Cross-Layer Failure Restoration for a Robust IPTV Service

Murat Yuksel (yuksem@cse.unr.edu)
University of Nevada – Reno

K. K. Ramakrishnan (kkrama@research.att.com)
Robert D. Doverspike (rdd@research.att.com)
AT&T Labs Research

IPTV Today

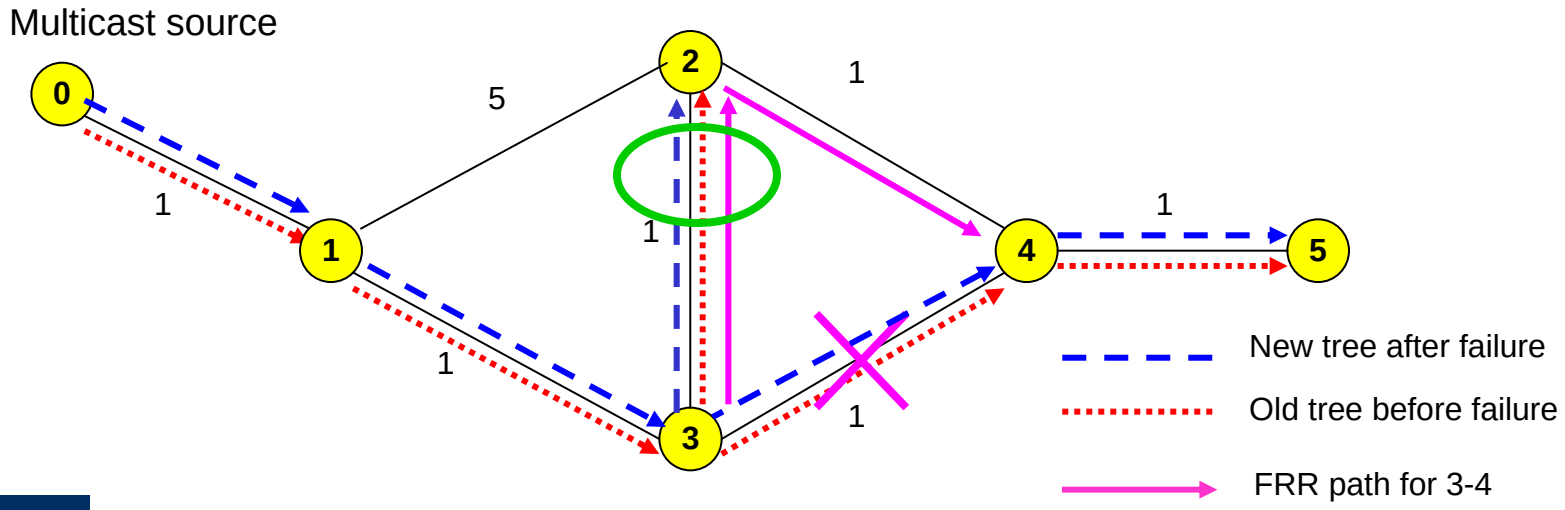
- "Rich Media" applications like IPTV require significant capacity
  - The capacity requirement keeps increasing with more and more TV channels carried over the IP backbone, and metro area network
  - Over 70% of raw link capacity is needed in a typical system
- System typically organized as:
  - a small set of centralized content acquisition sites (head-ends);
  - large number of media distribution sites in metropolitan cities;
  - Redundant set of routers and a number of servers at distribution sites
  - a metro and neighborhood area network to reach the home
- Uses IP multicast for distribution
  - PIM-SSM (source specific mode) is the multicast protocol used
  - Per "channel" tree from source (central acquisition) to receivers
  - Typically a group extends all the way to the consumer

Backbone Failures

- IPTV and other multimedia performance requirements are very stringent
  - E.g., ITU requirements for packet loss probability for video distribution is less than 10^-8
- Failures in a long distance backbone are not rare
- Even multiple failures are not rare..
- Depending solely on Layer 3 Recovery from a failure can take from tens of seconds up to several minutes
  - For example:
    - IGP can take tens of seconds to reconverge
      - Timers are set conservatively, in the interest of stability and scalability
    - PIM typically refreshes (and thus reconverges) its tree on the order of minutes
  - Such recovery times are not tolerable
- Recovery times greater than 50-100 msecs are difficult to treat using FEC and Resilient UDP

Existing Failure Restoration Approaches

- Link-level Fast Re-route (FRR) – *pure layer 2 approach*
  - Idea: Reroute traffic on the backup path of a failing link
  - IGP and PIM are not informed about the failure
  - Pros: Higher layers are not bothered/aware of failure being restored; local decision; fast restoration (primarily failure detection time) ~50 msecs
  - Cons: Traffic overlaps and hence significant loss are possible
    - Overlaps can last a long time (until failure is repaired) – several hours



Existing Failure Restoration Approaches

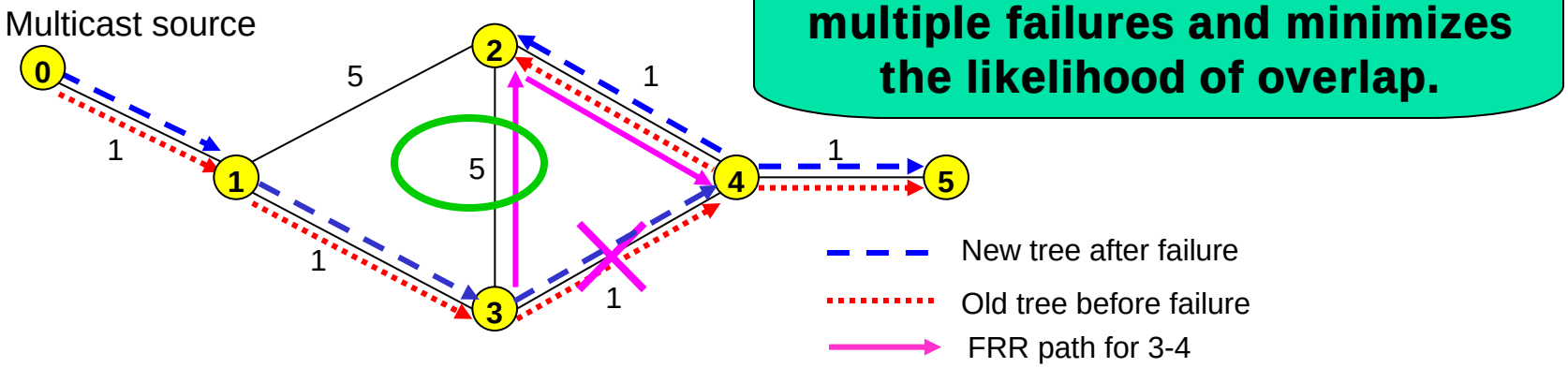
- Depend on pure Layer 3 mechanisms
- PIM Rejoin – *a pure multicast layer approach*:
  - A "passive" approach with standard PIM timers. Each PIM router resends a join on the upstream interface periodically, every 30secs or more, to refresh soft state.
  - IGP is exposed to the failures.
  - Pros: Standard definition of multicast. No need for extra implementation complexity.
  - Cons: When FRR is not used, significant loss takes place. When FRR is used, traffic overlaps can occur. During switchover to the new tree significant loss can occur.

We solve these issues without causing any significant state or messaging overhead.

Existing Failure Restoration Approaches

- FRR + IGP: Careful setting of IGP link weights
  - Idea: Set IGP link weights such that overlaps are avoided
  - Again, IGP and PIM are not bothered with failures
  - Pros: It is feasible to find such link weights for single failures [INFOCOM'07]
  - Cons: Overlaps are still possible for multiple failures

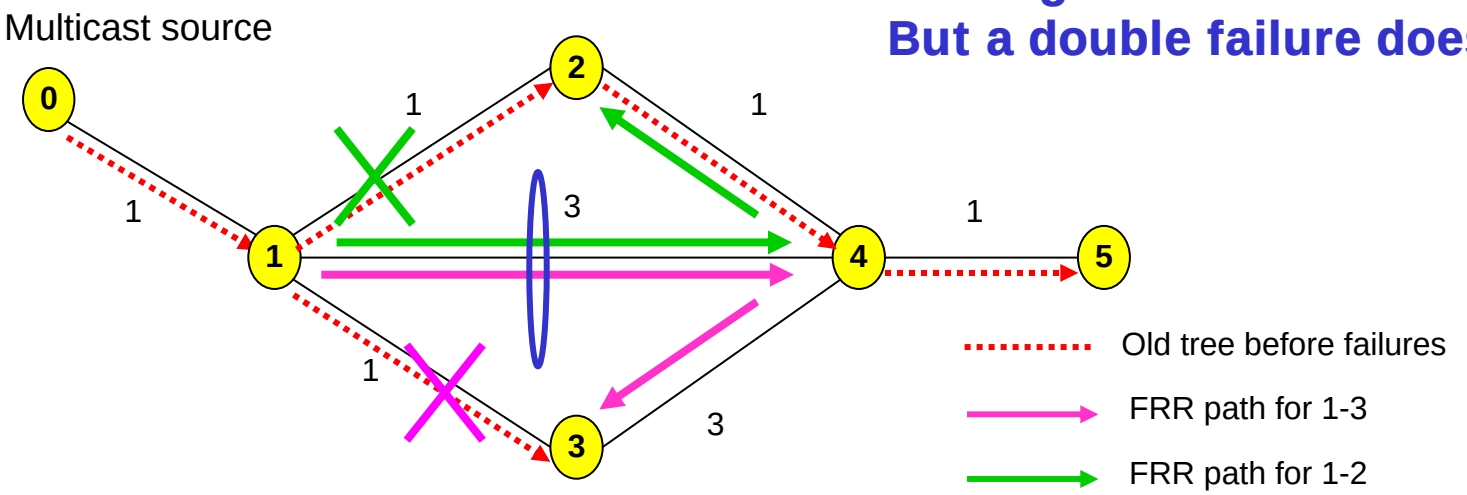
Our method can work over multiple failures and minimizes the likelihood of overlap.



Multiple Failures

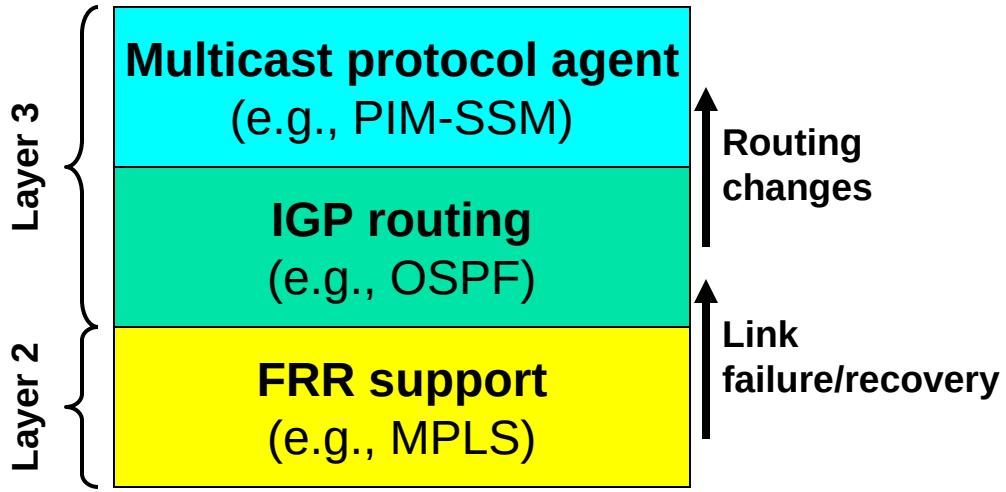
- None of the existing approaches can reasonably handle multiple failures.
  - Multiple failures can cause FRR traffic to overlap.
  - PIM must be informed about the failures and should switchover to the new tree as soon as it is possible.
    - So that overlaps due to multiple failures are minimized.

No single failure causes an overlap.
But a double failure does..



Our Approach: FRR + IGP + PIM

- Key contributions of our approach:
  - It guarantees reception of all data packets even after a failure (except the packets in transit) – hitless
  - It can be initiated when a failure is detected locally by the router and does not have to wait until routing has converged network-wide – works with local rules
  - It works even if the new upstream router is one of the current downstream routers – prevents loops during switchover



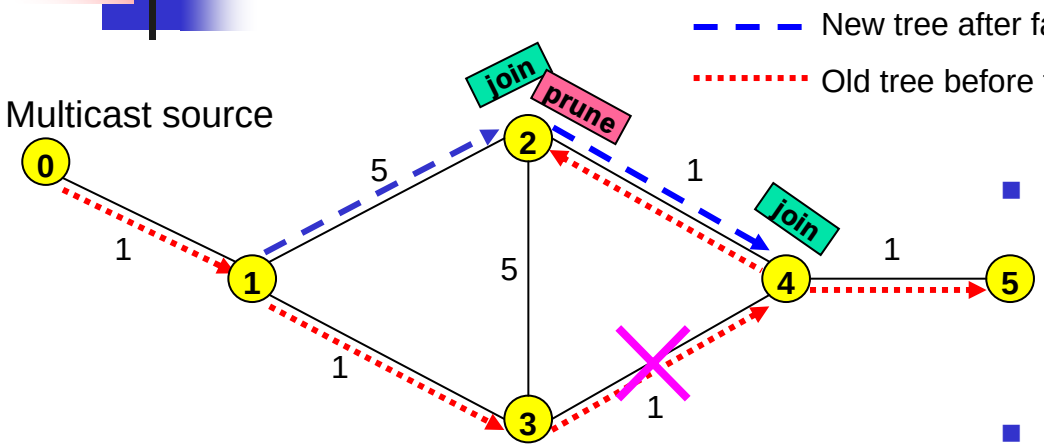
IGP-aware PIM: Key Ideas

Our key ideas as "local" rules for routers:

- Rule #1: Immediately try sending a join message if upstream has changed.
  - If IGP routing has changed, PIM will be notified.
  - PIM will evaluate and see if any of its (S,G) upstream nodes has changed. If so, it will try sending a join to the new upstream node. Two possibilities:
    - #1.a New upstream node is NOT among current downstream nodes → Just send the join immediately.
    - #1.b New upstream node is among current downstream nodes → Move this (S,G) into **"waiting-to-send-join" state** by marking a binary flag.
  - Do not remove the **old upstream node's state** info yet.
- Rule #2: Prune the old upstream when data arrives on the new tree.
  - Send prune to the old upstream node when you receive a data packet from the new upstream node.
  - Remove the old upstream node's state info.
- Rule #3: Move out of the transient "waiting-to-send-join" state upon prune reception.
  - When a prune arrives from a node on which we have been in the "waiting-to-send-join" state, then:
    - Send the joins for all (S,G)s that have been "waiting-to-send-join" on the sender of the prune.
    - Execute the prune normally.

Very minimal additional multicast state.

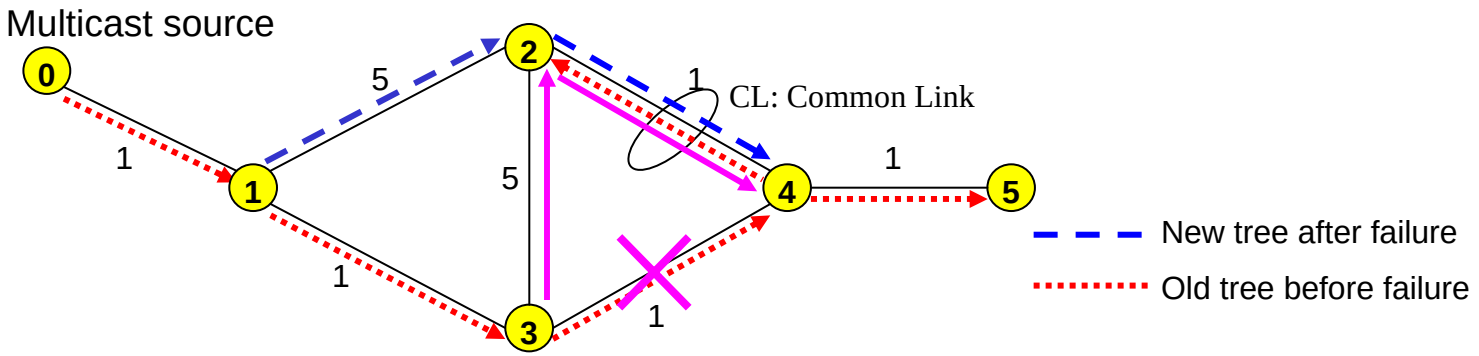
IGP-aware PIM Switchover: A sample scenario, No FRR yet



- Node 4:
  - detects the routing change after SPF and tries to send a join message to 2 (#1)
    - moves to "waiting-to-send-join" state (#1.b)
- Node 2:
  - hears about the failure and does SPF
  - detects the routing change after SPF and tries to send a join message to 1 (#1)
    - sends the join to 1 (#1.a)
    - but does not install the 2→1 interface yet
- Node 1:
  - receives the join message from 2
  - adds the 1→2 downstream interface and data starts flowing onto the new tree
- Node 2:
  - receives data packets from new tree and sends a prune to old upstream node (#2)
- Node 4:
  - receives prune from 2 and moves out of "waiting-to-send-join" state by sending the join to 2 (#3)
  - processes the received prune
- Node 2:
  - receives the join message from 4
  - adds the 2→4 downstream interface and data starts flowing onto the new tree

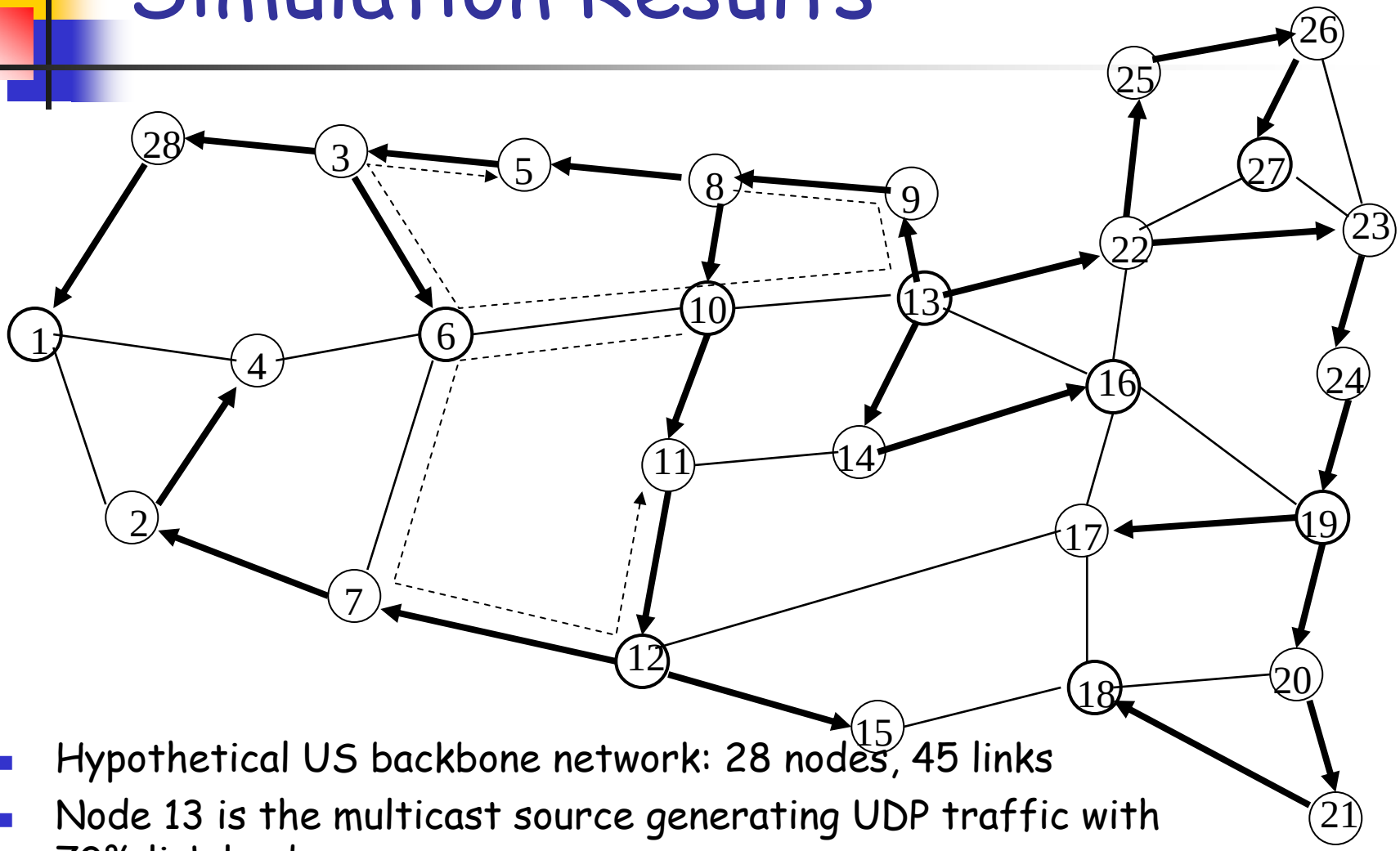
FRR Support → Congested Common Link

- When there is FRR support, common links (i.e., overlaps) may happen.
- Common Link (CL):
  - During a switchover, the new tree might overlap with the FRR path of the link that failed.



- **Issue: Congested Common Link**
  - CL might experience congestion and data packets on the new tree (blue) might never arrive at the node 4?
- **Solution**: Allow CLs, but **prioritize the traffic** on the new tree
  - After link failure, mark the data traffic on the new tree with a higher priority and FRR packets with lower priority.

Simulation Results

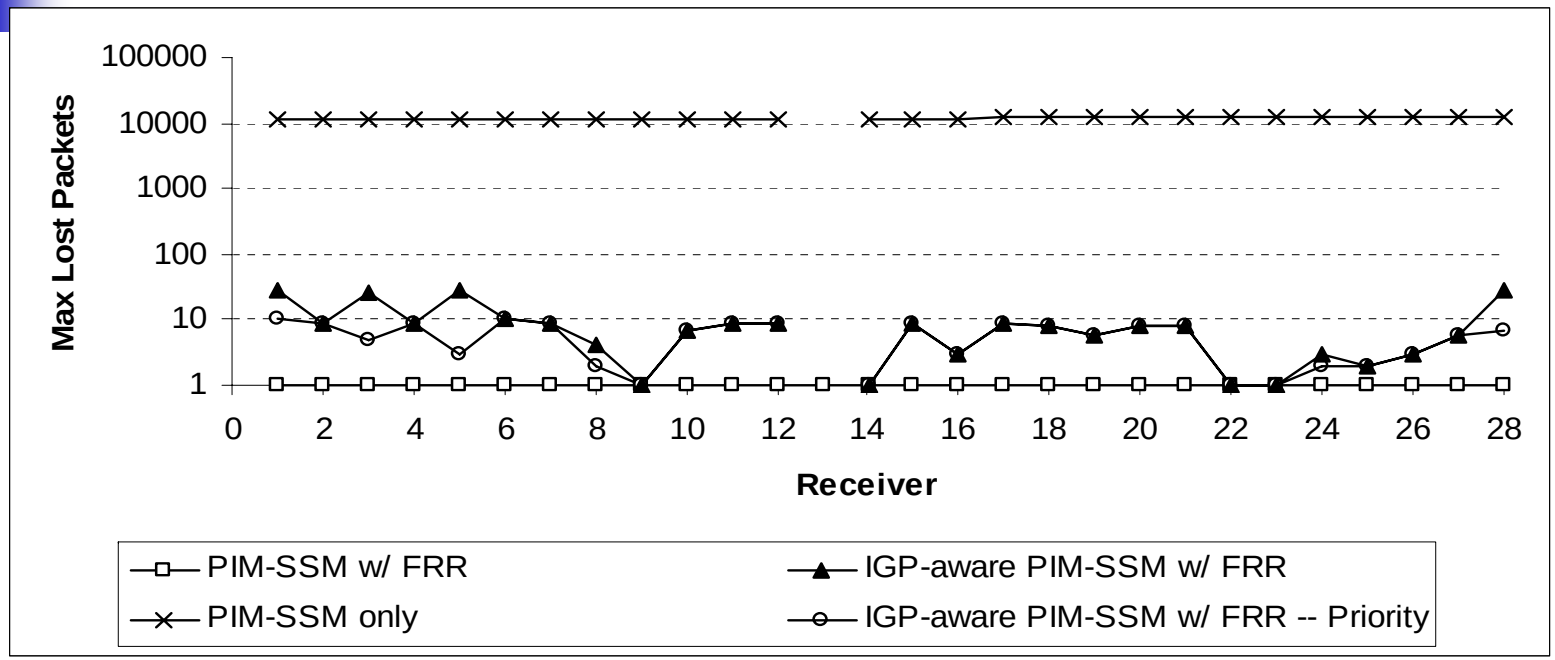


- Hypothetical US backbone network: 28 nodes, 45 links
- Node 13 is the multicast source generating UDP traffic with 70% link load.

Simulation Results

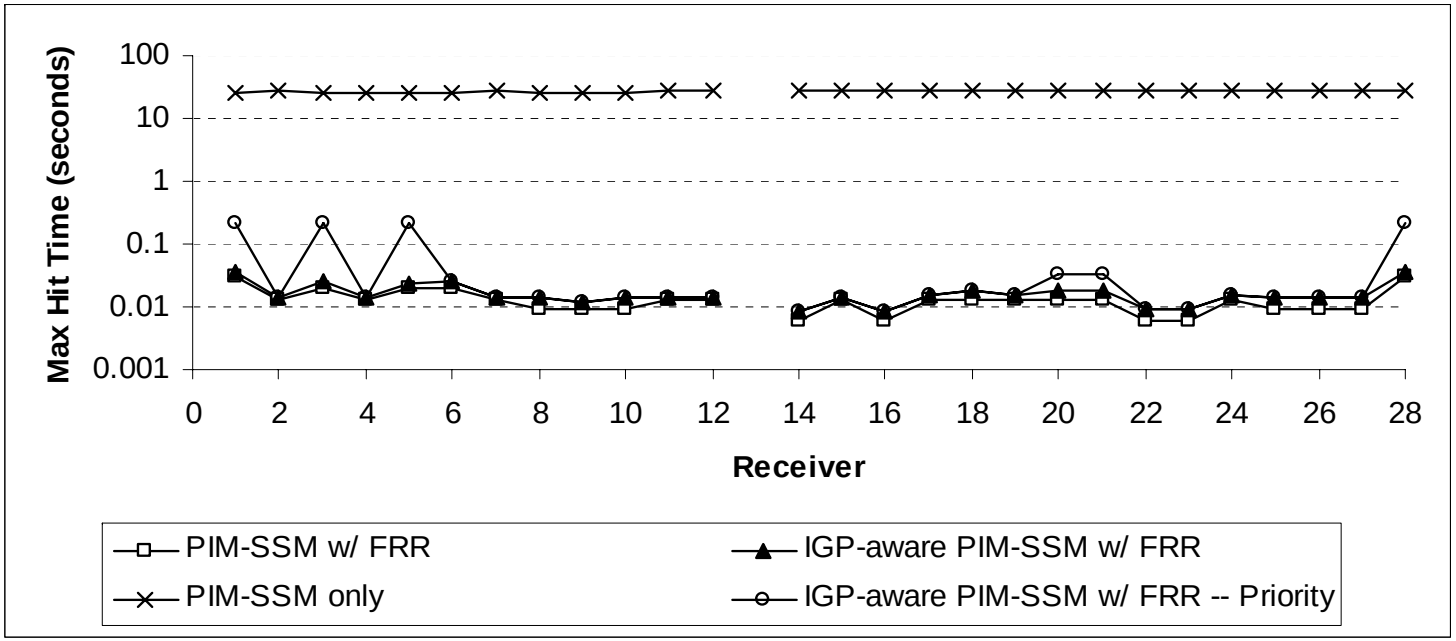
- ns-2 simulation of OSPF as the IGP, PIM-SSM as the multicast, and MPLS for FRR support
- Comparative evaluation of:
  - PIM-SSM Only
    - The standard IP multicast with PIM rejoin
  - PIM-SSM w/ FRR
    - Only FRR is used for restoration
  - IGP-aware PIM-SSM w/ FRR
    - Our multicast tree switchover protocol
  - IGP-aware PIM-SSM w/ FRR – Priority
    - Our multicast tree switchover protocol with low-priority forwarding of FRR traffic
- 120ms buffer time, 5secs spfDelayTime, and 10secs spfHoldTime
- 30secs of PIM rejoin time
- Failed each link on the tree and observed hit time and lost packets

Simulation Results



- With a pure Layer 3 (PIM-SSM only) solution, far too many packets are lost
- Maximum Lost packets goes down dramatically with a Layer 2 recovery mechanism like FRR
- Our IGP aware mechanism introduces no further hits
  - Primary loss is packets in "flight" and queued on outbound interface

Simulation Results (contd.)



- "PIM-SSM only" experience outages of tens of seconds - unacceptable
- IGP aware PIM with FRR has about the same time for "hit" as a single failure recovery time with FRR
  - Failure detection time dominates

Summary

- A method to make PIM-SSM re-convergence aware of the underlying network failure conditions.
- The method allows Fast Reroute support at the link layer.
- We are currently experimenting with multiple failure scenarios.

THE END

Thank you!