A Predictive Lossless Queue for Extremely Large Dataset Transfer in Markov Communication Systems

Samuel Nlend[1], **Theo G. Swart**[1†], and **Bhekisipho Twala**[2], Non-members

ABSTRACT

In this paper, a parametric prediction model is proposed for a queuing system driven by the Markov process. The aim of the model is to minimize the packet loss caused by time dependency characterized by the queue tail being too long, resulting in a time-out during the transfer of a large dataset. The proposed technique requires the key parameters influencing the queue content to be determined prior to its occupation regardless of the server capacity definition, using Bayesian inference. The subsequent time elapsing between the arrival and departure of a packet in the system is given, as well as the system load. This queue content planning is considered as the Markov birth-death chain; a type of discretization that characterizes almost all queuing systems, leading to an exponential queue, and captured herein using beta distribution. The inference results obtained using this exponential queue indicate that the queue with predictive parameters employing beta distribution, even when dealing with a loss system queue, involves less transition time and a greater load than a queue with coarse parameters; hence, preventing a long tail in the queue which is the cause of packet loss.

Keywords: Markov Chain, Geometric Distribution, Bayesian Inference, Beta Distribution, Queue Length, Waiting Time, Load

1. INTRODUCTION

The Markov birth-death chain has been used extensively to characterize queue behavior, but various solutions have been brought to the fore, resulting in different queuing systems. Most queue models are based on *a priori* probabilities or proportions, resulting in parameters which are actually coarse estimates, particularly when moving from the infinite to the finite queue server, or simply dealing with a finite queue server. As a result, the Markov queue (M), characterized by a constant rate captured through the Poisson process, has evolved toward the general consideration of queue distribution (G) to avoid packet loss caused by tails or time dependency.

However, all analyses using solutions such as the Markov matrix [1, 2], moment generating [3–6], and fluid method [7–9], developed to solve the queue problem, have certain similarities:

- A type of in-out queue consideration often labeled as arrival-departure, which can be considered as the Markov birth-death chain.
- Results that achieve a geometric server distribution $p_n = (1-\rho)\,\rho^n$, $n = 0, 1, 2, \ldots, \infty$, with ρ being the load and n the number of packets.

In fact, from this derived geometric distribution queue length, there are some queue solutions that require approximation and discretization to achieve a geometric queue distribution by letting $\rho = \exp(-\theta t)$, with the load being embedded in the parameter θ. However, since this type of approximation is achieved for general distribution (G) at the system equilibrium (when $t \to \infty$), the queue always empties itself as t elapses regardless of service completion. This results in a renewal process which entails extra provisions to deal with packet loss in the aggregate heavy traffic.

The contribution of this paper is therefore to address the data transfer time-out issue through the beta inference of parameter ρ in the queue which is likely to avoid packet loss as the time elapses ($t \to \infty$).

In this regard, several studies have been carried out to estimate ρ. The direct and indirect approaches capture ρ through moment solutions from general distribution based on the arrival process in G/M/1 queues, the service process in M/G/1 queues or both processes in G/G/1 queues. According to the direct approach statistics, the mean and variance of the parameters are captured through solutions such as moment and Kingman's bounds [10] to the spectral bounds [11] based on inter-arrival time conditions [12]. Better accuracy is provided by [13] and [14] who, in developing the indirect approach, used estimators from the geometric distribution proposed by [15] and [16] and followed the approaches of [17] and [18]. However, the results are then obtained based on *a priori* probability, and moment averages of the general distribution (G) queue, resulting again in parameters which are coarse in nature.

This paper considers these probabilities and parame-

Manuscript received on October 26, 2021; revised on January 28, 2022; accepted on March 7, 2022. This paper was recommended by Associate Editor Nattapong Kitsuwan.

[1]The authors are with the Center of Telecommunications, University of Johannesburg, South Africa.

[2]The author was with The Institute of Intelligent System, University of Johannesburg, South Africa.

[†]Corresponding author: tgswart@uj.ac.za

Digital Object Identifier: 10.37936/ecti-eec.2022203.247513

ters as variables by applying the Bayesian inference since it is useful for evaluating proportions, percentages, or probabilities that may vary with the beta distribution. In this regard, [19] and [20] used Bayesian inference to predict M/M/1 and M/G/1 queues respectively, relying on Monte Carlo simulation since some analytical moments were not tractable for [20]. While the researchers in [19] reveal that the intensity takes an exponential form of the inference parameters $\Gamma(a, b)$: $\rho(t) = bMt^{b-1}$ where $M = (1 \,/\, a)^b$, and $\Gamma(a, b)$ is a gamma function with parameters a, b.

In this paper, for better sizing of the queuing system for dealing with time-out issues, predictive queue behavior is proposed through Bayesian inference that considers the queue's parameters ρ, or θ for M/M/1, M/G/1, G/M/1, and G/G/1 queues as random variables. Since the beta distribution with parameters a, b, denoted by $B(a, b)$ is continuous, it can be used in prior statistical analysis for geometric distributions with a finite size; we make use of it and give the necessary conditions for which the predictive queue parameters, namely ρ, or θ, can prevent tails from forming in the queue.

This paper is organized as follows: Section 2 reviews the current queue solutions from the M/M/1 queue to the G/G/1 systems, emphasizing the approximations that lead to their exponential behaviors. This allows the profound understanding of each queue prior to its system parameterization, as explained in Section 3, where the system parameters are determined and presented using the Beta distribution. Section 4 provides the results based on a comparison between ρ from the coarse probability approach and the one suggested in this paper.

2. CURRENT QUEUE DISCRETIZATION METHODS

2.1 The M/M/1 Queue

The M/M/1 queue is modeled as a single-server with an infinite capacity queuing system. This results in the packets arriving according to the Poisson process (M) at an arrival rate of λ packets per unit time and an exponential (M) service time distribution at a service rate of μ packets per unit time. The memoryless property of the exponential distribution, here referred to as M allows the birth-death process as a modeling tool for this queue (Figs. 1 and 2).

When a packet enters an empty system, its service starts immediately. When dealing with a full system, the incoming packet is queued and enters the service facility once the service of the packet ahead has been completed.

The birth-death process for queue modeling illustrated in Fig. 2 is considered a Markov chain with time-homogeneous transition probabilities, characterized by the fact that the discrete state variable changes by one at most during an infinitesimal time interval. The resulting probability distributions governing the number of births-deaths in a specific time interval depend on the length of that interval and not on its starting point. These probability distributions are given in [1], and

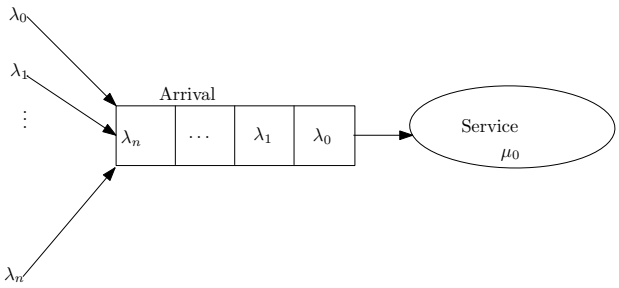

***Fig. 1**: Queue FCFS facility.*

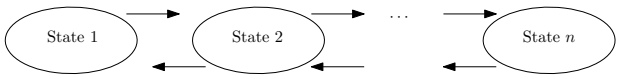

***Fig. 2**: Birth-death Markov chain.*

the transition probabilities obtained using Chapman-Kolmogorov's equation in [21], given by

$$p_{ij}(t + h) = \sum_{k=0}^{\infty} p_{ik}(h)p_{kj}(t), \tag{1}$$

which is combined with the Bayesian formula of [22] and applied to the state probabilities P_n of [1] to yield

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t) + \lambda_{n-1}P_{n-1}(t) \tag{2}$$

with λ_n and μ_n being the arrival and service rates at state n, which results in the transition matrix of [1].

Adan and Resing [2] take into account the state probabilities $P_n(t)$ at $t \to \infty$, $\Delta t \to 0$ under the assumption $dP_n(t) \,/\, dt = 0$ of Eq. (2) with $n = 0, 1, 2, \ldots$ and $P_n(t) \to P_n$, they then reach the queue length distribution

$$p_n = (1 - \rho)\rho^n, \quad n > 0, \quad \rho < 1. \tag{3}$$

2.2 G/M/1 and M/G/1 Queues

The G/M/1 or M/G/1 queue results from the difficulties of achieving an acceptable arrival or service process in the M/M/1 queue. Its analysis relies on the embedded Markov process that yields the matrix solution combined with the probability density function (PDF). This section presents a G/M/1 queue acknowledging its duality with an M/G/1 queue [23] where the transition matrix is obtained using [24].

2.2.1 Matrix solution

The queue parameters are obtained by applying the Markov chain to arrival instants. Hence, the system state to reach is given by the number in the system n_i with service S_i immediately before the arrival instants, as shown in Fig. 3.

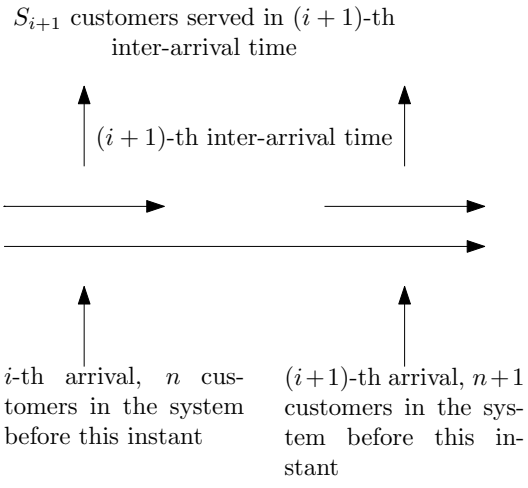

***Fig. 3**: Embedded Markov chain for G/M/1 queue.*

If the system has served S_{i+1} packets between the i-th and the $(i+1)$-th arrival, then $n_{i+1} = n_i + 1 - S_{i+1}$, $n = 0, 1, 2, \ldots$ and $S_{i+1} \leq n+1$. The equilibrium state of this embedded Markov chain is reached when $i \to \infty$. Under this condition, let $n_{i+1} = k$ and $n_i = j$; the one-step probability transition of the chain is applied to produce a quasi-similar expression of [1] with probability which is now in the form

$$P_{jk} = \begin{cases} P[n_{i+1} = k | n_i = j], & k > j+1 \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where P_{jk} is the probability that $(k-(j+1))$ packets are served between inter-arrival times. The equilibrium state probability is derived from the Chapman-Kolmogorov equation in the M/M/1 queue as

$$P_k = \sum_{j=0}^{\infty} P_j P_{jk} \quad \text{for} \quad k = 0, 1, 2, \ldots, \tag{5}$$

with P_j the probability to find j packets in the system state considered but with an arbitrary arrival instant for $j = 0, 1, 2, \ldots, \infty$.

To obtain the transition probabilities, [25] introduces α_n defined as the probability that n packets with distribution $a(t)$, are served during an interval time with at least n packets in the system state

$$\alpha_n = \int_0^{\infty} \frac{(\mu t)^n}{n!} \exp(-\mu t) a(t) dt. \tag{6}$$

The moment-generating solution is obtained by considering the probability expression α_n in Eq. (6) as the coefficients of z^n in the Laplace-Stieltjes transform of $a(t)$, denoted as $L_A(\mu - \mu z)$, which is the PDF of the inter-arrival times [26]. This implies that

$$\sum_{j=0}^{\infty} \alpha_j z^j = \int_0^{\infty} a(t) dt = L_A(\mu - \mu z). \tag{7}$$

Then, the series expansion of $L_A(\mu - \mu z)$ yields the coefficients α_j of the z^j, which are used to give a solution to the balance equations in [25], thereafter the probabilities p_j, $j = 0, 1, 2, \ldots, \infty$.

Letting σ be the unique root of the equation $\sigma = L_A(\mu - \mu z)$, the solution for the equilibrium state probabilities is given by [27]:

$$p_j = (\sigma - 1)\sigma^j, j = 0, 1, 2, \ldots, \infty. \tag{8}$$

As long as $0 < \sigma < 1$, the number of packets in the system at the embedded arrival instants is a geometric distribution with parameter σ which is the modulated load. The relation between the modulated load σ and ρ is given by Cruz's bounds as $\mathfrak{R}(\rho, \sigma)$ [28] which can also be exploited using the PDF in [29].

2.3 G/G/1 Queue

G/G/1 stands for a general distribution arrival, a general distribution service time, and a single-server with infinite capacity. Earlier solutions of G/G/1 queues yielded bound parameters of the system, obtained through a series of linked solutions.

2.3.1 Bound methods

The bound method is the result of successive bound solutions for G/G/1 queues.

A. Lindley's equations

The foundation is laid down using Lindley's equations in [30], which consider the relation between the waiting time w_n for the n-th packet, and the waiting time for the $(n+1)$-th packet as

$$w_{n+1} = \begin{cases} w_n + x_n - t_{n+1}, & w_n + x_n - t_{n+1} \geq 0 \\ 0, & w_n + x_n - t_{n+1} < 0 \end{cases} \tag{9}$$

with the equilibrium condition $\lim_{n\to\infty} E[u_n] < 0$. Defining $C_n(u)$, the n-th packet at time u as the PDF for u_n, such that

$$C_n(u) = P[u_n = x_n - t_{n+1}], \tag{10}$$

and for the waiting time when $y \geq 0$

$$W_{n+1}(y) = P[w_n + u_n \leq y], \tag{11}$$

which at the equilibrium is such that $W(y) = \lim_{n\to\infty} P[w_n \leq y]$, taking into account certain considerations in [23], as in the following equation

$$W_{n+1}(y) = \int_{w=0^-}^{\infty} C_n(y-w) dW_n(w), y \geq 0. \tag{12}$$

Eq. (11) yielding large n gives rise to Lindley's integral equation with $W(y) = 0$ when $y < 0$.

The integration of the equation, provided that $C(y - w) = 0$ as $w \to \infty$, and $W(0^-) = 0$, reaches the second form of Lindley's equation

$$w(y) = \begin{cases} -\int_{0^-}^{\infty} W(w)dC(y-w), & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (13)$$

letting $u = y - w$ and bearing in mind that $C(u)$ is a distribution, while $c(u)$ is the density function, produces

$$w(y) = \begin{cases} \int_{0^-}^{\infty} W(y-u)dC(u), & y \leq 0 \\ 0, & y < 0. \end{cases} \quad (14)$$

B. Spectral method

The spectral method applied by Liu [30] characterized the waiting time in Eq. (14) using $c(u)$ and the Laplace transform to achieve the spectral solution

$$L\left[W(y) + W^c(y)\right] = L\left[\int_{u-\infty}^{y} W(y-u)c(u)du\right], \quad (15)$$

by letting $L[c(u)] = L[a(-u)]L[b(u)]$ or $C^*(s) = A^*(-s)B^*(s)$, knowing that the probability function $W(y)$ is bound with its density function $w(y)$ in the Laplace domain by $s\phi(s) = L[w(y)] = W^*(s)$.

Furthermore, Liu [30] considered a queuing system for which $A^*(-s)$ and $B^*(s)$ are rational functions in the Laplace domain, yielding $A^*(-s)B^*(s) - 1 = {}_+(s)/\varphi_-(s)$ where

$$\begin{cases} \lim_{|s|\to\infty} \varphi_+(s) = s, & \text{for } \mathcal{R}\{s\} > 0 \\ \lim_{|s|\to\infty} \varphi_-(s) = s, & \text{for } \mathcal{R}\{s\} < 0. \end{cases} \quad (16)$$

Liu [30] further applied the theorem of Liouville [31], requiring a constant for any bounded analytic function within all the finite values of its support,

$$\phi^c(s)\varphi_-(s) = \phi(s)\varphi_+(s) = K, \quad (17)$$

which yields $\phi(s) = K/\varphi_+(s)$, where K is the constant to be determined through

$$s\phi(s) = W^*(s) = \int_{0^-}^{\infty} \exp(-sy)dW(y). \quad (18)$$

Applying the initial conditions, one gets

$$K = \lim_{s\to 0} \frac{\varphi_+(s)}{s}, \quad (19)$$

to yield $W^*(s) = s\phi(s)$.

C. Kingman's bounds

Applying Kingman's bounds, Kobayashi [32] carries on with the result of [30] $K = \lim_{s\to 0} \varphi_+(s)/s$ where $K = W(0) = \left[(1-\rho)/\lambda\right]\varphi_-(0)$ to derive an upper bound on the complementary waiting time $W_n^c(t) = P[w_n > t]$, with W_n being the waiting time of the n-th packet in a busy cycle.

By recursion on w_n and applying Kolmogorov's inequality theory [33], which is the generalized form of Chebyshev's inequality [34], extending it to the notion of martingales and semi-martingale [32] (or sub-martingale) variables using the inequality proposed by Feller [35]. Furthermore, applying the Laplace-Stieltjes transform of $C(u)$, one obtains a tighter complementary upper bound at the steady state for a small n and θ a positive real number, satisfying $\exp(\theta W) = \max(y_0, y_1, \ldots, y_n)$ for large n. The upper bound of the tail of the complementary waiting time is given by:

$$W_n^c \leq \exp(-\theta_0 t). \quad (20)$$

However, these results of queue bounds are better dealt with using the exponential queue, which is achieved using fluid and diffusion solutions.

2.3.2 Fluid solution

The idea behind fluid approximation is that the discrete arrival is sufficiently insignificant to cause congestion in the queue, as the rain drops are for a flood [36]. This concept of a leaky bucket considers $\{Y(t), t \geq 0\}$ to be the continuous time the Markov chain takes values $\{0, 1, 2, \ldots, N\}$ and the infinitesimal generator to be the matrix $Q = \{q_{ij}\}$. When in state j, fluid arrives at the rate a_j, and is served at a constant rate c with a net rate change in the queue of $r_j = a_j - c$. The continuous random variable $X(t)$ is the queue size at time t satisfying $0 \leq X(t) \leq K$, where K is the buffer size with overflow if the probability is given by $G(x) = \lim_{x\to\infty} P[X(t) > x]$ or in a bivariate stochastic process $(X(t), Y(t))$ reaches the equilibrium overflow if and only if:

$$F_j(x) = \lim_{t\to\infty} F_j(x,t) = P[X \leq x, Y = j]. \quad (21)$$

The time evolution of $F_j(x,t)$ is governed by the following equation [8]:

$$F_j(x, t+\Delta t) = \sum_{i,j\neq j} q_{ij}\Delta t F_i(x - r_{ij}\Delta t, t) + \left(1 - \sum_{i,i\neq j} q_{ij}\Delta t\right) F_j(x - r_j\Delta t, t) + O(\Delta t). \quad (22)$$

Knowing that $O(\Delta t)$ goes faster to 0 as Δt tends to 0, subtracting $F_j(x,t)$ from both sides, we get the differential equation:

$$\frac{\partial F_j(x,t)}{\partial t} + r_j \frac{\partial F_j(x,t)}{\partial x} = \sum_i q_{ij} F_j(x,t). \quad (23)$$

Introducing the rate matrix $R = \text{diag}(r_j)$ and the generator matrix $Q = \{q_{ij}\}$, the solution of Eq. (23) at the equilibrium is given by

$$F(x) = C\exp(QR^{-1}), \quad (24)$$

where C is a vector of constants determined by the initial conditions.

2.3.3 Diffusion solution

The diffusion approximation is characterized by the following differential equation [37]

$$dx(t) = \beta dt + z(t)\sqrt{\alpha}dt, \tag{25}$$

where $dx(t)$ represents the incremental changes in the continuous path of the process $x(t)$, β is the drift, α is the variance and $z(t)$ is the white Gaussian processor which can become a Brownian motion.

Then, $x(t)$ is a Brownian motion with drift satisfying the probability:

$$\frac{\partial f(x,t)}{\partial t} = -\beta\frac{\partial f(x,t)}{\partial x} + \frac{\alpha}{2}\frac{\partial^2 f(x,t)}{\partial x^2}. \tag{26}$$

This equation and applications of the diffusion approximation have been discussed by Cox and Miller [14], Newell [38], Gaver and Shelder [39], and Kobayashi [40]. Nonaka and Nogami [37] evaluated the heavy traffic download on the server access operation for Web traffic where the solution was obtained under boundary conditions and discretization using the reflecting barrier at the equilibrium. The result was a growing exponential, leading to the elementary return boundary condition. The most significant work on G/G/1 queue diffusion solution was carried out by Reiser and Kobayashi [41], which gives accurate results for the classical queue model, and Kobayashi and Ren [42, 43], giving rise to exponential bound methods.

Accuracy is provided by Reiser and Kobayashi [41]. By introducing the appropriate boundary conditions, the type 1 solution was obtained in the form of

$$f(x) = \frac{2\alpha}{\beta}\exp\left(-\frac{2\alpha x}{\beta}\right), \tag{27}$$

which, for a small queue size, is the exponential process. A similar result using the Lamperti transform appears in [44], from which a discrete process is then interpreted as a geometrical distribution of the queue size variable n with the same decrement factor $\rho_{es} = \exp(-2\alpha x/\beta)$.

Accordingly, for the G/G/1 queue, the geometrical distribution is written as:

$$p_n = \begin{cases} 1-\rho, & n = 0, \\ \rho(1-\rho_{es})\rho_{es}^n, & n \geq 1. \end{cases} \tag{28}$$

In the previous works, it can be observed that all the queue solutions achieve an exponential server which ultimately yields a geometric distribution. Hence, the use of the beta distribution as an inference tool to size up the queue.

3. PREDICTION METHOD: QUEUE PARAMETERIZATION

3.1 Beta Distribution

The beta function is employed to carry out this operation relying on the computation of conditional probabilities by mixing continuous and discrete random variables. Theoretically, the PDF $f_x(x)$ of a continuous random variable X and probability mass functions $P_N(n)$ of a discrete random variable can yield a joint conditional probability as:

$$f_X(x|n) = \frac{P_{N|X}(n|x)f_X(x)}{P_N(n)}, \tag{29}$$

or

$$P_{N|X}(n|x) = \frac{f_{X|N}(x|n)P_N(n)}{F_X(x)}. \tag{30}$$

The first equation can be written using Bayes' theorem as:

$$f_{N|X}(x|n) = \frac{P[N=n|X=x]P[X=x]}{P[N=n]}, \tag{31}$$

and if the PDF is considered as uniform, it leads to the probability function [45]

$$f_{X|n}(x|n) = \left(\int_0^1 x^n(1-x)^{N-n}dx\right)^{-1} x^n(1-x)^{N-n}, \tag{32}$$

which contains the beta distribution property.

In general, for any α, β positive and random variable X, the beta distribution is given by

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx. \tag{33}$$

This integral has so far been difficult to evaluate. However, theoretically, the gamma distribution can be used as an alternative.

The gamma distribution is given by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \tag{34}$$

which evaluated at the limit using L'Hôspital's rule yields the recurrence $\Gamma(x+1) = x\Gamma(x)$ [19]. As a result, $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. In the following, it can be noted that $a = \alpha, b = \beta$.

3.2 Bayesian inference

As stated earlier, the beta distribution can be used as a prior conjugate of geometric distribution. The previous results for queue modeling suggest that queuing systems can be modeled as a birth-death chain, resulting in a geometric distribution of the queue size or waiting time of the form [46]

$$p_n = (1-\rho)\rho^n, n = 0, 1, 2, \ldots, \infty. \tag{35}$$

The parameter ρ embraces several solutions, from the Markov queue to the diffusion queue, but with the load as a common denominator. This load can be considered as a ratio between the arrival rate and the service rate or as the exponential decrement of the processing proportion. It is therefore important to estimate ρ as a new variable to provide better sizing in the queue.

Taking into account the geometric queue size of previous results, the random variable can be expressed as ρ when the queue is occupied by x packets. The conditional geometric probability can be expressed as

$$P_{X|N}(x|\rho) = \rho^x(1-\rho)^x. \tag{36}$$

Let us anticipate the queue size using the beta distribution prior to the likelihood $P_{X|N}(x|\rho)$. This results in $\rho \sim B(a,b)$, approximation, which can be written as [19]:

$$f(\rho,a,b) = \begin{cases} \dfrac{1}{B(a,b)}\rho^{a-1}(1-\rho)^{b-1}, & 0 \le \rho \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

The predictive density function be expressed as

$$f(x|\rho) = \frac{1}{B(a+x,b+1)}\rho^{a+x-1}(1-\rho)^b. \tag{38}$$

The mean of this distribution is given by

$$E[X] = \frac{a+x}{a+x+b+1}, \tag{39}$$

which can be considered as geometric of mean $1/p_x$.

3.3 Evaluation of the Queue Parameters

To evaluate the value of ρ, it must be remembered that the queue is stable when $0 < \rho < 1$ and that equilibrium is reached when $\rho = 1$. It represents the system load which is defined as the ratio between the arrival rate and service rate.

In the case of an exponential server, there are two parameters to estimate, ρ_{es} and ρ, according to the equation $\rho_{es} = \exp(-\rho t)$ at the limit of the boundaries. According to the predictive probability density function, the following recurrence can be obtained for $x = 0, 1, 2, \ldots$:

$$p_{0|\rho} = \frac{a+b+1}{a}\rho^{a-1}(1-\rho)^b \tag{40}$$

$$p_{1|\rho} = \frac{a+b+2}{a+1}\rho^{a}(1-\rho)^b \tag{41}$$

$$p_{2|\rho} = \frac{a+b+3}{a+2}\rho^{a+1}(1-\rho)^b \tag{42}$$

$$\vdots \quad = \quad \vdots \tag{43}$$

$$p_{n|\rho} = \frac{a+b+n+1}{a+n}\rho^{a+n-1}(1-\rho)^b \tag{44}$$

$$\sum_{0}^{n} P_{k|\rho} = (a+b+1)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}\frac{\rho^{a-1}(1-\rho)^b}{1-\rho} \tag{45}$$

The queue of one unit length is therefore given by:

$$L = (a+b+1)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}\frac{\rho^{a-1}(1-\rho)^b}{1-\rho}. \tag{46}$$

Taking into account the Little theorem in [47] $L = \lambda W$, with W being the waiting time and $\rho = \lambda/\mu$ the load or traffic intensity, these parameters can be obtained as follows:

$$W = \frac{1}{\lambda}(a+b+1)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}\frac{\rho^{a-1}(1-\rho)^b}{1-\rho}. \tag{47}$$

From this waiting time expression, one can state the following condition without proof.

Proposition 1: The necessary and sufficient condition for the queue to be lossless $b = 1$, $a > 1$.

The waiting time and the queue length under this condition become:

$$L = (a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}\rho^{a-1}, \tag{48}$$

$$W = \frac{1}{\lambda}(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}\rho^{a-1}. \tag{49}$$

One can therefore deduce the load or traffic intensity using the length or waiting time as a log probability:

$$\ln\rho = \frac{1}{a-1}\log\frac{L}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}, \tag{50}$$

or

$$\ln\rho = \frac{1}{a-1}\log\frac{W\lambda}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}. \tag{51}$$

The queue parameters resulting from the inference of the different previous waiting times are given in Table 1.

4. SIMULATION AND RESULTS ANALYSIS

The use of the beta distribution under a certain condition yields the parameter ρ of different queuing systems which are given in Table 1 and we make use of two cases from them for simulation. We consider, in this simulation of the theoretical results, the equilibrium $\rho < 1$ and Cruz's bounds are reduced to an exponential server. One must keep in mind $b = 1$ as a necessary condition to avoid packet loss and $a = 2$ since $a > 1$.

The load in the scenario of a predictive queue is investigated using beta distribution and a non-predictive queue since the waiting time elapses for exponential queues. In this test, 1 unit queue length and 1 unit waiting time are considered according to the analysis in Section 2.3.1-C, involving the arrival of 100 and 1000 packets, respectively. The MATLAB outcomes of these theoretical results are shown in Figs. 4 and 5.

These theoretical results are further simulated using two different queues, namely an exponential M/M/1 and M/G/1 with a general diffusion service. Since the general

***Table 1**: Previous queue results and their estimated parameter.*

Queue System	Waiting time (W)	Load estimate (ρ_{es})
M/M/1 Matrix	$\frac{1}{\mu}\frac{\rho}{1-\rho}$	$\frac{1}{a-1}\log\frac{1}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}$
G/M/1 Moment	$\frac{1}{\lambda}\frac{\sigma}{\beta(1-\sigma)}$	
Exponential		$\frac{1}{a-1}\log\frac{1}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}$
Hyper-exponential		$\left\lceil\frac{1}{a-1}\log\frac{\lambda W}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}\right\rceil$
Shifted exponential		$\left\lfloor\frac{1}{a-1}\log\frac{\lambda W}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}\right\rfloor$
G/G/1 Diffusion	$2\frac{\rho_{es}}{\mu(1-\rho_{es})}$	$\rho_{es}=\frac{1}{a-1}\log\frac{\lambda W}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}$
Kingman	$\frac{\alpha Var[\rho]}{2(1-\rho)}$	$\rho=\frac{1}{L}\log(\rho_{es})\frac{1}{a-1}\log\frac{\lambda W}{(a+2)(n+1)\sum_{k=0}^{k=n}\frac{1}{a+k}}$

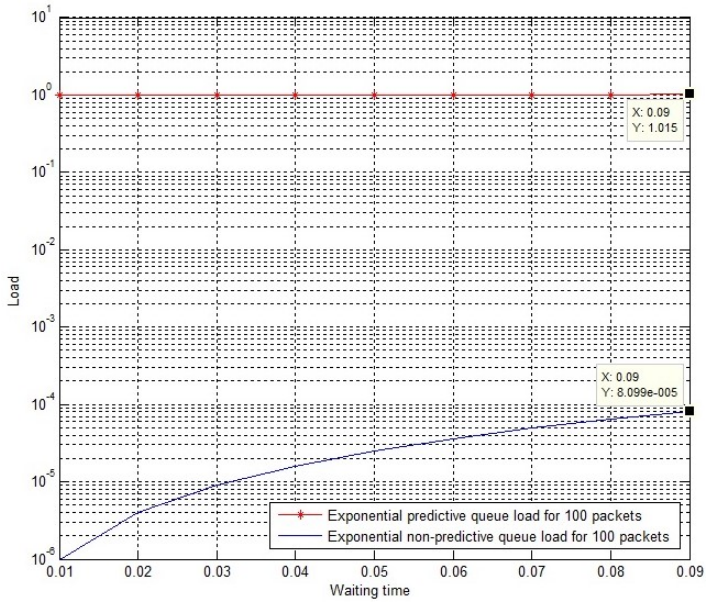

***Fig. 4**: Load of a predictive and non-predictive exponential queue for $n = 100$.*

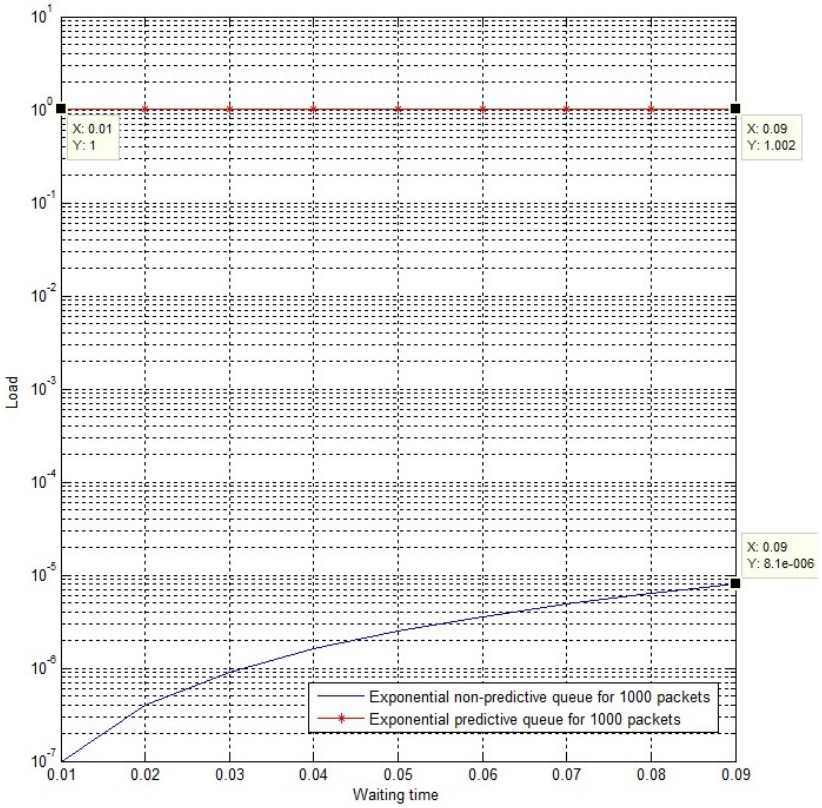

***Fig. 5**: Load of a predictive and non-predictive exponential queue for $n = 1000$.*

service i approximated as a geometric distribution allows the calculation of the decrement factor r_i to analyze the queue occupation from the traffic intensity r_0, the geometric distribution characterizing the queue behavior is accordingly given by: $p(n) = r_i^{n-1}(1 - r_0),\ 0 < r_i < 1$.

4.1 Queue Simulation

This considers a queuing system in which the inputs are made up of two incoming links with rates r_1 and r_2, namely G.711 A as source 1 and ADPCM as source 2, provided both sources can be used for different types of end equipment. The objective of this study is to prevent the queue service process from going beyond its waiting time while transferring the entire traffic.

1. Source 1: G.711 A algorithm, the parameters for this source are as follows:
   - $T_1 = 1$ ms, standard value of packetization delay.
   - $R_1 = 64$ kbits/s, standard bit rate.
   - $N_1 = 30$, (E1 standard) number of channels.
   - $a_1 = 22$, number of active sources.
   - $\beta_1^{-1} = a_1 \cdot T_1$, the mean active time.
   - $\alpha_1^{-1} = 100$ ms, the mean silence time.
2. Source 2: ADPCM algorithm, the parameters of which are as follows:
   - $T_2 = 1$ ms, standard value of packetization delay.
   - $R_2 = 32$ kbits/s, standard bit rate.
   - $N_2 = 24$, (T1 standard) number of channels.
   - $a_2 = 16$, the number of active sources.

***Table 2**: System load and waiting time.*

W ($\times 10^{-2}$ ms)	1	2	3	4	5	6	7	8	9
ρ_{es} ($\times 10^{-6}$), for $L = 100$	1	4	9	16	25	36	49	64	81
ρ_{es} ($\times 10^{-7}$), for $L = 1000$	1	4	9	16	25	36	49	64	81

   - $\beta_2^{-1} = a_2 \cdot T_2$, the mean active time.
   - $\alpha_2^{-1} = \alpha_1^{-1}$, the mean silence time.

Figs. 6 and 7, as well as Table 2, show the results from the queue simulation. Table 2 provides a better view of the predictive queue if it was taken alone in the figures.

4.2 Results

Figs. 4 and 5 provide the results for 100 and 1000 packet arrivals, respectively; the predictive queue serves 100% for 1/100 unit of waiting time, while the non-predictive model for the same duration shows 0.0001% of packets served. At the same waiting time and arrival rate, this difference observed in load is only explained by the better processing rate or departure rate. This means that the predictive queue provides greater processing capability for the same waiting time and queue length than the non-predictive queue.

This theoretical result is confirmed by the simulation results obtained by evaluating the queue occupation

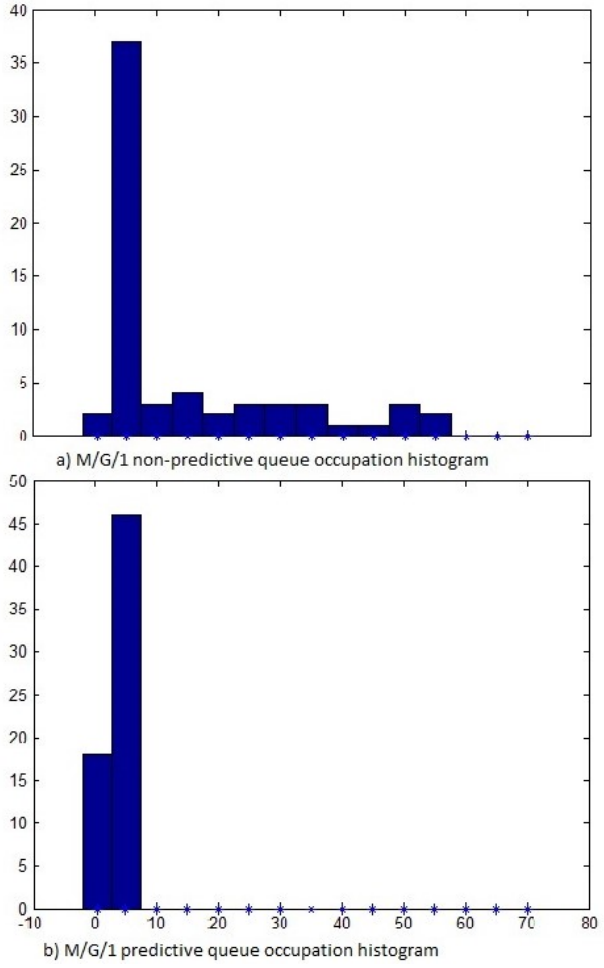

***Fig. 6**: Histogram of the occupation level of non-predictive vs predictive diffusion queue.*

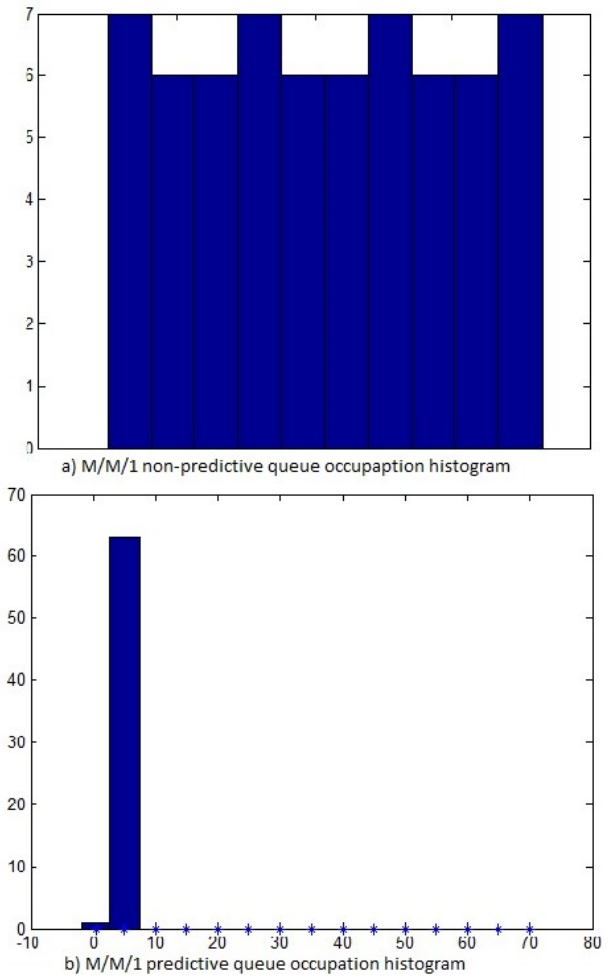

***Fig. 7**: Histogram of the occupation level of non-predictive vs predictive exponential queue.*

against the cumulative transition times according to the histograms presented in Figs. 6 and 7. The predictive queue will need one and a fraction of transition times to throughput or process the entire load, while the non-predictive queue necessitates 13 cumulative transition times to process the same load. According to these figures, the occupation level of the preventive queue is almost ten times more than its non-preventive counterpart, confirming the better service rate of the predictive queue, a characteristic which prevents queue tailing and, most importantly, helps to avoid delay variations, also known as jitter.

As a result, at any given waiting time, queue length, and a constant arrival rate, the predictive model has the provision to serve the offered load on time, preventing a tail from forming as the time elapses, while the non-predictive model does not have this advantage. This statement is confirmed by Table 2, which indicates that the load to be processed by the predictive queue is inversely proportional to the queue length L, and proportional to the quadratic waiting time W. For clarity, the load is defined as $\rho_{es} = W^2 / L$, where W is the waiting time and L the queue length. It can also be said that to overcome time-out issues and consequently packet loss for an efficient, extremely large data transfer, the waiting time W in the queue for a given processing load ρ_{es} and queue length of L, must be at most equal to $W = \sqrt{\rho_{es} L}$. It can also be said that the throughput or the departure rate for a service rate μ must be equal to $\rho_{es}\mu$. It is important to mention that this departure rate or throughput can be either fixed or dynamic.

Some may argue from this analytical result of a dynamic load estimate that even a queue management system could be implemented. Though it is not the focus of this paper, this point of view can be expressed in the following way for First Come First Served/First in First Out (FCFS/FIFO) or non-preemptive queues. The system evaluates the queue occupation at time t, $L(t)$, its potential waiting time $W(t)$, and sets the required processing rate or departure rate (also known as throughput). In this type of dynamic process, some form of queue management such as Short Processing Time First (SPTF) or Processor Sharing (PS) can be abandoned or improved, respectively.

Whether fixed or dynamic, these results stipulate that

one must discard them from the coarse queue model. By applying the same conditions ($b = 1$, $a = 2$) to the results of [19] $\rho(t) = bMt^{b-1}$, the system experiences a constant load $M = 1/2$ regardless of time t, which again yields a coarse parameter, discarding the possibility of capturing the queue behavior as is the case with the solution provided in this study.

5. CONCLUSION

The objective of this paper is to predict a queue model that avoids packet loss through time-out issues caused by tails in the queue. This is achieved in this study through system parameterization using Bayesian inference. The randomness of queue behavior is dealt with by conveniently sizing the queue through its parameters using beta distribution and setting the necessary lossless condition in an assumed loss queue system. The predictive queue is shown to fill up continuously, regardless of the variation in the load, and offers better performance than its non-predictive counterpart based on coarse parameters. This better performance is a result of an improved processing rate, crucial for avoiding tail or time dependency, time-out, and ultimately packet loss in the queue. This characteristic is important for systems where the load and time variations matter, such as "big data storage" communication in a cloud system. However, the implementation of such queue applications is left for further study.

ACKNOWLEDGMENTS

The authors would like to thank B. Twala (previously with the Institute for Intelligent Systems, University of Johannesburg), C. Mbhowa (previously with the Faculty of Engineering and Built Environment, University of Johannesburg), and S. Motala (with the Postgraduate School, University of Johannesburg) for their financial support.

REFERENCES

[1] A. T. Bharucha-Reid, *Elements of the Theory of Markov Processes and Their Applications*. New York, USA: McGraw-Hill, 1960.

[2] I. Adan and J. Resing. (2015). Queueing systems [Online]. Available: https://www.win.tue.nl/~iadan/queueing.pdf

[3] I. Adan and J. Resing, *Queueing Theory*. Eindhoven, The Netherlands: Department of Mathematics and Computing Science, Eindhoven University of Technology, 2001.

[4] J. F. C. Kingman, "On queues in heavy traffic," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 24, no. 2, pp. 383–392, 1962.

[5] H. Heffes and D. Lucantoni, "A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 856–868, Sep. 1986.

[6] W. Whitt, "Approximating a point process by a renewal process, I: Two basic methods," *Operations Research*, vol. 30, no. 1, pp. 125–147, Feb. 1982.

[7] M. Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis*. Reading, Massachusetts, USA: Addison-Wesley, 1987.

[8] M. Luhanga, "A fluid approximation model of an integrated packet voice and data multiplexer," in *IEEE INFOCOM'88, Seventh Annual Joint Conference of the IEEE Computer and Communcations Societies. Networks: Evolution or Revolution?*, 1988, pp. 687–692.

[9] J. M. Harrison and A. Zeevi, "Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime," *Operations Research*, vol. 52, no. 2, pp. 243–257, Apr. 2004.

[10] J. F. C. Kingman, "Some inequalities for the queue GI/G/1," *Biometrika*, vol. 49, no. 3/4, pp. 315–324, Dec. 1962.

[11] K. T. Marshall and R. V. Evans, "Some inequalities in queuing," *Operations Research*, vol. 16, no. 3, pp. 651–668, Jun. 1968.

[12] A. I. Elwalid and D. Mitra, "Statistical multiplexing with loss priorities in rate-based congestion control of high-speed networks," *IEEE Transactions on Communications*, vol. 42, no. 11, pp. 2989–3002, Nov. 1994.

[13] W. Whitt, "Approximating a point process by a renewal process: The view through a queue, an indirect approach," *Management Science*, vol. 27, no. 6, pp. 619–636, Jun. 1981.

[14] D. Cox and H. Miller, *The Theory of Stochastic Processes*. Hoboken, New Jersey, USA: John Wiley & Sons, 1965.

[15] R. B. Cooper, *Introduction to Queueing Theory*. New York, USA: Macmillan, 1972.

[16] A. Kuczura, "The interrupted poisson process as an overflow process," *Bell System Technical Journal*, vol. 52, no. 3, pp. 437–448, Mar. 1973.

[17] B. Wallström, "Congestion studies in telephone systems with overflow facilities," *Ericsson Technics*, vol. 22, no. 3, pp. 190–351, 1967.

[18] R. I. Wilkinson, "Theories for toll traffic engineering in the U. S. A." *Bell System Technical Journal*, vol. 35, no. 2, pp. 421–514, Mar. 1956.

[19] C. Armero and M. J. Bayarri, "Bayesian prediction in M/M/1 queues," *Queueing Systems*, vol. 15, pp. 401–417, Mar. 1994.

[20] F. Ruggeri, M. Wiper, and D. R. Insua. Bayesian model selection for M/G/1 queues [Online]. Available: ftp://ftp.isds.duke.edu/pub/WorkingPapers/97-31.ps

[21] P. Nain. (1998) Basic elements of queuing theory: Application to the modelling of computer systems [Online]. Available: http://zyurvas.narod.ru/knigi/Nain.pdf

[22] F. B. Nilsen, "Queuing systems: Modeling, analysisand simulation," Department of

Informatics, University of Oslo, Norway, Research Report 259, Apr. 1998. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76.353&rep=rep1&type=pdf

[23] S. K. Bose. (2002) The G/M/1, G/G/1, G/G/m and M/G/m/m queues [Online]. Available: http://home.iitk.ac.in/~skb/qbook/Slide_Set_12.PDF

[24] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. Hoboken, New Jersey, USA: Wiley, 1975.

[25] I. Adan, O. Boxma, and D. Perry, "The G/M/1 queue revisited," *Mathematical Methods of Operations Research*, vol. 62, pp. 437–452, Nov. 2005.

[26] D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *The Annals of Mathematical Statistics*, vol. 24, no. 3, pp. 338–354, Sep. 1953.

[27] L. Takács, *Introduction to the Theory of Queues*. New York, USA: Oxford University Press, 1962.

[28] R. Cruz, "A calculus for network delay. I. network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114–131, Jan. 1991.

[29] D. R. Cox, "The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 433–441, Jul. 1955.

[30] J. C. Lui. G/G/1 queueing systems [Online]. Available: http://www.cse.cuhk.edu.hk/~cslui/CSC5420/GG1.pdf

[31] D. Sloughter. (2004) Liouville's theorem [Online]. Available: http://math.furman.edu/~dcs/courses/math39/lectures/lecture-32.pdf

[32] H. Kobayashi, "Bounds for the waiting time in queueing systems," in *Computer Architectures and Networks: Modelling and Evaluation : Proceedings of an International Workshop*, E. Gelenbe and R. Mahl, Eds. Amsterdam, The Netherlands: North-Holland, 1974, pp. 263–274.

[33] S. L. Brumelle, "Some inequalities for parallel-server queues," *Operations Research*, vol. 19, no. 2, pp. 402–413, Apr. 1971.

[34] M. M. Marjanović and Z. Kadelburg, "A proof of Chebyshev's inequality," *The Teaching of Mathematics*, vol. 10, no. 2, pp. 107–108, 2007.

[35] W. Feller, *An Introduction to Probability Theory and Its Applications*. Hoboken, New Jersey, USA: John Wiley & Sons, 1957, vol. 1.

[36] H. Michiel and K. Laevens, "Teletraffic engineering in a broad-band era," *Proceedings of the IEEE*, vol. 85, no. 12, pp. 2007–2033, Dec. 1997.

[37] Y. Nonaka and S. Nogami, "Evaluation of diffusion approximation for the G/G/1 queuing model," in *8th Asia-Pacific Symposium on Information and Telecommunication Technologies*, 2010.

[38] G. F. Newell, *Applications of Queueing Theory*. London, U.K.: Chapman & Hall, 1971.

[39] D. P. Gaver and G. S. Shedler, "Approximate models for processor utilization in multiprogrammed computer systems," *SIAM Journal on Computing*, vol. 2, no. 3, pp. 183–192, 1973.

[40] H. Kobayashi, "Application of the diffusion approximation to queueing networks I: Equilibrium queue distributions," *Journal of the ACM*, vol. 21, no. 2, pp. 316–328, Apr. 1974.

[41] M. Reiser and H. Kobayashi, "Accuracy of the diffusion approximation for some queuing systems," *IBM Journal of Research and Development*, vol. 18, no. 2, pp. 110–124, Mar. 1974.

[42] H. Kobayashi and Q. Ren, "A diffusion approximation analysis of an ATM statistical multiplexer with multiple types of traffic. - I. equilibrium state solutions," in *Proceedings of ICC'93 - IEEE International Conference on Communications*, Geneva, Switzerland, 1993, pp. 1047–1053.

[43] Q. Ren and H. Kobayashi, "Diffusion approximation modeling for Markov modulated bursty traffic and its applications to bandwidth allocation in ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 679–691, Jun. 1998.

[44] S. Nlend, "Optimization of resources allocation for H.323 endpoints and terminals over VoIP networks," M.S. Thesis, Department of Electrical and Electronic Engineering Science, University of Johannesburg, South Africa, 2013.

[45] A. Mahanta. Beta distribution [Lecture notes].

[46] C. Piech. (2018). Beta distribution [Online]. Available: https://web.stanford.edu/class/archive/cs/cs109/cs109.1192/lectureNotes/17-Beta.pdf

[47] J. D. C. Little, "A proof for the queuing formula: $L = \lambda W$," *Operations Research*, vol. 9, no. 3, pp. 383–387, May 1961.

Samuel Nlend received his B.Sc. (Physics) degree in 1994 and M.Phil. (Electrical and Electronic) degree in 2013. Currently he is a Ph.D. candidate and assistant lecturer in the Department of Electrical and Electronic Engineering Science at the University of Johannesburg, South Africa.He worked from 1996 to 2010 respectively as the Regional Head of Department of Statistics and QoS and of Technical Department at Cameroon Telecommunications, Cameroon, and as a Technical Engineer at Phonelines International, South Africa. He is a member of the Engineering Council of South Africa since 2017 and a student member of IEEE since 2019.

Theo G. Swart received his B.Eng. and M.Eng. degrees (both cum laude) in electrical and electronic engineering from the Rand Afrikaans University, South Africa, in 1999 and 2001, respectively, and D.Eng. degree from the University of Johannesburg, South Africa, in 2006. He is an associate professor in the Department of Electrical and Electronic Engineering Science and a member of the UJ Center for Telecommunications. His research interests include digital communications, error-correction coding, constrained coding and power-line communications. He is a senior member of the IEEE, and was previously the chair of the IEEE South Africa Chapter on Information Theory. He is a specialist editor for the SAIEE Africa Research Journal.

Bhekisipho Twala is the Deputy Vice-Chancellor of the University of Pretoria. Before then, he was the Executive Dean of Engineering and Built Environment and Professor in Artificial Intelligence and Data Science at the Durban University of Technology, South Africa, Director of the School of Engineering at the University of South Africa and founder of the Institute for Intelligent Systems at the University of Johannesburg. He completed his Ph.D. at the Open University, UK in 2005, and was a post-doctoral researcher at Brunel University in the UK, mainly focussing on empirical software engineering. Before then, he did his M.Sc. (Statistics) from Southampton University, UK and a B.A. (Economics & Statistics) at the University of Swaziland in 1992. His current research includes promoting and conducting research in Artificial Intelligence within the Big Data Analytics field and developing novel and innovative solutions to key research problems in this area. He has a world-class track record of high-quality research and scholarship as evidenced by 180 publications in internationally leading journals and conferences. He is currently an associate editor of the Information Sciences Journal, Intelligent Data Analysis Journal, Journal of Computers, International Journal of Advanced Information Science and Technology, International Journal of Big Data Intelligence, Journal of Image and Data Fusion, Journal of Information Processing Systems, and a fellow of the Royal Statistical Society. Other professional memberships include the Association of Computing Machinery (ACM); the Chartered Institute of Logistics and Transport (CIT), South Africa and a Senior Member of the Institute of Electrical and Electronics Engineers (SMIEEE).