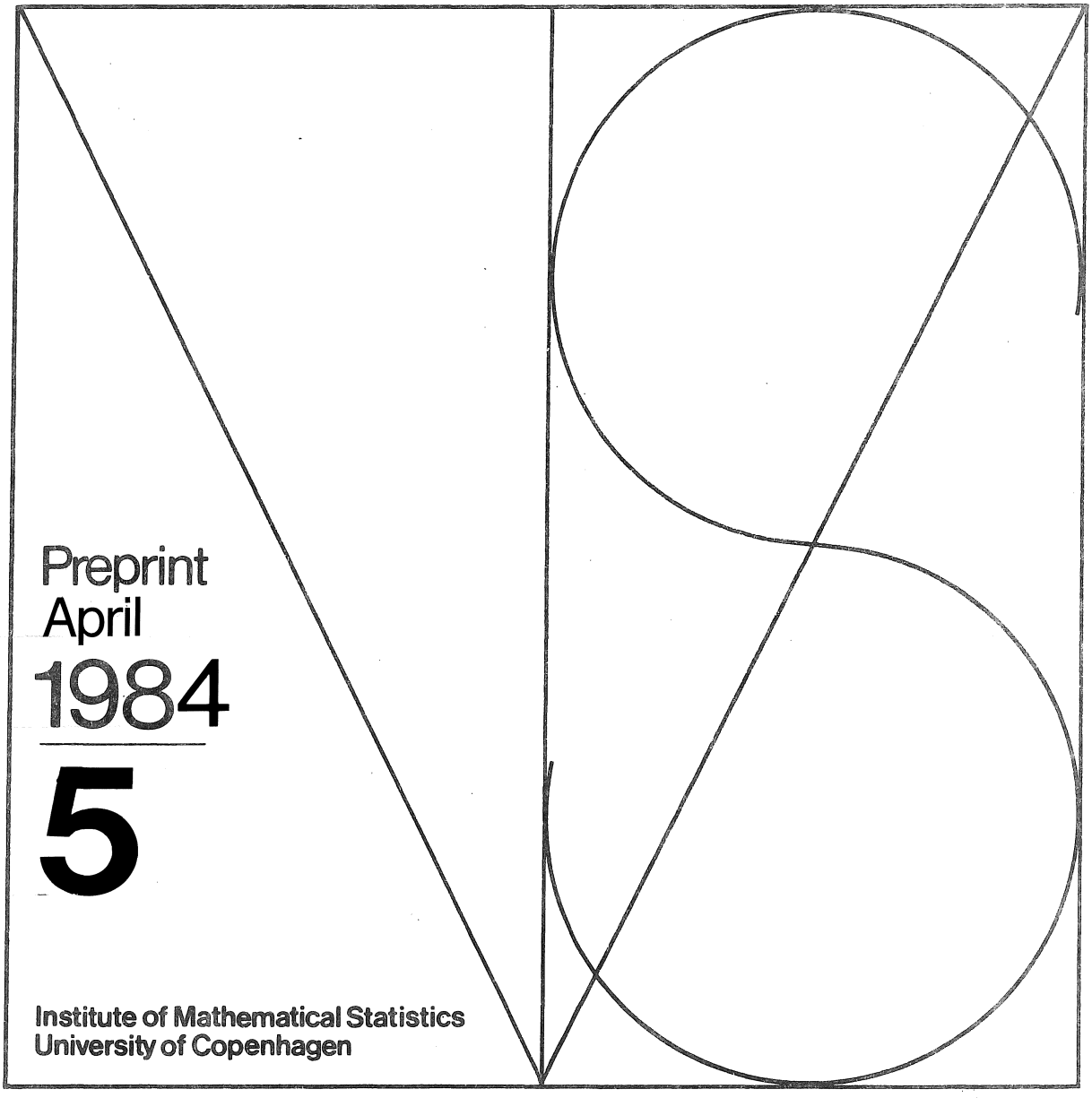
Søren Asmussen Hermann Thorisson

Boundary Problems and Large Deviation Results for Queue Length Processes



Preprint
April
1984
5

Institute of Mathematical Statistics
University of Copenhagen

Søren Asmussen and Hermann Thorisson*

BOUNDARY PROBLEMS AND LARGE DEVIATION RESULTS FOR QUEUE LENGTH PROCESSES

Preprint 1984 No. 5

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

April 1984

Boundary problems and large deviation results for queue length processes

Søren Asmussen, University of Copenhagen

and

Hermann Thorisson,* Chalmers University of Technology and the University of Göteborg

Abstract

Let for $N = 0,1,2,\ldots$ $\tau(N)$ be the time of first passage of the double-ended GI/G/1 queue length process with positive drift to level N and let $\xi(N)$ be the waiting time from $\tau(N)$ until the next service event. Then $\{\xi(N)\}$ is a Markov chain on $(0,\infty)$, the ergodic behaviour of which is found and shows some interesting connections to random walks. These observations are the key step in establishing an approximation of the form

$$P(Q(T) > N) \cong C\ \delta^N\ \Phi((T-\lambda N)/\kappa N^{\frac{1}{2}})$$

for the usual (one-sided) GI/G/1 queue length process with negative drift. Extensions to more general models are discussed.

---

*) Supported by the Swedish Natural Science Research Council and by the Icelandic Science Foundation.

1. Introduction

The present paper is concerned with the derivation of certain limit results for queue length processes. Particular attention is made to the finite time case, an area little developed compared to equilibrium, or steady state, theory.

Queueing problems in finite time seem certainly not unimportant from an application point of view: the steady state solution in many cases only plays the role of an approximation, the validity of which is not always easy to judge. However, the analytical difficulties are quite considerable. Closed form solutions are complicated even for such simple models as M/M/1, and can typically at best be found in some special cases in terms of double transforms, which provide little insight in the behaviour of the queueing probabilities themselves. Therefore approximation methods play an important role, the main classical results being based on the concepts of relaxation time and heavy traffic. To state these more precisely, let $\{Q_t\}_{t\geq 0}$ be the GI/G/1 queue length process which shall be our main example for a while. Assume that the traffic intensity ρ is less than one and that the interarrival distribution is non-lattice. Then the queue is stable and attains a limiting steady state, $Q_t \to Q_\infty$ in distribution, and for suitable constants it holds that

$$(1) \qquad P(Q_t \geq N) \cong P(Q_\infty \geq N) - C_1(N)t^{-3/2}e^{-\eta t}, \quad t\to\infty,$$

$$(2) \qquad P((1-\rho)Q_{t/(1-\rho)^2} \geq N) \cong e^{-\xi N}G(t), \quad \rho \uparrow 1,$$

where G is a first passage time distribution of a Brownian motion with drift. See e.g. [8] for the M/M/1 case of (1) and [8], [15] for waiting time analogues in M/G/1 (to our knowledge, (1) has not been proved for G1/G/1). For (2), see [16] and references therein, in particular [6].

We are here concerned with yet another type of approximation,

$$(3) \qquad P(Q_t \geq N) \cong C\,\delta^N\,\Phi\left(\frac{t-\lambda N}{\kappa N^{\frac{1}{2}}}\right), \quad N\to\infty,$$

i.e. a result of large deviation type. For $t = \infty$, $\Phi(\cdot) = 1$ so that (3) is a geometric approximation for the tail of the steady state queue length distribution exploited in [11] in a somewhat different setting and pointed out in [1] as an easy consequence of the classical exponential tail property of the waiting time (a recent result of this type is in [12]). For $t < \infty$, waiting time analogues were proved in [2] and in the unpublished technical report [3], queue lengths were treated for GI/M/m and a number of variants of M/G/1 (the first relation of this type seems to have occured in risk theory, [13], [5]).

This set of results leaves open the GI/G/1 case of (3) as an important gap. Though the method of [2], [3] (developed in Section 2) is quite general, it is non-trivial to fill out one of the steps, to establish ergodic properties of a certain process describing excess over the boundary in a two-dimensional problem in renewal theory. The purpose of the present short note is in part to resolve this problem (Section 3) and in part to present some of the material of [3] in a form stressing basic ideas rather than technical details.

We finally mention that [4] contains numerical comparisons of (1), (2), (3), for waiting times though, but we would not expect queue lengths to behave intrinsically different. None of the results are outstanding as approximations but (3) compares rather favourably for the purpose of giving some rough estimate of the correction to the equilibrium value. Unfortunately it does not seem easy to adapt to the present case the very accurate corrected diffusion approximation, based on [14] and essentially a variant and refinement of (2). It is strongly suggested from [14], however, that the boundary problem treated here would be one of the steps in that direction.

2. General method and its implementation

The technique of [2], [3] is based on representing $Q = \{Q_t\}_{t\geq 0}$ in terms of what is called a governing process $Q^* = \{Q^*_t\}_{t\geq 0}$ in [7]. This is a process evolving in a spatial homogeneous manner on the whole line, such that Q behaves as Q^* for large or moderate queue lengths. Once Q^* hits or becomes close to zero, Q is restarted in a regenerative way, say at time c.

Thus by means of standard formulas for regenerative processes, $P(Q_T > N)$ is obtained by convolving $f(t) = P(Q^*_t > N,\ t < c)$ with the renewal measure U associated with c. For asymptotic purposes, U is just replaced by the limiting normalized Lebesgue measure whereas to estimate $f(t)$, we split the path of Q^* in $[0, c)$ up in two segments separated by the first passage time $\tau(N) = \inf\{t\geq 0: Q^*_t > N\}$. One then has to find a supplementary variable $\xi(N)$ completely describing the initial conditions for the post-$\tau(N)$ process and next

1^o Prove an ergodic theorem for ξ, $\xi(N) \to \xi(\infty)$ in distribution, which ensures a regular behaviour of the post-$\tau(N)$ process;

2^o Prove that $\tau(N)$ is asymptotically normal conditionally upon $\{\tau(N) < \infty\}$ as $N\to\infty$. This is done by exploiting some random walk structure of Q^* which permits to involve what is called the associated process ${}^aP$ in [9], [2] and prove an unconditional CLT w.r.t. ${}^aP$ by Anscombe's theorem, using ${}^aP$-asymptotic normality of Q^* and $\tau(N)/N \overset{{}^aP}{\to} C_1$.

3^o Combine 1^o, 2^o to show

$$(4)\qquad P\left(\frac{\tau(N)-\lambda N}{\kappa N^{1/2}} \leq y,\ \xi(N) \leq x,\ \tau(N) < c\right) \cong C_2\ \delta^N\ \Phi(y)\ {}^aP(\xi(\infty) \leq x)$$

Having passed these points, (3) then follows rather easily up to the value of C. Thus one more step is needed,

4^o Give an alternative derivation of (3) for $T = \infty$, which permits to calculate C.

In the present case, Q^* is the so-called double-ended queue, i.e. the difference $N^A - N^B$ between two independent renewal processes governed by

the distributions A, B of the interarrival time, resp. the service time, and c is the length of the first busy cycle. The initial conditions corresponding to a customer having arrived just before zero, i.e. a busy cycle starting at $t = 0$, means N^A being zero-delayed and N^B having B itself as delay distribution. At time $\tau(N)$, an arrival has just occured and so the post-$\tau(N)$ N^A-process is again zero-delayed. However, to determine the post-$\tau(N)$ N^B-process we need to know the delay, that is, the time $\xi(N)$ until the next epoch of N^B.

Of the above steps, 4^o can be found in [1]. The expression for C is somewhat complicated, but can at least be made explicit in any model with imbedded Markov chain of matrix-geometric type, cf. [11]. Compared to [2], [3], 3^o follows in just the same way and 2^o is sketched in Sect. 4, requiring some technical variants. Where a really new idea is required is in 1^o since ξ fails to be regenerative, and we proceed to study this problem.

3. The Markov chain ξ

For each $y > 0$, we let P_y refer to the case where N^A is zero-delayed ($N^A(0) = 1$) and N^B has delay y. That is, the epochs of N^A, N^B are $\{S^A(n)\}_{n \geq 0}$, $\{y + S^B(n)\}_{n \geq 0}$ where $S^A(0) = 0$, $S^A(n) = X^A(1) + \ldots + X^A(n)$ etc with the obvious notation and independence assumptions. Define for $N = 0,1,2,\ldots$

$$\beta(N) = \inf\{n \geq N: S^A(n) < y + S^B(n-N)\},$$

$$\tau(N) = \inf\{t \geq 0: N^A(t) - N^B(t) = N + 1\} = S^A(\beta(N)),$$

$$\xi(N) = y + S^B(\beta(N) - N) - S^A(\beta(N))$$

(then $\beta(0) = \tau(0) = 0$, $\xi(0) = y$). Let further $\eta(u)$ be the overshot of level u for the random walk $\{y + (X^B(1) - X^A(2)) + \ldots + (X^B(n) - X^A(n+1))\}_{n \geq 0}$ and observe that $\eta(\cdot)$ is independent of $X^A(1)$ and

$$\xi(1) = \eta(X^A(1)).$$

THEOREM 1. *Suppose that* $EX^A(n) < EX^B(n)$ *and that* A *or* B *are non-lattice. Then* $\beta(N)$, $\tau(N)$, $\xi(N)$ *are all proper. Let further* $Y_1^A, Y_2^A, \ldots$ *be independent of the* $X^A(n), X^B(n)$ *and each distributed as* $X^A(n)$, *and define* $\tilde{\xi}(N) = \eta(Y_1^A + \ldots + Y_N^A)$. *Then* ξ *and* $\tilde{\xi}$ *are identically distributed Markov chains on* $(0,\infty)$. *For all* $y > 0$, *the limiting distribution of* $\xi(N)$, $\tilde{\xi}(N)$ *as* $N \to \infty$ *exists and is that of the weak limit* $\eta(\infty)$ *of* $\eta(u)$ *as* $u \to \infty$.

Proof: The first statement is easy since the law of large numbers implies that $\beta(N) < \infty$. Also the development of $N^A - N^B$ above the level N after time $\tau(N)$ is governed by just the same conditions as if we start $N^A - N^B$ at time zero with instead $y = \xi(N)$. Hence ξ is time-homogeneous Markov. That also $\tilde{\xi}$ is so follows by checking that observation of a Markov process $\{\eta(u)\}$ at the epochs $\{Y_1^A + \ldots + Y_N^A\}_{N \geq 0}$ of an independent renewal process yields

a homogeneous Markov chain. This structure shows also that if a limit $\eta(\infty)$ of $\eta(u)$ exists (which is a standard random walk result in the present case), then $\tilde{\xi}(N) \to \eta(\infty)$ weakly. Hence it only remains to check that ξ and $\tilde{\xi}$ have the same initial values and the same transition functions: The first statement follows from $\xi(0) = y = \tilde{\xi}(0)$ and the second from

$$\xi(1) = \eta(X^A(1)) \overset{\mathcal{D}}{=} \eta(Y_1^A) = \tilde{\xi}(1). \qquad \square$$

Remarks: The explicit form of the transition function for ξ is not needed but comes out quite easily in terms of the transition semigroup for η as

$$P(\xi(1) \leq x \mid \xi(0) = y) = P_y(\eta(Y_1^A) \leq x) = \int_0^\infty P_y(\eta(z) \leq x) dA(z)$$

$$= \int_0^\infty P(\eta(z) \leq x \mid \eta(0) = y)\, dA(z)$$

The process ξ provides a further example of a Lindley process with replacement, cf. [10]. Indeed, for all N with $S_N^A < \xi(0)$ we have $\xi(N) = \xi(0) - S^A(N)$ so that ξ evalues with the same increments as the random walk $\{-S^A(N)\}$ until the first negative value $-\eta^A(\xi(0)) = -z$ occurs: Here $\eta^A(u)$ is the overshot of level u for $S^A(n)$. Then $-z$ is replaced by a positive value depending on the past only through z and distributed as the overshot of level z for the random walk $S^A(n) - S^B(n)$.

4. The large deviation result

Let $\hat{A},\hat{B}$ denote the Laplace transforms of A,B. We shall need:

CONDITION (a) A is non-lattice and the equation $\hat{A}(\gamma)\hat{B}(-\gamma) = 1$ admits a solution $\gamma > 0$ with the additional property $\hat{B}''(-\gamma) < \infty$.

Define now the associated process ([9],[2]) by ${}^aA(dx) = \delta^{-1}e^{-\gamma x}A(dx)$, ${}^aB(dx) = \delta e^{\gamma x}B(dx)$ where $\delta = \hat{A}(\gamma) = \hat{B}(-\gamma)^{-1}$, and let ${}^aP$, ${}^aE$ etc refer to this set of distributions. It is a standard fact that for a stable queue ($EX^B(n) < EX^A(n)$) the associated queue is transient, ${}^aEX^B(n) > {}^aEX^A(n)$. Let further $\mathcal{F}_n = \sigma(X^A(k),\ X^B(k): k \le n)$, $\mathcal{F}_n^{(N)} = \sigma(X^A(k),\ X^B(\ell):\ k \le n,\ \ell \le N-n)$. For $G \in \mathcal{F}_{\beta(N)}^{(N)}$, $G \subseteq \{\beta(N) < \infty\}$ we then have $G \in \mathcal{F}_{\beta(N)}$ as well and hence in the same way as in [5],[14],[2],[4] that

$$(5)\qquad P_yG = {}^aE_y[\exp\{-\gamma(S_{\beta(N)}-y)\};\ G]$$

$$= e^{\lambda y}\,{}^aE[\exp\{-\gamma(\xi(N)-S^B(\beta(N)-N) + S^B(\beta(N)))\};\ G]$$

$$= e^{\eta y}\,\delta^{N+1}\,{}^aE[\exp\{-\gamma\xi(N)\};\ G],$$

using the independence of the $X^B(\beta(N)-N+k)$ of $\mathcal{F}_{\beta(N)}^{(N)}$ and ${}^aEe^{-\gamma X^B(k)} = \delta$.

LEMMA 1: As $N\to\infty$, $\tau(N)$ is asymptotically normal w.r.t. ${}^aP_y$, with mean λN and variance $\kappa^2 N$ where

$$(6)\qquad \lambda^{-1} = \frac{1}{{}^a\mu_A} - \frac{1}{{}^a\mu_B},\quad \kappa^2 = \lambda^3\left(\frac{{}^a\sigma_A^2}{{}^a\mu_A^3} + \frac{{}^a\sigma_B^2}{{}^a\mu_B^3}\right),$$

$${}^a\mu_A = {}^aEX^A(n),\ {}^a\sigma_B^2 = {}^aVarX^B(n) \text{ etc.}$$

Proof: It is well-known from renewal theory that

$$(7)\qquad \frac{N^A(t)-N^B(t)-\lambda^{-1}t}{(\lambda^{-3}\kappa^2 t)^{\frac{1}{2}}}$$

is asymptotically standard normal. Furthermore, from $N^A(\tau(N)) = N+1+N^B(\tau(N))$ it follows by division by $\tau(N)$ and $N^A(t)/t \to 1/\mu_A$, $N^B(t)/t \to 1/\mu_B$ that

$$(8) \qquad \frac{\tau(N)}{N} \to \left(\frac{1}{a_{\mu_A}} - \frac{1}{a_{\mu_B}}\right)^{-1} = \lambda \quad {}^{a}P_y - \text{a.s.}$$

Now Anscombe's condition for (7) was shown in [3] Lemma 4.4. Hence letting $t = \tau(N)$ and applying Anscombe's theorem we get that

$$\frac{N+1 - \lambda^{-1}\tau(N)}{(\lambda^{-3}\kappa^2\tau(N))^{\frac{1}{2}}}$$

is asymptotically standard normal. Applying (8) once more, the proof is complete. □

One can now carry out the step 3^{o} in Section 2 exactly as in [2], and up to the value of C, one gets

THEOREM 2: Consider a stable GI/G/1 queue length process $\{Q_t\}$. Then if condition (a) holds,

$$(9) \qquad P(Q_t > N) \tilde{=} C\,\delta^{N}\Phi\left(\frac{t - \lambda N}{\kappa N^{\frac{1}{2}}}\right), \quad N \to \infty,$$

where $\delta = \hat{A}(\gamma)$, λ, κ are given by Lemma 1 and

$$(10) \qquad C = \frac{1-\delta}{\gamma\,\delta\,\mu_A}\,D$$

where D is the constant in the standard asymptotic relation $P(W > X) \tilde{=} De^{-\gamma x}$ for the equilibrium actual waiting time.

[the relation (10) follows from the validity of (9) for $t = \infty$ and [1]. For a check, consider M/M/1 with $\mu_B = 1/\beta, \mu_A = 1/\alpha$. Then $\gamma = \beta - \alpha = (1-\rho)\beta$, $D = \delta = \rho$ and hence $C = \rho$ in agreement with $P(Q_\infty > N) = \rho^{N+1}$].

5. Service and arrivals in groups

It is natural to ask whether our large deviation result can be extended to more general models, and we shall here briefly review some of the discussion in [3] for the case of service and/or arrivals in groups. A survey of such models can be found in [8] and a variety of practical important cases arise according to how the server behaves near zero, i.e. if his service capacity exceeds the number of customers present. However, the governing process $\{Q_t^*\}$ is in all cases the same, the difference between two compound renewal processes, cf. Fig. 1, and there is no intrinsic difficulty in defining the associated

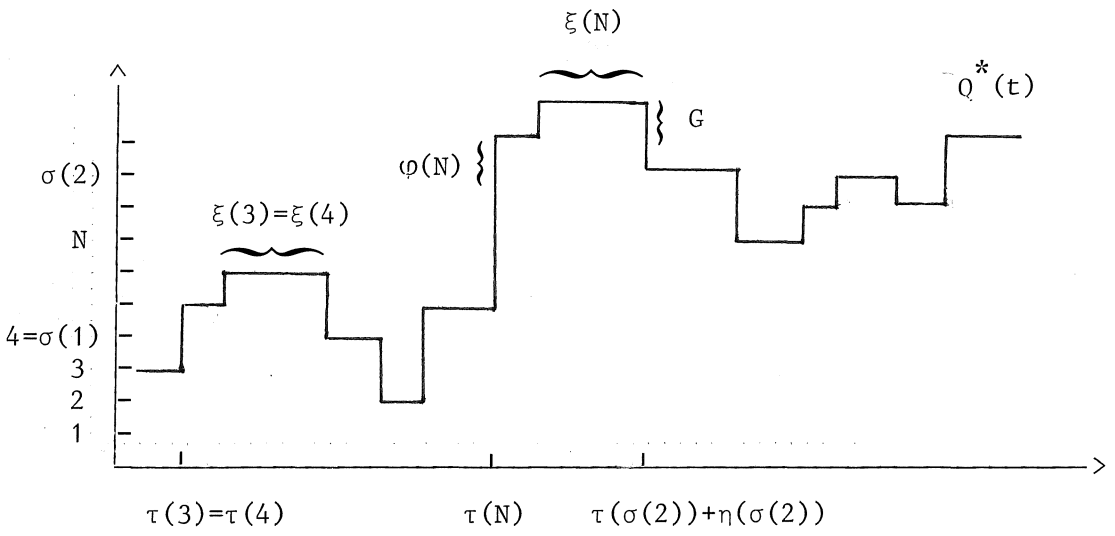


Figure 1

process and thereby assessing the values of the parameters δ,λ,κ in (5). The problem is again to compute C (which we shall not discuss here, see [3]) and, for a rigorous proof, to establish ergodicity of a suitable supplementary variable chain.

We define again $\tau(N) = \inf\{t \geq 0: Q^*(t) > N\}$, $\xi(N)$ as the residual service time at time $\tau(N)$, but it is seen that to determine the behaviour of the post-$\tau(N)$ process we need also to know the overshot $\varphi(N) = Q^*(\tau(N))-N-1$ (on Fig. 1, $\varphi(3) = 1, \varphi(4)=0$). Again, clearly $\{\xi(N), \varphi(N)\}$ is a Markov chain, but to show ergodicity seems considerably more complicated than in the case of single arrivals and services (where $\varphi(N)\equiv 0$). One possible approach would be by imposing conditions on absolutely continuous components and show Harris recurrence. In fact, this was our initial approach to Th. 1 before the particular structure of the problem was realized, but is heavily technical already for that case.

We shall instead here just point out that an easy alternative approach is available for the particular important case of compound Poisson arrivals. To this end, let

$$\sigma(k) = \text{the k'th } n \text{ such that } \varphi(n) = 0 \text{ and } N^A(\tau(n), \tau(n) + \xi(n)] = 0$$

cf. Fig. 1, and consider the evolvement of $\{\xi(n), \varphi(n)\}_{n > \sigma(k)}$. Since $Q^*(\tau(\sigma(k)))-1 = \sigma(k)$ the value of $Q^*(t)$ at $t = \tau(\sigma(k)) + \xi(\sigma(k))$ is $\sigma(k)+1-G$ where G is the number in the next group of services hence independent of $\xi(\sigma(k))$, $\varphi(\sigma(k))$. Also the next arrival occurs at time $\tau(\sigma(k)) + \xi(\sigma(k)) + U$ where U is exponentially distributed and independent of $\xi(\sigma(k))$. This shows that (ξ,φ) regenerates at $n = \sigma(k)$ and standard methods are applicable to derive convergence in distribution.

References

1. S. Asmussen, Equilibrium properties of the M/G/1 queue, Z. Wahrscheinlichkeitsth.verw.Geb. 58, 267-281 (1981).

2. S. Asmussen, Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the G1/G/1 queue, Adv.App.Probab. 14, 143-170 (1982).

3. S. Asmussen, Time-dependent approximation in some queueing systems with imbedded Markov chains related to random walks, Preprint 1981 no. 6, Institute of Mathematical Statistics, University of Copenhagen.

4. S. Asmussen, Approximations for the probability of ruin within finite time, Scand.Act.J. (1984).

5. B. von Bahr, Ruin probabilities expressed in terms of ladder height distributions, Scand.Act.J., 190-204 (1974)

6. A.A. Borovkov, Some limit theorems in the theory of mass service I-II, Th.Probab: Appl. 9, 550-565 (1964); ibid 10, 375-400 (1965).

7. A.A. Borovkov, Stochastic Processes in Queuing Theory, Springer Verlag, New York, Heidelberg, Berlin (1976).

8. J.W. Cohen, The Single Server Queue, North-Holland, Amsterdam (1969).

9. W. Feller, An Introduction to Probability Theory and its Applications 2 (2nd. Ed.), Wiley, New York (1971).

10. J. Keilson and J. Sumita, Evaluation of the total time in system in a preempt/resume priority queue via a modified Lindley Process, Adv.Appl. 15, 840-856 (1983).

11. M.F. Neuts, The caudal characteristic curve of queues, Tech.Rep. 82B, Applied Mathematics Institute, University of Delaware (1982).

12. M.F. Neuts and Y. Takahashi, Asymptotic behaviour of the stationary distributions in the G1/PH/C queue with heterogeneous servers, Z.Wahrscheinlichkeitsth.verw.Geb. 57, 441-452 (1981).

13. C.-O. Segerdahl, When does ruin occur in the collective theory of risk?, Skand.Aktuar Tidskr., 22-36 (1955).

14. D. Siegmund, Corrected diffusion approximation in certain random walk problems, Adv.Appl.Probab. 11, 701-719 (1979).

15. J.L. Teugels, Estimation of ruin probabilities, Insurance: Mathematics and Economics 1, 163-175 (1982).

16. W. Whitt, Heavy traffic limit theorems for queues: A Survey, Mathematical Methods in Queueing Theory (Clarke ed.), Lecture notes in Economics and Mathematical Systems 98, 307-350 (1974). Springer, New York.

PREPRINTS 1983

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Jacobsen, Martin: Two Operational Characterizations of Cooptional Times.

No. 2 Hald, Anders: Nicholas Bernoulli's Theorem.

No. 3 Jensen, Ernst Lykke and Rootzén, Holger: A Note on De Moivre's Limit Theorems: Easy Proofs.

No. 4 Asmussen, Søren: Conjugate Distributions and Variance Reduction in Ruin Probability Simulation.

No. 5 Rootzén, Holger: Central Limit Theory for Martingales via Random Change of Time.

No. 6 Rootzén, Holger: Extreme Value Theory for Moving Average Processes.

No. 7 Jacobsen, Martin: Birth Times, Death Times and Time Substitutions in Markov Chains.

No. 8 Hougaard, Philip: Convex Functions in Exponential Families.

PREPRINTS 1984

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Rootzén, Holger and Sternby, Jan: Consistency in Least Squares Estimation: A Bayesian Approach.

No. 2 Hougaard, Philip: Parameter Transformations in Multiparameter Nonlinear Regression Models.

No. 3 Jacobsen, Martin: Cooptional Times and Invariant Measures for Transient Markov Chains.

No. 4 Rootzén, Holger: Attainable Rates of Convergence of Maxima.

No. 5 Asmussen, Søren and Thorisson, Hermann: Boundary Problems and Large Deviation Results for Queue Length Processes.