CS 152
Computer Architecture and Engineering

Lecture 25 – Routers

2005-4-21

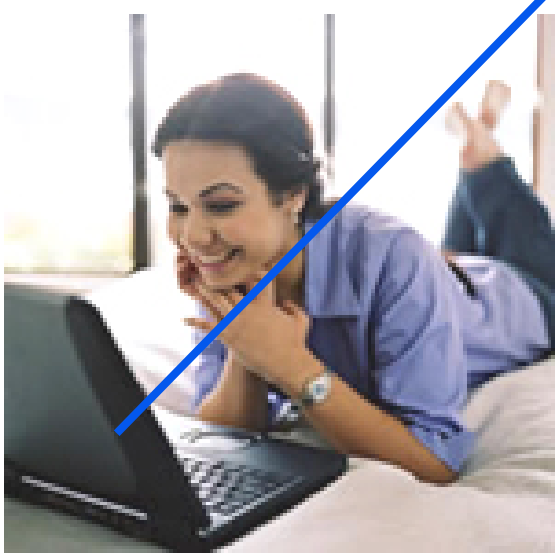
John Lazzaro
(www.cs.berkeley.edu/~lazzaro)

TAs: Ted Hong and David Marquardt

www-inst.eecs.berkeley.edu/~cs152/

Last Time: Internet Architecture

ISO Layer Names:
- IP packet: "Layer 3"
- WiFi and Cable Modem packets: "Layer 2"
- Radio/cable waveforms: "Layer 1"



801.11b WiFi packet

IP Packet

For this "hop", IP packet sent "inside" of a wireless 802.11b packet.

Cable modem packet

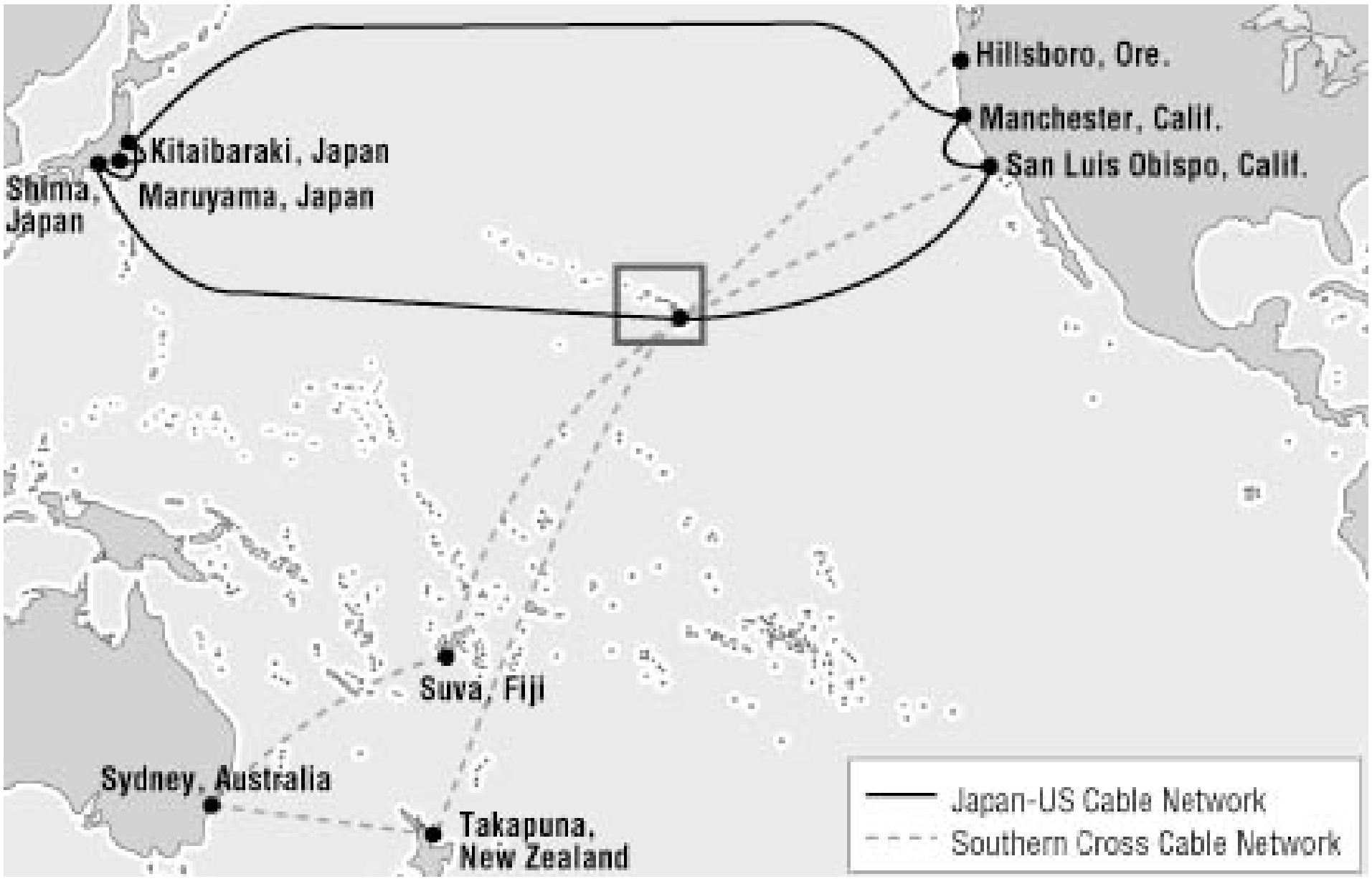
IP Packet

For this "hop", IP packet sent "inside" of a cable modem DOCSIS packet.

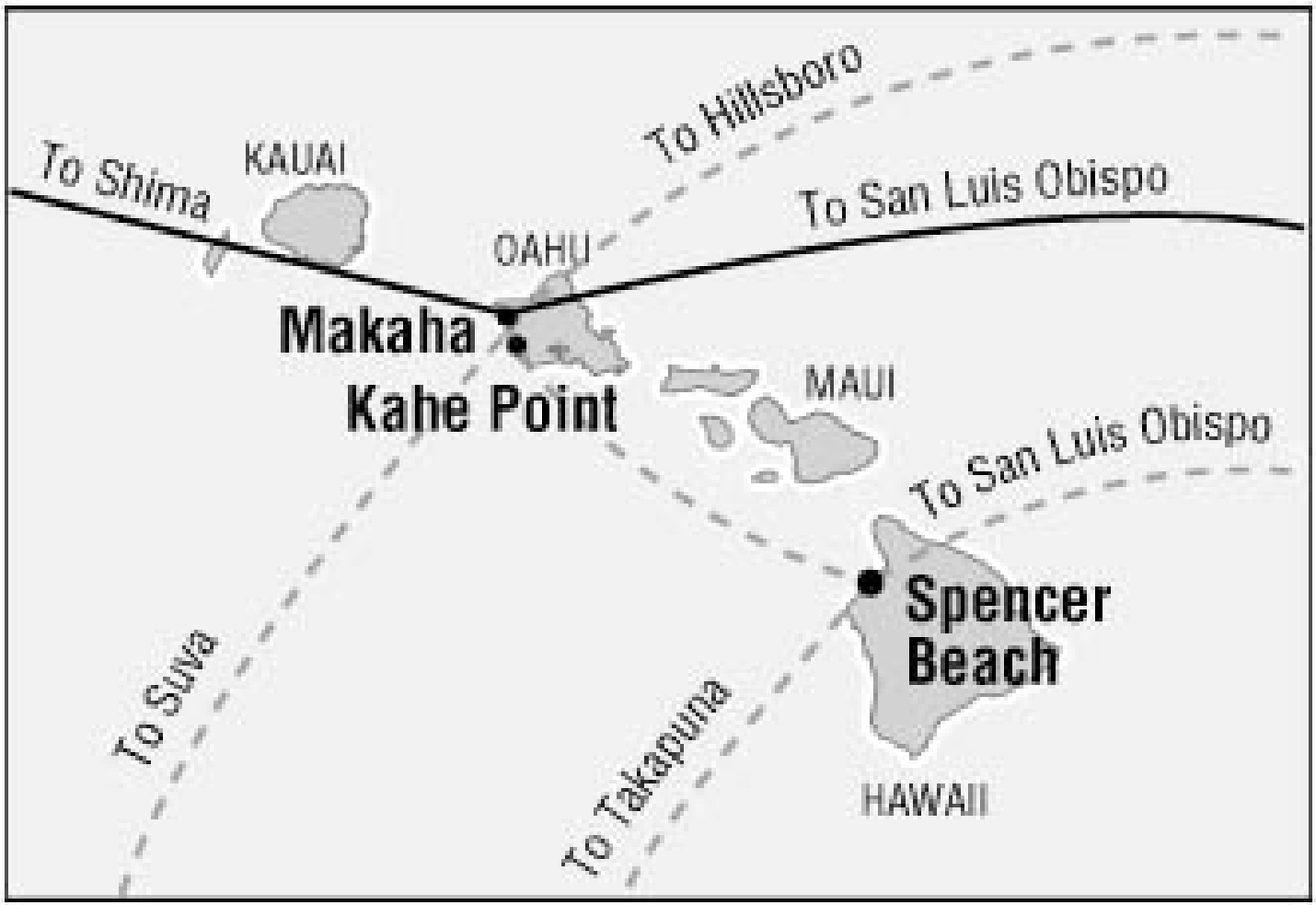
Today: Router Design

- **Router architecture:** What's inside the box?
- **Forwarding engine:** How a router knows the "next hop" for a packet.
- **Switch fabric:** When buses are too slow ... replace it with a switch!

Last time: Cables meet in Hawaii ...



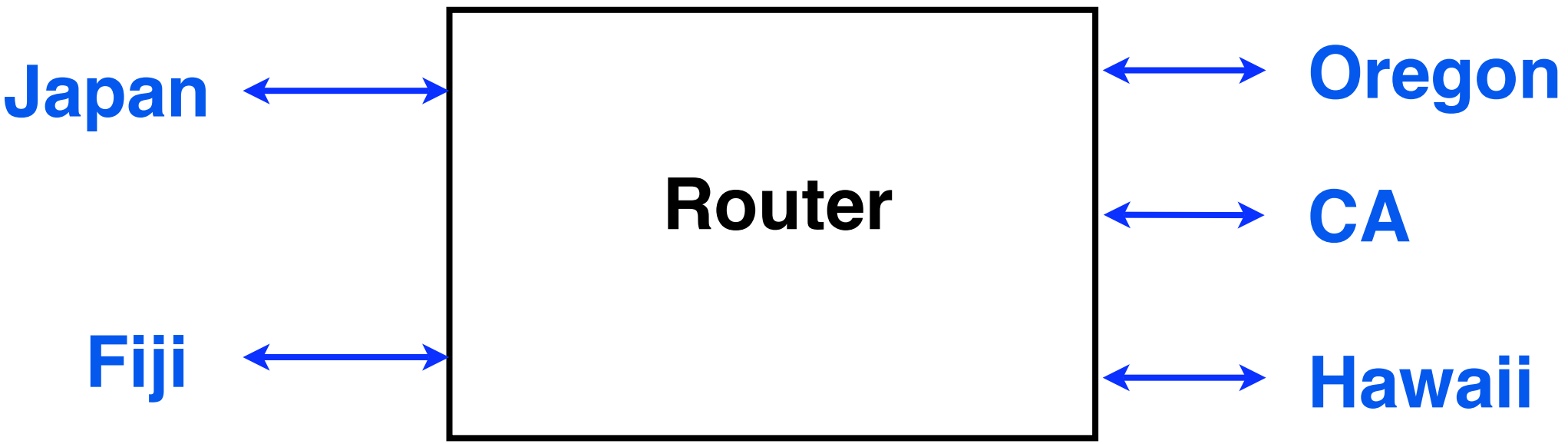
Last time: Routers are like hub airports



In Makaha, a router takes each Layer 2 packet off the San Luis Obispo (CA) cable, examines the IP packet destination field, and forwards to Japan cable, Fiji cable, or to Kahe Point (and onto big island cables).

The Oahu router ...

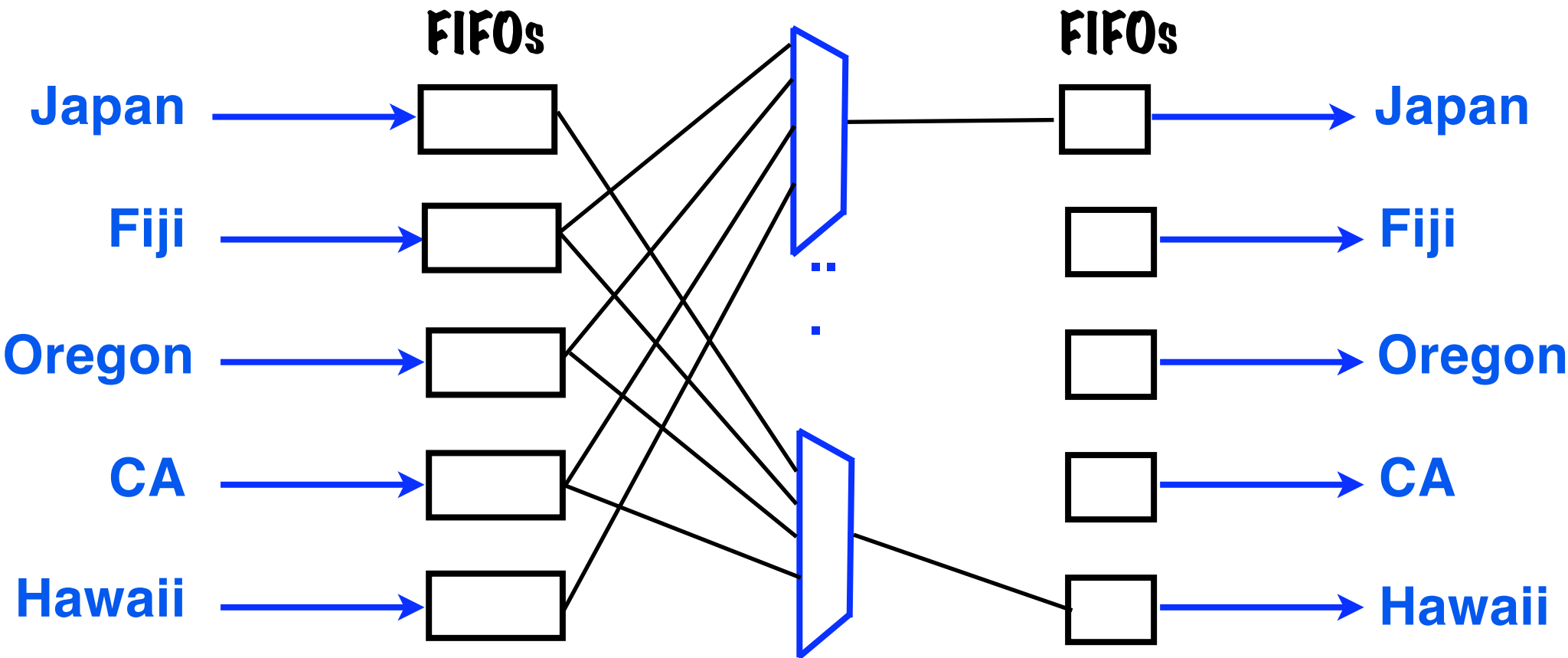
Assume each "line" is 160 Gbits/sec each way.



IP packets are forwarded from each inbound Layer 2 line to one of the four outbound Layer 2 lines, based on the destination IP number in the IP packet.

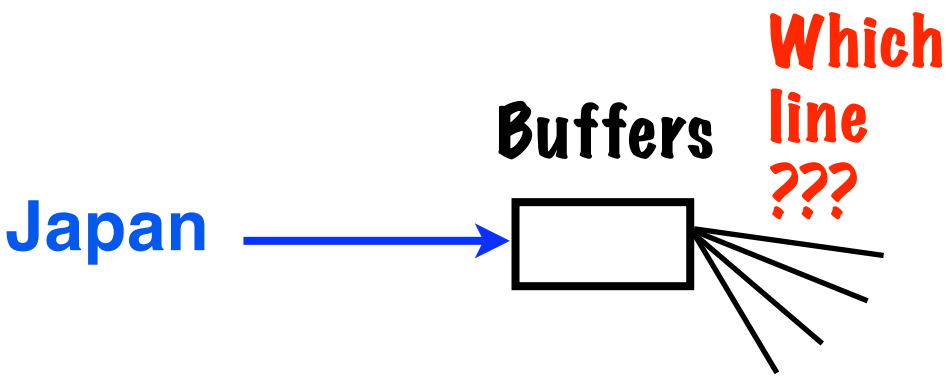
Challenge 1: Switching bandwidth

At line rate: 5*160 Gb/s = 100 GB/s switch!
Latency not an issue ... wide, slow bus OK.



FIFOs (first-in first-out packet buffers) help if an output is sent more bits than it can transmit. If buffers "overflow", packets are discarded.

Challenge 2: Packet forwarding speed



For each packet delivered by each inbound line, the router must decide which outbound line to forward it to. Also, update IP header.

Line rate: 160 Gb/s

Average packet size: 400 bits

Packets per second per line: 400 Million

Packets per second (5 lines): 2 Billion

Thankfully, this is trivial to parallelize ...

Challenge 3: Obeying the routing "ISA"

Network Working Group
Request for Comments: 1812
Obsoletes: 1716, 1009
Category: Standards Track

F. Baker, Editor
Cisco Systems
June 1995

Requirements for IP Version 4 Routers

Internet Engineering Task Force (IETF) "Request for Comments" (RFC) memos act as the "Instruction Set Architecture" for routers.

RFC 1812 (above) is 175 pages, and has 100 references which also define rules ...

The MGR Router: A case study ...

A 50-Gb/s IP Router

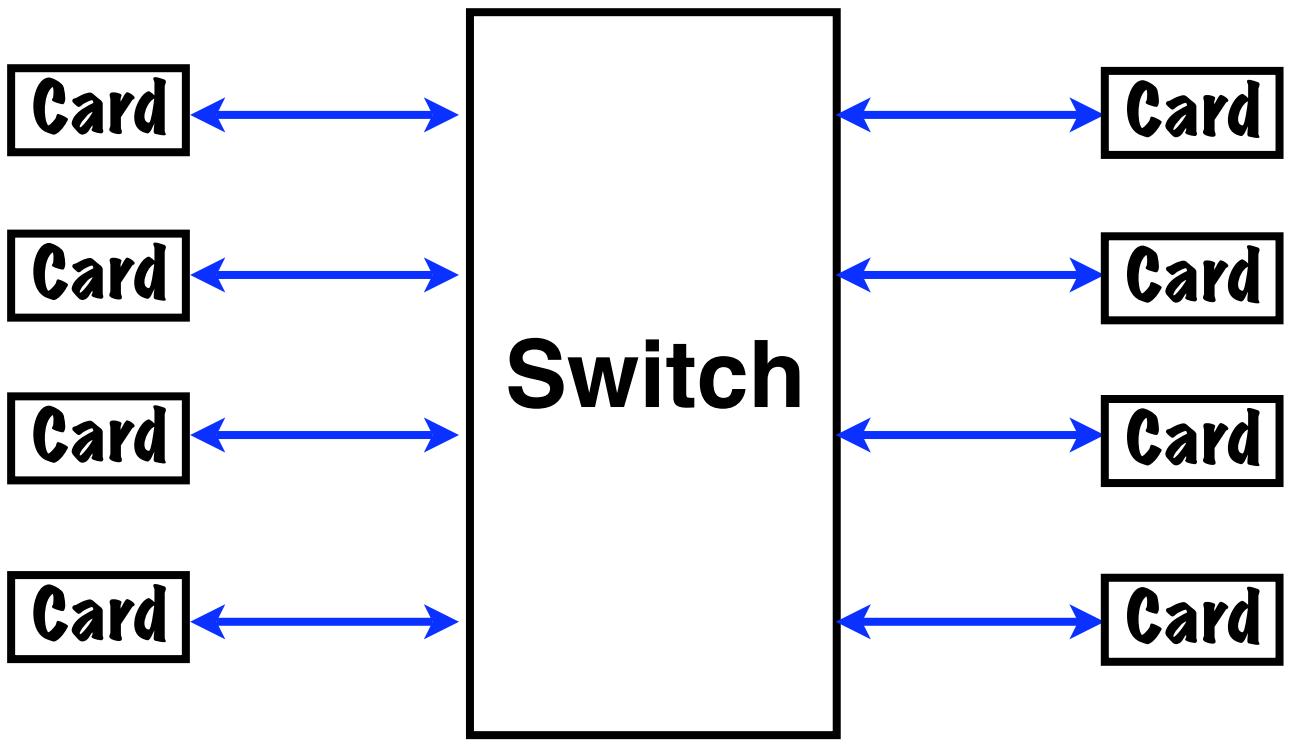
Craig Partridge, *Senior Member, IEEE*, Philip P. Carvey, *Member, IEEE*, Ed Burgess, Isidro Castineyra, Tom Clarke, Lise Graham, Michael Hathaway, Phil Herman, Allen King, Steve Kohalmi, Tracy Ma, John Mcallen, Trevor Mendez, Walter C. Milliken, *Member, IEEE*, Ronald Pettyjohn, *Member, IEEE*, John Rokosz, *Member, IEEE*, Joshua Seeger, Michael Sollins, Steve Storch, Benjamin Tober, Gregory D. Troxel, David Waitzman, and Scott Winterble

The "MGR" Router was a research project in late 1990's. Kept up with "line rate" of the fastest links of its day (OC-48c, 24 Gb/s optical).

Architectural approach is still valid today ...

MGR top-level architecture

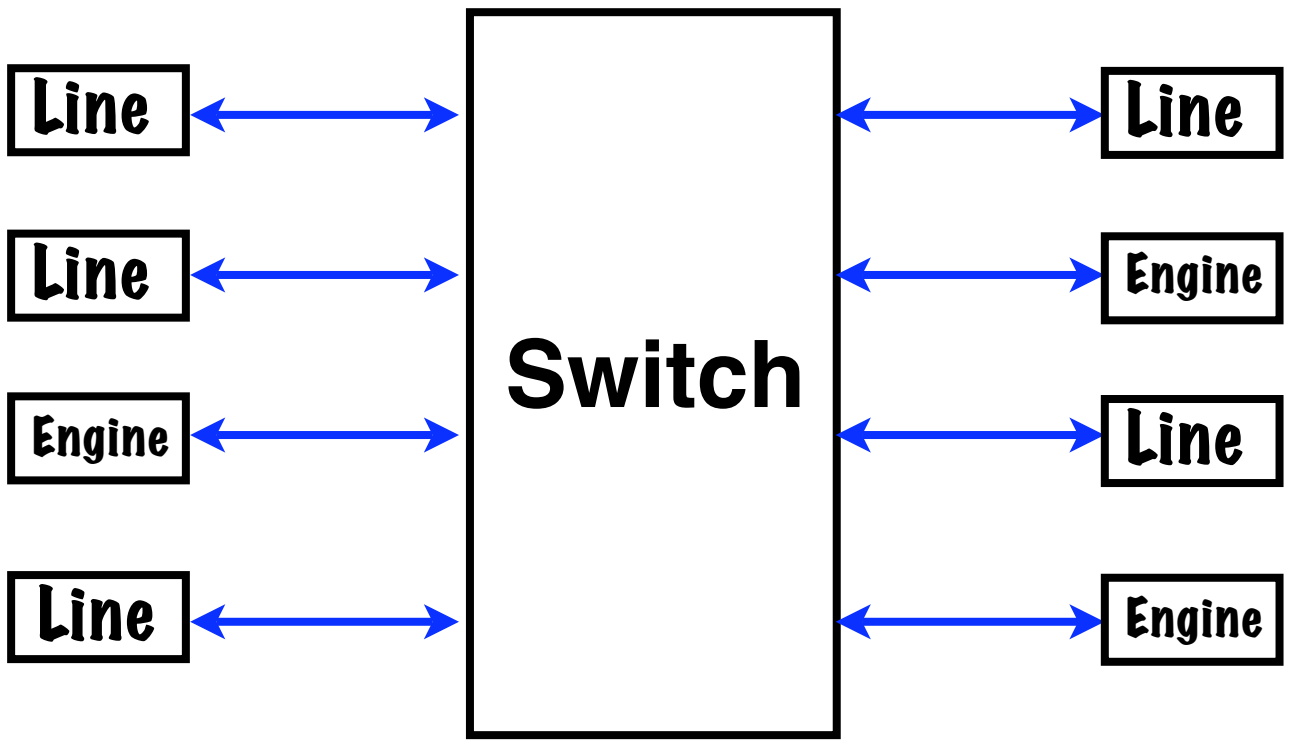
A 50 Gb/s switch is the centerpiece of the design. Cards plug into the switch.



In best case, on each switch "epoch" (transaction), each card can send and receive 1024 bits to/from one other card.

MGR cards come in two flavors ....

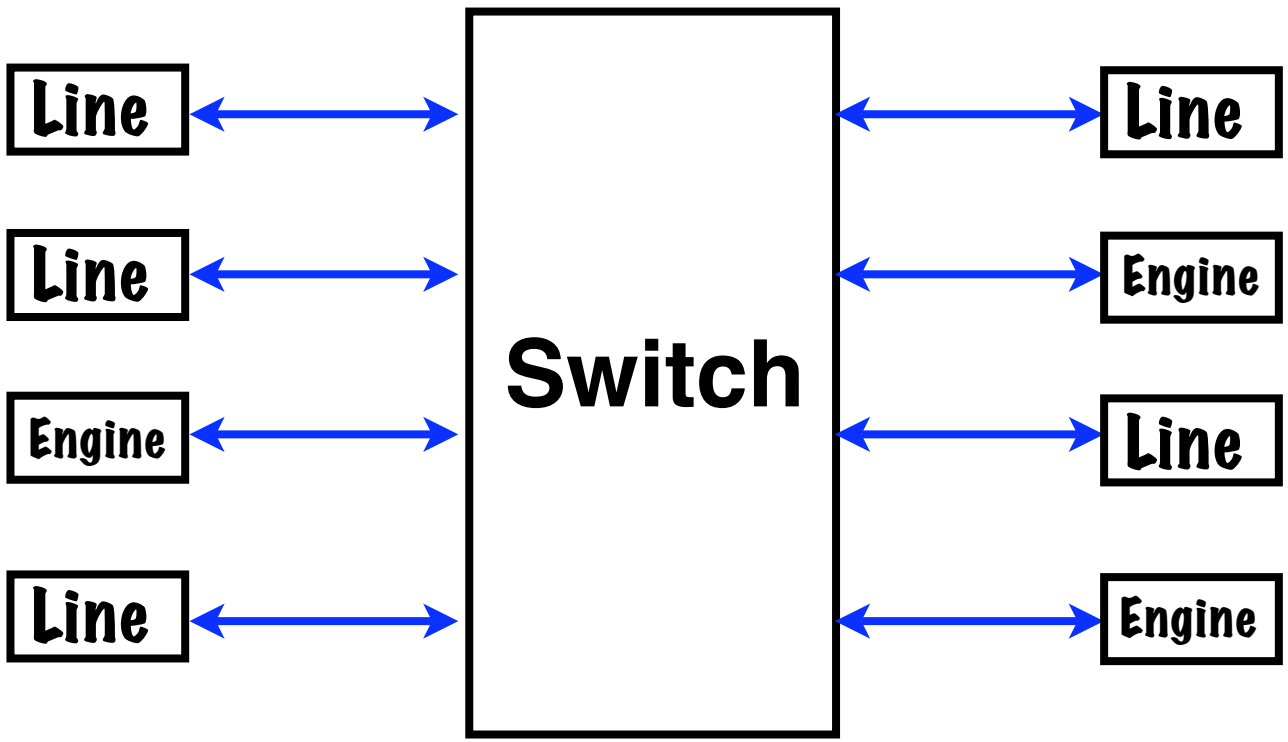
Line card: A card that connects to Layer 2 line. Different version of card for each Layer 2 type.



Forwarding engine: Receives IP headers over the switch from line cards, and returns forwarding directions and modified headers to line card.

A control processor for housekeeping

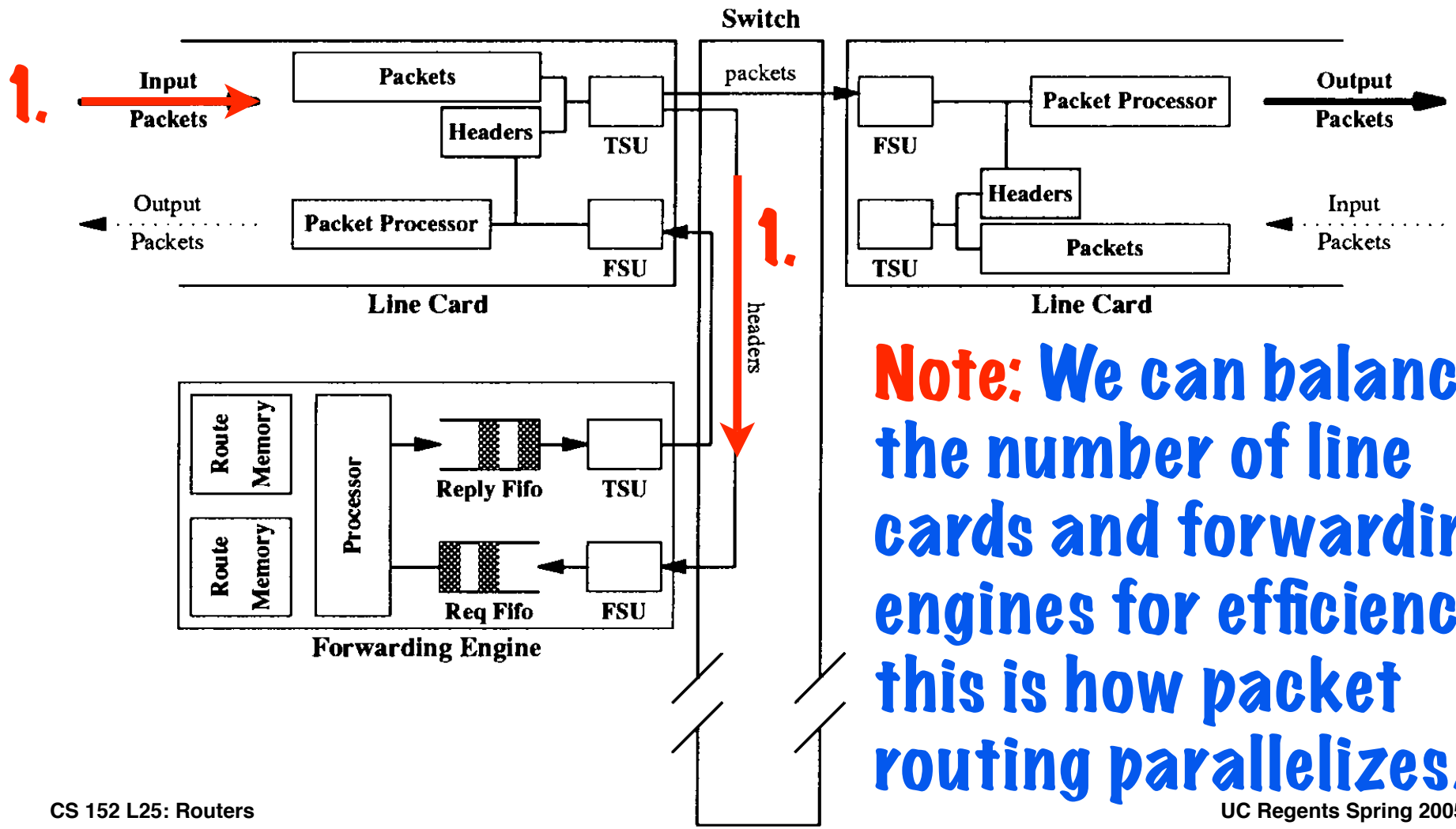
Forwarding engine handles fast path: the "common case" of unicast packets w/o options. Unusual packets are sent to the control processor.



Control processor

The life of a packet in a router ...

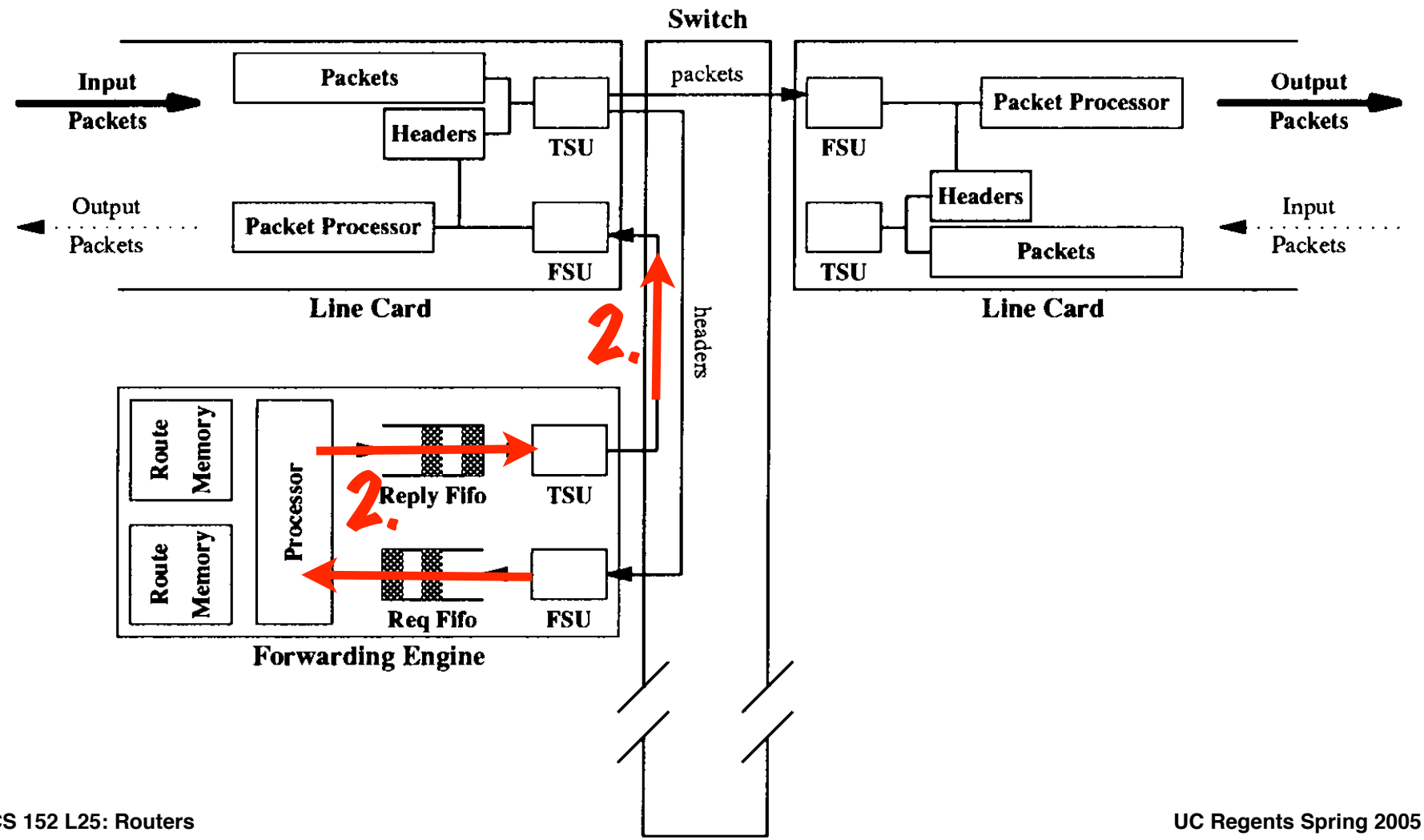
1. Packet arrives in line card. Line card sends the packet header to a forward engine for processing.



Note: We can balance the number of line cards and forwarding engines for efficiency: this is how packet routing parallelizes.

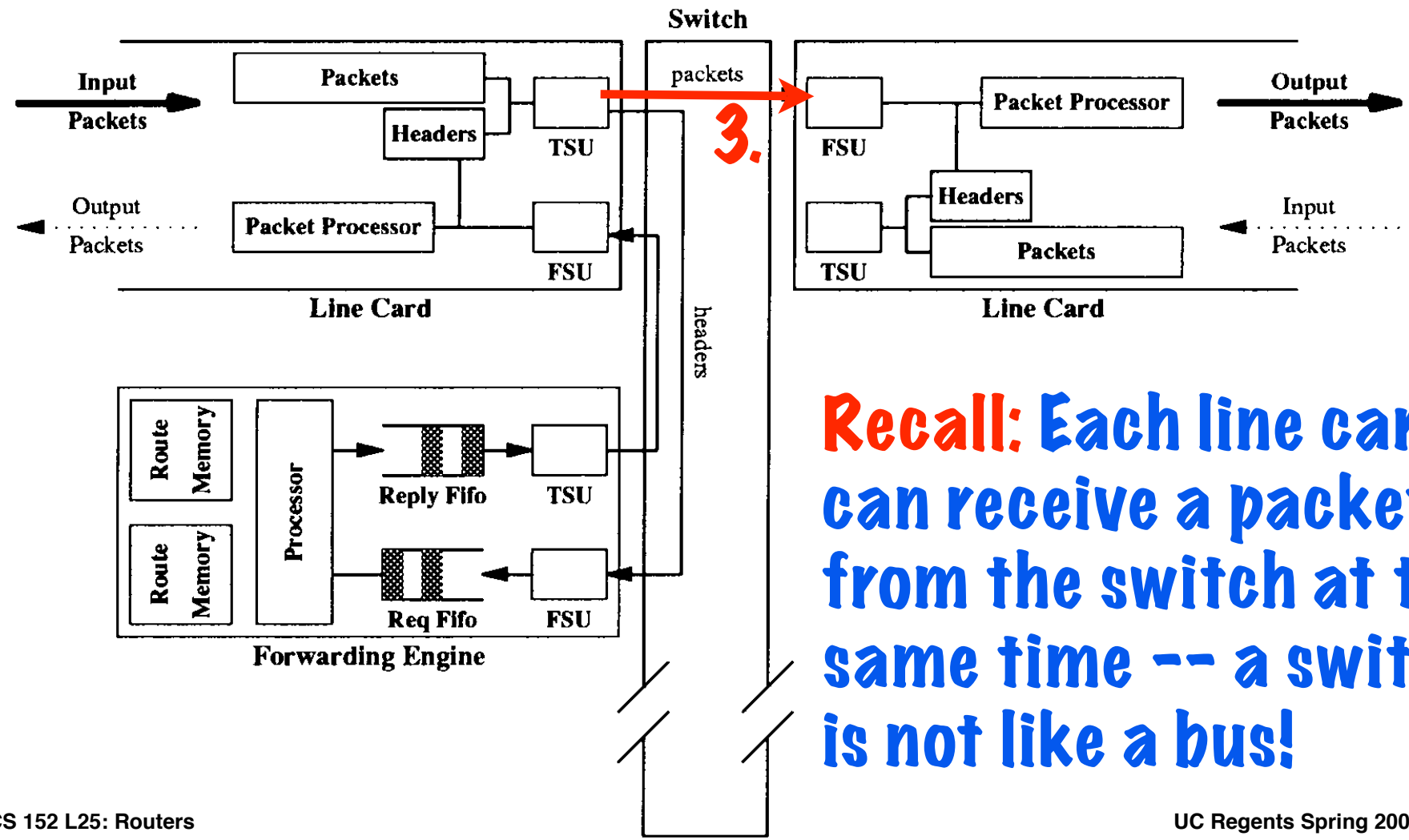
The life of a packet in a router ...

2. Forwarding engine determines the next hop for the packet, and returns next-hop data to the line card, together with an updated header.



The life of a packet in a router ...

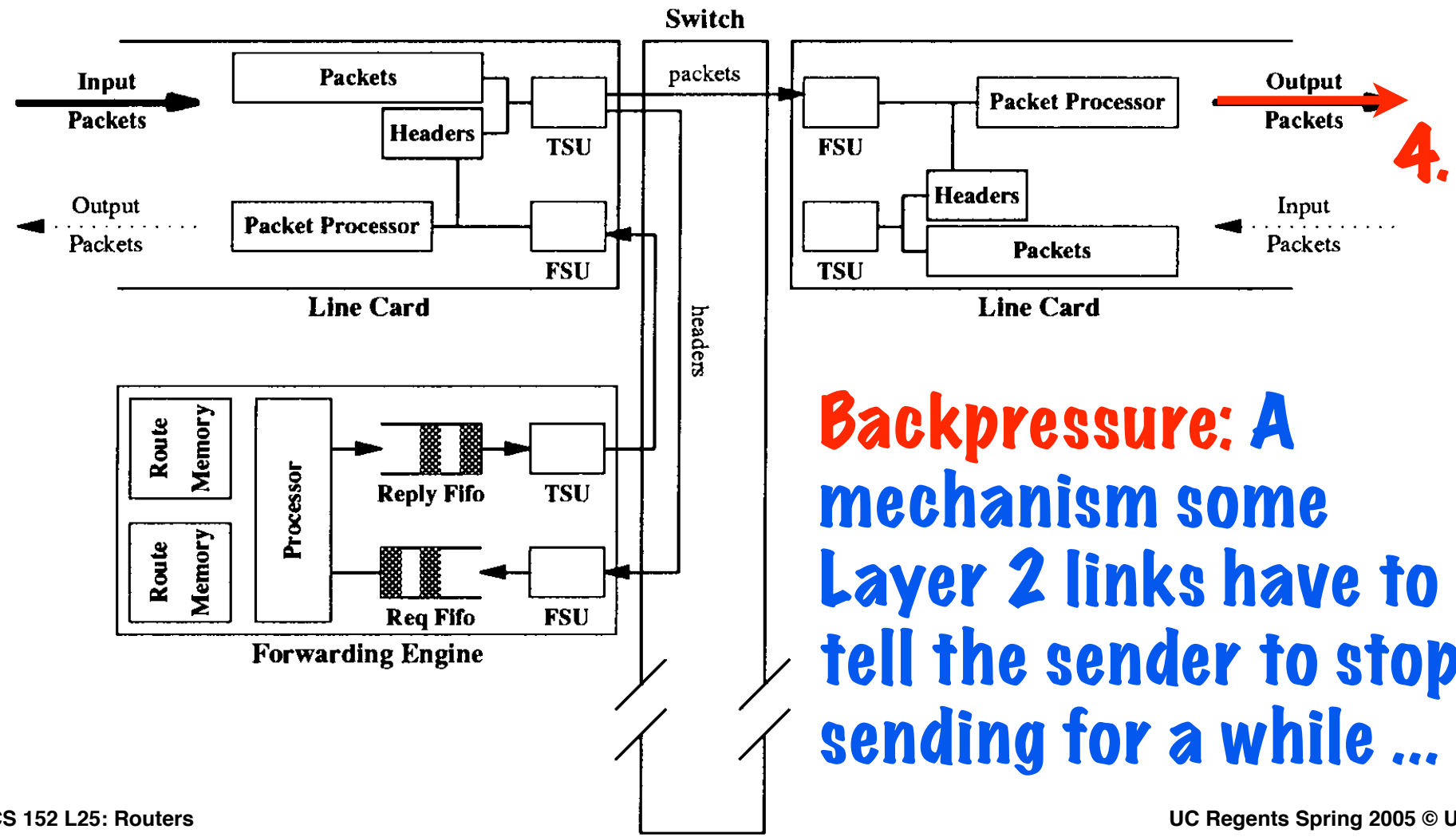
3. Line card uses forwarding information, and sends the packet to another line card via the switch.



Recall: Each line card can receive a packet from the switch at the same time -- a switch is not like a bus!

The life of a packet in a router ...

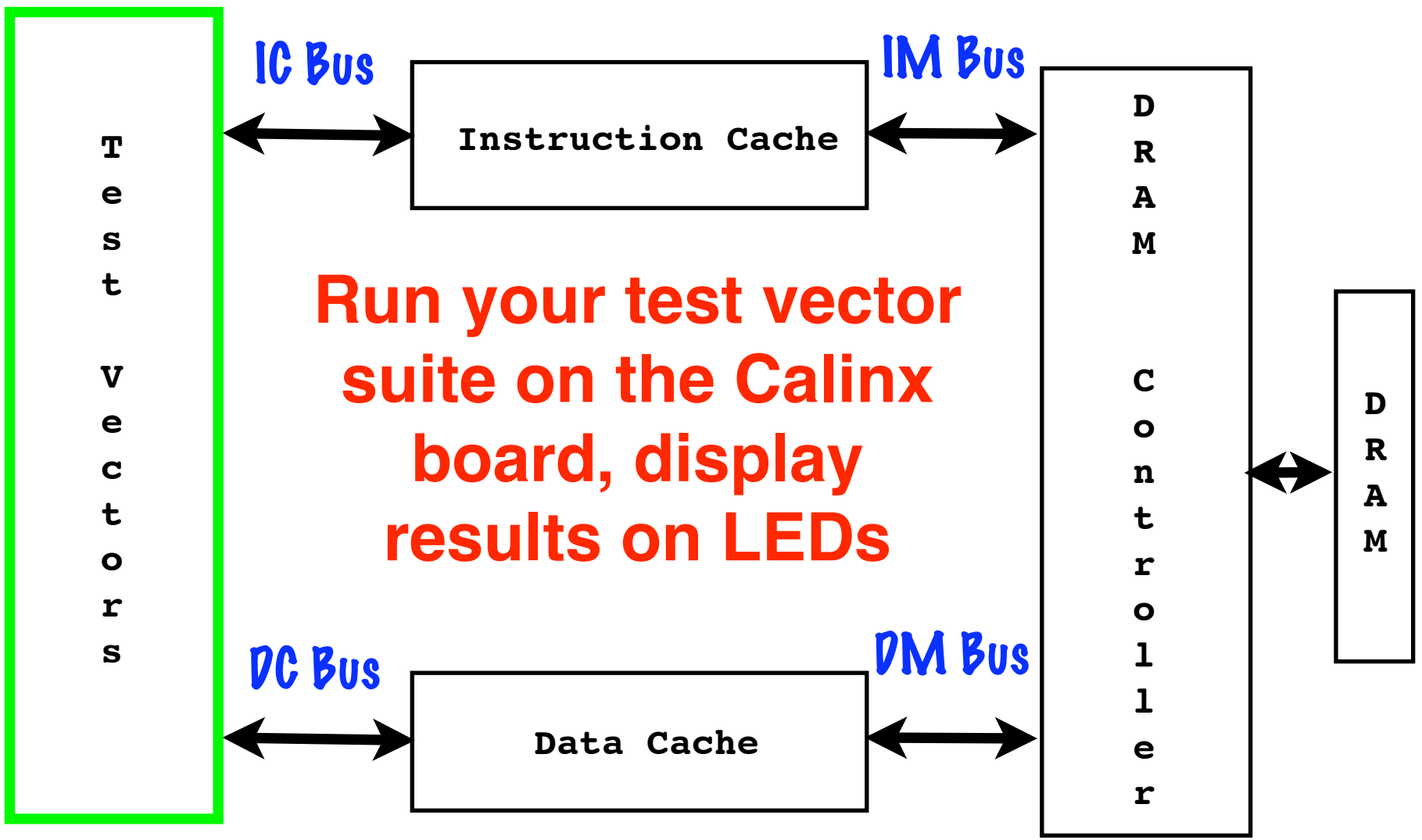
4. Outbound line card sends packet on its way ...



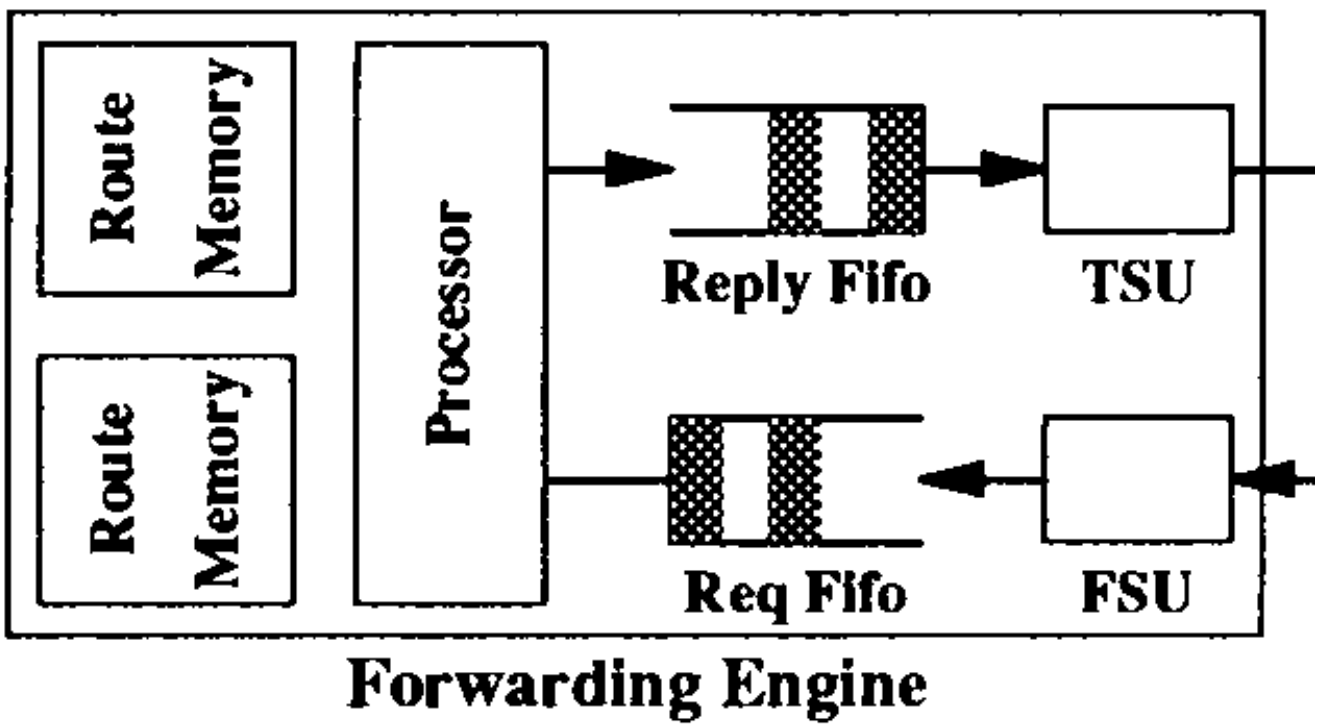
Backpressure: A mechanism some Layer 2 links have to tell the sender to stop sending for a while ...

This Friday: Memory System Checkoff

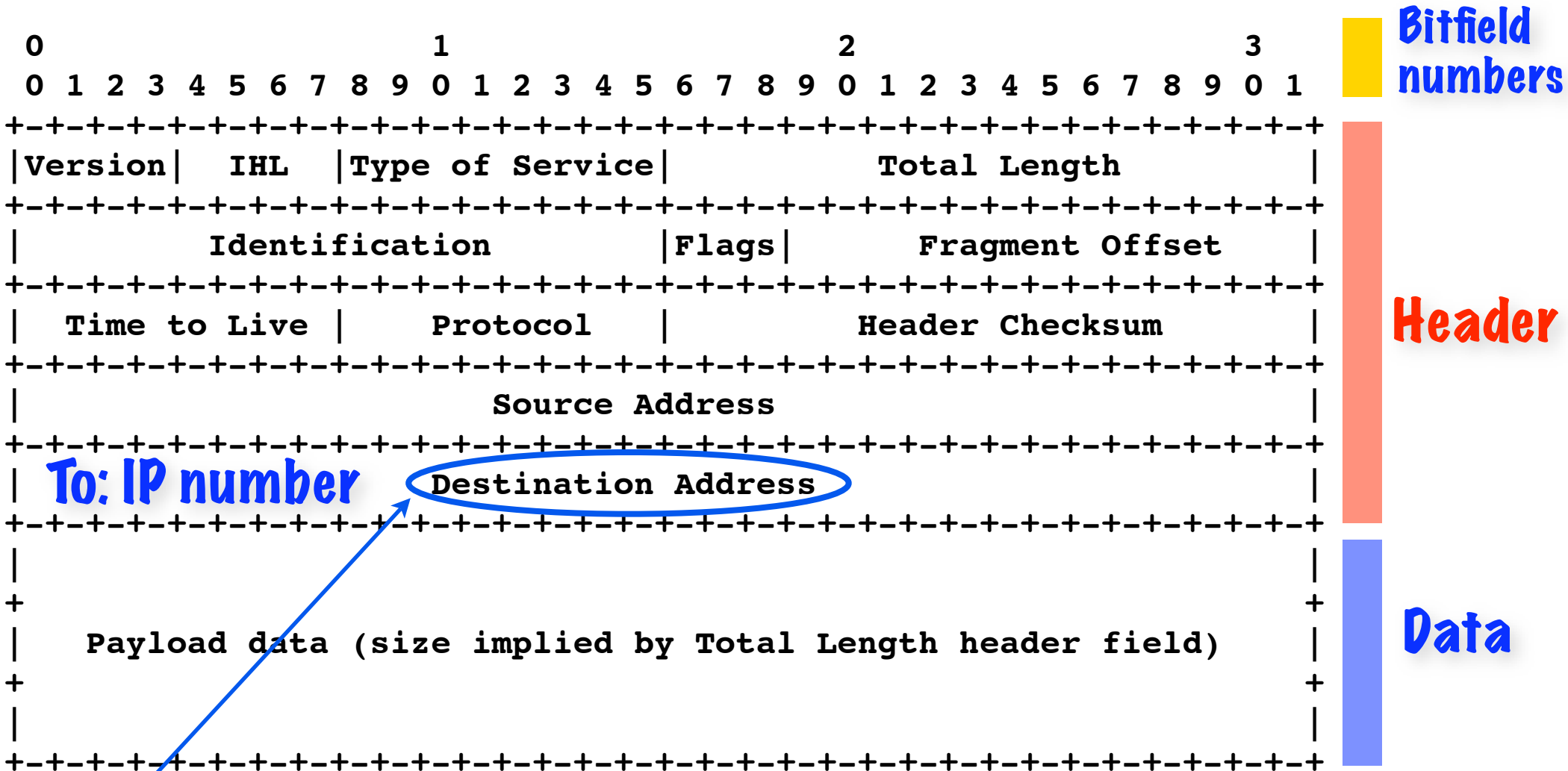
F 4/22				Final Project: Memory System Xilinx Checkoff Midterm Review Homework Due in Section



Packet Forwarding



Forwarding engine computes "next-hop"

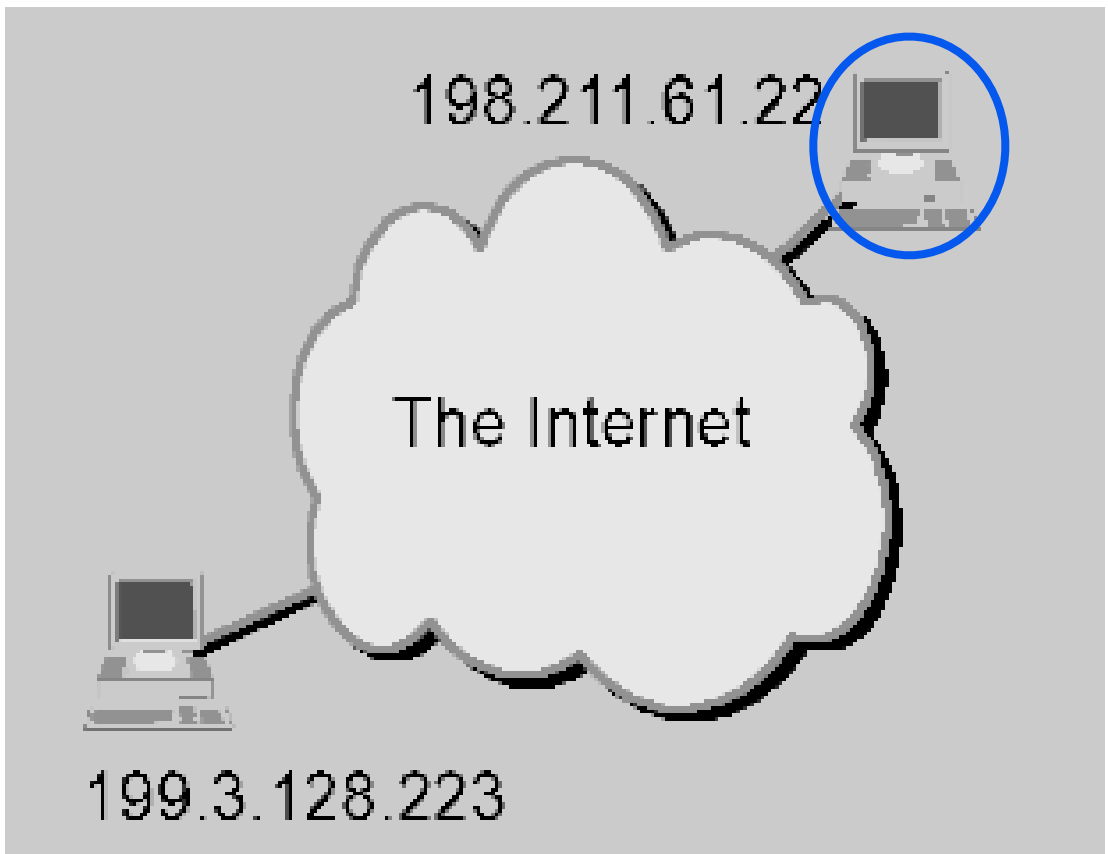


Forwarding engine looks at the destination address, and decides which outbound line card will get the packet closest to its destination. How?

Recall: Internet IP numbers ...

IP4 number for this computer: 198.211.61.22

198.211.61.22 == 3335732502 (32-bit unsigned)



Every directly connected host has a unique IP number.

Upper limit of 2^32 IP4 numbers (some are reserved for other purposes).

BGP: The Border Gateway Protocol

Routers use BGP to exchange routing tables. Tables code if it is possible to reach an IP number from the router, and if so, how "desirable" it is to take that route.

Network Working Group — Y. Rekhter
Request for Comments: 1771 — T.J. Watson Research Center, IBM Corp.
Obsoletes: 1654 — T. Li
Category: Standards Track — cisco Systems
Editors
March 1995

A Border Gateway Protocol 4 (BGP-4)

Routers use BGP tables to construct a "next-hop" table. Conceptually, forwarding is a table lookup: IP number as index, table holds outbound line card.

A table with 4 billion entries ???

Tables do not code every host …

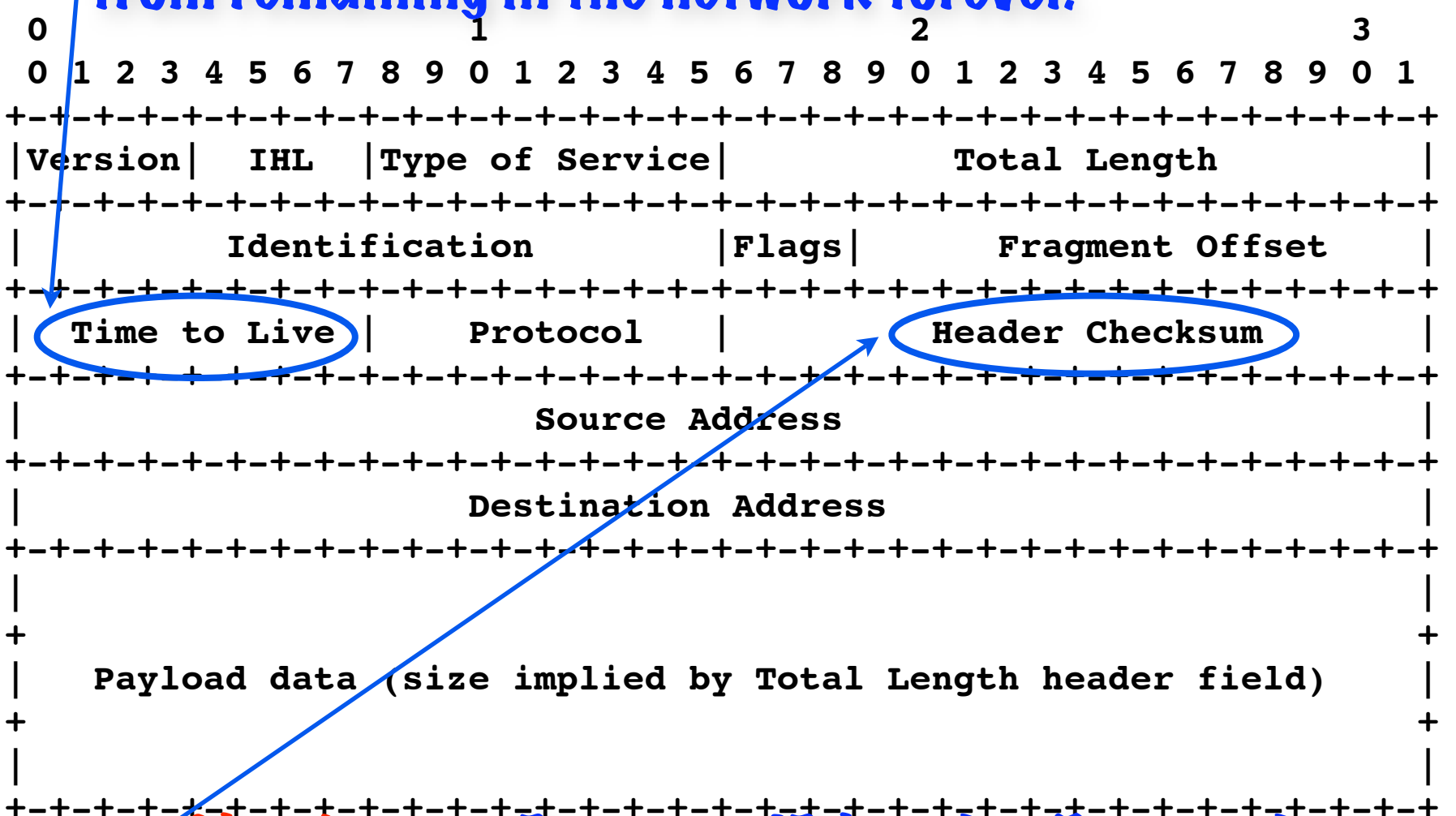
Routers route to a "network", not a "host". /xx means the top xx bits of the 32-bit address identify a single network.

Network	IP Address Range		Comment
	From:	To:	
128.32.0.0/16	128.32.0.0	128.32.255.255	UCB Local Area Networks *
136.152.0.0/16	136.152.0.0	136.152.255.255	UCB Local Area Networks and Home IP Service #
169.229.0.0/16	169.229.0.0	169.229.255.255	UCB Local Area Networks
131.243.52.0/24	131.243.52.0	131.243.52.255	UCB Melvin Calvin Lab. building
192.101.42.0/24	192.101.42.0	192.101.42.255	UCB Local Area Networks
199.133.139.0/24	199.133.139.0	199.133.139.255	USDA/UCB Joint Local Area Network

Thus, all of UCB only needs 6 routing table entries.
Today, Internet routing table has about 100,000 entries.

Forwarding engine: Also updates header

Time to live. Sender sets to a high value. Each router decrements it by one, discards if 0. Prevents a packet from remaining in the network forever.



Bitfield numbers

Header

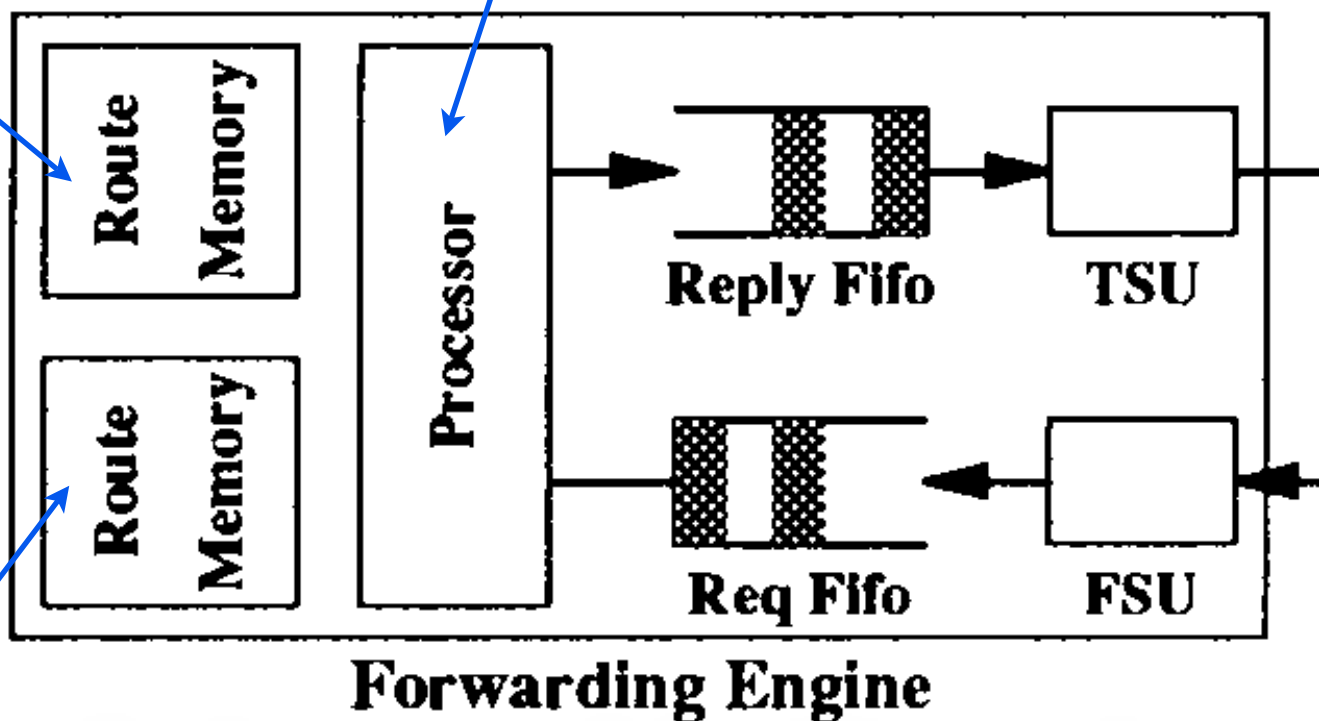
Data

Checksum. Protects IP header. Forwarding engine updates it to reflect the new Time to Live value.

MGR forwarding engine: a RISC CPU

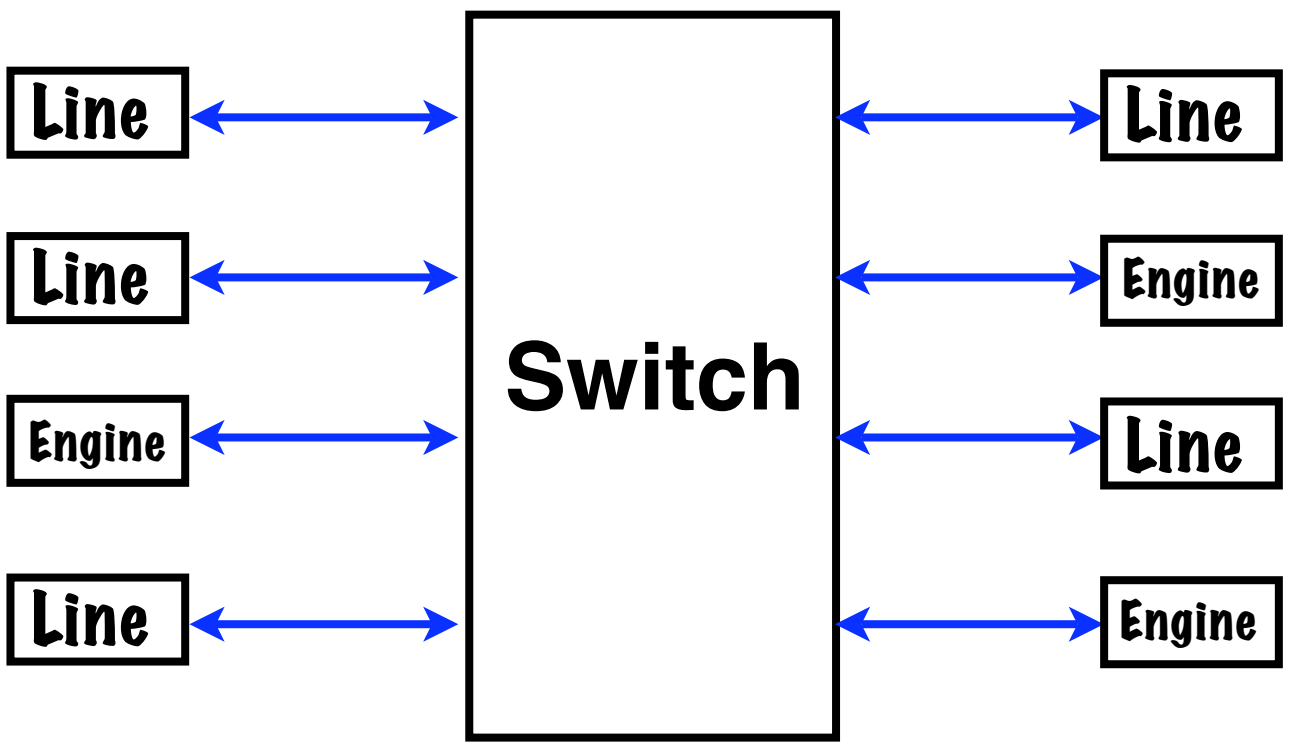
Off-chip memory in two 8MB banks: one holds the current routing table, the other is being written by the router's control processor with an updated routing table. Why??? So that the router can switch to a new table without packet loss.

85 instructions in "fast path", executes in about 42 cycles. Fits in 8KB I-cache

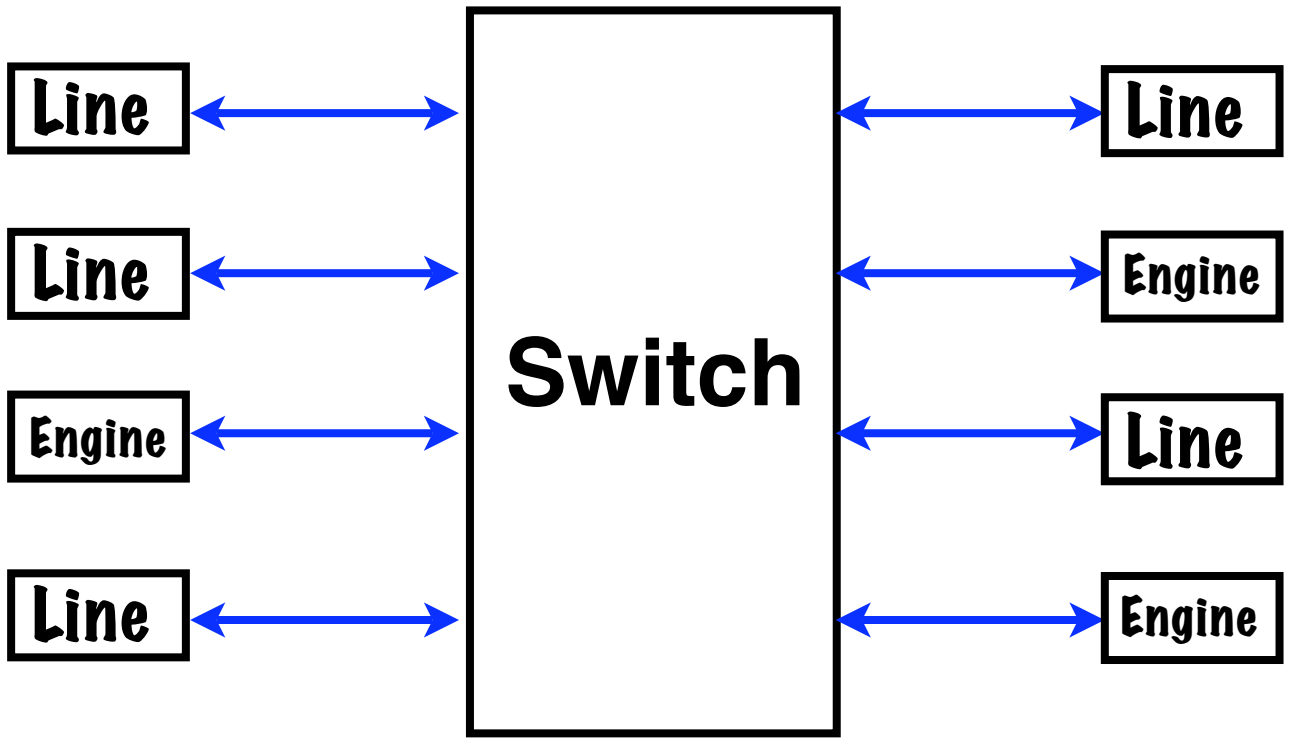


Performance: 9.8 million packet forwards per second. To handle more packets, add forwarding engines. Or use a special-purpose CPU.

Switch Architecture



What if two inputs want the same output?



A pipelined arbitration system decides how to connect up the switch. The connections for the transfer at epoch N are computer in epochs N-3, N-2 and N-1, using dedicated switch allocation wires.

A complete switch transfer (4 epochs)

- **Epoch 1:** All input ports ready to send data request an output port.
- **Epoch 2:** Allocation algorithm decides which inputs get to write.
- **Epoch 3:** Allocation system informs the winning inputs and outputs.
- **Epoch 4:** Actual data transfer takes place.

Allocation is pipelined: a data transfer happens on every cycle, as does the three allocation stages, for different sets of requests.

Epoch 3: The Allocation Problem

Output Ports (A, B, C, D)

Input Ports (A, B, C, D)

	A	B	C	D
A	0	0	1	0
B	1	0	0	1
C	0	1	0	0
D	1	0	1	0

A 1 codes that an input has a packet ready to send to an output. Note an input may have several packets ready.

Allocator returns a matrix with one 1 in each row and column to set switches. Algorithm should be "fair", so no port always loses ... should also "scale" to run large matrices fast.

	A	B	C	D
A	0	0	1	0
B	0	0	0	1
C	0	1	0	0
D	1	0	0	0

Recall: The IP "non-ideal" abstraction

- A sent packet may never arrive ("lost")

Router drops packets if too much traffic destined for one port, or if Time to Live hits 0, or checksum failure.

- If packets sent P1/P2/P3, they may arrive P2/P1/P3 ("out of order").
- Relative timing of packet stream not necessarily preserved ("late" packets).

This happens when the packet's header forces the forwarding processor out of the "fast path", etc.

- IP payload bits received may not match payload bits sent.

Usually happens "on the wire", not in router.

Conclusions: Router Design

- **Router architecture:** The "ISA" for routing was written with failure in mind -- unlike CPUs.
- **Forwarding engine:** The computational bottleneck, many startups target silicon to improve it.
- **Switch fabric:** Switch fabrics have high latency, but that's OK: routing is more about bandwidth than latency.