First in Line Waiting Times as a Tool for Analysing Queueing Systems

G.M. Koole[†], B.F. Nielsen[⋆], T.B. Nielsen[⋆]

[†]Dept. Mathematics
VU University Amsterdam
De Boelelaan 1081, 1081 HV, the Netherlands.

[⋆]Dept. Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads, 2800 Kgs. Lyngby, Denmark.

May 5, 2012

Abstract

We introduce a new approach to modelling queueing systems where the priority or the routing of customers depends on the time the first customer has waited in the queue. This past waiting time of the first customer in line, W^{FIL}, is used as the primary variable for our approach. A Markov chain is used for modelling the system where the states represent both the number of free servers and a discrete approximation to W^{FIL}. This approach allows us to obtain waiting time distributions for complex systems, such as the N-design routing scheme widely used in e.g. call centers and systems with dynamic priorities.

Keywords: Waiting time distribution; Call centers; Priority queues; Deterministic threshold; Erlang distribution; Dynamic priority; Due-date.

1 Introduction

The traditional approach to modelling queueing systems developed by Erlang, Engset, Fry, and Molina has been to look at the number of customers in queue, see [10] for original references. In this paper we introduce an alternative way of modelling queueing systems, which is useful for priority rules often used in real scenarios.

Prioritization of customers in real queueing systems is often implemented as a function of the waiting time of the customer first in line (FIL). An example of this is the use in call centers to route calls between agent pools in order to meet service levels of different customer classes [8]. The service levels are usually characterized by the telephone service factor (TSF), i.e. the fraction of calls answered within a certain time. The use of TSF thus motivates the examination of waiting time distributions rather than just average values. Prioritization based on waiting times is also used in the health care sector for e.g. operating room scheduling [4].

The traditional approach of modelling the number of customers in a system is not useful for modelling these systems as no information about the FIL waiting time, W^{FIL}, is available when only counting the number of customers. Instead we introduce the Erlang Approximation (EA) as a way of modelling the waiting time of the customer first in line.

The EA is based on a Markov chain where the states constitute a discrete representation of $W^{\mathrm{FIL}}$when one or more customers are waiting. When queues are empty, the states represent the number of free servers.

We show how the EA can be used for finding the waiting time distribution for different queueing systems. The approach taken is to first use the EA for a simple $M/M/n$ queueing system where the results can be verified by solving the equilibrium equations explicitly. The approach is then applied to more complex systems where explicit solutions are not obtainable. In these cases, the approximate nature of the EA is induced by the finite state spaces used for modelling the systems.

The fundamental idea of approximating a continuous variable using a number of discrete states and inferring a continuous distribution from the probability distribution of being in the states using the Erlang distribution is not new. This approach has for example been taken in [1]. Indeed, finding a desired distribution by summing over a set of Erlang distributions is a fundamental tool in stochastic modelling and can e.g. be used for determining the waiting time distribution of an $G/M/1$-system [9].

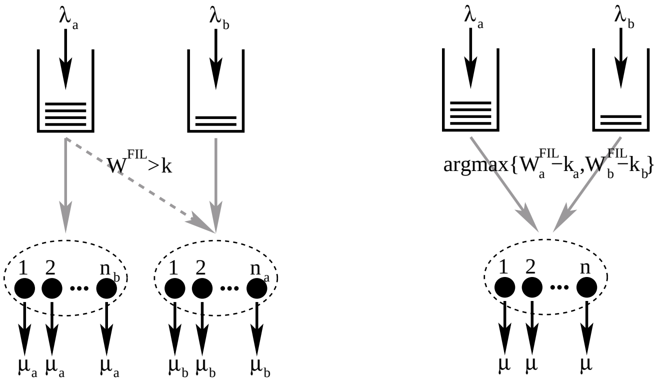
The use of waiting time to describe a stochastic process is also known from the standard method used for modeling $GI/G/1$ systems [9]. Here the waiting time of the $n+1$'th customer is given by Lindley's equation as $W_{n+1} = \max\{0,\ W_n + S_n - A_{n-1}\}$, where A is the interarrival time and S is the service time. Whenever both A and S are exponential, there is no need to consider the elapsed service time or interarrival time and W_n is sufficient to describe a Markov chain. The novelty of the EA lies in the use of W^{FIL}together with the discrete approximation, which allows for obtaining waiting time distributions for complex systems.

The literature on this kind of routing is limited despite its widespread use in industry. In [3] a system is considered where a single queue is served by two servers of which one is only allowed to take in customers when they have waited a given amount of time, and an expression for the waiting time distribution is found. In [2] a similar time-dependent overflow model is approximated by state-dependent overflow rates where states represent the number of customers in the system. A system less closely related to ours is given in [16] where service times are exponentially distributed with parameters which depend on the waiting time experienced by the customer entering service.

In Section 2 the EA is presented for the simple case of an $M/M/n$ queue and it is shown how the waiting time distribution can be found. The principles behind the EA are easier to explain using the simple system and it also facilitates verification of the approximation by comparison with theoretical values.

In Section 3 it is shown how the EA can be used for modelling the N-design system with deterministic thresholds. The N-design system is illustrated in Figure 1a. It gets its name from the way the arrows representing the possible routing options form an "N". It is composed of a group of flexible servers and a group of specialized servers and two classes of customers namely a-customers and b-customers. The system treated here is special in the regard that only a-customers having an elapsed waiting time, W^{FIL}, of more than a given deterministic threshold k, are allowed to overflow to b-servers. There exist multiple papers about the N-design: In [19], performance measures, such as average queueing delay and queue length distribution, are derived for a bilingual call center where a fraction of the agents are bilingual and the rest unilingual. In [18] the N-design is also examined and it is shown that a few flexible servers can improve mean service time substantially. However, neither of these papers deal with the deterministic overflow threshold.

An example of the general nature of the EA is given in Section 4 where the approach is used for a system with low and high priority customers and scheduling according to due dates, also referred to as dynamic priority [13], [7]. Scheduling according to due date is a generalization of strict non-preemptive prioritization where high priority customers are not always given service first. This prioritization scheme is for example used for operating rooms scheduling to ensure lower priority operations will not experience exceedingly long waiting times while still keeping high priority patients' waiting times relatively low [4].



(a) N-design

(b) V-design

Figure 1: Systems analyzed in Section 3 (Fig. 1a) and Section 4 (Fig. 1b).

Possible extensions of the EA are also discussed in Section 5. Only non-preemptive cases are considered here although our method could also be used for preemptive (non-resume) prioritization. Finally a conclusion is reached in Section 6.

2 The Erlang approximation

The fundamental idea of the Erlang approximation is to describe the FIL waiting time in a discretized form as states in a continuous time Markov chain. The waiting time distribution experienced by customers is derived by considering the states from which the customers enter service. This approach is explained and applied to an $M/M/n$-system in this section.

In order for the system to be Markovian, we need to have information about the waiting time of the second customer in line. However, we can find the waiting time of the second in line as $W^{\text{FIL}} - A$, where A is the interarrival time of customers. In the case $A > W^{\text{FIL}}$, there will be no second customer and a server will be released.

We define W^{FIL} at time t as $t - t_A$, where t_A is the time at which, the entity first in line arrived. Thus, W_t^{FIL} is the waiting time of the first customer in line at time t, with the convention $W_t^{\text{FIL}} = 0$ when the queue is empty.

As W^{FIL} is continuous, so will the state space of the resulting Markov process be. In order to make the analysis easier, we introduce the Erlang Approximation, where the continuous time is modelled by an Erlang distribution with rate γ.

In the EA we let states numbered $\{1, 2, ...\}$ represent $W_t^{\text{FIL}} > 0$. Transitions from state i to $i + 1$ with rate γ represent the linearly increasing FIL waiting time.

Whenever the queue is empty we have to differentiate between the number of free servers. This is done by letting states with negative numbers represent the number of free servers. Thus, state $-n$ corresponds to all n servers being vacant and state 0 to all servers being busy and the queue being empty. Lumping together the states representing free servers and W^{FIL}in one dimension can be done as servers can not be free while customers are waiting.

The underlying continuous time Markov chain can now be summarized by the set of transition rates from states i to j, $g(i,j)$, given in Equation (1)-(2), where the $p_{i,j}$'s are given in Theorem 2.1.

$$g_{i,i+1} = \begin{cases} \lambda, & \text{for } -n \leq i \leq 0 \\ \gamma, & \text{for } 0 < i \end{cases} \tag{1}$$

$$g_{i,i-j} = \begin{cases} p_{i,i-j}(n+i)\mu, & \text{for } -n \leq i \leq 0 \\ p_{i,i-j}n\mu, & \text{for } 0 < i, \ 0 \leq j \leq i \end{cases} \tag{2}$$

Theorem 2.1 *The discretization of the interarrival distribution is given as:*

$$p_{i,i-h} = \begin{cases} 1 - \sum_{h=0}^{i-1} \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h, & \text{for } i = h, \ 0 < i \\ \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^j, & \text{for } 0 \leq h < i \\ 1, & \text{for } i \leq 0, h = 1 \\ 0, & \text{else,} \end{cases}$$

where $p_{i,i-h}$ is the probability of the next state being state $i-h$, given a service completion happens in state i.

Proof For $i > 0$, the theorem follows directly from the geometric distribution, where the probability of having a γ-transition in a given state is $\gamma/(\lambda+\gamma)$. This means that the probability of going from state i to $i-h$ equals the probability of having had h, γ-transitions since the last λ-transition. The probability of a transition to state 0 is thus 1 minus the sum of probabilities of transitions to the other states. For $i \leq 0$, a service completion leads to one more server becoming vacant, thus $p_{i,i-1} = 1$. □

Arrivals occur with rate λ, see Equation (1) for $-n \leq i \leq 0$. In the states, $i < 0$, where servers are unoccupied, an arrival goes to a free server and i increases by one. In state $i = 0$, were all servers are busy and the queue empty, an arriving customer enters the queue and becomes the first customer in line. This means i becomes 1 and W_t^{FIL} starts to increase, as a customer is now waiting in the queue.

The increasing waiting time, W^{FIL}, is represented by the γ-transitions, Equation (1) for $i > 0$. Thus, γ-transitions do not represent a particular event, but the continuously evolving time. Customers arriving when $i > 0$ are not seen by the model as these arrivals do not affect W_t^{FIL}.

Service completions occur with rate $n\mu$ for $i \geq 0$ and rate $(n+i)\mu$ for $i < 0$, as this corresponds to the number of working servers times the service rate. For $i \leq 0$, a service completion leads to a transition to state $i-1$, see Equation (2), as this means an additional server becomes vacant.

Whenever a service completion occurs at time t and the queue is not empty, i.e. $i > 0$, W^{FIL}decreases with $min\{A, W_t^{\mathrm{FIL}}\}$, where A is a random variable describing the inter-arrival time of customers. That is, if $W^{\mathrm{FIL}} < A$ the queue becomes empty and i becomes 0. If $W^{\mathrm{FIL}} > A$, the second customer in line moves up and becomes the first in line and $W^{\mathrm{FIL}}$is decreased by A.

In the $M/M/n$ queue considered, A is exponentially distributed as the inter-arrival times in a Poisson arrival process are exponential. This reduction of W^{FIL}corresponds to a set of transitions from state i to states $i-j,\ j \in \{0,\ 1, ...,\ i\}$ as described in Equation (2). In the discrete setting of the EA, the exponential interarrival times translates to the length of these transitions following a geometric distribution as given in Theorem 2.1.

A finite state space is needed in order to determine the state probabilities at stationarity numerically. This is done by truncating the number of phases describing the waiting time at state D, so that $i \in \{-n, ..., D\}$. By solving $\boldsymbol{\pi}\mathbf{G} = \mathbf{0}$, where $\mathbf{G}$ is a generator matrix containing the transition rates, $g_{i,j}$, $i \neq j$, given in Equations (1)-(2) and diagonal elements $g_{i,i} = -\sum_{j=-n,\ j\neq i}^{D} g_{i,j}$.

The truncation of the state space introduces the risk of having a large probability mass in the truncated state, particularly if $\gamma \gg n\mu$. The value of γ has a significant influence on the approximation. Increasing it means that more states are required for the truncation of states not to have a too significant influence on the precision of the approximation. However, at the same time having γ large improves the approximation as it then better represents the continuously elapsing time. This suggests having a threshold on probability mass in the truncated state, e.g. $\pi_D < 0.001$, otherwise γ as large as possible. The structure of $\mathbf{G}$ further suggests that the value of γ should be increasing with μ, n, and D.

2.1 Waiting time distribution

In order to determine the waiting time distribution, the embedded Markov chain at service initiations is considered. Service initiations occur at λ-transitions from states with vacant servers, i.e. $i < 0$ and μ-transitions from states $i > 0$.

The state probability just before a service initiation, i.e. in the embedded Markov chain, from state i is denoted $\alpha_{\mu\lambda}(i)$. We let the distribution $\alpha_{\mu\lambda}(i)$ be given in the vector $\boldsymbol{\alpha}_{\mu\lambda}$ which can be found as:

$$\alpha_{\mu\lambda}(i) = \frac{\pi(i)\Lambda_{\mu\lambda}(i)}{\sum_{j=-n}^{D}\pi(j)\Lambda_{\mu\lambda}(j)}, \tag{3}$$

where $\pi(i)$ is the steady-state probabilities. $\Lambda(i)_{\mu\lambda}$ is the sum of the transition intensities from state (i) that result in a service initiation and the subscripts show the transitions that are considered. For the $M/M/n$ system $\Lambda(i)_{\mu\lambda}$ becomes

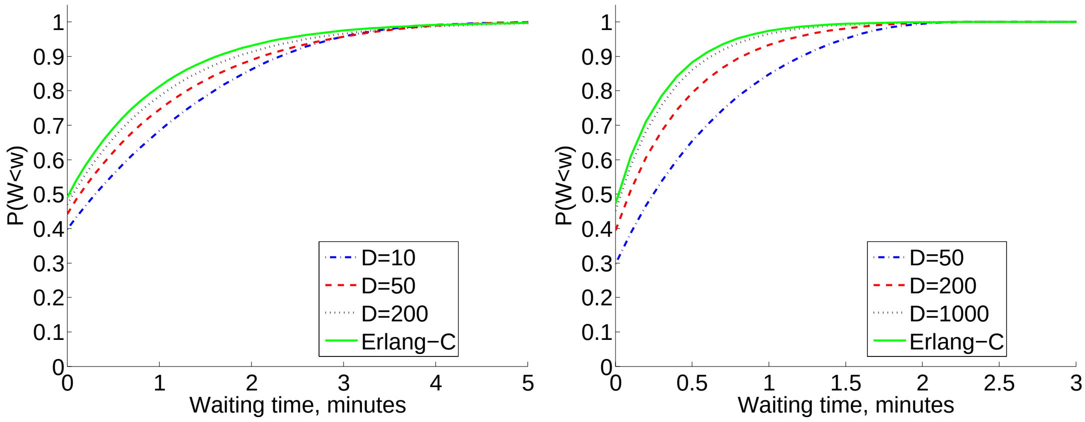
$$\Lambda_{\mu\lambda}(i) = \begin{cases} \lambda & \text{for } i < 0; \\ 0 & \text{for } i = 0; \\ n\mu & \text{for } 0 < i \leq D, \end{cases} \tag{4}$$

where we should note that $n\mu = \sum_{j=0}^{i} p_{i,j} n\mu$. The waiting time distribution can now be found from (3) and (4). A customer entering service when $i < 0$ goes directly to a free server and experiences no waiting time, this is represented by the first sum in

Equation (5). When a customer enters service from a state $i > 0$, he/she has waited a sum of i exponentially distributed time periods, each with mean $1/\gamma$. Let $F_\Gamma(t; i, \gamma) = 1 - \sum_{h=0}^{i-1} \frac{(\gamma t)^h}{h!} e^{-\gamma t}$ be the cdf of an Erlang-distribution with shape parameter $i \in \mathbb{N}$ and scale parameter $\gamma \in \mathbb{R}_+$, then we have the second sum in Equation (5). The waiting time distribution, as approximated by the EA, of a customer entering the system then becomes:

$$P(W \leq t) \approx \sum_{i=-n}^{-1} \alpha_{\mu\lambda}(i) + \sum_{i=1}^{D} F_\Gamma(t; i, \gamma)\alpha_{\mu\lambda}(i), \tag{5}$$

Figure 2 shows the waiting time distribution as found by the EA for a small (2a) and a large system (2b) respectively, both compared to the theoretical distribution given as $F(t) = 1 - E_{2,n}(A)e^{-(n-A)\mu t}$, $A < n$, $t \geq 0$, where $A = \lambda/\mu$ is the offered traffic and $E_{2,n}(A)$ is the delay probability as given by Erlang's C-formula [5]. It can be seen that the approximation does indeed converge for $D \to \infty$, as desired.



(a) $n = 4$, $\lambda = 3$, $\mu = 1$, $\gamma = \frac{2\mu}{3\lambda}(Dn)^{\frac{3}{4}}$

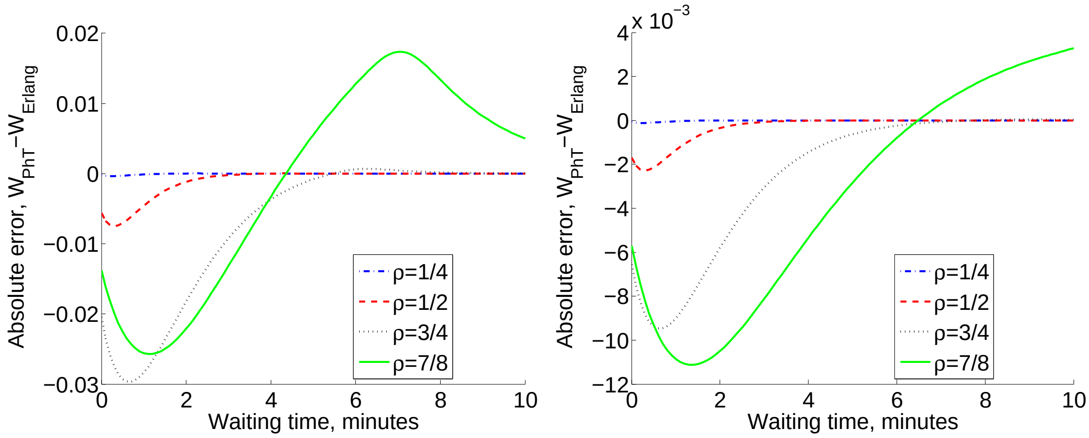
(b) $n = 40$, $\lambda = 37$, $\mu = 1$, $\gamma = \frac{3\mu n}{4\lambda}D^{0.9}$

Figure 2: The waiting time distribution as found by the EA compared to Erlang-C for a small (2a) and large system (2b) respectively. The basic time unit for the transitions can be chosen arbitrarily, however for the application to a call center setting, minutes would be the obvious choice.

In Figure 3 the absolute error for different values of the load on the system, $\rho = \lambda/(\mu n)$, is shown for two different model sizes, $D = 200$ (3a) and $D = 1000$ (3b). It is seen that the EA is able to handle different values of the load, ρ, well.

2.2 Infinite state space

Instead of truncating the state space, we can solve the EA of the $M/M/1$ queue for an infinite number of states. Solving the EA for $n > 1$ would also be straightforward, however, the calculations in e.g. Equation (5) becomes more involved and the additional insight gained is limited. The generator matrix, G, constructed from Equations (1)-(2) for $i \geq 1$ bears a strong resemblance to the corresponding matrix of the embedded Markov chain at arrivals for a $GI/M/1$ system, see [9]. In the $GI/M/1$ case, the number of customers is given by a geometric distribution at stationarity, hence the analogy and resemblance of the



(a) $D = 200$, $n = 4$, $\mu = 1$, $\gamma = \frac{2\mu}{3\lambda}(Dn)^{\frac{3}{4}}$

(b) $D = 1000$, $n = 4$, $\mu = 1$, $\gamma = \frac{2\mu}{3\lambda}(Dn)^{\frac{3}{4}}$

Figure 3: The absolute error of the waiting time distribution as found by the EA is compared to the corresponding distribution from Erlang-C for different loads. Fig. 3a shows a coarser approximation than Fig. 3b. Note that as, without loss of generality, $\mu = 1$ we have the load, ρ, equals λ/n.

generator matrices. This motivates guessing a solution to the steady state probabilities in the form of $\pi(i) = \theta\beta^i(1-\beta)$, where θ is a normalizing constant. Solving the infinite set of equations $\boldsymbol{\pi}\mathbf{G} = \mathbf{0}$ yields an expression for the steady state probabilities;

$$\pi_i = \begin{cases} \dfrac{\mu\gamma(\mu-\lambda)}{\lambda^3+\mu^2\gamma} & \text{, for } i = -1 \\ \dfrac{\lambda\gamma(\mu-\lambda)}{\lambda^3+\mu^2\gamma} & \text{, for } i = 0 \\ \dfrac{\lambda^2(\mu-\lambda)}{\lambda^3+\mu^2\gamma}\left(\dfrac{\lambda+\gamma}{\mu+\gamma}\right)^i & \text{, for } i \geq 1. \end{cases} \tag{6}$$

We can find the waiting time distribution from the steady state probabilities (6) in the same way as for the truncated system described in Section 2.1, the resulting expression becomes

$$P(W \leq t) = 1 - \frac{\lambda(\lambda+\gamma)}{\lambda^2+\mu\gamma} e^{\frac{\gamma}{\mu+\gamma}(\lambda-\mu)t}.$$

This converges to the known expression for the $M/M/1$ queue [9] for $\gamma \to \infty$, which proves that the EA does indeed converge to the correct solution in this simple case.

3 Call center modelling

In this section it is shown how the Erlang approach introduced in Section 2 can be used for modelling the system shown in Figure 1a. This system is referred to as an N-design due to the outline of the routing scheme [8]. The system is motivated by the call handling often seen in call centers, where the servers would be agents answering calls from customers with different needs or importance.

In the setup, two job types, a and b, arrive at queue a and queue b with rates λ_a and λ_b respectively. Queue a is served by a group of n_a servers each handling jobs with service rate μ_a; queue b is similarly served by n_b servers each working with service rate μ_b. When the waiting time, $W_{t,a}$, of the first job in queue a exceeds a limit, k, the job is allowed to go to server group b with non-preemptive priority over jobs in queue b. This means that there may be vacant b-servers even if there are a-customers waiting. As before γ-transitions represent the increasing FIL-waiting time with the addition that we let FIL-waiting times for type a and b jobs increase simultaneously when jobs are queued in both queues. The Erlang model used to describe this system is defined by Equations (7)-(12).

$$g_{(i,j),(i+1,j)} = \begin{cases} \lambda_a, & \text{for } i \leq 0 \\ \gamma, & \text{for } 0 < i < m,\ 0 < j \\ 0, & \text{else} \end{cases} \tag{7}$$

$$g_{(i,j),(i+1,j+1)} = \begin{cases} \gamma, & \text{for } 0 < i,\ 0 < j \\ 0, & \text{else} \end{cases} \tag{8}$$

$$g_{(i,j),(i,j+1)} = \begin{cases} \lambda_b, & \text{for } j \leq 0 \\ \gamma, & \text{for } 0 < j \end{cases} \tag{9}$$

$$g_{(i,j),(i-k,j)} = \begin{cases} p_{i,i-k}\mu_a(n_a + i), & \text{for } i \leq 0 \\ p_{i,i-k}\mu_a n_a, & \text{for } 0 < i < m \\ p_{i,i-k}(\mu_a n_a + \mu_b n_b), & \text{for } m < i \end{cases} \tag{10}$$

$$g_{(i,j),(i,j-k)} = \begin{cases} p_{j,j-k}\mu_b(n_b + j), & \text{for } j \leq 0,\ i \leq m \\ p_{j,j-k}\mu_b n_b, & \text{for } 0 < j,\ i \leq m \\ 0, & \text{for } m < i \end{cases} \tag{11}$$

$$g_{(m,j),(m,j+1)} = \begin{cases} p_{m,m-k}\gamma, & \text{for } j < 0 \\ 0, & \text{else} \end{cases} \tag{12}$$

Here, the Erlang approximation introduced in Section 2 is extended to two dimensions, one for each queue and server group. States are denoted (i, j), where i and j represent the a and b jobs respectively in the same way as for the $M/M/n$ case described in Section 2. We let $\mathcal{X}$ denote the complete set of states. The deterministic threshold, k, on the overflow is modelled by m phases after which the a-calls are allowed to go to the b-server group. In this 2-dimensional model the FIL-waiting times for queues a and b are truncated at states D_a and D_b respectively.

When at least one of the servers in group b is vacant and queue a reaches state m, the next γ-transition will result in the first job in queue a being allowed to go to a free server in group b thus decreasing the waiting time of the first in line in queue a and decreasing the number of vacant servers in group b by one. This situation corresponds to the diagonal double arrows originating from states $(m, j < 0)$.

Extra care has to be taken when determining the waiting time distribution for the two job types. The transitions leading to service initiations in the N-design model are λ_a and λ_b-transitions from states with vacant servers, i.e. $i < 0$ and $j < 0$ respectively, γ-transitions from states where $i = m$ and $j < 0$, and finally μ_a and μ_b-transitions from states $i > 0$ and $j > 0$ respectively.

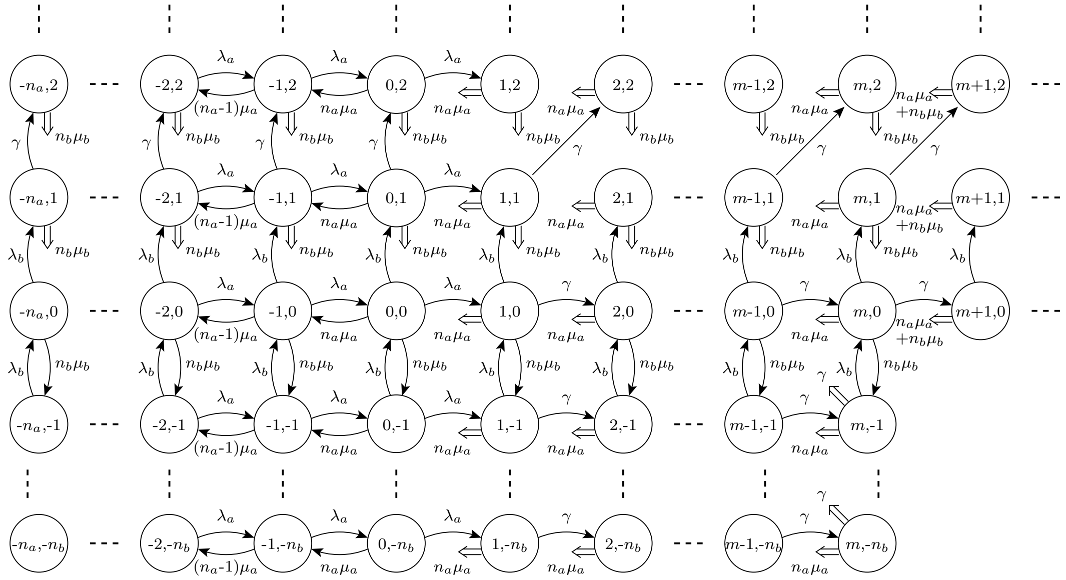


Figure 4: Illustration of the Erlang model. The horizontal and vertical directions describe queues a and b respectively. Negative indices refer to the number of vacant servers in each server group and positive indices refer to the waiting time of the first customer in line. The double arrows illustrate a distribution of transition intensities to a number of states in the given direction with total intensity as given next to the arrow. This distribution is given in Theorem 2.1.

Besides the obvious differentiation between service initiations for a and b jobs, it is also necessary to deal with service initiations for a jobs due to γ transitions in a special way. This is due to γ-transitions representing customers having waited exactly the time until the fixed threshold k and not a gamma-distributed amount of time.

The state probabilities just before service initiations are denoted $\alpha^a_{\mu\lambda}(i,j)$ and $\alpha^b_{\mu\lambda}(i,j)$ for a and b-calls respectively. The distribution of $\boldsymbol{\alpha}_{\mu\lambda}$ can be found as:

$$\alpha^a_{\mu\lambda}(i,j) = \frac{\pi(i,j)\Lambda^a_{\mu\lambda}(i,j)}{\sum_{(i,j)\in\mathcal{X}} \pi(i,j)\Lambda^a_{\mu\lambda\gamma}(i,j)}, \tag{13}$$

$$\alpha^a_{\gamma}(m,j) = \frac{\pi(m,j)\Lambda^a_{\gamma}(m,j)}{\sum_{(i,j)\in\mathcal{X}} \pi(i,j)\Lambda^a_{\mu\lambda\gamma}(i,j)}, \qquad j<0. \tag{14}$$

Here $\Lambda^a(i,j)$ is the sum of the transition intensities from state (i,j) that result in a service initiation for a-calls. The subscripts on Λ are used to differentiate between the different transitions that lead to service initiations. For a-calls, Λ is given as:

$$\Lambda^a_{\mu\lambda}(i,j) = \begin{cases} \lambda_a & \text{for } i < 0 \\ 0 & \text{for } i = 0 \\ n_a\mu_a & \text{for } 0 < i \leq m \\ n_a\mu_a + n_b\mu_b & \text{for } m < i \end{cases} \tag{15}$$

$$\Lambda^a_{\gamma}(i,j) = \begin{cases} \gamma & \text{for } i = m,\ j < 0; \\ 0 & \text{else.} \end{cases} \tag{16}$$

$$\Lambda^a_{\mu\lambda\gamma}(i,j) = \Lambda^a_{\mu\lambda}(i,j) + \Lambda^a_{\gamma}(i,j) \tag{17}$$

For the b-calls the expressions are somewhat simpler:

$$\Lambda^b_{\mu\lambda}(i,j) = \begin{cases} \lambda_b & \text{for } j < 0,\ i \leq m; \\ n_b\mu_b & \text{for } 0 < j,\ i \leq m \\ 0 & \text{else.} \end{cases} \tag{18}$$

Using the same approach as in Section 2, the waiting time distribution for type a jobs can be found from (13) and (14) as:

$$P(W_a \leq t) \approx \begin{cases} \sum_{i=-n_a}^{-1} \sum_{j=-n_b}^{D_b} \alpha^a_{\mu\lambda}(i,j) + \sum_{i=1}^{D_a} \left[F_\Gamma(t;i,\gamma) \sum_{j=-n_b}^{D_b} \alpha^a_{\mu\lambda}(i,j) \right], & t < k, \\ \sum_{i=-n_a}^{-1} \sum_{j=-n_b}^{D_b} \alpha^a_{\mu\lambda}(i,j) + \sum_{j=-n_b}^{-1} \alpha^a_{\gamma}(m,j) & \\ \qquad + \sum_{i=1}^{D_a} \left[F_\Gamma(t;i,\gamma) \sum_{j=-n_b}^{D_b} \alpha^a_{\mu\lambda}(i,j) \right], & t \geq k, \end{cases}$$

The waiting time distribution for type b jobs can be found in a similar fashion as:

$$P(W_b \leq t) \approx \sum_{i=-n_a}^{D_a} \sum_{j=-n_b}^{-1} \alpha^b_{\mu\lambda}(i,j) + \sum_{j=1}^{D_b} \left[F_\Gamma(t;j,\gamma) \sum_{i=-n_a}^{D_a} \alpha^b_{\mu\lambda}(i,j) \right],$$

where the α^b can be found in the same way as for the a-jobs from Equation (13) using (18) instead of (15). γ-transitions never lead to service initiations for b-jobs thus leading to the simpler expression for $P(W_b \leq t)$.

In Figure 5, the EA is compared to simulations for a large system. The simulations are carried out for the exact continuous time model in order to validate the discrete approximation of the EA. The EA is generally more computationally efficient than doing the amount of simulations required to obtain the tight 95% confidence intervals used for Figure 5. The waiting time distribution is used for the comparison to better illustrate the special behavior with the discontinuity at k, than would be the case if the density was used for the figure. The fact that it is hard to distinguish the lines for the simulations and the model only serves to illustrate how well the EA is able to approximate the waiting time distributions. Exactly solving the large system to which the EA has been applied

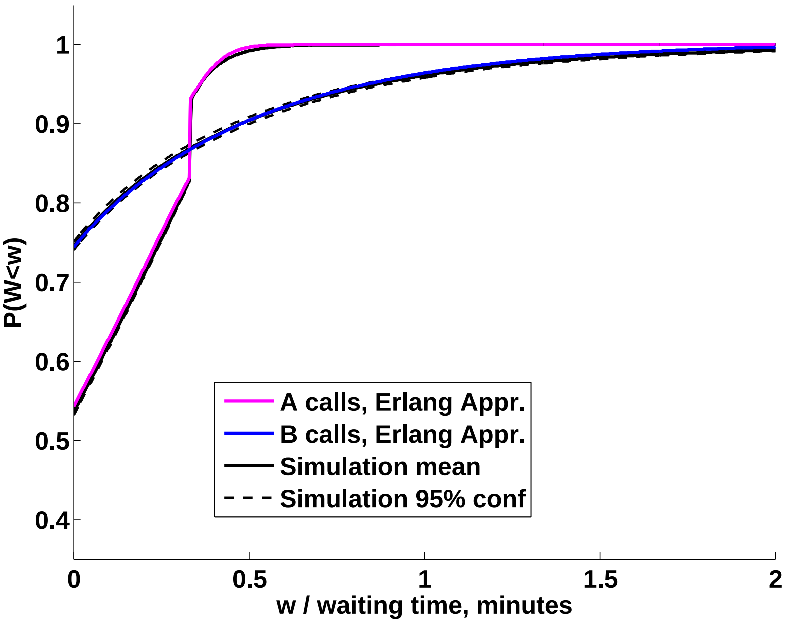


Figure 5: The Erlang approximation compared to simulations with 95% confidence intervals. The parameters used are $k = 0.33min$, $\lambda_a = 7min^{-1}$, $\lambda_b = 10min^{-1}$, $\mu_a = 0.33min^{-1}$, $\mu_b = 0.5min^{-1}$, $\gamma = 90min^{-1}$, $n_a = 20$, $n_b = 27$, $m = 30$, $D_a = 36$ and $D_b = 200$.

for Figure 5 is not feasible. This is discussed in detail in [3] and would require solving a system with a continuous density function for each of the servers.

Worth noting in Figure 5 is the change in convexity of the waiting time distribution for A-calls. This phenomenon was also observed for a simple two-server system treated in [3]. It originates in, that for $W^{\text{FIL}} < k$, the a-customers are served only by a-servers, and since $\lambda_a > n_a\mu_a$, the waiting time distribution for $w < k$ becomes convex. When $w > k$, a-customers are also served by the b-server, and as $\lambda_a < n_a\mu_< + n_b\mu_b$, the distribution becomes concave.

4 Dynamic prioritization

The possibilities of the EA are not limited to the N-design presented in Section 3. In this section we show how the EA, introduced in Section 2, can be used for a system with dynamic priority. In this case the next customer to be served when a server becomes available at time t, is selected from two queues, a and b, by $max\{W_{t,a} - k_a, W_{t,b} - k_b\}$, where the k's are constants and the W_t's are FIL-waiting times. Arrivals are assumed to occur according to a Poisson process with rates λ_a and λ_b to the two queues. Jobs from both queues are served by a joint group of n servers that each handle the jobs with exponential service time with mean $1/\mu$. If one queue is empty, jobs from the other will always be taken into service immediately if a server is vacant. The system is illustrated in Figure 1b.

As is shown in [6], the average waiting time of the low priority customers in a static priority system with two classes grows quickly when the relative frequency of high-priority customers increases. This prioritization scheme is thus desirable when exceedingly long waiting times should be avoided for all classes while still giving some priority to certain customers.

The exact formulation of the additive prioritization scheme may vary as only the difference of the constants matters, adding k_a and k_b to $W_{t,b}$ and $W_{t,a}$ respectively would be equivalent to the formulation used here. Here it is formulated by subtraction of a constant for each queue which especially for systems with more than two queues makes things clearer as it can be interpreted as a target waiting time for each queue.

This prioritization scheme was first introduced in [11] and is referred to as dynamic priority or sequencing according to due-date in the literature. In [12], Jackson's conjecture is given, stating that the tails of the waiting time distributions for the different priority classes are exponential with the same shape, i.e. decay rate. Indeed the shape of the tails are the same as for an FCFS system, only shifted $k_a - k_b$ from each other. The conjecture doesn't say anything about the part of distribution closer to the origin where much of the probability mass lies, which is where the EA can give a good approximation.

The state space of the EA for the V-design system with dynamic priority becomes two-dimensional as the FIL-waiting times of both queues need to be taken into account. States $i < 0$ keep track of the number of vacant servers in the same way as for the $M/M/n$ system treated in Section 2. Whenever a server finishes a job and $i > 0$, the next customer taken into service must be chosen according to $max\{W_{t,a} - k_a, W_{t,b} - k_b\}$. This is implemented in the model by dividing the state space into two parts representing either option as defined by Equations (19)-(20).

$$g_{(i,j),(i-k,j)} = \begin{cases} p_{i,i-k}(n+i)\mu, & \text{for } i \leq 0,\ j = 0 \\ p_{i,i-k}n\mu, & \text{for } 0 < i,\ \wedge i - m_a \geq j - m_b \\ 0, & \text{else,} \end{cases} \tag{19}$$

$$g_{(i,j),(i,j-k)} = \begin{cases} p_{j,j-k}n\mu, & \text{for } 0 < j\ \wedge i - m_a < j - m_b \\ 0, & \text{else} \end{cases} \tag{20}$$

where the state wise interpretation of the target waiting times, m_a and m_b, need to be inferred from the actual target waiting times, k_a and k_b, as $m_a = \gamma k_a$ and $m_b = \gamma k_b$.

Equations (21)-(23) defines the rest of the EA implementation for the system with dynamic priority.

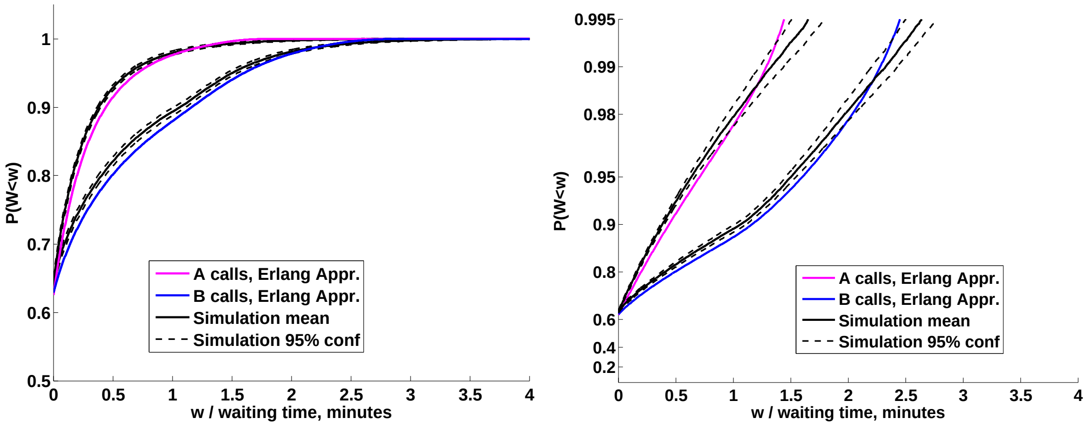
$$g_{(i,j),(i+1,j)} = \begin{cases} \lambda_a + \lambda_b, & \text{for } i < 0,\ j = 0 \\ \lambda_a, & \text{for } i = 0 \\ \gamma, & \text{for } 0 < i, j = 0 \\ 0, & \text{else} \end{cases} \tag{21}$$

$$g_{(i,j),(i+1,j+1)} = \begin{cases} \gamma, & \text{for } 0 < i,\ 0 < j \\ 0, & \text{else,} \end{cases} \tag{22}$$

$$g_{(i,j),(i,j+1)} = \begin{cases} \lambda_b, & \text{for } 0 \leq i,\ j = 0 \\ \gamma, & \text{for } i = 0,\ 0 < j \\ 0, & \text{else,} \end{cases} \tag{23}$$

In Figure 6, a plot of the waiting time distribution for a dynamic priority two class system is shown as found by the EA. Customers are always taken into service immediately when a server is vacant, thus the distributions start at the same value in 0. Figure 6 illustrates Jackson's conjecture [12] as it plots the waiting time distributions on a logarithmic scale thus producing straight lines for the exponential tails. Simulation results with confidence intervals are shown for validation.

The apparently large divergence of the tails in Figure 6b are due to the truncated state space, however though the divergence are highly magnified by the logarithmic plot, the absolute differences are small. The significant widening of the confidence intervals is a result of few events in the simulation experiencing long waiting times, as well as the previously mentioned magnifying effect of the logarithmic plot.



(a) Waiting time distribution.

(b) Waiting time distribution, logarithmic plot.

Figure 6: The Erlang approximation compared to simulations for a system with dynamic priority. The shorter waiting times of the high priority a-customers are clearly seen as that curve lies above the one for b-customers. The parameters used are $k_a = 0.25$, $k_b = 1.25$, $\lambda_a = 3.5$, $\lambda_b = 2.5$, $\mu = 1$, $n = 8$, $D_a = 90$, $D_b = 150$, and $\gamma = 60$.

A possible extension to the dynamic priority model would be to keep servers free to cater for potential a-customers even though b-customers may be in line. This could be treated as an optimization problem and approached using Markov Decision Processes. It would be somewhat analogous to the slow-server problem [14], [15] and [17].

5 Further work

In this section we discuss how discretionary prioritization and the inclusion of abandonments in models could be approached. A setup with overflow to a back office, such as the one dealt with in [2], could also be analyzed with the EA.

The discretionary priority discipline is a mix of the preemptive and non-preemptive disciplines, [13]. The idea is that high priority customers are only allowed to preempt lower priority customers under certain circumstances, e.g. that a low priority customer has just started service. In this case the EA could be used in an inverted way by modelling the elapsed service time of the last low priority customer who entered service and thus only allowing high priority customers to preempt when this value is under a given threshold.

The inclusion of abandonments is often considered essential when dealing with call centers [20]. It should be possible to extend the model to take abandonments into account by introducing transitions representing abandonments of the FIL customer. A negative binomial distribution should be used in Theorem 2.1 instead of the geometric distribution in order to account for the possibility of customers further back in the line having abandoned the system.

Another possible expansion of the EA is to use it as a foundation for optimizing the overflow and prioritization schemes described in this paper. For the fixed threshold model treated in Section 3, this could be done by introducing a choice of whether to allow overflow in each state where b-customers are waiting ($j > 0$) and using Markov Decision Processes to find an optimal policy. This could not only be used to verify the appropriateness of the commonly used fixed threshold policy, but also to find better policies.

6 Conclusion

We introduced a new approach to modelling queueing systems by using a discrete approximation of the waiting time of the customer first in line as the primary variable. This has proved useful when dealing with systems where routing or priority of customers depends on this waiting time as seen in many real scenarios such as call centers.

The Erlang Approximation (EA) was introduced in Section 2 and we showed that the resulting waiting time distribution indeed converges to the theoretical value when the state space increases in size.

In Section 3 we implemented the EA for an N-design routing scheme with deterministic threshold on the overflow between server groups, a design often used in call centers. We showed that it is possible to get a good approximation of the waiting time distribution and that the EA can thus be an alternative approach to examining complicated systems where simulation studies have otherwise been seen as the only viable approach. Indeed, it is possible to obtain computational advantages by using the EA as compared to simulations, depending on the variables used. Also, an interesting phenomenon where the convexity of the waiting time distribution changes at the fixed threshold was observed and discussed.

Further possibilities of the EA were discussed in Sections 4-5 including how abandonments could be taken into account and how the routing policies could be optimized. We also showed how the EA can be used to model a system with dynamic priority, thus showing the flexibility of the EA.

References

[1] Adan, Ivo and Resing, Jacques. A two-level traffic shaper for an on-off source. *Performance Evaluation*, 42, 4:279–298, 2000.

[2] Wolfgang Barth, Michael Manitz, and Raik Stolletz. Analysis of Two-Level Support Systems with Time-Dependent Overflow - A Banking Application. *Production and Operations Management*, 19: 757-768, 2010.

[3] René Bekker, Ger Koole, Bo Friis Nielsen, and Thomas Bang Nielsen. Queues with waiting time dependent service. *Queueing Systems*, 68(1): 61–78, 2011.

[4] Margaret L. Brandeau, François Sainfort, and William P. Pierskalla. *Operations Research and Health Care: A Handbook of Methods and Applications*. Springer, illustrated edition, 2004.

[5] E. Brockmeyer, H.L. Halstrøm, and Arne Jensen. The Life and Works of A.K. Erlang. *Transactions of the Danish Academy of Technical Sciences*, No. 2, 1948.

[6] Alan Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

[7] Richard W. Conway, William .L. Maxwell, and Louis W. Miller. *Theory of Scheduling.* Addison-Wesley Publishing Company, 1967.

[8] Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003.

[9] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes.* Oxford University Press, third edition, 2001.

[10] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory.* Wiley-Interscience, third edition, 1998.

[11] James R. Jackson. Some problems in queueing with dynamic priorities. *Nav. Res. Logistics Quart.*, 7:235–249, 1960.

[12] James R. Jackson. Queues with dynamic priority discipline. *Management Science*, 8(1):18–34 and 2627272, 1961.

[13] N.K. Jaiswal. *Priority Queues.* Academic Press, New York and Londons, 1968.

[14] G.M. Koole. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems & Control Letters*, 26:301–303, 1995.

[15] W. Lin and P.R. Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Automat. Control*, 29:696–703, 1984.

[16] M.J.M Posner. Single-server queues with service time dependent on waiting time. *Operations Research*, 21:610–616, 1973.

[17] M. Rubinovitch. The slow server problem. *Journal of Applied Probability*, 22:205–213., 1985.

[18] Robert A. Shumsky. Approximation and analysis of a call center with flexible and specialized servers. *OR Spectrum*, 26(3):307–330, 2004.

[19] D.A. Stanford and W.K. Grassmann. The bilingual server system: a queueing model featuring fully and partially qualified servers. *INFOR*, 31(4):261–277, 1993.

[20] W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, 2005.