Impact of arrival and service characteristics on the packet queueing performance

ANDRZEJ CHYDZINSKI
Institute of Informatics
Faculty Of Automatic Control, Electronics And Computer Science
Silesian University of Technology
Akademicka 16, Gliwice
POLAND
andrzej.chydzinski@polsl.pl

Abstract: It is known that the performance of queueing mechanisms in computer network nodes deteriorates when traffic is strongly autocorrelated, or has the batch structure, or when the service time is of high variability. It is not obvious however, which of those three factors plays the dominant role in this deterioration. In this paper, using realistic traffic parameterizations with different system loads and buffer sizes, we are trying to determine, which of the three prevails in making queues longer and packet losses higher.

Key–Words: packet queueing, queue size, packet losses, traffic autocorrelation, bursty traffic

1 Introduction

The performance of queueing mechanisms in computer network nodes (switches and routers) depends on characteristics of the processed traffic. Unfortunately, several phenomenons in the network traffic, which may cause significant deterioration of the performance of these mechanisms, have been noticed.

Firstly, it has been observed (see e.g. [1, 2, 3]) that packet interarrival times can be strongly autocorrelated, up to the lag of 10000. The fact that this phenomenon makes queues longer and packet losses higher has been widely disputed in many theoretical, simulation and experimental papers.

Secondly, traffic is often bursty, [4]. The term 'bursty' is associated with the appearance of packets in batches, rather than as single units. Such batch structure of traffic is explained by the design of the TCP protocol and its congestion control mechanism, which uses the variable called window. This variable determines, how many packets can be injected to the network one after another, without a confirmation of the delivery. Again, it is well known that the batch structure of traffic makes queues longer and packet losses higher.

Finally, packets can have differently distributed sizes in different IP networks. As the packet size translates directly to the service (forwarding) time, we have different distributions of the service time. The Pollaczek-Khinchine formula, [5], states that the average queue size in the $M/G/1$ system is equal to:

$$\overline{\Phi} = \rho + \frac{\rho^2 + \lambda^2 Var(F)}{2(1-\rho)},$$

where λ is the arrival rate of the Poisson process, $\rho < 1$ is the load of the system and $Var(F)$ is the variance of the service time distribution. Therefore, in some systems the average queue size may grow linearly with the variance of the service time.

Summarizing, there are at least three important factors that may cause deterioration of the performance of queueing mechanisms, known to be present in real networks: the interarrival time autocorrelation, batch arrivals and variable service times.

In theory, each of the three can play the dominant role in deteriorating the queueing performance. We can obtain an arbitrary large queue size by assuming the service time of large enough variance. We can obtain an arbitrary large queue size by using the arrival process of strong enough, long-range autocorrelation. And we can obtain an arbitrary large queue size by using heavy-tailed batch sizes of large average value. Therefore, searching for the dominant factor in the general case makes no sense.

In this paper, we are trying to find which of the three factors prevails in real networks, i.e. given the *realistic* parameterizations of the network queueing mechanisms. In particular, we know that the typical range of size of an IP packet is 40-1500 bytes. We know also, that in real networks the traffic autocorrelation is observed up to the fourth time scale, not further. We know also, which batch sizes can be expected in the traffic, given the design of the TCP window mechanisms.

To accomplish the goal of the paper, four distinct queueing models will be studied. The first, reference model, will incorporate neither the variable service time, nor the batch structure, nor the traffic autocorrelation. The second model will incorporate the variable service time, but not the batch structure, nor the traffic autocorrelation. The third model will incorporate the batch structure, but not the variable service time, nor the traffic autocorrelation. Finally, the fourth model will incorporate the traffic autocorrelation, but not the variable service time, nor the batch structure.

As was said, in every model, realistic parameters of the traffic and the service process will be used. In addition, each model will be studied in three load scenarios: underloaded system ($\rho = 0.9$), critically loaded system ($\rho = 1$) and overloaded system ($\rho = 1.1$). Moreover, each model will be studied separately with a small buffer ($N = 100$) and a large buffer ($N = 1000$).

The remaining part of the paper is structured as follows. In Section 2, the details of the four queueing models, as well as their parameterizations, are given. Section 3 discusses how the performance of these models can be evaluated. In Section 4, average queue lengths and packet loss ratios are presented and discussed for all the models, in six distinct load and buffer size scenarios. Finally, remarks concluding the paper are gathered in Section 5.

2 Queueing models

The following four models of queueing systems will be considered.

Model (a): Poisson arrivals and constant packet sizes. In Kendall's notation, this the $M/D/1/N$ system. Packet interarrival times are independent and exponentially distributed with parameter λ. Parameter λ will be changed according to the required system load ($\rho = 0.9$, $\rho = 1$ or $\rho = 1.1$). The packet size is constant and equal to 770 bytes, which translates to constant service time (transmission through the 1Gb/s interface). The buffer size is finite and equals N packets, meaning that no more than N packets can be stored, including the service position. If a packet arrives when the buffer is full, it is deleted and lost. Two buffer sizes will be considered: $N = 100$ and $N = 1000$.

This is the reference model – it incorporates neither the variable service time, nor the batch structure, nor the traffic autocorrelation.

Model (b): Poisson arrivals and uniform packet sizes. In Kendall's notation, this the $M/U/1/N$ system. Packet interarrival times are independent and exponentially distributed with parameter λ, while the packet size is uniformly distributed in the range [40;1500] bytes, which translates to uniform distribution of the service time (transmission through the 1Gb/s interface). The buffer size is finite and equals N. Two buffer sizes will be considered, $N = 100$ and $N = 1000$, as well as three load values, $\rho = 0.9$, $\rho = 1$ or $\rho = 1.1$.

This model incorporates the variable service time, but not the batch structure of the traffic, nor the traffic autocorrelation. Therefore, it is designed to test the influence of the bare variability of the packet size on the queueing performance.

Model (c): batch Poisson arrivals and constant packet sizes. In Kendall's notation, this the $M^X/D/1/N$ system. Packets arrive in batches. A realistic batch size distribution was taken from Fig. 5 of [4] and is the following:

$$b_1 = 0.296,\;\; b_2 = 0.231,\;\; b_3 = 0.034,\;\; b_4 = 0.181,$$

$$b_5 = 0.023,\;\; b_6 = 0.021,\;\; b_7 = 0.017,\;\; b_8 = 0.107,$$

$$b_9 = 0.0075,\;\; b_{10} = 0.007,\;\; b_{11} = 0.006,$$

$$b_{12} = 0.006,\;\; b_{13} = 0.005,\;\; b_{14} = 0.005,$$

$$b_{15} = 0.004,\;\; b_{16} = 0.042,\;\; b_{32} = 0.0075.$$

The average batch size is 4.1825. Interarrival times between batches are independent and exponentially distributed with parameter λ, which will be changed according to the required system load. The packet size is constant and equal to 770 bytes, which translates to constant service time (transmission through the 1Gb/s interface). The buffer size is finite and equal to N packets. Two buffer sizes will be considered, $N = 100$ and $N = 1000$, as well as three load values, $\rho = 0.9$, $\rho = 1$ or $\rho = 1.1$.

This model incorporates the batch structure of the traffic, but not the variable service times, nor the traffic autocorrelation. Therefore, it is designed to study the influence of the realistic batch structure on queueing characteristics.

Model (d): MAP arrivals and constant packet sizes. In Kendall's notation, this the $MAP/D/1/N$ system. The interarrival times follow the Markovian arrival process with the following parameters, taken from [6]:

$$D_{1a} = \left[\begin{array}{cc} 2.5582 \cdot 10^0 & 4.3951 \cdot 10^{-2} \\ 1.1369 \cdot 10^{-2} & 6.6173 \cdot 10^{-1} \end{array} \right],$$

$$D_{1b} = \left[\begin{array}{cc} 2.6769 \cdot 10^0 & 6.6924 \cdot 10^{-5} \\ 4.2706 \cdot 10^{-5} & 1.7082 \cdot 10^0 \end{array} \right],$$

$$D_{1c} = \left[\begin{array}{cc} 4.3309 \cdot 10^{0} & 2.7061 \cdot 10^{-4} \\ 6.7564 \cdot 10^{-2} & 2.2578 \cdot 10^{-2} \end{array} \right],$$

$$D_{1d} = \left[\begin{array}{cc} 3.5552 \cdot 10^{1} & 2.9355 \cdot 10^{-1} \\ 2.6962 \cdot 10^{0} & 4.8230 \cdot 10^{0} \end{array} \right],$$

$$D_0 = -D_{0a} \otimes D_{0b} \otimes D_{0c} \otimes D_{0d},$$
$$D_1 = D_{1a} \otimes D_{1b} \otimes D_{1c} \otimes D_{1d},$$

where each D_{0x} matrix is negative and diagonal with i-th element equal in modulus to the sum of the i-th row of the associated D_{1x} matrix. (This MAP process is able to mimic very well the strong autocorrelation observed in the famous Bellcore trace file). The arrival process is then scaled to obtain the required system loads. The packet size is constant and equal to 770 bytes, which translates to constant service time (transmission through the 1Gb/s interface). The buffer size is finite and equal to N packets. Two buffer sizes will be considered, $N = 100$ and $N = 1000$, as well as three load values, $\rho = 0.9$, $\rho = 1$ or $\rho = 1.1$.

This model incorporates the traffic autocorrelation, but not the variable service time, nor the batch structure. Therefore, it is meant to test the influence of the realistic, strong autocorrelation, on the queueing performance.

In every model and every load-buffer scenario, the stationary average queue size, the standard deviation of the queue size size and the loss ratio will be checked. Obviously, the loss ratio is the long-run fraction of packets lost due to the buffer overflow.

3 Models' solution

Each of the aforementioned queueing models is a special case of the $BMAP/G/1/N$ model, i.e. the model with batch Markovian arrival process [7], general distribution of the service time and finite buffer of size N. Therefore, the analytical solution of $BMAP/G/1/N$ can be used for solving each of the models (a)-(d). We will recall this solution in this section.

Therefore, in this section we assume that the arrivals form the BMAP process with m states, parameters $D_0, D_1, D_2, \ldots$, and transition probabilities:

$$p_i(0,i) = 0,$$

$$p_i(0,k) = \frac{1}{\lambda_i}(D_0)_{ik}, \qquad 1 \le i,k \le m, \quad k \ne i,$$

$$p_i(j,k) = \frac{1}{\lambda_i}(D_j)_{ik}, \qquad 1 \le i,k \le m, \quad j \ge 1.$$

where $\lambda_i = -(D_0)_{ii}$.

Let $\mathbb{P}(\cdot)$ denote probability, $X(t)$ denote the state of the modulating chain at time t, $J(t)$ denote the state of the modulating chain at time t, $F(\cdot)$ denote the distribution function of the service time distribution, $f(s)$ denote the Laplace-Stieltjes transform of $F(\cdot)$, $N(t)$ denote the total number of arrivals in $(0,t)$, δ_{ij} denote the Kronecker symbol and

$$P_{i,j}(n,t) = \mathbb{P}(N(t){=}n, J(t){=}j | N(0){=}0, J(0){=}i),$$

$$a_{k,i,j}(s) {=} \int_0^\infty e^{-st} P_{i,j}(k,t) dF(t),$$

$$z(s) = ((s+\lambda_1)^{-1}, \ldots, (s+\lambda_m)^{-1})^T,$$

$$\mathbf{1} = (1, \ldots, 1)^T.$$

Moreover, the following $m \times m$ matrices will be needed:

$$\mathbf{0} - m \times m \text{ matrix of zeroes},$$

$$A_k(s) = [a_{k,i,j}(s)]_{i,j}, \; Y_k(s) = \left[\frac{\lambda_i p_i(k,j)}{s+\lambda_i}\right]_{i,j},$$

$$\overline{D}_k(s) = \left[\int_0^\infty e^{-st} P_{i,j}(k,t)(1-F(t))dt\right]_{i,j},$$

$$\overline{A}_k(s) = \sum_{i=k}^{\infty} A_i(s),$$

$$B_k(s) = A_{k+1}(s) - \overline{A}_{k+1}(s)(\overline{A}_0(s))^{-1},$$

$$R_0(s) = \mathbf{0}, \quad R_1(s) = A_0^{-1}(s),$$

$$R_k(s) {=} R_1(s)(R_{k-1}(s) {-} \sum_{i=0}^{k-1} A_{i+1}(s) R_{k-i}(s)), \; k \ge 2.$$

$$M_N(s) {=} R_{N+1}(s)A_0(s) {+} \sum_{k=0}^{N} R_{N-k}(s)B_k(s) {-} \sum_{k=N+1}^{\infty} Y_k(s)$$
$$- \sum_{k=0}^{N} Y_{N-k}(s)[R_{k+1}(s)A_0(s) {+} \sum_{i=0}^{k} R_{k-i}(s)B_i(s)].$$

Now, let

$$\phi_{n,i}(s,l) = \int_0^\infty e^{-st} \Phi_{n,i}(t,l) dt,$$

and

$$\phi_n(s,l) = (\phi_{n,1}(s,l), \ldots, \phi_{n,m}(s,l))^T,$$

where

$$\Phi_{n,i}(t,l) = \mathbb{P}(X(t) = l | X(0) = n, J(0) = i).$$

In [8], the following theorem was proven.

Theorem 1. *The Laplace transform of the queue length distribution in the $BMAP/G/1/N$ system equals:*

$$\phi_n(s,l) = \sum_{k=0}^{N-n} R_{N-n-k}(s) g_k(s,l) + [R_{N-n+1}(s) A_0(s) + \sum_{k=0}^{N-n} R_{N-n-k}(s) B_k(s)] M_N^{-1}(s) m_N(s,l), \quad (1)$$

with

$$g_k(s,l) = \overline{A}_{k+1}(s)(\overline{A}_0(s))^{-1} r_N(s,l) - r_{N-k}(s,l),$$

$$r_n(s,l) = \begin{cases} \mathbf{0}\cdot\mathbf{1}, & \text{if } l < n, \\ \overline{D}_{l-n}(s)\cdot\mathbf{1}, & \text{if } n \le l < N, \\ \frac{1-f(s)}{s}\cdot\mathbf{1} - \sum_{k=0}^{N-n-1} \overline{D}_k(s)\cdot\mathbf{1}, & l = N. \end{cases}$$

$$m_N(s,l) = \sum_{k=0}^{N} Y_{N-k}(s) \sum_{i=0}^{k} R_{k-i}(s) g_i(s,l) - \sum_{k=0}^{N} R_{N-k}(s) g_k(s,l) + \delta_{0l} z(s).$$

This theorem can be used for solving all four queueing models, (a)-(d), exploited in this paper. Alternatively, the simple models of $M/G/1/N$ type and $M^X/G/1/N$ type can be solved using simplified formulas given in [9]. The loss ratio can be obtained using formula (see e.g. [10]):

$$L = 1 - \frac{1 - \Phi_{0,1}(\infty, 0)}{\rho},$$

where ρ is the system load.

Alternatively, the performance of queueing systems (a)-(d) can be studied using simulations, e.g. with the help of Omnet++ simulator, [11], and the simulation of the BMAP process presented in [12].

4 Results and discussion

The results are presented in Tables 1-6.

In particular, in Tab. 1, the results for an underloaded system and a small buffer are presented. As we can see, distributed packet sizes in (b) have a surprisingly small impact on the queueing performance - the queue and the losses are only slightly higher that in (a). On the other hand, both the batch structure and the autocorrelation enlarge the queue size several times and the losses by several orders of magnitude. Comparing models (c) and (d), we can conclude that both give results in the same ballpark (the autocorrelation has a slightly deeper influence of the queueing performance).

In Tab. 2, the results for an underloaded system and a large buffer are presented. Distributed packet sizes in (b) have a very small impact on the queueing performance. Contrary to Tab. 1, the impact of the batch structure and the autocorrelation is very different now. The batch structure enlarges the queue size 8 times, while the autocorrelation 72 times. As for the loss ratio, it is not significantly increased in (c), and drastically increased in (d).

In Tab. 3, the results for a critically loaded system and a small buffer are presented. As we can see, neither the distributed packet sizes, nor the batch structure, nor the autocorrelation, have a significant impact on the queue size in this scenario. As regards the loss ratio, it does not change for distributed packet sizes, but enlarges 8 times in the case of the batch structure, and 24 times in the case of the autocorrelation. The latter is especially bad, as it means a lot of packet losses – as much as 12%.

In Tab. 4, the results for a critically loaded system and a large buffer are presented. They are similar to those in Tab. 3 in the sense that neither the distributed packet sizes, nor the batch structure, nor the autocorrelation has a deep impact on the queue length. The loss ratio is almost the same in (a) and (b), then increases 8 times in (c) and 192 times in (d).

In Tab. 5, the results for an overloaded system and a small buffer are presented. As we can see, distributed packet sizes in (b) have a negligible impact on the queueing performance. The loss ratio is a little higher in (c) and (d) than in (a). Surprisingly, the queue size decreases a little in (c), if compared with (a), and almost twice in (d). This counterintuitive phenomenon is causes by higher loss ratio, which makes the actual, carried load, lower in (c) and (d), than in (a). This phenomenon was studied in detail in [13].

Finally, in Tab. 6, the results for an overloaded system and a large buffer are presented. In this case, both distributed packet sizes in (b) and the batch structure have a negligible impact on the queue size and loss ratio. The autocorrelation, on the other hand, increases the losses, but decreases the queue size almost two times. This is again connected with the decreased carried load of the system in (d).

Summarizing, we see that distributed packet sizes have a minor or negligible influence on the queueing performance in all the considered scenarios. Both the batch structure and the autocorrelation have a negative impact on the loss ratio in all the considered scenarios, but the impact of the autocorrelation is usually stronger. It can be extremely strong, when the load is small and the buffer is large (Tab. 2). The batch structure and the autocorrelation have also a negative impact on the queue size, but in an underloaded system only. (The impact of the autocorrelation is stronger).

system type	avg. queue length	stddev. queue length	loss ratio
(a) Poisson arrivals, constant packet sizes	4.95	4.89	$1.1{\cdot}10^{-10}$
(b) Poisson arrivals, uniform packet sizes	6.15	6.23	$1.0{\cdot}10^{-8}$
(c) batch arrivals, constant packet sizes	29.2	26.4	$1.2{\cdot}10^{-2}$
(d) MAP arrivals, constant packet sizes	35.1	35.7	$8.5{\cdot}10^{-2}$

Table 1: Average queue length, its standard deviation and the loss ratio, ρ =0.9 and N =100.

system type	avg. queue length	stddev. queue length	loss ratio
(a) Poisson arrivals, constant packet sizes	4.95	4.89	$< 10^{-10}$
(b) Poisson arrivals, uniform packet sizes	6.15	6.23	$< 10^{-10}$
(c) batch arrivals, constant packet sizes	40.7	45.2	$5.6{\cdot}10^{-10}$
(d) MAP arrivals, constant packet sizes	358	427	$5.5{\cdot}10^{-2}$

Table 2: Average queue length, its standard deviation and the loss ratio, ρ =0.9 and N =1000.

system type	avg. queue length	stddev. queue length	loss ratio
(a) Poisson arrivals, constant packet sizes	50.0	28.8	$5.0{\cdot}10^{-3}$
(b) Poisson arrivals, uniform packet sizes	49.7	28.9	$6.4{\cdot}10^{-3}$
(c) batch arrivals, constant packet sizes	47.7	30.0	$4.2{\cdot}10^{-2}$
(d) MAP arrivals, constant packet sizes	44.5	37.8	$1.2{\cdot}10^{-1}$

Table 3: Average queue length, its standard deviation and the loss ratio, ρ =1.0 and N =100.

system type	avg. queue length	stddev. queue length	loss ratio
(a) Poisson arrivals, constant packet sizes	478	290	$5.1{\cdot}10^{-4}$
(b) Poisson arrivals, uniform packet sizes	492	283	$5.6{\cdot}10^{-4}$
(c) batch arrivals, constant packet sizes	484	288	$4.2{\cdot}10^{-3}$
(d) MAP arrivals, constant packet sizes	409	440	$9.8{\cdot}10^{-2}$

Table 4: Average queue length, its standard deviation and the loss ratio, ρ =1.0 and N =1000.

system type	avg. queue length	stddev. queue length	loss ratio
(a) Poisson arrivals, constant packet sizes	94.8	5.21	$9.0 \cdot 10^{-2}$
(b) Poisson arrivals, uniform packet sizes	93.3	6.80	$9.0 \cdot 10^{-2}$
(c) batch arrivals, constant packet sizes	66.5	26.4	$1.0 \cdot 10^{-1}$
(d) MAP arrivals, constant packet sizes	52.9	38.2	$1.7 \cdot 10^{-1}$

Table 5: Average queue length, its standard deviation and the loss ratio, ρ =1.1 and N =100.

In the overloaded system, we can encounter reverse effect – the queue size might be smaller due to smaller load carried by the system.

system type	avg. queue length	stddev. queue length	loss ratio
(a) Poisson arrivals, constant packet sizes	994	10.7	$9.0{\cdot}10^{-2}$
(b) Poisson arrivals, uniform packet sizes	993	11.8	$9.0{\cdot}10^{-2}$
(c) batch arrivals, constant packet sizes	955	43.3	$9.0{\cdot}10^{-2}$
(d) MAP arrivals, constant packet sizes	471	427	$1.3{\cdot}10^{-1}$

Table 6: Average queue length, its standard deviation and the loss ratio, ρ =1.1 and N =1000.

5 Conclusions

In this paper, we tried to determine, which factor among the three: the variable service time, the batch traffic structure and the traffic autocorrelation, causes the worst deterioration of the queueing performance. We used four models of queueing systems with realistic traffic parameterizations, system loads and buffer sizes.

The variable service time has a minor impact on the packet loss ratio. On the other hand, the autocorrelation has a deep, negative impact on this characteristic. The batch structure enlarges the number of losses as well, but not that strongly, as the autocorrelation.

The variable service time has a minor impact on the queue size. The batch structure and the autocorrelation have a deep, negative impact on this characteristic (the impact of the autocorrelation is stronger), but only when the system is underloaded. If the system is overloaded, a reverse effect can be observed - the queues can get shorter when the traffic has the batch structure of when it is autocorrelated.

Acknowledgements: This work was conducted in within the statutory research project of the Institute of Informatics, BK-213/RAU2/2018.

References:

[1] Leland W., Taqqu M., Willinger W. and Wilson D. On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Transactions on Networking 2(1): pp. 1–15, (1994).

[2] Willinger W., Taqqu M. S., Sherman R. and Wilson D. V. Self-Similarity through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. Proceedings of ACM SIGCOMM95, pp. 100–113, (1995).

[3] Crovella M. and Bestavros A. Self-similarity in World Wide Web traffic: Evidence and possible causes. IEEE/ACM Transactions on Networking 5(6), pp. 835-846, (1997).

[4] Garetto M. and Towsley D. An efficient technique to analyze the impact of bursty TCP traffic in wide-area networks. Performance Evaluation 65, pp. 181-202, (2008).

[5] Takagi, H. *Queueing analysis - Vacation and Priority System*. North-Holland, Amsterdam, (1991).

[6] Casale, G., Zhang, E. Z. and Smirni, E. Trace data characterization and fitting for Markov modeling. Performance Evaluation, Volume 67, Issue 2, pp. 61–79, (2010).

[7] Lucantoni D. M. New results on the single server queue with a batch Markovian arrival process. Commun. Stat., Stochastic Models 7, no. 1, pp. 1–46, (1991).

[8] Chydzinski A. Queue Size in a BMAP Queue with Finite Buffer. Lecture Notes in Computer Science, vol. 4003, pp. 200–210, (2006).

[9] Takagi H. *Queueing analysis - Finite Systems*. North-Holland Amsterdam, (1993).

[10] Chydzinski A. and Mrozowski P. Queues with dropping functions and general arrival processes, PLoS ONE, vol. 11, issue 3, e0150702. doi:10.1371/journal.pone.0150702, (2016)

[11] http://www.omnetpp.org/

[12] Chydzinski A. and Winiarczyk R. On the Blocking Probability in Batch Markovian Arrival Queues. Microprocessors and Microsystems, 32/1, pp. 45–52, (2008).

[13] Chydzinski, A. On the infuence of traffic autocorrelation on the queueing performance. Proc. of World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2016), pp. 37-42, Orlando, July (2016).