Moment properties and long-range dependence of queueing processes

Prof. Evsei V. Morozov, Alexander S. Rumyantsev

Institute of Applied Mathematical Research,
Karelian Research Centre, RAS

E-mail: {emorozov, ar0}@krc.karelia.ru

Abstract

The aim of this study is to empirically extend known results on long range dependence for a separated system to 2-station tandem system. More exactly, we study connection of a long range dependence effect at the second station in a tandem network with moment properties of the input and service times in both stations. Simulation results are presented, and some related difficulties are discussed.

1 Introduction

The main motivation of this study is to empirically verify results obtained in the paper [4]. This verification for a separate system has been presented in [2].

We briefly recall related definitions and results. Consider a single-server queueing system GI/G/1 with a renewal input with arrival epochs $\{t_i\}$ and the i.i.d. interarrival times $\{T_i = t_{i+1} - t_i\}$ with distribution $A(x) = P(T \leq x)$ (T denotes generic interarrival). It is assumed that service times $\{S_i\}$ are i.i.d with distribution $B(x) = P(S \leq x)$. Denote by W_i the waiting time of customer i in queue. Recall famous Lindley's recursion which defines the sequence $\{W_i\}$:

$$W_{i+1} = (W_i + S_i - T_i)^+,$$

where $(\cdot)^+ = \max(0, \cdot)$. Let the stability condition holds, i.e. $ES/ET < 1$, or, equivalently, $E(S - T) < 0$. Then the sequence $\{W_i\}$ has a weak limit $W_i \Rightarrow W$, $i \to \infty$. Moreover, the following stochastic equality connects this stationary limit and supremum of the associated random walk with negative drift:

$$W = \sup_{i \geqslant 1}(S_i - T_i).$$

The stationary waiting time (or *delay*) W is widely used as a QoS parameter. In particular, it is important to know the properties of the delay for delay-sensitive systems, such as real-time voice and video traffic. We give the following well known result from [4] defining *long range dependence* of the sequence of delays in a GI/G/1 system. More exactly, provided that $ET < \infty, ES^3 < \infty$ and $ES^4 = \infty$,

$$\sum_{i=1}^{\infty} corr(W_0, W_i) = \infty. \tag{1.1}$$

In practice the result (1.1) means that the waiting time process stays above or below it's mean value unexpectedly long. It makes difficult the use of sample mean based estimator to estimate required parameter with a given accuracy in reasonable simulation time [9]. An extended discussion of this topic is in the work [8]

It is obvious that $t_i + W_i + S_i$ is the departure instant of customer i. Hence, these instants define an inter-departure process,

$$\begin{aligned} D_i &= t_{i+1} + W_{i+1} + S_{i+1} - t_i - W_i - S_i \\ &= S_{i+1} + (T_i - W_i - S_i) + W_{i+1} = S_{i+1} + (T_i - W_i - S_i)^+, \, i \geq 1. \end{aligned}$$

Consider a two-station tandem, where after being served in the first single-server queue, the customer (task) enters the second system. Thus, the output from first system is an input to the second one. For node j, denote by $\{T_i^{(j)}\}$ interarrival times, by $\{S_i^{(j)}\}$ service times and by $\{W_i^{(j)}\}$ waiting times, $j = 1, 2$, and note that $T_i^{(2)} = D_i^{(1)}$. It is interesting to estimate the impact of characteristics of the first station on the delay at the second one. The main difference from first station, is that the inter-arrival times for second station form not a renewal but rather a regenerative process. Results on the output process can be found in [3].

The stability condition for the whole network is [12],

$$ET^{(1)} > \max(ES^{(1)}, ES^{(2)}),$$

while condition

$$\max(ET^{(1)}, ES^{(1)}) > ES^{(2)},$$

implies stability of the second station solely, $W_i^{(2)} \Rightarrow W^{(2)}$ (with proper limit $W^{(2)}$), leaving a possibility of instability of first station, $W^{(1)} \to \infty$ (in probability of with probability 1).

In a stable system, the knowledge of moment properties of delay may be useful for instance, to approximate the tail of the delay via Chebyshev's

inequality. A well-known (for separated station) result that the finiteness of $r+1$-th moment of service time implies finiteness of the r-th moment of waiting time is extended to a tandem network in [12]. More exactly, for a two-station case, if $ES^{(1)} > ES^{(2)}$, then (similar to one station case) sufficient stability condition holds:

$$E\left(S^{(2)}\right)^{r+1} < \infty \quad \Rightarrow \quad E\left(W^{(2)}\right)^{r} < \infty.$$

However, if $ES^{(1)} \leqslant ES^{(2)}$, then an additional condition is placed on moments of service times at the first station:

$$E\left(S^{(j)}\right)^{r+1} < \infty \quad \Rightarrow \quad E\left(W^{(2)}\right)^{r} < \infty, \quad j = 1, 2.$$

It turns out to be that the latter assumption is not only technical one caused by the method of the proof (as it has been conjectured in [12]), but as has been shown in [10], violation of the assumption may lead to *infinite mean stationary delay* at the second node.

Even more surprising dependence of moment properties at a given station on the properties of other stations in tandem-like networks is found in [7] for the so-called *heavy-tailed* case. First recall that a random variable (r.v.) X with distribution F is called subexponential if asymptotic equivalence holds $P(X_1 + X_2 > x) \sim 2(1 - F(x)) := 2\bar{F}(x)$, where $X_{1,2}$ are i.i.d. copies of X. A particular case is Pareto r.v. with tail distribution (for $x \geqslant x_0 > 0$)

$$\overline{F}(x) = x^{-\alpha}, \quad \alpha > 0. \tag{1.2}$$

If $\overline{B}(x)$ is the tail distribution of service time S, then an integrated tail distribution (of a stationary remaining service time S_e) is defined as

$$\overline{B}_e(x) := P(S_e > x) = \frac{1}{ES} \int_x^\infty \overline{B}(x)\, dx, \ x \geq 0.$$

When both $\overline{B}(x)$ and $\overline{B}_e(x)$ are subexponential, then we call that distribution B belongs to a useful subclass $\mathcal{S}^*$.

The crucial result of [7] (for two-station tandem) is as follows. Assume stability, that is $\rho_i := ES^{(i)}/ET^{(1)} < 1$ for $i = 1, 2$, and let the service time distribution at the second station belong to $\mathcal{S}^*$. Also assume that $P(S^{(1)} > x) = o(P(S^{(2)} > x))$ and that service time at the first station also belongs to $\mathcal{S}^*$ or is *light-tailed* [7]). Then

$$P(W^{(2)} > x) \sim \frac{\rho_2}{1 - \rho_2} P(S_e^{(2)} > x).$$

In other words, if service time at first station has lighter tail then the tail of delay asymptotically behaves like in a single station, and previous station does not matter.

Note that some of given above results hold for some more general networks.

2 Long range dependence in tandem

In this section we discuss some difficulties which arise when we try to empirically verify the theoretical results mentioned in the previous section.

2.1 Pareto tail modeling

To simulate the i.i.d r.v. $X_1, \dots, X_n$ with a given distribution F, we sample i.i.d. pseudo-random numbers $U_1, \dots, U_n$ and then use inverse transform. More exactly, we use the inverse function $F^{-1}(U_i)$ to get the sample values X_i:

$$X_i = \overline{F}^{-1}(U_i),\ i = 1, \dots, n.$$

(It ie easy to check that obtained r.v. indeed have distribution F.) In particular, for Pareto (tail) distribution (1.2),

$$X_i = U_i^{-1/\alpha},\ i = 1, \dots, n.$$

The problem which arises in practice is that the values of U_i have limited accuracy, say, $U_i \geqslant 10^{-\beta}$ for some $\beta > 1$. Then the maximum value $x_{\max}$ obtained by inverse transform sampling is

$$x_{\max} \leqslant 10^{\beta/\alpha}.$$

(Note that in this case the sample size has to be approximately 10^{β}.) Thus, instead of sampling from Pareto distribution we in fact obtain *truncated Pareto* distribution [5], that is

$$\overline{F}(x) = \frac{(x_0 x_{\max})^{\alpha}}{x_{\max}^{\alpha} - x_0^{\alpha}}(x^{-\alpha} - x_{\max}^{-\alpha}), \quad x_0 \leqslant x \leqslant x_{\max} < \infty,$$

with $\overline{F}(x) = 0$ for $x \geq x_{\max}$ and $\overline{F}(x) = 1$ for $x \leq x_0$. (We mention an asymptotic level-q test in [1] to verify the hypothesis about the truncated Pareto distribution.) We recall that for classical Pareto (1.2) $EX^n = \infty$ for $n \geqslant \alpha$. However, in our case,

$$EX^n = \int_{x_0}^{x_{\max}} x^n \frac{\alpha(x_0 x_{\max})^{\alpha}}{x_{\max}^{\alpha} - x_0^{\alpha}} x^{-\alpha-1}\, dx = \frac{\alpha(x_0 x_{\max})^{\alpha}}{n - \alpha} \frac{x_{\max}^{n-\alpha} - x_0^{n-\alpha}}{x_{\max}^{\alpha} - x_0^{\alpha}}.$$

Hence, if $x_{\max} = 10^{\beta/\alpha}$ and $x_0 = 1$ (for standard Pareto),

$$EX^n = \frac{\alpha 10^{\beta}}{n-\alpha} \frac{10^{(n-\alpha)\beta/\alpha} - 1}{10^{\beta} - 1} \approx \frac{\alpha}{n-\alpha} 10^{\beta\cdot(n-\alpha)/\alpha}. \tag{2.1}$$

For instance, let $\beta = 16$ (the double precision accuracy of C language defined variable), $\alpha = 3.5$ and $n = 4$. Substitution this in (2.1) implies

$$EX^n \approx 7 \cdot 10^{2.2857}.$$

This value is far from being infinite. Moreover, for our case, to have EX^n at least an order of 10^{10}, one needs $\beta \approx 70$, see from (2.1). (The so-called *long arithmetics* provides an arbitrary order of accuracy but sample size 10^{70} hard to get in a reasonable simulation time.)

Nevertheless, note that if $n > \alpha$ and α approaches zero, then $(n-\alpha)/\alpha$ increases. Thus, for $0 < \alpha < 2$ one may get reasonable results for the value of EX^n.

2.2 Numerical results

The experiments were carried out on a High-Performance cluster [6]. The autocorrelation coefficients were calculated by formulae

$$\hat{\rho}_i = \frac{M\sum_{j=1}^{M} W_0(j)W_i(j) - \sum_{j=1}^{M} W_0(j)\sum_{j=1}^{M} W_i(j)}{M\sum_{j=1}^{M} (W_0(j))^2 - \left(\sum_{j=1}^{M} W_0(j)\right)^2},$$

where $W_i(j)$ corresponds to the waiting time for task i in the independent run j. Note that independent runs are preferable than a single long run in the presence of long-range dependence [11].

The problem discussed in the previous subsection means that if in distribution (1.2) $\alpha < 4$, we in fact obtain empirically finite forth moment implying convergence of autocorrelation series. A possibility to obtain (quasi)divergence in simulation is to take coefficient $\alpha < 2$, in which case the variance of stationary delay is (theoretically) infinite. Thus the main conclusion is that it is difficult to verify long-range dependence of the workload (delay) process neither in single-server, nor in tandem case, applying divergence of the autocovariance series stated in [4].

Nevertheless, an interesting case that leads to the divergence of autocorrelation series is an instability of a station. Consider an M/Pareto/1 $\rightarrow$ /Pareto/1 tandem system renewal input with interarrival time T and with (corresponding) Pareto service time (in more convenient for simulation form)

$$P(S^{(i)} > x) = (1+x)^{-\alpha_i},\ x \geq 0,\ i = 1, 2.$$

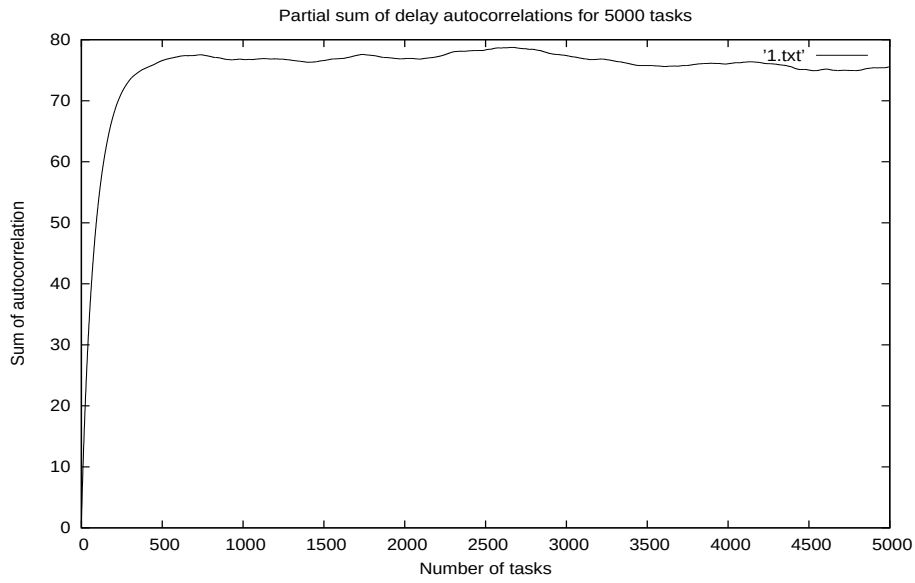

Figure 1: Convergence of autocorrelations, $\lambda = 3, \mu_1 = 3.5, \mu_2 = 4.2$.

Then the mean service time is $ES^{(i)} = (\alpha_i - 1)^{-1}$, $i = 1, 2$. Denote $\lambda = 1/ET$, $\mu_i = 1/ES^{(i)}$ and assume that $\mu_1 < \lambda < \mu_2$. Then first station becomes overloaded (because $\rho_1 = \lambda/\mu_1 > 1$) and in the limit has the output with rate μ_1. But because $\mu_1/\mu_2 < 1$, then the second station is stable (in limit) and we do not observe divergence of autocorrelations of waiting times, as Fig. 1 shows. If $\mu_2 < \lambda < \mu_1$, then the first station is stable, but the second station is unstable (since $\rho_2 > 1$). In this case the delays on the second station may become arbitrary high implying the divergence of autocorrelation series, see Fig. 2.

3 Conclusion

Detection of the long-range dependence in the networking traffic is extremely important to estimate QoS provided in the network. In this note, we verify by simulation this (second-order) property of the workload process in the second station of a two-station tandem network. We discuss the difficulties (caused by technical limitations) which arise when we apply simulation to establish (under appropriate moment conditions) divergence of the autocorrelation series, indicating theoretically the long-range dependence.

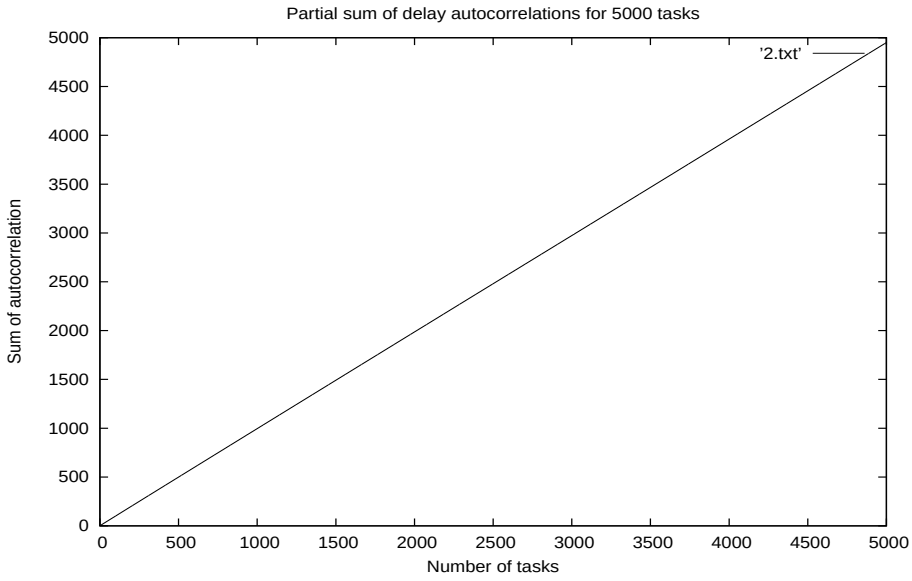

Figure 2: Divergence of autocorrelations, $\lambda = 3, \mu_1 = 4.2, \mu_2 = 3.5$.

4 Acknowledgments

This work is supported by Russian Foundation for Basic research, project No 10-07-00017 and done in the framework of the Strategy development Program for 2012-2016 "PetrSU complex in the research-educational space of European North: strategy of the innovation development.?

Bibliography

[1] I. Aban, M. Meerschaert, A. Panorska *Parameter Estimation for the Truncated Pareto Distribution* Journal of the American Statistical Assoc. vol. 101, 2006. pp. 270–277.

[2] D. V. Bodyonov, and E. V. Morozov, *Regenerative simulation of a tandem network with long-range dependent workload process.* Proceedings of FDPW'2004, vol. 6, Petrozavodsk State University, 2005. pp. 170–180.

[3] D. Daley *Queueing Output Process* Adv. Appl. Prob., vol. 8, Applied Probability Trust, 1976. pp. 395–415.

[4] D. Daley *The serial correlation coefficients of waiting times in a stationary single server queue.* J. Austr. Math. Soc. vol. 8, 1968. pp. 683–699.

[5] M. Harchol-Balter *The Effect of Heavy-Tailed Job Size Distribution on Computer System Design.* Proceedings of ASA-IMS Conference on Applications of Heavy Tailed Distributions in Economics, Engineering and Statistics, Washington, DC. 1999.

[6] *High-performance Data Center, KRC RAS* http://cluster.krc.karelia.ru/

[7] T. Huang, K. Sigman. *Steady-state asymptotics for tandem, split-match and other feedforward queues with heavy tailed service* Queueing Systems, vol. 33, 1999. pp. 233–259.

[8] E. Morozov, A. Rumyantsev *Regeneration and correlation properties of stationary delay in single-server queue (in Russian)* Proceedings of the International Wokrshop on Distributed Computer and Communication Networks (DCCN-2010). Moscow: R&D Company "Information and Networking Technologies", 2010. pp. 58–67.

[9] G. Samorodnitsky. *Long Range Dependence* Foundations and Trends in Stochastic Systems. vol. 1, No. 3, 2006. pp. 163–257.

[10] A. Scheller-Wolf, K. Sigman *Moments in Tandem Queues* Operations Research, vol. 46, 1996. pp. 378–380.

[11] W. Whitt. *The Efficiency of One Long Run versus Independent Replications in Steady-State Simulation* Management Science. vol. 37, No. 6, 1991. pp. 645–666.

[12] B. Wolfson *Some Moment Results for Certain Tandem and Multiple-Server Queues* J. Appl. Prob., vol. 21, Applied Probability Trust, 1984. pp. 901–910.