Efficient Renaming in Conflict-free Replicated Data Types (CRDTs)

Matthieu Nicolas
matthieu.nicolas@loria.fr
Université de Lorraine, CNRS, Inria, LORIA, F-54500
Nancy, France

Gérald Oster
gerald.oster@loria.fr
Université de Lorraine, CNRS, Inria, LORIA, F-54500
Nancy, France

Olivier Perrin
olivier.perrin@loria.fr
Université de Lorraine, CNRS, Inria, LORIA, F-54500
Nancy, France

Abstract

To achieve high availability, large-scale distributed systems have to replicate data and to minimise coordination between nodes. The literature and industry increasingly adopt Conflict-free Replicated Data Types (CRDTs) to design such systems. CRDTs are data types which behave as traditional ones, e.g. the Set or the Sequence. However, compared to traditional data types, they are designed to support natively concurrent modifications. To this end, they embed in their specification a conflict-resolution mechanism.

To resolve conflicts in a deterministic manner, CRDTs usually attach identifiers to elements stored in the data structure. Identifiers have to comply with several constraints such as uniqueness or being densely ordered according to the kind of CRDT. These constraints may prevent the identifiers' size from being bounded. As the number of the updates increases, the size of identifiers grows. This leads to performance issues, since the efficiency of the replicated data structure decreases over time.

To address this issue, we propose a new CRDT for Sequence which embeds a renaming mechanism. It enables nodes to reassign shorter identifiers to elements in an uncoordinated manner. Obtained experiment results demonstrate that this mechanism decreases the overhead of the replicated data structure and eventually limits it.

Keywords CRDT, real-time collaborative editing, eventual consistency, memory-wise optimisation

ACM Reference format:
Matthieu Nicolas, Gérald Oster, and Olivier Perrin. 2020. Efficient Renaming in CRDTs. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 7 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

- Real-time collaborative text editing
- Operational Transform (OT)
- CRDT [1]

2 Background

To solve conflicts deterministically and ensure the convergence of all nodes, CRDTs relies on metadata. In the context of Sequence CRDTs, two different approaches were proposed, each trying to minimise the overhead introduced. The first one [2–4] attaches fixed size identifiers to each element in the sequence and uses them to represent the sequence as a linked list. The downside of this approach is an ever growing overhead, as it needs to keep removed elements to deal with potential concurrent updates, effectively turning them into tombstones. The second one [5–7] avoids the need of tombstones by instead attaching identifiers from a dense totally ordered set to elements. It is then able to order elements into the sequence by comparing their respective identifiers. However this approach also suffers from an ever-increasing overhead, as the size of such dense totally ordered identifiers is variable and grows over time.

In the context of this paper, we focus on the later approach.

2.1 LogootSplit

Proposed by André et al. [7], LogootSplit is the state of the art of the variable-size identifiers approach of Sequence CRDT. As explained previously, it uses identifiers from a dense totally ordered set to position elements into the replicated sequence.

To this end, LogootSplit attaches identifiers made of a list of tuples to elements. These tuples have four components: 1. the *priority*, which embodies the intended position of the element 2. the *node identifier*, 3. the *node sequence number* and 4. the *offset*, which are combined to make identifiers unique. By comparing identifiers using the lexicographical order, LogootSplit is able to determine of the element's position relatively to others. In this paper, we represent identifiers using the following notation : $\text{priority}_{\text{offset}}^{\text{node id},\text{node seq}}$ where priority is a lowercase letter, node id an uppercase one and both node seq and offset integers.

Instead of storing an identifier for each element of the sequence, the main insight of LogootSplit is to aggregate

(a) Elements with their corresponding identifiers **(b)** Elements grouped into a block

Figure 1. Representation of a LogootSplit sequence containing the elements "helo"

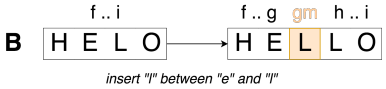


Figure 2. Insertion leading to longer identifiers

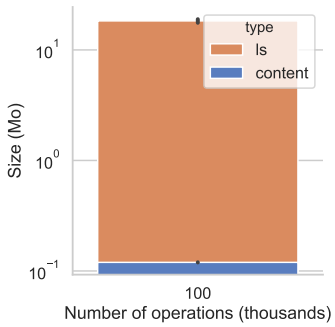


Figure 3. Footprint of the data structure

dynamically elements into blocks. Grouping elements into blocks enables LogootSplit to assign logically an identifier to each element while effectively storing only the block's length and its first element's identifier. LogootSplit gathers into a block elements with *contiguous* identifiers. We call *contiguous* two identifiers which are identical but for their last offset, and with the first one's offset being the predecessor of the second one's. Figure 1 illustrates such a case : in Figure 1a, the elements' identifiers form a chain of contiguous identifiers. LogootSplit is then able to group them into one block to minimise the metadata stored, as shown in Figure 1b.

This feature allows to shift the cause of metadata growth from the number of elements to the number of blocks. As blocks can contain an arbitrary number of elements, it enables to reduce significantly the memory overhead of the data structure.

2.2 Limits

As stated previously, the size of identifiers from a dense totally ordered set is variable. When nodes insert new elements between two others with the same *priority* value, LogootSplit has no other option than to increase the size of the resulting identifiers. Figure 2 illustrates such cases. In this example, since the node inserts a new element between contiguous identifiers, LogootSplit is not able to generate a fitting identifier of the same size. To comply with the intended order, LogootSplit generates the new identifier by appending to the predecessor's one a new tuple.

As a result, the size of identifiers tends to grow as the collaboration progresses. This growth impacts negatively the performances of the data structure on several aspects. Since identifiers attached to values become longer, the memory overhead of the data structure increases accordingly. This also increases the bandwidth consumption as nodes have to broadcast identifiers to others.

Additionally, as the lifetime of the replicated sequence increases, the number of blocks composing it grows as well. Indeed, several constraints on identifier generation prevent nodes from adding new elements to existing blocks. For example, only the block's author can append or prepend elements to it. These limitations cause the generation of new blocks. Since no mechanism to merge blocks a posteriori is provided, the sequence ends up fragmented into many blocks. The efficiency of the data structure decreases as each block introduces its own overhead.

In our benchmark, we measure that these issues eventually lead to the content representing a fraction of the whole data structure's size, less than 1%, as shown in Figure 3. It is thus necessary to address them.

2.3 Related works

Over the years, several works were presented to reduce the growth of variable-size identifiers. However, to the best of our knowledge, no works have been presented to decrease the number of blocks generated.

In [8, 9], authors design for Treedoc [5] a renaming mechanism to reassign shorter identifiers to elements. Nodes rely on a consensus mechanism to trigger the renaming and a catch-up protocol to handle concurrent updates. Since consensus algorithms are costly in large-scale distributed systems with churn, Letia et al. [8] introduce a two-tier architecture. Nodes are splitted between the *core*, a set of stable and highly connected nodes, and the *nebula* made of the remaining ones. Every node can update the sequence but only members of the *core* participate in the consensus leading to renaming. However the *core-nebula* approach is not suited to all kinds of applications. In fully distributed systems, there is no central authority to provide the set of stable nodes acting as the *core*.

In [10, 11], Nédelec et al. introduce another approach to address the identifiers growth issue : LSEQ. Its insight consists in using several strategies to select the *priority* value of a new tuple. These strategies prevent nodes from quickly saturating the space between two existing identifiers. This approach enables to reduce the growth of identifiers. And although the LSEQ approach has been designed for Logoot [6], it can also be applied to LogootSplit. However, its impact would be diminished as insertions into blocks result in new identifiers with increased sizes nonetheless.

3 Proposed approach

We propose a new Sequence CRDT belonging to the variable-size identifiers approach : *RenamableLogootSplit*.

To address the limitations of LogootSplit, we embed in this data structure a renaming mechanism. The purpose of this mechanism is to reassign shorter identifiers to elements and to regroup them into blocks to minimise the memory overhead of the whole sequence.

To avoid costly and not scalable consensus algorithms, we instead adopt the optimistic replication approach for this mechanism. Nodes perform renaming without any coordination. However, this operation is not intrinsically commutative with others. If conflicts arise, we use operational transformations to enable nodes to resolve them deterministically. While transforming *insert* and *remove* operations against *rename* one is straightforward, dealing with concurrent *rename* operations is more tedious. For the sake of simplicity and brevity, we focus in this paper only on the first case. We leave the presentation and evaluation of required additional steps to handle the concurrent *rename* operations to a future work.

3.1 System Model

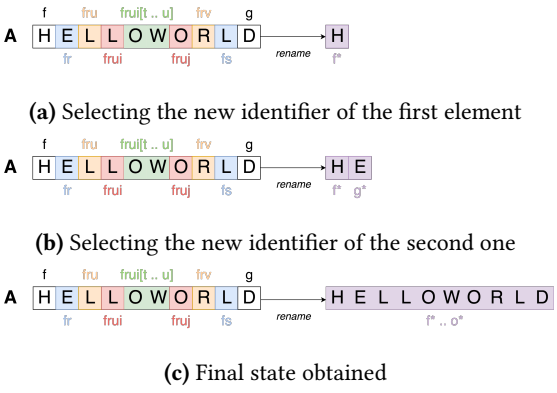
The system is composed of a dynamic set of nodes, as nodes join and leave dynamically the collaboration during its lifetime. Nodes collaborate to build and maintain a sequence using RenamableLogootSplit. Each node owns a copy of the sequence and edit it without any kind of coordination with others.

Nodes communicate through a Peer-to-Peer (P2P) network, which is unreliable. Messages can be lost, re-ordered or delivered multiple times. The network is also vulnerable to partitions, which split nodes into disjoined subgroups. To overcome the failures of the network, nodes rely on a message-passing layer. As RenamableLogootSplit is built on top of LogootSplit, it shares the same requirements for the operation delivery. This layer is thus used to deliver messages to the application exactly-once. The layer also ensures that *remove* operations are delivered after corresponding *insert* operations. Nodes use an anti-entropy mechanism to synchronise in a pairwise manner, by detecting and re-exchanging lost operations.

3.2 *rename* operation

RenamableLogootSplit enables nodes to reduce the overhead of their replica by the means of a new operation : the *rename* operation. This operation reassigns arbitrary identifiers to elements.

Its behavior is illustrated in Figure 4 and can be described as follow : 1. It reuses the id of the first element of the sequence, but modified with the node's own id and sequence number (Figure 4a) 2. It generates contiguous identifiers for all following elements (Figure 4b). As we assign contiguous identifiers to all elements of the sequence, we can eventually group them into one block as illustrated in Figure 4c. It allows nodes to benefit the most from the block feature and to minimise the overhead of the resulting state.



(a) Selecting the new identifier of the first element

(b) Selecting the new identifier of the second one

(c) Final state obtained

Figure 4. Renaming the sequence

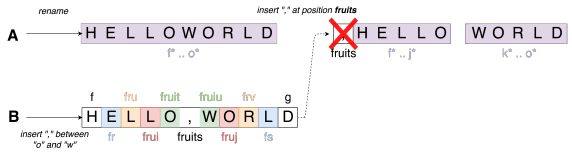
In order for the system to eventually converge, other nodes have to rename their state identically. To achieve this, the node issuing the *rename* operation broadcasts its former state to others. Using the former state, others compute the new identifier of each renamed identifier. However, other nodes' states may contain concurrently inserted identifiers. We will explain in subsection 3.3 how to rename them deterministically.

Broadcasting the *rename* operation embedding the former state may be quite bandwidth consuming since the size of identifiers and the number of blocks are not bounded. To partially adress this issue, we propose a compression mechanism which sends only the necessary components to identify uniquely a block instead of whole identifiers. This compression mechanism allows to reduce to a fixed amount per block the metadata to broadcast.

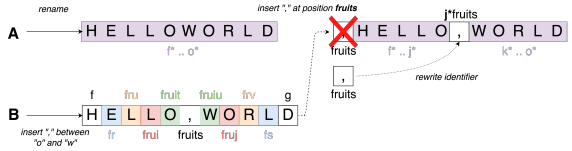
3.3 Dealing with concurrent updates

As *rename* operations can be issued without any kind of coordination, it is possible for other nodes to perform updates concurrently. Since identifiers are modified by the *renaming* mechanism, applying concurrent updates as they are would result in inconsistencies as illustrated in Figure 5a. It is thus necessary to handle concurrent operations to *rename* operations in a particular manner.

To detect them, we use an *epoch-based* system. We add an *epoch* to the sequence as a property. Each time a *rename* operation is applied, the sequence progresses to a new epoch. When nodes issue operation, they tag them with their current epoch. Upon the reception of an operation, nodes compare their current epoch to the operation's one. If they differ, nodes have to transform the operation before applying it.



(a) Concurrent update leading to inconsistency



(b) Fixed scenario

Figure 5. Dealing with concurrent updates

To transform an operation, nodes use the algorithm described in Algorithm 1. This algorithm enables nodes to transform identifiers against the *rename* operation. The main idea of this algorithm is to use the predecessor of the given identifier to do so. The algorithm consists mainly in 1. finding the predecessor of the given id in the former state 2. computing its counterpart in the renamed state 3. prepending it to the given id to generate the renamed id. An example of its usage is illustrated in Figure 5b.

Nodes applying remote *rename* operations use the same algorithm to rename identifiers from their state which do not appear in the propagated state, i.e. identifiers which has been inserted concurrently to the renaming.

Since nodes rely on the former state to transform concurrent operations to a *rename* operation to preserve their semantics, nodes has to store it. Nodes need it until each of them can not longer issue concurrent operations to the corresponding *rename* operation. In other words, nodes can safely garbage collect the former state once the *rename* operation became causally stable [12]. Meanwhile, nodes can offload it onto the disk as it is only required to handle concurrent operations.

Algorithm 1 Rename position

function RENAMEPOS($pos : P, newId : \mathbb{I}, newSeq : \mathbb{N}, renamedBlocks : \{b_i \in B\}_{i \in \mathbb{N}}$): P
  $firstPos \leftarrow posBegin(renamedBlocks[0])$
  $lastPos \leftarrow posEnd(renamedBlocks[renamedBlocks.length - 1])$

  $newPriority \leftarrow priority(firstPos)$
  $newFirstPos \leftarrow new\ P(newPriority, newId, newSeq, 0)$
  $newLastPos \leftarrow new\ P(newPriority, newId, newSeq, length)$

  if $firstPos < pos$ **and** $pos < lastPos$ **then**
    $predecessor \leftarrow findPredecessor(pos, renamedBlocks)$
    $indexOfPredecessor \leftarrow findIndex(predecessor, renamedBlocks)$
    $newPredecessor \leftarrow new\ P(newPriority, newId, newSeq, indexOfPredecessor)$
    return $concat(newPredecessor, pos)$
  else if $lastPos < pos$ **and** $pos < newLastPos$ **then**
    return $concat(newLastPos, pos)$
  else if $newFirstPos < pos$ **and** $pos < firstPos$ **then**
    $predecessorOfNewFirstPos \leftarrow new\ P(newPriority, newId, newSeq, -1)$
    return $concat(predecessorOfNewFirstPos, pos)$
  else
    return pos ▷ Return the position unchanged as it does not interfere with the renaming
  end if
end function

4 Evaluation

4.1 Simulations and benchmarks

To validate the proposed renaming mechanism, we performed an experimental evaluation to measure its performances on several aspects: 1. the size of the data structure 2. the integration time of *insert* and *remove* operations 3. the integration time of the *rename* operation. In cases 1 and 2, we use LogootSplit as the baseline data structure to compare results.

Since we were not able to retrieve an existing dataset of traces of realtime collaborative editing sessions, we ran simulations to generate traces to evaluate our data structure. The simulations depict the following scenario: several authors collaborate in order to write an article. Initially, they prioritise adding content as everything remains to be done. Thus they mainly insert elements into the document during this first phase. A few *remove* operations are still issued to simulate spelling mistakes. Once the document approaches the critical length, the collaborators switch to the second phase. From this point, they stop adding new content and focus on revamping existing parts instead. This is simulated by balancing the ratio between *insert* and *remove* operations. Each author has to perform a given number of operations and the collaboration ends once all of them received all operations. We take snapshots of the document at given steps of the collaboration to follow the evolution of the document.

We ran these simulations with the following experimental settings : we deployed 10 bots as separate Docker containers on a single workstation. Each container corresponds to a single mono-threaded Node.js process (version 13.1.0) simulating an author. The bots share and edit collaboratively the document using either LogootSplit or RenamableLogootSplit according to the session. In both cases, each bot performs an *insert* or a *remove* operation locally every 200 ± 50ms. During the first phase, the probabilities for each operation of being an *insert* or a *remove* are respectively of 80% and 20%. Once the document reaches 60k characters (around 15

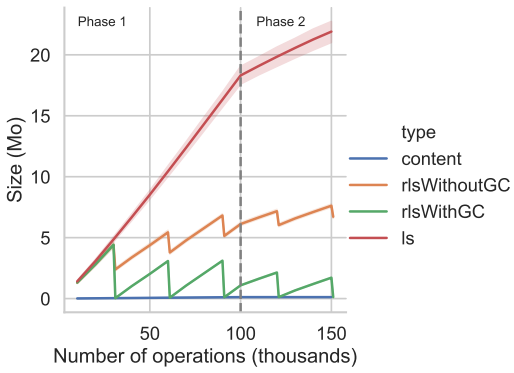


Figure 6. Evolution of the size of the document

pages), the probabilities are both set to 50%. The generated operation is then broadcast to others using a P2P full mesh network. After issuing an operation, there are 5% of chances that the bot moves its cursor to another position in the document. Each bot performs 15k operations. Snapshots are taken every 10k operations overall. Additionally, in the case of RenamableLogootSplit, one bot is arbitrarily designated as the master. It performs *rename* operations every 30k operations overall.

The code of the simulations is available at the following address: https://github.com/coast-team/mute-bot-random/. This repository also contains the code corresponding to the benchmarks described in the next subsections as well as the results computed.

Meanwhile, our implementation of LogootSplit and RenamableLogootSplit are available at https://github.com/coast-team/mute-structs. Both implementations use an AVL Tree, a self-balancing binary search tree, to represent the sequence. This data structure enables us to achieve *insert* and *remove* operations in logarithmic time.

4.2 Results

Memory overhead Using the snapshots generated, we compare the evolution of the size of the data structure in collaborative editing session. The results are displayed in Figure 6. On this plot, the blue line corresponds to the size of the content while the red one exhibits the growth of the LogootSplit data structure.

The green line illustrates the growth of the RenamableLogootSplit document in its best-case scenario. In this scenario, *rename* operations become stable as soon as they are issued. Former states can then be garbage collected safely, maximising the benefits of the *renaming* mechanism. In this case, we observe that *rename* operations reset the overhead of the data structure and eventually reduce by hundred times the document size compared to LogootSplit equivalent one.

As for the orange line, it represents RenamableLogootSplit worst-case scenario. Here, we assume that *rename* operations never become stable and that nodes have to store former states forever. However, obtained results show that RenamableLogootSplit outperforms LogootSplit and reduce by 66% the size of the data structure, even in this case. This outcome is explained by the fact that the AVL does not only store the content and blocks corresponding to the sequence. Some metadata is actually added to the state to browse the sequence more efficiently when performing updates. When a *rename* operation is applied, nodes only keep the sequence of blocks from the former state as an array to be able to transform concurrent operations. Other metadata is scrapped, which results in this memory gain.

Integration times of standard operations We set up benchmarks to measure the impact of the renaming mechanism on the integration times of *insert* and *remove* operations. The obtained results are presented in Figure 7.

Figure 7a displays the integration times of local operations while Figure 7b exhibits remote ones. In both cases, the orange boxplots correspond to LogootSplit's integration times while blue ones to RenamableLogootSplit's ones. The results show that the *renaming* mechanism allows to reduce the integration times of future operations.

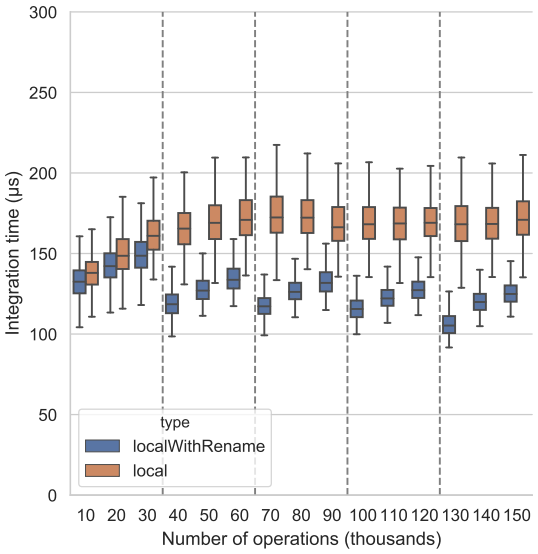
In Figure 7b, the green boxplots display the integration times of concurrent operations to a *rename* one. As illustrated in subsection 3.3, these operations require to be transformed before being applied to the renamed state. The results presented here show that this is actually faster than applying them directly on the former state.

Integration time of* rename *operation Finally, we measured the integration time of the *rename* operation according to the size of the document. Results are displayed in Figure 8. In this figure, the blue line corresponds to the integration time of a *local* rename operation while the orange one corresponds to the integration time of a remote one.

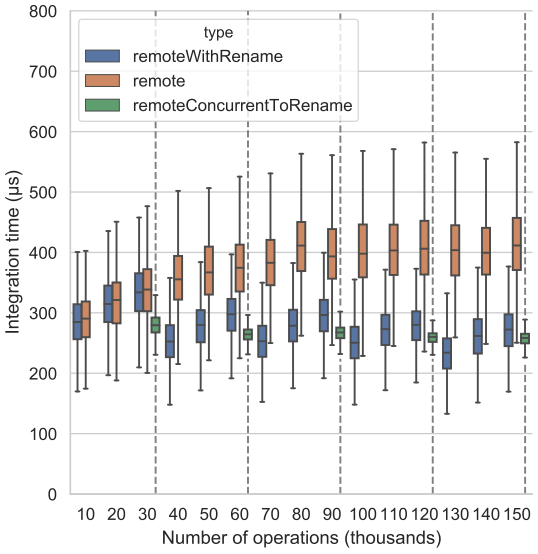
The main result of this benchmark is that the unit of time used when applying *rename* operations is in hundreds of milliseconds. However other operations can not be integrated during the processing of *rename* operations : remote operations won't be displayed to the user while local ones won't be propagated to others. *Rename* operations can thus be perceived as spikes of latency by users and degrade their experience if they are too long to process. It is necessary to take this concern into account when designing the strategy used to trigger *rename* operations to avoid such cases.

5 Conclusions and future work

- Designed new dense identifier-based Sequence CRDT embedding a renaming mechanism : *RenamableLogootSplit*
- Achieved better performances memory-wise than state of the art, even in worst-case scenario...



(a) Local operations



(b) Remote operations

Figure 7. Evolution of the integration time of standard operations

- ... at the cost of expensive but infrequent *rename* operations
- Study user behavior to set a limit to integration time of *rename* operations
- Present and validate mechanism to handle concurrent *rename* operations
- Design strategies to limit likelyhood of concurrent *rename* operations
- Design strategies to limit overall workload in case of concurrent *rename* operations

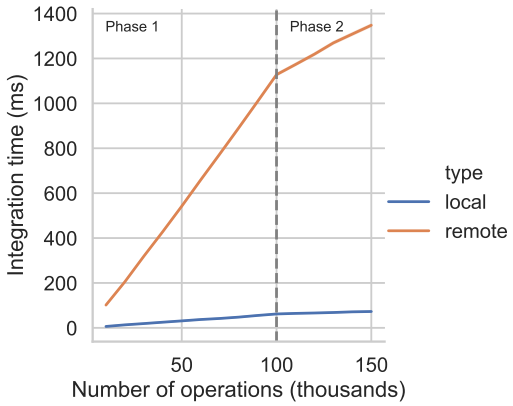


Figure 8. Evolution of the integration time of rename operations

References

[1] Marc Shapiro, Nuno M. Preguiça, Carlos Baquero, and Marek Zawirski. Conflict-free replicated data types. In *Proceedings of the 13th International Symposium on Stabilization, Safety, and Security of Distributed Systems*, SSS 2011, pages 386–400, 2011. doi: 10.1007/978-3-642-24550-3_29.

[2] Gérald Oster, Pascal Urso, Pascal Molli, and Abdessamad Imine. Data Consistency for P2P Collaborative Editing. In *ACM Conference on Computer-Supported Cooperative Work - CSCW 2006*, Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, pages 259 – 268, Banff, Alberta, Canada, November 2006. ACM Press. URL https://hal.inria.fr/inria-00108523. http://portal.acm.org/.

[3] Hyun-Gul Roh, Myeongjae Jeon, Jin-Soo Kim, and Joonwon Lee. Replicated abstract data types: Building blocks for collaborative applications. *Journal of Parallel and Distributed Computing*, 71(3):354 – 368, 2011. ISSN 0743-7315. doi: https://doi.org/10.1016/j.jpdc.2010.12.006. URL http://www.sciencedirect.com/science/article/pii/S0743731510002716.

[4] Loïck Briot, Pascal Urso, and Marc Shapiro. High Responsiveness for Group Editing CRDTs. In *ACM International Conference on Supporting Group Work*, Sanibel Island, FL, United States, November 2016. doi: 10.1145/2957276.2957300. URL https://hal.inria.fr/hal-01343941.

[5] N. Preguica, J. M. Marques, M. Shapiro, and M. Letia. A commutative replicated data type for cooperative editing. In *2009 29th IEEE International Conference on Distributed Computing Systems*, pages 395–403, June 2009. doi: 10.1109/ICDCS.2009.20.

[6] Stéphane Weiss, Pascal Urso, and Pascal Molli. Logoot : A scalable optimistic replication algorithm for collaborative editing on P2P networks. In *Proceedings of the 29th International Conference on Distributed Computing Systems - ICDCS 2009*, pages 404–412, Montreal, QC, Canada, June 2009. IEEE Computer Society. doi: 10.1109/ICDCS.2009.75. URL http://doi.ieeecomputersociety.org/10.1109/ICDCS.2009.75.

[7] Luc André, Stéphane Martin, Gérald Oster, and Claudia-Lavinia Ignat. Supporting adaptable granularity of changes for massive-scale collaborative editing. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing - CollaborateCom 2013*, pages 50–59, Austin, TX, USA, October 2013. IEEE Computer Society. doi: 10.4108/icst.collaboratecom.2013.254123.

[8] Mihai Letia, Nuno Preguiça, and Marc Shapiro. Consistency without concurrency control in large, dynamic systems. In *LADIS 2009 - 3rd ACM SIGOPS International Workshop on Large Scale Distributed Systems and Middleware*, volume 44 of *Operating Systems Review*, pages 29–34, Big Sky, MT, United States, October 2009. Assoc. for Computing

Machinery. doi: 10.1145/1773912.1773921. URL https://hal.inria.fr/hal-01248270.

[9] Marek Zawirski, Marc Shapiro, and Nuno Preguiça. Asynchronous rebalancing of a replicated tree. In *Conférence Française en Systèmes d'Exploitation (CFSE)*, page 12, Saint-Malo, France, May 2011. URL https://hal.inria.fr/hal-01248197.

[10] Brice Nédelec, Pascal Molli, Achour Mostéfaoui, and Emmanuel Desmontils. LSEQ: an adaptive structure for sequences in distributed collaborative editing. In *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng 2013, pages 37–46, September 2013. doi: 10.1145/2494266.2494278.

[11] Brice Nédelec, Pascal Molli, and Achour Mostéfaoui. A scalable sequence encoding for collaborative editing. *Concurrency and Computation: Practice and Experience*, n/a(n/a):e4108. doi: 10.1002/cpe.4108. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4108. e4108 cpe.4108.

[12] Carlos Baquero, Paulo Sérgio Almeida, and Ali Shoker. Making operation-based crdts operation-based. In Kostas Magoutis and Peter Pietzuch, editors, *Distributed Applications and Interoperable Systems*, pages 126–140, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.