Consensus Routing: The Internet as a Distributed System

John P. John\textsuperscript{1}, Ethan Katz-Bassett\textsuperscript{1}, Arvind Krishnamurthy\textsuperscript{1}, Thomas Anderson\textsuperscript{1}, Arun Venkataramani\textsuperscript{2}

\textsuperscript{1}Dept. of Computer Science, Univ. of Washington, Seattle
\textsuperscript{2}University of Massachusetts Amherst

NSDI 2008

Presented by: Ahmed Khurshid

CS 598 PBG Fall 2010
Advanced Computer Networks
Outline

• Routing pathologies affecting reliability
• Effect of delayed routing convergence
• Existing proposals for improving network reliability
• Consensus Routing – The Internet as a Distributed System
What do we expect from a network?

• **Common expectation**
  – Seamless connectivity from the source to the destination

• **Application specific expectations**
  – End-to-end reliability
  – Sufficient throughput
  – Low latency, jitter, etc.

• However, the network does not always behave the way we want
Routing Pathologies

• Distributed nature of Internet routing results in unpredictable behavior

• Not all the routers have a consistent view of the network all the time
  – Results in delayed routing convergence

• This causes
  – Black holes
  – Routing loops (eventually creating black holes)
  – Sub-optimal routing
Routing loop Examples

- **Link failure** causing BGP loops at 2 and 3
- **Policy change** causing BGP loops at 2 and 3 when 4 withdraws a prefix from 2 and 3 but not 6
Black hole Example

To reach P, AP is preferred over CD

iBGP link recovery causing black hole
Effect of Delayed Routing Convergence (Labovitz et al.)

- Tshort: represents both a route repair and failover
- Tlong: represents both a route failure and failover

Ref: Labovitz et al., “Delayed Internet Routing Convergence”, SIGCOMM 2000
Existing proposals for improving network reliability
Achieving Convergence-Free Routing (Lakshminarayanan et al.)

- A reactive approach
- Packets carry failure information
- Routers compute fault-free path on-the-fly
- No routing update is exchanged among the routers

SafeGuard: Safe Forwarding during Route Changes (Li et al.)

- Packets carry remaining path cost
- Change in path cost indicates change of route
- Approximates the effect of a full source route

Ref: Li et al., “SafeGuard: Safe Forwarding during Route Changes”, CoNEXT 2009
RBGP: Staying Connected In a Connected World (Kushman et al.)

- A proactive approach
- ASes advertise pre-computer backup paths for failover
- Increases processing and control overhead

Ref: Kushman et al., “RBGP: Staying Connected In a Connected World”, NSDI 2007
Consensus Routing
Motivation

• Internet routing protocols (both intra and inter domain) usually favors responsiveness over consistency
  – A new route is incorporated in the forwarding table before propagating the same to neighbors

• Results in routing loops and blackholes

• Usually there is no extra effort to ensure consensus
  – Solutions have been proposed for intra-domain routing
Consensus Routing

• A consistency first approach that cleanly separates safety and liveness of routing
  – Safety: All the routers use a consistent route towards a destination (i.e., no loops)
  – Liveness: Quick reaction to failures and policy changes

• Ensure both consistent behavior and quick reaction
  1. Runs a distributed coordination algorithm to ensure globally consistent view of routing state
  2. Forwards packets using one of two logically distinct modes
Stable Mode

• Consensus routing does not immediately incorporate a newly learned route into the forwarding table

• Periodically, all routers engage in a distributed coordination algorithm

• The coordination is based on
  • Chandy-Lamport snapshot algorithm
  • Paxos

• Output of the coordination is used to compute a set of stable forwarding tables (SFTs) that are guaranteed to be consistent
  • SFTs replace traditional FIBs (Forwarding Information Base)
Stable Mode – Update Log

Store updates into the update log without modifying the SFT
Stable Mode – Distributed Snapshot

Updates in the snapshot may be complete or incomplete
Stable Mode – Aggregation

- Tier-1 ASes are good candidates for being consolidators

Why?

- Better reachability
- Longevity
- Full mesh topology among the ASes

Consolidators

Snapshots

Users

Tier-1

Tier-2

Tier-3 (Stub)
Consolidators run Paxos to agree upon a global view by extracting incomplete updates from the reported snapshots.
Stable Mode – Flood

Message contains the set of incomplete updates (I) and the set of ASes (S) that successfully responded to the snapshot.
Stable Mode

- **SFT Computation**
  - SFT is computed using the global set of incomplete updates (I) and local logs
  - Routes involving ASes not present in S are not placed in the SFT

What happens to those ASes?

How does this strategy achieve consensus in an asynchronous distributed system?
Use of two SFTs

<table>
<thead>
<tr>
<th>Prefix - Y</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k^{th}$ SFT</td>
<td>B-&gt;C-&gt;D</td>
<td>C-&gt;D</td>
<td>D</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>$(k+1)^{th}$ SFT</td>
<td>B-&gt;C-&gt;E</td>
<td>C-&gt;E</td>
<td>E</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>
Transient Mode

• Consensus routing switches to this mode when
  – The next-hop router along a stable route is unreachable
  – A stable route is not available

• Uses several known schemes
  – Routing deflection
  – Detour Routing
  – Backup route
Route Deflection

- After encountering a failed link, deflect the packet to a neighboring AS after consulting RIB.
- If no neighbor can be chosen, then deflect the packet back to the sending AS (backtracking).
  - However, backtracking alone is not sufficient to guarantee reachability.

Limitations of backtracking.
Other Transient Schemes

• Detour Routing
  – After encountering a failed link, select a neighboring AS (arbitrarily) and tunnel transient packets to it
  – Tier-1 ASes are good choices in this selection

• Backup Routes
  – Use pre-computed backup routes to forward packets during failure (e.g., R-BGP)
Evaluation

• Simulation Methodology
  – CAIDA AS-level graphs gathered from RouteViews BGP tables
    • Includes 23,390 ASes and 46,095 links annotated with inferred business relationships of the linked ASes

• Using XORP prototype to measure implementation overhead

• Using PlanetLab nodes to measure the cost of consensus
Link Failure

- One of the links of a multi-homed stub AS is failed during each experiment

Consensus routing provides significantly higher levels of connectivity than BGP

Figure 6: Loops and disconnectivity in BGP following a failure.

Figure 7: Disconnectivity in consensus routing following a failure.
Effect of Traffic Engineering

• Withdraw a subprefix from all but one of the providers (3 or more) of a multi-homed AS

Consensus routing does not affect routing in case of policy changes.
In terms of bandwidth and time, consensus routing incurs little overhead.
Thanks

Questions and Comments?