CS 162 Operating Systems **Fall 2017**

Queuing Theory - Analyses of M/M/1 and M/G/1

Lecturer: Prof. Stoica Scribe: Antares Chen

1 Introduction

Given a queueing system, we have been interested in (1) the average number of items within the queue, (2) the average waiting time for each item, and (3) the average rate at which items can be processed. If we are given knowledge of two of these quantities, then the third can be easily determined using Little's Law. However, in the case where we only know the value of one of these metrics then to determine the other values we will need to make assumptions regarding how the queueing system behaves – specifically how items arrive to and are processed by the queue. In these notes, we will studying two types of queueing systems: M/M/1 systems and M/G/1 systems.

Our analysis will rely heavily on probabilistic tools, so we devote the first section towards a rigorous review of probabilistic concepts such as Poisson processes and continuous-time Markov chains. If the reader is familiar with these concepts then this section can be omitted. All relevant equations are numbered so that they can be referred to quickly when performing queuing analyses.

2 Review of Probability

Our analyses for M/M/1 and M/G/1 queuing models will depend heavily on probability. To that extent, we dedicate this section towards reviewing concepts regarding Poisson point processes and continuous-time markov chains. The content from this section is an adaptation of that presented here[1]. The reader may refer to this as a secondary source.

2.1 Poisson Process

We will model how items arrive and are serviced by the system as Poisson processes. Suppose we are interested in counting the number of times an event occurs during a time interval Δt. This counting process is *Poisson* if it abides by the following definition.

Definition 1 (Poisson Process)**.** *A counting process $\{N(t) : t \geq 0\}$ where $N(t)$ denotes the number of occurrences up to time t is Poisson with parameter $\lambda > 0$ if $N(0) = 0$ and the following conditions hold.*

1. *The probability that an event occurs between time t and $t + \Delta t$ is $\lambda \Delta t + o(\Delta t)$.*
2. *The probability that more than one event occurs between t and $t + \Delta t$ is $o(\Delta t)$.*
3. *The number of occurrences in non-overlapping time intervals are statistically independent.*

Intuitively, λ should be thought of as the rate at which an event occurs. Additionally, we use $o(\Delta t)$ to denote a function $f(\Delta t)$ that is *strictly* upper-bounded by Δt. Its use is similar to Big-O notation[2] where $f(\Delta t) \in o(\Delta t)$ if

$$\lim_{\Delta t \to 0} \frac{f(\Delta t)}{\Delta t} = 0$$

The $o(\Delta t)$ terms for a Poisson process can be thought of as negligent lower-order terms. We need not consider these terms since Δt dominates $o(\Delta t)$ asymptotically as for example in the occurence probability.

[1] `http://www.math.uchicago.edu/~may/VIGRE/VIGRE2011/REUPapers/Constantin.pdf`

[2] In fact this is called little-O notation

2.1.1 The Poisson Distribution

Given a Poisson process with parameter λ, we may be interested in determining the distribution of occurrences across a time interval. For an integer $n \geq 0$, define $p_n(t)$ to be the probability that an event occurs n times during time interval t. With additional an additional (sequential) time interval Δt, we may compute $p_n(t + \Delta t)$ in the following manner.

$$p_n(t + \Delta t) = \sum_{i=0}^{n} \Pr\{n - i \text{ occurrences during interval } t \text{ and } i \text{ during } \Delta t\}$$

By condition (3) of the definition above, we know that non-overlapping time intervals are independent. As t and Δt are non-overlapping:

$$p_n(t + \Delta t) = \sum_{i=0}^{n} \Pr\{n - i \text{ occurrences during interval } t\} \cdot \Pr\{i \text{ occurences during } \Delta t\}$$

By condition (1) we know that the probability of 1 event occurring between t and $t + \Delta t$ is $\lambda\Delta t + o(\Delta t)$. Additionally by (2), the probability of more than one event occurring is $o(\Delta t)$. Hence we have the following.

$$\begin{aligned} p_n(t + \Delta t) &= p_n(t)\big(1 - (\lambda\Delta t + o(\Delta t))\big) + p_{n-1}(t)(\lambda\Delta t + o(\Delta t)) + p_{n-2}(t)o(\Delta t) + \ldots + p_0(t)o(\Delta t) \\ &= p_n(t)\big(1 - \lambda\Delta t - o(\Delta t)\big) + p_{n-1}(t)(\lambda\Delta t + o(\Delta t)) + o(\Delta t) \\ &= p_n(t) - p_n(t)\lambda\Delta t - p_n(t)o(\Delta t) + p_{n-1}(t)\lambda\Delta t + p_{n-1}(t)o(\Delta t) + o(\Delta t) \\ &= p_n(t) - \lambda\Delta t\, p_n(t) + \lambda\Delta t\, p_{n-1}(t) + o(\Delta t) \end{aligned}$$

However, this is equivalent to the following:

$$\begin{aligned} & p_n(t + \Delta t) = p_n(t) - \lambda\Delta t\, p_n(t) + \lambda\Delta t\, p_{n-1}(t) + o(\Delta t) \\ \Longleftrightarrow \quad & \frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} = -\lambda p_n(t) + \lambda p_{n-1}(t) + \frac{o(\Delta t)}{\Delta t} \end{aligned}$$

Taking the limit as $\Delta t \to 0$, we derive the following relation.

$$\lim_{\Delta t \to 0} \frac{p_n(t + \Delta t) - p_n(t)}{\Delta t} = -\lambda p_n(t) + \lambda p_{n-1}(t) + \lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} \qquad \Longleftrightarrow \qquad \frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t)$$

Notice that this constitutes the following system of infinite linear ordinary differential equations

$$\frac{dp_n(t)}{dt} = \begin{cases} -\lambda p_0(t) & \text{for } n = 0 \\ -\lambda p_n(t) + \lambda p_{n-1}(t) & \text{for } n \geq 1 \end{cases}$$

For $n = 0$, the general solution for the above equation is $p_0(t) = Ce^{-\lambda t}$. Since $p_n(0) = 0$ (recall that the probability of an event with a continuous distributions over a 0 interval is always 0) for every $n \geq 0$, we have that $C = 1$.

$$p_0(t) = e^{-\lambda t}$$

For $n = 1$, we have the following.

$$\begin{aligned} \frac{dp_1(t)}{dt} = -\lambda p_1(t) + \lambda p_0(t) \qquad & \Longleftrightarrow \qquad \frac{dp_1(t)}{dt} = -\lambda p_1(t) + \lambda e^{-\lambda t} \\ & \Longleftrightarrow \qquad p_1(t) = Ce^{-\lambda t} + \lambda te^{-\lambda t} \end{aligned}$$

Here $C = 0$ as again $p_1(0) = 0$. And for $n = 2$, we have the following.

$$\begin{aligned} \frac{dp_2(t)}{dt} = -\lambda p_2(t) + \lambda p_1(t) \qquad & \Longleftrightarrow \qquad \frac{dp_2(t)}{dt} = -\lambda p_2(t) + \lambda^2 te^{-\lambda t} \\ & \Longleftrightarrow \qquad p_2(t) = Ce^{-\lambda t} + \frac{(\lambda t)^2}{2!}e^{-\lambda t} \end{aligned}$$

Again $C = 0$. We say that $p_n(t)$ defines the probability density function of a *Poisson distribution* with parameter $\lambda > 0$ and fixed number of occurrences n. Indeed, our observations above holds in general.

$$p_n(t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \tag{1}$$

2.2 Continuous-Time Markov Chains

We will also require the use of continuous-time Markov chains.

Definition 2 (Continuous-Time Markov Chain)**.** *Given some state space $\mathcal{X}$ and function $X : [0, \infty) \to \mathcal{X}$ where $X(t)$ denotes the current state at time t, then a continuous-time Markov chain is given by*

$$\{X(t) : t \in \mathbb{R}, 0 \leq t < \infty\}$$

Additionally, the transition probabilities are defined as follows. Let $\Delta t > 0$ be some time interval and suppose $X(t) = i$ for some state $i \in \mathcal{X}$.

1. *$X(t + \Delta t)$ is independent of all previous states $X(s)$ where $s < t$.*
2. *For any $j \in \mathcal{X}$, let $q_{ij}(t)$ be a non-negative function that measures how quickly the transition between states i, j happens. The probability of transitioning from i to j where $i \neq j$ during time t and $t + \Delta t$ is given by the following.*

$$\Pr\{X(t + \Delta t) = j \mid X(t) = i\} = q_{ij}(t)\Delta t + o(\Delta t)$$

As an example, we can consider a Poisson process with from the previous section as a CTMC. With parameter $\lambda > 0$, we define our state space to be $\mathcal{X} = \{i \in \mathbb{Z} : i \geq 0\}$ where state i denotes the event occurring i times. For some fixed time t and an interval Δt, we have the following transition probabilities.

$$\Pr\{X(t + \Delta t) = i \mid X(t) = j\} = \begin{cases} \lambda\Delta t - o(\Delta t) & \text{if } j = i - 1 \\ 1 - \lambda\Delta t - o(\Delta t) & \text{if } j = i \\ o(\Delta t) & \text{otherwise} \end{cases}$$

Observe that similar to how it was in the previous section, λ denotes the rate at which we transition from $i - 1$ to i occurrences – that is the rate at which an event occurs.

2.2.1 State Distribution Rate of Change

Given a continuous-time Markov Chain, we may be interested in the rate at which the probability we are in a current state changes with respect to time. For some time t, we denote $p_i(t)$ for any $i \in \mathcal{X}$ as the following.

$$p_i(t) = \Pr\{X(t) = i\}$$

That is, $p_i(t)$ denotes the probability that the Markov chain is in state i at time t. For time u, s where $u < s$, we also define $p_{ij}(u, s)$ for any $i, j \in \mathcal{X}$ as the following.

$$p_{ij}(u, s) = \Pr\{X(s) = j \mid X(u) = i\}$$

Here $p_{ij}(u, s)$ is the probability that the chain transitions from state i to j between time u and s. For any time $u < t$, notice that we can now split the equation for $p_i(t)$ by summing over a condition.

$$p_i(t) = \Pr\{X(t) = i\} = \sum_{r \in \mathcal{X}} \Pr\{X(u) = r\} \cdot \Pr\{X(t) = i \mid X(u) = r\} = \sum_{r \in \mathcal{X}} p_r(u)p_{ri}(u, t)$$

Now recall condition (2) of the definition for continuous time Markov chains states that $p_{ij}(t, t + \Delta t)$ for some interval Δt is given by the following. We will denote $q_i(t) = q_{ii}(t)$ for cleanliness.

$$p_{ij}(t, t + \Delta t) = \begin{cases} q_{ij}(t)\Delta t + o(\Delta t) & \text{if } i \neq j \\ 1 - q_i(t)\Delta t - o(\Delta t) & \text{otherwise} \end{cases}$$

With these identities in mind, we may now concern ourselves with the quantity $\frac{d}{dt}\, p_i(t)$. Notice that with the equation for $p_i(t)$ above and some small time period Δt, we have the following.

$$\begin{aligned} p_i(t+\Delta t) &= \sum_{r\in\mathcal{X}} p_r(t)p_{ri}(t,t+\Delta t) \\ &= p_i(t)p_{ii}(t,t+\Delta t) + \sum_{r\neq i} p_r(t)p_{ri}(t,t+\Delta t) \\ &= p_i(t)\big(1-q_i(t)\Delta t - o(\Delta t)\big) + \sum_{r\neq i} p_r(t)\big(q_{ri}(t)\Delta t + o(\Delta t)\big) \\ &= p_i(t) - p_i(t)q_i(t)\Delta t - o(\Delta t) + \sum_{r\neq i}\big(p_r(t)q_{ri}(t)\Delta t + o(\Delta t)\big) \end{aligned}$$

If we subtract $p_i(t)$ from both sides and divide by Δt, we derive the following.

$$\begin{aligned} & p_i(t+\Delta t) = p_i(t) - p_i(t)q_i(t)\Delta t - o(\Delta t) + \sum_{r\neq i}\big(p_r(t)q_{ri}(t)\Delta t + o(\Delta t)\big) \\ \Longleftrightarrow \quad & p_i(t+\Delta t) - p_i(t) = -p_i(t)q_i(t)\Delta t - o(\Delta t) + \sum_{r\neq i}\big(p_r(t)q_{ri}(t)\Delta t + o(\Delta t)\big) \\ \Longleftrightarrow \quad & \frac{p_i(t+\Delta t) - p_i(t)}{\Delta t} = -p_i(t)q_i(t) - \frac{o(\Delta t)}{\Delta t} + \sum_{r\neq i}\left(p_r(t)q_{ri}(t) + \frac{o(\Delta t)}{\Delta t}\right) \end{aligned}$$

Now let us take the limit as $\Delta t \to 0$.

$$\lim_{\Delta t\to 0} \frac{p_i(t+\Delta t) - p_i(t)}{\Delta t} = \lim_{\Delta t\to 0}\left(- p_i(t)q_i(t) - \frac{o(\Delta t)}{\Delta t} + \sum_{r\neq i}\left(p_r(t)q_{ri}(t) + \frac{o(\Delta t)}{\Delta t}\right)\right)$$

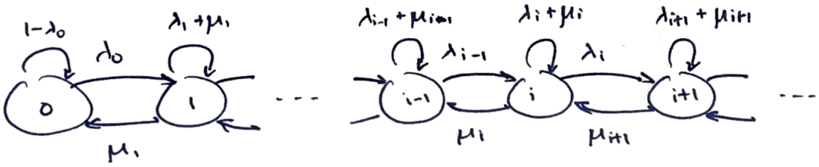
The right side becomes a derivative with respect to t and $\frac{o(\Delta t)}{\Delta t} \to 0$ as $\Delta t \to 0$. We thus have the following.

$$\frac{d}{dt}\, p_i(t) = -p_i(t)q_i(t) + \sum_{r\neq i} p_r(t)q_{ri}(t) \tag{2}$$

This differential equation is exactly what we want as it characterizes the instantaneous rate of change of the probability that our Markov chain is in state i at time t.

2.2.2 Birth-Death Process

As a final part of our review, we will look at a specific continuous-time Markov chain called the *birth-death process*. Our analysis here will also be useful when we consider queuing equations for an M/M/1 system later on. The states of a birth-death process are labeled by non-negative integers $\mathcal{X} = \{i \in \mathbb{Z} : i \geq 0\}$ where the transition probabilities are defined as follows.



For any given state i, the probability of transitioning to state $i+1$ is given by the following.

$$\Pr\{\text{population increases from } i \to i+1 \text{ in time } (t,\Delta t)\} = \lambda_i\Delta t + o(\Delta t) \qquad \forall i \geq 0$$

The probability of transitioning to state $i-1$ is then

$$\Pr\{\text{population decreases from } i \to i-1 \text{ in time } (t, \Delta t)\} = \mu_i \Delta t + o(\Delta t) \qquad \forall i \geq 1$$

The probability of remaining at state i is given by

$$\Pr\{\text{population remains at } i \text{ in time } (t, \Delta t)\} = (\lambda_i + \mu_i)\Delta t + o(\Delta t) \qquad \forall i \geq 0$$

Finally, the probability of transitioning to any state $j \notin \{i-1, i, i+1\}$ is 0. The rate at which we transition from one state $q_{ij}(t)$ is given by the following.

$$q_{ij}(t) = \begin{cases} \lambda_i & \text{if } j = i+1 \\ \mu_i & \text{if } j = i-1 \\ \lambda_i + \mu_i & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

Notice that $q_{ij}(t)$ is constant with respect to t. We often say that a Markov chain like this is *time-homogeneous* and simply denote $q_{ij} = q_{ij}(t)$. Now suppose we are interested in the steady-state distribution. When the system has reached a steady-state, then $p_i(t)$ will no longer change with respect to t.

It thus suffices to equate $\frac{dp_i(t)}{dt} = 0$. We can write equation 2 for the birth-death process as follows.

$$\frac{d}{dt}\, p_i(t) = -p_i(t)q_i(t) + \sum_{r \neq i} p_r(t)q_{ri}(t) = \begin{cases} -(\lambda_i + \mu_i)p_i(t) + \lambda_{i-1}p_{i-1}(t) + \mu_{i+1}p_{i+1}(t) & \text{if } i \geq 1 \\ -\lambda_0 p_0(t) + \mu_1 p_1(t) & \text{if } i = 0 \end{cases}$$

Now let $\frac{dp_i(t)}{dt} = 0$. Writing $p_i = p_i(t)$, we notice the following equation for $i = 1$.

$$p_1 = \left(\frac{\lambda_0}{\mu_1}\right) p_0$$

However, if we take any $i > 1$, we have the following equation.

$$p_{i+1} = \left(\frac{\lambda_i + \mu_i}{\mu_{i+1}}\right) p_i - \left(\frac{\lambda_{i-1}}{\mu_{i+1}}\right) p_{i-1}$$

Let us consider $i = 2$. This gives the following equation.

$$p_2 = \left(\frac{\lambda_1 + \mu_1}{\mu_2}\right) p_1 - \left(\frac{\lambda_0}{\mu_2}\right) p_0 = \left(\frac{\lambda_1 + \mu_1}{\mu_2}\right)\left(\frac{\lambda_0}{\mu_1}\right) p_0 - \left(\frac{\lambda_0}{\mu_2}\right) p_0 = \left(\frac{\lambda_1 \lambda_0}{\mu_2 \mu_1}\right) p_0$$

And for $i = 3$...

$$p_3 = \left(\frac{\lambda_2 + \mu_2}{\mu_3}\right) p_2 - \left(\frac{\lambda_1}{\mu_3}\right) p_1 = \left(\frac{\lambda_2 + \mu_2}{\mu_3}\right)\left(\frac{\lambda_1 \lambda_0}{\mu_2 \mu_1}\right) p_0 - \left(\frac{\lambda_1}{\mu_3}\right)\left(\frac{\lambda_0}{\mu_1}\right) p_0 = \left(\frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1}\right) p_0$$

Indeed, this pattern holds in general allowing us to demonstrate the following theorem.

Theorem 3. *For any $n \geq 1$, the steady-state solution for p_n in a birth-death process is given by the following.*

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \tag{3}$$

And as $\sum_{i \in \mathcal{X}} p_i = 1$, we have that p_0 is given by the following.

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right)^{-1} \tag{4}$$

3 Analysis of M/M/1 Systems

We now proceed with the analysis of M/M/1 queuing systems. We model a queuing system is an M/M/1 system by assuming that the arrival process is (M)emoryless, the service process is (M)emoryless, and there is (1) queue within the system. The arrival and service process are memoryless in the sense that the arrival and service times are both Poisson processes.

There are a handful of parameters that characterize important information regarding our system.

- λ denotes the *arrival rate*, or the average number of arriving items per unit time.
- T denotes the average time to service an item in the queue.
- $\mu = \frac{1}{T}$ denotes the *service rate*, or the rate at which items are serviced per unit time.
- $u = \frac{\lambda}{\mu} = \lambda T$ denotes the *server utilization*. This represents the average number of items to arrive during the time it takes to service one item.
- W denotes the average amount of time an item spends waiting in the system's queue.
- L denotes the average number of items currently inside the system.

If we are given information regarding two of L, W, λ then Little's Law, which states $L = \lambda W$, easily allows us to calculate the missing third value. However, our goal in modeling a queuing system as M/M/1 is to determine what these parameters are if we are given only one of the parameters.

3.1 Average Waiting Time

Suppose we are given the values for λ and T. The values of μ and u can then be easily derived but W and L may not. We calculate W with respect to the given parameters λ, T, μ, u so that we may then use Little's Law to determine L.

The equation for W is given by theorem 4 below. Before proceeding with the proof, let us sketch the approach that the proof takes. To calculate W, we first begin by modeling the queuing system as a birth-death process which is possible as the arrival and departure (servicing) processes are both Poisson and with only one queue any discrete moment of time can change the net number of items inside the system by 1 item.

Using the equations 3 and 4, we can determine the average number of items in the system. Little's Law will then give us the average time an item spends inside the system, but this must be the sum of average time spent waiting in the queue W and the time it takes to be serviced T. Immediately from this, we will have W as required.

Theorem 4. *Given a M/M/1 queuing system with arrival rate λ, average service time T, average service rate μ, and server utilization u, the average time an item spends in the queue W is given by the following.*

$$W = T\left(\frac{u}{1-u}\right) \tag{5}$$

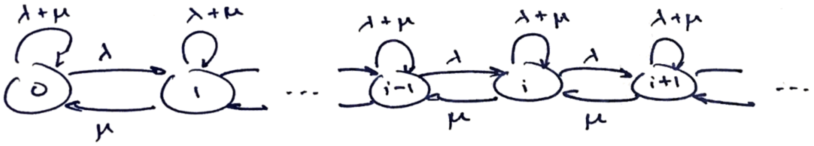
Proof. Consider that the M/M/1 queuing system has a Poisson arrival process. For any time t and an interval Δt, the arrival rate λ admits the following probabilities.

$$\begin{aligned}&\Pr\{\text{an item arrives during the interval } (t, t+\Delta t)\} = \lambda \Delta t + o(\Delta t)\\&\Pr\{\text{more than one arrival occurs during } (t, t+\Delta t)\} = o(\Delta t)\end{aligned}$$

Similarly, as the servicing process is Poisson with rate μ, the probability that an item is serviced and more than one item are serviced are the following.

$$\begin{aligned}&\Pr\{\text{an item is serviced during the interval } (t, t+\Delta t)\} = \mu \Delta t + o(\Delta t)\\&\Pr\{\text{more than one item is serviced during } (t, t+\Delta t)\} = o(\Delta t)\end{aligned}$$

Because the number of arrivals and serviced items are both independent values within overlapping time intervals, the M/M/1 queuing system is a birth-death process where an arrival signifies a birth and an item being serviced signifies a death. The states $\mathcal{X}$ determine the number of items within the queuing system and the birth and death rates are $\lambda_i = \lambda$ and $\mu_i = \mu$ respectively for every state $i \in \mathcal{X}$.



We are interested in determining N the average number of items in the queuing systems, or in this case the average population size in the birth-death process. Consider the steady-state probability of being in state i. If $i = 0$, then via equation 4, this is the following.

$$p_0 = \left(1 + \sum_n^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right)^{-1} = \left(1 + \sum_n^{\infty} \prod_{i=1}^{n} \frac{\lambda}{\mu}\right)^{-1} = \frac{1}{1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n}$$

Notice that $u = \frac{\lambda}{\mu}$ and if $0 \leq u < 1$, then the sum in the denominator converges to the following.

$$p_0 = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} = 1 - \frac{\lambda}{\mu} = 1 - u$$

For any other $n > 0$, equation 3 dictates the steady-state probability of being in state n as given by the following.

$$p_n = p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = p_0 \left(\frac{\lambda}{\mu}\right)^n = (1-u)u^n$$

Thus for any $i \in \mathcal{X}$ we have $p_i = (1-u)u^i$. Since we know the probability that there are i items inside the system for any i, we can calculate N the expected number of items inside the queue as follows.

$$\begin{aligned}
\mathbf{E}\{\#\text{ items inside queue}\} &= \sum_{i=0}^{\infty} i p_i = \sum_{i=0}^{\infty} i u^i (1-u) \\
&= u(1-u) \sum_{i=0}^{\infty} i u^{i-1} = u(1-u) \sum_{i=0}^{\infty} \frac{d}{du}(u^i) \\
&= u(1-u)\, \frac{d}{du}\left(\sum_{i=0}^{\infty} u^i\right) = u(1-u)\, \frac{d}{du}\left(\frac{1}{1-u}\right) \\
&= u(1-u)\, \frac{1}{(1-u)^2} = \frac{u}{1-u}
\end{aligned}$$

Now by Little's Law, the average time spent inside the queuing system is $\tau = \frac{N}{\lambda}$ or...

$$\tau = \frac{N}{\lambda} = \frac{u}{\lambda(1-u)} = \frac{1}{\lambda - \mu}$$

Finally, the time spent inside the system is simply the time spent waiting in the queue and the time spent being serviced. That is the following.

$$\tau = W + T \qquad \Longleftrightarrow \qquad W = \tau - T = \frac{1}{\lambda - \mu} - \frac{1}{\mu} = T\left(\frac{u}{1-u}\right)$$

We thus have $W = T(\frac{u}{1-u})$ as required. □

As a final remark, notice that in the proof of theorem 4 we assume that $0 \leq \frac{\lambda}{\mu} = u < 1$. Notice that this implies that at most one item arrives into the queuing system per amount of time it takes to process an item already in the queue. Under the case where this is true, then we say that our system is *stable*. It makes sense to label it this way as an *unstable* system would imply that the birth-death stable-state probabilities would not converge.

$$p_0 = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n}$$

Specifically we have the denominator of p_0 would diverge if $\frac{\lambda}{\mu} \geq 1$.

4 Analysis of M/G/1 Systems

It is convenient to assume that the arrival and servicing processes are Poisson, but this might not always be the case. For example, priority based thread schedulers are definitely not Poisson as there is some degree of dependence between the threads that currently exist within the system.

In an M/G/1 queuing system, the arrival process is (M)emoryless or Poisson, the servicing process is (G)eneral in that we assume a generic distribution in servicing times, and there is (1) queue within the system. The generic distribution for servicing times can indeed be characterized by the average time to service the item T, but it will also require knowledge of the standard deviation in service times σ. For this reason, we will define a new parameter for our system called the *squared coefficient of variance* C defined as follows.

$$C = \frac{\sigma^2}{T^2}$$

A critical assumption that we will make regarding M/G/1 systems is that the arrival of items is *Ergodic*. Specifically if v_i denotes the number of items to arrive during the time it takes to process the i-th item, then the following holds.

$$\lim_{i \to \infty} \mathbf{E}\{v_i\} = \mathbf{E}\{v\}$$

Here $\mathbf{E}\{v\}$ denotes the average number of arrivals during a service interval.

4.1 Pollaczek-Khinchine Formula

We will first derive a formula that relates the average number of jobs called the *Pollaczek-Khinchine formula*. The proof of this formula requires two lemmas that determine the average and variance of v, the number of items to arrive while an item is being serviced by the system.

Lemma 5. *Given an M/G/1 system with an arrival rate of λ and server utilization rate of u, the average number of items to arrive during the time it takes to service one element is given below.*

$$\mathbf{E}\{v\} = u$$

Proof. Since our system is M/G/1, the arrival process is Poisson. Thus for any integer $k \geq 0$, the conditional probability that k items arrive given that it takes a time length of t to service one item is given by equation 1.

$$\Pr\{v = k \mid \text{interval of length } t\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

On the contrary, the servicing process is general so suppose that it is characterized by the probability density function $f(t)$. We can then compute $\Pr\{v = k\}$ for any $k \geq 0$ in the following manner.

$$\Pr\{v = k\} = \int_0^{\infty} \Pr\{v = k \mid \text{interval of length } t\} \cdot f(t)\, dt = \int_0^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} f(t)\, dt$$

We may now calculate $\mathbf{E}\{v\}$ as follows.

$$\begin{aligned}\mathbf{E}\{v\} &= \sum_{k=0}^{\infty} k \cdot \Pr\{v = k\} \\ &= \sum_{k=0}^{\infty} k\left(\int_0^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} f(t)\, dt\right) \\ &= \sum_{k=0}^{\infty} \int_0^{\infty} \frac{(\lambda t)^k}{(k-1)!} e^{-\lambda t} f(t)\, dt \\ &= \int_0^{\infty} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{(k-1)!} e^{-\lambda t} f(t)\, dt \\ &= \int_0^{\infty} (\lambda t) f(t) e^{-\lambda t} \left(\sum_{k=0}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!}\right) dt\end{aligned}$$

Notice that the infinite sum is exactly the Taylor expansion of $e^{\lambda t}$ thus we have that given below.

$$\mathbf{E}\{v\} = \int_0^{\infty} (\lambda t) f(t) e^{-\lambda t} \left(\sum_{k=0}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!}\right) dt = \int_0^{\infty} \lambda t f(t) e^{-\lambda t} e^{\lambda t}\, dt = \lambda \int_0^{\infty} t f(t)\, dt = \lambda \mathbf{E}\{t\}$$

However, the expected service time is given by $T = \mathbf{E}\{t\}$ and as $\lambda T = u$, we have that $\mathbf{E}\{v\} = u$ as required. □

Using a proof similar to lemma 5, we can derive the variance of v as the following.

Lemma 6. *Given an M/G/1 system with an arrival rate of λ, arrival time variance of σ^2 and server utilization rate of u, then the variance of v is*

$$\mathbf{Var}\{v\} = u + \lambda^2\sigma^2$$

We can now demonstrate the proof for the Pollaczek-Khinchine formula.

Theorem 7 (Pollaczek-Khinchine)**.** *Suppose we are given a M/G/1 queuing system with server utilization m, arrival rate λ, and arrival time variance σ^2. The average number of items inside the queuing system N is then*

$$N = \frac{u}{2} + \frac{u + \lambda^2\sigma^2}{2(1-u)} \tag{6}$$

Proof. Let n_i denote the number of items in the queuing system when the i-th item is finished being processed by the system and recall that v_i denotes the number of items to arrive while the i-th item is being processed. Notice that these two variables are related as follows.

$$n_i = \begin{cases} v_i & \text{if } n_{i-1} = 0 \\ (n_{i-1} - 1) + v_i & \text{otherwise} \end{cases}$$

The first condition means that the if the queuing system has no items after processing item $i-1$, then after "processing" i, there will be v_i items in the queue. The second condition states that if there are items in the queue, then one is removed by being processed and v_i new items will arrive. This equation can be written more compactly as:

$$n_i = (n_{i-1} - \mathbb{I}\{n_{i-1} \geq 1\}) + v_i \tag{7}$$

With $\mathbb{I}\{n_{i-1} \geq 1\}$ as the indicator random variable that $n_{i-1} \geq 1$, this follows since if $n_{i-1} = 0$, then $\mathbb{I}\{n_{i-1} \geq 1\} = 0$ and $n_i = v_j$. Otherwise, $n_{i-1} \geq 1$ and $n_i = n_{i-1} - 1 + v_i$. We now claim that $\mathbf{E}\{\mathbb{I}\{n \geq 1\} = \mathbf{E}\{v\}$. To see, this observe that as arrivals are Poisson, we have that $\mathbf{E}\{n_i\} = \mathbf{E}\{n_{i-1}\}$ as $i \to \infty$. Taking the expectation above, we derive:

$$\mathbf{E}\{n_i\} = \mathbf{E}\{(n_{i-1} - \mathbb{I}\{n_{i-1} \geq 1\}) + v_i\} = \mathbf{E}\{n_{i-1}\} - \mathbf{E}\{\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{v_i\}$$

Then taking the limit

$$\begin{aligned}\lim_{i\to\infty} \mathbf{E}\{n_i\} &= \lim_{i\to\infty} \mathbf{E}\{n_{i-1}\} - \mathbf{E}\{\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{v_i\} \\ \mathbf{E}\{n\} &= \mathbf{E}\{n\} - \mathbf{E}\{\mathbb{I}\{n \geq 1\}\} + \mathbf{E}\{v\} \\ \mathbf{E}\{v\} &= \mathbf{E}\{\mathbb{I}\{n \geq 1\}\}\end{aligned}$$

Now to calculate $\mathbf{E}\{n\}$, we first square both sides of equation 7.

$$\begin{aligned}n_i^2 &= \big((n_{i-1} - \mathbb{I}\{n_{i-1} \geq 1\}) + v_i\big)^2 \\ &= n_{i-1}^2 - 2n_{i-1}\mathbb{I}\{n_{i-1} \geq 1\} + \mathbb{I}\{n_{i-1} \geq 1\}^2 + 2v_i n_{i-1} - 2v_i\mathbb{I}\{n_{i-1} \geq 1\} + v_i^2\end{aligned}$$

Observe that for any x, we have $\mathbb{I}\{x \geq 1\} = \mathbb{I}\{x \geq 1\}^2$ and thus taking the expectation, we have the following.

$$\begin{aligned}\mathbf{E}\{n_i^2\} &= \mathbf{E}\{n_{i-1}^2 - 2n_{i-1}\mathbb{I}\{n_{i-1} \geq 1\} + \mathbb{I}\{n_{i-1} \geq 1\})^2 + 2v_i n_{i-1} - 2v_i\mathbb{I}\{n_{i-1} \geq 1\} + v_i^2\} \\ &= \mathbf{E}\{n_{i-1}^2\} - 2\mathbf{E}\{n_{i-1}\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{\mathbb{I}\{n_{i-1} \geq 1\}^2\} + 2\mathbf{E}\{v_i n_{i-1}\} - 2\mathbf{E}\{v_i\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{v_i^2\} \\ &= \mathbf{E}\{n_{i-1}^2\} - 2\mathbf{E}\{n_{i-1}\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{\mathbb{I}\{n_{i-1} \geq 1\}\} + 2\mathbf{E}\{v_i n_{i-1}\} - 2\mathbf{E}\{v_i\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{v_i^2\}\end{aligned}$$

Since arrivals are also Poisson, v_i and n_j for any i, j are independent. Thus the above equation reduces to the following.

$$\mathbf{E}\{n_i^2\} = \mathbf{E}\{n_{i-1}^2\} - 2\mathbf{E}\{n_{i-1}\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{\mathbb{I}\{n_{i-1} \geq 1\}\} + 2\mathbf{E}\{v_i\}\mathbf{E}\{n_{i-1}\} - 2\mathbf{E}\{v_i\}\mathbf{E}\{\mathbb{I}\{n_{i-1} \geq 1\}\} + \mathbf{E}\{v_i^2\}$$

And if we take the limit as $i \to \infty$, this equation becomes:

$$\begin{aligned}\mathbf{E}\{n^2\} &= \mathbf{E}\{n^2\} - 2\mathbf{E}\{n\mathbb{I}\{n \geq 1\}\} + \mathbf{E}\{\mathbb{I}\{n \geq 1\}\} + 2\mathbf{E}\{v\}\mathbf{E}\{n\} - 2\mathbf{E}\{v\}\mathbf{E}\{\mathbb{I}\{n \geq 1\}\} + \mathbf{E}\{v^2\} \\ &= \mathbf{E}\{n^2\} - 2\mathbf{E}\{n\mathbb{I}\{n \geq 1\}\} + \mathbf{E}\{v\} + 2\mathbf{E}\{v\}\mathbf{E}\{n\} - 2\mathbf{E}\{v\}\mathbf{E}\{v\} + \mathbf{E}\{v^2\}\end{aligned}$$

Notice that $\mathbf{E}\{n\mathbb{I}\{n \geq 1\}\} = \mathbf{E}\{n\}$. This follows as

$$\mathbf{E}\{n\mathbb{I}\{n \geq 1\}\} = \sum_{k=0}^{\infty} k\mathbb{I}\{k \geq 1\}\Pr\{n = k\} = \sum_{k=1}^{\infty} k\Pr\{n = k\} = \mathbf{E}\{n\}$$

We can then write

$$\begin{aligned}\mathbf{E}\{n^2\} &= \mathbf{E}\{n^2\} - 2\mathbf{E}\{n\} + \mathbf{E}\{v\} + 2\mathbf{E}\{v\}\mathbf{E}\{n\} - 2\mathbf{E}\{v\}^2 + \mathbf{E}\{v^2\} \\ \iff \quad 2\mathbf{E}\{n\}(1 - \mathbf{E}\{v\}) &= \mathbf{E}\{v\} - \mathbf{E}\{v\}^2 + \mathbf{E}\{v^2\} - \mathbf{E}\{v\}^2\end{aligned}$$

Recall that $\mathbf{Var}\{v\} = \mathbf{E}\{v^2\} - \mathbf{E}\{v\}^2$, thus the above reduces to the following.

$$2\mathbf{E}\{n\}(1 - \mathbf{E}\{v\}) = \mathbf{E}\{v\} - \mathbf{E}\{v\}^2 + \mathbf{E}\{v^2\} - \mathbf{E}\{v\}^2 \qquad \iff \qquad 2\mathbf{E}\{n\}(1 - \mathbf{E}\{v\}) = \mathbf{E}\{v\} - \mathbf{E}\{v\}^2 + \mathbf{Var}\{v\}$$

Finally, by lemmas 5 and 6, we can replace $\mathbf{E}\{v\}$ and $\mathbf{Var}\{v\}$ with the following.

$$\begin{aligned}2\mathbf{E}\{n\}(1 - \mathbf{E}\{v\}) = \mathbf{E}\{v\} - \mathbf{E}\{v\}^2 + \mathbf{Var}\{v\} \qquad &\iff \qquad 2\mathbf{E}\{n\}(1 - u) = u - u^2 + u + \lambda^2\sigma^2 \\ &\iff \qquad \mathbf{E}\{n\} = \frac{u}{2} + \frac{u + \lambda^2\sigma^2}{2(1 - u)}\end{aligned}$$

Which is as required. □

4.2 Average Waiting Time

The Pollaczek-Khinchine formula with Little's Law immediately implies a formula for the average waiting time.

Theorem 8. *The average time an item waits in a queue within an M/G/1 system is given by W below.*

$$W = \frac{T}{2}\left(\frac{u}{1-u}\right)(1+C) \tag{8}$$

Proof. By theorem 7, the average number of items within an M/G/1 system is given by

$$\mathbf{E}\{n\} = \frac{u}{2} + \frac{u+\lambda^2\sigma^2}{2(1-u)}$$

Let τ denote the average time an item spends within the system. Little's theorem states that this is the following.

$$\tau = \frac{\mathbf{E}\{n\}}{\lambda} = \frac{u}{2\lambda} + \frac{u+\lambda^2\sigma^2}{2(1-u)\lambda} = \frac{T}{2} + \frac{u+\lambda^2\sigma^2}{2(1-u)\lambda}$$

Finally, as $\tau = W + T$, we have that W is given by the following.

$$W = \tau - T = \frac{T}{2} + \frac{u+\lambda^2\sigma^2}{2(1-u)\lambda} - T = \frac{u+\lambda^2\sigma^2}{2(1-u)\lambda} - \frac{T}{2} = \frac{T}{2}\left(\frac{u+\lambda^2\sigma^2}{T\lambda(1-u)} - 1\right)$$

Recall that $u = \lambda T$. We thus have

$$W = \frac{T}{2}\left(\frac{u+\lambda^2\sigma^2}{u(1-u)} - 1\right) = \frac{T}{2}\left(\frac{u}{1-u}\right)\left(\frac{u+\lambda^2\sigma^2}{u^2} - \frac{1-u}{u}\right) = \frac{T}{2}\left(\frac{u}{1-u}\right)\left(\frac{1}{u} + \frac{\lambda^2\sigma^2}{(\lambda T)^2} - \frac{1}{u} + 1\right)$$

This reduces to $\frac{T}{2}\left(\frac{u}{1-u}\right)\left(1+\frac{\sigma^2}{T^2}\right)$ and since $C = \frac{\sigma^2}{T^2}$, we thus have $W = \frac{T}{2}\left(\frac{u}{1-u}\right)(1+C)$ as required. □