NTNU | Norwegian University of Science and Technology

QUEUE THEORY

Jørn Vatn/September-2025

Iintroduction

Queue theory deals with modelling the situation where one or more “servers" serve “customers" arriving for service

- X_i = Inter-arrival times
- S_i = service times
- Inter-arrival and service times are governed by so-called *arrival processes* and *service processes* respectively

A letter is used to specify the distributions, and the following convention is usually applied:

- M for exponentially distributed variables, i.e., the Markov property holds
- E_k for Erlang-k distributed variables
- D for deterministic or fixed variables
- GI for general distributed interarrival times
- G for general distributed service times

Notation

A special notation is used for specification of queue models. A set of parameters are separated by the the slash symbol ("/") where the order of variables are:

1. The arrival process, M, E_k, D or GI.
2. The service process, M, E_k, D or G.
3. The # of servers
4. Limit on # in system. Default is infinite
5. # in the source. Default is infinite
6. Queue discipline. Default: first come, first served (FCFS)

Example: $M/M/1/3/\infty/\mathrm{FCFS}$

Objectives of the modelling

- Determine optimal number of servers
- Determine if it pays off to increase the maximum number of customers allowed in the system

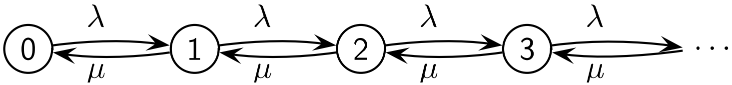
Performance

When analysing queue models the most interesting output variables to consider are:

- $P_i(t)$ = Probability of being in the various system states
- L = Expected number in the system under steady-state conditions.
- L_q = Expected number in the queue under steady-state conditions, i.e., customers being served are excluded.
- W expected time spent in the system by a customer under steady-state conditions.
- W_q expected time spent in the queue by a customer under steady-state conditions.

The $M/M/1$ queue

- For the $M/M/1$ queue we assume that customers arrive according to a Markov process
- The rate of arrivals is constant equal to λ independent of the number of customers being served
- There is only one server which has a completion rate of μ



The $M/M/1$ queue, cont

The transition matrix $\mathbf{A}$ is thus given on the following form where indexing of $\mathbf{A}$ starts at 0:

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & 0 & \dots & \dots & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \dots & \dots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \dots \\ \vdots & & & & & \end{bmatrix}$$

The $M/M/1$ queue, steady state solution

To obtain the steady state solution we have

$$-\lambda P_0 + \mu P_1 = 0$$

for the first equation, and

$$\lambda P_0 - (\lambda + \mu) P_1 + \mu P_2 = 0$$

for the second equation, and in the general case we have:

$$\lambda P_{j-1} - (\lambda + \mu) P_j + \mu P_{j+1} = 0, j > 0$$

The $M/M/1$ queue, steady state solution

We now obtain the steady state probabilities by:

$$P_1 = \frac{\lambda}{\mu} P_0$$

and by substituting P_1 into the second equation:

$$P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0$$

and in general

$$P_j = \left(\frac{\lambda}{\mu}\right)^j P_0$$

The $M/M/1$ queue, steady state solution

Since all probabilities sum to unity, we have

$$1 = P_0 + P_1 + \cdots = P_0 \left[1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2 + \dots \right]$$

Inserting the general rule $\sum_{k=0}^{\infty} r^k = 1/(1-r)$ into the brackets gives $1/(1-\frac{\lambda}{\mu})$ provided that $\lambda/\mu < 1$. Hence,

$$1 = P_0/(1-\frac{\lambda}{\mu})$$

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

$\rho = \frac{\lambda}{\mu}$ = *traffic intensity*

The $M/M/1$ queue, steady state solution & # of customers

Substituting for the other P_j's we get:

$$P_j = \left(\frac{\lambda}{\mu}\right)^j (1 - \frac{\lambda}{\mu}) = \rho^j (1 - \rho)$$

The mean number of customers in the system is given by:

$$L = \sum_{j=0}^{\infty} jP_j = \sum_{j=1}^{\infty} j\rho^j (1 - \rho) = \rho(1 - \rho) \sum_{j=1}^{\infty} j\rho^{j-1} = \frac{\rho}{1 - \rho}$$

$\left(\sum_{k=1}^{\infty} kr^{k-1} = 1/(1 - r)^2 \text{ for } r < 1\right)$

The $M/M/1$ queue, # of customers

To find the expected number of customers in the *queue*, L_q:

- We distinguish between those customers that are in the queue, and those being served
- Since there is only one server we have the following:
  - If there are zero or one customers, there are no customers in the queue,
  - If there are $j > 1$ customers in the system there will be $j-1$ customers in the queue

$$
\begin{aligned}
L_q &= 0(P_0 + P_1) + 1P_2 + 2P_3 + 3P_4 + \dots \\
&= 1\rho^2(1-\rho) + 2\rho^3(1-\rho) + 3\rho^4(1-\rho) + \dots \\
&= \rho^2(1-\rho)\left[1\rho^0 + 2\rho^1 + 3\rho^2 + \dots\right] = \rho^2/(1-\rho)
\end{aligned}
$$

The $M/M/1$ queue, waiting times

The waiting time, W is the expected time a customer stays in the queue, i.e., from he or she arrives until service is completed. It follows that:

$$W = \sum_{j=0}^{\infty}(j+1)\frac{1}{\mu}P_j = \frac{1}{\mu}\left(\sum_{j=0}^{\infty} jP_j + 1\right) = \frac{1}{\mu}\left(\frac{\rho}{1-\rho} + 1\right) = \frac{1}{\mu(1-\rho)}$$

Inserting $\mu = \lambda/\rho$ we observe that $W = L/\lambda$ which is known as *Little's formula*. Although not proved here, we have:

The $M/M/1$ queue, waiting times, cont

In general we can use *Little's formula* to obtain waiting times. The formula reads:

$$L = \lambda W$$

or, equivalently:

$$W = L/\lambda, \text{ For : } M/M/1 = \frac{1}{\mu(1-\rho)}$$

A similar result also applies for the mean time spent in the queue, i.e.,

$$W_q = L_q/\lambda, \text{ For : } M/M/1 = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\rho}{\mu(1-\rho)}$$

The $M/M/1/N$ queue

The $M/M/1/N$ queue differs from the $M/M/1$ queue in that there is a limited capacity to hold new customers. The transition matrix is now finite, and given by:

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & 0 & \dots & \dots & \dots & 0 \\ \mu & -(\lambda+\mu) & \lambda & 0 & \dots & \dots & 0 \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \dots & 0 \\ \vdots & & & & & & \\ 0 & \dots & \dots & \dots & \dots & \mu & -\mu \end{bmatrix}$$

The $M/M/1/N$ queue, results

$$P_j = \rho^j \frac{1-\rho}{1-\rho^{N+1}}, \rho \neq 1$$

and

$$P_j = \frac{1}{N+1}, \rho = 1$$

The probability that the system is in a state where the waiting room is full is given by P_N, hence the rate of lost customers equals

$$\eta(N) = \lambda \rho^N \frac{1-\rho}{1-\rho^{N+1}}, \rho \neq 1$$

The $M/M/C/N$ queue

The $M/M/C/N$ queue opens for more than one server which affects the below-diagonal elements:

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & 0 & \dots & \dots & \dots & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \dots & \dots & \dots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & \dots & \dots \\ 0 & 0 & 3\mu & -(\lambda+3\mu) & \lambda & 0 & \dots \\ \vdots & & & & & & \end{bmatrix}$$

VBA code for the $M/M/C/N$ queue

```
Function P_MMCN(lambda , mu , C , N )
Const o As Integer=1
For i = 0 To N - 1
  A(i + o, i + o + 1) = lambda
Next i
For i = 1 To N
  If i >= C Then
    A(i + o, i + o - 1) = C * mu
  Else
    A(i + o, i + o - 1) = i * mu
  End If
Next
P_MMCN = MarkovAsymptotic(A)
End Function
```

The $M/M/C/N$ queue

Use the VBA code or other code to verify the special case for $C = N$ we have:

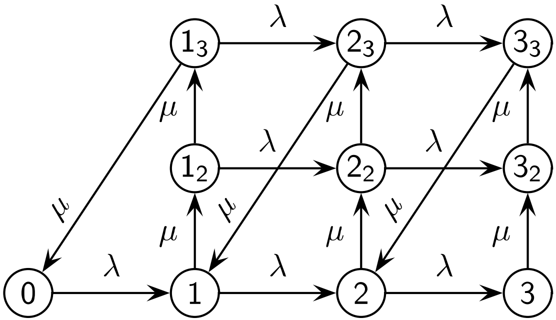
$$P_j = \frac{\rho^j/j!}{\sum_{i=0}^{C} \rho^i/i!}$$

▸ VBA code in Excel

The $M/E_k/1/N$ queue

- Assume that service times are Erlang-k distributed
- Erlang-k corresponds to a sum of k exponential distributed stochastic variables
- We may think of this as the server is then conducting k independent tasks
- In the modelling we now introduce $k - 1$ additional artificial states to each “physical " state j
- Let the artificial states linked to state $j = 1$ be denoted $1_2, 1_3, \ldots, 1_k$
- Similarly the artificial states linked to state $j = 2$ are denoted $2_2, 2_3, \ldots, 2_k$ etc

The $M/E_3/1/3$ queue



A state with no index should be renamed with index 1, i.e., state 1 is renamed to 1_1 when implemented in the computer code.

The $M/E_k/1/N$ queue, Markov equations

We now have the following transitions in the system:

$$
\begin{aligned}
a_{j_l,(j+1)_l} &= \lambda, 0 \le j < N, 1 \le l \le k \\
a_{j_l,j_{l+1}} &= \mu, 1 \le j \le N, 1 \le l < k \\
a_{j_k,(j-1)_1} &= \mu, 1 \le j \le N
\end{aligned}
$$

where elements not specified equal zero

The $M/E_k/1/N$ queue, converting regime

To hold all transition rates we need an $(N+1)k$ times $(N+1)k$ matrix. We may use the index conversion formula for an index j_l:

$$\text{INDEX}(j, l) = kj + l$$

Now, assume that the transition matrix has been specified by using the indexing formula. Let further $\tilde{\mathbf{P}}$ be the resulting vector of steady state probabilities from the $(N+1)k$ times $(N+1)k$ system with artificial nodes. There are k elements in $\tilde{\mathbf{P}}$ that represent the “physical" state j:

$$P_j = \sum_{l=kj+1}^{k(j+1)} \tilde{\mathbf{P}}_l$$

The expected number of customers in the system is given by:

$$L = \sum_{j=1}^{N} j \left[\sum_{l=kj+1}^{k(j+1)} \tilde{\mathbf{P}}_l \right]$$

The $M/E_k/1/N$ queue, Pollaczek-Khinchine

It is relatively easy to apply these formulas i to find a numerical solution to the expected number of customers. The result may then be compared to the famous Pollaczek-Khinchine formula that applies for an $M/G/1$ queue. The formula reads:

$$L = \rho + \frac{\rho^2 + \lambda^2 \mathrm{Var}(S)}{2(1-\rho)}$$

where $\mathrm{Var}(S)$ is the variance of the service times. Note that for the Erlang-k distribution with parameters k and β the mean value and variance is given by k/β and k/β^2 respectively. Here “β” represents the intensity in each “subtask" so that $\beta = k\mu$ where μ is the intensity for the “entire task". Hence, $\mathrm{Var}(S) = 1/(k\mu^2)$ which could be inserted in the Pollaczek-Khinchine equation.

Thank you for your attention

NTNU | Norwegian University of Science and Technology