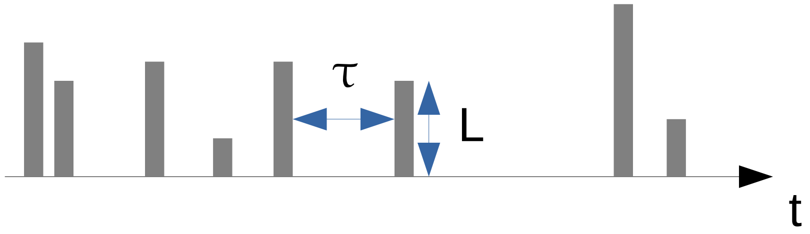
Network Engineering
M/G/1 queues

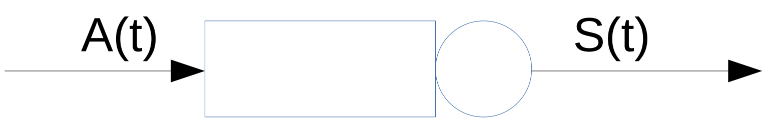
Boris Bellalta
boris.bellalta@upf.edu

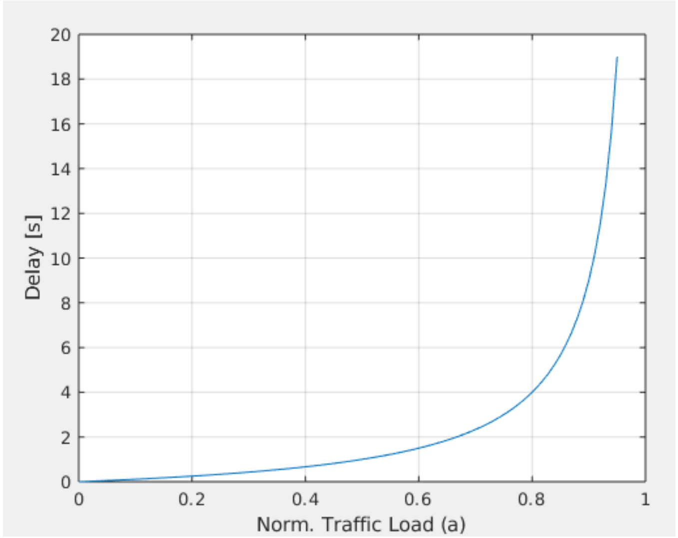
M/M/1

A(t) [arrival process]: L~expo, τ~expo



S(t) [Service process]: D_s~expo



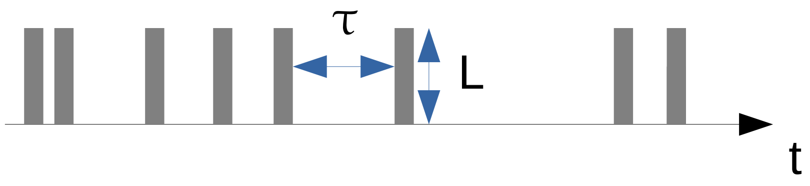


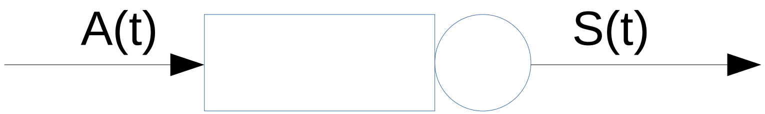
M/M/1

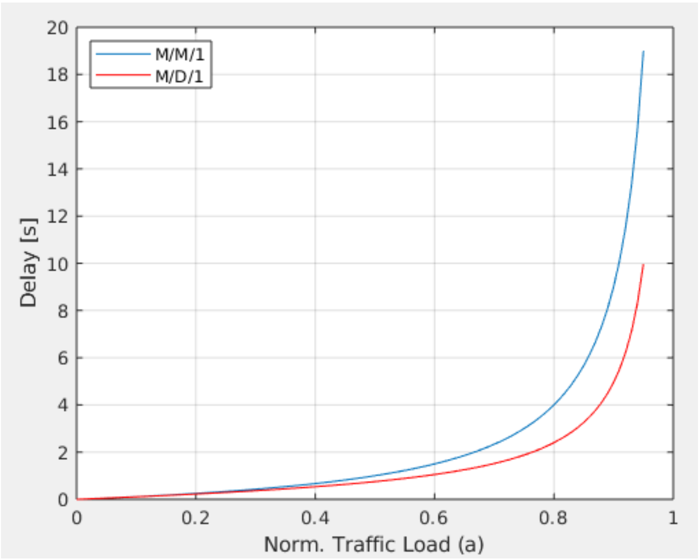
- What about if packet sizes are not exponentially distributed?

A(t) [arrival process]: L~det, τ~expo

S(t) [Service process]: D_s~det



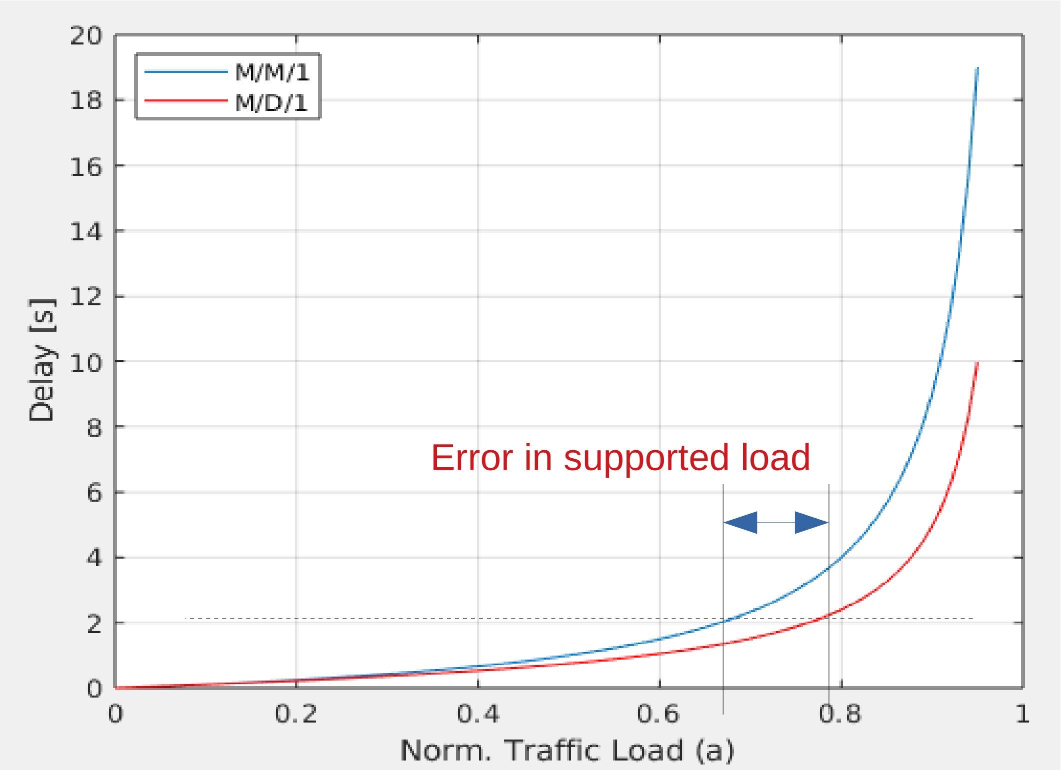




Our model is not accurate, there is an 'error' between the M/M/1 and the real case (M/D/1)

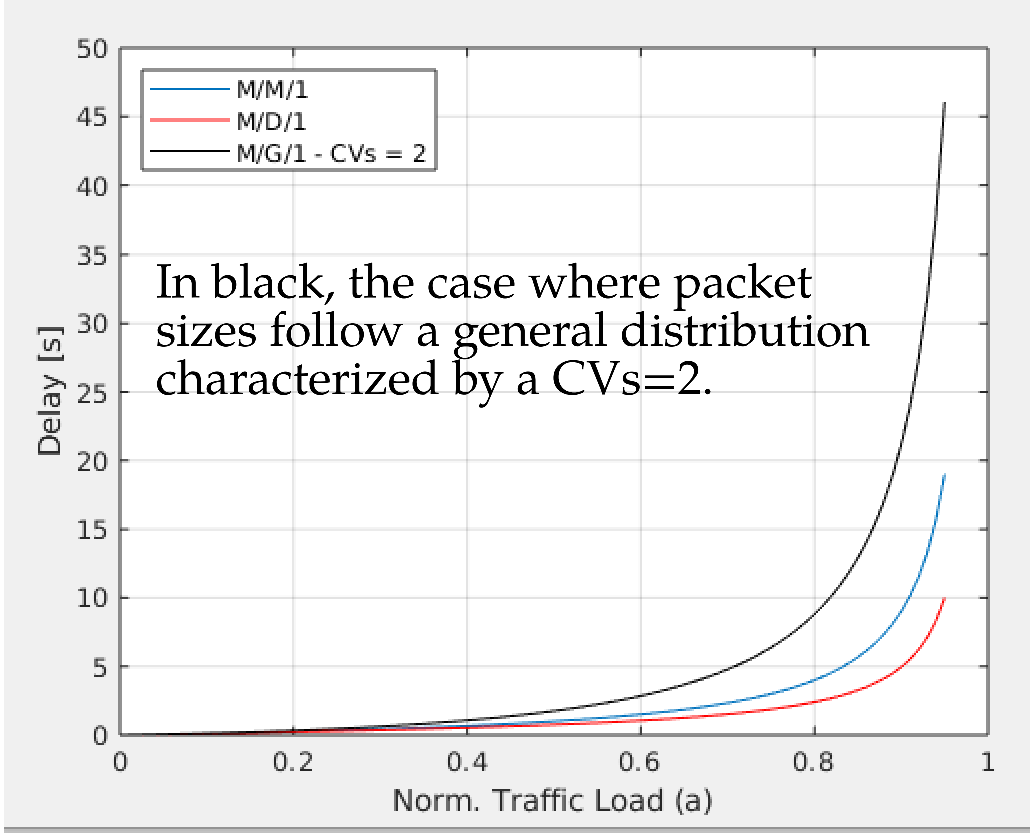
Example: maximum E[D] is 2 seconds

Using an M/M/1 model overestimates the system delay, so we are not able to use all system resources. (What is the maximum load to guarantee that E[D] is below 2 seconds?)



M/G/1

We don't know which distribution follows the packet size, but we can represent it through its:
- Coefficient of variation
- Second moment
- Variance



M/G/1 with CVs=1 → M/M/1
M/G/1 with CVs=0 → M/D/1

CVs increases → System performance decreases (more resources are needed to achieve the same performance)

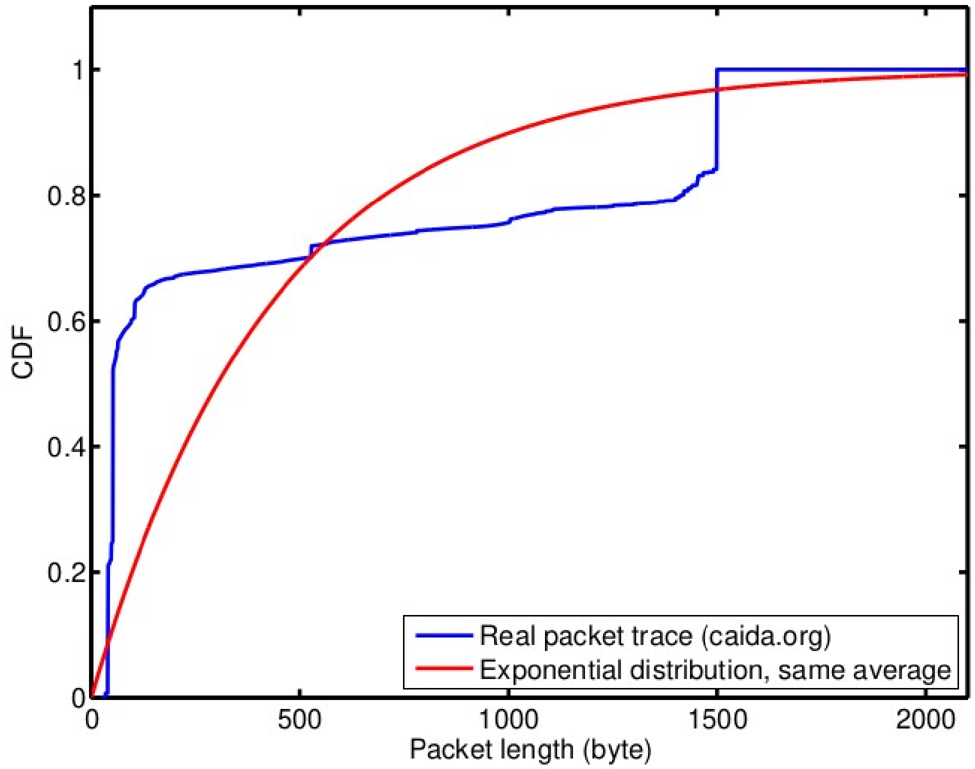
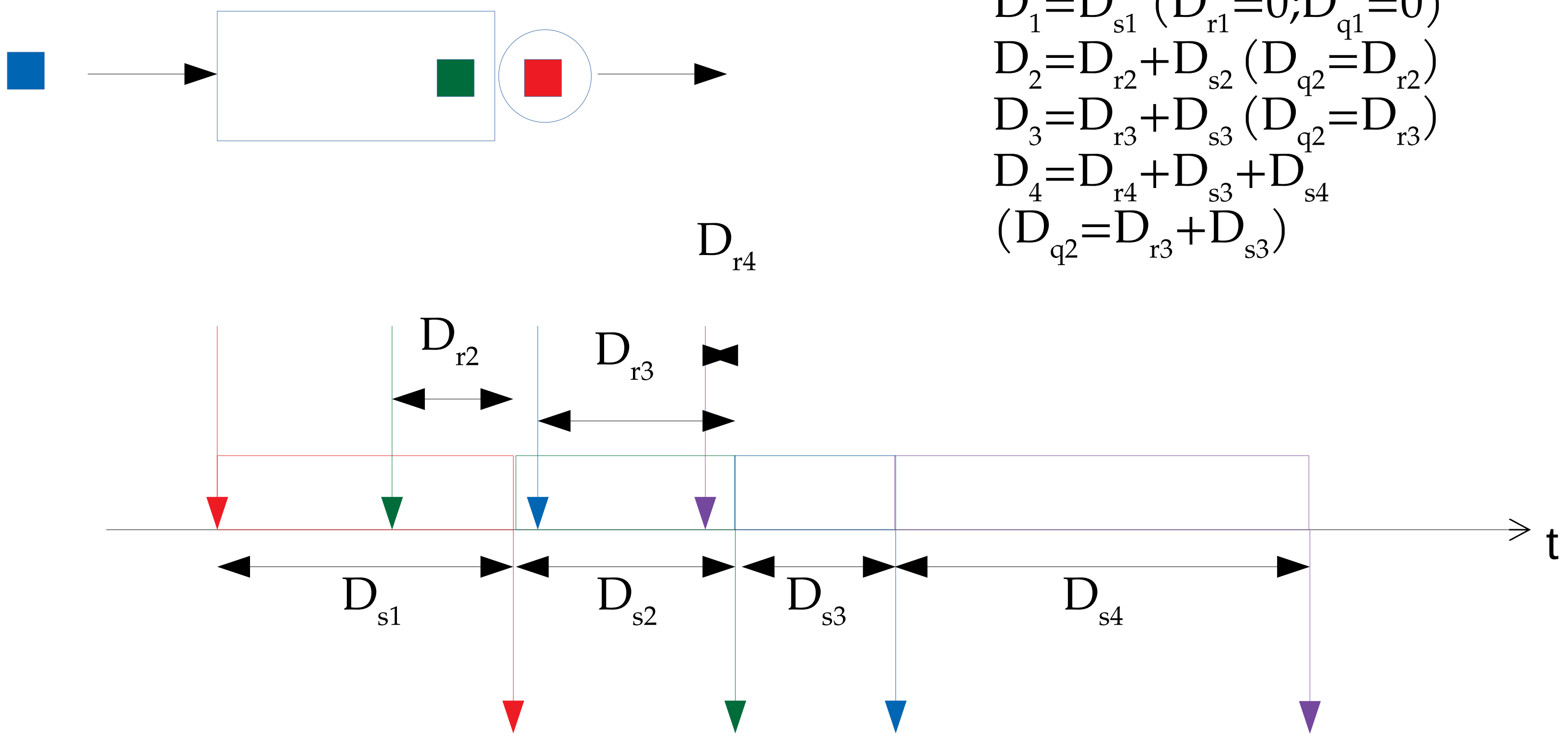


Figure 7.1: Real packet size distribution vs. exponential distribution with same mean

M/G/1



$D_1 = D_{s1}$ $(D_{r1} = 0; D_{q1} = 0)$

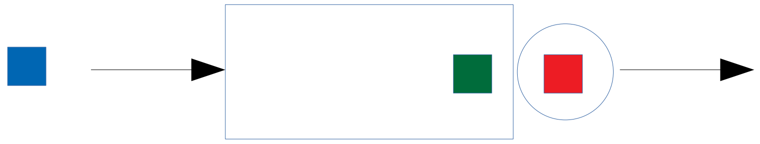
$D_2 = D_{r2} + D_{s2}$ $(D_{q2} = D_{r2})$

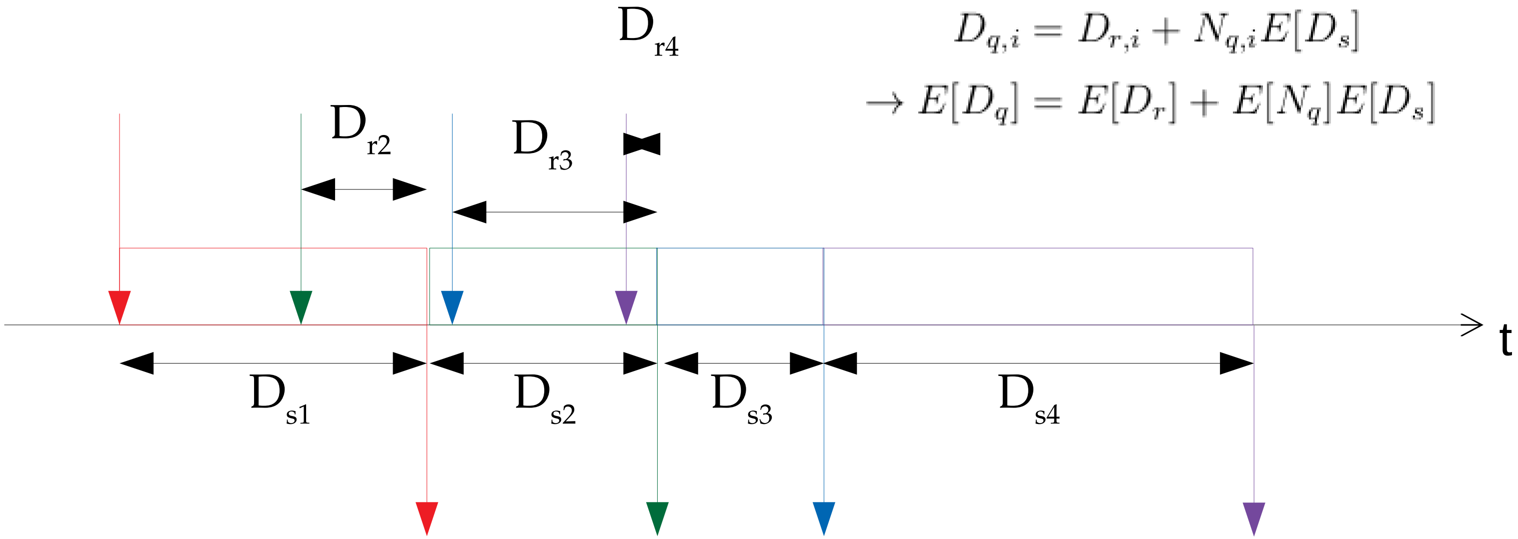
$D_3 = D_{r3} + D_{s3}$ $(D_{q2} = D_{r3})$

$D_4 = D_{r4} + D_{s3} + D_{s4}$

$(D_{q2} = D_{r3} + D_{s3})$

M/G/1





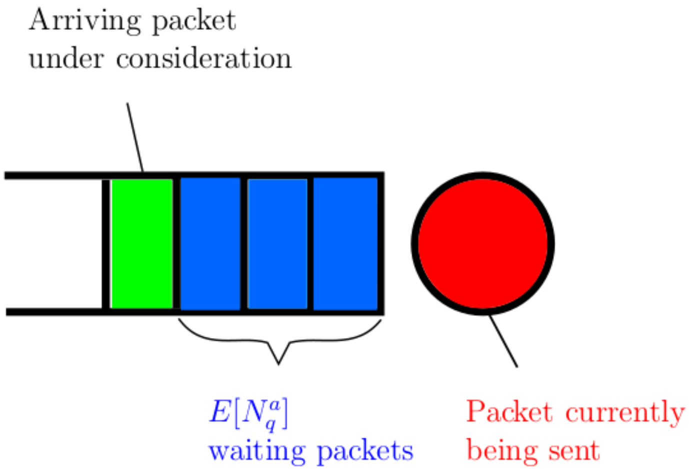
$$D_{q,i} = D_{r,i} + \sum_{n=1}^{N_{q,i}} D_{s,n}$$

$$D_{q,i} = D_{r,i} + N_{q,i}\left(\frac{1}{N_{q,i}}\sum_{n=1}^{N_{q,i}} D_{s,n}\right)$$

$$D_{q,i} = D_{r,i} + N_{q,i}E[D_s]$$

$$\rightarrow E[D_q] = E[D_r] + E[N_q]E[D_s]$$

M/G/1 – Queueing Delay



$$E[D_q] = \overbrace{E[N_q^a] \cdot E[D_s]} + \overbrace{E[D_r]}$$

Figure 7.3: Consideration for the average waiting time

$$\begin{aligned}
E[D_q] &= \lambda \cdot E[D_q] \cdot E[D_s] + E[D_r] \\
E[D_q] &= \rho \cdot E[D_q] + E[D_r] \\
E[D_q](1-\rho) &= E[D_r] \\
E[D_q] &= \frac{E[D_r]}{(1-\rho)}
\end{aligned}$$

$$\rho = a = \lambda \cdot E[D_s].$$

M/G/1 – Residual time

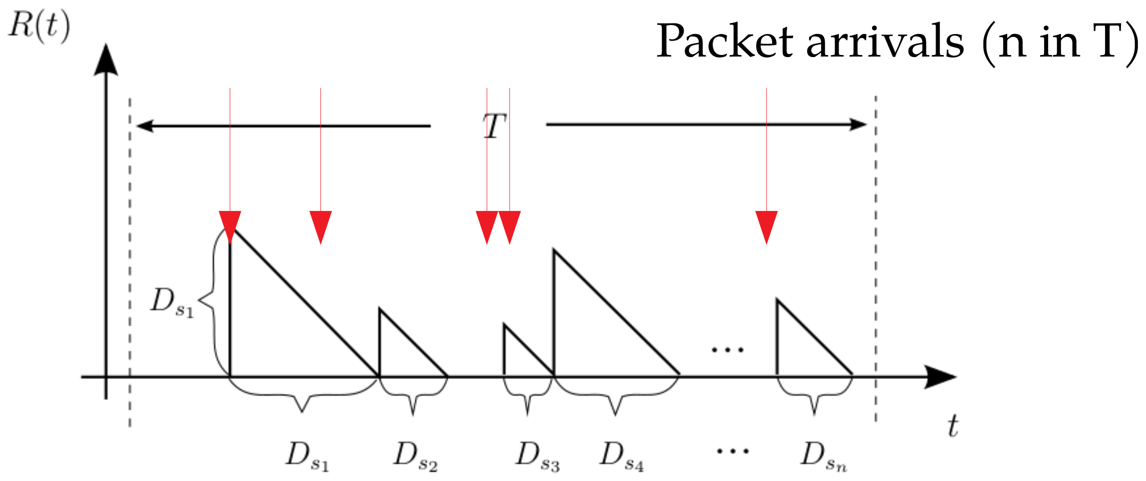


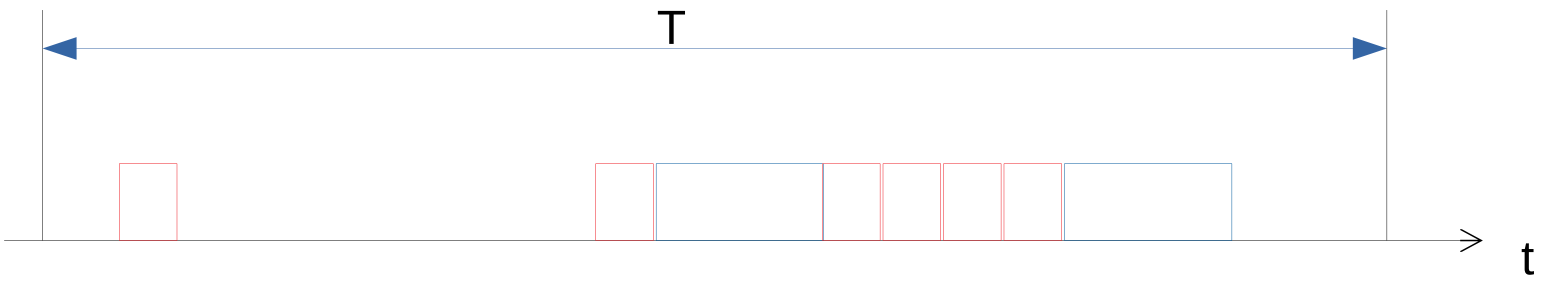
Figure 7.4: Residual service time process

For the average residual service time, we consider the system over a long timespan T. In this interval, we will see on average $\lambda \cdot T = n$ packets arriving. Then,

$$E[D_r] = \frac{1}{T}\int_0^T D_r(t')dt' = \frac{1}{T}\sum_{i=1}^{n}\frac{1}{2}D_{s_i}^2 = \underbrace{\frac{n}{T}}_{\to\lambda} \cdot \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n}\frac{1}{2}D_{s_i}^2}_{\to\frac{1}{2}E[D_s^2]} .$$

M/G/1 – Residual time

- Alternative way to calculate the residual time:
  - What is the probability to arrive when a packet ‘i’ is in service given that the ‘transmitter’ is occupied?
    - p_i E[D_{s,i}] / sum_{\forall j}{p_j E[D_{s,j}]}
  - We tradeoff the effect of amount of arrivals of ‘i’ packets and their size.
  - In average, the residual time when a packet arrives during the service time of a packet ‘i’ is E[D_{s,i}]/2.
- So, we get: \rho E[D^2_{s}]/(2E[D_{s}]), as we need to make explicit the condition that the transmitted is occupied.



Poisson arrivals → arrivals may happen at any instant of time with the same prob.

Given the server is busy:

p_red = 6 * T_red / (6*T_red + 2*T_blue) = 6/8 T_red / E[T]

p_blue = 2 * T_blue / (6*T_red + 2*T_blue) = 2/8 T_blue / E[T]

E[D_r|busy] = (T_red / 2)*p_red + (T_blau / 2)*p_blau =

= ((½)/E[T]) * ((6/8) T_red * T_red + (2/8) T_blue * T_blue) =

= (1/2)*E[T^2]/E[T]

E[D_r] = a * E[D_r|busy] + (1-a)*0 = lambda * (½) * E[T^2]

M/G/1

$$E[D_q] = \frac{E[D_r]}{(1-\rho)} = \frac{\lambda \cdot E[D_s^2]}{2(1-\rho)} \quad \left(= \frac{1+CV[D_s]^2}{2} \cdot \frac{\rho}{1-\rho} \cdot E[D_s] \right)$$

$$E[D] = E[D_q] + E[D_s] = \frac{\lambda \cdot E[D_s^2]}{2(1-\rho)} + E[D_s] \tag{7.5}$$

Applying Little's Law again using (7.4) and (7.5), we get the average number of packets in the queue and in the system, respectively:

$$E[N_q] = \lambda \cdot E[D_q] = \frac{\lambda^2 \cdot E[D_s^2]}{2(1-\rho)} = \frac{1+CV[D_s]^2}{2} \cdot \frac{\rho^2}{1-\rho} \tag{7.6}$$

$$E[N] = \lambda \cdot E[D] = \frac{\lambda^2 \cdot E[D_s^2]}{2(1-\rho)} + \lambda \cdot E[D_s] = \frac{1+CV[D_s]^2}{2} \cdot \frac{\rho^2}{1-\rho} + \rho \tag{7.7}$$

Application to M/M/1 waiting systems

Since the M/M/1 waiting system is a special case of the M/G/1 waiting system, we can apply the results for the latter and compare it with the results gained by the Markov chain-based approach. Since $CV[D_s] = 1$ in a M/M/1 system, we get

$$E[D_q] = \frac{1 + CV[D_s]^2}{2} \cdot \frac{\rho}{1-\rho} \cdot E[D_s] = \frac{\rho}{1-\rho} \cdot E[D_s],$$

and

$$E[N] = \frac{1 + CV[D_s]^2}{2} \cdot \frac{\rho^2}{1-\rho} + \rho = \frac{\rho^2}{1-\rho} + \rho = \frac{\rho^2}{1-\rho} + \frac{\rho(1-\rho)}{1-\rho} = \frac{\rho}{1-\rho},$$

which are the known formulas for M/M/1.

Application to M/D/1 waiting systems

As a second case for a specific class of service time distribution, we apply the M/G/1 analysis to a M/D/1 waiting system. Here, $CV[D_s] = 0$ due to the deterministic service process. Therefore,

$$E[D_q] = \frac{1 + CV[D_s]^2}{2} \cdot \frac{\rho}{1-\rho} \cdot E[D_s] = \frac{1}{2} \cdot \frac{\rho}{1-\rho} \cdot E[D_s],$$

or exactly half the average waiting time of a M/M/1 waiting system with the same average service time and load. Similarly,

$$E[N] = \frac{1 + CV[D_s]^2}{2} \cdot \frac{\rho^2}{1-\rho} + \rho = \frac{1}{2} \cdot \frac{\rho^2}{1-\rho} + \rho.$$

Traffic flows with multiple packet sizes - Exercise

- A traffic flow of load B= 8 Mbps contains the following packet sizes:
  - $L1=64$ Bytes (deterministic), with $p1=0.45$;
  - $L2=800$ Bytes (deterministic), with $p2 = 0.2$;
  - $L3 = 1500$ Bytes (deterministic), with $p3 = 0.35$
- The traffic flow arrives to a network interface, that transmits at a rate of $R=10$ Mbps.
- Calculate the waiting packet delay, and the total delay in the network interface.
- Compare the results if the M/M/1 queue was used.

The expected packet size is

$$E[L] = p_1(64 \cdot 8) + p_2(800 \cdot 8) + p_3(1500 \cdot 8) = 5710.4 \text{ bits}.$$

The expected service time is given by

$$E[D_s] = \frac{E[L]}{R} = p_1 \frac{64 \cdot 8}{10 \cdot 10^6} + p_2 \frac{800 \cdot 8}{10 \cdot 10^6} + p_3 \frac{1500 \cdot 8}{10 \cdot 10^6} = 0.571 \text{ ms}.$$

The second moment of the service time is given by

$$E[D_s^2] = p_1 \left(\frac{64 \cdot 8}{10 \cdot 10^6}\right)^2 + p_2 \left(\frac{800 \cdot 8}{10 \cdot 10^6}\right)^2 + p_3 \left(\frac{1500 \cdot 8}{10 \cdot 10^6}\right)^2 = 0.5871 \ \mu s^2. \quad (7.14)$$

The expected residual time is then given by

$$E[D_r] = \frac{\lambda E[D_s^2]}{2} = \frac{8 \cdot 10^6}{2 \cdot 5710.4} 0.5871 \cdot 10^{-6} = \frac{0.8225 \text{ ms}}{2} \quad (7.15)$$

The 2 was missing

Note that the residual time can be decomposed in the contributions of each packet size:

$$E[D_r] = E[D_{r,1}] + E[D_{r,2}] + E[D_{r,3}] = \frac{\lambda_1 E[D_{s,1}^2]}{2} + \frac{\lambda_2 E[D_{s,2}^2]}{2} + \frac{\lambda_3 E[D_{s,3}^2]}{2} \quad (7.16)$$

where $E[D_{s,i}^2] = \left(\frac{E[L_i]}{R}\right)^2 (1 + CV[D_{s,i}]^2)$.

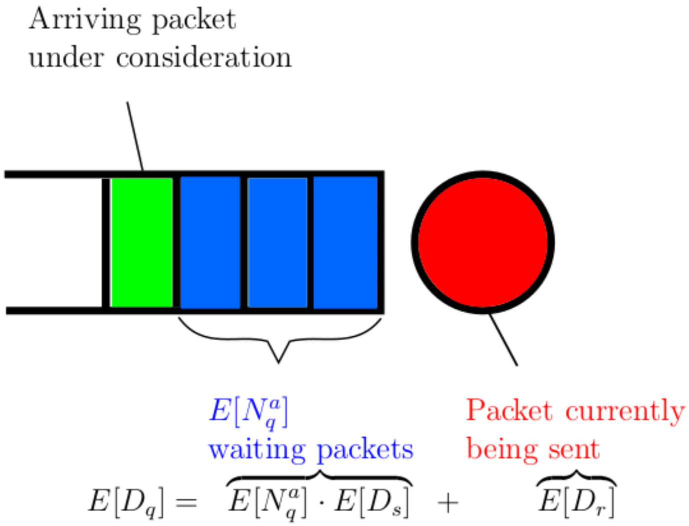


Figure 7.3: Consideration for the average waiting time

The expected queueing delay is

$$E[D_q] = \frac{E[D_r]}{1-\rho} = \frac{0.8225 \cdot 10^{-}3}{1-\rho} /2 = \frac{0.0021 \text{ s}}{2}$$

with $\rho = \lambda E[D_s] = \frac{8 \cdot 10^6}{E[L]} E[D_s] = 0.8225$

The expected system delay is

$$E[D] = \frac{E[D_r]}{1-\rho} + E[D_s] = \sout{0.0026 \text{ s}}\ 0.00155 \text{ s}$$

The CV of D_s is given by

$$CV[D_s] = \frac{\sqrt{V[D_s]}}{E[D_s]} = \frac{\sqrt{E[D_s^2] - E^2[D_s]}}{E[D_s]} = 0.8947$$

If we compare the obtained delay with the delay of a M/M/1 queue, we obtain:

$$E[D] = \frac{1}{\mu - \lambda} = \frac{1}{\frac{1}{E[D_s]} - \frac{B}{E[L]}} = 0.0029 \text{ s}$$

Traffic flows with multiple packet sizes - Exercise

- A traffic flow of load B= 8 Mbps contains the following packet sizes:
  - L1: E[L1]=64 Bytes (general, with CV[L1]=0.5), with p1=0.45;
  - L2: E[L2]=800 Bytes (general, with CV[L2]=1.2), with p2 = 0.2;
  - L3: E[L3] = 1500 Bytes (general, with CV[L3]=0.7), with p3 = 0.35
- The traffic flow arrives to a network interface, that transmits at a rate of R=10 Mbps.
- Calculate the waiting packet delay, and the total delay in the network interface.
- Compare the results if the M/M/1 queue was used.

Here, it's exactly the same as before. We first calculate the second moment of each type of size:

$$E[D_{s,i}^2] = E[D_{s,i}]^2(1 + CV[D_{s,i}]^2) \tag{7.21}$$

where $CV[D_{s,i}] = CV[L_i]$ as $D_{s,i} = \frac{L_i}{R}$, with R a constant.

Then, we can calculate the residual time:

$$E[D_r] = E[D_{r,1}] + E[D_{r,2}] + E[D_{r,3}] = \frac{\lambda_1 E[D_{s,1}^2]}{2} + \frac{\lambda_2 E[D_{s,2}^2]}{2} + \frac{\lambda_3 E[D_{s,3}^2]}{2} \tag{7.22}$$

Note that in previous exercise, since all packet sizes where deterministic, we just considered that $CV[L_i] = 0$ in all cases.