Contents

1 Introduction

A retrial queue consists of server(s) and a queue. Customers arrive according to a poisson process. These are called primary calls. If the server is free at the time of primary call, they start getting served immediately. Otherwise , they are added to the queue, and serve as a source of repeated calls. Every source produces a poisson process of repeated calls. Of course, if there exists the limit.

2 Notation

- λ : arrival rate of primary calls
- μ : rate of repeated calls
- $B(x)$: service distribution
- $C(t)$: no of busy servers at time t
- $N(t)$: no of sources of repeated calls
- $\xi(t)$: age of current process
- $\beta(t) = \int_0^\infty e^{-sx} dB(x)$: Laplace transform of service time
- $b(x) = \frac{B'(x)}{1-B(x)}$: Hazard rate
- $k(z) = \sum_0^\infty k_n z^n = \beta(\lambda - \lambda z)$

  $$k_n = \int_0^\infty \frac{\lambda x^n}{n!} e^{-x} dB(x)$$

  is distribution of number of primary calls that arrive during service time of a call

3 M/M/1 Retrial Queue

In M/M/1 retrial queue, the service time distribution is

$$B(x) = 1 - e^{-\nu x}$$

3.1 CTMC

The states of CTMC are $(C(t), N(t))$.
$C(t) = 1$ or 0 in case of single server

3.1.1 Limiting Distribution

For an M/M/1 retrial queue in the steady state, the joint distribution of server state $C(t)$ and queue length $N(t)$

$$p_{in} = P\{C(t) = i, N(t) = n\}$$

is given by

$$\begin{aligned} p_{0n} &= \frac{\rho^n}{n!\mu^n} \prod_{i=0}^{n-1} (\lambda + i\mu) \cdot (1-\rho)^{\frac{\lambda}{\mu}+1}, \\ p_{1n} &= \frac{\rho^{n+1}}{n!\mu^n} \prod_{i=1}^{n} (\lambda + i\mu) \cdot (1-\rho)^{\frac{\lambda}{\mu}+1}. \end{aligned}$$

The corresponding partial generating functions are given by

$$\begin{aligned} p_0(z) &\equiv \sum_{n=0}^{\infty} z^n p_{0n} = (1-\rho) \left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\mu}}, \\ p_1(z) &\equiv \sum_{n=0}^{\infty} z^n p_{1n} = \rho \left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\mu}+1}. \end{aligned}$$

Proof:

From a state $(0, n)$, only transitions into the following states are possible:

1. $(1, n)$ with rate λ;
2. $(1, n-1)$ with rate ν.

The first transition is due to the arrival of a primary call, and the second is due to the arrival of a repeated call. Since state $(0, n)$ means that the server is free, there is no transition corresponding to the service completion.

Reaching state $(0, n)$ is possible only from state $(1, n)$ with rate ν.

From a state $(1, n)$, only transitions into the following states are possible:

1. $(1, n+1)$ with rate λ;
2. $(0, n)$ with rate ν.

The first transition is due to the arrival of a primary call, and the second is due to a service completion. Since in state $(1, n)$ the server is busy, there is no transition corresponding to the arrival of a repeated call.

Reaching state $(1, n)$ is possible only from the states:

1. $(0, n)$ with rate λ;
2. $(0, n+1)$ with rate $(n+1)\mu$;
3. $(1, n-1)$ with rate λ.

Thus the set of statistical equilibrium equations for the probabilities p_{0n}, p_{1n} is

$$(\lambda + n\mu)p_{0n} = \nu p_{1n},$$

$$(\lambda + \nu)p_{1n} = \lambda p_{0n} + (n+1)\mu p_{0,n+1} + \lambda p_{1,n-1}$$

We use partial generating functions to solve these equations

$$p_0(z) \equiv \sum_{n=0}^{\infty} z^n p_{0n}$$

$$p_1(z) \equiv \sum_{n=0}^{\infty} z^n p_{1n}.$$

For them the above equations become:

$$\lambda p_0(z) + \mu z p_0'(z) = \nu p_1(z)$$

$$(\nu + \lambda - \lambda z)p_1(z) = \lambda p_0(z) + \mu p_0'(z).$$

Eliminating $p_1(z)$ we get the following differential equation for

$$p_0(z);$$

with solution

$$p_0'(z) = \frac{\lambda \rho}{\mu(1 - \rho z)} p_0(z)$$

$$p_0(z) = \frac{Const}{(1 - \rho z)^{\frac{\lambda}{\mu}}}.$$

$$p_1(z) \quad = \quad \rho p_0(z) + \tfrac{\mu z}{\nu} p_0'(z) = \rho p_0(z) + \tfrac{\rho^2 z}{1 - \rho z} p_0(z)$$

where $\rho = \frac{\lambda}{\mu}$

$$p_1(z) = \frac{\rho}{(1 - \rho z)} p_0(z)$$

The constant can be found with the help of the normalizing condition

$$\sum_{n=0}^{\infty} (p_{0n} + p_{1n}) = p_0(1) + p_1(1) = 1,$$

which implies that Const $= (1 - \rho)^{\frac{\lambda}{\mu} + 1}$. ■

3.1.2 Performance Metrics

- **Mean number of jobs in queue**

  $E[N(t)] = \dfrac{\rho(\lambda + \rho\mu)}{(1-\rho)\mu}$

- **Mean number of jobs in system**

  $E[K(t)] = \dfrac{\rho(\lambda + \mu)}{(1-\rho)\mu}$

- **Mean sojourn time**

  $W = \dfrac{\rho(\lambda + \mu)}{(1-\rho)\mu\lambda}$

- **Blocking probability**

  $p_b = \rho = \dfrac{\lambda}{\nu}$

- **Recurrence condition**

  $\rho < 1$ for positive recurrence
  $\rho = 1$ for null recurrence
  $\rho > 1$ for transience

- **Mean number of retrials per job**

  $E[R] = \dfrac{\rho(\lambda + \rho\mu)}{(1-\rho)\lambda}$

Proofs

(a) The stationary distribution of the number of sources of repeated calls $q_n = PN(t) = n$ has the generating function

$$p(z) = p_0(z) + p_1(z) = (1 + \rho - \rho z)(\frac{1-\rho}{1-\rho z})^{\frac{\lambda}{\mu}+1}.$$

$$E[N(t)] = p'(1)$$

(b) The stationary distribution of the number of customers in the system $Q_n = P\{K(t) = n\}$ has the generating function

$$Q(z) = p_0(z) + zp_1(z) = (\frac{1-\rho}{1-\rho z})^{\frac{\lambda}{\mu}+1}$$

.

$$E[K(t)] = Q'(1)$$

(c) The mean sojourn time W can be found by Little's Law

$$W = \frac{E[K(t)]}{\lambda}$$

(d) Blocking probability

$$p = p_1(1)$$

(e) Recurrence conditions can be calculated by finding condition fro mean sojourn time to be finite

(f) If a job spends time T inside the queue, then it retries according to a poisson process with rate μT

$$E[R|t = T] = \mu T$$

Using law of iterated expectations,

$$\begin{aligned} E[R] &= E_T[E[R|T]] \\ &= E_T[\mu T] \\ &= \mu E[T] \end{aligned}$$

3.1.3 Generator matrix

The state of the server can be described in more detail. Namely, let $M(t)$ be the total number of arrivals (both primary and repeated) since the last departure time. This means that if $M(t) = 0$ then the server is free at time t. But if $M(t) = m \geq 1$, then the server is occupied at time t and during the elapsed service time there were exactly $m - 1$ unsuccessful attempts to get service. The process $(M(t), N(t))$ is a Markov process with Z_+^2 as the state space. The rates of transitions of the process $((M(t), N(t))$ are

1. if $m = 0$ then

$$q_{(m,n)(i,j)} = \begin{cases} \lambda, & \text{if } (i,j) = (1,n) \\ n\mu, & \text{if } (i,j) = (1,n-1) \\ -(\lambda + n\mu), & \text{if } (i,j) = (0,n) \\ 0, & \text{otherwise} \end{cases}$$

2. if $m \geq 1$ then

$$q_{(m,n)(i,j)} = \begin{cases} \lambda, & \text{if } (i,j) = (m+1,n+1) \\ n\mu, & \text{if } (i,j) = (m+1,n) \\ \nu, & \text{if } (i,j) = (0,n) \\ -(\lambda + \nu + n\mu), & \text{if } (i,j) = (m,n) \\ 0, & \text{otherwise} \end{cases}$$

$C(t) = \delta(M(t))$, where $\delta(n)$ is the indicator function of positive integers, and thus the process $(C(t), N(t))$ can be thought of as a function of the process $(M(t), N(t))$.

3.2 DTMC

3.2.1 Transforming CTMC into Embedded DTMC

We convert the CTMC to an Embedded DTMC by taking $N_i = N(\eta_i)$ i.e no of calls in orbit at the time η_i of i^{th} departure.

Let $N_i = N(\eta_i)$ be the number of calls in orbit at the time η_i of the i th departure. It is easy to see that

$$N_i = N_{i-1} - B_i + \nu_i$$

where B_i is the number of sources which enter service at time ξ_i (i.e. $B_i = 1$ if the i th call is a repeated call and $B_i = 0$ if the i th call is a primary call) and ν_i is the number of primary calls which arrive in the system during the service time S_i of the i th call.

The Bernoulli random variable B_i depends on the history of the system before time η_{i-1} only through N_{i-1} and its conditional distribution is given by

$$\left.\begin{array}{l} \mathrm{P}\left\{B_i = 0 \mid N_{i-1} = n\right\} = \frac{\lambda}{\lambda + n\mu}, \\ \mathrm{P}\left\{B_i = 1 \mid N_{i-1} = n\right\} = \frac{n\mu}{\lambda + n\mu}. \end{array}\right\}$$

The random variable ν_i does not depend on events which have occurred before epoch ξ_i and has distribution

$$k_n = \mathrm{P}\left\{\nu_i = n\right\} = \int_0^\infty \frac{(\lambda x)^n}{n!} e^{-\lambda x} dB(x)$$

with generating function

$$k(z) \equiv \sum_{n=0}^{\infty} k_n z^n = \beta(\lambda - \lambda z)$$

and mean value

$$\mathrm{E}[\nu_i] = \sum_{n=0}^{\infty} n k_n = \rho$$

3.2.2 Transition Probabilities

The above remarks imply that the sequence of random variables $\{N_i\}$ forms a Markov chain, which is the embedded chain for our queueing system. Its one-step transition probabilities $r_{mn} = \mathrm{P}\left\{N_i = n \mid N_{i-1} = m\right\}$ are given by the formula

$$r_{mn} = \frac{\lambda}{\lambda + m\mu} k_{n-m} + \frac{m\mu}{\lambda + m\mu} k_{n-m+1}$$

$r_{mn} \neq 0$ only for $m = 0, 1, \ldots, n+1$.

Proof

With probability $\frac{\lambda}{\lambda+m\mu}$ the i call is a primary call (and the number of sources does not change) and with probability $\frac{m\mu}{\lambda+m\mu}$ the i th call is a repeated call (and the number of sources decreases by 1). To have n sources in the system at time η_i we need $n-m$ new arrivals during the service time of the i th call in the first case (probability of this event is k_{n-m}) and $n-m+1$ new arrivals during the service time of the i th call in the second case (probability of this event is k_{n-m+1}).

3.2.3 Mean Length of Queue

$$N_i = N_{i-1} - B_i + \nu_i$$

Taking mean values of both sides

$$E[N_i] = E[N_{i-1}] - E[B_i] + E[\nu_i]$$

Since in the steady state EN_i does not depend on i and $E\nu_i = \rho$, we get:

$$E[B_i] = \rho$$

Taking mean values of the squares of both sides of

$$\begin{aligned} E[N_i^2] &= E[N_{i-1}^2] + E[B_i^2] + E[\nu_i^2] \\ &\quad - 2E[N_{i-1}B_i] + 2E[N_{i-1}\nu_i] - 2E[B_i\nu_i] \end{aligned}$$

In the steady state $E[N_i^2] = E[N_{i-1}^2]$. Besides,

- since B_i is a Bernoulli random variable, $E[B_i^2] = E[B_i] = \rho$;
- $E[\nu_i^2] = k''(1) + k'(1) = \lambda^2\beta_2 + \rho$;
- since ν_i does not depend on N_{i-1} and B_i, we have:
  - and finally:

$$E[B_i\nu_i] = E[B_i] \cdot E[\nu_i] = \rho^2;$$

$$E[N_i^2] = E[N_{i-1}^2] + E[B_i^2] + E[\nu_i^2]$$

$$E[N_{i-1}\nu_i] = E[N_{i-1}] \cdot E[\nu_i] = \rho \cdot E[N_i]$$

$$\begin{aligned} E[N_{i-1}B_i] &= \sum_{n=0}^{\infty} E[N_{i-1}B_i | N_{i-1} = n] \cdot P(N_{i-1} = n) \\ &= \sum_{n=0}^{\infty} nE[B_i | N_{i-1} = n]\pi_n \\ &= \sum_{n=0}^{\infty} n\frac{n\mu}{\lambda+n\mu}\pi_n = \sum_{n=0}^{\infty} n(1 - \frac{\lambda}{\lambda+n\mu})\pi_n \end{aligned}$$

$$= \sum_{n=0}^{\infty} n\pi_n - \frac{\lambda}{\mu}\sum_{n=0}^{\infty}\frac{n\mu}{\lambda + n\mu}\pi_n$$

$$= E[N_i] - \frac{\lambda}{\mu}]E[B_i]$$

$$= E[N_i] - \frac{\lambda\rho}{\mu}$$

Substituting these values above, we get

$$E[N] = \rho + \frac{\lambda^2}{1-\rho}\left(\frac{\beta_1}{\mu} + \frac{\beta_2}{2}\right)$$

For the exponential case

$$E[N] = \rho + \frac{\lambda\rho}{1-\rho}\left(\frac{1}{\mu} + \frac{1}{\nu}\right)$$

4 M/M/K

4.1 Description of the model

Consider a group of c fully available servers in which a Poisson flow of calls with rate λ arrives. These calls are called primary calls

If an arriving primary call finds some server free it immediately occupies a server and leaves the system after service. Otherwise, if all servers are engaged, it produces a source of repeated calls. Every such source after some delay produces repeated calls until after one or more attempts it finds a free server, in which case the source is eliminated and the call receives service and then leaves the system.

We assume that periods between successive retrials are exponentially distributed with parameter μ, and service times are exponentially distributed with parameter ν. Without loss of generality we may assume that $\nu = 1$.

4.2 Generator Matrix

The functioning of the system can be described by means of a bivariate process $(C(t), N(t))$, where $C(t)$ is the number of busy servers and $N(t)$ is the number of sources of repeated calls (queue length) at time t. Under the above assumptions the bivariate process $(C(t), N(t))$ is Markovian with the lattice semi-strip $S = \{0, 1, \ldots, c\} \times Z_+$as the state space. Its infinitesimal transition rates $q_{(ij)(nm)}$ are given by:

1. for $0 \leq i \leq c - 1$

$$q_{(ij)(nm)} = \begin{cases} \lambda, & \text{if } (n,m) = (i+1, j), \\ i, & \text{if } (n,m) = (i-1, j), \\ j\mu, & \text{if } (n,m) = (i+1, j-1), \\ -(\lambda + i + j\mu), & \text{if } (n,m) = (i,j), \\ 0 & \text{otherwise.} \end{cases}$$

2. for $i = c$

$$q_{(cj)(nm)} = \begin{cases} \lambda, & \text{if } (n,m) = (c, j+1) \\ c, & \text{if } (n,m) = (c-1, j) \\ -(\lambda + c), & \text{if } (n,m) = (c,j) \\ 0 & \text{otherwise.} \end{cases}$$

This assumption allows extensive mathematical analysis of both stationary and transient behavior of the process. In contrast to this, for the retrial queue under consideration (as well as for other retrial queues) rates of transition from a point (i, j) of the semistrip $\{0, 1, \ldots, c\} \times Z_+$depend on the second coordinate j. The main difficulties in analysis and the most interesting properties of retrial queues are connected with this fact.

From a practical point of view the most important characteristics of the quality of service to subscribers are:

- the stationary blocking probability $B = \lim_{t\to\infty} \mathrm{P}\{C(t) = c\}$;
- the mean queue length in the steady state $N = \lim_{t\to\infty} \mathrm{E}N(t)$;
- the stationary carried traffic (which is equal to the mean number of busy servers) $Y = \lim_{t\to\infty} \mathrm{E}C(t)$;
- the mean waiting time W, which by Little's formula equals $\frac{N}{\lambda}$;
- the mean waiting time for customers which are really waiting for service (i.e. their first attempt was blocked) $W_B = \frac{W}{B}$.

4.3 Special Case: $M/M/2$

4.3.1 Obtaining partial generating functions

Let

$$\begin{aligned} p_{0j} &= \mathrm{P}\{C(t) = 0, N(t) = j\} \\ p_{1j} &= \mathrm{P}\{C(t) = 1, N(t) = j\} \\ p_{2j} &= \mathrm{P}\{C(t) = 2, N(t) = j\} \end{aligned}$$

be the joint distribution of the number of busy servers and the number in orbit in the steady state. These probabilities satisfy the following set of Kolmogorov equations:

$$\begin{aligned}(\lambda + j\mu)p_{0j} &= p_{1j} \\ (\lambda + 1 + j\mu)p_{1j} &= \lambda p_{0j} + (j+1)\mu p_{0,j+1} + 2p_{2j} \\ (\lambda + 2)p_{2j} &= \lambda p_{1j} + (j+1)\mu p_{1,j+1} + \lambda p_{2,j-1}\end{aligned}$$

and the normalizing condition

$$\sum_{j=0}^{\infty}(p_{0j} + p_{1j} + p_{2j}) = 1$$

Equation (2.2) in fact expresses p_{1j} through p_{0j}. Using this, from (2.3) we can also express p_{2j} through p_{0j} :

$$2p_{2j} = \left[(\lambda + j\mu)^2 + j\mu\right] p_{0j} - (j+1)\mu p_{0,j+1}.$$

Substituting expressions (2.2) and (2.5) into (2.3) we get the following recursive relation for probabilities p_{0j} :

$$\begin{aligned}&\lambda\left[(\lambda + j\mu)^2 + j\mu\right] p_{0j} - (j+1)\mu[2 + 3\lambda + 2(j+1)\mu]p_{0,j+1} \\ =&\lambda\left[(\lambda + (j-1)\mu)^2 + (j-1)\mu\right] p_{0,j-1} - j\mu[2 + 3\lambda + 2j\mu]p_{0j}\end{aligned}$$

This yields that for all j we have:

$$\lambda\left[(\lambda + (j-1)\mu)^2 + (j-1)\mu\right] p_{0,j-1} - j\mu[2 + 3\lambda + 2j\mu]p_{0j} = 0$$

or equivalently,

$$p_{0j} = \frac{\lambda}{j\mu} \cdot \frac{(\lambda + (j-1)\mu)^2 + (j-1)\mu}{2 + 3\lambda + 2j\mu} \cdot p_{0,j-1}.$$

This formula allows us to express all probabilities p_{0j} through p_{00} :

$$p_{0j} = \frac{\lambda^j}{j!\mu^j} \cdot \prod_{k=0}^{j-1} \frac{(\lambda + k\mu)^2 + k\mu}{2 + 3\lambda + 2\mu + 2k\mu} \cdot p_{00}$$

Now we can find probabilities p_{1j} and p_{2j} :

$$p_{1j} = (\lambda + j\mu)\frac{\lambda^j}{j!\mu^j} \cdot \prod_{k=0}^{j-1} \frac{(\lambda + k\mu)^2 + k\mu}{2 + 3\lambda + 2\mu + 2k\mu} \cdot p_{00}$$

$$p_{2j} = (1 + \lambda + (j+1)\mu)\frac{\lambda^j}{j!\mu^j} \cdot \prod_{k=0}^{j} \frac{(\lambda + k\mu)^2 + k\mu}{2 + 3\lambda + 2\mu + 2k\mu} \cdot p_{00}$$

The initial probability p_{00} can be obtained with the help of the normalizing condition as

$$
\begin{aligned}
p_{00}^{-1} =& \sum_{j=0}^{\infty} \frac{\lambda^j}{j!\mu^j} \prod_{k=0}^{j-1} \frac{(\lambda+k\mu)^2+k\mu}{2+3\lambda+2\mu+2k\mu} \\
&\times \left[1+\lambda+j\mu+\frac{(1+\lambda+(j+1)\mu)\left((\lambda+j\mu)^2+j\mu\right)}{2+3\lambda+2(j+1)\mu} \right].
\end{aligned}
$$

In fact probability p_{00}, as well as generating functions $p_i(z) = \sum_{j=0}^{\infty} z^j p_{ij}$ and the main performance characteristics, can be expressed in terms of the hypergeometric functions

$$
F(a,b,c;x) \equiv \sum_{j=0}^{\infty} \frac{x^j}{j!} \prod_{k=0}^{j-1} \frac{(a+k)(b+k)}{c+k}
$$

4.3.2 Useful Theorems

Theorem 1
For the $M/M/2$ retrial queue, the joint distribution of the number of busy servers and the number of sources of repeated calls in the steady state is given by the following partial generating functions:

$$
\begin{aligned}
p_0(z) =& F\left(a,b,c;\frac{\lambda z}{2}\right) \cdot p_{00}, \\
p_1(z) =& \left\{ \lambda F\left(a,b,c;\frac{\lambda z}{2}\right) + \frac{\lambda^3}{2+3\lambda+2\mu} \right. \\
&\left. \times F\left(a+1,b+1,c+1;\frac{\lambda z}{2}\right) \right\} \cdot p_{00}, \\
p_2(z) =& \left\{ \frac{\lambda^2}{2-\lambda z} F\left(a,b,c;\frac{\lambda z}{2}\right) + \frac{\lambda^3(\lambda z-1)}{(2+3\lambda+2\mu)(2-\lambda z)} \right. \\
&\left. \times F\left(a+1,b+1,c+1;\frac{\lambda z}{2}\right) \right\} \cdot p_{00},
\end{aligned}
$$

where:

$$
\left.
\begin{aligned}
a &= \tfrac{2\lambda+1+\sqrt{4\lambda+1}}{2\mu}, \\
b &= \tfrac{2\lambda+1-\sqrt{4\lambda+1}}{2\mu}, \\
c &= \tfrac{2+3\lambda+2\mu}{2\mu},
\end{aligned}
\right\}
$$

Theorem 2
For the $M/M/2$ retrial queue, the blocking probability and the mean queue length are given by

$$B = \frac{\lambda^2 + (\lambda - 1)g}{2 + \lambda + g}$$
$$N = \frac{1 + \mu}{\mu} \cdot \frac{\lambda^3 + \left(\lambda^2 - 2\lambda + 2\right) g}{(2 - \lambda)(2 + \lambda + g)}$$

where

$$g = \frac{\lambda^3}{2 + 3\lambda + 2\mu} \cdot \frac{F\left(a + 1, b + 1, c + 1; \frac{\lambda}{2}\right)}{F\left(a, b, c; \frac{\lambda}{2}\right)}$$

and the variables a, b, c were defined in Theorem 1.

Using the above two Theorems, B and N can be computed for an $M/M/k$ retrial queue.
The proofs to the theorems is out of the scope of this project.

5 Simulations

Parameters: $\lambda = 5, = 7, \mu = 1$

Parameter	Expected Values	Simulated Values
Mean Sojourn Time	3	3.16
Blocking Probability	0.71	0.705
Mean No of jobs	15	15.85
Mean No of pings	3	3.02

Expected values vs scheduled values

6 Graphs

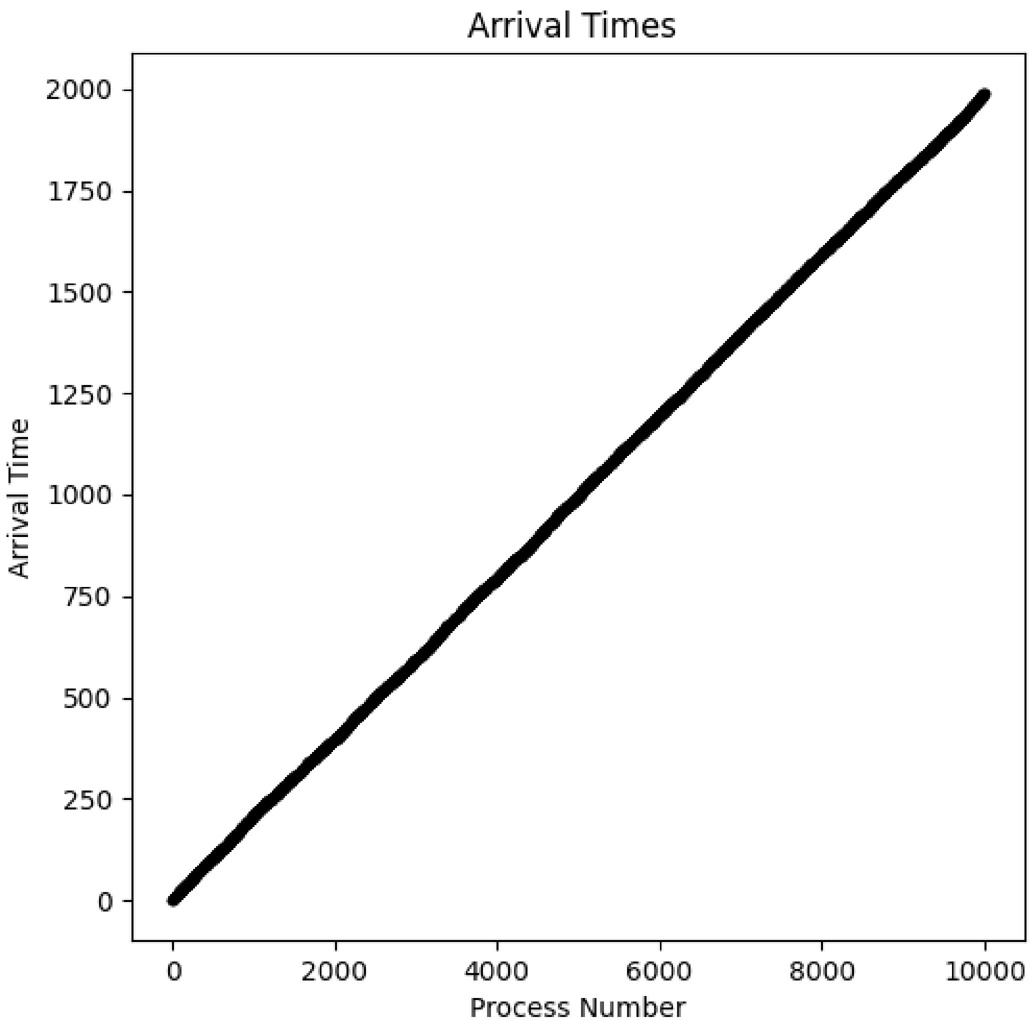

Figure 1: Arrival Times of the Jobs
The arrival times increase linearly with time.

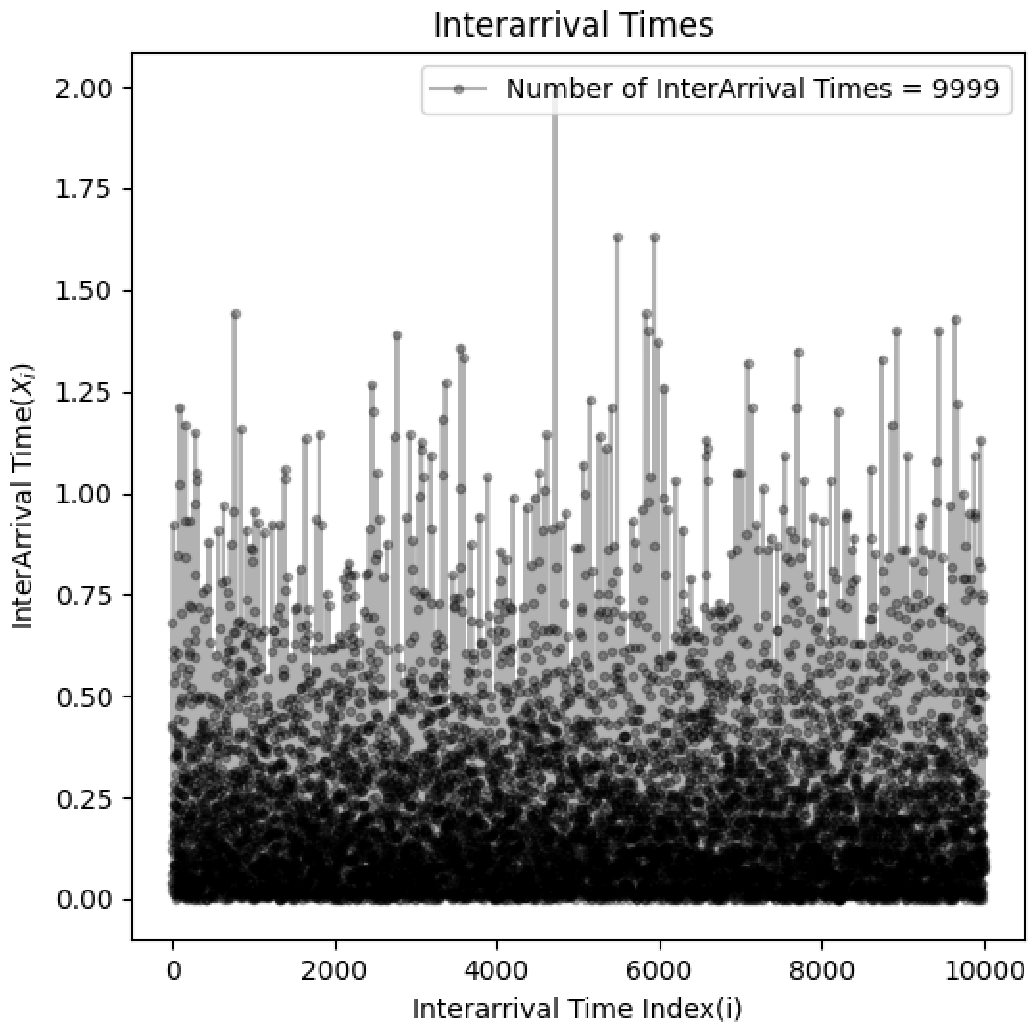

Figure 2: Interarrival Times

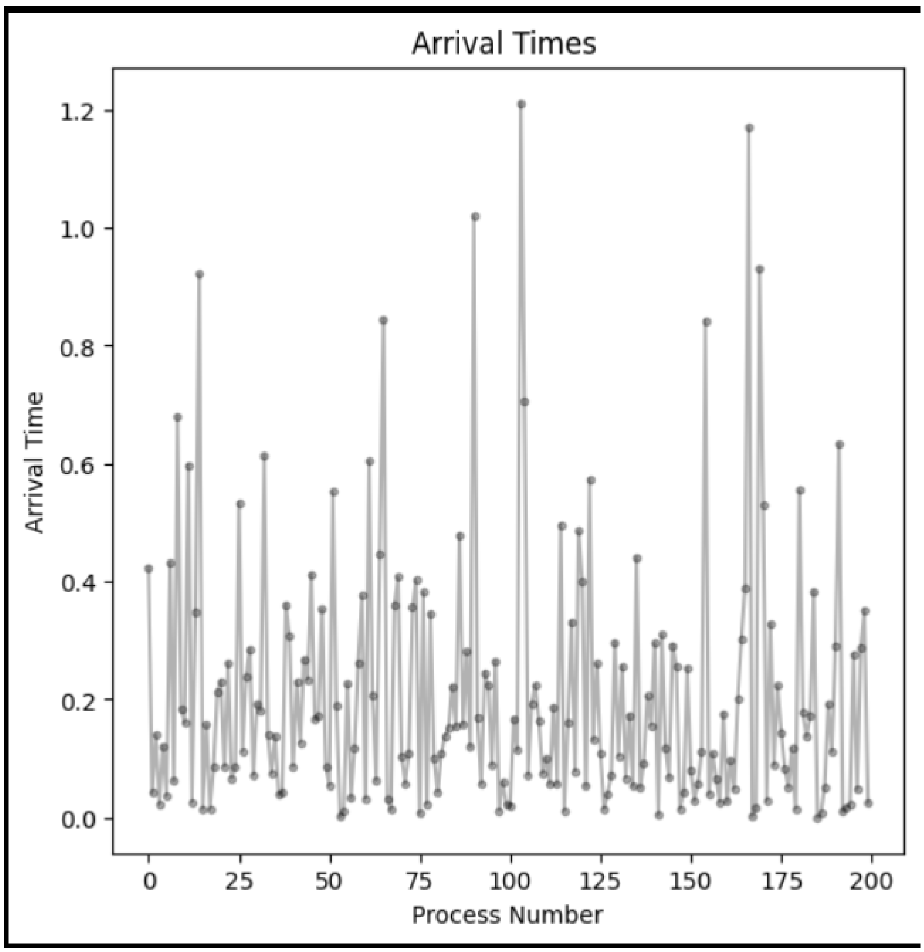

Figure 3: Your Image Caption
Clearly, the interarrival times are independent of eachother.

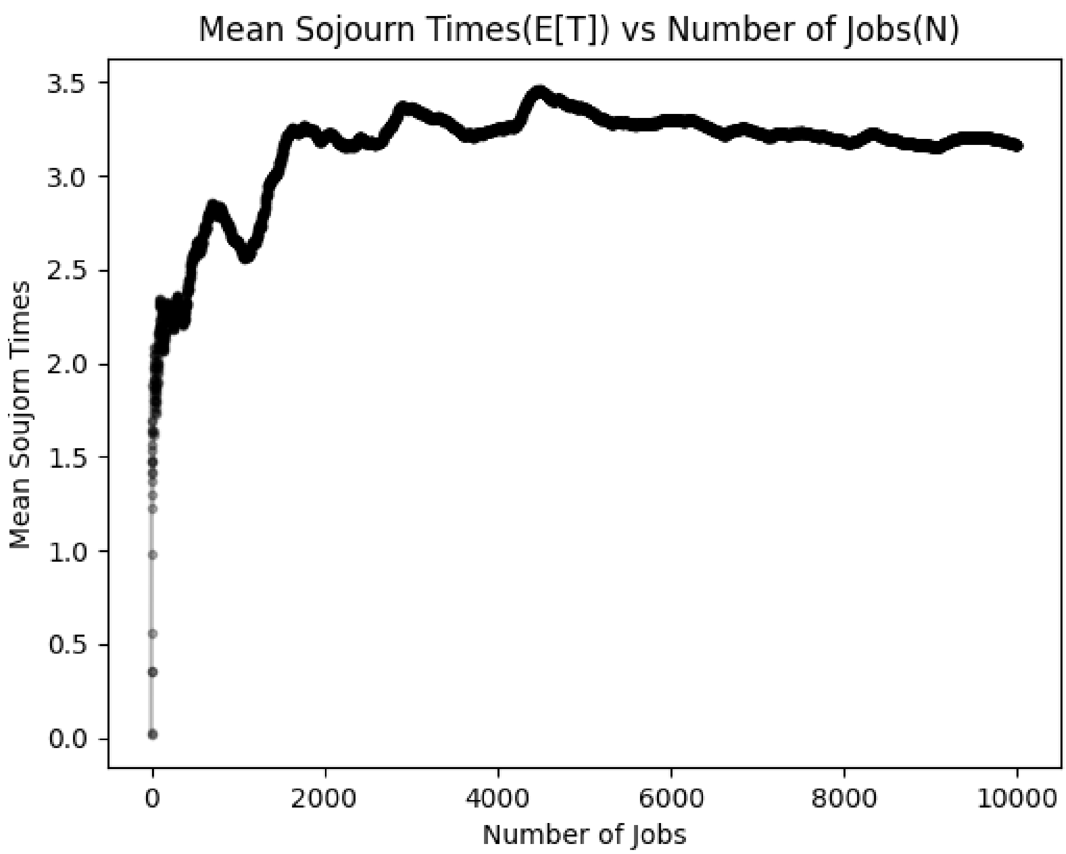

Figure 4: E[T] vs N
Mean Sojourn Time converges to the the value 3.16 as number of jobs arrived increases.

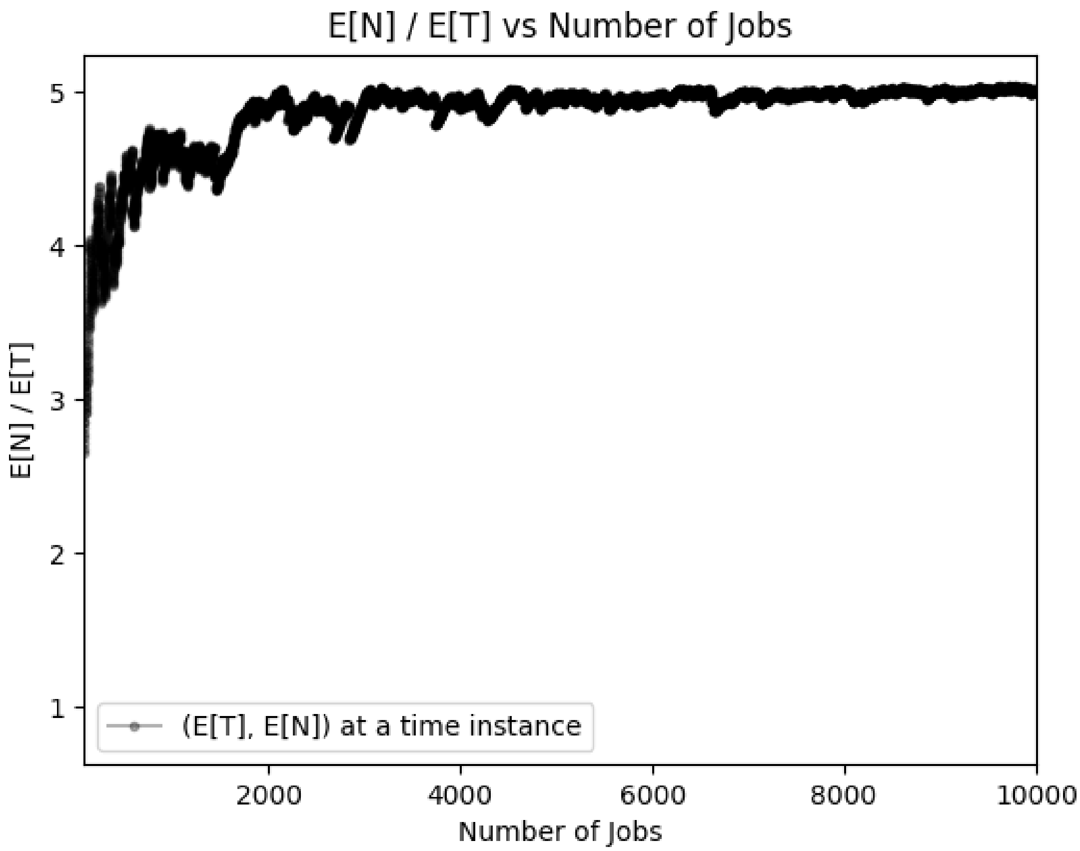

Figure 5: Verification of Little's Law
Little's Law is verified by the above plot as when the number of jobs arrived increases, the ratio

$$E[N]/E[T]$$

converges to λ

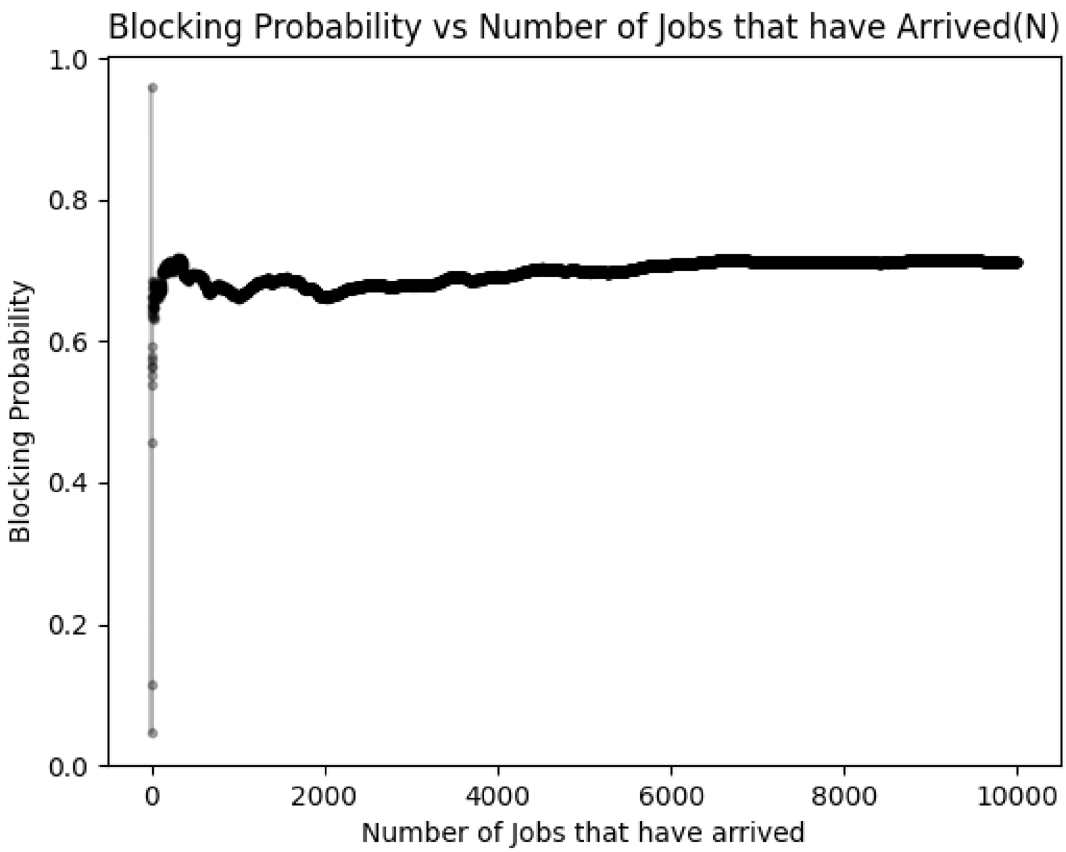

Figure 6: Blocking Probability vs Number of Jobs arrived(N)
The Blocking Probability stagnates to 0.70. Notice that the Blocking Probability does not depend on number of jobs that have arrived.