EECS 489
Computer Networks

Winter 2023

Z. Morley Mao

Material with thanks to Aditya Akella, Sugih Jamin, Philip Levis, Sylvia Ratnasamy, Peter Steenkiste, and many other colleagues.

Logistics

- Open book/text/notes, but OFFLINE
  - Except for taking the exam over the Internet
- You’re NOT allowed to write/run any programs
- You’re NOT allowed to collaborate with anyone

General guidelines (1)

- Test only assumes material covered in lecture, sections, and assignments after midterm
  - Text: only to clarify details and context for the above
- The test doesn't require you to do complicated calculations
  - Use this as a hint to determine if you're on right track
- You don't need to memorize anything
- You do need to understand how things work

General guidelines (2)

- Be prepared to:
  - Weigh design options outside of the context we studied them in
  - Contemplate new designs we haven't covered in detail but can be put together
    - e.g., I introduce a new IP address format; how does this affect..”
  - Reason from what you know about the pros/cons of solutions we did study

General guidelines (3)

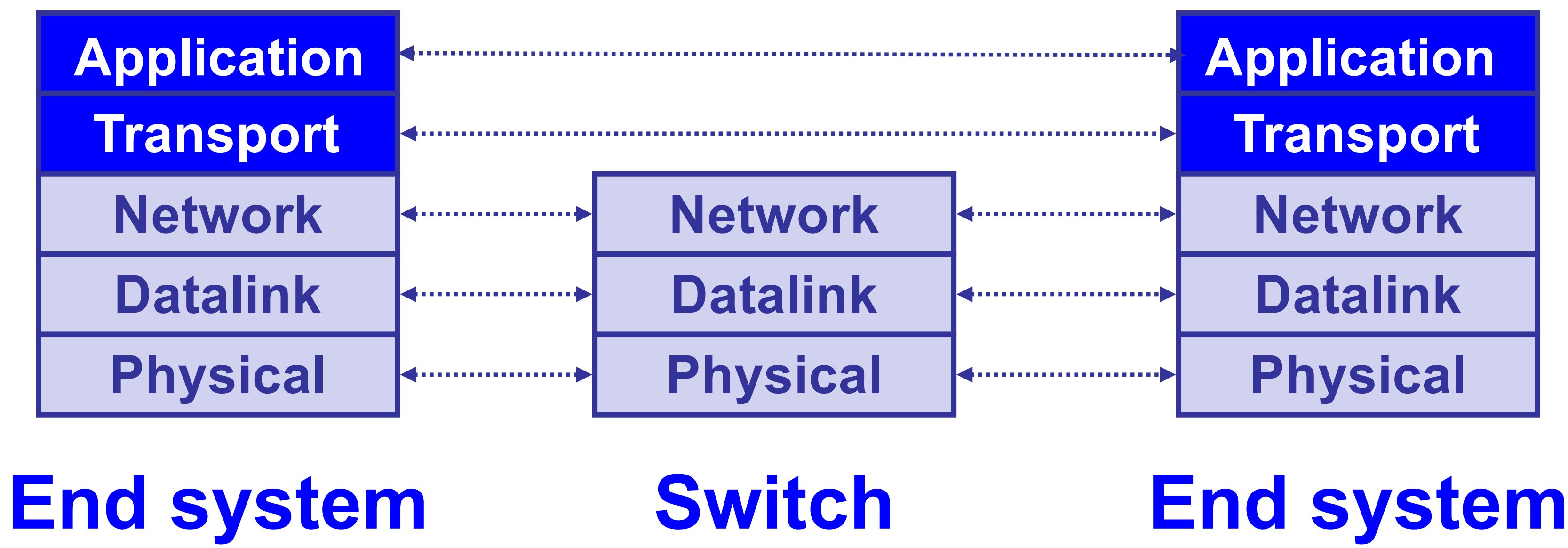
- Exam format
  - Like midterm, but we're working to avoid cascading mistakes
- Questions not ordered in terms of complexity
  - » Read all carefully
- Pace yourself accordingly!

This review

- Walk through what you're expected to know since the midterm: key topics, important aspects of each
- Not covered in review does NOT imply you don't need to know it
  - But if it's covered today, you should know it
- Summarize, not explain
  - Stop me when you want to discuss something further!

The networking stack

- Lower three layers implemented everywhere
- Top two layers implemented only at hosts

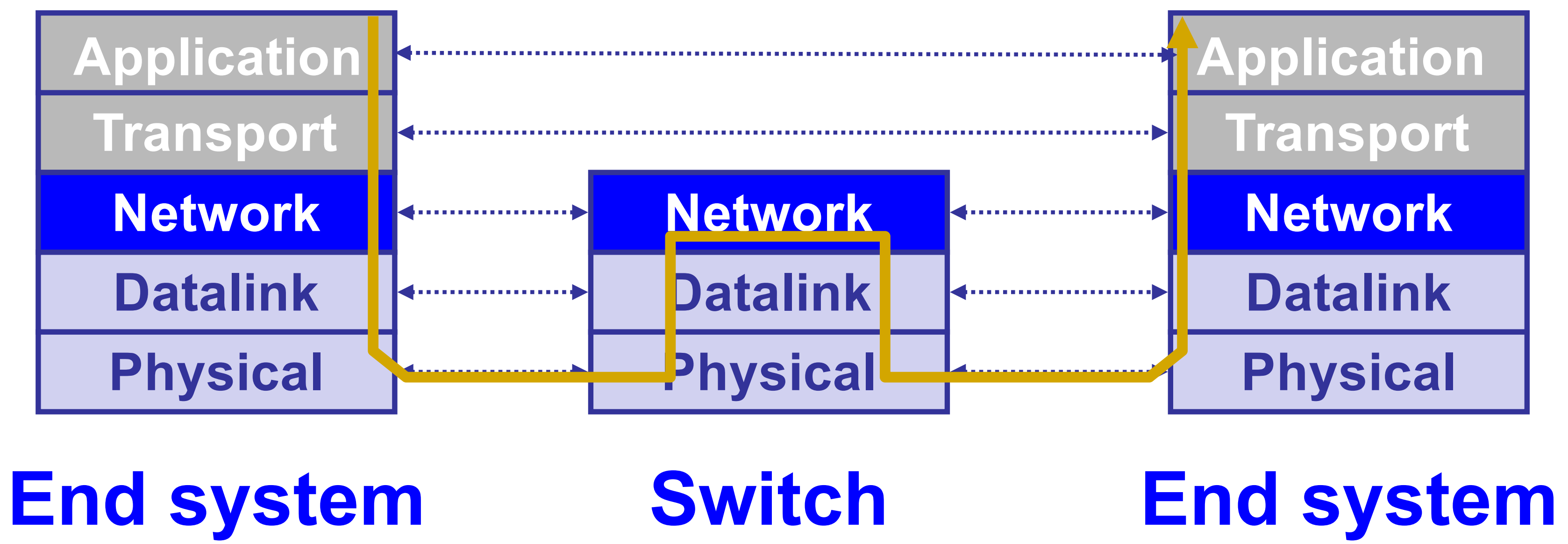


Topics

- Network layer (lectures 12–16)
  - Intra-domain routing
  - Inter-domain routing
  - SDN
- Link layer (lectures 17–19)
  - Ethernet
  - Wireless
- Datacenter networking (lectures 20)

Network layer

- Present everywhere
- Performs addressing, forwarding, and routing, among other tasks

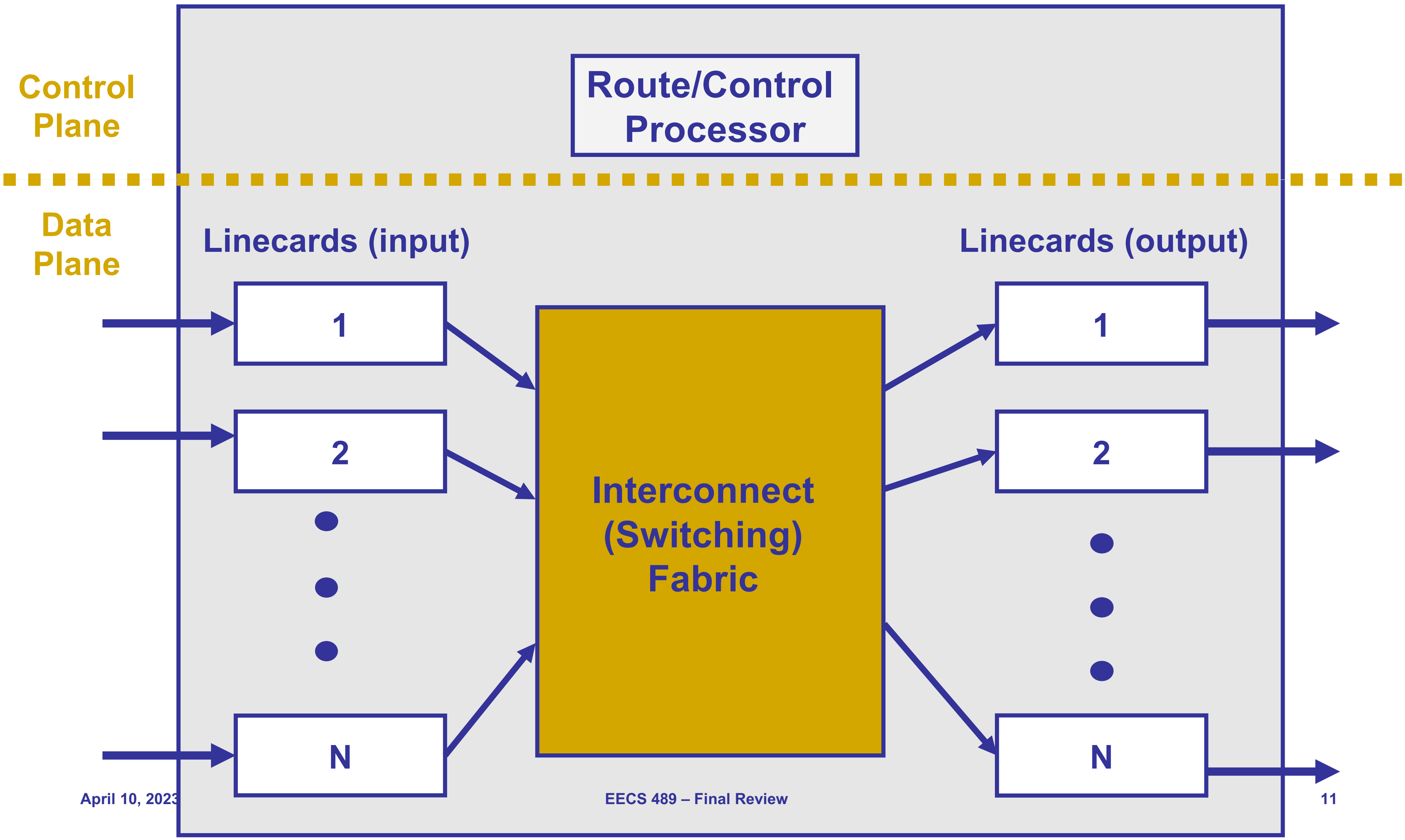


Forwarding vs. routing

- Forwarding: "data plane"
  - Directing one data packet
  - Each router using local routing state
- Routing: "control plane"
  - Computing the forwarding tables that guide packets
  - Jointly computed by routers using a distributed algorithm

- Very different timescales!

What's inside a router?



Routing: Local vs. global view

- *Local* routing state is the forwarding table in a single router
  - By itself, the state in a single router cannot be evaluated
  - It must be evaluated in terms of the global context
- *Global* state refers to the collection of forwarding tables in each of the routers
  - Global state determines which paths packets take

“Valid” routing state

- Global state is “valid” if it produces forwarding decisions that always deliver packets to their destinations
- Goal of routing protocols: compute valid state
  - How can we tell if routing state if valid?

Necessary and sufficient condition

- Global routing state is valid *if and only if*:
  - There are no dead ends (other than destination)
  - There are no loops
- A dead end is when there is no outgoing link (next-hop)
  - A packet arrives, but the forwarding decision does not yield any outgoing link
- A loop is when a packet cycles around the same set of nodes forever

Least-cost routes

- Least-cost routes provide an easy way to avoid loops
  - No reasonable cost metric is minimized by traversing a loop
- Least-cost paths form a spanning tree for each destination rooted at that destination

Intra-domain routing

- Link-state (LS) routing protocol
  - Dijkstra's algorithm
  - Broadcast neighbors' info to everyone
- Distance vector (DV) routing protocol
  - Bellman-Ford algorithm
  - Gossip to neighbors about everyone

Similarities between LS and DV routing

- Both are shortest-path based routing
  - Minimizing cost metric (link weights) a common optimization goal
    - Routers share a common view as to what makes a path “good” and how to measure the “goodness” of a path
- Due to shared goal, commonly used inside an organization
  - RIP and OSPF are mostly used for intra-domain routing

Comparison of LS and DV routing

Messaging complexity

- LS: with N nodes, E links, O(NE) messages sent
- DV: exchange between neighbors only

Speed of convergence

- LS: relatively fast
- DV: convergence time varies
  - **Count-to-infinity** problem

Robustness: what happens if router malfunctions?

- LS:
  - Node can advertise incorrect link cost
  - Each node computes its *own* table
- DV:
  - Node can advertise incorrect path cost
  - Each node's table used by others (errors propagate)

Addressing is key to scalable inter-domain routing

- Ability to aggregate addresses is crucial for
  - State: Small forwarding tables at routers
    - Much less than the number of hosts
  - Churn: Limited rate of change in routing tables

Classful addressing

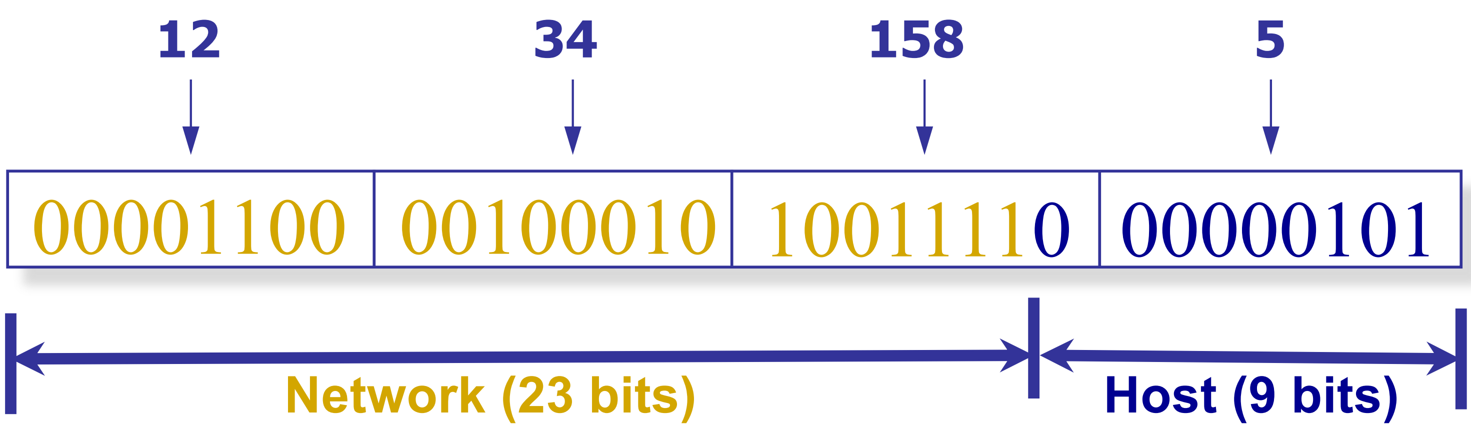
- Three classes
  - 8-bit network prefix (Class A),
  - 16-bit network prefix (Class B), or
  - 24-bit network prefix (Class C)
- Example: an organization needs 500 addresses.
  - A single class C address is not enough (<500 hosts)
  - Instead, a class B address is allocated (~65K hosts)
    - » Huge waste!

CIDR: Classless inter-domain routing

- Flexible division between network and host addresses
- Offers a better tradeoff between size of the routing table and efficient use of the IP address space

Hierarchy in IP addressing

- 32 bits are partitioned into a prefix and suffix components
- Prefix is the network component; suffix is the host component

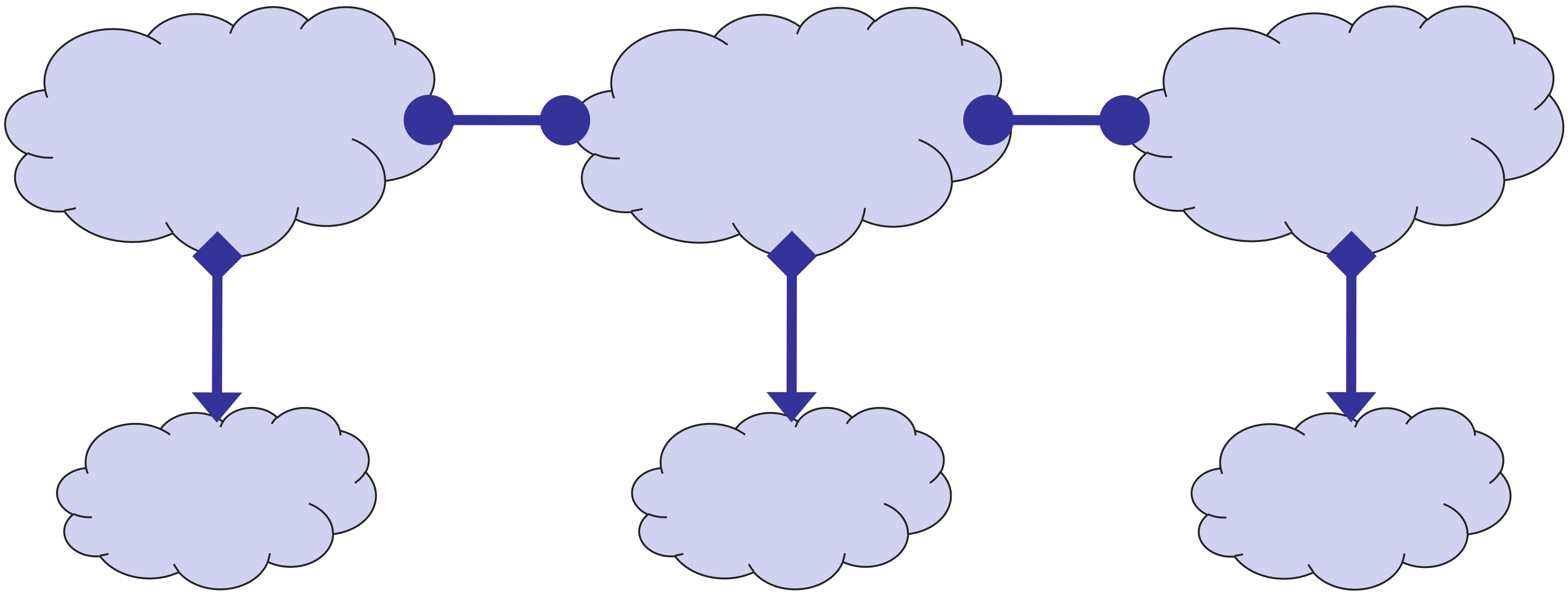


- Inter-domain routing operates on network prefix

Administrative structure shapes Inter-domain routing

- ASes want freedom to pick routes based on policy
- ASes want autonomy
- ASes want privacy

Business relationships



Relations between ASes

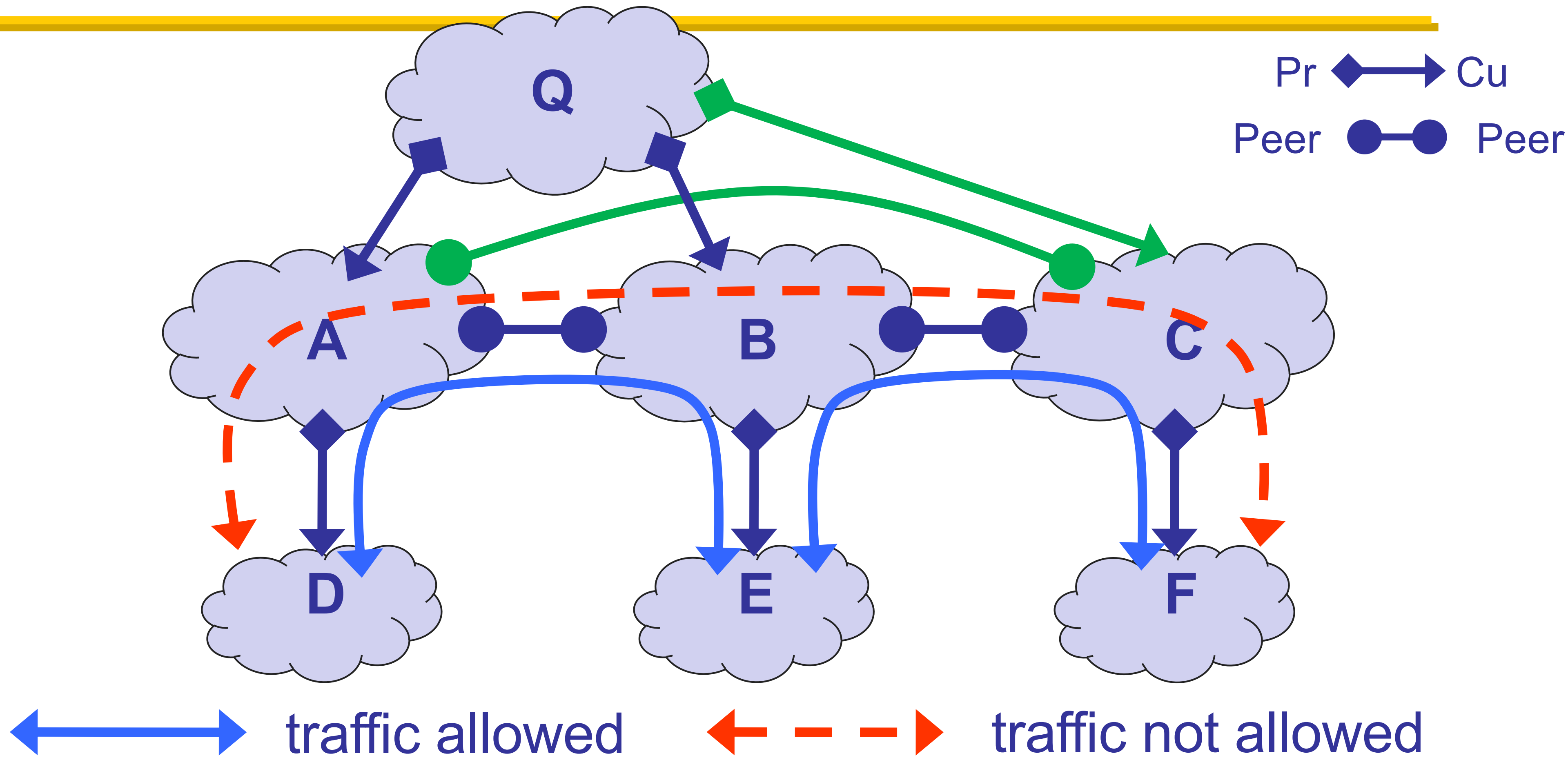
provider ◆⟶ customer

peer ●—● peer

Business implications

- Customers pay provider
- Peers don't pay each other

Routing follows the money!



- ASes provide “transit” between their customers
- Peers do not provide transit between other peers

BGP inspired by Distance-Vector with four differences

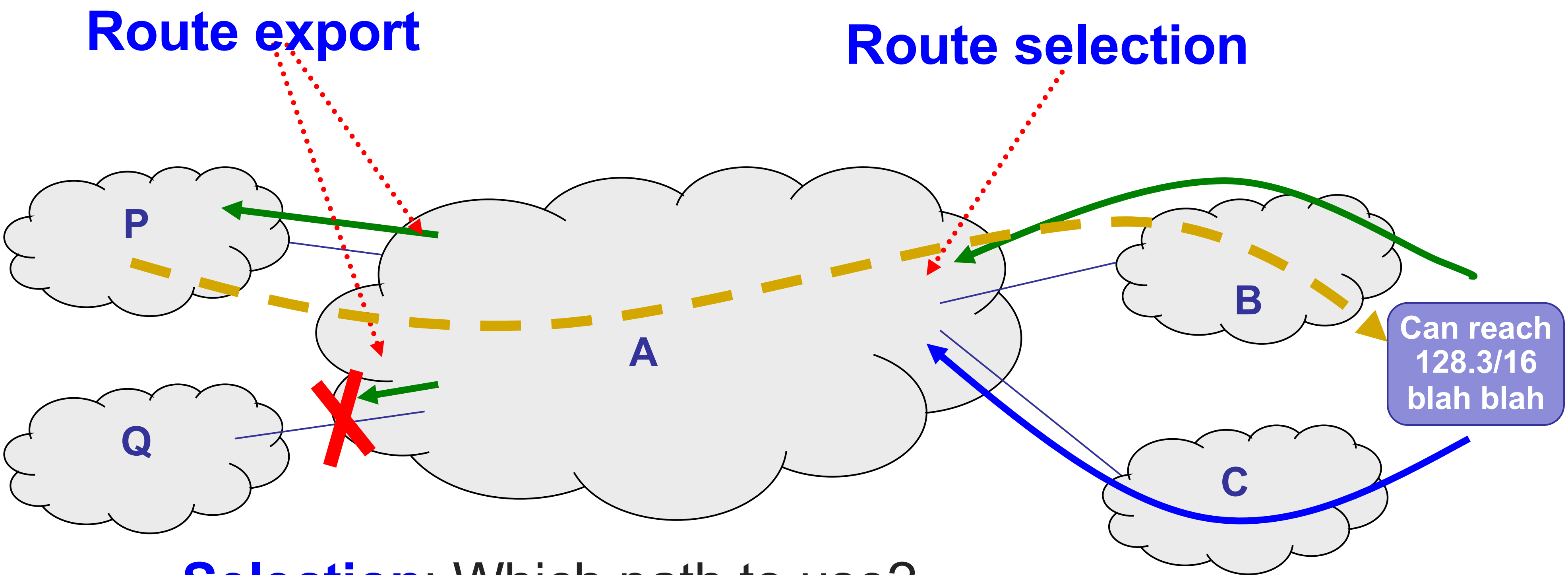
- Shortest-path routes may not be picked to enforce policy
- Path-Vector routing to avoid loops
- Selective route advertisement may affect reachability
- Routes may be aggregated for scalability

BGP: Basic idea

An AS advertises ("exports") its best routes to one or more IP prefixes

Each AS selects the "best" route it hears advertised for a prefix

Policy dictates how routes are "selected" and "exported"



- **Selection**: Which path to use?
  - Controls whether/how traffic leaves the network
- **Export**: Which path to advertise?
  - Controls whether/how traffic enters the network

Typical export policy

Destination prefix advertised by…	Export route to…
Customer	Everyone (providers, peers, other customers)
Peer	Customers
Provider	Customers

We'll refer to these as the "Gao-Rexford" rules (capture common – but not required! – practice)

Selection using attributes

- Rules for route selection in priority order

Priority	Rule	Remarks
1	LOCAL PREF	Pick highest LOCAL PREF
2	ASPATH	Pick shortest ASPATH length
3	MED	Lowest MED preferred
4	eBGP > iBGP	Did AS learn route via eBGP (preferred) or iBGP?
5	iBGP path	Lowest IGP cost to next hop (egress router)
6	Router ID	Smallest next-hop router's IP address as tie-breaker

eBGP, iBGP, and IGP

- eBGP: BGP sessions between border routers in different ASes
  - Learn routes to external destinations
- iBGP: BGP sessions between border routers and other routers within the same AS
  - Distribute externally learned routes internally
- IGP: “Interior Gateway Protocol” = Intra-domain routing protocol
  - Provide internal reachability via shortest path
  - E.g., OSPF, RIP

5-MINUTE BREAK!

Announcements

- Final exam date and time:
  - 90 Min; Online (same setup as Midterm exam)
  - Friday April 21: 1:30 pm - 3:00 pm
  - From registrar's final exam schedule
  - For those needing extra time: same start time (longer exam time)

“The Power of Abstraction”

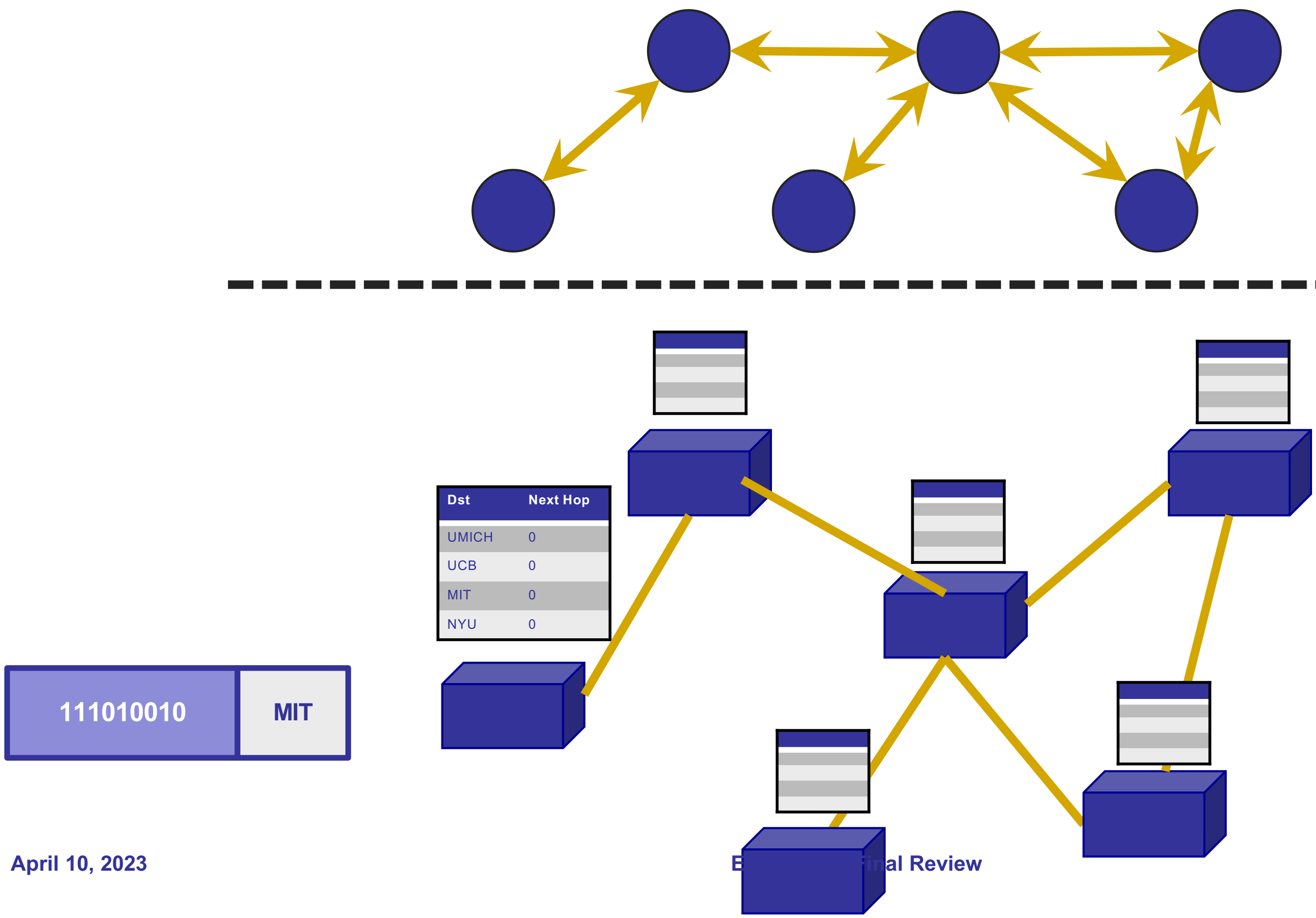
- “Modularity based on abstraction is the way things get done”
  - Barbara Liskov
- Abstractions → Interfaces → Modularity

Separate concerns with abstractions

- Be compatible with low-level hardware/software
  - Need an abstraction for general forwarding model
- Make decisions based on entire network
  - Need an abstraction for network state
- Compute configuration of each physical device
  - Need an abstraction that simplifies configuration

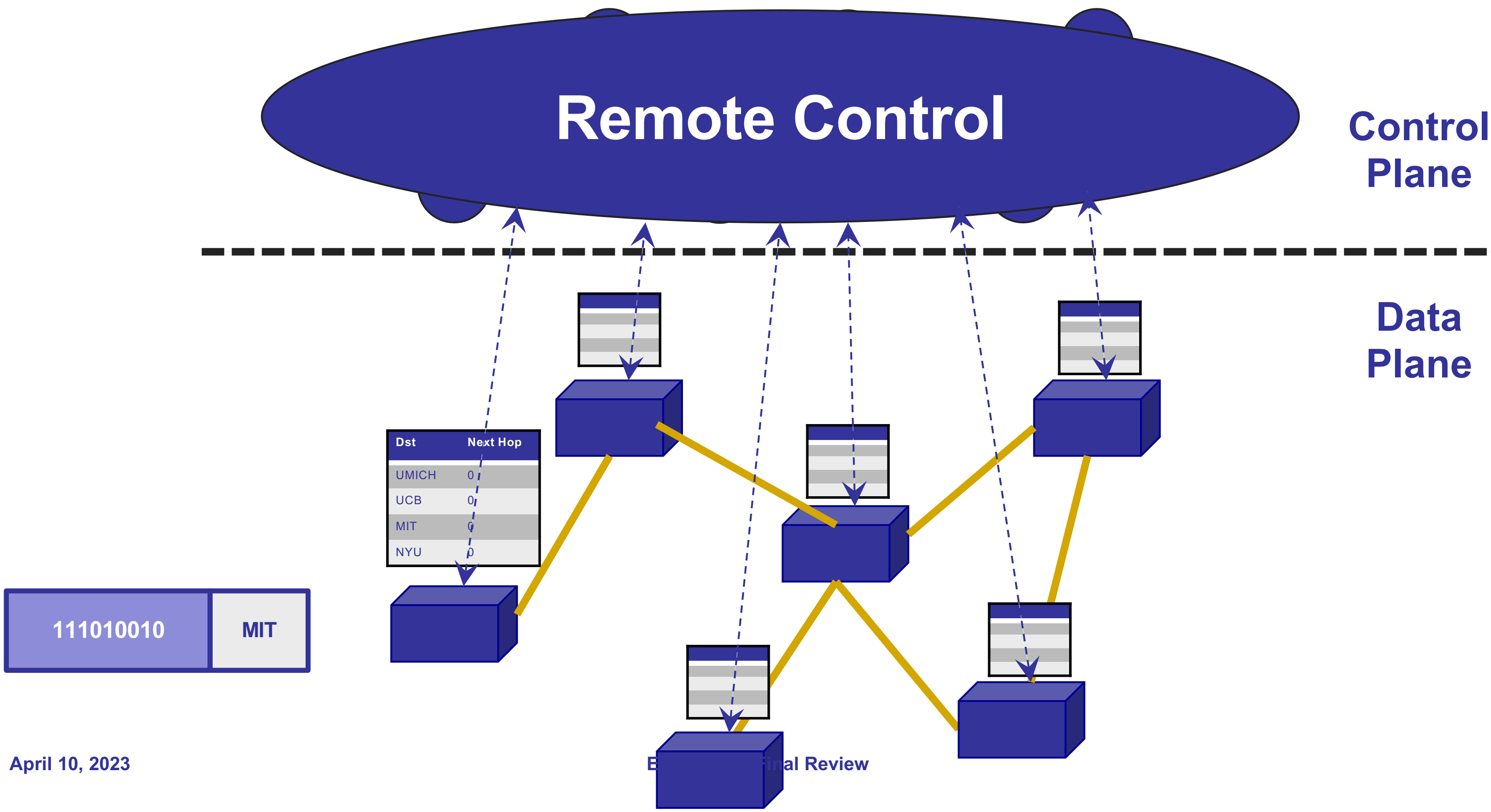
Traditional fully decentralized control plane

- Individual routing algorithm components in every router interact in the control plane



Logically centralized control plane

- A distinct (typically remote) controller interacts with local control agents (CAs)



SDN: Many challenges remain

- Hardening the control plane: dependable, reliable, performance-scalable, secure distributed system
  - Robustness to failures: leverage strong theory of reliable distributed system for control plane
  - Security: “baked in” from day one?
- Networks, protocols meeting mission-specific requirements
  - E.g., real-time, ultra-reliable, ultra-secure
- Internet-scaling

Fixed-function data plane

- Traditional switches are fixed-function
  - They can do whatever they can do at birth, but they cannot change!
  - Bottom-up design

- Even OpenFlow was designed to be a fixed protocol
  - With a fixed table format
  - Capable of doing limited things

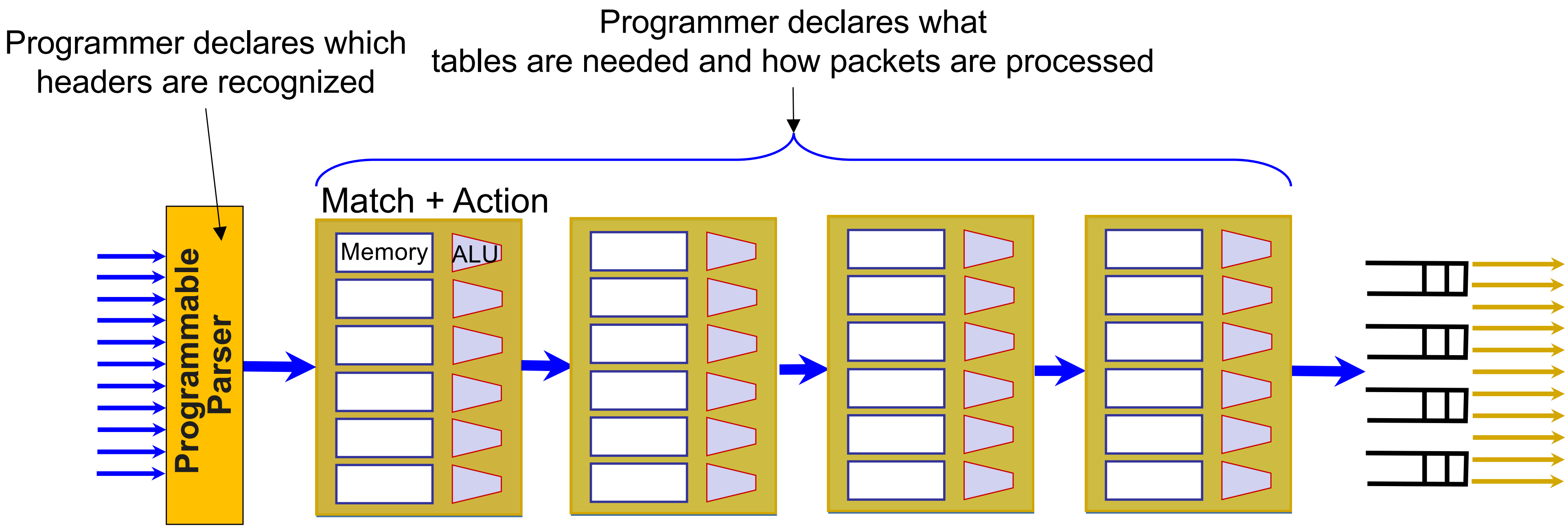
Programmable data plane

- What if we could tell switches exactly what we want?
  - What table to keep?
  - What rules to use?
  - What data to keep track of?
  - …

Top-down workflow

- Precisely specify using a well-defined language
- Compile it down to run on a standardized hardware (e.g., using P4)
- Run at line speed

PISA: Protocol Independent Switch Architecture



All stages are identical – makes PISA a good "compiler target"

Topics

- Network layer (lectures 12–16)
  - Intra-domain routing
  - Inter-domain routing
  - SDN
- Link layer (lectures 17–19)
  - Ethernet
  - Wireless
- Datacenter networking (lectures 20)

Data link layer

- Provides four primary services
  - Framing
    - Encapsulates network layer data
  - Link access
    - Medium access control (MAC) protocol defines when to transmit frames
  - Reliable delivery
    - Primarily for mediums with high error rates (e.g., wireless)
  - Error detection and correction

Point-to-point vs. broadcast medium

- Point-to-point: dedicated pairwise communication
  - E.g., long-distance fiber link
  - E.g., Point-to-point link b/n Ethernet switch and host
- Broadcast: shared wire or medium
  - Traditional Ethernet (pre ~2000)
  - 802.11 wireless LAN

Random access MAC protocols

- When node has packet to send
  - Transmit at full channel data rate **w/o** coordination
- Two or more transmitting nodes $\Rightarrow$ collision
  - Data lost
- Random access MAC protocol specifies
  - How to detect and recover from collisions
- Examples
  - ALOHA and Slotted ALOHA
  - CSMA, CSMA/CD, CSMA/CA (wireless)

CSMA (Carrier Sense Multiple Access)

- CSMA: listen before transmit
  - If channel sensed idle: transmit entire frame
  - If channel sensed busy, defer transmission
- Human analogy: don’t interrupt others!
- Does not eliminate all collisions
  - Why?

CSMA/CD (Collision Detection)

- CSMA/CD: carrier sensing, deferral as in CSMA
  - Collisions detected within short time
  - Colliding transmissions aborted, reducing wastage
- Collision detection easy in wired (broadcast) LANs
  - Compare transmitted, received signals
- Collision detection difficult in wireless LANs

Limits on CSMA/CD network length

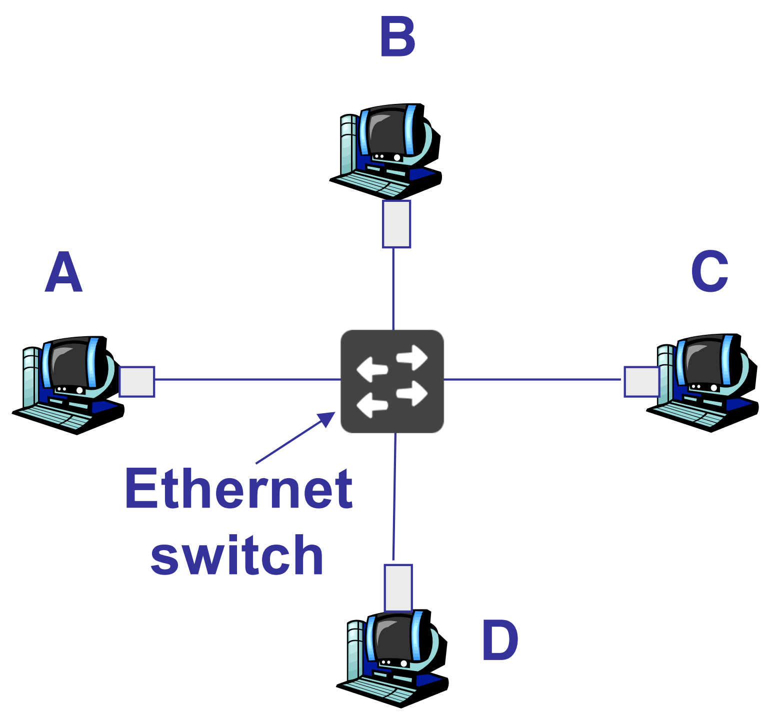
- Latency depends on physical length of link
  - Time to propagate a frame from one end to other
- Suppose A sends a frame at time **t**
  - And B sees an idle line at a time just before **t + d**
  - … so B happily starts transmitting a frame
- B detects a collision, and sends jamming signal
  - But A cannot see collision until **t + 2d**

Limits on CSMA/CD network length

- A needs to wait for time **2d** to detect collision
  - So, A should keep transmitting during this period
  - AND keep an eye out for a possible collision
- Imposes restrictions; e.g., for 10 Mbps Ethernet
  - Maximum length of the wire: 2,500 meters
  - Minimum length of a frame: 512 bits (64 bytes)

Why switched Ethernet?

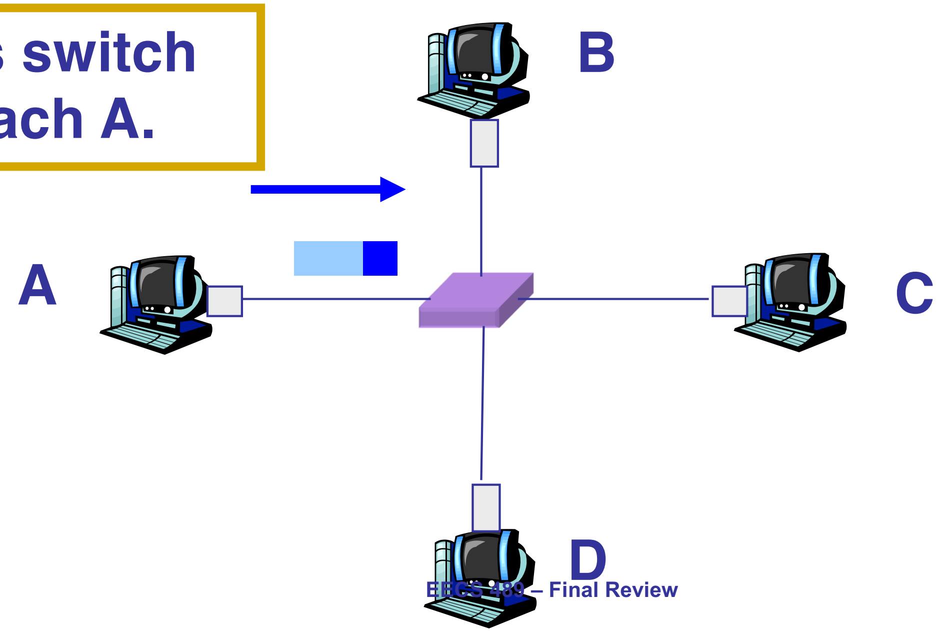
- Enables concurrent communication
  - Host A can talk to C, while B talks to D
  - No collisions and no need for CSMA/CD
  - No constraints on link lengths, etc.



Ethernet switches are "self learning"

- When a packet arrives:
  - Inspect source MAC address, associate with incoming port
  - Store mapping in the switch table
  - Use time-to-live field to eventually forget mapping

Packet tells switch how to reach A.



ARP and DHCP

- Link layer discovery protocols
  - ARP → Address Resolution Protocol
  - DHCP → Dynamic Host Configuration Protocol
  - Confined to a single local-area network (LAN)
  - Rely on broadcast capability

Key ideas in both ARP and DHCP

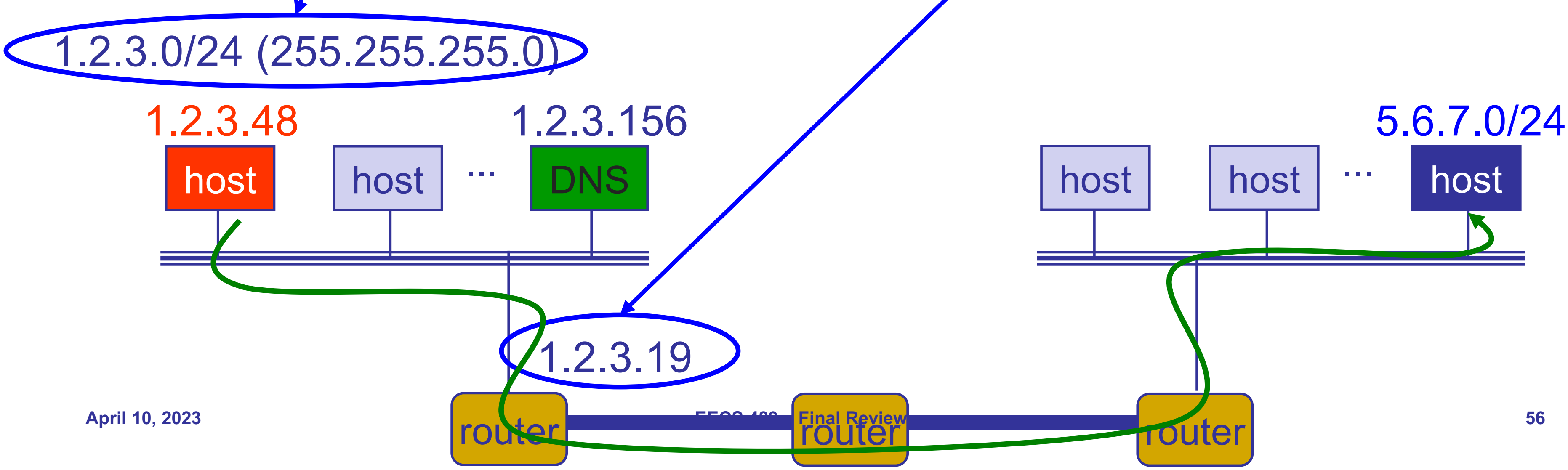
- Broadcasting: Can use broadcast to make contact
  - Scalable because of limited size
- Caching: remember the past for a while
  - Store the information you learn to reduce overhead
- Soft state: eventually forget the past
  - Associate a time-to-live field with the information
  - … and either refresh or discard the information
  - Key for robustness in the face of unpredictable change

ARP: Address Resolution Protocol

- Every host maintains an ARP table
  - List of (IP address → MAC address) pairs
- Consult the table when sending a packet
  - Map dest. IP address to dest. MAC address
  - Encapsulate (IP) data packet with MAC header; xmit
- What if IP address not in the table?
  - Sender broadcasts: Who has IP address 1.2.3.156?
  - Receiver replies: MAC address 58-23-D7-FA-20-B0
  - Sender caches result in its ARP table

What if the destination is remote?

- Look up the MAC address of the first hop router
  - 1.2.3.48 uses ARP to find MAC address for first-hop router **1.2.3.19** rather than ultimate destination IP address
- How does the red host know the destination is not local?
  - Uses netmask (discovered via DHCP)
- How does the red host know about 1.2.3.19?
  - Also DHCP



Wireless link characteristics

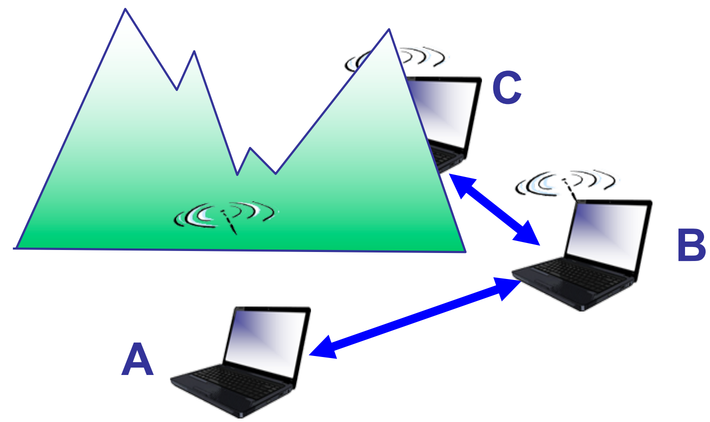
- Three important differences from wired link …
  - Decreased signal strength: Radio signal attenuates as it propagates through matter (path loss)
  - Multipath propagation: Radio signal reflects off objects ground, arriving at destination at slightly different times
  - Interference from other sources: Standardized wireless network frequencies (e.g., 2.4 GHz) shared by other devices (e.g., phone); devices (motors) interfere as well
- … make communication across (even a point-to-point) wireless link much more “difficult”

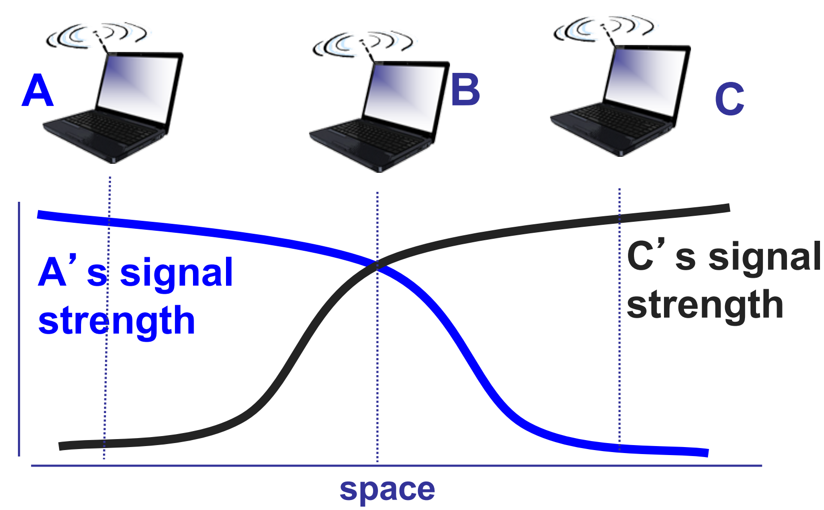
Wireless network characteristics

- Multiple wireless senders and receivers create many problems
  - Multiple access issues (we've seen this before)
  - Hidden terminal problem

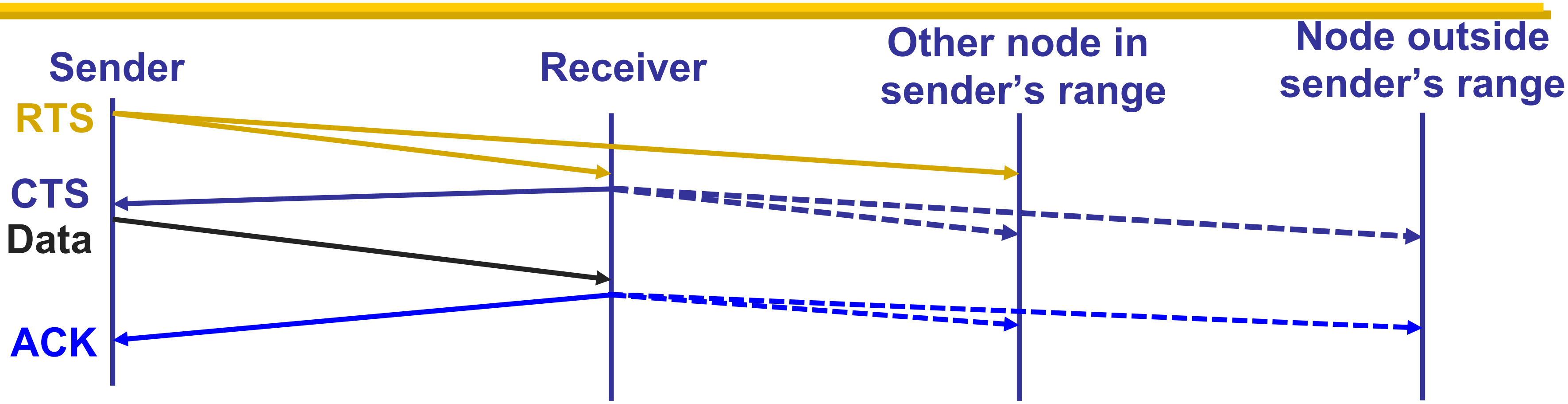
Hidden terminal problem

- B, A hear each other
- B, C hear each other
- A, C can not hear each other
- Hence, A, C are unaware of their interference at B





CSMA/CA



- Before every data transmission
  - Sender sends a Request to Send (RTS) frame with the length of transmission and the destination
  - Receiver respond with a Clear to Send (CTS) frame
  - Sender sends data
  - Receiver sends an ACK
- If sender doesn't get a CTS back, it assumes collision

Topics

- Network layer (lectures 12–16)
  - Intra-domain routing
  - Inter-domain routing
  - SDN
- Link layer (lectures 17–19)
  - Ethernet
  - Wireless
- Datacenter networking (lectures 20)

Datacenter applications

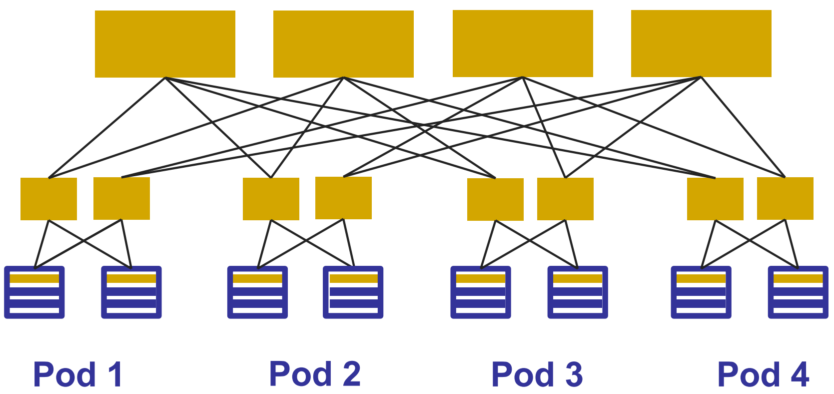
- Common theme: parallelism
  - Applications decomposed into tasks
  - Running in parallel on different machines
- Two common paradigms
  - Partition-Aggregate
  - Map-Reduce

Datacenter traffic characteristics

- Two key characteristics
  - Most flows are small
  - Most bytes come from large flows
- Applications want
  - High bandwidth (large flows)
  - Low latency (small flows)

Clos topology

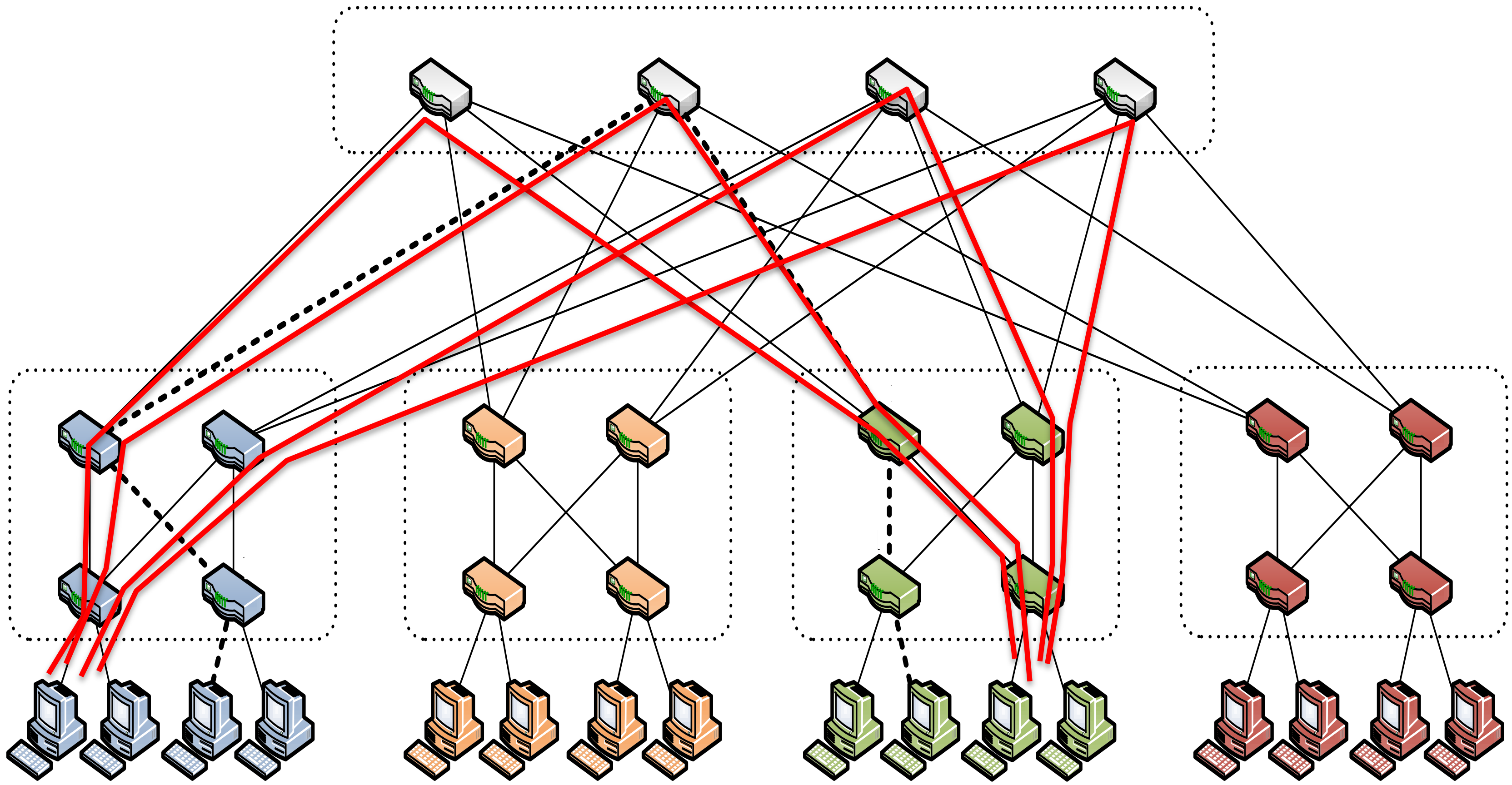
- Multi-stage network
- k pods, where each pod has two layers of k/2 switches
  - k/2 ports up and k/2 down
- All links have the same b/w
- At most $k^3/4$ machines
- Example
  - k = 4
  - 16 machines
- For k=48, 27648 machines



Datacenter networking stack

- Networking in modern datacenters
  - L2/L3 design
    - Addressing / routing / forwarding in the Fat-Tree
  - L4 design
    - Transport protocol design (w/ Fat-Tree)
  - L7 design
    - Exploiting application-level information (w/ Fat-Tree)

Using multiple paths well



L2/L3 highlights

- Load balancing while forwarding
  - Per-packet
  - Per-flow
- Hard-coded addressing or via indirection
- Modified LS/DV or source routing

L4 highlights

- Tension between high throughput and low latency requirements
  - Deep queues vs shallow queues
- DCTCP
  - React early, quickly, and with certainty using ECN
  - React in proportion to the extent of congestion, not its presence

L7 highlights

- What do applications care about?
  - Flow completion time (FCT)
  - Coflow completion time (CCT)
    - A coflow is a collection of flows with a shared application-level objective
  - We should strive to optimize as close an objective as possible to the application

Summary

- THANK YOU SO MUCH!!!