FlexSync: An aspect-oriented approach to Java synchronization

Charles Zhang
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
charlesz@cse.ust.hk

Abstract

Designers of concurrent programs are faced with many choices of synchronization mechanisms, among which clear functional trade-offs exist. Making synchronization customizable is highly desirable as different deployment scenarios of the same program often prioritize synchronization choices differently. Unfortunately, such customizations cannot be accomplished in the conventional non-modular implementation of synchronization. To enable customizability, we present FlexSync, an aspect oriented synchronization library, to enable the modular reasoning and the declarative specification of synchronization. Complex Java systems can simultaneously work with multiple synchronization mechanisms without any code changes. The FlexSync load-time weaver performs deployment time optimizations and ensures these synchronization mechanisms consistently interact with each other and with the core system. We evaluated FlexSync on commercially used complex Java systems and observed significant speedups as a result of the deployment-specific customization.

1 Introduction

In Java programs, synchronization is commonly referred to as the coordination of multiple threads in accessing shared program states. As concurrency becomes a common programming practice in the multi-core era, software designers are faced with many choices of synchronization mechanisms such as the use of locks, atomic blocks [7, 8], and, more recently, software transactional memory [10, 21]. For their distinctive operational differences, clear functional trade-offs exist among these synchronization mechanisms. This is problematic for building general-purpose and reusable Java systems as, in conventional approaches, synchronization mechanisms are "hardwired" to the application logic through the use of library APIs or specialized language constructs. At the same time, choosing the most appropriate mechanism is increasingly a decision about how reusable systems are being integrated in diversified composition contexts. Let us further elucidate this issue through a simple example.

Our example examines a general-purpose data structure, `Buffer`, used by multiple threads in a concurrent program. Each thread makes accessor calls to store and to retrieve data from the buffer, under the constraint that these accessor calls (`get` or `set`) can only proceed when the state of the `Buffer` is valid (`full` or `empty`). Inconsistency can happen if, for example, thread A empties the buffer after thread B verifies that the buffer has data and before it retrieves the data. We implemented three versions of the `Buffer` example using the following popular synchronization facilities: the Java monitors through the `synchronized` keyword (`lock`), the block-level atomicity support (BA) using two-phase-locking (2PL) as in [1], and the software transactional memory library(`stm`) exemplified by dstm2 [10] (Please refer to Section 2 for more introduction on 2PL-based BA and dstm2). In Figure 1[1], we plot against the number of concurrent threads the time taken by each version to complete a batch of identical buffer operations . For the lock version, we count the number of successful operations and perform a retry if any inconsistency is detected. The BA and STM versions produce no inconsistencies. One can easily observe that the lock-based approach has the fastest response time, whereas the BA implementation is slightly slower. The performance of the dstm2 version experiences significant fluctuations. It can be as fast as the BA approach or 5-6 times slower.

The example shows that the choice of synchronization mechanism for `Buffer` is dependent on what matters the most to the domain of its application. For instance, the use of locks is preferred if one desires high performance and is capable of tolerating inconsistencies. The STM approaches are appealing to domains requiring transactional semantics on the shared states and not sensitive to the fluctuations of the processing time. If only requiring atom-

[1] All measurements are collected on a dual-core Linux workstation with 4GB of physical memory. The number of total threads ranges from 10 to 1000. Each point is taken as the shortest time of five runs.

ICSE'09, May 16–24, 2009, Vancouver, Canada
978-1-4244-3452-7/09/$25.00

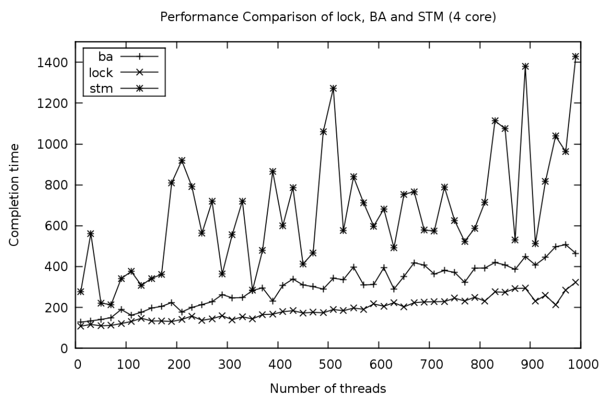

Figure 1. A comparison of response time among locks, BA and STM

icity, applications would prefer the lock-based atomicity support, which enlarges critical regions to trade the degree of concurrency with the consistency of states. Therefore, the buffer code, if it were to be used as a general-purpose building lock of other concurrent application, cannot be hardwired with any particular synchronization mechanism afore-listed. The practice of client-side locking, such as the `synchronizedMap` method of the `Collection` class, is effective in treating this problem for types representing data structures. However, we only use buffer as a problem illustration. Our work considers complex reusable Java programs, many are concurrent themselves.

We use the buffer example to illustrate that, because architectural compositions are difficult to predict, there are no clear winners among different synchronization choices when designing individual participating components. However, synchronization, once coded, is conventionally a non-customizable feature because it requires systematic code-level bindings to either language keywords or library APIs. This inflexibility can incur as much as 40% performance overhead, as our case studies of real commercial middleware systems show. This problem will exacerbate drastically because the degree of reuse and integration will increase dramatically in future software systems [17].

To modularize synchronization, we designed and evaluated FlexSync, including both an aspect-oriented library and a load-time weaver, to enable the modular reasoning of synchronization and the code-level separation of its mechanisms in reusable Java programs. Since synchronization is a classic example of crosscutting concerns [12], we are not the first [15, 5] to attempt a solution in the aspect oriented programming (AOP) paradigm. The essential differentiator for our approach is that we fully preserve how synchronization is conventionally reasoned in design and only change how it is expressed in code. Our approach is based on the observation that, conventionally, the expression of synchronization intentions, i.e., declaring regions of the program logic that require special synchronization attention, is an implicit outcome of the direct use of specific synchronization mechanisms, i.e., library APIs or language keywords. If these intentions have explicit and well-defined program structures, they can be reasoned and manipulated by meta-programming such as AOP [12] as to *externalize* reusable *feature interactions* between synchronization and the core system. Therefore, the design of FlexSync library APIs emphasizes on the ease of "picking out" the intentions where common interaction logic, encapsulated by the library, can be automatically applied. The FlexSync aspect weaver, an extension to the AspectJ aspect weaver, use static analysis, such as the control-flow analysis and the escape analysis, to automatically reason about the global composition and the interaction of synchronization mechanisms. With FlexSync, the synchronization code is modular and lives separately from the operational code. We show that, through FlexSync, sophisticated Java systems can simultaneously work with multiple synchronization mechanisms of very different genres. The flexibility and the deployment time optimizations, made possible by using FlexSync, can significantly improve the performance for large complex systems.

We make the following contributions in this paper:

1. We first present the concept of the explicit separation of intention and mechanism in the context of synchronization design. We empirically show that such separation can be achieved for large-scale and complex Java systems.

2. We present the FlexSync aspect synchronization library, which encapsulates patterns of interactions between Java code and the synchronization mechanisms and expose these patterns through a programming process called "tagging". We explain how the tagging process can attach different synchronization mechanisms onto the same code structure.

3. We present the FlexSync load-time synchronization weaver which supports the global reasoning of synchronization mechanisms in the scenarios of unanticipated composition of reusable systems.

4. We present a thorough evaluation of FlexSync-based synchronization implementations, covering the programming effort and its functional characteristics. We contribute[2] the source of the FlexSync library and the systems we experiment with for the interested readers to inspect and to perform further evaluations.

The rest of the paper is organized as follows: Section 2 introduces atomicity and the dstm2 implementation of software transactional memory; Section 3 presents the design methodology embodied in FlexSync. Section 4 evaluates FlexSync through standard benchmarks and case studies.

[2]The FlexSync Project. URL:`http://www.cse.ust.hk/~charlesz/flexsync`

2 Background

Block-level atomicity

In the presence of multiple threads, the block-level atomicity means a group of program executions, scoped lexically within a block, is to be carried out serially without the interference from the inter-leavings of other threads. In Java systems where objects are dynamically allocated, we use a two-phase-lock mechanism to acquire the associated locks of all dynamic objects in the control flow of the atomic block. These locks are released after the atomic block exists. Our implementation is based on the `cflow` constructs of AspectJ. Please refer to our source release for the details of the implementation.

Software transactional memory

Software transactional memory (STM) provides runtime infrastructures to keep track of the reads and writes to the shared program states. In STM, a *collision* can happen when thread B performs writes on the shared data after they are read by thread A. Then, the shared states are to be *rolled back* and the operation is re-executed. This can cause the large performance fluctuations because the chance of collision is subjective to the number of threads and the thread scheduling that can be non-deterministic. The particular STM library used in this research, dstm2, makes copies of the shared states to support rollbacks. This technique is reported to have the fastest runtime performance [10]. STM offers programmers a simpler concurrency control mechanism compared to the direct use of locks. In our application of dstm2, we replaced the automatic state copy capability in the original implementation with a callback method requiring the manual implementations. This is because many reads or writes in the systems that we have experimented with are not performed by "setters" and "getters" as required by the original scheme.

3 FlexSync: the modular and the global reasoning of synchronization

Following the definitions in [13], the general design goal of Flexsync is to first foster the modular reasoning of synchronization by using the FlexSync API to explicitly express how synchronization mechanisms interact with the operational logic. In addition, we address the global reasoning in the context of unanticipated program compositions, through the Flexsync loadtime weaver. The rest of the section present how we achieve these design goals in detail.

3.1 The separation of intention and mechanism

Synchronization is typically programmed as block-level lexical scopes demarcated by either language keywords, such as `synchronized` or `atomic`, or library calls, such as the `lock/unlock` pairs of `Lock` objects. We say that these *scopes* represent the synchronization *intention*s of the developers, which identify the code regions requiring special synchronization treatments. Meanwhile, we refer to language keywords or API calls used in the demarcations as the *mechanism*s, concerning the specific choices of synchronization facilities. These two concepts are usually undistinguished in conventional approaches.

In our current design, the separation of intention and mechanism is achieved by assuming that synchronization regions are method-like: the data flow in these regions can be re-expressed following the input/output model of a function. This is an effective way of making them structurally explicit. This assumption is certainly true for `synchronized` methods. In the case of synchronization blocks, we performed a study of if these blocks can be automatically transformed into methods using the *Extract method* facility of the Eclipse JDT refactoring library. We use an AST walker to retrieve the synchronization blocks and ask the JDT API to return the refactoring status. We studied four open source programs covering four types of servers in which concurrency is extensively used: RPC middleware (ORBacus[3]), JMS broker (OpenJMS[4]), web server (Jigsaw[5]), and database server (Derby[6]). In our study, we encountered three common causes of automatic refactoring failures: *early return* (ER), where a `return` statement is nested inside the block, *multiple variable assignment* (MV), where multiple local variables are written, and *branch selection* (BR), where the block resides in a branching block of either `if` or `switch`. Among these failures, the ER case can be generically treated with by setting a condition variable to true inside the refactored method and checking this condition after the method returns. It requires trivial source rewriting and, thus, can be automatically treated.

In Table 1, we report the size of each program, the usage statistics of synchronization, as well as the results of invoking the Eclipse refactoring APIs. Our observation is that ER accounts for the majority of the refactoring failures, and over 97% of synchronization blocks in all of the four programs can be automatically re-expressed using functions. The non-automatic blocks require the manual inspection and the code restructuring. However, they are small in number. Our study validates our design assumption that synchronization intentions can be enclosed by method boundaries.

[3]ORBacus: URL:`http://www.iona.com/orbacus`
[4]OpenJMS. URL:`http://openjms.sourceforge.net`
[5]Jigsaw. URL:`http://www.w3.org/Jigsaw`
[6]Derby URL:`http://db.apache.org/derby/`

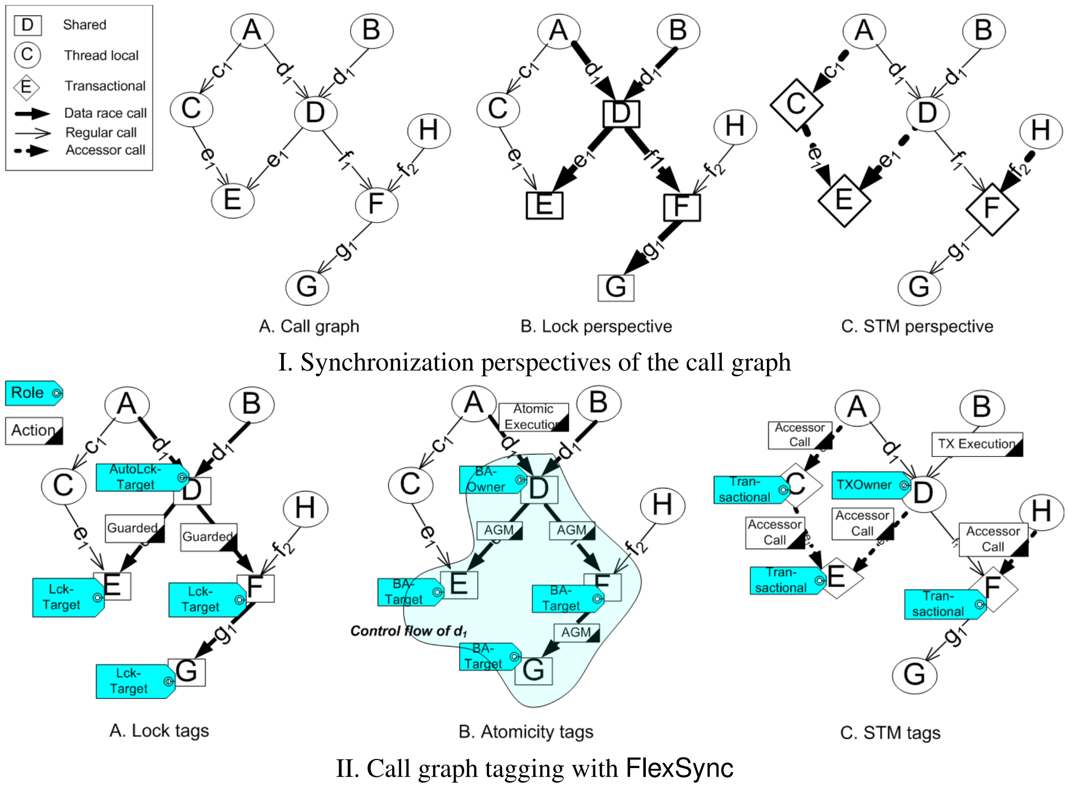

Figure 2. Perspectives and tagging

Program	Derby	Jigsaw	OpenJMS	ORBacus
Size	915320	160740	112968	189852
Synchronization usages				
Method	294	637	204	466
Block	604	149	400	262
Total	898	786	604	728
Method Extraction Failures				
ER	90	11	30	29
MV	18	2	9	3
BR	8	0	6	2
Degree	87.1%	99.7%	92.5%	95.3%
Degree with ER	97.1%	98.3%	97.5%	99.3%

Table 1. Summary of Server Synchronization Usage

3.2 The FlexSync synchronization library

Synchronization perspectives

The design of the FlexSync APIs is based on our observation that the interactions with different synchronization mechanisms can be projected as different interpretations of the same call graph. Suppose that Figure 2 (I-A) represents the call graph of a fictitious Java program. To program the lock-based approaches, including the atomicity support, we need to identify class types encapsulating shared program states as well as the method interfaces that could lead to data races. This interpretation is illustrated in Figure 2 (I-B). The use of STM is not generally concerned with the state sharing and data races. We need to instead identify transactional executions and the methods that cause reads or writes to data involved in the transactional executions. The corresponding representation of the original call graph as depicted in Figure 2 (I-C).

Synchronization specification

These different interpretations of the calling relationships are supported by FlexSync through a design process which we characterize as "tagging". There are two types of conceptual tags: the *role tag* operates at the class level for the declaration of what synchronization facilities to apply; the *action tag* operates at the method level for where the facilities are applied. Although we use the callgraph to explain the tagging process, the use of FlexSync does not require the knowledge of callgraphs. Role tags and action tags require only the local reasoning about a particular type. The tags for specific synchronization mechanisms are as follows:

Lock The tag AutoLckTarget identifies class types for which the execution of all of its methods are always protected by locks and, as default, by Java monitors. The tag LckTarget identifies types that are caller-synchronized. The methods susceptible to data races are tagged using Guarded. For the call graph given in Figure 2 (I-A), class type D is tagged as an AutoLckTarget as calls to all of its methods are unconditionally synchronized. E, F and G, are LckTargets, as they are selectively synchronized in the caller's lexical context. The tagged version of the call graph is presented in Figure 2 (II-A).

Block atomicity The tag BAOwner identifies the class types which the executions of one or more of its methods are atomic. We identify these methods with the Atomic-Execution tag. Classes having shared states, in this case, are identified with the BATarget tag. We use the AGM (atomic group member) tag on methods defined in BATargets if, first, they cause reads or writes to shared states, second, they are in the control flow of Atomic-Executions. In our example (Figure 2 (II-B)), class D has an atomic method d_1. The control flow of d_1 includes calls to methods e_1, f_2 and g_1. Therefore, these methods are tagged with AGM and the corresponding class types E, F and G with AtomTarget. Note that the use of AGM tags does not require any knowledge of the control flow information, and they are semantically identical to Guarded tags in the *Lock* scenario. We use a different tag to allow the FlexSync runtime automatically manage their relationships with the control flow of the atomic executions.

STM execution The tag Transactional identifies class types of which the states require transactional support. Their "accessor"[7] methods are identified with AccessorCall[8]. TXExecution identifies methods to be executed transactionally. In our example (Figure 2 (II-C)), AccessorCall identifies accessor methods c_1, e_1 and f_2. We thus identify types C, E, F as Transactional and the transactional method d_1 as TXExecution.

Implementation of tags

In FlexSync, the tagging process is translated into the programmatic mappings of Java interfaces and the AspectJ abstract pointcuts to the corresponding elements of Java programs. In the AspectJ nomenclature, the mapping of role tags is accomplished declaratively through the inter-type declarations(ITD) and the actions tags through "concretizing" the abstract pointcuts. Both techniques are commonly used in aspect library implementations [11, 9, 23]. Using the tags as the conceptual references, the FlexSync library encapsulates the reusable interaction logic for each mechanisms, which is explained as follows:

Lock 1. Protect the executions of the Guarded methods of the AutoLckTargets by the monitor of the corresponding AutoLckTarget instances. 2. Protect the call sites of the Guarded methods by the monitor of the callee instances of LckOwner.

Atomicity: 1. If a AGM is within the control flow of the AtomicExecution of the BAOwner, acquire the monitor lock of its corresponding BATarget and register the lock with the BAOwner (acquiring phase). 2. When a AtomicExecution completes, release the monitor locks of all BATargets registered with the BAOwner (releasing phase).

STM: 1. As Transactional instances initialize, set up their per-instance duplicates, as required by dstm2, to allow state rollbacks. 2. At the call sites of accessors, signal the dstm2 runtime to verify if the on-going transaction can proceed or must be aborted. 3. Repeat the TX execution of the TXOwner until the transaction successfully commits.

We emphasize the fact that the interaction logic can be well hidden behind our "tag" abstractions and implemented through AspectJ. This is a salient property of modular reasoning as pointed out in [13]. We will not bore the readers here with the implementation details and encourage the interested readers to download[9] a copy of the library for further references.

We now come back to the *Buffer* example presented in Section 1. The Buffer class contains four methods: setData, getData, isFull, and isEmpty. These methods are invoked by the doWork method of the class BufferUser. Before invoking the accessor methods, the method isFull or isEmpty is called to check the state of the buffer. The state is validated again inside the accessor methods for consistency checking. Figure 3 presents the AspectJ implementation of the three supported synchronization mechanisms through the FlexSync APIs. The use of the FlexSync tags are underlined. Despite its simplicity, we want to demonstrate the high degree of declarativeness in the implementation of synchronization enabled by FlexSync. As our evaluation in Section 4 shows, this property still holds for complex Java server systems. The implementation of the copyTo method is, however, non-declarative and enforced by the AspectJ compiler in the case of the STM support.

[7]Quotation here to entail that, in practice, not all methods that read or write the state of the object would lexically start with "set" or "get".

[8]The actual FlexSync APIs distinguish between *read* tags and *write* tags. We use a general name, AccessorCall, for the conciseness of the presentation.

[9]FlexSync. URL: http://www.cse.ust.hk/~charlesz/sync

Limitations

To use FlexSync, refactoring is still needed to transform blocks into methods. We are currently working on an automated solution to make the process transparent. The `wait/notify` semantics are also to be treated case by case, as they often intertwine with the application logic. We provide a replacement of `wait` by releasing the object lock in the case of BA and using an "abort→re-execution" sequence in the case of STM. The use of `wait`, however, will break the atomicity guarantees of BA in general as its Java semantic mandates the release of the monitor lock. In addition, the FlexSync-adaption of the dstm2 library requires FlexSync users to manually specify how program states are duplicated. From our experience, this manual process can be tedious. Our on-going work is trying to provide simplification solutions. Finally, our lock implementation does not handle the use of library-based locks, such as the `ReentrantLock` in the Java 5 library, that do not necessarily conform to the same lock/unlock interface and, hence, require new library code to be created. However, the dominating majority of lock uses in the Java programs that we have studied do not use library locks.

3.3 Global reasoning of tags

The design of synchronization using FlexSync allows a program to work with multiple synchronization mechanisms through configuration. However, when we integrate these programs to build complex systems, we must ensure the consistent and optimized interactions of locally specified synchronization mechanisms from the global perspective. In FlexSync, the global reasoning is carried out at the start-up time of Java programs by the FlexSync aspect weaver. Before the first class is loaded for execution, the weaver, as an extended AspectJ bytecode weaver, carries out static analysis over the bytecode of the entire system to check for inconsistencies and optimization opportunities. It also carries out load-time weaving to specifically treat reflective loading, a common way of Java composition. We now present these capabilities in detail.

Lock optimization

A synchronization mechanism is *redundant* if a type uses lock-based tags is never shared among threads in a particular compositional scenario. Such scenarios are often difficult to anticipate from the perspectives of individual programs. The FlexSync load-time weaver first leverage the techniques of escape analysis [19, 4] to detect, on the per-composition basis, the sharing status for every "tagged" type. Conventional escape analysis techniques reason about object instances. Our approach is more conservative as we define that a type escapes if any of its instances escape. Our

```
public aspect LckBuffer extends Lock {
      declare parents: Buffer implements AutoLockOwner;
}
```

A. Thread-safe

```
public aspect ATBuffer extends Atomicity {
      declare parents: Buffer implements BATarget;

      public pointcut agm():
            call(* Buffer.setData(..))||
            call(* Buffer.getData(..))||
            call(* Buffer.isFull(..))||
            call(* Buffer.isEmpty(..));

      public pointcut atomicexecution():
            execution(* BufferUser.doWork());
}
```

B. Atomic

```
public aspect STMBuffer extends ASTM {
      declare parents: Buffer implements Transactional;
      declare parents: BufferUser implements TXOwner;

      public pointcut txexecution():
            execution(* BufferUser.doWork());

      public pointcut reads():
            execution(* Buffer.getData(..))||
            execution(* Buffer.is*(..));

      public pointcut writes():
            execution(* Buffer.setData(..));

      public void Buffer.copyTo(Buffer copy){
            //copy the states. Code omitted
      }
}
```

C. STM

Figure 3. Implementing synchronization with FlexSync

implementation is based on the Indus project[10], which is an enhanced version of the equivalence-class-based analysis [19, 18]. The escape analysis goes through all classes seen on the Java class path, and the results are stored in a hash table maintained by the FlexSync weaver to decided whether the weaving is necessary, if a lock-based synchronization mechanism is to be used.

We claim no significant extensions to the Indus escape analysis algorithm other than the treatment of reflective class loading. The reflective class loading is referred to as instantiating a class by its lexical name through the `Class.forName` Java API combined with a type cast. In this case, we simply look up all subtypes of the type used in the type cast and store the results in the mapping table of the weaver. The weaver will have the concrete information

[10] Indus project. URL:`http://indus.projects.cis.ksu.edu/`

after the actual subtype is loaded.

Consistency checking

Two synchronization mechanisms can be *inconsistent* if, for instance, a type tagged with `BATarget` is not in the control flow of any atomic blocks, which causes incorrect holding of locks, or, a `TXTarget` is also tagged separately with `LckTarget`, which violates the assumption in STM of no direct use of locks. To prohibit erroneous usage of tags, FlexSync uses consistency rules presented in Table 2[11] as set predicates. These rules enforces the following usages of tags: 1. If a type is declared to use many mechanisms, a default choice needs to be specified; 2. The support of atomic or transaction behaviors ("targets") cannot be specified without the matching atomic or transactional executions ("owners") (rule 2); 3. A type cannot be mapped to more than one "owner" or "target" roles in any composition scenarios (rule 3 and 4); 4. As a current implementation limitation, no nested atomic regions and transactions are permitted (from rule 5 and 6); 5. No explicit use of locks for objects that are in the control flow of STM-based ones (rule 1).

To perform the consistency checking, we build a simple control-flow graph consisting of only the tagged types from the analysis information collected by the escape analysis of the Indus framework. Since the number of nodes in these graphs is generally small (in the order of hundreds), a simple depth-first graph traversal can accomplish the checking of the rules very quickly. We leave the guarantee of the completeness and the soundness of these rules to our future research. We currently rely on FlexSync to perform runtime checks in case our weaver fails to detect usage inconsistencies.

4 Evaluation

The assessment of FlexSync consists of two studies, one related to the programming effort of using FlexSync in modularizing synchronization, and the other to the functional characteristics of FlexSync. The target systems of study is specjbb2005[12], a popular benchmark for transactional enterprise Java applications, OpenJMS[13], an open source implementation of the JMS 1.1 specification, and ORBacus[14], an open source commercial implementation of the CORBA 2.4 specification. To experiment with software compositions, we switched the RPC engine of OpenJMS from Java RMI to ORBacus. This is a fully functional replacement as verified by the Sonic JMS benchmark[15].

4.1 Programming with FlexSync

In this evaluation, we want to first find out if FlexSync is capable of supporting synchronization in commercially used complex systems. We also want to study the programming characteristics of implementing synchronization in FlexSync APIs. We first remove the monitor-based synchronization from the original implementations. We then perform necessary refactorings to enclose synchronization blocks within method definitions. We use FlexSync to support the locking, block-level atomicity, and the dstm2-based STM. We cannot use STM on ORBacus because ORBacus involves network I/O operations which cannot have *rollback* semantics. This is a typical limitation of STM in general. The OpenJMS server poses the same limitations, however, since we use ORBacus as its remote procedure call (RPC) engine, the non-RPC part of the OpenJMS can have STM-compatible behaviors. This is part of our composition case study which will be presented shortly.

The total size of the FlexSync library is less than 60KB of Java bytecode. The sizes of specjbb2005, ORBacus and OpenJMS are listed tn Table 1. In Table 3, we quantify four aspects of the FlexSync-based implementation for each of the studied system: number of modules, i.e., class types, where synchronization is implemented in the original application (*orig*); the number of modules, i.e., aspects, for the FlexSync-based implementation (*flex*); the total size, in lines of code (LOC), of declarative(*dec*) and non-declarative(*ND*) portion in the FlexSync user code; and the non-declarative portion of STM implementation(ND'). We define the declarative degree, α, as the ratio between the declarative code and the total size of synchronization implementation. α' is computed without the STM code. Here we treat the STM as a special case because its non-declarative code almost exclusively involves the state duplications, i.e., copying class variables. We are working on automating this process.

The FlexSync-based approach fully exhibits the benefit of the aspect oriented approach as, the synchronization implementations is not only the much modular (9 modules in FlexSync vs. 37 modules in the original implementations), the task of programming is also simpler for two reasons: 1. the coding effort is largely declarative in nature, meaning interactions patterns are widely used; 2. the "owner/target" relationship, which is latent and spread-out in conventional implementations, is explicit and local in FlexSync-based approaches, allowing easier reasoning

[11]Note that the control flow definition($cflow$) is the conservative control flow computed statically from the bytecode where all possible branches of the call flow are explored

[12]Specjbb. URL: `http://www.spec.org/jbb2005/`

[13]OpenJMS URL: `http://openjms.sourceforge.net`

[14]http://www.iona.com/orbacus

[15]Sonic JMS Benchmark. URL:`http://www.sonicsoftware.com/products/sonicmq/performance_benchmarking/index.ssp`

Definitions	*Consistency rules*
1. τ : the type variable. T: the set of all types.	1. $\lvert tag(\tau)\rvert > 1 \wedge pref(\tau) := \emptyset$
2. $tag(\tau)$: the set of tags on τ	2. $cflow(\tau) \cap owners := \emptyset \wedge pref(\tau) \in \{TXTarget, BATarget\}$
3. $pref(\tau)$: tag indicated by design as the preferred tag of τ	3. $\lvert tag(\tau)\rvert >= 2 \wedge tag(\tau) \subseteq owners$
4. $cflow(\tau)$: the set of tags in control flows above τ	4. $\lvert tag(\tau)\rvert >= 2 \wedge tag(\tau) \subseteq targets$
5. $cover(\tau_i) := \{tag(\tau_k) \mid \tau_k \neq \tau_i, cflow(\tau_k) \cup tag(\tau_k) \equiv cflow(\tau_i)\}$	5. $cover(\tau) := \{BAOwner\} \wedge pref(\tau) \cap \{BATarget\} = \emptyset$
	6. $cover(\tau) := \{TXOwner\} \wedge pref(\tau) \cap \{TXTarget\} = \emptyset$
7.$targets := \{BATarget, TXTarget, LockTarget, AutoLockTarget\}$	7. $cflow(\tau) \cap \{TXOwner, TXTarget\} \neq \emptyset \wedge tag(\tau) \cap \{LckTarget, BATarget, AutoLckTarget\} \neq \emptyset$
8.$owners := \{BAOwner, TXOwner\}$	

Table 2. Consistency checking rules

and modification. The declarative degree of ORBacus is low due to the treatments of `wait/notify` semantics, a limitation we discussed previously.

App	*Orig*	*flex*	*dec*	*ND*	*ND'*	α	α'
specjbb	9	3	130	178	168	42%	93%
ORBacus	5	2	50	69	36	42%	60%
OpenJMS	23	3	313	159	109	66%	86%
Total	37	8	493	406	312	55%	84%

Table 3. Static assessment of FlexSync **implementations**

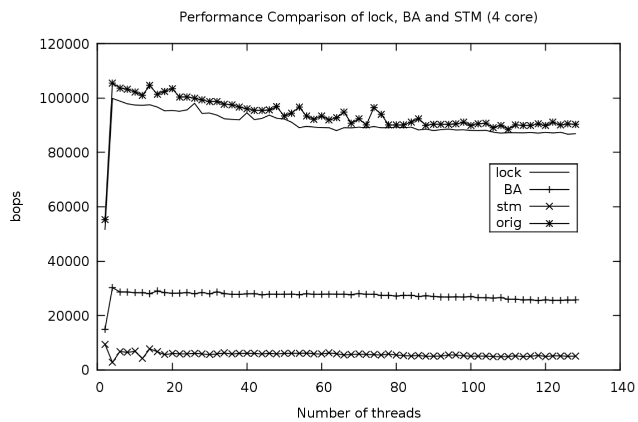

Figure 4. specjbb2005 FlexSync **performance**

4.2 Functional characteristics of FlexSync

Performance of FlexSync

To assess the performance of FlexSync, we use the specjbb2005 benchmark as an example of stand-alone Java applications. To introduce transactional behaviors in specjbb, we label the group of class types representing business transactions as `TXOwner` or `BAOwner`. We label all types declaring `synchronized` methods with `TXTarget` or `BATarget`, respectively. In Figure 4, we plot the benchmark score[16], bops (business operations per second), against the number of threads.

Our first observation from Figure 4 is that, for the lock version, the FlexSync approach does not incur significant runtime overhead (about 5%). This is generally true in all our experiments as will be shown shortly. The general performance profile of specjbb2005 is the same as our motivating buffer example that the lock-based approach gives the best performance (highest score), followed by BA, then by STM. The difference is that the BA score is at about 30% of that of the lock-based, and the STM version is about 20%-25% of the BA version. The reason for this dramatic slowdown is that each specjbb2005 test involves a large amount of synchronized data operations. As verified by our runtime profiling, in the BA version, over 20% of the CPU time is spent on lock contention which seriously limit the concurrency of the system. For the STM version, over 50% of the time is spent on backing up the data by the dstm2 runtime. The memory requirement for executing STM on specjbb ranges from 1 to 40 times larger as compared to the original version.

This experiment supports our motivation that each synchronization mechanism has its unique strengths and weaknesses. Clear trade-offs exist among properties such as correctness, safety, and performance. The FlexSync approach gives the customization options to the users of Java programs and let them decide which properties should take the precedence in their application domain.

Case studies of compositions

In this study, we use OpenJMS as an example of complex program composed from reusable systems and capable of supporting multiple architectural configurations. OpenJMS uses the remote procedure call (RPC) as its trans-

[16] All experiments are conducted on a 4-core Intel CPU running 2.6 Linux kernels using the JRockit R27 64-bit JVM. Each data point is an average of 3 identical runs, measured after warm-up periods.

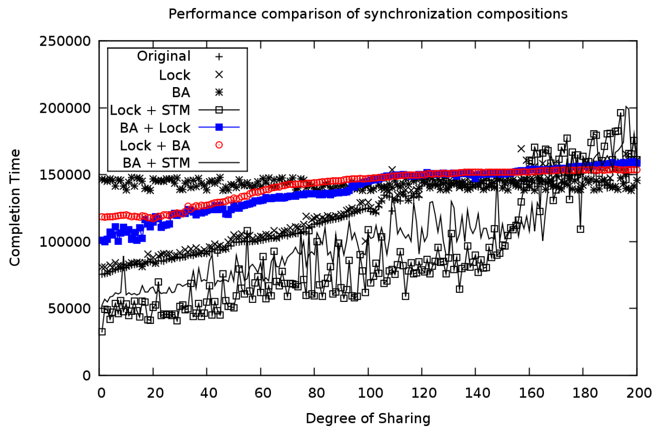

Figure 5. Combinatorial synchronization

port level mechanism, which is supported by ORBacus, a general purpose RPC middleware. To support the transactional and atomic operations, we mark types contains `synchronized` methods as `TXTarget` or `BATarget`, and we make the starting point of RPC invocations in ORBacus and of the message dispatching in OpenJMS as where the transactional or the atomic executions start.

In the first case study, we first gain a general perspective of the combinatorial complexity by quantifying all possible choices of synchronization mechanisms in the case of OpenJMS/ORBacus system. To obtain these points, we measure the response time for the JMS server in receiving a fixed number of messages into a set of message queues. We define *degree of sharing* as the average number of clients sharing each message queue. We achieve this by generating the client/queue association before each test and hardwiring the client/queue relationships during the run. We tested 7 possible configurations, reported in Figure 5. We make the following observations: 1. the lock version of FlexSync approach does not incur the performance overhead compared to the original version; 2. multiple synchronization mechanisms can coexist in providing JMS services; 3. the responses involving STM, although oscillating significantly, are faster than the lock-based versions, which is contradictory to our previous studies; 4. the versions involving BA have the worst performance compared to other versions. The surprising results about STM is due to the fact that, as each transactional operation results from a round of client-server communication, the state duplication in OpenJMS is far less frequent compared to that of specjbb2005 and our buffer example. In the case of BA, the use of 2PL significantly limits the concurrent degree of the system, which proves that the degree of liveness is vital to the performance of server-type systems

Our second case study illustrates the scenarios of the unanticipated compositions and how the FlexSync-based synchronization understands these cases and achieves performance gains. The canonical concurrency policy used by the OpenJMS/ORBacus system is that each connected client is assigned a dedicated thread on the server side. Data structures storing the RPC targets and the message queues are shared among these threads.

Scenario one: Queue dedication If the physical capacity allows, the server side can dedicate a separate messaging stack for each OpenJMS client simply by publishing each queue with a unique RPC address. This set-up alone can improve the processing throughput from 7% to 25% by our measurements. This is unanticipated scenario that can be achieved purely through deployment configurations. In such configurations, since there are no shared states, we can simply instruct the FlexSync weaver not to weave any synchronization mechanisms to achieve further speed-ups.

Scenario two: Event-driven RPC The RPC engine can make use of the reactor-based [20] event-driven concurrency models for its capability of handling problems such as C10k[17]. Since such models typically make no use of threads, the RPC engine serially dispatches requests, rendering the thread-safety property of the upper layer messaging mechanism in OpenJMS redundant. Again, this is a per-deployment scenario hard to be anticipated by the design of OpenJMS. This scenario can be created by using the CAL-based ORBacus implementation from our earlier research [23]. The FlexSync weaver scans through the bytecode image of the entire system and is able to detect that all shared states of the messaging layer does not escape from the executing thread of the reactor. Therefore, no synchronization mechanisms are applied as the result.

In Figure 6, we compare the following configurations: the canonical concurrency model (*shared original*), original OpenJMS configured to run dedicated queues (*dedicated original*), original OpenJMS using reactor (*shared reactor*), dedicated queue using reactor (*dedicated reactor*), and the afore-mentioned two optimized versions using FlexSync (*dedicated flex* and *dedicated reactor flex*). Our measurements first justify the validity of the case study where queue dedications and the use the reactor can produce speedups from approximately 7% to 25% at 500 clients. In queue dedication scenarios, the FlexSync-optimized version produces 23% speedups compared to the original version. For the use of the single-threaded reactor, the FlexSyncversion produces 25% speedups. The largest speedup, considering all configurations, is around 40%. These measurements prove the functional advantage of FlexSync-based synchronization implementations.

[17] The C10K problem. `http://www.kegel.com/c10k.html`

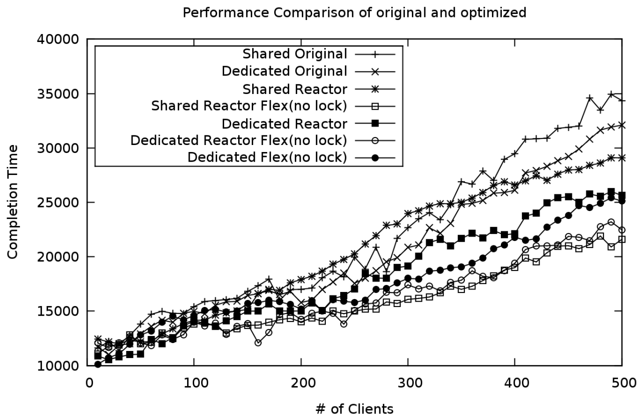

Figure 6. Performance optimization

5 Related Work

As a very active research area, research projects in the context of lock, atomicity and software transactional memory are beyond enumeration. We focus on presenting research addressing the programming aspect of synchronization challenges. We first covers the aspect oriented approaches to synchronization. We then present the research work on the synthetic approaches for conventional programming languages. We conclude with the discussion on various new language proposals.

AOP implementations Lopes and Lieberherr [15] presented one of the earliest AOP treatments of synchronization using *adaptive programming* [14]. In their approach, the structure of a program and its behavior, including the lock-based synchronization, are expressed in separate modules. Code generation is required to produce the final executable system. As a pioneering work, they focused on illustrating a benefit of AOP-based synchronization implementation as compared to the conventional approach. Our work is built on these insights and going one step further in considering how different synchronization mechanisms can coexist, be customized, and interact with the core program consistently in complex Java systems.

SyncGen [5] focuses on generating synchronization implementation from high-level specifications. These specifications (aspects) are invariant formulae, which are translated into byte-code instructions that are inserted into (weaving) the demarcation points of the synchronized region. Compared to our approach, aside from the lock-only approach, SyncGen introduce a new programming paradigm of specifying synchronization in high-level logic formulae. Our approach relies on FlexSync APIs to re-express the synchronization intention, therefore, does not fundamentally deviate from how the synchronization design is reasoned conventionally.

Lock synthesis Emmi et al [6] presented an automatic technique that takes a program annotated with atomic sections and produces a lock assignment for global variables that provides atomicity and deadlock free guarantees. Their work provides evidence that synchronization can be reasoned independently if we can know the programmers' intentions, in their case, through annotations. Research such as [4, 2] eliminates unnecessary lock placements through static analysis techniques. Our work directly leverages these results in performing selective aspect weaving.

Language approach There have been a proliferation of new language proposals, such as [16, 8, 22, 3] and many others, that provide new language design and the semantic guarantees to help programmers in writing safe, correct, and performant synchronization code. Rewriting complex applications with new languages is not always straightforward. The majority of the language proposals, being focusing on specific synchronization mechanisms, also inherit their limitations. We believe that the capability of customization is still a desired property for systems written in these new languages.

6 Conclusion

In the multi-core era, concurrency plays critical roles in improving software efficiency. The synchronization mechanisms of reusable Java systems are challenging to build because each of these mechanisms has unique strengths and weaknesses which are sensitive to specific usage requirements. In conventional approaches, synchronization is reasoned locally within the designed application and in a non-modular way. As a result, applications pay significant performance costs due to the mismatch between the non-flexibility of the systems and the diversity of deployment scenarios.

We have presented FlexSync, an aspect oriented synchronization library, to alleviate this problem by physically decoupling the synchronization implementation from the operational logic of Java systems. This is based on our observation that the design intention of synchronization and the specific choice of synchronization mechanisms can be explicitly separated and, in practice, most of the intentions can be represented by function-like structures. The FlexSync API fosters the modular reasoning of synchronization by essentially enabling programmers to give different interpretations of the same program structure according to the different synchronization semantics. The FlexSync library encapsulates reusable logic about how synchronization mechanisms and the operational logic interact. The FlexSync loadtime weaver performs the global reasoning of synchronization by applying the system-wide deployment-time analysis to achieve consistency and optimization.

We evaluated FlexSync with commercially used complex Java concurrent systems, and FlexSync is capable of supporting the functionalities of these systems. We quantify both the programming effort and the functional characteristics of FlexSync-based implementations. We found that programming synchronization in FlexSync is commonly declarative and specification-like. The FlexSync approach in general does not incur significant runtime overhead. In addition, systems using FlexSync also has the capability of making customization choices regarding domain-specific or deployment-specific requirements. We showed that, the FlexSyncapproach does not significant incur runtime overhead and can produce speed ups as much as 40% in various deployment-time configurations.

As future work, we aim to make programming with FlexSync a lot easier by focus on the automation of intention refactoring and state duplication. We also plan to study the synergistic effects among customizable concurrency models [23] and synchronization mechanisms. Our long term research goal is to significantly increase the customization capability of complex systems.

Acknowledgements

We thank all the anonymous reviewers for providing insightful comments, suggestions, and feedbacks. These reviews have significantly helped us improving this paper.

References

[1] Ali-Reza Adl-Tabatabai, Brian T. Lewis, Vijay Menon, Brian R. Murphy, Bratin Saha, and Tatiana Shpeisman. Compiler and runtime support for efficient software transactional memory. In *PLDI*, pages 26–37, New York, NY, USA, 2006. ACM.

[2] Jonathan Aldrich, Craig Chambers, Emin Gun Sirer, and Susan Eggers. Static analyses for eliminating unnecessary synchronization from java programs. In *SAS*, page pages, 1999.

[3] David F. Bacon, Robert E. Strom, and Ashis Tarafdar. Guava: a dialect of java without data races. In *OOPSLA*, pages 382–400, New York, NY, USA, 2000. ACM.

[4] Jong-Deok Choi, Manish Gupta, Mauricio J. Serrano, Vugranam C. Sreedhar, and Samuel P. Midkiff. Stack allocation and synchronization optimizations for java using escape analysis. *ACM TOPLAS*, 25(6):876–910, 2003.

[5] Xianghua Deng, Matthew B. Dwyer, John Hatcliff, and Masaaki Mizuno. Invariant-based specification, synthesis, and verification of synchronization in concurrent programs. In *ICSE*, pages 442–452. ACM Press, 2002.

[6] Michael Emmi, Jeffrey S. Fischer, Ranjit Jhala, and Rupak Majumdar. Lock allocation. In *POPL*, pages 291–296, New York, NY, USA, 2007. ACM.

[7] Cormac Flanagan and Stephen N Freund. Atomizer: a dynamic atomicity checker for multithreaded programs. In *POPL*, pages 256–267, New York, NY, USA, 2004. ACM.

[8] Cormac Flanagan and Shaz Qadeer. A type and effect system for atomicity. In *PLDI*, pages 338–349, New York, NY, USA, 2003. ACM.

[9] Jan Hannemann and Gregor Kiczales. Design Pattern Implementation in Java and AspectJ. In *ACM OOPSLA*, pages 161–173. ACM Press, 2002.

[10] Maurice Herlihy, Victor Luchangco, and Mark Moir. A flexible framework for implementing software transactional memory. In *OOPSLA*, pages 253–262, New York, NY, USA, 2006. ACM.

[11] Elizabeth A. Kendall. Role model designs and implementations with aspect-oriented programming. In *ACM OOPSLA*, pages 353–369. ACM Press, 1999.

[12] Gregor Kiczales, John Lamping, Anurag Menhdhekar, Chris Maeda, Cristina Lopes, Jean-Marc Loingtier, and John Irwin. Aspect-oriented programming. In Mehmet Akşit and Satoshi Matsuoka, editors, *ECOOP*, volume 1241, pages 220–242, Berlin, Heidelberg, and New York, 1997. Springer-Verlag.

[13] Gregor Kiczales and Mira Mezini. Aspect-oriented programming and modular reasoning. In *ICSE*, pages 49–58, New York, NY, USA, 2005. ACM.

[14] Karl Lieberherr, Doug Orleans, and Johan Ovlinger. Aspect oriented programming with adaptive methods. In *Communications of the ACM*, volume 10. ACM, 2001.

[15] Cristina Videira Lopes and Karl J. Lieberherr. Abstracting process-to-function relations in concurrency object-oriented applications. In *ECOOP*, pages 81–99, London, UK, 1994. Springer-Verlag.

[16] Bill McCloskey, Feng Zhou, David Gay, and Eric Brewer. Autolocker: synchronization inference for atomic sections. In *POPL*, pages 346–358, New York, NY, USA, 2006. ACM.

[17] L. Northrop, P. Feiler, R. P. Gabriel, J. Goodenough, R. Linger, T. Longstaff, R. Kazman, M. Klein, D. Schmidt, K. Sullivan, and K. Wallnau. Ultra-Large-Scale Systems: The Software Challenge of the Future. Software Engineering Institute, 2006.

[18] Venkatesh Prasad Ranganath and John Hatcliff. Pruning interference and ready dependences for slicing concurrent java programs. In *CC*, pages 39–56. Springer, 2004.

[19] Erik Ruf. Effective synchronization removal for java. In *PLDI*, pages 208–218, New York, NY, USA, 2000. ACM.

[20] Douglas Schmidt, Michael Stal, Hans Rohnert, and Frank Buschmann. *Pattern-Oriented Software Architecture Patterns for Concurrent and Networked Objects*, volume 2 of *Software Design Patterns*. John Wiley & Sons, Ltd, 1 edition, 1999.

[21] Nir Shavit and Dan Touitou. Software transactional memory. In *PODC*, pages 204–213, New York, NY, USA, 1995. ACM.

[22] Mandana Vaziri, Frank Tip, and Julian Dolby. Associating synchronization constraints with data in an object-oriented language. In *POPL*, pages 334–345, New York, NY, USA, 2006. ACM.

[23] Charles Zhang and Hans-Arno Jacobsen. Externalizing Java Server Concurrency with CAL. In *ECOOP*, pages 362–386. Lecture Notes in Computer Science 5142. Springer, 2008.