Queueing Model M/M/s

Using Birth & Death process

Ishita Gupta, Rohith Krishna,
Shravanth J, Karnam Yogesh.

Madras School of Economics

October 21, 2020

Introduction to the M/M/s queueing model

- Customers are generated over time by an input process. They arrive at the queue in a **Poisson process**.
- The time between consecutive arrivals is called **interarrival time**.
- The time elapsed between commencement of service to a customer to its completion is called **service time**.
- The interarrival times and service times are **exponentially distributed.**
- There are a finite number of servers in the queue.
- The **queue discipline** is the order in which customers are selected for service, usually assumed to be *first-come-first-served*.
- The queue has an infinite **system capacity** and can accommodate *infinite* customers.
- The population from which arrivals occur are called **calling population size**, which is assumed to *infinite*.

M/M/s queueing model

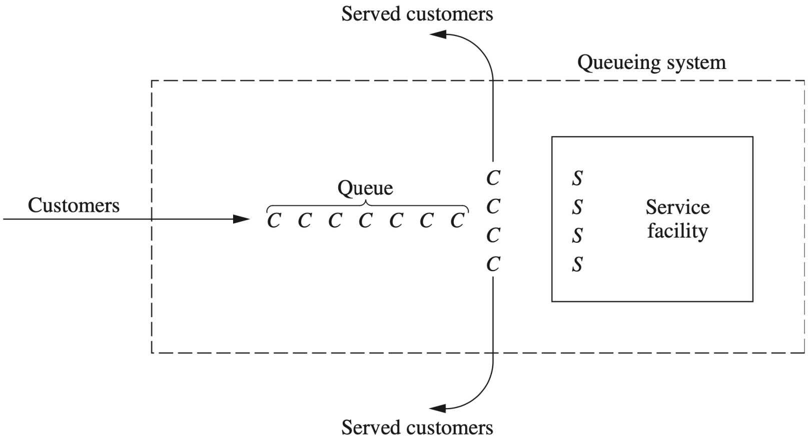

Figure: M/M/s queueing model.

Notations and Terminology

- State of system = number of customers in queueing system
- Queue length = number of customers waiting for service to begin = state of system minus number of customers being served
- $N(t)$ = number of customers in queueing system at time t
- $P_n(t)$ = probability of exactly n customers in the system at time t, given customers at time 0
- s = number of servers (parallel service channels) in the system
- λ_n = mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in the system
- μ_n = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in the system
- $\rho = \lambda / s\mu$ = utilisation factor for the service facility or the traffic intensity factor

Notations and Terminology

- The system is under **steady-state condition** (meaning, state is independent of the initial state and time elapsed).
- Thus, $P_n =$ probability of exactly n customers in the system.
- $L =$ expected number of customers in the queueing system.
- $L_q =$ expected queue length (excluding the customers currently being served).
- $w =$ waiting time in system (including service time) for each individual customer
- $W = E(w) =$ average waiting time in system.
- $w_q =$ waiting time in system (excludes service time) for each individual customer
- $W_q = E(w_q) =$ average waiting time in the queue (excluding the service time).

Little's formula

- Assume that λ_n is a constant for all n. Under steady state queueing process, it can be proved that:

$$\boxed{L = \lambda W} \qquad \boxed{L_q = \lambda W_q}$$

- If the λ_n are not equal, then it can be replaced by long run average arrival rate $\bar{\lambda}$.
- If μ is a constant, then the mean service time is a constant $1/\mu$ for all $n \geq 1$. Then,

$$\boxed{W = W_q + \frac{1}{\mu}}$$

- **Balk**: customers *refusing to enter* the system because the queue is too long.
- **Renege**: customers leaving the system *without being served*.
- We assume that balking and reneging *do not occur* in the system.

Birth & Death Process

- **Birth** - *arrival* of a new customer into the queueing system.
- **Death** - *departure* of a served customer from the system.
- Birth and Death process describes *probabilistically* how $N(t)$ changes as t increases.
- Individual births and deaths occur randomly, but mean occurrence rates depend only upon the current state of the system.
- Assumptions:
  - Given $N(t) = n$, the current probability distribution of *interarrival times* (remaining time until next birth) and *service completion times* (remaining time until next death) are exponentially distributed.
  - Interarrival: M with parameter λ_n.
  - Service: M with parameter μ_n.
  - Interarrival times and service completion times are mutually independent.

Birth & Death Process

- In queueing systems with n customers, λ_n denotes *mean arrival rate* and μ_n denotes *mean service completion rate*.
- When $n \to n+1$ then single birth occurs (arrival).
- When $n \to n-1$ then single death occurs (departure).
- If λ_n depends on n - arriving customers likely to balk.
- If μ_n depends on n - waiting customers likely to renege.

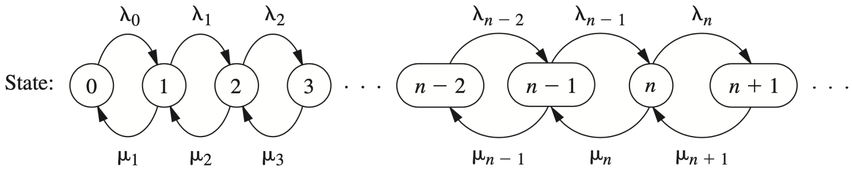

Figure: Rate diagram for the Birth & Death process

Rate In = Rate Out Principle

Let the system be in state n ($n = 0, 1, 2, \ldots$). We count the number of times that the process (queue) enters and the number of times it leaves the state:

$$E_n(t) = \text{number of times that process enters state } n \text{ by time } t$$

$$L_n(t) = \text{number of times that process leaves state } n \text{ by time } t$$

The system can go in state n only from state $n-1$ (arriving of a customer) or from state $n+1$ (departure of a customer). Hence, the two types of events (entering and leaving) alternate. These must always either be equal or differ by just 1:

$$|E_n(t) - L_n(t)| \leq 1 \xrightarrow[t \to \infty]{divide \;\; by \;\; t} \left| \frac{E_n(t)}{t} - \frac{L_n(t)}{t} \right| \leq \frac{1}{t}$$

$$\lim_{t \to \infty} \left| \frac{E_n(t)}{t} - \frac{L_n(t)}{t} \right| \leq 0$$

Rate In = Rate Out Principle

Dividing by t and let $t \to \infty$ gives us the mean rate of events per unit time.

$$\lim_{t\to\infty} \frac{E_n(t)}{t} = \text{mean rate at which process enters state } n$$

$$\lim_{t\to\infty} \frac{L_n(t)}{t} = \text{mean rate at which process leaves state } n$$

Thus,

$$\boxed{\textit{mean entering rate} = \textit{mean leaving rate}}$$

which is the Rate In = Rate Out Principle. This equation is called the **balance equation** for state n.

Balance Equations

State	Rate In = Rate Out
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$
3	$\lambda_2 P_2 + \mu_4 P_4 = (\lambda_3 + \mu_3) P_3$
$\vdots$	$\vdots$
n - 1	$\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$
$\vdots$	$\vdots$

Recursion Relations

State 0:

$$P_1 = \frac{\lambda_0}{\mu_1} P_0$$

State 1:

$$\begin{aligned} P_2 &= \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) \\ P_2 &= \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0 \end{aligned}$$

State 2:

$$\begin{aligned} P_3 &= \frac{\lambda_2}{\mu_3} P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) \\ P_3 &= \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0 \end{aligned}$$

Recursion Relations

State $n-1$:

$$
\begin{aligned}
P_n &= \frac{\lambda_{n-1}}{\mu_n} P_{n-1} + \frac{1}{\mu_n}(\mu_{n-1} P_{n-1} - \lambda_{n-2} P_{n-2}) \\
P_n &= \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n \mu_{n-1}\ldots\mu_1} P_0
\end{aligned}
$$

State n:

$$
\begin{aligned}
P_{n+1} &= \frac{\lambda_n}{\mu_{n+1}} P_n + \frac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1} P_{n-1}) \\
P_{n+1} &= \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_n \lambda_{n-1}\ldots\lambda_0}{\mu_{n+1}\mu_n\ldots\mu_1} P_0
\end{aligned}
$$

Steady-state probabilities

Let $C_n = \dfrac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_1}$ for $n = 1, 2, \ldots$ and

Define $C_n = 1$ for $n = 0$.

The steady-state probabilities are

$$\boxed{P_n = C_n P_0} \quad \text{for } n = 1,2, \ldots$$

We require that

$$\sum_{n=0}^{\infty} P_n = 1 \implies \left(\sum_{n=0}^{\infty} C_n\right) P_0 = 1$$

$$\boxed{P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1}}$$

Cost Equations

The probability that there are n customers in the queueing system P_n is computed using the recursion relation for the birth and death process. We obtain our cost equations using:

$$L = \sum_{n=0}^{\infty} nP_n \qquad L_q = \sum_{n=s}^{\infty} (n-s)P_n$$

$$W = \frac{L}{\bar{\lambda}} \qquad W_q = \frac{L_q}{\bar{\lambda}}$$

where, $\bar{\lambda}$ is the *long run average arrival rate*. The mean arrival rate of the system in state n is λ_n is associated with a probability P_n. Thus the long run average arrival rate is the weighted average over all n.

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$

Assumptions of M/M/s model

- $M/M/s$ queueing model is a special case of the birth & death process.
- Interarrival and service times are independent and identically distributed according to an exponential distribution (M).
- Arrival of customers is a Poisson process.
- The queueing system's *mean arrival rate* and *mean service rate* per busy server are constant (λ and μ respectively) regardless of the state of the system.
- System has finite multiple servers $s > 1$ with service rates:

  $\mu_n = n\mu$, when $n \leq s$
  $\mu_n = s\mu$, when $n \geq s$,
- When $s\mu$ exceeds the mean arrival rate λ, queueing system eventually reaches steady-state,

$$\boxed{\rho = \frac{\lambda}{s\mu} < 1} \qquad \rho - \text{utilization factor}$$

M/M/s as a Birth & Death process

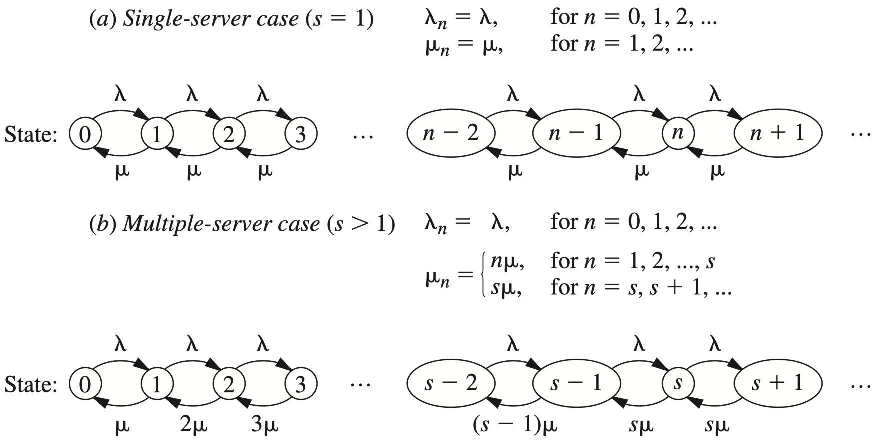

Figure: M/M/s as a Birth & Death process

Results for multiple-server case ($s > 1$)

As the λ's are equal, we find the value of C_n reduces to:

$C_0 = 1$ for $n = 0$ case

$$C_n = \begin{cases} \frac{\lambda}{\mu}\frac{\lambda}{2\mu}\frac{\lambda}{3\mu}\cdots\frac{\lambda}{n\mu}, & \text{for } n = 1, 2, \ldots, s \\ \frac{\lambda}{\mu}\frac{\lambda}{2\mu}\frac{\lambda}{3\mu}\cdots\frac{\lambda}{s\mu}\frac{\lambda}{s\mu}\frac{\lambda}{s\mu}\cdots\frac{\lambda}{s\mu}, & \text{for } n = s, s+1, \ldots \end{cases}$$

$$C_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!}, & \text{for } n = 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s} = \dfrac{(\lambda/\mu)^n}{s!s^{n-s}}, & \text{for } n = s, s+1, \ldots \end{cases}$$

Results for the multiple server case

We substitute the values for C_n in the expression $P_0 = 1/\left(\sum_{n=0}^{\infty} C_n\right)$

$$P_0 = 1 \Bigg/ \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{\infty} \left(\frac{\lambda}{s\mu}\right)^{n-s}\right]$$

The summation in the last term is a geometric series,

$$P_0 = 1 \Bigg/ \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\lambda/(s\mu)}\right]$$

These C_n factors also give,

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0, & \text{if } 0 \leq n \leq s \\ \dfrac{(\lambda/\mu)^n}{s!s^{n-s}} P_0 & \text{if } n \geq s \end{cases}$$

Average number of customers in the queue, L_q

$$
\begin{aligned}
L_q &= \sum_{n=s}^{\infty}(n-s)P_n = \sum_{j=0}^{\infty} jP_{s+j}, \quad j = n-s \text{ is the dummy index} \\
L_q &= \sum_{j=0}^{\infty} j\frac{(\lambda/\mu)^s}{s!}\rho^j P_0 = P_0\frac{(\lambda/\mu)^s}{s!}\rho\sum_{j=0}^{\infty}\frac{d}{d\rho}(\rho^j) \\
L_q &= P_0\frac{(\lambda/\mu)^s}{s!}\rho\frac{d}{d\rho}\left(\sum_{j=0}^{\infty}\rho^j\right) = P_0\frac{(\lambda/\mu)^s}{s!}\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) \\
L_q &= \frac{P_0(\lambda/\mu)^s\rho}{s!(1-\rho)^2}
\end{aligned}
$$

Results for W_q, W & L

Average waiting time in the queue W_q,

$$W_q = \frac{L_q}{\lambda}$$

Average time spent in the system W,

$$W = \text{Avg. time in the queue} + \text{ Avg. service time}$$

$$W = W_q + \frac{1}{\mu}$$

Average number of customers in the system L,

$$L = \lambda \left(W_q + \frac{1}{\mu} \right) = L_q + \frac{\lambda}{\mu}$$

Problem: County Hospital emergency room

- The County Hospital has an emergency room with one doctor always on duty. Due to the increase in emergency patient arrivals, the hospital is considering hiring a second doctor. The hospital's management models this problem as an M/M/s queueing model and gathers relevant information.
- They find that patients arrive on an average of 2 every hour. A doctor requires 20 minutes to treat a patient. Thus $\lambda = 2$ patients per hour and $\mu = 3$ patients per hour.
- If 1 hour is the unit of time, $1/\lambda = 1/2$ hours per patient and $1/\mu = 1/3$ patients per hour.
- The alternatives considered are whether to hire an extra doctor ($s = 2$) or not ($s = 1$).
- In both cases the utilization factor $\rho < 1$ so that system approaches steady-state condition.

Problem Solution

	s = 1	s = 2
ρ	2/3	1/3
P_0	1/3	1/2
P_1	2/9	1/3
P_n for $n \geq 2$	$\frac{1}{3}.\frac{2}{3}^n$	$\frac{1}{3}^n$
L_q	4/3	1/12
L	2	3/4
W_q	2/3 hour	1/24 hour
W	1 hour	3/8 hour

Efficiency of the hospital increases with an additional doctor.

References

- Hillier, Lieberman. Introduction to Operations Research. 9th Edition.
- Sheldon Ross. Introduction to Probability Models. 10th Edition.

Thank you!