Consensus Routing: The Internet as a Distributed System

John P. John, Ethan Katz-Bassett, Arivnd Krishnamurthy, Thomas Anderson, Arun Venkataramani

Presented by: Ruba Merza
I Present To You... The Internet! (.. sort of)
What’s wrong with BGP, anyway?

- Causes inconsistency in the network
- Leading to loops, blackholes
Loops!
Blackholes!
Aren’t there existing solutions to these problems?

- Different solutions solve problems in specific scenarios.
- No single mechanism that addresses them all.
Consensus Routing

• “Cleanly separate safety and liveness concerns in policy routing.”

• Safety:
  forwarding tables are always consistent and policy compliant.

• Liveness:
  the system quickly adapts to routing policies/failures in the network.
Safety Mechanism (Stable Mode)

• Routers Running Regular BGP (kind of)
• A distributed snapshot is taken periodically to figure out which BGP updates are incomplete.
• Snapshots are sent to consolidators
• Consolidators run a consensus algorithm (Paxos)
• Consolidators broadcast consistent views of updates back to ASes.
Liveness Mechanism (Transient Mode)

• What happens when a router can’t find a stable route?
• Consensus Routing implements transient forwarding schemes:
  1. Deflection routing
  2. Detour routing
  3. Backup routing
Routing Deflections

- Deflect packet to nearest AS neighbor
- Can’t find one? Backtrack!
- Backtracking has problems :(
Detour Routing

- Select an AS and forward packets to it.
- Now they’re the new AS’s problem, not yours!
Backup Routing

• Use a pre-computed backup route
• One way of doing this: RBGP
  - requires slight modifications to BGP
  - protects against single link failures
Evaluation (Link Failures)

Cumulative fraction of failure cases

Fraction of ASes disconnected

BGP

Loops

Loops/No Route
Evaluation (Link Failures)

![Consensus Routing Graph]

- Cumulative fraction of failure cases
- Fraction of ASes disconnected
- Consensus Routing
- Detour: closest tier1
- Backup route
- Backtrack

14
Overhead (volume of control traffic)
## Overhead (cost of consensus)

<table>
<thead>
<tr>
<th>Number of nodes</th>
<th>Time when first node learns value</th>
<th>Time when last node learns value</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>434 ms</td>
<td>490 ms</td>
</tr>
<tr>
<td>18</td>
<td>485 ms</td>
<td>1355 ms</td>
</tr>
<tr>
<td>27</td>
<td>590 ms</td>
<td>1723 ms</td>
</tr>
</tbody>
</table>
Overhead (path dilation)

![Graph showing cumulative fraction of failure cases vs. average path dilation. The graph includes lines for Backtrack, Backup route, and Detour: closest tier 1.]
Summary

• Consistency is needed when routing packets.
• Consensus routing takes care of both consistency and responsiveness.
• However! There are problems!
• There’s processing overhead on Tier-1 ASes.
• No mention of who’s in control of deciding how long epochs are.
• The custom simulator used isn’t available for the public to see!
• Probably won’t be used in the internet tomorrow :)}