Информационные процессы, Том 5, № 1, 2005, стр. 40–46.

Asymptotic Delay Distribution and Burst Size Impact on a Network Node Driven by Self-similar Traffic

C.D'Apice, R.Manzo

Department of Information Engineering and Applied Mathematics,
University of Salerno, Italy

Received December 1, 2004

Abstract—It was shown recently that under self-similar traffic the delay distribution function can decrease very slowly, so in order to guaranty the Quality of Service (QoS) in communication networks, burst size is usually bounded by some value using, for example, leaky-bucket mechanism.
In this paper we consider a discrete-time queue with M types of independent input processes. Each input process is the aggregation of sessions (bursts) arrived by a Poisson process. Asymptotic delay distribution at network node driven by self-similar traffic and its effects on burst size bound have been analysed. It is also found the critical value of the burst size at which delays start to increase considerably.

1. INTRODUCTION

Recent network traffic studies have shown that network arrival processes are more adequately modelled using self-similar processes. Self-similar models, in fact, are able to capture the burstiness and the long-range dependence characteristics, which means that significant correlations are present in arbitrarily large time scales. The presence of long-range dependence in traffic processes has a strong impact on queueing performance and buffer engineering, for example it completely alters the tail of queue waiting times.

Traffic network quality of service can be estimated evaluating packets delay distribution and packets loss probability. Typically packet delay or loss arise at network node, so the traffic performance analysis of a network can be reduced to the investigation of a node performance. A network node can be represented as a $G/D/1$ queueing system so a crucial matter is to find an adequate distribution for the input process G of this queueing model.

Two basic approaches have been developed in the investigation of packet loss phenomenon and overflow probability, i.e. the probability of exceeding a given level in a finite buffer queue: the so-called many sources asymptotics and the large buffer asymptotics. The many sources asymptotics, investigated by many authors ([1], [4], [8], [15]) is valid when a large number of sources access a buffer which is drained by a very high capacity server. The large buffer asymptotics approach, instead, is used when there are a few sources or some sources which utilize a significant amount of server capacity. It has been studied under various hypotheses on the source characteristics, see [5] for the case of light-tailed sources.

Recently empirical studies by Willinger *et al* ([7], [14]) have shown the presence of sources with heavy-tailed characteristics. It means that sources transmit for long periods of time when they come on and the activity periods have heavy-tailed characteristics in time, i.e. the complementary distribution of the activity period has a decay which obeys a power law rather than an exponential one.

The large buffer asymptotics for the case of heavy-tailed source activity periods has been studied under various assumptions on the input streams. These range from queues with fractional Brownian motion inputs ([12]), general Gaussian processes with negative drifts ([3]), ON-OFF inputs with long-tailed ON periods ([6]) and $M/G/\infty$ type of inputs with long-tailed G distributions ([10], [11], [13]). An excellent account can be found in the survey [1].

In this paper we will focus on the model proposed and studied in [9]. In this model the input process for a network node is the sum of sessions (bursts) arrived by a Poisson process but with length distributed according to a power law. There are many type of sources which can differ by the length distribution exponent, the arrival intensity, the rate in active period. Since the sum of different types of sources has been considered, the model is sufficiently general and give us the possibility to analyze how sources parameters can change server performance characteristics. In the paper [9] the influence of the rate of active source period and the length distribution exponent on system performance has been investigated. In this paper we will find how the value of the maximum length of active period (burst length) can change performance characteristics of the server.

The organization of the paper is as follows: in Section 2 we give notations, formulation of the model, and asymptotic definition. Section 3 presents the main result and its proof in brief form. In Section 4 we consider the simple homogeneous case and we give interpretation of the developed results.

2. TRAFFIC MATHEMATICAL MODEL

We consider a discrete-time queue with M types of independent input processes $Y_{t,i}$. Each input process $Y_{t,i}$ is the aggregation of sessions (bursts) arrived by a Poisson process with intensity λ_i. Sessions arrive independently of each other and a session of type i transmits at the rate $r_i > 0$ for a duration of time τ_i, called the session length or duration. We assume that the session lengths have a long-tailed distribution, i.e. for large values of x, the complementary distribution of τ_i is given by

$$\Pr\{\tau_i > x\} \sim \begin{cases} \alpha_i x^{-1-\beta_i}, \text{ if } x < B, \\ 0, \text{ if } x \geq B, \end{cases}$$

where α_i,$\beta_i > 0$ are some constants and the maximum burst size B will be defined later. If $\beta_i \in (0, 1]$ (the sessions are referred to as heavy-tailed) and $B = \infty$ the input process $Y_{t,i}$ will be asymptotically self-similar.

Throughout the paper the notation $A(x) \sim B(x)$ means that

$$\lim_{x\to\infty} \frac{A(x)}{B(x)} = 1.$$

It is assumed that the buffer is drained at a rate of C units per time. The sessions and their lengths are assumed to be mutually independent.

Let $\theta_{t,i}$ denote the number of sessions of type i which arrive into the system at time t. We assume that the r.v's $\theta_{t,i}$ are i.i.d. We have

$$\Pr\{\theta_{t,i} = n\} = \frac{{\lambda_i}^n}{n!} e^{-\lambda_i}, \ \ \lambda_i > 0.$$

The input process of the queue will be equal to

$$Y_t = \sum_{i=1}^{M} Y_{t,i}$$

and each component $Y_{t,i}$ can be expressed as

$$Y_{t,i} = \sum_{n=t}^{-\infty} \sum_{j=1}^{\theta_{n,i}} r_i I(\tau_{n,i,j} \geq t - n)$$

where $I(A)$ denotes the indicator function of the event A.

We denote the average system input load by

$$\rho = \mathbf{E}[Y_t] = \sum_{i=1}^{M} \lambda_i r_i \mathbf{E}[\tau_{t,i,1}]$$

and we will assume that $\rho < C$.

Denote by W_t the stationary buffer occupancy (the workload) assuming an infinite buffer. W_t is given by the following formula

$$W_t = \max(W_{t-1} + Y_t - C, 0).$$

In this article we are interested in studying the behavior of the tail probability $\Pr\{W_t > z\}$ for large z i.e. as $z \to \infty$, with maximum burst size $B \sim bz$, where $b > 0$ is some constant. More precisely our aim is to find the probability for the system to reach overload period. which can be defined as a period when packets delay is greater than z/C. Taking into account that for large z this probability is sufficiently small, it can be defined as

$$F_{ov} = \Pr\{W_t > z, \quad W_i \le z, \quad t - \delta z < i < t\}$$

where $\delta > 0$ is some sufficiently small constant.

To state the main results we need to define the following random variables:

- J denotes a set $(j_1, j_2, \cdots, j_M)$ of M non negative integers.
- $\kappa_J = \sum_{i=1}^{M} \beta_i j_i$ corresponds to the decay exponent associated to the set J.
- $R_J = \sum_{i=1}^{M} r_i j_i$ is the rate corresponding to the set J of sessions.
- $J_0 = \arg\min_J \{\kappa_J \ : R_J - (C - \rho) > \frac{1}{b}\}$.

3. EVALUATION OF OVERLOAD PROBABILITY BEHAVIOUR

Now we are ready to formulate the main result.

Theorem 1. *For large z, the system overload probability can be found as*

$$F_{ov} \sim z^{-\kappa_{J_0}+1} \prod_{i=1}^{M} P_i^{j_i^{(0)}} / j_i^{(0)} \tag{1}$$

where

$$P_i = \frac{\lambda_i \alpha_i}{\beta_i} \left((R_{J_0} - (C - \rho))^{-\beta_i} - b^{-\beta_i} \right).$$

In order to prove this result, first, for $b > \varepsilon > 0$ we define the processes Y_t^l and Y_t^h as follows. The process Y_t^l corresponds to the number of active sessions which have session lengths at most εz. This is given by:

$$Y_t^l = \sum_{n=t}^{t-\varepsilon z} \sum_{i=1}^{M} \sum_{j=1}^{\theta_{n,i}} r_i I(\varepsilon z \ge \tau_{n,i,j} \ge t - n). \tag{2}$$

The superscript l denotes that these sessions are "light-tailed".

The process Y_t^hcorresponds to the number of active sessions which have session lengths greater than εz. This is given by:

$$Y_t^h = \sum_{n=t}^{-\infty} \sum_{i=1}^{M} \sum_{j=1}^{\theta_{n,i}} r_i I(\tau_{n,i,j} \ge t - n, \tau_{n,i,j} > \varepsilon z). \tag{3}$$

The superscript h indicates that the sessions are "long-tailed".

We can see that the processes Y_t^l and Y_t^h are mutually independent and $Y_t = Y_t^l + Y_t^h$. Since we put $\varepsilon < b$, the process Y_t^l will have the same properties as in the unbounded case when $B = \infty$.

We start our analysis separately for the processes Y_t^l and Y_t^h and then we combine them to get the final result. For the process Y_t^l we can use directly two lemmas from the paper [9]. These Lemmas can be formulated in the following form. Let

$$X_k^l = \sum_{n=t-k}^{t} Y_n^l.$$

Lemma 1. *For any given* $\delta_2 > 0$, $c_1 > 0$, $0 < \varepsilon < \min\left\{\frac{\delta_2}{\rho}, \frac{\beta_{\min}}{c_1}, b\right\}$ *and sufficiently large* z

$$\Pr\{\sup_{k\geq 0}\{X_k^1 - \rho k\} > \delta_2 z\} \leq z^{1-c_1\tilde{\delta}_2}$$

where $\tilde{\delta}_2 = \delta_2 - (\rho + \delta_1)\varepsilon$, $\rho = \mathbf{E}[Y_t]$.

Lemma 2. *For any given* $\delta_1 > 0$, $\delta_2 > 0$, *and sufficiently small* $b > \varepsilon > 0$, *as* $z \to \infty$

$$\Pr\{\inf_{k\geq 0}\{X_k^1 - (\rho - \delta_1)k\} < -\delta_2 z\} \leq \ e^{-O(z)}.$$

First we need to find asymptotics for the probability $\Pr\{W_t > z\}$.

To prove upper bound, we consider the process Y_t^l as the input to a queue with service rate ρ and the process Y_t^h as the input to another queue with service rate $C - \rho$. Then both queues are stable and let W_t^l denote the stationary workload for the first queue and W_t^h denote the stationary workload for the second queue. If we define $X(-t,k) = \sum_{j=-t}^{k} Y_j$, and $X^l(-t,k) = \sum_{j=-t}^{k} Y_j^l$, we get

$$W_t = \sup_{k\geq 0}\{X(t-k,t) - Ck\},$$

$$W_t^l = \sup_{k\geq 0}\{X^l(t-k,t) - \rho k\},$$

$$W_t^h = \sup_{k\geq 0}\{X^h(t-k,t) - (C-\rho)t\}.$$

It is easy to see that

$$\Pr\{W_t > z\} \leq \Pr\{W_0^l > \delta z\} + \Pr\{W_t^h > (1-\delta)z\}.$$

Now from Lemma 2 we get

$$\Pr\{W_t > z\} \leq \Pr\{W_t^h > (1-\delta)z\}(1 + o(z)).$$

In this way, the upper bound for the system W_t^h is established.

Now, to get lower bound, with similar notation as above, we have

$$\sup_{k\geq 0}\{X(t-k,t) - Ck\} \geq \sup_{k\geq 0}\{X^h(t-k,t) - (C-\rho)k\} + \inf_{k\geq 0}\{X^l(t-k,t) - \rho k\}$$

and again, as $z \to \infty$ and using Lemma 1 , we get

$$\Pr\{W_t^h > z\}(1 + o(z)) \leq \Pr\{W_t > z\}$$

establishing that the probability $\Pr\{W_t > z\}$ is determined by the system W_t^h.

To analyze W_t^h we will also use lower and upper bounds arguments. To compute lower bound we just need to find any configuration of sessions which arise system overflow: $W_t^h > z$. To construct upper bound we need to find what kind of busy periods will dominate. Let us start with lower bound. Define by $A_{J,t}$ the event that at time t there are j_i active sessions of type i. Using Poisson sessions arrivals, we have

$$\Pr\{A_{J,t}\} \sim const.\ z^{-\kappa_J}$$

or more exactly

$$\Pr\{A_{J,t}\} \sim z^{-\kappa_J} \prod_{i=1}^{M} \frac{(P_{J,i})^{l_i}}{j_i!}$$

where

$$P_{J,i} = \frac{\lambda_i \alpha_i}{\beta_i} \left((R_J - (C-\rho))^{-\beta_i} - b^{-\beta_i} \right).$$

Since we compute $\Pr\{A_{J,t}\}$, the probability of the busy period with a given configuration of J_0 active sessions which are simultaneously active during the time interval at least equal to

$$l_0 = (R_{J_0} - (C-\rho))\, z$$

can be computed.

To get upper bound, first we find what the typical busy period is. Recall two definitions:

Definition 1. We define a congestion period in a queue as the duration within a busy period when the net rate is positive, i.e., the total input rate exceeds the server rate. Conversely we define a post-congestion period as the duration within a busy period when the net rate is negative, i.e., total input rate is less than the server rate.

Definition 2. We define an Isolated Typical Busy Period (ITBP) as a busy period during which there is one congestion period, one post-congestion period and that the congestion period is caused by exactly J_0 $h-$type sources arriving to begin the busy period.

Following the proof of the Lemma 2 from [9] we can find that typical busy period at which overflow occur will be isolated and will not contain any sessions except the sessions from configuration J_0 and

$$\Pr\{W_t^h > z,\ t \in ITBP\} = O(z^{-\kappa_{J_0}}),$$

$$\Pr\{W_t^h > z,\ t \notin ITBP\} = o(z^{-\kappa_{J_0}}).$$

Combing these results with lower bound and taking into account above arguments concerning the dominating of W_t^h system in overflow probability, we get equation (1).

4. HOMOGENEOUS CASE

In the simple case, when we have only one class of sources $M = 1$, easy calculations can be made to understand how sessions finite length can effect overflow probability. In the homogeneous case we have $r_i = r$, $\beta_i = \beta$, $J_0 = \{j_0\}$, $R_{J_0} = rj_0$, $\kappa_{J_0} = \beta j_0$. Thus

$$j_0 = \lfloor \frac{(C-\rho) + \frac{1}{b}}{r} \rfloor + 1$$

and

$$F_{ov} = const.\ z^{-\beta j_0}.$$

As we can see, in sense of overflow probability, the critical value of burst size bz is about $\frac{z}{C-\rho}$: critical delay value divided by $C-\rho$. It means that increasing burst size more than the value bz, critical delay probability or overflow probability does not change significantly. In this case only the session average rate r takes an important role. Meanwhile, bounding the burst size to a value less than bz we can significantly decrease overflow probability or probability to exceed critical delay. Roughly speaking, the absolute value of the delay distribution exponent will increase two times if we decrease the maximum burst size two times.

5. CONCLUSIONS

In this paper large buffer asymptotics has been considered in order to investigate the influence of the traffic burst size limit on the overload probability in a data network node. Limit on burst size means that the burst length distribution function is cut at some point to zero with appropriate normalization. The results have been obtained for the so called Poisson/Pareto traffic model, where sessions (bursts) arrive according to a Poisson process and session lengths are distributed by a power law, which implies that the probability of long length sessions has a significant value.

Let us discuss briefly the derived results. In the situation of large buffer asymptotics, as it was shown analytically, typical overload scenario arises due to a certain number of bursts with large length which are active in the same time period. It means that typical overload is not due to the fluctuation of sources or bursts number in the system but to bursts length. As a result, overload probability decreases according to the burst length distribution (by power law), while in the case of overload due to the sources number fluctuations, overload probability decreases exponentially. If we introduce a limit on the maximum burst length, overload probability behaviour depends on the maximum accepted delay value. If this limit is bigger than the maximum accepted delay value, overload probability asymptotically does not change since typical scenario is still the same. Otherwise, when the limit value is less than the maximum delay value, extra sources arrivals with larger bursts length are needed to reach overload in the system and in the case of homogeneous sources overload probability decreases exponentially over the limit on burst size. In this way, maximum burst size starts to play an important role for overload probability evaluation.

Within the analyzed model, another traffic parameter which plays a critical role is the source peak rate. Sources peak rate means the rate inside the burst generated by this source. As it was shown previously and in this paper, overload probability decreases exponentially decreasing the peak rate. In some sense, in the overload probability evaluation, it is not so important the peak rate or the burst length but the product of the burst length by the peak rate. Of course, this is true only in the case in which the burst length is less than the maximum accepted delay value. In a practical sense it means that we can increase the sources peak rate keeping the burst size (expressed in bits) constant and the overload probability on the node will not increase if the bursts length (in units of time) is less than the critical value.

We should keep in mind that the presented results are asymptotic. We considered asympotic while buffer size and burst length go to infinity (the so called large buffer asymptotics). It means that in the case of not limited values, the derived results can be considered only as some approximations which can be good or not depending on how the particular system is close to the limited one. For example, overload scenario in a real system can arise not only due to the long lengths sessions like in asymptotic case, but also to the active sources number fluctuations.

Next, we should keep in mind, that different models for network traffic can be constructed. For example, we can consider asymptotics when the sources number in the system goes to infinity. In this case the overload probability behaviour will be quite different. Overload probability will decrease exponentially over the buffer size and typical overload scenario is due to the sources number fluctuation. Practically, overload will arise when the average load will exceed the system service rate. The validity of one or another model depends strongly on the statistical properties of the particular network traffic.

Acknowledgements

We are grateful to Dr. N. Likhanov for useful discussions and substantive suggestions concerning the present manuscript.

REFERENCES

1. A. Botvich, A. and N. G. Duffield, Large deviations, economies of scale and the shape of the loss curve in large multiplexers, *Queueing Systems*, 1995, no. 20, pp. 293–320

1. O. J. Boxma and V. Dumas, Fluid queues with long-tailed activity period distributions, *Computer Communications,* 1998, no. 21, pp. 1509–1529.

3. J. Choe and N.B. Shroff, On the supremum distribution of integrated stationary Gaussian processes with negative linear drift, *Advances in Applied Probability*, 1999, no. 31, pp. 135–157.

4. C. Courcoubetis and R. Weber, Buffer overflow asymptotics for a switch handling many traffic sources, *J. Appl. Prob.*, 1996, vol. 33, no. 3, pp. 886–903

5. N. G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue with applications, *Math. Proc. Camb. Phil. Soc.*, 1995, no.118(1), pp. 363–374.

6. P. R. Jelenkovic and A. A. Lazar, asymptotics results for multiplexing subexponential on-off sources, *Advances in Applied Probability*, 1999, no. 31, pp. 394–421.

7. W. E. Leland, M.S. Taqqu, W. Willinger and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. on Networking*, 1994, vol. 2, no. 1, pp. 1–15.

8. N. Likhanov and R. Mazumdar, Cell loss asymptotics in buffers fed with a large number of independent stationary sources, *Journal of Applied Probability*, 1999, no. 36, pp. 86–96.

9. N. Likhanov and R. Mazumdar, Loss asymptotics in large buffers fed by heterogeneous long-tailed sources, *Advances in Applied Probabili*ty, 2000, no. 32, pp. 1168–1189.

10. N. Likhanov, Bounds on the buffer occupancy probability with self-similar input traffic, in *Self-similar network traffic and performance evaluation*, K.Park and W.Willinger eds., Wiley, 2000, pp. 193–214.

11. Z. Liu, P. Nain, D. Towsley and Z-L. Zhang, asymptotics behavior of a multiplexer fed by long-range dependent process, *Journal of Applied Probability*, 1999, no. 36, pp. 105–118.

12. I. Norros, A storage model with self-similar input, *Queueing Systems*, 1994, no. 16, pp. 387–396.

13. M. Parulekar and A. M. Makowski, Tail probabilities for $M/G/\infty$ input processes, *Queueing Systems*, no. 27, pp. 271–296.

14. V. Paxon and S. Floyd, Wide area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. on Networking*, 1993, no. 3, pp. 226–244.

15. A. Simonian and J. Guibert, Large deviations approximation for fluid queues fed by a large number of ON/OFF sources, *IEEE J. Sel. Areas Commun.*, 1995, vol. 13, no. 6, pp. 1017–1027.