Modelling issues of integrating services in Next Generation Networks

S.N.Stepanov[1], V.B.Iversen[2]

[1]Institute for Problems of Information Transmission
Russian Academy of Sciences

[2]COM Center, Technical University of Denmark

Paper outline

1. Motivation
2. Model description and performance measures
3. Exact evaluation of performance measures
4. Approximate estimation of performance measures
5. Numerical results
6. Conclusion

Motivation

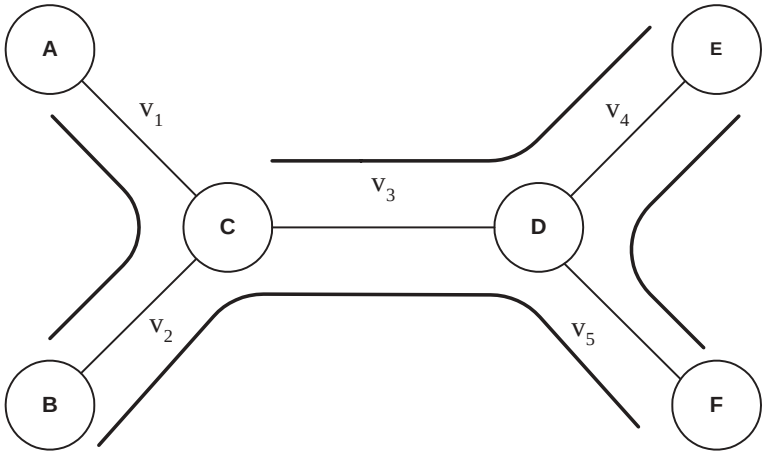
- Traffic sources under MPLS technology can be separated into two groups:
  - — the calls of the first group can not wait at points of traffic concentration (real-time traffic of interactive communications like Internet telephony, packetized video)
  - — the calls of the second group can tolerate some waiting (Best Effort traffic such as transfer of stored information)
- The term **QoS** traffic will be used for the first type of load and **Best Effort** traffic for the second reflecting properties of corresponding traffic
- The efficiency of link usage can be increased if transmission of queueable Best Effort traffic will be organized at the moments when part of link capacity is free from transmission of QoS traffic having preemptive priority for bandwidth usage
- The aim of the paper is to develop efficient tools for estimation of performance measures of such systems

Model description: QoS traffic

- We consider a network with some number of nodes and available transmission facilities between nodes which we call links
- Let J will be the total number of links and v_j will be the available bandwidth of link number j expressed in basic bandwidth units
- Let n will be the number of QoS flows that are transmitted from one node to another
- Flow number k is characterized by intensity $\lambda_{c,k}$, by mean time of bandwidth occupation $\frac{1}{\mu_{c,k}}$, by number of bandwidth units needed for call transmission b_k and by route R_k that consists of links numbers used by k-th flow. Resource requirements and link usage are expressed by demand-matrix D.
- Let $a_k = \frac{\lambda_{c,k}}{\mu_{c,k}}$ be the offered load of k-th flow.
- Let us suppose that arriving calls for each flow are Poissonian and transmission times are exponentially distributed.

Example of network serving QoS traffic

$J = 5$ links, $n = 4$ priority flows



Demand matrix D (link/flow) has components

$$D = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{vmatrix}$$

QoS traffic: performance measures

- Let $i_k(t)$ be the number of calls of k-th flow being served at time t. The model dynamic is described by a Markov process $r(t) = (i_1(t), \ldots, i_n(t))$. The process $r(t)$ takes values in the finite set of states S defined in accordance with links capacities and call requirement for bandwidth.
- Let $P(i_1, \ldots, i_n)$ be the model's stationary probabilities.
- The process of transmission of QoS traffic will be characterized by blocking probability π_k and by mean number of occupied bandwidth units M_k

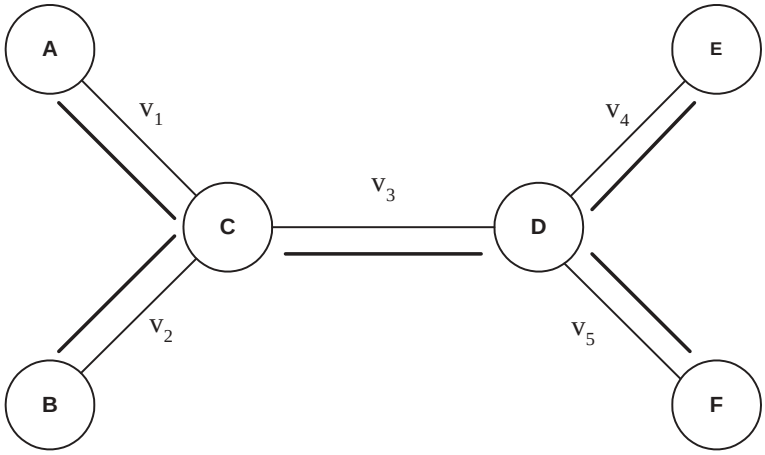
$$\pi_k = \sum_{i_1 b_1 + \ldots + i_n b_n + b_k > v} P(i_1, \ldots, i_n),$$

$$M_k = \sum_{i_k > 0} P(i_1, \ldots, i_n) i_k b_k.$$

Model description: Best Effort traffic

- Best Effort traffic consists of packets. When such a packet arrives to link number j it is accepted for transmission if there is at least one free basic bandwidth unit, otherwise the packet is send to the unlimited capacity buffer associated with j-th link.
- After finishing service on link j, the packet with probability one leaves the network.
- Let us suppose that the flow of Best Effort packets arriving to the j-th link is Poissonian with intensity $\lambda_{d,j}$, $j = 1, 2, \ldots, J$ and that the duration of time needed for servicing one Best Effort packet has an exponential distribution with parameter equal to μ_d.
- If QoS call does not find enough capacity it can interrupt transmission of necessary amount of Best Effort packets which are moved to a buffer. Dismissed Best Effort packet restarts transmission from the beginning. It means that copy of transmitted packet should be kept in special buffer associated with the j-th link up to the moment when transmission is successful.

Example of network serving Best Effort traffic



number of flows is five according to the number of links $J = 5$

Best Effort traffic: performance measures

- The process of servicing of a Best Effort packet on j-th link is characterized by mean delay T_j obtained with Little's formula $T_j = \frac{y_j}{\lambda_{d,j}}$, where y_j is the mean number of Best Effort packets served or waiting at the j-th link.
- The model dynamic is described by a Markov process

$$r(t) = (i_1(t), \ldots, i_n(t), i_{d,1}(t), \ldots, i_{d,J}(t))$$

- The value of y_j is defined as

$$y_j = \sum_S i_{d,j}\, P(i_1, \ldots, i_n, i_{d,1}, \ldots, i_{d,J}),\ j = 1, 2, \ldots, J$$

Exact evaluation of performance measures

- Because QoS service packets have absolute priority exact values of their performance measures π_k and M_k, $k = 1, 2, \ldots, n$ can be found numerically or analytically independently of Best Efforts performance measures with the help of a number of algorithms based on the product form relations valid for all $(i_1, \ldots, i_n) \in S$.

$$P(i_1, \ldots, i_n) = P(0, \ldots, 0)\frac{a_1^{i_1}}{i_1!}\frac{a_2^{i_2}}{i_2!}\ldots\frac{a_n^{i_n}}{i_n!}.$$

- Exact values of performance measures of Best Effort packets T_j, y_j can be found numerically by solving system of state equations (only for small cases) or generally by simulation technique (general case).

The idea of approximate algorithm

The idea of approach is based on the following notions

- If all intensities of Best efforts traffic equal to zero we obtain model of network with losses and product form solution for routes. Performance measures of each route can be found with help of number effective algorithms.
- If all intensities of QoS traffic equal to zero we obtain model of network with waiting also with product form solution but for links. Performance measures of each link can be found separately with help of simple queueing formulas derived in queueing theory for models of full availability group of servers with unlimited buffer and waiting of unsuccessful calls.

Outline of approximate algorithm

Main steps of realization are as follows

1. Using priority property of QoS traffic find its performance measures: portion of lost calls and mean usage of capacity.

2. Dimension link capacities to make small the probability of loss for QoS traffic.

3. Construct the auxiliary model for estimation of performance measures of Best Effort traffic. The foundation of auxiliary model is based on the basic property of exponentially distribution. According to it the moving to waiting of Best Effort call dismissed from service by blocked QoS call can be replaced by moving to waiting QoS call without interrupting service for Best Effort call. It allows to reconstruct the filling of each queue with help of auxiliary model of the link where all traffic flows considered as one flow.

Approximate estimation of QoS flows

- For simplicity, we suppose that all $b_k = 1$, $k = 1, 2, \ldots, n$, but the results obtained can be generalized to the case of arbitrary values of b_k.
- Values of π_k, $k = 1, 2, \ldots, n$ will be found approximately with the help of reduced load approximations

$$\pi_k \approx 1 - \prod_{j \in R_k} (1 - L_j), \quad k = 1, \ldots, n,$$

where values of L_j can be found as solution of the system of implicit equations

$$L_j = E\left(v_j, \sum_{\ell \in N_j} a_\ell \prod_{i \in R_\ell \setminus \{j\}} (1 - L_i)\right), \quad j = 1, \ldots, J.$$

- The values $\Lambda_j = \sum\limits_{\ell \in N_j} a_\ell \prod\limits_{i \in R_\ell \setminus \{j\}} (1 - L_i)$ can be used for estimation of the intensity of QoS traffic offered to the link number j.
- When π_k is found we can estimate M_k from relation $M_k = a_k b_k (1 - \pi_k)$.

Approximate estimation of Best Effort flows

- Let us suppose that the intensity of service are identical for all type of calls, $\mu_{c,k} = \mu_d = \mu$, $(k = 1, \ldots, n)$.
- For link number j the auxiliary model is constructed by considering flows of QoS and Best Effort traffic as one flow intensity $\lambda_j = \Lambda_j + \lambda_{d,j}$ without priority served by v_j servers with waiting of all blocked calls.
- The model is described by Markov process of the type $\ell(t)$, where $\ell(t)$ is the total number of occupied bandwidth units and waiting positions.
- Let $P(\ell)$ be the stationary probabilities of $\ell(t)$.

Approximate estimation of Best efforts flows

- Values $P(\ell)$ are found by the recurrence relations obtained from the system of state equations

$$P(\ell)\,\ell\mu = P(\ell-1)\,\lambda_j, \qquad \ell = 1, 2, \ldots, v_j$$

$$P(\ell)\,v_j\mu = P(\ell-1)\,\lambda_j, \qquad \ell = v_j+1, v_j+2, \ldots$$

- Knowing values $P(\ell)$ we can estimate the mean waiting time T_j of a Best Effort traffic on j-th link and w_j the mean number of Best Effort packets being waiting on the j-th link

$$T_j = \frac{\sum\limits_{\ell=1}^{\infty} P(\ell)\ell - \Lambda_j\left\{1 - E(v_j, \Lambda_j)\right\}}{\lambda_{d,j}},$$

$$w_j = \sum_{\ell=v_j+1}^{\infty} P(\ell)(\ell - v_j).$$

- For different values of service time it is necessary to change the mean value of service time for QoS flows $\frac{1}{\mu_{c,k}}$ to $\frac{1}{\mu_d}$ and proportionally change intensities of QoS flows from $\lambda_{c,k}$ to $\lambda_{c,k}\frac{\mu_d}{\mu_{c,k}}$, $(k = 1, \ldots, n)$. According to the product form property this transition does not change the values of performance measures of QoS flows.

Numerical examples

Input parameters: $J = 5$ links, $n = 4$ priority flows,

$v_1 = 10$, $v_2 = 30$, $v_3 = 35$, $v_4 = 25$, $v_5 = 30$,

$\lambda_{c,1} = 4$, $\lambda_{c,2} = 11$, $\lambda_{c,3} = 8$, $\lambda_{c,4} = 6$,

$\mu_{c,k} = 1$, $k = 1, 2, 3, 4$, and $\mu_d = 1$.

The values of $\lambda_{d,j}$ (scenario I and II) are presented at the bottom of the table Exact (found by simulation) and approximate (found by decomposition algorithm) values of T_j (upper part), y_j (middle part), w_j (lower part). Demand matrix D (link/flow) has components

$$D = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{vmatrix}.$$

Numerical output

Input	Output	Mean delay for Best Effort packets depending on link number				
		T_1	T_2	T_3	T_4	T_5
I	Exact	2.06±0.06	1.32±0,02	1.35±0,02	1.62±0.03	1.36±0,02
	Approx	2.13	1.32	1.36	1.65	1.39
II	Exact	1.58±0.03	1.24±0,03	1.29±0,03	1.48±0.02	1.30±0,02
	Approx	1.66	1.25	1.29	1.50	1.31

Input	Output	Mean number of Best Effort packets being on servicing and waiting depending on link number				
		y_1	y_2	y_3	y_4	y_5
I	Exact	10.30±0.09	15.79±0,08	17.56±0,09	14.53±0.08	13.57±0,09
	Approx	10.63	15.85	17.64	14.84	13.90
II	Exact	7.29±0.08	14.21±0,06	16.01±0,08	12.14±0.08	12.37±0,09
	Approx.	7.49	14.40	16.11	12.75	12.44

Input	Output	Mean number of Best Effort packets being on waiting depending on link number				
		w_1	w_2	w_3	w_4	w_5
I	Exact	5.28±0.09	3.80±0,07	4.58±0,08	5.67±0.06	3.72±0,07
	Approx.	5.62	3.87	4.67	5.84	3.90
II	Exact	2.77±0.06	2.74±0,05	3.60±0,06	4.08±0.06	2.87±0,07
	Approx.	2.97	2.90	3.62	4.23	2.93

I		$\lambda_{d,1} = 5$	$\lambda_{d,2} = 12$	$\lambda_{d,3} = 13$	$\lambda_{d,4} = 9$	$\lambda_{d,5} = 10$
II		$\lambda_{d,1} = 4.5$	$\lambda_{d,2} = 11.5$	$\lambda_{d,3} = 12.5$	$\lambda_{d,4} = 8.5$	$\lambda_{d,5} = 9.5$

Conclusions and further results

- An approach that can be used for description of telecommunication networks with differentiated services is described. In the model we distinguish between QoS traffics for real-time services, such as Internet telephony or video, and Best Effort traffic for relatively delay insensitive services such as data transfer.
- It is shown that performance measures of the differentiated services can be found with the help of approximate algorithm based on a decomposition technique.
- The established algorithm is very easy to implement and it provides good loss estimate over a wide range of parameters.
- Obtained results plan to be generalized for a number of other situations, for example, for models when after finishing service time on one link Best Effort call with some probability can leave the network or with another probability moves to other link.