Chapter 17

Performance Bounds

A complementary approach to exact or approximation techniques for the analysis of queueing network models is the computation of *bounds* for their performance measures. Performance bounds are useful in the preliminary phases of the design of a system, in which many parameters are not known accurately. Several alternatives for those parameters should be quickly evaluated, and rejected those that are clearly bad. Exact (and even approximate) solutions would be computationally very expensive. Bounds become useful in these instances since they usually require much *less computation effort*.

In this chapter, we concentrate in *net-driven* techniques for the computation of bounds for the main performance indices of timed Petri net models. Previous works on bounds computation for classical queueing networks are not included here and the interested reader is referred to the bibliographic remarks in Section 17.5. The presented techniques are characterized by their interest in stressing the intimate relationship between qualitative and quantitative aspects of Petri nets. In particular, the intensive use of *structure theory* of net models allows to obtain very *efficient computation* techniques.

The organization of the chapter is the following. In Section 17.1, a general approach for the computation of upper and lower bounds for arbitrary linear functions of average marking of places and throughput of transitions of both timed Place/Transition nets and timed Well-Formed Coloured nets is presented. Section 17.2 includes a more intuitive approach for the computation of throughput upper bounds, even if it is valid only for restricted Petri net subclasses. The relation of this technique with the general approach and the attainability of the bound for a particular subclass of nets is included. In Section 17.3, two possible improvements of upper and lower throughput bounds are presented using implicit places and liveness bounds of transitions, respectively. All the bounds computed in Sections 17.1, 17.2 and 17.3 are *insensitive* to the timing probability distributions since they are based only on the knowledge of the average service times. Section 17.4 includes a brief overview of three additional techniques for the improvement of the bounds. Since some additional assumptions on the form of the probability distribution functions associated to the service of transitions or on the conflict resolution policies are done, the obtained bounds are non-insensitive. Finally, some bibliographic remarks are included in Section 17.5.

17.1 Insensitive Performance Bounds

In this section, we present a general approach for the computation of bounds for performance indices of both *timed Place/Transition nets* and *timed Well-Formed Coloured nets*. The bounds are computed from the solution of proper linear programming problems, therefore they can be obtained in polynomial time on the size of the net model, and they depend only on the mean values of service time associated to the firing of transitions and the routing rates associated with transitions in conflict and not on the higher moments of the probability distribution functions of the random variables that describe the timing of the system. This notion of independence of the computed mean measures on the form of the probability distribution functions is known as the *insensitivity property* in queueing networks literature. The independence of the probability distribution can be viewed as a practical estimation of the performance results, since higher moments of the delays are usually unknown for real cases, and difficult to estimate and assess.

Unless otherwise explicitly stated, in this chapter we assume an *infinite-server semantics* for timed transitions, in other words, a transition t enabled K times in a marking $\mathbf{m}$ (i.e., $K = \max\{k \mid \mathbf{m} \geq k\mathbf{Pre}[\cdot, t]\}$) works at speed K times that it would work in the case it was enabled only once. Of course, an infinite-server transition can always be constrained to a "k–server" behaviour by adding one place that is both input and output (self-loop with multiplicity one) for that transition and marking it with k tokens. Other kinds of marking or time dependency of service times are forbidden.

17.1.1 General Statement Based on Linear Programming

In Chapter 8 (cfr. Section 8.3.2) several linear operational laws that relate the average marking of places and the throughput of transitions were derived from the definition of these performance measures. Those equations and inequalities can be used to compute upper and lower bounds for the throughput of transitions or for the average marking of places for general timed Petri nets using *linear programming* techniques. The idea is to compute vectors $\boldsymbol{\chi}$ and $\overline{\boldsymbol{\mu}}$ that maximize or minimize the throughput of a transition or the average marking of a place among those verifying the operational laws and other linear constraints that can be easily derived from the net structure.

A first set of linear equality constraints can be derived from the fact that the average marking vector $\overline{\boldsymbol{\mu}}$ is an average weight of reachable markings:

$$\overline{\boldsymbol{\mu}} = \sum_{\mathbf{m_r} \in \mathrm{RS}(\mathbf{m_0})} \beta_r \mathbf{m_r}$$

Since for each reachable marking,

$$\mathbf{m_r} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma}_\mathbf{r}$$

we obtain that also the average marking must satisfy the same linear equation:

$$\overline{\boldsymbol{\mu}} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma}$$

where

$$\boldsymbol{\sigma} = \sum_{\mathbf{m_r} \in \mathrm{RS}(\mathbf{m_0})} \beta_r \boldsymbol{\sigma}_\mathbf{r}$$

The following set of linear inequalities imposes that for each place the token flow out is less than or equal to the token flow in:

$$\sum_{t \in {}^\bullet p} \boldsymbol{\chi}[t] \; \mathbf{Post}[p,t] \geq \sum_{t \in p^\bullet} \boldsymbol{\chi}[t] \; \mathbf{Pre}[p,t], \quad \forall p \in P$$

If place p is known to be bounded, then the above inequality becomes an equality which represents the classical *flow balance* equation:

$$\mathbf{C}[p,\cdot] \cdot \overline{\boldsymbol{\mu}} = 0$$

On the other hand, for each pair of transitions t_i, t_j in (behavioural) free conflict (i.e., such that they are always simultaneously enabled or disabled the following equation is verified:

$$\frac{\boldsymbol{\chi}[t_i]}{r_i} = \frac{\boldsymbol{\chi}[t_j]}{r_j}$$

where r_i, r_j are the routing rates that define the resolution of the conflict between t_i and t_j.

Additionally, most of the operational inequality laws that were derived in Chapter 8 linearly relate the average marking of places with the throughput of their output transitions. Hence they can be considered as constraints for the following linear programming problem:

maximize [or minimize] $f(\overline{\boldsymbol{\mu}}, \boldsymbol{\chi})$ with f a linear function of $\overline{\boldsymbol{\mu}}, \boldsymbol{\chi}$
subject to

$$
\begin{array}{lll}
(c_1) & \overline{\boldsymbol{\mu}} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma}; & \\
(c_2) & \displaystyle\sum_{t \in {}^\bullet p} \boldsymbol{\chi}[t]\ \mathbf{Post}[p,t] \geq \sum_{t \in p^\bullet} \boldsymbol{\chi}[t]\ \mathbf{Pre}[p,t], & \forall p \in P; \\
(c_2') & \displaystyle\sum_{t \in {}^\bullet p} \boldsymbol{\chi}[t]\ \mathbf{Post}[p,t] = \sum_{t \in p^\bullet} \boldsymbol{\chi}[t]\ \mathbf{Pre}[p,t], & \forall p \in P \text{ bounded}; \\
(c_3) & \dfrac{\boldsymbol{\chi}[t_i]}{r_i} = \dfrac{\boldsymbol{\chi}[t_j]}{r_j}, & \forall t_i, t_j \in T : \text{behav. free choice (e.g. } \mathbf{Pre}[p,t_i] = \mathbf{Pre}[p,t_j], \forall p \in P); \\
(c_4) & \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \leq \dfrac{\overline{\boldsymbol{\mu}}[p]}{\mathbf{Pre}[p,t]}, & \forall t \in T, \forall p \in {}^\bullet t; \\
(c_5) & \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \geq \dfrac{\overline{\boldsymbol{\mu}}[p] - \mathbf{Pre}[p,t] + 1}{\mathbf{Pre}[p,t]}, & \forall t \in T \text{ persistent, age memory or immediate (p/a-m/i): } {}^\bullet t = \{p\}; \\
(c_5') & \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \geq k \dfrac{\overline{\boldsymbol{\mu}}[p] - k\ \mathbf{Pre}[p,t] + 1}{\mathbf{b}[p] - k\ \mathbf{Pre}[p,t] + 1}, & \forall t \in T \text{ p/a-m/i: } {}^\bullet t = \{p\} \ \wedge\ k \in \mathbb{N} :\ k\ \mathbf{Pre}[p,t] \leq \mathbf{b}[p] < (k+1)\mathbf{Pre}[p,t]; \\
(c_6) & \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t]\ \mathbf{Pre}[p_1,t] \geq \overline{\boldsymbol{\mu}}[p_1] - \mathbf{Pre}[p_1,t] + 1 & \\
 & \qquad -\mathbf{b}[p_1] \cdot \left(1 - \dfrac{\overline{\boldsymbol{\mu}}[p_2] - \mathbf{Pre}[p_2,t] + 1}{\mathbf{b}[p_2] - \mathbf{Pre}[p_2,t] + 1}\right), & \forall t \in T \text{ p/a-m/i: } {}^\bullet t = \{p_1, p_2\}, \mathbf{b}[p_1] \leq \mathbf{b}[p_2]; \\
(c_7) & \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t]\ \mathbf{Pre}[p_1,t] \geq \overline{\boldsymbol{\mu}}[p_1] - \mathbf{Pre}[p_1,t] + 1 & \\
 & \qquad -\mathbf{b}[p_1] \displaystyle\max_{1<j\leq k} \left(1 - \dfrac{\overline{\boldsymbol{\mu}}[p_j] - \mathbf{Pre}[p_j,t] + 1}{\mathbf{b}[p_j] - \mathbf{Pre}[p_j,t] + 1}\right), & \forall t \in T \text{ p/a-m/i: } {}^\bullet t = \{p_1, \cdots, p_k\}, \mathbf{b}[p_1] \leq \mathbf{b}[p_j]; \\
(c_8) & \overline{\boldsymbol{\mu}}, \boldsymbol{\chi}, \boldsymbol{\sigma} \geq \mathbf{0} &
\end{array}
\tag{17.1}
$$

As we remarked before, constraint (c_2) becomes an equality for bounded places (c_2'). Constraints (c_4) to (c_7) correspond to the operational inequality laws that were derived in Chapter 8. The equality sign also holds true in (c_4) if $\sum_{p \in {}^\bullet t} \mathbf{Pre}[p,t] = 1$ since in this case it may be combined with the opposite inequality (c_5). The constraint labelled with (c_5) can be improved if the input place to t is bounded, by introducing the additional constraint (c_5') (where $\mathbf{b}$[p] is the marking bound of p).

The above linear programming problem provides a general method to compute upper and lower bounds for arbitrary linear functions of average marking of places and throughput of transitions. For instance, if $f(\overline{\boldsymbol{\mu}}, \boldsymbol{\chi}) = \boldsymbol{\chi}[t]$, then the problem can be used to compute an upper or a lower bound (depending on the selection of "max" or "min" optimization for the objective function) for the throughput of transition t. In an analogous way, upper or lower bounds for the average marking of a given place p can be derived by solving the above linear programming problem for the objective function $f(\overline{\boldsymbol{\mu}}, \boldsymbol{\chi}) = \overline{\boldsymbol{\mu}}[p]$. The bounds are insensitive to the timing probability distributions since they are based only on the knowledge of the average service times.

The reader should notice also that most equalities and inequalities contain coefficients that depend only on the net structure and on the (known) average transition firing times (or probabilities in case of free choice immediate conflicts). The only coefficients that may be unknown at the time of the formulation of the model are the actual bounds for places $\mathbf{b}[p]$. If the modeller has no more *a priori* precise knowledge of these bounds, notice that an upper bound for them that can be used in the linear programming problem in (17.1) may be computed from a simplified linear programming problem that contains only constraint (c_1) (structural marking bound).

An improvement of the proposed bounds can be obtained if additional constraints that improve the linear characterization of the average marking in terms of the equation $\overline{\boldsymbol{\mu}} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma}$ are considered. For instance, if a *trap* Θ (i.e., $\Theta \subseteq P, \Theta^\bullet \subseteq {}^\bullet\Theta$) is not a P-semiflow, the net is live, and we are interested only in the steady state performance, then we can add the constraint:

$$\sum_{p \in \Theta} \overline{\boldsymbol{\mu}}[p] \geq 1$$

Similarly, if a *siphon* Σ (i.e., $\Sigma \subseteq P, {}^\bullet\Sigma \subseteq \Sigma^\bullet$) is not a P-semiflow and the net is live, then we can add the constraint:

$$\sum_{p \in \Sigma} \overline{\boldsymbol{\mu}}[p] \geq 1$$

A systematic method for the improvement of linear characterization of reachable markings [16] based on the addition of *implicit places* can be also applied and will be presented later in Section 17.3.1.

We remark that linear programming problems can be solved in *polynomial time* [26], therefore the above presented method for the computation of (upper and lower) bounds for the throughput and for the average marking of general timed nets has a polynomial complexity on the number of nodes of the net. Moreover, the well-known *simplex* method for the resolution of linear programming problems proceeds in linear time in most cases even if it has a theoretically exponential complexity.

17.1.2 Extension to Timed Well-Formed Coloured Nets

For Timed Well-Formed Coloured Nets (TWN's) it is possible to derive, directly from the inequalities developed in Chapter 8, operational relations allowing an efficient computation of performance bounds. Given a TWN, the basic idea is to consider the corresponding unfolded net and to apply the relations previously developed. The relations for the TWN are then obtained combining the partial results for the unfolded one.

A fundamental property that TWN's must have in order to be able to combine the results for the unfolded one is the *symmetry*, meaning that in the unfolded nets obtained from the Well-Formed ones all colour instances of a given place and of a given transition must be equivalent. To be more precise, if a transition t has average service time $\overline{\mathbf{s}}[t]$, then all of its instances have the same average service time. Moreover if a place p is bounded, then we assume that the maximum number of tokens that each of its instances can contain is the same.

In the following paragraphs we show, as an example, the derivation of *Minimum Throughput Law for single input arc* for TWN's.

Firstly, we recall some basic notations used in the derivations of relations for TWN's. A *generic function* is $f = \sum_{j=1}^{k} F_j$, where F_j is the j^{th} tuple and its arity l is given by the number of colour classes composing the colour domain of the place. This definition of function is slightly different from the classical one, since here we allow linear combinations only outside the tuples (i.e., each tuple is composed only by elementary functions). For example the function $F = \langle S - x, y\rangle$ is written as $F' = \langle S, y\rangle - \langle x, y\rangle$.

The *cardinality of a function* is defined as $|f| = \sum_{j=1}^{k} |F_j|$, where $|F_j| = \alpha_j \times \prod_{i=1}^{l} |(F_j)_i|$ is the cardinality of the j^{th} tuple. The coefficient α_j denotes the product of the coefficients of the elementary functions composing the tuple and $(F_j)_i$ is the i^{th} function of the j^{th} tuple. For example if $F_j = \langle 3x, 2y\rangle$, then $\alpha_j = 6$.

Each tuple F_j of a function f identifies a set of arcs, or *family of arcs*, (with weight α_j), whose cardinality is $A(F_j) = \prod_{i=1}^{l} |(F_j)_h|$. The global number of arcs corresponding to function f is $A(f) = \sum_{j=1}^{k} A(F_j)$, where each $A(F_j)$ has the sign of the corresponding tuple F_j. When $A(f) = 1$, then we denote as α_f the weights associated to the unique family of arcs corresponding to f.

If t is an input transition of place p (with function f), then $IN(p,t) = \frac{|cd(t)|}{|cd(p)|} A(f)$ is the number of input instances of t for each instance of p. Similarly if t is an output transition of place p, then $OUT(p,t) = \frac{|cd(t)|}{|cd(p)|} A(f)$ is the number of output instances of t for each instance of p.

To apply the Minimum Throughput Law for single input arc to an unfolded net, the conditions for its applicability must be met for all transition instances. This means that each instance t_i of a coloured transition t must have only one input place. This condition is met if the function f labelling the arc contains only *projection* and *successor* elementary functions (that is $A(f) = 1$).

Property 17.1 (Minimum Throughput Law for single input arc) *For all transition $t \in T$ with ${}^\bullet t = \{p\}$, $\mathbf{Pre}[p,t] = f$, and $A(f) = 1$:*

$$\alpha_f \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \geq OUT(p,t)\overline{\boldsymbol{\mu}}[p] - |cd(t)|(\alpha_f - 1)$$

Proof:
Assume to have a portion of a TWN containing transition t and its input place p and that $|cd(t)| = n$ and $|cd(p)| = m$. Considering the n instances of t we can write the following set of inequalities

$$\forall i \in \{1, \ldots, n\}\ \ \alpha_f \boldsymbol{\chi}[t_i]\ \overline{\mathbf{s}}[t_i] \geq \overline{\boldsymbol{\mu}}[p_{t_i}] - \alpha_f + 1$$

where p_{t_i} is the unique input place of transition instance t_i. Summing the left-hand sides and the right-hand sides of the above inequalities we obtain:

(c_1)	$\overline{\boldsymbol{\mu}}[p] = \mathbf{m_0}[p] + \sum_{t_i \in {}^\bullet p} \lvert f_i \rvert \boldsymbol{\sigma}[t_i] - \sum_{k_j \in p^\bullet} \lvert g_j \rvert \boldsymbol{\sigma}[k_j]$,	$\forall p \in P : \mathbf{Post}[p, t_i] = f_i$, $\mathbf{Pre}[p, k_i] = g_j$;
(c_2)	$\sum_{t_i \in {}^\bullet p} \lvert f_i \rvert \, \boldsymbol{\chi}[t_i] \geq \sum_{k_j \in p^\bullet} \lvert g_j \rvert \, \boldsymbol{\chi}[k_j]$,	$\forall p \in P : \mathbf{Post}[p, t_i] = f_i$, $\mathbf{Pre}[p, k_i] = g_j$;
(c_2')	$\sum_{t_i \in {}^\bullet p} \lvert f_i \rvert \, \boldsymbol{\chi}[t_i] = \sum_{k_j \in p^\bullet} \lvert g_j \rvert \, \boldsymbol{\chi}[k_j]$,	$\forall p \in P$ bounded;
(c_3)	$\frac{\boldsymbol{\chi}[t_i]}{r_i} = \frac{\boldsymbol{\chi}[t_j]}{r_j}$,	$\forall t_i, t_j \in T$: behav. free choice;
(c_4)	$\lvert f \rvert \, \boldsymbol{\chi}[t] \, \overline{\mathbf{s}}[t] \leq OUT(p,t) \, \overline{\boldsymbol{\mu}}[p]$,	$\forall t \in T, \forall p \in {}^\bullet t : \mathbf{Pre}[p,t] = f$;
(c_5)	$\alpha_f \, \boldsymbol{\chi}[t] \, \overline{\mathbf{s}}[t] \geq OUT(p,t) \, \overline{\boldsymbol{\mu}}[p] - \lvert cd(t) \rvert (\alpha_f - 1)$,	$\forall t \in T$ persistent, age memory or immediate: ${}^\bullet t = \{p\}$, $\mathbf{Pre}[p,t] = f$, $A(f) = 1$;
(c_5')	$\boldsymbol{\chi}[t] \, \overline{\mathbf{s}}[t] \geq k \frac{OUT(p,t)\overline{\boldsymbol{\mu}}[p] + \lvert cd(t) \rvert (1 - k\alpha_f)}{OUT(p,t) + \lvert cd(t) \rvert (1 - k\alpha_f)}$,	$\forall t \in T$ persistent, age memory or immediate: ${}^\bullet t = \{p\}$, $\mathbf{Pre}[p,t] = f$, $A(f) = 1$, $\wedge \; k \in \mathbb{N} : k\alpha_f \leq \mathbf{b}[p] \leq (k+1)\alpha_f$;
(c_6)	$\alpha_f \boldsymbol{\chi}[t] \, \overline{\mathbf{s}}[t] \geq OUT(p,t)\overline{\boldsymbol{\mu}}[p] + \lvert cd(t) \rvert (1 - \alpha_f) - OUT(p,t)\mathbf{b}[p] f_q$, where $f_q = \lvert cd(t) \rvert - \frac{OUT(q,t)\overline{\boldsymbol{\mu}}[q] + \lvert cd(t) \rvert (1-\alpha_g)}{OUT(q,t)\mathbf{b}[q] + \lvert cd(t) \rvert (1-\alpha_g)}$,	$\forall t \in T$ persistent, age memory or immediate: ${}^\bullet t = \{p, q\}$, $\mathbf{b}[p] \leq \mathbf{b}[q]$, $\mathbf{Pre}[p,t] = f, \mathbf{Pre}[q,t] = g, A(f) = A(g) = 1$;
(c_7)	$\alpha_1 \boldsymbol{\chi}[t] \, \overline{\mathbf{s}}[t] \geq OUT(p1,t)\overline{\boldsymbol{\mu}}[p1] - \lvert cd(t) \rvert (-\alpha_1 + 1) - OUT(p_1,t)\mathbf{b}[p_1] \max_{1 \leq j \leq n} f_j$, where $f_j = 1 - \frac{OUT(p_j,t)\overline{\boldsymbol{\mu}}[p_j] + \lvert cd(p_j) \rvert (1-\alpha_j)}{\mathbf{b}[p_j]/\lvert cd(p_j) \rvert - \alpha_j + 1}$,	$\forall t \in T$ persistent, age memory or immediate: ${}^\bullet t = \{p_1, \ldots, p_n\}, \mathbf{b}[p_1] \leq \mathbf{b}[p_j], j \in \{2, \ldots, n\}$, $\mathbf{Pre}[p_i, t] = f_i$, $A(f_1) = 1$;
(c_8)	$\overline{\boldsymbol{\mu}}, \boldsymbol{\chi}, \boldsymbol{\sigma} \geq \mathbf{0}$	

Table 17.1: Constraints for a linear programming problem for TWN's.

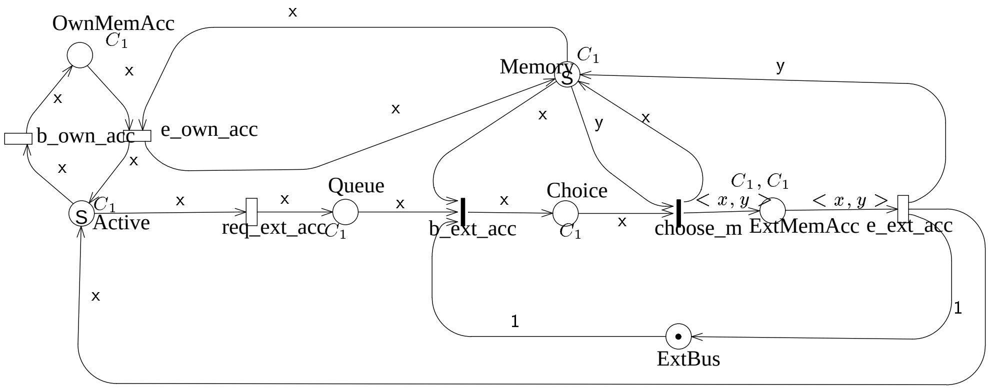

Figure 17.1: TWN model of a shared-memory multiprocessor.

$$\alpha_f \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \geq \left(\sum_{i=1}^{n} \overline{\boldsymbol{\mu}}[p_{t_i}] - |cd(t)|(\alpha_f - 1) \right) \tag{17.2}$$

Since each instance of p appears exactly $OUT(p,t)$ times in the summation of the above expression we can rewrite inequality (17.2) as

$$\alpha_f \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \geq \left(OUT(p,t) \sum_{i=1}^{m} \overline{\boldsymbol{\mu}}[p_i] - |cd(t)|(\alpha_f - 1) \right)$$

and the result follows. $\diamondsuit$

In a similar way it is possible to derive, for TWN's, the equivalent of relations devised for timed Petri nets. Therefore, performance bounds for TWN's can be computed solving linear programming problems with constraints included in Table 17.1 and any linear function of $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\chi}$ as objective function.

The average marking equation is written here in explicit form, but it could be written also in matricial form. Moreover relation (c_7) has been derived for TWN's under the hypothesis of strong symmetries. In particular we assumed that, for each input place of transition t in inequality (c_7), the weights of the arcs belonging to the families corresponding to the function labelling the arc are the same. Obviously the uncoloured version of (c_7) has no such restriction.

As we remarked in the case of timed Petri nets, also for TWN's constraint (c_2) becomes an equality for bounded places (c_2'). The equality sign also holds true in (c_4) if $\alpha_f = 1$ (i.e., the unique family of arcs corresponding to function f have weight 1) since in this case it may be combined with the opposite inequality (c_5). The constraint labelled with (c_5) can be improved if the input place to t_i is bounded, by introducing the additional constraint (c_5').

$$
\begin{array}{ll}
(c_1) & \overline{\boldsymbol{\mu}}[Active] = 4 + \boldsymbol{\sigma}[e\_e\_a] + \boldsymbol{\sigma}[e\_o\_a] - \boldsymbol{\sigma}[r\_e\_a] - \boldsymbol{\sigma}[b\_o\_a]; \\
& \overline{\boldsymbol{\mu}}[Memory] = 4 + \boldsymbol{\sigma}[e\_e\_a] - \boldsymbol{\sigma}[b\_e\_a]; \\
& \overline{\boldsymbol{\mu}}[OwnMemAcc] = \boldsymbol{\sigma}[b\_o\_a] - \boldsymbol{\sigma}[e\_o\_a]; \\
& \overline{\boldsymbol{\mu}}[Queue] = \boldsymbol{\sigma}[r\_e\_a] - \boldsymbol{\sigma}[b\_e\_a]; \\
& \overline{\boldsymbol{\mu}}[Choice] = \boldsymbol{\sigma}[b\_e\_a] - \boldsymbol{\sigma}[c\_m]; \\
& \overline{\boldsymbol{\mu}}[ExtMemAcc] = \boldsymbol{\sigma}[c\_m] - \boldsymbol{\sigma}[e\_e\_a]; \\
& \overline{\boldsymbol{\mu}}[ExtBus] = 1 + \boldsymbol{\sigma}[e\_e\_a] - \boldsymbol{\sigma}[b\_e\_a]; \\
(c_2') & \boldsymbol{\chi}[e\_e\_a] + \boldsymbol{\chi}[e\_o\_a] = \boldsymbol{\chi}[r\_e\_a] + \boldsymbol{\chi}[b\_o\_a]; \\
& \boldsymbol{\chi}[b\_e\_a] = \boldsymbol{\chi}[c\_m] = \boldsymbol{\chi}[e\_e\_a] = \boldsymbol{\chi}[r\_e\_a]; \\
(c_3) & \boldsymbol{\chi}[b\_o\_a] = \boldsymbol{\chi}[r\_e\_a]; \\
(c_4 \& c_5) & \boldsymbol{\chi}[b\_o\_a]\ \overline{\mathbf{s}}[b\_o\_a] = \overline{\boldsymbol{\mu}}[Active]/2; \\
& \boldsymbol{\chi}[r\_e\_a]\ \overline{\mathbf{s}}[r\_e\_a] = \overline{\boldsymbol{\mu}}[Active]/2; \\
& \boldsymbol{\chi}[e\_e\_a]\ \overline{\mathbf{s}}[e\_e\_a] = \overline{\boldsymbol{\mu}}[ExtMemAcc]; \\
(c_4) & \boldsymbol{\chi}[e\_o\_a]\ \overline{\mathbf{s}}[e\_o\_a] \le \overline{\boldsymbol{\mu}}[OwnMemAcc]; \\
& \boldsymbol{\chi}[e\_o\_a]\ \overline{\mathbf{s}}[e\_o\_a] \le \overline{\boldsymbol{\mu}}[Memory]; \\
(c_6) & \boldsymbol{\chi}[e\_o\_a]\ \overline{\mathbf{s}}[e\_o\_a] \ge \overline{\boldsymbol{\mu}}[OwnMemAcc] + \frac{\mathbf{b}[OwnMemAcc]}{\mathbf{b}[Memory]}\overline{\boldsymbol{\mu}}[Memory] \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - \mathbf{b}[Memory]; \\
(c_7) & 4\left(\overline{\boldsymbol{\mu}}[ExtBus] - \mathbf{b}[ExtBus]\left(1 - \dfrac{\overline{\boldsymbol{\mu}}[Memory]}{\mathbf{b}[Memory]}\right)\right) \le 0; \\
& 4\left(\overline{\boldsymbol{\mu}}[ExtBus] - \mathbf{b}[ExtBus]\left(1 - \dfrac{\overline{\boldsymbol{\mu}}[Queue]}{\mathbf{b}[Queue]}\right)\right) \le 0;
\end{array}
$$

Table 17.2: Constraints for the model in Figure 17.1.

17.1.3 An Example of Application

Let us present an example of application for the computation of bounds in the case of the TWN model of a shared-memory multiprocessor depicted in Figure 17.1. The architecture comprises a set of processing modules interconnected by a common bus called the "external bus". A processor can access its own memory module directly from its private bus through one port, or it can access non-local shared-memory modules by means of the external bus. In case of contention for the access to one shared-memory module, preemptive priority is given to external access through the external bus with respect to the accesses from the local processor. The experiments on the shared-memory model have been carried out assuming to have 4 processors and that the average service time of all the transitions are equal to 0.5.

According to the arguments presented in the previous sections, bounds can be computed solving linear programming problems with constraints included in Table 17.2, where the first letters of each transition name have been used for reasons of space. The solution for the linear programming problems leads to upper and lower bounds, for the throughput of transitions, given by

$$\frac{8}{11} \leq \boldsymbol{\chi}[e\_e\_a] \leq 2$$

while the "exact" solution with exponential distribution is

$$\boldsymbol{\chi}[e\_e\_a] = 1.71999$$

An improvement in the lower bound can be obtained observing that when a token arrives in place Choice transition choose_m is enabled at least for one transition instance. This implies that the average marking of place Choice is equal to 0 (transition choose_m is immediate), so

$$\overline{\boldsymbol{\mu}}[Choice] = 0; \quad \mathbf{b}[Choice] = 0$$

(only tangible markings are considered) can be added to the set of constraints. Moreover place Memory is implicit with respect to the enabling of transition b_ext_acc, so we can consider this transition as having only two input places, so constraint (c_6) can be applied instead of constraint (c_7). Finally,

$$\mathbf{b}[Queue] = 3$$

can be added since the output transition of place Queue is immediate, and from the behaviour of the model it is clear that at most 3 processors can be waiting in the queue. The relations (c_7) in the above linear programming problem can thus be replaced with the new constraint:

$$4\left(\overline{\boldsymbol{\mu}}[ExtBus] + \frac{\mathbf{b}[ExtBus]}{\mathbf{b}[Queue]}\overline{\boldsymbol{\mu}}[Queue] - \mathbf{b}[ExtBus]\right) \leq 0$$

where $\mathbf{b}[Queue] = 3$. Solving this reduced linear programming problem the values obtained for the upper and lower bounds are:

$$1 \leq \boldsymbol{\chi}[e\_e\_a] \leq 2$$

17.2 Reinterpretation of the Insensitive Bounds for Net Subclasses

This section includes a more intuitive approach for the computation of throughput upper bounds for particular net classes than the general technique presented above. It makes use of an *implicit decomposition* of the net model into P-semiflows. The details of the technique are presented in Section 17.2.1. Section 17.2.2 addresses the relationship between the general technique and this more intuitive approach. Section 17.2.3 shows that the bound is tight for marked graphs.

17.2.1 Little's Law and P-Semiflows

Three of the most significant performance measures for a closed region of a network in the analysis of queueing systems are related by Little's formula: the average number of customers, the output (or input) rate (throughput), and the average time spent by a customer within the region. For the case of timed Petri nets, if the involved *limit average performance indices* exist, then Little's formula can be applied, in particular, to each place of the system as follows (cfr. Chapter 8):

$$\overline{\boldsymbol{\mu}}[p] = (\mathbf{Pre}[p, \cdot] \cdot \boldsymbol{\chi})\overline{\mathbf{r}}[p]$$

where $\mathbf{Pre}[p, \cdot]$ is the row of the pre-incidence matrix of the underlying Petri net corresponding to place p, thus $\mathbf{Pre}[p, \cdot] \cdot \boldsymbol{\chi}$ is the output rate of place p, and $\overline{\mathbf{r}}[p]$ is the limit average token residence time at p.

In the study of computer systems, Little's law is frequently used when two of the related quantities are known and the third one is needed. This is not exactly the case here. In this case, $\overline{\mathbf{r}}[p]$ and $\overline{\boldsymbol{\mu}}[p]$ are unknown, while partial information about $\boldsymbol{\chi}$ can be easily computed only for some (interesting) net subclasses. Let us recall the definition of the *relative throughput vector* or *vector of visit ratios* to transitions (cfr. Chapter 8):

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

where $\Gamma[t_i] = 1/\boldsymbol{\chi}[t_i]$ represents the *average interfiring time* of transition t_i (i.e., the inverse of its throughput). Here we consider *live and bounded* timed net systems whose vector of visit ratios to transitions can be computed in polynomial time from the net structure $\mathcal{N}$ and from the relative frequency of conflict resolutions $\mathcal{R}$ (i.e., the routing rates associated with decisions). In Chapter 8, a class of net models for which such computation is possible was presented, as well as some interesting subclasses were identified. As an example, let us consider the net system depicted in Figure 17.2. For this net, the vector of visit ratios for transitions can be computed by solving the following linear system of equations:

$$\begin{array}{rcl} \mathbf{C} \cdot \mathbf{v} & = & \mathbf{0}; \\ r_1 \mathbf{v}[t_2] & = & r_2 \mathbf{v}[t_1]; \\ r_3 \mathbf{v}[t_4] & = & r_4 \mathbf{v}[t_3]; \\ \mathbf{v}[t_1] & = & 1 \end{array} \tag{17.3}$$

where r_1 and r_2 (r_3 and r_4) are the routing rates used for the resolution of the conflict between t_1 and t_2 (respectively, t_3 and t_4). The first set of equations (implying that $\mathbf{v}$ is a T-semiflow) are the *flow balance* equations written for each place (input and output flows of tokens are equal, provided that the places are bounded). The second (third) equation is directly derived from the fact that conflict between t_1 and t_2 (respectively, t_3 and t_4) is free and rates r_1 and r_2 (respectively, r_3 and r_4) are fixed. The fourth equation is the normalization for transition t_1.

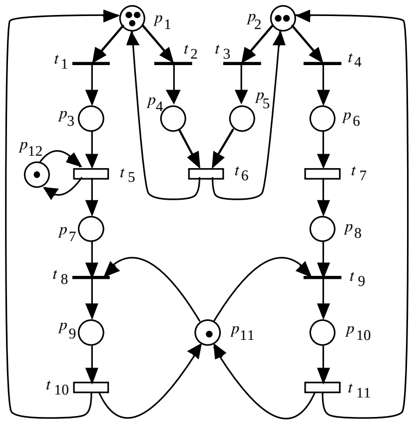

Figure 17.2: A live and bounded stochastic Petri net system.

In Chapter 8, equations like (17.3) have been shown to characterize the vector of visit ratios for important net subclasses such as, for instance, live and bounded *mono-T-semiflow systems* and live and bounded *free choice systems*. Unfortunately, for other subclasses like *simple net systems*, the relative throughput vector also depends on the initial marking and on the service times of transitions.

In the sequel of this chapter (unless otherwise explicitly stated), *we assume that timed transitions may never be in conflict.* For the modelling of conflicts we use immediate transitions with the addition of (marking and time independent) routing rates $\mathcal{R}$. In other words, for the subset of immediate transitions $\{t_1, \ldots, t_k\} \subset T$ being in conflict at each reachable marking, we suppose that the constants $r_1, \ldots, r_k \in \mathbb{R}^+$ are explicitly defined in the system interpretation in such a way that when $t_1, \ldots, t_k$ are simultaneously enabled, transition t_i $(i = 1, \ldots, k)$ fires with relative rate $r_i/(\sum_{j=1}^{k} r_j)$. In this way, routing is completely decoupled from duration of activities. The only restriction that this decoupling imposes to the system is that *preemption* cannot be modelled with two timed transitions (in conflict) competing for the tokens (i.e., *race policy* cannot be modelled; this constraint is equivalent to the use of a *preselection policy* for the resolution of conflicts among timed transitions).

If timed transitions are never in conflict, either all output transitions of a place p are immediate or p has a unique output transition, say t_1, and t_1 is timed. Then, in the later case:

$$\begin{aligned}\overline{\boldsymbol{\mu}}[p] \quad &= \quad (\mathbf{Pre}[p,\cdot] \cdot \boldsymbol{\chi})\overline{\mathbf{r}}[p] = \mathbf{Pre}[p,t_1]\boldsymbol{\chi}[t_1]\overline{\mathbf{r}}[p] \\ &\geq \mathbf{Pre}[p,t_1]\boldsymbol{\chi}[t_1]\overline{\mathbf{s}}[t_1] = \sum_{j=1}^{m} \mathbf{Pre}[p,t_j]\boldsymbol{\chi}[t_j]\overline{\mathbf{s}}[t_j]\end{aligned}$$

The inequality follows from the fact that the residence time $\overline{\mathbf{r}}[p]$ of a token at place p with only one output transition is greater than or equal to the service time $\overline{\mathbf{s}}[t_1]$ of that transition. So that:

$$\Gamma[t_1]\overline{\boldsymbol{\mu}}[p] \geq \sum_{j=1}^{m} \mathbf{Pre}[p,t_j]\Gamma[t_1]\boldsymbol{\chi}[t_j]\overline{\mathbf{s}}[t_j] = \sum_{j=1}^{m} \mathbf{Pre}[p,t_j]\mathbf{v}[t_j]\overline{\mathbf{s}}[t_j]$$

hence:

$$\Gamma[t_1]\overline{\boldsymbol{\mu}} \geq \mathbf{Pre} \cdot \overline{\mathbf{D}} \tag{17.4}$$

where $\overline{\mathbf{D}}$ is the vector of *average service demands* of transitions, with components:

$$\overline{\mathbf{D}}[t] = \overline{\mathbf{s}}[t]\mathbf{v}[t] \tag{17.5}$$

If all output transitions of place p are immediate, then $\overline{\boldsymbol{\mu}}[p] = \mathbf{Pre}[p,\cdot] \cdot \overline{\mathbf{D}} = 0$, thus inequality (17.4) holds for all place p.

P-semiflows $\mathbf{y}$ are non-negative left annullers of the incidence matrix $\mathbf{C}$ (i.e., $\mathbf{y} \cdot \mathbf{C} = \mathbf{0}$), thus $\forall \mathbf{m_0} : \mathbf{y} \cdot \mathbf{m} = \mathbf{y} \cdot \mathbf{m_0}$ for all reachable marking $\mathbf{m}$. Therefore, $\mathbf{y} \cdot \overline{\boldsymbol{\mu}} = \mathbf{y} \cdot \mathbf{m_0}$. Now, premultiplying by $\mathbf{y}$ the relation (17.4), the following lower bound for the average interfiring time of transition t_1 can be derived:

$$\Gamma[t_1] \geq \max_{\mathbf{y} \in \{P-\text{semiflow}\}} \frac{\mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}}}{\mathbf{y} \cdot \mathbf{m_0}} \tag{17.6}$$

Of course, un upper bound for the throughput of t_1 can be computed taken the inverse. From that bound and from the knowledge of the vector of visit ratios, upper bounds for the throughput of the other transitions can be derived.

Let us formulate the previous lower bound for the average interfiring time of t_1 in terms of a particular class of optimization problems called *fractional programming problems* [26]:

$$\begin{array}{rll} \Gamma[t_1] \geq & \text{maximum} & \dfrac{\mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}}}{\mathbf{y} \cdot \mathbf{m_0}} \\ & \text{subject to} & \mathbf{y} \cdot \mathbf{C} = \mathbf{0} \\ & & \mathbf{1} \cdot \mathbf{y} > 0 \\ & & \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.7}$$

where $\mathbf{1}$ is a vector with all entries equal to one. The above problem can be rewritten as follows:

$$\begin{array}{rll} \Gamma[t_1] \geq & \text{maximum} & \dfrac{\mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}}}{q} \\ & \text{subject to} & \mathbf{y} \cdot \mathbf{C} = \mathbf{0} \\ & & \mathbf{1} \cdot \mathbf{y} > 0 \\ & & q = \mathbf{y} \cdot \mathbf{m_0} \\ & & \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.8}$$

Then, because $\mathbf{y} \cdot \mathbf{m_0} > 0$ (guaranteed for live systems), we can change $\mathbf{y}/q$ to $\mathbf{y}$ and obtain the linear programming formulation stated in the next theorem (in which $\mathbf{1} \cdot \mathbf{y} > 0$ is removed because $\mathbf{y} \cdot \mathbf{m_0} = 1 \Longrightarrow \mathbf{1} \cdot \mathbf{y} > 0$).

Theorem 17.2 *For any live and bounded system, a lower bound for the average interfiring time $\Gamma[t_1]$ of transition t_1 can be computed by the following linear programming problem:*

$$\Gamma[t_1] \geq \begin{array}{ll} \textit{maximum} & \mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}} \\ \textit{subject to} & \mathbf{y} \cdot \mathbf{C} = \mathbf{0} \\ & \mathbf{y} \cdot \mathbf{m_0} = 1 \\ & \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.9}$$

If the solution of the above problem is unbounded and since it is a lower bound for the average interfiring time of transition t_1, the non-liveness can be assured (infinite interfiring time). If the visit ratios of all transitions are non-null (i.e., $\mathbf{v} > \mathbf{0}$), the unboundedness of the above problem implies that a total deadlock is reached by the net system. Anyhow, the unboundedness of the solution of (17.9) means that there exists an unmarked P-semiflow, and obviously the net system is non-live: if $\mathbf{y} \cdot \mathbf{C} = \mathbf{0}$ and $\mathbf{y} \cdot \mathbf{m_0} = 0$, then $\forall \mathbf{m}$ $\forall p \in \|\mathbf{y}\|$: $\mathbf{m}[p] = 0$, and the input and output transitions of p are never firable.

The basic advantage of Theorem 17.2 lies, again, in the fact that the *simplex method* for the solution of a linear programming problem has almost linear complexity in practice, even if it has exponential worst case complexity. In any case, algorithms of polynomial worst case complexity can be found in [26].

In order to interpret Theorem 17.2, let us consider again the net system of Figure 17.2. Assuming, for instance, that all routing rates associated with output transitions at conflicts in p_1 and p_2 are equal to one, then the system (17.3) gives $\mathbf{v} = \mathbf{1}$. Therefore, according to (17.5), the vector of average service demands for transitions (normalized for t_1) is $\overline{\mathbf{D}} = (0, 0, 0, 0, \overline{\mathbf{s}}[t_5], \overline{\mathbf{s}}[t_6], \overline{\mathbf{s}}[t_7], 0, 0, \overline{\mathbf{s}}[t_{10}], \overline{\mathbf{s}}[t_{11}])$, because transitions t_1, t_2, t_3, t_4, t_8, and t_9 are assumed to be immediate.

The *minimal* P-semiflows (minimal support solutions of $\mathbf{y} \cdot \mathbf{C} = \mathbf{0}, \mathbf{y} \geq \mathbf{0}$) of this net are:

$$\begin{array}{l} \mathbf{y_1} = (1,0,1,1,0,0,1,0,1,0,0,0) \\ \mathbf{y_2} = (0,1,0,0,1,1,0,1,0,1,0,0) \\ \mathbf{y_3} = (0,0,0,0,0,0,0,0,1,1,1,0) \\ \mathbf{y_4} = (0,0,0,0,0,0,0,0,0,0,0,1) \end{array} \tag{17.10}$$

and the application of (17.9) gives:

$$\Gamma[t_1] \;\; \geq \;\; \max\{ \begin{array}{l} (\overline{\mathbf{s}}[t_5] + \overline{\mathbf{s}}[t_6] + \overline{\mathbf{s}}[t_{10}])/3, \\ (\overline{\mathbf{s}}[t_6] + \overline{\mathbf{s}}[t_7] + \overline{\mathbf{s}}[t_{11}])/2, \\ \overline{\mathbf{s}}[t_{10}] + \overline{\mathbf{s}}[t_{11}], \\ \overline{\mathbf{s}}[t_5] \; \} \end{array} \tag{17.11}$$

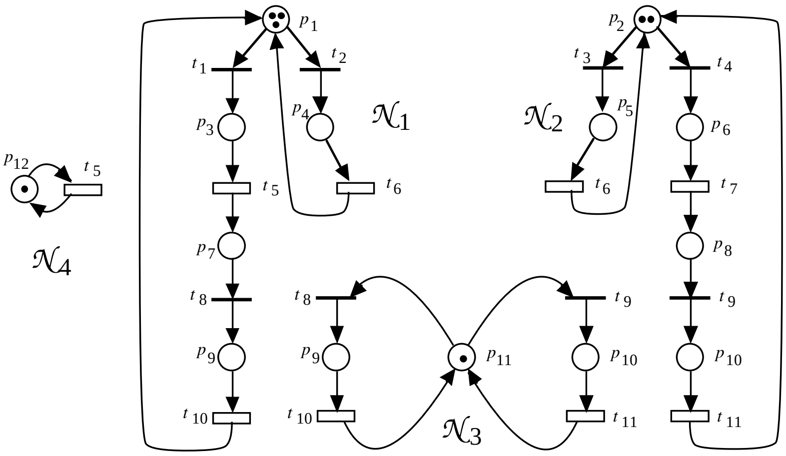

Figure 17.3: Embedded queueing networks of the net in Figure 17.2 generated by minimal P-semiflows.

Now, let us consider the P-semiflow decomposed view of the net: the four *subnets generated by* $\mathbf{y_1}$, $\mathbf{y_2}$, $\mathbf{y_3}$, and $\mathbf{y_4}$ are depicted in isolation in Figure 17.3. Formally speaking, if $\mathbf{y_i}$ is a minimal P-semiflow of a net $\mathcal{N} = \langle P, T, \mathbf{Pre}, \mathbf{Post} \rangle$, the subnet generated by $\mathbf{y_i}$ is $\mathcal{N}_i = \langle P_i, T_i, \mathbf{Pre}_i, \mathbf{Post}_i \rangle$ where $P_i = \|\mathbf{y_i}\|$ (the support of the P-semiflow), $T_i = {}^\bullet P_i \cup P_i{}^\bullet$ (i.e., the subset of input or output transitions of places belonging to P_i), and $\mathbf{Pre}_i$, $\mathbf{Post}_i$ are the functions $\mathbf{Pre}$, $\mathbf{Post}$ restricted to $P_i \times T_i$.

The quantities under the max operator in (17.11) represent, for this particular case, the average interfiring time of a transition of each of the four subnets (embedded queueing networks) assuming that all the nodes are delay stations (infinite-server semantics). Therefore, the lower bound for the average interfiring time of t_1 in the original net system given by (17.11) is computed *looking at the "slowest subsystem" generated by the P-semiflows*, considered in *isolation* (with delay nodes).

We remark that in this case, since $\mathbf{v} = \mathbf{1}$, the throughput of all transitions is equal and it is not necessary to weight the average interfiring time of transitions computed in isolated subnets in order to get a bound for transition t_1.

17.2.2 Equivalence with the General Statement

In this section, we study the relation between the general technique for the computation of bounds presented in Section 17.1.1 and the particular technique presented in the previous section (cfr. Theorem 17.2). More precisely, we show that the bound given by Theorem 17.2 never improves the bound given by the general technique presented in Section 17.1.1.

Let us consider live systems built on structurally live and structurally bounded nets. With respect to the timing interpretation, we are still considering infinite-server semantics for transitions and, as in the previous section, we assume that timed transitions may never be in conflict (conflicts resolution is quantified with routing rates associated to immediate transitions).

If $\mathcal{N}$ is a structurally live and structurally bounded net and $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ is a live system, Theorem 17.2 gives the following lower bound for the average interfiring time $\Gamma[t_1]$ (i.e., the inverse of the throughput, $\Gamma[t_1] = 1/\boldsymbol{\chi}[t_1]$) of transition t_1:

$$\begin{array}{rll} \Gamma_1^{\min} = & \text{maximum} & \mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}} \\ & \text{subject to} & \mathbf{y} \cdot \mathbf{C} = \mathbf{0};\ \mathbf{y} \cdot \mathbf{m_0} = 1;\ \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.12}$$

where $\overline{\mathbf{D}}$ is the vector of average service demands of transitions, with components $\overline{\mathbf{D}}[t] = \overline{\mathbf{s}}[t]\mathbf{v}[t]$, and $\mathbf{v}$ is the vector of visit ratios to transitions, with components $\mathbf{v}[t] = \frac{\boldsymbol{\chi}[t]}{\boldsymbol{\chi}[t_1]} = \Gamma[t_1]\boldsymbol{\chi}[t]$.

Since the net is structurally live, in particular, it is *structurally repetitive* (i.e., $\exists \mathbf{x} \geq \mathbf{0}:\ \mathbf{C} \cdot \mathbf{x} \geq \mathbf{0}$), thus by *Minkowski-Farkas lemma* [26]:

$$\nexists \mathbf{y} \geq \mathbf{0}:\ \mathbf{y} \cdot \mathbf{C} \leq \mathbf{0}\ \wedge\ \mathbf{y} \cdot \mathbf{C} \neq \mathbf{0}$$

Then, the solution of (17.12) is the same as that of:

$$\begin{array}{rll} \Gamma_1^{\min} = & \text{maximum} & \mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}} \\ & \text{subject to} & \mathbf{y} \cdot \mathbf{C} \leq \mathbf{0};\ \mathbf{y} \cdot \mathbf{m_0} = 1;\ \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.13}$$

Since the system is live:

$$\mathbf{y} \neq \mathbf{0}\ \wedge\ \mathbf{y} \cdot \mathbf{C} = \mathbf{0}\ \Longrightarrow\ \mathbf{y} \cdot \mathbf{m_0} \geq 1$$

Then, the solution of (17.13) is the same as that of:

$$\begin{array}{rll} \Gamma_1^{\min} = & \text{maximum} & \mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{D}} \\ & \text{subject to} & \mathbf{y} \cdot \mathbf{C} \leq \mathbf{0};\ \mathbf{y} \cdot \mathbf{m_0} \leq 1;\ \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.14}$$

Let us consider the *dual* linear programming problem[1] [26] of (17.14):

$$\begin{array}{rll} \gamma^* = & \text{minimum} & \gamma \\ & \text{subject to} & \gamma \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma} \geq \mathbf{Pre} \cdot \overline{\mathbf{D}};\ \gamma \geq 0, \boldsymbol{\sigma} \geq \mathbf{0} \end{array} \tag{17.15}$$

The *Strong Duality Theorem* of linear programming [26] states that if $\Gamma_1^{\min}$ or γ^* is finite, then both (17.14) and (17.15) have finite optimal value and $\Gamma_1^{\min} = \gamma^*$.

On the other hand, let us consider the general statement presented in Section 17.1.1 applied for the computation of a lower bound for the average interfiring time $\Gamma[t_1]$ of transition t_1. If we consider only constraints c_1, c_2', c_3, and c_4, the following system is obtained:

[1]The dual problem of $\max\{\mathbf{c} \cdot \mathbf{x}:\ \mathbf{A} \cdot \mathbf{x} \leq \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\}$ is $\min\{\mathbf{y} \cdot \mathbf{b}:\ \mathbf{y} \cdot \mathbf{A} \geq \mathbf{c},\ \mathbf{y} \geq \mathbf{0}\}$.

$$\begin{array}{rll} \gamma^* = & \text{minimum} & \dfrac{1}{\boldsymbol{\chi}[t_1]} \\ & \text{subject to} & \overline{\boldsymbol{\mu}} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma} \\ & & \displaystyle\sum_{t \in {}^\bullet p} \boldsymbol{\chi}[t]\ \mathbf{Post}[p,t] = \sum_{t \in p^\bullet} \boldsymbol{\chi}[t]\ \mathbf{Pre}[p,t],\ \forall p \in P \\ & & \dfrac{\boldsymbol{\chi}[t_i]}{r_i} = \dfrac{\boldsymbol{\chi}[t_j]}{r_j},\ \forall t_i, t_j \in T : \mathbf{Pre}[p,t_i] = \mathbf{Pre}[p,t_j], \forall p \in P \\ & & \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t] \le \dfrac{\overline{\boldsymbol{\mu}}[p]}{\mathbf{Pre}[p,t]},\ \forall p \in {}^\bullet t,\ \forall t \in T \\ & & \overline{\boldsymbol{\mu}}, \boldsymbol{\chi}, \boldsymbol{\sigma} \ge \mathbf{0} \end{array} \tag{17.16}$$

The first and the fourth sets of constraints in the above problem can be combined to eliminate variables $\overline{\boldsymbol{\mu}}$, leading to:

$$\begin{array}{rll} \gamma^* = & \text{minimum} & \dfrac{1}{\boldsymbol{\chi}[t_1]} \\ & \text{subject to} & \displaystyle\sum_{t \in {}^\bullet p} \boldsymbol{\chi}[t]\ \mathbf{Post}[p,t] = \sum_{t \in p^\bullet} \boldsymbol{\chi}[t]\ \mathbf{Pre}[p,t],\ \forall p \in P \\ & & \dfrac{\boldsymbol{\chi}[t_i]}{r_i} = \dfrac{\boldsymbol{\chi}[t_j]}{r_j},\ \forall t_i, t_j \in T : \mathbf{Pre}[p,t_i] = \mathbf{Pre}[p,t_j], \forall p \in P \\ & & \mathbf{m_0}[p] + \mathbf{C}[p,T] \cdot \boldsymbol{\sigma} \ge \mathbf{Pre}[p,t]\ \boldsymbol{\chi}[t]\ \overline{\mathbf{s}}[t],\ \forall p \in {}^\bullet t,\ \forall t \in T \\ & & \boldsymbol{\chi}, \boldsymbol{\sigma} \ge \mathbf{0} \end{array} \tag{17.17}$$

Now, the first and the second sets of constraints in the above problem state that vector $\boldsymbol{\chi}$ is proportional to the visit ratios vector, i.e., $\boldsymbol{\chi} = \mathbf{v}/\gamma$ (with $\mathbf{v}[t_1] = 1$), then the vector of variables $\boldsymbol{\chi}$ can be reduced to a single variable γ, and by definition of $\overline{\mathbf{D}}$, we get:

$$\begin{array}{rll} \gamma^* = & \text{minimum} & \gamma \\ & \text{subject to} & \gamma \mathbf{m_0}[p] + \mathbf{C}[p,T] \cdot \boldsymbol{\sigma} \ge \mathbf{Pre}[p,t]\ \overline{\mathbf{D}}[t],\ \forall p \in {}^\bullet t,\ \forall t \in T \\ & & \gamma \ge 0, \boldsymbol{\sigma} \ge \mathbf{0} \end{array} \tag{17.18}$$

Since we are assuming that timed transitions may never be in conflict, the above problem is equivalent to (17.15). Therefore, the bound (17.12) derived from Theorem 17.2 can be obtained also from the general approach (17.16) presented in Section 17.1.1.

17.2.3 Reachability of the Throughput Upper Bound for Marked Graphs

A particular interesting case of nets whose vector of visit ratios is fixed by the structure is that of marked graphs (MG's) (cfr. Chapter 8). Since MG's are *consistent* nets and their unique minimal P-semiflow is $\mathbf{1}$, their vector of visit ratios is also $\mathbf{1}$; therefore, $\Gamma[t] = \Gamma, \forall t \in T$, Γ is called the *average cycle time* of the MG and it can be bounded (assuming the net is strongly connected thus structurally bounded) by solving the following linear programming problem:

$$\begin{array}{rll} \Gamma \geq & \text{maximum} & \mathbf{y} \cdot \mathbf{Pre} \cdot \overline{\mathbf{s}} \\ & \text{subject to} & \mathbf{y} \cdot \mathbf{C} = \mathbf{0} \\ & & \mathbf{y} \cdot \mathbf{m_0} = 1 \\ & & \mathbf{y} \geq \mathbf{0} \end{array} \tag{17.19}$$

For deterministically timed nets, the *attainability* of this bound was shown by C. Ramchandani in his Ph.D. dissertation, meaning that the value computed in (17.19) is in fact the actual average cycle time of the MG. Even more, the next result shows that the previous bound cannot be improved only on the basis of the knowledge of the coefficients of variation for transition service times.

Theorem 17.3 *For live strongly-connected MG's with arbitrary values of mean and variance for transition service times, the bound for the average cycle time obtained from (17.19) cannot be improved.*

Proof:
We know from Ramchandani's work [27] that for deterministic timing the bound is reached. Only "max" and sum operators are needed to compute the average cycle time in case of MG's. Therefore, let us construct a family of random variables with arbitrary means and variances behaving in the limit like deterministic timing for both operators (max and sum).

This is the case for the following family of random variables, for varying values of the parameter α $(0 \leq \alpha \leq 1)$:

$$X_{\mu,\sigma}(\alpha) = \begin{cases} \mu\alpha & \text{with probability } 1-\epsilon \\ \mu\left(\alpha + \frac{1-\alpha}{\epsilon}\right) & \text{with probability } \epsilon \end{cases} \tag{17.20}$$

where

$$\epsilon = \frac{\mu^2(1-\alpha)^2}{\mu^2(1-\alpha)^2 + \sigma^2} \tag{17.21}$$

These variables are such that

$$\mathrm{E}[X_{\mu,\sigma}(\alpha)] = \mu; \quad \mathrm{Var}[X_{\mu,\sigma}(\alpha)] = \sigma^2$$

and they satisfy:

$$\lim_{\alpha\to 1} \mathrm{E}[\max(X_{\mu,\sigma}(\alpha), X_{\mu',\sigma'}(\alpha))] = \max(\mu, \mu') \tag{17.22}$$

and, of course,

$$\forall\ 0 \leq \alpha < 1 : \mathrm{E}[X_{\mu,\sigma}(\alpha) + X_{\mu',\sigma'}(\alpha)] = \mu + \mu'$$

Then, if random variables $X_{\overline{\mathbf{s}}[t],\sigma_t}(\alpha)$ are associated with transitions $t \in T$, taking α closer to 1, the average cycle time tends to the bound given by (17.19).

$\diamondsuit$

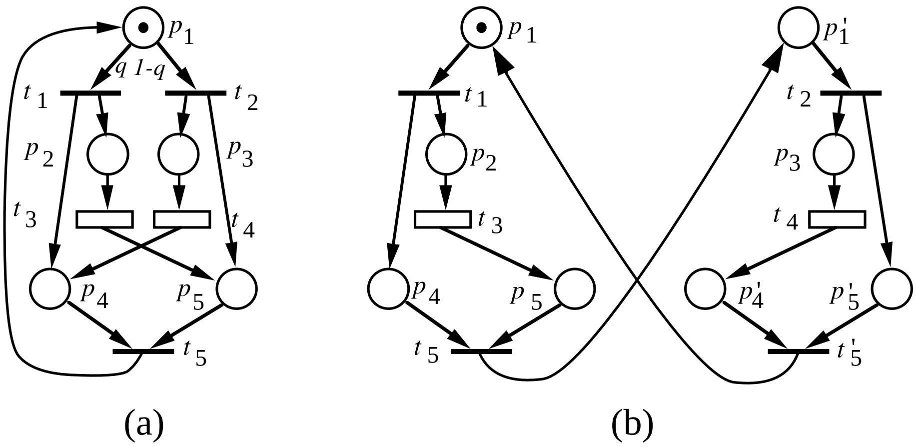

Figure 17.4: (a) A live and safe Free Choice net and (b) its behaviourally equivalent safe marked graph for deterministic resolution of conflict and $q = 1/2$.

We remark that the significance of the above theorem is the attainability of the bound for *any given means and variances* of involved random variables. In other words, even with the knowledge of second order moments, it is not possible to improve the bound given by (17.19), computed only with mean values.

17.3 Structural Improvements of the Insensitive Bounds

This section includes two approaches for improving the throughput bounds presented above. The first of them makes use of *implicit places* to improve the throughput upper bound presented in Section 17.2.1. The basic idea is that the addition of implicit places may increment the number of minimal P-semiflows of the net. Therefore, the computed bound (basically obtained by searching the slowest subsystem among those defined by the minimal P-semiflows) can be improved. The second technique allows to get a tight throughput lower bound for marked graphs, using the *structural liveness bound* of transitions.

17.3.1 The Role of Implicit Places

For strongly connected marked graphs, the bound derived in Theorem 17.2 has been shown to be reachable for arbitrary mean values and coefficients of variation associated with transition service times (cfr. Theorem 17.3). Unfortunately, this is not the case for more general net subclasses. Let us consider, for instance, the live and safe (1–bounded) Free Choice net in Figure 17.4.a. Let $\overline{\mathbf{s}}[t_3]$ and $\overline{\mathbf{s}}[t_4]$ be the average service times associated with t_3 and t_4, respectively. Let t_1, t_2, and t_5 be immediate transitions (i.e., they fire in zero time). Let

$q, 1-q \in (0,1)$ be the probabilities defining the resolution of conflict between t_1 and t_2. The relative (to t_5, i.e., such that $\mathbf{v}[t_5] = 1$) throughput vector is $\mathbf{v} = (q, 1-q, q, 1-q, 1)$. The elementary P-semiflows are $\mathbf{y_1} = (1,1,0,0,1)$ and $\mathbf{y_2} = (1,0,1,1,0)$. Then, applying the linear programming problem in (17.9) to this net, the following lower bound for the average interfiring time of transition t_5 is obtained:

$$\Gamma[t_5] \geq \max\{q\overline{\mathbf{s}}[t_3], (1-q)\overline{\mathbf{s}}[t_4]\}$$

while the actual interfiring time for this transition is

$$\Gamma[t_5] = q\overline{\mathbf{s}}[t_3] + (1-q)\overline{\mathbf{s}}[t_4]$$

independently of the higher moments of the probability distribution functions associated with transitions t_3 and t_4. Therefore the bound given by Theorem 17.2 is non-reachable for the net in Figure 17.4.a.

As was announced in Section 17.1.1, an improvement of the bound can be obtained if additional constraints that improve the linear characterization of the average marking are considered. Now, we exploit the linear information that can be obtained from *traps* of the net and later we reinterpretate the obtained improvement of the bound in terms of *implicit* places.

A trap in a Petri net $\mathcal{N}$ is a subset of places Θ such that $\Theta^\bullet \subseteq {}^\bullet\Theta$. A well-known property of these structural elements is recalled now.

Property 17.4 *Let $\mathcal{S}$ be a Petri net system and $\Theta \subseteq P$ a trap. If Θ is initially marked, then Θ is marked throughout the net's evolution.*

This property can be expressed in algebraic terms considering the vector $\mathbf{y_\Theta}$ associated with a given trap Θ, and defined as

$$\mathbf{y_\Theta}[p] = \begin{cases} 1, & \text{if } p \in \Theta \\ 0, & \text{otherwise} \end{cases}$$

(i.e., the *characteristic function* of the set Θ). The next inductive invariant is true: if $\mathbf{y_\Theta} \cdot \mathbf{m_0} \geq 1$ then $\mathbf{y_\Theta} \cdot \mathbf{m} \geq 1$ for all reachable marking $\mathbf{m}$.

Let us consider the vector $\mathbf{y_\Theta}$ associated with a trap Θ of a net, and a P-semiflow $\mathbf{y}$ such that $\mathbf{y} - \mathbf{y_\Theta} \geq 0$ (it always exists for conservative nets). The following linear relations can be derived for all reachable marking $\mathbf{m}$ and for the average marking vector $\overline{\boldsymbol{\mu}}$:

$$(\mathbf{y} - \mathbf{y_\Theta}) \cdot \mathbf{m} \leq \mathbf{y} \cdot \mathbf{m_0} - 1$$

$$(\mathbf{y} - \mathbf{y_\Theta}) \cdot \overline{\boldsymbol{\mu}} \leq \mathbf{y} \cdot \mathbf{m_0} - 1$$

Premultiplying inequality (17.4) by $\mathbf{y} - \mathbf{y_\Theta}$, a lower bound for $\Gamma[t_1]$ is derived:

Theorem 17.5 *For any net $\mathcal{N}$ and for any trap Θ of $\mathcal{N}$, a lower bound for the average interfiring time $\Gamma[t_1]$ of transition t_1 is given by:*

$$\begin{array}{rll} \Gamma[t_1] \geq & \textit{maximum} & \dfrac{(\mathbf{y}-\mathbf{y_\Theta})\cdot \mathbf{Pre}\cdot \overline{\mathbf{D}}}{\mathbf{y}\cdot \mathbf{m_0}-1} \\ & \textit{subject to} & \mathbf{y}\cdot \mathbf{C} = \mathbf{0} \\ & & \mathbf{y}-\mathbf{y_\Theta} \geq \mathbf{0} \\ & & \mathbf{y_\Theta}[p] = \texttt{ if } p\in\Theta \texttt{ then } 1 \texttt{ else } 0 \end{array} \qquad (17.23)$$

Going back to the net in Figure 17.4.a, the unique minimal trap different from the P-semiflows is $\Theta = \{p_1, p_4, p_5\}$. Considering the P-semiflow $\mathbf{y} = (2,1,1,1,1)$, we have $\mathbf{y} \geq \mathbf{y_\Theta} = (1,0,0,1,1)$, and Theorem 17.5 can be applied:

$$\Gamma[t_5] \geq q\overline{\mathbf{s}}[t_3] + (1-q)\overline{\mathbf{s}}[t_4] \qquad (17.24)$$

Therefore the bound obtained in Section 17.2.1 using only P-semiflows has been improved (in fact the bound computed now coincides with the actual interfiring time for this particular example).

In order to explain in an intuitive way (with the example) the reason of the previous improvement, let us derive a behaviourally equivalent safe marked graph (Figure 17.4.b) for the safe Free Choice net of Figure 17.4.a, assuming for the sake of simplicity that the resolution of conflict at place p_1 is deterministic with $q = 1/2$ (i.e., transitions t_1 and t_2 fire once each one, alternatively). The lower bound for the average cycle time of this MG based on Theorem 17.2 (i.e., using the P-semiflows) is

$$\Gamma_{\mathrm{MG}} \geq \overline{\mathbf{s}}[t_3] + \overline{\mathbf{s}}[t_4]$$

(in fact it is reached) and corresponds to the circuit $\langle p_1, p_2, p_5, p_1', p_3, p_4' \rangle$. Since transition t_5 appears instantiated twice in the MG, the obtained bound for the interfiring time of this transition is

$$\Gamma[t_5] \geq (\overline{\mathbf{s}}[t_3] + \overline{\mathbf{s}}[t_4])/2$$

In the original Free Choice net there does not exist any minimal P-semiflow including both p_2 and p_3 in its support, thus the previous bound is not obtained.

Now, we reinterpretate the linear marking relations derived from traps using *implicit places* (cfr. Chapter 6). Let us consider again the net in Figure 17.4.a and its behaviourally equivalent (for $q = 1/2$) marked graph depicted in Figure 17.4.b. The elementary circuits (P-semiflows) of this MG are

$$\begin{aligned} c_1 &= \langle p_1, p_2, p_5, p_1', p_5' \rangle \\ c_2 &= \langle p_1, p_4, p_1', p_3, p_4' \rangle \\ c_3 &= \langle p_1, p_2, p_5, p_1', p_3, p_4' \rangle \\ c_4 &= \langle p_1, p_4, p_1', p_5' \rangle \end{aligned}$$

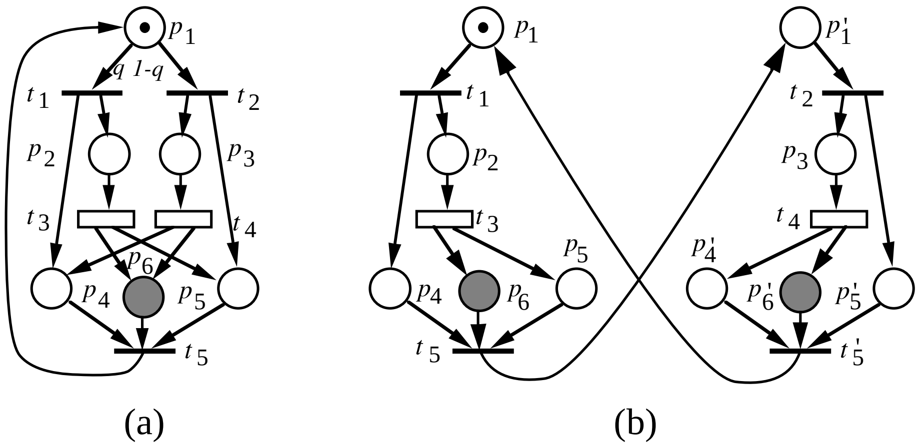

Figure 17.5: The same systems (a) and (b) of Figure 17.4 with the addition of implicit places (a) p_6 and (b) p_6 and p'_6, respectively.

The circuits c_1 and c_2 correspond to the elementary P-semiflows of the original net $\mathbf{y_1}$ and $\mathbf{y_2}$, respectively. Thus, these circuits cannot contribute to the improvement of the bound computed for the original net based on the P-semiflows. This is not the case for the circuits c_3 and c_4. These circuits add linear information which is not reflected by P-semiflows in the original net. The circuit c_4 does not include any timed transition and must not be considered. On the other hand, the circuit c_3 reflects the sequentialization of transitions t_3 and t_4, and it gives the actual cycle time of the net system.

A given elementary circuit of the derived MG does not correspond to any elementary P-semiflow of the original Free Choice net when it includes several instances of a unique transition and each instance has as input (or output) places which are instances of different original places. This is the case, for example, for the circuit c_3 of the MG of Figure 17.4.b. It includes instances t_5 and t'_5 of a unique transition, and the input places of these transitions in circuit c_3 are p_5 and p'_4, respectively, which are instances of different original places.

Now, let us increment the number of circuits of the MG of Figure 17.4.b, by adding the places p_6 and p'_6, as depicted in Figure 17.5.b. Places p_6 and p'_6 are replicas of places p_5 and p'_4, respectively (thus they are implicit), and can be supposed to be different instances of a new (implicit) place in the original net (place p_6 of the net in Figure 17.5.a). The addition of this place generates a new elementary P-semiflow $\mathbf{y_3} = (1, 1, 1, 0, 0, 1)$. With this P-semiflow, the lower bound for the interfiring time computed with the linear programming problem in (17.9) is

$$\Gamma[t_5] \geq q\overline{\mathbf{s}}[t_3] + (1-q)\overline{\mathbf{s}}[t_4]$$

which is the same obtained in (17.24), using relations derived from trap structures stated in Theorem 17.5.

Let us remark that the relation between the implicit place p_6 of the net in Figure 17.5.a and the trap $\Theta = \{p_1, p_4, p_5\}$ considered previously is straightforward: $\mathbf{C}[p_6, \cdot] = \mathbf{y_\Theta} \cdot \mathbf{C}$, that is, the incidence vector of p_6 is the sum of those of places p_1, p_4, and p_5.

The following linear relation can be derived from the trap Θ (and the P-semiflow $\mathbf{y} = \mathbf{y_1} + \mathbf{y_2}$):

$$(\mathbf{y} - \mathbf{y_\Theta}) \cdot \mathbf{m} = \mathbf{m}[p_1] + \mathbf{m}[p_2] + \mathbf{m}[p_3] \leq 1, \forall \mathbf{m} \in \mathrm{RS}(\mathcal{N}, \mathbf{m_0}) \qquad (17.25)$$

While this one follows from the new P-semiflow $\mathbf{y_3}$ that includes the implicit place p_6:

$$\mathbf{y_3} \cdot \mathbf{m} = \mathbf{m}[p_1] + \mathbf{m}[p_2] + \mathbf{m}[p_3] + \mathbf{m}[p_6] = 1, \forall \mathbf{m} \in \mathrm{RS}(\mathcal{N}, \mathbf{m_0}) \qquad (17.26)$$

It can be pointed out that the information given by relation (17.25) is included in that given by the new P-semiflow (equation (17.26)), because $\mathbf{m}[p_6] \geq 0$.

Now, technical details related with the addition of implicit places which improve the throughput upper bound computed by means of P-semiflows and traps are considered.

Let us consider an initially marked trap Θ of a given net $\mathcal{N}$, and its associated vector $\mathbf{y_\Theta}$ defined as in previous paragraphs. The following result, which follows from Property 6.1.2 in Chapter 6, assures that a structurally implicit place p_Θ associated with Θ, can be added to $\mathcal{N}$.

Property 17.6 *Let Θ be an initially marked trap of $\mathcal{N}$, $\mathbf{y_\Theta}[p] =$ `if` $p \in \Theta$ `then` 1 `else` 0, $\mathbf{y_\Theta} \cdot \mathbf{m_0} \geq 1$, and $\mathcal{N}^{p_\Theta}$ the net resulting from the addition of place p_Θ with incidence vector $\mathbf{C}[p_\Theta, \cdot] = \mathbf{y_\Theta} \cdot \mathbf{C}$ to $\mathcal{N}$. Then p_Θ is structurally implicit in $\mathcal{N}^{p_\Theta}$.*

The importance of the previous structural implicit place lies on the fact that, if a marking makes it implicit (e.g., the marking given by Property 6.1.2 in Chapter 6), then the lower bound for the average interfiring time of a transition computed using P-semiflows of the augmented net can improve the bounds based on P-semiflows of the original net (Theorem 17.2) and on the trap Θ (Theorem 17.5). Before to state this result we firstly present a technical lemma.

Lemma 17.7 *Let Θ be an initially marked trap of $\mathcal{N}$ and $\mathbf{y_\Theta}[p] =$ `if` $p \in \Theta$ `then` 1 `else` 0. Let p_Θ be a place defined as $\mathbf{C}[p_\Theta, \cdot] = \mathbf{y_\Theta} \cdot \mathbf{C}$. Then the pair composed by $\mathbf{y_\Theta}$ and $\mu_\Theta = -1$ is a feasible solution of the linear programming problem of Property 6.1.2 in Chapter 6, and $\mathbf{m_0}[p_\Theta] \leq \mathbf{y_\Theta} \cdot \mathbf{m_0} + \mu_\Theta$ (place p_Θ is assumed to be pure, i.e., selfloop-free).*

Proof:
$\mathbf{y_\Theta} \cdot \mathbf{C} = \mathbf{C}[p_\Theta, \cdot] \Rightarrow \forall t \in {}^\bullet p_\Theta : \mathbf{y_\Theta} \cdot \mathbf{Post}[\cdot, t] - \mathbf{y_\Theta} \cdot \mathbf{Pre}[\cdot, t] = -\mathbf{Pre}[p_\Theta, t]$.

Taking into account that $\mathbf{y_\Theta}$ is the characteristic function of a trap we have in the last equality that $\mathbf{y_\Theta}\cdot\mathbf{Pre}[\cdot,t] > 0$ if and only if $\mathbf{y_\Theta}\cdot\mathbf{Post}[\cdot,t] > 0$. Therefore, $\forall t \in {}^\bullet p_\Theta$: $\mathbf{y_\Theta} \cdot \mathbf{Pre}[\cdot,t] > \mathbf{Pre}[p_\Theta,t]$ and from the linear programming problem of Property 6.1.2 in Chapter 6 we conclude that $\mathbf{y_\Theta}$ and $\mu_\Theta = -1$ is a feasible solution. From the same linear programming problem we also conclude directly that $\mathbf{m_0}[p_\Theta] \leq \mathbf{y_\Theta} \cdot \mathbf{m_0} + \mu_\Theta$ because $\mathbf{y_\Theta} \cdot \mathbf{m_0} \geq 1$. $\diamondsuit$

Theorem 17.8 *Let $\langle \mathcal{N}, \mathbf{m_0}\rangle$ be a net system, Θ an initially marked trap of $\mathcal{N}$, $\mathbf{y_\Theta}[p] =$ `if` $p \in \Theta$ `then` 1 `else` 0, and $\langle \mathcal{N}^{p_\Theta}, \mathbf{m_0}^{p_\Theta}\rangle$ the net system resulting from the addition to the original net of the structural implicit place p_Θ with incidence vector $\mathbf{C}[p_\Theta,\cdot] = \mathbf{y_\Theta}\cdot\mathbf{C}$ and with $\mathbf{m_0}^{p_\Theta}[p_\Theta]$ given by Property 6.1.2 in Chapter 6.*

1. *A lower bound $\Gamma^{p_\Theta}[t_1]$ for the average interfiring time $\Gamma[t_1]$ of transition t_1 in $\langle \mathcal{N}, \mathbf{m_0}\rangle$ can be computed applying Theorem 17.2 to the system $\langle \mathcal{N}^{p_\Theta}, \mathbf{m_0}^{p_\Theta}\rangle$.*

2. *If $\Gamma^{PS}[t_1]$ and $\Gamma^{\Theta}[t_1]$ are the lower bounds of $\Gamma[t_1]$ derived from the direct application of Theorems 17.2 and 17.5, respectively, to the original system, then $\Gamma^{p_\Theta}[t_1] \geq \Gamma^{PS}[t_1]$ and $\Gamma^{p_\Theta}[t_1] \geq \Gamma^{\Theta}[t_1]$.*

Proof:
$\Gamma^{p_\Theta}[t_1]$ is a lower bound for the average interfiring time of t_1 in $\langle \mathcal{N}^{p_\Theta}, \mathbf{m_0}^{p_\Theta}\rangle$ by Theorem 17.2. Since p_Θ is implicit, t_1 has the same average interfiring time in $\langle \mathcal{N}, \mathbf{m_0}\rangle$ and in $\langle \mathcal{N}^{p_\Theta}, \mathbf{m_0}^{p_\Theta}\rangle$. Then, $\Gamma^{p_\Theta}[t_1]$ is a lower bound for the average interfiring time of t_1 in $\langle \mathcal{N}, \mathbf{m_0}\rangle$.

$\Gamma^{p_\Theta}[t_1] \geq \Gamma^{PS}[t_1]$ because if $\mathbf{y}$ is a P-semiflow of $\mathcal{N}$, then $\mathbf{z} = [\mathbf{y} \mid 0]$ is a P-semiflow of $\mathcal{N}^{p_\Theta}$.

Finally, we prove that $\Gamma^{p_\Theta}[t_1] \geq \Gamma^{\Theta}[t_1]$. Let $\mathbf{y}$ be a P-semiflow of $\mathcal{N}$ such that $\mathbf{y} - \mathbf{y_\Theta} \geq \mathbf{0}$. Then $\mathbf{z} = [(\mathbf{y} - \mathbf{y_\Theta}) \mid 1]$ is a P-semiflow of $\mathcal{N}^{p_\Theta}$. Now, applying equation (17.6) for $\Gamma^{p_\Theta}[t_1]$:

$$\begin{aligned}\Gamma^{p_\Theta}[t_1] &\geq \frac{[(\mathbf{y} - \mathbf{y_\Theta}) \mid 1]\cdot\mathbf{Pre}^{p_\Theta}\cdot\overline{\mathbf{D}}}{\mathbf{y}\cdot\mathbf{m_0} - \mathbf{y_\Theta}\cdot\mathbf{m_0} + \mathbf{m_0}^{p_\Theta}[p_\Theta]} = \\ &= \frac{(\mathbf{y} - \mathbf{y_\Theta})\cdot\mathbf{Pre}\cdot\overline{\mathbf{D}}}{\mathbf{y}\cdot\mathbf{m_0} - \mathbf{y_\Theta}\cdot\mathbf{m_0} + \mathbf{m_0}^{p_\Theta}[p_\Theta]} + \frac{\mathbf{Pre}[p_\Theta,\cdot]\cdot\overline{\mathbf{D}}}{\mathbf{y}\cdot\mathbf{m_0} - \mathbf{y_\Theta}\cdot\mathbf{m_0} + \mathbf{m_0}^{p_\Theta}[p_\Theta]}\end{aligned} \tag{17.27}$$

And this value is greater than or equal to that given by equation (17.23) in Theorem 17.5 because the second term of the above sum is non negative and the first term in the above sum is greater than or equal to that given by equation (17.23) in Theorem 17.5 (take into account that $\mathbf{m_0}^{p_\Theta}[p_\Theta] \leq \mathbf{y_\Theta}\cdot\mathbf{m_0} - 1$, by Lemma 17.7, and that the denominator is less than or equal to $\mathbf{y}\cdot\mathbf{m_0} - 1$). $\diamondsuit$

Theorem 17.8.2 tells that the addition of implicit places allows to obtain better bounds than those computed using traps or the P-semiflows of the original net. The problems that remain are how to add implicit places (i.e., which ones allow to improve the bounds) and when no more improvements are possible with the technique.

Previously, the net system of Figure 17.4.a has been considered as an example in which the bound computed using the trap $\Theta = \{p_1, p_4, p_5\}$ is tight because it reaches the actual value of the average interfiring time. It is also shown that the same value can be derived, after the addition, in Figure 17.5.a, of the associated implicit place p_6, considering the new P-semiflow $(1,1,1,0,0,1)$.

Let us consider the same net system of Figure 17.4.a, but assuming now that transition t_5 is not immediate but timed, with average service time equal to $\overline{\mathbf{s}}[t_5]$. If the linear programming problem in (17.9) is applied to the net, the following bound is obtained:

$$\Gamma^{PS}[t_5] = \max\{q\overline{\mathbf{s}}[t_3] + \overline{\mathbf{s}}[t_5], (1-q)\overline{\mathbf{s}}[t_4] + \overline{\mathbf{s}}[t_5]\}$$

If trap $\Theta = \{p_1, p_4, p_5\}$ and P-semiflow $\mathbf{y} = (2,1,1,1,1)$ are considered, inequality (17.23) gives the bound:

$$\Gamma^{\Theta}[t_5] = q\overline{\mathbf{s}}[t_3] + (1-q)\overline{\mathbf{s}}[t_4]$$

If the implicit place associated with Θ is added to the net, Theorem 17.8 gives the bound:

$$\Gamma^{p_\Theta}[t_5] = q\overline{\mathbf{s}}[t_3] + (1-q)\overline{\mathbf{s}}[t_4] + \overline{\mathbf{s}}[t_5]$$

(for the P-semiflow $(1,1,1,0,0,1)$), which improves both $\Gamma^{PS}[t_5]$ and $\Gamma^{\Theta}[t_5]$, and, in fact, it gives the actual interfiring time of transition t_5 (i.e., it is tight). Note that, in this case, the improvement is due to the non-null second term of the expression (17.27).

17.3.2 The Role of Liveness Bounds of Transitions

In a classical product-form QN, the number of servers at each station is explicitly given as a modelling choice (e.g., it can be said that a certain station has two servers). Stations may vary between *single* server and *delay* node (infinite server). In the second case, the maximum number of servers that can be working at such delay node is exactly the number of customers in the whole net system.

Since in this chapter we assume infinite server semantics for transitions, several instances of a same transition can work in parallel at a given marking. How many of them? The answer is given by the *degree of enabling* of a transition, t, at a given marking, $\mathbf{m}$, defined in Chapter 8 as:

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

Therefore it can be said that at $\mathbf{m}$, in transition t, $\mathbf{e}[t](\mathbf{m})$ servers work in parallel. This value can be eventually reduced by a design choice adding a self-loop place around t with q tokens: it is obvious that in this case $\mathbf{e}[t](\mathbf{m}) \leq q$.

The maximum number of servers working in parallel clearly influences the performance of the system. This value, in net systems terms, has been called the *enabling bound* of a transition.

Definition 17.9 *Let $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ be a net system. The enabling bound of a given transition t of $\mathcal{S}$ is*

$$\mathbf{eb}[t] = \sup\{k \in \mathbb{N} : \exists\ \mathbf{m_0} \xrightarrow{\sigma} \mathbf{m}, \forall p \in {}^\bullet t, \mathbf{m}[p] \geq k\ \mathbf{Pre}[p,t]\} \tag{17.28}$$

The enabling bound is a quantitative generalization of the basic concept of enabling, and is closely related to the concept of *marking bound of a place* (see Chapter 6).

Since we are interested in the steady-state performance of a model, one can ask the following question: how many servers can be available in transitions in any possible steady-state condition? The answer is given by the definition of the *liveness bound* concept.

Definition 17.10 *Let $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ be a net system. The liveness bound of a given transition t of $\mathcal{S}$ is*

$$\mathbf{lb}[t] = \sup\{k \in \mathbb{N} : \forall \mathbf{m}', \mathbf{m_0} \xrightarrow{\sigma} \mathbf{m}', \exists \mathbf{m}, \mathbf{m}' \xrightarrow{\sigma'} \mathbf{m} \wedge \forall p \in {}^\bullet t, \mathbf{m}[p] \geq k\ \mathbf{Pre}[p,t]\} \tag{17.29}$$

The above definition generalizes the classical concept of liveness of a transition. In particular, a transition t is live if and only if $\mathbf{lb}[t] > 0$, i.e., if there is at least one working server associated with it in any steady-state condition. The following is also obvious from the definitions.

Property 17.11 *Let $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ be a net system. For any transition t in $\mathcal{S}$, $\mathbf{eb}[t] \geq \mathbf{lb}[t]$.*

The definition of enabling bound refers to a behavioural property. Since we are looking for computational techniques at the structural level, we define also the structural counterpart of the enabling bound concept. Essentially, the reachability condition is substituted by the (in general) weaker (linear) constraint that markings satisfy the net state equation: $\mathbf{m} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma}$, with $\mathbf{m}, \boldsymbol{\sigma} \geq \mathbf{0}$.

Definition 17.12 *Let $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ be a net system. The structural enabling bound, $\mathbf{seb}[t]$, of a given transition t of $\mathcal{S}$ is*

$$\begin{array}{rll} \mathbf{seb}[t] = & \textit{maximum} & k \\ & \textit{subject to} & \mathbf{m_0}[p] + \mathbf{C}[p,T] \cdot \boldsymbol{\sigma} \geq k\ \mathbf{Pre}[p,t],\ \forall p \in P \\ & & \boldsymbol{\sigma} \geq \mathbf{0} \end{array} \tag{17.30}$$

Note that the definition of structural enabling bound reduces to the formulation of a linear programming problem, that can be solved in polynomial time.

Now let us remark the relation between behavioural and structural enabling bound concepts that follows from the implication "$\mathbf{m_0} \xrightarrow{\sigma} \mathbf{m} \ \Rightarrow \ \mathbf{m} = \mathbf{m_0} + \mathbf{C} \cdot \boldsymbol{\sigma} \ \wedge \ \boldsymbol{\sigma} \geq \mathbf{0}$".

Property 17.13 *Let $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ be a net system. For any transition t in $\mathcal{S}$, $\mathbf{seb}[t] \geq \mathbf{eb}[t]$.*

As we remarked before, the concept of enabling bound of transitions is closely related to the marking bound of places. In an analogous way, the structural enabling bound is closely related to the structural marking bound of places (see Chapter 6).

For the particular case of live and bounded free choice systems (thus, in particular, for live and bounded marked graphs), the following result holds.

Theorem 17.14 *Let $\mathcal{S} = \langle \mathcal{N}, \mathbf{m_0} \rangle$ be a live and bounded free choice system. For any transition t in $\mathcal{S}$, $\mathbf{seb}[t] = \mathbf{eb}[t] = \mathbf{lb}[t]$.*

A "trivial" lower bound in steady-state performance for a live net system with a given vector of visit ratios for transitions is of course given by the inverse of the sum of the services times of all the transitions weighted by the vector of visit ratios. Since the net system is live, all transitions must be firable, and the sum of all service times multiplied by the number of occurrences of each transition in the average cycle of the model corresponds to any *complete sequentialization* of all the transition firings.

Theorem 17.15 *For any live and bounded system, an upper bound for the average interfiring time $\Gamma[t_1]$ of transition t_1 can be computed as follows:*

$$\Gamma[t_1] \leq \sum_{t \in T} \mathbf{v}[t] \overline{\mathbf{s}}[t] = \sum_{t \in T} \overline{\mathbf{D}}[t] \tag{17.31}$$

This *pessimistic* behaviour is always reached in a marked graph consisting on a single loop of transitions and containing a single token in one of the places, independently of the higher moments of the probability distribution functions (this observation can be trivially confirmed by the computation of the lower bound given by (17.19), which in this case gives the same value).

Before trying to improve this trivial bound let us first consider the case of 1–live (i.e., all its transitions have a liveness bound equal to 1) strongly-connected MG's. If we specify only the mean values of the transition service times and not the higher moments, we may always find an stochastic model whose steady-state throughput is arbitrarily close to the trivial lower bound, independently of the topology of the MG (only provided that it is 1–live). The formal proof of this (somewhat counter-intuitive) result stated in the next theorem is based on the definition of the family of random variables:

$$X_\mu^i(\epsilon) = \begin{cases} 0, & \text{with probability } 1-\epsilon^i \\ \mu/\epsilon^i, & \text{with probability } \epsilon^i \end{cases} \tag{17.32}$$

for $\mu \geq 0; 0 < \epsilon \leq 1; i \in \mathbb{N}$. It is straightforward to see that

$$\mathrm{E}\left[X_\mu^i(\epsilon)\right] = \mu \quad \text{and} \quad \mathrm{E}\left[X_\mu^i(\epsilon)^2\right] = \mu^2/\epsilon^i$$

This implies that the coefficient of variation is 0 for $\epsilon = 1$, and that it tends to ∞ as $\epsilon \to 0$ provided that $i > 0$. Then, the proof derives from considering that each transition t_j in the MG has $X_{\overline{\mathbf{s}}[t_j]}^{j-1}(\epsilon)$ as random service time distribution.

Theorem 17.16 *For any 1–live strongly-connected MG with a given specification of the average service times $\overline{\mathbf{s}}[t_j]$ for each $t_j \in T$, it is possible to assign probability distribution functions to the transition service times such that the average cycle time is $\Gamma = \sum_j \overline{\mathbf{s}}[t_j] - O(\epsilon)$, $\forall \epsilon : 0 < \epsilon \leq 1$, independently of the topology of the net (and thus independently of the potential maximum degree of parallelism intrinsic in the MG)* [2].

Proof:
By construction, we will show that the association of the family of random variables $X_{\overline{\mathbf{s}}[t_j]}^{j-1}(\epsilon)$, defined in (17.32), with each transition $t_j \in T$ yields exactly the cycle time Γ claimed by the theorem. To give the proof, we will consider a sequence of models ordered by the index of transitions, in which the qth model of the sequence has transitions $t_1, t_2, \ldots, t_q$ timed with the random variables $X_{\overline{\mathbf{s}}[t_j]}^{j-1}(\epsilon)$, and all other transitions immediate (firing in zero time); the $|T|$th model in the sequence represents an example of attainability of the lower bound on throughput (upper bound for the average cycle time), independent of the net topology. Now we will prove by induction that the qth model in the sequence has a cycle time $\Gamma_q = \sum_{j=1}^q \overline{\mathbf{s}}[t_j] - O(\epsilon)$.

Base ($q = 1$): trivial since the repetitive cycle that constitutes the steady-state behaviour of the MG contains only one (single-server) deterministic transition with average service time $\Gamma_1 = \overline{\mathbf{s}}[t_1]$.

Induction step ($q > 1$): taking the limit $\epsilon \to 0$, the newly timed transition t_q will fire most of the time with time zero, thus normally not contributing to the computation of the cycle time, that will be just $\Gamma_{q-1} = \sum_{j=1}^{q-1} \overline{\mathbf{s}}[t_j] - O(\epsilon)$ (as in the case of model $q-1$) with probability $1-\epsilon^{q-1}$. On the other hand, the newly timed transition has a (very small) probability ϵ^{q-1} of delaying its firing by a time $\overline{\mathbf{s}}[t_q]/\epsilon^{q-1}$, which is at least of order $1/\epsilon$ bigger than any other service time in the circuit, so that in this case all other transitions will wait for the firing of t_q, after having completed their possible current service in a time which is $O(\epsilon)$ lower than the service time of t_q itself (i.e., $\overline{\mathbf{s}}[t_q]/\epsilon^{q-1} = \Gamma_{q-1}/O(\epsilon)$). Therefore we obtain that $\Gamma_q = (1-\epsilon^{q-1})\Gamma_{q-1} + \epsilon^{q-1}(\frac{\overline{\mathbf{s}}[t_q]}{\epsilon^{q-1}} - O(\epsilon)) = \sum_{j=1}^q \overline{\mathbf{s}}[t_j] - O(\epsilon)$. $\diamondsuit$

[2] We use here the notation $O(f(x))$ to indicate any function $g(x)$ such that $\lim_{x\to 0} \frac{g(x)}{f(x)} \leq k \in \mathbb{R}$.

In the previous result, the upper bound for the average cycle time (thus lower bound on throughput) is reached in a limit case ($\epsilon \rightarrow 0$) in which the random variables associated with transitions have infinite coefficient of variation. This is a way to obtain the minimum throughput if service times associated with transitions are assumed mutually uncorrelated. It can be shown that it is also possible to reach the lower bound in performance for finite coefficient of variation if a *maximum negative correlation* is assumed among the service times of transitions.

Until now, we have shown that the trivial sum of the average service times of all transitions in the system constitutes a tight (attainable) lower bound for the performance of a live and safe MG (or more generally of a 1–live strongly-connected MG, but otherwise independently of the topology) in which only the mean values and neither the probability distribution functions nor the higher moments are specified for the transition service times. Let us now extend this result to the more general case of k–live strongly-connected MG's.

An intuitive idea is to derive a lower bound on throughput for an MG containing transitions with liveness bound $k \geq 1$ (remember that, for MG's, $\mathbf{lb} = \mathbf{seb}$) by taking the method used for the computation of the throughput upper bound in Section 17.2.3, and substitute in it the "max" operator for the sum of the service times of all transitions involved. After some manipulation to avoid counting more than once the contribution of the same transition, one can arrive at the formulation of the following value for the maximum cycle time:

$$\Gamma \leq \sum_{t \in T} \frac{\overline{\mathbf{s}}[t]}{\mathbf{seb}[t]} \tag{17.33}$$

The proof of this result requires the following Lemma.

Lemma 17.17 *Any strongly-connected MG with arbitrary initial marking can be constrained to contain a main circuit including all transitions, without changing their liveness bound. This main circuit (which, in general, is not unique) contains a number of tokens equal to the maximum of the liveness bounds among all transitions. In addition there are other minor circuits that preserve the liveness bounds for transitions with bound lower than the maximum.*

Proof:
To construct an MG of the desired form we can apply the following iterative procedure that interleaves two non-disjoint circuits into a single one. Since the MG is strongly-connected each node belongs to at least one circuit; moreover, since the original MG is finite and each circuit cannot contain the same node more than once, this circuit interleaving procedure must terminate after a finite number of iterations. To reduce the number of circuits, implicit places created after each iteration can be removed. The iteration step is the following:

1. Take two arbitrary non-disjoint circuits (unless the MG already contains a main circuit including all nodes, there always exists such a pair of circuits because the MG is strongly-connected).

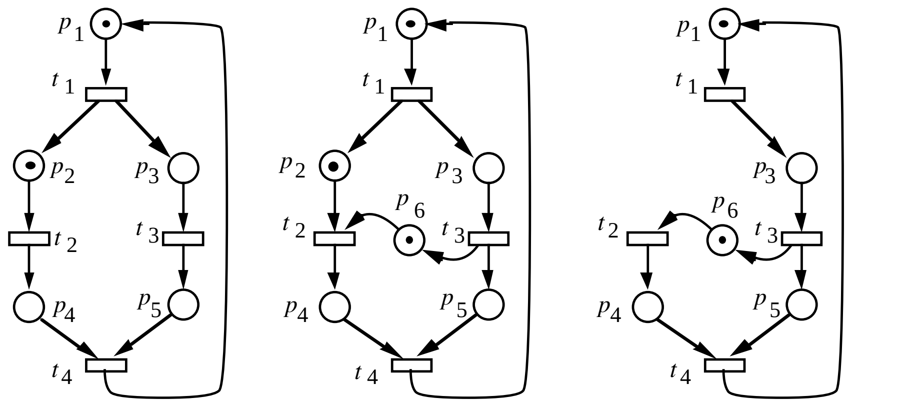

Figure 17.6: Example of structural sequentialization.

2. Combine them in a single circuit in such a way that the partial order among transitions given by the two original circuits is substituted by a compatible but otherwise arbitrary total order. This combination can be obtained by adding new places that are connected as input for a transition of one circuit and output for a transition of the other circuit that we decide must follow in the sequence determined by the new circuit we are creating.

3. Mark the new added places in such a way that the new circuit contains the same number of tokens as the maximum of the number of tokens in the two original circuits.

The above procedure is applied iteratively until all transitions are constrained into a single main circuit. At this point, we can identify and eliminate the implicit places that have been created during the circuits interleaving procedure. We obtain then an MG composed of one main circuit containing $c = \max_{t \in T} \mathbf{seb}[t]$ tokens that connects all transitions, and a certain number of minor circuits containing less tokens than c that maintain the liveness bound of the other transitions. $\diamondsuit$

The idea behind this constraint is to introduce a structural sequentialization among all transitions, thus potentially reducing the degree of concurrency between the activities modeled by the transitions. In other words, from the partial order given by the initial MG structure, we try to derive a total order without changing the liveness bound.

An example of application of the Lemma follows, in order to clarify the procedure. Consider the system depicted in Figure 17.6.a. This system contains only two circuits, namely $\langle t_1, t_2, t_4 \rangle$, and $\langle t_1, t_3, t_4 \rangle$; we can then add either the circuit $\langle t_1, t_2, t_3, t_4\rangle$ or $\langle t_1, t_3, t_2, t_4\rangle$; Figure 17.6.b depicts the resulting system in case we choose to add the second circuit. In this case only place p_6 (from t_3 to t_2) needs to be added to obtain the longer circuit, and it should be marked with one token, so that the new circuit comprising places $\langle p_1, p_3, p_6, p_4\rangle$ contains two tokens, as the original circuit $\langle p_1, p_2, p_4\rangle$ (while the other original circuit $\langle p_1, p_3, p_5\rangle$ contained only one). In our example, we need not iterate the procedure since we have already obtained a circuit containing all transitions of the MG. At this point we can identify and eliminate the implicit places that have been created during the circuits interleaving procedure. In the present example, we can easily see that place p_2 becomes implicit in Figure 17.6.b, so that it can be eliminated, finally leading ourselves to the MG depicted in Figure 17.6.c.

It should be evident that the MG transformed by applying the above Lemma has an average cycle time which is greater than or equal to the average cycle time of the original one, since some additional constraints have been added to the enabling of transitions: hence the average cycle time of the transformed MG is a lower bound for the performance of the original one. Now if $c = \max_{t\in T} \mathbf{seb}[t] = 1$ in the above Lemma, we re-find the lower bound of Theorem 17.16. In the case of $c > 1$, we can show that the average cycle time of the transformed system cannot exceed $\sum_{t\in T} \overline{\mathbf{s}}[t]/\mathbf{seb}[t]$.

Theorem 17.18 *For any live and bounded marked graph, an upper bound for the average cycle time* Γ *can be computed as follows:*

$$\Gamma \leq \sum_{t\in T} \frac{\overline{\mathbf{s}}[t]}{\mathbf{seb}[t]} \tag{17.34}$$

Moreover, this upper bound for the average cycle time is reachable for any MG topology and for some assignment of probability distribution functions to the service time of transitions (i.e., the bound cannot be improved).

Proof:
Without loss of generality, assume that transitions in the system resulting from the application of Lemma 17.17 are partitioned in two classes S_2 and S_1, with liveness bounds $K_2 = c > 1$ and $K_1 < c$ (where $c = \max_{t\in T} \mathbf{seb}[t]$), respectively (the proof is easily extended to the case of more than two classes). Construct a new model containing only K_1 tokens in the main circuit; at this point all transitions behave as K_1–servers, so that the cycle time is given by the sum of the service times of all transitions, divided by the total number of customers in the main loop K_1; moreover, the delay time for the transitions belonging to class S_1 is simply given by $D_1 = \sum_{t\in S_1} \overline{\mathbf{s}}[t]$. Now if we increase the number of tokens in the main loop from K_1 to K_2, the delay time of S_1 cannot increase, so that the contribution of S_1 to the cycle time cannot exceed D_1 for each of the first K_1 tokens. Under the hypothesis that the throughput of the system $\boldsymbol{\chi}[t_1]$ is given by the inverse of $\sum_{t\in T} \overline{\mathbf{s}}[t]/\mathbf{seb}[t]$, the average number of tokens of the main loop computed using Little's formula cannot exceed $N_1 = \boldsymbol{\chi}[t_1] D_1$, therefore the average number of tokens available to fire transitions in S_2 cannot be lower than

$$N_2 = K_2 - N_1 = K_2 \frac{\frac{K_2-K_1}{K_1}\sum_{t\in S_1}\overline{\mathbf{s}}[t] + \sum_{t\in S_2}\overline{\mathbf{s}}[t]}{\sum_{t\in S_2}\overline{\mathbf{s}}[t] + \frac{K_2}{K_1}\sum_{t\in S_1}\overline{\mathbf{s}}[t]}$$

On the other hand, we need only

$$N_2 = \boldsymbol{\chi}[t_1]D_2 = K_2 \frac{D_2}{\sum_{t\in S_2}\overline{\mathbf{s}}[t] + \frac{K_2}{K_1}\sum_{t\in S_1}\overline{\mathbf{s}}[t]}$$

tokens to sustain throughput $\boldsymbol{\chi}[t_1]$ in subnet S_2, so that we are assuming a delay in S_2:

$$D_2 \leq \frac{K_2 - K_1}{K_1}\sum_{t\in S_1}\overline{\mathbf{s}}[t] + \sum_{t\in S_2}\overline{\mathbf{s}}[t]$$

Now we claim that this is the actual maximum delay because the first K_1 tokens can proceed at the maximum speed in the whole system, thus experiencing only delay $\sum_{t\in S_2}\overline{\mathbf{s}}[t]$ in subnet S_2, while the remaining $K_2 - K_1$ tokens can also queue up for traveling through S_1, thus experiencing an additional delay of $\frac{1}{K_1}\sum_{t\in S_1}\overline{\mathbf{s}}[t]$ each.

With respect to the reachability of the bound, we proceed by construction, in a way very similar to that of Theorem 17.16. The only technical difference is that now, without any loss of generality, we assume first of all to enumerate transitions in non-increasing order of liveness bound (or, equivalently, of structural enabling bound) i.e., rename the transitions in such a way that $\forall t_i, t_j \in T$, $i > j \implies \mathbf{seb}[t_i] \leq \mathbf{seb}[t_j]$. Then, as in the case of Theorem 17.16, we can show that the association of the family of random variables $X^{j-1}_{\overline{\mathbf{s}}[t_j]}(\epsilon)$ with each transition $t_j \in T$ yields exactly the bound for the cycle time claimed by the theorem. To give the proof we consider a sequence of models ordered by the index of transitions, in which the qth model of the sequence has transitions $t_1, t_2, \ldots, t_q$ timed with the random variables $X^{j-1}_{\overline{\mathbf{s}}[t_j]}(\epsilon)$, and all other transitions immediate (firing in zero time); the $|T|$th model in the sequence represents the resulting model that is expected to provide the example of attainability of the lower bound. By induction we prove that the qth model in the sequence has a cycle time:

$$\Gamma_q = \sum_{j=1}^{q}\frac{\overline{\mathbf{s}}[t_j]}{\mathbf{seb}[t_j]} - O(\epsilon)$$

Base ($q = 1$): trivial since the repetitive cycle that constitute the steady-state behavior of the MG contains only one ($\mathbf{seb}[t_1]$–server) deterministic transition with average service time $\Gamma_1 = \overline{\mathbf{s}}[t_1]/\mathbf{seb}[t_1]$.

Induction step ($q > 1$): taking the limit $\epsilon \to 0$, each server of the newly timed transition t_q will fire most of the times with time zero, thus normally not disturbing the behavior of the other timed transitions, and not contributing to the computation of the cycle time, that will be just $\Gamma_q = \sum_{j=1}^{q-1} \frac{\overline{\mathbf{s}}[t_j]}{\mathbf{seb}[t_j]} - O(\epsilon)$ (as in the case of model $q-1$) with probability $1 - \epsilon^{q-1}$. On the other hand, each of the servers of the newly timed transition has a (very small) probability ϵ^{q-1} of delaying its firing of a time $\overline{\mathbf{s}}[t_q]/\epsilon^{q-1}$, which is at least order of $1/\epsilon$ bigger than any other service time in the circuit. Now if $\mathbf{seb}[t_q] = 1$, then the proof is completed, since also $\forall j > q$, $\mathbf{seb}[t_j] = 1$ by hypothesis, and we reduce to the induction step of the proof of Theorem 17.16. Instead if $\mathbf{seb}[t_q] > 1$ then we can consider $\mathbf{seb}[t_q]$ consecutive firings of t_q, and compute the average service time as the total time to fire $\mathbf{seb}[t_q]$ times the transition, divided by $\mathbf{seb}[t_q]$. Now if we consider m consecutive firings of instances of transition t_q, we obtain an average delay:

$$\sum_{j=0}^{m-1} (1 - \epsilon^{q-1})^j \epsilon^{(q-1)(m-j)} \frac{(m-j)\overline{\mathbf{s}}[t_q]}{\epsilon^{(q-1)}} = \overline{\mathbf{s}}[t_q](1 + O(\epsilon))$$

Therefore, the average cycle time of the qth model will be:

$$\Gamma_q = (1 - O(\epsilon^{q-1}))\Gamma_{q-1} + \frac{\overline{\mathbf{s}}[t_q]}{\mathbf{seb}[t_q]}(1 + O(\epsilon)) = \sum_{j=1}^{q} \frac{\overline{\mathbf{s}}[t_j]}{\mathbf{seb}[t_j]} - O(\epsilon).$$

$\diamondsuit$

The same idea for the improvement of the lower bound based on liveness bounds of transitions that has been presented for marked graphs can be applied for live and bounded Free Choice systems in order to improve the trivial bound of Theorem 17.15.

Theorem 17.19 *For any live and bounded Free Choice system, an upper bound for the average interfiring time $\Gamma[t_1]$ of transition t_1 can be computed as follows:*

$$\Gamma[t_1] \leq \sum_{t \in T} \frac{\mathbf{v}[t]\overline{\mathbf{s}}[t]}{\mathbf{seb}[t]} = \sum_{t \in T} \frac{\overline{\mathbf{D}}[t]}{\mathbf{seb}[t]} \qquad (17.35)$$

Proof:
Let us consider a deterministic conflicts resolution policy. A strongly connected MG with the same relative throughput vector can be constructed as follows (in fact, since for the MG $\mathbf{v} = \mathbf{1}$, what can appear are several instances of transitions to get the $\mathbf{v}$ of the original net):

1. Steady-state markings must be home states. Let $\mathbf{m_h}$ be one of the home states (there always exist some for live and bounded free choice systems, and substitute it to the initial marking (i.e., $\langle \mathcal{N}, \mathbf{m_h} \rangle$ is reversible).

2. From the live and bounded free choice system, a safe marking can be derived preserving liveness, removing tokens from $\mathbf{m_h}$.

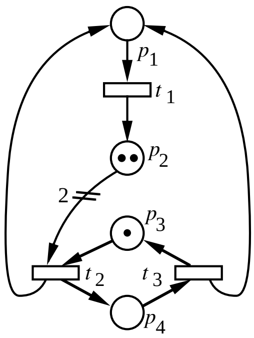

Figure 17.7: “Non-trivial” upper bound for the average interfiring time cannot be applied.

3. Develop the process, resolving the conflicts with the deterministic given policy, until cyclicity appears and the relative firing frequency holds. A safe MG is obtained in which transitions appear according to their relative firing frequencies.

4. The rest of tokens at each place in $\mathbf{m_h}$ in the original live and bounded free choice system can be added now in the corresponding place of the MG.

The actual interfiring time of the original free choice system (with deterministic conflicts resolution policy) is less than or equal to the one of the derived MG because the behaviour of the system has been constrained. Now, apply the bound obtained in Theorem 17.18. Different instances of a given transition are considered in the relative rate of the corresponding component in the relative firing frequency vector. Thus, the bound obtained for the derived MG applying Theorem 17.18 coincides with the bound obtained for the original system using the formula stated in this theorem. The theorem follows because $\mathbf{lb} = \mathbf{seb}$ for live and bounded free choice systems. $\diamondsuit$

Concerning non-free choice systems, only the trivial bound, given by the sum of the average service times of all transitions weighted by the vector of visit ratios, can be computed.

An example showing that the bound presented in Theorem 17.19 is not valid for non-free choice systems is depicted in Figure 17.7, where $\overline{\mathbf{s}}[t_1], \overline{\mathbf{s}}[t_2], \overline{\mathbf{s}}[t_3]$ are the average service times of transitions t_1, t_2, t_3, respectively. For this system, the vector of visit ratios normalized for transition t_2 (i.e., such that $\mathbf{v}[t_2] = 1$) is $\mathbf{v} = (2, 1, 1)$ and the liveness bounds of transitions are given by $\mathbf{lb}[t_1] = 2$, $\mathbf{lb}[t_2] = 1$, and $\mathbf{lb}[t_3] = 1$. Thus, the Theorem 17.19 would give the bound:

$$\Gamma[t_2] \leq \overline{\mathbf{s}}[t_1] + \overline{\mathbf{s}}[t_2] + \overline{\mathbf{s}}[t_3] \tag{17.36}$$

If exponentially distributed random variables (with means $\overline{\mathbf{s}}[t_1], \overline{\mathbf{s}}[t_2], \overline{\mathbf{s}}[t_3]$;

$\overline{\mathbf{s}}[t_1] \neq \overline{\mathbf{s}}[t_3]$) are associated with transitions, the average interfiring time for transition t_2 is

$$\Gamma[t_2] = \overline{\mathbf{s}}[t_1] + \overline{\mathbf{s}}[t_2] + \overline{\mathbf{s}}[t_3] + \frac{\overline{\mathbf{s}}[t_1]^2}{2(\overline{\mathbf{s}}[t_1] + \overline{\mathbf{s}}[t_3])} \tag{17.37}$$

which is greater than the value obtained from the Theorem 17.19, thus the "non-trivial" bound does not hold in general.

17.4 Additional Improvements: Non-Insensitive Bounds

A brief overview of three additional techniques for the improvement of the bounds presented before are included in this section. The common factor to these new techniques is that all of them need additional assumptions on the form of the probability distribution functions associated to the service of transitions or on the conflict resolution policies. In this sense, the obtained bounds are non-insensitive.

17.4.1 Linear Relations between Second Order Moments

In this section, some linear equations are derived between second order moments of marking of places for the particular case of *Exponential Petri Nets* (EPN). In this context, EPN are timed PN such that:

- all the transitions have (independent) exponential service time (in particular, immediate transitions are not allowed);
- a single-server semantics is assumed for transitions (or, an infinite-server semantics but with the assumption that every transition of the net has a self-loop place with multiplicity one and initially marked with one token);
- a race policy is assumed for the firing of transitions.

The obtained linear equations can be added as new constraints to the linear programming problem (17.1) of Section 17.1.1, therefore, the bounds for linear functions of average marking and throughput presented there can eventually be improved.

In order to get the new linear relations, we apply the *uniformization technique* [21] to the stochastic marking process $\{\overline{\boldsymbol{\mu}}(\tau)\}_{\tau\geq 0}$ (a more detailed discussion can be found in [24]).

Consider that each transition $t \in T$ of the EPN is continuously working with independent and exponentially distributed service time of rate $\lambda_t = 1/\overline{\mathbf{s}}[t]$. When a service at transition t is completed at instant τ there are two possibilities:

- either $\mathbf{e}[t](\tau) = 1$, i.e., t is enabled at instant τ and the completion of the service corresponds with a (real) firing of the transition; or

- $\mathbf{e}[t](\tau) = 0$, i.e., t is disabled at instant τ therefore nothing happens and we say that a *fictive firing* occurs.

Let $\{\tau_n\}_{n>0}$ be the sequence of time epochs of real or fictive service completions in the EPN. Then $\{\tau_n\}_{n>0}$ is a Poisson process with parameter $\lambda = \sum_{t\in T} \lambda_t$.

Denote $A_t(n)$ the indicator function defined as:

$$A_t(n) = \begin{cases} 1, & \text{if the } n\text{th real or fictive service completion occurs at } t \in T \\ 0, & \text{otherwise} \end{cases}$$

Then, $\sum_{t\in T} A_t(n) = 1$ and, for any $t \in T$, $\{A_t(n)\}_{n>0}$ is a sequence of independent and identically distributed random variables, independent of $\{\overline{\boldsymbol{\mu}}(\tau_n)\}_{n>0}$, and such that $\Pr\{A_t(n) = 1\} = \lambda_t/\lambda$.

If the system is in steady state then *PASTA property* (Poisson process see time average) [1] holds and in order to study the limit expected value $\overline{\boldsymbol{\mu}}$ of $\{\overline{\boldsymbol{\mu}}(\tau)\}_{\tau\geq 0}$, defined as:

$$\overline{\boldsymbol{\mu}}[p] = \lim_{\tau\to\infty} \mathrm{E}\left[\overline{\boldsymbol{\mu}}[p](\tau)\right]$$

it is enough to analyse the process $\{\overline{\boldsymbol{\mu}}(\tau_n)\}_{n>0}$.

First, notice that the evolution of $\{\overline{\boldsymbol{\mu}}(\tau_n)\}_{n>0}$ is determined by the following equation:

$$\overline{\boldsymbol{\mu}}[p](\tau_{n+1}) = \begin{cases} \overline{\boldsymbol{\mu}}[p](\tau_n), & \text{if } A_t(n) = 1 \text{ and } \mathbf{e}[t](\tau_n) = 0 \\ \overline{\boldsymbol{\mu}}[p](\tau_n) + \mathbf{C}[p,t], & \text{if } A_t(n) = 1 \text{ and } \mathbf{e}[t](\tau_n) = 1 \end{cases} \tag{17.38}$$

From this basic evolution equation, it is possible to compute linear relations between second order moments of $\{\overline{\boldsymbol{\mu}}(\tau_n)\}_{n>0}$.

For any pair of places $p_1, p_2 \in P$, the expectation of the product of $\overline{\boldsymbol{\mu}}[p_1](\tau_{n+1})$ and $\overline{\boldsymbol{\mu}}[p_2](\tau_{n+1})$ conditioned to $\mathcal{F}_n$ (where $\mathcal{F}_n$ is the σ-field generated by the events up to τ_n) can be computed from (17.38):

$$\begin{aligned}
&\mathrm{E}\left[\overline{\boldsymbol{\mu}}[p_1](\tau_{n+1})\overline{\boldsymbol{\mu}}[p_2](\tau_{n+1}) \mid \mathcal{F}_n\right] = \\
&\quad \sum_{t\in T} \frac{\lambda_t}{\lambda}(1 - \mathbf{e}[t](\tau_n))\overline{\boldsymbol{\mu}}[p_1](\tau_n)\overline{\boldsymbol{\mu}}[p_2](\tau_n) + \\
&\quad \sum_{t\in T} \frac{\lambda_t}{\lambda}\mathbf{e}[t](\tau_n)\left(\overline{\boldsymbol{\mu}}[p_1](\tau_n) + \mathbf{C}[p_1,t]\right)\left(\overline{\boldsymbol{\mu}}[p_2](\tau_n) + \mathbf{C}[p_2,t]\right) = \\
&\quad \overline{\boldsymbol{\mu}}[p_1](\tau_n)\overline{\boldsymbol{\mu}}[p_2](\tau_n) + \sum_{t\in T} \frac{\lambda_t}{\lambda}\mathbf{e}[t](\tau_n)\mathbf{C}[p_1,t]\mathbf{C}[p_2,t] + \\
&\quad \sum_{t\in T} \frac{\lambda_t}{\lambda}\mathbf{e}[t](\tau_n)\overline{\boldsymbol{\mu}}[p_1](\tau_n)\mathbf{C}[p_2,t] + \sum_{t\in T} \frac{\lambda_t}{\lambda}\mathbf{e}[t](\tau_n)\overline{\boldsymbol{\mu}}[p_2](\tau_n)\mathbf{C}[p_1,t]
\end{aligned}$$

Then, taking the expectation in the above equation and later taking the limit $n \to \infty$ we obtain:

$$\sum_{t\in T} \lambda_t \mathbf{e}[t]\mathbf{C}[p_1,t]\mathbf{C}[p_2,t] + \sum_{t\in T} \lambda_t \mathbf{y}[p_1,t]\mathbf{C}[p_2,t] + \sum_{t\in T} \lambda_t \mathbf{y}[p_2,t]\mathbf{C}[p_1,t] = 0$$

where

$$\mathbf{e}[t] = \lim_{\tau\to\infty} \mathrm{E}\left[\mathbf{e}[t](\tau)\right]$$

$$\mathbf{y}[p,t] = \lim_{\tau\to\infty} \mathrm{E}\left[\overline{\boldsymbol{\mu}}[p](\tau)\mathbf{e}[t](\tau)\right]$$

Now, since $\lambda_t \mathbf{e}[t] = \boldsymbol{\chi}[t]$ (*utilization law*) and changing $\mathbf{y}[p,t]$ to

$$\mathbf{z}[p,t] = \lambda_t \mathbf{y}[p,t]$$

we get:

$$\sum_{t\in T} \boldsymbol{\chi}[t]\mathbf{C}[p_1,t]\mathbf{C}[p_2,t] + \sum_{t\in T} \mathbf{z}[p_1,t]\mathbf{C}[p_2,t] + \sum_{t\in T} \mathbf{z}[p_2,t]\mathbf{C}[p_1,t] = 0$$

In the particular case that $p_1 = p_2$, the above equation takes the form:

$$\sum_{t\in T} \boldsymbol{\chi}[t]\mathbf{C}[p,t]^2 + 2\sum_{t\in T} \mathbf{z}[p,t]\mathbf{C}[p,t] = 0$$

In summary, the following set of linear constraints can be added to the linear programming problem (17.1) of Section 17.1.1 for the computation of upper or lower bounds for linear functions of average marking of places and throughput of transitions:

(c_9) $\quad \sum_{t\in T} \boldsymbol{\chi}[t]\mathbf{C}[p,t]^2 + 2\sum_{t\in T} \mathbf{z}[p,t]\mathbf{C}[p,t] = 0, \qquad \forall p \in P$

(c_{10}) $\quad \sum_{t\in T} \boldsymbol{\chi}[t]\mathbf{C}[p_1,t]\mathbf{C}[p_2,t]$

$\qquad + \sum_{t\in T} \mathbf{z}[p_1,t]\mathbf{C}[p_2,t] + \sum_{t\in T} \mathbf{z}[p_2,t]\mathbf{C}[p_1,t] = 0, \ \ \forall p_1, p_2 \in P, p_1 \neq p_2$

(c_{11}) $\quad \mathbf{z} \geq \mathbf{0}$

The reader should notice that the new variables $\mathbf{z}[p,t], p \in P, t \in T$ have been added also to the linear programming problem (17.1).

17.4.2 Embedded PF-QN's

Insensitive lower bounds for the average interfiring time of transitions were introduced in Section 17.2.1 looking for the maximum of the average interfiring time of transitions of isolated subsystems generated by elementary P-semiflows. A more realistic computation of the average interfiring time of transitions of these subsystems than that obtained from the analysis in complete isolation is considered now using, once more, the concept of liveness bound of transitions. The number of servers at each transition t of a given system in steady state is limited by its corresponding liveness bound $\mathbf{lb}[t]$ (or by its structural enabling bound which can always be computed in an efficient manner), because this bound is the *maximum reentrance* (or maximum self-concurrency) that the net structure and the marking allow for the transition.

The technique we are going to briefly present (a more detailed discussion can be found in [14]) is based on a *decomposition* of the original model in subsystems. In particular, we look for *embedded product-form closed monoclass queueing networks*. Well-known efficient algorithms exist for the computation of exact values or bounds for the throughput of such models [28, 32, 18].

Therefore, let us concentrate in the search of such subsystems. How are they structurally characterized? From a topological point of view, they are *P-components*: strongly connected State Machines. Timing of transitions must be done with exponentially distributed services. Moreover, conditional routing is modelled with decisions among immediate transitions, corresponding to generalized free conflicts in the whole system. In other words, if t_1 and t_2 are in conflict in the considered P-component, they should be in generalized free conflict in the original net: $\mathbf{Pre}[\cdot, t_1] = \mathbf{Pre}[\cdot, t_2]$. The reason for this constraint is that since we are going to consider P-components as product-form closed monoclass queueing networks with limited number of servers at stations (transitions), the throughput of these systems is *sensitive to the conflict resolution policy*, even if the relative firing rates are preserved. Therefore, conflicts in the P-component must be solved with exactly the same marking independent discrete probability distributions as in the whole net system, in order to obtain an optimistic bound for the throughput of the original net system. We call *RP-components* the subnets verifying the previous constraints.

Definition 17.20 *Let $\mathcal{N}$ be a net and $\mathcal{N}_i$ a P-component of $\mathcal{N}$ (strongly connected State Machine subnet). $\mathcal{N}_i$ is a routing preserving P-component, RP-component, iff for any pair of transitions, t_j and t_k, in conflict in $\mathcal{N}_i$, they are in generalized free (equal) conflict in the whole net $\mathcal{N}$:* $\mathbf{Pre}[\cdot, t_j] = \mathbf{Pre}[\cdot, t_k]$.

An improvement of the insensitive lower bound for the average interfiring time of a transition t_j computed in Theorem 17.2 can be eventually obtained computing the exact average interfiring time of that transition in the RP-component generated by a minimal P-semiflow $\mathbf{y}$, with $\mathbf{lb}[t]$–server semantics for each involved transition t (in fact, it is not necessary that t_j belongs to the P-component; the bound for other transition can be computed and then weighted according to the visit ratios in order to compute a bound for t_j). The P-semiflow $\mathbf{y}$ can be selected among the optimal solutions of (17.9) or it can be just a feasible *near-optimal* solution.

As an example, let us consider the net system depicted in Figure 17.8. Assume that routing probabilities are equal to $1/3$ for t_1, t_2, and t_3, and that t_7, t_8, t_9, t_{10}, t_{11}, t_{12} have exponentially distributed service times with mean values $\overline{\mathbf{s}}[t_7] = \overline{\mathbf{s}}[t_8] = \overline{\mathbf{s}}[t_9] = 10$, $\overline{\mathbf{s}}[t_{10}] = \overline{\mathbf{s}}[t_{11}] = \overline{\mathbf{s}}[t_{12}] = 1$. The elementary P-semiflows of the net are:

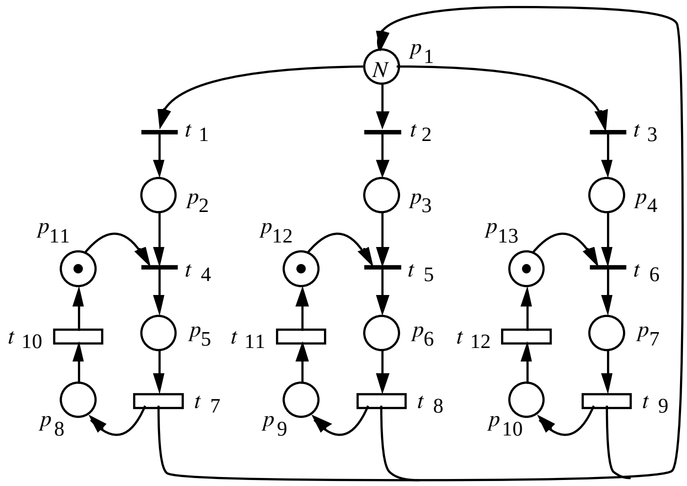

Figure 17.8: A live and bounded Free Choice system.

$$\begin{aligned} \mathbf{y_1} &= (1,1,1,1,1,1,1,0,0,0,0,0,0) \\ \mathbf{y_2} &= (0,0,0,0,1,0,0,1,0,0,1,0,0) \\ \mathbf{y_3} &= (0,0,0,0,0,1,0,0,1,0,0,1,0) \\ \mathbf{y_4} &= (0,0,0,0,0,0,1,0,0,1,0,0,1) \end{aligned} \tag{17.39}$$

Then, if the initial marking of p_{11}, p_{12}, and p_{13} is 1 token, and the initial marking of p_1 is N tokens, the lower bound for the average interfiring time derived from (17.9) is

$$\Gamma[t_1] \geq \max\{30/N, 11, 11, 11\} \tag{17.40}$$

For $N = 1$, the previous bound, obtained from $\mathbf{y_1}$, gives the value 30, while the exact average interfiring time is 31.06. For $N = 2$, the bound is 15 and it is derived also from $\mathbf{y_1}$ (average interfiring time of the P-component generated by $\mathbf{y_1}$, considered in isolation with infinite server semantics for transitions). This bound does not take into account the queueing time at places due to synchronizations (t_4, t_5, and t_6), and the exact average interfiring time of t_1 is $\Gamma[t_1] = 21.05$. For larger values of N, the bound obtained from (17.9) is equal to 11 (and is given by P-semiflows $\mathbf{y_2}$, $\mathbf{y_3}$ and $\mathbf{y_4}$). This bound can be improved if the P-component generated by $\mathbf{y_1}$ is considered with liveness bounds of transitions t_7, t_8, and t_9 reduced to 1 (which is the liveness bound of these transitions in the whole net).

The results obtained for different values of N are collected in Table 17.3. Exact values of average interfiring times for the P-component generated by $\mathbf{y_1}$ were computed using the *mean value analysis* algorithm [28]. This algorithm has $O(A^2B)$ worst case time complexity, where $A = \mathbf{y} \cdot \mathbf{m_0}$ is the number of

N	$\Gamma_{(17.9)}[t_1]$	$\Gamma_{(\mathbf{y_1})_{\mathrm{lb}}}[t_1]$	$\Gamma[t_1]$
1	30	30	31.06
2	15	20	21.05
3	11	16.67	17.71
4	11	15	16.03
5	11	14	15.03
10	11	12	13.02
15	11	11.34	12.35

Table 17.3: Bounds $\Gamma_{(17.9)}[t_1]$ obtained using (17.9), improvements for the bounds $\Gamma_{(\mathbf{y_1})_{\mathrm{lb}}}[t_1]$ presented in this section, and the exact average interfiring time $\Gamma[t_1]$ of t_1, for different initial markings N of p_1 in the net system of Figure 17.8.

tokens at the P-component and $B = \mathbf{y} \cdot \mathbf{Pre} \cdot \mathbf{1}$ is the number of involved transitions ($\mathbf{1}$ is a vector with all entries equal to 1). Exact computation on the original system takes several minutes in a *Sun SPARC Workstation* while bounds computation takes only a few seconds.

We also remark that other techniques for the computation of throughput upper bounds (instead of exact values) of closed product-form monoclass queueing networks could be used, such as, for instance, *balanced throughput upper bounds* [32] or *throughput upper bounds hierarchies* [18]. Hierarchies of bounds guarantee different levels of accuracy (including the exact solution), by investing the necessary computational effort. This provides also a hierarchy of bounds for the average interfiring time of transitions of Markovian Petri net systems.

Finally, the technique sketched in this section can be applied to the more general case of *Coxian* distributions (instead of exponential) for the service time of those transitions having either liveness bound equal to one (i.e., single-server stations) or liveness bound equal to the number of tokens in the RP-component (i.e., delay stations). The reason is that in these cases the embedded queueing network has also product-form solution, according to a classical theorem of queueing theory: the *BCMP theorem* [2].

17.4.3 Reduction and Transformation Techniques

The lower bounds for the throughput of transitions presented in previous sections are valid for any probability distribution function of service times but can be very pessimistic in some cases. In this section, an improvement of such results is briefly explained for the case of those net systems in which the following *performance monotonicity* property holds: *a local pessimistic transformation leads to a slower transformed net system* (i.e., a pessimistic local transformation guarantees a pessimistic global behaviour). Using the concept of *stochastic ordering* [29], a pessimistic transformation is, for example, to substitute the probability distribution function of a service (or token-subnet traversing) time

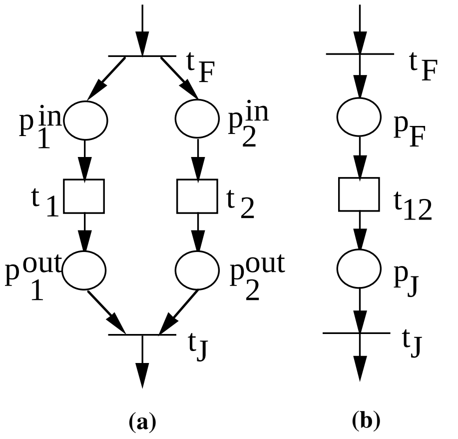

Figure 17.9: (a) Elementary fork-join and (b) its reduction.

by a *stochastically greater* probability distribution function. Live and bounded Free Choice is a class of systems for which the above performance monotonicity property holds. Details about the techniques presented here can be found in [11]. The basic ideas are:

1. To use local pessimistic transformation rules to obtain a net system "simpler" than the original (e.g., with smaller state space) and with equal or less performance.

2. To evaluate the performance for the derived net system, using insensitive bounds presented in previous sections, exact analysis, or any other applicable technique.

In order to obtain better bounds (after these two steps) than the values computed in previous sections, at least one of the transformation rules of item 1 must be less pessimistic than a total sequentialization of the involved transitions. We present first a rule whose application allows such *strict* improvement: the *fork-join rule*. Secondly, a rule that does not change at all the performance (*deletion of multistep preserving places*) is presented. Finally, a rule that does not follow the above ideas is also presented: the goal of this rule (*split of a transition*) is to make reapplicable the other transformation rules.

The most simple case of fork-join subnet that can be considered is depicted in Figure 17.9.a. In this case, if transitions t_1 and t_2 have exponential services X_1 and X_2 with means $\overline{\mathbf{s}}[t_1]$ and $\overline{\mathbf{s}}[t_2]$, respectively, they are reduced to a single transition (Figure 17.9.b) with exponential service time and mean:

$$\overline{\mathbf{s}}[t_{12}] = \mathrm{E}\left[\max(X_1, X_2)\right] = \overline{\mathbf{s}}[t_1] + \overline{\mathbf{s}}[t_2] - \left(\frac{1}{\overline{\mathbf{s}}[t_1]} + \frac{1}{\overline{\mathbf{s}}[t_2]}\right)^{-1}$$

Therefore, even if the *mean traversing time* of the reduced subnet by a single token has been preserved, it has been substituted by a stochastically greater

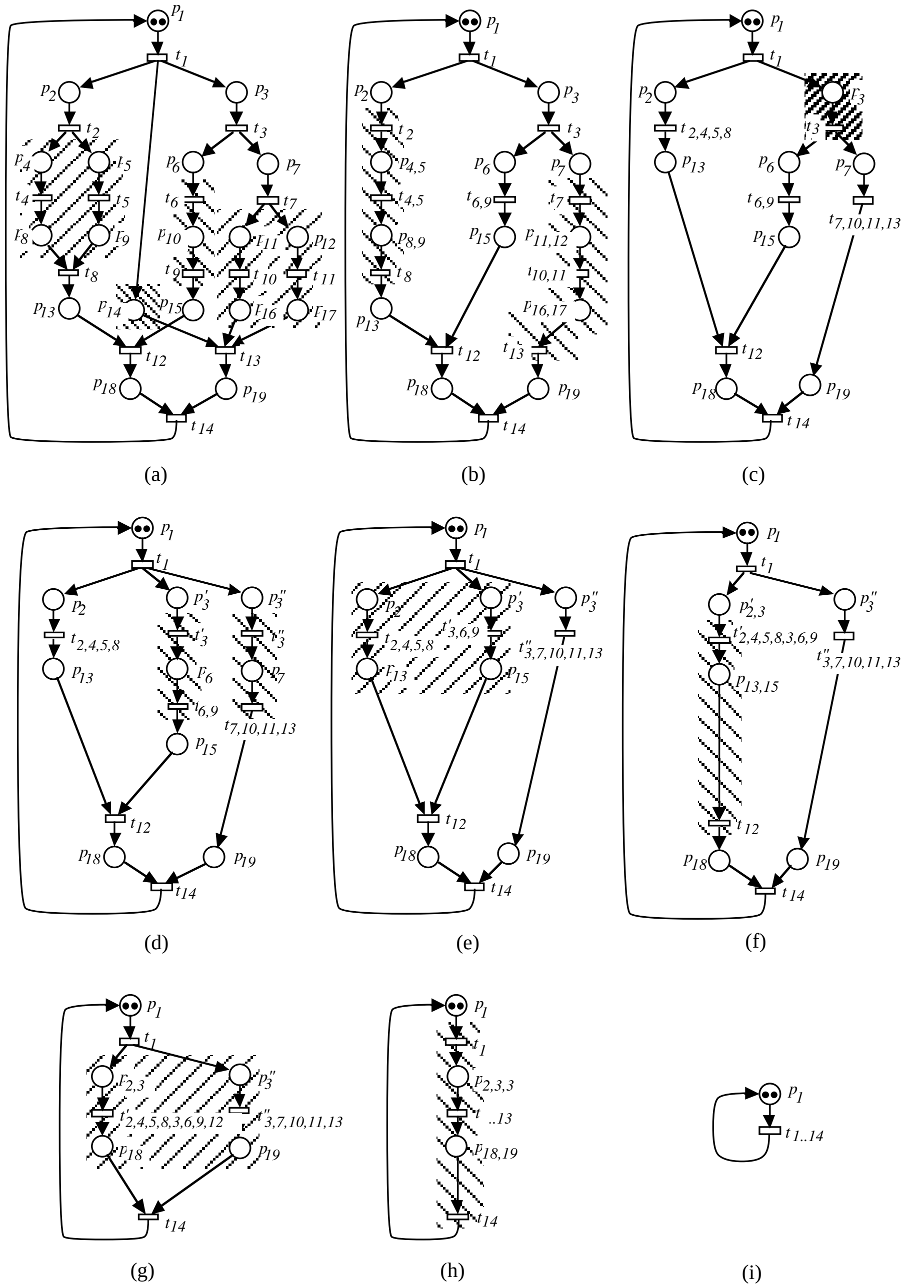

Figure 17.10: A complete reduction process. The relative error between insensitive bound and exact value diminishes from 140% to 35%.

variable. A trivial extension can be applied if the fork-join subnet includes more than two transitions in parallel.

Other transformation rules that have been presented in [11] are:

Deletion of a multistep preserving place: allows to remove some places without changing the *exact* performance indices of the stochastic net system. In fact the places that can be deleted are those whose elimination preserves the multisets of transitions simultaneously firable in all reachable markings (e.g., place p_{14} in Figure 17.10.a). The size of the state space of the model is preserved and also the exact throughput of transitions of the system.

Reduction of transitions in sequence: reduces a series of exponential services to a single exponential service with the same mean. Intuitively, this transformation makes indivisible the service time of two or more transitions representing elementary actions which always occur one after the other and lead to no side condition (e.g., transitions t_6 and t_9 in Figure 17.10.a). Therefore, the state space of the model is reduced. The throughput of transitions is, in general, reduced.

Split of a transition: this is not a state space reduction rule since it increases the state space of the transformed net system. The advantage of the rule is that it allows to proceed further in the reduction process using again the previous rules (e.g., transition t_3 in Figure 17.10.c).

An example of application of all above transformation rules is depicted in Figure 17.10 for a strongly connected marked graph with exponential timing. Let us assume that average service times of transitions are: $\overline{\mathbf{s}}[t_i] = 1$, $i = 1, 2, 3, 7, 8, 12, 13, 14$, and $\overline{\mathbf{s}}[t_i] = 10$ otherwise.

In order to compute firstly the insensitive lower bounds on throughput introduced in Section 17.3.2, it is necessary to derive the liveness bounds of transitions. In this case it is easy to see that $\mathbf{lb}[t_j] = 2$ for every transition t_j.

The vector of visit ratios of an MG is the unique minimal T-semiflow of the net: $\mathbf{v} = \mathbf{1}$. Therefore, the insensitive upper bound (valid for any probability distribution function of service times) of the average cycle time of the MG is $\Gamma \leq 34$. This value can be reached for some distributions of service times (see Theorem 17.18). Nevertheless, if services are exponential the exact average cycle time of the MG is $\Gamma = 14.15$.

The quantitative results of the transformation process illustrated in Figure 17.10 are shown in Table 17.4. We remark that the bound has been improved in polynomial time from 34 to 19.2.

17.5 Bibliographic Remarks

The general approach for the computation of insensitive performance bounds presented in Section 17.1 was introduced in [15].

The reinterpretation using Little's law and P-semiflows presented in Section 17.2.1 is in fact historically previous to the general approach, since it was

System	bound	error
Fig. 17.10.a	34	140 %
Fig. 17.10.b	29	105 %
Fig. 17.10.c	29	105 %
Fig. 17.10.d	29.5	108 %
Fig. 17.10.e	29.5	108 %
Fig. 17.10.f	24.8	75 %
Fig. 17.10.g	24.8	75 %
Fig. 17.10.h	19.2	35 %
Fig. 17.10.i	19.2	35 %

exact value: 14.15

Table 17.4: Successive improvements of the upper bound for the average cycle time of the MG in Figure 17.10 and relative errors with respect to the exact value $\Gamma = 14.15$.

firstly introduced in [5] for marked graphs and in [7] for Petri nets with *unique consistent firing count vector*. Improved versions of those papers are [6] and [8], respectively. The technique was extended to Free Choice systems in [9] and to *FRT-net systems* (cfr. Chapter 8) in [4] and [13].

The relation, presented in Section 17.2.2, between the general technique of Section 17.1 and the P-semiflows based technique of Section 17.2.1 is original from this chapter.

The results on the reachability of the throughput upper bound for marked graphs of Section 17.2.3 have been taken from [5, 6].

Concerning the improvements of the bounds, that based on implicit places (Section 17.3.1) was published in [10]; the use of liveness bounds of transitions presented in Section 17.3.2 was introduced for marked graphs in [5, 6] and later extended to Free Choice systems in [9]. The uniformization technique used to compute linear relations between second order moments in Section 17.4.1 was proposed in this framework in [24]. The improvement of the bounds based on the consideration of embedded product-form queueing networks presented in Section 17.4.2 was published in [12] and later improved in [14]. Finally, the reduction and transformation techniques briefly presented in Section 17.4.3 have been taken from [11].

Concerning other works that are not considered at all in this chapter, a large number of bounding techniques have been proposed for the performance measures of classical (*synchronization-free*) queueing networks. The first family is that of *asymptotic bound analysis* [22, 17]. Asymptotic bounds are obtained by considering two extreme situations: (1) no queueing takes place at any node, and (2) at least one station is saturated. These bounds do not require the product form property to hold and their computation is very fast, but they are not accurate in general. The rest of bounds that have been introduced are tighter but do require the product form assumption. This is the case of *balanced job bounds* [32, 23], which are based on the *mean value theorem* [28]. Finally, several schemes for the construction of *hierarchies of bounds* have been developed that guarantee any level of accuracy (including the exact solution), by investing the necessary computational effort: *performance bound hierarchies* [18, 19], *successively improving bounds* [30], *generalized quick bounds* [31]. All these techniques are derived from mean value theorem, thus they are valid only for product form networks.

With respect to timed Petri nets, M. Molloy [25] noted that the average token flows in an ordinary Markovian network at steady-state are conserved. Therefore, a series of *flow balance equations* can be written. Token flows are conserved in places so the sum of all flows into a place equals the sum of all flows out of the place. On the other hand, all token flows on the input and output arcs of a transition are equal. These equations determine the average token flows in the cycles of the net to within a constant. This constant cannot be determined without Markovian analysis at the reachability graph level. However, limit flows when the number of tokens tends to infinity can be computed. In order to do that, bottleneck transitions must be first located. Then, the actual flow through a bottleneck transition is (under saturation conditions) equal to its potential firing rate.

S. Bruell and S. Ghanta [3] developed algorithms for computing upper and lower bounds for the throughput of a restricted subclass of generalized stochastic Petri nets (with immediate and exponentially timed transitions). The considered nets include *control tokens* to model a physical restriction, such as semaphores, which is not a design parameter. The rest of tokens of such nets, grouped in *classes*, correspond to the notion of a job or customer in a monoclass queueing network, and its number is treated as a parameter of the net. The upper and lower bounds on throughput are computed hierarchically estimating maximum and minimum time of the path followed by each class of jobs.

In the paper of S. Islam and H. Ammar [20], methods to compute upper and lower bounds for the steady-state token probabilities of a subclass of generalized stochastic Petri nets were presented. The considered nets are obliged to admit a *time scale decomposition*. This means that the transitions of the net are supposed to be divided into two classes: slow and fast transitions, with several orders of magnitude of difference in the duration of activities. Moreover, the subnets obtained after removing all slow transitions with their input and output arcs must be conservative and admit a reversible initial marking. The computation is based on *near-completely decomposability* of Markov chains.

Bibliography

[1] F. Baccelli and P. Bremaud. *Elements of Queueing Theory.* Springer-Verlag, 1994.

[2] F. Baskett, K. M. Chandy, R. R. Muntz, and F. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, April 1975.

[3] S. C. Bruell and S. Ghanta. Throughput bounds for generalized stochastic Petri net models. In *Proceedings of the International Workshop on Timed Petri Nets*, pages 250–261, Torino, Italy, July 1985. IEEE Computer Society Press.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Campos, J. M. Colom, and M. Silva. Improving throughput upper bounds for net based models of manufacturing systems. In J. C. Gentina and S. G. Tzafestas, editors, *Robotics and Flexible Manufacturing Systems*, pages 281–294. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, The Netherlands, 1992.

[11] J. Campos, B. Sánchez, and M. Silva. Throughput lower bounds for Markovian Petri nets: Transformation techniques. In *Proceedings of the 4^{rd} International Workshop on Petri Nets and Performance Models*, pages 322–331, Melbourne, Australia, December 1991. IEEE-Computer Society Press.

[12] J. Campos and M. Silva. Throughput upper bounds for Markovian Petri nets: Embedded subnets and queueing networks. In *Proceedings of the 4^{rd} International Workshop on Petri Nets and Performance Models*, pages 312–321, Melbourne, Australia, December 1991. IEEE-Computer Society Press.

[13] 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.

[14] J. Campos and M. Silva. Embedded product-form queueing networks and the improvement of performance bounds for Petri net systems. *Performance Evaluation*, 18(1):3–19, July 1993.

[15] 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.

[16] J. M. Colom and M. Silva. Improving the linearly based characterization of P/T nets. In G. Rozenberg, editor, *Advances in Petri Nets 1990*, volume 483 of *Lecture Notes in Computer Science*, pages 113–145. Springer-Verlag, Berlin, 1991.

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

[18] D. L. Eager and K. C. Sevcik. Performance bound hierarchies for queueing networks. *ACM Transactions on Computer Systems*, 1(2):99–115, May 1983.

[19] D. L. Eager and K. C. Sevcik. Bound hierarchies for multiple-class queueing networks. *Journal of the ACM*, 33(1):179–206, January 1986.

[20] S. M. R. Islam and H. H. Ammar. On bounds for token probabilities in a class of generalized stochastic Petri nets. In *Proceedings of the* 3^{rd} *International Workshop on Petri Nets and Performance Models*, pages 221–227, Kyoto, Japan, December 1989. IEEE-Computer Society Press.

[21] J. Keilson. *Markov Chain Models. Rarity and Exponentiallity.* Springer-Verlag, 1979.

[22] L. Kleinrock. *Queueing Systems Volume II: Computer Applications.* John Wiley & Sons, New York, NY, 1976.

[23] J. Kriz. Throughput bounds for closed queueing networks. *Performance Evaluation*, 4:1–10, 1984.

[24] Z. Liu. Performance bounds for stochastic timed Petri nets. In G. De Michelis and M. Diaz, editors, *Application and Theory of Petri Nets 1995*, volume 935 of *Lecture Notes in Computer Science*, pages 316–334. Springer-Verlag, Berlin, 1995.

[25] M. K. Molloy. Fast bounds for stochastic Petri nets. In *Proceedings of the International Workshop on Timed Petri Nets*, pages 244–249, Torino, Italy, July 1985. IEEE-Computer Society Press.

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

[27] C. Ramchandani. *Analysis of Asynchronous Concurrent Systems by Petri Nets.* PhD thesis, MIT, Cambridge, MA, USA, February 1974.

[28] M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queueing networks. *Journal of the ACM*, 27(2):313–322, April 1980.

[29] S. M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, NY, 1983.

[30] M. M. Srinivasan. Successively improving bounds on performance measures for single class product form queueing networks. *IEEE Transactions on Computers*, 36:1107–1112, September 1987.

[31] R. Suri. Generalized quick bounds for performance of queueing networks. *Computer Performance*, 5(2):116–120, June 1984.

[32] J. Zahorjan, K. C. Sevcik, D. L. Eager, and B. Galler. Balanced job bound analysis of queueing networks. *Communications of the ACM*, 25(2):134–141, February 1982.