Notes on different types of delays

In packet switched networks, there are four types of commonly identified delays – processing, queuing, transmission and propagation delays.

Processing delay is the CPU cycles needed to look at the packet headers and decide what to do with the packet, and do it – basically the time needed to process the packet.

The *propagation delay* is the time a signal takes to traverse the medium. In a metal wire, such as a copper wire, this is the time needed for an electrical signal to be propagated from one host to another. An Ethernet card transmits bits by varying the electric potentials at its pins, and all other Ethernet cards attached to it by a wire will detect this change in electric potential and correctly interpret the bits (Ethernet protocol, IEEE 802.3 standard). In a fiber-optic cable, the time needed to propagate a light signal from end to end is the propagation delay. Processing and propagation delays are often considered negligible.

Transmission delay is related to *transmission rate* of an interface. A transmission rate of R bps means that the interface can push R bits to the interface per second. If a packet has length L bits, L/R is the transmission delay *for that packet*, which is the time it takes for the interface to push the whole packet to the wire.

When a packet arrives at a router interface to be forwarded, it has to wait its turn in a queue. Earlier packets need to be sent out first. This results in *queuing delay,* which varies depending on traffic conditions. Because each packet might suffer a different delay due to queuing, statistical measures such as average delay, variance of delay, probability that delay exceeds a certain threshold, are commonly used. There is an area of study in Statistics call Queuing Theory that deals with queuing issues.

Suppose packets arrive at a router queue at a rate of *a* packets per second, and that each packet is *L* bits long. Further suppose that the outgoing transmission rate at the router interface is *R* bits/sec. So *a*L* bits arrive at the interface per second, and *R* bits leave the interface. The ratio *Incoming Rate/Outgoing Rate = (a * L)/ R* is called the *traffic intensity.* If *(a * L)/ R* >1, incoming rate is more than outgoing rate, and queuing delay will build up. Ideally, *(a * L)/ R* should be less than 1.

Routers generally have small finite queues (why not very large queues?) to hold packets to be forwarded. If an incoming packet finds a queue full, the router drops the packet, resulting in a lost packet. The lost packet will need to be re-transmitted by the sending Application or Transport layer.

The "average queuing delay per packet" is the average of the delay for each bit in the packet. This delay happens because each bit takes a certain amount of time, D, to be put on the wire. So, the first bit has 0 delay, the next bit has to wait D, the third bit has to wait 2*D, etc. and the last bit has to wait 9999*D. So we average over these delays, (0*D+1*D+...+9999*D)/10000 to get the average delay per bit in the packet. Remember that these are rough estimates because of uncertainties in each value which we assume to be constant.

Generally, propagation delays are considered negligible, because propagation of signals on a Ethernet wire or over radio takes place at near the speed of light, 3*10^10 cm/sec. Similarly, in most cases, processing delays are also considered negligible, in view of the current speed of microprocessors. Thus, the most significant delays are generally the transmission and queuing delays.