Performance of Computer System

M/M/1 Queuing Model

Dr. Lina Xu
lina.xu@ucd.ie

School of Computer Science,
University College Dublin

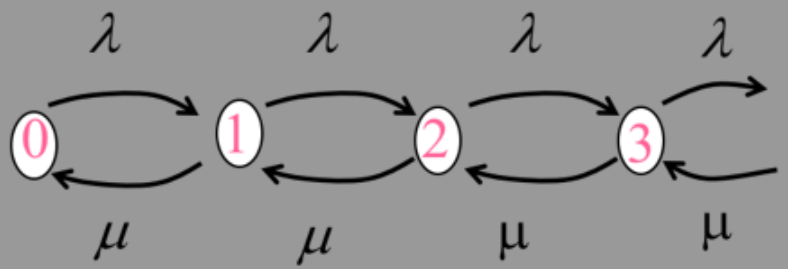
October 9, 2018

M/M/1 Queue

$$M \;/\; M \;/\; 1$$

arrival process	service process	# of servers

"*M*" stands for Markovian, meaning
exponential inter-arrival & exponential service times.



$\ldots$ so, $\lambda_k = \lambda, \quad k = 0,1,\ldots$

$\mu_k = \mu, \quad k = 1,2,\ldots$

where $\rho = \lambda/\mu$

Utilization factor or
traffic intensity

Derivation of Steady-State Probabilities

$$\text{Recall } \sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \text{ provided } |x| < 1$$

$$\text{Now } \pi_k = C_k \pi_0, \text{ where } \pi_0 = 1 / \sum_{k=0}^{\infty} C_k$$

$$\sum_{k=0}^{\infty} C_k = \sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k = \sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho}$$

Provided $\rho < 1$, i.e.,$\lambda < \mu$ Thus, $\pi_0 = 1 - \rho$ and $\pi_k = \rho_k(1-\rho)$.

Performance Measures for M/M/1 Queue

$$L = \frac{\lambda}{\mu - \lambda} \quad \text{provided } \lambda < \mu$$

$$L_q = L - (1 - \pi_0) = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

From Little's law, we know

$$W = \frac{1}{\lambda} L \quad \text{and} \quad W_q = \frac{1}{\lambda} L_q \quad \text{or}$$

$$W = \frac{1}{\mu - \lambda} \quad \text{and} \quad W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

Methodology vs. Formulas

The important thing is not the specific M/M/1 formulas, but the methodology used to find the results.

- Model the system as a birth-and-death process and construct the rate diagram.
- Depending on the system, defining the states may be the first challenge.
- Develop the balance equations.
- Solve the balance equations for π_k, $k = 0, 1, 2, \ldots$
- Use the steady-state distribution to derive L and L_q and, use Little?s law to get W and W_q .

M/M/1 Queue Example

On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per second (pps) and the gateway takes about two milliseconds to forward them.
Using an M/M/1 model, analyse the gateway.

- What is the probability of buffer overflow if the gateway had only 13 buffers?
- How many buffers do we need to keep packet loss below one packet per million?

M/M/1 Queue Example

On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per second (pps) and the gateway takes about two milliseconds to forward them.
Using an M/M/1 model, analyse the gateway.

- What is the probability of buffer overflow if the gateway had only 13 buffers?
- How many buffers do we need to keep packet loss below one packet per million?

Arrival rate $\lambda = 125$ pps

Service rate $\mu = 1/.002 = 500$ pps

Gateway Utilization $\rho = \lambda/\mu = 0.25$

Probability of n packets in the gateway $= (1-\rho)\ \rho^n = 0.75 * 0.25^n$

M/M/1 Queue Example

Mean Number of packets in the gateway

$= (\rho/(1-\rho) = 0.25/0.75 = 0.33)$

Mean time spent in the gateway

$= ((1/\mu)/(1-\rho) = (1/500)/(1-0.25) = 2.66$ milliseconds

Probability of buffer overflow

Probability > 13 packets

To limit the probability of loss to less than 10^{-6}:

$\rho^{n} <= 10^{-6}$

M/M/1 Queue Example

The last two results about buffer overflow are **approximate**.
Strictly speaking, the gateway should actually be modelled as a finite buffer **M/M/1/B** queue.
However, since the **utilisation is low** and the number of buffers is far above the mean queue length, the results obtained are a close approximation.

Example of M/M/1/2 Queue

A maintenance worker must keep 2 machines in working order.

- The 2 machines operate simultaneously when both are up.
- The time until a machine breaks has an exponential distribution with a mean of 10 hours.
- The repair time for the broken machine has an exponential distribution with a mean of 8 hours.
- The worker can only repair one machine at a time.

Evaluating Performance

- Model the system as a birth-and-death process.
- Develop the balance equations.
- Calculate the steady-state distribution π_k.
- Calculate and interpret L, L_q, W, W_q .
- What is the proportion of time the repairman is busy?
- What is the proportion of time that a given machine, e.g., machine No.1, is working ?

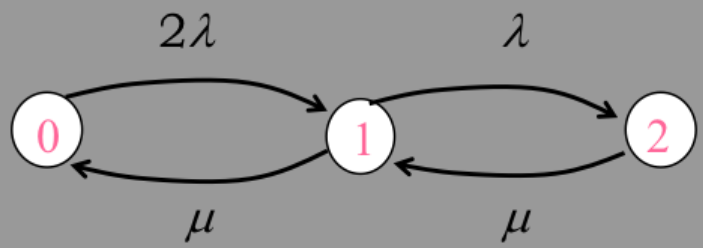
State -Transition Diagram

λ = **rate** at which a single machine breaks down
= 1/10 hr

μ = **rate** at which machines are repaired
= 1/8 hr

State of the system = # of broken machines.

Rate network:



Balance Equations for Repair Example

$$\mu\pi_1 = 2\lambda\pi_0 \qquad \text{state 0}$$

$$2\lambda\pi_0 + \mu\pi_2 = (\lambda + \mu)\pi_1 \qquad \text{state 1}$$

$$\lambda\pi_1 = \mu\pi_2 \qquad \text{state 2}$$

$$\pi_0 + \pi_1 + \pi_2 = 1 \qquad \text{normalize}$$

We can solve these balance equations for π_0, π_1 and π_2, but in this case, we can simply use the formulas for general birth-and-death processes:

$$C_k = \frac{\lambda_{k-1} \cdots \lambda_0}{\mu_k \cdots \mu_1} \qquad \pi_k = C_k \pi_0 \text{ and } \pi_0 = 1 \Big/ \sum_{k=0}^{2} C_k$$

Balance Equations for Repair Example

Here, $\lambda_0 = 2\lambda$ $\mu_1 = \mu$

$\lambda_1 = \lambda$ $\mu_2 = \mu$

$\lambda_2 = 0$

$C_1 = \frac{\lambda_0}{\mu_1} = \frac{2\lambda}{\mu}$, $C_2 = \frac{\lambda_1\lambda_0}{\mu_2\mu_1} = \frac{2\lambda^2}{\mu^2}$, and $C_0 = 1$ (by definition). Thus

$$\pi_0 = \frac{1}{1 + \frac{2\lambda}{\mu} + \frac{2\lambda^2}{\mu^2}} = 0.258, \quad \pi_1 = \frac{2\lambda}{\mu}\pi_0 = 0.412$$

$$\pi_2 = \frac{2\lambda^2}{\mu^2}\pi_0 = 0.330$$

$L = 0\pi_0 + 1\pi_1 + 2\pi_2 =$ **1.072** (avg # machines in system)

$L_q = 0\pi_1 + 1\pi_2 =$ **0.33** (avg # waiting for repair)

$$\text{average arrival rate } \bar{\lambda} = \sum_{k=0}^{2} \lambda_k \pi_k = \lambda_0 \pi_0 + \lambda_1 \pi_1 + \lambda_2 \pi_2$$

$$= (2\lambda)\pi_0 + \lambda \pi_1 = 0.0928$$

$$W = \frac{1}{\bar{\lambda}} L = \frac{1}{0.0928}(1.072) \qquad W_q = \frac{1}{\bar{\lambda}} L_q = \frac{1}{0.0928}(0.33)$$

$= \mathbf{11.55}$ hours

$= \mathbf{3.56}$ hours

Average amount of time that a machine has to wait to be repaired, including the time until the repairman initiates the work.

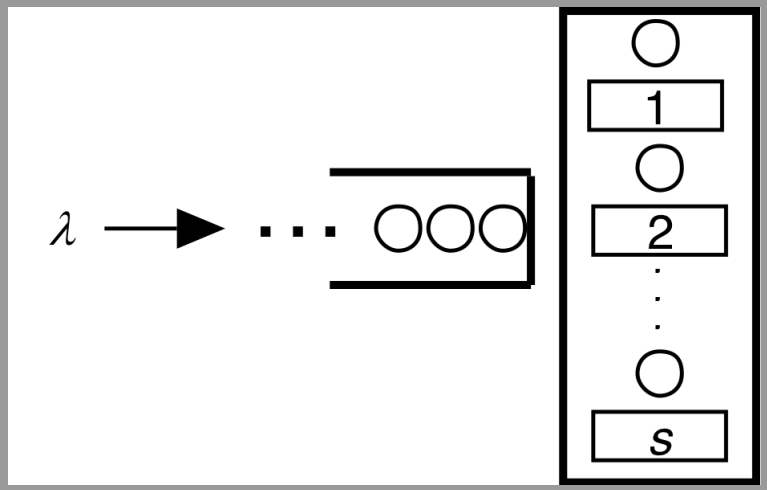
Average amount of time that a machine has to wait until the repairman initiates the work.

Proportion of time repairman is busy $= \pi_1 + \pi_2 = 0.742$

Proportion of time that machine #1 is working

$$= \pi_0 + \frac{1}{2}\pi_1 = 0.258 + \frac{1}{2}(0.412) = 0.464$$

Multi-Channel Queues - M/M/s



- Customers arrive according to a Poisson process with rate λ.
- The service times of customers are exponentially distributed with parameter μ.
- There are s servers, serving customers in order of arrival.

Stability condition: $\lambda < s\mu$

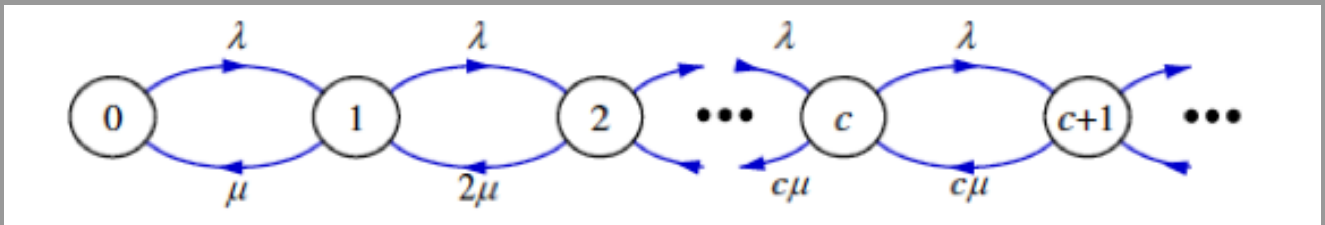
M/M/s: New Customer

Arriving customer finds n customers in system

- $n < c$: it is routed to any idle server
- $n \geq c$: it joins the waiting queue since all servers are busy

Birth-death process with state-dependent death rates

- $\mu_n = n\mu$, $1 \leq n \leq s$
- $\mu_n = c\mu$, $n \geq s$



Telephone Answering System Example

Situation:

- A utility company wants to determine a staffing plan for its customer representatives.
- Calls arrive at an average rate of 10 per minute, and it takes an average of 1 minute to respond to each inquiry.
- Both arrival and service processes are Poisson.

Problem:

- Determine the number of operators that would provide a "satisfactory" level of service to the calling population.

Analysis:

- *lambda* $= 10$, $1/\mu = 1$; $\rho = \lambda/(s\mu) < 1$ or $s > \lambda/\mu = 10$

Comparison of Multi-Server Systems

Measure	$M/M/11$	$M/M/12$	$M/M/13$
L_q	6.821	2.247	0.951
W_q	0.682	0.225	0.0.95
E	0.909	0.833	0.767
$\Pr\{ T_q = 0 \}$	0.318	0.551	0.715
$\Pr\{ T_q > 1 \}$	0.251	0.061	0.014

Machine Processing with Limited Space for Work in Process

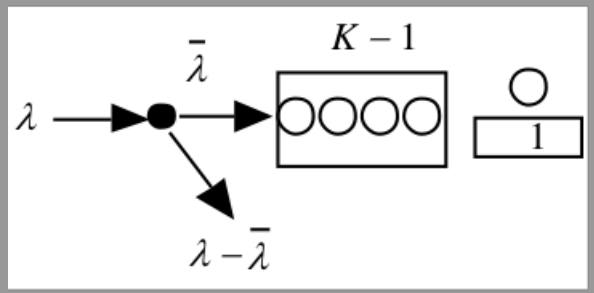
- Parts arrive at a machine station at the rate of 1.5/min on average.
- The mean time for service is 30 seconds.
- Both processes are assumed to be Poisson.
- When the machine is busy, parts queue up until there are 3 waiting. At that point arrivals sent for alternative processing.

Goals: Analyse the situation under the criteria that no more than 5% of arriving parts receive alternative processing and that no more than 10% of the parts that are serviced directly spend more than 1 minute in the queue.

Parameters for Machine Processing Queuing System

- Number of servers: $s = 1$
- Maximum number in system: $K = 4$
- Size of calling population: $N = \infty$
- Arrival rate: $\lambda = 1.5/\text{min}$
- Service rate: $\mu = 2/\text{min}$
- Utilisation: $\rho = \lambda/\mu = 0.75$
- Model: M/M/1/4

M/M/1/K Queue, $\rho \neq 1$



Measure	$M/M/1/K$	$M/M/1$
ρ	λ/μ	λ/μ
π_0	$\frac{1-\rho}{1-\rho^{K+1}}$	$1-\rho$
$\pi_n,\ 1 \le n \le K$	$\pi_0\rho^n$	$\pi_0\rho^n$
π_K	$\pi_0\rho^K$	0
P_B	$\frac{\rho(1-\rho^K)}{1-\rho^{K+1}}$	ρ
L_q	$\frac{\rho^2[(K-1)\rho^K - K\rho^{K-1}+1]}{(1-\rho)(1-\rho^{K+1})}$	$\frac{\rho^2}{1-\rho}$
L_s	$\rho(1-\pi_K)$	ρ
L	$\frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$	$\frac{\rho}{1-\rho}$
$\bar{\lambda}$	$\lambda(1-\pi_K)$	λ
E	$\rho(1-\pi_K)$	ρ

Solution for M/M/1/4 Model

Parameters

$\lambda = 1.5/\text{min}, \ \mu = 2/\text{min}$

$\rho = \lambda/\mu = 0.75$

$K = 4$

Steady state probabilities

$$\pi_0 = (1-\rho)/(1-\rho^{K+1}) = (1-0.75)/(1-0.75^5) = 0.25/0.7627$$
$$= 0.328$$

$$\pi_k = \pi_0 \rho^k = (0.328)(0.75^k); \ k = 1,\ldots,K$$

$$\pi_1 = 0.246, \ \pi_2 = 0.184, \ \pi_3 = 0.138, \ \pi_4 = 0.104$$

Solution for M/M/1/4 Model (cont'd)

Balking probability: $P_F = \pi_4 = 0.104$ or 10.4%

(does not meet 5% goal)

Average arrival rate: $\lambda = \lambda(1 - P_F) = 1.344/\text{min}$

$L = 1.444$

$W = 1.074 \text{ min}$

$E = \rho(1 - P_F) = 0.75(1 - 0.104) = 0.672$ or 67.2%

$\Pr\{T_q > 1\} = 0.225$

Add Second Machine - M/M/2/5 Model

New results

$P_F = 0.0068, \quad \lambda = \lambda(1 - \pi_5) = 1.49$

$L = 0.85, \quad W = 0.57 \text{ min}$

$E = 37.2\%$

$\Pr\{T_q > 1\} = 0.049 \quad \text{or} \quad 4.9\%$

This solution meets our original goals with the percentage of balking parts now less than 1% and the probability of a wait time greater than 1 minute less than 5%

Realistic Problems

Queue Scheduling

- First in first out
- Last in first out
- Processor sharing
- Priority
- Shortest job first
- Shortest remaining processing time
- Service facility

Realistic Problems

Service facility

- Single server: customers line up and there is only one server
- Parallel servers: customers line up and there are several servers
- Tandem queue: there are many counters and customers can decide going where to queue

Realistic Problems

Customer's behaviour of waiting

- Balking: customers deciding not to join the queue if it is too long
- Jockeying: customers switch between queues if they think they will get served faster by so doing
- Reneging: customers leave the queue if they have waited too long for service

It Is Useful...

Examples

- banks/supermarkets - waiting for service
- computers - waiting for a response
- failure situations - waiting for a failure to occur e.g. in a piece of machinery
- public transport - waiting for a train or a bus

It Is Useful...

What we can **Answer**?

- How long does a customer expect to wait in the queue before they are served, and how long will they have to wait before the service is complete?
- What is the probability of a customer having to wait longer than a given time interval before they are served?
- What is the average length of the queue?
- What is the probability that the queue will exceed a certain length?
- What is the expected utilisation of the server and the expected time period during which he will be fully occupied (remember servers cost us money so we need to keep them busy). In fact if we can assign costs to factors such as customer waiting time and server idle time then we can investigate how to design a system at minimum total cost.

It Is Useful...

What we can **Decide**?

- Is it worthwhile to invest effort in reducing the service time?
- How many servers should be employed?
- Should priorities for certain types of customers be introduced?
- Is the waiting area for customers adequate?