

## Lecture 9 — The $M/G/1$ System

In this lecture we move away from studying purely Markov systems and study the  $M/G/1$  queue and the special case of the  $M/D/1$  queue. (Note that we could see the  $M/M/1$  queue as a special case of the  $M/G/1$  queue). The result derived is known as the Pollaczek-Khinchin (P-K) formula. The formula we are working to prove is given by first defining:

$$\bar{X} = E[X] = \frac{1}{\mu} = \text{Average service time} \quad (1)$$

and

$$\overline{X^2} = E[X^2] = \text{Second moment of service time} \quad (2)$$

The P-K formula is then:

$$W = \frac{\lambda \overline{X^2}}{2(1 - \rho)} \quad (3)$$

where  $W$  is the expected customer waiting time in a queue and  $\rho = \lambda/\mu = \lambda\bar{X}$  the utilisation as usual.

This lecture we will derive and use the P-K formula and a simple variant.

First let us introduce some notation:

$W_i$  waiting time (in queue) for  $i$ th customer.

$X_i$  service time of the  $i$ th customer – we assume that these are independent and identically distributed (i.i.d) variables.

$N_i$  number of customers that is found in the queue (not yet being served) when the  $i$ th customer arrives.

$R_i$  residual service time found by the  $i$ th customer (defined below).

**Definition 1.** The residual time  $R_i$  is the service time remaining to the customer being served when the  $i$ th customer arrives at the queue. If no customer is being served ( $N_i = 0$ ) then  $R_i = 0$ .

A graph will help understand the concept of residual time. Figure 1 shows the residual time in a queue  $r(\tau)$  is the residual time remaining at time  $\tau$ .  $X_i$  is the service time of the  $i$ th customer (note that the slopes of all the diagonal lines on this graph are, obviously, one). If we take a time  $t$  where the system is empty (as shown in the diagram) then define  $M(t)$  as the number of customers who have been served and exited the system by time  $t$ .

The mean residual time in the interval  $[0, t]$  is clearly the average value on the  $y$  axis in the interval. This is the area under the curve divided by  $t$  which is given by:

$$\frac{1}{t} \int_0^t r(\tau) d\tau = \frac{1}{t} \sum_{i=1}^{M(t)} \frac{1}{2} X_i^2 \quad (4)$$

Which we can rewrite as:

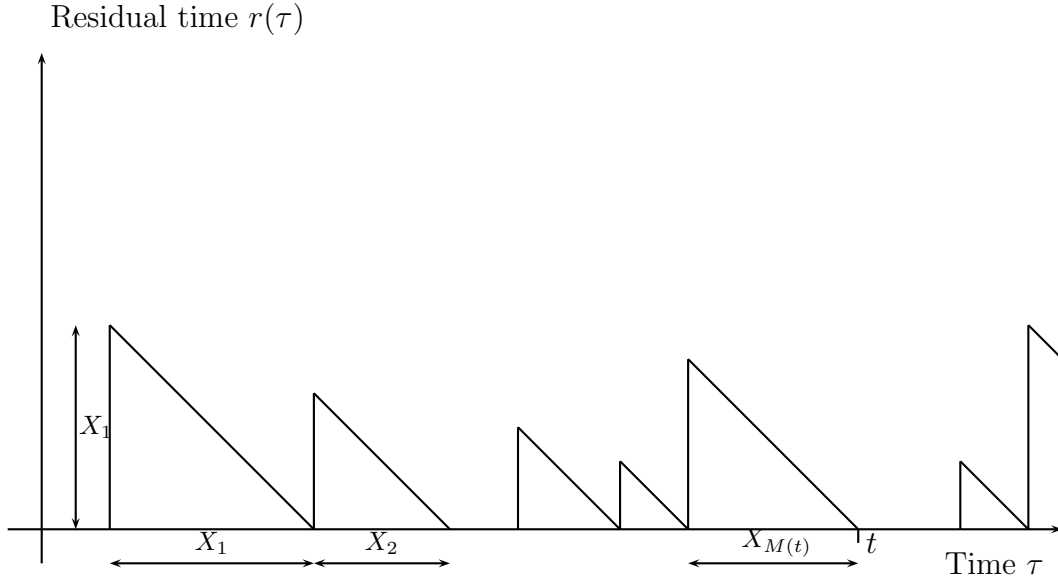


Figure 1: Service Time of Arrivals at an  $M/G/1$  queue.

$$\frac{1}{t} \int_0^t r(\tau) d\tau = \frac{1}{2} \frac{M(t)}{t} \frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} \quad (5)$$

Now, assuming the relevant limits exist we have:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t r(\tau) d\tau = \frac{1}{2} \lim_{t \rightarrow \infty} \frac{M(t)}{t} \cdot \lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} \quad (6)$$

Now, if we assume that the system is ergodic then we can replace these time averages with ensemble averages. In this case we define:

$$R = \text{Mean residual time defined as } R = \lim_{i \rightarrow \infty} E[R_i] \quad (7)$$

And, if the time average is the state space average then:

$$R = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t r(\tau) d\tau \quad (8)$$

Since the system is lossless (no customers ever vanish) then if the number of customers does not rise forever — the number queuing tends to a limit — we can say that the departure rate must equal the arrival rate. That is:

$$\lim_{t \rightarrow \infty} \frac{M(t)}{t} = \lambda \quad (9)$$

Therefore (6) becomes:

$$R = \frac{1}{2} \lambda \overline{X^2} \quad (10)$$

Now, we know that the waiting time for the  $i$ th customer is equal to the residual service time of the customer currently being served plus the total service times of those who are in the queue. This is given by:

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j \quad (11)$$

We note that the  $X_j$ s are i.i.d by hypothesis.  $N_i$  cannot possibly be affected by the  $X_j$  values in this sum since those are the service times of customers who are still waiting in this queue. Therefore,  $N_i$  is also independent from the  $X_j$  in the above. Therefore we may take expectations as follows:

$$E[W_i] = E[R_i] + E\left\{ \sum_{j=i-N_i}^{i-1} E[X_j | N_i] \right\} = E[R_i] + \overline{X} E[N_i] \quad (12)$$

Finally, taking the limit as  $i \rightarrow \infty$  and remembering that  $\overline{X} = \frac{1}{\mu}$  we have:

$$W = R + \frac{1}{\mu} N_Q \quad (13)$$

where  $N_Q$  is the limit as  $i \rightarrow \infty$  of the expected number found in the queue. By Little's theorem we get:

$$N_Q = \lambda W \quad (14)$$

and therefore:

$$W = R + \frac{\lambda}{\mu} W \quad (15)$$

rearranging and substituting  $\rho = \lambda/\mu$  and our expression for  $R$  from (10) we get:

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} \quad (16)$$

which is the P-K formula we required.

Let us remember the assumptions for this remarkably general formula.

1. The sending process was a Poisson process with parameter  $\lambda$ .
2. The steady state time averages  $R$ ,  $W$  and  $N_Q$  exist.
3. The long-term time averages correspond to the state-space averages.

4. The service times  $X_i$  are i.i.d. variables.

in our derivation we also assumed that the system was FIFO although this is not, in fact, necessary — it is only necessary that the order of service is independent of the required service time.

Note that the  $M/D/1$  queue is the special case of this when all service times are identical. In this case  $X_i = \frac{1}{\mu}$  and therefore  $\overline{X^2} = \frac{1}{\mu^2}$  and:

$$W = \frac{\rho}{2\mu(1-\rho)} \quad (17)$$

this is the lowest possible value of  $\overline{X^2}$  and therefore a lower bound for any  $M/G/1$  system. Compare it to the  $M/M/1$  system where  $\overline{X^2} = 2/\mu^2$  and therefore:

$$W = \frac{\rho}{\mu(1-\rho)} \quad (18)$$

In other words the  $M/M/1$  formula has twice the waiting time of the lower bound  $M/D/1$  waiting time. We should also note that there is no upper bound on  $\overline{X^2}$  therefore it is possible that queues which have a utilisation less than one have an infinite waiting time.

## Further $M/G/1$ information

### Question

What is the probability that the system is empty when a customer arrives?

### Answer

The expected time to serve  $n$  customers is  $\sum_{i=1}^n X_i$ . The expected time for  $n$  customers to depart is  $n/\lambda$  (since the customers are generated by a Poisson process with rate  $\lambda$  and are also departing at a similar rate as previously stated).

$$Pr\{\text{Empty}\} = \lim_{n \rightarrow \infty} \frac{\text{Time taken for } n \text{ customers to depart} - \text{Time serving } n \text{ customers}}{\text{Time taken for } n \text{ customers to depart}} \quad (19)$$

which is:

$$Pr\{\text{Empty}\} = \lim_{n \rightarrow \infty} \frac{n/\lambda - \sum_{i=1}^n X_i}{n/\lambda} \quad (20)$$

and therefore:

$$Pr\{\text{Empty}\} = 1 - \lambda \overline{X} \quad (21)$$

### Question

What is the average length between busy periods?

**Answer**

A period between busy periods begins when the last customer exits. It will end when the next customer is generated. Since the generating process is a Poisson and therefore memoryless, the expected time for the next arrival is after a time  $1/\lambda$ .

**Question**

What is the average length of a busy period?

**Answer**

If  $L$  is the average length of a busy period then:

$$Pr\{\text{Empty}\} = \frac{1/\lambda}{L + 1/\lambda} \quad (22)$$

Substituting from earlier and multiplying top and bottom of RHS by  $\lambda$ :

$$1 - \lambda\bar{X} = \frac{1}{\lambda L + 1} \quad (23)$$

rearranging:

$$\lambda L + 1 = \frac{1}{1 - \lambda\bar{X}} \quad (24)$$

final rearrangement gives us:

$$L = \frac{\bar{X}}{1 - \lambda\bar{X}} \quad (25)$$