Viewstamed Replication Revisited

Barbara Liskov and James Cowling

Overview

Viewstamped replication (VR) is a replication protocol that handles failures in which the nodes crash. In this paper, the authors propose a revised version of VR that incorporates Paxos in its functioning. The paper first introduces the concept of VR, which was actually first proposed in Barbara Liskov's 1988 paper about the topic. This paper is an improvement on the original, and the paper lists the key areas of improvement, and highlights VR's differences from Paxos. The paper then goes over some assumptions and key definitions that VR is built around. An overview of VR is presented after which the paper goes into detail about the functioning of VR in ideal circumstance. Then the paper discusses view changes, in which a new primary (leader) is chosen in case the old primary crashes. It then discusses the recovery protocol for a crashed replica. After this, the paper discusses actions to be taken in case of non-deterministic operations and the procedure for if a client crashes. Practical tweaks for this protocol are brought up, including modifying the recovery protocol with the help of checkpoints, state transfer to assist nodes that have fallen behind to get updated and a tweak to the view changes protocol. Some optimization strategies are also discussed.

The paper then discusses reconfiguration protocols for changing the number of replicas or any other change in the organization of replica groups. The paper then talks about the safety and progress properties with regards to the three main protocols discussed: view changes, recovery and reconfiguration before reaching its conclusion in which it states that this new version of VR is superior.

Thoughts on the Concepts Presented in the Paper

Reading this paper already knowing of the Paxos protocol made it easy to spot the DNA of Paxos in the three main protocols presented in the paper. The paper itself invites comparisons between VR and Paxos. The normal operation of VR essentially functions like single-decree Paxos, with the primary being the leader, the commands being processed by the leader, and consensus and fault tolerance being achieved by ensuring that the majority of replicas participate in each step. Even though three main protocols are not as heavily influenced by Paxos, their requirement of a quorum is reminiscent of Paxos. Indeed, one of the pillars of VR is the quorum intersection property which is responsible for the size of replica groups.

However, it is just as easy to see how VR is not Paxos. To paraphrase a line from the very first page, 'VR is a replication protocol while Paxos is a consensus protocol'. VR's priority is not consensus even though it is important for the nodes to execute the same command at every state. VR is much more concerned with fault tolerance, and the prioritizing of fault tolerance first and foremost through the paper is immediately noticeable.

Unlike Paxos, VR also involves the client to a much greater degree than Paxos did. Clients in VR have certain restrictions placed on them. For example, in VR clients can only have one outstanding request at a time. As mentioned earlier, client failures are also taken into account.

The three main protocols are discussed in great detail. However, some things remain unclear for me. One instance is in the view change protocol, in which in case a primary crashes, a new node is elected. Firstly, I'm not clear on why a Paxos-style leadership election can't take place, given that all the elements required for it are already present. Secondly, in the protocol in the paper, on Page 5, the paper says that on receiving the message to choose a new primary, a node will send a message to the node that is to be the new primary, without, as far as I can see, mentioning how the replica would know which node is going to be the new primary.

Additionally, in the section covering non-deterministic operations, the paper has brought up a concept that could potentially derail the operations of the protocol. In my opinion, the paper did not devote enough attention to this section, giving only one example regarding operations involving timestamps. I feel certain that this cannot be the extent of non-deterministic operations but am not well-versed enough in this topic yet to provide examples, and I wish the authors of the paper had done this, considering that their proposed solution to this (having the primary predict these values) feels vague and underdone compared to the detailed protocols in the prior sections.

Another observation that I have is that VR, or in any case, this paper, seems more practical and less theoretical, with its prioritization of fault tolerance, the consideration of clients in the process. This paper seems based on fewer assumptions that seem quite clear, for example assuming that messages that nodes send each other are not actually being sent by a malicious party.

Thoughts On This Paper

In my opinion, the organization of this paper is a bit strange. In the heart of the paper, the main protocols are discussed, i.e. the normal functioning, view change and recovery after which tweaks and optimizations are brought up, and then another set of protocols is brought up. I feel that it might have been better for the reconfiguration protocols to discussed before bringing up the pragmatics and optimization strategies.

I also believe that it would have been a vast improvement to have the discussion of safety and progress properties of each protocol to be discussed immediately after introducing that protocol, as was done in the Paxos Made Moderately Complex paper. Similar to that paper, it might have been better to plainly state the invariants so that the worst case of each protocol becomes clear.

The protocols in this paper are presented very methodically and systematically which made them easy to understand. To better understand VR I looked up the original 1988 paper about the same topic, but in comparison to this paper, it was much more difficult to follow. Regarding the original paper, I wonder how my understanding of VR would have changed had the differences between this version and the original version been expounded on more.

Conclusion

This paper was a succinct introduction to the concept of VR. The paper seems to be an improvement on the original paper about VR on all counts. Prior knowledge of Paxos helped illustrate its influence in the various protocols presented in the paper.