A QUEUE WITH STARTER AND A QUEUE WITH VACATIONS: DELAY ANALYSIS BY DECOMPOSITION

HANOCH LEVY
AT&T Bell Laboratories, Holmdel, New Jersey

LEONARD KLEINROCK
University of California, Los Angeles, California

(Received April 1983; revisions received January 1984, November 1984; accepted July 1985)

This paper analyzes both a queueing system that incurs a start-up delay whenever an idle period ends and one in which the server takes vacation periods. We show that the delay distribution in the queue with starter is composed of the direct sum of two independent variables: 1) the delay in the equivalent queue without starter, and 2) the additional delay suffered due to the starter's presence. Using this decomposition property, we easily derive the distribution of the delay suffered in the system with starter. This analysis is done for systems (both discrete and continuous time) whose interarrival times possess the memoryless property. Using this approach, we then analyze the $M/G/1$ system with vacation periods. First, we show that the $M/G/1$ with vacations can be considered as a special case of the $M/G/1$ with starter, so that the delay in the $M/G/1$ with vacations can be easily found by using the formula for the delay of the $M/G/1$ with starter. Second, using geometric arguments, we explain why the additional delay in the vacation system is distributed as the residual life of the vacation period.

We consider a first-come-first-served queueing system with a "starter." In such a system, the server is "turned off" whenever it becomes idle. When a customer arrives at an idle system, he cannot be served immediately; rather the system requires an additional (random) amount of time to start from "cold" before it can serve the new "first" customer. Customers who arrive to a "hot" system (i.e., one with at least one customer either in service or in the queue) will join the queue and be served in turn as in a simple queueing system.

The model for a queueing system with special considerations when the server becomes idle is not new. Miller (1964) analyzed a system whose server goes on a vacation (a "rest period") of random length whenever it becomes idle. Miller also considered a system whose server behaves normally but in which the first customer arriving at an empty system has a special service time. Scholl (1976) and Scholl and Kleinrock (1983) analyzed the "server with rest periods" using another approach. Scholl considered as a special case for rest periods, a queueing system with a starter (or "a system with initial set-up time"). Both papers analyze $M/G/1$ queues. These types of systems and similar ones were also reported by Avi-Itzhak, Maxwell and Miller (1965), Cooper (1970), Heyman (1977), Lemoine (1975), Levy and Yechiali (1975), Pakes (1973), Shanthikumar (1980), Van Der Duyn Schouten (1978), and Welsh (1964).

The need for studying *a queue with starter* for slotted (i.e., discrete time) systems, and the fact that previous studies analyzed only $M/G/1$ (i.e., continuous time) systems, motivated us to again study the queue with starter. The emphasis in this paper is on developing a novel approach to studying this system. This approach will compare the delay suffered by a customer in a usual queueing system with the delay in a system with starter. Instead of deriving the delay in the queue with starter directly, we find the *additional delay* suffered due to the presence of the starter. Moreover, we show that the *additional delay* in the system with starter is independent of the delay in the system without starter. Using this independence property, it is then easy to calculate the total delay in the system with starter: *it is simply the direct sum of the delay in the queue without starter plus the additional delay calculated above.*

This approach is very powerful in analyzing systems similar to the queue with starter. Levy (1984), using the same approach, analyzes a queueing system with starter where the *length* of the start-up period *depends* on the arrival process (unlike the system analyzed here, where the start-up time is independent of the arrival process). Levy also uses the results reported in

Subject classification: 688 busy period analysis.

0030-364X/86/3403-0426 $01.25

this paper to derive the delay in an exhaustive ALOHA system. The fact that the delay in the queue with starter can be calculated as the (independent) sum of two independent random variables makes the analysis of these systems relatively simple.

As stated above, in contrast to previous studies that analyzed $M/G/1$ systems, the emphasis in this paper is on studying *slotted* systems with memoryless arrival streams. In Section 2, we analyze the delay in a slotted queue with starter. In this analysis we derive the z-transform of the delay in this system. For the sake of completeness, we use our approach to rederive the delay in an $M/G/1$ queue with starter and find agreement with Scholl's results. in Section 3, we study a system with vacation periods. First, we show that a system with vacation periods can be considered as a special case of the queue with starter. The delay in this system can thus be easily found from the delay of the queue with starter. We then show that the delay of an $M/G/1$ with vacation periods *is exactly* the sum of two independent random variables:

- the delay in an $M/G/1$ without vacation periods;
- an additional delay distributed as the residual life of the vacation period.

Lastly, we mention that some of this work (as first reported in Levy and Kleinrock 1983) has been developed, in parallel, in two independent studies. In the first, Fuhrmann (1983, 1984) showed that the delay in the queue with vacation periods consists of the sum of two independent random variables:

- the delay in an $M/G/1$ without vacation periods;
- an additional delay distributed as the residual life of the vacation period.

Fuhrmann's (1983, 1984) result is identical to ours in Subsection 3.2. Nevertheless, his method of proving this property is rather different from ours. In the second, parallel paper, Doshi (1983) addresses the decomposition property in both the queue with starter and the queue with vacations. The model he uses is a continuous time model of a $GI/G/1$ queue. His emphasis is on studying the queue with vacation periods, while the queue with starter is considered as a special case of the queue with vacation periods. In addition, Gelenbe and Iasnogorodski (1980) have established the decomposition property for the $GI/G/1$ system with vacation periods.

1. Notation, Definitions and System Description

In this paper, we analyze our queueing system by means of the unfinished work in the system. We define:

$U(t) \triangleq$ unfinished work in the system at time t;
$\triangleq$ remaining time required to empty the system of all customers present at time t.

We use the usual notation:

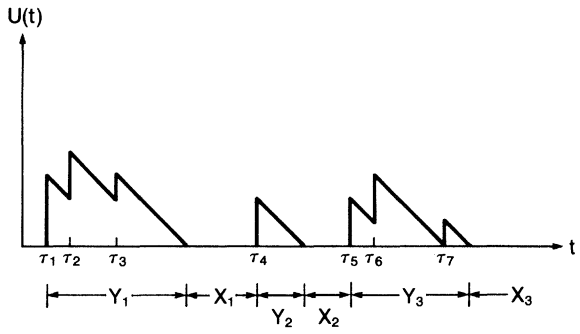
$C_n \triangleq$ the nth customer.

$\tau_n \triangleq$ arrival time of C_n.

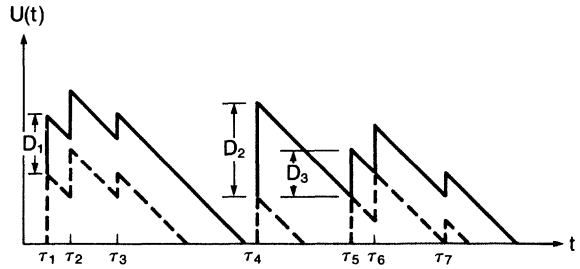
$t_n \triangleq \tau_n - \tau_{n-1}$
$=$ interarrival time between C_{n-1} and C_n.

$x_n =$ service time of C_n.

In Figure 1a we plot the behavior of $U(t)$ versus t in a simple queueing system. This system will be called *system-A*. As described in Kleinrock (1975), $U(t)$ can be viewed as the virtual waiting time, i.e., if the service policy is first-come-first-served, the waiting time of customer C_i is $U(\tau_i)$ (all the work residing in queue when C_i arrives). We also use the terms "busy period" and "idle period" to represent durations in which the server is continuously busy or idle (respectively). We denote the busy period durations by $Y_1, Y_2, Y_3, \ldots$ and the idle period durations by $X_1, X_2, X_3, \ldots$. Note that C_1, C_4 and C_5 initiate busy periods.



(a) System-A, a system without starter



(b) System-B, a system with starter

Figure 1. The unfinished work in the system (with and without starter).

We now switch to the queue-with-starter system and call it *system-B*. In order to analyze the queue-with-starter system, we construct system-B from system-A for each sample path by using the same arrival times and service times, and by adding the start-up delays (note that, logically, one could view this approach as constructing system-A from system-B by removing the start-up delays). This construction makes the sets of arrival instants ($\{\tau_i\}$) and service times ($\{x_i\}$) identical in both systems. In Figure 1b we plot $U(t)$ versus t in system-B. In this figure, the dashed line represents system-A and the solid line, system-B. The difference (denoted by D) represents the *additional delay* suffered in system-B.

In Figure 1b we note that customer C_1 arrives to an empty system and thus suffers an additional delay D_1 due to an independently selected *cold start*. Note that C_2 and C_3 suffer exactly the same additional delay. When C_4 arrives, he finds the system idle, and suffers the additional delay of a second independently selected cold start (D_2), which is not necessarily identical in length to D_1. However, we observe another behavior when C_5 arrives. Since $D_2 > X_2$, C_5 finds the system busy, and a cold start is not required. Nevertheless, C_5 is still subjected to an additional delay, $D_2 - X_2$. Again, we note that C_6 and C_7 suffer the same additional delay as C_5.

Keeping this in mind, we now turn to the analysis of system-B.

2. The Analysis of System-B, a Queue with Starter

As mentioned previously, this analysis will compare the behavior of systems A and B under the same arrival pattern.

In addition to the notation presented in Section 1, we define:

S_i = length of a cold start (if any) corresponding to the ith busy period.

D_i = actual additional delay suffered by the first customer of the ith busy period.

For convenience of notation, S_i is defined for *every* i. For a busy period i that suffers a cold start, S_i represents the length of the cold start. For other busy periods, S_i is a dummy variable that is not used in the analysis.

The reader should note that, even though we deal with system-B, we still consider busy and idle periods according to their appearance in system-A. This additional notation relates to busy (idle) periods as viewed in system-A, e.g., the ith busy period is the ith busy period in system-A.

2.1. The Basic Properties of the System

The following assumptions are required for the general analysis.

1. The length of an interarrival time t_i is independent of the length of any other interarrival time $t_j (i \neq j)$. The service time x_i of an arbitrary customer is independent of the service time x_j of any other customer ($i \neq j$). Service times are independent of interarrival times, so x_i is independent of t_j for all i and j.
2. The length S_i of a cold start is independent of the length S_j of any other cold start (for any $i \neq j$).
3. The length S_i of a cold start is independent of both the sequence $\{t_n\}$ and the sequence $\{x_n\}$.

The first assumption is very common for most queueing systems. The second and third assumptions simply state that the length of a cold start is chosen independently of system-A and of the length of other cold starts.

Next, we show how to calculate the additional delay suffered by the *first customer* of busy period i. The additional delay suffered by this *first customer* can be calculated recursively from the following equation:

$$D_1 = S_1 \tag{1a}$$

$$D_{i+1} = \begin{cases} D_i - X_i & \text{if } D_i \geq X_i \\ S_{i+1} & \text{if } D_i < X_i. \end{cases} \tag{1b}$$

The basis of the recursion D_1 is clearly the first cold start of the system. The first line in the recursion (1b) represents the case in which the first customer of busy period i (from system-A) finds system-B busy, while the second line represents the case in which this customer finds system-B idle, and his additional delay is due to an independent cold start. In the following subsections, we will use this recursion to calculate the limiting distribution of D_i.

While the additional delay suffered by a "first customer" is an important measure, our main interest is the additional delay suffered by an arbitrary customer. In the following development, we show that the distributions of these two delays are *identical*.

Theorem 1. *If customers C_i and C_j belong to the same busy period in system-A, they suffer exactly the same additional delay in system-B.*

Proof. Let $U_A(t)$ be the unfinished work in system-A and $U_B(t)$ be the unfinished work in system-B. Without loss of generality, let us assume that $j > i$. If $j = i + 1$ (C_j is the next customer arriving after C_i),

then we have

$$U_A(\tau_j) = U_A(\tau_i) - (\tau_j - \tau_i) + x_j,$$

$$U_B(\tau_j) = U_B(\tau_i) - (\tau_j - \tau_i) + x_j,$$

from which the claim follows. By a simple induction, the claim can be proved for arbitrary $j \neq i + 1$.

Theorem 2. *D_i is independent of X_i for every i.*

Proof. It is clear that D_i is a function only of $X_1, X_2, \ldots X_{i-1}$ and of $S_1, S_2, \ldots S_i$. Due to the memoryless property of the arrival process, X_i is independent of all these variables and thus it is also independent of D_i.

The following theorem states that the additional delay a customer suffers in the system with starter is actually independent of the delay he suffers in the system without starter.

Theorem 3. *Given that a customer is served in busy period j, the additional delay suffered by this customer in system-B is statistically independent of the delay he would suffer in the equivalent system-A.*

Proof. Let C_i be an arbitrary customer served in busy period j and let C_k ($k \leq i$) be the first customer served in this busy period. From Theorem 1, the additional delay suffered in system-B by C_i and C_k is the same. Thus, we must show that the additional delay suffered by C_k in system-B is independent of the delay C_i suffers in system-A. It is clear that the delay suffered by C_i is a function of only the interarrival times and the service times that "belong" to busy period j, namely, the series $t_{k+1}, t_{k+2}, \ldots t_i$ and the series $x_k, x_{k+1}, \ldots x_i$. On the other hand, the additional delay suffered by C_k is a function of only the system behavior prior to τ_k (the starting time of busy period j). Specifically, this delay is a function only of the sequence $t_2, t_3, \ldots t_k$, the sequence $x_1, x_2, \ldots x_{k-1}$ and the sequence $S_1, S_2, \ldots S_j$. Now, because the group of variables on which the delay (in system-A) depends and the group of variables on which the additional delay depends are mutually exclusive, and because of assumptions 1 and 3, these groups are statistically independent of each other. Thus, the additional delay suffered in system-B is independent of the delay suffered in system-A.

This theorem directly implies (see Doshi for details) that in *equilibrium* the additional delay suffered by an *arbitrary* customer in system-B is independent of the delay he would suffer in system-A.

This result now allows us to study in three steps the total delay suffered in the system with starter: 1) Derive the delay suffered in the system without starter. 2) Derive the additional delay suffered in the system with starter. 3) Convolve the distributions of the two delays to yield the total delay in the system with starter.

The next theorem states that the additional delay suffered in system-B by the customers of a given busy period (according to system-A) is independent of the number of customers served in this busy period.

Theorem 4. *D_i is independent of the number of customers served in busy period i.*

We omit the proof, due to its similarity to the proof of Theorem 3.

The following corollary is a direct result of Theorems 1 and 4.

Corollary 5. *The limiting distribution of the additional delay suffered by an **arbitrary** customer in system-B is identical to the limiting distribution of D_i.*

2.2 Discrete System with General Memoryless Arrivals

This section considers a discrete time model in which time is indexed by fixed length slots. The arrival process can be described as a renewal process, which means that the number of arrivals in slot i is independent of the number of arrivals in slot j for any $i \neq j$. Arrivals in slot i are not considered to be "in" the system until the *end* of slot i. The number of arrivals in a given slot is taken from a general distribution, and the service time is general.

In this section we are interested in the limiting behavior of D_i as i approaches infinity. We recall that time is measured in units of the fixed slot length and define:

$$d_i^j \triangleq \Pr[D_j = i], \quad D_j(z) \triangleq \sum_{i=0}^{\infty} d_i^j z^i, \quad d_i \triangleq \lim_{j\to\infty} d_i^j,$$

$$D(z) \triangleq \sum_{i=0}^{\infty} d_i z^i, \quad \bar{D} \triangleq \sum_{i=0}^{\infty} i d_i,$$

$$x_i^j \triangleq \Pr[X_j = i], \quad X_j(z) \triangleq \sum_{i=0}^{\infty} x_i^j z^i, \quad x_i \triangleq \lim_{j\to\infty} x_i^j,$$

$$X(z) \triangleq \sum_{i=0}^{\infty} x_i z^i, \quad \bar{X} \triangleq \sum_{i=0}^{\infty} i x_i,$$

$$s_i^j \triangleq \Pr[S_j = i], \quad S_j(z) \triangleq \sum_{i=0}^{\infty} s_i^j z^i, \quad s_i = \lim_{j\to\infty} s_i^j,$$

$$S(z) \triangleq \sum_{i=0}^{\infty} s_i z^i, \quad \bar{S} \triangleq \sum_{i=0}^{\infty} i s_i,$$

$a_i \triangleq \Pr[i \text{ arrivals in a given slot}]$.

In addition, $S^{(1)}(z)$ and $S^{(2)}(z)$, will denote, respectively, the first and second derivatives of $S(z)$, and $D^{(1)}(z)$ will denote the first derivative of $D(z)$.

With these assumptions, it is clear that the random variables X_j, representing the lengths of the idle periods, are independent and identically distributed. Thus, the limiting distribution of X_j is identical to the distribution of X_j. Since the number of arrivals in any slot is independent from slot to slot, X_j is geometrically distributed (shifted by a slot) with parameter a_0 (the probability of no arrival). For the sake of simplicity, let us use $x = a_0$; thus X_j is distributed as follows:

$$x_i = x_i^j = \Pr[X_j = i] = (1 - x) \cdot x^{i-1} \qquad i = 1, 2, 3 \ldots . \quad (2)$$

Similar arguments show that the limiting distribution of the length of a cold-start is identical to the distribution itself, so $s_i = s_i^j$.

In the following, we solve for $D(z)$. From Equations 1a and 1b we obtain:

$$d_i^{j+1} = \Pr[D_j - X_j = i \mid D_j \geq X_j] \cdot \Pr[D_j \geq X_j] + \Pr[S_{j+1} = i \mid D_j < X_j] \cdot \Pr[D_j < X_j] \qquad i = 0, 1, 2, \ldots . \quad (3)$$

Using the independence property between D_j and X_j (Theorem 2) and the independence between S_{j+1} and both D_j and X_j, and using the fact that $x_i^j = x_i$ and $s_i^j = s_i$, we compute d_i^{j+1}:

$$d_i^{j+1} = \sum_{k=1}^{\infty} x_k \, d_{k+i}^j + s_i \cdot \sum_{k=0}^{\infty} d_k^j \sum_{l=k+1}^{\infty} x_l \qquad i = 0, 1, 2, \ldots . \quad (4)$$

From (4), we compute the z-transform of D:

$$D_{j+1}(z) = \sum_{i=0}^{\infty} d_i^{j+1} z^i = \sum_{i=0}^{\infty} z^i \left[\sum_{k=1}^{\infty} x_k \, d_{k+i}^j + s_i \sum_{k=0}^{\infty} d_k^j \sum_{l=k+1}^{\infty} x_l \right]. \quad (5)$$

Substituting (2) into (5), and further manipulation, yields

$$D_{j+1}(z) = \frac{(1 - x)[D_j(x) - D_j(z)]}{x - z} + S(z) D_j(x). \quad (6)$$

Computing $D(z)$ by taking limits on j gives

$$D(z) = D(x) \left[\frac{1 - x + S(z)(x - z)}{1 - z} \right]. \quad (7)$$

At $z = 1$, we note that $D(1) = 1$, $S(1) = 1$, $S^{(1)}(1) = \bar{S}$; using these results and L'Hôpital's rule, we obtain

$$D(x) = \frac{1}{1 + (1 - x)\bar{S}} . \quad (8)$$

Substituting (8) into (7) gives us the important result:

$$D(z) = \frac{1}{1 + (1 - x)\bar{S}} \left[\frac{1 - x + S(z)(x - z)}{1 - z} \right]. \quad (9)$$

Equation 9 relates the z-transform of the additional delay of an arbitrary customer to the probability of no arrival (x), the z-transform of a cold start ($S(z)$) and the expected length of a cold start ($\bar{S}$). To calculate the z-transform of the *actual* delay suffered in the queue with starter, one must calculate the z-transform of the delay in the equivalent queue without starter, and multiply it by $D(z)$ (as given in (9)). This is true since the additional delay in the queue with starter is *independent* of the delay in the queue without starter (see Theorem 3).

Given Equation 9, it is now easy to compute the expected additional delay. From the relationship

$$\bar{D} = \left. \frac{dD(z)}{dz} \right|_{z=1}$$

and using L'Hôpital's rule, we obtain

$$\bar{D} = \frac{1}{1 + (1 - x)\bar{S}} \left[\frac{2\bar{S} - S^{(2)}(1)(x - 1)}{2} \right]. \quad (10)$$

Recalling that $S^{(2)}(1) = \overline{S^2} - \bar{S}$, we find that

$$\bar{D} = \frac{2\bar{S} + (\overline{S^2} - \bar{S})(1 - x)}{2 + 2\bar{S}(1 - x)} . \quad (11)$$

We note that the mean of the additional delay depends on the first and the second moments of the cold start and on the probability of at least one arrival $(1 - x)$ in a slot.

From Corollary 5, it is clear that (9) and (11) represent the additional delay and its expected value for an *arbitrary* customer in the system.

2.3. The Behavior of the Mean Additional Delay in the Discrete System

The purpose of this section is to examine the behavior of expression (11) for the expected additional delay suffered due to the existence of the start-up delays.

The behavior of (11) when arrivals are rare ($1 - x$ approaches 0) is $\bar{D} \approx \bar{S}$. In this situation the distance (in terms of time) between consecutive busy periods is very large, and almost every busy period suffers a cold start. Therefore, almost all customers will suffer a "cold start," so $\bar{D} \approx \bar{S}$ and $D(z) \approx S(z)$.

When arrivals are common ($1 - x \approx 1$), the length of idle periods is usually 1, and the expected value of

the additional delay is

$$\bar{D} \approx \frac{\bar{S} + \overline{S^2}}{2(1 + \bar{S})}.$$

This expression may be validated by calculating the expected value of the additional delay in a system when the length of every idle period is exactly one slot (see Levy for details).

From (11), we realize that $\bar{D}$ is monotonically increasing with $\bar{S}$ when $\overline{S^2}$ is held constant. Moreover, if instead we hold the squared coefficient of variation $(C_s^2 = (\overline{S^2} - (\bar{S})^2)/(\bar{S})^2)$ fixed, and let $\bar{S}$ approach infinity, $\bar{D}$ will also approach infinity.

While all the previous properties look intuitive, the following is very surprising: $\bar{D}$ is not necessarily smaller than $\bar{S}$, i.e., the mean of the additional delay seen by a customer may be larger than the expected length of a cold start. Take, for example, the following cold start distribution:

$$s_i = \begin{cases} 1 - \frac{1}{k} & i = 0 \\ \frac{1}{k} & i = k \\ 0 & \text{otherwise.} \end{cases}$$

So, $\bar{S} = 1$, $\overline{S^2} = k$. According to (11),

$$\bar{D} = \frac{2 + (k - 1)(1 - x)}{2 + 2(1 - x)}.$$

Clearly, if $k > 3$, then $\bar{D} > 1$; so $\bar{D} > \bar{S}$!

Once this property is noted, it is easily explained. The reason is that a short cold start affects only a few busy periods (in this extreme case, exactly one) and, therefore, only a few customers, while long cold starts affect many busy periods; therefore, many customers may see a large additional delay. Thus, if you average over all customers, the mean of the additional delay may exceed the average length of a cold start.

From this observation we realize that, even if we hold $\bar{S}$ fixed, $\bar{D}$ can approach infinity when the second moment of the cold start is large enough. This result is similar to the observation made about the mean delay suffered in an $M/G/1$ system (see, for example, Kleinrock); this delay increases linearly with the coefficient of variation of the service time, so the delay may be unbounded even if ρ is kept fixed and under unity. A similar well-known observation (again, see Kleinrock) is that the mean waiting time in the $M/G/1$ system may exceed the mean busy period duration.

We conclude that the additional delay may grow extremely large if either the expected value of the cold start or the second moment of the cold start is extremely large.

2.4 The Eigenfunctions of the Discrete System

In this section, we are interested in how the start-up delay distribution is transformed into the additional delay distribution. Mathematically, we may view Equation 9 as a transformation from $S(z)$ to $D(z)$ and express it as

$$D(z) = T(S(z)), \tag{12}$$

where T is the transformation expressed by (9).

We may now ask what the eigenfunction of this transformation is. The mathematical meaning of this eigenfunction is: find the solutions for the equation $S(z) = T(S(z))$. In other words, an eigenfunction of the system is an additional delay distribution $(D(z))$ that is identical to the cold start distribution $(S(z))$ causing it.

To solve for the eigenfunctions of our system, let us use (9) in (12), giving

$$S(z) = \frac{1}{1 + (1 - x)\bar{S}}\left[\frac{1 - x + S(z)(x - z)}{1 - z}\right]. \tag{13}$$

Solving (13) gives

$$S(z) = \frac{1/(1 + \bar{S})}{1 - (\bar{S}z/(1 + \bar{S}))}. \tag{14}$$

Inverting (14) yields

$$s_i = \frac{1}{1 + \bar{S}}\left(\frac{\bar{S}}{1 + \bar{S}}\right)^i \qquad i = 0, 1, 2, \ldots. \tag{15}$$

Yes!—the memoryless geometric distribution strikes again in queueing theory!

In conclusion, then, if the cold start is geometrically distributed, the distribution of the additional delay suffered by *all customers* is also geometrically distributed with the same parameter.

2.5. The *M/G/1* System with Bulk Arrivals—a Continuous Model

For the sake of completeness, we may essentially repeat the derivations made above for an $M/G/1$ system with bulk arrivals. The system is a first-come-first-served single-server system with exponential interarrival times (with parameter λ) and arbitrary service times. Now the interarrival times are continuous, whereas previously they were discrete. As in the discrete case, the arrivals themselves may consist of bulks of arbitrary size.

The basic notation is not changed: X_i, S_i, D_i have the same meaning as before, and Equation 1 still

holds. The probabilistic notation is the following:

$$D_i(t) \triangleq \Pr[D_i \leq t], \quad d_i(t) \triangleq \frac{dD_i(t)}{dt},$$

$$D_i^*(s) \triangleq \int_0^\infty e^{-st}\, d_i(t)\, dt,$$

$$X_i(t) \triangleq \Pr[X_i \leq t], \quad x_i(t) \triangleq \frac{dX_i(t)}{dt},$$

$$X_i^*(s) \triangleq \int_0^\infty e^{-st} x_i(t)\, dt,$$

$$S_i(t) \triangleq \Pr[S_i \leq t], \quad s_i(t) \triangleq \frac{dS_i(t)}{dt},$$

$$S_i^*(s) \triangleq \int_0^\infty e^{-st} s_i(t)\, dt.$$

The limits of $D_i^*(s)$, $X_i^*(s)$ and $S_i^*(s)$ are denoted by $D^*(s)$, $X^*(s)$ and $S^*(s)$.

Following the approach used above, it is easy to derive the Laplace–Stieltjes transform (LST) and the expected value of the additional delay (for a detailed derivation, see Levy):

$$D^*(s) = \frac{1}{1 + \lambda\overline{S}}\left[\frac{\lambda + S^*(s)(s - \lambda)}{s}\right] \tag{16}$$

$$\overline{D} = \frac{2\overline{S} + \lambda\overline{S^2}}{2 + 2\lambda\overline{S}}. \tag{17}$$

These expressions agree with Scholl's results, which he calculated using a different method.

The eigenfunction of the system is

$$S^*(s) = \frac{1}{s\overline{S} + 1}, \tag{18}$$

which is, as expected, the LST of the exponential distribution!

3. An M/G/1 with Vacation (Rest) Periods

Consider an $M/G/1$ system with unlimited storage. The arrival process is Poisson with arrival rate λ, and the service order is first-come-first-served. When the server becomes idle, it goes on a vacation of random length V. If, upon returning from a vacation, it finds any positive number of customers in the queue, it begins serving them as a regular $M/G/1$ system (until the next vacation). If, on the other hand, the server finds no customers in the queue, it takes another vacation. Vacations are identically distributed and independent of each other and of the arrival process or service times.

The $M/G/1$ system with vacation periods was first studied by Miller, who analyzed, in addition to other system properties, the delay in the system. This system and similar ones were reported and analyzed by Cooper, Gelenbe and Iasnogorodski, Heyman, Levy and Yechiali, Shanthikumar, and Van Der Duyn Schouten. Scholl, and Scholl and Kleinrock were the first to notice that the delay in an $M/G/1$ system with vacations has the same distribution as a random variable that is the sum of the following two independent, random variables:

- the time in system as if there were no vacation; plus
- an additional delay distributed as the residual life of the vacation period.

However, Scholl and Kleinrock emphasize that this is only an *observation* of the *expression* for the delay in the system with vacations. They were not able to show these properties directly (i.e., by analyzing the *system*).

In this section, we show, in a *direct* way, that the additional delay in a system with vacations is independent of the delay in a system without vacations, and that it is distributed as the residual life of the vacation distribution. First, using the queue with starter, we calculate the additional delay *directly* and find it to be as observed by Scholl. Second, we make a simple, direct queueing analysis of the additional delay in the system with vacations and show that it is distributed as the residual life of the vacation.

3.1. Solving a System-with-Vacations by a System-with-Starter

Consider a customer C_k who arrives to the system with vacations (which we refer to as system-B) and who finds this system empty. Let system-A be the equivalent system without vacations and let j be the busy period (according to system-A) in which C_k is served. Upon arriving to system-B, C_k must wait until the server returns from vacation. Let us call this delay the *return time* and denote it by R_j. It is obvious that the return time (R_j) observed in the system with vacations plays a role similar to that played by the start-up delay in the system with starter. We now use this similarity to show that the system with vacations can be considered as a system with starter whose start-up times (S_j) are the return times (R_j). It is clear that, in contrast to the cold starts, the return times are *not independent* of all interarrival times. This is true since the return time depends on the arrival process (for example, the return time R_j depends on the arrival epoch τ_k and therefore on the interarrival time t_k). For this reason, every theorem from Section 2 must

be checked again to make sure that each still holds when cold starts are replaced by return times. Even though the return times are not independent of all interarrival times, the following still holds:

Theorem 6. *The return time R_j is independent of all future interarrival times $(t_{k+1}, t_{k+2}, \ldots)$ and all future service times $(x_k, x_{k+1}, \ldots)$.*

This theorem can be proved using arguments similar to those used in the proof of Theorem 3.

In addition to Theorem 6, we next show, for systems whose arrival process possesses the memoryless property, that the return time is also independent of the system history.

Theorem 7. *Let t_0 be the moment at which system-B becomes idle and the server begins taking vacations. Let j be the first busy period (according to system-A) starting after t_0. If the interarrival times possess the memoryless property, then the return time R_j is independent of any property of system-B as observed prior to t_0.*

Proof. It is clear that the return time R_j depends on the lengths of the vacation periods taken after t_0 and on the timing of the next arrival after t_0. Since the interarrival times posseses the memoryless property, the time from t_0 to the next arrival is independent of the system history (prior to t_0). Since vacation lengths are also independent of the system behavior, the return time is independent of the system behavior prior to t_0.

From Theorems 6 and 7 it is now easy to see that, for an $M/G/1$ system, Section 2's theorems (and analysis) still hold if the cold start times are replaced by the return times. Therefore, the system with vacations can be considered as a system with starter with the role of the cold starts played by the return times. For this reason, we now abandon the term "return time" and the notation R_j and denote them, as we did for the queue with starter, by "cold starts" and S_j, respectively.

This discussion suggests an approach for solving the $M/G/1$ system with vacation periods:

Corollary 8. *An $M/G/1$ system with vacation periods can be solved as follows.*

1. *Compute the Laplace–Stieltjes transform (LST) for the distribution of a cold start (as seen by arrivals) resulting from the vacation periods.*
2. *Use the expression for the LST of the cold start distribution computed in Step* 1, *and plug it into expression* (16) *for $S^*(s)$.*
3. *The additional delay computed by this expression is the additional delay in the system with vacation periods.*

To adopt this approach, we first must calculate the distribution of a cold start. Keeping our old notation, we now add the vacation variable:

V = the length of a vacation period;

$v(t)$ = the probability density function of V;

$V^*(s)$ = the LST of $v(t)$.

We recall that the length of a cold start is denoted by S and that of an idle period by X. Since the arrival process is Poisson with rate λ, $x(t) = \lambda e^{-\lambda t}$. Moreover, due to the memoryless property of the arrival process, any time interval that starts at an arbitrary point, t_0, and ends with the first arrival after t_0, is also exponentially distributed with parameter λ (like $x(t)$).

To calculate the length of a cold start, we begin counting from the moment the system becomes idle; let us call this moment t_0. At t_0 the server goes on vacation, and the time elapsing until the server returns is V. The first arrival after t_0 occurs X time units after t_0. If $X \leqslant V$, then the server, on returning from vacation, finds a customer in the system, and the additional delay that this customer will suffer is $V - X$. If, on the other hand, $X > V$, then the returning server will take another vacation. Again, due to the memoryless property of the arrival process, the first arrival will occur X time units after the end of the first vacation. Thus, if $X > V$, we can calculate the length of the cold start recursively, as before. The following recursion summarizes these observations:

$$\Pr[S \leqslant t] = \begin{cases} \Pr[V - X \leqslant t] & \text{if } V \geqslant X \\ \Pr[S \leqslant t] & \text{if } V < X. \end{cases} \tag{19}$$

From this recursion we now solve for $S^*(s)$. From (19), and since V, X are independent,

$$s(t) = \int_{u=t}^{\infty} \lambda e^{-\lambda(u-t)} v(u)\, du + s(t) \int_{u=0}^{\infty} v(u) \int_{w=u}^{\infty} \lambda e^{-\lambda w}\, dw\, du. \tag{20}$$

Taking Laplace transforms on both sides of (20), and with further manipulation, we obtain

$$S^*(s) = \frac{\lambda \cdot [V^*(s) - V^*(\lambda)]}{\lambda - s} + V^*(\lambda) S^*(s). \tag{21}$$

Solving for $S^*(s)$ gives

$$S^*(s) = \frac{\lambda [V^*(s) - V^*(\lambda)]}{(\lambda - s)[1 - V^*(\lambda)]}. \tag{22}$$

From (22), and using

$$V^*(0) = 1, \quad \left.\frac{dV^*(s)}{ds}\right|_{s=0} = -\bar{V},$$

we obtain

$$\bar{S} = \left.\frac{-dS^*(s)}{ds}\right|_{s=0} = \frac{1 - \bar{V}\lambda - V^*(\lambda)}{\lambda(V^*(\lambda) - 1)}. \tag{23}$$

Now that we know the LST and the first moment of the starter distribution, we can compute the LST of the additional delay from the analysis of the queue with starter. This computation is done by substituting (22) and (23) into (16):

$$D^*(s) = \frac{1 - V^*(s)}{\bar{V}s}. \tag{24}$$

Yes! this is the residual life of the vacation period! We have thus shown that the additional delay in an $M/G/1$ system with vacation periods is independent of the original delay and is distributed as the residual life of the vacation period.

3.2. Direct Explanation for the Delay of a Queue with Vacations

In Section 3.1, we showed that the delay in a queue with vacations actually *is* (and not only "could be thought of as") the sum of two independent random variables:

- the delay in a queue without vacations;
- an additional delay distributed as the residual life of the vacation period.

Yet we did not give a direct queueing explanation for the fact that the additional delay is distributed as the residual life of the vacation period. We do so in this section.

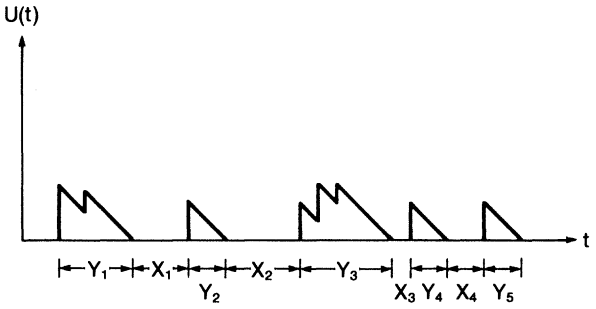
Consider the busy and idle periods in a regular $M/G/1$ system (denoted as system-A), as described in Figure 2a. We denote busy periods by $Y_1, Y_2, \ldots$ and idle periods by $X_1, X_2, \ldots$. Now let us impose vacations on this system (the new system is denoted as system-B). For "pedagogical" reasons, let us assume that the "vacation" is just another job the server must attend to. Thus, if we look from the *server's point of view*, we notice three properties:

1. The server consumes work at the rate of "one unit of work per unit of time."
2. At time points where a vacation V_i starts, additional work, equaling (in amount) the vacation length $|V_i|$, arrives to the system.
3. A new vacation starts if and only if the amount of work in the system is exactly zero. This means that the server takes a new vacation either when it finishes working in the $M/G/1$ system or when it returns from vacation and finds the $M/G/1$ system still empty of customers.

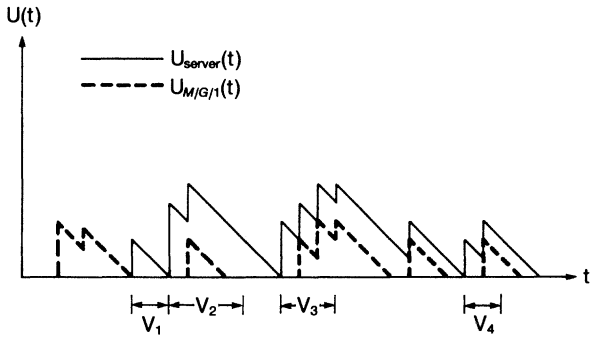
Figure 2b illustrates this situation. The solid line represents the total amount of work as seen by the server (denoted by $U_{\text{server}}(t)$), while the broken line represents the unfinished work in the $M/G/1$ system with no vacations (denoted by $U_{M/G/1}(t)$).

Next, we notice that the server system operates in a first-come-first-served (FCFS) fashion. This is true for the following reasons: 1) $M/G/1$ customers are served according to a FCFS policy. 2) "Vacation customers" arrive only when the system is empty. 3) The service discipline is non-preemptive for all types of customers. For this reason, the total time in system for an $M/G/1$ customer arriving at time t to system-B is exactly $U_{\text{server}}(t)$. Clearly, the time in system for the same customer in system-A is $U_{M/G/1}(t)$; thus, the additional delay suffered by this customer is given by $U_{\text{server}}(t) - U_{M/G/1}(t)$.

In Figure 3a, we plot the function difference $U_{\text{server}}(t) - U_{M/G/1}(t)$ (denoted by $D(t)$) versus t. From this figure

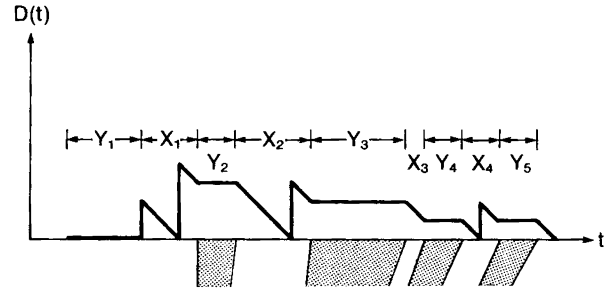


(a) The unfinished work in a regular $M/G/1$ system

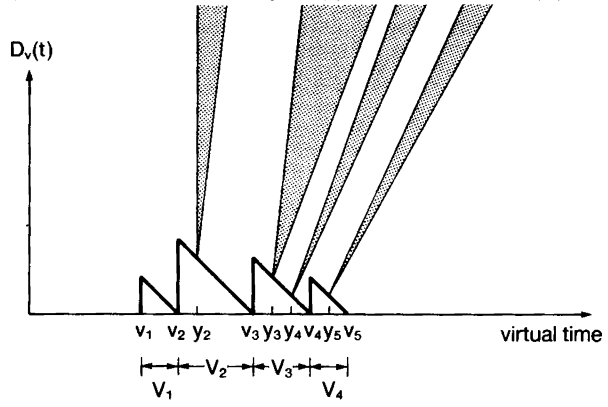


(b) Vacation periods "added" to a regular $M/G/1$ system

Figure 2. The unfinished work in a system with vacation periods.



(a) The additional delay: $D(t) = U_{server}(t) - U_{M/G/1}(t)$



(b) The additional delay versus virtual time

Figure 3. The additional delay in a system with vacations.

we can observe the following properties:

1. In time segments corresponding to idle periods in system-A (Figure 2a), $D(t)$ is consumed at the rate of "one work unit per time unit."
2. In time segments corresponding to busy periods in system-A, $D(t)$ remains constant.
3. The time epochs where $D(t)$ increases are those corresponding to the beginning of vacations. At such a moment, $D(t) = 0$ and discontinuously increases to the "height" of the vacation starting at that time.

In this figure, we observe that $D(t)$ is independent of any property of a system-A busy period (excluding its timing) since it stays constant during the duration of such periods. $D(t)$ is determined only by the *length of vacations* and the *length of system-A idle periods.* This observation explains why the additional delay in the queue with vacations (as in the queue with starter) is independent of the delay in the regular $M/G/1$ system. For this reason we can represent any system-A busy period by its starting point only. We do so by contracting to a point any flat segment of $D(t)$. This operation is done in Figure 3b, where the time axis becomes a virtual time axis and a segment Y_i from Figure 3a is contracted to a point y_i. The point corresponding to the beginning of a vacation, V_i, is denoted by v_i. For this figure, we define $D_v(t)$ as the (virtual) additional delay of virtual time t as seen in Figure 3b. In the transformation from 3a to 3b, we notice the following properties:

1. $D_v(y_i)$ equals the additional delay suffered by all customers of busy period (in Figure 2a) Y_i.
2. D_v continuously decreases at the rate of "one work unit per time unit." Whenever D_v becomes zero, it increases by a discontinuous increment.
3. The increments of D_v occur in epochs corresponding to vacation starts. The increment size is the vacation length.
4. Let t be an arbitrary time epoch on the virtual time axis and v_i be the epoch corresponding to the first vacation starting after t. From properties 2 and 3 and from the structure observed in Figure 3b, we may imply that $D_v(t) = v_i - t$.

From these arguments it becomes clear that, in order to find the additional delay suffered in system-A, one may compute D_v for the points $\{y_i\}$ in Figure 3b. This computation can be done as follows: We observe the time axis in Figure 3b and examine $D_v(y_i)$ for all y_i on this axis. We first note that the length of a subsegment (v_i, v_{i+1}) is distributed according to the distribution of the vacation length. Then, we notice that the intervals between the adjacent y-points represent lengths of idle periods; therefore, they are exponentially distributed, with parameter λ. Thus, in Figure 3b, the y-points behave like a stream of Poisson arrivals. Now, it is known that Poisson arrivals "see time averages" of continuous-time stochastic processes. Consequently, the fraction of arrivals observing a property of a given stochastic process is equal to the corresponding fraction of time this property is found on the time axis (see Wolff 1982). For this reason it is now clear that the distance from an arbitrary y_k point to the next v point is distributed as the residual life of the segments $\{(v_i, v_{i+1})\}$. Since these segments are distributed as the vacation length, $D_v(y_k)$ is distributed as the residual life of the vacation period.

We thus arrive at the promised conclusion: $D_v(y_i)$, the additional delay to customers in a system with vacation periods, is distributed as the residual life of a vacation period!

4. Summary

This paper studied queueing systems with starters and queueing systems with vacation periods. We showed that the delay distribution in the queue with starter is composed of the direct sum of two independent vari-

ables: 1) the delay in the equivalent queue without starter and 2) the additional delay suffered due to the presence of the starter. Using this decomposition property, we derived the Laplace transform of the additional delay, both for discrete systems with geometrically distributed interarrival times and for continuous systems with Poisson arrivals. Using the same approach, we then analyzed the $M/G/1$ system with vacation periods. We first showed that the $M/G/1$ system with vacations can be thought of as a special case of the $M/G/1$ system with starter, so that the delay in the $M/G/1$ system with vacations can be easily found by using the formula for the delay of the $M/G/1$ system with starter. Second, we explained, using geometric arguments, why the additional delay in the vacation system is distributed as the residual life of the vacation period.

Acknowledgment

This research was supported in part by the Defense Advanced Research Projects Agency of the Department of Defense, under contract MDA 903-82-C-0064.

This paper was written while Hanoch Levy was a Ph.D. student in Computer Science at the University of California, Los Angeles. The authors wish to thank B. Doshi, the referees and the Area Editor, Daniel Heyman, for their helpful comments.

References

AVI-ITZHAK, B., W. L. MAXWELL AND L. W. MILLER. 1965. Queueing with Alternating Priorities. *Opns. Res.* **13,** 306–318.

COOPER, R. B. 1970. Queues Served in Cyclic Order: Waiting Times. *Bell Syst. Tech. J.* **49,** 399–413.

DOSHI, B. T. 1983. Stochastic Decomposition in a $GI/G/1$ Queue with Vacations. Bell Laboratories, Holmdel, N.J. To appear in *J. Appl. Prob.*

FUHRMANN, S. W. 1983. A Note on the $M/G/1$ Queue with Server Vacations. Bell Laboratories, Holmdel, N.J.

FUHRMANN, S. W. 1984. A Note on the $M/G/1$ Queue with Server Vacations. *Opns. Res.* **32,** 1368–1373.

GELENBE, E., AND R. IASNOGORODSKI. 1980. A Queue with Server of Walking Type (Autonomous Service). *Ann. Inst. Henry Poincare* **16,** 63–73.

HEYMAN, D. P. 1977. The T-Policy for the $M/G/1$ Queue. *Mgmt. Sci.* **23,** 775–778.

KLEINROCK, L. 1975. *Queueing Systems, Vol I.: Theory.* Wiley-Interscience, New York.

LEMOINE, A. 1975. Limit Theorems for Generalized Single Server Queues: The Exceptional System. *SIAM J. Appl. Math.* **28,** 596–606.

LEVY, H. 1984. Non-Uniform Structures and Synchronization Patterns in Shared-Channel Communication Networks, CSD-840049, Computer Science Department, University of California, Los Angeles. Ph.D. dissertation.

LEVY, H., AND L. KLEINROCK. 1983. A Queue with Starter: Delay Analysis. ATS-83010. University of California at Los Angeles.

LEVY, Y., AND U. YECHIALI. 1975. Utilization of Idle Time in an $M/G/1$ Queueing System. *Mgmt. Sci.* **22,** 202–211.

MILLER, L. 1964. Alternating Priorities in Multi-Class Queues. Ph.D. dissertation, Cornell University, Ithaca, New York.

PAKES, A. G. 1973. On the Busy Period of the Modified $GI/G/1$ Queue. *J. Appl. Prob.* **10,** 192–197.

SCHOLL, M. 1976. Multiplexing Techniques for Data Transmission of Packet-Switched Radio Systems. UCLA-ENG-76123, University of California, Los Angeles, Computer Science Department. Ph.D. dissertation.

SCHOLL, M., AND L. KLEINROCK. 1983. On the $M/G/1$ Queue with Rest Periods and Certain Service-Independent Queueing Disciplines. *Opns. Res.* **31,** 705–719.

SHANTHIKUMAR, J. G. 1980. Some Analysis on the Control of Queues Using Level Crossing of Regenerative Processes. *J. Appl. Prob.* **17,** 814–821.

VAN DER DUYN SCHOUTEN, F. A. 1978. An $M/G/1$ Queueing Model with Vacation Times. *Z. Opns. Res.* Series A, **22,** 95–105.

WELCH, P. D. 1964. On a Generalized $M/G/1$ Queueing Process in which the First Customer of Each Busy Period Receives Exceptional Service. *Opns. Res.* **12,** 736–752.

WOLFF, R. W. 1982. Poisson Arrivals See Time Averages. *Opns. Res.* **30,** 223–231.