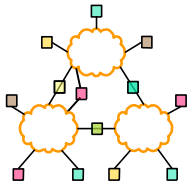


Computer Networks

(Graduate level)

Lecture 5: **Inter-domain Routing**

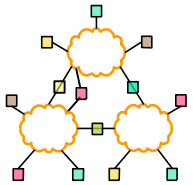
University of Tehran

Dept. of EE and Computer Engineering

By:

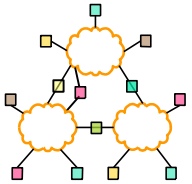
Dr. Nasser Yazdani

Inter-Domain Routing



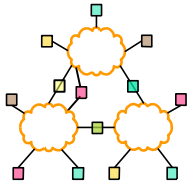
- Border Gateway Protocol (BGP)
- Sources
  - "Resolving Inter-Domain Policy Disputes", Cheng Tien Ee, atl.
  - Hari Balakrishnan notes about Interdomain from MIT
  - Craig Labovitz and .. "Delayed internet Routing Convergence"
  - HLP: A Next Generation Interdomain Routing Protocol
  - "Some Foundational Problems in Interdomain Routing"
  - Consensus Routing: The Internet as a Distributed System.
  - **BGP Security in Partial Deployment**

Outline



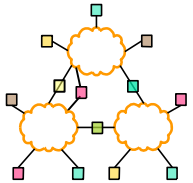
- External BGP (E-BGP)
- Internal BGP (I-BGP)
- Multi-Homing
- Stability Issues
- Scalability Issues
- Security

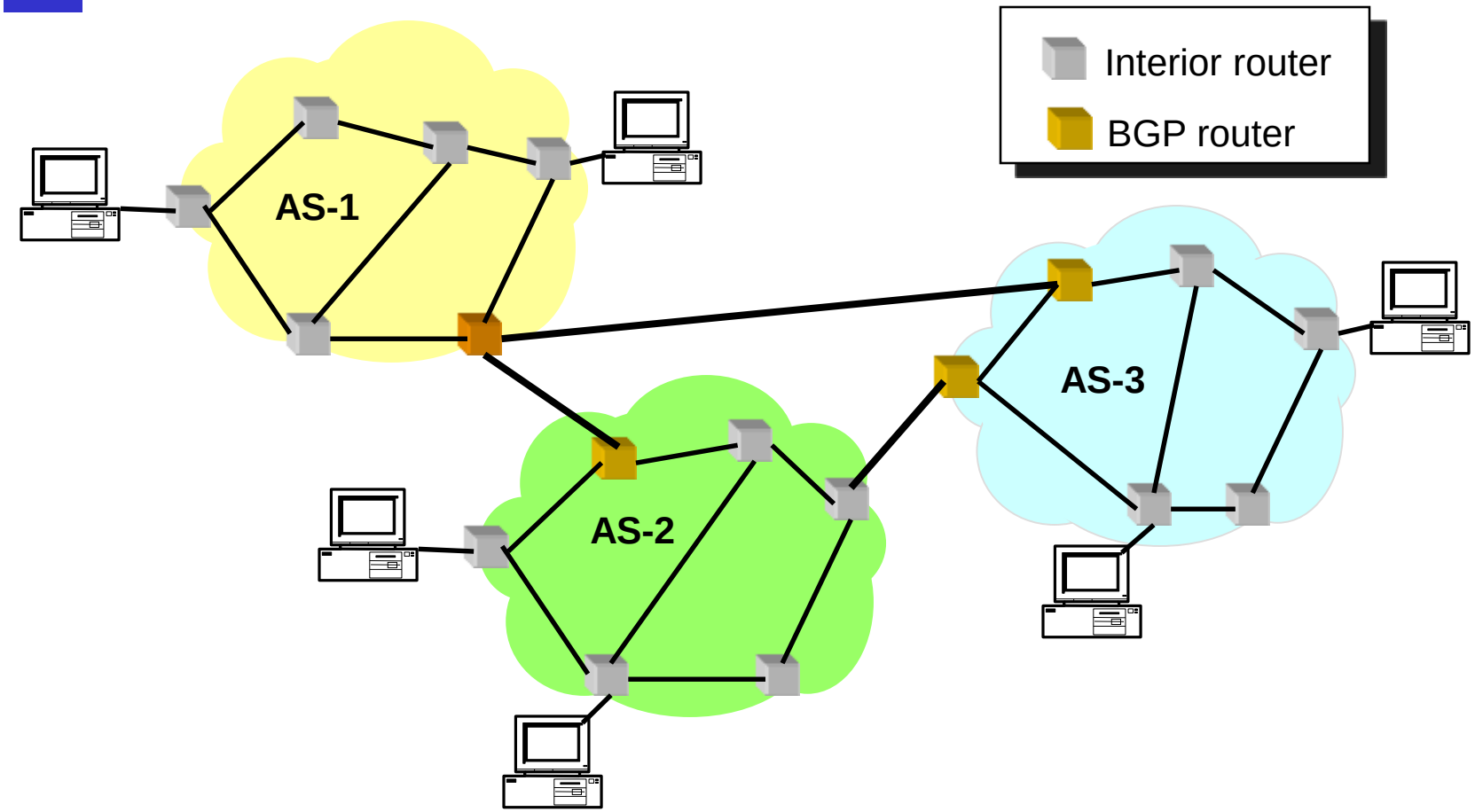
Internet Routing



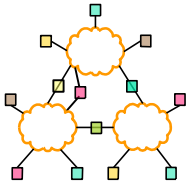
- Routing protocols like Distance vector or Link state suffer from different problems like convergence and Scaling due to the broadcasting routing packets.
- Internet organized as a two levels hierarchy
  - First level – autonomous systems (AS's)
    - AS – region of network under a single administrative domain
  - Each AS assigned unique ID
  - AS's peer at network exchange routing information.
  - AS's run an intra-domain routing protocols
    - Distance Vector, e.g., RIP
    - Link State, e.g., OSPF
  - Between AS's runs inter-domain routing protocols, e.g., Border Gateway Routing (BGP)
    - De facto standard today, BGP-4

Example

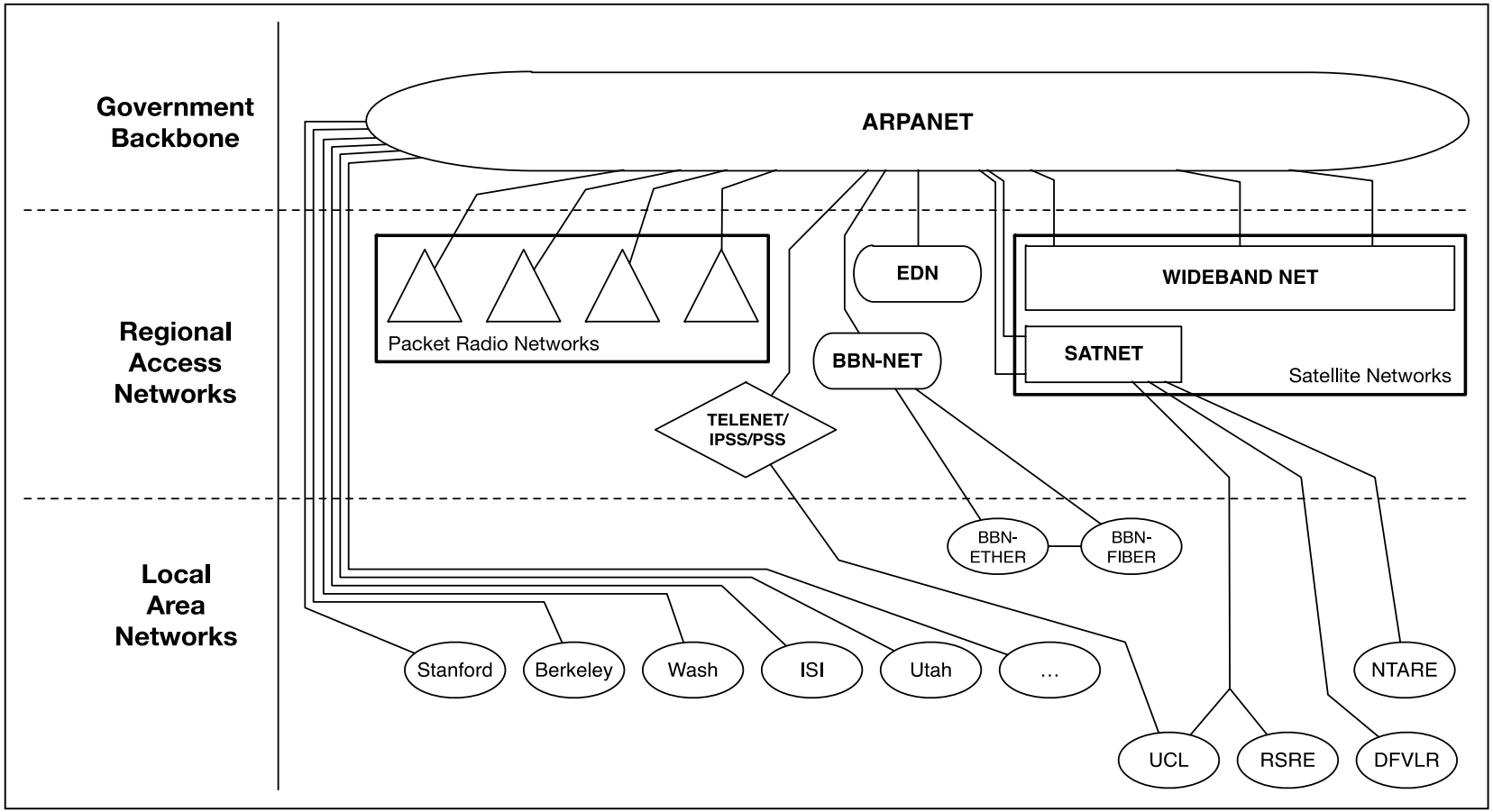




Interconnection Pre-1995

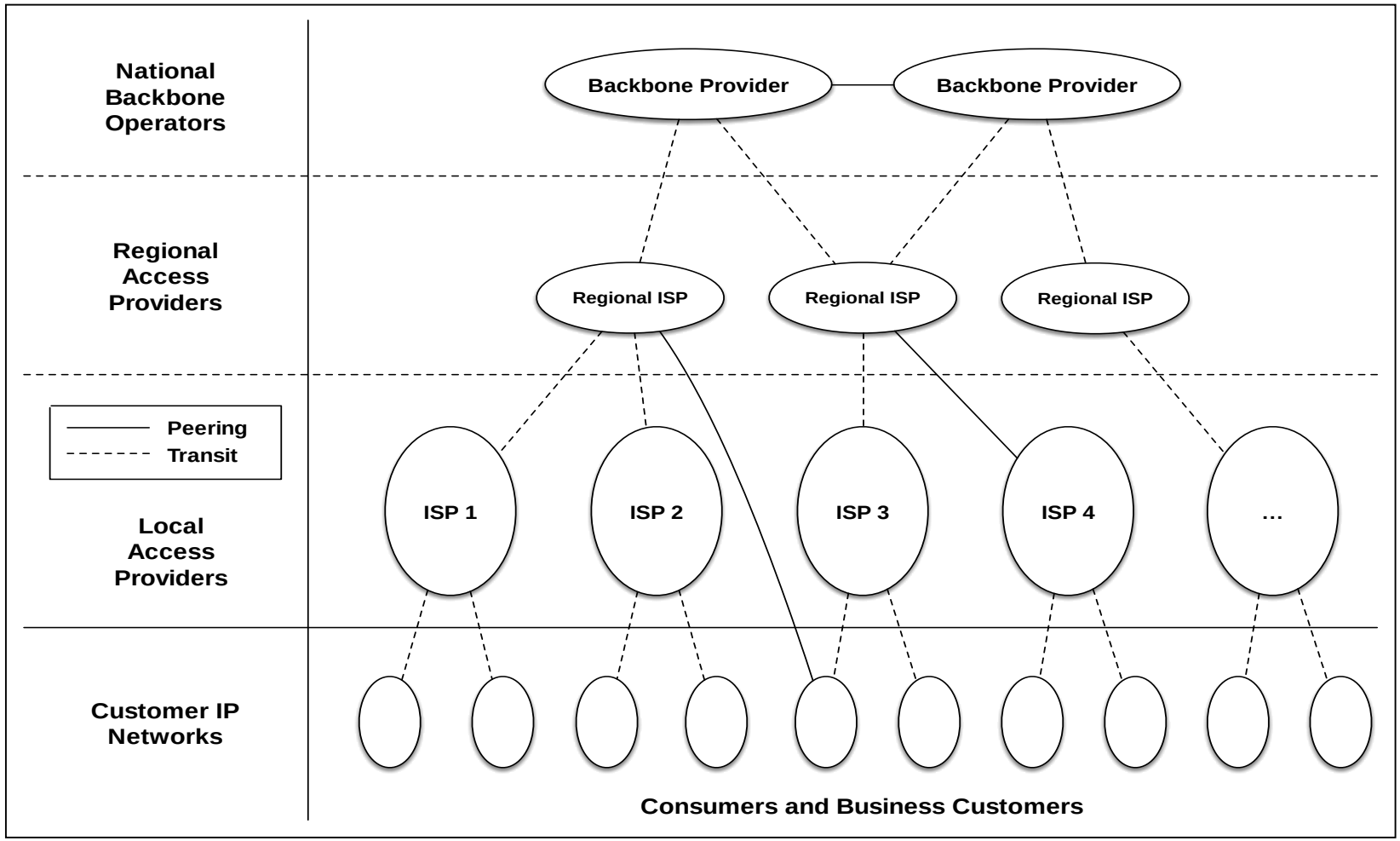


Network interconnection in the U.S. has evolved significantly since the early days of the Internet.

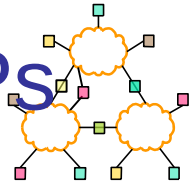


Interconnection Circa 1995-2005

The backbone eventually transitioned from a single government-operated backbone to a federated backbone model comprised of multiple commercial network operators.

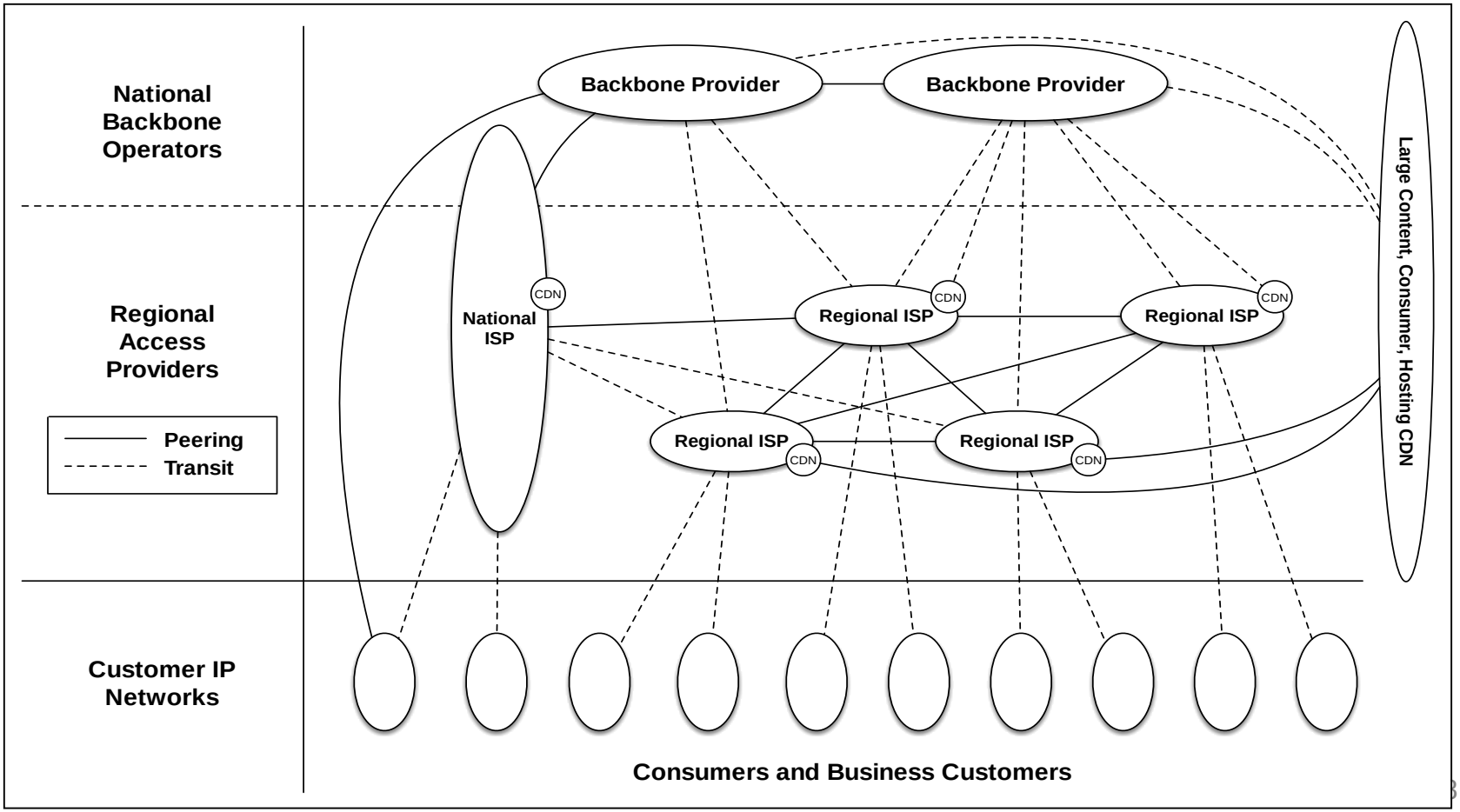


Today: Flattening due to CDNs and IXPs

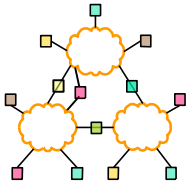


Evolved into a "complex amalgam of models incorporating new connectivity options, delivery options, traffic management requirements and business practices"

https://www.bitag.org/documents/Interconnection-and-Traffic-Exchange-on-the-Internet.pdf

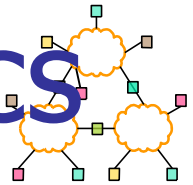


History



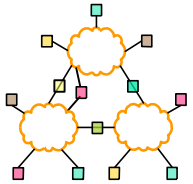
- Mid-80s: EGP
  - Reachability protocol (no shortest path)
  - Did not accommodate cycles (tree topology)
  - Evolved when all networks connected to NSF backbone
- Result: BGP introduced as routing protocol
  - Latest version = BGP 4
  - BGP-4 supports CIDR
  - Primary objective: connectivity not performance

Inter-domain Routing basics



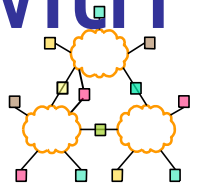
- Internet is composed of over more than 16000 autonomous systems
- **BGP** = **B**order **G**ateway **P**rotocol
  - Is a **Policy-Based** routing protocol
    - Each domain is free to specify inside its routing policy
  - Is the **de facto inter-domain routing protocol** of today's global Internet
- Relatively simple but configuration is complex and the entire world can see, and be impacted by, your mistakes.

Routing Choices



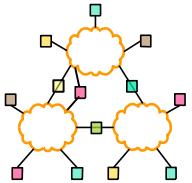
- Link state or distance vector?
  - No universal metric – policy decisions
- Problems with distance-vector:
  - Bellman-Ford algorithm may not converge
- Problems with link state:
  - Metric used by routers not the same – loops
  - LS database too large – entire Internet
  - May expose policies to other AS's

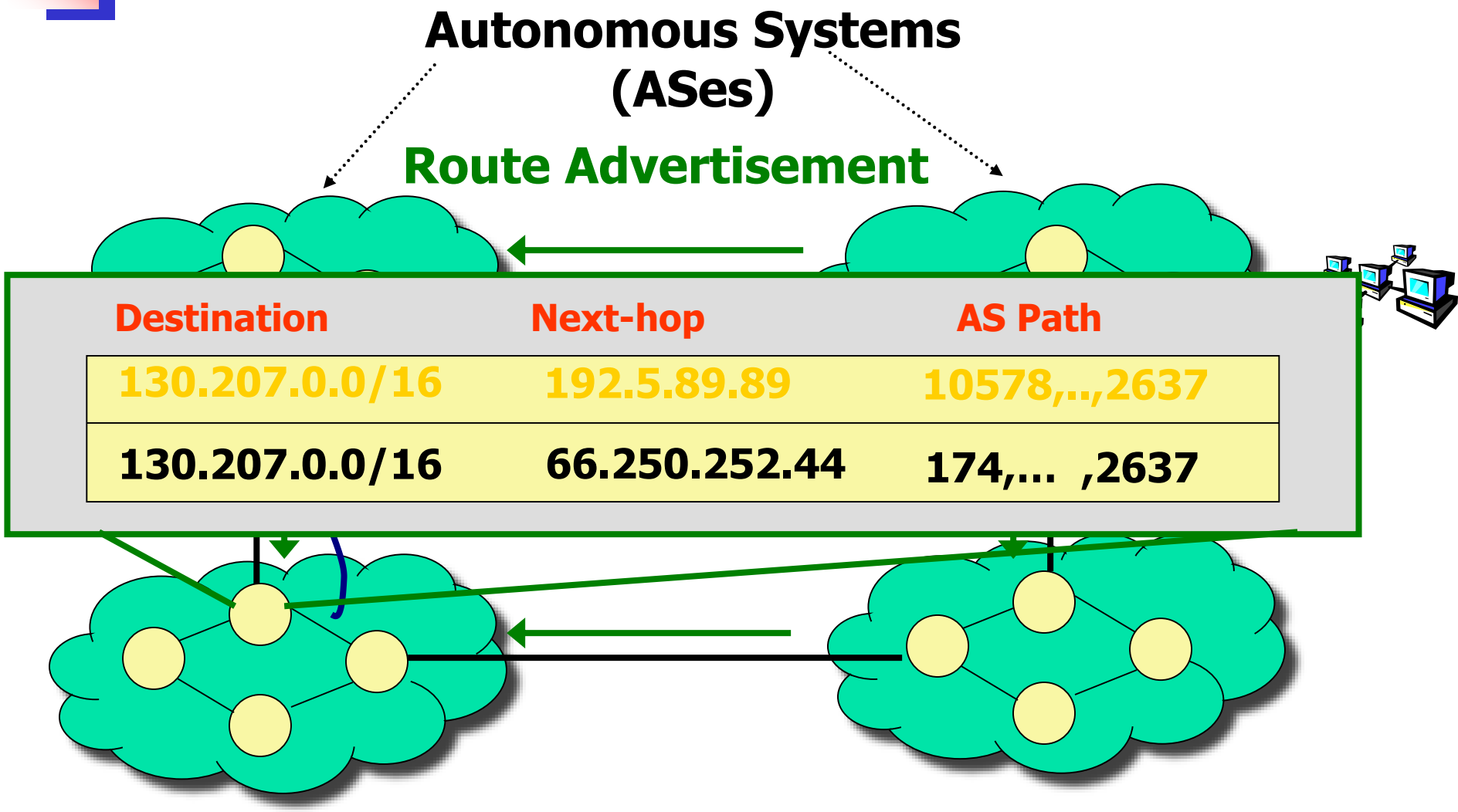
Solution: Distance Vector with Path



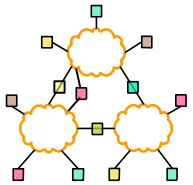
- Each routing update carries the entire path
- Loops are detected as follows:
  - When AS gets route check if AS already is in path
    - If yes, reject route
    - If no, add self and (possibly) advertise route further
- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops

Internet Routing Protocol: BGP



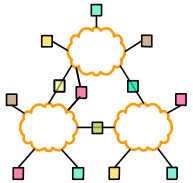


Interconnecting BGP Peers



- BGP uses TCP to connect peers
- AS's exchange reachability information through their BGP routers, only when routes change
- Advantages:
  - Simplifies BGP
  - No need for periodic refresh - routes are valid until withdrawn, or the connection is lost
  - Incremental updates
- Disadvantages
  - Congestion control on a routing protocol?
  - Poor interaction during high load

BGP Operations (Simplified)

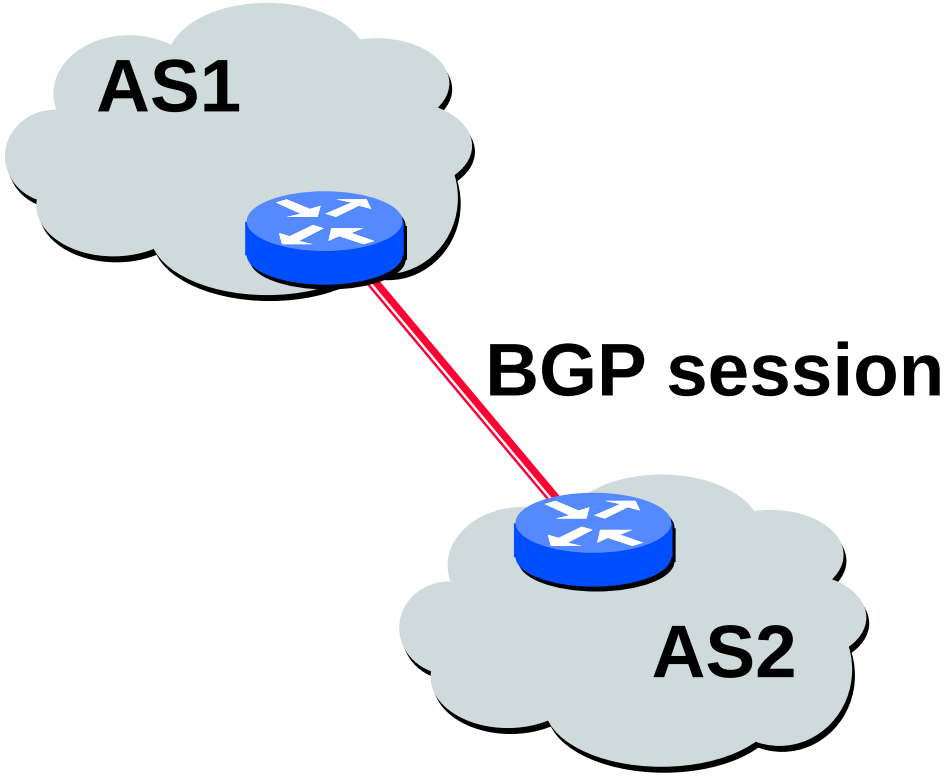


Establish session on TCP port 179

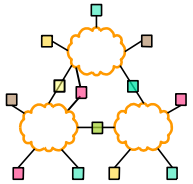
Exchange all active routes

Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages

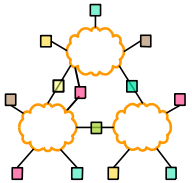


BGP Routing policies



- In theory BGP allows each domain to define its own routing policy...
- In practice there are two common policies
  - customer-provider routing
    - Customer c buys/pays Internet connectivity from provider P
  - shared-cost peering
    - Domains x and y agree to exchange packets by using a direct link or through an interconnection point.

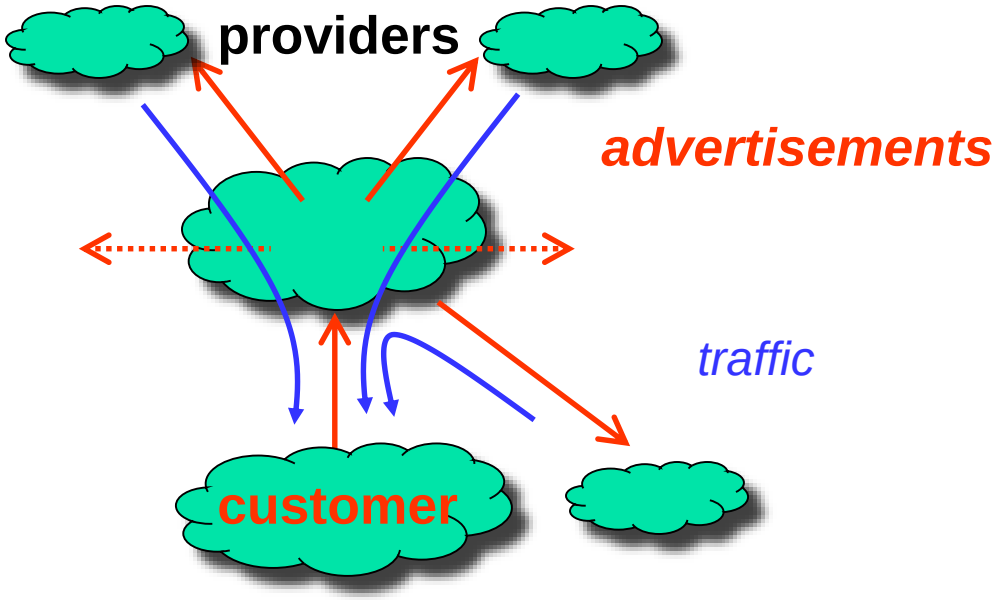
Implementing Transit



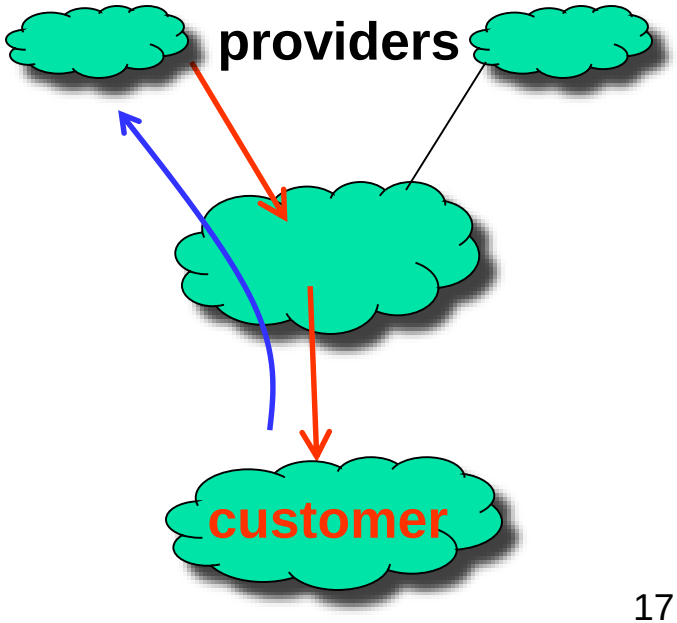
Filtering

- Routes from customer: to *everyone*
- Routes from provider: only to *customers*

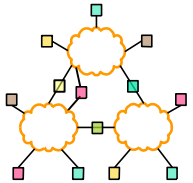
From other destinations
To the customer



From the customer
To other destinations

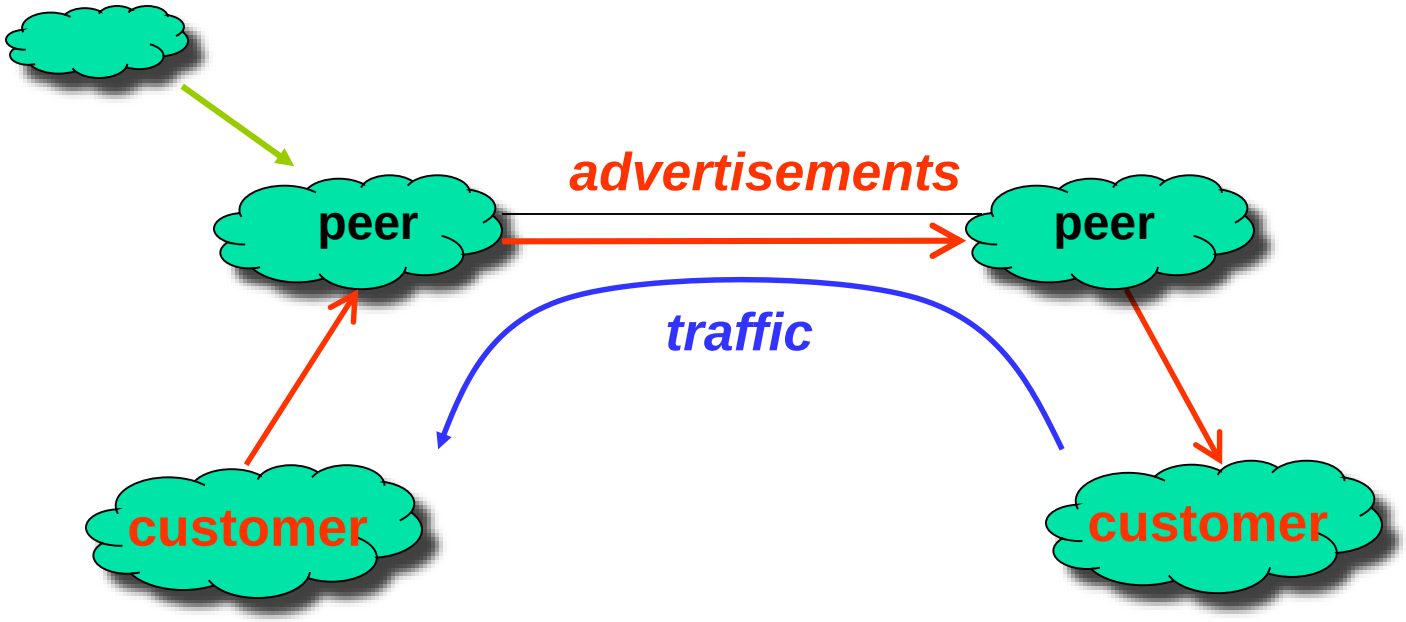


Implementing Peering

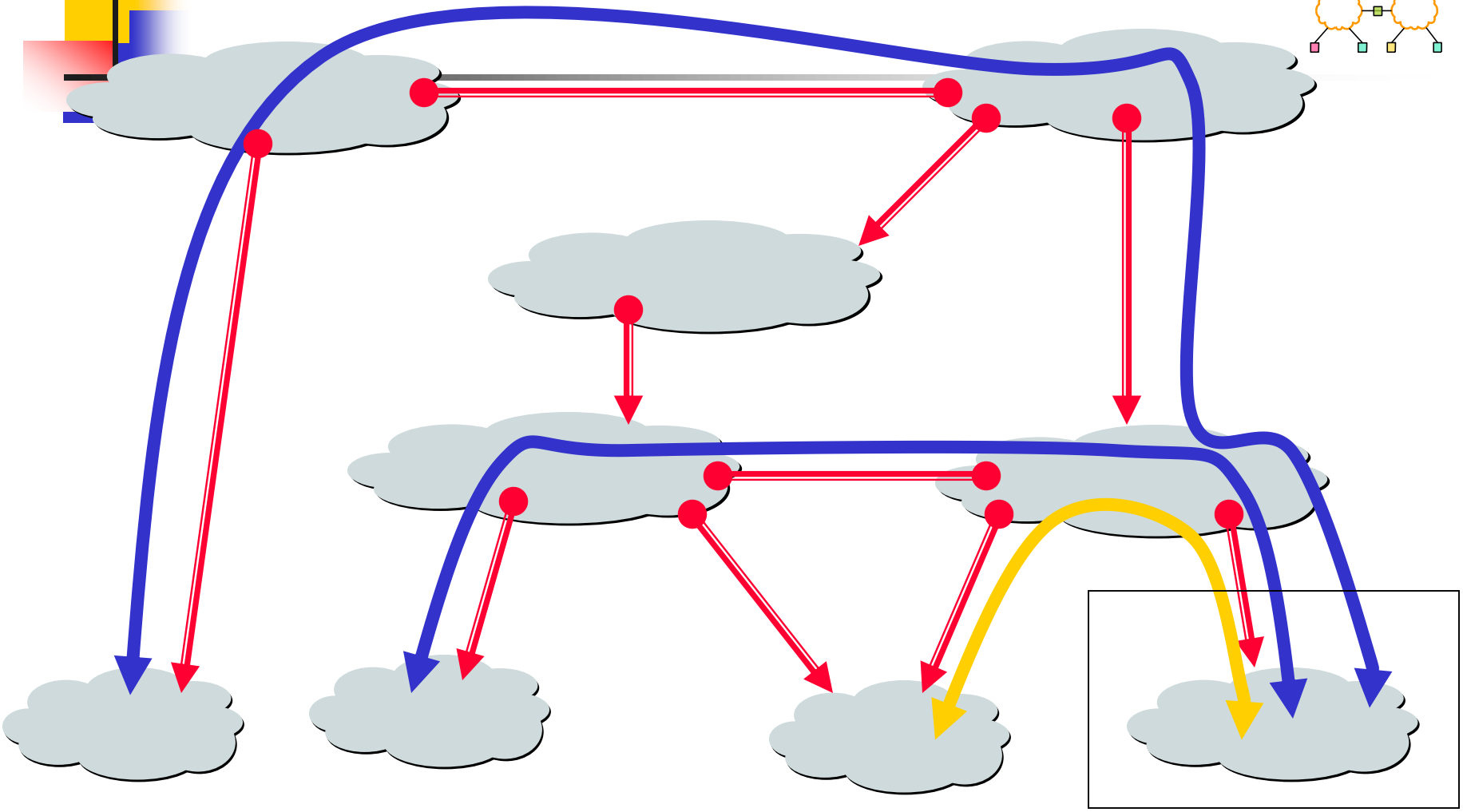


Filtering

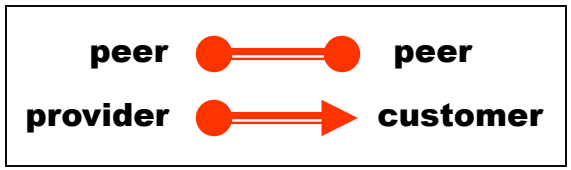
- Routes from peer: only to customers
- No routes from other peers or providers



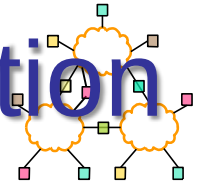
Peering Provides Shortcuts



Peering also allows connectivity between the customers of "Tier 1" providers.

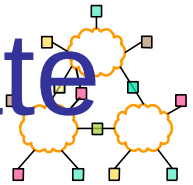


Physical Facilities for Interconnection



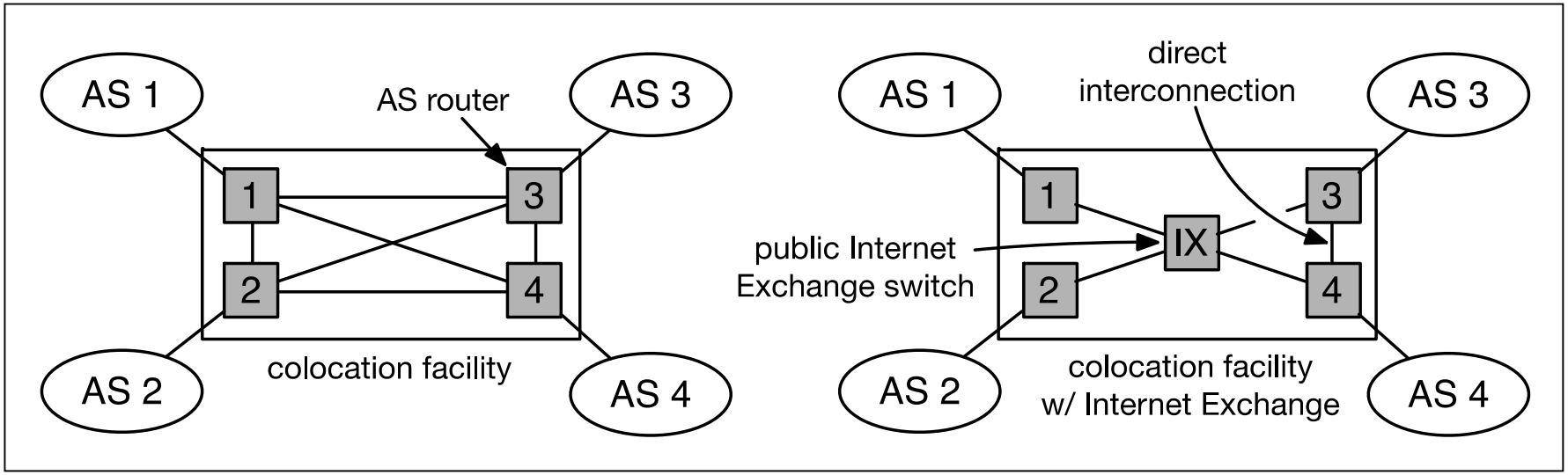
- **For networks to interconnect, they have to *physically* connect their networking equipment with each other.**
  - This requires the networks to meet in a common location, in facilities capable of supporting the equipment required for interconnection.
  - These colocation facilities lease their customers secure space to locate and operate equipment
- **Point of Presence (PoP)**
  - An access point to a communication provider's network.

Interconnection: Public & Private



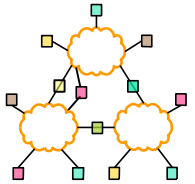
- **Interconnecting two networks requires both:**
  - (1) physical connectivity, *and*
  - (2) network connectivity.
- **Common options for interconnection are either:**
  - *Direct interconnection:*
    - Private bilateral arrangement between two networks using a dedicated physical connection
  - *Public connection:*
    - A multilateral arrangement where all networks connect into a public Internet Exchange switch.

Public and Private Interconnection



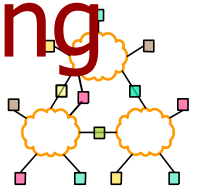
- **At left**: Simple colocation facility with direct interconnects
- **At right**: colocation facility that also offers IX through a public switch (or “switching fabric”)

Policy with BGP



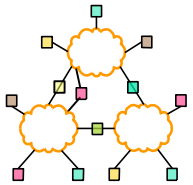
- BGP provides capability for enforcing policies.
- Policies are **not** part of BGP: they are configuration information
- BGP enforces choosing paths from alternatives and controlling advertisement to other AS's
- Example:
  - An AS refuses to act as transit, Limit path advertise
  - An AS can become transit for some AS's
    - Only advertise paths to some AS's
  - An AS can favor or disfavor certain AS's for traffic transit from itself

Implementation (RIB) Routing Information Bases



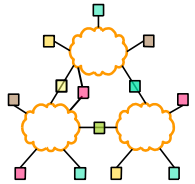
- Routes are stored in RIBs
  - Adj-RIBs-In: routing info that has been learned from other routers (unprocessed routing info)
  - Loc-RIB: local routing information selected from Adj-RIBs-In (routes selected locally)
  - Adj-RIBs-Out: info to be advertised to peers (routes to be advertised)
- routing policies enforced by defining two sets of filters for each peer
  - Import filter
  - Export filter

Implementing (2)



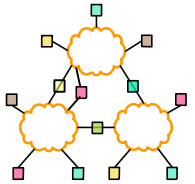
- Enforce transit relationships
  - Outbound route filtering
- Enforce order of route preference
  - provider < peer < customer
- Filters are defined in RPSL, Routing Policy Specification Language
  - Export: to Customer announce ANY
  - Export: to Peer announce Customer1 Customer2

Import and Export policies



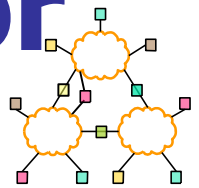
- **Import policy**
  - What to do with routes learned from neighbors?
  - Selecting best path
- **Export policy**
  - What routes to announce to neighbors?
  - Depends on relationship with neighbor

Export Policy

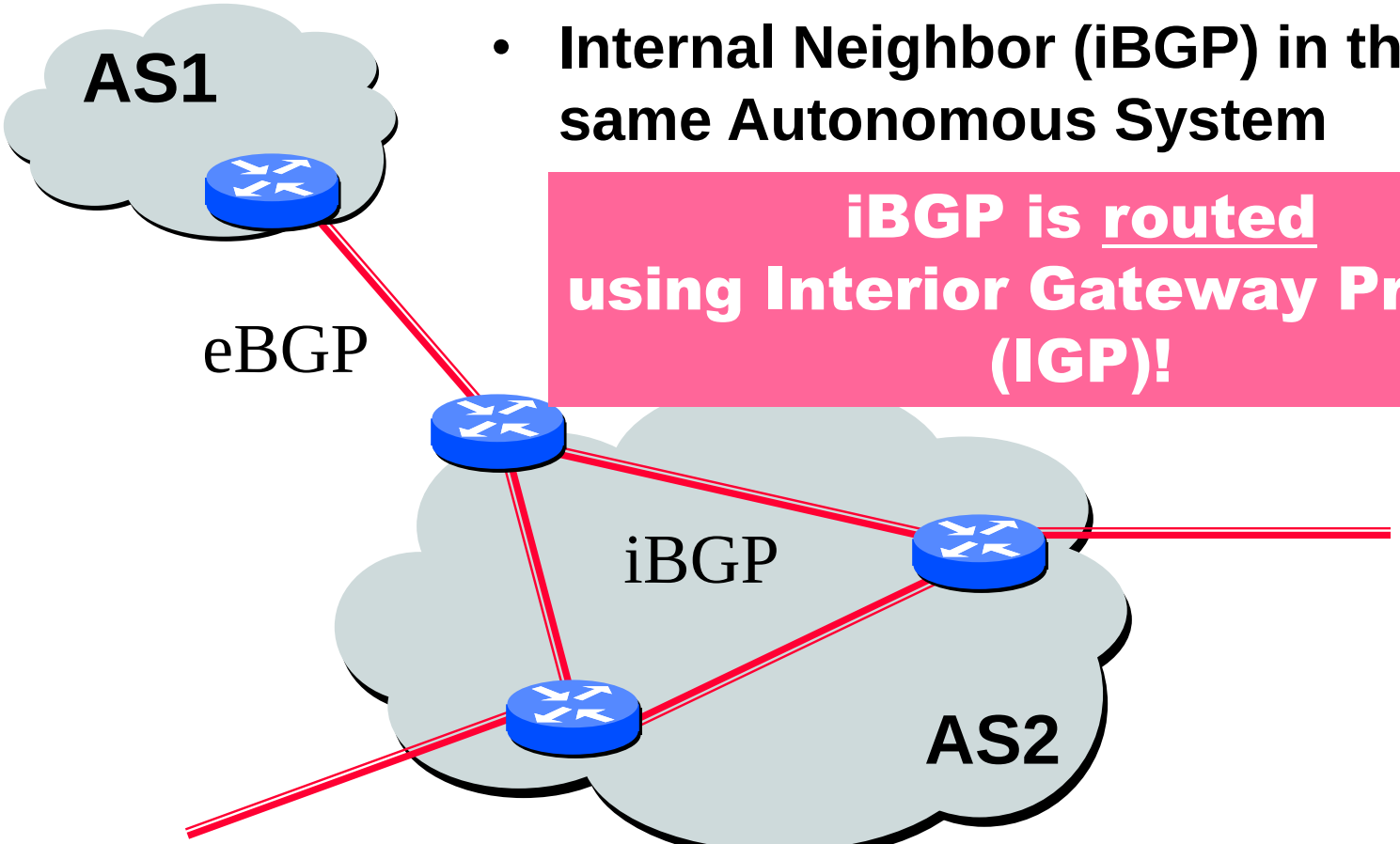


- An AS exports only best paths to its neighbors
  - Guarantees that once the route is announced the AS is willing to transit traffic on that route
- To Customers
  - Announce all routes learned from peers, providers and customers, and self-origin routes
- To Providers
  - Announce routes learned from customers and self-origin routes
- To Peers
  - Announce routes learned from customers and self-origin routes

Two Types of BGP Neighbor Relationships

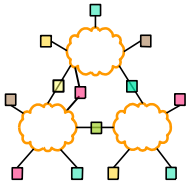


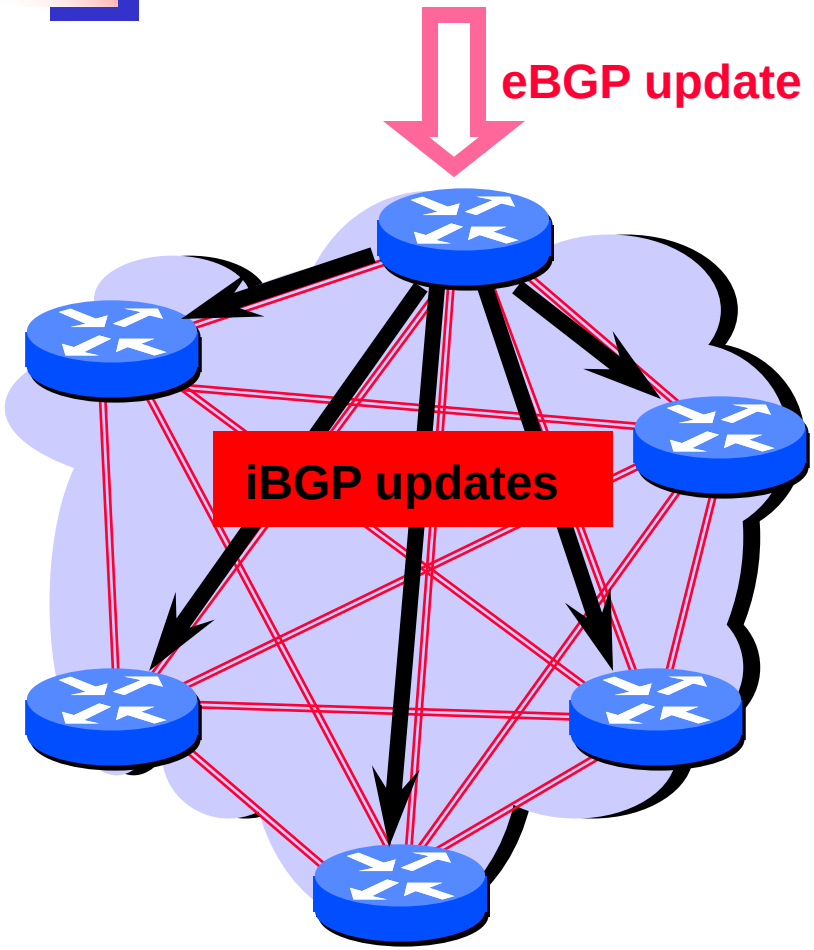
- External Neighbor (eBGP) in a different Autonomous Systems
- Internal Neighbor (iBGP) in the same Autonomous System



iBGP is routed using Interior Gateway Protocol (IGP)!

iBGP Peers Must be Fully Meshed

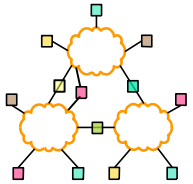




- **iBGP is needed to avoid routing loops within an AS**
- **Injecting external routes into IGP does not scale and causes BGP policy information to be lost**
- **BGP does not provide "shortest path" routing**

iBGP neighbors do not announce routes received via iBGP to other iBGP neighbors.

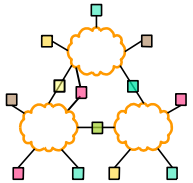
BGP Messages

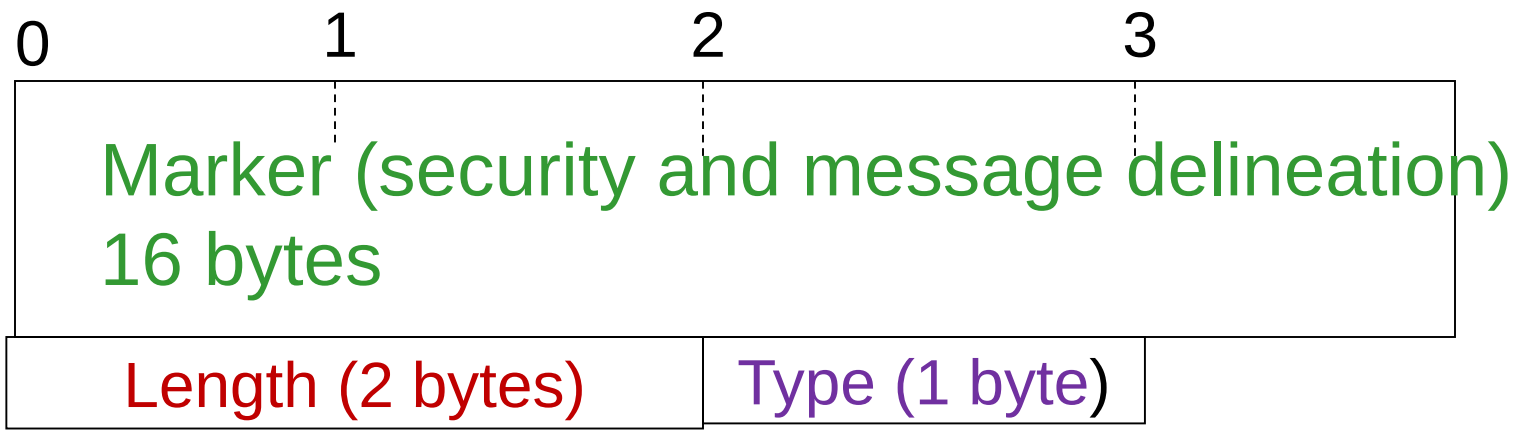


- **Open** : Establish a peering session.
- **Keep Alive** : Handshake at regular intervals.
- **Notification** : Shuts down a peering session.
- **Update** : Announcing new routes or withdrawing previously announced routes.

announcement
=
prefix + attributes values

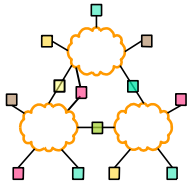
BGP Common Header

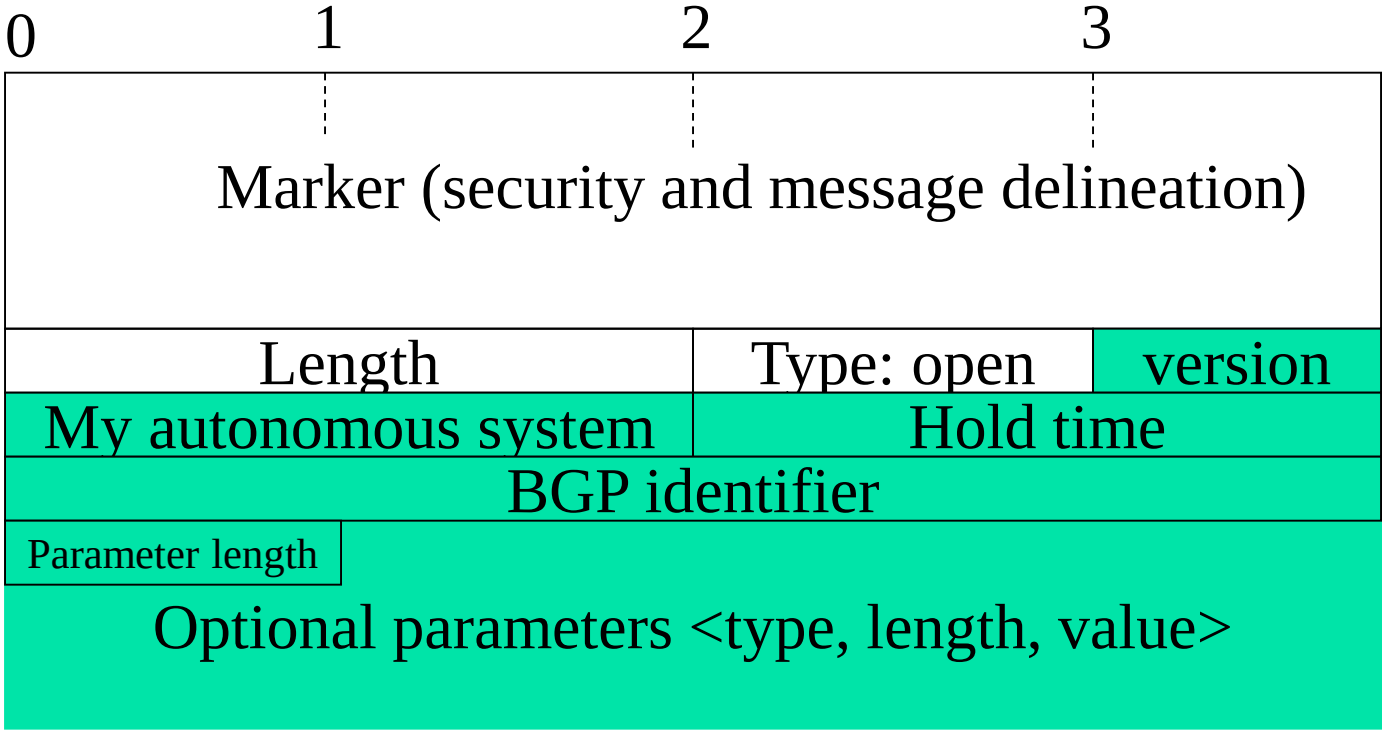




Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

BGP OPEN message





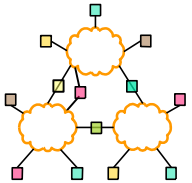
My AS: id assigned to that AS

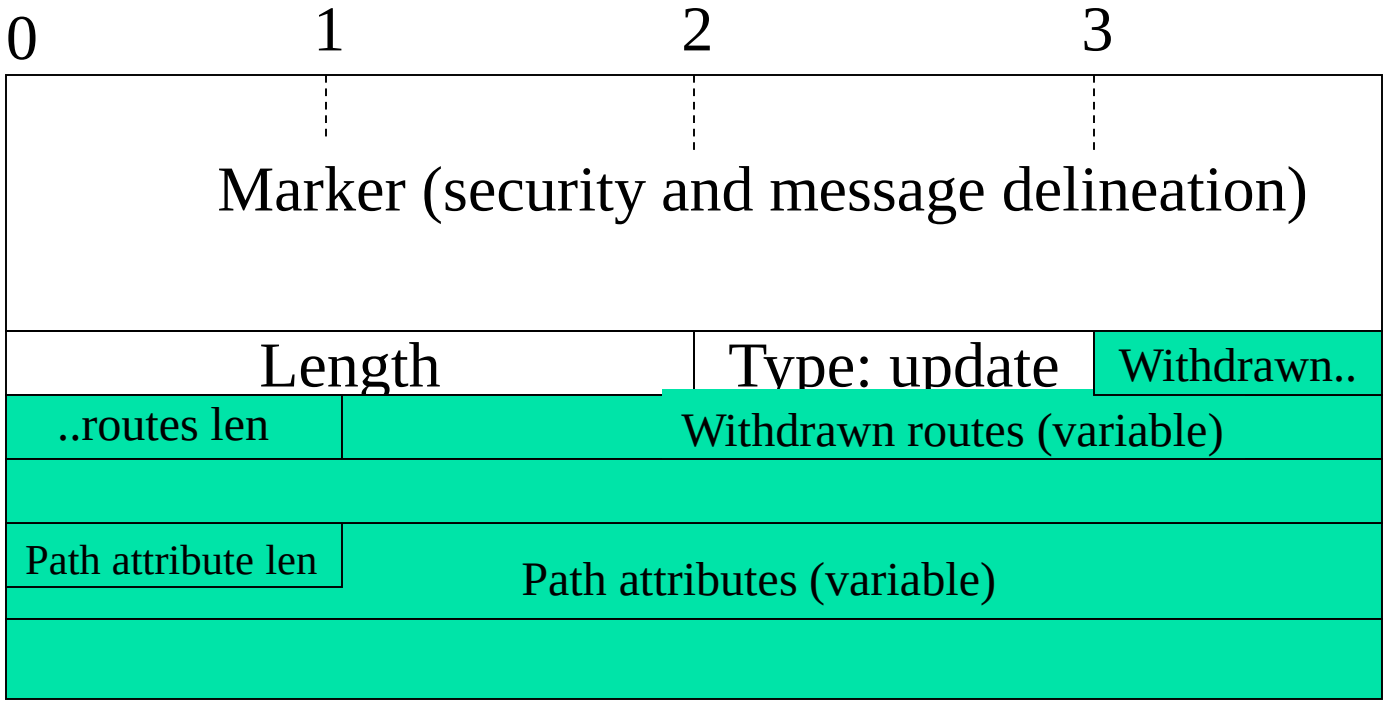
Hold timer: max interval between KEEPALIVE or UPDATE messages

interval implies no keep_alive.

BGP ID: IP address of one interface (same for all messages)

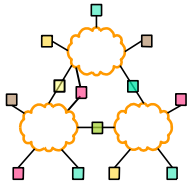
BGP UPDATE message





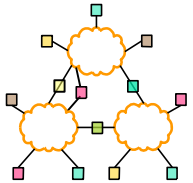
- Many prefixes may be included in UPDATE, but must share same attributes.
- UPDATE message may report multiple withdrawn routes.

BGP UPDATE Message

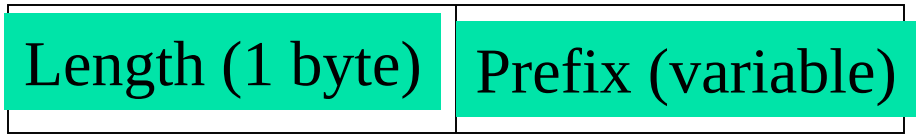


- List of withdrawn routes
- Network layer reachability information
  - List of reachable prefixes
- Path attributes
  - Origin
  - Path
  - Metrics
- All prefixes advertised in a message have same path attributes

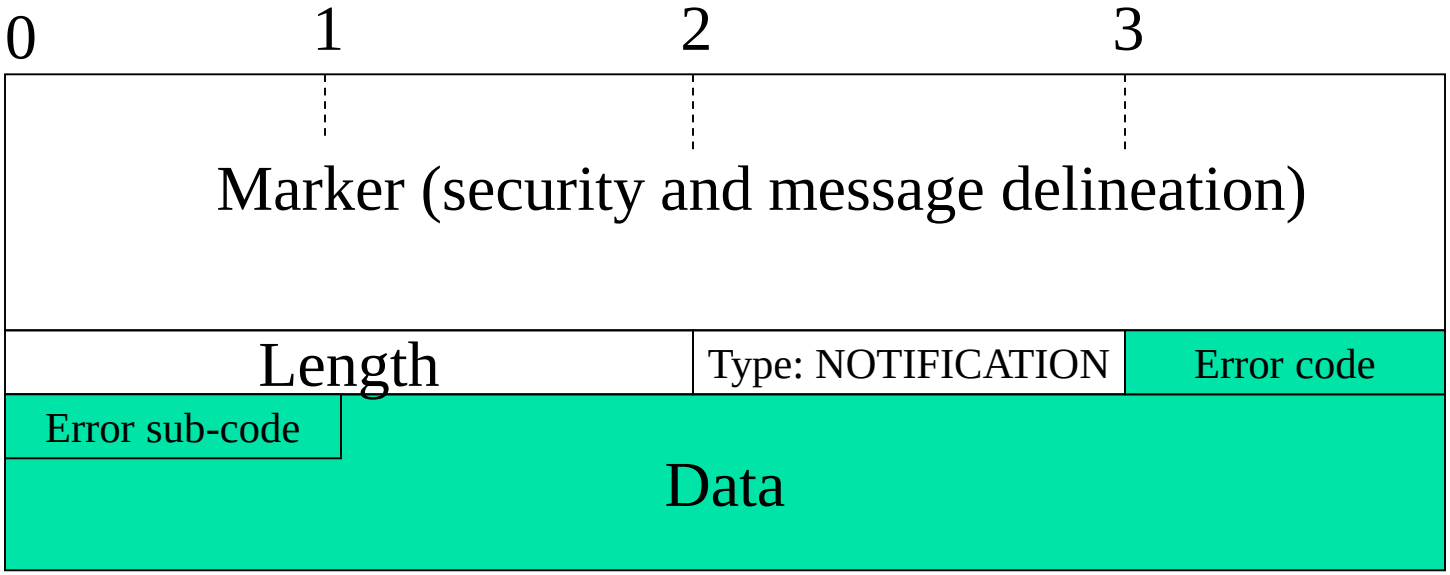
NLRI



- Network Level Reachability Information
  - list of IP address prefixes encoded as follows:

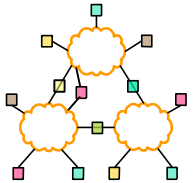


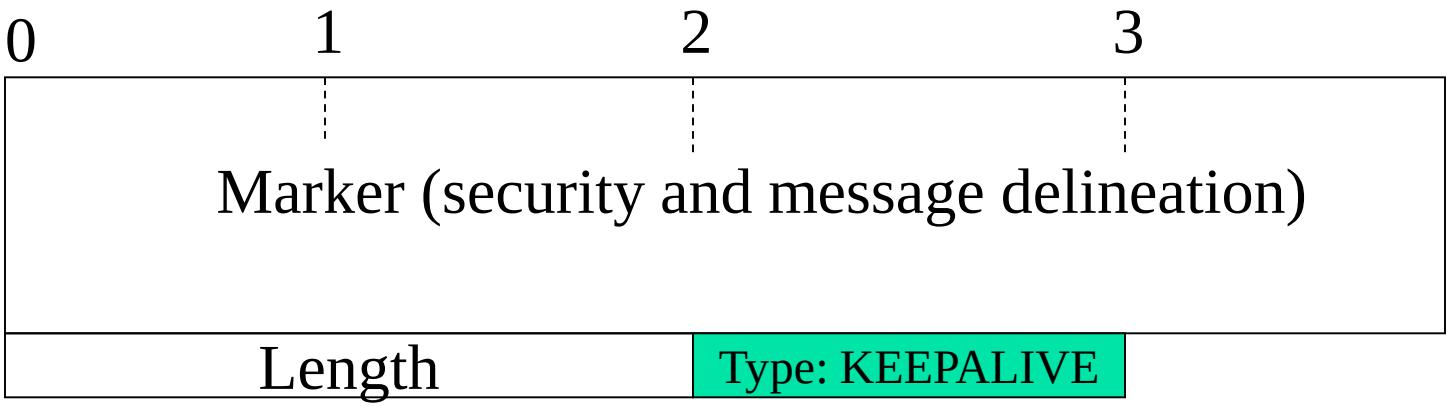
BGP NOTIFICATION message



- Used for error notification
  TCP connection is closed *immediately* after notification

BGP KEEPALIVE message



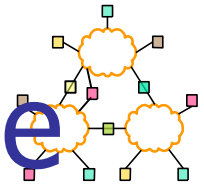


Sent periodically to peers to ensure connectivity.

If hold_time is zero, messages are not sent..

Sent in place of an UPDATE message

Example BGP Routing Table



The full routing table

```
> show ip bgp

 Network          Next Hop        Metric LocPrf Weight Path
*>i3.0.0.0        4.79.2.1             0    110      0 3356 701 703 80 i
*>i4.0.0.0        4.79.2.1             0    110      0 3356 i
*>i4.21.254.0/23  208.30.223.5        49    110      0 1239 1299 10355 10355 i
* i4.23.84.0/22   208.30.223.5       112    110      0 1239 6461 20171 i
```

Specific entry. Can do longest prefix lookup:

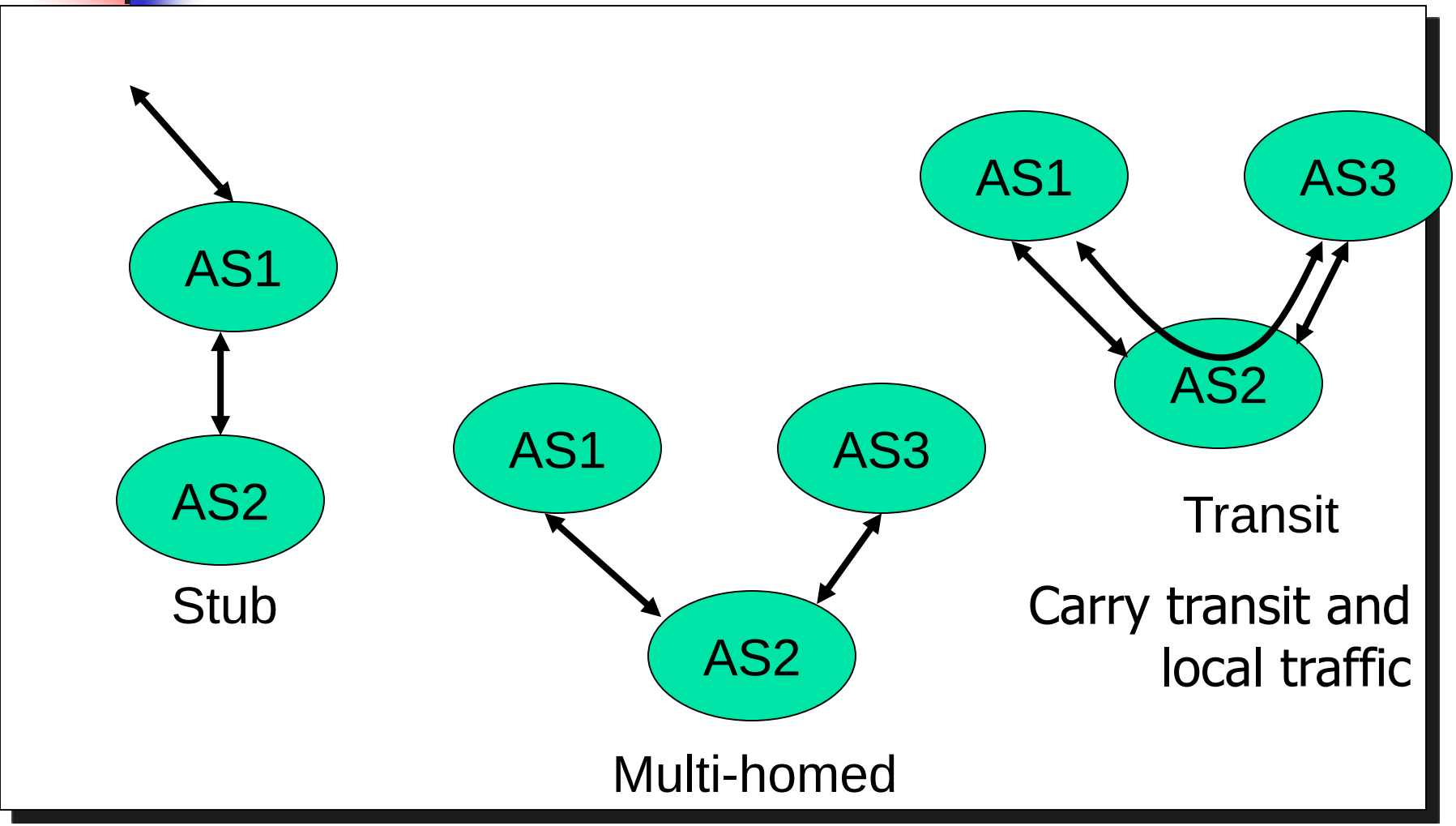
```
> show ip bgp 130.207.7.237
BGP routing table entry for 130.207.0.0/16
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  10578 11537 10490 2637
    192.5.89.89 from 18.168.0.27 (66.250.252.45)
      Origin IGP, metric 0, localpref 150, valid, internal, best
      Community: 10578:700 11537:950
      Last update: Sat Jan 14 04:45:09 2006
```

Prefix

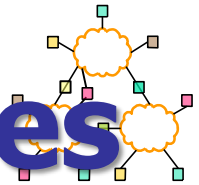
AS path

Next-hop

AS Categories

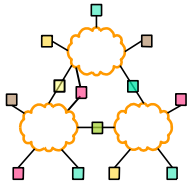


Important BGP attributes

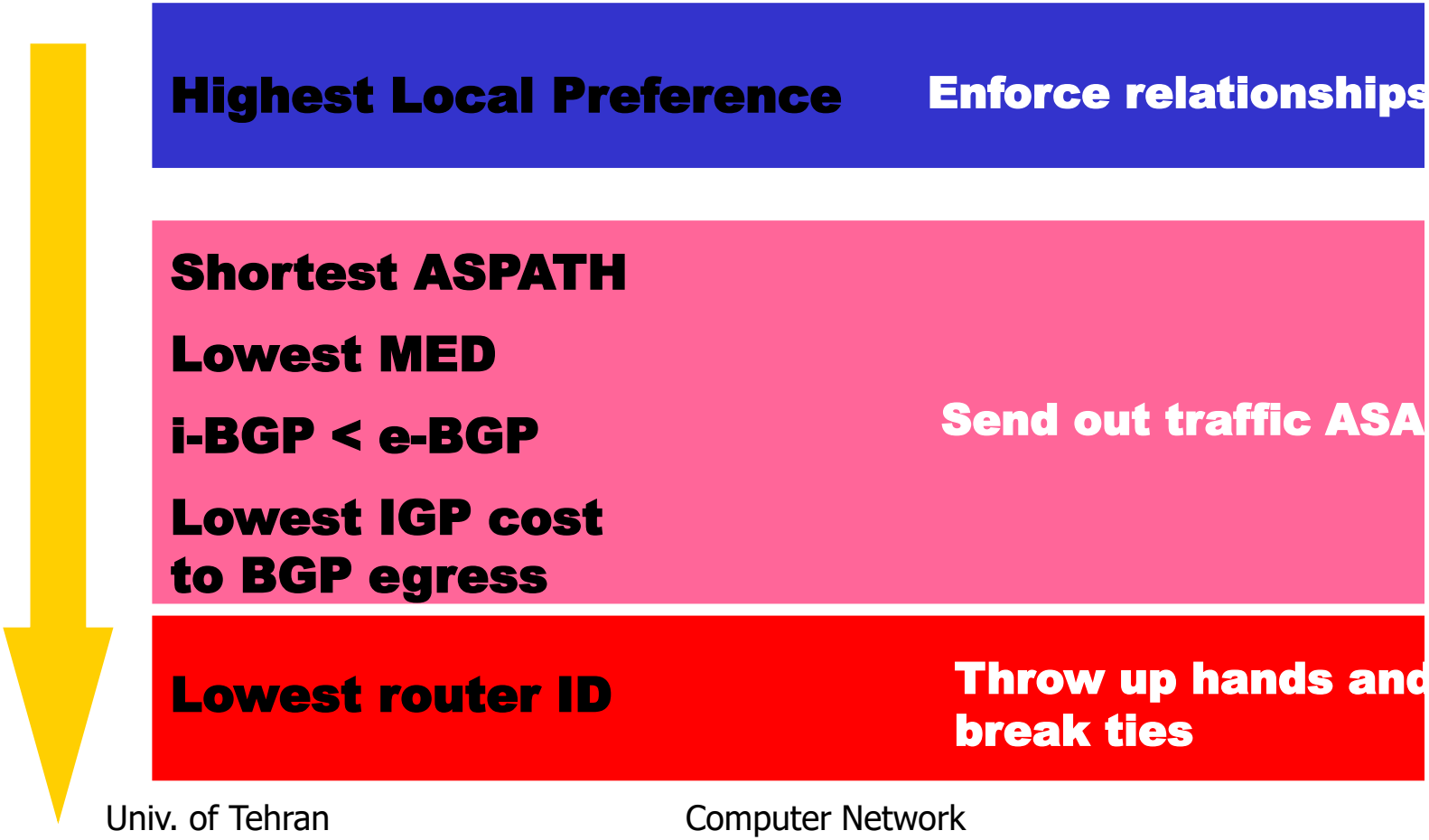


- LocalPREF
  - Local preference policy to choose "most" preferred route. Numerical value assigned by routing policy. Higher values are more preferred
- Multi-exit Discriminator (MED)
  - Which peering point to choose? One AS to specify that one exit point is more preferred than another. Lower values are more preferred.
- Import Rules
  - What route advertisements do I accept?
- Export Rules
  - Which routes do I forward to whom?

Route Selection Summary



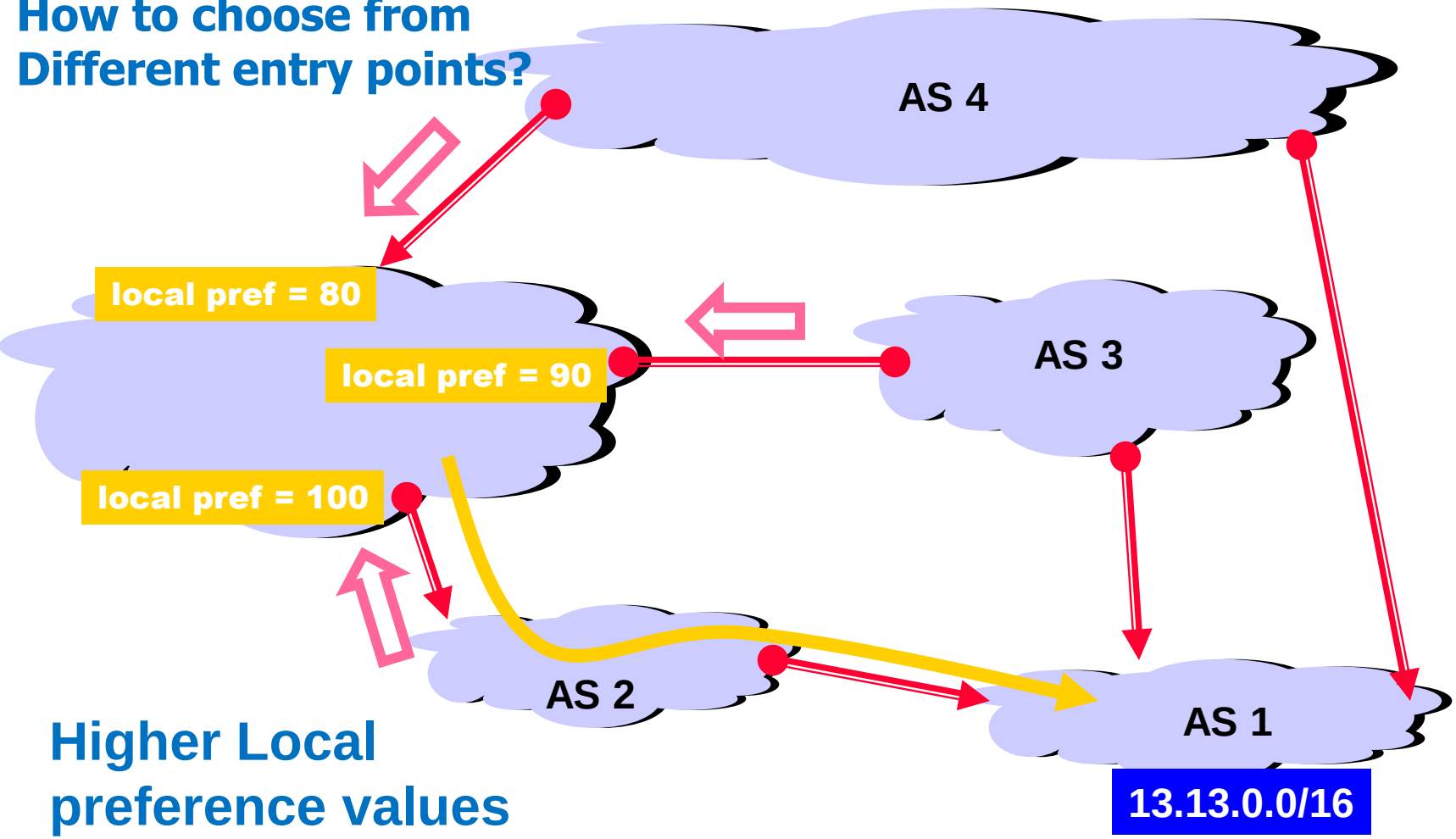
It is policy based and complex but in general



Local pref,

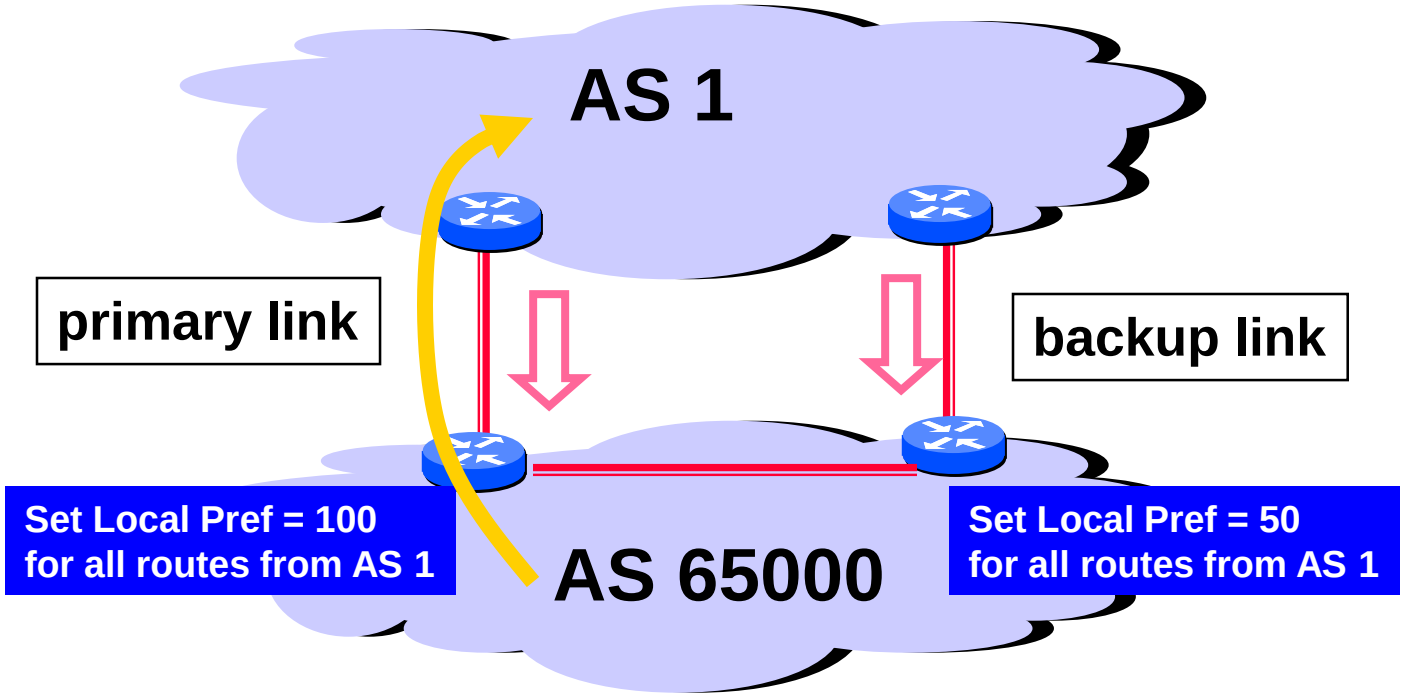
Local preference only used in iBGP

How to choose from Different entry points?



Higher Local preference values are more preferred

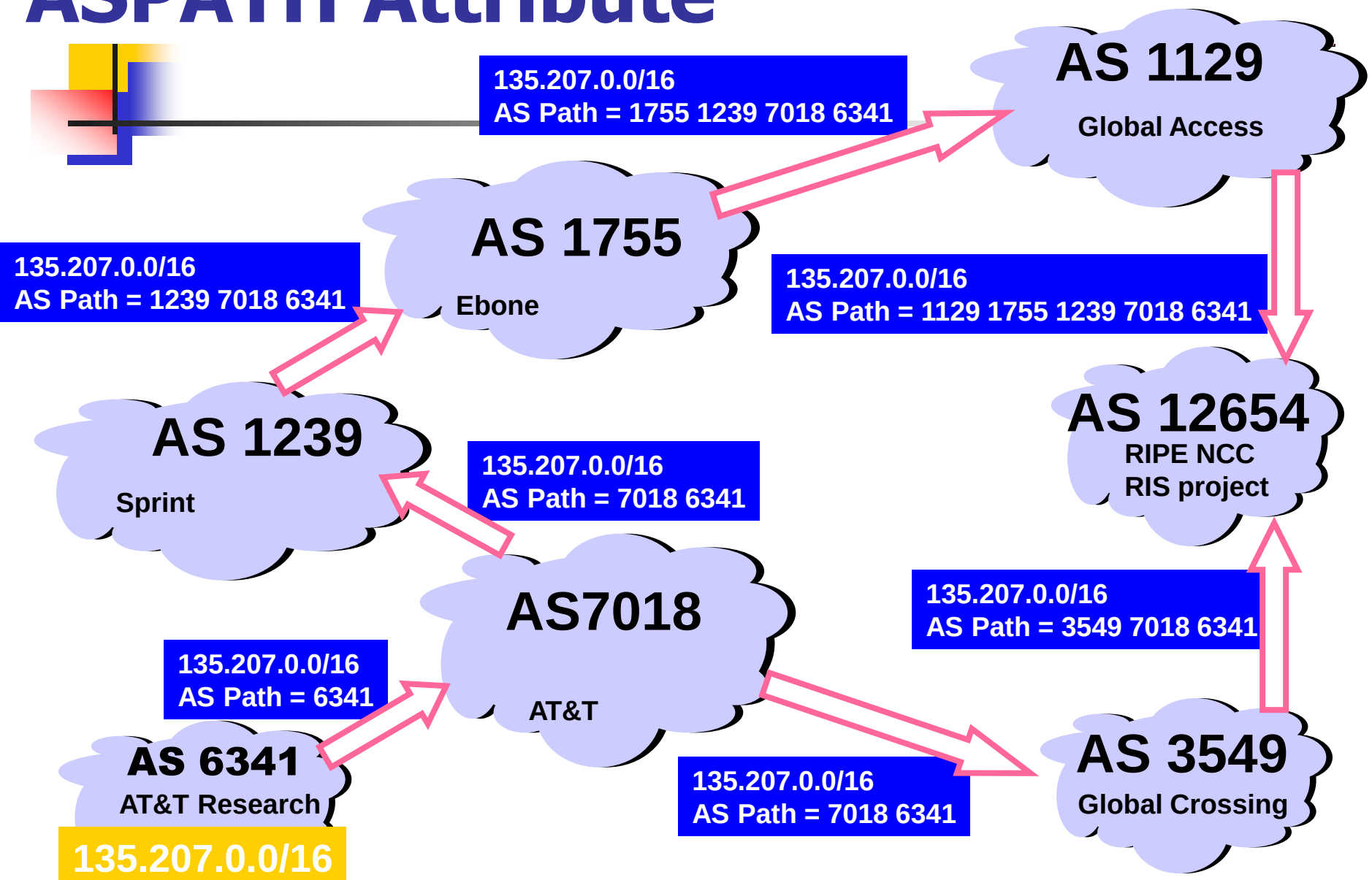
Backup Links with Local Preference (Outbound Traffic)



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

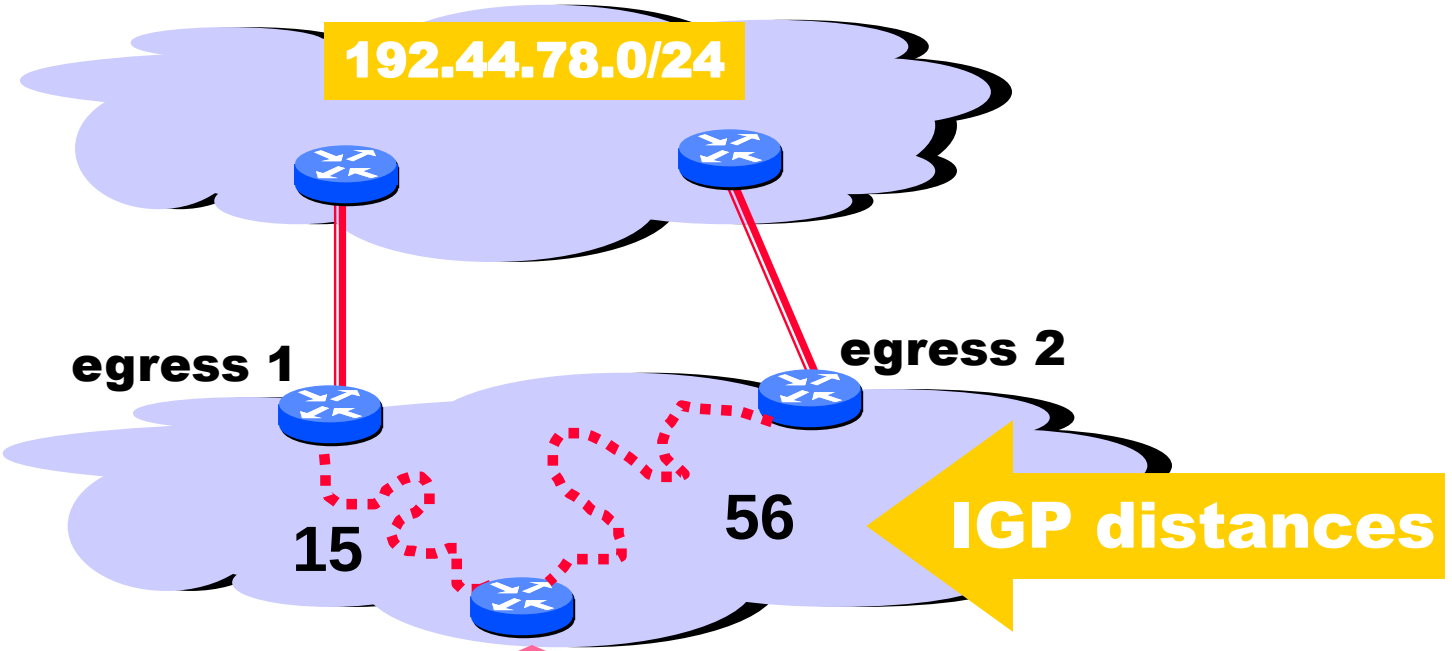
We'll talk about inbound traffic soon …

ASPATH Attribute



Shortest ASPATh wins

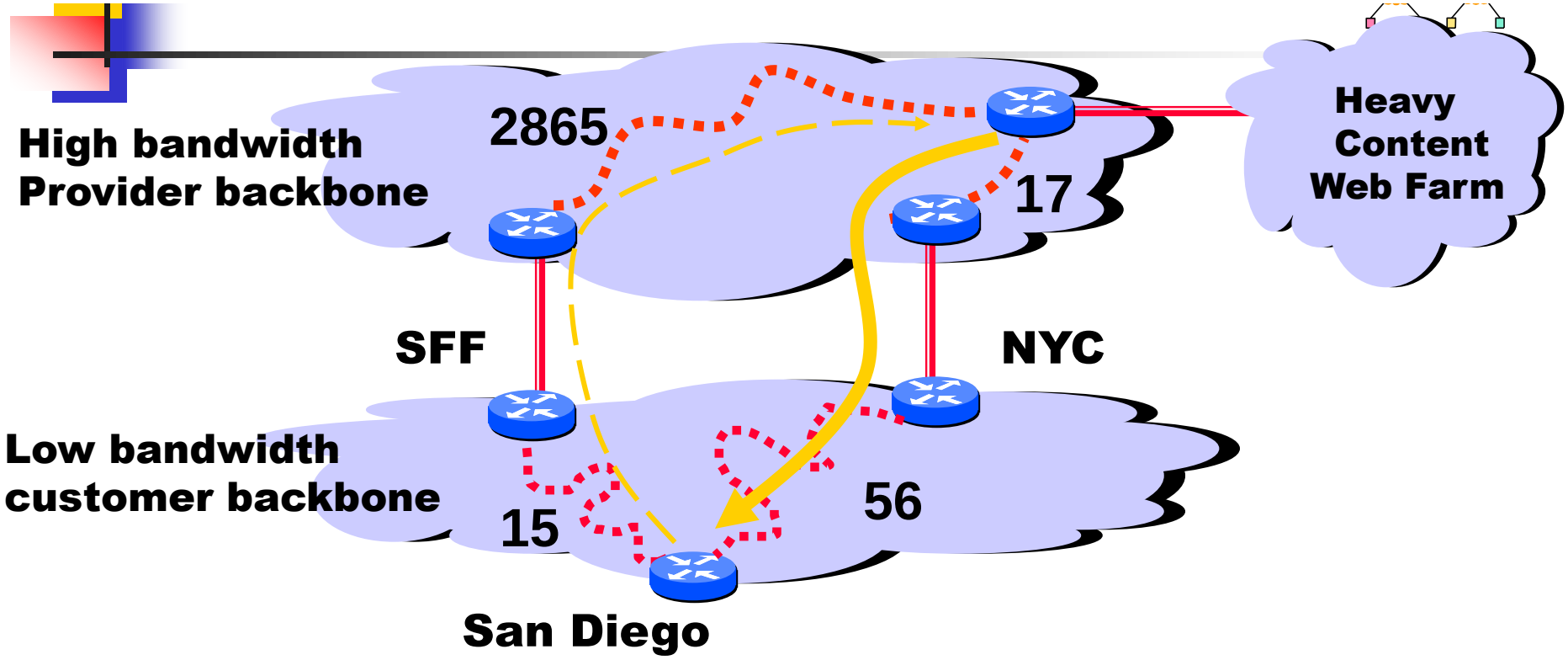
Hot Potato Routing



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!

Getting Burned by the Hot Potato

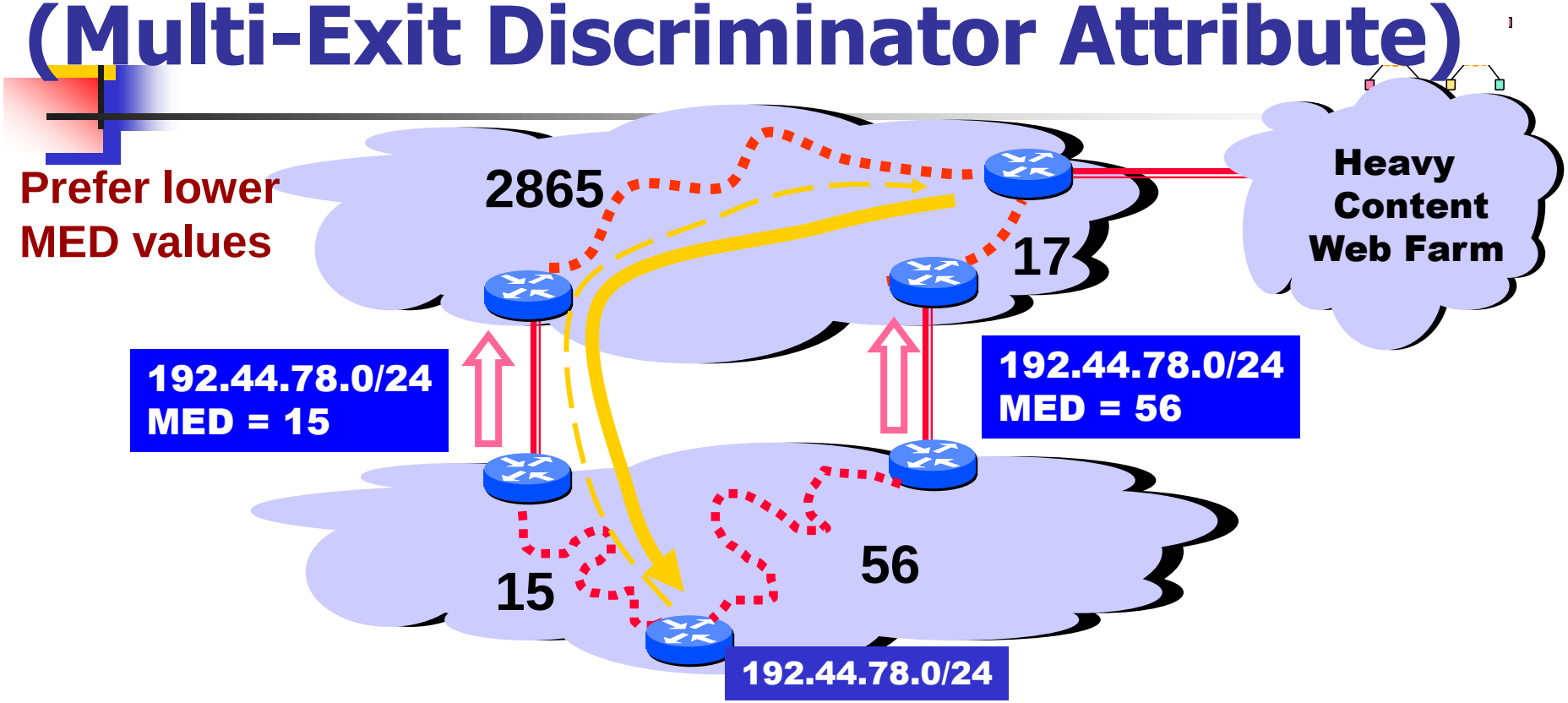


Many customers want their provider to carry the bits!

tiny http request

huge http reply

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

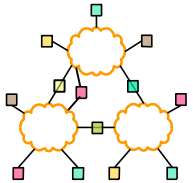


This means that MEDs must be considered BEFORE IGP distance!

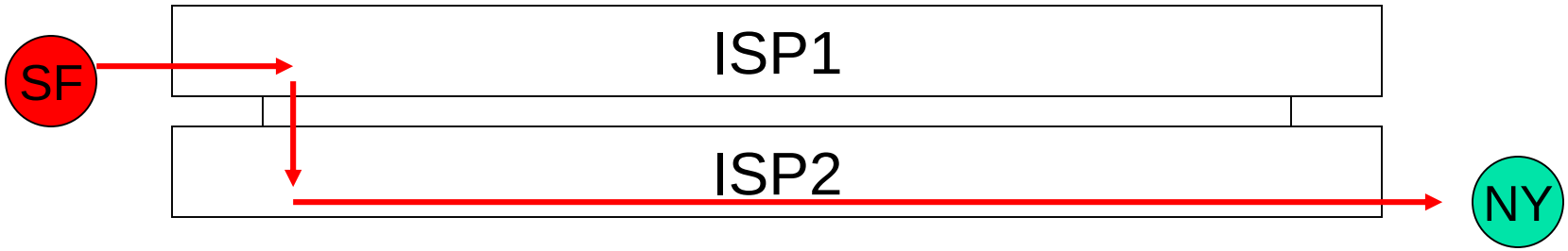
Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

MED

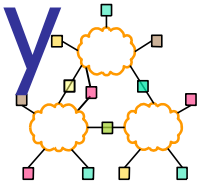


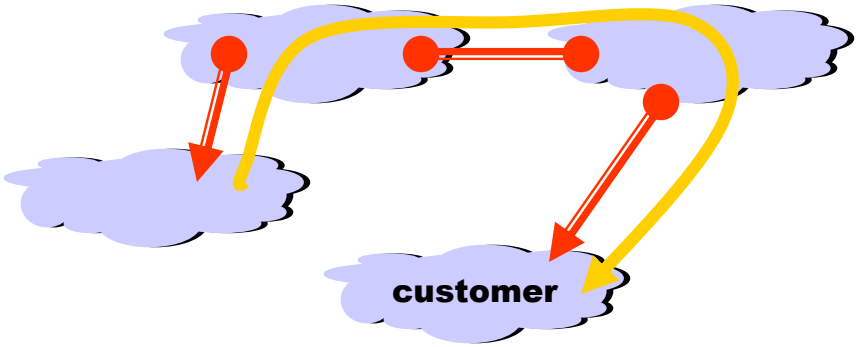
- MED is typically used in provider/subscriber scenarios
- It can lead to unfairness if used between ISP because it may force one ISP to carry more traffic:



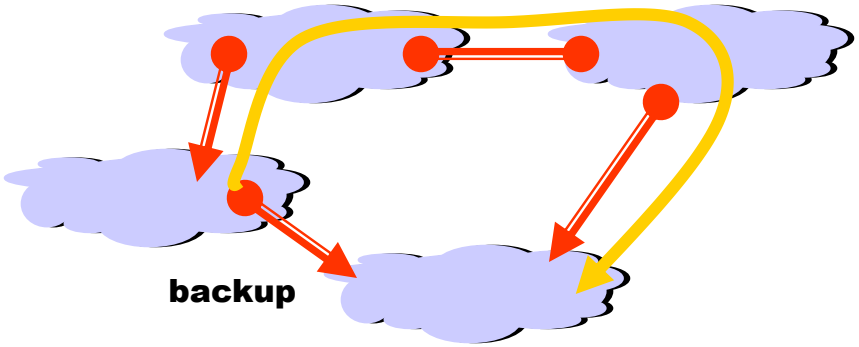
- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way

Policies Can Interact Strangely ("Route Pinning" Example)

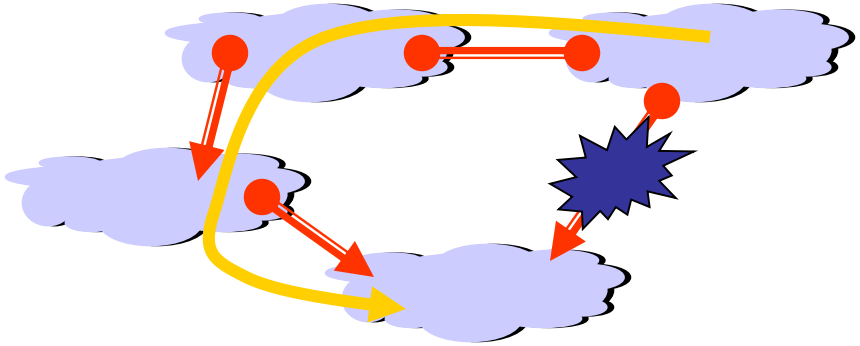




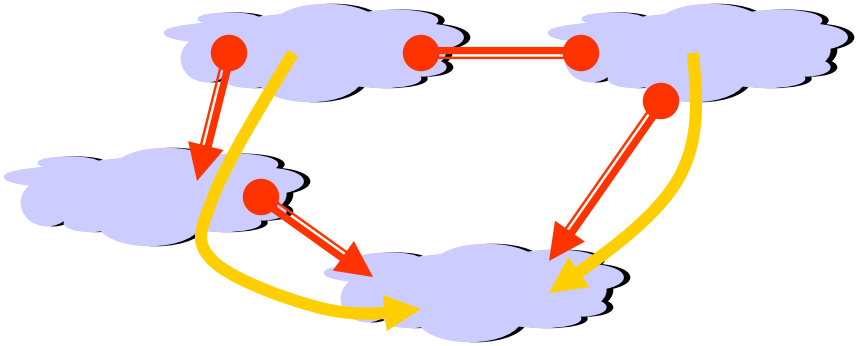
1



2 **Install backup link using community**

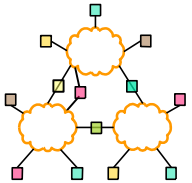


3 **Disaster strikes primary link and the backup takes over**



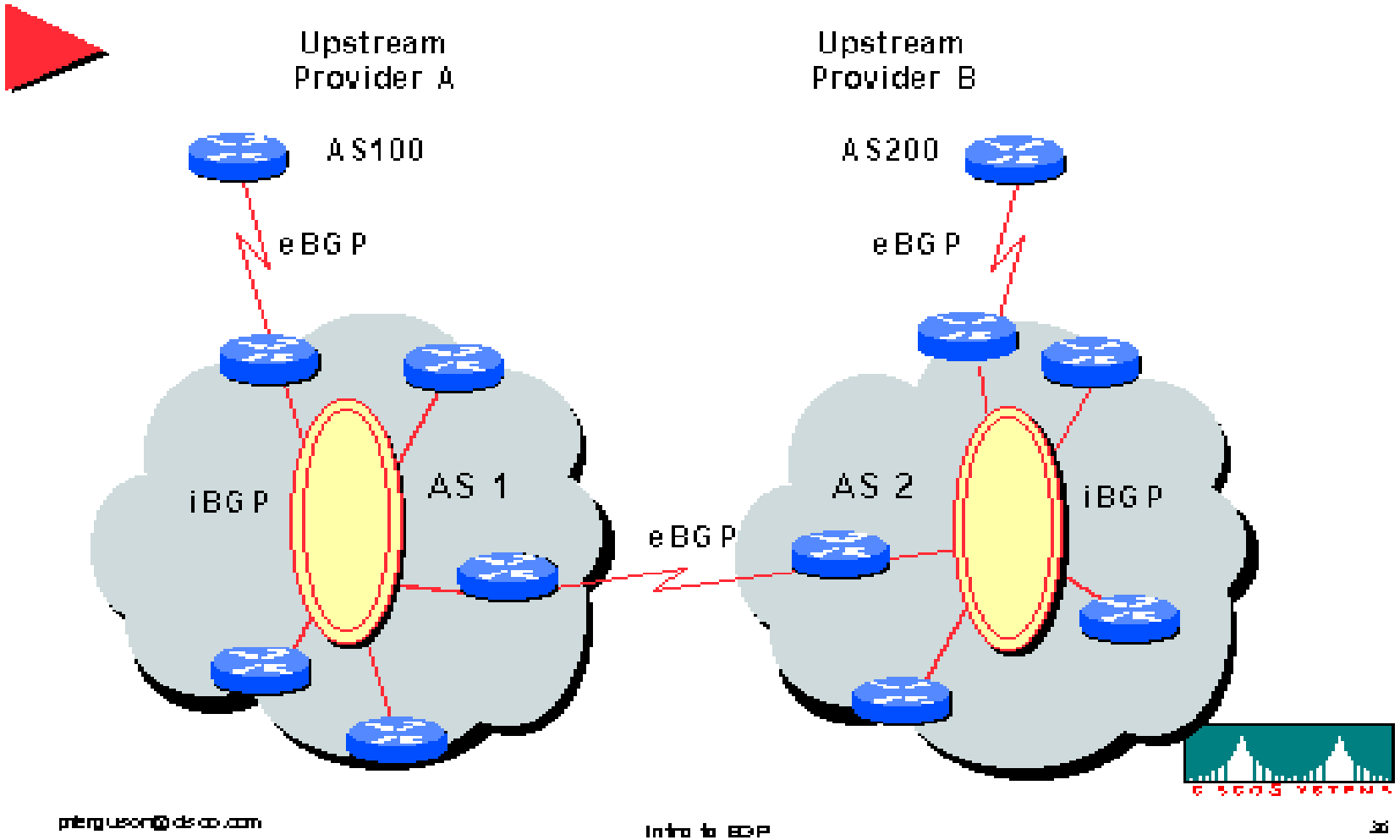
4 **Primary link is restored but some traffic remains *pinned* to backup**

Outline

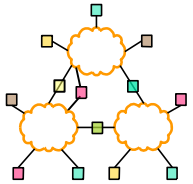


- External BGP (E-BGP)
- Internal BGP (I-BGP)
- Stability Issues

I-BGP

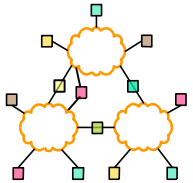

Upstream
Provider A
Upstream
Provider B
AS100
AS200
eBGP
eBGP
iBGP
AS 1
AS 2
eBGP
iBGP

Internal BGP (I-BGP)



- Same messages as E-BGP
- Different rules about re-advertising prefixes:
  - Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but
  - Prefix learned from one I-BGP neighbor cannot be advertised to another I-BGP neighbor
  - Reason: no AS PATH within the same AS and thus danger of looping.

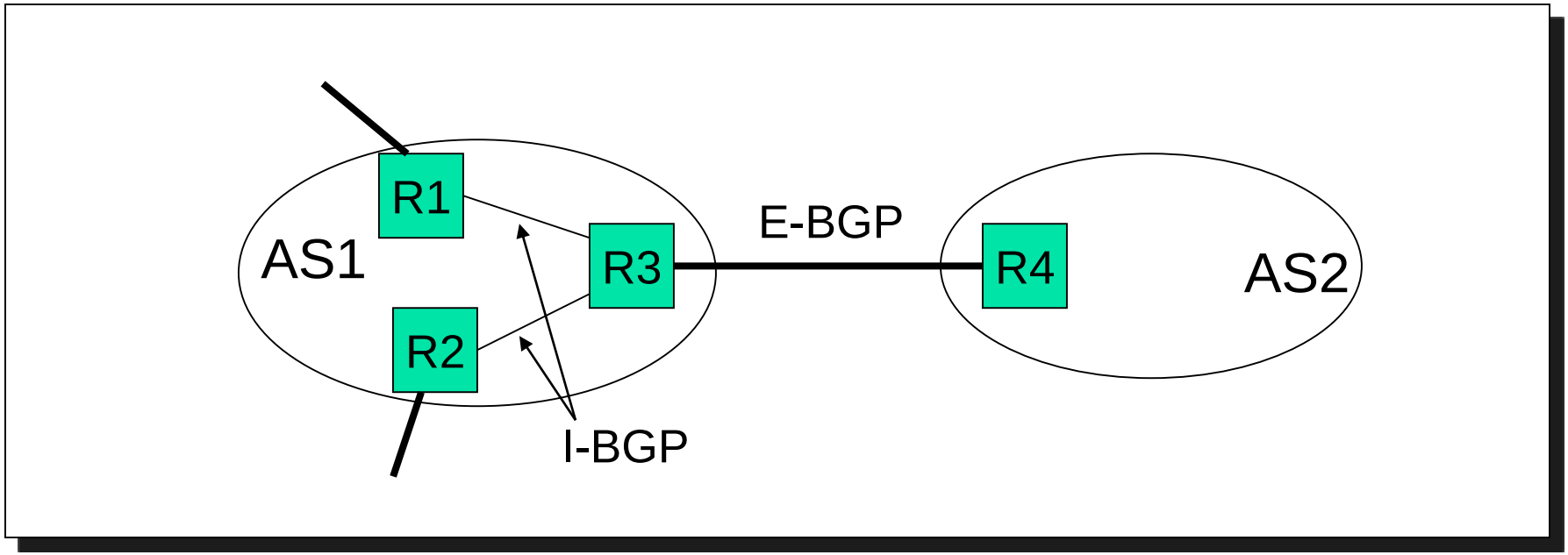
Internal BGP (I-BGP)



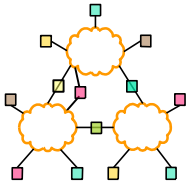
- R3 can tell R1 and R2 prefixes from R4
- R3 can tell R4 prefixes from R1 and R2
- R3 cannot tell R2 prefixes from R1

R2 can only find these prefixes through a *direct connection* to R1
Result: I-BGP routers must be fully connected (via TCP)!

- contrast with E-BGP sessions that map to physical links

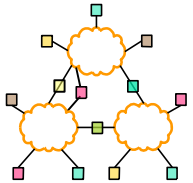


Link Failures

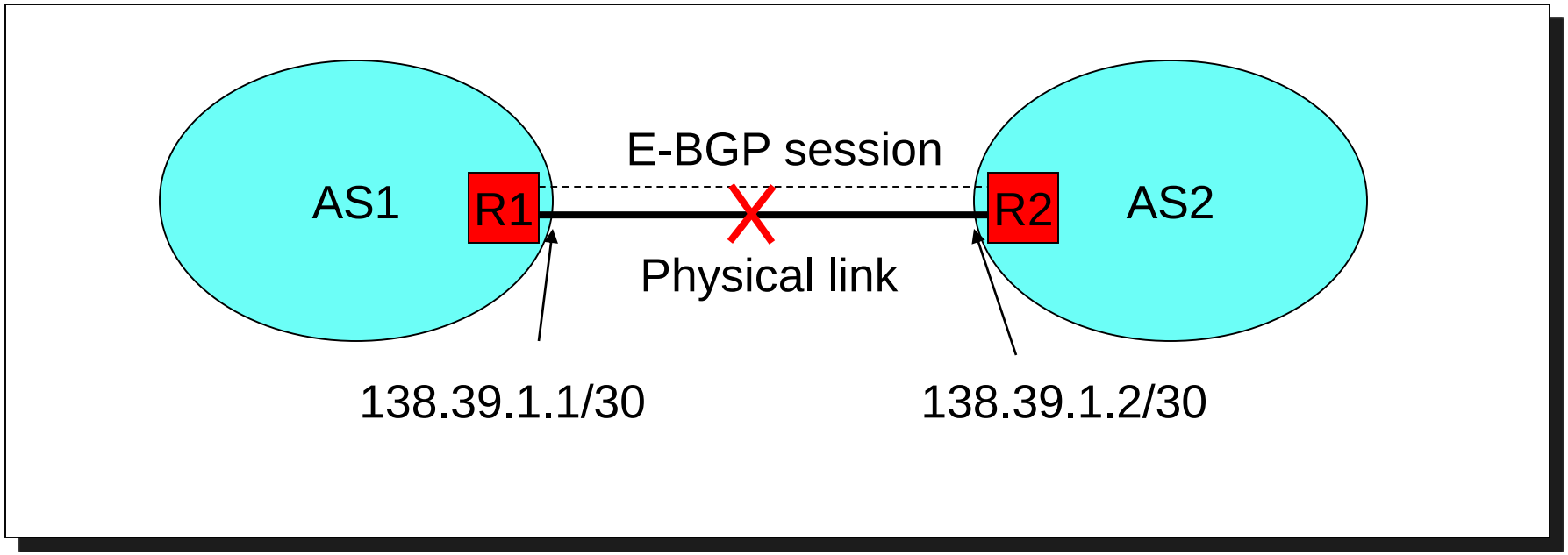


- Two types of link failures:
  - Failure on an E-BGP link
  - Failure on an I-BGP Link
- These failures are treated completely different in BGP
- Why?

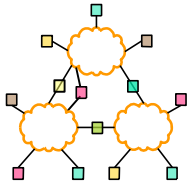
Failure on an E-BGP Link



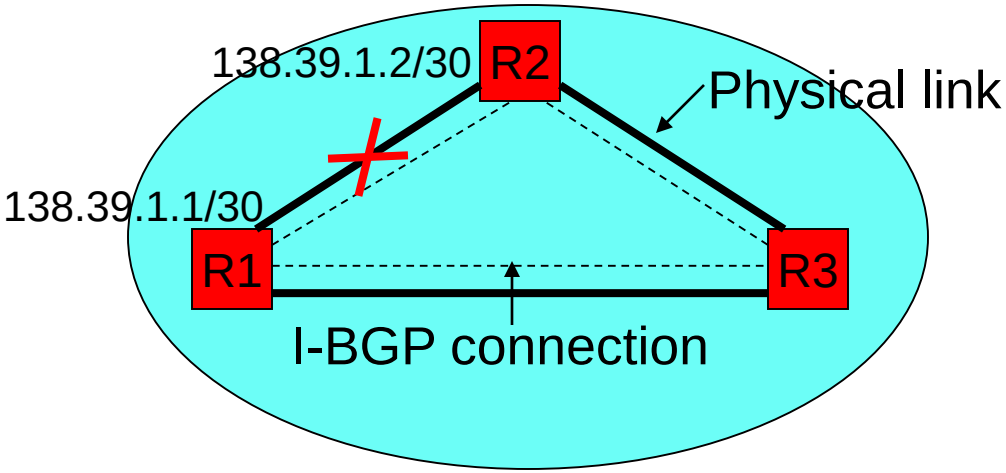
- If the link R1-R2 goes down
  - The TCP connection breaks
  - BGP routes are removed
- This is the *desired* behavior



Failure on an I-BGP Link

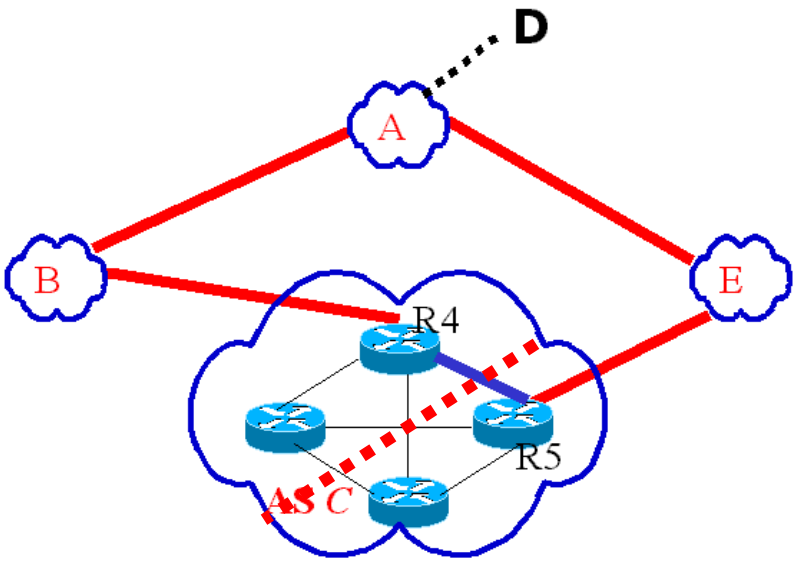


- If link R1-R2 goes down, R1 and R2 should still be able to exchange traffic
- The indirect path through R3 must be used
- Thus, E-BGP and I-BGP must use *different conventions* with respect to TCP endpoints



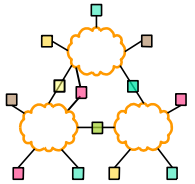
BGP Traffic Engineering

- We assumed an AS could be modeled as a node
  - with a single best path to the destination
- But a single AS may advertise more than one path.



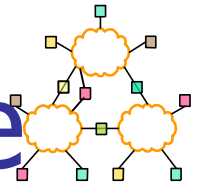
- BGP Traffic Engineering:
  - *R4* chooses path <*C B A*>
  - *R5* chooses path <*C E A*>
- Divide one AS into Logical ASes such that
  - All routers within a logical AS have the same best path→each logical AS can be modeled as a node.

Issues on Routing



- Challenges Facing Internet Routing
  - Internet Has Grown Dramatically
    - Large number of routing entries
    - High volumes of updates
    - Frequent topological changes
  - Fault-Model Has Changed Dramatically
    - More malfunctioning components
    - Intentional attacks
- Do we need a fundamentally new routing architecture?

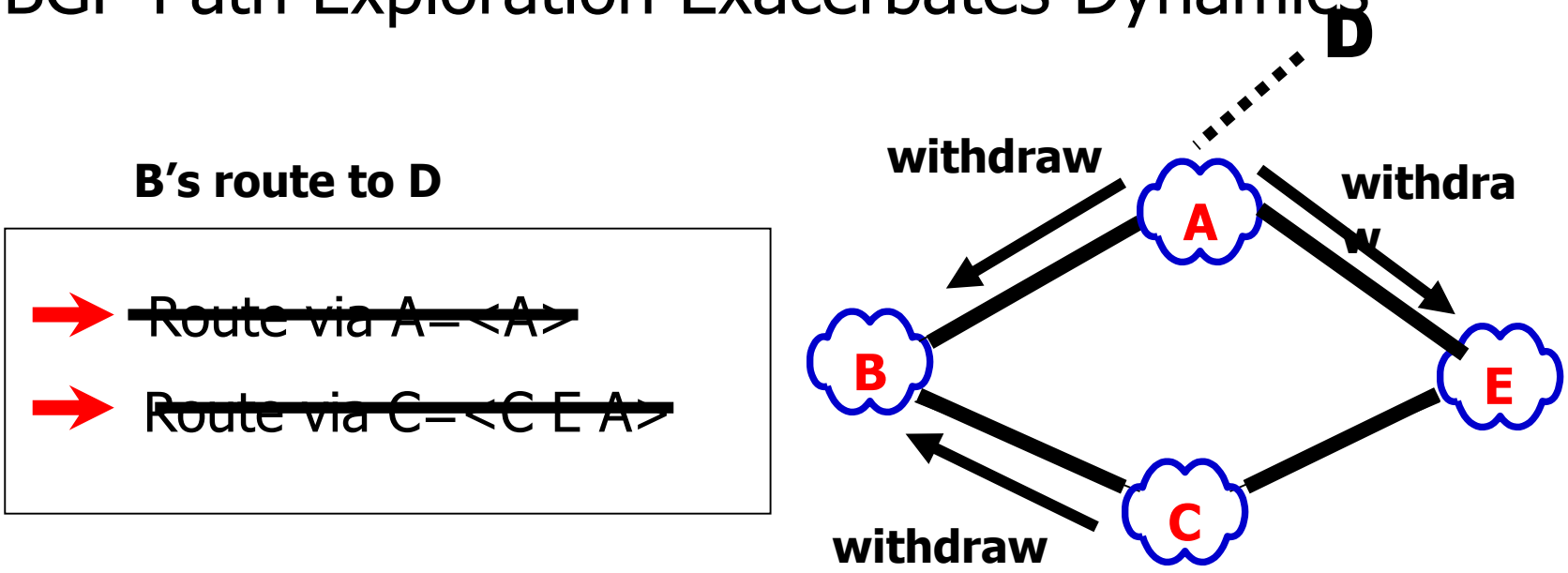
Toward a New Architecture



- One claim: BGP is nearing the end of its useful lifetime
  - The Internet will soon collapse unless we act!!
- Other claim: BGP is the best engineering solution we are likely to produce
  - We need incremental patches to new problems
- Who is right?
- We look some problems facing BGP and ask some fundamental questions regarding BGP.

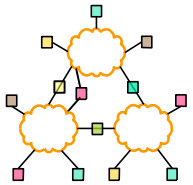
Path Vector Routing Changes

- Worms triggered edge instability
  - Routers crashed due to ARP cache overflow.
  - Links were congested by worm traffic.
- BGP Path Exploration Exacerbates Dynamics



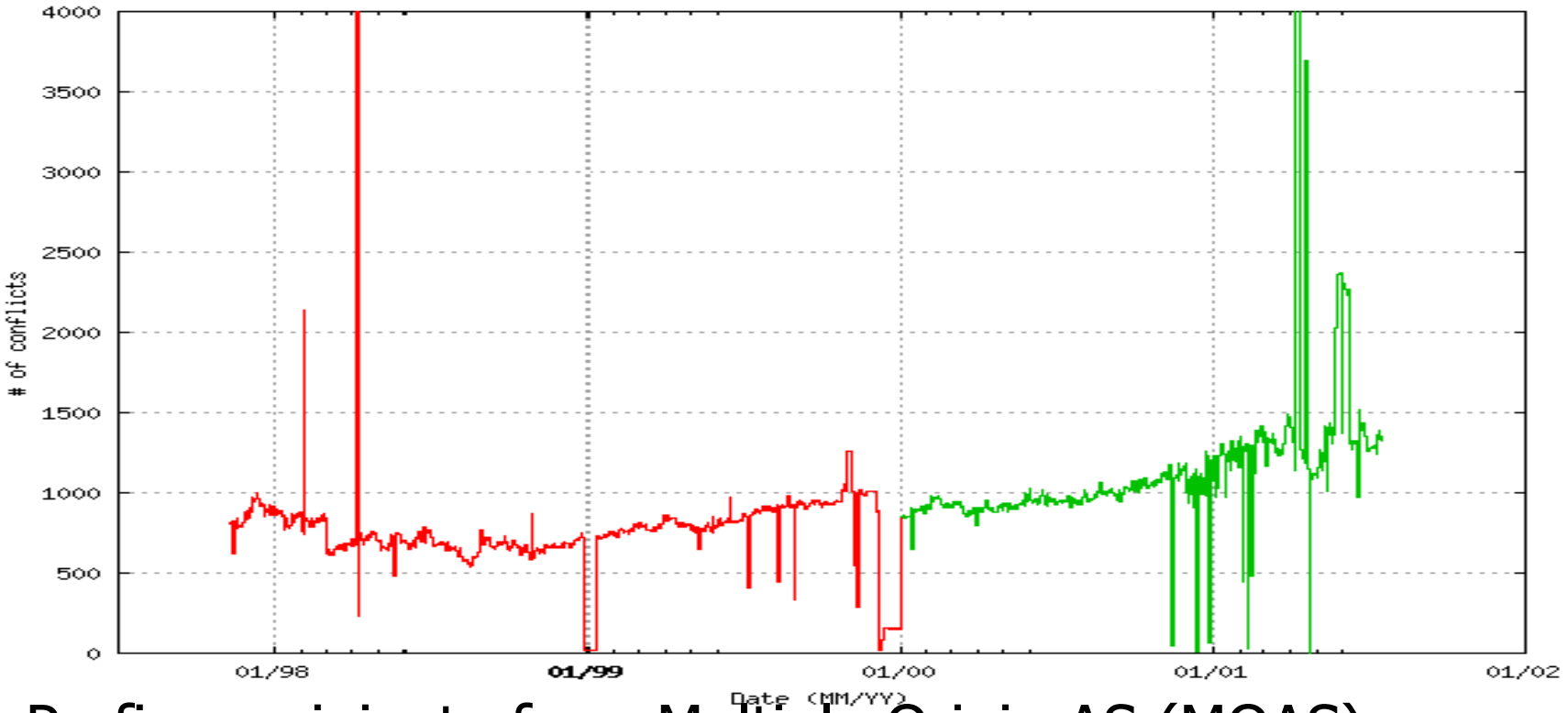
- Obsolete backup path <*C E A*>is used and convergence is delayed

What About Security?



- Convergence Discussion Neglects Security
  - What if routers send intentionally bad information?
- What is the Simplest Possible Attack?
  - Announce someone else routes
- Example: Suppose Univ. of Tehran announces it is the origin for 129.82.0.0/16
  - Can this Happen and/or What Would Prevent It?

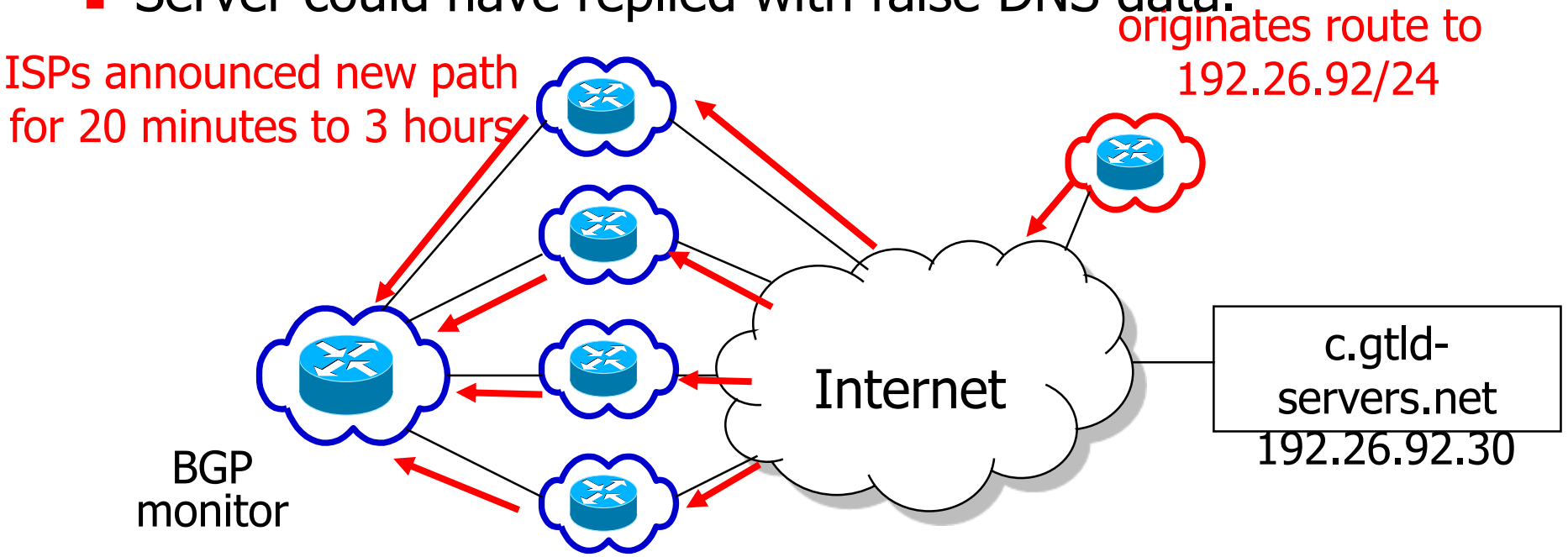
Multiple Origin AS (MOAS) Cases



- Prefixes originate from Multiple Origin AS (MOAS)
  - Lower curve likely due to valid operational needs
- Spikes are errors that disrupt routing to prefix
  - Includes loss of routes to top level DNS servers

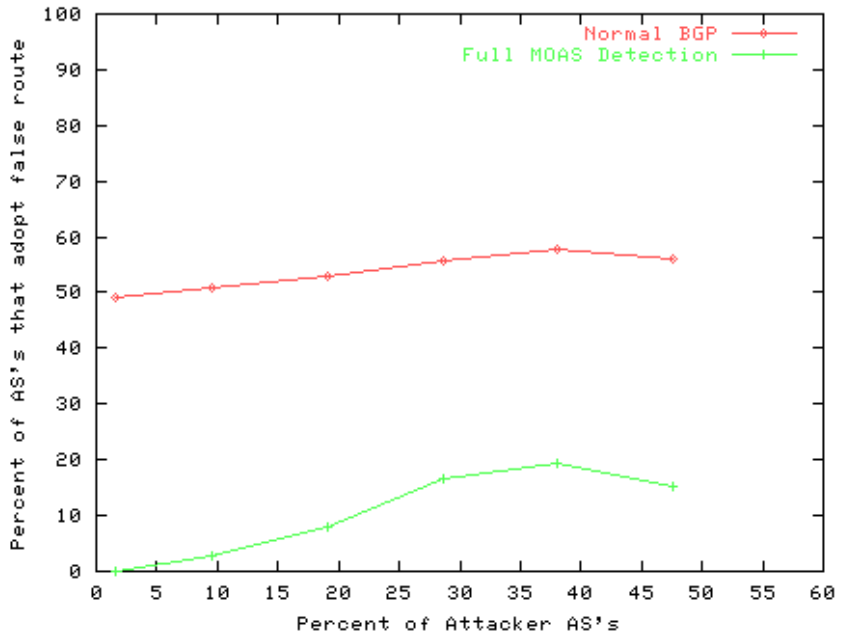
Infrastructure Faults and Attacks

- BGP and DNS Provide No Authentication
  - Faults and attacks can mis-direct traffic.
  - One (of many) examples observed from BGP logs.
  - Server could have replied with false DNS data.

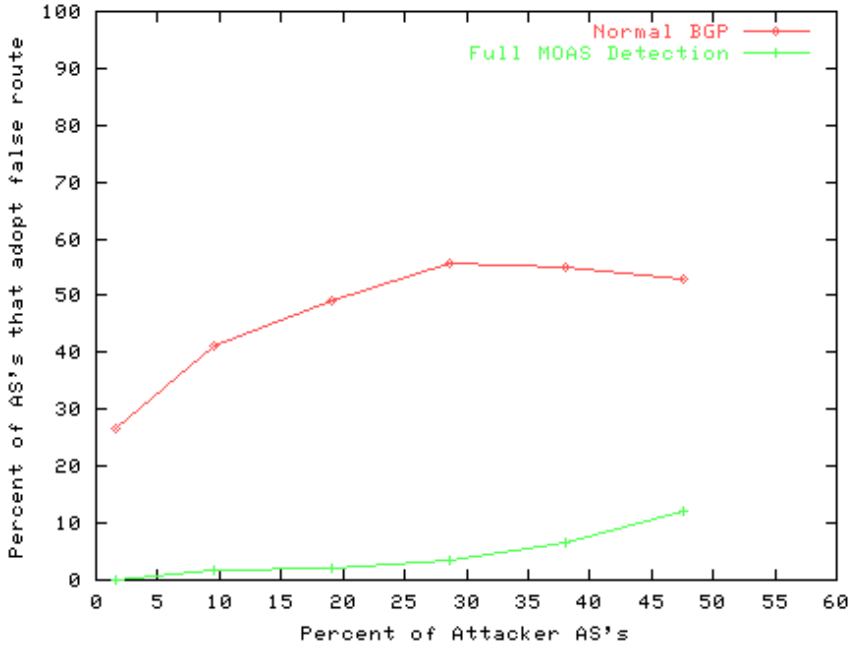


BGP false origin detection

Simulation Results

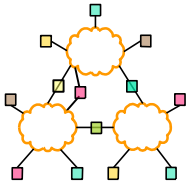


(a) One Origin AS



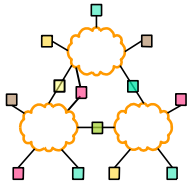
(b) Two Origin AS's

A Simple Filter



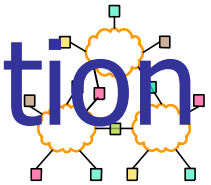
- Current BGP provides dynamic routes
  - Explore the opposite extreme...
- Select a single static route to each server.
  - Apply AS path filters to block all other announcements.
    - Also filter against more specifics.
- Route changes on a frequency of months, if at all.
  - Change in IP address, origin AS, or transit policy.
  - Adjust route only after off-line verification

General view



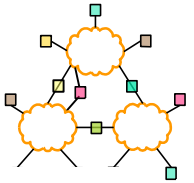
- Scale is limited to a small number of routes.
  - No exponential growth in top level DNS servers.
- Loss of a server is tolerable, invalid server is not.
  - Resolvers detect and time-out unreachable servers.
    - Provided surviving servers handle load, cost is some delay.
- Expect predictable properties and stable routes.
  - Servers don’t change without non-trivial effort.
  - Servers located in highly available locations.

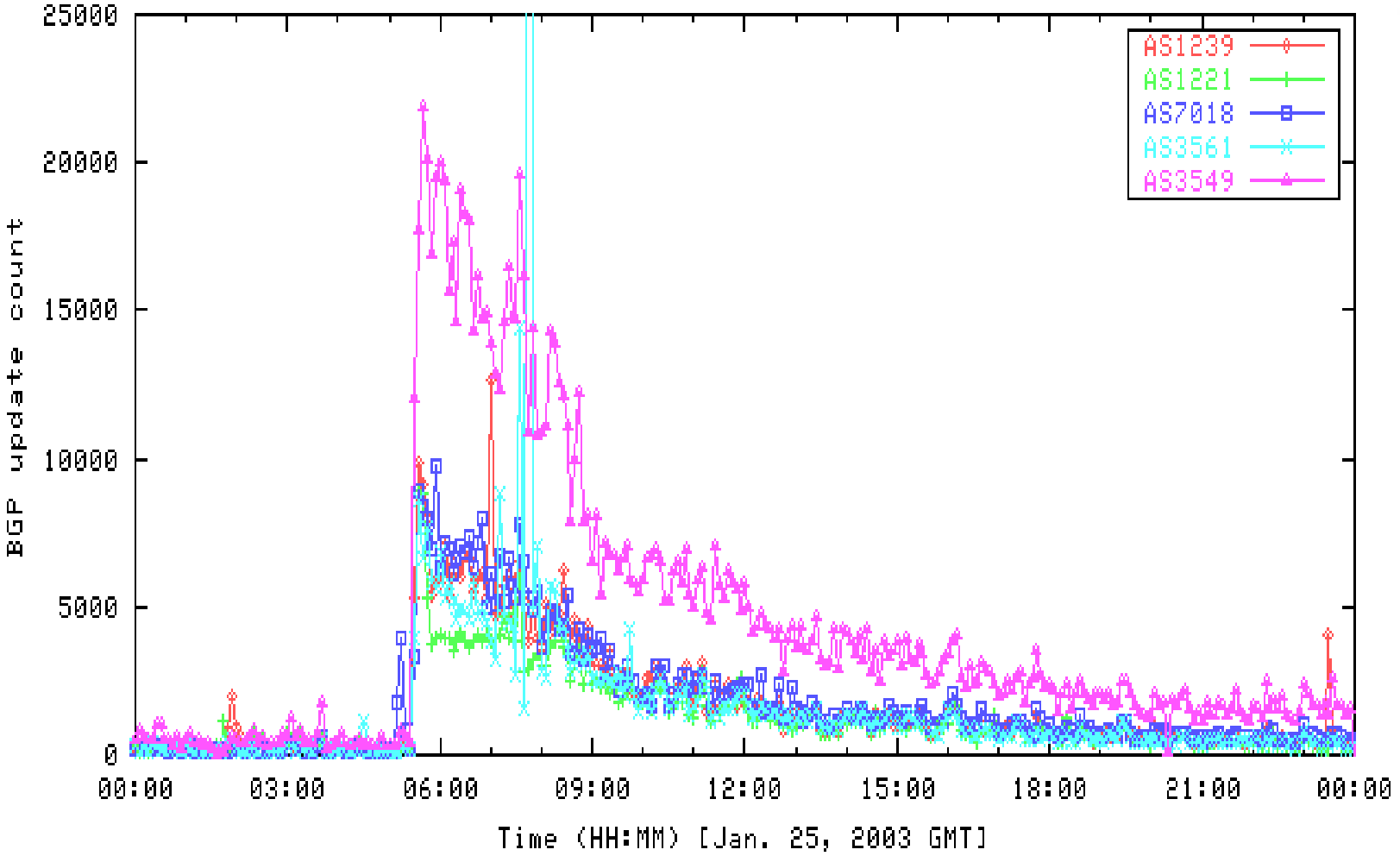
Convergence And Authentication



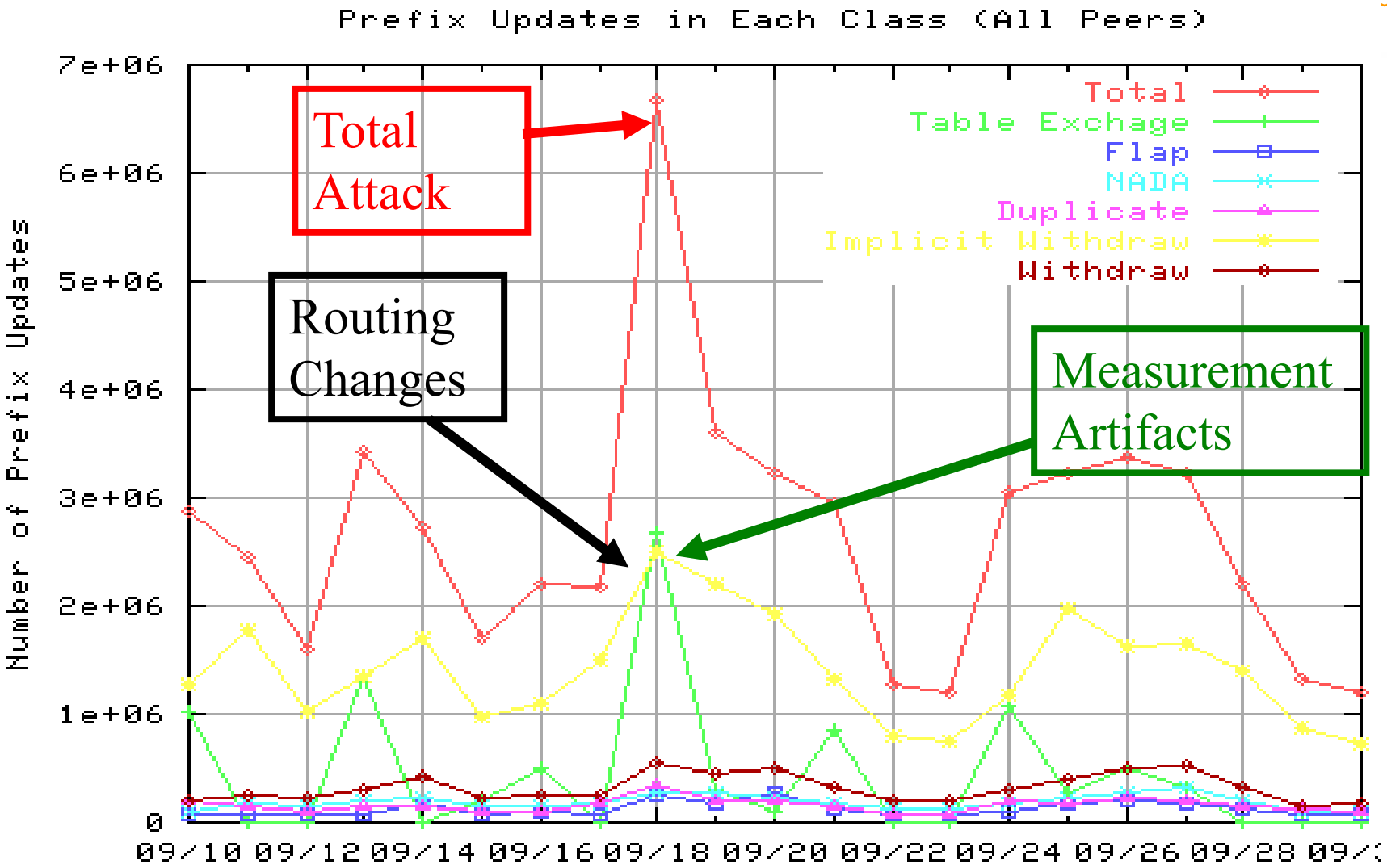
- BGP Suffers From Both Convergence Problems and Authentication Problems
  - Convergence fixes are good, if no attacks.
  - Authentication fixes work for redundant sites
- Can you improve both convergence and authentication in a realistic environment?
- Wide Variety of Other Routing Challenges
  - Check out and BBGP Project in Colorado state univ. if interested

BGP Updates During Worm

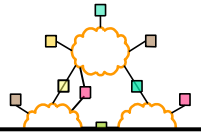


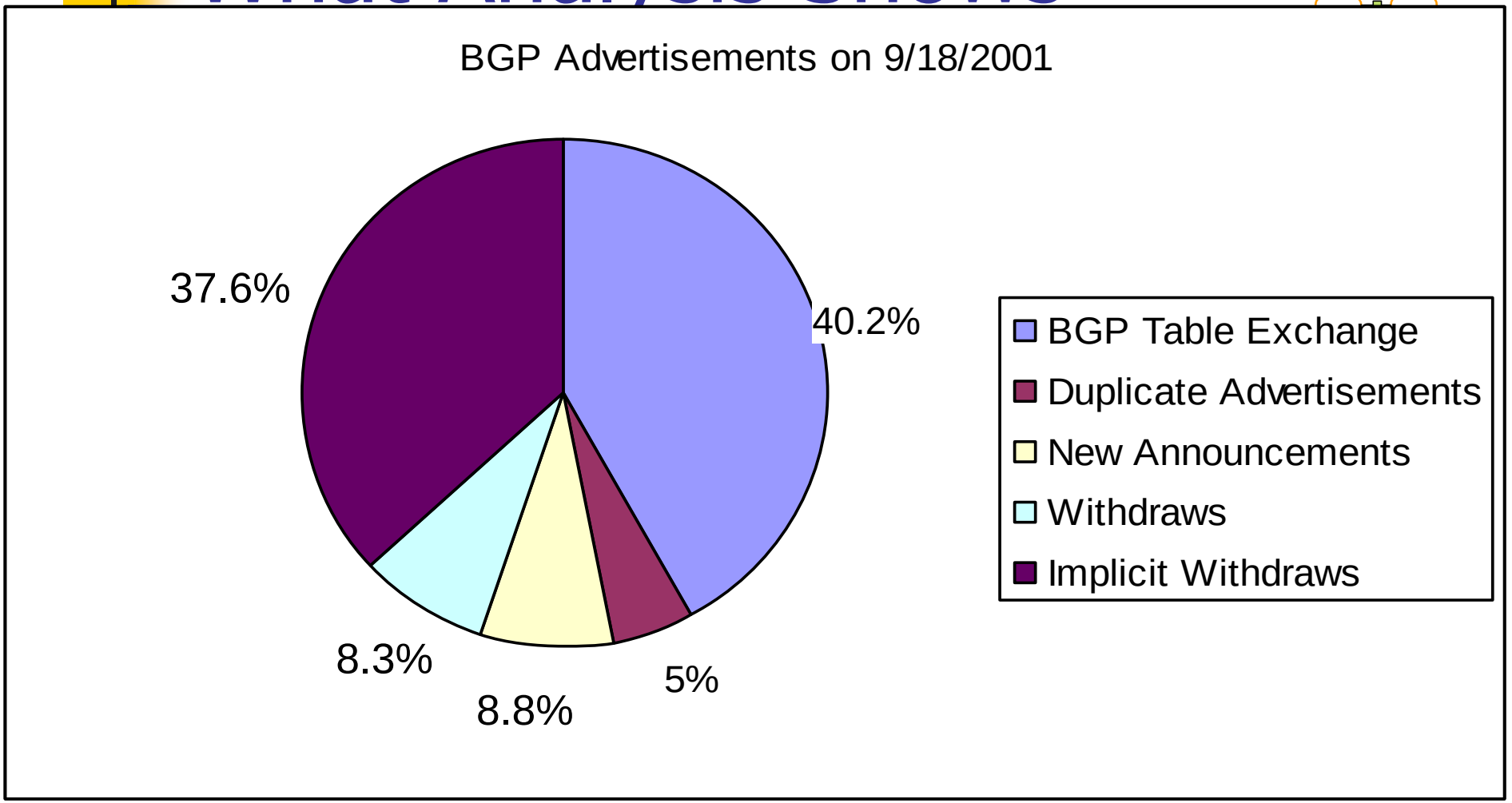


BGP Updates During Nimda Worm



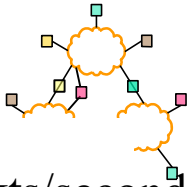
What Analysis Shows



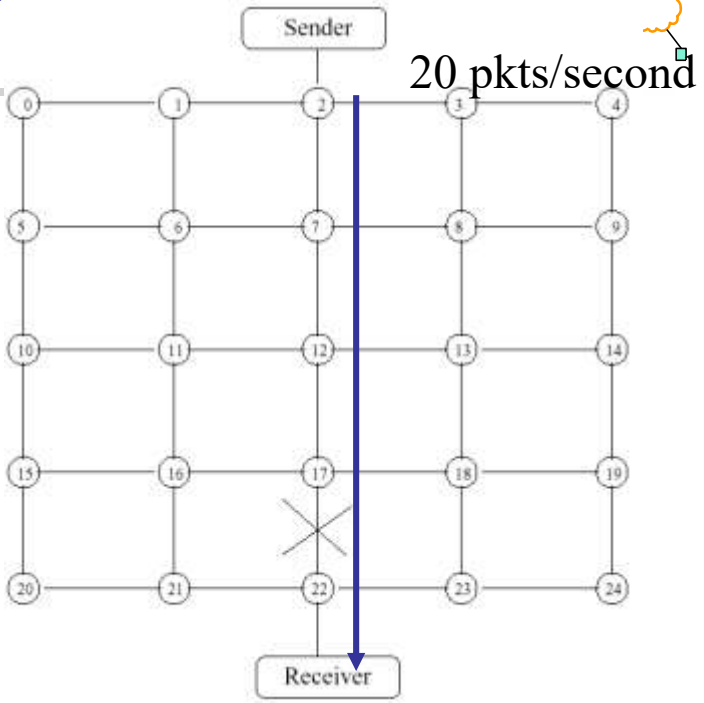


A *substantial* percentage of the BGP messages during the worm attack were not about route changes

Simulation conducted

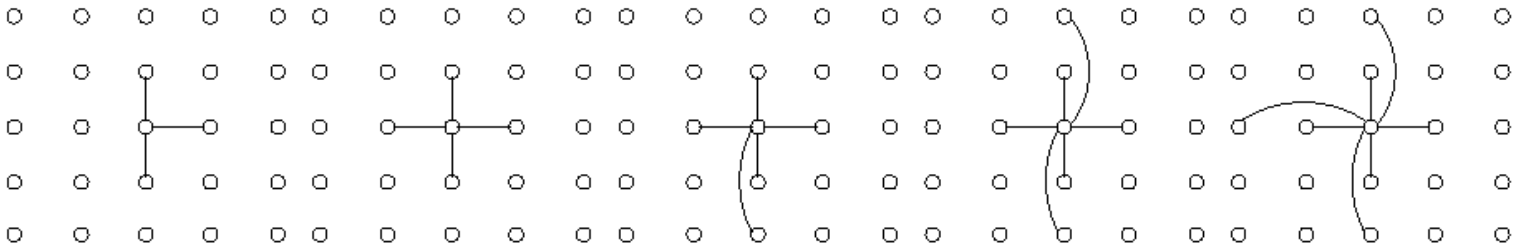


- 7 by 7 mesh topologies similar those in [Baran64]



- Simulated node degree range [3 ~ 16]

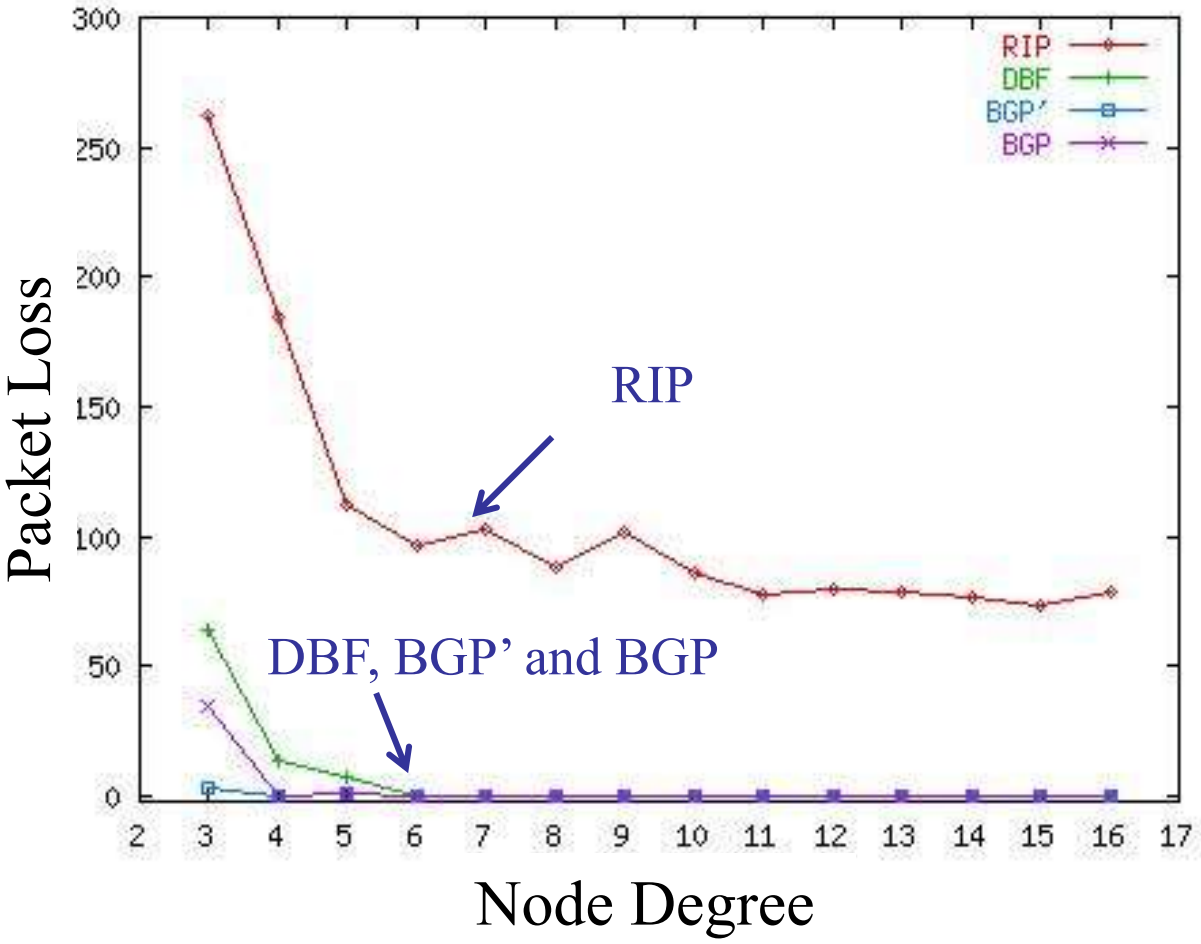
- Measure *Packet loss*, *loops*, path convergence time, throughput, and e2e delay.



(a) degree=3 (b) degree=4 (c) degree=5 (d) degree=6 (e) degree=7

twork

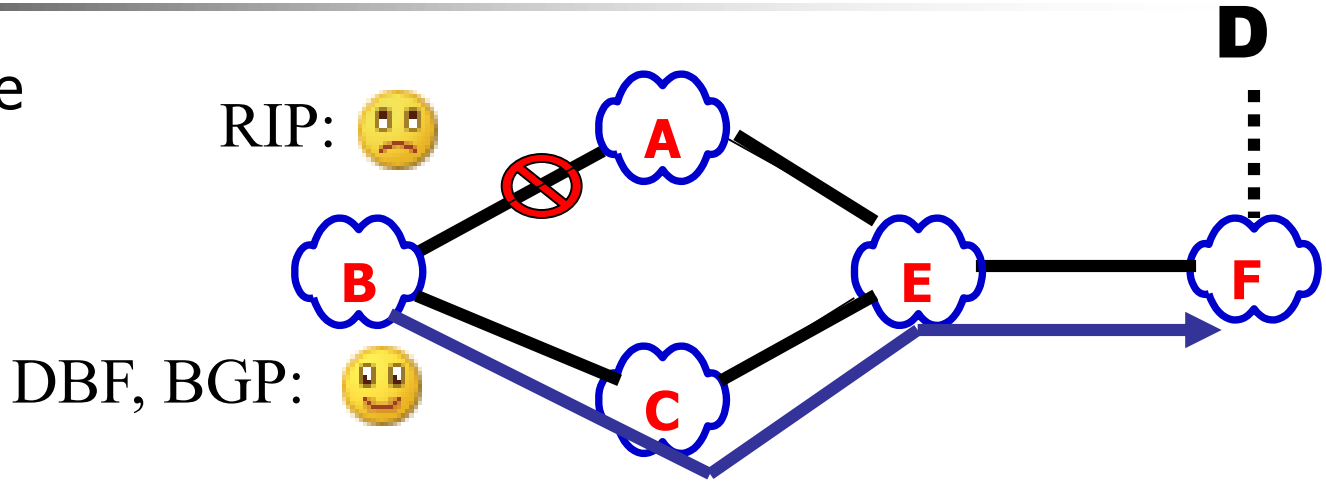
Packet Losses (I) : Observation



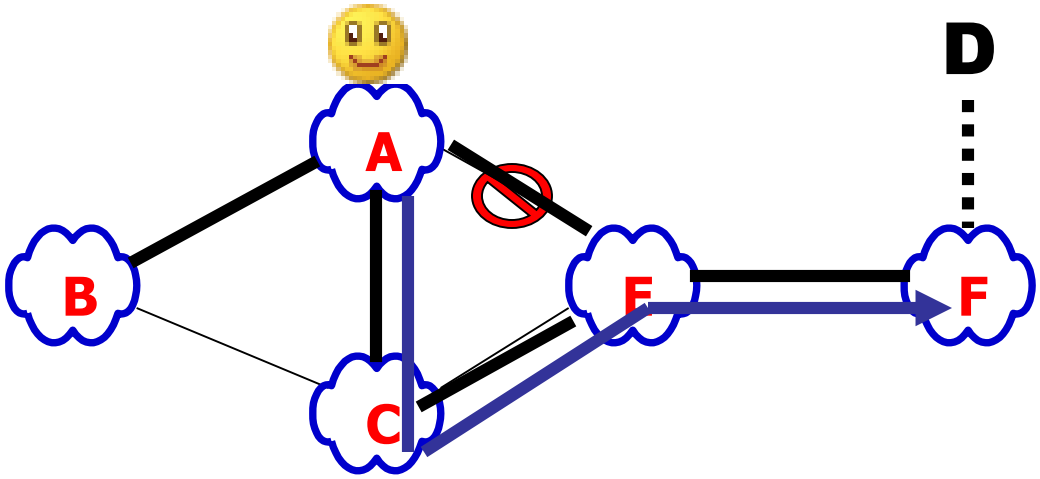
- Packet losses of DBF, BGP' and BGP decrease to zero at degree 6.
- Richer connectivity helps RIP little.
- DBF- Distributed Bellman-Ford

Packet Loss(II): Lessons Learned

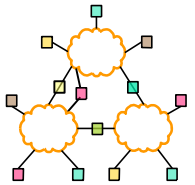
- Keeping alternate paths



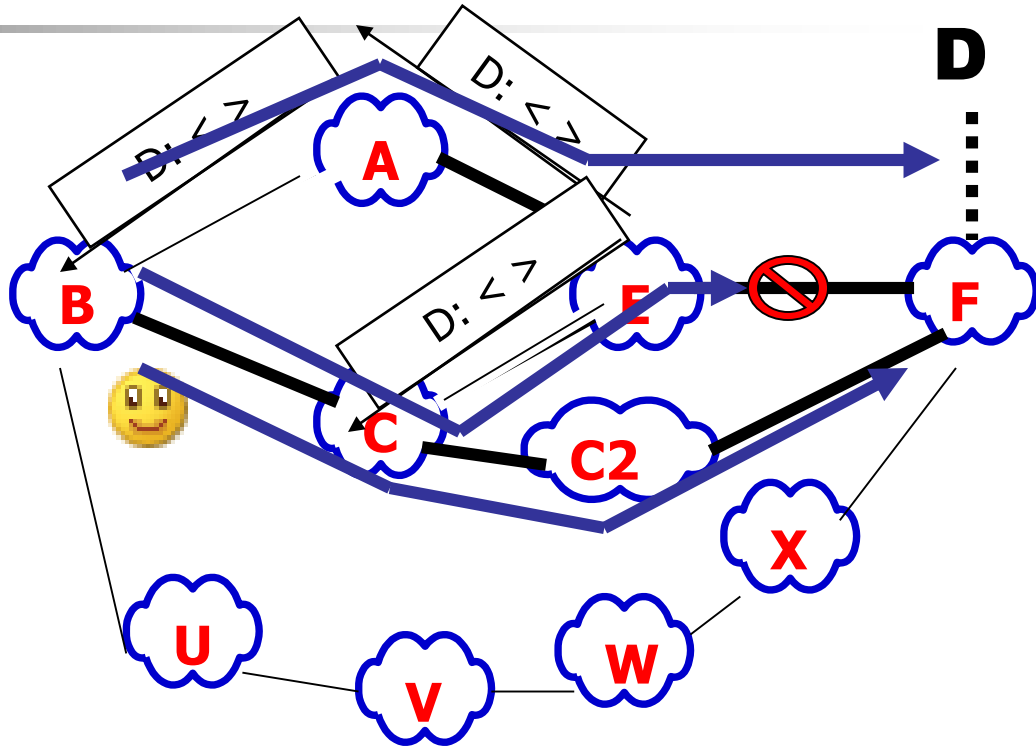
- Connectivity Matters
  - no immediate available alternative due to poor connectivity and poison reverse
  - alternative is more likely with richer connectivity



Is an alternate path valid?

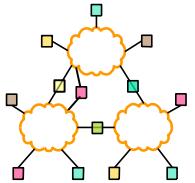


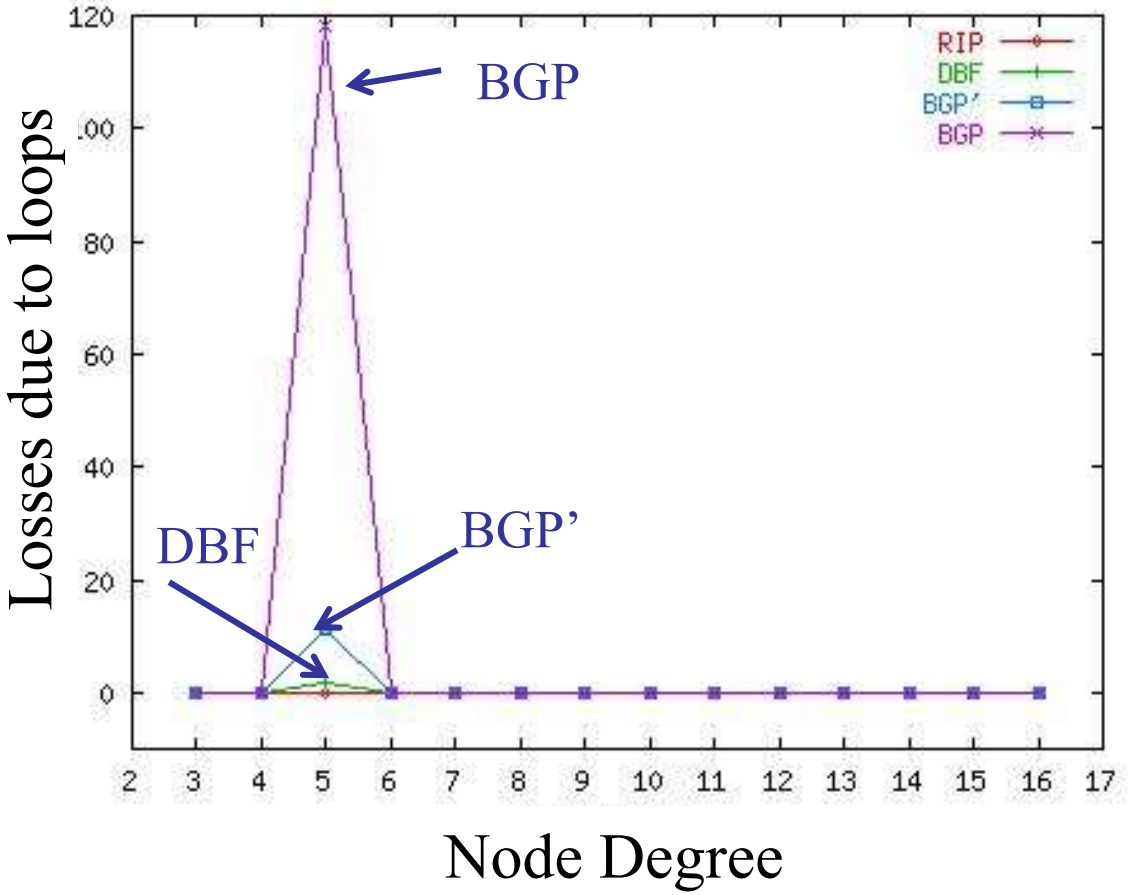
- Valid Alternate Paths: not using the failed link
  - Poison reverse and BGP's path information are not enough! [Pei:Infocom2002]



- Richer connectivity -->
  - reduces one single link's impact
  - better availability of valid(but may be suboptimal) path

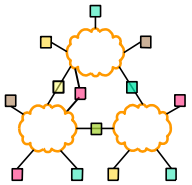
Transient Loops(I): Observation



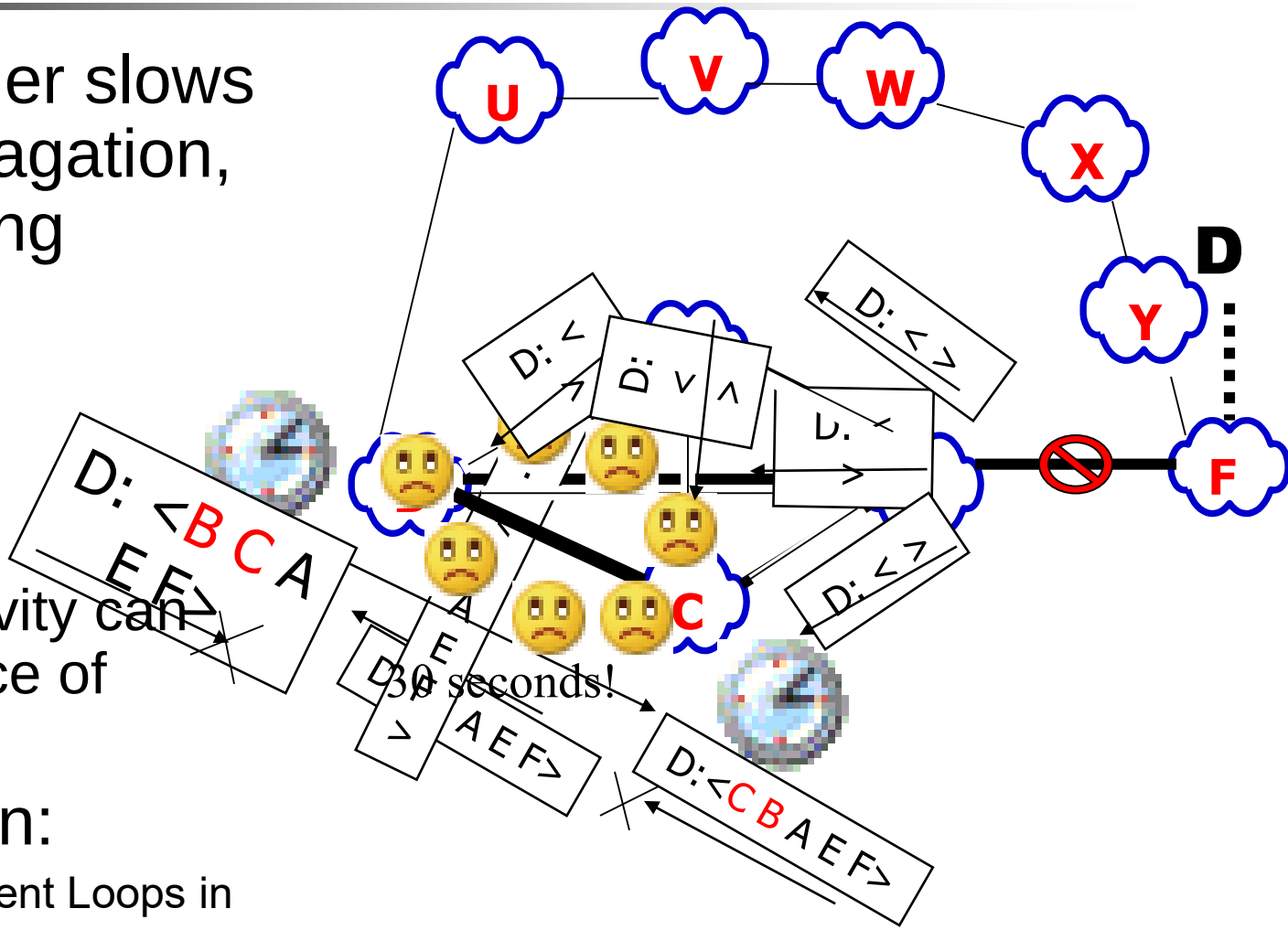


- BGP has the most loops!
- RIP has no loops
- Richer connectivity reduces the chance of looping.

Transient Loops(II): Msg Propagation



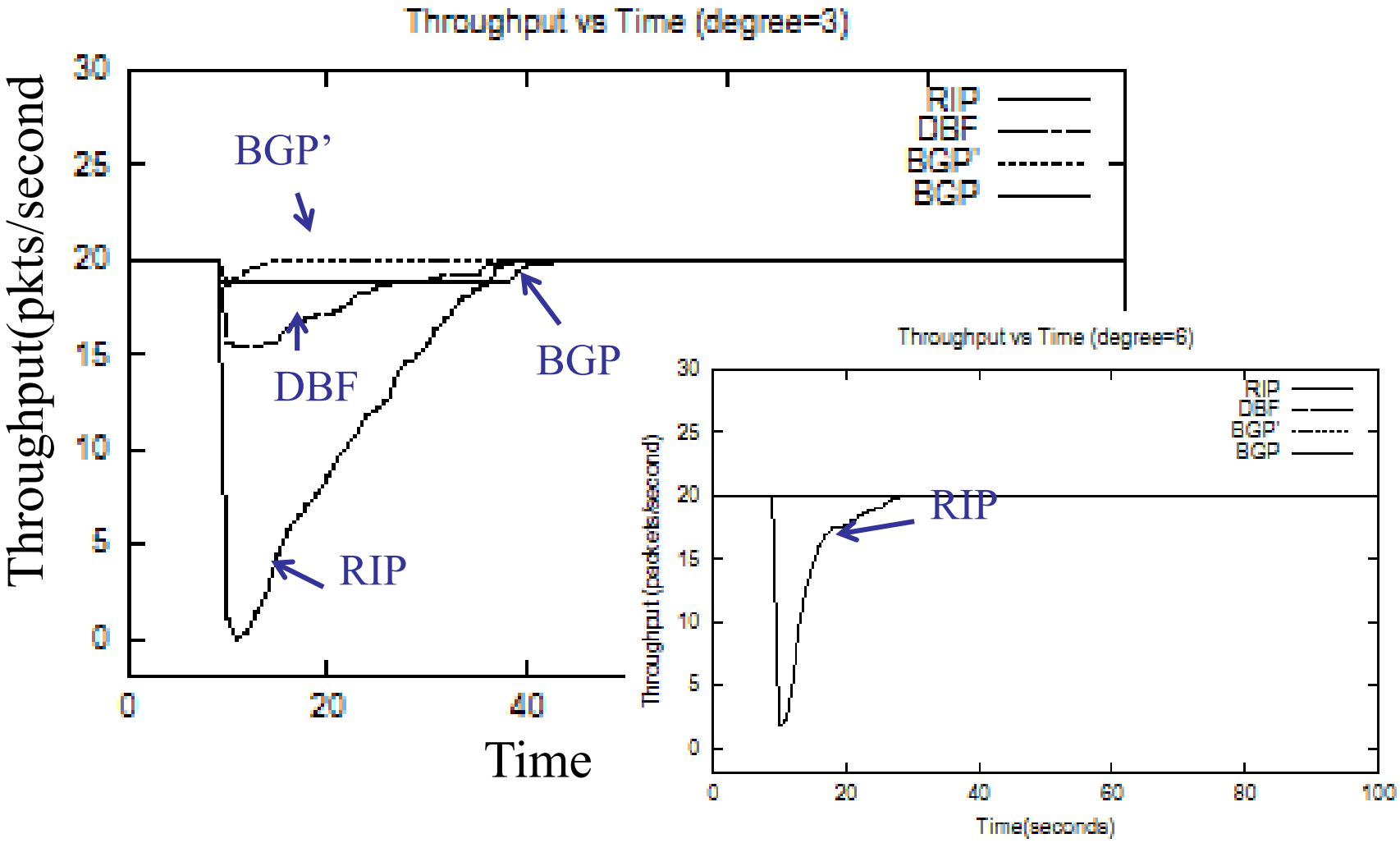
- Damping timer slows the msg propagation, causing looping



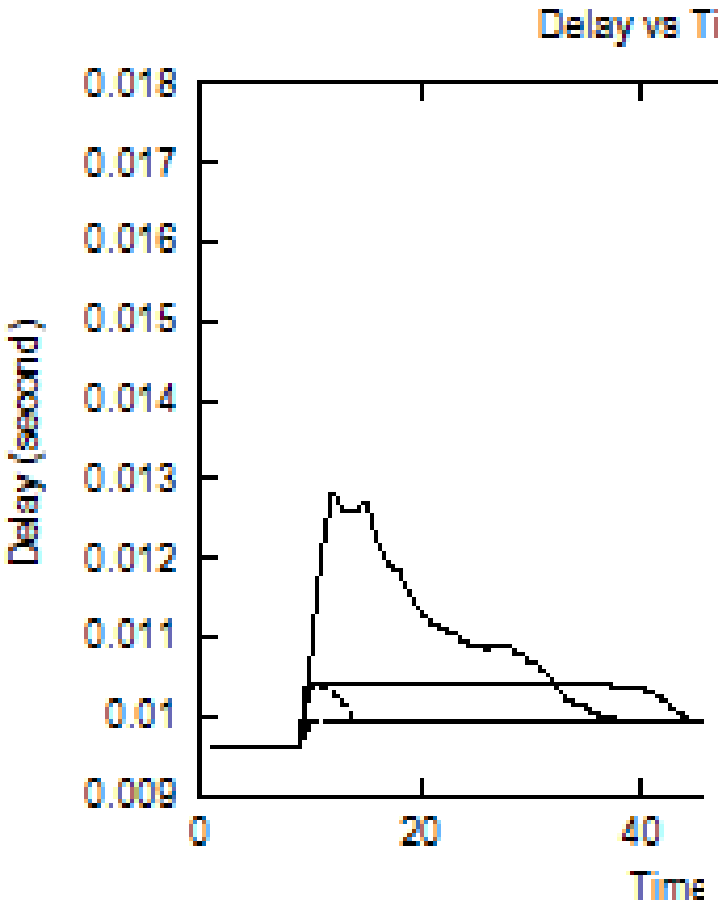
- Richer connectivity can reduce the chance of looping
- More details in:

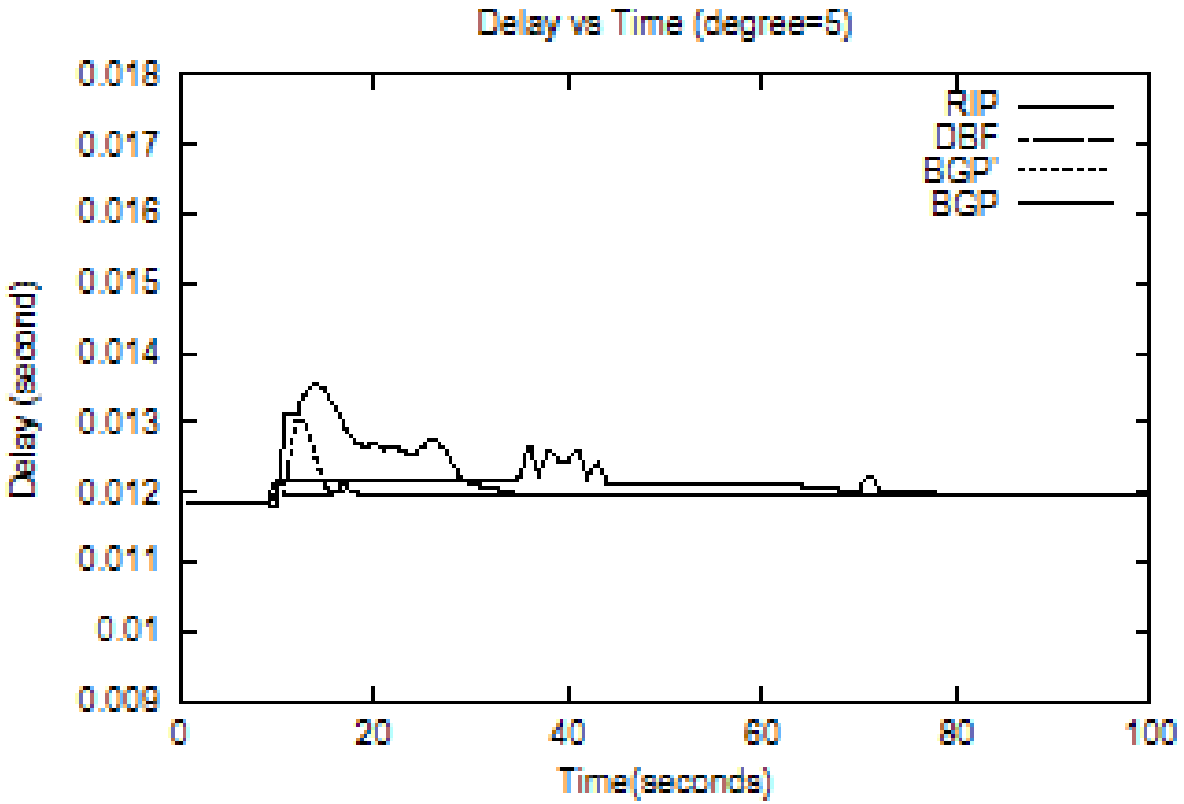
  "A Study of Transient Loops in BGP"

Instantaneous Throughput

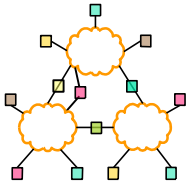


Packet Delay During Convergence



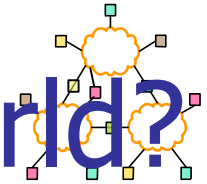


Outline



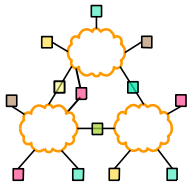
- External BGP (e-BGP)
- Internal BGP (i-BGP)
- Stability Issues
- Scalability Issues

Convergence in the real-world?



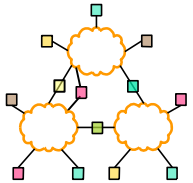
- [Labovitz99] Experimental results from two year study which measured 150,000 BGP faults injected into peering sessions at several IXPs
- Found
  - Internet averages 3 minutes to converge after failover
  - Some multihomed failovers (short to long ASPath) require 15 minutes

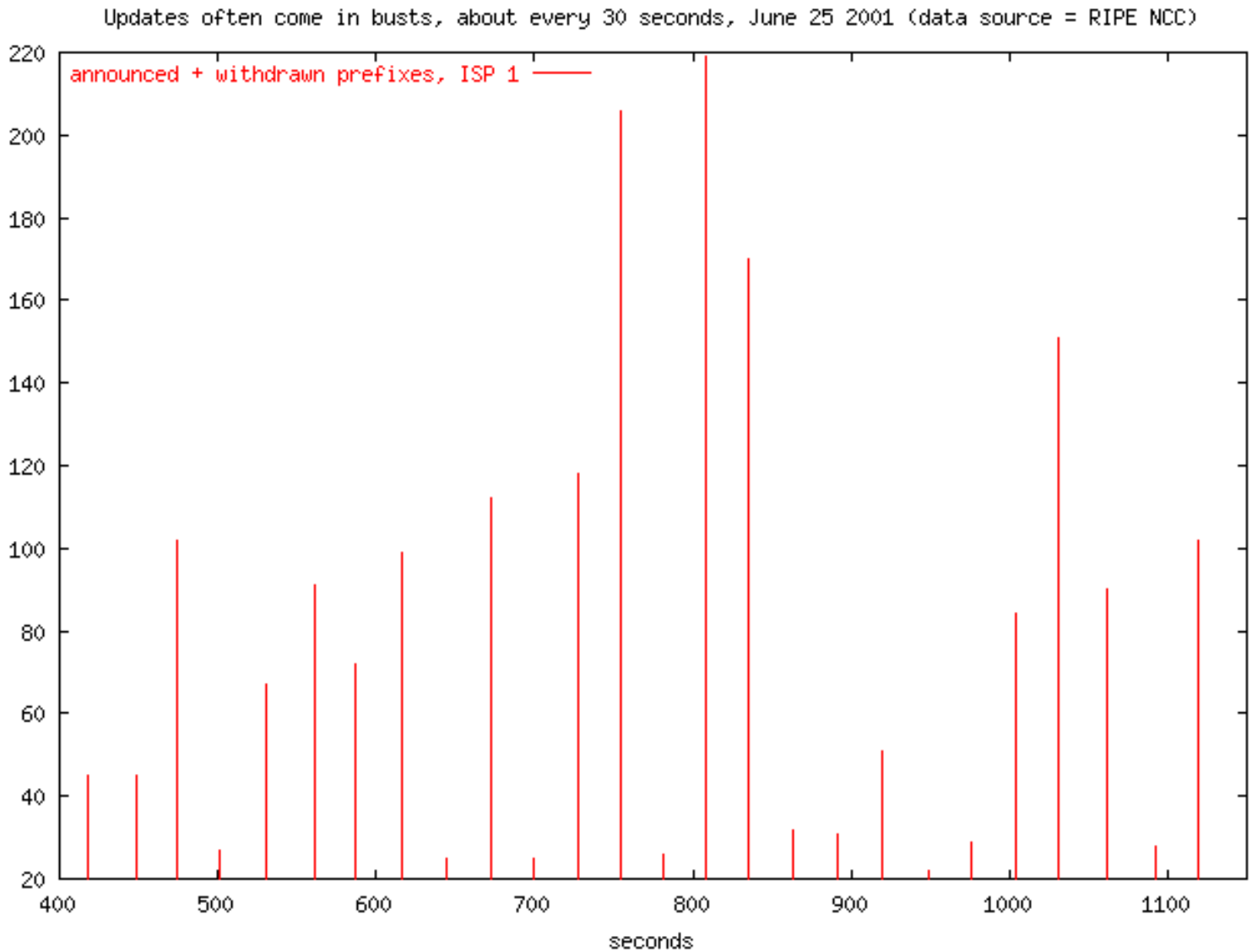
Signs of Routing Instability



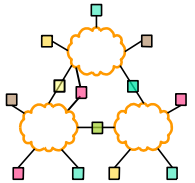
- Record of BGP messages at major exchanges, packet loss 30 times and delay 4.
- Discovered orders of magnitude larger than expected updates
  - Bulk were duplicate withdrawals
    - Stateless implementation of BGP – did not keep track of information passed to peers
    - Impact of few implementations
  - Strong frequency (30/60 sec) components
    - Interaction with other local routing/links etc.

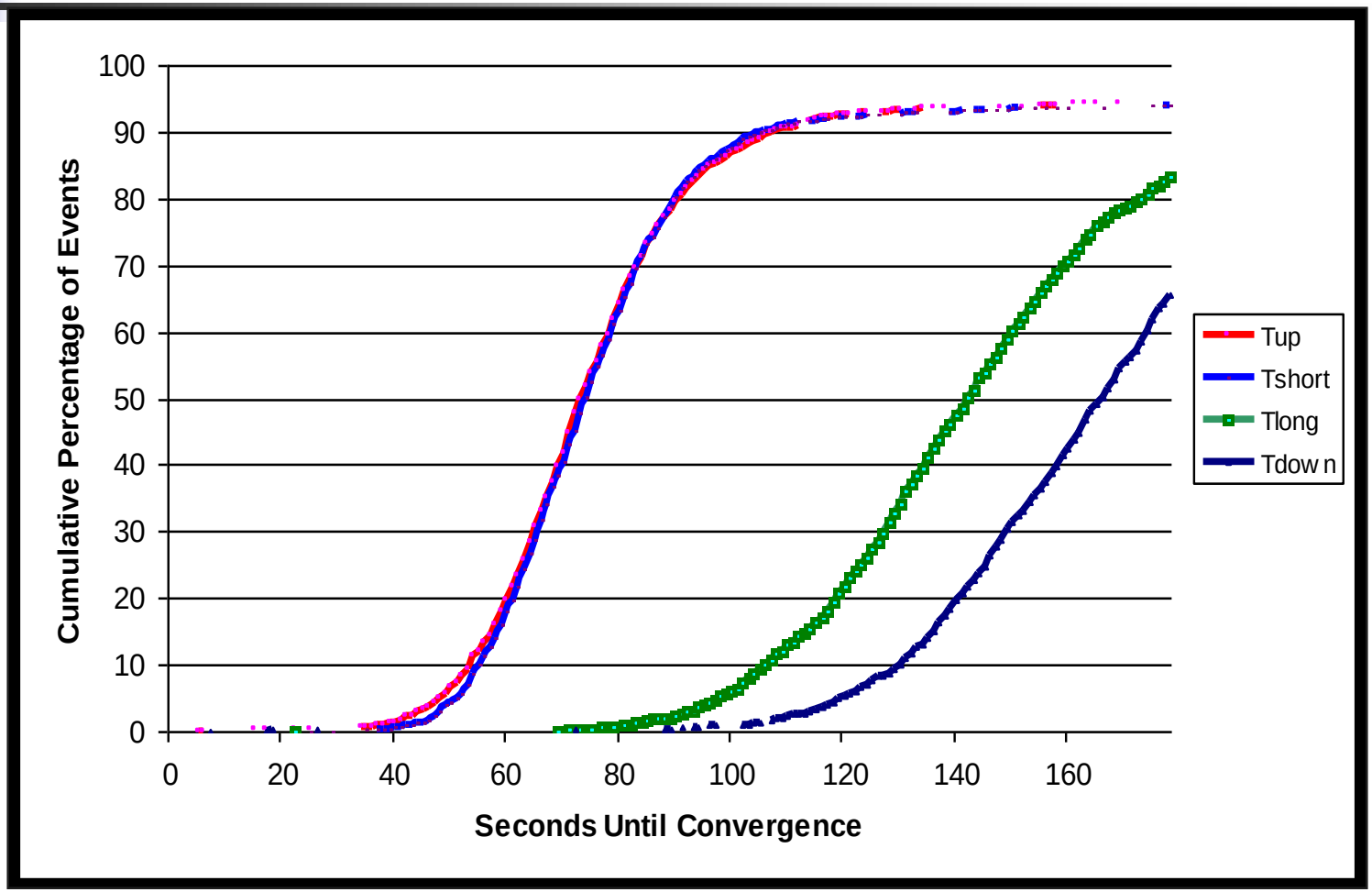
30 Second Bursts





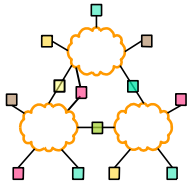
How Long Does BGP Take to Adapt to Changes?





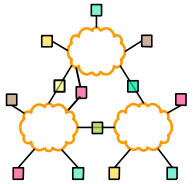
Thanks to Abha Ahuja and Craig Labovitz for this plot.

Route Flap Storm



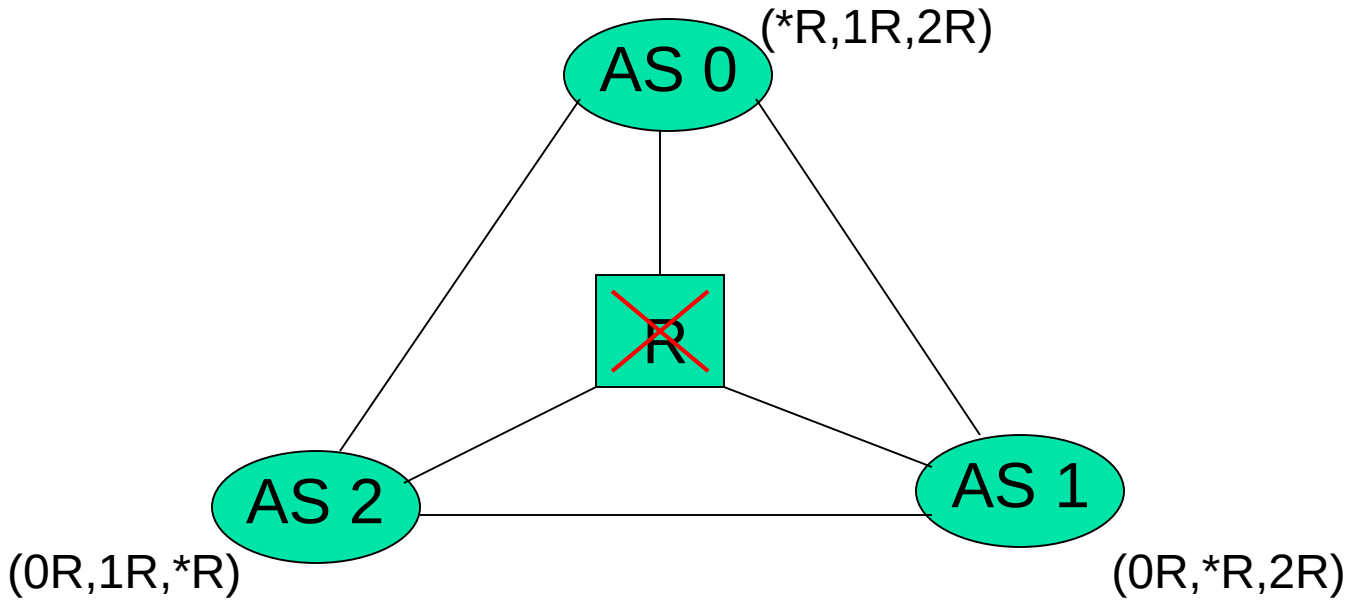
- Overloaded routers fail to send Keep_Alive message and marked as down
- I-BGP peers find alternate paths
- Overloaded router re-establishes peering session
- Must send large updates
- Increased load causes more routers to fail!

Route Flap Dampening

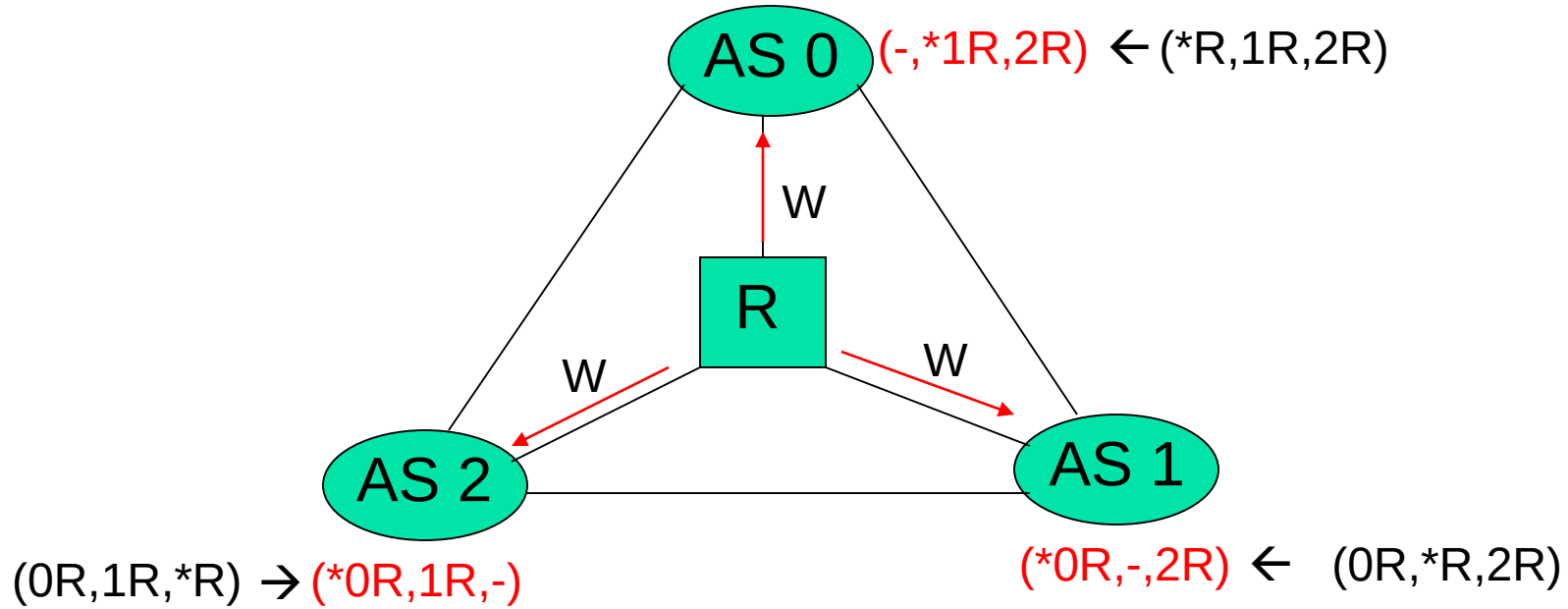


- Routers now give higher priority to BGP/Keep_Alive to avoid problem
- Associate a penalty with each route
  - Increase when route flaps
  - Exponentially decay penalty with time
- When penalty reaches threshold, suppress route

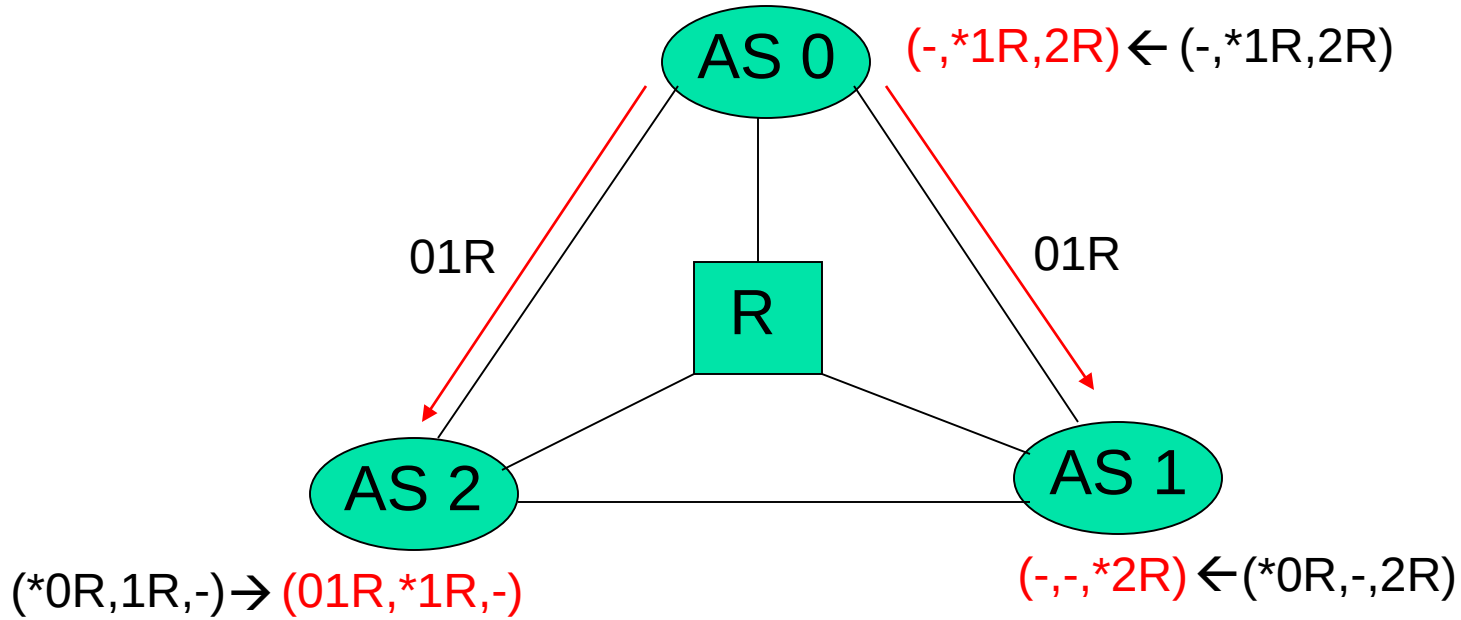
BGP Limitations: Oscillations



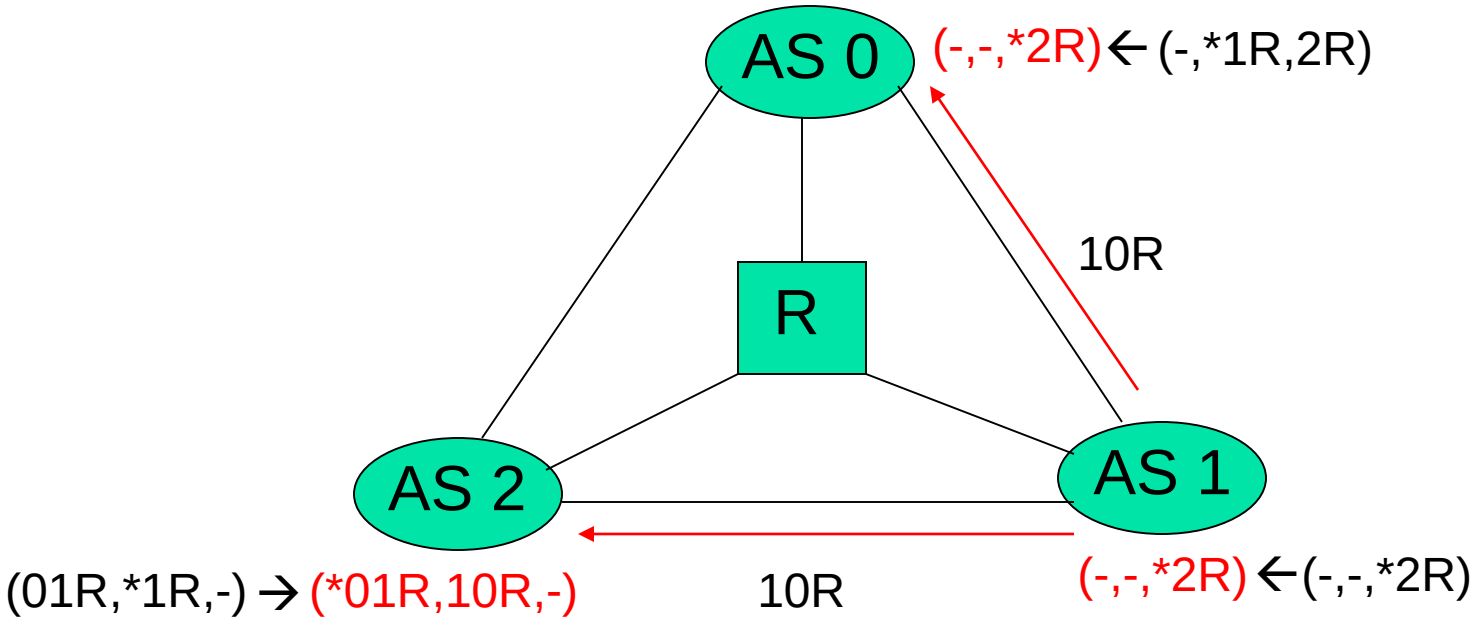
BGP Limitations: Oscillations



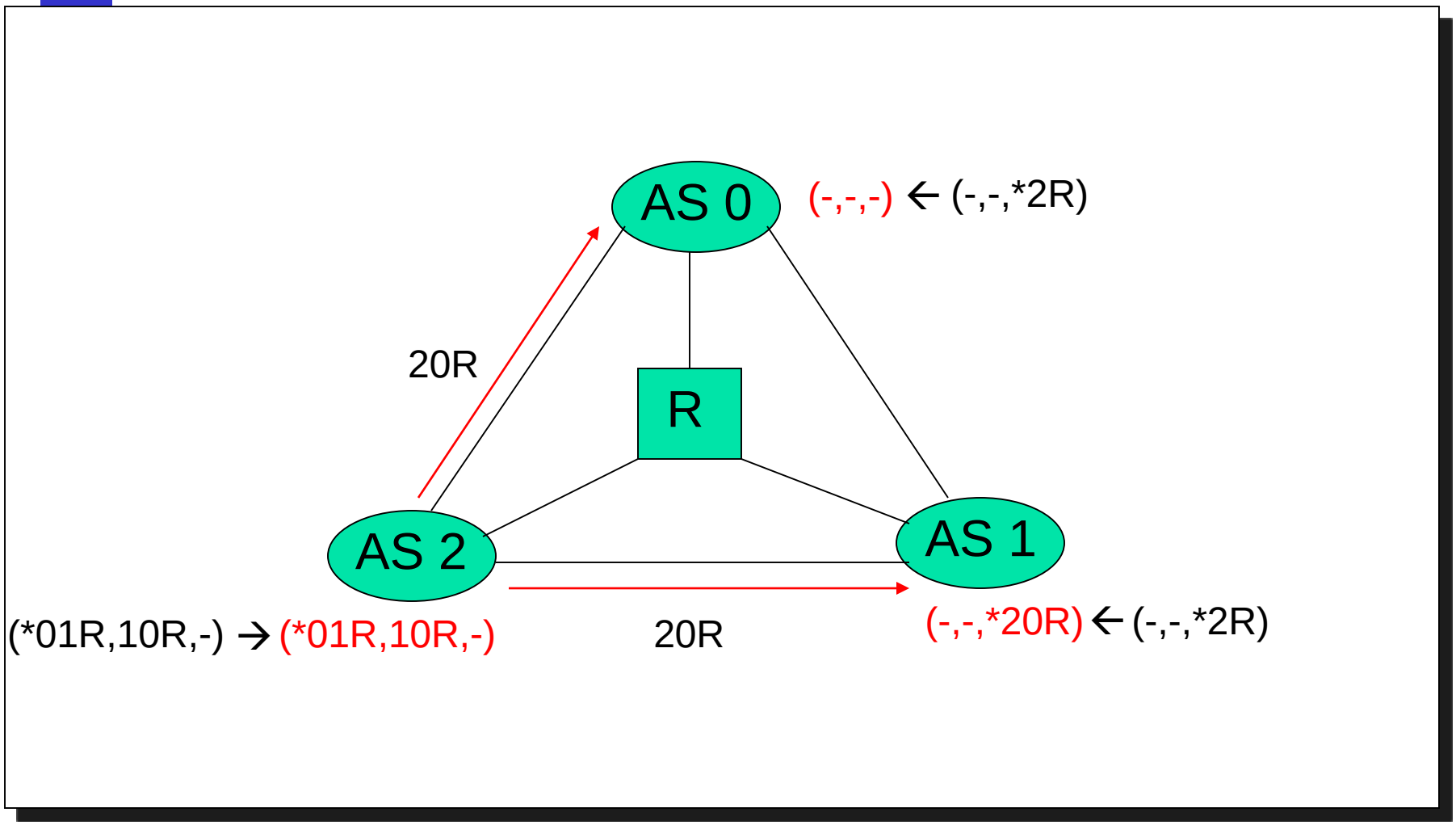
BGP Limitations: Oscillations



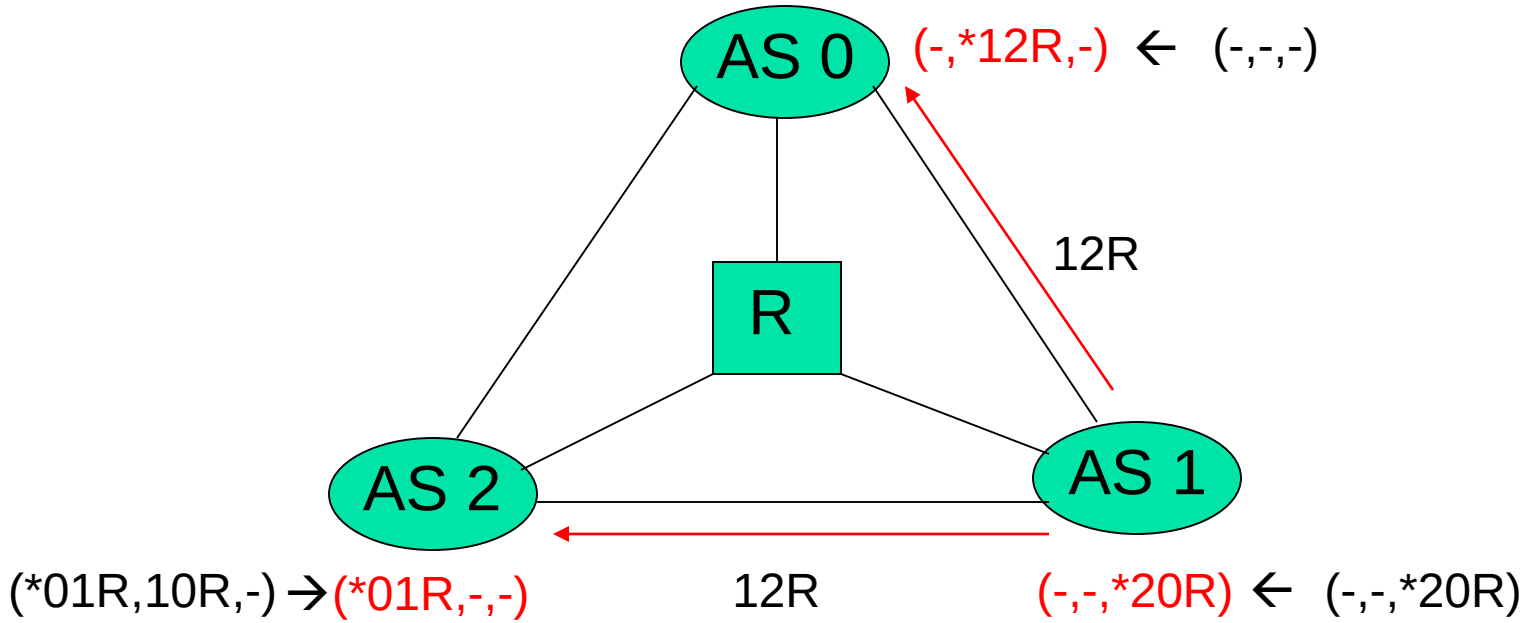
BGP Limitations: Oscillations



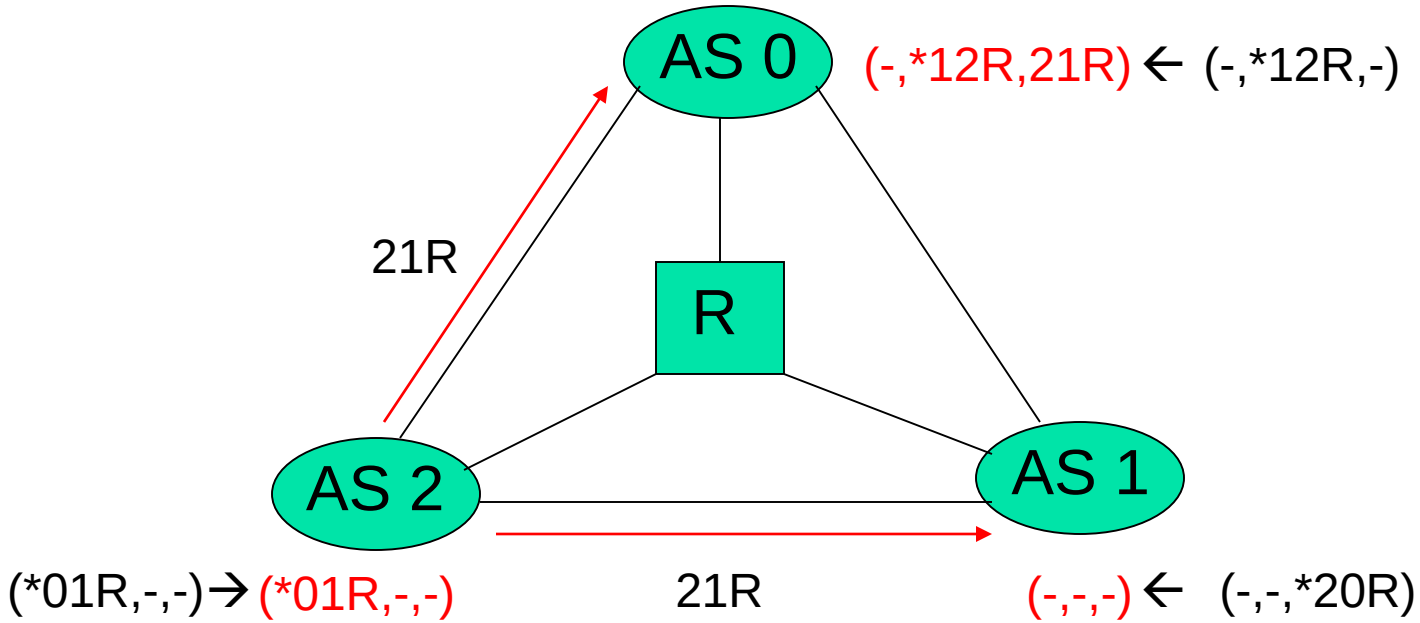
BGP Limitations: Oscillations



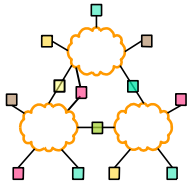
BGP Limitations: Oscillations



BGP Limitations: Oscillations

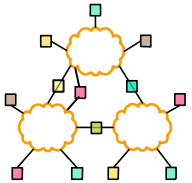


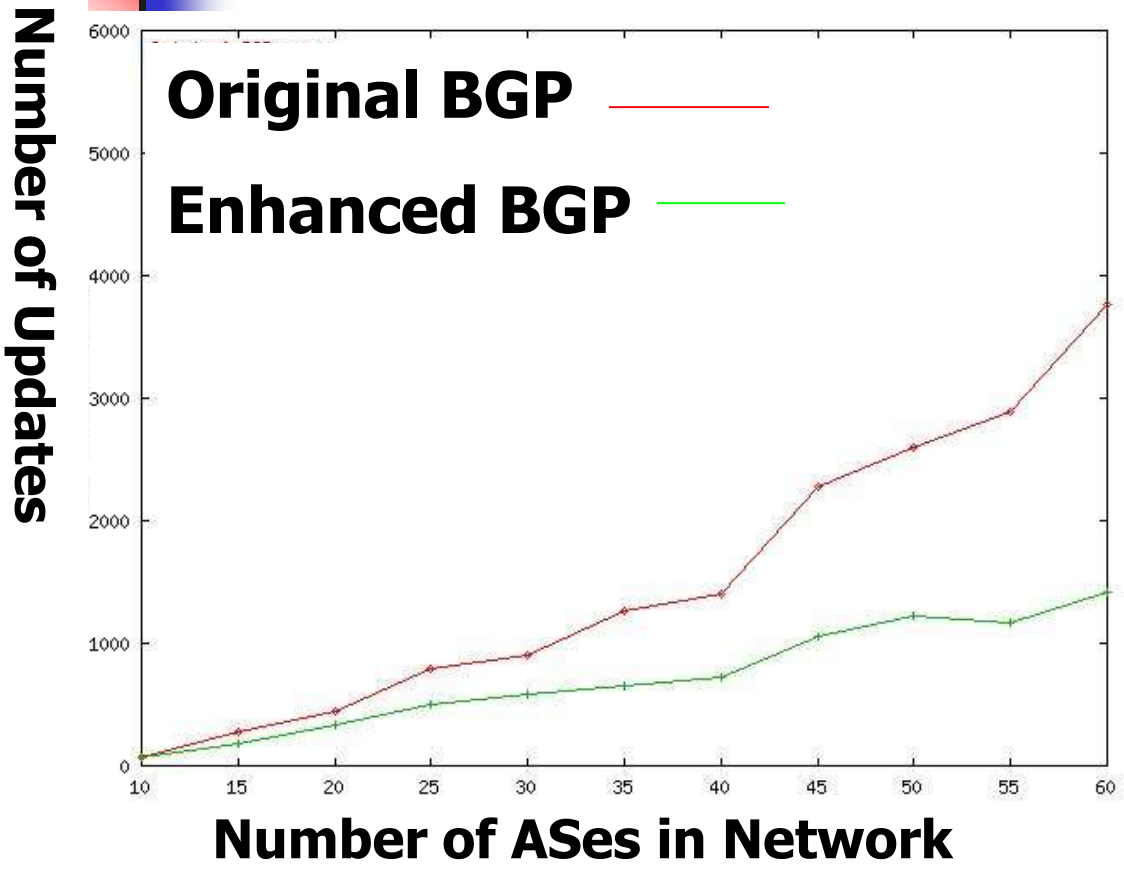
BGP Oscillations



- Can possible explore every possible path through network → (n-1)! Combinations
- Limit between update messages (MinRouteAdver) reduces exploration
  - Forces router to process all outstanding messages
- Typical Internet failover times
  - New/shorter link → 60 seconds
    - Results in simple replacement at nodes
  - Down link → 180 seconds
    - Results in search of possible options
  - Longer link → 120 seconds
    - Results in replacement or search based on length

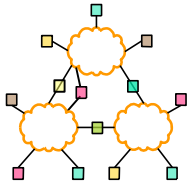
Number of Updates

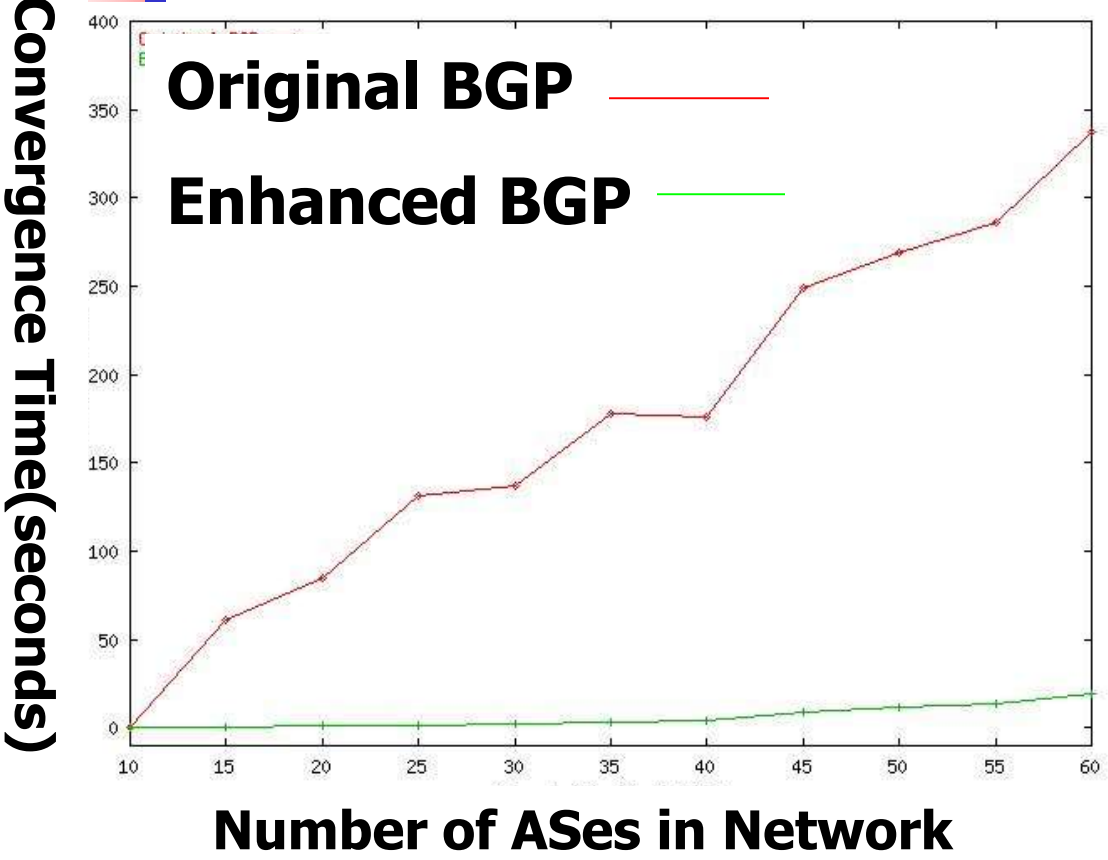




- Substantial reduction is achieved.
  - E.g. 3766 to 1419 in the 60-AS topology
- *MinRouteAdver* timer: within 30 seconds, only one advertisement is allowed.
- It "packs" consecutive changes into one update.

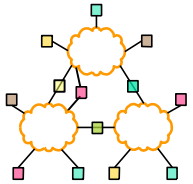
Convergence time





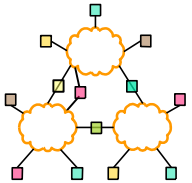
- Enhanced BGP reduces the convergence time substantially.
  - E.g. 337.0 seconds to 19.5 seconds in the 60-AS topology
  - Elimination of one advertisement can cut convergence time by 30 seconds

Problems



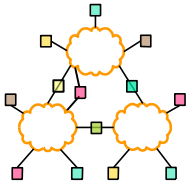
- Routing table size
  - Need an entry for all paths to all networks
- Required memory= O((N + M*A) * K)
  - N: number of networks
  - M: mean AS distance (in terms of hops)
  - A: number of AS's
  - K: number of BGP peers

Outline



- External BGP (e-BGP)
- Internal BGP (i-BGP)
- Stability Issues
- Scalability Issues

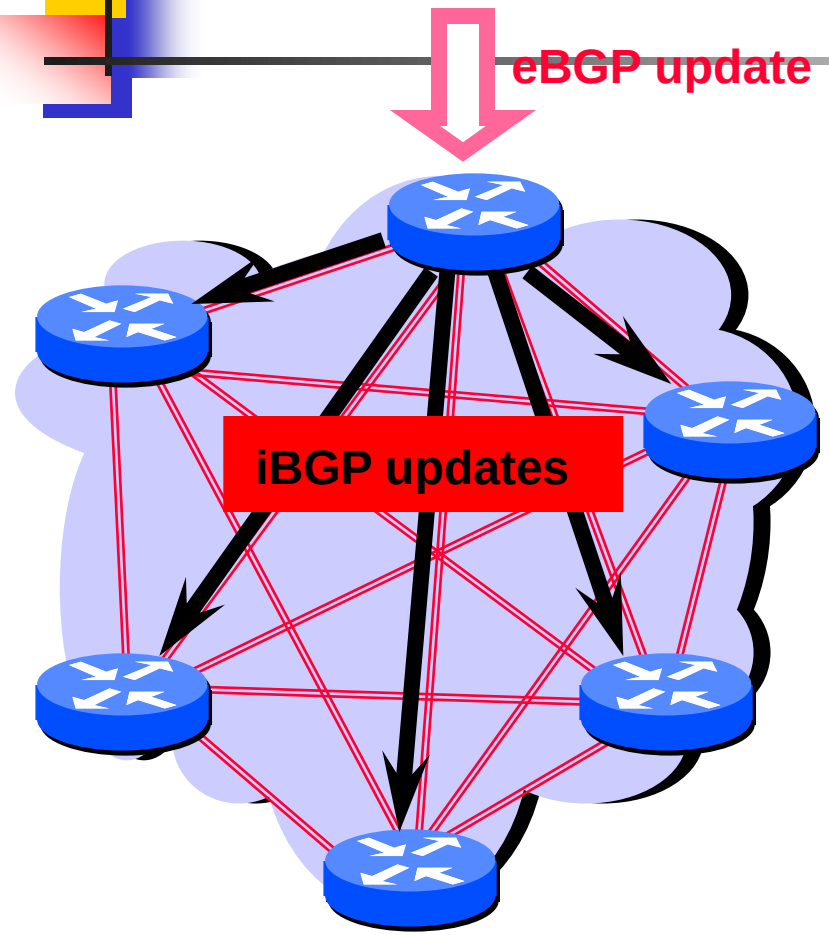
Big and Getting Bigger



- Scaling the iBGP mesh
  - Confederations
  - Route Reflectors
- BGP Table Growth
  - Address aggregation (CIDR)
  - Address allocation
- AS number allocation and use
- Dynamics of BGP
  - Inherent vs. accidental oscillation
  - Rate limiting and route flap dampening
  - Lots and lots of noise
  - Slow convergence time

Scale
Scale
Scale
Scale
Scale
Scale
Scale
Scale
Scale
Scale
Scale
Scale
Scale

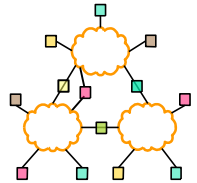
iBGP Mesh Does Not Scale



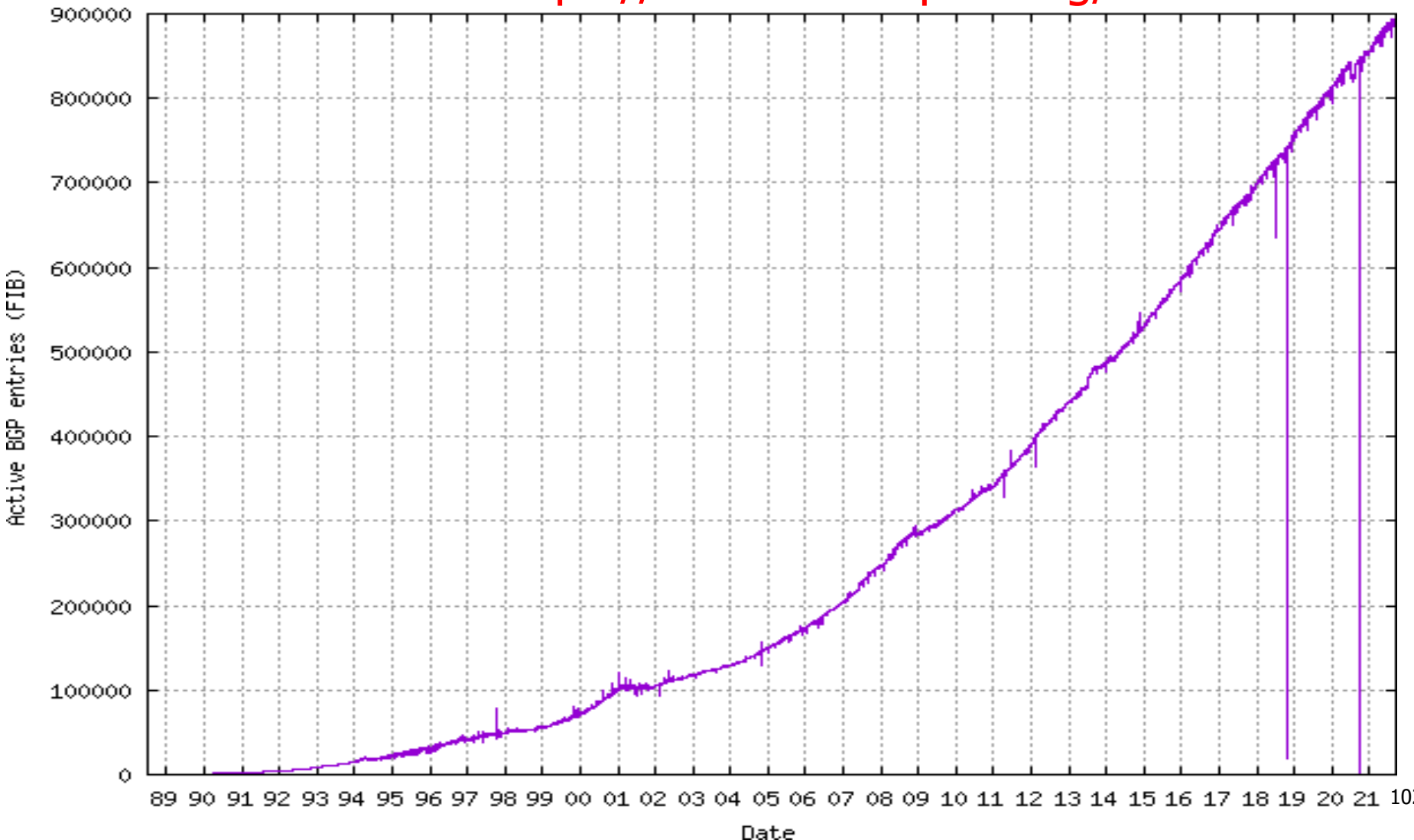
- N border routers means N(N-1)/2 peering sessions
- Each router must have N-1 iBGP sessions configured
- The addition a single iBGP speaker requires configuration changes to all other iBGP speakers
- Size of iBGP routing table can be order N larger than number of best routes (remember alternate routes!)
- Each router has to listen to update noise from each neighbor

Currently four solutions:
(0) Buy bigger routers!
(1) Break AS into smaller ASes
(2) BGP Route reflectors
(3) BGP confederations

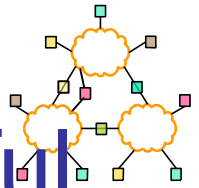
Routing Table Growth (IPv4)



https://www.cidr-report.org/



Large BGP Tables Considered Harmful

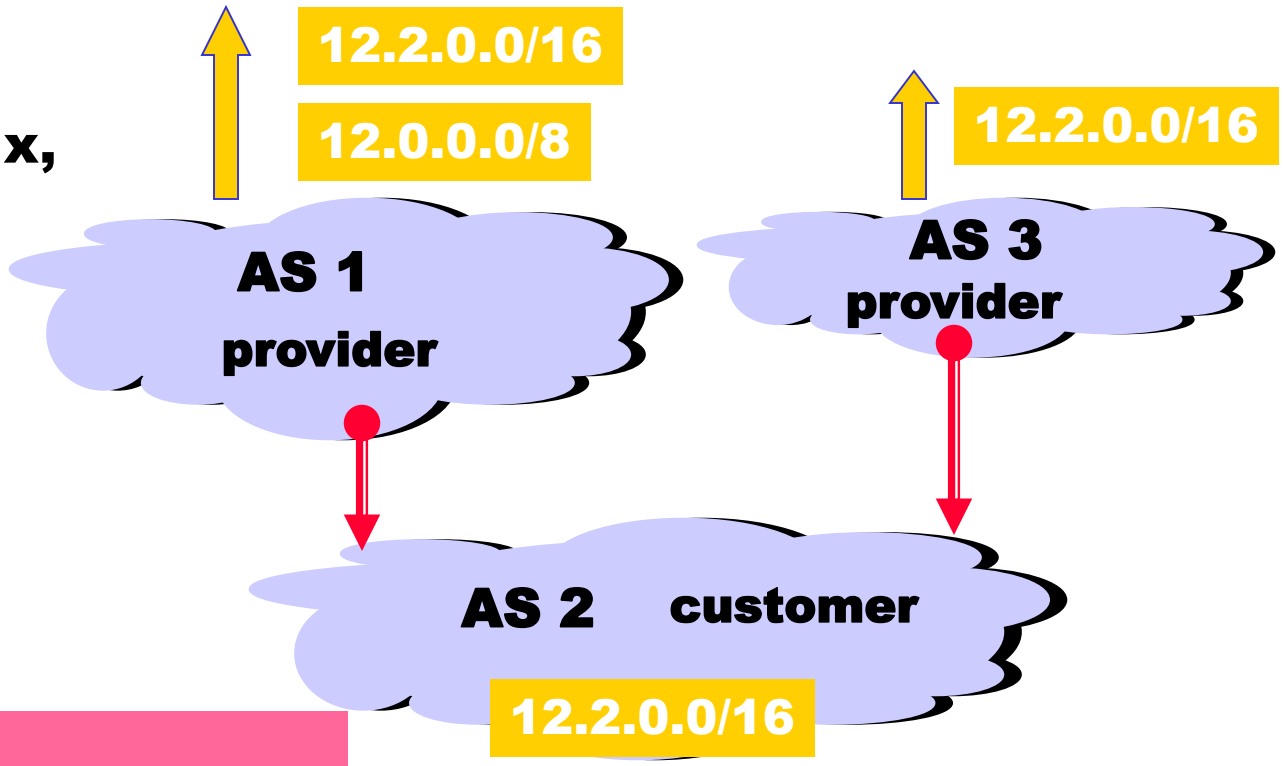


- **Routing tables must store best routes and alternate routes**
- **Burden can be large for routers with many alternate routes (route reflectors for example)**
- **Routers have been known to die**
- **Increases CPU load, especially during session reset**

Moore's Law may save us in theory. But in practice it means spending money to upgrade equipment ...

Deaggregation Due to Multihoming May be a Leading Cause

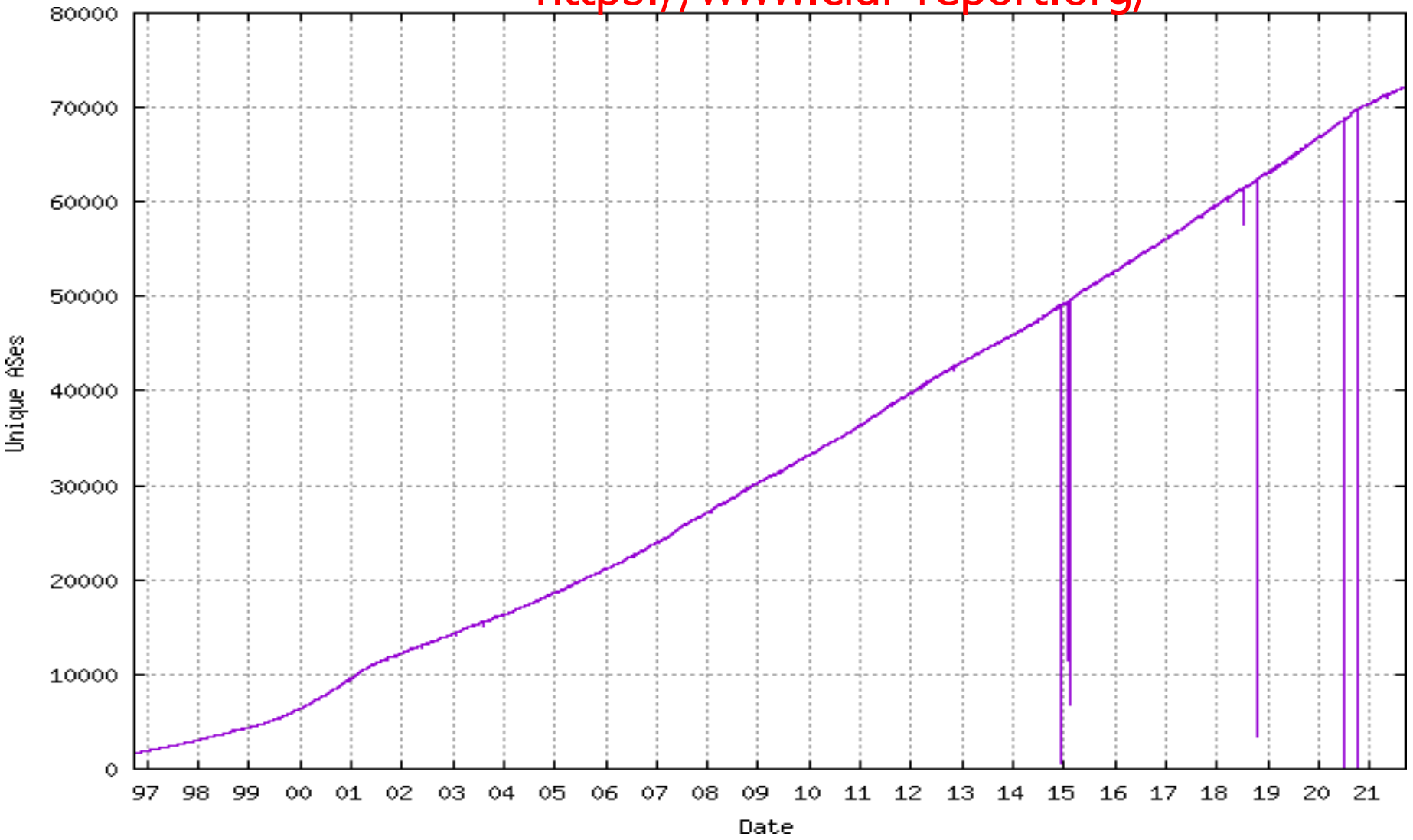
If AS 1 does not announce the more specific prefix, then most traffic to AS 2 will go through AS 3 because it is a longer match



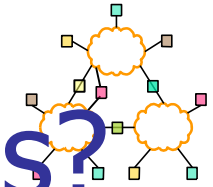
AS 2 is "punching a hole" in The CIDR block of AS 1

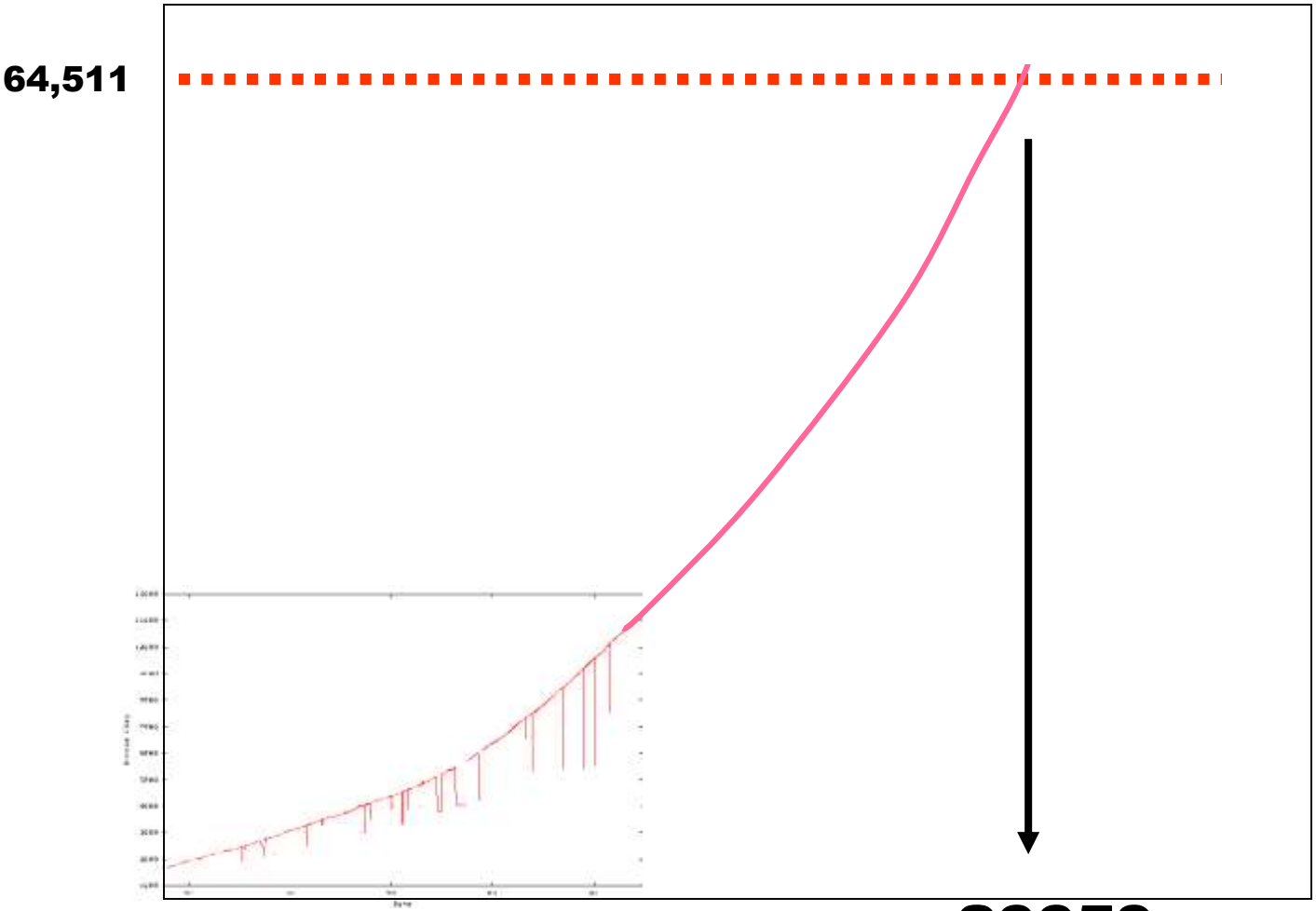
AS Growth with Time

https://www.cidr-report.org/

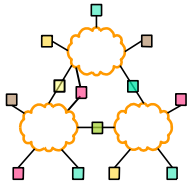


When will we run out of ASNs?



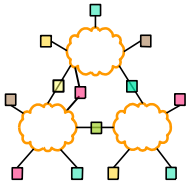


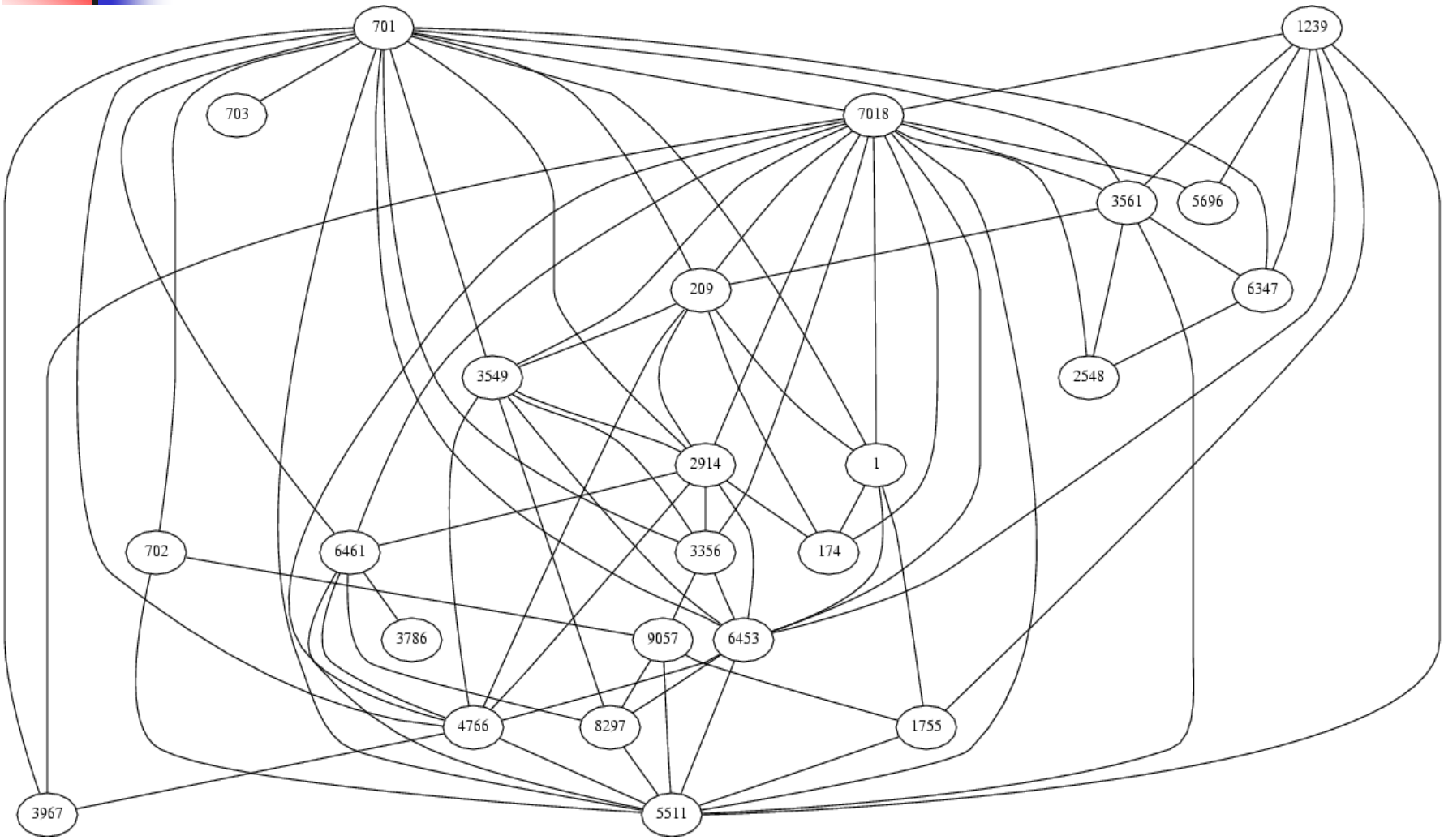
What is to be done?



- Make ASNs larger than 16 bits
  - How about 32 bits?
  - See Internet Draft: "BGP support for four-octet AS number space" (draft-ietf-idr-as4bytes-03.txt)
  - Requires protocol change and wide deployment
- Change the way ASNs are used
  - Allow multihomed, non-transit networks to use private ASNs
  - Uses ASE (AS number Substitution on Egress)
  - See Internet Draft: "Autonomous System Number Substitution on Egress" (draft-jhaas-ase-00.txt)
  - Works at edge, requires protocol change (for loop prevention)
  - Makes some kinds of debugging harder!

AS Graphs Can Be Fun

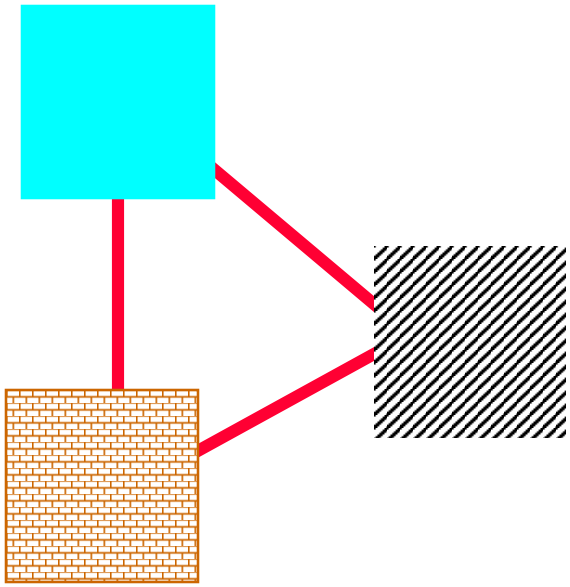




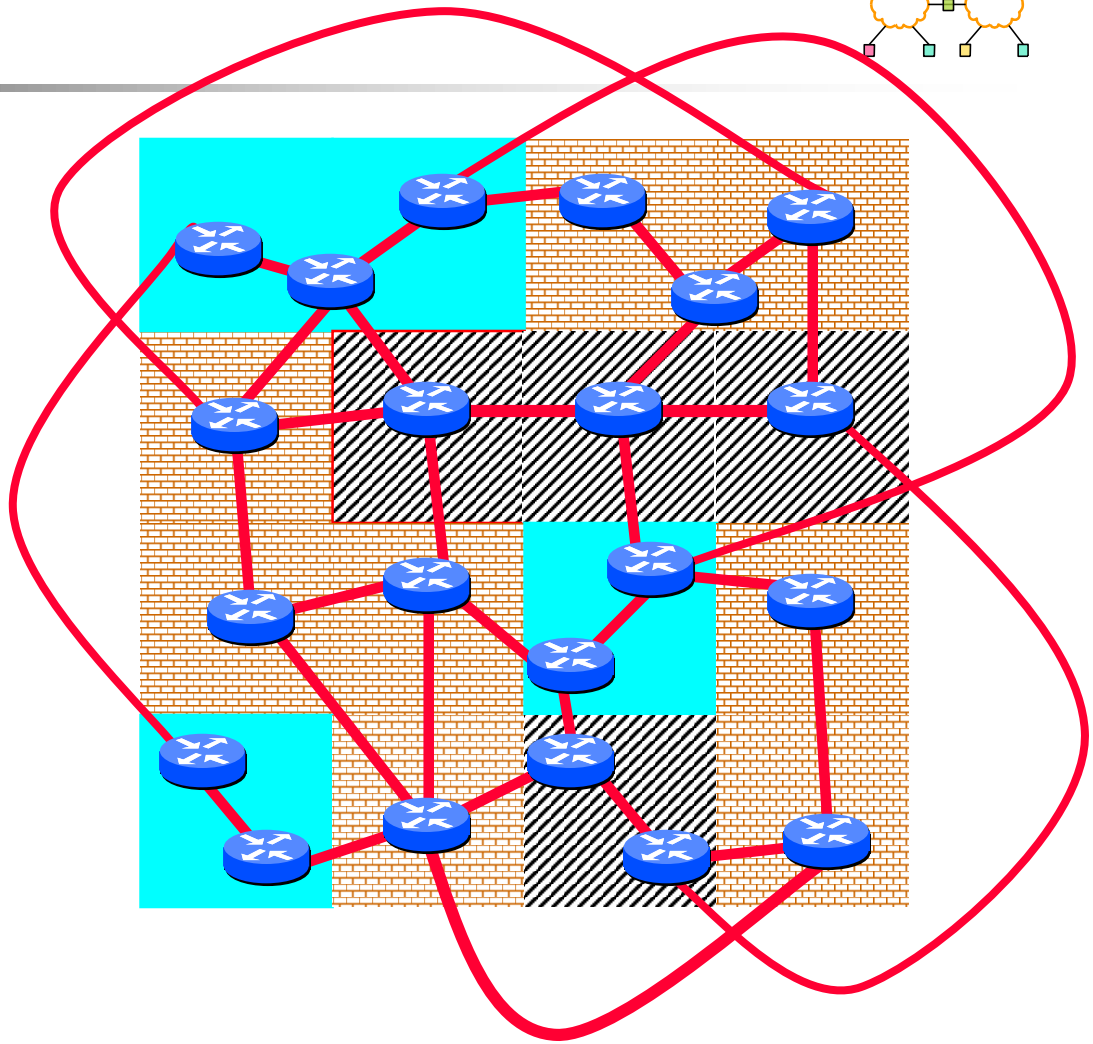
The subgraph showing all ASes that have more than 100 neighbors in full graph of 11,158 nodes. July 6, 2001. Point of view: AT&T route-server

AS Graphs Do Not Show Topology!

BGP was designed to throw away information!

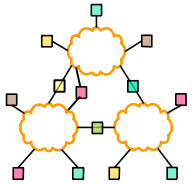


The AS graph may look like this.



Reality may be closer to this…

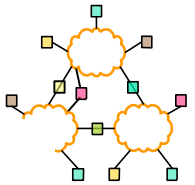
BGP Dynamics

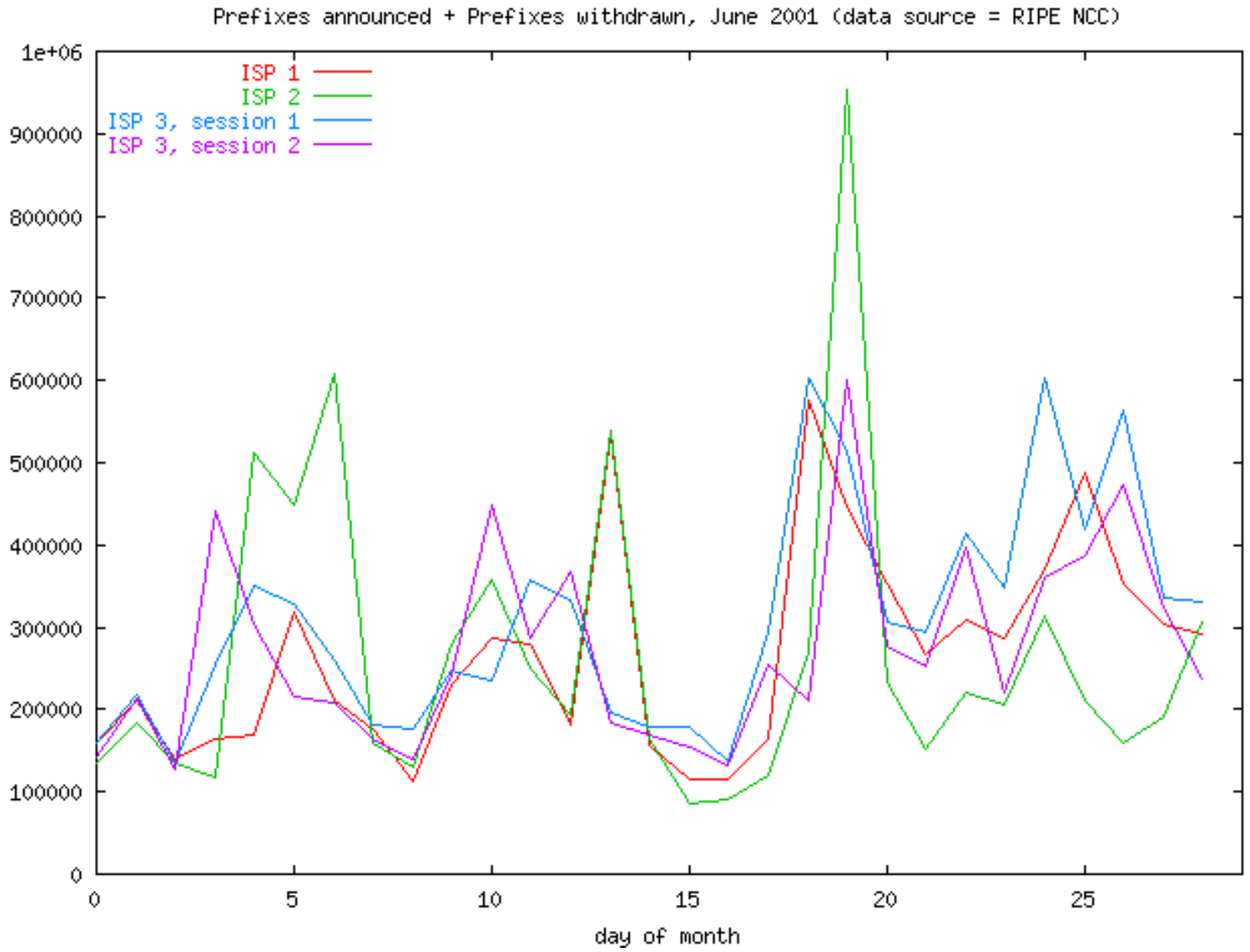


- How many updates are flying around the Internet?
- How long Does it take Routes to Change?

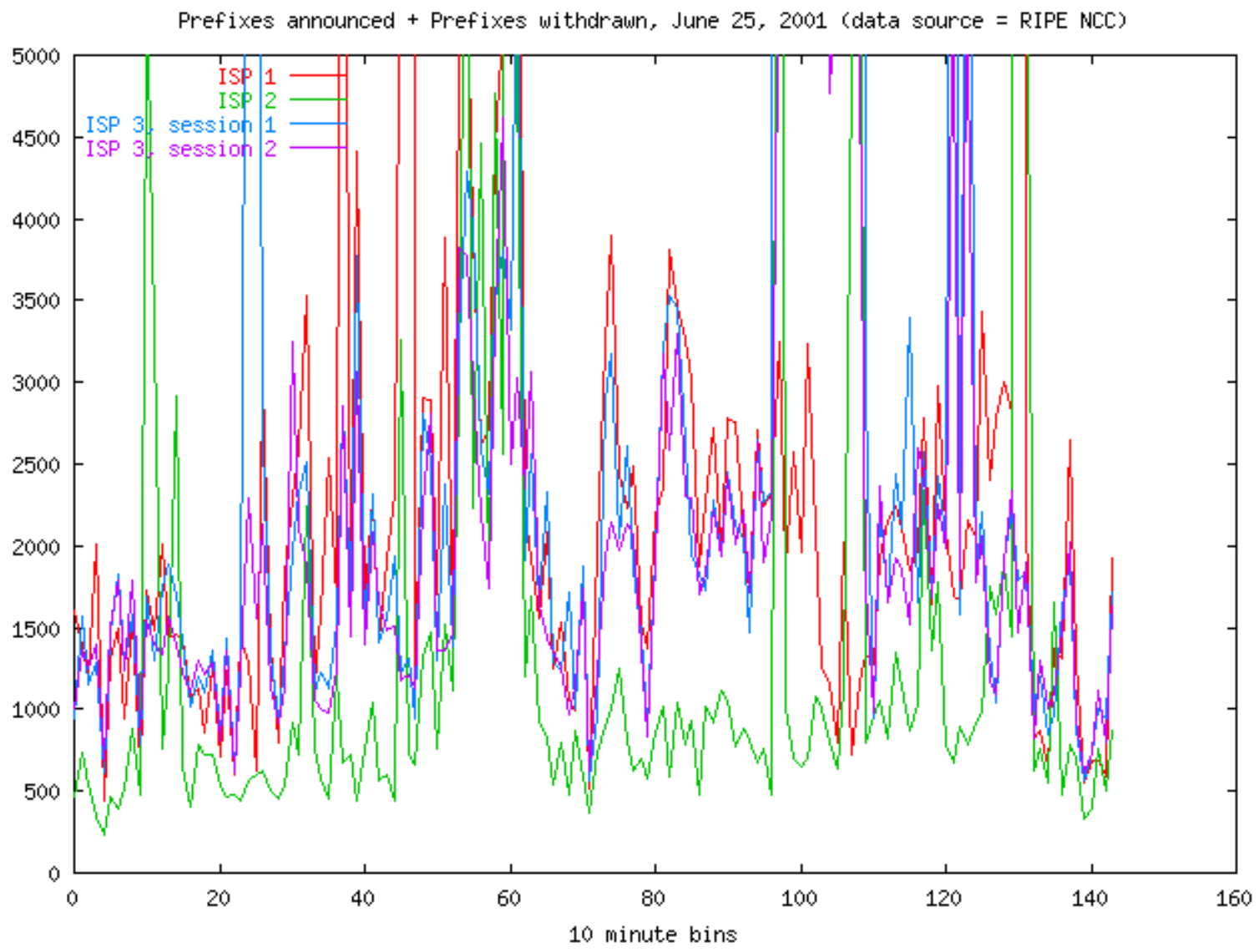
The goals of
(1) fast convergence
(2) minimal updates
(3) path redundancy are at odds

Daily Update Count

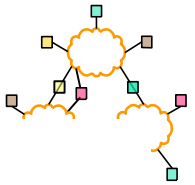


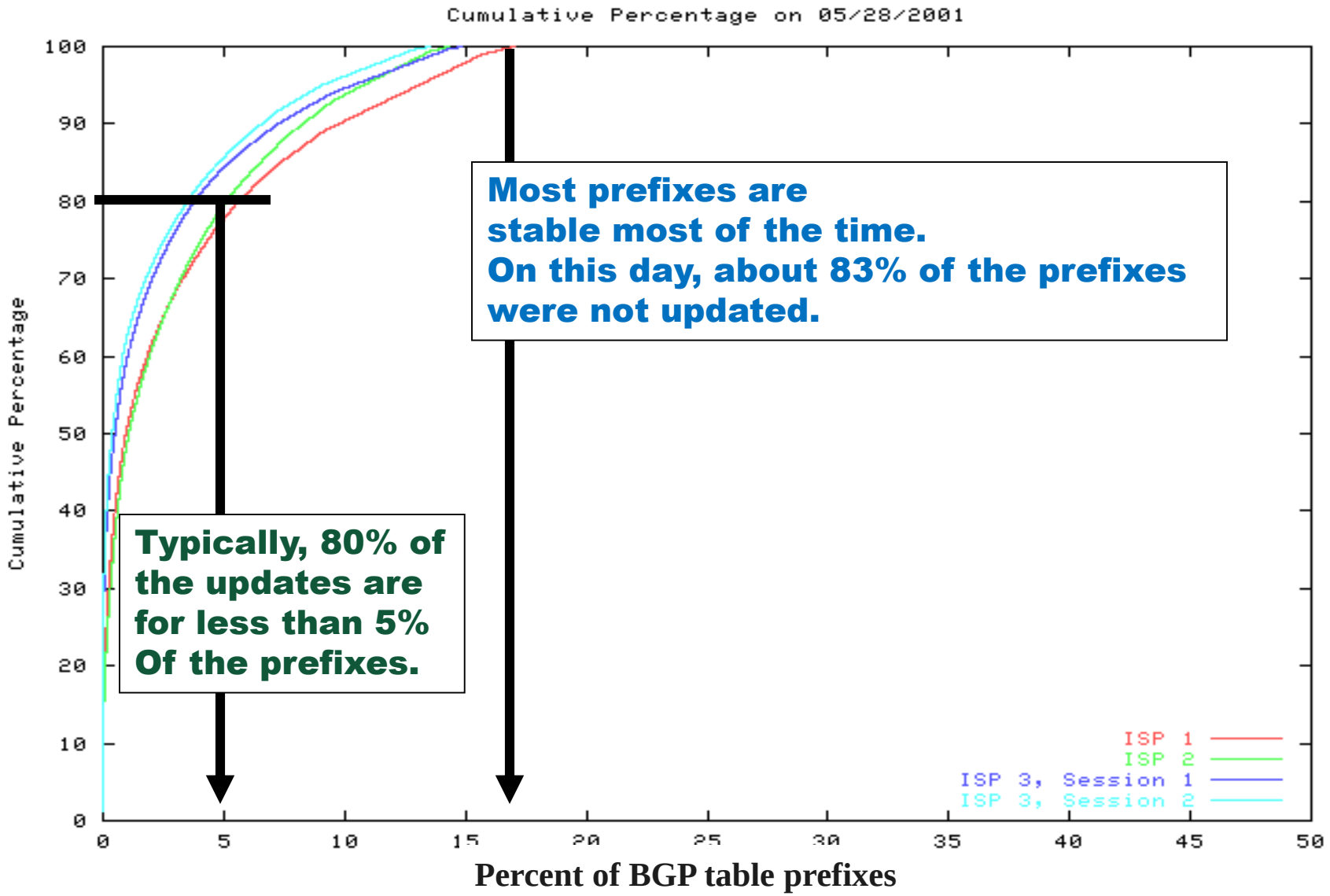


What is the Sound of One Route Flapping?



A Few Bad Apples …

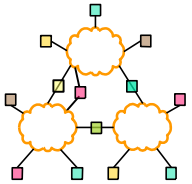




Thanks to Madanlal Musuvathi for this plot.

Data source: RIPE NCC

Two BGP Mechanisms for Squashing Updates

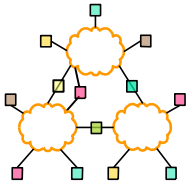


- Rate limiting on sending updates
  - Send batch of updates every MinRouteAdvertisementInterval seconds (+/- random fuzz)
  - Default value is 30 seconds
  - A router can change its mind about best routes many times within this interval without telling neighbors
- Route Flap Dampening
  - Punish routes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

Effective in dampening oscillations inherent in the vectoring approach

Must be turned on with configuration

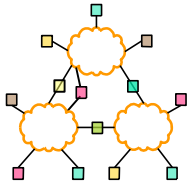
Q: Why All the Updates?



- Networks come, networks go
- There's always a router rebooting somewhere
- Hardware failure, flaky interface cards, backhoes digging, floods in Houston, …

This is "normal" --- exactly what dynamic routing is designed for…

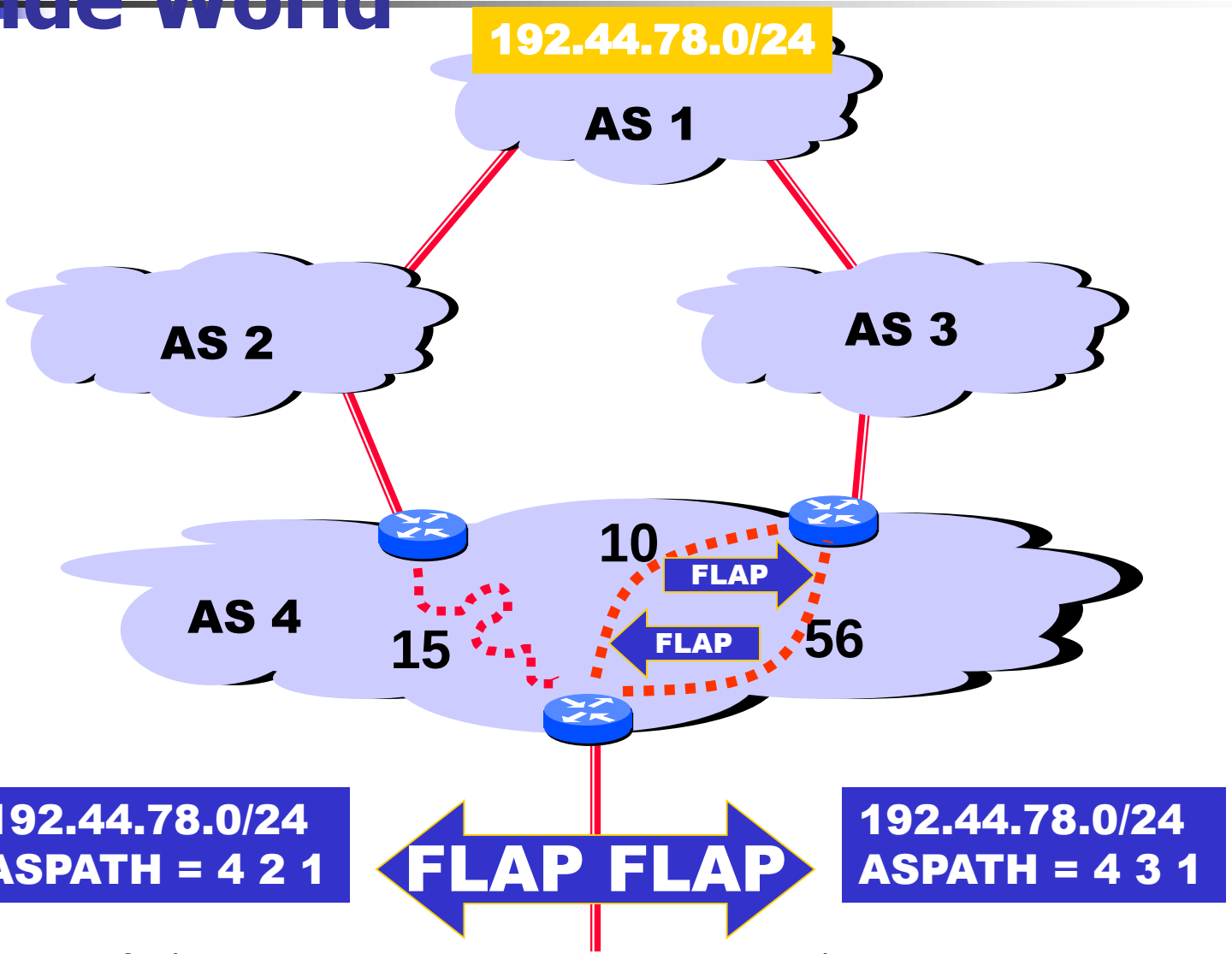
Q: Why All the Updates?



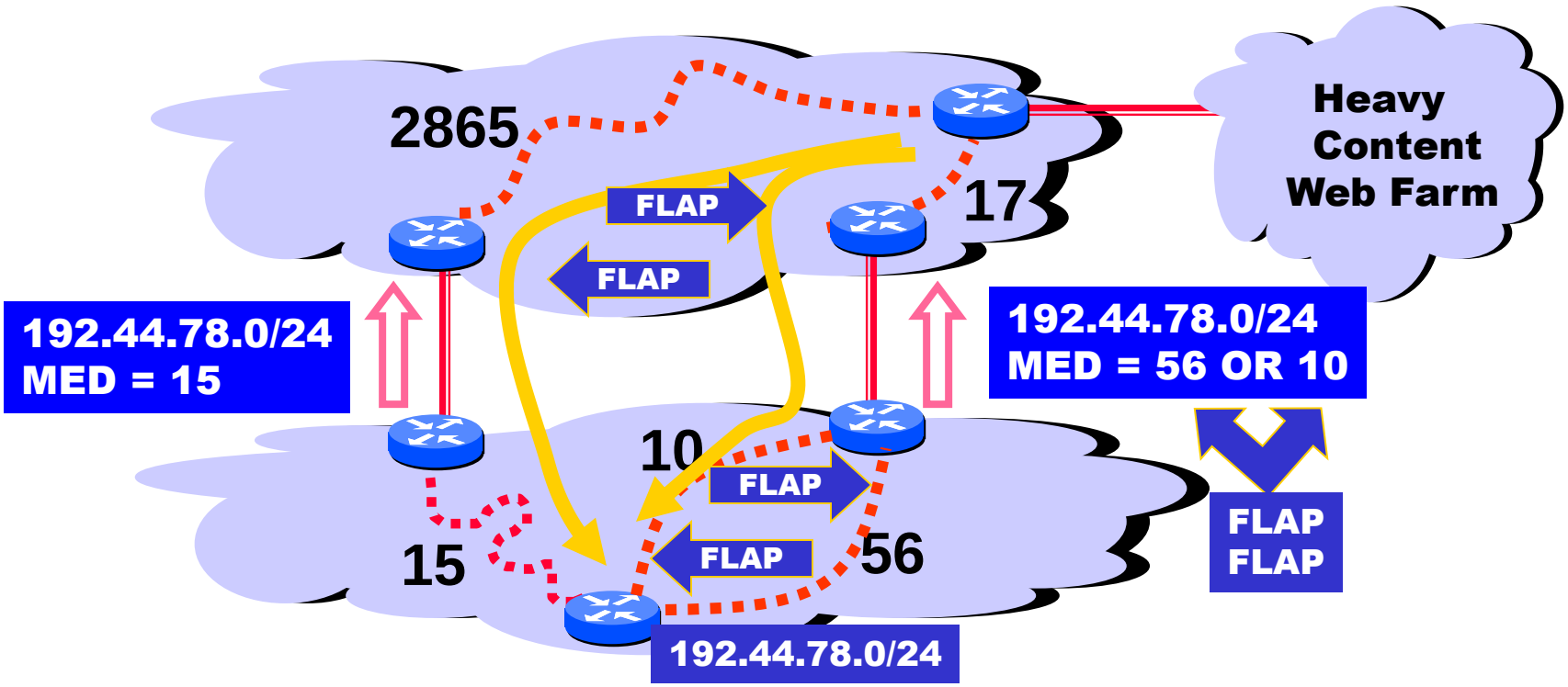
- Misconfiguration
- Route flap dampening not widely used
- BGP exploring many alternate paths
- Software bugs in implementation of routing protocols
- BGP session resets due to congestion or lack of interoperability: BGP sessions are brittle. One malformed update is enough to reset session and flap 100K routes. (Consequence of incremental approach)
- IGP instability exported by use of MEDs or IGP tie breaker
- Sub-optimal vendor implementation choices
- Secret sauce routing algorithms attempting fancy-dancy tricks
- Weird policy interactions (MED oscillation, BAD GADGETS??)
- Gnomes, sprites, and fairies
- ….

A: NO ONE REALLY KNOWS …

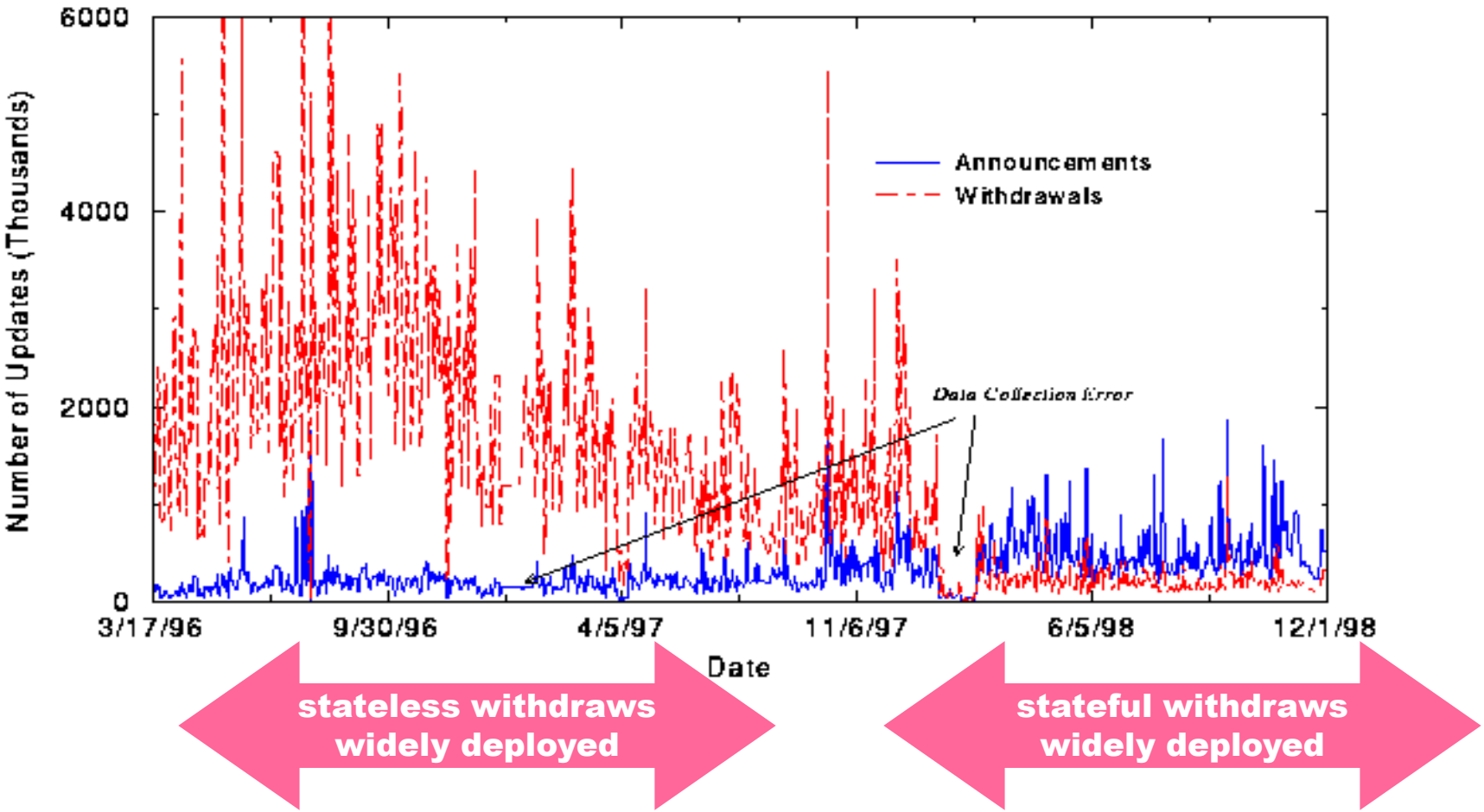
IGP Tie Breaking Can Export Internal Instability to the Whole Wide World



MEDs Can Export Internal Instability

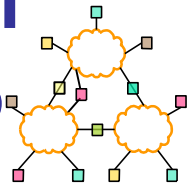


Implementation Does Matter!



Thanks to Abha Ahuja and Craig Labovitz for this plot.

How Long Will Interdomain Routing Continue to Scale?



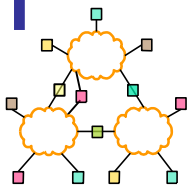
A quote from some recent email:

> ... the existing interdomain routing infrastructure is rapidly nearing the end of its useful lifetime. It appears unlikely that mere tweaks of BGP will stave off fundamental scaling issues, brought on by growth, multihoming and other causes.

Is this true or false? How can we tell?

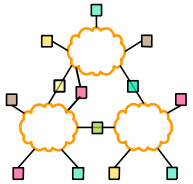
Research required...

HLP: Hybrid Link state path vector routing



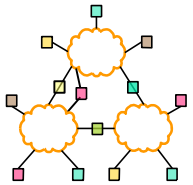
- Conflict between full path information and private policy.
- BGP problems:
  - Rapid grows (from 3000 AS to 17000, 1997-now) Scalability
  - Route Oscillations, %25 of prefixes flap and have 10 hours convergence problem
  - Poor fault isolation: most of routing event are globally visible

HLP: Solutions



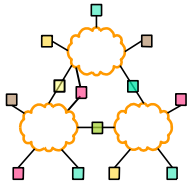
- Use two types of routing, Link state for hierarchy and path for peering or Fragmented Path Vector (FPV)
- Explicit information hiding
- Model complex relations as peering
- Policy variation as exceptions
- Result: Reduction of the size and domain of announcements.

Foundational Problems



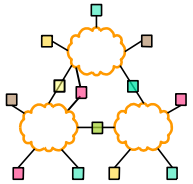
- Two complexity in BGP
  - Policy: Cause persistent oscillation, loops, slow convergence and network partitioning
  - Scalability:
- Two basic questions:
  - Which ones rooted in BGP design?
  - Which ones are fundamental limitation & challenging of scalable policy-based routing?

Policy-induced Problems



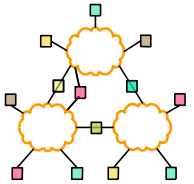
- Protocol oscillation: Stem from inability to satisfy group preference
  - Inter-As oscillation: Caused by policy dispute
    - Finding policy dispute is NP-complete
    - Even with stable input BGP might never converge.
  - Intera-As oscillation: Caused by non-monotonic ranking. MED from different Ases.
- Is it possible to have a protocol that always converges? Yes

Convergence Problems



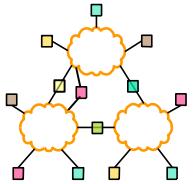
- Basic questions
  - Policy restriction: What restriction on policy such that with no topology and export policy it always converge?
  - Protocol change: With additional features and flexible policy guaranteeing convergence?
  - Total ordering : Are a set of preferences without pairwise ranking but always converge?

Security



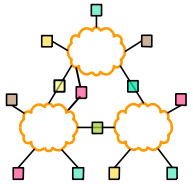
- BGP does not prevent an AS from advertising arbitrary perfixes
  - Works to determine out of band prefix ownerships.
- Is it possible to have
  - Decentralized verification?
  - In-band verification?
  - Detecting routes that violate polices?
  - Verifying the forwarding pathes?

Scaling in BGP

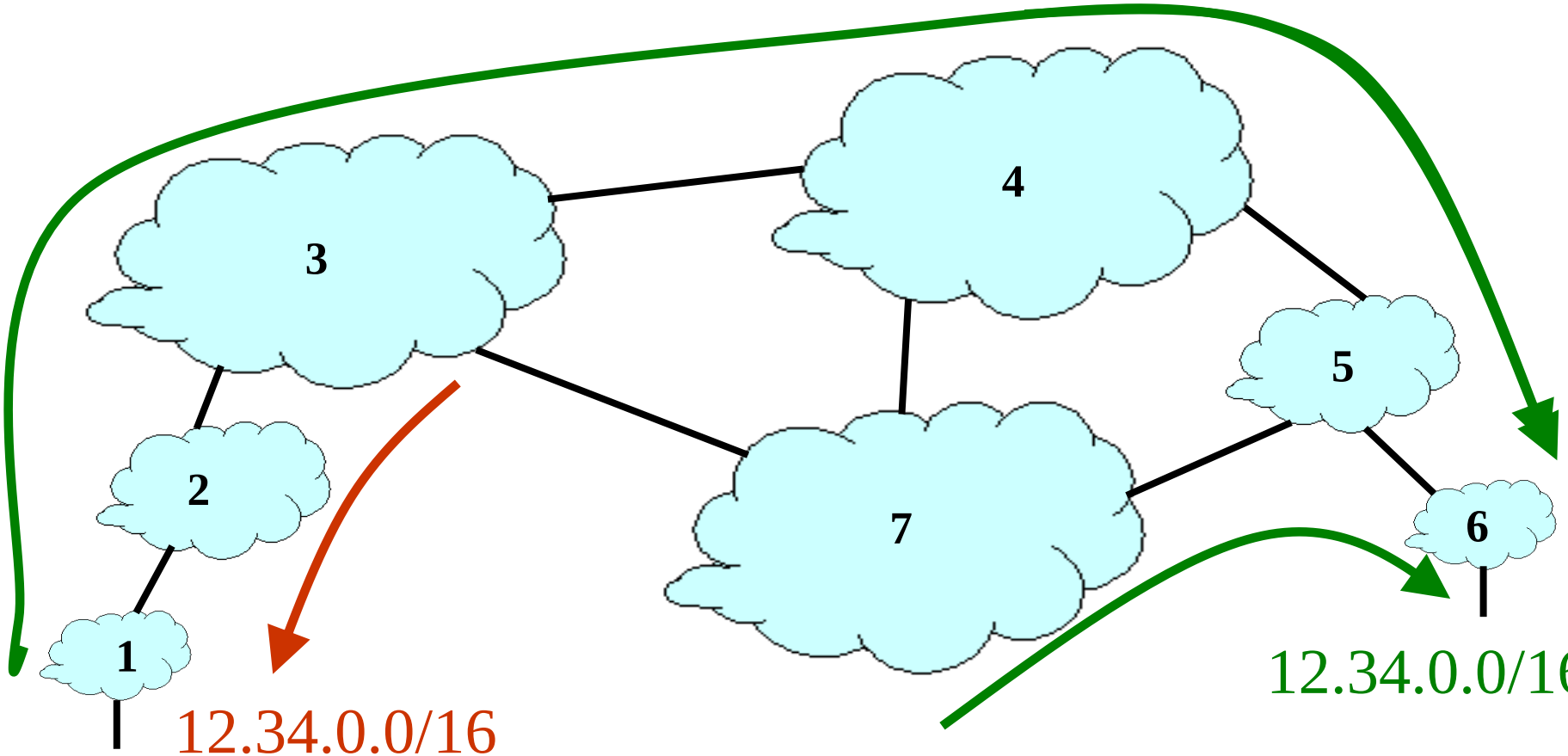


- Different scaling techniques
  - An AS as a single node.
  - Route reflection
  - Prefixes aggregation
- But all hide routing information which makes problem diagnosis difficult and sometimes impossible.
- There is a trade off between scalability and convergence.

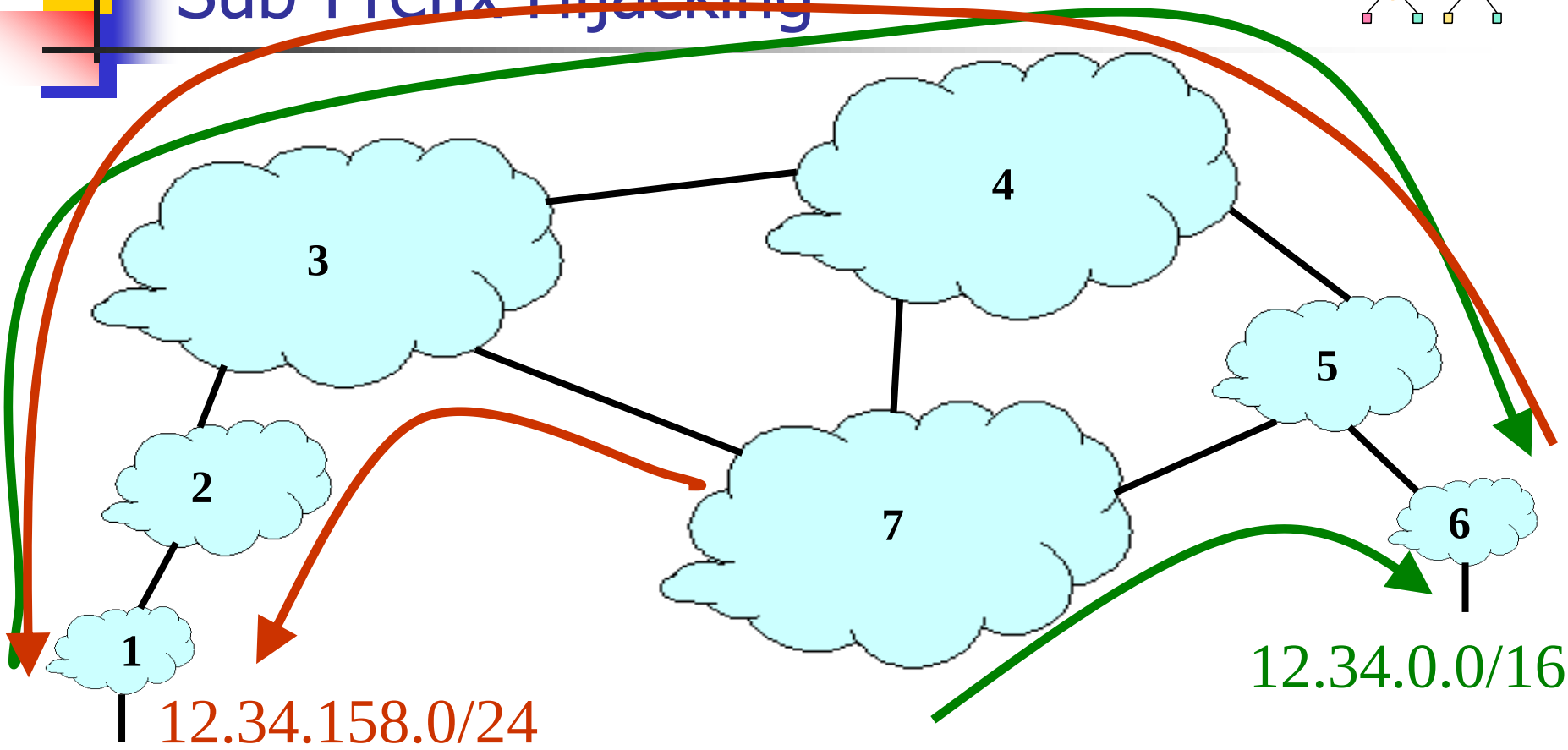
Prefix Hijacking



- Originating someone else's prefix
  - What fraction of the Internet believes it?

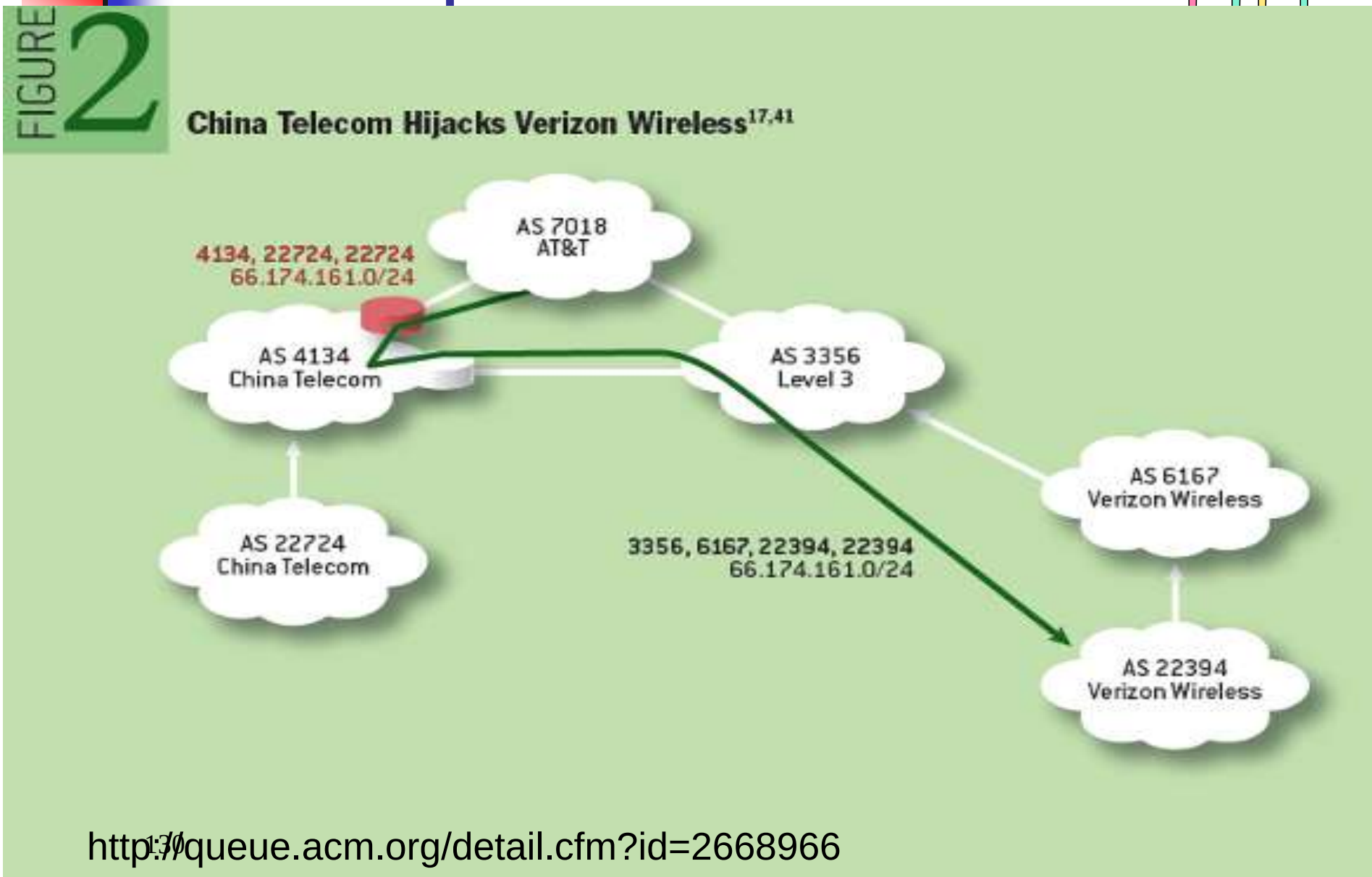


Sub-Prefix Hijacking



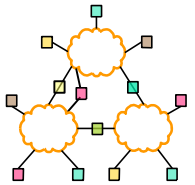
- Originating a more-specific prefix
  - Every AS picks the bogus route for that prefix
  - Traffic follows the longest matching prefix

Interception Attack

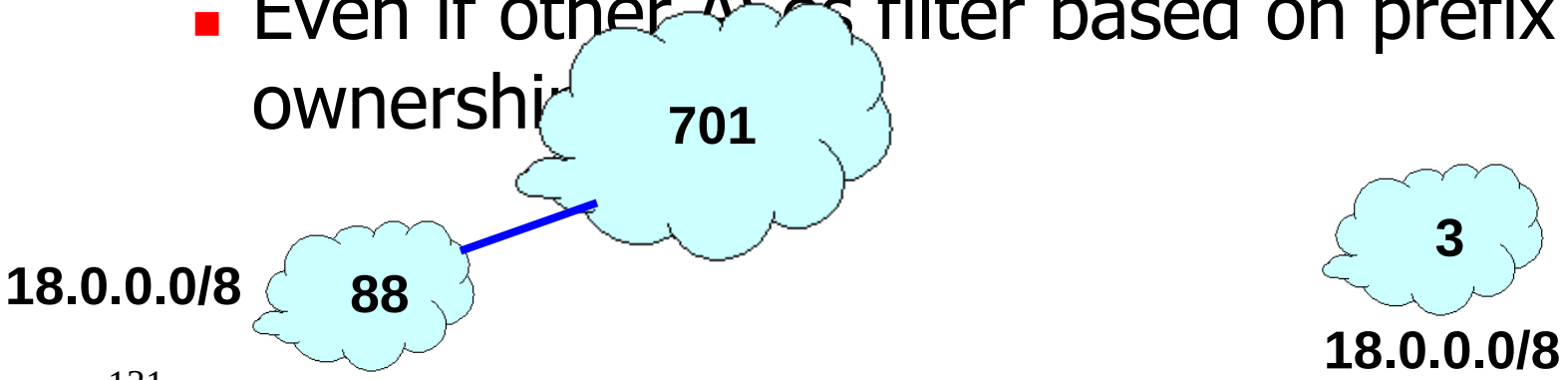


http://queue.acm.org/detail.cfm?id=2668966

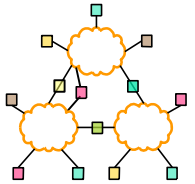
Bogus AS Paths to Hide Hijacking



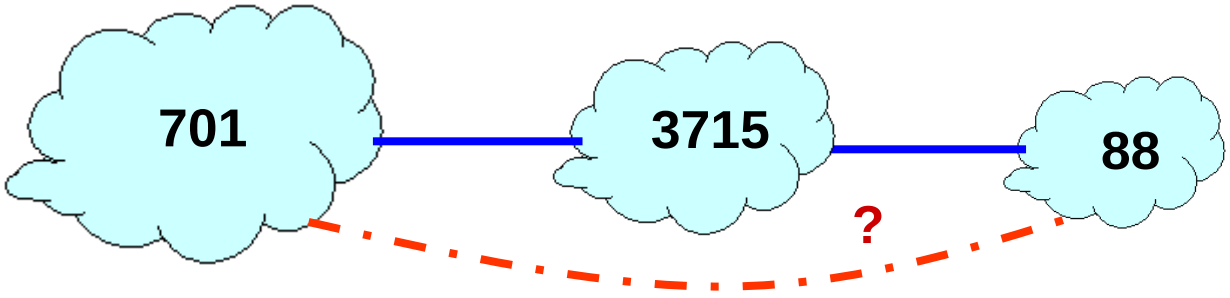
- Adds AS hop(s) at the end of the path
  - E.g., turns “701 88” into “701 88 3”
- Motivations
  - Evade detection for a bogus route
  - E.g., by adding the legitimate AS to the end
- Hard to tell that the AS path is bogus…
  - Even if other ASes filter based on prefix ownershi



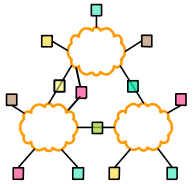
Path-Shortening Attacks



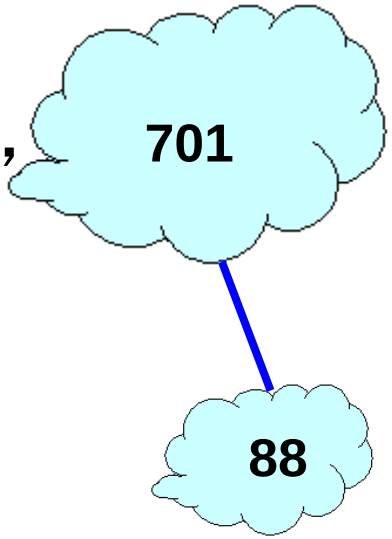
- Remove ASes from the AS path
  - E.g., turn “701 3715 88” into “701 88”
- Motivations
  - Make the AS path look shorter than it is
  - Attract sources that normally try to avoid AS 3715
  - Help AS 88 look like it is closer to the Internet’s core
- Who can tell that this AS path is a lie?
  - Maybe AS 88 *does* connect to AS 701 directly



Attacks that Add a Bogus AS Hop



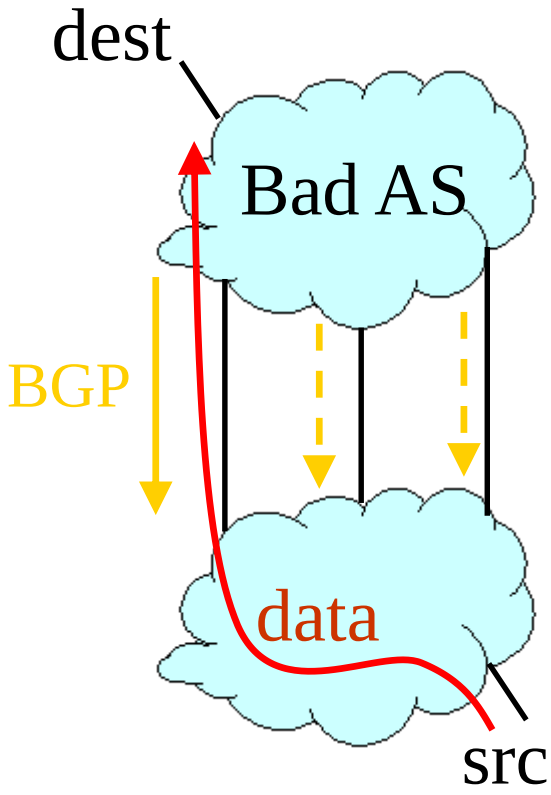
- Add ASes to the path
  - E.g., turn “701 88” into “701 3715 88”



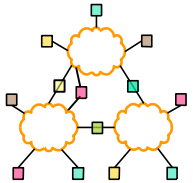
- Motivations
  - Trigger loop detection in AS 3715
    - Denial-of-service attack on AS 3715
    - Or, blocking unwanted traffic coming from AS 3715!
  - Make your AS look like is has richer connectivity
- Who can tell the AS path is a lie?
  - AS 3715 could, if it could see the route
  - AS 88 could, but would it really care as long as it received data traffic meant for it?

Violating "Consistent Export" to Peers

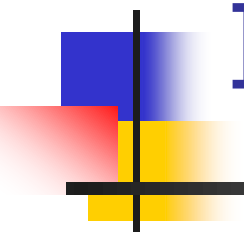
- Peers require consistent export
  - Prefix advertised at all peering points
  - Prefix advertised with same AS path length
- Reasons for violating the policy
  - Trick neighbor into "cold potato"
  - Configuration mistake
- Main defense
  - Analyzing BGP updates
  - … or data traffic
  - … for signs of inconsistency



Other Attacks

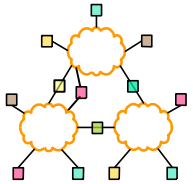


- Attacks on BGP sessions
  - Confidentiality of BGP messages
  - Denial-of-service on BGP session
  - Inserting, deleting, modifying, or replaying messages
- Resource exhaustion attacks
  - Too many IP prefixes (e.g., BGP "512K Day")
  - Too many BGP update messages
- Data-plane attacks
  - Announce one BGP route, but use another



Improving BGP Security

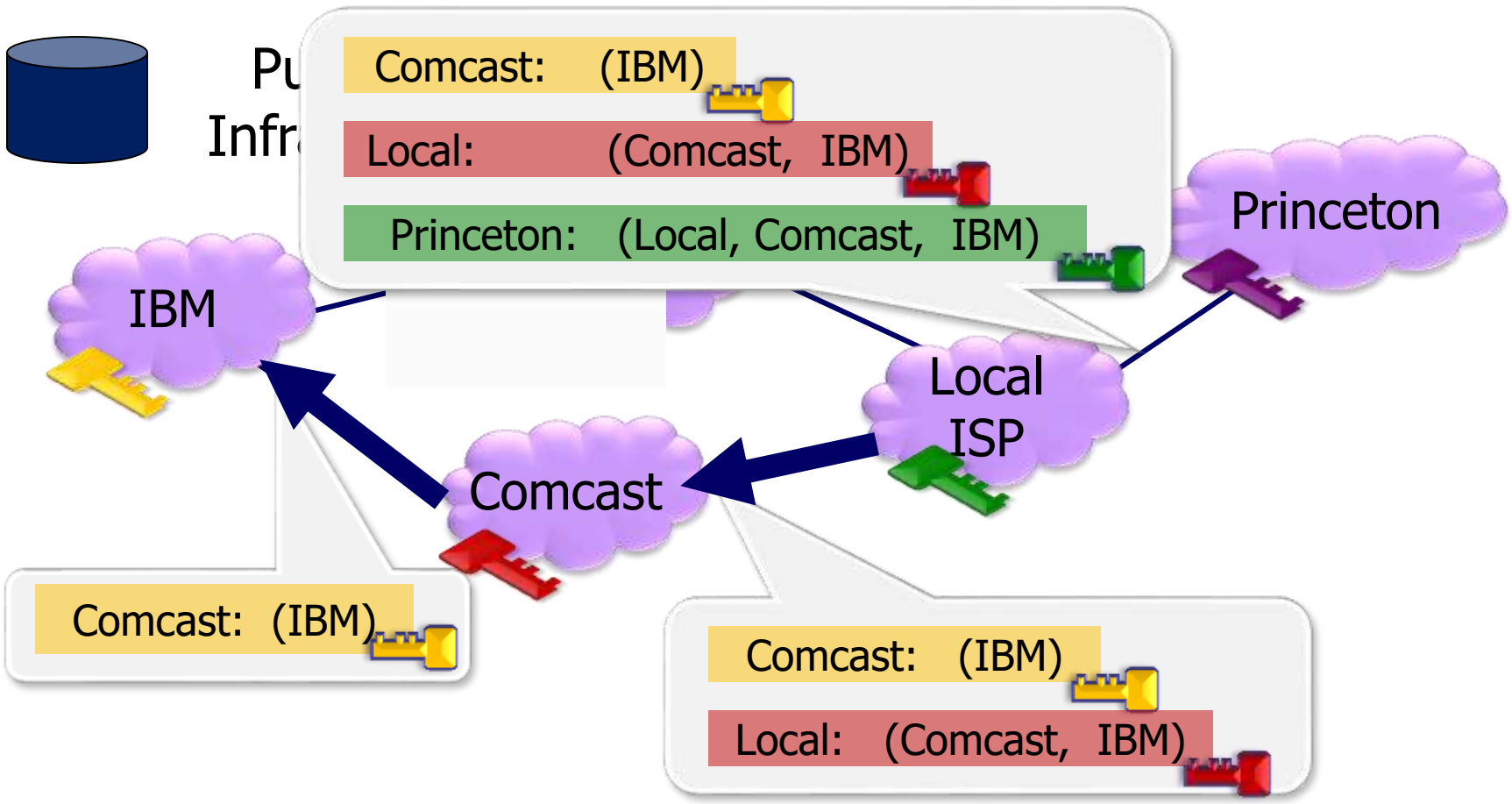
Solution Techniques



- Protective filtering
  - Know your neighbors
- Anomaly detection
  - Suspect the unexpected
- Checking against registries
  - Establish ground truth for prefix origination
- Signing and verifying
  - Prevent bogus AS PATHs
- Data-plane verification
  - Ensure the path is actually followed

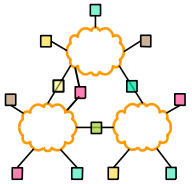
BGP

If AS a announced path abP then b announced bP to a



Public Key Signature: Anyone who knows IBM's public key can verify the message was sent by IBM.

Next Lecture: Routers



- How do you build a router
- References
  - Scaling Internet Routers Using Optics
  - [P+98] A 50 Gb/s IP Router