# Lecture-23: 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 ($\infty$).
- $B$ : 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 ($\infty$), 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 FCFS 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

A GI/GI/1 queue has *i.i.d.* inter-arrival time sequence, *i.i.d.* service time sequence, a single server, infinite buffer size, and FCFS queueing discipline.

#### 1.2.1 Fundamental processes

We denote the *i.i.d.* inter-arrival time sequence by $\xi : \Omega \to \mathbb{R}_+^{\mathbb{N}}$, where $\xi_n$ is the time duration between the $(n-1)$th and the $n$th arrival. We assume that $P\{\xi_1 > 0\} = 1$. The random service requirement sequence is denoted by $\sigma : \Omega \to \mathbb{R}_+^{\mathbb{N}}$, where $\sigma_n$ is the amount of service needed by $n$th arrival. The arrival rate is denoted by $\lambda \triangleq 1/\mathbb{E}\xi_1$, and the service rate is denoted by $\mu \triangleq 1/\mathbb{E}\sigma_1$. The average load on the system is denoted by $\rho \triangleq \mathbb{E}\sigma_n/\mathbb{E}\xi_n = \lambda/\mu$. Inter-arrival time sequence $\xi$ and service time sequence $\sigma$ are assumed to be independent.

#### 1.2.2 Derived processes

Given the inter-arrival time and service time processes, the number of servers, the buffer size, and the service discipline, we can derive arrival instant, departure instant, and waiting time sequence. We denote the random sequence of arrival instants by $A : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ where $A_n$ is the $n$th arrival instant defined as

$$A_n \triangleq \sum_{i=1}^{n} \xi_i.$$

Since the inter-arrival time sequence $\xi$ is *i.i.d.* , it follows that arrival instant sequence $A$ is a renewal sequence. Since $P\{\xi_1 > 0\} = 1$, the arrival point process $A : \mathbb{R}_+^{\mathbb{N}}$ is simple. The waiting time sequence is denoted by $W : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ where $W_n$ is the waiting time of $n$th arrival. We define $(x)_+ \triangleq x \vee 0$ and assume the initial waiting time $W_0 \triangleq w$. For each $n \in \mathbb{N}$, we can write the waiting time for $n$th arrival before it receives service, as

$$W_n = (W_{n-1} + \sigma_{n-1} - \xi_n)_+.$$

We define a step-size sequence $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ for step-size $n \in \mathbb{N}$ as $X_n \triangleq \sigma_{n-1} - \xi_n$. We observe that $W_n = (W_{n-1} + X_n)_+$ for each $n \in \mathbb{N}$. Since $\sigma$ and $\xi$ are individually *i.i.d.* and independent, it follows that $X$ is an *i.i.d.* sequence, and hence $W$ is a time homogeneous Markov sequence. Further, we can define a random walk $S : \Omega \to \mathbb{R}^{\mathbb{N}}$ for each time $n \in \mathbb{N}$ as $S_n \triangleq \sum_{i=1}^n X_i$. We note that waiting time sequence is a reflected random walk. We denote the departure instant sequence by $D : \Omega \to \mathbb{R}_+^{\mathbb{N}}$ where $D_n$ is the departure instant of $n$th arrival defined as

$$D_n \triangleq A_n + W_n + \sigma_n.$$

### 1.2.3 Intermediate processes

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

$$N_t^A \triangleq N^A(0,t] = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{A_n \leqslant t\}}, \qquad N_t^D \triangleq N^D(0,t] = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{D_n \leqslant t\}}.$$

Since $A_n \leqslant D_n$, we have $\mathbb{1}_{\{D_n\leqslant t\}} \leqslant \mathbb{1}_{\{A_n\leqslant t\}}$, and hence $N_t^D \leqslant N_t^A$. The buffer occupancy process is denoted by $L : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ where $L_t$ is the number of entities in the buffer at time $t \in \mathbb{R}_+$ is defined as

$$L_t \triangleq \sum_{n\in\mathbb{N}} \mathbb{1}_{[A_n,D_n)}(t) = \sum_{n\in\mathbb{N}} (\mathbb{1}_{\{A_n\leqslant t\}} - \mathbb{1}_{\{D_n\leqslant t\}}) = N_t^A - N_t^D \geqslant 0.$$

We are interested in the long term average of waiting time $\bar{W}$ for each arrival and the long term average of buffer occupancy $\bar{L}$, defined as

$$\bar{W} \triangleq \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^N W_n, \qquad \bar{L} \triangleq \limsup_{T\to\infty} \frac{1}{T} \int_0^T L_t dt.$$

## 1.3 Poisson arrivals see time averages (PASTA)

Consider a stochastic process $X : \Omega \to \mathcal{X}^{\mathbb{R}_+}$ and a homogeneous Poisson counting process $N : \Omega \to \mathbb{Z}_+^{\mathbb{R}_+}$ with rate $\lambda$ 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.** Let $B \in \mathcal{B}(\mathcal{X})$ be a Borel measurable set. We define a left continuous with right limits process $U : \Omega \to \{0,1\}^{\mathbb{R}_+}$ for each time $t \in \mathbb{R}_+$ as $U_t \triangleq \mathbb{1}_B(X_{t^-}) = \mathbb{1}_{\{X_{t^-}\in B\}}$. In term of $U$ and counting process $N$, we define two derived processes $V, Y : \Omega \to \mathbb{R}_+^{R_+}$ defined for each time $t \in \mathbb{R}_+$ as

$$V_t \triangleq \int_0^t U_s ds = \int_0^t \mathbb{1}_{\{X_{s^-}\in B\}} ds, \qquad Y_t \triangleq \int_0^t U_s dN_s = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{A_n\leqslant t\}} \mathbb{1}_{\left\{X_{A_n^-}\in B\right\}} = \sum_{n=1}^{N_t} \mathbb{1}_{\left\{X_{A_n^-}\in B\right\}}.$$

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

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

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

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

Step 1: We will show that $R$ is a continuous time martingale. Specifically, we will show that $\mathbb{E}\,|R_t| < \infty$ and $\mathbb{E}[R_{t+h} - R_t \mid \mathcal{F}_t] = 0$ for any $t, h \in \mathbb{R}_+$.

(a) Since $U_s \in \{0,1\}$ is an indicator function for all $s \in \mathbb{R}_+$, it follows that $0 \leqslant V_t \leqslant t$ and $0 \leqslant Y_t \leqslant N_t$. It follows that $|R_t| \leqslant Y_t + \lambda V(t) \leqslant N_t + \lambda t$, and hence $\mathbb{E}\,|R_t| \leqslant 2\lambda t$ for all $t \in \mathbb{R}_+$.

(b) For each $t, h \in \mathbb{R}_+$ and $n \in \mathbb{N}$, we define

$$Y^n_{t,h} \triangleq \sum_{k=0}^{n-1} U_{t+\frac{kh}{n}}(N_{t+\frac{(k+1)h}{n}} - N_{t+\frac{kh}{n}}) = Y_{t+h} - Y_t - \sum_{k=0}^{n-1} \int_{t+\frac{kh}{n}}^{t+\frac{(k+1)h}{n}} (U_s - U_{t+\frac{kh}{n}}) dN_s.$$

We can verify that $\left|Y_{t+h} - Y_t - Y^n_{t,h}\right| < \infty$ and $\lim_{n\in\mathbb{N}} U_{t+\frac{kh}{n}} = U_s$ for all $s \in t + \frac{kh}{n} + [0, \frac{1}{n}]$. Thus, exchanging limit and integration from dominated convergence theorem, we obtain that $\lim_{n\in\mathbb{N}} Y^n_{t,h} = Y_{t+h} - Y_t$ almost surely. Since $N$ is Poisson counting process it satisfies LAA assumption. Together with this fact, and that the counting process $N$ has rate $\lambda$, and $U_s$ is $\mathcal{F}_s$ measurable, we get

$$\mathbb{E}[Y^n_{t,h} \mid \mathcal{F}_t] = \lambda \mathbb{E}\Big[\frac{h}{n} \sum_{k=0}^{n-1} U_{t+\frac{kh}{n}} \mid \mathcal{F}_t\Big].$$

Taking limit $n \in \mathbb{N}$ on both sides of the above equation, and applying dominated convergence theorem to exchange limit and conditional expectation, we obtain

$$\mathbb{E}[Y_{t+h} - Y_t \mid \mathcal{F}_t] = \mathbb{E}[\lim_{n\in\mathbb{N}} Y^n_{t,h} \mid \mathcal{F}_t] = \lim_{n\in\mathbb{N}} \mathbb{E}[Y^n_{t,h} \mid \mathcal{F}_t] = \lambda \mathbb{E}\Big[\lim_{n\in\mathbb{N}} \frac{h}{n} \sum_{k=0}^{n-1} U_{t+\frac{kh}{n}} \mid \mathcal{F}_t\Big] = \lambda \mathbb{E}[V_{t+h} - V_t \mid \mathcal{F}_t].$$

Step 2: We will show that $\lim_{n\to\infty} \frac{R_{nh}}{n} = 0$ almost surely. We fix $h > 0$ and define a discrete filtration $\mathcal{G}_\bullet \triangleq (\mathcal{G}_n : n \in \mathbb{N})$ and a random sequence $Q : \Omega \to \mathbb{R}^{\mathbb{N}}$ where $\mathcal{G}_n \triangleq \mathcal{F}_{nh}$ and $Q_n \triangleq R_{nh} - R_{(n-1)h}$ for all $n \in \mathbb{N}$. We fix $n \in \mathbb{N}$, and observe that

(a) $\sigma(R_{nh}) \subseteq \mathcal{G}_n$,
(b) $\mathbb{E}\,|R_{nh}| \leqslant \mathbb{E}\,|R_{nh}| \leqslant 2\lambda nh$, and
(c) $\mathbb{E}[R_{nh} \mid \mathcal{G}_{n-1}] = \mathbb{E}[R_{nh} \mid \mathcal{F}_{(n-1)h}] = R_{(n-1)h}$.

It follows that $Q$ is a discrete time martingale difference sequence for martingale $(R_{nh} : n \in \mathbb{Z}_+)$ adapted to $\mathcal{G}_\bullet$. Fix $n \in \mathbb{N}$, and observe that $Q_n = (Y_{nh} - Y_{(n-1)h}) - \lambda(V_{nh} - V_{(n-1)h})$ where each term is positive. Therefore,

$$Q_n^2 \leqslant (Y_{nh} - Y_{(n-1)h})^2 + \lambda^2 (V_{nh} - V_{(n-1)h})^2 \leqslant (N_{nh} - N_{(n-1)h})^2 + \lambda^2 h^2.$$

Taking expectation on both sides for each $n \in \mathbb{N}$, and scaling with $1/n^2$ and summing over all $n \in \mathbb{N}$, we get $\mathbb{E} \sum_{n\in\mathbb{N}} \frac{Q_n^2}{n^2} \leqslant \lambda h(1 + 2\lambda h) \sum_{n\in\mathbb{N}} \frac{1}{n^2} < \infty$. It follows from Proposition A.2 that $\lim_{n\to\infty} \frac{R_{nh}}{n} = 0$ almost surely.

Step 3: We will show that $\lim_{t\to\infty} \frac{R_t}{t} = 0$ almost surely. Let $n \triangleq \lfloor \frac{t}{h} \rfloor + 1$, then $t \in [(n-1)h, nh)$. Since $R_t = Y_t - \lambda V_t$, we have $R_t - R_s \geqslant -\lambda(V_t - V_s)$ for all $t > s$ and $V_t - V_s \leqslant (t - s)$, we obtain

$$R_{nh} - R_t \geqslant -\lambda(V_{nh} - V_t) \geqslant -\lambda h, \qquad R_t - R_{(n-1)h} \geqslant -\lambda(V_t - V_{(n-1)h}) \geqslant -\lambda h.$$

Combining the two equations, we get $R_{(n-1)h} - \lambda h \leqslant R_t \leqslant R_{nh} + \lambda h$. Dividing both sides of the equation by $t$, taking limit $t \to \infty$, and using the fact that $\lim_{n\in\mathbb{N}} \frac{R_{nh}}{n} = 0$ almost surely, we obtain the result.

□

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

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

*Proof.* Recall that $L_u = \sum_{n\in\mathbb{N}} \mathbb{1}_{[A_n,D_n)}(u)$, and hence applying monotone convergence theorem to exchange integration and infinite sum, we obtain

$$\int_0^t L_u du = \sum_{n\in\mathbb{N}} \int_0^t \mathbb{1}_{[A_n,D_n)}(u)du = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{D_n\leqslant t\}} \int_0^t \mathbb{1}_{[A_n,D_n)}(u)du + \sum_{n\in\mathbb{N}} \mathbb{1}_{\{A_n\leqslant t<D_n\}} \int_0^t \mathbb{1}_{[A_n,D_n)}(u)du.$$

Using the fact that $D_n - A_n = W_n + \sigma_n$ for each $n \in \mathbb{N}$, we can write this integral and bound it as

$$\sum_{n\in\mathbb{N}} \mathbb{1}_{\{n\leqslant N_t^D\}}(W_n+\sigma_n) \leqslant \int_0^t L_u du = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{D_n\leqslant t\}}(W_n+\sigma_n) + \sum_{n\in\mathbb{N}} \mathbb{1}_{\{A_n\leqslant t<D_n\}}(t-A_n) \leqslant \sum_{n\in\mathbb{N}} \mathbb{1}_{\{n\leqslant N_t^A\}}(W_n+\sigma_n).$$

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

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

□

# A Strong law of large number for martingale difference sequence

**Definition A.1.** Consider a martingale $X : \Omega \to \mathbb{R}^{\mathbb{Z}_+}$ adapted to filtration $\mathcal{F}_\bullet \triangleq (\mathcal{F}_n : n \in \mathbb{Z}_+)$. We define the associated martingale difference sequence $Y : \Omega \to \mathbb{R}^{\mathbb{N}}$ as $Y_n \triangleq X_n - X_{n-1}$ for each $n \in \mathbb{N}$.

**Proposition A.2.** *Consider a martingale $X$ and associated martingale difference sequence $Y$. If $\sum_{n\in\mathbb{N}} \frac{Y_n^2}{n^2} < \infty$, then $\lim_{n\in\mathbb{N}} \frac{X_n}{n} = 0$ almost surely.*

*Proof.* We define another sequence $Z : \Omega \to \mathbb{R}^{\mathbb{Z}_+}$ adapted to $\mathcal{F}_\bullet$ where $Z_0 \triangleq 0$ and $Z_n \triangleq \sum_{i=1}^n \frac{Y_i}{i}$ for each $n \in \mathbb{N}$. We fix $n \in \mathbb{N}$, and observe that
(a) $\sigma(Z_n) \subseteq \mathcal{F}_n$,
(b) $\mathbb{E}\,|Z_n| \leqslant \sum_{i=1}^n \frac{1}{i}\mathbb{E}\,|X_n| + \mathbb{E}\,|X_{n-1}| < \infty$, and
(c) $\mathbb{E}[Z_n \mid \mathcal{F}_{n-1}] = \mathbb{E}[Z_{n-1} + \frac{Y_n}{n} \mid \mathcal{F}_{n-1}] = Z_{n-1}$ since $\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = 0$.
It follows that $Z$ is a martingale adapted to $\mathcal{F}_\bullet$ and

$$\mathbb{E}Z_n^2 = \sum_{i=1}^n \frac{1}{i^2}\mathbb{E}Y_i^2 + 2\mathbb{E}\sum_{i>j} \frac{1}{ij}\mathbb{E}[Y_iY_j \mid \mathcal{F}_j] = \mathbb{E}\sum_{i=1}^n \frac{Y_i^2}{i^2}.$$

From the hypothesis we have $\lim_{n\in\mathbb{N}} \mathbb{E}Z_n^2 < \infty$. From martingale convergence theorem, $\lim_{n\in\mathbb{N}} Z_n = Z_\infty$ exists and is finite almost surely. Further, we observe that

$$\frac{X_n}{n} = \frac{1}{n}\sum_{i=1}^n i\frac{Y_i}{i} = \frac{1}{n}\sum_{i=1}^n i(Z_i - Z_{i-1}) = \frac{1}{n}\Big(\sum_{i=1}^n iZ_i - \sum_{i=0}^{n-1}(i+1)Z_i\Big) = Z_n - \frac{1}{n}\sum_{i=0}^{n-1} Z_i.$$

The result follows from taking limit $n \to \infty$ on both sides and the existence of almost sure $Z_\infty$. □