Lecture-22: 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.

1.1 Notation

The notation $A/T/N/B/S$ for a queueing system indicates different components.

- A : stands for inter-arrival time distribution. Typical inter-arrival time distributions are general independent (GI) so that number of arrivals is a renewal counting process, memoryless (M) for Poisson arrivals, phase-type (PH), or deterministic (D).
- T : stands for service time distribution. Similar to inter-arrival time distribution, the typical service time distributions are general independent (GI), memoryless (M) for exponential service times, phase-type (PH), or deterministic (D).
- N : stands for number of servers. The number of servers could be one, finite (N), or countably infinite (∞).
- B : stands for the buffer size, or the maximum number of entities waiting and in service at any time. The buffer size is typically arbitrarily large (∞), or equal to the number of servers. If there is no buffer size specified, then it is countably infinite by default.
- S : stands for the queueing service discipline. 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). If there is no queueing discipline specified, then it is FIFO by default.

Typical performance metrics of interest are the sojourn times averaged over each arriving entity, and the number of entities in the queue as seen by an incoming arrival or outgoing departure from the system.

1.2 GI/GI/1 queue

We denote the random sequence of arrival instants by $A : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ where A_n is the arrival instant of nth entity. The inter-arrival time sequence is denoted by $\xi : \Omega \to \mathbb{R}_+^{\mathbb{N}}$, where $\xi_n \triangleq A_n - A_{n-1}$ is the duration between the $(n-1)$th and nth arrival instants. The random service requirement sequence is denoted by $\sigma : \Omega \to \mathbb{R}_+^{\mathbb{N}}$, where σ_n is the amount of service needed by nth arrival. For simplicity of analysis, one assumes that the random inter-arrival sequence $\xi : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ and random service time sequence $\sigma : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ are *i.i.d.* and independent to each other. 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 $\rho \triangleq \frac{\mathbb{E}\sigma_n}{\mathbb{E}\xi_n} = \frac{\lambda}{\mu}$.

We denote the random departure instant sequence by $D : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ where D_n is the departure instant of nth arrival, the random waiting time sequence by $W : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ where W_n is the waiting time of nth arrival, and the buffer occupancy process by $L : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ where L_t is the number of entities in the buffer at time $t \in \mathbb{R}_+$. These are derived processes from the arrival instant and service time processes, given the number of servers, the buffer size, and the service discipline. 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. When the interval is $(0,t]$ for some $t \in \mathbb{R}_+$, then we denote $N_t^A \triangleq N^A(0,t]$ and $N_t^D \triangleq N^D(0,t]$. Defining $(x)_+ \triangleq \max\{x,0\}$,

and for initial waiting time $W_0 \triangleq w$, we can write the waiting time for $(n+1)$th customer before it receives service, as

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

We define a random walk $S : \Omega \to \mathbb{R}^{\mathbb{N}}$ defined as $S_n \triangleq \sum_{i=1}^n X_i$ for all $n \in \mathbb{N}$ with $S_0 \triangleq 0$, where *i.i.d.* step-size sequence $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ is defined as $X_{n+1} \triangleq \sigma_n - \xi_{n+1}$ for the step-size $n \in \mathbb{N}$. For the random walk $S : \Omega \to \mathbb{R}^{\mathbb{N}}$, the history until nth step is denoted by $\mathcal{F}_n \triangleq \sigma(\sigma_0, \ldots, \sigma_{n-1}, \xi_1, \ldots, \xi_n)$. In terms of the *i.i.d.* step-size sequence $X : \Omega \to \mathbb{R}^{\mathbb{N}}$, we can write the waiting time sequence W as the reflected random walk, where $W_{n+1} = (W_n + X_{n+1})_+$ for each $n \in \mathbb{Z}_+$. 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. Since X is *i.i.d.* , it follows that W is time homogeneous.

1.3 Poisson arrivals see time averages (PASTA)

Consider a stochastic process $X : \Omega \to \mathcal{X}^{\mathbb{R}_+}$ and a homogeneous Poisson arrival counting process $N : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ with rate λ defined on the same probability space $(\Omega, \mathcal{F}, P)$, such that X_t is the system state at time t and N_t is the number of arrivals in the duration $(0,t]$. We define the natural filtration $\mathcal{F}_\bullet \triangleq (\mathcal{F}_t : t \in \mathbb{R}_+)$ for the joint process (X, N), such that $\mathcal{F}_t \triangleq \sigma(X_s, N_s, s \leqslant t)$ for all $t \in \mathbb{R}_+$.

Assumption 1.1 (Lack of anticipation (LAA)). Increment $N_s - N_t$ is independent of $\mathcal{F}_t$ for all $s \geqslant t$.

Definition 1.2. For a Borel measurable set $B \in \mathcal{B}(\mathcal{X})$, we define a left continuous with right limits process $U_t \triangleq \mathbb{1}_{\{X_t \in B\}}$ for all times $t \in \mathbb{R}_+$, to define two derived processes $t \mapsto V_t \triangleq \int_0^t U_s ds$ and $t \mapsto Y_t \triangleq \int_0^t U_s dN_s$. The asymptotic time average of system being in state B is defined as

$$\bar{\tau}_B \triangleq \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{1}_{\{X_u \in B\}} du = \lim_{t\to\infty} \frac{V_t}{t}.$$

We define the asymptotic average of the system being in state B as seen by an arriving customer as

$$\bar{c}_B \triangleq \lim_{n\in\mathbb{N}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\left\{X_{A_i^-} \in B\right\}} = \lim_{t\to\infty} \frac{Y_t}{N_t}.$$

Theorem 1.3 (PASTA). *Under LAA assumption, $\bar{\tau}_B = \bar{c}_B$ almost surely.*

Proof. We define a process $t \mapsto R_t \triangleq Y_t - \lambda V_t$ for all $t \in \mathbb{R}_+$. Since $\lim_{t\to\infty} \frac{t}{N_t} = \lambda$ almost surely, it suffices to show that $\lim_{t\to\infty} \frac{R_t}{t} = 0$ almost surely.

Step 1: We will show that $\mathbb{E}[R_{t+h} - R_t \mid \mathcal{F}_t]$ for any $t, h \in \mathbb{R}_+$. For each $t, h \in \mathbb{R}_+$ and $n \in \mathbb{N}$, we define $Y_{t,h}^n \triangleq \sum_{k=0}^{n-1} U_{t+\frac{kh}{n}} (N_{t+\frac{(k+1)h}{n}} - N_{t+\frac{kh}{n}})$. We observe that $\lim_{n\in\mathbb{N}} Y_{t,h}^n = Y_{t+h} - Y_t$ almost surely. From the LAA assumption, we get $\mathbb{E}[Y_{t,h}^n \mid \mathcal{F}_t] = \lambda \mathbb{E}\left[\frac{h}{n} \sum_{k=0}^{n-1} U_{t+\frac{kh}{n}} \mid \mathcal{F}_t\right]$. Applying bounded convergence theorem for the conditional expectation, we obtain

$$\mathbb{E}[Y_{t+h} - Y_t \mid \mathcal{F}_t] = \mathbb{E}[\lim_{n\in\mathbb{N}} Y_{t,h}^n \mid \mathcal{F}_t] = \lim_{n\in\mathbb{N}} \mathbb{E}[Y_{t,h}^n \mid \mathcal{F}_t] = \lambda \mathbb{E}[V_{t+h} - V_t \mid \mathcal{F}_t].$$

It follows that $\mathbb{E}[R_{t+h} - R_t \mid \mathcal{F}_s] = 0$ for any $s \leqslant t$ and R is continuous time martingale adapted to filtration $\mathcal{F}_\bullet$.

Step 2: We will show that $\lim_{t\to\infty} \frac{R_t}{t} = 0$ almost surely. Since $U_u \in \{0,1\}$ is an indicator function, it follows that $0 \leqslant V_t \leqslant t$ and $0 \leqslant Y_t \leqslant N_t$. Therefore, $\mathbb{E}R_t^2 \leqslant \mathbb{E}Y_t^2 + \lambda^2 \mathbb{E}V_t^2 \leqslant \lambda t + \lambda^2 t^2$. We observe that $n \mapsto R_{nh} - R_{(n-1)h}$ is discrete-time martingale adapted to filtration $\mathcal{F}_\bullet$, and $\sum_{n\in\mathbb{N}} \frac{1}{n^2} \mathbb{E}(R_{nh} - R_{(n-1)h})^2 \leqslant \lambda h(1+\lambda h) \sum_{n\in\mathbb{N}} \frac{1}{n^2} < \infty$. It follows that $\lim_{n\to\infty} \frac{R_{nh}}{n} = 0$ almost surely. The result follows from observing that $R_{nh} - \lambda h \leqslant R_t \leqslant R_{(n+1)h} + \lambda h$ for $t \in [nh, (n+1)h)$.

□

Theorem 1.4 (Little's law). *For a $GI/G/1$ queue with $\rho < 1$,*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t L_u du = \lambda \lim_{t\to\infty} \frac{\sum_{i=1}^{N_t^A} (W_i + \sigma_i)}{N_t^A}.$$

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

$$\sum_{i=1}^{N_t^D}(W_i+\sigma_i) \leqslant \int_0^t L_u du \leqslant \sum_{i=1}^{N_t^A}(W_i+\sigma_i).$$

Further, for a stable queue we have $\lim_{t\to\infty}\frac{N_t^D}{t} = \lim_{t\to\infty}\frac{N_t^A}{t}$. Hence,

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t L_u \mathrm{d}u = \lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{N_t^A}(W_i+\sigma_i) = \lim_{t\to\infty}\frac{1}{N_t^A}\frac{N_t^A}{t}\sum_{i=1}^{N_t^A}(W_i+\sigma_i).$$

Now, if $\lim_{t\to\infty}\frac{N_t^A}{t}$ and $\lim_{t\to\infty}\frac{1}{N_t^A}\sum_{i=1}^{N_t^A}(W_i+\sigma_i)$ exist, and $\lim_{t\to\infty}\frac{N_t^A}{t}=\lambda$ we get the result. □

1.4 M/M/1 queue

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 L_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, $L_t = 2$ means that there is one customer in service and one waiting in line.

1.4.1 Transition rates

Since the arrival and the service times are memoryless, the residual time for next arrival Y_t^A is identically distributed to ξ_1 and independent of past $\mathcal{F}_t$ and residual service time for entity in service Y_t^S is identically distributed to σ_1 and independent of past $\mathcal{F}_t$. We observe that L_t remains unchanged in the time $t + [0, \min\{Y_t^A, Y_t^S\})$. It follows that $L : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ is a right continuous process with left limits, and is piece-wise constant. Further, we observe that L_t can have a unit increase if $Y_t^A < Y_t^S$ corresponding to an arrival, and a unit decrease for $L_t \geqslant 1$ if $Y_t^A < Y_t^S$ corresponding to a departure. If $L_t = 0$, there can be no service and L_t remains 0 until $t + Y_t^A$, and has a unit increase at time $t + Y_t^A$. 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 $o(h)$. 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 $Q(0,0) = -\lambda$. It follows that $M/M/1$ queue occupancy is an irreducible CTMC.

1.4.2 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, $\Pi(z) = \frac{\pi(0)}{(1-z\rho)}$. Hence it follows from $\sum_{n\in\mathbb{Z}_+} \pi(n) = 1$ that

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

Example 1.5 ($M/M/1$ 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 $\rho = \frac{\mathbb{E}\sigma_1}{\mathbb{E}\xi_1} = \frac{\lambda}{\mu}$. Since $\sum_{n\in\mathbb{Z}_+} \pi_n = 1$, 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 $\rho = \frac{\lambda}{\mu}$. We say that the M/M/1 queue is in the *stable* regime when $\rho < 1$.

Corollary 1.6. *The number of customers in a stable M/M/1 queueing system at equilibrium is a reversible Markov process.*

Theorem 1.7 (Burke). *Departures from a stable M/M/1 queue are Poisson with same rate as the arrivals.*

Exercise 1.8. Directly characterize the departure process from a stable $M/M/1$ queue at stationarity.