Priority queues

Consider an $M/G/1$ queue where the customers are divided into K priority classes, $k = 1, \ldots, K$:

- class 1 has the highest priority and class K the lowest priority
- the arrival rates of different classes are $\lambda_1, \ldots, \lambda_K$ (Poissonian)
- the expectation and second moment of the service time of different classes: $\overline{S}_k$ ja $\overline{S_k^2}$, $k = 1, \ldots, K$.

We derive mean results of the Pollaczek-Khinchinin type for this kind of queueing system.

- Priority queues are becoming more important also in telecommunication systems
  - in computer systems (e.g. operating systems) they have been exploited for a long time
- We are focusing on the so called time priority, which defines the service order of different classes
  - by giving a customer a higher priority one can reduce delay and delay variation
- For finite size queues (e.g. a finite buffer), a separate issue arises regarding the control of losses (overflow); this gives rise to the notion of space priority.

Non-preemptive priority

- The service of the customer being served is completed even if customers of higher priority may arrive.
- Each priority class has a separate (logical) queue.
- When the server becomes free, customer from the head of the highest priority non-empty queue enters the server.

Notation

$$\begin{cases} \bar{N}_q^{(k)} &= \text{mean number of waiting class-}k\text{ customers in the queue} \\ \overline{W}_k &= \text{mean waiting time of class-}k\text{ customers} \\ \rho_k &= \text{the load of class } k,\ \rho_k = \lambda_k \overline{S}_k \\ \overline{R} &= \text{the mean residual service time in the server (upon arrival)} \end{cases}$$

The stability condition of the queue:

$$\rho_1 + \cdots + \rho_K < 1$$

If the condition is violated, the queues belonging to priority clasees lower than some limit k (i.e. with a higher class index k) will grow without bound.

Non-preemptive priority (continued)

In the same way is in the derivation of the Pollaczek-Khinchinin mean results we deduce for the highest priority class 1:

$$\overline{W}_1 = \overline{R} + \overline{S}_1 \bar{N}_q^{(1)}$$

where the latter term represents the average time needed to serve the class-1 customers ahead in the queue

By Little's result we have

$$\bar{N}_q^{(1)} = \lambda_1 \overline{W}_1 \quad \Rightarrow \quad \overline{W}_1 = \overline{R} + \rho_1 \overline{W}_1 \quad \Rightarrow \quad \boxed{\overline{W}_1 = \frac{\overline{R}}{1-\rho_1}}$$

For priority class 2 we get analogously

$$\overline{W}_2 = \overline{R} + \underbrace{\overline{S}_1 \bar{N}_q^{(1)} + \overline{S}_2 \bar{N}_q^{(2)}}_{\text{time needed to serve class-1 and class-2 customers ahead in the queue}} + \underbrace{\overline{S}_1 \lambda_1 \overline{W}_2}_{\text{time needed to serve those customers in higher classes that arrive during the waiting time of class-2 customer}}$$

By Little's result we again get

$$\bar{N}_q^{(2)} = \lambda_2 \overline{W}_2 \quad \Rightarrow \quad \overline{W}_2 = \overline{R} + \rho_1 \overline{W}_1 + \rho_2 \overline{W}_2 + \rho_1 \overline{W}_2 \quad \Rightarrow \quad \overline{W}_2 = \frac{\overline{R} + \rho_1 \overline{W}_1}{1-\rho_1-\rho_2}$$

Non-preemptive priority (continued)

By inserting in the expression for $\overline{W}_2$ the formula for $\overline{W}_1$ we get

$$\overline{W}_2 = \frac{\overline{R}}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

By continuing in the same way to lower priority classes (higher values of k) we get the general result

$$\boxed{\overline{W}_k = \frac{\overline{R}}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}}$$

The total time in system of class-k customers is on the average

$$\boxed{\overline{T}_k = \overline{W}_k + \overline{S}_k}$$

The mean residual service time $\overline{R}$ appearing in $\overline{W}_k$ can be derived by the same kind of graphical "triangle trick" as in the case of the Pollaczek-Khinchinin mean value formula:

$$\boxed{\overline{R} = \frac{1}{2}\sum_{k=1}^{K} \lambda_k \overline{S_k^2}}$$

Observations about the non-preemptive priority

- The analysis cannot easily be extended for multiserver systems
  - the residual service time $\overline{R}$ is difficult to determine
  - can, however, be determined if the service time of all the classes has an exponential distribution with the same mean
- The mean waiting time of customer can be controlled by the choice of the priority class
  - if highest priority is given to those customers which have the shortest service times, then the mean waiting time of the whole customer population will be minimized
  - cf. the queue at a copying machine, where ususally a person with only one page to copy is given priority ovar those taking copy of a whole report
  - in the case of two priority classes the mean sojourn time is

$$\overline{T} = \frac{\lambda_1 \overline{T}_1 + \lambda_2 \overline{T}_2}{\lambda_1 + \lambda_2}$$

  - one can show that if $\overline{S}_1 < \overline{S}_2$, then $\overline{T}$ is smaller than in the case where the priorities were interchanged (or no priorities were used)
- The waiting time of also the highest priority class 1 depends on the traffic of the lower classes (on the λ_k) because the non-preemptivity means that the lower classes are not completely "invisible" to the higher classes.

Kleinrock's conservation theorem for non-preemptive priority queues

Kleinrock's conservation theorem states

$$\boxed{\sum_{k=1}^{K} \rho_k \overline{W}_k = \frac{\rho \overline{R}}{1-\rho}}$$

where $\rho = \rho_1 + \cdots + \rho_K$ is the total load of the system.

- The weighted sum of the waiting times can never change no matter how sophisticated the queueing discipline may be.
- Any attempt to modify the queueing discipline so as to reduce one of the $\overline{W}_k$ will force an increase in some of the other $\overline{W}_k$.

In particular, if all the classes have the same average service time, $\overline{S}_k = s$, one obtains upon division by s

$$\frac{\lambda \overline{R}}{1-\rho} = \sum_{k=1}^{K} \lambda_k \overline{W}_k \overset{\text{(Little)}}{=} \sum_{k=1}^{K} \bar{N}_q^{(k)} = \overline{N}_q$$

The average total number of waiting customers $\overline{N}_q$ is a constant independent of the queueing discipline (under the assumption of the same average service time).

Proof of Kleinrock's conservation theorem

The average unfinished work $\overline{V}$ can be divided into two parts

$$\overline{V} = \overline{V}_q + \overline{R}$$

The average unfinished work in the queue (that of the waiting customers) $\overline{V}_q$ can be written with the aid of Little's theorem

$$\overline{V}_q = \sum_k \overline{N}_q^{(k)} \overline{S}_k = \sum_k \lambda_k \overline{W}_k \overline{S}_k = \sum_k \rho_k \overline{W}_k$$

Thus

$$\sum_k \rho_k \overline{W}_k = \overline{V} - \overline{R}$$

$\overline{V}$ is independent of the service order (holds for the whole process $V(t)$). So is $\overline{R}$ since all the customers ultimately go through the server. This shows that $\Sigma_k \, \rho_k \overline{W}_k$ is constant.

Without changing the value we can calculate $\overline{V}$ for any service discipline. Let us calculate it for an ordinary FIFO system ($M/G/1$ queue) which consists of a single class.

$$\overline{V} = \sum_k \rho_k \overline{W}_k + \overline{R} = \rho \overline{W} + \overline{R} \overset{\text{(PASTA)}}{=} \rho \overline{V} + \overline{R} \quad \Rightarrow \quad \overline{V} = \frac{\overline{R}}{1-\rho}$$

By substituting in the previous equation one obtains the desired result $\displaystyle \sum_k \rho_k \overline{W}_k = \frac{\rho \overline{R}}{1-\rho}$

Preemptive resume priority

The service of customer is interrupted when on arrival of customer belonging to a higher class arrives, and will be resumed from the point of interruption, when all the queues of the higher priority classes have been emptied

- In this case the lower priority class customers are completely "invisible" and do not affect in any way the queues of the higher classes.
- In the case of packet queues, the preemptive priority is approximated if the sending of the packets takes place in small pieces, e.g. ATM cells which have (non-preemptive priorities)
  - when a packet of a higher priority arrives, the sending of the cells belonging to the lower priority packet is interrupted and is continued first when all the higher priority packets have been fully sent

Preemptive resume priority (continued)

We wish to calculate the mean sojourn time $\overline{T}_k$ of class-k customers. It is composed of three parts:

1. The customer's own mean service time $\overline{S}_k$
2. The mean time to serve the customers in classes $1, \ldots, k$ ahead in the queue

   $$\frac{\overline{R}_k}{1 - \rho_1 - \ldots - \rho_k} \quad \text{, missä } \overline{R}_k = \tfrac{1}{2} \textstyle\sum_{i=1}^{k} \lambda_i \overline{S_i^2},$$

   which is the same as the mean waiting time in an $M/G/1$ queue arising solely from the customers in classes $1, \ldots, k$. This is due to the fact that the unfinished work of classes $1, \ldots, k$ (not affected by classes $k+1, \ldots, K$) does not at all depend on the mutual service order of customers in classes $1, \ldots, k$ (since the unfinished work is "anonymous"),

   $$\mathrm{E}^*[V_k(t)]_{\text{pr. queue}} = \mathrm{E}^*[V_k(t)]_{\mathrm{M/G/1}} = \mathrm{E}[W_k]_{\mathrm{M/G/1}}$$

3. The mean time it takes to serve those customers in the higher classes $1, \ldots, k-1$ which arrive during the time the class-k customer stays in the system

   $$\sum_{i=1}^{k-1} \overline{S}_i \lambda_i \overline{T}_k = \sum_{i=1}^{k-1} \rho_i \overline{T}_k, \quad k > 1 \quad (0, \text{ jos } k = 1)$$

Preemptive resume priority (continued)

By collecting the results together we obtain

$$\overline{T}_k = \overline{S}_k + \frac{\overline{R}_k}{1-\rho_1-\cdots-\rho_k} + \Big(\sum_{i=1}^{k-1}\rho_i\Big)\,\overline{T}_k$$

$$\begin{aligned}
\overline{T}_1 &= \frac{(1-\rho_1)\overline{S}_1 + \overline{R}_1}{1-\rho_1} \\
\overline{T}_k &= \frac{(1-\rho_1-\cdots-\rho_k)\overline{S}_k + \overline{R}_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}
\end{aligned}$$