S. Mirtchev, R. Goleva. Discrete Time Single Server Queueing Model whit a Multimodal Packet Size Distribution. In: T. Atanasova (Editor) Proceedings of a Conjoint Seminar "Modeling and Control of Information Processes", Sofia, Bulgaria, 2009, CTP, Sofia, 2009, ISBN: 978-954-9332-55-1, pp. 83-101

DISCRETE TIME SINGLE SERVER QUEUEING MODEL WITH A MULTIMODAL PACKET SIZE DISTRIBUTION

Seferin Mirtchev, Rossitza Goleva
Department of Communication Networks,
Technical University - Sofia
stm@tu-sofia.bg, rig@tu-sofia.bg

ABSTRACT

The importance of IP network will further increase and it will serve as a platform for more and more services, requiring different types and degrees of service quality. In this paper, we investigate the queueing behaviour found in input buffers in IP switch or router. To analyze these types of behaviour, we study the discrete-time version of the "classical" queue model M/M/1/k called Geo/MM/1/k and P/MM/1/k. We use a geometric (Geo) and Pareto (P) distributed inter-arrival time and a discrete multimodal packet length distribution (MM), defined on a range between a minimum and maximum value. In this model, the time unit is reduced from a packet service time to the time to transmit a byte. We develop balance equations for the state of the system, from which we derive packet delay and loss results. The finding that Geometric and Pareto distributions are adequate to model Internet packet inter-arrival times has motivated the proposal of methods to evaluate steady-state performance measures of Geo/MM/1/k queue. We propose a real time trace simulation model for estimating the steady-state probability showing the loss probability and delay of the P/MM/1/k queue. This model can be used to study the long-tailed queueing systems.

Categories and Subject Descriptors:
C.4 [PERFORMANCE OF SYSTEMS]: Modelling techniques
G.3 [PROBABILITY AND STATISTICS]: Queueing theory
I.6 [SIMULATION AND MODELING]: I.6.8 Types of Simulation, Discrete event
General Terms: *Performance, Design, Experimentation, Theory, Verification.*
Keywords: *Pareto distribution, discrete time queue, simulation model, queueing analyses, packet size distribution.*

INTRODUCTION

Managed IP networks have become a dominant factor in bringing information to users worldwide. Until recently, IP networks supported only a best effort service. This limitation has not been a problem for traditional Internet applications like web and email, but it does not satisfy the needs of many new applications like audio and video streaming, which demand high data throughput capacity (bandwidth) and have low-latency requirements. Thus, it is becoming increasingly important to provide Quality of Service (QoS) in managed IP networks.

As pointed out by several authors who have been collecting traffic data from the Internet, there is no a queueing theory method for queue analyses when one is given a set of packet inter-arrival times. Obviously, one could fit the resulting data to a distribution and then use a queueing model if it exists. There are some papers concerning batch arrivals like [1]. Traffic growth and its influence to the congestion management are demonstrated in [2]. Internet traffic can be described as having one or more of the following related characteristics [3], [4]: Self-similar (or fractal) traffic traces; Long-range dependence; Burstiness on multiple scales; Long- or heavy-tailed packet inter-arrival times or service requirements.

INTERNET TRAFFIC AND SELF-SIMILAR PROCESS

The Internet traffic data are well known to possess extreme variability and bursty structure in a wide range of time scales. This characteristic is not found on the Poisson process. The properties can be characterized by self-similar process. The large variation pertaining to the self-similar nature of data traffic causes congestion problems in the data network. The arrival process with Pareto distributed inter-arrival time is a popular model of self-similar processes.

The queue performance of P/M/1/k was studied by simulations in [5], [6]. They are investigated the queue behaviour with Pareto inter-arrival distribution. By numerical analysis and simulations, they have been analyzed the asymptotic and the exact loss probabilities of GI/M/1/k to show the big discrepancy between the asymptotic and the actual loss probability and propose a model for the loss probability of P/M/1/k as a function of the buffer size and the geometric parameter.

The Pareto distribution is a model for nonnegative data with a power law probability tail. A natural upper bound truncates the probability tail in many practical applications. An estimator is derived for the truncated Pareto distribution in [7]. They investigate distribution properties and illustrate its applicability in practice.

SYSTEM SIMULATION

Some limited analytical derivation for queueing models with Pareto distribution is proposed in the literature, but their solutions are often of a great mathematical challenge. To overcome such limitations, simulation tools that can deal with general

queueing systems have to be developed. Despite certain limitations, simulation algorithms provide a mechanism to obtain insight and good numerical approximation to parameters of networks of queues. The Internet traffic simulation is a difficult task due to the heterogeneous structure, immense size, and changing property of the network [8], [9].

Simulations that use traces generated by network traffic models usually examine a single node in the network, such as a router or switch; factors that depend on specific network topologies or routing information are specific to those topologies and simulations.

The simulation of systems using heavy-tailed distributions presents difficulties and needs efficient methods to study. In [10] there is a trial to go into insight nature of simulation difficulties of M/G/n queues with G heavy-tailed distribution. They have proposed and developed a method to speed up simulations and used M/G/1 systems as workbenches since they have some analytical results to check the simulation grades.

Stochastic simulation has become a well established paradigm used in performance evaluation of various complex dynamic systems. In [11] a method for estimating time evolution of several quantiles within some time interval is described. It is based on independent replications and its capability is demonstrated by simulating processes with different kinds of stationary, non-stationary or transient behaviour.

The concept of self-similarity (or fractal behaviour) is the best understood by looking at [12]. The use of synthetic self-similar traffic in computer networks simulation is of vital importance for the capturing and reproducing of actual Internet data traffic behaviour. Fernandes uses a technique for self-similar traffic generation that is achieved by aggregating On/Off sources where the active (On) and idle (Off) periods exhibit heavy tailed distributions. This work analyzes the balance between accuracy and computational efficiency in generating self-similar traffic and presents important results that can be useful to parameterize existing heavy tailed distributions such as Pareto, Weibull and Lognormal in a simulation analysis.

The Pareto distribution is a special heavy tailed distribution called a power-tailed distribution. It is found to serve as adequate model for many situations. In [13] many difficulties in simulating queues with Pareto service are investigated. They considered truncated Pareto service.

A method for studying Pareto queues is presented in [14]. The paper discusses the properties and use of the Pareto distribution. The method is used to study the Pareto/M/1 queue and look at the M/Pareto/1 queue. The first could be used to model arrivals of packets at a packet switched network, and the second, the time to transmit files through such a network.

The Pareto distribution has various forms. A one and two-parameter form is considered in [15]. The two Pareto forms are studied in detail. It is shown that the

usage of the two-parameter Pareto results in lower congestion than the comparable one-parameter Pareto.

A modelling and simulation approach using heavy-tailed mixture distributions is introduced in [16]. Their approach is used to build analytical models for random variables of several major Internet applications of a campus network. Several statistical features of an NS2 simulation are compared against those of the traffic traces being simulated. The comparison indicates that the simulation is statistically similar to the real traffic.

This paper presents a stochastic simulation method for studying Pareto queues. The paper discusses the properties and use of the Pareto distribution. We make the comparison between P/MM/1/k and Geo/MM/1/k and propose a real time trace simulation model for estimating the steady-state probability showing the tail-raising effect, the loss probability and delay. The data collected from network helps to do the evaluation correctly. The model can be used to study the long-tailed queueing systems.

PACKET LENGTH DISTRIBUTION

There are two major parameters generated by network traffic models: packet length distributions and packet inter-arrival distributions. Other parameters, such as routes, distribution of destinations, etc., are of less importance.

The packet queueing in an IP router arises because multiple streams of packets from different input ports are being multiplexed together over the same output port. A key characteristic is that the packets have different length. The minimum header size in IPv4 is 20 octets, and in IPv6 it is 40 octets. The maximum packet size depends on the specific sub-networks technology: 1500 octets in IPv4, 1280 octets in IPv6, more in Ethernet and Frame Relay, 1000 octets in X.25 networks.

The packet length distribution measured from the real traces exhibits the well-known multi-mode behaviour, with peaks for very short packets because of UDP traffic and acknowledgements. For the different maximum transfer units in the network the dominating peak is at 1500 bytes, due to the size of Ethernet frame and TCP services. This specific packet length distribution has a direct impact on the service time and we need a different approach to the queueing analysis in both cases.

The packet size distribution is one of the features that exhibit the most consistent behaviour, even among different networks. The packets are typically concentrated around the 40, 576, and 1500 byte regions and these values represent the IP packet header size and small UDP segments, the maximum transfer unit supported by most of IP routers and maximum packet size for IP networks.

DISCRETE TIME QUEUE

Discrete-time queueing systems have been a research topic for several decades now and there are many reference works on discrete-time queueing theory. Over the

years, different methodologies have been developed to assess the performance of queueing systems. The two main analytical approaches are the matrix analytic method and the transform method for discrete and for continuous-time analyses. Many authors have considered the Geo/G/1 queueing system [17], [18], [19], [20].

In [21] a complete study of a discrete-time single-server queue with geometrical arrivals of both positive and negative customers is carried out. Negative arrivals are used as a control mechanism in many telecommunication and computer networks. The study of a discrete-time single-server retrial queue with geometrical inter-arrival times and a phase-type service process is concerned in [22]. An iterative algorithm to calculate the stationary distribution of Markov chain is given.

Salvador in [4] proposes a traffic model and a parameter fitting procedure that are capable of achieving accurate prediction of the queuing behaviour for IP traffic exhibiting long-range dependence. The modelling process is a discrete-time Batch Markovian Arrival Process (dBMAP) that jointly characterizes the packet arrival process and the packet size distribution. In the proposed dBMAP, packet arrivals occur according to a discrete-time Markov Modulated Poisson Process (dMMPP) and each arrival is characterized by a packet size with a general distribution that may depend on the phase of the dMMPP.

In [3] Cao presents an introduction to bandwidth estimation and a solution to the problem of the best-effort traffic for the case where the quality criteria specify negligible packet loss. The solution is a simple statistical model, which is built and validated using queueing theory and extensive empirical study.

It has been shown [23] that in the case of real-time communications, for which small buffers are used for delay reasons, short range dependence dominates the loss process and so the Markov-Modulated Poisson Process (MMPP) might be a reasonable source model. They have presented an exact mathematical model for the loss process of a MMPP+M/Ek/1/K queue and have concluded that the packet size distribution affects the packet loss process and thus the efficiency of forward error correction.

In this paper, we investigate the basic queueing behaviour of packets found in IP output buffers. It is complicated because multiple streams of packets are being multiplexed together. The traffic is being generated from the packets of varying sizes that arrive for transmission on the link. The packets can wait or be dropped if their size is bigger than the free positions in the buffer. The quality metrics for the traffic on the Internet are the packets loss and delay. To analyze these types of behaviour, we study the discrete-time version of the “classical” queue model M/M/1/k called Geo/MM/1/k, where MM denotes a discrete multimodal packet length distribution.

BALANCE EQUATIONS FOR THE Geo/MM/1/K QUEUE

Let us consider a single server finite queue delay system Geo/MM/1/k with a geometric distributed inter-arrival time and a multimodal distributed packet length. The packet length distributions are defined on a range between a minimum and

maximum value.

We consider queueing phenomena in discrete-time queueing systems. We assume a fundamental time unit (time slot), the time to transmit an octet (byte), T_b. Customers arrive in the queueing system under consideration during the consecutive slots, but they can only start service at the beginning of slots. That is, service of customers is synchronized with respect to slot boundaries. Further, customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. During the consecutive slots, packets arrive in the system, are stored in a finite capacity queue and are served by a single server on a first in first out (FIFO) basis (Fig. 1).

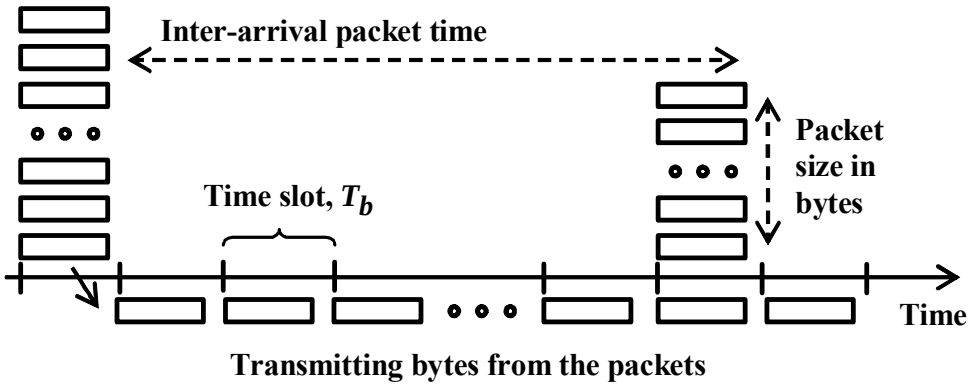

Fig. 1. Timing of events in the Geo/MM/1/k and P/MM/1/k queue.

We use a Bernoulli process for the packet arrivals, i.e. a geometrically distributed number of slots between arrivals. Let the probability that a packet arrives in an octet slot is p.

Thus we have a batch arrival process with geometrically distributed inter-arrival times. That is, the number of slots that separate consecutive slots where there are customer arrivals, constitute a series of independent and identically geometrically distributed random variables.

The probability of no octets arriving in a time slot is

$$a_0 = 1 - p \tag{1}$$

In this model, we assume three modal packet size distributions with a minimum value m_1 with probability b_1, a maximum value m_2 with probability b_2, and a value between minimum and maximum m_3 with probability b_3.

The mean number of bytes in the packet is:

$$b = m_1 b_1 + m_3 b_3 + m_2 b_2 \tag{2}$$

The probability that *n* octets arrive in a time slot is

$$a_n = p\, b_n, \qquad n = 1,\ 2 \ \ or \ \ 3 \tag{3}$$

The mean packet service time is the octet transmission time multiplied by the mean number of octets

$$\tau = T_b\, b, \ \ s \tag{4}$$

The mean arrival rate is

$$\lambda = p/T_b\ , \quad packets/s \tag{5}$$

Therefore, the offered traffic is given by

$$A = \lambda\tau = p\ \ b\ , \ \ Erl \tag{6}$$

The average inter-arrival time of the packets in time slots of the geometric distribution is

$$m_o = 1/p \tag{7}$$

The variance of the inter-arrival time of the packets is

$$d_i = (1-p)/p^2 \tag{8}$$

We define the state probability P_i of being of state *i*, as the probability that there are *i* octets in the system at the end of any time slot. For the system to contain *i* bytes at the end of any time slots it could contain any of *0, 1, 2,....., i+1* at the end of the previous slot. State *i* can be reached from any of the states *0* up to *i* by a precise number of arrivals. To move from *i+1* to *i* there should be no arrivals.

We can write the first equation by considering all the ways in which it is possible to reach the empty state

$$P_0 = \left(P_0 + P_1\right)a_0\ . \tag{9}$$

Similarly, we find a formula for the next state probabilities by writing the balance equations

$$P_i = P_{i+1}a_0\ , \qquad 1 \le i \le m_1 - 1 \tag{10}$$

We continue with this process and take into account that it is possible to enter a

packet in a time slot with length m_1, m_2 or m_3 bytes

$$
\begin{aligned}
P_{m_1} &= (P_0 + P_1)a_{m_1} + P_{m_1+1}a_0 \\
P_{m_1+1} &= P_2 a_{m_1} + P_{m_1+2}a_0 \\
&\quad o \quad o \quad o \\
P_{m_3} &= (P_0 + P_1)a_{m_3} + P_{m_3-m_1+1}a_{m_1} + P_{m_3+1}a_0 \\
P_{m_3+1} &= P_2 a_{m_3} + P_{m_3-m_1+2}a_{m_1} + P_{m_3+2}a_0 \\
&\quad o \quad o \quad o \\
P_{m_2} &= (P_0 + P_1)a_{m_2} + P_{m_2-m_3+1}a_{m_3} + P_{m_2-m_1+1}a_{m_1} + P_{m_2+1}a_0 \\
P_{m_2+1} &= P_2 a_{m_2} + P_{m_2-m_3+2}a_{m_3} + P_{m_2-m_1+2}a_{m_1} + P_{m_2+2}a_0 \\
&\quad o \quad o \quad o \\
P_{k-1} &= P_{k-m_2}a_{m_2} + P_{k-m_3}a_{m_3} + P_{k-m_1}a_{m_1} + P_k a_0 \\
P_k &= P_{k-m_2+1}a_{m_2} + P_{k-m_3+1}a_{m_3} + P_{k-m_1+1}a_{m_1} + P_{k+1}
\end{aligned}
\tag{11}
$$

Then using the fact that all the state probabilities must sum to 1 we write the last equation

$$\sum_{i=0}^{k+1} P_i = 1 \tag{12}$$

We can solve the system (9), (10), (11) and (12) and calculate the state probabilities.

GENERALISED PARETO DISTRIBUTION

The most common choice for telecommunication network design is based on the exponential assumption. Usual choice is the Poisson arrivals and exponential holding times. However, networks and applications of today generate a traffic that is bursty over a wide range of time scales. A number of empirical studies have shown that the network traffic is self-similar or fractal in nature.

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a power law probability distribution that coincides with social, scientific, geophysical, actuarial, and many other types of observable phenomena.

The family of Generalized Pareto Distributions (GPD) has three parameters: the location parameter μ, the scale parameter σ and the shape parameter ξ.

The cumulative distribution function of the GPD is:

$$F(x) = 1 - \left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-1/\xi} \quad . \tag{13}$$

We choose these substitutions

$$\eta_0 = \frac{\xi}{\sigma}; \quad \frac{1}{\xi} = 1 + \frac{\lambda}{\eta_o}; \quad \mu = 0 \quad . \tag{14}$$

Therefore, we receive another form of the generalized-Pareto distribution

$$F(t) = 1 - (1 + \eta_o t)^{-\left(1+\lambda/\eta_o\right)} \quad . \tag{15}$$

The mean value of the generalized-Pareto distribution is:

$$m_{GP} = 1/\lambda \quad . \tag{16}$$

The mean value is the average inter-arrival time for our study. The parameter λ is the packets arrival intensity.

The variance of the generalized Pareto distribution is:

$$d_{GP} = \frac{\lambda + \eta_o}{\lambda^2(\lambda - \eta_o)}, \quad 0 \le \eta_o \le \lambda \tag{17}$$

It follows that the probability density function of the GPD:

$$f(t) = (\eta_o + \lambda)(1 + \eta_o t)^{-\left(2+\lambda/\eta_0\right)} \quad . \tag{18}$$

It is convenient to define the mean value and variance of the arrival stream. We can easy calculate the parameter η_O (the ratio of the shape and scale parameter):

$$\eta_o = \lambda\left(1 - \frac{2}{d_{GP}\lambda^2 + 1}\right) \tag{19}$$

One can easily generate a random sample from Pareto distribution by using inverse distribution function. Given a random variable *U* with uniform distribution on the unit interval (0,1), the random variable *x* is Pareto-distributed.

$$x = \frac{U^{-\frac{\eta_o}{\eta_o+\lambda}} - 1}{\eta_o} \tag{20}$$

To receive the discrete inter-arrival time intervals we accept that packet arrived in the time interval will be served at the end of the time slot, T_b.

$$x_d = \left\lfloor \frac{U^{-\frac{\eta_o}{\eta_o+\lambda}} - 1}{\eta_o} \right\rfloor + 1 \tag{21}$$

SIMULATION MODEL DESCRIPTION

The problem of packet inter-arrival distribution is much more difficult. Understanding of network traffic has evolved significantly over the years, leading to a series of evolutions in network traffic models.

Simulations are the main tools for studying the performance of telecommunication networks. Our simulation model is used to study the P/MM/1/k queue. It could be used to model arrivals of packets at a packet switched network.

Let us consider a single server queue P/MM/1/k with a Pareto input stream, which is defined by packet arrival intensity λ, variance of the inter-arrival time d_{iGP}, three-modal service time with mean number of bytes in the packet b and limited waiting room k. This queueing system with peak input stream and three-modal service time is a non-Markovian model (Fig. 2). It is assumed that customers are served in FCFS order.

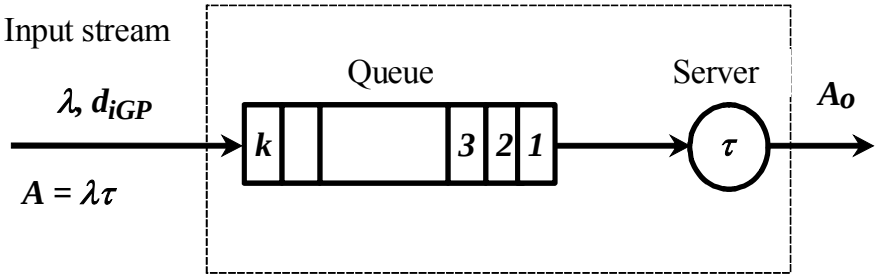

Fig. 2. Pareto input stream model with multimodal service time and finite waiting positions.

The arrival process in P/MM/1/k queue is considered to be renewal process. The queues with a heavy-tailed distribution of inter-arrival time are used to model queue systems where a range of values of the inter-arrival time, whose probability is very low, have a drastic impact on the overall performance of the system. The Pareto distribution is one of these heavy-tailed distributions and it is proposed to describe peak streams in the packet switched networks. The accurate analytical treatment of

P/MM/1/k systems is very difficult and in many cases it cannot be applied. Simulation is a possible method to study. Simulations with heavy-tailed random variables present some additional difficulties. A care must be taken during analyses of the results of these simulations. It is necessary to have accurate and efficient simulation methods. The efficacy is important because we need to generate big quantities of data for our simulation study and be accurate enough. The data accuracy can be estimated by means of comparisons with known results from simpler systems with analytical solution. One of these simpler queue systems that are studied analytically is the Geo/MM/1/k queue. This queue is used as a workbench for more efficient simulation methods that are able to deal with the heavy-tail difficulties.

We develop a real time trace simulation algorithm for evaluating the state probabilities of the queueing system, the packet congestion probability and the mean time in the queue. We use batch mean method for output results analysis and choose a confidence probability 95%. We define 20 batches and generate 20000 packets in every batch. We introduce an initial bias to eliminate the influence of the transient behaviour and time intervals between batches to received independent estimates of the packet congestion probability and the mean queueing time. We describe the accuracy of the estimates by means of a confidence interval, which with a given probability (95%) specifies how the estimate is placed relatively to the unknown theoretical value, using the Student's t-distribution with 19 degrees of freedom. This organization of our algorithm leads to good accuracy from a practical point of view. The relative errors of the presented results are less than 10%.

Random errors are caused by the stochastic variations of the simulation. They appear because every simulation is similar to a statistical experiment. The next source of error is the bias of the estimator itself, being often called the systematic error. This kind of error usually appears if assumptions about the analyzed data are true only approximately or asymptotically. If both the variance and the bias tend to zero for large number of observations the estimator is called consistent.

PERFORMANCE MEASURES

The carried traffic is equivalent to the probability that the system is busy

$$A_o = 1 - P_0 , \ \ erl \tag{22}$$

The packet congestion probability B is defined and evaluated by the simulation program as the ratio of lost and arrival packets. It can be calculated by offered and carried traffic (offered minus carried traffic) to offered traffic

$$B = \left(A - A_o\right) / A \tag{23}$$

The mean number of bytes and packets present in the system in steady state by definition is

$$L_b = \sum_{j=1}^{k+1} j\,P_j,\ bytes; \qquad L_p = L_b/b,\ packets \tag{24}$$

From the Little formula, we have the normalized mean queueing time of the bytes (time is measured in time slots)

$$\frac{W_{bq}}{T_b} = \frac{L_b}{T_b \lambda\, b} - b = \frac{L_b}{A} - b \tag{25}$$

The real time simulation gives us possibility to calculate the queueing time for every arrival packets and it is easy to obtain the mean queueing time.

ANALYTICAL AND SIMULATION RESULTS

In this section, we give numerical results obtained. The described models are tested on a computer over a wide range of arguments.

Figure 3 shows the stationary probability distribution in a single server queue Geo/MM/1/k with 0.7 erl offered traffic, 199 waiting positions in bytes, 5 bytes minimum packet length with probability 0.3, 10 bytes packet length as a second mode, 20, 40 or 80 bytes maximum packet length and 15 mean number of bytes in the packet. We can see that the probability distributions are almost linear decreasing in logarithmic scale and the influence of the maximum packet length on the stationary probability is significantly.

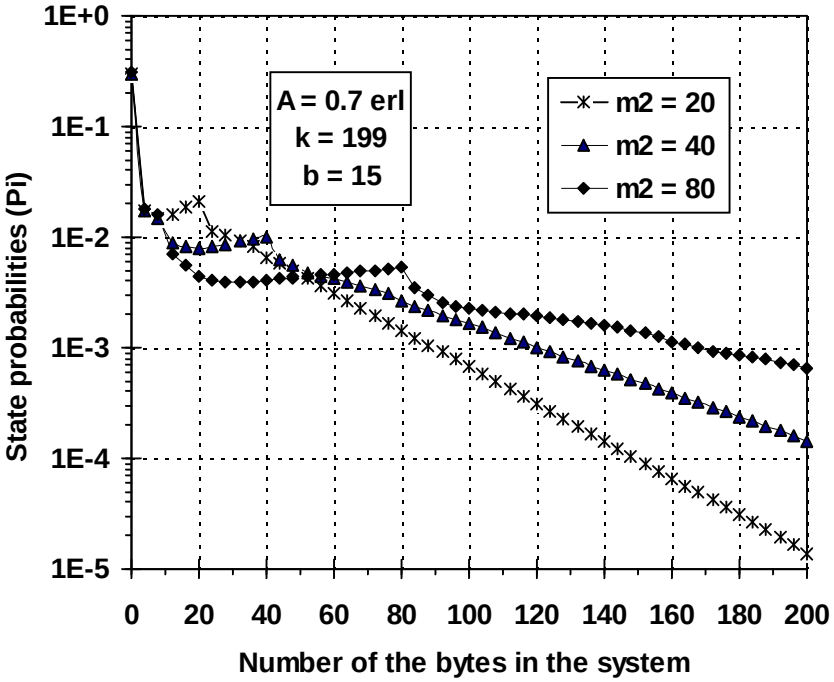

Fig. 3. The stationary probability distribution in a single server queue Geo/MM/1/k with different maximum packet size.

Figure 4 illustrates the dependence on the packet congestion probability from

the offered traffic and the same parameters of the multimodal distributed packet length. We can see that the influence of the maximum packet length on the packet congestion probability is immense.

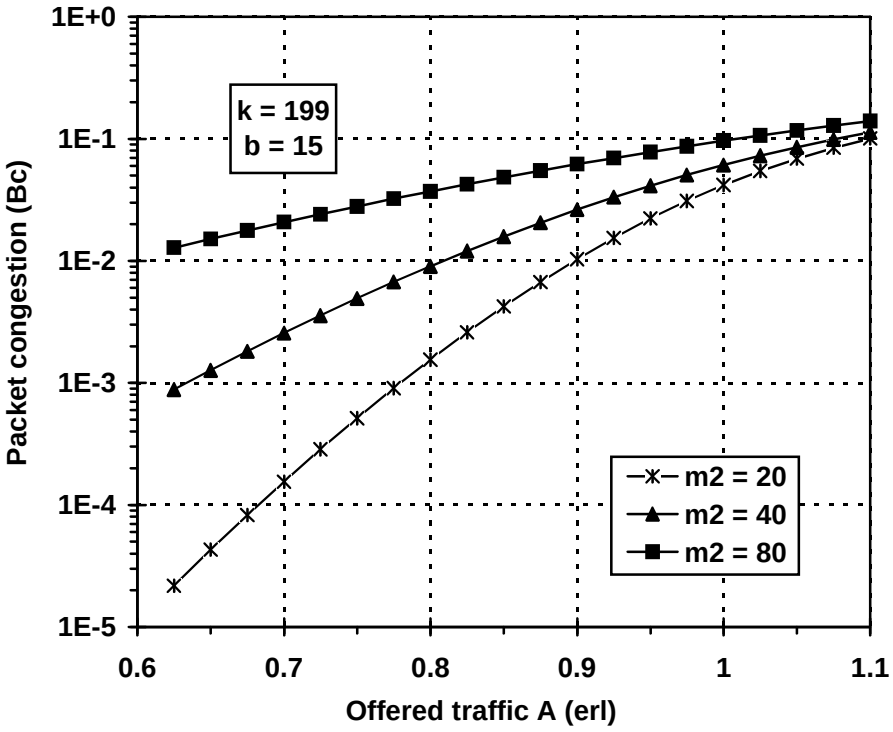

Fig. 4. Packet congestion probability in a single delay system with geometrically distributed inter-arrival and multimodal service time.

Figure 5 presents the normalized mean queueing time of the bytes (W/T_b) as function of the traffic intensity when the queue length and the packet length distribution is the same as in the Figure 3. The influence of the maximum packet length on the mean queueing time is significant when the offered traffic is smaller then 0.9 erl.

Figure 6 illustrates the stationary probability distribution in a single server queue P/MM/1/k with a Pareto input stream, 0.8 erl offered traffic, 199 waiting positions and different variance of the inter-arrival time. It is seen that when the variance increases the probability that the queue is full increases significantly.

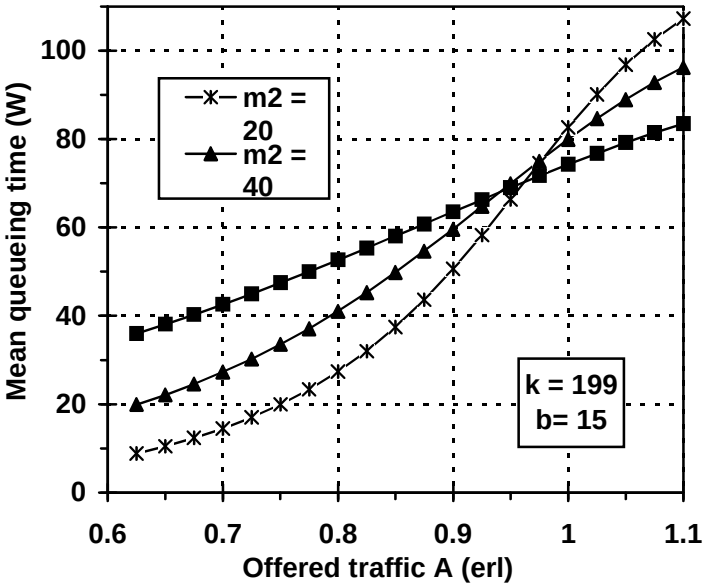

Fig. 5. Mean queueing time in time slots for a single delay system with geometrically distributed inter-arrival time and multimodal service time.

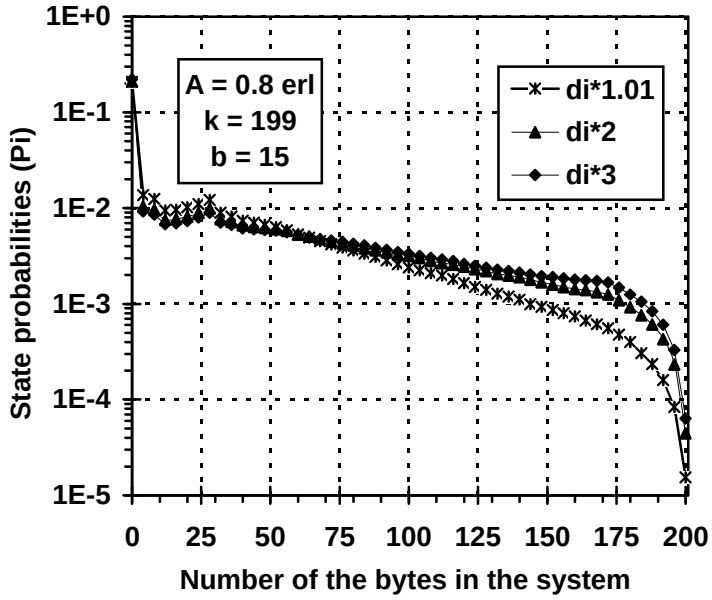

Fig. 6. Steady-state distribution of the number of bytes in a single server queue P/MM/1/k with different peakedness of the input stream.

Figure 7 presents the packet congestion probability in a single delay system with 199 waiting positions, different offered traffic and different variance of the inter-arrival time. When the offered traffic is comparatively small (0.8 erl) the influence of the variance of the call congestion probability is big.

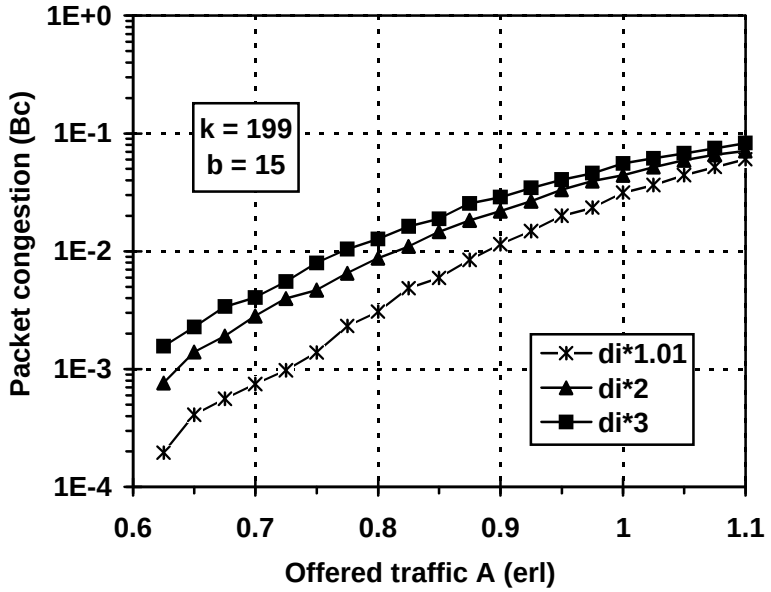

Fig. 7. Packet congestion probability in a single delay system with a Pareto input stream and multimodal service time.

Figure 8 shows the normalized mean queueing time of the packets as function of the offered traffic when the number of queueing positions is 199, the service time is a three-modal distributed and different variance of the inter-arrival time.

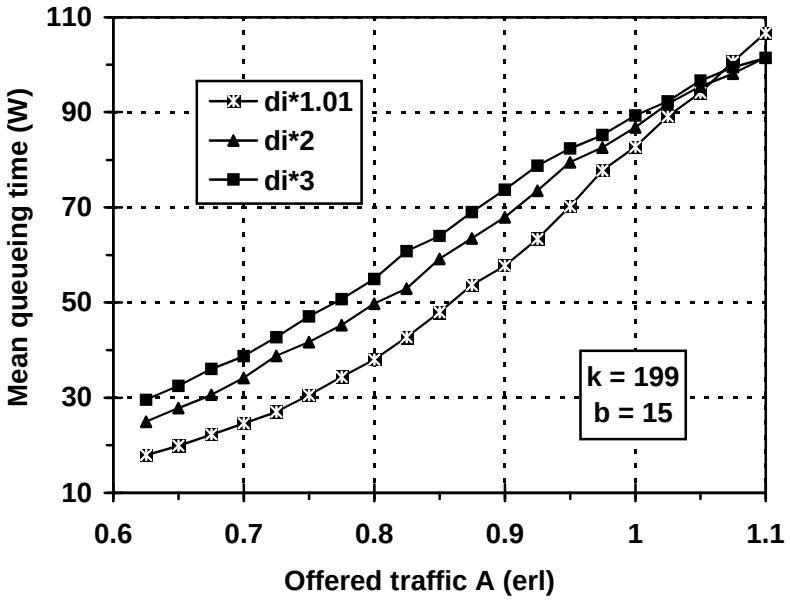

Fig. 8. Mean queueing time in time slots for a single delay system with a Pareto input stream and multimodal service time.

It is shown that the influence of the variance of the input stream over the

performance measures is significant. The heavy-tailed condition decisively contributes to raise the congestion and waiting time. The results are highly dependable from the number of packets in the queue. In this simulation we have relatively small number of packets of small size in the queue.

The computer simulation of P/MM/1/k queues presents important difficulties due to the slow decaying tail of the Pareto distribution. This makes extremely high values, with great influence on the statistical figures of the system. The probabilities are so low that in case we want to simulate the physical underlying processes, generating demanded times and arrival intervals, the cost in time will probably be prohibitive if we want accurate results. This forces to use all our knowledge of the statistics of the system inner processes, so the simulation can noticeably speed up. The results shown are applicable to UDP traffic that is carried usually with higher priority in the buffers and has small packets. The same derivates can be applied also to TCP traffic by simple scale multiplication of the queue length and packet size. In this case we will consider bigger queues with bigger packets but with approximate equal number of packets in the queue. TCP traffic is scheduled usually in different queues with less priority in comparison to the UDP traffic.

CONCLUSION

The importance of a discrete time single server queue in a case of a geometric and Pareto input stream and multimodal service time comes from its ability to describe behaviour that is to be found in more complex real queueing systems. It is one of the cases in a general teletraffic system that is important in telecommunication systems design.

In this paper, a basic discrete-time single server teletraffic system Geo/MM/1/k and P/MM/1/k are examined in detail.

We have presented an analytical model for evaluating the Geo/MM/1/k queue and a simulation method for studying the P/MM/1/k queueing systems. We have demonstrated its use by presenting numerical results. These results have shown that the Pareto distribution change significantly the queue behaviour. In the case of P/MM/1/k, congestion occurred even when the load is sufficiently small. But for that queue, the long-tailed nature of the Pareto helps to clear out congestion when a large inter-arrival time occurred. Our model can be applied for all P/G/1/k systems independently of the value of the parameters.

The proposed approach provides a unified framework to model discrete-time single server queue and peak input traffic. A generalised Pareto distribution is introduced and explained. Numerical results and subsequent experience have shown that this model is accurate and useful in analyses of teletraffic systems.

Our model permits us to look at the queueing behaviour. We show that as the load increases, the long-tailed nature of the queue brings to big losses and delays.

Comparisons with Poisson arrivals showed that the simple Markovian models seriously underestimate the performance of such systems. In a sense, our results help solidify those statements being made by other authors.

The simulation method we have presented could certainly be used to study congestion in the Next Generation Networks. Our method generates a complete probabilistic analysis of the queues we study. The method is quick and its accuracy can be easy evaluated. We have used the method with the Pareto only, but are investigating its use with other distributions and many different packet sizes.

We feel that our simulation method has excellent promise to analyze the type of congestion problems and delays seen on the Internet. Thus, we are continuing our research using the simulation method for a larger class of queueing systems. Our further study also concerns the behaviour of the common output interface of the routers where traffic of different priorities is mixed. In this new approach we combine TCP and UDP traffic and estimate per service, per transport protocol, per queue, per interface Quality of Service and performance parameters.

The results presented here add a new aspect to the evaluation of the discrete-time queueing system, and serve as a basis for future research on guaranteeing the quality of service. We consider this study important for Next Generation Networks where different services are mixed together and where the Quality of Service requirements should be covered per service and per interface.

REFERENCES

[1] L. P. Khadjiivanov, B. T. Taskov, A. A. Aliazidi, B. P. Tsankov, "Application of priority queueing mechanisms to ATM multiplexing and traffic control," Proc. of Integrated Broadband Communications Networks and Services, Copenhagen, Denmark, April 20 – 23, 1993, pp. 33.3.1 – 33.3.11.

[2] B. Tsankov, R. Pachamanov, D. Pachamanova, "Modified brady voice traffic model for WLAN and WMAN," Electronics Letters, vol.43, issue 23, Nov. 2007, pp. 1295-1297.

[3] J. Cao, W. Cleveland, D. Sun, "Bandwidth estimation for best-effort Internet traffic," Statist. Sci., Volume 19, Number 3 (2004), pp. 518-543.

[4] P. Salvador, A. Pacheco, R. Valadas, "Modelling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs," Computer Networks, Volume 44, Issue 3 , 2004, pp. 335-352.

[5] Y. Koh, Kim Kiseon, "Loss probability behaviour of Pareto/M/1/K queue," Communications Letters, IEEE, Volume 7, Issue 1, pp. 39-41, Jan 2003.

[6] Y. Koh, Kim Kiseon, "Evaluation of steady-state probability of Pareto/M/1/K experiencing tail-raising effect," Lecture Notes in Computer Science, Volume 2720, pp. 561-570, 2003.

[7] A. Inmaculada, M. Meerschaert, A. Panorska, "Parameter estimation for the truncated Pareto distribution," Journal of the American Statistical Association, Volume 101, Number 473, pp. 270-277, 2006.

[8] Ron Addie, “Snapshot simulation of internet traffic: fast and accurate for heavy-tailed flows,” In QoSim 2008: 1st International Workshop on the Evaluation of Quality of Service through Simulation in the Future Internet, 2008, Marseille, France.
[9] M. Ismail, A. Zin, “Measurement and characterization of network traffic utilization between real network and simulation modeling in heterogeneous environment,” IJCSNS International Journal of Computer Science and Network Security, Vol. 8 No. 3, 2008, pp. 326-337.
[10] L. P. Argibay, A. Suárez González, C. López García, R. Rodríguez Rubio, J. López Ardao, D. Teijeiro Ruiz, “On the simulation of queues with Pareto service,” Proc. 17th European Simulation Multiconference, pp. 442-447, 2003.
[11] M. Eickhoff, D. McNickle, K. Pawlikowski, “Analysis of the time evolution of quantiles in simulation,” International Journal of Simulation, Vol. 7, No 6, pp. 44-55, 2006.
[12] S. Fernandes, C. Kamienski, D. Sadok, ”Accuracy and computational efficiency on the fractal traffic generation,” Proc. of the 3rd IASTED International Multi-Conference on Wireless and Optical Communications-WOC, 2003.
[13] D. Gross, M. Fischer, D. Masi, J. Shorte, D. Gross, “Difficulties in simulating queues with Pareto service,” Proceedings of the Winter Simulation Conference, pp. 407-415, 2003.
[14] M. Fischer, H. Cart, “A method for analyzing congestion in Pareto and related queues,” Telecommunications Review, pp. 15-27, 1999.
[15] M. Fischer, D. Bevilacqua Masi, D. Gross, J. Shorte, “One-parameter Pareto, two-parameter Pareto, three-parameter Pareto: Is there a modeling difference?” Telecommunications Review, pp. 79-92, 2005.
[16] S. Luo, G. Marin, “Realistic Internet traffic simulation through mixture modeling and a Case Study,” Proceedings of the Winter Simulation Conference, pp. 2408 – 2416, 2005.
[17] J. Pitts, J. Schormans, “Introduction to IP and ATM design and performance,” John Wiley&Sons, Ltd., 2000.
[18] S. Mitrtchev, “Study of queueing behaviour in IP buffers,” Information Technologies and Knowledge magazine (IJ ITK) Vol.2., 2008, pp. 187-192.
[19] N. Vicari, P. Tran-Gia, “A numerical analysis of the Geo/D/N queueing system,” Technical Report No. 151, September 1996.
[20] Bo Chen, Xiaotie Deng, Wenan Zang, “On-line scheduling a batch processing system to minimize total weighted job completion time,” ISAAC 2001, pp. 380-389
[21] I. Atencia, P. Moreno, “A single-server G-queue in discrete-time with geometrical arrival and service process”, Performance Evaluation 59, 2005, pp. 85-97.
[22] I. Atencia, P. Moreno, “Geo/G/1 retrial queue with 2nd optional service,” International Journal of Operational Research 1, 2006, pp. 340-362.
[23] Dan Wang, Funda Ergün, Zhan Xu, “Unicast and multicast QoS routing with multiple constraints,” QoS-IP 2005, pp. 481-494.

Модел на едноканална телетрафична система с дискретно време и мултимодално разпределение на размера на пакетите

Сеферин Мирчев, Росица Голева
Катедра „Комуникационни мрежи”,
Технически университет - София
stm@tu-sofia.bg, rig@tu-sofia.bg

РЕЗЮМЕ

Важността на IP мрежата ще нараства в бъдеще и ще бъде платформа за все повече и повече услуги, изискващи различни видове и качество на обслужване. В тази статия изследваме поведението на опаката на входящ IP комутатор или маршрутизатор. За да анализираме това поведение, изучаваме дискретен вариант на класическия модел M/M/1/k, наречен Geo/MM/1/k и P/MM/1/k. Ние използваме геометрично (Geo) и Парето (P) разпределени интервали между моментите на постъпване и дискретно мултимодално (MM) разпределена продължителност на пакетите, дефинирана в интервал от минимална и максимална стойност. В този модел единицана за измерване на времето се намалява от времето за обслужване на пакет до времето за предаване на един байт. Фактът, че геометричното и на Парето разпределенията са подходящи за моделиране на интервалите между моментите на постъпване на пакетите в Интернет, ни мотивира да предложим метод за оценка на характеристиките, свързани с вероятностите на състоянията на системата Geo/MM/1/k. Ние предлагаме симулационен модел в реално време за оценка на вероятностите на състоянията, показващ закъсненията и вероятността за загуби на системата P/MM/1/k. Този модел е подходящ за изучаване на системите с дълговременна зависимост.