Lecture-20: Queues

1 Continuous time queues

A queueing system consists of arriving entities buffered to get serviced by a collection of servers with finite service capacity. The notation $A/T/N/B/S$ for a queueing system indicates

$A:$ inter-arrival time distribution,

$T:$ service time distribution,

$N:$ number of servers,

$B:$ buffer size, or the maximum number of entities waiting and in service at any time (∞ by default),

$S:$ queueing service discipline (FIFO by default).

Typical inter-arrival times are general independent (*GI*) so that number of arrivals is a renewal counting process, memoryless (*M*) for Poisson arrivals, phase-type, or deterministic (*D*). Similarly, the typical service times are general independent (*GI*), memoryless (*M*) for exponential service times, phase-type, or deterministic (*D*). The number of servers could be one, finite, or countably finite. The buffer size is typically arbitrarily large, or equal to the number of servers. Service discipline is usually first-come-first-served (FCFS), last-come-first-served (LCFS), or priority-ordered with or without pre-emption, or processor-shared (PS).

Typical performance metrics of interest are the sojourn times of each arriving entity, and number of entities in the queue as seen by the arriving/departing customer or by the system.

1.1 GI/G/1 queue

The nth entity arrives at instant A_n and requires service σ_n, and the duration between $(n+1)$th and nth entity is denoted by $\xi_n = A_n - A_{n-1}$. The random inter-arrival sequence $\xi : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ and random service times sequence $\sigma : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ are assumed to be *i.i.d.* and independent. The arrival point process $A : \mathbb{R}_+^{\mathbb{N}}$ is assumed to be simple, that is $P\{\xi_1 > 0\} = 1$, and hence this point process is a renewal process. The arrival rate is denoted by $\lambda \triangleq \frac{1}{\mathbb{E}\xi_1}$, and the service rate is denoted by $\mu \triangleq \frac{1}{\mathbb{E}\sigma_1}$. The average load on the system is denoted by $\rho \triangleq \frac{\mathbb{E}\sigma_n}{\mathbb{E}\xi_n} = \frac{\lambda}{\mu}$.

The number of arrivals and departures in a time duration $I \subseteq \mathbb{R}_+$ are denoted by $N_A(I)$ and $N_D(I)$ respectively. The departure instant and waiting time for the start of the service of the nth customer are denoted by D_n and W_n respectively. The number of entities in the buffer at time t is denoted by L_t, and hence $L : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ is a random process. Defining $(x)_+ \triangleq \max\{x, 0\}$, and letting $W_0 = w$, we can write the waiting time for $(n+1)$th customer before it receives service, as

$$W_{n+1} = (W_n + \sigma_n - \xi_{n+1})_+,\ n \in \mathbb{N}.$$

We define nth step-size $X_n = \sigma_n - \xi_{n+1}$ for a random walk $S_n = \sum_{i=1}^n X_i$ with $S_0 = 0$. For the random walk $S : \Omega \to \mathbb{R}^{\mathbb{Z}_+}$, the history of until nth step is denoted by $\mathcal{F}_n \triangleq \sigma(\sigma_1, \dots, \sigma_n, \xi_1, \dots, \xi_{n+1})$. In terms of the *i.i.d.* step-size sequence $X : \Omega \to \mathbb{R}^{\mathbb{N}}$, we can write $W_{n+1} = (W_n + X_n)_+$ for each $n \in \mathbb{N}$. From the independence of sequence $((\sigma_n, \xi_{n+1}) : n \in \mathbb{N})$, it follows that reflected random walk $W : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ is a Markov process.

Theorem 1.1 (Poisson arrivals see time averages (PASTA)). *At any time t, we denote a system state by X_t. Let $B \in \mathcal{B}(\mathbb{R}_+)$ a Borel measurable set, then*

$$\bar{\tau}_B \triangleq \lim_{t \in \mathbb{R}_+} \frac{1}{t} \int_0^t \mathbb{1}_{\{X_u \in B\}} du = \lim_{n \in \mathbb{N}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\left\{X_{t_i^-} \in B\right\}} \triangleq \bar{c}_B.$$

Proof. We will show the special case when $X_t = L_t$ is the number of customers in the system at time t, and $B = \{n\}$. Using continuity of probability, we define for $n \in \mathbb{Z}_+$

$$\pi_n \triangleq \lim_{t \to \infty} P\{L_t = n\}, \qquad \alpha_n \triangleq \lim_{k \in \mathbb{N}} P\{L_{t_k^-} = n\} = \lim_{k \in \mathbb{N}} \lim_{h \downarrow 0} P\{L_{t_k - h} = n | L_{t_k} = n+1\}.$$

Using independent increment property of Poisson arrivals, Baye's rule, and the fact that $\lim_{k\in\mathbb{N}} t_k = \infty$, we can write the second limiting probability as

$$\alpha_n = \lim_{k\in\mathbb{N}} \lim h \downarrow 0 \frac{P\{L_{t_k-h} = n, N_A(t_k-h,t_k] = 1\}}{P\{N_A(t_k-h,t_k] = 1\}} = \lim_{t\to\infty} P\{L_t = n\} = \pi_n.$$

□

Theorem 1.2 (Little's law). *For a $GI/G/1$ queue with $\rho < 1$,*

$$\lim_{t\to\infty} \frac{\int_0^t L_u du}{t} = \lim_{t\to\infty} \frac{\sum_{i=1}^{N_A(0,t]}(W_i+\sigma_i)}{N_A(0,t]}.$$

Proof. The key observation follows from looking at the piecewise constant curve L_t, to conclude

$$\sum_{i=1}^{N_D(0,t]}(W_i+\sigma_i) \le \int_0^t L_u du \le \sum_{i=1}^{N_A(0,t]}(W_i+\sigma_i).$$

Further, for a stable queue we have $\lim_{t\to\infty}\frac{N_D(0,t]}{t} = \lim_{t\to\infty}\frac{N_A(0,t]}{t}$. Combining these two results, the theorem follows from renewal reward theorem. □

1.2 M/M/1 queue

We consider the simplest continuous time queueing system with Poisson arrivals of homogeneous rate $\lambda = \frac{1}{\mathbb{E}\xi_1}$, independent *i.i.d.* exponential service time of rate $\mu = \frac{1}{\mathbb{E}\sigma_1}$ for each arrival, single server with infinite buffer size, and FCFS service discipline. It is clear that $L : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ is a right continuous process with left limits, and is piece-wise constant. We observe that L_t remains unchanged in the time $t + [0, \min\{Y_A(t), Y_S(t)\})$. Further, L_t can have at most one transition in an infinitesimally small interval $(t,t+h]$ with high probability, since the probability of two or more transitions is of order $o(h)$. Further, we observe that L_t can have a unit increase if $Y_A(t) < Y_S(t)$ and a unit decrease otherwise, for $L_t \geqslant 1$. If $L_t = 0$, there can be no service and L_t remains 0 until $t+Y_A(t)$, and has a unit increase at time $t+Y_A(t)$.

Since the arrival and the service times are memoryless, the residual time for next arrival $Y_A(t)$ is identically distributed to ξ_1 and independent of past $\mathcal{F}_t$ and residual service time for entity in service $Y_S(t)$ is identically distributed to σ_1 and independent of past $\mathcal{F}_t$. It follows that L is a homogeneous CTMC, and we can write the corresponding generator matrix as

$$Q(n,m) = \lambda \mathbb{1}_{\{m-n=1\}} + \mu \mathbb{1}_{\{n-m=1, m\geqslant 0\}}.$$

We observe that $Q(n,n) = -(\lambda+\mu)$ for $n \in \mathbb{N}$ and $Q(0,0) = -\lambda$.

The M/M/1 queue is the simplest and most studied models of queueing systems. We assume a continuous-time queueing model with following components.

- There is a single queue for waiting that can accommodate arbitrarily large number of customers.
- Arrivals to the queue occur according to a Poisson process with rate $\lambda > 0$. That is, let A_n be the arrival instant of the nth customer, then the sequence of inter-arrival times ξ is *i.i.d.* exponentially distributed with rate λ.
- There is a single server and the service time of nth customer is denoted by a random variable σ_n. The sequence of service times $\sigma : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ is *i.i.d.* exponentially distributed with rate $\mu > 0$, independent of the Poisson arrival process.
- We assume that customers join the tail of the queue, and hence begin service in the order that they arrive *first-in-queue-first-out* (FIFO).

Let X_t denote the number of customers in the system at time $t \in \mathbb{R}_+$, where "system" means the queue plus the service area. For example, $X_t = 2$ means that there is one customer in service and one waiting in line. Due to continuous distributions of inter-arrival and service times, a transition can only occur at customer arrival or departure times. Further, departures occur whenever a service completion occurs. Let D_n denote the nth departure from the system. At an arrival time A_n, the number $L_{A_n} = L_{A_n^-} + 1$ jumps up by the amount 1, whereas at a departure time D_n, then number $L_{D_n} = L_{D_n^-} - 1$ jumps down by the amount 1.

For the M/M/1 queue, one can argue that $L : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ is a CTMC on the state space $\mathbb{Z}_+$. We will soon see that a *stable* M/M/1 queue is time-reversible.

1.2.1 Transition rates

Given the current state $\{X_t = i\}$, the only transitions possible in an infinitesimal time interval are (a) a single customer arrives, or (b) a single customer leaves (if $i \geq 1$). It follows that the infinitesimal generator for the CTMC $\{X_t\}_t$ is

$$Q_{ij} = \begin{cases} \lambda, & j = i+1, \\ \mu, & j = i-1, \\ 0, & |j-i| > 1. \end{cases}$$

Since $\lambda, \mu > 0$, this defines an irreducible CTMC.

1.2.2 Equilibrium distribution and reversibility

We can define the load $\rho = \frac{\lambda}{\mu}$, and find the stationary distribution π by solving the global balance equation $\pi = \pi Q$ which gives

$$\pi_{n-1} Q_{n-1,n} + \pi_{n+1} Q_{n+1,n} = -\pi_n Q_{nn}, \qquad \pi_1 Q_{1,0} = -\pi_0 Q_{00}.$$

Taking the discrete Fourier transform $\Pi(z) = \sum_{n \in \mathbb{Z}_+} z^n \pi_n$ of the distribution π, we get $z\lambda\Pi(z) + z^{-1}\mu(\Pi(z) - \pi(0)) = (\lambda+\mu)\Pi(z) - \mu\pi(0)$. That is, $\Pi(z) = \pi(0)/(1-z\rho)$. Hence it follows from $\sum_{n \in \mathbb{Z}_+} \pi(n) = 1$ that

$$\pi(n) = (1-\rho)\rho^n, \; n \in \mathbb{Z}_+.$$

Example 1.3 ($M/M/1$ queue). The M/M/1 queue's generator defines a birth-death process. Hence, if it is stationary, then it must be time-reversible, with the equilibrium distribution π satisfying the detailed balance equations $\pi_n \lambda = \pi_{n+1}\mu$ for each $n \in \mathbb{Z}_+$. This yields $\pi_{n+1} = \rho\pi_n$ for the system load $\rho = \mathbb{E}\sigma_1/\mathbb{E}\xi_1 = \lambda/\mu$. Since $\sum_{i \geq 0} \pi = 1$, we must have $\rho < 1$, such that $\pi_n = (1-\rho)\rho^n$ for each $n \in \mathbb{Z}_+$. In other words, if $\lambda < \mu$, then the equilibrium distribution of the number of customers in the system is geometric with parameter $\rho = \lambda/\mu$. We say that the M/M/1 queue is in the *stable* regime when $\rho < 1$.

Corollary 1.4. *The number of customers in a stable M/M/1 queueing system at equilibrium is a reversible Markov process.*

Further, since M/M/1 queue is a reversible CTMC, the following theorem follows.

Theorem 1.5 (Burke). *Departures from a stable M/M/1 queue are Poisson with same rate as the arrivals.*