Performance Evaluation of Queuing Systems

Introduction to Queuing Systems
System Performance Measures & Little's Law
Equilibrium Solution of Birth-Death Processes
Analysis of Single-Station Queuing Systems
Single station queuing models are useful for performance evaluation of many devices, sub-systems and communications nodes that make up communications networks.

We introduce the different types of queuing systems that are possible and then present Birth-Death models that allow us to compute performance measures, such as mean queue length and delay, for such systems.

**Queuing System Models**

In a single station queuing model, the queuing system consists of a buffer for queuing customers and one or more identical servers. The queue may be of zero, finite or infinite size.
A server may only serve one “customer” (packet, call, request, etc.) at a time and so, at any given time, it is either “busy” or “idle”.

A customer represents some unit of work that keeps a server busy for some amount of time - (e.g. transmission of a packet of some length, carrying of a call for some amount of time, processing of a request, etc.).

If all servers in the system are busy when a new customer arrives then the customer joins the queue, if there is remaining space in the queue.

When a customer finishes service, it departs the system and a new customer is selected from the queue (if there are any waiting) to begin its service.

There are two sources of randomness in such a system:

1. customers arrive at random times, that is, the *inter-arrival-time* of customers is described by a random variable.
2. the amount of time required to service a customer is random, that is, the *service time* is described by a random variable.

It is often assumed that arrivals are independent events and that the service times of different customers are also independent.
Kendall’s Notation for Queues

We use the following notation to describe different types of queuing systems:

\[ A/B/m/K/p – queuing \text{ discipline} \]

\( A \) is the distribution of inter-arrival times and \( B \) is the distribution of service times.

The distributions \( A \) and \( B \) may be indicated as:

- \( M \) - Exponentially distributed inter-arrival times or service times (where the \( M \) indicates Memoryless)
- \( D \) - Deterministic distribution (that is, constant)
- \( G \) - General distribution (of unknown distribution)
- \( GI \) - General independent (unknown distribution but with independent arrivals and service times).

For the distributions \( A \) and \( B \) above, the mean arrival rate is normally denoted as \( \lambda \) and the mean service time as \( 1/\mu \), that is, \( \mu \) is the mean service rate.
Queuing Systems - Kendall’s Notation

$m$ indicates the number of servers in the system.

$K$ indicates the capacity of the overall system (the number of queuing places plus the number of servers). If $K$ is not specified in the Kendall notation, then the buffering capacity (queue length) is taken as infinite.

$p$ indicates the number in the population of customers that may arrive to the system, that is, the number of users of the system. If $p$ is not specified in the notation, then the population is taken as infinite.

The queuing discipline determines which customer is selected from the queue for processing when a server becomes available. Examples of different queuing disciplines are:

- **FCFS** - First-Come-First-Served - If no queuing discipline is stated in the Kendall description, this one is assumed. All the systems we consider have a FCFS queuing discipline.
- **LCFS** - Last-Come-First-Served
- **PS** - Processor-Sharing: All customers are served simultaneously where processing of all customers is equally spread across all servers.
Performance Evaluation of Queuing Systems

Introduction to Queuing Systems

System Performance Measures & Little’s Law

Equilibrium Solution of Birth-Death Processes

Analysis of Single-Station Queuing Systems
We now introduce some notation for the performance measures of interest in queuing systems.

**Number of Customers in the System:** In steady-state, the expected value of the state distribution vector \( \{ \pi_k \} \) gives the mean number of customers in the system. We will find a solution for the stationary distribution \( \{ \pi_k \} \) for continuous-time Birth-Death processes (queuing systems) later in this section. Given \( \{ \pi_k \} \), many of the other performance measures of interest may be derived.

**Utilisation (\( \rho \)):** For a queuing system with a single server, utilisation \( \rho \) is the fraction of time the server is busy. When there is no limit on the capacity of the system, then

\[
\rho = \frac{\text{mean service time}}{\text{mean inter-arrival time}} = \frac{\text{arrival rate}}{\text{service rate}} = \frac{\lambda}{\mu}.
\]

The utilization \( \rho \) when there are multiple servers, is the mean fraction of busy servers. Since \( m\mu \) is the overall service rate, in this case we have

\[
\rho = \frac{\lambda}{m\mu}.
\]

For a stable (ergodic) system, the condition for stability is \( \rho < 1 \).
Queuing Systems - Performance Measures

- **Throughput** ($\lambda$): The throughput for a queuing system with infinite capacity is the mean number of customers processed in a unit of time, i.e. the departure rate. Since the departure rate is equal to the arrival rate (and assuming $\rho < 1$), the throughput is

  $$\lambda = m \cdot \rho \cdot \mu.$$ 

  For a queuing system with finite capacity, there can be loss in the systems, and so the throughput can be less than the arrival rate. In this case, throughput is often denoted differently (e.g. as $S$) to distinguish it from the arrival rate $\lambda$.

- **Response Time** ($T$): (or sojourn time) is the total time a customer spends in the system.

- **Waiting Time** ($W$): is the time a job spends in the queue waiting to be serviced. Therefore, response time is the sum of the waiting time and the service time for a customer. $T$ and $W$ are, in general, random variables, so the expected values ($\bar{T}$ and $\bar{W}$) are often used and we may write:

  $$\bar{T} = \bar{W} + \frac{1}{\mu}.$$ 

- **Number of Customers in the System** ($N$): Again $N$ is a random variable and its expected value $\bar{N}$ is often of interest:

  $$\bar{N} = \sum_{k=1}^{\infty} k \cdot \pi_k.$$
Little’s Law

The mean number of customers in the system \( \bar{N} \) and the mean response time \( \bar{T} \) can be related using one of the most important theorems of queuing theory, Little’s Law:

\[
\bar{N} = \lambda \bar{T}
\]

We outline a proof of the Little’s theorem as follows:

Let \( \alpha(t) \triangleq \) the number of arrivals in the time interval \((0, t)\)
Let \( \delta(t) \triangleq \) the number of departures in the time interval \((0, t)\)
Little’s Law

We define $N(t)$ as the stochastic process representing the number in the system at time $t$, where

$$N(t) = \alpha(t) - \delta(t).$$

Let $\gamma(t) = \text{the total time all customers have spent in the system up to time } t = \text{the area between the two graphs } \alpha(t) \text{ and } \delta(t)$.

Let $\lambda_t = \text{the average arrival rate during interval } (0, t)$.

$$\lambda_t \triangleq \frac{\alpha(t)}{t} \quad (1)$$

Let $\bar{T}_t = \text{the average total time spent by a customer in the system over the interval } (0, t)$

$$\bar{T}_t = \frac{\gamma(t)}{\alpha(t)} \quad (2)$$

Let $\bar{N}_t = \text{the average number of customers in the system (during interval } (0, t))$

$$\bar{N}_t = \frac{\gamma(t)}{t} \quad (3)$$
Little’s Law

Equations (1),(2) and (3) can be combined:

\[
\frac{\gamma(t)}{t} = \frac{\alpha(t)}{t} \cdot \frac{\gamma(t)}{\alpha(t)}
\]

\[
\bar{N}_t = \lambda_t \bar{T}_t
\]

Let \( \lambda = \lim_{t \to \infty} \lambda_t \) and \( \bar{T} = \lim_{t \to \infty} \bar{T}_t \)

\( \bar{N} \) will then also exist giving \( \bar{N} = \lambda \bar{T} \).

The above result does not depend on the inter-arrival or service time distributions or on the number of servers in the system. Also, the law applies to any arbitrary boundary around part of the system, for example if only the queue is being considered then

\[
\bar{N}_q = \lambda \bar{W},
\]

where \( \bar{N}_q \) is the average number of customers in the queue and \( \bar{W} \) is the average time spent waiting in the queue. If only the server is being considered then:

\[
\bar{N}_s = \frac{\lambda}{\mu}.
\]
Performance Evaluation of Queuing Systems

Introduction to Queuing Systems
System Performance Measures & Little’s Law
Equilibrium Solution of Birth-Death Processes
Analysis of Single-Station Queuing Systems
We may model many different types of queuing systems by applying our continuous-time Birth-Death (Markov chain) analysis methods.

Calculation of the performance measures of interest depends on resolution of the stationary distribution of the states of the process, \( \{\pi_k\} \).

We have previously resolved the transient solution of the Birth-Death system for the simple case of the Poisson process. We now turn to the equilibrium solution and, given that the equations are more straight-forwards, we may find a general solution to \( \{\pi_k\} \) in terms of the birth and death coefficients.

Recall that the stationary distribution of an ergodic, homogeneous, continuous-time Markov chain is given by the solution to the system of linear equations:

\[
0 = \pi Q
\]  \hspace{1cm} (4)

coupled with the normalisation condition:

\[
\pi 1^T = \sum_{\forall i \in S} \pi_i = 1
\]
Birth-Death Systems in Equilibrium

From our definition of a Birth-Death process, recall that the matrix of transition rates $Q$ takes the form:

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

And so, equation (4) gives the set of linear equations:

$$0 = - (\lambda_k + \mu_k) \pi_k + \lambda_{k-1} \pi_{k-1} + \mu_{k+1} \pi_{k+1} \quad k \geq 1 \quad (5)$$
$$0 = -\lambda_0 \pi_0 + \mu_1 \pi_1 \quad k = 0 \quad (6)$$

So as to eliminate the equation for $k = 0$, we put

$$\lambda_{-1} = \lambda_{-2} = \lambda_{-3} = \ldots = 0$$
$$\mu_0 = \mu_{-1} = \mu_{-2} = \ldots = 0$$

Since a negative number of customers in the system is not possible, we have

$$\pi_{-1} = \pi_{-2} = \pi_{-3} = \ldots = 0.$$
Birth-Death Systems in Equilibrium

Thus, equations (5) and (6) can be combined to give the equilibrium equation:

\[ 0 = - (\lambda_k + \mu_k) \pi_k + \lambda_{k-1} \pi_{k-1} + \mu_{k+1} \pi_{k+1} \quad k = \ldots - 2, -1, 0, 1, 2, \ldots \] (7)

We find a solution for \( \pi_k \) as follows:

Solving for \( \pi_1 \)

\[ \pi_1 = \frac{\lambda_0}{\mu_1} \pi_0 \]

Solving for \( \pi_2 \)

\[ 0 = - (\lambda_1 + \mu_1) \pi_1 + \lambda_0 \pi_0 + \mu_2 \pi_2 \]
\[ 0 = - (\lambda_1 + \mu_1) \frac{\lambda_0}{\mu_1} \pi_0 + \lambda_0 \pi_0 + \mu_2 \pi_2 \]
\[ 0 = - \frac{\lambda_1 \lambda_0}{\mu_1} \pi_0 - \lambda_0 \pi_0 + \lambda_0 \pi_0 + \mu_2 \pi_2 \]
\[ \Rightarrow \pi_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} \pi_0 \]
Birth-Death Systems in Equilibrium

Continuing in this way:

\[ \pi_k = \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots \mu_k} \pi_0 \]

Thus the equilibrium probabilities may be expressed as:

\[ \pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_i+1} \quad k = 0, 1, 2, \ldots \tag{8} \]

And using the conservation (normalisation) equation, and noting that an empty product is equal to 1:

\[ \pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_i+1}} \tag{9} \]

Equations (8) and (9) give the general equilibrium (stationary) solution of the Birth-Death process.

We may model different types of queuing systems (as previously described by the Kendall notation) simply by setting particular sets of values for the coefficients \( \lambda_i, \mu_i \).
Performance Evaluation of Queuing Systems

Introduction to Queuing Systems
System Performance Measures & Little’s Law
Equilibrium Solution of Birth-Death Processes
Analysis of Single-Station Queuing Systems
In this case of an M/M/1 queuing system, the Birth-Death coefficients are:

\[ \lambda_k = \lambda \quad k = 0, 1, 2, \ldots \]
\[ \mu_k = \mu \quad k = 1, 2, 3, \ldots \]

Solving with these coefficients using the general solution equation (8):

\[ p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} \]
\[ p_k = p_0 \left( \frac{\lambda}{\mu} \right)^k \quad k \geq 0 \]

Note the change in notation to agree with Kleinrock: \( p_k \equiv \pi_k \)
Calculating $p_0$ using the normalisation equation (9):

$$p_0 = \left[1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k\right]^{-1}$$

$$p_0 = \left[1 + \frac{\lambda/\mu}{1 - \lambda/\mu}\right]^{-1}$$

$$p_0 = 1 - \lambda/\mu$$

Combining:

$$p_k = (1 - \rho)\rho^k \quad k = 0, 1, 2, \ldots \quad \text{where } \rho = \lambda/\mu$$

Note that for stability (ergodicity) in this system: $\rho < 1$.

To calculate $\bar{N}$ the average number of customers in the system:

$$\bar{N} = \sum_{k=0}^{\infty} k p_k$$

$$\bar{N} = \frac{\rho}{1 - \rho}$$
Figure: The steady state probability of $k$ customers in the system ($p_k$) for a given value of $\rho < 1$

We can now calculate the average time spent in the system using Little’s Law:

$$\bar{T} = \frac{\bar{N}}{\lambda} = \frac{\rho}{1 - \rho \lambda}$$

$$\bar{T} = \frac{1/\mu}{1 - \rho}$$
From the definition of the system, the arrivals to the M/M/1 system form a Poisson process, that is, the inter-arrival times are exponentially distributed.

We note (without proof) that the output process of the M/M/1 system is also a Poisson process, that is, the ‘inter-departure’ times are also exponentially distributed, with the same parameter $\lambda$ as the inter-arrival times.

We further note that the combination of Poisson traffic streams is also a Poisson traffic stream (see Tutorial 2).

These facts make it easy to analyse networks of connected M/M/1 systems, where the arrival process to one node is formed by the departure streams from one or more other nodes.

For other types of queuing systems, it is generally not the case that the departure process is the same as the arrival process and network analysis problems then involve resolving the joint distributions of the states of multiple queuing stations (generally an intractible problem to solve exactly for a large number of nodes).
Consider a queuing system where arrivals are discouraged when more customers are present in the system. This may be represented using state-dependent arrival (birth) rates in the Birth-Death system:

\[ \lambda_k = \frac{\alpha}{k + 1} \quad k = 0, 1, 2, \ldots \]
\[ \mu_k = \mu \quad k = 1, 2, 3, \ldots \]

Solving for \( p_k \) and \( p_0 \):

\[ p_k = p_0 \prod_{i=0}^{k-1} \frac{\alpha / (i + 1)}{\mu} = p_0 \left( \frac{\alpha}{\mu} \right)^k \frac{1}{k!} \]
\[ p_0 = \left[ 1 + \sum_{k=1}^{\infty} \left( \frac{\alpha}{\mu} \right)^k \frac{1}{k!} \right]^{-1} = e^{-\frac{\alpha}{\mu}} \]

Combining gives: \( p_k = \frac{(\alpha/\mu)^k}{k!} e^{-\frac{\alpha}{\mu}} \quad k = 0, 1, 2, \ldots \)
Consider a system with unlimited buffering capacity and \( m \) servers. The Birth-Death coefficients are:

\[
\lambda_k = \lambda \quad k = 0, 1, 2, \ldots \\
\mu_k = \begin{cases} 
  k\mu & 0 \leq k \leq m \\
  m\mu & m \leq k
\end{cases}
\]

Note that, for an ergodic system

\[
\frac{\lambda}{m\mu} < 1
\]

For the range \( k \leq m \):

\[
p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i + 1)\mu} = p_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!}
\]

For the range \( k \geq m \):

\[
p_k = p_0 \prod_{i=0}^{m-1} \frac{\lambda}{(i + 1)\mu} \prod_{j=m}^{k-1} \frac{\lambda}{m\mu} = p_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{m! \, m^{k-m}}
\]
Combining we have:

\[ p_k = \begin{cases} 
    p_0 \frac{(m\rho)^k}{k!} & k \leq m \\
    p_0 \frac{\rho^k m^m}{m!} & k \geq m
\end{cases} \]

Where \( \rho = \frac{\lambda}{m\mu} < 1 \)

\[ p_0 = \left[ \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \sum_{k=m}^{\infty} \frac{\rho^k m^m}{m!} \right]^{-1} = \left[ \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right) \right]^{-1} \]

The probability that an arrival joins the queue (is not served immediately) is:

\[ P[\text{queuing}] = \sum_{k=m}^{\infty} p_k = \sum_{k=m}^{\infty} p_0 \frac{\rho^k m^m}{m!} = \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right) \]

\[ P[\text{queuing}] = \frac{\left[ \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right) \right]}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right)} \]
The above equation is known as **Erlang’s C Formula**.

As it is can be cumbersome to calculate, charts (or software) are used to find values for the probability of delay (queuing).

We may also use the formula (by way of charts or software) to find the number of servers required given the maximum allowable probability of delay for a given the arrival intensity.

The Erlang-C formula is charted on the next slide. The x-axis is the **traffic intensity**, a measure of the rate of work offered to the system.

For a single server system:

\[
\text{Traffic Intensity} = \rho = \lambda \cdot \frac{1}{\mu}
\]

For an \( m \) server system:

\[
\text{Traffic Intensity} = m \cdot \rho = \lambda \cdot \frac{1}{\mu}
\]

The unit of traffic intensity is the **Erlang**, a dimensionless quantity.
M/M/m Queuing System

Erlang C - Probability of Delay (Queuing) Vs Traffic Intensity (Erlangs)
In this case, a maximum number of customers may be queued. Assume at most $K$ (including the customer in service). Any further arrivals (when there are $K$ in the system) are lost. The Birth-Death coefficients are:

$$\lambda_k = \begin{cases} \lambda & k < K \\ 0 & k \geq K \end{cases}$$

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

There are only $K + 1$ states in the system

$$p_k = p_0 \prod_{i=1}^{k-1} \frac{\lambda}{\mu} \quad k \leq K$$

$$p_k = p_0 \left( \frac{\lambda}{\mu} \right)^k \quad k \leq K$$

$$p_k = 0 \quad k > K$$
To solve for $p_0$:

$$p_0 = \left[ 1 + \sum_{k=1}^{K} \left( \frac{\lambda}{\mu} \right)^k \right]^{-1}$$

$$p_0 = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}$$

Combining $p_0$ with the previous equation, we have:

$$p_k = \begin{cases} 
\frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}} (\lambda/\mu)^k & 0 \leq k \leq K \\
0 & otherwise 
\end{cases}$$
M/M/m/m Loss System

In this case, there are \( m \) servers in this system and when a customer arrives it is assigned to a server or lost if all servers are busy.

The system parameters are:

\[
\lambda_k = \begin{cases} 
\lambda & k < m \\
0 & k \geq m 
\end{cases}
\]

\[
\mu_k = k\mu \quad k = 1, 2, \ldots, m
\]
M/M/m/m Loss System

\[ p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} \quad k \leq m \]

\[ p_k = \begin{cases} 
  p_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} & k \leq m \\
  0 & k > m 
\end{cases} \]

Solving for \( p_0 \)

\[ p_0 = \left[ \sum_{k=0}^{m} \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} \right]^{-1} \]
M/M/m/m Loss System

\[ p_k = \frac{(\lambda/\mu)^k}{k!} \sum_{i=0}^{m} \frac{(\lambda/\mu)^i}{i!} \]

Of particular importance for this system is \( p_m \), the fraction of the time that all \( m \) servers are busy:

\[ p_m = \frac{(\lambda/\mu)^m}{m!} \sum_{k=0}^{m} \frac{(\lambda/\mu)^k}{k!} \]

We note that this is the same probability as the probability that an arriving customer finds the system busy and is blocked (the call blocking probability).

This equation is known as **Erlang’s B Formula**. It can be graphed as the set of functions \( E_m(\lambda/\mu) \) (figure below) for use in system performance calculations.

As for the M/M/m/m chart, the x-axis is the traffic intensity = \( \lambda/\mu \).
Erlang B - Probability of Blocking (Loss) Vs Traffic Intensity (Erlangs)
In this case there is a finite number of system users ($M$). The user is either ‘arriving’ or already in the queue or being served. The average arrival rate for each customer is $\lambda$.

The state probabilities are:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda(M-i)}{\mu}$$

$$= p_0 \left( \frac{\lambda}{\mu} \right)^k \frac{M!}{(M-k)!}$$

for $0 \leq k \leq M$

with

$$p_0 = \left[ \sum_{k=0}^{M} \left( \frac{\lambda}{\mu} \right)^k \frac{M!}{(M-k)!} \right]^{-1}$$