ANNALES DE L'I. H. P., SECTION B

J. W. COHEN

On the tail of the stationary waiting time distribution and limit theorems for the M/G/1 queue

Annales de l'I. H. P., section B, tome 8, nº 3 (1972), p. 255-263

<http://www.numdam.org/item?id=AIHPB_1972__8_3_255_0>

Ann. Inst. Henri Poincaré,
Vol. VIII, n° 3, 1972, p. 255-263.

Section B :
Calcul des Probabilités et Statistique.

On the tail of the stationary waiting time distribution and limit theorems for the M/G/1 queue

by

J. W. COHEN

SUMMARY. — For the M/G/1 queueing system with traffic intensity $a < 1$ it will be shown that if the tail of the service time distribution is a regular varying function at infinity, so is the tail of the stationary waiting time distribution $W(t)$, also $dW(t)/dt$ varies regularly at infinity. Applying these results extreme value theorems are derived for the actual waiting time of the n^{th} arriving customer, for the virtual waiting time and for the supremum of the virtual waiting time, for the maximum of the actual waiting time and for that of the number of simultaneously present customers during a busy cycle.

Keywords: queueing theory, regular variation, supremum of actual, virtual waiting time, limit theorems, extreme value theorems.

1. For the M/G/1 queueing system with service time distribution $B(t)$, $B(0+) = 0$ and traffic intensity $a = \beta/\alpha < 1$, where β is the average service time and $1/\alpha$ the average interarrival rate, the Laplace-Stieltjes transform of the stationary waiting time distribution $W(t)$ is given by

$$\int_{0-}^{\infty} e^{-\rho t} dW(t) = (1-a)\frac{\alpha\rho}{\beta(\rho)+\alpha\rho-1}, \qquad \mathrm{Re}\,\rho \geqq 0, \tag{1}$$

$$\beta(\rho) = \int_0^{\infty} e^{-\rho t} dB(t).$$

Starting from this relation it will be shown that if $1 - B(t)$ is a function of regular variation at infinity so is $W(t)$ and conversely. In a forthcoming paper [*1*], it will be shown that this statement is also valid for the GI/G/1 queueing system. It is shown below that for the M/G/1 system regular variation of $1 - W(t)$ at infinity implies that the tails of the distribution of $\underline{v}_{\max}$ and $\underline{w}_{\max}$ also vary regularly at infinity, here $\underline{v}_{\max}$ is the supremum of the virtual waiting time during a busy cycle, whereas $\underline{w}_{\max}$ is the supremum of the actual waiting times of customers served in one busy cycle. Using these results limit theorems are obtained for

$$\underline{W}_n \overset{\text{def}}{=} \max_{1 \leqq m \leqq n} \underline{w}^{(m)}, \qquad \underline{V}_n \overset{\text{def}}{=} \max_{1 \leqq m \leqq n} \underline{v}^{(m)},$$
$$\underline{w}_n^+ \overset{\text{def}}{=} \max_{1 \leqq m \leqq n} \underline{w}_m, \qquad \underline{v}_t^+ \overset{\text{def}}{=} \sup_{0 \leqq \tau \leqq t} \underline{v}_\tau,$$

for $n \to \infty$ and $t \to \infty$. Here, $\underline{w}^{(m)}(\underline{v}^{(m)})$ is the maximum actual (virtual) waiting tome in the m^{th} busy cycle, so that $\underline{W}_n$ is the maximum actual waiting time during the first n busy cycles, $\underline{V}_n$ the supremum of the virtual waiting time during the first n busy cycles; $\underline{w}_m$ is the actual waiting time of the m^{th} arriving customer, $\underline{v}_t$ is the virtual waiting time at time t, so that $\underline{w}_n^+$ is the maximum actual waiting time of the first n arriving customers, whereas $\underline{v}_t^+$ is the supremum of the virtual waiting time during $[0, t]$. It is always assumed that $\underline{w}_1 = 0$.

In the case that $B(t)$ has an exponential tail limit theorems for the quantities mentioned above have been obtained by Cohen [*2*], [*3*], van Vugt [*4*] and Iglehart [*5*].

2. For $F(t)$ a defective probability distribution with

$$F(0) = 0, \qquad F(\infty) = \delta, \qquad 0 < \delta < 1, \tag{2}$$

and

$$U(t) = \sum_{n=1}^{\infty} F^{n*}(t), \qquad t > 0, \tag{3}$$
$$= 0, \qquad t < 0,$$

we have (cf. [*6*]) the following result:

LEMMA. — For $\rho \geqq 0$ and $S(t)$ a slowly varying function at infinity

$$U(\infty) - U(t) \sim \frac{\delta}{1 - \delta} t^{-\rho} S(t) \qquad \text{for} \qquad t \to \infty,$$

if and only if

$$F(\infty) - F(t) \sim \delta(1 - \delta) t^{-\rho} S(t) \qquad \text{for} \qquad t \to \infty.$$

For the right-hand side of (1) we have, since $a < 1$,

$$(4)\quad (1-a)\frac{\alpha\rho}{\beta(\rho)+\alpha\rho-1} = (1-a)\sum_{n=0}^{\infty} a^n(\beta\rho)^{-n}\{1-\beta(\rho)\}^n, \quad \operatorname{Re}\rho \geqq 0.$$

Since

$$\int_0^\infty e^{-\rho t}d\mathrm{H}(t) = \frac{1}{\beta\rho}\{1-\beta(\rho)\}, \quad \operatorname{Re}\rho \geqq 0,$$

where

$$(5)\qquad \begin{aligned} \mathrm{H}(t) &= \frac{1}{\beta}\int_0^t \{1-\mathrm{B}(\tau)\}\,d\tau, && t>0,\\ &= 0, && t<0, \end{aligned}$$

the well-known relation

$$(6)\qquad \begin{aligned} \mathrm{W}(t) &= 0, && t<0,\\ &= (1-a)\sum_{n=0}^{\infty} a^n \mathrm{H}^{n*}(t), && t>0, \end{aligned}$$

follows from (1), (4) and (5).

With

$$(7)\qquad \delta = a, \qquad \mathrm{F}(t) = a\mathrm{H}(t),$$

we have from (3) and (6),

$$(8)\qquad 1-\mathrm{W}(t) = (1-a)\{\mathrm{U}(\infty)-\mathrm{U}(t)\}, \qquad t>0.$$

THEOREM. — For the queueing system M/G/1 with traffic intensity $a < 1$ and $\mathrm{L}(t)$ slowly varying at infinity the following three relations imply each other for $k > 0$,

$$(9)\qquad 1-\mathrm{B}(t) \sim k(\beta/t)^{k+1}\mathrm{L}(t), \qquad t\to\infty,$$

$$(10)\qquad 1-\mathrm{W}(t) \sim \frac{a}{1-a}(\beta/t)^k\mathrm{L}(t), \qquad t\to\infty,$$

$$(11)\qquad \beta\frac{d\mathrm{W}(t)}{dt} \sim \frac{ak}{1-a}(\beta/t)^{k+1}\mathrm{L}(t), \qquad t\to\infty.$$

Proof. — If (9) holds then (cf. [7], p. 273),

$$(12)\quad 1-\frac{1}{\beta}\int_0^t\{1-\mathrm{B}(\tau)\}\,d\tau \sim \frac{1}{\beta}\int_t^\infty\{1-\mathrm{B}(\tau)\}\,d\tau \sim (\beta/t)^k\mathrm{L}(t), \quad \text{for } t\to\infty,$$

hence from (5) and (7),

$$F(\infty) - F(t) \sim a(\beta/t)^k L(t) \quad \text{for} \quad t \to \infty,$$

so that by applying the lemma above the relation (10) follows.

Since $H(t)$ has a density, $H^{n*}(t)$, $n = 1, 2, \dots$, has a density, $h_n(t)$, $n = 1, 2, \dots$, and from

$$\begin{aligned} h_1(t) &= 0, && t < 0, \\ &= \frac{1}{\beta}(1 - B(t)), && t > 0, \end{aligned}$$

$$h_{n+1}(t) = \frac{1}{\beta}\int_0^t h_n(t-\tau)\{1 - B(\tau)\}\,d\tau, \qquad t > 0, \qquad n = 1, 2, \dots,$$

it is readily seen that $h_n(t)$ is uniformly bounded by $1/\beta$ for $t \geqq 0$, $n = 1, 2, \dots$. Hence from (6), $dW(t)/dt$ exists for $t > 0$, since the series

$$\frac{d}{dt}W(t) = (1 - a)\sum_{n=1}^{\infty} a^n h_n(t), \tag{13}$$

converges uniformly in $t \in [0, \infty)$.

From

$$-\log W(t) = \sum_{n=1}^{\infty} \frac{1}{n}\{1 - W(t)\}^n, \qquad 0 \leqq t < \infty,$$

it follows

$$|-\log W(t) - \{1 - W(t)\}| < \{1 - W(t)\}^2/W(t), \qquad 0 < t < \infty.$$

Hence from (10) and the last inequality

$$-\log W(t) \sim \frac{a}{1-a}(\beta/t)^k L(t) \qquad \text{for} \qquad t \to \infty. \tag{14}$$

Denote by $\underline{v}_{\max}$ the supremum of the virtual waiting time during a busy cycle of the M/G/1 queueing system with $a < 1$, then (cf. [2], p. 606),

$$\Pr\{\underline{v}_{\max} \geqq t\} = \alpha \frac{d}{dt}\log W(t), \qquad t > 0. \tag{15}$$

Hence, the right-hand side of (15) is monotone.

Since

$$G(t) = t^{-\rho}L(t) \qquad \text{for} \qquad t \to \infty, \qquad G(t) = \int_t^{\infty} f(\tau)d\tau,$$

with $\rho > 0$ and $f(t)$ ultimately monotone for $t \to \infty$ implies

$$f(t) \sim \rho t^{-(\rho+1)} \mathrm{L}(t) \qquad \text{for} \qquad t \to \infty,$$

(cf. [8]), it follows from (14) and (15),

$$\beta \frac{d\mathrm{W}(t)}{dt} / \mathrm{W}(t) \sim \frac{ak}{1-a} (\beta/t)^{k+1} \mathrm{L}(t) \qquad \text{for} \qquad t \to \infty,$$

from which (11) follows since $\mathrm{W}(\infty) = 1$. The proof is complete, since (11) always implies (10) and (10) implies (9) on behalf of the lemma and the monotone density of $\mathrm{H}(t)$.

Note: in [1] it is proved that for the GI/G/1 queueing system with $a < 1$ the relations (9) and (10) imply each other.

3. The theorem above has a number of interesting consequences, which will be discussed below.

i) If

$$\lim_{t \to \infty} (t/\beta)^k (1 - \mathrm{B}(t)) = b, \qquad k > 1, \qquad b > 0, \tag{16}$$

then from (11) and (15)

$$\lim_{v \to \infty} (v/\beta)^k \Pr \{ \underline{v}_{\max} \geqq v \} = \frac{b}{1-a}.$$

Since the suprema of the virtual waiting times in successive busy cycles are i. i. d. variables it follows (cf. [7], p. 271) that: if (16) holds then

$$\lim_{n \to \infty} \Pr \left\{ \left(\frac{1-a}{bn} \right)^{1/k} \underline{\mathrm{V}}_n / \beta < x \right\} = \exp(- x^{-k}), \qquad x > 0, \tag{17}$$

with $\underline{\mathrm{V}}_n$ defined as in section 1.

ii) For the supremum $\underline{w}_{\max}$ of the actual waiting times in a busy cycle of a M/G/1 system with $a < 1$, we have [cf. [2], p. 606],

$$\Pr \{ \underline{w}_{\max} \geqq w \} = \alpha \frac{d}{dw} \log \Pr \{ \underline{w} < w + \underline{\sigma} \}, \qquad w > 0, \tag{18}$$

with $\underline{w}$ and $\underline{\sigma}$ independent, nonnegative variables with distribution $\mathrm{W}(t)$ and $1 - e^{-t/\alpha}$, $t \geqq 0$, respectively.

If

$$\Pr \{ \underline{w} \geqq w \} = w^{-k} \mathrm{L}(w), \qquad w \to \infty, \tag{19}$$

with $\mathrm{L}(w)$ slowly varying at infinity, then from

$$\Pr \{ \underline{w} \geqq w \} \geqq \Pr \{ \underline{w} - \underline{\sigma} \geqq w \} \geqq \Pr \{ \underline{w} \geqq w(1+\varepsilon) \} \Pr \{ \underline{\sigma} < w\varepsilon \}$$

for $\varepsilon > 0$, $w > 0$, we have

$$1 = \lim_{w\to\infty} \frac{\Pr\{\underline{w} \geqq w\}}{w^{-k}L(w)} \geqq \lim_{w\to\infty}\sup \frac{\Pr\{\underline{w} - \underline{\sigma} \geqq w\}}{w^{-k}L(w)}$$

$$\geqq \lim_{w\to\infty}\inf \frac{\Pr\{\underline{w} - \underline{\sigma} \geqq w\}}{w^{-k}L(w)}$$

$$\geqq \frac{1}{(1+\varepsilon)^k} \lim_{w\to\infty} \frac{\Pr\{\underline{w} \geqq w(1+\varepsilon)\}}{((1+\varepsilon)w)^{-k}L(w(1+\varepsilon))} = \frac{1}{(1+\varepsilon)^k},$$

since $L(w(1+\varepsilon))/L(w) \to 1$ for $w \to \infty$.

Consequently, the behaviour of $\Pr\{\underline{w} - \underline{\sigma} \geqq w\}$ for $w \to \infty$ is the same as that of $\Pr\{\underline{w} \geqq w\}$ if (19) holds. Since the righthand side of (18) is monotone in w the behaviour of the tail of the lefthand side of (18) follows by the same argument as in the proof in (11), and hence it results: if (16) holds then

$$\lim_{w\to\infty} (w/\beta)^k \Pr\{\underline{w}_{\max} \geqq w\} = \frac{b}{1-a}.$$

Consequently: if (16) holds then (17) is valid with $\underline{V}_n$ replaced by $\underline{W}_n$, the supremum of the actual waiting times during n successive busy cycles, i. e.

$$\lim_{n\to\infty} \Pr\left\{\left(\frac{1-a}{bn}\right)^{1/k} \underline{W}_n/\beta < x\right\} = \exp(-x^{-k}), \qquad x > 0. \tag{20}$$

iii) Denoting by $\underline{x}_{\max}$ the maximum number of customers simultaneously present during a busy cycle of the M/G/1 queue with $a < 1$, we have (cf. [*2*], p. 606) for $x \geqq 2$,

$$\Pr\{\underline{x}_{\max} > x\} = \frac{\displaystyle\int_0^\infty \frac{(t/\alpha)^{x-1}}{(x-1)!} e^{-t/\alpha} dW(t)}{\displaystyle\int_0^\infty \frac{(t/\alpha)^{x-1}}{(x-1)!} e^{-t/\alpha} W(t) dt/\alpha}. \tag{21}$$

If (16) applies then it is readily derived from (21) and (11) that

$$\lim_{x\to\infty} x^k \Pr\{\underline{x}_{\max} > x\} = \frac{b}{1-a} a^k, \tag{22}$$

by noting that $dW(t)/dt$ is uniformly bounded for $t \geqq 0$ and for every finite $T > 0$,

$$\lim_{x\to\infty} \int_0^T \frac{(t/\alpha)^{x-1}}{(x-1)!} e^{-t/\alpha} dW(t) = 0,$$

see also [*2*], p. 610 for a similar derivation.

From (22) it follows, if $\underline{X}_n$ denotes the supremum of the number of customers simultaneously present during n successive busy cycles that if (16) holds then

$$\lim_{n\to\infty} \Pr\left\{\frac{1}{a}\left(\frac{1-a}{bn}\right)^{1/k} \underline{X}_n < x\right\} = \exp(-x^{-k}), \qquad x > 0. \tag{23}$$

iv) At the moment of the m^{th} arriving customer $\underline{\mu}_m$ will denote the number of busy cycles preceding the moment of the m^{th} arrival, with $\underline{\mu}_m \overset{\text{def}}{=} 0$ if the m^{th} arrival belongs to the first busy cycle. Obviously, $\underline{\mu}_m$, $m = 1, 2, \ldots$, is the discrete renewal function of the renewal process with renewal distribution the distribution of the number $\underline{n}$ of customers served in a busy cycle. It is wellknown that with probability one

$$\lim_{m\to\infty} \frac{\underline{\mu}_m}{m} = 1/\mathrm{E}\{\underline{n}\} = 1 - a. \tag{24}$$

Evidently, we have for the maximum actual waiting time $\underline{w}_m^+$ of the first m arriving customers

$$\underline{W}_{\underline{\mu}_m} \leqq \underline{w}_m^+ \leqq \underline{W}_{\underline{\mu}_m+1}, \tag{25}$$

with $\underline{W}_0 \overset{\text{def}}{=} 0$. Berman (cf. [9], th. 3.2) showed that if (20) and (24) hold then for $x > 0$,

$$\lim_{m\to\infty} \Pr\left\{\left(\frac{1-a}{bm}\right)^{1/k} \underline{W}_{\underline{\mu}_m}/\beta < x\right\} = \{\exp(-x^{-k})\}^{1-a}.$$

Hence on behalf of (25) we obtain: if (16) holds then for $x > 0$,

$$\lim_{n\to\infty} \Pr\{(bn)^{-1/k}\underline{w}_n^+/\beta < x\} = \exp(-x^{-k}). \tag{26}$$

v) In the same way as (26) has been derived we obtain for $\underline{x}_n^+$, the maximum number of customers simultaneously present during the time interval of the first n arrivals: if (16) holds then for $x > 0$,

$$\lim_{n\to\infty} \Pr\{a^{-1}(bn)^{-1/k}\underline{x}_n^+ < x\} = \exp(-x^{-k}). \tag{27}$$

vi) With $\underline{\nu}_t$, $t > 0$, the number of busy cycles preceding time t, and $\underline{\nu}_t \overset{\text{def}}{=} 0$ if the first busy cycle exceeds t, it is evident that $\underline{\nu}_t$ is the renewal function of the renewal process with renewal distribution the distribution of the busy cycle $\underline{c}$. It is wellknown that with probability one

$$\lim_{t\to\infty} \frac{\underline{\nu}_t}{t} = 1/\mathrm{E}\{\underline{c}\} = \frac{1-a}{\beta}, \tag{28}$$

so that for $t \to \infty$

(29) $$\underline{\nu}_t/[t/\alpha] = \frac{\underline{\nu}_t}{t}(t/[t/\alpha]) \to \frac{1-a}{a}$$

with probability one, here $[t/\alpha]$ is the entier of t/α. Applying Berman's theorem (cf. [9], th. 3.2) we have from (29) and (17): if (16) holds

$$\lim_{t\to\infty} \Pr\left\{\left(\frac{1-a}{b[t/\alpha]}\right)^{1/k} \underline{V}_{\underline{\nu}_t}/\beta < x\right\} = \exp\left(-\frac{1-a}{a}x^{-k}\right),$$

so that by noting that $[t/\alpha]$ may be replaced by t/α,

(30) $$\lim_{t\to\infty} \Pr\{(bt/\beta)^{-1/k}\underline{V}_{\underline{\nu}_t}/\beta < x\} = \exp(-x^{-k}), \qquad x > 0.$$

Since

$$\underline{V}_{\underline{\nu}_t} \leqq \underline{v}_t^+ \leqq \underline{V}_{\underline{\nu}_t+1},$$

it follows from (30) for the supremum of the virtual waiting time during $[0, t]$ that if (16) holds then for $x > 0$,

(31) $$\lim_{t\to\infty} \Pr\{(bt/\beta)^{-1/k}\underline{v}_t^+/\beta < x\} = \exp(-x)^{-k}.$$

From the results above it is seen that we have proved the following theorem.

THEOREM. — For the queueing system M/G/1 with $a < 1$ if

$$\lim_{t\to\infty} (t/\beta)^k(1 - B(t)) = b, \qquad k > 1, \qquad b > 0,$$

then the limit theorems for $\underline{V}_n$ and $\underline{W}_n$ are given by (17) and (20), those for $\underline{X}_n$, $\underline{w}_n^+$ and $\underline{v}_t^+$ are given by (23), (26) and (31).

Finally, is is noted that if $\underline{\tau}_m$ represents the service time of the m^{th} arriving customer then for the sequences $\underline{w}_m$, $m = 1, 2, \ldots$, and $\underline{\tau}_m$, $m = 1, 2, \ldots$ the same extreme value theorem holds.

REFERENCES

[1] J. W. COHEN, Regular variation of the stationary waiting time distribution for the GI/G/1 queueing system, to appear in *J. Appl. Prob.* 1972.

[2] J. W. COHEN, *The Single Server Queue.* North Holland Publ. Co., Amsterdam, 1969, chapter III.7.

[3] J. W. COHEN, The suprema of the actual and virtual waiting times during a busy cycle of the $K_m/K_n/1$ queueing system, to appear in *Advances Appl. Prob.*, t. 4, 1972.

[4] Van VUGT, Limit theorems for the distribution of the supremum of the actual waiting times in a $K_m/G/1$ and a $G/K_n/1$ queueing system. Report Technological Univ. Delft, 1969, in Dutch.

[5] D. L. IGLEHART, Extreme values in the GI/G/1 queue. Report Stanford Univ., 1971, to appear in *Ann. Math. Stat.*
[6] H. CALLAERT and J. W. COHEN, A lemma on regular variation of a transient renewal function. Report Technological Univ. Delft, 1971, to appear in *Zeitschrift für Wahrscheinlichkeitsrechnung.*
[7] W. FELLER, *An Introduction to Probability Theory and its Applications.* Wiley, New York, vol. II, 1966.
[8] J. F. TEUGELS, Regular variation of Markov renewal theorems. *London Math. Soc.*, t. **2**, 1970, p. 179-190.
[9] S. M. BERMAN, Limiting distributions of the maximum term in sequences of dependent variables. *Ann. Math. Stat.*, t. **33**, 1962, p. 894-908.

(Manuscrit reçu le 22 février 1972).