Queueing Review
(mostly from BCNN)

Christos Alexopoulos and Dave Goldsman

Georgia Institute of Technology, Atlanta, GA, USA

10/25/17

Outline

Introduction

Goal: Review basics of Queueing Theory for very simple systems, mostly in steady-state.

Interested in the usual system performance measures that we've already discussed in earlier modules:

- server utilization (% of time server is busy)
- length of waiting lines
- delays of customers

For *simple* queueing systems, these measures can be computed *analytically*. But simulation is the method of choice for more-complicated models.

Some Definitions

- The *calling population* is the set of potential customers that will arrive to the system. It is usually regarded as *infinite*. Sometimes it's *finite*, in which case the arrival rates of the customers may depend on the current size.
- The *system capacity* is the limit on the number of customers that the system can contain. The system includes all servers and their associated queues. This limit can be infinite or finite.
- The *arrival process* can consist of *scheduled* arrivals (like a doctor's office) or *random* arrivals like a Poisson process or a nonhomogeneous Poisson process.
  - Recall that the times between arrivals of a Poisson process are i.i.d. exponential.
  - Arrivals can also occur in *batches* of either constant or random size.

- *Queue behavior* describes possible customer actions while waiting in line for service.
  - *Balking* — the customer leaves when he sees that the queue is too long.
  - *Reneging* — the customer leaves when he feels that the line is progressing too slowly.
  - *Jockeying* — the customer moves from line to line when he thinks he can do better.
- *Queue discipline* describes how customers are ordered in the queue. Who gets served first?
  - *FIFO* — first-in-first-out.
  - *LIFO* — last-in-first-out.
  - *SPT* — shortest processing time.
  - Random order of service, service by priority class, etc., etc.

- *Service times* are usually modeled as i.i.d. something with mean $1/\mu$ and variance σ^2, but not necessarily exponential. A service station can have multiple parallel servers of varying skills and schedules.
- For most of this module, we'll only look at individual queue/service stations, but of course these can be networked together in arbitrary generality.
- Students are encouraged to read up on open and closed *Jackson networks*, which are simplified network structures for which there are analytical results.
- For the rest of this module, we assume that all queues are FIFO.

Outline

Queueing Notation

$A/B/c/N/K$

- A is the interarrival-time distribution.
- B is the service-time distribution.

  Common symbols for A and B
  - M = exponential (memoryless or Markov)
  - D = deterministic (constant)
  - E_k = Erlang of order k
  - G = general distribution with mean $1/\mu$ and variance σ^2
- c = the number of parallel servers.
- N = system capacity (queue + server)
- K = size of calling population

Example: $M/M/1 = M/M/1/\infty/\infty$

$P_n(t) =$ probability of n customers in system at time t.

$P_n = \lim_{t\to\infty} P_n(t) =$ steady-state probability of n customers.

$\lambda =$ arrival rate

$\mu =$ service rate, and $1/\mu =$ mean arrival time

$\rho =$ server utilization $=$ traffic intensity $= \%$ time server is busy.

$I_n =$ interarrival time between customers $n-1$ and n

$A_n =$ arrival time of customer n.

$S_n =$ service time of customer n.

W_n = cycle time = time in system (Q + server) of customer n.

W_n^Q = waiting time in queue of customer n.

$L(t)$ = number of customers in system at time t.

$L_Q(t)$ = number of customers in queue at time t.

Steady-State Stuff:

$w = \lim_{n\to\infty} \mathrm{E}[W_n], \quad w_Q = \lim_{n\to\infty} \mathrm{E}[W_n^Q]$

$L = \lim_{t\to\infty} \mathrm{E}[L(t)], \quad L_Q = \lim_{t\to\infty} \mathrm{E}[L_Q(t)]$

$L(\infty)$ = steady-state number of customers in the system (this is a random variable, not an expected value)

Outline

Transient System Example

Most of what we'll talk about in the remainder of this module concerns steady-state analysis, but for some simple models, it is possible to undertake exact transient analysis (in addition to the hand-simulation analysis that we worked on in earlier modules).

Example: $M/M/1/1$ system in which an arriving customer will leave if he sees someone already in service. It can be shown (after solving certain differential equations) that

$$P_0(t) = \frac{\mu}{\lambda+\mu} + \left[P_0(0) - \frac{\mu}{\lambda+\mu}\right] e^{-(\lambda+\mu)t},$$

and $P_1(t) = 1 - P_0(t)$, where $P_0(0)$ is an initial condition and where the last term disappears as $t \to \infty$. Thus, in steady-state, we have

$$P_0 = \frac{\mu}{\lambda+\mu} \quad \text{and} \quad P_1 = \frac{\lambda}{\lambda+\mu}.$$

Outline

Little's Law (also known as the *Conservation Equation*) is

$$L \;=\; \lambda w \quad \text{or} \quad L_Q \;=\; \lambda w_Q,$$

so that the long-run average number of customers equals the product of the arrival rate and the long-run average time in system. If you let $a = \lambda$, you have the amazing form $L = aw$.

"Proof": Suppose N customers go through the system during time interval $[0, T]$. Then the obvious estimator for L (from our hand simulation module) is

$$\hat{L} \;=\; \frac{1}{T}\int_0^T L(t)\,dt \;=\; \frac{N}{T}\,\frac{1}{N}\int_0^T L(t)\,dt \;=\; \frac{N}{T}\,\frac{1}{N}\sum_{i=1}^{N} W_i,$$

since both $\int_0^T L(t)\,dt$ and $\sum_{i=1}^{N} W_i$ = total time in system.

As $T \to \infty$, we have $\hat{L} \to L$, $N/T \to \lambda$, and $\sum_{i=1}^{N} W_i/N \to w$. $\square$

$M/G/1$: Here we have i.i.d. service times from some general distribution with mean $1/\mu$ and variance σ^2.

The traffic intensity $\rho = \lambda/\mu$, and we will assume that $\rho < 1$, so that the queue is stable. Then the *Pollaczek-Khintchine Equation* gives

$$\begin{aligned} \rho &= \lambda/\mu \\ L &= \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} \\ L_Q &= L - \rho \quad \text{(obvious)} \\ w &= L/\lambda \quad \text{(Little)} \\ w_Q &= L_Q/\lambda \;=\; w - \frac{1}{\mu} \\ P_0 &= 1 - \rho \end{aligned}$$

Remark: The $M/G/1$ is actually easier to analyze than the $G/M/1$.

Example: Arrivals occur as a Poisson process with $\lambda = 1.5$/hour.

The service rate is $\mu = 2$/hr, and the variance of the i.i.d. service times is $\sigma^2 = (1/9)$ hr^2.

Then $\rho = 1.5/2 = 0.75$, so the queue is stable.

The steady-state probability that the server is idle is $P_0 = 1 - \rho = 0.25$.

The steady-state average number of customers in the system is

$$L = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} = \frac{3}{4} + \frac{\frac{9}{16}\left(1+\frac{4}{9}\right)}{2/4} = 2.375,$$

and (try this at home) the avg wait in queue is $w_Q = 1.08$ hours. □

$M/M/1$: This is a special case of the $M/G/1$ with Exp(μ) services so that the mean is still $1/\mu$, while $\sigma^2 = 1/\mu^2$.

Plugging these into the P-K Equation and simplifying gives

$$\begin{aligned} L &= \frac{\rho}{1-\rho} \\ L_Q &= L - \rho - \frac{\rho^2}{1-\rho} \\ w &= L/\lambda = \frac{1}{\mu(1-\rho)} \\ w_Q &= L_Q/\lambda = \frac{\rho}{\mu(1-\rho)} \\ P_n &= (1-\rho)\rho^n, \end{aligned}$$

where the result for P_n is a bonus because the $M/M/1$ is so simple.

Example: Arrivals occur at a one-man barbershop as a Poisson process with $\lambda = 2$/hour.

The services are i.i.d. Exp($\mu = 3$/hr).

Then $\rho = 2/3 = 0.75$.

The steady-state probability of n customers in the system is $P_n = \frac{1}{3}(\frac{2}{3})^n$, $n = 0, 1, \ldots$.

So the s-s $P(\text{barber is idle}) = P_0 = 1/3$.

Moreover, $P(\text{at least 4 guys in the shop}) = 1 - \sum_{n=0}^{3} P_n = 16/81$.

The steady-state average number of customers in the system is

$$L = \frac{\rho}{1-\rho} = 2. \quad \square$$

$M/D/1$: This is a special case of the $M/G/1$ with mean service time $1/\mu$ and $\sigma^2 = 0$. Then we have

$$L = \rho + \frac{\rho^2}{2(1-\rho)}, \quad L_Q = L - \rho, \quad w = \frac{L}{\lambda}, \quad w_Q = \frac{L_Q}{\lambda}$$

Remark: Less variability in the $M/D/1$ compared to the $M/M/1$ makes for better performance. For instance, for a given λ and μ, the $M/D/1$'s value of L is $<$ the $M/M/1$'s.

Proof: Need to show that

$$\rho + \frac{\rho^2}{2(1-\rho)} < \frac{\rho}{1-\rho},$$

which easily follows after a little algebra. $\square$

$M/E_k/1$: This is another special case of the $M/G/1$. Let's suppose $E_k \sim \text{Erlang}_k(\nu)$ denotes the sum of k i.i.d. Exp(ν) random variables. Then the mean service time is

$$\frac{1}{\mu} = \text{E}[E_k] = \frac{k}{\nu},$$

and the variance is

$$\sigma^2 = \text{Var}(E_k) = \frac{k}{\nu^2}.$$

Note that $\rho = \lambda/\mu = \lambda k/\nu$. Then

$$L = \rho + \frac{1+k}{2k}\frac{\rho^2}{1-\rho}, \quad L_Q = L - \rho, \quad w = \frac{L}{\lambda}, \quad w_Q = \frac{L_Q}{\lambda}$$

$M/M/c$: Multiple servers with exponential interarrivals and services. P-K doesn't apply here, so tables get a little messy, but not too bad.

$$\rho = \lambda/(c\mu)$$

$$P_0 = \left\{ \left[\sum_{i=0}^{c-1} \frac{(c\rho)^i}{i!} \right] + \left[\frac{(c\rho)^c}{(c!)(1-\rho)} \right] \right\}^{-1}$$

$$P(L(\infty) \geq c) = \frac{(c\rho)^c P_0}{(c!)(1-\rho)}$$

$$L = c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2}$$

$$L_Q = L - c\rho = \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2}$$

$$w = L/\lambda$$

$$w_Q = L_Q/\lambda = w - \frac{1}{\mu}$$

Example: Suppose we have a Poisson arrival process with $\lambda = 5$/hr who go to a station with $c = 2$ parallel servers, who can provide services that are i.i.d. Exponential at the rate of 5 per hour. What's the steady-state expected number of customers that will be in the queue?

This is an $M/M/2$ with $\lambda = 5$ and $\mu = 5$. Here $\rho = \lambda/(2\mu) = 0.5$, and the queueing table (with $c = 2$) gives

$$P_0 = \left\{ \left[\sum_{i=0}^{1} \frac{1}{i!} \right] + 1 \right\}^{-1} = 1/3.$$

Then

$$L_Q = L - c\rho = \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = 1/3. \quad \square$$

$M/M/1/N$: P-K doesn't apply, since now we only have *finite* capacity. Just believe the table abstracted from BCNN, where we use $a = \lambda/\mu$ and define λ_e as the *effective* arrival rate, i.e., the rate of arrivals who enter and *remain* in the system. Of course, $\lambda_e \leq \lambda$.

$$
\begin{aligned}
P_n &= \begin{cases} \frac{(1-a)a^n}{1-a^{N+1}} & \text{if } \lambda \neq \mu \\ 1/(N+1) & \text{if } \lambda = \mu \end{cases}, n = 0, 1, 2, \ldots, N \\
\lambda_e &= \lambda(1 - P_N) = \mu(1 - P_0) = \mu_e \\
\rho &= \lambda_e/\mu = 1 - P_0 \\
L &= \begin{cases} \frac{a[1-(N+1)a^N+Na^{N+1}]}{(1-a^{N+1})(1-a)} & \text{if } \lambda \neq \mu \\ N/2 & \text{if } \lambda = \mu \end{cases} \\
w &= L/\lambda_e, \qquad w_Q = w - \frac{1}{\mu} \\
L_Q &= \lambda_e w_Q = L - (1 - P_0)
\end{aligned}
$$

Example: Suppose we're dealing with a single-server queueing system in which the arrivals are Poisson with a mean of 12 minutes, and the services are exponential with rate 10 per hour. Suppose that the system can only hold 4 customers (in line + being served). Find the expected number of customers in line.

This is an $M/M/1/4$ with $a = \lambda/\mu = 5/10$. Then from the table,

$$P_0 = \frac{1-a}{1-a^5} = 0.516,$$

and

$$L = \frac{a[1-5a^4+4a^5]}{(1-a^5)(1-a)} = 0.839,$$

so that $L_Q = L - (1 - P_0) = 0.355$. $\square$

$M/M/c/N$: Use $a = \lambda/\mu$ and define λ_e as the *effective* arrival rate.

$$
\begin{aligned}
\rho &= \lambda/(c\mu) \\
P_0 &= \left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c}\right]^{-1} \\
P_N &= \frac{a^N P_0}{c! c^{N-c}} \\
L_Q &= \frac{a^c \rho P_0}{c!(1-\rho)^2}\left[1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho)\right] \\
\lambda_e &= \lambda(1 - P_N) \\
w_Q &= L_Q/\lambda_e \\
w &= w_Q + \frac{1}{\mu} \\
L &= \lambda_e w
\end{aligned}
$$

$M/M/\infty$: Think of this as a self-service system.

$$\begin{aligned} L &= \lambda/\mu \\ L_Q &= 0 \\ w &= 1/\mu \\ w_Q &= 0 \\ P_n &= \frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!},\ n = 0, 1, 2, \ldots \end{aligned}$$

Remark: Note the Poisson probabilities, $P_n = P\big(\text{Pois}(\lambda/\mu) = n\big)$.

Remark: By the way, you might want to go to BCNN's website www.BCNN.net to take a look at their QueueingTools.xls. Lots of good stuff there.