Chapter 8

Performance Measures and Basic Properties

A goal of performance modelling with timed Petri nets (TPN's) is the estimation of some quantifiable performance measures of the system under consideration by the simulation or analysis of the model of the system behaviour. In order to do that, *responsiveness* and *utilization* performance measures of the system must be described in terms of average values of operational quantities defined on the TPN model, like the *average marking* of a place or the *firing frequency* of a transition.

In Section 8.1, the usual average performance indices for TPN models are derived operationally from the basic observable events that were defined in Chapter 3.

Sections 8.2 and 8.3 are devoted to exploit the operational approach for the definition of performance measures. In Section 8.2, *operational analysis* techniques are used to partially characterize the behaviour of TPN models. Classical operational laws, like *Little's law* are stated in Petri net terms. Additionally, some *operational inequalities* are derived that are typical of the presence of synchronization and that have not been considered before in the framework of queueing networks models.

Section 8.3 illustrates another aspect of the deep gap, from the analytical point of view, existing between classical monoclass product-form queueing networks and TPN's. While in *Gordon-Newell* networks the vector of *visit ratios* to the stations can be easily (efficiently) computed from the routing information, in the case of bounded TPN's the computability of such index is a quite complex problem. This problem leads to a classification of net models attending to the dependency of the visit ratios on the net structure and the probabilistic routing, on the initial marking, and on the firing time of transitions.

Since few operational analysis techniques can presently be used to compute the performance indices of interest, the classical approach based on *stochastic processes* must be considered. Section 8.4 introduces also the definition of performance parameters on the TPN model as expected values of the random variables that characterize the behaviour of the model, under some stochastic assumptions. As the reader surely knows, under *ergodicity* conditions the average (operationally defined) performance indices are equal to the expected (stochastically defined) indices, thus the techniques that will be presented later on for the computation of expected values will be useful to compute the average performance measures of interest.

Some bibliographic remarks are summarized in Section 8.5.

8.1 Performance Measures Defined Operationally

The basic idea for the operational definition of average performance indices is to derive them algebraically from the atomic events observed in the behaviour of the TPN model. For instance, from the instantaneous marking in place p at time τ, $\boldsymbol{\mu}[p](\tau)$, the *average marking* in place p during the experiment interval $(0, \theta)$ is defined as[1]:

$$\overline{\boldsymbol{\mu}}[p] = \frac{1}{\theta} \int_0^\theta \boldsymbol{\mu}[p](\tau)\; d\tau$$

Let us assume in the rest of this chapter an *infinite server semantics* for the timed transitions. The *instantaneous enabling degree* of transition t at instant τ represents the internal concurrency of that transition at time τ:

$$\mathbf{e}[t](\tau) = \sup\{k \in \mathbb{N} : \forall p \in {}^\bullet t, \boldsymbol{\mu}[p](\tau) \geq k \; \mathbf{Pre}[p,t]\}$$

The following relation holds by definitions:

$$\mathbf{e}[t](\tau) = \min_{p \in {}^\bullet t} \left\lfloor \frac{\boldsymbol{\mu}[p](\tau)}{\mathbf{Pre}[p,t]} \right\rfloor, \quad \forall t \in T, \forall \tau \tag{8.1}$$

The *average enabling degree* of transition t during the interval $(0, \theta)$ is defined as:

$$\overline{\mathbf{e}}[t] = \frac{1}{\theta} \int_0^\theta \mathbf{e}[t](\tau)\; d\tau$$

The average enabling degree represents the average number of active servers associated with the transition firing during the experiment.

In cases where transition t is persistent (i.e., never disabled before firing) or t has *age memory* policy, the *average firing time* of transition t, $\overline{\mathbf{s}}[t]$, can be expressed as the quotient between the *total enabling work* of that transition and the total number of firings, $F_t(\theta)$, of t observed from 0 to θ (assuming

[1]The reader should notice that we write $\overline{\boldsymbol{\mu}}[p]$ for short, while for being more precise we should write $\overline{\boldsymbol{\mu}}[p](\theta)$, since the average is computed between 0 and θ.

$F_t(\theta) > 0$). The total enabling work of transition t is obtained by integration over the experiment interval of the instantaneous enabling degree of t at instant τ, $\mathbf{e}[t](\tau)$.

$$\overline{\mathbf{s}}[t] = \frac{\int_0^\theta \mathbf{e}[t](\tau)\, d\tau}{F_t(\theta)}$$

The *throughput* of transition t is the firing frequency of that transition observed during the interval $(0, \theta)$:

$$\boldsymbol{\chi}[t] = \frac{F_t(\theta)}{\theta}$$

With queueing systems terminology, the *relative throughput* between transitions is expressed by the vector of *visit ratios*, normalized, for instance, for transition t_1:

$$\mathbf{v}[t] = \frac{\boldsymbol{\chi}[t]}{\boldsymbol{\chi}[t_1]}$$

The quantity $\mathbf{v}[t]$ represents the number of times that transition t is fired per each firing of trasition t_1 (in queueing systems terminology, it represents the number of "visits" of a customer to station t per each visit to the reference station t_1). A detailed analysis of the computability of visit ratios for different net classes is presented in Section 8.3.

If $\alpha(p, i)$ denotes the arrival time of the i-th token in p, then the *total number of tokens visiting place* p during the experiment interval is:

$$N_p(\theta) = \sum_{i=1}^{\infty} \#(\alpha(p, i) \leq \theta)$$

where for an arbitrary predicate P, the function $\#$ is defined as:

$$\#(P) = \begin{cases} 1 & \text{if } P = \text{true} \\ 0 & \text{otherwise} \end{cases}$$

If $\delta(p, i)$ denotes the departure time of the i-th token from p, then for all place p such that $N_p(\theta) > 0$, the *average token residence time* in place p during the interval $(0, \theta)$ is the following:

$$\overline{\mathbf{r}}[p] = \frac{1}{N_p(\theta)} \sum_{k=1}^{N_p(\theta)} \int_0^\theta (\#(\alpha(p, k) \leq \tau) - \#(\delta(p, k) < \tau))\, d\tau$$

An alternative definition of the above indices may be given in a unified way in terms of *reward functions*, or index functions defined over the reachable markings of the PN system. First, let us denote by $\pi_i(\theta)$ the *proportion of time spent by the system in a given reachable marking,* $\mathbf{m_i}$*, during the interval* $(0, \theta)$:

$$\pi_i(\theta) = \frac{1}{\theta} \int_0^\theta \#(\boldsymbol{\mu}(\tau) = \mathbf{m_i})\ d\tau$$

A reward function over reachable markings is a function:

$$\begin{array}{rccc} r: & \mathrm{RS} & \longrightarrow & \mathrm{I\!R}^+ \\ & \mathbf{m} & \leadsto & r(\mathbf{m}) \end{array}$$

From that function, an *operational average reward* between 0 and θ can be computed as the weighted sum:

$$\overline{r}(\theta) = \sum_{\mathbf{m_i} \in \mathrm{RS}} r(\mathbf{m_i})\ \pi_i(\theta)$$

Different average performance indices can be computed as operational average rewards by giving different interpretations to the reward function. For instance, the *proportion of time during the experiment interval that a given predicate on the marking, $P(\mathbf{m})$, holds* can be defined using the following reward function:

$$r(\mathbf{m}) = \begin{cases} 1 & \text{if } P(\mathbf{m}) = \text{true} \\ 0 & \text{otherwise} \end{cases}$$

The average marking in place p during the interval $(0, \theta)$ can be redefined using the following reward function:

$$r(\mathbf{m}) = n \text{ if and only if } \mathbf{m}[p] = n$$

The equivalence of both definitions of average marking can be easily shown:

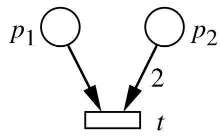
$$\begin{aligned} \overline{r}(\theta) &= \sum_{\mathbf{m_i} \in \mathrm{RS}} r(\mathbf{m_i})\ \pi_i(\theta) = \sum_{\mathbf{m_i} \in \mathrm{RS}} \mathbf{m_i}[p]\ \frac{1}{\theta} \int_0^\theta \#(\boldsymbol{\mu}(\tau) = \mathbf{m_i})\ d\tau = \\ &= \frac{1}{\theta} \int_0^\theta \sum_{\mathbf{m_i} \in \mathrm{RS}} \mathbf{m_i}[p]\ \#(\boldsymbol{\mu}(\tau) = \mathbf{m_i})\ d\tau = \frac{1}{\theta} \int_0^\theta \boldsymbol{\mu}[p](\tau)\ d\tau = \\ &= \overline{\boldsymbol{\mu}}[p](\theta) \end{aligned}$$

In a similar way, other average performance indices can be defined as operational average rewards.

8.2 From Performance Measures to Operational Laws

Operational analysis is a conceptually very simple way of deriving mathematical equations relating observable quantities in queueing systems. In this section operational analysis techniques are applied to derive linear equations and inequalities relating interesting performance measures in timed Petri net models.

The first result is the *enabling law* which is the Petri net counterpart of the well-known "utilization law" in queueing systems literature.



$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] = \overline{\mathbf{e}}[t] \leq \left(\overline{\boldsymbol{\mu}}[p_1], \frac{\overline{\boldsymbol{\mu}}[p_2]}{2}\right)$$

Figure 8.1: Maximum throughput law.

Property 8.1 (Enabling Law) *For all transition $t \in T$ with at least one input place and of type either persistent or age memory or immediate and for all experiment interval $(0, \theta)$ with $\theta > 0$ and $F_t(\theta) > 0$*

$$\overline{\mathbf{e}}[t] = \boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t]$$

Proof Sketch:
Trivial substitution of the definitions. ♢

From the enabling law it follows that if the average firing time of a transition is known, then its throughput is proportional to its average enabling degree. Of course in case of immediate transitions $\overline{\mathbf{s}}[t] = 0$, so immediate transitions are never enabled for non-null intervals of time.

The main conceptual difference between *basic queueing networks* (using the terminology introduced in Chapter 1) and Petri net models (or *extended queueing networks*) is the presence of a synchronization primitive in the latter, therefore the following operational inequalities derived for synchronization elements have no counterpart in operational laws for basic queueing networks.

Property 8.2 (Maximum Throughput Law) *For all transition $t \in T$ with at least one input place and of type either persistent or age memory or immediate and for all experiment interval $(0, \theta)$ with $\theta > 0$ and $F_t(\theta) > 0$*

$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] \leq \min_{p \in {}^\bullet t} \left(\frac{\overline{\boldsymbol{\mu}}[p]}{\mathbf{Pre}[p,t]} \right)$$

Proof:
Equation (8.1), that is valid in each instant of the experiment, implies that $\forall p \in {}^\bullet t$, $\forall \tau : 0 \leq \tau \leq \theta$, $\mathbf{e}[t](\tau) \leq \frac{\boldsymbol{\mu}[p](\tau)}{\mathbf{Pre}[p,t]}$.

Therefore $\forall p \in {}^\bullet t$, $\overline{\mathbf{e}}[t] \leq \frac{\overline{\boldsymbol{\mu}}[p]}{\mathbf{Pre}[p,t]}$, and applying the enabling law the result follows. ♢

This inequality establishes an upper bound for the average enabling (hence for the transition throughput once the firing time is defined) in the case of transitions with more than one input place that model a synchronization. As depicted in figure 8.1, the instantaneous enabling degree of a transition cannot be bigger than the marking of its input places (divided by the arc weight), therefore the same applies for the average values.

Property 8.3 (Minimum Throughput Law for single input transitions) *For all transition $t \in T$ with exactly one input place p and of type either persistent or age memory or immediate,*

$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] \geq \frac{\overline{\boldsymbol{\mu}}[p] - \mathbf{Pre}[p,t] + 1}{\mathbf{Pre}[p,t]} \tag{8.2}$$

Proof:
First define some auxiliary punctual marking functions:

$$\forall p \in P, \ \ \forall \tau, \quad \boldsymbol{\mu}^v[p](\tau) \ = \ \max(0, \boldsymbol{\mu}[p](\tau) - v)$$

$$\forall p \in P, \ \ \forall \tau, \quad \boldsymbol{\mu}_l^u[p](\tau) \ = \ \boldsymbol{\mu}^l[p](\tau) - \boldsymbol{\mu}^u[p](\tau)$$

Consider now the following properties of the auxiliary function $\forall k \in \mathbb{N} : k > 0$,

$$0 \ \leq \ \boldsymbol{\mu}_{kw-1}^{(k+1)w-1}[p](\tau) \ \leq \ w$$

Moreover, notice that the k-th server in transition t is enabled if and only if $\boldsymbol{\mu}_{kw-1}^{(k+1)w-1}[p](\tau) \geq 1$ in case $w = \mathbf{Pre}[p,t]$. Therefore,

$$\forall t \in T : {}^{\bullet}t = \{p\}, \quad \forall k \geq 1, \quad \forall \tau, \quad \mathbf{e_k}[t](\tau) \ \geq \ \frac{1}{\mathbf{Pre}[p,t]} \ \boldsymbol{\mu}_{k\mathbf{Pre}[p,t]-1}^{(k+1)\mathbf{Pre}[p,t]-1}[p](\tau)$$

where $\mathbf{e_k}[t](\tau)$ is the instantaneous enabling of k-th server in t, i.e., the characteristic function that evaluates to 1 if and only if the k-th server in transition t is busy at time τ ($\mathbf{e_k}[t](\tau) =$ if $\mathbf{e}[t](\tau) \geq k$ then 1 else 0).

Hence, $\forall t \in T : {}^{\bullet}t = \{p\}$:

$$\forall \tau, \ \mathbf{e}[t](\tau) \ \geq \ \frac{1}{\mathbf{Pre}[p,t]} \sum_{k=1}^{\infty} \boldsymbol{\mu}_{k\mathbf{Pre}[p,t]-1}^{(k+1)\mathbf{Pre}[p,t]-1}[p](\tau) \ = \ \frac{\boldsymbol{\mu}[p](\tau) - \mathbf{Pre}[p,t] + 1}{\mathbf{Pre}[p,t]}$$

Finally, taking the average over the experiment interval and applying the enabling law, the result follows. $\diamondsuit$

Observe that in the case that the right-hand expression in (8.2) is negative, a trivial inequality holds: $\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] \geq 0$.

Corollary 8.4 (Throughput Law for ordinary, single input transitions) *For all transition $t \in T$ with exactly one input place p and of type either persistent or age memory or immediate, and such that* $\mathbf{Pre}[p,t] = 1$*,*

$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] = \overline{\boldsymbol{\mu}}[p]$$

Proof Sketch:
Combining Properties 8.2 and 8.3. $\diamondsuit$

The above result is nothing more than the well-known expression for the average number of customers at delay stations in queueing networks terminology (remember that since we are considering infinite server semantics, the case of a transition with only one input place with ordinary incidence arc models a delay server).

In the particular case of bounded nets, the above minimum throughput law can be improved, as stated in the next results.

Property 8.5 (Minimum Throughput Law for single input transitions in bounded nets) *For all transition $t \in T$ with exactly one input place p and of type either persistent or age memory or immediate, if $\forall \tau$: $\boldsymbol{\mu}[p](\tau) \leq b_p$ and $\exists k \in \mathbb{N} : k\ \mathbf{Pre}[p,t] \leq b_p < (k+1)\mathbf{Pre}[p,t]$, then*

$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] \geq \ k\ \frac{\overline{\boldsymbol{\mu}}[p] - k\ \mathbf{Pre}[p,t] + 1}{b_p - k\ \mathbf{Pre}[p,t] + 1}$$

Proof:
Firstly note that $\forall t \in T$, $\forall j \in \mathbb{N}$

$$\overline{\mathbf{e}}[t] \geq \ j\ \frac{1}{\theta} \int_0^\theta \mathbf{e_j}[t](\tau)\ d\tau$$

where $\mathbf{e_j}[t](\tau)$ is defined as in the proof of Property 8.3.

Secondly, note that the marking in the input place p can be expressed as the sum of two components:

$$\forall \tau, \quad \boldsymbol{\mu}[p](\tau) \ = \ \mathbf{Pre}[p,t] \sum_{j=1}^{\infty} \mathbf{e_j}[t](\tau) \ + \boldsymbol{\nu}[p](\tau)$$

where the component $\boldsymbol{\nu}[p](\tau) \leq \mathbf{Pre}[p,t] - 1$ represents the portion of marking not used to enable the transition. Now taking the integral and dividing by θ one obtains:

$$\overline{\boldsymbol{\mu}}[p] \ = \ \mathbf{Pre}[p,t]\ \overline{\mathbf{e}}[t] + \overline{\boldsymbol{\nu}}[p]$$

This equation shows that the average enabling depends only on the mean values of the input place marking and of the unused portion of the marking.

The worst case from the point of view of enabling the transition k times occurs when the place is marked with $\mathbf{Pre}[p,t]k - 1$ tokens most of the time and with b_p tokens for the rest of the time, since this case maximizes the unused portion of the average marking in the input place. From these considerations the result follows. $\diamondsuit$

Property 8.6 (Minimum Throughput Law for binary synchronizations with ordinary arcs in bounded nets) *For all transition $t \in T$ of type either persistent or age memory or immediate with ${}^\bullet t = \{p_1, p_2\}$ and $\mathbf{Pre}[p_1,t] = \mathbf{Pre}[p_2,t] = 1$, if $\forall \tau$: $\boldsymbol{\mu}[p_1](\tau) \leq b_{p_1}$ and $\boldsymbol{\mu}[p_2](\tau) \leq b_{p_2}$, then*

$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] \geq \ \overline{\boldsymbol{\mu}}[p_1] + \frac{b_{p_1}}{b_{p_2}} \overline{\boldsymbol{\mu}}[p_2] - b_{p_1}$$

Proof:
Similarly to the previous case, two equations can be written relating the average marking, the average enabling, and the average portion of unused marking for each of the two input places:

$$\begin{aligned}\overline{\mathbf{e}}[t] &= \overline{\boldsymbol{\mu}}[p] - \overline{\boldsymbol{\nu}}[p_1] \\ \overline{\mathbf{e}}[t] &= \overline{\boldsymbol{\mu}}[p_2] - \overline{\boldsymbol{\nu}}[p_2]\end{aligned}$$

Now, upper bounds on the unused part of the marking can be computed as follows. The maximum fraction of time during which $\boldsymbol{\nu}[p_1](\tau)$ may be greater than zero is equal to the minimum time during which $\boldsymbol{\mu}[p_2](\tau) = 0$ (otherwise the transition would be enabled and the marking of p_1 would contribute to the enabling instead); since place p_2 is b_{p_2}–bounded, this fraction of time is less than or equal to $1 - \frac{\overline{\boldsymbol{\mu}}[p_2]}{b_{p_2}}$; moreover during this maximum time, the maximum value of the marking in p_1 is less than or equal to b_{p_1}. Hence

$$\overline{\boldsymbol{\nu}}[p_1] \;\leq\; b_{p_1} \left(1 - \frac{\overline{\boldsymbol{\mu}}[p_2]}{b_{p_2}}\right)$$

and from this the result trivially follows. ♢

Property 8.7 (Minimum Throughput Law in bounded nets) *For all transition $t \in T$ of type either persistent or age memory or immediate with ${}^{\bullet}t = \{p_1, p_2, \ldots, p_k\}$, if $\forall j : 1 \leq j \leq k$, $\forall \tau :$ $\boldsymbol{\mu}[p_j](\tau) \leq b_{p_j}$ and $b_{p_1} \leq b_{p_j}$, then*

$$\boldsymbol{\chi}[t] \cdot \overline{\mathbf{s}}[t] \geq \frac{\overline{\boldsymbol{\mu}}[p_1] - \mathbf{Pre}[p_1, t] + 1 - b_{p_1} \ \max_j(f_j)}{\mathbf{Pre}[p_1, t]}$$

where $\forall j : 2 \leq j \leq k,\ f_j = 1 - \frac{\overline{\boldsymbol{\mu}}[p_j] - \mathbf{Pre}[p_j, t] + 1}{b_{p_j} - \mathbf{Pre}[p_j, t] + 1}$

Proof Sketch:
Similar to the previous ones writing the upper bound for the quantity $\overline{\boldsymbol{\nu}}[p_1]$. ♢

The next property is the analogous of Little's law in queueing systems.

Property 8.8 (Little's Law) *For all place $p \in P$:*

1. *For all experiment interval $(0, \theta)$ with $\theta > 0$ and $F_t(\theta) > 0$:*

$$\overline{\mathbf{r}}[p] \left(\frac{\mathbf{m_0}[p]}{\theta} + \sum_{t \in {}^{\bullet}p} \mathbf{Post}[p, t] \boldsymbol{\chi}[t] \right) = \overline{\boldsymbol{\mu}}[p]$$

Notice that, even if it is not made explicit in the notation, $\overline{\mathbf{r}}[p]$, $\boldsymbol{\chi}[t]$, and $\overline{\boldsymbol{\mu}}[p]$ are functions of the duration of the experiment interval θ.

2. If the limits $\overline{\boldsymbol{\mu}}^*[p] = \lim_{\theta\to\infty} \overline{\boldsymbol{\mu}}[p]$ *and* $\boldsymbol{\chi}^*[t] = \lim_{\theta\to\infty} \boldsymbol{\chi}[t]$ *exist, then:*

$$\overline{\mathbf{r}}^*[p] = \lim_{\theta\to\infty} \overline{\mathbf{r}}[p] = \frac{\overline{\boldsymbol{\mu}}^*[p]}{\sum_{t\in{}^\bullet p} \mathbf{Post}[p,t]\boldsymbol{\chi}^*[t]}$$

Proof Sketch:
It follows taking into consideration the corresponding definitions and exchanging sum and integral signs in the definition of average token residence time. $\diamondsuit$

8.3 Visit Ratios Computability: A Net Hierarchy

In a classical monoclass queueing network, the following system can be derived by equating the rate of *flow of customers* into each station to the rate of flow out of the station:

$$\chi_j = \chi_{0j} + \sum_{i=1}^{m} \chi_i \; r_{ij}, \qquad j = 1, \ldots, m$$

where χ_i is the limit *throughput* of station i, i.e., the average number of service completions per unit time at station i, $i = 1, \ldots, m$; r_{ij}, is the probability that a customer exiting center i goes to j ($i, j = 1, \ldots, m$); and χ_{0j} is the external arrival rate of customers to station j ($j = 1, \ldots, m$), i.e., the average number of customers arriving at station j per unit of time.

If the network is open (i.e., if there exists a station j with positive external arrival rate, $\chi_{0j} > 0$ and also customers can leave the system), then the above m equations are linearly independent, and the exact throughput of stations can be derived (independently of the service times, s_i, $i = 1, \ldots, m$). This is not the case for closed networks. If $\chi_{0j} = 0$, $j = 1, \ldots, m$, then only $m - 1$ equations are linearly independent, and thus only relative throughputs can be determined, i.e., the *visit ratios*, denoted as v_i, for each station i.

Unfortunately, concerning timed Petri net models, the introduction of synchronization schemes can lead to the "pathological" behaviour of models reaching a total deadlock, thus with null visit ratios for all transitions, in the limit. In other words, for these models it makes no sense to speak about steady-state behaviour. Therefore, in the rest of this chapter only deadlock-free Petri nets are considered. Even more, in most systems that will be studied, deadlock-freeness implies liveness of the net, in other words, the existence of an infinite activity of all the transitions is assured.

8.3.1 The Visit Ratios Computation Problem

The counterpart of routing of customers in queueing networks consists both on the *net structure* $\mathcal{N}$ and the relative *routing rates at conflicts*, denoted $\mathbf{R}$, in

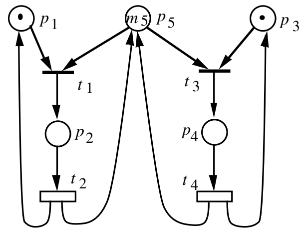


Figure 8.2: A net system whose visit ratios depend on the structure, on the routing at conflicts, on the initial marking, and on the firing times.

strongly connected marked graphs	$\mathbf{v} = \mathbf{1}$ {constant}
mono-T-semiflow nets	$\mathbf{v} = \mathrm{f}(\mathcal{N})$
live and bounded free choice nets	$\mathbf{v} = \mathrm{f}(\mathcal{N}, \mathbf{R})$
simple nets	$\mathbf{v} = \mathrm{f}(\mathcal{N}, \mathbf{R}, \mathbf{m_0}, \overline{\mathbf{s}})$

Table 8.1: Computability of the vector of visit ratios and net subclasses.

timed Petri nets. Unfortunately, in the general Petri net case it is not possible to derive the visit ratios only from $\mathcal{N}$ and $\mathbf{R}$. Net systems can be constructed such that the visit ratios for transitions do depend on the net structure, on the routing rates at conflicts, but also on the initial marking (distribution of customers), and on the distribution firing time (thus, in particular, on the average firing time) of transitions:

$$\mathbf{v} = \mathrm{f}(\mathcal{N}, \mathbf{R}, \mathbf{m_0}, \overline{\mathbf{s}})$$

where $\mathbf{v}$ is normalized, for instance, for transition t_1.

As an example, let us consider the net system depicted in Figure 8.2. Transitions t_1 and t_3 are *immediate* (i.e., they fire in zero time). The constants $r_1, r_3 \in \mathbb{N}^+$ define the conflict resolution policy, i.e., when t_1 and t_3 are simultaneously enabled, t_1 fires with relative rate $r_1/(r_1+r_3)$ and t_3 with $r_3/(r_1+r_3)$. Let s_2 and s_4 be the average firing times of t_2 and t_4, respectively. If $m_5 = 1$ (initial marking of p_5) then p_1 and p_3 are *implicit*, hence they can be deleted without affecting the behaviour! Thus a closed queueing network topology is derived. A product form queueing network can be obtained and the visit ratios, normalized for transition t_1 can be computed: $\mathbf{v} = (1, 1, r_3/r_1, r_3/r_1)$. If $m_5 = 2$ (different initial marking for p_5) then p_5 is now implicit, hence it can be deleted; two isolated closed tandem queueing networks are obtained and $\mathbf{v}' = (1, 1, s_2/s_4, s_2/s_4)$. Obviously $\mathbf{v} \neq \mathbf{v}'$, in general.

In this section, we consider those nets whose associated vector of visit ratios for transitions can be computed from the net structure and the routing rates at conflicts. Since the existence of strictly positive visit ratios requires liveness of the net and we are looking for an structural computation, the presentation is restricted to *structurally live* Petri nets. On the other hand, unless otherwise explicitly stated, *structurally bounded* nets are considered. Under these

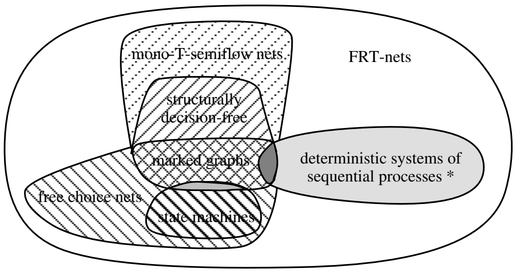


Figure 8.3: Inclusion relations among FRT-net subclasses (* these are marked nets).

restrictions, a characterization of the class of nets for which the vector of visit ratios can be computed *without any behavioural analysis* (i.e., from structure and routing rates at conflicts) in terms of a *rank condition* over the incidence matrix of the net is presented. These nets are defined as having *freely related T-semiflows* and denoted, for short, as *FRT-nets*. This means that they can have several independent T-semiflows but the vector of visit ratios, which is always a linear combination of the minimal T-semiflows, is computed as an "average T-semiflow" from local free choices among transitions, governed by the routing rates.

A particular class of FRT-nets is that of *mono-T-semiflow nets*, which have a unique minimal T-semiflow. They include *choice-free nets* (if strongly connected and consistent) and, in particular, the well-known net subclass of *marked graphs*.

FRT-nets include also *free choice nets*. This last class includes marked graphs, thus intersecting with mono-T-semiflow nets. *State machines* are also included in the class of free choice nets and constitute the Petri net counterpart of monoclass queueing networks.

Another well-known subclass of marked FRT-nets is that of *deterministic systems of sequential processes*, that is, 1–bounded state machines communicating through private buffers non-disturbing local decisions at state machines. In fact, *FRT-nets communicating through private buffers non-disturbing local decisions* are also FRT-nets. In this sense, FRT-nets can be recursively defined.

Inclusion relations among above mentioned subclasses are depicted in Figure 8.3.

8.3.2 FRT-Nets

In the following paragraphs the class of structurally live and structurally bounded nets whose vector of visit ratios for transitions can be computed just from the net structure and the routing rates at conflicts is defined. Nets belonging to this subclass are said to have *freely related T-semiflows* and called *FRT-nets*.

Before giving a formal definition, let us remark that the vector of visit ratios for transitions of any net should verify the two following conditions:

1. The vector of visit ratios $\mathbf{v}$ (normalized, for instance, for transition t_1) must be a non-negative right annuller of the incidence matrix:

$$\mathbf{C} \cdot \mathbf{v} = \mathbf{0}$$

2. The components of $\mathbf{v}$ must verify the following relations with respect to the routing rates for each subset of transitions $T_i = \{t_1, \ldots, t_k\} \subset T$ in *generalized free (or equal) conflict* (i.e., having equal pre-incidence function: $\mathbf{Pre}[\cdot, t_1] = \ \ldots \ = \mathbf{Pre}[\cdot, t_k]$):

$$\begin{array}{rcl} r_2\mathbf{v}[t_1] - r_1\mathbf{v}[t_2] & = & 0 \\ r_3\mathbf{v}[t_2] - r_2\mathbf{v}[t_3] & = & 0 \\ & \cdots & \\ r_k\mathbf{v}[t_{k-1}] - r_{k-1}\mathbf{v}[t_k] & = & 0 \end{array}$$

   Expressing the former homogeneous system of equations in matrix form: $\mathbf{R}[T_i] \cdot \mathbf{v} = \mathbf{0}$, where $\mathbf{R}[T_i]$ is a $(k-1) \times m$ matrix. Now, by considering all generalized free conflicts $T_1, \ldots, T_r$: $\mathbf{R} \cdot \mathbf{v} = \mathbf{0}$, where $\mathbf{R}$ is a matrix:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}[T_1] \\ \vdots \\ \mathbf{R}[T_r] \end{pmatrix}$$

$\mathbf{R}$ is a *routing matrix* with δ rows and $m = |T|$ columns, where δ is the difference between the number of transitions in generalized free conflict and the number of subsets of transitions in generalized free conflict ($\delta < m$) or, in other words, the number of independent relations fixed by the routing rates at conflicts. Given that $r_i \neq 0$ for all i, it can be observed that, by construction, $\text{rank}(\mathbf{R}) = \delta$. The above remarked conditions together with the normalization constraint for transition t_1, $\mathbf{v}[t_1] = 1$, characterize a unique vector if and only if the number of independent rows of the matrix

$$\begin{pmatrix} \mathbf{C} \\ \mathbf{R} \end{pmatrix}$$

is $m - 1$.

The class of structurally live and structurally bounded nets verifying the previous rank condition is introduced now. In order to do that, an equivalence relation on the set of T-semiflows of the net is defined. After that, the class of FRT-nets will be defined as nets having only one equivalence class for this relation.

Definition 8.9 (Freely connected T-semiflows) *Let $\mathcal{N}$ be a Petri net and $\mathbf{x_a}, \mathbf{x_b}$ two different T-semiflows of $\mathcal{N}$. $\mathbf{x_a}$ and $\mathbf{x_b}$ are said to be freely connected by places $P' \subset P$, denoted as $\mathbf{x_a} \overset{P'}{\wedge} \mathbf{x_b}$, iff $\exists t_a \in \|\mathbf{x_a}\|, t_b \in \|\mathbf{x_b}\|$ such that: $\mathbf{Pre}[\cdot, t_a] = \mathbf{Pre}[\cdot, t_b]$ and ${}^\bullet t_a = {}^\bullet t_b = P'$.*

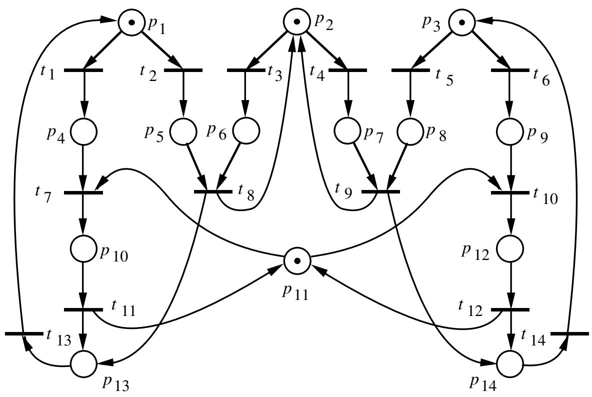


Figure 8.4: A live and structurally bounded FRT-net.

Definition 8.10 (Freely related T-semiflows) *Let $\mathcal{N}$ be a Petri net and $\mathbf{x_a}$, $\mathbf{x_b}$ two T-semiflows of $\mathcal{N}$. $\mathbf{x_a}$ and $\mathbf{x_b}$ are said to be freely related, denoted as $\langle \mathbf{x_a}, \mathbf{x_b} \rangle \in \mathrm{FR}$, iff one of the following conditions holds:*

1. $\mathbf{x_a} = \mathbf{x_b}$,
2. $\exists P' \subset P$ *such that* $\mathbf{x_a} \overset{P'}{\wedge} \mathbf{x_b}$, *or*
3. $\exists \mathbf{x_1}, \ldots, \mathbf{x_k}$ *T-semiflows of $\mathcal{N}$ and* $P_1, \ldots, P_{k+1} \subset P$, $k \geq 1$, *such that* $\mathbf{x_a} \overset{P_1}{\wedge} \mathbf{x_1} \overset{P_2}{\wedge} \ldots \overset{P_k}{\wedge} \mathbf{x_k} \overset{P_{k+1}}{\wedge} \mathbf{x_b}$.

From the above definition the next property trivially follows:

Property 8.11 FR *is an equivalence relation on the set of T-semiflows of a net.*

The introduction of this equivalence relation on the set of T-semiflows induces a partition into equivalence classes. FRT-nets are defined as follows:

Definition 8.12 (FRT-nets) *A Petri net $\mathcal{N}$ is a net with freely related T-semiflows (FRT-net, for short) iff the introduction of the freely relation on the set of its T-semiflows induces only one equivalence class.*

Note that FRT-nets are necessarily connected. Therefore, in what follows, unless otherwise explicitly stated, we consider only connected nets.

As an example, let us consider the net depicted in Figure 8.4. It is a live and structurally bounded net. Its minimal T-semiflows are:

$$\begin{aligned} \mathbf{x_1} &= (1,0,0,0,0,0,1,0,0,0,1,0,1,0) \\ \mathbf{x_2} &= (0,1,1,0,0,0,0,1,0,0,0,0,1,0) \\ \mathbf{x_3} &= (0,0,0,1,1,0,0,0,1,0,0,0,0,1) \\ \mathbf{x_4} &= (0,0,0,0,0,1,0,0,0,1,0,1,0,1) \end{aligned} \tag{8.3}$$

Then, the net is an FRT-net because:

$$\mathbf{x_1} \overset{\{p_1\}}{\wedge} \mathbf{x_2} \overset{\{p_2\}}{\wedge} \mathbf{x_3} \overset{\{p_3\}}{\wedge} \mathbf{x_4}$$

From the definition of FRT-nets, it may appears that a direct checking of the membership of a given net to this net subclass is not a polynomial problem on the net size. This is because the number of T-semiflows of a net can grow exponentially with the number of places and transitions. However, if structural liveness and structural boundedness are assumed, a nice characterization of the FRT-nets subclass can be obtained and checked in polynomial time. Before the presentation of that result, a second equivalence relation, now on the set of transitions of the net, is introduced.

Definition 8.13 (Equality conflict relation) *Two transitions t_a and t_b are said to be in equality conflict relation, denoted by $\langle t_a, t_b \rangle \in$ ECR, iff $\mathbf{Pre}[\cdot, t_a] = \mathbf{Pre}[\cdot, t_b]$.*

Since the equality conflict relation is based on the equality of vectors, the next property follows:

Property 8.14 ECR *is an equivalence relation on the set of transitions.*

Each equivalence class will be called *equality conflict set*, ECS for short. Let D be an ECS, the number $\delta_D = |D| - 1$ is called *number of non-redundant free conflicts of D*. The reason of the name lies on the fact that δ_D is exactly the number of independent relations among the throughput of transitions belonging to D that can be derived from the routing rates defining the resolution of the conflict. The *number of non-redundant free conflicts of a net*, denoted as δ, is the sum of all δ_D corresponding to the ECS's of the net: $\delta = \sum_{D \in T/ECR} \delta_D$.

Theorem 8.15 *Let $\mathcal{N}$ be a structurally live and structurally bounded net. Then $\mathcal{N}$ is a FRT-net if and only if $\text{rank}(\mathbf{C}) = m - \delta - 1$, where $\mathbf{C}$ is the incidence matrix of $\mathcal{N}$, $m = |T|$, and δ is the number of non-redundant free conflicts of the net.*

Due to its considerable extension, we do not include here the proof of the above Theorem. The interested reader is referred to the bibliography.

An important conclusion is stated in the next Corollary.

Corollary 8.16 *If $\mathcal{N}$ is structurally live and structurally bounded, deciding if $\mathcal{N}$ belongs to the class of FRT-nets is polynomial on the net size.*

Structural boundedness of a net can always be checked in polynomial time (iff $\exists \mathbf{y} \geq \mathbf{1}$ such that $\mathbf{y} \cdot \mathbf{C} \leq \mathbf{0}$). Unfortunately, structural liveness of FRT-nets cannot be decided (so far) efficiently. Nevertheless, a *necessary condition* for a net to be structurally live structurally bounded and FRT-net can be checked in polynomial time, looking for the consistency, conservativeness, and rank condition over the incidence matrix, because structural liveness and structural boundedness implies consistency and conservativeness.

The next result gives a *polynomial time* method for the computation of the vector of visit ratios for transitions of a live and structurally bounded FRT-net, from the knowledge of the net structure and the routing rates at conflicts.

Theorem 8.17 *Let $\langle \mathcal{N}, \mathbf{m_0} \rangle$ be a live and structurally bounded FRT-net system. Let $\mathbf{C}$ be the incidence matrix of $\mathcal{N}$ and $\mathbf{R}$ the previously introduced routing matrix. Then, the vector of visit ratios $\mathbf{v}$ normalized for transition t_1 can be computed from $\mathbf{C}$ and $\mathbf{R}$ by solving the following linear system of equations:*

$$\begin{pmatrix} \mathbf{C} \\ \mathbf{R} \end{pmatrix} \cdot \mathbf{v} = \mathbf{0}, \quad \mathbf{v}[t_1] = 1 \tag{8.4}$$

Proof:
We only have to check that the above system has a unique solution. By Theorem 8.15, the number of independent rows of matrix $\mathbf{C}$ is $m - \delta - 1$. Therefore, the $m - \delta - 1$ independent conditions given by $\mathbf{C} \cdot \mathbf{v} = \mathbf{0}$ plus the δ independent conditions given by $\mathbf{R} \cdot \mathbf{v} = \mathbf{0}$ plus the normalization condition $\mathbf{v}[t_1] = 1$ are enough to determine exactly the m components of the vector $\mathbf{v}$. ♢

The reader can notice that a *rank condition* over the incidence matrix $\mathbf{C}$ exists underlying Theorem 8.17: the system of equations (8.4) has a unique solution $\mathbf{v}$ if and only if $\text{rank}(\mathbf{C}) = m - \delta - 1$, where δ is the rank of $\mathbf{R}$.

From the above theorem, for structurally live and structurally bounded FRT-nets we have:

$$\mathbf{v} = \mathrm{f}(\mathcal{N}, \mathbf{R})$$

and the next complexity result follows:

Corollary 8.18 *The computation of the vector of visit ratios for transitions of a structurally live and structurally bounded FRT-net is polynomial on the net size.*

As an example, let us consider again the net depicted in Figure 8.4. The vector of visit ratios must be a right annuller of the incidence matrix, hence a linear combination of a basis of T-semiflows:

$$\mathbf{v} = \sum_{i=1}^{4} \alpha_i \mathbf{x_i} \tag{8.5}$$

where $\mathbf{x_i}$, $i = 1, \ldots, 4$, are the minimal T-semiflows (8.3) of the net. If r_1, r_2 are the routing rates of t_1, t_2 in the conflict at p_1; r_3, r_4 the routing rates of t_3, t_4 in the conflict at p_2; and r_5, r_6 the routing rates of t_5, t_6 in the conflict at p_3, then $\mathbf{v}$ must satisfy:

$$\begin{aligned} r_2 \mathbf{v}[t_1] &= r_1 \mathbf{v}[t_2] \\ r_4 \mathbf{v}[t_3] &= r_3 \mathbf{v}[t_4] \\ r_6 \mathbf{v}[t_5] &= r_5 \mathbf{v}[t_6] \end{aligned}$$

And together with the normalization requirement:

$$\mathbf{v}[t_1] = 1 \tag{8.6}$$

the four parameters α_i, $i = 1, \ldots, 4$, can be determined.

Another interesting qualitative property follows from Theorem 8.17, that does not hold for general nets:

Property 8.19 *Let $\mathcal{N}$ be a structurally live and structurally bounded FRT-net. Then $\mathcal{N}$ is live if and only if it is deadlock-free.*

Proof Sketch:
The "only if" direction is always true by definitions of liveness and deadlock-freeness. The other direction follows from Theorem 8.17. The vector of visit ratios for transitions is univocally determined in that theorem and all its components are non-null. If an infinite behaviour of the net always occurs (i.e., if the net is deadlock-free) the limit throughput of transitions must be proportional to the vector of visit ratios, which is strictly positive. Therefore, the net is live. ♢

8.4 Ergodicity and Stochastic Definition of Performance Indices

In Section 8.1, performance measures were defined *operationally* from the basic observable events in the evolution of the model behaviour. In particular, we saw how *transient* performance indices can be defined from the following *time average*:

$$\pi_i(\theta) = \frac{1}{\theta} \int_0^\theta \#(\boldsymbol{\mu}(\tau) = \mathbf{m_i})\ d\tau \tag{8.7}$$

The above value represents the proportion of time spent by the system in a given reachable marking, $\mathbf{m_i}$, during the interval $(0, \theta)$.

If instead of a single execution of the model, an infinite number of sample paths is considered, a *sample average* between 0 and θ, $\tilde{\pi}_i(\theta)$, can be defined as the average of $\pi_i(\theta)$ over all the sample paths.

Even if the above operational definition is the natural way of defining performance indices, few analysis techniques for the computation of such indices are presently known (one of them will be presented in Chapter 17). An alternative (classical) approach is to quantify the ocurrence of the observable events as *random variables*, therefore the sequences of observable events are *stochastic processes* for which, under certain assumptions, efficient analysis techniques exist (they will be presented in Chapters 9, 10, and 11). Thus, the stochastic counterpart of the above time average (8.7) can be defined as:

$$\eta_i(\theta) = \mathrm{E}[\boldsymbol{\mu}(\theta)] \tag{8.8}$$

where $\eta_i(\theta)$ represents the *expected value* (mathematical expectation operator) of $\boldsymbol{\mu}(\theta)$, the state distribution of the system at time θ.

An important fact is that the sample average and the expected value are equal:

$$\eta_i(\theta) = \tilde{\pi}_i(\theta)$$

The above values correspond to the transient behaviour, i.e., to the state of the system in a finite interval $(0, \theta)$. The corresponding limit or *steady-state* behaviours can be defined also in both the operational and the stochastic framework. Concerning the operational framework, the limit $(\theta \to \infty)$ of the time average can be defined as:

$$\pi_i = \lim_{\theta \to \infty} \pi_i(\theta) \tag{8.9}$$

On the other hand, within the stochastic framework, the existence of the following limit expected value can be considered:

$$\eta_i = \lim_{\theta \to \infty} \eta_i(\theta) \tag{8.10}$$

If the system is *ergodic*, the above limit exists and it is independent of the initial state. Moreover, if the system is ergodic the limit time average (8.9) and the limit expected value (8.10) are equal:

$$\pi_i = \eta_i$$

An important consequence of the above equality is that a single (infinite) sample path characterizes the expected behaviour of the system.

The steady-state probability distribution η_i can be used to compute the expected performance measures as average rewards, if the corresponding reward functions are properly defined.

The *expected marking* in place p can be computed as the following average reward:

$$\overline{\boldsymbol{\mu}}[p] = \sum_{\mathbf{m_i} \in \mathrm{RS}} r(\mathbf{m_i})\, \eta_i$$

where the reward function is:

$$r(\mathbf{m}) = n \text{ if and only if } \mathbf{m}[p] = n$$

The *throughput* of transition t can be derived from the following reward function

$$r(\mathbf{m}) = \begin{cases} \lambda_t & \text{if } t \text{ is enabled in } \mathbf{m} \\ 0 & \text{otherwise} \end{cases}$$

(where λ_t is the firing rate of transition t, i.e., the inverse of its average service time) as the average reward:

$$\boldsymbol{\chi}[t] = \sum_{\mathbf{m_i} \in \mathrm{RS}} r(\mathbf{m_i})\, \eta_i$$

In general, the *steady-state probability of a given predicate on the marking*, $P(\mathbf{m})$, can be computed from the corresponding reward function:

$$r(\mathbf{m}) = \begin{cases} 1 & \text{if } P(\mathbf{m}) = \text{true} \\ 0 & \text{otherwise} \end{cases}$$

In Chapters 9, 10, and 11, the computation of the steady-state distribution η_i will be addressed, for the particular case where the marking process of the model is a Continuous Time Markov Chain.

8.5 Bibliographic Remarks

Operational analysis is a conceptually very simple way of deriving mathematical equations relating observable quantities in queueing systems. A classical work is [13]. In [12] the reader can find some nice examples of how the application of operational analysis techniques can help in explaining and providing fundamental results in queueing network analysis.

The first specific work on the operational analysis topic within the timed Petri nets context appeared in [10] and it was applied to the computation of performance bounds in [9] (see Chapter 17). Both the operational definition of performance indices and the derivation of operational laws presented in Sections 8.1 and 8.2 have been taken from those papers.

The study of the computability of the visit ratios for Petri nets and the definition of the FRT-nets class, presented in Section 8.3, were considered firstly in [2] and later in [8]. The particular cases of marked graphs, mono-T-semiflow nets, and free choice nets were previously studied in [3, 4, 5, 6, 7].

Concerning ergodicity concept and ergodic theory, the reader is referred to the literature on stochastic processes. A good introduction is the book by Karlin and Taylor [14]. With respect to the definition of performance measures as reward functions over the marking, the reader is referred to [11] or [1].

Bibliography

[1] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets.* J. Wiley, 1995.

[2] J. Campos. *Performance Bounds for Synchronized Queueing Networks.* PhD thesis, Departamento de Ingeniería Eléctrica e Informática, Universidad de Zaragoza, Spain, October 1990. Research Report GISI-RR-90-20.

[3] J. Campos, G. Chiola, J. M. Colom, and M. Silva. Tight polynomial bounds for steady-state performance of marked graphs. In *Proceedings of the 3^{rd} International Workshop on Petri Nets and Performance Models*, pages 200–209, Kyoto, Japan, December 1989. IEEE Computer Society Press.

[4] J. Campos, G. Chiola, J. M. Colom, and M. Silva. Properties and performance bounds for timed marked graphs. *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, 39(5):386–401, May 1992.

[5] J. Campos, G. Chiola, and M. Silva. Properties and steady-state performance bounds for Petri nets with unique repetitive firing count vector. In *Proceedings of the 3^{rd} International Workshop on Petri Nets and Performance Models*, pages 210–220, Kyoto, Japan, December 1989. IEEE Computer Society Press.

[6] J. Campos, G. Chiola, and M. Silva. Ergodicity and throughput bounds of Petri nets with unique consistent firing count vector. *IEEE Transactions on Software Engineering*, 17(2):117–125, February 1991.

[7] J. Campos, G. Chiola, and M. Silva. Properties and performance bounds for closed free choice synchronized monoclass queueing networks. *IEEE Transactions on Automatic Control*, 36(12):1368–1382, December 1991.

[8] J. Campos and M. Silva. Structural techniques and performance bounds of stochastic Petri net models. In G. Rozenberg, editor, *Advances in Petri Nets 1992*, volume 609 of *Lecture Notes in Computer Science*, pages 352–391. Springer-Verlag, Berlin, 1992.

[9] G. Chiola, C. Anglano, J. Campos, J. M. Colom, and M. Silva. Operational analysis of timed Petri nets and application to the computation of performance bounds. In *Proceedings of the* 5^{th} *International Workshop on Petri Nets and Performance Models*, pages 128–137, Toulouse, France, October 1993. IEEE-Computer Society Press.

[10] G. Chiola, J. Campos, J. M. Colom, and M. Silva. Operational analysis of timed Petri nets. In *Proceedings of the* 16^{th} *International Symposium on Computer Performance Modelling, Measurement and Evaluation (Performance'93)*, Roma, Italy, September 1993.

[11] G. Ciardo, J. Muppala, and K. S. Trivedi. On the solution of GSPN reward models. *Performance Evaluation*, 12(4):237–253, July 1991.

[12] Y. Dallery and X. R. Cao. Operational analysis of stochastic closed queueing networks. *Performance Evaluation*, 14(1):43–61, January 1992.

[13] P. J. Denning and J. P. Buzen. The operational analysis of queueing network models. *ACM Computing Surveys*, 10(3):225–261, September 1978.

[14] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, New York, NY, 1975.