Lecture 7

M/M/c-delay

7.1 Probability of delay

Again the arrival process is Poissonian with arrival rate ρ. There are c servers. The distribution of holding-times is exponential (1). There is an infinite queue.

When there are r demands that have not yet been completely served, the system is said to be "in state $[r]$". When $r \leq c$, there are r engaged servers, whilst the queue is empty. If $r = c + m$ $(m > 0)$, all servers are engaged, whilst there are m queued demands. It is assumed that there is stationarity. The probability of the system being in $[r]$ is supposed to be time-independent and is denoted by p_r.

When the system is in $[r]$, there is a transition rate ρ to $[r+1]$. The transition rate to $[r-1]$ equals the number of engaged servers, i.e. r when $r \leq c$ and c otherwise. Hence, the states and the transition rates may be given as follows:

$$[0] \underset{1}{\overset{\rho}{\longleftrightarrow}} [1] \underset{2}{\overset{\rho}{\longleftrightarrow}} [2] \ldots \underset{c-1}{\overset{\rho}{\longleftrightarrow}} [c-1] \underset{c}{\overset{\rho}{\longleftrightarrow}} [c] \underset{c}{\overset{\rho}{\longleftrightarrow}} [c+1] \underset{c}{\overset{\rho}{\longleftrightarrow}} [c+2] \ldots$$

probability of states are respectively: $p_0, p_1, p_2, \ldots, p_{c-1}, p_c, p_{c+1}, p_{c+2}$

Just as in Sections 2.1 and 2.2 the probability of a transition $[r] \to [r+1]$ should equal the probability of a reverse transition. Hence,

$$-\rho p_r + (r+1)\, p_{r+1} = 0, \tag{7.1}$$

$r = 0, 1, \ldots, c-1$ and

$$-\rho p_r + c p_{r+1} = 0 \tag{7.2}$$

$r = c, c+1, \ldots$.

Those equations should be completed by

$$\sum_{i=0}^{\infty} p_r = 1. \tag{7.3}$$

The equations. (7.1) and (7.2) are suitable for expressing p_{r+1} in terms of p_r, this again in terms of p_{r-1},..., and ultimately in terms of p_0. The result is

$$p_r = \begin{cases} \frac{\rho^r}{r!} p_0 & for \quad r = 1, 2, \ldots, c \\ \left(\frac{\rho}{c}\right)^{r-c} p_c & for \quad r > c. \end{cases} \tag{7.4}$$

When $\rho \geq c$ and $p_0 > 0$, the array $p_0, p_1, \ldots$. would be never-decreasing. This is incompatible with (7.3). When $p_0 = 0$, all p_r, would be zero, which does not constitute a good solution. Hence, for $\rho \geq c$ no solution is found. This stems from the fact that for $\rho \geq c$ stationarity is impossible. For the present we assume $q < c$. When the values (7.4) for the p_r are inserted into (7.3), the value for p_0 follows from

$$p_0 \left[\sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \left(1 + \frac{\rho}{c} + \ldots\right) \right] = 1$$

yielding

$$\begin{aligned} p_0 &= 1 / \left[\sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \frac{c}{c-\rho} \right] \\ &= 1 / \left[\sum_{i=0}^{c} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \frac{\rho}{c-\rho} \right]. \end{aligned}$$

Note that for $c = 1$ we get

$$p_0 = 1 - \rho.$$

7.2 Average waiting-time for the system

The probability D that a demand is delayed equals the probability that all c servers are busy, i.e.

$$D = \sum_{r=c}^{\infty} p_r = p_0 \frac{\rho^c}{c!} \sum_{r=c}^{\infty} \left(\frac{\rho}{c}\right)^{r-c} = p_c \frac{c}{c-\rho}$$

or

$$D = \frac{\frac{\rho^c}{c!} \frac{c}{c-\rho}}{\sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\rho^c}{c!} \frac{c}{c-\rho}} \, (= E_{2,c}) \, .$$

This is the Erlang C-formula. The function $E_{2,c}$ is called the second *Erlang function*. Note that for $c = 1$ we have:

$$D = \rho.$$

Let us calculate L the average number of customers in the system. Obviously it is equal to

$$\begin{aligned}
L &= \mathbb{E}\left(\sum_{r=1}^{\infty} r p_r\right) = \sum_{r=1}^{c} r\frac{\rho^r}{r!}p_0 + \sum_{r=c+1}^{\infty} r\left(\frac{\rho}{c}\right)^{r-c} p_c = \\
&= p_0\rho\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + p_c c\sum_{i=1}^{\infty}\left(\frac{\rho}{c}\right)^i + p_c\sum_{i=1}^{\infty} i\left(\frac{\rho}{c}\right)^i = \\
&= p_0\rho\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + p_c c\left(\frac{1}{1-\rho/c} - 1\right) + p_c\frac{\rho/c}{(1-\rho/c)^2} = \\
&= p_0\rho\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + p_c\left(\frac{c\rho}{c-\rho} + \frac{\rho c}{(c-\rho)^2}\right) \\
&= p_0\rho\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + p_0\frac{\rho^c}{c!}c\rho\frac{c-\rho+1}{(c-\rho)^2} = \\
&= p_0\rho\left(\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + \frac{\rho^c}{c!}\frac{c}{c-\rho}\right) + p_0\rho\left(\frac{\rho^c}{c!}c\frac{c-\rho+1}{(c-\rho)^2} - \frac{\rho^c}{c!}\frac{c}{c-\rho}\right) \\
&= \rho + c\rho\rho^c\frac{p_0}{c!\,(c-\rho)^2} = \rho\left(1 + \frac{\rho^c p_0}{(c-1)!\,(c-\rho)^2}\right).
\end{aligned}$$

Again for $c = 1$ we have very simple formula:

$$L = \rho + \frac{\rho^2}{(1-\rho)^2}p_0 = \rho + \frac{\rho^2}{1-\rho} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

Hence average time spent in the system: $W = L/\lambda$ and for $c = 1$ we get $W = 1/(\mu-\lambda)$ and consequently W_Q average time spent in the queue $W_Q = W - 1/\mu = 1/(\mu-\lambda) - 1/\mu = \lambda/(\mu(\mu-\lambda)) = \rho/(1-\rho)$. Note that W_Q is equal to w the average waiting time! obtained below.

The quantity p_r, is the expected time the system is in $[r]$ per unit of time. The average waiting-time w per demand equals the expectation of total waiting-time (i.e. incurred by all demands) divided by the average number of demands, both per unit of time:

$$w = \sum_{r=c}^{\infty}(r-c)\,p_r/\rho = p_c\sum_{s=0}^{\infty} s\left(\frac{\rho}{c}\right)^s/\rho = p_c\frac{c}{(c-\rho)^2}$$

The average waiting-time EW per delayed demand equals

$$EW = \frac{w}{D} = \frac{1}{c-\rho}.$$

It is obvious that these averages are independent of queue-discipline, in so far as the latter does not depend on advance knowledge of holding-times of waiting demands. The distribution of waiting-times, however, is not independent of this discipline.

7.3 The Distribution of waiting-times in the system M/M/c- delay

The average waiting-time W of delayed demands does not depend on the order in which waiting demands are given service, as long as this order is not influenced by some advance knowledge of the holding-times of waiting demands. The distribution of the waiting-time W', however, depends on this order. Let $q(\tau)$ and $Q(\tau)$ be the p.d.f. and c.d.f. of W for delayed demands. In the next three sections those quantities will be evaluated for the queue-disciplines "first-come-first-served", "last-come-first-served" and "random".

7.3.1 First-come-first-served

When a demand arrives that finds all servers occupied, the number s of demands ahead of the new demand in the queue may have any value. First of all we shall evaluate $Q_s(\tau)$, the c.d.f. of the waiting-time W of a demand that finds s queued demands ahead of it. Now, in order that W for the arriving demand be less than τ, all s waiting demands plus the newly arriving one must have left the queue during τ. So this is equivalent to the ending of at least $s+1$ occupations during τ. As the servers are fully occupied during this process, the ends of occupations form a Poisson process with a rate c (bearing in mind that the holding-time is exponential (1)). So, according to , changing ρ into c, one obtains:

$$Q_s(\tau) = \sum_{i \geq s+1} \frac{\exp(-c\tau)\,(ct)^i}{i!},$$

The probability that a demand suffering delay finds s queued demands ahead of it can be evaluated with the aid of (7.4). It is

$$p_{s+c} / \sum_{m=0}^{\infty} p_{c+m} = (1-\eta)\,\eta^s$$

with $\eta = \rho/c$.

Hence the unconditional c.d.f. $Q(\tau)$ of the waiting-time W is:

$$\begin{aligned}
Q(\tau) &= \sum_{s=0}^{\infty} (1-\eta)\,\eta^s Q_s(\tau) \\
&= \sum_{s=0}^{\infty} (1-\eta)\,\eta^s \sum_{i \geq s+1} \frac{\exp(-c\tau)\,(ct)^i}{i!} \\
&= \exp(-c\tau) \sum_{i=1}^{\infty} \frac{(c\tau)^i}{i!} \sum_{s=0}^{i-1} (1-\eta)\,\eta^s \\
&= \exp(-c\tau) \sum_{i=1}^{\infty} \frac{(c\tau)^i}{i!} (1-\eta^i) \\
&= 1 - \exp(-(c-\rho)\,\tau)
\end{aligned}$$

Under the condition "first-come-first-served" the waiting-time W is exponential with average

$$EW = 1/(c-\rho).$$

7.3.2 Last-come-first-served

A newly arriving demand that cannot be served immediately is observed; let it be denoted by A. At some instant at which A is still waiting, let s be the number of waiting demands that have arrived after A. Let the situation be described by saying that "the system is in $[s]$". When A arrives the system is in $[0]$. Let $q_s(r)$ be the p.d.f. for the further waiting-time for A, when the system is in $[s]$. Obviously, equals the p.d.f. $q_0(\tau)$ for the (total) waiting-time for A.

Consider the event , " $f = \tau + \Delta\tau\,(+d\tau)$ whilst the system is in $[s]$". During the first part $\Delta\tau$ of the further waiting-time three events are possible when $s > 0$ (neglecting higher-order effects):

(i) one arrival (probability $\rho\Delta\tau$); new state $[s+1]$;

(ii) one termination (probability $c\Delta\tau$); new state $[s-1]$;

(iii) no change (probability $1-\rho\Delta\tau-c\Delta\tau$); new state $[s]$.

In order for f to be $\tau + \Delta\tau\,(+d\tau)$ in total, those events should be followed by the event that the further waiting-time is $\tau(+\mathrm{dr})$, taken conditional on the new state. The corresponding conditional probabilities are $q_{s+1}(r)d\tau$, $q_{s-1}(r)d\tau$ and $q_s(\tau)d\tau$, respectively. Hence:

$$q_s(\tau+\Delta\tau)d\tau = \rho\Delta\tau q_{s+1}(\tau)d\tau + c\Delta\tau q_{s-1}(\tau)d\tau + (1-\rho\Delta\tau-c\Delta\tau)\,q_s(\tau)d\tau$$

or, passing to the limit at $\Delta\tau \to 0$,

$$\frac{dq_s(\tau)}{d\tau} = \rho q_{s+1} - (\rho+c)\,q_s + cq_{s-1} \tag{7.5}$$

$s = 1, 2, \ldots$.

For $s = 0$ the event (ii) (termination) would lead to a situation with $f = 0$. Hence, for $s = 0$ and $\tau > 0$ the last term in (7.5) is absent (i.e. $q_{-1} = 0$).

The initial conditions are as follows. When the system is in $[s]$, with $s > 0$, it is impossible for the further waiting-time to end within the next interval $d\tau$. When the system is in $[0]$ this probability is $cd\tau$,

Hence

$$q_s\,(0) = c\delta_s^0.$$

Solving this differential equation leads to the following formula

$$q\,(\tau) = \sqrt{\frac{c}{\rho}}\frac{\exp\,(-\,(\rho+c)\,\tau)}{\tau} I_1\,(2\sqrt{\rho c}\tau)\,,$$

where I_1 is a modified Bessel function of order 1.

7.3.3 "Random" condition

Again, a newly arriving demand A, that cannot be served immediately, is observed. The situation in the queue at some instant after arrival of A will be described by the following complete set of "states of the system":

$$[s]\begin{cases} s=0,1,2,\ldots & A \quad \textit{is waiting together with others} \\ s=* & A \qquad \textit{is no longer waiting} \end{cases}$$

Let $Q_s(\tau)$ be the c.d.f. of the further waiting-time f for A, under the condition that the system is in state $[s]$. Obviously $Q_*(\tau)=1$ for $\tau>0$.

Consider the event "$f\le\tau+\Delta\tau$, whilst the system is in $[s]$, $s\ge 0$. During the first part at of $\Delta\tau$ the further waiting-time three events are possible (apart from higher-order effects):

(i) one arrival (probability $\rho\Delta\tau$); new state $[s+1]$;

(ii) a server becomes free (probability $c\Delta\tau$) and is seized by one of the other waiters (conditional probability $s/(s+1)$); new state $[s-1]$;

(iii) a server becomes free (probability $c\Delta\tau$) and is seized by A (conditional probability $1/(s+1)$); new state $[*]$;

(iv) no change (probability $1-\rho\Delta\tau-c\Delta\tau$); new state $[s]$.

In order for $f$to be $\le\tau+\Delta\tau$ in total, these events should be followed by the event that the further waiting-time is $\le\tau$, taken conditional on the new state. The corresponding conditional probabilities are $Q_{s+1}(\tau)$, $Q_{s-1}(\tau)$, and $Q_s(\tau)$, respectively. Hence:

$$Q_s(\tau+\Delta\tau)\cong\rho\Delta\tau Q_{s+1}(\tau)+c\Delta\tau\frac{s}{s+1}Q_{s-1}(\tau)+c\Delta\tau\frac{1}{s+1}+(1-\rho\Delta\tau-c\Delta\tau)\,Q_s(\tau),$$

or equivalently after passing to the limit with $\Delta\tau$ to zero

$$\frac{dQ_s}{d\tau}=\rho Q_{s+1}-(\rho+c)\,Q_s+\frac{cs}{s+1}Q_{s-1}+\frac{c}{s+1} \tag{7.6}$$

$$Q(\tau)=(1-\eta)\sum_{s=0}^{\infty}\eta^s Q_s(\tau).$$

Equation (7.6) was derived by Vaulot (1946) and Palm (1946, 1957). An analytical solution stems from Pollaczek (1946). Approximation* are given in Pollaczek (1946) and Riordan (1953).

7.3.4 Comparison of the cases "first-come-first-served", "last-come-first-served" and "random"

In general it cannot be decided whether "first-come-first-served" or "last-come-first-served" queue-discipline is better. Mostly, however, It is important to guarantee that some rather large critical value of wailing time is not exceeded too frequently. In this case the right-hand portion* of the curves should be compared, and a " first-come-first-servedl" discipline is best.