ASYMPTOTIC RELATIONS IN QUEUEING THEORY*

J.W. COHEN

Technological University, Delft, The Netherlands

Received 16 September 1972

Abstract. For the GI/G/1 queueing system a number of asymptotic results are reviewed. Discussed are asymptotics related to the time parameter for $t \to \infty$, relaxation times, heavy traffic theory, restricted accessibility with large bounds, approximation by diffusion processes, exponential and regular variation of the tail of the waiting time distribution, limit theorems and extreme value theorems.

<table>
<thead>
<tr>
<th>Queueing</th>
<th>restricted accessibility</th>
<th>regular variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI/G/1</td>
<td>diffusion</td>
<td>exponential tail</td>
</tr>
<tr>
<td>asymptotics</td>
<td>weak convergence</td>
<td>waiting time</td>
</tr>
<tr>
<td>relaxation</td>
<td>limit theorems</td>
<td>queue length</td>
</tr>
<tr>
<td>heavy traffic</td>
<td>extreme value theorems</td>
<td>busy period</td>
</tr>
</tbody>
</table>

1. Introduction

It may be said that until the late fifties research in Queueing Theory was mainly directed to the derivation of analytic expressions for the distributions of quantities like waiting time, queue length, busy period and so on, for a great variety of queueing models. Very often the investigator had to be satisfied with the Laplace–Stieltjes transforms or generating functions of the distributions he intended to find. Quite often indeed the researcher could be proud of such results, since they could be only obtained at the cost of hard labour and much ingenuity. It was certainly not unjustified that Kendall [22] in his 1963 review of queueing theory mentioned the Laplacian curtain in queueing theoretic results, although at that time the curtain had many (La)places with holes to peep through.

The appearance of Kingman's paper [23] in 1961 may be considered as a turning point; from that time on, asymptotic methods played an important role in research of queueing theory and provided a large number of applicable results.

Queueing theory offers a large field for the application of asymptotic methods. We mention here:

(i) asymptotics with respect to the time parameter for \( t \to \infty \), and the speed of convergence (the relaxation time);
(ii) asymptotics related to the nearly saturated system, i.e., for the single server model, traffic intensity approaching 1, heavy traffic theory;
(iii) models with restricted accessibility if the restriction becomes weak, e.g., models with a finite number of waiting places \( K \) if \( K \) becomes large;
(iv) the approximation of the actual process by a diffusion process;
(v) asymptotics with respect to the tail of queueing time distributions;
and last but not least:
(vi) the large field of the limit theorems, e.g., central limit theorems and extreme value theorems.

In the present lecture we shall try to review the field of asymptotic relations in queueing theory without aiming at a complete account, e.g., asymptotic results for passage time distributions (cf. e.g. [20]) are hardly discussed. The main ideas and type of results obtained up to now in the existing literature will be considered, restricting the discussion for the sake of simplicity and space to the single server model GI/G/1. Possibly the section on extreme value theorems, being part of the author's present research interest, got somewhat too much attention when compared to other sections, particularly the section on weak convergence.

The list of references, far from being complete, has compiled in such a way that the interested reader has good access to the existing literature. A great deal of the papers mentioned contains a comprehensive list of references to particular topics.

2. The GI/G/1 queueing model

In this paper, the notation as described in [5] will be used. \( \sigma_n, n = 1, 2, ... \) will denote the interarrival time between the \((n-1)\)th and \( n \)th arriving customer; \( \sigma_n, n = 1, 2, ... \), is a sequence of positive i.i.d. variables.
3. The behaviour of the system for \( t \to \infty \), the relaxation time

with

\[
A(\sigma) = \Pr\{\sigma_n < \sigma\}, \quad n = 1, 2, \ldots,
\]

\[A(0+) = 0, \quad \alpha = \mathbb{E}\{\sigma_n\}, \quad \alpha_2 = \mathbb{E}\{\sigma_n^2\},\]

\[
\alpha(s) = \int_0^\infty e^{-st} dA(t), \quad \Re s \geq 0.
\]

The service time of the \( n \)th arriving customer is indicated by \( \tau_n \), the sequence \( \tau_n, n = 1, 2, \ldots \), is a sequence of positive i.i.d. variables with

\[
B(\tau) = \Pr\{\tau_n < \tau\}, \quad n = 1, 2, \ldots,
\]

\[B(0+) = 0, \quad \beta = \mathbb{E}\{\tau_n\}, \quad \beta_2 = \mathbb{E}\{\tau_n^2\},\]

\[
\beta(s) = \int_0^\infty e^{-st} dB(t), \quad \Re s \geq 0.
\]

The families \( \{\sigma_n, n = 1, 2, \ldots\} \) and \( \{\tau_n, n = 1, 2, \ldots\} \) are assumed to be stochastically independent families.

The number of customers present in the system at time \( t \) is denoted by \( x_t \). \( v_t \) is the virtual waiting time at time \( t \). \( w_n \) is the actual waiting time of the \( n \)th arriving customer, the traffic intensity is given by \( \lambda = \beta/\alpha \). Further

\[
\rho_n \equiv \tau_n - \sigma_{n+1}, \quad \mathbb{E}\{\rho_n\} = \beta - \alpha,
\]

\[
\sigma^2 \equiv \text{var} \rho_n = \beta_2 - \beta^2 + \alpha_2 - \alpha^2, \quad b \equiv \sigma^2/(2\alpha).
\]

Finally, \( \Phi(\cdot) \) is the normal distribution with mean zero and variance 1.

3. The behaviour of the system for \( t \to \infty \), the relaxation time

A fundamental problem in queueing theory is the behaviour of the involved stochastic process for \( t \to \infty \). Of basic importance is here Lindley’s theorem [26] concerning the limit of the distribution of \( w_n \) for \( n \to \infty \) (cf. also Loynes [27]). Starting from this theorem, the behaviour of the distributions of \( x_t \) and of \( v_t \) for \( t \to \infty \) can be found. In the case that \( \lambda < 1 \), these limiting distributions exist.
An important question is here the speed of convergence to the limiting distribution. As an example, we mention (cf. [5]); for $t \to \infty,$

$$
\Pr \{\nu_t > 0\} = \begin{cases} 
\frac{a-C(\rho_0)}{\sqrt{2\pi}} \{1 + O(1/t)\}, & a < 1, \\
\frac{\exp(t\rho_0)}{\sqrt{2\pi t}} \{1 + O(1/t)\}, & a = 1, \\
(1-C(\rho_0)) \frac{\exp(t\rho_0)}{t^{3/2} \sqrt{2\pi}} \{1 + O(1/t)\}, & a > 1, 
\end{cases}
$$

(3.1)

where $\rho_0$ is that value of $\rho$ for which $1 - \beta(s) \alpha(\rho - s)$ has a real zero of multiplicity two. Conditions for the validity of (3.1) are:

(i) $\Pr\{\tau_n - \omega_{n+1} \neq 0\} > 0$;

(ii) $\alpha(s)$ and $\beta(s)$ are analytic in a region containing the imaginary axis;

(iii) $A(\cdot)$ and $B(\cdot)$ are not lattice distributions;

(iv) $1 - \alpha(s)$ has no zeros for $\rho_0 < \text{Re } s < 0.$

Here $\rho_0$ is negative and $|\rho_0|^{-1}$ represents the relaxation time; it is a measure of the speed of convergence.

For the M/M/1 model

$$
1/\rho_0 = -\beta/(1 - \sqrt{a})^2, \quad a \neq 1.
$$

(3.2)

The method used for the derivation of (3.1) is saddle point asymptotics.

The condition (iv) is of great interest; it states that the approach to stationarity of the arrival process is faster than that of the queueing process. The latter phenomenon is extremely interesting. A quite unexpected result of a similar type occurred in a study of Callaert [4]. Here the author investigated the speed of convergence to the limiting distribution in a many server queue, the behaviour of which could be described by a birth and death process; the analysis uses the spectral representation of birth and death processes as initiated by Karlin and McGregor [18]. It turned out that for low traffic intensities the relaxation time is that of a many server system without queueing facilities; for higher intensities, however, the relaxation time is analogous with the one for the M/M/1 queue.

For processes in discrete time, we should mention here the concept of geometric ergodicity as introduced by Kendall [21] (cf. also [22]).
An extremely interesting study of the approach to equilibrium has been given by Newell [29] (see also Section 6 of this paper).

4. Heavy traffic theory

The term "heavy traffic" has been coined by Kingman, who in a series of papers (cf. [23, 24]) started the investigation of the behaviour of the stationary actual waiting time distribution for \( a \downarrow 1 \). The main and characteristic result in this field of asymptotics in queueing theory is

\[
\lim_{a \uparrow 1} \Pr\{(1-a) w < x\} = 1 - e^{-x/b}, \quad x > 0,
\]

\( w \) being a stochastic variable whose distribution is the stationary (actual) waiting time distribution. The relation (4.1) may be derived by starting from the Laplace–Stieltjes transform of the waiting time distribution, i.e., by using a purely analytic technique. Other derivations use probabilistic asymptotic methods. Relatively little is known about the speed of convergence of the left-hand side in (4.1) to its limit. Also interesting is the question regarding the behaviour of the relaxation time for \( a \downarrow 1 \).

Important contributions to heavy traffic theory have been made by a number of Russian authors, particularly Prokhorov, and Borovkov [21], and more recently by Whitt [36] and Kyprianou [25]. The approach followed by Prokhorov, Whitt, Kyprianou and partially by Borovkov consists of the investigation of a sequence of queueing systems \( \{Q_m, m = 1, 2, \ldots\} \) with traffic intensity \( a_m \), and \( a_m \uparrow 1 \). They then apply the method of weak convergence. Kyprianou’s paper is very illustrative. His results for the virtual waiting time process may be reviewed as follows: There are 5 cases to be distinguished, viz.

1. \( (a_m - 1) t^{1/2} \to -\infty \),
2. \( \to -\delta \),
3. \( \to 0 \),
4. \( \to \delta \),
5. \( \to \infty \), \( a_m \downarrow a > 1 \), or \( a_m = a > 1 \),

for \( t \to \infty \), \( m \to \infty \) with \( a_m \to a \) and \( \delta > 0 \).
(1) Heavy traffic: for \( t \to \infty, a_m \uparrow 1 \),
\[
\lim \Pr \{ \nu_t^{(m)}(1-a_m) \leq x \} = 1 - e^{-x^b}, \ x \geq 0;
\]
(2)-(4) for \( t \to \infty, a_m \uparrow 1 \),
\[
\lim \Pr \{ \nu_t^{(m)}(2bt)^{-1/2} \leq x \} = \int_0^x g(z) \, dz, \ x \geq 0,
\] where
\[
g(z) = \sqrt{\frac{2}{\pi t}} \exp\left[-\left(\frac{1}{2} t^{-1}(z-\delta^t)\right)^2\right]
+ 2\delta' \exp(2\delta' z) \{ 1 - \Phi(t^{-1/2}(z + \delta^t)) \},
\]
with \( \delta' = (2b)^{-1/2} \delta, 0 \leq |\delta| < \infty \). These cases show a behaviour like Brownian motion with reflecting barrier at zero, with drift proportional to \( \delta^t \) and sign that of \( \delta \). In particular for \( \delta = 0 \), we have:
(3) for \( t \to \infty, a_m \uparrow 1 \),
\[
\lim \Pr \{ \nu_t^{(m)}(2bt)^{-1/2} \leq x \} = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-\frac{1}{2} z^2) \, dz, \ x \geq 0.
\]
(5) for \( t \to \infty, a_m \downarrow a > 1 \),
\[
\lim \Pr \{ \nu_t^{(m)} - (a_m - 1)t \} (2bt)^{-1/2} \leq x \} = \Phi(x), \ -\infty < x < \infty.
\]

The results above hold under the conditions:
(i) \( E \{ \tau_n^{(m)} - \sigma_n^{(m)} \} \to \beta - \alpha, \ m \to \infty \),
(ii) \( \text{var} \{ \tau_n^{(m)} - \sigma_n^{(m)} \} \to \sigma^2 < \infty, \ m \to \infty \),
(iii) \( E \{ |\tau_n^{(m)} - \sigma_n^{(m)}|^2 \epsilon \} \) is bounded in \( m \) for some \( \epsilon > 0 \).

Concerning the diffusion approach to heavy traffic theory, see Section 6.

5. Queueing systems with restricted accessibility

A queueing system has restricted accessibility if the entry or non-entry of an arriving customer depends on the state of the queueing system at the moment of his arrival. If the customer is not admitted, he
disappears and never returns; examples are systems with a limited number of waiting places, say \( m \), or a system which accepts customers only if their actual waiting time is less than \( K \) (cf. [6]).

For \( m \to \infty \), and similarly for \( K \to \infty \), the classical model is again obtained. It is of great practical as well as theoretical interest to obtain the asymptotic relations of the relevant distributions for \( m \) or \( K \to \infty \); the more so since systems with restricted accessibility possess stationary distributions for all values of the traffic parameter, and the behaviour of these distributions for \( m \to \infty \), or \( K \to \infty \), if the unrestricted system is unstable, is important for applications.

Literature on asymptotic problems of this type is relatively scarce; as examples we mention here the study by Tomko [33] and by Cohen [8]. A related asymptotic problem concerning a many server queue with the number of servers approaching \( \infty \) and the offered traffic also increasing to \( \infty \) has been discussed by Iglehart [14].

As an illustration, we quote a result from Tomko. Consider the M/M/1 queue with \( m \) waiting places, let \( w^{(m)} \) be a stochastic variable whose distribution is the stationary (actual) waiting time distribution of this system. Tomko proved the following:

(i) for \( a < 1 \),

\[
|\Pr\{w^{(m)} < x\} - W(x)\} | \leq c_1 a^m \quad \text{for } x \geq 0, \ m \geq 1,
\]

\( W(\cdot) \) being the stationary waiting time distribution for the unrestricted system M/M/1;

(ii) for \( a = 1 \),

\[
|\Pr\{w^{(m)} \leq x(m-1)\} - V(x)\} | \leq c_2 / \sqrt{m} \quad \text{for } x \geq 0, \ m \geq 1,
\]

\( V(x) \) being the uniform distribution on \([0, 1]\);

(iii) for \( a > 1 \),

\[
|\Pr\{(w^{(m)} - m) / \sqrt{m} \leq x\} - \Phi(x)\} | \leq c_3 / \sqrt{m}
\]

\[
\text{for } -\infty < x < \infty, \ m \geq 1;
\]

here \( c_1, c_2, c_3 \) are constants independent of \( x \) and \( m \).
From [8], we quote the following further results:

(iv) For the queueing system $K_m/K_n/1$ with $a = 1, \gamma_{v_0}$ the duration of the busy cycle initiated by an arriving customer with initial waiting time $\nu$ we have for $0 \leq q \leq 1, x \geq 0$,

$$\lim_{K \to \infty} \Pr \{ bK^{-2} \gamma_{qK,0} < x, \sup \{ v_{\tau} < K : 0 < \tau < \gamma_{qK,0} \} \} =$$

$$1 - q + 4 \sum_{h=1}^{\infty} (-1)^h (\pi h)^{-1} \exp(-x \pi^2 h^2) \sin((1-q)\pi h);$$

(v) for the queueing system $K_m/K_n/1$ with uniformly bounded virtual waiting time (cf. [5], also finite dam, capacity $K$), with $a = 1, \gamma_{v_0}$ defined as in (iv), we have for $0 \leq q < 1, x \geq 0$,

$$\lim_{K \to \infty} \Pr \{ bK^{-2} \gamma_{qK,0} < x \} =$$

$$1 - 2 \sum_{h=-\infty}^{\infty} \frac{\cos((2h+\frac{1}{2})\pi(1-q))}{(2h+\frac{1}{2})\pi} \exp\{-x(2h+\frac{1}{2})^2 \pi^2 \}. $$

So far for the examples. Sufficient material is available in the literature to start a study of asymptotic relations of the types indicated above not only for the stationary state but also for the time dependent situation.

6. Diffusion approximations

Gaver [12] and Newell [28, 29] analysed the M/G/1 system with $a \sim 1$ by using diffusion theory. The aim of their analysis is the investigation of the time dependent behaviour of nearly saturated or slightly oversaturated queues taking into account the effect of time dependent traffic parameters. These studies may be considered as highlights in the field of applied queueing theory. The results obtained are extremely well suited for direct application.

To provide a flavor of their approach, we start from Takács' integro-differential equation for the M/G/1 system with time independent parameters (the last assumption is not at all essential)

$$(6.1) \quad \frac{\partial}{\partial t} V(t, x) = -\frac{1}{a} V(t, x) + \frac{\partial}{\partial x} V(t, x) + \frac{1}{a} \int_{y=0}^{\infty} V(t, x-y) dB(y),$$

for $t \geq 0, x \geq 0$, with
§ 6. Diffusion approximations

(6.2) \[ V(0^+, x) = \begin{cases} 1 & \text{for } x > v > 0, \\ 0 & \text{for } x \leq v. \end{cases} \]

\[ V(t, x) = 0 \quad \text{for } x < 0, \ t \geq 0, \]

and

(6.3) \[ V(t, x) = \Pr \{ V_t < x | V_{0^+} = v \}. \]

Assume that we may write for \( x \geq y \),

(6.4) \[ V(t, x-y) = V(t, x) - y \frac{\partial}{\partial x} V(t, x) + \frac{y^2}{2!} \frac{\partial^2}{\partial x^2} V(t, x) \]

\[ - \frac{y^3}{3!} \frac{\partial^3}{\partial x^3} V(t, x) + \ldots. \]

Substituting (6.4) into (6.1) and taking \( x \) so large that we may write

\[ \int_0^x d\beta(y) \sim 1, \quad \int_0^x y^k d\beta(y) \sim \beta_k, \quad k = 1, 2, 3, \ldots \]

we get

(6.5) \[ \frac{\partial}{\partial t} V(t, x) = \left\{ (1-a) \frac{\partial}{\partial x} + \frac{\beta_2}{2a} \frac{\partial^2}{\partial x^2} - \frac{\beta_3}{6a} \frac{\partial^3}{\partial x^3} + \ldots \right\} V(t, x). \]

Putting

(6.6) \[ \tau = t/T, \quad \xi = x/X, \quad \nu(\tau, \xi) = V(\tau T, \xi X), \]

and taking the unit of time \( T \) and the scale unit \( X \) such that

\[ |1-a| T/X = 1, \quad (\beta_2/2a) T/X^2 = 1, \]

that is,

(6.7) \[ X = |1-a|^{-1} (\beta_2/2a), \quad T = (1-a)^{-2} (\beta_2/2a), \]

it follows from (6.5) that

(6.8) \[ \frac{\partial}{\partial \tau} \nu(\tau, \xi) = \left\{ \text{sgn}(1-a) \frac{\partial}{\partial \xi} + \frac{\partial^2}{\partial \xi^2} - \frac{1}{3} \beta_3 \frac{a}{\beta_2^2} (1-a) \frac{\partial^3}{\partial \xi^3} + \ldots \right\} \nu(\tau, \xi). \]
Suppose that \( a \sim 1 \) and that \( B(x) = 1 \) for \( x > X \) (cf. the derivation of (6.5)). Neglecting in (6.8) all terms containing the factor \((1-a)\) and its higher powers (i.e., looking for solutions of (6.8) for which third and higher derivatives with respect to \( \xi \) are bounded), we obtain the diffusion equation

\[
\frac{\partial}{\partial \tau} u(\tau, \xi) = \left\{ \text{sgn}(1-a) \frac{\partial}{\partial \xi} + \frac{\partial^2}{\partial \xi^2} \right\} u(\tau, \xi),
\]

with \( \text{sgn}(1-a) = 0 \) for \( a = 1 \). The solution of (6.9) can be easily obtained for the conditions (6.2). For details and particularly for the quality of the numerical approximation see [12].

If \( a \sim 1 \) and \( 1-a > 0 \), then (6.9) possesses a stationary solution \( u(\xi) \), the limit for \( \tau \to \infty \) of the time dependent solution \( u(\tau, \xi) \). This solution satisfies

\[
0 = \left\{ \frac{\partial}{\partial \xi} + \frac{\partial^2}{\partial \xi^2} \right\} u(\xi), \quad u(\xi) = 0 \text{ for } \xi \leq 0, \quad u(\xi) \to 1 \text{ for } \xi \to \infty.
\]

From (6.10) it is readily found that

\[
v(\xi) = V(\infty, (1-a)^{-1}(\beta_2/2\alpha)\xi) = \lim_{t \to \infty} \Pr \{ (1-a)(\beta_2/2\alpha)^{-1} \nu_t \to \xi \} \\
= 1 - e^{-\xi}, \quad \xi \geq 0,
\]

which represents the heavy traffic solution for the M/G/1 system. Obviously, the solution (6.11) has bounded derivatives.

It is of some interest to consider the solution of

\[
0 = \left\{ \frac{\partial}{\partial \xi} + \frac{\partial^2}{\partial \xi^2} - \gamma \frac{\partial^3}{\partial \xi^3} \right\} u(\xi),
\]

with

\[
\gamma = \frac{2}{3} \beta_2^{-2} \beta_3 \alpha (1-a) \quad \text{for} \quad a < 1, \: a \sim 1,
\]

and with the same boundary conditions as (6.10). The general solution in this case is given by

\[
u(\xi) = C_1 + C_2 \exp(\lambda_1 \xi) + C_3 \exp(\lambda_2 \xi),
\]

with \( \lambda_1, \lambda_2 \) the zeros of \( \lambda^2 - \lambda/\gamma - 1/\gamma \). Obviously, one of the zeros is positive (say \( \lambda_2 \)), so that in view of the norming condition, \( C_3 = 0 \). For
the other zero \( \lambda_1 \), we have in first approximation \( \lambda_1 = 1 + \gamma \). Applying the conditions of (6.10), we obtain

\[
v(\xi) = 1 - \exp \left\{ - (1 - \gamma) \xi \right\}, \quad \xi \geq 0.
\]

The latter result, when compared with the expression (6.11), gives a qualitative insight concerning the diffusion approximation of Takács' integro-differential equation.

The heuristic approach sketched above reveals something of the power of the diffusion approximation; in particular, the expressions for the time unit and scale unit are important, and the relation between \( T \) and the relaxation time given by (3.2) should be noted.

Finally we mention here Keilson's papers [19] concerning diffusion and birth-death processes, which contain results and ideas that are valuable for the approximation of queueing processes by diffusion processes. This field is still not sufficiently explored; in particular, more numerical results concerning the approximation are badly needed.

7. The approach by weak convergence methods

In a series of interesting papers, Iglehart and Whitt (cf. [15, 17, 35, 37]) apply weak convergence concepts for the derivation of limit theorems in queueing theory. Their approach is very fruitful and yields a vast number of new results as well as new derivations of known limit theorems for queues (cf. e.g. [13, 32]); as examples of known theorems which they proved we mention the following:

(i) For \( a > 1 \),

\[
\lim_{n \to \infty} \Pr \{(an^{1/2})^{-1} [w_n - n(\beta - \alpha)] < x \} = \Phi(x), \quad -\infty < x < \infty.
\]

(ii) For \( a = 1 \),

\[
\lim_{n \to \infty} \Pr \{(an^{1/2})^{-1} w_n < x \} = \sqrt{2} \int_{0}^{x} \exp(-\frac{1}{2} y^2) \, dy, \quad x \geq 0.
\]
\[
\lim_{n \to \infty} \Pr \{ \max \{(n^{1/2} a)^{-1} w_k : 0 \leq k \leq n \} < \cdot \} = \\
= 1 - \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp[-\pi^2 (2k-1)^2 (8x^2)^{-1}], \quad x \geq 0.
\]

Analogous theorems for the virtual waiting time \( v_t \) and queue length \( x_t \) can be derived.

Interesting new results are discussed in Iglehart's paper [15], from which we quote the following typical results for the case \( a < 1 \).

(iii) For \( a < 1 \),
\[
\lim_{t \to \infty} \Pr \{ (c t^{1/2})^{-1} \left[ \int_0^t x_s \, ds - t \bar{E}(x) \right] < x \} = \Phi(x), \quad -\infty < x < \infty,
\]
\[
\limsup_{t \to \infty} \left\{ \left[ 2t \log t \right]^{-1/2} \left[ \int_0^t x_s \, ds - t \bar{E}(x) \right] \right\} = c \quad \text{a.e.,}
\]
\( c \) being a positive constant, and \( \bar{E}(x) \) the mean of the stationary queue length distribution.

(iv) For \( a < 1 \), functional central limit theorems are derived as follows: with
\[
W_n(t) \overset{\text{def}}{=} \sum_{k=0}^{nt} w_k, \quad V_u(t) \overset{\text{def}}{=} \int_0^t v_s \, ds,
\]
it is shown that for \( n \to \infty \),
\[
\{ a_1 m^{-1} n \}^{-1/2} \left[ W_n(t) - n t \bar{E}(w/\beta) \right] \Rightarrow \xi,
\]
and for \( u \to \infty \),
\[
\{ a_0^2 m^{-1} u \}^{-1/2} \left[ V_u(t) - u E(v) \right] \Rightarrow \xi,
\]
where \( \Rightarrow \) stands for weak convergence, \( E\{w\} \) and \( E\{v\} \) are the averages of the stationary actual and virtual waiting distributions, respectively, \( a_1 \), \( a_0 \) and \( m \) are constants, and \( \xi \) stands for the standard Brownian motion process.

Whitt [37] also presents the following result concerning the rate of convergence: If for \( a = 1 \), \( E\{ |x_n - E_{n+1}^{2+\delta} | \} < \infty \) for some \( \delta > 0 \), there exists a constant \( A \) such that for all \( x \geq 0, n \geq 0 \),
\[
\]
§7. The approach by weak convergent methods

\[ \Pr \{ an^{1/n} \to x \} = \sqrt{\frac{2}{\pi}} \int_0^x \exp \left( -\frac{1}{2} y^2 \right) dy \leq A(\log n)^{3/2} n^{-2/3}, \]

with

\[ \lambda = \left( 1 + \frac{1}{2} a \right) (a + 3)^{-1} < \frac{1}{2}, \quad \mu = \min \{ a, \frac{1}{2} (1 + \frac{1}{2} a) (a + 3)^{-1} \}. \]

To provide an idea about the application of the weak convergence method (cf. Billingsley [1]), we give a sketch of the derivation of the result (ii) above as presented by Whitt [37]. With \( s_0 \overset{\text{def}}{=} 0, s_n \overset{\text{def}}{=} \tau_2 + \ldots + \tau_n - \sigma_{n+1} \), it is well-known that if \( w_1 = 0 \),

\[ w_{n+1} = s_n - \min \{ s_k : 0 \leq k \leq n \}, \quad n \geq 1. \]

Consider a sequence \( Q^{(m)}, m = 1, 2, \ldots \), of single server queues GI/G/1 with actual waiting times \( \{ w_n^{(m)}, n = 1, 2, \ldots \} \), \( m = 1, 2, \ldots \), and

\[ s_n^{(m)} = \tau_1^{(m)} - \sigma_2^{(m)} + \ldots + \tau_n^{(m)} - \sigma_{n+1}^{(m)}, \]

\[ \mathbb{E} \{ \tau_n^{(m)} - \sigma_{n+1}^{(m)} \} = \mathbb{E} \{ \tau_n^{(m)} - \sigma_{n+1}^{(m)} \} = 0. \]

Define

\[ \xi_m(t) \overset{\text{def}}{=} (\sigma m^{1/2})^{-1} s_{[mt]}^{(m)}, \quad \eta_m(t) \overset{\text{def}}{=} (\sigma m^{1/2})^{-1} w_{[mt]}^{(m)}, \]

for \( m = 1, 2, \ldots \), and \( t \in [0, 1] \). The processes \( \xi_m \) and \( \eta_m \) are stochastic and elements of the function space \( D = D[0, 1] \) of all real valued right continuous functions on \( [0, 1] \) which have limits from the left. This space is metrized by the Skohorod metric (cf. [1]). \( P_m \to P \) denotes weak convergence of a sequence of probability measures defined on a \( \sigma \)-field in \( D \) to a probability measure on this \( \sigma \)-field. A sequence \( X_m \) of elements of \( D \) converges in distribution to \( X \), \( X_m \Rightarrow X \in D \), if the distribution \( P_m \) of \( X_m \) converges weakly to the distribution \( P \) of \( X \).

Whitt shows that if \( \xi_m \Rightarrow \xi \) with \( \xi \) an arbitrary element of \( D \) then

\[ \eta_m \Rightarrow f(\xi) \]

with \( f : D \to D \), the uniform continuous function defined for any \( \chi \in D \) by

\[ f(\chi)(t) = \chi(t) - \inf \{ \chi(\tau) : 0 \leq \tau \leq t \}, \quad 0 \leq t \leq 1. \]

By applying Donsker's theorem (see [1]), it follows that \( \xi = \xi \), the Brownian motion in \( D \) and that \( f(\xi) = | \xi | \), the one-dimensional Bessel
process in $D$, so that for $a = 1$ and $m \to \infty$,
\[
(\sigma m^{1/2})^{-1} \frac{W(m)}{mt} \Rightarrow |\xi|.
\]

For $t = 1$, this leads to the result (ii) above. Along the same lines, it is shown that for $a > 1$ and $m \to \infty$,
\[
(\sigma m^{1/2})^{-1} \left[ \frac{W(m)}{mt} - (a - 1)mt \right] \Rightarrow \xi,
\]
from which the result (i) above follows by taking $t = 1$.

8. Regular variation and extreme value theorems

To obtain extreme value theorems for queueing processes, the behaviour of the tails of the relevant distributions should be known. A first result in this direction has been obtained by Smith [31], who proved that the Laplace–Stieltjes transform $\omega(s)$ of the stationary waiting time distribution ($a < 1$) is a rational function if the Laplace–Stieltjes transform of the service time distribution is a rational function (cf. [5, 30]); the converse statement is true if $\omega(s)$ has no zeros for $\text{Re } s > 0$. Roszberg [30] proved that $\omega(s)$ is rational if and only if

\[
E\{\exp\{-s(\tau_n - \sigma_{n+1})\} | \tau_n > \sigma_{n+1}\}
\]
is rational, and that the Laplace–Stieltjes transform of the idle period is rational if and only if

\[
E\{\exp\{-s(\tau_n - \sigma_{n+1})\} | \tau_n < \sigma_{n+1}\}
\]
is rational; here ($A$) stands for the indicator function of the event $A$.

\[
\omega(s) = [1 - \beta(s) \alpha(-s)]^{-1} e^{-B} \{1 - E\{e^{\tau s}\}\}, \quad \text{Re } s = 0,
\]
which holds if

\[
B \overset{\text{def}}{=} \sum_{n=1}^{\infty} n^{-1} \Pr\{\tau_n > \sigma_{n+1}\} < \infty,
\]

it is readily seen by analytic continuation that if the service time distri-
bution $B(\cdot)$ has an exponential tail, then so has $W(\cdot)$. This relation has a number of far-reaching conclusions and leads to important limit theorems.

Regular variation of the tail of $B(\cdot)$ is another type of tail behaviour which, as will be discussed below, leads to interesting limit theorems.

A positive function $f(t)$, $t \geq 0$, is said to be regularly varying at infinity with exponent $k$, $-\infty < k < \infty$, if for every $x > 0$,

$$\lim_{t \to \infty} \frac{f(xt)}{f(t)} = x^k.$$ 

In the case that $k = 0$, $f(t)$ is said to be slowly varying at infinity. If $L(t)$ is slowly varying at infinity, then the function $t^k L(t)$ is of regular variation at infinity with exponent $k$.

Recently, the concept of regular variation got much attention in probability theory (cf. [10, 11]). In queueing theory, we have for the GI/G/1 queue the interesting result that the tail of the stationary actual waiting time distribution $W(\cdot)$ has regular variation at infinity if and only if the tail of the service time distribution $B(\cdot)$ varies regularly at infinity.

$$1 - B(t) = k(\beta/t)^{k+1} L(t) \Rightarrow 1 - W(t) \sim a(1-a)^{-1}(\beta/t)^k L(t),$$

$$t \to \infty,$$

$$1 - W(t) = a(1-a)^{-1}(\beta/t)^k L(t) \Rightarrow 1 - B(t) \sim k(\beta/t)^{k+1} L(t),$$

$$t \to \infty,$$

for $k > 0$ and $L(\cdot)$ slowly varying at infinity. The if part is due to Borovkov [3], the only if part to the author. Both statements also hold if $W(\cdot)$ is replaced by the stationary virtual waiting time distribution.

By $w_{i \max}^1$ we shall denote the supremum of the actual waiting times during the $i$th busy period and by $W_n$, the maximum of $w_{i \max}^1$, $i = 1, 2, \ldots, n$, all busy periods starting with an empty system. Further $w_{n \max}^1$ is the maximum of $w_m^1$, $m = 1, \ldots, n$, with $w_1 = 0$.

We shall discuss a number of asymptotic relations and limit theorems for the quantities just introduced. We shall restrict this to the M/G/1 queueing system, since for this system the most complete results are available in literature.

**Case 1**: $a < 1$.

(A) Let $1 - B(t) \sim c_1 \exp(-t/c_2)$ for $t \to \infty$, $c_1 > 0$, $c_2 > 0$. Then:
(i) $\Pr\{w^{(i)}_{\text{max}} < t\}$ has an exponential tail,
\[
1 - \Pr\{w^{(i)}_{\text{max}} < t\} \sim b_2 e^{e t} \quad \text{for} \quad t \to \infty, \quad b_2 = \frac{\alpha - \beta}{\alpha + \beta'(\epsilon)} \frac{\alpha e}{1 - \alpha e},
\]
with $\epsilon$ the zero of $\beta(s) + \alpha s - 1$, $\Re s < 0$, which is nearest to the imaginary axis ($\epsilon$ is uniquely determined in this way);
(ii) $\lim_{n \to \infty} \Pr\{\beta^{-1} W_n < (-\epsilon \beta)^{-1} [x + \log(n b_2)]\} = \exp[-e^{-x}], \quad -\infty < x < \infty$;
(iii) $W_n (\beta \log n)^{-1} \to - (\epsilon \beta)^{-1}$ in probability;
(iv) $\lim_{n \to \infty} \Pr\{\beta^{-1} w^+_n < (-\epsilon \beta)^{-1} [x + \log(n(1-a)b_2)]\} = \exp[-e^{-x}], \quad -\infty < x < \infty$;
(v) $w^+_n (\beta \log n)^{-1} \to - (\epsilon \beta)^{-1}$ in probability.

(B) Let $1 - B(t) \sim b(\beta(t))k^k$ for $t \to \infty$, $b > 0$, $k > 1$. Then:
(i) $1 - \Pr\{w^{(i)}_{\text{max}} < t\} \sim (1-a)^{-1} b(\beta(t))k^k$ for $t \to \infty$,
\[
1 - B(t) \sim (\beta(t))^k L(t) \sim \beta \frac{dW(t)}{dt} \sim (1-a)^{-1} (\beta(t))^k L(t), \quad \text{for} \quad t \to \infty \quad \text{and} \quad L(\cdot) \text{ slowly varying at infinity};
\]
(ii) $\lim_{n \to \infty} \Pr\{(bn)^{-1} (1-a)^{1/k} \beta^{-1} < x\} = \exp[-x^{-k}], \quad x > 0$;
(iii) $\lim_{n \to \infty} \Pr\{(bn)^{-1/2} w^+_n (\beta^+)^{-1/2} < x\} = \exp[-x^{-k}], \quad x > 0$.

Case 2: $a = 1$, $\beta_2 < \infty$.
(i) $1 - \Pr\{w^{(i)}_{\text{max}} < t\} \sim \beta/t$ for $t \to \infty$;
(ii) $\lim_{n \to \infty} \Pr\{(n \beta)^{-1} W_n < x\} = \exp[-x^{-1}], \quad x > 0$;
(iii) $\lim_{n \to \infty} \Pr\{(n \beta_2)^{-1/2} w^+_n < x\} = 1 - (4/\pi) \Sigma_{k=1}^\infty (-1)^k (2k + 1)^{-1} \exp[-\pi^2 (2k-1)^2 (3x^2)^{-1}], \quad x \geq 0$.

Case 3: $a > 1$, $\beta_2 < \infty$.
\[
\lim_{n \to \infty} \Pr\{[n(\beta_2 - \beta^2 + a^2)]^{-1/2} [w^+_n - \beta n(a-1)] < x\} = \Phi(x), \quad -\infty < x < \infty.
\]

8.1. Comments

(1) The relations 1A (i)–(iii), 2(i)–(ii) have been obtained by the author (cf. [5]), for the queueing systems $K_m/K_n/1$, $G/K_n/1$, $K_m/G/1$, similar relations have been obtained in [8] and [34]. Iglehart generalized cases 1A (i)–(v) and 2(iii) to the GI/G/1 system. Actually he proved instead of 2(iii) the stronger functional limit theorem (see the end of section 7).

(2) The relations 1B (i)–(iii), obtained by the author (cf. [7]), are at present only known for the M/G/1 system. However, the author strongly conjectures that they are also valid for the GI/G/1 system with $1-a$ in Case 1B (ii) replaced by $e^{-B}$.
References

(3) Relations of the type above have also been derived for the virtual waiting time process \( \{v_t, t \in [0, \infty) \} \) (cf. [5, 8, 16, 34]). For the queue-length process \( \{w_t, t \in [0, \infty) \} \), the situation is different. For this process, results analogous to Case 1A are known for the M/G/1 and G/M/1 systems with the exception of Cases 1A(ii)–(iv), where the limits do not exist, but tight upper bounds are available for limsup and liminf. The relations analogous to 1B(i)–(iii) for the M/G/1 \( x_t \)-process have been proved in [7], and those analogous to 2(ii)–(iii) and 3(i) for the GI/G/1 \( x_t \)-process appear in Iglehart [16].

(4) It is noted that if \( a < 1 \) and \( 1 - B(t) \sim b(t/\beta)^k \) for \( t \to \infty, k > 1 \), then \( x_n^* = \max\{x_m : 1 \leq m \leq n\} \) (\( x_m \) being the service time of the \( n \)th arriving customer) has the same extreme value distribution as \( w_n^* \) with the same norming constants (cf. Case 1B(iii)).

(5) Finally, we mention as an open problem the heavy traffic approximation of the relations in Cases 1A, B, that is, the study of these relations for \( a \uparrow 1 \).

References

[34] E.M. van Vugt, Limit theorems for the distribution of the actual waiting time in a K_m/G/1 and a G/K_n/1 queueing system, Master Thesis, Technol. Univ., Delft (1969) (in Dutch).
[37] W. Whitt, Complements to heavy traffic limit theorems for the GI/G/1 queue, J. Appl. Probability 9 (1972) 185–191.