Article

Time to Critical Condition in Emergency Services

Pedro A. Pury [ID]

Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba, Ciudad Universitaria, Córdoba X5000HUA, Argentina; pedro.pury@unc.edu.ar

Abstract: Providing uninterrupted response service is of paramount importance for emergency medical services, regardless of the operating scenario. Thus, reliable estimates of the time to the critical condition, under which there will be no available servers to respond to the next incoming call, become very useful measures of the system's performance. In this contribution, we develop a key performance indicator by providing an explicit formula for the average time to the shortage condition. Our analytical expression for this average time is a function of the number of parallel servers and the inter-arrival and service times. We assume exponential distributions of times in our analytical expression, but for evaluating the mean first-passage time to the critical condition under more realistic scenarios, we validate our result through exhaustive simulations with lognormal service time distributions. For this task, we have implemented a simulator in *R*. Our results indicate that our analytical formula is an acceptable approximation under any situation of practical interest.

Keywords: first-passage time; Markov chain; queueing theory; simulation; OR in health services; KPI

MSC: [2010] 90B22; 60K25

1. Introduction

The problem of assigning resources to respond to a stochastic demand is a ubiquitous topic in operational research. The trade-off between service quality and operational efficiency is a crucial aspect of the Emergency Medical Services (EMS), where the lives of patients depend on the timeliness of care. Thus, the development of Key Performance Indicators (KPIs) to objectively quantify the performance across the operational, clinical and financial departments is a current demand of an industry that is becoming increasingly data-driven. KPIs are the basic tools for planners, and the nature of each KPI selects a particular feature of the system and determines its data gathering strategy.

Among the most intuitive and used operational KPIs in EMS are the successive times involved in the service cycle: call reception, patient triage, dispatch, ambulance turnout, travel from the base to the emergency site, paramedic care, eventual transfer of the patient to a hospital and return of the ambulance to its base. Response time (the interval between the reception of an emergency call and the arrival of a paramedic at the scene of the event) is a common operational metric of EMS, and it is considered a good indication of the quality offered by the service [1]. One reason for its popularity as a KPI resides in the fact that it is directly quantifiable and easily understood by the public and policy makers. Additionally, the EMS industry has the goal of providing care within eight minutes for cardiac arrest [2] and major trauma [3]. However, there is evidence that exceeding that response time criterion does not affect patient survival after a traumatic impact injury [4,5]. Moreover, solutions that only focus on shortening the response time are cost prohibitive and put the safety of patients, attendant crew and the public at risk [6]. A rational approach to the ambulance business process should simultaneously consider multiple metrics and operational trade-off between administrator-oriented and patient-centered KPIs [7].

One of the most important aspects of emergency medical management is avoiding the oversaturation of the system. Therefore, in this work, we consider the First-Passage Time (FPT) [8] to the critical condition, under which there will be not available servers to respond to the next incoming call. Criticality prediction is of special interest for the quality of the medical service (response time off-target) as well as for the financial management of the service, given that, as we will see, the critical condition strongly depends on the number L of ambulances simultaneously in service and because queueing a call may involve transferring it to another EMS. Any system operating under fixed conditions with a given number of servers and a First-Come First-Served (FCFS) discipline is a discrete one-dimensional stochastic process over the occupation states of servers. Therefore, the first-passage time to the state of oversaturation, in which there are no available servers, is finite. The question is how long is that time. Thus, the mean first-passage time (MFPT) becomes a relevant key performance indicator for the operational condition in EMS.

In urban emergency services logistics, there are two distinct fields: capacity planning and location analysis. Both fields are interrelated in the districting problem or how the region should be partitioned into areas of primary responsibility (districts) [9,10]. Here, MFPT can be applied to an operational subarea or district, preferably intended to be independently served by a subset of ambulances (intradistrict dispatches). In this case, MFPT gives us the average time to request an ambulance from another operational zone to answer the next emergency call when the primary equipment is busy (interdistrict dispatches). MFPT can also be a useful KPI in the decision-making process of emergency departments [11,12], where MFPT provides the average time to a shortage of intensive therapy beds when the FCFS discipline is used after the patient's triage.

Queueing theory has been widely applied in health care in the last 70 years [13]. Since Larson's seminal article [9], quite a few queueing models have been developed to incorporate the intrinsic probabilistic nature of urban EMS derived from the Poisson nature of the call arrival process and the variability in service times. Multiple queueing systems have been developed that respond with different emphasis on the KPIs selected in each case. In this contribution, we use a birth–death process to properly analyse the dependence of MFPT on the number of servers and the rest of the system's operational parameters. The birth–death process is basic to queueing models involving exponential inter-arrival and service time distributions. In the Kendall–Lee notation, we have $M/M/L/FCFS/\infty/\infty$ [8]. Thus, our analytical model is based only on two average times: T_C, the mean inter-arrival time, and T_S, the mean service time of a single server, that is, the time it takes for an ambulance to complete a trip from the instant a call is assigned until the release of this server. Several analytical results are well known in operational research under that assumption [8].

However, experimental evidence from an emergency service indicates that service time distributions are well fitted by lognormal distributions [12,14–16]. Hence, our objective is also to numerically evaluate the deviation between analytical and simulated results for MFPT. Thus, we present an *R*-simulator for a system of L servers in parallel with general distributions for inter-arrival and service times and FCFS discipline: $GI/G/L/FCFS$ [8]. This tool allows the user to calculate the key performance indicators of direct interest in the industry beyond the known analytical results, which are limited to exponential distributions. Particularly, we show our work on Mean First-Passage Time (MFPT) to system critical condition.

In this way, the motivation of our work is two-fold. On the one hand, we provide an explicit analytical expression for the MFPT to the critical condition, and, on the other hand, under more realistic conditions, we analyse the validity of our assumptions through exhaustive simulations. In Section 2, we provide our analytical expression for MFPT and explore the generic nature of the method, postponing detailed mathematical derivations to the appendices. Additionally, in that section, we describe the simulation framework for experimentation. Section 3 deals with the numerical results, and last, in Section 4, we discuss the importance of our contribution.

2. Model and Simulator

In this section, we develop an analytical closed-form solution for the MFPT and present a simulation framework based on discrete events.

2.1. Markov Chain Model for Servers in Parallel

We consider a stochastic continuous-time birth–death process [8] that describes the time evolution of the occupation state of a set of L servers in parallel. Changes in the state of the Markov chain imply the release of a server or putting one into action (if at least one is available). The state with occupation n corresponds to n received calls not completely served yet. Thus, when $n = 0$, all the servers are free, and there are neither trips in process nor calls in queue. For $0 < n \leq L$, there are no waiting calls, and n servers are in course of action. In an equivalent way, we can say that n calls are simultaneously being served. Particularly, when $n = L$, the system is saturated, that is, all servers are occupied. Even though there is not any call in the waiting queue, all servers have been assigned to calls, and consequently, there are not any servers available to process the next eventual incoming call. For $n > L$, the system is oversaturated, and there are $n - L$ calls in the waiting queue. At any time, the system can change its state of occupation between its nearest neighbours. Therefore, we denote the transition rate from the state n to $n + 1$ by ω_n^+, whereas the transition rate toward the lower occupation state ($n \rightarrow n - 1$) is ω_n^-. We assume that the time between calls is an exponential random variable with the mean number of calls per unit time $\lambda = 1/T_C$. On the other hand, the service time is also an exponential variable with the rate or mean number of services per unit time and per server $\mu = 1/T_S$. Thus, random call arrival times and service times consumed in each trip are generated from continuous time distributions. T_C and T_S are the average inter-arrival and service time, respectively. In our particular problem with L servers, the transition probabilities rates of the birth–death process are defined by [17]

$$
\begin{aligned}
\omega_n^+ &= \lambda \quad \forall n, \\
\omega_n^- &= \begin{cases} n\,\mu & \text{for } n \leq L, \\ L\,\mu & \text{for } n \geq L. \end{cases}
\end{aligned}
\tag{1}
$$

In this manner, ω_n^+ results constant and only ω_n^- depends on the state of the system. Then, all the experimental information needed to characterize our theoretical model are the average times T_C and T_S.

For fixed values of T_C and T_S, the Markov process always reaches the state $L + 1$, that is, the critical condition in which the incoming call could not be served. Therefore, we focus our interest in the first-passage time (FPT) to the state $L + 1$. That is the time needed by the system to reach the critical situation (first call is derived to the waiting queue) given an initial state without queue ($0 \leq n \leq L$). Following previous experience [18], the MFPT, $T(n)$, from the initial state $n = 0, \ldots, L$, can be written as

$$
\begin{aligned}
T(0) &= T_C \left(L + 1 + \sum_{k=0}^{L-1} \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{L} i!\,\gamma^i \right), \\
T(1) &= T(0) - T_C, \\
T(n) &= T(0) - T_C \left(n + \sum_{k=0}^{n-2} \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{n-1} i!\,\gamma^i \right) \quad \text{for } 2 \leq n \leq L,
\end{aligned}
\tag{2}
$$

where the parameter $\gamma = \mu/\lambda = T_C/T_S$ is the inverse of Erlang's rate [8]. The derivation and mathematical details of Equation (2) are worked out in Ref. [18] and Appendix A. In this way, knowing γ and T_C, the expressions of Equation (2) can be numerically evaluated in a very direct way.

The average involved at this stage is over realizations of the stochastic process. Under actual operating conditions, where the dispatcher knows the system stress in real-time, Equation (2) makes it possible to predict the MFPT to respond accordingly. However, if we

want to predict the MFPT under prospective stress conditions, we request for a quantity independent of the initial state in order to define a performance measure. Therefore, we need an average over $n = 0, \ldots, L$. For this purpose, we define

$$< T > = \sum_{n=0}^{L} P(n)\, T(n)\,, \tag{3}$$

where $P(n)$ is the probability of residence in the state n. To perform this calculation, we need to know the conditional probability of being at state n at time t given the initial state m, $P(n|m)(t)$. In order to simplify the problem, we propose to calculate $P(n)$ in the steady state regime, which is independent of the initial condition [8]: $P(n) = \lim_{t\to\infty} P(n|m)(t)$. Moreover, given that we are interested in the FPT to the critical condition (first jump to state $L + 1$), we will approximate $P(n)$ by working with the finite Markov chain with reflecting boundaries at sites 0 and L. It is important to note that the steady state probabilities of the finite chain are approximately the same as the residence probabilities of our unbounded chain, since in the lapse before FPT there are no calls in the queue. Under these assumptions, we obtain

$$P(n) = \frac{1}{S}\, \frac{\gamma^{-n}}{n!}\,, \quad \text{for } 0 \le n \le L\,, \tag{4}$$

where S is given by the normalization condition, $S = \displaystyle\sum_{n=0}^{L} \frac{\gamma^{-n}}{n!}$. The mathematical derivation of these expressions is relegated to the Appendix B. The truncated Poisson distribution given in Equation (4) is known as the Erlang *B*-formula, and it has been proven that this equilibrium distribution of the number of occupied servers is independent of the form of the service time distribution [19]. Moreover, the Erlang *B*-formula is also valid for heterogeneous servers, provided however, that all servers have equal mean service times T_S [20].

Equation (3) plus Equations (2) and (4) provide us with a closed form expression for calculating the MFPT averaged on initial states.

2.2. Simulation Framework

To compare and analyse the prediction of Equations (2) and (4) with realistic situations, we have developed a flexible discrete-event simulator (DES) [21] with a process-oriented approach that implements several features inherent to EMS management.

The architecture of our simulator is outlined in Figure 1. The input parameters configure the statistical distributions and set the number of servers (L). The proposed simulator consists of three main modules. The first module, called `simserveRs`, is the simulator kernel. Each thread of the simulation is triggered with a new call arrival. Using a pseudo-random number generator (RNG), it draws a call time, summing an exponential random value to the time of the previous call, and compares it with the release times of busy servers. Then, the kernel iteratively puts the older calls in the queue in service, whereas the release times are less than the last call time (FCFS discipline) [22]. Afterward, if there are no available servers, the incoming call is derived to the waiting queue; otherwise, the call is dispatched to a server, picking it at random among the free ones. At last, `simserveRs` assigns a service time drawn from the desired distribution. Thus, at each event, the simulation engine updates the system state composed by the servers and the queue.

The module `simcritical` launches `simserveRs` with the wanted initial condition and stops the execution when the first call is derived to the queue. Then, the average given by Equation (3) is calculated, and the aggregation over simulations is performed. The last module reckons the MFPT from each initial condition using the analytical expressions given by Equation (2).
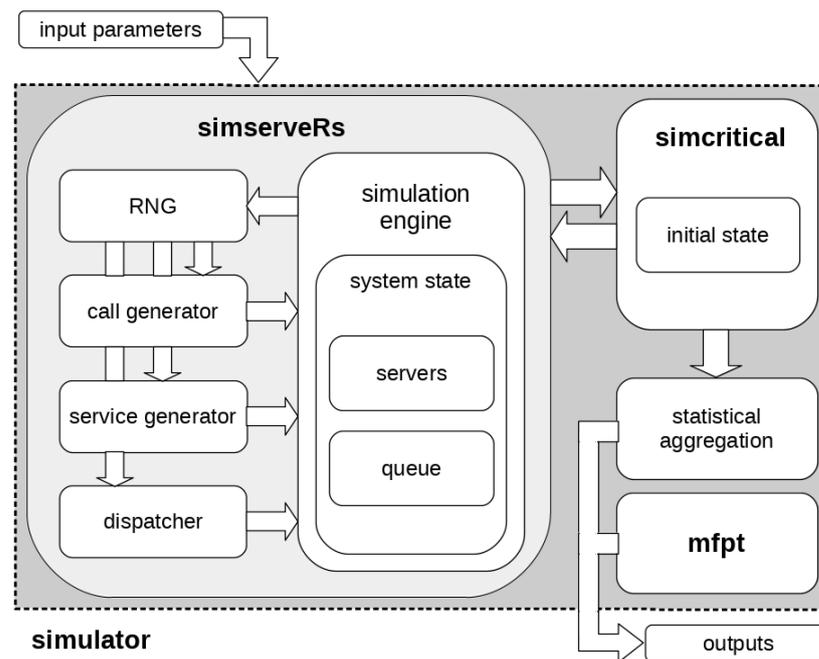
Figure 1. Architecture scheme of the discrete-event simulator.

The software is implemented in *R*, and it is available as open source (see details in Ref. [23]). The simulator can be simply adapted to incorporate most of the features of an actual EMS operator, such as disaggregating the service time into its components (e.g., preparation of ambulance at base, transit time, attention time and transit time to hospital) and implementing dispatching policies with distance or traffic time criteria. Additionally, the extensions of the simulator to any kind of distributions for inter-arrival times (e.g., Erlang) and service times (e.g., gamma or lognormal) are direct. Our simulator was developed in the second half of 2017 for a project related to process optimization in EMS management. The comparison between our simulation outputs and the historical data from an EMS operator showed a satisfactory statistical agreement in several operating scenarios. Over time, several generic DES frameworks for queueing systems have been developed, which deliver such functionalities and implement more efficient methods (see Ref. [24] and references therein). However, for the practical reason of evaluating the results of Section 2.1, the open source version of our simulator becomes an appropriate tool.

3. Results

In Figure 2, we show the non-linear behaviour of $< T >$ as a function of T_C and T_S and its strong dependence on *L*, as they are derived from Equations (2)–(4).
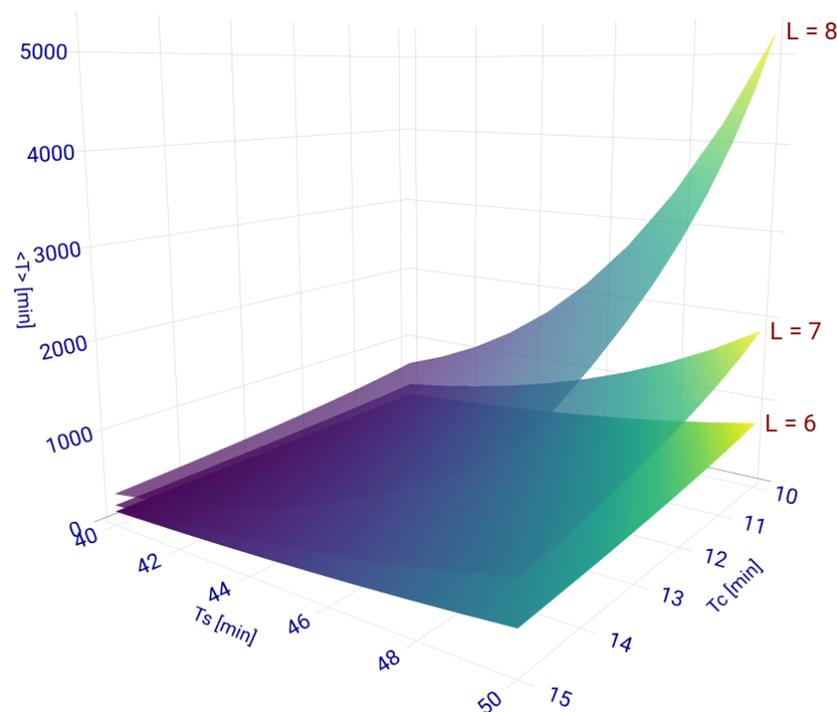
Figure 2. Average MFPT as function of T_C and T_S for three values of L.

The non-linear nature of $< T >$ is given by the powers of γ in Equation (2) and implies a sensitive dependence of the time to the critical condition on its variables. Thus, on the scale of the figure, variations in the order of one minute in T_C or T_S may represent variations of tens to thousands of minutes in $< T >$. The value of L determines the number of terms in Equations (2) and (3). Therefore, adding or removing a server, with the same values of T_C and T_S, can make a substantial change in the value of the average MFPT, as can be seen in Figure 2. The sensitivity of our problem on its parameters is very difficult to grasp intuitively and to predict from past experiences in practice.

The main reason for using the birth and death queueing model ($M/M/L$) is the fact that we can derive the closed-form result given in Section 2. However, the analytical prediction of any performance measure needs to be contrasted with numerical outputs from more realistic scenarios. As an illustration, we take values of inter-arrival and service times measured by the EMS *Sistema de Urgencias del Rosafe* (URG) [25] in Córdoba, Argentina. URG is one of several private EMS operators in a city of about 1.39 million inhabitants. In 2016, URG operated a fleet of nine ambulances that were usually stationed in predefined parking spots scattered throughout the metropolitan zone. The operating scenario distinguishes several daily time bands within which the mean values T_C and T_S are relatively constant, although these, in turn, show seasonal changes throughout the year.

In Figure 3, we show histograms of real data corresponding to 2568 calls received by URG between May 1 and October 31, 2016, late evening (20:00 to 23:00 h). In this time period, we found good stationary statistics in the data, and no critical condition was reported in which there were no ambulances available to serve a call. The service time value measures the time elapsed from dispatch until the ambulance is released. Very low service times correspond to situations that were quickly resolved at scenes close to the ambulance base, whereas the distribution tail involves complex cases with the transfer of the patient to a hospital.
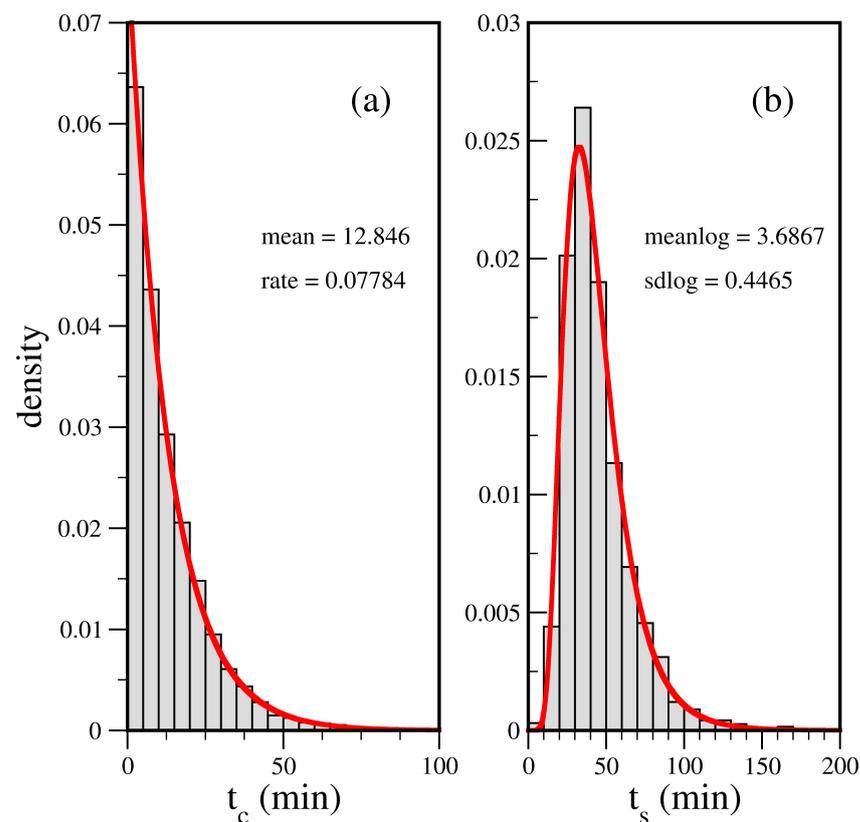
Figure 3. Histograms of real data corresponding to 2568 calls: (**a**) inter-arrival times; (**b**) service times. Solid lines are the best fits: (**a**) exponential; (**b**) lognormal. The insets show the fitting parameters.

From the figure, we can see that the inter-arrival times fit perfectly with an exponential distribution ($T_C = 12.85$ min, goodness-of-fit tests: Kolmogorov–Smirnov p-value $= 0.73$, Cramer–von Mises p-value $= 0.75$), but the best fit for service times is with a lognormal distribution ($T_S = 44.1$ min, goodness-of-fit tests: Kolmogorov–Smirnov p-value $= 0.017$, Cramer–von Mises p-value $= 0.052$). Thus, in this case, it is apparent that the main limitation of our analytical model is the assumption that the distribution of service times is exponential.

The input traffic to a call centre is a nonstationary Poisson process [7,14,26]. However, the arrival rate function $\lambda(t)$ is roughly constant over periods of a few hours [11,14,16,27]. Lognormal distributions for service times have already been reported in the literature [12,14–16]. The fact that the distribution of the sum of a few independent, but not necessarily identical, lognormal random variables could be approximated by a lognormal distribution [28] may explain the experimental findings when we only use two fitting parameters.

In Table 1, we show a comparison between analytical results given by Equation (2) and simulated MFPT from each initial state of a system with seven servers to the critical condition. The simulations were performed using the fitted distributions in Figure 3 and also using an exponential distribution for service times with a mean value equal to that of the fitted distribution in the right panel.

Table 1. Comparison among analytic and simulated values of MFPT (in minutes) for a system with seven servers. The simulated values are reported with its standard error for 10,000 simulations. In both cases, the mean service time is $T_S = 44.1$ min, whereas, in the first case, the probability distribution of service time is *exponential*, and in the second case, it is *lognormal* with `meanlog` $= 3.6867$ and `sdlog` $= 0.4465$. The percentage errors are that of the simulated values with respect to the analytical predictions.

		$T_C = 5$ min			
Initial		Exponential		Lognormal	
State	Analytic	Simulated	$\varepsilon\%$	Simulated	$\varepsilon\%$
0	67.6	67.2 ± 0.4	0.6	49.8 ± 0.3	35.7
1	62.6	62.2 ± 0.4	0.5	44.8 ± 0.3	39.7
2	57.1	56.7 ± 0.4	0.7	39.8 ± 0.3	43.5
3	50.8	50.5 ± 0.4	0.6	34.7 ± 0.3	46.4
4	43.7	43.4 ± 0.4	0.7	29.4 ± 0.2	48.6
5	35.4	35.1 ± 0.3	0.9	23.7 ± 0.2	49.4
6	25.8	25.4 ± 0.3	1.6	17.1 ± 0.2	50.9
7	14.2	14.1 ± 0.2	0.7	9.3 ± 0.2	52.7
		$T_C = 10$ min			
Initial		Exponential		Lognormal	
State	Analytic	Simulated	$\varepsilon\%$	Simulated	$\varepsilon\%$
0	315.2	310.9 ± 2.5	1.4	244.2 ± 2.1	29.1
1	305.2	300.7 ± 2.5	1.5	234.2 ± 2.1	30.3
2	292.9	288.4 ± 2.5	1.6	223.9 ± 2.1	30.8
3	277.3	272.7 ± 2.5	1.7	212.0 ± 2.1	30.8
4	256.7	252.3 ± 2.5	1.7	197.7 ± 2.1	29.8
5	228.1	224.3 ± 2.5	1.7	176.9 ± 2.0	28.9
6	185.6	182.1 ± 2.4	1.9	145.6 ± 2.0	27.5
7	117.7	115.2 ± 2.0	2.2	92.1 ± 1.7	27.8
		$T_C = 15$ min			
Initial		Exponential		Lognormal	
State	Analytic	Simulated	$\varepsilon\%$	Simulated	$\varepsilon\%$
0	1378.4	1380 ± 13	0.1	1201 ± 12	14.8
1	1363.4	1365 ± 13	0.1	1186 ± 12	15.0
2	1343.3	1345 ± 13	0.1	1169 ± 12	14.9
3	1314.6	1317 ± 13	0.2	1147 ± 12	14.6
4	1270.4	1272 ± 13	0.1	1114 ± 12	14.0
5	1195.1	1198 ± 13	0.2	1056 ± 12	13.2
6	1052.2	1058 ± 12	0.5	937 ± 12	12.3
7	745.4	751 ± 12	0.7	670 ± 11	11.3

The analytical predictions agree perfectly with the simulations for the exponentially distributed service times in all cases. For the lognormal distribution of service times, the deviations between the prediction and simulation are about 50% in the worst situations. However, these cases are of little practical interest. Given a fixed number of servers, low values of T_C imply very low MFPT values that are incompatible with EMS response times, while situations with very high values of MFPT are often related with idle infrastructure, which are also avoided in practice.

In order to study the differences in the values of FPT under different service time distributions, in Figure 4, we superimpose two histograms of simulated values for a system with seven servers. Both cases correspond to an initial condition with four occupied servers and the same exponential inter-arrival time distribution, but we use two different service time distributions with the same mean value. The FPT distribution with exponential service times is broader than the lognormal counterpart. Therefore, the MFPT under these

conditions is shorter for the lognormal service time distribution (also see entry for initial state equal to 4 in Table 1).
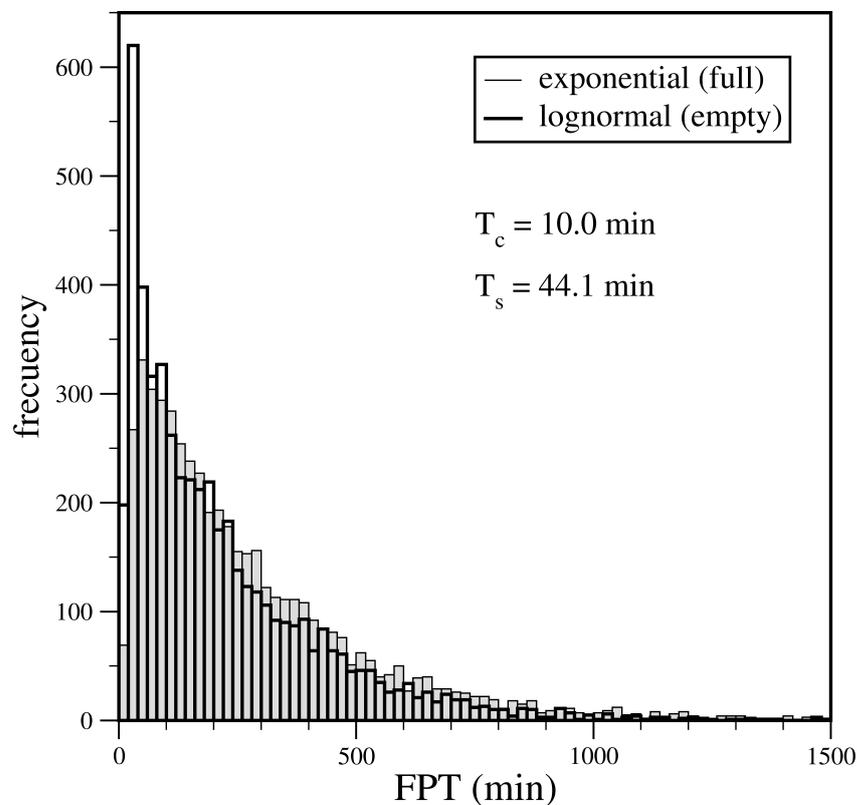
Figure 4. Histograms of FPT to a critical condition based on 5000 simulations from an initial state with 4 of 7 servers occupied. We compare two service time distributions with the same value of T_S: exponential (parameter $= 1/T_S$) and lognormal (`meanlog` $= 3.6867$ and `sdlog` $= 0.4465$).

Now, we investigate what happens with the residence probabilities in the long time limit. The analytical result given by the Erlang formula has been proven for systems without queue or $M/G/L$ blocking systems [19,20]. That is, if every server is busy when a call arrives, the call is lost; however, it is instructive to analyse the situation of a system with queue. The probability of residence in the queue, $P(n > L)$, is equivalent to the probability absorbed in the site $L + 1$ of the Markov chain. From Equation (A11), we can see that the probabilities in Equation (4) are the renormalized residence probabilities of a system with a queue in the interval $[0, L]$. To find out if this is also the case for lognormal service time distributions, we run our simulator 10^7 min in order to visit each state several times. In the first column of Table 2, we show the analytic result of Equations (A11)–(A13) and compare this prediction with the simulated values using four different service time distributions: an exponential and three different lognormal service time distributions—all of these with the same value $T_S = 44.1$ min ($T_C = 10$ min in all cases). The value `sdlog` $= 0.25$ corresponds to the almost symmetric case of the lognormal distribution, and `sdlog` $= 1$ corresponds to the more asymmetric situation (see Figure 6b). The exponential distribution is the case completely asymmetric, where the distribution does not have a maximum value. For comparison, we also introduce the middle value `sdlog`$= 0.625$.

Table 2. Comparison among analytic and simulated long-run probabilities of residence in each state of a system with seven servers, where q denotes the state with queued calls. All simulations correspond with a simulated running time of 1×10^7 min. In all cases, $T_C = 10$ min and $T_S = 44.1$ min, but the parameters in of the lognormal distributions are given by the pairs (`meanlog`, `sdlog`): (a) (3.75521, 0.25), (b) (3.59115, 0.625) and (c) (3.28646, 1.0). See Figure 6b.

State	Analytic	Exponential	Lognorm (a)	lognorm (b)	lognorm (c)
0	0.012	0.011	0.011	0.011	0.012
1	0.051	0.050	0.049	0.050	0.052
2	0.112	0.112	0.109	0.111	0.113
3	0.165	0.165	0.164	0.164	0.165
4	0.182	0.182	0.184	0.182	0.181
5	0.160	0.160	0.167	0.163	0.158
6	0.118	0.118	0.127	0.122	0.115
7	0.074	0.074	0.085	0.081	0.075
q	0.126	0.127	0.104	0.114	0.130

The difference among analytical and simulated values are less than 5% in all states with the exception of the queue. The analysed case with queue is a worse situation than the finite chain with both reflecting ends, used in the derivation of Equation (4), because of the probability of residence in the queue may be greater than the saturated state. Therefore, we find it is valid to use the analytic result of Equation (4) for taking the average in Equation (3) and also in the simulation of MFPT with lognormal service time distributions in a system with queue.

In Figure 5, we show the plots of MFPT averaged over the initial conditions, $< T >$, as a function of the mean inter-arrival time T_C using the probabilities given by Equation (4). We have sketched a characteristic situation for the service time, and we considered several numbers of servers $L = 5, \dots, 9$. The curves clearly show the non-linear behaviour of $< T >$ and allow us to evaluate the quality of the fit for the simulated situations achieved with our analytical expression.
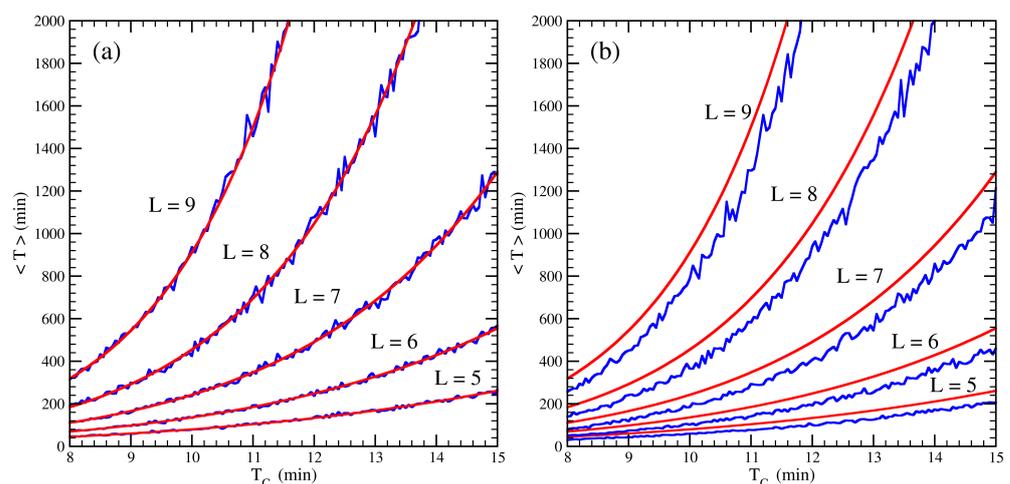
Figure 5. Analytic (red) and simulated (blue) average MFPT with (**a**) exponential and (**b**) lognormal distributed service times (`meanlog` = 3.6867 and `sdlog` = 0.4465). In both cases, $T_S = 44.1$ min. Simulated values for each value of T_C correspond to an average based on 1000 executions.

Again, in the left panel, we find excellent agreement among the analytical predictions and the simulations for exponentially distributed service times. In contrast, in the right panel, for lognormal distributed service times, we see that the analytical curves always run over the corresponding simulation. This fact has just been seen in Figure 4, where the FPT distribution has a longer tail in the case of service times distributed exponentially with respect to lognormal service times. Thus, the mean value of the FPT distribution (MFPT)

is higher for exponential service times. Therefore, the analytic model underestimates the number of necessary servers under a given stress condition. Drawing a horizontal line in Figure 5b, when we move in the direction of increasing demand (that is, lowering T_C), we first cross the blue (simulated) curves in all the considered cases. This is an important fact to keep in mind if we want to use the analytical model to find the best number of servers in a particular operational scenario. However, the discrepancies observed between the simulations and the closed form expression for the average MFPT are not significant enough to cause a considerable effect in the estimation of the optimal number of servers, given the separation among the curves for different values of L. Thus, for example, from Figure 5b, we can see that to obtain an average MFPT of 6 hs, given $T_C = 14$ min, six servers are needed, whereas for $T_C = 10$, eight servers are needed under the same service conditions.

We also analyse the effect of asymmetry in the service time distribution on average MFPT. For this purpose, we simulate the average MFPT for a whole family of service time distribution with fixed T_S but changing the parameter `sdlog` in the interval $[0.25, 1]$, where the minimum value corresponds to the almost symmetric case and the maximum corresponds to the more asymmetric situation. The average MFPT as a function of the `sdlog` parameter is shown in Figure 6.
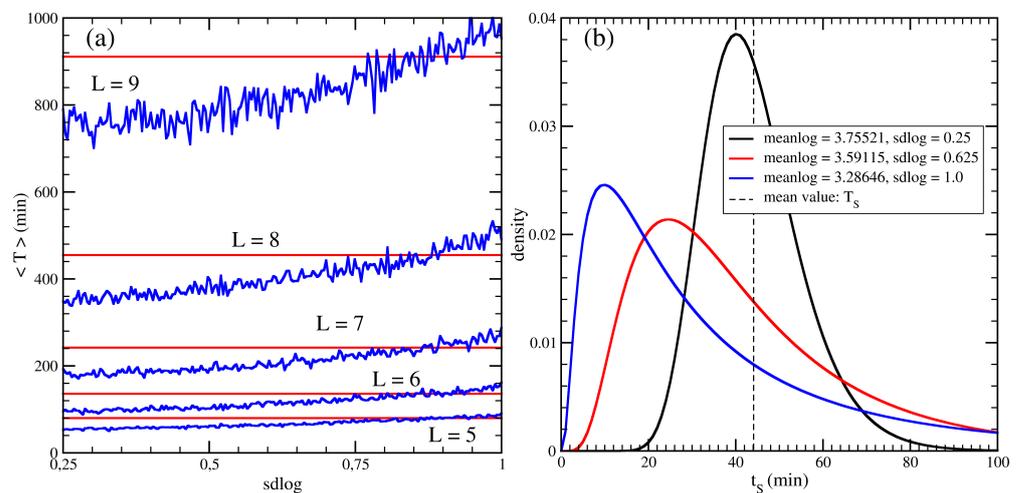
Figure 6. (**a**) Analytic (red) and simulated (blue) average MFPT with lognormal distributed service times with fixed $T_S = 44.1$ min (`sdlog` $\in [0.25, 1]$ and `meanlog` $= \ln(T_S) - $ `sdlog`$^2/2$) and $T_C = 10$ min. Simulated points for each value of `sdlog` are averages based on 1000 runs. (**b**) Sketches of lognormal densities corresponding to the extreme values (black and blue) and to the middle value (red) of `sdlog` in panel (**a**).

In all cases, the analytical prediction gives an acceptable approximation for the simulated data.

Finally, in the left panel of Figure 7, we plot the probability that one or more servers are available at the instant of a emergency call, $P(n < L)$ [27], as a function of the time between calls for a fixed mean service time. We are using Equation (4) for the calculation of this probability. In the right panel, we plot the average MFPT as a function of this availability probability for the same values of T_C given in the left panel. We can observe a very strong sensitivity of the average MFPT to the availability probability for high values of system availability. Thus, only controlling the availability of the system is not enough to assure a long enough time to the critical condition. A dispatcher observing a high availability probability value might conclude that the system is running unreasonably idle; however, the time to critical condition may be shorter than the expectation. More interesting, however, is that the function $< T >$ vs. $P(n < L)$ is practically independent of the number of servers.
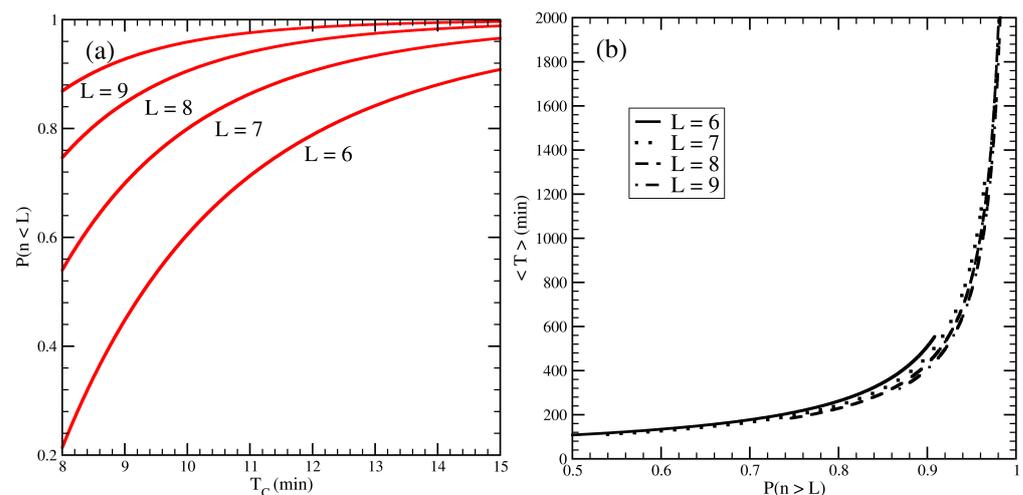
Figure 7. (**a**) The probability of one or more servers are free vs T_C. (**b**) Average MFPT as a function of the probability of servers' availability for $T_C \in [8, 15]$ min. In both panels, $T_S = 44.1$ min.

4. Concluding Remarks

MFPT is a useful KPI that allows estimating the running operative lapse of a system, under a given stress condition, before a service disruption. In this work, we have presented a closed-form expression to calculate the MFPT for a system of servers in parallel, and we also provide a simulation framework for the MFPT. Our formula, based on a birth–death process, only uses the average time between demands and the average service time. Our results make it possible to predict the MFPT given the stress of the system at a particular moment or to analyse the servers shortage time under generic operating conditions by averaging over the initial states of the system. The main limitation of our results, as is often with analytical exact results in queueing theory, is the assumption of an exponential distribution for service times. The impact of this limiting assumption is confronted with results of simulations using more realistic service distributions. Our results indicate that our analytical formula is an acceptable approximation under practical situations. Interesting potential future work may be to consider the implementation of accurate approximations for the $M/G/L$ problem [29] to the MFPT calculation. In addition, our simulation scheme allows us to evaluate the MFPT in any $GI/G/L/FCFS$ server configuration.

Conflicts of Interest: The author declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DES	Discrete-Event Simulator
EMS	Emergency Medical Services
FCFS	First-Come First-Served
FPT	First-Passage Time
KPI	Key Performance Indicator
MFPT	Mean First-Passage Time
RNG	pseudo-Random Number Generator
URG	Sistema de Urgencias del Rosafe

Appendix A. MFPT

For a birth–death process with asymmetric and site-dependent transition probabilities ($w_k^+ \neq w_k^-$), the analytical expression for the MFPT with a reflecting boundary condition at origin and an absorbing one at $L + 1$ is given by

$$
\begin{aligned}
T(0) &= \sum_{k=0}^{L} \frac{1}{w_k^+} + \sum_{k=0}^{L-1} \frac{1}{w_k^+} \sum_{i=k+1}^{L} \prod_{j=k+1}^{i} \frac{w_j^-}{w_j^+} , \\
T(1) &= T(0) - \frac{1}{w_0^+} , \\
T(n) &= T(0) - \sum_{k=0}^{n-1} \frac{1}{w_k^+} - \sum_{k=0}^{n-2} \frac{1}{w_k^+} \sum_{i=k+1}^{n-1} \prod_{j=k+1}^{i} \frac{w_j^-}{w_j^+} \quad \text{for } 2 \le n \le L .
\end{aligned}
\tag{A1}
$$

For details and derivation of Equation (A1), see Section 6 in Ref. [18]. In our model with L servers, using Equation (1) and the parameter γ defined in the text, we can recast the products in Equation (A1) as

$$
\prod_{j=k+1}^{i} \frac{w_j^-}{w_j^+} = \gamma^{i-k} \prod_{j=k+1}^{i} j = \gamma^{i-k} \frac{i!}{k!} .
\tag{A2}
$$

Thus, we can also recast the sums in Equation (A1) as

$$
\sum_{i=k+1}^{n-1} \prod_{j=k+1}^{i} \frac{w_j^-}{w_j^+} = \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{n-1} i! \, \gamma^i ,
\tag{A3}
$$

and

$$
\sum_{k=0}^{n-2} \frac{1}{w_k^+} \sum_{i=k+1}^{n-1} \prod_{j=k+1}^{i} \frac{w_j^-}{w_j^+} = \frac{1}{\lambda} \sum_{k=0}^{n-2} \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{n-1} i! \, \gamma^i .
\tag{A4}
$$

Replacing the last expression in Equation (A1), we obtain in a direct manner Equation (2) in Section 2.1.

Appendix B. Steady State

Following Ref. [8], we can construct the steady state for the problem of a Markov chain with a reflecting boundary condition at the origin. In the long-run, the time-independent probability of residence at state n, π_n, must satisfy

$$
\begin{aligned}
\omega_1^- \pi_1 - \omega_0^+ \pi_0 &= 0 , \\
\omega_{n+1}^- \pi_{n+1} + \omega_{n-1}^+ \pi_{n-1} - (\omega_n^+ + \omega_n^-) \pi_n &= 0 , \quad \text{for } n \ge 1 .
\end{aligned}
\tag{A5}
$$

Thus, we can prove by induction that

$$
\pi_n = \frac{\omega_{n-1}^+ \dots \omega_0^+}{\omega_n^- \dots \omega_1^-} \pi_0 = \prod_{j=1}^{n} \frac{\omega_{j-1}^+}{\omega_j^-} \pi_0 , \quad \text{for } n \ge 1 .
\tag{A6}
$$

From the normalization condition, $\sum_{n=0}^{\infty} \pi_n = 1$, results $\pi_0 = 1/S_\pi$, where

$$
S_\pi = 1 + \sum_{n=1}^{\infty} \prod_{j=1}^{n} \frac{\omega_{j-1}^+}{\omega_j^-} .
\tag{A7}
$$

The existence of the steady state is then determined by the convergence of the series in Equation (A7).

For the model given by Equation (1), the products in Equation (A6) and (A7) can be written as,

$$
\prod_{j=1}^{n} \frac{\omega_{j-1}^{+}}{\omega_{j}^{-}} =
\begin{cases}
\left(\dfrac{\lambda}{\mu}\right)^{n} \displaystyle\prod_{j=1}^{n} \dfrac{1}{j} = \dfrac{\gamma^{-n}}{n!}\,, & \text{for } 0 \le n \le L\,, \\[2ex]
\left(\dfrac{\lambda}{\mu}\right)^{n} \displaystyle\prod_{j=1}^{L} \dfrac{1}{j} \prod_{j=L+1}^{n} \dfrac{1}{L} = \dfrac{\gamma^{-n}}{L!}\,\dfrac{1}{L^{n-L}}\,, & \text{for } n \ge L\,.
\end{cases}
\tag{A8}
$$

Substituting Equation (A8) into Equation (A7), we obtain

$$
S_{\pi} = \sum_{n=0}^{L} \frac{\gamma^{-n}}{n!} + \frac{L^{L}}{L!} \sum_{n=L+1}^{\infty} (L\,\gamma)^{-n}\,.
\tag{A9}
$$

The convergence of the series in the last expression only occurs if $L\,\gamma > 1$. In this case,

$$
\sum_{n=L+1}^{\infty} (L\,\gamma)^{-n} = \frac{(L\,\gamma)^{-L}}{L\,\gamma - 1}\,.
\tag{A10}
$$

Substituting Equations (A8)–(A10) into Equation (A6), we obtain

$$
\pi_{n} = \frac{1}{S_{\pi}}
\begin{cases}
\dfrac{\gamma^{-n}}{n!}\,, & \text{for } 0 \le n \le L\,, \\[2ex]
\dfrac{L^{L}}{L!}\,(L\,\gamma)^{-n}\,, & \text{for } n \ge L\,,
\end{cases}
\tag{A11}
$$

where

$$
S_{\pi} = \sum_{n=0}^{L} \frac{\gamma^{-n}}{n!} + \frac{\gamma^{-L}}{L!\,(L\gamma - 1)}\,.
\tag{A12}
$$

In this manner, the long-term probability of having the system *with calls in the waiting queue* results in

$$
\pi_{q} = \sum_{n=L+1}^{\infty} \pi_{n} = \frac{\gamma^{-L}}{S_{\pi}\,L!\,(L\gamma - 1)}\,.
\tag{A13}
$$

The last expression is also is known as Erlang *C*-formula.

We now consider the case of a *finite* Markov chain with reflecting boundaries at the origin and at site L. The time-independent probabilities of residence, $P(n)$, in the each state n satisfy Equation (A5) but are supplemented with the additional reflecting condition at L,

$$
\begin{aligned}
\omega_{1}^{-}\,P(1) - \omega_{0}^{+}\,P(0) &= 0\,, \\
\omega_{n+1}^{-}\,P(n+1) + \omega_{n-1}^{+}\,P(n-1) - (\omega_{n}^{+} + \omega_{n}^{-})\,P(n) &= 0\,, \;\; \text{for } n \ge 1\,, \\
\omega_{L-1}^{+}\,P(L-1) - \omega_{L}^{-}\,P(L) &= 0\,.
\end{aligned}
\tag{A14}
$$

Following the above procedure, we find the first line of Equation (A8) again, which directly leads to Equation (4) in Section 2.1, where S is the normalization on the interval $[0, L]$.

References

1. Dos Santos Cabral, E.L.; Castro, W.R.S.; de Medeiros Florentino, D.R.; de Araújo Viana, D.; da Costa Junior, J.F.; de Souza, R.P.; Rêgo, A.C.M.; Araújo-Filho, I.; Medeiros, A.C. Response time in the emergency services. Systematic review. *Acta Cirúrgica Brasileira* **2018**, *33*, 1110–1121. [CrossRef] [PubMed]
2. Eisenberg, M.; Bergner, L.; Hallstrom, A. Cardiac Resuscitation in the Community: Importance of Rapid Provision and Implications for Program Planning. *J. Am. Med. Assoc.* **1979**, *241*, 1905–1907. [CrossRef]
3. Feero, S.; Hedges, J.R.; Simmons, E.; Irwin, L. Does out-of-hospital EMS time affect trauma survival? *Am. J. Emerg. Med.* **1995**, *13*, 133–135. [CrossRef]

4. Pons, P.T.; Markovchick, V.J. Eight minutes or less: Does the ambulance response time guideline impact trauma patient outcome? *J. Emerg. Med.* **2002**, *23*, 43–48. [CrossRef]
5. McCoy, C.E.; Menchine, M.; Sampson, S.; Anderson, C.; Kahn, C. Emergency Medical Services Out-of-Hospital Scene and Transport Times and Their Association with Mortality in Trauma Patients Presenting to an Urban Level I Trauma Center. *Ann. Emerg. Med.* **2013**, *61*, 167–174. [CrossRef] [PubMed]
6. Al-Shaqsi, S.Z.K. Response time as a sole performance indicator in EMS: Pitfalls and solutions. *Open Access Emergercy Med.* **2010**, *2*, 1–6. [CrossRef]
7. Singer, M.; Donoso, P. Assessing an ambulance service with queuing theory. *Comput. Oper. Res.* **2008**, *35*, 2549–2560. [CrossRef]
8. Winston, W.L. *Operations Research: Applications and Algorithms*, 4th ed.; Duxbury Press: Belmont, CA, USA, 2003; Chapter 20.
9. Larson, R.C. A hypercube queuing model for facility location and redistricting in urban emergency services. *Comput. Oper. Res.* **1974**, *1*, 67–95. [CrossRef]
10. Takeda, R.A.; Widmer, J.A.; Morabito, R. Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Comput. Oper. Res.* **2007**, *34*, 727–741. [CrossRef]
11. Uriarte, A.G.; Zúñiga, E.R.; Moris, M.U.; Ng, A.H.C. How can decision makers be supported in the improvement of an emergency department? A simulation, optimization and data mining approach. *Oper. Res. Health Care* **2017**, *15*, 102–122. [CrossRef]
12. Cildoz, M.; Ibarra, A.; Mallor, F. Accumulating priority queues versus pure priority queues for managing patients in emergency departments. *Oper. Res. Health Care* **2019**, *23*, 100224. [CrossRef]
13. Lakshmi, C.; Sivakumar, A.I. Application of queueing theory in health care: A literature review. *Oper. Res. Health Care* **2013**, *2*, 25–39. [CrossRef]
14. Brown, L.; Gans, N.; Mandelbaum, A.; Sakov, A.; Shen, H.; Zeltyn, S.; Zhao, L. Statistical Analysis of a Telephone Call Center. *J. Am. Stat. Assoc.* **2005**, *100*, 36–50. [CrossRef]
15. Ingolfsson, A.; Budge, S.; Erkut, E. Optimal ambulance location with random delays and travel times. *Health Care Manag. Sci.* **2008**, *11*, 262–274. [CrossRef]
16. Van Buuren, M.; Kommer, G.J.; van der Mei, R.; Bhulai, S. A simulation model for emergency medical services call centers. In Proceedings of the 2015 Winter Simulation Conference (WSC), Huntington Beach, CA, USA, 6–9 December 2015; pp. 844–855. [CrossRef]
17. Cox, D.R.; Smith, W.L. *Queues*; Methuen: London, UK, 1961.
18. Pury, P.A.; Cáceres, M.O. Mean first-passage and residence times of random walks on asymmetric disordered chains. *J. Phys. A Math. Gen.* **2003**, *36*, 2695–2706. [CrossRef]
19. Takacs, L. On Erlang's Formula. *Ann. Math. Stat.* **1969**, *40*, 71–78. [CrossRef]
20. Fakinos, D. The $M/G/k$ Blocking System with Heterogeneous Servers. *J. Oper. Res. Soc.* **1980**, *31*, 919–927. [CrossRef]
21. Banks, J.; Carson II, J.S.; Nelson, B.L.; Nicol, D.M. *Discrete-Event System Simulation*, 5th ed.; Pearson Education Limited: Harlow, UK, 2014.
22. Krivulin, N.K. A recursive equations based representation for the $G/G/m$ queue. *Appl. Math. Lett.* **1994**, *7*, 73–77. [CrossRef]
23. Pury, P.A. simserveRs: R Simulator for Parallel Servers with FCFS Queueing Discipline. Available online: https://bitbucket.org/realhubot/simservers/ (accessed on 22 August 2021).
24. Ucar, I.; Smeets, B.; Azcorra, A. simmer: Discrete-Event Simulation for R. *J. Stat. Softw.* **2019**, *90*, 1–30. [CrossRef]
25. Sistema de Urgencias del Rosafe SA. Available online: https://www.urg.com.ar/ (accessed on 22 August 2021).
26. Sze, D.Y. OR Practice—A Queueing Model for Telephone Operator Staffing. *Oper. Res.* **1984**, *32*, 229–249. [CrossRef]
27. Hall, W.K. Management science approaches to the determination of urban ambulance requirements. *Socio-Econ. Plan. Sci.* **1971**, *5*, 491–499. [CrossRef]
28. Wu, J.; Mehta, N.B.; Zhang, J. Flexible lognormal sum approximation method. In Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '05), St. Louis, MO, USA, 28 November–2 December 2005; Volume 6, pp. 3413–3417. [CrossRef]
29. Hokstad, P. Approximations for the $M/G/m$ Queue. *Oper. Res.* **1978**, *26*, 510–523. [CrossRef]