Network Engineering

M/G/1 queues

Boris Bellalta
boris.bellalta@upf.edu

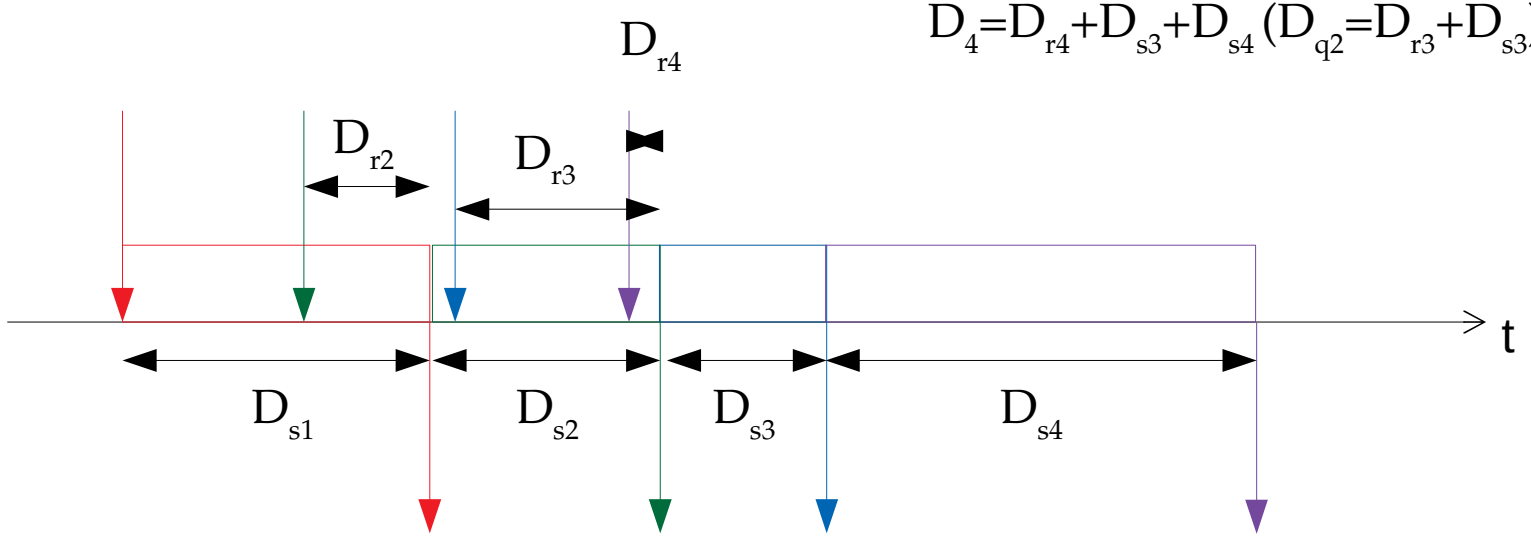
M/G/1

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

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

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

$D_4 = D_{r4} + D_{s3} + D_{s4}\ (D_{q2}=D_{r3}+D_{s3})$



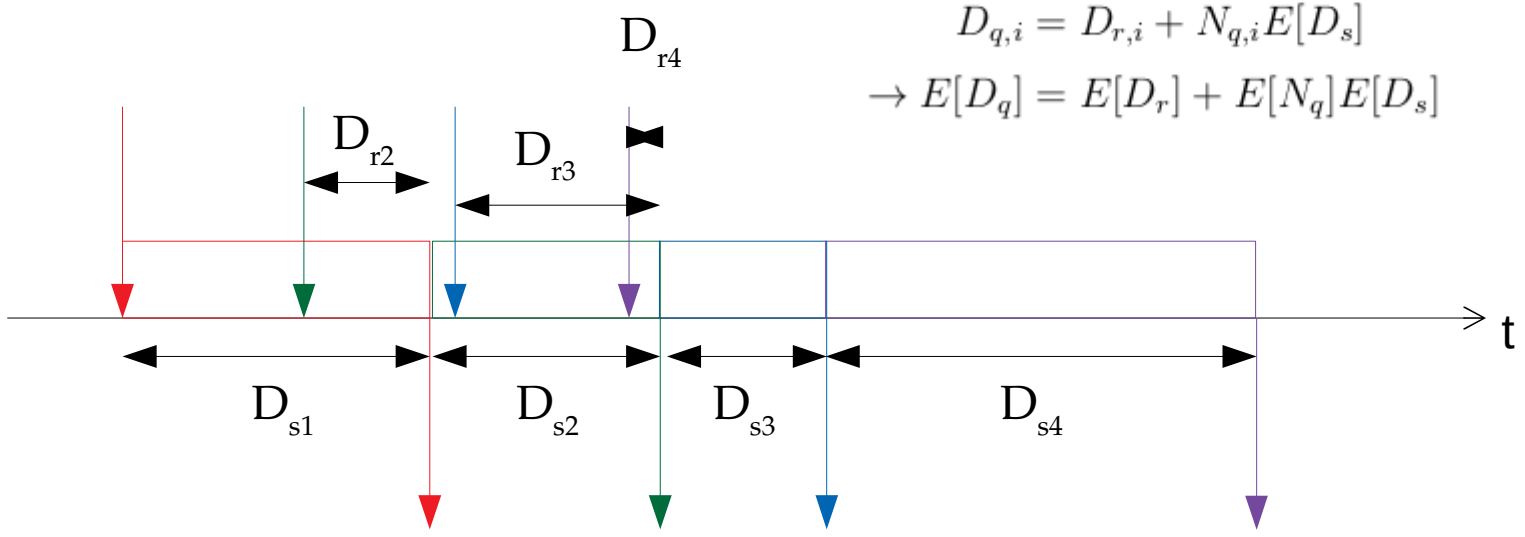
M/G/1

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

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

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

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



M/G/1 – Queueing Delay

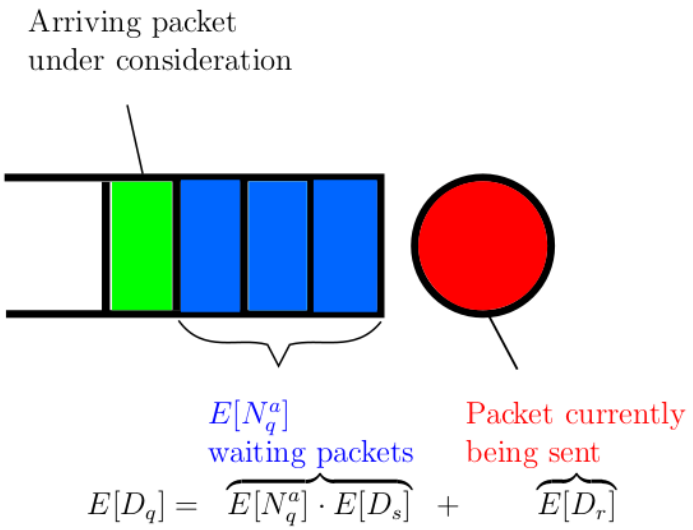


Figure 7.3: Consideration for the average waiting time

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

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

M/G/1 – Residual time

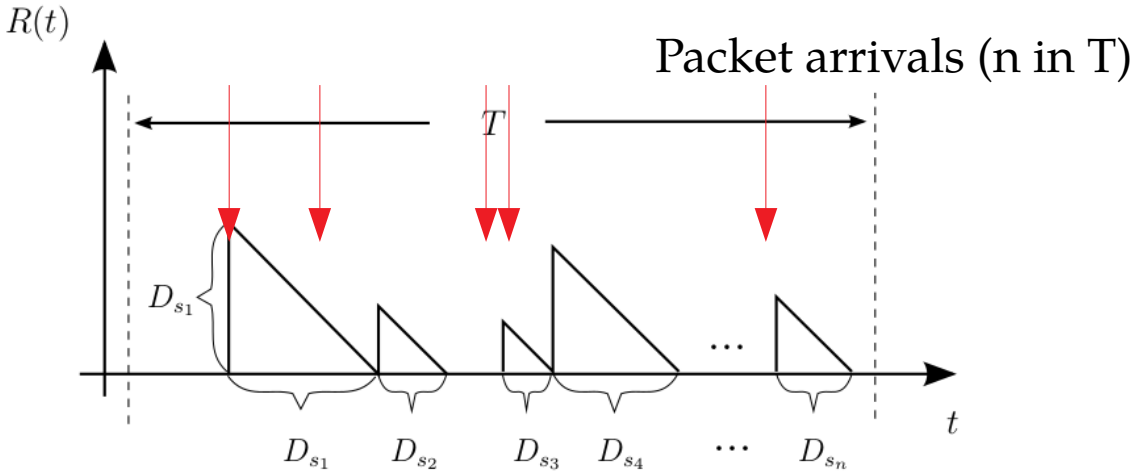
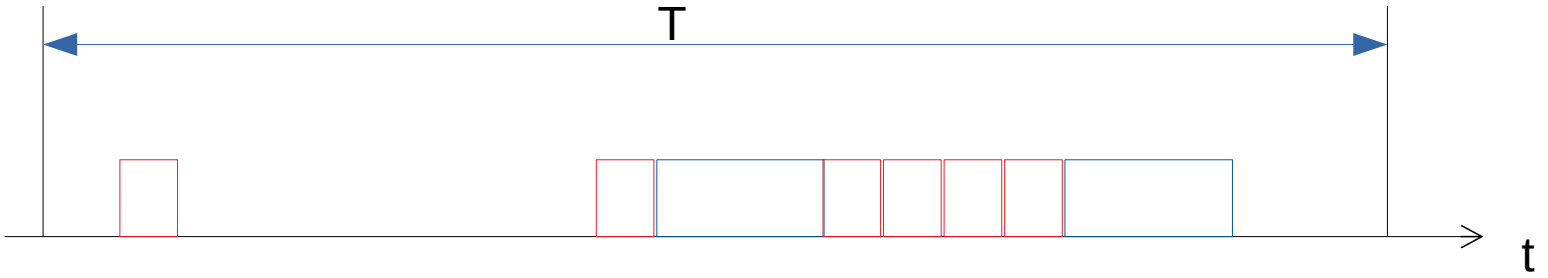


Figure 7.4: Residual service time process

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,

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



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_blue / 2)*p_blau =

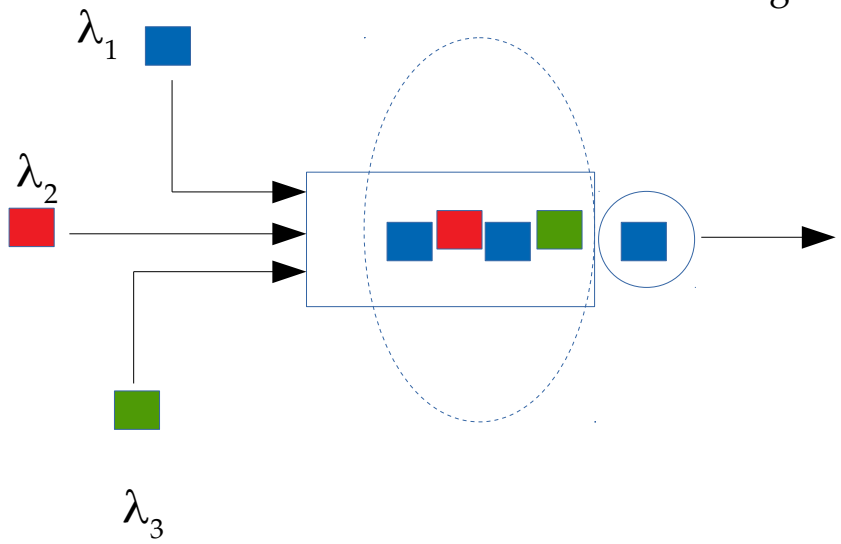
= ((½)/E[T]) * ((6/8) T_red * T_red + (2/8) T_blue * T_blue) =

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

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

Multiple Flows

An arriving packet finds in average the same mix of packets from other flows
in the buffer → The waiting delay in the buffer is the same in average for all the packets of all flows



F= 3 flows

$$E[D_q] = \sum_{f=1}^{F} E[N_{q,f}]E[D_{s,f}] + \sum_{f=1}^{F} a_f E[D_{r,f}|a_f]$$

$$E[D_q] = \sum_{f=1}^{F} E[N_{q,f}]E[D_{s,f}] + \sum_{f=1}^{F} a_f \frac{E[D_s^2]}{2E[D_s]}$$

$$E[D_q] = \sum_{f=1}^{F} E[N_{q,f}]E[D_{s,f}] + \sum_{f=1}^{F} \lambda_f E[D_{s,f}] \frac{E[D_{s,f}^2]}{2E[D_{s,f}]}$$

$$E[D_q] = \sum_{f=1}^{F} E[N_{q,f}]E[D_{s,f}] + \sum_{f=1}^{F} \lambda_f \frac{E[D_{s,f}^2]}{2}$$

$$E[D_q] = \sum_{f=1}^{F} E[N_{q,f}]E[D_{s,f}] + \sum_{f=1}^{F} E[D_{r,f}]$$

Multiple Flows - Example

- 2 Flows: λ_1 , λ_2, $E[D_{s,1}]$, $E[D_{s,2}]$, $CV[D_{s,1}]$, $CV[D_{s,2}]$

All waiting delays are the same

$$E[D_q] = E[N_{q,1}]E[D_{s,1}] + E[N_{q,2}]E[D_{s,2}] + E[D_{r,1}] + E[D_{r,2}]$$

$$E[D_q] = \lambda_1 E[D_q]E[D_{s,1}] + \lambda_2 E[D_q]E[D_{s,2}] + E[D_{r,1}] + E[D_{r,2}]$$

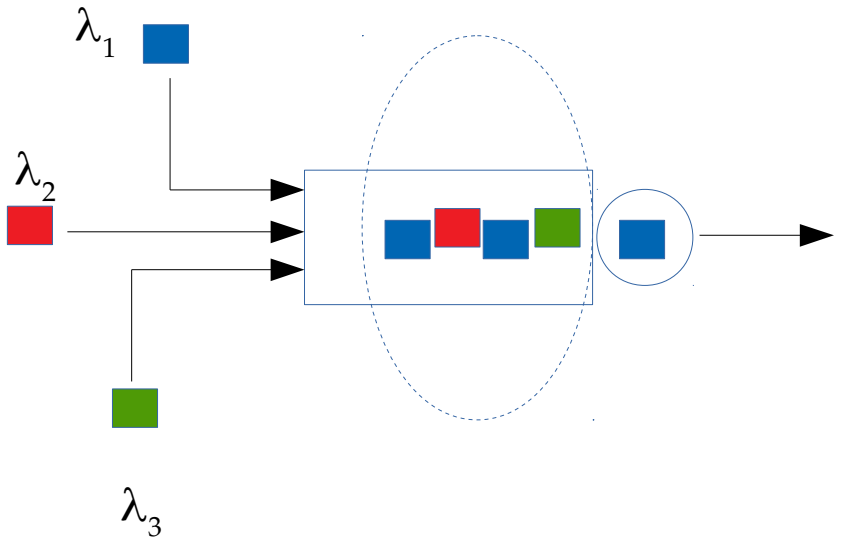
$$E[D_q](1 - a_1 - a_2) = E[D_{r,1}] + E[D_{r,2}]$$

$$E[D_q] = \frac{E[D_{r,1}] + E[D_{r,2}]}{1 - a_1 - a_2} = \frac{\lambda_1 \frac{E[D_{s,1}^2]}{2} + \lambda_2 \frac{E[D_{s,2}^2]}{2}}{1 - a_1 - a_2}$$

$$E[D_q] = \frac{\lambda \frac{E[D_s^2]}{2}}{1 - a_1 - a_2} = \frac{\lambda \frac{E[D_s^2]}{2}}{1 - a}$$

We consider the aggregate traffic

Average number of packets in the buffer



$$E[N_q] = \sum_{f=1}^{F} E[N_{q,f}]$$

$$E[N_{q,f}] = \frac{\lambda_f}{\lambda} E[N_q]$$

Example

As an example, we can imagine a link carrying file transfer and VoIP traffic, cf. Figure 8.1.

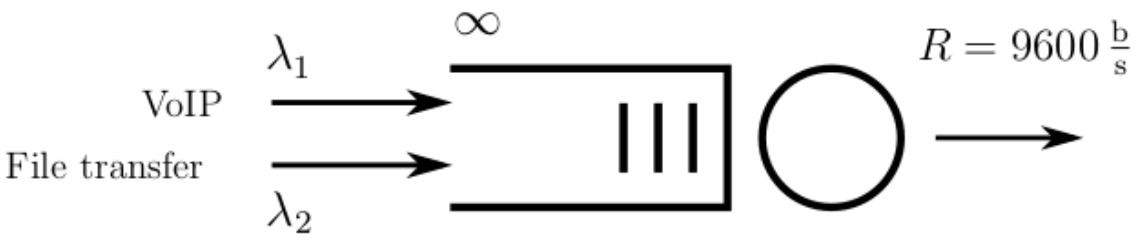
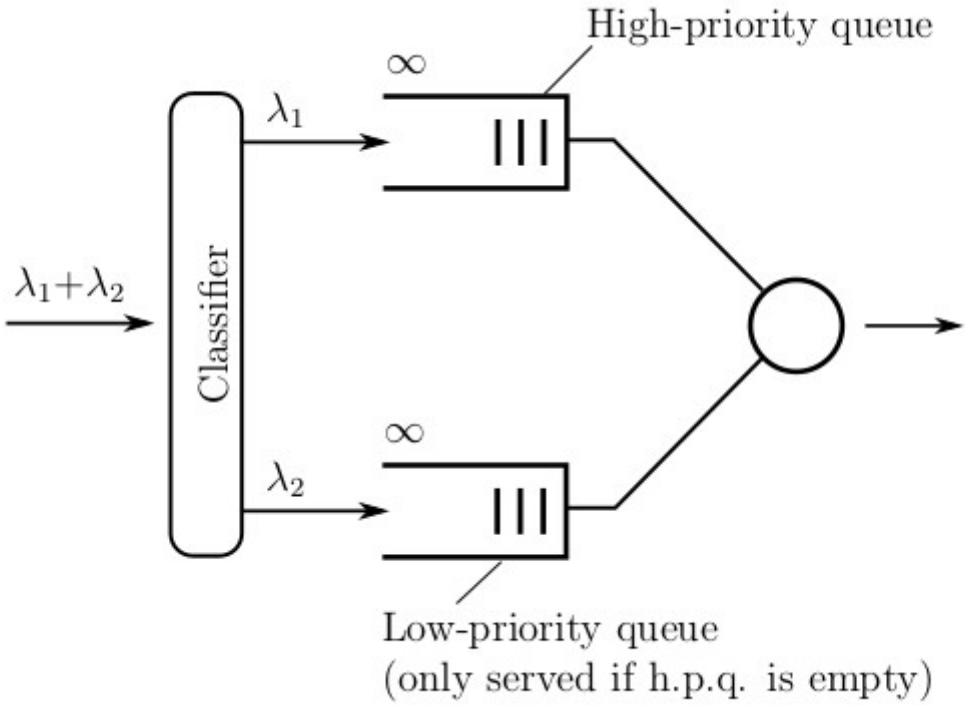


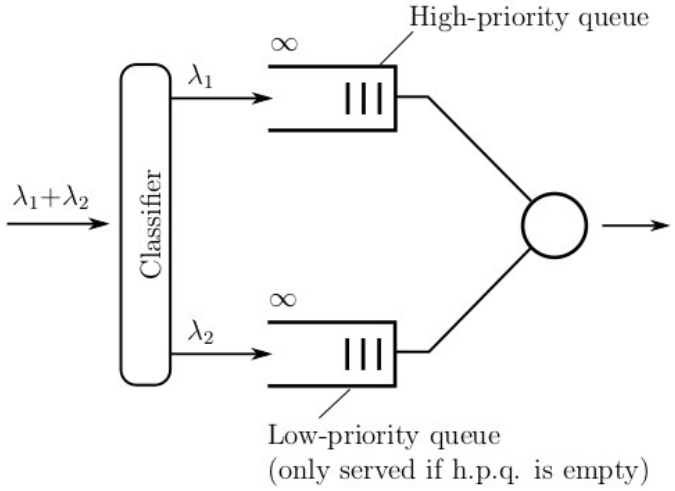
Figure 8.1: Link with two classes of packets

With a normal FIFO scheduling strategy, all packets would experience the same average waiting time, regardless of their type. For example, assume values of $L_1 = 48\,\text{bit}$, $\lambda_1 = 1.21\,\frac{1}{\text{s}}$ and $CV[D_{S_1}] = 0$ for the average packet length, arrival rate, and coefficient of variation for the VoIP packets, respectively, and the according values $L_2 = 960\,\text{bit}$, $\lambda_2 = 4.91\,\frac{1}{\text{s}}$ and $CV[D_{S_2}] = 1$ for the file transfer packets. Then, the average waiting time for all packets equals, using the M/G/1 waiting model, $E[D_q] = 97.64\,\text{ms}$.

Traffic Differentiation | Priority Systems



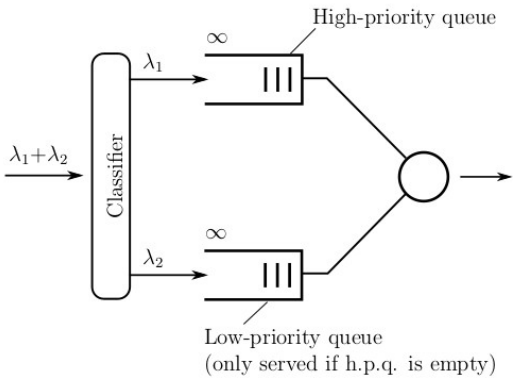
We now have to distinguish between two cases, i.e., whether the packet that arrived was a high-priority packet (priority class 1), or a low-priority packet (priority class 2). In the former case, the average waiting time of the packet is



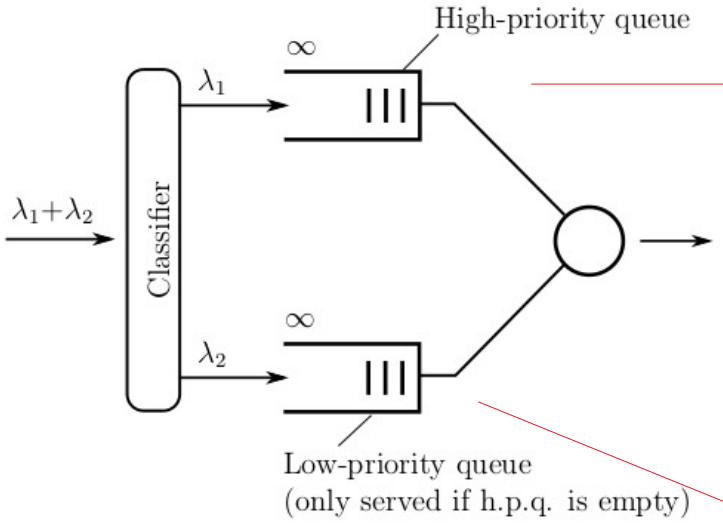
$$
\begin{aligned}
E[D_{q_1}] &= E[N_{q_1}] \cdot E[D_{S_1}] + E[D_r] \\
&= \lambda_1 \cdot E[D_{q_1}] \cdot E[D_{S_1}] + E[D_r] \\
&= \rho_1 \cdot E[D_{q_1}] + \frac{\lambda}{2} E[D_s^2] \\
\Rightarrow E[D_{q_1}] &= \frac{\lambda E[D_s^2]}{2(1-\rho_1)}
\end{aligned}
$$

For an arriving packet of low priority, the situation is slightly more complex. These packets have to wait until

1. the packet currently being serviced has finished
2. the packets of high priority found in the queue at arrival have been processed
3. the packets of low priority found in the queue at arrival have been processed
4. and finally, the packets of high priority that arrive during the waiting period have been processed as well (since they 'overtake' the packet under consideration).



$$
\begin{aligned}
E[D_{q_2}] &= \overbrace{E[N_{q_1}] \cdot E[D_{s_1}]}^{2.} + \overbrace{E[N_{q_2}] \cdot E[D_{s_2}]}^{3.} \\
&\quad + \underbrace{\lambda_1 \cdot E[D_{q_2}] \cdot E[D_{s_1}]}_{4.} + \underbrace{E[D_r]}_{1.} \\
&= \lambda_1 \cdot E[D_{q_1}] \cdot E[D_{s_1}] + \lambda_2 \cdot E[D_{q_2}] \cdot E[D_{s_2}] \\
&\quad + \lambda_1 \cdot E[D_{q_2}] \cdot E[D_{s_1}] + \frac{\lambda}{2} E[D_s^2] \\
&= \rho_1 E[D_{q_1}] + \rho_2 E[D_{q_2}] + \rho_1 E[D_{q_2}] + \frac{\lambda}{2} E[D_s^2]
\end{aligned}
$$



$$
\begin{aligned}
E[D_{q_1}] &= E[N_{q_1}] \cdot E[D_{S_1}] + E[D_r] \\
&= \lambda_1 \cdot E[D_{q_1}] \cdot E[D_{S_1}] + E[D_r] \\
&= \rho_1 \cdot E[D_{q_1}] + \frac{\lambda}{2} E[D_s^2]
\end{aligned}
$$

$$
\Rightarrow E[D_{q_1}] = \frac{\lambda E[D_s^2]}{2(1-\rho_1)}
$$

$$
\begin{aligned}
E[D_{q_2}](1-\rho_2-\rho_1) &= \rho_1 E[D_{q_1}] + \frac{\lambda}{2} E[D_s^2] \\
E[D_{q_2}] &= \frac{\rho_1 E[D_{q_1}] + \frac{\lambda}{2} E[D_s^2]}{(1-\rho_2-\rho_1)} \\
&= \frac{\frac{\rho_1 \frac{\lambda}{2} E[D_s^2]}{(1-\rho_1)} + \frac{\lambda}{2} E[D_s^2]}{(1-\rho_2-\rho_1)}
\end{aligned}
$$

$$
E[D_{q_2}] = \frac{\lambda E[D_s^2]}{2(1-\rho_2-\rho_1)(1-\rho_1)}
$$

Using the example from the beginning of the section, we can now calculate the waiting times for the VoIP and file transfer packets if we prioritize the VoIP packets. With the given values, we get $\rho_1 = 0.006$ and $\rho_2 = 0.491$, and thus $E[D_{q_1}] = 49.41\,\text{ms}$ and $E[D_{q_2}] = 98.24\,\text{ms}$. Comparing this with the result for a system without priorities, the VoIP packets, which only make up a small part of the load, now experience much shorter waiting times, while the data packets are affected by only slightly longer waits.

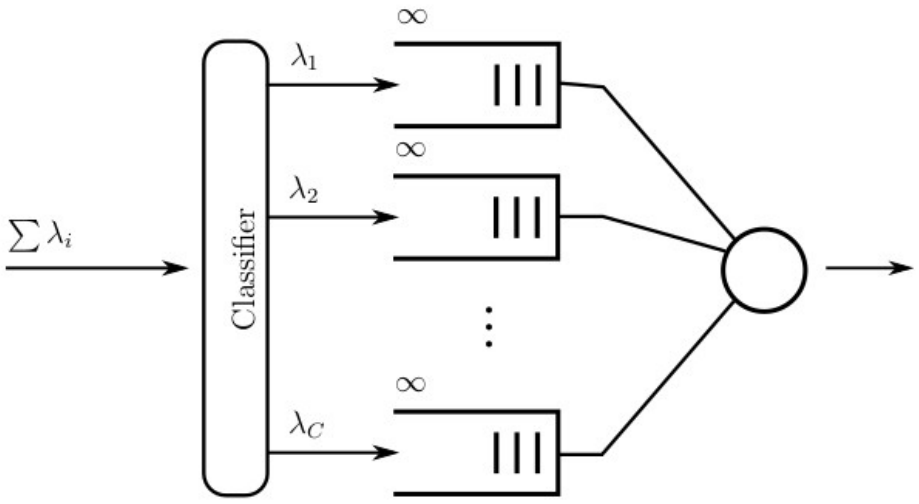


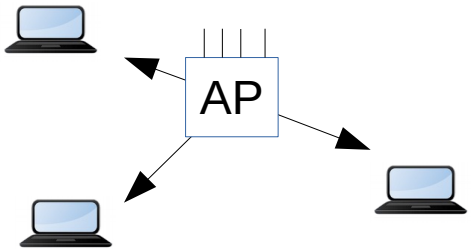
Figure 8.3: Generalized system with priority scheduling

$$E[D_{q_i}] = \frac{\lambda E[D_s^2]}{2\prod_{j=i-1}^{i}\left(1-\sum_{z=1}^{j} a_z\right)}$$

System delay (waiting + service)

- $E[D_1]=E[D_{s,1}]+E[D_{q,1}]$
- $E[D_2]=E[D_{s,2}]+E[D_{q,2}]$
- …
- $E[D_C]=E[D_{s,C}]+E[D_{q,C}]$

Example – DL Traffic Differentiation



- Calculate the average total delay for the packets of each flow:
- a) No traffic differentiation is applied
  - Single buffer for all flows
- b) Traffic differentiation is applied
  - Three buffers (2 per A, 1 per B and C)

- Station A has higher priority than B and C
- B and C have the same priority
- Flow 1 from A has higher priority than flow 2
- B_{A1}=0.1 Mbps; B_{A2}=2 Mbps;
- B_B=3 Mbps
- B_C=2 Mbps
- $E[D_{sA1}]$=0.5 ms; $CV[D_{sA1}]$=0;
- $E[D_{sA2}]$=2 ms; $CV[D_{sA2}]$=4;
- $E[D_{sB}]$=1.5 ms; $CV[D_{sB}]$=2.5;
- $E[D_{sC}]$=0.9 ms; $CV[D_sC]$=0.7;
- $E[L_{A1}]$=400 bits
- $E[L_{A2}]$=1500 Bytes
- $E[L_B]$=10000 bits
- $E[L_C]$=800 Bytes
- Transmission errors can be considered negligible