Diffusion Approximation Models for Transient States and their Application to Priority Queues

Tadeusz Czachórski
IITiS PAN
Polish Academy of Sciences
44-100 Gliwice, ul. Baltycka 5, Poland
tadek@iitis.gliwice.pl

Tomasz Nycz
Centrum Komputerowe
Politechnika Slaska
44-100 Gliwice, ul. Akademicka 16, Poland
tomasz.nycz@polsl.pl

Ferhan Pekergin
LIPN
Université Paris-Nord
93430 Villetaneuse, France
pekergin@lipn.univ-paris13.fr

***Abstract*—The article presents a diffusion approximation model applied to investigate the behavior of priority queues. We discusses the use of the diffusion approximation in transient analysis of queueing models in the case of a single station and of a queueing network presenting the solutions. We emphasize the numerical aspect of the solution and analyze the errors. In classical queuing theory, the analysis of transient states is complex and practically does not go far beyond M/M/1 queue and its modifications. However, the time dependent flows in computer networks and especially in Internet focus our interest on transient-state analysis, which is necessary to investigate the dynamics of TCP flows cooperating with active queue management or to see the changes of priority queues which assure the differentiated QoS. With the use of G/G/1/N and G/G/1/N/PRIOR models, we present the potentials of the diffusion approximation and in conclusions we compare it with alternative methods: Markovian queues solved numerically, fluid-flow approximation and simulation. Diffusion approximation allows us to include fairly general assumptions in queueing models. Besides the transient state analysis, it gives us a tool to consider input streams with general interarrival time distributions and servers with general service time distributions. Single server models can be easily incorporated into the network of queues. Here we apply the diffusion approximation formalism to study transient and steady-state behavior of G/G/1 and G/G/1/N priority preemptive models. The models can be easily converted to non-preemptive queueing discipline. The introduction of self-similar traffic is possible as well. The models can be useful in performance evaluation of mechanisms to differentiate the quality of service e.g. in IP routers, WiMAX, metro networks, etc.**

***Index terms* — diffusion approximation, transient states, priority queues.**

I. Introduction

The paper extends results presented earlier in [11]. Classical queueing models of priority queues are practically limited to steady-state analysis of M/G/1 queues with non-preemptive or preemptive resume priorities, see e.g. [18], [23], [19]. It is not enough to analyze today mechanisms to ensure the quality of service inside e.g. IP routers or in access networks where the load is changing dynamically and the traffic is entirely different from Poisson streams. Therefore we adapt the method of diffusion to consider transient states in the case of priority queues. The method is based on Gelenbe's model of G/G/1 and G/G/1/N queue supplemented with our approach [5] to solve transient states using this model. The single server models are summarized in Section II, in Section III they are extended to open network queueing models. We tested this approach several times in other non-priority models, considering e.g. the dynamics of FIFO queues in ATM routers [1], the dynamics of queues in ATM multiplexers in the case of self-similar traffic [6], the stability of TCP connections in the presence of AQM (RED queues) inside IP routers [8], investigating transmission time in ad-hoc networks [9] or modeling traffic control by leaky-bucket algorithm [10]. Section IV presents diffusion approximations of busy periods distributions which are important for priority queues presented in Section V.

II. Diffusion approximation of a FIFO station

Let $A(x)$, $B(x)$ denote the interarrival and service time distributions at a service station and $a(x)$ and $b(x)$ be their density functions. The distributions are general but not specified, the method requires only the knowledge of their first two moments. The means are denoted as $E[A] = 1/\lambda$, $E[B] = 1/\mu$ and variances are $\mathrm{Var}[A] = \sigma_A^2$, $\mathrm{Var}[B] = \sigma_B^2$. 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 replaces the process $N(t)$ by a continuous diffusion process $X(t)$, e.g. [25], 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

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

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 density of the diffusion process approximates the distribution of $N(t)$: $p(n,t;n_0) \approx f(n,t;n_0)$, and in steady state $p(n) \approx f(n)$.

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 [17]. 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.

A. Unlimited queue: G/G/1 station, transient solution

The process $N(t)$ is never negative, hence $X(t)$ should be also restrained to $x \geq 0$. A simple solution is to put a *reflecting barrier* at $x = 0$, see [22]. In this case

$$\int_0^\infty f(x,t;x_0)dx = 1\,,$$

and

$$\frac{\partial}{\partial t}\int_0^\infty f(x,t;x_0)dx = \int_0^\infty \frac{\partial f(x,t;x_0)}{\partial t}dx = 0\,.$$

Replacing $\partial f(x,t;x_0)/\partial t$ in the above integral by the right side of the diffusion equation we obtain the boundary condition corresponding to the reflecting barrier at zero:

$$\lim_{x\to 0}[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)] = 0\,. \qquad (2)$$

The solution of Eq. (1) with conditions (2) is, cf. [22]

$$f(x,t;x_0) = \frac{\partial}{\partial x}\left[\Phi\left(\frac{x-x_0-\beta t}{\alpha t}\right) - \mathrm{e}^{\frac{2\beta x}{\alpha}}\,\Phi\left(\frac{x+x_0+\beta t}{\alpha t}\right)\right]$$

where $\Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\Pi}}e^{-t^2/2}dt$ is the PDF of standard normal distribution.

The reflecting barrier excludes the zero value of the process: the process is immediately reflected. Therefore, this version of diffusion process is a heavy-load approximation: it gives reasonable results if the utilization of the investigated station is close to 1, i.e. probability $p(0)$ of the empty system is negligible.

This inconvenience can be removed by the introduction of another limit condition at $x = 0$: *a barrier with instantaneous (elementary) jumps* [14]. When the diffusion process comes to $x = 0$, it remains there for a time exponentially distributed with a parameter λ_0 and then returns to $x = 1$. The time when the process is at $x = 0$ corresponds to the idle time of the system.

The diffusion equation becomes

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

where $p_0(t) = P[X(t) = 0]$. 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 barrier. The second equation makes the balance of the $p_0(t)$: the term $\lim_{x\to 0}[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)]$ gives the probability flow *into* the barrier and the term $\lambda p_0(t)$ represents the probability flow *out* of the barrier.

Our approach, see [5], 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 barrier at $x = 0$. This process starts at $t = 0$ from $x = x_0$ and ends when it attains the barrier. Its probability density function is easier to determine and has the following form [4],

$$\phi(x,t;x_0) = \frac{e^{\frac{\beta}{\alpha}(x-x_0)-\frac{\beta^2}{2\alpha}t}}{\sqrt{2\Pi\alpha t}}\left[e^{-\frac{(x-x_0)^2}{2\alpha t}} - e^{-\frac{(x+x_0)^2}{2\alpha t}}\right]\,. \qquad (3)$$

The density function of the first passage time from $x = x_0$ to $x = 0$ is

$$\begin{aligned}\gamma_{x_0,0}(t) &= \lim_{x\to 0}[\frac{\alpha}{2}\frac{\partial}{\partial x}\phi(x,t;x_0) - \beta\phi(x,t;x_0)] = \\ &= \frac{x_0}{\sqrt{2\Pi\alpha t^3}}e^{-\frac{(\beta t+1)^2}{2\alpha t}}\,. \qquad (4)\end{aligned}$$

Suppose that the process starts at $t = 0$ at a point x with density $\psi(x)$ and every time it comes to the barrier it stays there for a time given by a density function $l_0(x)$ and then reappears at $x = 1$. The total stream $\gamma_0(t)$ of probability mass that enters the barrier is

$$\begin{aligned}\gamma_0(t) &= p_0(0)\delta(t) + [1-p_0(0)]\gamma_{\psi,0}(t) + \\ &\quad \int_0^t g_1(\tau)\gamma_{1,0}(t-\tau)d\tau \qquad (5)\end{aligned}$$

where

$$\begin{aligned}\gamma_{\psi,0}(t) &= \int_0^\infty \gamma_{\xi,0}(t)\psi(\xi)d\xi\,,\\ g_1(\tau) &= \int_0^\tau \gamma_0(t)l_0(\tau-t)dt\,.\end{aligned}$$

The density function of the diffusion process with instantaneous returns is

$$f(x,t;x_0) = \phi(x,t;\psi) + \int_0^t g_1(\tau)\phi(x,t-\tau;1)d\tau\,. \qquad (6)$$

For Laplace transforms of these equations we have

$$\begin{aligned}\bar\gamma_0(s) &= p_0(0) + [1-p_0(0)]\bar\gamma_{\psi,0}(s) + \bar g_1(s)\bar\gamma_{1,0}(s)\,,\\ \bar g_1(s) &= \bar\gamma_0(s)\bar l_0(s) \qquad (7)\end{aligned}$$

where

$$\bar\gamma_{x_0,0}(s) = e^{-x_0\frac{\beta+A(s)}{\alpha}}\,, \qquad \bar\gamma_{\psi,0}(s) = \int_0^\infty \bar\gamma_{\xi,0}(s)\psi(\xi)d\xi\,,$$

and then

$$\bar g_1(s) = \Big[p_0(0) + [1-p_0(0)]\bar\gamma_{\psi,0}(s)\Big]\frac{\bar l_0(s)}{1-\bar l_0(s)\bar\gamma_{1,0}(s)}\,. \qquad (8)$$

Equation (6) in terms of Laplace transform becomes

$$\bar f(x,s;x_0) = \bar\phi(x,s;\psi) + \bar g_1(s)\bar\phi(x,s;1)\,,$$

where

$$\begin{aligned}\bar\phi(x,s;x_0) &= \frac{\mathrm{e}^{\frac{\beta(x-x_0)}{\alpha}}}{A(s)}\left[\mathrm{e}^{-|x-x_0|\frac{A(s)}{\alpha}} - \mathrm{e}^{-|x+x_0|\frac{A(s)}{\alpha}}\right]\,,\\ \bar\phi(x,s;\psi) &= \int_0^\infty \bar\phi(x,s;\xi)\psi(\xi)d\xi\,, \quad A(s) = \sqrt{\beta^2+2\alpha s}\,.\end{aligned}$$

The inverse transforms of these functions could only be found numerically. For this purpose we use the Stehfest's algorithm

[28]: for any fixed argument t, the function $f(t)$ is obtained from its transform $\bar{f}(s)$ as

$$f(t) = \frac{\ln 2}{2} \sum_{i=1}^{N} V_i \, \bar{f}\left(\frac{\ln 2}{t} i\right) , \qquad (9)$$

where

$$V_i = (-1)^{N/2+i} \sum_{k=\lfloor \frac{i+1}{2} \rfloor}^{\min(i,N/2)} \frac{k^{N/2+1}(2k)!}{(N/2-k)!k!(k-1)!(i-k)!(2k-i)!} .$$

N is an even integer and its choice depends on a computer precision; we used $N = 12 - 40$.

The above transient solution of $G/G/1$ model assumes that the parameters of this model are constant. If they are evolving, we should define the time-periods where they can be considered constant and solve diffusion equation within these intervals separately. A transient solution obtained at the end of an interval serves as the initial condition for the next interval.

B. Limited queue: G/G/1/N station, transient solution

In the case of G/G/1/N station, the second barrier should be placed at $x = N$. When the process comes to this barrier, it stays there for a time corresponding to the period when the queue is full and incoming customers are lost and then, after the completion of the current service, the process jumps to $x = N - 1$.

The model equations become [14]

$$\begin{aligned}
\frac{\partial f(x,t;x_0)}{\partial t} &= \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta \frac{\partial f(x,t;x_0)}{\partial x} + \\
&\quad + \lambda_0 p_0(t)\delta(x-1) + \lambda_N p_N(t)\delta(x-N+1) , \\
\frac{dp_0(t)}{dt} &= \lim_{x\to 0}\left[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)\right] - \lambda_0 p_0(t) , \\
\frac{dp_N(t)}{dt} &= \lim_{x\to N}\left[-\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} + \beta f(x,t;x_0)\right] - \\
&\quad \lambda_N p_N(t) , \qquad (10)
\end{aligned}$$

where $\delta(x)$ is Dirac delta function.

The density function $f(x,t;x_0)$ is obtained in the similar way as previously. First we obtain the density $\phi(x,t;x_0)$ of the diffusion process with two absorbing barriers at $x = 0$ and $x = N$, started at $t = 0$ from $x = x_0$, cf. [4]

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

where

$$\begin{aligned}
a_n &= \exp\left[\frac{\beta x_n'}{\alpha} - \frac{(x - x_0 - x_n' - \beta t)^2}{2\alpha t}\right] \\
b_n &= \exp\left[\frac{\beta x_n''}{\alpha} - \frac{(x - x_0 - x_n'' - \beta t)^2}{2\alpha t}\right]
\end{aligned}$$

and $x_n' = 2nN$, $x_n'' = -2x_0 - x_n'$.

If the initial condition is defined by a function $\psi(x)$, $x \in (0,N)$, $\lim_{x\to 0}\psi(x) = \lim_{x\to N}\psi(x) = 0$, 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

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

Densities $\gamma_0(t)$, $\gamma_N(t)$ of the probability that at time t the process enters to $x = 0$ or $x = N$ are

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

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

$$\gamma_{1,0}(t) = \lim_{x\to 0}\left[\frac{\alpha}{2}\frac{\partial\phi(x,t;1)}{\partial x} - \beta\phi(x,t;1)\right] . \qquad (11)$$

The functions $\gamma_{\psi,0}(t)$, $\gamma_{\psi,N}(t)$ denote densities of the 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 $x = 0$ or $x = N$.

Finally, we can express $g_1(t)$ and $g_N(t)$ with the use of functions $\gamma_0(t)$ and $\gamma_N(t)$:

$$\begin{aligned}
g_1(\tau) &= \int_0^\tau \gamma_0(t) l_0(\tau - t)dt , \\
g_{N-1}(\tau) &= \int_0^\tau \gamma_N(t) l_N(\tau - t)dt ,
\end{aligned}$$

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 solutions, given by [14]:

$$f(x) = \begin{cases}
\frac{\lambda p_0}{-\beta}(1 - e^{zx}) & \text{for} \quad 0 < x \le 1 , \\
\frac{\lambda p_0}{-\beta}(e^{-z} - 1)e^{zx} & \text{for} \quad 1 \le x \le N-1 , \\
\frac{\mu p_N}{-\beta}(e^{z(x-N)} - 1) & \text{for} \quad N-1 \le x < N ,
\end{cases} \qquad (12)$$

where $z = \frac{2\beta}{\alpha}$ and p_0, p_N are determined through normalization

$$\begin{aligned}
p_0 &= \lim_{t\to\infty} p_0(t) = \{1 + \varrho e^{z(N-1)} + \frac{\varrho}{1-\varrho}[1 - e^{z(N-1)}]\}^{-1} , \\
p_N &= \lim_{t\to\infty} p_N(t) = \varrho p_0 e^{z(N-1)} .
\end{aligned}$$

Customer classes. As proposed in [15], the input stream λ can be composed of K classes of customers and $\lambda = \sum_{k=1}^{K} \lambda^{(k)}$ (all parameters concerning class k have an upper index with brackets) then the joint service time pdf is defined as

$$b(x) = \sum_{k=1}^{K} \frac{\lambda^{(k)}}{\lambda} b^{(k)}(x) \, ,$$

hence

$$\frac{1}{\mu} = \sum_{k=1}^{K} \frac{\lambda^{(k)}}{\lambda} \frac{1}{\mu^{(k)}} \, ,$$

and

$$C_B^2 = \mu^2 \sum_{k=1}^{K} \frac{\lambda_{(k)}}{\lambda} \frac{1}{\mu_{(k)}^2} ({C_B^{(k)}}^2 + 1) - 1 \, .$$

We assume that the input streams of different class customers are mutually independent, the number of class k customers that arrived within sufficiently long period of time is normally distributed with variance $\lambda^{(k)} {C_A^{(k)}}^2$; the sum of independent randomly distributed variables also has normal distribution with variance which is the sum of composing variances, hence

$$C_A^2 = \sum_{k=1}^{K} \frac{\lambda^{(k)}}{\lambda} {C_A^{(k)}}^2 \, . \qquad (13)$$

The above parameters yield α, β of the diffusion equation; function $f(x)$ approximates the distribution $p(n)$ of customers of all classes present in the queue: $p(n) \approx f(n)$ and the probability that there are $n^{(k)}$ customers of class k is

$$p_k(n^{(k)}) = \sum_{n=n^{(k)}}^{N} \left[p(n) \binom{n}{n^{(k)}} \left(\frac{\lambda^{(k)}}{\lambda} \right)^{n^{(k)}} \left(1 - \frac{\lambda^{(k)}}{\lambda} \right)^{n-n^{(k)}} \right] \, ,$$

$$k = 1, \ldots, K \, .$$

Numerical examples. We consider a G/G/1/30 queue (in fact, it is M/M/1/30 queue, as we assume $C_A^2 = C_B^2 = 1$). In Example 1 the input rate $\lambda(t)$ is varying in time as presented in Fig. 1. It represents a typical TCP flow with additive increases and multiplicative decreases in the case of packet losses, the range of time is [0,100] time units. In computations, the values of diffusion parameters are changed each 0.5 time unit. Figs. 2 - 5 present the main results of the diffusion model compared with the simulation results (in the latter case it is the average of 500 000 independent runs).

In Figs. 2 and 3 display the same numerical results concerning the values of the mean queue. Fig. 2 displays them in linear scale and Fig. 3 does it in logarithmic scale, to see better the errors of the diffusion approximation: they become visible for very small mean queue values, i.e. less then 0.001 at this model. Next figures present the probability $p(0,t)$ of the empty queue as a function of time, following time-dependent

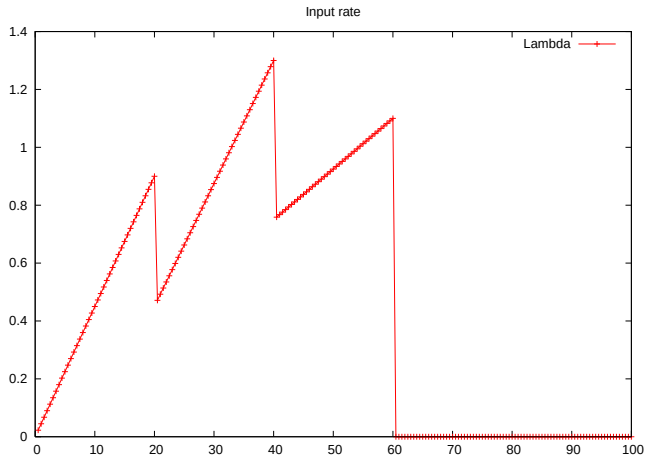

Fig. 1. Example 1: Input traffic intensity $\lambda(t)$.

input, Fig. 4, and the probability $p(N,t)$ that the queue is full, i.e. saturated, and rejects the arriving customers, Fig. 5.

In Example 2 the input rate is periodically varying between values 0.25 and 5. Figs. 6, 7, 8 display the same kind of results as previously: the mean number of customers, the probability of the empty and the probability of the saturated queue as a function of time. All results prove an almost perfect match of diffusion and simulation results. All simulations in the article have been performed with the use of OMNET++ [26].

III. Open network of G/G/1/N queues

The diffusion steady state model of an open network of G/G/1 or G/G/1/N queues was presented in [15]. Below we present its short summary. Let M be the number of stations, the throughput of station i is, as usual, obtained from traffic equations

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^{M} \lambda_j r_{ji} \, , \qquad i = 1, \ldots, M, \qquad (14)$$

where r_{ji} is routing probability between station j and station i; λ_{0i} is external flow of customers coming from outside of network.

The second moment of interarrival time distribution is obtained from two systems of equations; the first defines C_{Di}^2, the squared coefficient of variation of interdeparture times distribution at station i, as a function of C_{Ai}^2 and C_{Bi}^2; the second defines C_{Aj}^2 as another function of C_{D1}^2, ..., C_{DM}^2 :

1) The formula (15) defining the density function $d_i(x)$ of interdeparture times at station i is exact for $M/G/1$, $M/G/1/N$ stations and is approximate in the case of non-Poisson input [3]

$$d_i(x) = \varrho_i b_x(t) + (1 - \varrho_i) a_i(x) * b_i(x) \, , \qquad i = 1, \ldots, M, \qquad (15)$$

where * denotes the convolution operation. From (15) we get

$$C_{Di}^2 = \varrho_i^2 C_{Bi}^2 + C_{Ai}^2 (1 - \varrho_i) + \varrho_i (1 - \varrho_i) \, . \qquad (16)$$

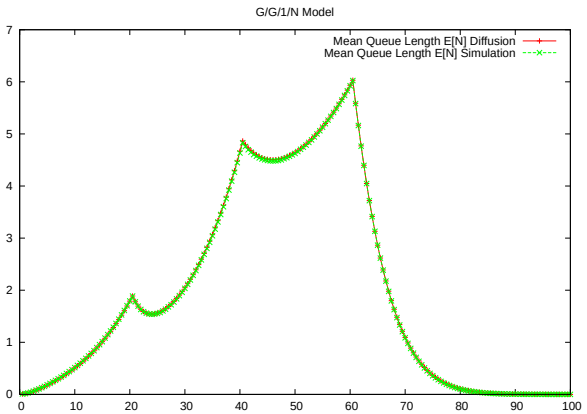

Fig. 2. Example 1: The mean number of customers as a function of time; diffusion approximation and simulation results. .

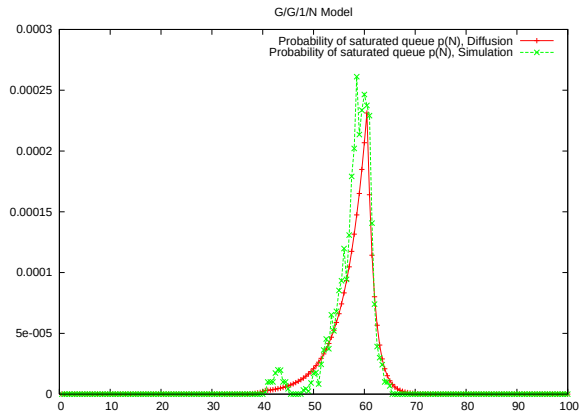

Fig. 5. Example 1: The probability $p(N,t)$ of the saturated queue, diffusion approximation and simulation results.

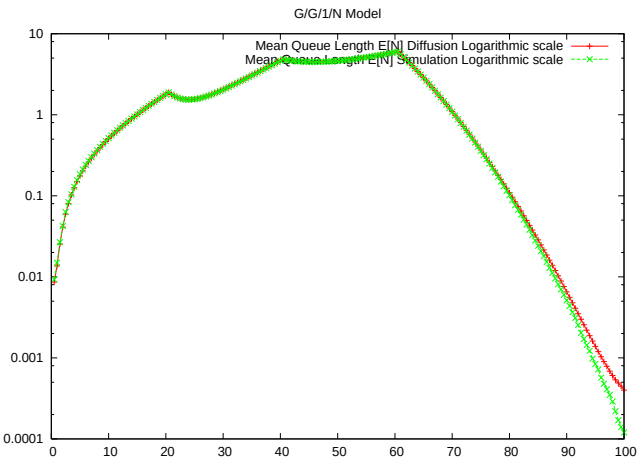

Fig. 3. Example 1: The mean number of customers (logarithmic scale) as a function of time; diffusion approximation and simulation results. .

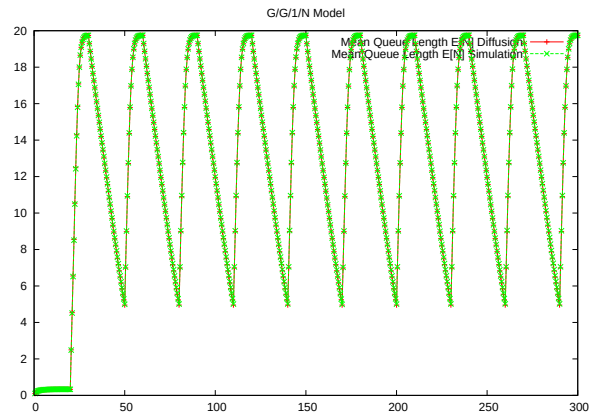

Fig. 6. Example 2: The mean number of customers as a function of time; diffusion approximation and simulation results. .

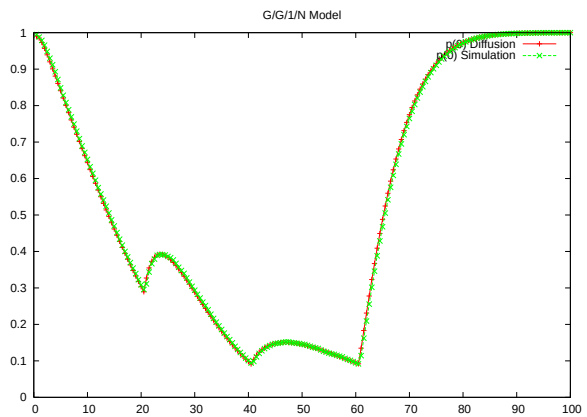

Fig. 4. Example 1: The probability $p(0,t)$ of the empty queue, diffusion approximation and simulation results.

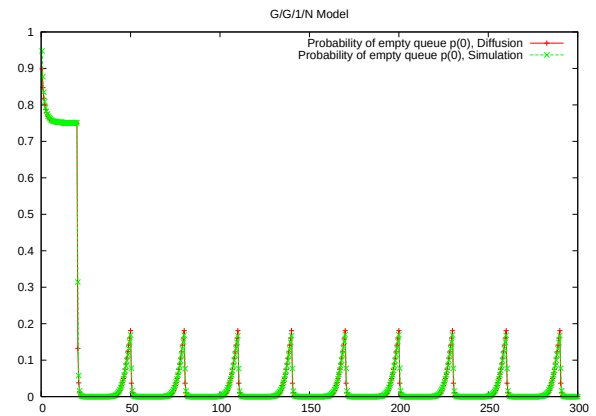

Fig. 7. Example 2: The probability $p(0,t)$ of the empty queue, diffusion approximation and simulation results.

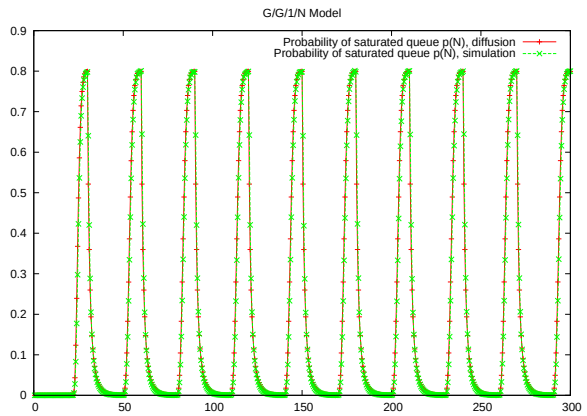

Fig. 8. Example 2: The probability $p(N,t)$ of the saturated queue, diffusion approximation and simulation results.

2) Customers leaving station i choose station j with probability r_{ij}: intervals between customers passing this way have the pdf $d_{ij}(x)$

$$\begin{aligned} d_{ij}(x) &= d_i(x)r_{ij} + d_i(x) * d_i(x)(1-r_{ij})r_{ij} + \\ &\quad + d_i(x) * d_i(x) * d_i(x)(1-r_{ij})^2 r_{ij} + \cdots \end{aligned}$$

or, after Laplace transform,

$$\begin{aligned} \bar{d}_{ij}(s) &= \bar{d}_i(s)r_{ij} + \bar{d}_i(s)^2(1-r_{ij})r_{ij} + \\ &\quad + \bar{d}_i(s)^3(1-r_{ij})^2 r_{ij} + \cdots = \frac{r_{ij}\bar{d}_i(s)}{1-(1-r_{ij})\bar{d}_i(s)} , \end{aligned}$$

hence

$$E[D_{ij}] = \frac{1}{\lambda_i r_{ij}} , \qquad C^2_{Dij} = r_{ij}(C^2_{Di} - 1) + 1 . \tag{17}$$

$E[D_{ij}]$, C^2_{Dij} refer to interdeparture times; the number of customers passing from station i to j in a time interval t has approximately normal distribution with mean $\lambda_i r_{ij} t$ and variation $C^2_{Dij}\lambda_i r_{ij} t$. The sum of streams entering station j has normal distribution with mean

$$\lambda_j t = [\sum_{i=1}^{M} \lambda_i r_{ij} + \lambda_{0j}]\, t$$

and variance

$$\sigma^2_{Aj} t = \{\sum_{i=1}^{M} C^2_{Dij}\lambda_i r_{ij} + C^2_{0j}\lambda_{0j}\} t ,$$

hence

$$C^2_{Aj} = \frac{1}{\lambda_j}\sum_{i=1}^{M} r_{ij}\lambda_i[(C^2_{Di}-1)r_{ij}+1] + \frac{C^2_{0j}\lambda_{0j}}{\lambda_j} . \tag{18}$$

Parameters λ_{0j}, C^2_{0j} represent the external stream of customers.

For K classes of customers with routing probabilities $r^{(k)}_{ij}$ (let us assume for the sake of simplicity that the customers do not change their classes) we have

$$\lambda^{(k)}_i = \lambda^{(k)}_{0i} + \sum_{j=1}^{M} \lambda^{(k)}_j r^{(k)}_{ji} , \qquad i = 1,\ldots,M; \;\; k = 1,\ldots,K, \tag{19}$$

and

$$\begin{aligned} C^2_{Di} &= \lambda_i \sum_{k=1}^{K} \frac{\lambda^{(k)}_i}{{\mu^{(k)}_i}^2}[{C^{(k)}_{Bi}}^2 + 1] + 2\varrho_i(1-\varrho_i) + \\ &\quad + (C^2_{Ai}+1)(1-\varrho_i) - 1 . \end{aligned} \tag{20}$$

A customer in the stream leaving station i belongs to class k with probability $\lambda^{(k)}_i/\lambda_i$ and we can determine ${C^{(k)}_{Di}}^2$ in the similar way as it has been done in Eqs. (17-18), replacing r_{ij} by $\lambda^{(k)}_i/\lambda_i$:

$${C^{(k)}_{Di}}^2 = \frac{\lambda^{(k)}_i}{\lambda_i}(C^2_{Di} - 1) + 1 ; \tag{21}$$

then

$$\begin{aligned} C^2_{Aj} &= \frac{1}{\lambda_j}\sum_{l=1}^{K}\sum_{k=1}^{K} r^{(k)}_{ij}\lambda_i \left[\left(\frac{\lambda^{(k)}_i}{\lambda_i}(C^2_{Di}-1)\right) r^{(k)}_{ij} + 1\right] \\ &\quad + \sum_{k=1}^{K} \frac{{C^{(k)}_{0j}}^2 \lambda^{(k)}_{0j}}{\lambda_j} . \end{aligned} \tag{22}$$

Eqs. (16), (18) or (20), (22) form a linear system of equations and allow us to determine C^2_{Ai} and, in consequence, parameters β_i, α_i for each station.

In the case of transient analysis, the time axis is divided into small intervals (equal e.g to the smallest mean service time) and at the beginning of each interval the Eqs. (14), (16), (18) are used to determine the input parameters of each station based on the values of $\varrho_i(t)$ obtained at the end of the precedent interval. A software tool was prepared and the examples below, concerning 2 network topologies, see Fig. 9 a,b, are computed with its use.

Example 3. The network is composed of the source and three stations in tandem, Fig. 9a. The source parameters are: $\lambda = 0.1$ $t \in [0,10]$, $\lambda = 4.0$ $t \in [10,20]$. Parameters of all stations are the same: $N_i = 10$, $\mu_i = 2$, $C^2_{Bi} = 1$, $i = 1,2,3$.

Fig. 10a presents mean queue lengths of stations in Model 1 as a function of time. Diffusion approximation is compared with simulation.

Example 4. The network topology is as in Fig. 9b. The characteristics of three sources and of one station are changing with time in the following pattern:
source A: $\lambda_A = 0.1$ for $t \in [0,10]$, $\lambda_A = 4.0$ for $t \in [10,21]$, $\lambda_A = 0.1$ for $t \in [21,40]$,
source B: $\lambda_B = 0.1$ for $t \in [0,11]$, $\lambda_B = 4.0$ for $t \in [11,20]$, $\lambda_B = 0.1$ for $t \in [20,40]$,
source C: $\lambda_C = 0.1$ for $t \in [0,15]$, $\lambda_C = 2.0$ for $t \in [15,22]$, $\lambda_C = 4.0$ for $t \in [22,30]$, $\lambda_C = 2.0$ for $t \in [30,31]$, $\lambda_C = 0.1$ for $t \in [31,40]$.

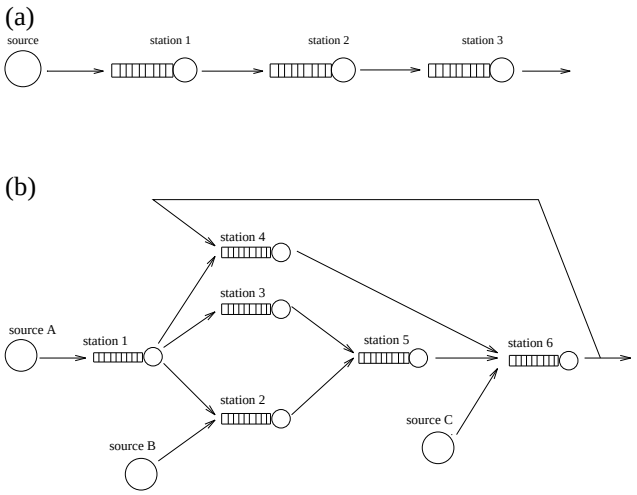

Fig. 9. Example 3 and 4 network topologies

Station 6: $\mu_6 = 2$ for $t \in [10, 15]$ and $t \in [31, 40]$; $\mu_6 = 4$ for $t \in [15, 31]$.

Other parameters are constant: maximum queue lengths $N_1 = N_4 = 10$, $N_3 = 5$, $N_2 = N_6 = 20$, $\mu_1 = \cdots = \mu_5 = 2$. Routing probabilities are: $r_{12} = r_{13} = r_{14} = 1/3$, $r_{64} = 0.8$. Initial state: $N_1(0) = 5$, $N_1(0) = 5$, $N_2(0) = 10$, $N_3(0) = 10$, $N_4(0) = 5$, $N_5(0) = 5$, $N_6(0) = 10$. The results in the form of mean queue lengths are presented and compared with simulation in Figs. 10, 11.

We observe that the output of queueing network models is not as good as in the case of single station models, but still reasonable.

IV. Diffusion approximation of the G/G/1 and G/G/1/N busy periods

Busy periods play an important role in the description of priority queues. During a busy period of higher priority customers, the sever is not available for lower priorities.

A. G/G/1 station

Let $\Gamma(t)$ and $\gamma(t)$ denote PDF and pdf of the busy period duration; for the M/M/1 system the function $\gamma(t)$ is known explicitly [19]

$$\gamma(t) = \frac{1}{t\sqrt{\varrho}} e^{(\lambda+\mu)t} I_1(2t\mu\varrho)$$

where $I_1(x) = \frac{x}{2} \sum_{k=0}^{\infty} \frac{1}{k!(k+1)!} \left(\frac{x}{2}\right)^{2k}$ is the modified Bessel function of the first kind and of order one. For M/G/1 system, a functional equation,

$$\bar{\gamma}(s) = \bar{B}(s + \lambda + \lambda\bar{\gamma}(s))$$

where $\bar{\gamma}(s) = \int_0^\infty e^{-st}\gamma(t)dt$, $\bar{B}(s) = \int_0^\infty e^{-st}b(t)dt$ are the Laplace transforms of $\gamma(t)$, $b(t)$, although impossible to invert in most cases, enables us to compute the moments of $\gamma(t)$, e.g.

$$E[\gamma] = -\frac{d}{ds}\bar{\gamma}(s)_{s=0} = \frac{1/\mu}{1-\varrho} \quad (23)$$

$$E[\gamma^2] = \frac{d^2}{ds^2}\bar{\gamma}(s)_{s=0} = \frac{E[b^2]}{(1-\varrho)^3} \quad (24)$$

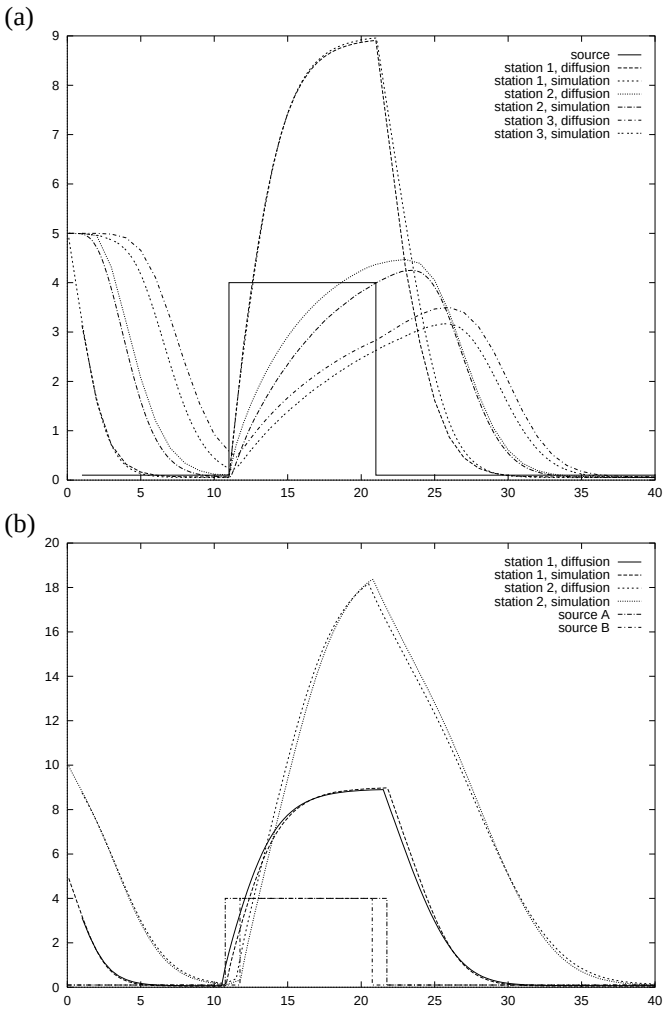

Fig. 10. (a) Example 3: The mean queue lengths of station1, station2 and station3 as a function of time — diffusion and simulation (100 000 repetitions) results; the source intensity $\lambda(t)$ is indicated. (b) Example 4: The mean queue lengths of station1 and station2 as a function of time — diffusion and simulation (100 000 repetitions) results; the source intensities $\lambda_A(t)$, $\lambda_B(t)$ are indicated.

The expression

$$\Gamma(t) = \int_0^t \sum_{n=1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{n-1}}{n!} b^{*n}(t)dt$$

where $b^{*n}(t)$ is the n-fold convolution of $b(t)$ with itself, could be helpful in numerical evaluation of the busy time distribution for the M/G/1 queue [19]. We know virtually nothing about $\Gamma(t)$, $\gamma(t)$ for the G/G/1 system. In the diffusion approximation, the busy period has a simple interpretation. In the case of G/G/1 system, it is just the first passage time from $x = 1$ to the absorbing barrier at $x = 0$, its pdf is given by Eq. (4) that yields

$$E[\gamma_{dif}] = \frac{1}{-\beta}, \qquad E[\gamma_{dif}^2] = -\frac{\alpha}{\beta^3} + \frac{1}{\beta^2}$$

which are exact results in the case of M/M/1 and M/G/1 systems.

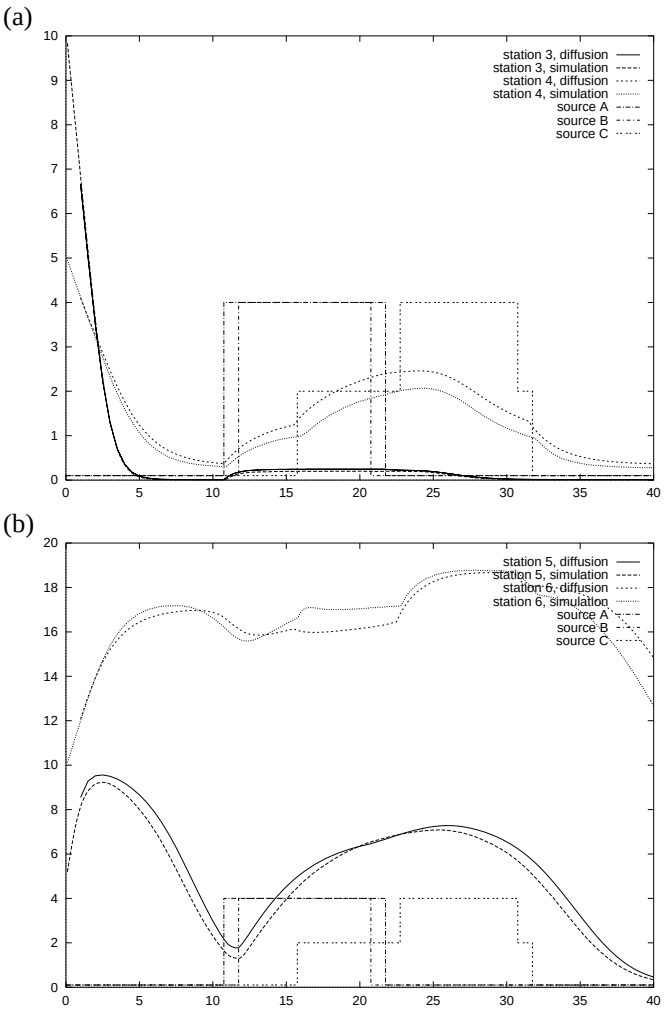

Fig. 11. Example 4: The mean queue lengths of station3 and station4 (a) and of station5 and station6 (b)

B. Busy period at G/G/1/N

As the process starting at $x = 1$ may visit the barrier at $x = N$ an unlimited number of times before coming to $x = 0$, the density function of the busy period is, preserving the previously used notation

$$\begin{aligned}\gamma(t) &= \gamma_{1,0}(t) + \gamma_{1,N}(t) * l_N(t) * \gamma_{N-1,0}(t) + \\ &+ \gamma_{1,N}(t) * l_N(t) * \gamma_{N-1,N}(t) * l_N(t) * \gamma_{N-1,0}(t) + \dots\end{aligned} \tag{25}$$

where * denotes the convolution operator, or

$$\begin{aligned}\bar{\gamma}(s) &= \bar{\gamma}_{1,0}(s) + \bar{\gamma}_{1,N}(s) l_N(s) \bar{\gamma}_{N-1,0}(s) + \\ &+ \bar{\gamma}_{1,N}(s) \bar{l}_N(s) \bar{\gamma}_{N-1,N}(s) \bar{l}_N(s) \bar{\gamma}_{N-1,0}(s) + \dots \\ &= \bar{\gamma}_{1,0}(s) + \frac{\bar{\gamma}_{1,N}(s) l_N(s) \bar{\gamma}_{N-1,0}(s)}{1 - \bar{\gamma}_{1,N}(s) l_N(s)} .\end{aligned} \tag{26}$$

Fig. 12 presents the comparison of busy period pdf given by diffusion approximation and simulation.

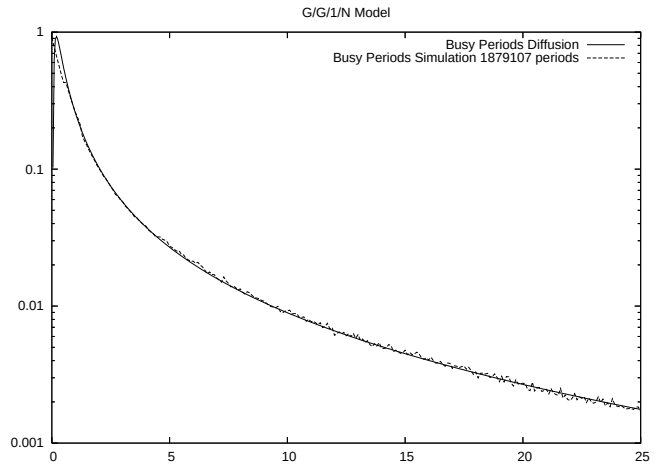

Fig. 12. M/M/1/20 queue, $\varrho = 0.75$, busy period diffusion approximation compared with simulation (histogram of over 1.8 million samples).

V. Diffusion approximation of preemptive - resume priority system

This paragraph introduces a diffusion model of a single server with priority preemptive - resume queuing discipline. Customers arriving to the system are divided into a certain number, say K, of classes. Each class is distinguished by its index k, $k = 1, \dots, K$, and has its own priority. The lower the number of the index, the higher the priority of the class. When a customer of class k is being served and a customer of class l, $l < k$ arrives, the current service is suspended and the service of the newcomer begins. After the completion of this service and the service of other, more privileged than class k customers, who have arrived meanwhile, the interrupted service is resumed at the point of suspension. Customers of the same priority class are served in the order of arrival. The presence of lower class customers is transparent to customers of a given class. We assume that interarrival times in the particular stream are characterized by parameters $\lambda^{(k)}$, $\sigma_A^{(k)^2}$ having the same meaning as λ, σ_A^2 in the case of one-class system. The service time of customers of class k has mean value $1/\mu^{(k)}$ and variance $\sigma_B^{(k)^2}$.

Following exactly the same procedure as for the FIFO system, we define: input process $E^{(K)}(t)$ as the total number of customers of all K classes who arrived to the system during the time period $[0, t]$, and the output process $H^{(K)}(t)$ as the number of customers of all K classes who left the system in $[0, t]$. Applying the central limit theorem and using the same arguments as for the first-come-first-served discipline, we can prove that these processes have approximately normal distributions if the period $[0, t]$ is sufficiently long and within a busy period of the server. The input process $E^{(K)}(t)$ consists of separate input processes $\varepsilon^{(k)}(t)$ for each class of customers:

$$E^{(K)}(t) = \sum_{k=1}^{K} \varepsilon^{(k)}(t)$$

The output process $H^{(K)}(t)$ can be described as

$$\begin{aligned} H^{(K)}(t) &= \sum_{k=1}^{K} \eta^{(k)}(t) \frac{\varrho^{(k)}}{1-R^{(k-1)}}(1-R^{(k-1)})\frac{1}{R^{(k)}} \\ &= \sum_{k=1}^{K} \frac{\varrho^{(k)}}{R^{(k)}}\eta^{(k)}(t) \end{aligned} \quad (27)$$

where $\eta^{(k)}(t)$ is the output process for the k-priority stream in the absence of other classes, $R^{(k)} = \sum_{t=1}^{(k)} \varrho^{(l)}$ and $\varrho^{(l)} = \lambda^{(l)}/\mu^{(l)}$. $R^{(k)}$ denotes the probability that the busy period for customers of classes $1, ..., k$ taken altogether is in progress, $\frac{\varrho^{(k)}}{1-R^{(k)}}$ denotes the probability that a customer of class k is present in the system. Class k has the $1 - R^{(k-1)}$ part of the server time at its disposal. $1 - R^{(k-1)}$ denotes the probability that there are no customers of priority higher than k present in the system. The total number of customers of classes $1, ..., K$ present in the system

$$N^{(K)}(t) = E^{(K)}(t) - H^{(K)}(t)$$

is changing and its changes during the time period $[0, t]$ have the mean $\beta^{(K)}t$ and the variance $\alpha^{(K)}t$,

$$\begin{aligned} \beta^{(K)} &= \sum_{k=1}^{K} \lambda^{(k)} - \sum_{k=1}^{K} \frac{\varrho^{(k)}}{R^{(k)}}\mu^{(k)}, \\ \alpha^{(K)} &= \sum_{k=1}^{K} \lambda^{(k)} {C_A^{(k)}}^2 + \sum_{k=1}^{K} \frac{\varrho^{(k)}}{R^{(k)}}\mu^{(k)} {C_B^{(k)}}^2, \\ {C_A^{(k)}}^2 &= {\lambda^{(k)}}^2 {\sigma_A^{(k)}}^2, \qquad {C_B^{(k)}}^2 = {\mu^{(k)}}^2 {\sigma_B^{(k)}}^2 \end{aligned}$$

and are approximately normally distributed. We replace the discrete-state process $N^{(K)}(t)$ by the continuous-state process $X^{(K)}(t)$ whose infinitesimal changes have normal distribution with the mean $\beta^K dt$ and the variance $\alpha^K dt$. Solving the diffusion equation with the same type of boundary conditions as defined earlier with the intensity of jumps from $x = 0$: $\Lambda^{(K)} = \sum_{k=1}^{K} \lambda^{(k)}$ we obtain the density function $f^{(K)}(x, t; x_0)$ for all classes considered together.

Let $v^{(K)}(n)$ denote the probability that n customers of class K are present in the system and $p^{(K-1)}(N-n)$ denote the probability that $N-n$ customers of all other classes are present in the system. Obviously,

$$p^{(K)}(N) = \sum_{n=0}^{N} p^{(K-1)}(N-n)v^{(K)}(n)$$

and similarily,

$$p^{(k)}(n) = \sum_{\nu=0}^{n} p^{(k-1)}(n-\nu)v^{(k)}(\nu), \quad k = 2, ..., K$$

or

$$v^k(n) = \frac{p^{(k)}(n) - \sum_{\nu=0}^{n-1} p^{(k-1)}(n-\nu)v^{(k)}(\nu)}{p^{(k-1)}(0)}, \quad k = 2, ..., K.$$

For the highest priority class

$$v^{(1)}(n) = p^{(1)}(n).$$

Thus, we know the distribution $v^k(n)$, the mean number of customers present in the system

$$E[n^{(k)}] = \sum_{\nu=0}^{\infty} v^{(k)}(\nu)\nu$$

and, by Little's result, the mean time they spend in the system

$$E[\mathcal{T}^{(k)}] = \frac{E[n^{(k)}]}{\lambda^{(k)}} \quad k = 1, ..., K$$

for each class of customers.

The inconvenience of this approach is the propagation of errors of the method. An alternative approach is to study the diffusion processes corresponding to the number of each class customers separately and to see the influence of higher classes on the queues of lower classes through the probability that the system is occupied by higher classes and thus is not able to serve the lower ones.

For example, if we take two classes, the first diffusion process corresponding to the priority class has parameters $\beta^{(1)} = \lambda^{(1)} - \mu^{(1)}$ and $\alpha^{(1)} = {\sigma_A^{(1)}}^2 {\lambda^{(1)}}^3 + {\sigma_B^{(1)}}^2 {\mu^{(1)}}^3$ and the second one, corresponding to the lower class which is served only in absence of the higher class, has the parameters

$$\begin{aligned} \beta^{(2)} &= \lambda^{(2)} - \mu^{(2)} p^{(1)}(0,t) \\ \alpha^{(2)} &= {\sigma_A^{(1)}}^2 {\lambda^{(1)}}^3 + {\sigma_A^{(2)}}^2 {\lambda^{(2)}}^3 + {\sigma_B^{(2)}}^2 {\mu^{(2)}}^3 p^{(1)}(0,t) + \\ &\quad + {\sigma_B^{(1)}}^2 {\mu^{(1)}}^3. \end{aligned}$$

Before the waiting time can be considered, we have to define the distribution of the completion time. The completion time is the time period between the beginning and the end of the service of any customer. On the highest priority level the completion time is equal to the service time, for the other classes it additionally includes the breaks caused by the service of more privileged customers. Let T be the service time of a customer of class k. If n customers of classes $1, ..., k-1$ arrive during the time T, the service will be interrupted n times, n has approximately normal distribution with the mean $\Lambda^{(k-1)}$ and the variance $\sum_{l=1}^{k-1} \lambda^{(l)} {C_A^{(l)}}^2 T$.

The duration of any of n breaks is distributed like the busy period $\gamma^{(k-1)}$ of the system serving customers of classes $1, ..., k-1$. The total time of breaks in T has the pdf

$$\varphi^{(k)}(t \mid T) = \sum_{n=0}^{\infty} p_{n|T} \gamma^{(k-1)(*n)}(t)$$

where $p_{n|T}$ is the probability of n breaks in T, $\gamma^{(k-1)(*n)}(t)$ is the n-fold convolution of $\gamma^{(k-1)}(t)$ with itself. Thus the pdf $c^{(k)}(t)$ of the completion time is

$$c^{(k)}(t) = \int_0^{\infty} b^{(k)}(t) \varphi^{(k)}(t - T \mid T) \mathbf{1}(t-T) dT,$$

where $\mathbf{1}(t) = 0$ for $t < 0$ and $\mathbf{1}(t) = 1$ for $t \geq 0$, and from its Laplace transform

$$c^{(k)}(s) = \int_0^{\infty} b^{(k)}(T) e^{-sT} \sum_{n=0}^{\infty} \{p_{n/T} [\bar{\gamma}^{(k)}(s)]^n\} dT$$

we obtain its moments $E = [c^{(k)}]$ and $E[(c^{(k)})^2]$:

$$E = [c^{(k)}] = -\frac{d}{ds}c^{(k)}(s)_{s=0} = \{E[\gamma^{(k-1)}]\Lambda^{(k-1)} + 1\}\frac{1}{\mu^{(k)}},$$

$$\begin{aligned} E[(c^{(k)})^2] &= \frac{d^2}{ds^2}c^{(k)}(s)_{s=0} = \\ &= E[\gamma^{(k-1)}]^2\Big[\Big(\sum_{l=1}^{(k-1)}\lambda^{(l)}{C_A^{(l)}}^2\Big)\frac{1}{\mu^{(k)}} + \\ &- \Lambda^{(k-1)}\frac{1}{\mu^{(k)}}\Big] + E[(\gamma^{(k-1)})^2]\Lambda^{(k-1)}\frac{1}{\mu^{(k)}} + \\ &+ E[\gamma^{(k-1)}]E[(b^{(k)})^2]\Lambda^{(k)} \cdot \\ &\cdot\{E[\gamma^{(k-1)}]\Lambda^{(k-1)} + 2\} + E[(b^{(k)})^2]. \end{aligned}$$

When all input streams are Poisson, i.e. ${C_A^{(l)}}^2 = 1$, $l = 1, ..., k$ the results are identical to the exact formula given for this case in [18]. Finally, we can define the mean waiting time for every priority level as

$$E[w^{(k)}] = \frac{E[n^{(k)}]}{\lambda^{(k)}} - E[c^{(k)}].$$

If we intend to consider a network of servers, we are obliged to determine the output stream of each server. In the case of priority queues we extend the approach used previously for a network of G/G/1/N stations, see Eqs. (16), (18) or (20), (22).

Let us denote $d^{(k)}(t)$ the pdf of interdeparture times in the stream of class k customers, it can be expressed as

$$\begin{aligned} d^{(k)}(t) &= \frac{\varrho^{(k)}}{1-R^{(k-1)}}c^{(k)}(t) + \left(1 - \frac{\varrho^{(k)}}{1-R^{(k-1)}}\right) \\ &\times[(1-R^{(k-)1})a^{(k)}(t) * c^{(k)}(t) \\ &+R^{(k-1)}a^{(k)}(t) * \gamma^{(k-1)}(t) * c^{(k)}(t)]. \quad (28) \end{aligned}$$

The components of this expression correspond to three situations, possible after the departure of any customer of class k:
– the next customer of the class k is in the system (it occurs with probability $\frac{\varrho^{(k)}}{1-R^{(k-1)}}$) and will leave it after its completion time,
– there are no customers of this class in the system and we shall wait the time described by $a^{(k)}(t)$, the interarrival time pdf (when the input is non-Poisson it is merely an approximation) until it appears and enters the server,
– no customer of class k is present in the system and a customer of higher class comes before him, so the busy period $\gamma^{(k-1)}$ must be terminated first.
The mean interdeparture time is obviously the same as the mean interarrival time and from the above (28), where in turn the densities of busy periods at each priority level are given by expressions of the type (25) and their moments are obtained from (26), we calculate the squared coefficient of the variation of interdeparture times at each priority customers, needed to integrate a single priority station into a network of such stations. The final formula is as follows:

$${C_D^{(k)}}^2 = \sum_{l=1}^{k} h^{(k,l)}{C_A^{(l)}}^2 + \psi^{(k)} \quad (29)$$

where

$$h^{(k,l)} = \begin{cases} \left(\zeta^{(k,l)} + \frac{1-R^{(k)}}{1-R^{(k-1)}}R^{(k-1)}g^{(k-1,l)}\right)(\lambda^{(k)})^2, & l < k, \\ \frac{1-R^{(k)}}{1-R^{(k-1)}}, & l = k, \end{cases}$$

and

$$\begin{aligned} \zeta^{(k,l)} &= \frac{\lambda^{(l)}}{\mu^{(k)}(\beta^{(k-1)})^2} + g^{(k-1,l)}\frac{\Lambda^{(k-1)}}{\mu^{(k)}}, \\ g^{(k,l)} &= \frac{1}{(\beta^{(k)})^3}, \end{aligned}$$

$$\begin{aligned} \psi^{(k)} &= \chi^{(k)}(\lambda^{(k)})^2 + \frac{1-R^{(k)}}{1-R^{(k-1)}}\Big\{1 + R^{(k-1)}e^{(k-1)}(\lambda^{(k)})^2 \\ &+2\varrho^{(k)}\left(1 - \frac{\Lambda^{(k-1)}}{\beta^{(k-1)}}\right) + \\ &-\frac{\lambda^{(k)}R^{(k-1)}}{\beta^{(k-1)}}\left[1 + 2\varrho^{(k)}\left(1 - \frac{\Lambda^{(k-1)}}{\beta^{(k-1)}}\right)\right]\Big\} - 1, \\ \chi^{(k)} &= \frac{{C_B^{(k)}}^2 + 1}{(\mu^{(k)})^2}\frac{\Lambda^{(k-1)}}{\beta^{(k-1)}}\left(\frac{\Lambda^{(k-1)}}{\beta^{(k-1)}} - 2\right) - \\ &\frac{\Lambda^{(k-1)}}{(\beta^{(k-1)})^2\mu^{(k)}} + e^{(k-1)}\frac{\Lambda^{(k)}}{\mu^{(k)}}\frac{{C_B^{(k)}}^2 + 1}{(\mu^{(k)})^2}, \\ e^{(k)} &= \frac{1}{(\beta^{(k)})^2} - \frac{1}{(\beta^{(k)})^3}\sum_{l=1}^{k}\frac{\varrho^{(l)}}{R^{(k)}}\mu^{(l)}{C_B^{(l)}}^2. \end{aligned}$$

The equation (29) corresponds to (16): it defines how the variation of the interdeparture times of the class-*k* customers depends on the variations of the interarrival times of all classes that may influence the output of this class. The parameters of service time distributions are hidden in the coefficients of the equation. Similarly, the extension of the equation (18) defining the squared coefficient of variation of interarrival times in the flow of class-*l* customers coming to station j has the following form

$${C_{Aj}^{(l)}}^2 = \frac{1}{\lambda_j^{(l)}}\sum_{i=1}^{M}\sum_{k=1}^{K} r_{ij}^{(kl)}\lambda_i^{(k)}[({C_{Di}^{(k)}}^2 - 1)r_{ij}^{(kl)} + 1] + \frac{{C_{0j}^{(l)}}^2\lambda_{0j}^{(l)}}{\lambda_j^{(l)}}, \quad (30)$$

where $r_{ij}^{(kl)}$ is the probability that a class-*k* customer leaving station i goes directly to station j having there class-*l* priority. Equations (29) and (30) taken together determine the input flow parameters for each class and each station, allowing us to analyze each station separately. As usual, in the case of transient states, all parameters should be considered constant at small intervals and all model equations should be solved for these parameters to define conditions at the beginning of the next interval.

Numerical examples Consider a server with two priority levels. In Example 5, the priority customers come with intensity $\lambda^{(1)} = 0.4$ during intervals $t \in [0, 10], [20, 30], [40, 50]$, etc. Otherwise $\lambda^{(1)} = 0$. The intensity of non-priority customers is constant, $\lambda^{(2)} = 0.4$. In Example 6 the server utilization is higher and the bursts of priority input stream are longer, $\lambda^{(1)} = 1.2$ during intervals $t \in [0, 20], [40, 60], [80, 100]$,

and $\lambda^{(1)} = 0$ between these intervals while the second class intensity is constant, $\lambda^{(2)} = 0.5$.

The queue capacities are in both cases limited: $N^{(1)} = N^{(2)} = 20$. To validate the diffusion model by the comparison of its results with the exact ones obtained with Markov chain model solved numerically, we assume Poisson input streams and exponential service time distributions for both types of customers, $\mu^{(1)} = \mu^{(2)} = 1$. Figs. 13, 14 refer to the Example 5. Fig. 13 displays the mean number of customers of each class as a function of time, given by diffusion model and by the corresponding Markov model. Fig. 14 compares the total number of customers of both classes.

Figs. 15, 16 give the same results for the Example 6.

Of course, diffusion approximation gives not only the mean values but also the queue distributions; e.g. Fig. 17 presents, for Example 5, exact and estimated probabilities $p^{(1)}(0,t)$, $p^{(2)}(0,t)$ that the queues of class 1 or class 2 are empty. Fig. 18 presents for Example 6 exact and estimated probabilities $p^{(1)}(N^{(1)},t)$, $p^{(2)}(N^{(2)},t)$ that the queues of class 1 or class 2 are saturated. For all computations we considered constant parameters inside subintervals of the 0.1 time unit length. In all cases the errors observed for priority queues are smaller then for non-priority ones. It is natural, the errors of the second class queue accumulate the errors of both classes.

VI. Conclusions

The article presents an adaptation of the diffusion approximation model with absorbing barriers to the analysis of transient states of queueing models. The method was applied previously to G/G/1/N service stations, here we also present the case of preemptive-resume priority queues.

In the article, we demonstrate how the diffusion approximation formalism is applied to study transient and behavior of G/G/1 and G/G/1/N non-priority and preemptive-priority models. The way we switch from one model to another demonstrates the flexibility of the method. Also the preemptive discipline can be easily converted to non-preemptive queueing discipline. Also the introduction of self-similar traffic is possible: as we change the diffusion parameters each small time-interval, we can modulate them to reflect self-similarity and long-term correlation of the traffic. Some other applications may be considered: recently we have used the diffusion approximation to estimate transfer times inside a sensor network [9], to model the performance of leaky-bucket algorithm as well as to study the work of call centers [10], and to investigate the stability of TCP connections with IP routers having AQM queues [8]. In the first case the diffusion process reflects the distance defined as the number of hops between the transmitted packet and its destination (sink). Owing to the introduction of the transient state analysis, the model captures more parameters (time-dependent and heterogeneous transmission, the presence of losses specific to each hop) of a sensor network transmission time than the already existing models of this type, also based on the diffusion approximation [16]. It also gives more detailed results: the density function of a packet travel time instead of its mean value. In the second

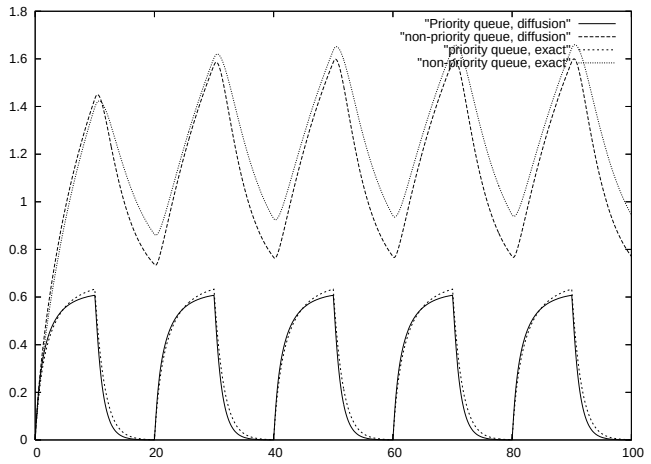

Fig. 13. Example 5: Mean number of customers as a function of time for the first and second priority levels, diffusion approximation and exact results

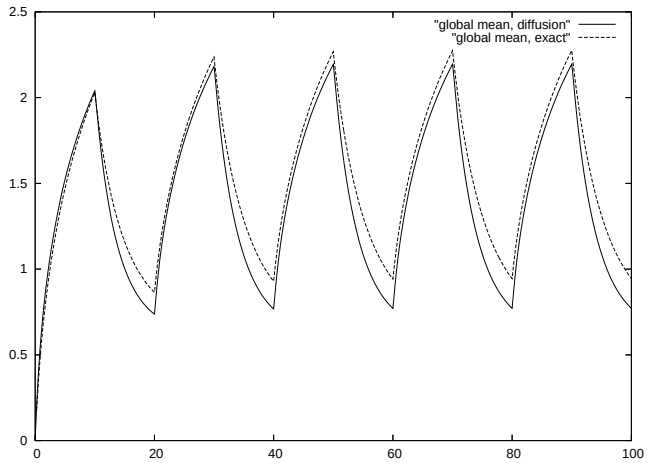

Fig. 14. Example 5: Global mean number of customers as a function of time, diffusion approximation and exact results.

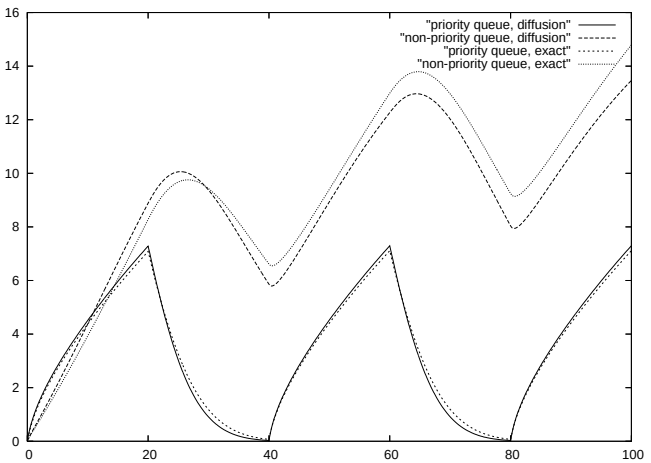

Fig. 15. Example 6: The mean number of customers as a function of time for the first and second priority levels, diffusion approximation and exact results.

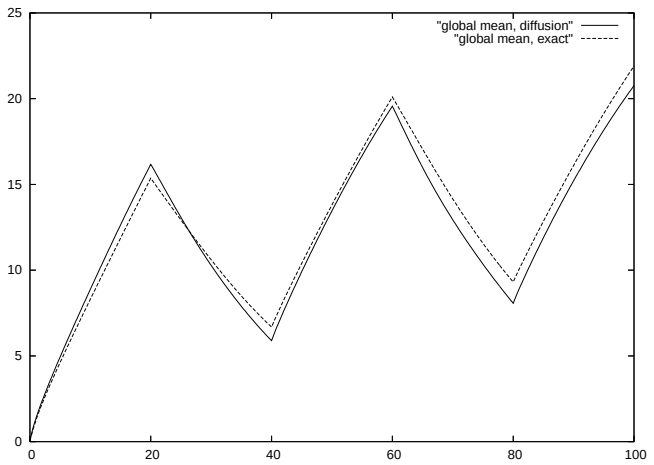

Fig. 16. Example 6: Global mean number of customers as a function of time, diffusion approximation and exact results.

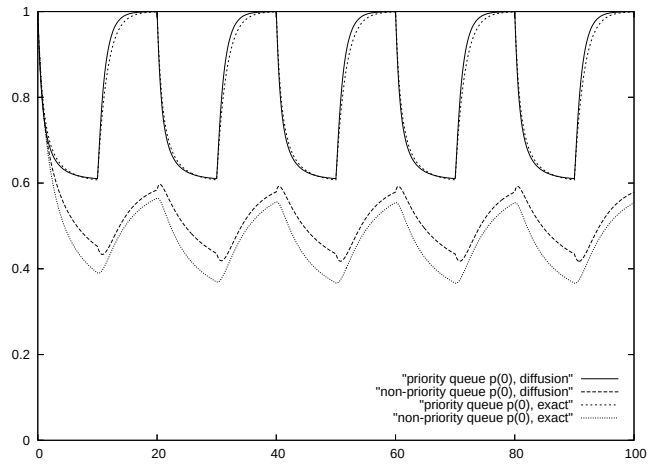

Fig. 17. Example 5: Probabilities $p^{(1)}(0,t)$, $p^{(2)}(0,t)$ that the queues of class 1 or class 2 are empty, diffusion approximation and exact results.

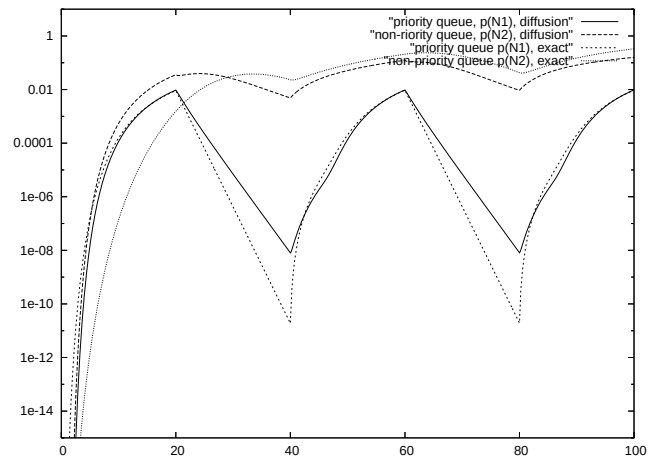

Fig. 18. Example 6: Probabilities $p^{(1)}(N^{(1)},t)$, $p^{(2)}(N^{(2)},t)$ that the queues of class 1 or class 2 are saturated, diffusion approximation and exact results. Logarithmic scale is chosen for better presentation of very small probabilities.

case, the introduction of state-dependent diffusion coefficients enables us to study transient states of parallel stations of G/G/N/N type. The third position proposes a diffusion model of a queue with RED (Random Early Detection) mechanism and considers the dynamics of TCP connections having RED queue in the congested router. Also the application of diffusion approximation to model wireless networks based on IEEE 802.11 standard gives promising results, [12].

Numerical examples, where the quantitative results of diffusion approximations are compared with simulations or the numerical solutions of corresponding Markov chain models, indicate acceptable level of errors of the proposed approach.

Of course, there are several ways we can analyze transient states in queueing models. In recent years we have put a considerable effort to master their use as efficient tools that give sound numerical results. Each of them has its advantages and disadvantages. Firstly, we can use simulation models. In this purpose we have developed an extension of OMNET++ (a popular simulation tool written in C++, [26]) allowing the simulation of transient state models. In particular, random generators were modified to make the changes of their parameters as a function of time possible, a new software was added to collect the statistics of multiple runs and to aggregate it. We used this module to validate the diffusion approximation results. Basically, the simulation run in a transient state investigation should be repeated sufficient number of times (e.g. 500 000 in our examples) and the results for a fixed time should be averaged. As the number of repetitions is high, the estimation of errors is easy (confidence interval) on the basis of normal distribution. However, the number of repetitions is related to the value of the investigated probabilities and in the case of rare events should be high and it increases the simulation time (typically in some of our examples, 5 minuts of computations for a diffusion model are compared to 24 hours of simulations, on a standard PC station).

The other way to model transient states is to create a Markov chain model and to solve it numerically. This approach, also combined with the use of stochastic Petri nets, gained already a considerable attention of researches, e.g. [13], [30], [27] and a number of software tools, e.g. SHARP, PEPSY, SNMP, MOSES [2] or XMARCA [20] was implemented. The numerical problems of solving very large systems of equations related to Markov models were thoroughly studied, e.g. [21]. This effort concerned mainly steady-state models. For several years we have developed a software to construct and to solve very large (having millions of states) Markov chains relating to queueing models and we have adopted suitable numerical methods and distributed algorithms. In the case of transient states, the implementation is based on the reduction of state space due to Arnoldi's orthogonal projection into the Krylov subspace [29]. We have also used Markov model to evaluate the errors in the case of the priority model presented here. Naturally, the usability of the approach depends on the size of the considered model, and it is relatively easy to go beyond the limit number, i.e. some tens of millions, of tractable states. We are still working on more powerful Markovian modules using

distributed algorithms and run on a cluster architecture.

Another well-known approach of modeling is the fluid-flow approximation where only the mean values of traffic intensity and service intensity are considered. Compared to the diffusion approximation, the model is simple: instead of partial differential equations of second order, the ordinary first-order linear differential equations are used. Due to its simplicity, it gained much interest in the analysis of transient states in Internet and in investigation of stability of its connections, e.g. [24]. However, as we tested in [7], the errors of the fluid-flow approximation in modeling queues dynamics are considerably larger than in the case of diffusion approximation which is a second-order approximation, where not only the mean values but also the variances of flow changes and of service times are considered.

Therefore we consider the diffusion approximation as a very convenient tool in the analysis of transient states queueing models in performance evaluation of computer and communication networks.

VII. Acknowledgements

This research was partially financed by the Polish Ministry of Science and Education grant N517 025 31/2997. The Authors thank the Referees for their valuable remarks.

References

[1] T. Atmaca, T. Czachórski, F. Pekergin, "A Diffusion Model of the Dynamic Effects of Closed-Loop Feedback Control Mechanisms in ATM Networks", 3rd IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, UK, 4-7th July 1995.

[2] G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi, "Queueing networks and Markov chains: modeling and performance evaluation with computer science applications", Wiley-Interscience, New York, 1998.

[3] P.J. Burke, The Output of a Queueing System, Operations Research, vol. 4, no. 6, pp. 699-704.

[4] R. P. Cox, H. D. Miller, The Theory of Stochastic Processes, Chapman and Hall, London (1965).

[5] T. Czachórski, "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.

[6] T. Czachórski, F. Pekergin, "Transient diffusion analysis of cell losses and ATM multiplexer behaviour under correlated traffic", 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, UK, 21-23 july 1997.

[7] T. Czachórski, J. M. Fourneau, F. Pekergin, "The Dynamics of Cell Flow and Cell Losses in ATM Networks Modelled by a Diffusion Process", Bulletin of Polish Academy of Sciences, Technical Sciences vol. 43, no. 4, pp. 538-548, 1995.

[8] T. Czachórski, K. Grochla, F. Pekergin, "Stability and Dynamics of TCP-NCR(DCR) Protocol", LNCS no. 4396, Wireless Systems and Mobility in Next Generation Internet, Springer-Verlag 2007.

[9] T. Czachórski, K. Grochla, F. Pekergin, "Un modèle d'approximation de diffusion pour la distribution du temps d'acheminement des paquets dans les réseaux de senseurs" Proc. of CFIP'2008, Les Arcs, 25-28 mars 2008, proceedings, edition electronique http://hal.archives-ouvertes.fr/CFIP2008.

[10] T. Czachórski, J.-M. Fourneau, T. Nycz, F. Pekergin, "Diffusion approximation model of multiserver stations with losses", Proc. of Third International Workshop on Practical Applications of Stochastic Modelling PASM'2008, Palma de Mallorca, 23rd September 2008, to appear also as an issue of Elsevier's ENTCS (Electronic Notes in Theoretical Computer Science).

[11] T. Czachórski, T. Nycz, F. Pekergin, "Transient states of priority queues – a diffusion approximation study", Proc. of The Fifth Advanced International Conference on Telecommunications AICT 2009 May 24-28, 2009 - Venice/Mestre, Italy.

[12] T. Czachórski, K. Grochla, T. Nycz, F. Pekergin, "A diffusion approximation model for wireless networks based on IEEE 802.11 standard" submitted to COMCOM Special Journal Issue on Heterogeneous Networks: Traffic Engineering and Performance Evaluation of Computer Communications.

[13] J. B. Dugan, K. S. Trivedi, R. Geist, V. F. Nicola, "Extended Stochastic Petri Nets: Applications and Analysis", in: E. Gelenbe(Hrsg.), Performance, North-Holland, 1984.

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

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

[16] E. Gelenbe, "'Travel delay in a large wireless ad hoc network"', 2nd Workshop on Spatial Stochastic Models of Wireless Networks (SPASWIN), Boston, 7th April (2006).

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

[18] K. N. Jaiswal, Priority Queues, Academic Press, New York 1968.

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

[20] R. L. Klevans, W. J. Stewart, From queueing networks to Markov chains: the XMARCA interface", Performance Evaluation, vol. 24, no. 1-2, pp. 23-46, 1995.

[21] W. Knottenbelt, "Distributed task-based solution techniques for large Markov models", Proc. of NSMC '99 Zaragoza, Spain, 1999.

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

[23] S. S. Lavenberg, "Computer Performance Modeling Handbook", Academic Press, New York 1983.

[24] V. Misra, W.-B. Gong, D. Towsley: Fluid-based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED, ACM SIGCOMM 2000.

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

[26] OMNET++ Community Site www.omnetpp.org.

[27] R. A. Sahner, K. S. Trivedi, A. Puliafito, "Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the Sharpe Software Package", Kluwer 1996.

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

[29] W. Stewart, Introduction to the Numerical Solution of Markov Chains, Princeton University Press, Chichester, West Sussex 1994.

[30] K. S. Trivedi, A. Puliafito, D. Logothetis, "From Stochastic Petri Nets to Markov Regenerative Stochastic Petri Nets", in: P. W. Dowd (Hrsg.), E. Gelenbe, (Hrsg.): Proc. of MASCOTS, IEEE Computer Society, 1995.