Basic Queueing Theory

M/M/-/- Type Queues

Kendall’s Notation for Queues

\[ A/B/C/D/E \]

Shorthand notation where \( A, B, C, D, E \) describe the queue

Applicable to a large number of simple queueing scenarios
Kendall’s Notation for Queues

\[ A/B/C/D/E \]

A  Inter-arrival time distribution
B  Service time distribution
C  Number of servers
D  Maximum number of jobs that can be there in the system (waiting and in service)
   \textit{Default} \equiv \textit{for infinite number of waiting positions}
E  Queueing Discipline (FCFS, LCFS, SIRO etc.)
   \textit{Default is FCFS}

\[ \begin{align*}
M & \text{ exponential} \\
D & \text{ deterministic} \\
E_k & \text{ Erlangian (order } k) \\
G & \text{ general}
\end{align*} \]

M/M/1 or M/M/1/∞  Single server queue with Poisson arrivals, exponentially distributed service times and infinite number of waiting positions

\[ N = \lambda W \quad (2.9) \]
\[ N_q = \lambda W_q \quad (2.10) \]

\textbf{Little’s Result}

Result holds in general for virtually all types of queueing situations where
\[ \lambda = \text{Mean arrival rate of jobs that actually enter the system} \]
Jobs blocked and refused entry into the system will not be counted in \( \lambda \)
Consider the time interval \((0,t)\) where \(t\) is large, i.e. \(t \to \infty\).

\[
\text{Area}(t) = \text{area between } \alpha(t) \text{ and } \beta(t) \text{ at time } t = \int_{0}^{t} [\alpha(t) - \beta(t)] \, dt
\]

Average Time \(W\) spent in system = \(\lim_{t \to \infty} \frac{\text{Area}(t)}{\alpha(t)}\)

Average Number \(N\) in system = \(\lim_{t \to \infty} \frac{\text{Area}(t)}{t} = \lim_{t \to \infty} \frac{\alpha(t)}{t} \frac{\text{Area}(t)}{\alpha(t)}\)

Since, \(\lambda = \lim_{t \to \infty} \frac{\alpha(t)}{t}\)

Therefore, \(N = \lambda W\)
The PASTA Property

“Poisson Arrivals See Time Averages”

\[ p_k(t) = P\{\text{system is in state } k \text{ at time } t \} \]

\[ q_k(t) = P\{\text{an arrival at time } t \text{ finds the system in state } k \} \]

\( N(t) \) be the actual number in the system at time \( t \)

\( A(t, t+\Delta t) \) be the event of an arrival in the time interval \((t, t+\Delta t)\)

\[ q_k(t) = \lim_{\Delta t \to 0} P\{N(t) = k \mid A(t, t+\Delta t)\} \]

\[ = \lim_{\Delta t \to 0} \frac{P\{A(t, t+\Delta t) \mid N(t) = k\}P\{N(t) = k\}}{P\{A(t, t+\Delta t)\}} = p_k(t) \]

because \( P\{A(t, t+\Delta t) \mid N(t) = k\} = P\{A(t, t+\Delta t)\} \)

Equilibrium Solutions for M/M/-/- Queues

**Method 1:** Obtain the differential-difference equations as in Section 1.2 or Section 2.2. Solve these under equilibrium conditions along with the normalization condition.

**Method 2:** Directly write the flow balance equations for proper choice of closed boundaries as illustrated in Section 2.2 and solve these along with the normalization condition.

**Method 3:** Identify the parameters of the birth-death Markov chain for the queue and directly use equations (2.7) and (2.8) as given in Section 2.2.

*In the following, we have used this approach*
M/M/1 (or M/M/1/\infty) Queue

\[ \lambda_k = \lambda \quad \forall k \]
\[ \mu_k = \begin{cases} 0 & k = 0 \\ \mu & k = 1, 2, 3, \ldots \end{cases} \]

For \( \rho < 1 \)

\[ p_k = p_0 \left( \frac{\lambda}{\mu} \right)^k = p_0 \rho^k \]

\[ p_0 = (1 - \rho) \]

\[ N = \sum_{i=0}^{\infty} ip_i = \sum_{i=0}^{\infty} ip_i (1 - \rho) = \frac{\rho}{1 - \rho} \]

Using Little’s Result

\[ W = \frac{N}{\lambda} = \frac{1}{\mu (1 - \rho)} \]

\[ W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu (1 - \rho)} \]

\[ N_q = \lambda W_q = \frac{\rho^2}{1 - \rho} \]

Using Little’s Result

M/M/1/\infty Queue with Discouraged Arrivals

\[ \lambda_k = \frac{\lambda}{k + 1} \quad \forall k \]

For \( \lambda \mu \leq \infty \)

\[ p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu (i + 1)} = p_0 \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} \quad (2.14) \]

\[ p_0 = \exp\left( -\frac{\lambda}{\mu} \right) \quad (2.15) \]

\[ N = \sum_{k=0}^{\infty} kp_k = \frac{\lambda}{\mu} \]

\[ \lambda_{eff} = \sum_{k=0}^{\infty} \lambda_k p_k = \mu \left[ 1 - \exp\left( -\frac{\lambda}{\mu} \right) \right] \]

Little’s Result

\[ W = \frac{N}{\lambda_{eff}} = \frac{\lambda}{\mu^2 \left[ 1 - \exp\left( -\frac{\lambda}{\mu} \right) \right]} \]

Effective Arrival Rate
M/M/1/∞ Queue with Discouraged Arrivals

In this case, PASTA is not applicable as the overall arrival process is not Poisson

\[ \pi_r = P\{ \text{arriving customer sees } r \text{ in system} \} \]
\[ P(E_i) = p_i = e^{-\lambda/\mu} \left( \frac{\lambda}{\mu} \right)^1 \frac{1}{1!} \]
\[ \Delta E \text{ be the event of an arrival in } (t, t+\Delta t) \]
\[ E_i \text{ is the event of the system being in state } i \]
\[ P(\Delta E | E_i) = \frac{\lambda \Delta t}{i+1} \]
\[ \pi_r = P(E_i | \Delta E) = \frac{P(E_i)P(\Delta E | E_i)}{P(\Delta E)} = \frac{P(E_i)P(\Delta E | E_i)}{\sum_{i=0}^{\infty} P(E_i)P(\Delta E | E_i)} \]
\[ \pi_r = \left( \frac{\lambda}{\mu} \right)^{r+1} \frac{1}{(r+1)!} \left( \frac{e^{-\lambda/\mu}}{1-e^{-\lambda/\mu}} \right) \]
\[ W = \sum_{k=0}^{\infty} \frac{k+1}{\mu} \pi_k = \frac{\lambda}{\mu^2 (1-e^{-\lambda/\mu})} \]

Copyright 2002, Sanjay K. Bose

M/M/m/∞ Queue (m servers, infinite number of waiting positions)

\[ \lambda_k = \lambda \quad \forall k \]
\[ \mu_k = k\mu \quad 0 \leq k \leq (m-1) \]
\[ = m\mu \quad k \geq m \]

For \( \rho = \lambda / \mu < m \)
\[ p_k = p_0 \frac{\rho^k}{k!} \quad \text{for } k \leq m \quad (2.16) \]
\[ p_k = p_0 \frac{\rho^k}{m!m^{k-m}} \quad \text{for } k > m \]

Erlang’s C-Formula

\[ p_0 = \left( \sum_{k=0}^{m-1} \frac{\rho^k}{k!} + \frac{\rho^m}{m!(m-\rho)} \right)^{-1} \quad (2.17) \]

\[ P\{\text{queueing}\} = \sum_{k=m}^{\infty} p_k = C(m, \rho) = p_0 \frac{\rho^m}{m!(m-\rho)} \quad (2.18) \]

Copyright 2002, Sanjay K. Bose
M/M/m/m Queue (m server loss system, no waiting)

\[ \lambda_k = \begin{cases} \lambda & k < m \\ 0 & \text{otherwise} \end{cases} \]  
\[ \mu_k = \begin{cases} k\mu & 0 \leq k \leq m \\ 0 & \text{otherwise} \end{cases} \]

\[ p_k = \begin{cases} p_0 \frac{\rho^k}{k!} & \text{for } k \leq m \\ 0 & \text{otherwise} \end{cases} \]  
\[ \rho = \frac{\lambda}{\mu} < \infty \]

\[ p_0 = \frac{1}{\sum_{k=0}^{m} \frac{\rho^k}{k!}} \]  
\[ \sum_{k=0}^{m} \frac{p_0}{k!} = 1 \]  
\[ B(m, \rho) = p_0 \frac{\rho^m}{m!} \]  
\[ B(0, \rho) = 1 \]

Simple model for a telephone exchange where a line is given only if one is available; otherwise the call is lost

Blocking Probability \( B(m, \rho) = \text{P\{an arrival finds all servers busy and leaves without service\}} \)

\[ B(m, \rho) = p_0 \frac{\rho^m}{m!} \]  
Erlang’s B-Formula

\[ B(0, \rho) = 1 \]

\[ B(m, \rho) = \frac{m}{1 + \frac{\rho B(m-1, \rho)}{m}} \]  
\[ B(0, \rho) = 1 \]
M/M/1/K Queue (single server queue with K-1 waiting positions)

\[ \lambda_k = \begin{cases} \lambda, & k < K \\ 0, & \text{otherwise} \quad \text{(Blocking or Loss Condition)} \end{cases} \]

\[ \mu_k = \begin{cases} \mu, & k \leq K \\ 0, & \text{otherwise} \end{cases} \]

For

\[
p_k = p_0 \rho^k \quad \text{for} \quad k \leq K \tag{2.23}
\]

\[
p_k = 0 \quad \text{otherwise}
\]

\[
p_0 = \frac{(1-\rho)}{(1-\rho^{K+1})} \tag{2.24}
\]

M/M/1/\infty K Queue (single server, infinite number of waiting positions, finite customer population K)

\[ \lambda_k = \lambda(K - k) \quad k < K \]
\[ = 0 \quad \text{otherwise} \quad \text{(Blocking or Loss Condition)} \]

\[ \mu_k = \mu \quad k \leq K \]
\[ = 0 \quad \text{otherwise} \]

For

\[
p_k = p_0 \rho^k \frac{K!}{(K-k)!} \quad k=1,\ldots,K \tag{2.25}
\]

\[
p_0 = \frac{1}{\sum_{k=0}^{K} \rho^k \frac{K!}{(K-k)!}} \tag{2.26}
\]
Delay Analysis for a FCFS M/M/1/∞ Queue
(Section 2.6.1)

\( Q \): Queueing Delay (not counting service time for an arrival
\( f_Q(t), \ cdf \ F_Q(t), \ L_Q(s) = LT(f_Q(t)) \)

\( W = Q + T \)
\( W \): Total Delay (waiting time and service time) for an arrival
\( f_W(t), \ cdf \ F_W(t), \ L_W(s) = LT(f_W(t)) \)

\( T \): Service Time
\( f_T(t) = \mu e^{-\mu t} \quad F_T(t) = e^{-\mu t} \quad L_T(s) = \frac{\mu}{(s + \mu)} \)

Since
\( Q \perp T \)
\( L_W(s) = \frac{\mu}{(s + \mu)} L_Q(s) \quad f_W(t) = f_Q(t) * \left( \mu e^{-\mu t} \right) \quad (2.30) \)

Knowing the distribution of either \( W \) or \( Q \), the distribution of the other may be found

For a particular arrival of interest -
\( F_Q(t) = P\{\text{queueing delay} \leq t\} \)
\( = P\{\text{queueing time} = 0\} + \left[ \sum_{n=1}^{\infty} P\{\text{queueing time} \leq t \mid n \text{ jobs in system}\} \right] p_n \)

\( F_Q(t) = (1 - p) + (1 - p) \sum_{n=1}^{\infty} p^n \int_{0}^{t} \frac{\mu^n}{(n-1)!} e^{-\mu x} \ dx \)
\( = (1 - p) + (1 - p) \int_{0}^{t} \mu e^{-\mu x} \sum_{n=1}^{\infty} \frac{\mu^n x^{n-1}}{(n-1)!} \ dx \quad \text{Erlang-n distribution for sum of} \ n \text{ exponential r.v.s} \)
\( = (1 - p) + (1 - p) \mu p \int_{0}^{t} e^{-\mu x} \sum_{n=1}^{\infty} \frac{(n-1)!}{(n-1)!} (1 - p)^n \ dx \)
\( = (1 - p) + (1 - p) \mu p \int_{0}^{t} e^{-\mu x} (1 - p)^{n-1} \ dx = (1 - p) + \rho (1 - e^{-\mu(1-p)}) \quad (2.31) \)

\( f_Q(t) = \frac{dF_Q(t)}{dt} = \delta(t) (1 - p) + \lambda (1 - p) e^{-\mu(1-p)} \quad (2.32) \)

\( f_W(t) = (1 - p) \mu e^{-\mu t} + \lambda (1 - p) \mu \int_{0}^{t} e^{-\mu(1-p)(1-x)} e^{-\lambda x} \ dx = (\mu - \lambda) e^{-(\mu + \lambda) t} \)
Arrival/Departure of Customer/Job of Interest from a FCFS M/M/1 Queue

Let

\[ N^* = \text{Number in the system that a job will see left behind when it departs} \]

\[ p_n^* = P[N^* = n] \text{ for } N^* = 0, 1, ..., \infty \]

For a FCFS queue, number left behind by a job will be equal to the number arriving while it is in the system.

\[ G^*(z) = \sum_{n=0}^{\infty} z^n p_n^* = \sum_{n=0}^{\infty} z^n \int_{t=0}^{\infty} \frac{\lambda t^n}{n!} e^{-\lambda t} f_W(t) dt = \int_{t=0}^{\infty} e^{-\lambda t(1-z)} f_W(t) dt = L_W(\lambda - \lambda z) \]  

(2.36)

\[ E[N^*] = \left. \frac{dG^*(z)}{dz} \right|_{z=1} = -\lambda \left. \frac{dL_W(s)}{ds} \right|_{s=0} = \lambda W \]  

(2.37)
An important general observation can also be made along the lines of Eq. (2.36).

Consider the number arriving from a Poisson process with rate $\lambda$ in a random time interval $T$ where $L_T(s) = LT[f_T(t)]$. The generating function $G(z)$ of this will be given by

$$G(z) = L_T(\lambda - \lambda z)$$

and the mean number will be $E[N] = \lambda E[T]$.

This result will be found to be useful in various places in our subsequent analysis.

---

**Delay Analysis for the FCFS M/M/m/∞ Queue**

(Section 2.6.2)

Using an approach similar to that used for the M/M/1 queue, we obtain the following

\[
f_Q(t) = \left\{1 - p_0 \left[ \frac{m \rho^m}{m! (m - \rho)} \right] \delta(t) + \left[ \frac{\mu p_0 \rho^m e^{-\mu y (m - p)}}{(m - 1)!} \right] \mu(t) \right\}
\]

\[
f_W(t) = \left\{1 - p_0 \left[ \frac{m \rho^m}{m! (m - \rho)} \right] \mu e^{-\mu t} - \left[ \frac{\mu p_0 \rho^m (e^{-\mu y (m - p)} - e^{-\mu y})}{(m - 1)! (1 - m - \rho)} \right] \right\}
\]

See Section 2.6.2 for the details and the intermediate steps.