Riga Technical University
Faculty of Electronics and Telecommunications
Department of Transport Electronics and Telematics

Mihails Kulikovs
PhD Program "Computer science, information and electronics systems"

Research and development of effective managing algorithms in admission control systems for telecommunication networks

Doctoral diploma thesis

Supervisor
Dr. habil. sc. ing., professor
Ernests Petersons

Riga 2010

Abstract

The main purpose of the present research is the development of managing algorithms in the control system that admits data flows to communication resources, thus providing the required quality of service.

The algorithms covered in the study are admission control and flows redistribution, the algorithms of resources redistribution and maximization of system load that lead to fulfillment of quality of service requirements.

The research resulted into elaboration of recommendations regarding optimization of admission control system functioning. In order to test the recommendations an *OPNET* modeling system has been applied.

The dissertation consists of an introduction, five chapters, conclusion exposed on 200 pages and containing 17 tables, 112 pictures. Also, the paper contains a list of literature used composed of 175 items and 4 appendixes.

Anotācija

Šī darba pamatuzdevums – izstrādāt vadības algoritmus datu plūsmu piekļuves sistēmai pie komunikāciju resursiem, vienlaicīgi nodrošinot uzdoto apkalpošanas kvalitāti.

Tika pētīti sekojošie algoritmi: piekļuves vadības un plūsmu dispečerizācijas algoritmi, resursu pārdales algoritmi, sistēmas noslodzes maksimizācijas algoritmi ar apkalpošanas kvalitātes garantiju nodrošināšanu.

Pētījumu rezultātā tika izstrādās rekomendācijas piekļuves vadības sistēmas funkcionēšanas optimizācijai. Izstrādāto rekomendāciju pārbaudei tika izmantota *OPNET* modelēšanas sistēma.

Disertācija sastāv no ievada, piecām sadaļām, nobeiguma, kas izklāstīti 200 lappusēs, ar 17 tabulām, 112 attēliem, 4 pielikumiem, kā arī ar izmantotās literatūras sarakstu no 175 nosaukumiem.

Contents

1. Introduction

Research purpose: The purpose of present research is the elaboration of effective methods to manage flows and allocate resources for systems of admission control based on measurements in networks with traffic of bursty character.

The actuality of the work is connected with the current situation in the field of telecommunications where it promised that by the end of 2010 the need for resources will outreach global network capacity [167]. Moreover, it is stated that in order to overcome the increasing demand, $137 billion would be necessary.

The work presents recommendations for development of managing algorithms and optimal resources allocation that give a possibility to reduce investment expenses.

For the last years the character of services offered to Internet users has changed crucially. Multimedia services have become dominating over traditional plain text (web, mail) and data (ftp) services. Hence, one can argue the evolution of Internet has occurred. In particular, two fields have evolved:

- Traffic - its parameters have changed
- Services - multimedia applications began requiring quality of service guarantees

As far as traffic development is concerned, bursty flows has become prevailing over traditional Poisson ones [79, 80, 5]. In this kind of traffic packet income time distribution comply with long-tail distribution, Pareto for example. Analysis has shown

that this traffic can be characterized by self-similarity property. It results in keeping statistical parameters under varying time scales. This traffic character dramatically influences network performance. It increases delays and packet loss probability [143, 50, 112].

In present work in Chapter 2. the author examines the models that describe traffic and analyses the recognized models of traffic generation with self-similar character in the simulation systems of communication networks. As a result, the recommendations on generation of self-similar traffic in simulation systems have been elaborated. Multiple simulations and its analysis proved the existing opinion about high correlation of self-similar traffic, as well as its heavy influence on communication system performance. The negative effect of self-similar traffic on network performance may be significantly reduced by the means of decreasing overloads of communication system elements.

One of the options to reduce overloads during the network operation is taking into account traffic characteristics on the stage of design when structural network units such as buffer size and bandwidth are being chosen [163, 9, 134, 165, 135, 48, 44, 8]. Optimal values of network structural units' parameters are chosen employing a queuing model. Due to self-similar character of traffic, the traditional queuing models are not suitable as they are built around the assumption of Poisson character of the incoming flow, e.g. $M/M/1/K$ [23], or unlimited resources, e.g. buffer size - $P/M/1$ [115]. Therefore, the author has made an investigation of queuing models that match the qualities of modern traffic. During the study, the author has identified a queuing model [139] that is convenient for making the calculations. It has packet arrival process characterize by Pareto distribution low, requests service is distributed according to exponential law; it has one service node and limited memory size ($P/M/1/K$). The model offers numeric method of calculating the probability of packet loss (P_{Loss}) under the specified buffer size and system utilization (ρ). In addition, a method to find a required K when ρ and P_{Loss} are specified is shown. It is described in Chapter 3. Due to the lack of possibility to calculate the important parameters of packet delay and average queue length, as well as the limited model which is able only for numeric calculations gave reasons for further research.

Studies have shown that due to the absence of Laplace transform for the "long-tail" distributions, there are no analytical expressions for queuing model with self-similar character of incoming flow. The fact the self-similar traffic is bursty, draws attention to analytical expressions of the work [20] for the queuing model with exponential service with one service node and bursty income ($M^X/M/1$). The expressions of the average queue length under an assumption about unlimited buffer size and packet loss probability under limited buffer size have been explored. The work proves this model,

operating within an utilization range that is typical for telecommunications systems, is similar to $P/M/1/K$ model if to compare the numeric indicators of queue parameters. The model has been further developed and added by analytical expressions for the average queue length ($\bar{K}$) and average packet service waiting time ($\bar{W}$) under the limited buffer size assupmtion. Using the analytical expressions of the specified system indicators with the queue such as P_{Loss} and $\bar{W}$, it is possible to design system structural units. On the basis of the analytical expressions for the systems with bursty incoming flows, Section 6.1. presents recommendations regarding the structural units of systems like buffer size (K) and throughput under the specified P_{Loss} and minimal cost.

Another option to reduce the overload of structural elements and the whole system is employing the algorithms that manage the flows. The managing algorithm is dedicated to provision of Quality of Service (QoS) guarantees to network clients. A large part of the present work (Chapter 6.) is devoted to the recommendations about design of managing algorithms. There are several problems in design of managing algorithms:

- Functional
- Structural
- Technological

The functional problems are related to the development of decision making methods of management. Hereby, management is considered to be the decision making process regarding the necessary QoS guarantees.

In this work it is assumed that QoS requirements have to be satisfied in the End-to-End (e2e) mode. This mode includes the fulfillment of the requested QoS along the entire way from the data source to the destination. In order the specified QoS are guaranteed it is important that a newly incoming flow does not violate QoS guarantees for the existing flows. Therefore, Connection Admission Control (CAC) has to become the main target function of the managing algorithm [127, 155].

Admission control is a set of actions performed by the network at the stage of a new connection establishment (or connection restoration) for the making of decision regarding the possibility to support a new connection with the requested parameters and quality of service. The new connection may be supported by the network only if it has the requested resources, requirements to quality of service for the existing connections are met and under the condition the new connection does not disturb them.

In order to make a management decision, in case of CAC it is to admit or reject, it is proposed to use bursty queuing models mentioned forehand. Thus, if P_{Loss}

and $\bar{W}$, calculated using the queuing models described above, satisfy the new flow requirements, the flow gets admitted. Otherwise it gets rejected.

So that a decision can be made, the managing element needs information. The problem of getting and processing the information is widely regarded in the present research. The network management could be successful if the incoming flow includes the following information:

- Quality of service requirements - P_{Loss}, $\bar{W}$, etc.
- Resources requirements - Peak Bit Rate (PBR), Sustainable Bit Rat (SBR)

For real time applications, interactive video for example, it is impossible to estimate the resources that might be needed in advance. For this reason the required resources are raised so one could be confident in fulfillment of QoS. This action has caused the failure of CAC that uses the requirements set by the flows for utilization evaluation, for example, the requirements to bandwidth. Such CAC is called Parametric-based Admission Control (PBAC), as it uses application traffic parameters for admission control. PBAC can be easily realized. However it provides low network utilization due to the raised requirements.

In a chapter 6. of the present work it is suggested to use CAC that gets the information about utilization directly from the performed measurements of the existing flows. Such an approach is called Measurement-Based Admission Control (MBAC) and recently it has become popular in provision of statistical QoS guarantees [82, 81, 78, 63]. In this work a new structural and functional MBAC scheme is proposed.

Measurement and evaluation of both existing and incoming traffic parameters module is one of MBAC structural elements. Further, the main problems of this module elaboration are described, as well as the possible ways to overcome them.

Data measurements and processing for clarification of system load is a complicated task due to traffic being bursty [7, 39, 34, 103, 69, 78]. Accumulated and analyzed statistics of the traffic during the small time interval may show oversized values of traffic parameters in case the measurement interval gets into the interval of packet burst, which correspondingly, leads to the overestimation of the required resources. Also, underestimation of traffic parameters might happen. That is in case the measurement interval gets into the empty interval that leads to the underestimation of the required network resources.

At the same time the gathered and analyzed statistics of the given traffic for a large time period produces underestimated parameter rates as packet groups of high income intensity will be averaged along to the intervals with low incoming packets intensity.

In search of the compromise the author has proposed using an adaptive approach that denotes observation period and sampling period dependence on the traffic character. The suggested adaptive measuring method allows not only reducing the number of errors in traffic measurements, but what is extremely important when real time mode is employed - reducing the managing system load related to the measuring process. Thus, the Section 6.2.1. offers solutions to the functional problems occurred during the design of the managing algorithms.

The second problem during the process of managing algorithms design is a structural problem caused by maximum usage of the limited network resources. In case of CAC, it means the admission of the large number of flows maintaining QoS guarantees. In case of bursty traffic a necessary throughput capacity will be smaller than the sum of declared peak rate of separate flows [175]. That difference is a gain which can be used to connect additional flows. In the work the Section 6.2.2. presents the studies and recommendations regarding usage maximization of the limited system resources. They also consider the calculations of the maximum number of allowed flows under maintenance of QoS requirements.

The best usage of the limited resources can be reached by the algorithm of optimal dispatching proposed by the author. Flow dispatching is needed in the case when the tie of all incoming connections is impossible. The situation is easily manageable when there are priority flows among the incoming flows - the flow that is in priority gets connected. The algorithm proposed by the author and described in the Section 6.2.3. solves the dispatching problem when flows belong to the same priority class. This algorithm allows obtaining the maximum load fulfilling QoS guarantees.

The third problem of the managing algorithm is the technological one. Its solution effects the cost of the managing system. As it has been mentioned above, the author offers recommendations about optimal calculation of structural system units values at the stage of design. It is known that the parameters of flows quality of service depend on bandwidth of outcoming channel and buffer size of the communication node [95]. Herewith, these parameters influence level depends on the traffic character and communication system load level [100, 116, 151, 156]. Taking into account the fact that the costs of resources vary, the author of the present work solves the task of optimal resource reallocation which is described in the Section 6.2.4. Reallocation studies and recommendations concern such structural units as buffer size and bandwidth between the flows in the process of system functioning. This reallocation is intended for minimization of communication system costs in general keeping up QoS guarantees.

All the recommendations have the analytical justification. The author has tested his recommendations by realization of MBAC model in OPNET© network modeling framework. Modeling results are reflected in the Chapter 7. It shows that using

only some recommendations, already provides substantial increase in communication system performance.

The research of this work has been made after the analysis of many publications dedicated to the problems of traffic analysis and management in telecommunication systems. However, the author has not found the solution to the problem formulated above during the review of the dissertation chapters, and it has become a subject of the present work.

Contributions: The contributions of this work can be summarized as follows:

- The optimal buffer size and throughput together with the algorithm of estimating the set packet loss probability minimize the summarized system costs.
- The method of estimation the adoptive evaluation and sampling periods provides the possibility for decreasing the estimation error and false decision making.
- An additional channel multiplexing possibility with the decrease of packet arrival rate of bursty flow with keeping guarantee of packets lost probability.
- The method including the optimal dispatching of the same priority class flows allows to improve the quality of service.
- The algorithm of the system resources reallocation, such as buffer size and throughput, allows lowering the overall costs while securing the quality of service requirements.

Approbation The results of this work were conveyed and discussed at the following events:

1. 2010 IEEE Region 8 International Conference on Computational Technologies in Electrical and Electronics Engineering SIBIRCON-2010, Irkutsk Listvyanka, Russia, July 11-15, 2010
2. The 14^{th} International Conference ELECTRONCS 2010// Kaunas, Lithuania, May 18-20, 2010
3. The 19^{th} Annual Wireless and Optical Communication Conference// Shanghai, China, May 14-15, 2010
4. The Fifth Advanced International Conference on Telecommunications AICT 2009// Venice/Mestre, ITALY: IARIA, May 24-28, 2009
5. The 13^{th} International Conference ELECTRONCS// Kaunas, Lithuania, May 13, 2009

6. IADIS International Conference Telecommunications, Networks and Systems 2008// Amsterdam, Netherlands, July 22-24, 2008

7. World Congress on Science, Engineering and Technology// Paris, France, July 4-6, 2008

8. The 14^{th} Conference on Information and Software Technologies IT 2008// Kaunas, Lithuania, April 24-25, 2009

9. The 12^{th} International Conference ELECTRONCS// Kaunas, Lithuania, May 21, 2009

10. RTU 48^{th} International Science Conference// Riga, Latvia, October 11-13, 2007

11. The 11^{th} International Conference ELECTRONCS// Kaunas, Lithuania, May 13, 2009

Publications The following papers of the author have been published:

1. Mihails Kulikovs, Ernests Petersons, Sergeys Sharkovsky. Integral measurement process of incoming traffic for measurement-based admission control// Proceedings of the 2010 IEEE Region 8 International Conference on Computational Technologies in Electrical and Electronics Engineering SIBIRCON-2010, pp. 183-186

2. M. Kulikovs, S. Sharkovsky, E. Petersons. Comparative Studies of Methods for Accurate Hurst Parameter Estimation// ELECTRONICS AND ELECTRICAL ENGINEERING No. 7(103), 2010

3. Mihails Kulikovs, Ernests Petersons, Sergeys Sharkovsky. Adaptive Traffic Measurement for MBAC System// Proceedings of 2010 19th Annual Wireless and Optical Communications Conference (WOCC 2010), pp. 354-358

4. M. Kulikovs and E. Petersons,Optimal Dispatching of the Flows Falling in the Same Priority Class// Automatic Control and Computer Sciences, 2010, Vol. 44, No. 1, pp. 42-46. ©Allerton Press, Inc., 2010.

5. Куликовс, Э. Петерсонс, Оптимальная диспетчиризация потоков, принадлежащих одному классу // АВТОМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА. 2010. No.1. С.58-64.

6. Alfreds Asars, Mihails Kulikovs, Ernests Petersons. Buffer Size and Output Bandwidth Optimization in a MBAC system// Automatic Control and Computer Sciences, Vol. 43, No. 5, pp. 241-246.

7. А. Асарс , М. Куликовс, Э. Петерсонс. Оптимизация объема буферной памяти и пропускной способности выходного канала в MBAC системе. //АВТОМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА. N5, 2009, pp.24-31

8. Mihails Kulikovs, Ernests Peterson. Real-Time Traffic Analyzer for Measurement-Based Admission Control// Proceeding of the 2009 Fifth Advanced International Conference on Telecommunications, 24-28 May 2009, Venice/Mestre, Italy, pp 72-75

9. Mihails Kulikovs. Statistical parameters estimation of the self-similar input traffic for the Measurement-based Admission Control// Scientific Proceedings of Riga Technical University in series "Telecommunications and Electronics". -Riga, 2008, pp 37. - 42.

10. M.Kulikovs, E.Petersons, Modeling the On-line Traffic Estimator in OPNET// ELECTRONICS AND ELECTRICAL ENGINEERING No. 7(95) 2009, pp. 82-86

11. М. Куликовс, Э. ПЕТЕРСОНС. Оценка параметров имитационной модели системы доступа в сеть самоподобных входных потоков.// АВТОМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА. N2, 2009, pp. 47 - 56

12. M. Kulikovs and E. Petersons Parameter Estimations in a Model Imitating a Network Admission System with Self-Similar Arrival Streams// Automatic Control and Computer Sciences, 2009, Vol. 43, No. 2, pp. 88-95. ©Allerton Press, Inc., 2009.

13. Kulikovs M., Petersons E. Remarks Regarding Queuing Model and Packet Loss Probability for the Traffic with Self-Similar Characteristics// Proceeding of World Academy of Science, Engineering and Technology. Volume 30,PWASET, 2008, pp. 537- 543.

14. Kulikovs M., Petersons E. REMARKS ON PACKET LOSS PROBABILITY FOR THE NETWORK TRAFFIC WITH SELF-SIMILAR BEHAVIOUR// Proceedings of WIRELESS APPLICATIONS AND COMPUTING 2008 and TELECOMMUNICATIONS, NETWORKS AND SYSTEMS 2008. 3. IADIS Press, 2008, pp. 85-90.

15. Kulikovs M., Petersons E. Packet Loss Probability Dependence on Number of ON-OFF Traffic Sources in OPNET// ELECTRONCS AND ELECTRICAL ENGINEERING. 5. E2008, 2008, pp. 77-80 Kaunas

16. Kulikovs M., Petersons E. Estimation of buffer overflow probability by OPNET// Conference Proceedings. 14th Conference on Information and Software Technologies. - Kaunas, Lithuania: Tehnologija, 2008. pp. 145-149.

17. Asars A., Petersons E., Kulikovs M. Modeling of measurement based admission control algorithm, using OPNET// Scientific Proceedings of Riga Technical University in series "Telecommunications and Electronics". -Riga, 2007, pp 16-21

2. Traffic characteristics

Network management and control is impossible without understanding the notion of network traffic and its characteristics. This chapter is dedicated to description of network traffic. Traffic evolution and its parameters are described. Along to traffic evolution and its parameters there appeared a necessity of renewal and models that describe the traffic and is presented in Section 2.1.

Section 2.2. describes the parameters of modern traffic, generation ways and its analysis.

2.1. The Modern Traffic

The section describes the evolution of network traffic and its parameters. Section 2.1.1. considers the development of network traffic using source model. Along to advancement of network technologies the source model stopped to be acceptable. A network model took its place (described in Section 2.1.2.). Section 2.1.3. is dedicated to applications characteristics that are being used in the present paper.

2.1.1. Link/Source Model

The link/source model is one of the approaches to network resource management.

The models of source requirements (the effective bandwidth) and transmission link (link-capacity and link buffering) are used to measure the network in this kind of modeling. It can be easily applied to the POTS where the source traffic model is simple: a 64 Kbps CBR

source [67].

The talk spurt and periods of silence represent a speech as the ON/OFF source, and can be considered as a variable data-rate source.

[74, 28, 110] have shown that the ON/OFF source model of speech may be represented by 2-state ON-OFF Markov model. In [29] and [66] the mean periods for the talk and silence periods were derived as 650 ms and 352 ms respectively.

Removing the periods of silence a voice conversation can be compressed and less data needs to be transmitted. Therefore, it will require less network resources. The source model for the compressed voice cannot be represented as a CBR source but should be represented as VBR source model. Such model has three parameters in comparison to the one of CBR. The model of voice is described using:

- burst size: the mean talk period (ON- period)
- the peak bit rate, the rate at which data is transmitted when the model is talking or is ON
- the sustained bit rate which is estimated from the mean burst size and the mean silence period (OFF)

In [65, 31] the authors proposed to use the VBR descriptions of traffic to dimension link for a given number of VBR traffic sources. Such proposals are valid as long as the traffic sources may be represented using Markov models.

Long before silence suppression was an available compression technique, Markovian models were used to describe other aspects of the telephone network. The duration of the telephone call and the time between consecutive calls were presented as 2-state ON-OFF Markov models by [49] and [111]. Using these models, network providers were able to make the best use of telephone resources while providing a computable level of service to customers.

However, telephone use has developed along with the evolution of telephone networks technology. Since telephone calls started to use for computer data the models proposed become not valid. In [25, 24, 162] reported that call characteristics significantly differ and that the distribution of call duration has changed from an exponential to distributions with a ”heavy-tail” and the flow-lifetimes are better represented by log-normal or Pareto distributions. The [126, 41, 52] show that in a packet-based network such as the Internet, models of the traffic are now also best represented by ”heavy-tail” distributions.

Examples of VBR traffic, characterized by ”heavy-tail” distributions, include LAN traffic [96], and WWW transactions carried over TCP/IP [41]. The reason why the ”heavy-tail” distribution characterizes Internet traffic more accurately is the following. Each of these sources, when multiplexed together, exhibit the properties of self-similarity. Traces analysis shows that WAN and WWW traces exhibits self-similarity [52] and [41] respectively.

Firstly introduced in [105], multi-fractal analysis has been adopted to best characterize the self-similar traffic in order to create suitable models for the link/source approach. A multi-fractal analysis of Internet traffic traces was reported in [137] and later extended in [138] to create a framework that supported multi-scale modeling of network traffic.

There are several reports showing that self-similarity property is not unique to Internet. The self-similarity presence reported in VBR video traffic [58]. In [142] the researcher reported the studies of VBR video traffic in ATM and showed that the loss-rates are not immensely affected by the self-similarity properties. Also, Markovian models have been shown to be sufficient for performance analysis. This report shows that short-term traffic correlations had a more significant effect on buffer behavior for VBR video traffic than the effects of self-similarity. The findings of [142, 62] propose the existence of an event-horizon beyond which the self-similar characteristics of traffic do not depend on that time-scale. An example is the timescale determined by the buffer-size of the system self-similar properties of traffic at timescales beyond this will have little impact on this timescale.

It is obvious that the network traffic models must be developed together with analysis techniques and evolution of the network traffic itself. The source model does not provide secure means of required resource for traffic with complicated and less-tractable properties. The source model provides corruptible results in the multi-hop network as the traffic passes through several network nodes and may not be desired nor understood. If the source gets smoother as it passes through each switch, then effective bandwidth based upon *a priori* declarations will be overly pessimistic. Additionally, [40] showed that burst-compression is a common phenomenon in multi-hop networks, making the traffic increasingly bursty [79, 80, 5] and therefore more demanding on network resources. Such experience further draws into doubt the usefulness of traditional traffic models for the accurate description of any more information than a flow peak data rate.

Such difficulties in applying a link/source model to satisfactorily characterize evolving traffic or to accurately characterize the elastic traffic carried through the Internet, an approach called the network model has arisen.

2.1.2. Network Model

Instead of a link/source model, the model that could take into account the interaction with the network and which would be suitable for characterizing elastic traffic, such as a TCP/IP network, was needed. Such network model was supported in [161]. Moreover, [11] illustrated the significant difference between the link/source modeling approach and simulations of the behavior of elastic traffic. The link/source model approach may suit traffic sources that do not interact with or adapt to the current network's behavior. However, Internet traffic is fundamentally adaptive in nature because of the TCP protocol control [10, 71]. The source behavior cannot be discussed separately from the network configuration (buffer management,

network routing or packet-scheduling) because of the adaptive protocol interaction with the network. The link/source model is valid only for the static configuration. If the network configuration changes, so do the model observations of network behavior.

In [161], the author proposed that the TCP mechanism for congestion control is fundamentally chaotic in nature. This means that TCP-based networks are sensitive to initial conditions and have unpredictable behavior that should be considered in the model.

Consideration of the fundamental operation of TCP increases the complexity of the models enormously. The behavior analysis of TCP protocol under the light and moderate loss was presented in [107] where the author offered the effective bandwidth approximation formula.

In [117] the behavioral model relating effective throughput and loss are presented. The model taking particular note of the impact the timeout mechanism has on performance. Based on the report [117], a network-model for the short lifetime TCP flows was presented in [32]. The models presented in [107, 117] and [32] stand on fundamental assumptions that the TCP congestion control is behaving in a periodic/predictable fashion. A contradicted fundamental idea is that cooperating TCP congestion control processes will form a deterministic, but chaotic system [11].

It is necessary to mention that the TCP protocol is not the only adaptive protocol in Internet. A number of adaptive services are built upon the UDP protocol [129] which offers a datagram service and reliable and adaptive transport mechanisms be built upon it.

A few works on the use of wide-area UDP-based traffic are presented in [71, 120]. The study of [71] states that UDP is the principle carrier of real-time traffic. Such a statement is confirmed in [109].

2.1.3. Application Traffic Models

The application traffic models are categorized in to two types of traffic modeled: foreground and background traffic. The foreground traffic model represents a specific user behavior or interaction, or in other word, application traffic is generated mainly through a user interaction with a device. One of the goals of the simulator to implement the foreground traffic model is to evaluating a user perspective application performance in detail as specified in [89].

The background traffic, however, is not directly related to a user interaction. For certain application or service generates network traffic as soon as a session starts regardless of a user interaction with a device and the amount of network traffic is significantly larger than foreground traffic for the application or service. Instant Messaging and VPN service are the examples. The usages of these two levels of model are the following:

- User level traffic model:
  - User behavior interactions in an application is modeled, and this model can be used with a simulation which includes detailed application layer, transport layer

and IP layer model on top of the Layer 1(PHY) and Layer 2 (MAC) models.

  - Application performance metrics specific to the application can be evaluated (e.g., web page download time, email download time, FTP download time, etc), and a scheduling mechanism for the application QoS in the MAC layer can be evaluated or optimized.
  - Since the IP level traffic will be generated according to the upper layer protocols, the generated Uplink (UL) and Downlink (DL) traffic will be correlated.

- IP packet level traffic model:

  - This model generally obtained from a network traffic measurement and represented as statistical packet distributions, such as packet size distribution and packet inter-arrival time distribution at the IP layer, and this model can be used with a simulation which does not includes detailed protocol layers above Layer 2 (MAC) models.
  - The IP packet level traffic model can be directly applied to the Layer 2 and Layer 1 model, and evaluating both application performance and scheduling mechanism in the MAC layer are not easy task with this type of model.

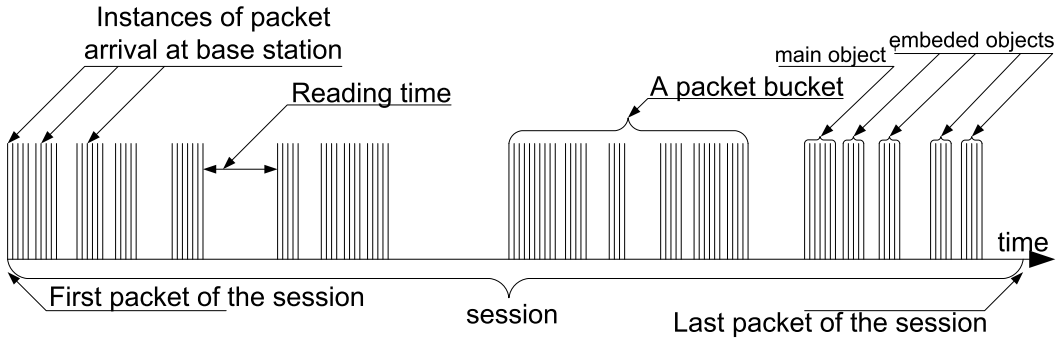
Following subsections is represented description for user level.

Web Browsing Traffic Model The web model is one of the most complex models of all. Measurements of HTTP traffic show that the large majority of page responses consist of relatively small objects. The distribution of page sizes however is ”heavy-tailed”, i.e., infrequent but very large page objects constitute a significant proportion of overall transmitted bytes. Each web page consists of a number of web objects such as a main page, embedded images, style sheets and executable java applets or plug-ins. The time between accessing two pages is denoted the think time (reading time) and includes time for the user to read all or parts of the page.

Since the web browsing model in [88] might be outdated, a recent study has been incorporated in [1].

Fig. 2.1. shows the packet trace of a typical web browsing session. The session is divided into ON/OFF periods representing web-page downloads and the intermediate reading times, where the web-page downloads are referred to as packet bucket. These ON and OFF periods are a result of human interaction where the packet bucket represents a user's request for information and the reading time identifies the time required to digest the web-page.

As is well known, web-browsing traffic is self-similar. In other words, the traffic exhibits similar statistics on different timescales. Therefore, a packet bucket, like a packet session, is divided into ON/OFF periods as in Fig. 2.1. Unlike a packet session, the ON/OFF periods



2.1. Figure Packet Trace of a Typical Web Browsing Session

Main object size (S_M)	Truncated Lognormal(η= $10710bytes$, σ= $25032bytes$, $min = 100bytes$, $max = 2Mbytes$)
Embedded object size (S_E)	Truncated Lognormal (η= $7758bytes$, σ= $126168bytes$, $min = 50bytes$, $max = 2Mbytes$)
Number of embedded objects per page (N_d)	Truncated Pareto (η= 5.64, $max = 165$)
Reading time (D_{pc})	Exponential (η= $30sec$)
Parsing time (T_p)	Exponential (η= $0.13sec$)

2.1. Table Web browsing traffic model

within a packet bucket are attributed to machine interaction rather than human interaction. A web-browser will begin serving a user's request by fetching the initial HTML page using an HTTP GET request. The retrieval of the initial page and each of the constituent objects is represented by ON period within the packet bucket while the parsing time and protocol overhead are represented by the OFF periods within a packet bucket.

HTTP/1.1 persistent mode transfer is used to download the objects, which are located at the same server and the objects are transferred serially over a single TCP connection. The distributions of the parameters for the web browsing traffic model are described in Table 2.1. which has been obtained through a set of measurements using a recent online-traffic analysis [1]. The parameters η and σ are the mean and the standard deviations.

File Transfer Traffic Model In FTP applications, a session consists of a sequence of file transfers, separated by reading times. The two main parameters of an FTP session are: the size of a file to be transferred, and reading time, i.e., the time interval between end of download of the previous file and the user request for the next file.

The parameters for the FTP application sessions are described in Table 2.2.

Main object size (S)	Truncated Lognormal(η= $2Mbytes$, σ= $0.722Mbytes$, $max = 5Mbytes$)
Reading time (D_{pc})	Exponential (η= $180sec$)

2.2. Table FTP traffic model parameters

Average Call Holding Time (sec)	Exponential: η= 210 sec
Voice CODEC	AMR (12.2kbps)
Frame Length	20 msec
Talk spurt length	Exponential: η= 1026 ms
Silence length	Exponential: η= 1171 ms
Silence suppression	Yes
Protocols	RTP/UDP/IP
Header compression	RTP/UDP/IP header compression
Speech Activity	47.17%
Uplink:Downlink Ratio	1:1
Total MAC PDU size during a talk spurt	42 bytes
Total MAC PDU size during a silence	16 bytes
Average BW usage at the MAC layer	9.25kbps (w/o HARQ CRC 2Bytes), 9.71kbps (w/ HARQ CRC 2Bytes)
VoIP characteristics may vary for a particular codec	

2.3. Table VoIP traffic model

VoIP Traffic Model There are a variety of encoding schemes for voice (i.e., G.711, G.722, G.722.1, G.723.1, G.728, G.729, and Adaptive Multi-Rate Audio (AMR)) that result in different bandwidth requirements. Including the protocol overhead, it is very common for a VoIP call to require between 5 Kbps and 64 Kbps of bi-directional bandwidth.

The VoIP traffic model in Table 2.3. assumes AMR codec and RTP/UDP/IP header is included in the packet size calculation. AMR codec [118] is the most important vocoder in wireless applications. Most likely, header compression will be used and hence the actual VoIP packet size will be 33 bytes plus 3 bytes compressed RTP/UDP/IP header. For the simplicity, signalling traffic is not modeled.

Voice call activities generate a pattern of talk spurt and silence (or ON and OFF) intervals by means of a speech activity detector so that it can be modeled as a two state Markov chain. The experimental measurements with ten conversations of 15 minutes in length showed the activity rate was 0.4717 with AMR, and the mean duration of the ON periods was $1026ms$, while the OFF periods mean duration was $1171ms$ [51]. In the voice traffic model $1026ms$ for a talk spurt and $1171ms$ silence period are used followed by [51].

Video Conference Traffic Model Video conferencing has differing bandwidth requirements for the audio and the video components.

For example, the audio component of a video-conference requires between 16 and 64 Kbps and the video component of a video-conference requires between 320 Kbps and 1 Mbps. The typical sustainable send/receive throughput requirements range is from a low of 32 Kbps to a high of 1 Mbps. A typical business-quality video conference runs at 384 Kbps and can deliver TV-quality video at 25 to 30 frames per second.

Moving Picture Experts Group (MPEG) compressed videos are composed of pictures (frames) that are separated into three different types: I, B, and P. I frames are intraframes that encode the current picture, while B and P frames interpolate from previous and future frames. When transmitted over an IP network, these frames are segmented into one or more IP packets.

Authors of [92] describe Moving Picture Experts Group 4 (Standard - Compressed Video at 64 Kbps) (MPEG4) video traffic modeling, and they separate the video trace into I, B, and P frames, model each set of frames, and combine them together to form a final model. In [108] authors describe MPEG4 video traffic modeling with separate set of I, B, and P frames.

The estimated values, for the parameters to model a video stream, vary from one trace to another. The scene lengths have a lognormal distribution, which means that the log of the scene length has a normal distribution.

For parameters associated with the lognormal distributions, the estimates depend strongly on the dimensions of the captured frames. For the video conferencing traffic model, high quality trace is described in Table 2.4. In the model two different resolutions for the display are considered: 176x144 for a small device and 320x240 for a large device. The required bandwidth for the uncompressed video stream with 176x144 pixels and 8-bit color depth is about 7.6 Mbps and with 320x240 pixels and 8-bit color depth is about 23 Mbps.

2.2. The Modern Traffic Parameters and Modelling

It was shown in previous section that currently there are various network services, such as WWW, FTP, VoIP, and video streaming. Because of that, the networks are full of streams with different characteristics. A number of empirical and analytical studies of traffic measurements from a variety of working packet networks have convincingly demonstrated that actual network traffic is self-similar or long-range dependent in nature - in sharp contrast to commonly made traffic modeling assumptions.

An analysis of network traffic has shown the presence of self-similarity both in local and in large-scale networks, which creates significant difficulties for a correct assessment and prediction of the traffic characteristics [41]. As has been shown in [2, 4, 99, 115, 121] the self-similarity has a negative influence on the network productivity. Various studies identified some evidences of self-similar behavior in computer network traffic, as well as its severe implications in network performance [122, 123]. Namely, due to presence of burst traffic in

Session Duration (sec)	3600 sec	
Video CODEC	MPEG4	
Protocols	RTP/UDP/IP	
Header compression	RTP/UDP/IP header compression	
Scene Length (sec)	Lognormal(η= 5.1 sec, σ= 9.05 sec)	
Direction	Bi-direction	
Display size	176x144	320x240
Color depth (bit)	8	
Sub-sampling method	4:1:1	
Mean bandwidth for Uncompressed stream	7.6 Mbps	23 Mbps
Compression ratio	13.95	
Mean bandwidth for compressed stream	0.54 Mbps	1.65 Mbps
I frame size	Lognormal(η= 210, σ = 1798)	Lognormal(η= 18793, σ= 5441)
P frame size	Lognormal(η= 2826, σ = 1131)	Lognormal(η= 8552, σ= 3422)
B frame size	Lognormal(η= 1998, σ = 716)	Lognormal(η= 6048, σ= 2168)
AR coefficient	$a_1 = 0.39$, $a_2 = 0.15$, $\sigma_\epsilon = 4.36$	

2.4. Table Video conference traffic model

several time-scales leading to an increase in end-to-end delay packet delays and their losses [94, 160, 73, 93].

Next, a mathematical description of self-similar effect will be described. Section 2.2.1. presents the main parameters characterizing the process as being self-similar. Section 2.2.2. is dedicated to the methods of self-similar traffic generation that are used in the current work, as well as the analysis of obtained sequences (Section 2.2.3.).

2.2.1. The Modern Traffic Parameters

The main characteristic parameter for the self-similar processes is the Hurst parameter - it shows the self-similarity level. The value H= 0.5 indicates the absence of self-similarity, while high (close to 1.0) values of H show a high degree of self-similarity.

Basically, speaking about self-similar precesses we mean second-order self-similarity. There are two classes of self-similar processes, namely exact self-similar and asymptotically self-similar processes [98]].

We consider a discrete-time stationary stochastic process $X = (\ldots, X_{-1}, X_0, X_1, \ldots)$. The autocorrelation function of process X is

$$R(k) = \mathbf{E}\left\{(X_i - \eta)(X_{i+k} - \eta)\right\}/\sigma^2 \tag{2.1}$$

The distribution density	$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$
The average value:	$\frac{\alpha x_m}{\alpha-1}$
The variance:	$Var \to \infty$ for $1 < \alpha < 2$

2.5. Table The statistical characteristics of the well-known Pareto distribution

Denote $X^{(m)} = (\ldots, X_{-1}^m, X_0^m, X_1^m, \ldots)$, where $X_i^m = \frac{1}{m}(X_{im+1} + \ldots + X_{im+m})$ and $\mathbf{E}X^{(m)} = \eta$.

The autocorrelation function of process $X_i^{(m)}$ is

$$R^{(m)}(k) = \mathbf{E}\left\{(X_i^m - \eta)(X_{i+k}^m - \eta)\right\}/VarX_i^{(m)}; k = 0, \pm 1, \pm 2, \ldots. \tag{2.2}$$

A stationary process X with finite mean $\eta = \mathbf{E}X_i < \infty$ and variance $\sigma^2 = VarX_i < \infty$ is called exactly second-order self-similar with parameter $0 < \beta < 1$ if

$$R(k) = \frac{1}{2}\left((k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}\right) \tag{2.3}$$

Parameter β and the Hurst parameter H are related as $H = 1 - \beta/2$, $\frac{1}{2} < H < 1$.

A stationary process X with finite mean $\eta = \mathbf{E}X_i < \infty$ and variance $\sigma^2 = VarX_i < \infty$ is called asymptotically second-order self-similar with parameter $0 < \beta < 1$ if

$$\lim_{m\to\infty} R^{(m)}(k) = \frac{1}{2}\left((k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}\right) \tag{2.4}$$

Self-similarity has been a dominant framework for modeling network traffic [125]. In order to model self-similar traffic, the time intervals between the packet arrivals are modeled by "heavy-tail" distributions. The Pareto distribution is an example of such a distribution. The tails of the corresponding distributions densities tend to zero at sub-exponential rates for large values of their arguments. The cumulative distribution function F(x) has a power tail if there exist positive constants c and a such that $\bar{F}(x) = 1 - F(x)$:

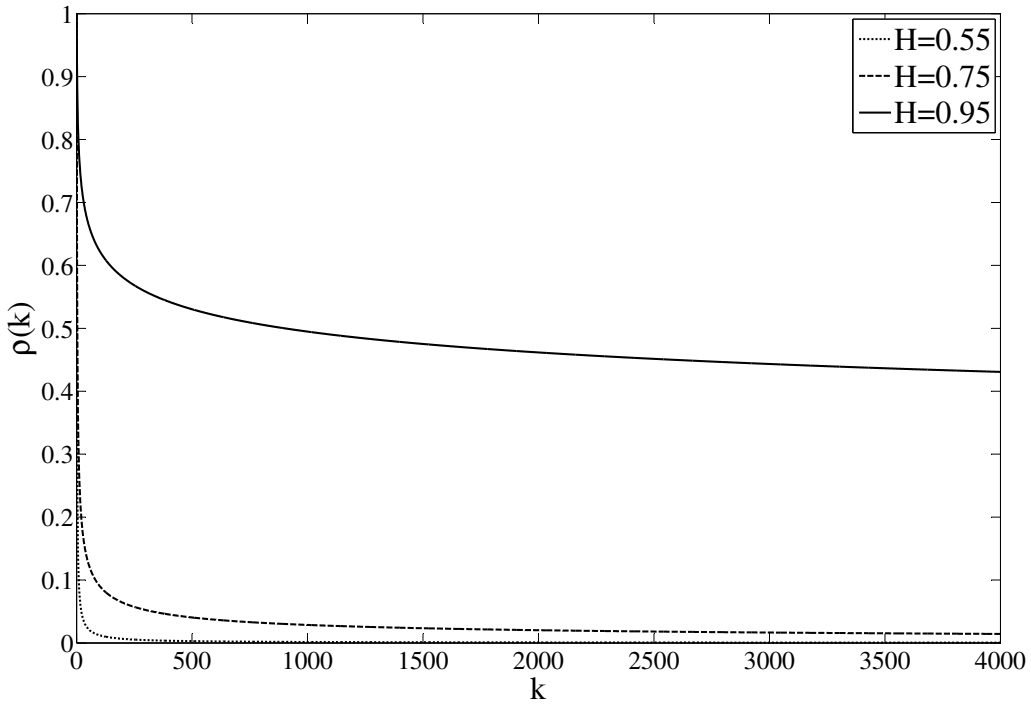
$$\lim_{x\to\infty}\left[x^\alpha \bar{F}(x)\right] = c. \tag{2.5}$$

The frequently used method of self-similarity assessment is connected with the definition of the Hurst parameter, which is closely connected with the shape parameter α of a "heavy-tailed" distribution function. The Hurst parameter for the Pareto distribution, which was recommended in many research papers as a possible candidate for description of Internet data, is connected with the parameter α by the relation $H = (3 - \alpha)/2$ [41].

There are five main versions of the Pareto distribution [53]. The statistical characteristics of well-known Pareto distribution are defined in Table 2.5.

For self-similar traffic analysis, the correlation function and the spectral density are more

important to us. Internet traffic exhibits pretty interesting temporal correlation properties, such as self-similarity and long memory (a slow decay of correlations) on various time scales [159, 122]. In contrast to the classical Poisson model, where $H = 0.5$, these properties stress slowly decaying correlations of the packet inter-arrival times. Such behavior is presented on Fig. 2.2.



2.2. Figure: The autocorrelation function of the self-similar processes with the different H parameter

Numerous investigations show that Internet traffic is characterized by values of the H parameters in the range $0.7 < H < 0.95$. Fig. 2.2. show the autocorrelation function for self-similar traffic modeled by a classical fractal Brownian motion and is given in [99] by the mentioned (2.3).

An interesting feature of self-similar processes is that the autocorrelation function does not degenerate as $m \to \infty$.

The estimation of traffic parameters is performed periodically. The required interval between the estimation can be computed using the information on the correlation interval of the data stream:

$$\tau_k = \frac{1}{2}\int_{-\infty}^{+\infty} |R(\tau)|d\tau = \int_{0}^{+\infty} |R(\tau)|d\tau, \tag{2.6}$$

which contains the corresponding autocorrelation function.

The quantity τ_k indicates approximately the length of time intervals where the correla-

tions of the random process are significant. On the basis of Fig. 2.2. and the correlation interval mentioned above, one can claim that a random process with a higher degree of self-similarity has a higher degree of correlation (long memory) and, hence, a longer correlation interval than processes with low self-similarity.

Analyzing the above-mentioned equation for the correlation interval and Fig. 2.2., we arrive at the problem of defining the correlation interval for self-similar traffic. Obviously, the area of the domain under the autocorrelation function of a random process with a high degree of self-similarity tends to infinity. Therefore, if we define the correlation interval in terms of that area, it would also tend to infinity and would not be useful for parameters estimation. Instead, we are going to define the correlation interval in terms of properties of the spectral density, and then use it for computation of the optimal observation window.

The correlation interval can be defined in terms of the spectral density function. In the simplest case of uniformly distributed traffic with the spectral range ΔF, the correlation coefficient can be expressed as $R(0) = S(0) \cdot \Delta F$, where $R(0)$ is the autocorrelation function for zero shift, and $S(0)$ is the constant value of the spectral density in the whole frequency range ΔF. This is described by the formula (2.7)

$$\tau_k = \frac{1}{2} \frac{S(0)}{R(0)} = \frac{1}{2 \cdot \Delta F}, \tag{2.7}$$

which is reminiscent of the famous Nyquist theorem.

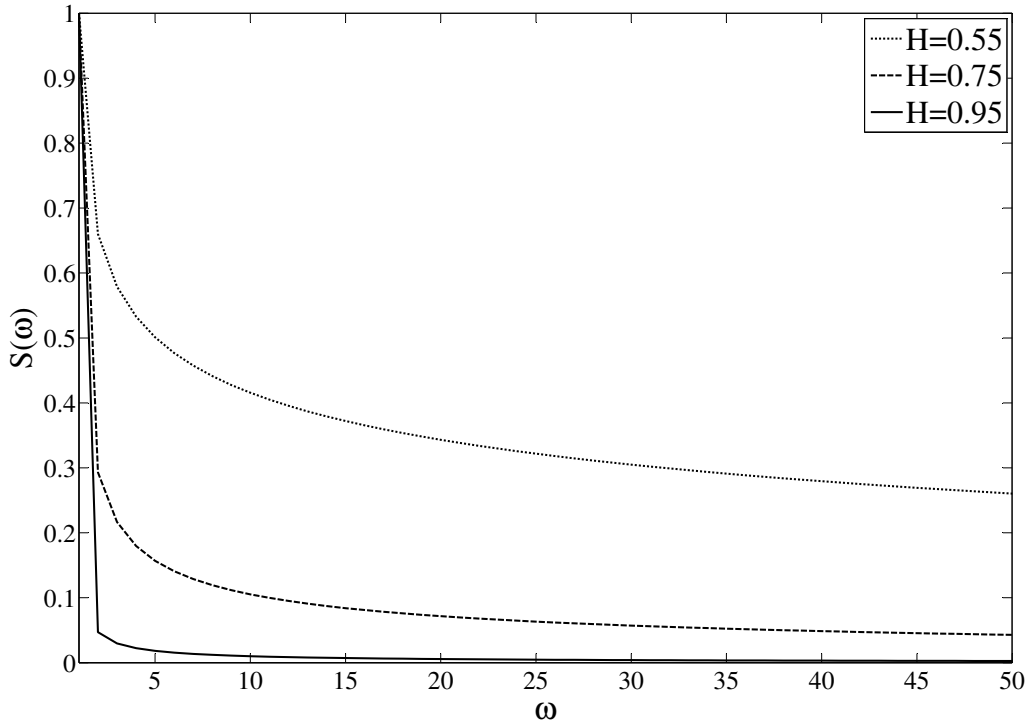
The last equation suggests the following simple scheme for defining the correlation interval. As one can see in Fig. 2.3., $\Delta F \rightarrow 0$ for traffic with a high self-similarity level ($H \rightarrow 1$). However, one can find the upper bound of the frequency range and thus determine the correlation interval τ_k.

2.2.2. Artificial traffic generation models

The traditional traffic source models such as Poisson which superposition does not exhibit self-similarity must be replaced with more accurate models in order to obtain reliable simulation results [41]. This section presents observation of two models for generating data streams with different predefined characteristics, including the self-similar parameter.

The first one is the widely known model of self-similar traffic based on creating independent ON/OFF sources. In order to simulate these sources, we use the OPNET environment. The method uses a superposition of many strictly alternating independent and identically distributed ON/OFF sources. The ON and OFF periods do not necessarily have identical distributions; however, their distributions must have "heavy-tails" as, for instance, the Pareto distribution has.

In the present research the simulation of the self-similar traffic is based on the ON/OFF model that originally was suggested by Mandelbrot [104].



2.3. Figure: The spectral density function of the self-similar processes with the different H parameter

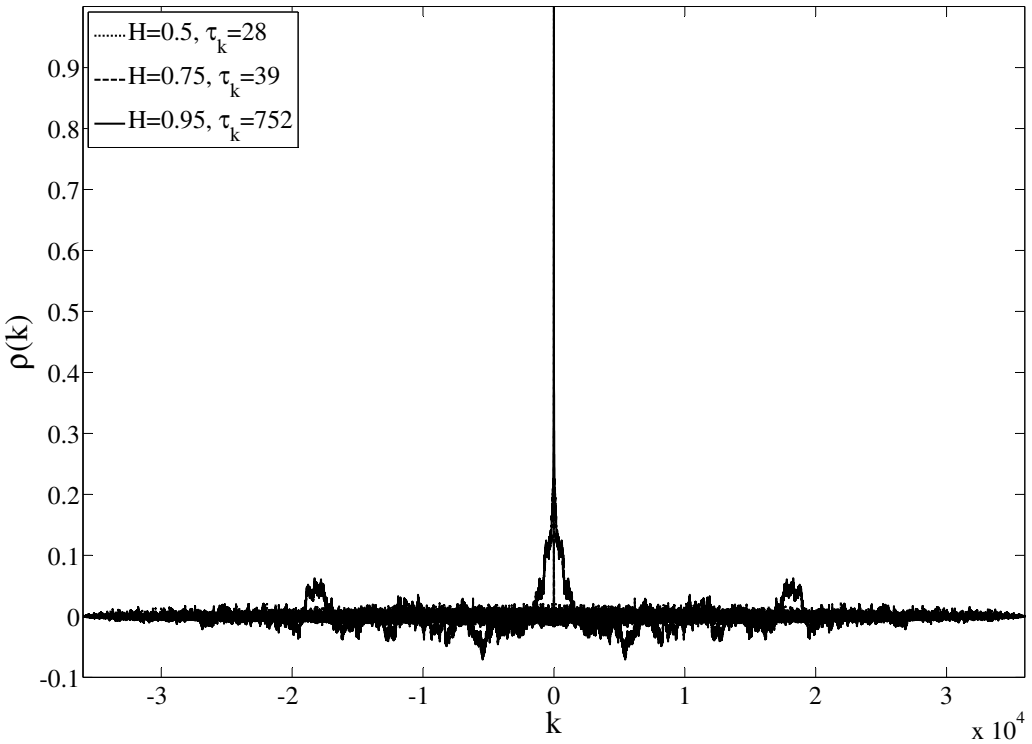
The network measurements show the reason to assume that in real traffic the OFF-period is longer then ON-period. ON-period corresponds to active time period. During that time transmitted packets are separated by small time intervals. It is reasonable to assume that packet sizes within a period remain constant. The ON period is termed packet "train". OFF-period corresponds to silent period, when no packet is transmitted.

The ON/OFF model was chosen for our simulation as it has been shown in the literature that self-similar network traffic can be generated by multiplexing several sources of Pareto-distributed ON- and OFF- periods [124]. Each source sends bursts with random duration distributed by Pareto distribution. The traffic generated by individual sources is independent and identically distributed.

The second model is a simplified ON/OFF model where every batch of packets contains exactly one packet, and the time intervals between the packet arrivals are described by the Pareto distribution, i.e., Fractal Renew Process with Pareto distributed packets arrival time.

2.2.3. Parameter Estimations in a Model with Self-Similar Arrival Streams

Together with the popular Hurst parameter, the autocorrelation functions and the spectral density are the main characteristics used for assessment of traffic parameters. The main attention in the current section is devoted to autocorrelation function.



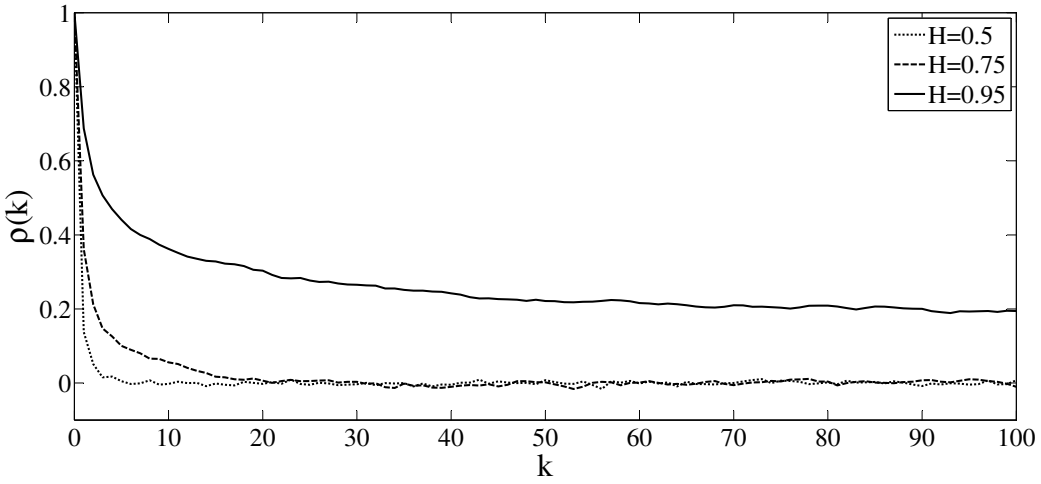
2.4. Figure: The autocorrelation function of the centered Fractal Renewal Process with Pareto distributed packets arrival time. Utilization $\rho = 0.75$

The streams were generated by the following four methods: two different methods of traffic generation (simple and ON/OFF) and two types of Pareto distribution (the two-parameter and the general Pareto distribution).

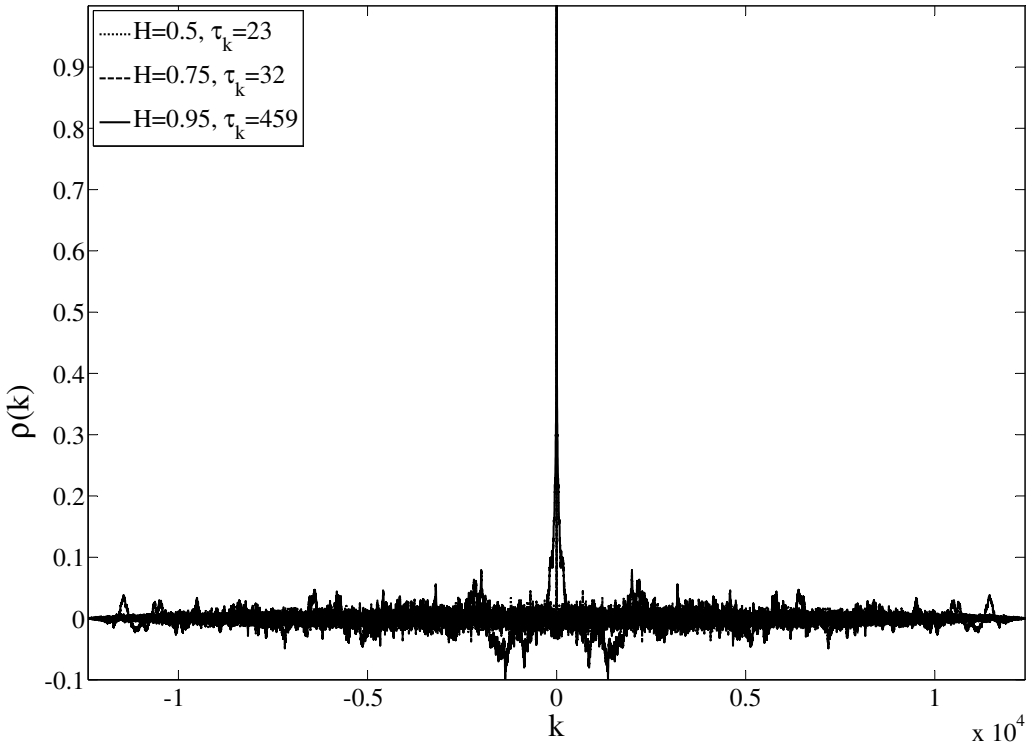
The investigation of the generated streams shows that the traffic where the packet inter-arrival times are distributed according to the Pareto law is not ergodic. The generated streams where packet arrives according to ON/OFF model are not ergodic also. Taking into account the absence of ergodicity in the self-similar traffic, one has to use the autocorrelation function described by Eq. 2.1.

Fig. 2.4. - Fig. 2.7. show the estimated autocorrelation function according Eq. 2.1 and correlation interval according to Eq. 2.6 of the artificial data. Fig. 2.4. and Fig. 2.5. represents autocorrelation function (hole view and short respectively) for the data obtained by using the Fractal Renew Process with two-parameter Pareto distributed packets inter-arrival time generation. Fig. 2.6. - Fig. 2.7. represents autocorrelation function (hole view and short respectively)for the data obtained by using ON/OFF model.

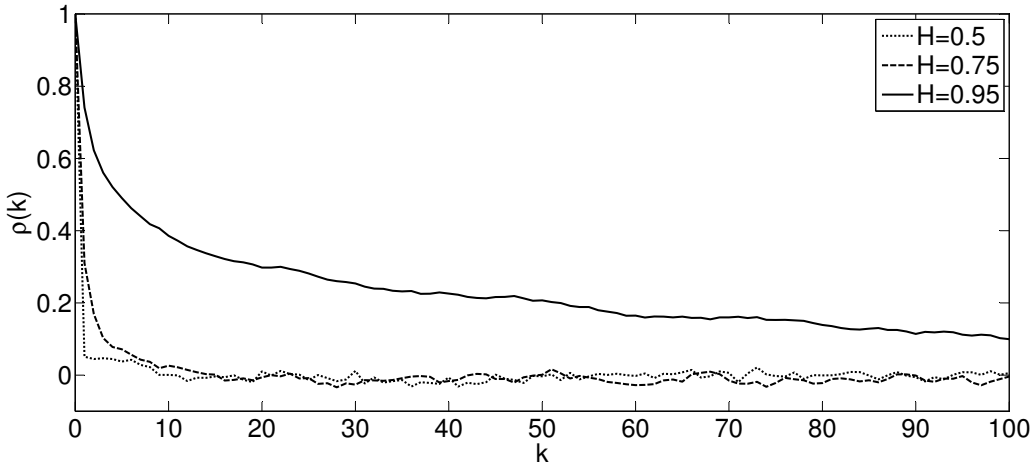
The obtained diagram suggests that the behavior of the autocorrelation functions of the generated stream agrees qualitatively with the theoretical results.



2.5. Figure: The short autocorrelation function of the centered Fractal Renewal Process with Pareto distributed packets arrival time. Utilization $\rho = 0.75$



2.6. Figure: The autocorrelation function of the centered ON/OFF Process with Pareto distributed packets arrival time and duration of ON- and OFF- periods. Utilization $\rho = 0.75$



2.7. Figure: The short autocorrelation function of the centered ON/OFF Process with Pareto distributed packets arrival time and duration of ON- and OFF- periods. Utilization $\rho = 0.75$

2.3. Summary

The chapter considered the network traffic and its parameters evolution. Due to traffic evolution, the models describing the traffic evolve accordingly, from source model to network model. Models development is also described in the current chapter.

The chapter examines the network application parameters that are being used in the dissertation for the results verification using the simulation approach.

The final section reviews the parameters of network traffic created by earlier mentioned applications. As it was shown above, the modern traffic has self-similarity qualities, whose level is characterized by H parameter. The last section emphasizes the basics of self-similarity theory, the models allowing generation of the self-similar traffic with a specified self-similarity parameter. In order to test the artificially generated traffic the autocorrelation function has been employed. That process is also described in the current section. Analysis has proved that using the described methods for the artificial traffic generation allows generating traffic with the specified self-similarity parameter.

As it was presented in "Introduction", traffic with self-similarity characteristics has dramatical influence on the network performance because of underestimated buffer size. Next chapter presents queuing models that take into account self-similarity property of modern traffic.

3. Buffer size evaluation

One of the important structural elements of network nodes is buffer capacity that can keep packets of incoming flows.

Buffer size has to be determined taking into consideration the limits of the packet loss probability of incoming flows in case the bandwidth is not sufficient. The buffer size has to be estimated on the basis of cost rates. The present chapter is dedicated to the issue of choosing the buffer size.

Starting with the work by Norros [115]], there has been mounting evidence that clearly shows that the performance of queuing models with self-similar inputs can be radically different from the performance predicted by traditional traffic models, especially related to Markovian models with short-range dependence. The practical effect of self-similarity presented in [54, 154, 143, 50, 112] and stated that the buffers needed at switches and multiplexers must be bigger than those predicted by traditional queuing analysis.

The adequate size of the buffer maintains Quality of Service (QoS) requirements within limited network capacity for as many users as possible. To get benefits the accurate model for the buffer size estimation should be used. Since classical queuing models do not suit modern packet switched networks the other models have to be used.

The chapter presents $M^X/M/1/K$ and $P/M/1/K$ queuing models that suit modern packet switched networks.

Section 3.1. describes the queuing model $P/M/1/K$. The model represents a numeric algorithm of necessary queue size calculation given the preset parameters of the incoming

traffic, requests' service intensity and the requested guaranteed loss level. The necessity of using the numeric method is clarified by the lack of Laplace transform for the distribution laws with "long-tails" which is necessary for packet loss probability estimation.

Using numeric methods has some restrictions. The methods mentioned need recursive calculations that use a large number of system resources and lots of time. This is not acceptable for the real time systems. For getting the results of $P/M/1/K$ model in real time systems the usage of in advance calculated parameters tables is more appropriate.

Another disadvantage of the model is the lack of expression to calculate such a parameter as average time of packet staying in the queue. Both are the important requirements of quality of service. The queuing model with a bursty incoming of requests gives a possibility to solve that problem. Analysis of $M^X/M/1/K$ model characteristics has showed that it has similar characteristics as the model described earlier. The model parameters can be gained analytically and be used in work to estimate the expected delays level. $M^X/M/1/K$ model description is given in section 3.2.

Effectiveness of buffer size selected on the basis of $P/M/1/K$ model checking is performed by means of $M^X/M/1/K$ model and is described in section 3.3., followed by conclusions.

3.1. Queuing Model - P/M/1/K

The lack of closed-form expression for their Laplace transform for most of heavy-tailed distributions forces the development of numerical techniques to analyze queuing systems with self-similar type of traffic. In the present study the size of the buffer was chosen according to [139], where the analytical expression for the derivative of the Laplace transform of the Pareto PDF is shown. Later it is used with a purpose to calculate the asymptotic packet loss probability.

According to [139], the packet loss probability P_{Loss} of the GI/M/1/K theorem of [35] can be written in closed form as following:

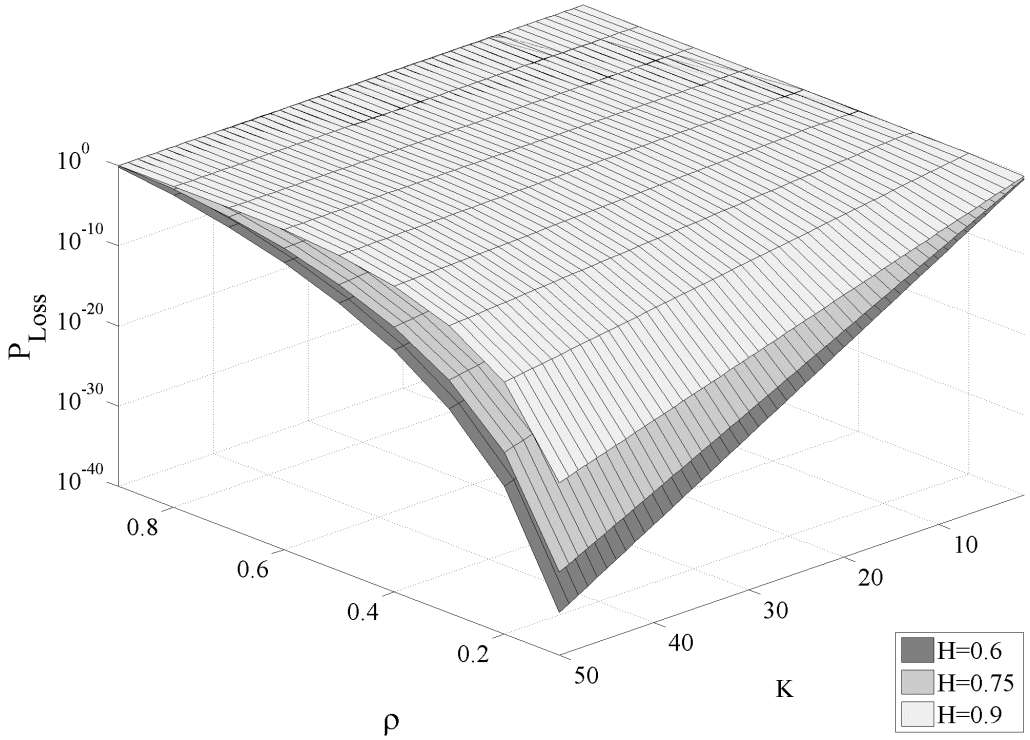
$$P_{Loss} = \left(1 - \frac{\alpha(\alpha-1)}{\rho} M^{\frac{\alpha}{2}-1} e^{\frac{M}{2}}\right) \left[\sqrt{M} W_{-\frac{\alpha+1}{2},-\frac{\alpha}{2}}(M) - W_{-\frac{\alpha}{2},\frac{1-\alpha}{2}}(M)\right] \sigma^K \quad (3.1)$$

where $M = \frac{(\alpha-1)(1-\sigma)}{\rho}$ and $W_{\eta,\xi}(\phi)$ - Whittaker's function.

It is important to keep in mind that this is an asymptotic result and it may not give feasible solutions for small values of the parameters involved.

The Fig. 3.1. presents the relation between utilization, buffer size and packet loss probability for different Hurst parameter of the queuing system evaluated according to [139]. Presented in paper results, as well our simulations show that the closed-form mathematical expressions for the performance measures in $P/M/1/K$ (where P means that the packet

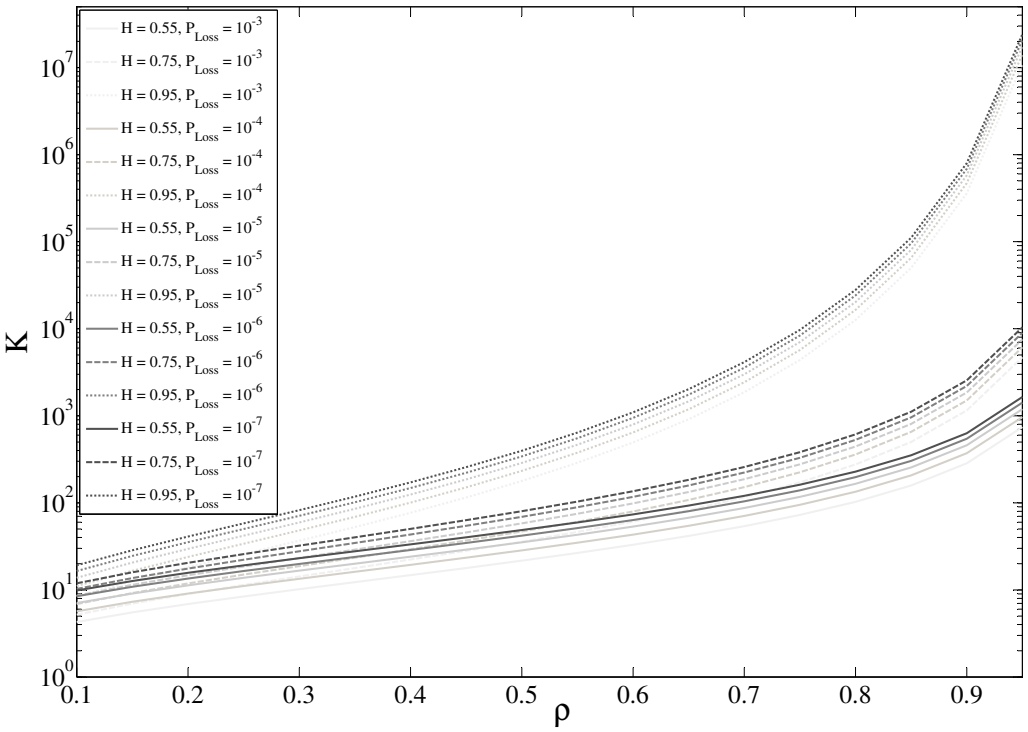
inter-arrival times are distributed according to the Pareto distribution low) queuing system gives appropriate results.



3.1. Figure: The relation between utilization, buffer size and packet loss probability for different Hurst parameter estimated according to [139]

It is commonly considered that a linear increase in buffer sizes will produce nearly exponential decreases in packet loss, and that an increase in buffer size will result in a proportional increase in the effective use of transmission capacity. As it can be seen in the Fig. 3.1. self-similar traffic do not hold these assumptions. The decrease in packet loss with buffer size is far less than expected, and as it can be seen, the buffer requirements begin to explode at lower levels of utilization for higher degrees of long-range dependence (higher values of H).

It is appropriate to use in advance calculated tables of necessary buffer size values in dependence on system load and self-similarity coefficient (H) of incoming flows, as well as preset probability of packet loss to gain result of $P/M/1/K$ model in real time systems. The table sample is given in appendix of the paper (Table A1. - A4.). Figure 3.2. graphically shows the relationship mentioned above. It is seen that the basic parameters that influence the calculated value of the necessary buffer size are the system utilization coefficient and self-similarity coefficient (H). The indicator of probability of the packet loss does not influence the calculated value of the buffer size greatly.



3.2. Figure The buffer value dependence on P_{Loss}

3.2. Queuing Model - $M^X/M/1/K$

As it has been discussed in the previous section, as well as reflected in [124, 171, 57] the analysis of communication system with self-similar incoming packet suggests using sufficiently complicated numeric and imitation analytical methods for evaluation of packet loss probability.

At the same time during information transfer in the system the groups of packets get formed. This observation motivated to search for correspondence between the models of queuing systems with group-type income of packets ($M^X/M/1$ system) and self-similar incoming data flow. It had to be taken into account that real time systems have the limited buffer size.

The author of [20] proposes the analytical model for the incoming flow with group-type (bursty) packet income.

For the analysis of that system it is assumed that it has Poisson incoming group flow, the group packets size is geometrical law distributed, has exponential service time variance law and one server.

Using these conditions, a formula for stationary probabilities p_n for $M^X/M/1$ model is found:

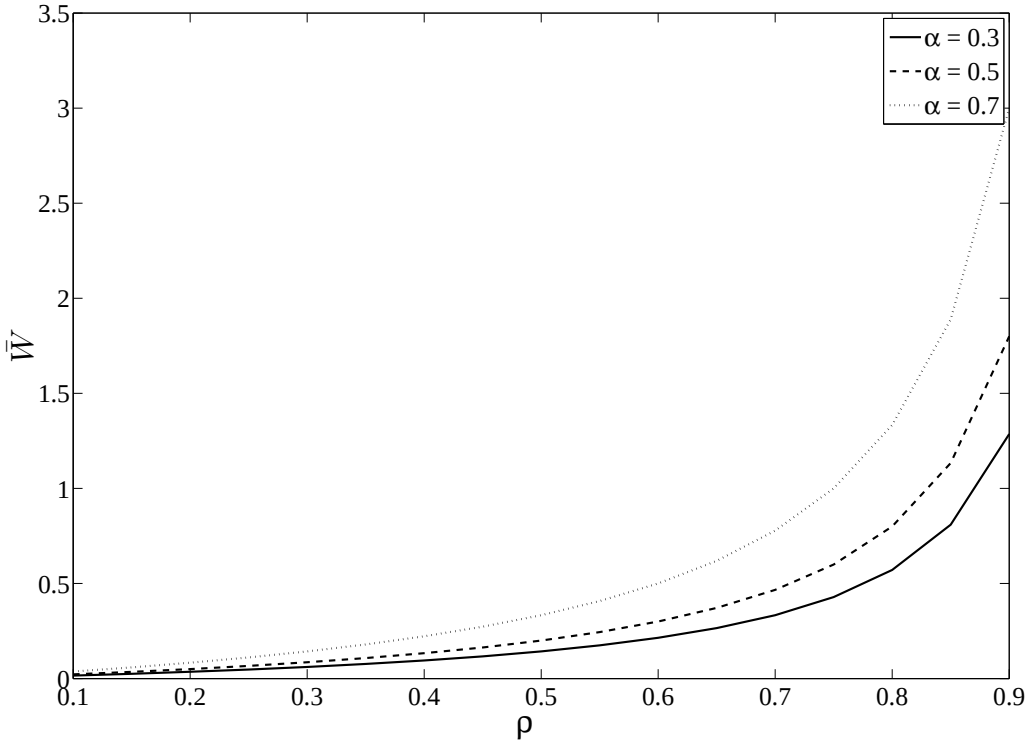
$$p_n = (1-\rho)(\gamma + (1-\gamma)\rho)^{k-1}((1-\gamma)\rho), n \geq 1 \tag{3.2}$$

where ρ is a system utilization coefficient and γ is geometrical variance that characterizes the number of packets in the burst.

Eq. 3.2 results in the expression for the estimation of average number of packets in the system:

$$\bar{K} = \frac{\rho}{(1-\gamma)(1-\rho)} \tag{3.3}$$

Using Little theorem, which can be described by $\bar{W} = \lambda\bar{K}$, the average time of request stay in the system. In Fig. 3.3. the mean time the packet stays within the system for $M^X/M/1$ queuing model is presented.



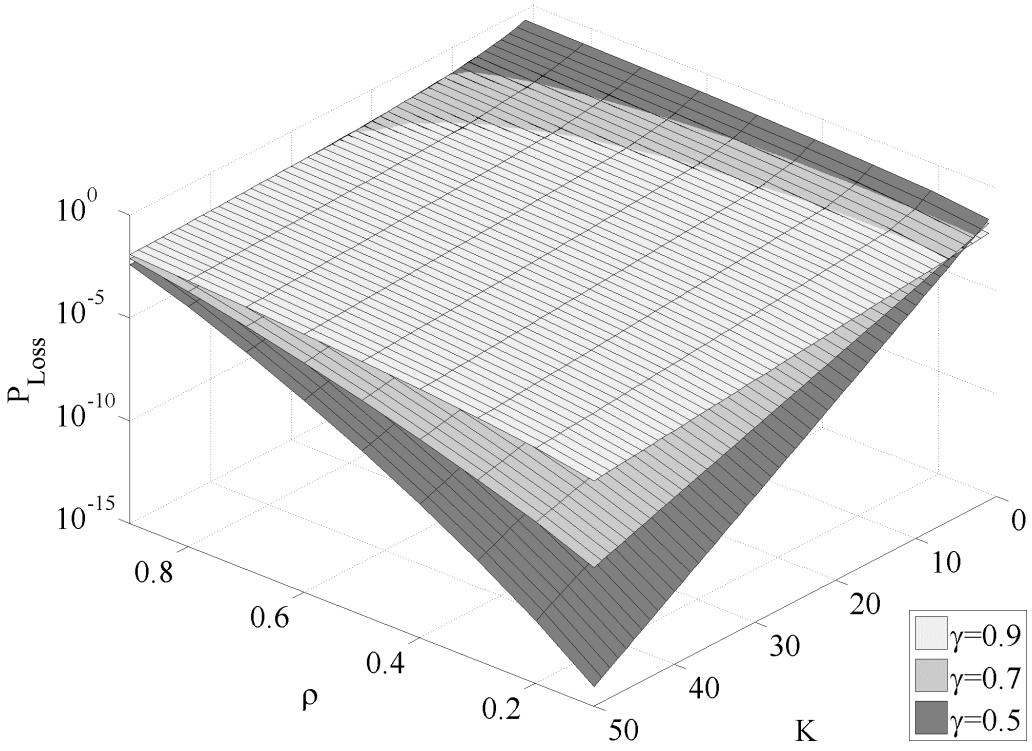
3.3. Figure: The mean time the packet stays within the system according to $M^X/M/1$ queuing model

Taking into consideration the number of system conditions with a limited buffer, it is possible to find the packet loss probability in the queuing $M^X/M/1/K$ system:

$$P_{Loss} = \frac{(1-\rho)(\gamma+(1-\gamma)\rho)^{K-1}((1-\gamma)\rho}{1-\rho(\gamma+(1-\gamma)\rho)^K} \tag{3.4}$$

In Fig. 3.4. the packet loss probability within the system for $M^X/M/1/K$ queuing model is presented.

Time of packet delay in a queue is an important parameter of the system performance. Mean waiting time in $M^X/M/1/K$ system can be evaluated using the previous results [20].



3.4. Figure The packet loss probability in the system $M^X/M/1/K$

For this reason an expression of an average packet number of $M^X/M/1/K$ queuing model has to be found. It can be done using Little theorem [23]. Thus we gain the stay time in the queue till the moment it is served for the queuing $M^X/M/1/K$ model.

In [20] it was found that steady-state probability of packet staying in the system K can be described by Eg. 3.2

An average packet number in the system with limited buffer size can be found as mentioned in [23]:

$$\bar{K} = \sum_{k=0}^{K} k \cdot q_K, \tag{3.5}$$

Where q_K is the normalized probability that K requests stay in the system. For the purpose of the normalized probability calculation it is needed to determine the normalizing polinom Z that can be presented as follows:

$$Z = \sum_{k=0}^{K} \pi_k = \sum_{k=0}^{K} \Big((1-\rho)(\gamma + (1-\gamma)\rho)^{k-1}((1-\gamma)\rho) \Big) = \frac{\rho\left(1-(\rho+\gamma-\rho\gamma)^{K+1}\right)}{\rho+\gamma-\rho\gamma} \tag{3.6}$$

Then the normalized probability that K requests stay in the system for the queuing $M^X/M/1/K$ model can be described by the following expression 3.7:

$$q_k = \frac{\pi_k}{Z} = \frac{(1-\rho)(\gamma+(1-\gamma)\rho)^{k-1}((1-\gamma)\rho)}{\frac{\rho(1-(\rho+\gamma-\rho\gamma)^{K+1})}{\rho+\gamma-\rho\gamma}} \tag{3.7}$$

Bearing in mind the gained results, the average queuing length can be calculated using the following expression (Eq. 3.8):

$$\bar{K} = \frac{K+1}{(\rho+\gamma-\rho\gamma)^{K+1}-1} - \frac{K(1-\rho-\gamma+\rho\gamma)+1}{(1-\rho)(\gamma-1)} \tag{3.8}$$

Fig. 3.5. and B1. - Fig. B4. present the graphs of relationships between the average queue length, system utilization, burstyness and buffer size. The curves disclose an interesting result. All the combinations of group forming and system performance rate indicators have some value of the average queue length that does not depend on the buffer size. With other words, there is a certain buffer size (K_{opt}) that is sufficient for provision of the average queue length. That means that calculating the system resources parameters in communication systems there is no sense to equip the system with the buffer size larger than K_{opt}. Similar approach can be applied whilst making decision about resources allocation in the dynamic system mode.

Combining Little theorem with Eq. 3.8 it is possible to calculate the average packet delay time in the queue for the bursty traffic (Fig. 3.6. and Fig. C1. - Fig. C4.).

The difference between models $M^X/M/1/K$ and $P/M/1/K$ is small in a wide range of the utilization coefficients of the system ρ and the Hurst parameter (H) that can be seen in the following Section 3.3.

3.3. Queuing Models P/M/1/K and M^X/M/1/K Comparision

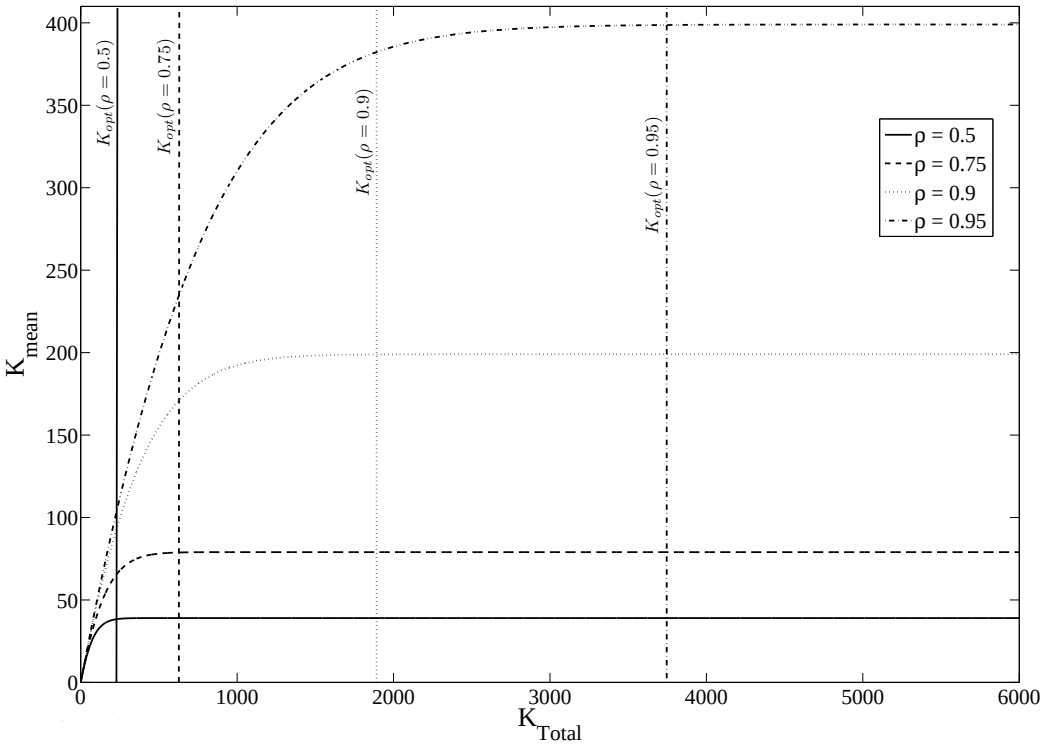
The section shows that the system queues with group-type packets income thoroughly evaluate the features of data transmission system functioning with limited buffer size and self-similar incoming traffic.

It is shown in [20] that the difference between the loss probability and the mean time the packets stay within the system for the models $M^X/M/1/K$ and $P/M/1/K$ is negligible.

In Fig. 3.1. and Fig. 3.4. the packet loss probability in the system $P/M/1/K$ and $M^X/M/1/K$ respectively are presented.

In [21] researched the queuing system with self-similar incoming traffic, $P/M/1$. The results gained allow making the comparison of characteristics of queuing systems $P/M/1$ and $M^X/M/1$, as well as showing the similarity of behavior of these characteristics.

Fig. 3.7. shows the corresponding curves of the mean time the packet stays in the system with self-similar incoming flow ($P/M/1$) and group flow ($M^X/M/1$). It can be observed that the corresponding characteristics of the systems are very similar. Also, the corresponding



3.5. Figure: The mean number of jobs in system for the $M^X/M/1/K$ queue model with $\gamma = 0.95$

characteristics are close to each other when $\rho = 0.1..0.8$.

In the end of the chapter we will consider the queue behavior under the condition of buffer size being chosen in accordance to $P/M/1/K$ model. Average queue length analysis will be presented using the expression gained for $M^X/M/1K$ model.

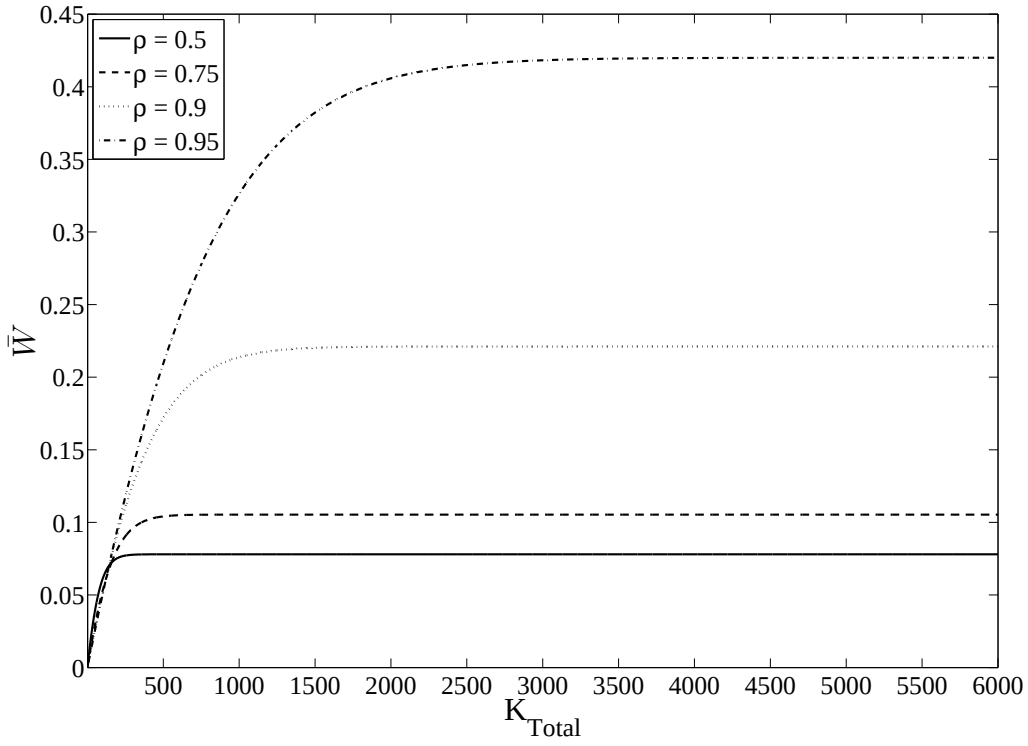
Fig. 3.8. presents the surfaces of allocated buffer size and average queue length. The graphs prove that for the increased values of utilization and self-similarity coefficients the average queue length tends to reach the full volume of buffer size.

Fig 3.9. shows the relationships between the average queue length and overall volume of buffer size. Such a relationship will be called the system robustness. It is seen that the main parameter affecting the system stability is self-similarity coefficient. System load rate almost does not influence the system stability, as well as the requested parameter of QoS, packet loss probability does not.

Evaluating the planes from Fig. 3.9. the following conclusions can be made:

1. The system is more robust at incoming data flows that are characterized by small values of self-similarity coefficients. It means that with the same level of robustness, hence, the same quality parameters guarantees, it possible to provide a higher system load.

2. A maximum system stability value does not exceed 0.6. that means the system with



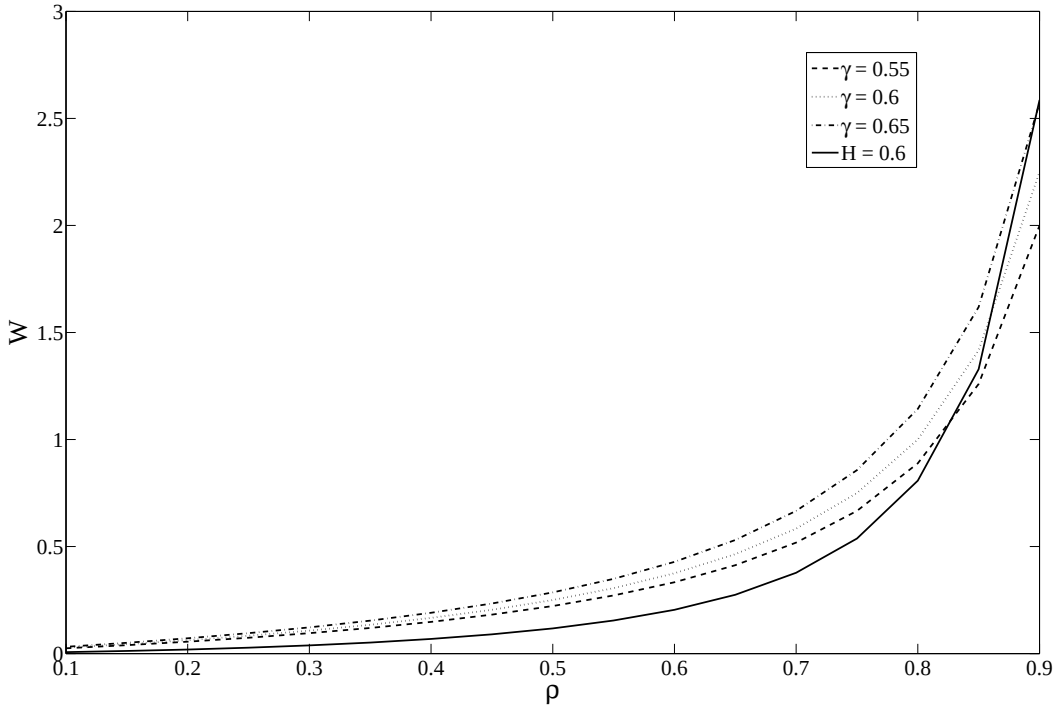
3.6. Figure: The mean waiting time of the job in system for the $M^X/M/1/K$ queue model with $\gamma = 0.95$

the volume of buffer size chosen in accordance to $P/M/1/K$ model is robust to the flows of self-similar character. Therefore, that model can be used for the calculation of parameters of buffer size for the modern telecommunication systems.

3.4. Summary

The present chapter considered the queuing models systems. It has presented the analytical and numeric methods for the calculation of queuing parameters such as the mean queue length, average awaiting time, packet loss probability, buffer size needed for provision of the requested packet loss probability. The main difference of the models from the traditional ones is that they take into account self-similar character for traffic of the incoming data flows. Thus, they can be applied for the calculation of queue parameters for the modern telecommunication systems.

$P/M/1/K$ queuing model estimation is a numeric method. P index shows that the distribution law of the packets inter-arrival time is Pareto. The model allows calculating the needed buffer size using the specified parameters of the incoming flow (system utilization created by the flow with H self-similarity coefficient) and the needed packet loss probability.

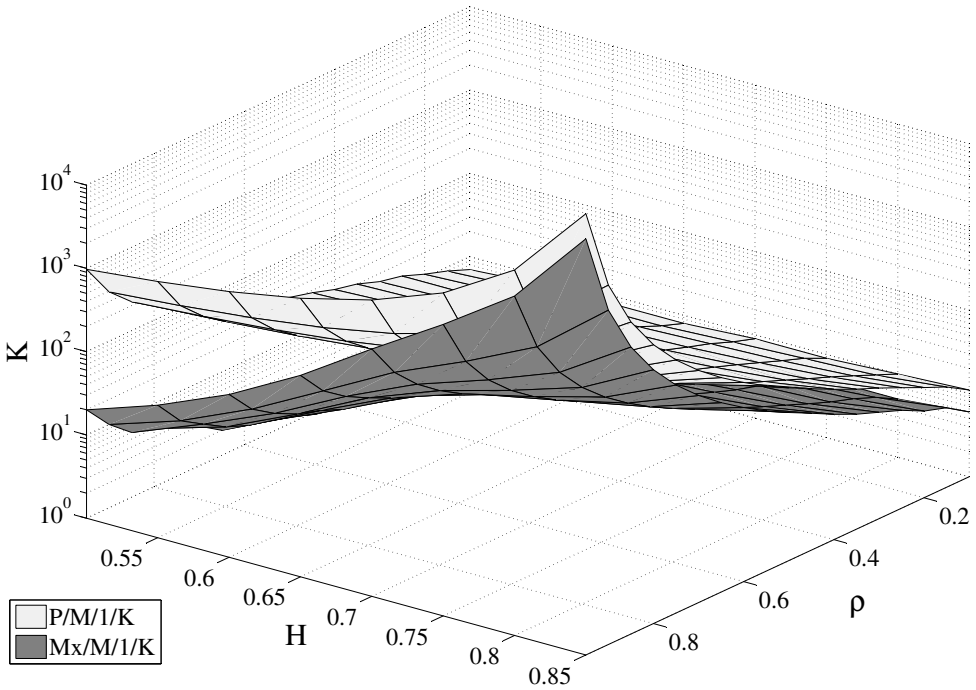


3.7. Figure: The mean time the packet stays within the system for $M^X/M/1$ and $P/M/1$ queuing model

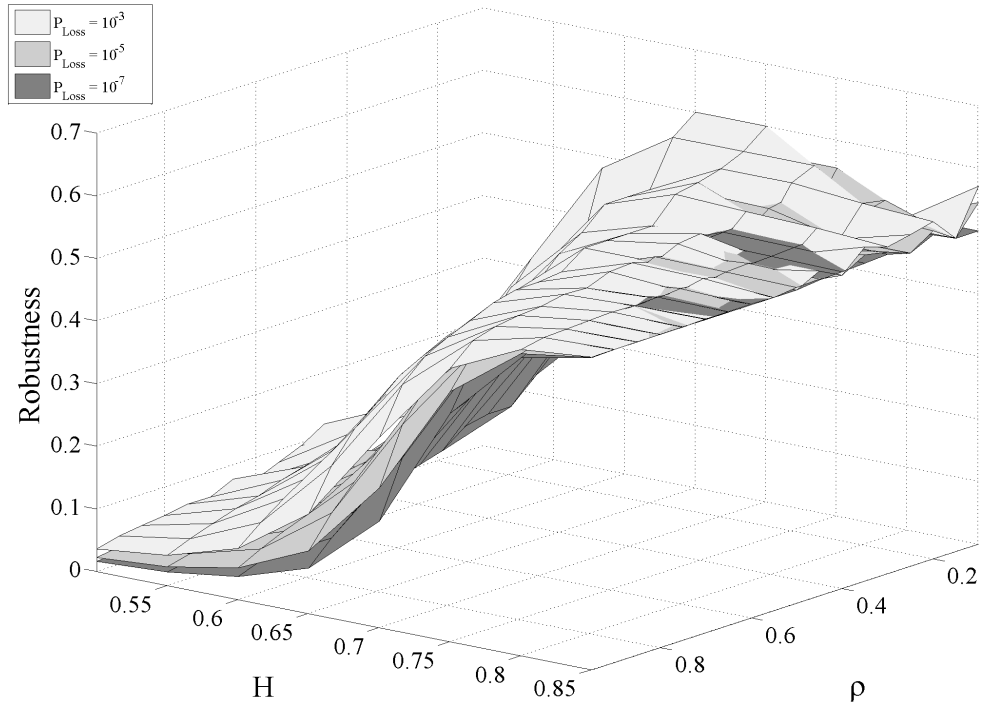
The $P/M/1/K$ queuing model was verified by simulation in OPNET framework.

$M^X/M/1/K$ queuing model represents an analytical model. It describes the queue behavior during the income of the flow with group-type packet formation into the system. $M^X/M/1/K$ queuing model gives a possibility for using analytical expressions for all parameters of the queue.

The chapter ends with the comparison of $P/M/1/K$ and $M^X/M/1/K$ models. The results show the models are similar and supply the system with a high rate of robustness that allows avoiding the overload of network. Next chapter presents a mechanism that can be used for increasing the performance of network.



3.8. Figure: The mean queue size for the $M^X/M/1/K$ model if the queue size is allocated according to $P/M/1/K$ queue model with $P_{Loss} = 10^{-5}$



3.9. Figure: The robustness of the system if the queue size is allocated according to $P/M/1/K$ queue model

4.

Quality of Service in Integrated-Service Networks

It was presented in Section 2.1.1. that packed-switched network traffic does not anymore consist only of data like Web/ftp/e-mail which are distinctive for traditional networks. As multimedia technologies like VoIP and video conferences develop, network traffic characteristics start to resemble more to the traffic of telephone and cable TV networks. Users get accustomed to use services that combine pure data, audio and video information. Requirements to the resources of such applications rise along to the increase of the throughput capability of connection channel. It is fairly predictable that in future the existing throughput size will not be enough to meet all the necessary requirements. Thus, a need for high speed convergence networks will become vital. There are obvious advantages of combination of heterogeneous applications into the same system. But at the same time it causes serious complications in terms of analysis, design and network management. Traditional packet-switched (IP-based) networks use Best Effort (BE) approach for the packet delivery. Such a method is not able to guarantee a required level of packet service [147]. Also, this method is not suitable for convergence networks where heterogeneous applications create different traffic and needs varying parameters of quality of service.

4.1. Intro to Quality of Service

QoS service enables necessary service parameters due to resource allocation [84]. The required QoS parameters are the following [168]:

- Service Availability: The reliability of users' connection to the Internet device
- Delay: The time taken by a packet to travel through the network from one end to another
- Delay Jitter: The variation in the delay encountered by similar packets following the same route through the network
- Throughput: The rate at which packets go through the network
- Packet loss rate: The rate at which packets are dropped, get lost or become corrupted

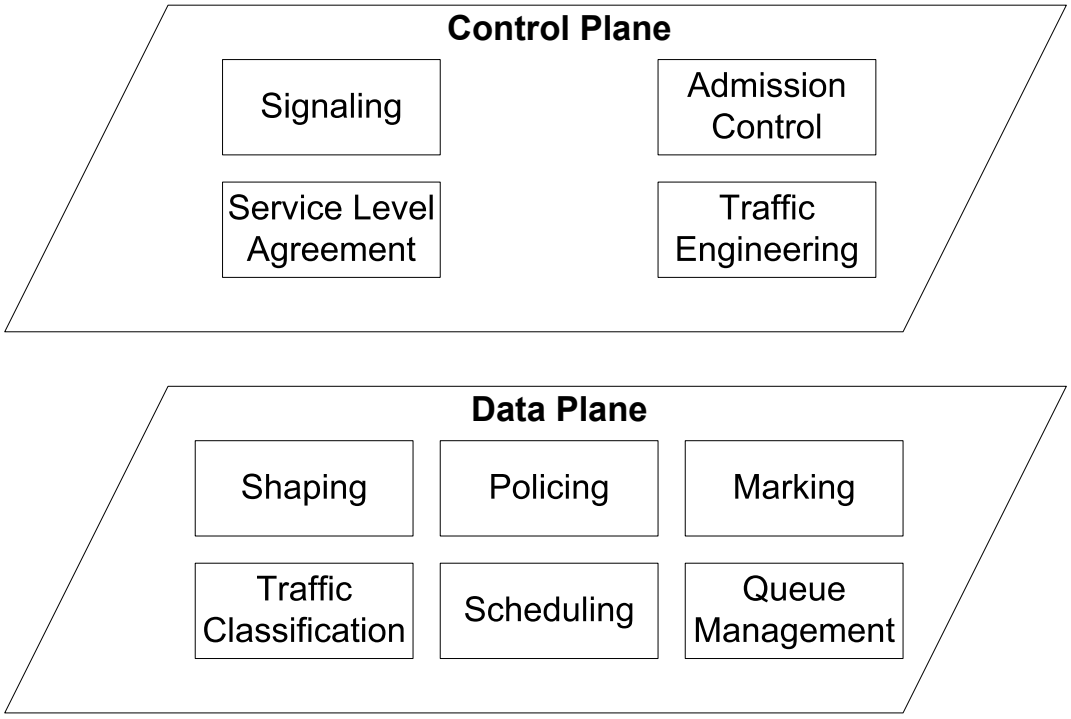
Application groups can be distinguished on the basis of the required parameters. First are the applications immune to delays, but intolerant to packet loss. Such services as ftp, www, e-mail can be considered as belonging to this group. The applications as Internet telephony and video conferencing are real-time by nature, so delay is the essential requirement for them. High delay in the real-time application makes the packet irrelevant and it can be considered as being lost. The applications of voice telephony and video conferencing can stay resistant at small coefficient of packet loss and maintain reasonable quality. At the same time for other critical real-time applications, tele-surgery for example, packet loss is unacceptable. Using advanced compression and coding techniques the applications become more sensitive to packet loss.

Real-time applications can be classified as one way communication (e.g. a video demand), and interactive communication (e.g. video telephony). For the first ones the range of delays jitter, is more important than the level of delays itself. For the second parameter both delays and jitter are important.

Packet loss primarily is connected to buffer overload at switching nodes. There are two basic approaches to cope with this problem. Firstly, it is a repeated transfer in case the packet is lost. Secondly, the allocation of buffer size and provision of priority service for the packets of some applications (Weighted round-robin [133, 164] and weighted fair-queue [43, 61, 153].

The purpose of QoS is the enhancement of service availability and throughput capability at minimal delays and elimination of jitter and losses [148, 101]. QoS for IP networks can be described using two planes which is shown on Fig. 4.1. Global and Local View: control plane and data plane respectively.

Control Plane The functional components of the control plane are described in [83] and [169] and are the following:



4.1. Figure Control and Data Plane of QoS

- Signaling: The information or message exchanged related to the establishment and control of a connection and the management of the network for guaranteed QoS, like RSVP.
- Admission Control (AC): The decision process of whether to accept the new flow of traffic or not according to given network resource and QoS requirement.
- Service Level Agreement (SLA): A service contract between the customer and the service provider that specifies the forwarding service the customer should receive.
- Traffic Engineering (TE):The process of arranging how traffic flows through the network so that congestion caused by uneven network utilization can be avoided. QoS routing belongs to this category.

The components located in the control plane of QoS provide utilization and resource allocation, load balance and the management of the traffic aggregation and path flexibility.

Data Plane In accordance to [83] and [169], the functional components of data plane are the following:

- Shaping: Delay the traffic so that it would conform to the predefined rate.

- Policing: Discard some packets when the incoming traffic violates the predefined rule.
- Marking: Set the Differentiated Services (DS) field in the packet, in which the class of differentiated service is encoded.
- Traffic Classification: Sort the packets based on the content of the header according to predefined rules.
- Scheduling: Send the packets in the order of the certain service rules.
- Queue Management: Control the length of queue by dropping the packets based on the defined rules.

Data plane performs data conditioning (shaping, policing, marking, etc) according to the QoS requirements. The components of data plane also schedule and manage the queue of priority and selected classes.

4.2. QoS Elaboration and Schemes

Further evolution of services that provide quality of service will be regarded.

As it was described above, IP has been developed for the purpose of ensuring best-effort service for delivering of the packet through any network transmission and system platform.

Popularity of real-time applications increased the necessity for secure maintenance of QoS in IP networks. The Internet engineering Task force (IETF) proposes true end-to-end QoS in the form of IntServ [27]. Later, [22] suggested an alternative - DiffServ. The time line of QoS mechanisms in IP networks is presented in Table 4.1.

4.2.1. Integrated Service

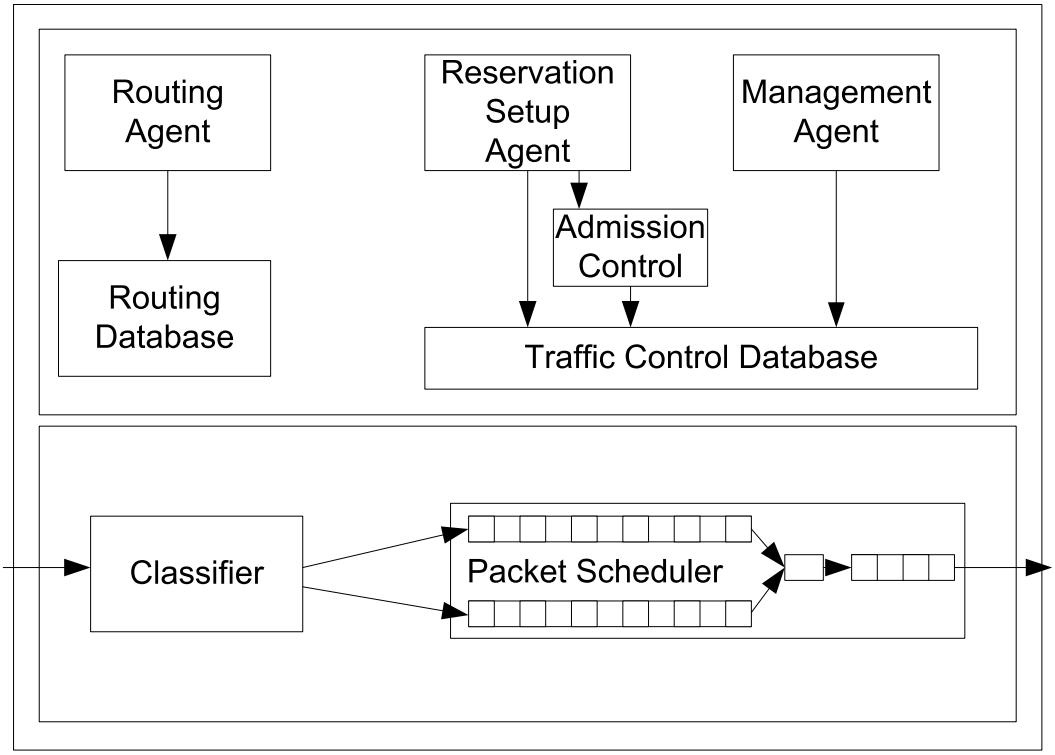
IntServ is orientated for long-lived unicast or multicast flows. The system reserves resources to satisfy QoS requirements for the flows. For signalization and reservation IntServ uses RSVP [170]. For the correct functioning of IntServ it is necessary to provide support for RSVP protocol at all nodes on the way of packet following from the source to the destination. In Figure 4.2. implementation framework proposed by [27] is depicted, showing the way to realize IntServ model.

Such an approach is able to provide strict guaranteed service [148] to ensure the specified borders of end-to-end delays and guaranteed bandwidth. Another options is the controlled load service realization [166] which guarantee better than best-effort and low delays in case of small and medium loads to the network.

IntServ guarantees the fulfillment of the required QoS for each flow under the condition the necessary resources are available. Traffic control [27] is implemented by the components: the packet scheduler, the classifier, and admission control.

Description	**RFC Number**	**Year**
IP Protocol	791 [152]	1981
Type of Service	1349 [6]	1989
IntServ	1633 [27]	1994
RSVP	2205 [170]	1997
Controlled Load Network Element Service	2211 [166]	1997
Guaranteed Service	2212 [148]	1997
DiffServ(DS Field)	2474 [113]	1998
DiffServ Architecture	2475 [22]	1998
Assured Forwarding PHB	2597 [68]	1999
Expedited Forwarding PHB (Revised)	3246 [42]	2002
2 bit DiffServ Architecture	2638 [114]	1999
IntServ over DiffServ	2998 [19]	2000
Aggregation of RSVP for IPv4 and IPv6	3175 [16]	2001
Multiprotocol Label Switching (MPLS)	3031 [140]	2001
Traffic Engineering	3272 [15]	2002

4.1. Table Time line for QoS developments in IP networks



4.2. Figure IntServ Implementation Model

bits	0-2	3-6	7
	IP-Precedence	Type of Service	Must be Zero
	111 Network control	0000 all normal	
	110 Internetwork control	1000 minimize delay	
	101 Critic	0100 maximize throughput	
	100 Flash Override	0010 maximize reliability	
	011 Flash	0001 minimize monetary cost	
	101 Immediate		
	001 Priority		
	000 Routine		

4.2. Table ToS byte as defined in original IPv4

- Packet Scheduler - manages the forwarding of different packet streams using a set of queues.
- Classifier - for the purpose of traffic control (and accounting), each incoming packet must be mapped into some class; all packets in the same class get the same treatment from the packet scheduler.
- Admission Control - Admission control implements the decision algorithm that a router or host uses to determine whether a new flow can be granted the requested QoS without impacting earlier guarantees.

IntServ is unsuitable for large-scale networks due to scalability and heterogeneity concerns [149]. IntServ is not scalable due to RSVP is not able to combine individually reserved session into a single class. An additional disadvantage is that the volume of per-flow states is enormous and all the nodes at the end-to-end path has to support the same reservation protocol.

4.2.2. Differentiated Service

As the realization of IntServ is a rather complicated task developers from IETF has suggested a DiffServ model. In this model the services with the same requirements are combined into one aggregated flow. It receives the necessary level of service in comparison to other flows.

A concept of Type of Service (ToS) priorities has already been included into the definition of IP V4 [152, 26]. In the estimation of IP protocol the lowest 3 bits from ToS byte can be used for packet classification on the edge of the network. The packet can be classified into one of eight categories as it is shown at Table 4.2. In case of congestion the packets of lower priority will be discarded in favor of the higher priority.

Bits	0-5	6-7
	Differentiated Code Point	Currently Unused

PHB	**Class**	**Drop Probability**		
		Low	**Medium**	**High**
Default		**000**000		
AF	1	**001**010	**001**010	**001**010
	2	**010**010	**010**010	**010**010
	3	**011**010	**011**010	**011**010
	4	**100**010	**100**010	**100**010
EF		**101**110		

4.3. Table Differentiated Services Code Point field (DSCP)

Each packet also can be labeled for receiving one of two levels of reliability, throughput and delay. Later, those bits were redefined for provision of additional classes of priorities [6]. Initially, the definite ToS precedence did not secure accurate ”differentiated classes”. Therefore, IETF proposed DiffServ architecture [113, 22]. ToS byte has been redetermined into DS field [22]. Now 6 bit is used for the classification of the packet, see Table 4.3. 3 bits IP precedence has been substituted by 6 bits and called Differentiated Services Code Point (DSCP). In this case using DSCP at the definite node the support up to 64 classes has become possible. Generally DSCP was supposed to be used for marking and a further transfer of the packet on the basis of their Behavior Aggregate (BA).

Later on, the Expedited Forwarding (EF) Per-Hop-Behavior (PHB) is being estimated [42]. It secures low loss, low delays, low jitter and assured bandwidth service. PHB is used for the support of real time applications, for example VoIP, and can be realized with the help of priority queues.

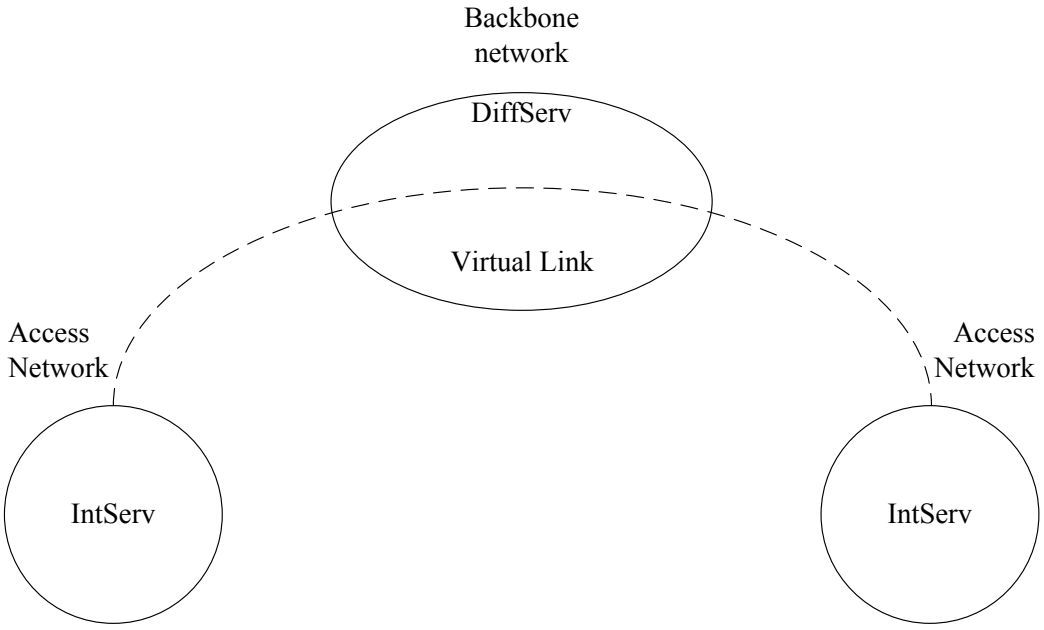
Another important PHB feature is Assured Forwarding (AF) that was described by [68]. AF is a rough equivalent of IntServ that Controlled Load Service. This defines a method by which different forwarding assurances can be given to different BA. Thus, traffic can be divided into three service classes - golden, silver and bronze. For the traffic belonging to golden class a higher service priority get assigned. It also means it has a higher probability of the prompt service than the silver one. The same relations exist between silver and bronze classes.

So that a customer could gain differentiated service he has to have SLA with service provider [169]. SLA shows the supported quantity of classes and the overall traffic volume allowed for a concrete class. Assurance service can be gained on the basis of static SLA. To receive the EF regime, further called as premium service, dynamic SLA is used. A signal protocol like RSVP has been used for the request for service on demand.

4.2.3. Integrated over Differentiated Service

A hybrid architecture has been suggested to escape the disadvantages of IntServ and DiffServ [19]. This architecture, IntServ over DiffServ, offers Scalable admission control methods for IP networks.

As it is shown in Fig. 4.3., IntServ is used in access networks, while DiffServ is used in the backbone network. IntServ access networks are connected through a virtual link provided by DiffServ cloud [106]. The task of DiffServ is the selection of resources of the backbone network to connect the access networks. In its turn, IntServ carry the function of allocation of the DiffServ selected resources to each call to satisfy the resource request. Data packets carry the signaling messages like RSVP $PATH$ and $RESV$ in the DiffServ backbone network.



4.3. Figure Framework for IntServ over DiffServ

The backbone can consist of one or more DiffServ domains. The resources of the aggregated flows transferred between the domains are selected on the basis of SLAs. But, of the resource allocation between the domains does not correctly reflect the characteristics of the aggregated traffic, admission control for each call in the access network may not be consistent with the virtual link congestion state, and individual QoS requirements may not be satisfied.

[16] suggested aggregated RSVP framework. The scalability has been provided by aggregating the stated in the router or employing resource reservations between subnets. But each flow is not completely isolated in the resource allocation since multiple flows share the same service class.

Application Layer	
Transport Layer	Integrated Service/ RSVP Differentiated Services

4.4. Table Relative positions of the different QoS schemes

4.3. Summary

The chapter presents overview of QoS schemes elaboration. IntServ, DifServ, IntServ over DiffServ are described.

In Table 4.4. relative positions of different QoS framework components are presented.

In present work the end-to-end QoS quaranty model will be used. This mode assumes fulfillment of necessary QoS requirements along the entire way from the data source to the destination. In this case it is important to make sure the new data flow does not violate QoS guarantees of the existing flows. CAC function can be used for this purpose. The next chapter describes the process of decision making about admission on the basis of measurements of network load, and does not need *a priori* description of the flows. The access control of this kind is named MBAC.

5. Measurement-Based Admission Control

In the process of quality of service guarantees provision it is important the new flow does not ruin the promised quality guarantee which already exists in the network. There are two different mechanisms to guarantee QoS parameters. The first mechanism is the distinguishing of resources providing service guarantee on the basis of a model and traffic parameters preset by a user or application. The mechanism was called ”static” [65, 85, 90] for the reason of the resources being allocated statically. This mechanism can be easily realized but has certain disadvantages why it has not become wide spread. Firstly, the traffic model which is the base for resource allocation is not able to describe the real network traffic accurately enough. Combination of several flows may create a new flow with unpredictable characteristics. Secondly, the connection parameters may be mentioned incorrectly. Overestimated parameters lead to under-utilization of communication system, while underestimated parameters lead to the overload of the system. Thirdly, the parameters of the traffic that is being transferred cannot be always identified. For example, it is impossible to know the traffic parameters beforehand as it is created by the application on-the-fly.

Those problems connected with the usage of static resource allocation can be eliminated using *Adaptive Bandwidth Control*. For dynamic management the necessary information in the state of system is received by measurement of the parameters of the current traffic. The values of current traffic parameters may be either measured or predicted. Depending on the approach used for traffic parameters determination, two schemes of resource management can be distinguished, closed-loop and open-loop. Immediate observation of the resource pa-

rameters and information submission to the block of resources redistribution are the features of direct management. This approach is orientated for the provision of QoS parameters guarantees such as average queue length [119, 128], loss [136, 72, 97, 150], and delay [86, 38].

The indirect method, the control over the parameters is based on the prediction of the incoming traffic parameters using the previous history. Resources management works by taking into consideration the predicted incoming traffic, for example, the intensity distribution of the exiting traffic to provide the preset parameters of QoS. As there is no correlation between the predicted traffic and needed QoS, it is complicated to provide the requested parameters of QoS. Most of the exciting work for open-loop only attempts to deliver low or zero packet loss rather than guarantee it quantatively [36, 3, 45].

Also, there is a hybrid solution for resources control. It gets applied as the direct and indirect control and allows eliminating the disadvantages of separate approaches [144].

5.1. Model of Measurement-Based Admission Control

It is necessary to fulfill the guarentees service quality for already admitted flows. For this reason it is necessary to have the mechanism that will fulfill CAC [127, 155].

For the last years the problems connected to the static allocation of resources have been solved using the method called MBAC [82, 81, 78, 63].

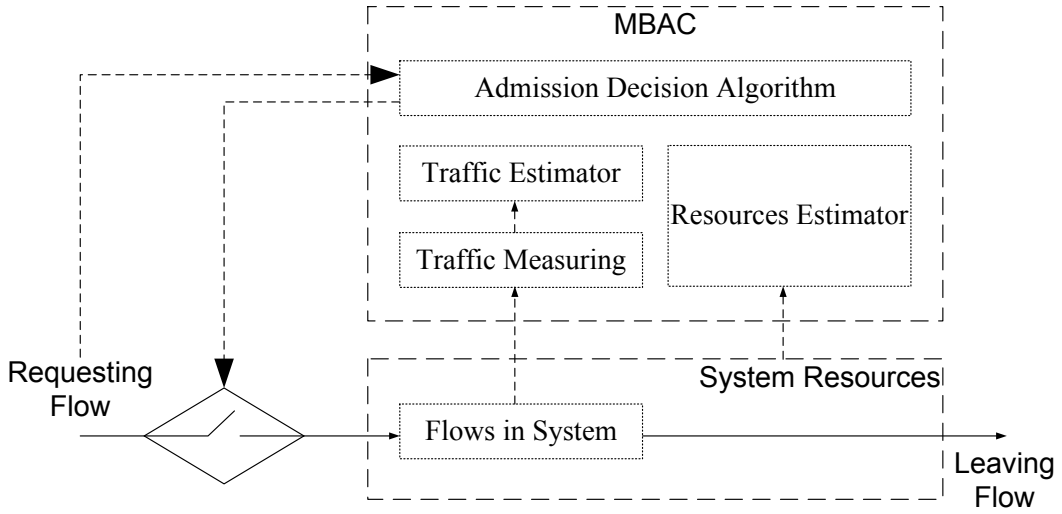
In this case resource allocation for the guarantee of QoS parameters occurs on the basis of combination of the measured traffic parameters (intensity of the received requests, variance and so on) and those requested by user [30]. It allows determining the needed resources more accurately without the usage of the predictions about the traffic model. That gives an opportunity to avoid errors related to the wrong declaration of application parameters.

Some works [85, 55, 59] suggest using access control based on the measurements of equivalent capacity [65] using real-time measurements. In other works [78, 63, 131, 46] authors suggest another approach when network load estimates to perform the admission control.

As Fig. 5.1. shows the MBAC mechanism consists of several parts. The highest level of processes detalization, the MBAC mechanism, can be described by the following processes:

- incoming flow traffic measurements,
- parameters estimator,
- admission decision algorithm - policy.

Not every AC algorithm has an estimator but all AC algorithms must have a policy. The policy of an algorithm is the procedure to follow at flow admission, whereas the role of estimator is to supply information for the use of the policy (procedure.)



5.1. Figure Model of Measurement-Based Admission Control

5.2. Measurements Module

The measurements process has to perform the traffic and available network resources measurements accurately. The most popular measurement mechanisms are the following:

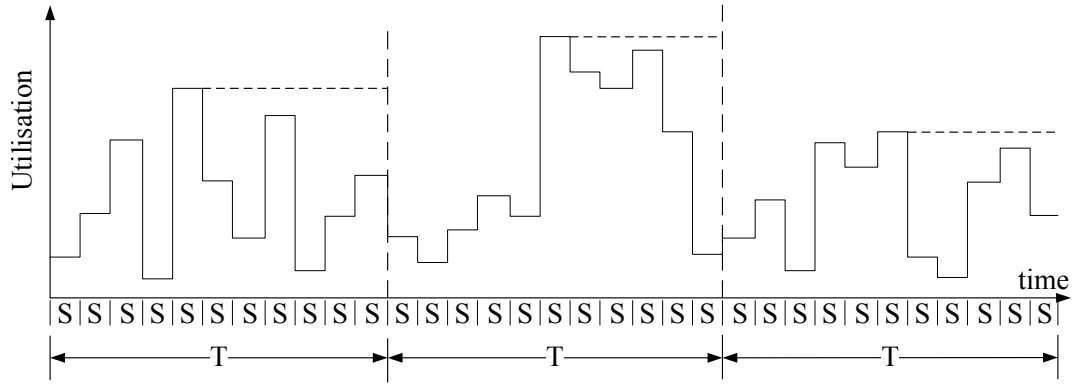
- Time Window
- Point-sampling
- Exponential averaging

Next, the mentioned techniques will be described.

Time Window Time Window scheme delivers measurements in certain time periods [65]. As Fig. 5.2. shows, the system load is assessed on each sampling period which occurs every S units. In the end of the measurement window T the load values or delays calculated during the previous period get renewed.

In work [65] suggested recommendations regarding the tuning of productivity of the admission control algorithm using the measurement parameters: S and T. Low values of S mean a high frequency of sample and lead to higher load averages, as well as results in a more conversive admission decision. And contrary to that the high values of S lower the measured load averages, resulting in the smoother traffic, hence permitting more flows.

T regulates the adaptability of the measurements. The small means frequent renewal of the parameters measurements used for decision making that makes the algorithm more conservative. The large T implies taking into account the affect of traffic burst. In [75] the following proportion $T/S \geq 10$ is recommended.



5.2. Figure Time window measurement of network load

Point Samples In [60] studied the memoryless measurement-based admission control in a decision theoretic framework. The impact of measurement errors on the performance is showed. However, compared to the approach in [158] which also focuses on a bufferless scheme, [60] builds a separation of time scales into their module whereas [158] deals directly with the interplay between the call, burst and cell time-scales. These techniques base their estimates on the current state of the network, without keeping any past history of the network.

Exponential Averaging In [55] another approach to estimate the average arrival rate (or the aggregate bandwidth) for a class of traffic is presented. The arrival rate r_i is measured once every A seconds.

The average arrival rate could then be calculated using an exponential-weight average with a weight $w : avg = (1-w) * avg + w * r_i$. The time constant for this is given in seconds as:

$$t = -1/ln(1-w) * A \tag{5.1}$$

The time constant t reflects the time scale. If t is too long the measurement will remember the flows that have terminated long ago. On the other hand if t is too short, then the potential traffic from the newly admitted connections will not be taken into account. As per [55] $t \geq -1/ln(1-w) * A$ and hence $w \geq 1 - e^{-A/t}$.

5.3. Estimator Module

For the algorithm to be able to provide the necessary QoS guarantees, it has to calculate resource requirements, typically bandwidth. The measurements can be performed on the basis if measurements, predictions or both.

Before discussion of estimators let's introduce the term of "effective bandwidth" whose

Technique		Requirements	
		Measurement	Per-class Declaration
I	Tangent at Peak	per-class measurements numbers of connections per class	peak rate
II	Tangent at Arbitrary Location	per-class measurements numbers of connections per class	peak rate, sustained rate
III	Tangent of Slope One	aggregate measurements numbers of connections per class	peak rate
IV	Tangent at Origin	aggregate measurement	peak rate

5.1. Table Measurement and declaration requirements of Chernoff Bound based estimators

formal definition is given in [85]. [85] may be interpreted to provide an effective bandwidth of any individual traffic source defined as the total bandwidth required to satisfy the QoS constraints of the total multiplexed traffic for a given buffer resource when divided among the number of traffic sources presented in the multiplexed flow.

Instantaneous Utilization (E-IU) The simple Instantaneous Utilization estimator calculates effective bandwidth on the basis of the last measurements. E-IU is a valuable estimator of a template which gives a push to the importance of measurement period while not being sufficiently effective itself. As it has been shown in the previous chapter the period over which a measurement is taken may have a considerable effect upon the computed effective bandwidth estimate.

The same simple estimator represents a starting point for the effectiveness and differences between the policy of AC comparison. If the activity over the interval τ is represented as $X[s, s+\tau]$ then effective bandwidth over that particular period at time is:

$$E = \frac{X[s, s+\tau]}{\tau} \tag{5.2}$$

Chernoff Bounds (E-CB) Chernoff Bounds estimator has been advanced by using Chernoff bounds. This algorithm has been suggested for limitation of tail probabilities of the sum of independent random variables [33]. These techniques may be applied to the curve of effective bandwidth versus mean rate for a traffic source.

Several estimators based on Chernoff Bounds and using bounds different information about the curve of effective bandwidth versus mean rate have been regarded in the work of [59] and are summarized in Table. 5.1. [55] discusses an MBAC algorithm based upon the Hoeffding bound, (of [70]). The approach discussed in [55] is considered in [59] as a specific example of the class of approaches that arise from the Chernoff Bound techniques.

The effective bandwidth requirement of the aggregate of traffic according to approach III is:

$$E = X + \frac{\zeta}{4} \sum_{k=0}^{K-1} p_k^2 n_k \tag{5.3}$$

where

- E - the estimate of the aggregate load
- K - the number of different types of flow
- n_k - the number of individual flows of a particular type
- p_k - the peak-rate for a particular flow-type
- X - the measured aggregate utilization

The scaling factor ζ may be adjusted to tune the algorithm's behavior.

Measured Sum (E-MS) Estimator uses a simpler approach with theoretical argumentation [78]. The estimator assesses the evaluation of the effective bandwidth based in the regular sampling if the measured aggregated load (Fig. 5.2.). [78] and [76] presented a number of functions for converting regular samples (time-window) into a load estimate. Such an approach for the computation of the current estimate is similar to one mentioned in [45].

An effective bandwidth is computed from the aggregate load then is:

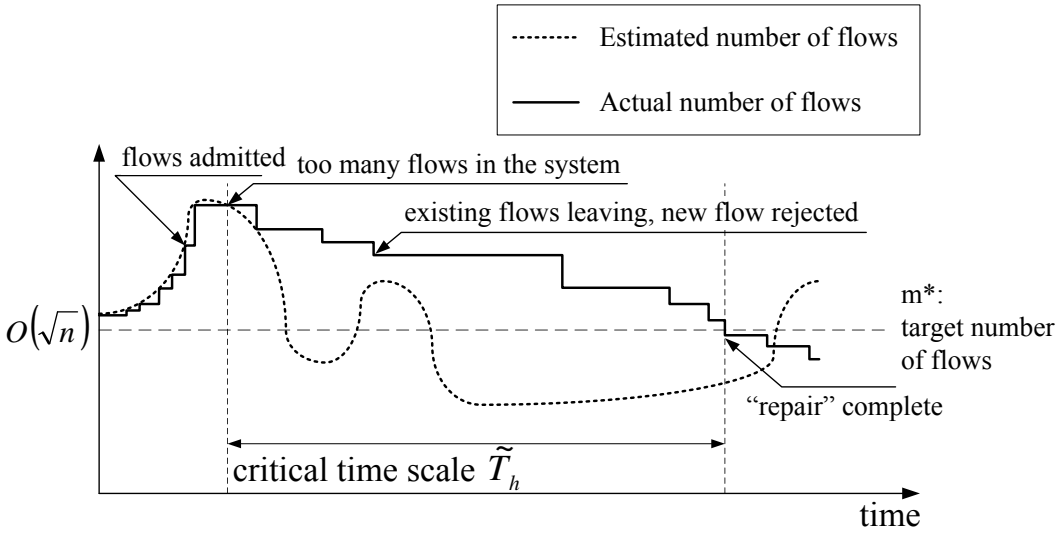
$$E = \frac{\hat{X}}{\mu} \tag{5.4}$$

where

- E - represents the estimate of effective bandwidth
- $\hat{X}$ - is the current aggregate load estimate
- μ - the utilization target that can be used to vary the QoS target in terms of amount of bandwidth to used by the admission control algorithm.

[63] presents an analytical specification for improvement of estimator quality. To contrast the effect of a memory window for measurements, the authors show a memoryless model, as shown in Fig. 5.3. It is shown that flow departures have a repair effect to past mistakes by the MBAC. The fluctuations of the estimated number of admissible flows around the perfect knowledge operating point is on the order of $\sqrt{n}$, where n is the normalized capacity, which means the system size in terms of the mean bandwidth of the flow. Thus it takes on the order of $\sqrt{n}$ flows to rectify past errors in accepting too many flows.

The repair time is of the order of $\tilde{T_h} = \sqrt{n}/(n/T_h)$, where T_h is the mean holding time of the flows and $\tilde{T_h}$ is called the critical time scale of the dynamic system. Thus $\tilde{T_h}$ is the natural time scale to analyze the full dynamics of the system. Fig. 5.4. shows the effect of memory on reducing the variation of the bandwidth estimator. In [63] presented



5.3. Figure $\tilde{T_h}$ is the time-scale for the system to recover from admission errors

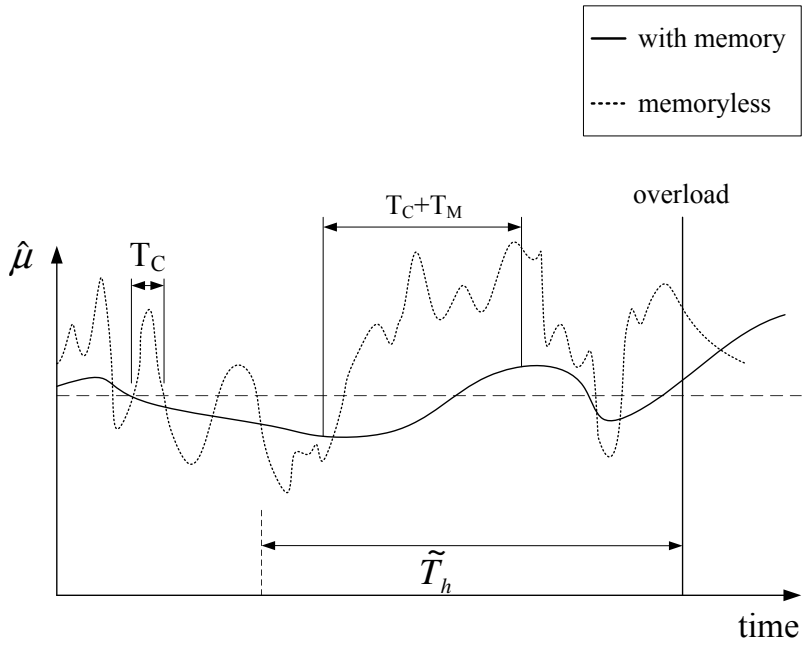
the parameter T_C that governs the exponential drop-off rate while T_M governs how the past bandwidth are weighted. Emphasizing the conclusion made in [63], the same time scale T_M of the order of $\tilde{T_h}$ is the time scale over which effect of past admissions persist and hence this leads itself naturally to a estimation window length. The traffic fluctuations on the time scale longer than the critical time scale fall into a repair regime; these fluctuations should be tracked by the MBAC so that they can be compensated with the flow admission or rejections. Also spare bandwidth should be set aside to absorb fluctuations at a time scale shorter that $\tilde{T_h}$ as these fluctuations are too fast to be compensated for by the repair effect. Thus a robust MBAC should predict the fluctuations statistics over a time scale of $\tilde{T_h}$, rather than estimate the long-term statistics of the traffic.

Per-Flow and Aggregate Measure Estimator (E-MPF and E-MA) Using a theory of large deviations the mechanisms of Measure Per-Flow Estimator and Measure Aggregate Estimator have been gained. These mechanisms vary only with its requests for measurements. Measure Per-Flow Estimator using measurements of each flow, while Measure Aggregate using aggregate measurements to avoid the management and computational overheads.

Each of the mechanisms can provide the requirements for bandwidth directly based on available buffer size and necessary relation to the packet loss.

An estimate of the effective bandwidth becomes the slope of the Scaled-Cumulative Generating Function (SCGF) for a particular set of (traffic) measurements constrained by a set of buffer characteristics (loss-ratio and buffer size). The definition of the SCGF, $\hat{\lambda}(\Theta)$ is:

$$\hat{\lambda}(\Theta) = \frac{1}{\tau} log \frac{1}{T} \sum_{t=1}^{T} e^{\Theta \hat{X}_t} \tag{5.5}$$



5.4. Figure Smoothing of the fluctuation on time scales of $T_C + T_M$

where

- T - is number of periodic measurements $\hat{X}_t$ each taken over the period τ
- $\Theta = \frac{-log\epsilon}{q}$
- ϵ - is the desired loss-ratio
- q - is the size of the buffer

The effective bandwidth is the the rate of change (slope) of this function for any given value of Θ.

[37] and [56] reported that effective bandwidth estimates computed using such asymptotic bounds may be either optimistic or conservative depending on the nature of the arrival streams.

Mean Variance Estimator (E-MV) Mean Variance Estimator is a simple estimator that uses the measurements of the average and variance of one time scale. Effective bandwidth of the measured traffic can be estimated by the following equation:

$$E = \bar{x} + \xi\sigma \tag{5.6}$$

where

- $\bar{x}$ - is the mean of the aggregate utilization

- σ - is the standard deviation of the aggregate utilization

The mean of the flow captures the long-term changes in the traffic, while the variance characterizes the variability of the traffic within the time-scale of the measurements. ξ allows the estimator to accommodate a range of variability in the traffic measurements made.

Apart from the selection of a value of ξ, such an estimator requires appropriate selection of the measurement period and the number of samples used to compute mean and variance of the measurements. Sampling error will arise in the estimation of mean and variance because the samples from which the estimates are computed are in themselves random variables. [45] incorporates a correction to ξ to avoid violation of QoS guarantees:

$$\xi' = max\{\xi, \sqrt{(T+1)(e^{\frac{\xi^2}{T}} - 1)}\} \tag{5.7}$$

Where T is the number of samples. The value of ξ' can then be substituted so that Eq. 5.6 becomes $E = \bar{x} + \xi'\sigma$.

Traffic Envelope (E-TE) Traffic Envelope is another estimator that is based in statistical information. Traffic Envelope Estimator introduces the notion of transport envelope and loss-boundary mechanism that allows effective calculating of bandwidth of transport envelope. The envelope can be described as the mean and deviation of traffic using a multiple time scale. As a result this measurement-based estimator, proposed by Knightly in [91] and further explored in [132] and [130] is able to characterize traffic over a series of time periods.

The proposed approach computes two estimates of effective bandwidth, for the two time-scales: short-term burstiness and long-term variance. The worst-case effective bandwidth estimate as max of them.

Time-scale Decomposition (E-TSD) Another statistics based estimator is Time-scale Decomposition mechanism. It can be differentiated by having no buffer. First introduced by [63] and extended to heterogeneous flow environment in [64]. The mechanisms suggest interesting ideas related to separation of time scale and especially regarding the measures to calculate the average and variance.

Loss-Based Estimator (E-LB) The last measurements based estimator is Loss-Based one. The algorithm is base on the current loss measurement. This mechanism does not offer the measurement of effective requirements of bandwidth but it provides the prediction of current loss relationships. The loss-based estimator presented here makes a prediction of long-term loss based upon measurements over the short-term, although it does not use the marginal distribution approach in [145]. The approach of using loss feedback is not new and has been used in a number of MBAC algorithms, such as that of [146]. Ideas incorporating delay as feedback in addition to loss were proposed in [77] and [157].

Equivalent Capacity (E-EC) The first algorithm in a group of estimators that use only the parameters, mentioned by the flow for the calculation of effective bandwidth, is Equivalent Capacity estimator. In [65] a system based upon fluid-flow approximations of the traffic multiplex is presented. Such approach provided with line speed, buffer capacity and a target loss probability along with *a priori* declarations about the traffic are able to compute an equivalent capacity for each source or multiplex of sources. Such an approach gives a possibility to determine the maximal quantity of flows acceptable to the system while satisfying the requirements for loss constraint.

Exponential Upper-Bounds (E-EUB) A previously described Equivalent Capacity estimator suggests approximation of per-source effective bandwidth. This approach does not take into account an additional benefit when the effective bandwidth estimate incorporates the buffer-space available at the point of multiplexing. [31] has introduced a coefficient to account for the additional gain that may be computed with prior knowledge of the buffer size. This allows for a refined computation of available capacity and thus a refined computation of the effective bandwidth per-source.

Effective Bandwidth Model (E-EBM) [47] offers a scheme of resource redistribution that is modeled by a shared buffer multiplexer fed by ON/OFF processes.

This algorithm is different from the previous ones for implying the periodicity of ON/OFF sources. In Equivalent Capacity and Exponential Upper-Bounds it was thought that the sources are regulated through ”leaky-bucket”. Thus, the source is described by the largest size of the burst, not by its average value.

For the provision of needed loss-ratio the estimator has to operate two parameters: service rate and buffer-size. The approach described by [47] allows reducing the two-resource allocation problem (buffer and bandwidth) to a single-resource allocation problem. It is implemented by the means of loss-probability of a buffer-less multiplexer calculation.

5.4. Policing Module

As it has been mentioned above, Admission Control Policy (Algorithm) is a procedure that is implemented at the income of a new request for the establishment of the connection. Such policies may incorporate the results of previous admissions or admission attempts as part of the flow-regulation process.

Target (P-T) Target - a simple admission policy that uses no estimator. It allows a nominated number of flows of a particular type into the multiplex. This policy allows construction of an AC similar to the algorithm used by [30]] to derive the performance values for MBAC

AC	Key Idea	Policy	Estimator	Descriptors
AC-PRA	Declared Peak	P-PA	-	peak rate
AC-ST	Simple Threshold	P-TO	E-IU	-
AC-AR	Acceptance Region	P-AR	E-IU	peak rate, mean rate
AC-CB	Chernoff Bounds	P-BP	E-CB	peak rate
AC-MS	Measured Sum	P-PA	E-MS	peak rate
AC-MPF	Large-Deviation Theory	P-PA	E-MPF	peak rate
AC-MA	Large-Deviation Theory	P-PA	E-MA	peak rate
AC-MV	Mean-Variance Estimator	P-TO	E-MV	-
AC-TSD	Time-scale Decomposition	P-TO	E-TSD	-
AC-TE	Traffic Envelope	P-TO	E-TE	peak rate
AC-LB	Loss-ratio	P-TO	E-LB	-
AC-EC	Equivalent Capacity	P-T	E-EC	peak rate, mean rate, mean burst size
AC-EUB	Exponential Upper Bounds	P-T	E-EUB	peak rate, mean rate, mean burst size
AC-EBM	Effective Bandwidth Model	P-T	E-EBM	peak rate, mean rate, max. burst size
AC-T	Target	P-T	-	-

5.2. Table Admission Control algorithms as combinations of policy and estimator

behavior.

Threshold Only (P-TO) Threshold Only policy will allow new admissions if a current utilization value is below a defined threshold.

Back-off Period (P-BP) Back-off Period policy is based on works described in [17, 18]. In this policy if a flow is rejected, the algorithm does not admit another flow of the rejected type until an existing flow of that same type leaves the system.

Pessimistic Admission (P-PA) Under this policy, new flow requests are (pessimistically) assumed to be transmitting at that traffic's worst-case transmission rate. Such policy uses assumption about flow peak-rate contribution until the nearest measurement-based estimation. Subsequently the measurement-based estimated values are used instead the peak-rate.

Policy of AC-AR (P-AR) The admission decision is made based upon whether the current utilization plus the peak rate declared by the new flow is less than or equal to the line capacity. There is no memory exists from admission to admission, so the success or failure of a previous flow admission will have no impact upon the behavior of the policy for future admissions.

5.5. Admission Control Module

The present chapter includes the description of AC algorithms. Table 5.2. presents a list of the algorithms together with the description of the main idea used by the estimator, as well as admission policy used for its realization. The last column with the name "*Descriptor*" the per-flow-admission requirements are described.

While only the P-PA policy explicitly requires peak-rate, AC-TE and AC-CB also require the peak-rate of flows as part of the admission process. AC-AR requires *a priori* knowledge of the mean and peak rate traffic descriptors in order to compute the admission surface. The estimators E-EC and E-EUB each require descriptions of the Markovian characteristics of the traffic: the peak rate, sustained rate and mean burst size. In contrast E-EMW requires the parameters used to describe a traffic regulator: peak rate, sustained rate and maximum burst size.

AC-PRA - Peak-rate Allocation This algorithm is implemented as a useful comparison point: based upon the peak-rate declarations of flow-attempts, it will admit flows if the declared peak-rate of the new attempt can be admitted in the current allocation. Results gained using this mechanism can be considered as the lower utilization-bound of any AC algorithm - achieving the best preservation of QoS through worst-case assumptions of the traffic flows.

This algorithm can be considered a special case of the leaky-bucket based characterization mandated for ATM in [14] and proposed for the Internet IntServ [27, 166].

AC-ST - Simple Threshold Simple Threshold is the simplest possible MBAC algorithm. It is based on thresholding policy where the user sets an arbitrary admission threshold and combined with the simplest Instantaneous Utilization.

AC-AR - Acceptance Region The approach is taken from [60] and [87]. And driven by a measurement of current line utilization. The acceptance region is computed to maximize line utilization for a nominated packet loss, given a set of flows with a known declaration of peak and mean rates. One aspect of the approach proposed by [60, 87] is that the system is robust to mis-specification of the mean and peak rates as well as being robust to the errors in measurement.

AC-CB - Chernoff Bounds Proposed in [59] admission control algorithm uses "*Tangent at Slope One*" Chernoff Bounds estimator in conjunction with back-off period policy. For the admission decision, X represents the current aggregate-load measurement, ζ is the algorithm control parameter, p_k is the peak-rate of class k and n_k is the number of flows present in class k. This expression is compared against the line capacity C. If the estimate is less than or equal to the line capacity, the new flow is admitted. The formula itself is given as:

$$X + \frac{\zeta}{4} \sum_{k=0}^{K-1} p_k^2 n_k \leq C \tag{5.8}$$

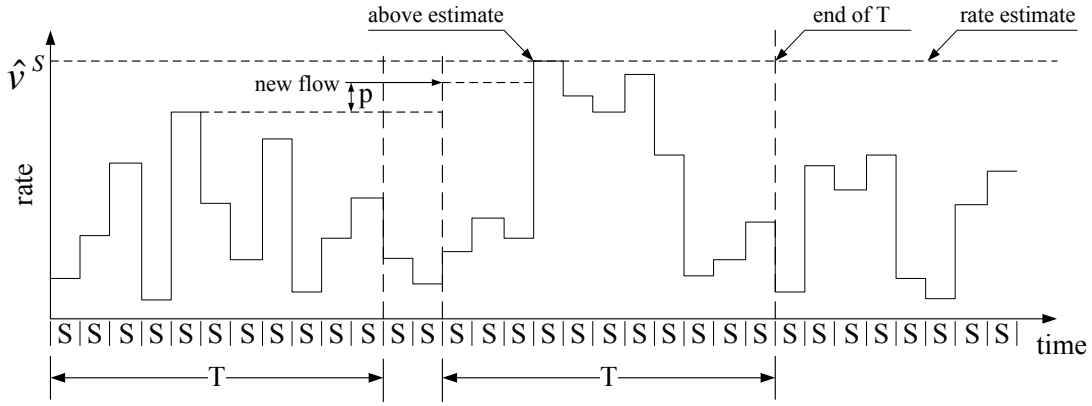
AC-MS - Measured Sum The algorithm of admission control is based on simple equation, Eq. 5.9. Given that $\hat{X}$ is the measured load of existing traffic, p is the peak-rate of the

new flow, C is the link capacity and υ is the user-defined utilization control, the admission decision based on the following:

$$\hat{X} < \upsilon C - p \tag{5.9}$$

Measured Sum MBAC algorithm algorithm uses a modified version of the load-estimator; recall that a procedure the algorithm authors call time-window is used to capture the maximum value during a characterization period.

Fig. 5.5. illustrated how the time-window mechanism worked. A new flow causes an instant reset of computation period T. At the same time working utilization value by the peak-rate of the new flow increases.



5.5. Figure Impact of new flow on utilization measurement

AC-MPF - Measure Per-Flow For the Measure per-flow MBAC algorithm the combination of the E-MPF and P-PA. E is estimated with MPF algorithm, p is the peak-rate declaration of the flow attempt and C is the total line capacity. The admission decision can be based upon:

$$E + p \leq C \tag{5.10}$$

In this equation the P-PA policy is implemented by increasing the comparison value by the peak-rate of the new flow. Once the new flow is admitted, the estimate, E, is also artificially increased by p until an up-to-date version of the current estimate is available.

AC-MA - Measure Aggregate Measure aggregate-based MBAC algorithm is similar to per-flow-based MBAC algorithm. The combination of the E-MA and P-PA uses an admission decision based upon Eq. 5.10.

AC-MV - Mean-Variance For the admission algorithm the E-MV estimator is combined with the simple P-TO, threshold-only policy.

AC-TE - Traffic Envelope The AC implementation of using the traffic envelop estimator. This worst-case description of the new flow, is computed from the peak rate of the new flow and the minimum measurement interval of the envelope. The traffic envelope approach separates the admission process into two parts. Firstly for the short time-scale events consider the buffer dynamics of the multiplexer and ensures the buffer resource is not over committed. The second part of this algorithm is the long time-scale, and a comparison is made between the mean resource requirements of the new flow and current flows. Interested readers are referred to [132] for details.

AC-TSD - Time-scale Decomposition The Time-scale Decomposition estimator detailed in [64], provides an estimate of the variance and mean of a single (theoretical) flow from the current aggregates.

AC-LB -Loss-Based Upon the current loss-estimate from the loss-based estimator E-LBE a simple admission algorithm can be constructed consisting of a test against a user-specified level of loss.

$$\epsilon \leq \hat{\epsilon} \tag{5.11}$$

AC-EC - Equivalent Capacity With the defining the configuration (buffer space, line speed) parameters, the QoS provision the user desires (the target loss-ratio), and needing the parameters describing the traffic in the multiplex, the maximum number of flows can be pre-computed. The estimator proposed in [65] an be converted into an admission algorithm. The AC process becomes a simple admission equation

$$n + 1 \leq N_{max} \tag{5.12}$$

where

- n - is the current number of flows in progress
- N - max is the pre-computed maximum.

AC-EUB - Exponential Upper-Bounds For the determination of the maximum number of the flows allowed, Exponential Upper-Bounds Estimator, uses the buffer-size and line capacity, along with *a priori* characterization of traffic and a user specified loss-ratio. This

pre-computed value for the maximum number of flows may then be used in an admission process based upon Eq. 5.12.

AC-EBM - Effective Bandwidth Model Effective Bandwidth Model specified values of buffer-size and line capacity, along with *a priori* characterization of traffic and a user specified loss-ratio, the maximum number of flows that may be admitted can be pre-computed. The estimate of the number of flows is computed off-line for the desired loss-ratio using the estimation procedure described in [47]. The maximum number of flows may then be used in an admission algorithm described by Eq. 5.12.

AC-T -Target Policy "Target" refers to a simple admission algorithm that will allow a nominated total number of flows in the multiplex. The Target AC allows the user to nominate the maximum number of flows to be admitted.

5.6. Summary

The model of MBAC is presented in the chapter. It is shown, that MBAC consists of three modules which are: measurements, estimator and policy module. Description of each module is presented in respective section. The main contribution of the chapter is presenting a number of different MBAC algorithms, noting the fundamental premise upon which each are based. A number of the MBAC algorithms have their basis in the solution or approximation of the Chernoff Bounds, while others approach the estimation problem from different theoretical backgrounds such as large-deviation theory or statistical analysis.

Four techniques are presented, each estimating the bound based upon different information about the curve of effective bandwidth versus mean rate.

Measured sum algorithm computes an estimate of effective bandwidth based upon regular sampling of measured aggregate loads and uses much simpler approach with little theoretical foundation combining a local-maximum prediction with a control over the level of line utilization.

Two presented techniques are based on large deviation theory. Both operate through the estimation of bandwidth requirements based directly upon available buffer-space and desired packet loss-ratio. In the chapter a family of algorithms based upon statistical information derived directly from the measurement of line utilization are introduced. An algorithm that for effective bandwidth computation uses traffic envelope is also presented. MBAC algorithms that use measured loss-ratio and given number of flows as an admission criteria are finalizing the section.

Within the presented algorithm some approaches does not take into account the effect of buffering, while others take into account gains made through buffering. For the last group of

algorithm the combined rate at which data enters the buffer may exceed the buffer service-rate for small periods before packet-loss will occur. That is why, the burst rate and burst duration of sources play an important contribution in the computation of the effective bandwidth of sources.

Next chapter presents evaluation of MBAC parameters and presents suggestions for efficiency improving.

6. MBAC parameters evaluation

The previous chapter was dedicated to the review of QoS parameters guarantee mechanisms. The most widely used in modern circumstances is the mechanism of access control based on measurements which has been described in details. The existing methods have been thoroughly discussed and possible reason to mistake occurring were mentioned. The present chapter shows the feasible solutions for prevention of possible errors.

Section 6.1. outlines recommendations for choosing the parameters of telecommunication system resources at the design stage. It presents the most optimal choice of parameters for the resources like buffer size and the output bandwidth if the total intensity of the arrival data traffic is defined. The optimal ratio of these two resources provides the necessary probability of packets loss at minimal costs.

Section 6.2. is focused on the issues that emerge during the process of MBAC mechanism functioning. The effectiveness of MBAC strongly depends on the accuracy of measurements. The accuracy is dependent on the estimation windows length (observation period) and sampling period. An excessively large observation period may result in underestimated values of measurements. At the same time the small observation period may cause the over-sized values. Thus, the sampling period affects the accuracy of calculated parameters.

Section 6.2.1. suggests the new adaptive approach to measurements. That method considers the connection of observation period and sampling period with the character of the value being measured which could be the intensity of the input data flow and its statistical parameters.

At the moment request for the introduction of a new connection is received, the access control has to determine the potential for this new connection support with QoS parameters guarantee. Section 6.2.2. presents the algorithm of access control which secures a high load of communication system, simultaneously providing the required probability of packet loss. The later is the significant parameter of Quality of Service.

The shortage of resources may lead to the situation when there is no possibility to provide the support to all incoming flows. In case the requests are of different priority classes, logically, the higher class is being served at first. Section 6.2.3. shows the algorithm of the effective management of the same priority class flows.

Section 6.2.4. contains recommendations regarding the dynamic redistribution of resources. It offers a substantial reduction of expenses during the communication system meeting the requirements of QoS. Conclusions and suggestions are given in the end.

6.1. The Design of MBAC Management System

Recently, the basic rule for choosing the buffer size [163] has been scrutinized [9, 134, 165]. The rule says the buffer size has too chosen as being able to incorporate the full Bandwidth-delay product (BDP) of data. [135, 48, 44] and [8] mention that realization of such large buffers for 40Gb/s channels is a complicated task for electronic routers and it is difficult to apply them for optical router of future. Moreover, the existing arguments for the rule [163] are no valid anymore due to the changes of the channels and its traffic size, as well as memory cost and bandwidth ratio. The following section presents the process of choosing the optimal buffer size and bandwidth of the outgoing channel maintaining the minimal costs and the specified QoS parameters at the stage of system design.

General statement of the problem and description of the optimization methods are described in the following section.

6.1.1. Buffer Size and Output Bandwidth Optimization in a MBAC System

We consider a packet switched system using the example of a data hub (switch) with i inputs and one output. We use λ_i to denote the intensity of the packet flow arriving to the hub through the i-th input ($i = 1, \bar{M}$).

The hub has a buffer of size K^* and the part of it of the size K_0 already occupied. Hence, $K^* - K_0 = K$ locations are involved in the decision-making process on allocating the sought buffer size for the arrival traffic.

The maximal bandwidth of the output is μ^*. The existing traffic flows have already taken part of the bandwidth of the size μ_0. Thus, the remaining part to be controlled is $\mu^* - \mu_0 = \mu$.

The arrival traffic can have optimal values of the buffer size K_{opt} and the bandwidth μ_{opt} allocated for it, for which the packet loss probability P_{Loss} does not exceed the given level.

To find these optimal values of the buffer size K_{opt} and the bandwidth μ_{opt}, we minimize the cost functional

$$C = c_1\mu + c_2K, \tag{6.1}$$

where c_1 is the cost of the unit bandwidth, and c_2 is the cost of the unit buffer.

Minimizing the cost functional, we need to meet the restriction

$$1 - P_{Loss} \geq 1 - \epsilon \tag{6.2}$$

where P_{Loss} is the packet loss probability, and ϵ is the upper admissible value of the loss probability.

Naturally, the resources allocated for the arrival traffic should not exceed the available ones, i.e.,

$$\mu_0 + \mu_{opt} \leq \mu^* \tag{6.3}$$

$$K_0 + K_{opt} \leq K^* \tag{6.4}$$

To find the optimal values μ_{opt} and K_{opt}, we form the Lagrangian

$$L = c_(\mu_0+\mu)+c_1(K_0+K)+\delta[(1-P_{Loss}-(1-\epsilon)]+\lambda(\mu_0+\mu-\mu^*)+\beta(K_0+K-K^*), \tag{6.5}$$

where δ, λ and β are the undetermined Lagrange multipliers.

We find the derivatives of the Lagrangian with respect to several variables, set them equal to zero, and solve the system of equations

$$\begin{cases} \dfrac{\partial L}{\partial \mu} = 0 \\ \dfrac{\partial L}{\partial K} = 0 \end{cases} \tag{6.6}$$

Obviously, it would be difficult to find the solution, which is caused by three undetermined Lagrange multipliers δ, λ and β and the lack of an analytical expression for P_{Loss}.

To solve the problem we use the fact that the self-similar traffic is of a group nature, which allows obtaining an analytical expression for the packet loss probability. In Chapter 3. was shown that the difference between models $M^X/M/1/K$ and $P/M/1/K$ is small in a wide range of the utilization coefficients of the system ρ and the Hurst parameter (H).

To simplify the solving procedure, we remove the resource restrictions as well. Again, to solve this problem, we introduce the cost function $C = c_1\mu + c_2K$, where c_1 is the cost of the unit intensity of the packet processing (or the unit output bandwidth), and c_2 is the cost of the unit buffer.

The packet loss probability is

$$P_{Loss} = \Phi(\lambda, \mu, K). \tag{6.7}$$

Then, to optimize the system's parameter, we can differentiate and zero out Lagrangian (6.8):

$$L = c_1\mu + c_2K - \delta[(1 - P_{Loss}) - (1 - P^*_{Loss})], \tag{6.8}$$

where δ is the undetermined Lagrange multiplier, and $(1 - P^*_{Loss})$ is the given restriction on the probability of the loss-free packet transmission.

To perform differentiation, we use the analytical expression for (6.7) gained in Chapter 3. Substituting Eq. 3.4 into the Lagrangian and differentiating and zeroing out derivatives Eq. 6.6, we obtain the system of equations

$$\begin{cases} c_1 + \delta\dfrac{\partial\Phi}{\partial\mu} = 0 \\ c_2 + \delta\dfrac{\partial\Phi}{\partial K} = 0 \end{cases} \tag{6.9}$$

This yields the optimality condition for the allocation of the bandwidth μ and the buffer size K:

$$\frac{\partial\Phi}{\partial\mu}\frac{1}{c_1} = \frac{\partial\Phi}{\partial K}\frac{1}{c_2} = -\frac{1}{\delta}. \tag{6.10}$$

This leads to the rule of the optimal bandwidth and buffer size allocation: "the decrease in the loss probability P_{Loss} per unit of cost should be the same for both types of resources viz. the bandwidth μ and the buffer size K".

We find the derivatives of the packet loss probability $P_{Loss} = \Phi(\lambda, \mu, K)$ with respect to μ and K:

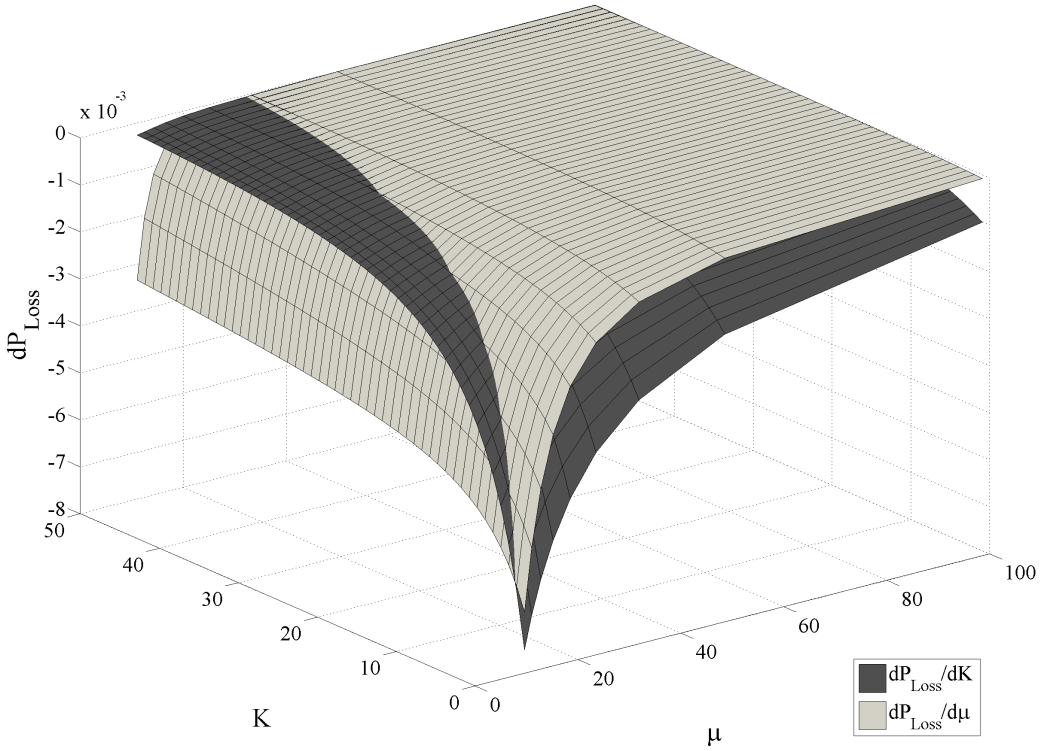
$$\frac{\partial\Phi}{\partial\mu} = \frac{\lambda(\gamma-1)\left(\frac{\lambda+\mu\alpha-\gamma\lambda}{\mu}\right)^K}{\left[\mu - \lambda\left(\frac{\lambda+\mu\alpha-\gamma\lambda}{\mu}\right)^K\right]^2} \times \\ \times \frac{\left[\lambda^2\left(\frac{\lambda+\mu\alpha-\gamma\lambda}{\mu}\right)^K - K\lambda^2 + \mu^2\alpha + \gamma\lambda^2 - \lambda^2 - 2\mu\alpha\lambda + K\alpha\lambda^2 + K\mu\lambda - K\mu\alpha\lambda\right]}{(\lambda + \mu\alpha - \gamma\lambda)^2} \tag{6.11}$$

and

$$\frac{\partial\Phi}{\partial K} = \frac{\lambda ln\left(\frac{\lambda+\mu\alpha-\gamma\lambda}{\mu}\right)(\mu-\lambda)(\gamma-1)\left(\frac{\lambda+\mu\alpha-\gamma\lambda}{\mu}\right)^{K-1}}{\left[\mu - \lambda\left(\frac{\lambda+\mu\alpha-\gamma\lambda}{\mu}\right)^K\right]^2} \tag{6.12}$$

We can construct two surfaces $\frac{\partial\Phi}{\partial\mu}\frac{1}{c_1}$ and $\frac{\partial\Phi}{\partial K}\frac{1}{c_2}$ in the three-dimensional space depending

on μ and K. The intersection of these two surfaces is presented in Fig. 6.1. and yields the optimal solution that corresponds to Eq. 6.10



6.1. Figure: $\frac{\partial \Phi}{\partial \mu}\frac{1}{c_1}$ and $\frac{\partial \Phi}{\partial K}\frac{1}{c_2}$ surfaces for $\rho = (0.1..0.9)$ represented by the inter-arrival rate $\lambda = 10$ and the coefficient of the geometric distribution $\gamma = 0.9$

Based on analytical expressions (Eq. 6.11) and (Eq. 6.12) is possible to quickly obtain the optimal solution for the real-time control of the network node parameters

The cases where one can apply the results of solving the optimization problem for the parameters of the communication system:

Optimization Solution Application In the course of designing the system, until there are no restrictions on the resources, we can choose the optimal value of the buffer size K^* and the optimal output bandwidth μ^* if the total intensity Λ of the arrival data traffic is known. To find the solution, we can analyze the graphs in Fig. 6.18. - Fig. 6.21., which show the intersection of the curves of the derivatives of P_{Loss} with respect to K and μ. Then, we determine the value that corresponds to the obtained K^* and μ^* on the graph in Fig. 3.4. having in mind that $\rho = \lambda/\mu$. If the values of the parameters do not ensure the given loss probability P_{Loss} we search for the next set of parameters for the increased value of the arrival traffic intensity $\Lambda > \Lambda_1$. The search process is convergent.

6.2. Control on Intellectual MBAC Management System

The optimal proportion of buffer size and bandwidth is strongly influenced by the traffic character, it's group forming quality in particular. For the effective system resources management one has to accurately determine traffic parameters in real time. Previously, various aspects of measurement process have been considered and a new model suggested. Also, the flows management and resources redistribution algorithms have been described.

6.2.1. Integral Measurement Process of Incoming Traffic

The real-time applications are both sensitive to delay and loss. As a result, quality of service guarantees have attracted a lot of research interest in the past decade. Dynamic management of system parameters like bandwidth and buffer capacity allows to gain essential performance benefits while quality of service guarantees are met. For the recent years a method of admission control based on measurements have received a wide recognition.

Measurements accuracy depends on the correct definition of sample and observation periods. [7] and [39] propose using three sampling methods - systematic, random and stratified. As the network traffic is a-periodical, aforementioned sampling unadaptive techniques frequency usually is based on the average network load [34], traffic distribution and the value causing a low overload rate [102]. In the case of actual traffic differing from the expected one, the resulting measurements can be inaccurate or having too many samples. In order to measure the traffic load, [69] suggests an adaptive sampling method. It has several disadvantages including calculative overhead. In the work of [78] a question about accurate observation period measurement is raised.

The present section shows several solutions to minimize possible errors of flows parameters estimation which is critical for making the correct decision about a new flow allowance.

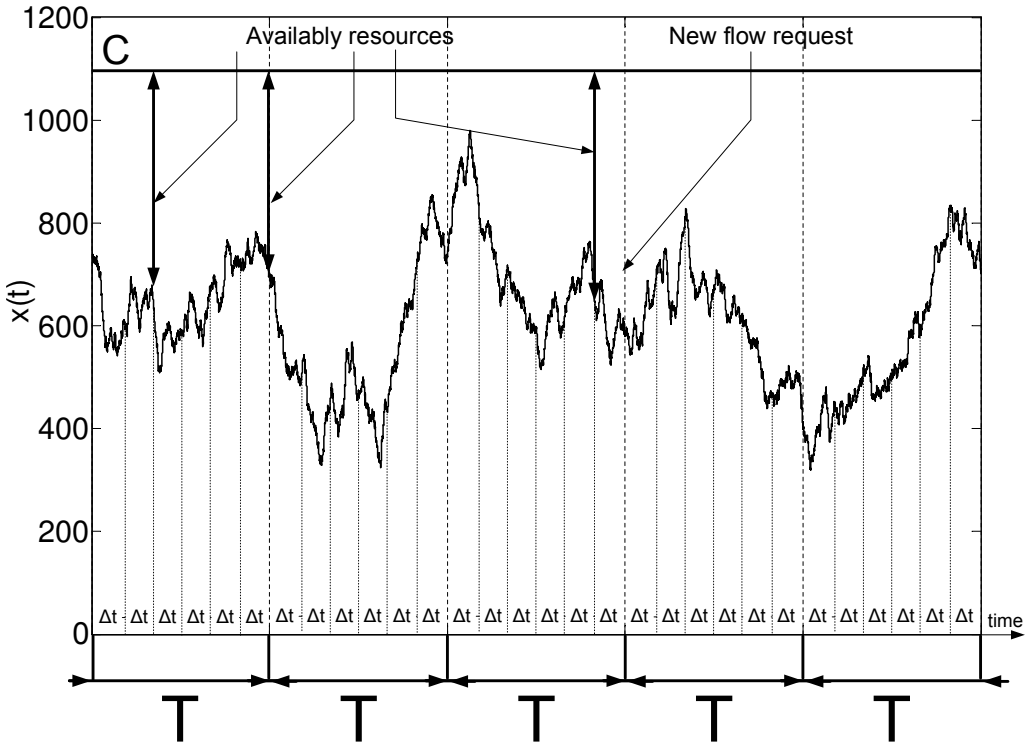
Firstly, recommendations for determination of observation period T are presented. It cannot be chosen too long as the character of current traffic may change suddenly and the system of this traffic will not be able to spot that hop. And it cannot be chosen too low as it will not provide credible results.

Second, calculation methods of such a significant MBAC parameter as sampling period are presented. If the parameter is too high, uncertainty in flow parameters evaluation will increase, as well as it will be impossible to observe each incoming packet because of lack of time spent for processing [141, 39].

The intensity of incoming packet flow and self-similarity parameter of incoming traffic need to be estimated so that calculation of outgoing channel load and buffer capacity is possible. These operations have to be done for further decision about allowance of the flow to the system with limited resources, e.g. bandwidth and buffer capacity.

The incoming flow is analyzed by accumulating data about it during the periods of time

T. Using these periods the system shows the results of incoming flow analysis which are further used to take a decision about the new flow access to the system. If these time periods (T) are fixed, the graph of incoming flow changes and measurements looks as presented in Fig. 6.2.



6.2. Figure The graph of incoming flow changes and measurements with the fixed periods.

The measurements of the flow parameters are taken during the discrete time periods Δt. The flow data during the time period Δt can be denoted as $x(i)$, where i is Δt period. For example, $x(2)$ is data of the aggregated number of packets in the second period Δt_i. For the outgoing channel the bandwidth capacity C is known. Thus, using the example given above, the remained resources are $C - \frac{X(2)}{\Delta t} = c_2$. The level the incoming flow influences the system corresponds to the utilization of channel ρ. In the case used before the utilization coefficient is the following:

$$\rho(2) = \frac{X(2)}{\Delta t}/c \tag{6.13}$$

It has to be mentioned that besides the last time period the system accumulates data about any parameter during a longer time period too. Thus, there is a probability this fact can be used.

The target parameter is intensity of incoming flow λ_i at period i. The intensity assigns the number of packets $x(i) = \lambda_i \Delta t$ that enter the system on i period within Δt time period. The flow intensity as well as the other incoming flow parameters is declared by the source at

the moment of request about available network channel resources.

The value declared in the request, the incoming flow intensity, for example, is not known for sure and get assigned roughly.

In general case, any declared parameter is designated as A. For it's uncertainty this parameter can be considered as probability density $P(A)$ with mathematical expectation A_0 and variance σ_A^2.

In fact communication system does not receive a true variable A_i but an accidental value $a_i = A_i + \xi_i$ influenced by inaccuracy and observation limits. Assume that these variables are independent and identically distributed. Also assume that "noise" of observation has probability density with the finite variance σ_a^2.

Taking into consideration these data the system determines a^* which is a evaluation. The discrepancy: $\epsilon = (a - a^*)^2$.

During r periods the system gets a vector of parameters values $\mathbf{a}^{<r>} = (a_1, a_2, ..., a_r)$. A specific risk SR while choosing the optimal value of a^* parameter:

$$SR = \int_{-\infty}^{+\infty} (a - a^*)^2 \cdot P(a/\mathbf{a}^{<r>}) da, \tag{6.14}$$

where $P(a/\mathbf{a}^{<r>})$ is an a-posteriori condition of probability distribution density for variable a with the preset vector $\mathbf{a}^{<r>}$. The optimal evaluation value a^*_{opt} is gained by minimization of the specific risk SR that leads to the following:

$$a_{opt} = \int_{-\infty}^{+\infty} a \cdot P(a/\mathbf{a}^{<r>}) da = M[a/\mathbf{a}^{<r>}]. \tag{6.15}$$

According to identification theory [172] the best distribution $P(a/\mathbf{a}^{<r>})$ with the limitations mentioned above should be normal. Using the parameters of a normal distribution in Eq. 6.15 it looks as following:

$$a_{opt} = M[a/\mathbf{a}^{<r>}] = \frac{\left[A_0 \frac{\sigma_a^2}{\sigma_A^2 \cdot r} + \frac{1}{r} \sum_{i=1}^{r} a_i\right]}{\left(\frac{\sigma_a^2}{\sigma_A^2 \cdot r} + 1\right)} \tag{6.16}$$

This evaluation expression results into a particular case when the measurements of parameter variables are produced without lapses, i.e. when $\sigma_a^2 = 0$.

In that case the optimal value of the parameter being evaluated will be:

$$a_{opt} = \frac{1}{r} \sum_{i=}^{r} a_i. \tag{6.17}$$

This value is formed at r period of flow observation and can be denoted as $a^*_{opt}(r)$. The expression can look differently:

$$a_{opt}(r) = \frac{1}{r}\sum_{i=1}^{r} a_i = \frac{r-1}{r} \cdot \frac{1}{r-1} \cdot \sum_{i=1}^{r-1} a_i + \frac{1}{r} a_i, \tag{6.18}$$

i.e.

$$a_{opt}(r) = a^*_{opt}(r-1) + \frac{1}{r}\left[a_r - a^*_{opt}(r-1)\right]. \tag{6.19}$$

The later equation leads to the algorithm of parameter evaluation: on another r step of observation the optimal value of the parameter turn into mathematical expectation of the parameter gained at the previous $r-1$ stage of observation, plus the correction that is equal to: $\frac{1}{r}(a_r - a^*_{opt}(r-1))$.

The recurrent expression that has been calculated above (Eq. 6.19) belongs to the class of stochastic approximation in its simplest form [173]. Thus, in the end of time period T the described approximation procedure results the formation of data about a parameter or parameters in the system.

Further, the evaluation procedures when a new flow is requested for the access to the operating flow will be described.

At the moment t_Z a request for the access of a new flow is received which needs resource more than the remained for the moment Δt_{Z-1} (a moment of giving the result for the previous measurements time period). In this case the system will not allow a new flow in spite at the moment T_Z the system has enough resources. Such a situation leads to the errors in the work of MBAC.

To solve this problem the observation period T has to be determined. It cannot be too long as the character of the current traffic may change suddenly and the system will not be able to note that change. Therefore, the appropriate value for the period T can be the correlation interval τ_k. The flow can change its statistical parameters after this interval. Thus:

$$T = \tau_k \tag{6.20}$$

The second parameter of MBAC has to be the period Δt. If it is high, a high uncertainty in the evaluation of the flow parameters appears. The most accurate option is the observation of each incoming packet. But it is impossible due to time losses of processing.

The definition of the time period has to take into account the flow character as this is observed in the self-similar traffic [13]. Using recommendations of the present work the following suggestions are proposed.

There is no necessity to expect the end of the period T at the incoming of a new flow at the moment T_Z. The evaluation of the working flow can be done using the accumulated data

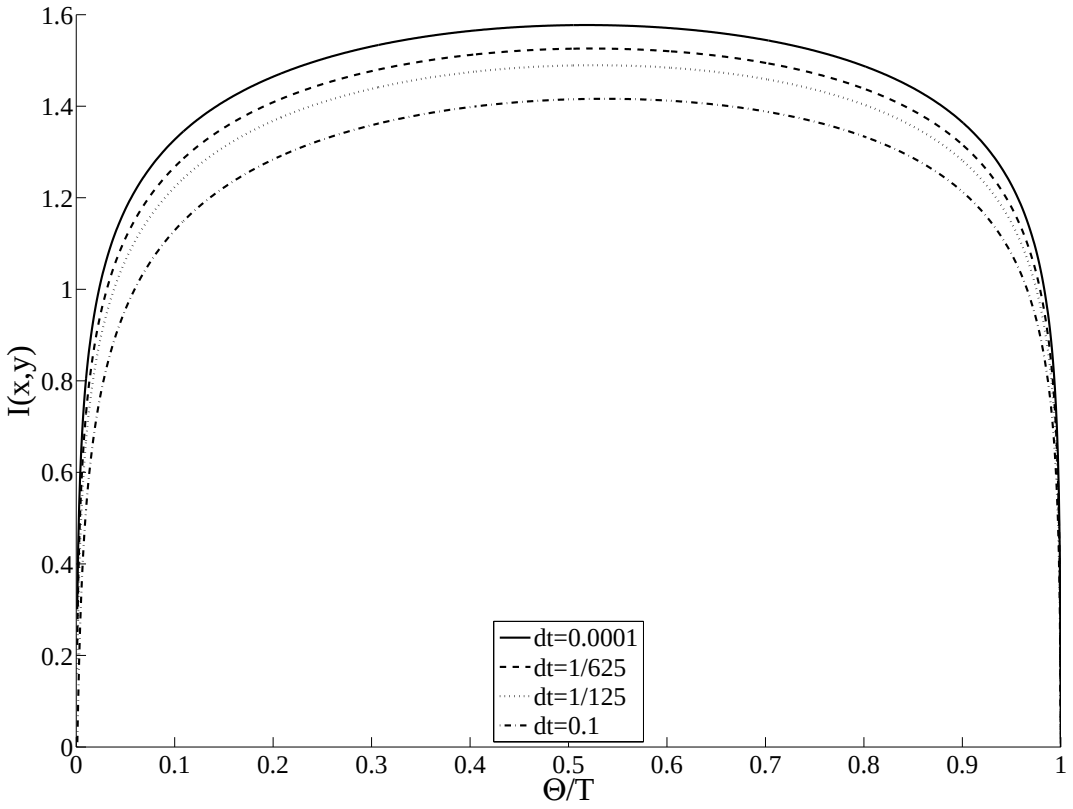
from periods $\Theta << T$ on the basis of the following expression:

$$I(x,y) = \frac{1}{2}\left\{1 + (t-\Theta)^{2H}\left[1 - \frac{[t^{2H} + (t-\Theta)^{2H} - \Theta^{2H}]^2}{4t^{2H}(1-\Theta)^{2H}}\right]\frac{(\Theta - \Delta t)}{b}\right\} \tag{6.21}$$

where $I(x,y)$ is the information volume that is gained by the destination y from the source x during the measurements for time Θ previous to the moment of taking the decision t, for example $t = T_Z$; H - Hurst parameter; Δt - period between the countdowns in the traffic $X(t)$, and $b = (\Theta - \Delta t)\sigma_z^2$, where σ_z^2 is the variance of packets number at single measurement.

Building the relationship $I(x,y)$ from Θ (period needed to check and process the measurements), it can be seen that $I(x,y) \to max$, if Θ is $\frac{1}{2}$ of the moment of the last evaluation t_i and the moment of taking the decision t. In the example used in the paper, $t = T_Z$ and is equal to the moment of getting the request for the new flow.

The constructed graph (Fig. 6.3. - Fig. 6.5.) shows that the necessary period of measurements and processing decreases while the self-similarity coefficient H increases. It is logical, as with the increase of H, flow correlation grows too. Therefore the flow can be observed using the shorter period of time.

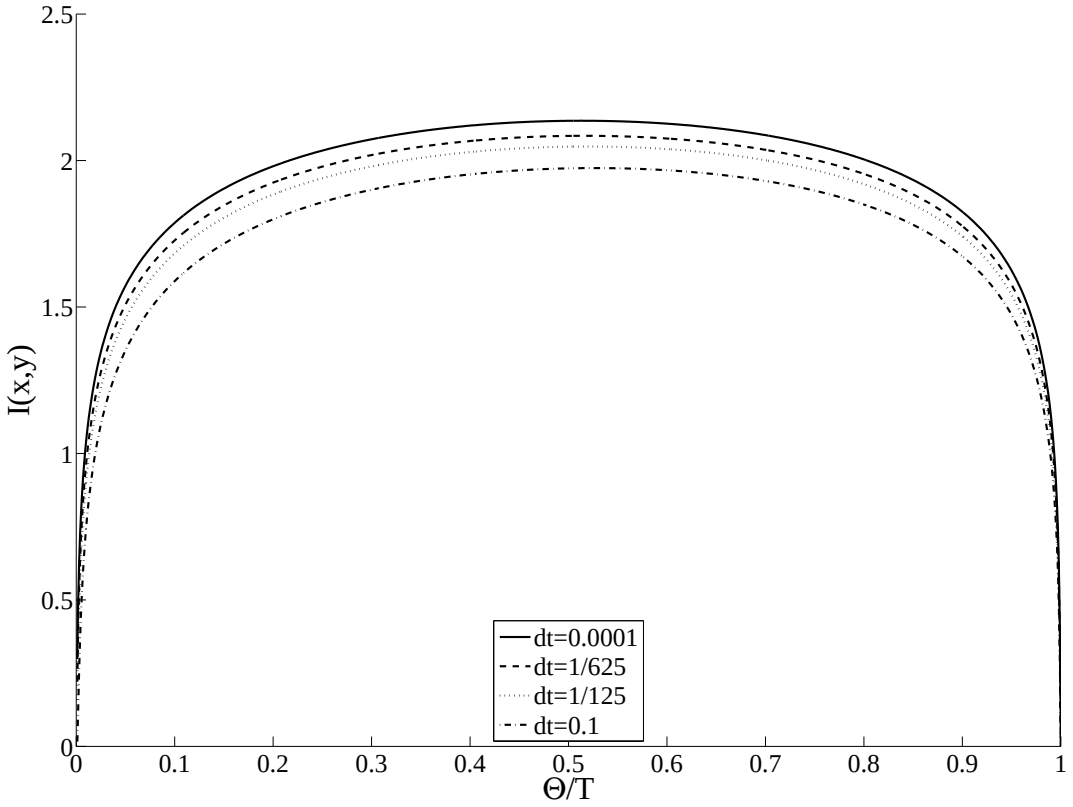


6.3. Figure $I(x,y)$ depending on Θ with various Δt for the $H = 0.5$.

At this point a conclusion about the time period Δt between the countdowns of incoming flow $X(t)$ can be made. It is possible to create a family of curves $I(x, y)$ depending on Θ with various Δt. Based on this argument it is possible to determine the period between the countdowns Δt^* while $I(x, y)$ are at their max.

The Fig. 6.3. - Fig. 6.5. shows the $I(x, y)$ depending on Θ with various Δt. For the charts Δt was chosen as follows: $10^{-4}, \frac{1}{625}, \frac{1}{125}, 10^{-1}$ that are depicted with solid, dashed, dotted and dotted-dashed curves respectively. The Fig. 6.3. - Fig. 6.5., presents depending for the H parameters $H = 0.5, 0.75, 0.95$ respectively.

The graphs show that Δt and Θ depend on the self-similarity power (H). The procedure described belongs to the case that evaluates the flows already existing in the system when the evaluation of aggregated current flow in the system takes place using the periods T with the gradual recurrent precision of the flow parameters with sub-periods $\Delta t << T$.



6.4. Figure $I(x, y)$ depending on Θ with various Δt for the $H = 0.75$.

Initial values Next, the case of MBAC starting phase is presented. The system starts to receive the requests for access of the first flow. Data about it declared by client can inform the system about an expected intensity of the flow (with some errors). The other flow parameters, a self-similarity degree, for example, might be unknown. In this case the initial period S of

the accumulated data about unknown flow and observation frequency are chosen.

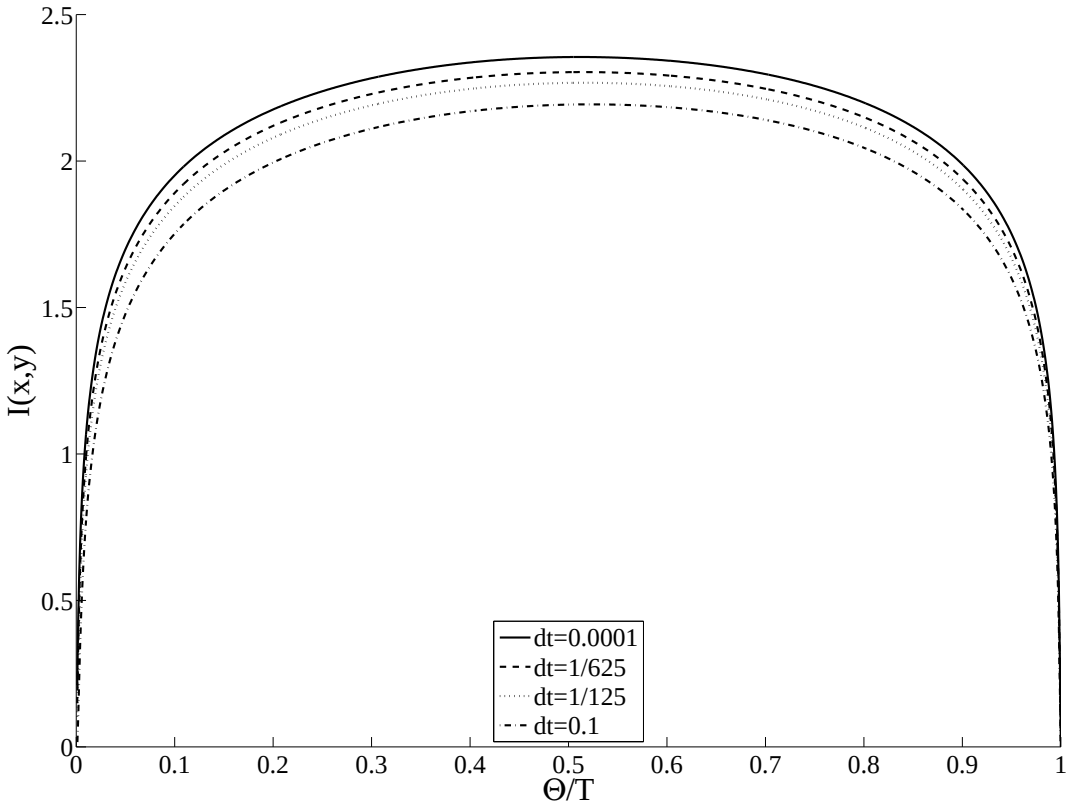
Clearly, the more measurements, the more accurate are the results. For this reason the initial period between the countdowns δ is thought to be reduced. However, there is no point in making δ too short. Therefore, the worst case is chosen, which the Poisson character of the flow is, i.e. it is not correlated. If the declared intensities of the flow λ are known, the period δ between the countdowns can be taken as equal to correlation period of Poisson flow. This period can be determined and equals to $\delta = \frac{1}{\lambda}$, and it's correlation function is $R(\tau) = e^{-2\lambda|\tau|}$.

The larger the observation period S is, the more accurate are the measurements. The measurements error $\beta = \frac{D}{(m-1)}$, where D is the process variance and m is the number of measurements ($m = \frac{S}{\delta} = S\lambda$). Poisson process variance represents $D = R_x(0) = 1$. Measurements error can be expressed in the following way:

$$\beta = \frac{D}{m-1} = \frac{1}{S\lambda - 1} \approx 1/S\lambda. \tag{6.22}$$

If the measurements error is preset $\beta^* = 1\%$, than $S = \frac{1}{0.01\lambda} = \frac{100}{\lambda}$.

The resulting graph of the observation and measurements process looks as it is shown on Fig. 6.6.



6.5. Figure $I(x, y)$ depending on Θ with various Δt for the $H = 0.95$.

At the initial stage of research that studies the flow with the length S, the observation are taken in the period $\delta << \Delta t$, i.e. in the periods between the incoming packets. Further, on the basis of gained statistics, the correlation coefficient τ_k which equals T_1 gets calculated. After the period T_1 ends, the flow analysis for time S is commenced again and a new correlation period is calculated, as well as a new period T_2 is assigned.

If before the end of period T_2 a new flow at the moment T_Z tries to enter the system, than on the basis of data gained for the past period Θ that can be estimated using the formula (Eq. **??**) the parameters of the current flow and resources remained for the new flow connection to the system are clarified.

The parameters of the flow that is being connected have to be declared. At least, their packet income intensity has to be preassigned. Further, the availability of throughput and buffer capacity resources have to be discovered. This would provide the parameter specified by QoS agreement, for examples the probability of packets loss P^*_{Loss}. The value of the Hurst coefficient (H) is set by the results of existing flow measurements. For P_{Loss} calculation on the basis of the declared parameters the tabulated correlations gained on the basis of article [139] can be applied.

If $P_{Loss} < P^*_{Loss}$, than the new flow is given an access into the system. Otherwise, it does not happen.

Taking into consideration a well known fact that the self-similar process has a group character, it is possible to produce dynamic management of the system parameters like throughput and buffer capacity on the basis of suggestions proposed in the articles by [12] and [13].

Conclusions Several solutions for the possible errors of flows parameters estimation minimization have been proposed in the section.

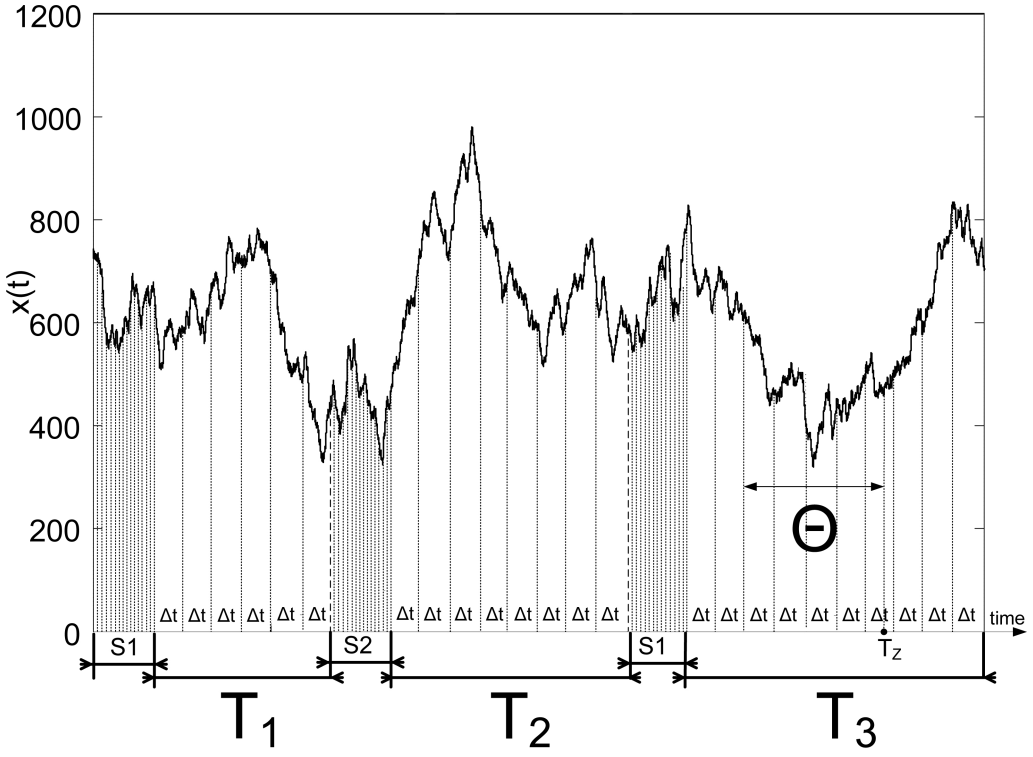
In the section the recurrent algorithm of parameter evaluation on r step of observation was presented. The recurrent expression belongs to the class of stochastic approximation. Thus, in the end of time period T the described approximation procedure results in the formation of data about a parameter or parameters in the system.

Solutions for observation and sampling period are proposed for both initial values selection and on-going estimation.

For the case when the flows are already admitted and the character of the current traffic may change suddenly the correlation interval τ_k has been proposed for use.

The maximization of $I(x, y)$ depending on Θ with various Δt grant the sampling period Δt estimation. It was shown that the necessary period of measurements and processing decreases while the self-similarity coefficient increases (H).

For the initial values of the system the important results have been gained. If the declared intensities of the flow λ are known, the period δ between the countdowns can be taken as equal to correlation period of Poisson flow. This period can be determined and equals to $\delta = 1/\lambda$.



6.6. Figure An integral measurement process of incoming traffic for MBAC.

The initial data flow observation period should be $S = \frac{1}{\beta \cdot \lambda}$, where β is a measurements error.

Interestingly, Θ remains unchangeable in a wide range. That fact has to be taken into consideration choosing the value for Θ. It does not have to be too small either too large.

6.2.2. Resource Allocation Method for Admission Control Algorithm

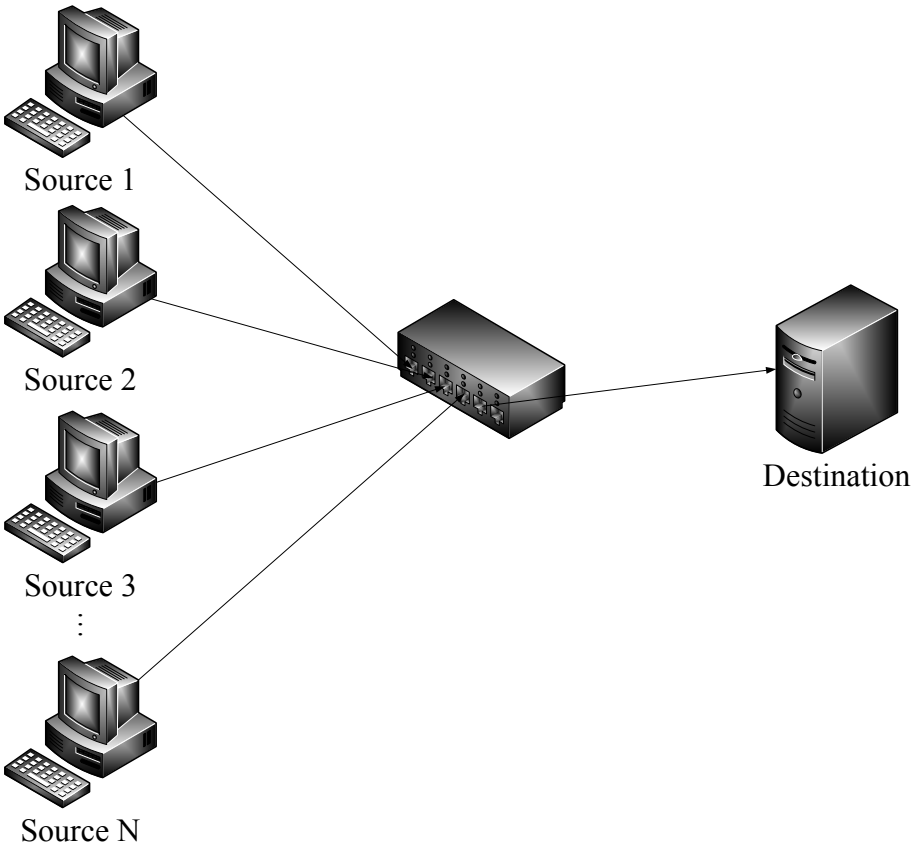
The present section concentrated on resource allocation in communication networks in order to assure specific QoS characteristics, such as packet loss probability for requested new connections.

To evaluate admission control decision algorithm based on the packet loss probability for we are going to estimate the memory volume according the method described in paper [139].

The decision for choosing call admission in our model is based on requested packet loss probability (P^*_{Loss}) assurance and available resource that will be allocated to guaranty required QoS parameters. To evaluate the resource allocation we have modeled a framework of multiple sources.

The framework of the model that were used for the simulation is presented as follows. The model consists of the multiple sources that belong to the same QoS priority level. It means that all sources have a similar right to be admitted and gain similar resources. Traffic

of multiple sources goes through the switch where the multiplexing is performed (Fig. 6.7.).



6.7. Figure The framework of the model

It is important to mention that the probability that arrival rate of the aggregated traffic would be equal to peak rate of the individual source approaches tends to zero. That is why it is possible to gain high network utilization while saving the QoS packet loss guaranties.

The traffic shaping is integrated into the switch. We considered the shaper policy as follows: if the limited resources of switch are not sufficient to guaranty the requested packet loss probability, then traffic shaper decreases the bandwidth of every established connection. For the case with multiple sources of the same QoS priority level it means that the bandwidth available for an individual source is inverse to the number of sources.

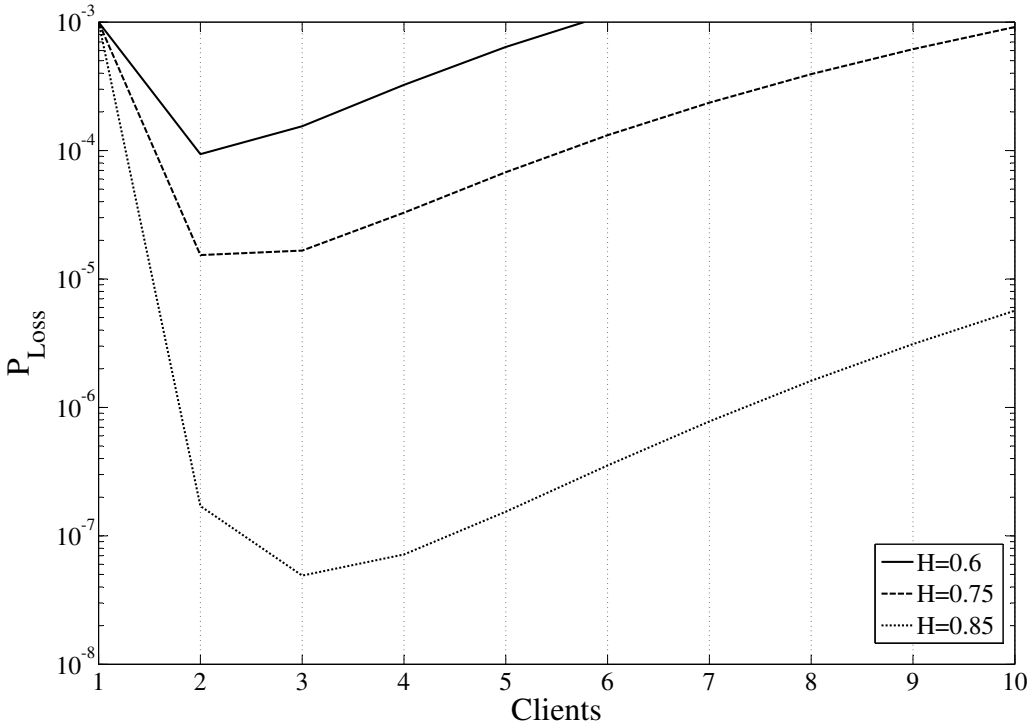
In the model we work with the buffer memory allocation proceeds in the following way. In our experiments we use the worst case of memory allocation which is allocation by zones. It means that each connection gets the proportional amount of the buffer memory.

Within the described model the packet loss probability (P_{Loss}) was estimated for different parameters of traffic inter-arrival rate/utilization ($\rho = \frac{\lambda}{\mu}$, where λ-packet inter-arrival rate and μ-packet service rate) and self-similarity parameter (H).

At first, the buffer capacity was estimated according to parameters of the individual

source and packet loss probability $P_{Loss} = 10^{-3}$. For the experiments the parameters of the single source model were chosen to create the utilization equal to $\rho_1 = 0.5, \rho_2 = 0.75$ and $\rho_3 = 0.8$. For the buffer capacity estimation the self-similarity H parameter considering as equal to $H_1 = 0.5, H_2 = 0.75$ and $H_3 = 0.85$.

Fig. 6.8. - Fig. 6.10. show the P_{Loss} probability for the connections. The packet inter-arrival rate decreases proportionally to the number of established connections. The buffer size of each source gets allocated proportionally to the number of connections in reference to initially estimated.

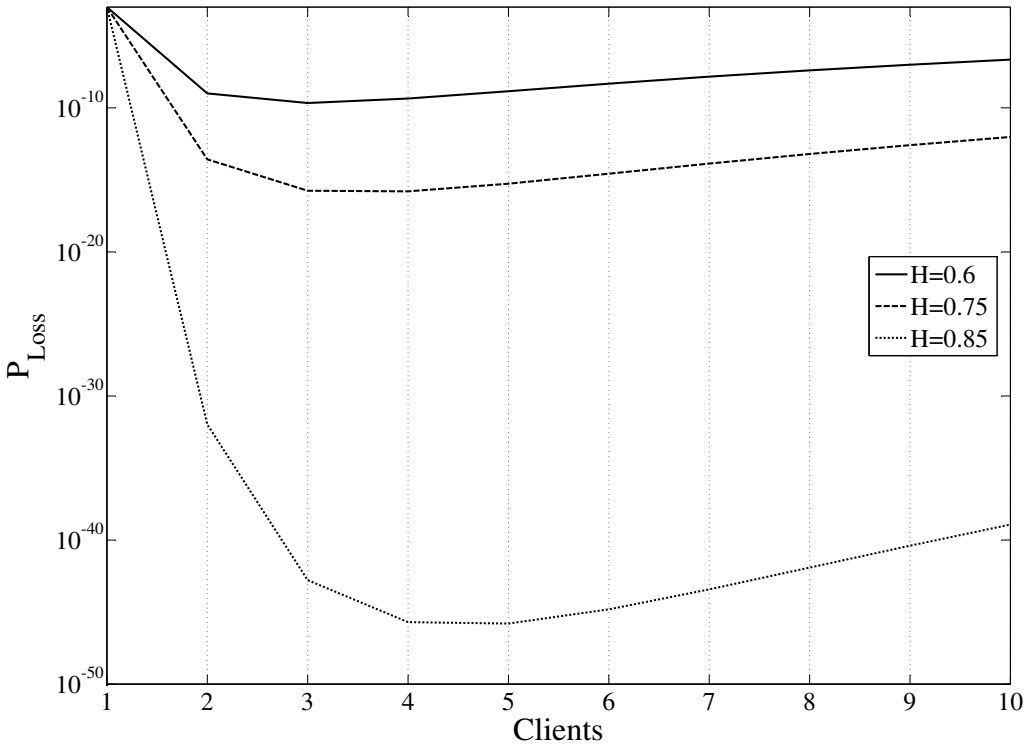


6.8. Figure: Packet loss probability for the source with decreasing inter-arrival rate proportional to the connected clients for the one client with arrival rate corresponding to $\rho = 0.5$

The solid line shows the packet loss probability for the inter-arrival rate of the one source that corresponds to long-range dependence parameter $H = 0.6$. The dashed line corresponds to $H = 0.75$, and the dotted line corresponds to $H = 0.85$.

In Fig. 6.8. - Fig. 6.10. it can be seen that packet loss probability decreases when the arrival rate decreases proportionally to the number of clients, and allocated memory decreases according to arrival rate. It decreases until the minimum and later on starts to increase. Respectively, cumulative packet loss probability $P^{\Sigma}_{Loss} = \sum_{i=1}^{ClientNumber} P_{Loss}$ for the multiplexed traffic of established client's connections decreases too. Such minimum, as you can see from the charts, depends on the input traffic parameters.

The multiplexing effect can be described as follows: for the one QoS class, leaving the



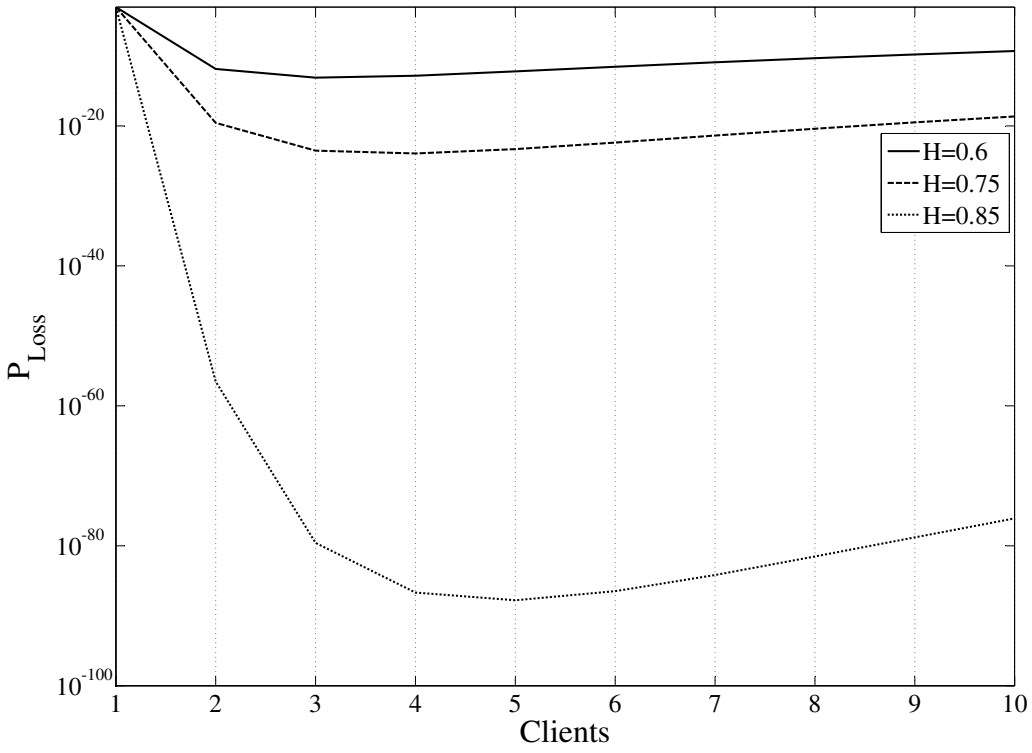
6.9. Figure: Packet loss probability for the source with decreasing inter-arrival rate proportional to the connected clients for the one client with arrival rate corresponding to $\rho = 0.75$

constant long-range dependent parameter and utilization with the number of clients increasing (decreasing the arrival rate of individual source) the aggregated traffic arise smaller packet loss probability. Fig. 6.8. - Fig. 6.10. demonstrate that greater the gain from the multiplexing could be achieved for the high utilization, or for the traffic with high long-range dependence parameter - H.

Moreover, comparing packet loss probability of the multiplexed traffic, single source traffic with similar long-range dependence parameter and produced similar network utilization the packet loss probability for the multiplexed traffic for some number of clients is smaller comparing to single source.

Fig. 6.11. illustrates the gained buffer capacity during the traffic aggregation. The graph shows the relationship between the gained and initial memory capacity taking into consideration the number of clients and the Hurst parameter. It is clearly seen that for the aggregated traffic that consists of multiple sources with high self-similarity parameter the benefit of the multiplexing is higher.

The advantage we gain of this result is the connection of more clients with the defined packet loss guaranty. In other words, we can have more available buffer memory for newly arrived connections. For such ”free” resources we use the ”over-utilization” term.



6.10. Figure: Packet loss probability for the source with decreasing inter-arrival rate proportional to the connected clients for the one client with arrival rate corresponding to $\rho = 0.8$

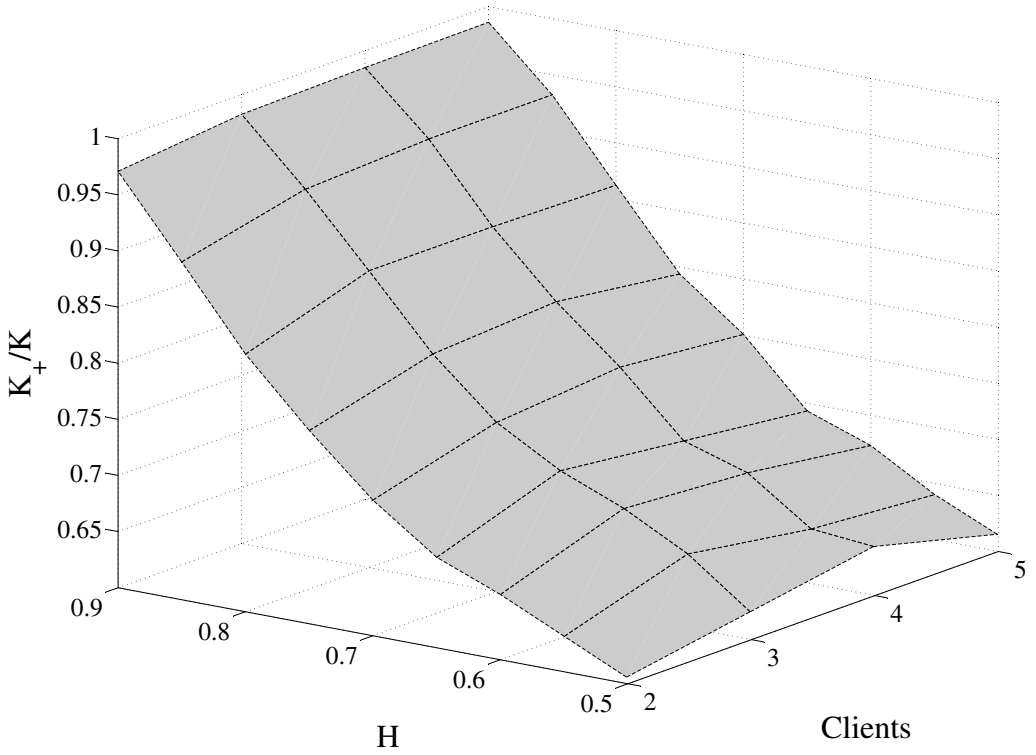
The "over-utilization", which is the bandwidth that has been achieved by multiplexing, should be taken into account by admission control as additional free resources in decision making phase.

If the admission control has enough resources to admit a connection with the parameters that meet the requirements of utilization $\rho \rightarrow 1$ and requested packet loss probability, the traditional admission control will admit only 10 sources with $\lambda = 0.2 * \mu$ to satisfy required packet loss probability. More than 5 connections could be admitted with packet loss guaranty in case of "over-utilization" parameter.

It has to be noted that "over-utilization" parameter depends on parameters of multiplexed connections and should be evaluated on the fly.

Conclusions The section presents the advantages of multiplexing according to packet loss probability. The results shows that the packet loss probability for the aggregated data flow decreases with the decreasing of the arrival rate of the individual source proportionally. The packet loss probability decreases until the minimum and later on starts to increase. Such a minimum value, as you can see from the charts, depends on the input traffic parameters.

In our experiments for different variety of parameters the minimum of cumulative packet loss is reached between 2 and 10 clients depending on the parameters.



6.11. Figure The buffer capacity during the traffic aggregation for $\rho = 0.75$

The experiments show that the probability that arrival rate of the aggregated traffic would be equal to peak rate of the individual source decreases with the number of individual sources. That is why it is possible to provide high network utilization while saving the QoS packet loss guaranties. Furthermore, the packet loss probability of the aggregated traffic comparing to the single source with the similar traffic characteristics for some number of clients is smaller in comparison to the single source. It offers to use gained buffer memory for the "over-utilization" - to admit additional connections. The "over-utilization" could be used either for admission control or for traffic control.

While working on admission control effectiveness improvement the following issues are useful to be taken into account. The large number of sources produce smaller packet loss probability as it would be evaluated for single source with respective arrival rate. This effect gives an opportunity to admit additional sources and to increase the network utilization. Generally, in the admission control field the "over-utilization" shows how many additional sources could be admitted into the system.

Traffic control could use the "over-utilization" parameter in the following way. Traffic control in case of collisions should decrease the input rate. Utilizing the "over-utilization" parameter the input rate will be decreased less than in the case when that parameter is not used. As the future work it is important to mention the work related to "over-utilization" parameter

on-line evaluation for the different number of input sources and its traffic parameters.

6.2.3. Optimal Dispatching of the Flows Falling in the Same Priority Class

In order to provide the required quality of service (QoS), in current data communication networks, a mechanism is intensively used allowing the data flow access to the communication system and further to the network. For efficient flow management, the flow classes and the priorities corresponding to them are introduced.

Four classes of flows are separated to which priorities are assigned. Those with the top priority are the flows that do not allow delays in the servicing of packets in the course of transmission, for instance, video and audio broadcasts. At the opposite end are the file data transmission flows, where the packet delay does not result in any negative consequences. On the basis of the priorities, the resources of the communication system are allocated, i.e., the buffer storage size and the output channel bandwidth. However, the problem of the resources allocation is complicated if two or more flows falling in the same priority class arrive at the input.

Therefore, in the present section, the case is considered of several competing flows falling in the same class. This section is dedicated to solving the question of which flow should be allowed to transmit data in the case of limited system resources.

The problem can be stated as follows. There is a communication node at the input of which n flows arrive with the intensities λ_i $(i = \overline{1,n})$. Each flow in the output channel is provided with the channel resources: the frequency band (pass band) or the servicing intensity μ_i $(\frac{1}{sec})$. If the rate of the transmission/servicing for the i-th flow comprises c_i $(\frac{bit}{sec})$ and the average packet length is l_i, then $\mu_i = \frac{c_i}{l_i}$ $(\frac{packet}{sec})$. Each i-th flow produces the load ρ_i of the communication node and the output channel out of the total load $\rho = \sum\limits_{i=1}^{n} \rho_i$.

Each flow is characterized by the servicing time variation coefficient v_i. It is assumed that the communication node has the total buffer storage capacity of r packets. The storage space is divided into regions so that each flow is provided with the capacity r_i. Naturally, $r \geq \sum\limits_{i=i}^{n} r_i$. It is assumed that maintenance procedures for servicing flows with identical priorities are used.

On complete filling of the storage, the newly arriving packets are lost regardless of their priority.

In accordance with [174], the probabilities of the packet losses for the flow of the k-th priority (the first priority is assigned to $k = 1$) are known:

$$P_{Loss_k} = \frac{\rho_{\Sigma_k}}{\rho_k} \frac{1 - \rho_{\Sigma_k}}{1 - \rho_{\Sigma_k}^{\frac{2r_k}{1+v_k^2}+1}} \rho_{\Sigma_k}^{\frac{2r_k}{1+v_k^2}}, \tag{6.23}$$

where $\rho_{\Sigma_k} = \sum_{i=1}^{k} \rho_i$ and r_k is the capacity of the buffer region for the k-th flow that has the servicing time variation coefficient v_k. Naturally, expression (Eq. 6.23) is obtained under some restrictions and assumptions [174].

Two problems are to be solved:

1. to determine which flow (within the class) should be assigned the highest priority and which the lowest;

2. to determine the optimum buffer storage capacity $r_i^*(i = \overline{1,n})$ for the competing flows of the same class.

The flow priority should be determined by the following algorithm. Consider two flows ($n = 2$). The first flow is assigned the highest priority $k = 1$. The packet loss probability for this flow is

$$P_{Loss_1} = \frac{1-\rho_1}{1-\rho_1^{\frac{2r_1}{1+v_1^2}+1}} \rho_1^{\frac{2r_1}{1+v_1^2}}. \tag{6.24}$$

The packet loss probability for the flow with the second priority is

$$P_{Loss_2} = \frac{\rho_1+\rho_2}{\rho_2} \frac{1-(\rho_1+\rho_2)}{1-(\rho_1+\rho_2)^{\frac{2r_2}{1+v_2^2}+1}} (\rho_1+\rho_2)^{\frac{2r_2}{1+v_2^2}}. \tag{6.25}$$

where $r_2 = r - r_1$, and r is the total buffer storage capacity of the communication node.

The total probability of the packet loss for two flows is

$$P_\Sigma = 1-(1-P_1)(1-P_2) \approx P_1 + P_2 \tag{6.26}$$

Take the ranking of the flows as unknown and therefore assume that there are two flows A and B that produce the loads ρ_A and ρ_B. Consider two cases: the case a is when the priority is given to flow A, and the case b is when the priority is given to flow B.

The packet loss probability for the case a is

$$P^a_{Loss_\Sigma} = \frac{1-\rho_A}{1-\rho_A^{\frac{2r_A}{1+v_A^2}+1}} \rho_A^{\frac{2r_A}{1+v_A^2}} + \frac{\rho}{\rho_B} \frac{1-\rho}{1-\rho^{\frac{2r_B}{1+v_B^2}+1}} \rho^{\frac{2r_B}{1+v_B^2}}, \tag{6.27}$$

where $\rho = \rho_A + \rho_B$. The packet loss probability $P^b_{Loss_\Sigma}$ for the case b is obtained by way of interchanging the positions of the indices A and B in Eq. 6.27.

If $P^a_{Loss_\Sigma} < P^b_{Loss_\Sigma}$ then the priority should be given to the flow with the index A, and vice versa.

In order to simplify the comparison procedure, one can take the variation coefficients as

equal: $v_A^2 = v_B^2 = 1$. Then,

$$P^a_{Loss_\Sigma} = \frac{1-\rho_A}{1-\rho_A^{r_A+1}}\rho_A^{r_A} + \frac{\rho}{\rho_B}\frac{1-\rho}{1-\rho^{r_B+1}}\rho^{r_B}, r_B = r - r_A. \tag{6.28}$$

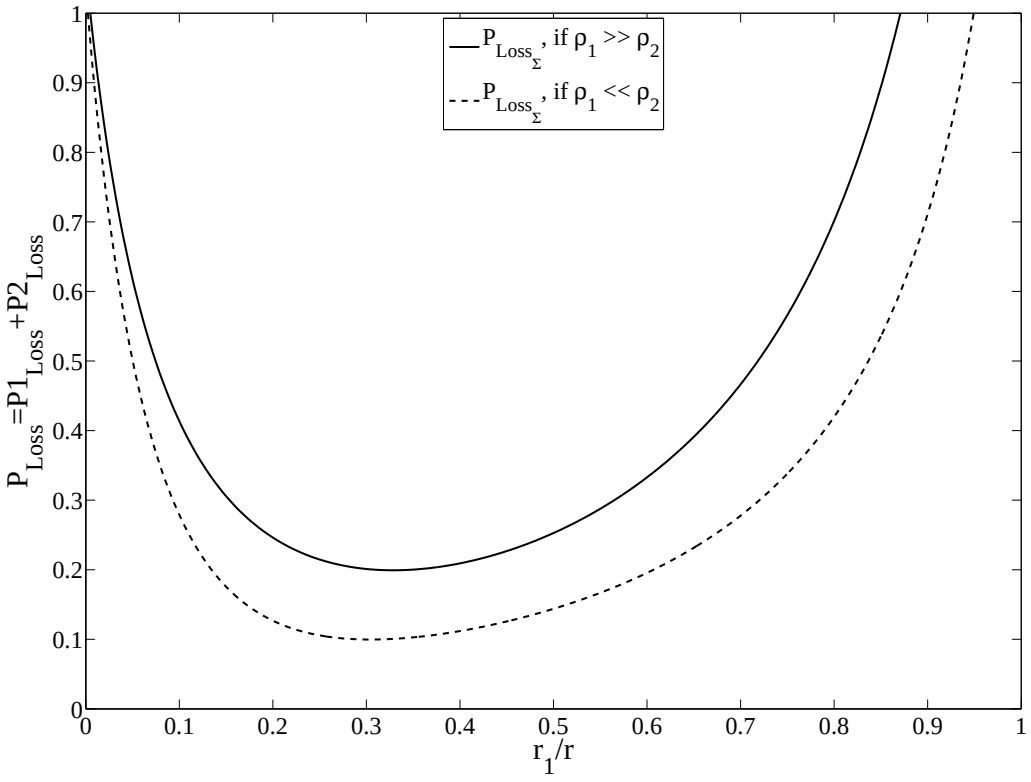
Consider the case of $\rho_A << 1$, while $\rho_B \to 1$, and assume that $r_A = r_B = r_x$. Then, the probability of losses can be estimated as

$$P^1_\Sigma \cong \rho_A^{r_x} + \frac{\rho}{\rho_B}\frac{1-\rho}{1-\rho^{r_x+1}}\rho_x^r. \tag{6.29}$$

If $\rho_A \to 1$ while $\rho_B << 1$, then

$$P^2_\Sigma \cong \frac{1-\rho_A}{1-\rho_A^{r_x+1}}\rho_A^{r_x} + \frac{\rho}{\rho_B}\frac{1-\rho}{1-\rho^{r_x+1}}\rho_x^r. \tag{6.30}$$

In the second case, the first addend is changed by the factor $\frac{1-\rho_A}{1-\rho_A^{r_x+1}}$, and the second one by the factor $\cong \frac{1}{\rho_B}$.



6.12. Figure: Total probability of the packet loss. The load produced by the first flow equals 0.5 and that by the second flow 0.3, where $r = 10$. The solid line stands for the priority being given to the first flow, and the dashed one, for the priority being given to the second flow.

Reasoning from this, the second strategy of assigning the priorities is disadvantageous.

Hence follows the conclusion: the priority should be given to the flow that produces less communication system load.

This conclusion is supported by the plots given in Fig. 6.12. - Fig. 6.14. In Fig. 6.12., the plots are given of the total (for the first and second flow) probability of the packet loss for the total storage capacity of ten packets and the servicing time variation coefficient equal to unity for both flows. The load produced by the first flow equals 0.5 and that by the second flow 0.3. On the plot, the packet loss probability in the case when the priority is given to the first flow is shown as a solid line, and the loss probability in the case when the priority is given to the second flow is shown as a dashed line. From the plot, it is obvious that the conclusion reached above is confirmed.

The advantage of giving the priority to the flow that produces a smaller load becomes more evident with the increase in the loads' ratio, as well as with the increase in the servicing time variation. For instance, in Fig. 6.13. the plots are constructed of the total packet losses for the flows producing loads equal to 0.8 and 0.1 for the first and second flow, respectively. The total buffer storage capacity equals $K = 31$. In Fig. 6.14., a plot is presented similar to the previous one but with the queries' servicing time variation increased to two ($v_1 = v_2 = v = 2$).

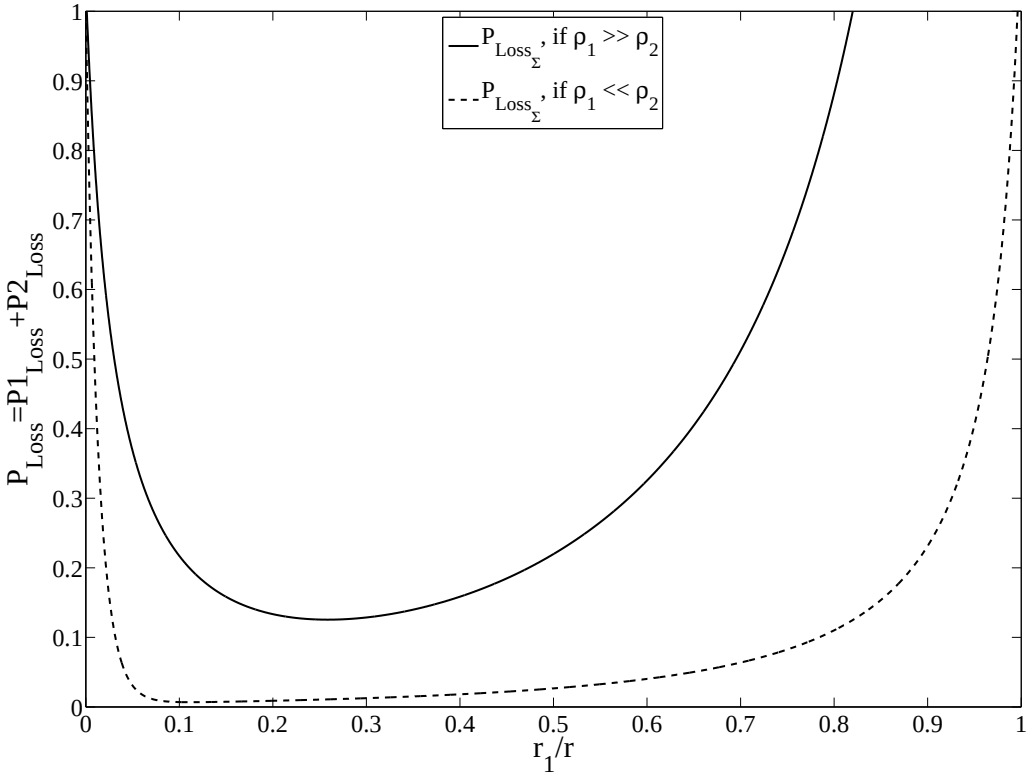
From the plot in Fig. 6.14., it follows that the increase in the servicing time variation coefficient results in the increase of the packet loss probability. Moreover, the total probability of losses differs essentially as compared to the first case, to which corresponds the variation coefficient $v = 1$, i.e., when the packets' servicing time is distributed exponentially.

The dependencies presented in Fig. 6.14. suggest that the occurrence at the input of a self-similar plot characterized by a large variation coefficient aggravates the situation still further: the wrong choice of the priorities leads to a drastic increase in the packet loss probability as compared to the case of the arrival of an elementary flow with an exponentially distributed packet servicing time.

Determining the optimum buffer storage capacity $r_i^*(i = \overline{1,n})$ for competing flows is presented below.

The aim of solving the problem in question is to determine the optimum storage capacities for the first and second flows bearing in mind that the total storage capacity is $r = r_A + r_B$. Assume that the priority is given to flow A, which produces the load $\rho_A << \rho_B$. It remains to find out what is the optimal storage space distribution r_A^* and r_B^* bearing in mind that $r_B^* = r - r_A^*$.

In order to do this, construct the plots of the dependency of the ratio of the storage capacity provided to the first flow to the total storage capacity $(r_{opt}/K(\rho))$, where $K(\rho)$ is the storage capacity computed from the $M/M/1/K$ model for the specified packet loss probability P_{Loss}. In Fig. 6.15. - Fig. 6.17., the buffer storage space allocation surface is presented, where the y axis (that is, the scale of the storage capacity $(K(\rho))$ is calibrated



6.13. Figure: Total probability of the packet loss. The load produced by the first flow equals 0.8 and that by the second flow 0.1, where $r = 31$. The solid line stands for the priority being given to the first flow, while the dashed one, to the second.

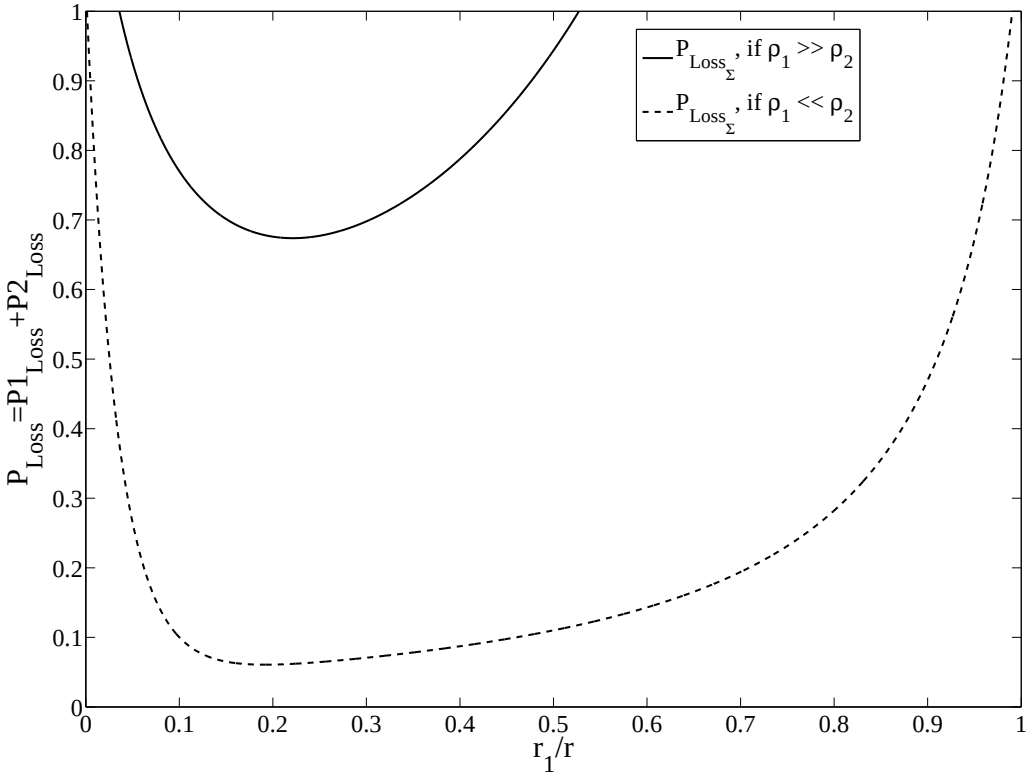
subject to the change of ρ used for computing the storage capacity. In Fig. 6.15. - Fig. 6.17. $(K(\rho))$, the plots are given of the optimal storage space partition for the packet loss probability $P_{Loss} = 10^{-5}$.

First, using the traditional $M/M/1/K$ model implying the arrival to the system of an elementary flow with an exponentially distributed servicing time, we compute the value of the buffer storage capacity K. We determine the capacity from the specified restriction on the packet loss probability P_{Loss}. Next, given the design value of the buffer storage capacity, the solutions surface is constructed, which shows how the ratio is changed of the optimal size r of the buffer storage space allocated to the preemptive flow to the capacity K of the buffer storage obtained from the $M/M/1/K$ model.

The optimal value of the buffer storage capacity allocated to the preemptive flow is determined from a family of plots similar to the ones presented in Fig. 6.12. and Fig. 6.13. For instance, in Fig. 6.12. it is shown that the optimal value is attained for the ratio $r_{opt}/r = 0.25$.

The absence of an analytical expression for computing the optimal value r_{opt} makes us either use a previously computed table of optimal values or employ numerical methods.

We demonstrate how r_{opt} is computed with an example. Assume that we have two flows



6.14. Figure: Total probability of the packet loss. The load produced by the first flow equals 0.8 and that by the second flow, 0.1, where $r = 31$, and $v_1 = v_2 = v = 2$. The solid line stands for the priority being given to the first flow, and the dashed one, to the second flow.

r_1	0	1	2	3	4	5	6	7	8	9	10	11	12
$r_2 = K - r$	12	11	10	9	8	7	6	5	4	3	2	1	0
P_Σ	1.00	0.26	0.102	0.067	0.067	0.082	0.106	0.142	0.195	0.277	0.41	0.711	1.00

6.1. Table Packet loss probabilities for the possible ways of the storage space allocation

producing the loads $\rho_1 = 0.5$ and $\rho_2 = 0.3$. Take the free buffer storage space to be certainly insufficient to provide data transmission without losses. In the present example, we use a storage size less than the required one. From the $M/M/1/K$ model, we compute the storage capacity K for the total $\rho = 0.6$ and the packet loss probability $P_{Loss} = 10^{-3}$. For the specified parameters $K = 12$, find how we should divide the storage space between the arriving flows. Using the assertion stated above that the priority should be given to the flow that produces a smaller load, we assign the highest priority to the second flow. Then, $\rho_A = 0.3$ and $\rho_B = 0.5$.

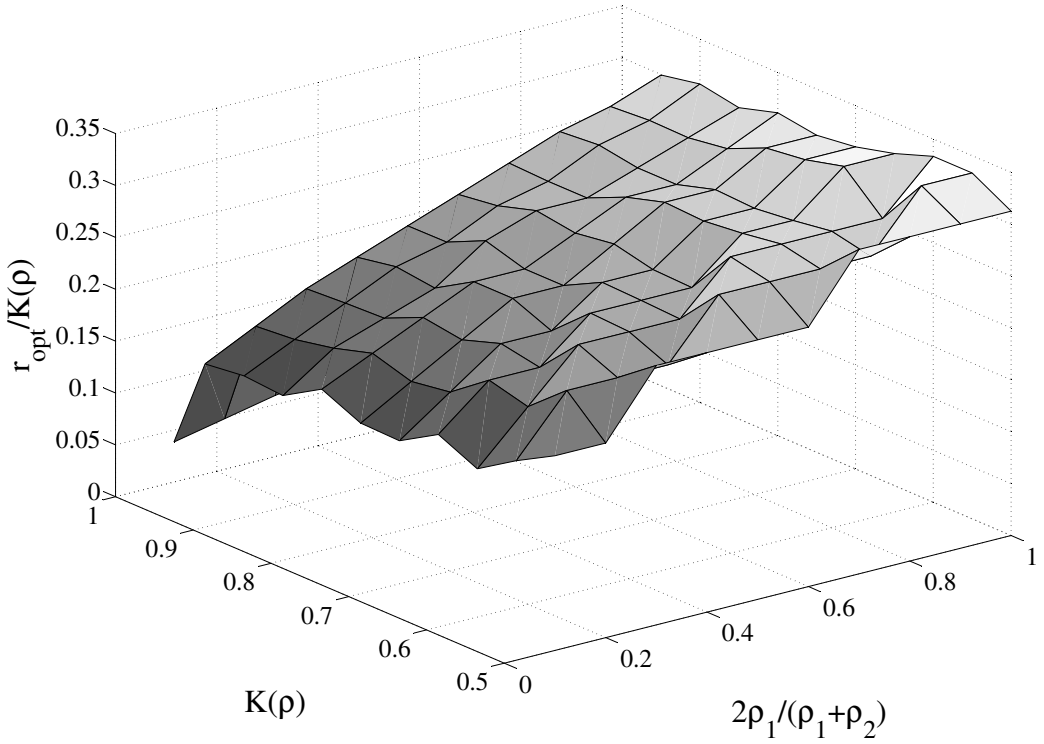
Calculate the probabilities of losses for all the possible ways of the storage space allocation. For the specified values, we obtain the following probabilities of the total packet losses Tab.6.1.

Thus, it is seen that the minimal total probability of the packet loss occurs for $r_1 = 3$ or

$r_{opt}/K = 0.25$.

Conclusions and Recommendations For the current section the following conclusions can be drawn. From the current charts the presented dependencies (Fig. 6.15. - Fig. 6.17.), it is seen that, with the increase in the ratio ρ_1/ρ_2 , the optimal size of the buffer storage space allocated to the preemptive flow increases.

More interesting is the fact that, for $K(\rho)$ falling into a sufficiently wide range, the ratio r_{opt}/K remains almost unchanged. This signifies that the choice of the optimal value is of a stable nature.

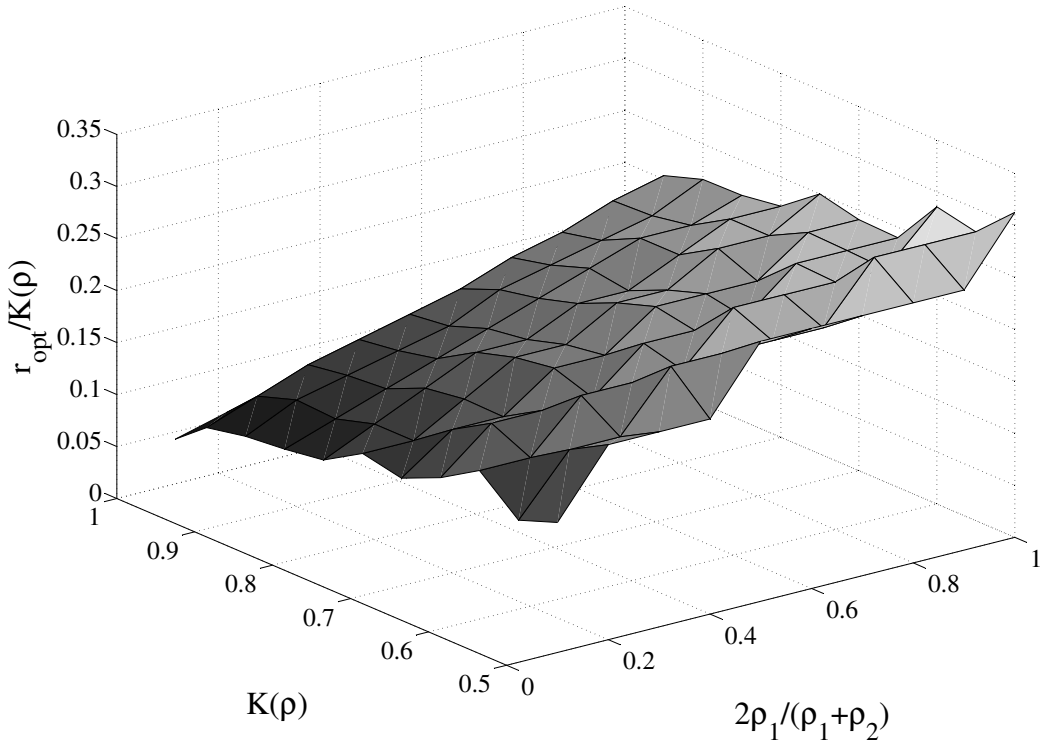


6.15. Figure: Surface of the ratio of the optimal r to K (r_{opt}/K) subject to the allocated buffer space, as well as to the ratio of the loads of the first and second flows $\rho_1 + \rho_2 = 0.6$.

6.2.4. Buffer Size and Output Bandwidth Optimal Relocation

MBAC systems measure and estimate the arrival traffic parameters in real time and use the obtained values and the given quality of service to dynamically allocate resources of the data communication system. The buffer size and output bandwidth allocated for the particular data flow according to its class of service act as such resources.

Several ATM adapted approaches have been proposed in [100, 116, 151, 156, 95]. The suggested algorithms increase network performance by correlating bandwidths and queue



6.16. Figure: Surface of the ratio of the optimal r to K (r_{opt}/K) subject to the allocated buffer space, as well as to the ratio of the loads of the first and second flows $\rho_1 + \rho_2 = 0.8$.

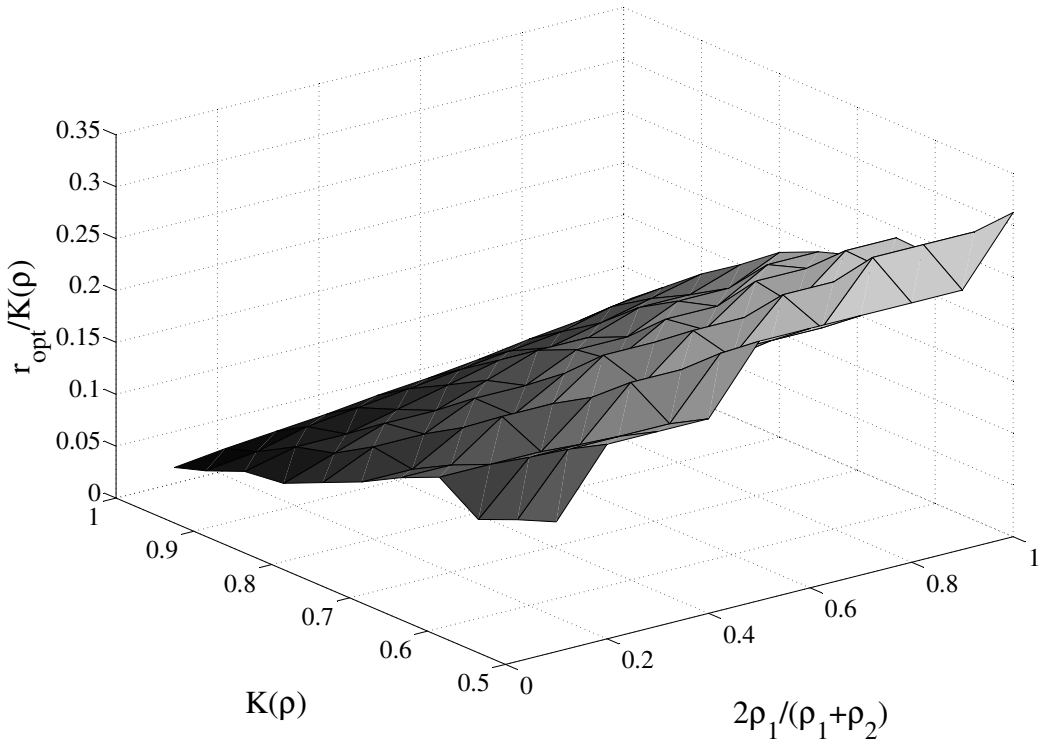
length. The mentioned methods performance metrics are Cell Loss Ratio, average cell delay and average queue length. Further, we suggest the method that besides performance metrics also uses the costs metrics.

We can apply the optimization rule obtained in Section 6.1.1. to the case of adjusting the system's parameters dynamically with the restrictions on the available resources taken into account. The essence is as follows.

Allocating an additional bandwidth or increasing the buffer size used in the communication system for the particular traffic flow, we reduce the loss probability $P_{Loss} = \Phi(\lambda, \mu, K)$. However, this makes the communication system's costs grow; therefore, we need to allocate resources so that relation (Eq. 6.31) is minimal:

$$\frac{\Delta P^i_{Loss}}{P^*_{Loss}} / \frac{c_i}{C}. \tag{6.31}$$

Here ΔP^i_{Loss} is the decrease of the loss probability caused by the allocation of the i-th resource ($i = 1$ if the bandwidth is increased by unity, and $i = 2$ if the buffer size is increased by unity), P_{Loss} is the packet loss probability prior to the resource allocation, C is the cost of the communication system prior to the resource allocation, and c_1 and c_2 are the specific cost



6.17. Figure: Surface of the ratio of the optimal r to K (r_{opt}/K) subject to the allocated buffer space, as well as to the ratio of the loads of the first and second flows $\rho_1 + \rho_2 = 0.5$.

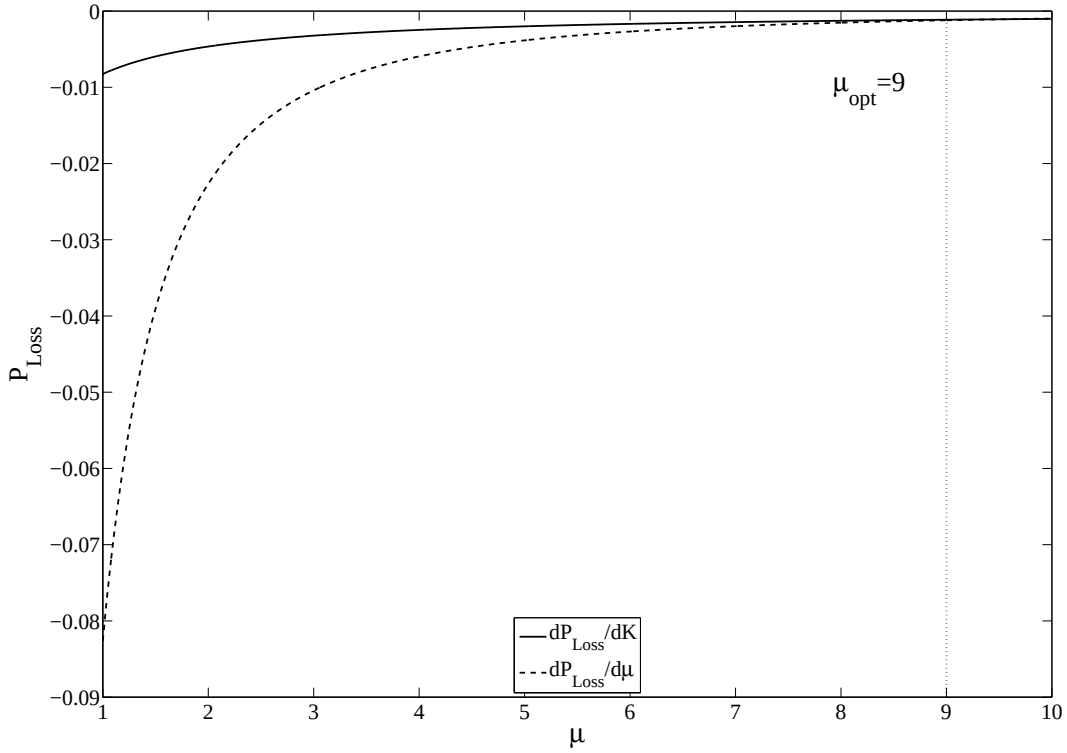
of the unit resource, the unit intensity of the packet processing and the buffer memory unit respectively.

Making successive decisions based on Eq. 6.31, we reduce the packet loss probability. We continue to perform these steps until $P_{Loss} \leq P^*_{Loss}$ where P^*_{Loss} is the given packet loss probability, and the allocation is performed among the available resources. If a new traffic flow arrives and the resource allocation does not ensure the given packet loss probability, the traffic is denied access.

In Fig. 6.18. - Fig. 6.21. show the curves of the derivatives of the packet loss probabilities with respect to K (dP_{Loss}/dK) and μ $(dP_{Loss}/d\mu)$ depending on such parameters as the packet arrival intensity (λ), the coefficient of the geometric distribution (γ), the packet processing intensity (μ), and the buffer size (there are 1 and 10 memory locations on the graphs).

Using the graphs in Fig. 6.18. and Fig. 6.20., we can find the optimal solution without taking into account the cost coefficients, and the graphs in Fig. 6.19. and Fig. 6.21. show the optimal solution for the coefficients taken into account.

There are several cases where one can apply the results of solving the optimization problem for the parameters of the communication system:



6.18. Figure: Optimal output bandwidth μ_{opt} without taking into consideration the price of resources. $\lambda = 1$, $K = 1$

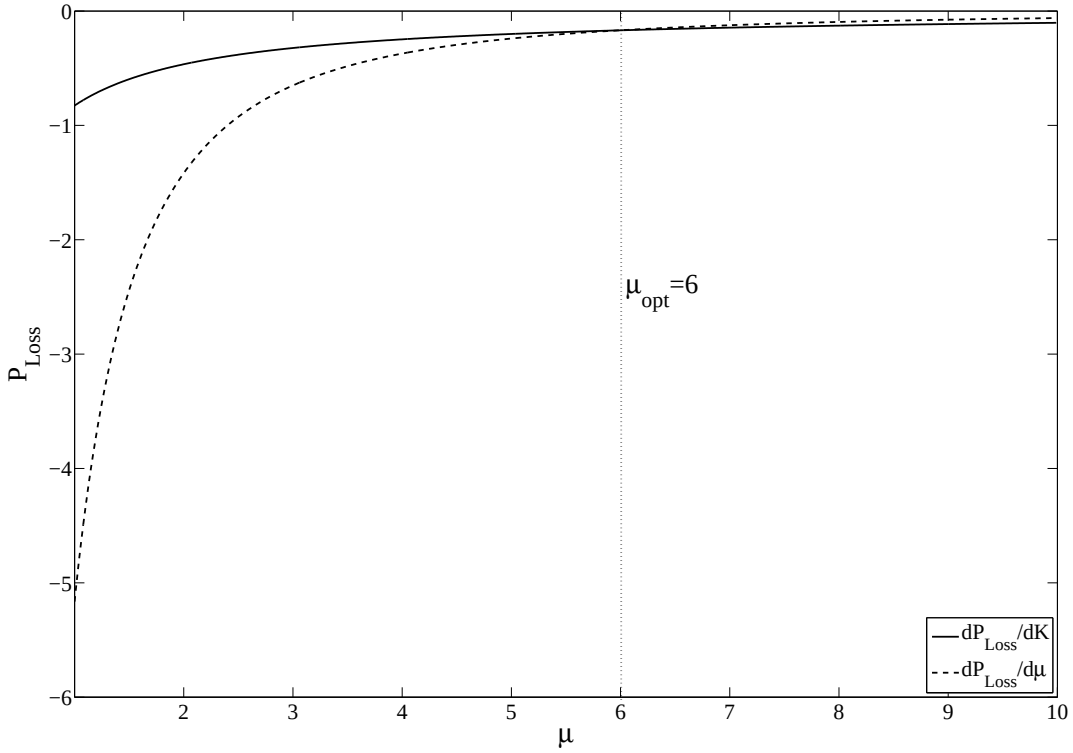
1. The operating system receives the request for additional data traffic with the *a priori* known intensity Λ_2. The system should make a decision on whether the new traffic can be admitted. In this case, we search the available resources $K^* - \Delta K$ and $\mu^* - \Delta\mu$ of the switching system for the optimal values of the buffer size and the output bandwidth, where ΔK and $\Delta\mu$ are the part of the resources already occupied by the existing traffic flows.

   We find the first solution similar to as for case described in 6.1.1. on page 65 as the intersection of the derivatives of the curves for $K \leq K^* - \Delta K$ and $\mu \leq \mu^* - \Delta\mu$. We use the obtained optimal values K_2^* and μ_2^* to find the loss probability P_{Loss} on the graph in Fig. 3.4. If it turns out to be less than the given P_{Loss} the solution is final.

   Otherwise, we search for new values of the parameters while assuming that the intensity of the arrival traffic $\Lambda > \Lambda_2$ stays within the available residual resources of the switching system.

   The process continues until the available resources are exhausted. If we fail to attain the optimal solution, the new traffic flow is not admitted.

2. Another case deals with the simultaneous request for resources from at least two com-



6.19. Figure: Optimal output bandwidth μ_{opt} with taking into consideration the price of resources. $\lambda = 1$, $K = 1$

peting traffic flows. In this case, we choose the priority traffic flow as a result of solving the optimal traffic control problem.

Then, we use the given intensity of the priority traffic flow to calculate the parameters of the resources by the algorithm from enumeration 1. If the traffic flow fails to get the required resources, we perform the second step; i.e., we analyze whether the second flow of less priority can be admitted. We apply the algorithm from enumeration 1 to perform the analysis.

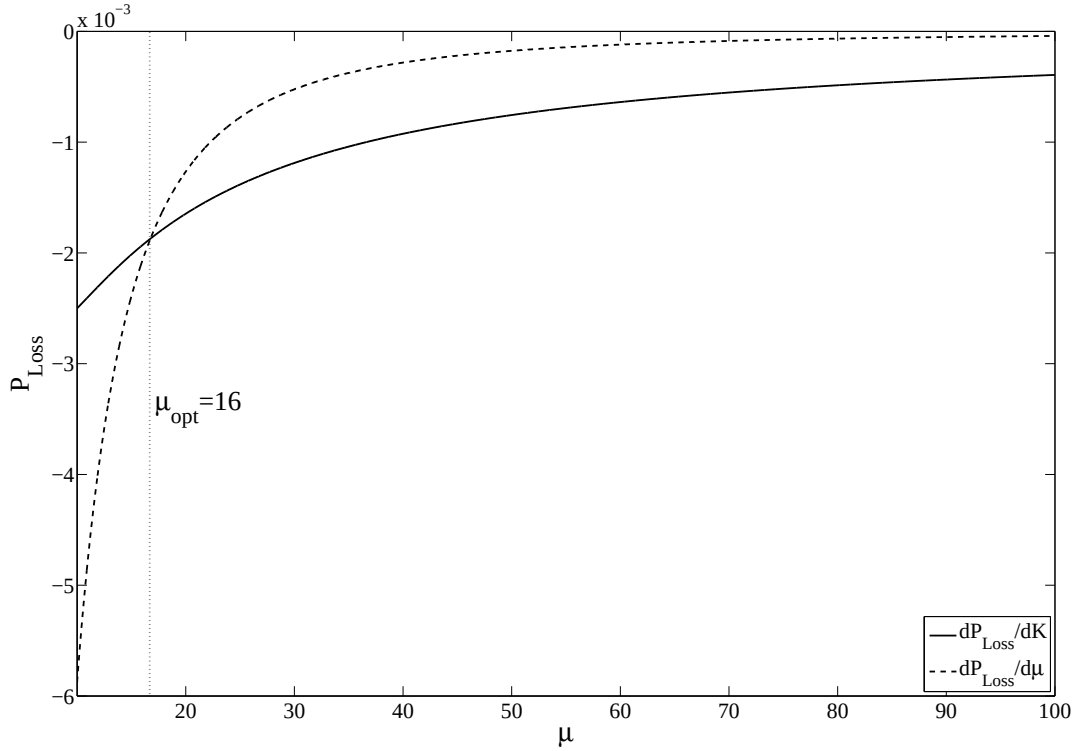
If the second traffic flow cannot get the required resources either, none of the flows are admitted.

As we can see from the graphs in Fig. 6.18. and Fig. 6.19. , the optimal solution exists even for the minimal buffer size $K = 1$, with the traffic intensity being low in this case.

If we take the cost indices into account, when c_1 is less than c_2, the value μ decreases from 9 to 6.

When the traffic intensity grows by ten times, the optimal solution is attained for the corresponding increased bandwidth and a greater buffer size $K = 10$, from 16 to 12.

Comparing the graphs obtained when we take into account the cost coefficients and when we do not, we can state that these coefficients are critical for the optimal choice of the re-



6.20. Figure: Optimal output bandwidth μ_{opt} without taking into consideration the price of resources. $\lambda = 10$, $K = 10$

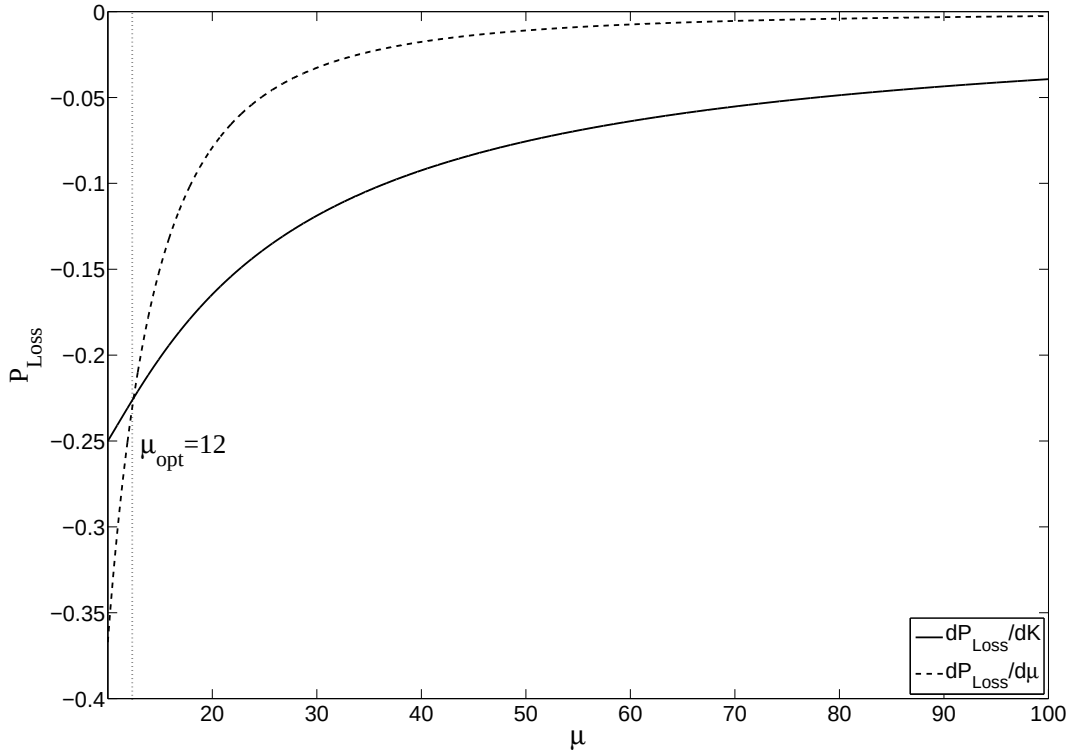
sources.

6.3. Summary

This chapter has been dedicated to the basic solutions which help the process of connection based on measurements support control to be effective. Section 6.1. of the present chapter was about making the choice of the optimal relation of buffer size and band width which allows the needed probability of packets loss at minimal costs. Such a choice of parameters has to be implemented at the stage of communication system design.

Section 6.2. considers the issues that rise during the process of MBAC mechanism functioning. In section 6.2.1. a new adaptive approach of incoming traffic measurement is proposed and the recurrent algorithm of parameter evaluation on r step of observation is presented. For the on going processes it is suggested to take observation period $T = \tau_k$, while for the sampling period Δt estimation is necessary to maximize the $I(x, y)$ depending on Θ. Initial values of the sampling and observation periods should be taken as equal to $\delta = 1/\lambda$ and $S = \frac{1}{\beta \cdot \lambda}$ respectively. β is a measurements error. It is shown that it improves the MBAC functioning as it provides more accurate measurements.

The probability that arrival rate of the aggregated traffic would be equal to peak rate of the



6.21. Figure: Optimal output bandwidth μ_{opt} with taking into consideration the price of resources. $\lambda = 10$, $K = 10$

individual source decreases with the number of individual sources. That is why it is possible to provide high network utilization while saving the QoS packet loss guaranties. Section 6.2.2. outlines the algorithm of access control that secures a high load of communication system while providing the probability of packet loss parameter of quality of service.

As the telecommunication systems have the connections that compete for free resources, the viable situation is that the flows belonging to the same priority class require connection. The method describes the optimal choice of the flow is given in Section 6.2.3. It shown that the admission priority receives the flow that rises lower utilization.

Pretty high load of system which also meets the requirements of quality of service can be gained by fulfilling the recommendations given in section 6.2.4. The main idea is the dynamic redistribution of resources with quality of service guarantee. It is shown that with the increase in the ratio ρ_1/ρ_2 , the optimal size of the buffer storage space allocated to the preemptive flow increases while the ratio $r_{opt}/K(\rho)$ remains almost unchanged.

Next chapter presents the integrated description of the algorithm of access control mechanism suggested by author (Section 7.1.) and description of MBAC model integration in OPNET framework.

7.

Simulation of Intellectual MBAC system

In order to verify the recommendation mentioned in Chapter 6., a simulation framework OPNET has been used. For the realization of MBAC in OPNET framework an algorithm has been elaborated and is presented in Section 7.1. For the realization of MBAC function the Real-Time Traffic Analyzer (RTTA) algorithm has been designed. Section 7.1. presents the complex description of the algorithm we propose for the measurement based access control mechanism. The algorithm takes into account the recommendations given above.

For the algorithm efficiency verification it was integrated as module in OPNET simulation framework. The main purpose of RTTA is traffic capturing, sniffing and storage. As it is shown in Section 7.2., the suggested data storing module is native for network.

For this reason the analysis of accumulated measurements can be performed very quickly that is the necessary requirement for the usage in real-time mode. Also, the proposed model takes into account the traffic character while saving data. It increases the accumulated statistics processing speed. Section 7.3. describes the scenarios used to test the performance of the MBAC methods that has been proposed. The chapter ends with the results of some methods proposed in this dissertation as well as methods that are already used at present. Also, the obtained results are discussed and conclusions are provided.

7.1. Intellectual MBAC Algorithm

The section describes structural system blocks of the complex of suggested algorithms for the realization of Intellectual MBAC (iMBAC) orientated for self-similar traffic. The characteristics of this traffic are inherent to the modern telecommunication systems traffic.

$iMBAC$ can be described using the following three modules:

1. Measurement module that fixes the inflow of packets to the system. At the moment of income the analysis of it's header occurs. Than a belonging to the current session gets determined. In case it is found, the information about the packet header is transferred to the third module, the one which accommodates the measurements. If the packet does not belong to any of existing connections, the packet require a new connection establishment. In this case the requirements for quality guarantees are transmitted to the second module.

2. The module that decides whether to support the incoming connection. It calculates the capability of the communication system to provide the quality guarantees necessary for the new connection without diminishing the quality of service of the existing connections. The decision is made on the base of accumulated statistics about system utilization. In case the support can be provided, the packet that initializes the connection gets forwarded and the information about it is sent to the third module.

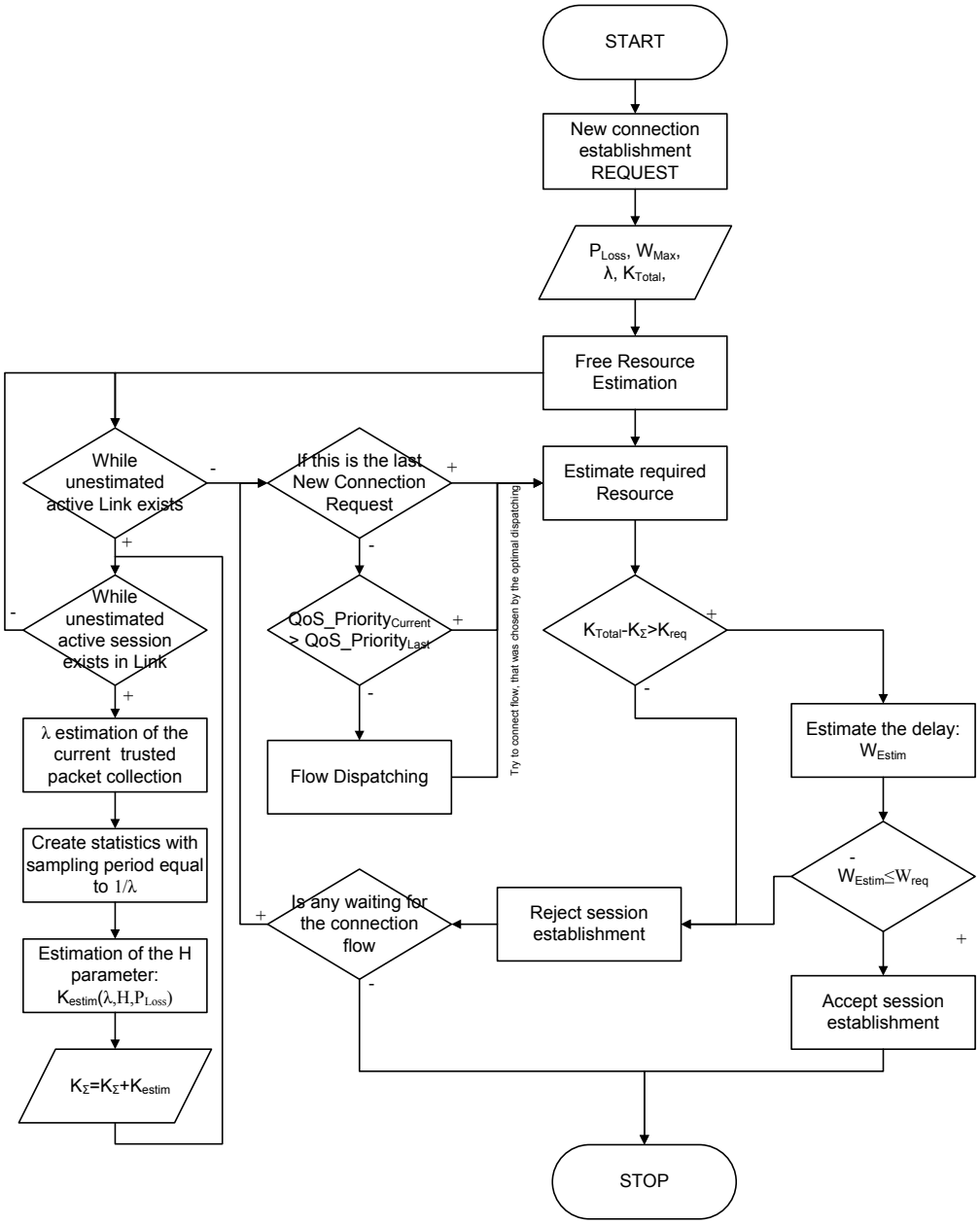
3. The measurement accumulating module. Receiving the information from the previous modules it makes a decision about the necessity of saving this information to provide the decision taking module with correct statistics. As it was described above in Section 2.2.3., only the accumulation of those data that overcome the borders of the previously accumulated data correlation interval is needed for accurate statistics.

Further goes the description of block schemes that realize the mentioned modules.

7.1.1. Decision Making Module

The block scheme of the decision making method is reflected in Fig. 7.1. During the request reception regarding the necessity to provide the support for the connection, the calculation of free available resources is determined. The purpose of the method suggested by us is the provision of such a quality of service parameter as packet loss probability (P_{Loss}). Therefore, the calculations of the needed buffer size for established connections are done. Due to heterogeneity of requirements for quality of service, the necessary buffer size is calculated separately for each connection. Calculations are performed by development of statistics and determining the parameters of active connections taking into account the existing measurements. The needed resource calculation is executed with the help of calculated intensity

of incoming packets (λ), and self-similarity coefficient (H), taking into consideration the requested quality of service. The difference of system buffer size (K_{Total}) and calculated basing on measurements produce a free amount of buffer space. The free amount of buffer size can be used for support of additional connections.



7.1. Figure iAdmission Control

If the system finds the request for the connection, then on the basis of traffic parameters and quality of service requirements the estimation of needed buffer size is performed. In case the required buffer size is smaller than the available one, a second check starts. This

additional check estimates the delays that will appear in the system during the support of the new connection. If the delay level gained does not exceed the requested level from both the established and new connections, then the connection gets accepted. It is important to mention that if the connection is supposed to be-directed, than the connection will be successfully established only in the case when the allowance for the establishment will be received for both directions.

As the telecommunications systems have connections that compete for resources, it may happen that more than one requirement for connection is sent at the same time. In this case the algorithm of optimal dispatching is enabled. Firstly, the check of possibility to provide the support for connection of the highest priority is performed. If there are requests for one class connections, the choice of connection is done in accordance to the algorithm described in chapter 6.2.3.

7.1.2. Measurements Module

Block scheme of the measurement module is shown in the Fig. 7.2. and can be described as follows.

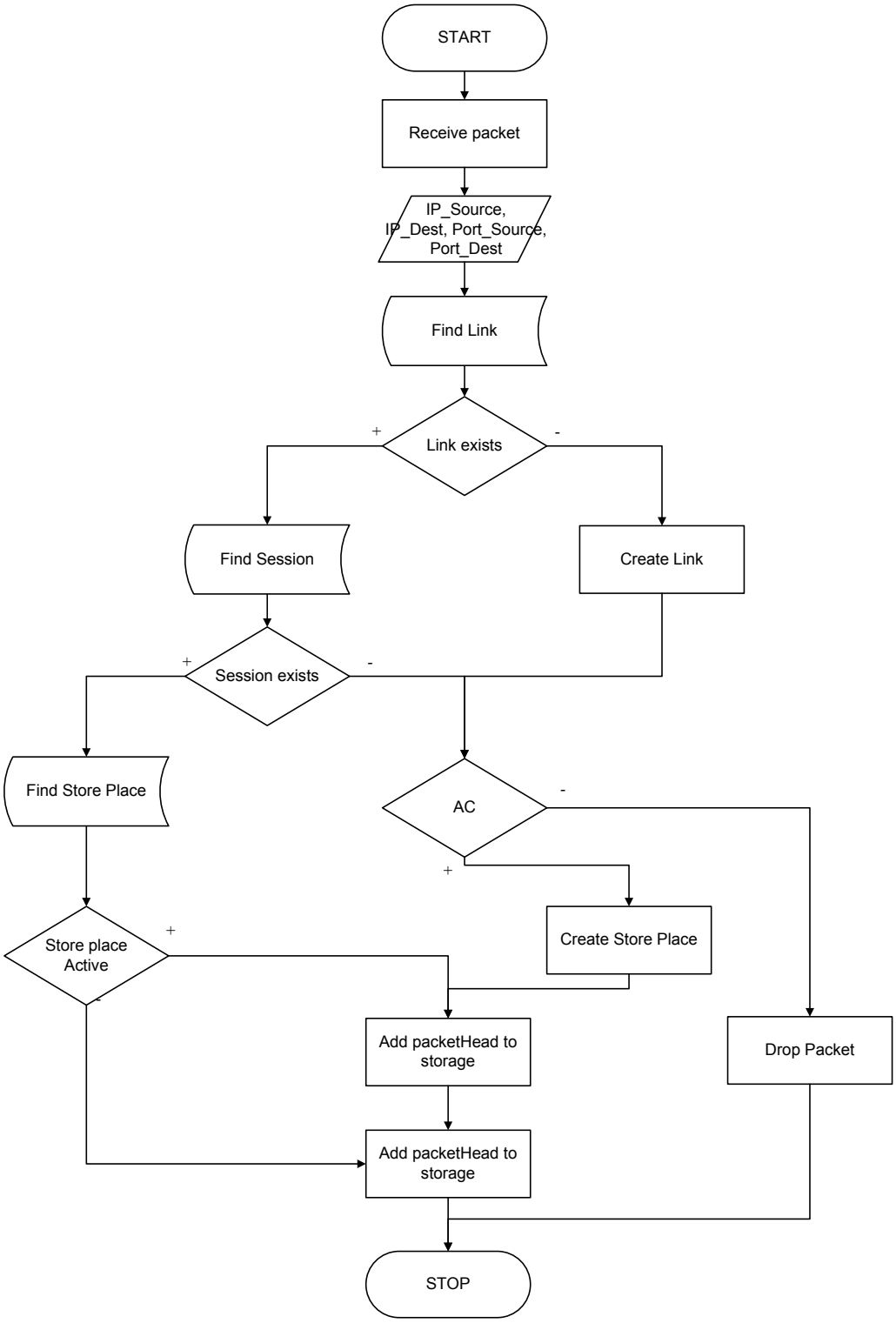
At the moment of packet incoming the system that supports $iMBAC$ service, the analysis of this packet is processed. The header reflects IP addresses and source ports of the receiver, packet size and the time of arrival.

The combination of IP addresses of the source and the receiver identify the connection of network level. The combination of IP address, source and receiver port describe the connection of transport level. It can be single or several in one connection of network level.

Connection access management described in the present paper belongs to the level of transport connection.

The following method is suggested for the purpose of sorting, saving and analyzing the measurements. When the packet joins the system and the necessary information from the header is being taken, the analysis of gained values goes on. Firstly, it is determined if that channel is known to the system, that is if the information between the mentioned source and the receiver has been transferred before. If such a channel exists in the system, the packet gets identified for its belonging to active connections going inside of the mentioned channel. If the connection is found, the packet belongs to the previously determined connection and has to be directed to the receiver which was preceded by the transfer of information about the packet to the module that saves the measurements.

In contrast, when the active connection cannot be discovered and thus the packet is initializing the new connection. If this is the initializing packet, than besides the useful information it includes the requirements for quality of service. While such a packet gets into the system, the decision about connection support has to be made. Likewise, if the incoming packet does not belong to the previously known channel, than this packet also is initializing

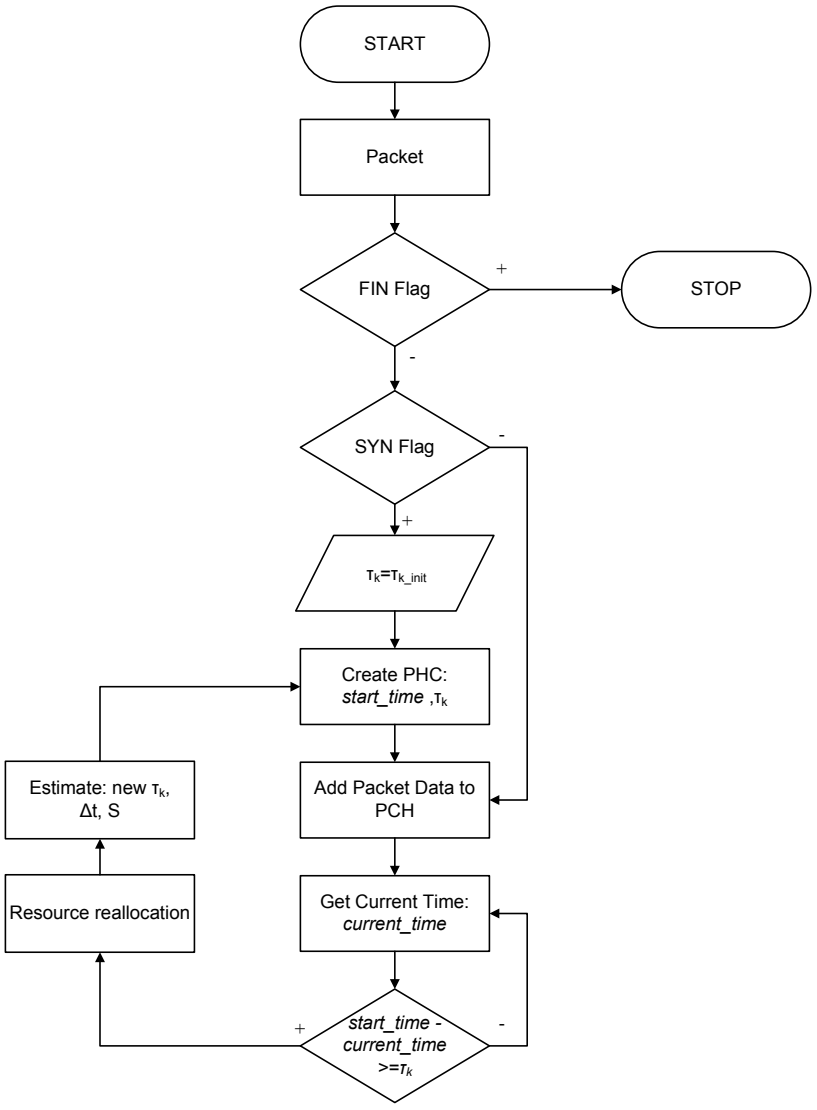


7.2. Figure iMeasurements

for the new connection and has to implement access control towards it. If the connection cannot be supported, the packet does not get transferred to the receiver and just get deleted from the system.

If the decision about the connection is positive, the space is allocated for keeping the measurements of this connection while keeping the information about the packet and the packet itself if forwarded to the receiver. It is necessary to clarify that in case bi-directed connection cannot be supported by the corresponding connection, the space allocated for the initializing connection becomes free.

7.1.3. Measurements accumulating module



7.3. Figure iPHC

The present section describes the third component of the $iMBAC$ service module. Data accumulated by this module get used for the development of statistics for the further calculation of traffic parameters. The block- scheme of the module is shown in Fig. 7.3. Besides the measurements accumulation the module allows to optimize the statistics and measurements process, as well as raise the speed of statistics analysis. The main idea was thoroughly described in Section 2.2.3.

All the measurements are kept not as a one set of data, but they are devided for definite time intervals. The intervals are chosen so that the statistical parameters in the interval remain unchanged and get estimated as the correlation interval. The interval will be named as a collection in further references. For example, when the initializing packet of the allowed connection enters the module, the first collection gets created. It belongs to a certain connection which is settled inside of the definite channel. The length of the first collection is set in accordance to the necessary error of measurements. Thus, to secure the inaccuracy that does not exceed 1%, the collection has to contain at least 100 gained packets. Hence, the interval of the first collection is chosen as being equal to "specified measurement error * average time of packet arrival".

The average arrival time of the packet is $1/\lambda$, where λ is the requested parameter of quality of service. At the time of packets arrival which belongs to the session of the collection, information about the packet gets allocated to the current collection. At the end of the interval of the first collection on the basis of gained measurements statistics is generated and correlation interval is calculated. During the calculated interval the data about the incoming packets do not get accumulated. Such an interval will be further called the passive interval. Respectively, the interval when the statistics is generated will be called active. At the end of the passive interval a new collection with active interval equaling the same calculated interval is created. In the case when the active interval of the collection does not imply the collection of sufficient quantity of measurements, for the purpose of generation of valid statistics with 1% error, the interval can be calculated as follows.

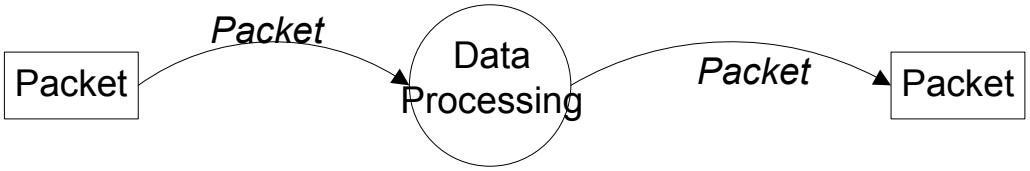
The interval is chosen so the collection includes at least 100 packets basing on the intensity of the incoming packets measured at the previous active interval. Estimating the intensity of the incoming packets in the previous collection it is possible to take into account even the pre-previous collections.

This approach and its influence on the accuracy of the calculated parameters is described in the previous Section 6.2.1.

7.2. Real-Time Traffic Analyzer for Measurement-Based Admission Control

This section presents the model of the traffic analyzer for real-time applications. It consists of two cross-dependent sub-modules: traffic measurement and traffic estimator. The model presented is characterized by low system overheads for real-time traffic parameters estimation and can be used either within IntServ or DiffServ approaches.

It is obvious that the packet is the main data unit in the network that influences network traffic characteristics and, consequently, the decision about admission. Fig. 7.4. depicts a packet moving throughout our data analyzer called *Data Processing*.



7.4. Figure A packet moving throughout the data analyzer

The packets belonging to the same established connection between a source and a destination are united into a session. The term of session implies that two computers establish a connection for a ”conversation” and allows larger messages to be handled with the provision of error detection and recovery.

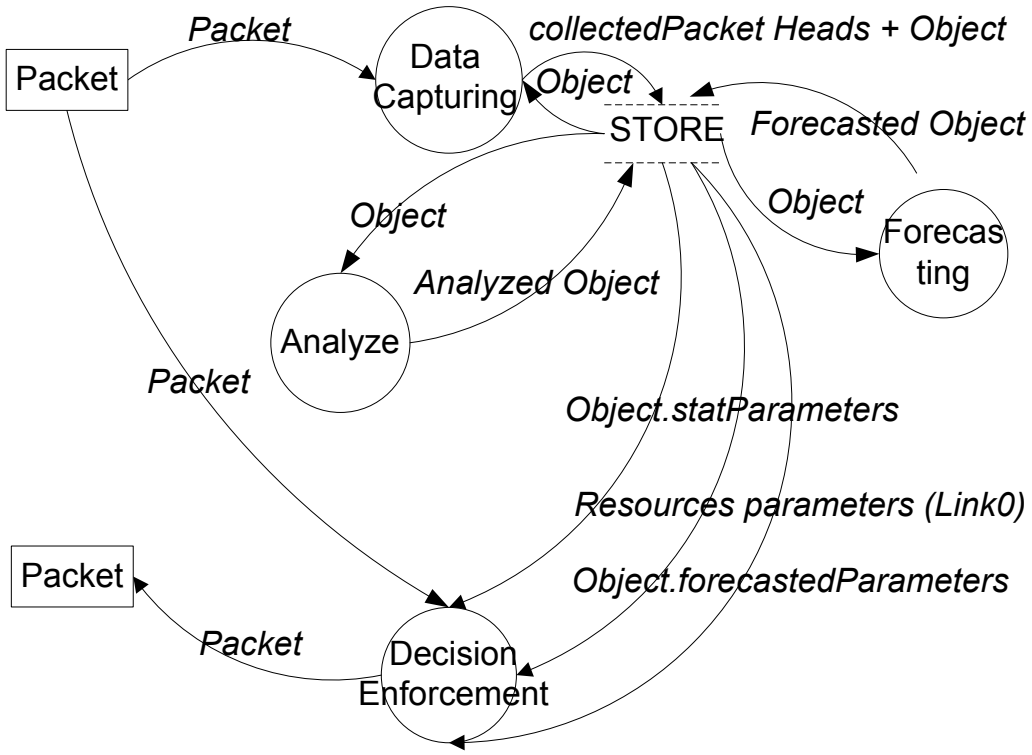
Multiple sessions can be established between a source and a destination. The aggregated traffic of the sessions is united into the link. To obtain the data storage model, we propose to use data flow diagram and select the essential processes and data.

Data are being processed as described in Fig. 7.5. Fig. 7.5. depicts a major data flow diagram of the data processing. This diagram presents a holistic view of the MBAC model. *Data Capturing* deals with measurement of the established flows. The responsibility of the *Analyze* is to estimate traffic parameters of the measured traffic. The function of a *Forecasting* module is to predict the network traffic. The *Decision Enforcement* module implements admission control functions. Admission decision is enforced basing on the estimated and forecasted traffic parameters. Decision enforcement regarding the packet affects if the packet will be transmitted or dropped.

In the proposed model a noticeable reduction of the overheads is achieved by *Data Capturing* process.

The key component in the model is *STORE* object. The object contains all the measured data that pass the system. The *STORE* structure is discussed below.

To estimate traffic parameters truthfully, the accurate measurement should be applied. Frequently the measurement unit collects all data going through it. As it was elaborated in



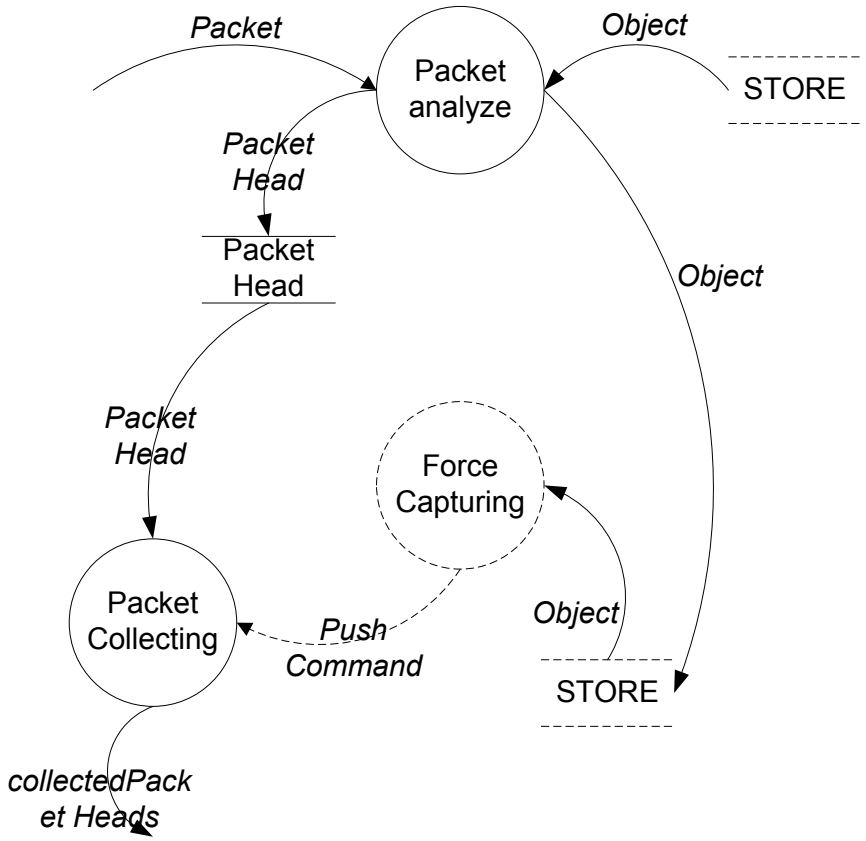
7.5. Figure Data processing DFD

the previous section (Section 7.1.) we propose to decrease system overheads by reduction of capturing procedures while the reliability to estimated parameters stays on the same level. We suggest creating interrelated *Data Capturing* and *Analyze* components in the way that estimated parameters have impact on the measurement process.

The capturing process is depicted on the Fig. 7.6. diagram. Receiving the packet the *Packet Analyze* module analyzes the head of the data. Based on the information from the packet head the *Packet Analyze* requests information from the main storage *STORE* about capturing state for that session where packet belongs. If the packet belongs to the data session with ”captured” status then the information about the packet will be collected. Otherwise the information of the packet will be ignored.

Only the information from the packet head is going to be stored in packet collection within the half of correlation interval: packet length, session identification, and traffic classification and arrival time.

The responsibility of *Force Capturing* component is to control corresponding time and correlation interval. If for the some session the time is equal to the half of its correlation interval *Force Capturing* produce *Push Command* to start measurement process. At time moment $t = T+\tau_k$ the measured traffic is analyzed in the *Data Estimator* and new correlation interval is estimated.



7.6. Figure Data capturing DFD

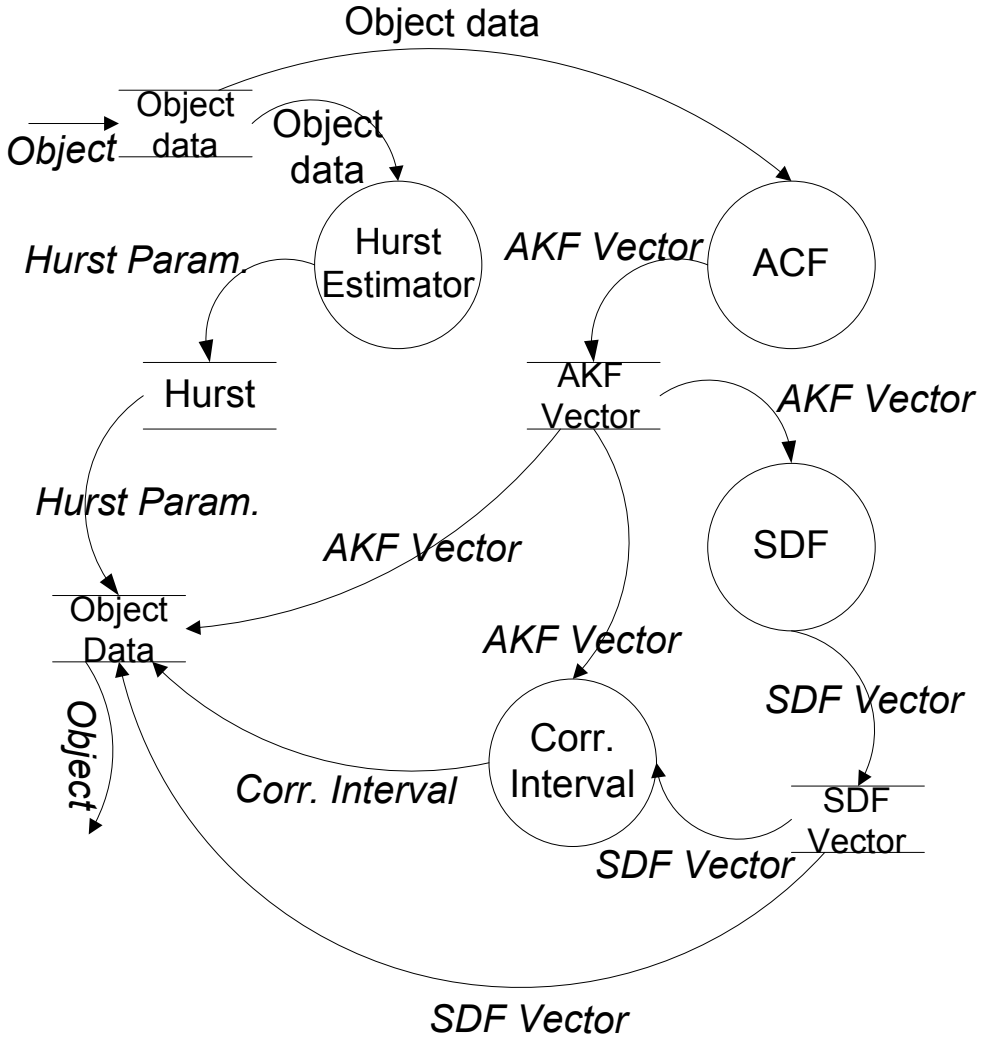
The model was designed taken into account that the parameters estimation could be applicable to the session or to the whole link. It corresponds to the *Object* that is transmitted to the Data Estimator and is depicted in the Fig. 7.7.

The following parameters are estimated for the data collected: AKF - autocorrelation function, SDF - spectral density function, CorrInterval - correlation interval for the input data.

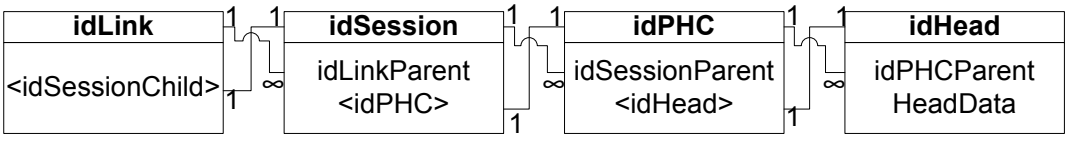
To estimate the degree of the self-similarity for the measured network traffic the Hurst Estimator is used.

The simplified picture of the model of data estimation representing Hurst parameter estimation, autocorrelation and spectral density functions, as well as a correlation interval is shown at Fig. 7.7. The inter-arrival rate, peak rate, mean arrival time, etc. are also estimated but not represented for the purposes of the picture clarity.

We suggest organizing *STORE* data storage element in the native network way where a variety of packets belongs to one session, whereas variety of session belongs to the link. The network equipment could have more than one link connection. Fig. 7.8. presents our



7.7. Figure Data analyzing DFD



7.8. Figure The main storage class

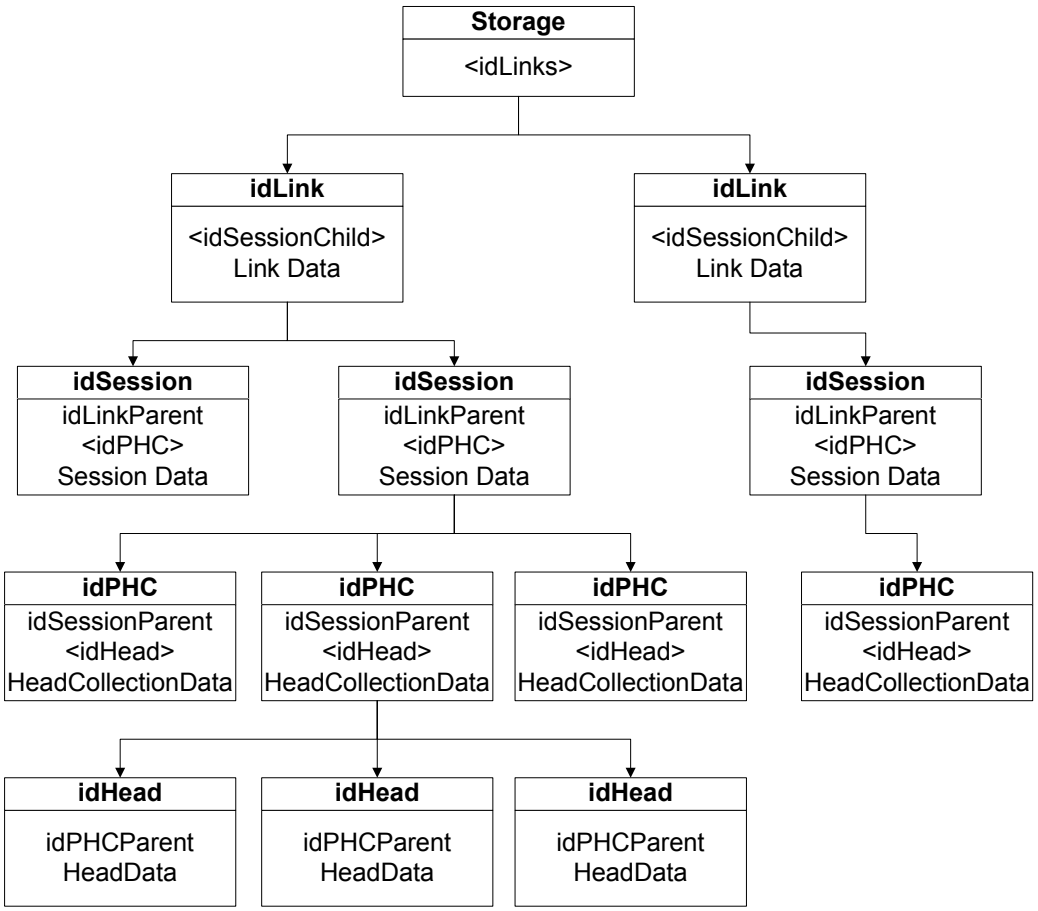
proposal for the STORE storage component.

Packet heads of the same session and on the same correlation interval are collected to the same *Packet Head Collection (PHC)*. Variety of *PHC* with the same session ID fully describes the characteristics of a session. And finally, the variety of the sessions with the same link ID can fully describe link characteristic.

It is necessary to mention, that cross-reference between objects in the storage element

plays important role. The child object could notify the parent object about its changes, and, similarly, the parent object is able to get information from the child object.

An example of object-oriented implementation is presented in Fig. 7.9. This way of storage organization provides great flexibility, however we have to recognize it is not the most efficient implementation. It has to be emphasized that the model is developed for the MBAC algorithm and is implemented in network simulation environment where the side procedure like storage element access time, does not influence the simulated objects. Out of these parameters the number of measurements and the confidential interval for the estimated parameters are the most significant.

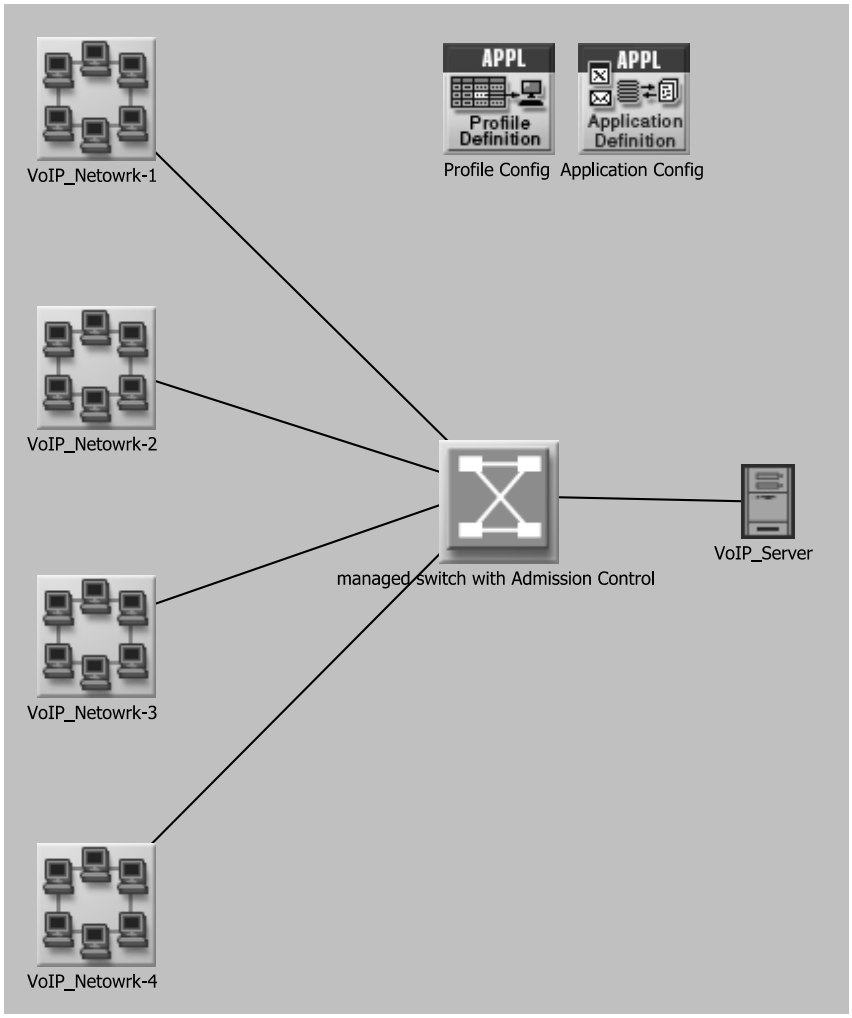


7.9. Figure The storage hierarchy

Conclusions This section presents the design of the traffic estimator model for the real-time applications. The model is designed to be implemented into MBAC algorithm. The main objective of the model is applicability for IntServ and DiffServ service guaranties mechanisms. Another contribution is degradation of the systems overheads related to network traffic capturing process.

7.3. Scenarious for Modeling in OPNET

This chapter is about the scenarios used for the verification of the basic recommendations from Chapter 6. Figure 7.10. presents the structural scheme of the scenario network where clients get connected to the server through the managed switch with implemented admission control.



7.10. Figure The OPNET Project for VoIP Scenario

The clients use VoIP applications in that scenario. Applications parameters are set in accordance to the description in Section 2.1.3. TCP transport protocol has been used in the scenario. 60 clients are united into 4 sub-networks via unmanageable switches that are connected to the managed switch. Analysis of traffic generated by clients in this scenario show a high rate of self-similarity - $H = 0.82$, that negatively influences network performance.

The present chapter reviews the results of the application of the elaborated managing algorithms that are intended to boost the performance of the network in general.

The admission control scenario mentioned above is realized on the managed switch. For testing of the effectiveness of recommendations for the managing MBAC algorithm suggested in Chapter 6., switch parameters, in particular bandwidth, were chosen as understated on purpose.

Central switch throughput for VoIP scenario is set as $3000 packet/sec$. Buffer size is calculated using $P/M/1/K$ model (for the detailed description see Section 3.1.) for $\rho = 0.9$ and self-similarity parameter $H = 0.8$.

The current version realized the following estimator and policy algorithms:

- Simple policy Policy of AC-AR (P-AR) based on the system load coefficient (described in Section 5.4.)
- Utilization parameter is estimated according to the Instantaneous Utilization (E-IU) method from Section 5.3.

The model realizes 4 measurement algorithms:

- Static - measurement interval is set before launching and remains unchanged all the time during the simulation. Time Window method is described in Section 5.2.
- Dynamic - measurement interval is chosen depending on the intensity of the incoming flow (described in Section 6.2.1.)
- Second dynamic method where measurement interval equals the correlation interval. This method description can be found in Section 7.2.
- Third dynamic method is the advanced version of the previously mentioned one. The improvements regard system load decrease related to the measurements. For more details, see Section 7.2.

Importantly, the modality of MBAC model utilized allows adding another new algorithms effortlessly.

7.4. Simulation Results

The current section presents the results proving some of the propositions made earlier on. It provides the comparative analysis of the network parameters without admission control and with different methods of admission control.

The method effectiveness will be evaluated according to the following criteria:

- Network utilization
- Packet losses
- Delays

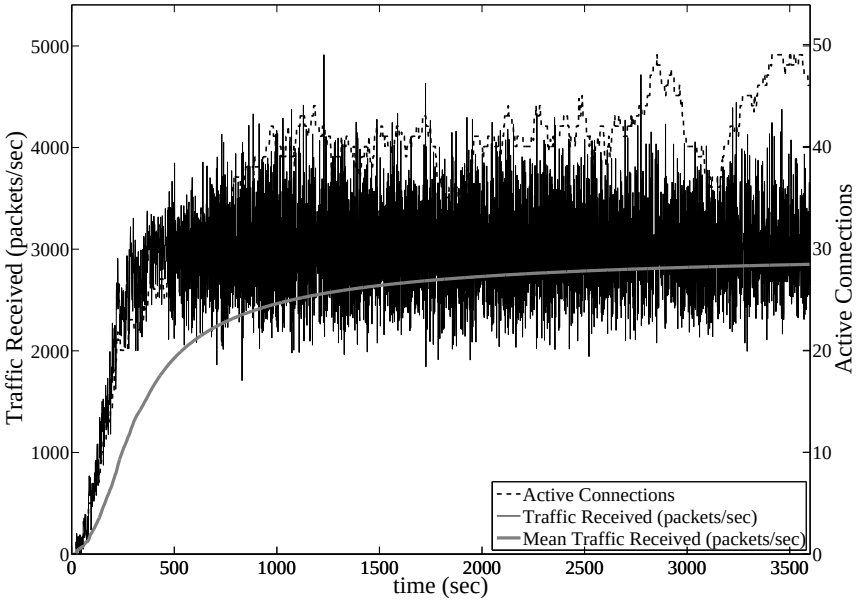
7.4.1. Without Admission Control

Fig. 7.11. - Fig. 7.16. present the results obtained during the simulation of the scenario of VoIP application without the managed control. To compare the modelling results the following parameters are used:

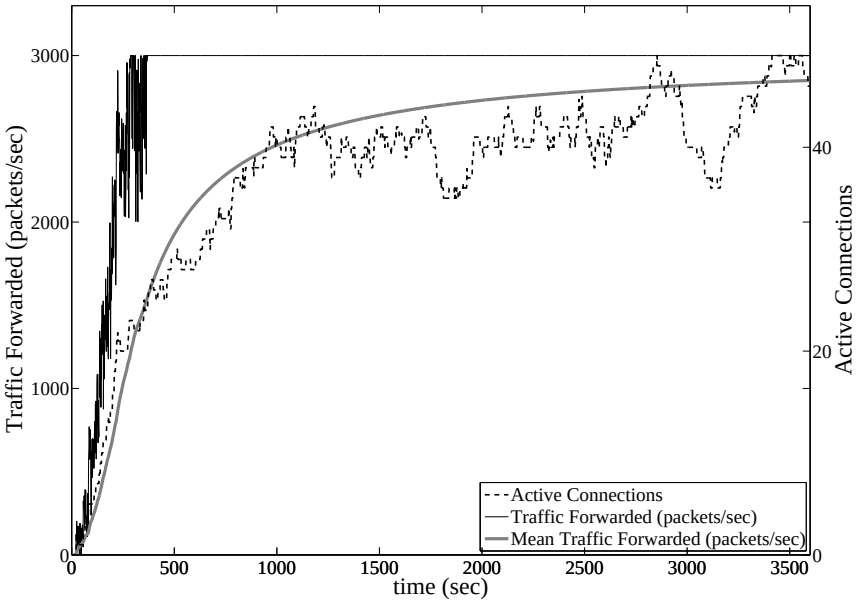
- Active Connection Count - total number of active TCP connections on the VoIP server.
- Traffic Recieved (packets/sec) - amount of traffic received in packet per sec.
- Traffic Forwarded (packets/sec) - number of packets forwarded by this switch per sec. Packets are forwarded on the appropriate output port, based on the configuration of the spanning tree and the address information in the filtering database.
- Size - this statistic represents the number of packets stored in the queue or subqueue.
- Delays - this statistic represents instantaneous measurements of packet waiting times in the queue or subqueue.
- Overflows - this statistic represents a running count of the number of times the queue has had to reject packet insertions due to lack of capacity.
- Packet end-to-end Delay (sec) - the total voice packet delay, called "analog-to-analog" or "mouth-to-ear" $delay = network_{delay} + encoding_{delay} + decoding_{delay} + compression_{delay} + decompression_{delay}$
- Jitter (sec) - If two consequetive packets leave the source node with time stamps t1 & t2 and are played back at the destination node at time t3 & t4, then: jitter = (t4 - t3) - (t2 - t1) Negative jitter indicates that the time difference between the packets at the destination node was less than that at the VoIP server.

The graphs reflect that the number of established sessions during the simulation time has reached 47, and the graph gained from the users exceeds the throughput of the central switch (Fig. 7.11.). Fig. 7.12. distinctly shows the situation while connecting 23 clients, the switch reaches the saturation mode that causes a severe increase of the queue length and delays. Fig. 7.13. and 7.14. show the queue reaction when the switch reaches the saturation mode. Such a queue behaviour automatically causes the increase in time delay for the voice application (Fig. 7.15.). One can see, the moment the switch gets saturated, the delay time increases under exponential law.

Taking into consideration the fact that for voice data transmission the acceptable delay is $200msec$, it can be argued the connection number > 23 leads to the quality of service disruption at this parameter. Fig. 7.16. shows that in this mode Jitter is located in the acceptable range, under $100msec$, thus it can be considered acceptable.

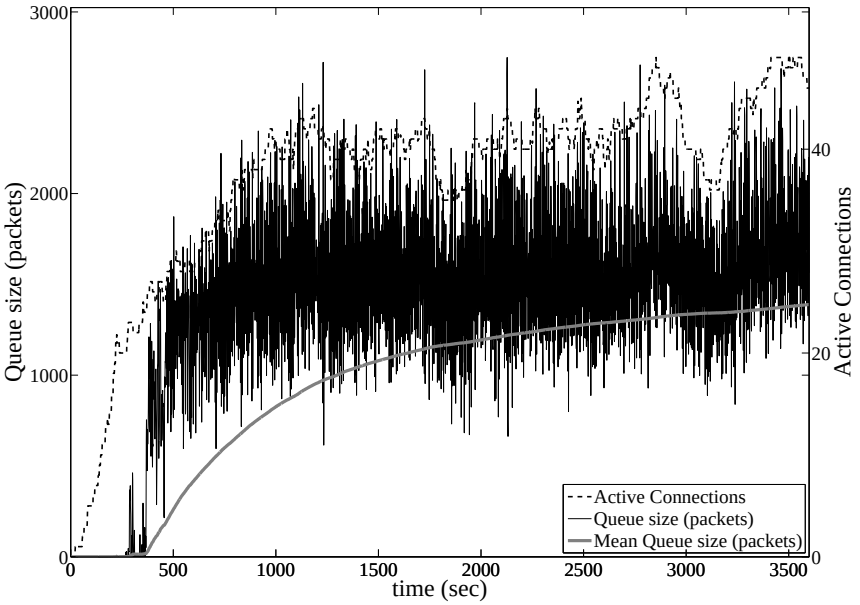


7.11. Figure Traffic Recieved with managed switch in scenario without AC

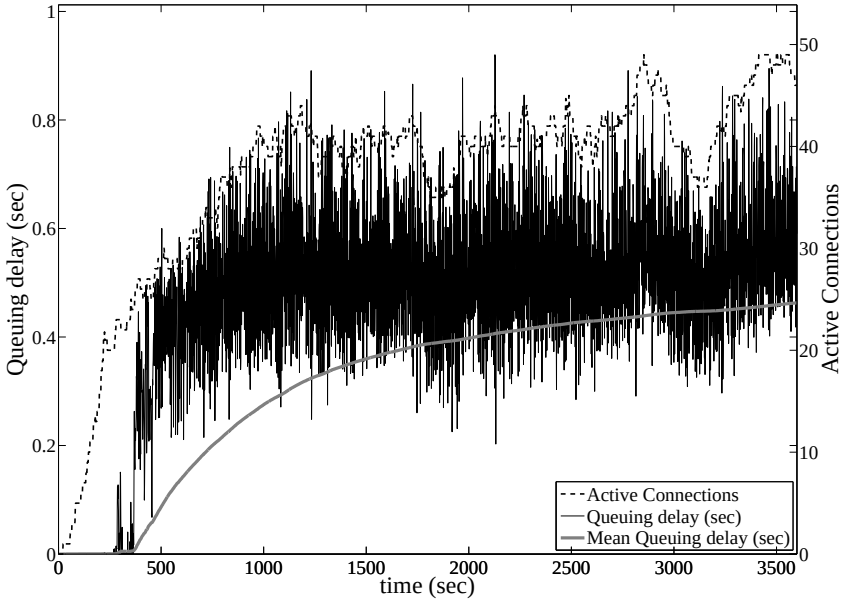


7.12. Figure Traffic Forwarded with managed switch in scenario without AC

Node functioning in the saturation mode also critically influences P_{Loss}. Sufficient buffer size is intended for the purpose of P_{Loss} decreasing. As the buffer size is limited, during the long time of the node being in the saturation mode, the buffer may get fully loaded. Herewith, the newly incoming packet will be lost. Fig 7.17. reflects this situation. Packet loss probability in this scenario is $P_{Loss} = Packets_{Rejected}/Packets_{Total} = 0.08$. It exceeds the necessary quality of service as for VoIP this parameter should $P_{Loss} < 0.01$.

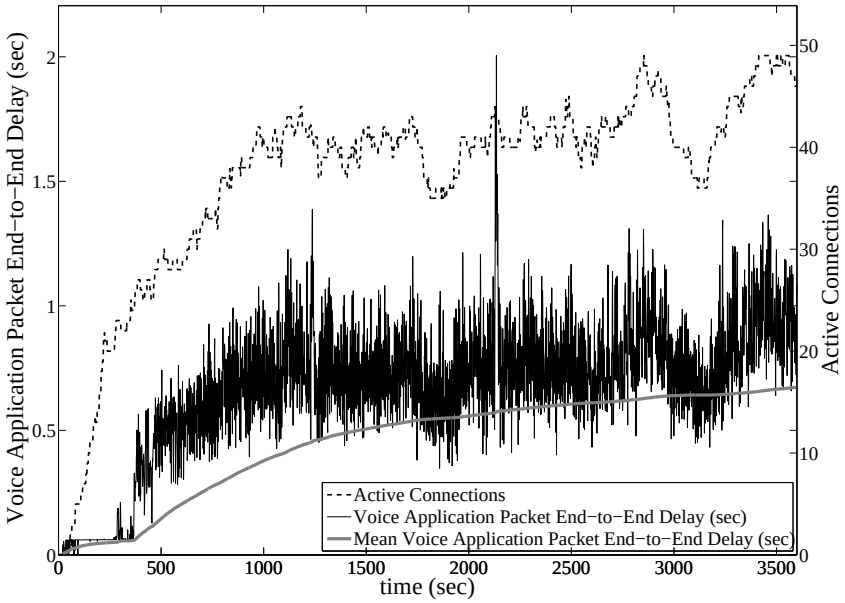


7.13. Figure Queuing Delay in managed switch in scenario without AC

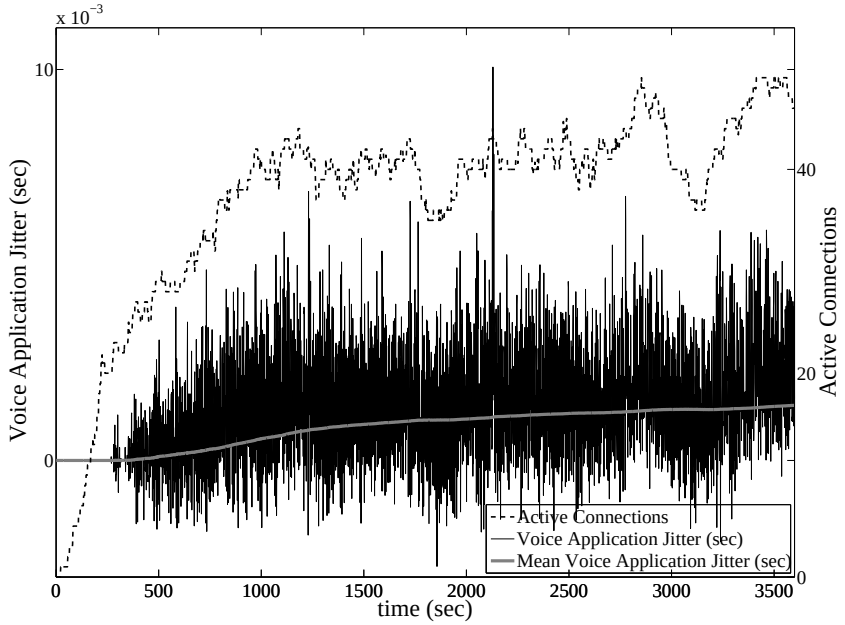


7.14. Figure Queuing Delay in managed switch in scenario without AC

The results prove the necessity to ensure admission control. The next sections will provide the results of simulations when admission control has been used using some of the recommendations given above.



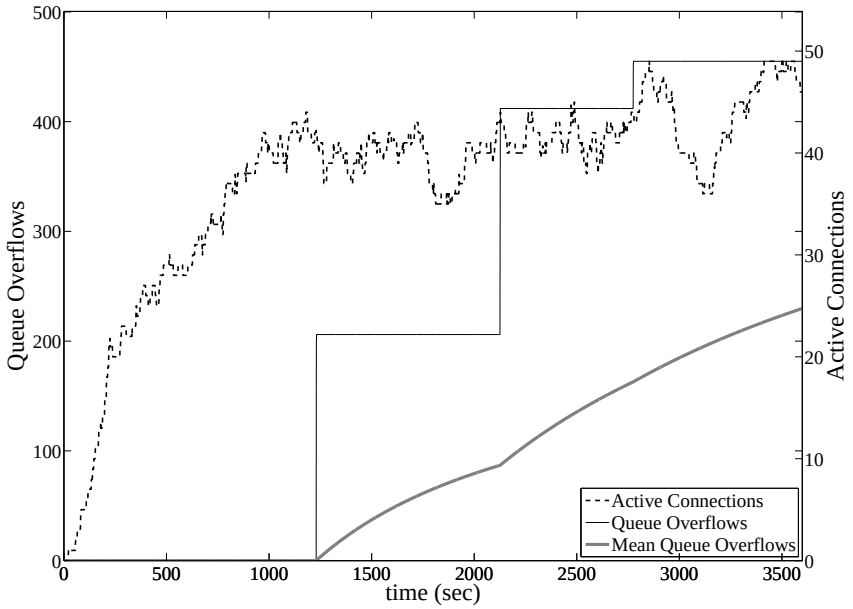
7.15. Figure Voice Application Delay in scenario withotu AC



7.16. Figure Voice Application Jitter in scenario without AC

7.4.2. Simple Admission Control

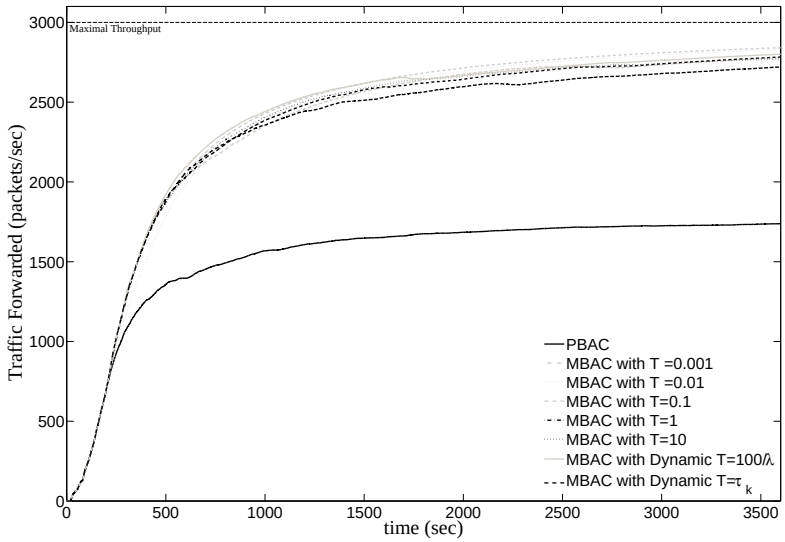
Fig. 7.18. and 7.19. present the modelling results and dynamic of changes of the average throughput and packet delay in the queue at the central managed switch. Simulated were the following admission control methods with simple policy P-AR and E-IU estimator:



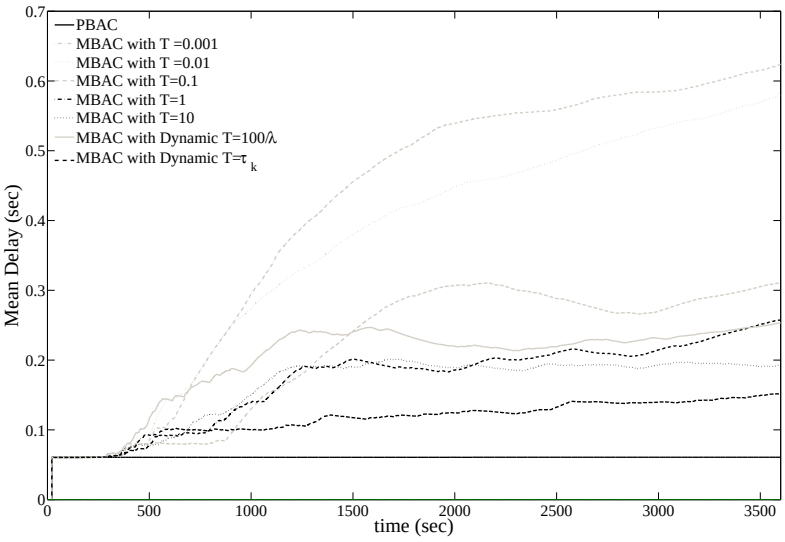
7.17. Figure Queue Overflows in managed switch in scenario without AC

- parametric-based admission control
- simple admission control based on measurements with measurement window $T = 0.001(sec)$
- simple admission control based on measurements with measurement window $T = 0.01(sec)$
- simple admission control based on measurements with measurement window $T = 0.1(sec)$
- simple admission control based on measurements with measurement window $T = 1(sec)$
- simple admission control based on measurements with measurement window $T = 10(sec)$
- simple admission control based on measurements with measurement window $T = 100/\lambda(sec)$
- simple admission control based on measurements with measurement window $T = \tau_k(sec)$

On the basis of the family of curves represented on Fig. 7.18. the opinion, existing in literature, that the parametric admission control provides low network load can be confirmed.



7.18. Figure The mean value of throughput for different AC parameters



7.19. Figure The mean value of delay in queue for different AC parameters

This curves family corresponds to the simple measurements based admission control and results in a similar high network load rate irrespective of measurement window (T).

For the qualitative analysis of the curves from Fig. 7.18. it has to be taken into account that the acceptable delays level for VoIP applications is $200msec$. Out of many methods being explored, the necessary QoS level is provided by PBAC and MBAC with measurement window that equals $T = \tau_k$. Putting together the analysis of Fig. 7.18. and 7.19. it can be argued that MBAC with measurement window that equals $T = \tau_k$, is the best algorithms for

admission control from all previously discussed.

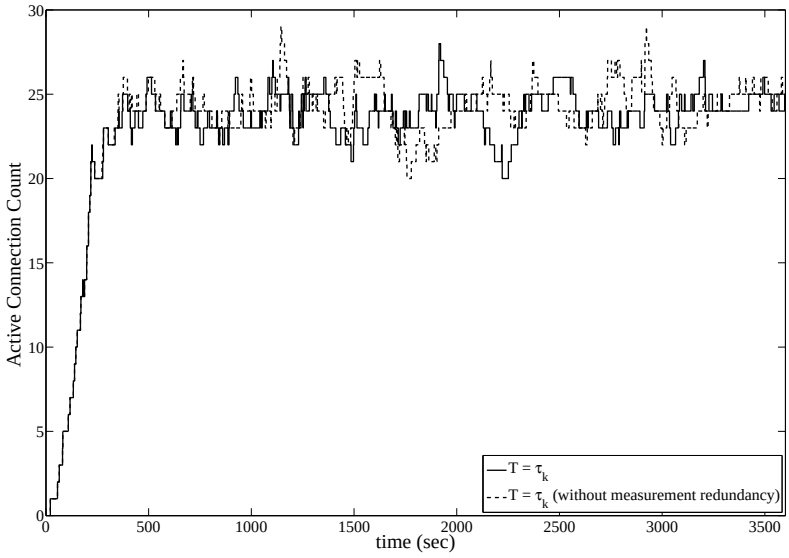
7.4.3. Simple Admission Control without measurements redundancy

In the previous section it has been proved that correctly chosen measurement interval is extremely important for measurement based admission control. It has been mentioned the managing algorithm gains the best result when choosing the measurement interval that equal correlation interval $T = \tau_k$.

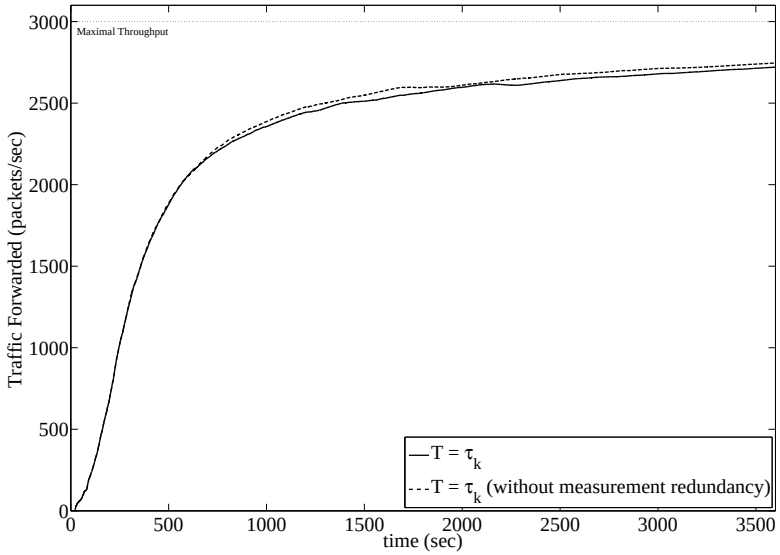
In Chapter 2.2. it has been mentioned that modern traffic is strongly correlated.

On the basis of that in Section 7.2. a solution has been proposed allowing decreasing the usage of system directly related to the measurement process. In the current section the comparative results of network performance under the usage of managing algorithm with measurement window equaling correlation interval with the same algorithm but decreasing the redundancy of measurements are described.

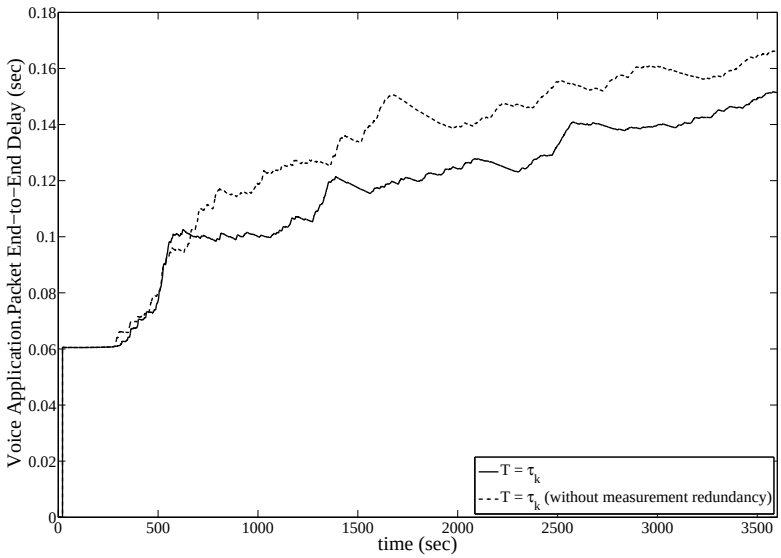
Fig. 7.20. and 7.21. show that for both cases the number of active sessions and utilization are very close. From Fig. 7.22. we can conclude that whilst the measurement redundancy decreases, the delay increases, however it remains acceptable. Evaluating the simulation results, we can argue that such an approach reduces the number of necessary measurements for $\approx 17\%$.



7.20. Figure The Active Session Number



7.21. Figure The Traffic Forwarded by managed switch



7.22. Figure The End-to-end Delay of Voice Application

7.5. Summary

The chapter reviewed the managing algorithm, measurement based admission control. *Decision Making*, *Admission Control*, *Measurements* and *STORE* module realization algorithms have been described in details in section 7.1. Data flow diagrams for this algorithm are given in Section 7.2.

The effectiveness of the proposed algorithm has been tested in OPNET modeling framework. For modeling according to Fig. 7.10. the network with managed switch has been

created. The algorithm has been realized as additional modules integrated in the switch. The performance has been tested on the basis of the scenario described in Section 7.3., where the traffic is created by the users of VoIP applications.

It has been shown that for measurement based admission control the critical factor affecting the accuracy of the decision making, is the measurement window. The last Section 7.4. presents qualitative comparisons of the network performance with different admission control methods measurement window. The results match the hypothesis that the best results can be gained while adaptive to traffic parameters window measurement is applied. The best window measurement equals correlation interval $T = \tau_k$. Also, simulation results prove that the costs of the correlated traffic, related to measuring process, can be decreased. That is supported by the proposed solution from section 7.2.

8. Conclusions

The work presents the recommendations for development of managing algorithms that provide the increase of network performance while ensuring quality of service guarantees.

The specialty of the proposed algorithms is that they take into account the bursty character of the traffic. The queuing model with the bursty incoming flow have been adopted. It is presented that this model can be used for description of systems with self-similar behavior of incoming flows. The initially presented model have been improved and hence starts to provide analytical expressions for queuing parameters such as the mean queue length, average awaiting time, packet loss probability, buffer size needed for provision of the requested packet loss probability. It allows to solve optimization task of choosing the structural unit at the stage of designing the communication system under the specified P_{Loss} while- minimizing the expenses.

To reduce the overload of structural elements and the whole system is employing the algorithms that manage the flows. The managing algorithm is dedicated to provision of QoS guarantees to network clients. The present work is devoted to the recommendations about design of managing algorithms. There are several problems in design of managing algorithms:

- Functional
- Structural
- Technological

The functional problems are related to the development of decision making methods of management. Hereby, management is considered to be the decision making process regarding the necessary QoS guarantees. In the present work it is suggested to use CAC for the QoS guarantee fulfillment. The CAC that gets the information about utilization directly from the performed measurements of the existing flows has recently become popular in provision of statistical QoS and called MBAC.

The intensity of incoming packet flow and self-similarity parameter of incoming traffic need to be estimated so that calculation of outgoing channel load and buffer capacity is possible. The measurements of the flow parameters are taken during the discrete sampling periods and is analyzed by accumulating data about it during the observation periods. Solutions for observation and sampling period values are proposed for measuring of already established connections and also for on-going estimation. In addition, the recommendations regarding the decrease of the system load which is directly related with the measurement process can be found in the research.

The structural problem caused by maximum usage of the limited network resources. In case of CAC, it means the admission of the large number of flows maintaining QoS guarantees. It has been claimed that the bursty traffic negatively influences the network performance. Also, multiplexing several flows with bursty character into one produces a significant gain, as the probability that the used bandwidth represent the sum of peak rates of the individual flows is very low. Therefore, the network utilization can be advanced by fulfilling quality guarantees due to the benefit gained from combining the flows which can be used for connection additional flows. In other words, we can have more available buffer memory for newly arrived connections.

In order to manage the flows effectively the flows, its classes have been introduced and the corresponding priorities. The task of resources allocation is complicated in the case when two or more flows of the same priority class arrive to the entrance. The connection of the flow is possible only in case of sufficient number of resources. The best usage of the limited resources can be reached by the algorithm of optimal dispatching proposed by the author. The work provides the algorithm of choosing the connected flow in the situation when there are no enough resources for connection of all the same priority simultaneously incoming flows. The work also present the algorithm for determination of the optimum buffer storage capacity for competing flows.

The technological problem of the managing algorithm effects the cost of the managing system. Reallocation studies and recommendations of the work concern such structural units as buffer size and bandwidth between the flows in the process of system functioning. The buffer size and output bandwidth allocated for the particular data flow according to its class of service act as such resources. Allocating an additional bandwidth or increasing the buffer size used in the communication system for the particular traffic flow, we reduce the loss

probability. However, this makes the communication system's costs grow. The work has presented recommendations how the mentioned algorithm may be applied to minimize the costs during the exploitation of the communication system.

The proposed recommendations were tested using the logical and structural MBAC scheme which has been realized in modeling framework OPNET. The modeling results have confirmed the recommendations elaborated previously.

Future Work The recommendations from Chapter 6. and results of the present work give significant scope for future work and especially in wireless environment.

A

Table of pre-Estimated Memory Volume

ρ \ H	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
0.5	24	30	36	48	74	179	2981	7.2288E+09
0.525	27	34	41	55	87	226	4757	7.2357E+09
0.55	30	38	47	63	104	289	7815	7.2430E+09
0.575	33	43	53	74	125	375	1.3262E+04	7.2506E+09
0.56	37	49	61	86	152	494	2.3333E+04	7.2586E+09
0.625	41	56	71	102	187	664	4.2749E+04	7.2669E+09
0.65	47	64	83	122	234	912	8.2002E+04	7.2757E+09
0.675	53	75	98	147	296	1285	1.6578E+05	7.2849E+09
0.7	61	87	117	180	381	1863	3.5608E+05	7.2947E+09
0.725	71	104	141	224	503	2797	8.2102E+05	7.3049E+09
0.75	83	124	172	285	680	4376	2.0592E+06	7.3158E+09
0.775	99	152	216	370	951	7200	5.7170E+06	7.3273E+09
0.8	120	189	276	496	1384	1.2606E+04	1.7990E+07	7.3394E+09
0.825	148	243	365	690	2121	2.3885E+04	6.6327E+07	7.3523E+09
0.85	189	322	502	1010	3477	5.0188E+04	3.0081E+08	7.3661E+09
0.875	250	447	731	1584	6255	1.2146E+05	1.8097E+09	7.3808E+09
0.9	350	667	1153	2746	1.2879E+04	3.6065E+05	7.5904E+09	7.3964E+09
0.925	536	1110	2070	5583	3.2854E+04	1.4796E+06	7.5771E+09	7.4132E+09
0.95	966	2256	4699	1.5209E+04	1.2401E+05	1.0945E+07	7.5644E+09	7.4313E+09
0.975	2575	7458	1.8953E+04	8.4825E+04	1.2203E+06	3.4138E+08	7.5521E+09	7.4507E+09

A1. Table Table of pre-estimated buffer size (K) dependence on utilization (ρ) and H parameter for loss probability $P_{Loss} = 10^{-3}$

d \ H	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
0.5	31	40	47	62	96	233	3890	9.5313E+09
0.525	35	44	54	71	114	295	6207	9.5383E+09
0.55	39	50	61	83	136	377	1.0196E+04	9.5456E+09
0.575	43	56	70	96	164	489	1.7300E+04	9.5532E+09
0.56	48	64	80	113	199	645	3.0432E+04	9.5611E+09
0.625	54	73	93	133	245	867	5.5751E+04	9.5695E+09
0.65	61	84	108	159	305	1190	1.0693E+05	9.5783E+09
0.675	70	98	128	192	386	1676	2.1616E+05	9.5875E+09
0.7	80	114	152	235	498	2430	4.6425E+05	9.5973E+09
0.725	93	135	184	293	656	3648	1.0704E+06	9.6075E+09
0.75	109	162	225	372	888	5707	2.6844E+06	9.6184E+09
0.775	129	198	281	483	1241	9389	7.4522E+06	9.6298E+09
0.8	156	247	360	647	1806	1.6437E+04	2.3448E+07	9.6420E+09
0.825	193	316	476	901	2766	3.1142E+04	8.6448E+07	9.6549E+09
0.85	246	420	655	1318	4534	6.5427E+04	3.9205E+08	9.6687E+09
0.875	326	584	953	2065	8155	1.5832E+05	2.3585E+09	9.6833E+09
0.9	456	870	1504	3580	1.6789E+04	4.7007E+05	9.8547E+09	9.6990E+09
0.925	699	1447	2698	7278	4.2822E+04	1.9283E+06	9.8670E+09	9.7158E+09
0.95	1259	2941	6124	1.9823E+04	1.6161E+05	1.4263E+07	9.8797E+09	9.7533E+09
0.975	3356	9719	2.4699E+04	1.1054E+05	1.5901E+06	4.4483E+08	9892991709	9.7339E+09

A2. Table Table of pre-estimated buffer size (K) dependence on utilization (ρ) and H parameter for loss probability $P_{Loss} = 10^{-4}$

d \ H	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
0.5	39	49	59	77	119	288	4799	1.1834E+10
0.525	43	55	66	88	141	364	7656	1.1841E+10
0.55	48	62	75	102	168	465	1.2576E+04	1.1848E+10
0.575	53	70	86	119	202	603	2.1337E+04	1.1856E+10
0.56	59	79	99	139	246	796	3.7532E+04	1.1864E+10
0.625	67	90	115	164	302	1069	6.8753E+04	1.1872E+10
0.65	76	104	134	196	376	1468	1.3186E+05	1.1881E+10
0.675	86	120	158	237	476	2068	2.6654E+05	1.1890E+10
0.7	99	141	188	290	614	2998	5.7243E+05	1.1900E+10
0.725	114	167	226	361	809	4500	1.3197E+06	1.1910E+10
0.75	134	200	278	459	1095	7039	3.3096E+06	1.1921E+10
0.775	160	245	347	596	1530	1.1578E+04	9.1874E+06	1.1932E+10
0.8	193	305	445	798	2227	2.0269E+04	2.8907E+07	1.1945E+10
0.825	239	390	587	1111	3411	3.8398E+04	1.0657E+08	1.1958E+10
0.85	304	517	808	1625	5591	8.0667E+04	4.8328E+08	1.1971E+10
0.875	402	720	1175	2547	1.0055E+04	1.9518E+05	2.9072E+09	1.1986E+10
0.9	563	1073	1854	4414	2.0699E+04	5.7948E+05	1.2157E+10	1.2002E+10
0.925	862	1784	3327	8973	5.2791E+04	2.3770E+06	1.2170E+10	1.2018E+10
0.95	1553	3625	7550	2.4437E+04	1.9921E+05	1.7580E+07	1.2182E+10	1.2036E+10
0.975	4137	1.1980E+04	3.0445E+04	1.3625E+05	1.9599E+06	5.4828E+08	1.2196E+10	1.2056E+10

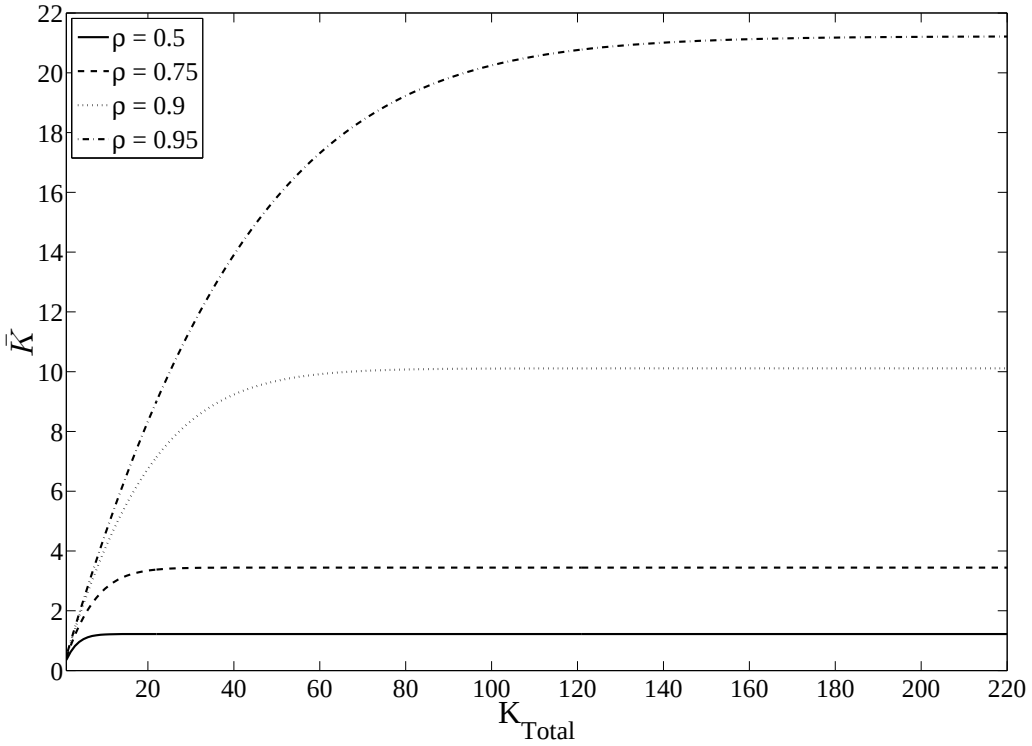
A3. Table Table of pre-estimated buffer size (K) dependence on utilization (ρ) and H parameter for loss probability $P_{Loss} = 10^{-5}$

d \ H	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
0.5	46	58	70	91	142	343	5708	1.4137E+10
0.525	51	65	79	105	168	433	9106	1.4143E+10
0.55	57	73	90	121	200	553	1.4957E+04	1.4151E+10
0.575	63	83	102	141	240	718	2.5375E+04	1.4158E+10
0.56	71	94	118	165	292	947	4.4632E+04	1.4166E+10
0.625	79	107	136	195	359	1272	8.1755E+04	1.4175E+10
0.65	90	123	159	233	448	1747	1.5680E+05	1.4183E+10
0.675	102	143	187	282	566	2459	3.1692E+05	1.4193E+10
0.7	117	168	223	345	730	3565	6.8061E+05	1.4202E+10
0.725	136	198	269	430	962	5351	1.5691E+06	1.4213E+10
0.75	160	238	330	545	1302	8370	3.9348E+06	1.4224E+10
0.775	190	291	413	709	1820	1.3768E+04	1.0923E+07	1.4235E+10
0.8	230	363	529	949	2648	2.4100E+04	3.4366E+07	1.4247E+10
0.825	284	464	698	1321	4056	4.5654E+04	1.2669E+08	1.4260E+10
0.85	361	615	961	1932	6648	9.5906E+04	5.7451E+08	1.4274E+10
0.875	478	856	1398	3028	1.1955E+04	2.3204E+05	3.4560E+10	1.4289E+10
0.9	669	1276	2205	5248	2.4609E+04	6.8889E+05	1.4498E+10	1.4304E+10
0.925	1025	2122	3955	1.0668E+04	6.2759E+04	2.8257E+06	1.4485E+10	1.4321E+10
0.95	1846	4310	8976	2.9050E+04	2.3682E+05	2.0898E+07	1.4472E+10	1.4339E+10
0.975	4917	1.4242E+04	3.6192E+04	1.6197E+05	2.3298E+06	6.5173E+08	1.4460E+10	1.4358E+10

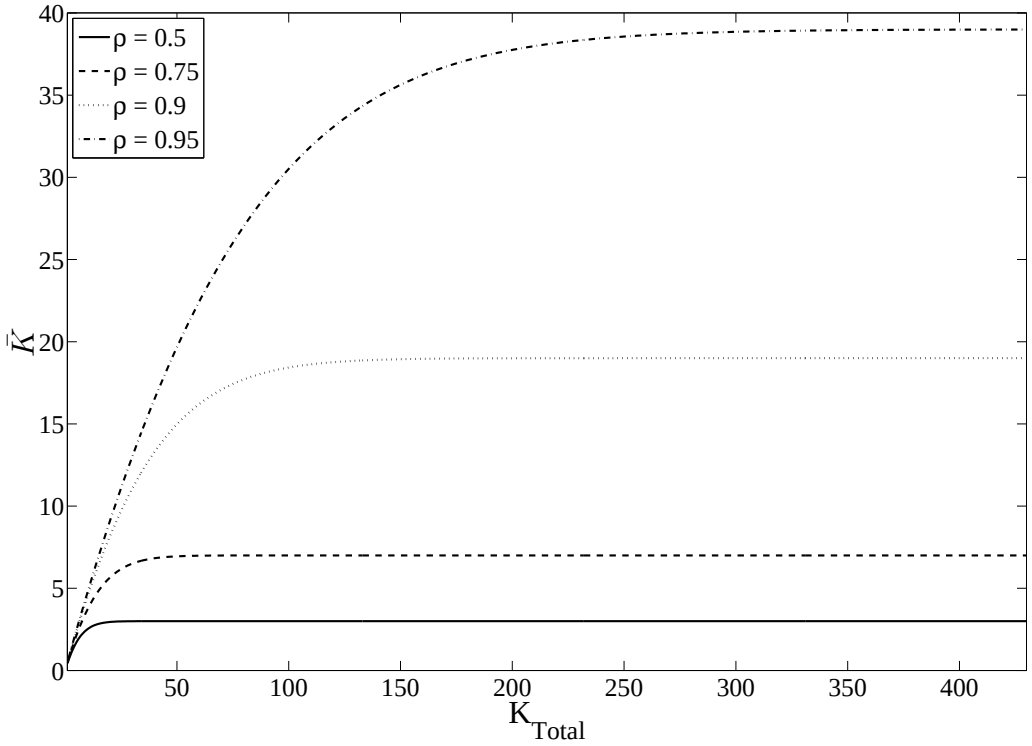
A4. Table Table of pre-estimated buffer size (K) dependence on utilization (ρ) and H parameter for loss probability $P_{Loss} = 10^{-6}$

B

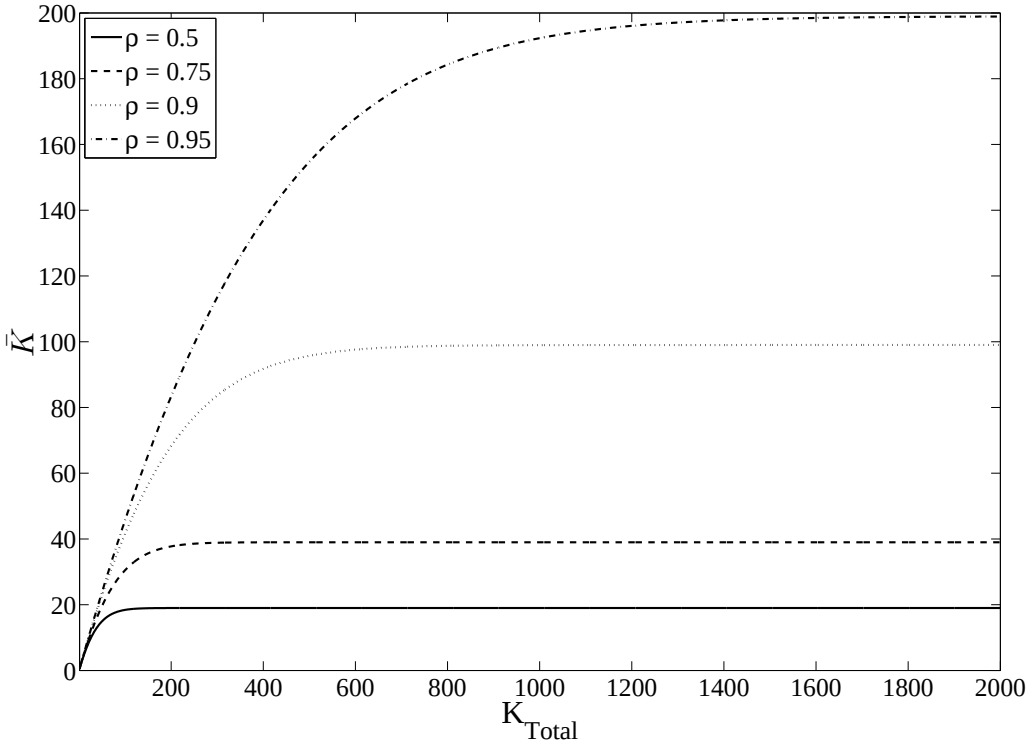
The mean number of Jobs in System for the $M^X/M/1/K$ queue model



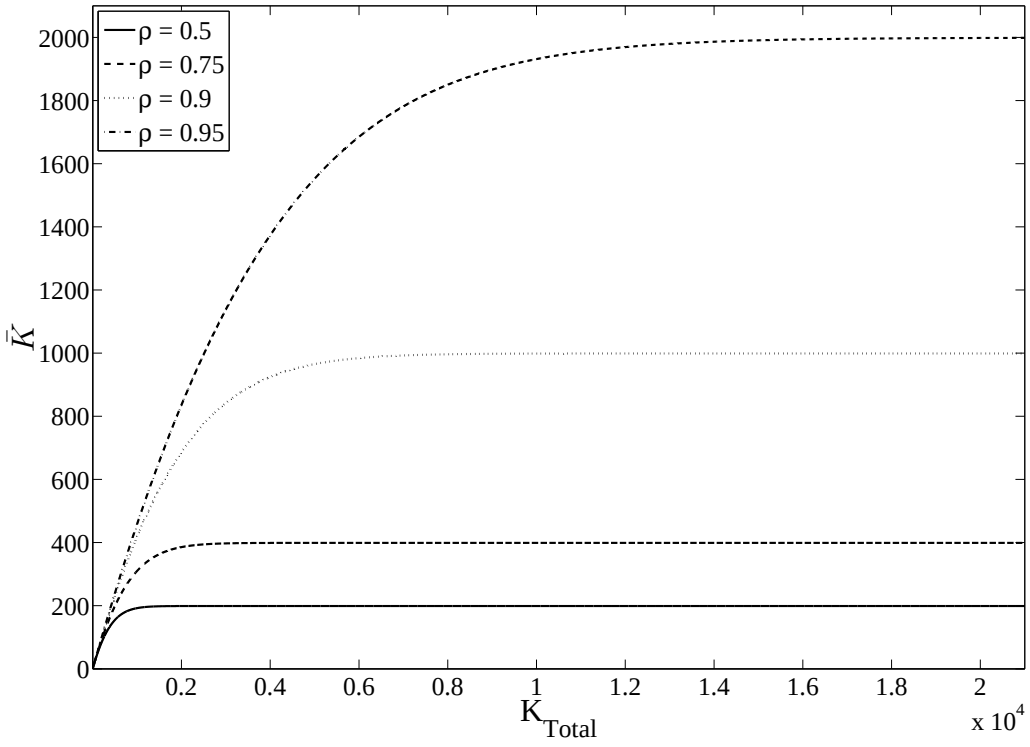
B1. Figure: The mean number of Jobs in System for the $M^X/M/1/K$ queue model with $\alpha = 0.1$



B2. Figure: The mean number of Jobs in System for the $M^X/M/1/K$ queue model with $\alpha = 0.5$



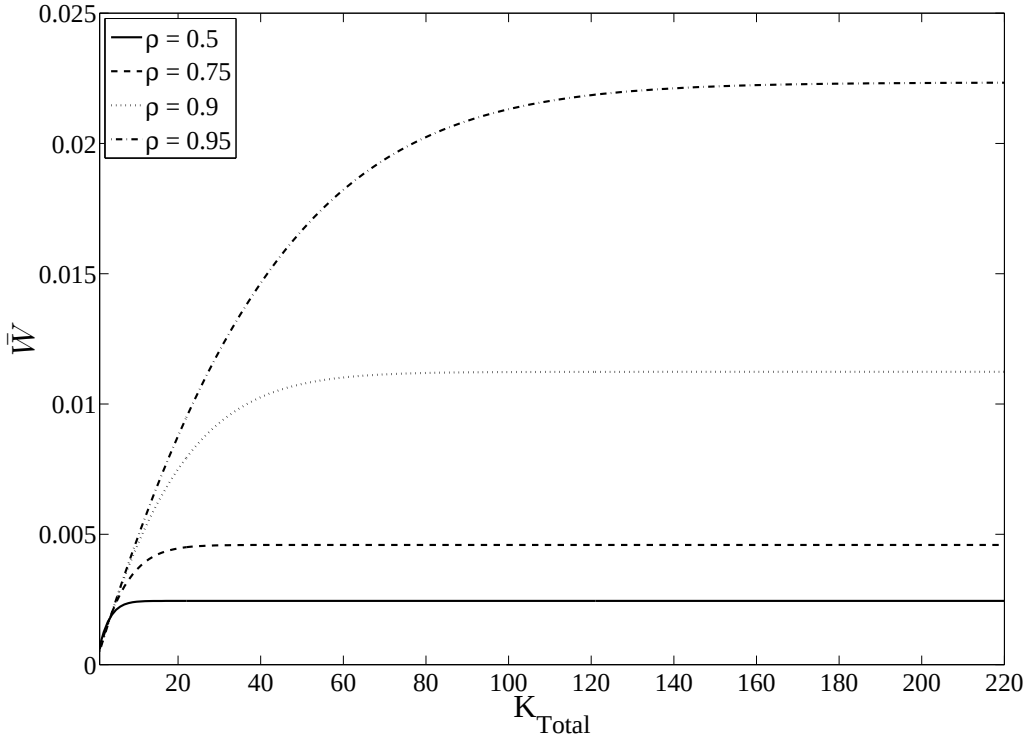
B3. Figure: The mean number of Jobs in System for the $M^X/M/1/K$ queue model with $\alpha = 0.9$



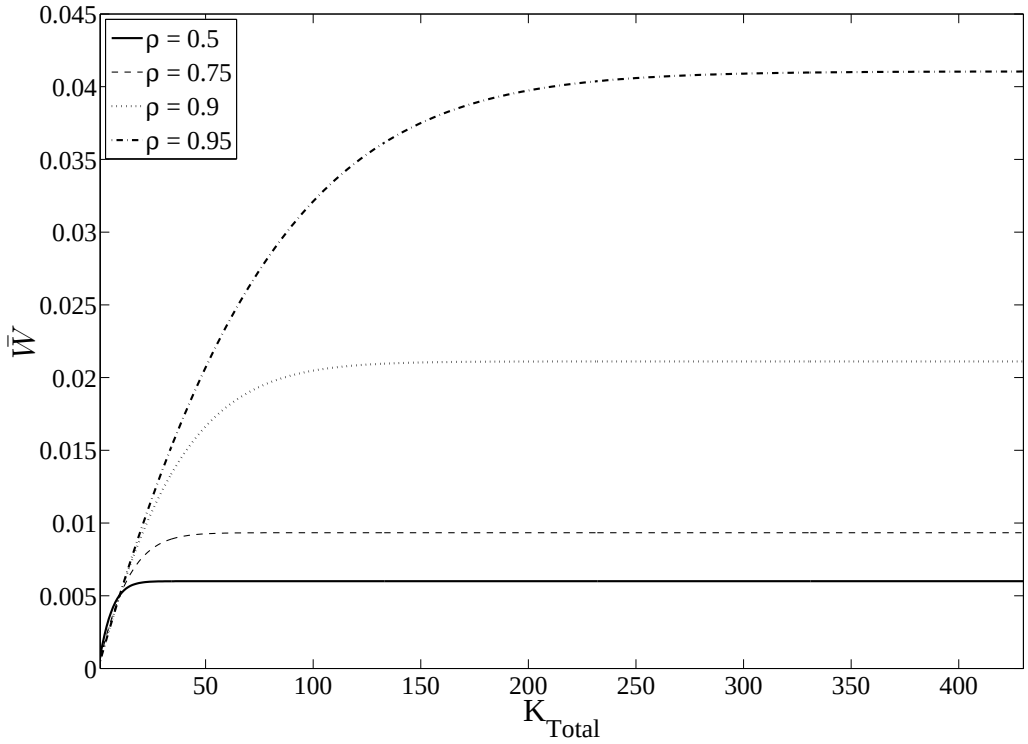
B4. Figure: The mean number of Jobs in System for the $M^X/M/1/K$ queue model with $\alpha = 0.99$

C

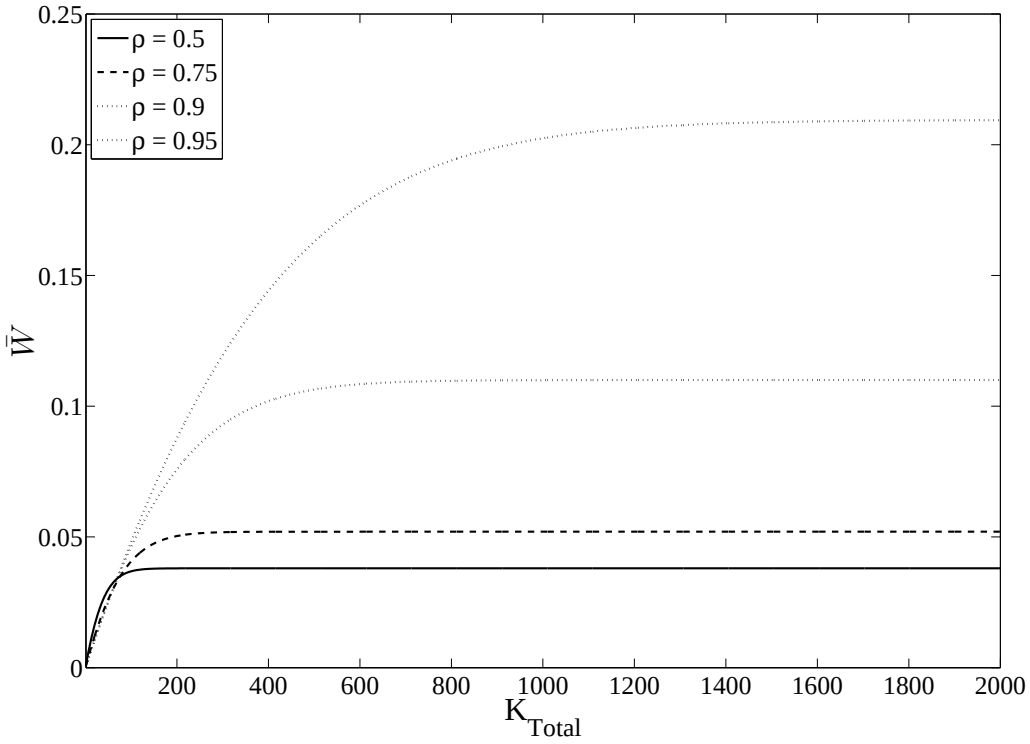
The mean waiting time of the Job in System for the $M^X/M/1/K$ queue model



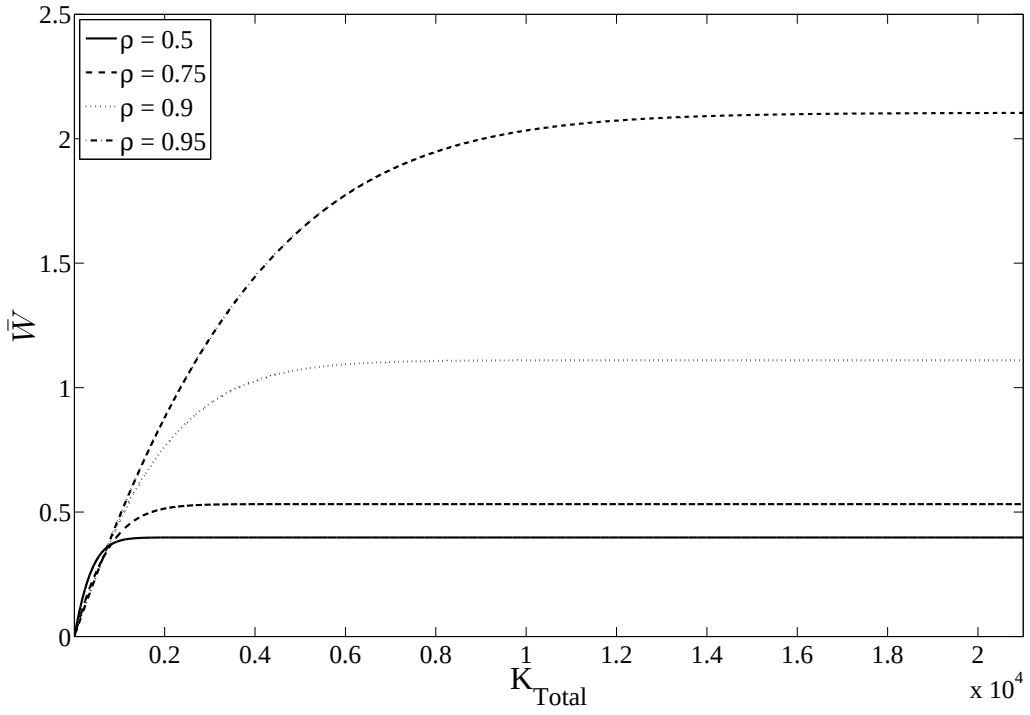
C1. Figure: The mean waiting time of the Job in System for the $M^X/M/1/K$ queue model with $\alpha = 0.1$



C2. Figure: The mean waiting time of the Job in System for the $M^X/M/1/K$ queue model with $\alpha = 0.5$



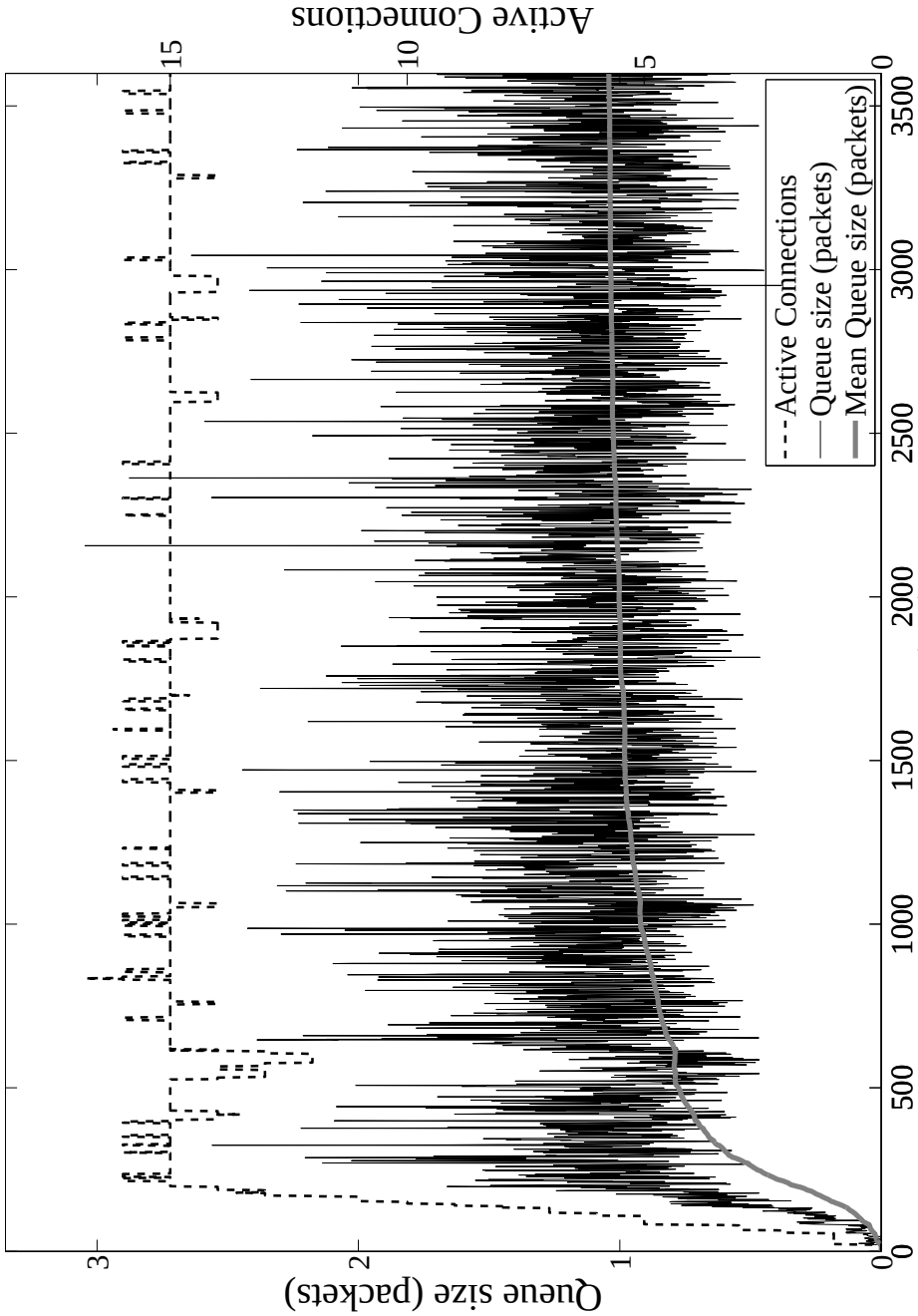
C3. Figure: The mean waiting time of the Job in System for the $M^X/M/1/K$ queue model with $\alpha = 0.9$



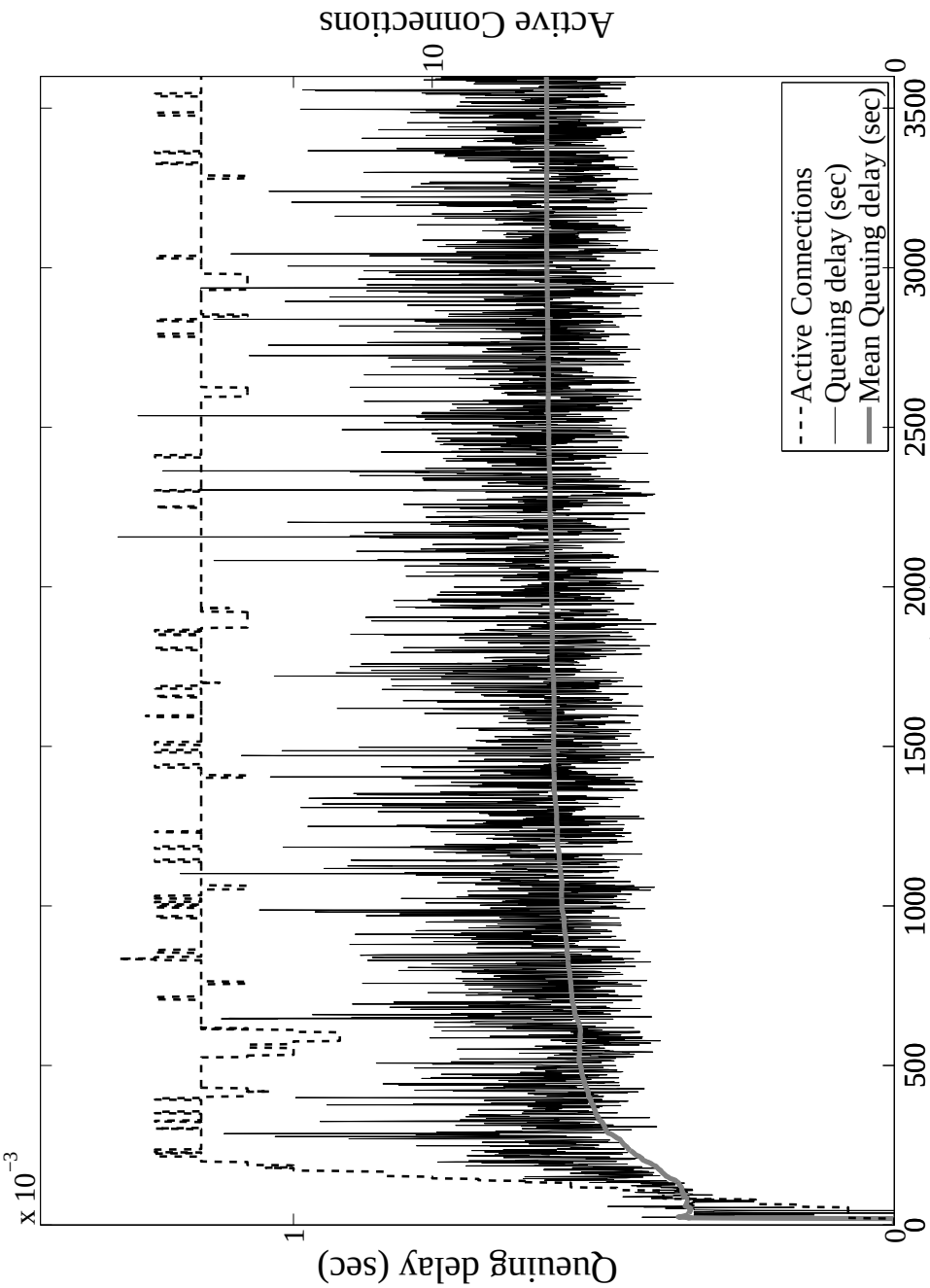
C4. Figure: The mean waiting time of the Job in System for the $M^X/M/1/K$ queue model with $\alpha = 0.99$

D

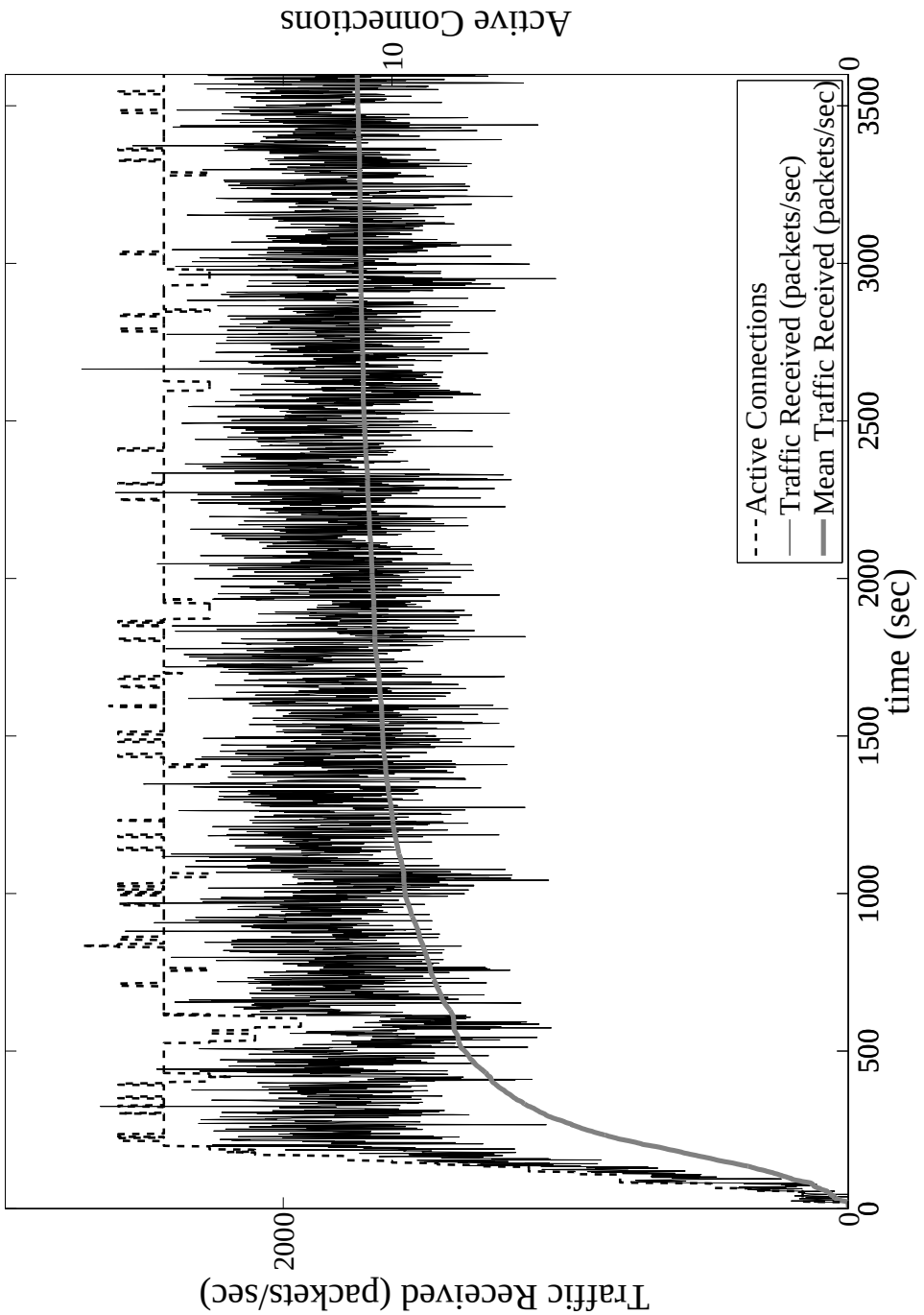
Simulation Results for Different Admission Conrols



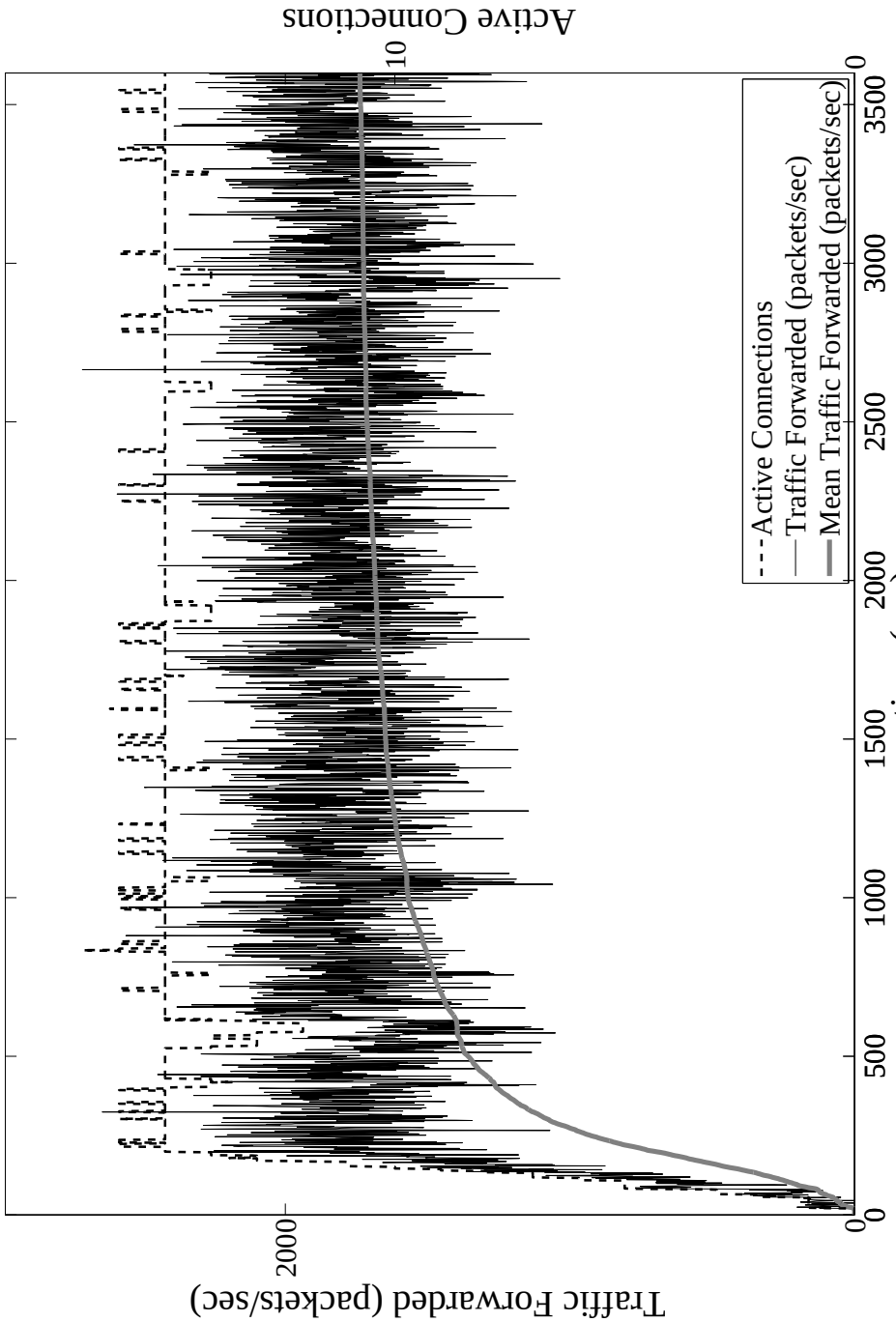
D1. Figure Queues Size for managed node under Parametric-Based Admission Control



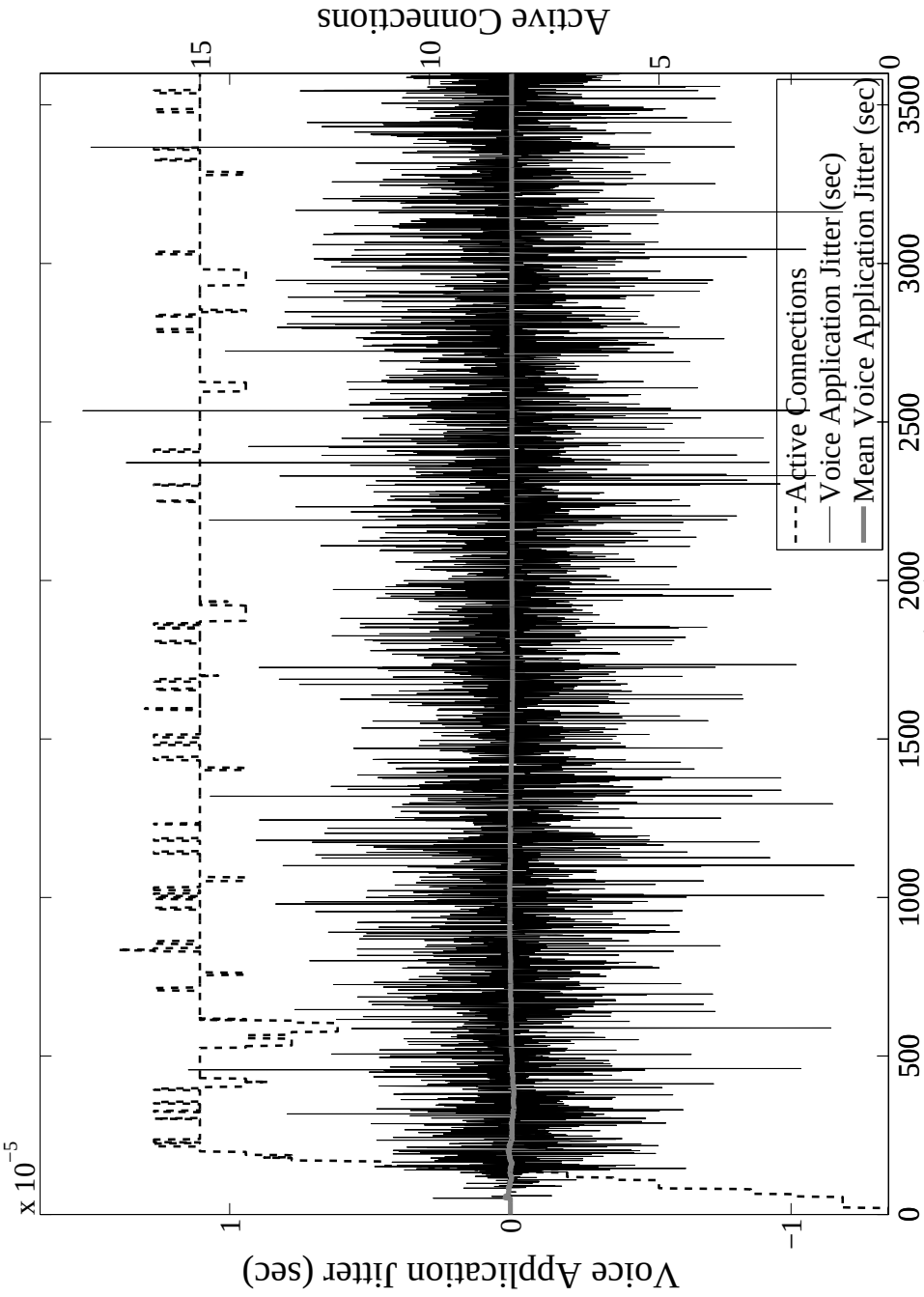
D2. Figure Queuing Delay for managed node under Parametric-Based Admission Control



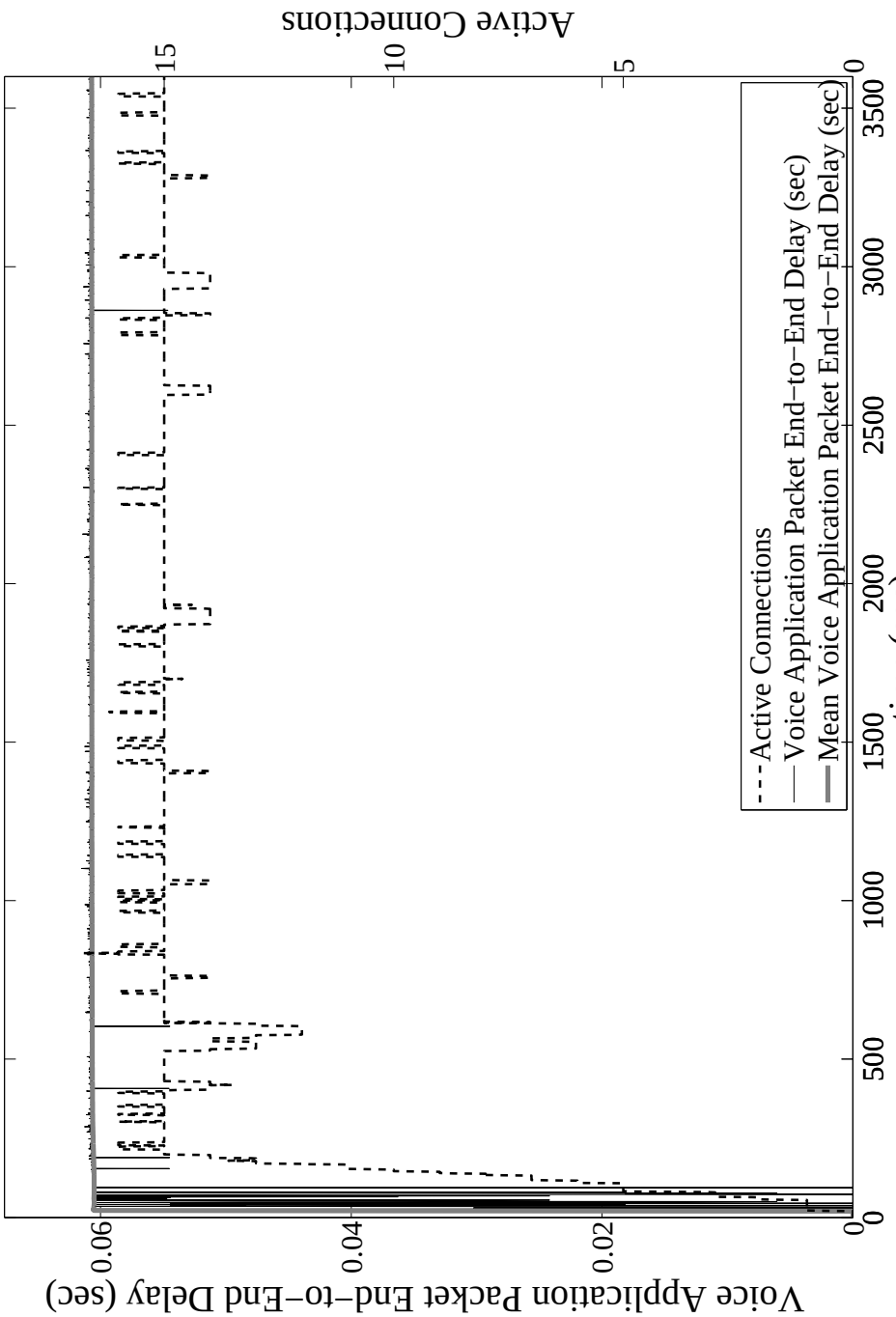
D3. Figure Traffic Received by managed node under Parametric-Based Admission Control



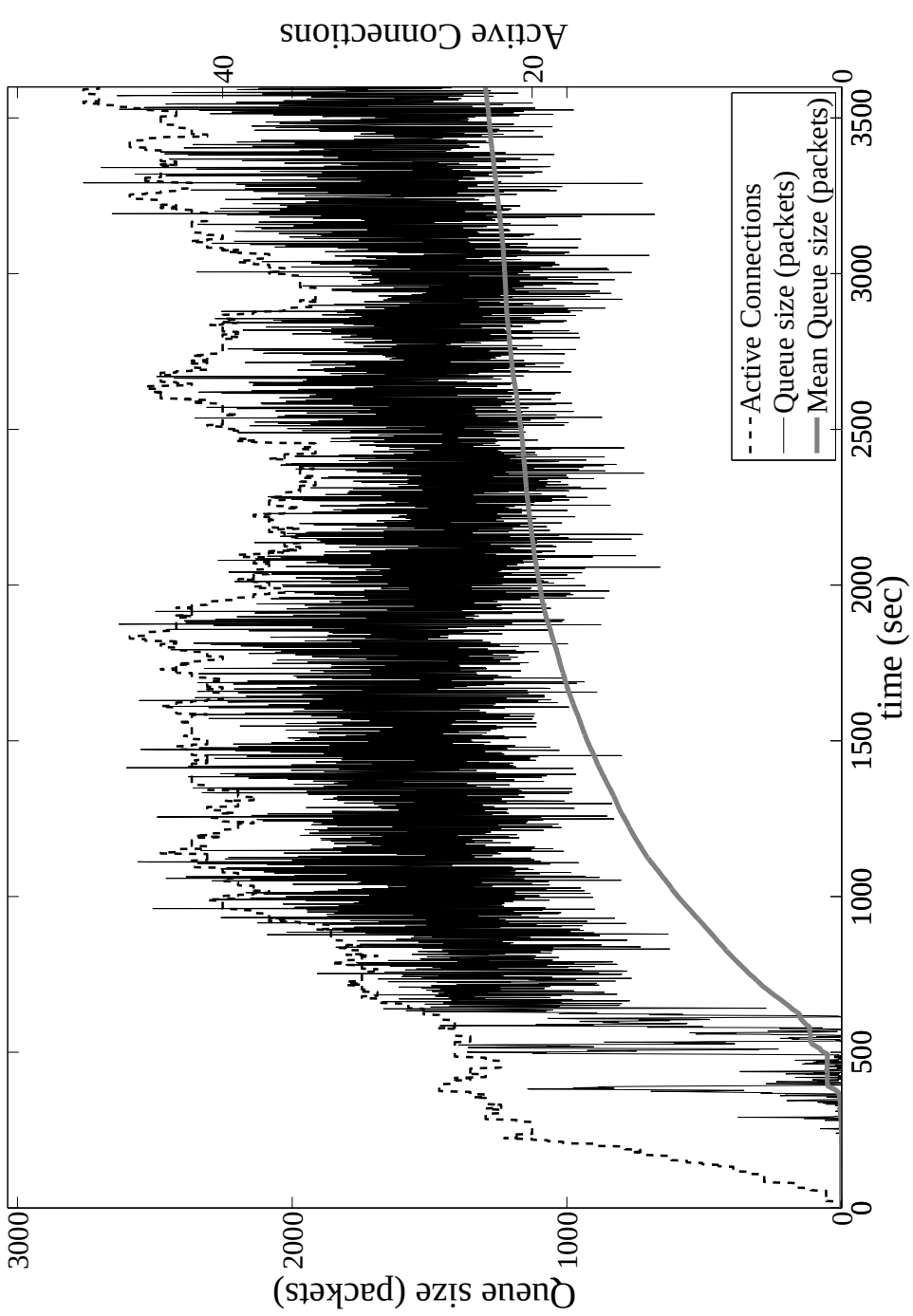
D4. Figure Traffic Forwarded by managed node under Parametric-Based Admission Control



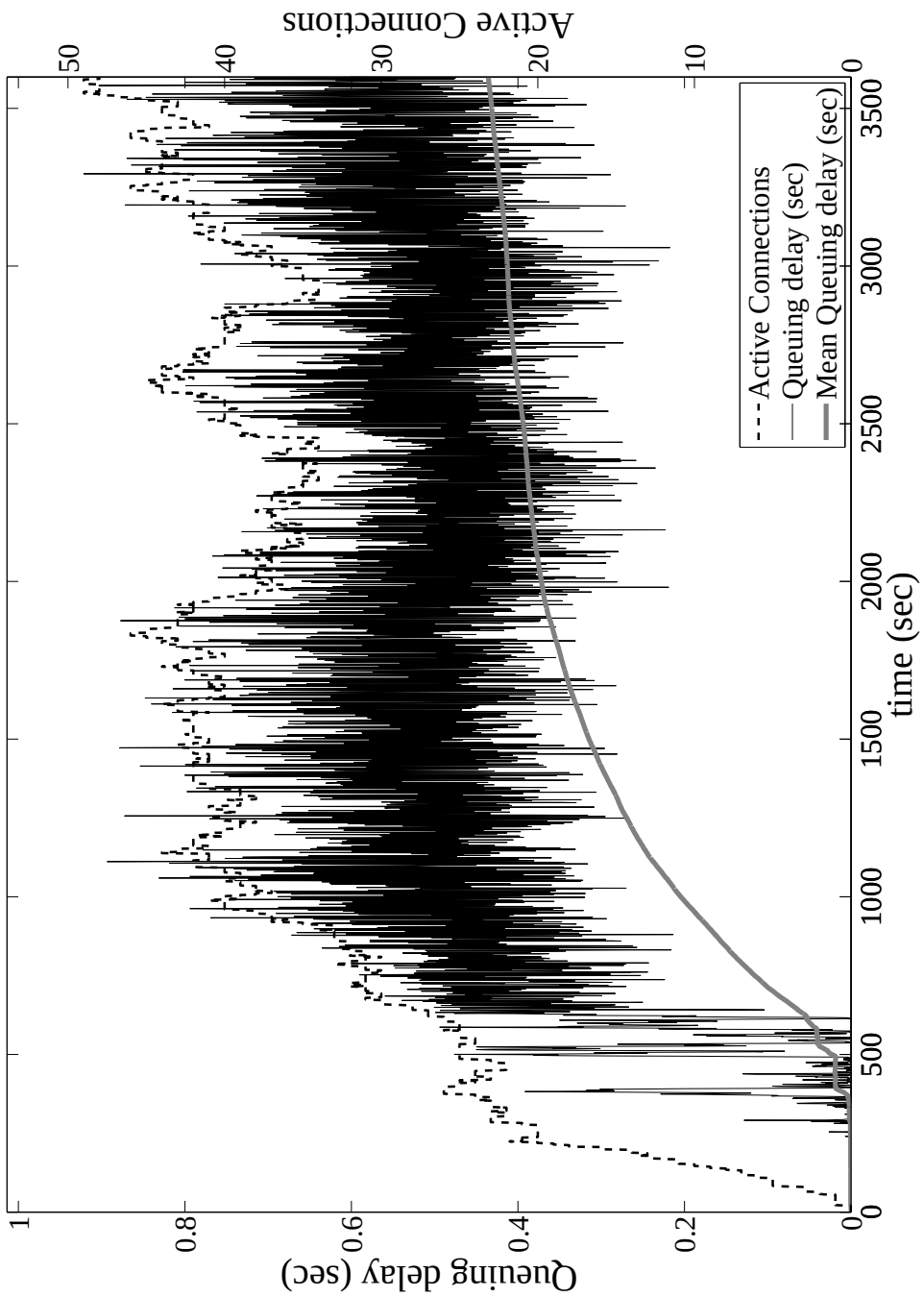
D5. Figure Voice Application Jitter for server node under Parametric-Based Admission Control



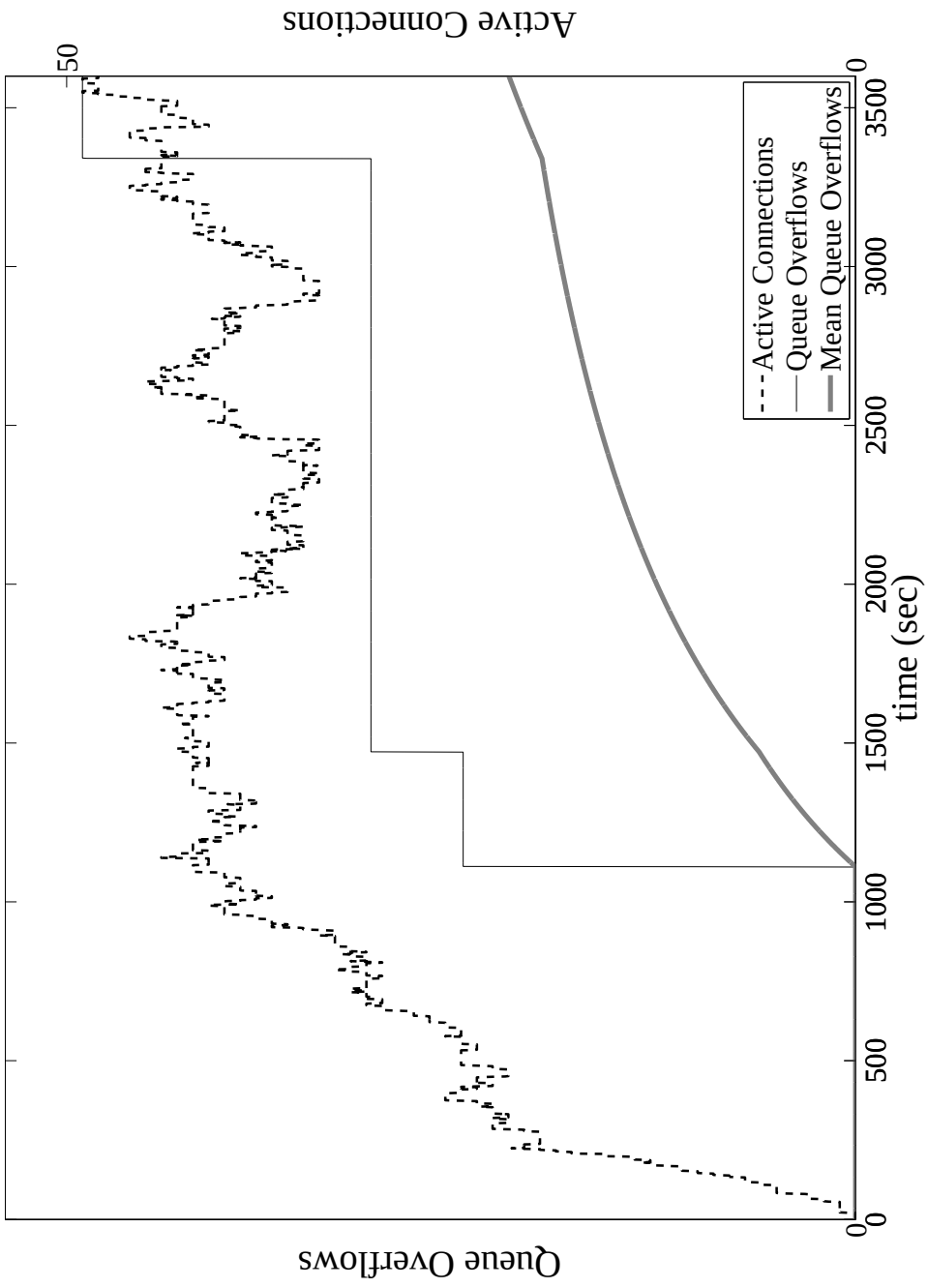
D6. Figure Voice Application Delay for server node under Parametric-Based Admission Control



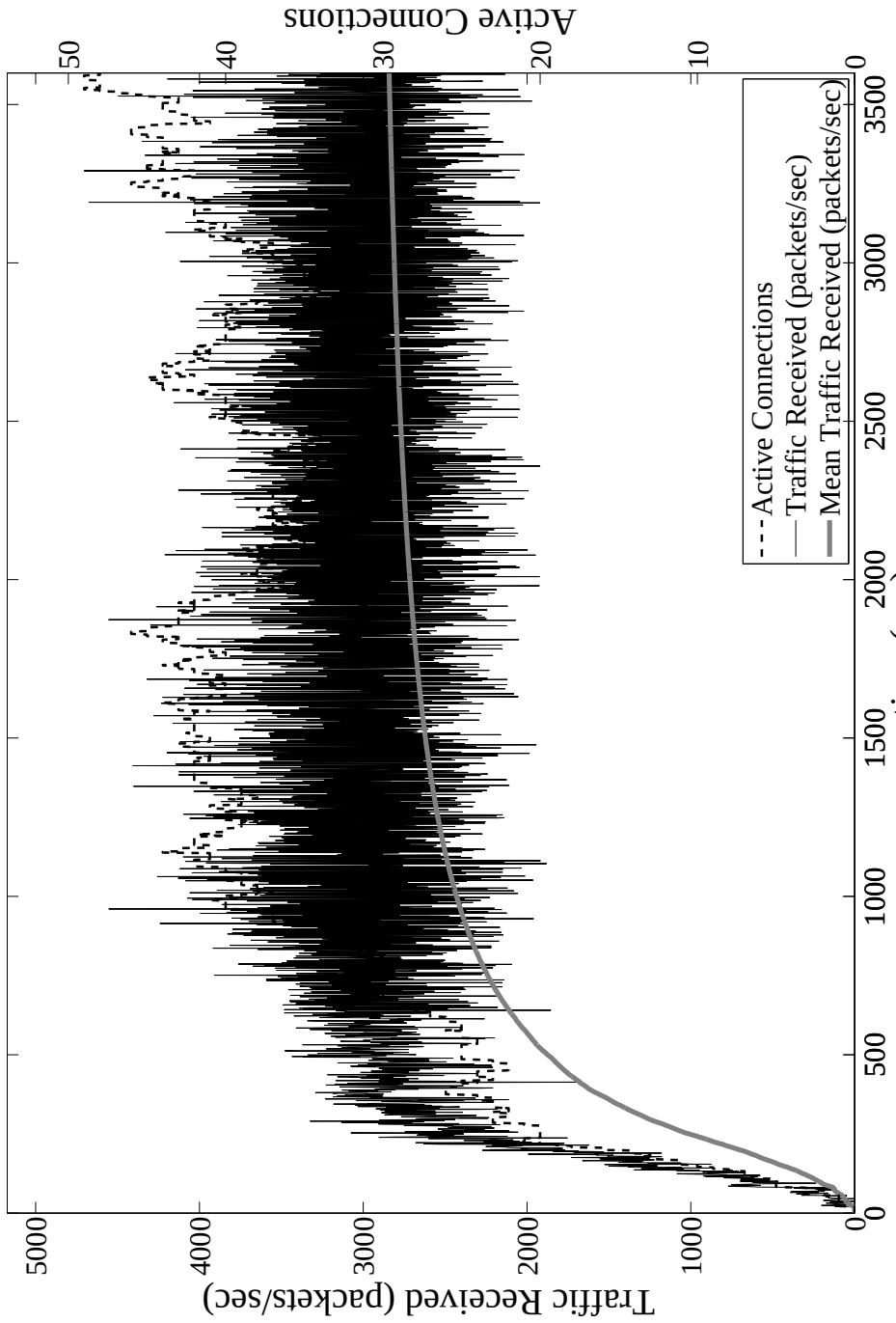
D7. Figure Queues Size for managed node under Measurement-Based Admission Control with observation window $T = 0.001$



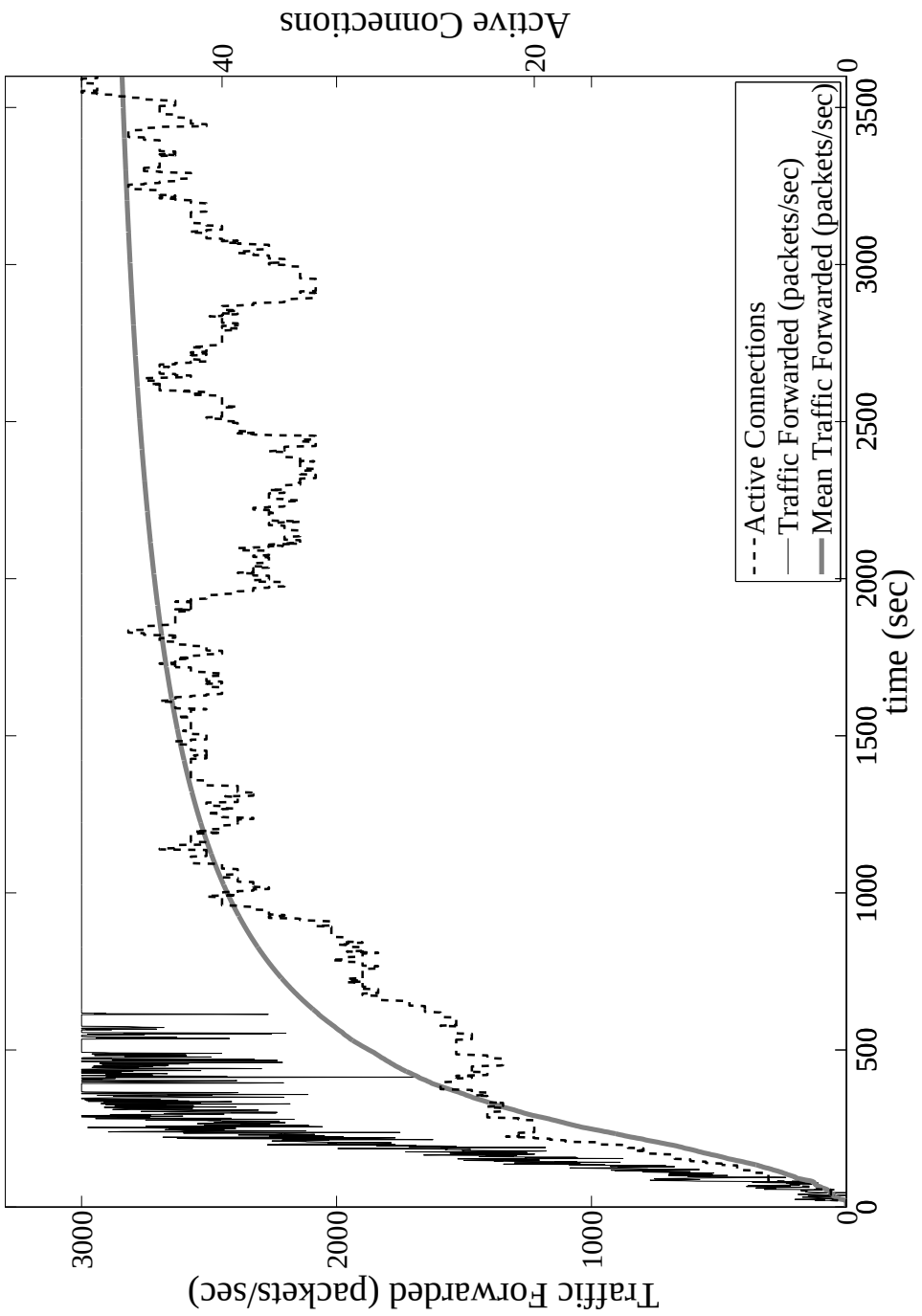
D8. Figure Queuing Delay for managed node under Measurement-Based Admission Control with observation window $T = 0.001$



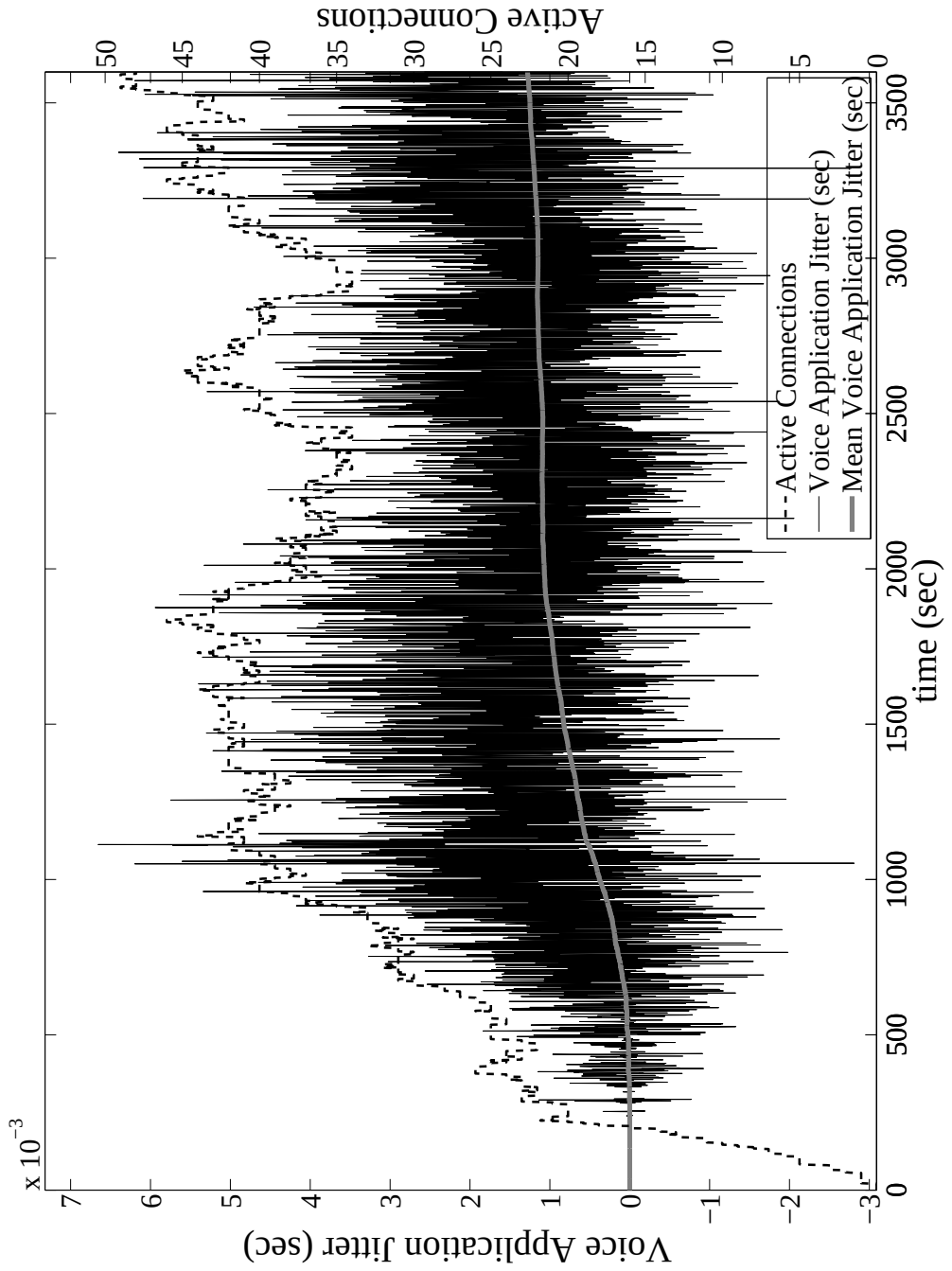
D9. Figure Queuing Delay for managed node under Measurement-Based Admission Control with observation window $T = 0.001$



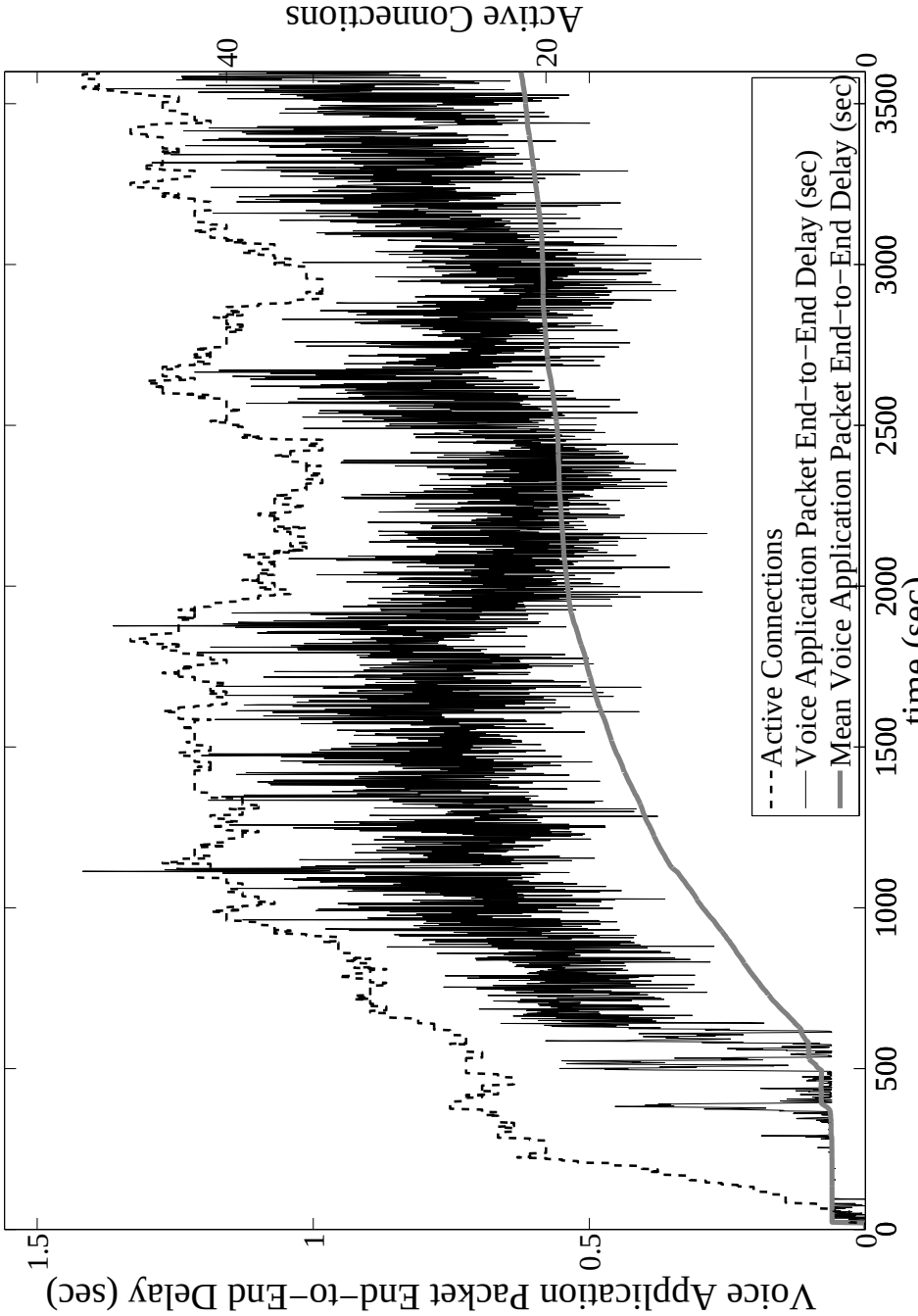
D10. Figure Traffic Received by managed node under Measurement-Based Admission Control with observation window $T = 0.001$



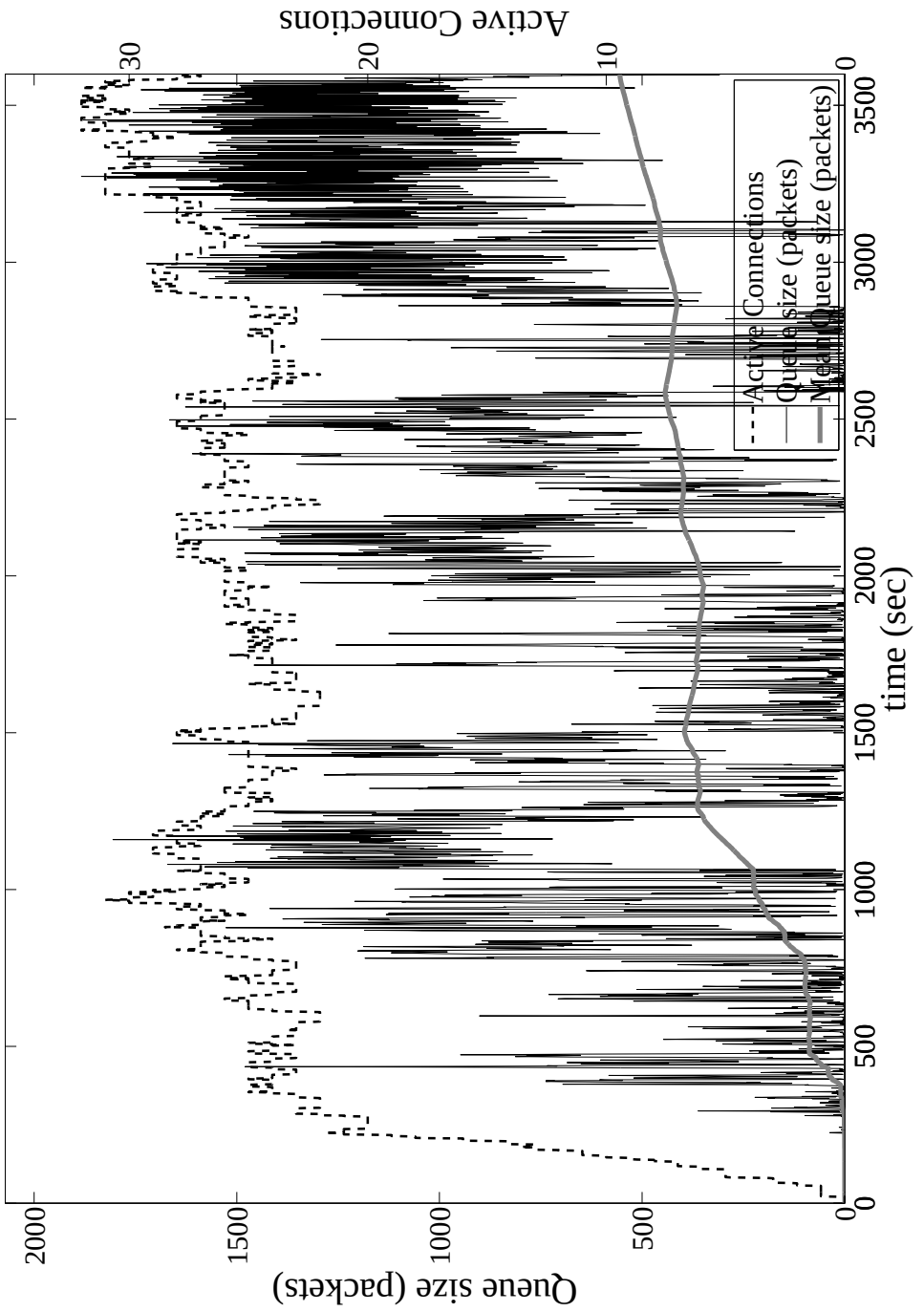
D11. Figure Traffic Forwarded by managed node under Measurement-Based Admission Control with observation window $T = 0.001$



D12. Figure Voice Application Jitter for server node under Measurement-Based Admission Control with observation window $T = 0.001$

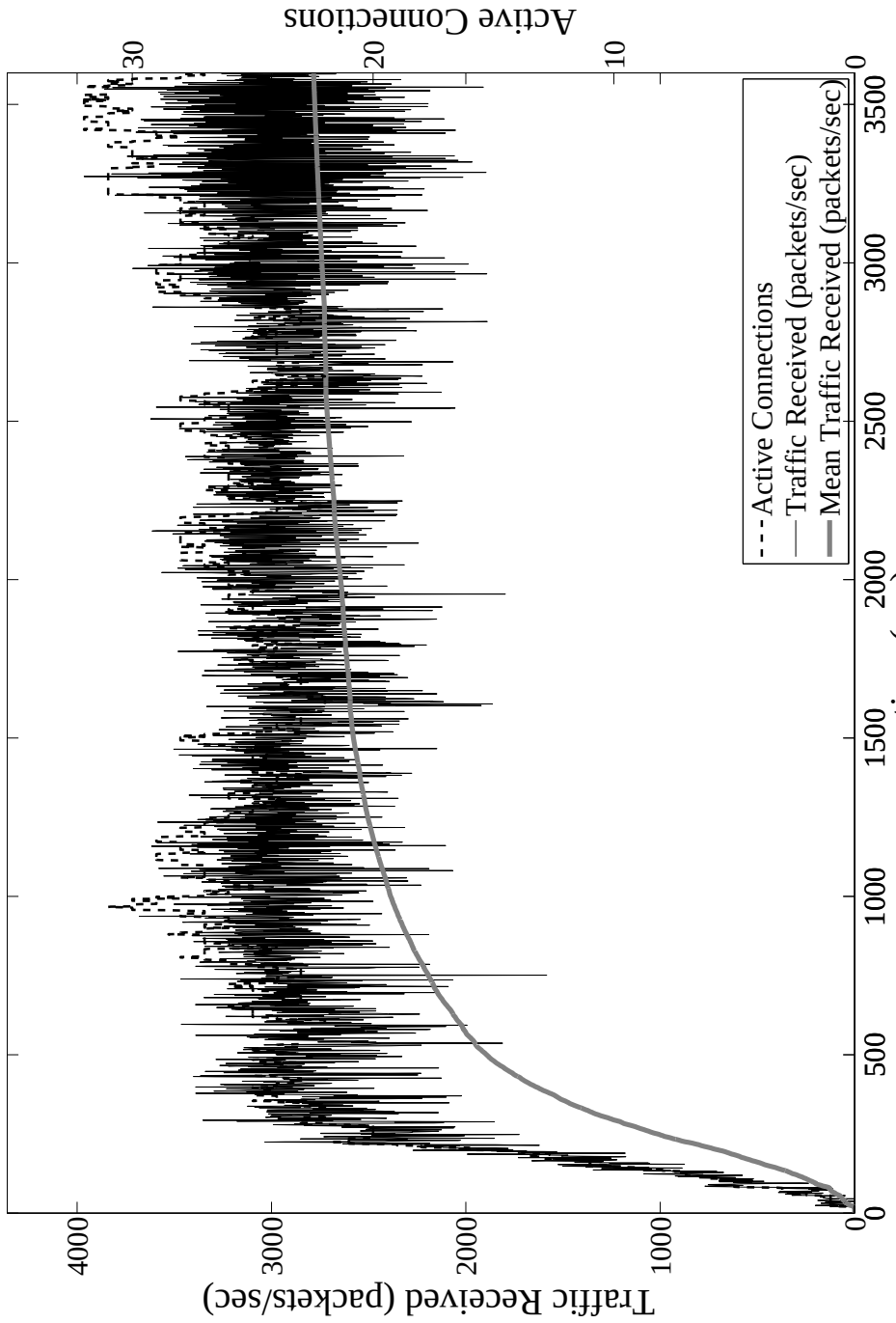


D13. Figure Voice Application Delay for server node under Measurement-Based Admission Control with observation window $T = 0.001$

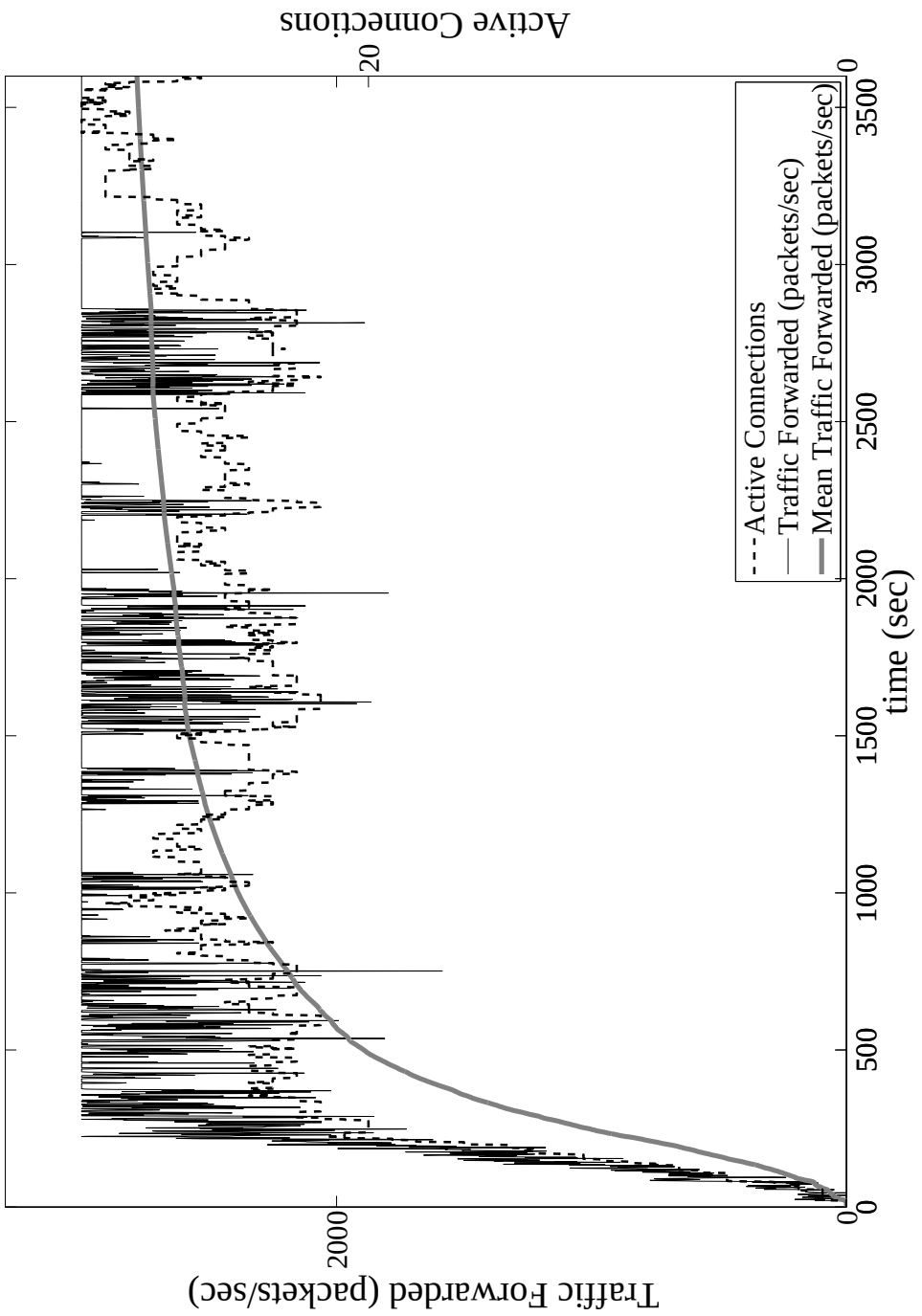


D14. Figure Queues Size for managed node under Measurement-Based Admission Control with observation window $T = 1$

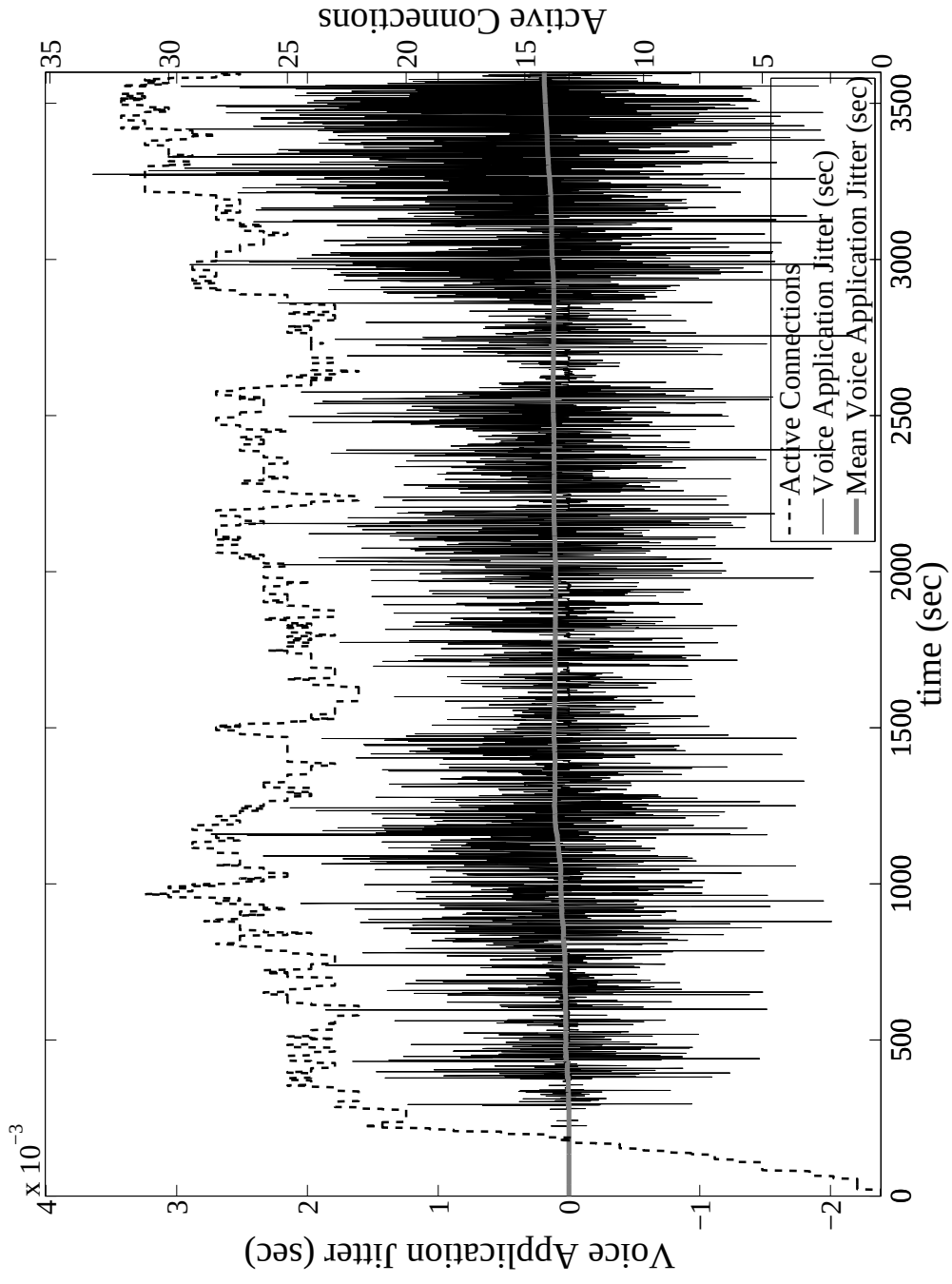
D15. Figure Queuing Delay for managed node under Measurement-Based Admission Control with observation window $T = 1$



D16. Figure Traffic Received by managed node under Measurement-Based Admission Control with observation window $T = 1$

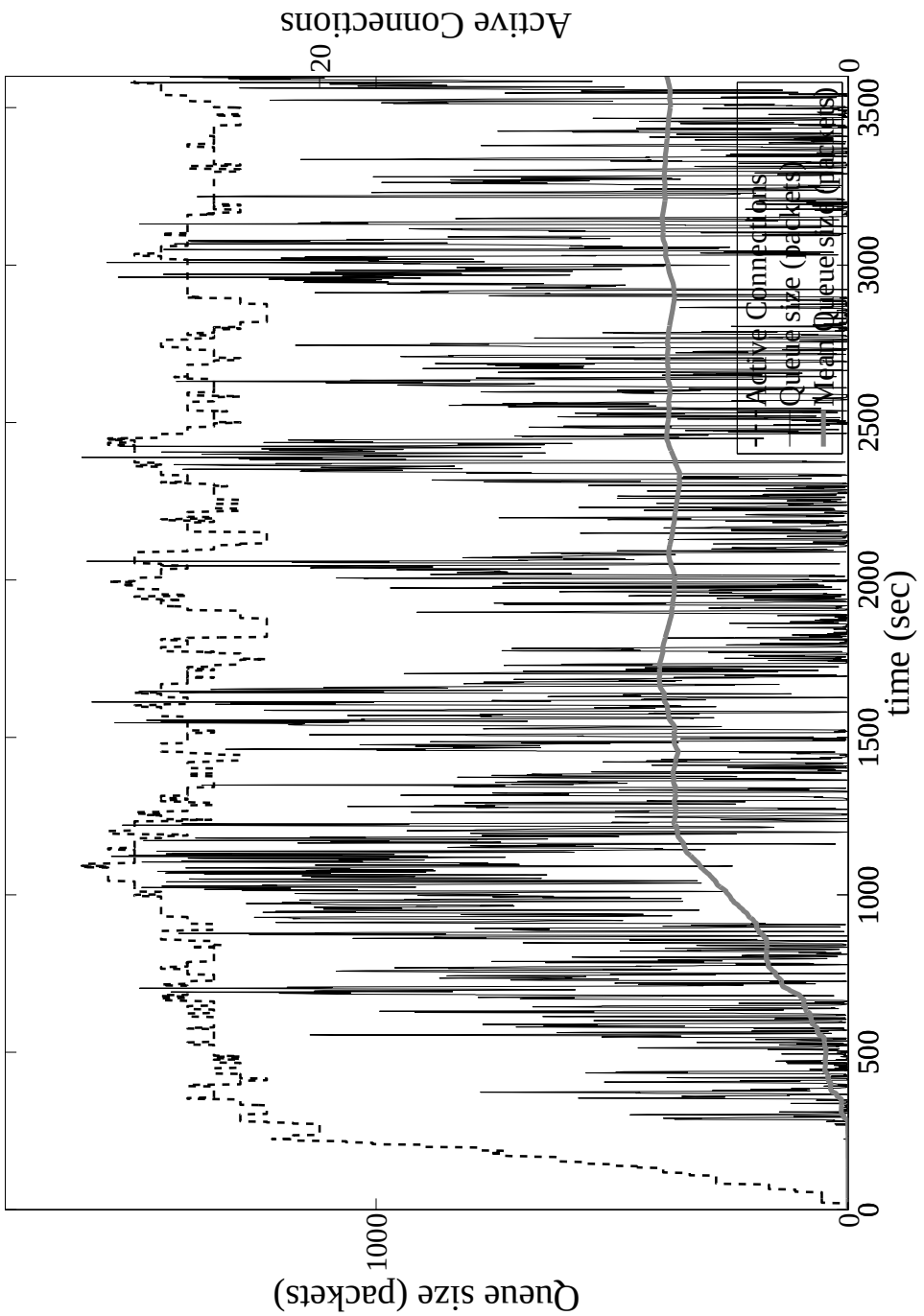


D17. Figure Traffic Forwarded by managed node under Measurement-Based Admission Control with observation window $T = 1$

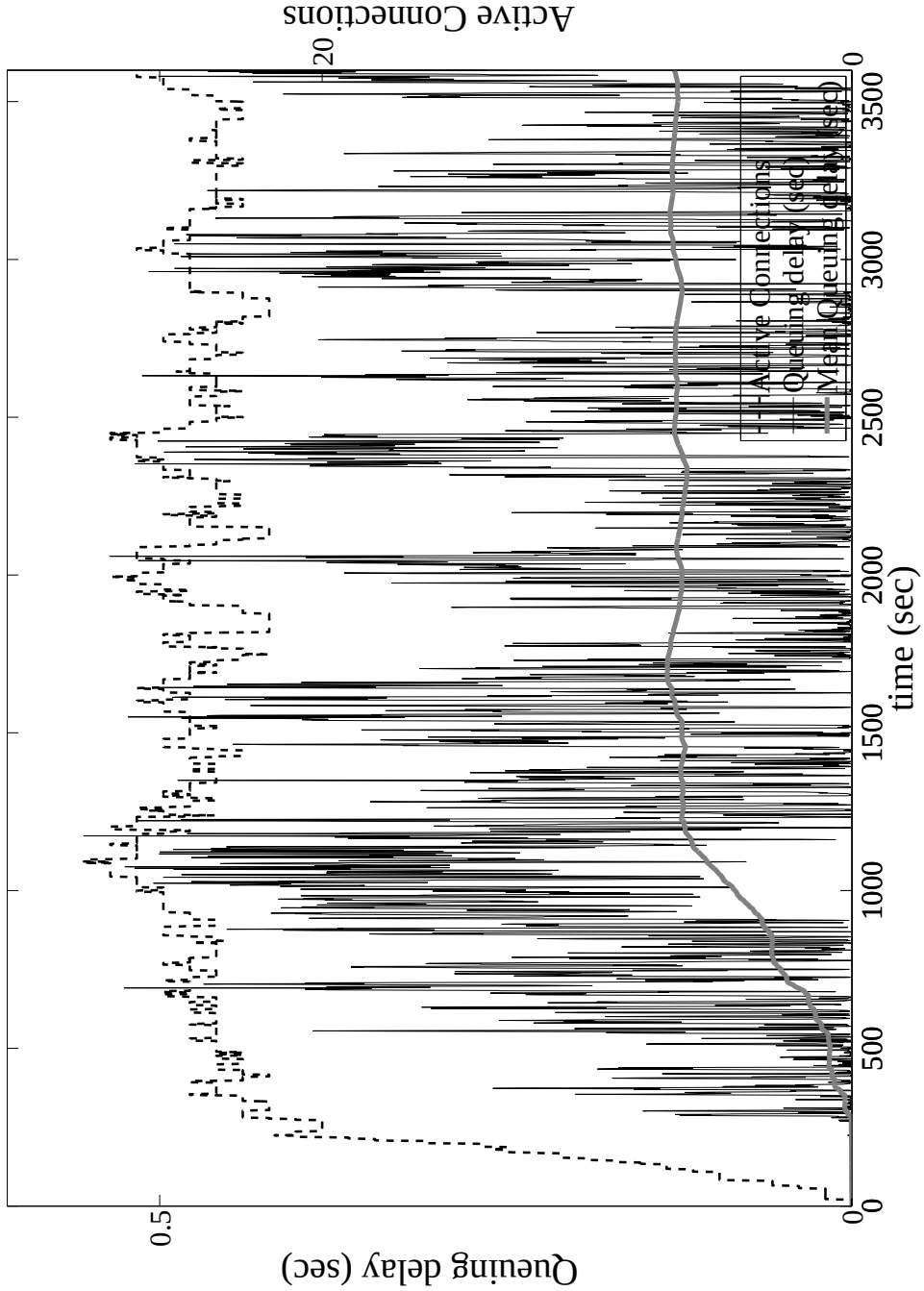


D18. Figure Voice Application Jitter for server node under Measurement-Based Admission Control with observation window $T = 1$

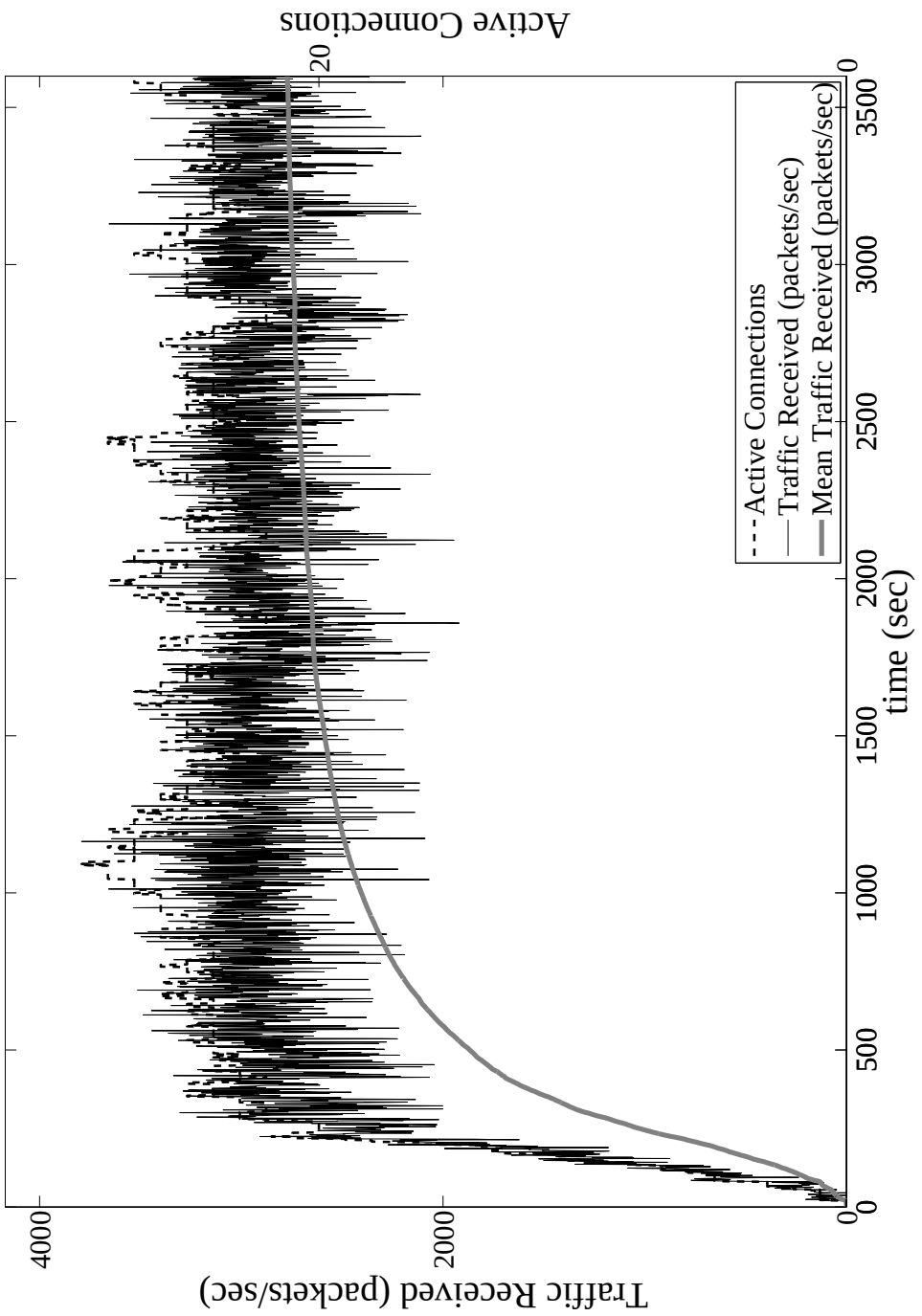
D19. Figure Voice Application Delay for server node under Measurement-Based Admission Control with observation window $T = 1$



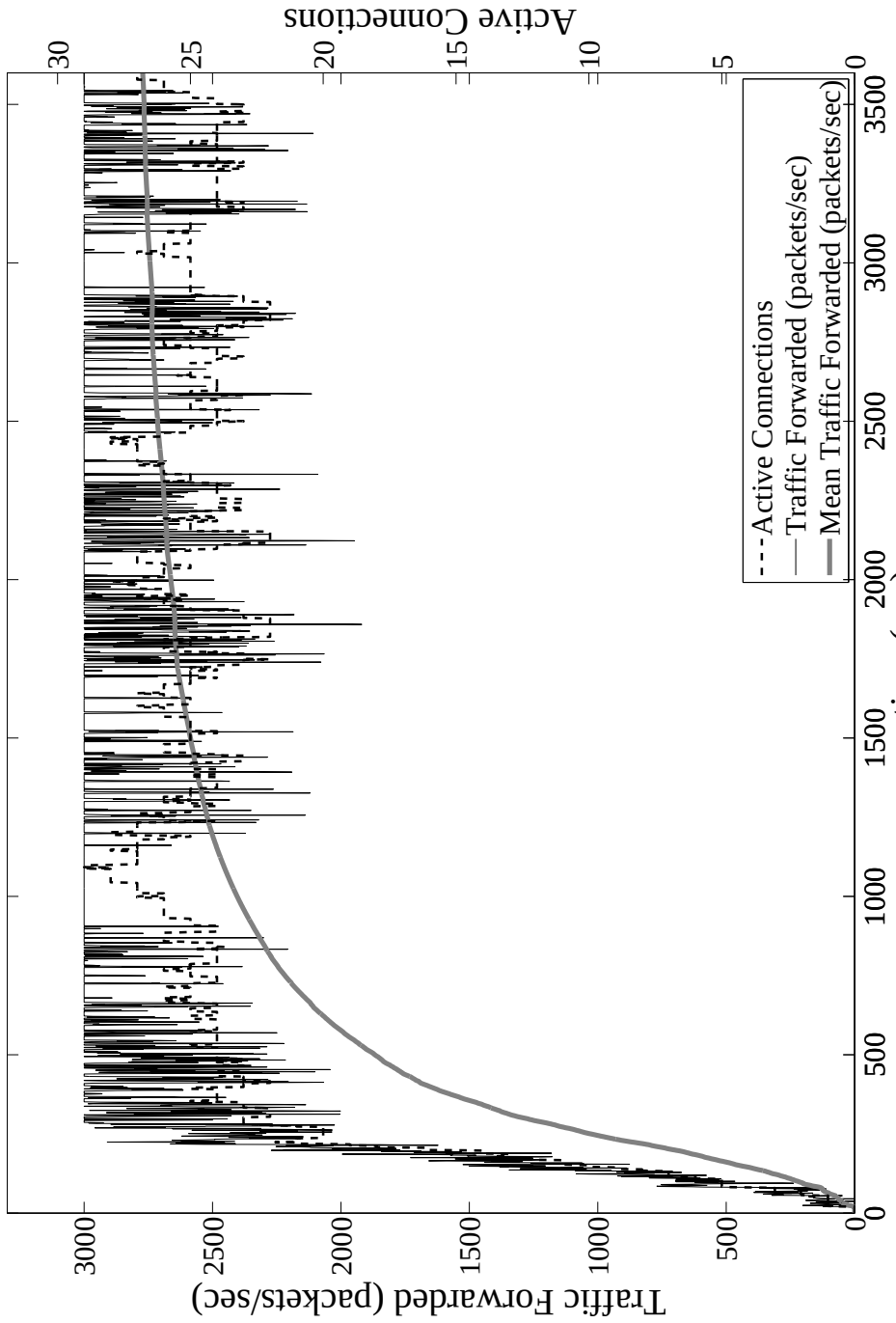
D20. Figure Queues Size for managed node under Measurement-Based Admission Control with observation window $T = 10$



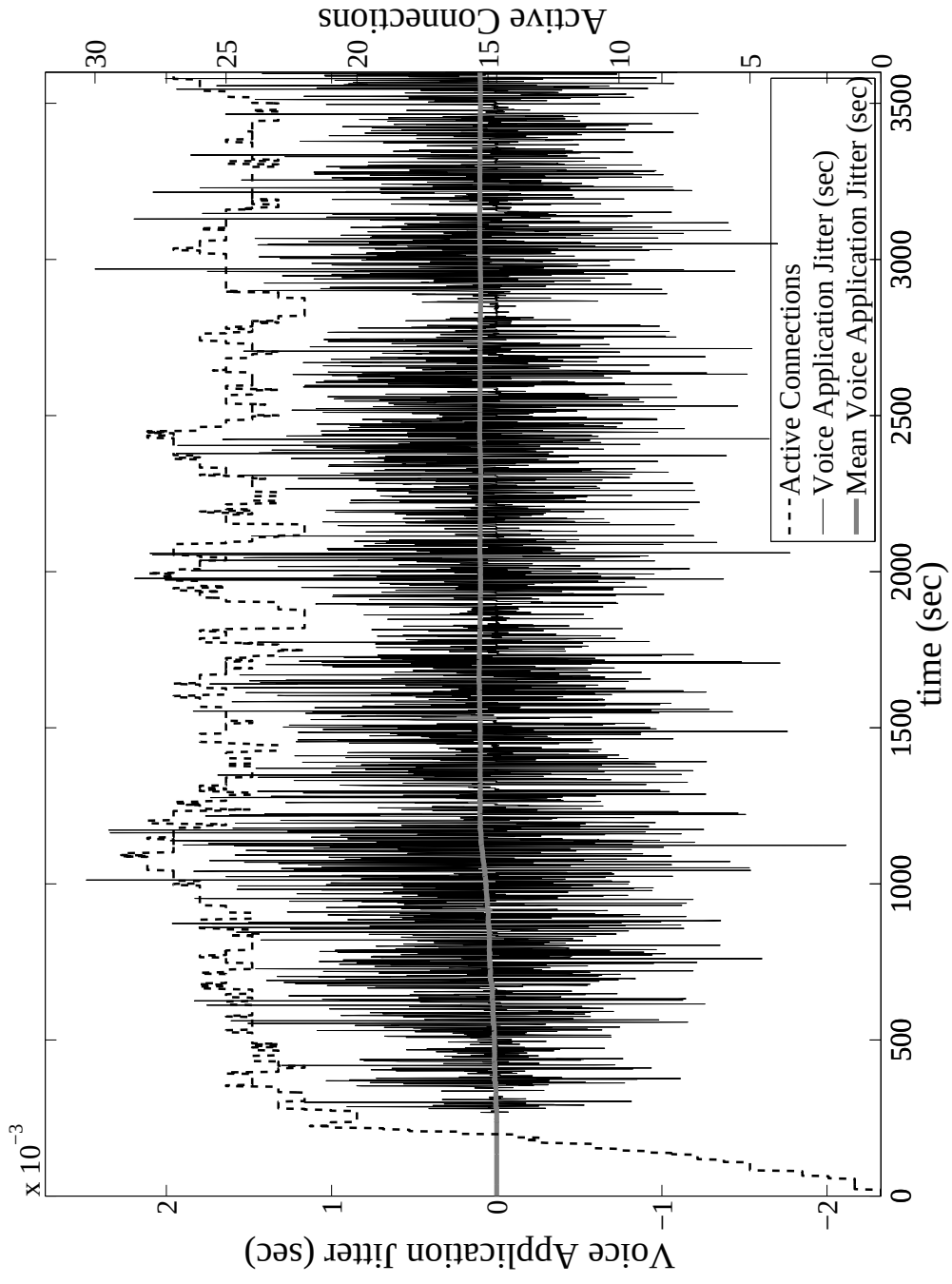
D21. Figure Queuing Delay for managed node under Measurement-Based Admission Control with observation window $T = 10$



D22. Figure Traffic Received by managed node under Measurement-Based Admission Control with observation window $T = 10$

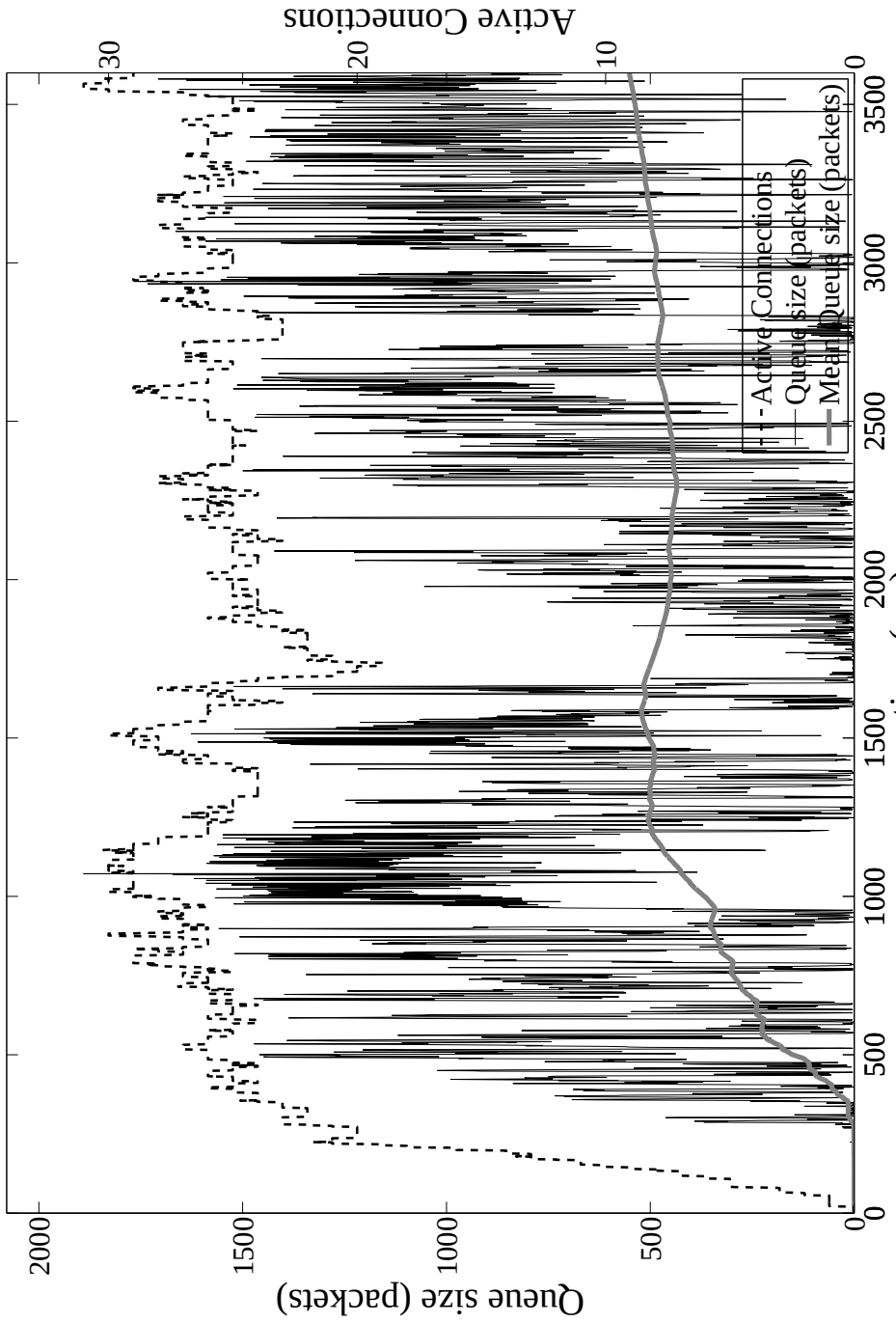


D23. Figure Traffic Forwarded by managed node under Measurement-Based Admission Control with observation window $T = 10$

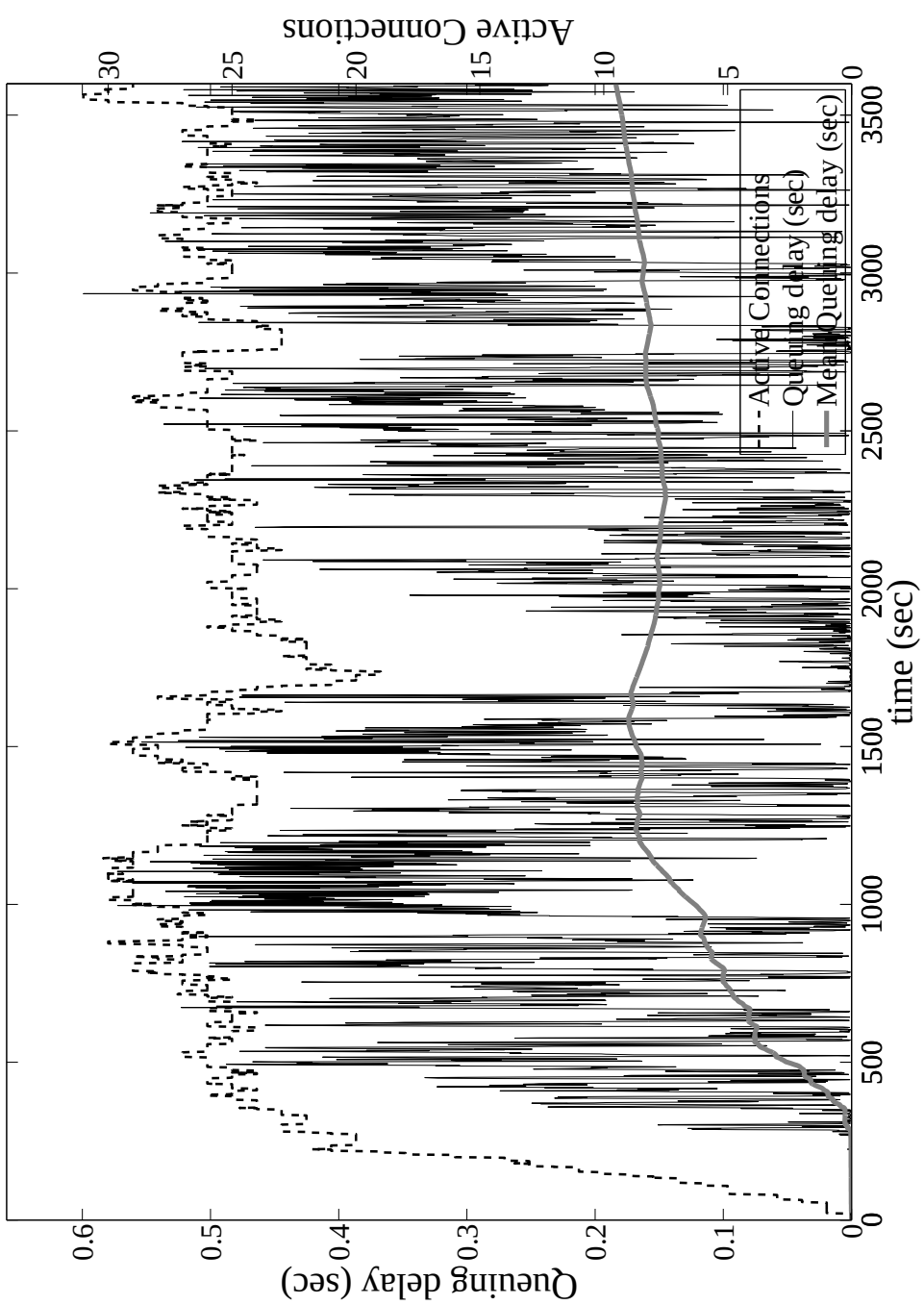


D24. Figure Voice Application Jitter for server node under Measurement-Based Admission Control with observation window $T = 10$

D25. Figure Voice Application Delay for server node under Measurement-Based Admission Control with observation window $T = 10$

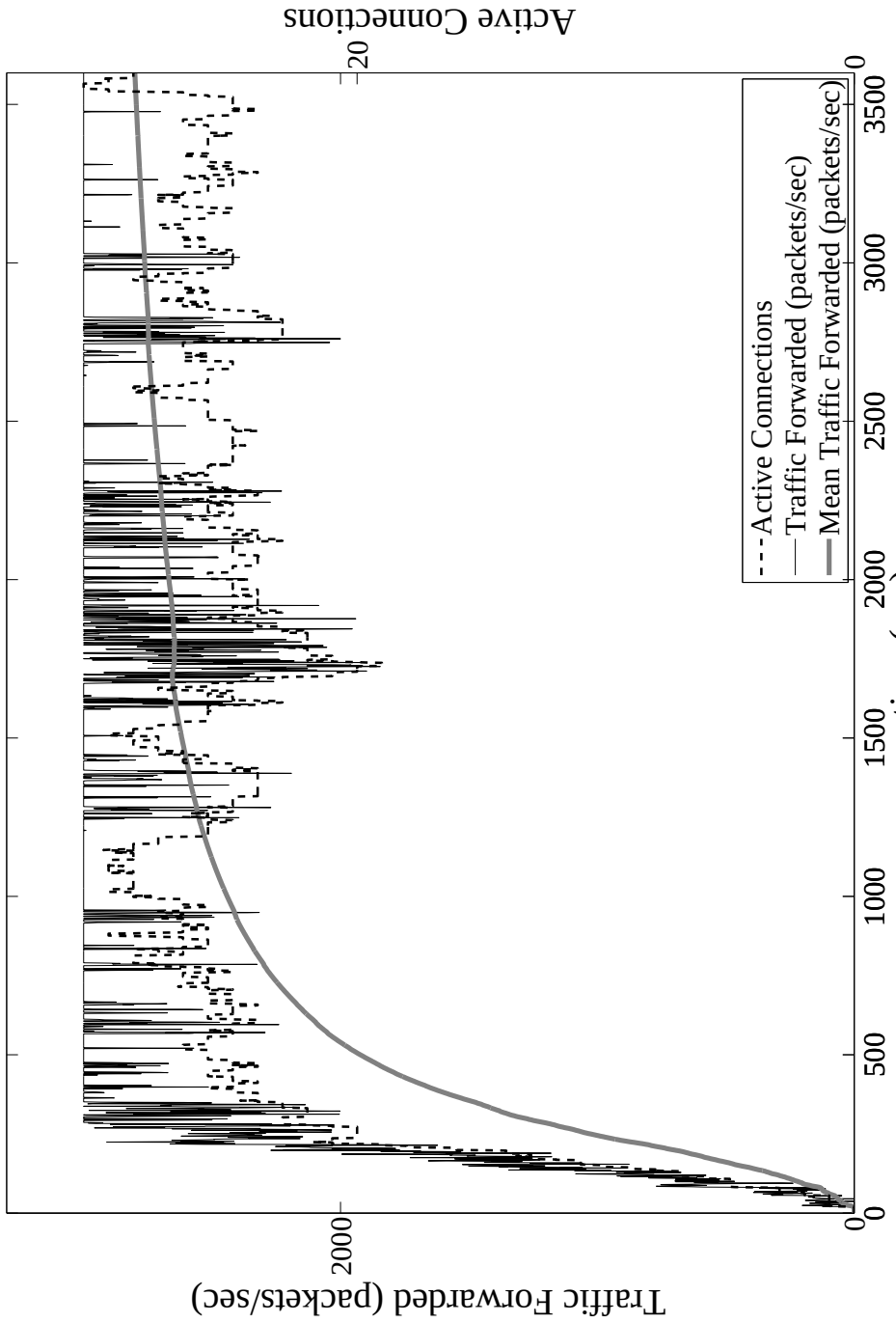


D26. Figure Queues Size for managed node under Measurement-Based Admission Control with observation window $T = 100/\lambda$

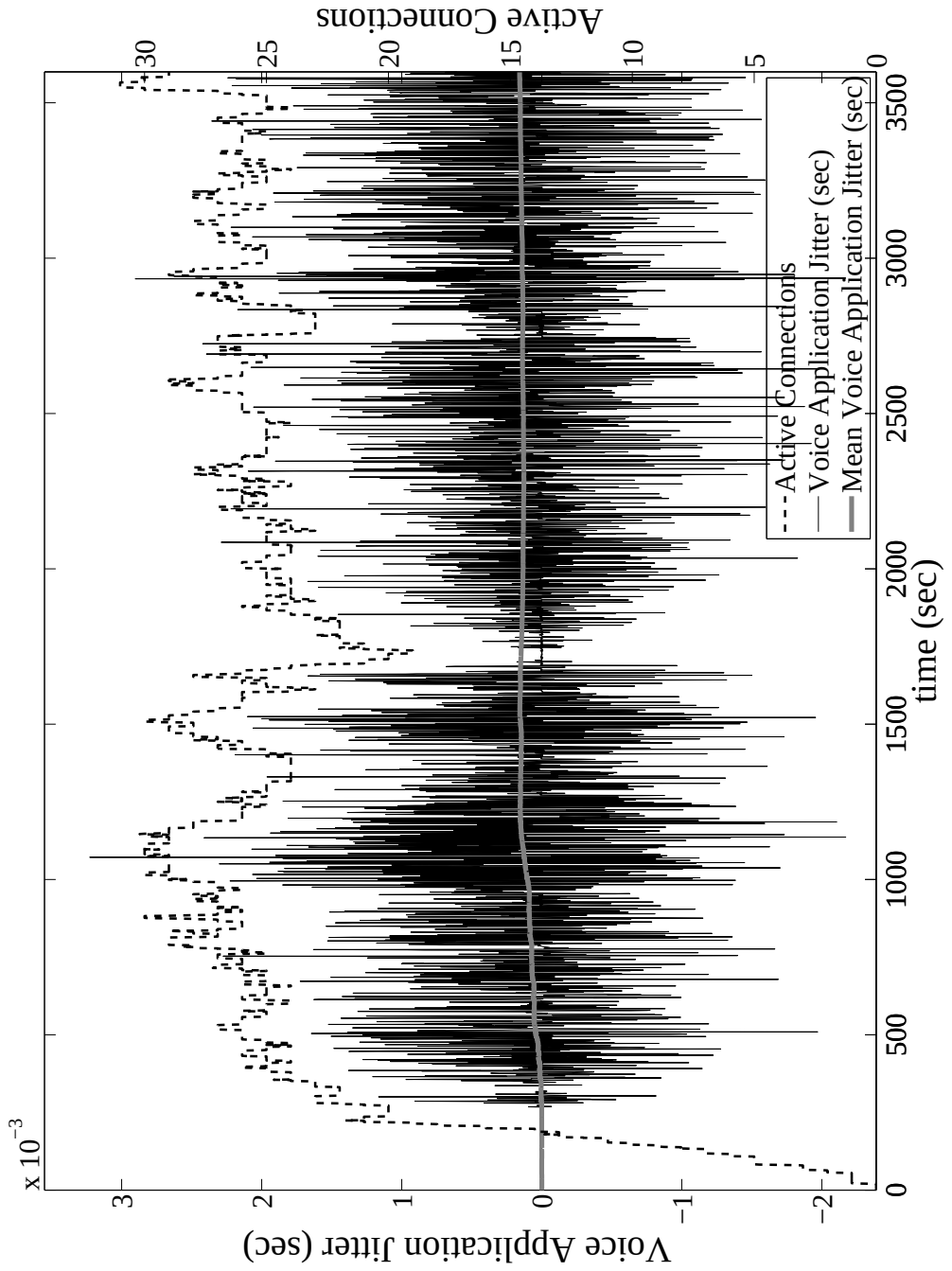


D27. Figure Queuing Delay for managed node under Measurement-Based Admission Control with observation window $T = 100/\lambda$

D28. Figure Traffic Received by managed node under Measurement-Based Admission Control with observation window $T = 100/\lambda$

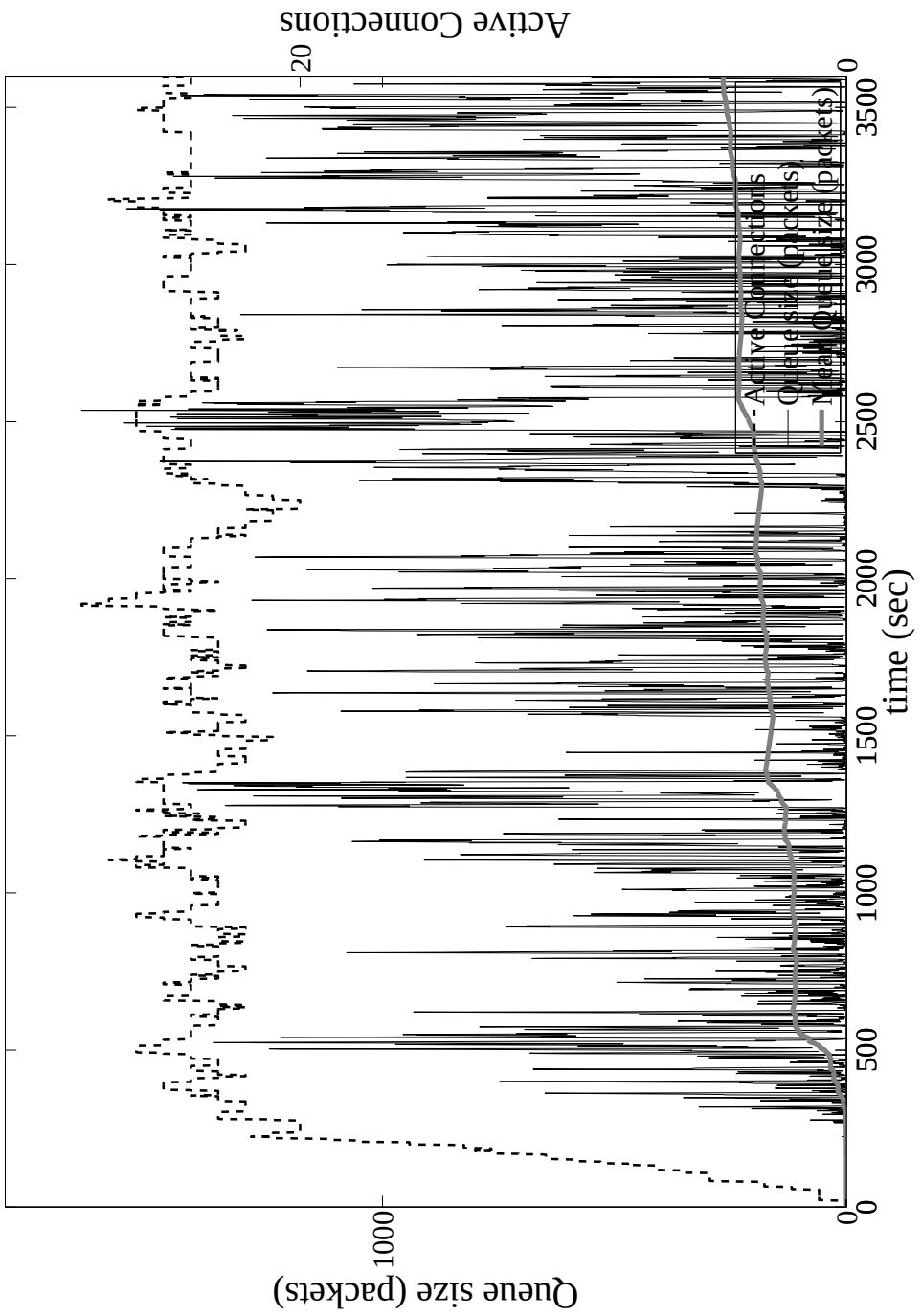


D29. Figure Traffic Forwarded by managed node under Measurement-Based Admission Control with observation window $T = 100/\lambda$

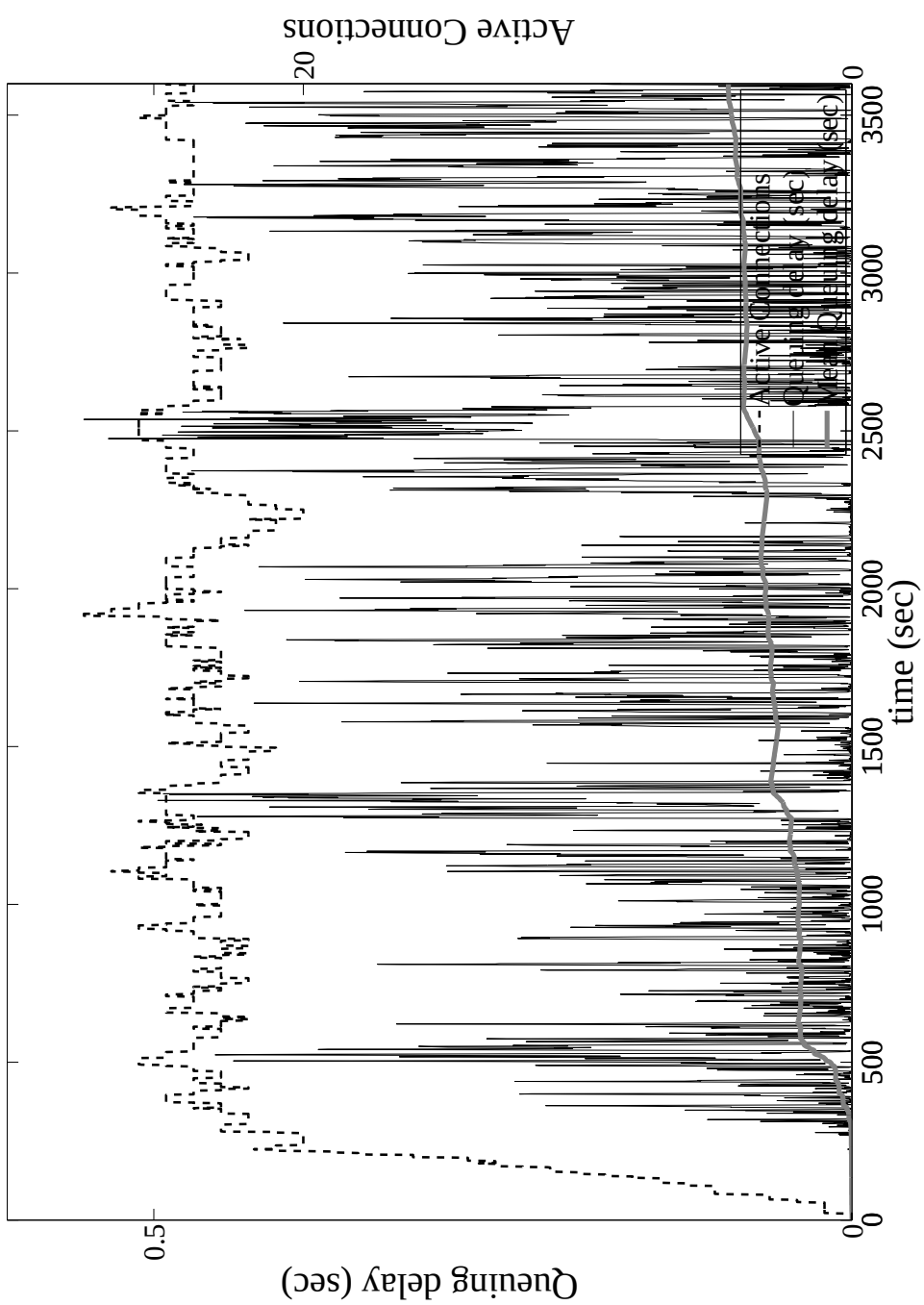


D30. Figure Voice Application Jitter for server node under Measurement-Based Admission Control with observation window $T = 100/\lambda$

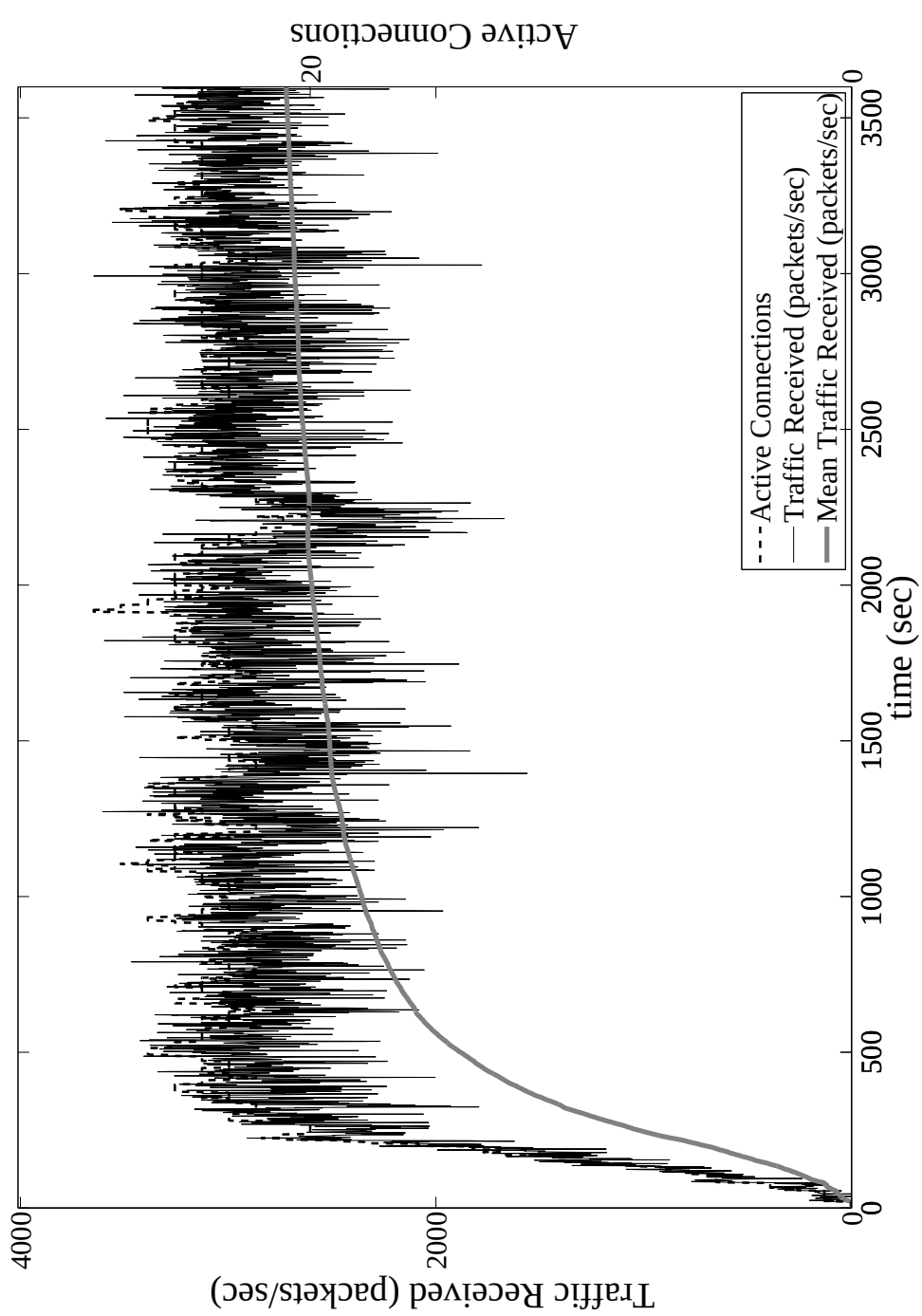
D31. Figure Voice Application Delay for server node under Measurement-Based Admission Control with observation window $T = 100/\lambda$



D32. Figure Queues Size for managed node under Measurement-Based Admission Control with observation window $T = \tau_k$

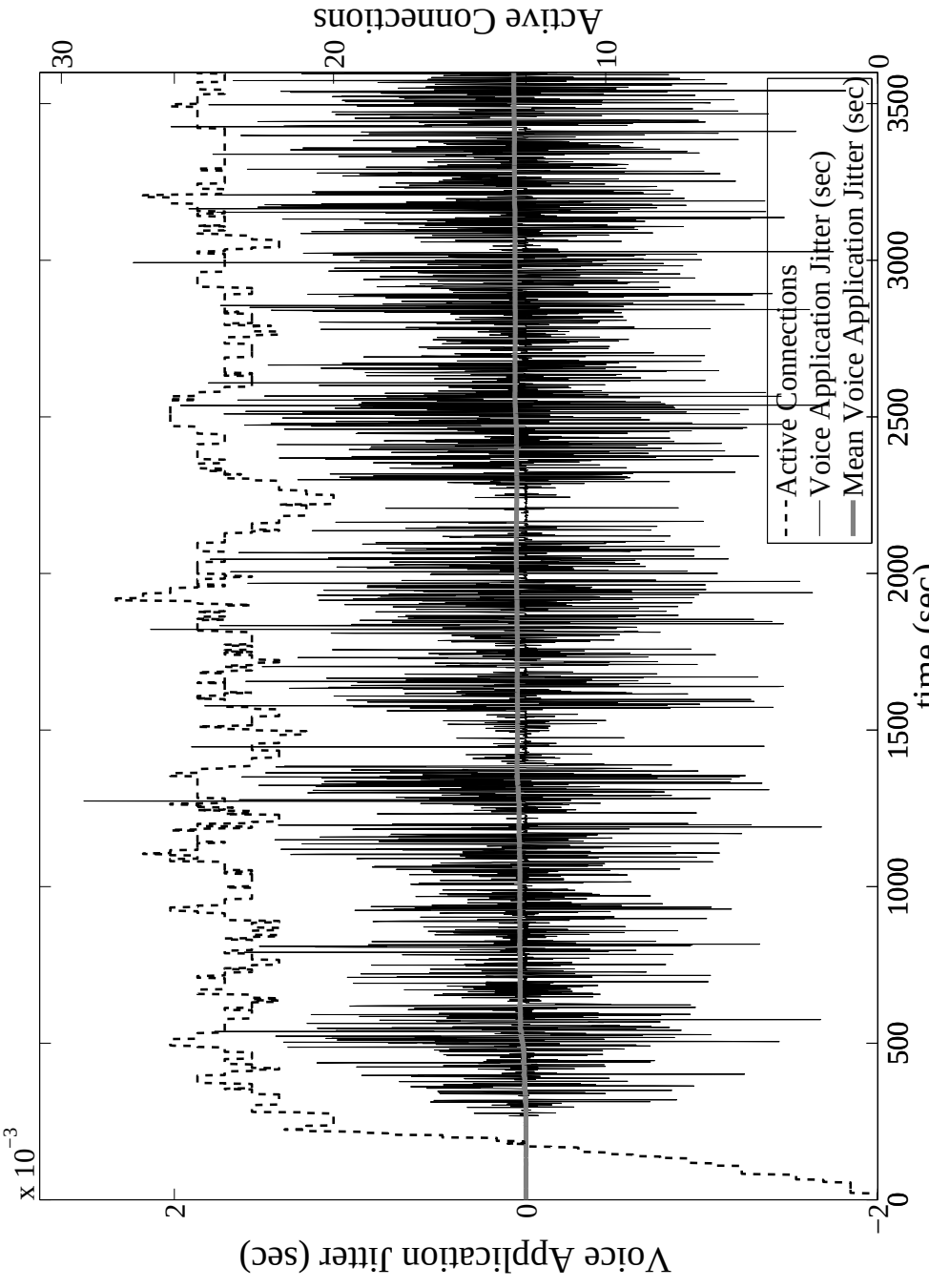


D33. Figure Queuing Delay for managed node under Measurement-Based Admission Control with observation window $T = \tau_k$

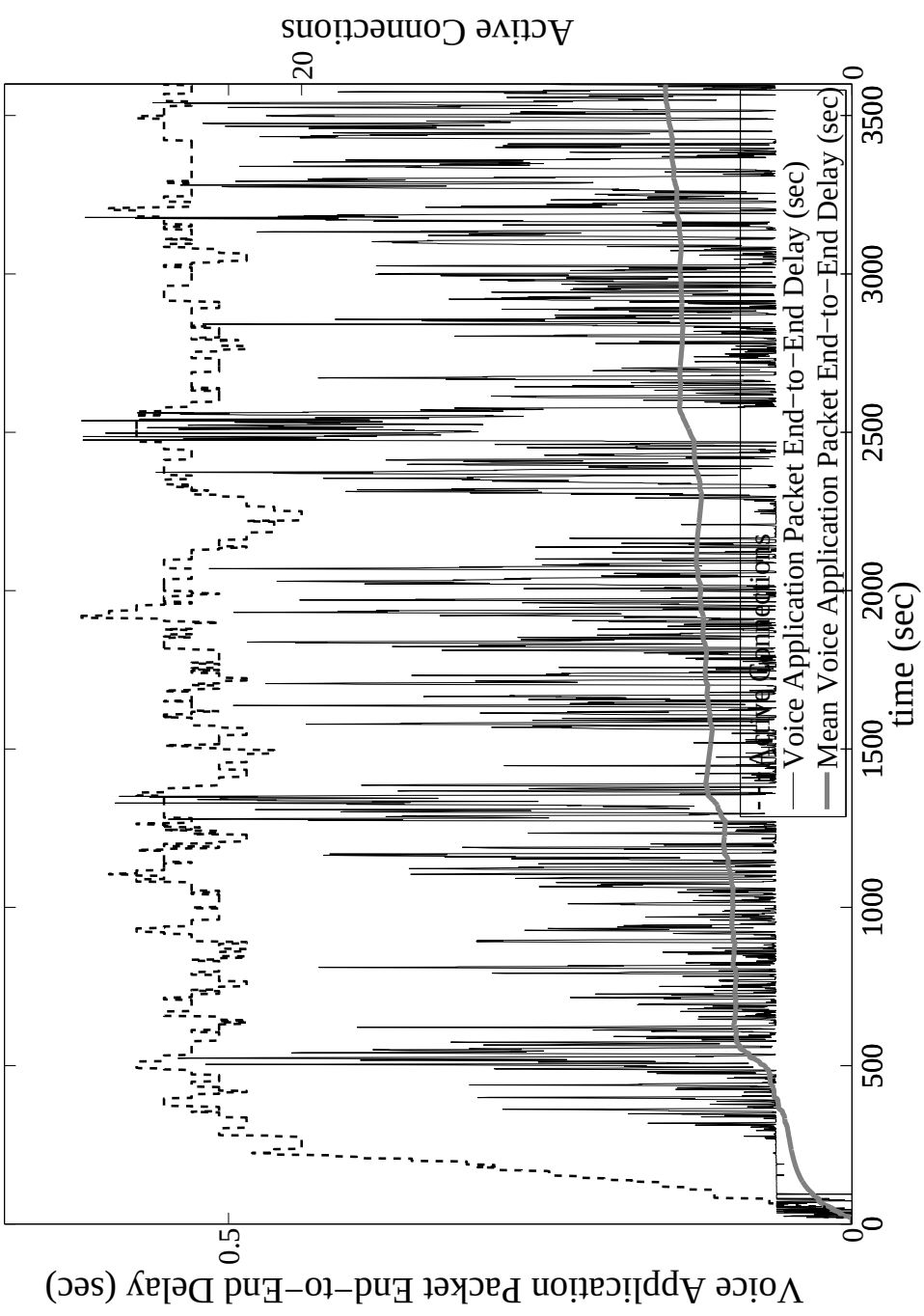


D34. Figure Traffic Received by managed node under Measurement-Based Admission Control with observation window $T = \tau_k$

D35. Figure Traffic Forwarded by managed node under Measurement-Based Admission Control with observation window $T = \tau_k$



D36. Figure Voice Application Jitter for server node under Measurement-Based Admission Control with observation window $T = \tau_k$



D37. Figure Voice Application Delay for server node under Measurement-Based Admission Control with observation window $T = \tau_k$

E

Definitions

Flow	A stream of packets with the same source IP address, source port number, destination IP address, destination port number and protocol ID.
Service Level Agreement	A service contract between a customer and a service provider that specifies the forwarding service a customer should receive. A customer may be a user organization or another provider domain (upstream domain).
Traffic Profile	A description of the properties of a traffic stream such as rate and burst size.
Differentiated Services field	The field in which the Differentiated Services class is encoded. It is the Type of Service (TOS) octet in the IPv4 header or the Traffic Class octet in the IPv6 header.
Per-Hop-Behavior	The externally observable behavior of a packet at a DS-compliant router
Mechanism	A specific algorithm or operation (e.g., queuing discipline) that is implemented in a router to realize a set of one or more per-hop behaviors.
Admission Control	The decision process of whether to accept a request for resources (link bandwidth plus buffer space)
Classification	The process of sorting packets based on the content of packet headers according to defined rules.
Behavior Aggregate Classification	The process of sorting packets based only on the contents of the DS field.
Multi-Field Classification	The process of classifying packets based on the content of multiple fields such as source address, destination address, TOS byte, protocol ID, source port number, and destination port number.
Marking	The process of setting the DS field in a packet
Policing	The process of handling out of profile traffic, e.g., discarding excess packets
Shaping	The process of delaying packets within traffic stream to cause it to conform to some defined traffic profile.
Scheduling	The process of deciding which packet to send first in a system of multiple queues
Queue Management	Controlling the length of packet queues by dropping packets when necessary or appropriate
Traffic Trunk	An aggregation of flows with the same service class that can be put into a MPLS Label Switched Path
Bandwidth-delay product	The product of a data link's capacity (in bits per second) and its end-to-end delay (in seconds).

E1. Table Definitions

List of Figures

List of Tables

Glossary

ATM Asynchronous Transfer Mode.

CBR Constant Bit Rate.

DiffServ Differentiated Service.

FTP File Transfer Protocol.

HTML HyperText Markup Language.

HTTP Hyper-Text Transfer Protocol.

IntServ Integrated Service.

IP Internet Protocol.

Kbps Kilo Bit Per Second.

LAN Local Area Network.

MAC Media Access Control.

Mbps Mega Bit Per Second.

ms millisecond.

PDF Probability density function.

PHY Physical Layer.

POTS Plain Old Telephone Service.

RSVP ReSource Reservation Protocol.

RTP Real-time Transport Protocol.

TCP Transmission Control Protocol.

TCP/IP Network protocols used on the Internet.

UDP User Datagram Protocol.

VBR Variable Bit Rate.

VoIP Voice over IP.

VPN Virtual Private Network.

WAN Wide Area Network.

WWW World Wide Web.

Acronyms

AC Admission Control.

AF Assured Forwarding.

AMR Adaptive Multi-Rate Audio.

BA Behavior Aggregate.

BDP Bandwidth-delay product.

BE Best Effort.

CAC Connection Admission Control.

DL Downlink.

DS Differentiated Services.

DSCP Differentiated Services Code Point.

e2e End-to-End.

EF Expedited Forwarding.

IETF Internet engineering Task force.

MBAC Measurement-Based Admission Control.

MPEG Moving Picture Experts Group.

MPEG4 Moving Picture Experts Group 4 (Standard - Compressed Video at 64 Kbps).

PBAC Parametric-based Admission Control.

PHB Per-Hop-Behavior.

QoS Quality of Service.

SCGF Scaled-Cumulative Generating Function.

SLA Service Level Agreement.

TE Traffic Engineering.

ToS Type of Service.

UL Uplink.

List of Symbols

α the shape parameter of Pareto distribution.

η the mean value.

γ the parameter of the geometric distribution.

H Hurst parameter.

K queue capacity.

$\bar{K}$ mean queue length.

λ packet inter-arrival rate.

P_{Loss} one of the QoS parameters - the packet loss probability.

ρ utilization.

$R(k)$ the autocorrelation function.

σ the standard deviation.

τ_k the correlation interval.

$\bar{W}$ mean service waiting time.

x_m the scale parameter of Pareto distribution.

Bibliography

[1] 3GPP2-TSGC5: *Http and ftp traffic model for 1xev-dv. simulations*. Technical report, TS TSGC5, 2001.

[2] Adas, A. and Amarnath Mukherjee: *On resource management and qos guarantees for long range dependent traffic*. In *in Proc. IEEE INFOCOM '95*, pages 779–787, 1994.

[3] Adas, Abdelnaser Mohammad: *Using adaptive linear prediction to support real-time vbr video under rcbr network service model*. IEEE/ACM Trans. Netw., 6(5):635–644, 1998, ISSN 1063-6692.

[4] Addie, Ronald G.: *Fractal traffic: measurements, modelling and performance evaluation*. In *in Proc. IEEE INFOCOM '95*, pages 977–984, 1995.

[5] Allman, Mark and Ethan Blanton: *Notes on burst mitigation for transport protocols*. SIGCOMM Comput. Commun. Rev., 35(2):53–60, 2005, ISSN 0146-4833.

[6] Almquist, P.: *Type of service in the internet protocol suite*. Technical report, IETF, United States, 1992.

[7] Amer, P.D. and L.N. Cassel: *Management of sampled real-time network measurements*. In *Proceedings 14th Conference on Local Computer Networks*, pages 62–68, Mineapolis, MN, USA, 1989. ISBN 0-8186-1968-6.

[8] Amogh, Dhamdhere and Dovrolis Constantine: *Open issues in router buffer sizing*. SIGCOMM Comput. Commun. Rev., 36(1):87–92, 2006, ISSN 0146-4833.

[9] Appenzeller, Guido, Isaac Keslassy, and Nick McKeown: *Sizing router buffers*. In *IN PROCEEDINGS OF ACM SIGCOMM*, pages 281–292, 2004.

[10] Applied Network Research (NLANR), National Laboratory for: *Fix-west statistics data summaries*. Technical report, NLANR, June 1995. `http://www.nlanr.net/`.

[11] Arvidsson, A. and P. Karlsson: *On traffic models for tcp/ip*. In *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, pages 457–466, 1999.

[12] Asars, A., M. Kulikovs, and E. Petersons: *Buffer size and output bandwidth optimization in a mbac system*. Journal Automatic Control and Computer Sciences, 43(5):241–246, October 2009, ISSN 0146-4116.

[13] Asars, A. and E. Petersons: *Determining the optimal interval of the parameter identification of self-similar traffic*. Journal Automatic Control and Computer Sciences, 43(4):211–216, August 2009, ISSN 0146-4116.

[14] Athanasiadis, Thanos, Yannis Avrithis, and Stefanos Kollias: *The atm forum. traffic management specification version 4.0*. In *in Proceedings of 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM '06*, 1996.

[15] Awduche, D., A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao: *Overview and principles of internet traffic engineering*. Technical report, IETF, United States, 2002.

[16] Baker, F., C. Iturralde, F. Le Faucheur, and B. Davie: *Aggregation of rsvp for ipv4 and ipv6 reservations*. Technical report, IETF, United States, 2001.

[17] Bean, N. G.: *Statistical Multiplexing in Broadband Communication Networks*. PhD thesis, Statistical Laboratory, University of Cambridge, Junw 1993. PhD Dissertation.

[18] Bean, N. G.: *Robust connection acceptance control for atm networks with incomplete source information*. Annals of Operations Research, 48(4):357–379, August 1994.

[19] Bernet, Y., P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawski, and E. Felstaine: *A framework for integrated services operation over diffserv networks*. Technical report, IETF, United States, 2000.

[20] Bistrov, V. and E. Peterson: *Analytic estimation of packet loss probability in communication systems with self-similar input flow*. Journal Automatic Control and Computer Sciences, 42(4):197–202, August 2008.

[21] Bistrov, V. and E. Peterson: *Comparing batch and self-similar arrivals in communication systems*. In *Proceedings of the International Conference "Electronics'2008"*, May 2008.

[22] Blake, S., D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss: *An architecture for differentiated service*. Technical report, IETF, United States, 1998.

[23] Bolch, Gunter, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi: *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. Wiley-Interscience, New York, NY, USA, 1998, ISBN 0-471-19366-6.

[24] Bolotin, V. A.: *New subscriber traffic variability patterns for network traffic engineering*. In *In Proceedings of the 15th International Teletraffic Congress (ITC15)*, Monterey, CA, USA, June 1997.

[25] Bolotin, V.A.: *Modeling call holding time distributions for ccs network design and performance analysis*. IEEE Journal on Selected Areas In Communications, 12(3):433–438, April 1994.

[26] Braden, R.: *Requirements for internet hosts - communication layers*. Technical report, IETF, United States, 1989.

[27] Braden, R., D. Clark, and S. Shenker: *Integrated services in the internet architecture: an overview*. Technical report, IETF, United States, 1994.

[28] Brady, P. T.: *A statistical analysis of on-off patterns in 16 conversations*. Bell System Technical, 47(1):73–91, 1968.

[29] Brady, P. T.: *A model for generating on-off speech patterns in two-way conversations*. Bell System Technical, 48(9):2445–2472, 1969.

[30] Breslau, Lee, Sugih Jamin, and Scott Shenker: *Comments on the performance of measurement-based admission control algorithms*. In *Proc. IEEE INFOCOM 2000*, pages 1233–1242, April 2000.

[31] Buffet, E. and N.G. Duttield: *Exponential upper bounds via martingales for multiplexers with markovian arrivals*. Journal of Applied Probability, 31:1049–1061, 1994.

[32] Cardwell, Neal, Stefan Savage, and Thomas Anderson: *Modeling tcp latency*. In *in IEEE INFOCOM*, pages 1724–1751, 2000.

[33] Chernoff, H.: *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*. Annals of Math. Stat., 23:493–509, 1952.

[34] Choi, Baek Young, Jaesung Park, and Zhi Li Zhang: *Adaptive random sampling for load change detection*. In *SIGMETRICS '02: Proceedings of the 2002 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 272–273, New York, NY, USA, 2002. ACM, ISBN 1-58113-531-9.

[35] Choi, Bong Dae, Bara Kim, and In Suk Wee: *Asymptotic behavior of loss probability in gi/m/1/k queue as k tends to infinity*. Queueing Syst. Theory Appl., 36(4):437–442, 2000, ISSN 0257-0130.

[36] Chong, Song, San qi Li, and Joydeep Ghosh: *Predictive dynamic bandwidth allocation for efficient transport of real-time vbr video over atm*. IEEE Journal on Selected Areas in Communications, 13:12–23, 1995.

[37] Choudhury, Gagan L., David M. Lucantoni, and Ward Whitt: *Squeezing the most out of atm*. IEEE Transactions on Communications, 44(2):203–217, February 1996.

[38] Christin, N., J. Liebeherr, and T. F. Abdelzaher: *A quantitative assured forwarding service*. In *In IEEE Infocom*, pages 864–873, 2001.

[39] Claffy, Kimberly C., George C. Polyzos, and Hans Werner Braun: *Application of sampling methodologies to network traffic characterization*. In *SIGCOMM '93: Conference proceedings on Communications architectures, protocols and applications*, pages 194–203, New York, NY, USA, 1993. ACM, ISBN 0-89791-619-0.

[40] Crosby, Simon: *Performance Management in ATM Networks*. PhD thesis, Cambridge University Computer Laboratory, May 1995. Technical Report 393.

[41] Crovella, Mark E. and Azer Bestavros: *Self-similarity in world wide web traffic: evidence and possible causes*. IEEE/ACM Trans. Netw., 5(6):835–846, 1997, ISSN 1063-6692.

[42] Davie, B., A. Charny, J. C. R. Bennet, K. Benson, J. Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, and D. Stiliadis: *An expedited forwarding phb (per-hop behavior)*. Technical report, IETF, United States, 2002.

[43] Demers, Alan, Srinivasan Keshavt, and Scott Shenker: *Analysis and simulation of a fair queueing algorithm*. Computer Communications Review, 1989.

[44] Dhamdhere, A., H. Jiang, and C. Dovrolis: *Buffer sizing for congested internet links*. In *In roceedings of IEEE INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, 2005.

[45] Duffield, N. G., Pawan Goyal, Albert Greenberg, Partho Mishra, K. K. Ramakrishnan, and Jacobus E. van der Merive: *A flexible model for resource management in virtual private networks*. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 95–108, New York, NY, USA, 1999. ACM, ISBN 1-58113-135-6.

[46] Duffield, N.G., J.T. Lewis, Neil O'Connell, Raymond Russell, and Fergal Toomey: *Entropy of atm traffic streams: A tool for estimating qos parameters*, 1995.

[47] Elwalid, Anwar, Debasis Mitra, and Robert H. Wentworth: *A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an atm node*. IEEE Journal on Selected Areas in Communications, 13:1115–1127, 1995.

[48] Enachescu, Mihaela, Yashar Ganjali, Ashish Goel, Nick McKeown, and Tim Roughgarden: *Part III: routers with very small buffers*. SIGCOMM Comput. Commun. Rev., 35(3):83–90, 2005, ISSN 0146-4833.

[49] Erlang, A. K.: *Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges*. Post Office Electrical Engineers Journal, 10:189–197, 1917.

[50] Erramilli, Ashok, Onuttom Narayan, and Walter Willinger: *Experimental queueing analysis with long-range dependent packet traffic*. IEEE/ACM Trans. Netw., 4(2):209–223, 1996, ISSN 1063-6692.

[51] Estepa, A., R. Estepa, and J. Vozmediano: *A new approach for voip traffic characterization*. IEEE COMMUNICATIONS LETTERS, 8(10):644–646, 2004.

[52] Feldmann, A., A. C. Gilbert, W. Willinger, and T. G. Kurtz: *The changing nature of network traffic: scaling phenomena*. SIGCOMM Comput. Commun. Rev., 28(2):5–29, 1998, ISSN 0146-4833.

[53] Fischer, Martin J., Denise M. Bevilacqua Masi, Donald Gross, and John F. Shortle: *One-parameter pareto, two-parameter pareto, three-parameter pareto: Is there a modeling difference?* Telecommunications Review, pages 79–92, 2005.

[54] Floyd, S.: *Highspeed tcp for large congestion windows*. Technical report, IETF, United States, 2003.

[55] Floyd, Sally: *Comments on measurement-based admissions control for controlled-load services*. Technical report, Lawerence Berkeley Laboratory, July 1996.

[56] Fowler, Thomas B.: *Large deviations, the shape of the loss curve, and economies of scale in large multiplexers*. Journal Queueing Systems, 20(3-4):293–320, September 1995.

[57] Fowler, Thomas B.: *The difficulties of analyzing nonlinear systems and heavy-tailed queuing systems with transform methods*. Telecommunications Review, pages 107–119, 2002.

[58] Garrett, Mark W. and Walter Willinger: *Analysis, modeling and generation of self-similar vbr video traffic*. In *SIGCOMM '94: Proceedings of the conference on Communications architectures, protocols and applications*, pages 269–280, New York, NY, USA, 1994. ACM, ISBN 0-89791-682-4.

[59] Gibbens, R. J. and F. P. Kelly: *Measurement-based connection admission control*. In *In 15th Internation Teletraffic Congress Proceedings*, 16 Mill Lane, Cambridge, CB2 1SB, Jun 1997.

[60] Gibbens, R. J., F. P. Kelly, and P. B. Key: *A decision-theoretic approach to call admission control in atm networks*. IEEE Journal on Selected Areas In Communications, 13(6):1101–1114, August 1995.

[61] Goyal, Pawan, Harrick M. Vin, and Haichen Cheng: *Start-time fair queuing: A scheduling algorithm for integrated servicespacket switching networks*. Technical report, University of Texas at Austin, Austin, TX, USA, 1996.

[62] Grossglauser, Matthias and Jean Chrysostome Bolot: *On the relevance of long-range dependence in network traffic*. IEEE/ACM Trans. Netw., 7(5):629–640, 1999, ISSN 1063-6692.

[63] Grossglauser, Matthias and David N. C. Tse: *A framework for robust measurement-based admission control*. IEEE/ACM Trans. Netw., 7(3):293–309, 1999, ISSN 1063-6692.

[64] Grossglauser, Matthias and David N. C. Tse: *A time-scale decomposition approach to measurement-based admission control*. IEEE/ACM Trans. Netw., 11(4):550–563, 2003, ISSN 1063-6692.

[65] Guerin, R., H. Ahmadi, and M. Naghshineh: *Equivalent capacity and its application to bandwidth allocation in high-speed networks*. IEEE Journal on Selected Areas In Communications, 9(7):968–981, September 1991.

[66] Habib, I.W. and T.N. Saadawi: *Multimedia traffic characteristics in broadband networks*. Communications Magazine, IEEE, 30(7):48–54, 1992, ISSN 0163-6804.

[67] Halsall, Fred: *Data Communications Computer Networks and Open Systems 3rd Ed.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1992, ISBN 0201565064.

[68] Heinanen, J., F. Baker, W. Weiss, and J. Wroclawski: *Assured forwarding phb group*. Technical report, IETF, United States, 1999.

[69] Hernandez, Edwin A., Matthew C. Chidester, and Alan D. George: *Adaptive sampling for network management*. J. Netw. Syst. Manage., 9(4):409–434, 2001, ISSN 1064-7570.

[70] Hoeffding, W.: *Probability inequalities for sums of bounded random variables*. Journal of the American Statistical Association, 58(301):13–30, March 1963.

[71] HORI, Yoshiaki, Hidenari SAWASHIMA, Hideki SUNAHARA, and Yuji OIE: *Performance evaluation of udp traffic affected by tcp flows*. IEICE TRANSACTIONS on Communications, 81(8):1616–1623, 1998.

[72] Hsu, Ivy and Jean Walrand: *Dynamic bandwidth allocation for atm switches*. Journal of Applied Probability, 33:758–771, 1995.

[73] Ilnickis, S. and E.Petersons: *Nonstacionary behavior research of terminal-server system with self-similar approximated input flow*. Journal Automatic Control and Computer Sciences, 39(4):48–59, 2005.

[74] Jaffe, Joseph, Louis Cassotta, and Stanley Feldstein: *Markovian model of time patterns of speech*. Science Magazine, 144(3620):1049–1061, May 1994.

[75] Jamin, Sugih, Peter B. Danzig, Scott Shenker, and Lixia Zhang: *A measurement-based admission control algorithm for integrated services packet networks*. In *SIGCOMM '95: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 2–13, New York, NY, USA, 1995. ACM, ISBN 0-89791-711-1.

[76] Jamin, Sugih and Scott Shenker: *Measurement-based admission control algorithms for controlled-load service: A structural examination*. In *IEEE/ACM Transactions on Networking*, pages 56–70, 1995.

[77] Jamin, Sugih, Scott Shenker, Lixia Zhang, and David D. Clark: *An admission control algorithm for predictive real-time service (extended abstract)*. In *Proceedings of the Third International Workshop on Network and Operating System Support for Digital Audio and Video*, pages 349–356, London, UK, 1993. Springer-Verlag, ISBN 3-540-57183-3.

[78] Jamin, Sugih, Scott J. Shenker, and Peter B. Danzig: *Comparison of measurement-based admission control algorithms for controlled-load service*. In *INFOCOM '97: Proceedings of the INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, page 973, Washington, DC, USA, 1997. IEEE Computer Society, ISBN 0-8186-7780-5.

[79] Jiang, Hao and Constantinos Dovrolis: *Source-level ip packet bursts: causes and effects*. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pages 301–306, New York, NY, USA, 2003. ACM, ISBN 1-58113-773-7.

[80] Jiang, Hao and Constantinos Dovrolis: *Why is the internet traffic bursty in short time scales?* In *SIGMETRICS '05: Proceedings of the 2005 ACM SIGMETRICS interna-*

tional conference on Measurement and modeling of computer systems, pages 241–252, New York, NY, USA, 2005. ACM, ISBN 1-59593-022-1.

[81] Jiang, Yuming, Peder J. Emstad, Anne Nevin, Victor Nicola, and Markus Fidler: *Measurement-based admission control for a flow-aware network*. In *EuroNGI 1st Conference on Next Generation Internet Networks (NGI)*, 2005.

[82] Jiang, Yuming, Peder J. Emstad, Victor Nicola, and Anne Nevin: *Measurement-based admission control: A revisit*. In *Seventeenth Nordic Teletraffic Seminar, Fornebu*, pages 25–27, 2004.

[83] Kalyanaraman, Shivkumar: *Better-than-best-effort: Qos, int-serv, diff-serv, rsvp, rtp.*, 2001. ip2001-Lecture14-6pp.pdf.

[84] Kamienski, C. A.: *An Architecture for Providing End-to-End QoS-based Advanced Services in the Internet*. PhD thesis, Federal University of Pernambuco, Brazil, February 2003. PhD Dissertation.

[85] Kelly, Frank: *Notes on effective bandwidths*. Stochastic Networks: Theory and Applications, pages 141–168, 1996.

[86] Kesidis, George: *Bandwidth adjustments using on-line packet-level measurements*. In *SPIE Conf. Performance and Control of Network Systems*, 1999.

[87] Key, P.B.: *Connection admission control in atm networks*. BT Technology Journal, 13(3):52–66, July 1995.

[88] Kim, Bong Ho: *Application traffic model*. In *WiMAX Forum AWG Contribution*, May 2007.

[89] Kim, Bong Ho, Jungnam Yun, Yerang Hur, Chakchai So-In, Raj Jain, and Abdel Karim Al Tamimi: *Capacity estimation and tcp performance enhancement over mobile wimax networks*. Comm. Mag., 47(6):132–141, 2009, ISSN 0163-6804.

[90] Kim, Han S. and Ness B. Shroff: *Loss probability calculations and asymptotic analysis for finite buffer multiplexers*. IEEE/ACM Trans. Netw., 9(6):755–768, 2001, ISSN 1063-6692.

[91] Knightly, E. W.: *Traffic Models and Admission Control for Integrated Services Networks*. PhD thesis, Department of EEC, University of California, 1996. PhD Dissertation.

[92] Krunz, Marwan and Satish K. Tripathi: *On the characterization of vbr mpeg streams*. In *in Proc. ACM SIGMETRICS*, pages 192–202, 1997.

[93] Kulikovs, M. and E. Petersons: *Packet loss probability dependence on number of on-off traffic sources in opnet*. ELECTRONICS AND ELECTRICAL ENGINEERING, 85(5):77–80, 2008.

[94] Kulikovs, Mihails and Ernests Petersons: *Remarks regarding queuing model and packet loss probability for the traffic with self-similar characteristics*. International Journal of Computer Science, 3(2):84–90, 2008.

[95] Kuokkwee, Wee, Mohamed Othman, Subramaniam Shamala, and Ahmad Ariffin: *Enhanced dynamic bandwidth allocation proportional to queue length with threshold value for vbr traffic*. The International Arab Journal of Information Technology, 4(2):117–124, 2007, ISSN 1683-3198.

[96] Leland, Will E., Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson: *On the self-similar nature of ethernet traffic (extended version)*. IEEE/ACM Trans. Netw., 2(1):1–15, 1994, ISSN 1063-6692.

[97] Liao, Raymond R. F. and Andrew T. Campbell: *Dynamic core provisioning for quantitative differentiated services*. IEEE/ACM Trans. Netw., 12(3):429–442, 2004, ISSN 1063-6692.

[98] Likhanov, N.: *Bounds on the buffer occupancy probability with self-similar input traffic*, pages 193–213. Self-Similar Network Traffic and Performance Evaluation. John Wiley & Sons, New York, K. Park and W. Willinger edition, 2000.

[99] Likhanov, N., B. Tsybakov, and N. D. Georganas: *Analysis of an atm buffer with self-similar ("fractal") input traffic*. In *INFOCOM '95: Proceedings of the Fourteenth Annual Joint Conference of the IEEE Computer and Communication Societies (Vol. 3)-Volume*, page 985, Washington, DC, USA, 1995. IEEE Computer Society, ISBN 0-8186-6990-X.

[100] Loh, Chee Heok: *Dynamic bandwidth allocation in atm network*. Master's thesis, University Putra, Malaysia, 2001.

[101] Ma, Qingming and Peter Steenkiste: *Quality-of-service routing for traffic with performance guarantees*. In *In Proc. IFIP International Workshop on Quality of Service*, pages 115–126, 1997.

[102] Ma, Wenhong, James Yan, and Changcheng Huang: *Adaptive sampling methods for network performance metrics measurement and evaluation in mpls-based ip networks*. In *In proceedings of Canadian Conference on Electrical and Computer Engineering, 2003. IEEE CCECE 2003*, volume 2, pages 1005–1008, 2003, ISBN 0-7803-7781-8.

[103] Ma, Wenhong, James Yan, and Changcheng Huang: *Adaptive sampling methods for network performance measurement under voice traffic*. In *Proc. of IEEE ICC*, pages 1129–1134, 2004.

[104] Mandelbrot, Benoit: *Long-run linearity, locally gaussian process, h-spectra and infinite variances*. International Economic Review, 10(1):82–111, February 1969. `http://ideas.repec.org/a/ier/iecrev/v10y1969i1p82-111.html`.

[105] Mandelbrot, Benoit B.: *Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier*. Journal of Fluid Mechanics, 62(2):331–358, 1974.

[106] Mase, K.: *Toward scalable admission control for voip networks*. IEEE Communications Magazine, 42(7):42–47, July 2004.

[107] Mathis, Matthew, Jeffrey Semke, Jamshid Mahdavi, and Teunis Ott: *The macroscopic behavior of the tcp congestion avoidance algorithm*. SIGCOMM Comput. Commun. Rev., 27(3):67–82, 1997, ISSN 0146-4833.

[108] Matrawy, Ashraf, Ioannis Lambadaris, and Changcheng Huang: *Mpeg4 traffic modeling using the transform expand sample methodology*. In *In Proc. of 4th IEEE IWNA4*, pages 249–256, 2002.

[109] Mena, A. and J. Heidemann: *An empirical study of real audio traffic*. In *Proc. IEEE INFOCOM 2000*, pages 1001–1010, April 2000.

[110] Minoli, D.: *Issues in packet voice communications*. In *Proceedings of the Institution of Electrical Engineers*, August 1979.

[111] Molina, E. C.: *Modeling call holding time distributions for ccs network design and performance analysis*. Bell System Technical, 6:461–494, 1927.

[112] Morin, Patrick R.: *The impact of self-similarity on network performance analysis*. Technical report, Carleton University, Dec 1995.

[113] Nichols, K., S. Blake, F. Baker, and D. Black: *Definition of the differentiated services field (ds field) in the ipv4 and ipv6 headers*. Technical report, IETF, United States, 1998.

[114] Nichols, K., V. Jacobson, and L. Zhang: *A two-bit differentiated services architecture for the internet*. Technical report, IETF, United States, 1999.

[115] Norros, I.: *A storage model with self-similar input*. Queueing System, 16:387–396, 1994.

[116] Othman, Mohamed, Chee Heok Loh, A. K. Ramani, D. Shyamala, and L. N. Abdullah: *Dynamic bandwidth allocation with low buffer storage in atm switch*. Journal of Electrical Engineering, 2(4):62–67, 2001, ISSN 0128-4428.

[117] Padhye, J., V. Firoiu, D. Towsley, and J. Kurose: *Modeling tcp throughput: A simple model and its empirical validation*. Technical report, University of Massachusetts, Amherst, MA, USA, 1998.

[118] Paksoy, E., J. Carlos de Martin, A. McCree, C. G. Gerlach, A. Anandakumar, Wai Ming Lai, and V. Viswanathan: *An adaptive multi-rate speech coder for digital cellular telephony*. In *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, pages 193–196, Washington, DC, USA, 1999. IEEE Computer Society, ISBN 0-7803-5041-3.

[119] Palazzo, R. C.: *A lyapunov approach to dynamic bandwidth allocation in b-isdn*. Master's thesis, School of Electrical Engineering, Clemson University, 1994.

[120] Pan, Davis Yen: *Digital audio compression*. Digital Tech. J., 5(2):28–40, 1993, ISSN 0898-901X.

[121] Park, Kihong: *Afec: An adaptive forward error-correction protocol and its analysis*. In *In Proc. IEEE IC3N*, pages 196–205, 1997.

[122] Park, Kihong, Gitae Kim, and Mark Crovella: *On the relationship between file sizes, transport protocols, and self-similar network traffic*. Technical report, Boston University, Boston, MA, USA, 1996.

[123] Park, Kihong, Gitae Kim, and Mark Crovella: *On the effect of traffic self-similarity on network performance*. In *In Proceedings of the SPIE International Conference on Performance and Control of Network Systems*, pages 296–310, 1997.

[124] Park, Kihong and Walter Willinger: *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., New York, NY, USA, 2000, ISBN 0471319740.

[125] Park, Kihong and Walter Willinker: *SELF-SIMILAR NETWORK TRAFFIC: AN OVERVIEW*, pages 1–38. Self-Similar Network Traffic and Performance Evaluation. John Wiley & Sons, New York, K. Park and W. Willinger edition, 2000.

[126] Paxson, Vern and Sally Floyd: *Wide-area traffic: The failure of poisson modeling*. IEEE/ACM Transactions on Networking, 3:226–244, 1995.

[127] Perros, H.G. and K.M.F. Elsayed: *Call admission control schemes: A review*. IEEE Magazine on Communications, 34(11):82–91, 1996.

[128] Pitsillides, Andreas, Petros Ioannou, and Loukas Rossides: *Congestion control for differentiated-services using non-linear control theory*. In *In Proceedings of the Sixth IEEE Symposium on Computers and Communications*, pages 726–733, Washington, DC, USA, 2001. IEEE Computer Society.

[129] Postel, J.: *User datagram protocol*. Technical report, IETF, United States, 1980.

[130] Qiu, Jingyu and Edward W. Knightly: *Qos control via robust envelope-based mbac*. In *In proceedings of Sixth International Workshop on Quality of Service (IEEE/IFIP IWQoS '98)*, pages 62–64, May 1998, ISBN 0-7803-4482-0.

[131] Qiu, Jingyu and Edward W. Knightly: *Measurement-based admission control with aggregate traffic envelopes*. IEEE/ACM Trans. Netw., 9(2):199–210, 2001, ISSN 1063-6692.

[132] Qiu, Jingyu and Edward W. Knightly: *Measurement-based admission control with aggregate traffic envelopes*. IEEE/ACM Trans. Netw., 9(2):199–210, 2001, ISSN 1063-6692.

[133] Rahat, Amitava, Nicholas Malcolm, and Wei Zhaot: *Hard real-time communications with weighted round robin service in atm local area networks*. In *ICECCS '95: Proceedings of the 1st International Conference on Engineering of Complex Computer Systems*, page 96, Washington, DC, USA, 1995. IEEE Computer Society, ISBN 0-8186-7123-8.

[134] Raina, G. and D. Wischik: *Buffer sizes for large multiplexers: Tcp queueing theory and instability analysis*. In *Next Generation Internet Networks '05*, pages 173–180, Rome, Italy, 2005.

[135] Raina, Gaurav, Don Towsley, and Damon Wischik: *Part II: control theory for buffer sizing*. SIGCOMM Comput. Commun. Rev., 35(3):79–82, 2005, ISSN 0146-4833.

[136] Rampal, S., D.S Reeves, and I. Viniotis: *Dynamic resource allocation based on measured qos*. Technical report tr 96-2, Center for Advanced Computing and Commun., North Carolina State University, United States, Jan 1997.

[137] Riedi, R. and J. Levy Vehel: *Tcp traffic is multifractal: a numerical study*, 1997.

[138] Riedi, Rudolf H., Matthew S. Crouse, Vinay J. Ribeiro, and Richard G. Baraniuk: *A multifractal wavelet model with application to network traffic*. IEEE TRANSACTIONS ON INFORMATION THEORY, 45:992–1018, 1998.

[139] Rodriguez-Dagnino, R. M.: *Some remarks regarding asymptotic packet loss in the pareto/m/1/k queueing system*. IEEE COMMUNICATIONS LETTERS, 9(10):927–929, 2005.

[140] Rosen, E., A. Viswanathan, and R. Callon: *Multiprotocol label switching architecture*. Technical report, IETF, United States, 2001.

[141] Rushby, John: *Systematic formal verification for fault-tolerant time-triggered algorithms*. IEEE Trans. Softw. Eng., 25(5):651–660, 1999, ISSN 0098-5589.

[142] Ryu, Bong K. and Anwar Elwalid: *The importance of long-range dependence of vbr video traffic in atm traffic engineering: myths and realities*. SIGCOMM Comput. Commun. Rev., 26(4):3–14, 1996, ISSN 0146-4833.

[143] Sahinoglu, Zafer and Sirin Tekinay New: *On multimedia networks: Self-similar traffic and network performance*. IEEE Communications Magazine, 37:48–52, 1999.

[144] Sahinoglu, Zafer and S. Tekinay: *A novel adaptive bandwidth allocation: wavelet-decomposed signal energy approach*. In *IEEE Global Telecommunications Conference, 2001. GLOBECOM '01*, pages 2253 – 2257, Murray Hill, NJ, 2001. IEEE Computer Society, ISBN 0-7803-7206-9.

[145] Saito, H.: *Call admission control in an atm network using upper bound of cell loss probability*. IEEE Transactions on Communications, 40(9):1512–1521, 1992.

[146] Saito, H. and K. Shiomoto: *Dynamic call admission control in atm networks*. IEEE Journal on Selected Areas In Communications, 9(7):982–989, September 1991.

[147] Sheldon, Tom: *McGraw-Hill's Encyclopedia of Networking and Telecommunications*. McGraw-Hill Professional, 2001, ISBN 0072120053.

[148] Shenker, S., C. Partridge, and R. Guerin: *Specification of guaranteed quality of service*. Technical report, IETF, United States, 1997.

[149] Siripongwutikorn, P., S. Banerjee, and D. Tipper: *A survey of adaptive bandwidth control algorithms*. IEEE Communication Surveys, 5(1):14–26, 2003.

[150] Siripongwutikorn, Peerapon, Sujata Banerjee, and David Tipper: *Adaptive bandwidth control for efficient aggregate qos provisioning*. SIGCOMM Comput. Commun. Rev., 32(3):19–23, 2002, ISSN 0146-4833.

[151] Sivaradje, G. and P. Dananjayan: *Efficient resource allocation scheme for real-time mpeg video traffic over atm networks*. In *ICCS '02: Proceedings of the The 8th International Conference on Communication Systems*, pages 747–751, Washington, DC, USA, 2002. IEEE Computer Society, ISBN 0-7803-7510-6.

[152] Southern California Information Sciences Institute & Defense Advanced Research Projects Agency, University of: *Internet protocol*. Technical report, IETF, United States, 1981.

[153] Stiliadis, Dimitrios and Anujan Varma: *Efficient fair queueing algorithms for packet-switched networks*. IEEE/ACM Trans. Netw., 6(2):175–185, 1998, ISSN 1063-6692.

[154] Takano, Ryousei, Tomohiro Kudoh, Yuetsu Kodama, Motohiko Matsuda, Hiroshi Tezuka, and Yutaka Ishikawa: *Design and evaluation of precise software pacing mechanisms for fast long-distance networks*. In *In Proceedings of PFLDNet 2005*, 2005.

[155] Tang, N., S. Tsui, and L. Wang: *A survey of admission control algorithms*. Technical report, Computer Science Department, University of California Los Angeles (UCLA), Dec 1998.

[156] Tanthawichian, P., A. Fujii, and Y. Nemoto: *Bandwidth allocation in atm networks: Heuristic approach*. In *IC3N '98: Proceedings of the International Conference on Computer Communications and Networks*, page 20, Washington, DC, USA, 1998. IEEE Computer Society, ISBN 0-8186-9014-3.

[157] Tedijanto, T.E. and L. Gun: *Effectiveness of dynamic bandwidth management mechanisms in atm networks*. In *In proceedings of Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM '93.*, volume 1, pages 358 – 367, 1993, ISBN 0-8186-3580-0.

[158] Tse, D. and M. Grossglauser: *Measurement-based call admission control: Analysis and simulation*. In *INFOCOM '97: Proceedings of the INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, page 981, Washington, DC, USA, 1997. IEEE Computer Society, ISBN 0-8186-7780-5.

[159] Tuan, Tsunyi and Kihong Park: *Multiple time scale congestion control for selfsimilar network traffic*. Performance Evaluation, 36:359–386, 1999.

[160] Ulanovs, P. and E. Peterson: *Modelling methods of self-similar traffic for network performance evaluation*. In *Scientific Proceedings of RTU, Series 7, Telecommunications and Electronics*, pages 40–49, 2002.

[161] Veres, Andras and Miklos Boda: *The chaotic nature of tcp congestion control*. In *Proc. IEEE INFOCOM 2000*, pages 1715–1723, April 2000.

[162] Vicari, N. and S. Köhler: *Measuring internet user traffic behavior dependent on access speed*. In *ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, USA, September 2000.

[163] Villamizar, Curtis and Cheng Song: *High performance tcp in ansnet*. SIGCOMM Comput. Commun. Rev., 24(5):45–60, 1994, ISSN 0146-4833.

[164] Wang, Yao Tzung, Tzung Pao Lin, and Kuo Chung Gan: *An improved scheduling algorithm for weighted round-robin cell multiplexing in an atm switch*. In *IEEE International Conference on Communications, ICC '94, SUPERCOMM/ICC '94, Conference Record, 'Serving Humanity Through Communications.'*, pages 1032–1037. IEEE Computer Society, 1994, ISBN 0-7803-1825-0.

[165] Wischik, Damon and Nick McKeown: *Part I: buffer sizes for core routers*. SIGCOMM Comput. Commun. Rev., 35(3):75–78, 2005, ISSN 0146-4833.

[166] Wroclawski, J.: *Specification of the controlled-load network element service*. Technical report, IETF, United States, 1997.

[167] Wucher, Karen: *The internet singularity, delayed: Why limits in internet capacity will stifle innovation on the web*. Technical report, Nemertes Research, November 2007.

[168] Xiao, Xipeng: *Providing quality of service in the internet*. PhD thesis, Michigan State University. Dept. of Computer Science and Engineering, East Lansing, MI, USA, 2000, ISBN 0-599-77251-4. Adviser-Ni, Lionel M.

[169] Xiao, Xipeng and Lionel M. Ni: *Internet qos: A big picture*. IEEE Network, 13:8–18, 1999.

[170] Zhang, L., S. Berson, S. Herzog, and S. Jamin: *Resource reservation protocol (rsvp) - version 1 functional specification*. Technical report, IETF, United States, 1997.

[171] Столлингс, В.: Современные компьютерные сети. СПб.: Питер, 2 редакция, 2003.

[172] Цилькин, Я.З.: Основы информационной теории информации. М.: Наука, 1984.

[173] Вазан, М.: Стохастическая аппроксимация. М.: Мир, 1973.

[174] Липаев, В. В. и С. Ф. Яшков: Эффективность методов организации вычислительного процесса в АСУ. М.: Статистика, 1975.

[175] Джамалипур, Аббас: Беспроводной мобильный Интернет. Архитектура, протоколы и сервисы. М.: Техносфера, 2009, ISBN 978-5-94836-115-4.