

Reviewer's opinion
on Ph.D. dissertation authored by

Jan Konczak

entitled:

Recovery Algorithms in State Machine Replication with Volatile and Non-volatile Main Memory

1. Problem and its impact

The thesis comprises an in-depth study of how the well-known Paxos distributed consensus protocol can be used and extended for implementing the State Machine Replication (SMR) approach in the crash-recovery fault model. Conceptually, the thesis improves a classical distributed algorithm that is important both in theory and practice. From a “systems” perspective, the work is significant due to the design and implementation of JPaxos and the extensive experimental evaluation of the proposed improvements. Further, the work is timely given the attention non-volatile memory has received in recent years.

2. Contribution

The work advances state of the art in at least three ways: (1) by proposing two new algorithms for recovering replicas when a minority of replicas fail; (2) by proposing two ways for employing non-volatile memory for efficiently support recovery in Paxos under catastrophic failures (e.g., when an arbitrary number of replicas - potentially all of them - fail); and (3) by presenting a detailed experimental evaluation of these proposals in several scenarios, comparing them with the standard way to implement failure recovery on Paxos, and analyzing the benefits and challenges of using non-volatile memory for supporting SMR replicas' recovery.

3. Correctness

In the following, I review each of the thesis chapters with my main impressions about their correctness and readability.

Chapter 1 presents the introduction of the thesis, setting up the context and problem statement. The intro is competent, but I miss a more explicit thesis statement, a list of contributions, and a section with an overview of the remaining of the document.

Chapter 2 describes the system model and some important definitions. Overall, everything is fine, and the content is appropriate. However, after reading the system model, it is still unclear whether a message can be lost forever. Typically, one assumes that after GST, a message sent is received after some time, but the text does not state that.



This chapter contains a couple of typos: byzantine instead of Byzantine and insolvable instead of unsolvable.

Chapter 3 describes the Paxos algorithm and its formulation for state machine replication (MultiPaxos), together with an overview of its implementation on the JPaxos library. The description of the protocols is excellent, showing the candidate mastered the details of Paxos (known to be a very complex protocol). Although brief, the description of the JPaxos implementation is also very good, focusing on the multi-threaded architecture and the many subtleties that need to be considered when implementing a high-performance SMR library.

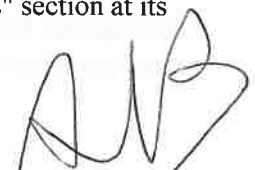
Chapter 4 describes the thesis' related work. The state of the art is clearly explained with sufficient detail. The only issue I have is with the description of DHL+18, in which it is said the work uses the consensus protocol of ABD95. The ABD construction is not a consensus protocol but a register emulation. Further, there is a typo here: ADB should be ABD.

Chapter 5 presents an overview of non-volatile/persistent memory technology. It's a good introduction to the theme, and the most important aspects used in the next chapters are introduced here. I found only two (minor) issues in this chapter: (1) the caption and explanation of Figure 5.3 are unclear about the difference between ordinary and non-temporal, and (2) it would be good to have a small excerpt of code exemplifying how PMDK works.

Chapter 6 contains the main contribution of the thesis. The chapter first describes three recovery algorithms supporting catastrophic failures, FullSS (from Paxos), mPaxos, and mPaxosSM, the latter two specifically designed (by the candidate) to work with persistent memory. Their description is quite detailed and very clear. My only question on this part is if it is true that, given the current limitations of existing persistent memory, mPaxosSM services need to be implemented using atomic transactions. The second part of the chapter describes algorithms that implement recovery as long as a minority of replicas fail but without requiring persistent memory. The critical issue these algorithms need to address is how to deal with stray messages. The two proposed algorithms, ViewSS and EpochSS, solves the problem by exploiting an existing element of Paxos (ballots) and by extending it with epochs, respectively. The chapter concludes with a detailed qualitative analysis of all these protocols, comparing their pros and cons. This analysis is certainly one of the high points of the document.

Chapter 7 presents an extensive experimental evaluation of the proposed algorithms. Two sets of experiments using different hardware are presented, shedding light on the pros and cons of every approach. The methodology is well explained and mostly clear. I would use another workload for the KV-Map experiments (not mixed, only writes), increase the state size (the ones used are too small), and add a setup with seven replicas. However, these are not critical issues, and the results are sound.

I have only minor issues with the chapter. For example, sometimes it is unclear if the candidate used three or five replicas (e.g., Section 7.2.4). I'm assuming it's always three unless stated otherwise. Further, regarding the presentation of the results, the graphs of Figure 7.1 are not very clear, as typical for 3D graphs. Figures 7.3 and 7.11, on the other hand, use a clever way to present the different stages of a replica recovery. The other graphs are common in this line of work and are competent in communicating the results. The chapter would also benefit from a "summary of results" section at its end or highlighted boxes with the main takeaways at the end of each section.

A handwritten signature in black ink, appearing to be 'AWB', located in the bottom right corner of the page.

Chapter 8 concludes the thesis and recalls the main contributions of the work. The only thing missing here is a discussion of potential future work and perspectives on the area.

4. Knowledge of the candidate

As described in previous section, Chapters 3-5 confirm unequivocally that the candidate mastered the subject area of the thesis, i.e., distributed consensus protocols and modern storage technology. The list of references and their discussion is – to the best of my knowledge – both complete and up to date.

5. Other remarks

See Section 3.

6. Conclusion

In the end, the minor typos and less clear text parts of the text do not impair the overall quality of the work submitted. As described in Section 1, the thesis contributions have both theoretical and practical impact, and they are validated through publications in a top journal (IEEE TDSC) and a reputed conference in the area (SRDS'21).

Considering what I have presented above, and the requirements imposed by Article 13 of *the Act of 14 March 2003 of the Polish Parliament on the Academic Degrees and the Academic Title* (with amendments)¹, my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with X)

Definitely YES

Rather yes

Hard to say

Rather no

Definitely NO

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of?

Definitely YES

Rather yes

Hard to say

Rather no

Definitely NO

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

Definitely YES

Rather yes

Hard to say

Rather no

Definitely NO



Alysson Bessani

¹ http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf

