

CSE 222

Graduate Networking

Fall 2001

Lecture 8: Inter-domain Routing

Stefan Savage

(thanks to Tim Griffin for the use of some slides)

Today

- Inter-domain routing
 - ◆ Problems
 - ◆ BGP & mechanisms
 - ◆ Transit/Peering Policies
- We're only going to scratch the surface here... could do a whole quarter on inter-domain routing

Recap

- Intra-domain routing
 - ◆ Determine where to send packets within a network
 - ◆ Optimize routes to follow *best* path according to some metric
 - ◆ Interior Gateway Protocols (IGPs)
- Options
 - ◆ Distance vector (RIP)
 - » Tell everything to neighbors, maintain shortest paths
 - » Convergence problems
 - ◆ Link state (OSPF, ISIS)
 - » Advertise your neighbors to everyone (flooding), compute shortest path
 - » Lots of state to maintain

Historic context

- Original ARPAnet had single routing protocol
 - ◆ Dynamic DV scheme, replaced with static metric LS algorithm
- New networks came on the scene
 - ◆ NSFnet, CSnet, DDN, etc...
 - ◆ With their own routing protocols (RIP, Hello, ISIS)
 - ◆ And their own rules (e.g. NSF AUP)
- Problem: how to deal with routing heterogeneity?

What to do?

- Some problems
 - ◆ **Consistency:** Network A uses hop count as a metric, Network B uses measured delay, Network C uses link capacity
 - ◆ **Policy:** Network A connects to Networks B and C. Network B is only allowed to carry network C's traffic?
- How would you resolve these problems?

One solution: Inter-domain routing

- Exterior Gateway Protocols (EGPs)
 - ◆ Only exchange **reachability** information (no metrics)
 - ◆ Decide what to do based on local policy
- Autonomous Systems (ASs)
 - ◆ Unit of abstraction in interdomain routing
 - ◆ Roughly, a network with common administrative control, a coherent internal routing policy, and presenting a **consistent** external view of connectivity
 - ◆ Represented by a 16-bit number
 - » Example: UUnet (701), Sprint (1239), UCSD (7377)
 - ◆ Run IGPs within an AS, EGPs between ASs

First attempt

- Protocol called EGP (can be confusing)
 - ◆ Connected NSFnet Backbone to regional networks, DDN/Milnet, etc..
 - ◆ EGP only provided reachability information (no metrics)
 - ◆ Assumed spanning tree topology based on single backbone
 - » No loops
- In 1995 NSFnet got out of the backbone business
 - ◆ Many backbones (MCI, Sprint, AT&T...)
 - ◆ Multiconnected regional networks
 - ◆ Meshed topology, loops...
- Need a new protocol

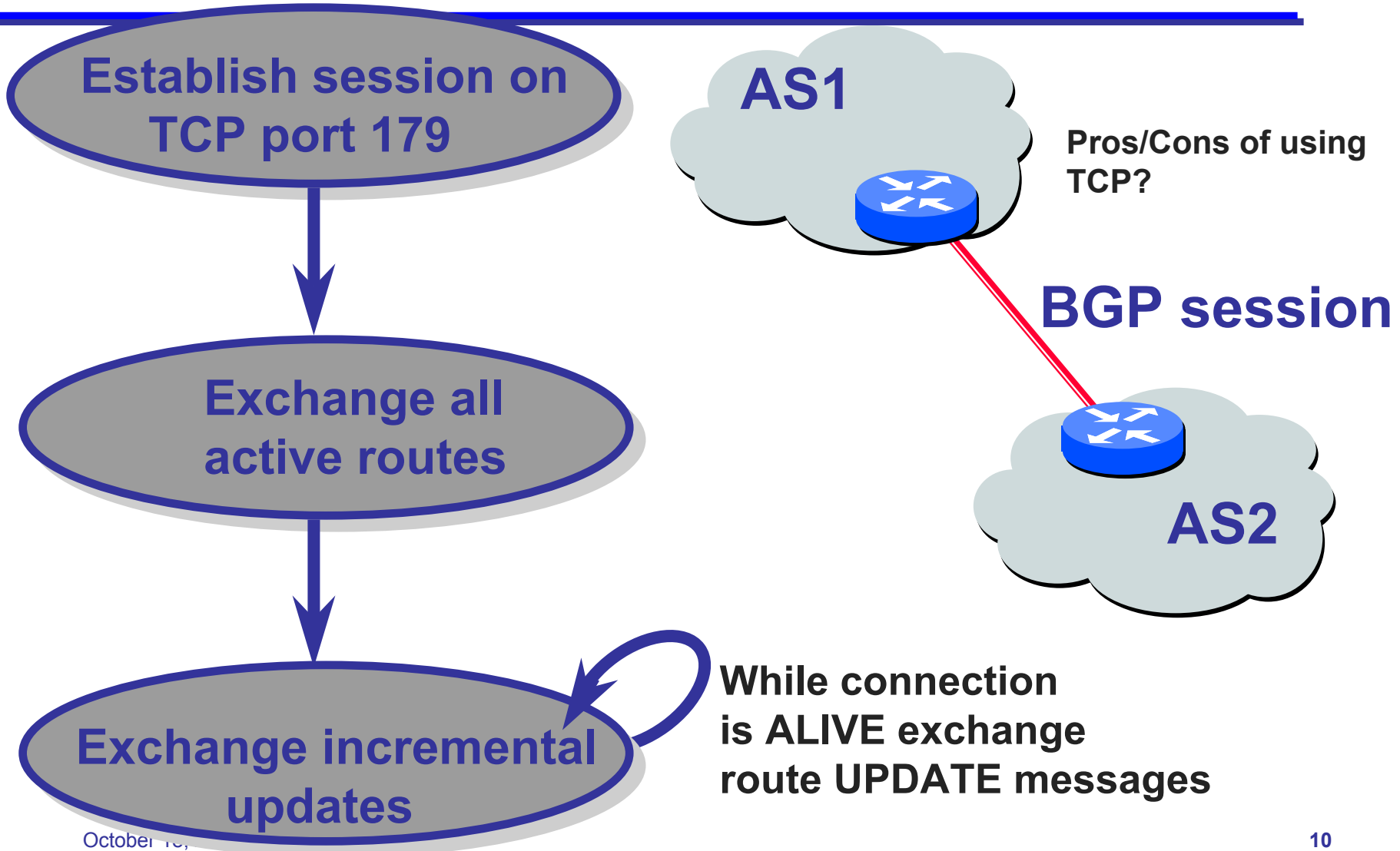
What kind of protocol?

- Link state?
 - ◆ Too much state
 - » Currently 11,000 ASs and > 100,000 networks
 - ◆ Relies on global metric & policy
- Distance vector?
 - ◆ May not converge; loops
- Solution: path vector
 - ◆ Reachability protocol, no metrics
 - ◆ Route advertisements carry list of ASs
 - » “I can reach 128.95/16 through this path: AS73, AS703, AS1”
 - » Automatic loop detection? How?

Border Gateway Protocol

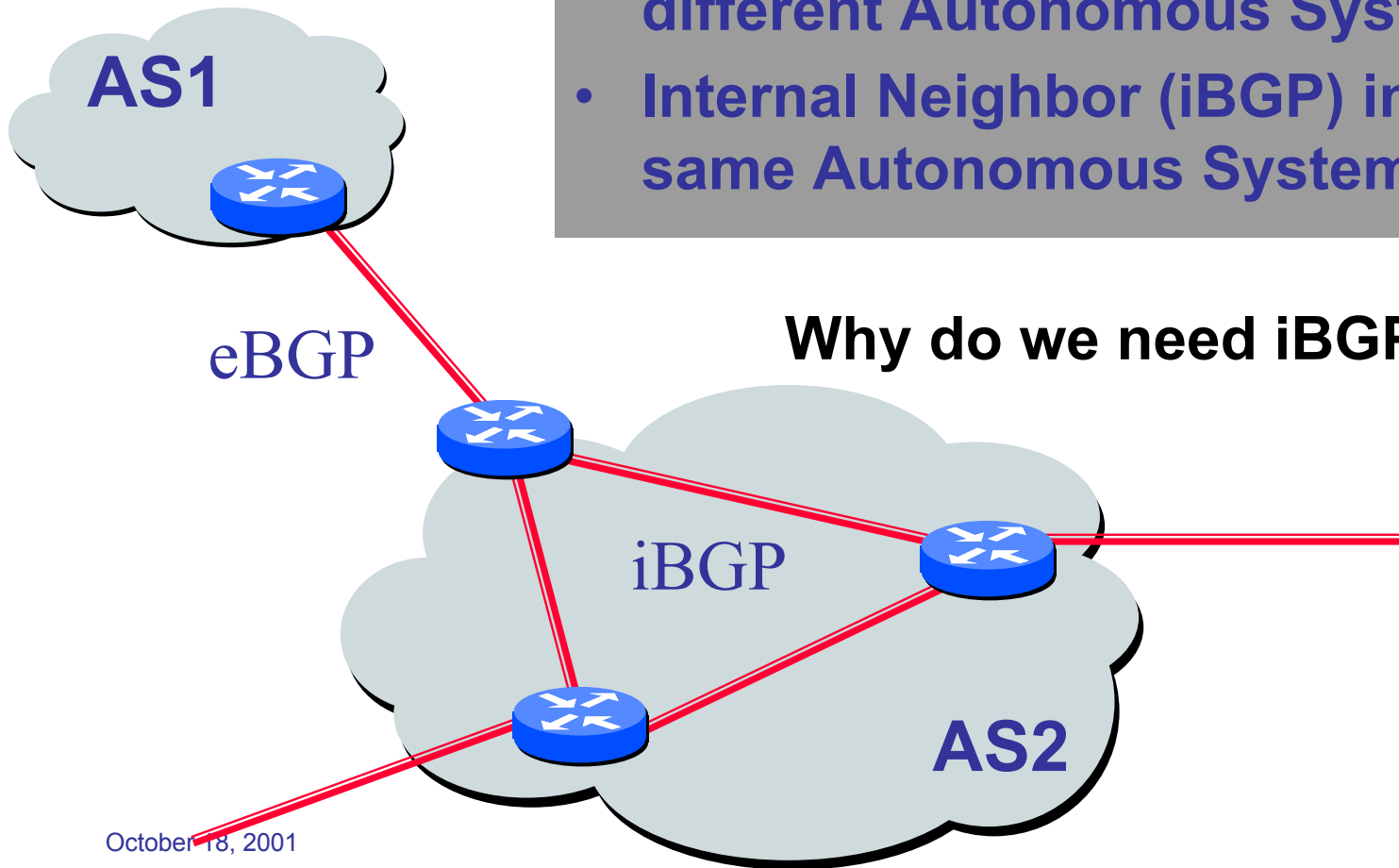
- Principal protocol used for routing across the Internet
 - ◆ Relatively simple protocol, complex usage
- Path vector protocol
 - ◆ Explicitly announce or withdraw routes
 - ◆ Routes include **attributes** in addition to path vector
 - ◆ Incremental updates (stateful)
- Policy is not part of protocol, but is built on top by filtering/mapping on attributes
 - ◆ Which routes do you listen to?
 - ◆ Which routes do you put in forwarding table?
 - ◆ Which routes do you advertise?

How BGP operates (roughly)

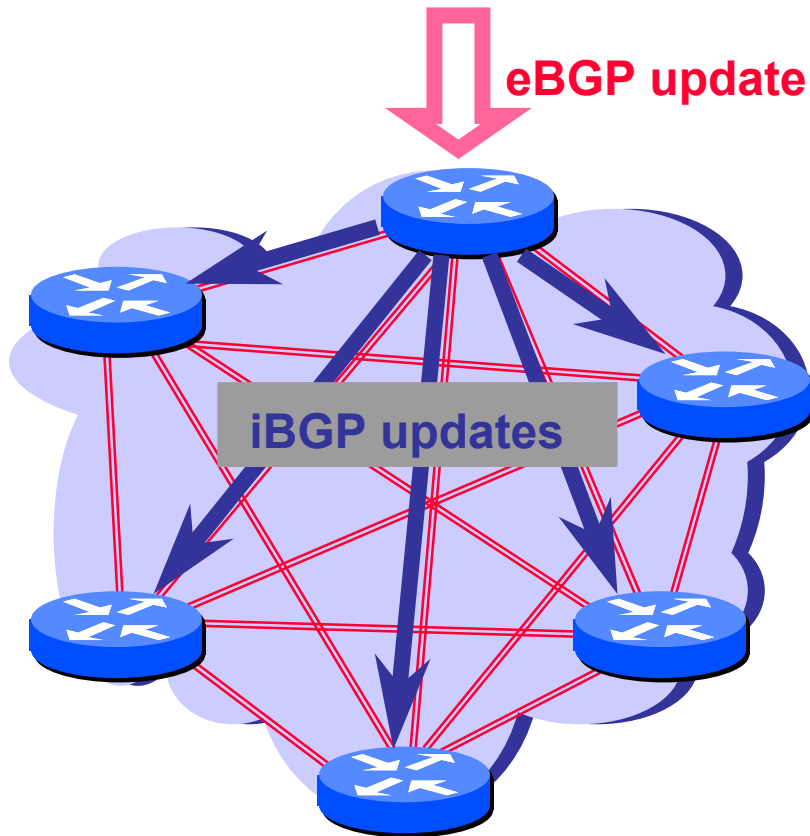


Two types of BGP neighbor relationships

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



iBGP keeps eBGP consistent



- **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**

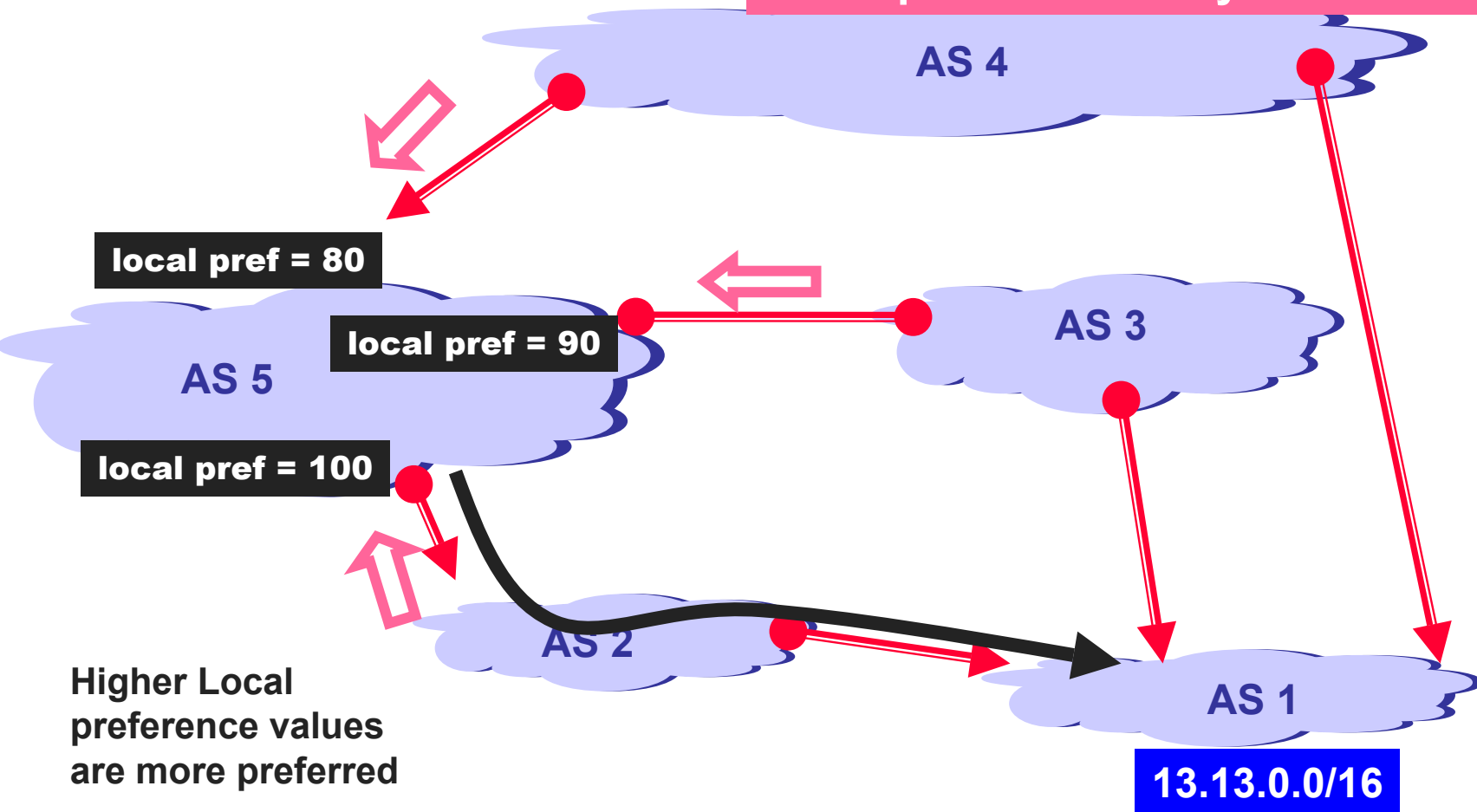
iBGP neighbors do not announce routes received via iBGP to other iBGP neighbors.

Important BGP attributes

- **Local pref:** Statically configured ranking of routes within AS
- **AS path:** ASNs the announcement traversed
- **Origin:** Route came from IGP or EGP
- **Multi Exit Discriminator:** preference for where to exit
- **Community:** opaque data used for inter-ISP policy
- **Next-hop:** where the route was heard from

Example: local pref

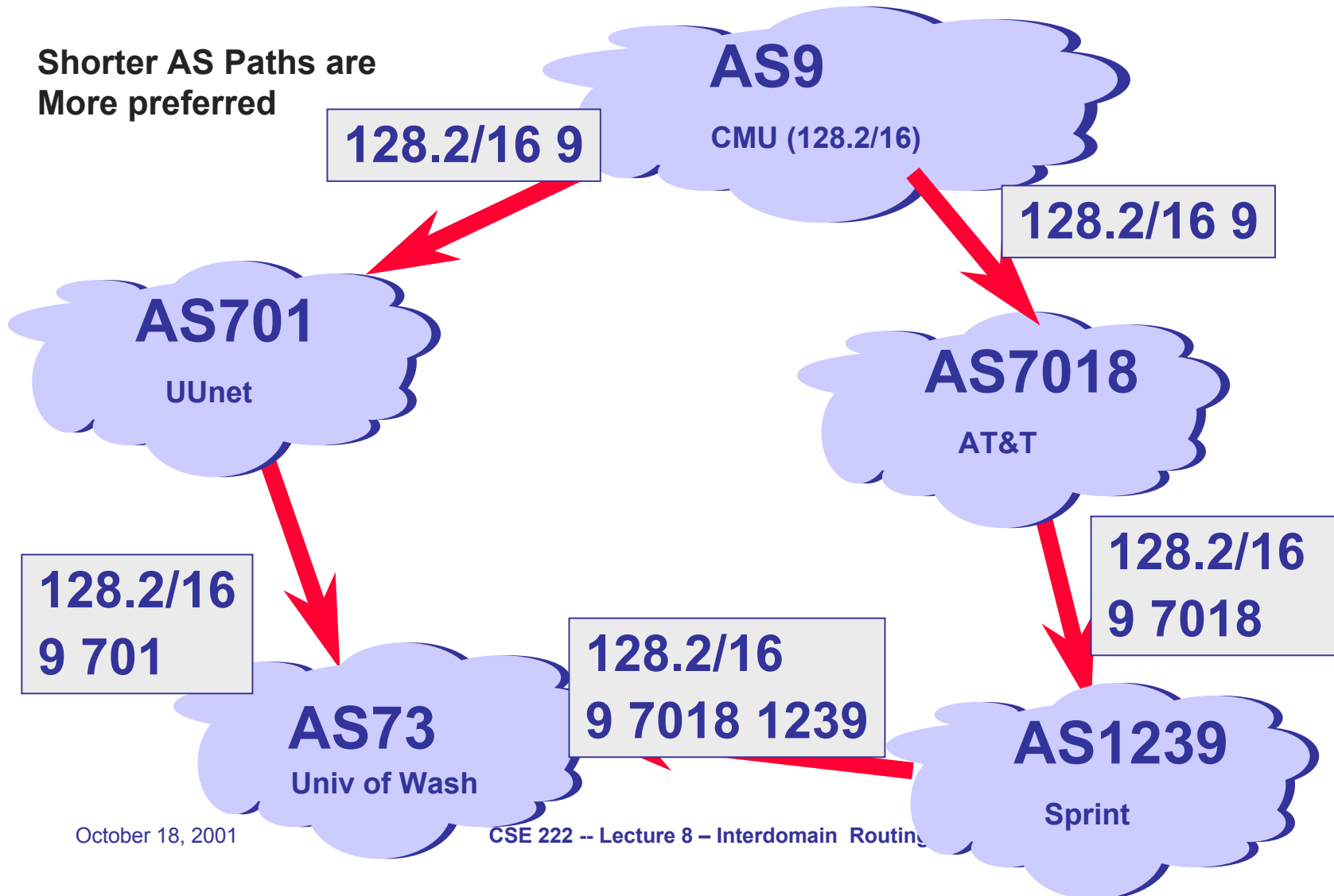
Local preference only used in iBGP



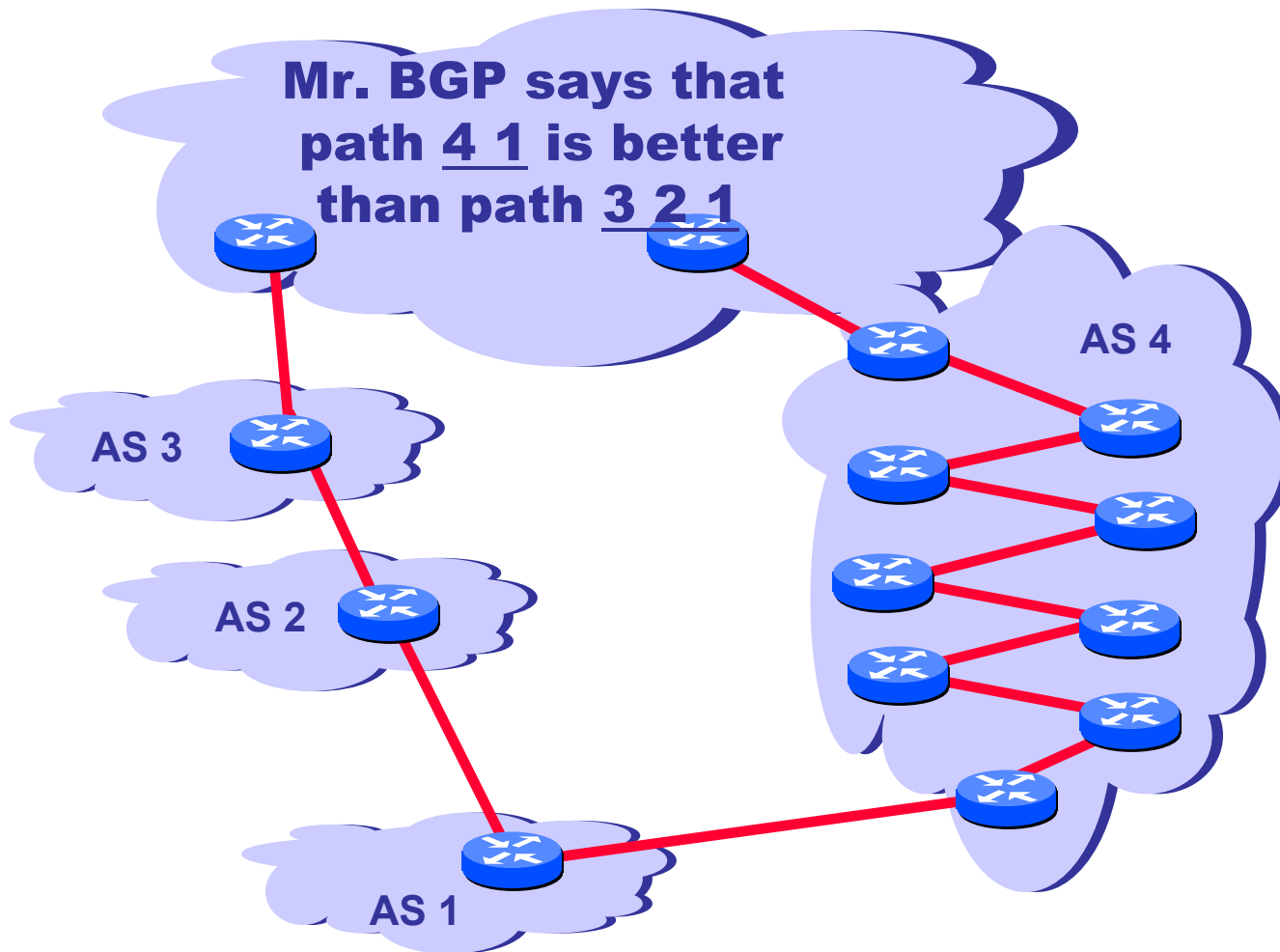
Higher Local preference values are more preferred

Example: AS Path

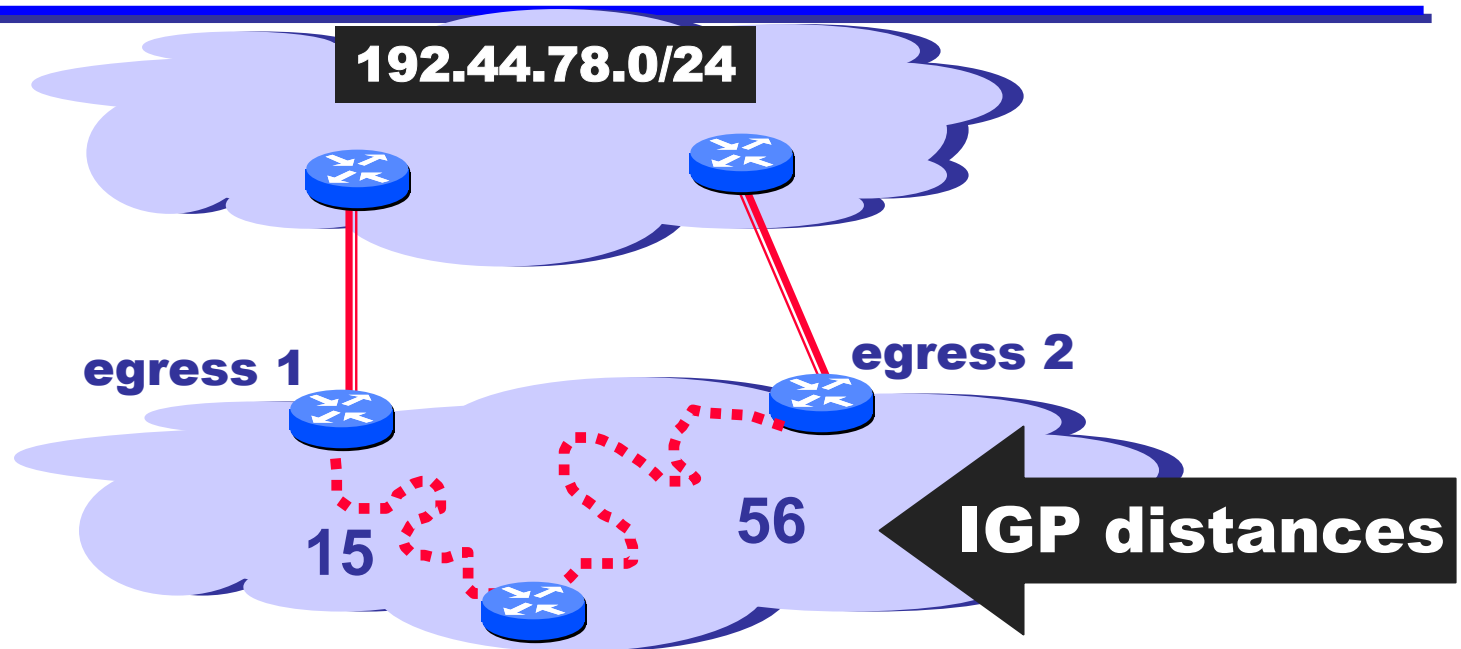
Shorter AS Paths are
More preferred



Shortest AS path doesn't mean best path



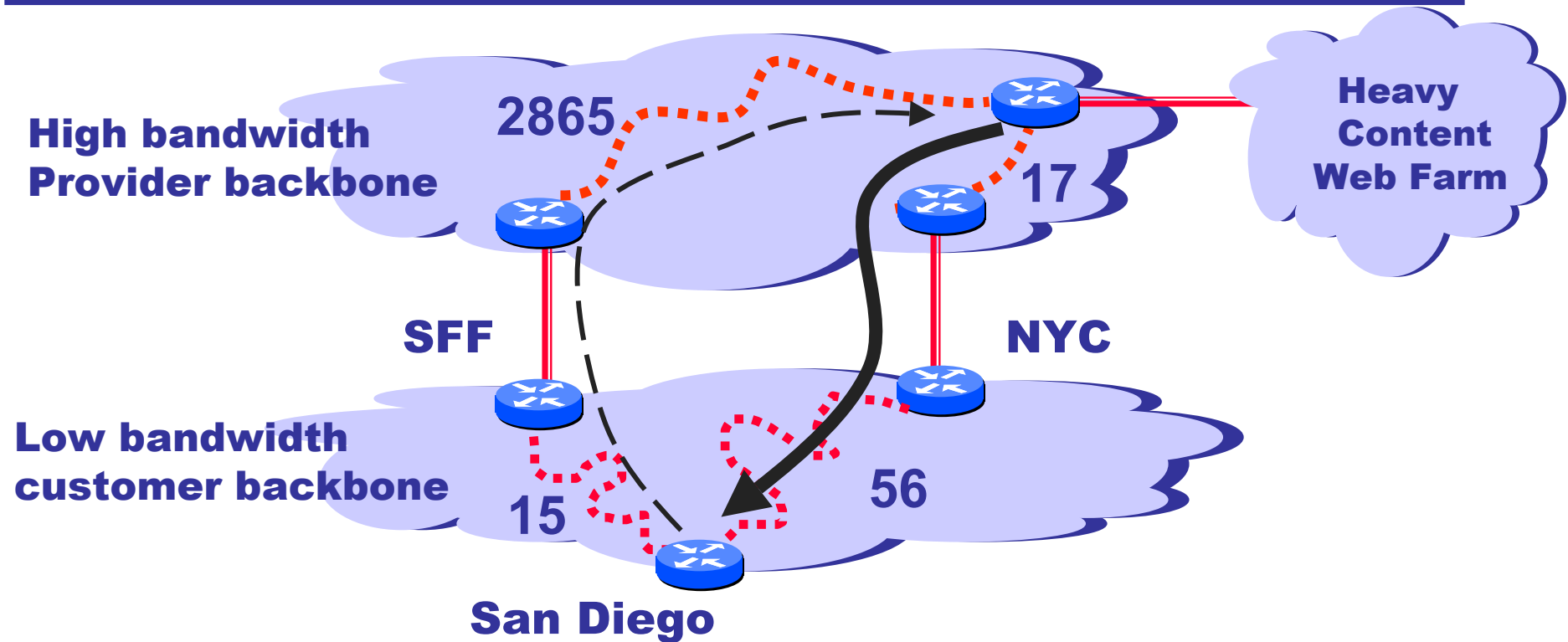
Example: Using IGP cost for Hot potato routing



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!

Problems with hot potato



Many customers want their provider to carry the bits!

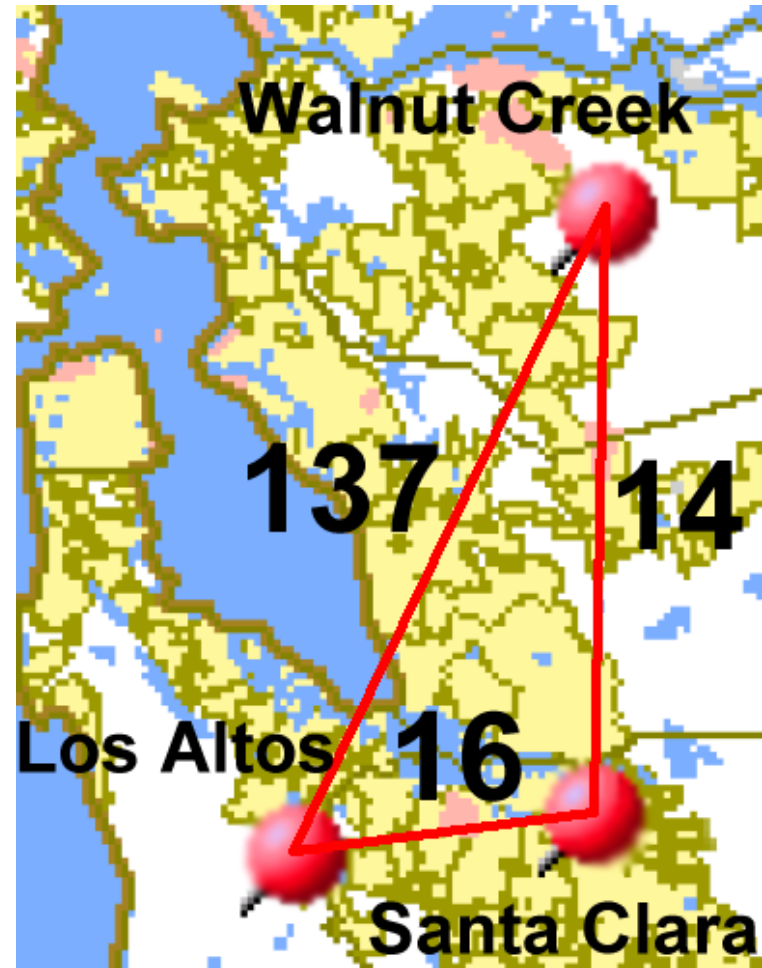


BGP Decision process

- Default decision for route selection
 - ◆ Highest local pref, shortest AS path, lowest MED, prefer eBGP over iBGP, lowest IGP cost, router id
- Many policies built on default decision process, but...
 - ◆ Possible to create arbitrary policies
 - » Any criteria: BGP attributes, source address, port # is prime, ...
 - » Can have separate policy for inbound routes, installed routes and outbound routes
 - ◆ Limited only by power of vendor-specific routing language
- Try to influence decision process at other ASs
 - ◆ AS padding, MEDs, Communities
 - ◆ More specific routes

BGP+policy does not arrive at shortest path

- Measured round-trip times between sites
- Pythagoras would have wept



(Times in milliseconds)

General Problems w/BGP

- Instability
 - ◆ Route flapping
 - ◆ Long AS-path decision criteria defaults to DV-like behavior (bouncing)
 - ◆ Not guaranteed to converge, NP-hard to tell if it does
- Scalability
 - ◆ ~100,000 network prefixes in default-free table today
 - ◆ Tension: Want to manage traffic to very specific networks (eg. multihomed content providers) but also want to aggregate information.

Routing policy

- So far we've discussed mechanism...
- How and why are basic routing policies decided?

History

- First policies for political reasons
 - ◆ NSFnet AUP (even today Internet2)
- Emergence of commercial policies
 - ◆ 1994-1995 NSFnet transition
 - » NSF ceases to run Internet backbone
 - » Commercial carrier (MCI, Sprint, ANS) start **selling** IP backbone service
 - » Interconnected with each other and regional networks at several public NAPs
 - » Everyone talks to everyone
 - ◆ Then five years went by...

Background – Settlement

- The telephone world
 - ◆ LECs (local exchange carriers)
 - ◆ IXC (inter-exchange carriers)
- LECs **MUST** provide IXCs access to customers; regulation
- When a call goes from one phone company to another:
 - ◆ Call billed to the caller
 - ◆ The money is split up among the phone systems – this is called “settlement”

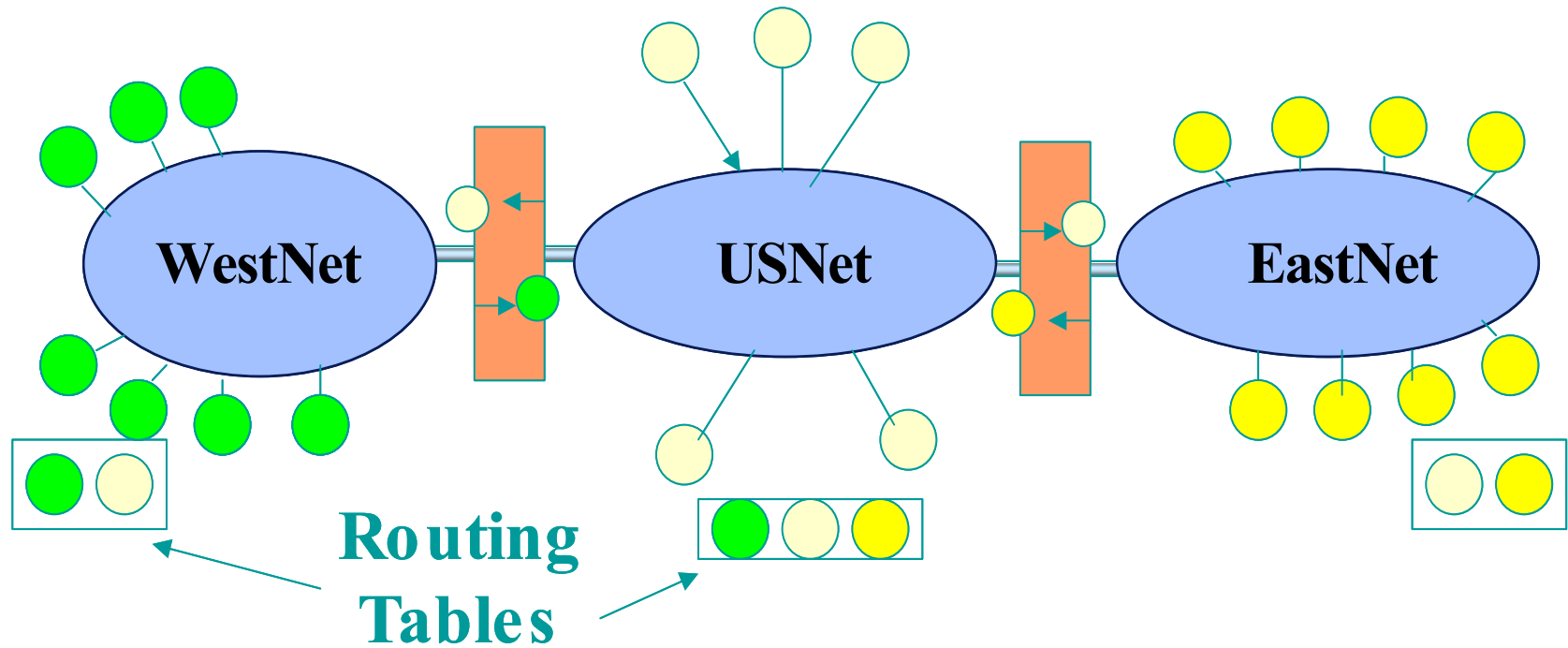
On the Internet...

- No regulation
 - ◆ One ISP doesn't have to talk to another
- Founded on “shared goodwill”
 - ◆ Pay for connectivity, not per packet
 - ◆ Not clear who should pay anyway
- No standard settlement

Peering vs Transit

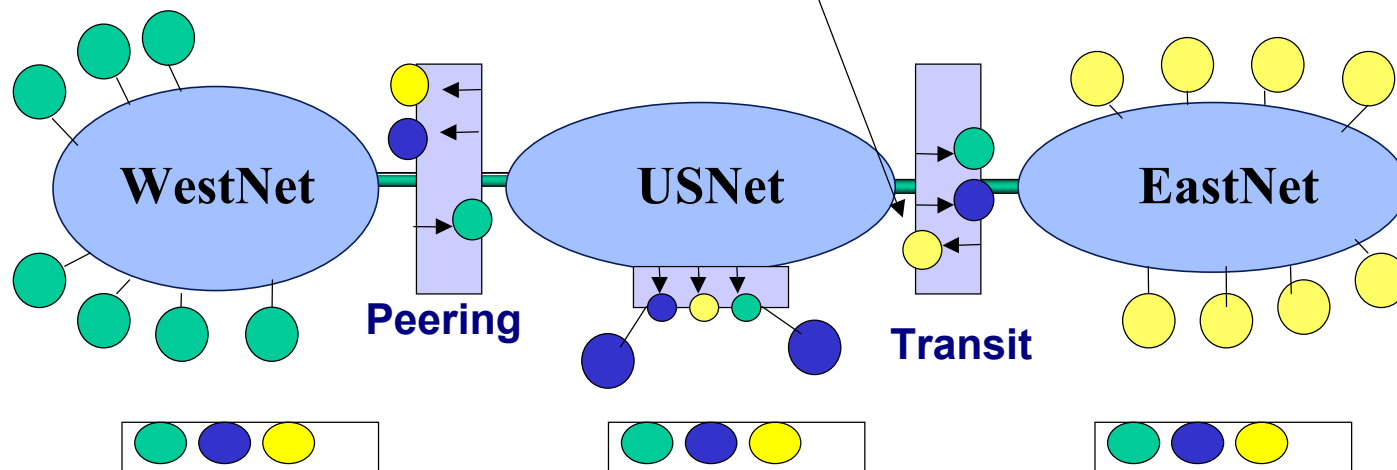
- Peering
 - ◆ Two ISPs provide connectivity to each others customers (traditionally for free)
 - ◆ Non-transitive relationship
- Transit
 - ◆ One ISP provides connectivity to every place it knows about (usually for money)

Example: peering



Example: transit

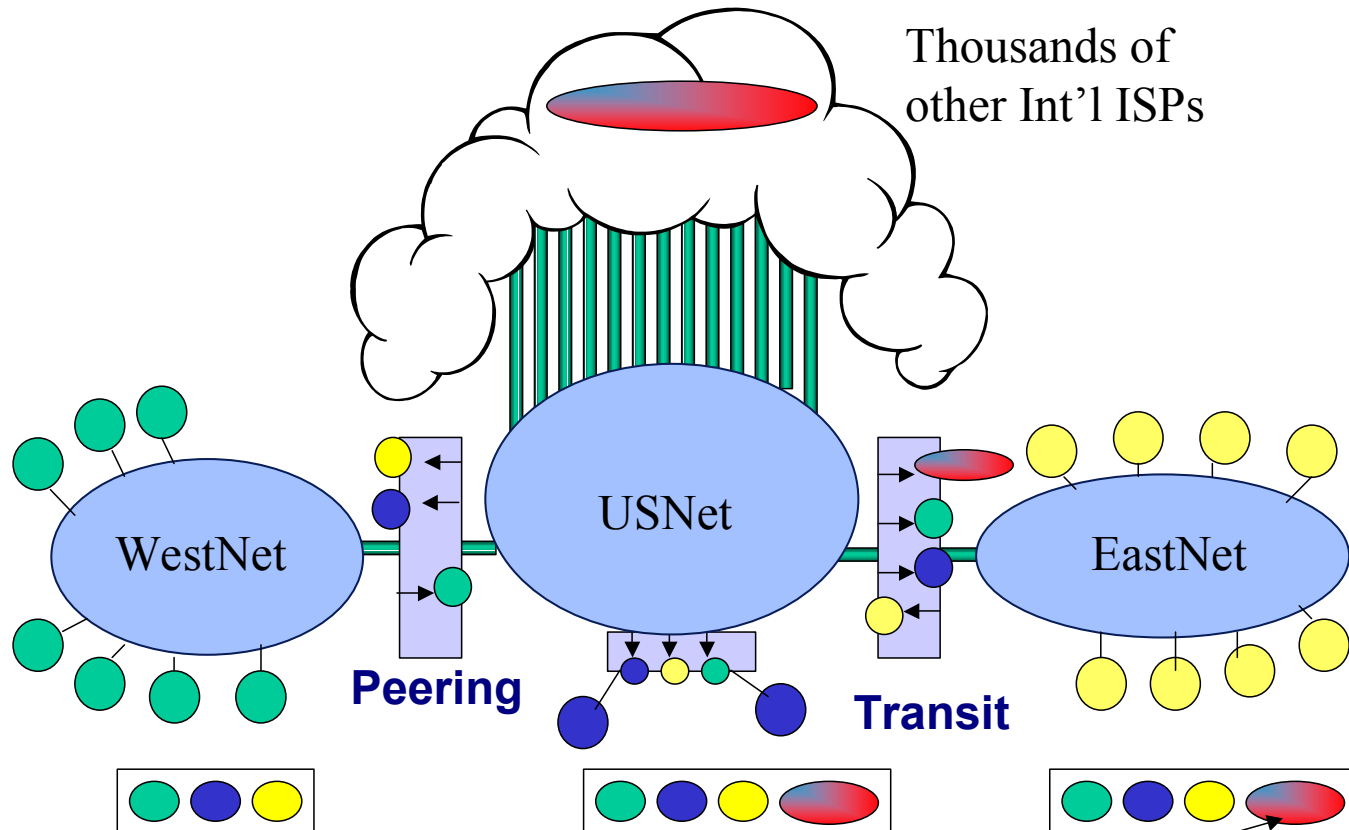
By EastNet purchasing transit,
Eastnet is announced by USNet to
USNet peering and transit interconnections alike.



The value of transit

- Not just paying for the fiber, but the connectivity
 - ◆ Remember, there is no single “backbone”
 - ◆ If you’re an ISP, how do your customers get to yahoo.com?
- Means big ISPs have more value to offer small ISPs than vice-versa

The value of transit (2)



The entire Internet
as known by USNet

Aside...

- Peering and transit are really two popular points on a continuum
- Some places sell “partial transit”
- Other places sell “usage-based” peering
- Principle issue is:
 - ◆ Which routes do you give away and which do you sell? To whom? Under what conditions?

Terminology 101: What's a Tier-1 ISP?

- My definition:
 - ◆ ISP big enough that they don't have to buy transit
 - ◆ AT&T, Sprint, Uunet, Genuity, etc.
- Tier-2 buy transit from Tier-1, etc.

- Increasingly worthless terms
 - ◆ Everyone claims to be Tier-1
 - ◆ More complicated forms of settlement
 - ◆ Leverage depends on business model

Terminology 101:

Public vs private peering

- Public peering
 - ◆ Connection via shared switch or network at “public” exchange point (place anyone can be if they pay money)
 - ◆ Still negotiated bilaterally
- Private peering
 - ◆ Private point-to-point link between peers

Why peer?

- Transit is very expensive
 - ◆ Was \$150,000 for an OC3 (155Mbps) transit link
- Peering with other ISPs can reduce the amount of traffic sent on transit link
 - ◆ Also lower latency?
- Communication patterns aren't uniform
 - ◆ More of your traffic is exchanged with some networks than others
 - ◆ Try to peer with other ISPs whose customers exchange traffic frequently with your customers...

Why not peer?

- Traffic asymmetry
 - ◆ More traffic goes one way than the other
 - ◆ Peer who carries more traffic feels cheated
- Hassle
- Top tier (big) ISPs have no interest in helping lower tier ISPs compete
 - ◆ The “Big Boys” all peer with each other at no/little cost
- Harder to deal with problems without strong financial incentive

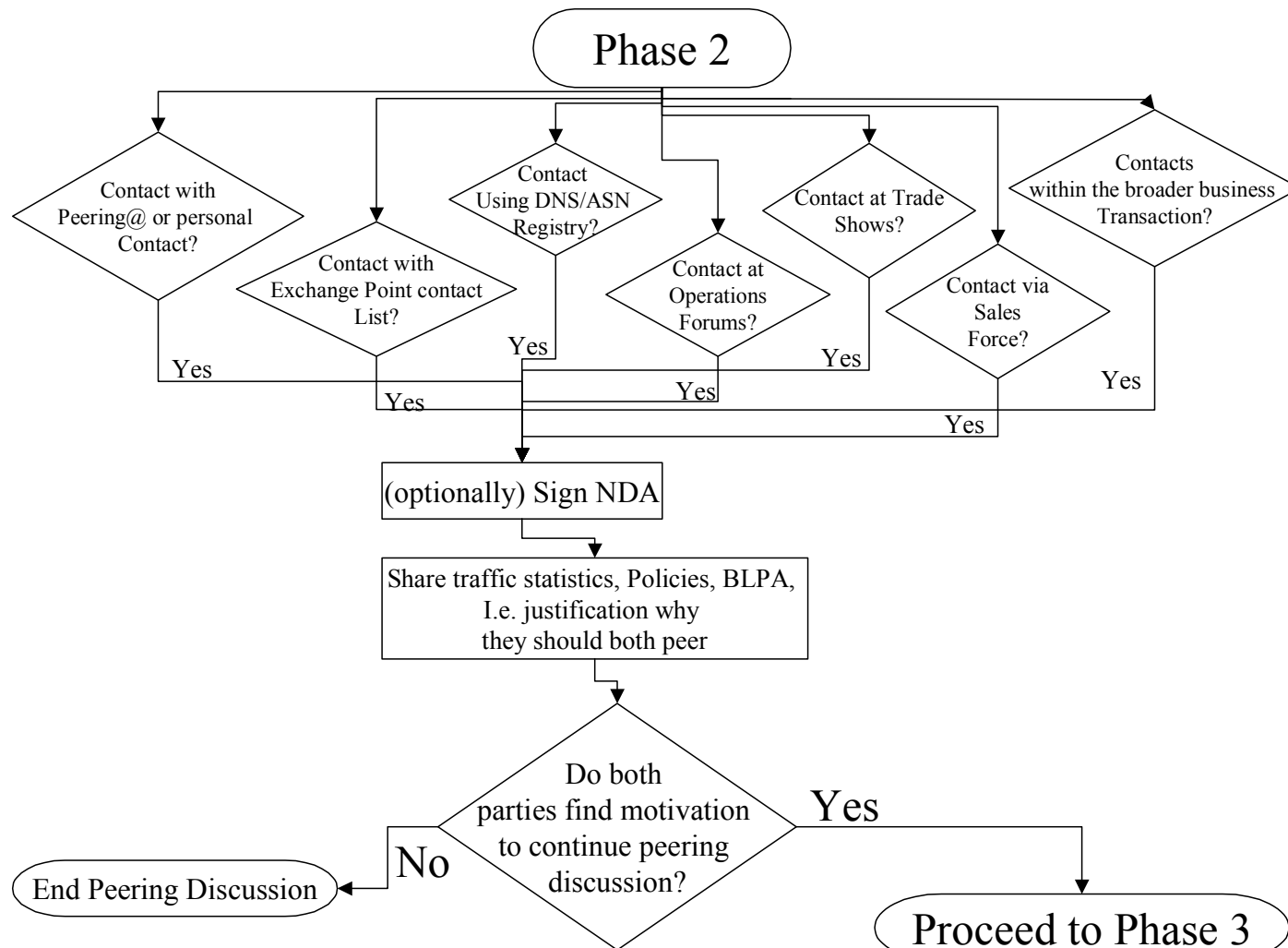
Who to peer with?

- Technical issues
 - ◆ Where does my traffic go? Biggest benefit?
- Business issues
 - ◆ Other partnerships/deals
 - ◆ Peering policies (once shrouded in mystery, now becoming more open)
 - ◆ Cost of exchanging traffic (i.e. where do we peer)

Model for Tier-2 ISPs

1. Buy transit from big provider
2. Peer at public exchange points to reduce transit cost
3. Establish private point-to-point peering with key ISPs
4. When you're big enough, negotiate peering with transit provider

Negotiating peering... (ugh)



How to interconnect?

- Direct connection
 - ◆ Cost of circuit lease (\$\$\$)
- Exchange-based interconnect
 - ◆ Exchange: place that houses equipment from multiple networks to exchange traffic
 - ◆ If you both already have equipment in the same building somewhere, then just run a cable between your machines (cheap)
 - ◆ Neutral exchanges vs affiliated exchanges

Summary

- Interdomain-routing
 - ◆ Exchange reachability information (plus hints)
 - ◆ Local policy to decide which path to follow
- Traffic exchange policies are a big issue \$\$\$
 - ◆ Complicated by lack of compelling economic model (who creates value?)
 - ◆ Very hard to be a small ISP
- Business issues can have serious operational/performance impact on the Internet

Discussion

- Performance impact peering vs transit and ratio-based peering
 - ◆ I have a funny story about @Home
- How do CDNs affect peering/transit issue?
- Implicit trust issues in transit routes
 - ◆ Will X really get my packets to Y who isn't X's customer?
- What if someone lies?

For next time...

- Multicast routing...
- Read Deering and Cheriton 90 and Almeroth 2000
- Chapter 4.4

Recap: Classless IP addressing

- Routes represented by tuple (network prefix/mask)
 - ◆ Allows arbitrary allocation between network and host address
 - ◆ e.g. 10.95.1.2/8: 10 is network and remainder (95.1.2) is host



- Route lookup: *longest prefix match*
 - ◆ For a given destination, find entry in route table that matches the most number of bits (i.e. with the largest mask)
 - ◆ Example: 128.95.4.1
 - » One route for 128.95.0.0/16 (CMU)
 - » One route for 128.95.4.0/24 (CMU SCS)