Throughput Upper Bounds for Markovian Petri Nets: Embedded Subnets and Queueing Networks

Javier Campos and Manuel Silva
Dpto. de Ingeniería Eléctrica e Informática *
Centro Politécnico Superior, Universidad de Zaragoza
María de Luna 3, 50015 Zaragoza, SPAIN

Abstract

This paper addresses the computation of upper bounds for the steady-state throughput of stochastic Petri nets with immediate and exponentially distributed service times of transitions. We try to deeply bridge stochastic Petri net theory to untimed Petri net and queueing network theories. Previous results for general service time distributions are improved for the case of Markovian nets by considering the slowest embedded subnet (generated by the support of left annullers of the incidence matrix of the net). The obtained results for the case of live and bounded free choice nets are of special interest. For such nets, the subnets generated by the left annullers of the incidence matrix can be seen as embedded product-form closed monoclass queueing networks, and efficient algorithms exist for their analysis.

1 Introduction

The computation of upper bounds for the *throughput of transitions*, defined as the average number of firings per unit time, of stochastic Petri nets with exponentially distributed service times of transitions is considered in this paper. Previous results for general service time distributions [2, 3, 4, 5, 6, 7], based only on the net structure, on the routing probabilities, and on the mean values of the service time of transitions, are newly interpreted and improved. Throughput lower bounds for Markovian nets are studied in a companion paper [8].

The improvement of the throughput upper bound, for the particular case of exponential distributions, is based on the consideration of waiting time in queues (places) due to limited number of servers at transitions (liveness bound of transitions) of the slowest subnet generated by P-semiflows. The computation of such bounds is, in general, much cheaper than the exact analysis of the whole net. If the subnets generated by P-semiflows have state machine topology, classical algorithms from closed queueing networks analysis, such as *balanced throughput upper bounds* [19], *throughput upper bounds hierarchies* [11], or exact *mean value analysis* [18], can be used for an efficient computation. The obtained results for live and bounded free choice nets [14] are of special interest. For these nets, all minimal P-semiflows generate state machines. Therefore, free choice nets can be seen as a collection of closed queueing networks synchronized through some transitions avoiding a direct interaction of choices and concurrency. In this case, the computation of bounds is based on the application of well-known algorithms from queueing networks literature to the *slowest closed queueing network embedded in the stochastic Petri net.*

We assume that the reader is familiar with the structure, firing rules, and basic properties of net models (see [16] for a recent survey). Let us recall some notation here: $\mathcal{N} = \langle P, T, Pre, Post\rangle$ is a net with $n = |P|$ places and $m = |T|$ transitions. If the Pre and $Post$ incidence functions take values in $\{0,1\}$, $\mathcal{N}$ is said ordinary. PRE, $POST$, and $C = POST - PRE$ are $n \times m$ matrices representing the Pre, $Post$, and global incidence functions. Vectors $Y \geq 0$, $Y^T \cdot C = 0$ ($X \geq 0$, $C \cdot X = 0$) represent P-semiflows, also called conservative components (T-semiflows, also called consistent components). M (M_0) is a marking (initial marking). Finally, σ represents a fireable sequence, while $\vec{\sigma}$ is the firing count vector associated to σ. If M is reachable from M_0 (i.e., $\exists\sigma$ such that $M_0[\sigma\rangle M$), then $M = M_0 + C \cdot \vec{\sigma} \geq 0$ and $\vec{\sigma} \geq 0$.

The introduction of timing specification is essential in order to use Petri net models for performance evaluation of distributed systems. We consider nets with exponentially distributed timed transitions with *one phase* firing rule, i.e., a timed enabling (called the *service time* of the transition) followed by an atomic firing. The service times of transitions are supposed

*This work was partially supported by the DEMON Esprit BRA 3148 and the PRONTIC 354/91.

to be mutually independent and time independent.

We assume that a transition t enabled K times in a marking M (i.e., $K = \max\{k | M \geq kPRE[t]\}$) works at conditional speed K times that it would work in the case it was enabled only once (*infinite-server semantics*). 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 1) for that transition and marking it with k tokens. Therefore, the infinite-server semantics appears to be the most general one, and for this reason it is adopted in this paper.

In order to avoid the coupling between resolution of conflicts and duration of activities, we suppose that transitions in conflict are *immediate* (they fire in zero time). Decisions at these conflicts are taken according to *routing rates* $\mathcal{R}$ associated with immediate transitions (*generalized stochastic Petri nets* [1]). In other words, each subset of transitions $\{t_1, \ldots, t_k\} \subset T$ that are in conflict in one or several reachable markings are considered immediate, and the constants $r_1, \ldots, r_k \in \mathbb{R}^+$ are explicitly defined in the net interpretation in such a way that when $t_1, \ldots, t_k$ are simultaneously enabled, transition t_i $(i = 1, \ldots, k)$ fires with probability $r_i / (\sum_{j=1}^{k} r_j)$. Note that the routing rates are assumed to be strictly positive, i.e., all possible outcomes of any conflict have a non-null probability of firing. This fact guarantees a *fair* behaviour for the non-autonomous Petri nets that we consider.

The paper is organized as follows. In section 2 we recall the throughput upper bounds for stochastic Petri nets with general service time distributions (including both deterministic and stochastic timing) derived in previous works. An interpretation in terms of subnets is given. In section 3, an improvement of the previous bounds is achieved for the case of exponential distributions for the service time of transitions, computing the throughput of the slowest subnet generated by P-semiflows with limited-server semantics for transitions. The case in which P-semiflows generate state machines is considered in section 4. Section 4.1 is devoted to the free choice nets case. In the case of other net subclasses, the addition of *implicit places* [10] can have an additional advantage, since new closed queueing networks embedded in the net are created (section 4.2). Conclusions are summarized in section 5.

2 Bounds for general service time distributions and their interpretation

In this section we consider the steady-state behaviour of stochastic Petri nets with general service time distribution or deterministic as a particular case, under *weak ergodicity* assumption for the firing and the marking processes [3]. Weak ergodicity of a process is considered instead of the usual *strong ergodicity* concept [13] because this one creates problems when we want to include the deterministic case as a special case of a stochastic model. In particular, the marking and the firing processes of a stochastic net are said to be weakly ergodic if the following limits exist (where τ represents the time):

$$\overline{M} \stackrel{\text{def}}{=} \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} M_s ds < \vec{\infty}; \ \vec{\sigma}^* \stackrel{\text{def}}{=} \lim_{\tau \to \infty} \frac{\vec{\sigma}_\tau}{\tau} < \vec{\infty}$$

Where $\overline{M}$ and $\vec{\sigma}^*$ are constants and *almost everywhere convergence* is assumed (in other words, all sample paths give the same estimation of average values). $\overline{M}$ and $\vec{\sigma}^*$ are called the limit average marking and the limit throughput vector, respectively.

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 [15]: the average number of customers, the output rate (throughput), and the average time spent by a customer within the region. In previous works [2, 3, 4, 5, 6, 7], we considered stochastic live and bounded Petri nets as synchronized queueing networks and applied Little's result to each place of the net. Let us denote $R(p_i)$ the average time spent by a token within the place p_i (response time at place p_i). Then the above mentioned relationship can be stated as follows:

$$\overline{M}(p_i) = (PRE[p_i] \cdot \vec{\sigma}^*) R(p_i) \qquad (1)$$

where $PRE[p_i]$ is the i^{th} row of the pre-incidence matrix of the underlying Petri net, thus $PRE[p_i] \cdot \vec{\sigma}^*$ is the output rate of place p_i.

The conditions under which equation (1) holds are very weak, thus this equation is widely applicable. 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, $R(p_i)$ and $\overline{M}(p_i)$ are unknown, while information about $\vec{\sigma}^*$ can easily be computed for interesting net subclasses. Let us define the relative firing frequency vector or *vector of visit ratios to transitions* as $\vec{v}^{(i)} \stackrel{\text{def}}{=} \Gamma^{(i)} \vec{\sigma}^*$, where $\Gamma^{(i)} = 1/\vec{\sigma}_i^*$ represents the *mean interfiring time* of transition t_i (i.e., the inverse of its throughput). Here we consider those stochastic nets whose vector of visit ratios to transitions can be computed in an efficient way from the net structure $\mathcal{N}$ and from the relative frequency of conflict resolutions $\mathcal{R}$ (i.e., the routing probabilities associated with decisions). This is the case, for instance, for strongly connected *marked graphs* [3], live and bounded *mono-T-semiflow nets* [5], live and bounded *free choice nets* [4], and *FRT-nets* [2]. Unfortunately, for other subclasses like *simple nets*, the relative firing frequency vector also depends on the initial marking M_0 and on the service times of transitions [4].

The *response* times at places $R(p_i)$ are unknown. In fact, they can be expressed as sums of the *waiting* times due to the synchronization schemes and the *service* times associated with transitions, and only service times are known: s_i, $i = 1, \ldots, m$. Thus a lower bound for the response time at a given place can be obtained from the knowledge of the average service time of its output transition (remember that if a place has several output transitions, they are assumed to be immediate), and the following system of inequalities can be derived from (1):

$$\Gamma^{(i)}\overline{M} \geq PRE \cdot \vec{D}^{(i)} \tag{2}$$

where $\vec{D}^{(i)}$ is the vector of *average service demands* of transitions, with components:

$$D^{(i)}(t_j) \stackrel{\text{def}}{=} s_j v^{(i)}(t_j) \tag{3}$$

P-semiflows Y are non-negative left annullers of the incidence matrix C (i.e., $Y^T \cdot C = 0$, thus $Y^T \cdot M = Y^T \cdot M_0$ for all reachable marking M and $Y^T \cdot \overline{M} = Y^T \cdot M_0$). Now, premultiplying by Y the relation (2), the following lower bound for the mean interfiring time of a given transition t_i can be derived:

$$\Gamma^{(i)} \geq \max_{Y \in \{P-semiflow\}} \frac{Y^T \cdot PRE \cdot \vec{D}^{(i)}}{Y^T \cdot M_0}$$

The previous lower bound can be formulated in terms of a fractional programming problem and later, after some considerations, transformed into a linear programming problem [17]:

Theorem 2.1 [5] *For any live and bounded net, a lower bound for the mean interfiring time $\Gamma^{(i)}$ of transition t_i can be computed by the following linear programming problem:*

$$\begin{array}{lll} \Gamma^{(i)} \geq & \mathit{maximum} & Y^T \cdot PRE \cdot \vec{D}^{(i)} \\ & \mathit{subject\ to} & Y^T \cdot C = 0 \\ & & Y^T \cdot M_0 = 1; Y \geq 0 \end{array} \tag{LPP1}$$

If the solution of the above problem is unbounded and since it is a lower bound for the mean interfiring time of transition t_i, the non-liveness can be assured (infinite interfiring time). If the visit ratios of all transitions are non-null, the unboundedness of the above problem implies that a total deadlock is reached by the net. This result has the following interpretation: if the problem (LPP1) is unbounded then there exists an unmarked P-semiflow, and the net is non-live.

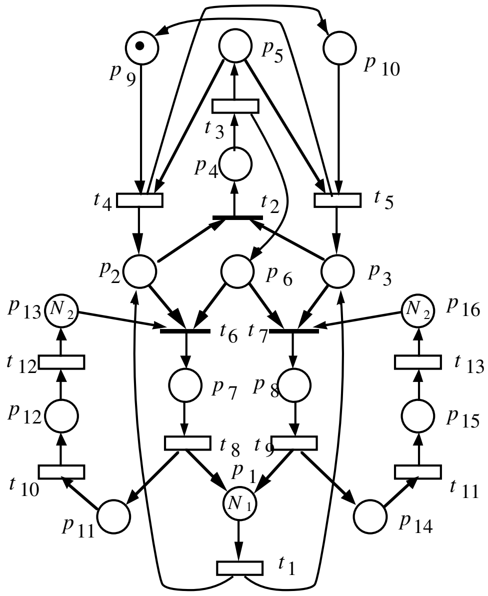

Figure 1: A live and bounded stochastic Petri net.

The basic advantage of theorem 2.1 lies 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 [17].

For strongly connected marked graphs, the bound derived from theorem 2.1 has been shown to be reachable for arbitrary mean values and coefficients of variation associated with transition service times [3]. This is not the case, in general, for live and bounded free choice nets. However, a reachable bound for live and safe free choice nets has also been obtained for arbitrary mean service times, by considering some *multisets of circuits* of the nets [6].

In order to interpret the theorem 2.1, let us consider the *mono-T-semiflow net* [5] depicted in figure 1. The unique minimal T-semiflow of the net is $X = (2,2,2,1,1,1,1,1,1,1,1,1,1)^T$. Therefore, according to (3), the vector of average service demands for transitions normalized, for instance, for t_4 is $\vec{D}^{(4)} = (2s_1, 0, 2s_3, s_4, s_5, 0, 0, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13})^T$, because the vector of visit ratios is $\vec{v}^{(4)} = X$ (see [5]) and transitions t_2, t_6, and t_7 are assumed to be immediate ($s_2 = s_6 = s_7 = 0$).

The minimal P-semiflows (minimal support solutions of $Y^T \cdot C = 0, Y \geq 0$) of this net are:

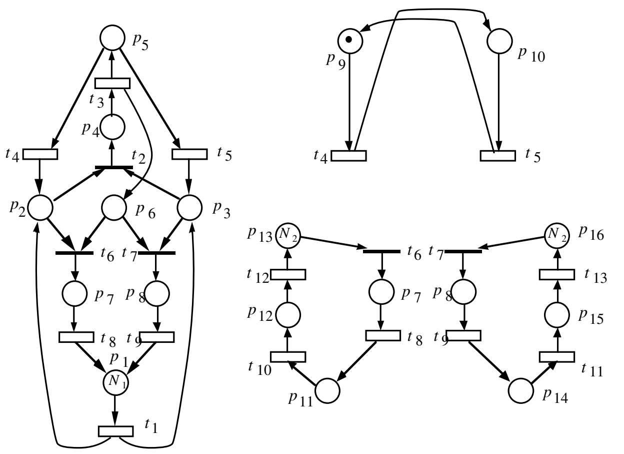

Figure 2: Subnets of the net in figure 1 generated by minimal P-semiflows.

$$\begin{aligned} Y_1 &= (2,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)^T \\ Y_2 &= (0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0)^T \\ Y_3 &= (0,0,0,0,0,0,1,0,0,0,1,1,1,0,0,0)^T \\ Y_4 &= (0,0,0,0,0,0,0,1,0,0,0,0,0,1,1,1)^T \end{aligned}$$

Then, the application of (LPP1) gives:

$$\begin{aligned} \Gamma^{(4)} \quad \geq \quad \max\{ \; & (4s_1 + 2s_3 + s_4 + s_5 + s_8 + s_9)/2N_1, \\ & s_4 + s_5, \\ & (s_8 + s_{10} + s_{12})/N_2, \\ & (s_9 + s_{11} + s_{13})/N_2 \; \} \end{aligned} \tag{4}$$

Where $N_1 > 0$ is the initial marking of place p_1, and $N_2 > 0$ is the initial marking of p_{13} and p_{16}. Now, let us consider the P-semiflow decomposed view of the net: the four subnets generated by Y_1, Y_2, Y_3, and Y_4 are depicted in isolation in figure 2. The exact mean interfiring times of the second, third, and forth subnets are s_4+s_5, $(s_8+s_{10}+s_{12})/N_2$, and $(s_9+s_{11}+s_{13})/N_2$, respectively (remember that infinite-server semantics is assumed). The exact mean intefiring time of t_4 in the first subnet (generated by Y_1) cannot be computed in a compact way (like the others), because it includes synchronizations (it has not queueing network topology). In any case, its mean interfiring time is greater than $(4s_1 + 2s_3 + s_4 + s_5 + s_8 + s_9)/2N_1$, because this would be the cycle time of a queueing network (without delays due to synchronizations) of infinite-server stations with the same average service demands and number of customers. Therefore, the lower bound for the mean interfiring time of t_4 in the original net given by (4) is computed looking at the slowest subnet generated by the P-semiflows, considered in isolation.

In the next section, we improve the previous bound by taking into account that the maximum number of servers that can be available at transitions of the subnets can be limited by the number of tokens in the other subnets.

3 Improvement for Markovian stochastic Petri nets

As stated erlier, the mean interfiring time of transitions of isolated subnets generated by P-semiflows is computed in (LPP1) assuming infinite-server semantics for the involved transitions. A more realistic computation of the mean interfiring time of transitions of these subnets than that obtained from the analysis in complete isolation is considered now. The performance of a net with infinite-server semantics of transitions depends on the maximum degree of enabling of the transitions, the *enabling bound*.

Definition 3.1 [5] *Let $\langle \mathcal{N}, M_0 \rangle$ be a marked Petri net. The enabling bound of a given transition t of $\mathcal{N}$ is* $E(t) \stackrel{\text{def}}{=} \max\{k \mid \exists M \in R(\mathcal{N}, M_0) : M \geq kPRE[t]\}$.

In particular, the steady-state performance does depend on the maximum degree of enabling of transitions in steady-state, which can be different from the maximum degree of enabling of a transition during all its evolution from the initial marking. Therefore, we also recall the concept of *liveness bound*, which allows to generalize the classical concept of liveness of a transition:

Definition 3.2 [5] *Let $\langle \mathcal{N}, M_0 \rangle$ be a marked Petri net. The liveness bound of a given transition t of $\mathcal{N}$ is* $L(t) \stackrel{\text{def}}{=} \max\{k \mid \forall M' \in R(\mathcal{N}, M_0), \exists M \in R(\mathcal{N}, M') : M \geq kPRE[t]\}$.

The definitions above refer to behavioural properties. Since we are looking for computational techniques at the structural level, we also recall the structural counterpart of one of the concepts.

Definition 3.3 [5] *Let $\langle \mathcal{N}, M_0 \rangle$ be a marked Petri net. The structural enabling bound of a given transition t of $\mathcal{N}$ is:*

$$\begin{aligned} SE(t) \stackrel{\text{def}}{=} \quad & \textit{maximum} \quad k \\ & \textit{subject to} \quad M_0 + C \cdot \vec{\sigma} \geq kPRE[t] \\ & \qquad\qquad\quad \vec{\sigma} \geq 0 \end{aligned} \tag{LPP2}$$

Note that the definition of structural enabling bound reduces to the formulation of a linear programming problem. The following result related to the above concepts can be obtained:

Theorem 3.1 [5] *Let $\langle \mathcal{N}, M_0 \rangle$ be a marked Petri net, then for all transition t of $\mathcal{N}$, $SE(t) \geq E(t) \geq L(t)$.*

The interest of the above theorem lies in the fact that for those nets whose exact liveness bounds of transitions cannot be computed efficiently, upper bounds can be always obtained by solving the linear programming problems (LPP2), i.e., by computing the structural enabling bounds.

Going back to the semantics of transitions, the number of servers at each transition t of a given net in steady-state can be supposed to be limited to its corresponding liveness bound $L(t)$ (or to 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.

For the case of exponential distributions of the service time of transitions, the knowledge of the liveness bound of transitions for a given net allows to improve the throughput upper bound computed in theorem 2.1 by investing an additional computational effort.

Theorem 3.2 *Let $\langle \mathcal{N}, M_0 \rangle$ be a live and bounded stochastic Petri net with constant routing probabilities defining the conflict resolution policy and exponential distributions for service time of non-immediate transitions. For each transition t, let $L(t)$ be its liveness bound. Let Y be a feasible solution of the problem (LPP1) and $\Gamma^{(i)}_{Y_\infty}$ the corresponding value of the objective function in (LPP1). Let $\Gamma^{(i)}_{Y_L}$ be the exact mean interfiring time of t_i computed for the isolated subnet generated by Y, with $L(t)$–server semantics for each involved transition t. Then:*

$$\Gamma^{(i)} \geq \Gamma^{(i)}_{Y_L} \geq \Gamma^{(i)}_{Y_\infty}$$

where $\Gamma^{(i)}$ is the exact mean interfiring time of t_i in the whole net.

We restrict ourselves to stochastic nets with null or exponentially distributed service times because in this case, there exists a well-known technique for computing the values of $\Gamma^{(i)}_{Y_L}$, by solving the embedded continuous time Markov chain [1].

We give now an interpretation of this improvement from a queueing theory point of view:

Interpretation. *Both the bound presented in section 2 and the presented in this section are based on the computation of the mean interfiring time of transitions of subnets generated by P-semiflows considered in isolation. In the first case, since infinite-server semantics is considered for the isolated subnet, the real (unknown) response time at places is lowerly bounded by the service time of transitions, but waiting time due to synchronizations is not considered at all. Now, the bound for the response time at places is improved taking into account not only the service time but also a part of the queueing time due to synchronizations: the time in queue when $L(t)$ servers is the maximum available at each transition t.*

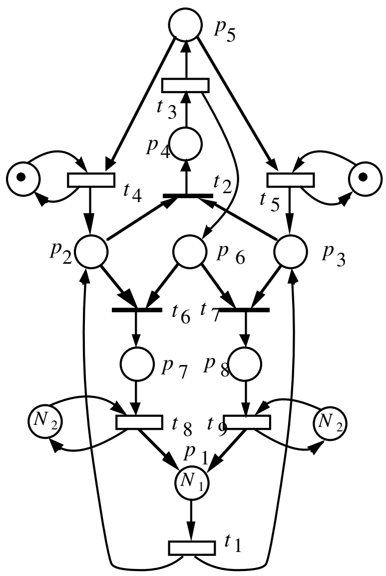

Figure 3: Subnet of the net in figure 1 generated by Y_1, with $L(t)$–server semantics for each involved transition t.

As an example, let us consider again the net depicted in figure 1. The liveness bounds of all transitions can be computed by solving the corresponding problems (LPP2) (in this case $L(t) = SE(t), t = 1, \ldots, 13$). They are $L(t_1) = L(t_2) = L(t_3) = N_1$, $L(t_4) = L(t_5) = 1$, $L(t_6) = L(t_7) = L(t_8) = L(t_9) = L(t_{10}) = L(t_{11}) = L(t_{12}) = L(t_{13}) = N_2$. Assume, for instance, that $N_1 \geq N_2$. Then the subnet generated by P-semiflow Y_1, considered with $L(t)$–server semantics for each of the transitions t, is the one depicted in figure 3. Now, theorem 3.2 can be applied for the improvement of the throughput upper bound of transitions derived from (LPP1). In general, the subnet in figure 3 has far less reachable markings than the original net. Therefore, the computation of the throughput bound takes considerably less computational effort than that needed for computing the exact throughput. Assume that exponentially distributed service times are associated to timed transitions, with means $s_1 = s_3 = s_4 = s_5 = s_8 = s_9 = s_{10} = s_{11} = s_{12} = s_{13} = 1$,

N_1/N_2	$\lvert\mathcal{T}_{\mathcal{N}}\rvert$	$\Gamma^{(4)}$	$\lvert\mathcal{T}_{\mathcal{N}_1}\rvert$	$\Gamma^{(4)}_{Y_L}$	$\Gamma^{(4)}_{Y_\infty}$
2/1	628	4.95	58	4.35	2.5
3/2	9144	2.98	184	2.94	2.0
4/2	23110	2.44	536	2.35	2.0
4/3	66630	2.35	440	2.35	2.0
5/4	> 250000	??	898	2.09	2.0

Table 1: Comparison between exact mean interfiring time of t_4, and the upper bounds obtained from theorems 2.1 and 3.2, for different values of N_1 and N_2, for the net in figure 1.

while transitions t_2, t_6, t_7 are immediate. The comparison between exact mean interfiring time of t_4 (for the case of equal routing rates of t_2, t_6 and t_7 when they are in conflict), and the upper bounds obtained from theorems 2.1 and 3.2 are collected in table 1, for different values of the initial marking. The table also compares the number of *tangible* markings [1] of the original net and of the considered subnet. The reader can notice that the obtained improvements are significant in all cases. Notice also that for the case of $N_1 = 5$, $N_2 = 4$, the number of tangible markings of the original net is greater than 250000, while the number of tangible markings of the subnet keeps "quite manageable" (898 markings). The computation of exact values takes several minutes in a *SPARC Workstation* while the bounds can be evaluated in a few seconds. All the computations have been obtained using the tool *GreatSPN* [9].

4 P-semiflows and embedded queueing networks

In this section, we study the particular case in which the subnets generated by P-semiflows of a given net have state machine topology. Such subnets are called *P-components*. The interest of that case lies in the fact that *Markovian Petri nets with state machine topology are product-form monoclass queueing networks.* Therefore, well-known efficient algorithms exist for the computation of exact values of the throughput (or bounds for it). First we present the particular case of live and bounded free choice nets. After that, the possible extensions to other net subclasses are considered.

4.1 Free choice nets case

Let $\mathcal{N} = \langle P, T, Pre, Post\rangle$ be a Petri net and $P' \subseteq P$. $\mathcal{N}' = \langle P', T', Pre', Post'\rangle$ is called a *P-component* of $\mathcal{N}$ iff $\mathcal{N}'$ is the subnet of $\mathcal{N}$ generated by P' (i.e., $T' \subseteq T$ and $Pre', Post'$ are the restrictions of $Pre, Post$ to P' and T') and $\forall t \in T'$: $|{}^\bullet t \cap P'| \leq 1 \wedge |t^\bullet \cap P'| \leq 1$ (i.e., $\mathcal{N}'$ has state machine topology).

An important result in the structure theory of nets assures that each minimal P-semiflow of a structurally live and structurally bounded free choice net generates a P-component:

Theorem 4.1 [12] *Let $\mathcal{N} = \langle P, T, Pre, Post\rangle$ be a structurally live and structurally bounded free choice net and $Y \geq 0$. Y is a minimal P-semiflow of $\mathcal{N}$ iff the two following conditions hold:*
a) $\forall p \in P$: $Y(p) \in \{0, 1\}$.
b) There exists a P-component of $\mathcal{N}$, $\mathcal{N}' = \langle P', T', Pre', Post'\rangle$, such that $||Y|| = P'$, where $||Y|| = \{p \in P | Y(p) > 0\}$ is the support of Y.

We remark that a P-component (state machine) has the topology of a monoclass queueing network. At this point, let us interpret again, from a queueing theory point of view, the linear programming problem (LPP1) for the case of live and bounded free choice nets:

Interpretation. *Let Y be a minimal P-semiflow of a live and bounded free choice net and*

$$\Gamma_Y^{(i)} = \frac{Y^T \cdot PRE \cdot \vec{D}^{(i)}}{Y^T \cdot M_0}$$

its corresponding value of the objective function in the problem (LPP1). Then $\Gamma_Y^{(i)}$ is the exact mean interfiring time for station (transition) t_i if the closed monoclass queueing network generated by Y (P-component) is considered in isolation, with delay stations (infinite-server semantics for all transitions) and general service time distributions. Moreover, each optimum solution Y^* of (LPP1) (note that more than one can exist) corresponds to a slowest closed queueing network (P-component) embedded in the net.

Let us remark that the throughput of transitions (with infinite-server semantics) in the isolated P-component is *insensitive* (i.e., independent) to the distribution of service time of transitions. This is not the case for multiple-server (but finite) semantics. For those servers, the response time at places depends not only on the service time but also on the waiting time in queue, and the actual throughput does depend on the number of servers at each station and on the form of the service distributions.

Now, we recall a result which provides an efficient method for the computation of liveness bounds of transitions, for the case of live and bounded free choice nets.

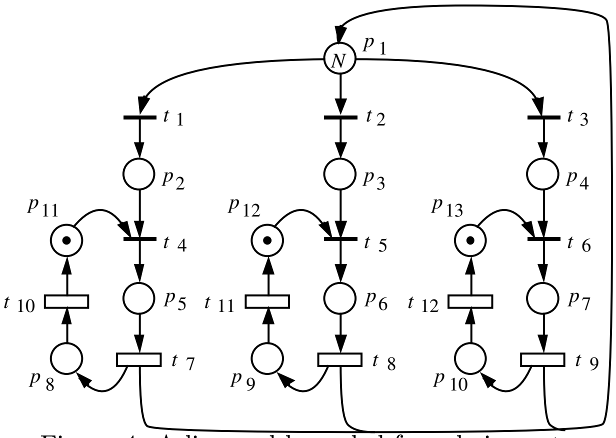

Figure 4: A live and bounded free choice net.

Theorem 4.2 [4] *Let* $\langle \mathcal{N}, M_0 \rangle$ *be a live and bounded free choice net. Then, for all transition* t *of* $\mathcal{N}$, $SE(t) = E(t) = L(t)$.

In other words, the liveness bounds of transitions (actual number of servers needed at transitions in steady-state) of live and bounded free choice nets can be obtained by solving the problem (LPP2).

Since P-components generated by P-semiflows of a live and bounded free choice net have monoclass queueing network topology and we are considering exponential servers, such P-components can be seen as product-form queueing networks (if *FCFS service discipline* is assumed) and well-known efficient algorithms for their analysis can be used [18] instead of the enumerative technique based on the derivation of the embedded continuous time Markov chains.

As an example of application of theorem 3.2 for the particular case of live and bounded free choice nets, let us consider the one depicted in figure 4. 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 $s_7 = s_8 = s_9 = 10$, $s_{10} = s_{11} = s_{12} = 1$. The P-semiflows of the net are:

$$\begin{aligned} Y_1 &= (1,1,1,1,1,1,1,0,0,0,0,0,0)^T \\ Y_2 &= (0,0,0,0,1,0,0,1,0,0,1,0,0)^T \\ Y_3 &= (0,0,0,0,0,1,0,0,1,0,0,1,0)^T \\ Y_4 &= (0,0,0,0,0,0,1,0,0,1,0,0,1)^T \end{aligned}$$

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 mean interfiring time derived from (LPP1) is $\Gamma^{(1)}_{(\mathrm{LPP1})} = \max\{30/N, 11, 11, 11\}$. For $N = 1$, the previous bound, obtained from Y_1, gives the value 30, while the exact mean interfiring time is 31.06. For $N = 2$, the bound is 15 and it is derived also

N	$\Gamma^{(1)}$	$\Gamma^{(1)}_{(\mathrm{LPP1})}$	$\Gamma^{(1)}_{(Y_1)_L}$
1	31.06	30	30
2	21.05	15	20
3	17.71	11	16.67
4	16.03	11	15
5	15.03	11	14
10	13.02	11	12
15	12.35	11	11.34

Table 2: Exact mean interfiring time of t_1, bounds obtained using (LPP1), and the improvements derived from theorem 3.2, for different initial markings of p_1 in the net of figure 4.

from Y_1 (mean interfiring time of the P-component generated by 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 mean interfiring time of t_1 is $\Gamma^{(1)} = 21.05$. For larger values of N, the bound obtained from (LPP1) is equal to 11 (and is given by P-semiflows Y_2, Y_3 and Y_4). This bound can be improved if the P-component generated by Y_1 is considered with liveness bounds of transitions t_4, t_5, t_6, 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 2. Exact values of mean interfiring times for the P-component generated by Y_1 were computed using the *mean value analysis* algorithm [18]. This algorithm has $O(A^2 B)$ worst case time complexity, where $A = (Y^*)^T \cdot M_0$ is the number of tokens at the P-component and $B = Y^T \cdot PRE \cdot \mathbb{1}$ is the number of involved transitions ($\mathbb{1}$ is a vector with all entries equal to 1). As for the results in table 1, exact computation takes several minutes of the CPU of a *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* [19] or *throughput upper bounds hierarchies* [11]. Hierarchies of bounds guarantee any level of accuracy (including the exact solution), by investing the necessary computational effort. This provides also a hierarchy of bounds for the mean interfiring time of transitions of live and bounded free choice nets.

Since the statement of the theorem 3.2 holds for every feasible solution Y of (LPP1), it holds for each optimum solution Y^* and the bound computed in theorem 2.1 can be eventually improved for exponential distributions as follows:

Corollary 4.1 *An improvement of the throughput upper bound computed in theorem 2.1 can be obtained computing the value* $\Gamma^{(i)}_{Y^*_L}$ *of theorem 3.2 for an optimum solution* Y^* *of the problem (LPP1). Moreover, the improvement is strict if and only if the P-component generated by* Y^* *contains more than* $\min\{L(t) \mid t \in P\text{-}component\}$ *tokens.*

Taking into account that the number of optimum solutions of (LPP1) (giving the same value of the objective function) can be theoretically exponential on the net size, the next question that must be answered is: Which optimum solution(s) of problem (LPP1) should be considered in order to obtain a greater improvement with the application of corollary 4.1?

We now present an algorithm for the computation of an improvement of bounds given by problem (LPP1), based on a possible heuristic for the selection of some optimum solutions of this problem. The heuristic gives the possibility of selecting an arbitrary number K of optimum solutions of (LPP1). The way of selecting only optimum solutions among all feasible solutions of that linear programming problem consists of considering the following constraints:

$$\begin{array}{l} Y^T \cdot PRE \cdot \vec{D}^{(i)} = \Gamma^{(i)}_{PS} \\ Y^T \cdot C = 0;\ Y^T \cdot M_0 = 1;\ Y \geq 0 \end{array} \tag{5}$$

where $\Gamma^{(i)}_{PS}$ is the optimum value of (LPP1), and must be computed before.

Now, it is easy to understand that, among the above optimum solutions, those with less liveness bounds for involved transitions should probably give slower embedded queueing networks. This is because, if only a few servers exist at a given transition, the waiting time of tokens in the input places will be larger. A "natural" way to select those P-components with expected less number of servers at involved transitions is to solve a linear programming problem with expresions (5) as constraints, and with the same objective function than in problem (LPP1) but modifying the vector $\vec{D}^{(i)}$, dividing the mean service time s_j of each transition t_j by its corresponding liveness bound $L(t_j)$. Intermediate situations can be considered dividing each s_j by a quantity ranging from $1+\delta$ (with $\delta > 0$) to $L(t_j)$.

An algorithm for the previously argumented heuristic can be as follows:

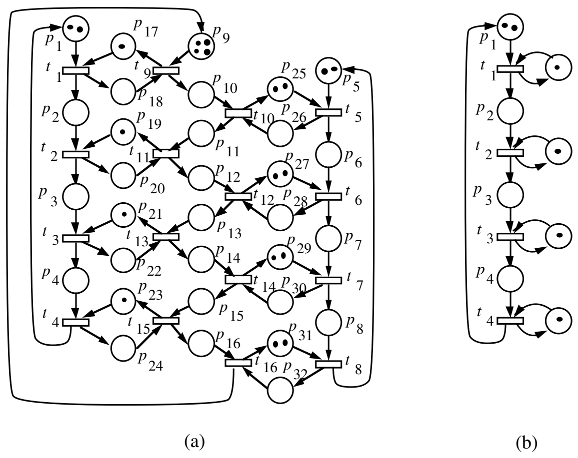

Figure 5: An example of application of the heuristic algorithm.

Step 0. Compute $L(t)$ for each t, solving the problem (LPP2).

Step 1. Solve the problem (LPP1). Let $\Gamma^{(i)}_{PS}$ be its optimum value.

Step 2. For $k := 1$ to K solve the linear programming problem (LPP$_{(k)}$):

$$\begin{array}{ll} \text{maximize} & Y^T \cdot PRE \cdot \vec{G}^{(i)}_k \\ \text{subject to} & Y^T \cdot PRE \cdot \vec{D}^{(i)} = \Gamma^{(i)}_{PS} \\ & Y^T \cdot C = 0;\ Y^T \cdot M_0 = 1 \\ & Y \geq 0 \end{array}$$

where $\vec{G}^{(i)}_k$ is a vector with dimension equal to the number of transitions and components

$$G^{(i)}_k(t_j) = \frac{s_j\ v^{(i)}(t_j)}{1 + k(L(t_j) - 1)/K}$$

Let Y_k be one optimum solution of (LPP$_{(k)}$), $k = 1, \ldots, K$.

Step 3. For $k := 1$ to K compute the mean interfiring time $\Gamma^{(i)}_k$ of the isolated P-component associated to Y_k assuming $L(t)$–server semantics for each transition t, using (for instance) the mean value analysis algorithm.

Step 4. Select $\max\{\Gamma^{(i)}_k \mid k = 1, \ldots, K\}$ as the new lower bound for the mean interfiring time of transition t_i.

As an example, let us consider the live and bounded free choice net in figure 5.a. In fact, we have selected a marked graph for simplicity: in this case the subnets generated by minimal P-semiflows are *elementary circuits*. From a queueing theory perspective, the example does not lose generality because even though the embedded queueing networks have visit ratios equal to one for all transitons, arbitrary average service demands can be obtained changing the associated mean service times. Assume that service times of transions are exponential with means $s_1 = s_2 = s_3 = s_4 = s_5 = s_6 = s_7 = s_8 = 2$ and $s_9 = s_{10} = s_{11} = s_{12} = s_{13} = s_{14} = s_{15} = s_{16} = 1$. Then, the application of (LPP1) gives $\Gamma^{(1)} \geq \frac{8}{2} = 4$. This optimum value is obtained for two different feasible solutions (circuits), generated by the P-semiflows:

$$\begin{aligned} Y_1 &= (1,1,1,1,0,0,0,0,0,\ldots,0)^T \\ Y_2 &= (0,0,0,0,1,1,1,1,0,\ldots,0)^T \end{aligned}$$

The application of the above heuristic for $K = 1$ selects the first one, because the liveness bounds of the involved transitions in Y_1 are equal to 1 (less than those in Y_2): ($\text{LPP}_{(1)}$) gives the optimum value equal to 4 for Y_1, while the other feasible solution Y_2 gives only 2 (the service times of the involved transitions in Y_2 are divided by their corresponding liveness bounds, which are all equal to 2). Then, the queueing network generated by Y_1, with liveness bounds of transitions equal to 1, must be solved (figure 5.b). Mean value analysis aplied to this network gives the following bound for the mean interfiring time of t_1: $\Gamma^{(1)} \geq 5$. While the exact mean interfiring time in the whole net (obtained solving the embedded continuous time Markov chain, with 10515 states) is $\Gamma^{(1)} = 5.87$.

Let us remark that, in the particular case in which the liveness bounds of all transitions were equal ($L(t) = L$, for all transition t), the problems ($\text{LPP}_{(k)}$) would not select any "better" solution. All the feasible solutions would give the same value for the objective function of each ($\text{LPP}_{(k)}$): $\Gamma_{PS}^{(i)}/(1 + k(L-1)/K)$. In this case the heuristic used by the above algorithm is not good. Fortunately, this case is easy to detect (at Step 0), and there exists an alternative heuristic for the selection of an optimum solution of (LPP1):

$$\begin{aligned} \text{maximize} \quad & Y^T \cdot M_0 \\ \text{subject to} \quad & Y^T \cdot (PRE \cdot \vec{D}^{(i)} - \Gamma_{PS}^{(i)} M_0) = 0 \\ & Y^T \cdot C = 0; Y \geq 0 \end{aligned} \qquad \text{(LPP3)}$$

That is, since all P-components include transitions with the same maximum number of servers, we can expect that the slowest P-component is the one with the maximum number of tokens, and thus with maximum response time at places, waiting for an available server.

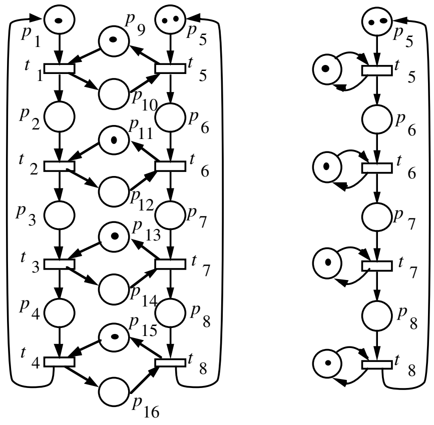

Figure 6: Application of the heuristic when the liveness bounds of all transitions are equal.

As an example, look at the net depicted in figure 6.a. Assume exponential service time distributions with means $s_1 = s_2 = s_3 = s_4 = 2$ and $s_5 = s_6 = s_7 = s_8 = 4$. The application of (LPP1) gives $\Gamma^{(1)} \geq \max\left\{\frac{8}{1}, \frac{16}{2}, \frac{6}{1}\right\} = 8$. The optimum value is reached with two different feasible solutions, the circuits generated by:

$$\begin{aligned} Y_1 &= (1,1,1,1,0,0,0,0)^T \\ Y_2 &= (0,0,0,0,1,1,1,1)^T \end{aligned}$$

The liveness bound of all transitions is equal to 1. Therefore the problem ($\text{LPP}_{(1)}$) does not select any P-component. However, the application of problem (LPP3) selects the circuit generated by Y_2, because it contains a greater number of tokens than the circuit generated by Y_1. Then, the application of the mean value analysis algorithm to the network generated by Y_2, with liveness bound of transitions equal to 1 (figure 6.b) gives the bound $\Gamma^{(1)} \geq 10$. And the exact mean interfiring time of transition t_1 in the original net is $\Gamma^{(1)} = 13.14$.

Last but not least, let us remark that also the net structure-based improvement derived in [7] can be taken into account before the application of the algorithm presented above. In other words, the addition of *implicit places* can generate new, slower P-components in the net [10], that must be considered as feasible solutions of problems ($\text{LPP}_{(k)}$) in the algorithm.

As an example, consider the net in figure 7.a. Exponential service times are associated with transitions t_3 and t_4 with means $s_3 = s_4 = 5$. Transitions

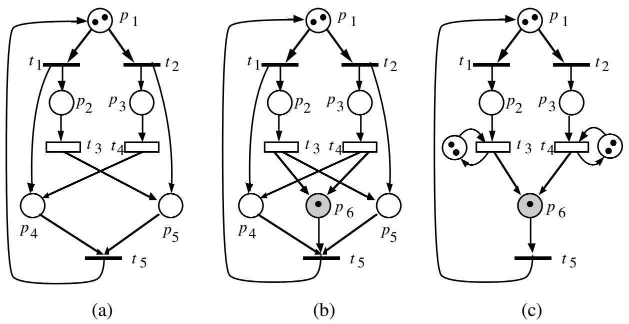

Figure 7: The addition of implicit places improves the bound.

t_1, t_2, and t_5 are immediate. Assuming that the conflict at p_1 is solved with equal probability for t_1 and t_2, the vectors of visit ratios and average service demands to transitions are (see [4]) $\vec{v}^{(1)} = (1,1,1,1,2)^T$ and $\vec{D}^{(1)} = (0,0,5,5,0)^T$. The minimal P-semiflows are $Y_1 = (1,1,0,0,1)^T$ and $Y_2 = (1,0,1,1,0)^T$. The problem (LPP1) gives $\Gamma^{(1)} \geq \max\left\{\frac{5}{2}, \frac{5}{2}\right\} = 2.5$.

Now, if the place p_6 is added to the net with initial marking equal to 1 in order to be implicit (see figure 7.b), the following P-semiflow is generated: $Y_3 = (1,1,1,0,0,1)^T$. The application of (LPP1) yields: $\Gamma^{(1)} \geq \max\left\{\frac{5}{2}, \frac{5}{2}, \frac{10}{3}\right\} = 3.3$.

An additional improvement can be obtained applying corollary 4.1. That is, considering the P-component generated by Y_3 with liveness bounds of t_3 and t_4 equal to 2 (see figure 7.c). The exact solution of this embedded queueing network gives the bound: $\Gamma^{(1)} \geq 3.75$.

The exact mean interfiring time is $\Gamma^{(1)} = 4$. Therefore, the relative error has been reduced from 37.5% of the first bound ($\Gamma^{(1)} \geq 2.5$) to 6.25% of the last one ($\Gamma^{(1)} \geq 3.75$).

4.2 Non-free choice nets case

For other subclasses of nets (e.g., mono-T-semiflow nets), P-semiflows in general do not correspond to P-components of the net. In this case, the subnets generated by the support of the P-semiflows do not have product-form solution and cannot be analysed using the mean value analysis algorithm. In order to solve this problem, a technique consisting of the addition of some implicit places can be used [10]. In fact, for structurally live and structurally bounded nets, a set of implicit places can be added to the net such that it can be covered by P-components, and the algorithm presented in the previous section can be applied to the P-semiflows corresponding to those P-components.

We refer the reader to [10]. Figure 8.a has been taken from that paper, and shows a live and bounded mono-T-semiflow net (which is not a state machine) with a unique minimal P-semiflow which covers all the places: $Y_0 = (2,1,1,1,1)^T$. Therefore, it does not generate a P-component. However, three implicit places can be added (figure 8.b) in such a way that four new P-semiflows are created:

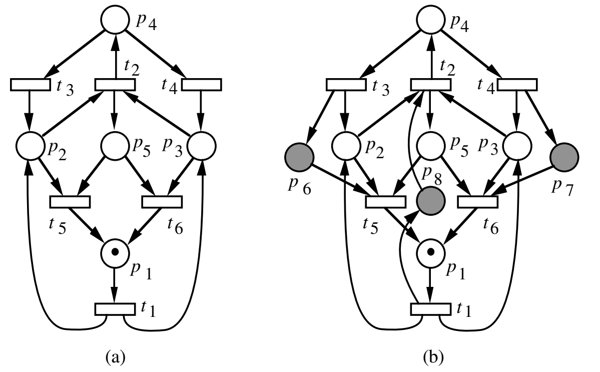

Figure 8: (a) Net non-covered by P-components, which is covered by four P-components after the addition of three implicit places (b).

$$
\begin{aligned}
Y_1 &= (1,0,1,1,0,1,0,0)^T \\
Y_2 &= (1,1,0,1,0,0,1,0)^T \\
Y_3 &= (1,0,0,0,1,0,0,1)^T \\
Y_4 &= (1,0,0,1,0,1,1,1)^T
\end{aligned}
$$

that generate four P-components which cover the whole net. The approach presented in the previous section can be applied to the resulting net.

5 Conclusions

We have addressed the problem of computing upper bounds for the throughput of transitions in Markovian Petri net models (or the corresponding synchronized queueing networks).

Until now, only the net structure, the routing probabilities, and the mean service time of transitions had been used in order to compute such bounds.

In the particular case of exponential distributions associated with timing of transitions, a considerable improvement can be obtained considering the throughput of the slowest subnet generated by a P-semiflow, considered in "partial isolation", i.e., considering the maximum reentrance in steady-state (or liveness bound) of their transitions, allowed by the rest of the net.

For the case of live and bounded free choice nets with Markovian timing (exponential distributions), since all minimal P-semiflows generate "state machine-topology" subnets, they can be seen as embedded closed product-form monoclass queueing networks, and efficient algorithms can be applied for the computation of their exact (or upper bound on) throughput.

For more general structurally live and structurally bounded net subclasses, the addition of a set of implicit places leads to nets that can be covered by subnets with state machine topology. Therefore, the same techniques as for free choice nets can be applied.

References

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

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

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

[4] J. Campos, G. Chiola, and M. Silva. Properties and performance bounds for closed free choice synchronized monoclass queueing networks. *IEEE Transactions on Automatic Control*, Special Mini-Issue on Modeling and Analysis of Multi-Dimensional Queueing Systems. To appear.

[5] 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.

[6] J. Campos and J. M. Colom. A reachable throughput upper bound for live and safe free choice nets. In *Proceedings of the Twelfth International Conference on Application and Theory of Petri Nets*, pages 237–256, Gjern, Denmark, June 1991. Selected for *Advances in Petri Nets 1991*, LNCS, Springer-Verlag.

[7] J. Campos, J. M. Colom, and M. Silva. Improving throughput upper bounds for net based models. In *Proceedings of the IMACS-IFAC SYMPOSIUM Modelling and Control of Technological Systems*, pages 573–582, Lille, France, May 1991. To be published by Elsevier Science Publishers B.V. (North-Holland).

[8] 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*, Melbourne, Australia, December 1991. IEEE-Computer Society Press.

[9] G. Chiola. A graphical Petri net tool for performance analysis. In *Proceedings of the* 3^{rd} *International Workshop on Modeling Techniques and Performance Evaluation*, Paris, France, March 1987. AFCET.

[10] 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 *LNCS*, pages 113–145. Springer-Verlag, Berlin, 1991.

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

[12] J. Esparza and M. Silva. On the analysis and synthesis of free choice systems. In G. Rozenberg, editor, *Advances in Petri Nets 1990*, volume 483 of *LNCS*, pages 243–286. Springer-Verlag, Berlin, 1991.

[13] G. Florin and S. Natkin. Les réseaux de Petri stochastiques. *Technique et Science Informatiques*, 4(1):143–160, February 1985. In French.

[14] M. H. T. Hack. Analysis of production schemata by Petri nets. M. S. Thesis, TR-94, M.I.T.,Boston, USA, 1972.

[15] J. D. C. Little. A proof of the queueing formula $L = \lambda W$. *Operations Research*, 9:383–387, 1961.

[16] T. Murata. Petri nets: Properties, analysis, and applications. *Proceedings of the IEEE*, 77(4):541–580, April 1989.

[17] 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, The Netherlands, 1989.

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

[19] 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.