EE4204 Homework 2

Zhuang Jianning
A0214561M

1. Retransmission of lost data can be done at the link, transport, and application layers. What are the pros and cons of doing it at each layer?

Link Layer

+ Link layer retransmissions only need to retransmit over the current link and not the entire end to end path, thus increasing performance with lower latency and less bandwidth wastage.

+ Link layer retransmissions can prevent error losses from being misinterpreted by TCP as congestion losses.

- End to end retransmission is still necessary and may cause retransmission at multiple layers simultaneously, which may interfere with each other.

Transport Layer

+ Transport layer retransmissions allow applications to use the same transport protocol and avoid having to reimplement retransmission functionality in applications. This can help reduce bugs and implementation time.

- Retransmissions can be inefficient and slow as a packet can only be retransmitted by the sender

Application Layer

+ Satisfies the end-to-end argument as the application is in the best position to know what piece of data needs to be retransmitted, and what are the constraints.

+ Data which may be lost above the transport layer can be retransmitted.

- Increases the complexity of applications.

2. Compare Implicit versus Explicit congestion signals. What are the advantages and disadvantages of each?

Implicit Congestion Signals

In implicit signaling, there is no communication between the congested nodes and the source. The source guesses that there is congestion in a network. For example when the sender sends several packets and there is no acknowledgement for a while, one assumption is that there is congestion.

+ Implicit congestion signals do not require router support

+ Transport protocols have to interpret implicit congestion signals

- Implicit congestion signals cannot be distinguished and can even be misinterpreted by the scheme

Explicit Congestion Signals

In explicit signaling, if a node experiences congestion it can explicitly send a packet to the source or destination to inform about congestion. The signal is included in the packets that carry data rather than creating a different packet as in case of choke packet technique. Explicit signaling can occur in either forward or backward direction.

+ Explicit congestion signals have lower delay when interpreting congestion compared to implicit congestion signals. E.g TCP only requires 3 duplicate ACKs to determine a packet is lost
+ Explicit congestion signals can determine what kind of loss occurred
+ Explicit congestion signals can be used for congestion avoidance

- Explicit congestion signals can be lost

3. What are the main functions of the transport layer? Describe briefly.

The Transport Layer ensures reliable transfer of data between end users by managing the delivery and error checking of data packets. It regulates the size, sequencing, and ultimately the transfer of data between systems and hosts to ensure end to end reliability. Layer 4 protocols include TCP (Transmission Control Protocol) and UDP (User Datagram Protocol).

The main functions include Multiplexing and Demultiplexing, Flow Control, Error Control, Segmentation and Reassembly and Connection Control

4. How does the transport layer perform multiplexing and demultiplexing?

Each TCP/UDP segment has source and destination port numbers. In multiplexing, data is encapsulated with transport headers which are differentiated by their port numbers. The resulting segments are then passed to the network layer. In demultiplexing, hosts use the IP address and port number to direct the segment to the appropriate socket.

Each socket in a host can be assigned a port number. In the TCP/IP protocol, port numbers are 16-bit numbers ranging from 0 to 65,535. Well known port numbers start from 0 to 1023 and they are reserved for well-known application protocols.

5. Why does TCP wait for three duplicate acknowledgments before retransmitting a packet? What do the triple duplicate acks represent?

In fast retransmission, which addresses the inefficiency of timeout based retransmission, the sender does not wait until the timer expires but only until 3 duplicate ACKs before retransmitting data.

If three or more duplicate ACKs are received in a row, it is a strong indication that a segment has been lost. TCP then performs a retransmission of what appears to be the missing segment, without waiting for a retransmission timer to expire.

6. How does TCP set its timeout value?

TCP sets its timeout value as a function of RTT. RTT can be estimated by watching the ACKs and by keeping a running average with the formula:

EstimatedRTT = a*EstimatedRTT + (1-a)*SampleRTT

Timeout can be set as 2*EstimatedRTT

To improve estimated RTT, only samples for segments sent a single time are collected so retransmissions causing large sample RTT and duplicate packets which cause small sample RTT will not be considered.

7. TCP congestion avoidance is done via AIMD. Explain.

TCP congestion avoidance is done via Additive Increase Multiplicative Decrease (AIMD). AIMD increases the congestion window by 1 if the entire window's worth of packets in a RTT is ACKed without error and cuts the congestion window in half for every packet that is lost which results in a timeout. This adaptively probes the network to expand the congestion window until a segment is lost, which contracts the congestion window. AIMD is usually used after slow start reaches threshold CW/2 after previous timeout occurs so as to be cautious about adding more data packets near the old congestion point. This allows for congestion avoidance by backing off before there are serious packet losses.

8. What is the goal of network fairness? Is TCP fair? If so, explain what resources TCP allocates in a fair manner.

Fairness in computer networks deals with the distribution of network resources among applications and is achieved when network resources are distributed in a fair way.

TCP fairness requires that a new protocol receive a no larger share of the network than a comparable TCP flow. This is important as TCP is the dominant transport protocol on the Internet, and if new protocols acquire unfair capacity they tend to cause problems such as congestion collapse.

TCP which uses AIMD ensures fairness because hosts increase their bandwidth additively, but when congestion occurs they drop their bandwidth multiplicatively. So hosts with higher share of bandwidth lose the most. As this additive increase and multiplicative decrease continues, bandwidths converge to fair amounts.

9. What is the throughput of TCP? The throughput is the average rate that packets are successfully decoded at the receiver. Note that the rate that packets are sent by the sender is an upper bound on the actual throughput and since it is easily computable, we use it to estimate the throughput.

Throughput is the number of bits transferred per unit time. The throughput is the average rate that packets are successfully decoded at the receiver. Note that the rate that packets are sent by the sender is an upper bound on the actual throughput and since it is easily computable, we use it to estimate the throughput.

Since TCP keeps a window size to control traffic and prevent congestion, a TCP flow with window size W allows W packets per RTT. Hence throughput is W packets per RTT time

10. We said that the goal of the transport layer (layer 4) was end-to-end reliability. Recall that layer 2 also has reliability, but it is hop-by-hop reliability. Why do we have reliability at both layers? Are they both necessary?

A link-layer reliability is often used for links that are prone to high error rates such as wireless links rather than forcing an end-end retransmission of data by transport or application layer protocol.

End-to-end reliability is necessary because not all network layers are reliable, and only the end systems have a clear idea of constraints and reliability requirements.

Pros and Cons of retransmissions at each level is also mentioned in Q1.