Queuing theory

Giovanni Righini

Università degli Studi di Milano

Logistics

Queues

Queues occur when a service is provided by a limited number of resources to a population of users or customers.

Queuing systems are common in logistics. For instance:

- jobs must be processed by a given machine in a production plant;
- people line up to obtain service service (bank, post-office,...);
- machines ask for repair operations by some specialized workers;
- patients wait for treatment in a First-Aid center;
- incoming calls are received by a call center;
- vehicles form lines when paying tolls to enter/leave highways or when traversing state borders;
- distributed processes ask for access to a central server in a computerized system;
- etc...

Queuing theory

Queuing theory aims at studying queuing systems in a scientific and quantitative way, to optimize their performance and cost.

It concerns both analysis and design of queuing systems.

Design is often done via simulation and repeated analysis of several alternative scenarios, because constraints and objectives are typically non-linear and non-deterministic.

Analysis of queuing systems (1)

A model of a queuing system is described by the following four main components.

- A source representing the generation of demand:
  - it may come from a *finite* or an *infinite* population;
  - it is described by the interarrival time, i.e. the time interval between two consecutive arrivals of customers in the system. The interarrival time is usually modeled with a random variable, described by a *probability distribution*.

Analysis of queuing systems (2)

- A queue with *finite* or *infinite* capacity.
- A discipline that is used to select which customer must be served among those in the queue: First-In-First-Out, priority-based,. . . .
- A service center represented by a number of parallel servers, providing the same service to the same queue. It is described by the service time, i.e., the time needed to provide service to each customer. The service time is usually modeled with a random variable, described by a *probability distribution*.

Classification

Queuing systems are classified with a three fields notation:

- the first field indicates the probability distribution of the interarrival time;
- the second field indicates the probability distribution of the service time;
- the third field indicates the number of parallel servers.

Typical distributions in the first two fields are:

- M: Markovian,
- D: degenerate,
- E_k Erlang with parameter k,
- G: generic.

Relevant quantities (1)

The main quantities that are relevant to the analysis of a queuing system are:

- s, the number of parallel servers;
- n, the number of users in the system; it includes those receiving service as well as those waiting in queue;
- $P_n(t)$, the probability that there are n users in the systems at time t;

Relevant quantities (2)

- λ_n, the mean arrival rate when there are n users in the system;
- λ, the mean arrival rate when it does not depend on n;
- μ_n, the mean completion rate for each server, when there are n users in the system;
- μ, the mean completion rate for each server, when it does not depend on n;
- $1/\lambda$, the mean arrival time;
- $1/\mu$, the mean service time for each server.

The utilization factor

When $n < s$ users are in the system, $s - n$ servers are idle and the system completion rate is $n\mu$.

When $n \geq s$ users are in the system, all servers are busy and the system completion rate is $s\mu$.

The utilization factor of a queuing system is

$$\rho = \frac{\lambda}{s\mu}.$$

It represents the mean fraction of time for which each server is busy.

Steady-state conditions

When a queuing system is in a transient condition, the relevant quantities may depend on time t.

When a queuing system is in a steady-state condition, the relevant quantities do not depend on time t.

In general a queuing system reaches a steady-state condition if and only if its utilization factor is strictly less than 1:

$$\rho < 1.$$

Otherwise the system explodes, i.e. the size of the queue goes to infinity.

Relevant performance indicators

When we analyze a queuing system in a steady-state condition, we are mainly interested in these five performance indicators:

- P_n: probability that n users are in the system;
- L: average number of users in the system;
- L_q: average number of users in the queue;
- W: average time to traverse the system;
- W_q: average waiting time in the queue.

From the probabilities P_n for each n, we can obtain all the others. In particular

$$L = \sum_n P_n n.$$

Relations between the indicators

The performance indicators L, L_q, W and W_q are linked by three main relations.

Little's Law (1961). If λ does not depend on n:

$$L = \lambda W$$

$$L_q = \lambda W_q.$$

If λ_n depends on n, Little's Law still holds, by replacing λ with its mean value

$$\overline{\lambda} = \sum_n P_n \lambda_n.$$

The third relation is fairly obvious: the overall time to traverse the system is given by the time spent in queue plus the service time:

$$W = W_q + 1/\mu.$$

Markovian models

The most widely used probability distribution is the exponential distribution. Its probability density function is:

$$P\{T \leq t\} = 1 - e^{-\alpha t}$$

$$P\{T > t\} = e^{-\alpha t}$$

The expected value is $1/\alpha$. The variance is $1/\alpha^2$.

Properties (1)

Property 1. $f_T(t)$ is decreasing.

$$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\} \quad \forall t > 0$$

The event T occurs more often before the expected value.

Service. It represents actions that are usually fast, occasionally very long (e.g. service at a First-Aid center). It does not represent well repetitive actions (e.g. toll payment at a highway barrier).

Arrivals. It represents independent interarrival times. Consecutive arrivals are usually close to each other, but sometimes there are long periods in between.

Properties (2)

Property 2. Lack of memory.

$$P\{T > t + \Delta t \,|T > \Delta t\} = P\{T > t\} \;\; \forall t > 0$$

The next event does not depend on the last one.

Service. The remaining service time cannot be better estimated from the knowledge of the amount of time already elapsed since its beginning.

Properties (3.1)

Property 3. Relation with Poisson processes. Let $X(t)$ be the number of events occurring in the time interval $[0, t]$.

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!} \quad n = 0, 1, 2, \ldots$$

$X(t)$ has a Poisson distribution with parameter αt. The average is $E\{X(t)\} = \alpha t$. The expected number of events per unit of time is α. The counting of the events is a Poisson process with parameter α.

Properties (3.2)

Service. When service time has an exponential distribution with parameter μ, the average number of service completions between 0 and t is μt. With s active servers, it is $s\mu t$.

Arrivals. If users' arrivals are Poisson events with parameter λ, every interval with a same width has the same probability of containing an arrival.

Properties (4.1)

Property 4. The composition of Poisson processes is a Poisson process. Given n random variables $T_1, T_2, \ldots, T_n$ with exponential distributions with parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$ the random variable $U = \min_{i=1,\ldots,n}\{T_i\}$ has exponential distribution with parameter $\sum_{i=1}^{n} \alpha_i$.

$$P\{U > t\} = e^{-(\sum_{i=1}^{n} \alpha_i)t}$$

Viceversa: if $\lambda_i = p_i \lambda$, every process i is a Poisson process.

Properties (4.2)

Arrivals. If the source population is composed of different types of users, it is possible to study the interarrival times for the whole population from the knowledge of the interarrival times for each user type.

Service. If the servers have different service rates, it is possible to study the service rate of the overall system from the knowledge of the service rates of the servers.

Properties (5)

In small intervals the probability of an event is about $\alpha\Delta t$:

$$P\{T \leq t + \Delta t | T > t\} \approx \alpha\Delta t \quad \text{for small} \Delta t.$$

Arrivals/Service. This gives the probability that an arrival or a completion occur in an interval of duration Δt.

Birth-and-death process

We consider the queuing system as a a dynamic system, whose state is represented by the number of customers in the system, $N(t)$.

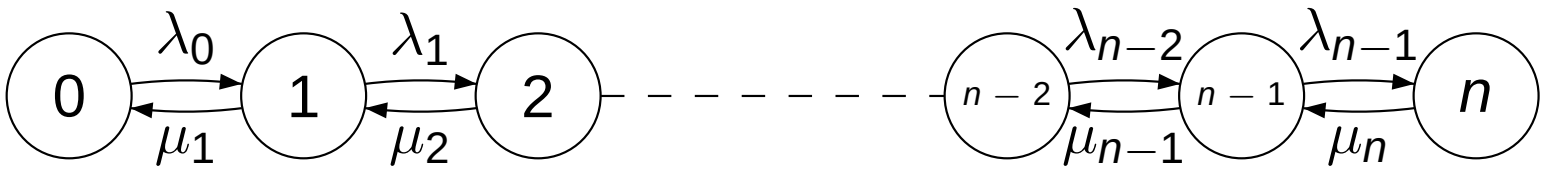
We assume that arrivals and completions are independent.

The interarrival time is assumed to be a random variable with exponential distribution with parameter λ_n.

The completion time for each server is assumed to be a random variable with exponential distribution with parameter μ_n.

It is a special case of continuous-time Markov chain.

Birth-and-death process: analysis



Define:

- $E_n(t)$: number of times the system enters state n up to time t.
- $L_n(t)$: number of times the system leaves state n up to time t.

Balance equation: $\lim_{t\to\infty} \frac{E_n(t)}{t} = \lim_{t\to\infty} \frac{L_n(t)}{t}$

Birth-and-death process: analysis

We obtain a system of linear equations:

$$\begin{cases} \mu_1 P_1 = \lambda_0 P_0 \\ \lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1 \\ \dots \\ \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_m) P_n \end{cases}$$

There is an equation for each state, but one equation is linearly dependent on the others.

The last equation is the normalization equation: $\sum_{n=0}^{\infty} P_n = 1$.

Birth-and-death process: analysis

We define

$$c_n = \frac{\lambda_{n-1}\lambda_{n-2}\dots\lambda_0}{\mu_n\mu_{n-1}\dots\mu_1} \quad n = 1, 2, \dots$$

and we obtain

$$P_n = c_n P_0 \quad \text{from which} \quad P_0 = \frac{1}{\sum_{n=0}^{\infty} c_n}$$

Then:

$$L = \sum_{n=0}^{\infty} nP_n \qquad L_q = \sum_{n=s}^{\infty} (n-s)P_n$$

$$W = \frac{L}{\overline{\lambda}} \qquad W_q = \frac{L_q}{\overline{\lambda}}$$

where $\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$.

Birth-and-death process: analysis

In some cases we can obtain the sums analytically; otherwise they must be approximated numerically.

These results hold when a steady state exists:

- $\lambda_{\overline{n}} = 0$ for some $\overline{n}$ or
- $\rho = \frac{\lambda}{s\mu} < 1$.

$M/M/1$

M stands for exponential distribution (of inter-arrival time and service time).

Assuming $\rho < 1$:

$$C_n = \frac{\lambda^n}{\mu^n} = \rho^n \quad n = 0, 1, \ldots$$

$$P_0 = 1 - \rho$$

$$P_n = \rho^n(1 - \rho)$$

$$L = \frac{\rho}{1-\rho}$$

$$L_q = L - (1 - P_0) = \frac{\rho^2}{1-\rho}$$

$$W = \frac{1}{\mu-\lambda}$$

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)}$$

$M/M/s$

We assume $\rho = \frac{\lambda}{s\mu} < 1$.
When $n \leq s$, then $\mu_n = n\mu$.
When $n \geq s$, then $\mu_n = s\mu$.

$$c_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} & n \leq s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} & n \geq s \end{cases}$$

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\frac{\lambda}{s\mu}}}$$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & n \leq s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0 & n \geq s \end{cases}$$

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}$$

$$L = L_q + \frac{\lambda}{\mu}$$

$$W_q = \frac{L_q}{\lambda}$$

$$W = W_q + 1/\mu$$

$M/M/1/K$ with finite capacity

The queue length is limited to K (e.g. a capacitated buffer). In this case the system reaches a steady-state condition even if $\lambda > \mu$.

$$\lambda_n = \begin{cases} \lambda & n < K \\ 0 & n \geq K \end{cases}$$

$$c_n = \begin{cases} \rho^n & n < K \\ 0 & n \geq K \end{cases}$$

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

$$P_n = \begin{cases} \frac{1}{K+1} & n \leq K \\ \frac{1-\rho}{1-\rho^{K+1}}\rho^n & n \leq K \end{cases} \quad \begin{array}{l} \text{if } \rho = 1 \\ \text{otherwise} \end{array}$$

$$L = \begin{cases} \frac{K}{2} & \text{if } \rho = 1 \\ \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & \text{otherwise} \end{cases}$$

$$L_q = L - (1 - P_0)$$

$$W = L/\overline{\lambda}$$

$$W_q = L_q/\overline{\lambda}$$

where $\overline{\lambda} = \lambda(1 - P_K)$.

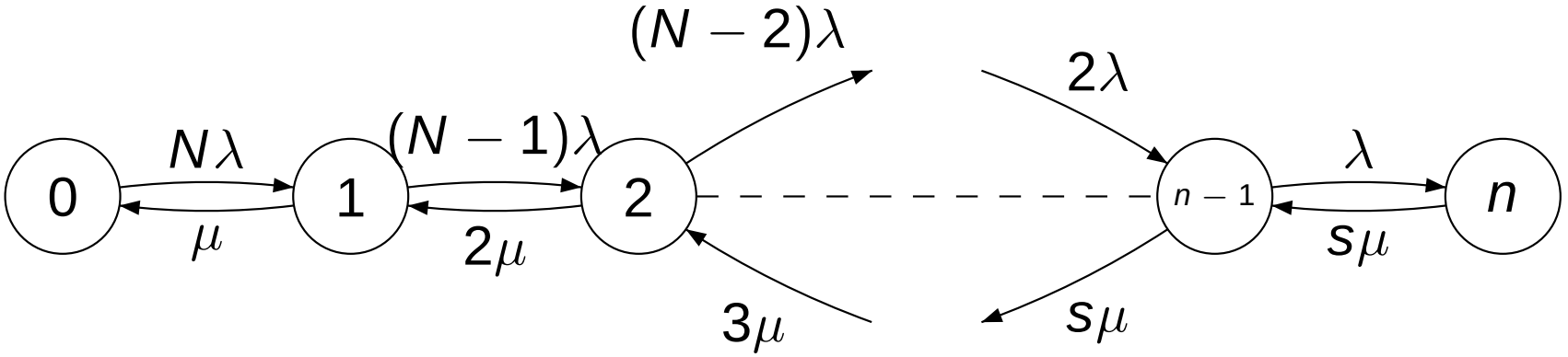
$M/M/s$ with finite population

We assume there exist only N potential customers: each of them can be in the system or not (e.g. machine maintenance).

We assume the time spent by each customer out of the system is a random variable with exponential distribution described by a parameter λ.

Therefore the next arrival time is the minimum among the $N - n$ arrival times of the customers out of the system.
Hence it is a random variable with exponential distribution with parameter $\lambda_n = (N - n)\lambda$.

$M/M/s$ with finite population



$$\lambda_n = \begin{cases} (N-n)\lambda & n \leq N \\ 0 & n > N \end{cases}$$

$$\mu_n = \begin{cases} n\mu & n < s \\ s\mu & n \geq s \end{cases}$$

The formulae hold for any probability distribution of the out-of-the-system time with expected value $1/\lambda$.

$M/M/1$ with state-dependent frequencies

Effect 1. The longer is the queue, the fewer new customers arrive.

$$\lambda_n = (n+1)^{-a}\lambda_0 \quad \text{with } a > 0$$

$$c_n = \frac{(\lambda_0/\mu)^n}{(n!)^a} \quad n = 0, 1, 2, \ldots$$

Effect 2. The longer is the queue, the faster is the service.

$$\mu_n = n^b \mu_1 \quad \text{with } b > 0$$

$$c_n = \frac{(\lambda/\mu_1)^n}{(n!)^b} \quad n = 0, 1, 2, \ldots$$

Both effects.

$$c_n = \frac{(\lambda_0/\mu_1)^n}{(n!)^{a+b}} \quad n = 0, 1, 2, \ldots$$

$M/M/s$ with state-dependent frequencies

We assume the frequencies depend on the number of customers in queue for each server, i.e. n/s.

Effect 1.

$$\lambda_n = \begin{cases} \lambda_0 & n \leq s-1 \\ (\frac{s}{n+1})^a \lambda_0 & n \geq s-1 \end{cases}$$

Effect 2.

$$\mu_n = \begin{cases} n\mu_1 & n \leq s \\ (\frac{n}{s})^b s\mu_1 & n \geq s \end{cases}$$

$M/E_k/1$

The Erlang distribution depends on an additional integer parameter k, which is called *shape parameter*.

The density function is:

$$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1} e^{-k\mu t} \quad t \geq 0$$

The shape parameter affects the variance $\frac{1}{k\mu^2}$, not the exp. value $\frac{1}{\mu}$.

The Erlang distribution describes the sum of k independent random variables with the same distribution with expected value $\frac{1}{k\mu}$.

Therefore it is used to represent service times when the service consists of several (identical) operations in a sequence.

$M/E_k/1$

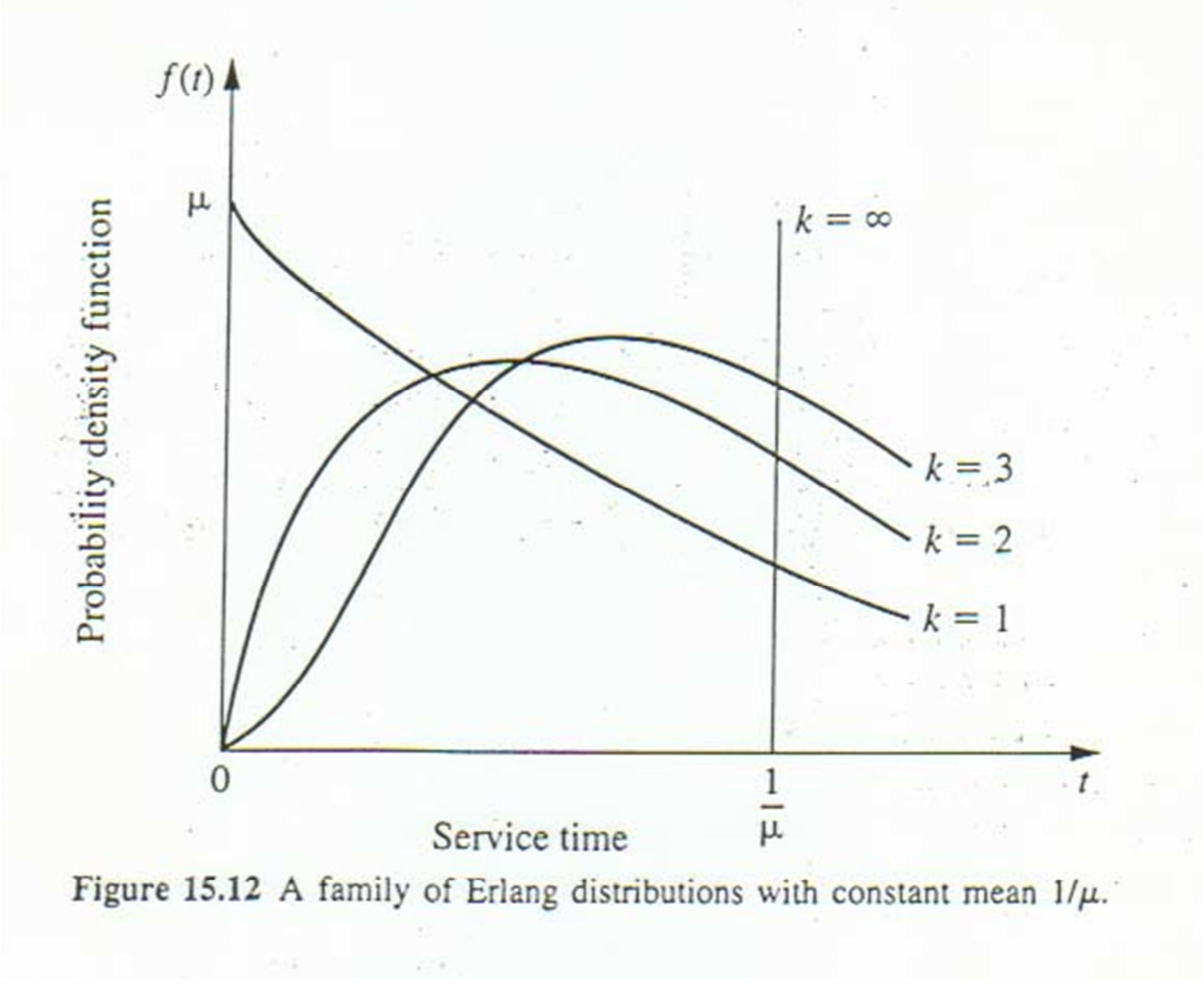


Figure 15.12 A family of Erlang distributions with constant mean $1/\mu$.

$M/E_k/1$

For $k = 1$, E_k reduces to an exponential distribution M.
For $k \to \infty$, E_k tends to a degenerate distribution D.

$$L_q = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu-\lambda)}$$

$$W = W_q + 1/\mu$$

$$W_q = \frac{1+k}{2k} \frac{\lambda}{\mu(\mu-\lambda)}$$

$$L = \lambda W$$

$M/G/1$

Assumption: service times are independent and they have the same probability distribution with expected value $1/\mu$ and variance σ^2.
Condition for convergence is $\lambda/\mu < 1$.

$$P_0 = 1 - \rho$$

Pollaczek − Khintchine formula :

$$L_q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)}$$

$$L = \rho + L_q$$

$$W_q = L_q/\lambda$$

$$W = W_q + \mu$$

All performance parameters increase with σ^2.
If G is exponential, then $\sigma^2 = 1/\mu^2$ and we obtain $M/M/1$.
For $s > 1$ few results are known.

$M/D/s$

In this case we have constant service time: $\sigma^2 = 0$.

For $s = 1$ we have $L_q = \frac{\rho^2}{2(1-\rho)}$, that is half the value of L_q in an $M/M/1$ system.

For $s > 1$ we must resort to tabulated values, as for many other cases, such as $G/M/s$, $D/M/s$, $E_k/M/s$, $E_m/E_k/s$, $E_k/D/s$, $D/E_k/s$.

Queuing systems with priorities

We assume to have k queues, each with FIFO discipline.

As soon as a server becomes idle, it starts serving the first customer from the highest priority non-empty queue.

Each queue is assumed to correspond to an input Poisson process, whose parameter λ_i can be different for each queue $i = 1, \ldots, k$.

Service times are assumed to have an exponential distribution with the same parameter μ for all queues.

Each queue h converges to a steady-state condition if $\sum_{i=1}^{h} \lambda_i < s\mu$.

Queuing systems with priorities

We can consider models with and without pre-emption.

Among those with pre-emption, we can distinguish between preemption-resume and pre-emption-restart.

For the overall system, the same results for L, L_q, W and W_q still hold as for the $M/M/s$ system.

We have different variance for the waiting times.

Usually one wants to study L_i, L_{qi}, W_i and W_{qi} for each queue i.

Designing a queueing system

To properly design a queuing system for a given population of customers, one has to decide:

- the number s of servers;
- their efficiency μ;
- the number of service centers λ.

The cost for the service provider increases with the level of service.
The cost for the customers decreases with the level of service.
The definition of the optimal trade-off between the two cost terms depends on the relation between the provider and the customers, the type of service and the cost function.

Designing a queueing system

Case 1. The waiting cost is a function $g(n)$ of the number n of customers in the system:

$$E[WaitingCost] = \sum_{n=0}^{\infty} g(n) P_n$$

Case 2. The waiting cost is a function $h(w)\lambda$ of the waiting time w:

$$E[WaitingCost] = \lambda \int_0^{\infty} h(w) f_w(w) dw$$