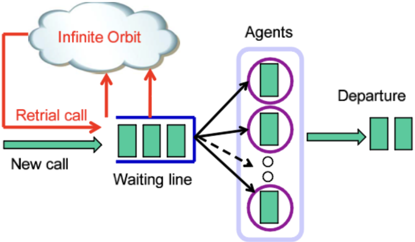
Retrial Queues

Tejas C., Sreyas A., Hardik K., Atidipt A., Aditya M.

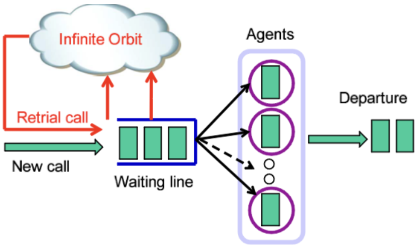
PMCS Project , Spring ’24

Analogy



Infinite Orbit
Agents
Departure
Retrial call
New call
Waiting line

Analogy



- Busy Call Centres
- TCP Packet Transmission
- LAN

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
- $k(z) = \sum_0^\infty k_n z^n = \beta(\lambda - \lambda z)$

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

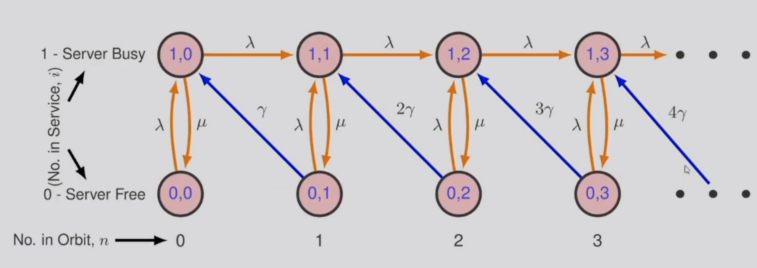
  is distribution of number of primary calls that arrive during service time of a call

M/M/1

- Service Time distribution

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

- State transitions



M/M/1 (State Transitions)

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.

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.

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.

M/M/1 (Limiting Distribution)

The statistical probability equations are given by

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

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

The partial generating functions are

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

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

$M/M/1$ (Performance Metrics)

- Mean number of jobs in queue

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

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

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

$$E[N(t)] = \sum n p_n = p'(1)$$

$M/M/1$ (Performance Metrics)

- Mean number of jobs in system

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

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

.

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

$M/M/1$ (Performance Metrics)

- Blocking probability

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

$P(\text{Server busy}) = \sum_n p_{1n} = p_(1)$

$M/M/1$ (Performance Metrics)

- Mean sojourn time

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

Found using Little's Law

$M/M/1$ (Performance Metrics)

- Recurrence Conditions

$$
\begin{aligned}
&\rho < 1 && \text{for positive recurrence} \\
&\rho = 1 && \text{for null recurrence} \\
&\rho > 1 && \text{for transience}
\end{aligned}
$$

Proved by examining mean sojourn time

$M/M/1$ (Performance Metrics)

- Mean no of retrials per job

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

If a job spends time T inside the system, it retries after X_i which are $Exp(\mu)$. The no of retries is a stopping time for X. $T = \sum_{i=0}^{R} X_i$

$$\implies E[T] = E[X]E[R]$$

$$\implies E[R] = \mu E[T]$$

$M/M/1$ (Embedded DTMC)

- General service times do not have the memoryless property.
- Convert 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.

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

- B_i is indicator for repeated calls
- ν_i is no of jobs that arrive during service

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

Embedded DTMC

- One-step transition probabilities $r_{mn} = \mathrm{P}\left\{N_i = n \mid N_{i-1} = m\right\}$ are given by the formula

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

- Mean Length of queue

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

Special Case: $M/M/2$

- Consider the basic case of multi server model (Taking ν=1)

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

be the limiting distributions. These probabilities satisfy the following statistic probability equations

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

and normalizing condition

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

Simulating $M/M/1$

- The arrival of jobs , the retrial of jobs in orbit and the processing of the jobs is simulated by generating random numbers.
- This gives us the pre-computed values of various parameters like arrival time , departure time, number of retrials etc . These will be used to compute various performence metrics.

Simulating $M/M/1$

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

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

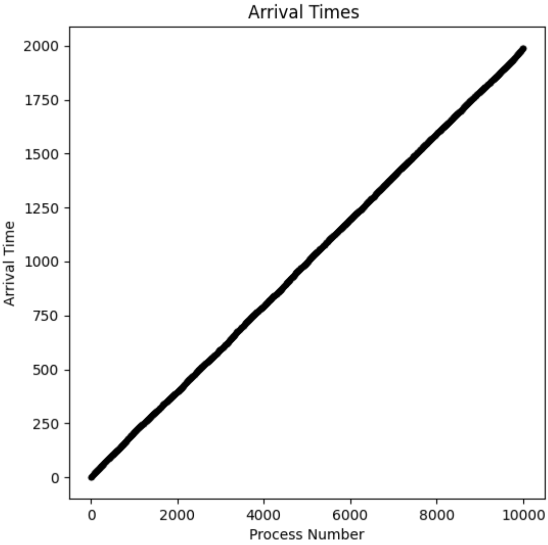


Figure: Arrival Times of the Jobs

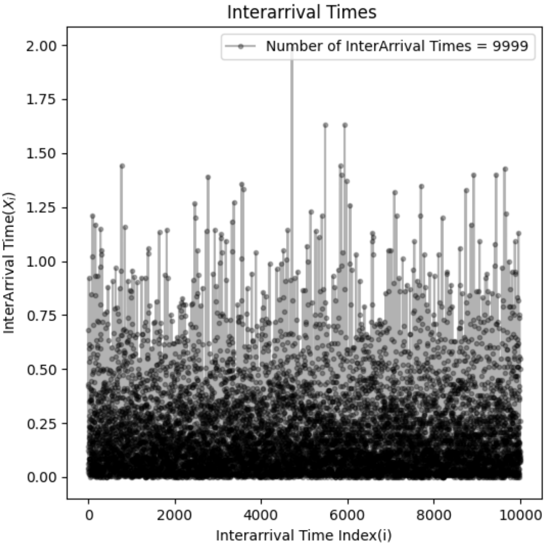


Figure: Interarrival Times

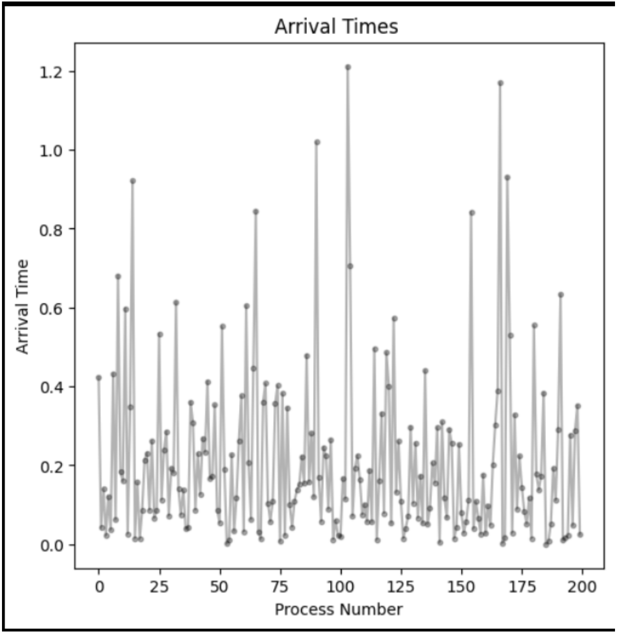


Figure: Interarrival Times

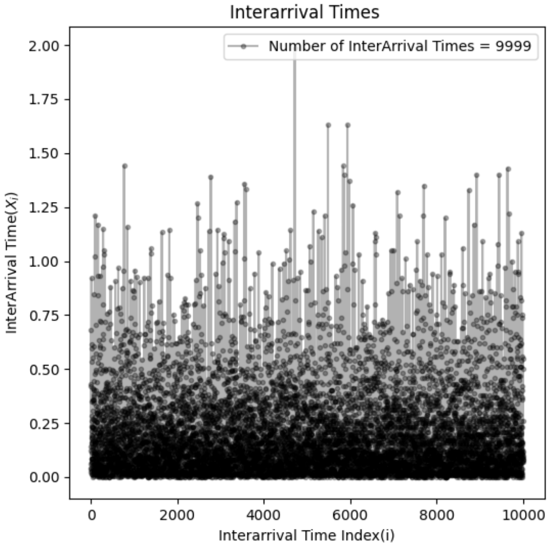


Figure: Interarrival Times

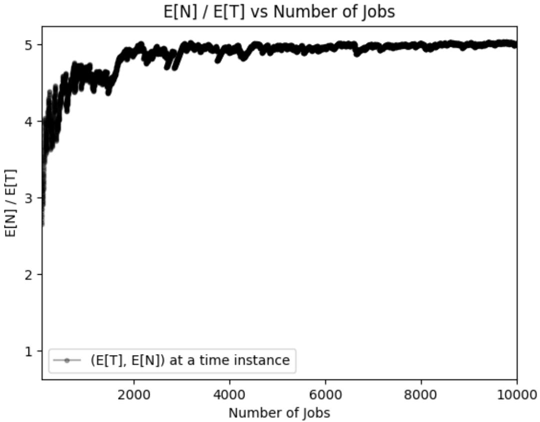


Figure: Verification of Little's Law

Little's Law is verified by:

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

converges to λ

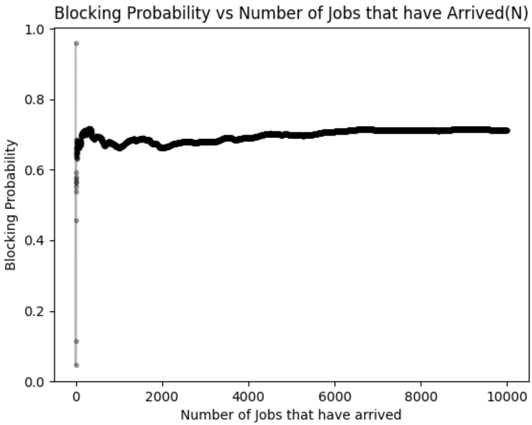


Figure: Blocking Probability vs Number of Jobs arrived(N)

The Blocking Probability stagnates to 0.70.