NetPaxos: Consensus at Network Speed

Huynh Tu Dang
with
Daniele Sciascia, Marco Canini*,
Fernando Pedone, Robert Soulé

Università della Svizzera italiana (USI)  * Université catholique de Louvain (UCL)
Software-Defined Networking

- Tailors the network to applications’ needs
  - e.g., Route selection, Traffic prioritization, Traffic aggregation, …
Motivation

SDN allows to dynamically reconfigure the network…
But can we go beyond that and co-design the network with the goals of application in mind?
Paxos: Fundamental Building Block of Fault-tolerant Applications

- Consensus among replicas
- Paxos performance is crucial for applications
- Paxos could benefit from increased network support
Paxos Background

- Proposers
- Coordinator
- Acceptors
- Learners
Paxos Background

- Proposers
- Coordinator
- Acceptors
- Learners

Propose a value
Paxos Background

- **Proposers**: Propose a value
- **Coordinator**: Coordinates proposers
- **Acceptors**
- **Learners**
Paxos Background

- **Proposers**: Propose a value
- **Coordinator**: Coordinates proposers
- **Acceptors**: Choose a single value
- **Learners**
Paxos Background

- **Proposers**: Propose a value
- **Coordinator**: Coordinates proposers
- **Acceptors**: Choose a single value
- **Learners**: Learn the chosen value
Basic Paxos

- Proposers
- Coordinator
- Acceptors
- Learners
Basic Paxos

Proposers

Coordinator

Acceptors

Learners

Phase 1 (Prepare Phase)
Basic Paxos

- Proposers
- Coordinator
- Acceptors
- Learners

Phase 1 (Prepare Phase)
Phase 2 (Accept Phase)
Basic Paxos

- Proposers
- Coordinator
- Acceptors
- Learners

Phase 1 (Prepare Phase)
Phase 2 (Accept Phase)
3 message delays
Fast Paxos

- Proposers
- Coordinator
- Acceptors
- Learners

2 message delays
More acceptors
Fast Paxos

- Proposers
- Coordinator
- Acceptors
- Learners

2 message delays
More acceptors
2 type of rounds:
- Fast round
- Classic round
Basic Paxos

- Proposers
- Coordinator
- Acceptors
- Learners

6 message delays

Switch

Switch

Switch

Switch
Fast Paxos

- Proposers
- Coordinator
- Acceptors
- Learners

4 message delays
Approaches to Improve Paxos Performance

- **Approach 1**: Extend SDN switches to implement Paxos logic inside the forwarding devices

- **Approach 2**: Don’t extend SDN switches, but rely on network ordering assumptions to optimize Paxos
Approach 1: Moving Paxos Logic in Forwarding Devices

Proposers

Coordinator

Acceptors

3 message delays

Switch

Learners
Generate round & sequence numbers
Sufficient Set of OpenFlow Extensions

Stateful Comparisons

Proposers

Coordinator

Acceptors

Learners
Sufficient Set of OpenFlow Extensions
Sufficient Set of OpenFlow Extensions
# Feasibility of Paxos in Hardware

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Device</th>
<th>Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round &amp; Sequence Generator</td>
<td>Netronome NFP-6xxx, NetFPGA, Arista 7124FX</td>
<td>stateful flow processing</td>
</tr>
<tr>
<td>Stateful Comparisons</td>
<td>Arista 7124FX</td>
<td></td>
</tr>
<tr>
<td>Persistent Storage</td>
<td>Arista 7124FX</td>
<td>50 GB SSD logging</td>
</tr>
<tr>
<td>Storage Cleanup</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Approach 2: Using Current Switches & Relying on Network Assumptions

- Avoid moving Paxos logic inside forwarding devices
- Rely on packet order assumptions in the network
- Achieve consensus at high speed
NetPaxos

Proposers

Serializer

Minion

Minion

Minion

Minion

Learners
NetPaxos

Proposers

Serializer

Minion

Minion

Minion

Minion

Learners
Packets From the Serializer to Minions Arrive In Order

Important for Performance
Packets From a Minion to Learners & Storage Server Arrive In Order

Important for Correctness
Evaluation

• How frequently are our ordering assumptions violated in practice with off-the-shelf SDN hardware?

• What are the expected performance benefits of NetPaxos?
Experiments

Proposers

Serializer

Minion

Minion

Minion

Minion

Learners
Experiments

Proposers

Pica 8

Learners

Pica 8

Pica 8

Pica 8

Pica 8

Pica 8
Experiments

- Send 500,000 messages (1470-byte long) at a target sending rate
- Range rate from ~ 1k to 70k msg/s
- Messages contain seq. no, proposer id, ts, value
Experiments

- 2 learners receive msgs on 4 NICs
- **deliver** if same seq. no on msgs from 3 NICs
- compare values to see disagreements
Evaluation

• How frequently are our ordering assumptions violated in practice with off-the-shelf SDN hardware?

• Evaluate by measuring:

  • **Indecision**: not every learner can deliver the chosen value
  • **Disagreement**: learners deliver different values
NetPaxos Assumptions Hold Most of The Time
NetPaxos Assumptions Hold Most of The Time

70 % link capacity
NetPaxos Assumptions Hold Most of The Time

A learner cannot decide a value
NetPaxos Assumptions Hold Most of The Time

0% disagreement

learners learn different values

correctness assumption
Evaluation

• What are the expected performance benefits of NetPaxos?

• Compare throughput and (one-way) latency of NetPaxos vs. server-based basic Paxos*

• Quantify performance improvements for a best-case scenario

* [https://bitbucket.org/sciascid/libpaxos](https://bitbucket.org/sciascid/libpaxos)
NetPaxos Performance Improvement

![Graph showing performance improvement of NetPaxos compared to Basic Paxos](image_url)
NetPaxos Performance Improvement

 através de uma ferramenta. A taxa de transmissão melhorou em 9X, e a latência reduziu em 90% no cenário de melhor caso.
Summary

• Beyond management simplification, SDN has potential for increasing application performance

• Two approaches leveraging SDN to improve Paxos performance:
  • Extend OpenFlow to implement Paxos logic inside SDN switches
  • Don’t extend OpenFlow, but rely on network assumptions to implement NetPaxos
Thank you for listening

Tu Dang

http://www.people.usi.ch/danghu/