Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

STOCHASTIC PETRI NETS FOR MODELLING AND SIMULATION

Peter J. Haas

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120, U.S.A.

ABSTRACT

Stochastic Petri nets (SPNs) have proven to be a powerful and enduring graphically-oriented framework for modelling and performance analysis of complex systems. This tutorial focuses on the use of SPNs in discrete-event simulation. After describing the basic SPN building blocks and discussing the modelling power of the formalism, we present elements of a steady-state simulation theory for SPNs. Specifically, we provide conditions on the building blocks of an SPN that ensure long-run stability for the underlying marking process (or for a sequence of delays determined by the marking process) and the validity of estimation procedures such as the regenerative method, the method of batch means, and spectral methods.

1 INTRODUCTION

Developing and analyzing stochastic simulation models of complex computer, manufacturing, telecommunication, workflow, or transportation systems is almost always a challenging task. Real-world systems usually comprise multiple activities or processes that proceed concurrently. Activities often have precedence relationships, e.g., assembly of a part in a manufacturing cell does not begin until assembly of each of its subparts has completed. Specified activities may be synchronized in that they must always start or terminate at the same time. Activities frequently compete for limited resources, and one activity may have either preemptive or nonpreemptive priority over another activity for use of a resource. The generalized semi-Markov process (GSMP) is the traditional framework for mathematical modelling of general discrete-event stochastic systems. Although useful for a unified theoretical treatment of discrete-event systems, the GSMP framework is not always well suited to practical modelling tasks. In particular, the modeller is forced to specify the "state of the system" directly as an abstract vector of random variables. Such a specification can be highly nontrivial: the system state definition must be as concise as possible for reasons of efficiency, but must also contain enough information so that a sequence of state transitions and transition times can be generated during a simulation run and system characteristics of interest can be determined from the sequence.

Stochastic Petri nets (SPNs), introduced by the computer science community in the early 1980s, are very appealing in that they not only have the same modelling power as GSMPs but also admit a graphical representation that is well suited to top-down and bottom-up modelling of complex systems. SPNs are a probabilistic extension of the original nets introduced by Carl Adam Petri in his 1962 Ph.D. dissertation. At present, the literature contains over 8800 books, papers, and reports dealing with Petri nets and their extensions. A variety of computer packages are available for simulation and analysis of Petri nets, and ISO/IEC standard 15909 for Petri net models is currently under development.

This tutorial provides an introduction to SPNs, with an emphasis on those aspects of SPNs that are pertinent to simulation. We first describe the basic SPN building blocks, illustrate the use of SPNs as models of discrete-event systems, and discuss the modelling power of the formalism. We then present elements of a steady-state simulation theory for SPNs by providing conditions on the building blocks of an SPN that ensure long-run stability for the key stochastic processes associated with the net and the validity of estimation procedures such as the regenerative method, the method of batch means, and spectral methods. Our presentation primarily follows the monograph of Haas (2002) as well as recent papers by Glynn and Haas (2004, 2005), and we refer the reader to these sources for further details of the results presented here as well as various extensions, refinements, and pointers to the SPN literature.

2 THE SPN MODEL

The SPN framework provides a powerful set of building blocks for specifying the state-transition mechanism and event-scheduling mechanism of a discrete-event stochastic system. We give an overview of the SPN building blocks and then formally define the "marking process" of the SPN,

which records the state of the net as it evolves over continuous time.

2.1 Building Blocks

An SPN is specified by a finite set of *places* and a finite number of *transitions* along with a *normal input function*, an *inhibitor input function,* and an *output function* (each of which associates a set of places with a transition). A *marking* of an SPN is an assignment of *token* counts (nonnegative integers) to the places of the net. A transition is *enabled* whenever there is at least one token in each of its normal input places and no tokens in any of its inhibitor input places; otherwise, it is *disabled*. An enabled transition *fires* by removing one token per place from a random subset of its normal input places and depositing one token per place in a random subset of its output places. An *immediate* transition fires the instant it becomes enabled, whereas a *timed* transition fires after a positive (and usually random) amount of time. In the context of discrete-event systems, the marking of the SPN corresponds to the state of the system, and the firing of a transition corresponds to the occurrence of an event.

SPNs have a natural graphical representation—see Figure 1—that facilitates modelling of discrete-event systems. This bipartite graph of the places and transitions of an SPN determines the event-scheduling mechanism. In the graphical representation of an SPN, places are drawn as circles, immediate transitions as thin bars, and timed transitions as thick bars. Directed arcs connect transitions to output places and normal input places to transitions; arcs terminating in open dots connect inhibitor input places to transitions. Tokens are drawn as black dots.

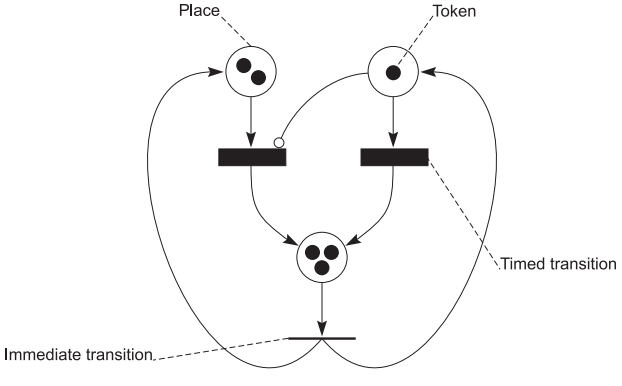

Figure 1: SPN Building Blocks

Heuristically, an SPN changes marking in accordance with the firing of a transition enabled in the current marking (or with the simultaneous firing of two or more transitions enabled in the current marking). Here the new marking may coincide with the current marking. The times at which transitions fire are determined by a stochastic mechanism. Specifically, a *clock* is associated with each transition. The *clock reading* for an enabled transition indicates the remaining time until the transition is scheduled to fire. Clocks run down at marking-dependent *speeds*, and a marking change occurs when one or more clocks run down to 0. The transitions enabled in a marking therefore compete to change the marking: the transitions whose clocks run down to 0 first are the "winners."

At time 0 the initial marking and clock readings are selected according to an initial probability distribution. At each subsequent marking change there are three types of transitions:

1. A *new* transition is enabled in the new marking and either is not enabled in the old marking—so that no clock reading is associated with the transition just before the marking change—or is in the set of transitions that triggers the marking change—so that the associated clock reading is 0 just before the marking change. For such a transition, a new clock reading is generated according to a probability distribution that depends only on the old marking, the new marking, and the set of transitions that triggers the marking change.
2. An *old* transition is enabled in both the old and new markings and is not in the set of transitions that triggers the marking change. The clock for such a transition continues to run down (perhaps at a new speed).
3. A *newly disabled* transition is enabled in the old marking and disabled in the new marking. If the transition is not in the set of transitions that triggers the marking change, then it is "cancelled" and its clock reading is discarded. Otherwise, the clock associated with the transition has just run down to 0 and no new clock reading is generated.

As mentioned before, we distinguish between immediate transitions that always fire the instant they become enabled and timed transitions that fire only after a positive amount of time elapses. The clock reading generated for a new immediate transition is always equal to 0 with probability 1, whereas the clock reading generated for a new timed transition is always positive with probability 1. If at least one immediate transition is enabled in a marking—as in Figure 1—then the marking is *immediate*; otherwise, the marking is *timed*. An immediate marking vanishes the instant it is attained.

SPNs are well-suited to modelling concurrency, synchronization, precedence, and priority, and are conducive to both bottom-up and top-down modelling. In bottom-up modelling, a detailed subnet is developed for each com-

ponent of a system, and then the subnets are combined to form the overall SPN model. In top-down modelling, a preliminary SPN model is developed that captures the main interactions between the components of the system without modelling each component in detail. Then the subnets corresponding to the system components are each progressively refined until the model is sufficiently detailed.

Example 1 (Cyclic queues with feedback) Consider a closed network of queues with two single-server service centers and N (≥ 2) jobs; see Figure 2. With fixed probability $p \in (0, 1)$, a job that completes service at center 1 moves to center 2 and with probability $1 - p$ joins the tail of the queue at center 1. A job that completes service at center 2 moves to center 1. The queueing discipline at each center is first-come, first-served. Successive service times at center i ($i = 1, 2$) are i.i.d. as a random variable L_i having a continuous distribution function. For future reference, we also assign a "position" to each job in the network, as indicated in the figure.

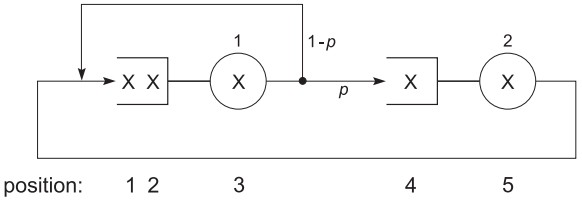

Figure 2: Cyclic Queues with Feedback (Five Jobs)

An SPN model of this system is displayed in Figure 3. The tokens in place $d_{1,i}$ correspond to the jobs waiting in queue at center i, and a token in place $d_{2,i}$ corresponds to a job that is undergoing service at center i. A marking of the net corresponds to a vector of the four token counts in places $d_{1,1}$, $d_{2,1}$, $d_{1,2}$, and $d_{2,2}$, respectively; e.g., the marking displayed in Figure 3 is denoted as $s = (2, 1, 1, 1)$.

$e_{1,i}$ = start of service at center i
$e_{2,i}$ = completion of service at center i

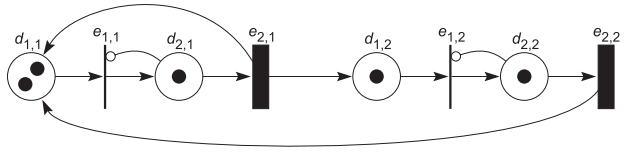

Figure 3: SPN Representation of Cyclic Queues with Feedback

Whenever transition $e_{2,2}$ fires, it removes a token from place $d_{2,2}$ and deposits a token in place $d_{1,1}$, reflecting the fact that a job that completes service at center 2 moves to center 1. Whenever transition $e_{2,1}$ fires, it removes a token from place $d_{2,1}$; moreover, it deposits a token in place $d_{1,2}$ with probability p and in place $d_{1,1}$ with probability $1 - p$. In this manner the SPN model captures the feedback mechanism in the network of queues.

Transition $e_{1,i}$ = "start of service at center i" is immediate for $i = 1, 2$, reflecting the fact that a job starts to undergo service at the same instant it is selected for service. Whenever transition $e_{1,i}$ fires, it removes a token from place $d_{1,i}$ and deposits a token in place $d_{2,i}$. Suppose, for example, that the marking is $s = (2, 1, 1, 1)$ as pictured in Figure 3 and transition $e_{2,2}$ fires, so that the marking changes to the immediate marking $s' = (3, 1, 1, 0)$. Then transition $e_{1,2}$ becomes enabled and fires immediately, changing the marking to $s'' = (3, 1, 0, 1)$. Transition $e_{2,2}$ becomes enabled at this marking change and a new clock reading is generated from the distribution of L_2. Observe that, due to the inhibitor arcs, transition $e_{1,i}$ never fires when place $d_{2,i}$ contains a token, reflecting the fact that at most one job at each center can undergo service at any time.

The SPN representation of a system need not be unique. For example, Figure 4 displays an alternative SPN representation of the network of queues. This SPN does not distinguish between a job undergoing service and jobs waiting in queue; that is, the tokens in place d_i correspond to all jobs (in queue or undergoing service) at center i for $i = 1, 2$. Although the original SPN in Figure 3 represents the service mechanism in greater detail, the SPN in Figure 4 is easier to work with in practice: the latter SPN has fewer places and transitions but can be used to study any performance characteristic that can be studied using the former SPN. □

e_1 = service completion at center 1
e_2 = service completion at center 2

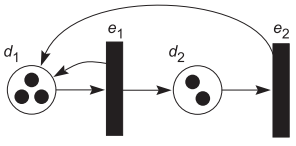

Figure 4: Alternative SPN Representation of Cyclic Queues with Feedback

In order to obtain SPN graphs that are more concise, we can associate "colors" with both tokens and transitions and work with "colored stochastic Petri nets" (CSPNs). Such nets are especially well suited for systems that are comprised of subsystems having similar structure or behavior. Chapter 9 in Haas (2002) describes extensions of the results in this paper to the CSPN setting, and shows how symmetry

with respect to color can be exploited to improve simulation efficiency.

2.2 The Marking Process

We now provide some notation for SPNs and formally define the marking process of the net in terms of an underlying Markov chain having a general state space. Let $D = \{d_1, d_2, \ldots, d_L\}$ be a finite set of *places*, $E = \{e_1, e_2, \ldots, e_M\}$ be a finite set of *transitions*, and $E' \subset E$ a (possibly empty) set of *immediate transitions*. The transitions in $E - E'$ are called *timed transitions*. Also let $I(e), L(e), J(e) \subseteq D$ be the sets of *normal input places*, *inhibitor input places*, and *output places*, respectively, for transition $e \in E$. Denote by G the finite or countably infinite set of *markings*. For $s \in G$ we write $s = (s_1, s_2, \ldots, s_L)$, where s_j is the number of *tokens* in place $d_j \in D$. Set $E(s) = \{e \in E\colon s_j \geq 1 \text{ for } d_j \in I(e) \text{ and } s_j = 0 \text{ for } d_j \in L(e)\}$, so that $E(s)$ (assumed nonempty) is the set of transitions that are *enabled* when the marking is s. A transition $e \in E - E(s)$ is *disabled* when the marking is s. Define the set S' of *immediate markings* by $S' = \{s \in G\colon E(s) \cap E' \neq \varnothing\}$ and the set S of *timed markings* by $S = G - S' = \{s \in G\colon E(s) \cap E' = \varnothing\}$.

For $E^* \subseteq E(s)$, denote by $p(s'; s, E^*)$ the probability that the new marking is s' given that the marking is s and the transitions in the set E^* fire simultaneously. The new-marking probabilities are constrained by the requirement that a transition remove at most one token from each normal input place and deposit at most one token in each output place when it fires. The token count of a place may increase or decrease by more than 1 when transitions fire simultaneously.

As mentioned previously, the clocks associated with the transitions of the net, along with the speeds at which the clocks run down, determine which of the enabled transitions trigger the next marking change. Denote by $r(s, e)$ (≥ 0) the *speed* (finite, deterministic rate) at which the clock associated with transition e runs down when the marking is s. The requirement that $r(s, e)$ be finite is needed to ensure that timed transitions never fire instantaneously. We do not allow zero speeds for immediate transitions; such transitions always fire the instant that they become enabled. By convention, $r(s, e) = 1$ if $e \in E' \cap E(s)$. We assume that $r(s, e) > 0$ for some $e \in E(s)$. Let $C(s)$ be the set of possible *clock-reading vectors* when the marking is s: $C(s) = \{c = (c_1, \ldots, c_M)\colon c_i \geq 0 \text{ and } c_i > 0 \text{ if and only if } e_i \in E(s) - E'\}$. (The ith component of a clock-reading vector $c = (c_1, \ldots, c_M)$ is the clock reading associated with transition e_i.) Beginning in marking s with clock-reading vector $c = (c_1, \ldots, c_M) \in C(s)$, the time $t^*(s, c)$ to the next marking change is given by $t^*(s, c) = \min_{\{i\colon e_i \in E(s)\}} c_i / r(s, e_i)$, where $c_i / r(s, e_i)$ is taken to be $+\infty$ when $r(s, e_i) = 0$. The set of transitions $E^*(s, c)$ that fire simultaneously and trigger the next marking change is given by $E^*(s, c) = \{e_i \in E(s)\colon c_i - t^*(s, c) r(s, e_i) = 0\}$. Observe that $E^*(s, c) = E(s) \cap E'$ whenever the marking s is immediate.

At a marking change from s to s' triggered by the simultaneous firing of the transitions in the set E^*, a finite clock reading is generated for each *new transition* $e' \in N(s'; s, E^*) = E(s') - (E(s) - E^*)$. Denote the *clock-setting distribution function* (that is, the distribution function of such a new clock reading) by $F(\cdot; s', e', s, E^*)$. For $e' \in E'$, we require that $F(0; s', e', s, E^*) = 1$ for s, s' and E^* so that immediate transitions always fire instantaneously. For $e' \in E - E'$, we require that $F(0; s', e', s, E^*) = 0$ for s, s' and E^* so that timed transitions never fire instantaneously. For each *old transition* $e' \in O(s'; s, E^*) = E(s') \cap \big(E(s) - E^*\big)$, the old clock reading is kept after the marking change. For $e' \in (E(s) - E^*) - E(s')$, transition e' (that was enabled before the transitions in E^* fired) becomes disabled and the clock reading is discarded.

Denote by $X(t)$ the marking of the SPN at time t. Formal definition of the *marking process* $\{X(t)\colon t \geq 0\}$ of an SPN with general firing times is in terms of a general state space Markov chain $\{(S_n, C_n)\colon n \geq 0\}$ that describes the net at successive marking changes; see Section 3.1 in Haas (2002) for a formal definition of this chain. Heuristically, $S_n = (S_{n,1}, \ldots, S_{n,L})$ represents the marking and $C_n = (C_{n,1}, \ldots, C_{n,M})$ represents the clock-reading vector just after the nth marking change. The chain takes values in the set $\Sigma = \bigcup_{s \in S} \big(\{s\} \times C(s)\big)$. Denote by μ the *initial distribution* of the chain; for a subset $B \subseteq \Sigma$, the quantity $\mu(B)$ represents the probability that $(S_0, C_0) \in B$. We assume throughout that the initial marking s_0 is selected according to a (possibly degenerate) initial-marking distribution function ν_0 and then, for each enabled transition $e_i \in E(s_0)$, the corresponding clock reading $c_{0,i}$ is generated according to an initial clock-setting distribution function $F_0(\cdot; e_i, s_0)$. Thus $\mu(A) = \nu_0(s_0) \prod_{e \in E(s_0)} F_0(a_i; e, s_0)$ for all sets $A = \{s_0\} \times \{(c_{0,1}, \ldots, c_{0,M}) \in C(s_0)\colon 0 \leq c_{0,i} \leq a_i \text{ for } 1 \leq i \leq M\}$. Let ζ_n ($n \geq 0$) be the nonnegative, real-valued time of the nth marking change: $\zeta_n = \sum_{j=0}^{n-1} t^*(S_j, C_j)$. The marking process is then defined by setting

$$X(t) = \begin{cases} S_{N(t)} & \text{if } N(t) < \infty; \\ \Delta & \text{if } N(t) = \infty, \end{cases}$$

where $\Delta \notin G$ and $N(t) = \sup\{n \geq 0\colon \zeta_n \leq t\}$ is the number of marking changes that occur in the interval $(0, t]$. By construction, the marking process takes values in the set $S \cup \{\Delta\}$ and has piecewise constant, right-continuous sample paths. We assume throughout that

$$P_\mu\left\{\sup_{n \geq 0} \zeta_n = \infty\right\} = 1 \tag{1}$$

so that only a finite number of marking changes occur in any finite time interval. (Here as elsewhere, we use the notation P_μ to emphasize the fact that the initial state (S_0, C_0) is distributed according to μ.) Some sufficient conditions for (1) to hold are given in Section 3.3 of Haas (2002). It follows that $P_\mu\{X(t) \neq \Delta \text{ for } t \geq 0\} = 1$, and hence the state space of the marking process can be taken simply as S.

We often write $E_n^* = E^*(S_n, C_n)$ for $n \geq 0$. We also define the *embedded chain* $\{(S_n^+, C_n^+) : n \geq 0\}$ of the marking process of an SPN to be the discrete-time process that records the marking and clock-reading vector just after each marking change at which the new marking is timed; it can be shown that $\{(S_n^+, C_n^+) : n \geq 0\}$ is indeed a Markov chain. Denote the state space of the embedded chain by Σ^+.

2.3 Modelling Power

Exactly how large a class of discrete-event systems can be modelled within the SPN framework? Although this question cannot be answered precisely, the modelling power of SPNs can usefully be compared with that of GSMPs. The GSMP is the traditional model for the underlying stochastic process of a discrete-event system, and a wide range of computer, communication, manufacturing, and transportation systems have been modelled as GSMPs. Thus GSMPs are a good benchmark for assessment of modelling power. GSMPs have a more general state-transition mechanism, event-scheduling mechanism, and form of the state space than SPNs. It might therefore be conjectured that SPNs have less modelling power than GSMPs.

It is shown in Chapter 4 of Haas (2002) that, on the contrary, SPNs have the same modelling power as GSMPs. More specifically, it is shown that for any GSMP there exists an SPN that "strongly mimics" the GSMP in the sense that the marking process and underlying chain of the SPN have the same finite-dimensional distributions as those of the GSMP and its underlying chain under an appropriate mapping between the state spaces. The converse result is also established. These results provide a justification for our SPN formulation and establish SPNs as an attractive general framework for modelling and simulation analysis. The methodology used to obtain the modelling-power results can also be used to assess the relative modelling power of different SPN formulations and the contribution of individual SPN building blocks to overall modelling power.

3 STABILITY AND SIMULATION

Engineers and systems designers are often interested in performance characteristics such as the long-run average operating cost for a flexible manufacturing system, the long-run fraction of time a database is accessible, or the long-run utilization of a communications link. When the system of interest is modelled as an SPN, each of these characteristics typically can be specified as a time-average limit of the form

$$r(f) = \lim_{t\to\infty} \frac{1}{t} \int_0^t f\big(X(u)\big)\, du, \tag{2}$$

where f is a real-valued function and $X(t)$ denotes the marking of the net at time $t \geq 0$. Other performance measures of interest can be expressed as functions of such time-average limits or as (functions of) time-average limits of the underlying chain used to define the marking process.

In this section we provide conditions on the building blocks of an SPN under which the marking process is stable, so that time-average limits are well-defined. We then provide additional conditions under which various simulation-based methods for estimating time-average limits are guaranteed to produce valid results. Although omitted for brevity, extensions of results given here can be applied to estimate other performance measures such as cumulative rewards (both continuous and impulse rewards) and gradients with respect to system parameters. Our results can also be applied in the setting of GSMPs.

3.1 Recurrence

Stability of the marking process typically follows from stability of the underlying general state-space Markov chain used to define the marking process. Perhaps the most basic notion of stability for such a chain is "Harris recurrence." Recall that a *measure* ϕ on subsets of the state space Γ of a Markov chain $\{Z_n : n \geq 0\}$ assigns to each subset a nonnegative real number in a manner such that $\phi(\varnothing) = 0$ and $\phi(\bigcup_n A_n) = \sum_n \phi(A_n)$ whenever $A_1, A_2, \ldots$ are disjoint. A measure is *nontrivial* if $\phi(A) > 0$ for some $A \subseteq \Gamma$. A typical example of a nontrivial measure is Lebesgue measure which, roughly speaking, maps the set of all points in a rectangular region to the length, area, or volume of the region, depending upon the dimensionality, and is also well defined for sets of points having more complex structure.

Definition 1 *The chain* $\{Z_n : n \geq 0\}$ *is* Harris recurrent with recurrence measure ϕ *if* ϕ *is nontrivial and* $P_z\{Z_n \in A \text{ i.o.}\} = 1$ *for all* $z \in \Gamma$ *and* $A \subseteq \Gamma$ *with* $\phi(A) > 0$.

A Harris recurrent chain has the property that any "dense enough" set of states (as measured by ϕ) is hit infinitely often (i.o.) with probability 1. Thus a Harris recurrent chain is stable in that it does not systematically drift off toward the outer reaches of the state space—fix a dense set of states that is compact, and observe that the chain repeatedly returns to this set. We require that each target set be dense because an individual state typically is hit with probability 0 when the state space of the chain is uncountably infinite.

A Harris recurrent chain admits an *invariant measure*, that is, a measure π_0 on subsets of Γ that satisfies

$$\int P(z, A)\, \pi_0(dz) = \pi_0(A) \tag{3}$$

for $A \subseteq \Gamma$. The measure π_0 is unique to within a multiplicative constant. If $\pi_0(\Gamma) < \infty$, then $\pi(\cdot) = \pi_0(\cdot)/\pi_0(\Gamma)$ is an invariant *probability* measure, and (3) can be rewritten as $P_\pi\{Z_1 \in A\} = \pi(A)$ for $A \subseteq \Gamma$. That is, if the initial state of the chain Z_0 is distributed according to π, then Z_1 is also distributed according to π. (It then follows from the Markov property that Z_k is distributed according to π for $k \geq 0$, and hence the chain is stationary.) Such a chain is called *positive* Harris recurrent, and it can be shown that the expected time between successive visits to a recurrent set A of states is finite.

We now give conditions—encapsulated in the "positive density assumption" PD(q) given in Definition 3 below—under which the embedded chain of the marking process of an SPN is positive Harris recurrent. To prepare for this definition, we first need to define a notion of irreducibility for SPNs. For $s, s' \in G$, write $s \to s'$ if either s is immediate and $p\big(s'; s, E(s) \cap E'\big) > 0$ or if s is timed and $p(s'; s, \{e\})r(s, e) > 0$ for some $e \in E(s)$. Next, write $s \rightsquigarrow s'$ if either $s \to s'$ or there exist markings $s^{(1)}, s^{(2)}, \ldots, s^{(n)} \in G$ ($n \geq 1$) such that $s \to s^{(1)} \to \cdots \to s^{(n)} \to s'$. Clearly, the relation $\rightsquigarrow$ is transitive.

Definition 2 *An* SPN *with marking set G is said to be* irreducible *if $s \rightsquigarrow s'$ for each $s, s' \in G$.*

Recall that a nonnegative function G is a *component* of a distribution function F if G is not identically equal to 0 and $G \leq F$. If G is a component of F and G is absolutely continuous, so that G has a density function g, then we say that g is a *density component* of F. If F is absolutely continuous with density function f, then f is trivially a density component of F.

Definition 3 Assumption PD(q) *is said to hold for a specified* SPN *if (i) the marking set G is finite, (ii) the* SPN *is irreducible, (iii) all speeds are positive, and (iii) there exists $0 < \bar{x} < \infty$ such that each clock-setting distribution function $F(\cdot; s', e', s, e^*)$ and $F_0(\cdot; e', s)$ with $e' \in E - E'$ has finite qth moment and has a density component that is positive and continuous on $(0, \bar{x})$.*

Observe that if Assumption PD(q) holds, then Assumption PD(q') holds for all $0 \leq q' \leq q$. Occasionally we need a strengthened version of Assumption PD(q), which we call *Assumption PDE*, in which the requirement of finite qth moment for distributions F and F_0 is replaced by the requirement that $F, F_0 \in \mathcal{G}^+$, where $\mathcal{G}^+$ is the set of distribution functions on $[0, \infty)$ that have a convergent LaPlace–Stieltjes transform in a neighborhood of the origin. That is, $F \in \mathcal{G}^+$ if and only if there exists $a_F > 0$ such that $\int_0^\infty e^{ux}\, dF(x) < \infty$ for $u \in [0, a_F]$. Observe that each distribution function $F \in \mathcal{G}^+$ has finite moments of all orders, so that if Assumption PDE holds, then Assumption PD(q) holds for all $q \geq 0$. Many common distribution functions belong to $\mathcal{G}^+$, for example, the uniform, exponential, gamma, beta, and truncated normal distributions.

We now give our key recurrence result. As before, denote by Σ and Σ^+ the state spaces of the underlying chain $\{(S_n, C_n): n \geq 0\}$ and embedded chain $\big\{(S_n^+, C_n^+): n \geq 0\big\}$, respectively. Whenever Assumption PD(q) holds, we define $\bar{\phi}$ to be the unique measure on subsets of Σ^+ such that

$$\bar{\phi}\big(\{s\} \times [0, x_1] \times [0, x_2] \times \cdots \times [0, x_M]\big) = \prod_{\{i:\, e_i \in E(s)\}} \min(x_i, \bar{x}) \tag{4}$$

for all $s \in S$ and $x_1, x_2, \ldots, x_M \geq 0$. If, for example, a set $B \subseteq \Sigma^+$ is of the form $B = \{s\} \times A$ with $E(s) = E$, then $\bar{\phi}(B)$ is equal to the Lebesgue measure of the set $A \cap [0, \bar{x}]^M$.

Theorem 1 *Suppose that Assumption PD(1) holds for an* SPN. *Then the embedded chain of the marking process is positive Harris recurrent with recurrence measure $\bar{\phi}$ given by* (4) *and hence admits an invariant probability measure π.*

In some contexts it suffices to show that a state $\bar{s}$ is recurrent in the sense the $P\{S_n = \bar{s} \text{ i.o.}\} = 1$. Theorem 1 can be specialized to establish the desired recurrence property for the specified set. Alternatively, a geometric trials technique can be used to establish recurrence. This technique, which is described in Section 5.2 of Haas (2002), exploits the detailed structure of the SPN model and avoids the somewhat restrictive positive density assumptions used in Theorem 1.

3.2 Regenerative Simulation of SPNs

Suppose that there exists a sequence of *regeneration points* for the marking process $\{X(t): t \geq 0\}$, that is, an increasing sequence $0 \leq T_0 < T_1 < T_2 < \cdots$ of a.s. finite random times such that, for $k \geq 1$, the post-T_k process $\{X(T_k + t): t \geq 0;\ \tau_{k+l}: l \geq 1\}$ is distributed as the post-T_0 process $\{X(T_0 + t): t \geq 0;\ \tau_l: l \geq 1\}$ and is independent of the pre-T_k process $\{X(t): 0 \leq t < T_k;\ \tau_1, \ldots, \tau_k\}$, where $\tau_k = T_k - T_{k-1}$. In other words, the process probabilistically "starts over from scratch" at each T_k, so that the sequence of random times $\{T_k: k \geq 0\}$ decomposes sample paths of the marking process into i.i.d. cycles. If $|G| < \infty$ and the expected cycle length $E_\mu[\tau_1]$ is finite, then the time-average limit $r(f)$ in (2) is well defined and finite for any function f. If, moreover, the "regenerative variance constant" $\sigma^2(f) = \mathrm{Var}_\mu\big[\int_{T_0}^{T_1} f\big(X(u)\big)\, du - r(f)\tau_1\big]$ is finite, then regenerative simulation methods can be used to obtain strongly consistent point estimates and asymptotic

confidence intervals for $r(f)$; see, e.g., Section 6.3 in Haas (2002).

Using results from Section 3.1, we can obtain conditions on the building blocks of an SPN under which the regenerative method is applicable. For a marking $\bar{s} \in G$ and transition $\bar{e} \in E(\bar{s})$, denote by $\{\theta(k): k \geq 0\}$ the indices of the successive marking changes at which the marking is $\bar{s}$ and transition $\bar{e}$ fires: $\theta(-1) = 0$ and

$$\theta(k) = \inf\left\{ n > \theta(k-1): S_{n-1} = \bar{s} \text{ and } E^*_{n-1} = \{\bar{e}\} \right\} \tag{5}$$

for $k \geq 0$. In accordance with our usual notation, we denote by $O(s'; \bar{s}, \bar{e})$ the set of transitions in $E - \{\bar{e}\}$ that are enabled both before and after a marking change from $\bar{s}$ to s' triggered by the firing of transition $\bar{e}$.

Theorem 2 *Let $\bar{s} \in S$ and $\bar{e} \in E(\bar{s})$. Suppose that Assumption PD(2) holds. Also suppose that for each s' such that $p(s'; \bar{s}, \{\bar{e}\}) > 0$ either*

(a) $O(s'; \bar{s}, \bar{e}) = \varnothing$ *or*

(b) $O(s'; \bar{s}, \bar{e}) \neq \varnothing$ *and the clock for each transition $e_i \in O(s'; \bar{s}, \bar{e})$ is always set according to an exponential distribution with fixed intensity $v(e_i)$.*

Then the random times $\{\zeta_{\theta(k)}: k \geq 0\}$ defined via (5) with $\bar{E} = \{\bar{e}\}$ form a sequence of regeneration points for the marking process $\{X(t): t \geq 0\}$. Moreover, the regenerative cycle length τ_1 has finite mean and the regenerative variance constant $\sigma^2(f)$ is finite for any real-valued function f defined on S.

The idea behind the theorem is that the conditions in (a) and (b) ensure that the marking process probabilistically restarts at certain marking changes, and Assumption PD(2) ensures that such marking changes occur infinitely often with probability 1. A marking $\bar{s} \in G$ is said to be a *single state* if $E(\bar{s}) = \{\bar{e}\}$ for some $\bar{e} \in E$. Observe that the condition in (a) always holds for a single state. Thus, if an SPN has a recurrent single state, then there exists a sequence of regeneration points for both the marking process and the underlying chain. In practice, regeneration points for SPNs with nonexponential clock-setting distributions are almost always defined in terms of a single state.

3.3 Standardized Time Series

This section deals with methods for estimation of time-average limits when regenerative methods are not applicable. This situation can occur either because there is no apparent sequence of regeneration points or because regenerations occur so infrequently that the method is impractical.

Our results rest on a strong law of large numbers (SLLN) and "functional" central limit theorem (FCLT) for the marking process. The SLLN given below provides conditions (in the absence of regenerative structure) under which time-average limits are well defined.

Theorem 3 *Suppose that Assumption PD(1) holds. Then $\lim_{t\to\infty}(1/t)\int_0^t f\big(X(u)\big)\,du = r(f)$ a.s. for any real-valued function f defined on S, where $r(f)$ is a finite constant.*

Roughly speaking, an output process $\{f\big(X(t)\big): t \geq 0\}$ with time-average limit $r(f)$ obeys a FCLT if the associated cumulative (i.e., time-integrated) process, centered about the deterministic function $g(t) = r(f)t$ and suitably compressed in space and time, converges in distribution to a Brownian motion as the degree of compression increases. For a real-valued function f defined on S and a finite constant $r(f)$ such that the assertion of Theorem 3 holds, set

$$U_v(f)(t) = \frac{1}{\sqrt{v}} \int_0^{vt} \Big(f\big(X(u)\big) - r(f) \Big)\, du$$

for $t \geq 0$ and $v > 0$. Also denote by W a standard Brownian motion on $[0, \infty)$ and by $\Rightarrow$ weak convergence on $C[0, \infty)$, the space of real-valued continuous functions. Weak convergence on $C[0, \infty)$ generalizes to a sequence of real-valued random functions—that is, a sequence of real-valued stochastic processes—the usual notion of convergence in distribution of a sequence of real-valued random variables.

Theorem 4 *Suppose that Assumption PD(2) holds and let f be an arbitrary real-valued function defined on S. Then there exists a nonnegative number $\sigma^2(f)$ such that $U_v(f) \Rightarrow \sigma(f)W$ as $v \to \infty$ for any initial distribution μ.*

Fix a real-valued function f and suppose that Assumption PD(2) holds for the SPN of interest. By the foregoing SLLN, the time average limit $r(f)$ is well defined and finite. Moreover, a strongly consistent point estimator for $r(f)$ is given by $\hat{r}_v = \bar{Y}_v(1)$, where

$$\bar{Y}_v(t) = \frac{1}{v} \int_0^{vt} f\big(X(u)\big)\, du$$

for $0 \leq t \leq 1$ and $v > 0$. Standardized time series (STS) methods are concerned with obtaining asymptotic confidence intervals for $r(f)$. To this end, denote by $C[0, 1]$ the set of continuous real-valued functions defined on $[0, 1]$. For a mapping ξ from $C[0, 1]$ to $\Re$, let $D(\xi)$ be the set of discontinuity points for ξ. That is, $x \in D(\xi)$ if $\lim_{n\to\infty} \xi(x_n) \neq \xi(x)$ for some sequence $x_1, x_2, \ldots \in C[0, 1]$ with $\lim_{n\to\infty} \sup_{0\leq t\leq 1} |x_n(t) - x(t)| = 0$. Next, denote by Ξ the set of mappings from $C[0, 1]$ to $\Re$ such that $\xi \in \Xi$ if and only if (i) $\xi(ax) = a\xi(x)$ for $a \in \Re_+$ and $x \in C[0, 1]$. (ii) $\xi(x - be) = \xi(x)$ for $b \in \Re$ and $x \in C[0, 1]$, where $e(t) = t$ for $0 \leq t \leq 1$. (iii) $P\{\xi(W) > 0\} = 1$. (iv) $P\{W \in D(\xi)\} = 0$. It can be shown that the convergence asserted by Theorem 4 implies that, for $0 < p < 1$, the interval $\big[\hat{r}_v - \xi_v z_p, \hat{r}_v + \xi_v z_p\big]$ is an asymptotic $100p\%$

confidence interval for $r(f)$, where $\xi_v = \xi(\bar{Y}_v)$ and z_p is a constant such that $P\{-z_p \le W(1)/\xi(W) \le z_p\} = p$. Different choices of the mapping ξ lead to different estimation procedures, such as the standard method of batch means (with the number of batches independent of the simulation run length) or the original versions of the STS area method and STS maximum method.

The foregoing estimation methods can be extended to deal with time-average limits expressed in terms of the underlying chain, and can be combined with jackknifing techniques to handle nonlinear functions of time-average limits.

3.4 Consistent Estimation Methods

Consider an SPN with an underlying chain $\{(S_n, C_n): n \ge 0\}$ having state space Σ, together with a real-valued function $\tilde{f}$ defined on Σ, such that

$$\lim_{n\to\infty} \bar{r}(n; \tilde{f}) = \tilde{r}(\tilde{f}) \text{ a.s.} \tag{6}$$

for some finite constant $\tilde{r}(\tilde{f})$ and

$$\frac{\sqrt{n}\big(\bar{r}(n; \tilde{f}) - \tilde{r}(\tilde{f})\big)}{\tilde{\sigma}(\tilde{f})} \Rightarrow N(0, 1) \tag{7}$$

as $n \to \infty$ for some constant $\tilde{\sigma}(\tilde{f}) \in (0, \infty)$, where $\bar{r}(n; \tilde{f}) = (1/n)\sum_{j=0}^{n-1} \tilde{f}(S_j, C_j)$. Suppose that we can find an estimator V_n that is *consistent* for the variance constant $\tilde{\sigma}^2(\tilde{f})$ that appears in the central limit theorem (CLT) in (7), that is, an estimator V_n that converges to $\tilde{\sigma}^2(\tilde{f})$ in probability as $n \to \infty$ or, equivalently, $V_n \Rightarrow \tilde{\sigma}^2(\tilde{f})$ as $n \to \infty$. Then an application of Slutsky's theorem shows that

$$\frac{\sqrt{n}\big(\bar{r}(n; \tilde{f}) - \tilde{r}(\tilde{f})\big)}{V_n^{1/2}} \Rightarrow N(0, 1),$$

so that

$$\left[\bar{r}(n; \tilde{f}) - \frac{z_p\, V_n^{1/2}}{\sqrt{n}}, \bar{r}(n; \tilde{f}) + \frac{z_p\, V_n^{1/2}}{\sqrt{n}}\right]$$

is an asymptotic $100p\%$ confidence interval for $\tilde{r}(\tilde{f})$, where z_p is the $(1+p)/2$ quantile of the standard normal distribution. This section is concerned with methods for obtaining point estimates and confidence intervals based on consistent estimation of the variance constant. Note that the regenerative method is one such "consistent estimation method." Our emphasis in this section is on alternative methods that do not require regenerative structure. When applicable, consistent estimation methods yield confidence intervals whose lengths are, asymptotically, both smaller in expectation and less variable than the lengths of confidence intervals based on "cancellation" methods such as STS.

For brevity, we deal with time-average limits of the underlying chain $\{(S_n, C_n): n \ge 0\}$. As described in Section 7.3.5 of Haas (2002), our results can be extended to obtain confidence intervals for time-average limits of the marking process.

We assume throughout that Assumption PDE holds. It follows that a SLLN and CLT as in (6) and (7) hold for any "polynomially dominated" function $\tilde{f}$, that is, any function $\tilde{f}$ such that there exists $q \ge 0$ for which $\sup_{(s,c)\in\Sigma} |\tilde{f}(s, c)|/|\tilde{g}_q(s, c)| < \infty$, where

$$\tilde{g}_q(s, c) = \begin{cases} 1 + \max_{1\le i\le M} c_i^q & \text{if } (s, c) \in \Sigma^+; \\ 1 & \text{if } (s, c) \in \Sigma - \Sigma^+ \end{cases}$$

for $s \in G$ and $c = (c_1, c_2, \ldots, c_M) \in C(s)$. Heuristically, a function $\tilde{f}$ is polynomially dominated if $|\tilde{f}|$ is bounded above on Σ^+ by a polynomial function of the maximum clock reading and is bounded above on $\Sigma - \Sigma^+$ by a constant. For example, the holding-time function t^* is polynomially dominated.

We also focus on "aperiodic" SPNs, defined as follows. A *d-cycle* of an SPN is a finite partition $\{G_1, G_2, \ldots, G_d\}$ of G such that $s' \in G_{i+1}$ whenever $s \in G_i$ and $s \to s'$. (Take $G_{d+1} = G_1$.) The *period* of the SPN is the largest d for which a d-cycle exists; the SPN is called *aperiodic* if $d = 1$ and *periodic* (with period d) if $d > 1$.

We now seek conditions on the building blocks of an SPN under which various "quadratic-form" estimators of the variance constant $\tilde{\sigma}^2(\tilde{f})$ are consistent. By "quadratic-form" estimators, we mean estimators of the form

$$V_n = V_n(\tilde{f}) = \sum_{i=0}^{n}\sum_{j=0}^{n} \tilde{f}(S_i, C_i)\tilde{f}(S_j, C_j)q_{i,j}^{(n)},$$

where each $q_{i,j}^{(n)}$ is a finite constant and $q_{i,j}^{(n)} = q_{j,i}^{(n)}$ for all i, j. We further focus on the subclass of "localized" quadratic-form estimators. A quadratic-form estimator V_n is said to be *localized* if there exist a constant $a_1 \in (0, \infty)$ and sequences $\{a_2(n): n \ge 0\}$ and $\{m(n): n \ge 0\}$ of nonnegative constants with $a_2(n) \to 0$ and $m(n)/n \to 0$ such that

$$|q_{i,j}^{(n)}| \le \begin{cases} a_1/n & \text{if } |i - j| \le m(n); \\ a_2(n)/n & \text{if } |i - j| > m(n). \end{cases}$$

The class of localized quadratic-form estimators includes both variable-batch-means and spectral estimators.

Our strategy is to invoke known results that establish the consistency of various quadratic-form estimators for stationary processes and then apply the following result, which is based on a "coupling" argument.

Theorem 5 *Let $\{(S_n, C_n): n \geq 0\}$ be the underlying chain of an aperiodic* SPN*, and let $\tilde{f}$ be a polynomially dominated real-valued function defined on Σ. Suppose that Assumption PDE holds, so that there exists an invariant distribution π for the chain and $\{\tilde{f}(S_n, C_n): n \geq 0\}$ obeys a* CLT *with variance constant $\tilde{\sigma}^2(\tilde{f})$. If a localized quadratic-form estimator $V_n(\tilde{f})$ satisfies $V_n(\tilde{f}) \Rightarrow \tilde{\sigma}^2(\tilde{f})$ when the initial distribution is π, then $V_n(\tilde{f}) \Rightarrow \tilde{\sigma}^2(\tilde{f})$ for any initial distribution.*

For example, we can establish sufficient conditions for consistency of the variable batch means estimator of $\sigma^2(\tilde{f})$. Here both the number of batches $b = b(n)$ and the batch length $m = m(n)$ increase as the simulation run length n increases. The conditions are that Assumption PDE holds, that $\tilde{f}$ is polynomially dominated, and that $b(n) \to \infty$ and $m(n) \to \infty$ as $n \to \infty$.

We can similarly obtain sufficient conditions for the validity of spectral methods. The variance estimators have the form $V_n^{(S)} = \sum_{h=-(m-1)}^{m-1} \lambda(h/m)\hat{R}_h$, where $\hat{R}_h$ is the sample estimate of the autocorrelation function of $\{\tilde{f}(S_n, C_n): n \geq 0\}$ at lag h and λ is a "window function." For a large class of window functions, it can be shown that $V_n^{(S)}$ is consistent provided that Assumption PDE holds, that $\tilde{f}$ is polynomially dominated, and that the spectral window length $m = m(n)$ satisfies $m(n) \to \infty$ and $m^2(n)/n \to 0$.

4 DELAYS

We now consider performance measures of the form $\lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} f(D_j)$, where f is a real-valued function and $D_0, D_1, \ldots$ is a sequence of delays determined by the marking changes of the net.

A delay in an SPN is computed as the length of a corresponding "delay interval"—that is, a random time interval—whose start (left endpoint) and termination (right endpoint) each coincide with a marking-change epoch. Sometimes the *limiting average delay* $\lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} D_j$ can be estimated indirectly, that is, without measuring lengths of individual delay intervals. For general time-average limits of a sequence of delays, however, individual lengths must be measured and then combined to form point and interval estimates. Specification and subsequent measurement of individual delays is a decidedly nontrivial step of the simulation: in general, there can be more than one ongoing delay at a time point and delays need not terminate in the order in which they start. In the following, we describe the method of "start vectors" for specification and measurement of delays, and provide conditions on the SPN and start-vector building blocks under which time-average limits exist and various simulation-based estimation methods are applicable.

4.1 Specification and Measurement of Delays

A sequence of delays in an SPN is specified in terms of *starts* $\{A_j: j \geq 0\}$ and *terminations* $\{B_j: j \geq 0\}$, i.e., $D_j = B_j - A_j$ for $j \geq 0$. The A_j's and B_j's are defined on the same probability space as the underlying chain $\{(S_n, C_n): n \geq 0\}$. We restrict attention to sequences $\{A_j: j \geq 0\}$ and $\{B_j: j \geq 0\}$ such that $A_j = \zeta_{\alpha(j)}$ and $B_j = \zeta_{\beta(j)}$ for $j \geq 0$, where $\alpha(j)$ and $\beta(j)$ are a.s. finite random indices. That is, we restrict attention to delays that start and terminate only at marking changes. We also focus on sequences for which the $\alpha(j)$'s are nondecreasing, so that delays are enumerated in start order. The $\beta(j)$'s need not be nondecreasing, however, reflecting the fact that there can be more than one ongoing delay at a time point and delays need not terminate in the order in which they start.

One approach to specification and measurement of delays "tags" various entities (such as jobs or customers) as they move through the system. The disadvantage of this approach is that it often requires either distinguishable tokens or a large number of additional places and transitions.

We now give an alternative method for specifying and measuring delays that avoids the need for tagging. The idea is to use a sequence of real-valued random vectors, called *start vectors*, to construct the sequences $\{A_j: j \geq 0\}$ and $\{B_j: j \geq 0\}$. The sequence $\{V_n: n \geq 0\}$ of start vectors is determined by the sample paths of the chain $\{(S_n, C_n): n \geq 0\}$ and provides the link between the starts and terminations of the individual delay intervals. The nth start vector V_n records the starts of delay intervals for all ongoing delays and newly started delays at time ζ_n, that is, all starts $A_j = \zeta_{\alpha(j)}$ such that $\alpha(j) \leq n < \beta(j)$. Usually (but not necessarily) the positions of the starts in the start vector correspond to the locations in the system of entities for whom a delay is underway. In general, the values of the starts and the order of the starts in the start vector together summarize the history of the net and comprise sufficient information to measure individual delays. Some components of V_n may be equal to -1. As discussed below, lengths are never computed for delay intervals with negative starts. The negative components typically serve as placeholders and correspond to entities in the system at time 0 for whom no delay is underway.

Whenever the transitions in the set E^* fire simultaneously and trigger a marking change from s to s', a new start vector is obtained from the current start vector by (i) inserting the current time at zero or more positions specified by an index vector $i_\alpha(s'; s, E^*)$, (ii) deleting components at zero or more positions specified by an index vector $i_\beta(s'; s, E^*)$, and (iii) permuting the components according to an index vector $i_\pi(s'; s, E^*)$. If the index value j (≥ 0) appears as an element of i_α, then the current time is inserted to the right of position j in the start vector V; insertion to the right of position 0 means insertion to the left of the leftmost

element of V. If $i_\pi = (i_1, i_2, \ldots, i_k)$, then the permuted start vector is $(v_{i_1}, v_{i_2}, \ldots, v_{i_k})$. Components are deleted one at a time in the order in which the indices appear in the vector $i_\beta(s'; s, E^*)$. For each nonnegative component that is deleted, the length of a delay interval is computed by subtracting the deleted component from the current time. These deleted components are the left endpoints of delay intervals for the delays that terminate at the current time. Deleted components equal to -1 are not used to compute lengths of delay intervals and are simply discarded. We assume that the current marking determines the length of the start vector and denote this length by $\psi(s)$ when the current marking is s. The initial start vector is a specified vector, denoted $v_0(S_0)$, that is determined by the initial marking S_0 and has components that are equal to 0 or -1. See Section 8.1.2 in Haas (2002) for further details of the start-vector mechanism.

Denote by $n_\alpha(s'; s, E^*)$ and $n_\beta(s'; s, E^*)$ the lengths of the vectors $i_\alpha(s'; s, E^*)$ and $i_\beta(s'; s, E^*)$, respectively, for each s', s, and E^*. The number of delays that start at time ζ_n is equal to $n_\alpha(S_n; S_{n-1}, E^*_{n-1})$ for $n \geq 1$. Denote by $V_{n,i}$ the ith component of the vector V_n for $1 \leq i \leq \psi(S_n)$, and set

$$K = \inf\left\{ n \geq 0 \colon V_{n,i} \neq -1 \text{ for } 0 \leq i \leq \psi(S_n) \right\}. \quad (8)$$

The number of delays that terminate at time ζ_n is less than or equal to $n_\beta(S_n; S_{n-1}, E^*_{n-1})$ for $1 \leq n \leq K$ and equal to $n_\beta(S_n; S_{n-1}, E^*_{n-1})$ for $n > K$. Similarly, the total number of newly started delays (of positive duration) and ongoing delays at the nth marking change is less than or equal to $\psi(S_n)$ for $0 \leq n < K$ and equal to $\psi(S_n)$ for $n \geq K$.

Example 2 (Cyclic queues with feedback) Consider the delay intervals from whenever a job completes service at center 2 to when the job next completes service at center 2, and suppose that we wish to estimate time-average limits of the sequence of delays for all N jobs. The method of start vectors can be used to specify and measure individual delays in the SPN of Figure 4.

The start vector V_n records for each of the N jobs in the network the most recent time during the interval $[0, \zeta_n]$ at which there was a completion of service at center 2 and the job moved to center 1. If a job has never moved from center 2 to center 1 during the interval $[0, \zeta_n]$, then the corresponding component of V_n is equal to -1. The components of the start vector are ordered from left to right according to increasing positions—see Figure 2—of the corresponding jobs in the network.

Formally, set $\psi(s) = N$ for $s \in G$. Also set $i_\alpha(s'; s, E^*)$ equal to (0) if $E^* = \{ e_2 \}$ and equal to $\varnothing$ otherwise, and set $i_\beta(s'; s, E^*)$ equal to $(N+1)$ if $E^* = \{ e_2 \}$ and equal to $\varnothing$ otherwise. Thus, whenever there is a completion of service at center 2 and a job moves to the tail of the queue at center 1, the new start vector is obtained from the current start vector by inserting the current time to the left of the first component, deleting the rightmost component, and then subtracting the latter component from the current time to compute a delay if the component is nonnegative. Next, for $s = (s_1, s_2)$, $s' = (s'_1, s'_2) \in G$ and $E^* \subseteq E(s)$, set $i_\pi(s'; s, E^*) = (s_1, 1, 2, \ldots, s_1 - 1, s_1 + 1, s_1 + 2, \ldots, N)$ if $E^* = \{ e_1 \}$ and $s'_1 = s_1 > 1$. Otherwise, set $i_\pi(s'; s, E^*) = \varnothing$, so that no permutation is performed. Thus, whenever there are s_1 (> 1) jobs at center 1 and a job completes service at center 1 and joins the tail of the queue at center 1, the new start vector is obtained from the current start vector by cyclically permuting the first s_1 components. Otherwise, the components are unchanged—in particular, no permutation is needed when $E^* = \{ e_2 \}$.

Suppose that at time 0 there is a completion of service at center 2 with all jobs at center 2, so that the initial marking is $s_0 = (1, N - 1)$ and a delay starts at time 0. We then set $v_0(s_0) = (0, -1, -1, \ldots, -1)$, where the vector on the right side is of length N. Because $N - 1$ components of $v_0(s_0)$ are equal to -1, there are $N - 1$ marking changes at which there is a completion of service at center 2 and no delay is computed. At the time ζ of each such marking change, the job completing service at center 2 has not previously completed service at center 2 during the interval $[0, \zeta]$ and ζ is not an element of the sequence $\left\{ B_j \colon j \geq 0 \right\}$ of terminations. □

4.2 Regenerative Methods for Delays

In this section we provide methods for estimating general time-average limits of the form $\lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} f(D_j)$, where the sequence of delays $\left\{ D_j \colon j \geq 0 \right\}$ is determined from the marking changes of an SPN by means of start vectors. We also provide specialized estimation methods in this setting for the limiting average delay $\lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} D_j$.

Our key assumption is that there exists a sequence of regeneration points for the marking process $\{ X(t) \colon t \geq 0 \}$ and for the underlying chain $\{ (S_n, C_n) \colon n \geq 0 \}$. In particular, we suppose throughout that there exists a recurrent single state $\bar{s}$, so that $E(\bar{s}) = \{ \bar{e} \}$ for some $\bar{e} \in E$ and $P_\mu \{ S_n = \bar{s} \text{ i.o.} \} = 1$. The regeneration points then correspond to the successive times at which the marking is $\bar{s}$ and transition $\bar{e}$ fires. That is, if we define $\{ \theta(k) \colon k \geq 0 \}$ as in (5), then the random indices $\{ \theta(k) \colon k \geq 0 \}$ form a sequence of regeneration points for $\{ (S_n, C_n) \colon n \geq 0 \}$ and the random times $\{ \zeta_{\theta(k)} \colon k \geq 0 \}$ form a sequence of regeneration points for $\{ X(t) \colon t \geq 0 \}$. Implicit in this definition is the assumption—made for convenience—that the net behaves as if at time 0 the marking is $\bar{s}$ and transition $\bar{e}$ fires. We also suppose that the starts $\left\{ A_j \colon j \geq 0 \right\}$, the terminations $\left\{ B_j \colon j \geq 0 \right\}$, and the random index K that is defined by (8) satisfy $P_\mu \{ K < \infty \} = 1$,

$P_\mu\{A_j < \infty\} = P_\mu\{B_j < \infty\} = 1$ for $j \geq 0$, and $P_\mu\{\lim_{j\to\infty} A_j = \infty\} = 1$.

When there are no ongoing delays at any regeneration point for the marking process—see Figure 5—it is intuitively clear that the regeneration points decompose the delays into i.i.d. blocks. The sequence of delays therefore is a regenerative process in discrete time, and we can estimate time-average limits using methods as in Section 3.2. This scenario holds, for example, whenever $\psi(\bar{s}) = 0$ or whenever all delays are of positive length and $n_\beta(s; \bar{s}, \{\bar{e}\}) = \psi(\bar{s})$ for all s such that $p(s; \bar{s}, \{\bar{e}\}) > 0$.

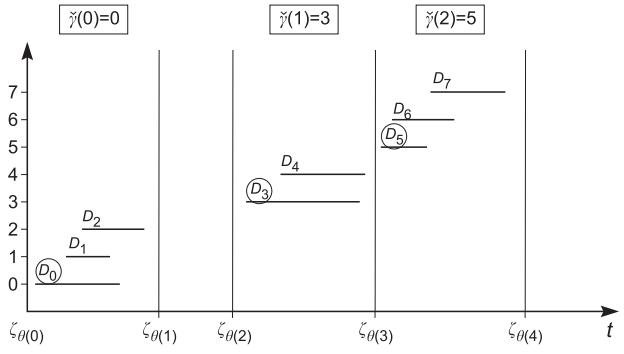

Figure 5: Regenerative Cycles for Delays

The situation is not so simple, however, when there are ongoing delays at each regeneration point. Our approach is to first select a random subsequence $\{\zeta_{\breve{\theta}(k)} : k \geq 0\}$ of the original regeneration points $\{\zeta_{\theta(k)} : k \geq 0\}$ for the marking process, as shown in Figure 6. In the figure, vertical solid lines indicate regeneration points corresponding to this subsequence, whereas vertical dashed lines indicate regeneration points that are in the original sequence but are not in the subsequence. The regeneration points in the subsequence $\{\zeta_{\breve{\theta}(k)} : k \geq 0\}$ also form a sequence of regeneration points, but with longer cycles. The subsequence is chosen such that all delays that start during one of these longer cycles terminate by the end of the next such cycle. Denote by $\breve{\gamma}(k)$ the index of the first delay to start after time $\zeta_{\breve{\theta}(k)}$. It can be shown that the random indices $\{\breve{\gamma}(k) : k \geq 0\}$ decompose sample paths of the process $\{D_n : n \geq 0\}$ into identically distributed, one-dependent cycles.

One of two methods can then be used to estimate time-average limits. The *extended regenerative method* is almost identical to the standard regenerative method, but replaces the usual variance constant $\breve{\sigma}^2(f)$ by a modified constant that takes into account the dependence between adjacent cycles. The *multiple runs method* simulates the first cycle to compute $\breve{Y}_1(f) = \sum_{j=\breve{\gamma}(0)}^{\breve{\gamma}(1)-1} f(D_j)$ and the cycle length $\breve{\delta}_1 = \breve{\gamma}(1) - \breve{\gamma}(0)$. The simulation is then restarted (with different random number seeds), and the procedure repeated. Multiple iterations result in a sequence

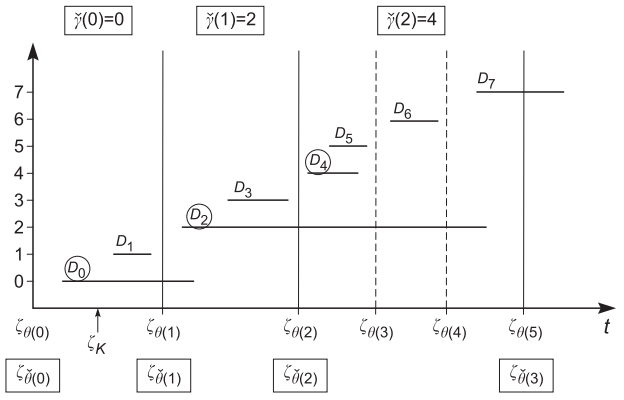

Figure 6: Definition of One-Dependent Cycles

$\left(\breve{Y}_{1,1}(f), \breve{\delta}_{1,1}\right), \left(\breve{Y}_{1,2}(f), \breve{\delta}_{1,2}\right), \ldots$ of i.i.d. pairs, and the standard regenerative method is applied to these pairs. Neither of these two methods dominates the other in terms of simulation efficiency; see Section 8.2.3 in Haas (2002).

For the special case of the limiting average delay, it is shown in Section 8.2.4 of Haas (2002) that, under appropriate regularity conditions,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} D_j = \frac{E_\mu[Z_1]}{E_\mu[\delta_1]} \text{ a.s.},$$

where $\delta_k = \sum_{n=\theta(k-1)}^{\theta(k)-1} n_\alpha(S_n; S_{n-1}, E^*_{n-1})$ and $Z_k = \int_{\zeta_{\theta(k-1)}}^{\zeta_{\theta(k)}} \psi\big(X(t)\big)\, dt$ for $k \geq 1$. It follows that the standard regenerative method can be used to obtain strongly consistent point estimates and asymptotic confidence intervals for the limiting average delay. Measurement of individual delays is not required, and there may be ongoing delays at each regeneration point of the marking process.

4.3 STS Methods for Delays

We conclude by giving conditions under which STS methods can be used to obtain point estimates and confidence intervals for time-average limits of a sequence of delays; see Haas (2002, 2003) for details. We require that the start-vector mechanism be *regular* in that (i) there exists $s \in S$ and $e^* \in E(s)$ such that $n_\alpha(s'; s, \{e^*\}) > 0$ for all s' with $p(s'; s, \{e^*\}) > 0$, and (ii) there exists a potential finite sequence of marking changes such that (a) each marking change is triggered by the firing of a single transition, and (b) all the components of the start vector at the beginning of the sequence are deleted by the end of the sequence. In analogy to our earlier definition, we define $f : \Re_+ \mapsto \Re$ to

be *polynomially dominated* if $\sup_x |f(x)|/(x^b + 1) < \infty$ for some $b \geq 0$.

Theorem 6 *Let* $\{ D_j : j \geq 0 \}$ *be a sequence of delays determined from the underlying chain of a marking process by means of a regular start-vector mechanism, and let* f *be a polynomially dominated function. Suppose that Assumption PDE holds. Then (i) there exists a finite real constant* $r(f)$ *such that* $\lim_{n\to\infty}(1/n)\sum_{j=0}^{n-1} f(D_j) = r(f)$ *a.s., and (ii) there exists a nonnegative number* $\sigma^2(f)$ *such that* $U_n(f) \Rightarrow \sigma(f)W$ *as* $n \to \infty$ *for any initial distribution* μ*, where* $\Rightarrow$ *denotes weak convergence on* $C[0,\infty]$*,* W *is a standard Brownian motion, and* $U_n(f)(t) = (1/\sqrt{n})\int_0^{nt}\big(f(D_{\lfloor u\rfloor}) - r(f)\big)\,du$ *for* $0 \leq t \leq 1$ *and* $n \geq 0$.

Under the conditions of Theorem 6, time-average limits are well defined. Just as in Section 3.3, it follows directly from Theorem 6 that STS methods can be used to obtain asymptotic confidence intervals for such limits.

REFERENCES

Glynn, P. W., and P. J. Haas. 2004. On functional central limit theorems for semi-Markov and related processes. *Communications in Statistics – Theory and Methods* 33:487–506.

Glynn, P. W., and P. J. Haas. 2005. A law of large numbers and functional central limit theorem for generalized semi-Markov processes. *Communications in Statistics – Stochastic Models*. To appear.

Haas, P. J. 2002. *Stochastic Petri nets: Modelling, stability, simulation*. New York: Springer-Verlag.

Haas, P. J. 2003. Estimation methods for delays in non-regenerative discrete-event systems. *Communications in Statistics – Stochastic Models* 19:1–35.

AUTHOR BIOGRAPHY

PETER J. HAAS has been a Research Staff Member at the IBM Almaden Research Center since 1987 and is also a Consulting Associate Professor in the Department of Management Science and Engineering at Stanford University. He is an Associate Editor (Simulation Area) for *Operations Research* and has recently joined the editorial board of *ACM TOMACS*. In 2003, his monograph, *Stochastic Petri Nets: Modelling, Stability, Simulation* won the Outstanding Simulation Publication Award from the INFORMS College on Simulation. He is a member of INFORMS.