

Lecture 19:

Border Gateway Protocol and Router Implementation

CSE 123: Computer Networks

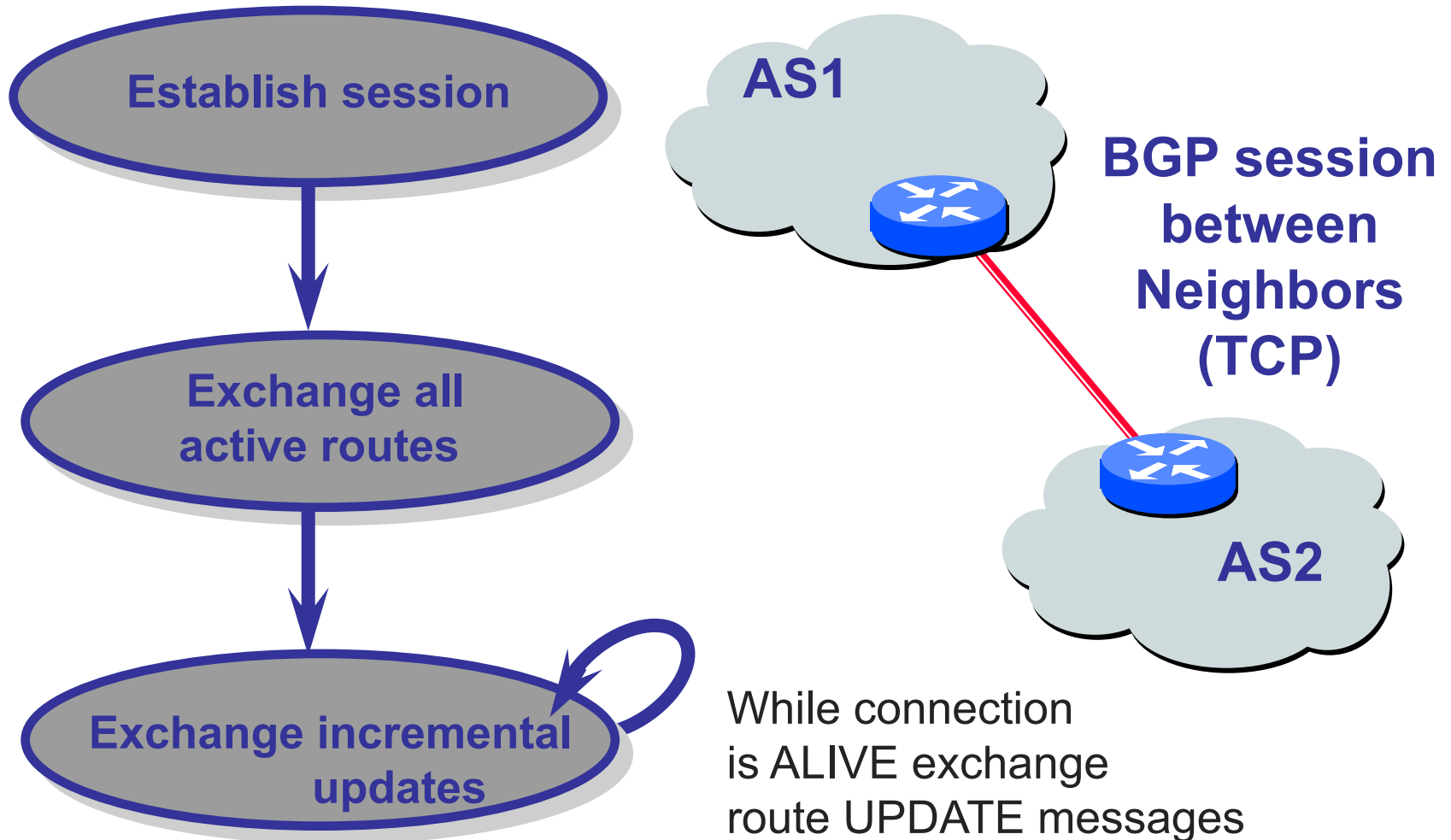
Aaron Schulman



Some figures courtesy Mike Freedman



Basic BGP Operation





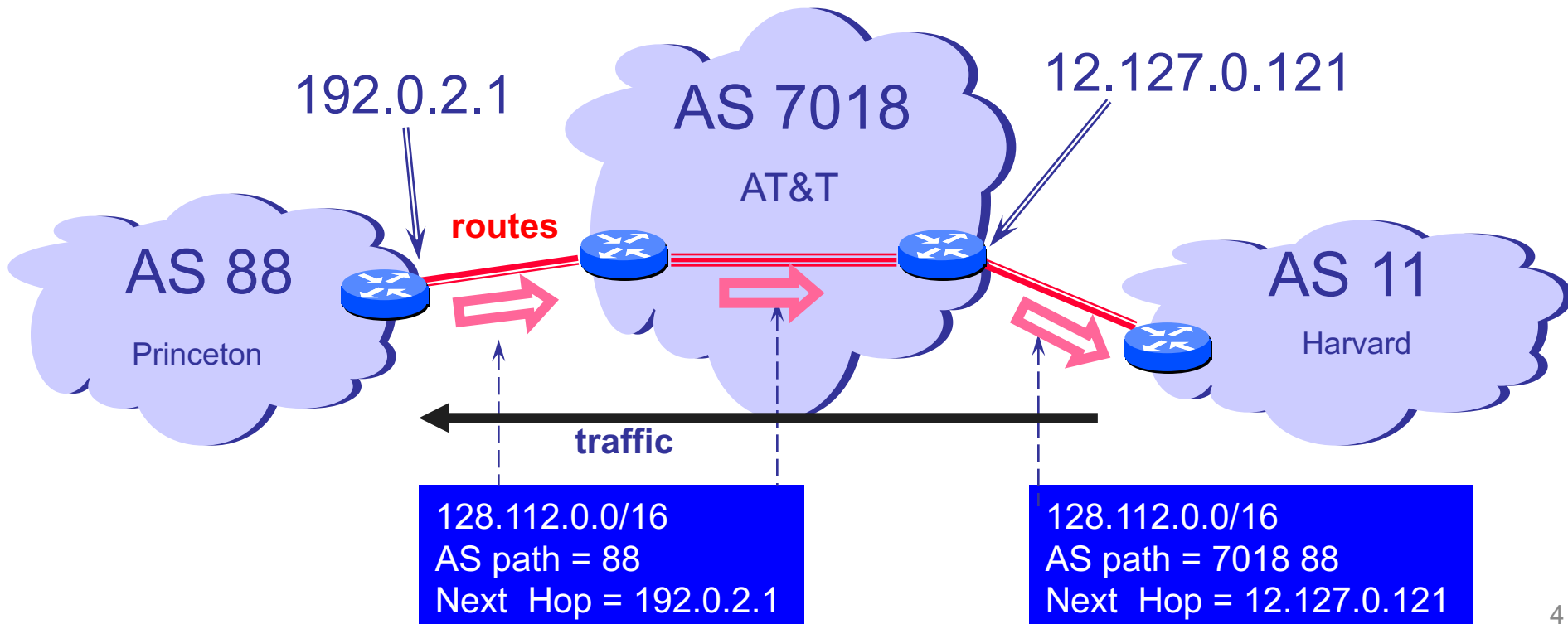
Step-by-Step

- A router 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
 - ◆ Announcement
 - » Upon selecting a new active route, add own AS 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)





(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 for inter-ISP policy



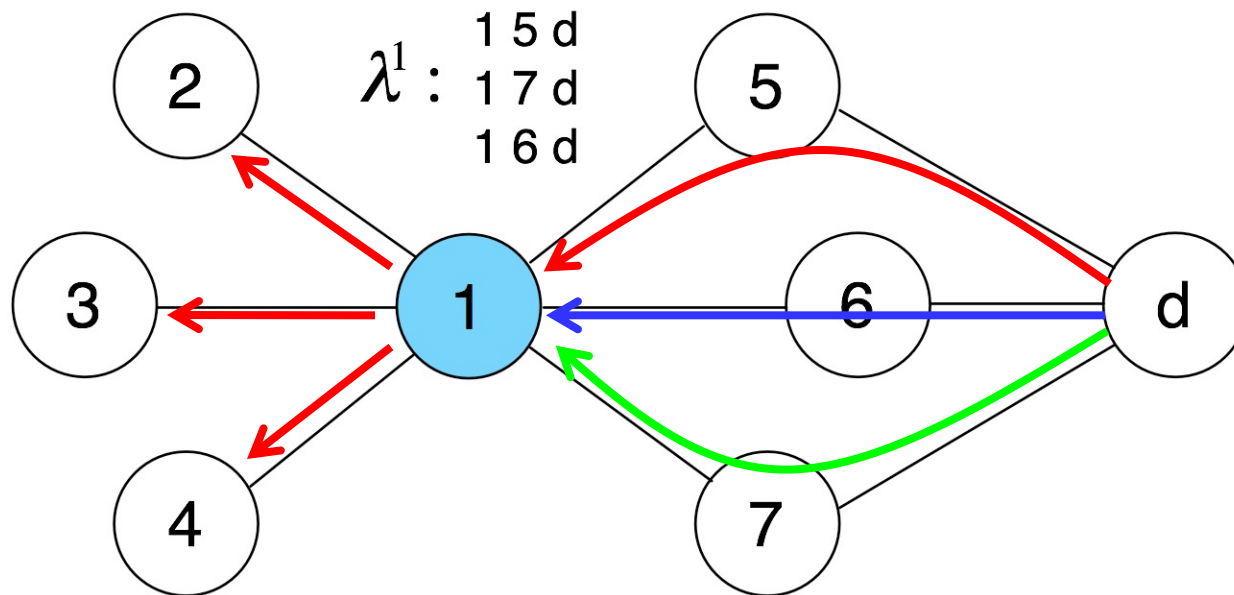
BGP Decision Process

- Default decision order 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
 - » Any criteria: BGP attributes, source address, prime number of bytes in message, ...
 - » Can have separate policy for inbound routes, installed routes and outbound routes
 - ◆ Limited only by power of vendor-specific routing language



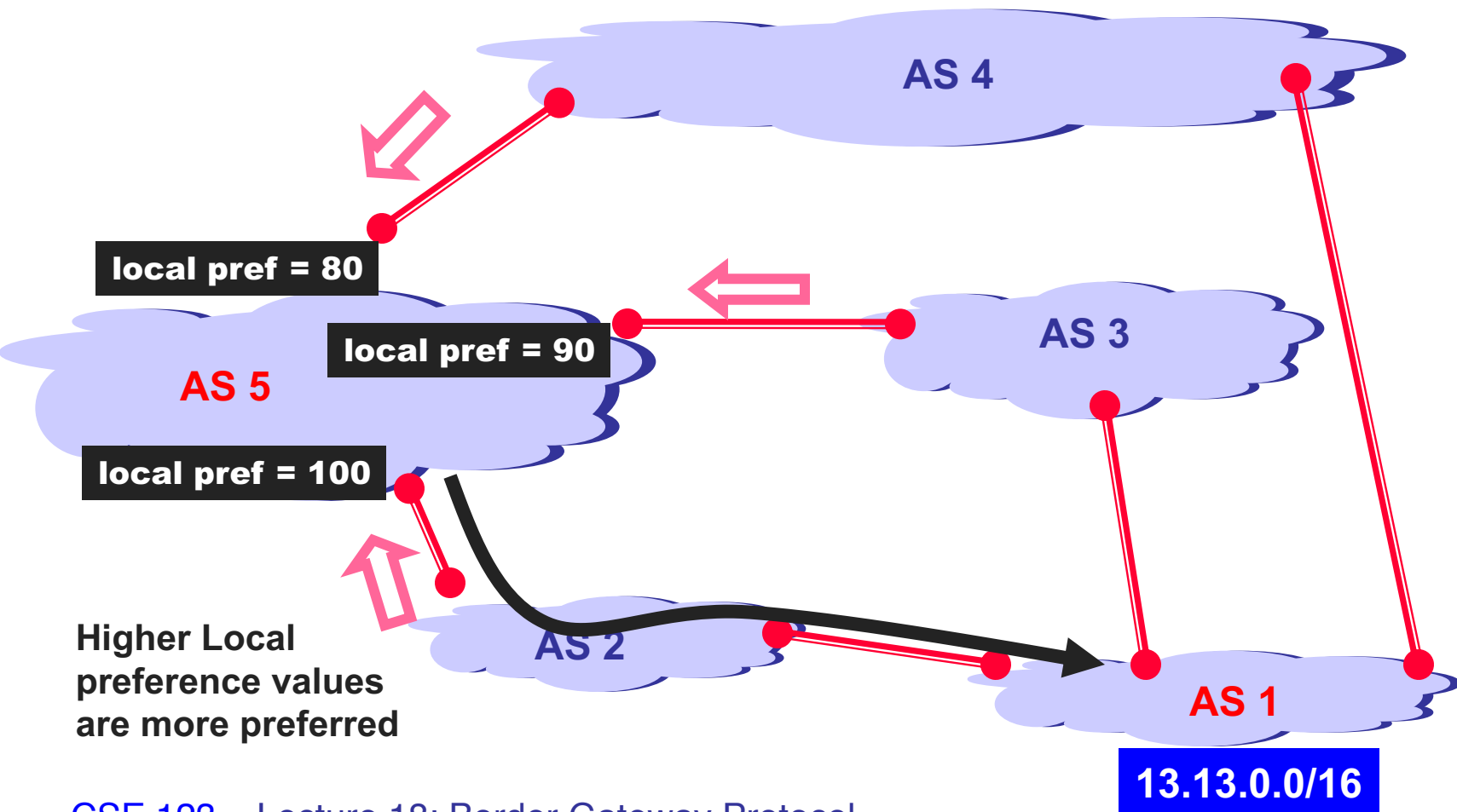
Export Only Active Routes

- In conventional path vector routing, a node has one **ranking function**, which reflects its routing policy





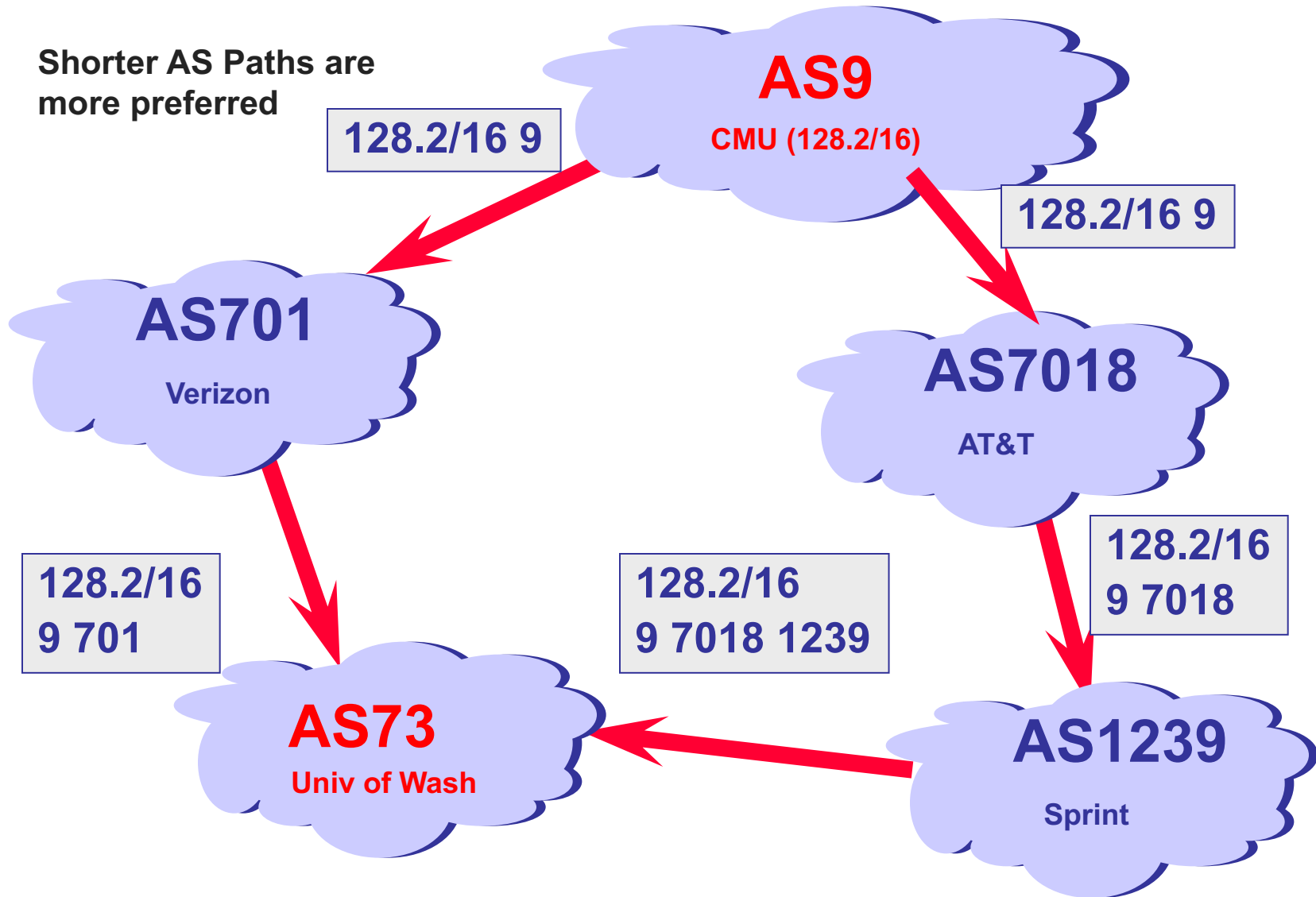
Example: Local Pref



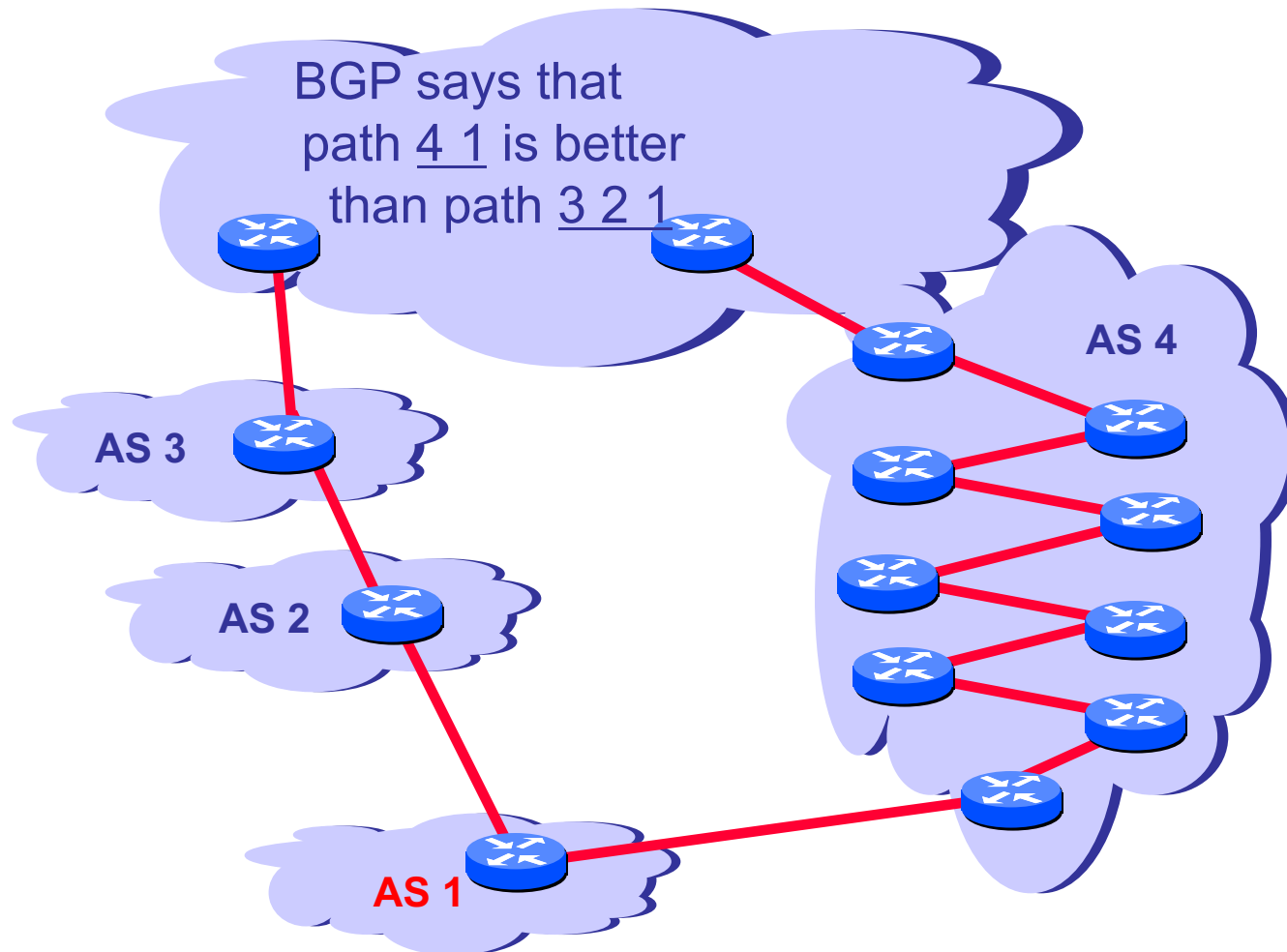


Example: Shortest AS Path

Shorter AS Paths are more preferred



AS Paths vs. Router Paths





BGP Has Lots of Problems

- Instability
 - ◆ Route flapping hurts (prefix goes up,down,up,down)
 - » Lots of recalculation!
 - ◆ Not guaranteed to converge, NP-hard to tell if it does
- Scalability still a problem
 - ◆ >500,000 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 so you don’t need info on all specific prefixes.
- Performance
 - ◆ Non-optimal, doesn’t balance load across paths

A History of Settlement (\$\$\$)



- The telephone world
 - ◆ LECs (local exchange carriers) (e.g., PacBell, NYNEX)
 - ◆ IXC (inter-exchange carriers) (e.g., Sprint, AT&T)
- LECs MUST provide IXCs access to customers
 - ◆ This is enforced by laws and 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”

AS Business Relationships



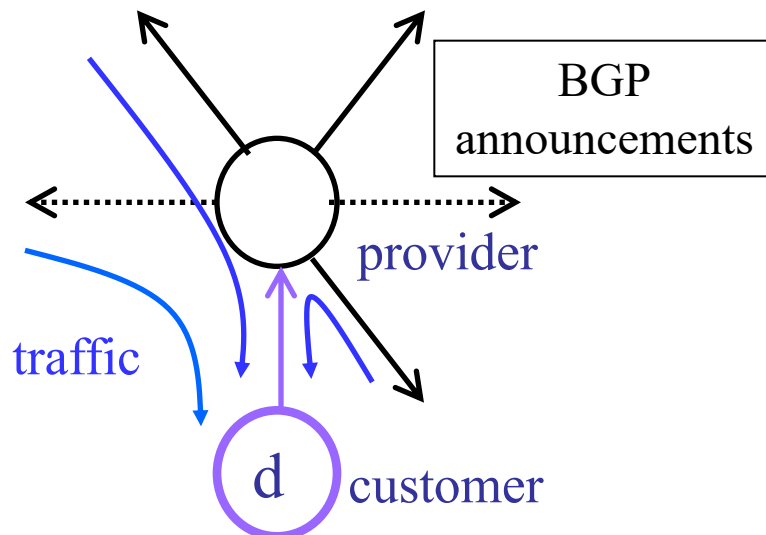
- Neighboring ASes have business contracts
 - ◆ How much traffic to carry
 - ◆ Which destinations to reach
 - ◆ How much money to pay
- Common business relationships
 - ◆ Customers pay providers (95% billing model)
 - » E.g., Princeton is a customer of USLEC
 - » E.g., MIT is a customer of Lumen
 - ◆ Peers don't pay peers (exchange equal traffic for free)
 - » E.g., Verizon is a peer of AT&T
 - » E.g., Harvard is a peer of Harvard Business School
- Routing follows the money (allow inexpensive paths)



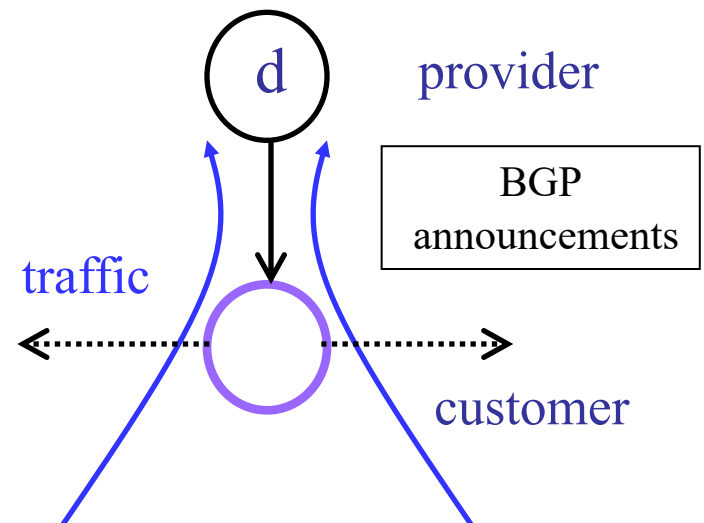
Customer/Provider

- Customer needs to be reachable from everyone
 - ◆ Provider tells all neighbors how to reach the customer
- Customer does not want to provide transit service
 - ◆ Customer does not let its providers route through it to other providers

Traffic to the customer



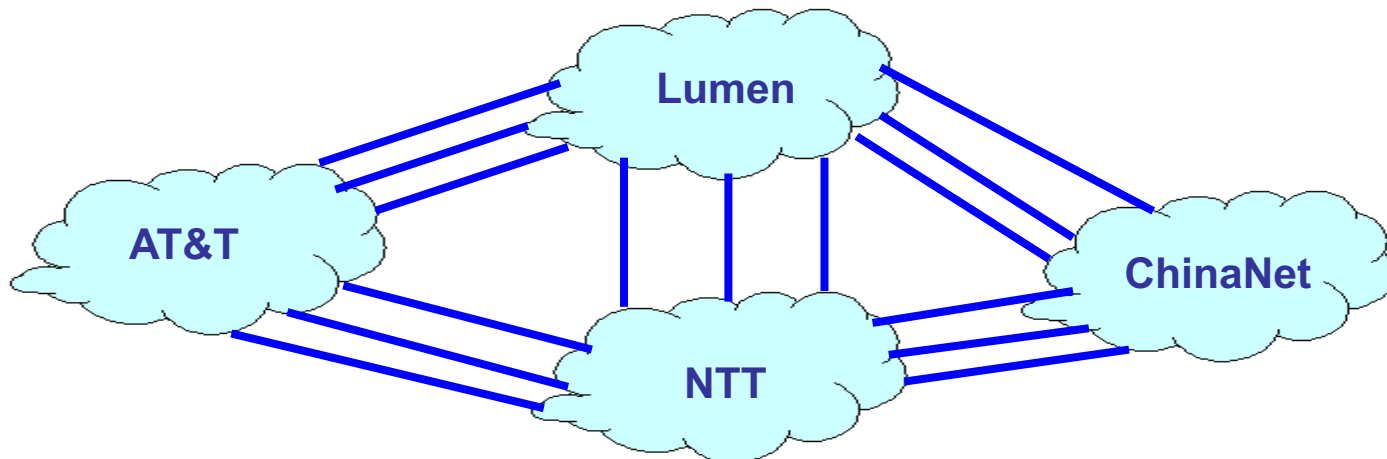
Traffic from the customer





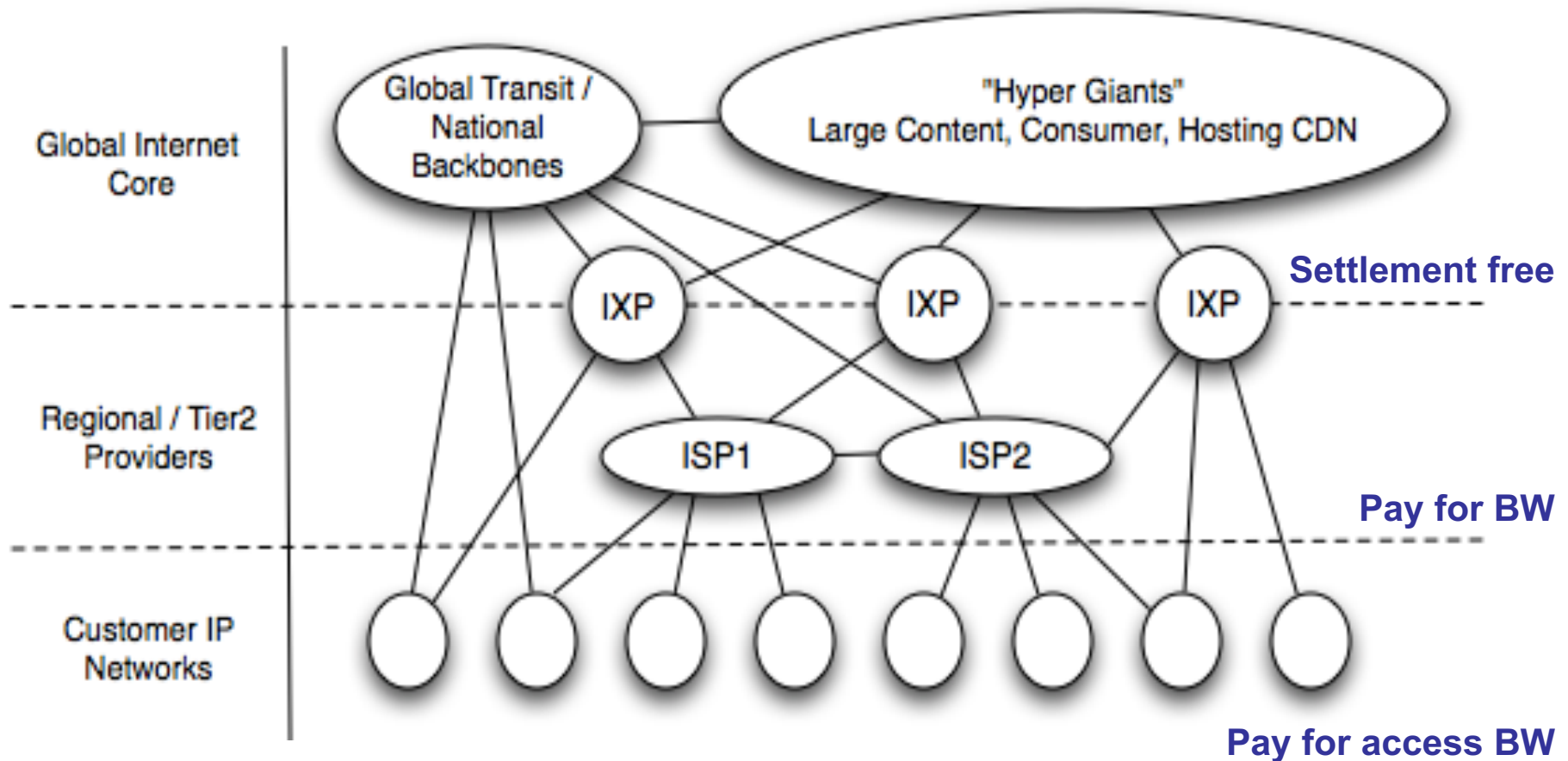
Tier-1 Providers

- Make up the “core” of the Internet
 - ◆ Has no upstream provider of its own
 - ◆ Typically has a national or international backbone
- Top of the Internet hierarchy of ~10-20 ASes
 - ◆ E.g., AT&T, Lumen (CenturyLink), NTT/Verio, Verizon, ChinaNet
 - ◆ Full peer-peer connections between tier-1 providers





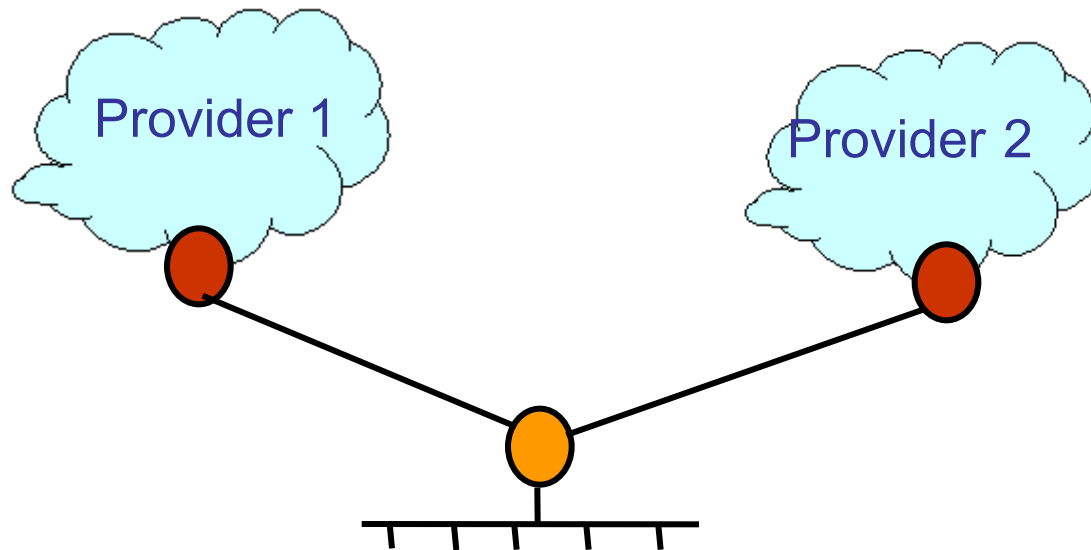
The Internet Hierarchy





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





BGP Summary

- Interdomain-routing
 - ◆ Exchange reachability information (plus hints)
 - ◆ BGP is based on path vector routing
 - ◆ 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?)
 - ◆ Can have significant impact on performance



What's in a Router?

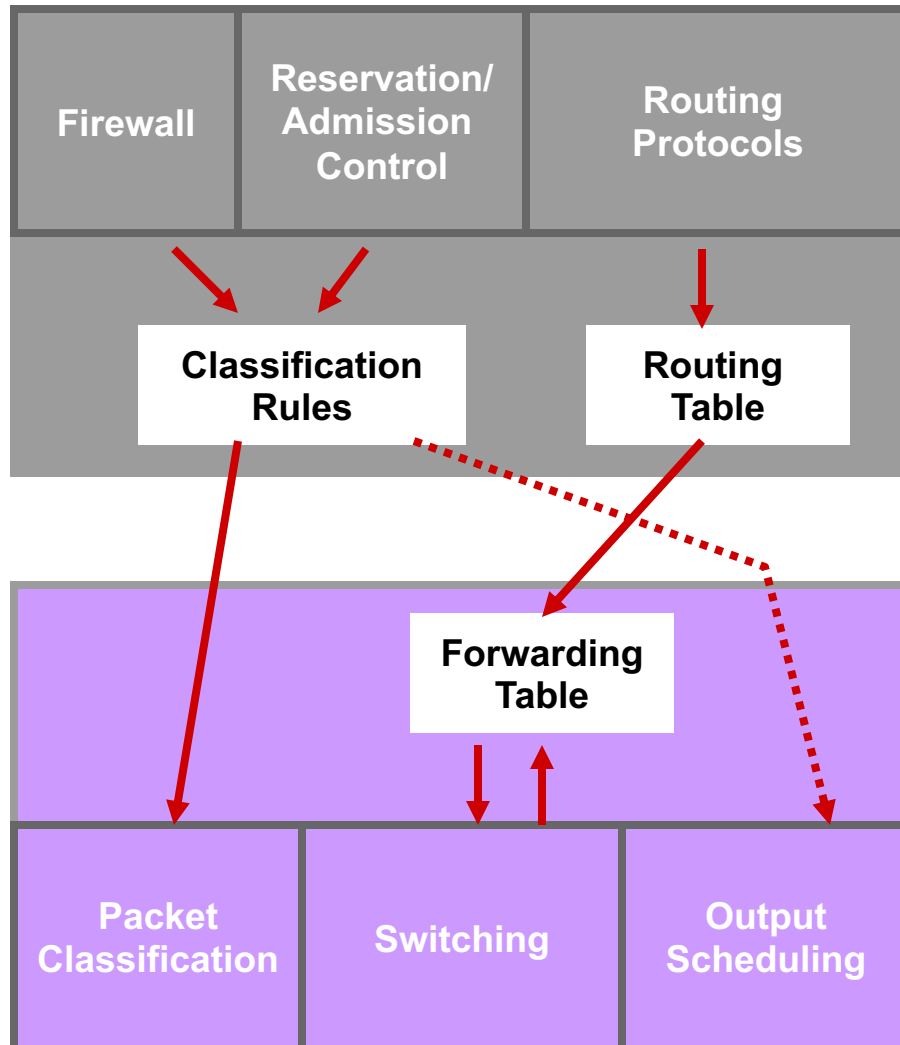
- Physical components
 - ◆ One or more **input interfaces** that receive packets
 - ◆ One or more **output interfaces** that transmit packets
 - ◆ **Computing platform** to run routing protocol
 - ◆ A **chassis** (box + power) to hold it all
- Functions
 - ◆ **Forward** packets
 - ◆ **Drop** packets (congestion, security, QoS)
 - ◆ **Delay** packets (QoS)
 - ◆ **Transform** packets? (Encapsulation, Tunneling)



Router Functions

1. Receive incoming packet from link input interface
2. Lookup packet destination in forwarding table
(destination, output port(s))
3. Validate checksum, decrement TTL, update checksum
4. Buffer packet in input queue
5. Send packet to output interface (interfaces?)
6. Buffer packet in output queue
7. Send packet to output interface link

Router functional architecture



Control Plane

- General-purpose CPU
- Complex software
- Per-control action
- May be slow

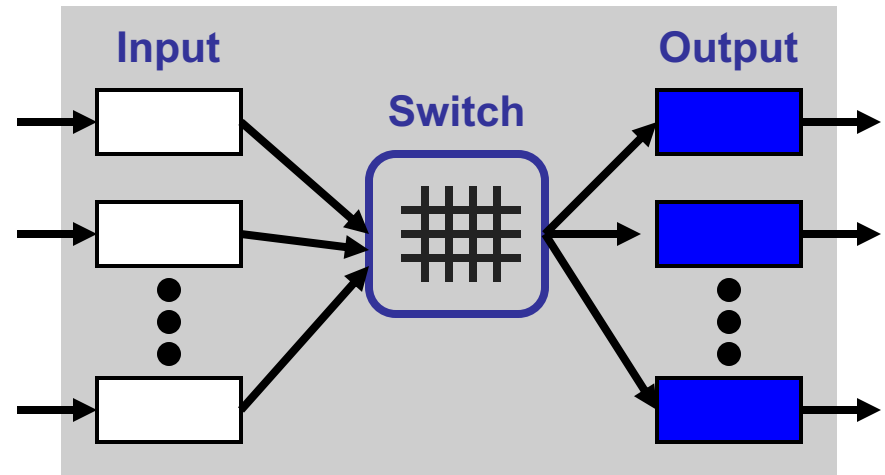
Data plane

- Simple
- Per-packet
- Must be incredibly fast (Tbps)

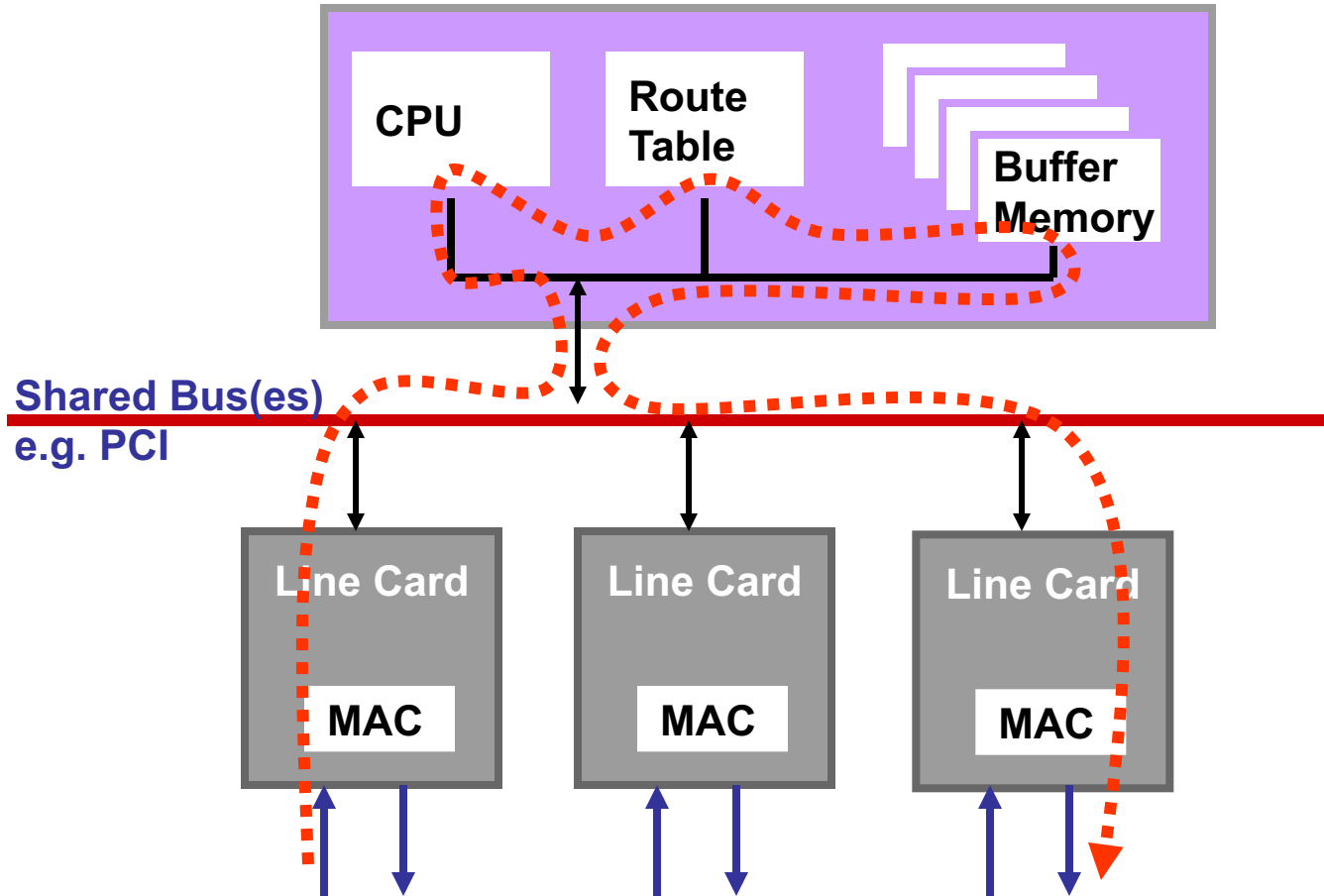


Interconnect architecture

- Input & output ports connected via switch fabric
 - ◆ Ports grouped together on “line cards”
- Kinds of switch fabric
 - ◆ Shared Memory
 - ◆ Bus
 - ◆ Crossbar
- How to deal with transient contention?
 - ◆ Input queuing
 - ◆ Output queuing



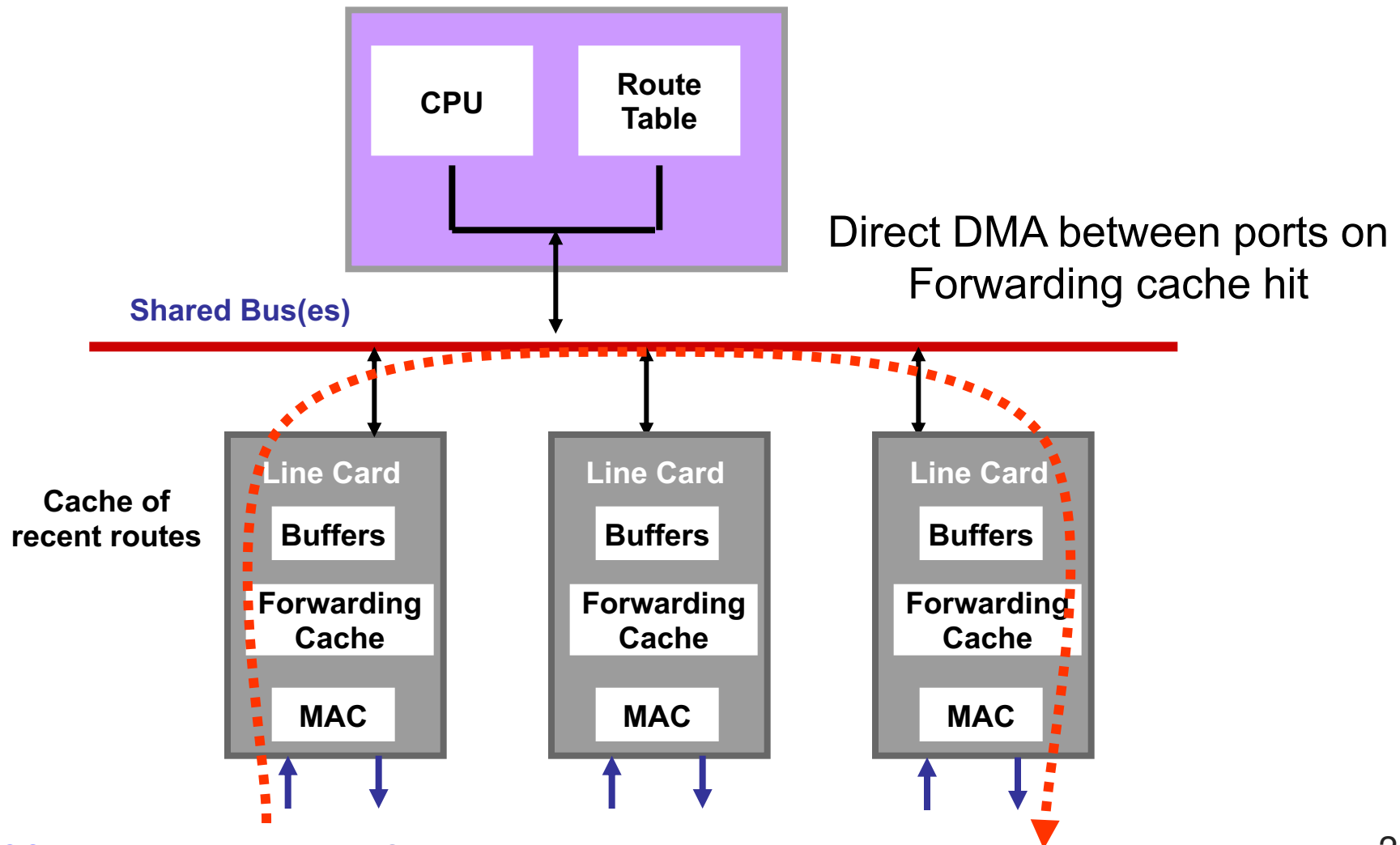
First Generation Routers (mostly software)



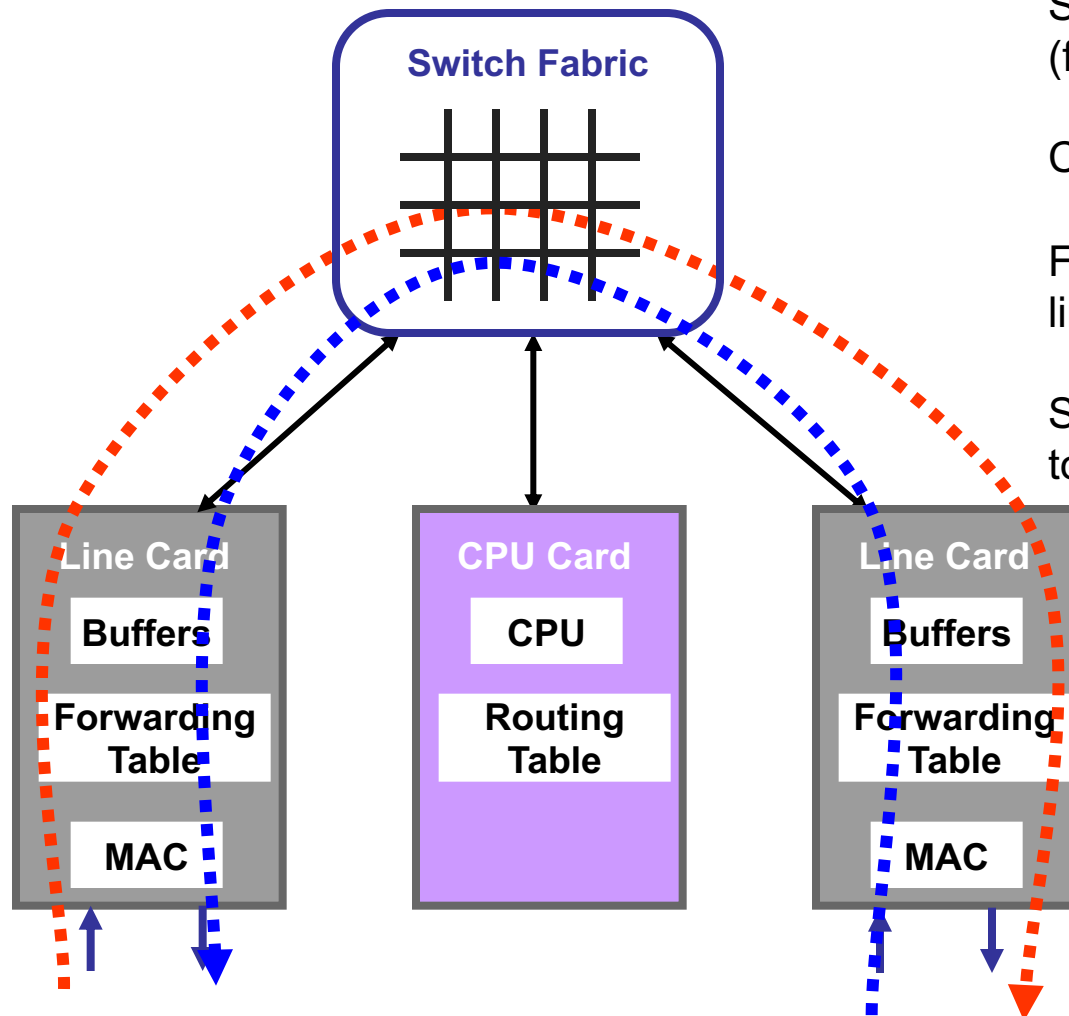
Single CPU and
shared memory;

All forwarding
hits main CPU

Second Generation Routers (HW caching forwarding table)



Third Generation Routers (improved HW shared bus)



Shared interconnect
(frequently crossbar)

Centralized scheduler

Full forwarding table in
line card

Switches fixed-sized cells
to improve performance



For Next Time

- Read P&D 3.5 (Router Implementation)