Queueing Theory Basics Goal: make an analytical model of customers needing service, and use that model to predict queue lengths and waiting times.

> *May not be accurate for real situations, but we get insights needed for studying networks.*

Terminology

Customers — independent entities that arrive at random times to a Server and wait for some kind of service, then leave.

Server — can only service one customer at a time; length of time to provide service depends on type of service; customers are served in FIFO order.

Time — real, continuous, time.

Queue — customers that have arrived at server but are waiting for their service to start are *in the queue*.

Queue Length at time t — number of customers in the queue at time t.

Waiting Time — for a given customer, how long that customer has to wait between arriving at the server and when the server actually starts the service (total time is waiting time plus service time).

Illustration of customers, queue and server.

a_1 is being served

a_2–a_4 are in queue

a_5–a_9 haven't yet arrived at server

Notation

T_i is arrival time (at server) of customer a_i

$$0 < T_1 < T_2 < T_3 < \cdots$$

Δ_i is *interarrival time* defined by

$$\Delta_i = T_i - T_{i-1}$$

S_i is service time for customer a_i

M/M/1 Queues

In order to make analysis possible, some assumptions about interarrival and service times.

1. The number of arrivals at server in any time interval of length τ is Poisson distributed with parameter $\lambda\tau$.

$$P(A(t+\tau) - A(t) = n) = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}$$

2. Each Δ_i is independently random with probability

$$P(\Delta_i = t) = \lambda e^{-\lambda t}$$

   which of course means that

$$P(\Delta_i > t) = e^{-\lambda t}$$

   What is average (expected) interarrival time?

3. Each S_i is independently random with probability

$$P(S_i = t) = \mu e^{-\mu t}$$

Theorem [Statistical Equilibrium]

Let x_t be the number of customers either in queue or being served at time t. Theorem states:

$$P(x_t = n) = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

(Proved by Markov chain analysis.)

Questions

1. What if $\lambda \geq \mu$?
2. What is expected queue length ?

Arrival Rate versus Interarrival Rate

Let A_t be the number of arrivals at the server from time zero up to time t. An intuitive definition of the arrival rate λ is

$$\lambda = \lim_{t\to\infty} \frac{A_t}{t} \qquad \text{customers/unit of time}$$

Let the first customer arrive at time τ_1, the second customer arrive at time $\tau_1+\tau_2$, the third customer arrive at time $\tau_1+\tau_2+\tau_3$, and so on. Then $\tau_1, \tau_2, \tau_3, \ldots$ are the lengths of time intervals *between* arrivals. The arrival rate up to the time when the k-th customer arrives is

$$\frac{k}{\tau_1+\tau_2+\cdots+\tau_k}$$

The average interarrival time (time between arrivals) up to the time when the k-th customer arrives is

$$\frac{\tau_1+\tau_2+\cdots+\tau_k}{k}$$

Therefore

$$\lambda = \lim_{k\to\infty} \frac{k}{\sum_{i=1}^{k}\tau_k} = \lim_{k\to\infty} \frac{1}{(\sum_{i=1}^{k}\tau_k)/k} = \frac{1}{\lambda^{-1}}$$

shows that the mean interarrival rate is λ^{-1}.

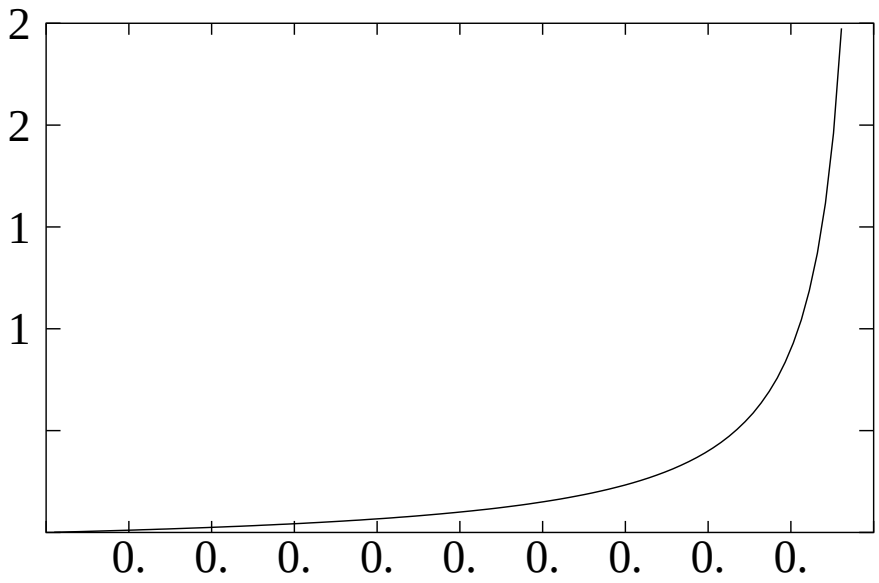
If arrival distribution is Poisson, then why is interarrival distribution exponential?

Actually it's the other way around: if we start by assuming that interarrival rates are exponential, then Poisson distribution of arrival rates can be proved (using repeated convolution).

Mean queue length N is:

$$N = \frac{\lambda}{\mu - \lambda}$$

Graph of λ/μ (x-axis) versus N (y-axis):



Recall that mean interarrival rate is λ^{-1} and mean service rate is μ^{-1}.

$$\frac{\lambda}{\mu} = \frac{\mu^{-1}}{\lambda^{-1}}$$

Therefore, when the average service rate gets larger and the (inter) arrival rate stays the same, the queue length increases!

Little's Theorem

$$N = \lambda T$$

where N is the mean queue length, T is the mean total delay (including both queuing time and service time), and λ is the (Poisson parameter) arrival rate.

Intuition

If average customer spends time T in system, then about λT customers are waiting behind, because arrival rate is λ.

Question

Customers arrive on average on every two minutes at a fast food restaurant. The mean time spent in the restaurant per customer is 20 minutes (waiting in line, paying, eating). On average, how many customers are in the restaurant?

Other Types of Queueing Models

$M/M/m$ — exponential arrival rate and service times, with m servers (like grocery store with many checkout lanes).

$M/M/m/m$ — exponential arrival rate and service times, with m servers, but nobody waits in queue (if all m servers are busy when a customer arrives, that customer gives up and leaves).

$M/M/\infty$ — exponential arrival rate and service times, with unlimited number of servers (customers never wait in queue).

$M/D/1$ — service times are deterministic (e.g. a constant, fixed service time regardless of customer).

$M/G/1$ — exponential arrival rate, but service rate has a "general" (arbitrary) probability distribution, and a single server.

$M/G/m$ — same as above, but with m servers.

For each of the above models, the questions are the same: what is mean waiting time, what is mean total time in system, what is mean queue length? How are these factors related?

The P-K Formula for $M/G/1$

The "G" in $M/G/1$ refers to a general distribution of service times. Let X_i be the service time for the i-th arriving customer. Suppose these times $X_1, X_2, \ldots$ are independent, identitically distributed, and independent of the interarrival times.

$$E(X) = \frac{1}{\mu} = \text{mean service time}$$

Also define the second moment of service time, which is

$$E(X^2) = \text{mean of squared service times}$$

For convenience,

$$\rho = \frac{\lambda}{\mu} = \lambda E(X)$$

The *Pollaczek-Khinchin* formula is:

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

where W is the expected customer time waiting in queue.

What is P-K formula if $G = M$, that is, for $M/M/1$?

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

What is P-K formula for $M/D/1$?

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

Networks of Queues We can also model a closed system of customers that travel from one server to another.

> A typical example would be a set of “jobs” (the customers) in a computer system, with a disk drive (one server), a CPU (a second server), and a network interface (a third server). Using queueing models, one can predict where bottlenecks occur, if enough is known about distributions of service times.