Rennes, 29 September, 2022

To Whom It May Concern,

I would like to start by thanking you for allowing me to review this Ph.D. thesis submitted by *Matthieu Nicolas in order to obtain a doctorate from the Université de Lorraine.* In the following, I provide a summary of the thesis, discuss some remarks and ideas for improvement, and provide my recommendation. I also provide a non-exhaustive list of more minor comments in the appendix.

Summary of the Manuscript

The Ph.D. thesis focuses on peer-to-peer collaborative editing platforms and proposes solutions to improve their scalability, particularly, by introducing a novel mechanism for reducing the cost associated with Sequence-oriented CRDTs such as those used in collaborative editing. Besides this main contribution, the thesis also includes a very detailed, instructive, and pedagogical state-of-the-art on CRDTs, and a discussion of the tradeoff and technical choices associated with the implementation of the proposed solution into a publicly available prototype. The topics addressed in this thesis are very timely, particularly if we consider the current interest in sovereign software solutions that can provide an alternative to the centralization of proprietary cloud-based applications.

The manuscript starts with an introduction motivating the need for decentralized applications and discussing the importance of the CRDT abstraction in this context. Chapter 2 then presents a detailed and comprehensive survey of CRDTs covering both the set and the sequence data types. Regarding sequence-oriented CRDTs, the chapter does a very good job of discussing available solutions and highlighting both the advantages and the limitations of the approaches based on densely-ordered identifiers. The chapter also presents an overview of Logoot-split, which constitutes the basis for the contributions of this thesis. Then it highlights its main limitation: the inevitable growth of metadata with respect to content. In particular, it observes that after some time the content of the data structure ends up representing only 1% of its required storage space, the remaining 99% being required by metadata. The chapter concludes by presenting the existing approaches for mitigating the cost of sequence-oriented CRDTs and by outlining the solution proposed in this thesis.

Chapter 3 presents the main contribution of this thesis: RenamableLogootSplit, a novel CRDT that augments existing CRDTs based on densely-ordered identifiers with a renaming operation that can reorganize and group the elements of a sequence. This renaming operation makes it possible to reduce the metadata associated with a sequence without changing its content. Renaming identifiers during operation presents significant challenges as new identifiers are continuously being generated. The chapter describes the candidate's original solution incrementally by first describing how renaming operations

Inria centre at Rennes University
Campus universitaire de Beaulieu
Avenue du Général Leclerc
35042 Rennes Cedex - France
Phone: +33 (0) 2 99 84 71 00

www.inria.fr/en

can handle concurrent insert and delete operations, then showing how to handle concurrent renaming operations issued by different nodes, and finally discussing how to garbage-collect old information. The chapter provides a complexity analysis of the proposed algorithms, and an experimental evaluation based on a simulated data trace. Finally, it concludes with some ideas for possible improvements.

Chapter 4 discusses the implementation of MUTE, a decentralized collaborative editor to which the candidate contributed by implementing both RenamableLogootSplit and several other components. The chapter starts by describing the requirements and architecture of MUTE and then discusses the implementation of the various components, focusing on those contributed by the candidate. These include the implementation of RenamableLogootSplit, and Dotted LogootSplit, as well as a message delivery layer, and a membership layer based on the SWIM membership protocol. In particular, the chapter discusses some changes made by the candidate in his implementation of the SWIM membership protocol. The first consists of a change to the ordering relation that controls SWIM's failure detection mechanism. The second involves the addition of an anti-entropy mechanism that allows nodes to discover new nodes. Finally, the chapter briefly presents MUTE's network and security components and highlights possible improvement directions.

Chapter 5 concludes the thesis by summarizing the contributions and outlining directions for future work.

As I mentioned above, I found the work presented in this thesis very relevant and timely. The main contribution is original and interesting as confirmed by its publication in the IEEE Transactions on Parallel and Distributed Systems. The state-of-the-art chapter provides a very nice introduction to CRDTs and to the challenge addressed in this thesis. Finally, the discussion of the implementation highlights some additional challenges and contributions. Both Chapter 2 and Chapter 4, therefore, constitute, in my opinion, important contributions of this thesis, and with some additional effort could result in publishable work.

Major Remarks and Questions

The introduction could be a little clearer in describing the contributions and providing a global picture. In particular, in Section 1.2 it is not clear which of the cited papers are the candidate's and which are related work.

Chapter 2, albeit very interesting contains some unclear sections (see details in the appendix). But what I found most disturbing is that the discussion about it Figure 2.28 states that metadata represents 99% of the storage space after some time, but it seems to me that this is the case even after less than 10 operations. Is the plot correct?

Chapter 3 contains the main contribution. Here I regretted the fact that the algorithms come with informal justifications and with an empirical evaluation but with no proofs. I understand that in the context of systems research, proofs are not very common, but I have the impression that the description of the algorithms in this thesis would benefit from the presence of a clear specification and proof. I also

found it unfortunate that there is no evaluation of the several optimizations and improvements proposed in Section 3.5, which remain mostly ideas for future work.

Second, I found the pseudocode particularly hard to read and not sufficiently discussed in the text. This may be because I am not an expert in CRDTs, but for example, it took me a while to understand that the variable pos in Algorithm 2 is element pos of the first tuple in the identifier $firstId$. A notation like firstId.pos would have been much easier to understand. Similarly, it took me a while to understand why the offset of the tuple generated at line 20 of Algorithm 2 is i, the index of id_i. I would appreciate seeing a clearer explanation of the algorithm since the Ph.D. thesis does not have a page limit like a published paper.

I also found section 3.2.3 particularly hard to read. I suggest referring directly to examples from the figures when describing the bullet points on page 79 and the management of cases that follows. I only understood those clearly when I got to the examples at the end of the section. Like before the pseudo-code and the description of the algorithm require some effort to follow. A clearer although possibly more verbose explanation would have been helpful.

Some descriptions could also be improved by restructuring the text, for example, the last paragraph of Section 3.3 states that the detection of causally stable epochs relies on the use of a group-membership protocol. It would have been better to state this requirement up front. Moreover, I could not help wondering about the impact of the membership protocol on the example in Figure 3.11.

Regarding Section 3.5, I appreciated the analysis of best and worst cases for memory consumption. Yet, I wonder if it wouldn't be possible and better to evaluate the actual ability of nodes to detect causally stable operations. Also, I was expecting Section 3.5 to evaluate the gain in storage space employed for metadata with respect to what was shown in Figure 2.28.

Finally, it would have been nice to have an empirical evaluation of the design choices and improvements presented in Chapter 4.

Conclusion and Assessment

To conclude, I believe this thesis comprises valuable contributions that go beyond what has currently been published. So, despite some minor presentation issues that I recommend the candidate address, this thesis constitutes a good manuscript. ***I am therefore in favor of the defense of this thesis by Matthieu Nicolas in order to obtain a doctorate from the Université de Lorraine.***

Best Regards

Davide Frey

APPENDIX: Minor Comments and typos.

page 13: Definition 10: need citation.

page 15: you should explain why you assume that a read is invoked after each add or remove.

page 23: is [44] a good citation in this context?

page 24: what is a Pure CRDT (synchronized by operation) [41]?
This part is unclear, first, you say you need two functions but then you say you can omit the first, but it's not clear when.

Page 26: what is a " mécanisme d'instantané"? is it a snapshot? A citation is needed here.

"Cependant, ce patron repose sur l'hypothèse d'une livraison causale des opérations et n'est donc pas optimal." Clarify what you mean by this.

page 27: unclear how delta is a state. A difference between states is not a state as far as I can tell

page 28: "convergence forte": recall the definition (Def. 15)

I do not understand this statement: le couple ⟨etats du CRDT,couche livraison⟩ qui forme un sup-demi-treillis dans le cadre de ce modèle de synchronisation. What does it mean for a bcast protocol to be part of a lattice?

page 33: the text after Def. 23 is unclear. You should say clearly why causal delivery does not scale.

page 50: the text mentions the concept of "dot" but this is never defined before Chapter 4.

Section 2.4.5 should be made clearer.

The synthesis in Section 2.6 ignores refs [82] and [83].

page 83: footnote 25. Missing full stop before "Ainsi"

It is unclear to say that the dataset is generated by simulation, say that it is a simulated execution maybe.

The distinction between local and remote operations is unclear in figure 3.13 and its description in the text. What is taken in into account in distant operations? Is dissemination to other nodes included? Or are distant operations those that have been issued by other nodes?

It is not clear where rename operations take place in Figure 3.13. Is it after 30k operations? This is never stated clearly.

The last sentence that starts on page 95 appears out of place. It repeats what was already said without adding any content.

I am not sure I understand the suggestion outlined in 4.3.1 of adding a meta-CRDT that controls the integration of additional CRDTs into a given shared document. How would these multiple CRDTs and the meta CRDT be visible to the user?

The use of the term state "etat" in the description of the anti-entropy mechanism generates confusion. First I understood the mechanism had the objective to discover new nodes and update node-membership lists, then the mention of state suggested that it is also used to synchronize the state of the shared document, but the last sentences confirmed my first interpretation.

The Synthesis of Section 4.3.2 mentions a mechanism whose spatial complexity is O(n), but the section makes no mention of spatial complexity. The only other mention of spatial complexity refers to the epoch-based delivery mechanism discussed in Section 4.4.3.

Regarding the delivery layer, I understood it implements the same guarantees for both protocols but the original version of LogootSplit does not have the same requirements. Did you use the same delivery layer for both protocols? Does this have an impact on the results?

4.4.1 Mentions the possibility of removing FIFO delivery by replacing the version vector with an internal version vector recording data for intervals (check). Would this lead to any advantage for MUTE?

Typos:

ne le fassent ressurgirent -> ressurgir
sup-demi-trellis instead of sup-demi-treillis,
on page 56: rmv abrege rmv

SIÈGE
Domaine de Voluceau
Rocquencourt – B.P. 105
78153 Le Chesnay – France
Phone: +33 (0)1 39 63 55 11
www.inria.fr