Donald L. Iglehart
Gerald S. Shedler

Simulation of Non-Markovian Systems

A generalized semi-Markov process provides a stochastic process model for a discrete-event simulation. This representation is particularly useful for non-Markovian systems where it is nontrivial to obtain recurrence properties of the underlying stochastic processes. We develop "geometric trials" arguments which can be used to obtain results on recurrence and regeneration in this setting. Such properties are needed to establish estimation procedures based on regenerative processes. Applications to modeling and simulation of ring and bus networks are discussed.

1. Introduction

It appears to be the rule rather than the exception that usefully detailed stochastic models for complex systems are such that it is extremely difficult or impossible to obtain an exact analytic solution. Simulation is essentially a controlled statistical sampling technique which can be used to study complex stochastic systems when analytic and/or numerical techniques do not suffice. We concentrate here on *discrete-event digital simulation* in which the behavior of a specified stochastic system is observed by sampling on a digital computer system and stochastic state transitions occur only at a set of increasing (random) epochs of time. In discrete-event simulations most of the stochastic processes that we encounter have piecewise-constant sample paths.

When simulating, we experiment with a stochastic system and observe its behavior. In the course of the simulation we measure certain quantities associated with the system, and using statistical techniques, draw inferences about characteristics of well defined random variables. The most obvious methodological advantage of simulation is that in principle it is applicable to stochastic systems of arbitrary complexity. It is, however, a decidedly nontrivial matter in practice to obtain from a simulation information which is both useful and accurate, and to obtain it at reasonable cost. The difficulties arise primarily from the inherent variability in a stochastic system, and it is necessary to seek theoretically sound and computationally efficient methods for carrying out the simulation. Apart from implementation considerations, important concerns for simulation relate to *generation methods* for sample paths of the stochastic system under study, the *design of simulation experiments*, and the *analysis of simulation output*. Since results of a simulation are based on observation of a stochastic system, it is absolutely essential that some assessment of the precision of results be provided.

Implicit in the implementation of any simulation is the definition of an appropriate "state" for the system. Heuristically, the system state maintains sufficient information about the system so that state transitions that occur over time completely determine the quantities of interest. This "state of the system at time t" constitutes a stochastic process in continuous or discrete time. When carrying out the simulation, we observe the behavior of this process as it evolves in time. In order to do so it is necessary to have a means of generating sample paths of this process and to have methods for obtaining valid estimates of the quantities of interest in the system.

In this paper we focus on simulation methods for non-Markovian systems in continuous time; i.e., systems whose state cannot be modeled as a Markov chain with countable state space. This is characteristic of local area computer network models (see, e.g., Loucks, Hamacher, and Preiss [1]) where it is important to incorporate system control algorithms explicitly into the simulation model. We restrict

attention to those discrete-event simulations whose underlying stochastic process can be represented as a generalized semi-Markov processs (GSMP) in the sense of Whitt [2].

In addition to providing a framework for generating sample paths of the underlying stochastic process of the simulation, the GSMP representation is particularly useful for simulation of non-Markovian systems in that it leads to methods for obtaining recurrence properties of the underlying stochastic process; cf. Glynn [3]. Such properties are needed to establish estimation procedures based on regenerative processes; cf. Fossett [4] and Iglehart and Shedler [5]. For specific non-Markovian systems (e.g., ring and bus network models) it can be difficult to determine conditions (distributional assumptions) under which the underlying stochastic process is regenerative. In this paper we develop "geometric trials" arguments (cf. Nummelin [6] and Tuominen and Tweedie [7]) which can be used to show the applicability of regenerative simulation methods.

2. Generalized semi-Markov processes

Heuristically, a GSMP (Matthes [8], König, Matthes, and Nawrotzki [9, 10]) moves from state to state in accordance with the occurrence of events associated with the occupied state. Each of the several possible events associated with a state compete to trigger the next transition, and each of these events has its own distribution for determining the next state. At each state transition of the GSMP, new events may be scheduled. For each of these new events, a clock indicating the time until the event is scheduled to occur is set according to an independent (stochastic) mechanism. If a scheduled event does not trigger a transition but is associated with the next state, its clock continues to run; if such an event is not associated with the next state, it is abandoned.

Following Whitt [2], formal definition of a GSMP is in terms of a general state space Markov chain (GSSMC) which describes the process at successive epochs of state transition. Let S be a finite or countable set of *states* and $E = \{e_1, e_2, \cdots, e_M\}$ be a finite set of *events*. For $s \in S$, $E(s)$ denotes the set of all events that can occur when the GSMP is in state s. When the process is in state s, the occurrence of an event $e \in E(s)$ triggers a transition to a state s'. We denote by $p(s'; s, e)$ the probability that the new state is s' given that event e triggers a transition in state s. For each $s \in S$ and $e \in E(s)$ we assume that $p(\cdot; s, e)$ is a probability mass function. The actual event $e \in E(s)$ which triggers a transition in state s depends on *clocks* associated with the events in $E(s)$ and the *speeds* at which these clocks run. Each such clock records the remaining time until the event triggers a state transition. We denote by r_{si} (≥ 0) the deterministic rate at which the clock c_i, associated with event e_i, runs in state s; for each $s \in S$, $r_{si} = 0$ if $e_i \notin E(s)$. We assume that $r_{si} > 0$ for some $e_i \in E(s)$. (Typically in applications all speeds r_{si} are equal to one. There are, however, models in which speeds other than unity as well as state-dependent speeds are convenient. For example, zero speeds are needed in queueing systems with service interruptions of the preemptive-resume type; cf. Shedler and Southard [11].)

For $s \in S$ define

$$C(s) = \{(c_1, \cdots, c_M): c_i \geq 0 \text{ and } c_i > 0 \text{ if and only if } e_i \in E(s); c_i r_{si}^{-1} \neq c_j r_{sj}^{-1} \text{ for } i \neq j \text{ with } c_i c_j r_{si} r_{sj} > 0\}. \quad (1)$$

The conditions in Eq. (1) ensure that no two events simultaneously trigger a transition (as defined below). The set $C(s)$ is the set of possible clock readings in state s. The clock c_i and event e_i are said to be *active* in state s if $e_i \in E(s)$. For $s \in S$ and $c \in C(s)$, let

$$t^* = t^*(s, c) = \min_{\{i: e_i \in E(s)\}} \{c_i r_{si}^{-1}\}, \quad (2)$$

where $c_i r_{si}^{-1}$ is taken to be $+\infty$ when $r_{si} = 0$. Also set

$$c_i^* = c_i^*(s, c) = c_i - t^*(s, c) r_{si}, \qquad e_i \in E(s) \quad (3)$$

and

$$i^* = i^*(s, c) = i \text{ such that } e_i \in E(s) \text{ and } c_i^*(s, c) = 0. \quad (4)$$

Beginning in state s with clock vector c, $t^*(s, c)$ is the time to the next state transition and $i^*(s, c)$ is the index of the unique triggering event $e^* = e^*(s, c) = e_{i^*(s,c)}$.

At a transition from state s to state s' triggered by event e^*, new clock times are generated for each $e' \in N(s'; s, e^*) = E(s') - (E(s) - \{e^*\})$. The distribution function of such a new clock time is denoted by $F(\cdot; s', e', s, e^*)$ and we assume that $F(0; s', e', s, e^*) = 0$. For $e' \in O(s'; s, e^*) = E(s') \cap (E(s) - \{e^*\})$, the old clock reading is kept after the transition. For $e' \in (E(s) - \{e^*\}) - E(s')$, event e' ceases to be scheduled after the transition.

Next consider a GSSMC $\{(S_n, C_n) : n \geq 0)\}$ having state space

$$\sum = \bigcup_{s \in S} (\{s\} \times C(s))$$

and representing the state (S_n) and vector (C_n) of clock readings at successive state transition epochs. (The ith coordinate of the vector C_n is denoted by $C_{n,i}$.) The transition kernel of the Markov chain $\{(S_n, C_n) : n \geq 0\}$ is

$$P((s, c), A) = p(s'; s, e^*) \prod_{e_i \in N(s')} F(a_i; s', e_i, s, e^*) \prod_{e_i \in O(s')} 1_{[0, a_i]}(c_i^*), \quad (5)$$

where $N(s') = N(s'; s, e^*)$, $O(s') = O(s'; s, e^*)$, and

$$A = \{s'\} \times \{(c_1', \cdots, c_M') \in C(s'): c_i' \leq a_i \text{ for } e_i \in E(s')\}.$$

The set A is the subset of $\sum$ which corresponds to the GSMP entering state s' with the reading c_i on the clock associated with event $e_i \in E(s')$ set to a value in $[0, a_i]$.

Finally, the GSMP is a piecewise constant continuous time process constructed from the GSSMC $\{(S_n, C_n) : n \geq 0\}$ in the following manner. Denote by ζ_n the time of the nth state transition, $n \geq 0$. [We assume that

$$P\{\sup_{n \geq 1} \zeta_n = +\infty \mid (S_0, C_0)\} = 1$$

for all initial states (S_0, C_0).] Then set

$$X(t) = S_{N(t)}, \tag{6}$$

where

$$N(t) = \max\{n > 0: \zeta_n \leq t\}.$$

The process $\{X(t) : t \geq 0\}$ is a GSMP.

The following examples illustrate the use of the GSMP model as a formal specification of a discrete-event simulation of a non-Markovian system.

- *Example 1*

Consider a unidirectional ring network having a fixed number of *ports*, labeled 1, 2, $\cdots$, N in the direction of signal propagation. At each port *message packets* arrive according to a random process and queue externally. A single *control token* circulates around the ring from one port to the next. The time for the token to propagate from port N to port 1 is a positive constant, R_N, and the time for the token to propagate from port $j-1$ to port j is a positive constant, R_{j-1}, $j = 2, 3, \cdots, N$. When a port observes the token and there is a packet queued for transmission, the port converts the token to a *connector* and transmits a packet followed by the token pattern; the token continues to propagate if there is no packet queued for transmission. By destroying the connector prefix the port removes the transmitted packet when it returns around the ring. Assume that the time for port j to transmit a packet is a positive random variable, L_j, with finite mean. Also assume that packets arrive at individual ports randomly and independently of each other; i.e., the time from end of transmission by port j until the arrival of the next packet for transmission by port j is a positive random variable, A_j, with finite mean. Note that there is at most one packet queued for transmission at any time at any particular port.

Set

$$X(t) = (Z_1(t), \cdots, Z_N(t), M(t), N(t)), \tag{7}$$

where

$$Z_j(t) = \begin{cases} 1 & \text{if there is a packet queued for transmission at port } j \text{ at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$M(t) = \begin{cases} j & \text{if port } j \text{ is transmitting a packet at time } t, \\ 0 & \text{if no port is transmitting a packet at time } t. \end{cases}$$

$N(t) = 1$ if at time t port N is transmitting a packet or the token is propagating to port 1, and $N(t) = j$ if at time t port $j-1$ is transmitting a packet or the token is propagating to port j, $j = 2, \cdots, N$.

The process $\{X(t) : t \geq 0\}$ defined by Eq. (7) is a GSMP with a finite state space, S, and event set, E. The events in the set E are "observation of token," "end of transmission," and "arrival of packet for transmission by port j," $j = 1, 2, \cdots, N$. For $s = (z_1, \cdots, z_N, m, n) \in S$, the event sets $E(s)$ are as follows. The events "end of transmission" $\in E(s)$ if and only if $m > 0$ and "observation of token" $\in E(s)$ if and only if $m = 0$. The event "arrival of packet for transmission by port j" $\in E(s)$ if and only if $z_j = 0$ and $m \neq j$.

If $s = (z_1, \cdots, z_N, m, m+1) \in S$ with $0 < m < N$, $s' = (z_1, \cdots, z_N, 0, m+1)$ [or if $s = (z_1, \cdots, z_N, N, 1) \in S$, $s' = (z_1, \cdots, z_N, 0, 1)$] and $e =$ "end of transmission," then the state transition probability $p(s'; s, e) = 1$. If $s = (z_1, \cdots, z_{n-1}, 1, z_{n+1}, \cdots, z_N, 0, n) \in S$ with $n < N$, $s' = (z_1, \cdots, z_{n-1}, 0, z_{n+1}, \cdots, z_N, n, n+1)$ [or if $s = (z_1, \cdots, z_{N-1}, 1, 0, N) \in S$ and $s' = (z_1, \cdots, z_{N-1}, 0, N, 1)$] and $e =$ "observation of token," then $p(s'; s, e) = 1$. If $s = (z_1, \cdots, z_{n-1}, 0, z_{n+1}, \cdots, z_N, 0, n) \in S$, $s' = (z_1, \cdots, z_{n-1}, 0, z_{n+1}, \cdots, z_N, 0, n+1)$, and $e =$ "observation of token," then $p(s'; s, e) = 1$. If $s = (z_1, \cdots, z_{j-1}, 0, z_{j+1}, \cdots, z_N, m, m+1) \in S$ with $m \neq j$ and $0 < m < N$, $s' = (z_1, \cdots, z_{j-1}, 1, z_{j+1}, \cdots, z_N, m, m+1)$ [or if $s = (z_1, \cdots, z_{j-1}, 0, z_{j+1}, \cdots, z_N, N, 1) \in S$ with $N \neq j$, $s' = (z_1, \cdots, z_{j-1}, 1, z_{j+1}, \cdots, z_N, N, 1)$], and $e =$ "arrival of packet for transmission by port j," then $p(s'; s, e) = 1$. All other state transition probabilities $p(s'; s, e)$ are equal to zero.

The distribution functions of new clock times for events $e' \in N(s'; s, e^*)$ are as follows. If $e' =$ "end of transmission" and $s' = (z_1, \cdots, z_N, m, n)$, then $F(x; s', e', s, e^*) = P\{L_m \leq x\}$ for all s and e^* such that $p(s'; s, e^*) > 0$. If $e' =$ "observation of token" and $s' = (z_1, \cdots, z_N, 0, n)$, then $F(x; s', e', s, e^*) = 1_{[R_{n-1}, \infty)}(x)$. If $e' =$ "arrival of packet for transmission by port j" and $s' = (z_1, \cdots, z_{j-1}, 0, z_{j+1}, \cdots, z_N, 0, j+1)$, then $F(x; s', e', s, e^*) = P\{A_j \leq x\}$.

- *Example 2*

Consider a ring network having a fixed number, K, of equal size *slots* and a fixed number of equally spaced *ports*, labeled

$1, 2, \cdots, N$ in the direction of signal propagation. At each port constant (slot size) length *message packets* arrive according to a random process and queue externally. The propagation delay from one port to the next is a positive constant, R. We assume that the number of ports N is a multiple of K and (so that there is no loss of utilization due to "unused bits") that the time to transmit a message packet is equal to NR/K. The lead "full/empty" bit maintains the status of each slot. Subject to the restriction that no port can hold more than one slot simultaneously, a port that has a packet queued for transmission and observes the status bit of an empty slot sets the bit to 1 ("full") and starts transmission (begins filling the slot). Transmission ends when the slot contains the entire packet. When the status bit of the filled slot propagates back to the sending port, the port resets the bit to 0 ("empty") and releases the slot. To ensure that all ports have an opportunity to transmit, a port which releases a slot passes the empty slot to the next port. Assume that packets arrive at individual ports randomly and independently of each other; i.e., the time from end of transmission by port j until the arrival of the next packet for transmission by port j is a positive random variable, A_j, with finite mean. Note that there is at most one packet queued for transmission at any time at any particular port.

Set

$$X(t) = (Z_1(t), \cdots, Z_N(t), M_1(t), \cdots, M_K(t), N_1(t), \cdots, N_K(t)), \quad (8)$$

where

$$Z_j(t) = \begin{cases} 1 & \text{if there is a packet queued for transmission at port } j \text{ at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \cdots, K$,

$$M_i(t) = \begin{cases} j & \text{if port } j \text{ holds slot } i \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

$N_i(t) = 1$ if at time t the status bit of slot i is propagating to port 1, and $N_i(t) = j$ if at time t the status bit of slot i is propagating to port $j, j = 2, 3, \cdots, N$. For any i $(1 \leq i \leq K)$ the vector $(Z_1(t), \cdots, Z_N(t), M_1(t), \cdots, M_K(t), N_i(t))$ contains the same information about the system as the vector $X(t)$. Incorporation of all the components $N_1(t), \cdots, N_K(t)$ into the state vector facilitates generation of the process.

The process $\{X(t) : t \geq 0\}$ defined by Eq. (8) is a GSMP with a finite state space S and event set E. The events in the set E are "observation of status bits by ports" and "arrival of packet for transmission by port j," $j = 1, 2, \cdots, N$. Let $s = (z_1, \cdots, z_N, m_1, \cdots, m_K, n_1, \cdots, n_K) \in S$. The event "observation of status bits by ports" $\in E(s)$ for all $s \in S$. The event "arrival of message for transmission by port j" $\in E(s)$ if and only if $z_j = 0$ and for each i either (1) $m_i \neq j$ or (2) $m_i = j$ and $n_i - 1 = j - 1 + \ell \pmod N$ for some integer ℓ such that $N/K < \ell < N$. Note that the ends of transmission coincide with the occurrence of particular "observation of status bits by ports" events. Suppose, for example, that there are $N = 4$ ports and $K = 2$ slots and that $s = (0, 0, 0, 0, 1, 0, 3, 1)$; i.e., port 1 is transmitting a packet in slot 1, slot 2 is empty, the status bit of port 1 is propagating to port 3, and the status bit of slot 2 is propagating to port 1. Then the occurrence of the event "observation of status bits by ports" in state s corresponds to an end of transmission by port 1.

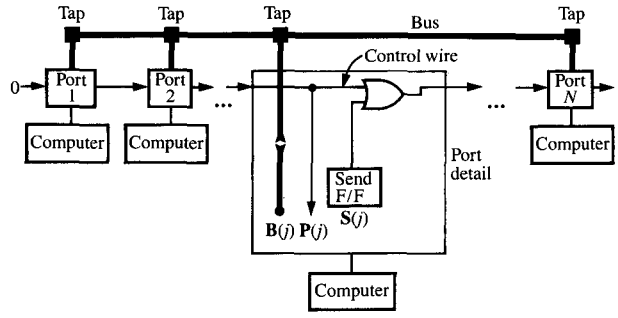

Figure 1 Bus network and ports.

• *Example 3*

Consider a collision-free bus network (cf. Eswaran, Hamacher, and Shedler [12]) with N *ports*, numbered 1, 2, $\cdots$, N from left to right; see Fig. 1. *Message packet* traffic on the passive bilateral bus is transmitted/received by port j at tap $\mathbf{B}(j)$. In addition to the bus, a one-way logic *control wire* also links the ports. Associated with each port j is a flip-flop, $\mathbf{S}(j)$, called the *send flip-flop*. The signal $\mathbf{P}(j)$, called the *OR-signal*, tapped at the control wire input to port j is the inclusive OR of the send flip-flops of all ports to the left of port j. Denote by T the end-to-end bus propagation delay. [For technical reasons, T actually must be the end-to-end propagation delay plus a small (fixed) quantity.] Denote the actual propagation delay along the bus between port i and port j by $T(i, j), i, j = 1, 2, \cdots, N$. Thus, $T(i, j) = T(j, i) < T$ for all i, j and $T(i, j) + T(j, k) = T(i, k)$ for all $i < j < k$. (We assume that $T(i, j) \neq T(k, j)$ for distinct i, k and all j.) Let $R(j)$ be the propagation delay (including gate delays) along the control wire from port j to port $N, j = 1, 2, \cdots, N$; thus, $R(1) \geq R(2) \geq \cdots \geq R(N) = 0$. Denote by $R(i, j)$ the propagation delay along the control wire from port i to port j. We assume that signal propagation along the control wire is slower than along the bus and that delays along shorter sections of each path scale proportionally; i.e., $R(1) > T$ and $R(i, j) > T(i, j)$ for all i, j.

Specification of distributed control scheme $A1$ is in terms of an algorithm for an individual port j. Packets (for trans-

mission by port j) which arrive while an execution of the algorithm by port j is in progress queue externally. Upon completion of this execution of the algorithm, one of any such packets immediately becomes available to port j for transmission and the next execution of the algorithm begins.

Algorithm A1

- Set $\mathbf{S}(j)$ to 1.
- Wait for a time interval $R(j) + T$.
- Wait until the bus is observed (by port j), to be idle AND $\mathbf{P}(j) = 0$; then start transmission of the packet, simultaneously resetting $\mathbf{S}(j)$ to 0.

For simplicity we assume that there can be at most one packet in queue at each port. Specifically, suppose that the time from end of transmission by port j until the arrival of a next packet for transmission by port j is a positive random variable, A_j, with finite mean. Also suppose that the time for port j to transmit a packet is a positive random variable, L_j, with finite mean.

Set

$$W(t) = (W_1(t), \cdots, W_N(t)), \tag{9}$$

where $W_j(t)$ equals 1 if at time t port j has set its flip-flop but has not yet completed the $R(j) + T$ wait, equals 2 if port j has completed the $R(j) + T$ wait but has not started transmission, equals 3 if port j is transmitting, and equals 4 otherwise. Next set

$$U(t) = (U_1(t), \cdots, U_N(t)), \tag{10}$$

where $U_j(t)$ equals 1 if port j observes the bus to be busy at time t and equals 0 otherwise. Also set

$$V(t) = (V_{2,1}(t), V_{3,1}(t), V_{3,2}(t), V_{4,1}(t), \cdots, V_{N,N-1}(t)), \tag{11}$$

where $V_{j,k}(t)$ equals 1 if port j has observed that port k has set its flip-flop and equals 0 otherwise. (Port j observes $\mathbf{P}(j) = 1$ at time t if and only if $V_{j,k}(t) = 1$ for some $k < j$.) Finally, set

$$X(t) = (W(t), U(t), V(t)).$$

Then the stochastic process $\{X(t) : t \geq 0\}$ is a GSMP with a finite state space S and event set E. The events in the set E are "end of transmission by port j," "end of wait for $R(j) + T$," "setting (to 1) of flip-flop by port j," "observation by port j of start of transmission," "observation by port j of end of transmission," "observation by port j of the setting (to 1) of flip-flop by port k to the left," and "observation by port j of the resetting (to 0) of a flip-flop by port k to the left," $j = 1, 2, \cdots, N$. For $s = (w_1, \cdots, w_N, u_1, \cdots, u_N, v_{2,1}, \cdots, v_{N,N-1}) \in S$ the event sets $E(s)$ are as follows. The event set $E(s)$ contains "end of transmission by port j" if and only if $w_j = 3$. The event "end of wait for $R(j) + T$" $\in E(s)$ if and only if $w_j = 1$. The event "setting of flip-flop by port j" $\in E(s)$ if and only if $w_j = 4$. The event "observation by port j of start of transmission" $\in E(s)$ if and only if $w_k = 3$ for some k and $u_j = 0$. The event "observation by port j of end of transmission" $\in E(s)$ if and only if $w_k = 3$ for some k and $u_i = 1$. The event "observation by port j of setting of flip-flop by port k to the left" $\in E(s)$ if and only if $w_k = 1$ for some $k < j$ and $v_{k,j} = 0$. The event "observation of resetting of flip-flop by port k to the left" $\in E(s)$ if and only if $w_k = 3$ for some $k < j$ and $v_{k,j} = 1$.

The distribution functions of new clock times for events $e' \in N(s'; s, e^*)$ are as follows. If e' = "end of transmission by port j" $\in E(s') - (E(s) - \{e^*\})$ and $p(s'; s, e^*) > 0$, the clock setting distribution function $F(x; s', e', s, e) = P\{L_j \leq x\}$. If e' = "end of wait for $R(j) + T$," the clock setting distribution function $F(x; s', e', s, e^*) = 1_{[R(j)+T,\infty)}(x)$. If e' = "setting of flip-flop by port j," the clock setting distribution function $F(x; s', e', s, e^*) = P\{A_j \leq x\}$. If e' = "observation by port j of start of transmission," the clock setting distribution function $F(x; s', e', s, e^*) = 1_{[T(k,j),\infty)}(x)$ if $w_k' = 3$. If e' = "observation by port j of end of transmission," the clock setting distribution function $F(x; s', e', s, e^*) = 1_{[T(k,j),\infty)}(x)$ if $w_k' = 4$. If e' = "observation by port j of setting of flip-flop by port k to the left," the clock setting distribution function $F(x; s', e', s, e^*) = 1_{[R(k,j),\infty)}(x)$ if $w_k' = 1$ $(k < j)$. If e' = "observation by port j of resetting of flip-flop by port k to the left," the clock setting distribution function $F(x; s', e', s, e^*) = 1_{[R(k,j),\infty)}(x)$ if $w_k' = 3$.

3. Returns to a fixed state

Recurrence properties of the underlying stochastic process of a discrete-event simulation are needed to establish estimation procedures based on regenerative processes. Lemma 4 is a special case of a generalized Borel-Cantelli lemma due to Doob [13, p. 324]. The elementary proof given below uses a "geometric trials" argument.

- *Lemma 4*

Let $\{Y_n : n \geq 0\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking on values in a set S. Let $s' \in S$. Suppose that there exists $\delta > 0$ such that

$$P\{Y_n = s' \mid Y_{n-1}, \cdots, Y_0\} \geq \delta \quad \text{a.s.} \tag{12}$$

for all $n \geq 1$. Then $P\{Y_n = s' \quad \text{i.o.}\} = 1$.

Proof

Let I be the index of first entrance time of $\{Y_n : n \geq 0\}$ to state s':

$$I = \min\{n \geq 1 : Y_n = s'\}.$$

Then

$$P\{I > n\} = P\{Y_n \neq s', \cdots, Y_1 \neq s'\}$$

and it is sufficient to show that $P\{I > n\} \leq (1 - \delta)^n$ for all $n \geq 1$. For any n,

$$\begin{aligned} P\{I > n\} &= P\{Y_n \neq s', \cdots, Y_1 \neq s'\} \\ &= E\{P\{Y_n \neq s', \cdots, Y_1 \neq s' \mid Y_{n-1}, \cdots, Y_1\}\} \\ &= E\{1_{\{Y_{n-1} \neq s', \cdots, Y_1 \neq s'\}} P\{Y_n \neq s' \mid Y_{n-1}, \cdots, Y_1\}\} \\ &\leq E\{1_{\{Y_{n-1} \neq s', \cdots, Y_1 \neq s'\}} (1 - \delta)\} \\ &= (1 - \delta) P\{I > n - 1\} \end{aligned}$$

and therefore $P\{I > n\} \leq (1 - \delta)^n$. □

Lemma 4 provides a means of showing that the underlying stochastic process of a simulation returns infinitely often to a fixed state. Specifically, let $\{X(t) : t \geq 0\}$ be a stochastic process with right-continuous and piecewise constant sample paths and countable state space, S. Let $s' \in S$ and suppose that $\{T_n : n \geq 0\}$ is an increasing sequence of finite ($T_n < \infty$ a.s.) state transition times for $\{X(t) : t \geq 0\}$ such that

$$P\{X(T_n) = s' \mid X(T_{n-1}), \cdots, X(T_0)\} \geq \delta \text{ a.s.}$$

for some $\delta > 0$. Then $P\{X(T_n) = s' \text{ i.o.}\} = 1$ by Lemma 4 [with $Y_n = X(T_n)$]. In practice, it can be difficult to show that $T_n < \infty$ a.s.

The argument used in Example 5 is due to Richard Tweedie.

• *Example 5*

In the token ring model of Example 1 let T_n be the nth time at which port 1 observes the token, $n \geq 0$. Then there is a packet queued for transmission at ports 2, 3, $\cdots$, N and port 1 starts transmission of a packet at time T_n if $X(T_n) = s'$, where $s' = (0, 1, \cdots, 1, 1, 2)$. Lemma 4 implies that $P\{X(T_n) = s' \text{ i.o.}\} = 1$ provided that

$$P\{A_j > x + y \mid A_j > y\} \leq P\{A_j > x\} \tag{13}$$

for all $x, y \geq 0$ and

$$P\{A_j \leq R_j + \cdots + R_N\} > 0, \tag{14}$$

$j = 1, 2, \cdots, N$. First observe that $T_n < \infty$ a.s. since

$$E\{T_n - T_{n-1}\} \leq R_1 + \cdots + R_N + \sum_{j=1}^{N} E\{L_j\} < \infty$$

for all $n \geq 1$. Now set

$$\delta = \prod_{j=1}^{N} P\{A_j \leq R_j + \cdots + R_N\}.$$

By Eq. (14), $\delta > 0$ and we claim that

$$P\{X(T_n) = s' \mid X(T_{n-1}), \cdots, X(T_0)\} \geq \delta. \tag{15}$$

To see this, let $T_n(j)$ be the first time after T_{n-1} that the token leaves port j; i.e.,

$$T_n(N) = \inf\{t > T_{n-1} : N(t) = 1 \text{ and } M(t) = 0\}$$

and

$$T_n(j) = \inf\{t > T_{n-1} : N(t) = j + 1 \text{ and } M(t) = 0\},$$

$j = 1, 2, \cdots, N - 1$. The definition of $T_n(j)$ implies that there is no packet queued for transmission at port j at time $T_n(j)$ and that $T_n - T_n(j) \geq R_j + \cdots + R_N$, the time for the token to propagate from port j to port 1. Equation (13) ensures that

$$P\{Z_j(T_n-) = 1 \mid X(T_{n-1}), \cdots, X(T_0)\} \geq P\{A_j \leq R_j + \cdots + R_N\}$$

for all j and therefore that

$$\begin{aligned} P\{X(T_n) = s' \mid X(T_{n-1}), \cdots, X(T_0)\} &= P\{Z_1(T_n-) \\ &= 1, \cdots, Z_N(T_n-) = 1 \mid X(T_{n-1}), \cdots, X(T_0)\} \\ &\geq \delta. \end{aligned}$$

4. Regenerative generalized semi-Markov processes

Heuristically, a regenerative stochastic process has the characteristic property that there exist random time points, referred to as *regeneration points* or *regeneration times*, at which the process probabilistically restarts. Typically, a regenerative process probabilistically starts afresh when the process returns to some fixed state. The essence of regeneration is that between any two successive regeneration points the evolution of the process is a probabilistic replica of the process between any other such pair of regeneration points.

In the presence of certain regularity conditions, a regenerative stochastic process $\{X(t) : t \geq 0\}$ has a limiting distribution provided that the time between regeneration points is finite. Furthermore, the regenerative structure ensures that the behavior of the process between two successive regeneration points determines the limiting distribution of the process as a ratio of expected values. A consequence of these results (Crane and Iglehart [14]) is that a strongly consistent point estimate and asymptotically valid confidence interval for the expected value of a general (measurable) function of the limiting random variable X can be obtained by observation of a finite portion of a single sample path of the regenerative process. This is accomplished by simulating the process in cycles and measuring certain quantities defined within the individual cycles.

Irreducible and positive recurrent continuous time Markov chains having a finite or countable state space are the most familiar examples of a regenerative process in continuous time. The successive entrances of such a Markov chain to any fixed state form a sequence of regeneration points. It is frequently difficult, however, to show that the underlying

stochastic process of a non-Markovian system is regenerative. Typically, the problem lies primarily in establishing conditions under which the process returns infinitely often to a fixed state.

The usual formal definition (cf. Smith [15]) of a regenerative process is in terms of the pasting together of so-called "tours." We give an equivalent definition.

• *Definition 6*

A *stopping time* for a stochastic process $\{X(t) : t \geq 0\}$ is a random variable T [taking values in $[0, \infty)$] such that for every finite $t \geq 0$, the occurrence or nonoccurrence of the event $\{T \leq t\}$ can be determined from the history $\{X(u) : u \leq t\}$ of the process up to time t.

• *Definition 7*

The real (possibly vector-valued) stochastic process $\{X(t) : t \geq 0\}$ is a *regenerative process in continuous time* provided that

1. There exists a sequence of stopping times $\{T_k : k \geq 0\}$ such that $\{T_{k+1} - T_k : k \geq 0\}$ are independent and identically distributed;
2. For every sequence of times $0 < t_1 < t_2 < \cdots < t_m$ $(m \geq 1)$ and $k \geq 0$, the random vectors $\{X(t_1), \cdots, X(t_m)\}$ and $\{X(T_k + t_1), \cdots, X(T_k + t_m)\}$ have the same distribution and the processes $\{X(t) : t < T_k\}$ and $\{X(T_k + t) : t \geq 0\}$ are independent.

According to Definition 7, every regenerative process has an embedded renewal process. The random times $\{T_k : k \geq 0\}$ are regeneration points for the process $\{X(t) : t \geq 0\}$, and the time interval $[T_{k-1}, T_k)$ is called the kth *cycle* of the process. The requirement that the regeneration points be stopping times means that for any fixed t the occurrence of a regeneration point prior to time t (i.e., $T_1 \leq t$) depends on the evolution of the process $\{X(t) : t \geq 0\}$ in the interval $(0, t]$ but not beyond time t.

Proposition 8 gives a set of conditions on the building blocks of a GSMP which ensures that the process is regenerative and that the expected time between regeneration points is finite. The latter result is due to Peter Glynn.

• *Proposition 8*

Let $\{X(t) : t \geq 0\}$ be a GSMP with a finite state space S and event set E. Suppose that there exist states $s_0, s_0' \in S$ and an event $e^* \in E$ such that $p(s_0'; s_0, e^*) > 0$ and $O(s_0'; s_0, e^*) = E(s_0') \cap (E(s_0) - \{e^*\}) = \varnothing$. Also suppose that there exists an increasing sequence of stopping times $\{T_n : n \geq 0\}$ that are finite ($T_n < \infty$ a.s.) state transition times at which e^* is the trigger event and $\delta > 0$ such that

$P\{V(T_n) = (s_0, s_0') \mid V(T_{n-1}), \cdots, V(T_0)\} \geq \delta$ a.s., where $L(t)$ is the last state occupied by the GSMP before jumping to $X(t)$ and $V(t) = (L(t), X(t))$. Then $\{X(t) : t \geq 0\}$ is a regenerative process in continuous time. Moreover, if

$$\overline{\lim_{n \to \infty}} \frac{T_n}{n} = \alpha < \infty \quad \text{a.s.},$$

then the expected time between regeneration points is finite.

Proof

Since $P\{V(T_n) = (s_0, s_0') \mid V(T_{n-1}), \cdots, V(T_0)\} \geq \delta > 0$ and $T_n < \infty$ a.s., Lemma 4 ensures that $\{V(T_n) : n \geq 0\}$ hits state (s_0, s_0') infinitely often with probability one. Furthermore, at such a time T_n, the only clocks that are active have just been set since $O(s_0'; s_0, e^*) = \varnothing$. The joint distribution of $X(T_n)$ and the clocks set at time T_n depends on the past history of $\{X(t) : t \geq 0\}$ only through s_0', the previous state s_0, and the trigger event e^*. Therefore, the subset of times T_n at which event e^* triggers a transition from state s_0 to state s_0' are regeneration points for the process $\{X(t) : t \geq 0\}$.

To show that the expected time between regeneration points is finite, let $\{S_n' : n \geq 1\}$ be the regeneration points; i.e.,

$$S_n' = \inf\{T_k > S_{n-1}' : X(T_k) = s_0', X(T_k-) = s_0\}.$$

Then $E\{S_n' - S_{n-1}'\} < \infty$ if and only if

$$\overline{\lim_{n \to \infty}} \frac{S_n'}{n} < \infty \quad \text{a.s.}$$

Next observe that $S_n' = T_{k(n)}$ for some sequence $\{k(n) : n \geq 1\}$ and that

$$\frac{S_n'}{n} = \frac{T_{k(n)}}{k(n)} \frac{k(n)}{n}.$$

Thus,

$$\overline{\lim_{n \to \infty}} \frac{S_n'}{n} \leq \alpha \overline{\lim_{n \to \infty}} \frac{k(n)}{n},$$

and it can be shown (using an argument similar to that in Lemma 4 and the Borel-Cantelli lemma) that

$$\overline{\lim_{n \to \infty}} \frac{k(n)}{n} < \infty \quad \text{a.s.}$$

so that

$$\overline{\lim_{n \to \infty}} \frac{S_n'}{n} < \infty \quad \text{a.s.}$$

and $E\{S_n' - S_{n-1}'\} < \infty$. □

• *Example 9*

In the token ring of Example 1, take $s_0' = (0, 1, \cdots, 1, 1, 2)$ and e^* = "observation of token." A transition to state s_0' can

occur when event e^* is the trigger event only if e^* occurs in state $s_0 = (1, \cdots, 1, 0, 1)$ and in this case the set $O(s_0', s_0, e^*) = \varnothing$. If T_n is the nth time that port 1 observes the token, there exists $\delta > 0$ such that $P\{X(T_n) = s_0' | X(T_n), \cdots, X(T_0)\} \geq \delta$ by the argument in Example 5. The successive times T_n at which e^* is the trigger event in state s_0 (and there is a transition to state s_0') are regeneration points for the process $\{X(t) : t \geq 0\}$.

Next observe that

$$T_n - T_{n-1} \leq R_1 + \cdots + R_N + L_{1n} + \cdots + L_{Nn},$$

where L_{jn} is distributed as L_j. Thus,

$$\frac{T_n}{n} \leq \sum_{k=1}^{n} \frac{U_k}{n},$$

where $U_k = R_1 + \cdots + R_N + L_{1k} + \cdots + L_{Nk}$. By the strong law of large numbers

$$\sum_{k=1}^{n} \frac{U_k}{n} \to E\{U\} < \infty.$$

Therefore,

$$\overline{\lim_{n\to\infty}} \frac{T_n}{n} < \infty \quad \text{a.s.}$$

and the expected time between regeneration points is finite.

5. Concluding remarks

Most discrete-event simulations can be modeled within the GSMP framework. In some stochastic systems, however, it is possible to define a system state which maintains sufficient information to determine the quantities of interest and to specify an algorithm for generating sample paths of the associated stochastic process, but the process does not have a GSMP representation. As an example, suppose that the state of the collision-free bus network at time t is defined to be

$$X(t) = (W(t), U(t), V(t)),$$

where $W(t)$ and $U(t)$ are as in Example 3, $V_j(t)$ is the number of ports to the left observed by port j to have set their flip-flop and $V(t) = (V_1(t), \cdots, V_N(t))$. The process $\{X(t) : t \geq 0\}$ has a finite state space, S. It does not appear to be possible, however, to specify an event set E such that $\{X(t) : t \geq 0\}$ is a GSMP with state space S and event set E. For example, suppose that E is the set of events: "end of transmission by port j," "end of a wait for $R(j) + T$," "setting of flip-flop by port j," "observation by port j of start of transmission," "observation by port j of end of transmission," "observation by port j of setting of flip-flop by port to the left," and "observation by port j of resetting of flip-flop by port to the left," $j = 1, 2, \cdots, N$. Then $\{X(t) : t \geq 0\}$ fails to be a GSMP because there are states for which it is not possible to determine whether or not the event "observation by port j of setting of flip-flop by port to the left" or the event "observation by port j of resetting of a flip-flop by port to the left" is active. (Select i, j, and k with $1 \leq i < j < k \leq N$ and take $s = (w_1, \cdots, w_N, u_1, \cdots, u_N, v_1, \cdots v_N)$ with $w_i = 3$, $w_k = 1$, and $v_j = 1$.)

Acknowledgments

The authors have benefitted from stimulating technical discussions with Carl Hamacher and Richard Tweedie and from the comments of anonymous referees. Both authors are grateful to the National Science Foundation for support under Grant MCS-8203483. In addition, Donald L. Iglehart gratefully acknowledges partial support under Office of Naval Research Contract N00014-76-C-0578 (NR 042-343).

References

1. W. M. Loucks, V. C. Hamacher, and B. Preiss, "Performance of Short Packet Local Area Rings," *Technical Report*, Departments of Electrical Engineering and Computer Science, University of Toronto, Ontario, Canada.
2. W. Whitt, "Continuity of Generalized Semi-Markov Processes," *Math. Oper. Res.* **5,** 494–501 (1980).
3. P. W. Glynn, Forthcoming technical report, Department of Industrial Engineering, University of Wisconsin, Madison, WI, 1983.
4. L. D. Fossett, "Simulating Generalized Semi-Markov Processes," *Technical Report No. 4*, Department of Operations Research, Stanford University, CA, 1979.
5. D. L. Iglehart and G. S. Shedler, "Statistical Efficiency of Regenerative Simulation Methods for Networks of Queues," *Adv. Appl. Prob.* **15,** 183–197 (1983).
6. E. Nummelin, "A Splitting Technique for ϕ-Recurrent Markov Chains," *Technical Report MAT A80*, Helsinki University of Technology, Finland, 1976.
7. P. Tuominen and R. L. Tweedie, "Exponential Ergodicity in Markovian Queueing and Dam Models," *J. Appl. Prob.* **16,** 867–880 (1979).
8. K. Matthes, "Zur Theorie der Bedienungsprozesse," *Trans. 3rd Prague Conference on Information Theory and Statistical Decision Functions*, Prague, 1962.
9. D. König, K. Matthes, and K. Nawrotzki, *Verallgemeinerungen der Erlangschen und Engsetschen Formeln*, Akademie-Verlag, Berlin, 1967.
10. D. König, K. Matthes, and K. Nawrotzki, "Unempfindlichkeitseigenschaften von Bedienungsprozessen," Appendix to *Introduction to Queueing Theory*, B. V. Gnedenko and I. N. Kovalenko, German edition, 1974.
11. Gerald S. Shedler and Jonathan Southard, "Regenerative Simulation of Networks of Queues with General Service Times: Passage Through Subnetworks, *IBM J. Res. Develop.* **26,** 625–633 (1982).
12. K. P. Eswaran, V. C. Hamacher, and G. S. Shedler, "Collision-Free Access Control for Computer Communication Bus Networks," *IEEE Trans. Software Engineering* **SE-7,** 574–582 (1981).
13. J. L. Doob, *Stochastic Processes*, John Wiley & Sons, Inc., New York, 1953.
14. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic Systems: III, Regenerative Processes and Discrete Event Simulation," *Oper. Res.* **23,** 33–45 (1975).
15. W. L. Smith, "Renewal Theory and Its Ramifications," *J. Roy. Statist. Soc. Ser. B* **20,** 243–302 (1958).

Received April 13, 1983; revised May 11, 1983

Donald L. Iglehart *Stanford University, Stanford, California 94305.* Dr. Iglehart is Professor of Operations Research in the Department of Operations Research at Stanford University. He received a B.S. in engineering physics from Cornell University, Ithaca, New York, and an M.S. and a Ph.D. in mathematical statistics from Stanford University. He has published papers on inventory theory, queueing theory, simulation methodology, and stochastic processes.

Gerald S. Shedler *IBM Research Division, 5600 Cottle Road, San Jose, California 95193.* Mr. Shedler has been a Research staff member at IBM since 1965, initially at the Thomas J. Watson Research Center, Yorktown Heights, New York, and since 1970, in the Computer Science Department at the Research laboratory in San Jose. During 1973–1974, while on sabbatical from IBM, he was associated with Stanford University as Acting Associate Professor in the Department of Operations Research, and subsequently has been Consulting Associate Professor in the same department. He has worked extensively on applications of stochastic processes, particularly to performance evaluation of computer systems. His current research is on discrete event methods for simulation of stochastic systems. Mr. Shedler is coauthor with Donald L. Iglehart of *Regenerative Simulation of Response Times in Networks of Queues*, published in 1980.