PUBLICATIONS DU DÉPARTEMENT DE MATHÉMATIQUES DE LYON

J. P. KELLER

The Intermittent Server

Publications du Département de Mathématiques de Lyon, 1978, tome 15, fascicule 1 , p. 83-95

<http://www.numdam.org/item?id=PDML_1978__15_1_83_0>

Publications du
Département de
Mathématiques
Lyon 1978 t. 15-1

THE INTERMITTENT SERVER

by J.P. KELLER

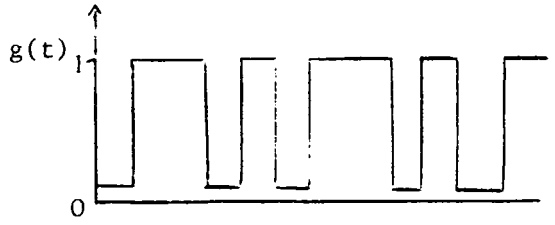
0. INTRODUCTION.

This note is devoted to the study of Queuing Systems in which some of the servers may work intermittently, ie, a server is allowed to work only when "given" time by some allocator. This case occurs, for instance, when the same server carries out more than one kind of services, and our system differentiates these services. Another example is a configuration with "mutually exclusive servers" when only one of the mutually exclusive servers may work a a time, the other being idle. This happens for example when several servers share the same ressource of which only one copy is available for all them. The obvious attitude to have in these situations is to assume that if μ is the serving rate of one working server. (ie, the serving rate if uninterrupted) and if the server works 1/n-th of the time, then the actual serving rate is μ/n, summarizing, intermittent servers ought to be treated as ordinary servers with "cut rates". A rigorous treatment of this folk-theorem presents more difficulties than one would expect. In this respect the intervention of Non-Standard Analysis to prove the basic theorem of this paper, we hope, will enlight the nature of the theorem - the dealings with infinitesimal oscillations - rather than bury the reader into a sea of epsilons.

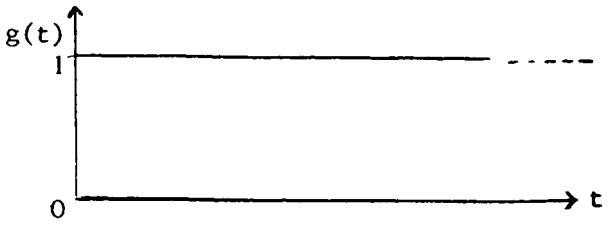
For expository reasons we will present first the case of a single server We will then easily generalize Jackson's theorem on networks of queues to include intermittent servers. We claim that a large number of systems usually investigated by programming methods - simulations - could be, with a minimum of algorithmic analysis, described by closed formulas. This could be the case of an Operating system, in one of its key section, the Monitor.

This report is only one side of a collective effort undertaken at the Courant Institute : examples illustrating the usefulness of our techniques by G. Belpaire (2), and verifications by means simulations by A. Borg (3) will be reported elsewhere.

The fundamental difficulty, for a classical queuing theorist, with our intermittent servers is : the time scale is neither memoryless nor uniform. For a given server let $g(t)$ be the serving time characteristic function (fig. 1 a, b).



(fig. 1a) g for an intermittent server.



(fig. 1b) g for a classical server.

When g(t) = 1, the server si actually working. When g(t) = o, the server is idle. Consider a single server queue and its discrete states Ei (= state when i people are in the system). If the server is *not intermittent*, the memoryless property of any state may be described as follows (4). Let τ_i be a random variable that represents the time which the process spends in state Ei. Then

(1) $$P\ (\tau_i > s + t \mid \tau_i > s) = h_i(t)$$

holds for the conditionnal probability. If our system has an intermittent server, the system has actually two different time clocks : the time clock to join the system (which in our example is uninterrupted) and the clock used to leave the system -represented by g(t). Thus (1) does not hold any more. We still retain a discrete state space representation for our stochastic process with P(k,t) being the probability of finding the system in state E_k at time t.

We are going to introduce a suitable averaging procedure to reduce this system to a Markov process. The crux of the method is as follows : instantaneous probabilities may depend on the current value of g (at least their time derivative does , as will be seen below). Thus there might not be any limiting probability distribution as $t \to +\infty$ in the sense that $\frac{\partial P}{\partial t}$ is never identically zero, i.e., P oscillates at all time. The key observation is : at $t \to +\infty$ these oscillations might very well be infinitesimal in amplitude (fig. 2).

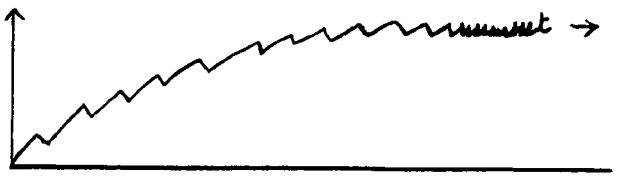


Fig. 2 . a function with a discontinuous derivative and infinitesimal oscillations as $t \to \infty$.

Anyone who has had Calculus via Non-standard Analysis (5) will not have any difficulty with our exposition of the solution. Other people are refered for a quick introduction - and that's all we need - to the introductory chapters of (6) or (7). Our reasoning has enough intuitive appeal for any reader to fell compelled to read it carefully, if not verify it (this part of the paper ought to be entitled : Another Act in the Revenge of Leibniz).

In section I we are going to establish the stochastic difference equation for the single intermittent server. In section II we will parallel Jackson's treatment of a "Job-shop-like Queuing System" (1) with groups of mutually exclusive servers and we expect the reader to be familiar with ref. (1). In section III we present the non-standard approach to the solution of such stochastic difference equations in the limit $t \to \infty$. In section IV we state some results and provide some examples.

It should be said that the author does not know of an adequate treatment, even of the single intermittent server through the usual M/G/I techniques, because of the lack of uniform service distribution : the service distribution depends on the absolute instants at which the service starts and restarts. The nicety of our result is : to reduce such a system to a Markovian one, all that is required is the existence of the limit :

$$\frac{1}{t}\int_0^t g(x)dx \to Cg \qquad (t \to \infty), \tag{2}$$

i.e. the existence of an average value of g (Ie. a limiting ratio : total working time/total time).

I - THE SINGLE INTERMITTENT SERVER.

The queue is a Poisson arrival system of rate λ . We assume that the queue is also Poisson of rate/μ for the servicing, *when it is not idle,* i.e., the probability of leaving the system at time t, $T \leqslant t \leqslant T+\Delta T$ for an infinitesimal ΔT , is

$$p_{\text{leaving}}(T \leqslant t \leqslant T+\Delta T) = \begin{cases} \mu \times \Delta T, & \text{if server is working during the interval } (T,T+\Delta T) ; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

i.e., for an arbitrary infinitesimal T :

$$p_{\text{leaving}}(T \leqslant t \leqslant T+\Delta t) = \mu g(T)\Delta T. \tag{4}$$

Let P(K,t) be the probability of having K customers in the system at time t. The usual analysis yields the following evolution equation, in the birth-death approximation (for h infinitesimal) :

$$(k > 1) \quad P(k,t+h) = 1-\lambda h+o(h)\ (1-\mu g(t)h+o(h))\ P(k,t) + (\lambda h+o(h))\ P(k-1,t) + (\mu g(t)h+o(h))\ P(k+1,t)\ ;$$

$$(k = o) \quad P(o,t+h) = (1-\lambda h+o(h))\ P(o,t) + (\mu g(t)h+o(h)\ P(1,t).$$

Introducing the generating function

$$(5) \quad P(Z,t) = \sum_{k=o}^{\infty} P(k,t)Z^k \ ,$$

We obtain the differential equation

$$(6) \quad \frac{\partial P}{\partial t}\ (Z,t) = \lambda(Z-1)\ P(Znt) + g(t)\ (\frac{1}{Z} -1)\ (P(Z,t)-P(b,t)).$$

The existence of a limiting distribution for P, as $t \to +\infty$ will be discussed later.

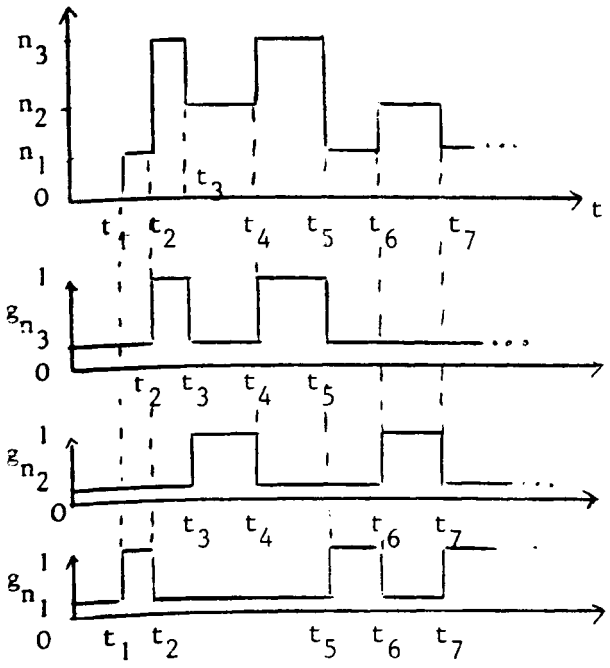
II - JOB SHOP-LIKE QUEUING SYSTEMS WITH GROUPS OF MUTUALLY EXCLUSIVE SERVERS.

We consider a network of queues and service stations numbered I,...,N and we assume a partition of these N servers into p+1 groups G_o, G_1, G_p :

G_o : the set of "independent" servers,

G_i, ..., G_p } groups of mutually exclusive servers.

Independent servers are the classical servers, who work all the time, that is, more precisely, as long as there are customer waiting at their stations. On the contrary, for the servers belonging to the same groupe , say $G_i = \{n_1, n_2, n_3\}$ there is a time allocation mechanism "giving" time in turn to n_1, n_2 or n_3. Such a time allocation function, called $G_i(t)$, typically looks like the one of fig. 3.



(Fig. 3) A typical time allocation function between servers n_1, n_2, n_3. In this example no one works during (t_1, t_2), then n_3 during (t_2, t_3) etc.

(Fig. 4) Time characteristic functions of servers n_1, n_2, n_3 corresponding to the allocation of fig. 3.

Correspondingly each server has its own time characteristic function $g_n(t)$ (see fig . 4) satisfying :

$$G_i(t) = \sum_{n \in G_i} n\, g_n(t)$$

the mutual exclusion translates immediately into :

$$m, n \in G_i,\ m \neq n : g_m(t) . g_n(t) = 0$$

It has to be said that $\sum_{n \in G_i} g_n(t)$ is not necessarily identically one. The intervals during which $\sum_{n \in G_i} g_n(t) = 0$ are those during which none of the servers in G_i work. Note that any server in the group G_0 has a constant time characteristic function $g = 1$ (see fig. 1.b.).

We assume at this point the reader familiar with ref. (1) and its notation.

Through a discussion similar to the one of section I we derive the probability of completing a service at center n, at time t : $T \leqslant t \leqslant T + \Delta T$, assuming there were Kn people at this center at time T :

$$(9)\quad \text{P completing service at center n } (T \leqslant t \leqslant T + \Delta T) = \begin{cases} \mu(n,k_n)\Delta T & \text{if server n is during } (T, T + \Delta T), \\ 0 & \text{otherwise,} \end{cases}$$

i.e.

P completing service $(T \leqslant t \leqslant T + \Delta T) = \mu(n,k_n)\ g(T)\ \Delta T$
at center n.

Then an analysis entirely similar to the one developped in ref (1) leads us to the time dependent state probability :

$$(10)\quad \frac{dP}{dt}(\vec{K},t) = -\Big[\lambda(S(\vec{K})) + \sum_{i=o}^{p} \sum_{Gi} \mu(n,k_n)\ (1-r(n,n))\ g_n(t)\ P(\vec{k},t)$$

$$+ \sum_{i=o}^{p} \sum_{n \in Gi} \lambda(S(\vec{k}) - 1)\ r(o,n)\ P(\vec{h}(n),t)\ .$$

(10)

$$+ \sum_{i=o}^{p} \sum_{n \in Gi} \mu(n,k_n+1)\ r(n,N+1)\ g_n(t)\ P(\vec{i}(n),t)$$

$$+ \sum_{i=o}^{p} \sum_{n \in Gi} \mu(n,d_n+1)\ r(n,m)\ g_n(t)\ P\ (\vec{j}(n,m),t).$$

with the same notations and restrictions as in (1). Formally one introduces two vectors

$$P(t) = (P(\vec{o},t),\ldots,P(\vec{k},t),\ldots)\ ,$$
$$G(t) = (g_1(t),\ldots,g_N(t))$$

and the (infinite) system (10) may be written

$$(11)\quad \frac{dP}{dt} = F_o\ (P) + F_1(P).G(t).$$

Provided that the rates $\lambda(g)$, $\mu(n,q)$ $q=0,1,2,\ldots,n=1,2,\ldots N$ remain bounded the linear operators F_o, F_1, are continuous in P. This observation, though trivial, plays an important role in the search of a solution to (11), as one might expect. On another hand G(t) is a discontinuous function though, it might be considered piecewise continuous, with at most countably many isolated discontinuity points (finite jumps) and remains bounded uniformly over t $(0,\infty)$, i.e., G is certainly measurable , of course the same observation holds of any of the individuals $g_1,\ldots,g_N$ or of the characteristic function of the single intermittent server of section I.

III - ASYMPTOTIC BEHAVIOR OF THE SOLUTION OF THE STOCHASTIC DIFFERENCE EQUATION FOR A SYSTEM WITH INTERMITTENT SERVERS.

Roughly put, to establish the equations for a system with intermittent servers, one uses the following recipe

(i) write down the equations for the classical (i.e. w/o intermittent servers) system

(ii) for each intermittent server, replace its service rate by the product $\mu x\ g(t)$ with its time characteristic function.

This leads to an explicit factorization of the time dependance as exhibited in (6) or (11). Without the explicit mention of parameters other than time and with an appropriate scalar product, the equation we have to study is of the type :

$$\frac{du}{dt} = f_1(u) + f_2(u).\ g(t), \tag{12}$$

where f_i $(i=1,2)$, g are either scalar functions or operators as seen above. All we require for our analysis is :

A - the continuity of f_1, f_2 ,

B - the continuity of the scalar product ,

C - the existence of $C_g = \lim_{t \to +\infty} C_{g(t)} = \lim_{t \to +\infty} (\frac{1}{t} \int_o^t g(X)dX)$.

Proceeding as usual to find an asymptotic behavior to a solution to (12), i.e., solving $\frac{du}{dt} = 0$ does not yield any clue as to what might be, if any, the asumptotic, hence, time independent, function since no time independent function, in general, will satisfy

$$f_1(u) + f_2(u).\ g(t) = o.$$

Let us assume the existence of an asymptotic behavior for a solution u of (12) and let U, F_1 , F_2 , G be the non-standard extensions of u, f_1 , f_2, g respectively. We are specially interested by the behavior of these extensions for infinite values of their argument, t. In particular, the existence of an asymptotic limit for u says :

for some real, infinite, t_o.

for all $t \geqslant t_o$ one has :

(14) (a) $\mathcal{U}(t) \approx V$ (V is the limit of $\mathcal{U}$ if any),

(b) $C_G(t) \approx C_g$,

(c) $F_i(\mathcal{U}(t)) \approx F_i(V)$.

By definition $\mathcal{U}$ is a solution of

$$(12) \qquad \frac{d\mathcal{U}}{dt} = F_1(\mathcal{U}(t)) + F_2(\mathcal{U}(t)).\ G(t)$$

or equivalently of

$$(15) \qquad \mathcal{U}(t) = \mathcal{U}(s) + \int_s^t (F_1(\mathcal{U}(x)) + F_2(\mathcal{U}(x).\ G(x))\ dX.$$

Using the continuity of F_i (i=1,2) and the fact that $\mathcal{U}$ is almost a constant over any interval (s,t) such that $t_o \leqslant s \leqslant t$, we will be able to treat $F_i(\mathcal{U}(x))$ as a constant and write

$$\int_s^t (F_1(\mathcal{U}(x)) + F_2(\mathcal{U}(x)).G(x))\,dx$$

$$\approx F_1(V) \int_s^t dx + F_2(V). \int_s^t G(x)dx$$

$$= (t-s)\left[F_1(V) + F_2(V). \frac{1}{t-s} \int_s^t G(x)\,dx\right]$$

and by a convenient choixe of (s,t) we are going to be able to verify :

$$\frac{1}{t-s} \int_s^t G(X)dx \approx C_g.$$

Whence turning (15) into :

$$\mathcal{U}(t) - \mathcal{U}(s) \approx (t-s).\ (F_1(V) + F_2(V).\ C_g).$$

Thus showing that the limit v of u, if any, satisfies

$$f_1(v) + f_2(v).\ C_g = 0.$$

The mathematical details follow.

The continuity of F_i (i=1,2) may be expressed as

$$x \approx y \rightarrow F_i(x) \approx F_i(y)$$

Let us choose some infinite interval (S,T) with $t_o \leqslant S \leqslant T$;

since $\mathcal{U}(x) \approx V$ on (S,T) we have

(16) $$F_i(\mathcal{U}(x)) - F_i(V) = \varepsilon_i(x) \approx 0 \qquad (i=1,2).$$

(S,T) being compact and $|\varepsilon_i|$ continuous, $|\varepsilon_i|$ reaches its maximum, itself an infinitesimal. Thus there is an *infinite* real Ω satisfying :

$$\Omega \leqslant T - S\ ,$$
$$\Omega \cdot \varepsilon_i(x) \approx 0 \quad \text{all} \quad x \in (X,T) \qquad (i=1,2).$$

For instance take

$$\underline{\Omega} = \inf \left(T-S, \frac{1}{\sqrt{\max_{x \in (S,T)} |\varepsilon_1(x)|}}, \frac{1}{\sqrt{\max_{x \in (S,T)} |\varepsilon_2(x)|}}\right).$$

Let us choose

$$\begin{cases} s = S \\ t = S + \underline{\Omega}, \end{cases}$$

or any other infinite sub-interval $(s,t) \leqslant S,T)$ of length $(t-s) \leqslant \underline{\Omega}$. For any such s, t, $s/t \approx o$ and by (16) :

$$\int_s^t (F_1(U(x)) + F_2(U(x)).\ G(x)dx \approx F_1(V) \int_s^t dx + F_2(V).\ \int_s^t G(x)dx ,$$

$$\int_s^t (F_1(U(x)) + F_2(U(x)).\ G(x))\,dx \approx (t-s)\left[F_1(V) + F_2(V).\ \frac{1}{t-s} \int_s^t G(x)dx\right].$$

More over we are going to verify that $\frac{1}{t-s} \cdot \int_s^t G(x)\,dx \approx C_g(t)$ hence also $\approx C_g$.

For : $\frac{1}{t-s} \int_s^t G(x)dx = \frac{1}{t-s} \left(\int_o^t G(x)dx - \int_o^s G(x)dx\right) = \frac{t}{t-s} C_g(t) - \frac{s}{t-s} C_g(s)$,

where we define, as in (13), $C_G(t) = \frac{1}{t} \int_o^t G(x)dx$;

whence : $\frac{1}{t-s} \int_o^t G(x)dx = \frac{1}{1-s/t} C_G(t) - \frac{s}{t} \cdot \frac{1}{1-s/t} C_G(s)$;

since $s/t \approx o$ and $C_G(s) \approx C_g$ is finite, we cancel the second term on the right and we get :

$$\frac{1}{t-s} \int_s^t G(s)dx \approx C_G(t) \approx C_g.$$

Summarizing :

$$\int_s^t F_1(U(x) + F_2(U(x)).G(x)dx \approx (t-s)\ (F_1(V) + F_2(V).C_g).$$

Plugging this back into (15) we obtain

(17) $\mathcal{U}(t) - \mathcal{U}(S) \approx (t-s).\ (F_1(V) + F_2(V).C_g).$

By (14,a), $\mathcal{U}(t) - \mathcal{U}(s) \approx 0.$

Whence (17) becomes :

$$0 \approx (t-s).(F_1(V) + F_2(V).C_g)$$

and (t-s) being infinite V satisfies

$$F_1(V) + F_2(V).\ C_g \approx 0\ .$$

Taking the standard parts we obtain :

(18) $$f_1(v) + f_2(v).C_g = 0\ .$$

IV - LIMITING STATE PROBABILITY DISTRIBUTION FOR QUEUING SYSTEMS WITH INTERMITTENT SERVERS.

In what follows one assumes that each server has a time characteristic function g(t) as represented in fig. 1 satisfying :

$$\lim_{t\to\infty} \frac{1}{t} \int_o^t g(x)dX = C_g < \infty.$$

The above condition is the one used in the previous section. It guarantees that the systems of section I and II, for instance if they have limiting state distributions, then these are the same as the limiting state distributions for :

(6') $$\frac{\partial p}{\partial t} = \lambda(Z-1)\ P(Z,t) + \mu cg\ (\frac{1}{Z} - 1)\ (P(Z,t) - P(o,t))$$

or

(11') $$\frac{dp}{dt} = F_o(P) + F_1(P).\ C_G.$$

a) SINGLE SERVER.

The limiting state probability distribution of the single server is the solution of :

(6") $\lambda(Z-1)\ P(Z,t) + \mu C_g\ (\frac{1}{Z} - 1)\ (P(Z,t) - P(o,t)) = o.$

Using the normalization condition $p_{Lim}(1) = 1$, (6") yields.

(19) $P_{Lim}(Z) = \frac{1-\rho g}{1-Z\rho g}$, $\rho_g = \frac{\lambda}{\mu C_g}$.

That is

(20) $P_{Lim}(K) = (1-\rho g)\rho g^K$.

As expected the utilization factor is adjusted from the classical single server, for wasted time : c_g is the limiting value of the ratio :

working time/total time.

Indeed $c_g \leqslant 1 \Rightarrow \rho_g \geqslant \lambda/\mu$. This represents an increase in the business of the server : to serve the same number of customers in less real time, the server will have to be busy more often.

b) NETWORKS, JOB-SHOP-LIKE SYSTEMS.

The equilibrium of the system with groups of mutually exclusive servers, if any, will be identical to the one of the system obtained by substituting Cg_n to $g_n(X)$. In order words, this solution is the solution given in réf (1) using for service rates, instead of $\mu(n,Kn)$, the real rates $\mu(n,Kn)Cg_n$.

Let $W(K) = \prod_{i=o}^{K-1} \lambda(i)$ for $K = o,1,2,\ldots$

$$\omega_G(K) = \prod_{i=o}^{P} \prod_{n \in G_i} \prod_{j=1}^{Kn} \left(\frac{e(n)}{\mu(n,j)Cg_n}\right) \qquad \text{for } \vec{K}=K_1,\ldots,K_n,\ldots,K_N ,$$

$$T_G(K) = \sum_{S(\vec{k})=K} \omega_G(\vec{k}) \qquad k = o,1,2,\ldots ,$$

$$\Pi_G = \frac{1}{\sum_{K=o}^{\infty} W(K)\ T_G(K)} \qquad \text{if sum converges and} \quad \Pi_G = o \text{ otherwise .}$$

Then if $\Pi_G > o$, a unique equilibrium exists and is given by

$$(21) \qquad P_{Lim}(\vec{K}) = \Pi_G \omega_G(\vec{K})\ W(s(\vec{k})).$$

REFERENCES AND NOTES.

(1) J.R. JACKSON, Jobshop-like Queuing systems, Management Sciences, 10 (1963).

(2) Currently associated with Citibank, N.Y.

(3) Currently at N.Y.U.

(4) L. KLEINROCK, Queuing systems, Vol. 1, J. Wiley and sons (1975).

(5) J. KEISLER, Elementary Calculus, Prindle and Co (1976).

(6) M. DAVIS, Applied Non-standard Analysis, J. Wiley and sons (1977).

(7) K.D. STROYAN, WAJ Luxemburg, Introduction to the theory or Infinitesimals Academic Press (1976).

The presentation given here owes much to the discussions I had in several places where these ideas were exposed. Amongst them at N.Y.U., Lyon's and Grenoble's University and at the IREM in Paris.

J.P. KELLER
Département de Mathématiques
Université Claude Bernard
43, bd du 11 novembre 1918
69621 VILLEURBANNE