Performance Analysis of Live and Bounded Free Choice Systems

Javier Campos*

1 Introduction

Timed and stochastic Petri nets constitute an adequate model for the evaluation of performance measures of concurrent and distributed systems. Nevertheless, one of the main problems in the actual use of these models for the evaluation of large systems is the *explosion of the computational complexity of the analysis algorithms*. In general, exact performance results are obtained from the numerical solution of a *Markov chain*, whose dimension is given by the size of the state space of the model. Only for very restricted net subclasses, exact analysis can be achieved with *structural techniques* at the net level (see, e.g., [11]).

Quick computation of performance bounds is a complementary approach to the exact analysis, specially useful in the preliminary phases of the design of a system. Structural techniques for the computation of bounds have been developed within the DEMON project. These results have been obtained by bridging two active fields:

- Qualitative theory of Petri nets (and in particular, structural analysis techniques).

- Theory of stochastic models (stochastic Petri nets and queueing networks).

Different net subclasses, such as live marked graphs [5], mono-T-semiflow nets [7], live and bounded free choice systems [6], and live and bounded FRT-nets (nets with *Freely Related T-semiflows*) [4], have been considered. The results include two levels of accuracy depending on the knowledge about the distributions associated with the enabling time of transitions: the first, less accurate, for *general service* distributions and the second, more tight, for exponentially distributed service times (*Markovian nets*).

2 Organisation

In this report, some of the obtained results for the case of live and bounded free choice (LBFC) systems with general service time distributions are collected. We first include a

*Departamento de Ingeniería Eléctrica e Informática, Universidad de Zaragoza, Maria de Luna 3, E-50015 Zaragoza, Spain. This work has been partially supported by the projects PA86-0028 of the Spanish Comisión Interministerial de Ciencia y Tecnología (CICYT), TIC 0358/89 of the Spanish Plan Nacional de Investigación, and the DEMON ESPRIT Basic Research Action 3148.

brief discussion on the timing semantics of net models and recall basic qualitative results and their crucial consequences for the performance analysis. In the next two sections, throughput bounds for general service time distributions and for exponential distributions are presented. Some concluding remarks and references to extensions of the work are contained in the final section.

3 Stochastic Petri nets and synchronized queueing networks

Stochastic Petri nets are usually introduced as autonomous net systems with the addition of a timing interpretation. From a different perspective, they can be seen as *queueing network models* extended with a general *synchronization primitive* (preserving the free choice decision scheme, in the cases that we consider here). Places of the net represent the *queueing rooms* while transitions represent the *service stations*. In this section, we briefly recall the different implications that the addition of a timing interpretation has in the net model, using classical queueing theory terminology.

3.1 Introducing the time

Since Petri nets are bipartite graphs, historically there have been two ways of introducing the concept of time in them, namely, associating a time interpretation (deterministic or stochastic) with either transitions [16] or places [19]. Since transitions represent activities that change the state (marking) of the net system, it seems natural to associate a duration with these activities (transitions). The latter has been our choice. In other words, from a queueing theory perspective, the waiting rooms are represented by places, the service stations are represented by timed transitions, and we denote by s_i the *average service time* of transition t_i.

In the case of timed transition models, two different firing rules have been defined:

1) "timed firing" of transitions in three phases which changes the firing rule of Petri nets introducing a timed phase in which the transition is "working" after having removed tokens from the input and before adding tokens to the output places, or a

2) "timed enabling" followed by an atomic firing which does not affect the usual Petri net firing rule.

These different timing interpretations have different implications on the resolution of *conflicts* [1]. On the one hand, using timed transition models with three phases firing we can define a policy for conflict resolution independent of the time specification but we cannot model *pre-emption*. On the other hand, using timed transition models with single phase firing we can model pre-emption but we cannot define conflict resolution policies independent of the timing specification (the conflicts are usually resolved in this case with *race policy*, i.e., the transition which samples the minimum service time is the one whose firing determines the change of marking).

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* associated with immediate transitions (like in *generalized stochastic Petri nets* [3]). In this way, pre-emption cannot be modelled. 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 $r_1, \ldots, r_k \in \mathbb{N}^+$ are explicitly defined in the net interpretation in such a way that when $t_1, \ldots, t_k$ are enabled, transition t_i, $i = 1, \ldots, k$, fires with probability (or with long run rate, in the case of deterministic conflicts resolution policy) $r_i / (\sum_{j=1}^{k} r_j)$. 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.

In summary, we model service stations by means of (deterministic or stochastic) timed transitions, routing by means of immediate transitions in conflict, and both kinds of transitions, timed and immediate, can be used as fork (split) nodes and join (synchronization) nodes.

Another possible source of confusion in the definition of the timed interpretation of a Petri net model is the concept of *degree of enabling* of a transition (or re-entrance). In the case of timing associated with places, it seems quite natural to define an unavailability time which is independent of the total number of tokens already present in the place, and this can be interpreted as an *infinite server* policy from the point of view of queueing theory. In the case of time associated with transitions, it is less obvious a-priori whether a transition enabled K times in a marking M (i.e., $K = \max \{ \; k \mid M \geq k \; PRE[t] \; \}$) should work at conditional speed 1 or K times that it would work in the case it was enabled only once. In the case of stochastic Petri nets with exponentially distributed service times associated with transitions, the usual implicit hypothesis is to have *single server* semantics (see, e.g., [15, 13]), and the case of *multiple server* is handled as a case of service rate dependent on the marking; this trick cannot work in the case of more general probability distributions. This is the reason why people working with deterministic timed transitions Petri nets prefer an infinite server semantics (see, e.g., [17, 14, 20]). 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 was adopted in this work.

3.2 Basic qualitative/quantitative results

In order to compute the steady-state performance of a system we have to assume that some kind of "average behaviour" can be estimated on the long run of the system we are studying. The usual assumption in this case is that the system model must be *ergodic*, meaning that at the limit, when the observation period tends to infinity, the estimates of average values tend (almost surely) to the theoretical expected values of the (usually unknown) probability distribution functions that characterize the performance indices of interest.

In the case of stochastic Petri nets, ergodicity condition can be considered for the two most important associated stochastic processes: the marking process and the firing process.

The usual ergodicity assumption is very strong and difficult to verify in general; moreover, it creates problems when we want to include the deterministic case as a special case of a stochastic model, since the existence of the theoretical limiting expected value can be hampered by the periodicity of the model. Thus we prefer the concept of *weak ergodicity* that allows the estimation of long run performance also in the case of deterministic models. For stochastic Petri nets, weak ergodicity of the marking and the firing processes can be defined in the following terms:

Definition 3.1 [6] *The marking process M_τ, where $\tau \geq 0$ represents the time, of a stochastic marked net is weakly ergodic iff the following limit exists:*

$$\lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau M_u \, du = \overline{M} < \vec{\infty}, \ \ a.s. \tag{1}$$

and the constant vector $\overline{M}$ is called the limit average marking.

The firing process $\vec{\sigma}_\tau$, where $\tau \geq 0$ represents the time, of a stochastic marked net is weakly ergodic iff the following limit exists:

$$\lim_{\tau \to \infty} \frac{\vec{\sigma}_\tau}{\tau} = \vec{\sigma}^* < \vec{\infty}, \ \ a.s. \tag{2}$$

and the constant vector $\vec{\sigma}^*$ is the limit of transition throughputs (or limit firing flow vector).

For the case of bounded systems, the existency of the limits $\overline{M}$ and $\vec{\sigma}^*$ is close related to the *home state* existence. Since LBFC systems have home state, the next statement holds:

Theorem 3.1 [6] *Let $\langle \mathcal{N}, M_0 \rangle$ be an LBFC system with deterministic or stochastic service times of transitions. Then, both the marking and the firing processes of $\langle \mathcal{N}, M_0 \rangle$ are weakly ergodic.*

The computation of bounds for the throughput σ_j^* of transition t_j or its inverse, the *mean interfiring time* $\Gamma^{(j)} = 1/\sigma_j^*$, requires to obtain before the *relative throughputs* of transitions that, in order to approach with queueing theory terminology, we call *visit ratios*. The vector of visit ratios, normalized for transition t_j, is a vector $\vec{v}^{(j)}$ with components:

$$v_i^{(j)} \stackrel{\text{def}}{=} \frac{\sigma_i^*}{\sigma_j^*}, \quad i = 1, \ldots, m \tag{3}$$

where σ_i^* is the i^{th} component of $\vec{\sigma}^*$.

The computability of the vector of visit ratios on different system parameters induces a hierarchy of nets where some well-known subclasses are re-encountered (see table 1). In [4], the most general class of structurally live and structurally bounded nets whose vector of visit ratios can be computed from the structure and the routing rates at conflicts has been defined and characterized.

For LBFC systems, the vector of visit ratios for transitions can be computed in polynomial time, from the net structure and the routing rates at conflicts. This computation takes into account that $\vec{v}^{(j)}$ must be a non-negative right annuller of the incidence matrix, with

strongly connected marked graphs	$\vec{v}^{(j)} = 1\!\!1$ (constant)
mono-T-semiflow nets	$\vec{v}^{(j)} = \vec{v}^{(j)}(\mathcal{N})$
live and bounded free choice systems	$\vec{v}^{(j)} = \vec{v}^{(j)}(\mathcal{N}, \mathcal{R})$
simple nets	$\vec{v}^{(j)} = \vec{v}^{(j)}(\mathcal{N}, \mathcal{R}, M_0, \vec{s})$

Table 1: Computability of the vector of visit ratios and net subclasses ($\mathcal{R}$ represents the routing rates at conflicts).

the j^{th} component equal to 1, and such that its components corresponding to any set of transitions $t_1, \ldots, t_k$ in structural conflict must verify the following relations with respect to the associated routing rates $r_1, \ldots, r_k$:

$$\begin{array}{rcl} r_2 v_1^{(j)} & = & r_1 v_2^{(j)} \\ r_3 v_2^{(j)} & = & r_2 v_3^{(j)} \\ & \cdots & \\ r_k v_{k-1}^{(j)} & = & r_{k-1} v_k^{(j)} \end{array} \tag{4}$$

Now, considering all structural conflict sets, the previous relations can be represented straightforward in a matrix form as: $R \cdot \vec{v}^{\,(j)} = 0$. Therefore, the following theorem can be stated:

Theorem 3.2 [6] *Let $\langle \mathcal{N}, M_0 \rangle$ be an LBFC system. Let C be the incidence matrix of $\mathcal{N}$ and R the previously introduced matrix. Then, the vector of visit ratios $\vec{v}^{\,(j)}$ normalized for transition t_j can be computed from C and R by solving the following linear system of equations:*

$$\begin{pmatrix} C \\ R \end{pmatrix} \cdot \vec{v}^{\,(j)} = 0, \quad v_j^{(j)} = 1 \tag{5}$$

Corollary 3.1 *The computation of the vector of visit ratios for transitions for LBFC systems is polynomial on the net size.*

The reader can notice that a rank condition over the incidence matrix C exists underlying the theorem 3.2: the system of equations (5) has a unique solution $\vec{v}^{\,(j)}$ if and only if $rank(C) = m - 1 - (a - n)$ with "a" the number of input arcs to transitions (because $a - n$ is the number of independent relations fixed by the routing rates at conflicts, i.e., the rank of matrix R). The subclass of structurally live and structurally bounded nets satisfying a rank condition like the above one is, in fact, more extensive than free choice nets. The definition of that subclass (FRT-nets) can be found in [4].

Unfortunately, in the general Petri net case, it is not possible to derive the visit ratios only from C and R. For instance, for *simple nets* the visit ratios depend on C, R, and also on M_0 and s_i, $i = 1, \ldots, m$ [6].

Once the vector of visit ratios has been derived, a compact way of including its information together with the mean service times of transitions is to define the vector of *average service demands* for transitions, $\vec{D}^{(j)}$, with components:

$$D_i^{(j)} \stackrel{\text{def}}{=} s_i v_i^{(j)} , \quad i = 1, \ldots, m \tag{6}$$

The performance of a model with infinite-server semantics depends on the maximum degree of enabling of the transitions; and in particular, the steady-state performance depends on the maximum degree of enabling of transitions in steady-state, which in general can be different from the maximum degree of enabling of transitions during its evolution starting from the initial marking (see an example in [7]). For this reason we recall here two concepts of degree of enabling of a transition t: the *enabling bound*, $E(t)$, and the *liveness bound*, $L(t)$. The last is obviously constrained to the steady-state. They allow the generalization of the classical concepts of enabling and liveness of a transition.

Definition 3.2 [7] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The enabling bound of a given transition t of $\mathcal{N}$ is:*

$$E(t) \stackrel{\text{def}}{=} \max\{k \mid \exists M \in R(\mathcal{N}, M_0) : M \geq kPRE[t]\} \tag{7}$$

Definition 3.3 [7] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The liveness bound of a given transition t of $\mathcal{N}$ is:*

$$L(t) \stackrel{\text{def}}{=} \max\{k \mid \forall M' \in R(\mathcal{N}, M_0), \exists M \in R(\mathcal{N}, M') : M \geq kPRE[t]\} \tag{8}$$

The liveness bound of a transition represents the maximum number of servers that are available in steady-state in that transition. The enabling and liveness bounds are behavioural properties, of complex computation. Since we are looking for computational techniques at the structural level, we can also introduce the structural counterpart of the enabling bound concept. To do this, the reachability condition is substituted by the weaker (linear) constraint that markings satisfy the *net state equation*: $M = M_0 + C \cdot \vec{\sigma}$, with $M, \vec{\sigma} \geq 0$.

Definition 3.4 [7] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The structural enabling bound of a given transition t of $\mathcal{N}$ is:*

$$SE(t) \stackrel{\text{def}}{=} \begin{array}{ll} \text{maximum} & k \\ \text{subject to} & M_0 + C \cdot \vec{\sigma} \geq kPRE[t]; \ \vec{\sigma} \geq 0 \end{array} \tag{LPP1}$$

The following result relates the three above presented concepts:

Theorem 3.3 [6] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system.*

1. *$SE(t) \geq E(t) \geq L(t)$, $\forall t \in T$.*
2. *If $\langle \mathcal{N}, M_0 \rangle$ is reversible (i.e., M_0 is a home state) then $E(t) = L(t)$, $\forall t \in T$.*
3. *If $\langle \mathcal{N}, M_0 \rangle$ is an LBFC system then $SE(t) = E(t) = L(t)$, $\forall t \in T$.*

Statement 3 in the above theorem allows to compute $L(t)$ in polynomial time for LBFC systems, by solving the linear programming problem (LPP1).

4 Throughput bounds for LBFC systems with general services

In this section we present upper and lower bounds for the steady-state throughput of transitions of LBFC systems, with general service time distributions, i.e., without any assumption on the probability distribution functions associated with the service time of transitions. Only mean values and not the higher moments of the random variables that describe the timing of the system are considered. This fact is useful in practice since higher moments of the delays are usually unknown for real cases, and difficult to estimate and assess.

4.1 Basic results

Basically, throughput upper bounds are computed by finding the slowest isolated subnet among those generated by P-semiflows of the net, and are presented in the next theorem.

Theorem 4.1 [6] *For LBFC systems, a lower bound for the mean interfiring time $\Gamma^{(j)}$ of transition t_j (or its inverse an upper bound for the throughput) can be computed by solving the following linear programming problem:*

$$\begin{array}{rll} \Gamma^{(j)} \geq & \text{maximum} & Y^T \cdot PRE \cdot \vec{D}^{(j)} \\ & \text{subject to} & Y^T \cdot C = 0 \\ & & Y^T \cdot M_0 = 1 \\ & & Y \geq 0 \end{array} \qquad \text{(LPP2)}$$

where PRE and C are matrices representing the Pre and global incidence functions of the net, M_0 is the initial marking, and $\vec{D}^{(j)}$ is the vector of average service demands for transitions.

We remark that the computation of the above bound has polynomial complexity on the net size. This is because the computation of vector $\vec{D}^{(j)}$ is polynomial (by definition and corollary 3.1) and because linear programming problems can also be solved in polynomial time.

If the solution of (LPP2) is unbounded and since it is a lower bound for the mean interfiring time of transition t_j, the non-liveness can be assured (infinite interfiring time). If the visit ratios of all transitions are non-null, the unboundedness of the problem (LPP2) implies that a total deadlock is reached by the net. This result has the following interpretation: if (LPP2) is unbounded then there exists an unmarked P-semiflow, and the net is non-live.

For live and bounded marked graphs, the bound derived from theorem 4.1 has been shown to be reachable, for arbitrary values of the means and variances of the random variables defining the services times [5]. Unfortunately, this is not the case for LBFC systems. However, some obtained improvements are reported in the next section.

Concerning throughput lower bounds, they can be derived by adding the service time of all transitions, divided by their corresponding liveness bounds, and weighted by the visit ratios. This computation implies an almost complete sequentialization of all the activities represented in the model (we say "almost" because self-concurrency at each transition is

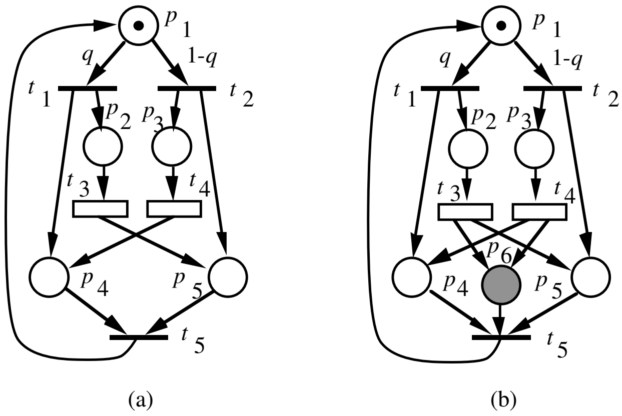

Figure 1: (a) An LBFC system and (b) the addition of the implicit place p_6.

considered, dividing by the liveness bound; i.e., each transition is entered in the accounts with a number of servers equal to its liveness bound).

Theorem 4.2 [6] *For LBFC systems, an upper bound for the mean interfiring time $\Gamma^{(j)}$ of transition t_j (or its inverse a lower bound for the throughput) is:*

$$\Gamma^{(j)} \leq \sum_{i=1}^{m} \frac{s_i}{L(t_i)} \; v_i^{(j)} = \sum_{i=1}^{m} \frac{D_i^{(j)}}{L(t_i)} \tag{9}$$

where s_i, $v_i^{(j)}$, $D_i^{(j)}$, and $L(t_i)$ are the mean service time, visit ratio, average service demand, and liveness bound, respectively, for transition t_i, $i = 1, \ldots, m$.

We remark that the above bound can also be computed in polynomial time, since both the vector of visit ratios and the liveness bounds of transitions can be computed with such complexity (by corollary 3.1 and theorem 3.3.3, respectively).

The bound given by theorem 4.2 has been shown to be reachable for live and bounded marked graphs [5], but not for LBFC systems.

4.2 Improvements

In this section we summarize some ideas for improving the throughput upper bound presented in theorem 4.1. The first one concerns the addition of *implicit places* to the LBFC system. Let us explain here this improvement by using an example. Consider the LBFC system depicted in figure 1.a. Let s_3 and s_4 be the mean service times of 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 routing probabilities defining the resolution of conflict at place p_1. The vector of visit ratios normalized for t_5 is:

$$\vec{v}^{\,(5)} = (q, 1-q, q, 1-q, 1)^T \tag{10}$$

All P-semiflows can be generated by non-negative linear combinations of:

$$\begin{aligned} Y_1 &= (1,1,0,0,1)^T \\ Y_2 &= (1,0,1,1,0)^T \end{aligned} \tag{11}$$

Then, applying the problem (LPP2) to this net, the following lower bound for the mean interfiring time of t_5 is obtained:

$$\Gamma^{(5)} \geq \max\{qs_3, (1-q)s_4\} \tag{12}$$

while the exact mean interfiring time for this transition is:

$$\Gamma^{(5)} = qs_3 + (1-q)s_4 \tag{13}$$

independently of the higher moments of the probability distribution functions associated with transitions t_3 and t_4. Therefore, the bound given by theorem 4.1 is non-reachable for the net in figure 1.a.

Now, let us consider the net in figure 1.b, where the implicit place p_6 has been added to the original net. The addition of implicit places generates the appearance of more elementary P-semiflows. In this case:

$$Y_3 = (1,1,1,0,0,1)^T \tag{14}$$

Then, the application of (LPP2) can eventually lead to an improvement of the previous bound. For this net:

$$\Gamma^{(5)} \geq \max\{qs_3, (1-q)s_4, qs_3 + (1-q)s_4\} = qs_3 + (1-q)s_4 \tag{15}$$

which is exactly the actual mean interfiring time of t_5.

More details for this technique can be found in [9]. In general, the obtained bound is non-reachable. As an example, look at the net in figure 2.a. The exact mean interfiring time of t_7 for deterministic timing is:

$$\Gamma^{(7)} = \max\{qs_3+s_6, (1-q)s_4+s_5, qs_3+(1-q)s_4+(1-q)s_5+qs_6, qs_5+(1-q)s_6\}+s_7 \tag{16}$$

and its clearly greater than the bound obtained after the addition of the implicit place p_8 (figure 2.b):

$$\Gamma^{(7)} \geq \max\{qs_3 + s_6, (1-q)s_4 + s_5, qs_3 + (1-q)s_4\} + s_7 \tag{17}$$

Another approach for improving the throughput upper bound of theorem 4.1 is presented in [8]. It is based on the consideration of some specific *multisets of circuits* of the net in which elementary circuits appear a number of times according to the visit ratios of the involved transitions. Basically, it is a generalization (in the *graph theory* sense) of the application of theorem 4.1 for the case of marked graphs, but now for LBFC systems, because circuits (P-semiflows) of the marked graph are substituted now by multisets of circuits.

In [8], a polynomial method for the computation of the above improvement, derived from multisets of circuits, is presented for the case of live and safe free choice systems. It is based on the application of (LPP2) to a net obtained from the original one after an *expansion* of linear size increasing. The expansion, that is a quick modification of the *Lautenbach transformation for the polynomial computation of minimal traps of a net*, is shown in figure 2.c for the net in figure 2.a. A bijection exists between the minimal P-semiflows of the expanded net and the specific multisets of circuits of interest of the original one. The application of theorem 4.1 to the expanded net gives the bound:

$$\Gamma^{(7)} \geq \max\{qs_3+s_6, (1-q)s_4+s_5, qs_3+(1-q)s_4+(1-q)s_5+qs_6, qs_5+(1-q)s_6\}+s_7 \tag{18}$$

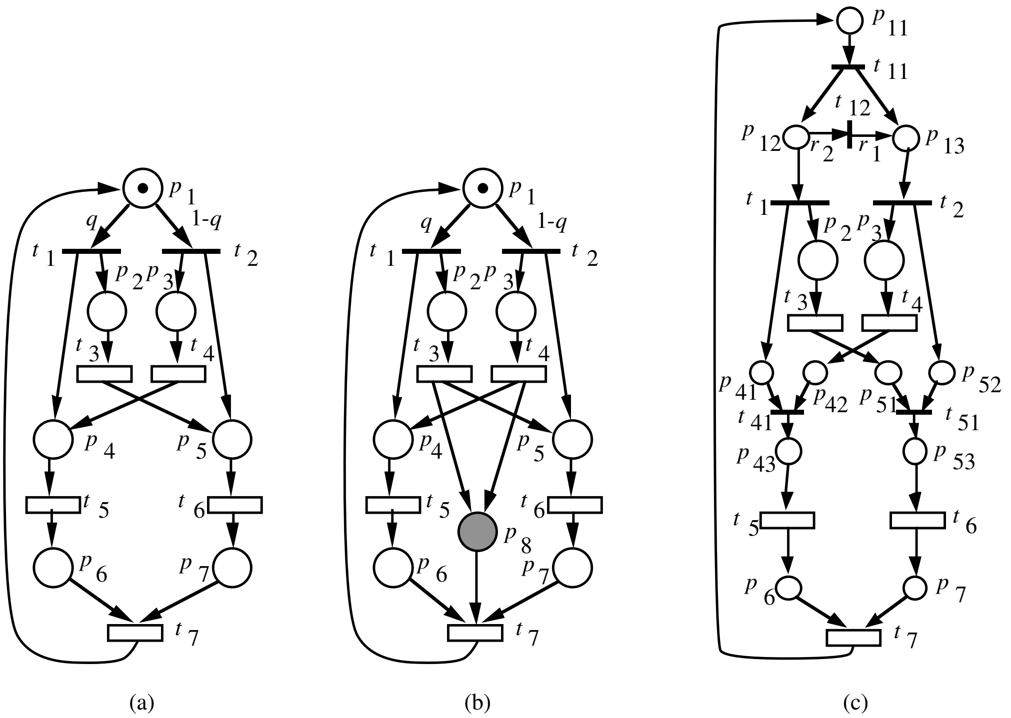

Figure 2: (a) An LBFC system, (b) the addition of the implicit place p_8, and (c) the expansion for obtaining a reachable bound (the weights r_1 and r_2 are such that $r_1/r_2 = q/(1-q)$).

which is the exact mean interfiring time of t_7, given by (16), for deterministic service times of transitions.

The bound obtained using this transformation technique has been shown to be reachable for live and safe free choice systems [8]. Work is in progress for the generalization of the technique to bounded (non-safe) net systems.

5 Throughput bounds for Markovian LBFC systems

Exponential distribution of service is one of the most usual in performance modelling of systems. The main reason is that the *memoryless property* greatly simplifies the analysis of models. Therefore, in this section we assume that timed transitions represent exponentially distributed services with infinite server semantics (*exponential delays*, with queueing network terminology). This particular case of stochastic Petri nets has been considered before in the literature [15, 13, 3] (with single server semantics and marking dependent rates, in the more general case), mainly concerning exact analysis in steady-state. The existing techniques are, in general, *enumerative* since they are based on the solution of an embedded *continuous time Markov chain* whose state space is the set of reachable markings of the net system (there exist efficient techniques only for very restricted net subclasses, e.g., [2]).

In summary, in this section we present better bounds for stochastic Petri nets with exponentially distributed timed transitions and infinite server semantics, that we call *Markovian Petri nets*, for short.

5.1 Embedded queueing networks

Lower bounds for the mean interfiring time of transitions for general service time distributions were introduced in previous section looking for the maximum of the mean interfiring time of transitions of isolated subnets generated by P-semiflows. 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, using once more the concept of liveness bound of transitions. The number of servers at each transition t of a given net in steady-state is limited by 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.

Theorem 5.1 [12] *Let* $\langle \mathcal{N}, M_0 \rangle$ *be an LBFC system 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 (LPP2) and* $\Gamma^{(j)}_{Y_\infty}$ *the corresponding value of the objective function in (LPP2). Let* $\Gamma^{(j)}_{Y_L}$ *be the exact mean interfiring time of* t_j *computed for the isolated subnet generated by* Y*, with* $L(t)$*–server semantics for each involved transition* t*. Then:*

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

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

We restrict ourselves to stochastic nets with null or exponentially distributed service times because in this case there exists a technique for computing the values of $\Gamma^{(j)}_{Y_L}$: Since P-components generated by P-semiflows of LBFC systems 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. Therefore, $\Gamma^{(j)}_{Y_L}$ can be efficiently computed.

Both the bound presented in previous section and the presented in this one 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.

5.2 Transformation techniques

The lower bounds for the throughput of transitions presented in theorem 4.2 are valid for any probability distribution function of service times but can be very pessimistic in some

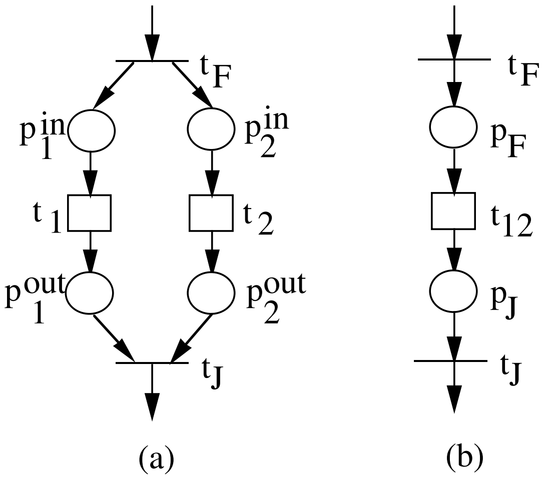

Figure 3: (a) Elementary fork-join and (b) its reduction.

cases, since they are computed assuming a total sequentialization of the work performed at different transitions. Now, an improvement of such results is briefly explained for the case of exponentially distributed service times (details about the techniques presented here can be found in [10]). The basic ideas are:

1. To transform a net into a "simpler" one whose throughput is less than or equal to that of the original net.

2. To evaluate the performance for the derived net, using the bounds presented in theorem 4.2, exact analysis, or any other applicable technique.

In order to obtain better bounds (after these two steps) than the values computed in theorem 4.2, at least one of the transformation rules of item 1 must be less pessimistic than a total sequentialization of the involved transitions. We present firstly a rule whose application allows such *strict* improvement: the *fork-join rule.*

The most simple case of fork-join subnet that can be considered is depicted in figure 3.a. In this case, if transitions t_1 and t_2 have exponential services X_1 and X_2 with means s_1 and s_2, they are reduced to a single transition (figure 3.b) with exponential service time and mean $E[\max\{X_1, X_2\}] = h(s_1, s_2)$, where h is the function:

$$
\begin{aligned}
h(s_1, \ldots, s_n) = & \\
& \sum_{i=1}^{n} s_i - \sum_{\substack{i,j=1 \\ i<j}}^{n} \left(\frac{1}{s_i} + \frac{1}{s_j} \right)^{-1} + \sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} \left(\frac{1}{s_i} + \frac{1}{s_j} + \frac{1}{s_k} \right)^{-1} + \qquad (19) \\
& \cdots + (-1)^{n+1} \left(\frac{1}{s_1} + \cdots + \frac{1}{s_n} \right)^{-1}
\end{aligned}
$$

Therefore, even if the *mean traversing time* of the reduced subnet by a single token has been preserved, the random service time of the new transition t_{12} has more variance than the reduced subnet. Thus, if the reduction rule is applied to an LBFC system, the throughput of the whole system has been reduced. This fact is due to a *monotonicity* result for LBFC systems: the increase of the service time of a transition never produces an increase of the throughput. A trivial extension can be applied if the fork-join subnet includes more than two transitions in parallel. For other net subclasses (different from LBFC), since the monotonicity property does not hold in general, the local reduction rule does not assure that the throughput decreases.

Other transformation rules that have been presented [10] 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. The size of the state space of the model is preserved.

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. Therefore, the state space of the model is reduced.

Split of a transition: this is not a state space reduction rule since it increases the state space of the transformed net. The advantage of the rule is that it allows to proceed further in the reduction process using again the previous rules.

Addition of an immediate transition: With the previous rules all the strongly connected marked graphs which have at least one transition belonging to all circuits of the net can be completely reduced (to a net having only one transition and one place). If this fact does not occur, an immediate transition can be added, synchronizing all circuits, and the complete reduction can be achieved.

Several examples of application of transformation rules and comparison of the obtained bounds with those derived from theorem 4.2 can be found in [10].

6 Conclusions

Within the DEMON project, we have attacked the performance analysis of Petri net systems deeply bridging qualitative and quantitative aspects of the model. From the qualitative point of view, structural techniques for the analysis of Petri net systems have been considered. From the quantitative side, classical queueing theory concepts and laws have been used.

The benefits have been for both the qualitative and quantitative understanding of such models. From the qualitative perspective, some fundamental new results have been obtained. We remark the appearance of original results about the rank of the incidence matrix of structurally live and structurally bounded nets and, in particular, of free choice nets. Several key results in the structure theory of these nets appear as corollaries of the rank theorem.

From the quantitative (performance analysis) point of view, fast algorithms (of polynomial complexity) allow to compute bounds for the throughput of transitions of LBFC systems and more general net subclasses (nets with freely related T-semiflows [4]).

References

[1] M. Ajmone Marsan, G. Balbo, A. Bobbio, G. Chiola, G. Conte, and A. Cumani. The effect of execution policies on the semantics and analysis of stochastic Petri nets. *IEEE Transactions on Software Engineering*, 15(7):832–846, July 1989.

[2] M. Ajmone Marsan, G. Balbo, G. Chiola, and S. Donatelli. On the product-form solution of a class of multiple-bus multiprocessor system models. *Journal of Systems and Software*, 6(1,2):117–124, May 1986.

[3] 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.

[4] 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.

[5] J. Campos, G. Chiola, J. M. Colom, and M. Silva. Properties and performance bounds for timed marked graphs. Research Report GISI-RR-90-17, Departamento de Ingeniería Eléctrica e Informática, Universidad de Zaragoza, Spain, July 1990.

[6] 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.

[7] 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.

[8] 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.

[9] 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 appear in the *IMACS Transactions.*

[10] 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.

[11] J. Campos and M. Silva. Steady-state performance evaluation of totally open systems of Markovian sequential processes. In M. Cosnard and C. Girault, editors, *Decentralized Systems*, pages 427–438. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, The Netherlands, 1990.

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

[14] M. A. Holliday and M. K. Vernon. A generalized timed Petri net model for performance analysis. In *Proceedings of the International Workshop on Timed Petri Nets*, pages 181–190, Torino, Italy, July 1985. IEEE-Computer Society Press.

[15] M. K. Molloy. Performance analysis using stochastic Petri nets. *IEEE Transaction on Computers*, 31(9):913–917, September 1982.

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

[17] R. R. Razouk and C.V. Phelps. Performance analysis using timed Petri nets. In *Proceedings of the International Conference on Parallel Processing*, pages 126–129, August 1984.

[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. Sifakis. Use of Petri nets for performance evaluation. *Acta Cybernetica*, 4(2):185–202, 1978.

[20] W. M. Zuberek. Performance evaluation using timed Petri nets. In *Proceedings of the International Workshop on Timed Petri Nets*, pages 272–278, Torino,Italy, July 1985. IEEE-Computer Society Press.