

MASARYK UNIVERSITY
FACULTY OF INFORMATICS



Real-time collaboration in Komodo

MASTER'S THESIS

Bc. Matúš Makový

Brno, Spring 2015

Declaration

Hereby I declare, that this paper is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Bc. Matúš Makový

Advisor: RNDr. Filip Nguyen

Acknowledgement

Abstract

Keywords

Contents

1	Introduction	1
2	Real-time collaboration	3
2.1	<i>Basic overview</i>	3
2.2	<i>Operational Transformation</i>	6
2.3	<i>Differential Synchronization</i>	15
2.4	<i>Commutative Replicated Data Types (CRDTs)</i>	21
3	Comparison of Techniques	28
3.1	<i>Comparison Criteria</i>	28
3.2	<i>Comparison</i>	30
4	Komodo	32
4.1	<i>Overview</i>	32
4.2	<i>Komodo Requirements for Real-time Collaboration</i>	32
4.3	<i>Best Technique for Komodo</i>	32
4.4	<i>Java Implementation</i>	32
5	Conclusion	33

1 Introduction

Many software solutions enable people to create new things in a better and faster way. In most cases the resulting product should be so complex that one person is not enough for the successful and fast creation. Creators try to collaborate to achieve a common goal. Among other opportunities and possibilities are collaboration and sharing the greatest benefits of the Internet.

We can identify two types of collaboration over the Internet, non real-time collaboration and real-time collaboration. At the beginning, as the Internet didn't have such capacity, people tried to use it just for sharing their drafts of work and sending them to each other, this type of collaboration is called non real-time. In non real-time collaboration users work on separate copies of a project and then need to merge their changes into one final project. In other words, they had to find differences between their drafts and reflect them to each other's version. It doesn't offer such flexibility as real-time collaboration and also it had many limitations, for example two people could not edit the same file in a project without having to resolve conflicts manually when they tried to merge their work with collaborator's version

Examples of non real-time collaboration could be Revision control (Git, SVN). With the development of the Internet came a reasonable solution called real-time collaboration. Using this principle, author can see what his collaborator is doing in real-time and the manual synchronization or manual conflict resolution is not necessary. Information technologies take care of this synchronization and conflict resolution for the users.

In the theoretical part of this thesis, author deals with principles of real-time collaboration and describes some of the techniques used to implement real-time collaboration in software over the Internet. There is also a comparison of these techniques from various aspects.

The practical part of this thesis deals with JBoss Data Virtualization and Komodo software as a new version of Teiid Designer developed by Red Hat that should use real-time collaboration in its upcoming release. Author recommends best technique for this authoring software regarding the requirements of data models and finds a suitable implementation in Java programming language.

2 Real-time collaboration

This chapter covers the basic overview of collaboration in general, description of real-time collaboration and description of difficulties with its implementation over the Internet. Last sections of this chapter describe three techniques used for implementation over the Internet and its properties. Techniques described in this chapter are: Operational Transformation, Differential Synchronization and Commutative replicated data types.

2.1 Basic overview

Collaboration, as defined by English dictionary, is an act of working with another or others on a joint project. Authoring Systems in IT that support collaboration are called Groupware.

"Groupware systems are computer-based systems that support two or more users engaged in a common task, and that provide an interface to a shared environment. These systems frequently require fine-granularity sharing of data and fast response times." [2]

There are many techniques used in groupware systems that are not suitable for real-time collaboration, for example, Locking or Single Active Participant technique. Fundamental principle of Locking and Single Active Participant is to lock data when it is being modified by someone. This principle is adopted for example by Microsoft, when users try to edit shared document.

When using real-time collaboration, authoring software creates an illusion that users are working on one common copy of a document online. There is no requirement to commit changes to some kind of shared repository and no need for a user to resolve conflicts. Changes should be reflected and saved immediately. Examples of real-time collaborative editors that we recognize today are Google Docs,

Etherpad and already terminated Google Wave.

Software engineer has different options for implementetation of real-time collaboration in software solution.

Requirements for a good technique are:

- speed
- latency tolerantion
- low data transfer
- consistency maintanance
- good conflict resolution

The speed requirement means that changes made on one side of the collaboration process need to be reflected to the other sides as soon as possible and vice versa. Changes should be sent to other collaborators as soon as they are done. If a collaboration technique is fast enough it is much easier to satisfy other requirements on this technique. Fast enough technique is able to maintain good consistency. "The system's response time is the time necessary for the actions of one user to be reflected by their own interface; the notification time is the time necessary for one user's actions to be propagated to the remaining users' interfaces." [2] Response time and notification time should be as short as possible.

One of the problems of real-time collabarative editing could be the latency of network. Implemented technique should be able to tolerate the latency of internet connections, because collaborators could be in very distinct parts of the world. It should be able to reconstruct the right order of operations because data sent over the Internet don't necessary come in right order and the order of the operations is very important to maintain consistency.

Low data transfer requirement is also very important. Different sides of the collaboration should send as few data as possible. Transferred data should only describe change that has been done on one side of collaboration, it is not necessary to transfer the whole project. The less data is needed to transfer the faster the whole protocol can be.

Consistency maintenance is necessary for the success, it has to be ensured, that users on both sides are looking at the same version of a document regardless of number and complexity of operations done on both sides of the collaboration. Lack of consistency could cause other problems and chaos in the document versions. According to [14] there are three problems encountered when trying to achieve consistency maintenance and they correspond to the properties of CCI consistency model proposed in [13]

1. **Casuality Preservation** - operation O_1 causally precedes operation O_2 if O_1 occurred locally before O_2 . The problem is to execute operations in right order on all sites.
2. **User Intension Preservation** - technique must preserve user's intension in the context of state, in which the operation was executed. This problem is in strong relation with conflict resolution requirement and occurs when it is not possible to determine if O_x causally precedes O_y or O_y causally precedes O_x .
3. **Convergence** - when same operations have been applied on every site, the documents are identical.

Because the real-time collaboration should be asynchronous there arises a problem of concurrency. This means, that changes can happen at the same time and in the same sections of project. Good conflict resolution requirement is present because of concurrency. Implementation techniques have to be able to identify and resolve a conflict when users are editing the same part of the project.

The concurrency control algorithms are separated in two classes: pessimistic and optimistic.

"Pessimistic algorithms require communication with other sites or with a central coordinator before making a change to data." [7] One example of pessimistic algorithm can be already mentioned Locking or Single Active Participant technique.

"Optimistic concurrency control, on the other hand, requires no communication before applying changes locally. The party making a change applies it immediately, then informs the other parties of the action. If more than one participant makes a change at the same time, a conflict resolution algorithm creates compensating changes to move everyone to the same final state." [7]

This implies that optimistic algorithms are better solution for Internet, because the latency can not be guaranteed.

2.2 Operational Transformation

Operational transformation (OT) is optimistic concurrency control algorithm used for real-time collaborative editing over the Internet.

It was first introduced in paper Concurrency Control in Groupware Systems [2] in 1989, together with The Distributed Optional Transformation (dOPT) Algorithm. "The algorithm has a number of properties which make it suitable for groupware. First, operations are performed immediately on their originating site, thus responsiveness is good. Secondly, locks are not necessary so all data remains accessible to group members. Finally, the algorithm is fully distributed, and resilient to site failure..."[2]

This first algorithm was able to process only plain text. Later some problems with correctness were discovered and resolved in following works. Over the years many other algorithms implementing

OT have been published. Algorithms used today are able to process also XML and other formats.

OT was used also in Jupiter Collaboration System in 1995 [7]. Jupiter's two-way algorithm is derived from the dOPT algorithm used by Grove in [2]. This algorithm is one of the improvements of dOPT.

Today is Google one of the most common implementators of this protocol. The main application was in Google Wave project, which is now terminated, although Operational Transformation found application also in Google Docs project. According to [12] Google is one of the biggest innovators of this approach and its biggest contribution is the idea of operation composition, which is described later in this section.

The approaches differ also in the architecture, dOPT doesn't involve server in its design, but the recent algorithms use a central server that maintains its version of the document, coordinates the communication and broadcasts operations.

This section will explain basic principles of Operational transformation using Google version of the control algorithm, introduce some improvements added by Google and then point out the differences between the presented version and other versions like dOTP, Jupiter and TIPS.

The basic idea is that data are replicated on every machine and the only information that is sent over the Internet are the operations. Data replication ensures good responsiveness in high latency environments. Wave's addition to OT are also annotations. "An annotation is some meta-data associated with an item range, i.e., a start position and an end position. This is particularly useful for describing text formatting and spelling suggestions, as it does not unnecessarily complicate the underlying structured document format." [15]

The main building block of this approach, as the name suggests,

is an operation. Here are operations defined for use in Google Wave [15]:

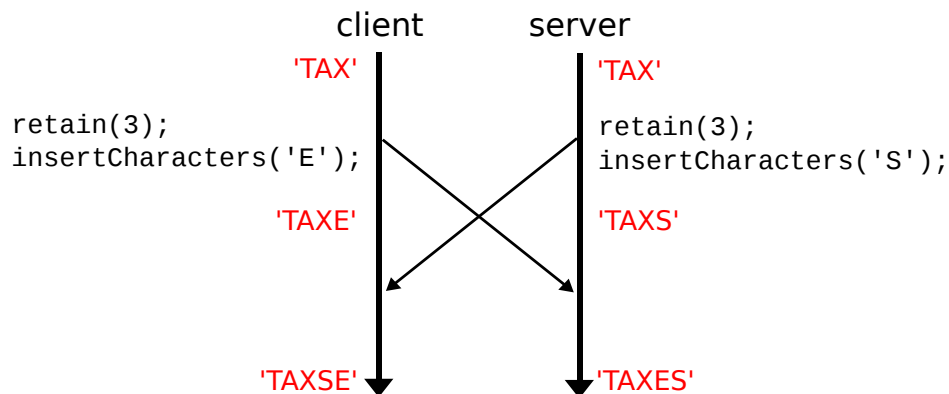
- `retain()` - move cursor
- `insertCharacters()` - inserts texts
- `insertElementStart()` - inserts starting tag
- `insertElementEnd()` - inserts end tag
- `deleteCharacters()` - deletes text
- `deleteElementStart()` - deletes starting tag
- `deleteElementEnd()` - deletes ending tag
- `replaceAttributes()` - replaces attributes in tag
- `updateAttributes()` - updates attributes in tag
- `annotationBoundary()` - describes changes in annotations

The operation is executed locally and then sent over the network to the server or to other peers, to execute the operation on their version of the document. Every object that can be changed in the document has its index for identification. Google Wave uses XML as format for storing information so as an object is considered a character in text and also a starting and terminating tag in XML structure (characters of tags are not considered elements).

The second major part of this technique is a transformation function, that transforms operations. Operations which are sent over the network and received by server can not be executed directly on the server's version of the document, because the index of the desired element could be different in the local copy and in the server copy of the document and this can cause inconsistency and violation of User Intension Preservation.

See Figure 2.1 with example for better explanation. For the purpose of this example, we will assume that we have one client that sends operations done on his local copy and a server that receives operations from all peers and maintains server version of the document.

Figure 2.1: Example Operational Transformation



Both of them start with text 'TAX' in their document. Pointer starts at index 0 and operation `retain()` shifts the pointer by number of indexes given as parameter. Client wants to add character 'E' so he executes operations:

```
retain(3);
insertCharacters('E');
```

In the meantime server received operation from other peer, that has inserted character 'S', so server executes operations:

```
retain(3);
insertCharacters('S');
```

Server and client exchange the operations. Different order and difference in indexes on both sides of the collaboration result in inconsistent document state. String in first document is 'TAXSE' and string in sec-

ond document is 'TAXES'. In order to preserve user's intension and to reach correct result we have to transform this operations. The correct result after inserting letters 'E' and 'S' should be 'TAXES'. Server has correct version of the document, but client doesn't so we have to transform the operation sent by server to look like this:

```
retain(4);
insertCharacters('S');
```

This transformation was helpful, but it is not enough. Now the client operation retains 3 characters, inserts new character and leaves cursor in front of last letter and the server operation retains 4 characters inserts character and leaves cursor behind the whole word. Client and server should end up in the same state. So it is necessary to transform also the client operation:

```
retain(3);
insertCharacters('E');
retain(1);
```

After this change the result of applying operations on both sides will be correct.

More gernerally, if we have two operations O_1 and O_2 we need a transformation function *trasform*, such that:

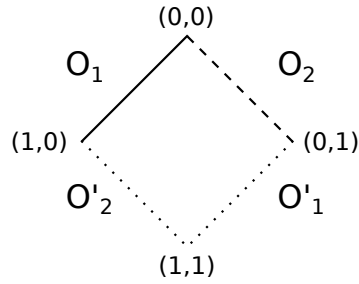
$$transform(O_1, O_2) = (O'_1, O'_2) \quad \text{where} \quad O'_2 \circ O_1 \equiv O'_1 \circ O_2$$

In other words, the function takes two operations and transforms them against each other and the output of the function are two transformed operations, such that, if we have same strings on both sides a we apply O_1 and O'_2 on one side, and O_2 and O'_1 on the other side, the resulting string is the same.

"If an OT system is to support particular functionality, then it must be able to support certain transformation properties. For group editing and consistency maintenance, the system must support a trans-

formation function known as Inclusion Transformation (IT). For group undo, where the effect of a previously executed operation is un-done at all sites, and all operations executed after it are all re-transformed, the system must support another transformation function known as Exclusion Transformation(ET). Inclusion Transformation transforms operation O_A against another operation O_B in such a way that the impact of O_B is effectively included. Exclusion Transformation transforms operation O_A against another operation O_B in such a way that the impact of O_B is effectively excluded." [5] Problem described in the example could be also visualized as a "diamond problem". See Figure 2.2

Figure 2.2: Diamond problem - 1



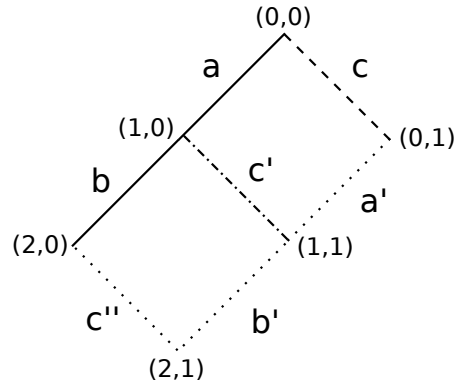
Operations done by peer take the document to the left side and operations done on server take the document to the right side. Each node in the diamond diagram represents the number of operations done by client and server in particular state.

As for the example, the correct scenario would be that peer and server exchange their operations, each one computes its transformation and applies the transformed operation on their versions of documents.

The problem presented by this example is actually the only one that is Operational Transformation capable of solving. Complicated situation occurs when sides of the collaboration diverge by more

then one step.

Figure 2.3: Diamond problem - 2



This example is used in [12] On the server side, transformation for operations a, c has to be computed, and server applies operation a' on its version of the document. The remaining operation c' has to be preserved for the next transformation. In the next step server must use b and c' as an input and computes b' that can be applied on its version of the document. Server's document is in the desired state.

On the client side, c has to be transformed against two operations to get to desired state. For two operations it is not a big problem, but as the number of operations could very quickly increase, because Operational Transformation is an optimistic technique, so it could potentially take long time to process. One of the advantages of Google Wave Operational Transformation is that it is able to compose the operations if they are compatible. This Composer is able to validate operations to make sure that they are compatible and compose them. One of the criteria for compatibility is that operations must span the same number of indexes. If operations a and b are composed into operation d , transformation function can take operations d, c and produce operation c'' that can be applied to peer's document and the document is in desired state.

This example is still simple. It does not point out some necessary parts of Operational transformation. The importance of preservation of operation c' was obvious from the diagram, but without the diagram it is hard to determine what operation should server or client preserve. In order to make it easier, metadata about the parent state of the operation is added. Every operation has an identifier of state in which it was executed. Google Wave uses a scalar version number, but [12] suggests to use a hash of the documents contents.

Having information about parent state, the server is able to determine that parent state of operation b is not in server history and it has to derive a new operation, that would take the document to state after application of operations a and c . According to [12] this operation is c' . This deriving of operation c' is called bridging.

With this new feature comes a new problem. In order to derive the operation c' server has to preserve also operation a . This could be a scalability problem in an environment with many clients editing the same document, since server has potentially preserve this data for every client. This is resolved by transferring all the responsibility to client and buffering the operations on client side. Client can send only operations that come from state in server's history and also he can send only one operation to the server and he has to wait for the acknowledgment of the operation. All following operations are preserved in buffer on client side. One of the advantages of buffering operations is that if these operations are compatible, they can be composed using the Composer. When the server acknowledges first operation, client can send next operation. This way client can predict server's path and send only operations that are on the server's path. Also it is much more complicated for the server to track every client, but it generally is much easier for client to track only one server. This way the server solves only one step diamond problem presented in the first example.

Daniel Spiewak in article Understanding and Applying Operational Transformation [12] describes some other improvements of Operational Transformation by Google Wave. For example, when client receives operation from server and has some operations in buffer, he can transform the whole buffer against this operation in order to accelerate the process on his site and also to ensure that operations in buffer have a parent state in servers history.

The main difference between Google control algorithm and the first dOPT algorithm is that dOPT doesn't use transformation function, it uses transformation matrix. If we have set of operations of cardinality m the matrix should be of size $m \times m$. The entries of the matrix correspond to all possible pairs of operations and contain functions which transform operations to other operations.

Another important part of the algorithm is the state vector. "Timestamps for each client are handled by a vector timestamp, where a state vector s_i for a client C_i will have at position j , the number of operations known to have been executed by client C_j ." [5]

According to this vector it is decided whether the received operation is executed or put in the request queue.

The difference between Google algorithm and Jupiter is not so significant. Actually, Jupiter also uses transformation function, it is called xform, although it is very similar to dOPT transformation matrix. Only difference is that Jupiter doesn't use the operation composition. According to [7] this algorithm assumes use of transport layer that delivers data in right order, such as TCP.

The last major algorithm implementing this technique is TIPS, which is modern and based on the admissibility-based transformation (ABT) framework. It is created on top of HTTP protocol. Main difference is that clients can join and leave session at any time and the algorithm is more reliable when network failures can be expected.

2.3 Differential Synchronization

Differential synchronization is the second technique used for real-time collaboration. It was fully described in 2009 by Neil Fraser in paper Differential Synchronization [3]. This section will describe and explain this technique, point out its advantages and disadvantages. As the main source of this information will be used paper mentioned above, because it is also the only article that fully describes this principle.

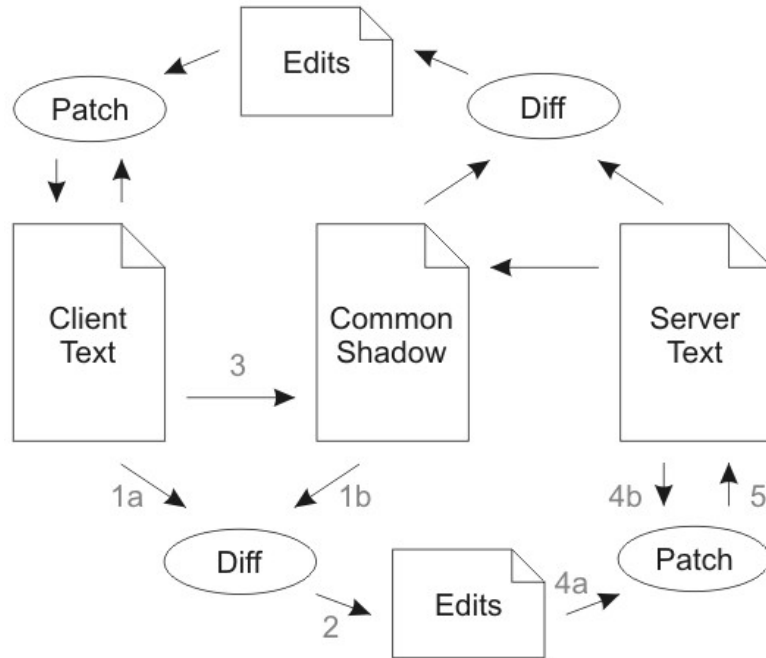
DS also replicates data on all sites of collaboration and identical code runs on both server and client, so it is symmetrical. The main building block is the presence of diff & patch algorithms. Diff algorithm is able to compute a difference between documents and save it in appropriate format for the patch algorithm, which is able to apply these changes on the other copy. The patch algorithm must be fuzzy, this means that the changes may be applied even if the document has changed in the meantime, because Differential Synchronization is an optimistic technique. The final version presented by Neil Fraser is suitable for unreliable networks and networks with high-latency. This property will be described later in this section.

"The key feature of DS is that it is simple and well suited for use in both novel and existing state-based applications without requiring application redesign." [3] Differential Synchronization is able to process variety of formats, not only plain text. As long as there is a diff and patch algorithm available for the desired format, DS is able to use it. It is implemented and used in MobWrite.

Basic principle of this technique can be described using data flow diagram in Figure 2.4 originally presented in [3].

In the beginning Client text, Server text and Common Shadow are the same. Client text and Server text represent two sites of the collaboration. The goal is to keep them updated. Client and Server are allowed to make changes in their documents at any time.

Figure 2.4: Differential synchronization - Basic Architecture [3]



After specified time interval (timeout) a snapshot of Client text is taken and using a diff algorithm the difference between client snapshot and common shadow is acquired. Common shadow represents the last common document state before any edits have been made. The output of these algorithms are changes that have been done by client. The client snapshot is copied over to the Common Shadow and the changes are patched on the Server text. Now the process repeats symmetrically using the server text in order to apply changes made by server or other clients on the client text. This time the Common shadow is the same as client snapshot, so diff algorithm returns changes made by server. The technique is very simple, it is obvious that the main parts of this technique are the diff a patch algorithms.

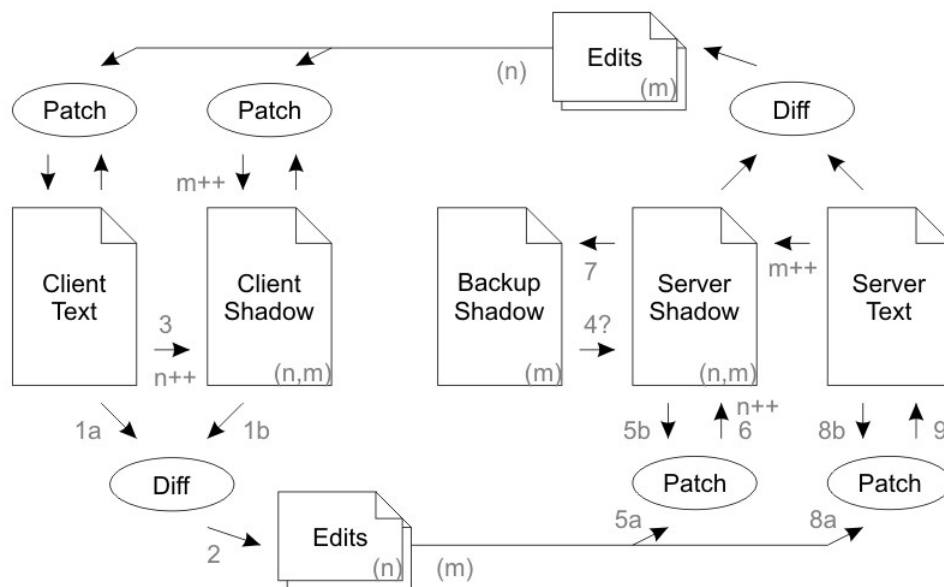
This example is suitable for explanation of the theory behind this technique, but not for use in practice. The problem is that the shadow

is common and there is no method to implement this in real life. Client and server must have separate copies of "common" shadow. These copies are called Client Shadow and Server Shadow.

We have to make sure that Client Shadow and Server shadow are the same after every synchronization. The delivery is not guaranteed and there is possibility that the shadows may be out of sync. Neil Fraser in [3] suggests sending a simple checksum of shadow together with the list of edits. After these edits are applied checksum of local shadow is computed and compared to the recieved checksum. If they do not match one of the sides has to send the whole document to get the shadows back to sync. This solution is able to detect this problem with packetloss, but it is not able to solve it without dataloss.

Differential synchronization has an architecture that enables recovering from this failure. Figure 2.5 shows this architecture.

Figure 2.5: DS - Guaranteed Delivery Method [3]



The main difference is that in this architecture server has another shadow text called Backup shadow to maintain the last version of server shadow and both client and server maintain list of edits that were sent and not acknowledged, in [3] is this list called outbound stack. Also the versions of Client shadow and Server Shadow are labeled with version numbers to be able to detect a failure in communication and acknowledge messages.

Basic scenario, that is presented on the picture can be described as follows:

1. Client text and Client Shadow are used as arguments for diff algorithm.
2. List of edits is produced and labeled with Client Shadow version n .
3. Client text is copied to Client Shadow and its version is incremented.
4. This step is used only in specific situations that are described later
5. List of edits together with last known Server Shadow version number m is sent to server. This list of edits and the current Server Shadow are used as arguments for patch algorithm.
6. Server patches edits to its shadow and increments last known client shadow version number.
7. Server takes backup of the current server shadow by copying it in the Backup Shadow. Backup shadow contains also m .
8. List of edits and Server Text are used as arguments for the patch algorithm.
9. Output of patch algorithm is used to patch the Server Text.

And again the process repeats simetrically, only exception is that client doesn't take a backup of his shadow. Every time one of the sites recieves an acknowledgement of its edits (m or n), it can delete these edits from outbound stack.

This was a description of a normal scenario without failures. To prove that this new architecture is more robust, in the following section some other scenarios will be described.

- **Lost data sent from client** - List of edits sent from client to server is not acknowledged. Client keeps edits in the stack and after the timeout, client sends new edits together with these old not-acknowledged edits. This repeats until server acknowledges some data.
- **Lost data sent from server** - Server sends list of its edits together with acknowledgement of last client edits. Packet never reaches the client. Client sends new edits together with old edits (because "old" edits were never acknowledged and also never deleted from outbound stack of client). Server spots that version number m recieved from client doesn't mach version number m on server, but it matches m in backup shadow. It copies the backup shadow to server shadow (step 4 from basic scenario above), applies new edits recieved from client and computes new diff for all edits made by server and never acknowledged by client and sends it.
- **Duplicate packet** - When server recieves two packets from client with same number n edits from first one are applied and second one is thrown away and the communication continues as usual.
- **Out of order packet** - The combination of scenarios mentioned above is used. One of lost data scenarios is used and then duplicate packet scenario.[3]

The last part of this section is dedicated to diff and patch algorithms.

Diff algorithm has two main roles in this technique. It is used to retrieve the list of edits made by user and also to save it in suitable format. Detection of user edits can be sometimes tricky and there arises the problem of semantic diff versus minimal diff. Minimal diff is in most cases not suitable for this implementation, there is a possibility that user's intension would not be preserved. For example, the change of word 'win' to 'kid' can be described as replacement of two letters 'w' and 'n' (minimal diff), but also as a replacement of whole word (semantical diff). When one side changes word 'win' to 'kid' and the other changes the word 'win' to 'won' the result should be either 'kid' or 'won', which are results of semantical diffs, the word should not be 'kod', which is result of minimal diffs. "An algorithm must be used to expand minimal diffs into semantically meaningful diffs." [3]

Patch algorithm takes the edits produced by the diff algorithm and patches them to the target text. This algorithm has a role of determining where to patch these edits, because the place where the desired word was at the time of edit could have changed. It has to solve a dilemma whether the patch should be applied at the section with smallest Levenshtein distance [6] between two versions of the document or it should be applied at the section close to the original location. "It is probably more correct to apply a patch onto a near-match at the expected location than to a perfect match at the other end of the document." [3]

Other important property of the algorithm that need to be defined in the implementation is the length of timeout. Short cycle has higher requirements on the network, because DS is constantly sending some information over the network. On the other hand, with long cycle it is necessary to transfer bigger amounts of data and the probability

of collision during patch is higher. Neil Fraser also mentions in his work that with rising number of clients scalability may become an issue.

2.4 Commutative Replicated Data Types (CRDTs)

The third technique discussed in this paper is called Commutative Replicated Data Types. Although [11] describes CRDTs that are used in various situations, for example as implementation of registers, counters, sets, graphs, and sequences, as the main source of information for this section serves the first and most detailed work Designing a commutative replicated data type [10] and later updates and improvements [9] [4] This papers suggest the use of Treedoc for cooperative editing, this data structure is described later in this section.

Interesting property of this technique is that CRDTs converge without any need of concurrency control algorithm. This is also the reason, why this section deals mainly with the architecture of the data type and not with the technique as a whole like in the previous sections. Many papers deal with this datatype and highlight its suitability for cooperative editing. [1] evaluated CRDTs for real-time collaboration and found out that this approach is suitable for real-time collaboration, in fact the paper says that they are better in some aspects than other algorithms. This approach is used in WOOT collaboration system [8] and also in Logoot [16], these two algorithms are briefly described at the end of this section.

"A Commutative Replicated Data Type (CRDT) is a data type where all concurrent operations commute with one another." [10] Operations commute when two sites that started in a same state apply the operations in different order and the resulting state is also the same. Achieving this property is equivalent to preserving user's

intention.

This approach also assumes the replication of document on all sites of the collaboration and unlike the first two techniques this one doesn't use a client-server model in its implementation, instead it uses peer-to-peer architecture.

The smallest unit that can be inserted or deleted is called atom. This can be a character but also a larger part of the structure, for example in XML could be as atoms considered whole tags. Supported operations are `insert(pos,atom)` and `delete(atom)`.

- `insert(pos,atom)` - inserts new atom at position in document
- `delete(pos)` - deletes atom at the position

Using these two operations it is possible to produce all other needed operations. It is also easy to implement transactions using CRDTs.

Every atom has its unique position identifier, it is called UID. This UID is used as parameter during execution of the operations. Properties of UIDs are:

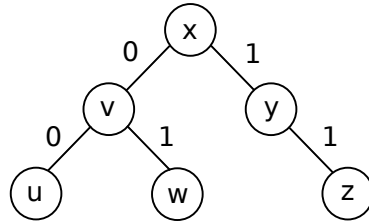
- Identifier is unique and stays the same for every atom for whole life of the document.
- There exists a total order of identifiers.
- It is always possible to generate new identifier between any two identifiers. Generally, for any two identifiers A and B , such that $A < B$ we are able to create new identifier C , such that $A < C < B$.

Fundamental principle of this technique is that operations are executed locally and sent over the network to all other peers. There should be no need to preserve the order of the operations, this is due to commutativity property and the uniqueness of identifiers. This uniqueness also ensures that there is no need to transform operations

or provide suitable context because it is clear what atom should be edited. [10] points out that real numbers are suitable for the role of identifiers, but for implementation it would be necessary to be able to work with infinite precision and that is impossible. The solution suggested by [10] is the use of Treedoc. Algorithms using CRDTs differ mostly in the way of generating/maintaining these identifiers.

Treedoc is an abstract data structure that is used in document editing. It is implemented as a binary tree and every node represents an atom in the structure. The UID is the path in the tree that starts in the root. Root's UID is empty string []. Figure 2.6 shows an example of this treedoc.

Figure 2.6: Treedoc



Treedoc in figure encodes string 'uvwxyz' and uses characters as atoms. The UIDs for the characters are: $u = [00]$, $v = [0]$, $w = [01]$, $x = []$, $y = [1]$, $z = [11]$. This structure ensures the property, that it is always possible to generate new identifier between two existing UIDs.

In the end, the whole document consists of pairs (UID,atom). These are ordered by UIDs. Partial order definition on identifiers can be found in [10] and also [9].

Delete operation is easy to implement. Deleting is simply replacing the atom in the node with null. If the deleted atom was a leaf it can be completely removed from memory, if it was an inner node it is kept in memory and garbage collected at some point in future.

However, the leaf can be completely deleted only when the delete operation is stable, which according to [10] means that has been executed on all sites of collaboration. This paper also introduces a new procedure $gc(N)$, which removes leaf N if it is stably deleted. This operation is executed only locally.

Insert operation is harder to implement. The basic algorithm implementing this operation is not trying to keep the tree balanced, that is resolved later by other suitable solutions. Full code of the algorithm can be seen in [10] [9].

Inserting new atom between to other atoms($atom_a$ and $atom_b$) is more complicated. First, the algorithm has to check if there is some atom between supplied atoms, the reason for this is that the algorithm has to insert new atom between two already present atoms that have nothing between them. In the next step, algorithm figures out whether one of atoms is ancestor of the other, there could be 3 situations:

- $atom_a$ is ancestor of $atom_b$: In this situation, $atom_b$ has no left child so this new atom will be new left child of $atom_b$. New UID will be $uid(b) \cdot 0$
- $atom_b$ is ancestor of $atom_a$: In this situation, $atom_a$ has no right child so this new atom will be new right child of $atom_a$. New UID will be $uid(a) \cdot 1$
- atoms are not in ancestry relation: In this situation, $atom_a$ has no right child so this new atom will be new right child of $atom_a$. New UID will be $uid(a) \cdot 1$

Ancestry relation is defined as expected. The node a is ancestor of node b if on the path from b to root node there is a . This also means that $uid(b) = uid(a) \cdot (0, 1)^k$ where k is the difference between depths of these two nodes.

The problem with concurrent inserts is resolved with extension of the nodes with disambiguators. These disambiguators are pairs of $idX = (siteID, counter)$ where $siteID$ represents the site that initiated the creation of this disambiguator and $counter$ variable represents counter that is maintained by each site. There is also a total order of disambiguators defined: $(s_1, c_1) < (s_2, c_2)$ iff $c_1 < c_2 \vee (c_1 = c_2 \wedge s_1 < s_2)$. This ensures that every disambiguator is unique and that the order of disambiguators is same on every site.

Figure 2.7: Treedoc

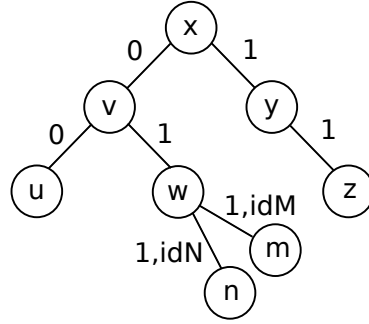


Figure 2.7 shows the use of disambiguators. As this concurrency can occur at every node, [10] suggests the use of disambiguators on every node when it is created and garbage collecting these disambiguators periodically. An insert operation is stable at some site, when this site received all operations that happened after this operation has been executed locally from all sites. At this point it is clear that there will be no other concurrent inserts.

The last operations designed for this abstract data structure are *explode(string)* and *flatten(path)*. Operation *explode(string)* accepts a string of atoms and returns a tree, that is constructed from supplied string and operation *flatten(path)* accepts a path from sub-tree and returns a string of atoms. Flatten operation skips the null nodes on the path. These nodes were produced by deleting inner nodes. Sequential ap-

plication (*explode(flatten(treedoc))*) of these two operations on an unbalanced treedoc with null nodes results in new treedoc that has the same content, but it is balanced and without null nodes in the structure. These operations don't commute with edit operations on the tree. The *explode* operation is not really necessary, it can be substituted with applying a path to a string. So there is no need to search for solution for its commutativity. The commutativity of *flatten* operation is ensured by voting. This operation can not be executed when there is another *insert* or *delete* in progress. Therefore, all the sites vote whether the flatten operation can be executed. If one of the sites spots an edit operations it votes "No" otherwise it votes "Yes". *Flatten* is executed only if all sites vote "Yes".

Interesting thing about CRDTs is that with this approach it is easy to implement transactions. As the operations commute all that is needed are labels of the beginning and the end of the transaction and a buffer. When a peer executes local transaction all operations received from network are buffered and vice versa. With transaction support it is easy to implement block operations such as copy and paste or search and replace. Because we can guarantee that during the transaction the state of the document won't change.

Last part of this section deals with Logoot and WOOT algorithms, which are alternatives to the original algorithm. They both use the idea of CRDTs, so every atom has its unique identifier and that is the reason why there is no conflict in the execution of the operations and the operations commute. They also use only 2 operations that were described earlier. The only difference is in the way of generating the identifiers.

Logoot tries to avoid the use of tombstones, which are deleted atoms that had to remain in the structure, but they are not visible. (e.g. null nodes in Treedoc). It has a bit more complicated structure for the UIDs and it uses whole lines as atoms. Every site has its site

identifier s and a "clock" h_s that is incremented every time a line is created. Every line has its position pid in the document. Position is a list of identifiers and identifier is a couple of $\langle p, s \rangle$ where p is an integer and s is the unique site identifier. Whole document consists of pairs $(pid, text)$. Also the first and the last line are uniquely identified by having l_B (beginning) and l_E (end) in the text variable. Example line generated by site s can look like this: $((\langle i, s \rangle, h_s), text)$. New UUIDs are generated by adding more identifiers $(\langle i, s \rangle)$ in the first part of the position tuple and by updating h_s . The order of the identifiers can be described as lexicographic. Authors of Logoot point out that the use of tombstones generates big overhead, so they tried to avoid it by using new format of identifiers. More information can be found in [16].

WOOT is one of the first algorithms that used CRDTs. The time complexity is not as good as Logoot's time complexity, this is caused mainly by the use of tombstones. WOOT identifies characters as atoms and every character has a unique identifier, these atoms are called W-characters. This W-character is a tuple $\langle id, \alpha, v, id_{cp}, id_{cn} \rangle$ where id is the identifier of the character, that is a couple constructed of unique identification number of the site and a logical clock. α is the actual value of the character. v is boolean value that represents visibility and id_{cp}, id_{cn} are identifiers of the previous and next character. These identifiers are the reason why there have to be the tombstones, they are necessary to identify the exact position of the character on each site. Partial order between the atoms is defined and uniqueness of the UUIDs is ensured by a unique identifier of the site and with the logical clock. The supported operations are also insert and delete.

3 Comparison of Techniques

The following chapter will present a comparison of these techniques using the criteria described in the following section. This chapter belongs to the teorecital part of the thesis, so the comparison is more general. It points out the general advantages and disadvantages of these techniques.

3.1 Comparison Criteria

This section describes the comparison criteria for the presented techniques and presents the reasons for choosing these particular criteria.

Comparison criteria include:

- client-server vs. peer to peer model
- requirements on local machine vs. remote machine
- amount of data transfer
- network requirements
- time complexity
- conflict resolution ability
- latency tolerance

It is important to determine whether the technique can be used only with client - server model or with peer to peer model or with both of them. With this requirement also related the scalability of the technique, how many participant could be in the collaboration. This is important for the use in practice, because this can be one of the most fundamental requirements of the system.

3. COMPARISON OF TECHNIQUES

One of the comparison criteria are performance requirements. These requirements should be reviewed as requirements on local machine vs. requirements on remote machine. Some techniques might have harder requirements on the user side and other on server side. Evaluation of these requirements is also dependent of the architecture.

The amount of data transfer is also important. The more data is necessary to transfer the harder it is to implement this technique in environments with collaboration sites in distinct parts of the world. Best practice is to transfer as few data as possible. This criterion should be reviewed in contrast with other techniques.

Network requirements should be taken into account, because it is necessary to know if one of the techniques requires particular protocol or some other network functionality. Some of the techniques may not be suitable for bigger networks like Internet, they may be suitable only for local area networks.

Time complexity is very closely related to the requirements on the machines. This criterion focuses on the time complexity of the algorithms that ensure consistency maintenance. Some of the techniques might do the majority of the work in the algorithm, other may have done some preparation and the following algorithm is not so complex.

Conflict resolution ability is also assessed. The main goal is to have as few conflicts as possible, but when some conflict occurs, it is necessary to resolve it as fast as possible and with the correct result. All of the techniques are able to resolve conflicts, but the important part is the correctness of the result.

The best technique should be also able to tolerate the latency of the network. The ability to recover from lost packet situation or recover from situation when the packets come out of order is important.

3.2 Comparison

This section will evaluate the techniques according to presented criteria. In the last part of the comparison a table should be presented to summarize the findings of this section.

Operational transformation

- client-server vs. peer to peer model - All articles about Operational Transformation used in this thesis suggest the client - server model for this technique. The main reason why this technique can not be implemented with peer to peer model is the fact, that client has to send operations that come from a state in server's history. This would imply that every peer would have to preserve history of every other peer to be able to send such operations. This problem already occurred in the design of Operational Transformation, and was resolved by moving the responsibility to track the history from server to clients.
- requirements local vs. remote - In this technique are requirements on local machine slightly higher than on remote server. As mentioned above, client has to track server's history and only particular type of operations. Server solves only the basic diamond problem and sends out operations that it receives.
- amount of data transfer - Operational Transformation sends only operations over the network. Only in network failures it is necessary to transfer the whole document.
- network requirements - This technique assumes the use of connection over TCP protocol, in order to ensure the delivery. Although the mechanism should be able to detect lost packets, because server has to acknowledge operations.
- time complexity - Time complexity for consistency maintenance is linear. For other more specific and more difficult functions, like undo can be the complexity non-linear.
- conflict resolution ability -
- latency tolerance - As it was mentioned above, server has to ac-

3. COMPARISON OF TECHNIQUES

knowledge user operations. This means that the chance that user operations will get lost is very small. This also implies that server preserves a version of the document on which all operations were applied, in case of inconsistency the whole document can be transferred to user.

4 Komodo

4.1 Overview

4.2 Komodo Requirements for Real-time Collaboration

4.3 Best Technique for Komodo

4.4 Java Implementation

5 Conclusion

Bibliography

- [1] AHMED-NACER, M., IGNAT, C.-L., OSTER, G., ROH, H.-G., AND URSO, P. Evaluating crdts for real-time document editing. In *Proceedings of the 11th ACM Symposium on Document Engineering* (New York, NY, USA, 2011), DocEng '11, ACM, pp. 103–112.
- [2] ELLIS, C. A., AND GIBBS, S. J. Concurrency control in groupware systems. *SIGMOD Rec.* 18, 2 (June 1989), 399–407.
- [3] FRASER, N. Differential synchronization. In *Proceedings of the 9th ACM Symposium on Document Engineering* (New York, NY, USA, 2009), DocEng '09, ACM, pp. 13–20.
- [4] LETIA, M., PREGUIÇA, N. M., AND SHAPIRO, M. Crdts: Consistency without concurrency control. *CoRR abs/0907.0929* (2009).
- [5] LEUNG, C. Operational transformation in cooperative software systems. *McGill Science Undergraduate Research Journal*, 8 (2013).
- [6] LEVENSHTAIN, V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (1966), 707.
- [7] NICHOLS, D. A., CURTIS, P., DIXON, M., AND LAMPING, J. High-latency, low-bandwidth windowing in the jupiter collaboration system. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology* (New York, NY, USA, 1995), UIST '95, ACM, pp. 111–120.
- [8] OSTER, G., URSO, P., MOLLI, P., AND IMINE, A. Data consistency for p2p collaborative editing. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (New York, NY, USA, 2006), CSCW '06, ACM, pp. 259–268.

-
- [9] PREGUIÇA, N., MARQUES, J. M., SHAPIRO, M., AND LETIA, M. A commutative replicated data type for cooperative editing. *2013 IEEE 33rd International Conference on Distributed Computing Systems 0* (2009), 395–403.
 - [10] SHAPIRO, M., AND PREGUIÇA, N. Designing a commutative replicated data type. Research Report RR-6320, 2007.
 - [11] SHAPIRO, M., PREGUIÇA, N., BAQUERO, C., AND ZAWIRSKI, M. A comprehensive study of Convergent and Commutative Replicated Data Types. Research Report RR-7506, Jan. 2011.
 - [12] SPIEWAK, D. Understanding and applying operational transformation, 2015.
 - [13] SUN, C., JIA, X., ZHANG, Y., YANG, Y., AND CHEN, D. Achieving convergence, causality preservation, and intention preservation in real-time cooperative editing systems. *ACM Trans. Comput.-Hum. Interact.* 5, 1 (Mar. 1998), 63–108.
 - [14] VIDOT, N., CART, M., FERRIÉ, J., AND SULEIMAN, M. Copies convergence in a distributed real-time collaborative environment. 171–180.
 - [15] WANG D., MAH A., L. S. Wave protocol google code, 2010.
 - [16] WEISS, S., URSO, P., AND MOLLI, P. Logoot: A Scalable Optimistic Replication Algorithm for Collaborative Editing on P2P Networks. In *29th IEEE International Conference on Distributed Computing Systems - ICDCS 2009* (Montreal, Canada, June 2009), 2009 29th IEEE International Conference on Distributed Computing Systems, IEEE, pp. 404–412.