

Lecture 12 (Transport 3)

Congestion Control, Part 1

CS 168, Fall 2026 @ UC Berkeley

Slides credit: Sylvia Ratnasamy, Rob Shakir, Peyrin Kao, Ion Stoica

How big should the window be?

Pick window size W to balance three goals:

1. Take advantage of network capacity ("fill the pipe")... (set $W = \text{RTT} \times \text{bandwidth}$)
2. ...but don't overload the recipient (flow control)... (set $W = \text{recipient's advertised window}$)
3. ...and don't overload links (congestion control). (set $W = \text{sender's congestion window}$)

So far, we've assumed that congestion control magically tells us the rate to send at

- Today, we'll understand how the transport layer (and TCP) determines this rate.

Congestion Control: Why do we need it?

Lecture 12, CS 168, Spring 2026

Congestion Control Principles

- **Why do we need it?**
- Why is it hard?
- Design Goals

Dynamic Adjustment

- Algorithm Sketch
- Reacting to Congestion (AIMD)
- Implementing Congestion Control with Event-Driven Updates

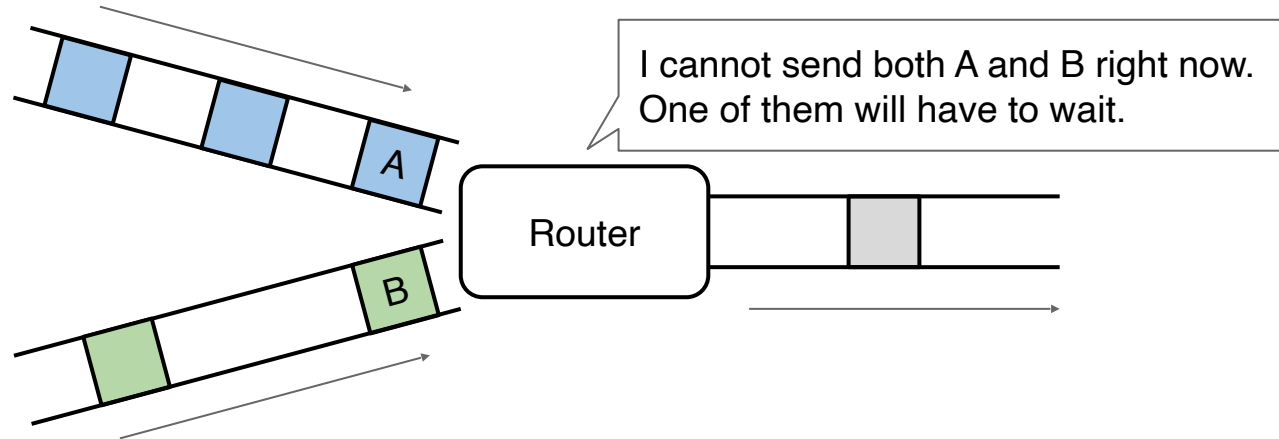
If two packets arrive at the same time, the router will send one, and buffer the other.

The router maintains a *queue* of packets that still need to be sent.

- If the queue is full, and a packet arrives, the packet gets dropped.

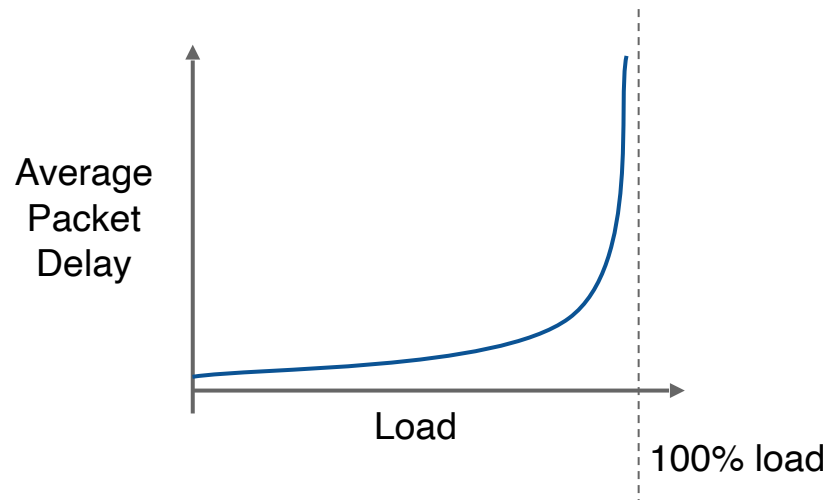
If too many packets arrive close in time:

- The router cannot keep up, and gets congested.
- Packets can be delayed (stuck waiting in queue) or dropped (queue is full).



For a typical queuing system with bursty arrivals:

- Higher load = more packets sent = higher packet delay.
- Delay gets really bad, even before we reach 100% load.
- Trade-offs: Need to balance high link utilization and low delay.



In the 1980s, TCP did not implement congestion control.

- Sending rate was only limited by flow control (recipient buffer capacity).
- If packets are dropped, resend them over and over, at the same fast rate.

This led to a *congestion collapse*.

- 1986: Capacity of a major link dropped to 0.1% of its original capacity.
- Why? Network was flooded with thousands of copies of every packet.

In October of '86, the Internet had the first of what became a series of 'congestion collapses.' During this period, the data throughput from LBL to UC Berkeley (sites separated by 400 yards and two IMP hops) dropped from 32 Kbps to 40 bps. We were fascinated by this sudden factor-of-thousand drop in bandwidth and embarked on an investigation of why things had gotten so bad. In particular, we wondered if the 4.3BSD (Berkeley Unix) TCP was mis-behaving or if it could be tuned to work better under abysmal network conditions. The answer to both of these questions was "yes".

– Karels (UCB) and Jacobson (LBL)

How was congestion collapse solved in the 1980s?

- Michael Karels and Van Jacobson were working on BSD (an early OS).
- They changed a few lines of TCP code to fix the problem.
 - Recall: TCP is implemented in everybody's OS.
 - The fix worked, and everybody quickly adopted it.

The history of congestion control had a big impact on its design.

- Invented as an ad-hoc patch to an existing system. (*Not in original TCP design.*)
- Implemented in the OS. (*Not in applications, routers, etc.*)

Fast-forward to today:

- Extensive research and improvements since the original patch.
- The core ideas in the original patch still persist.
- No major congestion collapses ever since.

Congestion Control: Why is it hard?

Lecture 12, CS 168, Spring 2026

Congestion Control Principles

- Why do we need it?
- **Why is it hard?**
- Design Goals

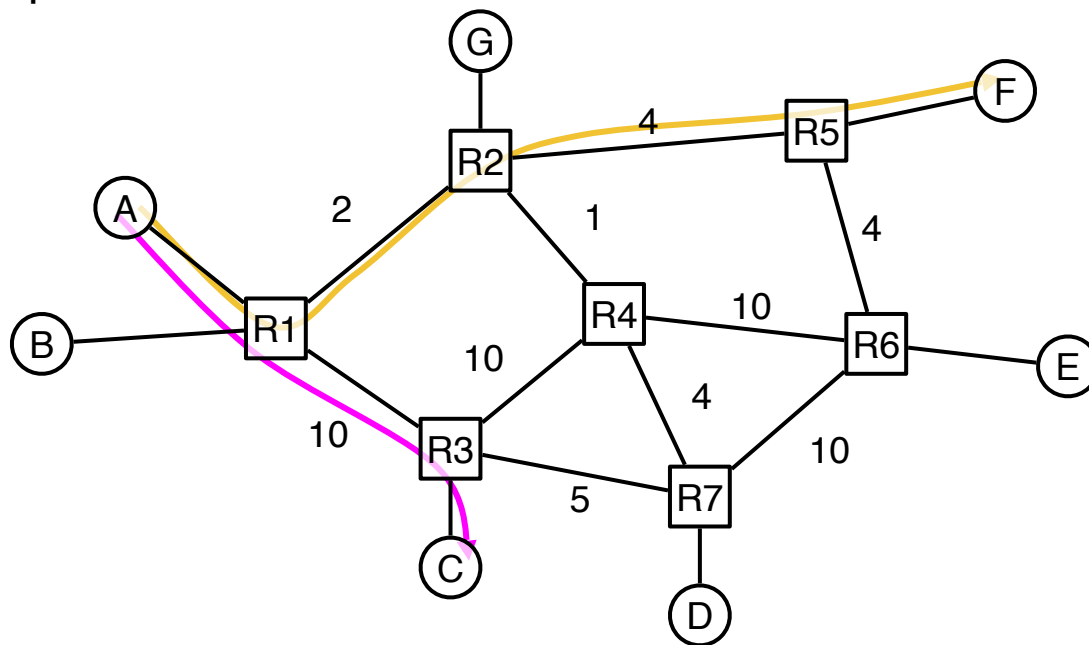
Dynamic Adjustment

- Algorithm Sketch
- Reacting to Congestion (AIMD)
- Implementing Congestion Control with Event-Driven Updates

At what rate should A send traffic?

It depends on the destination.

- $A \rightarrow C$ can send at 10 Gbps.
- $A \rightarrow F$ can only send at 2 Gbps.

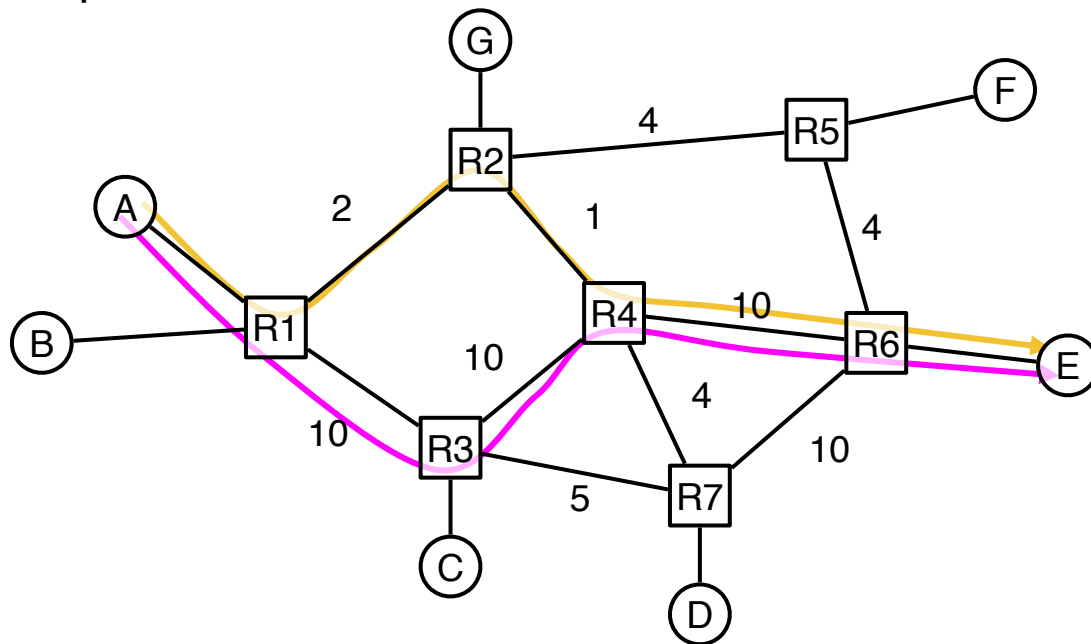


Numbers are bandwidths, in Gbps.
Ignore bandwidth of host-to-router links.

At what rate should $A \rightarrow E$ send traffic?

It depends on the path through the network.

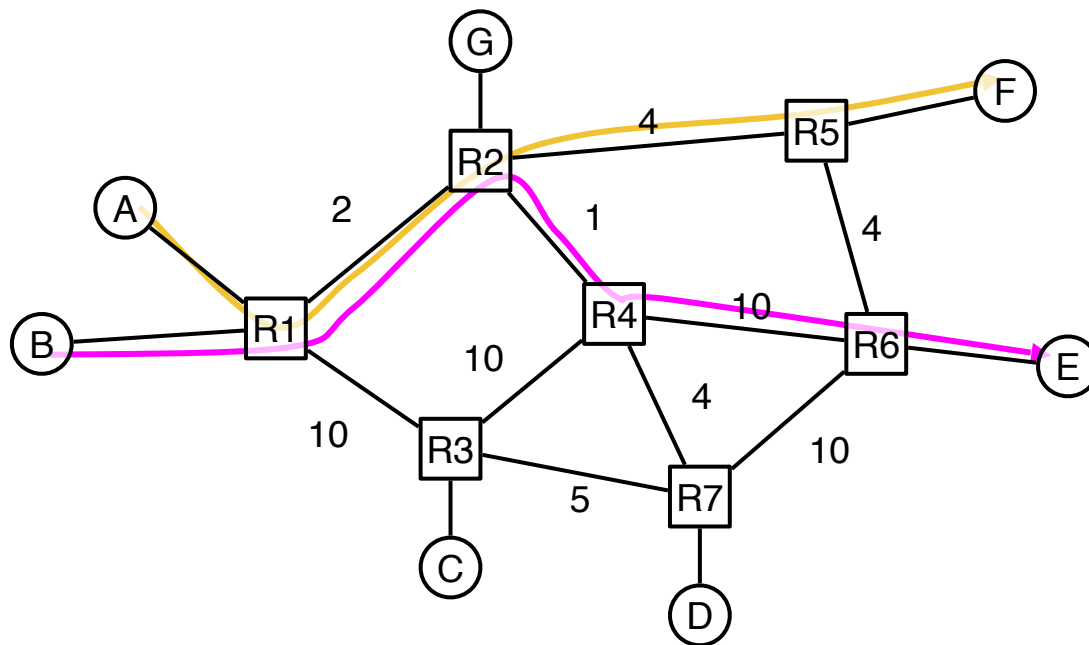
- $A \rightarrow E$ via R3 can send at 10 Gbps.
- $A \rightarrow E$ via R2 can send at 1 Gbps.



At what rate should $A \rightarrow F$ send traffic?

It depends on others using the network.

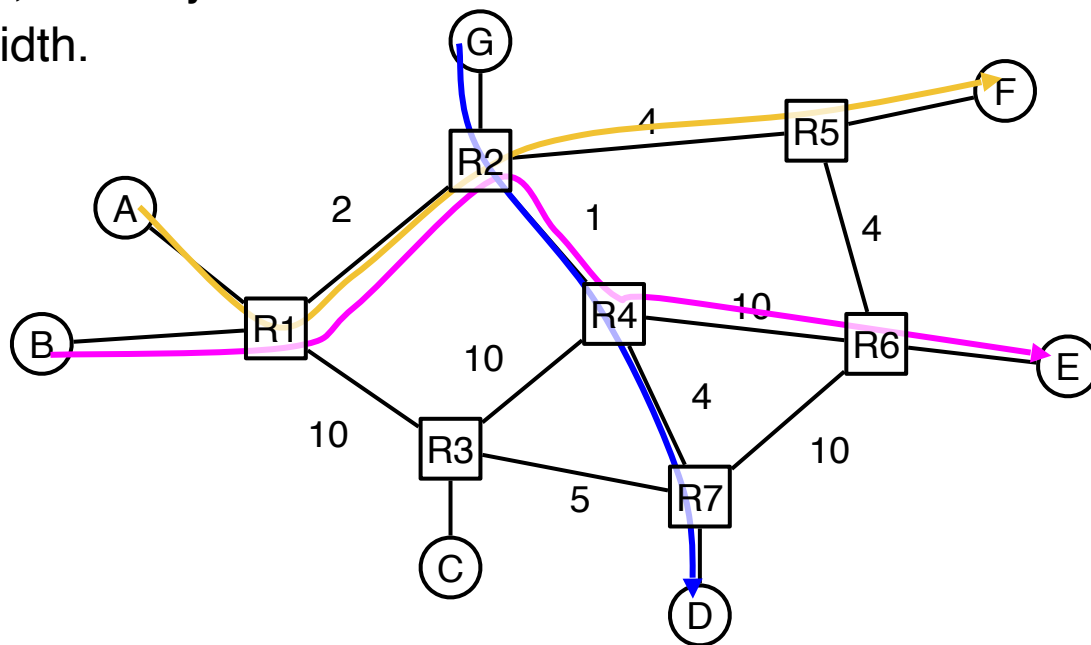
- $A \rightarrow F$ and $B \rightarrow E$ share a link.
- $B \rightarrow E$ can send at 1 Gbps.
- $A \rightarrow F$ can send at 1 Gbps.



What if $G \rightarrow D$ starts sending traffic?

- Maybe $B \rightarrow E$ should slow to 0.5 Gbps.
- Then $A \rightarrow F$ could increase to 1.5 Gbps.

$G \rightarrow D$ and $A \rightarrow F$ share no links, but they still affected each other's bandwidth.



The sender's optimal rate depends on:

- The destination.
- The path to the destination.
- Other connections sharing links along that path.
- Connections sharing links with those other connections.
- And so on...

Congestion control is a global problem.

- Connections in the network are highly interdependent.
- Solving it optimally requires a global view of the network.

Fundamentally, congestion control is a *resource allocation problem*.

- Bandwidth is a limited resource.
- Each connection demands a certain amount of bandwidth.
- We decide how much to allocate to each connection.

Resource allocation is a classic problem (e.g. CPU scheduling, memory allocation).

But, congestion control has additional challenges.

- Everybody must solve it locally. No global view of network.
- Network topology is constantly changing.

Design Goals

Lecture 12, CS 168, Spring 2026

Congestion Control Principles

- Why do we need it?
- Why is it hard?
- **Design Goals**

Dynamic Adjustment

- Algorithm Sketch
- Reacting to Congestion (AIMD)
- Implementing Congestion Control with Event-Driven Updates

Goals for a good congestion control algorithm:

1. Avoid congestion: Minimize packet delay and loss.
2. Efficiency: Maximize link utilization.
3. Fairness: Connections should share the available bandwidth.
4. Practical to implement: Scalable, decentralized, adaptive, etc.

Our ultimate goal is a good tradeoff between these metrics.

- Could avoid congestion (1) by sending really slowly.
- But then we aren't using all the bandwidth on the link (2).

Reservations:

- Connections reserve bandwidth at the start, and release bandwidth at the end.
- Requires connections to know bandwidth ahead of time.
- Other problems: Hard to implement, inefficient bandwidth usage, etc.

Pricing/priorities:

- Analogy: Express toll lanes on the highway. Price changes depending on traffic.
- Prioritize traffic from users who pay more.
- Requires a payment model.

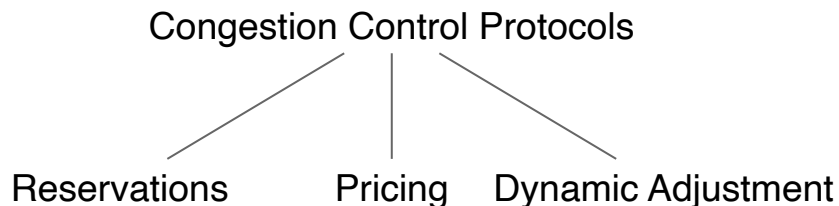
Dynamic adjustment:

- Learn the current congestion level, and adjust your rate accordingly.
- Requires good citizenship (no cheating).

3 possible approaches: reservations, pricing, and dynamic adjustment.

All the algorithms we'll see are based on dynamic adjustment.

- If we rebuilt the Internet from scratch, another approach might be feasible.
- But congestion control was invented as a patch on the existing Internet.
- Dynamic adjustment is the best fit for the existing Internet:
 - Doesn't require knowing bandwidth ahead of time.
 - Doesn't require a payment model.



Dynamic Adjustment: Algorithm Sketch

Lecture 12, CS 168, Spring 2026

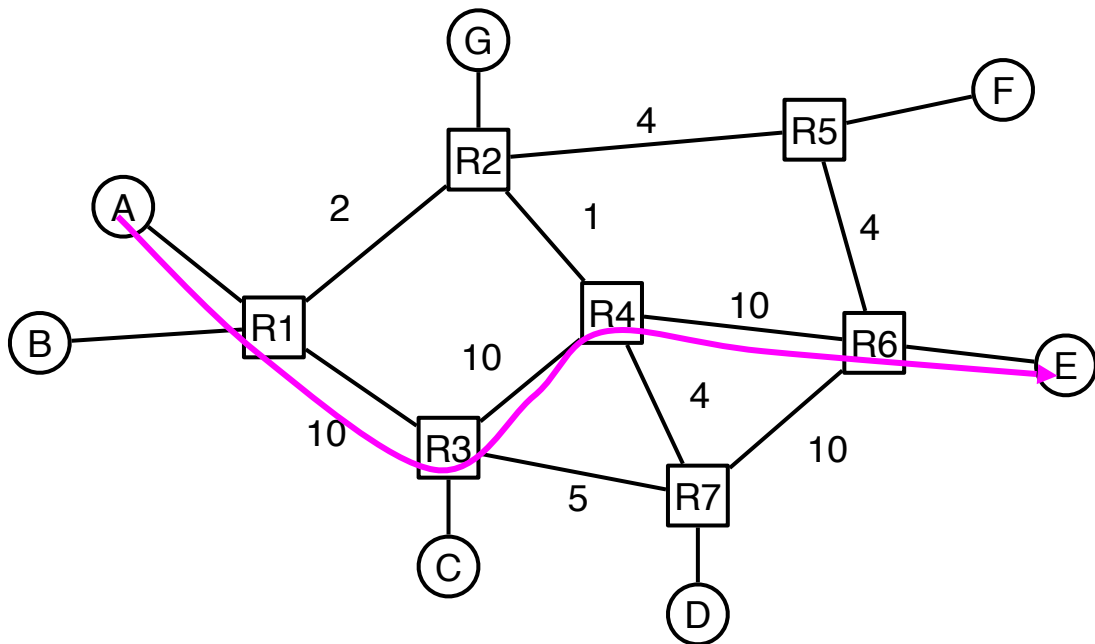
Congestion Control Principles

- Why do we need it?
- Why is it hard?
- Design Goals

Dynamic Adjustment

- **Algorithm Sketch**
- Reacting to Congestion (AIMD)
- Implementing Congestion Control with Event-Driven Updates

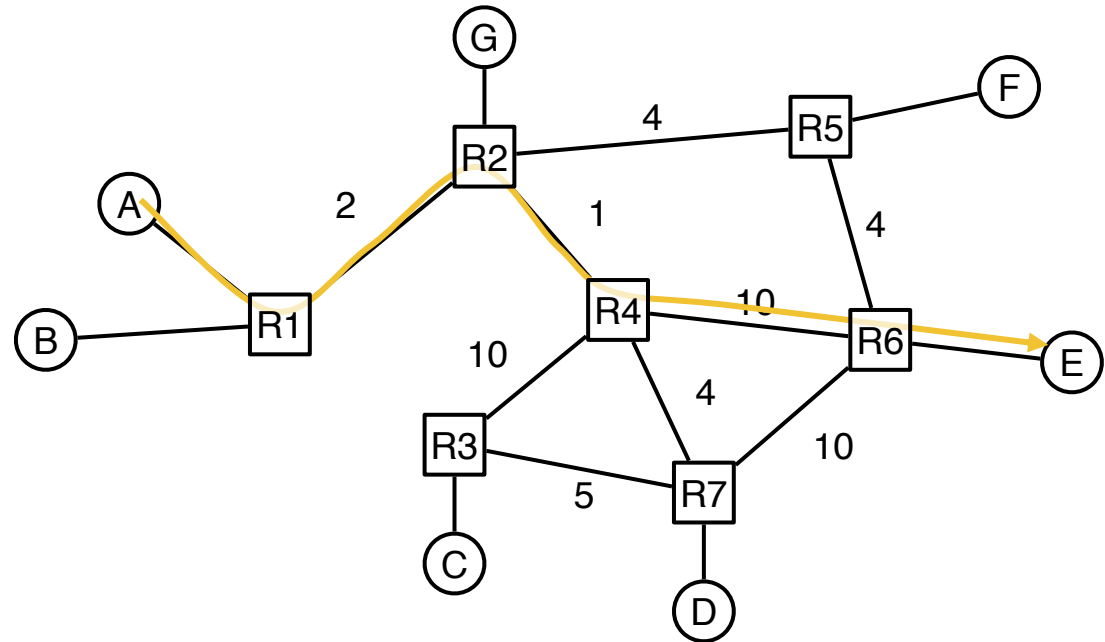
Discovery: Host A discovers it can send data at 10 Gbps.



Later, the $R1 \rightarrow R3$ link goes down. A's traffic is rerouted.

Detection: A notices that 10 Gbps is congesting the network.

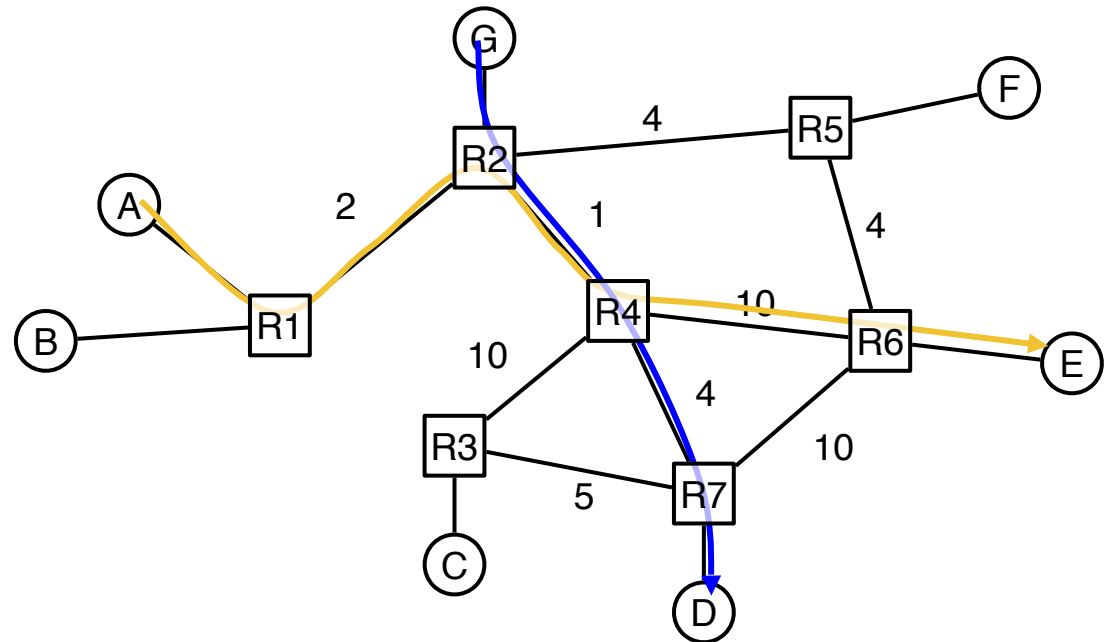
Adjustment: A decides to slow down to 1 Gbps.



Later, $G \rightarrow D$ starts sending traffic.

Detection: A notices that 1 Gbps is congesting the network.

Adjustment: A decides to slow down to 0.5 Gbps.



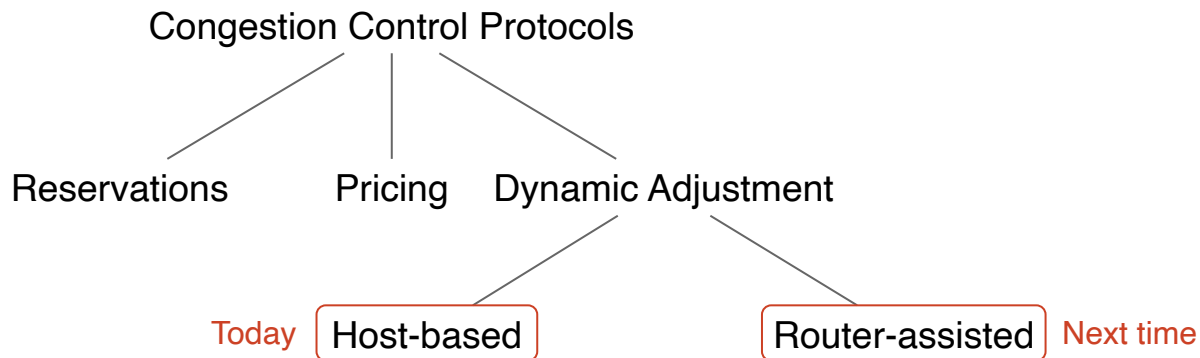
Solutions can be classified into 2 types, depending on where they're implemented:

Host-based congestion control: (*TCP uses this one*)

- Implemented at the end hosts. No special support from routers.
- Hosts adjust rate based on *implicit* feedback from routers.

Router-assisted congestion control:

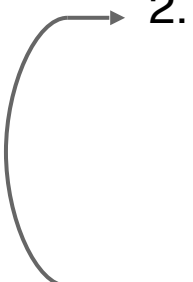
- Routers signal congestion back to end hosts.
- Hosts pick rate based on *explicit* feedback from routers.



Every host *independently* runs the same algorithm:

1. Pick an initial rate R .

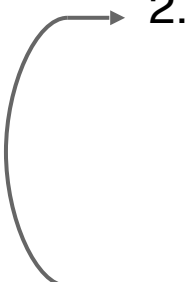
2. Try sending at rate R for some period of time.

- Did I experience congestion in this time period?
 - If no, increase R .
 - If yes, reduce R .
 - Repeat.
- 

Every host *independently* runs the same algorithm:

1. Pick an initial rate R . (*How do we pick the initial rate?*)

2. Try sending at rate R for some period of time.

- Did I experience congestion in this time period? (*How do we detect congestion?*)
 - If no, increase R .
 - If yes, reduce R . (*How much do we increase/decrease by?*)
 - Repeat.
- 

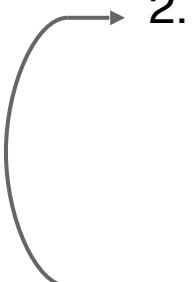
Components of a solution:

- Discovering an initial rate.
- Detecting congestion.
- Reacting to congestion (or lack thereof) by increasing/decreasing rate.

Every host *independently* runs the same algorithm:

1. Pick an initial rate R . (*How do we pick the initial rate?*)

2. Try sending at rate R for some period of time.

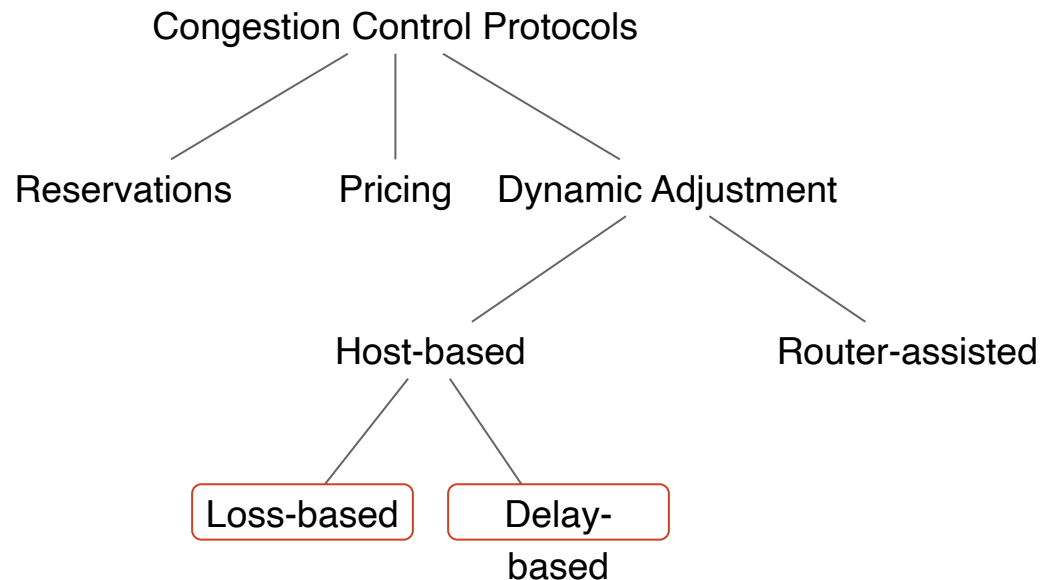
- Did I experience congestion in this time period? (*How do we detect congestion?*)
 - If no, increase R .
 - If yes, reduce R . (*How much do we increase/decrease by?*)
 - Repeat.
- 

Components of a solution:

- Discovering an initial rate.
- **Detecting congestion.**
- Reacting to congestion (or lack thereof) by increasing/decreasing rate.

There are 2 ways for the host to detect congestion.

Remember: The host can't see the whole network.



Loss-based detection: If we lose a packet, declare congestion.

- This is what TCP uses.

Benefits:

- Unambiguous signal: Every packet is either lost (resent), or not lost.
- TCP already detects loss for reliability.

Cons:

- Not all loss is caused by congestion. (e.g. checksum error).
- A delayed packet could be mistakenly declared lost.
- When we detect loss, the network is already congested (queues are full).

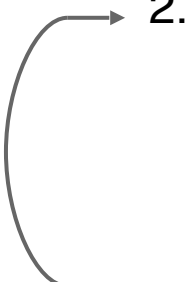
Delay-based detection: If packets are delayed, declare congestion.

- Challenge: Packet delay length varies with queue size and other traffic.
- Long considered tricky to get right.
- Google's BBR protocol (2016) is challenging this assumption.

Every host *independently* runs the same algorithm:

1. Pick an initial rate R . (*How do we pick the initial rate?*)

2. Try sending at rate R for some period of time.

- Did I experience congestion in this time period? (*How do we detect congestion?*)
 - If no, increase R .
 - If yes, reduce R . (*How much do we increase/decrease by?*)
 - Repeat.
- 

Components of a solution:

- **Discovering an initial rate.**
- Detecting congestion: Packet loss or delay.
- Reacting to congestion (or lack thereof) by increasing/decreasing rate.

Goal for discovering initial rate: Estimate available bandwidth.

Start slow for safety.

- Starting too fast would cause congestion.

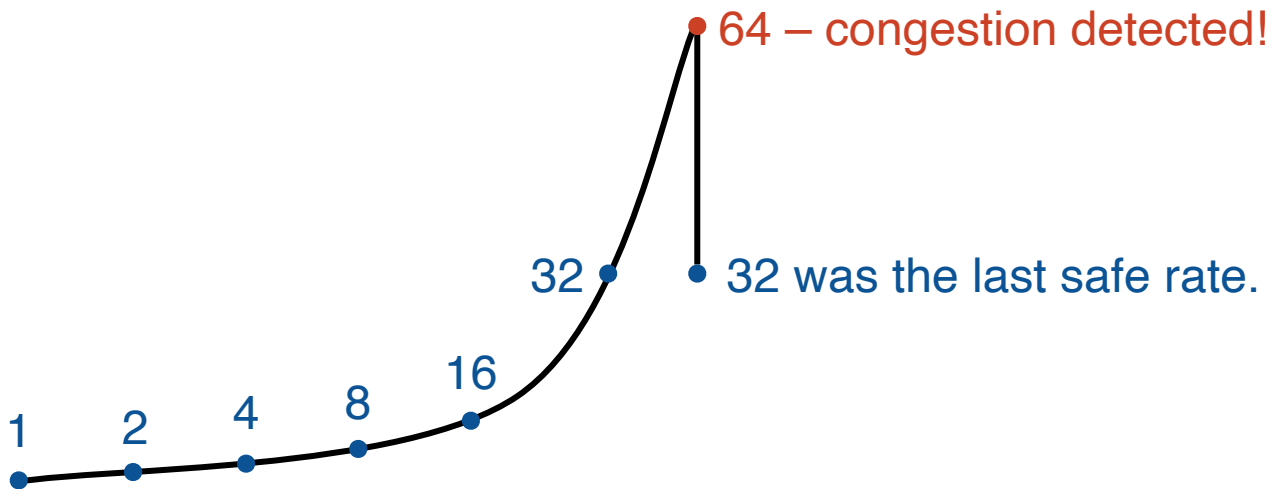
Ramp up quickly for efficiency.

- Example of inefficiency: Suppose we added 0.5 Mbps every 0.1 seconds.
- If 1 Mbps available, we'll reach the target in 0.2 seconds.
- If 1 Gbps available, we'll reach the target in 200 seconds.
 - Lots of time wasted sending at suboptimal rates.

What kind of mathematical function starts slow but ramps up quickly?

Slow start is an algorithm for discovering the initial rate:

- Start with a small rate (could be much less than actual bandwidth).
- Increase exponentially (e.g. double rate each time) until we detect loss.
- A safe rate is the most recent rate before detecting loss.
 - $0.5 \times$ rate when loss occurred.



Reacting to Congestion

Lecture 12, CS 168, Spring 2026

Congestion Control Principles

- Why do we need it?
- Why is it hard?
- Design Goals

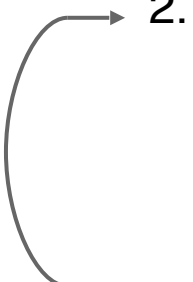
Dynamic Adjustment

- Algorithm Sketch
- **Reacting to Congestion (AIMD)**
- Implementing Congestion Control with Event-Driven Updates

Every host *independently* runs the same algorithm:

1. Pick an initial rate R . (*Slow start.*)

2. Try sending at rate R for some period of time.

- Did I experience congestion in this time period? (*Detect packet loss.*)
 - If no, increase R .
 - If yes, reduce R . (*How much do we increase/decrease by?*)
 - Repeat.
- 

Rate adjustment: How much do we increase/decrease by?

- Determines how quickly a host adapts to changes in available bandwidth.
- Determines how effectively bandwidth is consumed.
- Determines how bandwidth is shared (fairness).

Goals for rate adjustment:

- Efficiency: High utilization of link bandwidth.
- Fairness: Each flow gets an equal share of bandwidth.

At a high level, we can either adjust quickly or slowly.

- Fast: **Multiplicative** changes.
 - Increase by doubling: $R \rightarrow 2R$
 - Decrease by halving: $R \rightarrow R/2$
- Slow: **Additive** changes.
 - Increase by adding: $R \rightarrow R + 1$
 - Decrease by subtracting: $R \rightarrow R - 1$

We can combine these rates in four ways:

- | | | |
|---------|--------------------------|--------------------------|
| ● AIAD: | Additive Increase, | Additive Decrease. |
| ● AIMD: | Additive Increase, | Multiplicative Decrease. |
| ● MIAD: | Multiplicative Increase, | Additive Decrease. |
| ● MIMD: | Multiplicative Increase, | Multiplicative Decrease. |

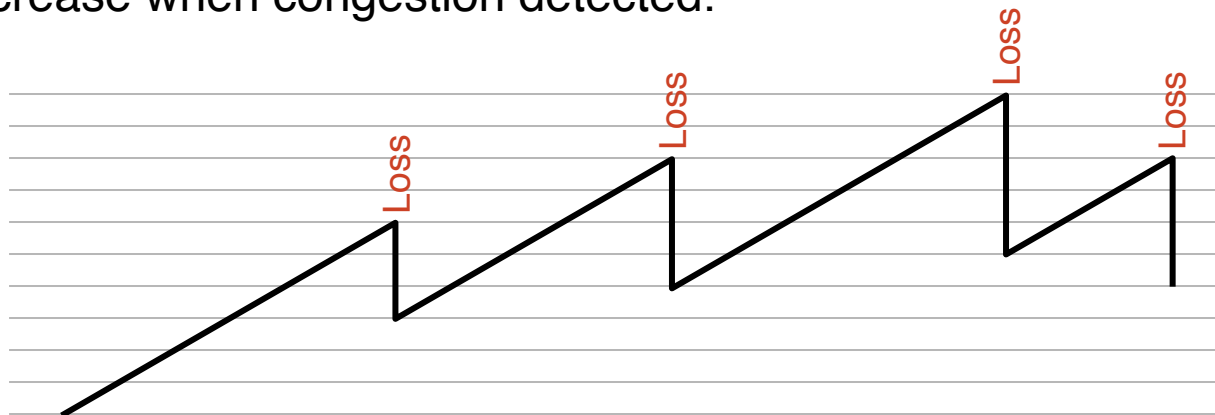
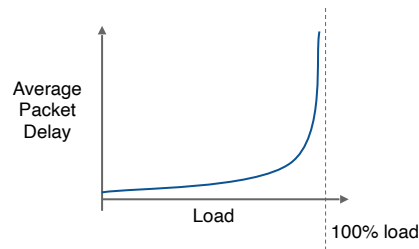
Non-obvious fact: AIMD (Additive Increase, Multiplicative Decrease) is the best option.

Intuition: Sending too much is worse than sending too little.

- Sending too much: Congestion, packets dropped and retransmitted.
- Sending too little: Somewhat lower throughput.

General approach:

- Gentle increase when uncongested. (exploration)
- Rapid decrease when congestion detected.



Non-obvious fact: Of the four options, AIMD is the only one that ensures fairness.

To see why, let's build a simple model:

- Two flows using a single link of fixed capacity C .
- Both flows are *independently* running *identical* dynamic adjustment algorithms.
- Flows are sending at rates X and Y .
 - If $X + Y > C$, the network is congested.
 - If $X + Y < C$, the network is underloaded.
- Goals:
 - Full link utilization: $X + Y = C$.
 - Fair sharing: $X = Y$.

Assume $C = 1$.

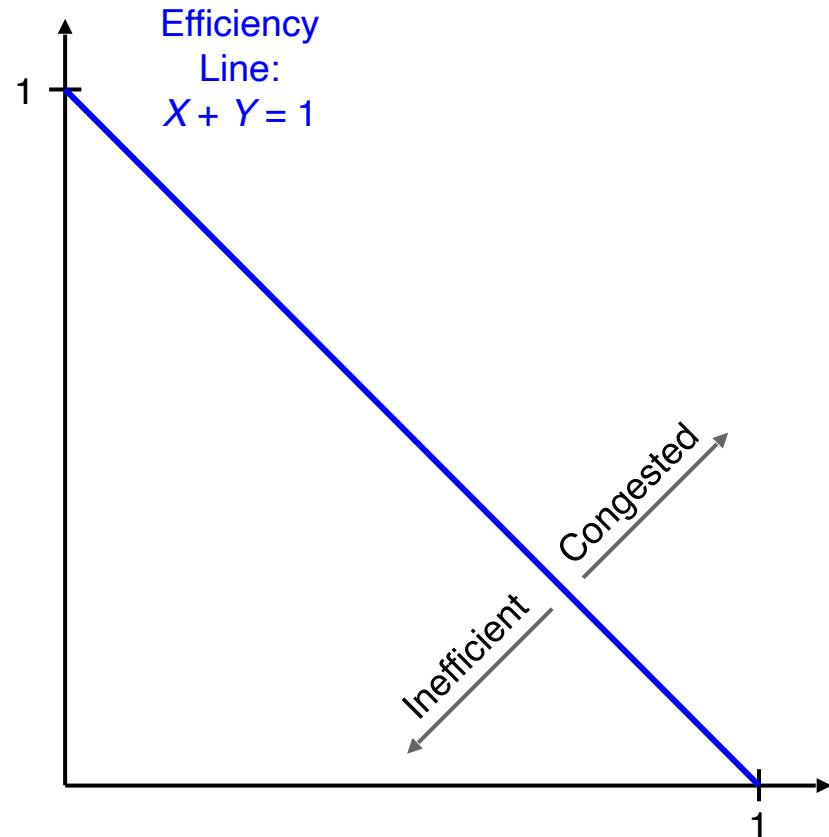
Efficiency line: $X + Y = 1$

Congestion: $X + Y > 1$

Everything above the efficiency line.

Unused capacity: $X + Y < 1$

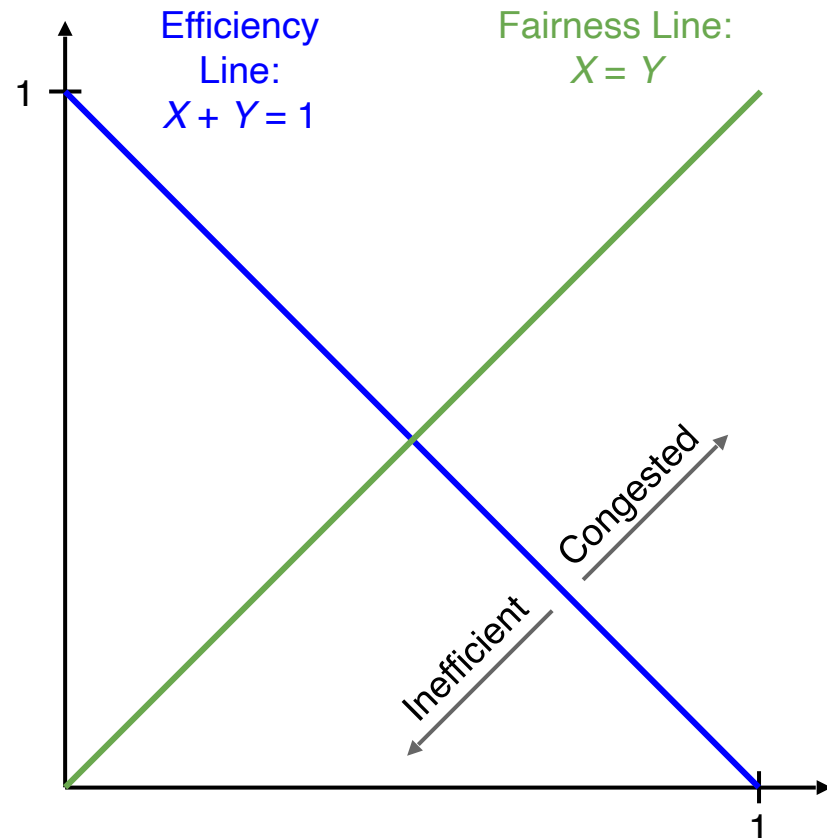
Everything below the efficiency line.



Fairness line: $X = Y$

The lines intersect at $(0.5, 0.5)$.

This allocation is both fair and efficient.



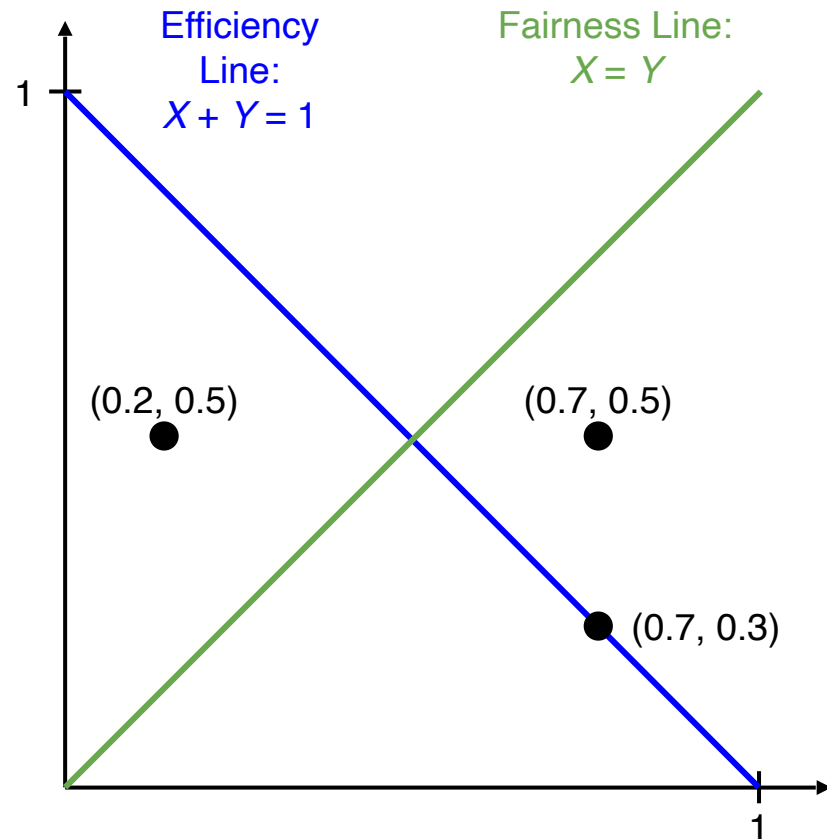
Rate Adjustment Graph

(0.2, 0.5) is inefficient (below blue line).

(0.7, 0.5) is congested (above blue line).

(0.7, 0.3) is efficient, but unfair.

On blue line, but not on green line.



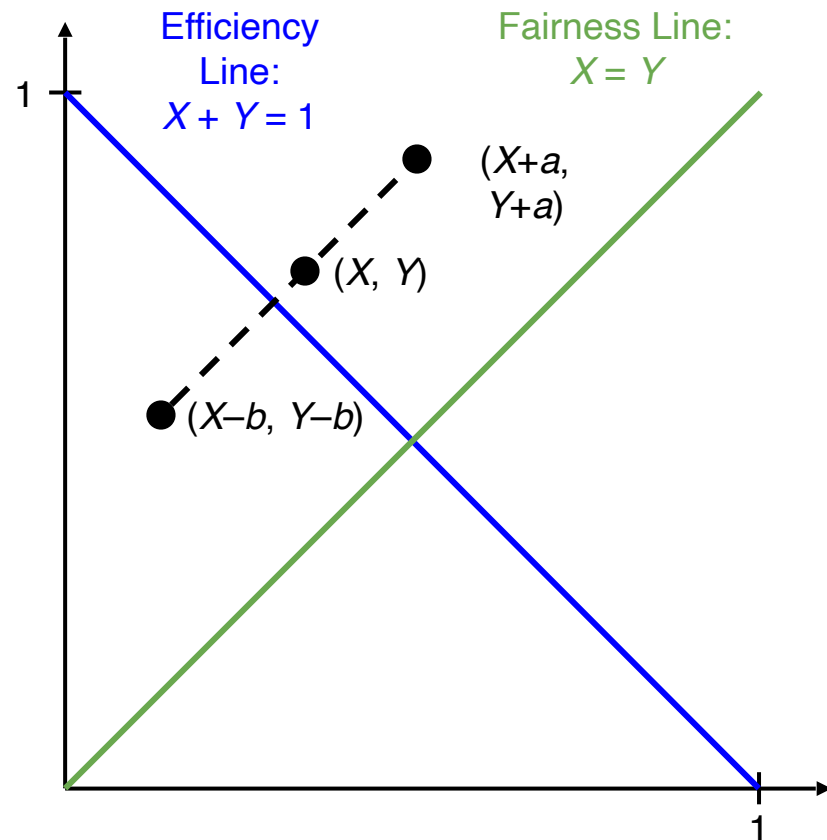
Recall: Both flows *independently* running *identical* algorithms.

Suppose current allocation is (X, Y) .

If both increase additively: $(X + a, Y + a)$.

If both decrease additively: $(X - b, Y - b)$.

Notice: Allocations always stay on a line with slope = 1.



Increase: $+a$

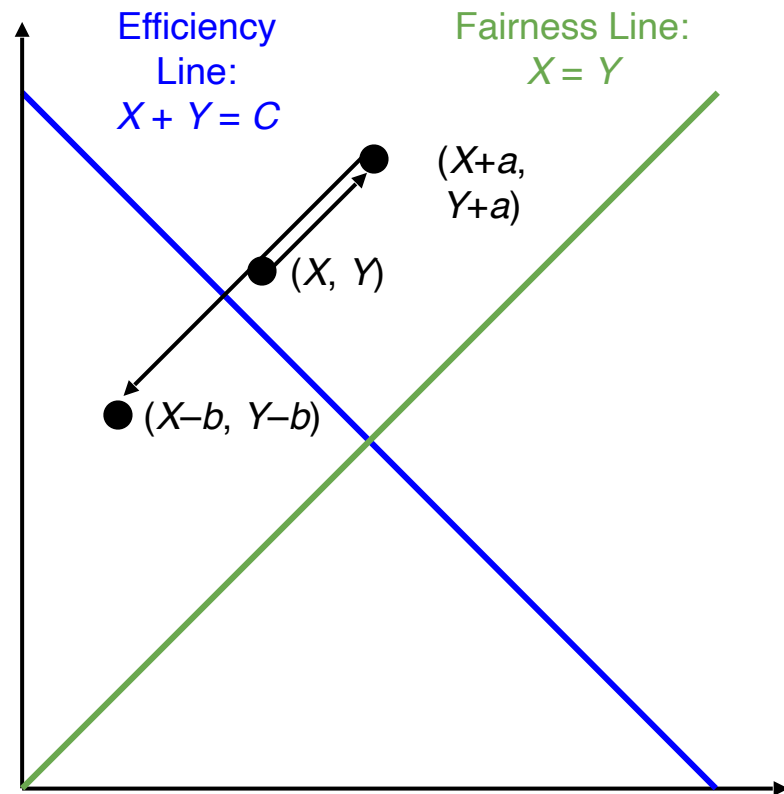
Decrease: $-b$

We keep moving along the line with slope 1.

The *difference* in allocations stays the same!

Here $Y-X$.

AIAD maintains the same unfairness across iterations and does not converge to fairness.



Recall: Both flows *independently* running *identical* algorithms.

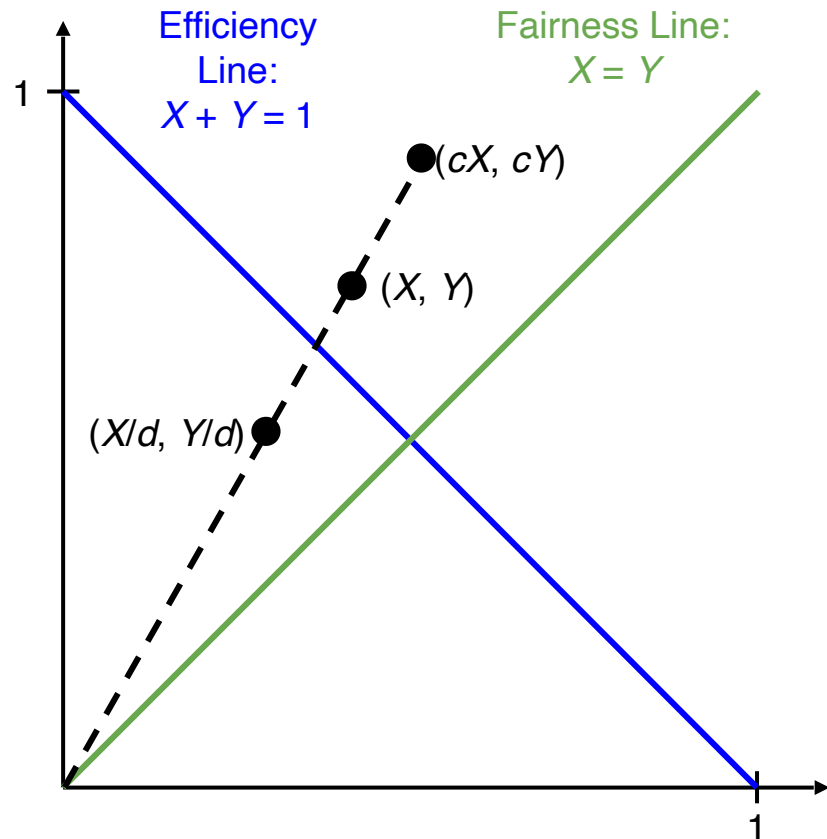
Suppose current allocation is (X, Y) .

If both increase multiplicatively: (cX, cY) .

If both decrease multiplicatively: $(X/d, Y/d)$.

Notice: Allocations always stay on a line with slope $= Y/X$, going through the origin.

$$\frac{\text{Rise}}{\text{Run}} = \frac{cY - Y}{cX - X} = \frac{Y(c-1)}{X(c-1)} = \frac{Y}{X}$$



MIMD (Multiplicative Increase, Multiplicative Decrease) Adjustments on Graph

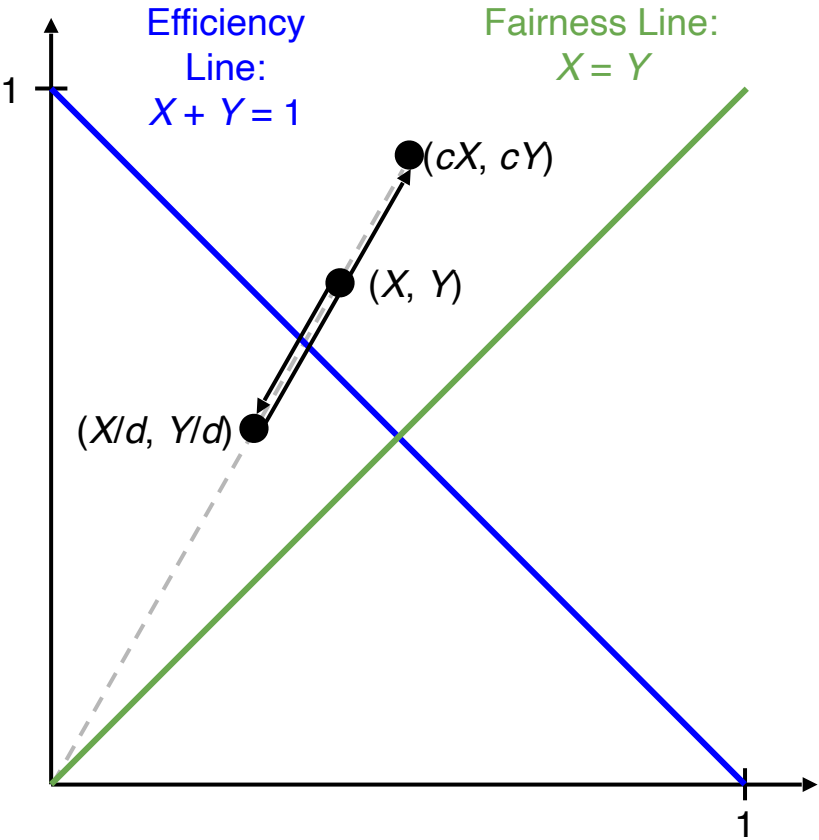
Increase: $\times c$

Decrease: $\div d$

We keep moving along the line connected to the origin.

The *ratio* of allocations stays the same!
Here, Y/X .

MIMD maintains the same unfairness across iterations and does not converge to fairness.



Increase: $\times 2$	Decrease: -1		$C = 5$	
Start:	$X = 1$	$Y = 3$	sum = 4, no congestion	$Y - X = 2$
First iteration:	$X = 2$	$Y = 6$	sum = 8, congestion	$Y - X = 4$
Second iteration:	$X = 1$	$Y = 5$	sum = 6, congestion	$Y - X = 4$
Third iteration:	$X = 0$	$Y = 4$	sum = 4, no congestion	$Y - X = 4$
Fourth iteration:	$X = 0$	$Y = 8$	X stuck at 0 forever!	

The gap in allocation multiplies when we increase. before: $Y - X$. after: $2Y - 2X = 2(Y - X)$

The gap in allocation stays the same when we decrease. (Same as AIAD.)

MIAD is *maximally unfair*. The gap increases until one person has all the bandwidth!

AIMD (Additive Increase, Multiplicative Decrease) Dynamics

Increase: +1		Decrease: ÷2		$C = 5$	
Start:		$X = 1$	$Y = 2$	sum = 3, ok	$Y - X = 1$
First iteration:	$X = 2$	$Y = 3$		sum = 5, ok	$Y - X = 1$
Second iteration:	$X = 3$	$Y = 4$		sum = 7, congestion	$Y - X = 1$
Third iteration:	$X = 1.5$	$Y = 2$		sum = 3.5, ok	$Y - X = 0.5$
Fourth iteration:	$X = 2.5$	$Y = 3$		sum = 5.5, congestion	$Y - X = 0.5$
Fifth iteration:	$X = 1.25$	$Y = 1.5$		sum = 2.75, ok	$Y - X = 0.25$
Sixth iteration:	$X = 2.25$	$Y = 2.5$		sum = 4.75, ok	$Y - X = 0.25$
Seventh iteration:	$X = 3.25$	$Y = 3.5$		sum = 6.75, congestion	$Y - X = 0.25$
Eighth iteration:	$X = 1.625$	$Y = 1.75$		sum = 3.375, ok	$Y - X = 0.125$
Ninth iteration:	$X = 2.625$	$Y = 2.75$		approaching 2.5 each	$Y - X = 0.125$

The gap in allocation stays the same when we increase.

The gap in allocation halves when we decrease.

- Before: $Y - X$
- After: $(0.5 Y - 0.5 X) = 0.5(Y - X)$

The gap approaches 0, so AIMD approaches fairness!

$$Y - X = 1$$

$$Y - X = 1$$

$$Y - X = 1$$

$$Y - X = 0.5$$

$$Y - X = 0.5$$

$$Y - X = 0.25$$

$$Y - X = 0.25$$

$$Y - X = 0.25$$

$$Y - X = 0.125$$

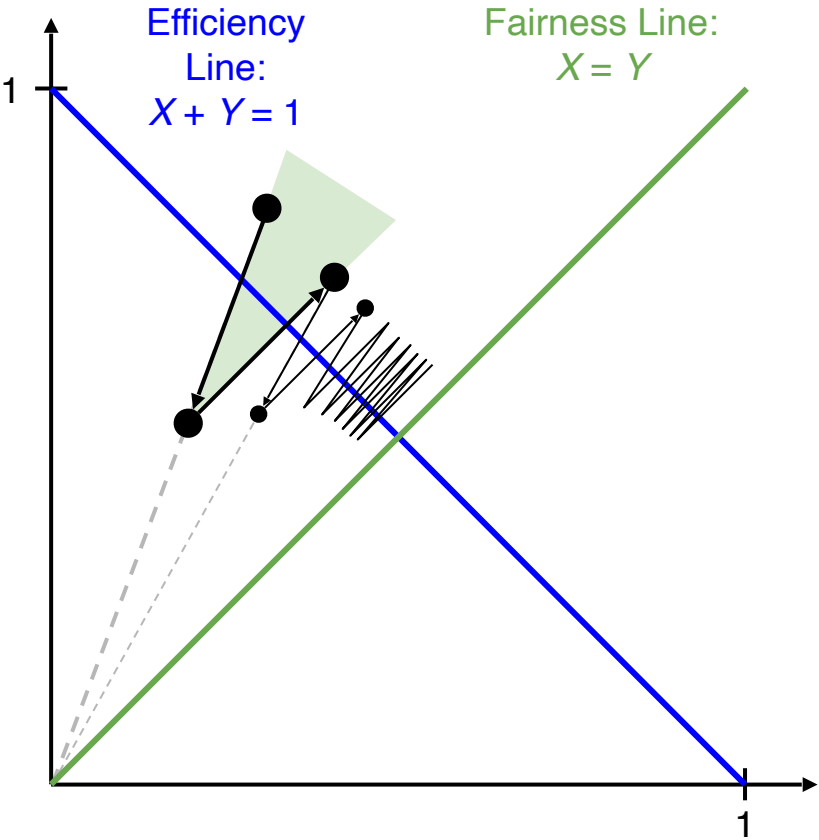
$$Y - X = 0.125$$

AIMD (Additive Increase, Multiplicative Decrease) Adjustments on Graph

Increase: $+a$

Decrease: $\div d$

AIMD converges to fairness!



Non-obvious fact: Of the four options, AIMD is the only one that ensures fairness.

↖ Hopefully more obvious now!

AIMD embodies gentle increase, rapid decrease.

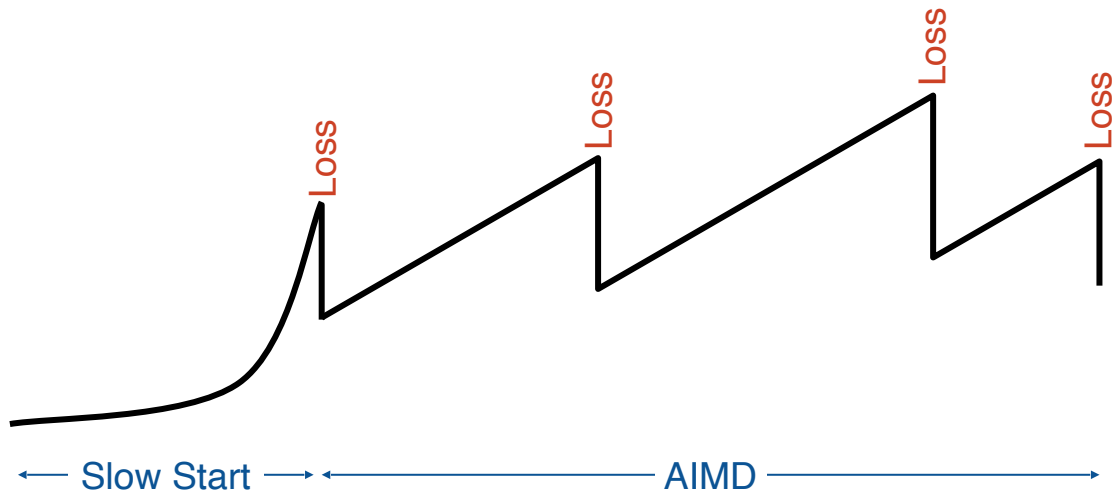
Out of the four options:

- AIAD and MIMD retain unfairness.
- MIAD approaches maximum unfairness (one person with all the bandwidth).
- AIMD approaches fairness across iterations.

Every host *independently* runs the same algorithm:

1. Pick an initial rate R . (*Slow start.*)
2. Try sending at rate R for some period of time.
 - Did I experience congestion in this time period? (*Detect packet loss.*)
 - If no, increase R . (*Additive increase.*)
 - If yes, reduce R . (*Multiplicative decrease.*)
 - Repeat.

This algorithm creates a sawtooth-shaped graph as the rate is adjusted.



Implementing Congestion Control with Event-Driven Updates

Lecture 12, CS 168, Spring 2026

Congestion Control Principles

- Why do we need it?
- Why is it hard?
- Design Goals

Dynamic Adjustment

- Algorithm Sketch
- Reacting to Congestion (AIMD)
- **Implementing Congestion Control with Event-Driven Updates**

The sender maintains a window of W packets in flight.

Pick window size W to balance three goals:

1. Take advantage of network capacity ("fill the pipe").
 - Not actually used (impractical to calculate, and never the minimum).
2. Don't overload the recipient.
 - Receiver window (RWND): Recipient reports how much space it has left.
3. Don't overload links.
 - Congestion window (CWND): Sender runs a congestion control algorithm.

W is the minimum of RWND and CWND.

- For today, we'll assume $RWND > CWND$, so window is determined by CWND.
- Often true in practice too.

Real TCP measures data in bytes.

- Thus, real TCP maintains CWND in bytes.

In this lecture, we'll measure CWND in packets for convenience.

- To convert to bytes, recall: 1 packet = MSS bytes.
- MSS = Maximum Segment Size.

Congestion control is used to set the window size (CWND).

We talked about congestion control as adjusting the rate.

Possibly confusing fact: Window size and rate are directly correlated.

- Recall: $\text{Rate} = \text{CWND} / \text{RTT}$.
- Key thing to remember: Larger window = higher rate.