Chapter 4
Impact of Variability of Interarrival and Service Times

4.1 Importance of Distributions: A Motivating Example

In this section we highlight the *significant errors* in the computation of performance indexes that are introduced when *only* the mean values are considered instead of the distributions of some input parameters. Consider a server that requires a *constant* `Service time` S of 1 s to execute a request. Assume that the requests arrive with rate $\lambda = 60$ req/min in groups (*bursts*) and that the requests of a burst arrive at the same instant of time. The time between consecutive bursts is *constant*. We will analyze the impact on `Queue time` and `Response time` of different burst lengths, ranging from 1 to 60, considering always the *same* arrival rate.

In the case shown in Fig. 4.1a, a request arrives at the server exactly every *second*. Since the time S required for its execution is always equal to 1 s, the queue will *never* take place (the `Queue time` is equal to zero) and thus the mean `Response time` (`Queue time` plus `Service time`) is exactly one *second* for all *requests*. In the other graphs it is assumed that the requests arrive at the server with burst of increasing dimensions.

In Fig. 4.1b a burst of size 2 arrives exactly every two *seconds*. The first request never waits in queue, while the latter waits for one *second*, that is, the execution time of the first. So the mean `Queue time` is 0.5 s. In the graph of Fig. 4.1c a burst of size 3 arrives exactly every three *seconds*. The first request never waits, the second waits a *second* and the third waits two *seconds*. So the mean `Queue time` is 1 s and the mean `Response time` is 2 s.

Finally, in Fig. 4.1d 60 *requests* arrive together in a single burst every sixty *seconds* (the rate is always 1 req/s). In this case the mean `Queue time` is 29.5 s. Let us remind that the sum of n positive consecutive integers starting from 1 is $n(n+1)/2$. In our case we have 60 requests, but *only* $n = 59$ of them wait from 1 to 59 s, respectively. Thus, the *mean* waiting time (`Queue time`) of the 60 requests is 29.5 s and the mean `Response time` is 30.5 s! The conclusion is

G. Serazzi, *Performance Engineering*,
https://doi.org/10.1007/978-3-031-36763-2_4

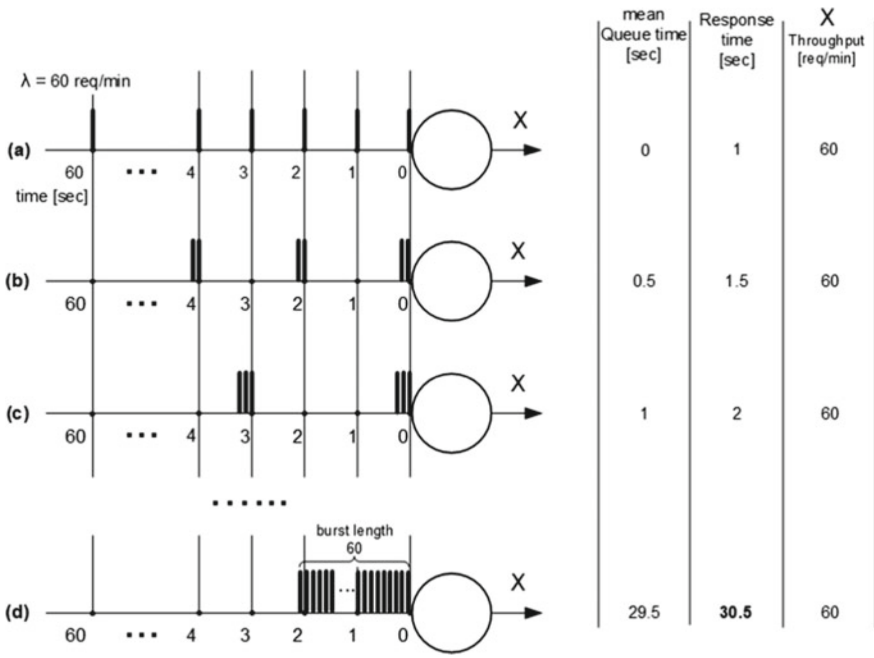


Fig. 4.1 Impact of different burst lengths on mean `Response time`

> *Even considering the* **same** *arrival rate* $\lambda = 1$ req/s *and the* **same** `Service times` *S=1* s, *depending on the arrival pattern of requests we could have a* **very high variability** *of mean* `Response times`: *from 1 to 30.5* s *in the example considered (and this is not the worst case!).*

4.2 Variability of Interarrival Times

tags: open, single class, Queue, Exp/Hypo-exp/Hyper-exp, JSIMg.

The objective of this case study is to emphasize the impact of the variance of `Interarrival times` on the performance of a system.

4.2.1 Problem Description

Consider a model of a web server that needs to execute an e-commerce application to sell equipment produced by a new company. While the mean and variance of `Service time` required to process a purchase order can be estimated with sufficient accuracy, the pattern of incoming requests is unknown as customers are located all over the world.

We use a simple model of the web server consisting of a single queue station. To account for the unknown patterns of the incoming requests we consider *five* distributions of interarrival times with the *same mean* and increasing variability. To describe the variance of interarrival times we use the *coefficient of variation* `c` of each distribution (given by the *standard deviation/mean* ratio). For a given value of arrival rate, the values of `c` are directly proportional to the variance of the five distributions since their means are the same.

The `Service times` are assumed *exponentially* distributed with the same mean $S = 1$ s for *all* the models.

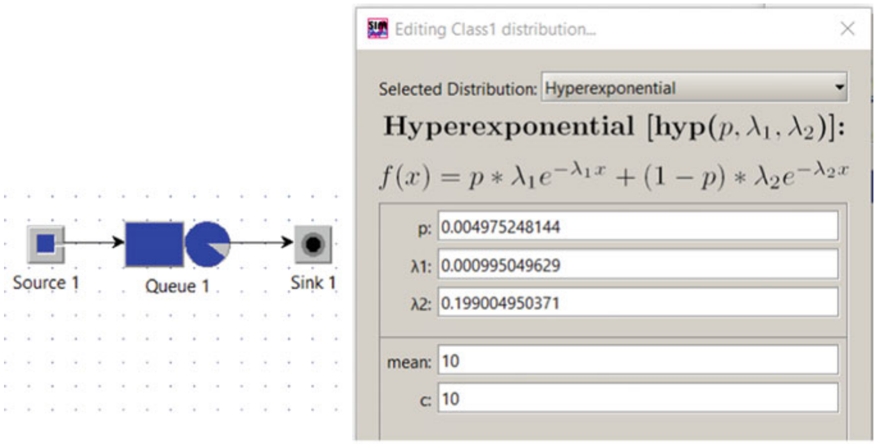
To analyze a wide range of traffic intensities we consider several arrival rates, ranging from light-load (10% of server utilization) to heavy-load (90% of server utilization) conditions. For *each* arrival rate we execute *five* models corresponding to the *five* interarrival time distributions. As a reference metric we consider the mean `Response times` of the models executed. The models are solved with JSIMg.

4.2.2 Model Implementation

We use a open model consisting of three stations: `Source1`, `Queue1`, and `Sink1`, Fig. 4.2a. The `Service times` of `Queue1`, with mean $S = 1$ s, used in all the models have the same *exponential* distribution. The five distributions considered of `Interarrival times`, in sequence of increasing variance are: *Constant* cv = 0, *Hypo-exponential* cv = 0.5 (`Hypo-exp`), *Exponential* cv = 1 (`Exp`), *Hyper-exponential* cv = 5 (`Hyper-exp`), *Hyper-exponential* cv = 10 (`Hyper-exp`). Figure 4.2b shows the window for setting the `mean = 10` (corresponding to $\lambda = 0.1$ req/s) and the `coefficient of variation cv = 10` of the *Hyper-exp* distribution.

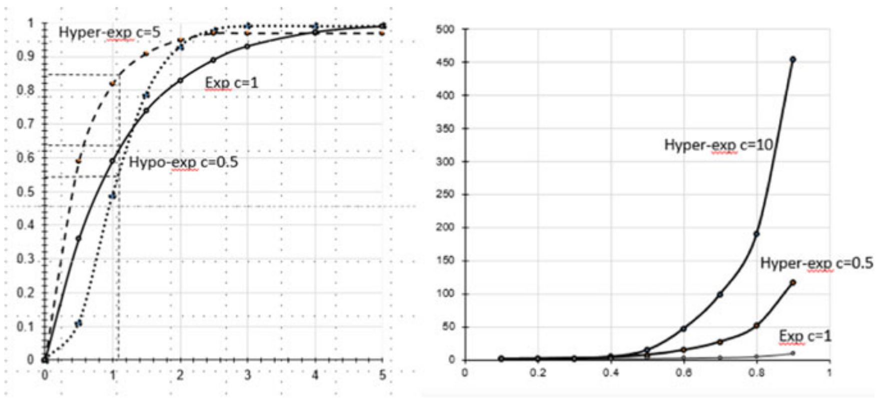
The differences between the distributions are emphasized in Fig. 4.3a (obtained with $\lambda = 0.9$ req/s), that shows the graphs relating to three of them: *Hypo-exp* cv = 0.5, *Exp* cv = 1, and *Hyper-exp* cv = 0.5. As can be seen, the percentages of `Interarrival times` (i.e., the *percentiles*) that are *less than* the mean value 1.111 s are very different: 56.8% for the *Hypo-exp*, 63.6% for the *Exp* (the exact analytical result is 0.6321), and 85% for the *Hyper-exp* with cv = 5 (and 91% for the *Hyper-exp* cv = 10, not shown in the figure). To obtain the percentiles of a metric with JSIMg see Sect. 2.2 and Figs. 2.10, 2.11.

The increase of variability also heavily influences the *maximum* values of the various distributions: 5.4 s for the *Hypo-exp*, 16.69 s for the *Exp*, 261.22 s for the *Hyper-exp* cv = 5, and 864.46 s for the *Hyper-exp* cv = 10. The number of samples needed to reach the equilibrium of the metric `Throughput` of `Source1`, that provides the `Interarrival times` with 99% Confidence Interval and 0.03 Max Rel. Err., ranges from 40960 of the *Hypo-exp* to 1063920 of the *Hyper-exp* cv = 10.



(a) Layout of the model (b) Parameters of the Hyper-exp distribution

Fig. 4.2 The model considered (**a**), Settings of the `mean = 10` and `coeff. of variation cv = 10` of the *Hyper-exponential* distribution of `Interarrival times` for `Arr.rate` 0.1 req/s (**b**)



(a) Three `Interarrival time` distr. (b) Response times [*sec*] *vs* Arr. rate

Fig. 4.3 `Interarrival time` distributions with *increasing variability* (cv = 0.5, 1, 5) obtained with $\lambda = 0.9$ req/s and the same mean 1.111 s (**a**); the corresponding `Response times` of `Queue1` for $\lambda = 0.1 \div 0.9$ req/s (**b**)

Table 4.1 `Response times` [s] with five `Interarrival time` distributions with increasing variance *vs* `Arrival rates`. `Service times` S = 1 s are *exponentially* distributed

`Arrival rate`	`Response times`				
	`Interarrival time` distributions				
	Const cv = 0	Hypo-exp cv = 0.5	Exp cv = 1	Hyper-exp cv = 5	Hyper-exp cv = 10
$\lambda = 0.1$ [req/s]	1.00	1.01	1.11	1.22	1.24
$\lambda = 0.3$ [req/s]	1.05	1.12	1.43	2.20	2.40
$\lambda = 0.6$ [req/s]	1.47	1.70	2.54	14.49	46.98
$\lambda = 0.9$ [req/s]	5.13	6.43	9.92	116.88	455.06

4.2.3 Results

To simulate the different traffic intensities we use, for each distribution, a `What-if` analysis, with *Arrival rate* as control parameter, that execute nine models with λ ranging from 0.1 (light load) to 0.9 (heavy load) req/s with increments of 0.1. Figure 4.3b shows how the `Response time` R varies with different arrival patterns and rates. To make it easier to understand the figure, only R obtained with three distributions are plotted: *Exp* cv = 1, and *Hyper-exp* with cv = 5 and cv = 10. As can be seen, the values of R grow very fast not only when the `Arrival rate` is approaching the saturation value $\lambda^{sat} = 1$ req/s (and expected) but also with the increase of the variability of the `Interarrival times` (and this is not so expected).

Table 4.1 shows the `Response Times` for the five distributions with `Arrival rates` $\lambda = 0.1, 0.3, 0.6, 0.9$ req/s.

Even if we do not consider the two extreme distributions (i.e., the Constant cv = 0 and the Hyper-exp cv = 10), the differences between the `Response times` corresponding to the same λ become greater as the utilization of the server increases. The values of the last row of the table, corresponding to the utilization of 90%, show a difference of *more than 18 times* between 6.43 s with Hypo-exp cv = 0.5 and 116.88 s with Hyper-exp cv = 5!

Since for a given arrival rate λ the server *utilization U* is the *same* for *all* distributions (it is $U = \lambda\, S$), we may conclude that:

> *measuring server* `Utilization` *is* **useless** *to predict* `Response times` *if it is not complemented with the knowledge of other metrics, such as the distributions of* `Interarrival` *and* `Service times`.

4.3 Variability of Service Times

tags: open, single class, Queue, Exp/Hypo-Exp/Hyper-Exp, JSIMg.

This case study has been purposely designed to highlight the impact of the variance of `Service times` on the performance of a system. The `Service times` follow five different distributions, while the `Interarrival times` are generated according to the same *Exponential* distribution. It can be considered the dual of the example discussed in the preceding section in which the opposite situation was evaluated.

4.3.1 Problem Description

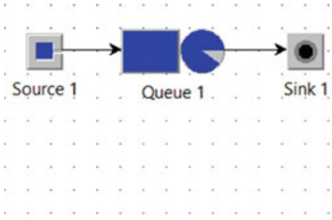
The scenario of this example is quite common in many practical problems in which the execution times of the applications are often *highly variable* based on input data and required functions (see, e.g., [21]).

We consider an application for the computation of the path between two geographical locations. The algorithms that compute the driving route from a source to a destination are computationally heavy and the `Service demands` are *highly variable* as a function of the locations considered. For these reasons the management decided to deploy the application on a dedicated server and to evaluate the impact on `Response time` of the different locations.

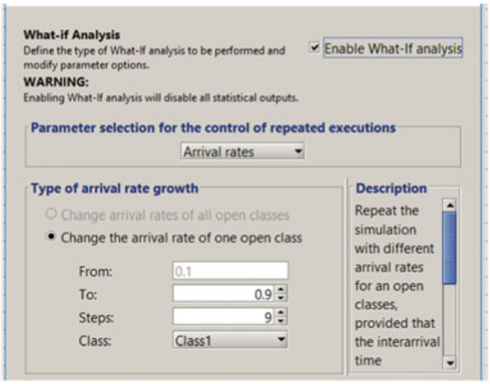
The `Interarrival times` of the route requests are assumed *Exponentially* distributed and different `Arrival rates`, that cover the range from light to heavy traffic, are considered. To account for the different fluctuations in execution times, *five* distributions with *increasing* variances, from zero to very high values, and the same mean were considered. For each `Arrival rate` we evaluate the `Response time` for the five `Service times` distributions. The models are solved with JSIMg.

4.3.2 Model Implementation

The layout of the open model used is shown in Fig. 4.4a. It consists of three stations: `Source1`, `Queue1`, and `Sink1`. The five distributions of the `Service times` considered, in sequence of increasing variance, are: *Constant* $\mathrm{cv} = 0$ (`Const`), *Hypo-exponential* $\mathrm{cv} = 0.5$ (`Hypo-exp`), *Exponential* $\mathrm{cv} = 1$ (`Exp`), *Hyper-exponential* $\mathrm{cv} = 5$ (`Hyper-exp`), *Hyper-exponential* $\mathrm{cv} = 10$ (`Hyper-exp`). The use the *coefficient of variation* `cv` of each distribution (given by the *standard deviation/mean* ratio) to describe the variance of `Service times` is convenient in this case as,



(a) Layout of the model



(b) Parameters of the `What-if`

Fig. 4.4 Model considered (**a**); `What-if` with `Arrival rates` $\lambda = 0.1 \div 0.9$ req/s (**b**)

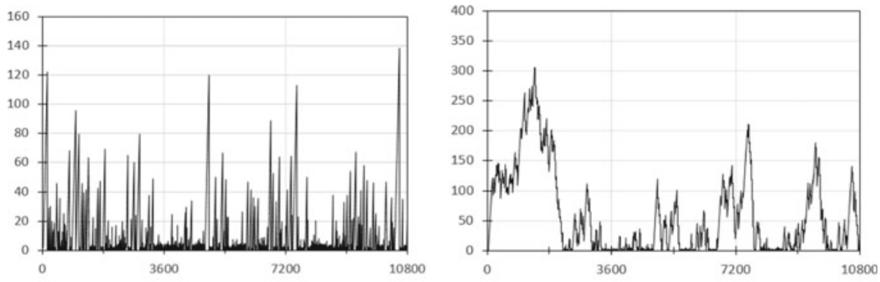
for a given `Arrival rate`, its values are directly proportional to the variance of the five distributions being their means the same (S = 1 s).

The same *Exponential* distribution of the `Interarrival times` generated by `Source1` is used in all the models. A `What-if` analysis is used to execute, for each distribution of `Service times`, 9 models with `Arrival rates` ranging from 0.1 to 0.9 req/s with increments of 0.1 (see Fig. 4.4b). Globally, five `What-if` analyses are required corresponding to the five distributions of `Service times` considered (in total 45 models are executed).

4.3.3 Results

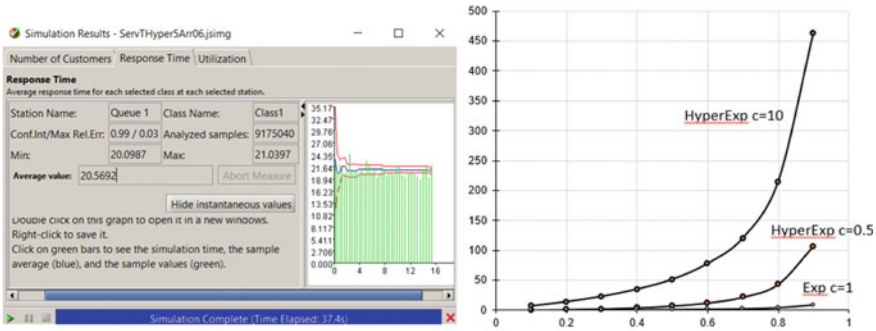
The objective of the two graphs of Fig. 4.5 is to provide a visual evidence of the negative effects of service time *fluctuations* on `Response times`. In Fig. 4.5a the `Service times` of a period of three hours (simulated time) with a `Hyper-exp cv = 5` distribution are shown. Remember that the mean is $S = 1$ s for all distributions! The `Response times`, with $\lambda = 0.9$ req/s, for the same period are shown in Fig. 4.5. The data for the plots of Fig. 4.5 are obtained from the CSV files generated by JSIMg.

The correlation between the bursts of high values of S and the peaks of `Response times` is evident and consistent with intuition. The *bursts* create a congestion of the server and small increases in arriving requests in this condition determine enormous increases in queue length, and in `Response times` together with it. For example, consider the initial period of half-hour, or the period of about 800 s centered at the end of two hours (7200 s), or the period starting at about 9000 s. It must be pointed out that the fluctuations of `Response times` are emphasized in our case due to the high `Utilization` of the server $U = \lambda S = 0.9$.



(a) Fluctuations of S *Hyper-exp* cv=5 (b) The corresponding `Response times` [*sec*]

Fig. 4.5 `Service times` generated with *Hyper-exp* distribution (S = 1 s and cv = 5) for a period of three hours (**a**); corresponding `Response times` with $\lambda = 0.9$ req/s (**b**)



(a) R with $\lambda = 0.6\ r/s$, cv=5, S=1 *sec* (b) R *vs* `Arrival rates` [*req/sec*]

Fig. 4.6 `Response Time` with *Hyper-exp* cv = 5 distrib. of S (**a**); R with three different `Service times` distributions and same mean 1 s, `Interarrival times` are *Exponentially* distributed

Figure 4.6a shows an example of the results provided by one of the 45 models executed: the behavior of the `Response times` obtained from a simulation run with $\lambda = 0.6$ req/s and *Hyper-exp* distribution of `Service times` with cv = 5. The mean value R = 20.56 s with the precision required (99% of conf. interval, 0.03 max error) is obtained with 9175040 samples.

The `Response times` obtained with three different distributions of `Service times` are shown in Fig. 4.6b. The `arrival rate` range from 0.1 to 0.9 req/s with step of 0.1. The variance of the three distributions increases from the *Exponential* (cv = 1) to the *Hyper-exp* (cv = 10).

The `Response times` obtained by JSIMg simulating five distributions of `Service times` and $\lambda = 0.1, 0.3, 0.6, 0.9$ req/s are given in Table 4.2. As can be seen, for the same `Arrival rate` there are huge differences between the values obtained with the five distributions. These differences increase as server `Utilization` increases. Even avoiding to consider the `Constant cv=0`

Table 4.2 `Response times` with five `Service times` distributions with increasing variance and same mean S = 1 s *vs* `Arrival rates`. `Interarrival times` are *Exponentially* distributed

`Arrival rate`	`Response time`				
	`Service time` distributions				
Exp cv = 1	Const cv = 0	Hypo-exp cv = 0.5	Exp cv = 1	Hyper-exp cv = 5	Hyper-exp cv = 10
$\lambda = 0.1$ [r/s]	1.05	1.06	1.11	2.42	6.67
$\lambda = 0.3$ [r/s]	1.21	1.26	1.43	6.66	22.62
$\lambda = 0.6$ [r/s]	1.76	1.95	2.54	20.56	77.15
$\lambda = 0.9$ [r/s]	5.53	6.51	9.92	119.17	453.36
$\boldsymbol{\lambda}$ **= 0.9** M/G/1	**5.5**	**6.625**	**10**	**118**	**455.5**

distribution, which provides a lower bound for all distributions, we can have enormous differences (up to **70 times** with $\lambda = 0.9$ req/s) between the `Response times` obtained with *Hypo-exp* cv = 0.5 (6.51 s) and those with *Hyper-exp* cv = 10 (453.36 s)! Let us remark that these differences occur even if the `Utilization` of the server is the same for all distributions.

Thus, we can conclude that:

> *to provide* **accurate** *performance forecast of a server it is essential to know the distributions of* `Interarrival` *and* `Service times`, *and not just their mean values and server* `Utilization`.

The model considered in this section could be solved analytically obtaining exact results. In fact it corresponds to a M/G/1 queue station (see the tutorial [32] and, e.g., [36]) having *Exponential* `Interarrival times`, i.e., the arrival process is Poisson (Markovian, M), `Service times` with *general* distribution (G) with given `mean` and `variance`, and a single server. The `Response time` of this station is given by:

$$R_{Queue1} = \textit{waiting time in queue } W + \textit{Service time } S = \frac{US(1+cv^2)}{2(1-U)} + S \quad (4.1)$$

where U is the server `Utilization` (U $= \lambda S$), and cv is the *coefficient of variation* of the *general* distribution of `Service times` with mean S. Note that both the mean and the variance of `Service times` *must be known* to compute the coefficient of variation. In the last row of Table 4.2 are reported the exact `Response times` computed with Eq. 4.1. As can be seen, the values obtained with JSIMg are very close to the exact ones, and are all within the 99% confidence intervals.

Let us remark that when the `Service times` are *Constant* it is cv = 0 and the model is identified as M/D/1 (D stands for *Deterministic* `Service times`). Its waiting time W (computed with Eq. 4.1) is half of that obtained with an *Exponential* distribution (in a M/M/1 model with cv = 1). For example, as shown in the last row of Table 4.2 with *Constant* distribution it is W = 4.5 s while with `Exponential` it is W = 9 s (with $\lambda = 0.9$ req/s and S = 1 s). The waiting time W of an M/D/1 station is the *lower bound* for any M/G/1 station with the same S and `Arr. rate`.