Border Gateway Protocol

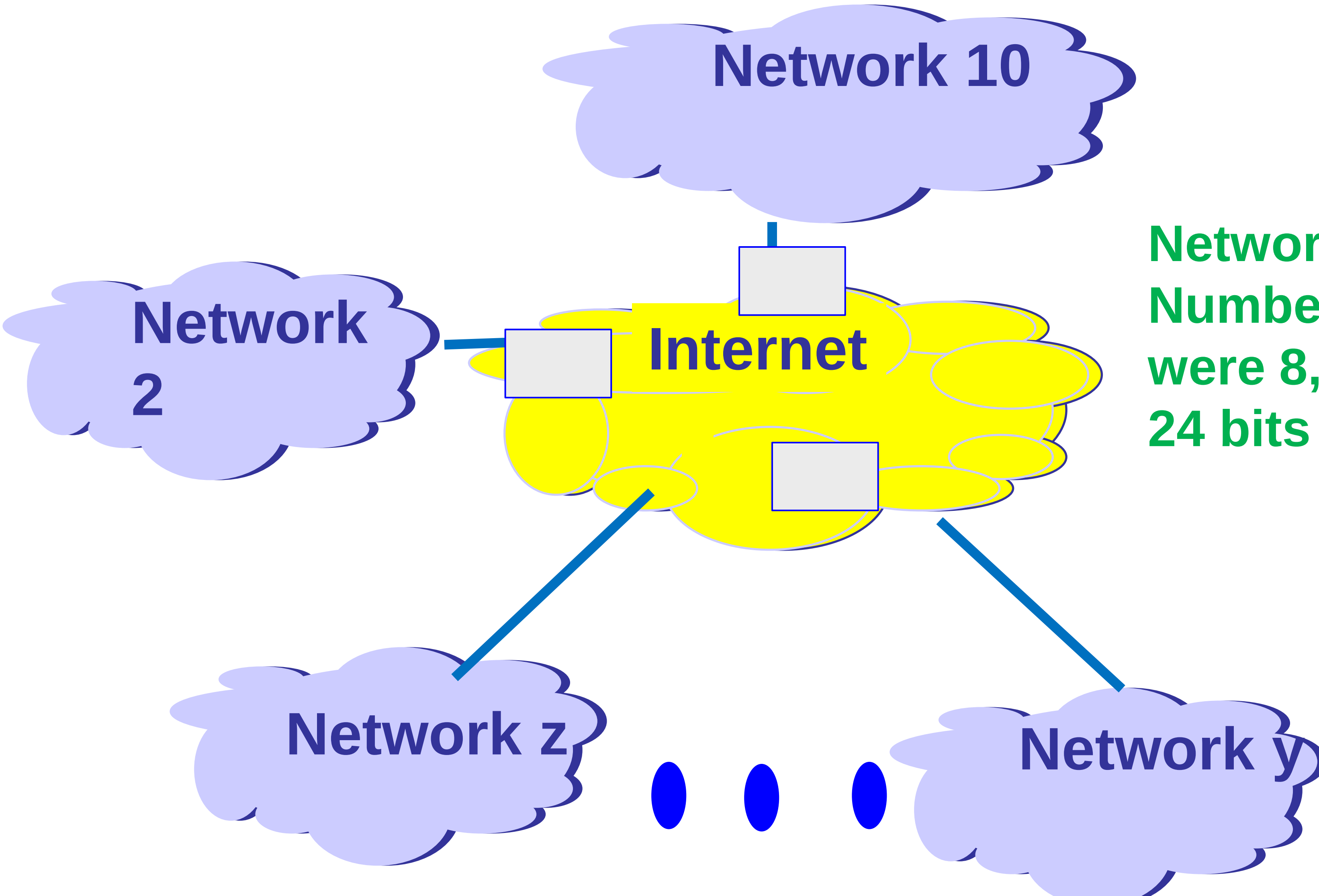
CSE 118: Computer Networks

George Varghese

Computing routes from UCLA to the world as opposed to computing routes within UCLA

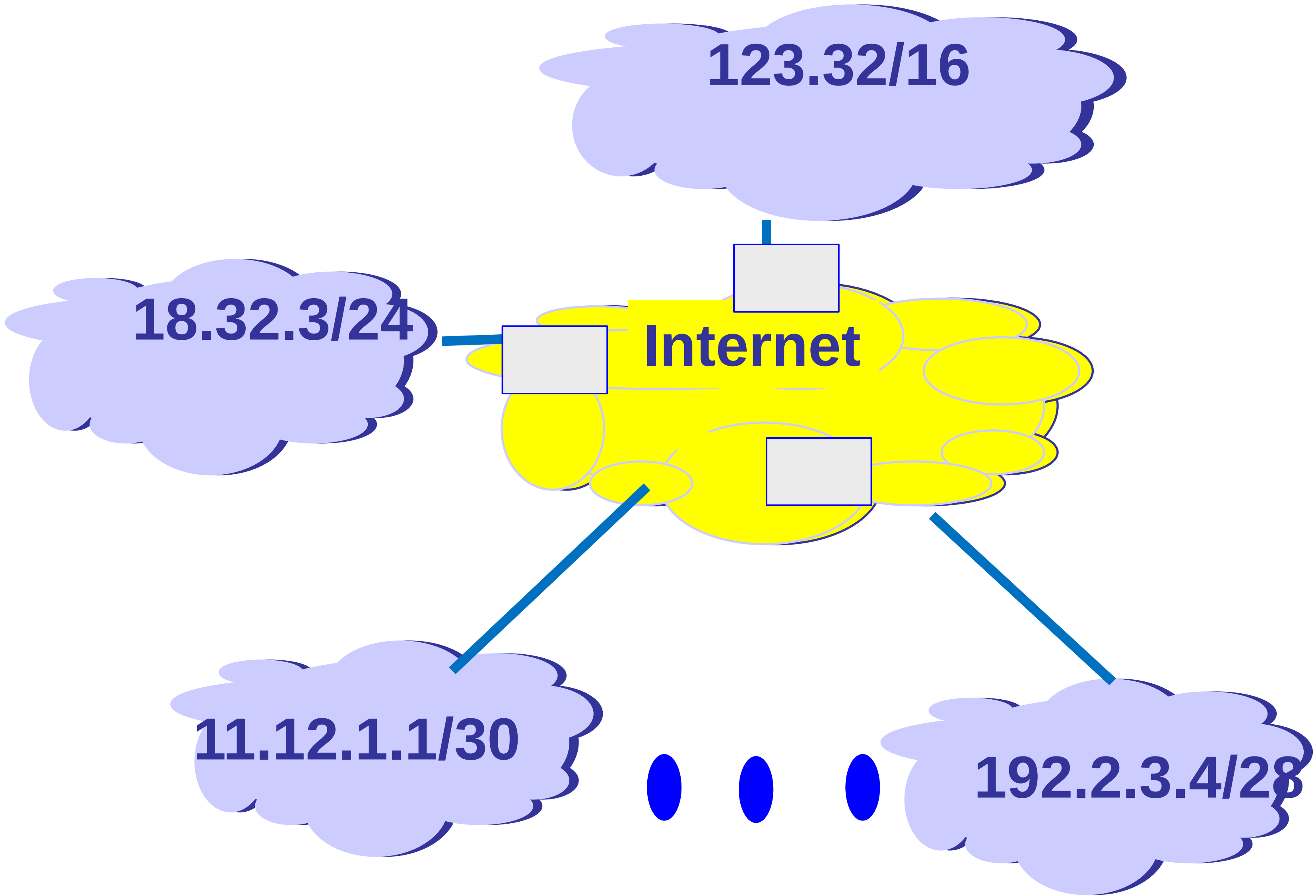
Many slides courtesy Alex Snoeren

IP's Abstract View of World



Network Numbers were 8, 16, or 24 bits

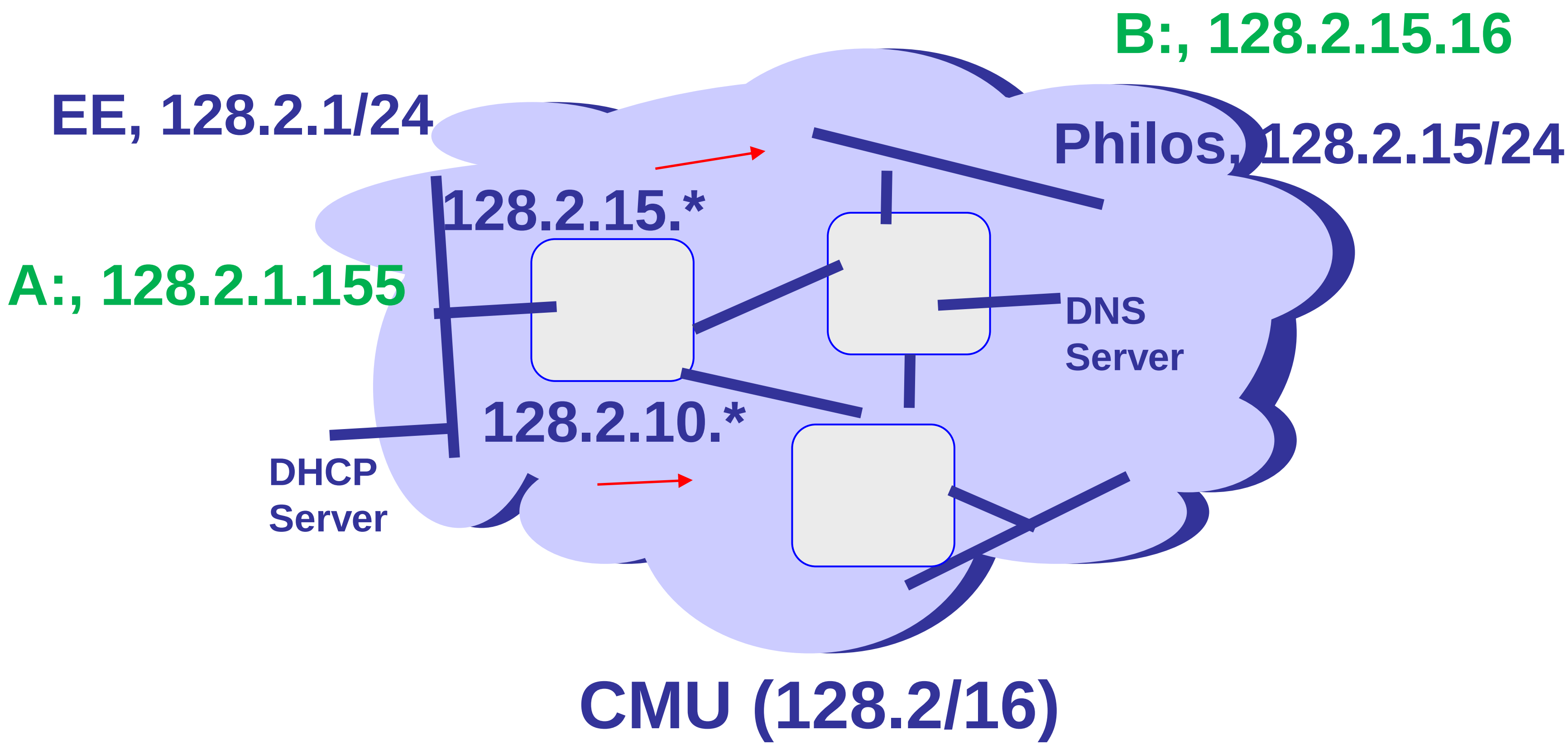
Classless New World: Prefixes



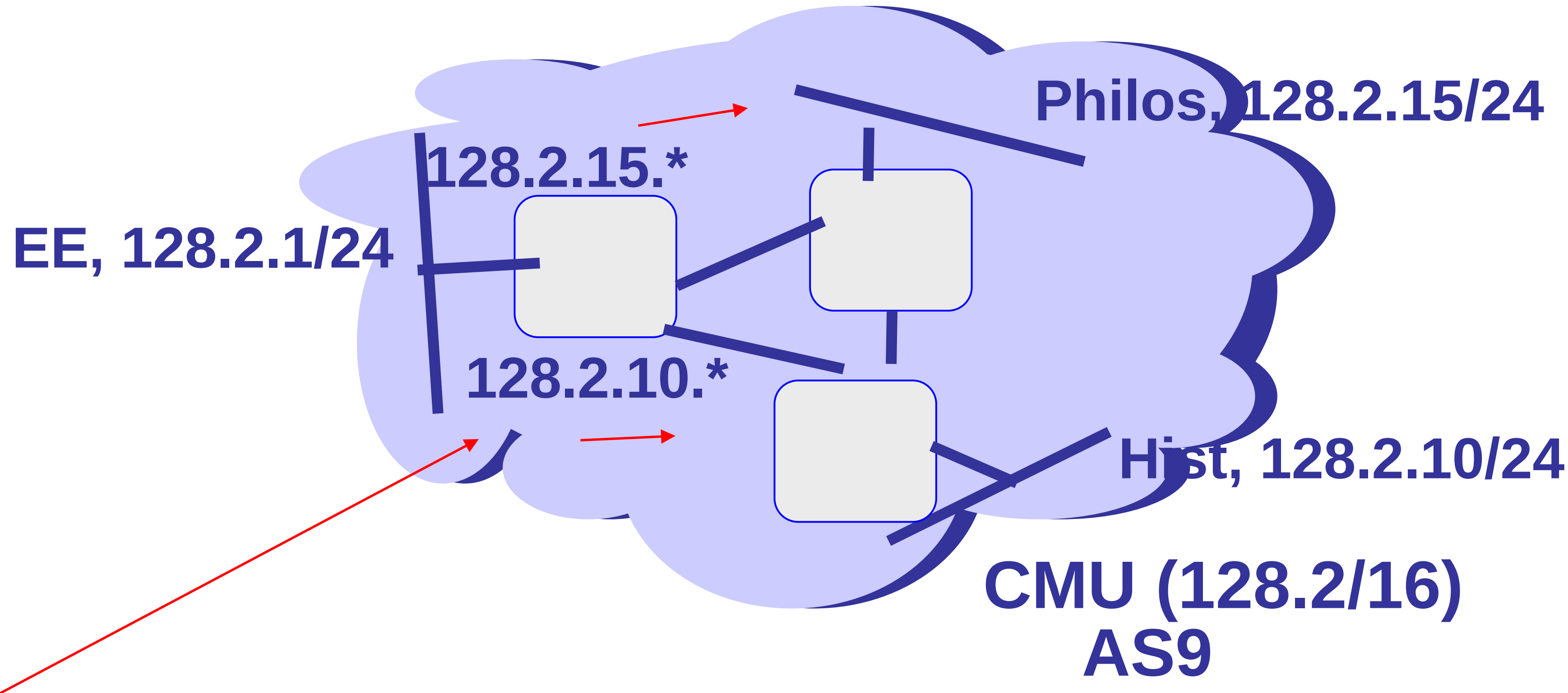
IP Addresses and Prefixes

- 32 bytes written as A.B.C.D, where A, B, C, D are integers from 1 to 255 representing one byte
- .For example, an EE server in CMU can be 128.2.1.155, first byte is 10000000 (remove dots)
- A Prefix is represented by slash or wildcard notation, For example CMU is 128.2/16 which means that all IP addresses in CMU start with 10000000 0000 0010 *
- Another way to encode prefixes is with a mask. Represent a /16 with a bit mask starting with 16 1's followed by 16 0's. Can AND with mask to find prefix

Get Started by DHCP and ARP



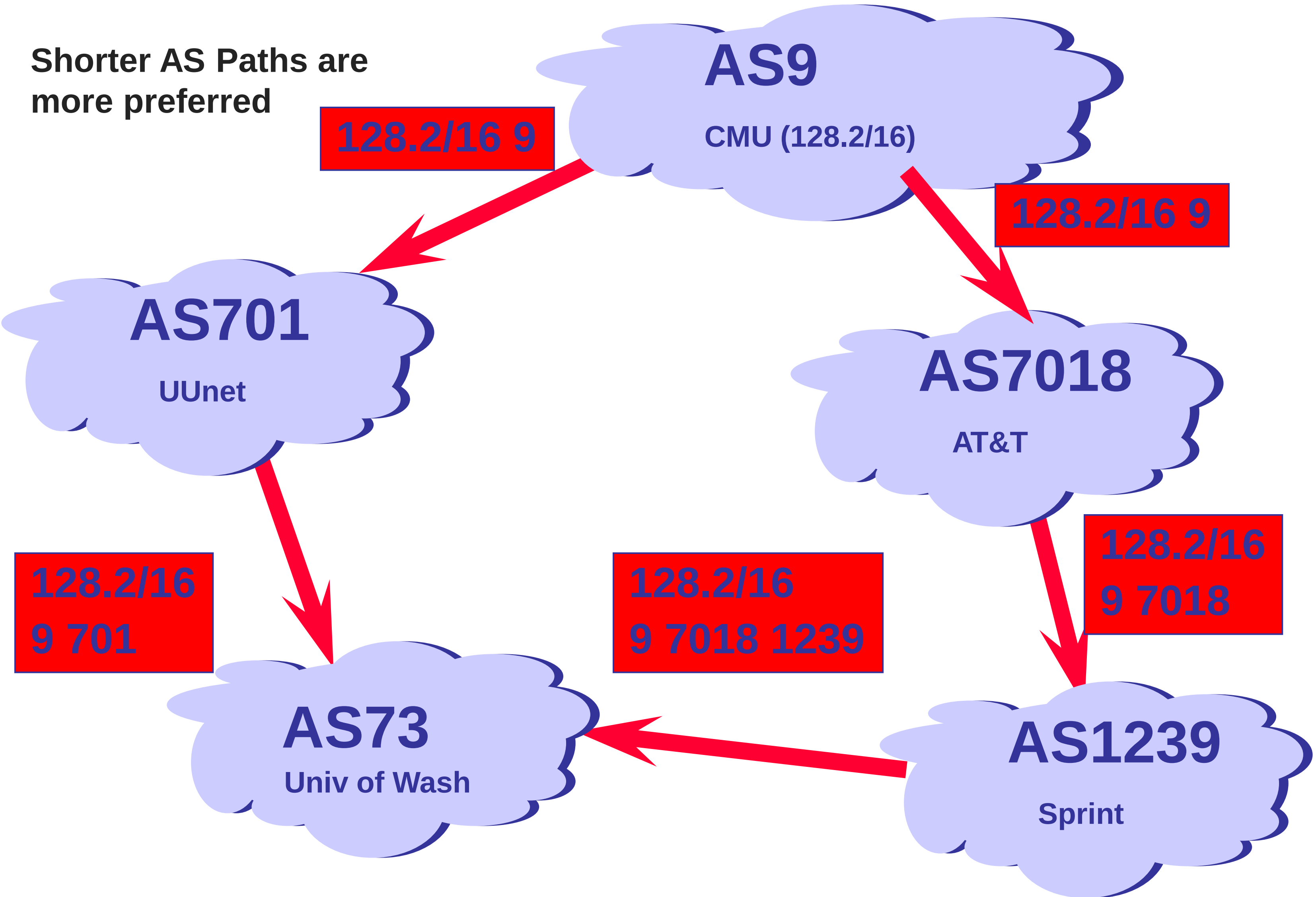
So far: Route Computation within an Autonomous System (AS)



Link State, or Distance Vector used within AS between routes to compute routes

BGP: Routing between ASes

Shorter AS Paths are more preferred



Why Interdomain Routing: Policy

Why not one happy melting pot of a network:

- Multiple providers (see IP evolution) implies need for independence and independent policies.
- Different metrics, trust patterns, different charging policies (hot potato, cold potato), different administrative and legal requirements (e.g., ARPANET only for government business, Canadian traffic stays within Canada).
- Not very well developed. Basic conflict between abstraction and hierarchies (for scaling) and ability to specify arbitrary policies.

Possible Polices

- Never use Routing Domain X for any destination.
- Never use domains X and Y.
- Don't use X to get to a destination in domain Y.
- Use X only as a last resort.
- Minimize number of domains in path.
- Government messages can traverse the ARPANET but not others.

BGP Overview

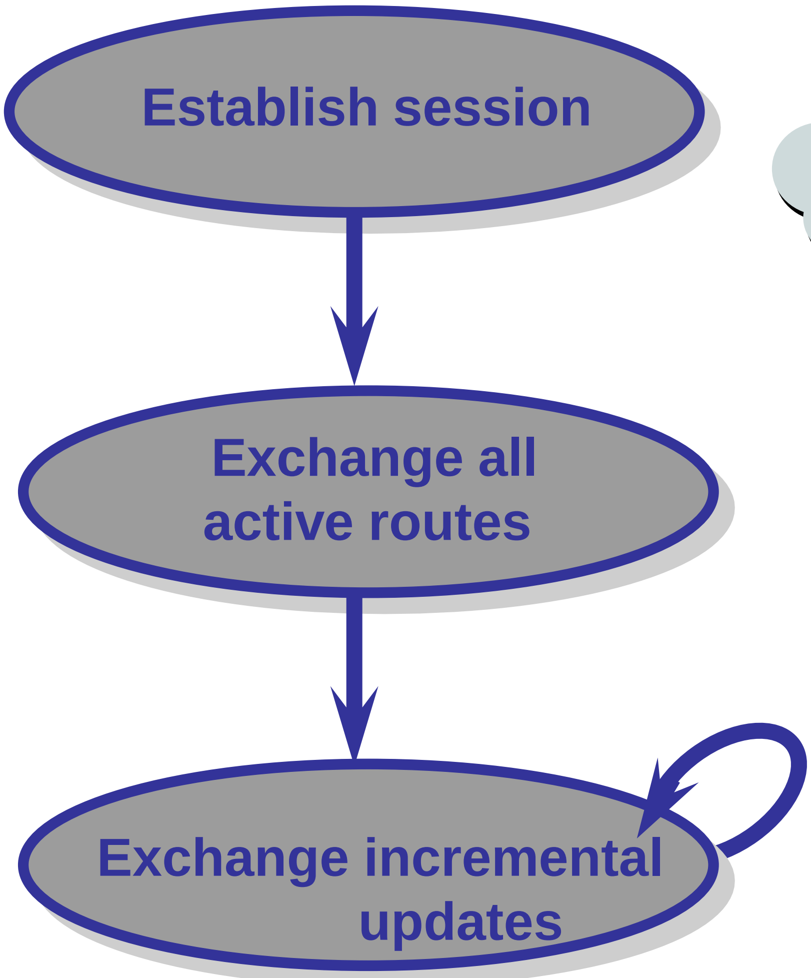
- Border Gateway Protocol (BGP)
  - The canonical path vector protocol
  - How routing gets done on the Internet today
- BGP operation
  - Basic BGP and differences from Distance vector
  - BGP features (Local Pref, MED, Community)
  - Issues with BGP
- BGP Alternatives

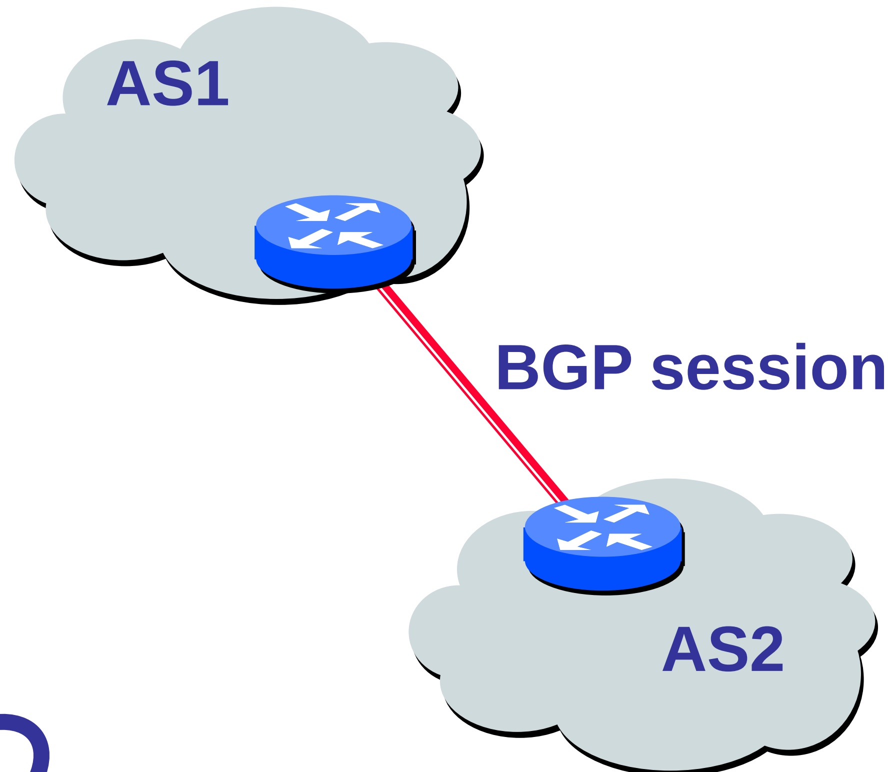
Border Gateway Protocol

- Interdomain routing protocol for the Internet
  - Prefix-based path-vector protocol
  - Policy-based routing based on AS Paths
  - Evolved during the past 28 years

- **1989 : BGP-1 [RFC 1105], replacement for EGP**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771], support for CIDR**
- **2006 : BGP-4 [RFC 4271], update**

Basic BGP Operation





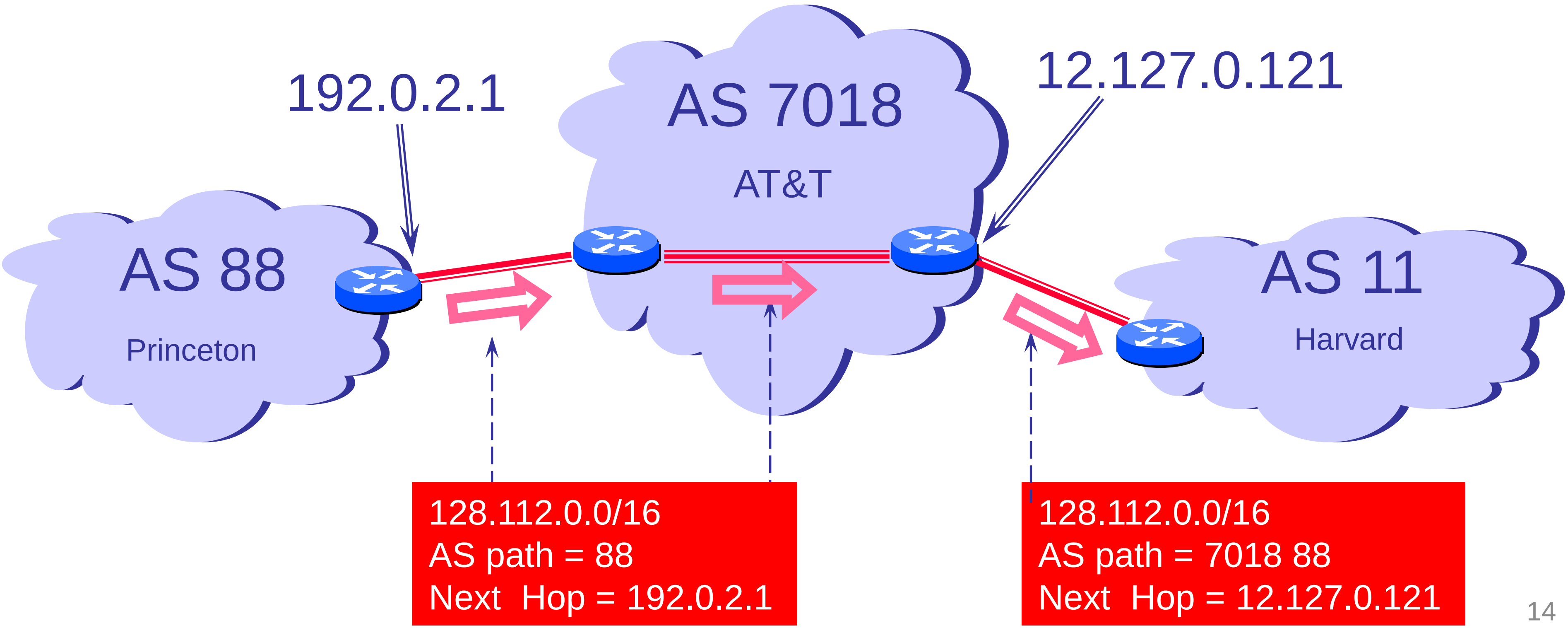
While connection is ALIVE exchange route UPDATE messages

Step-by-Step

- A node learns multiple paths to destination
  - Stores all of the routes in a routing table
  - Applies policy to select a single active route
  - … and may advertise the route to its neighbors
- Incremental updates unlike distance vector
  - Announcement
    - Upon selecting a new active route, add own node id to path
    - … and (optionally) advertise to each neighbor
  - Withdrawal
    - If the active route is no longer available
    - … send a withdrawal message to the neighbors

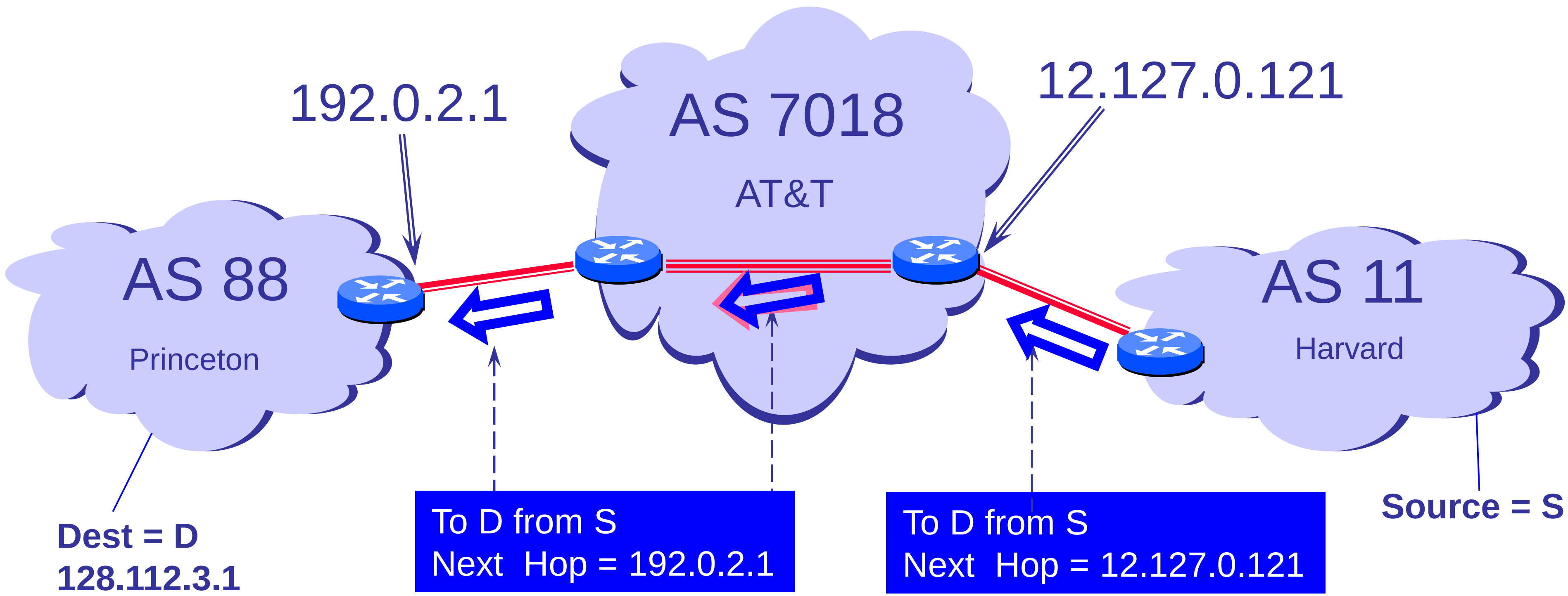
A Simple BGP Route

- Destination prefix (e.g., 128.112.0.0/16)
- Route attributes, including
  - AS path (e.g., “7018 88”)
  - Next-hop IP address (e.g., 12.127.0.121)



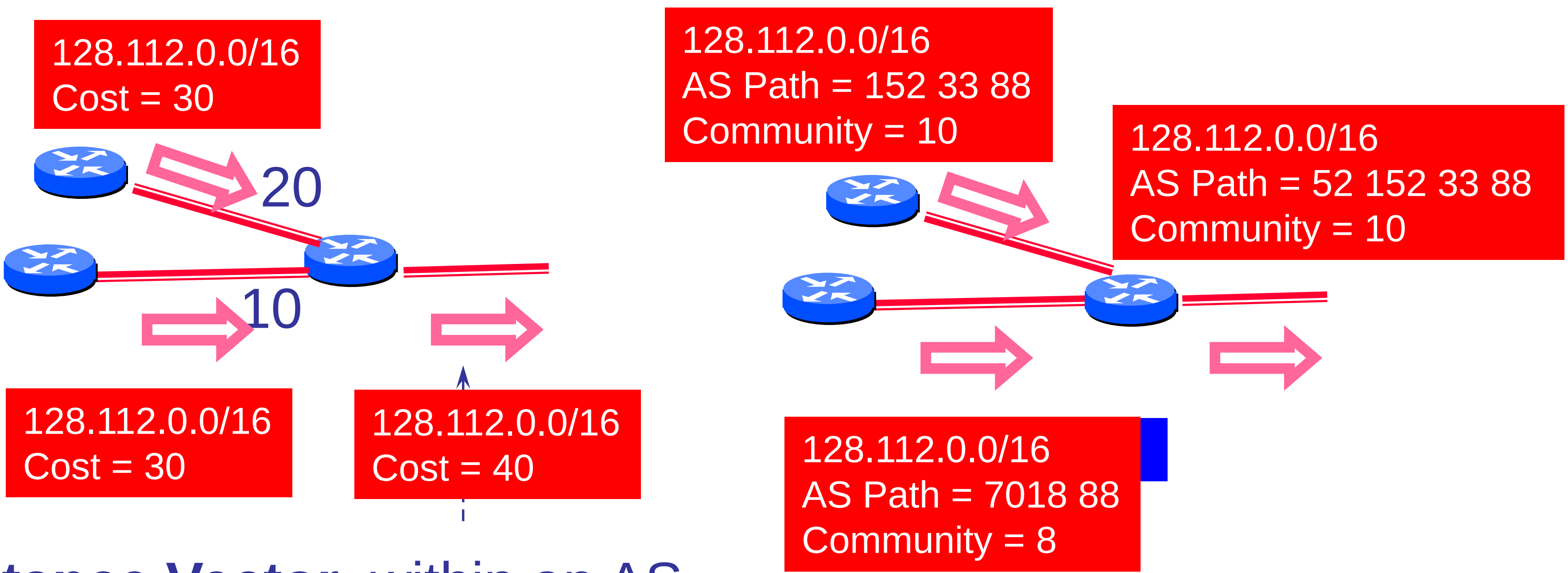
Data Packets flow in opposite direction from BGP updates

- Notice how Next Hop Info from last slide is crucial to build forwarding table at each route used to choose next hop
- Have to do ARP as well to get MAC address of next hop



Distance Vector versus BGP

- Only way in distance vector to tune routes is via cost
- In BGP, one can “control” routes in more complex ways



Distance Vector, within an AS, only attribute is cost, Always Pick & propagate shortest

Path Vector, between ASes, Multiple attributes, Complex Choices settable in config files

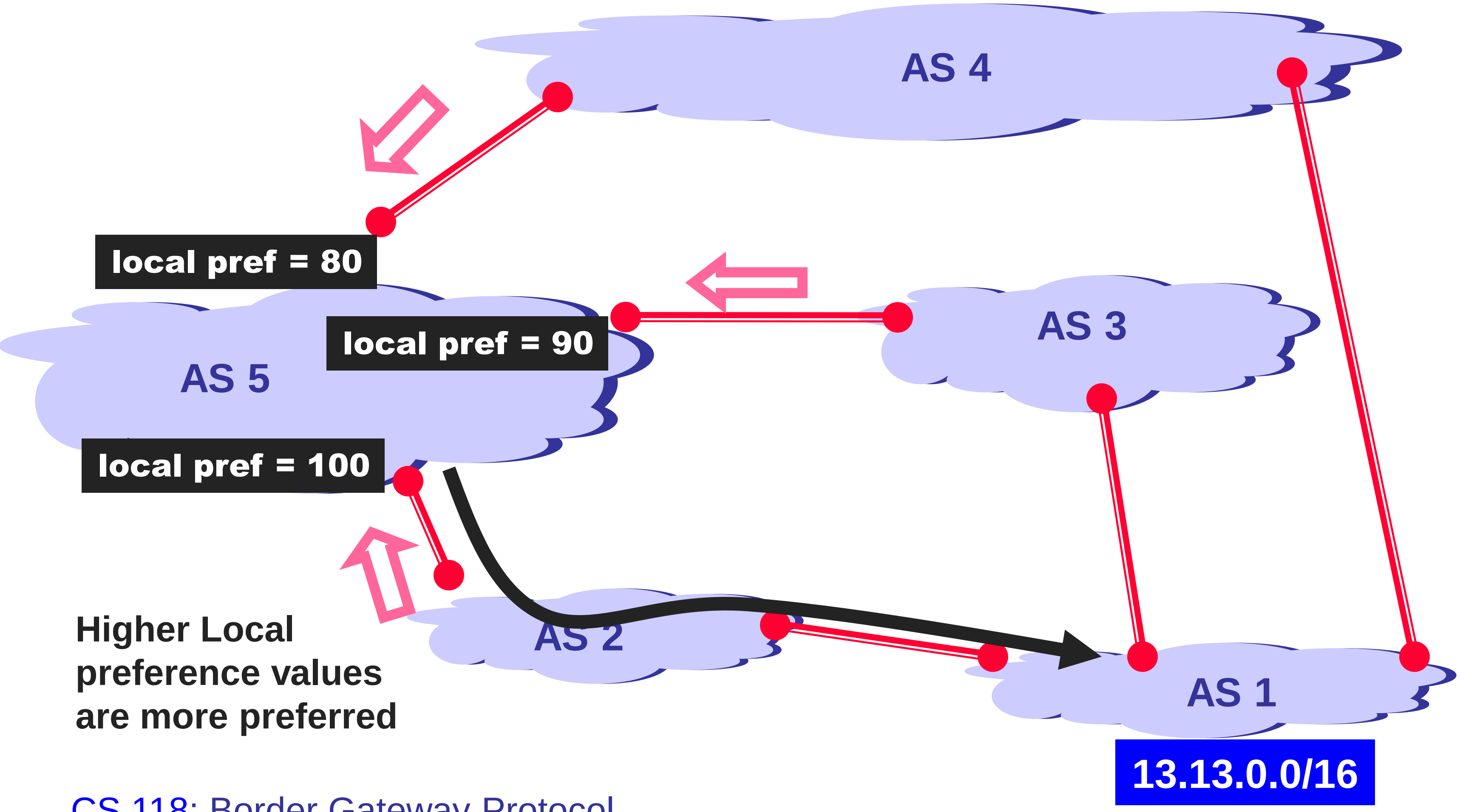
(some) BGP Attributes

- **AS path:** ASs the announcement traversed
- **Next-hop**: where the route was heard from
- **Origin:** Route came from IGP or EGP
- **Local pref:** Statically configured ranking of routes within AS
- **Multi Exit Discriminator:** preference for where to *exit* network
- **Community:** opaque data used to tag routes that are to be treated equivalently.

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 in principle
  - Limited only by power of vendor-specific routing language

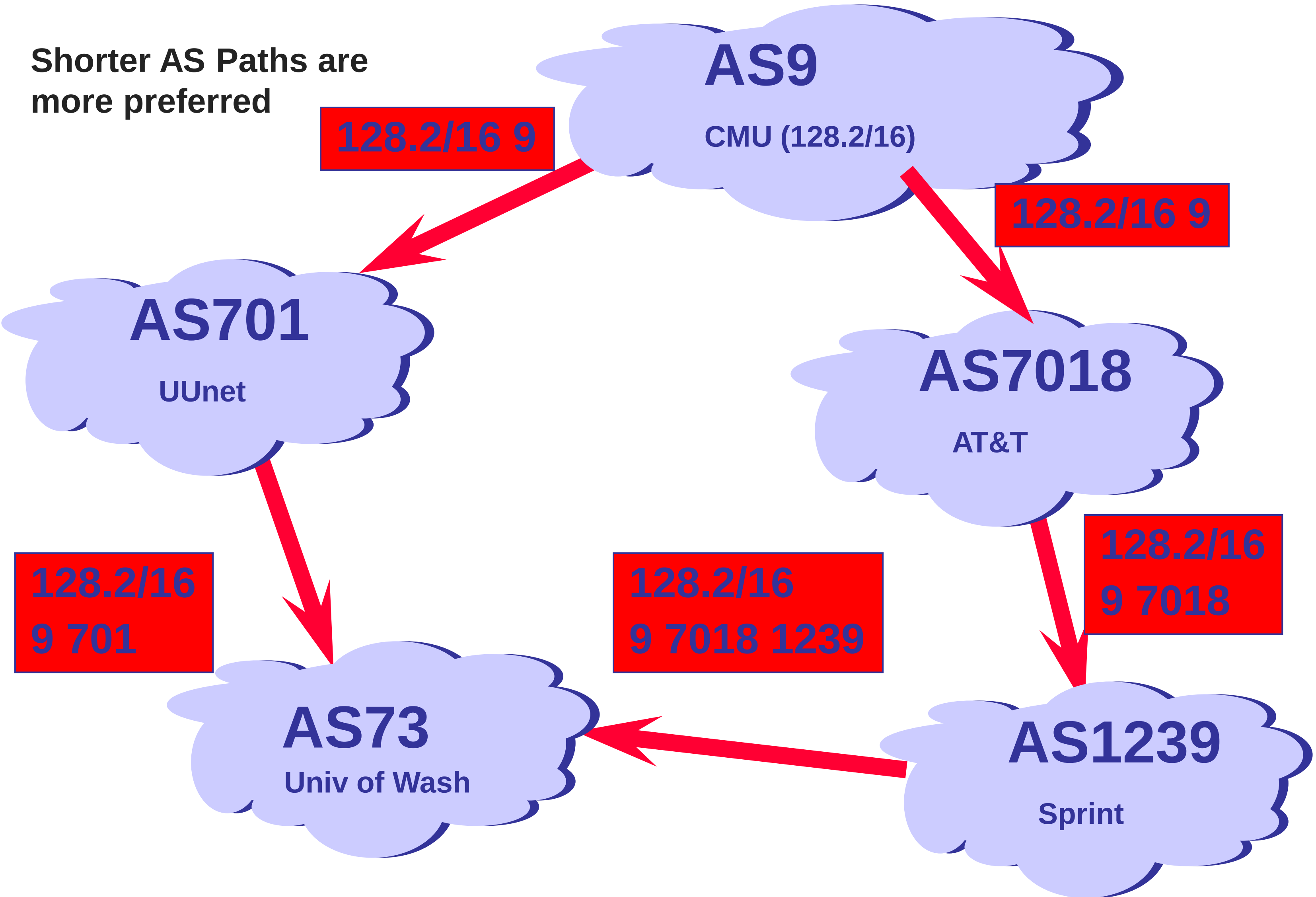
Example: Local Pref



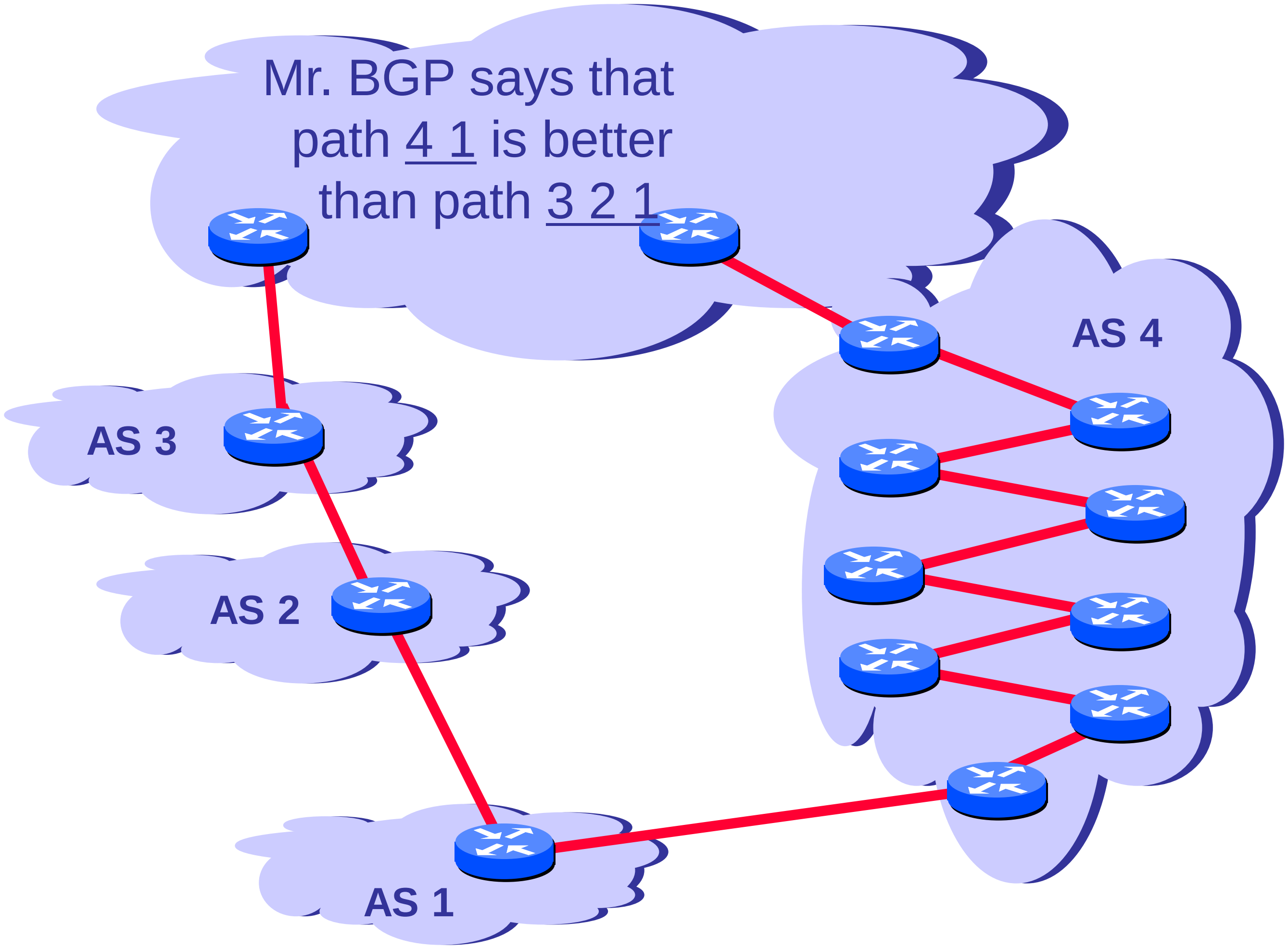
Higher Local preference values are more preferred

Example: Short AS Path

Shorter AS Paths are more preferred

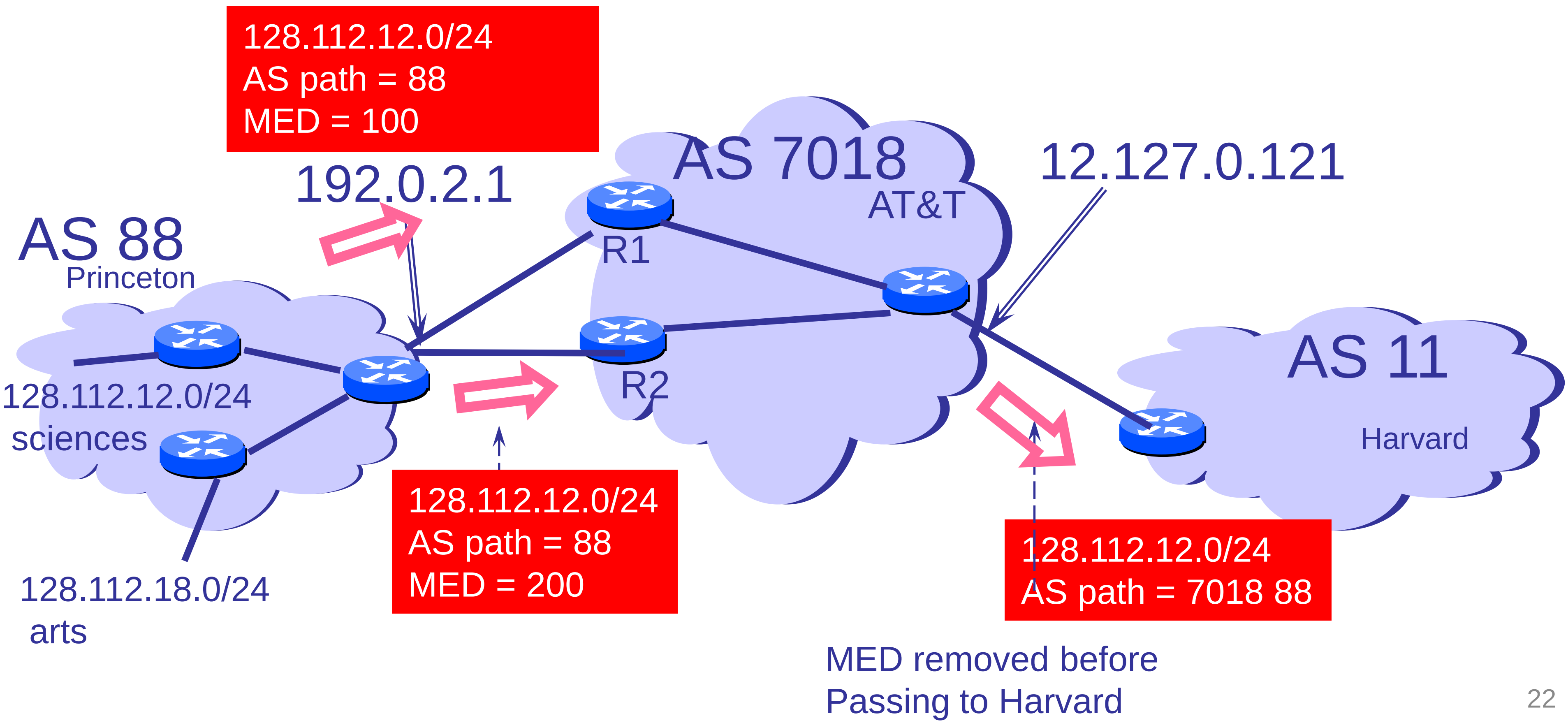


AS Paths *vs.* Router Paths



More intricate feature: MEDs

- Way to do load balancing by passing a hint to next AS
- Request Harvard send traffic to Princeton sciences via R2



Doing MEDs in Cisco router config at Princeton exit

neighbor R1 route-map setMED-R1 out
neighbor R2 route-map setMED-R2 out

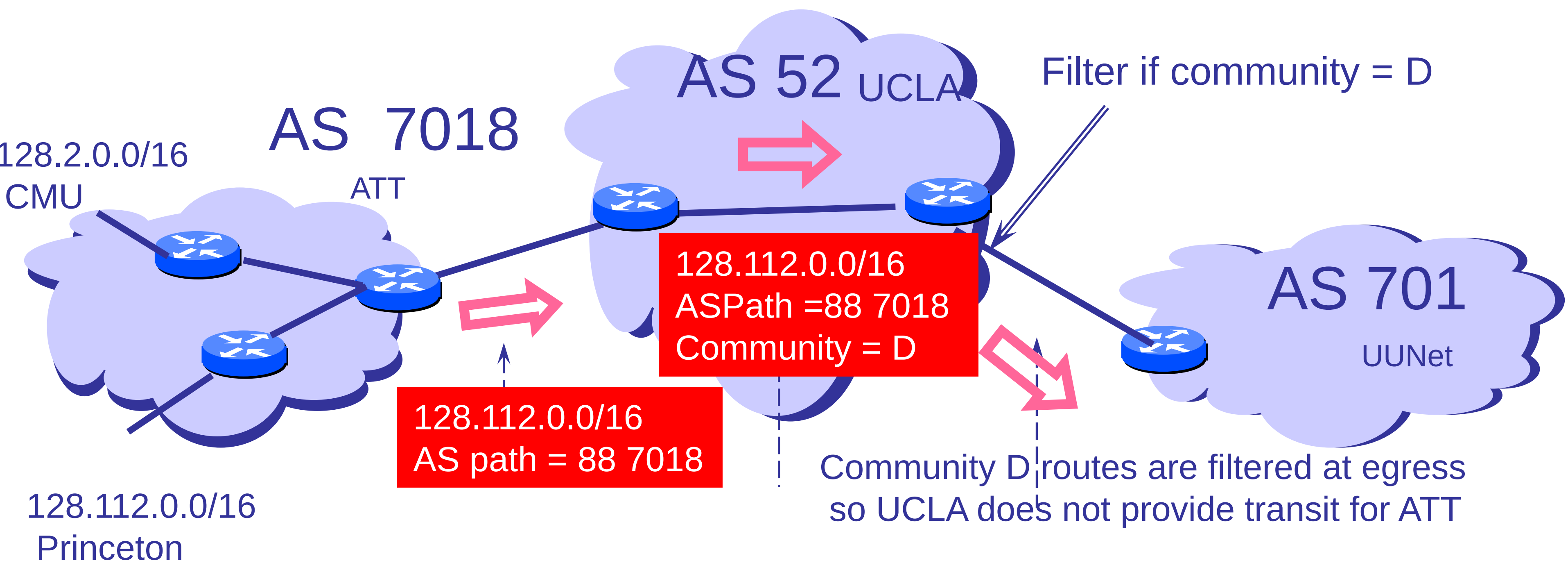
access-list 1 permit 128.112.12.0 255.255.255.0 // sciences
access-list 2 permit 128.112.18.0 255.255.255.0 // arts

route-map setMED-R1 … match ip address 1 set metric 100
// for R1 send science prefix with lower MED priority
route-map setMED-R1 … match ip address 2 set metric 200
// for R1 send arts prefix with higher MED prioriity

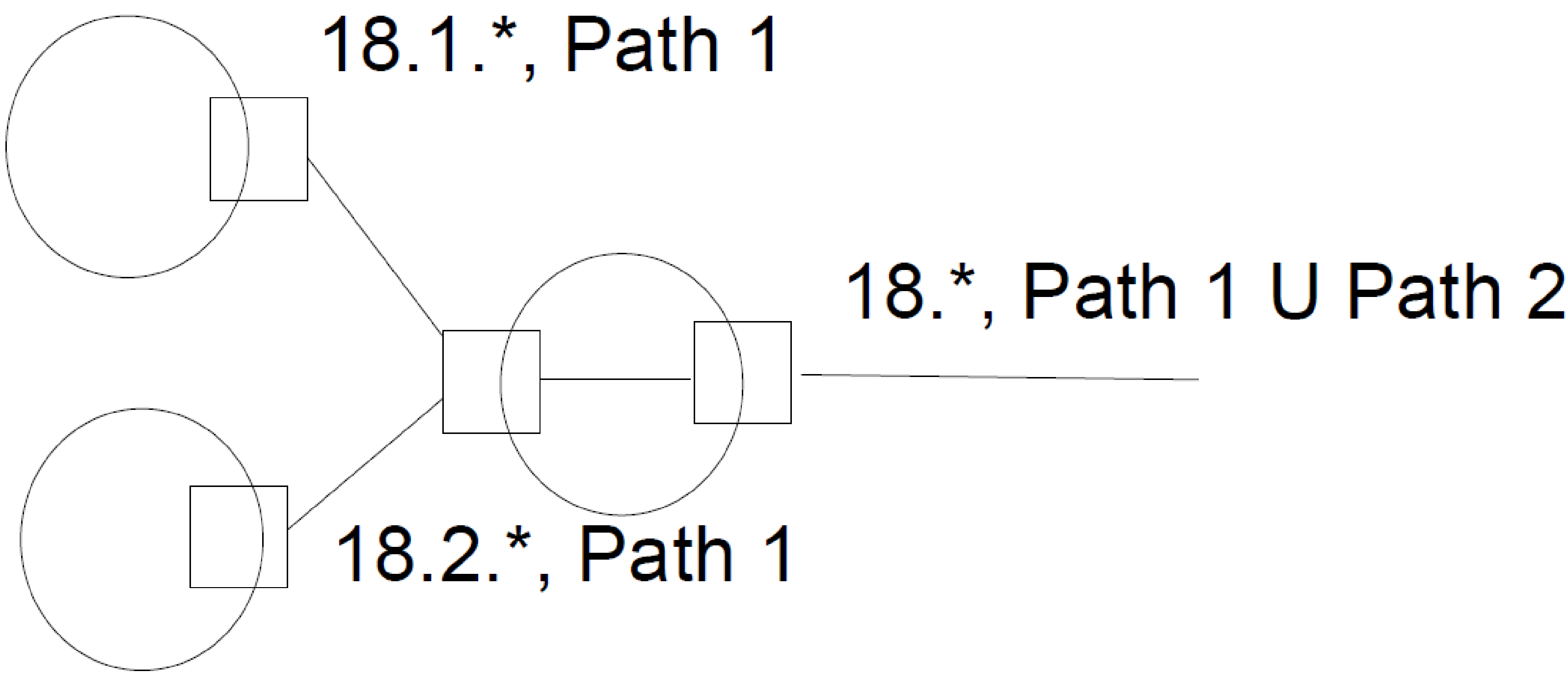
route-map setMED-R2 … match ip address 1 set metric 200
// for R2 send science prefix with higher MED priority
route-map setMED-R2 … match ip address 2 set metric 100
// for R2 send arts prefix with lower MED priority

Feature 2: community

- Way to tag multiple routes with same tag value
- Then remote routers can act on tag (e.g., filter)



Feature 3: Aggregation



BGP Has Lots of Problems

- Instability
  - Route flapping (network x.y/z goes down… tell everyone)
  - Not guaranteed to converge, NP-hard to tell if it does
- Scalability still a problem
  - >1,000,000 network prefixes in default-free table today
  - Tension: Want to manage traffic to very specific networks (eg. multihomed content providers) but also want aggregation
- Performance
  - Non-optimal, doesn't balance load across paths

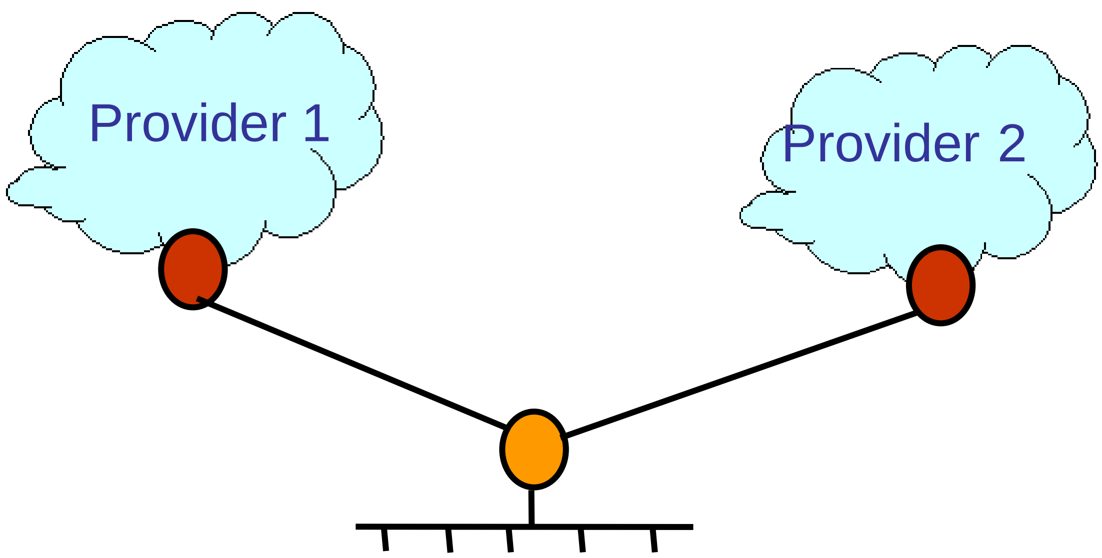
Business Relationships

- Neighboring ASes have business contracts
  - How much traffic to carry
  - Which destinations to reach
  - How much money to pay

- Common business relationships
  - Customer-provider
    - » E.g., Princeton is a customer of USLEC
    - » E.g., MIT is a customer of Level3
  - Peer-peer
    - » E.g., UUNET is a peer of Sprint
    - » E.g., Harvard is a peer of Harvard Business School

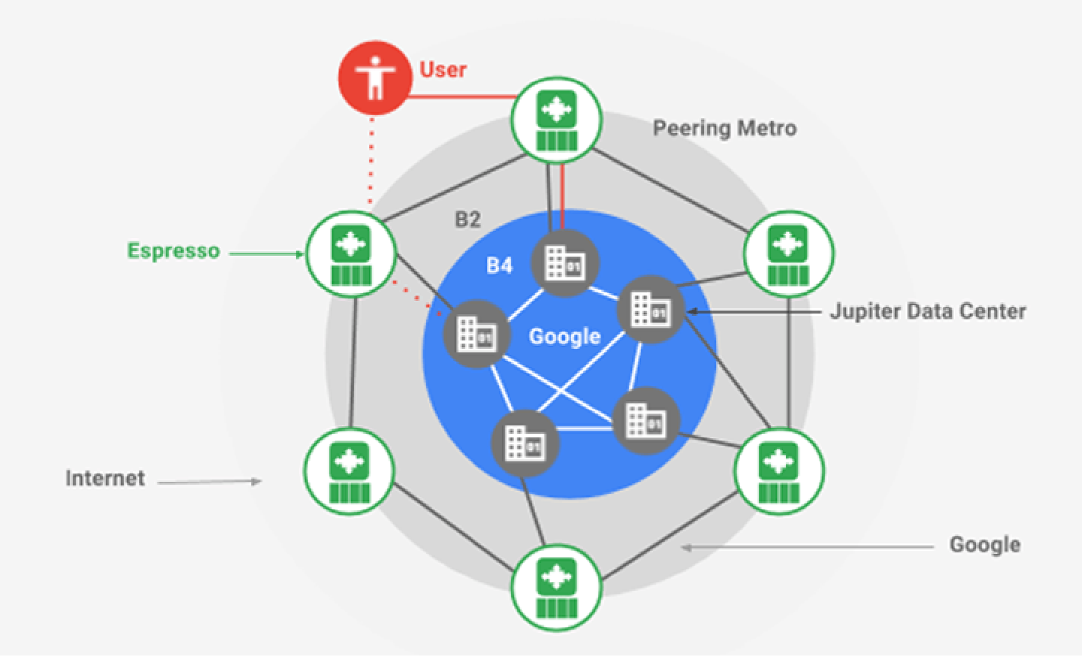
Multi-Homing

- Customers may have more than one provider
  - Extra reliability, survive single ISP failure
  - Financial leverage through competition
  - Better performance by selecting better path
  - Gaming the 95th-percentile billing model



Beyond BGP

- SDN inspired approaches like Google’s Espresso
- Link state versions of BGP (IDRP, Radia proposal)



Google Gives Last Mile a Shot of "Espresso"

Google border routers talk BGP to the outside world but send all BGP announcements to a service that also has latency information from Google Apps and so picks better routers to the external Internet

Conclusions

- Link State and Distance vector are used to route within a Domain/AS/ISP/Enterprise
- BGP is used to compute routes between ASes
- Basically like distance vector gossip except you add not just a total cost but list of all Ass in path so far.
- AS Path helps policy because any router can choose to drop based on AS's in path.
- AS Path also helps prevent loops without a hop count
- BGP has issues and there are alternatives to BGP