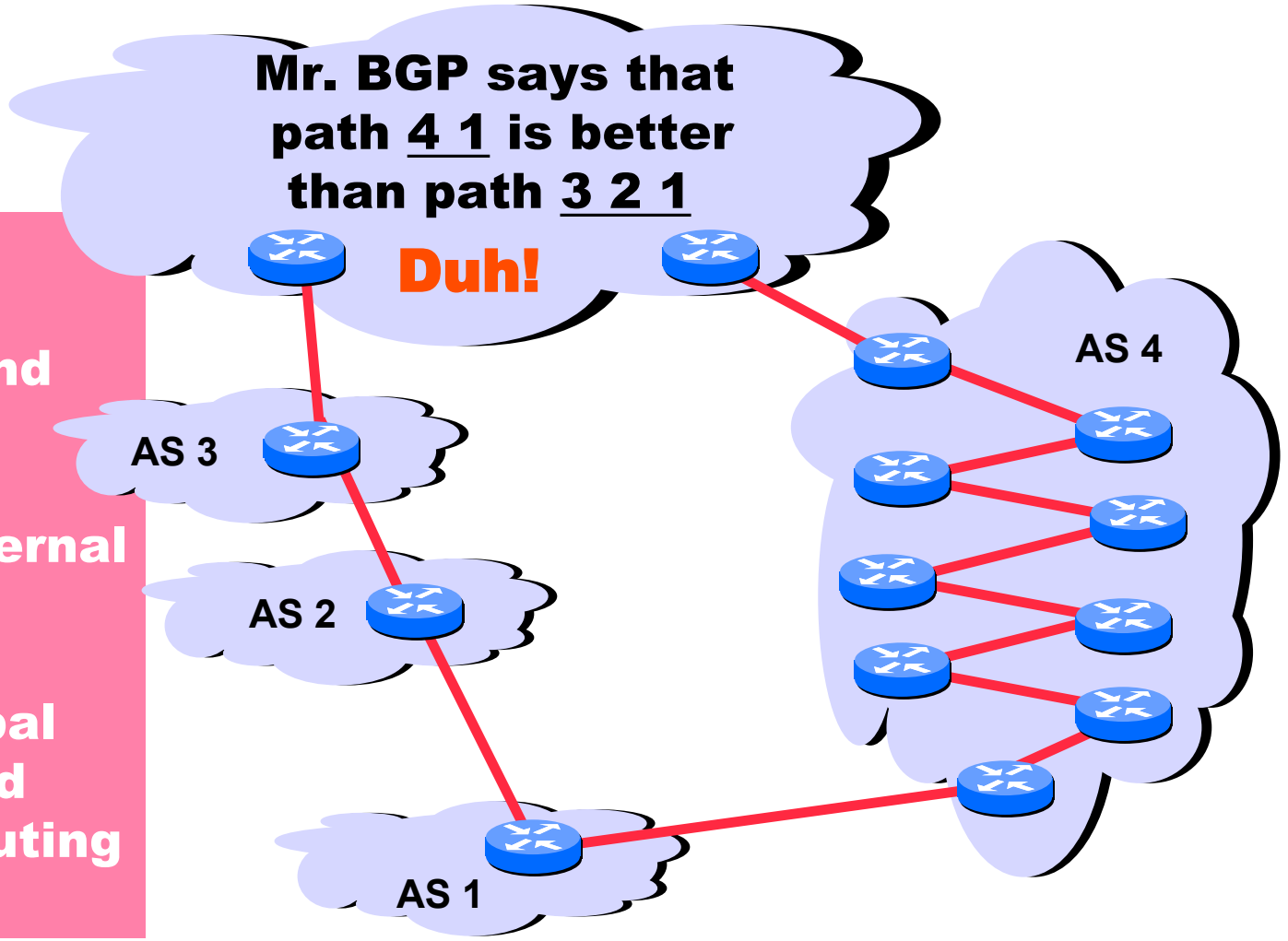
Internet Routing Protocols
Lectures 03 & 04
BGP Traffic Engineering and
BGP Dynamics

Advanced Systems Topics

Lent Term, 2010

Timothy G. Griffin
Computer Lab
Cambridge UK

Shorter Doesn't Always Mean Shorter



In fairness: could you do this "right" and still scale?

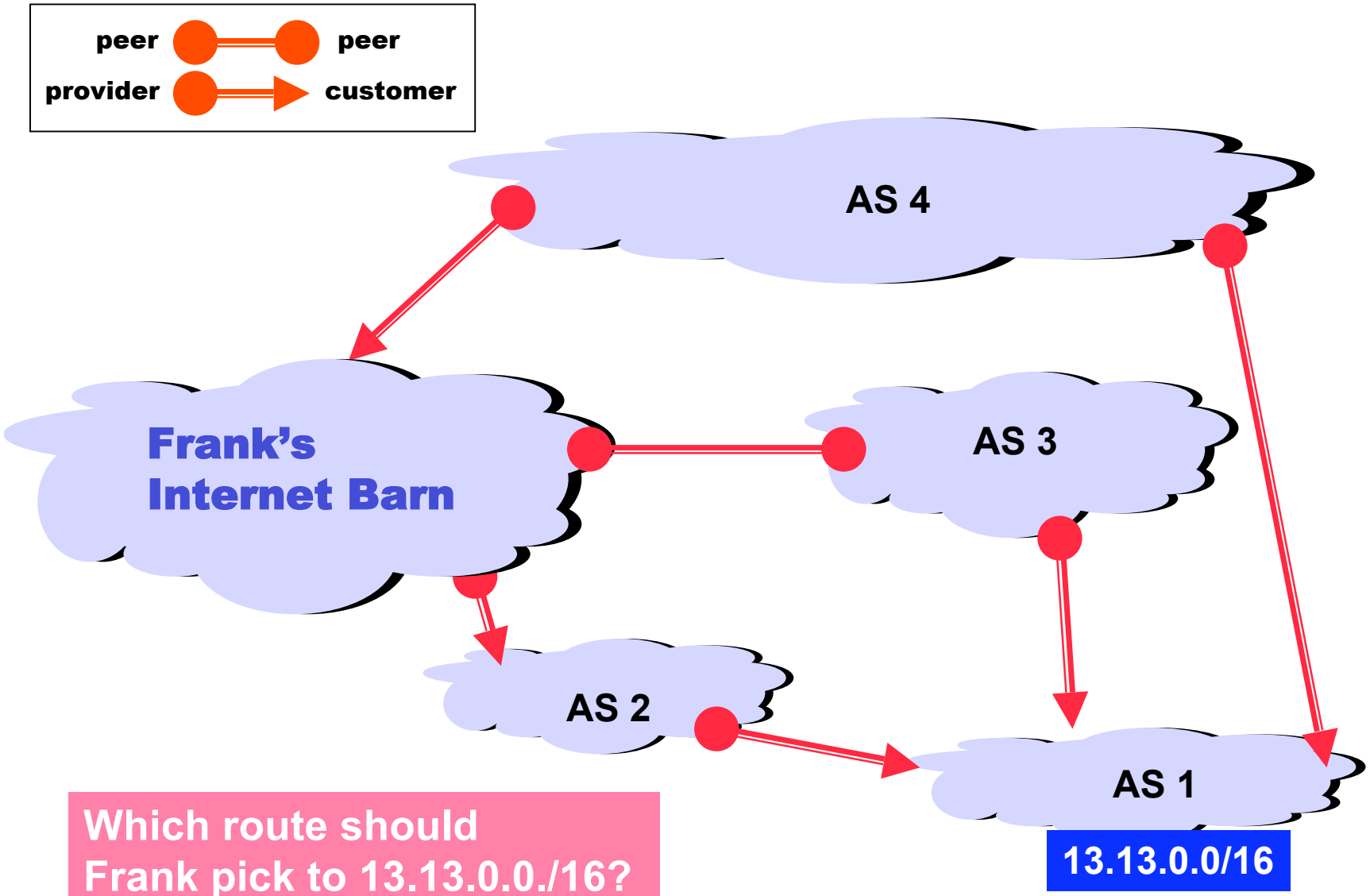
Exporting internal state would dramatically increase global instability and amount of routing state

Implementing Customer/Provider and Peer/Peer relationships

Two parts:

- Enforce transit relationships
  - Export all (best) routes to customers
  - Send only own and customer routes to all others
- Enforce order of route preference
  - provider < peer < customer

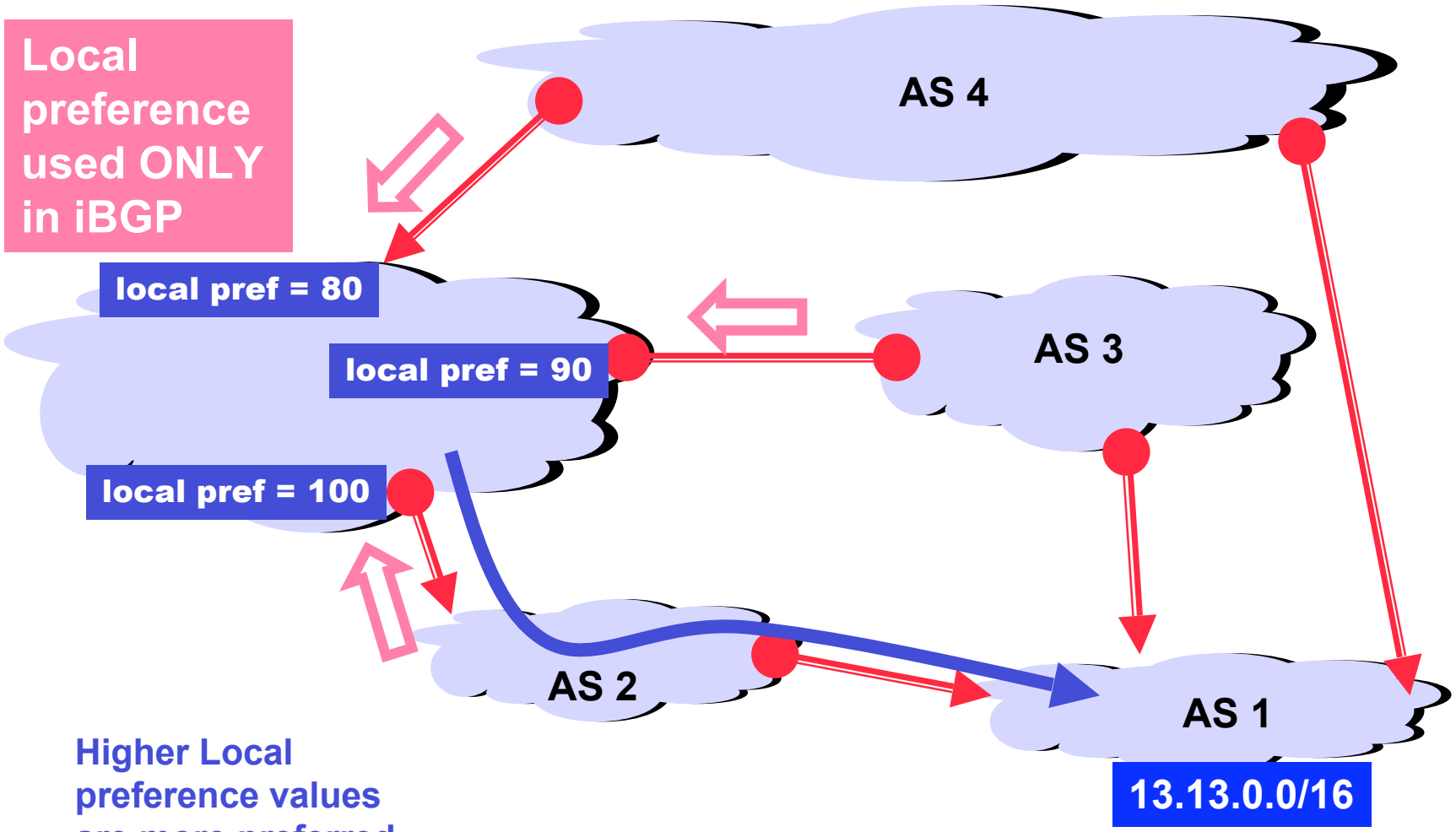
So Many Choices



Which route should Frank pick to 13.13.0.0./16?

LOCAL PREFERENCE

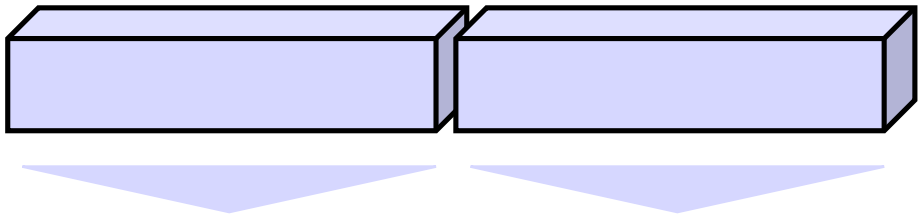
Local preference used ONLY in iBGP



Higher Local preference values are more preferred

How Can Routes be Classified? BGP Communities!

A community value is 32 bits



By convention, first 16 bits is ASN indicating who is giving it an interpretation

community number

Used for signally within and between ASes

Very powerful BECAUSE it has no (predefined) meaning

Community Attribute = a list of community values. (So one route can belong to multiple communities)

RFC 1997 (August 1996)

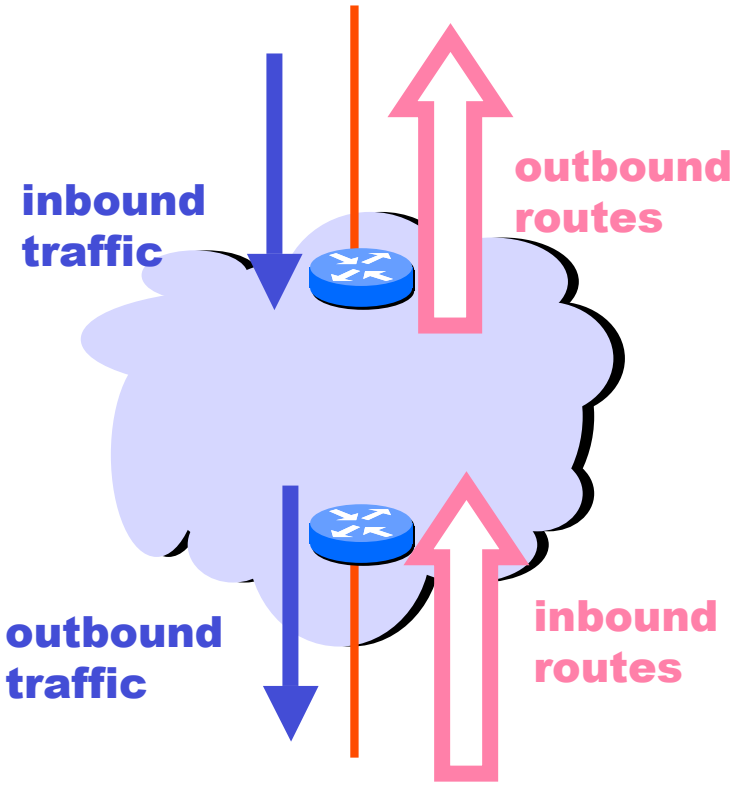
Reserved communities

no_export = 0xFFFFFF01: don't export out of AS

no_advertise 0xFFFFFF02: don't pass to BGP neighbors

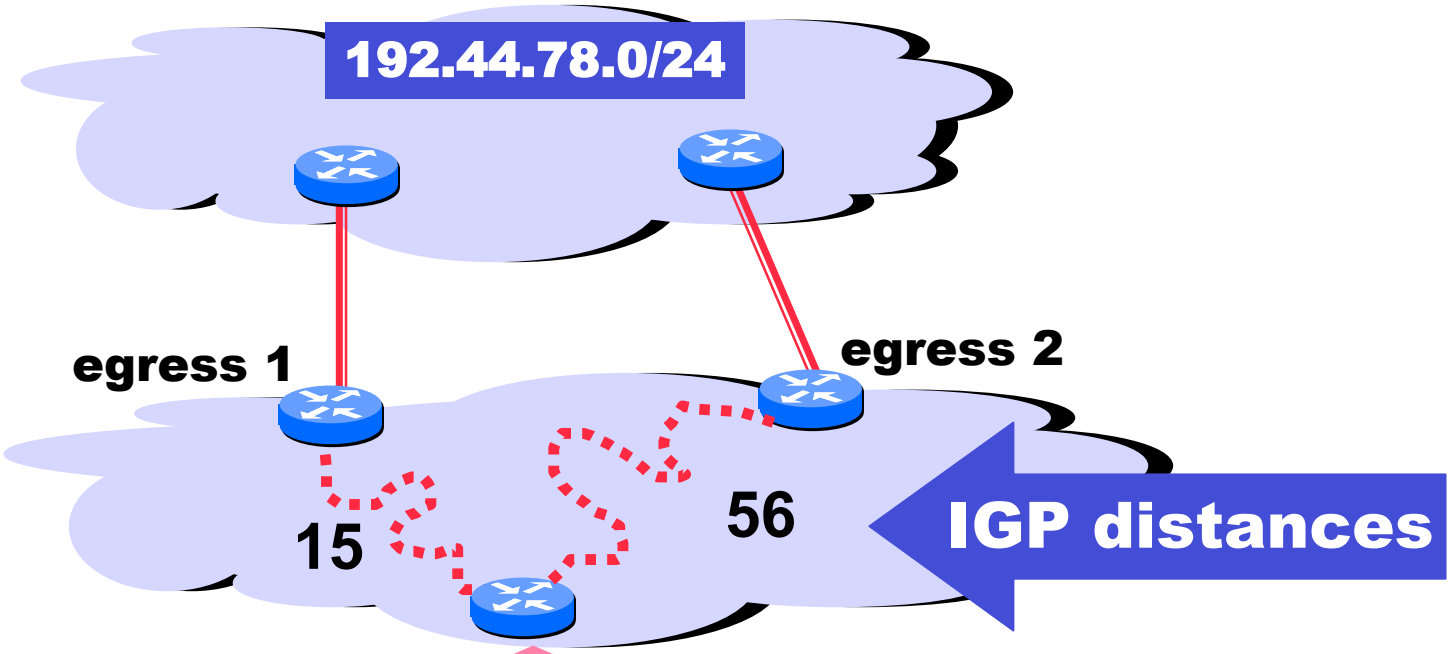
Tweak Tweak Tweak (TE)

- For inbound traffic
  - Filter outbound routes
  - Tweak attributes on outbound routes in the hope of influencing your neighbor's best route selection
- For outbound traffic
  - Filter inbound routes
  - Tweak attributes on inbound routes to influence best route selection



In general, an AS has more control over outbound traffic

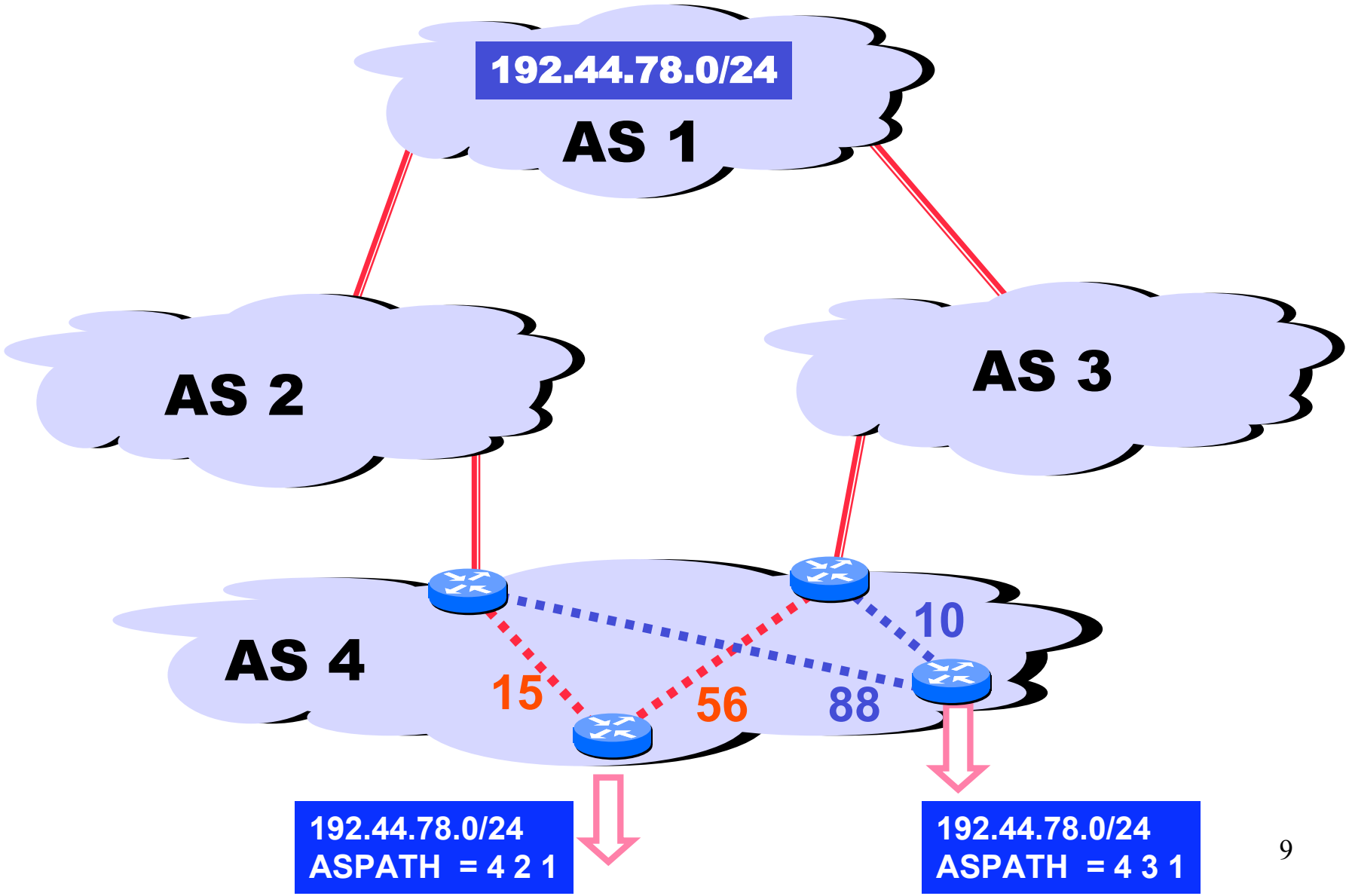
Hot Potato Routing: Go for the Closest Egress Point



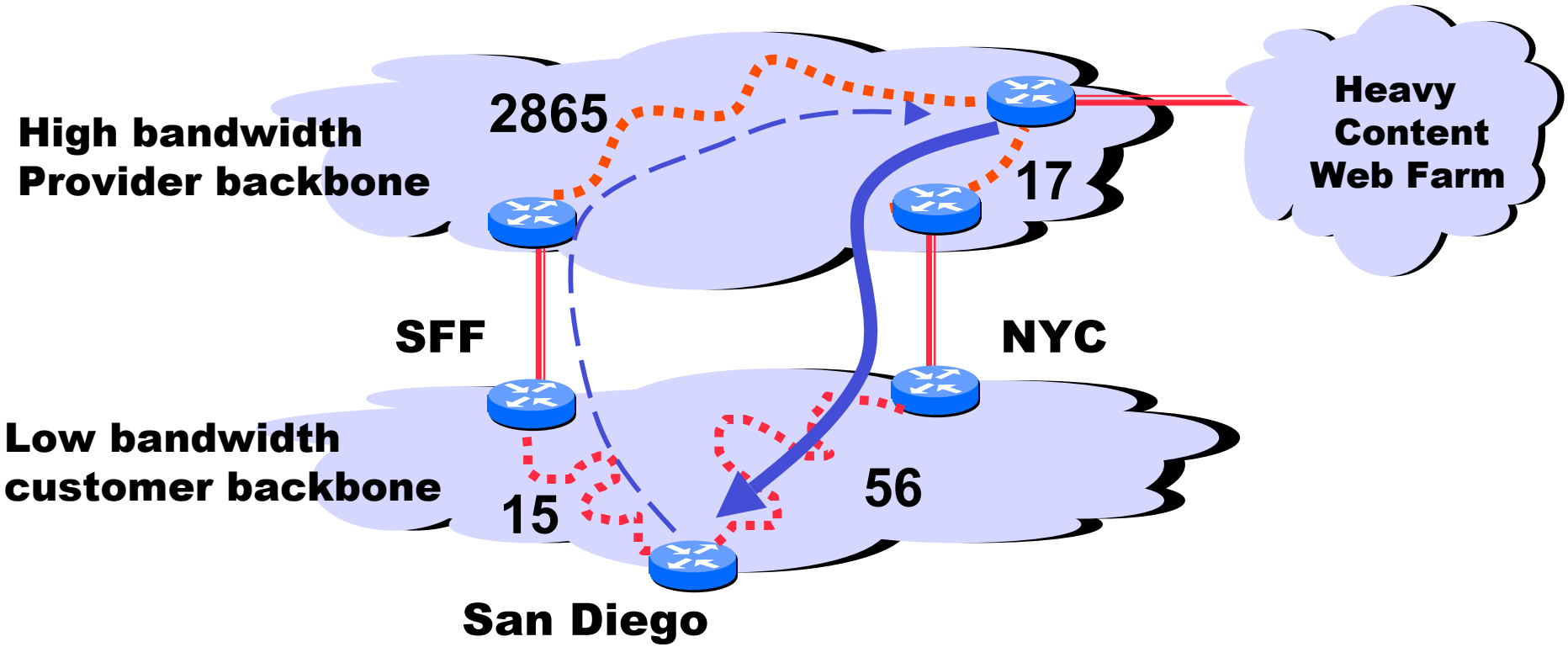
This Router has two BGP routes to 192.44.78.0/24.

Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!

Routers make independent selections!

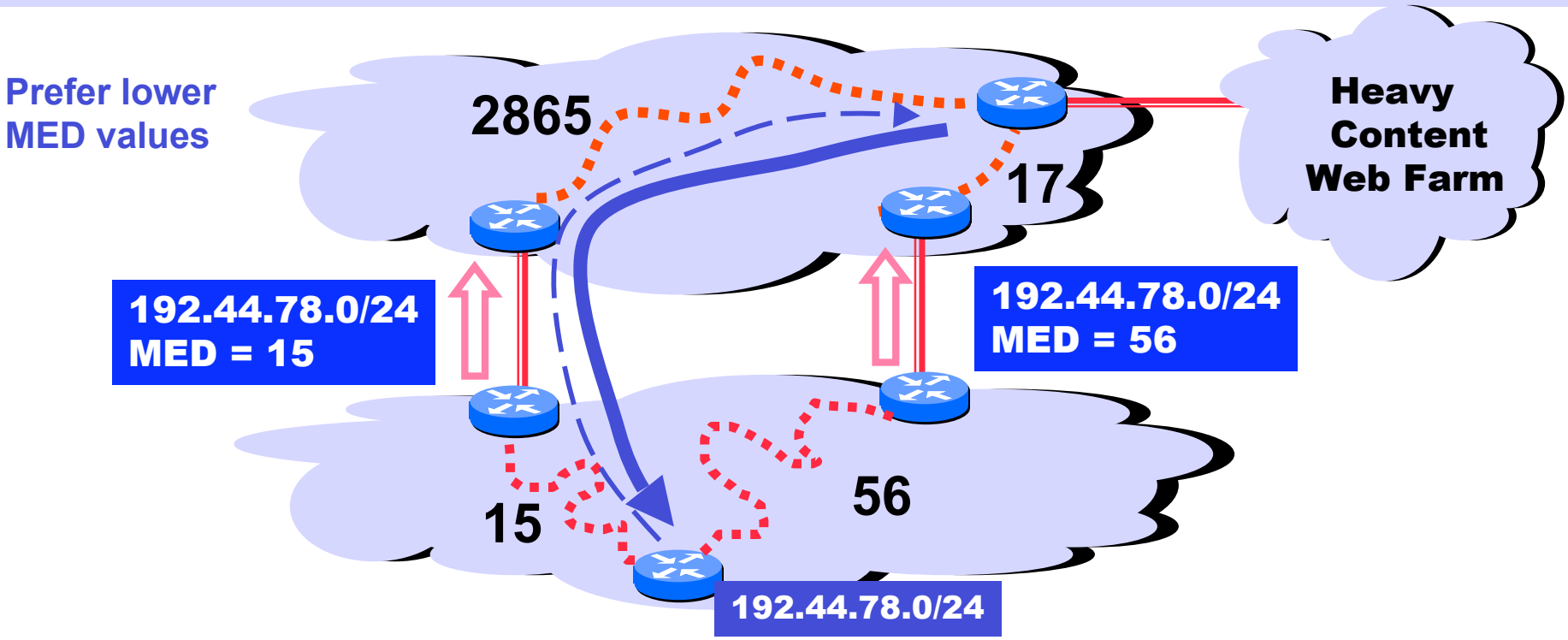


Getting Burned by the Hot Potato



Many customers want their provider to carry the bits!

Cold Potato Routing with MEDs (Multi-Exit Discriminator Attribute)

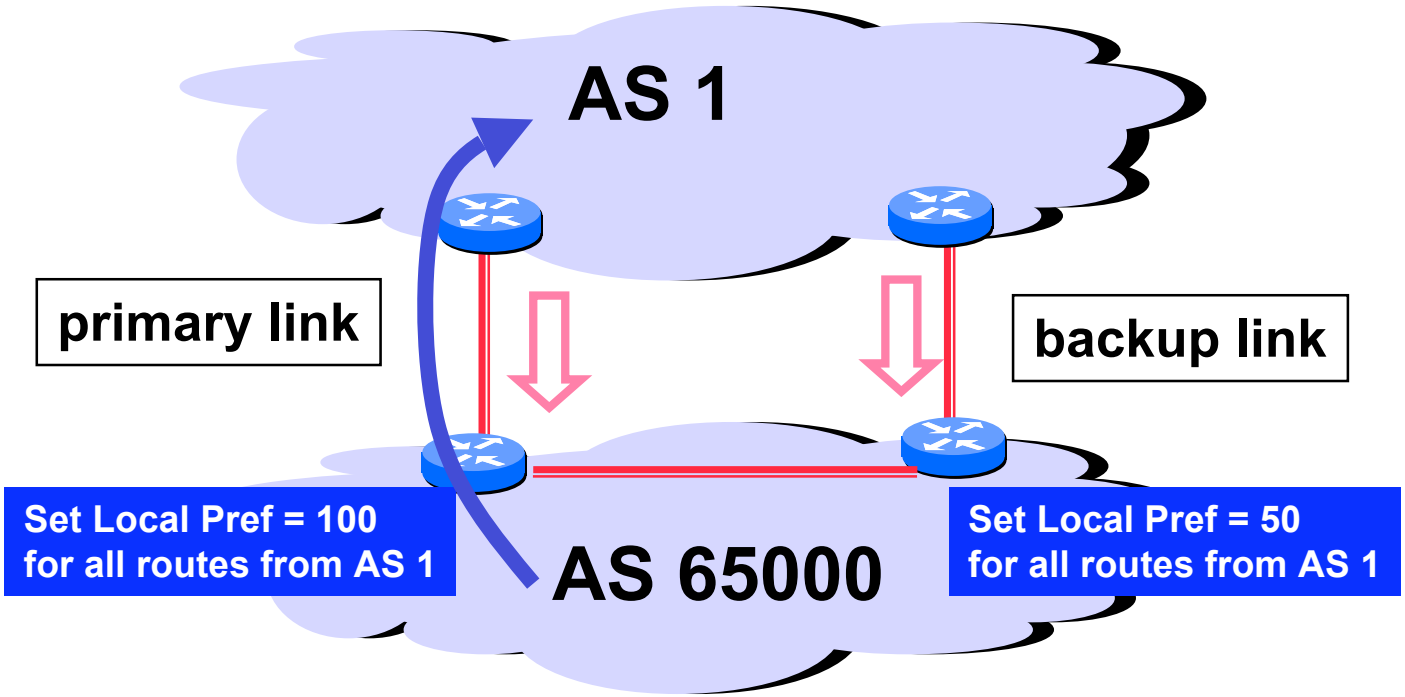


This means that MEDs must be considered BEFORE IGP distance!

Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

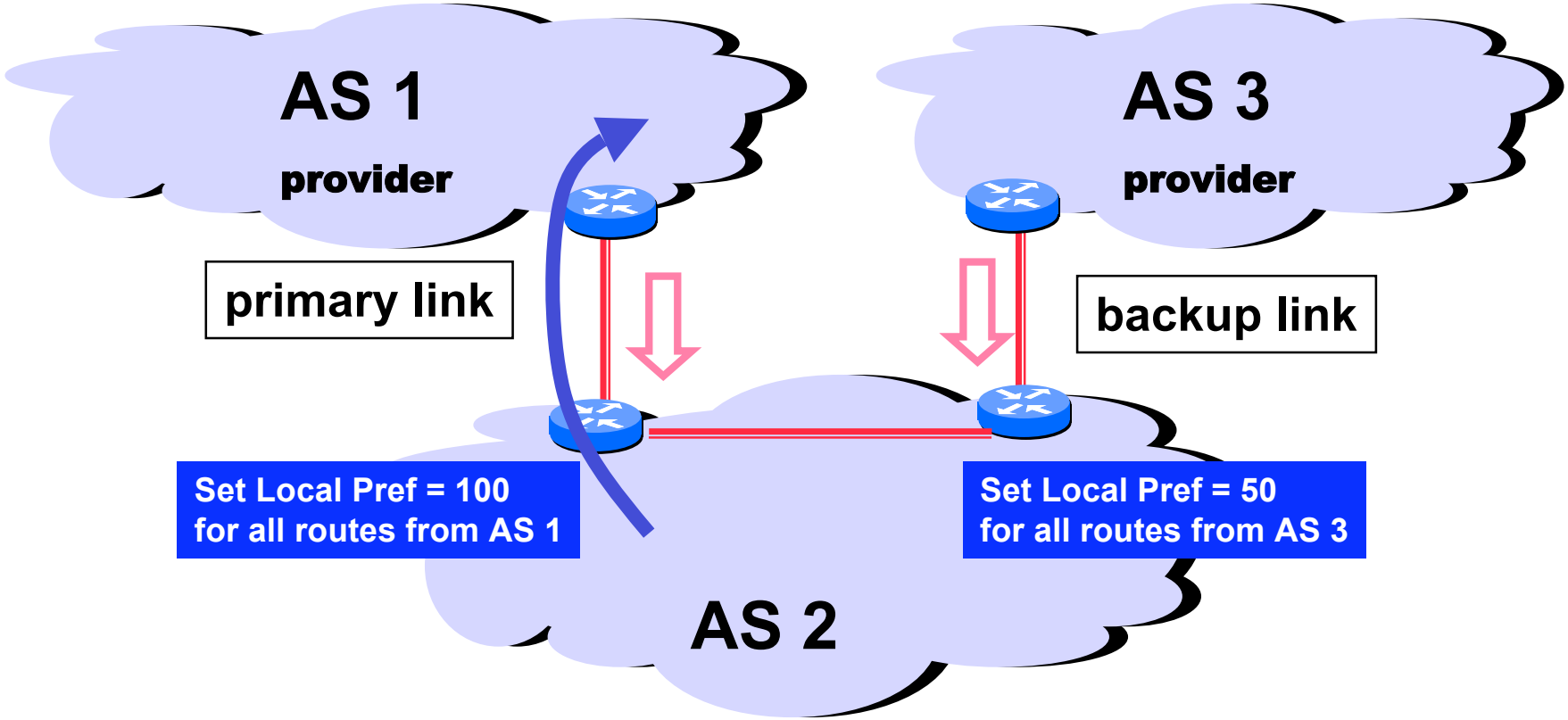
Implementing Backup Links with Local Preference (Outbound Traffic)



Forces outbound traffic to take primary link, unless link is down.

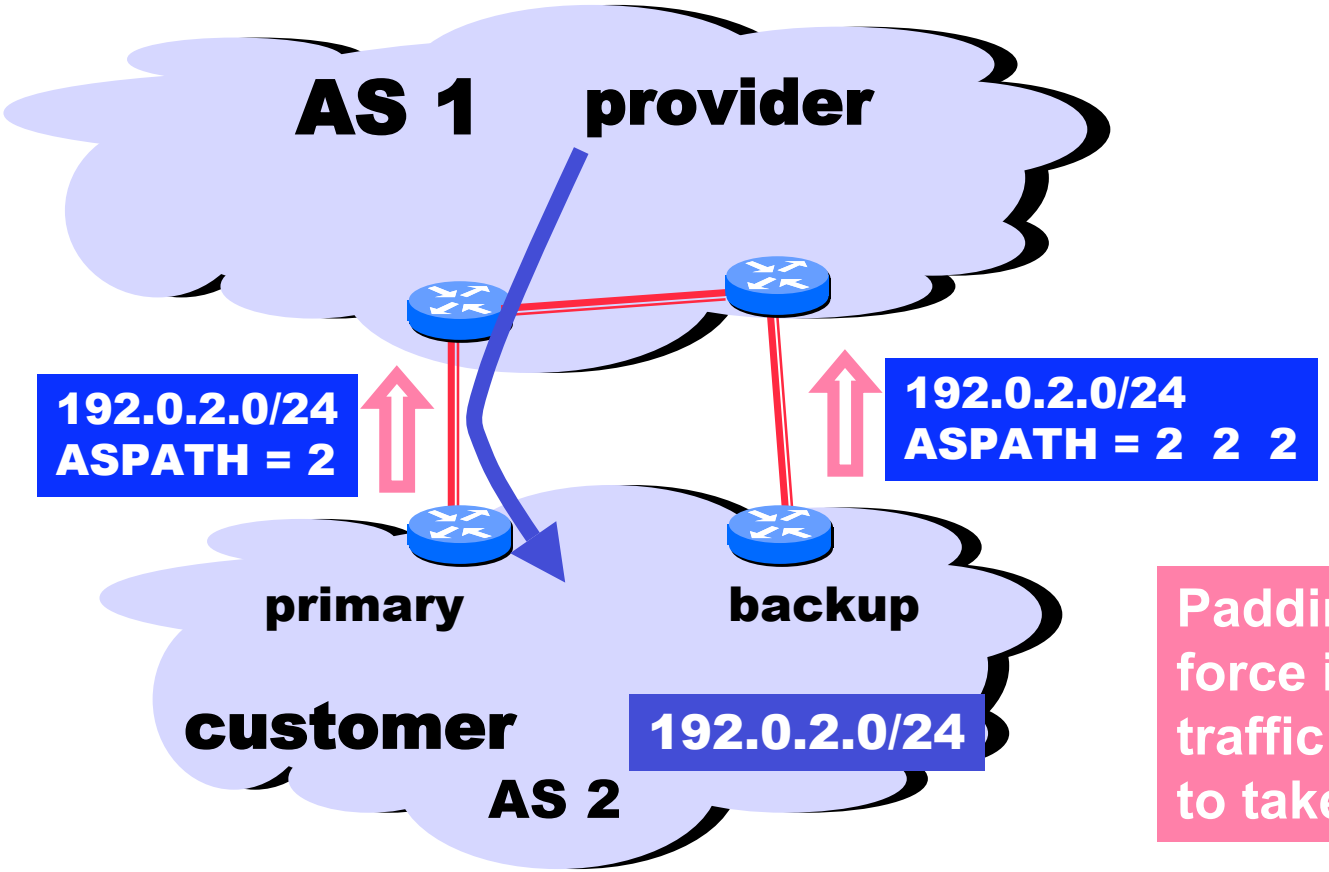
We'll talk about inbound traffic soon …

Multihomed Backups (Outbound Traffic)



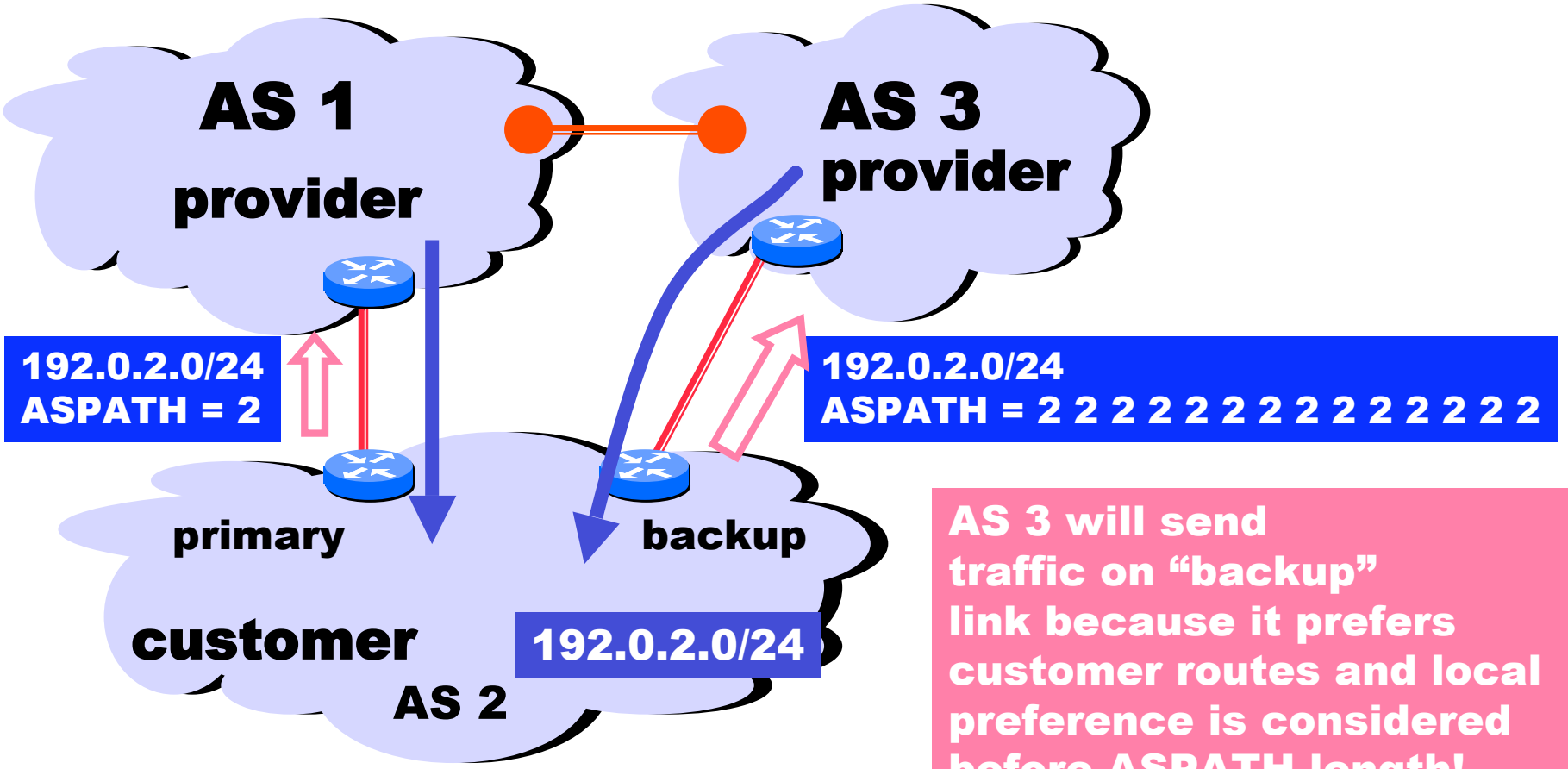
Forces outbound traffic to take primary link, unless link is down.

Shedding Inbound Traffic with ASPATH Padding. Yes, this is a Glorious Hack ...



Padding will (usually) force inbound traffic from AS 1 to take primary link

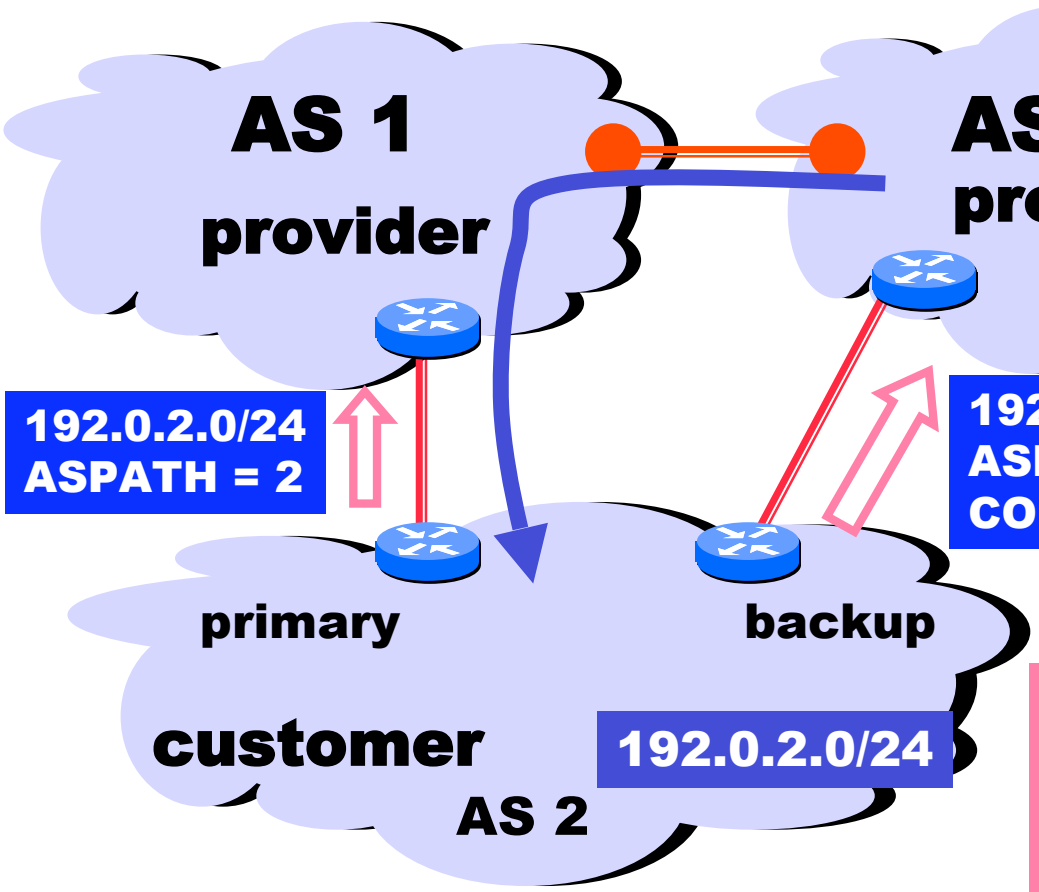
… But Padding Does Not Always Work



AS 3 will send traffic on "backup" link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing

COMMUNITY Attribute to the Rescue!



AS 3: normal customer local pref is 100, peer local pref is 90

Customer import policy at AS 3:
If 3:90 in COMMUNITY then
set local preference to 90
If 3:80 in COMMUNITY then
set local preference to 80
If 3:70 in COMMUNITY then
set local preference to 70

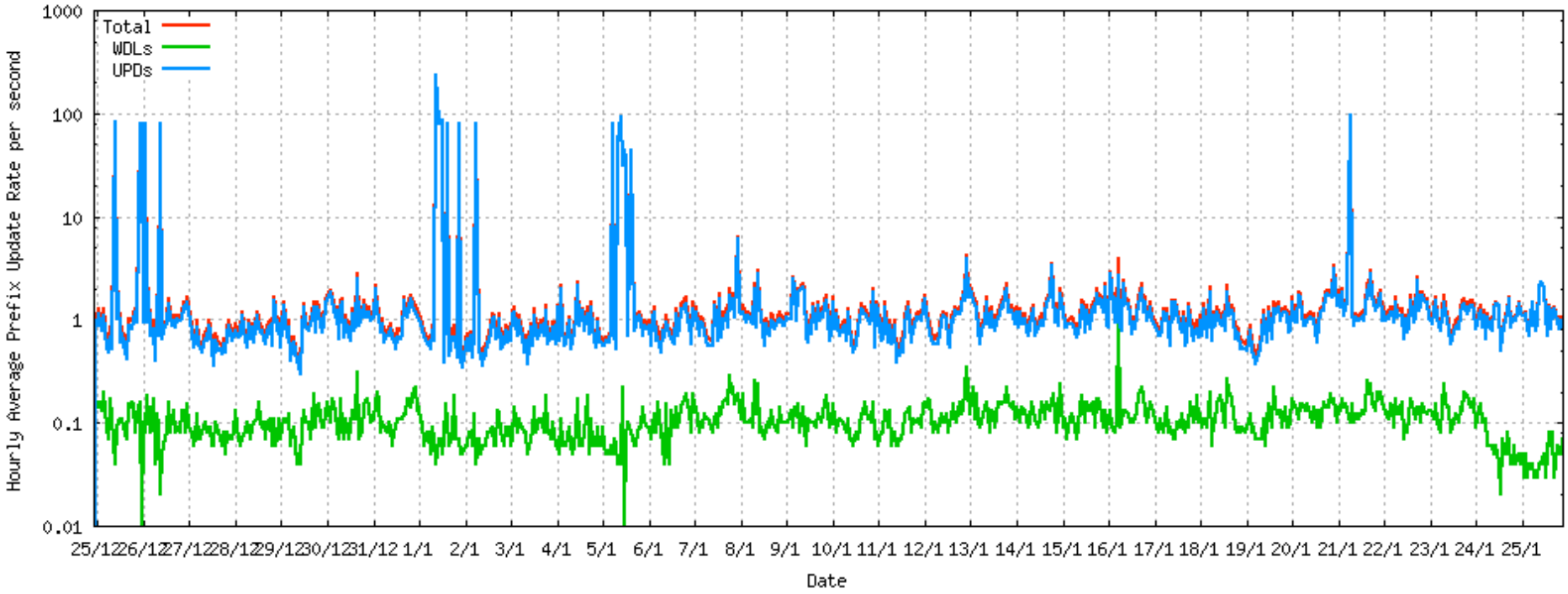
BGP Dynamics

- How many updates are flying around the Internet?
- How long Does it take Routes to Change?

The goals of
(1) fast convergence
(2) minimal updates
(3) path redundancy
are at odds.

Pick any two!!

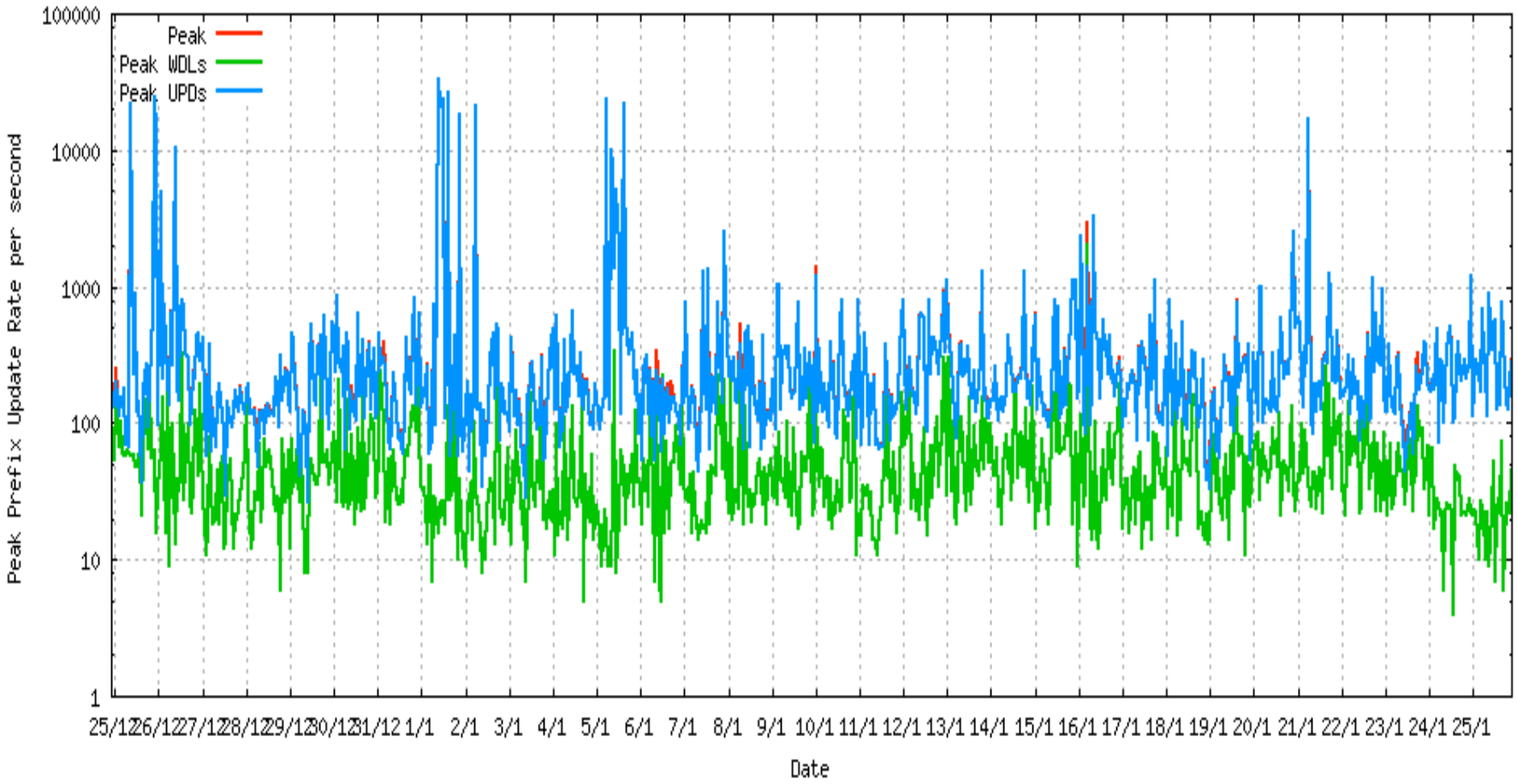
Hourly Average of Per-Second Updated and Withdrawn Prefix Rate



http://bgpupdates.potaroo.net

Jan 26 2009

Hourly Peak of Per-Second Updated and Withdrawn Prefix Rate



Results will vary depending on location…

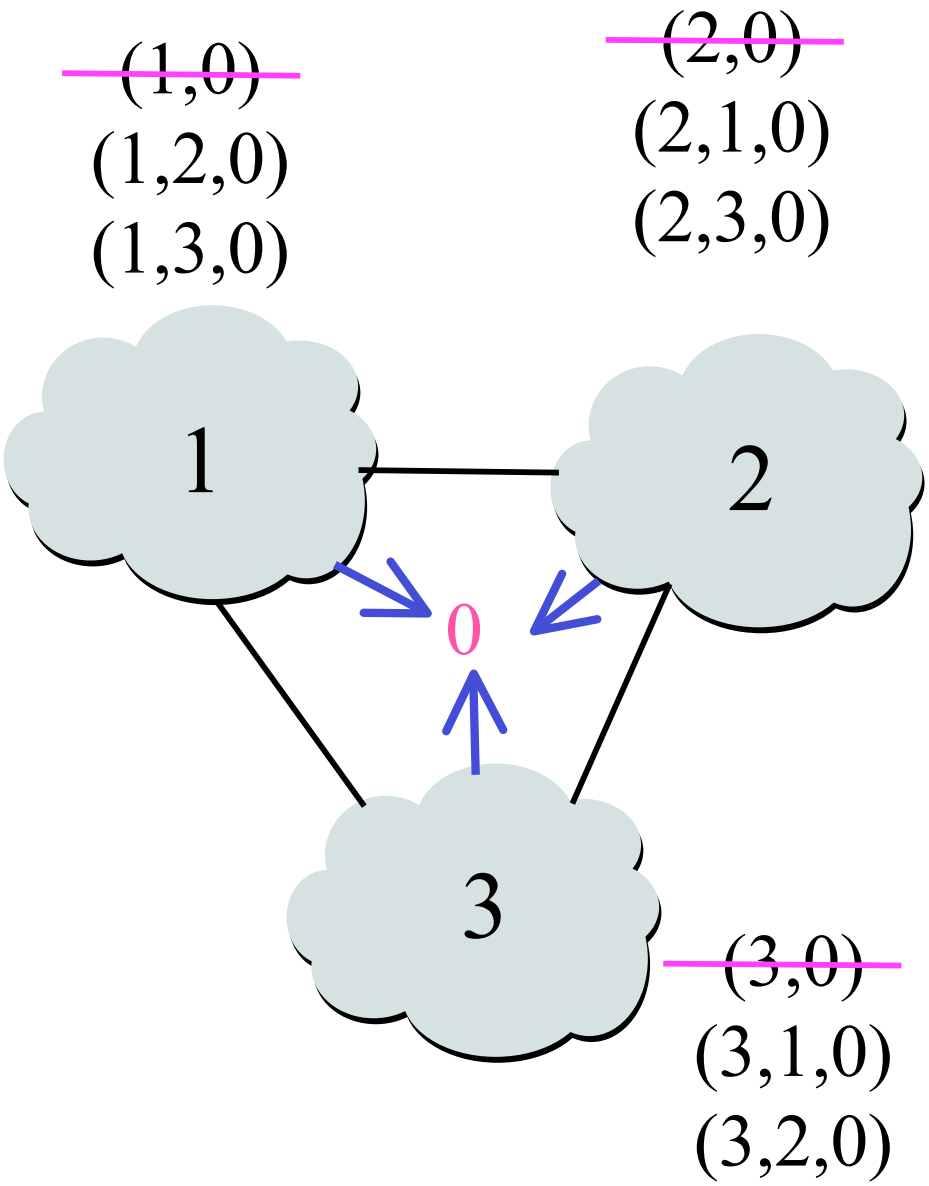
Q: Why All the Updates?

- The Internet is large, so isn't there always something going on somewhere? (That is, BGP is just doing a good job of keeping things connected!)
- Is BGP exploring many alternate paths during convergence?
- Are IGP instabilities are being exported to the interdomain world?
- Have bad tradeoffs beem made in router software implementation?
- Are BGP sessions being reset due to congestion?
- Weird policy interactions like MED oscillation?
- Gnomes, sprites, and fairies
- ….

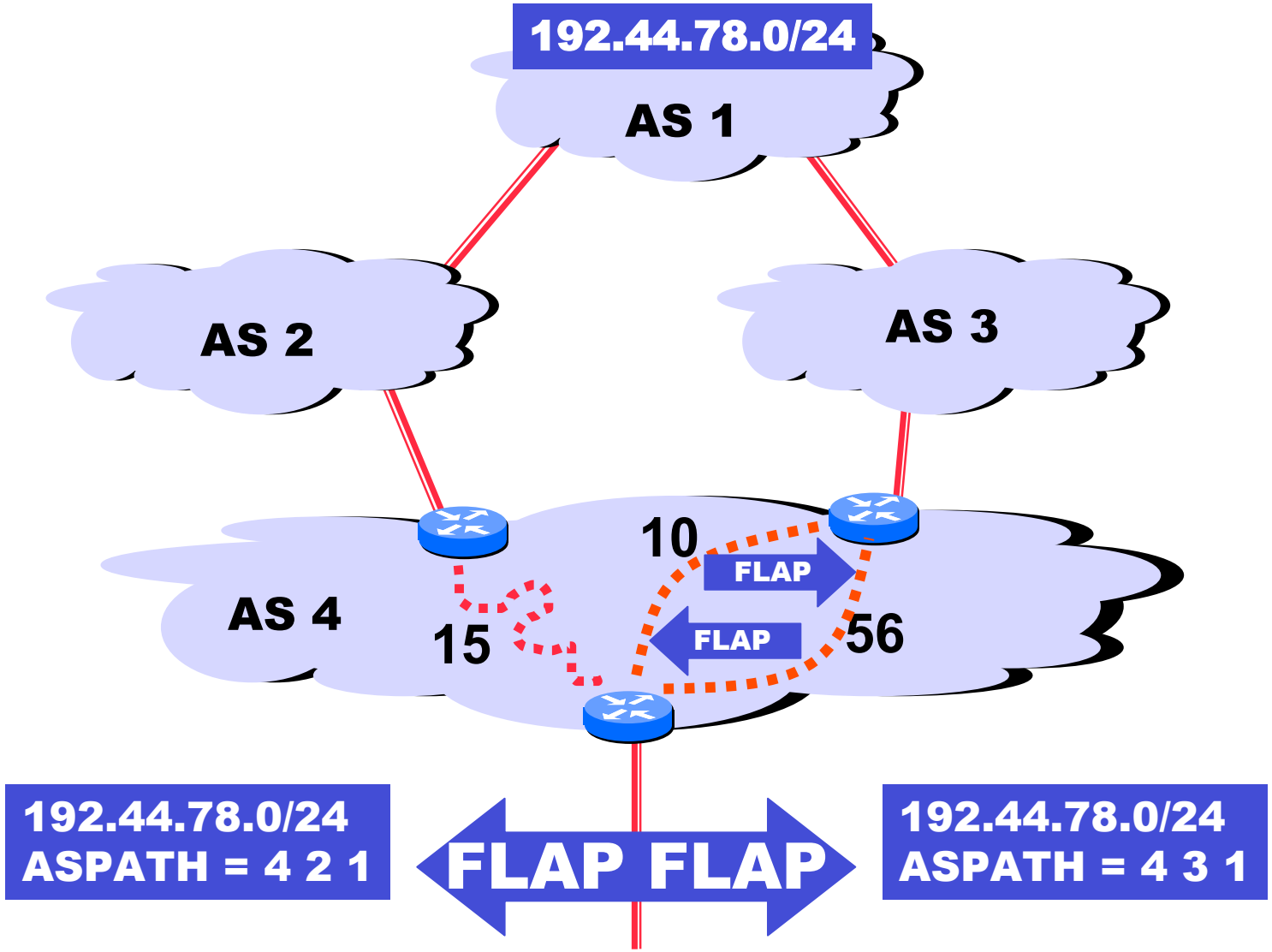
A: NO ONE REALLY KNOWS …
BGP does a very good job hiding information!

Routing Change: Path Exploration

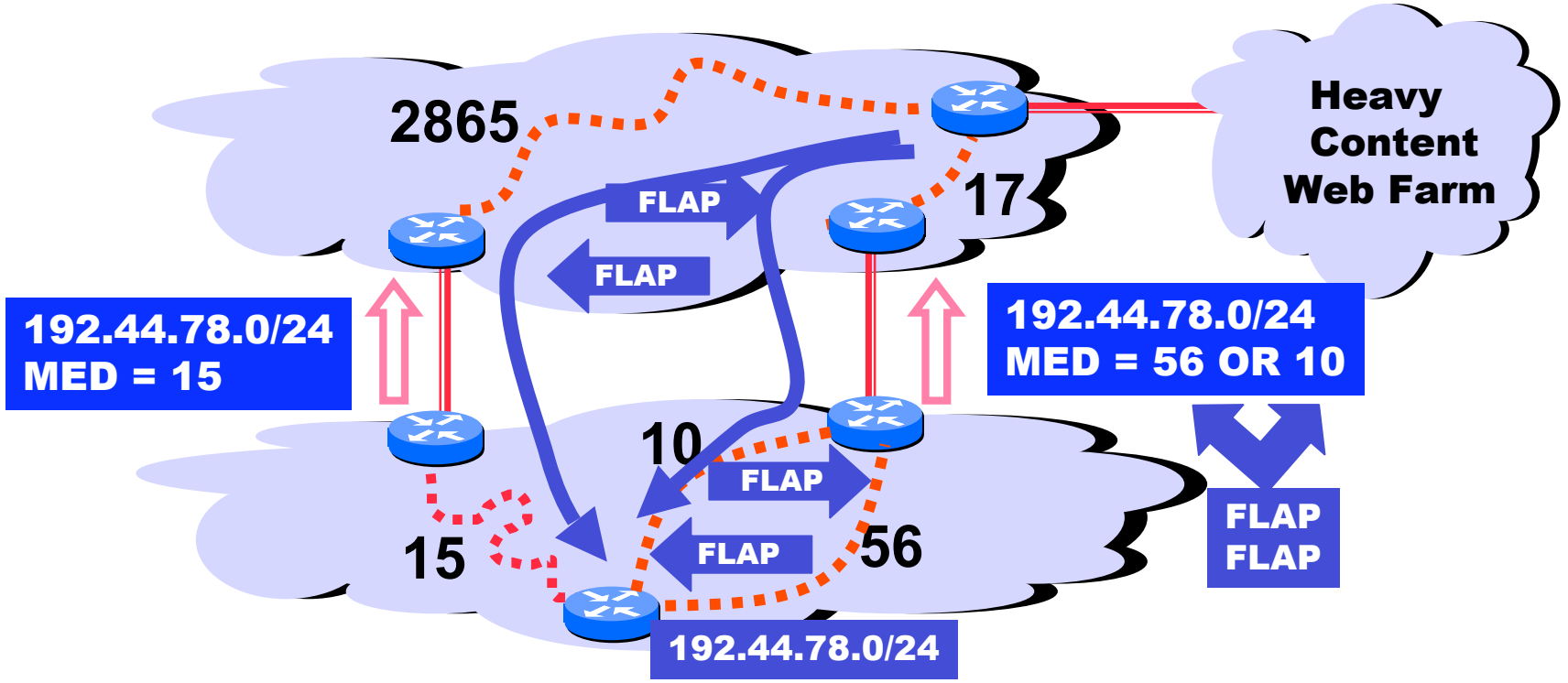
- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All switch to longer paths
  - Eventually withdrawn
- E.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



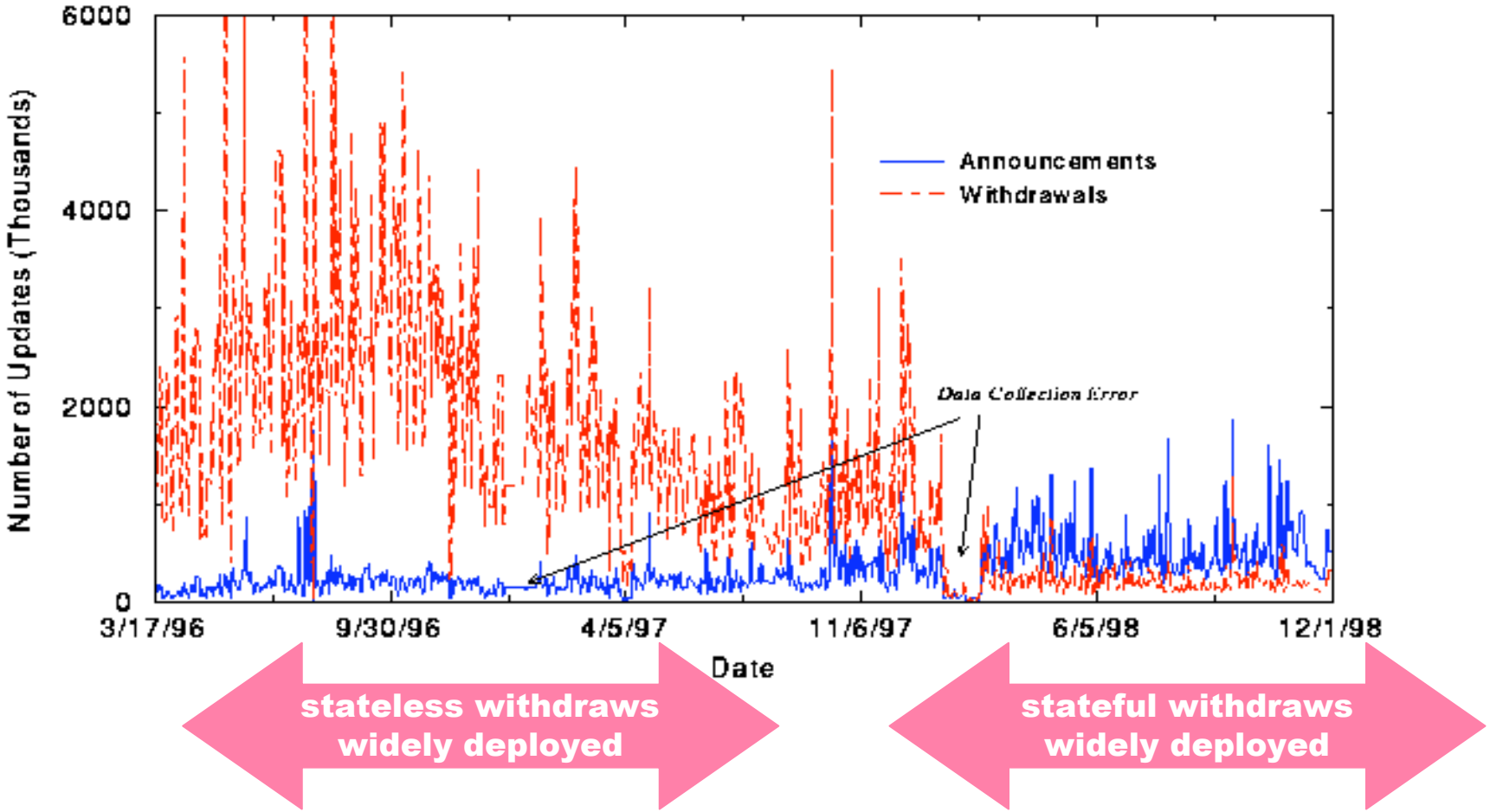
IGP Tie Breaking Can Export Internal Instability to the Whole Wide World



MEDs Can Export Internal Instability

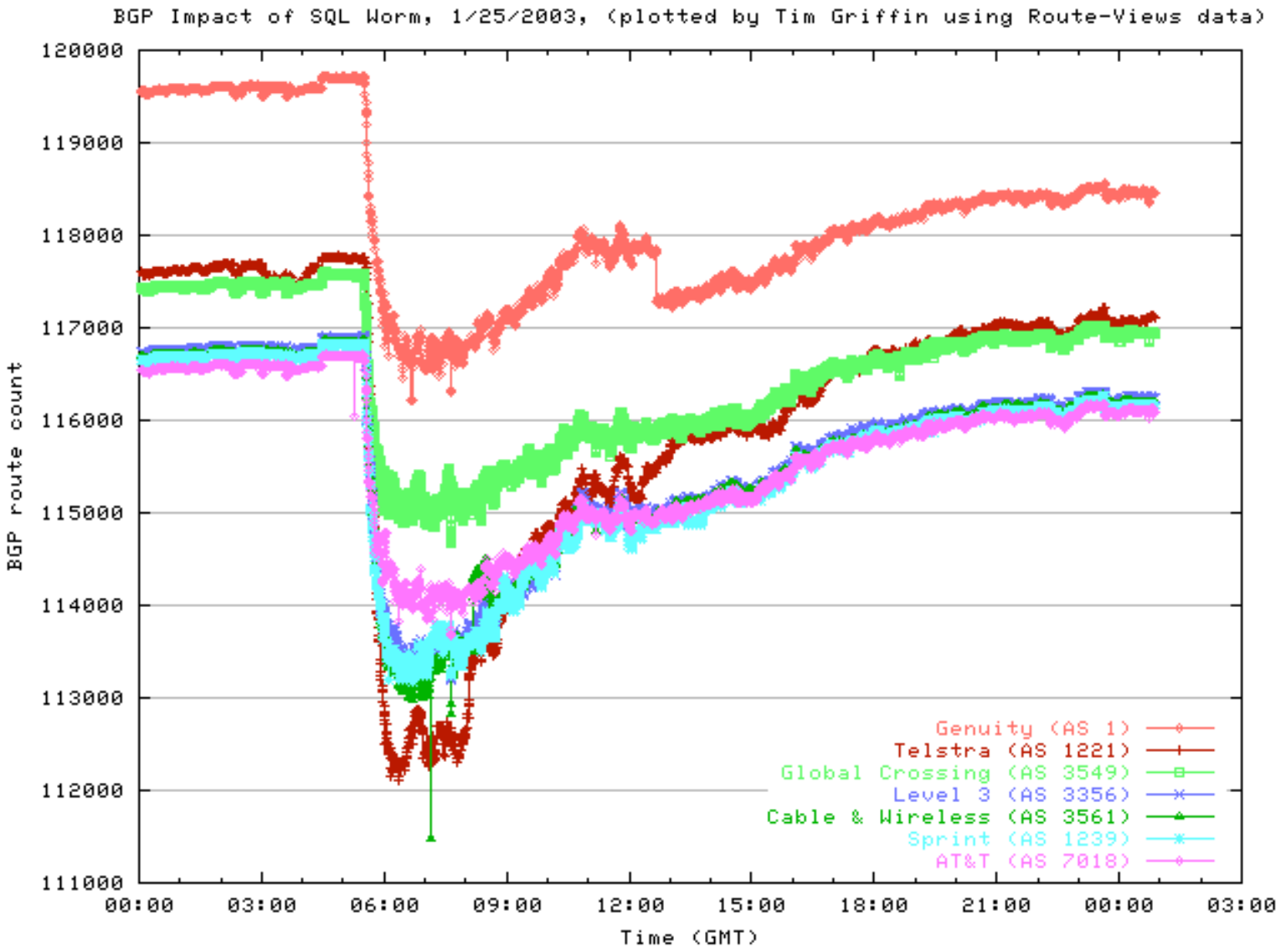


Implementation Does Matter!



Thanks to Abha Ahuja and Craig Labovitz for this plot.

Conjestion can take down BGP sessions! The SQL Slammer worm



Two BGP Mechanisms for Squashing Updates

- Rate limiting on sending updates
  - Send batch of updates every MinRouteAdvertisementInterval (MRAI) seconds (+/- random fuzz)
  - Default value is 30 seconds
  - A router can change its mind about best routes many times within this interval without telling neighbors

Effective in dampening oscillations inherent in the vectoring approach

- Route Flap Dampening
  - Punish routes for "misbehaving"

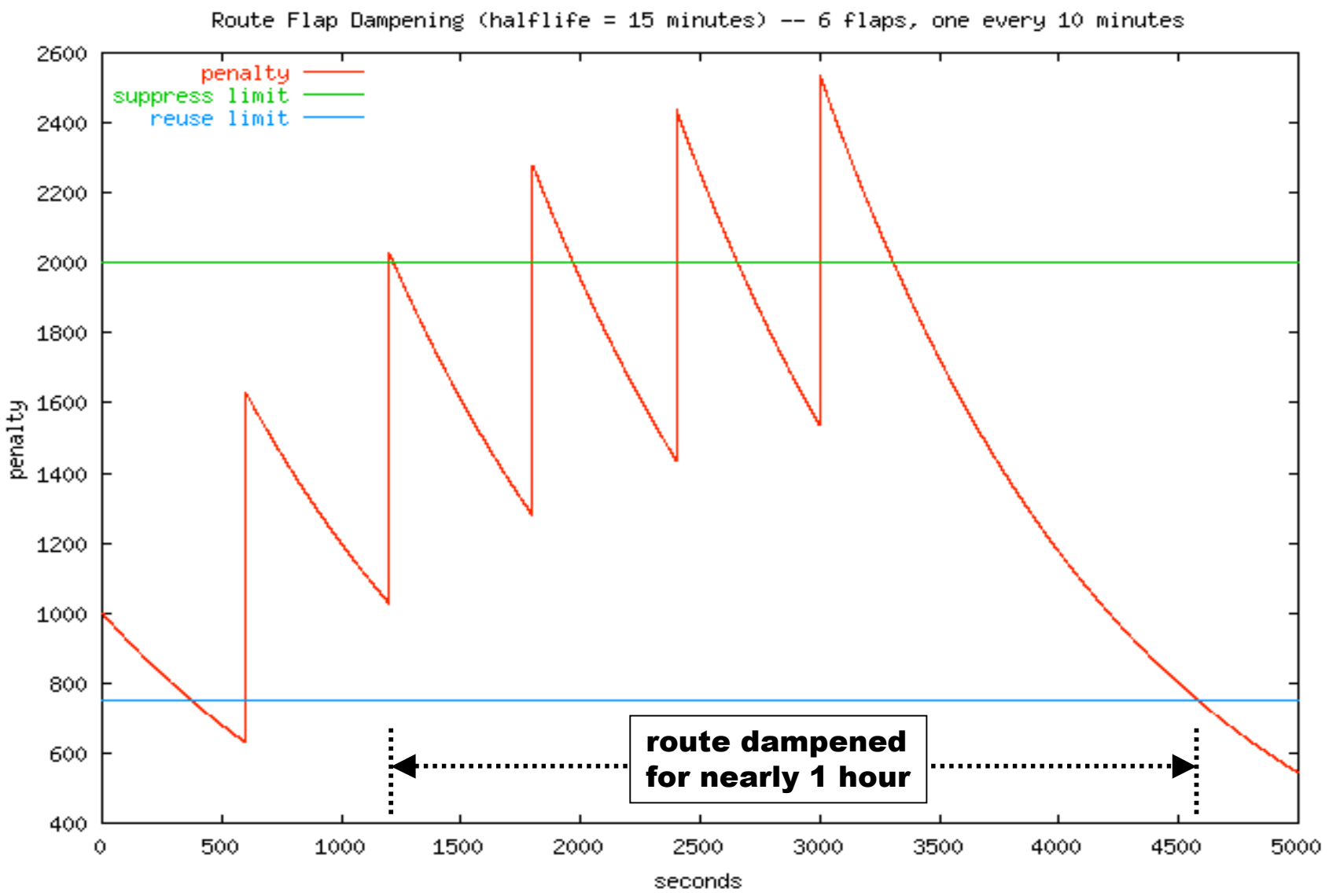
Must be turned on with configuration

Route Flap Dampening (RFC 2439)

**Routes are given a penalty for changing.
If penalty exceeds suppress limit, the
route is dampened. When the route is not changing,
its penalty decays exponentially. If the penalty goes
below reuse limit, then it is announced again.**

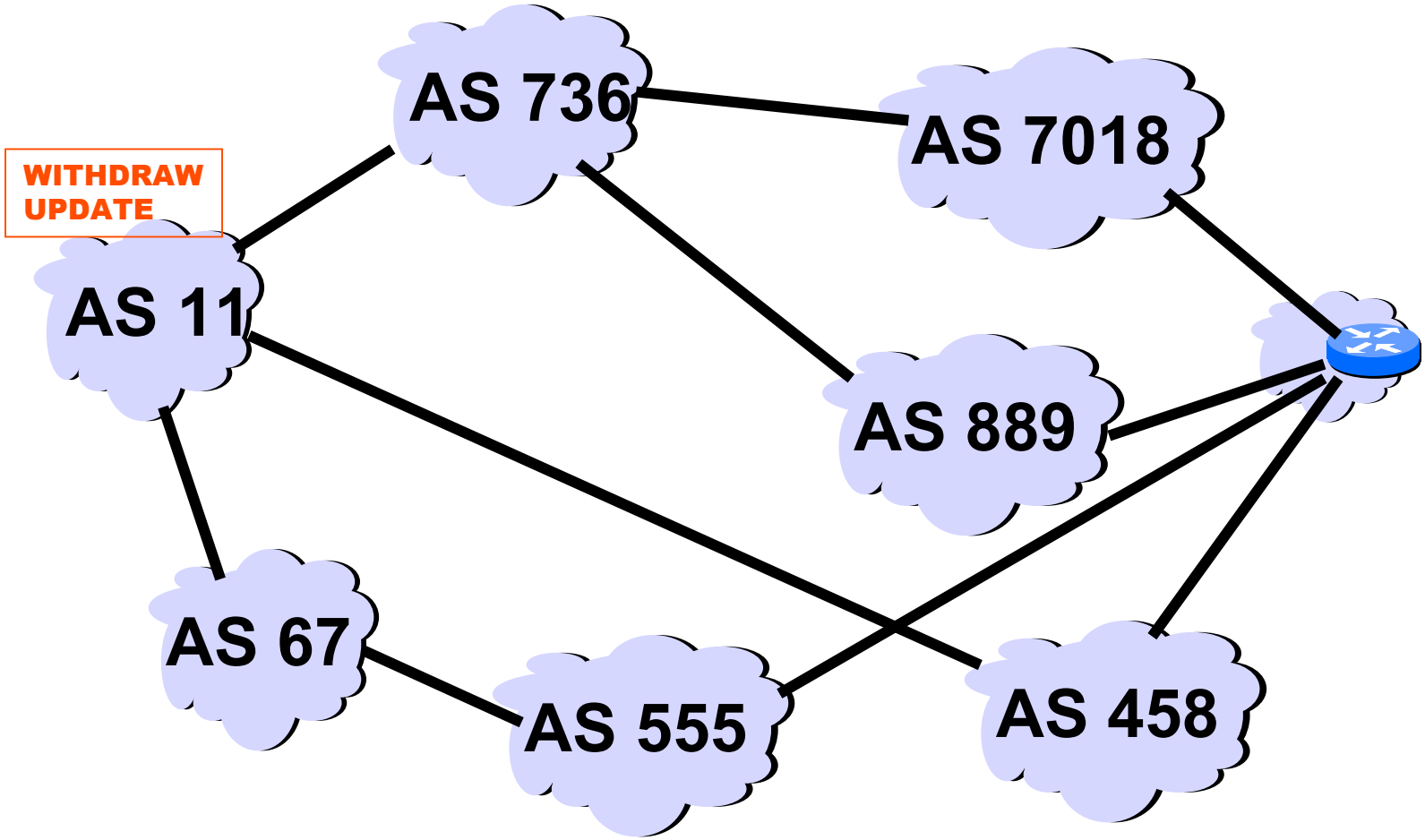
- **Can dramatically reduce the number of BGP updates**
- **Requires additional router resources**

How it works



penalty for each flap = 1000

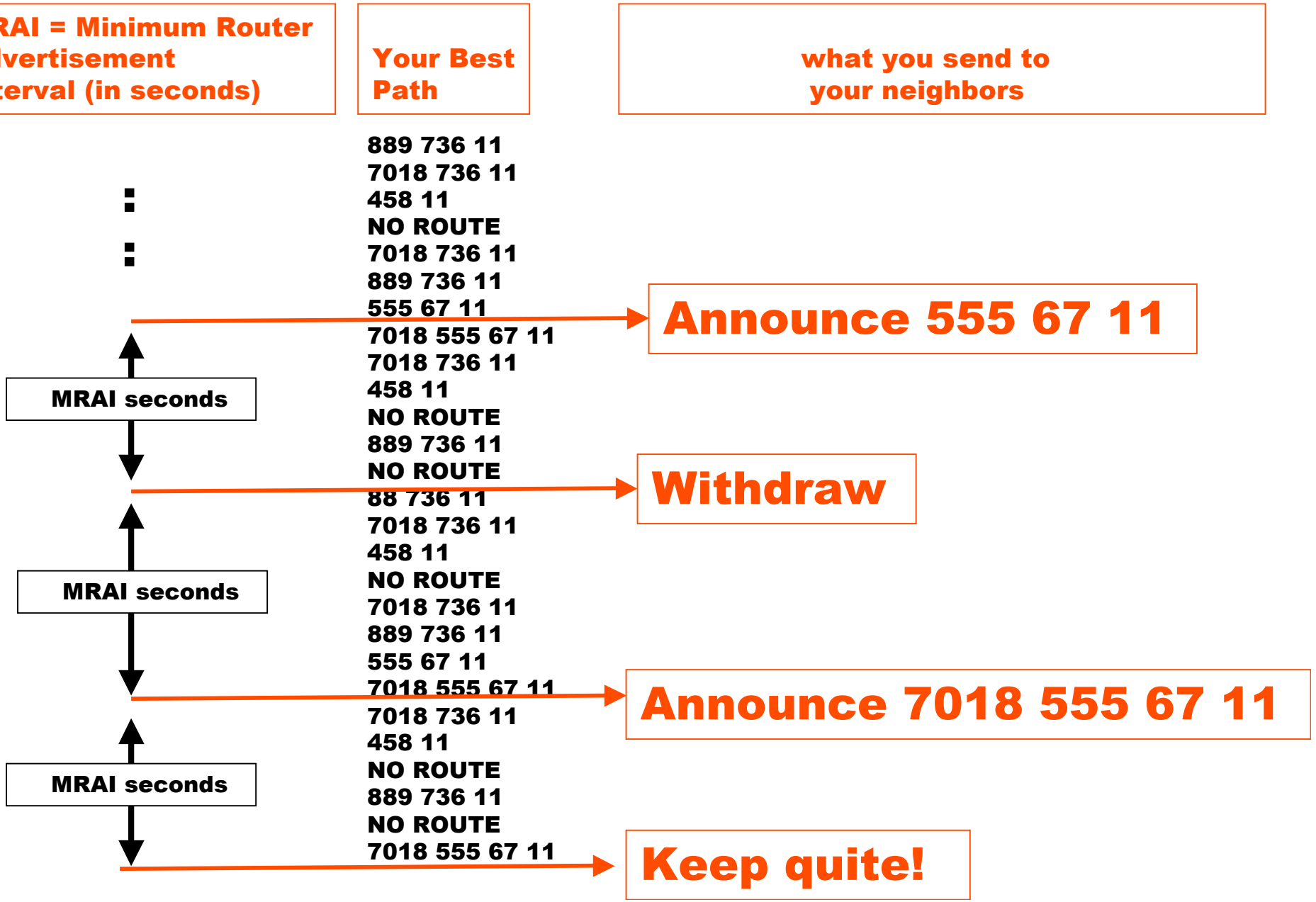
Problems with Flap Damping : punishes small updates for "well connected" destinations



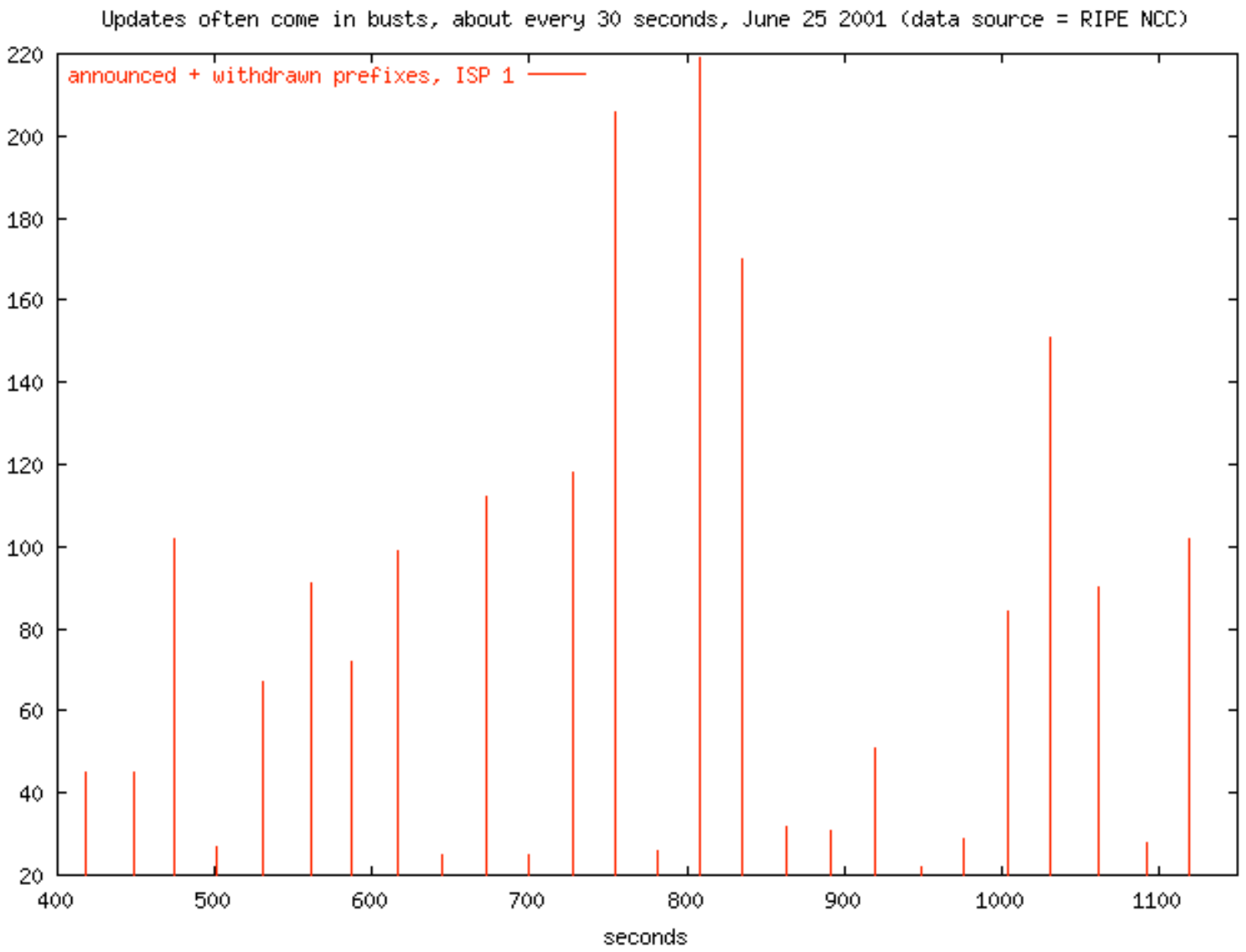
889 736 11
7018 736 11
458 11
NO ROUTE
7018 736 11
889 736 11
555 67 11
7018 555 67 11
7018 736 11
458 11
NO ROUTE
889 736 11
NO ROUTE
88 736 11
7018 736 11
458 11
NO ROUTE
7018 736 11
889 736 11
555 67 11
7018 555 67 11
7018 736 11
458 11
NO ROUTE
889 736 11
NO ROUTE
7018 555 67 11

The prefix is not "misbehaving" -- it is BGP!!

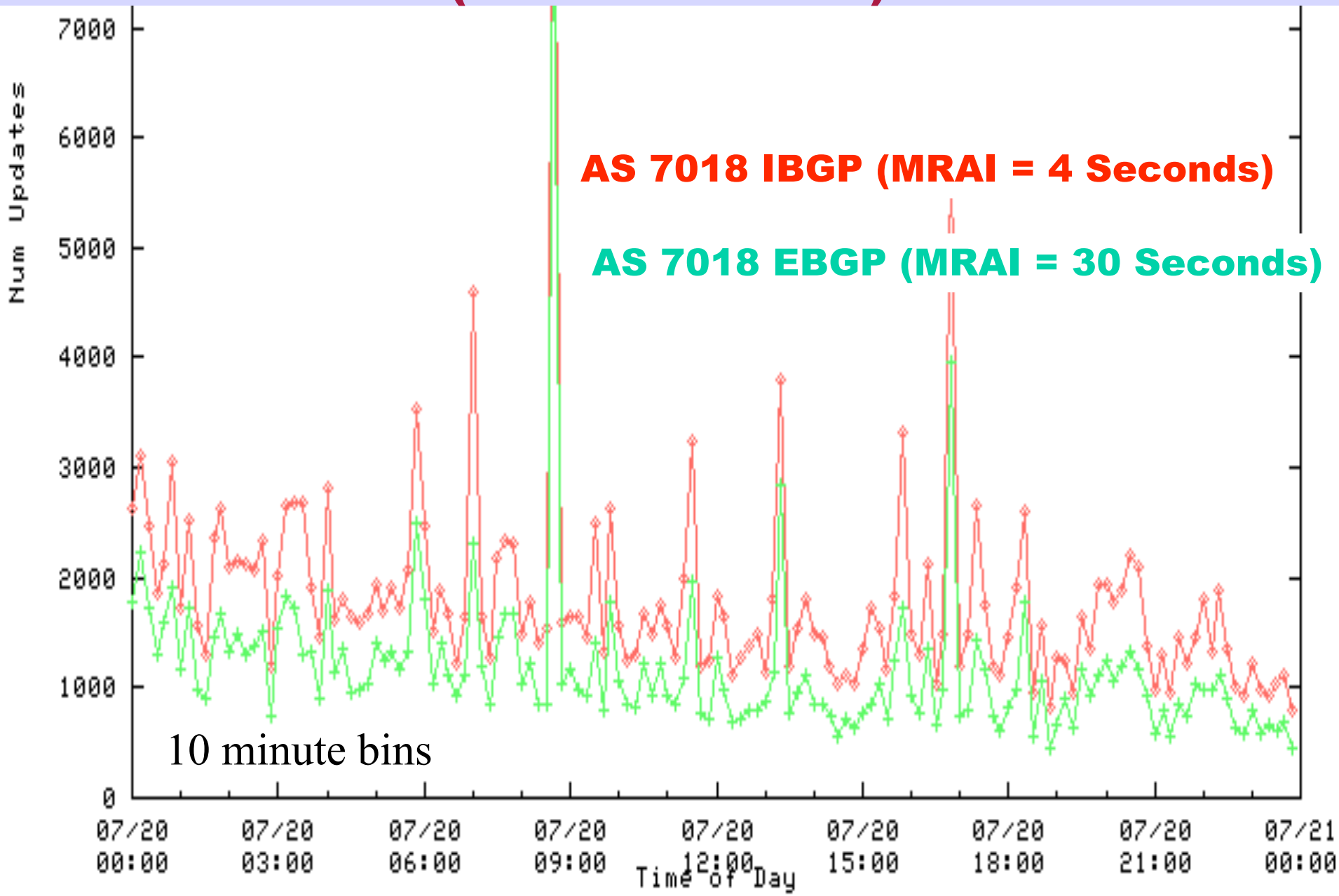
Rate Limiting



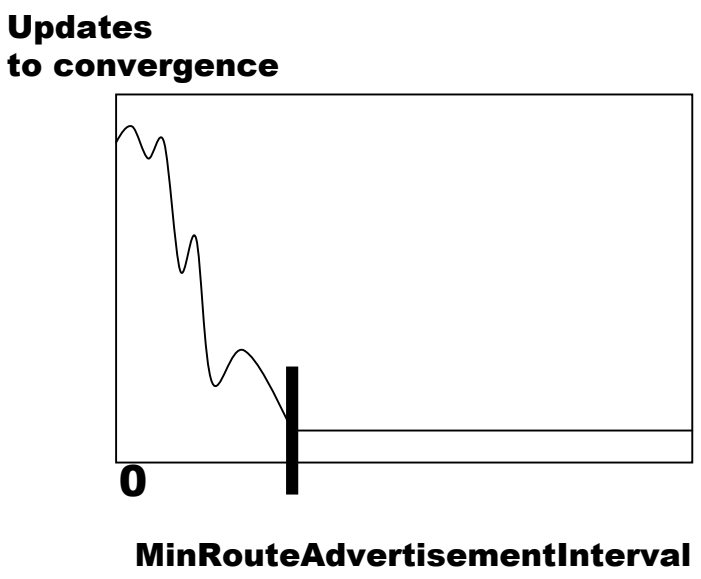
30 Second Bursts

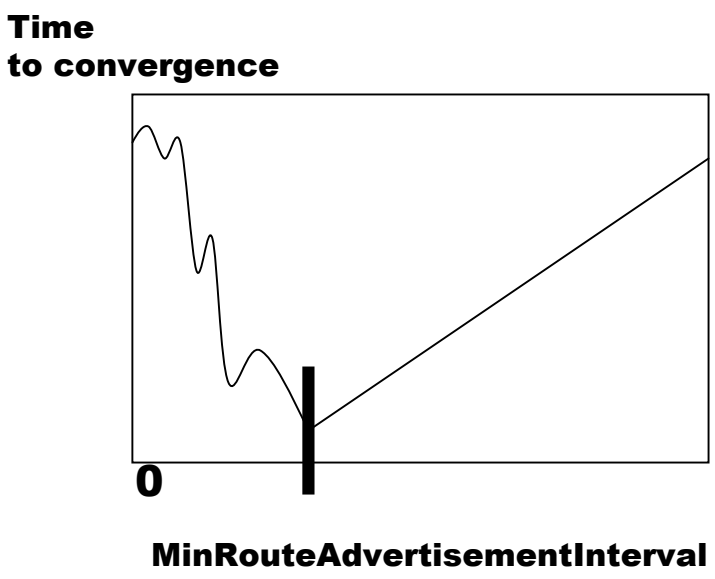


Rate limiting in action (IBGP vs EBGP)



Why is Rate Limiting Needed?





Rate limiting dampens some of the oscillation inherent in a vectoring protocol.

SSFNet (www.ssfnet.org) simulations,
An Experimental Analysis of BGP Convergence Time ---T. Griffin and B.J. Premore. ICNP 2001.

Two Main Factors in Delayed Convergence

- BGP can explore many alternate paths before giving up or arriving at a new path
  - No global knowledge in vectoring protocols
- Rate limiting timer slows everything down

Current interval (30 seconds) was picked "out of the blue sky"