Queue normalization methods in systems GI/GI/1/*m* with infinite variance of service time

V N Zadorozhnyi[1], T R Zakharenkova[1], M Pagano[2]

[1] Omsk State Technical University, 11, Mira Pr., Omsk 644050, Russia
[2] University of Pisa, 43, Lungarno Pacinotti, Pisa 56126, Italy

E-mail: zwn2015@yandex.ru, ZakharenkovaTatiana@gmail.com, Michele.Pagano@iet.unipi.it

Abstract. Queuing systems with an infinite variance of service time are considered. The average waiting time in such systems is equal to infinity at a stationary regime. We analyze the efficiency of introducing of absolute priorities with infinite number of priority classes determined by the special axis marking on intervals for possible values of service time. It is stated that queues in systems become normalized, i.e. the average queue length become finite, when using regular marking. Furthermore, request loss probabilities radically decrease when buffer size is finite. More efficient marking – exponential marking – is proposed for practical purposes in networks with fractal traffic. The optimization problems of regular and exponential markings are solved.

1. Introduction

In fractal traffic, increasing the buffer size of network devices is an inefficient tool for the loss probability reducing [1–6]. It is explained by infinite stationary average queue length in system GI/GI/1 when the variance Var(x) of request service time x equals infinity. Let us consider the M/Pa/1 queueing system with Pareto distribution of service time, the system commonly used for modeling network devices. In truth, let us consider, for instance, the M/Pa/1 queueing system with Pareto distribution of service time

$$F(t) = 1 - \left(\frac{K}{t}\right)^{\alpha}, \qquad K > 0, \quad t \geq K, \tag{1}$$

where K is the least value of the random variable x (scale parameter) and $\alpha > 0$ is shape parameter.
The mathematical expectation (m.e.) $E(x) = b < \infty$ when $\alpha > 1$, and the variance $\mathrm{Var}(x) = \infty$ at $\alpha \leq 2$. Therefore, the range $1 < \alpha \leq 2$ of parameter α values, frequently used at network devices modeling, determines finite m.e. b and infinite variance of the service time x.
Consequently, the second moment $b^{(2)}$ of time x is infinite as well.
Applying the Pollaczek-Khinchine formula to the considered system with $b^{(2)} = \infty$, average waiting time for the system is found to be

$$W = \frac{\lambda b^{(2)}}{2(1-\rho)} = \infty$$

at any load coefficient $\rho \in (0, 1)$ and any rate $\lambda > 0$ of arrival flow. In accordance with Little's formula $L = \lambda W$, the average queue length L in such system is also infinite. This very result was obtained in [6] by determination of the average queue length by means of moment generating function.
We found that if the discipline of absolute priorities with afterservice and with the infinite number of priority classes determined by requests service time is introduced in such system M/Pa/1, then the average waiting time W becomes finite. At the same time, the priority classes are set by the following regular marking (RM) of the range of possible service time x:

$$\{t_k\} = t_0, t_1, ..., t_k, ..., \tag{2}$$

where $t_0 = K$ and $t_k - t_{k-1} = \Delta = \text{const} > 0$ for any $k = 1, 2, \ldots$. Such marking divides the range of possible service time into the intervals $[t_{k-1}, t_k)$ with length Δ, therefore, the values t_k can be calculated by the formula $t_k = K + k\Delta$. If request entering the original non-priority system M/Pa/1 has a service time belonging to the k^{th} interval, it will be associated with the k^{th} priority in descending order. As the number of priority classes is infinite one can assign negative priorities to requests.

2. The queue normalization theorem in a system M/Pa/1 with finite m.e. and infinite variance of service time

Let us find average waiting time W in a system M/Pa/1 having arrival flow divided into priority components according to RM (2) of service time. Due to [7], average staying time U_k of the request with k^{th} priority class in the system can be expressed as follows:

$$U_k = \frac{b_k}{1-\sigma_{k-1}} + \frac{\sum_{i=1}^{k} \lambda_i b_i^{(2)}}{2(1-\sigma_k)(1-\sigma_{k-1})}, \quad k = 1, 2, \ldots, \tag{3}$$

where b_k is average service time for requests of the k^{th} priority class,

λ_i is the arrival rate of requests with the i^{th} priority class,

$b_i^{(2)}$ is the second service moment for the requests of the i^{th} priority class,

$\sigma_k = \sum_{i=1}^{k} \rho_i$ is the sum of system load coefficients due to the requests of priority classes from the 1st to the k^{th}, $\rho_i = \lambda_i b_i$, $\sigma_{k-1} = \sigma_k - \rho_k$.

Then average time U of requests staying in the system can be determined as the sum

$$U = \sum_k p_k U_k, \tag{4}$$

where $p_k = \mathrm{P}(t_{k-1} \le x < t_k)$ is the probability of arriving request classified as the k^{th} priority class, and average waiting time is a difference

$$W = U - \mathrm{M}(x) = U - b. \tag{5}$$

Indicators (4) and (5) of the considered system depend on the marking of $\{t_k\} = t_0, t_1, ..., t_k, ...$ service time axis. Let us find for the given marking $\{t_k\}$ the parameter values for the system M/Pa/1, included in the right side of the formula (3).

Probability $p_k = \mathrm{P}(t_{k-1} \le x < t_k)$ of interval $[t_{k-1}, t_k)$ at distribution (1) of service time x equals

$$p_k = \left(\frac{K}{t_{k-1}}\right)^{\alpha} - \left(\frac{K}{t_k}\right)^{\alpha} = K^{\alpha}(t_{k-1}^{-\alpha} - t_k^{-\alpha}), \tag{6}$$

hence, providing that $x \in [t_{k-1}, t_k)$ conditional distribution function for service time *x* has the form:

$$F_k(t) = F(t \mid t_{k-1} \le x < t_k) = \frac{F(t) - F(t_{k-1})}{p_k} = \frac{t_{k-1}^{-\alpha} - t^{-\alpha}}{t_{k-1}^{-\alpha} - t_k^{-\alpha}}, \qquad t_{k-1} \le x < t_k . \tag{7}$$

Therefore, for formula (3) we obtain:

$$b_k = \int_{t_{k-1}}^{t_k} t dF_k(t) = \int_{t_{k-1}}^{t_k} t \frac{\alpha t^{-\alpha-1}}{t_{k-1}^{-\alpha} - t_k^{-\alpha}} dt = \frac{\alpha}{\alpha - 1} \cdot \frac{t_{k-1}^{1-\alpha} - t_k^{1-\alpha}}{t_{k-1}^{-\alpha} - t_k^{-\alpha}}, \tag{8}$$

$$\lambda_i = p_i \lambda , \tag{9}$$

(all priority components of the arrival flow are Poisson),

$$b_i^{(2)} = \int_{t_{i-1}}^{t_i} t^2 dF_i(t) = \frac{\alpha}{\alpha - 2} \cdot \frac{t_{i-1}^{2-\alpha} - t_i^{2-\alpha}}{t_{i-1}^{-\alpha} - t_i^{-\alpha}}, \quad \alpha \ne 2, \tag{10}$$

(when $\alpha = 2$ the uncertainty of 0/0 type is disclosed by L'Hopital rule),

$$\sigma_k = \sum_{i=1}^{k} \rho_i = \sum_{i=1}^{k} \lambda_i b_i ,$$

$$\sigma_{k-1} = \sigma_k - \rho_k ,$$

$$\rho_k = \lambda_k b_k . \tag{11}$$

Theorem. Introducing into the system M/Pa/1 the discipline of absolute priorities with afterservice which are determined by the infinite RM with positive step Δ makes the average waiting time finite
Proof. With absolute priorities being introducing, the average staying time (4) is determined as follows:

$$U = \sum_k p_k U_k = W(\Delta) = \sum_{k=1}^{\infty} \frac{p_k b_k}{(1 - \sigma_{k-1})} + \sum_{k=1}^{\infty} \frac{p_k \sum_{i=1}^{k} \lambda_i b_i^{(2)}}{2(1 - \sigma_k)(1 - \sigma_{k-1})} . \tag{12}$$

Here the first sum on the right is finite due to the following relations:

$$S_1 = \sum_{k=1}^{\infty} \frac{p_k b_k}{1 - \sigma_{k-1}} \le \sum_{k=1}^{\infty} \frac{p_k b_k}{1 - \rho} = \frac{1}{1 - \rho} \sum_{k=1}^{\infty} p_k b_k = \frac{b}{1 - \rho}, \tag{13}$$

where $\rho < 1$ is the system load coefficient.

The second sum on the right (12) is limited above

$$S_2 = \sum_{k=1}^{\infty} \frac{p_k \sum_{i=1}^{k} \lambda_i b_i^{(2)}}{2(1 - \sigma_k)(1 - \sigma_{k-1})} \le \sum_{k=1}^{\infty} \frac{p_k \sum_{i=1}^{k} \lambda_i b_i^{(2)}}{2(1 - \rho)(1 - \rho)} = \frac{1}{2(1 - \rho)^2} \sum_{k=1}^{\infty} p_k \sum_{i=1}^{k} p_i \lambda b_i^{(2)}$$

$$= \frac{\lambda}{2(1 - \rho)^2} \sum_{k=1}^{\infty} p_k \sum_{i=1}^{k} p_i b_i^{(2)} = \frac{\lambda}{2(1 - \rho)^2} S_3 . \tag{14}$$

Hence, only the finiteness of the sum $S_3 = \sum_{k=1}^{\infty} p_k \sum_{i=1}^{k} p_i b_i^{(2)}$ remains to be proven.

By inserting the corresponding expressions (6) and (10), we obtain:

$$S_3 = \sum_{k=1}^{\infty} K^{\alpha}(t_{k-1}^{-\alpha} - t_k^{-\alpha}) \sum_{i=1}^{k} K^{\alpha}(t_{i-1}^{-\alpha} - t_i^{-\alpha}) \frac{\alpha}{\alpha - 2} \cdot \frac{t_{i-1}^{2-\alpha} - t_i^{2-\alpha}}{t_{i-1}^{-\alpha} - t_i^{-\alpha}}$$

$$= \frac{\alpha K^{2\alpha}}{\alpha - 2} \sum_{k=1}^{\infty} (t_{k-1}^{-\alpha} - t_k^{-\alpha}) \sum_{i=1}^{k} (t_{i-1}^{2-\alpha} - t_i^{2-\alpha}) < \frac{\alpha K^{2\alpha}}{\alpha - 2} \sum_{k=1}^{\infty} (t_{k-1}^{-\alpha} - t_k^{-\alpha}) \sum_{i=1}^{k} (t_{i-1} - t_i)$$

$$= \frac{\alpha K^{2\alpha}}{\alpha - 2} \sum_{k=1}^{\infty} (t_{k-1}^{-\alpha} - t_k^{-\alpha}) \sum_{i=1}^{k} \Delta = \frac{\alpha K^{2\alpha} \Delta}{\alpha - 2} \sum_{k=1}^{\infty} (t_{k-1}^{-\alpha} - t_k^{-\alpha}) k = \frac{\alpha K^{2\alpha} \Delta}{\alpha - 2} S_4, \quad (15)$$

Let us calculate the sum S_4, considering that the coordinates of the regular marking $(t_0, t_1, ..., t_k, ...)$ have the form $(K, K + \Delta, K + 2\Delta, ..., K + k\Delta, ...)$:

$$S_4 = \sum_{k=1}^{\infty} (t_{k-1}^{-\alpha} - t_k^{-\alpha}) k = \sum_{k=1}^{\infty} (k t_{k-1}^{-\alpha} - k t_k^{-\alpha})$$

$$= (1t_0^{-\alpha} - 1t_1^{-\alpha}) + (2t_1^{-\alpha} - 2t_2^{-\alpha}) + (3t_2^{-\alpha} - 3t_3^{-\alpha}) + ... + (kt_{k-1}^{-\alpha} - kt_k^{-\alpha}) + ... = \sum_{k=0}^{\infty} t_k^{-\alpha}$$

$$= \sum_{k=0}^{\infty} (K + k\Delta)^{-\alpha} = \Delta^{-\alpha} \sum_{k=0}^{\infty} \left(k + \frac{K}{\Delta} \right)^{-\alpha} = \Delta^{-\alpha} \zeta(\alpha, K/\Delta), \quad (16)$$

where $\zeta(s, q)$ – Hurwitz's zeta function [8]. Series (16) meets integral Cauchy test for convergence, thus, sum S_4, and, consequently, sums S_3, S_2, average staying time (12) and average waiting time W (5) are finite.
The theorem is proved.

3. Derivation of calculation formula for RM optimization

Since the RM is uniquely determined by the parameter Δ the optimization problem of RM is stated as follows:

$$W(\Delta) \to \min_{\Delta \geq 0}. \quad (17)$$

When using any numerical methods to solve this problem, the main problem is calculation of values $W(\Delta)$ at the given step Δ. And since the developed queue normalization method for systems with infinite variance of service time is targeted at its application by engineers who design data communication networks, numerical method for calculation of $W(\Delta)$, available to engineers, had be developed. This problem will be solved in the next section.
From (5) and (12) we have the following for the original M/Pa/1 system at RM (2) determining the priority classes

$$W(\Delta) = \sum_{k=1}^{\infty} \frac{p_k b_k}{(1 - \sigma_{k-1})} + \sum_{k=1}^{\infty} \frac{p_k \sum_{i=1}^{k} \lambda_i b_i^{(2)}}{2(1 - \sigma_k)(1 - \sigma_{k-1})} - b, \quad (18)$$

where

$$p_k \sum_{i=1}^{k} \lambda_i b_i^{(2)} = \lambda p_k \sum_{i=1}^{k} p_i b_i^{(2)} = \lambda p_k \sum_{i=1}^{k} K^{\alpha}(t_{i-1}^{-\alpha} - t_i^{-\alpha}) \cdot \frac{\alpha}{\alpha - 2} \cdot \frac{t_{i-1}^{2-\alpha} - t_i^{2-\alpha}}{t_{i-1}^{-\alpha} - t_i^{-\alpha}} = \frac{\lambda \alpha K^{\alpha}}{\alpha - 2} p_k \sum_{i=1}^{k} (t_{i-1}^{2-\alpha} - t_i^{2-\alpha})$$

$$= \frac{\lambda\alpha K^{\alpha}}{\alpha-2} p_k (t_0^{2-\alpha} - t_1^{2-\alpha} + t_1^{2-\alpha} - t_2^{2-\alpha} + ... + t_{k-1}^{2-\alpha} - t_k^{2-\alpha}) = \frac{\lambda\alpha K^{\alpha}}{\alpha-2} p_k (K^{2-\alpha} - t_k^{2-\alpha}), \quad (19)$$

$$\sigma_k = \sum_{i=1}^{k} \rho_i = \sum_{i=1}^{k} \lambda_i b_i = \lambda \sum_{i=1}^{k} p_i b_i = \lambda \sum_{i=1}^{k} K^{\alpha} (t_{i-1}^{-\alpha} - t_i^{-\alpha}) \cdot \frac{\alpha}{\alpha-1} \cdot \frac{t_{i-1}^{1-\alpha} - t_i^{1-\alpha}}{t_{i-1}^{-\alpha} - t_i^{-\alpha}} = \frac{\lambda\alpha K^{\alpha}}{\alpha-1} \sum_{i=1}^{k} (t_{i-1}^{1-\alpha} - t_i^{1-\alpha})$$

$$= \lambda b K^{\alpha-1} (t_0^{1-\alpha} - t_1^{1-\alpha} + t_1^{1-\alpha} - t_2^{1-\alpha} + ... + t_{k-1}^{1-\alpha} - t_k^{1-\alpha}) = \rho K^{\alpha-1} (K^{1-\alpha} - t_k^{1-\alpha}) = \rho(1 - K^{\alpha-1} t_k^{1-\alpha}), \quad (20)$$

$$\sigma_{k-1} = \rho(1 - K^{\alpha-1} t_{k-1}^{1-\alpha}). \quad (21)$$

Taking into account (21), (6) and (8), the first sum in (18) we rewrite in the form:

$$S_1 = \sum_{k=1}^{\infty} \frac{p_k b_k}{(1-\sigma_{k-1})} = \sum_{k=1}^{\infty} \frac{\alpha K^{\alpha}}{\alpha-1} \cdot \frac{t_{k-1}^{1-\alpha} - t_k^{1-\alpha}}{1-\rho(1-K^{\alpha-1} t_{k-1}^{1-\alpha})} = \sum_{k=1}^{\infty} b K^{\alpha-1} \frac{t_{k-1}^{1-\alpha} - t_k^{1-\alpha}}{1-\rho(1-K^{\alpha-1} t_{k-1}^{1-\alpha})}$$

$$= b \sum_{k=1}^{\infty} \frac{[K + \Delta(k-1)]^{1-\alpha} - [K + k\Delta]^{1-\alpha}}{(1-\rho)K^{1-\alpha} + \rho [K + \Delta(k-1)]^{1-\alpha}} = b \sum_{k=1}^{\infty} \frac{(k+\gamma-1)^{1-\alpha} - (k+\gamma)^{1-\alpha}}{(1-\rho)\gamma^{1-\alpha} + \rho (k+\gamma-1)^{1-\alpha}}, \quad (22)$$

where $\gamma = \frac{K}{\Delta}$ is the dimensionless representation form of the step Δ.

The second sum in (18) with regard to (6), (9) and (19)–(21) we transform in a similar way:

$$S_2 = \sum_{k=1}^{\infty} \frac{p_k \sum_{i=1}^{k} \lambda_i b_i^{(2)}}{2(1-\sigma_k)(1-\sigma_{k-1})} = \sum_{k=1}^{\infty} \frac{\frac{\lambda\alpha K^{\alpha}}{\alpha-2} p_k (K^{2-\alpha} - t_k^{2-\alpha})}{2(1-\sigma_k)(1-\sigma_{k-1})}$$

$$= \frac{\lambda\alpha K^{2\alpha}}{2(2-\alpha)} \sum_{k=1}^{\infty} \frac{(t_{k-1}^{-\alpha} - t_k^{-\alpha})(t_k^{2-\alpha} - K^{2-\alpha})}{[1-\rho(1-K^{\alpha-1} t_k^{1-\alpha})] \cdot [1-\rho(1-K^{\alpha-1} t_{k-1}^{1-\alpha})]}$$

$$= \frac{\lambda\alpha K^{2\alpha}}{2(2-\alpha)} \sum_{k=1}^{\infty} \frac{[(K + \Delta(k-1))^{-\alpha} - (K + k\Delta)^{-\alpha}] \cdot [(K + k\Delta)^{2-\alpha} - K^{2-\alpha}]}{[(1-\rho) + \rho K^{\alpha-1} (K + k\Delta)^{1-\alpha})] \cdot [(1-\rho) + \rho K^{\alpha-1} (K + \Delta(k-1))^{1-\alpha})]}$$

$$= \frac{\lambda\alpha K^{2}}{2(2-\alpha)} \sum_{k=1}^{\infty} \frac{[(k+\gamma-1)^{-\alpha} - (k+\gamma)^{-\alpha}] \cdot [(k+\gamma)^{2-\alpha} - \gamma^{2-\alpha}]}{[(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma)^{1-\alpha})] \cdot [(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma-1)^{1-\alpha}]}. \quad (23)$$

Then, in general, the dependence $W(\Delta)$ defined by relation (18) can be rewritten in explicit form:

$$W(\gamma) = S_1 + S_2 - b = b \sum_{k=1}^{\infty} \frac{(k+\gamma-1)^{1-\alpha} - (k+\gamma)^{1-\alpha}}{(1-\rho)\gamma^{1-\alpha} + \rho (k+\gamma-1)^{1-\alpha}}$$

$$+ \frac{\lambda\alpha K^{2}}{2(2-\alpha)} \sum_{k=1}^{\infty} \frac{[(k+\gamma-1)^{-\alpha} - (k+\gamma)^{-\alpha}] \cdot [(k+\gamma)^{2-\alpha} - \gamma^{2-\alpha}]}{[(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma)^{1-\alpha})] \cdot [(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma-1)^{1-\alpha}]} - b, \ (\alpha \le 2), \quad (24)$$

where the sum S_2 at $\alpha = 2$ after taking the limit is expressed as:

$$S_2\big|_{\alpha=2} = \lim_{\alpha \to 2} S_2 = \lambda K^2 \sum_{k=1}^{\infty} \frac{[(k+\gamma-1)^{-2} - (k+\gamma)^{-2}] \cdot [\ln(k+\gamma) - \ln(\gamma)]}{[(1-\rho)\gamma^{-1} + \rho(k+\gamma)^{-1})] \cdot [(1-\rho)\gamma^{-1} + \rho(k+\gamma-1)^{-1}]}. \quad (24\,a)$$

4. Simplification of the calculation formula with error control

When calculating $W(\gamma)$ by the formulae (24), (24 *a*), sufficiently large number of initial members of series have to be summed because these series converge very slowly. For instance, at values α far enough from unity (for example, at $\alpha \geq 1.8$), one can still calculate $W(\gamma)$ with an accuracy of enough significant digits by summing up several million first members of the corresponding series belonging to (24), (24 *a*). But at values α close to unity (which sometimes come close to $\alpha = 1.2$ and even to $\alpha = 1.1$ while modeling actual traffic) these series converge so slowly that their calculation by means partial sums becomes unacceptably costly.

Example 1. Calculating a part of the sum S_1 (11) with $\lambda = 1$, $\alpha = 8/7$, $K = 0.0625$, $b = \rho = 0.5$, $\Delta = 0.2$, $\gamma = K/\Delta = 0.3125$ using large number of N first summands, we obtain the following partial sums $S_1(N)$ (rounded up to 6 significant digits): $S_1(10^6) = 0.576005$, $S_1(10^7) = 0.605963$, $S_1(10^8) = 0.628093$, $S_1(10^9) = 0.644327$, $S_1(10^{10}) = 0.656177$, $S_1(10^{11}) = 0.664792$, $S_1(10^{12}) = 0.671039$. Such powerful mathematical service as site wolframalpha.com cannot cope with calculation of sum $S_1 = S_1(\infty)$.

Therefore, for approximate calculation of W it is proposed to use partial sums $S_1(N)$, $S_2(N)$ (for some sufficiently large N) along with integral estimates $I_1(N+1)$, $I_2(N+1)$ for the remainders of the series:

$$W(\gamma) \approx S_1(N) + I_1(N+1) + S_2(N) + I_2(N+1) - b\,, \qquad (24\,b)$$

where $S_1(N) = b\sum_{k=1}^{N} \frac{(k+\gamma-1)^{1-\alpha} - (k+\gamma)^{1-\alpha}}{(1-\rho)\gamma^{1-\alpha} + \rho\,(k+\gamma-1)^{1-\alpha}}$, $I_1(N+1) = b\int_{N+1}^{\infty} \frac{(k+\gamma-1)^{1-\alpha} - (k+\gamma)^{1-\alpha}}{(1-\rho)\gamma^{1-\alpha} + \rho\,(k+\gamma-1)^{1-\alpha}}\,dk$,

$$S_2(N) = \frac{\lambda\alpha K^2}{2(2-\alpha)} \sum_{k=1}^{N} \frac{[(k+\gamma-1)^{-\alpha} - (k+\gamma)^{-\alpha}]\cdot[(k+\gamma)^{2-\alpha} - \gamma^{2-\alpha}]}{[(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma)^{1-\alpha})]\cdot[(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma-1)^{1-\alpha}]},$$

$$I_2(N+1) = \frac{\lambda\alpha K^2}{2(2-\alpha)} \int_{N+1}^{\infty} \frac{[(k+\gamma-1)^{-\alpha} - (k+\gamma)^{-\alpha}]\cdot[(k+\gamma)^{2-\alpha} - \gamma^{2-\alpha}]}{[(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma)^{1-\alpha})]\cdot[(1-\rho)\gamma^{1-\alpha} + \rho(k+\gamma-1)^{1-\alpha}]}\,dk\,.$$

For improper integral I_1 to be calculated by means of mathematical packages let us approximate the numerator of expression under integral sign applying to it a Taylor series expansion in powers of k:

$$(k+\gamma-1)^{1-\alpha} - (k+\gamma)^{1-\alpha} = \left[k^{1-\alpha} - (\gamma-1)(\alpha-1)k^{-\alpha} + 0.5(\gamma-1)^2\alpha(\alpha-1)k^{-\alpha-1} - \ldots\right]$$
$$-\left[k^{1-\alpha} - \gamma(\alpha-1)k^{-\alpha} + 0.5\gamma^2\alpha(\alpha-1)k^{-\alpha-1} - \ldots\right] \approx (\alpha-1)k^{-\alpha}\,. \qquad (25)$$

Example 2. Performing a routine analysis of those errors in calculating the sum S_1 that are caused by substituting part of the sum S_1 with improper integral I_1 and substituting the numerator of integral function with expression (25), we determine that sum $S_1(10^8) + I_1(10^8+1)$ coincides with sum S_1 with an accuracy at least 4 significant digits. Calculating it with a help of wolframalpha.com, we found that:

$$S_1 \approx S_1(10^8) + I_1(10^8+1) = b\sum_{k=1}^{1000000} \frac{(k+\gamma-1)^{1-\alpha} - (k+\gamma)^{1-\alpha}}{(1-\rho)\gamma^{1-\alpha} + \rho\,(k+\gamma-1)^{1-\alpha}}$$
$$+\,b\int_{1000001}^{\infty} \frac{(\alpha-1)k^{-\alpha}}{(1-\rho)\gamma^{1-\alpha} + \rho\,(k+\gamma-1)^{1-\alpha}}\,dk \approx 0.687258\,.$$

This example shows that partial sum $S_1(10^{12}) = 0.671039$ calculated in Example 1 is a bad approximation of sum S_1.

Similarly, in the numerator of the second integral we substitute the difference $[(k+\gamma-1)^{-\alpha} - (k+\gamma)^{-\alpha}]$ with value $\alpha k^{-\alpha-1}$.

Moreover, in the numerator of the second integral we substitute the second multiplier by the first one since their difference is negligible: it is the difference of two power functions that can be easily estimated by expanding it into a Taylor series.

Therefore, average waiting time can be calculated with controlled errors by the formula:

$$W(\gamma) \approx b\sum_{k=1}^{N}\frac{(k+\gamma-1)^{1-\alpha}-(k+\gamma)^{1-\alpha}}{(1-\rho)\gamma^{1-\alpha}+\rho\,(k+\gamma-1)^{1-\alpha}} + b\int_{N+1}^{\infty}\frac{(\alpha-1)k^{-\alpha}}{(1-\rho)\gamma^{1-\alpha}+\rho\,(k+\gamma-1)^{1-\alpha}}dk$$

$$+\frac{\lambda\alpha K^2}{2(2-\alpha)}\sum_{k=1}^{N}\frac{[(k+\gamma-1)^{-\alpha}-(k+\gamma)^{-\alpha}]\cdot[(k+\gamma)^{2-\alpha}-\gamma^{2-\alpha}]}{[(1-\rho)\gamma^{1-\alpha}+\rho(k+\gamma)^{1-\alpha})]\cdot[(1-\rho)\gamma^{1-\alpha}+\rho(k+\gamma-1)^{1-\alpha}]}$$

$$+\frac{\lambda\alpha^2K^2}{2(2-\alpha)}\int_{N+1}^{\infty}\frac{(k+\gamma)^{2-\alpha}k^{-1-\alpha}-\gamma^{2-\alpha}k^{-1-\alpha}}{[(1-\rho)\gamma^{1-\alpha}+\rho(k+\gamma)^{1-\alpha})]^2}dk - b\,. \qquad (26)$$

The simplest way to control errors obtained using formula (26) is the test calculations with gradual increase of the number N of summands in the partial sums S_1 и S_2 because these errors decrease quite quickly with the growth of N.

When γ are large, i.e. at small values Δ with respect to K, for W to be calculated one may use the exact formula of limit:

$$\lim_{\Delta\to 0} W(\Delta) = \alpha K^{\alpha}\int_{K}^{\infty}\frac{t^{-\alpha}}{\left[1-\frac{\lambda\alpha K^{\alpha}}{\alpha-1}\left(K^{1-\alpha}-t^{1-\alpha}\right)\right]}dt + \frac{\lambda\alpha^2K^{2\alpha}}{2(2-\alpha)}\int_{K}^{\infty}\frac{t^{-1-\alpha}\left(t^{2-\alpha}-K^{2-\alpha}\right)}{\left[1-\frac{\lambda\alpha K^{\alpha}}{\alpha-1}\left(K^{1-\alpha}-t^{1-\alpha}\right)\right]^2}dt - b\,, \qquad (27)$$

obtained from (5) and (12) taking the limit.

5. Example of optimization problem solving for RM

Table 1 shows the calculation of the dependence $W(\Delta)$ according to formulae (26), (27) with parameters stated in Example 1 for the original system M/Pa/1.

Table 1. The calculation of $W(\Delta)$ by parts of sum (26).

N	Δ	γ	$S_1(N)$	$I_1(N+1)$	$S_2(N)$	$I_2(N+1)$	W
–	0	–	–	–	–	–	0.206297
10^{10}	0.01	6.25	0.645499	0.047295	0.013239	2.29E-05	0.206056
10^{9}	0.02	3.125	0.633278	0.059165	0.013379	3.58E-05	0.205858
10^{8}	0.05	1.25	0.619756	0.071669	0.013938	5.22E-05	0.205415
	0.1	0.625	0.624742	0.065127	0.015142	4.32E-05	0.205055
	0.15	0.416667	0.626916	0.061573	0.016432	3.87E-05	0.204960
	0.2	0.3125	0.628093	0.059165	0.017778	3.58E-05	0.205072
	0.3	0.208333	0.629215	0.055920	0.020566	3.20E-05	0.205733
	0.4	0.15625	0.629632	0.053734	0.023422	2.95E-05	0.206817
10^{6}	0.5	0.125	0.583584	0.098250	0.026030	9.73E-05	0.207961
	1	0.0625	0.586893	0.089389	0.040552	8.08E-05	0.216915
	1.5	0.041667	0.587976	0.084565	0.054862	7.24E-05	0.227475
	10	0.00625	0.586728	0.065127	26.41610	4.32E-05	26.56800
	100	0.000625	0.576914	0.047295	206.8180	2.29E-05	206.9422
	1000	6.25E-05	0.563815	0.034262	1604.210	1.21E-05	1604.308

The figure 1 depicts the part of the dependence $W(\Delta)$ in a sufficiently small neighbourhood of the optimal step Δ_{opt}.

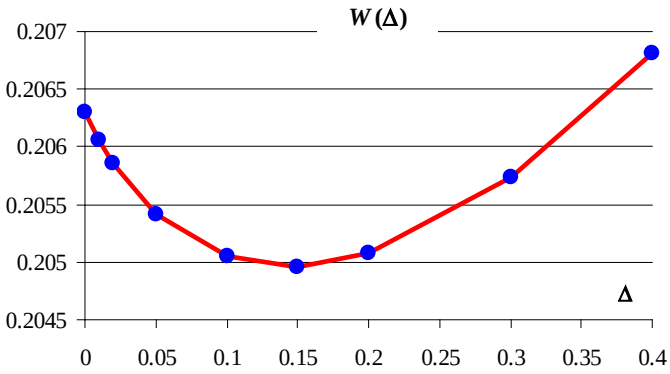

Figure 1. Calculated dependence $W(\Delta)$.

As one can see in figure 1, $\Delta_{opt} \approx 0.15$. In the neighbourhood of the optimum the line graph is slightly sloping, and if one overestimates or underestimates value Δ_{opt} even twofold, the time W will not change significantly.
Thus, introducing absolute priorities with infinite optimal RM in the system M/Pa/1 reduces the average waiting time and average queue length from infinity to 0.2 in considered example.

6. Exponential marking

Numerical experiment with a large number of other markings differing from a regular one showed that the most efficient marking is infinite exponential $(t_0, t_1, ..., t_k, ...)$, in which points t_k are distributed as follows:

$$t_0 = K, \quad t_k = K + ce^{ak}, \qquad (k = 1, 2, ...), \tag{28}$$

where c, a are coefficients that can be optimized to minimize the average waiting time W at given λ, K, α.
In addition to this exponential marking (28) being able to reduce the time W to finite values, its advantage is that due to the rapid growth of the interval lengths between the points t_k the partial sums in (18), which are calculated with only several tens of initial summands, coincide with corresponding sums of infinite series with an accuracy at enough significant digits. Because of this, the calculation of W according to formula (18) can be performed in a few rows of Excel spreadsheets, and at the same time it is possible to perform precise optimization of the marking parameters c, a with the gradient method built into add-in 'Solver'.
For such optimization it is required several seconds of computer time. Experiments with optimal exponential markings conclude that optimal exponential markings are as good as optimal regular markings in terms of reducing W.
A small number of practically realizable infinite exponential marking levels (28) are also distinguish beneficially it from RM for the practical implementation of the infinite markings method.

7. The loss probabilities reducing at finite buffer

The developed in the article normalization method of «fractal» queues allows a radical reduction of the loss probability in queueing systems with heavy tailed distributions and finite buffer. The developed method consists in introducing absolute priorities defined by infinite markings in systems with a finite buffer. It takes into account the results of previously published empirical studies of systems operating under fractal traffic [9–11] and achievements for the classical theory of priority disciplines optimization [6, 7, 12, 13]. However, the proposed method differs significantly from well-known ones by considering of infinite markings. It is proposed to optimize the parameters of the used

markings by numerical methods based on the exact formulae of the queueing theory. We optimize the markings under condition that a system has an infinite buffer. And the subsequent buffer limit occurs under conditions where the unlimited queue becomes on average as short as possible, and therefore the loss probability should be reduced as much as possible. This heuristic justification of the developed method also explains the independence of the selected marking from the buffer length and, therefore, indirectly justifies the calculation of the rapid decrease in the loss probability with the increasing buffer size.

Figure 2 shows the results of simulation experiments which were carried out to compare the efficiency of the proposed method and the usually recommended method of reducing the loss probability, consisting in a simple increase in the buffer size without introducing absolute priorities. Simulation experiments were carried out with the system M/Pa/1, in which $\alpha = 1.5$, $\rho = 0.5$. The optimal value of Δ for the corresponding regular marking, determined up to two significant decimal digits, is 0.30. The optimal values of the exponential marking parameters are as follows: $a = 1.055$, $c = 0.08424$. The horizontal coordinate axis corresponds to the buffer size m, and the vertical axis corresponds to the loss probability P. In all experiments 10 million requests were passed through the system.

A continuous, sharply downward, red line is the result of the buffer build-up, using absolute priorities due to optimal exponential marking. The line converges to a straight line that at the logarithmic scale of the ordinate axis indicates that the dependence $P(m)$ represented by this line is asymptotically exponential.

The continuous green line at the top of the diagram is calculated by simulating the initial system that does not use priorities. Since the corresponding dependence is power-law [6], here the line goes down with deceleration. In simulation experiments at the buffer size $m = 100$, the probability P decreases only to 0.029, at $m = 1000$ we get $P = 0.00764$, and at $m = 10\,000$ we get $P = 0.00125$. As you can see from Figure 2, when using absolute priorities, such loss probability is achieved already at the buffer size $m = 5$. Moreover, using buffers designed to store 10,000 packets in real network devices is sure to make no sense because of the high cost of the corresponding equipment and the large delays occurring in the corresponding queues.

The markers on the top line indicate the results obtained when using relative priorities (with the same set of priority classes as with absolute priorities).

It is easy to explain such high efficiency of the absolute priorities introducing. At infinite variance of service time x, the system occasionally receives requests with very high, "catastrophic" service time. If such a request enters the system and occupies a channel in non-priority mode or relative priority mode, a long queue is created at the system input during its service, which leads to an overflow of even a very large buffer. In the mode of absolute priorities, incoming "non-catastrophic" requests with a higher priority "do not notice" a catastrophic request and simply push it out into the queue, consequently they are served as if there were no catastrophic request. As a result, long queues accumulating does not occur.

This reasoning make us suppose that the proposed method will be as effective in any systems GI/GI/1/m, where service time x distribution is a heavy tailed distribution (HTD) with infinites (or simply large) variance.

This assumption is confirmed by simulation experiments with a sufficiently large number of such systems. As an example figure 7 shows the results of modeling a system in which the arrival flow is set by the gamma distribution of the intervals of request arrivals and the service time x has a lognormal distribution. The gamma distribution has a m.e. 1 and a relatively high variance of 16 (the parameters β and α of the distribution are chosen to be 16 and 1/16, respectively). Lognormal distribution has the parameters $\mu = -10$, $\sigma = 4.3$. At such parameters of the lognormal distribution, for the service time x we get $M(x) = b = 0.47$ $M(x) = b = 0.47$, $D(x) = 5.938 \cdot 10^{42}$.

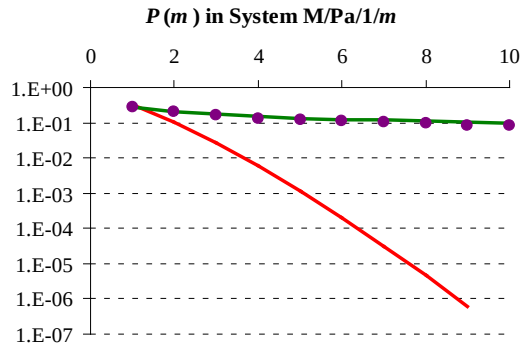

Figure 2. Dependences $P(m)$ in non-priority and priority modes of system M/Pa/1/m.

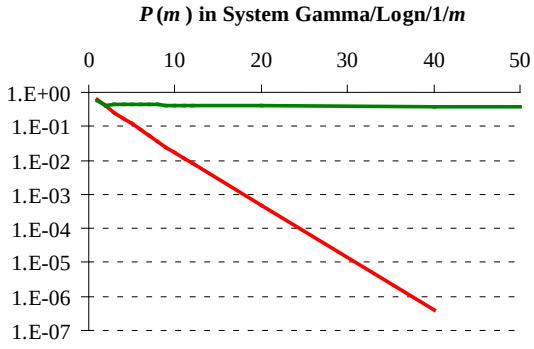

Figure 3. Dependences $P(m)$ in system Gamma/Logn/1/m.

The upper almost horizontal line in figure 3 shows the dependence $P(m)$ in non-priority mode: P decreases very slowly with the growth of m. At m = 10 000, the probability P = 0.25436 was obtained, at m = 100 000, the probability P = 0.234502. In figure 3 this sequence of probabilities begins with values of the order 0.5...0.4.

The lower line in the diagram corresponds to the dependence $P(m)$ obtained in the experiment when using absolute priorities with infinite marking of complexities, i.e. the method developed in this article.

Note that to implement the proposed queue normalization method, it is not required additional hardware costs, a rather small modification of the network software is sufficient.

In addition, it should be noted that in the packet switching networks, applying relative priorities to packets, which are defined by considered methods in accordance with a size of the file divided into packets, is almost equivalent to the absolute priorities introducing for files

8. Conclusions

To solve the problems, which in practice result from an infinite variance Var(x), the article develops a method based on introducing absolute priorities for requests with an infinite number of priority classes.

In the course of the research carried out in the article, the following main results were obtained.

1) The regular marking, in which all intervals $[t_{k-1}, t_k)$ have the same length Δ, of semi-infinite range of service time values is investigated. The theorem is formulated and proved that if the absolute priorities determined by the regular marking are introduced into the system M/Pa/1/∞ with infinite average waiting time, its average waiting time W becomes finite.

2) The problem of optimizing regular marking by criterion $W \to \min$ is set and solved by numerical methods. In the framework of this problem, the calculation formula to determine the average waiting time W is obtained at $\Delta \to 0$.

3) Among a large number of irregular markings, the most effective one is found, i.e. the exponential marking, in which the length of successive intervals grows as an exponent with two constant coefficients. It is shown that by reducing the average waiting time W an exponential marking is not inferior to regular marking. At the same time, exponential marking is much more economical: at $1 < \alpha \leq 2$, when tens of millions of priority classes are implemented by the optimal regular marking, the optimal exponential marking requires the implementation no more than 50 priority levels.

4). The efficiency of applying the developed method to the systems M/Pa/1/m with a finite buffer and Var(x) = ∞ is studied. It is established that the introduction of absolute priorities determined by the considered infinite markings drastically reduces the loss probability. In this case, the slow decrease (with power speed) of $P(m)$ turns into a rapid one occurring at an exponential speed.

5). We put the hypothesis and founded it at the level of physical sense, that the method will also be equally effective in other systems GI/GI/1/m with infinite or very large variance Var(x).

In order to verify this hypothesis, a series of experiments with such systems was carried out, and the experiments have convincingly confirmed the proposed hypothesis.

9. References

[1] Leland W E, Willinger W, Taqqu M S and D V Wilson 1993 On the Self-Similar Nature of Ethernet Traffic, *Proc. of SIGCOMM '93* September 13–17 San Francisco vol 23 ed D Oran (New York: ACM) pp 183-193
[2] Paxon V and Floyd S 1995 Wide area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on networking* **3** 226-44
[3] Park K and Willinger W 2000 *Self-Similar Network Traffic and Performance Evaluation* (New York: Wiley-Interscience) pp 558
[4] Erramilli A, Narayan O and Willinger W 1996 Experimental queueing analysis with long range dependent packet traffic, *IEEE/ACM Transactions on networking* **4** 209-23
[5] Zadorozhnyi V N and Zakharenkova T R 2017 Minimization of Packet Loss Probability in Network with Fractal Traffic *Proc. of Information Technologies and Mathematical Modelling* September 29 – October 3 Kazan vol 800 ed A Dudin et al (Cham: Springer) pp 168-83
[6] Likhanov N, Tsybakov B, Georganas N 1995 Analysis of an ATM Buffer with Self-Similar («Fractal») Input Traffic *Proc. of INFOCOM '95* 2-6 April Boston vol 3 (IEEE publisher) pp 985–992
[7] Kleinrock L 1976 *Queueing Systems Volume* II: *Computer Applications* (New York: Wiley Interscience) p 400
[8] Adamchik V S 1998 *Some Series of the Zeta and Related Functions* Analysis 18 pp 131-144
[9] Zadorozhnyi V N, Zakharenkova T R, Tulubaev D A 2018 Estimation of Prioritized Disciplines Efficiency Based on the Metamodel of Multi-flows Queueing Systems *Proc. of Information Technologies and Mathematical Modelling* September 10 –15 Tomsk vol **912** ed A Dudin et al (Cham: Springer) pp 290-304
[10] Neame T, Zukerman M, Addie R 1999 *A practical approach for multi-media traffic modeling* Proc. of Broadband Communications 10–12 November, Hong Kong pp 73-82
[11] Zadorozhnyi V N 2016 *Queues and power distributions* (Novosibirsk: SB RAS) p 161 (In Russian)
[12] Bronstein O I, Rykov V V 1965 *On optimal priority rules in queueing system* Izv. AS USSR Techn. Cibern 6 pp 28–37 (in Russian)