Chapter 9

Stochastic Petri Nets

Serge Haddad , Patrice Moreaux[1]

1. Introduction

One of the main interests of Petri nets is to combine qualitative analysis (*i.e.* the property verification) and quantitative one (*i.e.* performance evaluation) [FLO 78, MOL 81, REI 98a, REI 98b]. By comparison, concurrency models like process algebra [HIL 96] have only recently been extended with stochastic features and if first results are promising, there are still more research about performance evaluation of stochastic Petri nets. Similarly, the usual model for performance evaluation like queueing networds [KLE 75] do not include synchronization mechanisms and adding them by ad hoc constructions [FDI 86, DAL 97] do not reach the generality and the simplicity of concurrency modeling by Petri nets.

Stochastic Petri nets have been introduced in a pragmatic way at the end of the seventies, in order to take benefit from the evaluation methods of Markov chains. This approach leads to immediate results but occults the semantical features underlying the definition of stochastic Petri nets and cannot be easily generalized to different probability distributions. So along the three chapters devoted to stochastic Petri nets, we proceed as follows. First we tackle the semantical level *i.e.* the level of the stochastic processes and we study their properties with the aim to design analysis methods. Here we have chosen to

[1]LSV, ENS Cachan, 61, avenue du Président Wilson 94235 Cachan Cedex France, (haddad@lsv.ens-cachan.fr)
LISTIC - BP 80439 - F74944 Annecy le Vieux Cedex (patrice.moreaux@univ-savoie.fr)

emphasize the principles which characterize every method. So we omit the programming features related to numerical computations. Indeed, these features are not specific to stochastic Petri nets and are covered by excellent books [STE 94, BOL 98]. Then we present models (and extensions) of stochastic Petri nets which generate such stochastic processes. Thus the constraints that appear in the definition have an intuitive explanation. At last, we develop the analysis methods at the stochastic Petri net level, which points out the links with the qualitative methods described in other parts of the book.

This chapter begins by the characterization of stochastic processes associated with discrete event systems. We describe the family of random variables of these processes and interpret them w.r.t. a realization of an execution. Then we briefly recall renewing theory which ensures under weak conditions, the existence of a stationary distribution of the discrete event system. Indeed, we mainly cover the study of systems on the long run. Then we enumerate by increasing complexity order typical processes for which the renewing theory can be applied, beginning by Markov chains.

Afterwards we develop the key points of a stochastic semantic for Petri nets. This includes the specification of a random variable associated with the firing delay of a transition, the choice criteria between enabled transitions, the handling of the firing degree in the samplings of the random variable associated with a transition and the memorization of the previous samplings, once the firing is performed. Then we restrict the type of distributions, which leads to stochastic processes previously studied. Among the different families of stochastic nets, Petri nets with exponential and immediate distributions, called generalized stochastic Petri nets, are considered as the standard model [Ajm 95]. We indicate, for every family, how to to compute the stationary distribution based on the reachability graph (when it is finite).

The basic algorithms have a complexity of the same magnitude order as the reachability graph size for the simple models and greater for models with more general distributions. Thus the more elaborated techniques split in two families: the first one aims at obtaining a complexity smaller than the size of the graph (*e.g.* by restricting the class of Petri nets) and the second one aims at obtaining the same order of complexity than the size of the graph but for extended models.

The last section describes some of these methods in order to cover the diversity of the approaches. Two other methods and applicable to well formed nets are presented in the next chapters. Those covered in this section are:

- the research of a product form: a formula that expresses the stationary probability of a marking including the net parameters and the place marking as variables of the formula. This method illustrates the extension of

a technique first applied in queueing networks.

- the research of bounds (*e.g.* on the rates) using the structure of the net. Here one observes that linear programming eases the quantitative and qualitative analysis of Petri nets.
- approximation methods that take advantage of a net decomposition or of a transformation of the stochastic process.
- a resolution method for nets whose a single place is unbounded. The application of this method show that conditions on the structure of Markov chains can be naturally translated in terms of Petri nets.

2. A stochastic semantic for the discrete event systems

2.1. *The stochastic model*

We assume that the bases of probability theory are known by the reader [FEL 68, FEL 71, TRI 82].

Notations:

- $\Pr(E)$ denotes the probability of event E and $\Pr(A \mid B)$ the probability of A knowing B.
- $\mathbb{R}$ (resp. $\mathbb{R}^+$) denotes reals (resp. non negative reals)
- A measure on $\mathbb{R}$ is given by a function F, increasing, right continuous such that $lim_{x \to -\infty} F(x) = 0$. $F(x)$ represents the measure of interval $]-\infty, x]$. The mass of the measure (finite or not) is $lim_{x \to \infty} F(x)$
- A distribution is a measure with mass 1.
- Usual integration is denoted ds where s is the integration variable. Integration w.r.t. a measure F is denoted $F\{ds\}$.
- The word *almost*, in expressions like *almost everywhere* or *almost surely*, means "for a set of probability 1".

An execution of a discrete event system (DES) is characterized by a sequence of events $\{e_1, \ldots, e_n, \ldots,\}$ (sequence supposed to be infinite) occurring after time delays. Only events can change the state of the system.

Formally, the stochastic behavior of a DES is determined by two families of random variables:

- $X_0, \ldots, X_n, \ldots$ taking their values in the (discrete) state of the system $\{s_1, \ldots, s_k, \ldots\}$. X_0 represents the initial distribution of the system and X_n $(n > 0)$ the distribution after the n^{th} event. The occurrence of an event does not necessarily modify the state of the system, consequently X_{n+1} may be equal to X_n.
- $T_0, ..., T_n, ...$ taking their values in $\mathbb{R}^+$ where T_0 represents the time interval before the first event and T_n, $(n > 0)$ represents the time interval between the n^{th} and $(n+1)^{th}$ event. Observe that this interval can be null (*e.g.* a sequence of instructions considered as instantaneous compared to database transactions including inputs/outputs).

A priori, no restriction should be required for these families of random variables. However to avoid the pathologic character of some executions, we exclude the possibility for a DES to execute an infinite number of actions in finite time. Otherwise stated, we establish sufficient conditions in order to fulfill the following equation:

$$\sum_{n=0}^{\infty} T_n = \infty \; almost\; surely \qquad [1]$$

This restriction enables us to define the state of the system at any instant. Let $N(t)$, be the random variable defined by:

$$N(t) = \inf\{n \text{ such that } \sum_{k=0}^{n} T_k > t\}$$

Using equation 1, $N(t)$ is defined *almost everywhere*. As can be seen in figure 1, $N(t)$ presents jumps with amplitude greater than 1. The state $Y(t)$ of the system at time t, is $X_{N(t)}$. Observe that $Y(t)$ is not equivalent to the stochastic process, but that it allows, in most of the cases, to proceed to standard analyses. The scheme of figure 1 presents a possible execution of the process and illustrates the interpretation of the random variables previously introduced. In this example, the process is initially in state s_4 and remains in it until t_0 when it moves to state s_6. At time $t_0 + t_1$, the system successively visits in a null time, states s_3 and s_{12} before reaching state s_7 where it sojourns some time. The observation $Y(t)$ in continuous time occults the vanishing states s_3 and s_{12} of the process.

2.2. *Analysis with renewing theory*

The performance evaluation of a DES leads to two kinds of analysis:

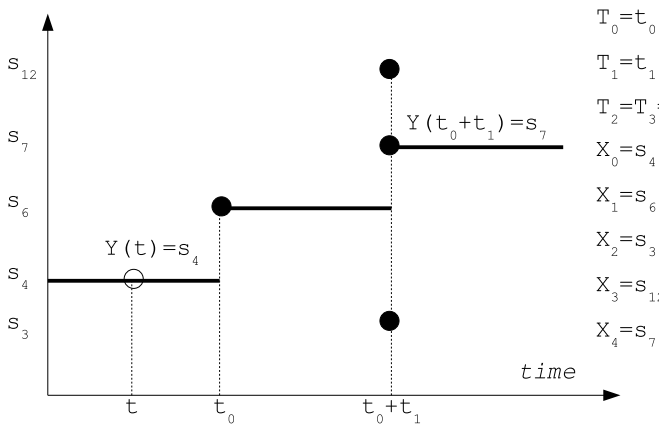

Figure 1: A realization of the stochastic process

- The study of the transient behavior, *i.e.* the computation of measures for performance indices depending on elapsed time since the initial state. This study covers the initializing stage of systems and terminating systems. Among application areas, one can cite dependability and safety analyses [LAP 95, MEY 80, TRI 92].
- The study of the stationary behavior. For numerous applications, what interests the modeler is the behavior of the system once the initial stage is left and that it stabilizes.

This supposes that such a stationary behavior exists. Which can be resumed, denoting $\boldsymbol{\pi}(t)$ the distribution of $Y(t)$, by:

$$\lim_{t\to\infty} \boldsymbol{\pi}(t) = \boldsymbol{\pi} \qquad [2]$$

where $\boldsymbol{\pi}$ is also a distribution, called stationary distribution . In this case, one calls the process an *ergodic process*. A sufficient condition for this asymptotic behavior is the existence of a repetitive phenomenon corresponding to some event occurrences such that the process identically behaves after every such occurrence.

Definition 1 *A stochastic process is a renewing process if there exists a family of random variables:* $I_1, \ldots, I_k, \ldots$ *(defined almost everywhere) taking values in* $\mathbb{N}$ *such that:*

- $I_k < I_{k+1}$
- $\forall k, k',\ \{X_{I_k+n}, T_{I_k+n}\}_{n\in\mathbb{N}}$ *and* $\{X_{I_{k'}+n}, T_{I_{k'}+n}\}_{n\in\mathbb{N}}$ *are probabilistic replicates of a fixed process.*

A renewing process is fully determined by its behavior between two renewing instants. Let us call:

- F the distribution of time between two renewing instants
- d the mean, supposed to be finite, of this distribution
- $p_k(t)$ the probability that, t time units after a renewing instant, a new renewing instant has not occurred and that the process is in state s_k

Observe that $1 - F(t) = \sum_k p_k(t)$. Hence family $\{p_k\}_{k\in\mathbb{N}}$ determines the stochastic process.

We must distinguish two cases depending on the type of distribution F since some distributions lead to periodic behaviors. Let us pick an elementary DES (a semi-Markovian process with deterministic delays, see later on), visiting three states s_1, s_2, s_3. We consider entrance in s_1 as a renewing instant. The DES sojourns 1 t.u. in state s_1 then moves to state s_2 where it remains during 2 t.u.. With probability 1/2, it returns in state s_1 or moves to state s_3 where it remains 3 t.u., before returning to s_1.

Assume that the process starts in state s_1. Then every $3n$ t.u., the process either in s_1, or in s_3 and every $(3n+1)$ t.u. the process is either in s_2 or in s_3. So there is no stationary distribution. These periodic behaviors are generated by arithmetic distributions.

Definition 2 *A distribution is arithmetic if it is concentrated on points* $\{n.\tau\}_{n\in\mathbb{N}}$ *for some* τ*. The period of an arithmetic distribution is the greatest* τ *fulfilling this property.*

Below is the main result about the existence of a stationary behavior.

Theorem 3 ([FEL 68, FEL 71]) *Given a renewing process defined by* $\{p_k\}_{k\in\mathbb{N}}$*. If* F *is not arithmetic then,*

$$\lim_{t\to\infty} \boldsymbol{\pi}(t)[s_k] = \frac{1}{d}.\int_0^\infty p_k(t)dt$$

If F *is arithmetic with period* τ *and if the process starts at a renewing instant,*

$$\lim_{n\to\infty} \boldsymbol{\pi}(t+n.\tau)[s_k] = \frac{\tau}{d}.\sum_{i=0}^{\infty} p_k(t+i.\tau)$$

If these two formulas are relatively simple, their application requires to know more about the stochastic process as will be seen in the next sections. These formula are generalizable to every performance index independent of the behavior of the process before the last renewing instant (*e.g.* the elapsed time since the last renewing instant).

2.3. *Discrete Time Markov chain*

2.3.1. *Presentation*

A discrete time Markov (*Discrete Time Markov Chain*, DTMC) has the following characteristics:

- The time interval between instants T_n is the constant 1.
- The next state following the current state only depends on this state and the transition probabilities are constant over the time:

$$\Pr(X_{n+1} = s_j \mid X_0 = s_{i_0}, ..., X_n = s_i) =$$

$$\Pr(X_{n+1} = s_j \mid X_n = s_i) = p_{ij} = \mathbf{P}[i, j]$$

  We will use both notations for state transitions.

The process is characterized by its initial distribution $\boldsymbol{\pi}_0$ and matrix $\mathbf{P}$. Let $\boldsymbol{\pi}_n$ the distribution de X_n, then $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0.\mathbf{P}^n$

2.3.2. *Conditions for a stationary distribution*

It is clear that every entrance in some given state constitutes a renewing instant. However it remains two conditions to check:

- There must be *almost surely* an infinite number of renewing instants.
- The mean time between two renewing instants must be finite.

This leads to a classification of states:

- A state is *transient* if the probability of return is less than 1. Such a state cannot constitutes a renewing process since the number of returns is *almost surely* finite. For obvious reasons, its occurrence probability goes to 0. A state is called *recurrent* if it is not transient.

- A state is *null recurrent* if the mean time for return is infinite. This state cannot ensure the existence of a stationary distribution. Intuitively, once reached, this state will occur after intervals whose mean length will go to infinity and consequently its occurrence probability goes to 0. This intuitive reasoning is mathematically sound.
- A state is *non null recurrent* if the mean time for return is finite. It is then sufficient that the state can be reached *almost surely* from the initial distribution in order to ensure the existence of a stationary distribution.

Let us detail this point and consider the graph (possibly infinite) built as follows:

- The set of vertices is the set of states.
- There is an arc from s_i to s_j if $p_{ij} > 0$.

Let us study the strongly connected components (s.c.c.) of this graph. If a s.c.c. has an exit arc, then necessarily, the states of this s.c.c. are transient. If there exists two terminal s.c.c. (*i.e.* without exit arcs) then the stationary distribution depends on the probability to reach them. Consequently, the independence of the stationary distribution from the initial distributions requires a single terminal s.c.c. reachable *almost surely*.

One calls *irreducible* chain, a terminal s.c.c. In a irreducible chain all states are of the same kind. Let us examine the irreducible chain defined by:

$$\forall i, p_{i\,i+1} = 1 - e_i \quad \text{and} \quad p_{i\,1} = e_i, \quad \text{with } 0 < e_i < 1$$

Then Pr(to not return in s_1) $= \prod_{i=1}^{\infty}(1 - e_i)$. A logarithmic transformation shows that this probability is non null iff $\sum e_i$ is convergent. Assume that states are recurrent; the mean time for a return in s_1 is equal to $1 + \sum_{k=1}^{\infty} \prod_{i=1}^{k} e_i$. Thus states are null recurrent if this sum is divergent and non null recurrent otherwise.

Classification criteria exist (see below). In a finite graph, the existence and unicity (whatever the initial distribution) of a stationary distribution is ensured as soon as there is a single terminal s.c.c.

In a discrete time Markov chain, the distribution of return to a state is arithmetic and its period is a multiple of 1. However since the sojourn in a state lasts at least 1 t.u., if the period is 1 the arithmetic case of theorem 3 is reducible to the general case and there is a stationary distribution. One calls such states *ergodic* and the chain is said *aperiodic*. If the period (k) is greater than 1, one can partition the states into subsets $S_0, S_1, \ldots, S_{k-1}$ such that from states of S_i one reaches states of $S_{(i+1) \bmod k}$. If one considers the state changes

every k t.u. (transition matrix $\mathbf{P}^k$) one obtain k independent chains over states S_i with period 1.

2.3.3. *Computation of the stationary distribution*

Once the existence of the stationary distribution is ensured, the computation is relatively easy. Indeed, one has $\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n.\mathbf{P}$. Taking the limits (which is sound here), one obtains $\boldsymbol{\pi} = \boldsymbol{\pi}.\mathbf{P}$. Furthermore if the chain is aperiodic then $\boldsymbol{\pi}$ is the single distribution of:

$$\mathbf{X} = \mathbf{X}.\mathbf{P} \qquad [3]$$

and the existence of a solution which is a distribution ensures that the irreducible chain is ergodic.

In the finite case, in order to solve equation [3], one can perform to a direct computation by adding the normalization equation $\mathbf{X}.\mathbf{1}^T = 1$ où $\mathbf{1}^T$ denotes the column vector whose every component is 1. But iterative computations are more interesting, the simplest consisting in iterating $\mathbf{X}_{n+1} \leftarrow \mathbf{X}_n.\mathbf{P}$[STE 94].

2.4. *Continuous time Markov chains*

2.4.1. *Presentation*

A continuous time Markov chain (CMTC) has the following characteristics:

- The time interval between instants T_n is a negative exponential random variable whose rate depends only on state X_n. Otherwise stated

$$\Pr(T_n \leq t \,|\, X_0 = s_{i_0}, ..., X_n = s_i, T_0 \leq t_0, ..., T_{n-1} \leq t_{n-1}) =$$

$$\Pr(T_n \leq t \,|\, X_n = s_i) = 1 - e^{\lambda_i.t}$$

- The state following the current state depends only on this state and transition probabilities are constant over time:

$$\Pr(X_{n+1} = s_j \,|\, X_0 = s_{i_0}, ..., X_n = s_i, T_0 \leq t_0, ..., T_{n-1} \leq t_{n-1}) =$$

$$\Pr(X_{n+1} = s_j \,|\, X_n = s_i) = p_{ij} = \mathbf{P}[i, j]$$

In continuous time Markov chains, due to the lack of memory of the exponential distribution, the evolution of the DES depends only the current state.

Contrary to discrete time chains, the stochastic process may present the pathologic behavior discussed at the beginning of the chapter. This behavior is excluded if, for instance, there exists a finite upper bound to the set of λ_i or if the discrete chain defined by matrix $\mathbf{P}$ and called *embedded chain*, is irreducible and recurrent.

The process is characterized by its initial distribution $\boldsymbol{\pi}(0)$, matrix $\mathbf{P}$ and the λ_i's. let us call $\boldsymbol{\pi}(t)$ the distribution of Y_t and $\pi_k(t) = \boldsymbol{\pi}(t)[s_k]$. If δ is small, between t and $t+\delta$ the probability of occurrence of more than an event is negligible and the probability of occurrence of a state change from k to k' is approximatively equal to $\lambda_k.\delta.p_{kk'}$

$$\pi_k(t+\delta) \approx \pi_k(t).(1-\lambda_k.\delta) + \sum_{k' \neq k} \pi_{k'}(t).\lambda_{k'}.\delta.p_{k'k}$$

Consequently

$$\frac{\pi_k(t+\delta) - \pi_k(t)}{\delta} \approx \pi_k(t).(-\lambda_k) + \sum_{k' \neq k} \pi_{k'}(t).\lambda_{k'}.p_{k'k}$$

And finally:

$$\frac{d\pi_k}{dt} = \pi_k(t).(-\lambda_k) + \sum_{k' \neq k} \pi_{k'}(t).\lambda_{k'}.p_{k'k}$$

Let us define matrix $\mathbf{Q}$ by: $q_{kk'} = \lambda_k.p_{kk'}$ for $k \neq k'$ and $q_{kk} = -\lambda_k (= -\sum_{k' \neq k} q_{kk'})$. Then the previous equation can written:

$$\frac{d\boldsymbol{\pi}}{dt} = \boldsymbol{\pi}.\mathbf{Q} \qquad [4]$$

Taking the limits, one obtains the transient behavior of the process:

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0).\sum_{n=0}^{\infty} \frac{t^n}{n!}.\mathbf{Q}^n = \boldsymbol{\pi}(0).e^{t.\mathbf{Q}} \qquad [5]$$

where the second equality is a definition. One calls $\mathbf{Q}$ the *infinitesimal generator* of the process.

2.4.2. *Existence and computation of a stationary distribution*

Assume that $\boldsymbol{\pi}(t)$ converges towards a stationary distribution. It is reasonable to suppose that $\frac{d\boldsymbol{\pi}}{dt}$ goes to 0. Hence equation [4] becomes $\boldsymbol{\pi}.\mathbf{Q} = 0$.

Indeed, the existence of a distribution, solution of equation:

$$\mathbf{X}.\mathbf{Q} = 0 \quad \text{and} \quad \mathbf{X}.\mathbf{1}^T = 1 \qquad [6]$$

is a necessary and sufficient condition and here again this equation admits at most a solution (if the embedded chain is irreducible). The computation is performed similarly to the one for discrete time chains.

2.5. *Semi-Markovian processes*

2.5.1. *Presentation*

Here we describe a restricted notion of semi-Markovian process and this for two reasons. First, this definition allows a simplified computation of stationary distributions; and, the next section presents a family of processes more general than the family of semi-Markovian processes. A semi-Markovian process is an extension of CTMCs where sojourn time in states may have any distribution. This process has the following characteristics:

- The time interval between instants T_n is a random variable that only depends on state X_n. Otherwise stated:

$$\Pr(T_n \leq t \,|\, X_0 = s_{i_0}, ..., X_n = s_i, T_0 \leq t_0, ..., T_{n-1} \leq t_{n-1}) =$$

$$\Pr(T_n \leq t \,|\, X_n = s_i) = \Pr(D_i \leq t)$$

  where D_i is a random variable with a finite mean, denoted d_i.

- The state following the current state only depends on this state and transition probabilities are constant over time:

$$\Pr(X_{n+1} = s_j \,|\, X_0 = s_{i_0}, ..., X_n = s_i, T_0 \leq t_0, ..., T_{n-1} \leq t_{n-1}) =$$

$$\Pr(X_{n+1} = s_j \,|\, X_n = s_i) = p_{ij} = \mathbf{P}[i, j]$$

Observe that here again, the sequence of states (X_n) constitutes a DTMC embedded in the process.

2.5.2. *Existence and computation of a stationary distribution*

Here we only state a sufficient condition for the existence of the stationary distribution which covers the most frequent cases. First we assume that the

embedded chain is irreducible with a distribution solution of $\mathbf{X}.\mathbf{P} = \mathbf{X}$ and that one of the distribution D_i is not arithmetic.

Here again the entrances in a state (s_i) can constitute a renewing process. Given some state ; the fact that it occurs infinitely only depends on transition probabilities p_{ij} and is ensured by our first hypothesis. The mean return time must be carefully examined. Indeed, every visit in s_i gives place to a sojourn with mean time d_i. Thus although the mean number of visits before a return is finite, the mean return time could be infinite. Let us call $\boldsymbol{\pi}'$ ($\pi'_k = \boldsymbol{\pi}'[s_k]$) the distribution solution of equation [3]. Then the mean number of visits of s_k between two visits of s_i is $\frac{\pi'_k}{\pi'_i}$. Consequently, the mean return time to s_i is equal to:

$$d_i + \sum_{k \neq i} d_k . \frac{\pi'_k}{\pi'_i} = \frac{1}{\pi'_i} . \sum_k d_k . \pi'_k$$

Otherwise stated, the existence of a stationary distribution is ensured if $\sum_k d_k.\pi'_k$ is finite. Since D_i is not arithmetic, one easily deduces that the distribution of return is not arithmetic.

With the same reasoning, one concludes that the ratio $\frac{\pi_k}{\pi_i}$ corresponds to the mean sojourn time in s_k between two returns in s_i divided by the mean sojourn time in s_i:

$$\frac{\pi_k}{\pi_i} = \frac{d_k . \frac{\pi'_k}{\pi'_i}}{d_i} = \frac{d_k . \pi'_k}{d_i . \pi'_i}$$

This leads to the stationary distribution:

$$\pi_k = \frac{\pi'_k . d_k}{\sum_{k'} \pi'_{k'} . d_{k'}} \qquad [7]$$

Observe that the way we have proceeded allows some distributions D_i to be concentrated in 0 (see section 3.3).

2.6. *Regenerative Markovian processes*

2.6.1. *Presentation*

A regenerative Markovian process (or semi-regenerative process) includes a subset of states, called regenerative states since entrance in any on this state constitutes a renewing process. We call S' this subset. Such a process is fully determined by its behavior between two consecutive entrances in regenerative states. Formally, one defines for every $k, k' \in S'$ and every $i \in S$ the following quantities:

- $F_k(t)$ represents the distribution of time between entrance in s_k and entrance in a new regenerative state,
- d_k the mean of this distribution supposed to be finite,
- $f_{ki}(t)$ the probability that, after t t.u. since entrance in s_k, there have been no new entrances in some regenerative state and that the process is in state s_i,
- $G_{kk'}(t)$ the probability that after entrance in s_k, the process has reached a new regenerative state $s_{k'}$ in time $\leq t$. $G_{kk'}$ represents a measure of mass ≤ 1.

2.6.2. *Existence and computation of a stationary distribution*

We simultaneously describe sufficient conditions of existence for a stationary distribution and its computation. First we suppose that the probability to reach in the future a new regenerative state from any regenerative state is always equal to 1, which means that F_k is a probability distribution.

One studies the embedded Markov chain representing the visits to regenerative states whose matrix is $\mathbf{P}$. This matrix can be computed by:

$$p_{kk'} = G_{kk'}(\infty) = \int_0^\infty G_{kk'}\{dt\}$$

We assume that this chain is ergodic and we note the distribution solution $\boldsymbol{\pi}'$. We need to compute the mean sojourn time in a state s_i between the entrance in s_k and the entrance in a new regenerative state (noté d_{ki}) :

$$d_{ki} = \int_0^\infty f_{ki}(t)dt$$

The stationary distribution ($\boldsymbol{\pi}$) is now deduced by weighting these sojourn times by the visit probabilities to the regenerative states.

$$\pi_i = \frac{\sum_{k\in S'} \pi'_k.d_{ki}}{\sum_{k\in S'} \pi'_k.d_k}$$

The main difficulty is related to the determination of the f_{ki}'s and the $G_{kk'}$'s. In some cases, this can be performed by the transient analysis of a Markov chain (see section 3.4).

3. Stochastic Petri nets

3.1. *Stochastic Petri nets with general distributions*

The stochastic feature of Petri nets is introduced by considering that a transition has a random firing delay (taking values in $\mathbb{R}^+$). The different families of stochastic Petri nets are defined by restricting the type of distributions. For the moment, we do not make any hypothesis on distributions. The definition of distributions is not sufficient to characterize the stochastic process. We are going to successively study the problems related to this characterization.

Remark: Most of the parameters of the process can depend on the current marking. For sake of simplicity, we will not mention it in the sequel.

3.1.1. *Choice policy*

Given the initial marking, we need to determine the next transition to fire among the fireable ones. There are two possible strategies:

- a probabilistic choice w.r.t. a distribution associated with the subset of fireable transitions. This is a *preselection* since the choice takes place before the sampling of the delay.
- an independent sampling for every delay followed by the choice of the shortest delay. In case of equal delays, one also performs a probabilistic choice called *post-selection*.

The second solution is always chosen since on the one hand it corresponds to a more natural modeling and on the other hand since with the help of immediate transitions (see section3.3) preselection can be simulated by post-selection. Observe that except if the distributions are continuous, one needs to specify the distributions of selections.

3.1.2. *Service policy*

If a transition has an enabling degree $e > 1$, one can consider that the marking *provides* e clients to the transition viewed like a server. So when sampling the delay, three options are possible depending on the event modeled by the transition:

- a single sampling is performed, the transition offers only one service at a time (*single-server* policy)

- e samplings are performed, the transition is a "parallel" server (*infinite-server* policy)

- $Min(e, deg(t))$ samplings are performed, the transition can offer at most $deg(t)$ simultaneous services; this case generalizes the other ones (with $deg(t) = 1$ or ∞) (*multiple-server* policy). The modeller must specify $deg(t)$ for every transition.

3.1.3. *Memory policy*

Once transition t is fired, what is the effect of a sampling that has not be chosen for another transition t' for the next firings?

The first possibility consists in forgetting the sampling that has been performed. If transition t' remains fireable, this takes place to a new sampling (*resampling memory*). With such a semantic, t could model the failure of a service specified by t'.

The second possibility consists in memorizing the sampling decremented by the sampling of t, but only if t' remains fireable (*enabling memory PRD* (*Preemptive Repeat Different*)). If t' is disabled, this mechanism models a time-out (t') disarmed by t.

The third possibility is as the previous one for a transition still fireable but let the sampling unchanged if t' is disabled. This sampling will be used again when t' will be fireable (mode *enabling memory PRI* (*Preemptive Repeat Identical*)). A disabled transition t' could model a job aborted by t that should be restarted.

The forth possibility consists in memorizing the sampling decremented by the sampling of t. A disabled transition t' could model a job suspended by t (*age memory* also called *PRS* (*Preemptive ReSume*)).

To complete this policy, we must take into account the case of multiple-server transitions, which requires to choose which samplings should be memorized, decremented or forgotten. The simplest solution is a *FIFO* policy for samplings. The last performed sampling is the first forgotten. Other policies (like suspend or forget the client the least engaged) are not necessarily compatible with some analysis methods.

It is clear that once these three policies are defined, the stochastic process is fully détermined. So we now focus on the distributions for transition delays.

3.2. *Stochastic Petri nets with exponential distributions*

In the basic model [FLO 85, MOL 81] every transition (t) has an exponential distribution with rate $\mathbf{w}[t]$ (that will be denoted $w_k = \mathbf{w}[t_k]$).

Let us examine the stochastic process generated by a stochastic Petri net with policy *single-server*. Let m be some marking, $t_1, \ldots, t_k$ the fireable transitions from m. One fulfills that:

- the sojourn time is an exponential with rate $w_1 + \ldots + w_k$
- the probability to pick t_i as the next firing is equal to $\frac{w_i}{w_1+\ldots+w_k}$ and it is independent from the sojourn time in the marking.
- the distribution of the remaining firing delay of t_i if t_j is fired is equal to the initial distribution (absence of memory)

Otherwise stated, only the new marking determines the future behavior of the stochastic process. Thus it is a continuous time Markov chain, isomorphic to the reachability graph of the Petri net, whose all parameters are given by states (*i.e.* the markings). This reasoning is also valid for other service policies.

If the graph is finite the formula [5] gives the transient behavior of the net and if furthermore it has a single terminal s.c.c. then the resolution of equation [6] provides the stationary distribution of the net.

Using the stationary distribution, other performance indices can be computed as the mean throughput (number of firings per time units) of transitions given by:

$$\overline{\chi}_k = \sum_{m \text{ reachable}} \pi_m.\text{services}(m, t_k).w_k \qquad [8]$$

where services(m, t_k) indicates the number of clients in state m served by transition t_k; this number depends on the enabling degree and the service policy of the transition.

3.3. *Generalized stochastic Petri nets*

Modeling an algorithm or a protocol requires to represent choices, loops and other control structures. These actions are logical operations and have a negligible duration w.r.t. a data transmission for instance. Modeling them by an exponential distribution with a high rate is unsatisfactory since, on the one hand the choice of the rate is arbitrary and on the other hand numerical computations suffer from values with very different magnitude order. To overcome

this difficulty, immediates transitions (*i.e.* with a distribution concentrated in 0) have been introduced. In this new model [Ajm 84], called GSPN for *Generalized Stochastic Petri Net*, the markings are partitioned in two categories: the tangible markings from which no immediate transition is fireable and the vanishing markings.

Let us examine the stochastic process generated by a GSPN from a given marking m. If m is tangible then the process is identical to the one of a Markovian SPN. Let us examine the case of a vanishing marking; there is at least one fireable immediate transition. *Almost surely* the sampling of exponential transitions is > 0. Thus the choice of the transition is done by a post-selection between immediate transitions. Since the delay of immediate transitions is null and the distributions of other transitions are without memory, the remaining delay are identical to the initial delays and the state of the process only depends on the new marking.

So this is a semi-Markovian process whose sojourn times in tangible markings follow an exponential distribution and sojourn times in vanishing markings are null. The transition probabilities (matrix $\mathbf{P}$) are obtained either from the rates, or from parameters of post-selection.

The analysis of section 2.5.2 is applicable here. However, in this particular case, an improvement is possible. Observe that in the stationary distribution (see equation [7]), the vanishing markings have a null occurrence probability. Thus one wants to eliminate them before the resolution of the embedded chain. To this aim, one considers the process as a Markovian regenerative process whose regenerative states are the tangible markings. We need to compute the transition probabilities between regenerative states. So we decompose matrix $\mathbf{P}$ in sub-matrices:

- $\mathbf{P}_{VV}$, transitions between vanishing markings
- $\mathbf{P}_{TT}$, transitions between tangible markings
- $\mathbf{P}_{VT}$, transitions from vanishing markings to tangible markings
- $\mathbf{P}_{TV}$, transitions from tangible markings to vanishing markings

Reasoning on the number of encountered vanishing markings, when going from a tangible marking to another tangible marking, one checks that the new transition matrix ($\mathbf{P}'$) is given by:

$$\mathbf{P}' = \mathbf{P}_{TT} + \sum_{n=0}^{\infty} \mathbf{P}_{TV}.(\mathbf{P}_{VV})^n.\mathbf{P}_{VT} = \mathbf{P}_{TT} + \mathbf{P}_{TV}.(\mathbf{Id}_{VV} - \mathbf{P}_{VV})^{-1}.\mathbf{P}_{VT}$$

where $\mathbf{Id}_{VV}$ is the identity matrix on vanishing markings.

If $\mathbf{Id}_{VV} - \mathbf{P}_{VV}$ is not invertible, it means a pathological behaviour (*i.e.* a non null probability to infinitely remain in the vanishing states). Otherwise the two expressions can be used to compute $\mathbf{P}'$.

This kind of elimination of vanishing states is applicable to more general models (*e.g.* deterministic SPNs) under some hypotheses.

The GSPN model is the one that has yielded the greater number of specifications and analyses of systems [Ajm 95]. Tool GreatSPN [CHI 95] has contributed to its expansion. Observing that these analyses are based on a finite reachability graph, several extensions have been introduced: inhibitor arcs, guards on transitions, rates depending on the current marking (called *functional dependencies*), etc. When we will mention the consequences of these extensions on the more elaborated methods.

3.4. *Deterministic stochastic Petri nets*

If the exponential distributions are appropriate for modeling events whose temporal distribution is unknown, some operations have a duration included in a time interval or even assimilable to a constant. In this case, the choice of an exponential distribution leads to very approximative results. So an extension of the GSPN model including deterministic transitions has been introduced [Ajm 87, LIN 98].

These nets are usually called Deterministic stochastic Petri nets (DSPN) . Several variations have been successively proposed for covering different situations. Here we only describe a basic version in order to more easily explain the stochastic process. We exclude immediate transitions since their handling is performed by the technique of the previous section. We also forbid functional dependencies: in particular rates of exponential transitions and delays of deterministic transitions are independent of the current marking. The net executes with single-server and enabling-memory PRD policies. The main hypothesis is that at any time *at most one deterministic transition is fireable.* which restricts the application area of this model. Recent works do not rely on this hypothesis at the price of increasing complexity.

Different methods have been proposed for the analysis of these nets [GER 99, LIN 98, LIN 99]. We present such a method whose efficiency has been experimentally proved. Let us examine the stochastic process generated by the net. Wa can characterize regenerative points:

- Every time that the process reaches a marking where no deterministic transition is fireable, the future of the process only depends on the marking.

- Every time that the process fires a deterministic transition and that a deterministic transition t_k is then fireable, the future of the process only depends on the marking since the firing delay of t_k is the initial delay, denoted d_k.
- Every time that the process fires a exponential transition and that a deterministic transition t_k is then fireable, the future of the process only depends on the marking since the firing delay of t_k is the initial delay, denoted d_k.

Be careful: in the last two cases, the regenerative point is characterized both by the reached marking and by the conditions of the firing since the same making can be reached by a firing that does not lead to a regenerative point. To distinguish regenerative points from markings, we note m^r, the regenerative point associated with a marking m and m^c a state reached with marking m and which is not a regenerative point.

Let us compute the parameters of the process behavior between two regenerative points in order to apply the results of section 2.5. Observe that when on enters a regenerative point, one fires a sequence of exponential transitions possibly ended by a firing of the active deterministic transition. For the first kind of regenerative point, every firing leads to a new point. The parameters of the behavior are thus given by the rates of the fireable transitions.

For the other kinds of regenerative points (m^r), the firing of exponential transitions corresponds to the evolution of a Markov chain whose states are some m_i^c and which ends by:

- either the firing of the deterministic transition d_k t.u. later,
- either the firing of an exponential transition at most d_k t.u. later that leads to m_j^r.

Let us call C_{m^r} the Markov chain composed by m^c (considered as initial state) and these m_i^c and m_j^r (these last ones being without successors in the chain). Let us note $\mathbf{Q}_{m^r}$ its infinitesimal generator which is directly obtained by the rates of exponential transitions. One says that this is a *subordinated chain*. $\boldsymbol{\pi}_t^{m^r}$ is the distribution at instant t of this chain knowing that the initial distribution $\boldsymbol{\pi}_0^{m^r}$ is concentrated in m^c. Let us recall that $\boldsymbol{\pi}_t^{m^r} = \boldsymbol{\pi}_0^{m^r}.e^{t.\mathbf{Q}_{m^r}}$

The transition probabilities between regenerative points (matrix $\mathbf{P}$) are deduced from the state of this chain at instant d_k :

$$\mathbf{P}[m^r, m_1^r] = \boldsymbol{\pi}_{d_k}^{m^r}[m_1^r] + \sum_{m_2[t_k>m_1} \boldsymbol{\pi}_{d_k}^{m^r}[m_2^c]$$

which means that m_1^r has been the first regenerative point reached no later than d_k or that no regenerative point has been reached before d_k and then that the firing of the deterministic transition has led to m_1. The mean sojourn time in a marking before reaching the new regenerative point is given by:

$$\text{sojourn}_{m^r}(m_1) = \int_0^{d_k} \boldsymbol{\pi}_t^{m^r}[m_1^c]dt$$

To efficiently perform these computations, different techniques are possible [JEN 53, GRO 84]. However this resolution is most time and space expensive than the one of GSPNs.

3.5. *Phase-type stochastic Petri nets*

A phase-type distribution [NEU 81] is defined by a Markov chain with an absorbing state (*i.e.* without successor) and an initial distribution. If F denotes the distribution then $F(t)$ is the probability to be in the absorbing state at time t. Using 2.3.2, F is a probability distribution iff the absorbing state is the single terminal s.c.c. of the graph associated with the chain. In this case, F is defined by equation [5]. The states of the chaine (except the last one) are called stages.

It has been established that in some sense, every distribution is a limit of phase-type distributions [COX 55]. For instance, an exponential distribution is a phase-type distribution with a single stage and an immediate distribution is a phase-type distribution without stage. A deterministic distribution with duration d is approximated by a distribution with n consecutive stages whose rate is $\frac{n}{d}$.

So the phase-type stochastic Petri nets (PH-SPN) have a great expressive power. However, such a net generates a stochastic processus of the same kind as the one of GSPN. Indeed, the sampling of a phase-type distribution can be seen as a random sampling of the choice of the first stage, a sampling of the exponential distribution of the stage, a new random sampling of the choice of the next stage, etc. until one reaches the absorbing state. So, rather than considering transition firings as the events of the SED, one selects a more elementary step: the stage change of the distributions. This requires to complete the state of the SED. A state is defined by:

- a marking,
- for every transition, a descriptor which includes a sequence of samplings not yet used to fire the transition. For every sampling, one memorizes its current stage. If the transition works with the *enabling memory* policy,

the number of firings is exactly the number of services offered by the transition. If it works with the *age memory* policy, this number can be greater since one takes into account the suspended services.

Every step that reaches an absorbing state is an *external* transition since it updates the marking. The new descriptor is computed w.r.t. the different policies of the net. The *internal* transitions let the marking unchanged and in the descriptor a single sampling is updated.

One builds the semi-Markovian process as a reachability graph starting from the initial state and *firing* the internal and external transitions. More elaborate constructions are possible by noting that, for instance, some markings lead to the same set of descriptors.

However the problem is the number of states of this process which has the same magnitude order as the product of the size of the reachability space and of the number of descriptors. We will see in the following chapters how to obtain the stationary probabilities of the net without building the process.

4. Some standard analysis methods

4.1. *Research of a product form*

We describe this method in the framework of exponential SPNs through an example in order to illustrate its principles without entering the algorithmic details.

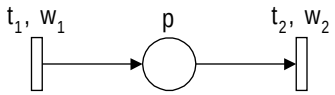

Figure 2: Modeling of a queue by a Petri net

Let us look at the net of figure 2, which models a queue. The steady-state distribution of this (unbounded) net is given by (for $w_1 < w_2$) :

$$\boldsymbol{\pi}[n.p] = \left(1 - \frac{w_1}{w_2}\right) . \left(\frac{w_1}{w_2}\right)^n$$

In this equation, we observe that the marking n of the place appears as exponent in the distribution.

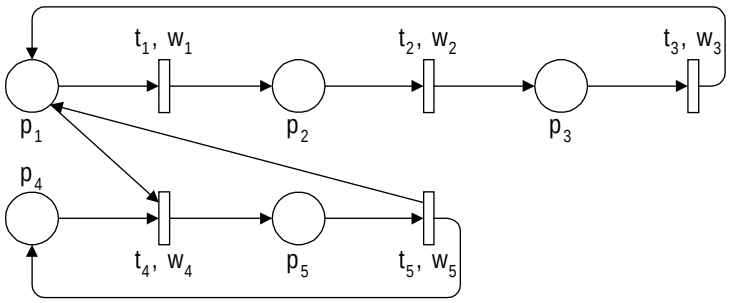

Figure 3: A Petri net with a product form

So the first idea is to generalize this formula as a product whose terms are expressions obtained from the rates of k' transitions and whose exponents are markings of k places.

$$\boldsymbol{\pi}[\sum_{i=1}^{k} n_i.p_i] = \frac{1}{G}.\prod_{i=1}^{k}(f_i(w_1,...,w_{k'}))^{n_i}$$

G, the normalizing constant, is defined as the sum over the set of reachable markings, of products occurring in the right hand term of the equation.

Let us examine the net of figure 3. One observes that the transitions can be partitioned in two subsets $T_a = \{t_1, t_2, t_3\}$ and $T_b = \{t_4, t_5\}$. Inside every subset, preconditions of a transition are postconditions of another one and vice versa. For instance, from marking m one can fire t_1, iff m is obtained by firing t_3 from another marking. Let us note $\mathbf{Q}^a$ (resp. $\mathbf{Q}^b$) matrix $\mathbf{Q}$ where all rates are cancelled except those of T_a (resp. T_b), $\mathbf{Q} = \mathbf{Q}^a + \mathbf{Q}^b$.

The second idea is to substitute equation $\mathbf{X}.\mathbf{Q} = 0$ by two equations $\mathbf{X}.\mathbf{Q}^a = 0$ and $\mathbf{X}.\mathbf{Q}^b = 0$. Solving these two systems is not equivalent but a solution of the second system provides a solution of the first system. In this context, the first system is called global balance equations and the second one is called local balance equations.

Combining these two ideas, we look for a product form as follows:

$$\boldsymbol{\pi}(m) = \frac{1}{G}.f_1(m).f_2(m)$$

where f_1 (which depends on w_1, w_2, w_3) is unchanged by a firing in T_b and f_2 (which depends on w_4, w_5) is unchanged by a firing in T_a. Assume that this form exists and detail a local balance equation.

$$\sum_{m'[t'>m,\ t'\in T_a} \frac{1}{G}.f_1(m').f_2(m').\mathbf{w}[t'] = \sum_{m[t''>m'',\ t''\in T_a} \frac{1}{G}.f_1(m).f_2(m).\mathbf{w}[t'']$$

Since $f_2(m) = f_2(m')$, the equation is equivalent to:

$$\sum_{m'[t'>m,\ t'\in T_a} f_1(m').\mathbf{w}[t'] = \sum_{m[t''>m'',\ t''\in T_a} f_1(m).\mathbf{w}[t''] \qquad [9]$$

The main difficulty to obtain a solution is the constraint that f_1 is unchanged by a firing in T_b. In the example, markings of p_2 and p_3 are unchanged by a firing in T_b and those of p_4 and p_5 are unchanged by a firing in T_a.

So we write:

$f_1(m) = (w_1)^{a_1.m(p_2)+b_1.m(p_3)}.(w_2)^{a_2.m(p_2)+b_2.m(p_3)}.(w_3)^{a_3.m(p_2)+b_3.m(p_3)}$

Let us recall that from m:

- one fires t_1 iff m is reached by a firing of t_3
- one fires t_2 iff m is reached by a firing of t_1
- one fires t_3 iff m is reached by a firing of t_2

Equalizing these terms in equation [9], leads to (after simplification):

$$\begin{array}{rcl} (w_1)^{b_1}.(w_2)^{b_2}.(w_3)^{b_3+1} & = & w_1 \\ (w_1)^{-a_1+1}.(w_2)^{-a_2}.(w_3)^{-a_3} & = & w_2 \\ (w_1)^{a_1-b_1}.(w_2)^{a_2-b_2+1}.(w_3)^{a_3-b_3} & = & w_3 \end{array}$$

The only solution (for any possible value of rates) is then:

$$f_1(m) = \left(\frac{w_1}{w_2}\right)^{m(p_2)} \cdot \left(\frac{w_1}{w_3}\right)^{m(p_3)}$$

Similarly:

$$f_2(m) = \left(\frac{w_4}{w_5}\right)^{m(p_4)}$$

In the general case, conditions for the existence of this decomposition are fulfilled by a subclass of nets called *Product Form Stochastic Petri Net*, PF-SPN) [HEN 90]. Furthermore, [HAD 01, HAD 05] establish a necessary and sufficient condition fully structural that such a net admits a product form whatever its stochastic parameters. It remains a last difficulty. If the computation of the normalizing constant is naively performed, this requires to enumerate reachable states. So it reduces the interest of the method. Fortunately, the presence of invariants characterizing the reachability space greatly simplifies this computation [SER 93].

To conclude, methods based on product form have a weak computational complexity but they are applicable on models whose components have simple synchronizations.

4.2. *Bound computations*

The stochastic bounds that we state here are valid for every distribution (with finite mean) of transition delays [CHI 93]. They only rely on the existence of a steady-state distribution of markings and on steady-state throughput of transitions. Hence these bounds are valid for all nets presented in this chapter including the nets with an infinite state space. On the other hand, these bounds will be accurate when the performance measures are insensible (*i.e.* only depend on the mean of distributions) or weakly sensible. This analysis is extended in [LIU 95] where the author inserts constraints related to the variance of distributions.

The general idea is:

- to represent the performance indices by variables,
- to establish linear constraints between variables,
- to maximize or minimize, with linear programming, a linear function of variables (that represents performance index to be evaluated) submitted to the previous constraints.

Numerous algorithms are possible for this last step among them the more efficient ones perform in polynomial time w.r.t. the size of constraints [NEM 89]. So we only describe the first two points.

For every place p, variable $\overline{m}_p$ denotes the mean marking of p. For every transition t, variable $\overline{\chi}_t$ denotes the mean throughput of t. At last, variable σ_t is the number of occurrences of t in a pseudo firing sequence since this number is not necessarily an integer. These variables occur in the constraint but not in the function to be optimized.

The first constraint is defined by:

$$\forall p, \forall t, \;\; \overline{m}_p \geq 0, \;\; \overline{\chi}_t \geq 0, \;\; \sigma_t \geq 0$$

The mean marking $\overline{m}$ is a mean, weighted par $\boldsymbol{\pi}$, of reachable markings m, every marking being reached by a sequence σ_m (with occurrence vector $\overrightarrow{\sigma}_m$) from m_0. Consequently:

$$\overline{m} = m_0 + \sum_{m \text{ reachable}} \pi_m.(\mathbf{Post}[p,t] - \mathbf{Pre}[p,t]).\overrightarrow{\sigma}_m$$

The second term is the product of the incidence matrix by a weighted mean of sequences that can be substituted by variables σ_t. This leads to the following

constraint:

$$\overline{m}_p = m_0(p) + \sum_t (\mathbf{Post}[p,t] - \mathbf{Pre}[p,t]).\sigma_t$$

There are two other kinds of constraints: the first ones are obtained by an analysis of the steady-state situation and the second ones by establishing a relation related to the behavior of the process until a arbitrary instant and studying the asymptotic behavior of this relation when time goes to infinity. This last technique is called *operational analysis* .

We exhibit few examples of these two kinds of constraints. Assume that two transitions t and t' are simultaneously fireable or not and that the choice probabilities between t and t' are constant. This is the case for instance with exponential transitions whose rates are constant and that follow a *single-server* policy. Let us note r_t and $r_{t'}$ these probabilities. Obviously:

$$\frac{\overline{\chi}_t}{r_t} = \frac{\overline{\chi}_{t'}}{r_{t'}}$$

Since the steady-state distribution of the marking exists, the input flows of a place must be equal to the output flows. Consequently:

$$\forall p, \ \sum_t \mathbf{Pre}[p,t].\overline{\chi}_t = \sum_t \mathbf{Post}[p,t].\overline{\chi}_t$$

Let t be a transition with mean delay $\frac{1}{w_t}$, working with *infinite-server* policy and having only a single input arc labelled by 1 and connected to place p which has only t as output. Observe the process between 0 and θ. Let $m_p(s)$ denote the number of tokens at time s, let $\overline{m}_p(\theta)$ the mean number of tokens between 0 and θ :

$$\overline{m}_p(\theta) = \frac{1}{\theta} \int_0^\theta m_p(s) ds$$

Let us introduce $V(\theta)$ the sum of sojourn time of tokens in p that are consumed before θ and $U(\theta)$ the sum of sojourn time of tokens in p that are produced before θ or present in the initial marking.

To establish a relation between these quantities, assume that the tokens client of the transition pay a uniform cost with rate 1: otherwise stated a token present in an interval with length ds pays ds euros. If tokens arrived before θ pay their presence until time θ, $\int_0^\theta m_p(s)ds$ is the amount that the transition has won between 0 et θ. If the tokens arrived before θ pay their presence when they leave, $V(\theta)$ is the amount that the transition has won between 0 et θ. If the tokens arrived before θ pay their presence when they arrive, $U(\theta)$ is the amount that the transition has won between 0 et θ. Consequently:

$$V(\theta) \leq \theta.\overline{m}_p(\theta) \leq U(\theta) \ \Leftrightarrow \ \frac{V(\theta)}{\theta} \leq \overline{m}_p(\theta) \leq \frac{U(\theta)}{\theta} \qquad [10]$$

Let us analyze the asymptotic behavior. Let $N(\theta)$ be the number of tokens arrived between 0 et θ. Since the net has a steady-state distribution, the input flow of p is equal to the output flow.

$$\lim_{\theta \to \infty} \frac{N(\theta)}{\theta} = \overline{\chi}_t$$

Since there is a steady-state distribution of the marking,

$$\lim_{\theta \to \infty} \overline{m}_p(\theta) = \overline{m}_p$$

Furthermore every token has a sojourn time equal to the firing delay of the transition. Let us call d_n the sojourn time of the n^{th} token. Using the law of great numbers:

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} d_i}{n} = \frac{1}{w_t}$$

Let us establish a relation between these quantities.

$$\frac{\overline{\chi}_t}{w_t} = \lim_{\theta \to \infty} \left(\frac{N(\theta)}{\theta} \right) . \left(\frac{\sum_{i=1}^{N(\theta)} d_i}{N(\theta)} \right) = \lim_{\theta \to \infty} \frac{U(\theta)}{\theta}$$

Using an analytical reasoning [STI 74] one proves that:

$$\lim_{\theta \to \infty} \frac{U(\theta)}{\theta} = \lim_{\theta \to \infty} \frac{V(\theta)}{\theta}$$

Passing to the limit, equation [10] provides another constraint:

$$\frac{\overline{\chi}_t}{w_t} = \overline{m}_p$$

This constraint is a variation of Little formula. Let us remark that the bound computation has also been applied to stochastic well-formed nets, since symmetries of the model ease specification of constraints.

4.3. *Approximation methods*

4.3.1. *Approximation by decomposition*

Here we are interested in a subclass of SPNs called *Stochastic Marked Graph*, SMG whose underlying net is an event graph (*c.f.* chapter 3). This class

is often used for modeling product flows like for instance in manufacturing systems [SIL 97, VAL 97].

In these nets, every place is output of a single transition and input of a single transition. Moreover the net, viewed like a graph, is strongly connected. We assume that the net works with an *infinite-server* policy. There are similar methods for the other policies.

All throughputs of transitions in steady-state behavior (*i.e.* the number of firings per time unit) are equal. Indeed, a path links every pair of transitions (t_1, t_2) and the (new) tokens in this path are produced by t_1 and consumed by t_2. If the throughput of t_1 would be greater than the one of t_2, the number of tokens would infinitely grow which is impossible since this number is bounded by the initial number of tokens of a circuit including this path. By symmetry, one deduces that the throughput of t_1 is equal to the one of t_2. So we can say that the net has a throughput. To compute the throughput of the net, one can establish the steady-state distribution of the net, pick a transition and apply formula [8].

The goal of the approximation by decomposition is to substitute to the construction of the reachability graph of the net, the construction of graphs for subnets obtained by decomposition [CAM 94]. Indeed, for appropriate decompositions, the size of the whole graph has the same magnitude order as the product of sizes of the graphs of the subnets. The algorithm includes two steps:

- the decomposition in subnets,
- the approximate computation of the throughput.

In order to obtain a decomposition, one chooses a set of places called a cut which partitions the net in two connected components. This choice must be guided by the behaviour of the system. As a general rule, one wants to minimize the size of the cut. In the figure 4, the cut is the set of places pa, pb et pc. Let R_1 and R_2 the two subnets; each one consists in a connected component, the cut and transitions connected to the cut. To complete these nets, places p_{61}, p_{62} and p_{43} are added to the subnets. Every such place corresponds to a pair of transitions at the boundary of a subnet. For instance, p_{61} corresponds to pair (t_6, t_1).

Places p_{61}, p_{62} (resp. p_{43}) represent an abstraction of the first (resp. second) component. The initial marking of p_{61} is defined as the minimum of the marking of a path from t_6 to t_1 in this component. We apply the same process for the other places.

Net R_{12} constitutes a full abstraction of the initial net. Due to the structural properties of event graphs, all these abstractions do not modify languages

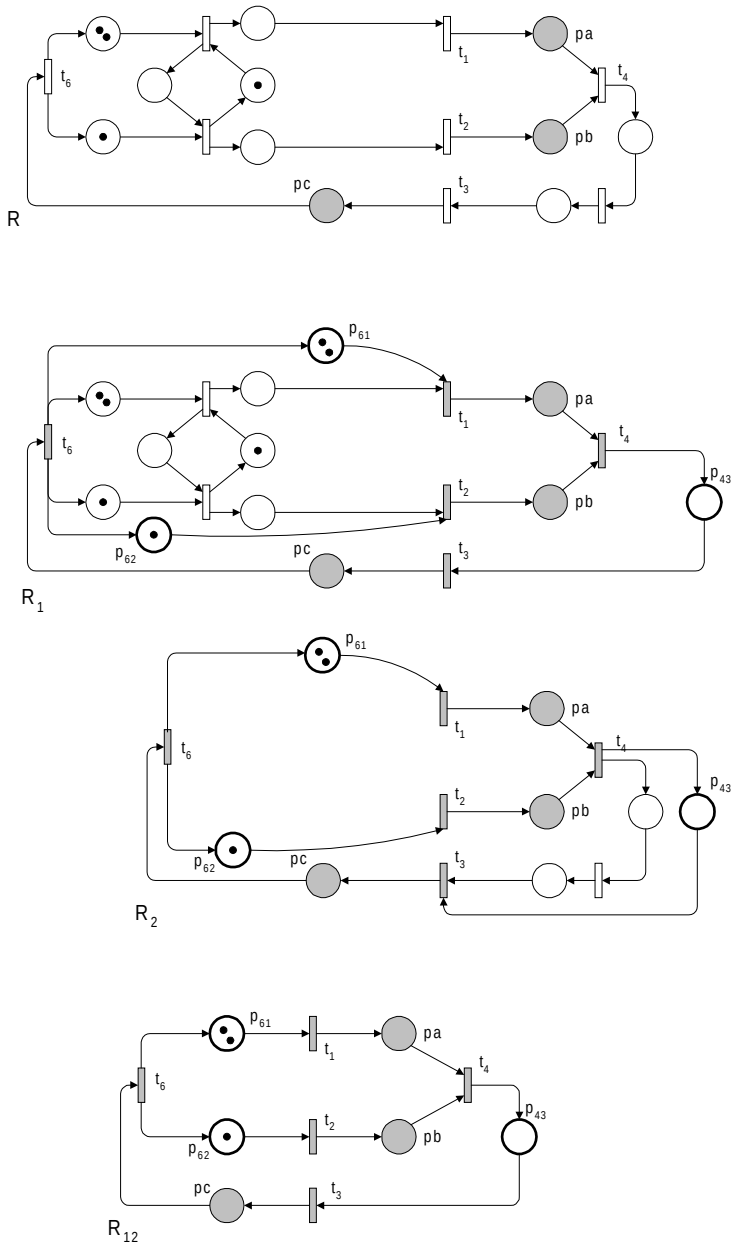

Figure 4: Décomposition and abstraction of a stochastic event graph

(omitting the transitions that do not occur in the abstraction) and the set of reachable markings (omitting the places that do not occur in the abstraction).

The computation of the throughput of the net is performed iteratively. We explain it on the example. The rate of t_3 in net R_1 will be updated at the beginning of each stage in order to express the activity of the other component. We do it similarly for rates of t_1 and t_2 in R_2. One initially selects a rate for t_3 in R_1, for instance its value in R. Then each stage includes four steps:

- using the steady-state distribution, one computes the throughput of net R_1, denoted χ_1, and the mean marking of p_{61} and p_{62}. Observe that these mean markings are proportional to the *service* time of t_1 and t_2 for these places (*i.e.* the mean time to consume a token of p_{61} et p_{62}) since the production is simultaneous and that we have chosen an *infinite-server* policy.
- the ratio of rates of t_1 and t_2 in R_2 is now determined by the previous step. It remains to compute the scaling factor. This is done in R_{12} where for different values of this factor, one computes the steady state distribution and the throughput of the net in order to be as close as possible as χ_1. The graph of R_{12} has a very small size. Hence this step has a reasonnable complexity.
- the third step is symmetrical w.r.t. the first step. One computes the throughput χ_2 of R_2 with rates of t_1 and t_2 obtained by the previous step. In this example, since t_3 is the single transition which leads from R_2 to R_1, the computation of ratios is useless.
- one examines again net R_{12} to compute the rate of t_3 to be used in R_1. One tries different values in order that the throughput of R_{12} is as close as possible of χ_2.

One finishes the iterations when values χ_1 are χ_2 enough close that one can assume that they correspond to the throughput of net R. There is no theoretical guarantee for convergence, neither for the precision of the result. This lack of guarantee usually holds for almost all approximation methods. However, experimentations show a very fast convergence (less than 10 iterations) and an error less than 1%. These good results are surely due to the fact that the quantitative approximation is based on an appropriate functionnal decomposition.

4.3.2. *Approximation by mean values*

This analysis is applicable a priori on any kind of Petri nets although the tools are limited to uniform, exponential, deterministic distributions and some

of their combinations. We present it for transitions working with *single-server* and *enabled memory PRD* policies.

The method is based on the construction of a graph called *probabilistic state graph* [JUA 91]. Every vertex of the graph includes a marking and distributions associated with every transition. the initial vertex includes the initial marking and the specified distributions.

One determines the fireable transitions $\{t_k\}_{k\in K}$ and one computes the probability to fire every transition. In the sequel, we assume that the distributions are continuous. It is equivalent to compute the probability that the sampling of every transition is the smallest one and this is expressed by the following formula:

$$\Pr(t_k \text{ fired}) = \int_0^\infty \prod_{k'\neq k} (1 - F_{k'}(s))F_k\{ds\}$$

where F_k is the distribution associated with t_k. In the general case of discontinuous distributions, one must include the parameters of post-selection which complicates the expressions but does not change the principle of the method.

The mean sojourn time in the vertex is similarly expressed by the formula:

$$\int_0^\infty \prod_{k\in K} (1 - F_k(s))ds$$

One builds a successor per possible transition firing. Her is the approximation since one considers that the transition is fired after a deterministic time that is computed by:

$$\theta_k = \frac{1}{\Pr(t_k \text{ fired})} \int_0^\infty s. \prod_{k'\neq k} (1 - F_{k'}(s))F_k\{ds\}$$

The random variable has been substituted by its mean. The new distribution of a transition still fireable is:

$$F'_{k'}(t) = \frac{F_{k'}(t + \theta_k) - F_{k'}(\theta_k)}{1 - F_{k'}(\theta_k)}$$

The other distributions are the initial distributions. If the reachability graph and the intermediate distributions are finite, then the probabilistic state graph is finite. There exist sufficient conditions on the Petri net for this property. In case the graph is infinite, a stopping mechanism for cutting branches is introduced which takes into account branching probabilities to eliminate vertices supposed to be reached with a weak probability.

This graph is now viewed like a regenerative process specified by the branching probabilities and the sojourn times. The resolution is performed as indicated in section 2.5.1.

Unfortunately it is difficult to establish criteria which ensure a good approximation. For instance, numerous uniform distributions diminish the accurateness of the approximation. Furthermore, some reachable states can be missed by the construction even without the stopping mechanism.

4.4. *Unbounded Petri nets*

Until now, most of the methods that we have presented apply to bounded nets. We end this chapter by an exact method applicable to infinite state systems [FLO 89, HAV 95]. Here we only study SPNs but this analysis is also applicable (with some adaptations) to GSPNs and even to PH-SPNs. When the nets have a single unbounded place p, it is possible to compute the steady-state distribution with additional weak constraints (other conditions are also possible):

- Arcs connected to place p are labelled by 1.
- Two arbitrary values of $m(p)$ greater than some threshold k_0 yield the same transition rates and the same firing conditions in the possible functional dependencies.

In that case, reachable markings are partitioned depending on the marking of p in a family $\{S_k\}_{k=0}^{\infty}$ where S_k is the subset of reachable markings such $m(p) = k$. The constraints imply that from S_k, either one reaches S_{k-1}, S_{k+1} or one remains in S_k. Also, if $m(p) > k_0$, place p has no more effect on the behavior of the net. So matrix $\mathbf{Q}$ of the infinitesimal generator of the net presents beyond k_0 regularities that are expressed by the existence of three matrices:

- $\mathbf{A}_0$ is the transition submatrix from S_k to S_{k+1}
- $\mathbf{A}_1$ is the transition submatrix from S_k to S_k
- $\mathbf{A}_2$ is the transition submatrix from S_k vers S_{k-1}

Assume that the chain is irreducible and ergodic with a steady-state distribution $\boldsymbol{\pi}$ and note $\boldsymbol{\pi}_k$ the distribution over states S_k. The equilibrium equation can be rewritten for $k > k_0$:

$$\boldsymbol{\pi}_k.\mathbf{A}_0 + \boldsymbol{\pi}_{k+1}.\mathbf{A}_1 + \boldsymbol{\pi}_{k+2}.\mathbf{A}_2 = 0$$

We want to establish a recurrence between $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_{k+1}$. Using the structure of $\mathbf{Q}$, $\mathbf{A}_1$ is invertible and $-\mathbf{A}_1^{-1}$ is a positive matrix. Consequently :

$$\boldsymbol{\pi}_{k+1} + \boldsymbol{\pi}_k.\mathbf{A}_0.\mathbf{A}_1^{-1} = -\boldsymbol{\pi}_{k+2}.\mathbf{A}_2\mathbf{A}_1^{-1} \geq 0$$

One improves this relation by defining a sequence of increasing matrices:

$$\mathbf{R}_0 = -\mathbf{A}_0.\mathbf{A}_1^{-1} \text{ et } \mathbf{R}_{n+1} = -(\mathbf{A}_0 + (\mathbf{R}_n)^2.\mathbf{A}_2).\mathbf{A}_1^{-1}$$

One shows by recurrence that the left hand term below remain positive (and decreases since the $\mathbf{R}_n$ are increasing). Indeed:

$$\begin{aligned}\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k.\mathbf{R}_{n+1} &= \boldsymbol{\pi}_{k+1} + \boldsymbol{\pi}_k.\mathbf{A}_0.\mathbf{A}_1^{-1} + \boldsymbol{\pi}_k.(\mathbf{R}_n)^2.\mathbf{A}_2.\mathbf{A}_1^{-1} \\ &= -(\boldsymbol{\pi}_{k+2} - \boldsymbol{\pi}_k.(\mathbf{R}_n)^2).\mathbf{A}_2\mathbf{A}_1^{-1} \\ = -(\boldsymbol{\pi}_{k+2} - \boldsymbol{\pi}_{k+1}.\mathbf{R}_n + (\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k.\mathbf{R}_n).\mathbf{R}_n).\mathbf{A}_2\mathbf{A}_1^{-1} &\geq 0\end{aligned}$$

Since every component of $\boldsymbol{\pi}_k$ is non null, matrices $\mathbf{R}_n$ are bounded and the sequence converges to a matrix $\mathbf{R}$ which fulfills (by passing to the limit) :

$$\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k.\mathbf{R} \geq 0 \text{ pour } k > k_0 \quad [11]$$

$$\mathbf{R} = -(\mathbf{A}_0 + \mathbf{R}^2.\mathbf{A}_2).\mathbf{A}_1^{-1} \quad [12]$$

Let us define a vector $\boldsymbol{\pi}'$ by:

$$\forall\, k \leq k_0,\ \ \boldsymbol{\pi}'_k = \boldsymbol{\pi}_k \quad \text{et} \quad \forall\, k > k_0,\ \ \boldsymbol{\pi}'_k = \boldsymbol{\pi}_{k_0+1}.\mathbf{R}^{(k-k_0-1)}$$

Then, using equation [12], this vector is solution of equation $\mathbf{X}.\mathbf{Q} = 0$ and using equation [11], it is smaller or equal than $\boldsymbol{\pi}$ component per component. Hence the sum of its components is finite. Normalizing it (*i.e.* dividing it by this sum), one obtains $\boldsymbol{\pi}''$ a distribution solution. But the distribution solution is unique. So $\boldsymbol{\pi} = \boldsymbol{\pi}'' = \boldsymbol{\pi}'$.

The normalizing equation can be written as:

$$\sum_{k=0}^{k_0} \boldsymbol{\pi}_k.1^T + \sum_{k=k_0+1}^{\infty} \boldsymbol{\pi}_{k_0+1}.\mathbf{R}^{(k-k_0-1)}.1^T = \sum_{k=0}^{k_0} \boldsymbol{\pi}_k.1^T + \boldsymbol{\pi}_{k_0+1}.(\mathbf{Id}-\mathbf{R})^{-1}.1^T = 1$$

Once matrix $\mathbf{R}$ (*e.g.* by approximating it with $\mathbf{R}_n$ for large n) and $(\mathbf{Id} - \mathbf{R})^{-1}$ are computed, we proceed to a linear resolution in a finite space. The system to be solved is $\mathbf{X}.\mathbf{Q} = 0$, reduced to states of $\{S_k\}_{k \leq k_0+1}$ and completed by the normalisazing equation [HAV 98].

This procedure has also been successfully employed to approximate finite systems where a place reaches huge values. The threshold k_0 is often much smaller than the bound of the place. If furthermore the system fulfills a set of conditions for quasi-reversibility [KEL 79], this approximation becomes an exact result [HAV 93].

5. Conclusion

Stochastic Petri nets have been initially introduced as another formalism to represent stochastic DES with exponential distributions. Along the thirty last years, the modeling needs have led to extend this model to more general distributions (deterministic, phase type, arbitrary) including null delay. These extensions generate families of SPNs whose properties depend on multiple choices, sometimes subtle, related to the stochastic semantic of the model. Most of these nets generate stochastic processes that are renewing processes. Furthermore research has also developed appropriate analysis methods for these processes. Some of them adapt results obtained in the queuing network theory. However most of them, are partly based on structural properties related to the ordinary nets. The two next chapters present two characteristic examples of such approaches: the stochastic well formed nets and the tensorial methods for SPNs.

References

[Ajm 84] M. Ajmone Marsan, G. Balbo and G. Conte. A class of generalized of stochastic Petri nets for the performance evaluation of multiprocessor systems. *ACM Transactions on computer systems*, 2(2):93–122, May 1984.

[Ajm 87] M. Ajmone Marsan and G. Chiola. On Petri nets with deterministic and exponentially distributed firing times. In G. Rozenberg, ed, *Advances in Petri Nets 1987*, number 266 in LNCS, pages 132–145. Springer–Verlag, 1987.

[Ajm 95] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets*. Wiley series in parallel computing. John Wiley & Sons, England, 1995.

[BOL 98] G. Bolch, S. Greiner, H. de Meer and K.S. Trivedi. *Queueing networks and Markov chains*. John Wiley & Sons, New-York, 1998.

[CAM 94] J. Campos, J.M. Colom, H. Jungnitz and M. Silva. Approximate throughput computation of stochastic marked graphs. *IEEE Transactions on Software Engineering*, 20(7):526–535, July 1994.

[CHI 93] G. Chiola, C. Anglano, J. Campos, J.M. Colom and M. Silva. Operationnal analysis of timed Petri nets and application to computation of performance bounds. In *Proc. of the 5th International Workshop on Petri Nets and Performance Models*, pages 128–137, Toulouse, France, October 19–22 1993. IEEE Computer Society Press.

[CHI 95] G. Chiola, G. Franceschinis, R. Gaeta and M. Ribaudo. GreatSPN 1.7: GRaphical Editor and Analyser for Timed and Stochastic Petri nets. *Performance Evaluation*, 24(1,2):47–68, 1995.

[COX 55] D. R. Cox. A use of complex probabilities in the theory of stochastic processes. *Proc. Cambridge Philosophical Society*, pages 313–319, 1955.

[DAL 97] Y. Dallery, Y. Liu and D. Towsley. Properties of fork/join queueing networks with blocking under various operating mechanisms. *IEEE Transactions on Robotics and Automation*, 13(4):503–518, 1997.

[FDI 86] S. Fdida, G. Pujolle and D. Mailles. Réseaux de files d'attente avec sémaphores. *TSI*, 5(3):187–196, mai-juin 1986.

[FEL 68] W. Feller. *An introduction to probability theory and its applications. Volume I.* John Wiley & Sons, 1968. (third edition).

[FEL 71] W. Feller. *An introduction to probability theory and its applications. Volume II.* John Wiley & Sons, 1971. (second edition).

[FLO 78] G. Florin and S. Natkin. Évaluation des performances d'un protocole de communication à l'aide des réseaux de Petri et des processus stochastiques. In *Journées AFCET Multi-ordinateurs, multi-processeurs en temps réel*, CNRS, Paris, France, mai 1978.

[FLO 85] G. Florin and S. Natkin. Les réseaux de Petri stochastiques. *TSI*, 4(1):143–160, 1985.

[FLO 89] G. Florin and S. Natkin. Necessary and sufficient ergodicity condition for open synchronized queuing networks. *IEEE Transactions on software engineering*, 15(4):367–380, 1989.

[GER 99] R. German and M. Telek. Formal relation of Markov renewal theory and supplementary variables in the analysis of stochastic Petri nets. In P. Buchholz and M. Silva, eds, *Proc. of the 8th International Workshop on Petri Nets and Performance Models*, pages 64–73, Zaragoza, Spain, September 8–10 1999. IEEE Computer Society Press.

[GRO 84] D. Gross and D.R. Miller. The randomization technique as a modeling tool an solution procedure for transient markov processes. *Oper. Res.*, 32:345–361, 1984.

[HAD 01] S. Haddad, P. Moreaux, M. Sereno and M. Silva. Structural charaterization and behavioural properties of product form stochastic Petri nets. In *Proc. of the 22th International Conference on Application and Theory of Petri Nets*, volume 2075 of *LNCS*, pages 164–183, NewCastle Upon Tine, UK., June 25–29 2001.

[HAD 05] S. Haddad, P. Moreaux, M. Sereno and M. Silva. Product-form and stochastic Petri nets: a structural approach. *Perform. Eval.*, 59:313–336, 2005.

[HAV 93] B.R. Haverkort. Approximate performability and dependability modelling using generalized stochastic Petri nets. *Performance Evaluation*, 18(1):61–78, 1993.

[HAV 95] B.R. Haverkort. Matrix-geometric solution of infinite stochastic Petri nets. In *Proc. of the International Performance and Dependability Symposium*, pages 72–81. IEEE Computer Society Press, 1995.

[HAV 98] B.R. Haverkort. *Performance of Computer Communication Systems*. John Wiley & Sons, 1998.

[HEN 90] W. Henderson, C.E.M. Pearce, P.G. Taylor and N.M. van Dijk. Closed queueing networks with batch services. *Queueing Systems*, 6:59–70, 1990.

[HIL 96] J. Hillston. *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[JEN 53] A. Jensen. Markov chains as an aid in the study of Markov processes. *Skand. Aktuarietidskrift*, 3:87–91, 1953.

[JUA 91] G. Juanole and Y. Atamna. Dealing with arbitrary time distributions with the stochastic timed Petr net model – application to queueing systems. In *Proc. of the fourth International Workshop on Petri Nets and Performance Models*, pages 32–51, Melbourne, Australia, December 2–5 1991. IEEE Computer Society Press.

[KEL 79] F.P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons, England, 1979.

[KLE 75] L. Kleinrock. *Queueing systems. Volume I: Theory*. Wiley-Interscience, New-York, 1975.

[LAP 95] J.C. Laprie, ed. *Guide de la sûreté de fonctionnement*. Cépaduès - Éditions, Toulouse, France, 1995.

[LIN 98] C. Lindemann. *Performance Modelling with Deterministic and Stochastic Petri Nets*. John Wiley & Sons, 1998.

[LIN 99] C. Lindemann, A. Reuys and A. Thümmler. DSPNexpress 2.000 performance and dependability modeling environment. In *Proc. of the* 29^{th} *Int. Symp. on Fault Tolerant Computing*, Madison, Wisconsin, June 1999.

[LIU 95] Z. Liu. Performance bounds for stochastic timed Petri nets. In *Proc. of the 16th International Conference on Application and Theory of Petri Nets*, number 935 in LNCS, pages 316–334, Turin, Italy, June 26–30 1995. Springer–Verlag.

[MEY 80] J.F. Meyer. On evaluating the performability of degradable computing systems. *IEEE Transactions on Computers*, 29(8):720–731, August 1980.

[MOL 81] M. K. Molloy. *On the integration of delay and throughput in distributed processing models*. PhD dissertation, University of California, Los Angeles, CA, USA, September 1981.

[NEM 89] G.L. Nemhauser, A.H.G. Rinnooy Kan and M.J. Todd, eds. *Optimization*, volume 1 of *Handbooks in Operations research and Management Science*. North-Holland, Amsterdam, 1989.

[NEU 81] M. F. Neuts. *Matrix-geometric solutions in stochastic models - an algorithmic approach*. The John Hopkins University Press, London, 1981.

[REI 98a] W. Reisig and G. Rozenberg, eds. *Lectures on Petri Nets I: Basic models*. Number 1491 in LNCS. Springer–Verlag, June 1998. Advances in Petri nets.

[REI 98b] W. Reisig and G. Rozenberg, eds. *Lectures on Petri Nets II: Applications*. Number 1492 in LNCS. Springer–Verlag, June 1998. Advances in Petri nets.

[SER 93] M. Sereno and G. Balbo. Computational algorithms for product form solution stochastic Petri nets. In *Proc. of the 5th International Workshop on Petri Nets and Performance Models*, pages 98–107, Toulouse, France, October 19–22 1993. IEEE Computer Society Press.

[SIL 97] M. Silva and E. Teruel. Petri nets for the design and operation of manufacturing systems. *European journal of control*, 3(3), 1997.

[STE 94] W. J. Stewart. *Introduction to the numerical solution of Markov chains*. Princeton University Press, USA, 1994.

[STI 74] S. Stidham. A last word on $L = \lambda W$. *Operations research*, 22(2):417–421, 1974.

[TRI 82] K. S. Trivedi. *Probability & statistics with reliability, queueing, and computer science applications*. Prentice Hall, Englewood Cliffs, NJ, USA, 1982.

[TRI 92] K. S. Trivedi, J. K. Muppala, S. P. Woole and B. R. Haverkort. Composite performance and dependability analysis. *Performance Evaluation*, 14(3–4):197–215, February 1992.

[VAL 97] R. Valette. Some issues about Petri net application to manufacturing and process supervisory control. In *Proc. of the 18th International Conference on Application and Theory of Petri Nets*, number 1248 in LNCS, pages 23–41, Toulouse, France, June 23-27 1997. Springer–Verlag.

Index