Queueing Model

1 Introduction

Queueing models are mathematical representations of systems in which entities (such as customers, data packets, or jobs) wait in line for service. The key components of queueing models:

- **Arrival Process:** Describes how entities arrive at the queue. Common arrival processes include Poisson processes (exponential interarrival times) and deterministic arrivals (fixed intervals).
- **Service Process:** Describes how entities are served. This can be represented by various distributions, such as exponential, deterministic, or general service times
- **Number of Servers:** Represents the resources providing the service. A single-server queue has one server, while a multi-server queue has multiple servers.
- **Queue Discipline:** Rules determining the order in which entities are served. Common disciplines include First-Come-First-Served (FCFS), Last-Come-First-Served (LCFS), and priority-based systems.
- **System Capacity:** Maximum number of entities that can be in the system, including those in service and waiting. Some systems have infinite capacity, while others are limited.

2 Kendall Notation

Kendall notation is a standardized system used to describe and classify queueing models succinctly. The notation consists of six symbols, commonly written in the form:

$$A/B/c/K/N/D \tag{1}$$

where each symbol represents a different aspect of the queueing system.

2.1 A: Arrival Process

This describes the statistical distribution of the time between arrivals of entities to the system. It includes several options:

- **M (Markovian)**: Exponential interarrival times (Poisson process).

- **D (Deterministic)**: Fixed interarrival times.
- **G (General)**: General distribution with a known mean and variance.
- **Ek (Erlang-k)**: Erlang distribution with shape parameter k.

2.2 B: Service Time Distribution

This describes the statistical distribution of service times. It includes several options:

- **M (Markovian)**: Exponential service times (Poisson process).
- **D (Deterministic)**: Fixed service times.
- **G (General)**: General distribution with a known mean and variance.
- **Ek (Erlang-k)**: Erlang distribution with shape parameter k.

2.3 c: Number of Servers

The number of parallel servers providing service in the system.

2.4 K: System Capacity

The maximum number of entities allowed in the system, including those in service and in the queue. If not specified, it is assumed to be infinite.

2.5 N: Calling Population Size

The size of the population from which the arriving entities are drawn. If noot specified, it is assumed to be infinite.

2.6 D: Queue Discipline

The order in which entities are served. The common disciplines include:

- **FCFS**: First-Come-First-Served
- **LCFS**: Last-Come-First-Served
- **SIRO**: Service In Random Order
- **PR**: Priority

3 Performance Metrics

3.1 Notations

Some important performance metrics of queueing model include:

1. **Queue length:** The number of entities in the queue.
2. **Waiting time**: The time an entity spends waiting in the queue.
3. **Utilization:** The fraction of time the servers are busy.
4. **Throughput:** The rate at which entities are served.

These metrics include some mathematical relations. First, introduce the notations of these variables:

- λ: arrival rate
- μ: service rate
- y(Y if random variables): time that customer spends in the system
- $w(W)$: time customer waits in the queue
- $b(B)$: time customer spends being served. By intuition, we know that:

$$y = w + b \tag{2}$$

$$E[Y] = E[W] + E[B] \tag{3}$$

- $f(F)$: number of customers in the system. If $f = 0$, then the queue is empty. If $f = 1$, then we have $w = 0$ and $y = b$, which means the customer is served right away.
- $d(D)$: throughput, which has the unit of [# customers]/[time]. This depends on λ and μ. If $\lambda < \mu$, then $d = \lambda$. If $\lambda \geq \mu$, then $d = \mu$.
- c: max throughput
- ρ: utilization, or relative throughput. This is defined by $\rho = \frac{d}{c}$. When $\rho = 1$, we have a full queueing system.

3.2 Laws

There are several empirical laws coould be derived. Based on the definition of customers in the system:

$$E[F] = E[D] \cdot E[Y] \tag{4}$$

The number of customers in the queue could be expressed as:

$$E[F_Q] = E[D] \cdot E[W] \tag{5}$$

The number of customers in the serving could be expressed as:

$$E[F_S] = E[D] \cdot E[B] \tag{6}$$

Finally we have:

$$E[F] = E[D]E[Y] = E[D](E[W] + E[B]) = E[F_Q] + E[F_S] \tag{7}$$

4 M/M/1 Probability Calculation

The M/M/1 queueing model is one of the simplest and most commonly analyzed models in queueing theory. It represents a system with a single server where both the arrival process and the service process follow exponential (memoryless) distributions.

4.1 Derivation

Before everything starts, we make an important assumption: the arrival time of the customers are independent from each other. In math, we can divide the timeline into many subintervals. In each subinterval, it is like flipping coin, with some certain samll probability p there is a customer coming into the queue. If we divide 1 unit of time into n subintervals, then the average number of customers coming at this unit of time will be np.

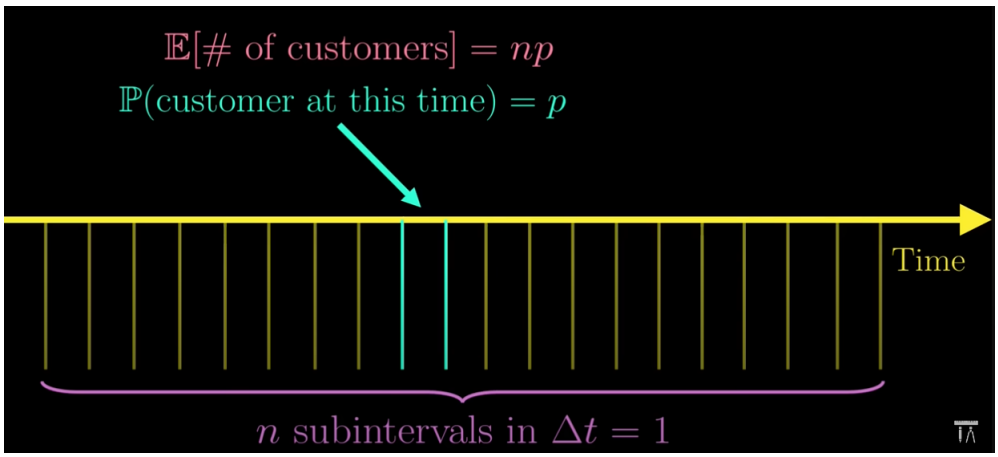

Figure 1: Independent Arrival Time Assumption

However, finally we want to let $n \to \infty$ rather than having discrete intervals, without having infinite number of customers, **we adjust p to $\frac{\lambda}{n}$**. Therefore, **in 1 unit of time**:

$$E[\text{number of customers}] = np = \lambda \tag{8}$$

Now we are interested in the probability that k customers in total arriving in the queue in unit of time. This needs to include the choice of k among n, the probability that k arrives and the probability that $n - k$ not arrives. So the final expression is:

$$P(\text{k customers}) = \binom{n}{k}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \tag{9}$$

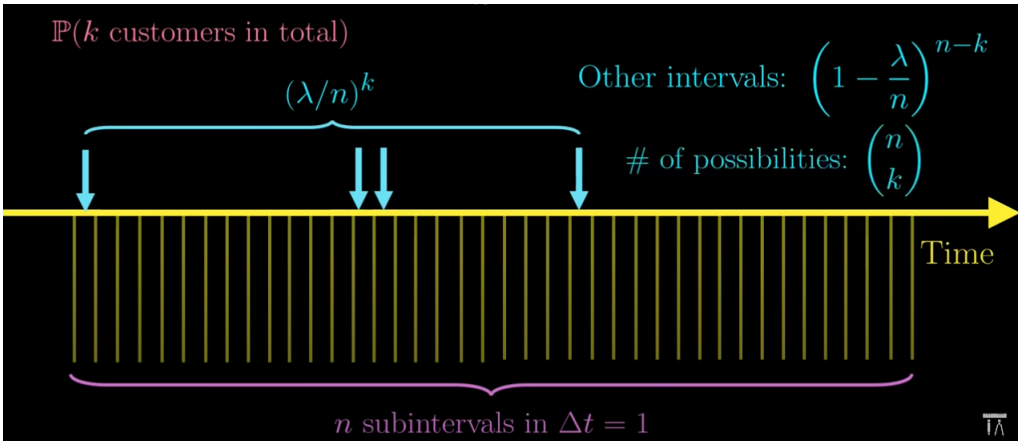

Figure 2: k customers arrival

When $n \to \infty$, after some magic math, we have:

$$P[\text{k customers, with expectation as } \lambda] = \frac{e^{-\lambda}\lambda^k}{k!} \tag{10}$$

This will become a distribution about the number of the customers, which is called **Poisson distribution**. Similarly, we define the service rate (or departure rate) as μ, also follow the Poisson distribution. The whole process is called Poisson process, with numerous states changed by customers' arrival and departure.

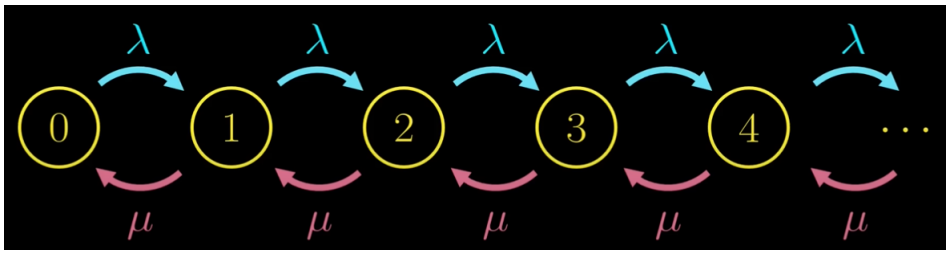

Figure 3: Poisson Process

Now we are interested in the evolution of the probability of each state. Suppose the initial time as t, and we want to know the probability of state 2 after one time step $\frac{1}{n}$ $(p_2(t+\frac{1}{n}))$:

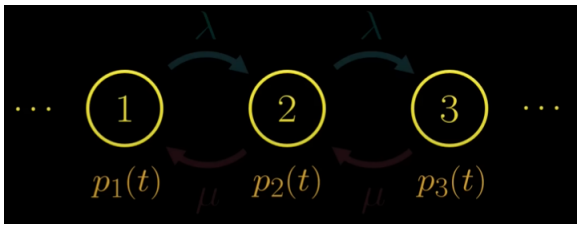

Figure 4: State Evolution Probability

The first possible case is that at time t, it was already at state 2, which means no arrival from state 1 and no departure from state 3. Therefore, the expression is:

$$p_2(t)(1-\frac{\lambda}{n})(1-\frac{\mu}{n}) \tag{11}$$

Another possible case is that at time t, it was at state 3, having departure from state 3, so the expression is:

$$p_3(t)(\frac{\mu}{n}) \tag{12}$$

Another possible case is that at time t, it was at state 1, having arrival from state 1, so the expression is:

$$p_1(t)(\frac{\lambda}{n}) \tag{13}$$

So the final expression will be:

$$p_2(t+\frac{1}{n}) = p_2(t)(1-\frac{\lambda}{n})(1-\frac{\mu}{n}) + p_3(t)(\frac{\mu}{n}) + p_1(t)(\frac{\lambda}{n}) \tag{14}$$

Similarly, we can get the general case expression:

$$p_k(t+\frac{1}{n}) = p_k(t)(1-\frac{\lambda}{n})(1-\frac{\mu}{n}) + p_{k+1}(t)(\frac{\mu}{n}) + p_{k-1}(t)(\frac{\lambda}{n}) \tag{15}$$

Rearrange this equation, we have:

$$\frac{p_k\left(t+\frac{1}{n}\right) - p_k(t)}{\frac{1}{n}} = \left(-\lambda-\mu+\frac{\lambda\mu}{n}\right)p_k(t) + \mu p_{k+1}(t) + \lambda p_{k-1}(t) \tag{16}$$

When $n \to \infty$, the LHS of the equation is approximate with the derivative of p_k, and the final expression becomes:

$$p_k'(t) = -(\lambda+\mu)p_k(t) + \mu p_{k+1}(t) + \lambda p_{k-1}(t) \tag{17}$$

But notice that this expression only holds for the situation when $k \neq 0$. When $k = 0$, the states are shown below:

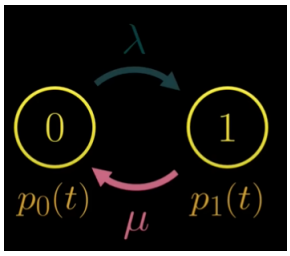

Figure 5: State 0

One possible case is that it was already at state 0 at time t, without arrival to state 1. And the other possible case is that it was at state 1, with departure. Therefore the final expression is:

$$p_0(t+\frac{1}{n}) = p_0(t)(1-\frac{\lambda}{n}) + p_1(t)(\frac{\mu}{n}) \tag{18}$$

Rearrange the equation, we have:

$$\frac{p_0\left(t+\frac{1}{n}\right) - p_0(t)}{\frac{1}{n}} = -\lambda p_0(t) + \mu p_1(t) \tag{19}$$

Similarly, when $n \to \infty$, we have:

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \tag{20}$$

4.2 Stable State Calculation

Assume that the system becomes stable in long term, then we have:

$$p_k(t) \to \pi_k \,\text{as}\, t \to \infty \tag{21}$$

And assume $p_k'(t), p_0'(t) = 0$ in long run, then the previous expressions will become:

$$0 = -(\lambda+\mu)\pi_k + \mu\pi_{k+1} + \lambda\pi_{k-1} \tag{22}$$

$$0 = -\lambda\pi_0 + \mu\pi_1 \tag{23}$$

Rearrange the $k \neq 0$ case expression, we have:

$$\pi_{k+1} - \pi_k = \frac{\lambda}{\mu}(\pi_k - \pi_{k-1}) \tag{24}$$

Which means that the differences are in scaling law, and the final probability could be expressed as:

$$\pi_k = \pi_0 + (\pi_1 - \pi_0)\left(1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2 + \cdots + \left(\frac{\lambda}{\mu}\right)^{k-1}\right) \tag{25}$$

Now take a look at the $k = 0$ case, we have:

$$\pi_1 = (\frac{\lambda}{n})\pi_0, \;\; \pi_1 - \pi_0 = (\frac{\lambda}{\mu} - 1)\pi_0 \tag{26}$$

So we update the expression:

$$\pi_k = \pi_0 + \pi_0(\frac{\lambda}{\mu} - 1)\left(1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2 + \cdots + \left(\frac{\lambda}{\mu}\right)^{k-1}\right) \tag{27}$$

After some magic math, we finally have:

$$\pi_k = \pi_0(\frac{\lambda}{\mu})^k \tag{28}$$

Based on the definition of the probability, we know:

$$\pi_0 + \pi_1 + \cdots = 1 \tag{29}$$

Also from some magic math, we get:

$$\frac{\pi_0}{1 - (\lambda/\mu)} = 1 \tag{30}$$

$$\pi_0 = 1 - \frac{\lambda}{\mu} \tag{31}$$

This only works when the expression converges, or in other words:

$$\frac{\lambda}{\mu} < 1 \tag{32}$$

This also makes sense by intuition. If the arrival rate is always faster than the departure rate, the queue will never end, so the system is not stabilized. We call the stabilized case as **invariant distribution**.

Another bizarre property this model has is that, whether we start our service at any time, there is no impact on the remaining service time. In other words, the service does not care about the history, always stay $\frac{\mu}{n}$. Therefore, we say the departure process is **memoryless**. Similarly, we say that the arrival process is also **memoryless**.

4.3 Laws Calculation

Recall that utilization measures the fraction of time the servers are actively serving customers. In math, it is defined as:

$$\rho = \frac{d}{c} = \frac{E[D]}{c} \tag{33}$$

And based on the definition, π_0 is the idle probability, which is the probability that there are zero customers in the system. If we define another probability π_{busy}, which

occurs when there is at least one entity in the system, then we have the following relation:

$$\pi_0 + \pi_{busy} = 1 \tag{34}$$

Based on the definition of utilization, we have:

$$\rho = \pi_{busy} = 1 - \pi_0 = \frac{\lambda}{\mu} \tag{35}$$

Therefore, the expected number of customers in the system will be:

$$E(F) = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} i\pi_0\rho^i = \pi_0\rho\sum_{i=0}^{\infty} i\rho^{i-1} \tag{36}$$

Now construct a derivative:

$$E(F) = \pi_0\rho\frac{d}{d\rho}\sum_{i=0}^{\infty}\rho^i = \pi_0\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \tag{37}$$

The throughput of M/M/1 system is:

$$E(D) = \lambda \tag{38}$$

By Little's Law, the average time in the system is:

$$E(Y) = \frac{E(F)}{\lambda} = \frac{1}{\mu-\lambda} \tag{39}$$

And the average wait time and the number of customers:

$$E(W) = E(Y) - E(B) = \frac{1}{\mu-\lambda} - \frac{1}{\mu} \tag{40}$$

$$E(F_Q) = \lambda E(W) = E(F) - \rho \tag{41}$$

5 M/M/m Probability Calculation

The M/M/m queueing model is an extension of the M/M/1 model, where there are multiple servers (denoted by m) providing service to arriving entities. This model is used to analyze systems with more than one server.

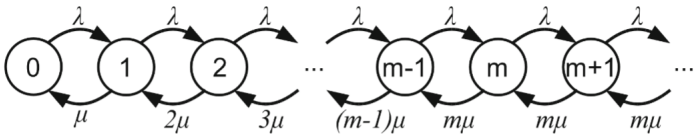

Figure 6: M/M/m system

Several important variables are shown below. The utilization is:

$$\rho = \frac{\lambda}{m\mu} \tag{42}$$

The state probability is:

$$\pi_i = \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i \pi_0, \quad 0 \le i < m \tag{43}$$

$$\pi_i = \left(\frac{\lambda}{m\mu}\right)^{i-m} \pi_m = \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \left(\frac{\lambda}{m\mu}\right)^{i-m} \pi_0, \quad m \le i \tag{44}$$

The idle probability is:

$$\pi_0 = \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1} \tag{45}$$

Expected number of customers in the system:

$$E[F] = m\rho + \frac{\rho(m\rho)^m}{m!(1-\rho)^2}\pi_0 \tag{46}$$

Number of customers in the queue:

$$E[F_Q] = \frac{\rho(m\rho)^m}{m!(1-\rho)^2}\pi_0 \tag{47}$$

If $m \to \infty$, then:

$$\pi_0 = e^{-\lambda/\mu} \tag{48}$$

$$E(F) = \frac{\lambda}{\mu} \tag{49}$$

$$E(Y) = \frac{1}{\mu} \tag{50}$$