Lecture 11
CS/EE 5516 - Spring 1994
Performance Modeling

References:
[BG] Chapter 3: 3.1, 3.2.1, 3.2.3, 3.3.1, 3.3.2, ...
C. H. Sauer and K. M. Chandy, *Computer Systems Performance Modeling*, 2.4.2, 3.2.1-3.2.2

Topics:
1. Performance measures of interest
2. Little's Law and its application
3. Poisson process and the exponential distribution
4. Memoryless property
5. Markovian queue
6. Networks of Markovian queues

Topic 1: Performance measures of interest ([BG] 3.1)

Delivery of a packet from source to destination (at the network layer) involves four delays:

1. Processing delay (at a node)
2. Queueing delay
3. Transmission delay
4. Propagation delay

Consider a *single* transmission line with:

C (bits/sec) transmission capacity: number of bits that can be transmitted per second over a bit pipe

L (bits) message length

Q. What is the delay to deliver one packet with:

a. statistical mulitiplexing -- either
   - merge inbound streams into one queue serviced in FCFS order, or
   - put each inbound stream in its own FCFS queue, but serve queues in sequence one packet at a time
b. time division multiplexing (TDM) with *m* time slots
c. frequency division multiplexing (FDM) with *m* channels, each with 1/m-th of the bandwidth of the entire channel

A.
a. SM:
L/C [bits/(bits/sec)=sec], because entire pipe is used for one packet at a time, and it takes L/C seconds to transmit a message of length L bits

b. Lm/C, or m times larger than SM. View the channel as having capacity C/m.

- If slot duration is short compared to packet length, then it takes many slots to transmit packet
- Otherwise, the packet waits (m-1) transmission times between packets of the same stream, but requires time L/C to send packet when its turn arrives

c. see TDM

Fact #1: SM has a smaller average delay per packet than TDM or FDM. Consider case of idle network and just one packet to send. SM starts right way (vs. TDM wait for slot), and then sends at rate C (vs. rate C/m for FDM).

Fact #2: With regular arrival from all streams (in each stream, all packets arrive sufficiently apart so that no packet has to wait while the preceeding packet is transmitted), TDM/FDM have smaller *variance* in delay.

Topic 2. Little's Law and its application ([BG] 3.2.1, [SC] 2.4.2)

We'll think in terms of *customers,*

- messages in transmit in a bit pipe
- connections in a virtual circuit network
- calls in a telephone network

service time,

- transmission time (L/C)

and a *system*:

- one link in a netwwork
- one switch in a network
- one switch + link in a network
- one path in a network
- the entire network

General problem:

Given:

- customer *arrival rate*

(the "typical" number of customers entering the system per unit time)
- customer *service rate*
(the "typical" number of customers the system serves per unit time when it is *constantly busy*)

Find:
- the average *number of customers in the system*
(the typical number of customers either waiting in queue or undergoing service)
- the average *delay per customer*
(the "typical" time a customer spends waiting in queue plus service time)

Let's see how far we can go without assuming any statistical information about the arrival or service processes.

Review of random variables:

Random variable X maps experiment outcomes to integers or reals.

If X is continuous:
$F(x) = Prob[X \leq x]$ (the prob. distribution or cumulative distr. function)
$f(x) = F'(x)$
$0 \leq F(x) \leq 1$

If X is discrete:
p(x) is the probability mass function
$F(x) = \Sigma\ ip(i)$, where the sum limits are $-\infty$ to x.

Little's Law:
Consider two random variables:

$N(t)$ the number of customers in the system at time t

T_i the total time spent in the system by the i-th customer to arrive

$\alpha(t)$ the number of customers who arrived in interval [0,t]

Note that N(t) is a discrete RV, and T_i is a continuous RV.

Let:

N_t be the mean number of customers in the system, in time interval [0,t]

T_t be the mean delay that a customer experiences, in time interval [0,t]

We will only consider systems that are *ergodic*. This term has a precise meaning (take ISE 5414 to learn it), but intuitively it means that the RV N(t) reaches a *limiting distribution* as $t \to \infty$.

For an ergodic system, N_t and T_t converge on a limiting value as $t \to \infty$, denoted N and T, respectively:

$N = \lim(t \to \infty)\ N_t$

$T = \lim(t \to \infty)\ T_t$

If our system is a campus network observed between midnight and 9am, it is *not* ergodic, because the number of customers grows rapidly after 8a.m. However if our system is a large (100's of users) campus network observed for a month period during the semester (and no new hosts are added or removed), it *can* be regarded as ergodic.

Recall how to find the *expected value* of a RV: $\Sigma(np(n))$ or $tf(t)dt$, where the limits are 0 and ∞, and p(n) is the probability of the RV having value n, and f(t) is the probability density function.

Thus:

$N = E[n] = \Sigma(ip(i))$

$T = E[t] = \tau f(\tau)d\tau$

where the limits are 0 and ∞, p(i) is the prob. of i customers in the system, and f is the density function of the delay per customer.

Little's law relates Nand T. Let's guess the relation:

N = ? T

(jobs) = ? (sec)

Let ? be a scalar constant, namely λ.. Then λ must have units of (jobs/sec). Thus λ can be interpreted as:

- arrival rate
- throughput
- departure rate

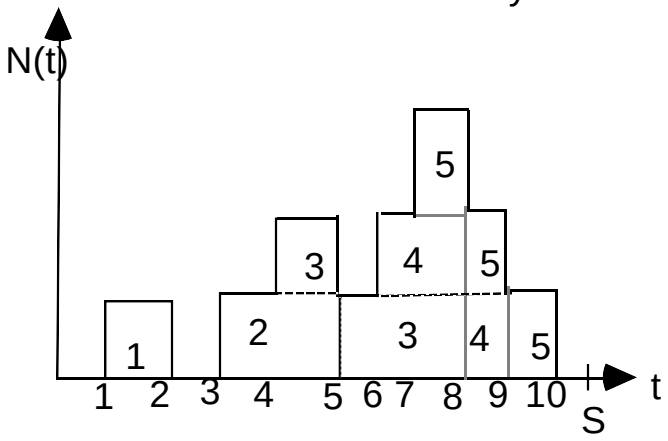
If the limits N and T exist, then the three quantities above *must* be equal! (After all, the departure rate of a system must equal its arrival rate, or queue length grows without bound. Thus arrival rate equals throughput.)

Little's Law:	$N = \lambda T$

Thus a system with a big waiting time must have a big storage capacity for lots of waiting customers (consider a busy airport with delayed flights). In contrast, a system with small per-customer delay needs a small waiting area (consider a gas station).

Informal argument to demonstrate why Little's law holds:

Consider a queueing system with FCFS discipline. Consider a period starting at time 0 and stopping at time S, where the system is *empty* at time S. Let N(t) denote the number of customers in the system at time t.



Thus:

$\lambda = \alpha(\Sigma)/S = 5/10$

Let A(S) be the area under N(t) up to time S, i.e.,

$A(S) = \int N(t)dt = 1+2+4+3+3 = 13$

where the limits are 0 and S.

We can easily show that

$N_S = A(S)/S = 13/10 = 1.3$

and

$T_S = A(S)/\alpha(\Sigma) = 13/5 = 2.6$

Thus

$\lambda = \alpha(\Sigma)/S = 5/10 = 0.5$

$= (A(S)/T_S) / (A(S)/N_S) = N_S/T_S = 1.3/2.6 = 0.5$

Example application of Little's law:

Example 1 (based on [BG] Ex. 3.1):

Q. Suppose messages arrive at a switch at the average rate of 3000/second, and the average queue length is 1000 packets. What is the average delay per packet (queueing plus switch processing time)?

A. λ = 3000 packets/second
N = 1000 packets
Thus $T=N/\lambda$ = 1/3 seconds.

Utilization

Q. If we consider just a *server*, there can be at most one customer in the system. What is the intepretation of N in this case?

A. N is the *utilization* (U), or the fraction of time that the server is busy:
utilization = throughput * delay
Let μ be 1/T (the *service time*):
$\rho = \lambda / \mu$

Q. What is the domain of ρ?

A. $0 \leq \rho \leq 1$
because the arrival rate cannot exceed the service rate in an ergodic system, or an infinite length queue will develop.

Example applications of Little's law:

Example 2 (based on [BG] Ex. 3.1):

Q. Suppose the switch in Example 1 sends all outgoing traffic on a single link. Suppose the transmission time per packet is 10^{-4} seconds. What is the proportion of time that the link is busy carrying a packet?

A. $\lambda = 3 \times 10^3$ packets/second
$T = 10^{-4}$ second
Thus $\rho = \lambda T$ = 30% (i.e., there is 3/10 of a customer on the link on the average)

Example 3 (based on [BG] Ex. 3.2):

Q. Suppose there are n hosts in the Internet, and the arrival rate of packets to the hosts is $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Suppose that the average *total* number of packets in the Internet is N. What is the average delay per packet?

A. $T = N/\Sigma\lambda_i$, where the sum is from 1 to n.

Example 4 (based on [BG] Ex. 3.4):

Q. Consider a go-back-n protocol used at the DLC layer with n=256. Measurement reveals that the average time from when a packet is accepted by the sending DLC until the packet is released by the receiving DLC is 10^{-2} seconds. Find an upper bound on the rate at which the sending DLC accepts packets.

A. The number of customers in the DLC cannot exceed the window size. Thus $n \geq \lambda T$, where $T=10^{-2}$. Thus $\lambda \leq 256/0.01$, or $\lambda \leq$ 25,600 packets/second.

Topic 3. Poisson distribution

Reference: Law and Kelton, *Simulation Modeling and Analysis*, 2nd. ed, pp. 405-406.

Let N(t) denote the number of events that occur at or before time t.

The stochastic process { N(t) | $t \leq 0$ } is a *Poisson process* iff:

1. Customers arrive one at a time.

2. The number of event occurrences in disjoint time intervals is independent:

   N(t+s) - N(t) [the number of events in time interval (t,t+s]] is independent of { N(u) | $0 \leq u \leq t$ }.

3. The distribution of N(t) is stationary:

The distribution of N(t+s) - N(t) is independent of t for all $t,s \geq 0$.

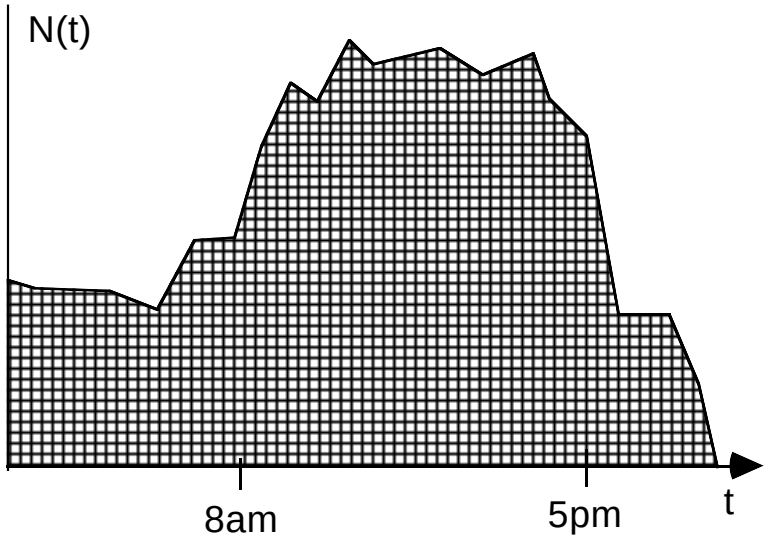
Are these reasonable assumptions for networks, where "event" is "arrival"?

1. *Yes:* a network or switch is rarely viewed as having "simultaneous arrivals". (In real life, a mob may storm a building. So the arrival process is a *bulk arrival*.)

2. *No* for the destination node running a go-back n protocol, because the *number* and *rate* of past messages sent affects when the window closes and hence a gap in arrivals occurs.

   Yes for a switch in a datagram network.

3. *No* for the total traffic arriving to a campus network over one day:



   The number of arrivals is perhaps stationary between 8 am and 5 pm, but not over the entire day.

Let's think for a moment about <u>what distribution in the world could possibly satisfy 2 & 3. Is there just one? Are there many such distributions?</u>

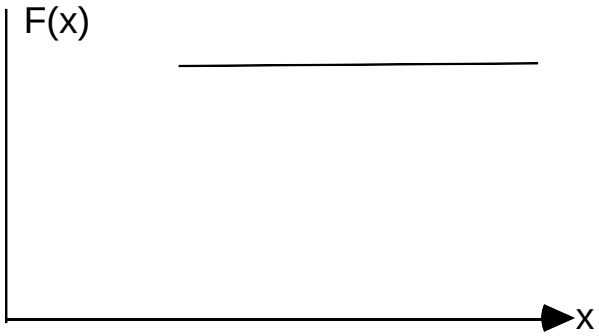
Let's just consider two common distributions: uniform and exponential.

I assume you are familiar with the uniform distribution, so I will not draw it.

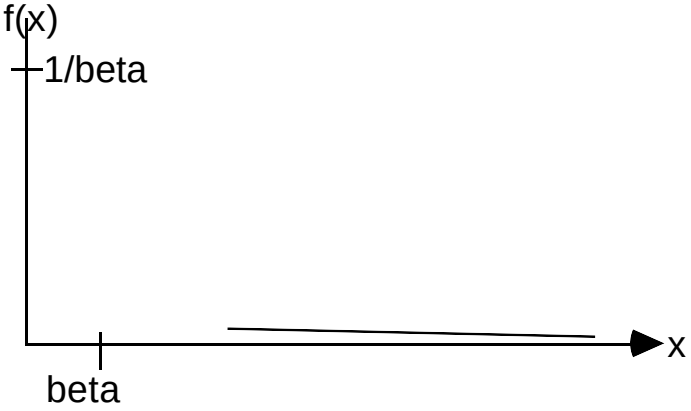
However, the exponential distribution with rate β may be unfamiliar:

Mean=β
Variance=β^2

$F(x) = 1-e^{-x/\beta}$ if $x \geq 0$ (0 otherwise)



$f(x) = (1/\beta)e^{-x/\beta}$ if $x \geq 0$ (0 otherwise)



Q. Consider arrivals of packets to a network. Suppose the distribution of arrival time has a mean of 10 seconds. Suppose you monitor the network and see that I last sent a packet 10 second ago. When do you expect the next packet?

(I am asking you to think about the infamous *interarrival time* of packets to a network)

A. For a uniform(0,20) distribution, you expect it in another 5 seconds. (Proof: see HW9, problem 2.)

For an exponential distribution, the fact you've waited 10 minutes is irrelevant -- you expect next packet in 10 minutes!

Proof that the expected time is 10, given we waited 10 time units already:

Let b be the arrival rate. Then we want to calculate the probability distribution of $Pr[T \leq 10+t \mid T>10]$ (for $0 \leq t$):

$$Pr[T \leq 10+t \mid T>10] = Pr[10<T \leq 10+t] \,/\, 1-Pf[T \leq 10]$$

$$= (F(10+t) - F(10)) / (1-Pr[T \leq 10])$$

$$= \frac{(1-e^{-((10+t)/10)}) - (1-e^{-(10/10)})}{1 - (1-e^{-(10/10)})}$$

$$= \frac{e^{-10/10} - e^{-((10+t)/10)}}{e^{-10/10}}$$

Dividing the numberator and denominator by 1/e:

$$= 1 - e^{-t/10}$$

Thus *the conditional distribution is the same as the unconditional distribution!*

Do you see that *only* the exponential distribution first property 2 stated above in the definition of a Poisson process?

Theorem 1: If { N(t) | t ≤ 0 } is a Poisson process with rate λ, then its corresponding interarrival times A_1, A_2, ... are IID[1] exponential random variables with mean $1/\lambda$.

Theorem 2: If { N(t) | t ≤ 0 } is a Poisson process then the number of arrivals in any time interval of length s is a Poisson random variable with parameter λs (where λ is a positive real number).

Theorem 3: Merging *n* Poisson processes with rates λ_1, λ_2, ..., λ_n produces a Poisson stream with rate $\lambda_1+\lambda_2+...+\lambda_n$.

Let's prove Theorem 3 for n=2 [See Chandy/Sauer pp. 34-35]. First take on faith the fact that the Poisson process with rate λ implies that, for a sufficiently small time interval:

Pr[1 arrival in interval Δt]	$= \lambda\Delta t$
Pr[0 arrivals in interval Δt]	$= 1- \lambda\Delta t$
Pr[>1 arrival in interval Δt]	$= O(\Delta t^2) \approx 0$

If two Poisson streams with rates λ_1 and λ_2 are merged, then:

[1] IID = independent and identically distributed

Pr[1 arrival in interval Δt] $= (\lambda_1 \Delta t)(1- \lambda_2 \Delta t)+ (1- \lambda_1 \Delta t)(\lambda_2 \Delta t)$
$= (\lambda_1+\lambda_2)\Delta t$
Pr[0 arrivals in interval Δt] $= (1- \lambda_1 \Delta t)(1- \lambda_2 \Delta t)$
Pr[2 arrival in interval Δt] $= (\lambda_1 \Delta t)(\lambda_2 \Delta t)$
$= O(\Delta t^2) \approx 0$

SImilarly, *splitting* a Poisson process into two streams by flipping a coin results in Poisson processes.

Theorem 3 is *very* useful in networks -- consider the aggregate traffic in a network switch from multiple inbound links.

Theorem 4: For sufficiently large *n*, merging *n* IID arrival processes (*not Poisson!)*, each with rate λ/n, produces an aggregate stream that is Poisson with rate λ.

Topic 4. Memoryless property

The exponential distribution is *memoryless*, which will simplify analysis that uses it, because the *accumulated service time* can be ignored.

Another motivation for the exponential distribution:

If we model a system we can now use *one dimension*:
(Number of customers present)
rather than
(number of customers present, remaining service time)

This will lead naturally to the Markov process.

The question of whether the memoryless property is reasonable for networks is equivalent to whether the Poisson process is reasonable.

Topic 5. Markovian queue

A *stochastic process* is a set of random variables indexed by time:
{ N(t) }

We'll consider continuous time *chains*, which are stochastic processes composed of discrete random variables.

In particular, out stochastic process will be { N(t) }, and we will assume Poisson arrivals and exponentially distributed service times.

Recall that the probability that one event occurs in a small time Δt in a Poisson process with rate λ is $\lambda\Delta t$. The probability that 0 events occur is $1-\lambda\Delta t$. The probability that more than one event occurs is order Δt^2 0.

Let's draw a diagram to see how N(t) evolves and the transition probabilities. Here we are considering *discrete time*. Later we'll change to continuous time.

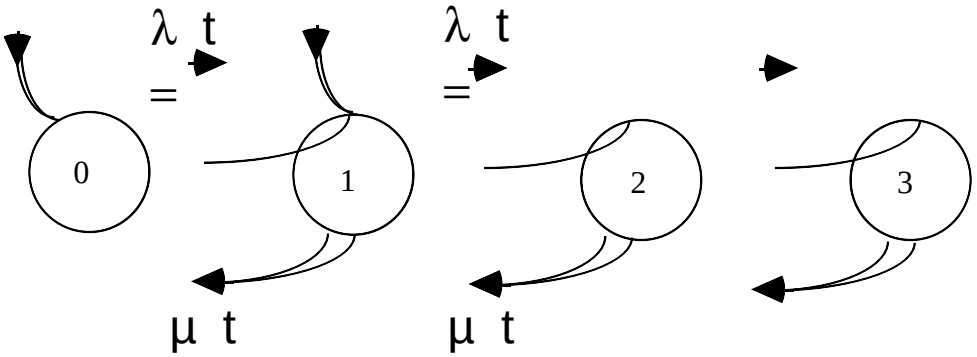
(Insert [BG] Fig. 3.6)

Write and solve balance equations to derive M/M/1 solution.

Topic 6. Networks of Markovian queues

Solution of M/M/1 Queue for time in system & # customers

Markov Chain



N(t) = # customers in system

$P_n = \lim_{t \to} P_n\{N(t) = n\}$

n 0:

$P_n \lambda = P_{n+1} \mu t$ $P_1 = \rho P_0$ $P_2 = \rho^2 P_0$

=> $P_{n+1} = \rho P_n$

=> $P_{n+1} = \rho^{n+1} P_0$

$(n = 0, \ ) \ P_n = 1$

$P_0 = 1 - \rho$

$$\boxed{P_n = \rho^n(1 - \rho)}$$

$N = \quad (n=0, \ ) \ n\rho^n(1 - \rho)$

$$\boxed{N = \rho/(1 - \rho) = \lambda/(\mu - \lambda) = N}$$

$\rho = 1/2$: $\mu = 1$, $\lambda = .5$: $N = (1/2)/(1/2) = 1$

$T = N/\lambda$

$\boxed{\Rightarrow T = 1/(\mu - \lambda)}$

$\Rightarrow 1/(1/2) = 2$

$N_Q = \lambda/(\mu - \lambda) - \lambda/\mu$

$T_Q = 1/(\mu - \lambda) - 1/\mu$