Journal of the Nigerian Association of Mathematical Physics
Volume 31, **(July, *2015*), pp 9 – 18**

Analysis of Diffusion Approximation for Queues and Its Application

S. S. Daodu[1], J. N. Ugwu[2], K.C Ukaoha[3] and O. Folorunsho[4]

[1,3]University of Benin, Nigeria.
[2,4]Federal University, Oye-Ekiti, Nigeria.

Abstract

Diffusion approximation is an improvement to the fluid approximation by permitting $\alpha(t)$ and $\delta(t)$ to have variations about the mean, where $\alpha(t)$ represents the total number of arrivals upto time t and $\delta(t)$ the total number of departures. It is a second-order approximation to queueing systems. In this paper we provide a detailed analysis of the diffusion approximation, and also present some specific examples of the application of the diffusion model.

Keywords: Diffusion, Heavy-Traffic, Birth-Death, Flexible manufacturing Systems, Renewal Process.

1.0 Introduction

The fluid flow approximation to queues is in fact a first-order approximation for queues in which the arrival and departure processes are replaced by their mean values, thereby creating a deterministic continuous process. Kleinrock [1] stated that diffusion approximation is an improvement to the fluid approximation by permitting $\alpha(t)$ and $\delta(t)$ to have variation about the mean, where $\alpha(t)$ represents the total number of arrivals up to time t and $\delta(t)$ the total number of departures. We introduce the variances $\sigma^2\alpha(t)$ and $\sigma^2\delta(t)$ for the arrival and departure processes respectively in order to represent the random fluctuations of these processes about their means. To introduce these fluctuations about the mean value of the process is to represent these fluctuations by normal distribution. This assumption of normality for $\alpha(t)$ {and for $\delta(t)$ } is the cornerstone of the diffusion approximation.

He further stated that, for the diffusion approximation, it is proposed that the arrival process $\alpha(t)$ and the departure process $\delta(t)$ are both to be approximated by continuous random process (with independent increments) which at time t are normally distributed with means $\alpha(t)$ and $\delta(t)$ and variances $\sigma^2\alpha(t)$ and $\sigma^2\delta(t)$ respectively. This approximation is intended to be used to make statements about number of customers in the system N(t) and the unfinished work in the system U(t). As it is well known, if we have two independent normally distributed random processes, say $\alpha(t)$ and $\delta(t)$, then any linear combination of these two is also a normally distributed process (with some appropriate mean and variance). One linear combination we are interested in is $\alpha(t)-\delta(t)$, which represents N(t)the backlog expressed in number of customers. We are also very much interested in the unfinished work U(t), which represents the backlog in units of time. Thus when N(t) is large, we have a departure process that is approximately independent of the arrival process; and it is this case which is of interest to us. The approximation that we make when the system is lightly loaded should be expected to be poor, thus we have the framework for a second-order approximation (the diffusion approximation) to queueing systems. If we replace $\alpha(t)$ by its mean $\overline{\alpha(t)}$ and its variance $\sigma^2{}_{\alpha(t)}$, then it is equivalent to making a Taylor expansion

$$F(w,t;y,\tau)-F(x,t;y,\tau)=(w-x)\,\mathrm{x}\,\frac{\partial F}{\partial x}+\frac{1}{2}(w-x)^2\frac{\partial^2 F}{\partial x^2}+o\left[(w-x)^2\right] \qquad (1.1)$$

If we substitute equation (1.1) into the equation $F(x,t-\Delta t;y,\tau)-F(x,t;y,\tau)$

Corresponding author: S. S. Daodu, E-mail: sege.daodu@gmail.com, Tel.: +234-8130762937 & 9023508022

$$= \frac{1}{\Delta t}\int_{-\infty}^{\infty}\left[F(w,t;y,\tau)-F(x,t;y,\tau)\right]dwF(x,t-\Delta t;w,t) \quad (1.2)$$

and take the limit as $\Delta t \to 0$, then from the following equations:

$$m(x,t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty}(y-x)^2 d_y\, F(x,t-\Delta t;y,t) \quad (1.3)$$

and $\sigma^2(x,t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{\infty}(y-x)^2 d_y\, F(x,t-\Delta t;y,t) \geq 0$ (1.4)

we arrive at the following partial differential equation for F:

$$-\frac{\partial F}{\partial t} = m(x,t)\frac{\partial F}{\partial t} + \frac{1}{2}\sigma^2(x,t)\frac{\partial^2 F}{\partial t^2} \quad (1.5)$$

This is the backward Kolmogorov equation for a continuous time continuous state Markov process and F will satisfy this equation except at points of accumulation (such as the origin, y=0).

Keinrock [1] also stated that the diffusion equation (also known as the Fokker-Planck equation) is the forward equation for the diffusion process. He showed that if we take the first two terms in the series

$$\frac{\partial F}{\partial t} = \sum_{n-1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial w^n}\left[An(w,\tau)f\right] \quad (1.6)$$

to be non zero and assume $An(w,t) = 0$ for n=3, 4,5…

we have

$$\frac{\partial F}{\partial t} = \frac{\partial}{\partial w}\left[m(w,\tau)f\right] + \frac{1}{2}\frac{\partial^2}{\partial w^2}\left[\sigma^2(w,\tau)f\right] \quad (1.7)$$

Equation (1.7) is known as a one-dimensional Fokker-Planck equation. Both equations (1.5) and (1.7) are referred to as diffusion equations.

In fluid approximation, we take the unfinished work U(t) as the related stochastic process of interest instead of the number in system N(t). The diffusion approximation to the equilibrium for the waiting time is given by the equation

$$F(w) = 1 - e^{2mw/\sigma^2}, w \geq 0 \quad (1.8)$$

where $m = N(t)/t = (\rho - 1)/\bar{x}, \rho > 1$

This paper is divided into 5 sections. Section 1 is the introduction, section 2 is the review of relevant literature, while section 3 deals with areas of application. Section 4 deals with specific examples of application, and section 5 is the conclusion.

2.0 Relevant Literature Review

Kimura [2] studied a diffusion model for a work station in a flexible manufacturing system. He noted that flexible manufacturing systems (FMS_s) are a class of automated systems being used in many industries to improve productivity and that a typical FMS work station has a set of parallel machines with general multi-server queues with service times modeled as a general processing time and a limited buffer, so that it can be modeled as a general multi-server queue with finite waiting space. He developed a diffusion approximation model for the standard GI/G/s/s+r queue, having a general independent arrival process, general service times, s servers, s extra waiting spaces and abandonment rates r, (the +r). The author also noted that this model refines some defects in another models of Yao and Buzacott [3]. He approximated the process of the number of customers by a time-homogeneous diffusion process in a closed interval on the nonnegative real line.

Kimura [4] considered base family of state-dependent queues whose queue-length process can be formulated by a continuous-time Markov process. He developed a piecewise-constant diffusion model for an enlarged family of queues, each of whose members has arrival and service time distributions generalized from those of the model. He stated that this is an extension as well as a refinement of the M/M/s-consistent diffusion model for the GI/G/s queue developed by Kimura [5], where the base was a birth-and-death process. As a typical base, he focused on birth-and-death processes, and also considered a class of continuous-time Markov processes with lower-triangular infinitesimal generators. He noted that there have been various diffusion models for stable queues [1, 6]. Also, see Whitt [7] and Kimura [8] for brief reviews of the previous diffusion models for single-and multi-server queues, respectively. Kimura [4] considered a base family of state-dependent queues arrival and service distributions generalized from those of the associated queue in M. The Birth-Death (BD) family M is a special subset of Markovian queues whose queue-length process is a continuous-time Markov process.

The model is applicable to queues with finite waiting spaces [3,9-13]. The model is also a refinement of the M/M/s-consistent model in the sense that it satisfies a conservation law for the steady-state distribution.

To obtain the steady-state distribution,

he let p(x, t|x^o) be the probability density function (pdf) of X(t) starting from X(0)=x^o, i.e., p(x, t|x^o)dx=P{x< X(t) < x + dx| X(0) = x^o} and pdf p(x, t | x^o) to satisfy the Kolmogorov forward (or Fokker-Planck) equation

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{a(x)p(x,t \mid x^o)\} - \{b(x)p(x,t \mid x^o)\}, \tag{2.1}$$

so that the steady-state pdf $p(x) = \lim_{t\to\infty} p(x,t \mid x^o)$, if it exists, satisfies the ordinary differential equation

$$\frac{1}{2}\frac{d^2}{dx^2}\{a(x)p(x)\} - \frac{d}{dx}\{b(x)p(x)\} = 0, \quad 0 < x < x_N. \tag{2.2}$$

At the origin x=0 and x_N, the pdf p(x) also satisfies the boundary conditions

$$\frac{1}{2}\frac{d^2}{dx^2}\{a(x)p(x)\} - b(x)p(x)\Big|_{X=0,X_N} = 0, \tag{2.3}$$

Solving (2.2) together with (2.3) and using the Piecewise Constant (PC) assumption, he obtained

$$p(x) = Cq(x), \quad 0 \le x \le x_N, \tag{2.4}$$

where C is the normalization constant,

$$q(x) = \frac{1}{a_k}\exp\left\{\frac{2b_k}{a_k}(x - x_{k-1})\right\}\prod_{j=1}^{k-1}\gamma_j \quad x \in \mathrm{I}_k,\ k = 1,...,N, \tag{2.5}$$

and

$$\gamma_j = \exp\left\{\frac{2b_j}{a_j}(x_j - x_{j-1})\right\}, \quad j = 1,...,N \tag{2.6}$$

To extract an approximation for the steady-state distribution $\{p_k\}$ from the steady-state pdf p(x), he used the pointwise discretization defined by

$$Pk = P_o q(x) = \frac{Po}{a_k}\prod_{j=1}^{k}\gamma j, \quad k = 1,...,N-1. \tag{2.7}$$

$$p_0 = \frac{\mu_N}{\sum_{j=0}^{N-1}(\lambda_j + \mu_N - \mu_j)\prod_{i=1}^{j}\gamma i} \tag{2.8}$$

and

$$p_N = \frac{\sum_{j=0}^{N-1}(\lambda_j - \mu_j)\prod_{i=1}^{j}\gamma_j}{\sum_{j=0}^{N-1}(\lambda_j + \mu_N - \mu_j)\prod_{i=1}^{j}\gamma_j,} \tag{2.9}$$

where we set $\mu_0 \equiv 0$.

Kimura [4] noted that (2.8) is an approximation for queues in G, i.e. queues with general or general independent arrival process. In general, $\{Pk\}$ in the left-hand side of (2.7) should be replaced by the steady-state probabilities just before arrivals. This means that (2.7) is still true for queues with Poisson arrivals.

Kimura [12] developed a diffusion approximation for finite-capacity multiserver queues with finite waiting space. He focused mainly on the steady-state distribution and congestion measures for the number of customers in the system. He observed that finite waiting spaces have been useful models of computer, communication, and manufacturing systems experiencing congestion due to irregular flows. He assumed that the limited waiting room corresponds to a local storage or buffer for waiting customers, considering the local storage at a work station in a flexible manufacturing system (FMS) that typically has a small number waiting spaces. He considered the model; GI/G/s/s + r queuing system with: s≥1 identical servers in parallel, r≥0, extra waiting spaces, and the FCFS (first-come first-served) discipline.

He assumed if: N(t) be the number of customers (either waiting or being served) in the system at time t≥0; the generic random variable N (N^o) to indicate the number of customers in the system at an arbitrary time (just before an arrival epoch) in equilibrium.

For k = 0, …, s+r, let pk = P(N=k) and qk = P(N^o= k). (2.10)

He generalized that it is quite difficult to obtain an analytical solution for the distribution {pk} (and also {qk}) except for a few special cases such as the M/M/s/s + r queue and the GI/M/s/s + r queue, [13]. Kimura [12] observed from Kimura [8], that generalizing the M/M/s-consistent diffusion model for the GI/G/s queue has limitations and refined the model as GI/G/s/s + r queue without these defects: which is consistent with the exact results for the M/G/s/s and M/M/s/s + r queues, and it satisfies the exact relation between ps+ r and qs+r for the GI/M/s/s + r queue. From his basic assumption, the GI/G/s/s + r queueing system was considered and specified by the following assumptions: let F (G) denote the interarrival-time (service-time) cumulative distribution function (CDF) with mean $X-1/(\rho-1)$, for $\rho>1$ and let $\mathrm{ci}=\sigma^2/\bar{x}$ be the squared coefficient of variation (SCV, that is, variance divided by the square of the mean) of F(G). Let $\rho=x/sp$ be the traffic intensity and assume that the system is in steady state. In addition, let A(t), D(t), and L(t) denote the cumulative numbers of arrivals, departures, and lost customers during the time interval (0, t], respectively. Then, the number of customers N(t) was represented as

N(t) = N(0) + A(t) - D(t) - L(t), t > 0 (2.11)

He maintained that the diffusion approximations for finite-capacity queues is to approximate the discrete-valued process {N(t); t≥0} by an appropriate time - homogeneous diffusion process {X(t); t≥0} on a finite subset of $R_+ = (0,\infty)$, utilizing asymptotic properties of the counting processes A(.), D(.), and L(.) in (2.11). Kimura [12] defined an interval Z_k of R_+ corresponding to the event. {N = k}, (k = 0,. . , s + r). In the SBN model, they proposed that I_k(SBN) = [k - 0.5, k + 0.5] for k = 0,…, s+r-1 and Z_{s+r}(SBN) = [s + r - 0.5, s + r + 0.5], whereas, in the YB model, the irregular intervals such that Z_0(YB) = {0}, Z_1(YB) = [0,[(s+r)/(s+r-1)], Z_k(YB) = [(k - l),([s+r]/[s+r-1]), k[(s+r)/(s+r-1)] for k = 2,3,…, s + r – 1 and (s+r)/(s+r-1)>1, and Z_{s+r}(YB) = {s + r } were used, for k = 2,3,… s+r-1, and (s + r)/(s + r - 1) > 1. Kimura [12] also suggested the use of intervals defined by

$$\left.\begin{array}{ll}\{0\}, & k=0,\\ (x_{k-1}, x_k), & k=1,2,\ldots,s+r\end{array}\right\} \qquad (2.12)$$

Whitt [14], developed a diffusion approximation for the queue-length stochastic process in the *G/GI/n/m* queuing model, for large n. He noted that the rapid growth of telephone call centers and more general customer contact centers has generated renewed interest in the performance of multiserver queuing models when the number of servers is large. He further noted that the primary focus of the approximation is for the steady-state delay probability and the steady-state probability that all servers are busy in the G/GI/n/∞ model. Whit [14] focused on the steady-state distribution of the diffusion process to obtain approximations for steady-state performance measures of the queuing model, especially upon the steady-state delay probability. The approximations are based on heavy-traffic limits in which *n* tends to infinity as the traffic intensity increases. Thus, the approximations are intended for large *n* and observed that Halfin and Whitt [15] showed that scale version of the queue-length for *GI/M/n/*∞ converge to a diffusion process when the traffic intensity ρ_n approaches 1 with $(1-\rho_n)\sqrt{n}\to\beta$ for $0<\beta<\infty$. Also, Whitt [16], extended that limit to a special class of *G/GI/n/m_n* models in which the number of waiting places depends on *n* and the service-time distribution is a mixture of an exponential distribution with probability *p* and a unit point mass at 0 with probability 1 − *p*. He maintained that finite waiting rooms are treated by incorporating the additional limit $m_n/\sqrt{n}\to k$ for $0<k\leq\infty$. From the heuristic diffusion approximation for the G/GI/n/∞, he obtained the approximation for the delay probability as

$$\alpha_{G/GI/n/\infty}\equiv\alpha_{G/GI/n/\infty}(\beta,z)\approx\alpha(\beta/\sqrt{z}) \qquad (2.13)$$

where α is the M/M/n/∞ asymptotic-delay-probability function defined by

$$\alpha\equiv\alpha(\beta)=[1+\beta\Phi(\beta)/\varphi(\beta)]^{-1} \qquad (2.14)$$

where β is the limit in

$$\begin{aligned}&\sqrt{n}(1-\rho_n)\to\beta\ ,\quad for\ \ 0<\beta<\infty\ \ and\\ &z\cong z(c_a^2,G)\cong 1+(c_a^2-1)\eta(G),\end{aligned} \qquad (2.15)$$

where G is the service-time DF, assumed to have finite mean $1/\mu$, $G^c\equiv 1-G$ is the associated complementary DF,

$$\eta(G) \equiv \mu \int_0^\infty G^c(x)^2 dx \equiv \frac{\int_0^\infty G^c(x)^2 dx}{\int_0^\infty G^c(x) dx} \tag{2.16}$$

and, is in (2.17),

$$z \cong z\left(c_a^2, P\right) = 1 + \frac{P\left(c_a^2 - 1\right)}{2} = \frac{P\left(c_a^2 + c_s^2\right)}{2} \tag{2.17}$$

c_a^2 is the normalization constant in a FCLT for the arrival process (assumed to hold), which requires that c_a^2 be finite.

To seek an approximation for the queue-length process and its steady-state distribution in the general G/GI/n/m model, the author used the approximation (2.15) for z to get:

$$v \cong \frac{z}{P} = \frac{c_a^2 + c_s^2}{2} \tag{2.18}$$

for v, and then obtained the associated approximation for ρ as:

$$P = \frac{z}{v} = \frac{2z}{c_a^2 + c_s^2}, \tag{2.19}$$

where z is given by (2.15) C_a^z is the scaling constant in the FCLT, as in

$$C_n(t) \equiv \left[C(t) - nt\right] \Big/ \sqrt{nc_a^2}, \quad t \geq 0 \tag{2.20}$$

(for some nonnegative scaling constant c_a^z)

and

$$C_n \Rightarrow B \ in \ (D, J_1) \tag{2.21}$$

(where B is standard (zero drift, unit diffusion coefficient, Brownian motion), and C_s^2 is the square of the co-variance (scv) of the service-time distribution.

Whit [16] noted that:

(1) the analysis of superposition arrival processes leads to approximations of the form

$$V = \frac{\left(C_a^2 + wC_s^2 + 1 - w\right)}{2} \tag{2.22}$$

where the weight w is a strictly decreasing function of $\beta = \sqrt{n}(1-\rho)$ with $w(0) = 1$ and $w(\infty) = 0$

(2) a specific function based on simulation experiments by Albin [29, 30] is

$$w \equiv w(\beta) = \left[1 + 4\beta^2\right]^{-1} \quad for \quad \beta = \sqrt{n}(1-\rho), \tag{2.23}$$

(3) a direct application of (2.23) is effective.

Whitt [16] established heavy-traffic stochastic process limits for a class of G/GI/n/m queues in which the number of servers is allowed to increase along with the traffic intensity, and noted that Puhalaki and Reiman [17], already established many-server heavy-traffic limits for the GI/PH/n/∞ model with phase-type service-time distributions, but the limit process there is a complicated multidimensional diffusion process, whose steady-state distribution remains to be determined. Whitt [16] formulated the heavy-traffic stochastic-process limits for the $G/H_2^*/n/m$ model and also stated and proved the following theorem:

Theorem: Given the family of $G/H_2^*/n/m$ models, where the rate-1 arrival process obeys the Functional Central Limit Theorem (FCLT) in

$$Cn \Rightarrow C \equiv \sqrt{c_a^2 B} \text{ in } (\mathrm{D}, \mathrm{J}_1) \tag{2.24}$$

and suppose that the arrival rate λ_n and the number of waiting spaces M_n change with n so that;

$$\sqrt{n}(1-\rho_n) \to \beta \quad for \quad -\infty < \beta < \infty \tag{2.25}$$

and

$$m_n / \sqrt{n} \to k \quad for \quad 0 < k \le \infty \tag{2.26}$$

hold with $-\infty < \beta < \infty \quad and \quad 0 < k \le \infty$. In addition, suppose that the initial conditions are as specified in (2.27) – (2.29) as follows

$$0 \le Q_n(0) \le n + m_n \tag{2.27}$$

$$Q_n(0) \Rightarrow Q(0) \quad as \quad n \to \infty \tag{2.28}$$

$$Q_n(0) \equiv [Q_n(0) - n]/\sqrt{n} \tag{2.29}$$

Then,

$$(Q_n, Q_n^a) \Rightarrow (Q, Q^a) \ in \ (D, J_1)^2 \quad as \quad n \to \infty \tag{2.30}$$

where

$$Q(t) \equiv g(Q^\rho(t)) \quad , \quad t \ge 0 \tag{2.31}$$

$$g(x) \equiv \begin{cases} x, & x<0 \\ x/\rho, & 0 \le x \le Pk \end{cases} \tag{2.32}$$

$$Q^a \equiv Q(o)\mu_e^{-1} \tag{2.33}$$

and Q^p is a diffusion process starting at $Q^P(0) = g^{-1}(Q(0))$ with a reflecting upper barrier at pk if k<∞ and an inaccessible upper boundary at infinity if k=∞.

The diffusion process Q^p has infinitesimal mean (drift function).

$$m_{Q^p} = \begin{cases} -P\mu\beta, & 0 \le x < Pk \\ -P\mu(x+\beta), & x < 0 \end{cases} . \tag{2.34}$$

and infinitesimal variance (diffusion function)

$$\sigma^2_{Q^p}(x) = P^2 N(c_a^2 + (2/P) - 1) = P^2 \mu(c_a^2 + c_s^2) \ , -\infty < x < Pk \tag{2.35}$$

Note: $C_n(t)$ denotes the arrival process, B is the standard (zero drift, unit diffusion coefficient) Brownian motion, and J_1 is the customary Skorohod topology.

Choi et al. [18] presented a diffusion approximation of the first overflow time in the GI/G/m system with finite capacity and derived the Laplace-Stieltjes transform of the first passage time of the diffusion process which approximates the system size. The authors noted that numerical results showed that the diffusion approximation is a good approximation for heavy traffic systems.

Chen and Ye [19], studied methods in diffusion approximation for multiserver systems on sandwich, uniform attraction and state-space collapse models. They observed that so many authors have worked on this area since the pioneer work of Kigman [20], and Iglehart and Whitt [21, 22]. Chen and Ye [19] stated that fluid approximation resembles the strong law of large numbers (SLLNs) and the central limit theorem (CLT) to the random sequences. Chen and Ye [19] gave examples of summation of X(n); of n independent and identically distributed random variables. Chen and Ye [19] stated that the strong law of large numbers suggests that X(n)/n converges almost surely to a constant m which is the common mean of the random variables, and central limit theorem suggests that $\sqrt{n}[X(n) - m]$ converge weakly (or in distribution) to a normal distribution. Chen and Ye [19] observed that the limiting result is fundamental to many applications; matching SLLNs as fundamental to the point of estimate and CLT to the confidence interval in statistics. Chen and Ye [19] stated that fluid approximation is about the convergence of the fluid scale process, $\bar{X}^n(n) ::= X(nt)/n$, as n→ ∞; when if exists, its limit, denoted as $\bar{X}(t)$, is referred to as the fluid limit and affirmed that fluid limit is a deterministic process which might in special cases is a linear process or a piecewise linear process. Chen and Ye [19] also observed that the standard procedure in establishing the diffusion approximation for a single scale server queuing system is through the use of the reflection mapping. This reflection mapping is used to characterize the dynamics of queueing length or the work load process and the commutative idle time process. The authors assumed φ^k to denote space k-dimension on RCLL (Right Continuous with Left Limit) function on [0,∞] endowed with uniform norm.

Given that $\varphi^k = \{x \in \varphi^k : x(0) \ge 0\}$, Chen and Ye [19] affirmed that a sequence of the process x_n in φ^k that converges to a process x under the uniform norm is the same as x_n that converges to x on any compact set, and denoted it by x→X, uniform on any compact set.

Chen and Ye [19] also stated that the sequence of the stochastic process x_n converges to x weakly as n→ ∞, and denoted this by x_n=X as n→ ∞.

3.0 Areas of Application

The diffusion model is applicable to:

- flexible manufacturing systems (FMS), which are used in many industries to improve productivity
- a communication system with a trunk reservation in scheme
- machine interference problem

4.0 Some specific examples of the application of the diffusion model

Example 1: A finite GI/G/s queue

As an example of BD-based diffusion model, Kimura [2] applied it to the GI/G/s queue with finite waiting space, which is an extension of the M/M/s-consistent diffusion model for the GI/G/s queue with infinite capacity in Kimura [8]. He considered the standard GI/G/s/N system with s (≥ 1) identical servers in parallel, N-s≡r(≥ 0) extra waiting spaces, and FCFS (first-come first-served) discipline. The system can be specified by the following notations: Let F(G) denote the interarrival-time (service-time) distribution function (DF) with mean $\lambda^{-1}\left(\mu^{-1}\right)$, and let $c_a^2\left(c_s^2\right)$be the squared coefficient of variation (scv, i.e., variance divided by the square of mean) of F (G). Let $\rho = \lambda / s\mu$ be the traffic intensity and assume that the system is in steady-state, so that $\rho < 1 \; if \; N \to \infty$.

Since the arrival process A(·) in the GI/G/s queue is a renewal process that is independent of the event {L = k} (i.e., $A_k(\cdot)$ = A(·) and hence $\alpha_k^2 = c_a^2$ for $k \in S_d$), we need to focus only on the conditional departure process $D_k(\cdot)$ for obtaining the expressions of the infinitesimal parameters $\{b_k\}$ and $\{a_k\}$. Following Kimura [8], for the case $N \to \infty$, we briefly summarize the key ideas to obtain these parameters. He suggested approximating the conditional departure process $D_k(\cdot)$ by the superposition or sum of min (k, s) independent and continuously busy service processes, i.e.,

$$D_k(t) \approx \sum_{j=1}^{\min(k,s)} S_j(t), \quad t \geq 0, \; k = 1,..., N-1, \tag{4.1}$$

where $S_j(.)$ (j=1,…, s) is the counting process whose renewal points are generated by the service-time DF G. Note that this approximation assumes the conditional departure process $D_k(\cdot)$ to be (approximately) independent of the arrival process A(.), and that the superposition process in (4.1) is not a renewal process in general. Since the conditional departure rate μ_k is the intensity of the superposition process, we have $\mu_k = \min(k,s)\mu(k \geq 1)$, and hence, from the expressions

$$b_k = \lambda_k - N_k, \; k = 1,..., N \tag{4.2}$$

and

$$b_k = \lambda - \min(k,s)\mu, \quad k = 1,..., N-1. \tag{4.3}$$

As shown in the central limit theorem,

$$\frac{A_k(t) - \lambda_k t}{\sqrt{\lambda_k \alpha_k^2 t}} \Rightarrow N(0,1) \; and \; \frac{D_k(t) - N_k t}{\sqrt{\mu_k \delta_k^2 t}} \Rightarrow N(0,1) \tag{4.4}$$

and δ_k^2 denote the scv of the approximately renewal DF of the process $D_k(.)$ in (4.4). Then, by virtue of the renewal theory and the basic relation, the infinitesimal variance $\{a_k\}$ can be written as

$$L(t) = L(0) + A(t) - D(t), \; t \geq 0 \tag{4.5}$$

$$a_k = \lambda c_a^2 + \min(k,s)\mu\delta_{\min(k,s)}, \quad k = 1,..., N-1 \tag{4.6}$$

Kimura [2] then noted there are two basic methods for obtaining δ_k^2, i.e., the asymptotic method (AM) and the stationary-interval method (SIM) [23].

While Kimura [8] used a hybrid approximation generated by combining these two methods, Kimura [4] used a simpler approximation based on the AM, which is

$$\delta_k^2 \approx \begin{cases} 1, & k = 1,..., s-1, \\ 1 + \min(\rho,1)\left(c_s^2 - 1\right), & k = s, \end{cases} \tag{4.7}$$

due to the empirical facts that the AM approximations are accurate in heavy traffic and that a delay system behaves like a loss system in light traffic.

Example 2:

Secondary servers with a buffer:

Kimura [4] also noted that a second example can be found in the work of Browne and Whitt [24] which may be found in a communication system with a trunk reservation scheme. Consider a service facility with one primary server plus a buffer of capacity $r_1(\geq 0)$. There are $s\ (\geq 1)$ secondary servers that accept overflows from the primary buffer, which have an additional buffer of capacity $r_2\ (\geq 0)$ to hold arrivals when all servers are busy. Assume that a customer in service in the secondary system immediately leaves and enters the primary buffer, whenever space opens up in the primary buffer. Kimura [4] noted that in the original example of Browne and Whitt [24], inter-arrival and service times are assumed to be exponentially distributed. In his paper, however, the author considered the model in more general settings, assuming that they are iid random variables with general distributions. Let F denote the interarrival-time DF with mean λ^{-1} and the scv c_a^2, and let G_i (i=1, 2) denote the service-time cdf at the ith server(s) with mean η_i^{-1} and the scv c_i^2.

From these assumptions, Kimura obtained the parameters of the BD-based diffusion model, i.e., $N = r_1 + r_2 + s + 1, \lambda_k = \lambda,\ \mu_k = \eta_1 + \max\{0, \min(k - r_1 - 1, s)\}\eta_2,$ and hence,

$$b_k = \begin{cases} \lambda - \eta_1, & k = 1,..., r_1, \\ \lambda - \eta_1 - (k - r_1 - 1)\eta_2, & k = r_1 + 1,..., r_1 + s \\ \lambda - \eta_1 - s\eta_2 & k = r_1 + s + 1,..., r_1 + r_2 + s. \end{cases} \quad (4.8)$$

to obtain $\{a_k\}$, he decomposed the conditional departure process $D_k(.)$ into two streams from the primary single server and the secondary multiple servers. Applying the same renewal-theoretic argument as in the previous example to these streams, he obtained

$$a_k = \begin{cases} \lambda c_a^2 - \eta_1 d_1^2, & k = 1,..., r_1, \\ \lambda c_a^2 - \eta_1 d_1^2 + (k - r_1 - 1)\eta_2, & k = r_1 + 1,..., r_1 + s \\ \lambda c_a^2 - \eta_1 d_1^2 + s\eta_2 d_2^2 & k = r_1 + s + 1,..., r_1 + r_2 + s. \end{cases} \quad (4.9)$$

where d_i^2 is given by

$$d_i^2 = 1 + \min(p, 1)(c_i^2 - 1), \quad i = 1, 2. \quad (4.10)$$

Example 3:

A GI/G/s machine interference problem:

As a third example, Kimura [4], showed how the BD-based diffusion model can be applied to the so-called machine interference problem. In the queueing context, this system is classified into finite-source queues and is denoted by GI/G/s/./K. In particular, if the running-time distribution is exponential, the arrival process is called quasi-random input [see Cooper [25], section 3.5]. Kimura [4] stated that Benson and Cox [26], obtained the steady-state distribution for the M/M/s/K case by using the BD formulation; Bunday and Scraton [27], analysed the GI/M/s/K case to find the insensitivity property that the steady-state distribution of the number of failed machines is the same in the M/M/s/K and GI/M/s/K cases. Kimura [4] further noted that Halachmi and Franta [11], proposed a basic diffusion model for the GI/G/s/K queue [10]. Sivazlian and Wang [28] proposed a similar diffusion model for a generalized machine interference problem with warm-standby spare machines. All of these diffusion models have the inconsistency defect.

To specify the BD-based diffusion model for the GI/G/s/./K queue precisely, he claimed that he needs some notations: He let F(G) denote the running-time (repair-time) DF with mean $\lambda^{-1}(\mu^{-1})$ and the scv $c_a^2(c_s^2)$. In addition, Kimura [4] let $B_j(t)(j = 1,..., K)$ denote the counting process whose renewal points are generated by the running-time DF F, which counts the number of breakdowns of the jth machine in the time interval (0, t], assuming that the repair times are virtually ignored. Then, the author approximated the conditional arrival process $A_k(.)$ by the sum of K-k independent running processes, i.e.,

$$A_k(t) \approx \sum_{j=1}^{K-k} B_j(t), \quad t \geq 0, k = 0,..., K - 1. \quad (4.11)$$

From (4.11), Kimura [4] showed that $\lambda_k = (K-k)\lambda(k=0,...K-1)$. Hence, by using the same renewal-theoretic argument as in the first example, he gave the infinitesimal parameters as

$$b_k = (K-k)\lambda - \min(k,s)\mu, and$$
$$a_k = (K-k)\lambda\alpha_k^2 + \min(k,s)\mu\delta^2_{\min(k,s)} \quad (4.12)$$

for k=1,…,K, where α_k^2 denotes the scv of the approximately renewal DF of $A_k(.)$ in (4.11) and $\delta_k^2\,(k=1,...,s)$ is given by

$$\delta_k^2 \approx \begin{cases} 1, & k=1,...,s-1 \\ 1+\min(\rho,1)(c_s^2-1), & k=s \end{cases} \quad (4.13)$$

For the scv α_k^2 in $\{a_k\}$, the AM approximation $\alpha_k^2 \approx c_a^2(k=1,...,K-1)$ has been used in all of the previous diffusion models. However, the author proposed in his paper a much simpler approximation, which is given by

$$\alpha_k^2 \approx 1, \quad k=1,...,K-1. \quad (4.14)$$

This approximation is due to the insensitive property for the exponential repair case, and it is partially supported by simulation results in [9] which show that the steady-state distribution is nearly independent of c_a^2 even if the repair-time distribution is non-exponential. Clearly, the BD-based diffusion model with this approximation is completely insensitive to the distribution form of F.

Finally, Kimura [4] noted that the insensitivity in GI/M/s/./K queue enables us to develop a modified diffusion approximation which depends on the DF F only through its scv $\{\alpha_k^2\}$ and stated that between the two basic approximations for $\{\alpha_k^2\}$, the SIM approximation is more appropriate than the AM one, since the SIM approximation satisfies the well-known property that a super-position arrival process converges to a Poisson process as the number of component processes tend to infinity, i.e., $\lim_{K\to\infty}\alpha_k^2 = 1 \;\; for \;\; k<\infty;\; cf$, [8]. On the other hand, the AM approximation $\alpha_k^2 \approx c_a^2$ does not satisfy this property except for the exponential case. From (4.11), the SIM approximation for $\{\alpha_k^2\}$ can be written as

$$\alpha_k^2(SIM) = 2(K-k)\int_0^\infty \{1-F_e(t)\}^{K-k}\,dt - 1, \quad k=1,...,K-1, \quad (4.15)$$

where F_e is the stationary-excess DF associated with the interarrival-time DF F. He used α_k^0 defined by

$$\alpha_k^0 = (K-k)\lambda\alpha_k^2(SIM) + \min(k,s)\mu,\; k=1,...,K-1, \quad (4.16)$$

instead of α_k^2 in the expression

$$\gamma_k = \left(\frac{a_k^*}{a_{k-1}^*}, \frac{\lambda_{k-1}}{\mu_k}\right)^{\theta_k} \quad with \quad \theta_k = \frac{a_k^*}{a_k}, \quad k=1,...,N \quad (4.17)$$

and hence, Kimura [4] obtained

$$P_k = \begin{cases} \dfrac{\mu_N \in_k}{\sum_{j=0}^{N=1}(\lambda_j + \mu_N - \mu_j)\in_j}, & k=0,...,N-1 \\ \dfrac{\sum_{j=0}^{N-1}(\lambda_j - \mu_j)\in_j}{\sum_{j=0}^{N-1}(\lambda_j + \mu_j - \mu_j)\in_j}, & k=M \end{cases} \quad (4.18)$$

with

$$\in_j = \begin{cases} 1, & j=0 \\ \dfrac{1}{a_j}\prod_{i=1}^{j}\left(\dfrac{a_i^*}{a_{i-1}^*}, \dfrac{\lambda_{i-1}}{\mu_i}\right)^{\theta_i}, & j=1,...,N-1 \end{cases} \quad (4.19)$$

Kimura [4] noted that $\{P_k\}$ in (4.18) is another approximation for the steady-state distribution that is consistent with the exact result for the GI/M/s/./K case.

5.0 Conclusion

We have shown that the diffusion approximation is an improvement to the fluid approximation. It is a second-order approximation. It is a good approximation for heavy-traffic system.

We also showed that it is applicable to a number of systems, even though it is an approximation method.

6.0 References

[1] Kleinrock I. (1975), Queueing Systems, **1**: Theory, (Wiley, New York).

[2] Kimura T. (1997), "A Diffusion Model for a Work Station in Flexible Manufacturing systems," APORS' 97: the fourth conference of the association of Asian – pacific operational research societies within IFORS, Melbourne, Australia.

[3] Yao D. D. and Buzacott J. A. (1985), "Workstation Models of Flexible Manufacturing Systems Part I: The Diffusion Approximation," EJOR **19**. No. 2; 233 – 240.

[4] Kimura T. (2002), "Diffusion Approximations for Queues with Markovian Bases", Annals of Operations Research, **113**, 27 – 40.

[5] Kimura T. (1995), An M/M/s-consistent diffusion model for the GI/G/s queue, Queueing Systems, **19,** 377-397.

[6] Newell G. F. (1971), Applications of Queueing Theory (Chapman and Hall, London).

[7] Whitt W. (1982), "Refining diffusion approximations for queues", Operations Research Letters, **30**, 165-169.

[8] Kimura T. (1995), An M/M/s-consistent diffusion model for the GI/G/s queue, Queueing Systems, **19,** 377-397.

[9] Sunaga T., Biswas S. K. and Nishida, N. (1982), "An approximation method using continuous models for queueing problems. II (Multi-server finite queue)", Journal of the Operations Research Society of Japan **25**, 113-128.

[10] Biswas S. K. and Sunaga T. (1981), "Diffusion approximation method for the solution of two-stage cyclic queueing problems, Memoirs of the Faculty of Engineering, Kyushu University, **41**, 17-32.

[11] Halachmi B. and Franta W.R. (1977), "A diffusion approximate solution to the G/G/k queueing system", Computers and Operations Research, **4**, 37-46.

[12] Kimura T. (2003). A Consistent Diffusion Approximation for Finite-Capacity Multiserver Queues. Mathematical and Computer Modelling, **38,** 1313-1324. www.Elseviermathematics.com/locate/mcm

[13] Hokstad T. (1975) The G/M/m queue with finite waiting room, Journal of Applied Probability, **12,** 779-792.

[14] Whitt W. (2004). A Diffusion Approximation for the G/GI/n/m Queue. Operations Research. INFORMS. **52**, No. 6, 922–941

[15] Halfin S. and Whitt, W. (1981), "Heavy-Traffic limits for queues with many exponential servers", Opns Res., **29**, No 3 567-588.

[16] Whitt W. (2005). Heavy-traffic Limits for the $G/H_2/n/m$ queue. Mathematics of operation Research, **30**: 1-27

[17] Puhalaki, A.A. and Reiman, M.I (2000), "The Multi-class GI/PH/N queue in the Halfin-Whitt regime", Adv. Probab; **32**, 564-595.

[18] Choi B. D., Lee Y.W. and Shim, Y.W. (1995), "Diffusion approximations for first overflow time in GI/G/m system with finite capacity", Journal of Applied Mathematics and Stochastic Analysis, **8,** Issue 1, 11-28.

[19] Chen H. and Ye H. (2014) Methods in Diffusion Approximation for Multiserver Systems on Sandwich. Available online at http://myweb.polyu.edu.hk/~lgtyehq/papers/ChenYe10-HTmethod.pdf Cited on 9-12-2014.

[20] Kingman J. F. C. (1965). The heavy traffic approximation in the theory of queues. In Proceedings of Symposium on Congestion Theory, D. Smith and W. Wilkinson (eds.), University of North Carolina Press, Chapel Hill, 137-159

[21] Iglehart D. L. and Whitt W. (1970a). Multiple channel queues in heavy traffic, I. Advances in Applied Probability. **2,** 150-177.

[22] Iglehart D. L. and Whitt W. (1970b). Multiple channel queues in heavy traffic, II. Advances in Applied Probability. **2,** 355-364

[23] Whitt W. (1982), "Refining diffusion approximations for queues", Operations Research Letters, 165-169.

[24] Browne S. and Whitt W. (1995), "Piecewise-linear diffusion process", in: Advances in Queueing: Theory, Methods, and Open Problems, ed. Dshalalow, J.H. (CRC Press, Boca Raton, FL,) 463-480

[25] Cooper R. B. (1981), Introduction to Queueing Theory, 2nd ed, (North-Holland, New York).

[26] Benson, F. and Cox, D.R.(1951), "The productivity of machines requiring attention at random interval", J.R. Statist. Soc B, **13,** 65-82.

[27] Bunday B.D. and Scraton R.E. (1980), "The G/M/r machine interference model", European Journal of Operational Research **4**, 399-402.

[28] Sivazlian, B.D and Wans, K.H (1990), "Diffusion approximation to the G/G/R machine repair problem with warm standby spares". Naval Research Logistics, **37**, 753-772.

[29] Albin S.L. (1982), On Poisson Approximations for superposition arrival processes in queues. Management Sci. 28, 126 – 137.

[30] Albin S.L. (1984). Approximating a point process by a renewal process II: Superpostion arrival processes. Oper. Res. 32, 1133 – 1162.