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 Retransmission
  - Pro: Avoids resending over end-to-end path, may be more efficient
  - Pro: Hide congestion losses from higher layers
  - Con: Still need end-to-end retransmission
- Transport Layer Retransmission
  - Pro: Do not need to implement retransmission in applications
- Application Layer Retransmission
  - Pro: Some applications can tolerate loss but not delay, so it may be better to let them make the decision.
  - Con: Application complexity is increased.

2. Implicit versus Explicit congestion signals?

- Implicit:
  - An Implicit congestion signal does not require router support.
  - Transport protocols always need to adapt to implicit congestion signals anyway. Explicit congestion signals can be lost, so a lack of an explicit signal does not imply a lack of congestion.
  - Some explicit signals require sending an extra packet to signal congestion. This can cause control traffic congestion collapse.
  - Some explicit signaling schemes set a congestion bit in packets. Selfish receivers can subvert this by not sending the bit back the sender.
- Explicit:
  - An explicit congestion signal can be used for congestion avoidance. Routers can signal congestion when their queues grow long, but before they have to drop packets.
  - An explicit congestion signal requires lower latency to interpret than implicit congestion signals. For example, TCP requires 3 duplicate acknowledgements to determine that a packet has been lost.
  - An explicit congestion signal distinguishes congestion loss from other loss. Wireless networks may drop packets because of signal fading or interference. Route fluttering can send packets into oblivion. This sources of packet loss can be misinterpreted by an implicit congestion signal scheme.

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

   The main function of the Transport Layer is end-to-end reliability. This is achieved primarily through retransmissions, acknowledgments, flow control and congestion control. See the notes for more details.

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

   Through the use of port numbers. Processes transmit and listen on certain port numbers and the computer operating system will send packets with that port number to that process. Recall that port numbers are also used by the NAT protocol (which sits at the network layer). What does NAT use port numbers for?

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

   Duplicate acknowledgements are a sign of packet loss. TCP waits for three duplicate acks (which means four acks for a certain missing packet) to ensure the packet is not in flight and is likely lost. Remember there is nothing magical about the number “three”; it is more art than science.

6. How does TCP set its timeout value?

   TCP provides reliability by requiring the receiver to send an acknowledgement to the sender for every packet it receives. Since packets and acks can get lost, TCP sets a timer at the sender and if the data is not acknowledged before the timer expires, the sender resend the data. The timeout duration is a function of how long the sender expects the acknowledgement to arrive and this is a function of the Round Trip Time (RTT). It is the job of the TCP process at the sender to estimate the RTT, as well as its statistics (e.g., mean and variance).

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

   TCP congestion avoidance is based on AIMD, which stands for Additive Increase Multiplicative Decrease. This means that TCP increases the congestion window for every window of packets acknowledged and cuts the congestion window in half for every packet that is lost (or inferred lost due to receiving triple duplicate acks). The idea of AIMD is to probe the network congestion limits by increasing the congestion window and then back off when loss occurs. This cycle of probe and backoff is the main principle behind TCP congestion avoidance.

   This cycle is shown in the TCP sawtooth behavior slide of the Transport Layer Protocol notes.

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

   Network fairness is a complicated issue and mostly depends on your perspective. One common notion of fairness is equal division of resources, for example, fair allocation of bandwidth amongst competing flows. However, flows may have differing bandwidth requirements, so "fair" could also mean allocation of resources proportional to demand. Another notion of fairness is max-min fairness, which aims to maximize the minimum resource that any flow gets.

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.

   TCP controls the amount of traffic by adjusting the transmit window size W at the sender. The TCP algorithm tries to give approximately the same window size W to all competing flows. In other words, the resources that TCP allocates fairly is the storage/buffer space in the downstream routers.

   The sending rate of a flow is computed as follows. A flow with a window size of W packets has a throughput of (W/RTT) packets/second. The RTT here is the total round trip time (in seconds), consisting of all possible delays, such as those due to processing, propagation, queuing and transmission delays.