Module 10: Semi-Markovian Queueing Systems

Lecture 38

M/G/1 Queues, The Pollaczek-Khinchin Mean Formula

Semi-Markovian Queueing Systems

- We have so far considered queueing models where the underlying distributions were assumed to be related to exponential and hence the resulting model was a CTMC and hence a Markov analysis was employed.
- We will now relax the exponential distribution assumptions and involve general distributions.
- The queues can no longer be modelled as CTMCs and the corresponding analysis will not be applicable anymore.
- We will relax the exponential assumption on either the arrival process or the service process (and not both) for now. The resulting queueing models are called as semi-Markovian queueing models (because of the underlying processes).
- Within the context of semi-Markovian queues, we will come across embedded Markov chains embedded within a continuous-time non-Markov process and we can employ some of the theory of Markov chains to analyze the queues.
- While we use only the embedded Markov chain technique for our analysis, there are also other methods available for the analysis (like the supplementary variable technique, etc.).

$M/G/1$ Queues

- The first model that we consider is an $M/G/1$ queue.
- All the assumptions of an $M/M/1$ queue are intact, except that the service time distribution is a general distribution (instead of exponential).
- Let λ be the arrival rate (of the Poisson process).
- Let S denote a (nonnegative) random service time with a general distribution and $\mu = \dfrac{1}{E[S]}$ be the service rate.
- The assumption $\rho = \dfrac{\lambda}{\mu} < 1$ is essential for our equilibrium analysis of the $M/G/1$ queue.
- The analysis of $M/G/1$ is substantially difficult as compared to $M/M/1$.
  ▶ The state description of $M/M/1$ is simple as one needs just the number in the system (helped by the memoryless property of exponential distribution).
  ▶ For an $M/G/1$ queue, its general state description would require specification of both the number in the system and the amount of service already provided to a customer being served when an arrival takes place.

Expected-Value Measure of Effectiveness: The Pollaczek-Khinchin (PK) Formula

- We will now consider deriving a collection of formulas for the expected-value measures of performance for the $M/G/1$ queue: L, L_q, W and W_q.
  - ▶ These formulas are known as Pollaczek-Khinchin (PK) formulas or specifically as Pollaczek-Khinchin (PK) mean formulas.
  - ▶ We first obtain one of these four measures and obtain the others using the relationships among them in the usual way.
- The first derivation obtains the results by considering the system at times when customers arrive at the system.
- The second derivation obtains the (same) results by considering the system at times when customers depart from the system.

Derivation Using Arrival Times

- Consider a customer arriving to the queue. Her delay is determined by the customers who are already in the system when she arrives.
- In particular, there may be customers in the queue, and there may be a customer already in service.
- Let us consider the customers in the queue at the time of her arrival.
  - ► Each customer who is in the queue ahead of her contributes, on average, $E[S]$, to her delay.
  - ► There are, on average, L_q customers in the queue when she arrives. (This holds under Poisson arrivals and, by the PASTA property, an arrival sees L_q in front of her)
  - ► Thus, her average delay due to these customers is $L_q E[S]$.
- Now, the customer who is in service (if there is any) when she arrives contributes a different amount to her delay. This customer has completed some of his service already, so his contribution to her delay is his remaining service time, not his total service time.

- Combining these two, the average queue wait for the arriving customer is

$$W_q = L_q E[S] + P\{\text{server busy}\} \cdot E[\text{residual service time} \mid \text{server busy}].$$

Using $L_q = \lambda W_q$ to eliminate L_q and then rearranging terms gives

$$W_q = \frac{P\{\text{server busy}\} \cdot E\left[\text{residual service time} \mid \text{server busy}\right]}{1-\rho}.$$

- In the above, $P\{\text{server busy}\}$ is the probability that the arriving customer finds the server busy. By the PASTA property, this is the same as the fraction of time the server is busy, so $P\{\text{server busy}\} = \rho$.

- We need to find the expected residual service time, conditional on the arrival finding the server busy. It can be shown that

$$E\left[\text{residual service time} \mid \text{server busy}\right] = \frac{E[S^2]}{2E[S]} = \frac{1+C_B^2}{2}E[S],$$

where C_B^2 is the squared coefficient of variation (SCV) of the service distribution, namely $\frac{Var(S)}{E^2(S)}$.

♦ Recall that the above result is the average excess or average residual time of a renewal process.

- Unless $C_B^2 = 0$, the expected remaining service time as seen by a customer arriving to a busy server is more than half of the expected service time (this is an example of the inspection paradox or paradox of residual life).

► This counter-intuitive result is because customers are more likely to arrive during long service intervals compared with shorter ones, bringing the average above $E[S]/2$.

- Finally, combining the preceding results gives us

$$W_q = \frac{1 + C_B^2}{2} \cdot \frac{\rho}{1 - \rho} \cdot E(S)$$

- This formula has three terms: a variability term, a utilization term, and a time scale term.
- *The first term* $(1 + C_B^2)/2$ involves the squared coefficients of variation of the service distribution C_B^2. For exponential service, $C_B^2 = 1$ so $(1 + C_B^2)/2 = 1$ and, in this case, formula for W_q reduces to the analogous formula for the $M/M/1$ queue.
- *The second term* $\rho/1 - \rho$ involves the queue utilization and increases to infinity as $\rho \to 1$.
- *The last term* $E(S)$ has units of time and can be thought of as a time-scale factor.
- Thus, W_q is the product of two time quantities that are independent of the time scale chosen and the time-dependent quantity $E(S)$.

- The formula for W_q is a powerful result. Only three parameters are needed to compute W_q:
  - The arrival rate λ.
  - The mean $E(S) = 1/\mu$ of the service distribution.
  - The squared coefficient of variation (SCV) C_B^2.

- Other measure of effectiveness can easily be obtained from W_q.
  Using Little's law and/or $W = W_q + E(S)$, we obtain

$$L_q = \lambda W_q, \quad W = W_q + 1/\mu, \quad L = \lambda W = L_q + \rho.$$

Table of PK Mean Formulas for M/G/1 Queue

The following table shows several different ways to express the results. The first column gives the performance mearues and the second column express the measure using the SCV of the service distribution C_B^2, the third column uses the second moment of the service distribution $E(S^2)$, and fourth column uses the variance of the service distribution σ_B^2.

L_q	$\frac{1+C_B^2}{2} \cdot \frac{\rho^2}{1-\rho}$	$\frac{\lambda^2 E\left[S^2\right]}{2(1-\rho)}$	$\frac{\rho^2+\lambda^2\sigma_B^2}{2(1-\rho)}$
W_q	$\frac{1+C_B^2}{2} \cdot \frac{\rho}{\mu-\lambda}$	$\frac{\lambda E\left[S^2\right]}{2(1-\rho)}$	$\frac{\rho^2/\lambda+\lambda\sigma_B^2}{2(1-\rho)}$
W	$\frac{1+C_B^2}{2} \cdot \frac{\rho}{\mu-\lambda} + \frac{1}{\mu}$	$\frac{\lambda E\left[S^2\right]}{2(1-\rho)} + \frac{1}{\mu}$	$\frac{\rho^2/\lambda+\lambda\sigma_B^2}{2(1-\rho)} + \frac{1}{\mu}$
L	$\frac{1+C_B^2}{2} \cdot \frac{\rho^2}{1-\rho} + \rho$	$\frac{\lambda^2 E\left[S^2\right]}{2(1-\rho)} + \rho$	$\frac{\rho^2+\lambda^2\sigma_B^2}{2(1-\rho)} + \rho$

Example

Consider an $M/E_k/1$ queue. The SCV of an E_k distribution equals $1/k$. Therefore,

$$W_q = \frac{1+1/k}{2}\frac{\rho}{1-\rho}E(S)$$

which coincides with earlier results.

Similarly, we can obtain the results of $M/D/1$ either from this by letting $k \to \infty$ or by taking $C_B^2 = 0$ in the PK formula in the table.