A Discrete Time Queueing Model for End-to-end Delay and Jitter Analysis

Olav Østerbø

Abstract— **In this paper we consider a discrete time queueing model to calculate end-to-end delays and jitter in packet networks. The arrival process is formed by a discrete time renewal (foreground stream) and a batch process (background stream). By this model we are able to trace the foreground stream as it passes through a multiplexer where it will be disturbed by crossing packet streams (background traffic). In particular we have analyzed the distribution of the time between two successive departures. By approximating the output foreground stream with a renewal stream and applying this as the input to the next node, by applying the queueing model recursively, we get an end-to-end description. Especially the evolution of the jitter, but also the end-to-end delay will be analyzed more accurately than given by the conventional convolutions approach. Some numerical examples are given and discussed for both evolution of jitter and end-to-end delay. For the end-to-end delay we also compare the results with those obtained by convolution.**

Index Terms— **End-to-end delay, jitter, discrete time queueing.**

I. Introduction

IT is reason to believe that real time services will be a significant part of the traffic offered in future multi service networks. Real time services require a quite strict control regime to be able to maintain QoS (Quality of Service). It is well known that networks based on statistical multiplexing (like IP-networks) will introduce certain disturbances (delay and jitter) in the bit stream mainly due to queuing in routers (or switches). The end-to-end QoS is realized through the contributions from the different domains and the QoS guarantees end-to-end will be realized through different SLAs (Servic Level Agreement) between the customer and the access network and/or between different core network domains. For each domain it will be important to estimate the contribution to the QoS parameters since each administrative domain will be responsible for their own contribution through the SLA. The most important parameters will typically be delay, jitter and information losses due to buffer overflow. It will be important for a network operator to be able to estimate the QoS parameters in his own domain to set the appropriate parameters in the SLAs. In addition it will be of vital importance to implement the necessary control structures that makes it possible to maintain the guarantees, especially in own domain.

Generally it is a challenging task to analyze end-to-end delay and jitter in communication networks. Often one limits the effort to consider a particular network node and then applies the single node results to obtain some kind of end-to-end estimates. However, this approach yields only approximation of the end-to-end delay and the accuracy is difficult to predict. On the other hand a complete description with several traffic types with corresponding analysis of end-to-end delay in a communication network with inter node dependencies is not feasible. In this paper we describe a method that is somehow in between the single node approach and a complete network description to obtain end-to-end results. This is done by performing an exact analysis of a particular traffic stream as it passes through a multiplexer in the network and then apply these results recursively to obtain the end-to-end results. The only approximation done by this approach is that we approximate the output process by a renewal process, i.e. we neglect the dependencies inflicted the output stream.

The rest of the paper is organized as follows. In section 2 we briefly discuss end-to-end delay models based on convolution. We then describe and analyze the discrete time queueing model which is the basis for the end-to-end analysis in section 3. In section 4 the end-to-end delay and jitter is derived on the basis of the discrete time queueing model. Then some numerical results are discussed in section 5 and finally in section 6 some concluding remarks are given.

II. End-to-end delay models based on convolutions

In previous papers [1] and [2] we calculated end-to-end delay in large scale IP-networks by assuming that the delay in each router is independent and hence the end-to-end delay can be obtained by convolution. For FCFS (First-Come First-Served) exact results is only available for acyclic type of Jackson Networks (where a packet visits a node at most once), see [3]. In [4], however, it is argued that if the load from a particular flow only is a small fraction of the total load in a node and the input processes to the network is 'smoother than Poisson' (i.e. with less variability) then the independent assumption will be quite reasonable and will represent a worst case scenario. We therefore used the M/G/1 queueing model to obtain the waiting time distribution in each node and then applied the convolution to obtain the end-to-end delay. It

Manuscript received January 24, 2009; accepted May 4 2009.
Olav N. Østerbø is with Telenor R&I, 1331 Fornebu, Norway (phone: +47 48212596; fax: +47 67891812; e-mail: olav-norvald.osterbo at telenor.com).

turned out the convolution may be heavily simplified for some cases of particular interest. In [1] we proved that the convolution of the waiting times of N identical M/G/1 queues may be written in terms of the waiting time distribution of a single queue through partial derivative with respect to the load:

$$D^N(x;\rho)=\frac{(1-\rho)^N}{(N-1)!}\frac{\partial^{N-1}}{\partial\rho^{N-1}}\left\{\frac{\rho^{N-1}}{1-\rho}D(x;\rho)\right\} \quad (1)$$

where $D(x;\rho)$ is the corresponding waiting time distribution for a single M/G/1 queue with load ρ, i.e. it is possible to obtain the convolution by plain derivatives. For deterministic service times, i.e. the M/D/1 case, we obtained the following expression for the convolution:

$$D^N(x;\rho)=(1-\rho)^N\sum_{k=0}^{\lfloor x\rfloor}\sum_{l=0}^{N-1}\frac{(-1)^l}{l!k!}\binom{N+k-1}{k+l}(\rho(k-x))^{k+l}e^{-\rho(k-x)} \quad (2)$$

(where the service times is scaled to unity).

In applications the delay expression (2) may be used to find the α-quantile by solving for $t^N(\alpha,\rho)$ from the equation $D^N(t;\rho)=1-\alpha$ or alternative one may fix the α-quantile t and solve for the maximum possible load $\rho^N(\alpha,t)$.

III. Discrete time queueing models

The "critical assumption" for models based on convolution is of course the independence assumption needed to obtain the corresponding Laplace transform on product form. One other approach would be to analyze a particular traffic stream (flow) as it traverses a multiplexer and try to capture the characteristics of that particular traffic process (flow) at the output. This particular output process will then mingle with other traffic streams and will constitute the input to the next multiplexer in the chain. By this approach we are able to trace a particular stream (flow) and describe the distortion as it passes through a particular path through the network. Similar approach to study end-to-end behavior is well documented in the literature. (See [5], [6].)

It turns out that discrete time queueing model is easier to analyze than the corresponding continuous time counterpart, due the discrete nature of the corresponding models, and the corresponding analyze tool will be based on generating function rather than Laplace transforms usually applied for continuous time models. In the following we consider a discrete (slotted) queueing model where we put the main emphasis of tracing a particular traffic stream as it passes through a multiplexer where it may be disturbed by crossing packet streams (background traffic). We are particularly interested in describing the output process of that particular stream which in turn will be part of the input traffic to the next multiplexer.

A. Transient queueing analysis

The queueing model taken as basis of the analysis is depicted in Fig. 3 below. It is a single server, infinite capacity queue operating in discrete (or slotted) time with two classes of customers. Any activities in the system, e.g. arrivals, departures, etc., are assumed to occur at slot boundaries.

The arrival process is formed by superposing of a discrete time renewal process, foreground stream (***FS***), and a (discrete) batch arrival process (e.g. Poisson or Bernoulli process), background stream (***BS***). We let the slots be successively numbered $k=0,1,2,...$ and we assume that the batch size in slot k, generated by the ***BS***, B_k is independent and follows a general (discrete) distribution $b(i)=P(B_k=i)$ with generating function $B(z)$. The ***FS*** renewal process is characterized by the distribution of the number of slots between arrivals $A_n=T_{n+1}-T_n$ where T_n (we take $T_0=0$) is the slot number for the n'th arrival of the FS; $n=0,1,2,...$, and we assume A_n that is independent (of n and of ***BS***) and follows a general (discrete) distribution $a(i)=P(A_n=i)$ with generating function $A(z)$.

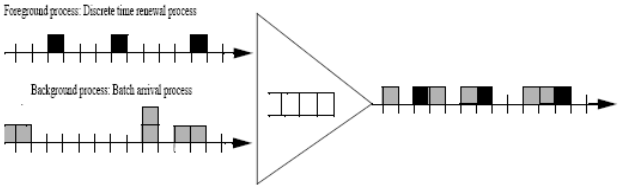

Fig. 1. The queueing model for the packet multiplexer.

The total load on the multiplexer is $\rho=\rho_{FS}+\rho_{BS}$ where $\rho_{FS}=1/A'(1)$ is the load of ***FS***, and $\rho_{BS}=B'(1)$ is the load of ***BS***, and we also assume that $\rho<1$ to secure that the queueing system is stabile.

Below we provide a complete transient analyze of the packet multiplexer described above where we consider joint distribution of the number of packets in the queue just prior to the n'th arrival from the ***FS***, Q_n, and the corresponding slot number T_n. (See also [7] chapter 9)

Theorem I: The generating function of the double z-transform $Q_n(z,x)=E\left[z^{Q_n}x^{T_n}\right]$, $Q(z,x,s)=\sum_{n=1}^{\infty}s^{n-1}Q_n(z,x)$ is given as the contour integral:

$$Q(z,x,s)=\frac{1}{2\pi i}\frac{z}{B(z)}\int_{C_u}\frac{\left(1-\frac{1}{z}\right)}{\left(1-\frac{1}{\zeta}\right)}\frac{\zeta^m A\left(x\frac{B(\zeta)}{\zeta}\right)\left(\frac{1}{B(\zeta)}-\frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{\left(\frac{\zeta}{B(\zeta)}-\frac{z}{B(z)}\right)\left(1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)\right)}\exp[I(z,x,s)-I(\zeta,x,s)]d\zeta \quad (3)$$

where m is the number of packets in the queue at (the end of) the slot just after the arrival of the first packet of the ***FS*** (i.e. $n=0$), and C_u is the disc $\{|\zeta|=u\}$ with $u<1$ and where the radius u is chosen so large that both $\varsigma=z$ and all roots of $1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)=0$ inside the unit disc are included in C_u. And further

$$I(z,x,s)=\frac{1}{2\pi i}\int_{C_r}\frac{\log\left[1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)\right]}{\left(\frac{\zeta}{B(\zeta)}-\frac{z}{B(z)}\right)}\left(\frac{1}{B(\zeta)}-\frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)d\zeta \quad (4)$$

where C_r is the disc $\{|\zeta|=r\}$ with $1<r<r_M$, where r_M is the roots of $1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)=0$ outside the unit disc with smallest modulo. (It is possible to show that the root r_M is real.) Further the z-transform of the steady state queue length distribution $Q_0(z)$ is given by:

$$Q_0(z)=(1-B'(1))\frac{(z-1)}{(z-B(z))}\exp[I(z,1,1)-I(1,1,1)] \quad (5)$$

Proof: We observe the queue size at the end of each slot, and we define the following stochastic variables: Q_n-the number of packets in the queue at the end of the slot just prior to the n'th arrival of a packet from the ***FS***, and by conditioning on $A_n=T_{n+1}-T_n=k$ we let Q_n^i -the number of packets in the queue in the end of the slot T_n+i (for $i=1,...,k$). If we denote B_n^i the number of packets arriving from ***BS*** during slot T_n+i, then we have the following relation between the queue lengths in the different slots: $Q_n^1=Q_n+B_n^1$ and $Q_n^i=\left[Q_n^{i-1}+B_n^i-1\right]^+$ for $i=2,...,k$ and we also obviously have $Q_{n+1}=Q_n^k$. We now define the joint generating functions:

$Q_n^i(z,x)=E[z^{Q_n^i}x^{T_n}|T_{n+1}-T_n=k]$ for $i=1,...,k$.

By applying the queueing relations above we find the following recursions:

$Q_n^i(z,x)=Q_n(z,x)B(z)$ and

$$Q_n^i(z,x)=Q_n^{i-1}(z,x)\frac{B(z)}{z}+(1-\frac{1}{z})q_n^{i-1}(x) \text{ for } i=2,...,k$$

where $q_n^i(x)=E[x^{T_n}1_{\{Q_n^i+B_n^{i+1}=0\}}|T_{n+1}-T_n=k]$ is the boundary transform, taking into account the probability of having an empty queue in slot T_n+i+1 $(i=1,...,k-1)$. Hence, by solving recursively we obtain $Q_n^i(z,x)$ in terms of $Q_n^1(z,x)$:

$$Q_n^k(z,x)=Q_n^1(z,x)\left(\frac{B(z)}{z}\right)^{k-1}+(1-\frac{1}{z})\sum_{l=1}^{k-1}q_n^l(x)\left(\frac{B(z)}{z}\right)^{k-l-1}$$

for $k\geq 2$. For $n=0$ we make the conditions that $T_0=0$ and $Q_0^1=m$ implying:

$$Q_0^k(z,x)=z^m\left(\frac{B(z)}{z}\right)^{k-1}+(1-\frac{1}{z})\sum_{l=1}^{k-1}q_0^l(x)\left(\frac{B(z)}{z}\right)^{k-l-1} \text{ and}$$

$$Q_n^k(z,x)=zQ_n(z,x)\left(\frac{B(z)}{z}\right)^{k-1}+(1-\frac{1}{z})\sum_{l=1}^{k-1}q_n^l(x)\left(\frac{B(z)}{z}\right)^{k-l-1}$$

for $n\geq 1$. Now since $Q_{n+1}(z,x)=\sum_{k=1}^{\infty}a(k)x^kQ_n^k(z,x)$ we find:

$$Q_1(z,x)=z^mA\left(x\frac{B(z)}{z}\right)+(1-\frac{1}{z})\sum_{k=0}^{\infty}\tilde{q}_0^k(x)\left(\frac{B(z)}{z}\right)^k \text{ and}$$

$$Q_{n+1}(z,x)=zQ_n(z,x)A\left(x\frac{B(z)}{z}\right)+(1-\frac{1}{z})\sum_{k=0}^{\infty}\tilde{q}_n^k(x)\left(\frac{B(z)}{z}\right)^k$$

For $n\geq 1$, where we define the boundary transforms $\tilde{q}_n^k(x)=\sum_{l=1}^{\infty}a(k+l+1)x^{k+l+1}q_n^l(x)$. Finally by combining the transforms above by introducing the generating functions

$Q(z,x,s)=\sum_{n=1}^{\infty}s^{n-1}Q_n(z,x)$ and $q^k(x,s)=\sum_{n=0}^{\infty}s^n\tilde{q}_n^k(x)$ we obtain the following expression for $Q(z,x,s)$:

$$Q(z,x,s)=\frac{z^mA\left(x\frac{B(z)}{z}\right)\frac{z}{B(z)}+(1-\frac{1}{z})\sum_{k=0}^{\infty}q^k(x,s)\left(\frac{B(z)}{z}\right)^k}{1-szA\left(x\frac{B(z)}{z}\right)} \quad (6)$$

It remains to determine the unknown coefficients $q^k(x,s)$ in (6). To do so we shall first assume that there is a maximum number K so that $a(K+1)\neq 0$ but $a(i)=0$ for $i>K+1$. (This restriction we put on the discrete distribution is quite weak; since we always may approximate an infinite (countable) discrete distribution by a finite one by simply choosing K large enough. Alternative we may also take the limit $K\to\infty$ in the final result.) With this assumption we may apply the powerful method, often used to determine unknown coefficients in generating functions given as a fraction, by locating the zeros of the dominator inside the unit disc and claiming analytical behavior of the transforms in the same domain. From the definition we have $q^k(x,s)=0$ for $k>K-1$. Next by applying the famous Rouche's theorem we have that for $|s|<1$ and $|x|\leq 1$ that the equation $1-szA\left(x\frac{B(z)}{z}\right)=0$ will have exactly K distinct roots $r_j=r_j(x,s)$ inside the unit disc. Moreover, by letting $z\to r_j$ we therefore must have:

$$\sum_{k=0}^{K-1}q^k(x,s)\left(\frac{B(r_j)}{r_j}\right)^k=-\frac{{r_j}^m}{(1-\frac{1}{r_j})}A\left(x\frac{B(r_j)}{r_j}\right)\frac{r_j}{B(r_j)} \quad (7)$$

for $j=1,..,K$. (7) is a linear system of Vandermonde type that is easy to solve and the corresponding sought generating function is found in):

$$\sum_{k=0}^{K-1}q^k(x,s)\left(\frac{B(z)}{z}\right)^k=-\frac{{r_j}^m}{(1-\frac{1}{r_j})}A\left(x\frac{B(r_j)}{r_j}\right)\frac{r_j}{B(r_j)}\prod_{l=1,l\neq j}^{K}\frac{\frac{B(z)}{z}-\frac{B(r_l)}{r_l}}{\frac{B(r_j)}{r_j}-\frac{B(r_l)}{r_l}} \quad (8)$$

By *Lemma I* (given in the appendix) we have

$$\prod_{l=1,l\neq j}^{K}(\frac{z}{B(z)}-\frac{r_l}{B(r_l)})=\left(\frac{z}{B(z)}\right)^K\frac{h(z)}{\frac{z}{B(z)}-\frac{r_j}{B(r_j)}}\exp[I(z,x,s)] \quad \text{where}$$

$h(z)=1-szA\left(x\frac{B(z)}{z}\right)$ and $I(z,x,s)$ is given by (4). By taking the limit $z\to r_j$ we also find that

$$\prod_{l=1,l\neq j}^{K}(\frac{r_j}{B(r_j)}-\frac{r_l}{B(r_l)})=\left(\frac{r_j}{B(r_j)}\right)^K\frac{h'(r_j)}{\frac{1}{B(r_j)}-\frac{r_jB'(r_j)}{B(r_j)^2}}\exp[I(r_j,x,s)].$$

Finally, by inserting for (8) in (6) we find:

$$Q(z,x,s)=\frac{z}{B(z)}\left[\frac{z^m A\left(x\frac{B(z)}{z}\right)\frac{z}{B(z)}}{1-szA\left(x\frac{B(z)}{z}\right)}+\right.$$
$$\left.\sum_{j=1}^{K}\frac{(1-\frac{1}{z})r_j{}^m A\left(x\frac{B(r_j)}{r_j}\right)(\frac{1}{B(r_j)}-\frac{r_jB'(r_j)}{B(r_j)^2})}{(1-\frac{1}{r_j})(\frac{z}{B(z)}-\frac{r_j}{B(r_j)})h'(r_j)}\exp[I(z,x,s)-I(r_j,x,s)]\right] \quad (9)$$

We now recognize (9) as the residue expansion of the integral (3) (with residues $\varsigma=z$ and $\varsigma=r_j$ for $j=1,..,K$). For $\rho<1$ the steady state distribution exists and may be found by applying the Tauberian theorem [8]:
$Q_0(z)=E[z^{Q_0}]=\lim_{s\to 1}(1-s)Q(z,1,s)$.

For $x=1$ one of the roots $r_j=r_j(1,s)\to 1$ when $s\to 1$. We denote this root as r_1 and we find $\frac{\partial r_1}{\partial s}(1,1)=\frac{1}{A'(1)(1-\rho)}$. By applying (9) when taking the limit, we only get contribution from the root $j=1$ and we obtain (5) .■

Based on the transient transform (3) and (4) and the corresponding stationary transform (5) we may find the characteristics of the output process for the ***FS*** when the queueing system is in steady state. Although we consider FCFS (First-Come First-Serve) queueing discipline the ordering of between ***FS*** and ***BS*** packets arriving in the same slot has to be specified. Below we have analyzed three possible orderings for a ***FS*** packet when it arrives in a slot with (possible several) ***BS*** packets:

- ***F***-The ***FS*** packet is always served first when it arrives together with ***BS*** packets,
- ***R***-The ***FS*** packet and possible ***BS*** packets is served at random,
- ***L***-The ***FS*** packet is always served last when it arrives together with ***BS*** packets.

Among these three orderings the random will be the most important one since this requires no special treatment packets (from any arrival streams). However we include the ***F*** and ***L*** disciplines since they represent lower and upper bounds for the delay and jitter.

We denote D_n the delay (in slots) for the n'th packet arrival from the FS. Then we have $D_n=Q_n+U_n$ where Q_n is the queue length at the arrival instants, and U_n is the delay for a ***FS*** packet due to possible arrivals of ***BS*** packets in the same slot. We are interested in the joint distribution of D_n-D_0 and T_n and we define the double z-transform $W_n(z,x)=E[z^{D_n-D_0}x^{T_n}]$ and the generating function $W(z,x,s)=\sum_{n=0}^{\infty}s^{n-1}W_n(z,x)$.

Theorem II: Based on the three orderings ***F***, ***R*** and ***L***, and using the subscript ***F***, ***R*** and ***L*** to separate the three cases, we find the following contour integral of the transform $W(z,x,s)$:

$$W_F(z,x,s)=\frac{1}{2\pi i}\frac{z}{B(z)}\int_{C_w}\frac{\left(1-\frac{1}{z}\right)}{\left(1-\frac{1}{\zeta}\right)}\frac{B(\zeta)A\left(x\frac{B(\zeta)}{\zeta}\right)Q_0(\frac{\zeta}{z})\left(\frac{1}{B(\zeta)}-\frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{\left(\frac{\zeta}{B(\zeta)}-\frac{z}{B(z)}\right)\left(1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)\right)}\exp[I(z,x,s)-I(\zeta,x,s)]d\zeta \quad (10)$$

$$W_R(z,x,s)=\frac{1}{2\pi i}\frac{zBI(z)}{(z-1)B(z)}\int_{C_w}\frac{\left(BI(\zeta)-BI(\frac{\zeta}{z})\right)}{(\zeta-1)}\frac{A\left(x\frac{B(\zeta)}{\zeta}\right)Q_0(\frac{\zeta}{z})\left(\frac{1}{B(\zeta)}-\frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{\left(\frac{\zeta}{B(\zeta)}-\frac{z}{B(z)}\right)\left(1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)\right)}\exp[I(z,x,s)-I(\zeta,x,s)]d\zeta \quad (11)$$

$$W_L(z,x,s)=\frac{1}{2\pi i}\int_{C_w}\frac{(z-1)}{\left(1-\frac{1}{\zeta}\right)}\frac{B(\frac{\zeta}{z})A\left(x\frac{B(\zeta)}{\zeta}\right)Q_0(\frac{\zeta}{z})\left(\frac{1}{B(\zeta)}-\frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{\left(\frac{\zeta}{B(\zeta)}-\frac{z}{B(z)}\right)\left(1-s\zeta A\left(x\frac{B(\zeta)}{\zeta}\right)\right)}\exp[I(z,x,s)-I(\zeta,x,s)]d\zeta \quad (12)$$

where $BI(z)=\int_1^z B(x)dx$ is the integral of $B(z)$.

Proof: For the case ***F*** we have $D_n=Q_n+1$ and hence $D_n-D_0=Q_n-Q_0$. By conditioning on $Q_0=k$ and $Q_0^1=B_0^1+Q_0=m$ we have
$W_F(z,x,s)=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}z^{-k}Q(z,x,s)P(Q_0=k,Q_0^1=m)$. By applying the integral expression (3) for $Q(z,x,s)$ and the joint transform $E[z_1^{Q_0}z_2^{Q_0^1}]=E[(z_1z_2)^{Q_0}]E[z_2^{B_0^1}]=B(z_2)Q_0(z_1z_2)$ we obtain (10). For the case ***R*** we have $D_n=Q_n+U_n$ where U_n is the number of ***BS*** packets arriving in the same slot as a ***FS*** packet and are placed prior to the ***FS*** packets when the mutual position among them is chosen at random. Hence, $D_n-D_0=U_n+Q_n-D_0$. By conditioning on $D_0=k$ and $Q_0^1=B_0^1+Q_0=m$ we have
$W_R(z,x,s)=E[z^{U_n}]\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}z^{-k}Q(z,x,s)P(D_0=k,Q_0^1=m)$. Since $Q_0^1=B_0^1+Q_0$ and $D_0=U_0+Q_0$ the transform $E[z_1^{D_0}z_2^{Q_0^1}]=E[z_1^{U_0}z_2^{B_0^1}]E[(z_1z_2)^{Q_0}]=E[z_1^{U_0}z_2^{B_0^1}]Q_0(z_1z_2)$.

Now $P(U_n=i\,|\,B_n^1=j)=\frac{1}{j+1}$ for $i=1,....,j+1$ and the joint distribution $P(U_n=i,B_n^1=j)=\frac{1}{j+1}b(j)$ for $i=1,....,j+1$ and $j=0,1,..,$. Hence

$$E[z_1^{U_n}z_2^{B_n^1}]=\sum_{j=0}^{\infty}\sum_{i=1}^{j+1}z_1^{\,i}z_2^{\,j}\frac{b(j)}{j+1}=\frac{z_1}{z_2(1-z_1)}(BI(z_2)-BI(z_1z_2)) \quad (13)$$

where $BI(z)=\int_1^z B(x)dx$. By applying the integral expression (3) for $Q(z,x,s)$, and the joint transform
$E[z_1^{D_0}z_2^{Q_0^1}]=\frac{z_1Q_0(z_1z_2)}{z_2(1-z_1)}(BI(z_2)-BI(z_1z_2))$ and the marginal transform $E[z^{U_n}]=\frac{z}{z-1}BI(z)$ we obtain (11). For the case ***L*** we have $D_n=Q_n+B_n^1+1$ so $D_n-D_0=B_n^1+Q_n-Q_0^1$. By conditioning on $Q_0^1=m$ we have

$W_L(z,x,s) = E[z^{B_n^1}]\sum_{k=0}^{\infty}\sum_{m=0}^{\infty} z^{-k} Q(z,x,s) P(Q_0^1 = m)$. By applying the integral expression (3) for $Q(z,x,s)$, and the transform $E[z^{Q_0^1}] = E[z^{Q_0}]E[z^{B_0^1}] = B(z)Q_0(z)$ we obtain (12).

B. Jitter and inter-departure distributions

On the bases of the joint z-transforms (10), (11) and (12), one for each of queueing disciplines defined, makes it easy to find the distribution of inter-departure time, jitter and queueing delay. We have that the jitter J is the difference of the delay for two succeeding packets from the ***FS***, $J = D_1 - D_0$, and the corresponding z-transform is simply $J(z) = E[z^{D_1 - D_0}] = W(z,1,0)$. Similar the inter-departure time between succeeding packet G from the ***FS*** is simply the difference between departure times, $G = D_1 - D_0 + T_1$, and the z-transform is given by $G(z) = E[z^{D_1 - D_0 + T_1}] = W(z,z,0)$. We also denote $D(z) = E[z^{D_0}]$ the z-transform of the delay for a packet from the FS. We obtain the following expressions:

1. ***F***-The ***FS*** packet is always served first when it arrives together with ***BS*** packets:

$$J_F(z) = \frac{1}{2\pi i}\frac{z}{B(z)}\int_{C_u} \frac{\left(1-\frac{1}{z}\right)B(\zeta)A\left(\frac{B(\zeta)}{\zeta}\right)Q_0\left(\frac{\zeta}{z}\right)\left(\frac{1}{B(\zeta)} - \frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{\left(1-\frac{1}{\zeta}\right)\left(\frac{\zeta}{B(\zeta)} - \frac{z}{B(z)}\right)} d\zeta \quad (14)$$

$$G_F(z) = \frac{1}{2\pi i}\frac{z}{B(z)}\int_{C_u} \frac{\left(1-\frac{1}{z}\right)B(\zeta)A\left(z\frac{B(\zeta)}{\zeta}\right)Q_0\left(\frac{\zeta}{z}\right)\left(\frac{1}{B(\zeta)} - \frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{\left(1-\frac{1}{\zeta}\right)\left(\frac{\zeta}{B(\zeta)} - \frac{z}{B(z)}\right)} d\zeta \quad (15)$$

and $D_F(z) = zQ_0(z)$, (16)

2. ***R***-The ***FS*** packet and possible ***BS*** packets are served at random:

$$J_R(z) = \frac{1}{2\pi i}\frac{zBI(z)}{(z-1)B(z)}\int_{C_u} \frac{\left(BI(\zeta) - BI(\frac{\zeta}{z})\right)A\left(\frac{B(\zeta)}{\zeta}\right)Q_0\left(\frac{\zeta}{z}\right)\left(\frac{1}{B(\zeta)} - \frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{(\zeta-1)\left(\frac{\zeta}{B(\zeta)} - \frac{z}{B(z)}\right)} d\zeta \quad (17)$$

$$G_R(z) = \frac{1}{2\pi i}\frac{zBI(z)}{(z-1)B(z)}\int_{C_u} \frac{\left(BI(\zeta) - BI(\frac{\zeta}{z})\right)A\left(z\frac{B(\zeta)}{\zeta}\right)Q_0\left(\frac{\zeta}{z}\right)\left(\frac{1}{B(\zeta)} - \frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{(\zeta-1)\left(\frac{\zeta}{B(\zeta)} - \frac{z}{B(z)}\right)} d\zeta \quad (18)$$

and $D_R(z) = \frac{z}{z-1} BI(z)Q_0(z)$, (19)

3. ***L***- The ***FS*** packet is always served last when it arrives together with ***BS*** packets:

$$J_L(z) = \frac{1}{2\pi i}\int_{C_u} \frac{(z-1)B(\frac{\zeta}{z})A\left(\frac{B(\zeta)}{\zeta}\right)Q_0\left(\frac{\zeta}{z}\right)\left(\frac{1}{B(\zeta)} - \frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{(1-\frac{1}{\zeta})\left(\frac{\zeta}{B(\zeta)} - \frac{z}{B(z)}\right)} d\zeta \quad (20)$$

$$G_L(z) = \frac{1}{2\pi i}\int_{C_u} \frac{(z-1)B(\frac{\zeta}{z})A\left(z\frac{B(\zeta)}{\zeta}\right)Q_0\left(\frac{\zeta}{z}\right)\left(\frac{1}{B(\zeta)} - \frac{\zeta B'(\zeta)}{B(\zeta)^2}\right)}{(1-\frac{1}{\zeta})\left(\frac{\zeta}{B(\zeta)} - \frac{z}{B(z)}\right)} d\zeta \quad (21)$$

and $D_L(z) = zB(z)Q_0(z)$, (22)

where the contour C_u is taken as a circle $\{|\zeta| = u\}$ so that $\zeta = z$ is inside but not $\zeta = 1$ that is $|z| < u < 1$.

The representation of the z-transforms above in terms of complex contour integrals is quite beneficial since it is possible to deform the contour C_u by picking up the pole $\zeta = z$ or including the pole $\zeta = 1$. The corresponding expressions for the z-transforms (14)-(22) may be found in [7].

IV. End-to-end delay and jitter evaluation for a stream traversing a series of queueing nodes.

The main objective in this paper is to develop analytical models that are possible to extend to also cover end-to-end analysis that goes beyond the traditional models based on convolutions. With convolution models the distortion (coloring) of a particular packet stream as it passes a multiplexer is neglected. A more accurate approach will be to consider a particular packet stream as it passes through a network and describe the change in the stream as it traverses the nodes where it will be disturbed by other (background) traffic.

By the slotted model described in this paper we may analyze in detail the output process for a particular packet stream when the input is a renewal process. In particular we have analyzed the time distribution between two successive departures. If we further approximate the output stream by a renewal stream (which is fully characterized by the distribution between two successive renewals), we may take this renewal stream as the input to the next node and thereby apply the queueing model recursively to get an end-to-end description. By this method we may "track" a given packet stream from source to destination as it crosses a multiple of nodes. In Fig. 2 we have depicted the key idea behind the end-to-end model. It is, however, well known that the output process of the queueing model described in chapter 3 will not be exactly renewal. Nevertheless, simulation studies [6] indicate that this type approximation indeed is very good if we only consider the marginal distribution of the output processes. Especially, the evolution of the jitter, but also the end-to-end delay will therefore be analyzed more accurately than by the convolutions approach.

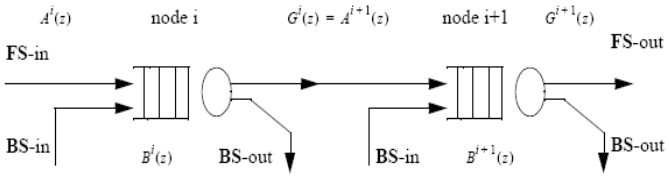

Fig. 2. The tandem queueing model for end-to-end modeling.

In the following we consider a chain of N queueing nodes and make the following assumptions:

- The streaming traffic (***FS***) enters node 1 according to a discrete time renewal process with distribution between

arrivals given by $a_1(m)$ and with generating function $A_1(z)$.

- The background traffic (***BS***) at node i enters (and leaves the node) according to a batch process with batch size distribution $b_i(k)$ and with generating function $B_i(z)$.
- The queueing discipline is FIFO for all the queues and the possible orderings when a streaming packet (***FS***) arrives in a slot with (possible several) background packets (***BS***) is either ***F***, ***R*** or ***L*** as described in chapter 3.

If we denote $d_i(m)$ distribution of the delay, with generating function $D_i(z)$, and denote distribution of the inter-departure time $g_i(m)$ with generating function $G_i(z)$ for the i 'th node, we may write the functional relations:

$$D_i(z) = F_I(A_i(z), B_i(z)),\ G_i(z) = F_{II}(A_i(z), B_i(z)) \qquad (23)$$

and $A_{i+1}(z) = G_i(z)$ (for $i = 1,2,...,N$)

where the functional entities F_I and F_{II} relate the delay distribution and inter-departure distribution as "functions" of the arrival processes. The actual form of F_I is given by the z-transform of the steady state queue length distribution $Q_0(z)$ (given by (5) and (4)) and the relations (16), (19) or (22), and further F_{II} is either the representations (15), (18) or (21) depending on the scheduling ***F***, ***R*** or ***L***. The jitter is found by inverting (14), (17) or (20).

Finally, if we denote $d^N(m)$ the distribution of the end-to-end delay for a chain of N successive nodes, then we obtain $d^N(m)$ by taking convolutions of the delay distributions at each node, i.e. we calculate $d^i(m) = \sum_{l=0}^{m} d^{i-1}(l) d_i(m-l)$ for $i = 2,...,N$. The corresponding z-transform is the product of the z-transforms in each node, i.e. $D^N(z) = \prod_{i=1}^{N} D_i(z)$.

Although the recursions given by (23) are analytical in nature, the corresponding procedure is highly numerical and contains some key assumptions: For each iteration both $d_i(m)$ (for $m = 0,1,..., N_{D_i}^{\max}$) and $g_i(m)$ (for $m = 0,1,..., N_{G_i}^{\max}$) are calculating numerically by inversion of the corresponding z-transforms, where we truncate the distributions when the probabilities are less than some quoted accuracy, and we use these (truncated) numerical transforms $D_i(z) = \sum_{m=1}^{N_{D_i}^{\max}} d_i(m) z^m$ and $G_i(z) = \sum_{m=1}^{N_{G_i}^{\max}} g_i(m) z^m$ as "input distributions" to the next iteration.

V. SOME NUMERICAL EXAMPLES

In the numerical examples below we have taken the following input streams:

- The ***FS*** is deterministic with mean time between arrivals equal $1/\rho_{FS}$.
- The ***BS*** is a Poisson stream with parameter ρ_{BS}.

For most of the examples we have taken the ***R***(random) queueing discipline where there is a random selection of all packets arriving in the same slot, but we have also given a few examples with the other two disciplines ***F***(first) and ***L***(last), more ore less to check out the numerical results and also see how sensitive the end-to-end delay and the evolution of the jitter are to the particular scheduling choice. We also assume that the load from the ***FS*** is relatively low, and we have taken the mean time between arrivals of the ***FS*** to be 10 slots i.e. $\rho_{FS} = 0.1$ for most of the examples, but we also have a few examples with mean time between arrivals of the ***FS*** to be 5 i.e. $\rho_{FS} = 0.2$.

The two main goals with the examples are

- Firstly, to compare this rather heavy numerical approach with the convolution approach,
- and secondly, to investigate the evolution of the jitter distribution as the ***FS*** traverses a chain of queues.

A. End-to-end delay

In Fig.3-Fig.6 we have depicted the CDF (Complementary Distribution Function) of the end-to-end delay for various parameter choices, and where we also have plotted the corresponding results obtained by the convolution approach.

In Fig.3 and Fig.4 we have compared the cases where we halve the mean time between arrivals of the ***FS*** (while not changing the total load). (In this example we have deterministic ***FS*** and ***R***(random) queueing discipline.) As expected the effect of increasing the load from a deterministic stream while keeping the total load constant will lead to output streams with less variance; and hence the queueing performance will improve.

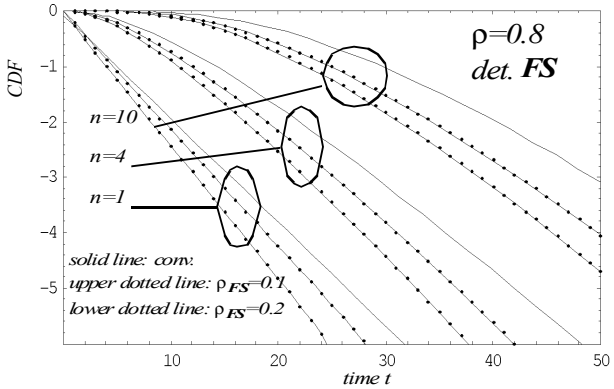

Fig. 3. Logarithmic plot of the CDF of end-to-end delay for ***R***(random)-queueing discipline and some different parameters as function of time (in slots).

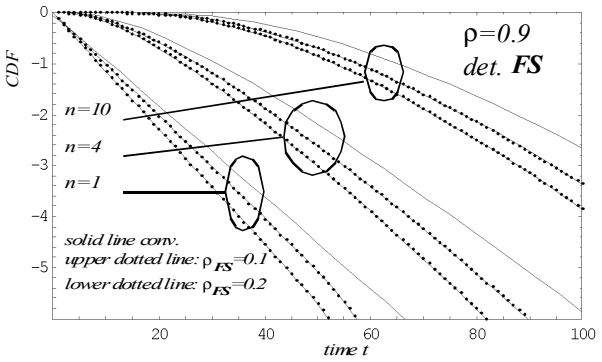

Fig. 4. Logarithmic plot of the CDF of end-to-end delay for ***R***(random)-queueing discipline and some different parameters as function of time (in slots).

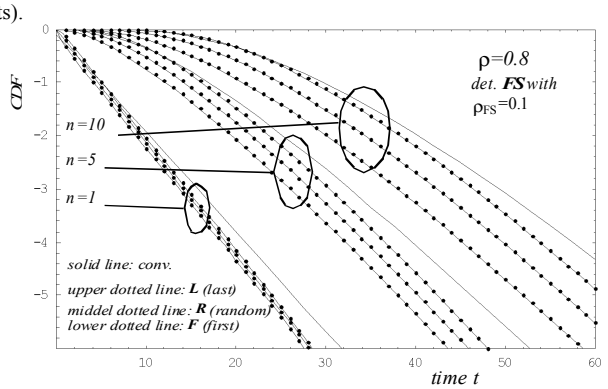

Fig. 5. Logarithmic plot of the CDF of end-to-end delay for different queueing discipline and some different parameters as function of time (in slots).

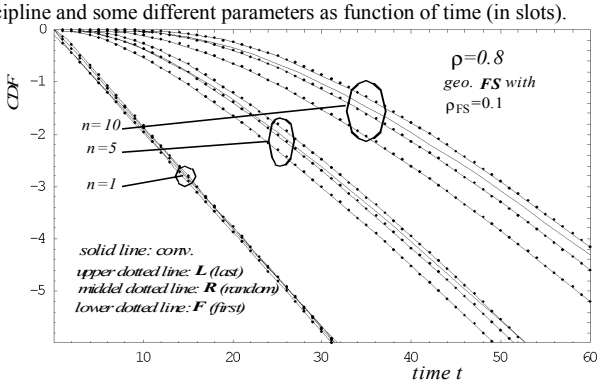

Fig. 6. Logarithmic plot of the CDF of end-to-end delay for different queueing discipline and some different parameters as function of time (in slots).

In Fig.5 and Fig.6 we have studied the effects of having different scheduling of packets (from the two input streams) arriving in the same slot. In Fig.5 we have deterministic ***FS*** while in Fig.6 geometrical ***FS*** is used. It seems that for deterministic ***FS*** all the three scheduling give end-to-end delay that are below that of the convolution. In Fig.6 we give one example where the slotted model gives worse performance than the convolution. This occurs when we have geometrical ***FS*** and ***L***(last) queueing principles (where the ***FS*** packet is placed behind all ***BS*** packets arriving in the same slot).

As a conclusion of the numerical examples for the end-to-end delay we have seen that the convolution approach for all, except for one particular case, will give an upper bound of the end-to-end delay compared with the slotted model considered in this chapter.

B. Jitter evolution

The jitter a packet stream is inflicted will be an important measure for the QoS in a communication network. For real time services the jitter will decide the dimension of the de-jitter buffer needed to obtain a regular bit stream at the receiver site. In Fig.7-Fig. 10 we have depicted a series of examples for the evolution of the jitter for the ***FS***. We have put main emphasis in the node-to-node evolution as the stream passes through a chain of nodes.

In Fig. 7 and Fig. 8 we have plotted the PDF of jitter where we look into the different scheduling strategies ***R***(random), ***F***(first) and ***L***(last) for a deterministic ***FS***. In these examples the load is set to 0.7 and the mean inter-arrival time for the ***FS*** is taken to be 10. As expected the scheduling ***F***(first) gives the most narrow jitter, and in between is the curves for ***R***(random) scheduling, while the ***L***(last) scheduling gives the broadest jitter. This is most evident at the first queues in the chain. As the number of passed queues increases the jitter get broader and the difference becomes less visible.

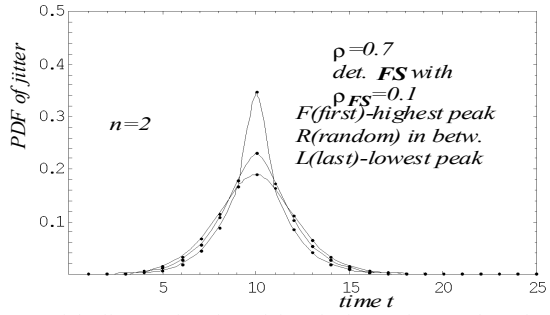

Fig. 7. PDF of the jitter as function of time for increasing number of nodes for the different queueing disciplines, ***F***(first), ***R***(random) and ***L***(last).

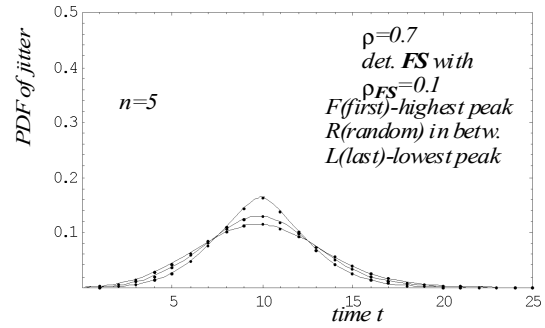

Fig. 8. PDF of the jitter as function of time for increasing number of nodes for the different queueing disciplines, ***F***(first), ***R***(random) and ***L***(last).

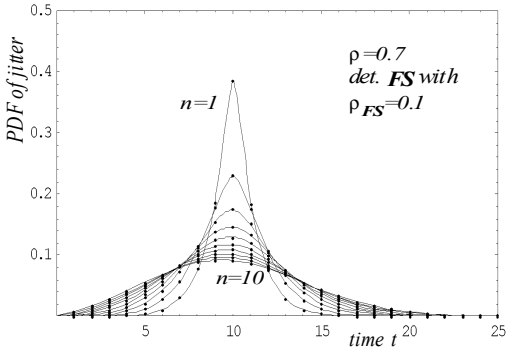

Fig. 9. PDF of the jitter as function of time for increasing number of nodes and deterministic distributed ***FS*** and ***R***(random) queueing discipline.

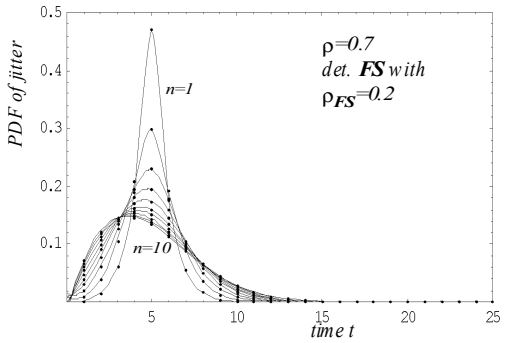

Fig. 10. PDF of the jitter as function of time for increasing number of nodes and deterministic distributed ***FS*** and ***R***(random) queueing discipline.

In Fig. 9 we have depicted all the PDFs of a deterministic ***FS*** with load 0.1, from node 1 up to the exit on node 10, using the ***R***(random) queueing discipline. Even though the jitter seems to be symmetrical for small number of queues, we observe that after passing the 10'th queue the jitter is not symmetrical any more. If we increase the load from the ***FS***, the mean inter-arrival time will decrease and the corresponding jitter will be more asymmetrical. This effect is clearly seen in Fig. 10 where the load from the ***FS*** is 0.2.

VI. Concluding Remarks

The methods proposed in this paper show that it is possible to obtain analytical results for quite complicated queueing models, even more important, to obtain numerical results from them. The aim has been to go beyond the assumption of product form solutions. The proposed model has the advantage that it is recursive, the output from a node constitutes the input to the next one, and in this way the end-to-end view is kept, and the changes of the packet stream from the input to the output are an important part of the analysis.

In addition to the end-to-end delay we have also analyzed the evolution of the jitter for a deterministic packet stream that passes through a series of queues. If all the nodes are identical we have also demonstrated that the jitter will converge to a given probability distribution.

Appendix

Lemma I: Let $r_j = r_j(x,s)$ be the distinct roots of the equation $h(z) = 1 - szA\left(x\frac{B(z)}{z}\right) = 0$ inside the unit disc where we assume that $A(z) = \sum_{i=1}^{K+1} a(i) z^i$ is a polynomial of degree $K+1$ (where $a(K+1) \neq 0$). If $|z| < 1$ then the product

$$\prod_{l=1}^{K}\left(\frac{z}{B(z)} - \frac{r_l}{B(r_l)}\right) = \left(\frac{z}{B(z)}\right)^K \left(1 - szA\left(x\frac{B(z)}{z}\right)\right)\exp[I(z,x,s)] \qquad (24)$$

where $I(z,x,s)$ the integral (4).

Proof: By changing the variable to $\zeta = z/B(z)$ the product $P = \prod_{l=1}^{k}(\zeta - \zeta_l)$ where ζ_l is the corresponding root of $g(\zeta) = 1 - s\zeta A(x/\zeta)B(z(\zeta)) = 0$ where $z = z(\zeta)$ is the inverse of $\zeta = z/B(z)$. We have the contour integral

$$\frac{1}{2\pi i}\int_\Gamma \log(\varsigma - \zeta)\frac{g'(\varsigma)}{g(\varsigma)}d\varsigma = \sum_{k=1}^{K}\log(\zeta - \zeta_k) - K\log\zeta$$

where Γ is a closed contour containing all the roots $\zeta_1,\ldots,\zeta_K$ of $g(\varsigma)$ inside the contour and also contains $\varsigma = 0$ (which is a pole of multiplicity K) and hence

$$\log P = \frac{1}{2\pi i}\int_\Gamma \log(\zeta - u)\frac{g'(\zeta)}{g(\zeta)}d\zeta + K\log u$$

By choosing $\Gamma = C \cup L_1 \cup L_2 \cup C_\varepsilon$ (see Fig. 11), and evaluate the contributions to the integral from the different parts of the contour and taking the limit $\varepsilon \to 0$ we find:

$$\log P = \log(\zeta^K) + \log g(\zeta) + \frac{1}{2\pi i}\int_C \frac{\log(g(\varsigma))}{\zeta - \varsigma}d\varsigma\,.$$

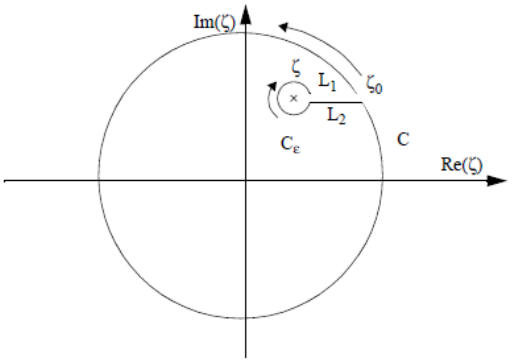

Fig. 11. The contour Γ

By changing the variable to $\zeta = z/B(z)$ we obtain (24).∎

References

[1] Østerbø O., Models for Calculating End-to-end Delay in Packet Networks; ITC-18, Berlin, Germany, August 31- September 5, 2003, pp 1231-1240.

[2] Østerbø O., End-to-End Delay Models with Priority; ITC-19, Beijing, China, August 29- September 2, 2005, pp 1049-1058.

[3] Hayes J. F., Modeling and Analysis of Computer Communications Networks 1984, Plenum Press, New York.

[4] De Vleeschauwer D.,Petit G.H., Steyaert B., Wittevrongel S., Bruneel H., Calculation of end-to-end delay quantile in network of M/G/1 queues, Electronics Letter, Vol. 37, 2001, pp 535-536.

[5] W. Matragi, K. Sohraby, C. Bisdikian, Jitter calculus in ATM Networks: Single node case, in Proc. IEEE INFOCOM'94, Toronto, Ont., Canada, June 12-16, 1994.

[6] W. Matragi, K. Sohraby, C. Bisdikian, Jitter calculus in ATM Networks: Multiple node case, in Proc. IEEE INFOCOM'94, Toronto, Ont.,Canada, June 12-16, 1994.

[7] Østerbø O. Mathematical Modelling and Analysis of Communication Networks: Transient Characteristics of Traffic Processes and Models foe End-to-end Delay and Delay-jitter; dr. philos thesis 2003, Department of Telematics, Faculty of Information Technology, Mathematics and Electrical Engineering, NTNU Trondheim Norwegian University of Science and Technology. (Theses may be downloaded from http://www.diva-portal.org/ntnu/abstract.xsql?dbid=6.)

[8] Feller W., An introduction to probability theory and its applications, Vol. II.