For the purpose of comparison we use four distributed concurrency control frameworks, including two distributed TM implementations that represent two main approaches to transactional synchronization, and two typical distributed locking schemes.

The first distributed TM we evaluate is HyFlow2 [14], a state-of-the-art optimistic data-flow distributed TM implemented in Scala. The optimistic approach to concurrency control means that when two transactions conflict on a shared object, one of them is aborted and re-executed. The data-flow model means that whenever a client accesses a shared object, the object is moved to the client's node for the duration of the access (but there is always exactly one copy of each object in the system). HyFlow2 implements the Transaction Forwarding Algorithm (TFA) [25] to handle synchronization (with the guarantee of opacity [26]) and uses the Akka library for networking. For the evaluation we configure HyFlow2 to use standard Java serialization.

The second distributed TM is Atomic RMI [16], a pessimistic control-flow distributed TM implemented in Java, on top of Java RMI. The pessimistic approach means transactions defer operations to avoid conflicts altogether and thus preventing transactions from aborting. The control flow model means that shared objects are immobile and execute any code related to accesses on the nodes they are on. Atomic RMI uses the Supremum Versioning Algorithm (SVA) [27] for concurrency control, which guarantees last-use opacity [28].

The two locking schemes used for the evaluation are fine grained locks (FGL) and a global lock (GLock). FGL simulate an expert implementation of a distributed locking scheme using mutual exclusion locks with one lock per shared object. Locks in FGL are always acquired and released according to two-phase locking (2PL) and used according to a predefined global locking order to prevent deadlocks. Locks are released in FGL after the last access to a particular object within a given transaction. Such an implementation of 2PL is trivially last-use opaque. GLock is a locking scheme where the system contains a single lock on one of the nodes that must be acquired at the beginning of each transaction and held throughout each transaction's duration. This means that only one transaction can be executed at a time in the entire system. Hence, this is an anchor implementation, roughly equivalent to using a sequential baseline execution in non-distributed TM evaluations. GLock is trivialy opaque. Both locking schemes used for the evaluation are built on top of Java RMI and use custom lock implementations.

4.2. Testing Environment

We perform our evaluation using a 16-node cluster connected by a 1Gb network. Each node is equipped with two quad-core Intel Xeon L3260 processors at 2.83 GHz with 4 GB of RAM each and runs a OpenSUSE 13.1 (kernel 3.11.10, x86_64 architecture). We use the 64-bit Java HotSpot(TM) JVM version 1.8 (build 1.8.0_25-b17).

4.3. Parameter Settings

As the first part of our evaluation we showcase the benchmark's configurability by displaying the behavior of the four frameworks within diverse workloads generated by manipulating the parameters of Helenos. The starting point for the reconfiguration is the *standard* workload, which has an 80–20% read-to-write-task ratio (the probabilities of executing a specific task are shown in Fig. 6) typically used for TM evaluation. The basic configuration of the standard scenario contains 1024 buckets deployed on 16 nodes, with 200 concurrent clients executing 3 consecutive tasks each. The message length is set to 8 words and the operation delay is set to 3ms. We treat these parameters as our reference for comparing TM implementations. The parameters were selected for the standard benchmark on the basis of experimentation to simulate a well-behaved distributed TM system. We then show how the manipulation of these reference values impacts the performance of the various concurrency control mechanisms. We test each of the parameters independently. If it is not explicitly stated otherwise the values of all parameters are exactly as in the standard scenario.

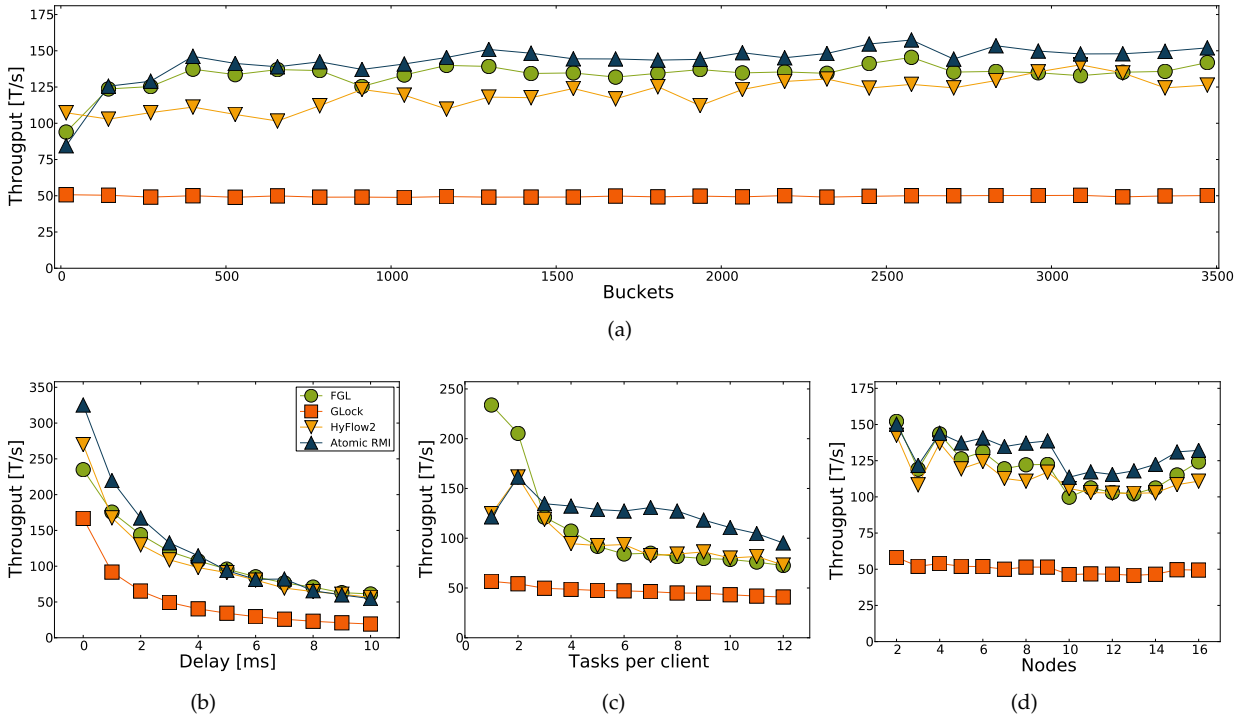


Figure 4: Impact of parameters on throughput.

Buckets. Fig. 4a shows the impact of manipulation of the number of buckets used within the system on throughput. As the number of buckets is increased, transactions are less likely to access the same bucket, so the system's contention decreases. A decrease in contention leads to an increase in throughput for all evaluated frameworks, but the change has varying impact. The evaluation shows that Atomic RMI and FGL both capitalize on higher contention in the range between 1 and 500 buckets, which is attributable to the higher costs of conflict avoidance in those frameworks. This is contrasted with HyFlow2, whose throughput increases more gradually in the same range, as contention lowers. When the number of buckets exceeds 1000, any further increase does not result in an improvement in throughput, showing that at this point contention is negligible and other factors, like operation overhead, dictate performance. Given these results, we use 1024 buckets as the reference value.

Delay. Fig. 4b shows the transactional throughput of the four frameworks as the length of the operations executed by transactions changes. Naturally, as the length of each operation increases, the throughput of the system decreases (even for a sequential execution simulated by GLock). Here, all frameworks behave similarly, showing a steady decrease in performance, that converges around the 10ms mark. The reference delay value we select is 3ms, at which point the throughput for FGL, HyFlow2, and Atomic RMI is in the vicinity of 175 transactions per second, with a slight divergence.

Tasks per client. Fig. 4c shows the change in transactional throughput as the number of consecutive tasks executed by each client increases. The results clearly show anomalous behavior for low task numbers, that significantly diverges from the results for higher task values. The divergence stems from the heterogeneity of tasks in Helenos, which means that for low task values some clients will finish their tasks much faster than their counterparts that execute longer tasks at the same time. If the number of consecutive tasks is low, the clients with shorter tasks finish execution early and leave the system with a diminished number of concurrent clients, thus randomly, and unwarrantably decreasing the system's contention. When the number of tasks per client is increased, the results show that performance steadies for all frameworks. On

the other hand, increasing the task count leads to a prohibitive total execution time for the entire benchmark (e.g. for FGL, the execution time for each data point is 35.2s for 3 consecutive tasks, 43.1s for 4, and 96.4s for 10). Hence, we select 3 tasks, the least value where the performance is not anomalous, for the reference value.

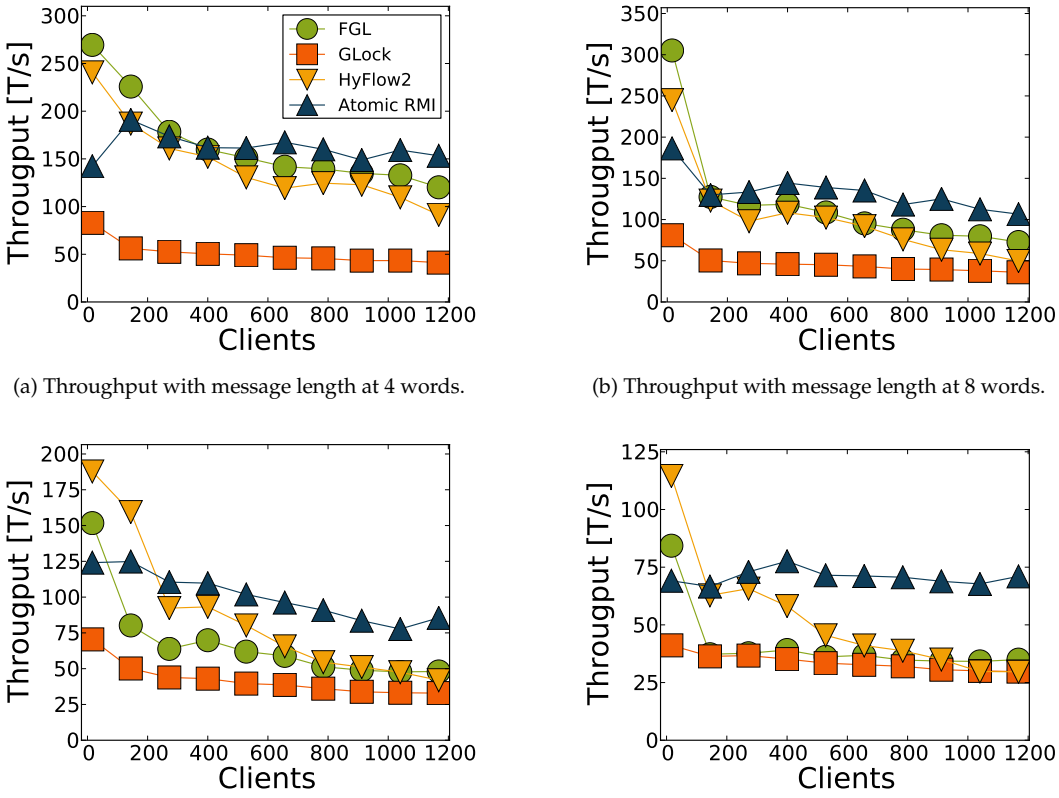
Nodes. Fig. 4d shows the reaction of the four frameworks to changes in network size. As nodes are introduced into the distributed system, the contention and transactional characteristics remain constant, which causes little change in the throughput values for all frameworks. The number of nodes, along with the number of buckets, decide how buckets are distributed in the system. Given a constant number of buckets, the lower the number of nodes, the greater the number of buckets that are hosted per node. This change itself does not impact the behavior of the TM frameworks themselves, however. What the fluctuation in throughput reflects is the impact of the physical properties of the testing environment itself in general, and the network topology as well as communication costs in particular, which is consistently more pronounced for certain configurations (i.e. for a system with 3 nodes or 10 nodes). The reference value for node count selected for the benchmark is 16, since the TM applications are meant to be used in larger rather than smaller systems, and that is the largest configuration available to us. There is no recommended number of nodes for the benchmark, however our experience shows that node counts below 4 can sometimes produce anomalous results. While using the benchmark we have also noted that only a small part of execution time was used by the benchmark itself and not by the tested algorithm. Additionally, this part was not increasing with heavier load or more nodes in the system. Therefore, we can speculate that the benchmark would not be a bottleneck and allow execution on much greater number of nodes than showed in this paper, provided the tested algorithm can scale up to this point as well.

Message length. Fig. 5 shows the impact of setting a different message length for the contents of messages, as it affects `TermTable`. An increasing message length has a visible effect on all those transactions that access `TermTable`, since it increases the access sets of transactions that read from that table (because parts of a message are distributed among a greater number of buckets) and the length of write operations on that table. Hence, increasing the message length impacts contention, and thus decreases the transactional throughput of each framework. The impact is most visible for FGL, where an increased message length causes a significant drop in the presence of high client numbers. The reference value for message length is set at 8 words.

4.4. Scenarios

Fig. 7 shows a comparison of the four benchmarks using different workloads, wherein clients execute different sets of tasks. The composition of particular scenarios in use is shown in Fig. 6, but generally, scenarios marked *small* consist primarily of short and medium tasks, with a minimal number of long tasks, whereas scenarios marked *large* consist primarily of medium and long tasks, with sporadic short tasks also occurring. Scenarios marked *R* are read-dominated, consisting predominantly of tasks executing read-only transactions, *W* are write-dominated, consisting mostly of tasks executing read/write transactions, and *R/W* scenarios are balanced. Comparing the performance of the four frameworks using a varied set of scenarios allows us to evaluate in a realistic environment and draw conclusions from their performance.

In scenario *Small R* (Fig. 7a), we see that the increase in contention driven by a growing number of clients impacts Atomic RMI least, whereas the impact on FGL and HyFlow2 is significantly more visible. HyFlow2 and FGL start off with an advantage in performance, with throughput as high as 250 transactions per second, until the number of clients reaches 200, when the performance suddenly degrades and stabilizes between 75 and 100 transactions per second for client counts larger than 300. In contrast, Atomic RMI retains a constant throughput between 100 and 125 transactions per second regardless of scale. We see, however, that the aforementioned frameworks manage to scale in the 300–1200 client range with reasonable throughput, and retain an advantage of a sequential execution of at least 25% and up to 150%. The benchmark then shows that for specific sets of parameters and a particular workload Atomic RMI maintains better scalability, whereas solutions like FGL and HyFlow2 have a definite edge in lower contention environments.



(a) Throughput with message length at 4 words.

(b) Throughput with message length at 8 words.

(c) Throughput with message length at 12 words.

(d) Throughput with message length at 16 words.

Figure 5: Impact of message length on throughput.

In scenario *Small R/W* (Fig. 7b) we see a general decrease in throughput, as a large contingent of read/write operations are introduced into the benchmark. These shape the transactional characteristic (e.g. for all frameworks there is a gradual degradation in performance; see Fig. 7b and Fig. 7c). Due to its operation-type agnostic concurrency control method and an automated method of early release, Atomic RMI handles the increase in write operations better than FGL and HyFlow until the contention increases significantly with the introduction at 800 clients, at which point all frameworks tend to converge and achieve a near-sequential performance (up to 30 transactions per second for Atomic RMI vs 20 transaction per second for GLock). Even though FGL is also pessimistic and has early release, releasing objects early incurs additional network messages in that scheme (since a lock has to be signaled), which, in turn, impacts performance in a highly saturated network such as the one in this scenario. On the other hand, a large number of write operations and growing contention cause HyFlow2 to achieve a significant abort rate and a high retry rate (see below), causing the transactions to often waste work and incur additional network traffic as data-flow objects are supplied to transactions. The results show that pessimistic distributed TM may be favorable to optimistic distributed TM in a high contention write-dominated distributed system.

The *Small W* scenario (Fig. 7c) is analogous to *Small R/W*, showing similar performance characteristics, but between 50% and 75% decreased. It is plainly visible that for relatively low numbers of clients (around 600) all frameworks tend to saturate and perform the same as or not much better than a sequential evaluation.

The *Large R* scenario (Fig. 7d) shows similar behavior to *Small R*, but longer transactions have an impact on the performance of all four frameworks, that is especially visible in the 1–400 client range when the benchmark has relatively low contention. Note that transactional throughput values for all frameworks are nevertheless much lower than in *Small R*, since the transactions are much longer on average. At the

Task type	Standard	Small R	Small R/W	Small W
term search	0.25	0.30	0.19	0.02
interaction search	0.20	0.30	0.19	0.02
send unicast	0.06	0.02	0.19	0.44
send multicast	0.04	0.02	0.02	0.02
batch import	0.04	0.02	0.02	0.02
clear inbox	0.06	0.02	0.18	0.44
association level	0.20	0.30	0.19	0.02
indexing	0.15	0.02	0.02	0.02
Description	80-20% read-to-write ratio, not biased towards either long or short transactions	Mainly short, read-only transactions (90% of all tasks)	Short transactions, both read-only and writing (94% of tasks in the scenario)	Write-heavy, but short transactions (88% composition)

Task type	Large R	Large R/W	Large W
term search	0.44	0.19	0.02
interaction search	0.02	0.02	0.02
send unicast	0.02	0.02	0.02
send multicast	0.02	0.19	0.30
batch import	0.02	0.18	0.30
clear inbox	0.02	0.19	0.30
association level	0.02	0.02	0.02
indexing	0.44	0.19	0.02
Description	Long and read-only transactions (88% of all tasks)	Both read-only and writing, long-lasting transactions (94% composition)	Long and write-heavy transactions (90% composition)

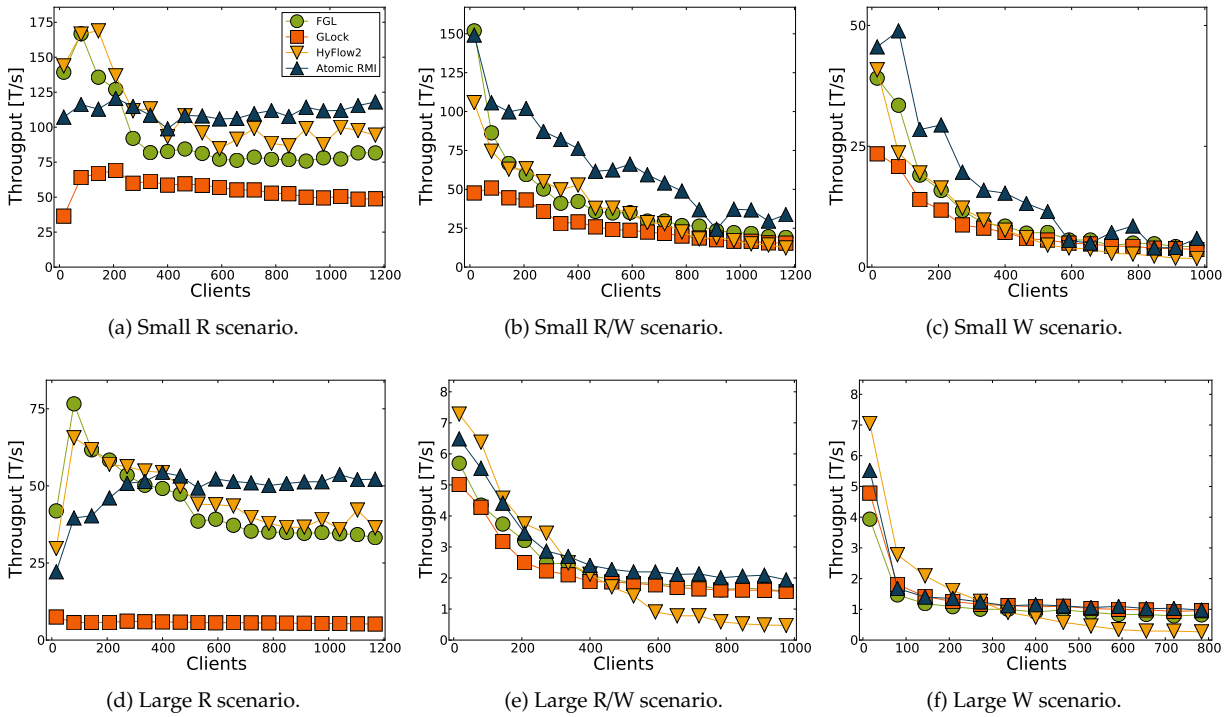
Figure 6: Descriptions of evaluation scenarios and task type probabilities in each of them.

outset, FGL, HyFlow2, and Atomic RMI all achieve comparatively low throughput, which increases as new clients are added. We attribute this to the initial undersaturation of the system, which means that adding new clients only increases tasks executed in parallel, without increasing the incidence of conflicts. Hence, Atomic RMI gradually increases throughput until it reaches 50 transactions per second at the 400 client mark. On the other hand, the throughput of HyFlow2 and FGL increases more rapidly with the introduction of new clients and peaks with around 100 clients at 70–75 transactions per second. This, however saturates both frameworks, and any further added clients cause a decline in performance, until both frameworks stabilize around the 600 client mark at about 30–40 transactions per second. Hence, as in *Small R*, there is range of load for which optimistic TM or a more subtle locking solution is preferable to pessimistic TM, but the distinction is much more pronounced. On the other hand, once contention reaches a certain threshold, pessimistic TM gains a stable advantage over the other two types of frameworks. Thus, the choice of paradigm must be tailored to the application and its workload, and Helenos provides important clues as to how this adjustment should be made.

When large read-write tasks are introduced in both *Large W* (Fig. 7f) as well as *Large R/W* (Fig. 7e), they throttle the throughput of all frameworks, since they have long execution times, as well as large R/W sets, and therefore typically cannot be executed in parallel to other tasks. Hence, throughput is below 8 transactions per second even for low contention situations. Furthermore, all frameworks perform similarly to GLock with Atomic RMI performing marginally better, FGL performing the same, and HyFlow2 performing notably worse. The performance of optimistic TM below GLock is attributable to the tendency of the workload to generate a large incidence of conflicts, leading to a high abort rate (52.5% for *Large R/W* and 68.8% for *Large W*) and a hight retry rate (211 retries per commit in *Large R/W* and 803 for *Large W*).

4.5. Aborts

We show the abort and retry rates for Fig. 8. The figure shows the results for HyFlow2, whereas the remaining frameworks had a consistent abort rate of 0 (i.e. no aborts at all) and a retry rate of 1 (i.e. one



(a) Small R scenario. (b) Small R/W scenario. (c) Small W scenario.

(d) Large R scenario. (e) Large R/W scenario. (f) Large W scenario.

Figure 7: Workload configurations.

attempt to execute a transaction per commit). The results for HyFlow2 show that the abort rate of that implementation tends to fluctuate around a certain stable level, despite increasing contention with the number of clients. However, the retry rate shows, that transactions that abort tend to abort more often on average as the contention increases, leading to a high number of aborts total.

4.6. Mean Flow Time

The benchmark allows to perform other measurements, including mean flow time. We show an example of that measurement in Fig. 9 for the standard scenario. Mean flow time indicates, that whereas the analysis of throughputs prefer Atomic RMI over FGL and HyFlow2 in the standard scenario, the average time between a transaction's start and commitment is lower in FGL and HyFlow2. Hence, for systems, where short response time is important (e.g. from the user experience perspective), those systems are preferable.

4.7. Discussion

The evaluation shows that the benchmark provides a wide range of articulation through its parameter selection and the composition of the workload. However, despite the achieved configurability, the benchmark remains rooted in a real-life application and reflects the complexity of the implementation of such a system, which differenciates it from microbenchmarks. This connection to reality makes the conclusions drawn on the basis of the results more meaningful and applicable to practical use cases.

For example, given a specific application like a distributed data warehouse for use in analytics and decision support, where it is characteristic to expect predominantly read-only transactions that aggregate data from various sources, the evaluation data gathered in the evaluation indicates that in this particular use-case pessimistic TM would perform better on average than other types of concurrency control schemes (Fig. 7d). However, we also note that this is the case only if the contention is fairly high. To the contrary, as the number of clients depreciates, an optimistic TM is the better choice to capitalize on low overhead. This data is therefore immediatelly useful to an architect involved in the design of such an application.

# Clients	Abort ratio	Retry rate
16	11.8%	1.255
80	22.4%	2.900
144	17.7%	4.516
208	26.1%	6.044
272	21.8%	7.744
336	30.2%	8.625
400	23.1%	10.278
464	21.5%	12.903
528	30.9%	13.909
592	29.5%	16.805
656	30.1%	18.522
720	23.3%	20.657
784	22.5%	23.772
848	21.8%	26.812
912	29.0%	27.750
976	25.9%	31.424
1040	26.0%	34.406
1104	21.8%	36.263
1168	26.8%	38.932

Figure 8: HyFlow2 abort ratio and retry rate for the standard scenario.

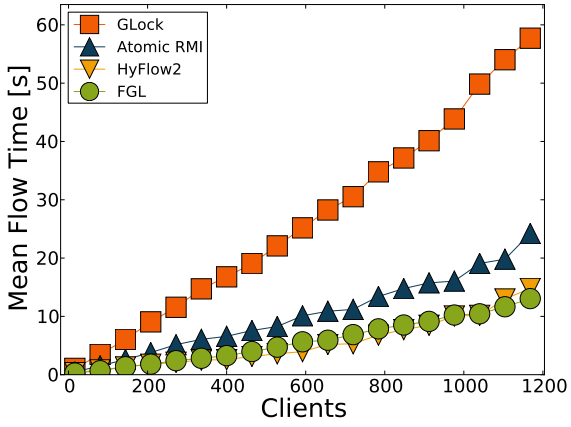


Figure 9: Mean flow time for the standard scenario.

Another example is that of a geographically distributed system of sensors periodically checked by a network of servers, that use transactions to maintain consistency. Here, a large quantity of small read operations is expected as in Fig. 7a. The case is similar to the one above, but while optimistic TM is significantly more performant in low contention, its throughput degrades in high contention, while pessimistic TM performs in a more stable fashion. This allows the architect to select either distributed TM based on the variability of the server network.

On the contrary to both previous system types, the data from Fig. 7f shows that when designing a datastore used primarily for archivization, i.e. one with a large amount of long and write-oriented transactions, one need not bother with sophisticated synchronization schemes, since the number of conflicts involved forces near-serial execution in any case.

Alternatively, developers of distributed TM systems can use the data provided by Helenos to analyze the systems and easily find their strong and weak points, whether inherent or ammeanable. For instance, as presented above HyFlow2 shows some room for improvement when subjected to write-heavy workloads with high retry rates. Knowing this, the developers of HyFlow2 can tune their contention management. They can still use Helenos for this purpose by creating some specialised scenarios for detailed examination of very specific situations and workloads. This can lead to the identification of the source of such behavior in the distributed TM design. As for Atomic RMI, its notably worse performance in terms of MFT comparing to the HyFlow2 can be at least partially attributed to the fact that this system does not distinguish between read and write operations. Introducing such distinction in the internal workings of this system and maximizing the parallelization of read operations can lead to closing the MFT performance gap between said TM systems, as shown in [29].

5. Conclusions

We presented Helenos, a benchmark for performing comprehensive evaluations and comparisons of distributed TM systems. It provides depth of analysis by its implementation of a complex use case. It also allows for analyzing many aspects of the memory using a wide variety of parameters and metrics. Finally,

it is based on a real application, meaning that the results of evaluations are more likely to be reflected in practice when the TM system is deployed in the real world. As such, it fills the vacuum caused by a lack of distributed TM benchmarks other than microbenchmarks. It should be especially pointed out that Helenos is particularly designed and built with evaluation of distributed TMs in mind. This characteristics make it a much anticipated tool for many distributed TM researchers and can contribute to rising the overall quality of distributed TM evaluation in their future work.

We showed that Helenos provides important and meaningful data regarding distributed TM system performance in near real-life situation. Helenos, in contrast to other benchmarks that have been used for testing distributed TMs, is built particularly for evaluating distributed TMs, and allows to do so with specific distributed applications in mind. Users of Helenos can use several, different parameters to create specialized scenarios to suit their specific needs and use them to perform extenvie and complex evaluation.

In our future work we wish to create a full suite of benchmarks for testing distributed TMs, to fulfill the requirement for broad evaluations, as well as deep ones. That is, in order to be able to evaluate distributed TMs in different applications from different fields. To that end we plan on distributing some applications from the STAMP benchmark suite, as well as to find additional applications that are specific to distributed systems. The work on these is already underway and the early results are promising.

Acknowledgments. The project was funded from National Science Centre funds granted by decision No. DEC-2012/07/B/ST6/01230.

References

[1] M. Herlihy, J. E. B. Moss, Transactional Memory: Architectural Support for Lock-free Data Structures, in: Proceedings of ISCA'93: the 20th International Symposium on Computer Architecture, 1993, pp. 289–300.

[2] V. Aslot, M. Domeika, R. Eigenmann, G. Gaertner, W. B. Jones, B. Parady, SPEComp: A new benchmark suite for measuring parallel computer performance, in: Proceedings of WOMPAT'01: the International Workshop on OpenMP Applications and Tools: OpenMP Shared Memory Parallel Programming, 2001.

[3] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, A. Gupta, The SPLASH-2 programs: characterization and methodological considerations, in: Proceedings of ISCA'95: the 22nd Annual International Symposium on Computer Architecture, 1995.

[4] R. Guerraoui, M. Kapałka, J. Vitek, STMBench7: A benchmark for software transactional memory, in: Proceedings of EuroSys'07: the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems, 2007.

[5] M. Ansari, C. Kotselidis, K. Jarvis, M. Luján, C. Kirkham, I. Watson, Lee-TM: A non-trivial benchmark for transactional memory, in: Proceedings of ICA3PP'08: the 8th International Conference on Algorithms and Architectures for Parallel Processing, 2008.

[6] C. C. Minh, J. Chung, C. Kozyrakis, K. Olukotun, STAMP: Stanford Transactional Applications for Multi-Processing, in: Proceedings of IISWC'08: the IEEE International Symposium on Workload Characterization, 2008.

[7] D. Peng, F. Dabek, Large-scale incremental processing using distributed transactions and notifications, in: Proceedings of OSDI '10: the 9th USENIX Symposium on Operating Systems Design and Implementation, 2010.

[8] J. C. Corbett, et al., Spanner: Google's Globally-Distributed Database, in: Proceedings of OSDI'12: the 10th USENIX Symposium on Operating Systems Design and Implementation, 2012.

[9] R. L. Bocchino, V. S. Adve, B. L. Chamberlain, Software Transactional Memory for Large Scale Clusters, in: Proceedings of PPoPP'08: the 13th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2008.

[10] M. Couceiro, P. Romano, N. Carvalho, L. Rodrigues, D2STM: Dependable Distributed Software Transactional Memory, in: Proceedings of PRDC'13: the 15th IEEE Pacific Rim International Symposium on Dependable Computing, 2009.

[11] T. Kobus, M. Kokociński, P. T. Wojciechowski, Hybrid replication: State-machine-based and deferred-update replication schemes combined, in: Proceedings of ICDCS'13: the 33rd International Conference on Distributed Computing Systems, 2013.

[12] S. Hirve, R. Palmieri, B. Ravindran, HiperTM: High Performance, Fault-Tolerant Transactional Memory, in: Proceedings of ICDCN'14: the 15th International Conference on Distributed Computing and Networking, 2014.

[13] M. M. Saad, B. Ravindran, HyFlow: A High Performance Distributed Transactional Memory Framework, in: Proceedings of HPDC '11: the 20th International Symposium on High Performance Distributed Computing, 2011.

[14] A. Turcu, B. Ravindran, R. Palmieri, HyFlow2: A High Performance Distributed Transactional Memory Framework in Scala, in: Proceedings of PPPJ'13: the 10th International Conference on Principles and Practices of Programming on JAVA platform: virtual machines, languages, and tools, 2013.

[15] K. Siek, P. T. Wojciechowski, Atomic RMI: a Distributed Transactional Memory Framework, in: Proceedings of HLPP'14: the 7th International Symposium on High-level Parallel Programming and Applications, 2014, pp. 73–94.

[16] K. Siek, P. T. Wojciechowski, Atomic RMI: a Distributed Transactional Memory Framework, International Journal of Parallel Programming 44 (3) (2015) 598–619. doi:10.1007/s10766-015-0361-x.

[17] Helenos Benchmark, `https://gitlab.cs.put.poznan.pl/helenos` (2016).

[18] W. Ruan, Y. Liu, M. Spear, Stamp need not be considered harmful, in: Proceedings of TRANSACT'14: the 9th ACM SIGPLAN Workshop on Transactional Computing, 2014.

[19] S. Mishra, A. Turcu, R. Palmieri, B. Ravindran, HyflowCPP: A distributed transactional memory framework for C++ (Aug. 2013).

[20] S. Hong, T. Oguntebi, J. Casper, N. Bronson, C. Kozyrakis, K. Olukotun, Eigenbench: A simple exploration tool for orthogonal tm characteristics, in: Proceedings of IISWC'10: the IEEE International Symposium on Workload Characterization, 2010.
[21] F. Zyulkyarov, V. Gajinov, O. S. Unsal, A. Cristal, E. Ayguadé, T. Harris, M. Valero, Atomic Quake: using transactional memory in an interactive multiplayer game server, in: Proceedings of PPoPP'09: the 14th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2009.
[22] A. Turcu, R. Palmieri, B. Ravindran, S. Hirve, Automated data partitioning for highly scalable and strongly consistent transactions, IEEE Transactions on Parallel and Distributed Systems 27 (1) (2016) 106–118.
[23] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, R. Sears, Benchmarking cloud serving systems with YCSB, in: Proceedings of the 1st ACM symposium on Cloud computing, ACM, 2010, pp. 143–154.
[24] A. Lakshman, P. Malik, Cassandra: a decentralized structured storage system, ACM SIGOPS Operating Systems Review 44 (2010) 25–40.
[25] M. M. Saad, R. B., Transactional forwarding: Supporting highly-concurrent STM in asynchronous distributed systems, in: Proceedings of SBAC-PAD'12: the 24th IEEE International Symposium on Computer Architecture and High Performance Computing, 2012.
[26] R. Guerraoui, M. Kapałka, Principles of Transactional Memory, Morgan & Claypool, 2010.
[27] P. T. Wojciechowski, Language Design for Atomicity, Declarative Synchronization, and Dynamic Update in Communicating Systems, Publishing House of Poznań University of Technology, 2007.
[28] K. Siek, P. T. Wojciechowski, Brief announcement: Relaxing opacity in pessimistic transactional memory, in: Proceedings of DISC'14: the 28th International Symposium on Distributed Computing, 2014.
[29] K. Siek, P. T. Wojciechowski, Atomic RMI 2: Highly parallel pessimistic distributed transactional memory, IEEE Transactions on Parallel and Distributed Systems, arXiv:1606.03928 [cs.DC] (submitted).