REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

M. KRAKOWSKI

Conservation methods in queuing theory

Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle, tome 7, n° V1 (1973), p. 63-83

<http://www.numdam.org/item?id=RO_1973__7_1_63_0>

R.A.I.R.O.
(7e année, V-1, 1973, p. 63-84)

CONSERVATION METHODS IN QUEUING THEORY

par M. KRAKOWSKI

Abstract. — *We apply a procedure, the method of invariants or conservation method, to derive several classical and new (as far as could be ascertained) results pertaining to the operating characteristics of stationary queuing systems. Thus, the relation* $P_0 = 1 - \rho$ *for single channel systems is a special instance of the Principle of Customers Conservation. This principle applies to each category of customers separately. The well known relation* $L = \lambda W$ *expresses the conservation of Total System Seniority, i.e. the cumulated time within the system of all present customers. A similar formula applies to each type of customers. The above conservation principles are applicable also to systems with reneging. A quantity whose conservation is examined is the sum of the squares of the sojourn times of the customers within the system. The conservation method is also applied to the service system. The steady-state equations of balance across suitable cuts are interpreted as conservation rules.*

NOTATION

(excluding several symbols used in a single section)

λ = frequency of arrivals

μ = frequency of servicing in a single channel under full-load conditions; $1/\mu$ is the average service time

ρ = λ/μ

α = reneging propensity within a queue; cf. section 3

S_k = state of the system in which there are k customers

P_k = probability of S_k

r.v. = abbreviation for « random variable »

$\tilde{L}$ = number of customers within the system (r.v.)

L = $E(\tilde{L})$

$\tilde{l}$ = number of customers in the queue (r.v.)

l = $E(\tilde{l})$

$\tilde{W}$ = sojourn of a customer in the system from entry till departure (r.v.)

W = $E(\tilde{W})$

$\tilde{w}$ = sojourn of a customer in the queue (r.v.)

w = $E(\tilde{w})$

$\tilde{t}_k$ = seniority of incumbent $\# k$ (r.v.); cf. section 2

$\tilde{T}$ = $\Sigma \tilde{t}_k$, Total System Seniority (r.v.); cf. section 2

T = $E(\tilde{T})$

$\tilde{S}$ = Total Seniority Within Service Channels (r.v.); cf. section 4

S = $E(\tilde{S})$

$\tilde{X}$ = service time (r.v.); $E(X) = 1/\mu$

$\tilde{R}$ = remaining service time of the customer in a given channel at a random instant; defined for busy channels only (r.v.)

R = $E(\tilde{R})$; R is also equal to the average sojourn time of the serviced customer at a random instant, as can be shown

c = number of channels, may be a random or a control variable

Q = probability that a given service channel is occupied.

The term « queue » in this paper refers to the part of the queuing system awaiting entry into the service channel(s).

The Kendall notation for the description of queuing systems is frequently used in this paper.

INTRODUCTION

Conservation principles have played a fundamental role in physical sciences, engineering, and economics. Conservation of mass, energy, momemtum, charge, etc. often provide first integrals to a system of equations. Where the equations are few in number the conservation laws alone may suffice to find the solution.

In stationary queuing systems there are obviously quantities which satisfy a conservation rule. A customer entering the system must leave it. The accumulated sojourn times of the customers must fluctuate so that an expected change during a random interval is zero. The same is true about many other system functions.

This paper attempts to systematize and exploit the basic conservation rules in order to derive gross characteristics of queuing systems and first integrals to the equations of state.

Several classical results are shown to be conservation statements or their consequences, e.g. $P_0 = 1 - \rho$, $L = \lambda W$, and the Pollaczek-Khinchine formula for the queue length in an $M/G/1$ system. Extensions of these results and several others are also derived. A more flexible way to set up the equations of balance for Markovian systems is shown in section 6.

No effort was made at mathematical rigor the stress being on the intuitive and perceptible even if this means that the status of some proofs is that of plausibility arguments to the purists.

SECTION 1. CONSERVATION OF CUSTOMERS

An assumption made more often implicitly than explicitly in the treatment of queuing systems is that a customer entering the service booth eventually leaves it, when the service is completed. Under stationary conditions the somewhat weaker statement that the *frequency of entries into a service channel equals the frequency of departures out of this channel* will be more suitable for our exposition.

The above statement will be referred to as the *Principle of Customer* (or entry) *Conservation*. Where a distinction is useful, one can refer to the preservation of individual customers as the Strong Principle of Customer Conservation, whereas the equality of entry and departure frequencies is the corresponding Weak Principle of Customer Conservation. In this paper only the weak form of the Principle will be used and the adjective « weak » will be omitted.

A remarkable thing about this principle is that, to the best of my knowledge, it has not been formally stated in any of the common references or text-books. This despite the fact that conservation principles play such a prominent role in physics, engineering, economics, and other branches of learning and of the arts.

The ease of application and insight provided will be elucidated by the consideration of several examples.

Unless otherwise specified, it will be assumed that no server is ever idle when there is an unserved customer in the system.

EXAMPLE 1.1. For a $G/G/1$ queuing system the frequency of arrivals is λ and the frequency of departures is the full-load frequency μ multiplied by the probability that the system is not empty, i.e. $1 - P_0$. Therefore, our conservation principle yields

$$\lambda = (1 - P_0)\mu, \tag{1.1}$$

and

$$P_0 = 1 - \lambda/\mu = 1 - \rho. \tag{1.2}$$

The same result is usually obtained for $M/M/1$ by expressing each P_n in terms of P_0 and then summing all the state probabilities to one.

EXAMPLE 1.2. Consider now a system $G/G/c$, all channels being identical. The frequency of entries into a given channel is ρ/c; the frequency of departures out of this channel is the full-load service frequency μ multiplied by the probability Q that this channel is in service. Therefore, under symmetrical operating conditions,

$$\lambda/c = \mu Q \tag{1.3}$$

and the probability of a given server being occupied is

$$Q = \lambda/c\mu = \rho/c. \tag{1.4}$$

Note that Q can also be interpreted as the expected number of customers in a given service booth since this number is a zero-one random variable.

Therefore the expected number of busy servers is cQ;

$$\text{expected } \# \text{ of busy servers} = cQ = \rho. \tag{1.4 a}$$

We can easily write down the probability that (under symmetric operating conditions) a given service station is free :

$$1 - Q = P_0 + \frac{c-1}{c} P_1 + \frac{c-2}{c} P_2 + \ldots + \frac{1}{c} P_{c-1}. \tag{1.5}$$

It follows from (1.5) and (1.4) that

$$\rho = c(1 - P_0) - (c-1)P_1 - (c-2)P_2 - \ldots - P_{c-1}. \tag{1.6}$$

When $c = 1$ we get (1.1).

The principle of conservation of customers applies separately to each category of customers, e.g. when these are classified by age or gender.

EXAMPLE 1.3. Let there be n types of customers, the kind k arriving with the frequency λ_k and its service time being of average duration $1/\mu_k$. In a $G/G/1$ system we can write

$$\lambda_k = Q_k \cdot \mu_k \tag{1.7}$$

and

$$Q_k = \lambda_k/\mu_k = \rho_k \tag{1.8}$$

where Q_k is the probability that the service booth is in state k, i.e. that it has a customer of type k in it.

The probability that the system is empty is

$$P_0 = 1 - \sum_{k=1}^{n} \lambda_k/\mu_k = 1 - \sum_{1}^{n} \rho_k. \tag{1.9}$$

EXAMPLE 1.4. We extend example 1.3 by having c identical channels operated under symmetrical conditions. The input intensity of the k-th type of customer to a given channel, say i, is λ_k/c.

Therefore

$$\lambda_k/c = Q_{ik} \cdot \mu_k \tag{1.10}$$

and

$$Q_{ik} = \frac{1}{c}\rho_k, \qquad \rho_k = \lambda_k/\mu_k, \tag{1.11}$$

where

Q_{ik} = probability that i-th channel has a customer of type k.

Note that (1.11) is independent of i.

The probability that a *given* channel is occupied is

$$Q = \sum_{k=1}^{n} \rho_k/c \qquad \text{where} \qquad \rho_k = \lambda_k/\mu_k. \tag{1.12}$$

Since Q can also be interpreted as the expected number of customers in a given channel the expected number of customers in all the service stations is cQ, and therefore

$$\text{the expected } \# \text{ of occupied service stations} = \sum_{1}^{n} \rho_k. \tag{1.13}$$

EXAMPLE 1.5. Consider three service booths in tandem, with a common server, the service in each booth requiring on the average $1/\mu_i$ time units, $i = 1, 2, 3$. The intensity of arrivals is λ.

Applying the principle of conservation of customers to each booth separately, and denoting by Q_i the probability that booth $\# i$ is occupied, we obtain

$$\lambda = Q_1 \cdot \mu_1 = Q_2 \cdot \mu_2 = Q_3 \cdot \mu_3. \tag{1.13}$$

Therefore

$$Q_i = \lambda/\mu_i \quad \text{with} \quad i = 1, 2, 3. \tag{1.13 a}$$

The probability that the server is idle is

$$P_0 = 1 - \lambda/\mu_1 - \lambda/\mu_2 - \lambda/\mu_3. \tag{1.14}$$

Of course, instead of three booths with a common server we could have one operation, in one booth, composed of three successive phases each of duration $1/\mu_i$.

EXAMPLE. 1.6. As in example 1.5 there are still three serviceb ooths in tandem but now each booth has its own server and waiting lines may form between the booths. Now the service system may contain up to three customers in service, although at most one in each booth.

If Q_i is the probability that booth i is occupied, $i = 1, 2, 3$. then the principle of customer conservation, applied to each booth, gives

$$\lambda = Q_1 \cdot \mu_1 = Q_2 \cdot \mu_2 = Q_3 \cdot \mu_3, \tag{1.15}$$

similar in form to 1.13, despite the different servicing system.

EXAMPLE 1.7. Consider now a queuing system $M/G/c$ symmetrically operated. A customer of rank n is defined, in an ad hoc manner, as one who encountered n other customers in the system upon arrival. *Question :* What is the probability Q_n that a *given* service channel has a customer of rank n ?

Solution. The total frequency of arrivals of rank-n customers is λP_n, and out of these $\lambda P_n/c$ reach the given channel, as follows from the symmetry of operations. The frequency of departures of rank-n customers out of our channel is $Q_n \cdot \mu$. Therefore, applying the principle of customer conservation to those of rank n, we get

$$\lambda P_n/c = Q_n\mu \tag{1.16}$$

and

$$Q_n = \lambda P_n/c\mu \tag{1.17}$$

The derivation of (1.17) by other means would be more laborious.

EXAMPLE 1.8. In this example we will show how the Principle of Customers Conservation can be used to reduce the recursion base. Consider a queuing system where the customers arrive in a Poissonian manner with frequency λ but are served only in couples. Thus, the server is idle when the system is in state zero or one. The service duration per couple is expected to be $1/\mu$ and is exponentially distributed.

The equations of balance are

$$\sum_0^\infty P_k = 1 \tag{1.18}$$

$$\lambda P_0 = \mu P_2 \tag{1.19}$$

$$\lambda P_1 = \lambda P_0 + \mu P_3 \tag{1.20}$$

$$(\lambda + \mu)P_k = \lambda P_{k-1} + \mu P_{k+2} \qquad \text{for} \qquad k \geqslant 2. \tag{1.21}$$

It can be easily verified that the above system of equations has a recursion base P_0 and P_1; in other words given P_0 and P_1 the remaining state probabilities can be computed recursively. This recursion base can be reduced to P_0 only, using the Principle of Customer Conservation. We have namely

$$\lambda = (1 - P_0 - P_1)\,\mu, \text{ i.e. } P_1 = 1 - P_0 - \rho. \tag{1.22}$$

The reduction of the recursion base to P_0 simplifies the numerical work connected with the determination of the state probabilities from the equations of balance.

SECTION 2. CONSERVATION OF SYSTEM SENIORITY

We define the Total System Seniority, or briefly System Seniority, as the cumulative time spent already in the system by all customers present there at the survey instant. Denoting this r.v. by $\tilde{T}$, and the time spent within the system by incumbent customer $\neq k$, as of the survey instant, by $\tilde{t}_k$, we have

$$\tilde{T} = \Sigma \tilde{t}_k \quad \text{(summing over all incumbent customers)} \tag{2.1}$$

$$\tilde{T} = 0 \quad \text{when the system is in state } S_0.$$

Let L and W be the average number of customers in the system and the average completed sojourn time of a customer, respectively. The sojourn can be terminated by completion of service, by reneging, or by any termination of service prior to its completion.

The expected change in the seniority $\tilde{T}$ during dt due to aging equals Ldt; the expected decrease in $\tilde{T}$ due to departing customers is $\lambda W dt$. The contribution to $\tilde{T}$ of new customers, arriving during dt is of the order dt^2, as is easy to verify. Under stationary conditions the net change in $\tilde{T}$ should vanish, on the average. Therefore

$$Ldt = \lambda W dt \quad \text{and} \quad L = \lambda W. \tag{2.2}$$

The derivation of

$$l = \lambda w \tag{2.3}$$

where l is the average queue length and where w is the expected time of sojourn in the queue, is analogous to that of (2.2) and will be omitted here.

(2.2) and (2.3.) are often referred to as Little's formulas ; cf. Bibliography.

It can be immediately recognized that (2.2) and (2.3) are valid for each type of customer separately when there are several such types. Then, for the k-th type of customers

$$L_k = \lambda_k W_k, \tag{2.2 a}$$

where λ_k is the arrival frequency of the customers of type k, and

$$l_k = \lambda_k w_k. \tag{2.3 a}$$

The proofs of (2.2 *a*) and of (2.3 *a*) follow from the conservation of the seniority of the type-k customer. They are virtually identical to those of (2.2) and (2.3) and will be omitted. We have proved the special case $k = 1$ instead of the more general one for the sake of clarity in exposition.

The physics and economics of (2.2), (2.3), (2.2 *a*), and (2.3 *a*) are perfectly clear. These formulas make still sense when the number of service channels, c, is a random or a control variable in a system $G/G/c$ and reneging causes no difficulty. All that is needed is stationarity in some very weak sense. The underlying physical interpretation, namely the conservation of the accumulated sojourn time, explains the wide validity of the formulas discussed in this section.

It is clear that $L = l +$ expected number of occupied channels. Taking into account (1.4 *a*) we have

$$L = l + \rho, \quad \text{where} \quad \rho = \lambda/\mu. \tag{2.4}$$

(Note that in our notation μ is the full-load frequency for a single channel.)

The same reasoning relating expected changes in $\Sigma\tilde{t}_k$ to expected changes in $\tilde{W}$ can be applied to changes in $\Sigma f(\tilde{t}_k)$ where $f(\ )$ is an arbitrary differentiable function. We then have

$$E[\Sigma f'(\tilde{t}_k) = E[f(\tilde{W}) - f(0)] \tag{2.5}$$

the summation being over all incumbent customers. In section 5 we will treat the case $f(t) = t^2$ in more detail.

EXAMPLE 2.1. Consider a system $M/M/c$ where the c channels have expected service times $1/\mu_j$, j = 1, ..., c. Under full-load conditions the interdeparture times are exponentially distributed with frequency $\sum_1^c \mu_j$.

The expected time, w which a newcomer spends awaiting service is composed of :

a) the expected time till a channel becomes available for the customer next in line ; this is equal to $P(n \geqslant c)/\sum_1^c \mu_j$, $P(n \geqslant c) = \sum_c^\infty P_n$ being the probability that all channels are occupied.

b) the expected service time of the customers encountered by the newcomer in the queue ; this time is $l/\sum_1^c \mu_j$.

We assumed in *a)* and *b)* that a newcomer encounters n other customers in the system with the probabilty P_n and finds a queue of expected length l ; cf. Réf. 3 for a proof that this holds in $M/G/c$.

It follows from *a)* and *b)* that

$$w = P(n \geqslant c)/\sum_1^c \mu_j + l/\sum_1^c \mu_j.$$

This equality aiong with $l = \lambda w$ resuits in

$$(2.6) \qquad l = \frac{\rho}{c - \rho} P(n \geqslant c) = \frac{\lambda}{\sum_1^c \mu_j - \lambda} P(n \geqslant c), \text{ where } \rho = \frac{c\lambda}{\sum_1^c \mu_j}.$$

The transitions from S_k to S_{k+1} being as frequent as from S_{k+1} to S_k (cf. Section 6) it follows that

$$P_{k+1} = \frac{\rho}{c} P_k \text{ for } k \geqslant c$$

and

$$P(n \geqslant c) = \frac{c}{c - \rho} P_c$$

In order to express P_c (and the other state probabilities) in terms of λ, μ_j, and c it is necessary, unless all the μ_j are equal, to know the server discipline, i.e. the precedence, preemptive or otherwise, among the servers of different working speeds. In the case where only the speediest servers are occupied when some channels are empty (a slower server will always turn over his job to an idle speedier one) we have a birth-and-death process which can be easily handled with the method of Section 6. This is also the case when all servers are of equal skill, i.e. when we have a system $M/M/c$.

When $c = 1$, i.e. when we deal with the case $M/M/1$, (2.6) becomes

$$(2.6\,a) \qquad l = \frac{\rho^2}{1 - \rho}.$$

since $P_1 = \rho P_o = \rho(1 - \rho)$ implies $P(n \geqslant c) = \rho$.

EXAMPLE 2.2. We will extend now the case $M/M/1$ to the system $M/G/1$. The argument for the duration of w runs as before in example 2.1 but the expected time interval for the service channel to be cleared, conditional upon its being occupied, is now designated as R ($R = 1/\mu$ in example 2.1, for the negative-exponential holding times). R is a holding time characteristic and we will have more to say about it in section 4.

We have now therefore for $M/G/1$

$$(2.7) \qquad w = \rho R + l\frac{1}{\mu} = l/\lambda$$

and

$$(2.8) \qquad l = \frac{\lambda \rho R}{1 - \rho}.$$

This is the well known Pollaczek-Khinchine formula in its linear form cf. ref. [3]. (Substituting (4.6) into (2.8) we get the usual quadratic form.)

It is easy to verify that for exponential holding times (2.8) becomes (2.6). For constant holding times of duration a we have $R = a/2$ and (2.8) becomes

$$l = \frac{1}{2} a\lambda\rho/(1-\rho) \quad ; \quad M/D/1. \tag{2.9}$$

The derivation of the Pollaczek-Khinchine formula in this section provides, in my opinion, a clearer intuitive grasp of the dependence of the waiting time upon the deviation of the holding time from a constant value than do the usual derivations of the quadratic form.

EXAMPLE 2.3. A customer of rank n is defined as in example 1.7. Let L_n be the expected number of rank-n customers in the queuing system. The problem is to determine L_n for $M/M/c$.

The solution is simple if one uses the formula (2.2 a) the rank-k customer playing the role of type k. The frequency of arrivals of rank-k customers is obviously λP_k. The sojourn time of a rank-k customer is expected to be

$$W_k = 1/\mu \quad \text{when} \quad k < c \tag{2.10}$$
$$= (k - c + 1)/c\mu + 1/\mu = (k+1)/c\mu, \quad \text{when} \quad k \geqslant c.$$

Therefore, with $L_k = \lambda_k W_k = \lambda P_k W_k$ we get

$$L_k = \rho P_k \quad \text{when} \quad k < c \tag{2.11}$$
$$= (k+1) \cdot \rho \cdot P_k/c \quad \text{when} \quad k \geqslant c.$$

For $k < c$ we have, as known from the balance equations (cf. section 6), $\lambda P_k = (k+1)\mu P_{k+1}$ and $\rho P_k = (k+1)P_{k+1}$.

For $k \geqslant c$ we have $\lambda P_k = c\mu P_{k+1}$, i.e. $P_{k+1} = \rho P_k/c$.

Therefore,

$$L_k = (k+1)P_{k+1} \quad \text{for} \quad k \geqslant 0. \tag{2.12}$$

Note that (2.12) is independent of c, the number of channels.

SECTION 3. CONSERVATION OF CUSTOMERS AND OF SENIORITY UNDER CONDITIONS OF RENEGING

We modify now the system $G/G/c$ by introducing a reneging propensity of Poissonian intensity α for customers in the queue. That is, each customer in the queue (but not in a service channel) has a probability αdt of reneging within the time dt.

The expected number of new arrivals during the interval dt is $\lambda\, dt$. The expected number of customers reneging during dt is $l\alpha\, dt$. The expected number

of customers to enter the service channels during dt is therefore $(\lambda - l\alpha)\,dt$; in other words, the frequency of entries into the service system is $\lambda - l\alpha$. The frequency of departures out of the service system is $cQ\mu$, where Q is the probability that a given channel is occupied. Therefore,

$$\lambda - l\alpha = cQ\mu \tag{3.1}$$

and

$$l = (\lambda - cQ\mu)/\alpha \tag{3.2}$$

where cQ is the expected number of occupied channels.

Q can be expressed in terms of the state probabilities P_i, $i \leqslant c - 1$, as was done in (1.5).

Thus, again a first integral was obtained connecting l and the state probabilities $P_0, P_1, \ldots P_{c-1}$.

During dt the expected increase in the total queue seniority due to aging is $l\,dt$; the expected decrease due to entries into the service booths is $(\lambda - l\alpha)w_1\,dt$; the expected decrease due to customers reneging out of the queue is $l\alpha w_2 dt$, where w_1 and w_2 are the expected sojourn times in the queue of customers who reach a service channel and those who reneg, respectively,

Therefore,

$$l = (\lambda - l\alpha)w_1 + l\alpha w_2. \tag{3.3}$$

Eliminating l from (3.2) and from (3.3) we get

$$\alpha cQ\mu w_1 = (\alpha w_2 - 1)(\lambda - cQ\mu), \tag{3.4}$$

a relation connecting w_1, w_2, and Q.

For a system $M/M/c$ modified by the reneging propensity α as defined above the transitions form a birth-and-death process, as can be easily seen. In particular, one finds that for $M/M/1$ with the reneging propensity α there is a recursive relation (cf. example 6.3)

$$P_{k+1} = P_k\lambda/(\mu + k\alpha). \tag{3.5}$$

This allows us to determine, algorithmically or explicitly, $Q = 1 - P_0$ and $l = \sum_1^\infty (k-1)P_k$.

Denoting by w the expected sojourn time of a customer in the queue, terminated by entrance into the service station or be reneging, we have

$$l = \lambda w. \tag{3.6}$$

w is a weighted average of w_1 and of w_2 and it is clearly

$$w = [cQ\mu w_1 + (\lambda - cQ\mu)w_2]/\lambda \tag{3.7}$$

Note that the above relations do not allow to solve for w_1 and for w_2 in terms of the state probabilities and of l (which itself is so expressible).

Equation (2.3 *a*) enables one to relate, under wide conditions, say $G/G/c$, the expected number l_1 of successful customers (those who will be serviced eventually), to their expected waiting time for service w_1. Similarly one can relate the average number l_2 of unsuccessful customers in the queue to their average sojourn before reneging, w_2. The arrival rate of successful customers equals the departure rate out of the service channels, $cQ\mu$, and the arrival rate of the unsuccessful customers is $\lambda - cQ\mu = \alpha l$. Therefore, for $G/G/c$

$$l_1 = cQ\mu w_1 = c(1 - P_0)\mu w_1 \tag{3.8}$$

$$l_2 = (\lambda - cQ\mu)w_2 = \alpha l w_2.$$

SECTION 4. CONSERVATION METHOD APLIED TO THE SERVICE SYSTEM

Let the random variable $\tilde{S}$ (total channel seniority) be the cumulated time spent already in the service channels by all customers who are being serviced at the survey instant.

The expected increase in $\tilde{S}$ during a « random » dt is σdt, σ being the average number of occupied channels; this increase is due to aging. The expected decrease during a « random » dt, due to departures, is the overall rate of departures (equal to the overall rate of arrivals $= \lambda$) multiplied by the average duration of a completed service span, i.e. $1/\mu$, and by dt. Other contributions are of the order dt^2, as is easy to see.

Therefore

$$\sigma = \lambda \frac{1}{\mu} = \rho \text{ (same as (1.4 } a\text{))} \tag{4.1}$$

This result was derived by using the principle of customer conservation in section 1. It is interesting to note that the conservation of different physical quantities may lead to the same result.

Similar considerations apply to the case when there are several types of customers. We then have a conservation rule for each type of customers. This will, in the case of seniority conservation, lead to results already obtained in section 1.

Consider now the random variable $\tilde{S}^n$, $n > 1$, and its expected change during dt. Let the random variable $\tilde{X}$ be the holding time in a channel; $E(\tilde{X}) = 1/\mu$, of course. Then, the expected increase in $\tilde{S}^n$ due to aging is $nE(\tilde{S}^{n-1})dt$, and the expected decrease due to departures is $\lambda E(\tilde{X}^n)dt$. Other

contributions are of the order dt^2. Therefore, stationarity implies the balancing of the two expected changes, and we have

$$nE(\tilde{S}^{n-1}) = \lambda E(\tilde{X}^n). \tag{4.2}$$

In particular, for $n = 2$ we get, with $S = E(\tilde{S})$

$$2S = \lambda\,(\mathrm{var}\,\tilde{X} + X^2) \quad ; \quad X = E\,(\tilde{X}) = 1/\mu. \tag{4.3}$$

(Note that $\tilde{S} = 0$ when the system is empty. We require $n > 1$ to assure $S^{n-1} \to 0$ as $S \to 0$.)

Consider now a *single channel under full-load conditions.* The successive service intervals form now a renewal process. Let $\tilde{X}$ be defined as above, and let the random variable $\tilde{R}$ be the sojourn time of the incumbent customer as of the survey instant. Let $R = E(\tilde{R})$. How does now $\tilde{R}^n$ change during a random dt? The increase due to aging is clearly expected to be $nE(\tilde{R}^{n-1})\,dt$; the decrease due to departures is expected to be $\mu E(\tilde{X}^n)\,dt$. (Note that under full-load conditions the rate of departures is μ, and not λ.)

Therefore, with $n > 1$,

$$nE(\tilde{R}^{n-1}) = \lambda E(\tilde{X}^n). \tag{4.4}$$

For $n = 2$ we get

$$2E(\tilde{R}) = 2R = \mu\,(\mathrm{var}\tilde{X} + X^2) \quad ; \quad X = 1/\mu. \tag{4.5}$$

and

$$R = (\mathrm{var}\,\tilde{X} + X^2)/2X. \tag{4.6}$$

The expectation R, like X and $\mathrm{var}\,\tilde{X}$, is a characteristic of the random variable $\tilde{X}$, of the corresponding renewal process, or of the corresponding stationary population, and not of the complete queuing system along with its discipline. The relation (4.6) can be applied to other renewal processes and their completed and uncompleted intervals, e.g. where $\tilde{X}$ is the total sojourn time in the queuing system of a departing customer and $\tilde{R}$ is his sojourn time as of a random survey instant. (Or, $\tilde{X}$ can be the interarrival time and $\tilde{R}$ the time length since last arrivals as of a random survey instant). In a stationary population the random variable $\tilde{X}$ is a complete life-span and $\tilde{R}$ is the age. Note that the random variable « remaining lifespan » has the same distribution as $\tilde{R}$ under stationary conditions.

There is a close relation between the probability density functions of the random variables $\tilde{X}$ and $\tilde{R}$. It can be derived from (4.4) but we will infer it from slightly more general considerations.

Let

$$\begin{aligned} p(t) &= \text{probability density function of } \tilde{R} \\ f(t) &= \text{probability density function of } \tilde{X}. \end{aligned} \tag{4.7}$$

Let $g(t)$ be differentiable and not increase too rapidly with increasing t; it will suffice if $g(t)p(t)$ tends to zero as t tends to infinity. Otherwise $g(t)$ is arbitrary. During dt the change in $g(\tilde{R})$ is expected to be $E[g'(\tilde{R})]\,dt$, as a result of the aging; the change due to departures is expected to be

$$\mu E[g(\tilde{X}) - g(0)]\,dt.$$

Other contributions being of order dt^2 the balance equation is

$$E[g'(\tilde{R})] = \mu E[g(\tilde{X}) - g(0)] \tag{4.8}$$

(The multiplier μ is used because we still refer to full-load conditions). Taking into account (4.7) we get from (4.8)

$$\int_0^\infty g'(x)p(x)\,dx = \mu \int_0^\infty g(x)f(x)\,dx - \mu g(0). \tag{4.9}$$

The first integral can be written $p(x)g(x)\Big|_0^\infty - \int_0^\infty g(x)p'(x)\,dx$. Since $p(x)g(x)$ is assumed to tend to zero as x tends to infinity (4.9) becomes

$$-p(0)g(0) - \int_0^\infty g(x)p'(x)\,dx = \mu \int_0^\infty f(x)g(x)\,dx - \mu g(0). \tag{4.10}$$

Since $g(x)$ is arbitrary (apart from a mild infinity and differentiability conditions) (4.10) can be satisfied only if

$$\mu = p(0) \tag{4.11}$$

and

$$p'(t) + \mu f(t) = 0. \tag{4.12}$$

SECTION 5. TOTAL SENIORITY AND VARIANCE OF SOJOURN TIMES

Consider a queuing system $G/G/c$ where c can be a random or a control variable and where reneging of waiting customers is possible. The random variables $\tilde{t}_k$ and $\tilde{T}$ are defined as in section 2, the numbering of the customers being for definiteness in order of descending seniority. When there happen to be m customers in the system then $\tilde{t}_a = 0$ for $a > m$.

Let's define

$$T = E(\tilde{T}) \tag{5.1}$$

and the random variable

$$\tilde{T}^{(2)} = \Sigma \tilde{t}_k^2 , \tag{5.2}$$

summed over the present customers ;

$$\tilde{T}^{(2)} = 0$$

when the system is empty.

The expected increase in $\tilde{T}^{(2)}$ during dt due to aging is $2T\,dt$. The expected decrease in $\tilde{T}^{(2)}$ due to departing customers is $\lambda E(\tilde{W}^2)\,dt$ where the r.v. $\tilde{W}$ is the completed sojourn time of a customer within the system; this sojourn may terminate due to completion of service, to dismissal, to reneging, or to quitting by the servers. Other contributions are of the order dt^2. Stationarity requires that $E[d\tilde{T}^{(2)}] = 0$ and we get the balance equation

$$2T = \lambda E(\tilde{W}^2) = \lambda[\text{var } \tilde{W} + W^2] \quad ; \quad W = E(\tilde{W}). \tag{5.3}$$

Let the random variable ${}^*t_k$ be the remaining sojourn time of customer $\neq k$; let ${}^*\tilde{T} = \Sigma\, {}^*\tilde{t}_k$ (summed over incumbent customers); let ${}^*\tilde{T}^{(2)} = \Sigma {}^*\tilde{t}_k^2$ (summed over incumbent customers). Then it is clear from symmetry (and is easy to prove directly) that

$$E[2({}^*\tilde{T})] = \lambda E(\tilde{W}^2) = \lambda\,[\text{var } \tilde{W} + W^2]. \tag{5.3 a}$$

Case G/G/1 (no reneging)

Let T_k be the expected remaining total service time at a « random » instant, conditional upon there being k customers in the system. It is easy to see that

$$T_k = kR + \frac{k-1}{\mu} + \frac{k-2}{\mu} + \ldots + \frac{1}{\mu} = kR + \frac{k(k-1)}{2\mu}. \tag{5.4}$$

The term kR is the expected remaining time R of service of customer $\neq 1$ (inside the booth), multiplied by the number of customers present in the system; the term $(k-1)/\mu$ is the expected service time of the customer $\neq 2$ (first in the queue) multiplied by the number of customers *not ahead of him* (this includes himself); etc. Therefore, denoting by $\tilde{L}$ the random variable « $\neq$ of customers in the system », we get

$$T = \sum_1^\infty P_k T_k = \sum_1^\infty P_k \left[kR + \frac{k(k-1)}{2\mu} \right] = RL + \frac{1}{2\mu}[E(\tilde{L}^2) - L]. \tag{5.5}$$

Combining (5.3) and (5.5) we get

$$\lambda E(\tilde{W}^2) = \frac{1}{\mu} E(\tilde{L}^2) + L\left(2R - \frac{1}{\mu}\right). \tag{5.6}$$

Multiplying (5.6) by λ we can write it in dimensionless form :

$$E[(\lambda\tilde{W})^2] = \rho E(\tilde{L}^2) + 2\lambda RL - \rho L \tag{5.6 a}$$

or

$$\text{var}\,(\lambda\tilde{W}) + L^2 = \rho\,\text{var}\,\tilde{L} + \rho L^2 + 2\lambda RL - \rho L \tag{5.7}$$

and hence

$$\rho\,\text{var}\,\tilde{L} - \text{var}\,(\lambda\tilde{W}) = (1-\rho)L^2 - L(2\lambda R - \rho). \tag{5.8}$$

Corresponding to (5.6 *a*) there is an expression in terms of the queue length $\tilde{l}$ and the waiting time for service $\tilde{w}$:

$$E[(\lambda\tilde{w})^2] = \rho E(\tilde{l}^2) - \rho l + 2\lambda lR\,;\ l = E(\tilde{l}) \text{ and } w = E(\tilde{w}). \tag{5.9}$$

Corresponding to (5.8) we have an expression in $\tilde{l}$ and $\tilde{w}$:

$$\rho\,\text{var}\,\tilde{l} - \text{var}\,(\lambda\tilde{w}) = (1-\rho)l^2 - l(2\lambda R - \rho). \tag{5.10}$$

For a system $G/D/1$ we have $2\lambda R = \rho$, and therefore (5.6 *a*) and (5.9) become respectively (5.11) and (5.12) :

$$E[(\lambda\tilde{W})^2] = \rho E(\tilde{L}^2), \qquad (G/D/1) \tag{5.11}$$

$$E[(\lambda\tilde{w})^2] = \rho E(\tilde{l}^2), \qquad (G/D/1). \tag{5.12}$$

For holding times other than deterministic $2\lambda R > \rho$ and we have inequalities corresponding to (5.11) and to (5.12) :

$$E[(\lambda\tilde{W})^2] \geqslant \rho E(\tilde{L}^2), \qquad (G/G/1) \tag{5.11 a}$$

$$E[(\lambda\tilde{w})^2] \geqslant \rho E(\tilde{l}^2), \qquad (G/G/1). \tag{5.12 a}$$

Consider now the average remaining sojourn time for a customer, conditional upon the system $G/G/1$ being non-empty, i.e. upon $\tilde{L} > 0$. That is, consider the expression

$$\frac{1}{1-P_0}\sum_1^\infty \frac{T_k}{k} P_k = \frac{1}{\rho}\sum_1^\infty P_k\left[R + \frac{k-1}{2\mu}\right] = R + \frac{1}{2}\frac{l}{\rho}\frac{1}{\mu}, \tag{5.13}$$

a most plausible result since l/ρ is the average queue size, conditional upon the system being non-empty, i.e. upon $\tilde{L} \geqslant 1$, at a random survey instant.

SECTION 6. CONSERVATION OF THE FLOW OF TRANSITIONS

Consider the birth-and-death process whose transition diagram appears below.

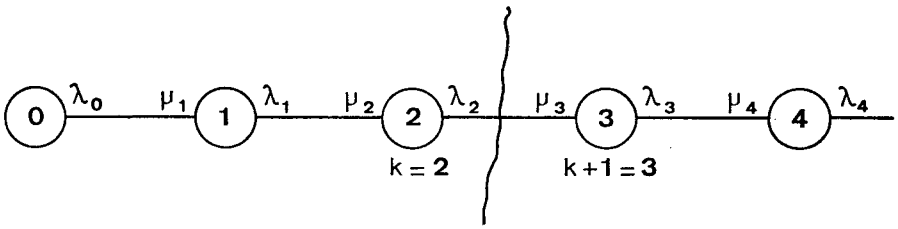

Figure 1

Notice the cut separating the « left » states from the « right » ones. It is clear, that under stationary conditions *the frequency of transitions across the cut from left to right must equal the frequency of transitions across this cut from right to left.* This statement is an instance of the *Conservation of the Transition Flow* or more briefly, of the *Conservation of Flow.* We can imagine a jumper moving from state to neighboring state, in the above diagram, in accordance with the history of the system. This jumper crosses a given cut in any of the two directions with equal frequency. This is a weak version of the property that across any cut, over any period of time, the left and right jumps are equal in number or differ by one.

In order to express the Conservation of Flow symbolically note that

(6.1) $\lambda_k dt$ = probability that the system will jump to S_{k+1} within dt conditional upon its being in S_k; $k \geqslant 0$

and

(6.2) $\mu_k dt$ = (conditional) probability that if the system is in state $k+1$ it will jump within dt to the state $k-1$; $k \geqslant 1$.

The conservation of flow states therefore that

$$P_k \cdot \lambda_k = P_{k+1} \cdot \mu_{k+1} \tag{6.3}$$

and

$$P_{k+1} = P_k \lambda_k / \mu_k \quad , \quad k \geqslant 0. \tag{6.4}$$

Our conservation rule leads directly to a recursion relation whereas the customary derivation produces a difference equation with three arguments.

It follows from (6.4) that

$$P_k = P_0 \prod_0^{n-1} \lambda_k \Big/ \prod_1^{n} \mu_k \tag{6.5}$$

and this along with $\sum_0^{\infty} P_k = 1$ yields all the state probabilities.

When $\lambda_k = \lambda$ and $\mu_k = \mu$ for all k, and using $\lambda/\mu = \rho$, (6.5) becomes

$$P_k = \rho^k \cdot P_0. \tag{6.5 a}$$

Since the transitions from left to right must be as frequent as transitions from right to left, we have for the birth-and-death process the *parity condition*

$$\sum_0^{\infty} \lambda_k P_k = \sum_1^{\infty} \mu_k P_k. \tag{6.6}$$

This gives norhing new informationally as it can also be obtained by summation from (6.3) but is convenient for the case of (6.5 *a*); (6.6) then becomes

$$\lambda = \mu(1 - P_0) \quad \text{and} \quad P_0 = 1 - \rho. \tag{6.6 a}$$

The statement (6.6 *a*) expresses also the conservation of customers. Thus there is a close relation between the geometry of the diagram and the mechanics of the queuing operation.

The above described procedure, the method of cuts, can be stated more generally for stochastic systems describable by means of « states », not cecessarily Markovian discrete systems.

Conservation of Transition Flow. Divide all states of the system into two non-overlapping groups *A* and *B*. Then, under stationary conditions, the frequency of transitions from group *A* into group *B* equals the frequency of transitions from *B* to *A*.

When, in a queuing system, *A* consists of a single state (and *B* of the remaining ones) the above conservation rule yields the usual balance equations. Thus, in place of the recursion relation (6.3) we would have

$$(\lambda_k + \mu_k)P_k = \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1}. \tag{6.7}$$

Limiting one of the two groups of states to a single point (in more general systems it may be a neighborhood in the phase space) may be referred to as the local conservation of flow. Informationally the general and local forms are equivalent but the general statement often allows a more skillful selection of the cuts and of the corresponding difference equations. The proof of the equivalence, in its basic outline, is quite simple. It starts with the observation that

the flow out of $(a \cup b)$ and the flow into it are equal, a and b being two groups of states, in particular non-overlapping groups. We will not pursue this topic here.

EXAMPLE 6.1. Customers arrive with frequency λ in a Poissonian manner but are served only in groups of size n. There is a single channel and the service time is exponentially distributed. Define the (super)state $\hat{S}_k$ as the situation where there are $mn + k$, $0 \leqslant k < n$, customers in the system. Since departures don't affect the state $\hat{S}_k$ of the system it is clear that the transitions and their frequencies are represented by the following diagram, where $n = 4$.

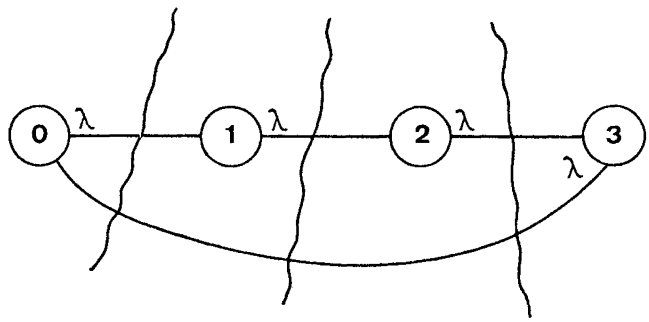

Figure 2

With A_k denoting the probability of the (super)state $\hat{S}_k$ the three cuts give the equations

$$\lambda A_0 = \lambda A_3; \quad \lambda A_1 = \lambda A_3; \quad \lambda A_2 = \lambda A_3 \tag{6.8}$$

and therefore

$$A_0 = A_1 = A_2 = A_3 = \tfrac{1}{4}. \tag{6.9}$$

More generally

$$A_k = 1/n \quad , \quad 0 \leqslant k \leqslant n - 1. \tag{6.10}$$

Note that $A_k = \sum_{i=0}^{\infty} P_{ni+k}$ and that therefore

$$\begin{aligned} &P_0 + P_n + P_{2n} + P_{3n} + + = 1/n \\ &P_1 + P_{n+1} + P_{2n+1} + P_{3n+1} + + = 1/n \\ &\text{etc.} \end{aligned} \tag{6.11}$$

EXAMPLE 6.2. Consider the Markovian transition diagram below. A queuing scenario can be made up, if desired, without difficulty.

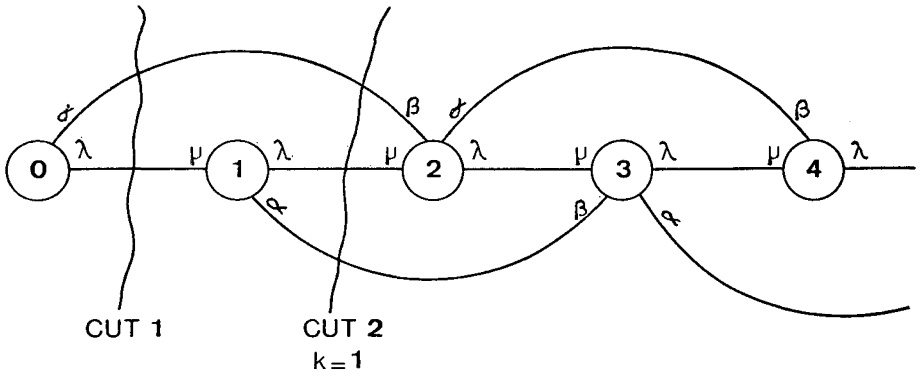

Figure 3

The balance equations are, in addition to $\sum_0^\infty P_k = 1$,

(6.11) $$(\lambda + \alpha)P_0 = \mu P_1 + \beta P_2, \text{ for cut } \# 1$$

(6.12) $$(\lambda + \alpha)P_1 + \alpha P_0 = (\mu + \beta)P_2 + \beta P_3, \text{ for cut } \# 2$$

(6.13) $$(\lambda + \alpha)P_k + \alpha P_{k-1} = (\mu + \beta)P_{k+1} + \beta P_{k+2}.$$

(6.12) is a spécial case of (6.13), namely when $k = 1$.

The recursion base of the above system of equations is P_0 and P_1. This recursion base can be reduced to P_0 by adding the parity equation (left-to-right transitions are as frequent as right-to-left ones), or, equivalently, the equation of customer conservation in a queue scenario. This equation is

(6.14) $$(\lambda + \alpha) = (\mu + \beta)\sum_2^\infty P_k + \mu P_1 = (\mu + \beta)(1 - P_0 - P_1) + \mu P_1.$$

EXAMPLE 6.3. A system $M/M/1$ is modified by a reneging propensity α on the part of the customers in the queue (but not in service). Thus a customer in the queue will reneg with probability $\alpha\, dt$ during dt. The transition diagram is shown below.

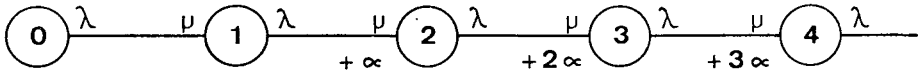

Figure 4

The balance of transitions across a cut separating the states S_k and S_{k+1} is

(6.15) $$P_k\lambda = P_{k+1}(\mu + k\alpha) \quad , \quad k \geqslant 0,$$

and therefore the recursion relation is

(6.16) $$P_{k+1} = P_k\lambda/(\mu + k\alpha).$$

Along with $\Sigma P_k = 1$, (6.16) allows to determine numerically all the state probabilities.

The rule of conservation of customers results in

$$\lambda = l\alpha + (1 - P_0)\mu, \tag{6.17}$$

relating the expected queue size l to P_0.

The generalization to $M/M/c$ is an easy exercise (cf. section 3).

BIBLIOGRAPHY

[1] Cox D. R. and Smith W. L., *Queues*, John Wiley & Sons, 1961.

[2] Feller William, *An Introduction to Probability Theory and Its Applications*, vol. 1, 3-rd edition, John Wiley & Sons, 1968 (or any prior edition).

[3] Krakowski Martin, *Arrival and Departure Processes in Queuing Systems and the Pollaczek-Khinchin Formula*, To appear in Revue Française d'Automatique, Informatique, et Recherche Opérationnelle.

[4] Little John, D. C., *A Proof for the Queuing Formula : $L = \lambda W$*, Oper. Res., vol. 9, N° 3, 383-387, 1961.

[5] Morse Philip M., *Queues, Inventories and Maintenance : The Analysis of Operations Systems with Variable Demand and Supply*, John Wiley & Sons, 1958.

[6] Prabhu N. U., *Queues and Inventories : A Study of the Basic Stochastic Processes*, John Wiley & Sons, 1965.

[7] Saaty Thomas L., *Elements of Queuing Theory with Applications*, McGraw-Hill, 1961.