Journal of Applied Mathematics and Stochastic Analysis
8, Number 4, 1995, 331-340

COMMENTS ON A SINGLE SERVER QUEUE[1]

RYSZARD SYSKI
University of Maryland
Department of Mathematics
College Park, MD 20742 USA

(Received May, 1995; Revised August, 1995)

ABSTRACT

A review of the work of V.E. Beněs and P. Brémaud on a single-server queue G/G/1 is interpreted in terms of point processes and associated martingales.

A concept of a discounted workload is introduced, and its connection with random Dirichlet series and associated semigroups is investigated.

Key words: Beneš integral equation, Dirichlet series, martingales, point processes, semi-groups, workload.

AMS (MOS) subject classifications: 60K25, 90B22, 60G44, 60G57, 60H220.

1. Introduction

This paper is devoted to comments on a topic of great interest which well illustrates general principles in Queueing Theory. This is the case of the waiting time in a single server queue, G/G/1 with a general input and a general service time, without any assumptions of independence.

It was V. Beněs who, in his rather thin, but delightful and very well-written book which was published in 1963, developed the general theory free of special assumptions on type of processes; see [1]. In a book published in 1981, P. Brémaud used martingale methods to study waiting time, queue length, and other problems in the G/G/1 system; see [2].

The present paper summarizes the work of Beněs and Brémaud, and compares their approaches. In addition, a new interpretation arising from this study is presented in the last section of the paper.

2. Beněs' Theory

The primary ingredient of Beněs' theory is the idea of describing traffic offered to a queue by a single-stochastic process $K = (K_t)$ where K_t represents the workload offered in the interval $[0, t]$, for each $t \geq 0$. The process K has nondecreasing and right-continuous paths, and can be re-

[1]The paper is an extended version of a talk presented at the 3rd INFORMS Telecommunications Conference in Boca Raton, Florida during March 1995.

presented in the following form:

$$K_t = \sum_{n=1}^{\infty} Z_n I_{(T_n \le t)}, \quad t \ge 0$$

where T_n is the instant of arrival of the nth customer and Z_n is the amount of service required by that customer; I_A denotes the indicator of an event A.

Hence, K is in fact a marked point process with mark space $\boldsymbol{R}$. Note that no assumptions on processes (T_n) and (Z_n) are imposed (in particular, no independence); therefore, the theory is free from effects of such assumptions.

Beněs targeted the virtual waiting time V_t (for strict order service) defined as the time a customer would have to wait for service if that customer arrived at time t. Evidently, the process (V_t) is a functional of the workload process (K_t), more specifically,

$$V_t = \tau(K_s, s \le t), \quad t \ge 0,$$

where τ is an operator. Thus, V_t is regarded as the residual work at time t. Paths of the process (V_t) are familiar saw-tooth right-continuous functions with jumps of magnitude Z_n occurring at times T_n, and decreasing with slope -1 between jumps.

In terms of this interpretation, Beněs derived the following stochastic integral equation for the virtual waiting time:

$$V_t = K_t - t + \int_0^t H(-V_u)du, \quad t \ge 0,$$

where H is the unit step, and the integral represents the total time during which server was idle up to t.

An explicit solution of the equation (obtained by E. Reich) has the form:

$$V_t = \sup_{0 \le s \le t} (\zeta_t - \zeta_s)$$

provided ζ_x has a zero in $[0,t]$, and $V_t = \zeta_t$ when $\zeta_x > 0$ for $x \in [0,t]$. Here $\zeta_t = K_t - t$, if positive, represents the excess of arriving load in the interval $(0,t]$ over the elapsed time t; it is therefore the overload up to t.

Beněs was interested in finding distribution of V_t, and showed that it can be obtained in terms of distribution of K_t and the conditional probability

$$R(t,u,w) = \mathbb{P}(\zeta_t - \zeta_u \le w \mid V_u = 0).$$

This leads to the following equation valid for $w > 0$:

$$\mathbb{P}(V_t \le w) = \mathbb{P}(\zeta_t \le w) - \frac{\partial}{\partial w} \int_0^t R(t,u,w)\mathbb{P}(V_u = 0)du$$

where $\mathbb{P}(V_u = 0)$; therefore, the probability that the server is idle at time u, must be determined from the separate Volterra integral equation of the first kind. It is the above integral equation whose modified form became the starting point of applications to fluid models in ATM systems; see Roberts [10] and Norros [8].

Explicit solutions of the Beněs integral equation for the distribution of the virtual waiting time process are hard to get, even by applying the Laplace-Stieltjes transform to this equation.

Of special interest is the recent study of the Beneš equation using Malliavin calculus, [3].

To achieve some simplification, Beneš introduced two assumptions:

i) "weak stationarity" expressed by the requirement that the function R depends only on time difference, $R(t,u,w) = R(t-u,w)$;

ii) "weak Markov property" expressed in terms of the first zero of ζ_t; this is weaker than the requirement that such a zero is a regenerative point.

His book is essentially devoted to the study of the effects of these assumptions. He illustrated his theory on the example of the M/G/1 queue, where the process K is Markovian. Beneš, however, did not use explicitly the usual terminology of marked point processes and did not consider martingales. These ideas were extensively utilized by Brémaud. He mentioned the same representation for a point process K, but preferred to rather work with a queueing process (Q_t) defined

$$Q_t = Q_0 + A_t - D_t, \quad t \geq 0,$$

where (A_t) and (D_t) are point processes without common jumps.

Here Q_t is the number of customers in the system at time t, and A_t and D_t denote the number of arrivals and departure up to t, respectively.

The study of the point process (K_t) by martingale approach is presented below in section 3.

3. Brémaud's Theory

The first application of martingales to the study of queueing systems was carried out by Brémaud in his doctoral thesis and subsequent papers; results were summarized in his book [2]. The following is an outline of this approach. For basic concepts of processes and properties of point processes used here, see [2], [4], [6], [7], and [9].

Consider the increasing right-continuous process $K = (K_t)$ already defined above. It follows from the general theory of processes that there exists a unique (up to equivalence) right-continuous predictable increasing process $B = (B_t)$, such that

$$\mathbb{E}\int_0^{T_n} C_s dK_s = \mathbb{E}\int_0^{T_n} C_s dB_s$$

for all non-negative predictable processes $C = (C_t)$. B is called a *dual predictable projection*, or a (*predictable*) *compensator* of K.

Moreover, the process $M = (M_t)$ defined by

$$M_t = K_t - B_t, \quad t \geq 0,$$

is a local martingale (with mean value zero), relative to the natural σ-fields of K, namely $F_t = \sigma(K_s, s \leq t)$ for each t. This means that for $T_n \to \infty$, the process $(M_{t \wedge T_n})$ is a uniformly integrable martingale for each $n \geq 1$.

The localization is needed because, in general, the process K is not integrable (only locally integrable, $\mathbb{E}K_{T_n} < \infty$ for each n).

In the case of the process K considered here, its jump times T_n are totally inaccessible, so the

compensator B is continuous. In particular, when B is absolutely continuous with respect to the Lebesgue measure, it Radon-Nikodym derivative (taken predictable) is called the *intensity* of K.

In the theory of point processes, it is natural to take smaller σ-fields (also called internal histories) for which explicit expressions for compensators can be obtained. Consider first a counting process defined by

$$N_t(A) = \sum_{n=1}^{\infty} I_{(Z_n \in A)} I_{(T_n \le t)}$$

for any Borel subset A of $\mathbb{R}_+$. Then, define σ-fields

$$F_t^{\mu} = \sigma[N_s(A),\ 0 \le s \le t, A \in \mathcal{B}_+] \text{ for each } t \ge 0,$$

where superscript μ indicates a random measure $\mu(dt, dz)$ such that

$$\mu\{(0,t], A\} = N_t(A).$$

The double sequence (T_n, Z_n) and the random measure $\mu(dt, dz)$ are identified and both are called a (*primitive*) *marked point process.* Note that F_t^{μ} is also generated by indicators of events $(Z_n \in A)$ and $(T_n \le t)$. One can show that for each T_n,

$$F_{T_n}^{\mu} = \sigma(T_i, Z_i, 0 \le i \le n) \text{ and } F_{T_n-}^{\mu} = \sigma(T_i, Z_i, 0 \le i \le n-1; T_n).$$

Let F_t be a history of $\mu(dt, dz)$ suitable completed such that

$$F_t = F_0 \vee F_t^{\mu}.$$

Suppose that for each $n \ge 1$ there exists a regular conditional distribution of $(T_{n+1} - T_n, Z_n)$ given F_{T_n} of the form:

$$\mathbb{P}(T_{n+1} - T_n \in ds, Z_{n+1} \in dz \mid F_{T_n}) = G^{(n+1)}(ds, dz) = g^{(n+1)}(s, dz)ds$$

with the marginal distribution

$$G^{(n+1)}(ds, \mathbb{R}_+) = G^{(n+1)}(ds).$$

Then, it is shown that there exists a positive predictable random measure

$$\nu(ds, dz) = \frac{g^{(n+1)}(s - T_n, dz)ds}{1 - G^{(n+1)}(s - T_n)} = \Lambda_s(dz)ds, \text{ for } T_n < s \le T_{n+1}$$

which defines intensity $\Lambda_s(dz)$. Thus, for any Borel set A, the intensity of the point process K may be written as $\Lambda_s(A)$, and the compensator B has the form:

$$B_t(A) = \int_0^t \Lambda_s(A)ds,$$

with left-continuous process $\Lambda_s(A)$, or alternatively:

$$B_t(A) = B_{T_n}(A) + \int_{T_n}^t \Lambda_s(A)ds, \quad \text{for } T_n < t \le T_{n+1}.$$

In other words, for each n and each Borel set A, the family of r.v.'s

$$N_{t\wedge T_n}(A) - \int_0^{t\wedge T_n} \Lambda_s(A)ds$$

is a martingale, relative to (F_t). Alternatively, the difference of random measures $\mu\{(0,t],A\} - \nu\{(0,t],A\}$ for $t \geq 0$ is a local martingale.

Finally, the following representation holds for any martingale (M_t) relative to (F_t) and vanishing at zero:

$$M_t = \int_0^t \int_0^\infty H(s,z)[\mu(ds,dz) - \nu(ds,dz)]$$

where H is a predictable process integrable with respect to $\Lambda_s(dz)ds$ for any $t \geq 0$. In particular, for $H(s,z) = z$,

$$\mathbb{E}K_t = \mathbb{E}B_t$$

as it should be.

For example, for the usual M/G/1/ queue with service time having arbitrary distribution function F with density f and mean m, we get

$$g(s,dz)ds = \lambda e^{-\lambda s} f(z)dz$$

for each n. Hence

$$\Lambda_s(dz) = \lambda f(z)dz,$$

and thus

$$B_t = \lambda mt.$$

4. Applications

As already indicated, the virtual waiting time as well as other quantities of interest, are essentially obtained by appropriate transformations of the workload process $K = (K_t)$. By using well-known methods, explicit results are rarely obtained, even in the form of various transforms. It is, therefore, of interest to consider other transformations of the process K which may lead to functionals useful in characterization of a queueing system.

There are several possibilities and the choice should be made based on the properties of a selected functional. Some examples given below are based on the martingale approach. It should be noted that this is not a reinterpretation of the virtual waiting time, but rather a study of another property of a G/G/1 system.

4.1 α-process

In general, K_t diverges when $t \to \infty$. To insure convergence, consider the "discounted" workload K_t^α defined by

$$K_t^\alpha = \int_0^t e^{-\alpha s} dK_s \quad \text{for } \alpha > 0 \text{ and } t \geq 0.$$

Then,

$$K_t^\alpha = \sum_{n=1}^{\infty} Z_n e^{-\alpha T_n} I_{(T_n \leq t)} \quad \text{for } t \geq 0,$$

so

$$K_\infty^\alpha = \sum_{n=1}^{\infty} Z_n e^{-\alpha T_n}.$$

Next, recall that

$$\nu(ds, dz) = \Lambda_s(dz)ds \text{ for } T_n < s \leq T_{n+1},$$

where intensity $\Lambda_s(dz)$ depends on n. Then, the compensator (B_t^α) of the process (K_t^α) has the form:

$$B_t^\alpha = \int_0^t \int_0^\infty e^{-\alpha s} z\nu(ds, dz) = \sum_{i=0}^{\infty} \int_{T_i \wedge t}^{T_{i+1} \wedge t} e^{-\alpha s} \Lambda_s^* ds,$$

where

$$\Lambda_s^* ds = \int_0^\infty z\nu(ds, dz).$$

Obviously,

$$B_\infty^\alpha = \sum_{i=0}^{\infty} \int_{T_i}^{T_{i+1}} e^{-\alpha s} \Lambda_s^* ds.$$

Consequently,

$$M_t^\alpha = K_t^\alpha - B_t^\alpha, \quad t \geq 0,$$

is a local martingale, in agreement with the representation of (M_t^α)

$$M_t^\alpha = \int_0^t e^{-\alpha s} dM_s, \quad t \geq 0$$

as a stochastic integral of a continuous function, with respect to a local martingale.

Assume now, that $\mathbb{E}K_\infty^\alpha < \infty$. Then, according to [4], the potential generated by an integrable increasing process (K_t^α) is given by

$$U_t^\alpha = \mathbb{E}(K_\infty^\alpha \mid F_t) - K_t^\alpha, \quad t \geq 0,$$

where the first term on the right is an uniformly integrable martingale, and (U_t^α) is a supermartingale with $\mathbb{E}U_t^\alpha \to 0$ as $t \to \infty$.

Next, denote by $K_t^{(2)}$ the convolution of K_t with itself, so $(K_\infty^\alpha)^2$ is the Laplace-Stieltjes transform of $K_t^{(2)}$. Then, the energy of a potential (U_t^α), or of an increasing process (K_t^α), is defined by the expression:

$$en = 2^{-1}\mathbb{E}(K_\infty^\alpha)^2;$$

see [11].

For example, in the M/G/1 queue,

$$\mathbb{E}K_\infty^\alpha = \mathbb{E}B_\infty^\alpha = \lambda m \alpha^{-1},$$

and energy is

$$en = \frac{m^{(2)}\lambda}{4\alpha} + \frac{m^2\lambda^3}{\alpha^2(2\lambda+\alpha)}$$

where m and $m^{(2)}$ are the first and the second moment of the service time distribution, respectively.

4.2 Dirichlet series

The series for K_∞^α, defined earlier in 4.1 with $Z_n > 0$ and $T_n < T_{n+1}$, $T_n \to \infty$, is known as a random Dirichlet series. Writing K_∞^α as

$$\sum_{n=1}^{\infty} Z_n e^{-\alpha T_n} = \int_0^\infty e^{-\alpha s} dK_s,$$

indicates that its properties are analogous to properties of Laplace-Stieltjes transforms.

As in the case of non-random series, if the series converges for some α_0, then it converges also for each α such that $\alpha > \alpha_0$. Infimum of such α that K_∞^α converges is called the *abscissa of convergence* of the series, and is a random variable denoted by

$$\Gamma = \inf(\alpha: K_\infty^\alpha < \infty).$$

From standard results (see [12]), one can deduce that

$$\Gamma = \limsup_n (\log S_n)/T_n,$$

where

$$S_n = Z_1 + \ldots + Z_n \text{ for } n \geq 1, \text{ and } S_0 = 0,$$

is the total service time up to n.

Of special interest is the case when Z_n is constant for all n, say c, and $T_n = \log(nT)$ where T is a positive random variable. Then,

$$K_\infty^\alpha = cT^{-\alpha}\zeta(\alpha) \text{ and } \Gamma = 1,$$

where $\zeta(\alpha)$ is the Riemann zeta function (converging for $\alpha > 1$).

4.3 Semigroup

Another interpretation of the "discounted" workload K^α may be obtained by considering an operator P^α acting on the space $\mathcal{S}$ of random sequences $S = (S_n)$ such that

$$Z_n = S_n - S_{n-1} > 0, \quad n \geq 1, \text{ with } S_0 = 0.$$

Assume that the norm on the space $\mathcal{S}$ is defined by:

$$\| S \| = \sup_n \mathbb{E} Z_n.$$

In particular, one can take for (S_n), a random sequence whose terms represent total service time

up to n, as defined in 4.2.

Suppose that

$$0 < T_1 < T_n < T_{n+1} \text{ and } T_n \to \infty.$$

Then, consider partial sums

$$K_n^\alpha = \sum_{k=1}^{n} (S_k - S_{k-1}) e^{-\alpha T_k} \text{ for } n \geq 1, \quad K_0^\alpha = 0,$$

and form a random sequence $K^\alpha = (K_n^\alpha)$. This sequence belongs to $\mathcal{S}$, and its norm is

$$\| K^\alpha \| = \sup_n \mathbb{E}(K_n^\alpha - K_{n-1}^\alpha).$$

Observe that for fixed n, $K_n^\alpha \to S_n$ as $\alpha \to 0$, but $\lim_n S_n = S_\infty$ is infinite by assumption. Furthermore, for fixed α, $\lim_n K_n^\alpha = K_\infty^\alpha$ coincides with K_∞^α defined earlier and is finite for $\alpha > \Gamma$.

Now define the transformation P^α which maps random sequences S on random sequences K^α by:

$$P^\alpha S = K^\alpha.$$

Thus, the nth component of $P^\alpha S$ is the nth partial sum of the Dirichlet series K_∞^α.

Theorem: *The family (P^α) forms a semigroup of linear transformations on the space $\mathcal{S}$ into itself*

$$P^{\alpha+\beta} = P^\alpha P^\beta, \quad \alpha > 0, \ \beta > 0,$$

with the norm not exceeding $\mathbb{E}e^{-\alpha T_1}$.

The proof may be patterned on that for the non-random case; see [5]. Here, only the semigroup property is verified. From the definition, $(P^\beta S)_n = K_n^\beta$, so

$$(P^\alpha K^\beta)_n = \sum_{i=1}^{n} (K_i^\beta - K_{i-1}^\beta) e^{-\alpha T_i} = \sum_{i=1}^{n} Z_i e^{-\beta T_i} e^{-\alpha T_i} = K_n^{\alpha+\beta}.$$

An equivalent definition of the transformation P^α is obtained by introducing the triangular matrix $\boldsymbol{T}^\alpha = (t_{ij}^\alpha)$, where

$$t_{ij}^\alpha = e^{-\alpha T_j} - e^{-\alpha T_{j-1}}, \ j < i, \ t_{ii}^\alpha = e^{-\alpha T_j}, \ t_{ij}^\alpha = 0, j > i.$$

Then, it is easy to verify that $\boldsymbol{T}^\alpha S = K^\alpha$.

Furthermore, the infinitesimal generator of the semigroup (P^α) is the operator $\boldsymbol{Q}$ taking random sequences S into random sequences $\boldsymbol{Q}S = (D_n)$ where

$$D_n = -\sum_{k=1}^{n} Z_k T_k.$$

This transformation is also based upon a sub-diagonal matrix $\boldsymbol{A} = (a_{ij})$ where

$$a_{ij} = T_{j+1} - T_j, j < i, a_{ii} = -T_j, a_{ij} = 0, \ j > 1,$$

for $i = 0, 1, 2, \ldots$ with state O absorbing. Note that

$$\boldsymbol{A}S = \boldsymbol{Q}S.$$

4.4 Example

It may be helpful to see an example of this approach to a non-conventional G/G/1 queue with some dependence.

Suppose that the joint density of the input $T_{n+1} - T_n$ and the service Z_{n+1} has the form

$$f(s,z) = \begin{cases} \mu(\lambda+\mu)e^{-\lambda s}e^{-\mu z} & \text{for } z > s \\ 0 & \text{for } z < s \end{cases}$$

independent of n. In agreement with standard notation, λ and μ are positive constants and μ should not be confused with measure μ introduced earlier.

The marginal densities are

$$f_1(s) = (\lambda+\mu)e^{-(\lambda+\mu)s} \text{ and } f_2(z) = (\mu/\lambda)(\lambda+\mu)e^{-\mu z}(1-e^{-\lambda z})$$

for $s > 0$ and for $z > 0$, respectively, indicating a dependence.

Take now

$$g^{(n+1)}(s,dz)ds = f(s,z)dsdz \quad \text{for } z > s.$$

Then, the random measure $\nu(ds,dz)$ has the form:

$$\nu(ds,dz) = \mu(\lambda+\mu)e^{\mu(s-T_n)}e^{-\mu z}ds\,dz, \text{ on } T_n < s \leq T_{n+1}.$$

Consequently, for $T_n < t \leq T_{n+1}$, the compensator B_t is

$$\begin{aligned} B_t &= \int_0^t \int_0^\infty z\nu(ds,dz) \\ &= (\lambda+\mu)t/\mu + (\lambda+\mu)2^{-1}\left(\sum_{k=0}^{n-1}(T_{k+1}-T_k)^2 + (t-T_n)^2\right). \end{aligned}$$

Rather tedious evaluation yields for $\alpha > 0$:

$$B_\infty^\alpha = (\lambda+\mu)/(\mu\alpha) + (\lambda+\mu)\alpha^{-2} - (\lambda+\mu)\alpha^{-1}\sum_{n=0}^{\infty}(T_{n+1}-T_n)e^{-\alpha T_n}.$$

On the other hand, the random measure $\mu(ds,dz)$ is

$$\mu(ds,dz) = \sum_{n=1}^{\infty} I_{(T_n \in ds)}I_{(Z_n \in dz)}.$$

Hence,

$$K_t = \int_0^t \int_0^\infty z\mu(ds,dz) = \sum_{n=1}^{\infty} Z_n I_{(T_n \leq t)},$$

and for $\alpha > 0$,

$$K_\infty^\alpha = \sum_{n=1}^{\infty} Z_n e^{-\alpha T_n}.$$

As already noted, for $\alpha \geq 0$:

$$\mathbb{E}K_t^\alpha = \mathbb{E}B_t^\alpha.$$

References

[1] Beneš, V.E., *General Stochastic Processes in the Theory of Queues*, Addison-Wesley 1963.

[2] Brémaud, P., *Point Processes and Queues*, Springer Verlag 1981.

[3] Decreusefond, L. and Ustunel, A.S., The Beneš equation and stochastic calculus of variations, *Stoc. Proc. Applications*, (presented at the 3rd INFORMS Telecom. Conf., Boca Raton, Florida 1995), to appear.

[4] Dellacherie, C. and Meyer, P.A., *Probabilities and Potential* (*B*), North-Holland 1982.

[5] Hille, E. and Phillips, R.S., *Functional Analysis and Semi-groups*, (Colloq. Publ.), Amer. Math. Soc. 1957.

[6] Karr, A.F., *Point Processes and their Statistical Inference*, Marcel Dekker, New York 1966.

[7] Metivier, M., *Semimartingales*, Walter de Gruyter, Berlin 1982.

[8] Norros, I., Studies on a model for connectionless traffic based on fractional Brownian motion, *COST* **242TDC(92)** (1992).

[9] Reiss, R.D., *A Course on Point Processes*, Springer Verlag 1993.

[10] Roberts, J.W. (editor), *Performance Evaluation and Design of Multiservice Networks*, COST **224**, Final Report, CEC 1992.

[11] Syski, R., *Passage Times for Markov Chains*, IOS Press 1992.

[12] Widder, D.V., *The Laplace Transform*, Princeton University Press, New Jersey 1946.