Consensus Routing: The Internet as a Distributed System

EECS 589 Paper Review

Shuang Qiu, Buting Ma

September 22, 2016

1 Paper Information

Title: Consensus Routing: The Internet as a Distributed System
Authors: John P. John, Ethan Katz-Bassett, Arvind Krishnamurthy, Thomas Anderson, and Arun Venkataramani.
Venue and Date: 5th USENIX Symposium on Networked Systems Design and Implementation (NDSI), 2008.

2 Summary of Paper

Responsiveness and consistency are both two critical aspects of designing a practical Internet routing protocols. However, many traditional routing protocols, especially interdomain protocols such as BGP, have already paid much more attention to the responsiveness instead of consistency. This will result in the problems of routing loops and black holes because each router does not have a consistent view of the whole system.

This paper argues that, in a distributed system, the consistency ensures its behavior more predictable and securable. And thus, the authors propose a consistency-first approach, named Consensus Routing, to pursue both consistency and responsiveness in the Internet routing. More specifically, regarding consistency and responsiveness as safety and liveness property respectively, consensus routing method separates safety and liveness concerns into two distinct modes of packet delivery, stable mode and transient mode. In the stable mode, the routing is done when all dependent routers agree upon a global state. If there is a problem with a router to possess stable route for a packet, this router will heuristically forward the packet using local information - this is the transient mode. This paper presents promising results of its simulation experiments. It shows that, when failures and policy changes happen, the consensus routing gains significant availability over BGP while ensuring that packets traverse adopted routes and that the overhead from consensus routing is almost negligible.

3 Novelty of Consensus Routing

The first novelty of this paper is that it eventually finds a way to balance both consistency and responsiveness instead of just favoring responsiveness over consistency as the traditional protocols did. This approach can address the issues of routing loops and black holes which can happen in previous responsiveness-first routing protocols. It actually offers another perspective of how to design a practical routing protocols.

The most inspirational point of this paper is that it incorporates some important ideas from distributed system design into developing their new method. Because the authors realize that the consistent state in a distributed system makes its behavior more predictable and securable, which properly matches their motivation in this paper. Consensus routing approach will first run a distributed coordination algorithm to ensure the consistency among all the dependent routers. And the coordination is further based on two classical algorithms in distributed system, namely, distributed snapshot [1] and consensus [2] algorithms. The combination of ideas and knowledge across different research fields can sometimes show great power.

Another very interesting idea of this paper is to apply the two different routing modes, especially the transient mode. The two-mode process makes the novel routing approach much more practical and feasible. This is because although the stable mode can work to ensure global consistency most of the time, the ideal situation may not always hold. Therefore, the transient mode gives another way to route packets in case of a problem. By this means, the liveness is also maintained. From this point of view, the transient mode could be viewed as a kind of complement or backup for the stable mode. This idea can show us how to design a practical system in real-world scenarios.

4 Advantages Not Recognized

One advantage of this consensus protocol is “versioning” on view change. It “quantizes” the network status by snapshot so that you can have a relatively static view of the whole network rather than a dynamic graph that is always in change.

This brings two benefits. First, this discrete view is good for logging. For example, instead of discarding the k^{th} SFT at the end of $(k+1)^{th}$ epoch, we can log the k^{th} SFT, then we are going to have a snapshot of the network status at that moment. However, in the traditional BGP, we may only be able to log the updates and only have a local and incomplete view of the network. Second, this logged snapshot is good for doing analysis, because it is consistent. Hence, we can check the logic to debug or adapt the routing policy to resolving congestion much easier on the basis of the consistent view than a dynamic one in the traditional BGP. Furthermore, this consistent view can be helpful for us to understand and build the realistic model for future research.

5 Shortcomings and Overlooked Points

5.1 Resonance

The key idea of this consensus routing protocol is that the routers will achieve an agreement on a consistent routing scheme on the snapshot (state of Internet at a particular moment) for a period of time, then adapt to changes, and then make the agreement again. Because the routers under this consensus protocol will "change view" at the end of each epoch. Although not synchronized, the routers "make this transition at a slightly different time". In our expectation, this may cause the system to enter a resonant phenomenon described in the paper [3]. Periodically, at each "view change", the understanding of the network changes. And this might incur some unexpected behavior such as a burst in traffic or transient package.

The solution to this problem proposed in the paper [3] is to randomly defer the outgoing adverting message so that the updates will be evenly distributed and will not resonate. However, this solution can not be adopted in the consensus routing, because it may destroy the consensus mechanism.

5.2 Scalability

This novel routing approach of this paper is based on consensus. Nonetheless, the consensus is somehow expensive. The paper argues that the cost brought by consensus is tolerable in their simulation, because only the transit ASes need to be involved in the consensus system and there are only 3000 such ASes. In this simulation, 8% overhead is induced.

In the subsection of the discussion on security, the authors advise byzantine fault-tolerant agreement techniques to exclude the failed or malicious nodes from the snapshot. The author didn't implement this in their simulation experiments. However, if this is applied, the cost of consensus will be $\Theta(n^2)$(three phases, each phase is $\Theta(n^2)$). Hence the cost(e.g. overhead) grows quadratically with the size, and it may not be tolerable any more if the number of transit ASes is too large.

References

[1] K. M. Chandy, and L. Lamport. *Distributed snapshots: Determining global states of distributed systems*. ACM Transactions on Computer Systems, 3(1):63-75, 1985.

[2] L. Lamport. *The part-time parliament*. ACM TOCS, 16(2):133-169, 1998.

[3] Floyd, and Jacobson. *The Synchronization of Periodic Routing Messages*. ACM/IEEE Transactions on Networking, 2(2):122-136, 1994.