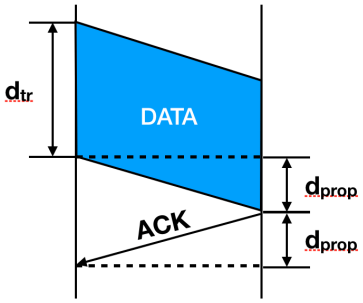
Transport Layer (II)

Utilization: Stop & Wait vs. Sliding Window (cont.)

- Utilization is also called *normalized throughput*.
  - It's the ratio between goodput and throughput.
    - goodput: "useful" bytes delivered per second
    - throughput: total number of bytes that can be delivered per second

- Stop and Wait Utilization



$$Utilization = \frac{d_{tr}}{d_{tr} + 2d_{prop}} = \frac{1}{1 + 2\frac{d_{prop}}{d_{tr}}}$$

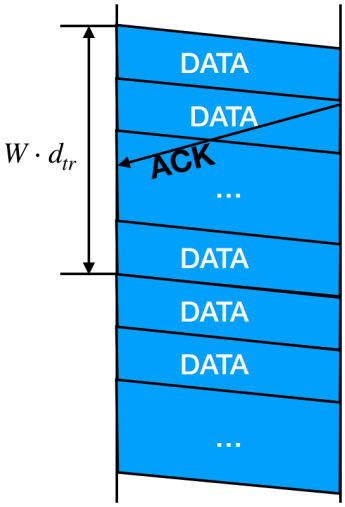
- Please note that the transmission rate is a constant, so the utilization can be calculated using transmission delay and propagation delay.

- Sliding Window Utilization

CASE 1

$$2d_{prop} < (W-1) \cdot d_{tr}$$

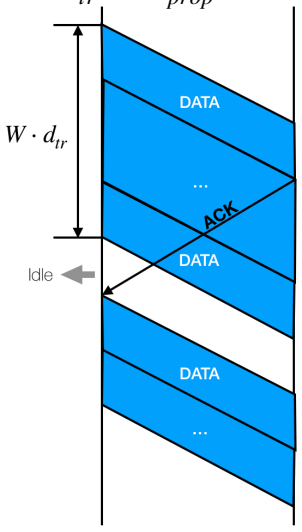
$$Utilization = 100\,\%$$



CASE 2

$$2d_{prop} \geq (W-1)d_{tr}$$

$$Utilization = \frac{W \cdot d_{tr}}{d_{tr} + 2d_{prop}} = W \cdot \frac{1}{1 + 2\frac{d_{prop}}{d_{tr}}}$$



- Exercise: Alice sends data to Bob over a 10 Mbps link. Each packet contains 1,000 bytes, and the one-way propagation delay is 20 ms. ACK transmission and processing times are negligible.
  - Calculate: 1) the transmission delay, 2) Stop-and-Wait utilization, 3) Sliding Window utilization for W=10, and 4) the minimum window needed for full utilization.

$$d_{tr} = \frac{8000}{10^7} = 0.8ms$$

$$U_{SW} = \frac{0.8}{0.8 + 2 \times 20} = 0.00196 = 1.96\,\%$$

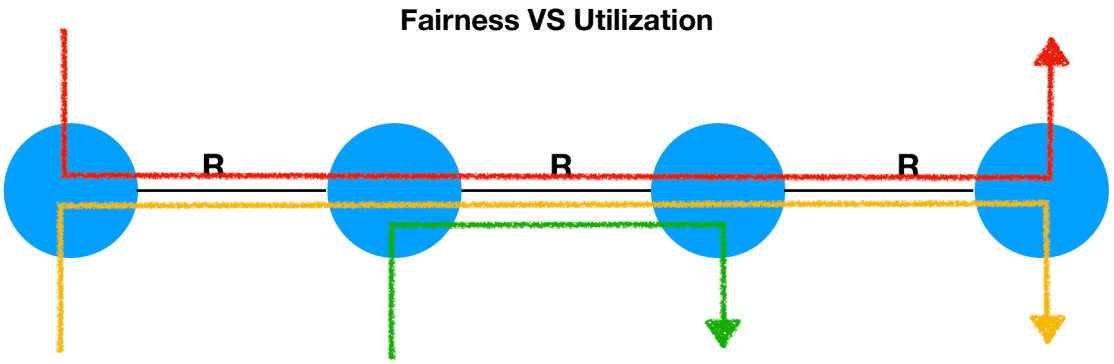
$$U_{wind,10} = \frac{10 \times 0.8}{0.8 + 2 \times 20} = 0.01961 = 19.61\,\%$$

$$W_{min} = \lceil \frac{0.8 + 2 \times 20}{0.8} \rceil = 51$$

  - Follow-up question: Why can a faster link actually make Stop-and-Wait utilization worse when packet size and propagation delay stay constant?
    - Faster transmission reduces d_{tr}, but propagation delay remains unchanged. The sender spends an even larger fraction of time waiting for acknowledgments.

Fairness and Utilization

- Flow control (protect receiver), congestion control (protect network): same fundamental issue
- Goal: Make the most effective use of the network capacity
  - avoid congestion
  - maximize utilization
  - maintain fairness (or deliver promised service level)



- If the red and yellow traffic join the network first, they will maximize each of their utilization while keeping the fairness of the usage. Therefore, each of them will get a half of the bandwidth. When the green traffic join the network, they have to adjust the usage to keep the fairness while maximizing utilization among them.

Traffic	Max Utilization	Fair
—	R/2	R/3
—	R/2	R/3
—	θ	R/3

- Exercise: Suppose three flows share a 12 Mbps bottleneck link. All three always have data to send, and they have equal priority.

  - If only Alice and Bob are transmitting, what is a fair allocation? What changes when Charlie joins?
    - 6 Mbps each initially; 4 Mbps each when all three compete under an equal-share fairness model.

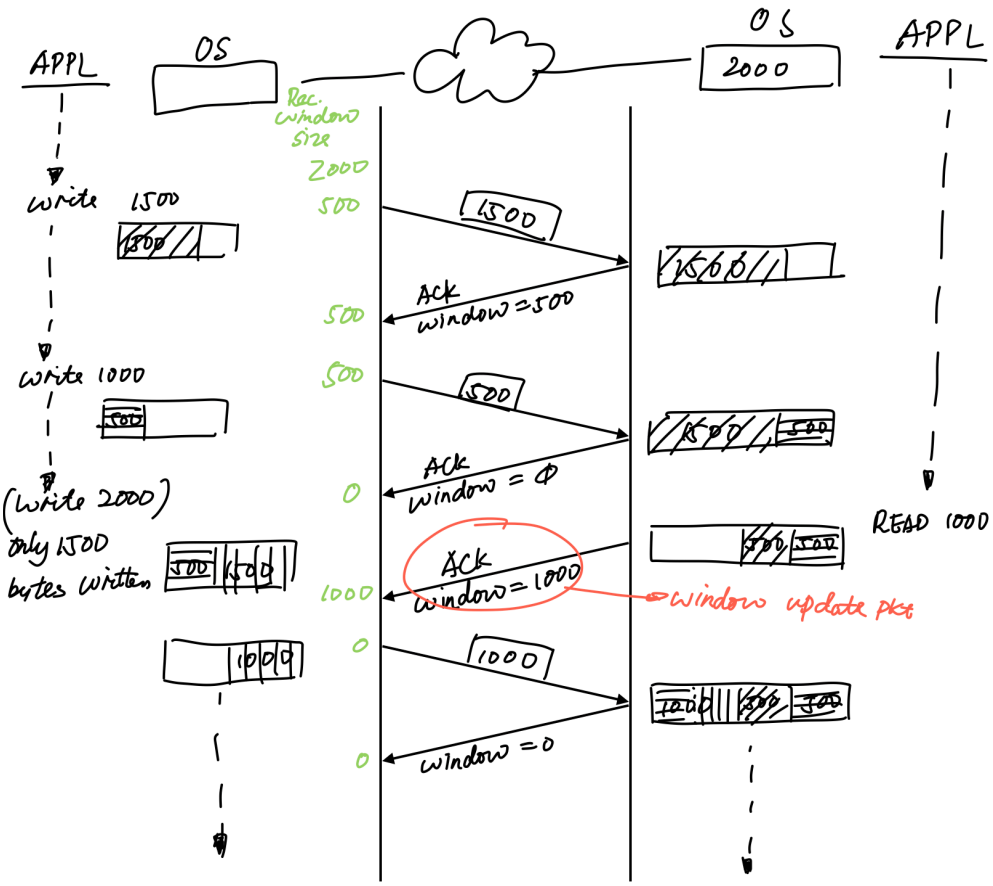
  - Is maximizing total throughput always the same as maximizing fairness?
    - No. A network can be fully utilized while allocating almost all capacity to one flow. Fairness is a separate objective.

Two Different Reasons a Sender Must Slow Down

- There are two different reasons a sender must slow down: Receiver Congestion and Network Congestion.

- Receiver Congestion:
  - receiver congestion happens when the receiver is unable to keep up with incoming data
  - this congestion can be solved by explicit feedback from receiver to sender

  - In the example below, after building up the connection, the sender will know that the receiver window size is 2000 bytes. So, the sender won't transmit more than the available receiver window size per time. Suppose the sender's window size is also 2000 bytes in this example. The sender wants to send three messages to the receiver in this example. The three message sizes are 1500 bytes, 1000 bytes, and 2000 bytes. The first message can be sent all at once, as the receiver has 2000-byte window available. After receiving the first message, the receiver only has 500-byte available window, and it will inform the sender the size of the available window via acknowledgment. So, when the sender sends the second



message, it can only send 500 bytes out of the 1000-byte message to the receiver first, and then wait for the window becomes available again. Once the receiver READs 1000 bytes from the winder, there are 1000 bytes in the receiver winder become available again. Meanwhile, the sender wants to send the third message that is 2000 bytes. The sender window is then filled up by the remaining 500 bytes from the second message and 1500 bytes from the third message. The remaining 500 bytes of the third message will be written into the sender window when there is available space. Since only 1000 bytes become available in the receiver window, the sender will transmit 1000 bytes of its window to the receiver. The remaining 1000 bytes in the sender's window coupled with the remaining 500 bytes of the third message will be transmitted when the receiver window becomes available.

- Network Congestion
  - network congestion happens when nodes or links of the network are overloaded
  - it can be solved by either explicit congestion notification or implicit congestion notification.
    - explicit congestion notification can be achieved via a few technologies
      - routers provide explicit feedback to the sender and/or receiver regarding the congestion state of the network. The feedback may be just a single bit.

    - a router inform the sender of the maximum host sending rate it can support on an outgoing link.
- implicit congestion notification is usually used by the Internet, which doesn't required explicit notification but can be inspected using other metrics such as RTT and packet loss.
    - Round Trip Time (RTT)
        - time between data packet transmission and reception of its acknowledgment
        - increase in RTT could be interpreted as due to an increase in queue lengths in nodes (congestion)
        - or it could be due to a route change …
    - Packet Loss
        - packet loss due to queue overflow (congestion)
        - packet loss due to random packet errors

Disclaimer: Notes adapted from the textbook and online resources.