ECE 333: Introduction to Communication Networks
Fall 2001

Lecture 12: Delay Models II

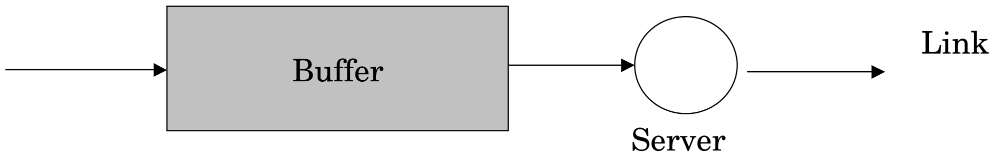
- **Queueing models**

In this lecture we will continue our discussion of delay in networks, specifically queueing delays. Recall queueing delays arise when packets have to wait in memory while other packets are being transmitted. Last time we introduced some basic components of queueing models and discussed Little's Theorem, which says that the average number of a customers in a system is equal to the product of the arrival rate and the average customer delay, *i.e.* $N=\lambda T$.

Little's theorem applies to a very broad class of systems. It tells us how to relate the average delay and the average occupancy, if either is given. Today we will look a more specific model of a queueing system and see how these quantities are related to the arrival rate and service time. This will help us answer questions like the following:

- *Suppose the arrival rate at a link is increased, what effect will this have on the delay?*
- *What if both the arrival rate and the transmission speed are both increased?*

In the following we study a model of a buffer attached to a transmission link with a transmission rate of C bps.



Packets randomly arrive at the buffer (from some application or from another link). A packet leaves the buffer and enters the server when it is to be transmitted. We think of a packet as being in the server until all of it is transmitted, *i.e.* the time in the server is equal to the transmission time, which is equal to L/C seconds for a packet of length L.

We consider the case where one packet is served at a time in FCFS order. We assume that the buffer has no capacity constraints and that no packets are blocked. Recall to formulate a mathematical model of this system we need to specify a model for the arrivals to the buffer and a model for the service times required by each packet (i.e. the packet lengths). We look at a model for the arrivals first.

Poisson Processes:

We will consider a particular model for the arrival of packets to the buffer, called a ***Poisson process***. A Poisson process is a common model for the arrivals in a queueing system. One reason for using this model is that it allows a tractable analysis. Also, Poisson processes are generally considered a good model for an aggregate of traffic from large number of similar and independent users.[1] These models are widely accepted as a model for the arrival of calls in a telephone network. However, in computer networks, Poisson models have been shown to not always be a good model for the arrivals of packets; a great deal of research is done on ***traffic modeling*** for computer networks.

Recall we denote the inter-arrival time between the packet n - 1 and n by A_n. and the number of arrivals between time 0 and t by α(t). Thus, α(t) = largest n such that $\sum_{i=1}^{n} A_i \leq t$).

[1] Recall from basic probability that the central limit theorem implies that a large number of independent random variables can be well approximated by a Normal distribution. With Poisson processes there is also a limit theorem that implies that these processes are a good approximation to the aggregation of a large amount of traffic.

Definition: α(t) is a ***Poisson process*** (with arrival rate λ) if the inter-arrival times, A_n, are independent random variables with an exponential distribution, with mean, $E(A_n) = 1/\lambda$,

$$i.e.\ \ P(A_n \le a) = 1 - e^{-\lambda a}\text{, for all } a \ge 0.$$

The p.d.f. of A_n is $p(a) = \lambda e^{-\lambda a}$, and $\mathrm{var}(A_n) = 1/\lambda^2$.

Several important properties of a Poisson process are discussed next.

Properties of Poisson process:

1. The number of arrivals in any interval of length T has a Poisson distribution with parameter λT,

$$i.e.\quad P(\alpha(t+T) - \alpha(t) = n) = e^{-\lambda T} \cdot \frac{(\lambda T)^n}{n!} \text{ for } n = 0,1,\ldots$$

Note the average number in an interval of length T is λT. Therefore it is correct to refer to λ as the average arrival rate.

2. Memoryless property:

$$P(A_n > r + t \mid A_n > t) = \frac{P(A_n > r + t)}{P(A_n > t)} = \frac{e^{-\lambda(r+t)}}{e^{-\lambda t}} = e^{-\lambda r}$$
$$= P(A_n > r)$$

This means that the probability of a new arrival is independent of that fact that an arrival has not occurred for t seconds.

3. Combining - The combination of two independent Poisson processes with rates λ_1 and λ_2 is a Poisson Process with rate $\lambda_1 + \lambda_{2.}$

4. Arrivals from a Poisson process see the system in a "typical" state, i.e.

$$P(N(t) = n) = P(N(t) = n \mid an\ \ arrival\ \ occured\ \ just\ \ after\ \ time\ \ t)$$

Proof of 4:

Let ***E(t)*** be the event that an arrival occurs in $(t,t+\delta)$.

then

$$P(N(t)=n\mid E(t))=\frac{P(N(t)=n,E(t))}{P(E(t))} \qquad \text{(Bayes' rule)}$$

$$=\frac{P(E(t)\mid N(t)=n)P(N(t)=n)}{P(E(t))}$$

$$=P(N(t)=n)$$

The last inequality follows since $E(t)$ is independent of $N(t)$. □

Service times:

The service time or transmission time of a packet is also modeled as a random variable. We denote the service time of the *n*th packet by X_n. We will consider the case where the sequence of random variables X_1, X_2 … are independent and identically distributed (i.i.d.). We denote by

$$\mu=\frac{1}{E(X_i)}$$

the service rate in packets per second. Note since the service time of the *i*th packet is equal L_i/C, we have $\mu=C/E(L_i)$. We will allow the service time to have an arbitrary distribution.

To summarize we are considering a single server, FCFS, queueing system with infinite capacity, a Poisson arrival process, and general i.i.d. service times.

Notation:

A queueing system is often described using the notation such as ***M/M/1.*** Here the first letter indicates the inter-arrival distribution, the second letter indicates the service time distribution, and the last number indicates the number of servers.

The letters commonly used for the arrivals and service times include

M stands for a memoryless (exponential) distribution.
G stands for General distribution
D stands for deterministic

Deterministic means the inter-arrival times or service times are all equal to the same value. Memoryless inter-arrivals correspond to a Poisson process.

Thus the model described on the previous page is a M/G/1 system.

Other possibilities include: M/M/n, G/G/1, M/D/1, ….

Pollaczek-Khinchin formula:

For a M/G/1 queue, the expected waiting time W (time in the queue), is given by the ***Pollaczek-Khinchin formula*** (P-K formula). Specifically, we have

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

where $\overline{X^2} = E(X^2)$ and $\rho = \lambda E(X_i) = \lambda / \mu$ is the utilization (as defined in the last lecture).

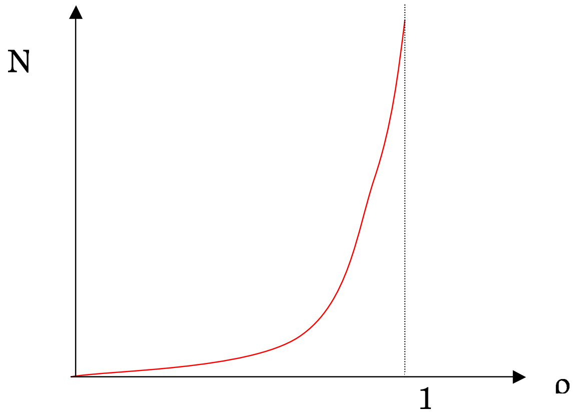
Thus the expected total time in the system, T, is given by

$$T = \overline{X} + \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

Using Little's theorem we have that the expected number in the queue, N_Q and the expected number in the system, N, are given by:

$$N_Q = \frac{\lambda^2 \cdot \overline{X^2}}{2(1-\rho)} \qquad\qquad N = \rho + \frac{\lambda^2 \cdot \overline{X^2}}{2(1-\rho)}$$

Thus as the utilization (or equivalently the arrival rate for a fixed transmission rate) increases, the average number in the system will increase as shown below:



Similar behavior can be seen in the other parameters.

Special cases:

Both the M/M/1 and M/D/1 are special cases of the M/G/1 queue, thus the PK formula can be applied to these.

(1) **M/M/1 system** - Here service times have an exponential distribution with parameter μ, so $\overline{X} = 1/\mu$ and $\overline{X^2} = 2/\mu^2$. Thus from the PK formula,

$$W = \frac{\rho}{\mu(1-\rho)}$$

(2) **M/D/1 system** - service times are deterministic, each packet has service time $\overline{X} = 1/\mu \Rightarrow \overline{X^2} = 1/\mu^2$. In this case, the PK formula reduces to

$$W = \frac{\rho}{2\mu(1-\rho)}$$

Note the expected waiting time is twice as long in (1) as in (2) for same ρ. This can be attributed to the randomness in the service times in the first case.

Also notice that for a given arrival rate and average service time, (2) is lower bound to average delay for any distribution. (Why?)

Example

Consider a buffer in a network that is modeled as an M/M/1 queue. Suppose the arrival rate is increased from λ packets per second to $K\lambda$. If the transmission rate is also increased from C bps to KC bps, how does the average delay change? What about the average number in the system?

Answer: Note since both the arrival rate and the service rate are increased by the same factor, the utilization, ρ will remain the same. Using the above formulas the average number in the system for a M/M/1 queue is

$$N = \frac{\rho}{1-\rho}$$

Therefore the average number in the system will be unchanged by the increase in rates. However the average delay per packet (from Little's theorem) is equal to N/λ. Thus the average delay will decrease. To summarize, a transmission line that is K times as fast will handle K times as many packets with K times smaller delay.

Justification for PK formula:

Let W_i be the waiting time for the i-th packet.

This can be written as

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j$$

Where R_i is the ***residual service time*** seen by the i-th packet - i.e. if the j-th packet is being transmitted when the i-th packet arrives, then R_i is the remaining time until the j-th packet is completely transmitted.

Taking expectations

$$E(W_i) = E(R_i) + \overline{X} \cdot E(N_i)$$

(Here we used the independence of the various random variables)
Assuming that in the limit as $i \to \infty$, the various expected values converge to their time average values[2] we have:

[2] This assumption is valid for most cases of interest.

$$W = R + \frac{1}{\mu} N_Q$$

where R= mean residual time = $\lim_{i \to \infty} E(R_i)$

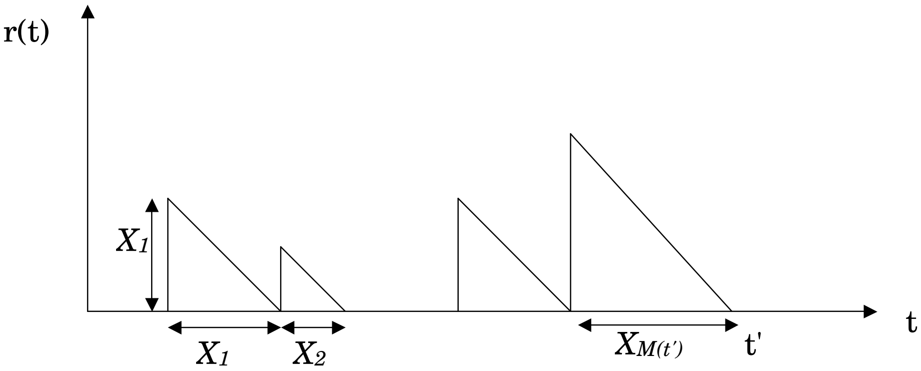
By Little's Theorem $N_Q = \lambda W$

$$\text{thus, } W = R + \rho W \quad \Rightarrow \quad \boxed{W = \frac{R}{1-\rho}. \qquad (*)}$$

Need to calculate R?

Since arrivals are Poisson, customers find systems in typical state. Thus R is the same as the mean residual time of the system in a typical state.

Below is plot of residual time $r(t)$



Consider a time t' for which $r(t)$=0. Then the time average of $r(t)$ up till time t' is

$$\frac{1}{t'} \int_0^{t'} r(t) dt = \frac{1}{t'} \sum_{i=1}^{M(t')} \frac{1}{2} X_i^{\,2}$$

where $M(t')$ = number of service completions by time t'.

can re-write this as

$$\frac{1}{t'}\int_0^{t'} r(t)dt = \frac{1}{2}\cdot\frac{M(t')}{t'}\cdot\left(\frac{1}{M(t')}\sum_{i=1}^{M(t')} X_i^{\;2}\right)$$

Taking limits (again assuming limits exist and that time averages equal to ensemble average)

$$\lim_{t'\to\infty}\frac{1}{t'}\int_0^{t'} r(t)dt = \frac{1}{2}\cdot\lim_{t'\to\infty}\frac{M(t')}{t'}\cdot\lim_{t'\to\infty}\left(\frac{1}{M(t')}\sum_{i=1}^{M(t')} X_i^{\;2}\right)$$

$$\Rightarrow R = \frac{1}{2}\lambda\overline{X^2}$$

Substituting this into (∗), the P-K formula follows.