ISSN 0032-9460, Problems of Information Transmission, 2018, Vol. 54, No. 2, pp. 176–190. © Pleiades Publishing, Inc., 2018.
Original Russian Text © A.V. Lebedev, 2018, published in Problemy Peredachi Informatsii, 2018, Vol. 54, No. 2, pp. 86–102.

Maximum Remaining Service Time in Infinite-Server Queues

A. V. Lebedev

*Department of Probability Theory, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, Moscow, Russia*
e-mail: avlebed@yandex.ru

Received April 13, 2017; in final form, July 21, 2017

Abstract—We study the maximum remaining service time in infinite-server queues of type $M|G|\infty$ (at a given time and in a stationary regime). The following cases for the arrival flow rate are considered: (1) time-independent, (2) given by a function of time, (3) given by a random process. As examples of service time distributions, we consider exponential, hyperexponential, Pareto, and uniform distributions. In the case of a constant rate, we study effects that arise when the average service time is infinite (for power-law distribution tails). We find the extremal index of the sequence of maximum remaining service times. The results are extended to queues of type $M^X|G|\infty$, including those with dependent service times within a batch.

DOI: 10.1134/S0032946018020060

1. INTRODUCTION

The classical infinite-server (infinite-line) queueing system $M|G|\infty$ is a queue with a Poisson arrival flow of rate λ and with infinitely many servers. Every arrival immediately gets service at one of free servers. Arrival times are independent and identically distributed. The service time distribution will be denoted by B.

Such queues have been studied since the paper [1]. Basic results for such queues can be found in textbooks [2, 3]. Further generalizations mostly concerned more complicated arrival flows. A vast bibliography is contained in [4, 5]. The author has previously studied batch arrival queues $M^X|G|\infty$ from the point of view of the maximum number of customers [4, 6, 7]. Now we consider another characteristic.

Assume that at initial time the system is free of customers. The maximum remaining service time W_T is understood to be the maximum of remaining service times over all servers that are busy at time T. From the practical point of view, this is the time required to correctly complete the system operation after terminating the arrival flow (for example, in computer data processing or in telecommunication data transmission). Denote the distribution of the random variable W_T by G_T, and the limit distribution as $T \to \infty$ (if exists), by G_∞.

The maximum remaining service time in infinite-server queues with doubly stochastic Poisson arrivals was considered in [8, 9]. Namely, it was assumed that the service time is distributed exponentially and the arrival flow rate switches between two values, λ_1 and λ_2, in exponentially distributed time intervals with parameters α_1 and α_2. Recursive formulas were derived that allow to estimate numerically the distribution density.

Below we consider the following cases for the arrival flow rate: (1) time-independent, (2) given by a function of time, (3) given by a random process. In the last case, concrete results are obtained for a stationary Gaussian process.

As for examples of service time distributions, we pay special attention to hyperexponential distributions [3, ch. 2, Section 8; 10; 11] and heavy-tail distributions (with power-law tails).

Note also that for the infinite mean service time (with a heavy enough tail), even in the case of constant service rate, the number of customers in the queue tends to infinity, and thus also the maximum remaining service time does, which therefore requires a normalization for convergence in distribution. Infinite-server queues with infinite mean service time were studied in [5, 12].

For a doubly stochastic Poisson arrival flow, we consider the case where the rate is described by a stationary Gaussian process. This approximation is widely used in telecommunication traffic models [13], in particular for modeling self-similarity and long-term dependence. Furthermore, such a rate can occur in the limit where the arrival flow is determined by the number of customers in another $M|G|\infty$ system with high load [14, ch. 2, Section 2].

In the case of a constant arrival rate, we find the extremal index of the sequence of maximum remaining service times. This special subject belongs to stochastic extreme value theory [15–17].

All results are obtained for an ordinary arrival flow but can be extended to $M^X|G|\infty$ batch arrival queues under the assumption that batch sizes are independent and service times can be dependent only within a batch. The maximum remaining service time for a batch is determined only by the current duration of its stay in the system and by the maximum service time of customers in the batch; we denote the distribution of the latter by B^*. Then all the obtained results on the maximum remaining service time for the $M|G|\infty$ queue with distribution B^* are valid for the $M^X|G|\infty$ queue (simply speaking, we may merely join all customers of a batch into a single new customer). Thus, the question reduces to finding B^*. This distribution is derived for both independent and some certain dependent service times.

2. THE CASE OF A TIME-INDEPENDENT RATE

First we consider the case of a time-independent arrival rate (we have a stationary Poisson flow). Let the rate be λ.

Theorem 1. *For* $\lambda = \text{const}$, $0 < T < \infty$, *we have*

$$G_T(x) = \exp\left\{-\lambda \int\limits_0^T \bar{B}(t+x)\,dt\right\},$$

where $\bar{B}(x) = 1 - B(x)$.

This formula can also be rewritten as

$$G_T(x) = \exp\left\{-\lambda \int\limits_x^{T+x} \bar{B}(t)\,dt\right\}.$$

Clearly, if there exists a finite mean service time

$$\mu = \int\limits_0^\infty \bar{B}(x)\,dx,$$

then there also exists a limit distribution

$$G_\infty(x) = \exp\left\{-\lambda \int\limits_x^\infty \bar{B}(t)\,dt\right\}.$$

It is convenient to introduce new distribution functions

$$F_T(x) = 1 - \frac{1}{\mu_T} \int\limits_0^T \bar{B}(t+x)\,dt, \quad \mu_T = \int\limits_0^T \bar{B}(t)\,dt, \quad T > 0;$$

then

$$G_T(x) = \exp\{-\lambda \mu_T \bar{F}_T(x)\},$$

in particular for $T = +\infty$ if $\mu < \infty$.

Example 1. Let the service time have an exponential distribution

$$B(x) = 1 - e^{-\beta x}, \quad x \geq 0, \quad \beta > 0; \tag{1}$$

then

$$G_T(x) = \exp\{-A(T)e^{-\beta x}\}, \quad x \geq 0, \quad 0 < T \leq \infty,$$

where

$$A(T) = \frac{\lambda}{\beta}(1 - e^{-\beta T}), \qquad A(\infty) = \frac{\lambda}{\beta}.$$

In fact, we obtain the Gumbel distribution truncated at zero, since there is positive probability $e^{-A(T)}$ that the system is free.

Example 2. Let the service time have a hyperexponential distribution

$$B(x) = 1 - \sum_{i=1}^n c_i e^{-\beta_i x}, \quad x \geq 0, \quad \beta_i > 0, \quad \sum_{i=1}^n c_i = 1, \quad \sum_{i=1}^n c_i \beta_i \geq 0; \tag{2}$$

then

$$G_T(x) = \exp\biggl\{-\sum_{i=1}^n A_i(T) c_i e^{-\beta_i x}\biggr\}, \quad x \geq 0, \quad 0 < T \leq \infty,$$

where

$$A_i(T) = \frac{\lambda}{\beta_i}(1 - e^{-\beta_i T}), \qquad A_i(\infty) = \frac{\lambda}{\beta_i}.$$

Example 3. Let the service time have a Pareto distribution

$$B(x) = 1 - (x+1)^{-\alpha}, \quad x \geq 0, \quad c > 0, \quad \alpha > 1; \tag{3}$$

then

$$G_\infty(x) = \exp\biggl\{-\frac{\lambda}{\alpha - 1}(x+1)^{-(\alpha-1)}\biggr\}, \quad x \geq 0,$$

which is a shifted Fréchet distribution truncated at zero.

Now let us prove a general limit theorem under high load conditions.

Theorem 2. *Let F_T belong to the domain of attraction of a maximum stable distribution H, i.e., there exist normalizing constants $a(s) > 0$ and $b(s)$, $s > 0$, such that*

$$F_T^s(a(s)x + b(s)) \to H(x), \quad s \to \infty; \tag{4}$$

then

$$G_T(a(\lambda)x + b(\lambda)) \to H^{\mu_T}(x), \quad \lambda \to \infty;$$

i.e., we have convergence to a distribution of the same extremal type.

The theorem is in particular valid for $T = +\infty$ if $\mu < \infty$.

For more details on maximum stable distributions, extremal types, and attraction domains, see [15–17].

Example 4. Let the service time have an exponential distribution (1); then

$$G_\infty\left(\frac{x+\ln\lambda}{\beta}\right) \to \exp\left\{-\frac{e^{-x}}{\beta}\right\}, \quad x \in \mathbb{R}, \quad \lambda \to \infty;$$

i.e., we have convergence to a Gumbel distribution.

Example 5. Let the service time have a hyperexponential distribution (2), and let $k = \arg\min \beta_i$; then $\bar{B}(x) \sim c_k e^{-\beta_k x}$, $x \to \infty$, and

$$G_\infty\left(\frac{x+\ln(\lambda c_k)}{\beta_k}\right) \to \exp\left\{-\frac{e^{-x}}{\beta_k}\right\}, \quad x \in \mathbb{R}, \quad \lambda \to \infty;$$

i.e., we have convergence to a Gumbel distribution.

Example 6. Let the service time have a Pareto distribution (3); then

$$G_\infty\left(\lambda^{1/(\alpha-1)}x\right) \to \exp\left\{-\frac{x^{-(\alpha-1)}}{\alpha-1}\right\}, \quad x > 0, \quad \lambda \to \infty;$$

i.e., we have convergence to a Fréchet distribution.

Example 7. Let the service time be uniformly distributed on $[0,1]$: $B(x) = x$, $x \in [0,1]$; then

$$G_\infty(x) = \exp\left\{-\frac{\lambda}{2}(1-x)^2\right\}, \quad x \in [0,1],$$

whence

$$G_\infty\left(\frac{x}{\sqrt{\lambda}} + 1\right) = e^{-x^2/2}, \quad x < 0;$$

i.e., we obtain a Weibull distribution on the negative semiaxis (here passing to the limit is even not needed).

Now let us pass to the case with infinite mean service time (for power-law tails).

Theorem 3. *Let*

$$\bar{B}(x) \sim cx^{-\alpha}, \quad x \to \infty, \quad c > 0, \quad \alpha \in (0,1];$$

then for $0 < \alpha < 1$ *we have*

$$G_T(T^{1/\alpha}x) \to \exp\{-c\lambda x^{-\alpha}\}, \quad x > 0, \quad T \to \infty, \tag{5}$$

and for $\alpha = 1$,

$$G_T(Tx) \to \left(\frac{x}{x+1}\right)^{c\lambda}, \quad x > 0, \quad T \to \infty. \tag{6}$$

3. THE CASE OF A RATE GIVEN BY A FUNCTION OF TIME

Let the arrival flow rate be given by a bounded nonnegative function $\lambda(t)$, $t \geq 0$ (we have a nonstationary Poisson flow). Then Theorem 1 can be generalized to the following.

Theorem 4. *For* $\lambda = \lambda(t)$, $0 < T < \infty$, *we have*

$$G_T(x) = \exp\left\{-\int\limits_0^T \lambda(t)\bar{B}(T-t+x)\,dt\right\}.$$

It is convenient to rewrite this formula in an equivalent form

$$G_T(x) = \exp\left\{-\int\limits_0^T \lambda(T-t)\bar{B}(t+x)\,dt\right\}.$$

Example 8. Let the service time have an exponential distribution (1); then

$$G_T(x) = \exp\{-A(T)e^{-\beta x}\}, \quad x \ge 0, \quad 0 < T < \infty,$$

where

$$A(T) = \int\limits_0^T \lambda(T-t)e^{-\beta t}\,dt.$$

Example 9. Let the service time have a hyperexponential distribution (2); then

$$G_T(x) = \exp\left\{-\sum_{i=1}^n A_i(T)c_i e^{-\beta_i x}\right\}, \quad x \ge 0, \quad 0 < T < \infty,$$

where

$$A_i(T) = \int\limits_0^T \lambda(T-t)e^{-\beta_i t}\,dt.$$

It is clear that even for a finite mean service time, it is not for every function $\lambda(t)$ that G_∞ exists. For instance, for a periodic arrival flow rate the distribution G_T also oscillates. However, one can consider its partial limits. Models with periodic arrival flows have been studied for long in queueing theory [18]. Also, in [19] there were introduced rates given by periodic random processes (in the sense of periodicity of finite-dimensional distributions in time). Papers [20, 21] considered sinusoidal rates, as in the following example.

Example 10. Let $\lambda(t) = \lambda_0 + \varepsilon \sin \omega t$, $\lambda_0, \omega > 0$, $|\varepsilon| < \lambda_0$; then in the case of an exponential service time distribution (1) we obtain

$$A(T) = \frac{\lambda}{\beta}(1 - e^{-\beta T}) + \varepsilon \frac{\beta \sin \omega T - \omega \cos \omega T + \omega e^{-\beta T}}{\omega^2 + \beta^2},$$

whence for $T = 2N\pi/\omega$ (at endpoints of periods) we get

$$A(T) \to \frac{\lambda}{\beta} - \varepsilon \frac{\omega}{\omega^2 + \beta^2}, \quad N \to \infty,$$

and for $T = (2N+1)\pi/\omega$ (at midpoints of periods) we have

$$A(T) \to \frac{\lambda}{\beta} + \varepsilon \frac{\omega}{\omega^2 + \beta^2}, \quad N \to \infty.$$

Similarly one can compute $A_i(T)$ for a hyperexponential distribution.

4. THE CASE OF A RATE GIVEN BY A RANDOM PROCESS

Let the arrival flow rate be given by a random process $\lambda(t)$, $t \ge 0$ (we have a doubly stochastic Poisson flow). Then Theorem 4 can be generalized to the following.

Theorem 5. *For $\lambda = \lambda(t)$, $0 < T < \infty$, we have*

$$G_T(x) = \mathbf{E}\exp\biggl\{-\int\limits_0^T \lambda(t)\bar{B}(T-t+x)\,dt\biggr\}.$$

Theorem 5 is proved based on Theorem 4 by simply averaging over all possible trajectories of the process $\lambda(t)$.

Assume that the rate is of the form

$$\lambda(t) = \lambda_0 + \sigma\xi(t), \quad \lambda_0, \sigma > 0,$$

where $\xi(t)$, $t \ge 0$, is a stationary zero-mean Gaussian process with covariance function $C(t)$, $t \ge 0$; $C(0) = 1$.

Strictly speaking, this is impossible, since in this case $\lambda(t)$ may take negative values. However, the probability of such values and their contribution to the result for $\sigma \ll \lambda_0$ will be very small.[1] In what follows, when writing approximate equalities, we mean that they are the more accurate the less the ratio σ/λ_0 and that they are upper estimates for the functions $G_T(x)$ and $G_\infty(x)$ (and, correspondingly, stochastic estimates from below for the distributions).

Corollary. *Under the above assumptions we have*

$$G_T(x) \approx \exp\biggl\{-\lambda_0\int\limits_0^T \bar{B}(t+x)\,dt + \frac{\sigma^2}{2}\int\limits_0^T\int\limits_0^T C(u-v)\bar{B}(u+x)\bar{B}(v+x)\,du\,dv\biggr\}$$

and

$$G_T(x) \le \exp\biggl\{-\lambda_0\int\limits_0^T \bar{B}(t+x)\,dt + \frac{\sigma^2}{2}\int\limits_0^T\int\limits_0^T C(u-v)\bar{B}(u+x)\bar{B}(v+x)\,du\,dv\biggr\}.$$

This obviously implies

$$G_\infty(x) \approx \exp\biggl\{-\lambda_0\int\limits_0^\infty \bar{B}(t+x)\,dt + \frac{\sigma^2}{2}\int\limits_0^\infty\int\limits_0^\infty C(u-v)\bar{B}(u+x)\bar{B}(v+x)\,du\,dv\biggr\},$$

or

$$G_\infty(x) \approx \exp\biggl\{-\lambda_0\int\limits_x^\infty \bar{B}(t)\,dt + \frac{\sigma^2}{2}\int\limits_x^\infty\int\limits_x^\infty C(u-v)\bar{B}(u)\bar{B}(v)\,du\,dv\biggr\}$$

if all the integrals converge.

Example 11. Let the service time have exponential distribution (1); then

$$G_T(x) \approx \exp\left\{-A(T)e^{-\beta x} + D(T)e^{-2\beta x}\right\}, \quad x \ge 0,$$

where

$$A(T) = \frac{\lambda_0}{\beta}(1 - e^{-\beta T}), \qquad D(T) = \frac{\sigma^2}{2}\int\limits_0^T\int\limits_0^T C(u-v)e^{-\beta(u+v)}\,du\,dv.$$

[1] For details, see the remark to the proof of the corollary.

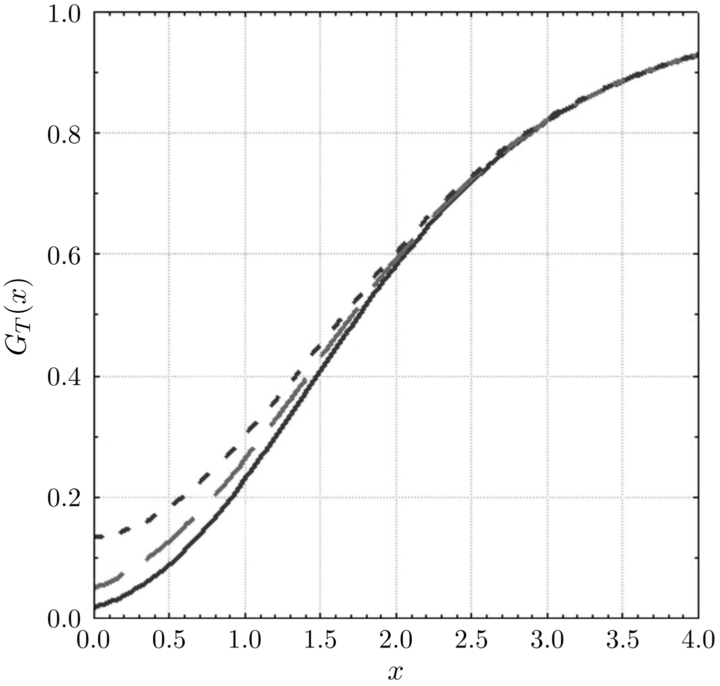

Fig. 1. Plot of the function $G_T(x)$ to Example 11.

This formula has a probabilistic sense only if $G_T(x)$ happens to be a nondecreasing function, which yields the conditions $A(T) \geq 2D(T)$.

In Fig. 1 we present a plot of the function for $\beta = 1$, $A(T) = 4$, and $D(T) = 0, 1, 2$ (from bottom to top). It is seen that discrepancies rapidly decay.

Let, for example, $C(t) = e^{-\delta|t|}$; then

$$\int_0^\infty \int_0^\infty e^{-\delta|u-v|-\beta(u+v)}\, du\, dv = \frac{1}{\beta(\delta+\beta)},$$

whence

$$D(\infty) = \frac{\sigma^2}{2\beta(\delta+\beta)}.$$

Example 12. Let the service time have hyperexponential distribution (2); then

$$G_T(x) \approx \exp\left\{-\sum_{i=1}^n A_i(T) c_i e^{-\beta_i x} + \sum_{i=1}^n \sum_{j=1}^n D_{ij}(T) c_i c_j e^{-(\beta_i+\beta_j)x}\right\}, \quad x \geq 0,$$

where

$$A_i(T) = \frac{\lambda_0}{\beta_i}(1 - e^{-\beta_i T}), \qquad D_{ij}(T) = \frac{\sigma^2}{2} \int_0^T \int_0^T C(u-v) e^{-(\beta_i u + \beta_j v)}\, du\, dv.$$

This formula has a probabilistic sense only if $G_T(x)$ happens to be a nondecreasing function, which yields the condition

$$\sum_{i=1}^n A_i(T) c_i \beta_i \geq \sum_{i=1}^n \sum_{j=1}^n D_{ij}(T) c_i c_j (\beta_i + \beta_j).$$

If $C(t)=e^{-\delta|t|}$, then

$$\int\limits_0^\infty\int\limits_0^\infty e^{-\delta|u-v|-(\beta_i u+\beta_j v)}\,du\,dv=\frac{1}{\beta_i+\beta_j}\left(\frac{1}{\delta+\beta_i}+\frac{1}{\delta+\beta_j}\right),$$

whence

$$D_{ij}(\infty)=\frac{\sigma^2}{2(\beta_i+\beta_j)}\left(\frac{1}{\delta+\beta_i}+\frac{1}{\delta+\beta_j}\right).$$

Example 13. Let the service time be uniformly distributed on the interval $[0,1]$, and let $C(t)=(1-\delta|t|)_+$, $\delta\in(0,1]$; then

$$\begin{aligned}\int\limits_x^1\int\limits_x^1(1-\delta|u-v|)(1-u)(1-v)\,du\,dv&=\int\limits_0^{1-x}\int\limits_0^{1-x}(1-\delta|u-v|)uv\,du\,dv\\&=\frac{1}{60}(1-x)^4(15-4\delta(1-x)),\quad x\in[0,1],\end{aligned}$$

whence for all $T\ge1$ we obtain

$$G_T(x)\approx\exp\biggl\{-\frac{\lambda_0}{2}(1-x)^2+\frac{\sigma^2}{8}(1-x)^4-\frac{\delta\sigma^2}{30}(1-x)^5\biggr\},\quad x\in[0,1].$$

This formula has a probabilistic sense only if $G_T(x)$ happens to be a nondecreasing function, which yields the condition

$$\lambda_0\ge\frac{\sigma^2}{6}(3-\delta).$$

In Fig. 2 we present the plot of the function for $\lambda_0=4$, $\sigma^2=8$, and $\delta=0,\frac{1}{2},1$ (from top to bottom). It is seen that discrepancies rapidly decay.

Since the terms under the exponent that are due to the random component of the rate decrease with x faster than those due to the constant component λ_0, the limit distribution in Examples 11–13 as $\lambda_0\to\infty$ and $\sigma/\lambda_0\to0$ with the corresponding linear normalization coincides with that in Examples 4, 5, and 7 (with λ replaced by λ_0).

5. EXTREMAL INDEX

The extremal index of a narrow-sense stationary sequence $\{\xi_n\}$, $n\ge1$, is defined as follows [16, Section 3.7].

Definition. Let random variables ξ_n, $n\ge1$, have distribution F, and let $M_n=\max\{\xi_1,\ldots,\xi_n\}$. If for each $\tau>0$ there exists a number sequence $u_n(\tau)$ such that $n\bar F(u_n(\tau))\to\tau$ and $\mathbf{P}(M_n\le u_n(\tau))\to e^{-\theta\tau}$, $n\to\infty$, then θ is said to be the *extremal index.*

Any value of $\theta\in[0,1]$ can be possible.

One of interpretations of the extremal index is as follows: high-level exceedances in a sequence occur not one by one but in groups (clusters) of average size $1/\theta$. In applications this can mean natural disasters, failures of engineering systems, data losses in information transmission, financial losses, etc. Clearly, if such events happen several times in success, this is much more dangerous than single events and must be taken into account in risk management.

For more details on extremal indices, see [16, 17].

Let us turn back to the case of a constant arrival flow rate. Denote by $W(t)$ the maximum remaining service time at time $t\ge0$ under the assumption that the system started its operation infinitely long ago and is in a stationary regime, so that $W(t)$ has distribution G_∞ for any $t\ge0$.

Consider the stationary sequence $W(n)$, $n\ge1$, and put the question of its extremal index.

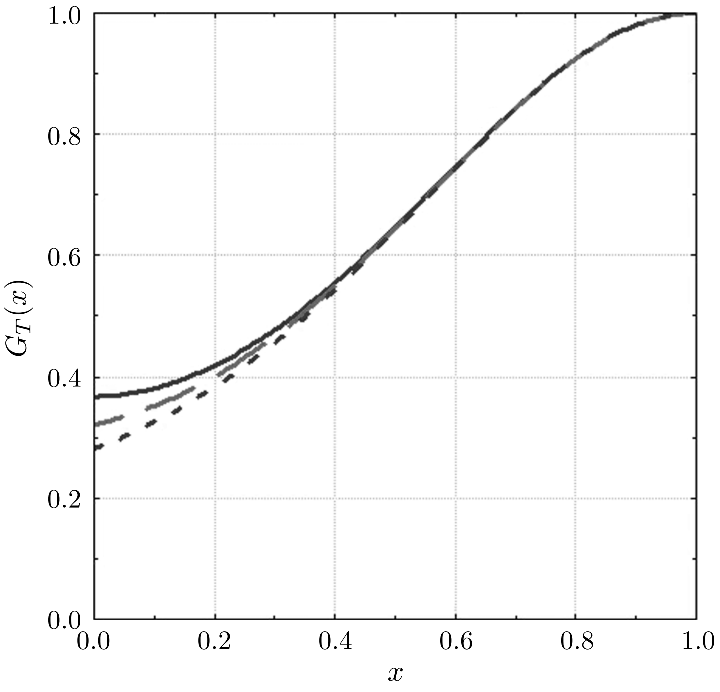

Fig. 2. Plot of the function $G_T(x)$ to Example 13.

Theorem 6. *If* $\bar{B}(x) \sim ce^{-\beta x}$, $x \to \infty$, *then* $\theta = 1 - e^{-\beta}$.

Hence, under the condition of the theorem, high-level exceedances in the sequence $W(n)$, $n \geq 1$, form clusters of average size $1/(1 - e^{-\beta})$.

By properly choosing a time unit, one can also obtain from this that for any $h > 0$ the sequence $W(nh)$, $n \geq 1$, has extremal index $\theta = 1 - e^{-\beta h}$.

It is also clear that if the tail of B decreases faster than any exponent, then $\theta = 1$, and if slower than any exponent (heavy tail, for instance, power-law one), then $\theta = 0$. In the former case, high-level exceedances occur asymptotically one by one, and in the latter, form series of infinite average length.

6. QUEUES WITH BATCH ARRIVALS

Assume that batch sizes are independent and have generating function $g(s)$ and that service times are independent of a batch size.

Theorem 7. *If service times of customers within a batch are independent, then*

$$B^*(x) = g(B(x)).$$

Example 14. Let batch sizes be upper bounded by a number $n > 1$, so that

$$g(s) = \sum_{m=1}^{n} p_m s^m, \quad p_m \geq 0, \quad \sum_{m=1}^{n} p_m = 1,$$

and let B be an exponential distribution (1); then B^* is a hyperexponential distribution (2):

$$B^*(x) = \sum_{m=1}^{n} p_m \left(1 - e^{-\beta x}\right)^m = 1 - \sum_{k=1}^{n} c_k e^{-k\beta x}, \quad c_k = (-1)^k \sum_{m=1}^{n} p_m C_m^k.$$

In a recent paper [21], a model with dependent service times within a batch was considered. To describe this dependence, a multivariate hyperexponential Marshall–Olkin distribution was introduced.

The main ideas behind are the following:

1. Besides independent factors influencing service times of each separate customer in a batch, there is a common factor influencing all of them together. In the Marshall–Olkin model, this is the termination factor. Namely, in a random time after servicing the batch begins, service of all customers in the batch that have not been served by this time is simultaneously terminated.

2. Each batch has a random type (independently of others), which specifies distributions of service times within the batch.

Following these ideas, we construct a more general model. Assume that the batch type and size are independent. The type takes values $1 \le i \le n$ with probabilities c_i. If a batch has size m and type i, then service times are given by $X_j = \min\{Y_j, Z\}$, $1 \le j \le m$, where Y_j, $1 \le j \le m$, are independent and have distribution B_i', and Z is independent of Y_j, $1 \le j \le m$, and has distribution B_i''.

Then service times in a batch of type i have distribution functions

$$B_i(x) = 1 - \bar{B}'{}_i(x)\bar{B}''{}_i(x), \tag{7}$$

and in an arbitrary batch,

$$B(x) = \sum_{i=1}^{n} c_i B_i(x). \tag{8}$$

Theorem 8. *In this model we have*

$$B^*(x) = 1 - \sum_{i=1}^{n} c_i(1 - g(B_i'(x)))\bar{B}_i''(x).$$

Example 15. In the model from [21] (up to changing the notation) we have the following: B_i' are exponential distributions with parameters $\beta_i' = \beta_i - \delta_i$, $0 < \delta_i < \beta_i$; B_i'' are exponential distributions with parameters δ_i. Then, by (7) and (8), service times in a batch of type i have exponential distribution with parameter β_i, and in an arbitrary batch, hyperexponential distribution, precisely according to (2).

As was noted in [21], to ensure a known correlation coefficient ρ_i between service times in a batch of type i, one should put

$$\delta_i = \frac{2\beta_i\rho_i}{1+\rho_i}.$$

In this example, by Theorem 6 we obtain

$$B^*(x) = 1 - \sum_{i=1}^{n} c_i(1 - g(1 - e^{-(\beta_i - \delta_i)x}))e^{-\delta_i x}, \quad x \ge 0.$$

If batch sizes are bounded from above, this formula, as well as in Example 14, also leads to some hyperexponential distribution.

7. CONCLUSION

We have studied the maximum remaining service time in $M|G|\infty$ infinite-server queues, which is understood as the maximum of remaining service times over all busy servers at a given time or in a stationary regime. From the practical point of view, this is the time required to correctly complete the system operation after terminating the arrival flow (for example, in computer data processing or in telecommunication data transmission). The following cases for the arrival flow rate have been considered: (1) time-independent, (2) given by a function of time, (3) given by a random process. In the last case, concrete results are obtained for a stationary Gaussian process. As examples of service time distributions, we have considered exponential, hyperexponential, Pareto, and uniform distributions. In the case of a constant rate, we have studied effects occurring when the mean service time is infinite (for distributions with power-law tails). We have found the extremal index of a sequence of maximum remaining service times. The results are extended to $M^X|G|\infty$ batch arrival queues, in particular with dependent service times within a batch (generalized Marshall–Olkin model).

APPENDIX

Proof of Theorem 1. We use the well-known property of a stationary Poisson flow: for a given number k of arrivals during the interval $[0,T]$, their arrival times (irrespective of the order) are independent and identically distributed in this interval. Given that an arrival appeared at time $T-t$, its probability to be served before time $T+x$ has distribution $B(t+x)$. Arrivals are served independently. We obtain

$$G_T(x)=\sum_{k=0}^{+\infty}\frac{(\lambda T)^k}{k!}e^{-\lambda T}\left(\frac{1}{T}\int\limits_0^T B(t+x)\,dt\right)^k=\exp\left\{-\lambda\int\limits_0^T \bar{B}(t+x)\,dt\right\}.\quad\triangle$$

Proof of Theorem 2. Taking the logarithm of (4), we get

$$s\ln F_T(a(s)x+b(s))\to\ln H(x),\quad s\to\infty,$$

whence

$$\bar{F}_T(a(\lambda)x+b(\lambda))\sim-\ln F_T(a(\lambda)x+b(\lambda))\sim-\frac{\ln H(x)}{\lambda},\quad\lambda\to\infty;$$

therefore,

$$G_T(a(\lambda)x+b(\lambda))=\exp\{-\lambda\mu_T\bar{F}_T(a(\lambda)x+b(\lambda))\}\to H^{\mu_T}(x).\quad\triangle$$

Proof of Theorem 3. For $0<\alpha<1$ we have

$$\int\limits_{xT^{1/\alpha}}^{T+xT^{1/\alpha}}\bar{B}(t)\,dt\sim c\int\limits_{xT^{1/\alpha}}^{T+xT^{1/\alpha}}t^{-\alpha}\,dt=c\frac{(T+xT^{1/\alpha})^{1-\alpha}-(xT^{1/\alpha})^{1-\alpha}}{1-\alpha}$$
$$=\frac{c(xT)^{1/\alpha-1}}{1-\alpha}\left(\left(1+\frac{T}{xT^{1/\alpha}}\right)^{1-\alpha}-1\right)\sim\frac{c(xT)^{1/\alpha-1}}{1-\alpha}\frac{1-\alpha}{xT^{1/\alpha-1}}=cx^{-1/\alpha},\quad T\to\infty,$$

which implies (5). For $\alpha=1$ we have

$$\int\limits_{xT}^{T+xT}\bar{B}(t)\,dt\sim c\int\limits_{xT}^{T+xT}t^{-1}\,dt=c(\ln(T+Tx)-\ln(Tx))=c\ln\frac{1+x}{x},\quad T\to\infty,$$

which implies (6). $\triangle$

Proof of Theorem 4. We use the well-known property of a nonstationary Poisson flow: for a given number k of arrivals during the interval $[0,T]$, their arrival times (irrespective of the order) are independent and identically distributed with density $\lambda(t)/\Lambda(T)$, $t\in[0,T]$, where

$$\Lambda(T)=\int\limits_0^T\lambda(t)\,dt.$$

Given that an arrival appeared at time t, its probability to be served before time $T+x$ has distribution $B(T-t+x)$. Arrivals are served independently. We obtain

$$\begin{aligned}G_T(x)&=\sum_{k=0}^{+\infty}\frac{(\Lambda(T))^k}{k!}e^{-\Lambda(T)}\Biggl(\frac{1}{\Lambda(T)}\int\limits_0^T\lambda(t)B(T-t+x)\,dt\Biggr)^k\\&=\exp\Biggl\{-\int\limits_0^T\lambda(t)\bar{B}(T-t+x)\,dt\Biggr\}.\quad\triangle\end{aligned}$$

Proof of the corollary. The random variable

$$\zeta=-\int\limits_0^T\lambda(t)\bar{B}(T-t+x)\,dt=-\int\limits_0^T\lambda(T-t)\bar{B}(t+x)\,dt\tag{9}$$

is normally distributed with parameters

$$\mathbf{E}\zeta=-\lambda_0\int\limits_0^T\bar{B}(t+x)\,dt,\qquad\mathbf{D}\zeta=\sigma^2\int\limits_0^T\int\limits_0^T C(u-v)\bar{B}(u+x)\bar{B}(v+x)\,du\,dv.$$

We use the well-know formula for the mathematical expectation of a normal random variable:

$$\mathbf{E}e^{\zeta}=\exp\Bigl\{\mathbf{E}\zeta+\frac{1}{2}\mathbf{D}\zeta\Bigr\}.$$

If we now take into account that the arrival flow rate cannot take negative values and replace $\lambda(t)$ with $\lambda_1(t)=\max\{\lambda(t),0\}$ in (9), we obtain $\zeta_1\le\zeta$, whence $G_T(x)=\mathbf{E}e^{\zeta_1}\le\mathbf{E}e^{\zeta}$. $\triangle$

Remark to the proof of the corollary. The approximation error due to nonnegativity of the arrival flow can be upper estimated as follows.

Using Hölder's inequality with some $p,q>1$, $1/p+1/q=1$, we obtain

$$\mathbf{E}e^{\zeta}-\mathbf{E}e^{\zeta_1}=\mathbf{E}e^{\zeta}(1-e^{-(\zeta-\zeta_1)})\le(\mathbf{E}e^{p\zeta})^{1/p}\left(\mathbf{E}(1-e^{-(\zeta-\zeta_1)})^q\right)^{1/q}.$$

The first factor on the right-had side is computed easily:

$$(\mathbf{E}e^{p\zeta})^{1/p}=\left(\exp\Bigl\{p\,\mathbf{E}\zeta+\frac{p^2}{2}\mathbf{D}\zeta\Bigr\}\right)^{1/p}=\exp\Bigl\{\mathbf{E}\zeta+\frac{p}{2}\mathbf{D}\zeta\Bigr\};$$

to estimate the second, we use the fact that the function $y(x)=1-e^{-x}$ satisfies for $x\ge0$ the inequality $0\le y(x)\le\min\{x,1\}$:

$$\mathbf{E}(1-e^{-(\zeta-\zeta_1)})^q\le\mathbf{E}(1-e^{-(\zeta-\zeta_1)})\le\mathbf{E}(\zeta-\zeta_1).$$

Denote

$$\mu_{T,x} = \int\limits_0^T \bar{B}(t+x)\,dt, \quad x \ge 0;$$

then $\mu_{T,x} \le \mu_T$.

Now we expand the difference according to (9), using the stationarity of $\lambda(t)$ and symmetry of the normal distribution:

$$\begin{aligned}\mathbf{E}(\zeta - \zeta_1) &= \int\limits_0^T \mathbf{E}\big(\lambda_1(T-t) - \lambda(T-t)\big)\bar{B}(t+x)\,dt \\ &= \mu_{T,x}\,\mathbf{E}(\lambda_1(0) - \lambda(0)) = -\mu_{T,x}\,\mathbf{E}\min\{\sigma\xi(0) + \lambda_0, 0\} \\ &= \mu_{T,x}\,\mathbf{E}\max\{\sigma\xi(0) - \lambda_0, 0\} = \mu_{T,x}\sigma\,\mathbf{E}\max\{\xi(0) - \lambda_0/\sigma, 0\}.\end{aligned}$$

Using the well-known upper estimate for the tail of the standard normal distribution Φ, we obtain

$$\mathbf{E}\max\{\xi(0) - \lambda_0/\sigma, 0\} = \int\limits_{\lambda_0/\sigma}^{\infty} \bar{\Phi}(x)\,dx \le \int\limits_{\lambda_0/\sigma}^{\infty} \frac{1}{x}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \le \frac{\sigma}{\lambda_0}\bar{\Phi}\Big(\frac{\lambda_0}{\sigma}\Big).$$

Combining the formulas, we come to the estimate

$$\mathbf{E}e^{\zeta} - \mathbf{E}e^{\zeta_1} \le \exp\Big\{\mathbf{E}\zeta + \frac{p}{2}\mathbf{D}\zeta\Big\}\left(\frac{\mu_{T,x}\sigma^2}{\lambda_0}\bar{\Phi}\Big(\frac{\lambda_0}{\sigma}\Big)\right)^{1/q}.$$

By varying p and q for concrete values of the parameters, one can further optimize the estimate numerically. Analytical refinement of the estimate is also possible.

Proof of Theorem 6. For the maximum remaining service times we have the recurrent formula

$$W(n) = \max\{W(n-1) - 1, V(n)\}, \quad n \ge 1, \tag{10}$$

where $V(n)$ is the maximum remaining service time at time n for arrivals that appeared during the interval $(n-1, n]$. The random variables $V(n)$, $n \ge 1$, are independent and have a common distribution G_1, and each $V(n)$ does not depend on $W(k)$, $k < n$.

Let $M_n = \max\{W(1), W(2), \ldots, W(n)\}$; then (10) implies

$$M_n = \max\{W(1), V(2), \ldots, V(n)\},$$

where all the variables under the maximum sign are independent. Using Theorem 1, we obtain

$$\mathbf{P}(M_n \le x) = \exp\bigg\{-\lambda\bigg(\int\limits_0^{\infty} \bar{B}(t+x)\,dt + (n-1)\int\limits_0^1 \bar{B}(t+x)\,dt\bigg)\bigg\}.$$

From $\bar{B}(x) \sim ce^{-\beta x}$, $x \to \infty$, it follows that $\bar{G}_{\infty}(x) \sim (\lambda c/\beta)e^{-\beta x}$, $x \to \infty$. Define $u_n(\tau)$, $\tau > 0$, $n \ge 1$, by the condition

$$e^{-\beta u_n(\tau)} = \frac{\beta\tau}{\lambda cn};$$

then $n\bar{G}_{\infty}(u_n(\tau)) \to \tau$, $n \to \infty$. Based on the asymptotics

$$\int\limits_0^{\infty} \bar{B}(t+x)\,dt \sim \frac{c}{\beta}e^{-\beta x}, \quad \int\limits_0^1 \bar{B}(t+x)\,dt \sim \frac{c(1-e^{-\beta})}{\beta}e^{-\beta x}, \quad x \to \infty,$$

we obtain

$$\mathbf{P}(M_n \le u_n(\tau)) = \exp\left\{-\lambda\left(\frac{c}{\beta}\frac{\beta\tau}{\lambda cn}(1+o(1)) + (n-1)\frac{c(1-e^{-\beta})}{\beta}\frac{\beta\tau}{\lambda cn}(1+o(1))\right)\right\}$$
$$\to \exp\{-(1-e^{-\beta})\tau\}, \quad n\to\infty,$$

whence $\theta = 1-e^{-\beta}$ by the definition. △

Proof of Theorem 7. Let

$$g(s) = \sum_{m=1}^{\infty} p_m s^m, \quad p_m \ge 0, \quad \sum_{m=1}^{\infty} p_m = 1.$$

Averaging the maximum service time over the batch size, we obtain

$$B^*(x) = \sum_{m=1}^{\infty} p_m B^m(x) = g(B(x)).$$

Proof of Theorem 8. If a batch is of size m and type i, the maximum service time for it amounts to

$$X_{\max} = \min\{\max\{Y_1,\dots,Y_m\}, Z\}$$

and has the distribution function

$$B^*_{m,i}(x) = 1-(1-(B'_i(x))^m)\bar{B}''_i(x).$$

Averaging over the batch size and type, we obtain

$$B^*(x) = \sum_{i=1}^{n} c_i \sum_{m=1}^{\infty} p_m(1-(1-(B'_i(x))^m)\bar{B}''_i(x))$$
$$= 1-\sum_{i=1}^{n} c_i(1-g(B'_i(x)))\bar{B}''_i(x). \quad \triangle$$

REFERENCES

1. Riordan, J., Telephone Traffic Time Averages, *Bell Syst. Tech. J.*, 1951, vol. 30, no. 4, pp. 1129–1144.

2. Afanas'eva, L.G. and Bulinskaya, E.V., *Sluchainye protsessy v teorii massovogo obsluzhivaniya i upravleniya zapasami* (Random Processes in Queueing and Inventory Management Theory), Moscow: Moscow State Univ., 1980.

3. Bocharov, P.P. and Pechinkin, A.V., *Teoriya massovogo obsluzhivaniya* (Queueing Theory), Moscow: Ross. Univ. Druzhby Narodov, 1995.

4. Lebedev, A.V., Extrema of Some Queueing Processes, *Cand. Sci. (Math.) Dissertation*, Moscow: Moscow State Univ., 1997.

5. Chernavskaya, E.A., Limit Theorems for Infinite-Server Queues with Heavy-Tail Service Time Distributions, *Cand. Sci. (Math.) Dissertation*, Moscow: Moscow State Univ., 2017.

6. Lebedev, A.V., Asymptotics of Maxima in an Infinite Server Queue with Bounded Batch Sizes, *Fundam. Prikl. Mat.*, 1996, vol. 2, no. 4, pp. 1107–1115.

7. Lebedev, A.V., Maxima in the $M^X|G|\infty$ System with "Heavy Tails" of Group Sizes, *Avtomat. i Telemekh.*, 2000, no. 12, pp. 115–121 [*Autom. Remote Control* (Engl. Transl.), 2000, vol. 61, no. 12, pp. 2039–2044].

8. Glukhova, E.V. and Orlov, A.B., Mean Busy Period of Infinite Multilinear Queues with a Doubly Stochastic Input Flow, *Izv. Vuzov, Ser. Fiz.*, 2003, no. 3, pp. 62–68 [*Russian Phys. J.* (Engl. Transl.), 2003, vol. 46, no. 3, pp. 287–295].

9. Orlov, A.B., Probability Density of the Maximum Remaining Service Time on Busy Servers, *Vychisl. Tekhnol.*, 2008, vol. 13, Special Issue 5 (Selected Talks of the VI Int. Conf. on Information Technologies and Mathematical Modeling, Anzhero-Sudzhensk, Russia, Nov. 9–10, 2007), pp. 93–98.

10. Tarasov, V.N., Analysis of Queues with Hyperexponential Arrival Distributions // *Probl. Peredachi Inf.*, 2016, vol. 52, no. 1, pp. 16–26 [*Probl. Inf. Trans.* (Engl. Transl.), 2016, vol. 52, no. 1, pp. 14–23].

11. Ushakov, V.G., Queueing System with Working Vacations and Hyperexponential Input Stream, *Inform. i ee Primen.*, 2016, vol. 10, no. 2, pp. 92–97.

12. Chernavskaya, E.A., Limit Theorems for an Infinite-Server Queuing System, *Mat. Zametki*, 2015, vol. 98, no. 4, pp. 590–605 [*Math. Notes* (Engl. Transl.), 2015, vol. 98, no. 3–4, pp. 653–666].

13. Shelukhin, O.I., Osin, A.V., and Smol'skii, S.M., *Samopodobie i fraktaly: telekommunikatsionnye prilozheniya* (Self-similarity and Fractals: Telecommunication Applications), Moscow: Fizmatlit, 2008.

14. Borovkov, A.A., *Asimptoticheskie metody v teorii massovogo obsluzhivaniya*, Moscow: Nauka, 1980. Translated under the title *Asymptotic Methods in Queuing Theory*, Chichester: Wiley, 1984.

15. Galambos, J., *The Asymptotic Theory of Extreme Order Statistics*, New York: Wiley, 1978. Translated under the title *Asimptoticheskaya teoriya ekstremal'nykh poryadkovykh statistik*, Moscow: Nauka, 1984.

16. Leadbetter, M.R., Lindgren, G., and Rootzén, H., *Extremes and Related Properties of Random Sequences and Processes*, New York: Springer, 1983. Translated under the title *Ekstremumy sluchainykh posledovatel'nostei i protsessov*, Moscow: Mir, 1989.

17. Embrechts, P., Klüppelberg, C., and Mikosh, T., *Modelling Extremal Events for Insurance and Finance*, New York: Springer, 2003, 4th ed.

18. Evdokimova, G.S., Multichannnel Queueing Systems with Periodic Input Flow, *Avtomat. i Telemekh.*, 1974, no. 4, pp. 62–65 [*Autom. Remote Control* (Engl. Transl.), 1974, vol. 35, no. 4, part 1, pp. 571–574].

19. Afanas'eva, L.G. and Bashtova, E.E., Limit Theorems for Queueing Systems with Doubly Stochastic Poisson Arrivals (Heavy Traffic Conditions), *Probl. Peredachi Inf.*, 2008, vol. 44, no. 4, pp. 72–91 [*Probl. Inf. Trans.* (Engl. Transl.), 2008, vol. 44, no. 4, pp. 352–369].

20. Eick, S.G., Massey, W.A., and Whitt, W., $M_t|G|\infty$ Queues with Sinusoidal Arrival Rates, *Manage. Sci.*, 1993, vol. 39, no. 2, pp. 241–252.

21. Pang, G. and Whitt, W., Infinite-Server Queue with Batch Arrivals and Dependent Service Times, *Probab. Engrg. Inform. Sci.*, 2012, vol. 26, no. 2, pp. 197–200.