POLITECNICO DI MILANO

Performance Evaluation and Applications





G/G/c queues

Motivating example

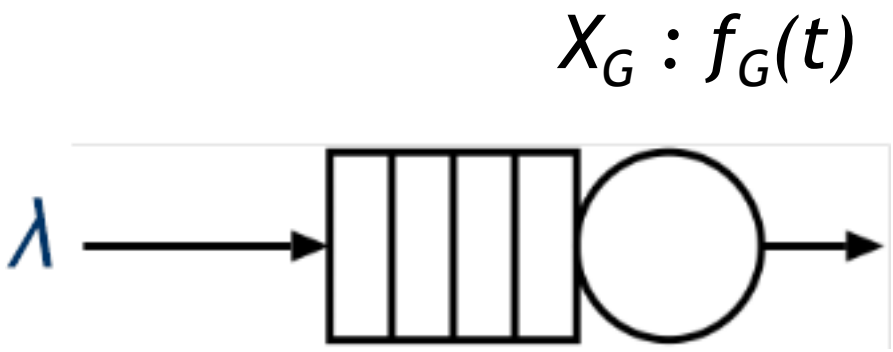
A work support center, receives costumers at the rate of 4 persons per hour. Such arrivals can be considered a Poisson process. Services instead can be of two types: 3/4 of them requires an average of 5 minutes, while the ¼ requires an average of 40 minutes. In this case, the exponential assumption is no longer be valid. Can we estimate analytically the average time each costumer will spend at the center, and on the average how many person will be inside the building?

?

M/G/1

M/G/1 systems are characterized by Poisson arrival rate λ and a general service time distribution.

Let us call X_G the general distribution, characterized by a PDF $f_G(t)$.



M/G/1

Let us define the *average service time D* and its *second moment* m_2 in the following way:

$$D = E[X_G] = \int_0^\infty t \cdot f_G(t) \cdot dt$$

$$m_2 = E[X_G{}^2] = \int_0^\infty t^2 \cdot f_G(t) \cdot dt$$

Let us also give the usual definition of traffic intensity ρ:

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

M/G/1

Let us recall how we computed moments of a distribution from a set of samples or measures:

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

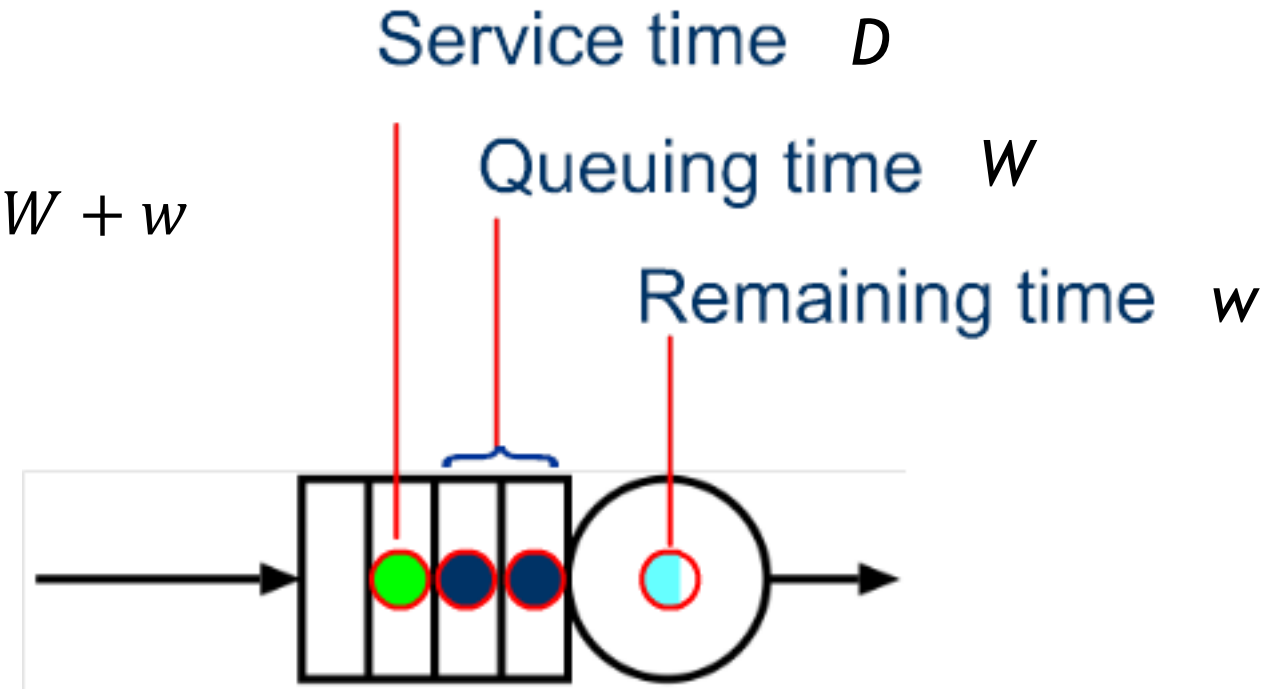
$$m_2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N} (X_G)^2\right] = E\left[{X_G}^2\right] = \int_0^{\infty} t^2 \cdot f_G(t) \cdot dt$$

M/G/1/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.

$$R = D + W + w$$

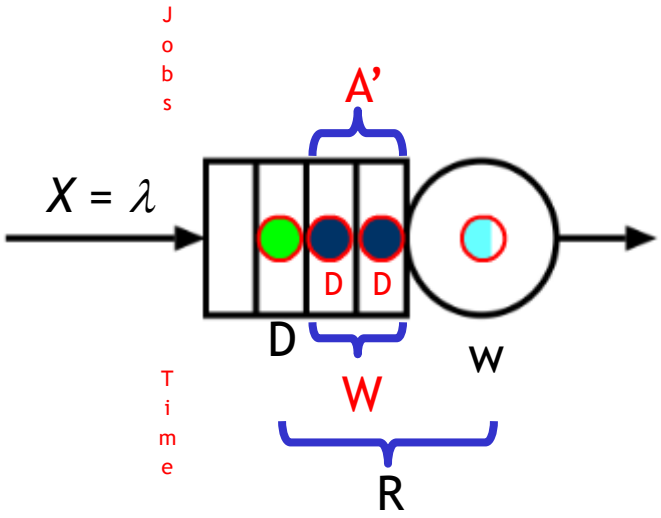


M/G/1

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:

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

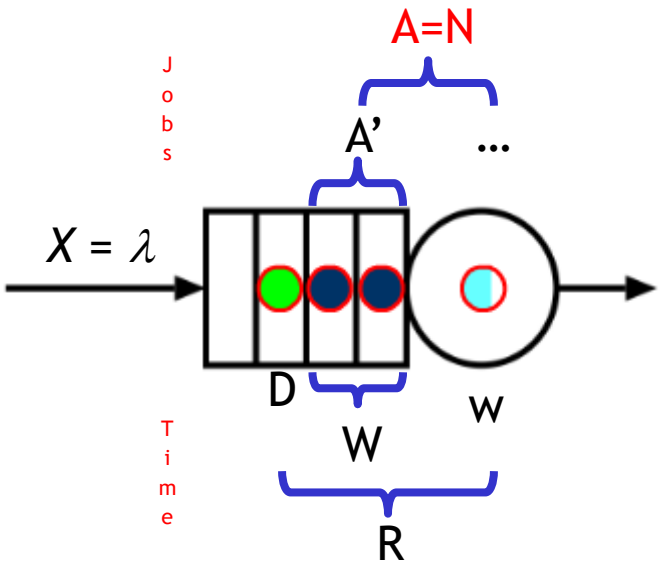
With a few derivations, *A'* can be expressed as function of *W*, leading to an equation from which *W* can be determined.



M/G/1

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.

$$A = N$$

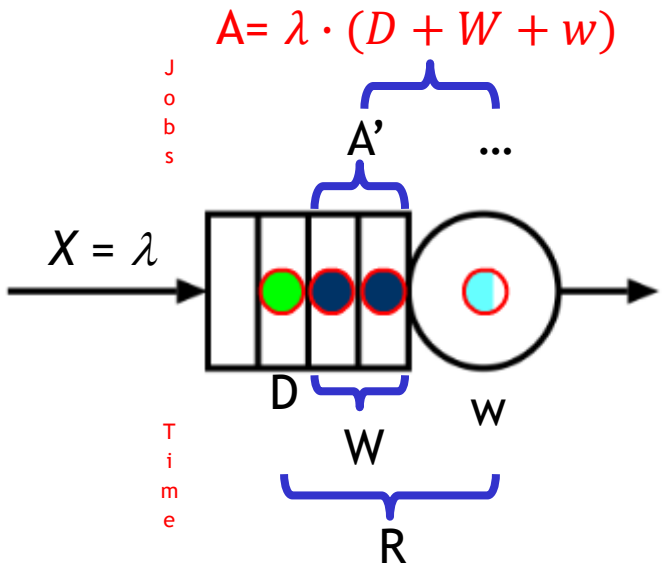


M/G/1

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

$$R = D + W + w$$

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



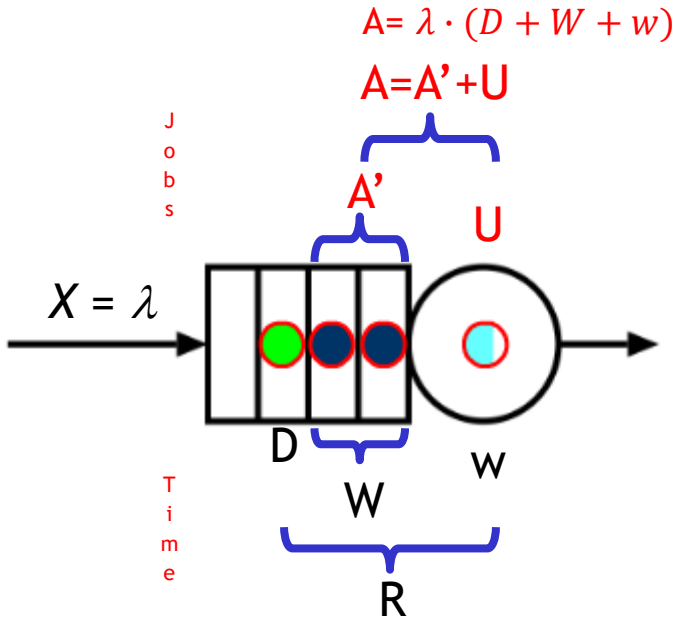
M/G/1

Since the fraction of job in service is, by definition, the utilization of the system, the total number of jobs found at the arrival can then further be expressed in two ways:

$$A = A' + U \qquad \text{and} \qquad A = \lambda \cdot (D + W + w)$$

Since by the *Utilization Law* we have U = XD = λD, we obtain:

$$A' = A - U$$
$$A' = \lambda \cdot (D + W + w) - \lambda D$$
$$A' = \lambda \cdot (W + w)$$



M/G/1

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.

$$W = A' \cdot D \qquad A' = \lambda \cdot (W + w) \qquad \rho = \lambda \cdot D$$

$$W = \lambda \cdot (W + w) \cdot D = \rho \cdot (W + w)$$

$$W \cdot (1 - \rho) = \rho \cdot w$$

$$W = \frac{\rho \cdot w}{1 - \rho}$$

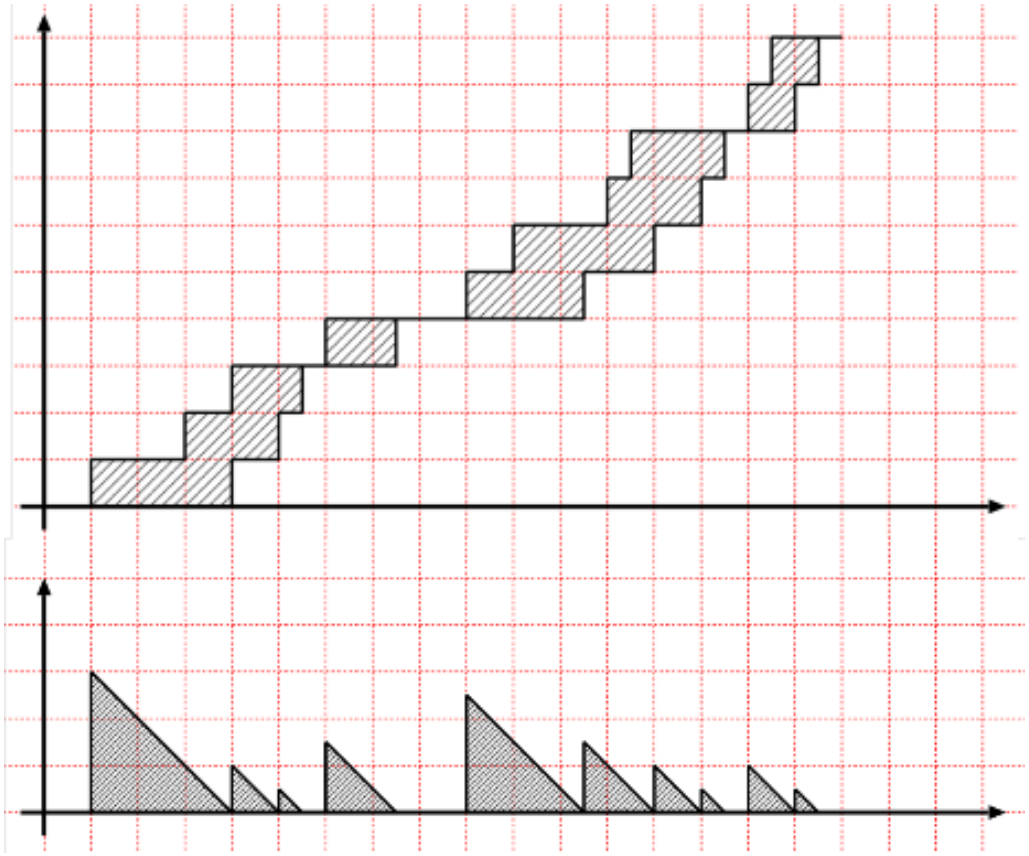
The average response time *R* can than be computed as function of the remaining service time *w*:

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

M/G/1

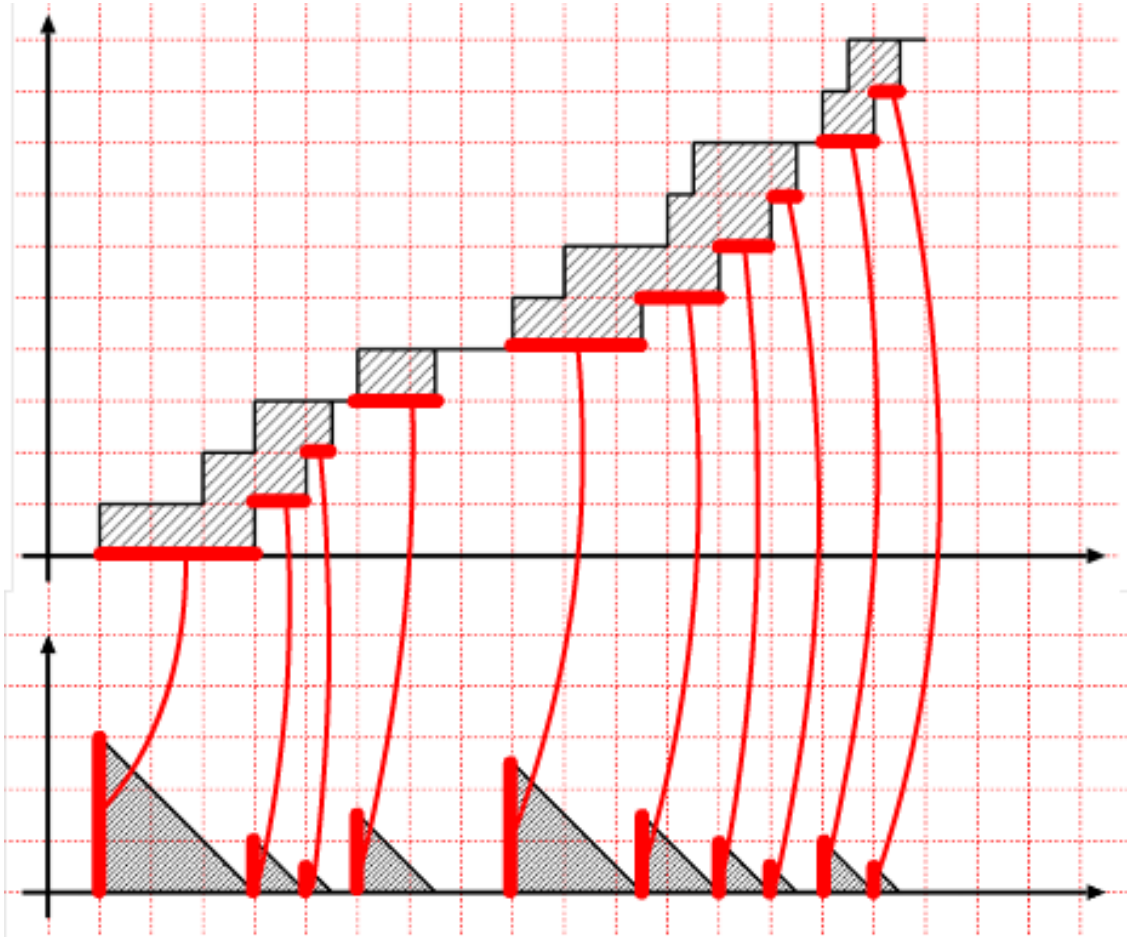
Let us consider a possible evolution of the system.

Thanks to the service in order assumption, we can determine the remaining service time for each instant from the arrivals and services plot.



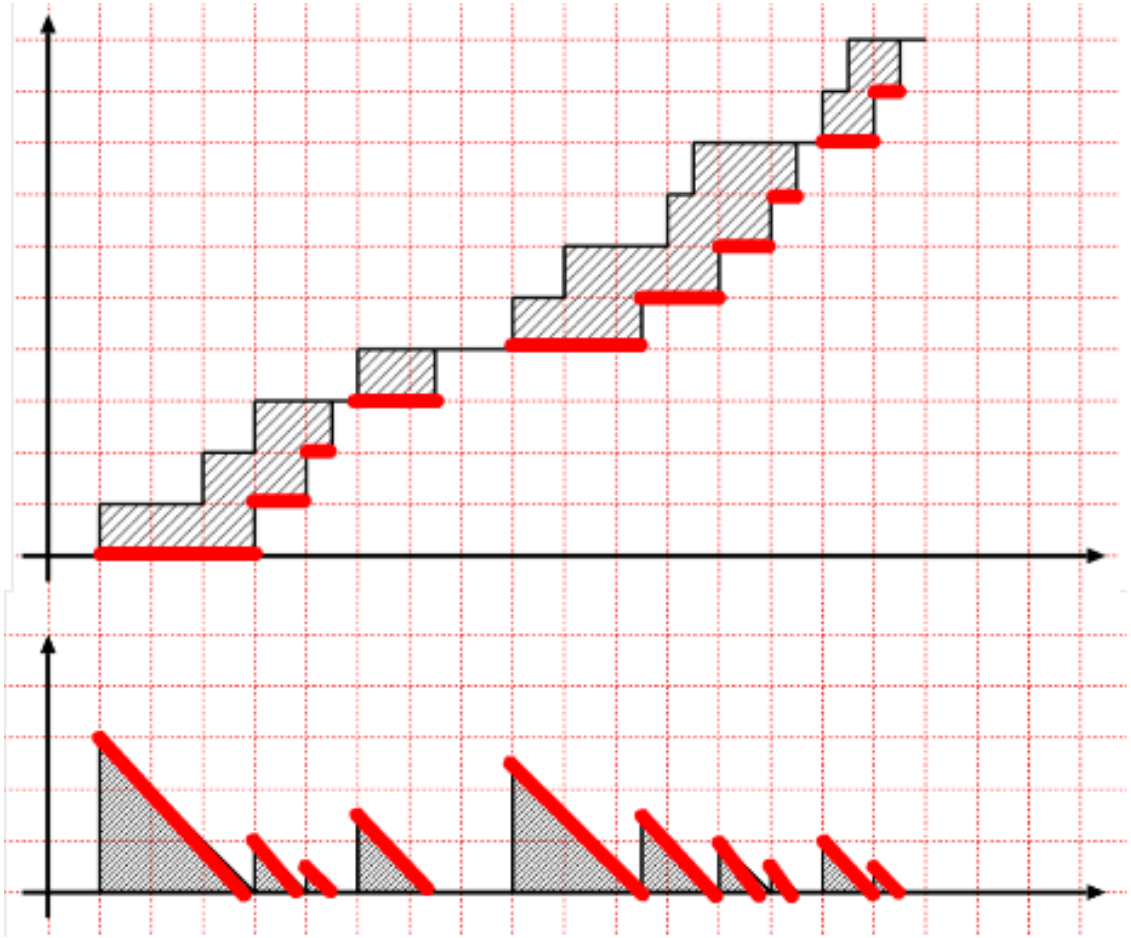
M/G/1

In particular, we have that after each service, the waiting time will have a jump corresponding to the service time of the next job.



M/G/1

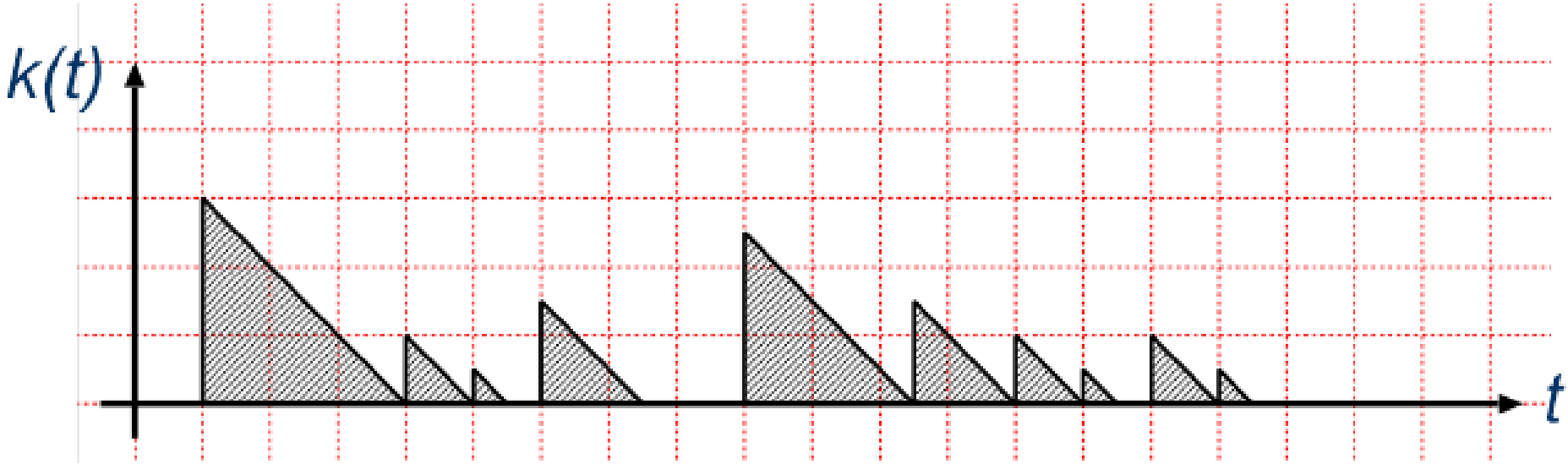
The jump will be followed by a linear decrease.



M/G/1

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)*.



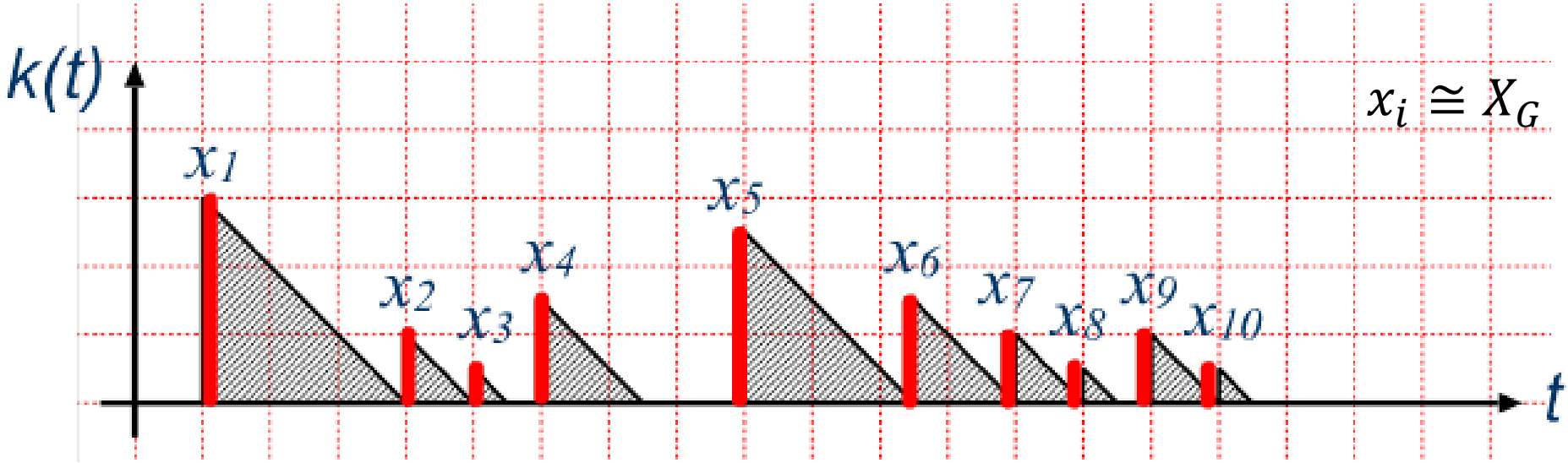
$$w = \lim_{T \to \infty} \left[\frac{1}{T} \cdot \int_0^T k(t) \cdot dt \right]$$

M/G/1

The jumps correspond to instances of the service time distribution.

In a time interval *T*, there will be *C(T)* jumps: one for each completion.

The integral of *k(t)* can then be computed as the area of the corresponding triangles (that have same base and height):



$$\int_0^T k(t) \cdot dt = \sum_{i=1}^{C(T)} \frac{{x_i}^2}{2}$$

M/G/1

If T tends to infinity, we have:

$$\int_0^T k(t) \cdot dt = \sum_{i=1}^{C(T)} \frac{{x_i}^2}{2}$$

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

Let us multiply and divide by *C(T)*.

Throughput of the system, or arrival rate, since it is stable

Second moment of the service distribution

$$w = \lambda \cdot \frac{E\left[{X_G}^2\right]}{2} = \frac{\lambda \cdot m_2}{2}$$

The first limit is the definition of throughput that, if a system is stable, corresponds to the arrival rate.

The second limit is the definition of the second moment of the service time distribution (divided by two).

M/G/1

The average response time *R* of a job is thus computed as:

$$w = \frac{\lambda \cdot m_2}{2}$$

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

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

Using Little's law, we can also compute the average number of jobs *N* in the system.

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

The previous relation is known as the *Pollaczek-Khinchine* formula.

M/G/1

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

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 = m_2 - D^2$$

$$c_v^2 = Var[X]/D^2 \qquad Var[X] = D^2 c_v^2$$

$$m_2 = D^2 + Var[X] \qquad m_2 = D^2(1 + c_v^2)$$

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

$\lambda^2 \cdot D^2 = \rho^2$

$$N = \rho + \frac{\rho^2 + \lambda^2 \cdot Var[X]}{2(1-\rho)} = \rho + \frac{\rho^2(1 + c_v^2)}{2(1-\rho)}$$

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

$$R = D + \frac{\rho^2 + \lambda^2 \cdot Var[X]}{2\lambda(1-\rho)} = D + \frac{\rho D}{(1-\rho)}\left[\frac{1 + c_v^2}{2}\right]$$

Average time spent in queue for an M/M/1

$\Theta = \frac{\rho D}{1-\rho}$

M/G/∞

$$U = \rho = \lambda \cdot E[X_G]$$

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

$$N = U$$

$$R = E[X_G]$$

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.

G/M/1 models can be analyzed by solving an equation that involves the Laplace $L_G(s)$ transform of the distribution of the inter-arrival time, and the service rate $\mu=1/D$.

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

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

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

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

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.

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).

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 and c_v.

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.

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

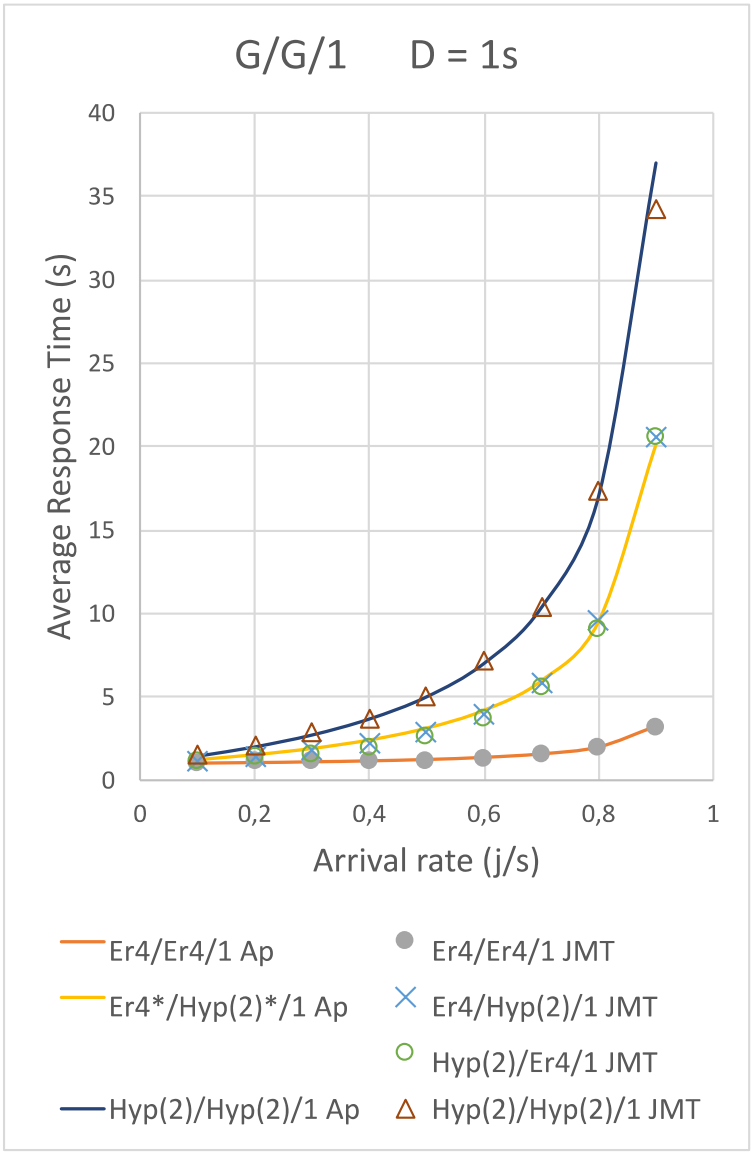
G/G/c approximation: the Kingsman formula

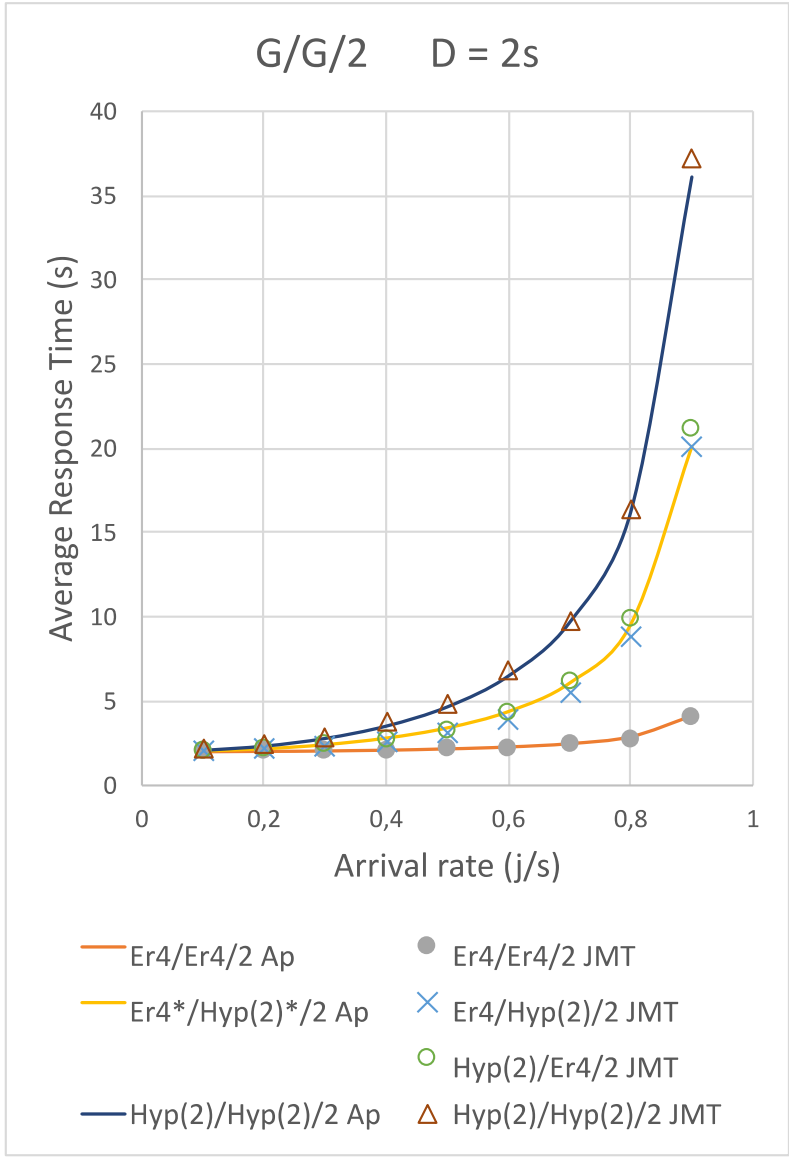
Examples:

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

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

G/G/c approximation: the Kingsman formula





Analysis of Motivating Example

We can assume and compute:

Arrivals: $\lambda = 4$ *cost. / h* *(exponential)*

Services: $\mu_1 = 12$, $\mu_2 = 1.5$, $p_1 = 0.75$ *(hyper-exponential)*

D = 0.75 / 12 + 0.25 / 1.5 = 0.2292 h = 13 min 45 sec

$m_2 = 2\,(0.75 / 12^2 + 0.25 / 1.5^2) = 0.23$

$\rho = 0.9167$

$$R = D + \frac{\lambda \cdot m_2}{2(1-\rho)} = 1.2652\ h$$

$$N = \lambda \cdot R = 5.06$$