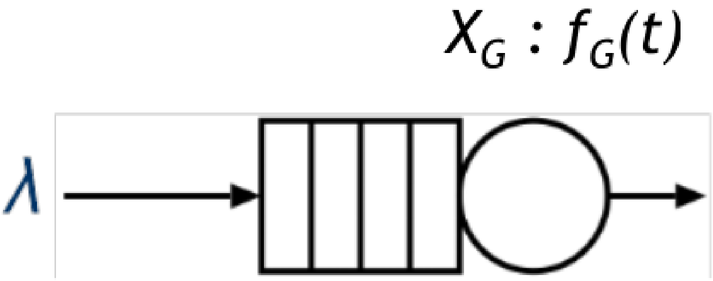
17.GGc queues

We can do very few things when there are non exponential distributions. One thing we can do is G/G/c queues. Usually in reality we have Poisson distribution.

M/G/1

- M: arrival to the system follow a poisson process and the service have a general distribution as long as the service time are independent



Service time(first moment of the system):

$$D = E[X_G] = \int_0^\infty t f_G(t)dt = \lim_{N\to\infty}[\frac{1}{N}\sum_{i=1}^{N} X_G]$$

Second moment:

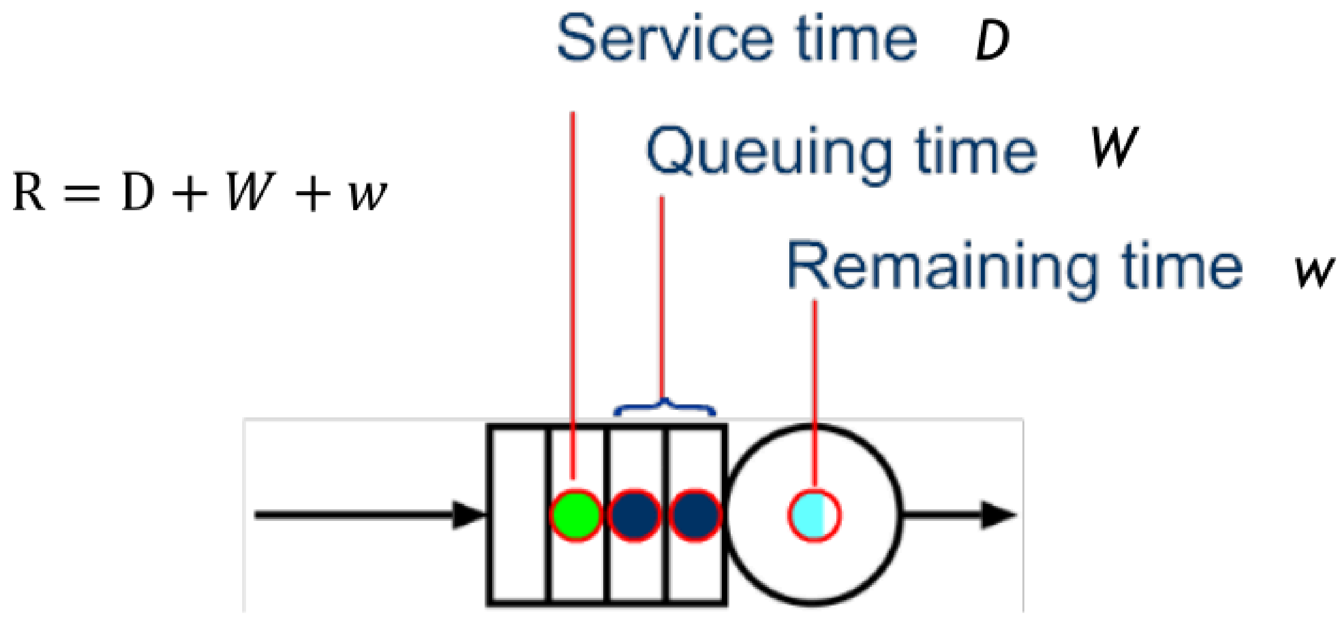
$$m_2 = E[X_G^2] = \int_0^\infty t^2 f_G(t)dt = \lim_{N\to\infty}[\frac{1}{N}\sum_{i=1}^{N} X_G^2]$$

Traffic intensity ρ(stable if $\rho < 1$):

$$\rho = \lambda D = \lambda E[X_G]$$

FCFS

Let us focus on First-Come-First-Served service center. The average response time of a station is the sum of three terms: Service Time, Queueing Time and Remaining Time of the job in service at the arrival. Queuing time and remaining service time of the job in service at the arrival determine the time a job will have to wait before being served.

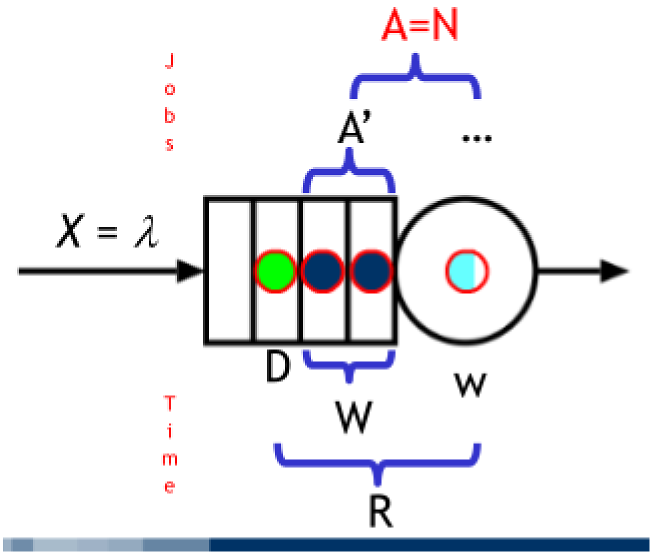


The waiting time in the queue W can be computed from A', the number of jobs found waiting (not in service) by one job at its arrival:

$$W = A'E[X_G] = A'D$$

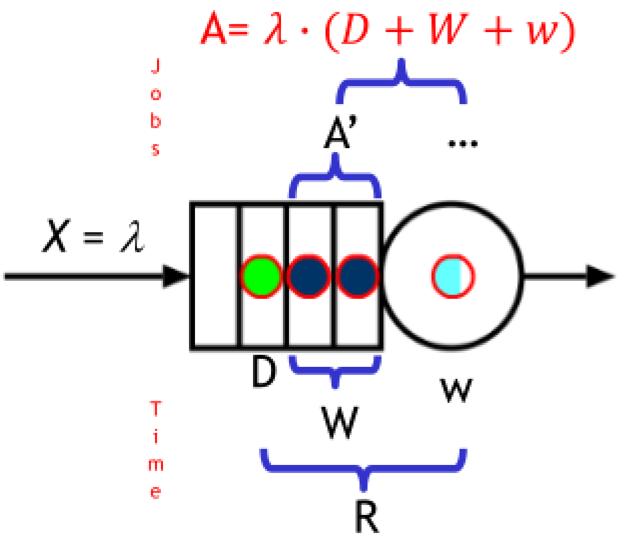
With a few derivations, A' can be expressed as function of W, leading to an equation from which W can be determined.
Since Poisson arrivals can basically occur at any time with the same probability, the number of jobs found at the arrival in a station is identical to the average number of jobs in that queue.

$$A = N$$



Applying Little's Law, this can be expressed as function of the Response Time:

$$A = N = XR = \lambda R = \lambda(D + W + w)$$



$$N_q = A' = N - U = A - U \Rightarrow A = A' + U$$

$$A' = A - U$$

$$A' = \lambda D + W + w - \lambda D$$

$$A' = \lambda W + w$$

We can use this result to express the waiting time due to other jobs in the queue as function of the average remaining time of the costumer currently in service.
The average response time R can than be computed as function of the remaining service time w:

$$R = D + W + w = D + \frac{w}{1-\rho}$$

Let us call k(t) the waiting time function. The average waiting time w can be computed as the time average of function k(t).

$$w = lim_{T\to\infty}[\frac{1}{T}\int_0^T k(t)dy] = \lim_{T\to\infty}\frac{C(T)}{T}\lim_{T\to\infty}\frac{1}{C(T)}\sum_{i=1}^{C(T)}\frac{x_i^2}{2}$$

This result permit the computation of the response time as:

$$R = D + \frac{\lambda m_2}{2(1-\rho)}$$

Using Little's law we can compute the average number of jobs in the system:

$$N = \rho + \frac{\lambda^2 m_2}{2(1-\rho)}$$

Remembering the relations between the first two moments, the variance and the coefficient of variation, we can write:

> $$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 = m_2 - D^2$$
>
> $$c_v^2 = Var[X]/D^2 \qquad Var[X] = D^2 c_v^2$$
>
> $$m_2 = D^2 + Var[X] \qquad m_2 = D^2(1 + c_v^2)$$
>
> $$N = \rho + \frac{\lambda^2 \cdot m_2}{2(1-\rho)}$$
>
> $\lambda^2 \cdot D^2 = \rho^2$
>
> $$N = \rho + \frac{\rho^2 + \lambda^2 \cdot Var[X]}{2(1-\rho)} = \rho + \frac{\rho^2(1 + c_v^2)}{2(1-\rho)}$$
>
> Average time spent in queue for an M/M/1
>
> $$R = D + \frac{\lambda \cdot m_2}{2(1-\rho)}$$
>
> $$\Theta = \frac{\rho D}{1-\rho}$$
>
> $$R = D + \frac{\rho^2 + \lambda^2 \cdot Var[X]}{2\lambda(1-\rho)} = D + \boxed{\frac{\rho D}{(1-\rho)}}\left[\frac{1 + c_v^2}{2}\right]$$

M/G/∞

M/G/∞ has results that are very simple and very similar to the ones for the M/M/∞. They are used a lot in telephony and traffic engineering. Results depend only on the mean of the distribution, and are insensitive to the higher moments.
Infinite servers, so when a job arrives it is served immediately. The average response time is equal to the average service time

$$U = \mu = \lambda E[X_G]$$

$$p_n = e^{-\rho}\frac{\rho^n}{n!}$$

$$N = U$$

$$R = E[X_G]$$

G/M/1

Arrival rate is $lambda = \frac{1}{E[X_A]}, U = \lambda D = \lambda/\mu$
G/M/1 models can be analyzed by solving an equation that involves the Laplace LG(s) transform of the distribution of the inter-arrival time, and the service rate m=1/D.

$$L_G(\sigma - \mu\sigma) = \sigma$$

Assuming that $E[X_A]$ is the average inter-arrival time, the various performance indices can then be computed in the following way:

owing way:

$$U = \frac{1}{\mu \cdot E[X_A]} \qquad X = \frac{1}{E[X_A]}$$

$$R = \frac{1}{(1-\sigma)\cdot\mu} \qquad N = \frac{1}{(1-\sigma)\cdot\mu\cdot E[X_A]} = \frac{U}{1-\sigma}$$

Since determining σ is extremely hard from a numerical point of view, this result has very few practical applications, and it is important mainly from the theoretical point of view.

G/M/c

G/M/c models are also characterized by analytical solutions (although quite complex).
The main idea is that the service process between two generally distributed arrivals follows a Markovian process, as for the G/M/1 queue.
In this case, however, the speed of service changes with the length of the queue, as it happens for the M/M/c queue.
When you do the proof of G/M/1 you can arrive at a point where you can also prove the G/M/c(even if this result is really difficult to find, completely useless).

G/G/1, M/G/c and G/G/c models

G/G/1, M/G/c and G/G/c do not have simple solution techniques.
Several bounding techniques are however available to determine upper and lower bounds for the considered systems (in particular for G/G/1).
We can write an equation to find the solution. This equation is an integral equation so the tools needed to resolve it are too complex. But this type of systems are really important for a theoretical point of view as we have a lot of system that can be described in these way there are simple bounding and generalization techniques.

G/G/c approximation: the Kingsman formula

The Kingsman formula gives an approximation of the G/G/c queue, starting from the average inter-arrival time $T = 1/\lambda$, average service time D, and their coefficient of variations,respectively c_a(arrival) and c_v(service). Here $E[\Theta_{M/M/c}]$ refers to the expected waiting time in the corresponding M/M/c queue with the same arrival rate and average service time.
Note that when arrivals are exponential and c_a=1, the formula corresponds exactly to the one for the M/G/1 queue.

$$R = D + [\frac{c_a^2 + c_v^2}{2}]E[\Theta_{M/M/c}]$$

Examples:

$$G/G/1 \qquad R \cong D + \left[\frac{c_a^2 + c_v^2}{2}\right]\frac{\rho D}{1-\rho} \qquad \rho = \frac{D}{T}$$

$$G/G/2 \qquad R \cong D + \left[\frac{c_a^2 + c_v^2}{2}\right]\frac{\rho^2 D}{1-\rho^2} \qquad \rho = \frac{D}{2T}$$

In the assignment we will use these formulas