A Model of a Source-Retrial Open Exponential Queuing Network with Finite Shared Buffers in Multi-Queue Nodes

Miron Vinarskiy
Institute of Control Sciences of Russian Academy of Science
Moscow, Russia
Corresponding address: 3709 Mariana Way, Santa Barbara, California, 93105 USA
e-mail: mironvin@yahoo.com

Abstract— **We study a model of an open exponential queuing network where each node comprises several M/M/1 queues that share a common waiting space (a buffer) of limited capacity. A customer arriving to a node with a fully occupied buffer is blocked and re-injected by the source after a delay into the network. The process is repeated until the customer completes his service in the network and exits it. Input flow to each node is a superposition of the external Poisson flow, the flows coming from other nodes, and the retrials. The assumption made is that input flow to a node is a Poisson process. Under this assumption, two results are presented: an analytical evaluation of the network throughput and a method of an approximate analysis of the network model. The approach for both is based on iteratively solving a system of non-linear equations for unknown nodal flow rates. Existence and uniqueness of the solutions, obtained by the iterative algorithms, are rigorously proven in both cases. Required network and node performance characteristics are presented. The method provides low bound estimates for a moderately loaded (non-congested) network.**

Keywords- queuing network; multi-queue node; finite buffer; retrial; delay.

I. INTRODUCTION

Limited waiting spaces (finite buffers) in real-life nodes (service centers) lead to a so-called “blocking” when a customer cannot get into a fully occupied buffer. In many applications, such as computer communications, telephone systems, and distributed data processing a blocked customer tries to re-enter a network after some random time. The framework of retrial queues and networks seems to be an adequate approach for these applications. Most of the work on retrial models has been done on single queues (see, e.g., [1], [2], [3]). Retrial queuing network models have mostly concentrated on tandem queues. An exact analysis of these network models does not seem to be feasible for the general case, and therefore almost all known retrial tandem models use approximate approaches (see, e.g., [4], [5]).

The works by Irland et al. [4] and Avrachenkov et al. [5] give some details for tandem queuing systems with retrials of blocked customers. Irland et al. [4] considered a single isolated source-destination path in a packet-switching network as a tandem of single-queue nodes with a limited waiting space in each node. They compared two retrial techniques for a blocked customer (packet): local retrials (switch-retransmission) and source retrials (host-retransmission). The former retransmits a customer backup copy from the preceding switch, while the latter resends it from the network subscriber. Assuming Poisson flows in each node, they used a decomposition of a tandem queue network into simple M/M/1/N node models to approximate the unknown node input rates.

Avrachenkov et al. [5] considered a tandem network of two M/M/1/1 queues with blocking and with an M/M/1/∞ source-retrial (orbit) queue. The model formalized the interaction of data flow generated by a short TCP connection with a network of finite buffers. Authors explicitly solved the model and derived a stability condition. For more complex networks, it was suggested to use a fixed point approximation [6] with an assumption of a Poisson flow in each queue. It was shown in [7] that a fixed point approximation for a retrial queue with a Poisson assumption works well only when the nominal load is small. This fact was confirmed in [8] for a tandem network with an arbitrary number of M/G/K/K queues.

Lam [9] studied a model of a packet-switching network with local retrials and multi-queue nodes. A blocked customer (packet) is unlimitedly retransmitted from an adjacent node until the nodal buffer becomes open. Under the Poisson flow assumption, a system of non-linear equations was built for the unknown nodal blocking probabilities, and solved iteratively. No proof of iterations convergence was presented.

The network model under study in this paper is an extension of the single-class queuing network model with losses and multi-queue nodes [10] to the case of source-retrials. The model description and solution methodology have a lot in common with the model in [10], but we focus specifically on the source-retrial. Adding retrials to the network model with losses makes flow balance equations more complex. In turn, it requires different approaches to

 ISBN: 978-1-61208-655-2

prove the solution. The goal of the paper is to show that the model can be solved analytically by an approximate numerical method.

Blocked customers are dispatched back to the network after a random delay in the M/M/∞ retrial queue with infinite exponential servers. Thus, the model uses the classic retrial policy: each blocked customer generates a stream of repeated requests independently of the rest of the customers in the retrial group.

The model can be used for performance evaluation of distributed data processing systems with nodes implemented as shared-memory architecture multiprocessor service centers, telecommunication systems and computer communication networks with source-retransmission of undelivered packets. In a distributed data processing system a customer (data request) can travel between nodes in order to get access to a distributed database. Upon completing its service by a node processor, a customer can leave the system from the node, or continue service at either the next node, or at the same node by a different processor.

Our network model is based on multi-queue nodes with a finite common buffer in each node. Buffer sharing policy is Complete Sharing (CS), where no restrictions on buffer occupancy are imposed for any queue. Output queuing structures in shared-memory switches/routers are good examples of such nodes [11]. In this application, a packet memory pool is shared among output ports.

Retrial queues are very complex objects. Even for a single retrial M/M/C queue, a closed form solution is only available for the number of servers $C \leq 2$. An approximate analysis for C>>2 is performed by replacing a retrial queue with a loss queue, under a Poisson input. The latter represents the mixture of a primary Poisson flow and retrials. This approximation works really well for not overloaded queue [1].

To make our network model analytically tractable, we also use a Poisson process to represent a node input. The input flow is a superposition of an external Poisson stream, a traffic coming from other nodes, and a retrial flow. Under this assumption, two results are presented: - 1) an analytical evaluation of the network throughput, which determines a permissible network load; - 2) a method of an approximate analysis of the network model. In both cases, the result is achieved by decomposing the network into separate simple nodal models and combining the nodal results in a system of non-linear equations for the unknown nodal flow rates. It is shown that the systems can be solved iteratively, and a proof is provided that the iterations converge to a unique solution. The solution for the nodal flow rates in the network model is used to receive several all-network and node performance measures.

The approach provides reasonable low bound estimates for a moderately loaded (non-congested) network. We use the term “moderately loaded” to approximately define a network mode, where an internal traffic, including retrials, loads any server in a node under 80% of capacity, and node blocking probabilities lower than 0.05.

The remainder of this paper is organized as follows. In Section II, we provide a formal description of the network model, including notation and the node product-form state distribution. In Section III, we present equations and a computational procedure for the network throughput. In Section IV, we concentrate on the network flow balance equations. Direct substitution iterations are used to solve the equations. In Section V, we define the required network performance measures. In Section VI, we present some numerical results computed by our analytic method in comparison with simulation results.

II. NETWORK MODEL

The network model under consideration here is a modification of the single-class queuing network model with losses [10]. Some model notation and description from [10] is included in this paper to provide a clear foundation for the model’s expansion.

Let us consider an open queuing network with W nodes. The retrial (orbit) queue Figure 1 formalizes the random delay associated with a retrial of a blocked customer. The queue has infinitely many exponential servers (M/M/∞) with service rate μ_0.

The node-i (i = 1, 2,.., W) comprises $Q_i > 1$ of M/M/1 queues sharing finite common buffer of size N_i units Figure 1. The buffer contains all Q_i queues, including customers in service. The queuing system q ($q = 1,2,..,Q_i$) is characterized by an exponentially distributed service time with mean μ_q^{-1}, and queuing discipline FCFS (first come first served).

A customer arriving at node-i when its buffer is fully occupied is blocked, and transferred to the orbit queue that dispatches the customer back to the network after some random time. Retrials are distributed between nodes with probabilities γ_{0i}, i=1,2,.., W, $\sum_{i=1}^{W} \gamma_{0i} = 1$.

If there are free spots in the buffer of node-i, then an arriving customer joins the q-th queue system with probability α_{iq}, q =1,2.., Q_i, $\sum_{q=1}^{Q_i} \alpha_{iq} = 1$. Customers initially arrive to the network from an external source, which generates a Poisson flow with rate λ_0. This flow is distributed between nodes according to probabilities p_{0i}, i = 1, 2,.., W, $\sum_{i=1}^{W} p_{0i} = 1$. A customer, that has completed his service in node-i, is either transferred to node-j with routing probability p_{ij} , i,j = 1, 2,.., W , $\sum_{j=1}^{W} p_{ij} < 1$, or completes his service in the network and leaves with probability $p_{iE} = (1 - \sum_{j=1}^{W} p_{ij}) > 0$.

Figure 1 shows traffic in a node in the network. Node-i receives an "original" Poisson flow from an external source with the rate $\lambda_0 p_{0i}$. A secondary flow (dashed lines) is produced by other nodes in the network and possibly by node-i itself, as well as by the source-retrials. Superposition of the original and secondary flows forms the node-i input flow with rate λ_i. A part of this flow with the rate $\lambda_{i0}^{(R)}$ is blocked, initiating the source-retrial. The rest goes through node-i and then splits into a secondary flow with probability $1 - p_{iE} = \sum_{j=1}^{W} p_{ij}$ and traffic with flow rate λ_{iE} exiting the network after node-i.

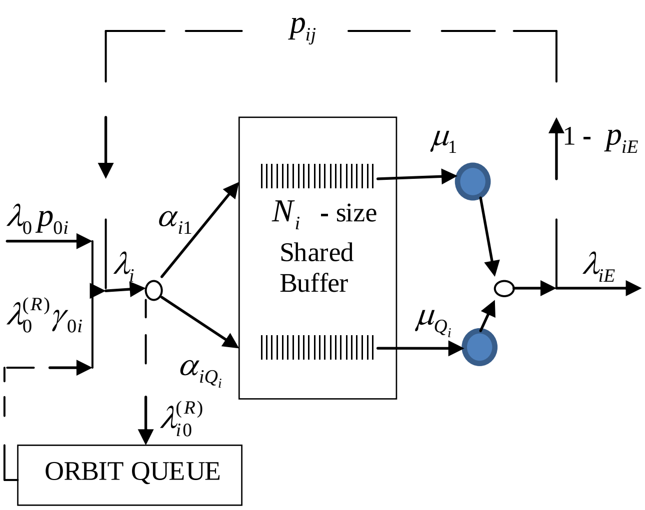

Figure1. Flows in the node-i of the network.

To determine traffic rates λ_i (i = 1, 2, .., W), we assume that the superposition of the original external Poisson flow and all secondary flows to the node-i is a Poisson process. This approach provides acceptable low bounds for moderately loaded networks (see Section VI). We have observed that the method works really well for networks where each node is connected to at least two nodes and traffic after a node splits according to a Markovian routing, merging with other flows as input arrivals.

It should be noted that the buffer overflow is a bursty stream. It can be efficiently approximated by Interrupted Poisson Process (IPP) [12] with a squared coefficient of variation $c^2 = (\mathrm{var}/mean^2) > 1$. However, even an individual queue with limited waiting space and IPP input does not have a closed form solution. Analysis is performed numerically. For a multi-queue node, even this approach fails in general because the number of the Markov equations grows exponentially with the number of queues.

A. Notation

The multi-queue node-i state is given by Q_i - dimensional vector $\mathrm{n}_i = (n_{i1}, n_{i2}, .., n_{iQ_i})$, where n_{iq} ($0 \le n_{iq} \le N_i$, $q = 1, 2, .., Q_i$) denotes the number of customers in the q -th queue system, including the customer in service. It is convenient to introduce the following notation:

$n_{i,k}^{m} = (n_{i1}, .., n_{i,k-1}, \mathbf{m}, n_{i,k+1}, .., n_{iQ_i})$,

$n_i = \sum_{q=1}^{Q_i} n_{iq}$ = total number of customers in node-i,

$\overline{n}_i(N_i)$ = mean number of customers in node-i.

Let

D_i = the set of permissible states that determined by CS buffer sharing policy in the node-i

$$D_i = \{\mathbf{n}_i : \sum_{q=1}^{Q_i} n_{iq} \le N_i,\ 0 \le n_{iq} \le N_i\}.$$

For queue- q we have

$\rho_{iq} = \lambda_i\ \alpha_{iq} / \mu_q$ = offered traffic intensity for queue- q in node-i.

The following is for the network,

$\mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_W)$` = column-vector of input flow rates in all $W$ nodes, (` denotes transposition).

$\| \mathbf{\Lambda} \| = \sum_{i=1}^{W} |\lambda_i|$ = norm of the vector $\mathbf{\Lambda}$.

B. The node product-form state distribution

The node-i model in Figure 1 can be considered as an open exponential queuing sub network under Poisson arrivals with input rate

$$\lambda_i(N_i) = \begin{cases} 0 & \text{– if buffer is fully occupied} \\ \lambda_i & \text{- otherwise.} \end{cases}$$

The equilibrium state probability distribution for this type of queuing network is given by product form [13], [14]

$$\boldsymbol{P}(\mathbf{n}_i) = G(N_i)^{-1} \prod_{q=1}^{Q_i} \rho_{iq}^{n_{iq}} , \qquad (1)$$

where

$$G(N_i) = \sum_{\mathbf{n}_i \in D_i} \prod_{q=1}^{Q_i} \rho_{iq}^{n_{iq}} \qquad (2)$$

is the normalization constant. From (1, 2) the stationary probability that the node-i is available for a customer is

$$\pi_i(\lambda_i) = G(N_i - 1)/G(N_i) \qquad (3)$$

The output flow rate from the node-i is

$$\lambda_i^{out} = \lambda_i \pi_i(\lambda_i) \qquad (4)$$

In the following statement, imported from [10, Proposition 3.2], the index i is dropped to simplify the notation.

Proposition 1. The node output $\lambda\pi(\lambda)$ is an increasing function of $\lambda > 0$, and $\lim_{\lambda\to\infty} \lambda\pi(\lambda) = \hat{G}(N-1) / \hat{G}(N)$, where $\hat{G}(N)$ is the normalization constant of the closed queuing network that is a node model under a constantly full buffer.

III. NETWORK THROUGHPUT

Let us consider the network output rate

$$O(\mathbf{\Lambda}) = \sum_{i=1}^{W} \lambda_i \pi_i(\lambda_i)\ (1 - \sum_{j=1}^{W} p_{ij}) . \qquad (5)$$

In the network's stationary mode, $O(\mathbf{\Lambda})$ is equal to the source flow rate λ_0. To determine the network's permissible load λ_0, let us find the network throughput $O_{\max}$, which has to satisfy the following inequality

$$O_{\max} \le \lim_{\mathbf{\Lambda}\to\infty} O(\mathbf{\Lambda}) = \sum_{i=1}^{W} a_i (1 - \sum_{j=1}^{W} p_{ij}), \qquad (6)$$

where $a_i = \lim_{\lambda_i\to\infty} \lambda_i \pi_i(\lambda_i) = \hat{G}(N_i - 1)/\hat{G}(N_i)$ (see Proposition 1).

To calculate $O_{\max}$ we assume the infinite network load $\lambda_0 = \infty$. Under this assumption, a group of nodes will have constantly full buffers. Among them will be the nodes that receive initial arrivals from an external source according to positive probabilities $p_{0i} > 0$. Also, the group will have nodes that receive retrials that are generated by all nodes, including those with always full buffers.

Let us assume that there will be v ($0 \le v < W$) nodes with not always full buffers and (W - v) nodes with constantly full buffers. Let $I_1 = \{1,2,..,v\}$ and $I_2 = \{v+1, v+2,..., W\}$. Then $O_{\max}$ can be expressed from (5 and 6) as

$$O_{\max} = [\sum_{i\in I_2} a_i + \sum_{i\in I_1} \lambda_i \pi_i(\lambda_i)](1 - \sum_{j=1}^{W} p_{ij}) . \qquad (7)$$

Unknown flow rates λ_i, $i \in I_1$, are solutions of the following system of non-linear flow balance equations

$$\lambda_i = \sum_{j\in I_2} a_j p_{ji} + \sum_{j\in I_1} \lambda_j \pi_j(\lambda_j) p_{ji} \qquad i \in I_1 . \qquad (8)$$

The structure of (8) is very similar to the flow balance equations in the single-class network model with losses [10, expression (2.6) for $R = 1$]. Thus, a positive unique solution of (8) $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*,.., \lambda_v^*)$ can be found by direct substitution iterations as in [10, expression (3.4)]. We omit here the proof of the iterations convergence. An interested reader is referred to [10, Theorem 3.1].

With vector $\boldsymbol{\lambda}^*$ the network throughput $O_{\max}$ is fully determined by (7), that in turn defines the network permissible load $\lambda_0 < O_{\max}$.

IV. NETWORK FLOW BALANCE EQUATIONS

The following system of non-linear equations establishes flow balance for nodes in the network

$$\lambda_j = \lambda_0 p_{0j} + \sum_{i=1}^{W} \lambda_i \pi_i(\lambda_i) p_{ij} + \gamma_{0j} \sum_{i=1}^{W} \lambda_i (1 - \pi_i(\lambda_i)),$$

$$j = 1, 2, .., W . \qquad (9)$$

The flow rate into the orbit queue is determined by λ_j (j =1, 2, .., W) as

$$\lambda_0^{(R)} = \sum_{j=1}^{W} \lambda_{j0}^{(R)} = \sum_{j=1}^{W} \lambda_j (1 - \pi_j(\lambda_j)) . \qquad (10)$$

It is convenient rewrite (9) in vector form

$$\mathbf{\Lambda} = \mathbf{\Psi}(\mathbf{\Lambda}), \qquad (11)$$

where $\mathbf{\Psi}(\mathbf{\Lambda}) = (\ \Psi_1(\mathbf{\Lambda}),...,\ \Psi_W(\mathbf{\Lambda})\ )`$,

$$\Psi_j(\mathbf{\Lambda}) = \lambda_0 p_{0j} + \sum_{i=1}^{W} \lambda_i \pi_i(\lambda_i) p_{ij} +$$

$$\gamma_{0j} \sum_{i=1}^{W} \lambda_i (1 - \pi_i(\lambda_i)), \qquad j = 1, 2,.., W . \quad (12)$$

Operator $\mathbf{\Psi}(\mathbf{\Lambda})$ is defined in $\Omega = \{ \mathbf{\Lambda} : \lambda_i \ge 0,$ $i = 1,2,.., W \}$ and maps $\Omega \to \Omega$.

Proposition 2. Operator $\mathbf{\Psi}(\mathbf{\Lambda})$ is an increasing operator.

Proof. Proposition immediately follows from [10, expression (D.1) in Appendix D] for node-k

$$\frac{\partial(\lambda_k \pi_k(\lambda_k))}{\partial \lambda_k} = \pi_k(\lambda_k)\ [\bar{n}_k(N_k-1)-\bar{n}_k(N_k)+1] > 0$$

and

$$\frac{\partial \Psi_i(\mathbf{\Lambda})}{\partial \lambda_k} = p_{ki}\pi_k(\lambda_k)[\bar{n}_k(N_k-1)-\bar{n}_k(N_k)+1] +$$

$$\gamma_{0i}\ \{1 - \pi_k(\lambda_k)[\bar{n}_k(N_k-1)-\bar{n}_k(N_k)+1]\} =$$

$$p_{ki}\omega_k + \gamma_{0i}(1-\omega_k) > 0, \qquad (13)$$

where $i,k = 1,2,..,W$, and

$$0 < \omega_k = \pi_k(\lambda_k)[\bar{n}_k(N_k-1)-\bar{n}_k(N_k)+1] < 1. \quad (14)$$

The system (11) can be solved iteratively by using the following relation

$$\mathbf{\Lambda}^{(m+1)} = \mathbf{\Psi}(\mathbf{\Lambda}^{(m)}) \qquad m = 0, 1, 2, \ldots, \qquad (15)$$

where vector $\mathbf{\Lambda}^{(m)} = (\lambda_1^{(m)},...,\lambda_W^{(m)})$\`is a result of the $m$-th iteration, and $\mathbf{\Lambda}^{(0)} = (0,0,..,0)$\`. Also, $\mathbf{\Lambda}^{(m+1)} \geq \mathbf{\Lambda}^{(m)}$ if $\lambda_i^{(m+1)} \geq \lambda_i^{(m)}$, and $\mathbf{\Lambda}^{(m)} > 0$ if $\lambda_i^{(m)} > 0$ for $i = 1, 2, .., W$.

Theorem. For network load $\lambda_0 < O_{\max}$, the sequence$\{\mathbf{\Lambda}^{(m)}, m \geq 0\}$,defined by (15),converges to $\mathbf{\Lambda}^*$, a positive unique solution of system (11).

Proof.

Existence of $\mathbf{\Lambda}^*$. Vector $\mathbf{\Lambda}^{(1)}$ has a positive component for node-i if $p_{0i} > 0$. Vector $\mathbf{\Lambda}^{(2)}$ can have more positive components if there are positive probabilities of transferring a customer from node-i to other nodes. From Proposition 2 and $\mathbf{\Lambda}^{(1)} = \mathbf{\Psi}(\mathbf{\Lambda}^{(0)}) \geq \mathbf{\Lambda}^{(0)}$ follows that the sequence $\{\mathbf{\Lambda}^{(m)}$, m=0, 1, 2,...} is a non-decreasing sequence.

Let us show that the sequence (15) is limited in Ω. From Proposition 1 follows that $O(\mathbf{\Lambda})$ (5) is an increasing function of $\mathbf{\Lambda}$, and consequently for $\lambda_0 < O_{\max}$ there is $\mathbf{\Lambda}^* \in \Omega$ that for any $\mathbf{\Lambda} > \mathbf{\Lambda}^*$ ($\mathbf{\Lambda} \in \Omega$)

$$O(\mathbf{\Lambda}) > \lambda_0. \qquad (16)$$

By summing (12) over $j = 1,2,\ldots, W$ we can get

$$\|\mathbf{\Psi}(\mathbf{\Lambda})\| = \lambda_0 - \sum_{i=1}^{W}\lambda_i\pi_i(\lambda_i)\ (1-\sum_{j=1}^{W}p_{ij}) + \|\ \mathbf{\Lambda}\ \|. \quad (17)$$

Applying (5) and (16) to (17) we have

$$\|\ \mathbf{\Lambda}\ \| > \|\mathbf{\Psi}(\mathbf{\Lambda})\| \text{ for } \mathbf{\Lambda} > \mathbf{\Lambda}^*\ (\mathbf{\Lambda} \in \Omega). \qquad (18)$$

Let us assume that the sequence $\{\mathbf{\Lambda}^{(m)}$, m=0, 1, 2,...} is not limited in Ω. Then there will be a number m, such that $\mathbf{\Lambda}^{(m)} > \mathbf{\Lambda}^*$, and according to (18) $\|\mathbf{\Lambda}^{(m)}\| > \|\mathbf{\Psi}(\mathbf{\Lambda}^{(m)})\|$.

Consequently, $\|\ \mathbf{\Lambda}^{(m+1)}\| = \|\mathbf{\Psi}(\mathbf{\Lambda}^{(m)})\| < \|\mathbf{\Lambda}^{(m)}\|$,

that contradicts the fact that the sequence $\mathbf{\Lambda}^{(m)}$ is a non-decreasing sequence. Thus, a positive vector $\mathbf{\Lambda}^* = \lim_{m\to\infty}\mathbf{\Lambda}^{(m)} < \infty$ is a solution of (11).

Uniqueness of the solution $\mathbf{\Lambda}^*$. Let us assume that there are two different solutions $\mathbf{\Lambda}^* > 0$ and $\mathbf{\Lambda}^{**} > 0$. Then, for the convex domain Ω, we have

$$\|\mathbf{\Lambda}^* - \mathbf{\Lambda}^{**}\| =$$

$$\|\mathbf{\Psi}(\mathbf{\Lambda}^*) - \mathbf{\Psi}(\mathbf{\Lambda}^{**})\| \leq \|\mathbf{\Psi}'(\mathbf{\Lambda})\|\ \|\mathbf{\Lambda}^* - \mathbf{\Lambda}^{**}\|, \quad (19)$$

where $\mathbf{\Lambda}^*, \mathbf{\Lambda}^{**} \in \Omega$, $\mathbf{\Lambda} = \mathbf{\Lambda}^* + \xi(\mathbf{\Lambda}^{**} - \mathbf{\Lambda}^*)$, $0 < \xi < 1$.

From (13) we have $\|\mathbf{\Psi}'(\mathbf{\Lambda})\| = \max_k \sum_{i=1}^{W}\left|\frac{\partial \Psi_i(\mathbf{\Lambda})}{\partial \lambda_k}\right| =$

$\max_k[1-\omega_k(1-\sum_{i=1}^{W}p_{ki})]$, where $0 < \omega_k < 1$ (see 14). From $\sum_{i=1}^{W}p_{ki} < 1$ for $k = 1,2,...,W$ follows that $\|\mathbf{\Psi}'(\mathbf{\Lambda})\| < 1$, and the inequality (19) can be valid only if $\mathbf{\Lambda}^* = \mathbf{\Lambda}^{**}$. Q.E.D. Computational complexity of (15) is $\sim O(W^2)$.

V. NETWORK PERFORMANCE MEASURES

A. Nodal measures

To simplify notation we drop index i for an arbitrary node in the network. Let us consider the following aggregate state for a node

$A(u) = \{\mathbf{n} \in D : \sum_{q=1}^{Q} n_q = u\}$, which comprises all states with the total population of u customers in the node. With this state we associate two functions:

$$g(u) = \sum_{\mathbf{n} \in A(u)} \prod_{q=1}^{Q} \rho_q^{n_q} \quad \text{and}$$

$$g^{-k}(u) = \sum_{\mathbf{n}_k^0 \in A(u)} \prod_{q=1}^{Q} \rho_q^{n_q}.$$

 ISBN: 978-1-61208-655-2

The normalization constant

$$G(N) = \sum_{u=0}^{N} g(u).$$

Distribution of the total number of customers in the node is given by

$$P_u = P_0 \cdot g(u) = g(u) / \sum_{z=0}^{N} g(z) \quad \text{for } 0 \le u \le N.$$

The mean number of customers in the node is

$$\bar{n}(N) = \sum_{u=0}^{N} u P_u .$$

Using Little's Law, the average time a customer spends in the node is

$$\bar{t} = \bar{n}(N) / \lambda\pi(\lambda) \qquad (20)$$

The marginal distribution of the number of customers in the queue-q is

$$P_{n_q}^{(q)} = \rho_q^{n_q} \sum_{u=0}^{N-n_q} g^{-q}(u) / \sum_{z=0}^{N} g(z).$$

The mean number of the queue-q customers is

$$\bar{n}_q = \sum_{n_q=1}^{N} n_q P_{n_q}^{(q)} .$$

The average delay (queue plus server) for a queue-q customer is given by Little's Law

$$\bar{d}_q = \bar{n}_q / \lambda\pi(\lambda)\alpha_q .$$

B. *All-network service measures*

On the average, λ_0 customers arrive at the network during a unit interval. Therefore, in stationary mode, when $\lambda_0 < O_{\max}$, the network output rate $O(\mathbf{\Lambda}) = \lambda_0$.

During this time interval $\sum_{k=1}^{W} \lambda_k \pi_k(\lambda_k)$ customers, on average, go through the service nodes. Thus, the average number of services received by a customer in the network is

$$\bar{s} = \sum_{k=1}^{W} \lambda_k \pi_k(\lambda_k) / \lambda_0 .$$

The average sojourn time for a customer in the network, including a retrial delay, is

$$\bar{T} = [\lambda_0^{(R)}(1/\mu_0) + \sum_{k=1}^{W} \lambda_k \pi_k(\lambda_k)\bar{t}_k] / \lambda_0 , \qquad (21)$$

where $\bar{t}_k$ is defined in (20) and $\lambda_0^{(R)}$ in (10).

VI. COMPARISON OF ANALYTIC AND SIMULATION RESULTS

In this section, we present some numeric results computed by our analytic method in comparison with simulation. The simulation code is written in C^{++}, and simulates a network processing of $\sim 10^6$ customers in one run. We experiment with two network topologies: a symmetric complete-graph network and a ring-type network.

The symmetric configuration includes five nodes; each has two identical single servers (two-queue node) and a buffer of size $N = 10$. Traffic arriving to node-i (i =1, 2, .., 5) splits equally between two queues, i.e., α_{iq} = 0.5, q = 1, 2. The orbit queue service rate is μ_0 = 0.1, while all other servers in the network have the same service rate μ =1. The external input flow with rate λ_0 is uniformly distributed between the nodes, i.e., p_{0i} = 0.2 (i =1, 2, .., 5). A customer that has completed his service in the node-i is either transferred to the node-j (j =1, 2, .., 5) with probability 0.1 or leaves the network with probability 0.5. Retrials are distributed into the network with probability γ_{0i} = 0.2 (i =1, 2, .., 5).

TABLE I presents results for λ_0 =2.0, 3.0, 3.6, and 4.0. Columns 3-5 have data for one separate node; column 2 presents the average sojourn time in the network, including retrials. The upper figure in each box has been received by the analytic method. The lower was obtained by simulation.

TABLE I. NETWORK CHARACTERISTICS OF THE SYMMETRIC 5-NODE NETWORK. ANALYTIC RESULTS VERSUS SIMULATION.

	(21)	(9)	(4)	(20)
λ_0	$\bar{T}$	λ_i	λ_i^{out}	$\bar{t}_i$
1	2	3	4	5
2.0	3.335	0.8	0.799	1.663
	3.34	0.8	0.799	1.666
3.0	5.025	1.212	1.198	2.393
	5.226	1.214	1.199	2.5
3.6	7.142	1.512	1.437	3.06
	8.3	1.516	1.439	3.62
4.0	9.914	1.805	1.599	3.67
	13.487	1.842	1.612	5.24

External arrival rates in the range λ_0 = 2.0 – 3.6 moderately load the network. We can observe that for these loads the node input rate is λ_i <1.6, and consequently ρ_{iq} < 0.8. Average sojourn times in network, calculated analytically, are lower than in simulation by 0.15% - 13.9%.

 ISBN: 978-1-61208-655-2

We can conclude that a Poisson assumption for a node input gives reasonable low bounds for this load range. Further increase of source arrival rate brings the network close to congestion, dramatically increasing the difference between analytical result and simulation one.

Another example is a 5-node ring-type network, where all five nodes are identical two-queue nodes, described above. The input flow with rate λ_0 is uniformly distributed between the nodes, i.e., p_{0i} = 0.2 (i =1, 2, .., 5). After completing his service in node-i, customer is either transferred to node-$(i+1)$ with probability 0.2, or to node -$(i-1)$ with probability 0.2, or exits the network with probability 0.6. Retrials are distributed between nodes with probability γ_{0i} = 0.2 (i =1, 2, .., 5). For node-1 the "left" neighbor is node-5. For node-5 the "right" neighbor is node-1. We assume p_{ii} = 0, i.e., a node may not route traffic to itself.

The computational results for λ_0 = 3.0, 3.5, 4.0, and 4.5 are shown in TABLE II, which has similar structure as TABLE I. For moderate network load λ_0 = 3.0 – 4.0, the node input rate λ_i < 1.6 and ρ_{iq} < 0.8. Comparison of the analytic results (upper figures in each box) with simulation ones (lower figures) shows that in this load range the analytic method provides acceptable low bound estimates. For instance, the error of calculating the average sojourn times in the network is in the 1.18% - 9.2% range. The network becomes congested under λ_0 = 4.5 and the error is increased to 20.7%. This example demonstrates that our Poissonian hypothesis works even for a weakly connected not congested network.

TABLE II. NETWORK MEASURES OBTAINED ANALYTICALLY AND BY SIMULATION FOR THE RING-TYPE 5-NODE NETWORK.

	(21)	(9)	(4)	(20)
λ_0	$\bar{T}$	λ_i	λ_i^{out}	$\bar{t}_i$
1	2	3	4	5
3.0	3.34	1.003	0.999	1.978
	3.38	1.004	1.001	1.998
3.5	4.019	1.178	1.166	2.3
	4.2	1.179	1.168	2.43
4.0	5.02	1.369	1.332	2.739
	5.531	1.372	1.334	3.03
4.5	6.643	1.607	1.499	3.272
	8.38	1.631	1.52	4.16

VII. CONCLUSION

We have extended the model of an open exponential single-class queuing network with losses due to limited shared waiting spaces in multi-queue M/M/1 nodes [10] to the case of the source-retrials, experienced by blocked customers. The goal of the paper is to show that the model can be solved approximately by an analytical numerical approach. Using the methodology outlined in [10], we have established an approximate numerical method that makes it possible to solve the model analytically. An analytical procedure to evaluate the network throughput that determines a permissible network load was received as well.

The main result of the paper is a method of an approximate analysis of the network model under a moderate load. The core of the approach is solving iteratively a system of non-linear equations for the unknown nodal flow rates. We have rigorously proven that the iterative algorithm converges to a unique solution, which is used to obtain several network and node performance measures.

The model can be used for performance evaluation of computer communication networks with adaptive or alternative routing and source–retransmission of undelivered packets. Also, the paper results can help to analyze different structures of distributed database systems with multiprocessor nodes.

Future work can consider a source-retrial multi-class queuing network with finite shared buffer in multi-queue nodes. The use of Interrupt Poisson Process as a node input might help to conduct an approximate analysis of an even congested network.

ACKNOWLEDGMENT

The author would like to thank anonymous referees for a number of useful comments and suggestions given to the original manuscript.

REFERENCES

[1] G. I. Falin, and J.G.C. Templeton, Retrial queues. Boca Raton, FL: CRC Press, 1997.

[2] J. R. Artalejo, and A. Gomez-Coral, Retrial queueing systems. A computational approach. Berlin: Springer Berlin Heidelberg, 2008.

[3] M. Jain, A. Bhagat, and C. Sherkhar, "Double orbit finite retrial queues with priority customers and service interruptions," Applied Mathematics and Computation, Vol. 253, pp. 324-344, 2015.

[4] M. Irland and G. Pujolle, "Comparison of two packet-retransmission techniques," IEEE Transactions on Information Theory, Vol. IT-26, No. 1, pp. 92-97, 1980.

[5] K. Avrachenkov and U. Yechiali, "Retrial networks with finite buffers and their application to Internet data traffic," Probability in the Engineering and Informational Sciences, Vol. 22, pp. 519-536, 2008.

[6] G. K. Takahara, "Fixed point approximation for retrial networks," Probability in the Engineering and Informational Science, Vol. 10, Issue 2, pp. 243-259, 1996.

 ISBN: 978-1-61208-655-2

[7] J. W. Cohen, “Basic problems of telephone traffic theory and the influence of repeated calls,” Philips Telecommunication Review, Vol. 18, pp. 49-100, 1957.

[8] K. Avrachenkov and U. Yechiali, “On tandem blocking queues with a common retrial queue,” Computer & Operations Research, Vol. 37, pp. 1174-1180, 2010.

[9] S. Lam, “Store-and-forward buffer requirements in a packet switching network,” IEEE Transactions on Communications, Vol. 24, No. 4, pp. 394-403, 1976.

[10] M. Vinarskiy, “A method of approximate analysis of an open exponential queuing network with losses due to finite shared buffers in multi-queue nodes,” European Journal of Operational Research, Vol. 258, Issue 1, pp. 207-215, 2017.

[11] W. Aello, A. Kesselman, and Y. Masour, “Competitive buffer management for shared-memory switches,” ACM Transactions on Algorithms, Vol. 5, No. 1, pp. 3.1-3.6, 2008.

[12] K. S. Meier-Hellstern, “The analysis of a queue arising in overflow models,” IEEE Transactions on Communications, Vol. 37, No. 4, pp. 367-372, 1989.

[13] S. Lam, “Queueing networks with population size constrains,” IBM Journal of Research and Development, Vol. 21, No. 4, pp. 370-378, 1977.

[14] P. G. Harrison, “Reversed processes, product form and some non-product forms,” Linear Algebra and Its Applications, Vol. 386, pp. 359-381, 2004.