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, collaborative editing

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),* 4 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

2 Background

To solve conflicts deterministically and ensure the convergence of all nodes, CRDTs relies on additional metadata. In the context of Sequence CRDTs, two different approaches were proposed, each trying to minimize the overhead introduced. The first one affixes constant-sized identifiers to each value in the sequence and uses them to represent the sequence as a linked list. The downside of this approach is an evergrowing overhead, as it needs to keep removed values to deal with potential concurrent updates, effectively turning them into tombstones. The second one avoids the need of tombstones by instead attaching densely-ordered identifiers to values. It is then able to order values into the sequence by comparing their respective identifiers. However this approach also suffers from an ever-increasing overhead, as the size of such densely-ordered identifiers is variable and grows over time.

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

2.1 LogootSplit

- Attach dense identifiers to elements
- Group elements with adjacent identifiers into blocks to reduce overhead

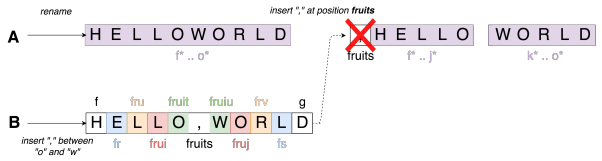
2.2 Limits

- Growth of identifiers
- Increasing number of blocks

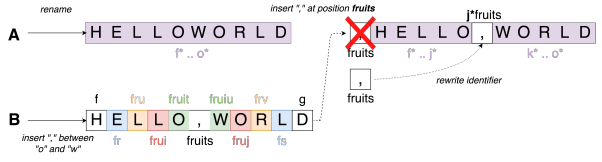
3 Proposed approach

We propose a new Sequence CRDT relying on dense identifiers to order elements : *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 values in such a manner that we are then able to aggregate them into one unique block. This allows to reduce the metadata of the whole sequence. However, as the goal is to reduce LogootSplit's evergrowing memory overhead, we have to



(a) Concurrent update leading to inconsistency



(b) Fixed scenario

Figure 1. Dealing with concurrent updates

design the renaming mechanism while minimizing its own footprint.

3.1 *rename* operation

- Reassign arbitrary identifiers to elements
- Such as elements can be grouped into only one block, independently of the current size of the sequence

3.2 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 1a. 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.

To transform an operation, nodes use the algorithm described in Algorithm 1. This algorithm enables nodes to rename identifiers according to 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 1b.

3.3 Downsides

The proposed approach presents several drawbacks.

The first one concerns the former state. As illustrated in subsection 3.2, nodes has to transform 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 $hasBeenRenamed(pos, renamedBlocks)$ **then**
    $index \leftarrow findIndex(pos, renamedBlocks)$
    return $new\ P(newPriority, newId, newSeq, index)$
  else 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

to a *rename* operation against it to preserve their semantics. To do so, nodes rely on the former state. Nodes has thus to store the former state 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. Meanwhile, nodes can offload it onto the disk as it is only used when dealing with concurrent operations.

The other issue regards the bandwidth consumption. As nodes rely on the former state to compute the renamed one and to handle concurrent operations, it is necessary to propagate it. The node issuing the *rename* operation has thus to broadcast it to other nodes. As the size of identifiers and the number of blocks are not bounded, broadcasting this operation may end up quite bandwidth consuming. 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.

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 the *rename* operation 3. the integration time of *insert* and *remove* operations. In cases 1 and 3, 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 prioritize 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 every 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 pages), the probabilities are both set to 50%. The generated operation is then broadcast to others using a Peer-to-Peer (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.

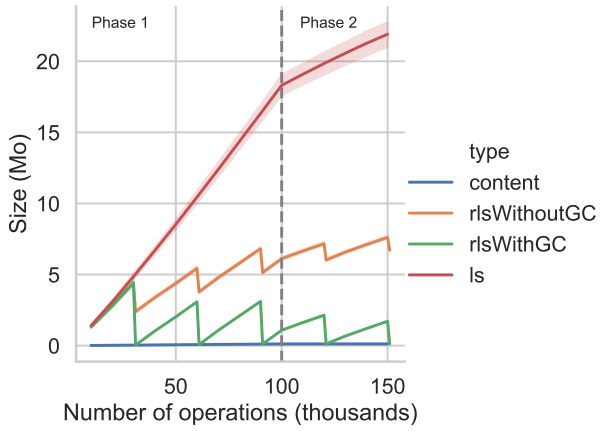


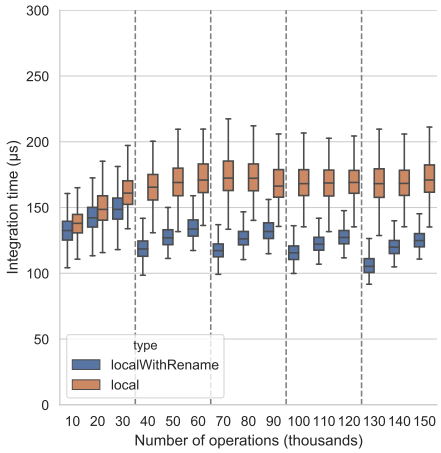
Figure 2. Evolution of the size of the document

4.2 Results

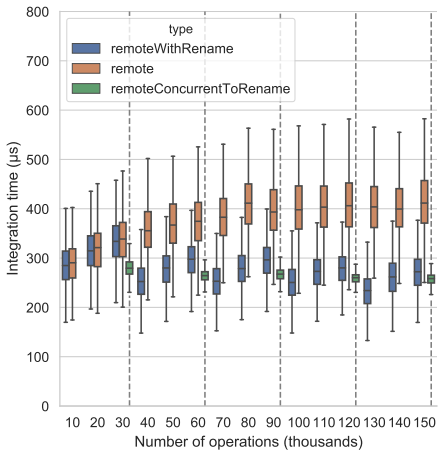
Memory overhead Using the snapshots generated, we compare the evolution of the size of the data structure in collaborative editing session. The results are displayed in Figure 2. On this plot, the blue line corresponds to the size of the content while the red one exhibits the growth of the LogootSplit 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, maximizing 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 has 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



(a) Local operations



(b) Remote operations

Figure 3. Evolution of the integration time of standard operations

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 3.

Figure 3a displays the integration times of local operations while Figure 3b 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 3b, the green boxplots display the integration times of concurrent operations to a *rename* one. As illustrated in subsection 3.2, these operations require to be transformed

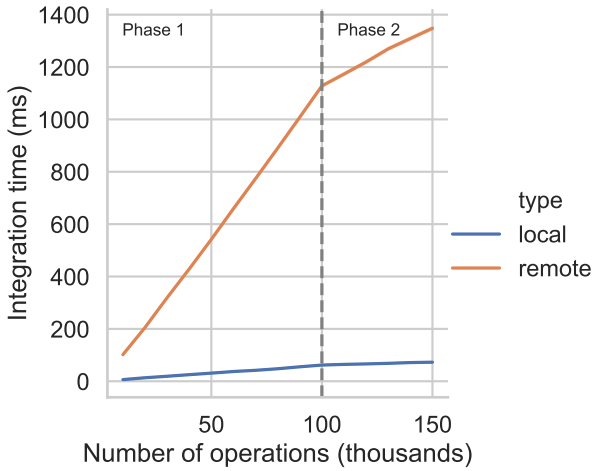


Figure 4. Evolution of the integration time of rename operations

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 4. 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