Access Convergence for Heavy Load Markov Ethernet Bursty Traffic Using Two-level Statistical Multiplexing

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

ABSTRACT

A method for modeling aggregated heavy Markov bursty Ethernet traffic from different sources is proposed in this paper, particularly that prevailing between gateway services and internet routing devices, with the aim of achieving rate accommodation. In other words, to accommodate different rates while filtering out delays in the queue, to achieve access network convergence. Although gateway functions can be used to achieve this by adapting service rates, as many gateways as services are required. Instead of considering the distributed gateway services method, statistical multiplexing is chosen for this study for cost efficiency in network resources. Unfortunately, statistical multiplexing exhibits greater packet variation (jitter) and transfer delay. These delays, basically resulting from positive correlations or time dependency in the queue system, are addressed through infinitesimal queue modeling, based on the diffusion process approximated by Ornstein-Uhlenbeck, which deals with infinitesimal changes in the Markov queue. The related analysis has resulted in an exponential queueing model for univariate and/or multivariate servers obtained through Markov Gaussian approximation. An experiment based on two different voice algorithms shows rate accommodation, and a fluid solution, which is dynamically outputted according to the transmission link availability during each transition time, without any significant delay. Hence, better transfer delay and rate control is obtained through the proposed two multiplexing levels within an Ethernet LAN.

Keywords: Statistical Multiplexing, Diffusion Model, Markov Gaussian Approximation, Gateway Distributed Function, Arrival Process, Departure Process, Access Convergence

1. INTRODUCTION

Nowadays, all network services are run as applications on top of TCP/IP protocol suites. However, the migration of real-time protocol (RTP) services, such as voice over the internet and live TV streaming, does not happen without challenges, particularly when it comes to providing an acceptable and commercial level of services as in the old legacy networks to end-users using a single platform. The distributed gateway service through rate adaptation offers a solution for providing all services on one platform. In that context, ITU-T, IEEE, and IETF have released a series of recommendations and protocols under several umbrellas (e.g., H.323, SAP, SIP) to guarantee the required “quality of service (QoS)” for distributed gateway services within a LAN, without a guarantee like for Ethernet.

Despite these provisions, the network continues to experience a low QoS index [1, 2] due to impairments, all of which basically result from packet loss, jitter, and transfer delay. In this regard, the study conducted by Sriram and Whitt [3] who characterize the aggregate arrival process from a voice source in the multiplexer, shows that the positive correlation or time dependency in the queue system is the cause of the significant delay at the multiplexer output, thereby reducing the efficient utilization of network resources.

To optimize resource utilization where smooth rate adaptation matters, two-level multiplexing is proposed in this paper. Firstly, a level 1 multiplexer is used, based on the M/G/1 queueing model, to accommodate the bursty traffic from different sources, emitted at different rates. Secondly, a level 2 multiplexer based on the G/G/1 queue.

The contribution of this paper is therefore to achieve access network convergence by replacing the costly distributed gateway services approach with two-level statistical multiplexing, modeled to achieve rate accommodation, rather than the rate adaptation solution which requires as many gateways as the potential services provided.

In a packet switching system, some nodes are usually encountered where the arriving packets are processed (routing analysis, rate adaptation, multiplexing) before continuing on their way. When the incoming traffic is greater than the processing capacity, the solution is to buffer the arriving packets accordingly and process them in compliance with the queue order. This is called statistical multiplexing. Statistical multiplexers

Manuscript received on July 17, 2021; revised on September 30, 2021; accepted on February 13, 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.2022203.247512

are therefore modeled as queueing systems with finite buffer space for the incoming traffic, served (on a service discipline basis) by one or more transmission links of fixed or varying capacity [4].

Two key concepts arise from this definition of the multiplexer, namely traffic and buffers. Traffic refers to teletraffic engineering and buffer refers to queueing theory. To achieve multiplexer, traffic, and queue models need to be developed, as demonstrated in this paper.

This paper is organized as follows: Section 2 presents the traffic and queue models. Section 3 explains the diffusion approximation process with the resulting Gaussian approximations, while the relevant multiplexing levels are presented in Section 4. Section 5 presents details of the simulations carried out along with results allowing the sizing of such buffers.

2. BACKGROUND

2.1 Traffic Modeling

Modeling the traffic involves characterizing the input of the multiplexer, also known as the arrival process. The arrival process analysis relies on the way the packets are emitted and the statistical properties of the inter-arrival times. The packet emission can be periodic, leading to a deterministic analysis; or continuous-time analyzed using Bernoulli or Poisson processes.

The statistical properties of transition epochs can include the "continuous-time Markov chain (CTMC)" [5], which can be simulated by a "Markov modulated Poisson process"; its discrete form being the "discrete-time Markov chain (DTMC)" [6] or the hybrid form classified as the "semi-Markov process (SMP)" family.

Many studies have been conducted in this regard. Frost and Melamed [7] characterized the arrival process as samples of packets arriving in a sequence of random arrival times associated with a random workload with a renewal process proposed as a solution. Gusella [8] studied Ethernet traffic, showing it to be non-stationary and characterized by a long-tailed inter-arrival distribution, while Leland *et al.* [9] studied Ethernet traffic scaled over several times, proposing a self-similar process as the model to overcome the tail. This paper considers overcoming the long-tailed queue through an infinitesimal analysis based on the diffusion process.

2.2 Queue Modeling

Queue modeling is a group of absolute QoS-defining network parameters that characterize bandwidth usage, transfer delay, and delay variation. The queue is modeled according to the specification $A/S/N/C/D$ describing the system, which stands for Arrival/Service/Number of servers/System Capacity/Discipline. This study focuses on the specification with one server linked to a system capacity C, resulting in $A/S/1/C$. Furthermore, an $A/S/1$ queue model is designed for which the system capacity C (queue content+server) is chosen such that the system is always stable or simply, a queue without loss.

To determine A and S, at least one method of queue solution is required. The most used are the matrix, moment generating function, and fluid methods. In this paper the diffusion approximation method is proposed, a limit of the fluid method, to model the M/G/1 and G/G/1 queues, where the service process is considered as general distribution for both queues and the arrival process as the Markov or general process, respectively.

Modeling M/G/1 or G/G/1 queues resort to approximations and many studies have been conducted in this field. The embedded Markov chain as the underlying process proposed by Kendall [10] is approximated by the moment generating function based on Laplace-Stieltjes transform [11]. This approximating solution follows Cruz bounds, commonly denoted by $R(\rho, \sigma)$ with ρ as the traffic intensity and σ as the unique root of the Laplace-Stieltjes transform resulting equation, and the probability density function from the indirect method of Whitt [12]. Lui [13] proposed a spectral analysis approximated from Lindley's equations, applying the theorem of Liouville [14] to derive a Poisson queue. Kingman [15] provides a more tractable method known as Kingman's bounds on the tail probability, based on Lindley's equations to derive the waiting time bounds in the queue. Kobayashi [16] extended Lui's solution [13] by applying Kolmogorov's inequality [17], which is the general form of Chebychev's inequality [18], the notion of martingale and sub-martingale variables [15] extended to the inequality of Feller [19], who proposed exponential waiting time bounds as a solution. Other exponential bound solutions come from Kobayashi and Ren for on-off sources [20] and Markov modulated rate processes [21] using univariate and multivariate Ornstein-Uhlenbeck diffusion approximations, respectively.

The most tractable exponential solutions were developed by Reiser and Kobayashi [22], who applied the central limit theorem in the diffusion equation to a small number of sources, and Schwartz [23] and Luhanga [24] who used the fluid solution for two types of traffic. They all achieved an exponential server as the solution.

Based on the Ornstein-Uhlenbeck (O-U) diffusion approximations and tractable solutions of the exponential server, here the following simpler approach of a queueing model is proposed from a synthesis of the works in [21–24], for two-level statistical multiplexing to accommodate services from multiple access networks in the Ethernet LAN, while nullifying the tail probability in the queue.

3. RATE ACCOMMODATION MODEL: DIFFUSION APPROXIMATION

The solution in this paper is from the normalized asymptotic approach of Schwartz [23] and Luhanga [24] as the sample size increases, and the diffusion method of Reiser and Kobayashi [22] as it converges in probability to the value being estimated. The maximum likelihood estimator is then chosen for that purpose and its central limit property applied to model the queues.

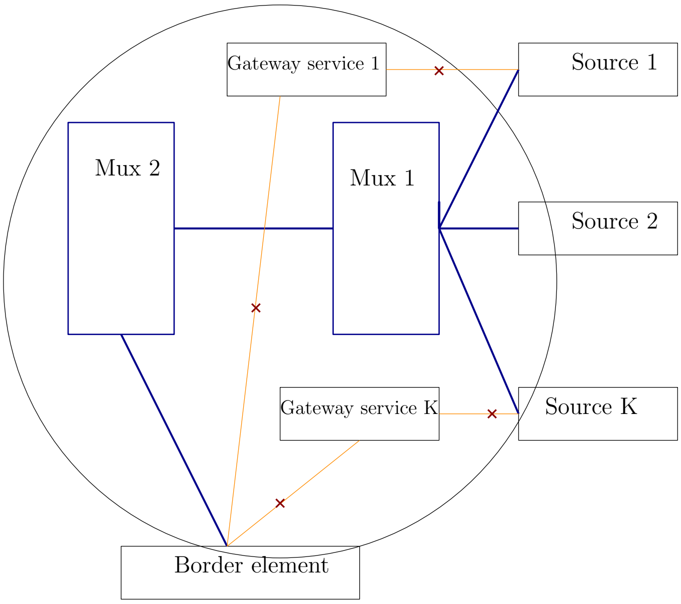


Fig. 1: Two-level multiplexing within a core network.

Two-level multiplexing within the Ethernet LAN is configured as shown in Fig. 1, where the gateway network services are replaced by the statistical multiplexer which outputs the required traffic.

3.1 Level 1 Statistical Multiplexing

Level 1 multiplexing concerns the sources operating with on-off signaling.

3.1.1 Traffic Model

The level 1 multiplexer (Mux 1) handles a sample of each active on-off source, each with its own emission rate. The Gaussian distribution is derived, with the resulting queue being MG/G/1, denoting the Markov Gaussian arrival (MG), general service time (G), and one server. The traffic in the multiplexer is in a super position for packet streams from many on-off sources. Since some sources are off when others are on, the input of the multiplexer queue, as well as the queue size, are variable. Markov analysis can only be conducted using the bivariate (active sources-queue size) process.

Let N_k be the number of sources of type k with a rate of R_k packets/second, and A_k the number of active sources with mean time α_k^{-1} ms, or $N_k - A_k$ the number of off sources with a mean silence time of β_k^{-1} ms. The multiplexer input is given by $\sum_{k=1}^{N} R_k A_k$, with traffic intensity of $(1/C)\sum_{k=1}^{N} R_k A_k$ where C is the system capacity.

3.1.2 Queue Model

To serve the A_k sources, $Q(t)$ is defined as the queue size of the buffer at time t with C packets/second being constant capacity of the transmission link. By definition, the statistical multiplexer allocates the capacity lying between the average and peak rates, buffering the traffic when the load exceeds capacity. Therefore, the changes in the multiplexer can be captured by the differential equation

$$\frac{dQ}{dt} = \begin{cases} R(t) - C, & Q(t) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let us consider the bivariate process $(A(t), Q(t))$ as previously mentioned, such that $(A(t), Q(t)) = \left\{A_k, 1 \le k \le K; Q(t)\right\}$ is a Markov process. One can define the underlying probability function of the process as $P(A, x, t) = Pr[Q(t) \le x]$, satisfying the stochastic differential equation in Eq. (2).

$$\frac{\partial p\left(A_k, x, t\right)}{\partial t} + \left(\sum_{k=1}^{K} R_k A_k(t) - C\right) \frac{\partial p\left(A_k, x, t\right)}{\partial x} = \sum_{k=1}^{K} \left[\left(N_k - A_k\right)\beta_k + A_k\alpha_k\right] p\left(A_k, x, t\right) + \sum_{k=1}^{K} \left(N_k - A_k + 1\right)\beta_k p\left(A_k - 1_k, x, t\right)$$
$$+ \sum_{k=1}^{K} \left(A_k + 1_k\right)\alpha_k p\left(A_k + 1_k, x, t\right) \tag{2}$$

As $x \to \infty$, Eq. (2) becomes a Markov birth-death process which is known to result in a Bernoulli solution

$$P(A,t) = \lim_{k\to\infty} P(A,x,t) = \prod_{k=1}^{K} P(A_k,t), \tag{3}$$

where

$$P(A_k,t) = \binom{N_k}{A_k} q_k^{A_k}(t)[1-q_k(t)]^{N_k - A_k} \tag{4}$$

also known to be a Gaussian distribution method for large numbers. Thus, the Gaussian parameters are obtained by transforming the on-off sources; a two-state Markov chain, into a Gaussian process using univariate diffusion approximation.

3.1.3 Univariate Diffusion Approximation

Let $(Y(t), Q(t))$ be the diffusion approximation of $(A(t), Q(t))$ with its probability function given by $f(y,x,t) = Pr\left[y_k \le y \le y_k + dy_k; Q(t) \le x\right]$. The resulting differential equation in the second order Taylor series representation is given by Eq. (5) [19].

Using an analogy to the order of Eqs. (2) and (5), the infinitesimal statistical properties of the process can be defined; namely the infinitesimal mean $m_k = N_k\beta_k - (\alpha_k+\beta_k)y_k$ and the infinitesimal variance by $v_k = N_k\beta_k - (\alpha_k - \beta_k)y_k$.

Considering only the infinitesimal arrival process of mean zero, gives $f(y,t) = \lim_{x\to\infty} f(y,x,t)$ and $y^* = N_k\beta_k / \left(\alpha_k + \beta_k\right)$. Therefore, the mean and the variance become $m_k = -(\alpha_k+\beta_k)(y-y_k^*)$, and $v_k(y_k^*) = 2N_k\beta_k\alpha_k / \left(\alpha_k + \beta_k\right)$. Diffusion equation in Eq. (5) taken for an individual source type can now be written as Eq. (6).

The diffusion process as characterized is called the Ornstein-Uhlenbeck (O-U) [19]. At the equilibrium state ($t \to \infty$), the approximated probability density function solution of this equation at the reflecting boundaries $y = 0$ and $y = N_k$, yields a Gaussian distribution defined by Eq. (7) where $\sigma_k^2 = v_k(y_k^*) / 2(\alpha_k + \beta_k)$, and is maximum if $y = y_k^*$.

3.2 Queue Solution Approximation

The changes in the queue ΔQ are given by the O-U differential equation and an analogy with the method of Reiser and Kobayashi [22], giving $t \to \infty$ Eq. (8) is obtained.

We can therefore deduce from the diffusion equation (Eq. 9)

$$\begin{aligned}\frac{\partial f\,(y,x,t)}{\partial t} + \sum_{k=1}^{k=K}(R_kY_k - C)\frac{\partial f\,(y,x,t)}{\partial x} &= -\sum_{k=1}^{K}\left[\left(N_k - y_k\right)\beta_k + y_k\alpha_k\right] f\left(y_k,x,t\right) \\ &\quad + \sum_{k=1}^{K}\left(N_k - y_k + 1_k\right)\beta_k\left[f\,(y,x,t) - \frac{\partial f\,(y,x,t)}{\partial x}\right] \\ &\quad + \frac{1}{2}\sum_{k=1}^{k}\left(N_k - y_k\right)\beta_k\frac{\partial^2 f\,(y,x,t)}{\partial x^2} + \sum_{k=1}^{K}\left(y_k + 1_k\right)\alpha_k\left[f\,(y,x,t) + \frac{\partial f\,(y,x,t)}{\partial x}\right] \\ &\quad + \frac{1}{2}\sum_{k=1}^{K} y_k\alpha_k\frac{\partial^2 f(y,x,t)}{\partial x^2} \\ &= -\sum_{k=1}^{K}\frac{\partial}{\partial x}[N_k\beta_k - (\alpha_k+\beta_k)f(y,x,t)] + \frac{1}{2}\sum_{k=1}^{K}\frac{\partial^2}{\partial x^2}[(N_k\beta_k - (\beta_k - \alpha_k)y_k)f(y,x,t)]\end{aligned} \tag{5}$$

$$\frac{\partial f_k(y,t)}{\partial t} = (\alpha_k+\beta_k)\frac{\partial}{\partial y}\left[(y-y_k^*)f_k(y,t)\right] + \frac{N_k\alpha_k\beta_k}{\alpha_k+\beta_k}\frac{\partial^2 f_k(y,t)}{\partial y^2} \tag{6}$$

$$\lim_{t\to\infty} f(y,t) = f(y) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left\{-\frac{(y-y_k^*)}{2\sigma_k^2}\right\} \tag{7}$$

$$\left(\sum_{k=1}^{K}\left(R_ky_k - C\right)\right)\frac{\partial f(y,x)}{\partial x} = \sum_{k=1}^{K}\left[\left(\alpha_k+\beta_k\right)\frac{\partial\left(y_k - y_k^*\right)f(y,x)}{\partial x} + \left(\frac{N_k\alpha_k\beta_k}{\alpha_k+\beta_k}\right)\frac{\partial^2 f(y,x)}{\partial x^2}\right] \tag{8}$$

$$\frac{\partial f(x,t)}{\partial t} = -\beta\frac{\partial f(x,t)}{\partial x} + \frac{\alpha}{2}\frac{\partial^2 f(x,t)}{\partial x^2}, \tag{9}$$

the mean queue length by $E[Q] = \beta = (\widehat{R}_k - C)\Delta t$, where $\widehat{R}_k - C$ is the drift of the diffusion process representing the variation of the mean and C the capacity transmission link, and the variance in the queue is given by

$$\mathrm{Var}[Q] = \alpha\Delta t = \sum_{k=1}^{K}\frac{2N_k\alpha_k\beta_k}{\alpha_k+\beta_k}\Delta t, \tag{10}$$

with

$$\alpha = \sum_{k=1}^{K}\frac{2N_k\alpha_k\beta_k}{\alpha_k+\beta_k}, \tag{11}$$

as the volatility in the variance of the diffusion process.

In [22], these two parameters are shown to be accurately taken as decrement factors denoted as $\hat{\rho} = \exp(-2\beta/\alpha)$, where β is the drift and α is the volatility of the process. Therefore, the probability distribution of the queue size is geometric and given by

$$\widehat{p_n} = \rho(1-\hat{\rho})\hat{\rho}^{n-1}, \quad n \geq 1, \tag{12}$$

where $\rho = \widehat{R}_k \,/\, C < 1$ is the traffic intensity.

3.3 Level 2 Multiplexing

3.3.1 Traffic Model

Multiplexing the traffic from the nodes in level 1 serving different access networks, leads to the finding of a parametric traffic model suitable for the most used link layer technologies in the LAN networks without a well-defined aggregated traffic model, namely Ethernet. In this paper, these nodes are modeled as “Markov modulated rate (MMR)” sources. Since the traffic at the multiplexer input is of multiple types, it is reasonable to re-scale the aggregated traffic using “carrier sense multiple access with collision detection (CSMA/CD)”, which governs Ethernet access.

The traffic is modeled according to the behavior of superposed K sources from the level 1 multiplexer, each one governed by an M-state Markov chain with the probability transition matrix $P = \{p_{ij}\}$, where $i, j = 0, 1, 2, \ldots, M-1$.

When a source is in state i, it generates packets at R_i packets/second. After a holding time, generally distributed with mean α_i^{-1} ms and variance σ_i^2, it leaves from state i to state j with transition probability p_{ij}. The rate accommodation implies the state transition of the $M \times K$ MMR sources by a closed queueing network with M servers and a total of K customers. Each source in state $i = (m,k)$ or node i, is considered as a customer k served by one of the $m = 0, 1, \ldots, M-1$ parallel servers. Thus, the arrival process at node i, is the sum of these departures from the level 1 multiplexer routing their traffic to that node, while the level 2 multiplexer input is the sum of the departures from all $i = (m,k)$ nodes routing their traffic during the holding time. The formulation of the process is summarized in Fig. 2 with the superposed traffic at the multiplexer input given by $R(t) = \sum_{i=0}^{I-1} R_i N_i(t)$, $I = 1, 2, \ldots, M \times K$.

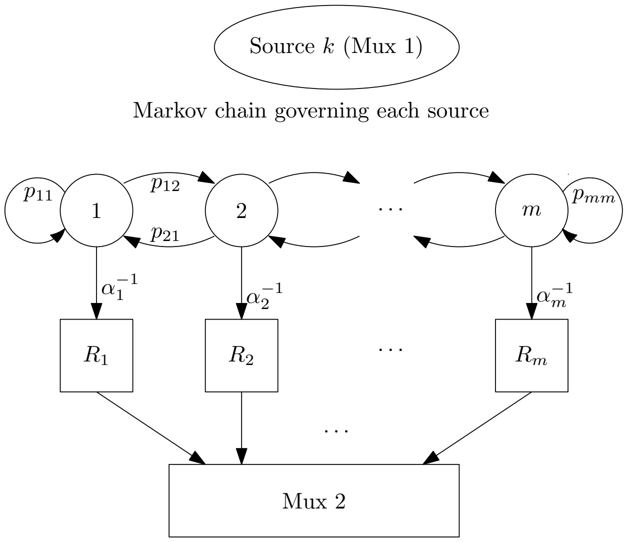


Fig. 2: *Level 2 multiplexer*

3.3.2 Queue Model and Diffusion Approximation

The diffusion model is formulated using a system composed of a statistical multiplexer and K independent sources. Each source is characterized by an M-state Markov chain (k-type sources in level 1), suggesting a $K \times M$ dimensional process. However, before going through the $K \times M$ analysis, let us first investigate the process used to determine the number of sources.

Let $N(t)$ be that process, defined by the transposed matrix $N(t) = [N_0(t), N_1(t), \ldots, N_{K-1}(t)]^{\mathrm{T}}$, where $N_k(t)$ is the number of sources served by server m at time t with rate R_m packets/second. The Markov chain state m is reached after leaving state j with the transition probability P_{jm}. This state has a holding time for the mean of α_m^{-1} and σ_m^2 for the variance. The mean departure rate from node j can be given by $\alpha_j N_{jk}$ while the counting arrival process at node k is equal to the aggregation of the departures from node j.

Given $N(t)$ and according to the model previously described, its diffusion approximation can be defined as $X(t) = [X_0(t), X_1(t), \ldots, X_{k-1}]$, which according to the asymptotic study by Halfin and Whitt [25], $X(t)$ follows the stochastic differential equation

$$dX(t) = b(X(t))dt + DdW(t), \tag{13}$$

$$\sum_k X_k(t) = M, \tag{14}$$

$$X(0) = x_0, \tag{15}$$

where $W(t)$ is the Brownian motion and $b(x)$ the infinitesimal mean, highlighting the drift in the process compared to the long-term equilibrium. The higher the $b(x)$, the faster the speed of the drift, which is the variation of the mean. $\boldsymbol{D}$ represents the randomness of the process, outlining its volatility. The higher the value of $\boldsymbol{D}$, the larger the magnitude of system volatility or variance.

From the differential stochastic equation, the conditional probability density function can be $f(x, t : x_0, 0) = Pr[x_k \leq X_k(t) \leq x_k + dx_k : X(0) = 0]$ satisfying the multidimensional differential equation

$$\frac{\partial f(x,t)}{\partial t} = -\sum_{i=0}^{K-1} \left[b_i(x) f(x,t)\right] + \sum_{i=0}^{K-1}\sum_{j=0}^{K-1} \frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}\left[a_{ij}(x) f(x,t)\right], \tag{16}$$

letting the multivariate mean and covariance to be $b_i(x)$ and $a_{ij}(x)$ respectively, where $b(x) = \boldsymbol{B}\mathbf{x}$ is the infinitesimal mean matrix with $B = [\beta_{ij}]$ and $\mathbf{x} = [x_0, x_1, \ldots, x_{K-1}]$ associated with $X(t)$. $A(x)$ is the infinitesimal covariance matrix such that $D = \sqrt{A(x)}$ with $A = [a_{ij}]$. The differential equation in Eq. (16) becomes [25]:

$$dX(t) = BX(t)dt + \sqrt{A(x)}dW(t), \tag{17}$$

with $\sum_{k=1}^{K} X_k(t) = K$, for which Gaussian approximation solution can be obtained asymptotically.

3.4 Asymptotic Diffusion Approximation

Let $x^* = [x_0^*, x_1^*, \ldots, x_{K-1}^*]$ be the asymptotic equilibrium state of the process $X(t)$, such that the reverting process is stable with drift $b(x) = B(x^* - \mathbf{x})$. In this particularly narrow region, the infinitesimal covariance is constant. At the equilibrium state, we get $dX(t) = B(x^* - X(t))dt + \sqrt{A}dW(t)$. This equation represents the multivariate Ornstein-Uhlenbeck process, satisfying the probability density function differential equation [21]

$$\frac{\partial f(x,t)}{\partial t} = -\sum_{i=0}^{K-1}\sum_{j=0}^{K-1} \beta_{ij}\frac{\partial}{\partial x_i}(x_i - x_i^*) f(x,t) + \sum_{i=0}^{K-1}\sum_{j=0}^{K-1} a_{ij}\frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}\left[a_{ij}(x) f(x,t)\right], \tag{18}$$

where β_{ij} and a_{ij} are the (i, j)-th entries of $K \times M$ matrices.

To obtain the maximum likelihood estimator, the transition log-likelihood proposed by Aït-Sahalia [26] is introduced. This approximation is based on Hermite series expansion and the change of variable in the Jacobian form. The strategy is designed to determine the Hermite series expansion of the transition function P_y, which is a normal distribution $N(0, 1)$, representing a reduced form of the transition function p_x derived from the diffusion process equation $dX_i = \mu(X_i)dt + \sigma(X_i)dW_i$.

The Hermite series approximation is in the form

$$P_y^j(y/y_0, \Delta) = \Delta^{-\frac{m}{2}} \phi\left(\frac{y - y_0}{\sqrt{\Delta}}\right) \times \sum_{t|h| \leq j} \eta_h(\Delta y_0) H_h\left(\frac{y - y_0}{\sqrt{\Delta}}\right) \tag{19}$$

where Δ is the sampling interval, $\phi(x)$ is the density of a normal distribution with a zero mean and identity covariance matrix, taking into account the sum of arrival distribution in zero mean while variance σ^2. H_h represents the Hermite polynomials associated with vector $h = [h_1, h_2, \ldots, h_m]^{\mathrm{T}}$. $\eta(\Delta, y_0)$ are the Hermite coefficients arising from the expansion in Δ and given by

$$\eta(\Delta, y_0) = \frac{1}{h_1! \cdots h_m!} E\left[H_h\left(\frac{y - y_0}{\Delta^{1/2}}\right) / Y_t = y_0\right], \tag{20}$$

where the expectation entity is evaluated by the Taylor expansion and given by

$$E\left[f(Y_\Delta, Y_0, \Delta)|Y_0\right] = \sum_k \frac{\Delta^k}{k!} A_y^k f(y, y_0, \psi\delta)\Big|_{y=y_0 \delta=0} + O(\Delta^{k+1}) \tag{21}$$

where A_y^k is the infinitesimal generator of the process Y, which by applying to f yields the solution of the diffusion differential equation [25]. The log-transition, for any given j where the convergence of the Hermite polynomials is verified as $j \to \infty$, the resulting log-expansion takes the form

$$l_y^k(y|y_0, \Delta) = -\frac{m}{2}\ln(2\pi) + \frac{C_y^k(y|y_0)}{\Delta} + \sum_{k=0}^{k=K} C_y^k(y|y_0)\frac{A^k}{k!}, \tag{22}$$

whose coefficients C_y^k for $k = -1, 0, 1, 2, \ldots, K$ are combinations of the coefficients identified in the Hermite series approximation and $l_v = \ln p$.

The coefficients are determined from the series expansion, satisfying Kolmogorov's equations describing the evolution of the process. Consider the following Kolmogorov equation [27]

$$\frac{\partial p_y(y|y_0, \Delta)}{\partial \Delta} = -\sum_{i=1}^{m} \frac{\partial[\mu_i(y) p_y(y|y_0, \Delta)]}{\partial y_i} + \frac{1}{2}\sum_{i=1}^{m} \frac{\partial^2 p_y(y|y_0, \Delta)}{\partial y_i^2}, \tag{23}$$

from which the equivalent form for the log-likelihood is given by Eq. (24).

Exploiting the log function in Eq. (22) and substituting it for the log-likelihood function in Eq. (24), we get

$$\frac{\partial l_y(y|y_0,\Delta)}{\partial} = -\sum_{i=1}^{m}\frac{\partial \mu y_i(y)}{\partial y_i} - \sum_{i=1}^{m}\mu_{y_i}(y)\frac{\partial l_y(y|y_0,\Delta)}{\partial y_i} + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2 l_y(y|y_0,\Delta)}{\partial y_i^2} + \frac{1}{2}\sum_{i=1}^{m}\left(\frac{\partial l_y(y|y_0,\Delta)}{\partial y_i}\right)^2 \tag{24}$$

$$\frac{\partial l_y^{(k)}(y|y_0,\Delta)}{\partial} = -\frac{C_y^{(-1)}(y|y_0)}{\Delta^2} - \frac{m}{2\Delta} + \sum_{k=1}^{K-1} C_y^k(y|y_0)\frac{\Delta^{k-1}}{(k-1)!} \tag{25}$$

$$\frac{\partial l_y^{(k)}(y|y_0,\Delta)}{\partial\Delta} = \frac{1}{\Delta}\frac{\partial C_y^{(-1)(y|y_0)}}{\partial y_i} + \sum_{k=0}^{K}\frac{\partial C_y^{(-1)}(y|y_0)}{\partial y_i}\frac{\Delta^k}{k!} \tag{26}$$

$$\frac{\partial^2 l_y^{(k)}}{\partial y_i^2} = \frac{1}{\Delta}\frac{\partial^2 C_y^{(-1)}(y|y_0)}{\partial y_i^2} + \sum_{k=0}^{K}\frac{\partial^2 C_y^{(-1)}(y|y_0)}{\partial y_i^2}\frac{\Delta^k}{k!} \tag{27}$$

By equating the coefficients of second order of Δ, we get the leading coefficient given by

$$C_y^{(-1)}(y|y_0) = -\frac{1}{2}\left(\frac{\partial C_y^{(-1)}(y|y_0)}{\partial y_i}\right)^{\mathrm{T}}\left(\frac{\partial C_y^{(-1)}(y|y_0)}{\partial y_i}\right) \tag{28}$$

By satisfying the condition that the density must approximate a Gaussian density as $\Delta \to 0$, the approximate solution is given by [27]

$$C_y^{(-1)}(y|y_0) = -\frac{1}{2}\|y - y_0\|^2 = -\frac{1}{2}\sum_{i=1}^{m}(y - y_{oi})^2 \tag{29}$$

Then by equating the terms of the first order, we get

$$\sum_{i=1}^{m}\frac{\partial C_y^{(0)}(y|y_0)}{\partial y_i}(y - y_0) = \sum_{i=1}^{m}\mu_{yi}(y)(y - y_0) \tag{30}$$

In the same way the higher coefficients are determined using the recurrence for the higher order [27, Th. 1],

$$C_y^{(0)}(y|y_0) = k\int_0^1 G_y^{(k)}(y_0 + u(y - Y_0)|y_0)u^{k-1}du, k \geq 1 \tag{31}$$

where $G_y^{(k)}$ is given by 1) for $K < 2$

$$G_y^{(k)}(y|y_0) = \sum_{i=1}^{m}\mu_H(y)\frac{\partial Ck_y^{k-1}(y|y_0)}{\partial y_i} + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2 C_y^{k-1}(y|y_0)}{\partial y_i^2} + \frac{1}{2}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\left(\frac{\partial^2 C_y^{(0)}(y|y_0)}{\partial y_i^2} + \left[\frac{\partial^2 C_y^{(0)}(y|y_0)}{\partial y_i}\right]^2\right) \tag{32}$$

and 2) for $k \geq 2$

$$G_y^{(k)}(y|y_0) = \sum_{i=1}^{m}\mu_H(y)\frac{\partial Ck_y^{k-1}(y|y_0)}{\partial y_i} + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2 C_y^{k-1}(y|y_0)}{\partial y_i^2} + \frac{1}{2}\sum_{i=1}^{m}\sum_{h=0}^{K-1}\binom{K-1}{h}\frac{\partial C_y^h C(y|y_0)\partial C_y^{(k+h)}(y|y_0)}{\partial y_i \partial y_i} \tag{33}$$

The change in variable based on the Jacobian matrix for a differentiable function $\gamma(x)$ is given by $\Delta\gamma(x) = \sigma^{-1}(x)$, then we can write $Y_t = \gamma(X_t) = \int_0^x (du / \sigma(u))$, with Y_t satisfying $dY_t = \mu_y(Y_t)dt + DW_t$. This change in variable reverts to the log-likelihood of the process and is called Lamperti transform. It should be noted that the covariance of the multivariate process is given by $\upsilon(x) = \sigma(x)\sigma^T(x)$ and letting $D_\upsilon(x) = (1/2)\ln[\mathrm{Det}[\upsilon(x)]]$, we can transform Y_t to X_t. Therefore, the log-likelihood of X denoted as $l_x = \ln p_x$ is given by

$$\begin{aligned} l_x(x|x_0),\Delta) &= -\frac{1}{2}\ln(\mathrm{Det}[\upsilon(x)] + l_y(\Delta,\gamma(x)|\gamma(x_0))) \\ &= D_\upsilon(x) + l_y(\Delta,\gamma(x)|\gamma(x_0)) \end{aligned} \tag{34}$$

From this, the approximation of order k in Δ can be used to reach the solution [24]

$$\begin{aligned} l_x^k(x|x_0,\Delta) = &-\frac{m}{2}\ln(2\pi\Delta) - D_\upsilon(x) + \frac{C_y^{(-1)}(\gamma(x)|\gamma(x_0))}{\Delta} \\ &+ \sum_{k=0}^{K} C_y^{(k)}(\gamma(x)|\gamma(x_0))\frac{\Delta^k}{k} \end{aligned} \tag{35}$$

This log probability is useful in determining the required parameters for the approximated Gaussian distribution using multivariate diffusion.

3.4.1 Multivariate Diffusion Approximation

The variation of the traffic in the queue is given by the differential stochastic equation

$$dX(t) = B(x^* - X(t))dt + \sqrt{A}dW(t), \tag{36}$$

where X being reducible, and applying the change in variable $\gamma(x) = \sigma^{-1}x$, the resulting reducible process differential equation is given by

$$\begin{aligned} dY_t &= (\sigma^{-1}Bx^*\sigma Bx^*Y_t)dt + dW_t \\ &= k(\eta - Y_t)dt + dW_t. \end{aligned} \tag{37}$$

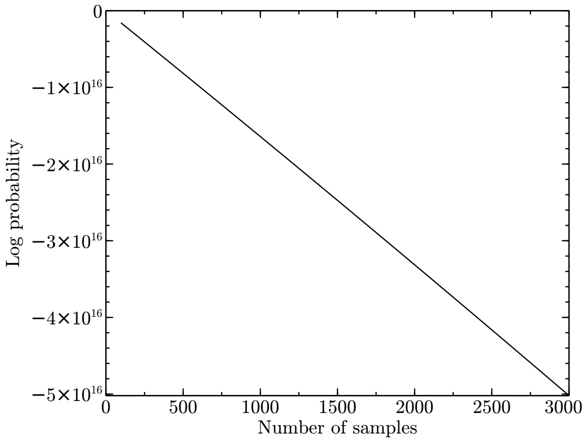


***Fig. 3**: Log-transition probability*

Furthermore, since the process is reducible, Ito's lemma of the variable γ satisfies $\nabla\gamma(x) = \sigma^{-1}(x)$.

Each element of m-dimension diffusion is reducible by means of a simple transformation known as Lamperti transform:

$$Y_t \equiv \gamma(X_t) = \int_x \frac{du}{\sigma(u)}. \tag{38}$$

Therefore, letting $X_t = \sigma Y_t$ allows the reversion to process X.

From the differential equation in Eq. (37), it can be noted that $\eta = \sigma^{-1}x^* = [\eta_i]_{i=1,2,\ldots}$, $k = \sigma^{-1}B\sigma = [k_{ij}]_{i,j=1,2,\ldots}$, and $\sigma = \sqrt{A}$.

The coefficients of the series expansion can be found in the appendix of [28] and are used accordingly to approximate the Gaussian distribution. To ensure no interference occurs, as previously mentioned, the non-diagonal of the matrix $[k_{ij}]$ denoting $\Theta = [\theta_{ij}]$ must be zero, resulting in the infinitesimal mean matrix having real eigenvalues. The conditions $\theta_{11} > 0$ and $\theta_{22} > 0$ are necessary for making the process stationary, so that the standard asymptotic provides the asymptotic distribution of the maximum likelihood estimators. Since MLE is of interest in this study, $X_e = [X_{e1}, X_{e2}]$ is defined as the asymptotic estimate value. Therefore, the drift reversion in the diffusion and diffusion equation processes becomes

$$dX(t) = B(X_t - X(t))dt + \sqrt{A}dW_t. \tag{39}$$

To approximate the distribution, coefficients in Eqs. (29), (31), (32), (33), and the log-transition probability in Eqs. (22), (24) are utilized to provide a solution through MATLAB, yielding the results shown in Fig. 3.

Fig. 3 shows the log-transition probability as a straight line with a negative gradient. Therefore, the probability of the reduced function is an exponential distribution with the rate given by the following Kolmogorov equation as

$$\frac{\partial L_y}{\partial \Delta} = \sum_i \frac{\partial \mu_y}{\partial y_i}. \tag{40}$$

Since the infinitesimal mean is $d\mu_y = k(\eta - y)dt$ as in Eq. (37), it can be observed that the rate given by the drift in the diffusion process is Θ. This implies that the transition probability is in the form of $P_y(t) \propto \exp(-k\Theta)$.

Since the purpose of diffusion approximation is to overcome the exponential server by considering the mean and variance of the service time distribution, Θ being a function of the mean and the variance, $\Theta = \sigma^{-1}\beta\sigma$ is sufficient to characterize the changes in the queue.

Since the change in variable is independent of Θ, the sampling interval, and according to Kolmogorov's equation in Eq. (40), related to the diffusion process X, it can be deduced that the probability transition of the diffusion process is $X(t)$, by $P_x(t) = \psi \exp(-\Theta t)$ where ψ is the normalization coefficient, and Θ is the diffusion coefficient of its drift and volatility with zero interference (diagonal matrix). As observed, the inference of these exponential increments can be written as in level 1, yielding the geometric distribution [29]

$$\hat{p}_n = \begin{cases} 1 - \rho, & n = 0, \\ \rho(1 - \hat{\rho})\hat{\rho}^{n-1}, & n \geq 1, \end{cases} \tag{41}$$

where $\hat{\rho} = \exp(-\Theta t)$.

4. QUEUE IMPLEMENTATION

4.1 Level 1 Multiplexer

Multiplexer implementation is carried out using two voice algorithms in MATLAB from two different sources, namely G.711 A as Source 1 and ADPCM as Source 2. Two voice algorithms are chosen for simplicity, while the choice of G.711 and ADPCM (G.721) is motivated by the fact that they can be used interchangeably in two different access layers (physical and the data link of the OSI model), namely, E1 and T1, since all are pulse-code modulation systems regardless of the version used (μ- or A-law).

G.711 is an ITU-T standard that uses a sampling rate of 8000 per second, with a tolerance of 50 parts per million (ppm). It uses non-uniform quantization where 8 bits are used to represent each sample, resulting in a 64-kbps bit rate. ADPCM is a variant of differential pulse-code modulation (DPCM) that alters the size of the quantization step to allow further reduction in the required data bandwidth for a given signal-to-noise ratio. Its algorithm maps a series of 8-bit (μ- or A-law) PCM samples into a series of 4-bit ADPCM samples, thereby doubling the line capacity. Starting with G.721, which is a 32-kbits/s scheme, and reinforcing the simulations by multiplexing the G.711 with other ADPCM schemes, namely the G.726 of 16 kbits/s and G.729 of 8 kbits/s, respectively, aims to demonstrate how the capacity can be improved with further processing techniques.

The objective of this study is to achieve a smooth transition with less delay in the waiting time of the length of the queue for both algorithms applied in two different access networks, namely T1 and E1. The specifications of these modulation schemes are as follows:

4.1.1 Source 1: G.711 A algorithm

The parameters for this source are:

- $T_1 = 1$ ms: standard value of packetization delay,
- $R_1 = 64$ kbits/s: standard bit rate,
- $N_1 = 30$ (E1 standard) , number of sources,
- $a_1 = 22$, number of active sources,
- $\beta_1 - 1 = a_1 T_1$, the mean active time,
- $\alpha_1 - 1 = 100$ ms, the mean silence time.

4.1.2 Source 2: ADPCM algorithm

The parameters for this source are:

- $T_2 = 16$ ms: standard value of packetization delay,
- $R_2 = 32$ kbits/s: standard bit rate,
- $N_2 = 24$, (T1 standard) number of sources,
- $a_2 = 16$, the number of active sources,
- $\beta_2 - 1 = a_2 T_2$, the mean active time,
- $\alpha_2 - 1 = \alpha_1 - 1$, the mean silence time.

4.2 Level 2 Multiplexer

As in the previous experiment, MATLAB is used but with a higher number of sources. The parameters then are:

- $C_1 = 3$, $C_2 = 2$, the squared coefficients of the two types of heavy traffic,
- $P_{11} = 0.6$, $P_{12} = 0.4$, $P_{21} = 0.5$, $P_{22} = 0.5$, the probability transition matrix,
- $\alpha_1 = 1/200 \times 10^{-3}$, $\alpha_2 = 1/100 \times 10^{-3}$, the holding time of the two states,
- $X_{01} = 1000$, $X_{02} = 1200$, the number in the queue at $t = 0$,
- $X_1 = 1400$, $X_2 = 1600$, the number for the two types of sources,
- $C = 100$ Mbits/s the transmission capacity link used for Ethernet,
- $X_{e1} = (1/5)C$, $X_{e2} = (1/10)C$, the transmission rate available for each source.

4.3 Simulation

A statistical multiplexer is considered in this study, whose inputs consist of two incoming links with rates r_1 and r_2. Diffusion approximation is also considered, resulting in a geometric distribution with a decrement factor r_i in the aggregated process to analyze the fluctuations in the queue caused by traffic intensity r_0. Accordingly, the geometric distribution characterizing the queue behavior is given by $p(n) = r_i^{n-1}(1 - r_0)$, $0 < r_i < 1$.

The exploitation of the equations previously mentioned requires the following inputs:

- the number and rate of type 1 sources;
- the number and rate of type 2 sources;
- and the transmission link capacity in bit/sec, and
- the number of samples.

The processing phase outputs:

- the number and the service rate in the queue.

Table 1: *Queue performance.*

ρ (%)	1	10	20	40	60	80
E[D] (Mb/s)	0.9	4.2	5.3	6.1	6.4	6.6
E[L] (packet)	593	60	29	15	10	7
E[S] (μs)	7.2	0.72	0.36	0.18	0.12	0.09

Two queues will be simulated, the resulting diffusion approximation queue and Poisson queue, and the results compared.

The results of the simulations previously described in MATLAB are presented in Figs. 4, 5, 6, and 7, representing comparisons between the Poisson and diffusion queues, as well as the impact of compression techniques on the queue content through a comparison between the G.711 and other ADPCM schemes. Figs. 4 and 5 present the simulation results of a comparison between the diffusion and Poisson queues, while Figs. 6 and 7 show how the capacity, and hence the bandwidth, can be improved efficiently. The X-axis represents the cumulative number of transitions in the simulation, while the Y-axis is the histogram representing the number of output levels (state-queue content), while the stairs represent the transition time (Figs. 6 and 7).

To gain a sense of what is happening in the queue, its performance is inspected based on the mean queue values obtained when utilization varies. The results are presented in Table 1, where ρ is the utilization, E[D] is the mean departure or throughput, E[L] is the mean queue length, and E[S] is the mean service time.

4.4 Analysis Results

Simulation of the M/G/1 queue using diffusion approximation is directly compared to the M/M/1 queue. The histogram in Fig. 4(a) shows the two levels of the M/M/1 queue where the traffic is aggregated prior to output. This complies roughly with the definition for the statistical multiplexer in that the resource allocation lies between the average and peak values, while in the diffusion queue, the multiplexer accommodates a variable rate as shown in Fig. 4(b) and hence, dynamic resources allocation. This is confirmed by the stair graphs in Fig. 5, whereby the transition times are variable for the diffusion approximation in Fig. 5(b).

However, the M/M/1 queue exhibits no flexibility in uniform time transitions (Fig. 5(b)). As demonstrated by previous studies, this lack of flexibility is more likely to experience a tail due to jitter, hence increasing delay and, consequently, packet loss. The rate accommodation is explained by the fact that the diffusion approximation process, exploiting the Gaussian property of MLE as approximated in this paper, results in smoothing the correlated non-renewal process into a Markov rate renewal process in the sequence $(X(t), Q(t))$ consisting of the phase in the Markov process affecting the number of sources $X(t)$ and queue content $Q(t)$ at the departure

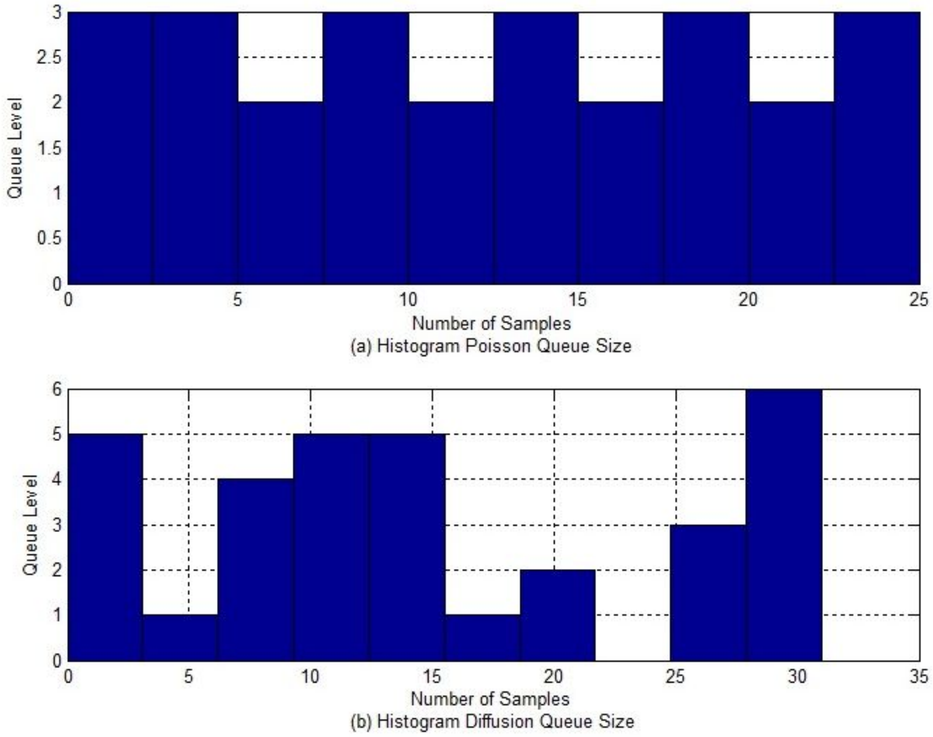


Fig. 4: Histogram for Poisson vs. diffusion queue size.

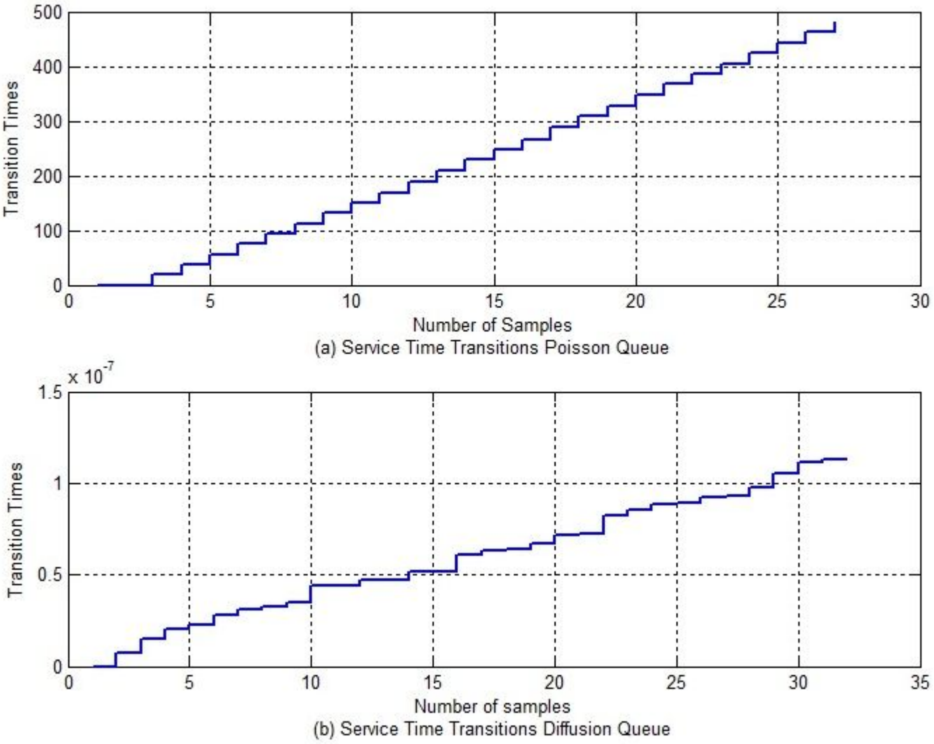


Fig. 5: Service time transitions Poisson vs. diffusion queue.

times $t > 0$. The results can be outlined as follows:

1. The variables $(X(t), Q(t))$ satisfy $N + 1$ levels as the fluid process aligns with the theory.
2. The transfer delay of the diffusion process needs an average of 3 ns times the transfer delay of the exponential queue, resulting in a smaller buffer size for the same incoming traffic. As in the simulation conducted on the M/G/1 diffusion and M/M/1 queues,

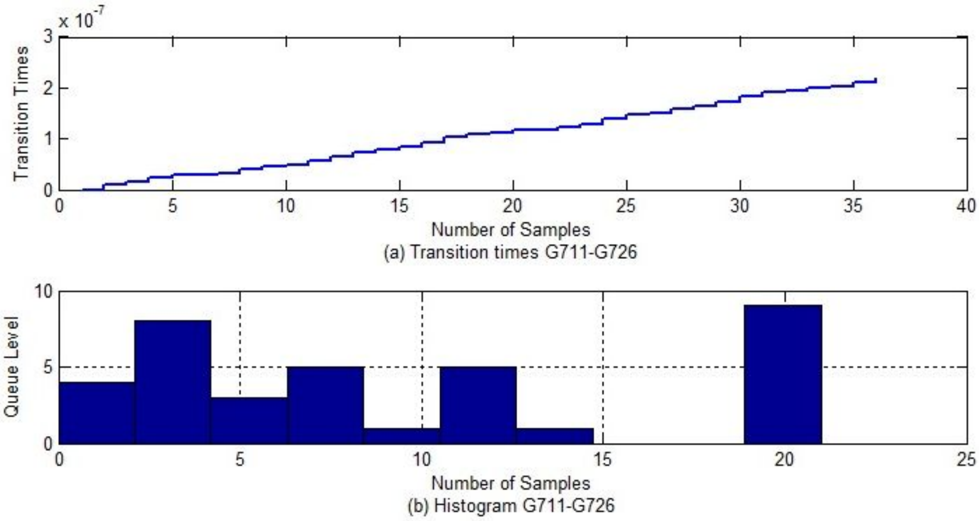


Fig. 6: Queue content G.711 and G.726 multiplexed.

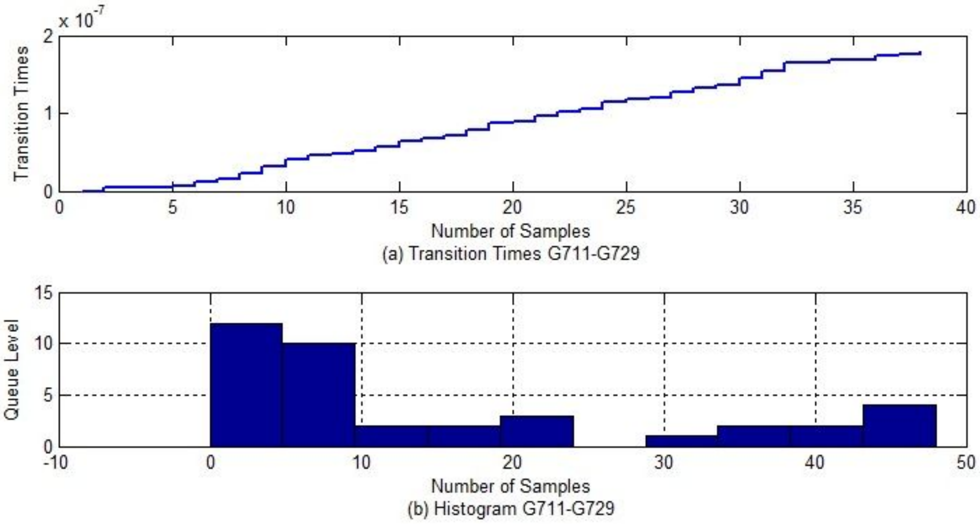


Fig. 7: Queue content G.711 and G.729 multiplexed.

if the LAN card of 100 Mbits/s provides 250 ms of buffering, the buffer size is less than 100 bytes for the diffusion queue, while about 7 Mbytes is required for the Poisson queue (8 kHz sampling rate).

3. Over the same service rate, the multiplexer also processes more packets when some bandwidth-saving techniques are employed, such as compression, allowing the queue length to adapt to the requirements of the traffic offered. In fact, the multiplexer outputs a modulated rate as it is inputted, depending only on the number of active sources X_i from a bursty arrival rate and the rates of the parallel servers $\boldsymbol{R}_i$ such that $\sum_{i=1}^{I} X_i \boldsymbol{R}_i < C$. This allows for the accommodation of all available rates and outputs them with negligible delay (less than 1 μs). This negligible delay plays a crucial role in nullifying the tail in the queue despite variations in the queue content.
4. This result is confirmed in Table 1. As previously mentioned, a packet in an Ethernet is 1518 × 8 bits long; one can realize that availing above 70% of the service rate requires a very small buffer with a service time of less than 120 ns when fewer than 10 packets are queued, while the queue length becomes important when the service time is longer than 30 ns since the server is only available for less than 20%.

This result is due to the fluid capability of the diffusion queue, hence has the potential to serve more packets offered at different rates within Ethernet LAN without any additional delay, thereby providing statistical multiplexing with multi-rate services to millions of users.

5. CONCLUSION

In this paper, a fluid solution is proposed. As with diffusion approximation, the rate adaptation method can be replaced using the distributed gateway service, employing the statistical multiplexer as the rate accommodation system. To achieve this, diffusion approximation provides an exponential server which, with its exponential decrements, yields geometric distribution. This distribution is not only suitable for Markov methods but also for fluid applications since it accommodates variable bit rates while smoothing randomness in the queue. This can therefore serve as a solution for statistical multiplexers, especially as, unlike the well-known exponential servers, the decrement rate of the exponential server achieved in this study decays faster, resulting in the probability of a full capacity queue equal to zero. As a sequence, the variable allows traffic to be smoothed like a fluid in the diffusion queue according to the space available in the server. This leads to the suggestion that the distributed network gateway services can be replaced with the more cost-effective two-level multiplexing as another way of achieving access network convergence. The initiative for setting up such a statistical multiplexer 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] E. T. Affonso, R. D. Nunes, R. L. Rosa, G. F. Pivaro, and D. Z. Rodriguez, "Speech quality assessment in wireless VoIP communication using deep belief network," *IEEE Access*, vol. 6, pp. 77 022–77 032, 2018.

[2] QoS for Video Conferencing: Critical Considerations for Empowered Communications. https://www.vonage.com/resources/articles/qos-video-conferencing-critical-considerations-empowered-communications/

[3] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 833–846, Sep. 1986.

[4] K. Chandra, "Statistical multiplexing," in *Wiley Encyclopedia of Telecommunications*, J. G. Proakis, Ed. Hoboken, New Jersey, USA: John Wiley & Sons, 2003.

[5] J. Medhi, *Stochastic Models in Queuing Theory.* San Diego, California, USA: Academic Press, 1991.

[6] J. J. Hunter, *Mathematical Techniques of Applied Probability: Discrete Time Models: Basic Theory.* San Diego, California, USA: Academic Press, 1983.

[7] V. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, vol. 32, no. 3, pp. 70–81, Mar. 1994.

[8] R. Gusella, "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 2, pp. 203–211, Feb. 1991.

[9] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] 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.

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

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

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

[20] 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*, 1993, pp. 1047–1053.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] Y. Aït-Sahalia, "Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach," *Econometrica*, vol. 70, no. 1, pp. 223–262, Jan. 2002.

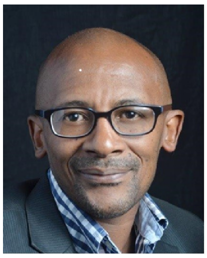
[27] A. Friedman, *Partial Differential Equations of Parabolic Type.* Englewood Cliffs, New Jersey, USA: Prentice-Hall, 1964.

[28] Y. Aït-Sahalia, "Closed-form likelihood expansions for multivariate diffusions," *Annals of Statistics*, vol. 36, no. 2, pp. 906–937, Apr. 2008.

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

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).