STUDIA UNIVERSITATIS
Revistă ştiinţifică a Universităţii de Stat din Moldova, 2008, nr.8(18)

HEAVY TRAFFIC ANALYSIS IN QUEUEING SYSTEMS

Olga BENDERSCHI
Catedra Informatică şi Optimizare Discretă

The traffic intensity is defined in the classical theory of queueing systems as the ratio of the expected service time and the expected interarrival time and is an important measure of the system performance. When the traffic intensity is close but less than one such approximations are called *heavy-traffic approximations*. Heavy traffic limits for queuing networks are a topic of continuing interest. However, these limits have been rigorously derived only for a few types of systems. The goal of this paper is to describe the heavy-traffic behavior of classical queuing systems.

1 Introduction

When performing the analysis of a service system, we are usually trying to describe the congestion experienced by a typical arrival, actual or virtual (at an arbitrary time), and, thus, to perform the standard steady-state analysis. In order to be able to effectively use extreme-value engineering in performance analysis, we need to be able to determine the distribution, or at least the mean, of the maximum congestion. This requirement represents a major difficulty, because distributions of maximum congestion measures in queueing models are unavailable except in very special cases. However, the extreme-value theory comes to our aid. Fundamental limit theorems in extreme-value theory imply that the extreme value distributions over suitably long intervals can be approximated by a few special distributions (see Castillo 1988 [1], Glynn and Whitt 1994 [2]).

The pioneering works in heavy traffic approximations to queues (Kingman 1961[3]) and queueing networks (Iglehart and Whitt 1970 [4], [5]) appeared a while ago. A detailed overview of the enormous literature body is given in Williams 1996 [6]. A classified list of research on heavy traffic limit theorems for queues is given in Kimura 1993 [7].

2 Queueing Systems $M|G|1$

2.1 Heavy traffic limits for waiting time

Assume that (see Whitt 2000 [8]) the queueing process is a discrete-time process satisfying the finite-capacity generalization of the classical Lindley recursion, i.e.,

$$Q(k) = max\Big\{0, min\{C, Q(k-1) + X(k)\}\Big\},$$

where $X(k)$ is the net input between periods $k-1$ and k; i.e.,

$$X(k) = A(k) - B(k), \quad k \geq 1,$$

$A(k)$ is a nonnegative input and $B(k)$ is a nonnegative potential (maximum possible) output. The variable $Q(k)$ depicts the queue (or buffer) content in period k. In the classical Lindley recursion associated with the $GI/GI/1/$ queue, $C = \infty$, $Q(k)$ is the waiting time of the k^{th} customer before beginning service, $A(k)$ is the service time of the $(k-1)^{st}$ customer and $B(k)$ is the interarrival time between the $(k-1)^{st}$ and k^{th} customers.

The first heavy-traffic limit theorem was obtained by Kingman J. F. G. 1961 [3], 1962 [9]. By applying asymptotics to the previously determined transform of the steady-state queue content $Q(\infty)$ in the case $C = \infty$, Kingman showed that the relatively complicated steady-state distribution is asymptotically exponential as the traffic intensity $\rho \equiv \frac{EA(k)}{EB(k)}$ approaches 1, leading to the approximations

$$P(Q(\infty) > x) \approx e^{-x/EQ(\infty)}$$

and

$$EQ(\infty) \approx \frac{EA(1)\rho(c_A^2 + c_B^2)}{2(1-\rho)}$$

where c_A^2 and c_B^2 are the squared coefficients of variation (SCV, variance divided by the square of the mean) of A(1) and B(1), respectively.

2.2 Heavy-traffic extreme-value limits for queues

Peter W. Glynn and Ward Whitt (1994)[2] constructed a sequence of queueing systems indexed by n whose traffic intensities ρ_n approach 1 from below as $n \to \infty$. The length of the interval t_n over which the maximum is taken, must also approach infinity, but neither too quickly nor too slowly. We need $(1-\rho_n)^2 t_n \to \infty$ as $n \to \infty$ to have the relevant time in reflected Brownian motion (RBM) go to infinity, but we also need to impose conditions on how fast t_n grows. These conditions allow the limit to hold even when the normalized maximum wait fails to have the customary extreme-value limit as $t \to \infty$ for fixed ρ.

The specific process we consider is the sequence of waiting times in the $GI/G/1$ queue (so that t_n above should be an integer), but the argument extends easily to other processes and models, given that corresponding strong approximations hold. The corresponding limit for the discrete queue-length process is interesting because no extreme-value limit holds for each fixed ρ.

Further are described the results received by Peter W. Glynn and Ward Whitt (1994)[2].

For each $n \geq 1$, let $W_n \equiv \{W_n(k) : k \geq 0\}$ be a waiting time sequence, defined by $W_n(0) = 0$ and

$$W_n(k+1) = [W_n(k) + \rho_n V_k - U_k]^+,$$

where $[x]^+ = max\{x, 0\}$, $U \equiv \{U_k : k \geq 1\}$ and $V \equiv \{V_k : k \geq 0\}$ are independent sequences of i.i.d. nonnegative random variables satisfying

$$EV_k = EU_k = 1,$$

$\sigma_v^2 \equiv VarV_k < \infty$ and $\sigma_u^2 \equiv VarU_k < \infty$, with at least one of $\sigma_v^2 > 0$ and $\sigma_u^2 > 0$.

Let $A_n = \sum_{k=1}^{n} U_k$ and $C_n = \sum_{k=0}^{n-1} V_k$. Then

$$W_n(k) = S_n(k) - \min_{0 \leq j \leq k} S_n(j),$$

where

$$S_n(k) = \rho_n C_k - A_k.$$

Let $B \equiv \{B(t) : t \geq 0\}$ be canonical (drift 0, variance 1) Brownian motion (BM) and let $R \equiv \{R(t) : t \geq 0\}$ be canonical RBM (with drift 0 and variance 1), i.e.,

$$R(t) = t + B(t) - \min_{0 \leq s \leq t} \{s + B(s)\}, \quad t \geq 0.$$

Let $M_n(k) = \max\limits_{0\le j\le k} W_n(j), k \ge 0$, and $M(t) = \max\limits_{0\le j\le k} R(s), t \ge 0$. Extreme-value limits for $M_n(k)$ as $k \to \infty$ for any fixed n are given in Iglehart 1972 [10] and Pakes [11]. These limits require the extra condition

$$E \quad exp(\varepsilon V_k) < \infty \quad for \quad some \quad \varepsilon > 0 \tag{1}$$

and more, and involve relatively complicated normalization constants. However, it is natural to expect that the situation should simplify in heavy traffic. Let $\Rightarrow$ denote convergence in distribution.

Theorem 1 *If $\rho \uparrow 1$ with $(1-\rho_n)\sqrt{n} \to c$ as $n \to \infty$, were $0 \le c < \infty$ then*

$$n^{-1/2} M_n(nt) \Rightarrow \left[\frac{\sigma_u^2 + \sigma_v^2}{c}\right] M\left[\frac{c^2 t}{\sigma_u^2 + \sigma_v^2}\right] \quad as \quad n \to \infty$$

Let Z be a random variable with the classical Gumbel extreme-value cdf, i.e., $P(Z \le x) = exp\,(-e^{-x})$, $-\infty < x < \infty$.

Theorem 2 *Suppose that $\rho_k \uparrow 1$ with $(1-\rho_n)\sqrt{t_n} \to \infty$ as $n \to \infty$*

(a) If $EV_k^p < \infty$ for $p > 2$ and $\overline{\lim\limits_{n\to\infty}}(1-\rho_n)t_n^{1/p} < \infty$ as $n \to \infty$ then

$$\frac{2(1-\rho_n)M_n(t_n)}{\sigma_u^2 + \sigma_v^2} - \log\left[\frac{2(1-\rho_n)^2 t_n}{\sigma_u^2 + \sigma_v^2}\right] \Rightarrow Z \quad as \quad n \to \infty \tag{2}$$

(b) If (1) holds and $(1-\rho_n) log\, t_n \to 0$ as $n \to \infty$, then (2) holds

2.3 Limits for the busy-period distribution

Limit theorems are established for the busy-period distribution in single-server queues in Abate and Whitt 1994 [12].

Consider the classical $M|G|1$ queue with one server, unlimited waiting space and some work-conserving discipline such as first-come first-served. Customers arrive according to a Poisson process, whose rate we take to be λ. The service times are independent and identically distributed, and independent of the arrival process. Let the service-time distribution have cdf (cumulative distribution function) $B(t)$ with mean $\mu = 1$ and finite second moment m_2. Thus the traffic intensity is $\rho = \lambda \cdot \mu$.

The busy period is the interval between the epoch of an arrival to an empty system and the next epoch at which the system is empty again.

Let $\Pi(t)$ be the cdf of the busy period. We assume that $\rho < 1$; then $\Pi(t)$ is proper.

For any cdf $F(t)$ with mean m, let $F(t) = 1 - F(t)$ be the complementary cdf (ccdf) and let

$$F_e(t) = m^{-1} \int_0^t F^c(u)du, t \ge 0 \tag{3}$$

We characterize the heavy-traffic limit as the density $h_1(t)$ of the first-moment cdf $H_1(t)$ of regulated or reflecting Brownian motion (RBM) investigated in Abate and Whitt 1987 [14]. In particular, $H_1(t)$ is the time-dependent mean of RBM starting empty, normalized by dividing by the steady-state limit. Its density $h_1(t)$ can be expressed explicitly as

$$h_1(t) = 2t^{-1/2}\phi(t^{1/2}) - 2\left[1 - \Phi(t^{1/2})\right] = 2\gamma(t) - \gamma_e(t), t \ge 0 \tag{4}$$

where $\Phi(t)$ is the cdf and $\phi(t)$ is the density of a standard normal random variable with mean 0 and variance 1, $\gamma(t)$ is gamma density with mean 1 and shape parameter 1/2, i.e.,

$$\gamma(t) = (2\pi t)^{-1/2} exp(-t/2), \quad t \geq 0, \tag{5}$$

and $\gamma_e(t)$ is the associated stationary-excess density.

Heavy-traffic limit is obtained by simply increasing the ρ. To obtain our heavy-traffic limit, we scale both inside (time) and outside the complementary cdf $\Pi^c_\rho(t)$. We introduce the subscript ρ to indicate the dependence upon ρ.

Theorem 3 *For each $t > 0$,*

$$\lim_{\rho \to 1} m_2(1-\rho)^{-1} \Pi^c_\rho(t m_2 (1-\rho)^{-2}) = h_1(t) \tag{6}$$

Let Π_ρ be the busy period in the model with traffic intensity ρ. The busy period is understood to mean the interval from when the server first becomes busy until the server is again idle. For models more general than $GI|G|1$, we can interpret this distribution as the long-run average of all such distributions over all busy periods.

Condition C.1. For some constant b, $(1-\rho)E\,\Pi_\rho \to b$ as $\rho \to 1$.

Let $\{W^*_\rho(t) : t \geq 0\}$ be the stationary workload process in the queue with traffic intensity ρ.

Here $W^*_\rho(t)$ should be interpreted as the time required for the system to become empty after time t if no new work were to arrive after time t. Let $\{R^*(t) : t \geq 0\}$ be a stationary version of canonical RBM with drift coefficient -1 and diffusion coefficient 1. The stationary version is initialized by the exponential steady-state distribution with mean $1/2$. Let $\Rightarrow$ denote convergence in distribution or weak convergence. Let $D[0,\infty)$ be the function space of right-continuous real-valued functions with left limits, endowed with the usual Skorohod J_1 topology; e.g., see Ethier and Kurtz 1986 [13].

Condition C.2. For some constant d,

$$\{W^*_\rho(dt(1-\rho)^{-2}) : t \geq 0\} \Rightarrow \{R^*(t) : t \geq 0\} \text{ in } D[0,\infty) \text{ as } \rho \to 1.$$

Theorem 4 *If conditions C1 and C2 hold, then*

$$(d/b)(1-\rho)^{-1} \Pi^c_\rho(dt(1-\rho)^{-2}) \to h_1(t) \quad as \quad \rho \to 1 \tag{7}$$

for each t.

For the $M|G|1$ queue, conditions $C1$ and $C2$ are known to hold with $b = 1$ and $d = m_2 = c_s^2 + 1$, where c_s^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of a service time. Hence, Theorem 4 actually contains Theorem 3 as a special case.

For $GI|G|1$ queues with mean service time 1, the mean busy period coincides with the reciprocal of the probability that an arrival finds an empty queue. For the $M|G|1$ queue, this probability is just $1-\rho$, but for other models it is more complicated. For the $GI|M|1$ queue, Halfin 1985 [15] showed that condition $C1$ holds with

$$b = \frac{c_a^2 + 1}{2}, \tag{8}$$

where c_a^2 is the SCV of the interarrival time.

3 Queueing Systems $M|G|1$ with priority classes

3.1 Heavy traffic limits for waiting time

Consider the $M|G|1$ queue with two priority classes. It is shown by O.J. Boxma, J.W. Cohen and Q. Deng [16] for heavy-tailed case that the waiting time distribution of the low-priority customers is regularly varying of index one degree higher than that of the service time distribution with the heaviest tail.

Consider the $M|G|1$ queueing model with two priority classes,with either the nonpreemptive or the preemptive resume discipline. We are interested in the effect of the priority structure on the tail of the low-priority waiting-time distribution.

The high-priority class is indexed by 1 and the low-priority class by 2. Let $B_j(t)$ denote the service time distribution function of class-j, λ_j the arrival rate of class-j and ρ_j the traffic load of class-j for $j = 1, 2$. The arrival processes of the two classes are independent. For j =1, 2; put

$$\beta_j := \int_0^\infty t dB_j(t) < \infty,$$

$$\beta_{j\,2} := \int_0^\infty t^2 dB_j(t) < \infty,$$

$$\rho_j := \lambda_j \beta_j,$$

$$\rho := \rho_1 + \rho_2,$$

and assume that $\rho < 1$.

Let W_2 denote the steady-state waiting time of the low-priority customers until start of the service (note that it has the same distribution for the nonpreemptive and the preemptive resume discipline). When $\beta_{j2} < \infty$ for $j = 1, 2$, the following heavy-traffic limit theorem for W_2 holds (cf. J. Abate, W. Whitt 1997[17]):

$$\lim_{\rho_2 \uparrow 1-\rho_1} Pr\{\Delta W_2 \le t\} = 1 - e^{-t}, t \ge 0, \tag{9}$$

where

$$\Delta := \frac{2(1-\rho_1)(1-\rho)}{\rho_1 \beta_{12}/\beta_1 + \rho_2 \beta_{22}/\beta_2}.$$

Assume that at least one of the service time distributions has a regularly varying tail with index $-v$, i.e.

$$1 - B_j(t) \sim L(t) t^{-v} \quad as \quad t \to \infty,$$

for $j = 1$ and/or $j = 2$, where $L(t)$ is a slowly varying function and $1 < v < 2$. Here $f(t) \sim g(t)$ as $t \to \infty$ stands for $\lim_{t\to\infty} f(t)/g(t) = 1$. A measurable positive function $L(t)$ defined on some neighborhood $[a, \infty)$ is called as slowly varying function if for all $x > 0$, $\lim_{t\to\infty} L(xt)/L(t) = 1$.

For $s \ge 0$ and j=1, 2; define the L-S transforms of the service time distributions and of the residual service time distributions.

$$\beta_j\{s\} := \int_0^\infty e^{-st} dB_j(t),$$

$$\beta_{je}\{s\} := \frac{1}{\beta_j} \int_0^\infty e^{-st}(1 - B_j(t))dt.$$

Concerning the service time distributions $B_j(\cdot)$ for $j = 1, 2$, we only introduce assumptions about their tails, i.e. about $1 - B_j(t)$ for $t \to \infty$. It is assumed that one of the service time j distributions has a regularly varying tail behavior, another one has less heavy tail behavior, or both of the service time distributions have a regularly varying tail with the same index. That is, one of the following assumptions holds,

$$(i) \quad 1 - B_1(t) \sim -\frac{1}{1-v}(t/\beta_1)^{-v} L(t/\beta_1) \quad as \quad t \to \infty, \tag{10}$$

$$M_{2\mu} := \int_0^\infty t^\mu dB_2(t) < \infty, \quad for \quad a \quad \mu > v$$

$$(ii) \quad 1 - B_2(t) \sim -\frac{1}{1-v}(t/\beta_2)^{-v} L(t/\beta_2) \quad as \quad t \to \infty$$

$$M_{1\mu} := \int_0^\infty t^\mu dB_1(t) < \infty, \quad for \quad a \quad \mu > v$$

$$(iii) \quad 1 - B_j(t) \sim -\frac{1}{1-v}(t/\beta_j)^{-v} L_j(t/\beta_j) \quad as \quad t \to \infty$$

$$L(t) := L_1(t) \quad for \quad t \geq 0,$$

$$\alpha := \lim_{t\to\infty} \frac{L_2(t)}{L(t)} < \infty;$$

$$(iv) \quad 1 - B_j(t) \sim -\frac{1}{1-v}(t/\beta_j)^{-v} L_j(t/\beta_j) \quad as \quad t \to \infty$$

$$L(t) := L_2(t) \quad for \quad t \geq 0,$$

$$\alpha := \lim_{t\to\infty} \frac{L(t)}{L_1(t)} = \infty;$$

where $1 < v < 2$, $L(\cdot)$, $L_1(\cdot)$ and $L_2(\cdot)$ are slowly varying functions. To obtain our heavy traffic limit theorem,we assume that $L(t)$ is continuous for sufficiently large t. Without loss of generality, we may assume $v < \mu < 2$.

Consider the contraction equation

$$\frac{Kx^{v-1}L(1/x)}{1-\rho} = 1, x > 0, \tag{11}$$

where K is a function of both ρ_1 and ρ_2 such that $K > c$ for some positive constant c, $L(x)$ is a slowly varying function, and denote by $\Delta(\rho_2)$ the unique root of (11) such that

$$\Delta(\rho_2) \downarrow 0 \quad for \quad \rho_2 \uparrow 1 - \rho_1,$$

cf. O.J. Boxma, J.W. Cohen 1997 [18].

Theorem 5 *For the stable $M|G|1$ queue with two priority classes, the service time distributions $B_1(t)$ and $B_2(t)$ satisfying one of the assumptions in (10), the "contracted" waiting time $\Delta(\rho_2)W_2/\beta_1$ converges in distribution for $\rho_2 \uparrow 1 - \rho_1$, and the limit distribution $R_{v-1}(t)$ is given by: for $t \geq 0$,*

$$R_{v-1}(t) = 1 - \sum_{n=0}^{\infty} (-1)^n \frac{t^{n(v-1)}}{(n(v-1)+1}. \tag{12}$$

The coefficient of contraction $\Delta(\rho_2)$ is that root of the equation (11) with the property that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1-\rho_2$, and with $K = K_1, ..., K_4$ corresponding to assumptions $(i), ..., (iv)$ in (11) respectively, where

$$K_1 = \frac{\rho_1}{(1-\rho_1)^{v-1}},\ K_2 = \frac{\rho_2(\beta_2/\beta_1)^{v-1}}{(1-\rho_1)^{v-1}},\ K_3 = \frac{\rho_1 + \rho_2\alpha(\beta_2/\beta_1)^{v-1}}{(1-\rho_1)^{v-1}} \text{ and } K_4 = K_2.$$

Moreover, the L-S transform of $R_{v-1}(t)$ is

$$\int_{0-}^{\infty} e^{-st} dR_{v-1}(t) = \frac{1}{1+s^{v-1}},\ s \geq 0.$$

Consider the queueing model with k priority classes where $k \geq 2$. Let the j-th priority class be indexed by j for $1 \leq j \leq k$. Denote by ρ_j the workload generated by class-j, λ_j the arrival rate of class-j, $B_j(t)$ the service time distribution of class-j, W_j the steady-state class-j waiting time for $1 \leq j \leq k$. To have a steady-state class-k waiting time distribution, we assume $\sum_{j=1}^{k} \rho_j < 1$.

Suppose one of the service time distributions has the following heavy tail behavior:

$$1 - B_i(t) \sim L(t)t^{-v}, \quad as \quad t \to \infty, \tag{13}$$

with $L(t)$ as lowly varying function and $1 < v < 2$, the other service time distributions being such that, for $j \neq i$, $1 \leq j \leq k$,

$$\int_0^{\infty} t^{\mu_j} dB_j(t) < \infty, \quad where \quad \mu_j > v,$$

or

$$1 - B_j(t) \sim L_j(t)t^{-v}$$

with $\lim_{t\to\infty} L_j(t)/L(t) < \infty$.

Let the first $k-1$ classes be the high-priority class, class-k the low-priority class in a queueing model with two priority classes. The service time distributions of the two classes in the new model are given by

$$\widetilde{B}_1(t) = \frac{\sum_{j=1}^{k-1} \lambda_j B_j(t)}{\sum_{j=1}^{k-1} \lambda_j} \tag{14}$$

$$\widetilde{B}_2(t) = B_k(t) \tag{15}$$

The above assumptions imply that one of the assumptions in (10) holds for $\widetilde{B}_1(t)$, $\widetilde{B}_2(t)$. Hence the following heavy-traffic limit theorem holds.

Theorem 6 *For the stable $M|G|1$ queue with k ($k \geq 2$) priority classes, the above assumptions for the service time distributions $B_j(t)$, $1 \leq j \leq k$, holding, the "contracted" waiting time $\Delta(\rho_k)W_k/\beta_1$ converges in distribution for $\rho \uparrow 1 - \sum_{j=1}^{k-1} \rho_j$; the limit distribution $R_{v-1}(t)$ is given by(12), and the coefficient of contraction $\Delta(\rho_k)$ is that root of the equation (11)with the property that $\Delta(\rho_k) \downarrow 0$ for $\rho \uparrow 1 - \sum_{j=1}^{k-1} \rho_j$.*

3.2 Priority Queueing Systems with Switchover Times

Consider a queueing system with a single server and r classes of incoming requests, each having its own flow of arrival and waiting line. Suppose that the time periods between two consecutive arrivals of the requests of the class i are independent and identically distributed with some common cumulative distribution function $A_i(t)$ with mean $\mathbb{E}[A_i]$, $i = 1, \ldots, r$. Similarly, suppose that the service time of a customer of the class i is a random variable B_i with a cumulative distribution function $B_i(t)$ having mean $\mathbb{E}[B_i]$, $i = 1, \ldots, r$.

It is assumed that the server needs some additional time to proceed with the switching from one priority waiting line of requests to another. This time is considered to be a random variable, and we say that C_{ij} is the time of switching from the service of i-requests to the service of j-requests, if $1 \leq i, j \leq r$, $i \neq j$.

This class of Priority Queueing systems is describe in [19, 20].

We discuss here the result obtained in Ciumac Mishkoy[21].

Let as denote by $P_m(t)$ the probability of the fact that in the moment t there are m customer in the system, were $m = (m_1, m_2, \ldots, m_r)$, $m_i \in Z$, $m_i \geq 0$, $i = \overline{1,r}$, (m_i is the number of the L_i flow customers).

Let $z = (z_1, \ldots, z_r)$ be the r-dimensional vector, $|z_i| \leq 1$ ($i = \overline{1,r}$ and $z^m = z_1^{m_1} \times z_2^{m_2} \ldots \times z_r^{m_r}$. Then the generating function of the $P_m(t)$ probabilities will be

$$P(z,t) = \sum_{|m| \geq 0} P_m(t) z^m, \text{ were } |m| = m_1 + m_2 + \ldots + m_r.$$

It is supposed that

$$\beta_{in} = \int_0^\infty t^n dB_i(t) < +\infty \text{ and } c_{in} = \int_0^\infty t^n dC_i(t) < +\infty \quad (i = \overline{1,r}; \quad n \geq 1)$$

By the customers system traffic intensity ρ_{ki} of the first k flows $L_1, \ldots, L_k$ we understand the mean time spend by the device for the generalized customers service of the service of the $L_1, \ldots, L_k$ flows, which arrive on the average in the unit of time.

We considering the following case the heavy traffic intensity all r costumers flows are divided into the l groups of flows, so that the i group ($i = \overline{2,l}$) belong the flows, for which difference between unit and device traffic intensity of the customers of this flow and of the higher priority customers is infinitesimal of higher order than for the flows of $\overline{1, i-1}$ groups, and of lower order than for the flows of $\overline{i+1, l}$ groups. Let p be the number for which $\rho_p \downarrow 0$ and $\rho_{p-1} \to \rho_{p-1}^* > 0$,$p \geq 2$, were $\rho_i = 1 - \rho_{i1}$, $i = \overline{1,r}$.

We introduce the following notations:

$d_{k,j}^* = \lim\limits_{\rho_{p1} \uparrow 1} d_{k,j}$ $d_{k,j} = \rho_k / \rho_j$, $d_k^* = d_{k,k-1}^*$

$\rho_{i2}^* = \lim\limits_{\rho_{p1} \uparrow 1} \rho_{1,2}$ $(k = \overline{p,r}; j = \overline{p-1, k-1}; i = \overline{1,r})$,

where ρ_{i2} quantities are found by double differentiation of the expressions for $\pi_k(s)$, $\nu_k(s)$ and $h_k(s)$ in the point $s = 0$ ($k = \overline{2,r}$).

Let $\sigma_k = a_1 + \ldots + a_k$, $\sigma = \sigma_r$, where $a_1, \ldots, a_r$ are the parameters of the $L_1, \ldots, L_k$ flows.

We note

$$|\sigma - az|_k = a_k(1 - z_k) + \ldots + a_r(1 - z_r),$$

s^* and z_i^* are given by

$$s^* = s\rho_r^2 / \rho_{r2}, \quad z_i^* = U(p-1)z_i + U(1-p+1)e^{-y_1 z_i},$$

were $y_i = \frac{\rho_{i-11}\rho_i}{a_i\rho_{i2}}$ $(i = \overline{1,r})$, $U(t) = \begin{cases} 1, \text{ for } t > 0 \\ 0, \text{ for } t \le 0. \end{cases}$

Then $z^* = (z_1^*, \dots, z_r^*)$. Let us introduce the following notations $(Res \ge 0; k = \overline{1,r})$

$$\mu_{k+1}(s) = s + \sigma_k - \sigma_k\pi_k,$$

$$\eta_k(s) = s - a_k + a_k h_k(s),$$

$$y_k(s) = s + a_k - a_k\overline{\pi_k},$$

were π_k, $h_k(s)$,$\overline{\pi_k}$ are the Laplace - Stieltjes transforms of the distribution functions of the auxiliary systems characteristics. Let $\nu_k$by the Laplace - Stieltjes transforms of the distribution function of the orientation cycle.

We present here the asymptotical expansions of the $\mu_{k+1}(s)$, $\eta_k(s)$, $y_k(s)$ and $\nu_k(s)$ functions.

Theorem 7 *Let* $\rho_{p1} \uparrow 1$ $(p \ge 2)$ *and* $\rho_{i2}^* < +\infty$ $(i = \overline{1,r}$. *Then*

a) $d_k^* = 0$ $(k = \overline{p,r})$

$$\delta_k\Phi_1 \times \dots \times \Phi_k/\rho_{k-1} - \mu_k(\delta_k) \sim Phi_1 \times \dots \times \Phi_k\rho_{k-12}^*\rho_k d_k s^2,$$

$$\eta_k(\delta_k) \sim \rho_k^2(s + \rho_{p2}^* s^2),$$

$$1 - \nu_k(\delta_k) \sim (q_k - 1)\Phi_1 \times \dots \times \Phi_k\rho_k s/\sigma_{k-1},$$

$$y_k(\nu_k)\rho_{k-1}\rho_k\Lambda(s),$$

were $\nu_k \sim s\rho_k^2$,$\delta_k \sim s\rho_{k-1}\rho_k$, $\Lambda(s) = (\rho_{p2}^*)^{-1}\Psi(s\rho_{p2}^*)$,

$$\Psi(s) = -\frac{1}{2}(1 - \sqrt{1+4s}),$$

$\Phi_i = 1 + \frac{\sigma_i - \sigma_{i-1}\pi_{i-1}(a_i)}{\sigma_{i-1}}(q_i - 1)$ $(1 = \overline{2,r})$, $\Phi_1 = 1$ *and* q_i *quantities* $(i = \overline{2,r})$ *were calculated in [19]. In the case when* $p = 2$ *and* $d_2^* = 0$, $y_1(\delta_2) \sim sp_2/(1 + a_1c_{11})$.

We introduce a series of notations, which will promote the presentation of the main result:

$$L_{p-1}(z^{(p-1)}) = \rho_{p-1}^*\big(1 + \sigma_{p-1}\pi_{p-1}(z^{(p-1)}; 0)\big)$$

$$\Delta_i(z,s) = (d_{r,i-1}^*)^2 s + \sum_{n=1}^{r}(d_{n-1,i-1}^*)^2 d_n^* z_n \ (i = \overline{p+1, r+1})$$

$$\overline{\Delta}_i(z,s) = (\rho_{p2}^*)^{-1}\Delta_i(z,s).$$

$$R(z,s) = (1 - U(d_k^*))\Phi_k(U(d_{k+1}^*)\Psi(\Delta_{k+1}(z,s)) - z_k)/(\Delta_{k+1}(z,s) - z_k - z_k^2)$$
$$+U(d_k^*)\Phi_k d_k^*\{(1 - d_k^*)U(d_{k+1}^*)\Psi(\Delta_{k+1}(z,s)) - z_k)\}/\big((t/d_k^* - 1)\Psi(\Delta_k(z,s)) - z_k\big).$$

Theorem 8 *Let* $\rho_{p1} \uparrow 1$ $(p \ge 2)$, $t \to \infty$ *so that* $\rho_k(1 - d_k)^{-1} \to 0$ $(k = \overline{p,r})$, $tp_r^2/\rho_{r2} \to r$ $(r < \infty)$, $\rho_{i2}^* < +\infty$ $(i = \overline{1,r})$. *Then* $(Re > 0)$,

$$P(z^*, t) \longrightarrow f(z,r),$$

were

$$s\int_0^\infty e^{-st}dt = L_{p-1}(z^{(p-1)})\frac{\sqrt{1/4+s} + 1/2}{\Phi_1 \times \dots \times \Phi_r}\prod_{k=p} R_k(z,s).$$

The proofs of these theorems are given in [21].

References

[1] **Castillo E.** *Extreme Value Theory in Engineering.* Academic Press, S. Diego, 1988.

[2] **Glynn P.W, Ward W.** *Heavy-traffic extreme-value limits for queues.* AT T Bell Laboratories Murray Hill, NJ 07974-0636, February 9, 1994.

[3] **Kingman J. F. C.** *The single server queue in heavy traffic.* Proc. Camb. Phil. Soc. 57, 902904, 1961.

[4] **Iglehart D. L. and Whitt W.** *Multiple channel queues in heavy trafffic I.* Adv. in Appl. Probab. 2 150-177, 1970.

[5] **Iglehart D. L. and Whitt W.** *Multiple channel queues in heavy trafffic II.* Adv. in Appl. Probab. 2 355-364, 1970.

[6] **Williams R. J.** *On the approximation of queueing networks in heavy traffic. In Stochastic Networks: Theory and Applications* (S. Zachary, F. P. Kelly and I. Ziedins, eds.) 35–56, Clarendon Press, Oxford, 1996.

[7] **Kimura T.** *A Bibliography of Research on Heavy Traffic Limit Theorems for Queues.* Economic Journal of Hokkaido University pp. 167-179, 1993.

[8] **Whitt W.** *An Overview of Brownian and Non-Brownian FCLTs for the Single-Server Queue.*AT&T Labs, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, 2000.

[9] **Kingman J. F. C.** *On queues in heavy traffic.* J. Roy. Statist. Soc. Ser B, 24, 383392, 1962.

[10] **Iglehart D. L.** *Extreme values in the GI/G/1 queue.* Ann. Math. Statist. 43, 627-635, 1972.

[11] **Pakes A. G.** *On the tails of waiting-time distributions.* J. Appl. Prob., 12, 555-564, 1975.

[12] **Abate J. and Whitt W.** *Limits and approximations for the busy-period distribution in single-server queues.* AT&T Bell Laboratories, September 23, 1994.

[13] **Ethier S. N. and Kurtz T. G.** *Markov Processes, Characterization and Convergence.* Wiley, New York, 1986.

[14] **Abate J. and Whitt W.** *Transient behavior of regulated Brownian Motion.*, I and II. Adv. Appl. Prob. 19, 560-631, 1987.

[15] **Halfin S.** *Delays in queues, properties and approximations.* Teletraffic Issues in an Advanced Information Society, Proceedings of *ITC-11*, M.Akiyama, ed., Elsevier, Amsterdam 47-52, 1985.

[16] **Boxma O.J., Cohen J.W. and Deng Q.** *Heavy-Tracffic Analysis of the M/G/1 Queue with Priority Classes.* Teletraffic Engineering in a Competitive World, Proceedings of the ITC-16, Edinburgh, UK, North-Holland, Amsterdam, 1999, pp. 1157-1167.

[17] **Abate J. and Whitt W.** *Asymptotics for M/G/1 low-priority waiting-time tail probabilities.* Queueing Systems 25, 173-233, 1997.

[18] **Boxma O.J., Cohen J.W.** *Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions.* Thechnical Report PNA-R9710, CWI, Amsterdam, 1997.

[19] **Klimov G. P., Mishkoy G. K. 1979.** *Prioritetnye sistemy obsluzhivaniya s orientatsiei* (Priority Queueing Systems with Switchover Times). Moscow University Press. In Russian.

[20] **Mishkoy Gh., Giordano S., Bejan A., Benderschi O.** *Priority queueing systems with switchover times: generalized models for QoS and CoS network technologies.* Comput. Sci. J. Moldova nr. 2, vol. 15(44) 217–242, 2007.

[21] **Ciumac V.P., Mishkoi G.K.** *Asymptotics of the queue length of the priority queuing system with orientation.* Kiev, Proceedings of the Sixth USSR-Japan Symposium "Probability theory and Mathematical Statistics, 1991.

Prezentat la 20.11.2008