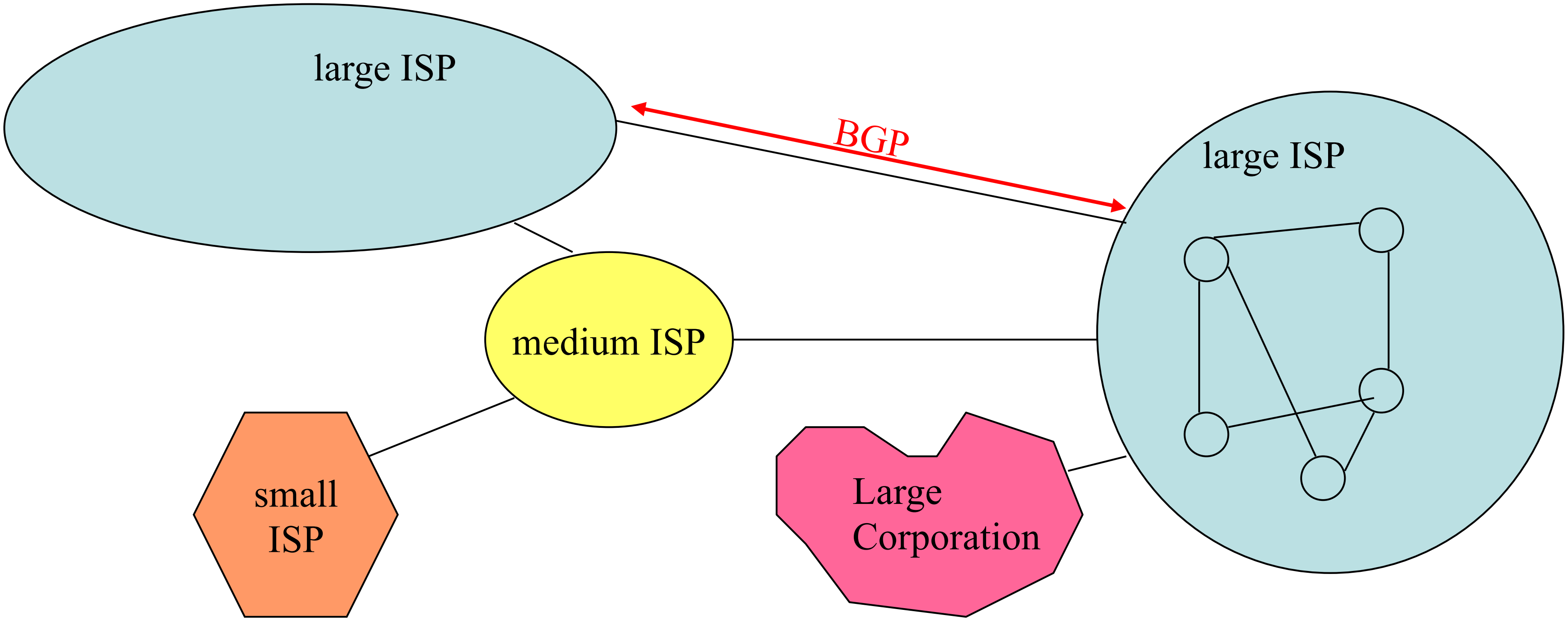
CSC 525:
Principles of Computer Networks

Inter-domain Routing



- AS (Autonomous System): a collection of routers under the same technical and administrative control. Identified by a 32-bit AS number.
- BGP (Border Gateway Protocol): between ASes to exchange inter-domain routing information.
- BGP decides the routes at the AS level, while intra-domain routing decides the routes within an AS

Why BGP?

- Why another layer of routing with a different protocol?
  - Accommodate different administrative domains
  - Allow flexible routing policies
  - Routing scalability

- **1989 : BGP-1 [RFC 1105]**
  - **Replacement for EGP (1984, RFC 904)**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771]**
  - **Support for Classless Interdomain Routing (CIDR)**

Path Vector Routing

- Similar to Distance Vector
  - receive paths from neighbors, choose the best paths, announce the best paths to neighbors.
- Major differences:
  - Spell out the entire AS path
    - E.g. (52, 2153, 11537, 1706)
    - Prevent loops, facilitate policies
  - Store all backup paths
    - Speed up failover
  - Hoping these changes will avoid most problems of the distance vector.

BGP Operations

- Establish a BGP peering session on top of *TCP*
  - Reliable delivery of routing messages between two routers
    - Simplify BGP operations
    - Save the periodic route refresh
  - Negative impact under congestion
  - Periodic Keep-Alive messages to maintain the session.
    - Keepalive timer 30 seconds, Hold timer 180 seconds.
- Initial Routing Table Exchange
  - Similar to OSPF's database synchronization
- Propagate incremental routing updates afterwards
  - Triggered by route changes.
  - No periodic refresh of routing updates, i.e., hard states.

BGP Message Types

- Open: Establish a peering session
- Keep Alive: handshake at regular interval
- Notification: shuts down a peering session
- Update:
  - **announcing** new routes or
  - **withdrawing** previously announced routes

Announce : prefix, path attributes

Withdraw : prefix

Prefix

- Represent the destination network
- In the old days, fixed boundary between the network and host parts of an address
  - Class A (/8), B(/16), C (/24)
- CIDR: arbitrary boundary
  - 131.179.96.0/23
  - More efficient address allocation, allow aggregation
  - Need explicit prefix length or mask
- Routing lookup uses longest prefix match
  - 131.179.0.0/24 vs. 131.179.0.0/16
- CIDR allows aggregation
  - 131.179.0.0/24 and 131.179.1.0/24 → 131.179.0.0/23

Attributes

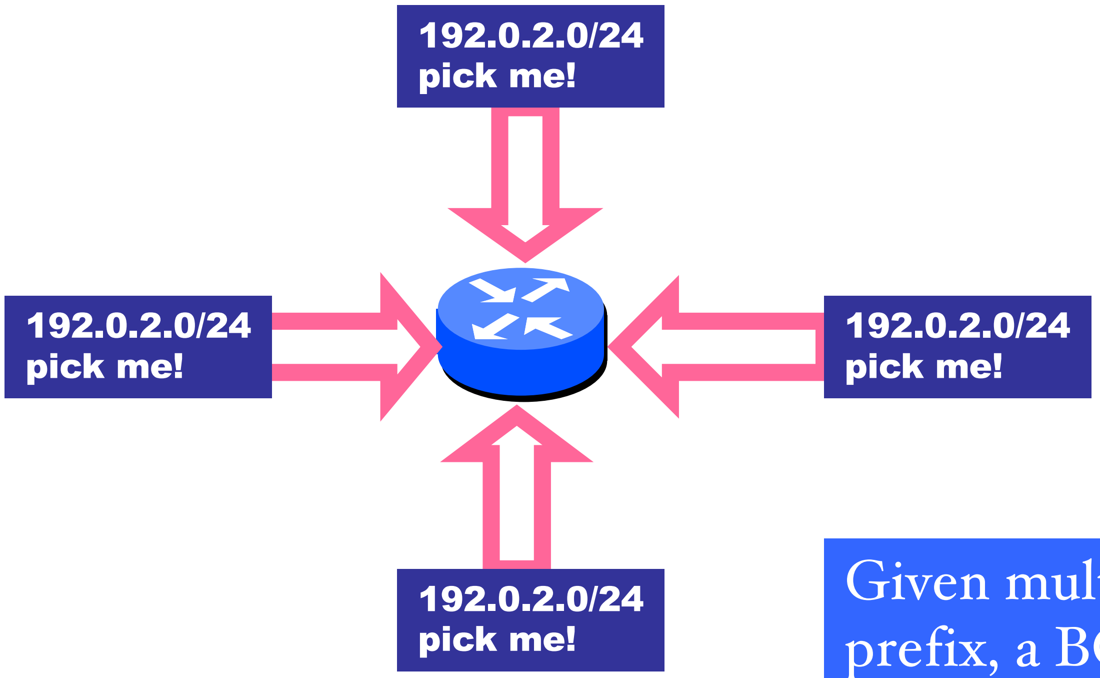
Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	[Rosen]
...		
255	reserved for development	

From IANA: http://www.iana.org/assignments/bgp-parameters

Most Important attributes

Not all attributes need to be present in every announcement

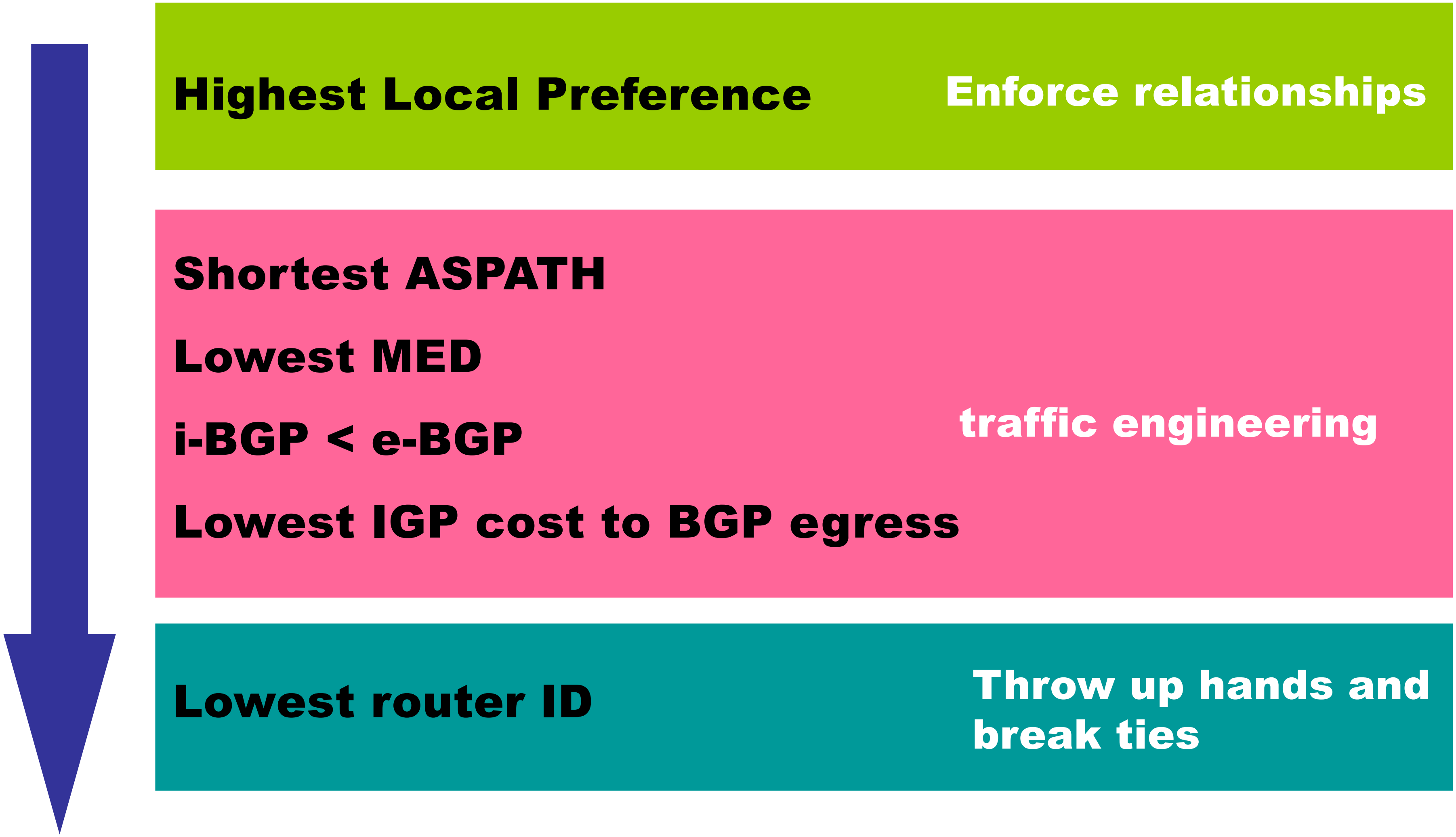
Attributes are used in best route selection



Given multiple routes to the same prefix, a BGP router must pick at most one best route

(Note: it could reject them all!)

Route Selection Summary



Highest Local Preference — Enforce relationships

Shortest ASPATH

Lowest MED

i-BGP < e-BGP

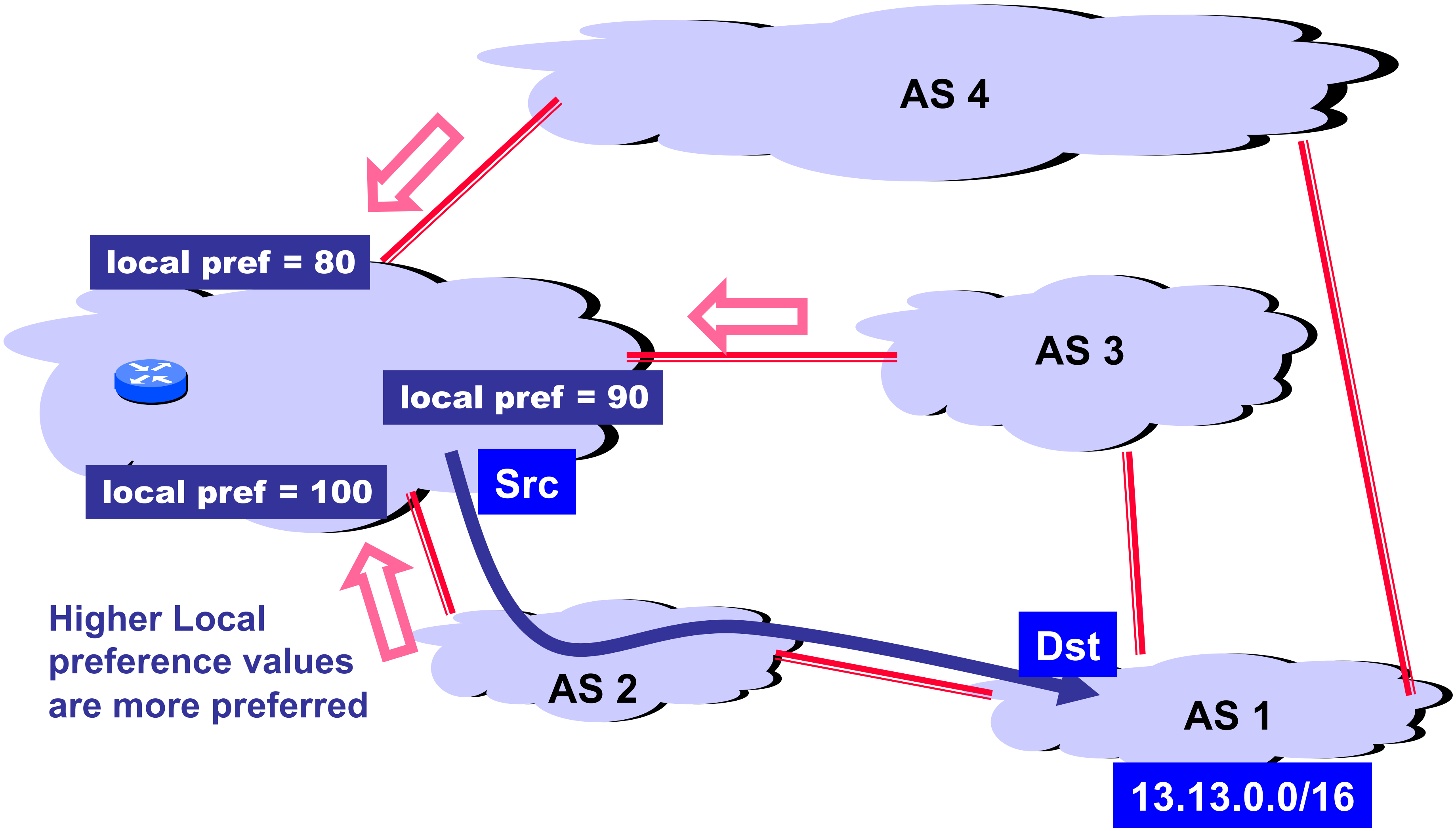
Lowest IGP cost to BGP egress

traffic engineering

Lowest router ID — Throw up hands and break ties

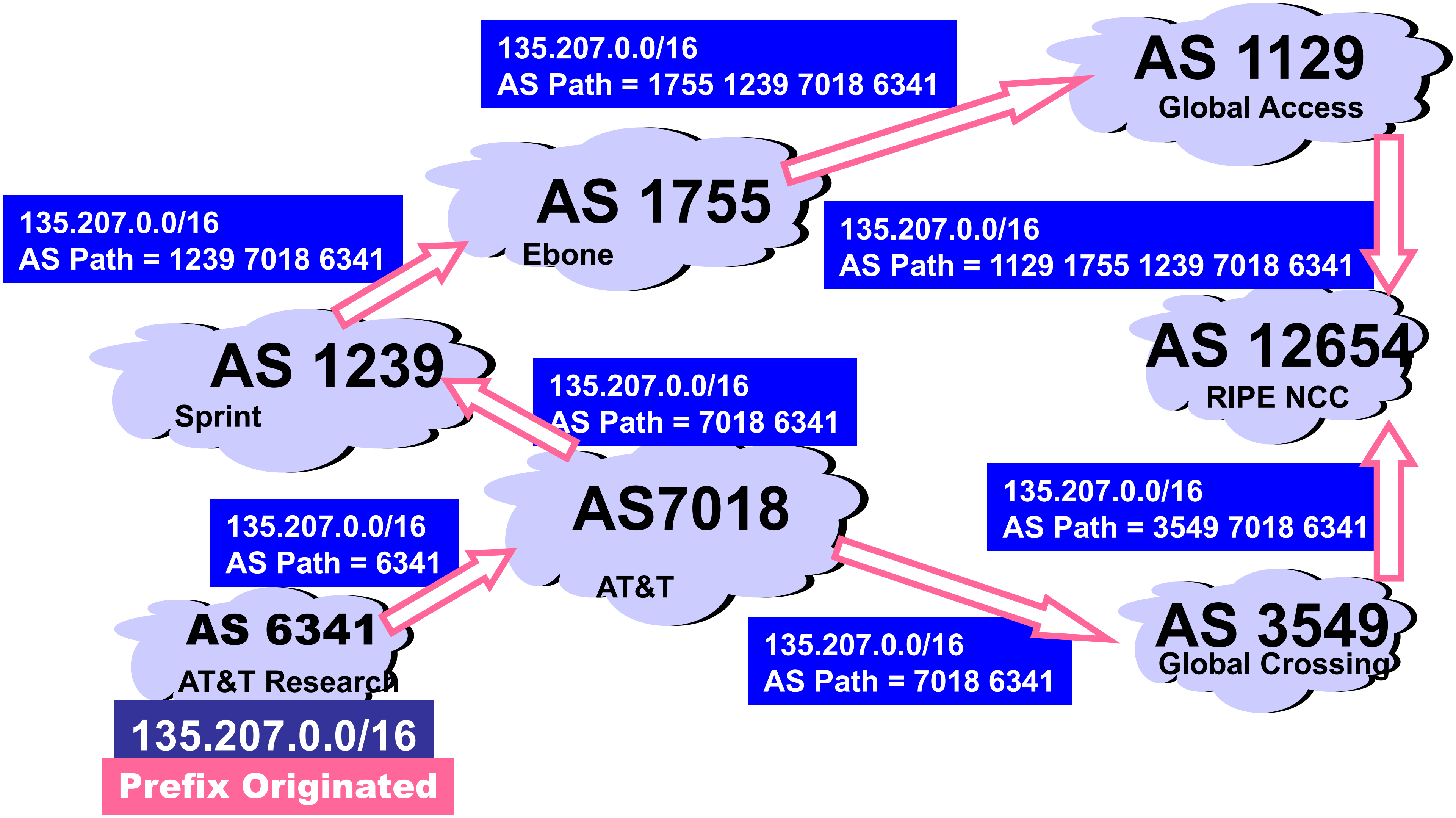
(Not using path cost like in the intra-domain routing)

Local Preference



Higher Local preference values are more preferred

ASPATH Attribute



135.207.0.0/16
AS Path = 1755 1239 7018 6341

AS 1129
Global Access

AS 1755
Ebone

135.207.0.0/16
AS Path = 1239 7018 6341

135.207.0.0/16
AS Path = 1129 1755 1239 7018 6341

AS 12654
RIPE NCC

AS 1239
Sprint

135.207.0.0/16
AS Path = 7018 6341

AS7018
AT&T

135.207.0.0/16
AS Path = 3549 7018 6341

135.207.0.0/16
AS Path = 6341

AS 6341
AT&T Research

AS 3549
Global Crossing

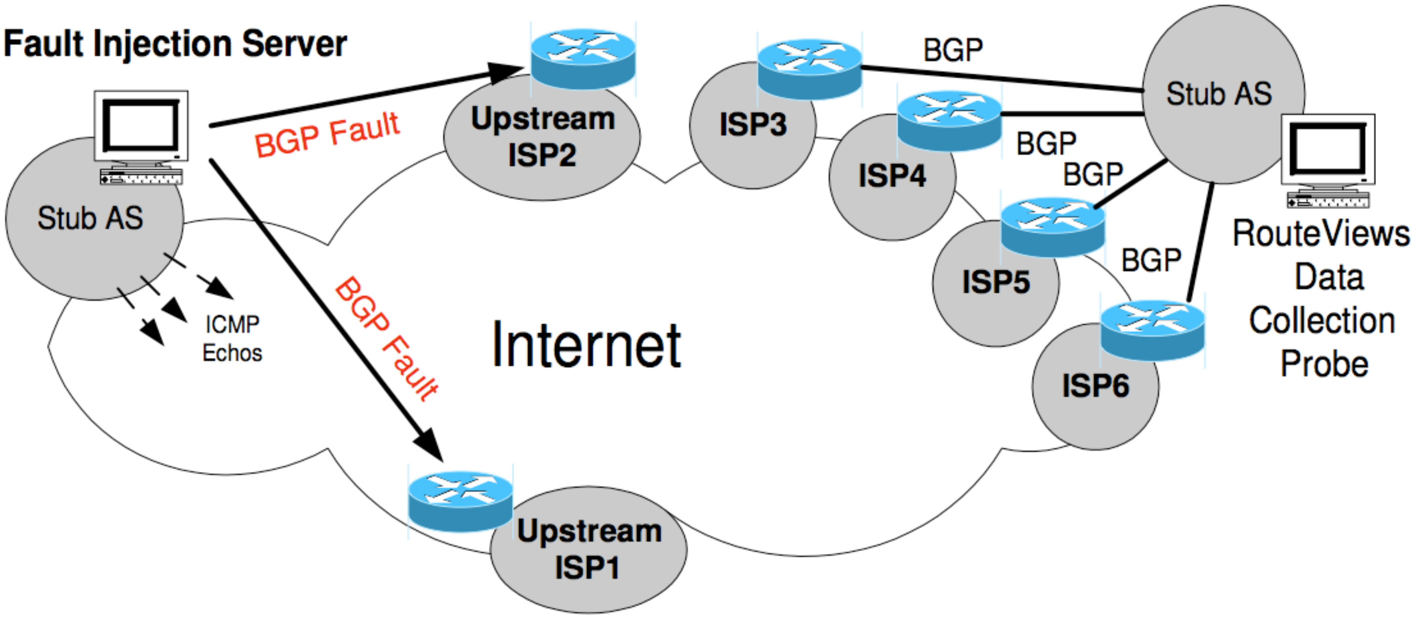
135.207.0.0/16
AS Path = 7018 6341

135.207.0.0/16
Prefix Originated

"Delayed Internet Routing Convergence"

- Previously held belief:
  - path vector should have no routing loops and should converge fast since it explicitly carries the entire path in the routing updates and stores backup paths.
- Discovery:
  - *Path exploration* causes longer convergence delay, packet loss, and extra latency to data traffic.
- Methodology:
  - Collect and analyze BGP routing updates from different vantage points on the Internet.

Measurement Framework

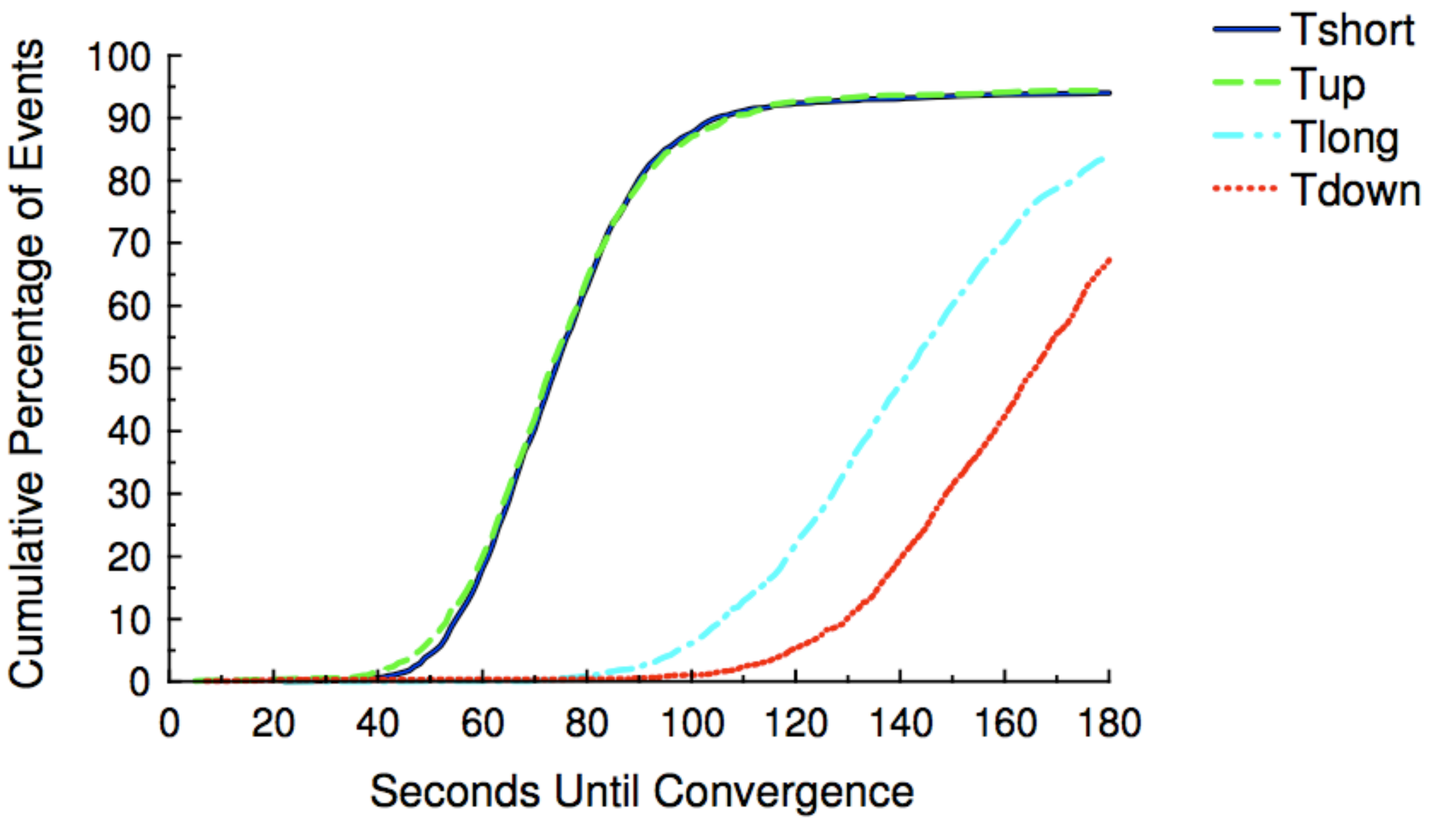


- Controlled experiments
  - RouteViews passive monitoring
  - BGP fault injection (sometimes called *beacons*)
- Active probing for end-to-end performance

Event Types and Performance Metrics

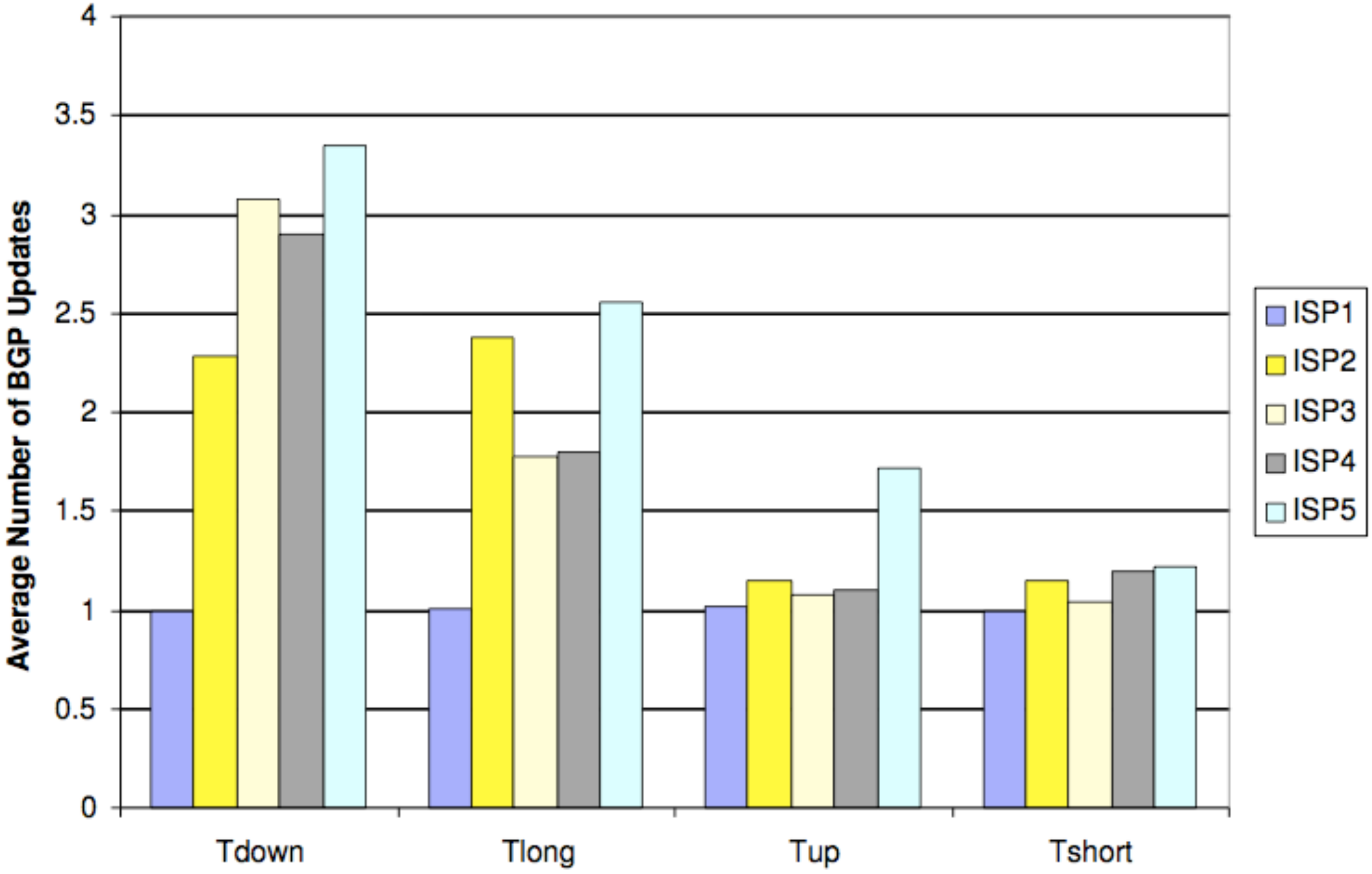
- Type of Events
  - Tup, Tdown, Tshort, Tlong
  - Equipment failures, maintenance, configuration changes, session failures etc.
- Routing Metrics
  - Convergence time
  - Number of routing messages
  - Network-wide vs. a single router's view
- End-to-end performance
  - Packet loss
  - Round trip delay

Convergence Time

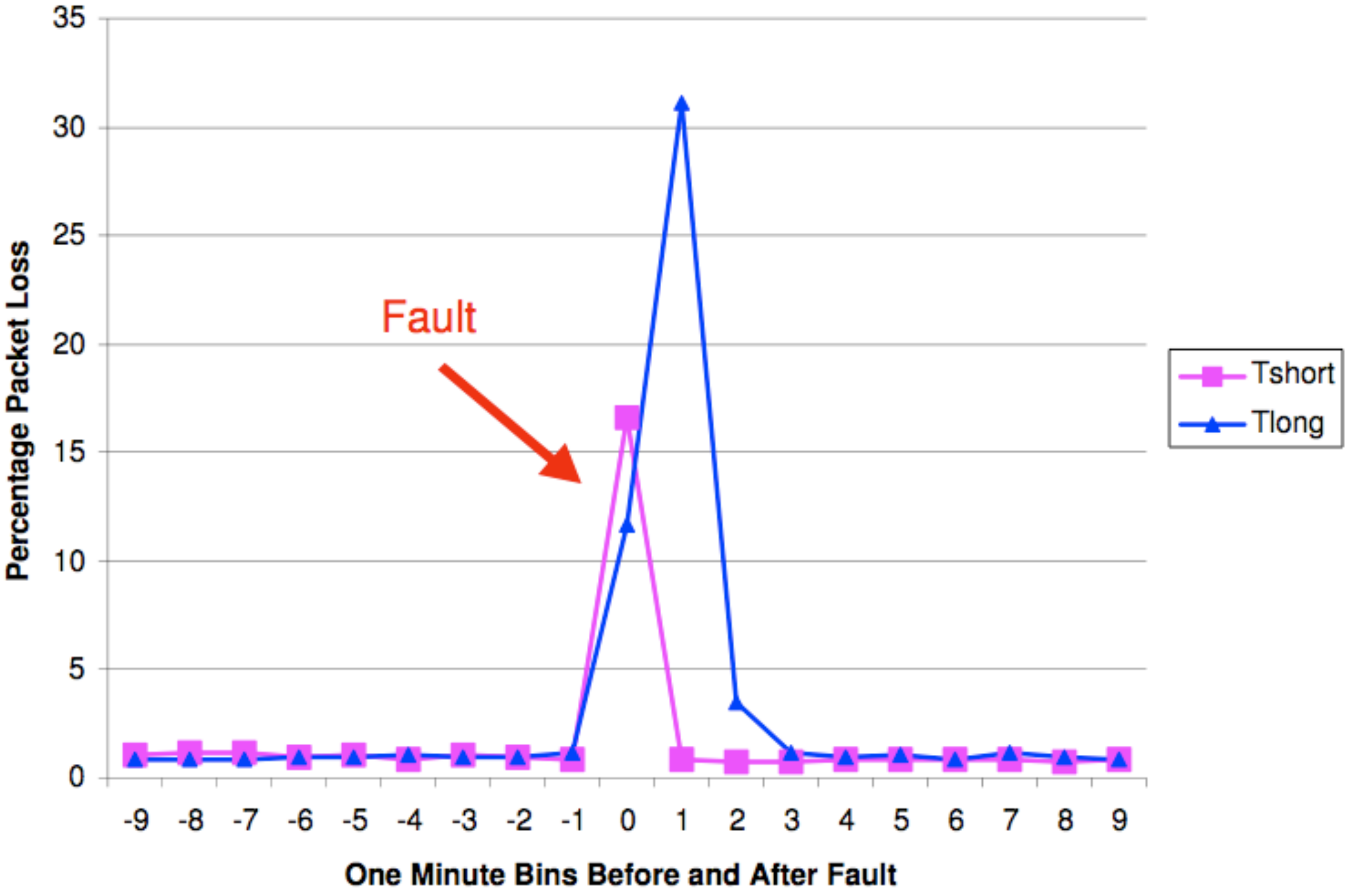


- Longer than expected
- T_{down} and T_{long} take much longer

Number of Updates

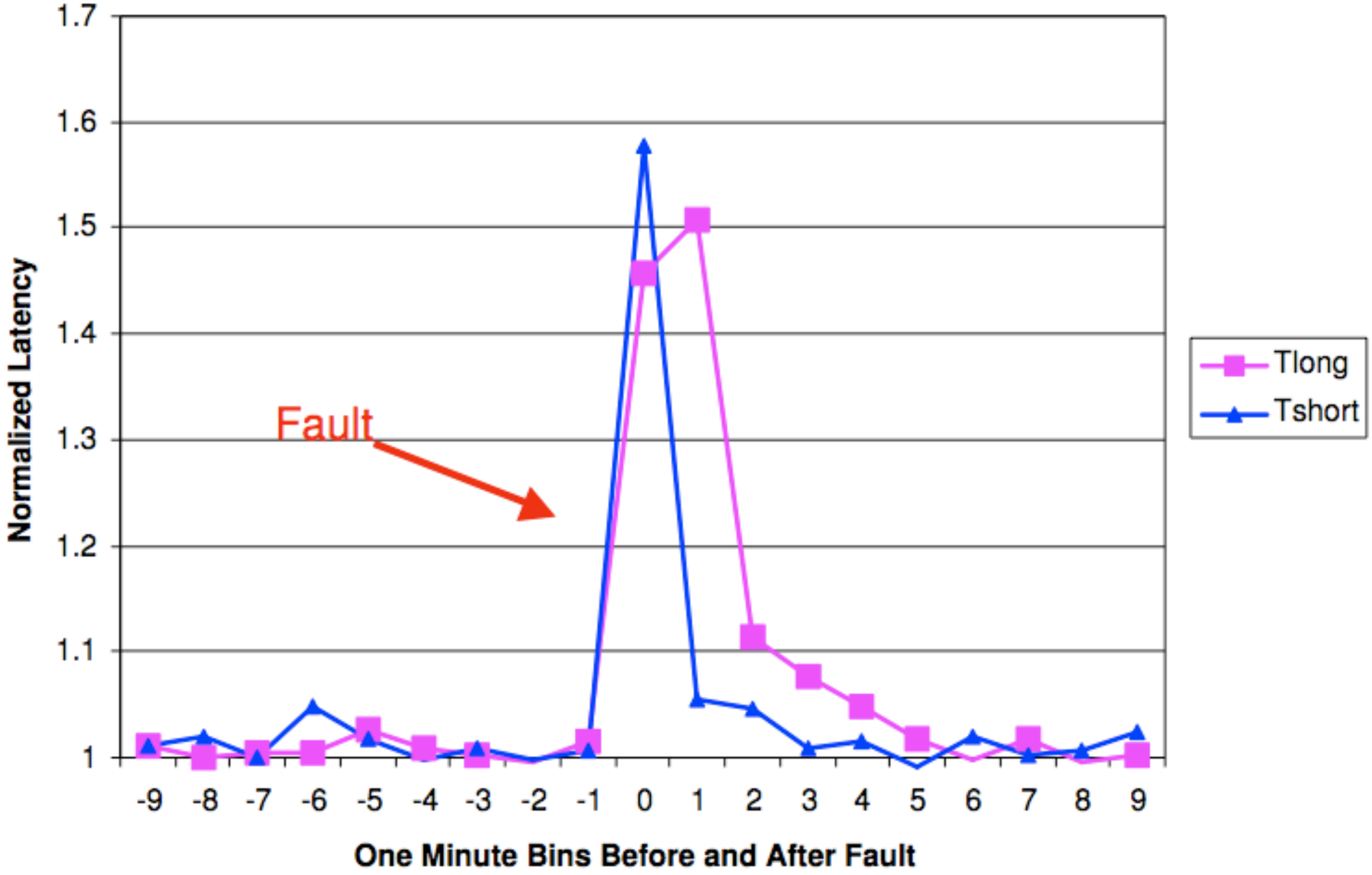


Packet Loss

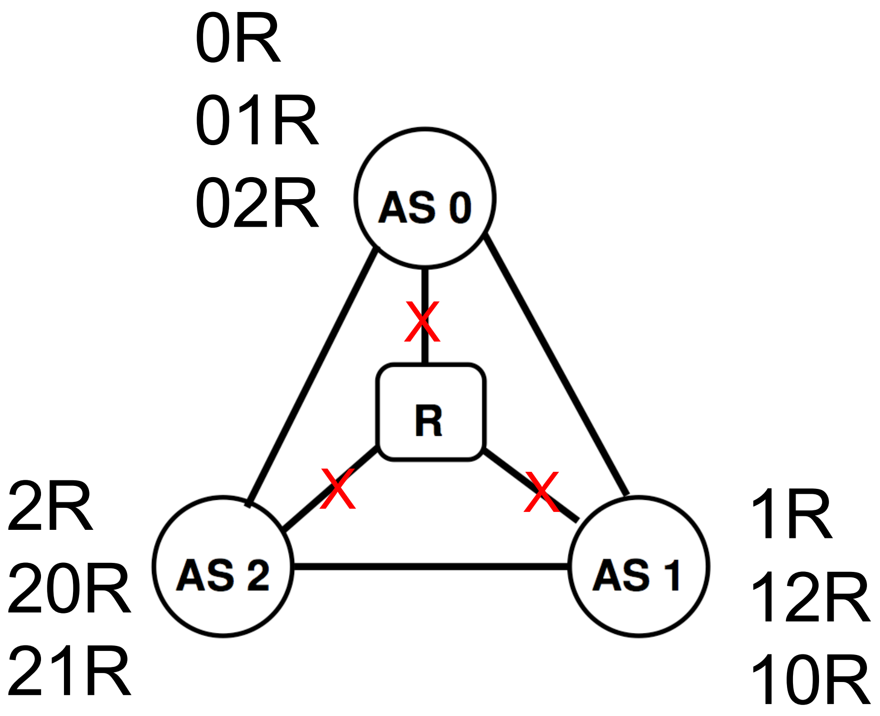


Percentage Packet Loss
35
30
25
20
15
10
5
0
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
One Minute Bins Before and After Fault
Fault
Tshort
Tlong

Round Trip Delay



Path Exploration



Path vector prevents loops, but is still vulnerable to invalid paths

- Initial situation
  - All ASes use direct path
- Then R fails
  - All ASes lose direct path
  - None knows R failed.
  - All switch to longer paths
  - Eventually withdrawn, but it takes time.
- E.g., AS 2
- (2,R) ➔ (2,0,R)
- (2,0,R) ➔ (2,1,R)
- (2,1,R) ➔ (2,0,1,R)
- (2,0,1,R) ➔ null

Path Exploration

- The theoretical worst convergence time is O(n!)
  - Full mesh, message ordering, process and propagate one update at a time, no MRAI
  - No route filtering
  - Shortest ASPath
  - Tdown
- In reality, the typical case is must faster than this worst case, but still can be too long for applications.

MRAI Timer

- Minimum Route Announcement Interval
  - default 30 seconds
- Space out the sending of consecutive updates
  - Not applied to the first update
  - Not applied to withdrawals
- Consolidate transient updates during this time.
  - thus reduce path exploration
- But also slow down the propagation of correct route
- There's an optimal MRAI value for fast convergence
  - The value varies for different topologies though.

Contributions

- Measured BGP convergence time and end-to-end packet loss for different routing events on the Internet.
- Discovered path exploration
- Propose SSLD (sender side loop detection) as a minor fix.
  - Only effective in small topology
- Impacts
  - The measurement framework, terminology, and metrics.
  - Started the research on inter-domain routing dynamics.