ACI650 - Modelos y Simulación

Queueing Processes

Mario González
Facultad de Ingeniería y Ciencias Ambientales
Centro de Investigación, Estudios y Desarrollo de Ingeniería (CIEDI)

May 15, 2017

Learning Objectives

- Employ modeling techniques for Queueing systems governed by the exponential process.
- Simulation examples.
- Hands-on workshop.

Basic definitions and notation I

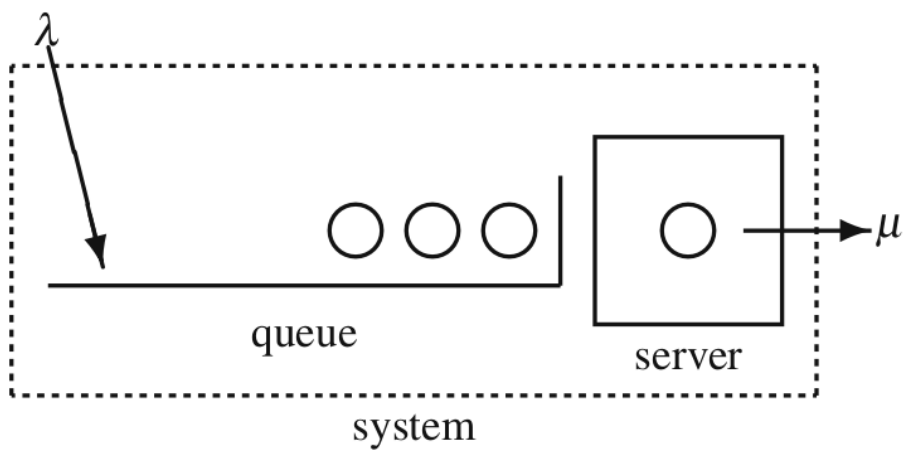
- Many phenomena for which mathematical descriptions are desired involve waiting lines either of people or material.
- A queue is a waiting line, and queueing processes are those stochastic processes arising from waiting line phenomena.
- For example:
  - banks/supermarkets - waiting for service,
  - the utilization of data processing services at a computer center,
  - failure situations - waiting for a failure to occur e.g. in a piece of machinery,
  - public transport - waiting for a train or a bus.

Basic definitions and notation II

- Queues are a common every-day experience.
- Queues form because resources are limited.
- It makes economic sense to have queues.
- For example:
  - How many supermarket cashiers would be needed to avoid queuing?
  - How many buses or trains would be needed if queues were to be avoided/eliminated?

Basic definitions and notation III

- A queueing process involves the arrival of customers to a service facility and the servicing of those customers.
- All customers that have arrived but are not yet being served are said to be in the queue.
- The queueing system includes all customers in the queue and all customers in service.
- The figure bellow represents a queueing system with a mean arrival rate of λ, and mean service rate of μ, four customers in the system, and three in the queue.



Basic definitions and notation IV

- The general notation of a queueing system has the following form (Kendall's notation):

$$\left(\begin{array}{c}\text{arrival}\\ \text{process}\end{array} \Big/ \begin{array}{c}\text{service}\\ \text{process}\end{array} \Big/ \begin{array}{c}\text{number}\\ \text{of servers}\end{array} \Big/ \begin{array}{c}\text{maximum}\\ \text{possible}\\ \text{in system}\end{array} \Big/ \begin{array}{c}\text{queue}\\ \text{discipline}\end{array}\right)$$

- Queueing symbols used with Kendall's notation:

Symbols	Explanation
M	Exponential inter-arrival or service time
D	Deterministic inter-arrival or service time
E_k	Erlang type k inter-arrival or service time
G	General inter-arrival or service time
$1, 2, \ldots \infty$	Number of parallel servers or capacity
FIFO	First in, first out queue discipline
LIFO	Last in, first out queue discipline
SIRO	Service in random order
PRI	Priority queue discipline
GD	General queue discipline

Basic definitions and notation V

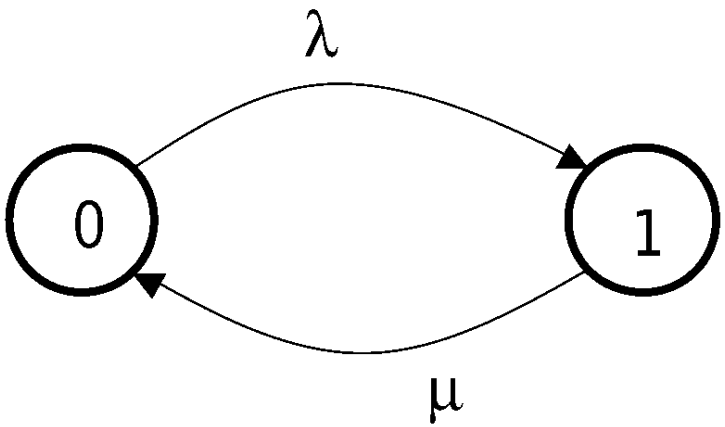
- The purpose in this presentation is to give an introduction to queueing processes.
- To maintain the introductory level of this material, arrival processes to the queueing systems will be assumed to be Poisson processes, and service times will be exponential.

Markov Process

- A **Markov Process** is a stochastic process satisfies the **Markov property**.
- A stochastic process has the Markov property if the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the sequence of events that preceded it.
- A Markov process has no memory (memoryless process).
- If the state space is countable, the process is called a Markov chain.
- If the time is continuous, the process is called Markov chain in continuous time.

Markov chains I

- A Markov chain is usually shown by a state transition diagram, or by a State Transition Matrix denoted by Q and known as the Infinitesimal generator (**G**) of a stochastic process.



$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

Markov chains II

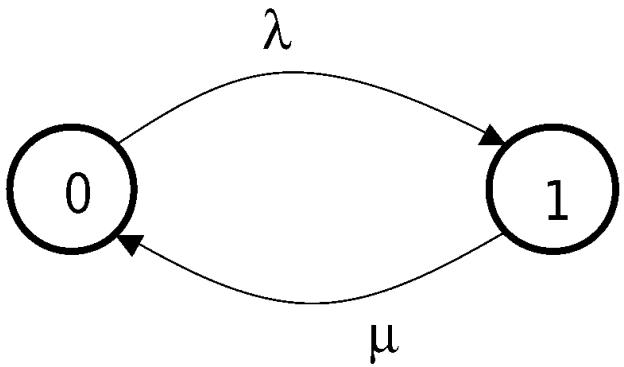
- When describing a Markov process we obtain the generator matrix.
- The usual technique is to obtain the off-diagonal elements first and then let the diagonal element equal the negative of the sum of the off-diagonal elements on the row.
- The generator is used to obtain the steady-state probabilities directly as follows.

  Let $Y = Y_t; t \geq 0$ be a Markov process with an irreducible, recurrent state space E and with generator matrix **G**. Furthermore, let **p** be a vector of steady-state probabilities. Then **p** is the solution to

$$pG = 0$$

$$\sum_{j \in E} p(j) = 1.$$

Markov chains III



$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

- Solving

$$\pi G = 0$$

$$\sum_j \pi(j) = 1$$

- We get the steady-state probabilities π:

$$\pi = \left(\frac{\mu}{\lambda + \mu}, \frac{\lambda}{\lambda + \mu} \right)$$

Single Server Systems

- The simplest queueing systems to analyze are those involving a Poisson arrival process and a single exponential server.
- We will start by considering a system that has unlimited space for arriving customers.

Infinite Capacity Single-Server Systems I

- We begin with an M/M/1 system, or equivalently, an M/M/1∞/FIFO system.
- The M/M/1 system assumes customers arrive according to a Poisson process with mean rate λ and are served by a single server whose time for service is random with an exponential distribution of mean $1/\mu$.
- If the server is idle and a customer arrives, then that customer enters the server immediately.
- If the server is busy and a customer arrives, then the arriving customer enters the queue which has infinite capacity. When service for a customer is completed, the customer leaves and the customer that had been in the queue the longest instantaneously enters the service facility and service begins again.

Infinite Capacity Single-Server Systems II

- The flow of customers through the system is a Markov process with state space $\{0, 1, \cdots\}$.
- The Markov process is denoted by $\{N_t;\ t \geq 0\}$ where N_t denotes the number of customers in the system at time t.
- The steady-state probabilities are $p_n = \lim_{t\to\infty} Pr\{N_t = n\}$.
- We let N be a random variable with probability mass function $\{p_0, p_1, \cdots\}$.
- The random variable N thus represents the number of customers in the system at steady-state, and p_n represents the long-run probability that there are n customers in the system.
- Another way to view p_n is as the long-run fraction of time that the system contains n customers.

Infinite Capacity Single-Server Systems III

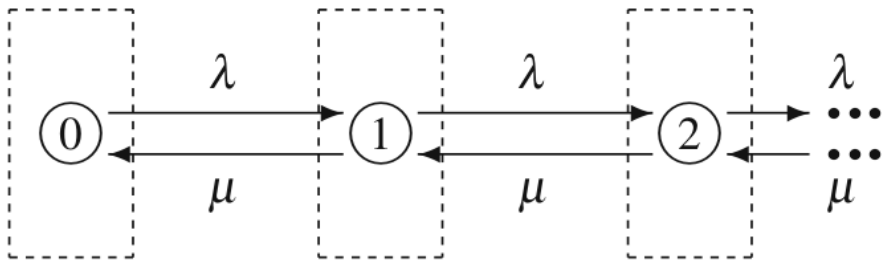
- Sometimes we will be interested in the number of customers that are in the queue and thus waiting for service.
- Let the random variable N_q denote the steady-state number in the queue.
- If the system is idle, $N_q = N$.
- When the system is busy, $N_q = N - 1$.

Infinite Capacity Single-Server Systems IV

- Our immediate goal is to derive an expression for $p_n, n = 0, 1, \cdots$, in terms of the mean arrival and service rates.
- This derivation usually involves two steps:
  1. Obtain a system of equations defining the probabilities and
  2. Solve the system of equations.
- Step (1) is relatively easy. It is usually Step (2) that is difficult. In other words, the system of equations defining p_n is not hard to obtain, but it is sometimes hard to solve.

Infinite Capacity Single-Server Systems V

- An intuitive approach for obtaining the system of equations is to draw a state diagram and then use a rate balance approach as shown in the following figure:



- The rate into the box around Node 0 is μp_1, the rate out of the box around Node 0 is λp_0, thus, “rate in” = “rate out” yields

$$\mu p_1 = \lambda p_0.$$

- The rate into the box around Node 1 is $\lambda p_0 + \mu p_2$, the rate out of the box around Node 1 is $(\mu + \lambda)p_1$, thus

$$\lambda p_0 + \mu p_2 = (\mu + \lambda)p_1.$$

Infinite Capacity Single-Server Systems VI

- Continuing in a similar fashion and rearranging, we obtain the system

$$p_1 = \frac{\lambda}{\mu} p_0 \text{ and}$$

$$p_{n+1} = \frac{\lambda + \mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1} \text{ for } n = 1, 2, \cdots$$

Infinite Capacity Single-Server Systems VII

- A more rigorous approach for obtaining the system of Equations is to first get the generator matrix for the M/M/1 system.
- Since the inter-arrival and service times are exponential, the queueing system is a Markov process.
- The rate at which the process goes from State n to State $n+1$ is λ, and the rate of going from State n to State $n-1$ is μ, thus, the generator is the infinite dimensioned matrix given as

$$\mathbf{G} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \mu & -(\mu+\lambda) & \lambda & \\ & & \ddots & \ddots & \ddots \\ & & & & \end{bmatrix}$$

- The system of equations formed by $\boldsymbol{pG} = 0$ then yields the equations presented before.

Infinite Capacity Single-Server Systems VIII

- The system of equations can be solved by successively forward substituting solutions and expressing all variables in terms of p_0.
- Since we already have $p_1 = (\lambda/\mu)p_0$, we look at p_2 and then p_3:

$$\begin{aligned} p_2 &= \frac{\lambda+\mu}{\mu}p_1 - \frac{\lambda}{\mu}p_0 \\ &= \frac{\lambda+\mu}{\mu}(\frac{\lambda}{\mu}p_0) - \frac{\lambda}{\mu}\frac{\mu}{\mu}p_0 = \frac{\lambda^2}{\mu^2}p_0\,, \end{aligned}$$

$$\begin{aligned} p_3 &= \frac{\lambda+\mu}{\mu}p_2 - \frac{\lambda}{\mu}p_1 \\ &= \frac{\lambda+\mu}{\mu}(\frac{\lambda^2}{\mu^2}p_0) - \frac{\lambda}{\mu}\frac{\mu}{\mu}(\frac{\lambda}{\mu}p_0) = \frac{\lambda^3}{\mu^3}p_0\,. \end{aligned}$$

- A pattern emerges and we can assert that:

$$p_n = \frac{\lambda^n}{\mu^n}p_0 \text{ for } n \geq 0.$$

Infinite Capacity Single-Server Systems IX

- The ratio λ/μ is called the traffic intensity for the queueing system and is denoted by ρ for the M/M/1 system.
- More generally, ρ is usually defined as the arrival rate divided by the maximum system service rate.
- Now we have p_n for all n in terms of p_0 so the long-run probabilities become known as soon as p_0 can be obtained.
- An expression for p_0 can be determined by using the norming equation, namely

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} \frac{\lambda^n}{\mu^n} = p_0 \sum_{n=0}^{\infty} \rho^n$$

$$= \frac{p_0}{1-\rho}.$$

Infinite Capacity Single-Server Systems X

- The equality in the above expression made use of the geometric progression so it is only valid for $\rho < 1$. If $\rho \geq 1$, the average number of customers and time spent in the system increase without bound and the system becomes unstable.
- One might be tempted to design a system such that the service rate is equal to the arrival rate, thus creating a "balanced" system. This is false logic for random interarrival or service times since in that case the system will never reach steady-state.
- The value derived above for p_0 can be expressed for the M/M/1 system as follows:

$$p_n = (1 - \rho)\rho^n \text{ for } n = 0, 1, \cdots,$$

where $\rho = \lambda/\mu$ and $\rho < 1$.

Infinite Capacity Single-Server Systems XI

- Let's review the former steps:

  1. Form the Markov generator matrix, **G**.
  2. Obtain a system of equations by solving $\boldsymbol{pG} = 0$.
  3. Solve the system of equations in terms of p_0 by successive forward substitution and induction if possible.
  4. Use the norming equation to find p_0.

- The most difficult step is the third step. It is not always possible to find a closed-form solution to the system of equations. and often techniques other than successive forward substitution must be used.

Infinite Capacity Single-Server Systems XII

- **Example.** An operator of a small grain elevator has a single unloading dock. Arrivals of trucks during the busy season form a Poisson process with a mean arrival rate of four per hour. Because of varying loads (and desire of the drivers to talk) the length of time each truck spends in front of the unloading dock is approximated by an exponential random variable with a mean time of 14 minutes. Assuming that the parking spaces are unlimited, the M/M/1 queueing system describes the waiting line that forms. Calculate:
  - The probability of the unloading dock being idle.
  - The probability that there are exactly three trucks waiting.
  - The probability that four or more trucks are in the system.

Performance measures I

- Expected number of customers in the system, denoted by L, and the expected number in the queue, denoted by L_q.

$$L = E[N] = \frac{\rho}{1-\rho}, \quad L_q = E[N_q] = \frac{\rho^2}{1-\rho}.$$

- Variance of the number in the system and queue

$$V[N] = \frac{\rho}{(1-\rho)^2}, \quad V[N_q] = \frac{\rho^2(1+\rho-\rho^2)}{(1-\rho)^2}.$$

Performance measures II

- Waiting times are another important measure of a queueing system.
- **Little's Law.** Consider a queueing system for which steady-state occurs. Let $L = E[N]$ denote the mean long-run number in the system, $W = E[T]$ denote the mean long-run waiting time within the system, and λ_e the mean arrival rate of jobs into the system. Also let $L_q = E[N_q]$ and $W_q = E[T_q]$ denote the analogous quantities restricted to the queue. Then

$$L = \lambda_e W, \quad L_q = \lambda_e W_q$$

- Here, λ_e refers to the effective mean arrival rate into the system; whereas, λ refers to the mean arrival rate to the system. λ includes those customers who come to the system but for some reason, like a finite capacity system that is full, they do not enter; λ_e only counts the customers who make it to the server.

Performance measures III

- For a M/M/1 system, the effective arrival rate is the same as the arrival rate (i.e., $\lambda_e = \lambda$):

$$W = E[T] = \frac{1}{\mu - \lambda},$$

$$W_q = E[T_q] = \frac{\rho}{\mu - \lambda},$$

where T is the random variable denoting the time a customer (in steady-state) spends in the system and T_q is the random variable for the time spent in the queue.

Performance measures IV

- When the arrival process is Poisson, there is a generalization of Little's formula that holds for variances.
- **Property.** Consider a queueing system for which steady-state occurs and with a Poisson arrival stream of customers entering the system. Let N denote the number in the system, T denote the customer waiting time within the system, and λ_e the mean arrival rate of jobs into the system. Also let N_q and T_q denote the analogous quantities restricted to the queue. Then the following hold:

$$V[N] - E[N] = \lambda_e^2 V[T],$$

$$V[N_q] - E[N_q] = \lambda_e^2 V[T_q].$$

- For the M/M/1 system with Poisson arrivals we obtain:

$$V[T] = \frac{1}{(\mu - \lambda)^2} = \frac{1}{\mu^2(1-\rho)^2},$$

$$V[T_q] = \frac{2\rho - \rho^2}{(\mu - \lambda)^2} = \frac{\rho(2-\rho)}{\mu^2(1-\rho)^2}.$$

Performance measures V

- Another relationship for queueing systems due to the definition of the system and the queue is that the mean waiting time in the system must equal the mean time in the queue plus the mean service time.
- **Property.** Consider a queueing system for which steady-state occurs. Let W denote the mean long-run waiting time within the system, W_q the mean long-run waiting time in the queue, and μ the mean service rate, then

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

Performance measures VI

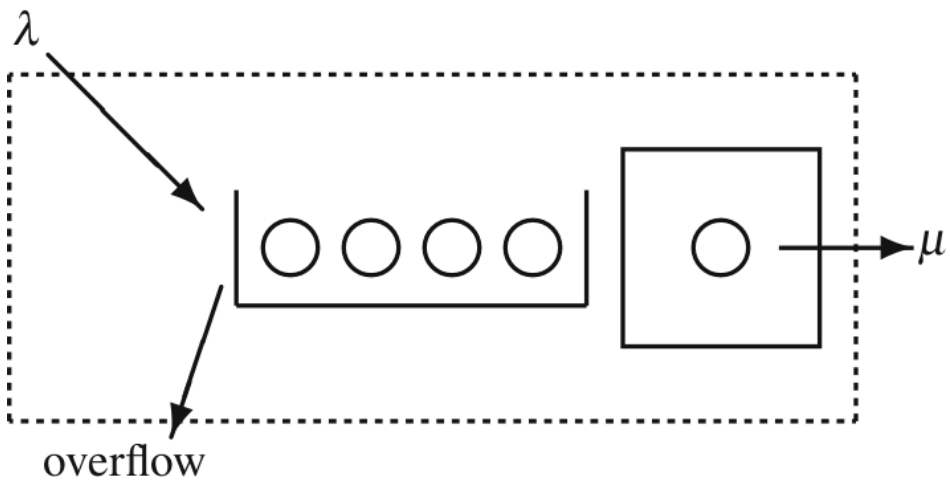
- For the example in slide 24 **calculate the performance measures** and check that the mean time each truck spends in the system is 3.5 hours with a standard deviation of 3.5 hours, and the mean time each truck waits in line until his turn at the dock is 3 hours and 16 minutes with a standard deviation of 3 hours and 29 minutes.

Simulation exercise

- Simulate a M/M/1 system and calculate the performance measures by simulation and check if it complies with the theory.

Finite Capacity Single Server Systems I

- The assumption of infinite capacity is often not suitable.
- When a finite system capacity is necessary, the state probabilities and measures of effectiveness presented before are inappropriate to use;
- thus, new probabilities and measures of effectiveness must be developed for the $M/M/1/K$ system.



Finite Capacity Single Server Systems II

- If K is the maximum number of customers possible in the system, the generator matrix is of dimension $(K+1) \times (K+1)$ and has the form

$$\mathbf{G} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \mu & -(\lambda+\mu) & \lambda \\ & & & \mu & -\mu \end{bmatrix}.$$

- Building the $\boldsymbol{pG} = 0$ systems, and using the norming equation to get p_0
- We have for an M/M/1/K system

$$p_n = \begin{cases} \rho^n \frac{1-\rho}{1-\rho^{K+1}} & \text{for } \rho \neq 1, \\ \frac{1}{K+1} & \text{for } \rho = 1\,, \end{cases}$$

Finite Capacity Single Server Systems III

- The mean for the number in the system L and queue L_q are

$$L = \begin{cases} \rho \frac{1+K\rho^{K+1}-(K+1)\rho^K}{(1-\rho)(1-\rho^{K+1})} & \text{for } \rho \neq 1 \\ \frac{K}{2} & \text{for } \rho = 1 \end{cases}$$

$$L_q = \begin{cases} L - \frac{\rho(1-\rho^K)}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \frac{K(K-1)}{2(K+1)} & \text{for } \rho = 1 \,. \end{cases}$$

- The variances for the above quantities for the $M/M/1/K$ system are

$$V[N] = \begin{cases} [\rho/(1-\rho^{K+1})(1-\rho)^2] \\ \quad \times [1+\rho-(K+1)^2\rho^K \\ \quad + (2K^2+2K-1)\rho^{K+1} - K^2\rho^{K+2}] - L^2 & \text{for } \rho \neq 1\,, \\ K(K+2)/12 & \text{for } \rho = 1\,. \end{cases}$$

$$V[N_q] = V[N] - p_0(L+L_q)$$

Finite Capacity Single Server Systems IV

- The probability that an arriving customer enters the system is the probability that the system is not full.
- Therefore, to utilize Little's formula, we set $\lambda_e = \lambda(1 - p_K)$ for the effective arrival rate to obtain the waiting time equations as follows:

$$W = \frac{L}{\lambda(1 - p_K)}$$

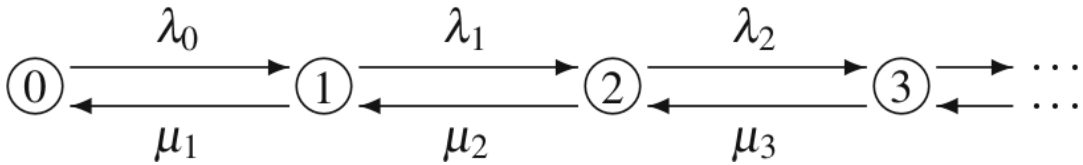
$$W = W_q + \frac{1}{\mu}.$$

Finite Capacity Single Server Systems V

- **Example**. A corporation must maintain a large fleet of tractors. They have one repairman that works on the tractors as they break down on a first-come first-serve basis. The arrival of tractors to the shop needing repair work is approximated by a Poisson distribution with a mean rate of three per week. The length of time needed for repair varies according to an exponential distribution with a mean repair time of 1/2 week per tractor. The current corporate policy is to utilize an outside repair shop whenever more than two tractors are in the company shop so that, at most, one tractor is allowed to wait. Each week that a tractor spends in the shop costs the company $100. To utilize the outside shop costs $500 per tractor. (The $500 includes lost time.) We wish to review corporate policy and determine the optimum cutoff point for the outside shop; that is, we shall determine the maximum number allowed in the company shop before sending tractors to the outside repair facility.

Multiple Server Queues I

- A birth-death process is a special type of Markov process which is applicable to many types of Markov queueing systems.
- The birth-death process is a process in which changes of state are only to adjacent states.



- The generator matrix for a general birth-death process is given by

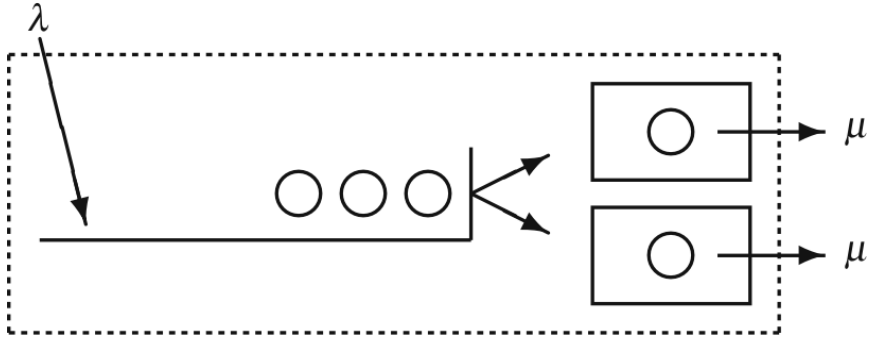
$$\mathbf{G} = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & & \\ & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

where λ_n and μ_n are the birth rate (arrival rate) and death rate (service rate), respectively, when the process is in state n.

Multiple Server Queues II

- The birth-death process can be used for many types of queueing systems.
- The M/M/1 system, arises by letting the birth rates be the constant λ and the death rates be the constant μ.
- The M/M/1/K system equations can be obtained by letting $\lambda_n = 0, \ \forall \ n > K$.

Multiple Server Queues III



- The M/M/c queueing system above is a birth-death process with the following values for the birth rates and death rates:

$$\lambda_n = \lambda \quad \text{for } n = 0, 1, \cdots,$$
$$\mu_n = \begin{cases} n\mu & \text{for } n = 1, \cdots, c-1 \\ c\mu & \text{for } n = c, c+1, \cdots. \end{cases}$$

  - The reason that $\mu_n = n\mu$ for $n = 1, \cdots, c-1$ is that when there are less than c customers in the system, each customer in the system is being served and thus the service rate would be equal to the number of customers since unoccupied servers remain idle (i.e., free servers do not help busy servers). If there are more than c customers in the system, then exactly c servers are busy and thus the service rate must be $c\mu$.

Multiple Server Queues IV

- For the M/M/c it holds,

$$p_n = \begin{cases} p_0 r^n / n! & \text{for } n = 0, 1, \cdots, c-1 \\ p_0 r^n / (c^{n-c} c!) & \text{for } n = c, c+1, \cdots, \end{cases}$$

$$p_0 = \left[\frac{c r^c}{c!(c-r)} + \sum_{n=0}^{c-1} \frac{r^n}{n!} \right]^{-1},$$

where $r = \lambda/\mu$ and $\rho = r/c < 1$.

- We have that the measures of effectiveness for the M/M/c queueing system

$$\begin{aligned} L_q &= \sum_{n=c}^{\infty} (n-c) p_n \\ &= \frac{p_0 r^c \rho}{c!(1-\rho)^2}. \end{aligned}$$

Multiple Server Queues V

- Little's formula can then be applied to obtain

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

and

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

and finally

$$L = L_q + r.$$

- The reason that we let $r = \lambda/\mu$ is because most textbooks and technical papers usually reserve ρ to be the arrival rate divided by the maximum service rate, namely, $\rho = \lambda/(c\mu)$.

Multiple Server Queues VI

- For completeness, we also give the formula needed to obtain the variances for the M/M/c system as

$$E[(N_q(N_q-1)] = \frac{2p_0 r^c \rho^2}{c!(1-\rho)^3}$$

$$E[T_q^2] = \frac{2p_0 r^c}{\mu^2 c^2 c!(1-\rho)^3}$$

$$V[T] = V[T_q] + \frac{1}{\mu^2}$$

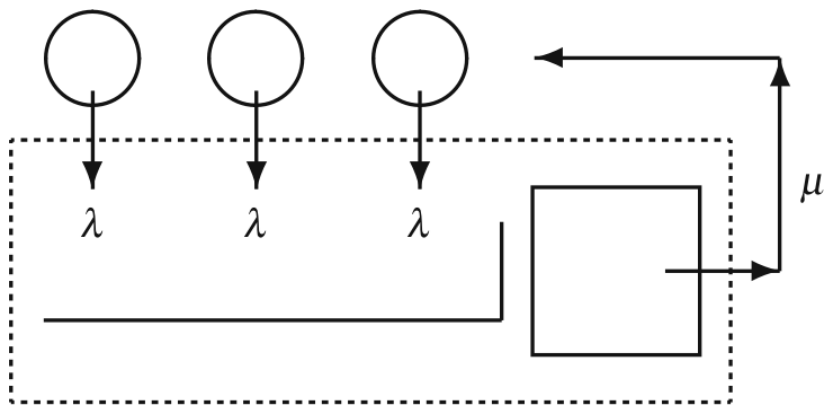
$$V[N] = \lambda^2 V[T] + L\,.$$

Multiple Server Queues VII

- **Example.** The corporation from the previous example has implemented the policy of never allowing more than three tractors in their repair shop. For $600 per week, they can hire a second repairman. Is it worthwhile to do so if the expected cost is used as the criterion? To answer this question, the old cost for the M/M/1/3 system is compared to the proposed cost for an M/M/2/3 system.

Multiple Server Queues VIII

- **Example.** A small corporation has three old machines that continually breakdown. Each machine breaks down on the average of once a week. The corporation has one repairman that takes, on the average, one half of a week to repair a machine. Assuming breakdowns and repairs are exponential random variables, write the birth-death equations for λ_n, μ_n and calculate p_n.



  - Let us further assume that for every hour that a machine is tied up in the repair shop, the corporation losses $25. Calculate the cost of this system per hour due to the unavailability of the machines.

Sources and resources

- Applied Probability and Stochastic Processes, Feldman, Richard M., Valdez-Flores, Ciriaco.
- Chapter 7 (pages 210-225), Applied Probability and Stochastic Processes, Feldman, Richard M., Valdez-Flores, Ciriaco.
- Geometric progression, Wikipedia.