Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 232 (2009) 125–143

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Diffusion Approximation Model of Multiserver Stations with Losses

Tadeusz Czachórski[1, 2]

Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
Gliwice, Poland

Jean-Michel Fourneau[3]

PRiSM
Université de Versailles-Saint Quentin
Versailles, France

Tomasz Nycz[1, 4]

Centrum Komputerowe
Politechnika Slaska
Gliwice, Poland

Ferhan Pekergin[1, 5]

PRiSM
Université de Versailles-Saint Quentin
Versailles, France

Abstract

The article presents a diffusion approximation model of a $G/G/N/N$ station – N parallel servers without queueing. Diffusion approximation allows us to include in queueing models fairly general assumptions. First of all it gives us a tool to consider in a natural way transient states of queues, which is very rare in classical queueing models. Then we may consider input streams with general interarrival time distributions and servers with general service time distributions. Single server models may be easily incorporated into a network of queues. Here, we apply the diffusion approximation formalism to study transient behaviour of $G/G/N/N$ station and use it to construct a model of a typical call centre and to study the sliding window mechanism, a popular Call Admission Control (CAC) algorithm.

Keywords: diffusion approximation, transient analysis, G/G/N/N/queue, call center, sliding window mechanism.

doi:10.1016/j.entcs.2009.02.054

1 Introduction

We use the diffusion approximation to study a classical queueing model of the type $G/G/N/N$, known in its Markovian and steady-state version since the times of Erlang. The use of diffusion approximation allows us to include general interarrival times, general service times and to solve the model also for transient states. This approach is then applied to model a sliding window mechanism and to study the performance of a call centre.

The diffusion approximation is a second-order approximation where the value of a diffusion process represents the number of customers present (waiting at a queue or being served) in a system. The solution of the diffusion equation which is a second-order partial differential equation, gives the density of the diffusion process which in turn is an approximation of the queue distribution.

The article is organised as follows: section 2 resumes the known results on diffusion approximation needed to proceed further, especially it presents the Gelenbe's model of $G/G/1/N$ station [6]. The main merit of this model is the introduction of a special type of barriers, so called *absorbing barriers with instantaneous jumps* limiting the diffusion process to the interval $x \in [0, N]$. Reflecting barriers at $x = 0$ used previously by Newell [12] and Kobayashi [11] meant that the probability that the process is at $x = 0$ is null, hence the obtained results might be applied only as heavy traffic approximations, in cases where the server utilisation is closed to one and the probability that the system is empty is negligible.

Gelenbe's barriers have the following character: when the diffusion process comes to $x = 0$, it remains there for a time exponentially distributed with a parameter λ_0 and then it returns to $x = 1$. The time when the process is at $x = 0$ corresponds to the idle time of the system. Similarly, if the process comes to the barrier at $x = N$, it stays there for a time which is exponentially distributed with parameter μ_0 and corresponds to the time needed to complete the service performed at the moment when the queue becomes saturated, and then jumps to $x = N - 1$. This allows the use the diffusion approximation in evaluation of service stations with any coefficient of utilisation. However, the introduced type of boundary conditions does not match the usual Dirichlet conditions for differential equations and Gelenbe gives only the steady-state solution to this model, that means it solves it in case when the diffusion partial differential equation becomes an ordinary one. After the description of this model we present its transient solution which was proposed in [2] and consists in representing the density function of the diffusion process having absorbing barriers with instantaneous returns by a superposition of the densities of the diffusion process with pure absorbing barriers (the process is ended when it comes to a barrier).

In the $G/G/1/N$ model the diffusion parameters do not depend on x. Section

[1] The research was partially financed by ...

[2] Email:tadek@iitis.gliwice.pl

[3] Email:jmf@prism.uvsq.fr

[4] Email:tomasz.nycz@polsl.pl

[5] Email:pekergin@lipn.paris13-univ.fr

3 extends it to the case of piecewise-constant diffusion parameters depending on the value of the diffusion process. The main concept is to introduce fictive barriers between subintervals within which the diffusion has constant parameters. We used already such barriers to manipulate the probability flows representing jumps of diffusion process related to the sending various size optical packets in a model of an electronic-optical node [3] and to represent space-heterogeneous failures in hop-by-hop transmission in sensor networks [4]. Here, in the optics of $G/G/N/N$ model, the subintervals separated by barriers have unitary length and correspond to a fixed number of busy service channels: the interval $(0, 1]$ corresponds to one busy channel, the interval $(1, 2]$ corresponds to two busy channels, and so on. In case of $G/G/N/N + m$ model, the last interval has the length of $m + 1$ units.

In section 4 we evaluate through some numerical computations the level of errors introduced by the method and demonstrate its use through two examples. The first of them refers to the sliding window mechanism which is a kind of connection admission control (CAC) policy: during a certain time T, only N packets are allowed to enter the network. At each arrival of a packet, the number of arrivals at the period $[t-T, t]$ is checked and if it is inferior to N, the packet passes, else, if it is equal to N, the packet is rejected or classified as best-effort. This mechanism may be modelled by a $G/D/N/N$ queue. It may be applied also in two-stage version: the packets rejected at first stage are reconsidered at the second, and if during the last period T there was fewer than N_1 arrivals at this stage, they may enter the network as best-effort packets, otherwise they are deleted. The example demonstrates how the variable intensity of the input traffic is polished by the activity of the mechanism.

The second example refers to the problem of modelling call centres which is now often studied in operations research, e.g. [5,8,14]. A pool of service providers responds to arriving demands. If all responders are occupied the calls are transferred to another group of agents, if it exists, otherwise the call is lost. A common feature of such service systems is the variation of the rate and often the type of the service demand in the time. This variation is a source of a supplementary complexity in the organization of service systems. The different types of demand are modelled by service classes and the agents with different skills may be seen as different service stations with multiple servers. The model presented here does not address the full complexity of some call centres, however it permits to take into account the variable arrival rates.

2 Diffusion approximation of the $G/G/1/N$ station

Let $A(t)$, $B(t)$ denote the interarrival and service time distributions at a service station and $a(t)$ and $b(t)$ be their density functions. The distributions are general but not specified, the method requires only the knowledge of their two first moments. The means are denoted as $E[A] = 1/\lambda$, $E[B] = 1/\mu$ and variances are $\text{Var}[A] = \sigma_A^2$, $\text{Var}[B] = \sigma_B^2$. Denote also squared coefficients of variation $C_A^2 = \sigma_A^2\lambda^2$, $C_B^2 = \sigma_B^2\mu^2$. $N(t)$ represents the number of customers present in the system at time t.

Diffusion approximation, e.g. [12] replaces the process $N(t)$ by a continuous diffusion process $X(t)$, the incremental changes $dX(t) = X(t+dt) - X(t)$ of which are normally distributed with the mean βdt and variance αdt, where β, α are coefficients of the diffusion equation

$$\frac{\partial f(x,t;x_0)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta\frac{\partial f(x,t;x_0)}{\partial x} . \tag{1}$$

This equation defines the conditional pdf of $X(t)$:

$$f(x,t;x_0)dx = P[x \le X(t) < x + dx \mid X(0) = x_0].$$

The both processes $X(t)$ and $N(t)$ have normally distributed changes; the choice $\beta = \lambda - \mu$, $\alpha = \sigma_A^2\lambda^3 + \sigma_B^2\mu^3 = C_A^2\lambda + C_B^2\mu$ ensures that the parameters of these distributions grow at the same rate with the length of the observation period.

More formal justification of the use of diffusion approximation lies in limit theorems for $G/G/1$ system given e.g. in [9]. If $\hat{N}_n$ is a series of random variables derived from $N(t)$:

$$\hat{N}_n = \frac{N(nt) - (\lambda - \mu)nt}{(\sigma_A^2\lambda^3 + \sigma_B^2\mu^3)\sqrt{n}} ,$$

then this series is weakly convergent (in the sense of distribution) to ξ, where $\xi(t)$ is a standard Wiener process provided that the system is overloaded and never attains equilibrium.

In case of $G/G/1/N$ queue, the diffusion process should be limited to the interval $[0,N]$ corresponding to possible number of customers inside the system. Therefore, two barriers should be placed at $x = 0$ and $x = N$. In Gelenbe's model when the diffusion process comes to $x = 0$, it remains there for a time exponentially distributed with a parameter λ_0 which is approximated by the intensity λ of the input stream, and then jumps instantaneously to $x = 1$. When the diffusion process comes to the barrier at $x = N$ it stays there for a time exponentially distributed with the parameter μ_N, that means the time for which the queue is saturated and which is approximated by the service intensity μ and then jumps instantaneously to $x = N - 1$.

The diffusion equation becomes

$$\begin{aligned}
\frac{\partial f(x,t;x_0)}{\partial t} &= \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta\frac{\partial f(x,t;x_0)}{\partial x} + \\
&\quad +\lambda p_0(t)\delta(x-1) + \mu p_N(t)\delta(x-N+1) , \\
\frac{dp_0(t)}{dt} &= \lim_{x\to 0}\left[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)\right] - \lambda p_0(t) , \\
\frac{dp_N(t)}{dt} &= \lim_{x\to N}\left[-\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} + \beta f(x,t;x_0)\right] - \mu p_N(t) ,
\end{aligned}$$

where $p_0(t) = P[X(t) = 0]$, $p_N(t) = P[X(t) = N]$ and $\delta(x)$ is the Dirac function. The term $\lambda p_0(t)\delta(x-1)$ gives the probability density that the process is started at point $x = 1$ at the moment t because of the jump from the left barrier and the term $\mu p_N(t)\delta(x-N+1)$ defines the probability to start the process at $x = N-1$ due to the jump from the right barrier. The second equation makes balance of probability flows that influence $p_0(t)$: the term $\lim_{x\to 0}\left[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)\right]$ gives the probability flow *into* the barrier and the term $\lambda p_0(t)$ represents the probability flow *out* of the barrier. The third equation makes the same balance for the right barrier.

Our approach, see [2], to obtain the function $f(x,t;x_0)$ of the process with jumps from the barrier is to express it with the use of another pdf $\phi(x,t;x_0)$ for the diffusion process with the absorbing barriers at $x=0$ and $x=N$. This process starts at $t=0$ from $x=x_0$ end ends when it attains either of the barriers. Its probability density function is easier to determine and has the following form, see e.g. [1]

$$\phi(x,t;x_0) = \frac{1}{\sqrt{2\Pi\alpha t}} \sum_{n=-\infty}^{\infty} (a_n - b_n)$$

where

$$a_n = \exp\left[\frac{\beta x'_n}{\alpha} - \frac{(x-x_0-x'_n-\beta t)^2}{2\alpha t}\right], \quad b_n = \exp\left[\frac{\beta x''_n}{\alpha} - \frac{(x-x_0-x''_n-\beta t)^2}{2\alpha t}\right]$$

and $x'_n = 2nN$, $x''_n = -2x_0 - x'_n$.

If the initial condition is defined by a function

$$\psi(x), \quad x \in (0,N), \quad \lim_{x\to 0}\psi(x) = \lim_{x\to N}\psi(x) = 0,$$

then the pdf of the process has the form $\phi(x,t;\psi) = \int_0^N \phi(x,t;\xi)\psi(\xi)d\xi$.

Then the pdf $f(x,t;\psi)$ of the diffusion process with elementary returns from both barriers is expressed as

$$f(x,t;\psi) = \phi(x,t;\psi) + \int_0^t g_1(\tau)\phi(x,t-\tau;1)d\tau + \int_0^t g_{N-1}(\tau)\phi(x,t-\tau;N-1)d\tau \ .$$

In order to determine the densities $g_1(t)$ and $g_{N-1}(t)$ we consider the densities $\gamma_0(t)$, $\gamma_N(t)$ of probability that at time t the process enters to $x=0$ or $x=N$ are

$$\gamma_0(t) = p_0(0)\delta(t) + [1-p_0(0)-p_N(0)]\gamma_{\psi,0}(t) + \int_0^t g_1(\tau)\gamma_{1,0}(t-\tau)d\tau + $$
$$+ \int_0^t g_{N-1}(\tau)\gamma_{N-1,0}(t-\tau)d\tau \ ,$$

$$\gamma_N(t) = p_N(0)\delta(t) + [1-p_0(0)-p_N(0)]\gamma_{\psi,N}(t) + \int_0^t g_1(\tau)\gamma_{1,N}(t-\tau)d\tau + $$
$$+ \int_0^t g_{N-1}(\tau)\gamma_{N-1,N}(t-\tau)d\tau \ ,$$

where $\gamma_{1,0}(t)$, $\gamma_{1,N}(t)$, $\gamma_{N-1,0}(t)$, $\gamma_{N-1,N}(t)$ are the densities of first passage times between corresponding points, e.g.

$$\gamma_{1,0}(t) = \lim_{x\to 0}\left[\frac{\alpha}{2}\frac{\partial\phi(x,t;1)}{\partial x} - \beta\phi(x,t;1)\right] \ . \qquad (2)$$

The functions $\gamma_{\psi,0}(t)$, $\gamma_{\psi,N}(t)$ denote densities of probabilities that the initial process, started at $t=0$ at the point ξ with density $\psi(\xi)$, will end at time t by entering respectively to $x=0$ or $x=N$.

Finally, we may express $g_1(t)$ and $g_{N-1}(t)$ with the use of functions $\gamma_0(t)$ and $\gamma_N(t)$:

$$g_1(\tau) = \int_0^\tau \gamma_0(t) l_0(\tau - t) dt \;, \qquad g_{N-1}(\tau) = \int_0^\tau \gamma_N(t) l_N(\tau - t) dt \;,$$

where $l_0(x)$, $l_N(x)$ are the densities of sojourn times in $x=0$ and $x=N$; the distributions of these times are not restricted to exponential ones.

The presented transient solutions tend as $t \to \infty$ to the known steady-state Gelenbe's solutions, cf. [6]. They assume that the parameters of the model do not vary with time. However, we are interested in time-dependent input stream, hence the model is applied to small time-intervals, typically of one time-unit length, where the parameters are constant and the solution at the end of each interval gives the initial conditions for the next one.

3 State-dependent diffusion parameters, $G/G/N/N$ transient model

In the case of $G/G/N/N$ model where there is no queueing, the value of the diffusion process or the number of customers in the system corresponds to the number of active parallel channels. The output stream depends on the number of occupied channels, hence the diffusion parameters depend also on the value of the diffusion process: $\alpha = \alpha(x,t)$, $\beta = \beta(x,t)$.

The diffusion interval $x \in [0,N]$ is divided into subintervals of unitary length and the parameters are kept constant within these subintervals. For each time- and space-subinterval with constant parameters, transient solution is obtained. The equations for space-intervals are solved together with balance equations for probability flows between neighbouring intervals. The results are given in the form of Laplace transforms of density functions of the investigated diffusion process and then inverted numerically.

If $n-1 < x < n$, it is supposed that n channels are busy (n customers inside the system), hence we choose

$$\text{(3)} \qquad \alpha(x,t) = \lambda(t) C_A^2(t) + n\mu C_B^2, \qquad \beta(x,t) = \lambda(t) - n\mu \quad \text{for} \quad n-1 < x < n.$$

Jumps from $x=N$ to $x=N-1$ are performed with density μ.

In transient state we should balance the probability flows between neighbouring intervals with different diffusion parameters. We put imaginary barriers at the borders of these intervals and suppose that the diffusion process entering the barrier at $x=n$, $n=1,2,\ldots N-1$, from its left side (the process is growing) is absorbed and immediately reappears at $x=n+\varepsilon$. Similarly, a process which is diminishing and enters the barrier from its right side reappears at its other side at $x=n-\varepsilon$, see Fig. 1. The numerical value of ε should be small, for example of the order of 2^{-10},

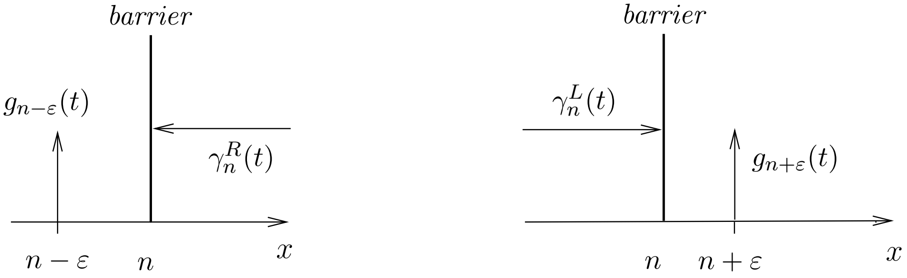

Fig. 1. Flow balance for the barrier at $x = n$

but – as we have checked – its exact value has not much influence on the solution.

The density functions $f_i(x,t;\psi_i)$, $i = 1,\dots N$, for the intervals $x \in]i-1, i[$ are as follows:

$$f_1(x,t;\psi_1) = \phi_1(x,t;\psi_1) + \int_0^t g_{1-\varepsilon}(\tau)\phi_1(x,t-\tau;1-\varepsilon)d\tau\,,$$

$$f_n(x,t;\psi_n) = \phi_n(x,t;\psi_n) + \int_0^t g_{n-1+\varepsilon}(\tau)\phi_n(x,t-\tau;n-1+\varepsilon)d\tau + \\ + \int_0^t g_{n-\varepsilon}(\tau)\phi_n(x,t-\tau;n-\varepsilon)d\tau, \qquad n = 2,\dots N-1,$$

$$f_N(x,t;\psi_N) = \phi_N(x,t;\psi_N) + \int_0^t g_{N-1+\varepsilon}(\tau)\phi_N(x,t-\tau;N-1+\varepsilon)d\tau \tag{4}$$

The relationships between the probability mass flows entering the barriers and reappearing at regeneration points are:

$$\gamma_n^R(t) = g_{n-\varepsilon}(t), \qquad \gamma_n^L(t) = g_{n+\varepsilon}(t), \qquad n = 1,\dots,N-1 \tag{5}$$

with two exceptions concerning flows coming from barriers at $x = 0$ and $x = N$: $g_{1+\varepsilon}(t) = \gamma_1^L(t) + g_1(t)$, $g_{N-1-\varepsilon}(t) = \gamma_{N-1}^R(t) + g_{N-1}(t)$.

The densities $g_1(t)$, $g_{N-1}(t)$ are the same as in $G/G/1/N$ model, Eq. (2). Also the densities $\gamma_n^R(t)$, $\gamma_n^L(t)$ are obtained in the similar way as in $G/G/1/N$:

$$\gamma_0(t) = p_0(0)\delta(t) + \gamma_{\psi_1,0}(t) + \int_0^t g_{1-\varepsilon}(\tau)\gamma_{1-\varepsilon,0}(t-\tau)d\tau\ ,$$

$$\gamma_1^L(t) = \gamma_{\psi_1,1}(t) + \int_0^t g_{1-\varepsilon}(\tau)\gamma_{1-\varepsilon,1}(t-\tau)d\tau\ ,$$

$$\gamma_n^L(t) = \gamma_{\psi_n,n}(t) + \int_0^t g_{n-1+\varepsilon}(\tau)\gamma_{n-1+\varepsilon,n}(t-\tau)d\tau +$$

$$+ \int_0^t g_{n-\varepsilon}(\tau)\gamma_{n-\varepsilon,n}(t-\tau)d\tau \ , \qquad n = 2, \ldots N-1$$

$$(6) \qquad \gamma_n^R(t) = \gamma_{\psi_{n+1},n}(t) + \int_0^t g_{n+\varepsilon}(\tau)\gamma_{n+\varepsilon,n}(t-\tau)d\tau \ +$$

$$+ \int_0^t g_{n+1-\varepsilon}(\tau)\gamma_{n+1-\varepsilon,n}(t-\tau)d\tau \ , \qquad n = 1, \ldots N-2$$

$$\gamma_{N-1}^R(t) = \gamma_{\psi_N,N-1}(t) + \int_0^t g_{N-1+\varepsilon}(\tau)\gamma_{N-1+\varepsilon,N-1}(t-\tau)d\tau$$

$$\gamma_N(t) = p_N(0)\delta(t) + \gamma_{\psi_N,N}(t) + \int_0^t g_{N-1+\varepsilon}(\tau)\gamma_{N-1+\varepsilon,N}(t-\tau)d\tau \ ,$$

where $\gamma_{i,j}(t)$ are the densities of first passage times between points i, j and are obtained as in $G/G/1/N$ model, Eq.(2).

This system of equations is transformed with the use of Laplace transform and solved numericaly to obtain the values of $\bar{f}_n(x, s; \psi_n)$. Then we use the Stehfest inversion algorithm [13] to compute $f_n(x, t; \psi_n)$; for a specified t, the values of $\bar{f}_n(x, s; \psi_n)$ are needed for several s. In the case of a $G/G/20/20$ station, for a single value of $f(x, t; x_0)$, we need 16 values of its Laplace transform and each of them is a solution of the system of equations (4), (5), (6) for 20 intervals. This analytical-numerical approach needs careful programming to assure numerical stability of computations. When time tends to infinity, and $\lambda(t) = \lambda$, the transient solution is convergent to the steady-state solution in the following form:

$$f_1(x) = \begin{cases} \frac{\lambda p_0}{-\beta_1}(1 - e^{z_1 x}) & \text{for } \beta_1 \neq 0 \\ \frac{2\lambda p_0}{-\alpha_1} x & \text{for } \beta_1 = 0 \end{cases} \qquad 0 < x \leq 1 \ ,$$

$$(7) \quad f_n(x) = \begin{cases} f_{n-1}(n-1)e^{z_n(x-n+1)} & \text{for } \beta_n \neq 0 \quad n-1 \leq x \leq n \ , \\ f_{n-1}(n-1) & \text{for } \beta_n = 0 \quad n = 2, \ldots, N-1 \ , \end{cases}$$

$$f_N(x) = \begin{cases} \frac{\mu p_N}{-\beta_N}\left[e^{z_N(x-N)} - 1\right] & \text{for } \beta_N \neq 0 \\ \frac{-2\mu p_N}{\alpha_N}(x - N) & \text{for } \beta_N = 0 \end{cases} \qquad N-1 \leq x < N \ .$$

where $z_n = \frac{2\beta_n}{\alpha_n}$.

The examples that follow demonstrate the quality of the approximation as well as of the algorithmic part of the solution.

4 Numerical examples

Let us consider a single $M/M/20/20$ station as a basic model. During the period $t \in [0, 60]$ the intensity of the input stream $\lambda_{in}(t)$ has a saw-teeth shape, typical for TCP flows, and then $\lambda_{in}(t) = 0$. This intensity may be seen in Fig. 2 where it is compared with the output flow from the station, computed as $\lambda_{out}(t) = \sum_{n=1}^{N} p(n,t)n\mu$. The service intensity of a single channel is $\mu = 1$.

Fig. 3 shows the mean number of working channels (number of customers inside the system) as a function of time (diffusion and simulation results). For each fixed time we take in simuation the average value of 100 000 independent values noted for this time during consecutive runs. As the number of the simulation repetitions is large (100 000), we may evaluate the relation between confidence level and the confidence interval Δ on the basis of normal distribution. Fig. 4 displays the confidence interval obtained for the mean number of occupied channels, as previously presented in Fig. 2, for the confidence level of 99%. The confidence interval is changing as a function of time because for each time the variance of all masurements is different. The intervals are very small compared to measured values, therefore below we consider the simulation results as exact.

There is a common opinion that diffusion approximation is satisfactory when the number of customers allowed to the system is large: the approximation of the discrete process by a continuous process seems more natural in that case. In our previous work we have observed this fact and it does not need to be emphasized. However, when we compare the mean queue length of a $M/M/2/2$ model in Fig. 5 to that of $M/M/20/20$ model, Fig. 3, we observe no visible differences in errors of the method. This observation which must be confirmed by more detailed studies allows us to predict a good level of robustness with respect to the model size.

Fig 6 illustrates the errors introduced by the method for the mean number of occupied channels (number of customers) at M/M/20/20 and M/D/20/20 stations. It displays the relative errors of the diffusion approximation; its results are compared with the simulation results and the latter are considered as exact. At the beginning of the execution time the relative error is comparatively large but it is not so significant due to small absolute values of errors. The sharp increases of errors in case of constant service come from fast changes of the input stream. The reduction of these errors may be done by a changing the computations steps but it is time-consuming and difficult to do in a general way. Excluding these special points, the errors are limited to few percent.

Fig. 7 shows for a fixed time ($t = 30$ time units) the influence of C_A^2 and C_B^2 on the mean value of occupied channels (number of customers) in $G/G/20/20$ model (λ and μ as previously). The presented function gives the ratio of the mean value $E[N(t = 30)]$ for $G/G/20/20$ station, for chosen values of C_A^2 and C_B^2 to $E[N(t = 30)]$ for $M/M/20/20$ station ($C_A^2 = 1$ and $C_B^2 = 1$). Unexpectedly, the influence is not so visible as in steady-state case and it is inverse to that one might think: larger C_A^2 and C_B^2 result in smaller mean queues. This fact needs further investigation.

Not only mean values are important, frequently we are interested in distributions. Fig. 8 displays the flow of rejected customers and Fig. 9 gives probability $p(20,t)$ that all channels are occupied – we see, as the input stream is Poisson at this example, an illustration of PASTA theorem. Of course, diffusion model is not restricted to Poisson streams.

Figures 10 – 12 compare the approximations $f(x,t;0)$ of the number of working channels with simulation histograms to demonstrate the magnitude of the approximation errors and almost perfect match of the predicted dynamics of changes to the reality. The comparison of the curves points out the good behaviour of the distribution obtained with the diffusin approximation. Even if some diffrences exist between the distribution obtained by simulation and the diffusion approximation, they are usually due to a small shift in a time. Indeed, distribution computed with the diffusion approximation is slightly late or in advance with respect to those obtained by the simulation of the same time-moment. However, when the input parameters remain unchanged for some time interval, these differences desappear and distribitions become more closed.

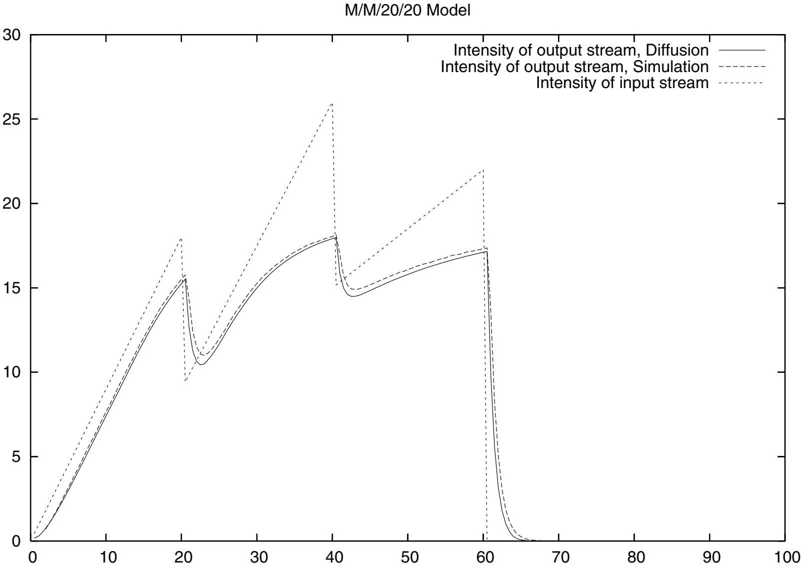

Fig. 2. Input stream and output streams in case of exponential service - diffusion and simulation

To investigate the behaviour of the model in larger time scale, Fig. 13 shows the input rate as a function of time and compares the mean queue length predicted by diffusion approximation with given by simulations of various length (10, 20 and 30 thousand of repetitions). We observe a systematic error for nearly steady-state (the interval $t \in [500, 1000]$, where input rate is fixed). The next Fig. 14 compares the density of diffusion process for $t = 1000$ time-units, obtained with described above procedure (the diffusion equation is solved inside intervals of the length of the half of time-unit, hence $t = 1000$ means 2000 of consecutive subintervals) with steady-state solution (7) and simulation histograms. All curves have the same shape

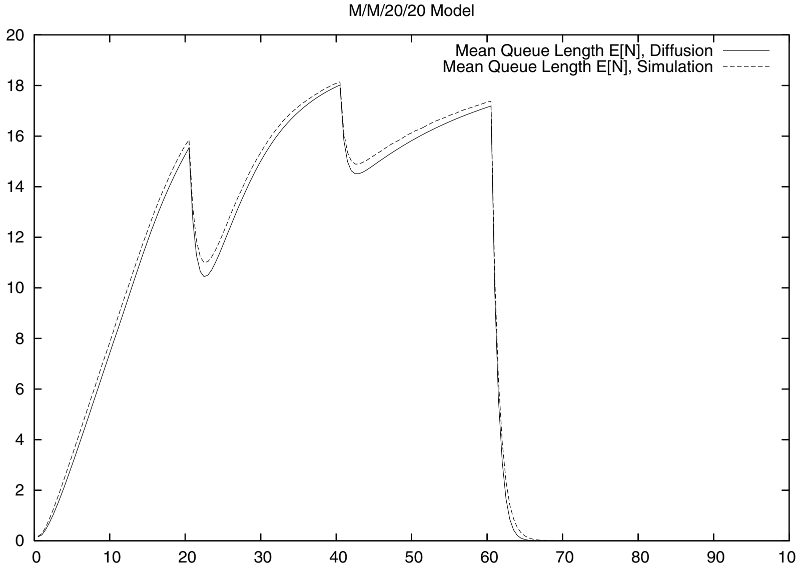

Fig. 3. Mean number of customers as a function of time, diffusion approximation and simulation results

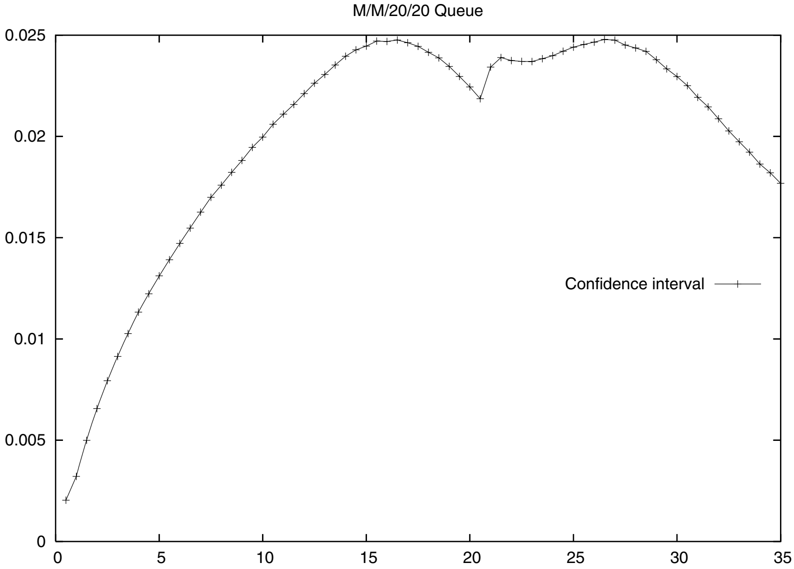

Fig. 4. Confidence interval for the mean value of occupied channels given by simulation of M/M/20/20 model, confidence level of 99%

but are slightly shifted.

Below, we sketch two possible applications of the developped $G/G/N/N$ model.

The first example refers to a single service station $M/D/20/20$ which may be considered as a model of sliding window mechanism, described in the introduction. The constant service time is $T = 1$ time unit, the input flow is the same as in Fig. 2. The output flow is computed as $\lambda_{out}(t) = \lambda_{in}(t - T)[(1 - p(N, t - T)]$. The

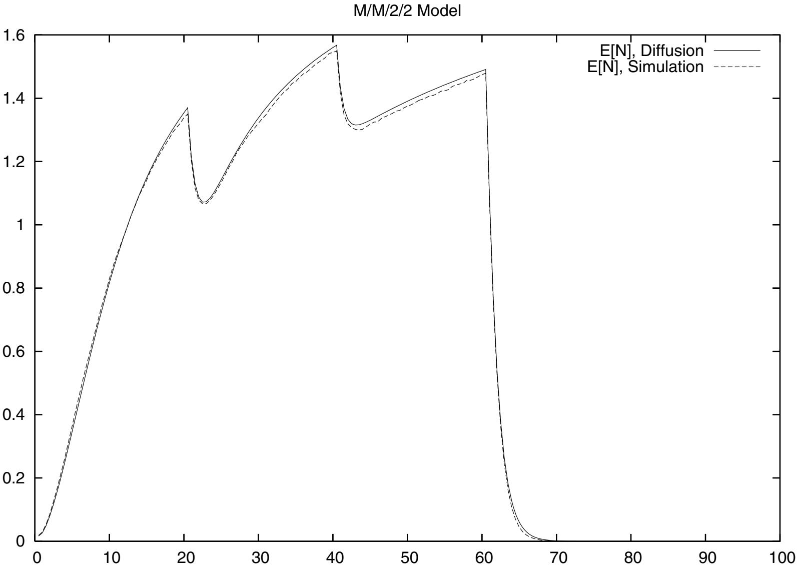

Fig. 5. Mean number of customers as a function of time, $N = 2$, diffusion approximation and simulation results

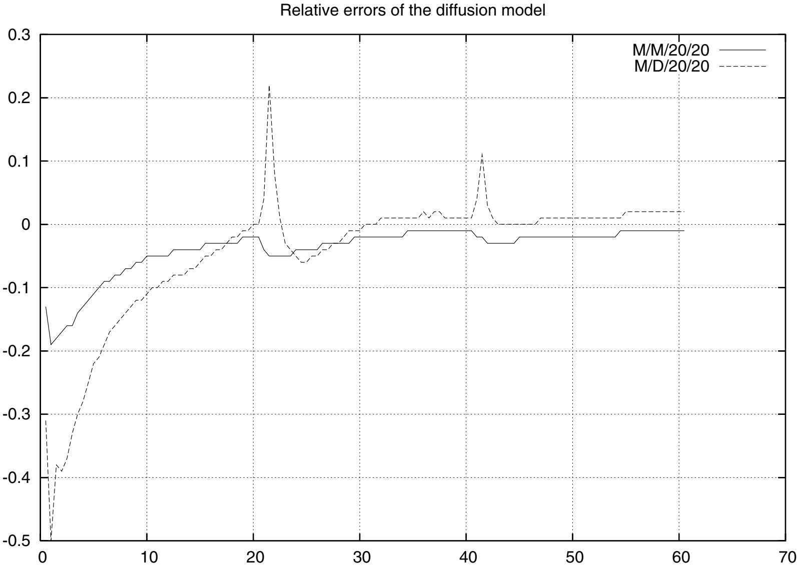

Fig. 6. Mean number of occupied channels at M/M/20/20 and M/D/20/20 stations: relative errors of the diffusion approximation model compared with the simulation model, the results of which are considered as exact

polishing effect of the sliding window mechanism is well visible. In the same figure the output flow given by the diffusion approximation is confronted with the results of simulation, obtained as previously as the mean of 100 000 independent runs. We see that the diffusion and simulation results are very close.

The second example refers to a simple model of a call centre. In station 1

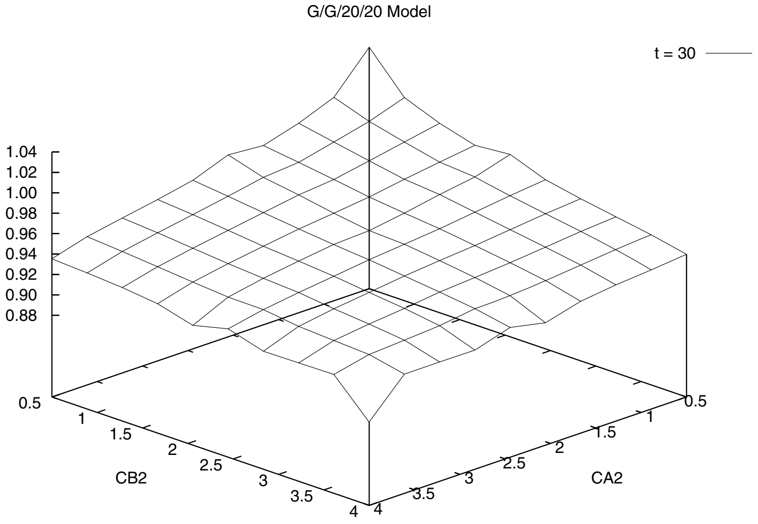

Fig. 7. The influence of C_A^2 and C_B^2 on the mean value of occupied channels in $G/G/20/20$ model: the ratio of $E[N(t=30)]$ for chosen values of C_A^2 and C_B^2 to $E[N(t=30)]$ for $M/M/20/20$ station ($C_A^2 = C_B^2 = 1$).

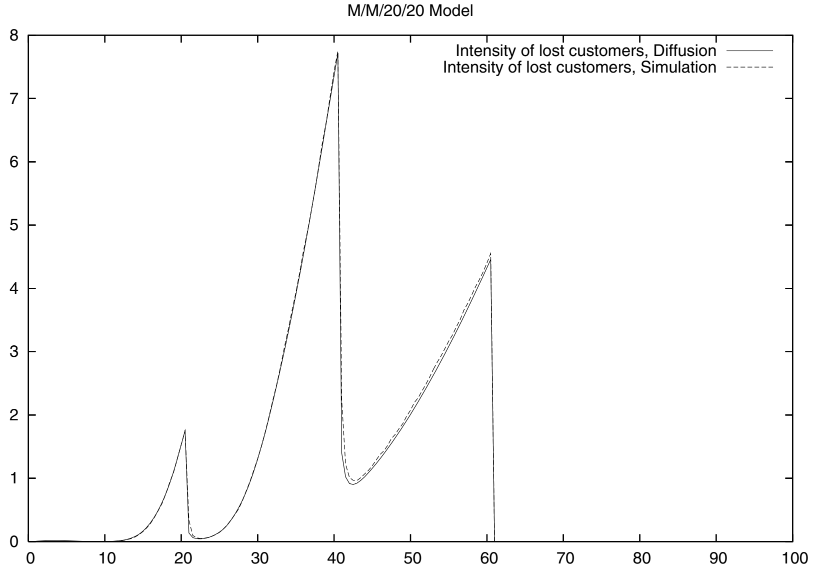

Fig. 8. The flow of rejected customers, case of exponential service, diffusion approximation and simulation results

there are N_1 parallel channels serving task type one (experts A), in station 2, N_2 parallel channels may serve task type two (experts B) and at station 3 there are N_3 channels to serve both types of tasks (experts A+B) in case when all N_1 or N_2 channels are busy. We assume $N_1 = N_2 = 20$ and $N_3 = 10$.

The stations are empty at $t = 0$. The input rate $\lambda_{1,in}(t)$ of station 1 flow starts

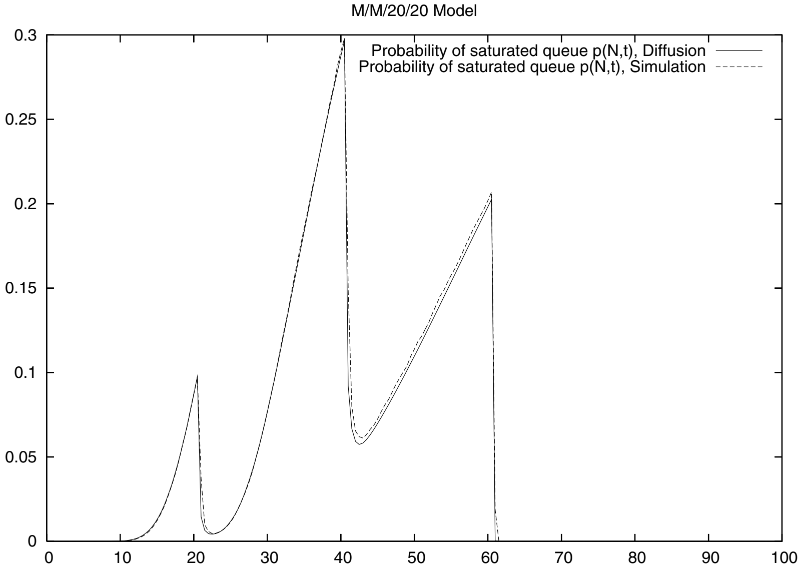

Fig. 9. Probability $p(N, t)$ that all channels are occupied, case of exponential service

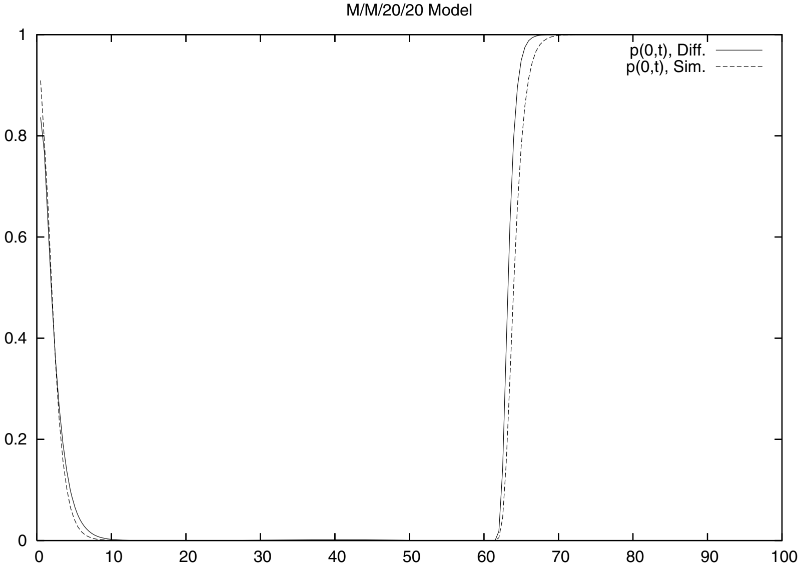

Fig. 10. Probability $p(0, t)$ that all channels are empty

at $t = 0$ and grows linearly from $\lambda_{1,in}(0) = 0$ to $\lambda_{1,in}(50) = 30$ and then decays also linearly to $\lambda_{1,in}(80) = 0$. The second station flow starts its linear growth from zero at $t = 20$ and attains the value $\lambda_{2,in}(60) = 30$ and then goes down linearly to $\lambda_{2,in}(100) = 0$, see Fig. 16.

Fig. 17 presents mean value of busy channels in all three stations, Fig. 18 presents the overflow streams redirected from station 1 to station 3 and from station 2 to station 3 and also the sum of these flows, i.e. entire input stream to station

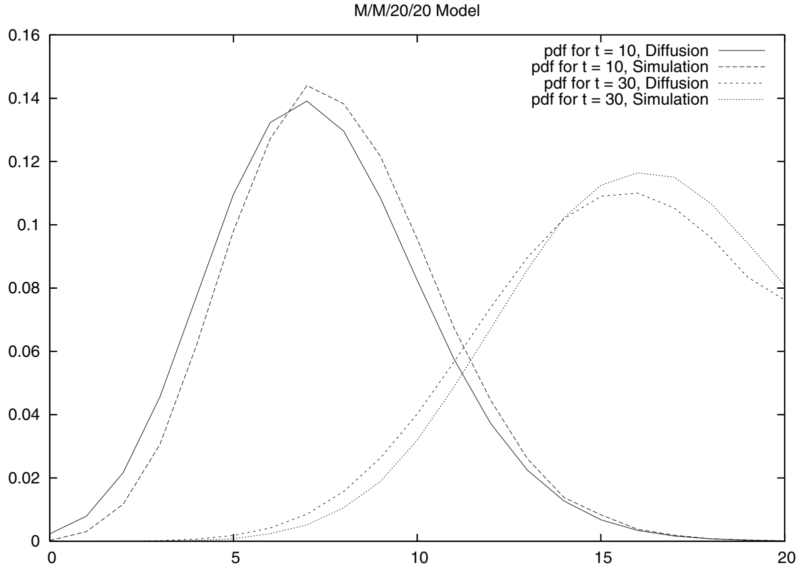

Fig. 11. Densities $f(x, t = 10; 0)$, $f(x, t = 30; 0)$ and corresponding simulation histograms

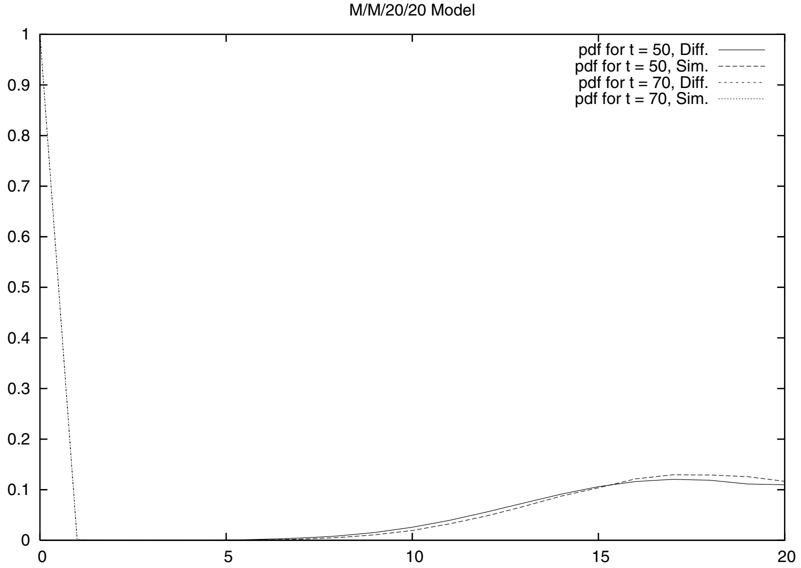

Fig. 12. Densities $f(x, t = 50; 0)$, $f(x, t = 70; 0)$ and corresponding simulation histograms

3. All diffusion results are compared with simulation (mean of 100 000 runs) and we observe that the results of both methods do not differ much. The good estimation of $p(N, t)$ and the intensity of rejected flows are particularly important in the dimensioning and optimization of call centres. In multistage call centres the input streams of subsequent stages correspond to overflow streams of previous stations. The saturation probability $p(N, t)$ of these low-level stations is another measure of main interest.

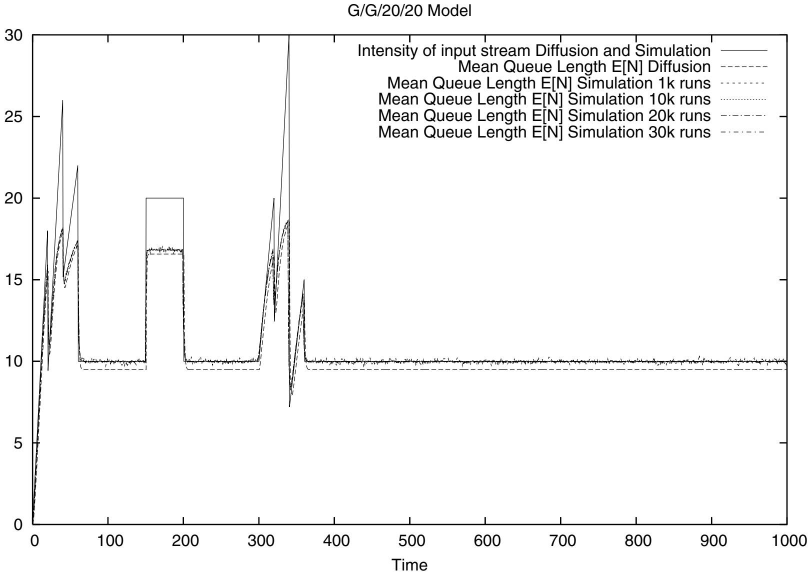

Fig. 13. Longer interval: $t \in [0, 1000]$, input rate (solid line) and the mean number of customers given by diffusion approximation and by simulations of various length (10, 20, and 30 thousand repetitions)

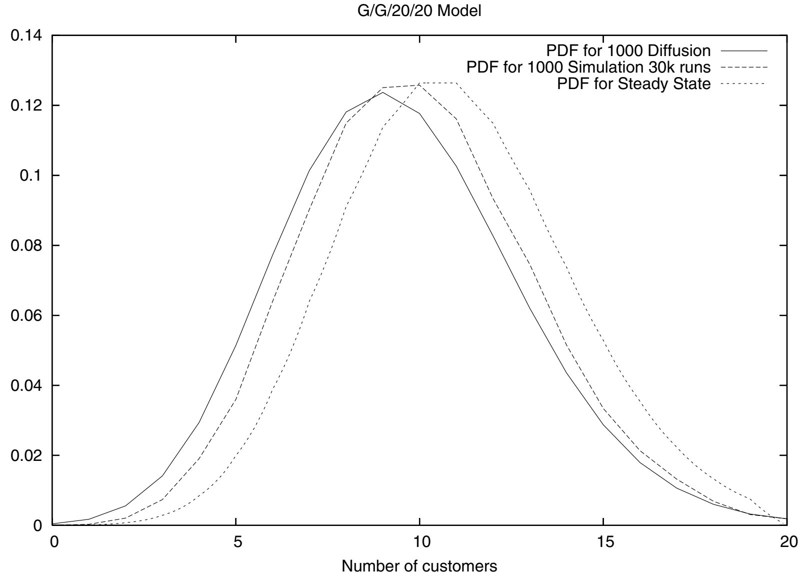

Fig. 14. Comparison of the density $f(x, t = 1000; 0)$ with steady-state diffusion solution and simulation histogram

5 Conclusions

In the article we develop the algorithmic and numerical side of the diffusion approximation in case of time-dependent and space-dependent parameters. The approach is based on constant parameter model that we proposed earlier and on the division of the time and space axes into small intervals with constant parameters. The main

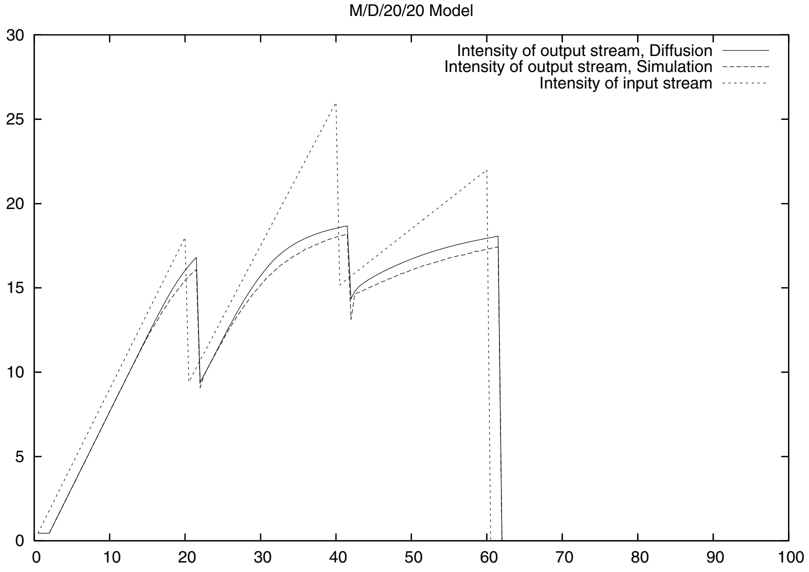

Fig. 15. Input stream and output streams in case of constant service, sliding window model – diffusion and simulation

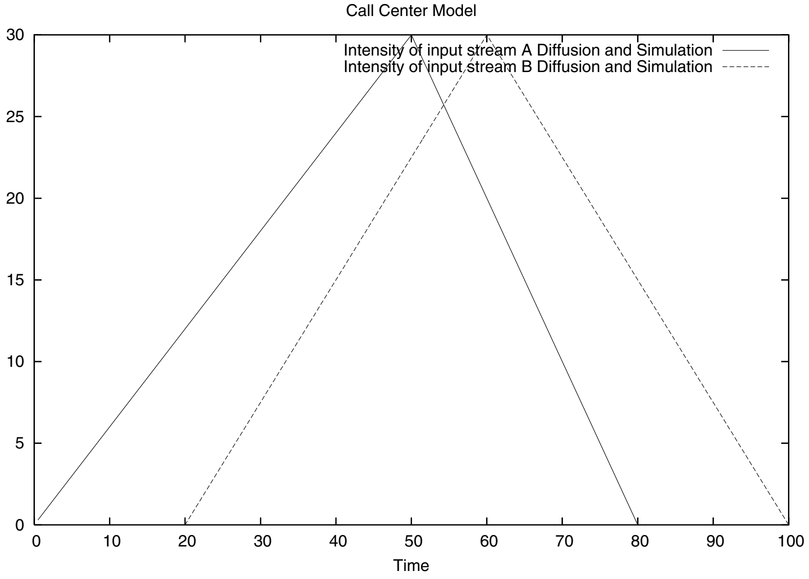

Fig. 16. Input streams in the call centre model

effort of this approach is to represent the interdependencies of the process evolving in the adjacent intervals (description of the probability flows coming from one interval into another). The system of introduced balance equations may be also used to represent complex jumps of the diffusion process extending thus the applicability of the diffusion approximation. As indicate several numerical examples, the apoach gives generaly good results and is an efficient and accurate tool to study the

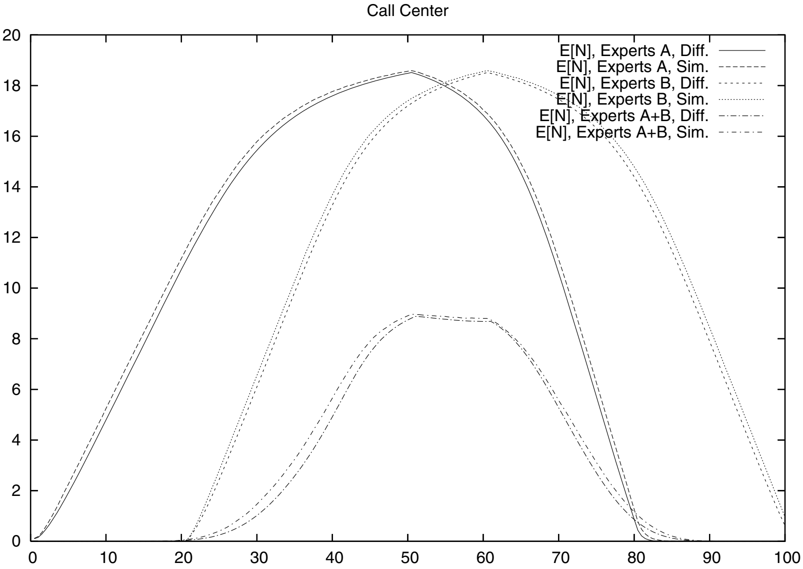

Fig. 17. Mean number of customers in three station of call centre model, diffusion and simulation

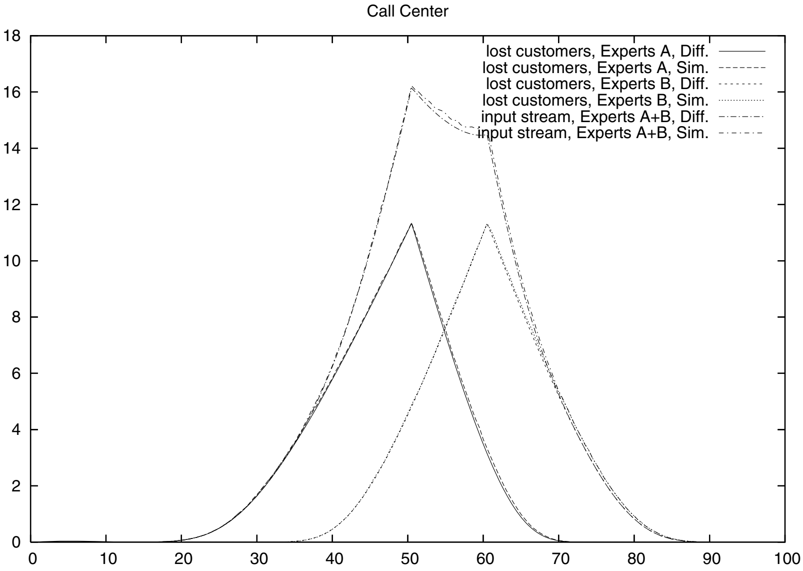

Fig. 18. Call centre model, flows redirected from station 1 and from station 2 to station 3, separately and the sum of both streams, diffusion and simulation results

transient behaviour of queueing systems, well predicting the dynamics of a queue changes, in terms of its mean value as well as of the distribution. The examples studied refer to G/G/N/N station and its applications.

The effort to develop the related software and the computational cost is not negligible, but seems smaller then in case of very large Markov models or repetitive simulations which might be the alternative approaches. The execution time of considered diffusion models is of the order of few minutes while for corresponding simulations on the same computer (PC) the time is of the order of hours or days. A more detailed study of the method accuracy and execution times is planned.

References

[1] Cox, R. P., and H. D. Miller. *The Theory of Stochastic Processes*, Chapman and Hall, London, UK, 1965.

[2] Czachórski, T. A method to solve diffusion equation with instantaneous return processes acting as boundary conditions, *Bulletin of Polish Academy of Sciences, Technical Sciences* vol. 41 (1993), no. 4.

[3] Czachórski, T., and F. Pekergin, A diffusion approximation model of an electronic-optical node *Proc. of 2nd European Performance and Engineering Workshop, EPEW 05*, Versailles, 1-3 September 2005, LNCS no. 3670, pp. 187-199, Springer Verlag 2005.

[4] Czachórski, T., K. Grochla, and F. Pekergin, Diffusion approximation model for the distribution of packet travel time at sensor networks *LNCS* no. 5122, 2008. (Title: Wireless Systems and Mobility in Next Generation Internet)

[5] Gans, N., G. Koole, and A. Mandelbaum, Telephone Call Centers Tutorial, Review, and Research Prospects, *Manufacturing & Service Operations Management*, Vol. 5, pp. 79-141, 2003.

[6] Gelenbe, E. On Approximate Computer Systems Models, *J. ACM*, vol. 22, no. 2, (1975).

[7] Gelenbe, E., and G. Pujolle. The Behaviour of a Single Queue in a General Queueing Network, *Acta Informatica*, Vol. 7, Fasc. 2, pp.123-136, 1976.

[8] Green, L. V. , P. J. Kolesar, and W. Whitt. Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System. *Production and Operations Management (POMS)*, vol. 16, No. 1, pp. 13-39, 2007,

[9] Iglehart, D. *Weak Convergence in Queueing Theory,* Advances in Applied Probability, vol. 5, pp. 570-594, 1973.

[10] Kleinrock, L. *Queueing Systems, vol. I: Theory, vol. II: Computer Applications*, Wiley, New York USA 1975, 1976.

[11] Kobayashi, H. *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*, Addison Wesley, Reading, Massachusetts USA 1978.

[12] Newell, G. F. *Applications of Queueing Theory,* Chapman and Hall, London 1971.

[13] Stehfest, H. Algorithm 368: Numeric inversion of Laplace transform, *Comm. of ACM*, vol. 13, no. 1, p. 47-49 (1970).

[14] Whitt, W. A Diffusion approximation for the G/G/n/m Queue, *Operations Research*, Vol. 52, no. 6, pp. 922-941, 2004.