Queueing Models for Performance Evaluation of Computer Networks—Transient State Analysis

Tadeusz Czachórski

Abstract Queueing theory is a useful tool in design of computer networks and their performance evaluation. The literature concerning this subject is abundant. However, it is in general limited to the analysis of steady states. It means that flows of customers considered in models are constant and obtained solutions do not depend on time. It is in glaring contrast with the flows observed in real networks where the perpetual changes of traffic intensities are due to the nature of users, sending variable quantities of data, cf. multimedia traffic, and also due to the performance of traffic control algorithms which are trying to avoid congestion in networks, e.g. the algorithm of congestion window used in TCP protocol which is adapting the rate of the sent traffic to the observed losses or transmission delays. We discuss here the means used to analyse transient states in queueing models. In computer applications a mathematical model is useful only when it furnishes quantitative results. Therefore practical issues related to numerical side of models are of importance and are here discussed. We present three approaches—Markov models solved numerically, fluid flow approximation and diffusion approximation. A particular importance is given to the latter as the author has here over 20 year experience in development and application of this method. He is also convinced of the qualities of this approach—its flexibility to treat various variants of queueing models. Traffic intensity observed in computer networks have a complex stochastic nature that influences the network performances. We discuss also this side of implemented queueing models.

1 Introduction

Queueing theory has many applications but performance evaluation of computer networks and computer systems seems to be the most important one. The origins of queueing theory were also related to the transmission of information: first queueing

T. Czachórski (✉)
Institute of Theoretical and Applied Informatics PAS, Ul. Baltycka 5, 44-100 Gliwice, Poland
e-mail: tadek@iitis.pl

V.V. Mityushev and M. Ruzhansky (eds.), *Analytic Methods in Interdisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 116,
DOI 10.1007/978-3-319-12148-2_4

models were proposed a hundred years ago by Agner Krarup Erlang to evaluate the performance of Copenhagen telephone exchange [1, 2] and by Tore Olaus Engset, traffic analyst and then director of Norvegian Televerket (now Telenor Group) [3]. Both of them were studying—already in these days of human operators and cord boards used to switch telephone calls by means of jack plugs—how many circuits were needed to provide an acceptable telephone service or how many telephone operators were needed to handle a given volume of calls. Their analysis was based on Markov models, they assumed that the new connection demands made Poisson process and the duration of connections was given by a negative exponential distribution.

In a generic queueing model *customers* arrive to a *service system* at random intervals and are served during a random time: if the server is busy serving other customers, the arriving ones are queued. The model is expected to determine the distribution—or at least the mean value—of the number of customers in the system and their waiting time. If the number of customers that may be present in the system is limited, we are also interested in probability of rejection of an arriving customer.

Kendall's notation [4] is used to classify standard queueing nodes: in "$A/B/c$" A denotes the type of distribution of interarrival times, B—the type of service time distribution, and c the number of parallel servers. The symbol M (memoryless) on the place of A or B means that the corresponding distribution is exponential, E_r, H_r C_r denote Erlang, hyperexponential, and Cox distributions of order r, D means deterministic, G is for general distribution, etc. The notation has since been extended to $A/S/c/N/H/R$ where N is the capacity of the queue, H is the size of the customer source (if there are n customers present in a queueing system, it means that $H-n$ may still arrive), and R is the queueing discipline, e.g. FIFO (First-in-First-out) means that the customers are served following the order of their arrival; when the final three parameters are not specified, it is assumed that $N = \infty$, $H = \infty$ and $R = FIFO$, see e.g. [5].

Following this notation, Erlang considered $M/M/c/c$ model to obtain probability that all c parallel channels are occupied and the new calls are rejected, so called Erlang B formula for blocking probability, and $M/M/c$ model to determine probability that calls are queued (Erlang C formula) if their queueing is possible. Engset used $M/M/c/c/H$ model with finite populaton of users. These formulae are still in common use in telecommunication, see e.g. [6] even it their assumptions on exponential distributions between arrivals and connection duration are not well fitting the reality.

Then many mathematicians, e.g. Kolmogorov, Khinchin, Crommelin, Palm, Takács contributed to the development of queueing theory and their models found many applications. Apart from the obvious—where clients are people and service points are real-life locations such as post office counters or supermarket checkouts—there are many other interpretations; the "*clients*" might be ships docking at ports (here, "*service time*" is the unloading time), parcels sent to warehouses (storage time), or vehicles arriving at junctions (the time taken to cross).

The era of computer systems opened new perspectives. Processes executed in a computer system are queued waiting for system resources and the time they use them is the service time. Maybe the first queueing model of a computer system was presented by Sherr [7], where $M/M/1/N/N$ machine-repairman model considered by Khinchin [8] was applied, only its interpretation was changed. The original model referred to a pool of working machines: the time between failures was the customers' interarrival time and the server represented a workshop where the machines were repaired, the time of repair being the service time. In the new model the machines were replaced by terminals, the time between failures became users' thinking time and the time of processing the task by the computer system was interpreted as the service time. The very good accordance of the model results (mean response time of the system as a function of the number of active users) with measurements gave a boost to the development of next queueing models of computer systems. The first monographs appeared in the United States, e.g. [5, 9, 10], and Europe, e.g. [11].

Then came the era of computer networks and their models. Vast volumes of data organized into packets and supervised by communications protocols are constantly flowing across telecommunications networks, usually via intermediate nodes. At nodes, the incoming packets are queued to be sent according to network availability. The total transmission time is composed of signal propagation time between nodes which is constant and determined by the length of links and the speed of light in the link and of the waiting time at each node which is unknown and depends on the highly irregular current load of the network.

Estimation of this waiting time is important, since users need the transmission time to be as short and repeatable as possible. Reliability of transmission is equally important; when a packet queue is full, further incoming packets will not be saved. Network operators must seek a compromise: on one hand, it is important that a network be used to the best of its capacity; on the other, the more a network is being charged, the more the quality of service may drop. These issues are studied with the use of queueing theory. Here customers represent packets, the queues are at routers, and the service time is the time to sent a packet, bit-by-bit through an output gate of a router, hence it is proportional to the size of the packet.

The community of researchers studying these problems is vast and organised in academic as well as in telecommunication research centres. The constant development of computer networks and of Internet in particular gives new problems to be studied. Some of topical issues at the moment are e.g. the design of "green", i.e. energy saving (economical) networks, performances of various topologies of optical networks, modelling wireless networks with mobile customers.

However, queueing models are usually limited to steady-state analysis, as modelling of transient states is difficult. It limits the quality and applicability of queueing models, as real traffic intensity is constantly varying, see a typical example in Fig. 1. Moroever, the change of time scale does not change the variability of the traffic—a feature which is called "self-similarity" of the process and is related to its long-range autocorrelation. This feature, noticed in network traffic only at 90-ties [12], influences greatly the queue lengths and loss probabilities and should be reflected in queueing models.

The article discusses most useful methods we may apply in practice to analyse in a quantitative way transient states of queues in presence of time varying input flows,

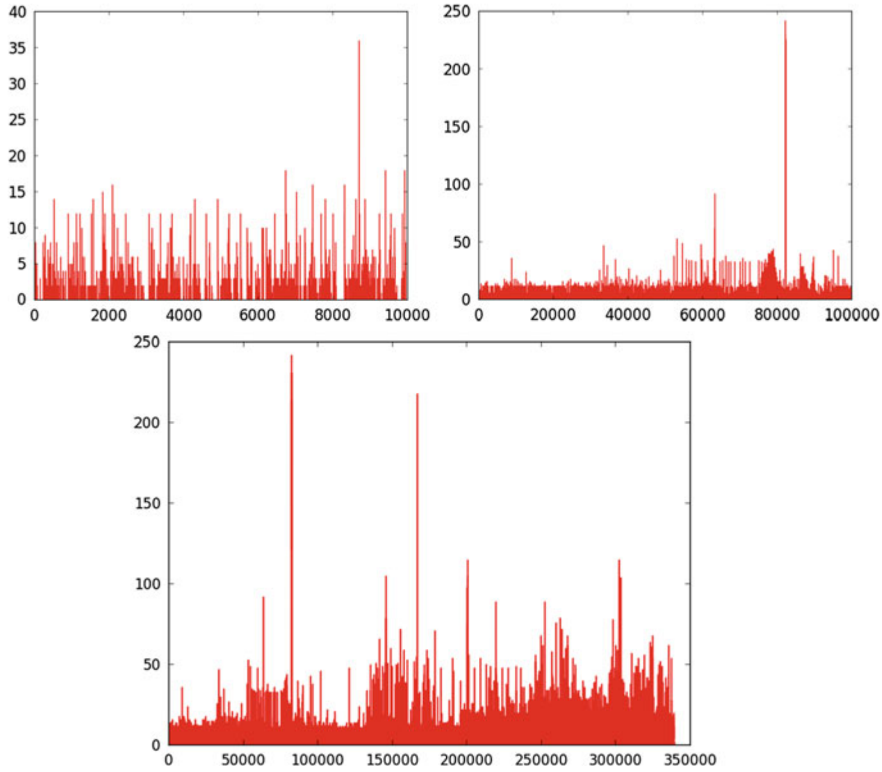

Fig. 1 Traffic intensity as a function of time in different time scales (data gathered at IITiS PAN)

namely: numerical solution of Chapman-Kolmogorov equations for continuous time Markov chains with very large state space, diffusion approximation, and fluid flow approximation. Markov models are essential for the evaluation of the performance of computer networks. In Markov chains the sojourn time at each state is exponentially distributed but it does not mean that in queueing models we are restricted to exponential distributions for interarrival times and service times. These distributions may be represented by linear combinations of exponentially distributed phases, e.g. by Cox distribution. However, Markov models are not scalable: the number of states is increasing rapidly with the complexity of a modelled object.

In diffusion approximation a diffusion equation (second order partial differential equation) defining the position of a particle in diffuse motion is used to describe the probability distribution of a queue length. This approach is merging states of the considered queueing system and needs much less computations than the Markov models. The principles of the approximation were given in [13] and then extended in [14] to the analysis of transient states with the use of semi-analytical, semi-numerical approach. Fluid-flow approximation is a simplified version of this method—only mean values of packet flows, queue lengths and service times are considered.

Differential equations (first order linear ordinary differential equations) involved here are simpler, and the computations can be completed in a reasonable time even for very large network topologies. It is also easier to model mechanisms used to control queues in nodes.

2 Markov Models

Steady state Markov models having a closed form solution and giving an explicit formula for state probabilities are limited to so called BCMP model [15]. It includes a multiclass open and closed network of any topology with four types of service stations. The types are: FIFO queue, IS (Infinite Server—there is enough parallel service channels to serve all customers, hence they are not queued), LIFO (Last In First Out—when a new customer comes, its service is started immediately and the current service is suspended and resumed when possible), PS (Processor Sharing—all customers are served in parallel but the service time is proportional to the number of present customers). In case of FIFO queue, service time distribution is limited to exponential one, the same for all classes of customers. For other types of stations it could be a Cox distribution, different for each class of customers. Note that Cox distribution may approximate any practically encountered distribution with infinitesimal error and there are tools able to match a Cox distribution to any empirical histogram, e.g. [16]. In more complex cases, e.g. $M/C_r/1$ FIFO queue with Cox service distribution, the underlying its performance Markov chain should be solved numerically.

For transient states, analytical solution is known only for M/M/1 and M/M/1/N single queues and even there it is complex. Transient states of these models were investigated more than half a century ago. Chapman-Kolmogorov equations (first-order linear differential equations) define state probabilities $p(n, t; n_0)$ of n customers present in the system at time t if $n = n_0$ at time $t = 0$. If we apply the Laplace transform to make these equations algebraic ones, solve them and then find the original functions of the solutions in time domain, we obtain [17]:

$$p(n, t; n_0 = i) = e^{(\lambda+\mu)t}[\rho^{\frac{n-i}{2}} I_{n-n_0}(at) + \rho^{\frac{n-n_0-1}{2}} I_{n+n_0+1}(at) + (1-\rho)\rho^n \sum_{j=n+n_0+2}^{\infty} \rho^{\frac{-j}{2}} I_j(at)] \quad (1)$$

where λ is input flow intensity, μ is service intensity (i.e. $1/\mu$ is mean service time), $\rho = \lambda/\mu$ is server utilisation factor, $a = 2\mu\sqrt{\rho}$ and $I_k(x)$ is the modified Bessel function of the first type and order k. Similarly, transient distributions for the limited queue M/M/1/N were derived [18, 19]. Some simplifications of the solution (1) were proposed, e.g. the generating function of the distribution $\bar{p}(n, s; n_0)$ may be replaced by expressions having simpler original functions in time domain [20] or Bessel functions may be replaced by easier to compute functions, [21].

These results do not fit well to the problem of modelling network routers, where the incoming streams are not Poisson and the size of packets is not exponentially distributed. Note that the solution (1) refers to transient states but it is assumed that the model parameters, λ in particular, are constant. Hence, in case of time dependent flows we should make them piecewise constant. We need models treating constantly changing non-Poisson flows and assuming general distributions of service times. We need also the possibility to include in these models the description of self-similarity of flows which enlarges mean queue lengths at buffers and increases packet loss probability, reducing this way the quality of services provided by a network. The models should also meet very large topologies characteristic to the Internet.

Markov models are very flexible and may reflect the mechanisms for regulating the intensity of Internet transmissions and mechanisms to ensure the quality of transmission services; they may also include self-similar flows. Let us summarise Markov approach to reflect self-similarity of network flows.

The term self-similar was introduced by Mandelbrot [22] for explaining water level pattern of river Nile observed by H. Hurst. This term was also known as *Hurst Effect*. The degree of self-similarity is expressed by *Hurst parameter*, denoted by H.

A real valued stochastic process:

$$X = \{X(t)\}_{t \in R}$$

is self-similar with $H > 0$, if for any $a > 0$,

$$\{X(at)\}_{t \in R} \stackrel{d}{=} \{a^H X(t)\}_{t \in R}$$

where $\stackrel{d}{=}$ denotes equality in finite dimensional distribution sense. This means that self-similar processes are scale invariant.

Mathematically, the difference between short-range dependent processes and long-range ones (self-similar) is as follows [23]:

For a short-range dependent process:

- $\sum_{r=0}^{\infty} \mathrm{Cov}(X_t, X_{t+\tau})$ is convergent,
- spectrum at $\omega = 0$ is finite,
- for large m, $\mathrm{Var}(X_k^{(m)})$ is asymptotically of the form $\mathrm{Var}(X)/m$,
- the aggregated process $X_k^{(m)}$ tends to the second order pure noise as $m \to \infty$;

For a long-range dependent process:

- $\sum_{r=0}^{\infty} \mathrm{Cov}(X_t, X_{t+\tau})$ is divergent,
- spectrum at $\omega = 0$ is singular,
- for large m, $\mathrm{Var}(X_k^{(m)})$ is asymptotically of the form $\mathrm{Var}(X)m^{-\beta}$,
- the aggregated process $X_k^{(m)}$ does not tend to the second order pure noise as $m \to \infty$,

where the spectrum of the process is the Fourier transformation of the autocorrelation function and the aggregated process $X_k^{(m)}$ is the average of X_t on the interval m:

$$X_k^{(m)} = \frac{1}{m}(X_{km-m+1} + \cdots + X_{km}) \quad k \geq 1.$$

Estimation of Hurst parameter is the most frequently used method to check if a process is self-similar: for non-self-similar processes $H = 0.5$; for $0.5 < H < 1$ process is self-similar; the closer H is to 1, the greater the degree of persistence of long-range dependence. Hurst parameter H can be estimated by various methods. The simplest one is based on the analysis of variance-time plot. The variation of aggregated self-similar process is equal to:

$$Var(X_k^{(m)}) = Var(X)m^{-\beta}$$

so the log-log plot of $\frac{Var(X_k^{(m)})}{Var(X)}$ versus m is a line with slope β.

Since the days the self-similarity of network flows was discovered, several non-Markov models have been introduced, e.g. with the use of fractional Brownian Motion [24], stable Levy Motion [25] chaotic maps [26], α-stable distribution [27], fractional Autoregressive Integrated Moving Average (fARIMA) [28] and fractional Levy Motion [29]. The advantage of these models is that they give a good description of the traffic with the use of few parameters. Their drawback consist in the fact that they cannot be incorporated in Markov queueing models.

A way to create a Markov model of a self-similar source is to use a Markov-modulated Poisson processes (MMPP) the parameter of which depends on a state of a separate Markov chain (modulator). In the simplest form we use a two state modulator. The superposition of MMPP's is also an MMPP which is a special case of Markov Arrival Process (MAP).

A MAP is defined by two square matrices $\mathbf{D_0}$ and $\mathbf{D_1}$ such that $\mathbf{Q} = \mathbf{D_0} + \mathbf{D_1}$ is an irreducible infinitesimal generator for the continuous-time Markov chain (CTMC) underlying the process, and $D_0(i, j)$ (respectively $D_1(i, j)$) is the rate of hidden (respectively observable) transitions from state i to state j [30]. Two-state MAP is a Markovian arrival process with square matrices as follows:

$$\mathbf{D_0} = \begin{bmatrix} -\sigma_1 & \lambda_{1,2} \\ \lambda_{2,1} & -\sigma_2 \end{bmatrix}, \qquad \mathbf{D_1} = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} \\ \mu_{2,1} & \mu_{2,2} \end{bmatrix}$$

where $\lambda_{i,j} \geq 0$, $\mu_{i,j} \geq 0$, for all i, j. The diagonal elements of matrix $\mathbf{D_0}$ are $\sigma_1 = \lambda_{1,2} + \mu_{1,1} + \mu_{1,2} > 0$ and $\sigma_2 = \lambda_{2,1} + \mu_{2,2} + \mu_{2,1} > 0$ such that underlying continuous-time Markov chain Matrix $\mathbf{Q}$ has no absorbing states.

Following the model proposed in [31], a LRD process can be modelled as the superposition of d two-state MMPPs. The ith MMPP ($1 \leq i \leq d$) can be parameterized by two square matrices:

$$\mathbf{D_0^i} = \begin{bmatrix} -(c_{1i} + \lambda_{1i}) & c_{1i} \\ c_{2i} & -(c_{2i} + \lambda_{2i}) \end{bmatrix}, \qquad \mathbf{D_1^i} = \begin{bmatrix} \lambda_{1i} & 0 \\ 0 & \lambda_{2i} \end{bmatrix}$$

The element c_{1i} is the transition rate from state 1 to 2 of the ith MMPP and c_{2i} is the rate out of state 2 to 1. λ_{1i} and λ_{2i} are the traffic rate when the ith MMPP is in state 1 and 2 respectively. The sum of $\mathbf{D_0}^i$ and $\mathbf{D_1}^i$ is an irreducible infinitesimal generator $\mathbf{Q}^i$ with the stationary probability vector:

$$\overrightarrow{\pi}_i = \left(\frac{c_{2i}}{c_{1i} + c_{2i}}, \frac{c_{1i}}{c_{1i} + c_{2i}} \right)$$

The superposition of these two-state MMPPs is a new MMPP with 2^d states and its parameter matrices, $\mathbf{D_0}$ and $\mathbf{D_1}$, can be computed using the Kronecker sum of those of the d two-state MMPPs [32]:

$$(\mathbf{D}_0, \mathbf{D}_1) = \left(\oplus_{i=1}^{d} \mathbf{D_0}^i, \oplus_{i=1}^{d} \mathbf{D_1}^i \right)$$

A procedure for fitting moments computed from the model and from measured data gives us the model parameters, [31]. We may also use an approach based on hidden Markov chains to model self-similar traffic [33].

The main disadvantage of Markov approach is its lack of scalability. Markov chains to be used become intractable because of the number of states which is growing very rapidly with the complexity of a modelled object (so called *state explosion*). Each state of the Markov chain corresponds to one state of the system. It is necessary to construct and solve the system of equations defining the probability of states—the number of equations equals the number of states. The existing solvers as e.g. Markov solver in QNAP [34], XMARCA [35], PEPS [36], PRISM [37] consider only steady state Markov chains and solve algebraic systems of equations.

Theoretically, for any continuous time Markov chain the Chapman-Kolmogorov equations with transition matrix Q

$$\frac{d\pi(t)}{dt} = \pi(t)\mathrm{Q}, \tag{2}$$

have the analytical transient solution:

$$\pi(t) = \pi(0)e^{\mathrm{Q}t}, \tag{3}$$

where $\pi(t)$ is the probability vector and $\pi(0)$ is the initial condition. However, it is not easy to compute the expression $e^{\mathrm{Q}t}$ where Q is a large matrix, see e.g. [38]. It may be done by its expansion to Taylor's series

$$e^{\mathrm{Q}t} = \sum_{k=0}^{\infty} \frac{(\mathrm{Q}t)^k}{k!}. \tag{4}$$

but the task is numerically unstable, especially for large Q. Additionally, to consider $\lambda(t)$, we should make the parameters of the model piecewise constant in small intervals and apply the solution (3) at each of these intervals.

In IITiS we are developing our own package Olymp [39]. It is a library generating transition matrices of continuous time Markov chains (CTMC), solving them. Olymp uses Java language to define network nodes and the interactions between them. Due to the potentially very large sizes of the models' transition matrices, their generation is parallelized, and they can be compressed on-the-fly using a dedicated compression based on finite-state automata. Olymp has a quite different approach to represent CTMC in the comparison to typical model checkers. A move to another state involves a transfer of a token. A node that sends the token initiates the move asynchronously, in moments of time that adhere to an exponential distribution. A node that receives the token can accept it validating the move. The negotiation can be thought of as for example an agreed transfer of a packet between these two nodes or as a synchronisation on a signal, distributed to the network by a clock node. At the moment we are able to generate and solve Markov chains of the 150 million of states. The method of solution used is one of projection methods based on Krylov subspace with Arnoldi process to project the exponential of a large matrix approximately onto a small Krylov subspace, see [35]; the transition matrix is then small and the computation of the expression (3) with the use of uniformization method and Padé approximations is much easier, [40–43]. This approach is supplemented by direct numerical solution of large systems of ordinary differential equations (ODE) using uniformization, i.e. discretization of the CTMC, that is replacing the CTMC by a DTMC (a discrete-time Markov chain) and a Poisson process.

We are increasing the size of tractable Markov chains by several orders through the use of a GPU-CPU (graphical processing unit) and a better design of computational algorithms for parallel computing and optimization of memory usage, [44]. GPU capabilities go far beyond the computer graphics. It is well known that a potential computational power of GPUs is much greater than that of contemporary CPUs (in a sense of the performance measured by number of floating point operations per second). Thus, it is possible to shorten the time of computations. Due to the enormous amount of the data to be processed, methods must be developed to store vectors and matrices with intelligent management of memory.

3 Diffusion Approximation

3.1 Diffusion Approximation of Single $G/G/1$, $G/G/1/N$ Queues

This approach is merging states of the considered queueing system and thus needs much less computations than the Markov models. We present here the principles of the method following [13] where steady-state solution of a single G/G/1/N model was given and then extended to the network of queues in [45]. We supplemented these results with semi-analytical, semi-numerical transient state solution [14] given for constant model parameters but it could be applied also in case of time-dependent parameters if we only make them constant within small intervals.

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 two first 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 $X(t)$, 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

$$\frac{\partial f(x,t;x_0)}{\partial t} = \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta\frac{\partial f(x,t;x_0)}{\partial x}. \tag{5}$$

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 increase at the same rate with the length of the observation period. In the case of G/G/1/N station, the process evolves between barriers placed at $x = 0$ and $x = N$ where barriers *with instantaneous jumps* are placed, [13]. 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. When the process comes to the barrier at $x = N$, it stays there for a time which is exponentially distributed with a parameter μ_0 which corresponds to the time when the system is full and do not accept new customers (the completion time of current service from the moment when the queue becomes full). The assumption on exponential sojourn times in barriers will be dropped below where transient model is presented. Diffusion equation becomes and is supplemented by balance equations for probabilities $p_0(t)$ and $p_N(t)$ of being at the barriers

$$\begin{aligned}
\frac{\partial f(x,t;x_0)}{\partial t} &= \frac{\alpha}{2}\frac{\partial^2 f(x,t;x_0)}{\partial x^2} - \beta\frac{\partial f(x,t;x_0)}{\partial x} \\
&\quad + \lambda_0 p_0(t)\delta(x-1) + \lambda_N p_N(t)\delta(x-N+1), \\
\frac{dp_0(t)}{dt} &= \lim_{x\to 0}\,[\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} - \beta f(x,t;x_0)] - \lambda_0 p_0(t), \\
\frac{dp_N(t)}{dt} &= \lim_{x\to N}\,[-\frac{\alpha}{2}\frac{\partial f(x,t;x_0)}{\partial x} + \beta f(x,t;x_0)] - \lambda_N p_N(t),
\end{aligned} \tag{6}$$

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

In stationary state, Eq. (6) become ordinary differential ones and their solution, if $\rho = \lambda/\mu \neq 1$, may be expressed as, see [13]:

$$f(x) = \begin{cases} \dfrac{\lambda p_0}{-\beta}(1 - e^{zx}) & \text{for } 0 < x \leq 1, \\ \dfrac{\lambda p_0}{-\beta}(e^{-z} - 1)e^{zx} & \text{for } 1 \leq x \leq N-1, \\ \dfrac{\mu p_N}{-\beta}(e^{z(x-N)} - 1) & \text{for } N-1 \leq x < N, \end{cases} \quad (7)$$

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

$$p_0 = \lim_{t\to\infty} p_0(t) = \{1 + \rho e^{z(N-1)} + \frac{\rho}{1-\rho}[1 - e^{z(N-1)}]\}^{-1}, \quad (8)$$

$$p_N = \lim_{t\to\infty} p_N(t) = \rho p_0 e^{z(N-1)}. \quad (9)$$

The model may include **classes** of customers, each having its interarrival and service time distributions and routing probabilities, [45]. When the input stream λ is 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

$$\frac{1}{\mu} = \sum_{k=1}^{K} \frac{\lambda^{(k)}}{\lambda} \frac{1}{\mu^{(k)}}, \quad \text{and} \quad C_B^2 = \mu^2 \sum_{k=1}^{K} \frac{\lambda_{(k)}}{\lambda} \frac{1}{\mu_{(k)}^2} ({C_B^{(k)}}^2 + 1) - 1. \quad (10)$$

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

$$C_A^2 = \sum_{k=1}^{K} \frac{\lambda^{(k)}}{\lambda} {C_A^{(k)}}^2 \, . \quad (11)$$

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

$$p_k(n^{(k)}) = \sum_{n=n^{(k)}}^{N} \left[p(n) \binom{n}{n^{(k)}} \left(\frac{\lambda^{(k)}}{\lambda} \right)^{n^{(k)}} \left(1 - \frac{\lambda^{(k)}}{\lambda} \right)^{n-n^{(k)}} \right] \quad k = 1, \ldots, K \ . \tag{12}$$

Our **transient solution** of Eq. 6 is based on the representation of the density function $f(x, t; x_0)$ of the diffusion process with barriers with jumps by a superposition of the density functions $\phi(x, t; x_0)$ of diffusion processes with absorbing barriers at $x = 0$ and $x = N$, which has the following form, see [46]

$$\phi(x, t; x_0) = \begin{cases} \delta(x - x_0) & \text{for } t = 0 \\ \frac{1}{\sqrt{2\Pi\alpha t}} \sum_{n=-\infty}^{\infty} \left\{ \exp\left[\frac{\beta x'_n}{\alpha} - \frac{(x - x_0 - x'_n - \beta t)^2}{2\alpha t} \right] \right. & \\ \qquad \left. - \exp\left[\frac{\beta x''_n}{\alpha} - \frac{(x - x_0 - x''_n - \beta t)^2}{2\alpha t} \right] \right\} & \text{for } t > 0, \end{cases} \tag{13}$$

where $x'_n = 2nN$, $x''_n = -2x_0 - x'_n$. 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 $\phi(x, t; \psi) = \int_0^N \phi(x, t; \xi)\psi(\xi)d\xi$.

The probability density function $f(x, t; \psi)$ of the diffusion process with elementary returns is composed of the function $\phi(x, t; \psi)$ which represents the influence of the initial conditions and of a spectrum of functions $\phi(x, t - \tau; 1)$, $\phi(x, t - \tau; N - 1)$ which are pd functions of diffusion processes with absorbing barriers at $x = 0$ and $x = N$, started at time $\tau < t$ at points $x = 1$ and $x = N - 1$ with densities $g_1(\tau)$ and $g_{N-1}(\tau)$:

$$f(x, t; \psi) = \phi(x, t; \psi) + \int_0^t g_1(\tau)\phi(x, t - \tau; 1)d\tau + \int_0^t g_{N-1}(\tau)\phi(x, t - \tau; N - 1)d\tau. \tag{14}$$

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

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

$$\gamma_N(t) = p_N(0)\delta(t) + [1 - p_0(0) - p_N(0)]\gamma_{\psi,N}(t) + \int_0^t g_1(\tau)\gamma_{1,N}(t-\tau)d\tau$$
$$+ \int_0^t g_{N-1}(\tau)\gamma_{N-1,N}(t-\tau)d\tau, \tag{15}$$

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

$$\gamma_{1,0}(t) = \lim_{x\to 0}[\frac{\alpha}{2}\frac{\partial\phi(x,t;1)}{\partial x} - \beta\phi(x,t;1)]. \tag{16}$$

For absorbing barriers

$$\lim_{x\to 0}\phi(x,t;x_0) = \lim_{x\to N}\phi(x,t;x_0) = 0,$$

hence $\gamma_{1,0}(t) = \lim_{x\to 0}\frac{\alpha}{2}\frac{\partial\phi(x,t;1)}{\partial x}$. 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 $x = 0$ or $x = N$.

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

$$g_1(\tau) = \int_0^\tau \gamma_0(t)l_0(\tau - t)dt, \quad g_{N-1}(\tau) = \int_0^\tau \gamma_N(t)l_N(\tau - t)dt, \tag{17}$$

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 as it is in Eq. (6).

The above equations are transformed by the Laplace transform, and the transform of $f(x,t,x_0)$ is obtained analytically and then its original is computed numerically using e.g. Stehfest algorithm [47].

In case of unlimited queue of G/G/1 type we just remove the barrier at $x = N$ and related to it terms and equations.

3.2 Open Network of G/G/1, G/G/1/N Queues, Steady State and Transient Solution

The steady-state open networks models of $G/G/1$ queues were studied in [45]. Let M be the number of stations and suppose at the beginning that there is one class of customers. The throughput of station i is, as usual, obtained from traffic equations

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^{M} \lambda_j r_{ji}, \quad i = 1, \ldots, M, \tag{18}$$

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

Second moment of interarrival time distribution is obtained from two systems of equations; the first defines C^2_{Di} as a function of C^2_{Ai} and C^2_{Bi}; the second defines C^2_{Aj} as another function of C^2_{D1}, ..., C^2_{DM} :

(1) The formula (19) is exact for $M/G/1$, $M/G/1/N$ stations and is approximate in the case of non-Poisson input [48]

$$d_i(t) = \rho_i b_i(t) + (1 - \rho_i) a_i(t) * b_i(t), \quad i = 1, \ldots, M, \tag{19}$$

where * denotes the convolution operation. From (19) we get

$$C^2_{Di} = \rho_i^2 C^2_{Bi} + C^2_{Ai}(1 - \rho_i) + \rho_i(1 - \rho_i). \tag{20}$$

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

$$\begin{aligned} d_{ij}(x) &= d_i(x) r_{ij} + d_i(x) * d_i(x)(1 - r_{ij}) r_{ij} \\ &\quad + d_i(x) * d_i(x) * d_i(x)(1 - r_{ij})^2 r_{ij} + \cdots \end{aligned} \tag{21}$$

hence

$$E[D_{ij}] = \frac{1}{\lambda_i r_{ij}}, \quad C^2_{Dij} = r_{ij}(C^2_{Di} - 1) + 1. \tag{22}$$

$E[D_{ij}]$, 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

$$\lambda_j t = [\sum_{i=1}^{M} \lambda_i r_{ij} + \lambda_{0j}] t \quad \text{and variance} \quad \sigma^2_{Aj} t = \{\sum_{i=1}^{M} C^2_{Dij} \lambda_i r_{ij} + C^2_{0j} \lambda_{0j}\} t,$$

hence

$$C^2_{Aj} = \frac{1}{\lambda_j} \sum_{i=1}^{M} r_{ij} \lambda_i [(C^2_{Di} - 1) r_{ij} + 1] + \frac{C^2_{0j} \lambda_{0j}}{\lambda_j} \,. \tag{23}$$

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

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

$$\lambda_i^{(k)} = \lambda_{0i}^{(k)} + \sum_{j=1}^{M} \lambda_j^{(k)} r_{ji}^{(k)}, \quad i = 1, \ldots, M; \; k = 1, \ldots, K, \tag{24}$$

and

$$C_{Di}^2 = \lambda_i \sum_{k=1}^{K} \frac{\lambda_i^{(k)}}{\mu_i^{(k)2}} [C_{Bi}^{(k)2} + 1] + 2\rho_i(1-\rho_i) + (C_{Ai}^2 + 1)(1-\rho_i) - 1\,. \tag{25}$$

A customer in the stream leaving station i belongs to class k with probability $\lambda_i^{(k)}/\lambda_i$ and we can determine $C_{Di}^{(k)2}$ in the similar way as it has been done in Eqs. (21)–(22), replacing r_{ij} by $\lambda_i^{(k)}/\lambda_i$:

$$C_{Di}^{(k)2} = \frac{\lambda_i^{(k)}}{\lambda_i}(C_{Di}^2 - 1) + 1; \tag{26}$$

then

$$C_{Aj}^2 = \frac{1}{\lambda_j} \sum_{l=1}^{K} \sum_{k=1}^{K} r_{ij}^{(k)} \lambda_i \left[\left(\frac{\lambda_i^{(k)}}{\lambda_i} (C_{Di}^2 - 1) \right) r_{ij}^{(k)} + 1 \right] + \sum_{k=1}^{K} \frac{C_{0j}^{(k)2} \lambda_{0j}^{(k)}}{\lambda_j}. \tag{27}$$

Equations (20), (23) or (25), (27) form a linear system of equations and allow us to determine C_{Ai}^2 and, in consequence, parameters β_i, α_i for each station.

In our approach to **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. (18), (20) and (23) are used to determine the input parameters of each station based on the values of $\rho_i(t)$ obtained at the end of the precedent interval. As the values of parameters are changed at each interval, also externel flows $\lambda_{0j}^{(k)}(t)$ may be controlled following any, possibly **self-similar process**.

3.3 G/G/c/c and G/G/c/c/H Stations

We come here back to mentioned in the introduction Erlang and Engset models playing historical role in telecommunication. We show their more general diffusion version including any interarrival and service time distributions. In case of multiple service channels as in $G/G/c/c$ and $G/G/c$ stations, the output flow is state-dependent, and in case of finite population $G/G/c/c/H$ model also the input flow is state-dependent.

In $G/G/c/c$ we distinguish c subintervals: if $n-1 < x < n$, it is supposed that n channels are busy (n customers inside the system), hence we choose

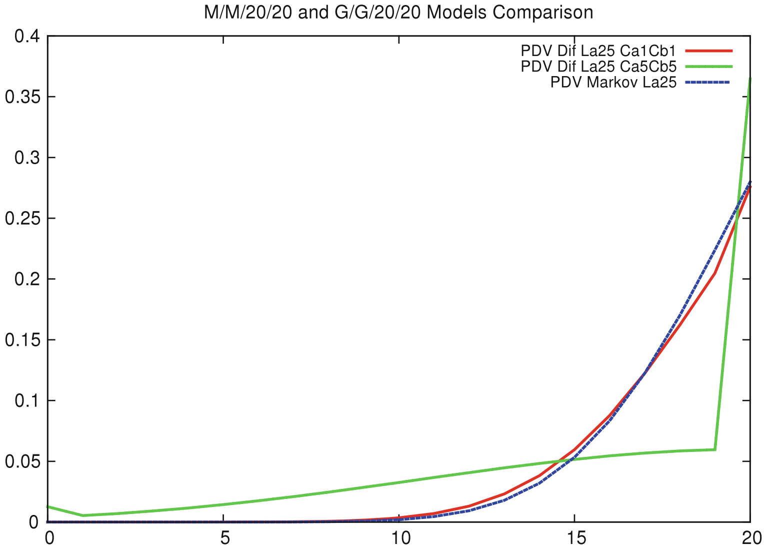

Fig. 2 A comparison of the distribution $p(n)$ at a M/M/20/20 model and its diffusion approximation (Eq. 29), $\lambda = 25$, $\mu = 1$ and a corresponding G/G/20/20 model with $C_A^2 = C_B^2 = 5$

$$\alpha_n = \lambda C_A^2 + n\mu C_B^2, \quad \beta_n = \lambda - n\mu \quad \text{for } n-1 < x < n, \; n = 1, 2, \ldots, c. \quad (28)$$

Jumps are performed from $x = 0$ to $x = 1$ with intensity λ and from $x = c$ to $x = c - 1$ with intensity $c\mu$. The solution is

$$\begin{aligned}
f_1(x) &= \begin{cases} \dfrac{\lambda p(0)}{-\beta_1}\left(1 - e^{z_1 x}\right) & \text{for } \beta_1 \neq 0, \\ \dfrac{2\lambda p(0)}{-\alpha_1} x & \text{for } \beta_1 = 0, \end{cases} \quad 0 < x \le 1\,, \\
f_n(x) &= \begin{cases} f_{n-1}(n-1) e^{z_n(x-n+1)} & \text{for } \beta_n \neq 0, \quad n-1 \le x \le n\,, \\ f_{n-1}(n-1) & \text{for } \beta_n = 0, \quad n = 2, \ldots, c-1\,, \end{cases} \\
f_c(x) &= \begin{cases} \dfrac{\mu p(c)}{-\beta_c}\left[e^{z_c(x-c)} - 1\right] & \text{for } \beta_c \neq 0, \\ \dfrac{-2\mu p(c)}{\alpha_c}(x - c) & \text{for } \beta_c = 0, \end{cases}, \quad c - 1 \le x < c\,.
\end{aligned} \quad (29)$$

where $z_n = \frac{2\beta_n}{\alpha_n}$, and $p(0)$, $p(c)$ come from normalisation. Figures 2 and 3 show the accuracy of this approach in case of heavy and light load by comparing the Markovian model with its diffusion approximation and demonstrates how non-exponential distributions influence the results. In case of $G/G/c$ system the barrier at $x = c$ is removed and the last interval with parameters β_c, α_c is extended: $x \in (c - 1, \infty)$.

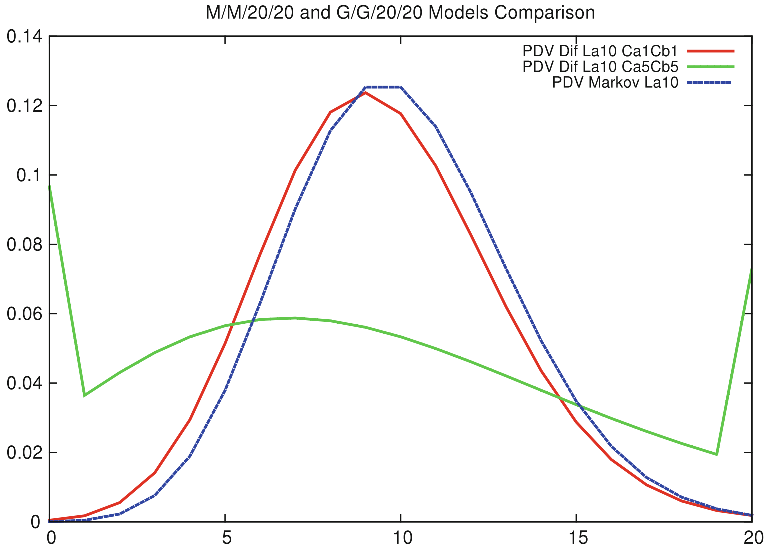

Fig. 3 A comparison of the distribution $p(n)$ at a M/M/20/20 model and its diffusion approximation (Eq. 29), $\lambda = 10$, $\mu = 1$ and a corresponding G/G/20/20 model with $C_A^2 = C_B^2 = 5$

In case of $G/G/c/c/H$ system with finite population,

$$\beta_n = (H - n + 1)\nu - n\mu, \quad \alpha_n = (H - n + 1)\nu C_A^2 + n\mu C_B^2, \quad 1 \le n \le c,$$
$$\beta_n = (H - n + 1)\nu - c\mu, \quad \alpha_n = (H - n + 1)\nu C_A^2 + c\mu C_B^2, \quad n \ge c.$$

where ν and C_A^2 refer to the sojourn time in the pool, and the solution is analogous to (29).

For the overflow traffic description we may use probabilities $p(n)$, $n \ge c$ given by $G/G/c$ or $G/G/c//H$ models to extend Riordan formulae but it would be more natural to compute the characteristics of the flow in terms of diffusion models. If the pdf $f_A(x)$ of interarrival times mean $1/\lambda$ and variance σ_A^2 (squared coefficient of variation C_A^2 and $p(c)$ is the blocking probability, then the interevent density $f_{overflow}(x)$ of times between events in the overflow traffic is

$$f_{overflow}(x) = f_A(x)p(c) + f_A(x) * f_A(x)(1 - p(c))p(c) + \cdots$$

giving mean $1/[\lambda p(c)]$ and squared coefficient of variation $C_{overflow}^2$ = p(c) $\left(C_A^2 - 1\right) + 1$. These parameters may be used while splitting and merging traffic flows in the same way as it is done in network diffusion models [45].

Similarly, the pdf $f_{acc}(x)$ of interevent times at the accepted traffic is

$$f_{acc}(x) = f_A(x)(1 - p(c)) + f_A(x) * f_A(x)p(c)(1 - p(c)) + \cdots .$$

There is no probability transfer between intervals in steady-state solution but we should take it into account in **transient solution** of $G/G/c/c$ and $G/G/c/c/H$ models. Inside each of c intervals of unitary length, the diffusion equation is solved assuming that the barriers at its left and right side act as absorbing ones. The density function $\phi(x,t;x_0)$ of a diffusion process limited by two absorbing barriers has the same form as we used previously in $G/G/1/N$ model. To balance probability flows between neighbouring intervals having different diffusion parameters, we put imaginary barriers between these intervals and suppose that the diffusion process which is entering a barrier at $x = n$, $n = 1, 2, \ldots, c-1$, from its left side (the process is increasing) is absorbed and immediately reappears at $x = n+\varepsilon$. Similarly, a process which is decreasing and enters the barrier from its right side reappears at its other side at $x = n-\varepsilon$. The value of ε should be small, for example of the order of 2^{-10}, but we checked that it has no significant impact on the solution.

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

$$
\begin{aligned}
f_1(x,t;\psi_1) &= \phi_1(x,t;\psi_1) + \int_0^t g_{1-\varepsilon}(\tau)\phi_1(x,t-\tau;1-\varepsilon)d\tau, \\
f_n(x,t;\psi_n) &= \phi_n(x,t;\psi_n) + \int_0^t g_{n-1+\varepsilon}(\tau)\phi_n(x,t-\tau;n-1+\varepsilon)d\tau \\
&\quad + \int_0^t g_{n-\varepsilon}(\tau)\phi_n(x,t-\tau;n-\varepsilon)d\tau, \quad n = 2, \ldots c-1, \\
f_c(x,t;\psi_c) &= \phi_c(x,t;\psi_c) + \int_0^t g_{c-1+\varepsilon}(\tau)\phi_c(x,t-\tau;c-1+\varepsilon)d\tau
\end{aligned} \tag{30}
$$

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

$$
\gamma_n^R(t) = g_{n-\varepsilon}(t), \quad \gamma_n^L(t) = g_{n+\varepsilon}(t), \quad n = 1, \ldots, c-1 \tag{31}
$$

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

The densities $g_1(t)$, $g_{N-1}(t)$ are the same as in $G/G/1/N$ model, and densities $\gamma_n^R(t)$, $\gamma_n^L(t)$ are obtained as

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

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

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

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

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

$$+ \int_0^t g_{n+1-\varepsilon}(\tau)\gamma_{n+1-\varepsilon,n}(t-\tau)d\tau, \quad n = 1, \dots c-2$$

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

$$\gamma_c(t) = p(c)(0)\delta(t) + \gamma_{\psi_c,c}(t) + \int_0^t g_{c-1+\varepsilon}(\tau)\gamma_{c-1+\varepsilon,c}(t-\tau)d\tau \ , \qquad (32)$$

where $\gamma_{i,j}(t)$ are the densities of first passage times between points i, j.

This system of equations is transformed with the use of Laplace transform and solved numerically to obtain the values of $\bar{f}_n(x, s; \psi_n)$. Then we use the Stehfest inversion algorithm [47] to compute $f_n(x, t; \psi_n)$; for a specified t.

The parameters of the above 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.

Example 3.1 (extended Erlang model) Consider single $M/M/20/20$ and $G/G/20/20$ stations, at each channel $\mu = 1$. At $t = 0$ the system is empty. During the period $t \in [0, 20]$ the intensity of the input stream $\lambda_{in}(t) = 10$, then for $t \in [20, 50]$ $\lambda_{in}(t) = 25$, and for $t \in [50, 80]$ again $\lambda_{in}(t) = 10$.

Figure 4 presents the mean number of occupied channels as a function of time, for exponential ($C_A^2 = C_B^2 = 1$) and non-exponential ($C_A^2 = C_B^2 = 5$) distributions, the impact of the variances of both considered distributions is visible.

Example 3.2 (extended Engset model) During the period $t \in [0, 20]$ there is a finite population of $H = 20$ connections, the activation intensity of each connection is $\nu = 1$. As previously, we consider a $M/M/20/20$ station, at each channel $\mu = 1$,

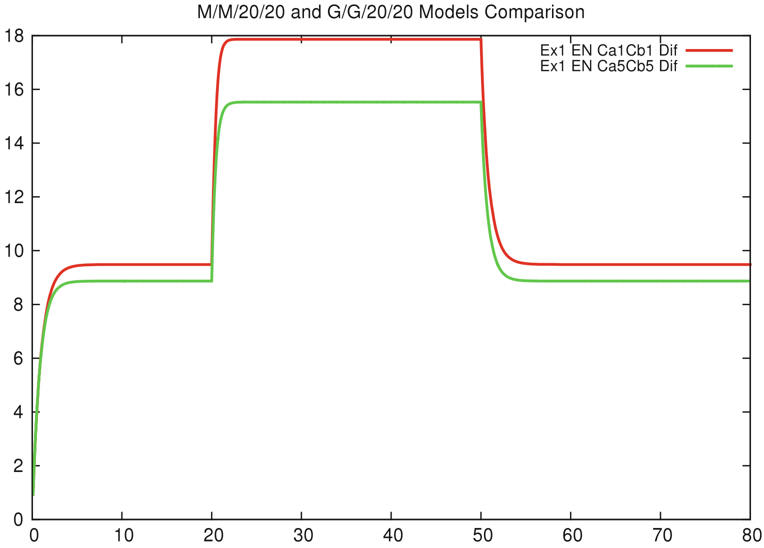

Fig. 4 Example 3.1 (infinite population): mean number of occupied channels as a function of time, $C_A^2 = C_B^2 = 1$ and $C_A^2 = C_B^2 = 5$,

and t $t = 0$ the system is empty. Then in $t \in [20, 50]$ the population is $H = 40$ and for $t \in [50, 80]$ the size of the pool is again $H = 20$.

Figure 5 presents probabilities $p(c)$ for the both examples (Erlang and Engset models), for exponential and non-exponential distributions. The influence of large variations is more important in finite population model. During peak load ($t \in [20, 50]$ it is 70 % greater for $C_A^2 = C_B^2 = 5$ then for $C_A^2 = C_B^2 = 1$ in case of Engset model and 32 % greater in case of Erlang model. During low traffic periods the relative difference of $p(c)$ in Markov and non-Markov models is much more distinct, as the value of $p(c)$ is closed to zero.

We do not compare here the approximate results with simulation but in general, the accuracy of these results is acceptable, see a discussion of diffusion approximation errors in [49].

3.4 Diffusion Approximation of Preemptive—Resume Priority System

This paragraph introduces a diffusion model of a single server with priority preemptive—resume queuing discipline, cf. [50]. Customers arriving to the system are divided into a number K classes. Each class is distinguished by its index k, $k = 1, \ldots, K$, 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

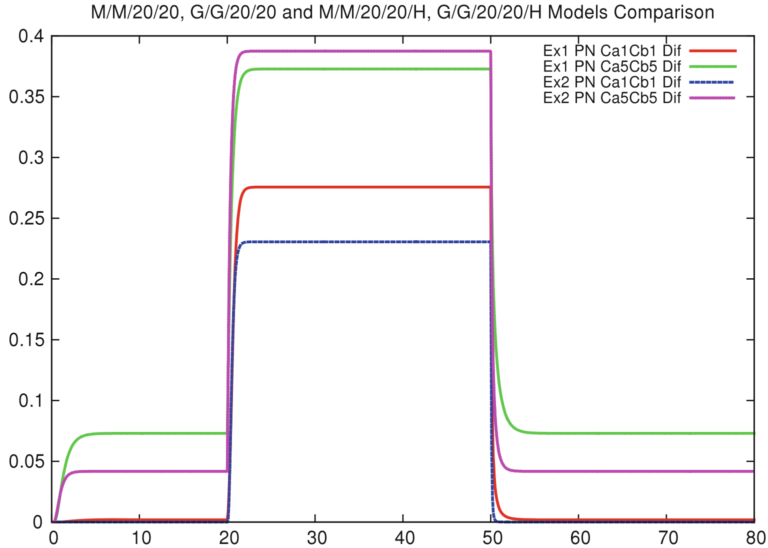

Fig. 5 Blocking probabilities as a function of time for both Examples 3.1, 3.2: finite and infinite population, $C_A^2 = C_B^2 = 1$ and $C_A^2 = C_B^2 = 5$

customer of class l, $l < k$ arrives, the current service is suspended and the service of the newcomer begins. After 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. The 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, there are two streams defined: 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 it can be proven that these processes have approximately normal distributions if the period $[0, t]$ is sufficiently long and within a busy period of the server. 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 $\alpha^{(K)}t$:

$$\beta^{(K)} = \sum_{k=1}^{K} \lambda^{(k)} - \sum_{k=1}^{K}((1 - p_0^{(k-1)}(t))\mu^{(k-1)} + p_0^{(k-1)}(t)\mu^{(k)}),$$

$$\alpha^{(K)} = \sum_{k=1}^{K} \lambda^{(k)} C_A^{(k)^2} + \sum_{k=1}^{K}((1 - p_0^{(k-1)}(t))\mu^{(k-1)} C_B^{(k-1)^2} + p_0^{(k-1)}(t)\mu^{(k)} C_B^{(k)^2}),$$

$$C_A^{(k)^2} = \lambda^{(k)^2}\sigma_A^{(k)^2}, \quad C_B^{(k)^2} = \mu^{(k)^2}\sigma_B^{(k)^2}, \quad p_0^{(0)}(t) = 1, \quad \mu^{(0)} = 0$$

and are approximately normally distributed. Probability $p_0^{(k-1)}(t)$ for classes 2, ..., K is the probability, that there are no customers of all higher priority classes (1, ..., $K - 1$) present in the system. Diffusion approximation method replaces 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 in earlier chapters with the intensity of jumps from $x = 0$: $\Lambda^{(K)} = \sum_{k=1}^{K} \lambda^{(k)}$ gives the density function $f^{(K)}(x, t; x_0)$ for all classes considered together.

This approach studies the diffusion processes corresponding to each class customers together, taking into account 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, in the case of 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 both higher and lower classes, where lower priority is served only in absence of the higher class, has the parameters

$$\beta^{(2)} = \lambda^{(1)} + \lambda^{(2)} - (1 - p^{(1)}(0, t))\mu^{(1)} - p^{(1)}(0, t)\mu^{(2)}$$

$$\alpha^{(2)} = \sigma_A^{(1)2}\lambda^{(1)^3} + \sigma_A^{(2)^2}\lambda^{(2)^3} + (1 - p_0^{(1)}(t))\sigma_B^{(1)^2}\mu^{(1)^3} + p_0^{(1)}(t)\sigma_B^{(2)^2}\mu^{(2)^3}.$$

Example 3.3 Consider a server with two priority levels. In the first example, the priority customers come with intensity $\lambda^{(1)} = 0.4$ during intervals $t \in [0, 10], [20, 30], [40, 50], \ldots$, otherwise $\lambda^{(1)} = 0$. The intensity of non-priority customers is constant, $\lambda^{(2)} = 0.4$. The queue capacities are limited to $N^{(1)} = N^{(2)} = 20$. The results of the diffusion model are validated by comparison with the exact ones obtained with OMNeT++ discrete network simulator. Poisson input streams and exponential service time distributions for both types of customers are assumed: $\mu^{(1)} = \mu^{(2)} = 1$. Figure 6 displays the mean number of customers of each class as a function of time, given by diffusion model and by the corresponding simulation result and Fig. 7 compares the total number of customers of both classes.

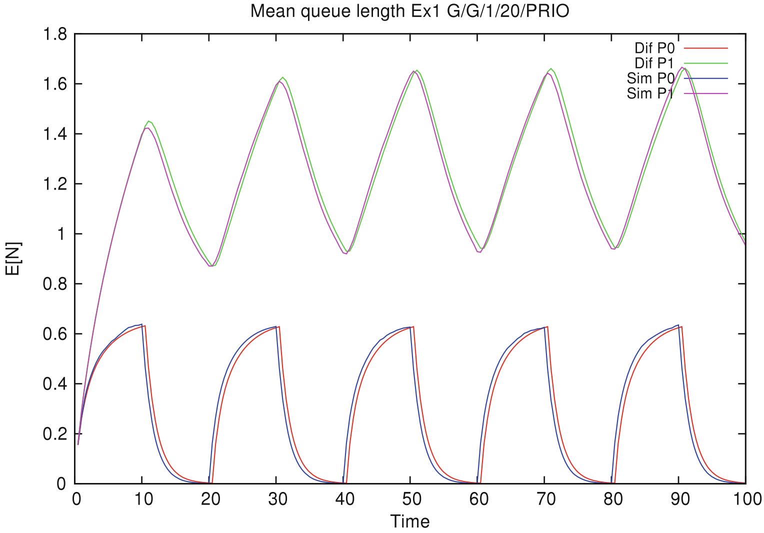

Fig. 6 Example 3.3: mean queue lengths of priority (P0) and non-priority (P1) classes

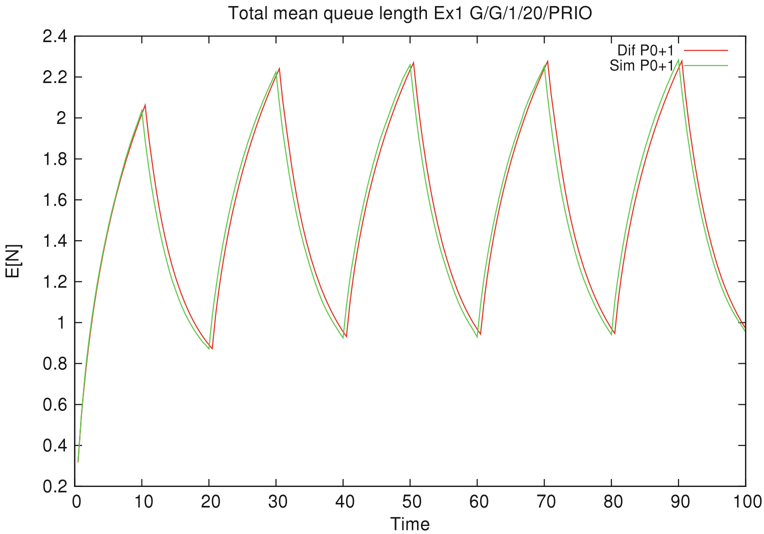

Fig. 7 Example 3.3: total mean queue length of both priority classes

4 Fluid-Flow Approximation

In this method, already adapted to model Internet transmissions [51–53], only the mean values of flow changes are considered and for this reason it should introduce larger errors than the diffusion approach which is a second order approximation.

The fluid approximation uses first-order ordinary linear differential equations to determine the dynamics of the average length of node queues and the dynamics of TCP congestion windows in a modelled network. The changes of a queue length at a station j, $dq_j(t)/dt$, Eq. (33), are defined as the intensity of the input stream, i.e. the sum of all flows $i = 1, \ldots, K$ traversing a particular node, minus the constant intensity of output flow C_j, i.e. the number of packets sent further in a time unit:

$$\frac{dq_j(t)}{dt} = \sum_{i=1}^{K} \frac{W_i(t)}{R_i(\boldsymbol{q}(t))} - \mathbf{1}(q_v(t) > 0)\ C_j\ . \tag{33}$$

A router allows reception of traffic from K TCP flows ($K \leqslant N$), where N is the entire number of flows in the network. Following the TCP congestion avoidance principles, each flow i ($i = 1, ..., N$) is determined by its time varying congestion window size W_i giving the number of packets that may be sent without waiting for acknowledgement of reception of previous packets. The window size, Eq. (34), increases by one at each RTT (round trip time) in the absence of a packet loss and decreases by half of its current value after every packet loss occurring in nodes on the flow path (the latter decision is taken after the time τ). Its size divided by RTT represents the flow intensity. The amount of loss for the entire TCP connection is defined as flow throughput intensity multiplied by total drop probability—the probability which specifies that the loss occurs in nodes on the route. It is based on a matrix $\boldsymbol{B}$ that stores drop probabilities in each router in all flows in the network.

$$\frac{dW_i(t)}{dt} = \frac{1}{R_i(\boldsymbol{q}(t))} - \frac{W_i(t)}{2} \cdot \frac{W_i(t-\tau)}{R_i(\boldsymbol{q}(t-\tau))} \cdot \left(1 - \prod_{j \in V_i} (1 - B_{ij})\right). \tag{34}$$

The values B_{ij} give drop probability $ploss$ at node j for packets of connection i; V_i is the set of nodes belonging to this connection, and $\boldsymbol{q}(t)$ is the vector of queues at these nodes. Delays R_i in the above formulas determine the time needed for the information on congestion and packet loss to propagate through the network back to the sender of a flow i, it consists of queue delays at all nodes j, defined as $q_j(t)/C_j$ along this connection and the propagation delay Tp_i:

$$R_i(\boldsymbol{q}(t)) = \sum_{j \in V_i}^{M} \frac{q_j(t)}{C_j} + Tp_i. \tag{35}$$

The drop probability $ploss_j(x_j)$, Eq. (36), in a single node is determined according to RED mechanism, e.g. [53] as a function on the moving average queue length $x_j(t)$ which is the sum of current queue $q_j(t)$ taken with a weight parameter w and previous average queue taken with $(1 - w)$ weight parameter,

$$ploss_j(x_j) = \begin{cases} 0, & 0 \leqslant x_j < t_{min_j} \\ \dfrac{x_j - t_{min_j}}{t_{max_j} - t_{min_j}} p_{max_j}, & t_{min_j} \leqslant x_j < t_{max_j} \\ 1, & t_{max_j} \leqslant x_j \ . \end{cases} \tag{36}$$

We used already fluid flow approximation to investigate other than RED algorithms in router queues, [54–56].

4.1 Comparison of Fluid Flow Approximation and Diffusion Approximation

We have already reported the ability to build and compute large network models, containing hundreds of thousands nodes in case of fluid flow and thousands of nodes in case of diffusion approximation, [57]. Here, we compare in a simple one-node model the errors of both approaches, cf. [58]. The examples differ in the choice of buffer length, input stream, and RED parameters. The numerical comparisons were conducted in a single node in two phases. The first one (Example 4.1) included the analysis of queue length given by the both methods when the traffic was a predefined function of time, the same for both methods, and the second one (Example 4.2) included the elements of congestion window size mechanism. The results of the approximations were compared to simulations obtained with the use of OMNET++ package [59] adapted by us to simulate transient states (automatic repetition of simulation runs 500,000 times, collection of histograms for a set of defined time moments, parameters of random number generators defined as functions of time). The number of simulation repetitions was sufficient to consider its results as almost accurate, the confidence intervals were negligible compared to the values of obtained queue lengths.

Example 4.1 The considered time-dependent input flow is presented in Fig. 8. The classical diffusion approximation model determines a router queue with the drop-tail (passive) algorithm. To compare this approach with the fluid flow approximation, we decided to disable the window mechanism in fluid flow node and by the choice of RED parameters—thresholds $t_{min} = 0$ packets, $t_{max} = 20$ packets and $pmax = 0$. The buffer length in both cases was set to 20 packets and the service intensity was $\mu = C = 1.5$ packets per time-unit.

The mean queue lengths given by approximation and simulation are displayed in Fig. 9. The diffusion approach is much more accurate than fluid flow and its results are practically the same as simulations, whereas in fluid flow only the general tendency of the queue changes is preserved.

Example 4.2 The fluid flow approximation assumes that the input traffic at a node is defined by congestion window size and RTT time of the flow which traverses that node, whereas in the diffusion model it is given explicitly. To compare the both

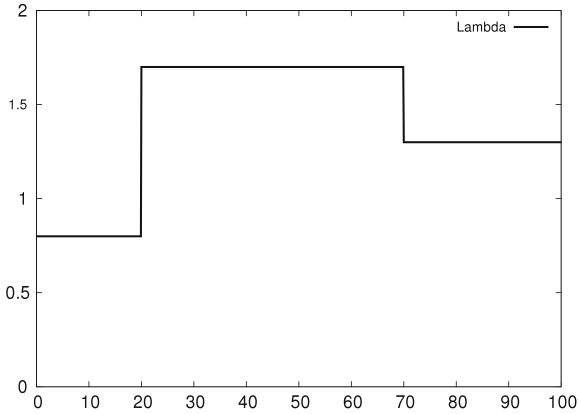

Fig. 8 Input stream as a function of time in Example 4.1—no window mechanism

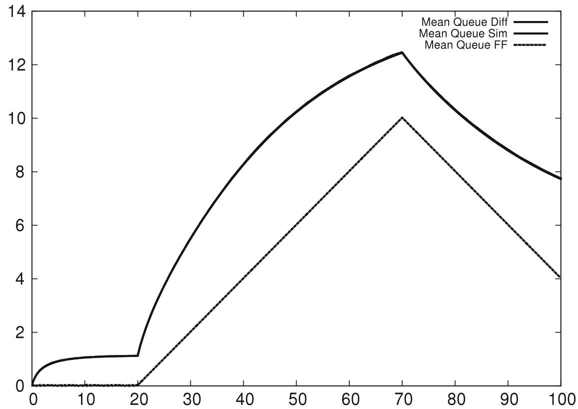

Fig. 9 Example 4.1: Comparison of the mean queue at a single node as a function of time for both approximations and simulation, results in case of disabled window mechanism

models, the time-dependent input stream as shown in Fig. 10 obtained from the fluid flow model where it was defined by the congestion window, was also given as the input flow to the diffusion model. The service intensity, C in fluid flow and μ in diffusion model, were set to the same value of 0.75 packets per time unit. The buffer size was set to 5 packets and RED linear increase range was $t_{min} = 0.75$, $t_{max} = 2.5$ packets, $pmax = 0.1$. The inclusion of the congestion window mechanism makes the generated fluid flow characteristics closer to the results obtained by diffusion approximation and the simulation. However, the error is still important (Fig. 11).

The obtained results demonstrate that the fluid flow method which is frequently used in modelling because of its simplicity generates much larger errors compared to the diffusion results. However, the diffusion calculations are more time-consuming and have also precision limits in its numerical computations that makes impossible to analyse this way networks with large buffers at nodes. Fluid flow model provides only a rough characteristics of network dynamics but can be used to model very large network.

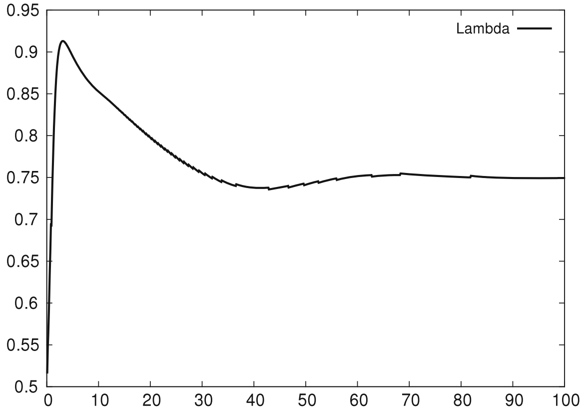

Fig. 10 Example 4.2: Input stream as a function of time for both models in case of window mechanism

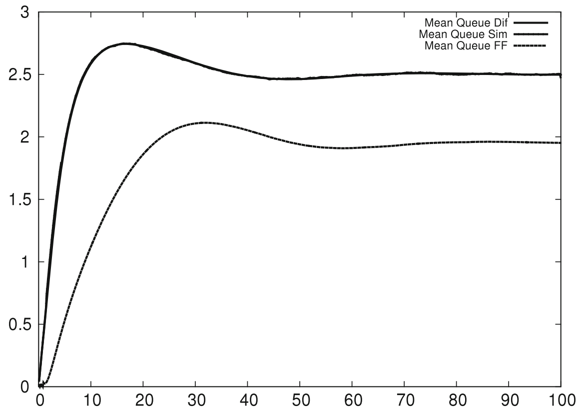

Fig. 11 Example 4.2 Comparison of the mean queues as a function of time for both approximations and simulations results in case of window mechanism

5 Conclusions

The size and complexity of models which may be analysed by diffusion and fluid flow approximations are much larger than in case of traditional Markovian models. Diffusion and fluid approximations are useful approaches that are complementary to Markov models if we do not need a detailed description including all events concerning singular packets and occurring in a real system. We investigated the limitations of the use of both approximations in the transient analysis of IP router queues in presence of input flow originating from TCP congestion window algorithm. Fluid flow approximation generates much larger errors but is very fast and may be applied to larger networks. Diffusion approximation is more accurate and may furnish not only the mean values of queues but also their distributions, therefore it is better adapted to estimate the packet losses. However, the calculations are more complex. An alternative to analytical models is discrete event simulation—also used here to evaluate results of diffusion and fluid flow approximations. We have developed an extension of OMNET++ (a popular simulation tool written in C++, [59]) allowing simulation of transient state models. In this case a simulation run should be repeated a sufficient number of times (e.g. 500 thousands in our examples) and the results for a fixed time should be averaged. It makes transient simulation models time-consuming.

We have developed our own tools for the three analytical methods and we are testing their possibilities. The models based on Markov chains are still essential in performance evaluation and supporting the design of new communication protocols, mechanisms for regulation of the intensity of Internet transmissions and mechanisms to ensure the quality of transmission services. However, they should be limited to small network configurations. The tools based on approximations are able to treat very large (up to millions of nodes) networks giving a software testbed to consider modifications of protocols or the choice of network topologies.

Acknowledgments This work was supported by Polish project NCN nr 4796/B/T02/2011/40 "*Models for transmissions dynamics, congestion control and quality of service in Internet*".

References

1. A.K. Erlang, The theory of probabilities and telephone conversations. Nyt Tidsskr. Mat. B **20**, 33–39 (1909)
2. A.K. Erlang, Solutions of some problems in the theory of probabilities of significance in automatic telephone exchnges. Electroteknikeren **13**, 5–13 (1917)
3. T.O. Engset, Die Wahrscheinlichkeitsrechnung zur Bestimmung der Whlerzahl in automatischen Fernsprechmtern, Elektrotechnische Zeitschrift, Heft 31 (1918)
4. D.G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann. Math. Stat. **24**(3), 338 (1953)
5. L. Kleinrock, *Queueing Systems*, volume I: Theory, volume II: Computer Applications (Wiley, New York, 1975/1976)
6. M. Stasiak, M. Gabowski, A. Wisniewski, P. Zwierzykowski, Modelling and Dimensioning of Mobile Networks, from GSM to LTE, (Wiley, 2011)
7. A.L. Sherr, An analysis of time-shared computer system, Ph.D. Thesis, Project MAC, MIT Press, Cambridge, 1967
8. A.Y. Khinchin, On the average stopping time of machines (in Russian). Mat. Sb. **40**, 119–123 (1933)
9. E.D. Lazowska, J. Zahorjan, G.S. Graham, K.C. Sevcik, *Computer System Analysis Using Queueing Network Models* (Prentice-Hall Inc, New Jersey, 1984)
10. H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evalution Methodology*, Quantitative System Performance (Addison Wesley, Reading, 1978.)
11. E. Gelenbe, *I. Mitrani Analysis and synthesis of computer systems* (Academic Press, London, 1980)
12. W. Willinger, W.E. Leland, M.S. Taqqu, On the self-similar nature of ethernet traffic. IEEE/ACM Trans. Netw. **2**, 1–15 (1994)
13. E. Gelenbe, On approximate computer systems models. J. ACM **22**(2), 261–269 (1975)
14. T. Czachórski, A method to solve diffusion equation with instantaneous return processes acting as boundary conditions. Bulletin of Polish Academy of Sciences. Tech. Sci. **41**(4), 417–451 (1993)
15. F. Baskett, M. Chandy, R. Muntz, J. Palacios, Open, closed and mixed networks of queues with different classes of customers. J. ACM **22**(2), 248–260 (1975)

16. P. Reinecke, T. Krauß, K. Wolter, Hyperstar: phase-type fitting made easy. in *9th International Conference on the Quantitative Evaluation of Systems (QEST) 2012*. 201202 Tool Presentation (September 2012)
17. D.C. Champernowne, An elementary method of solution of the queueing problem with a single server and constant parameters. J. R. Stat. Soc. **B18**, 125–128 (1956)
18. L. Takâcs, *Introduction to the Theory of Queues* (Oxford University Press, Oxford, 1960)
19. A.M.K. Tarabia, Transient analysis of M/M/1/N queue—an alternative approach. Tamkang J. Sci. Eng. **3**(4), 263–266 (2000)
20. T.C.T. Kotiah, Approximate transient analysis of some queueing systems. Oper. Res. **26**(2), 334–346 (1978)
21. S.K. Jones, R.K. Cavin, D.A. Johnston, An efficient computational procedure for the evaluation of the $M/M/1$ transient state occupancy probabilities. IEEE Trans. Commun. **COM–28**(12), 2019–2020 (1980)
22. B. Mandelbrot, J.V. Ness, Fractional brownian motions, fractional noises and applications. SIAM Review, vol. 10 (1968)
23. D.R. Cox, *Long-Range Dependance: A Review*, Statistics: An Appraisal (Lowa State University Press, Lowa, 1984)
24. I. Norros, On the use of fractional Brownian motion in the theory of connectionless networks. IEEE J. Sel. Areas Commun. **13**(6), 953–962 (1995)
25. T. Mikosch, S. Resnick, H. Rootzen, A. Stegeman, Is network traffic approximated by stable levy motion or fractional Brownian motion? Anal. Appl. Probab. **12**(1), 23–68 (2002)
26. A. Erramilli, R.P. Singh, P. Pruthi, An application of determinic chaotic maps to model packet traffic. Queueing Syst. **20**(1–2), 171–206 (1995)
27. J.R. Gallardo, D. Makrakis, L. Orozco-Barbosa, Use a α-stable self-similar stochastic processes for modeling traffic in broadband networks. Perform. Eval. **40**(1–3), 71–98 (2000)
28. F. Harmantzis, D. Hatzinakos, Heavy network traffic modeling and simulation using stable FARIMA processes. in *19th International Teletraffic Congress* (Beijing, 2005)
29. N. Laskin, I. Lambadatis, F.C. Harmantzis, M. Devetsikiotis, Fractional levy motion and its application to network traffic modeling. Comput. Netw. **40**(3), 363–375 (2002)
30. G. Casale, Building accurate workload models using Markovian arrival processes, SIGMETRICS'11, (San Jose, U.S.A., 2011), pp. 7–11
31. A.T. Andersen, B.F. Nielsen, A markovian approach for modeling packet traffic with long-range dependence. IEEE J. Sel. Areas Commun. **16**(5), 719–732 (1998)
32. W. Fischer, K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook. Perform. Eval. **18**(2), 149–171 (1993)
33. J. Domanska, A. Domanski, T. Czachorski, *Internet Traffic Source Based on Hidden Markov Model, NEW2AN 2011* (St. Petersburg, Russia, 2011)
34. D. Potier, New user's introduction to QNAP2, Rapport Technique no. 40, INRIA, Rocquencourt (1984)
35. W. Stewart, *Introduction to the Numerical Solution of Markov Chains* (Princeton University Press, Chichester, 1994)
36. PEPS, www-id.imag.fr/Logiciels/peps/userguide.html
37. PRISM—probabilistic model checker, www.prismmodelchecker.org/
38. C. Moler, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. **45**(1), 30–49 (2003)
39. P. Pecka, S. Deorowicz, M. Nowak, Efficient representation of transition matrix in the markov process modeling of computer networks, in *Man-Machine Interactions 2, Advances in Intelligent and Soft Computing no. 103*, ed. by T. Czachórski, et al. (Springer, 2011), pp. 457–464
40. C. Scientifique, B. Philippe, R.B. Sidje, Transient solutions of Markov processes by krylov subspaces. *2nd International Workshop on the Numerical Solution of Markov Chains* (1989)
41. R.B. Sidje, K. Burrage, S. McNamara, Inexact uniformization method for computing transient distributions of Markov chains. SIAM J. Sci. Comput. **29**(6), 2562–2580 (2007)
42. R.B. Sidje, W.J. Stewart, A numerical study of large sparse matrix exponentials arising in Markov chains. Comput. Stat. Data Anal. **29**, 345–368 (1999)

43. R.B. Sidje, Expokit: a software package for computing matrix exponentials. ACM Trans. Math. Softw. **24**(1), 130–156 (1998)
44. Numerical computation for Markov chains on GPU: building chains and bounds, algorithms and applications. Project POLONIUM 2012–2013, bilateral cooperation PRISM-Université de Versailles and IITiS PAN, Polish Academy of Sciences
45. E. Gelenbe, G. Pujolle, The behaviour of a single queue in a general queueing network. Acta Inform. **7**(Fasc. 2), 123–136 (1976)
46. R.P. Cox, H.D. Miller, *The Theory of Stochastic Processes* (Chapman and Hall, London, 1965)
47. H. Stehfest, Algorithm 368: numeric inversion of laplace transform. Commun. ACM **13**(1), 47–49 (1970)
48. P.J. Burke, The output of a queueing system. Oper. Res. **4**(6), 699–704 (1956)
49. T. Czachorski, J.-M. Fourneau, T. Nycz, F. Pekergin, Diffusion approximation model of multiserver stations with losses. Electron. Notes Theor. Comput. Sci. **232**, 125–143 (2009)
50. T. Czachrski, K. Grochla, T. Nycz, F. Pekergin, Diffusion approximation models for transient states and their application to priority queues. IARIA J. Int. J. Adv. Netw. Serv. **2**(3), 205–217 (2009)
51. K. Hollot, Y. Liu, V. Misra, D. Towsley, W.B. Gong, Fluid methods for modeling large heterogeneous networks. Technical report AFRL-IF-RS-TR-2005-282 (2005)
52. Y. Liu, F. Lo Presti, V. Misra, Y. Gu, Fluid Models and Solutions for Large-Scale IP Networks, ACM/SigMetrics (2003)
53. V. Misra, W. Gong, D. Towsley, A fluid-based Analysis of a network of AQM routers supporting TCP flows with an application to RED. in *Proceedings of the Conference on Applications, Technologies, Architectures and Protocols for Computer Communication (SIGCOMM 2000)*, pp. 151–160 (2000)
54. A. Domański, J. Domańska, T. Czachórski, *Comparison of CHOKe and gCHOKe Active Queues Management Algorithms with the use of Fluid Flow Approximation*, Communications in Computer and Information Science, vol 370 (Springer, Berlin, 2013)
55. A. Domański, J. Domańska, T. Czachórski, *Comparison of AQM Control Systems with the Use of Fluid Flow Approximation*, Communications in Computer and Information Science, vol 291 (Springer, Heidelberg, 2012)
56. J. Domańska, A. Domański, T. Czachórski, *Fluid Flow Analysis of RED Algorithm with Modified Weighted Moving Average*, Communications in Computer and Information Science, vol 356 (Springer, Berlin, 2013)
57. T. Czachórski, M. Nycz, T. Nycz, F. Pekergin, Analytical and numerical means to model transient states in computer networks', in *20th International Conference, CN 2013*, (Lwowek Slaski, Poland, June 17–21, 2013). Springer Proceedings Series: Communications in Computer and Information Science, Vol. 370, pp. 426–435, ISBN: 978-3-642-38864-4
58. T. Nycz, M. Nycz, T. Czachórski, A numerical comparison of diffusion and fluid-flow approximations used in modelling transient states of TCP/IP networks, in *Proceedings of Computer Networks*, ed. by A. Kwiecie, P. Gaj, P. Stera (Springer, Berlin, 2014)
59. OMNET++ Community Site, www.omnetpp.org

http://www.springer.com/978-3-319-12147-5

Analytic Methods in Interdisciplinary Applications
Mityushev, V.; Ruzhansky, M.V. (Eds.)
2015, X, 181 p. 61 illus., 40 illus. in color., Hardcover
ISBN: 978-3-319-12147-5