Structural Techniques and Performance Bounds of Stochastic Petri Net Models

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

In this paper we overview some recent results obtained by the authors and collaborators on the performance bounds analysis of some stochastic Petri net systems. The mathematical model can be seen either as a result of the addition of a particular random timing interpretation to an "autonomous" Petri net or as a generalization of classical queueing networks with the addendum of a general synchronization primitive. It constitutes an adequate tool for both the validation of logical properties and the evaluation of performance measures of concurrent and distributed systems.

Qualitative and quantitative understandings of Petri net models are stressed here making special emphasis on structural techniques for the analysis of logical and performance properties. Important aspects from the performance point of view, such as relative throughput of stations (transitions), and number of servers present at them, are related to Petri net concepts like P- or T-semiflows or liveness bounds of transitions. For the particularly interesting case of Markovian Petri net systems, some improvements of the bounds can be achieved. Marked graphs and free choice are net subclasses for which the obtained results have special quality, therefore an additional attention is focussed on them.

Keywords: graph theory, linear algebra and linear programming techniques, Markovian systems, performance evaluation, P- and T-semiflows, qualitative and quantitative analysis, stochastic Petri net systems, structural techniques, synchronized queueing networks, throughput bounds, transformation/reduction techniques.

Contents

1 Introduction

The increasing complexity of parallel and distributed systems is forcing the researchers to deeply improve the techniques for the analysis of *correctness* and *efficiency* using mathematical models. These two faces, the qualitative validation and the quantitative analysis of models, have been usually developed quasi-independently: "Stochastic Petri Nets (SPN) were initially proposed by researchers active in the applied stochastic modelling field as a convenient graphical notation for the abstract definition of Markovian models. As a consequence, the basic definitions of SPN (and of their variations as well) were originally more concerned with the characteristics of the underlying stochastic process, rather than with the structure of the underlying Petri net model" (quoted from [AM90]). Nevertheless it is easy to accept that SPN represent a meeting point for people working in Petri nets and Performance Evaluation.

Petri nets are a well known mathematical tool extensively used for the modelling and validation of parallel and distributed systems [Pet81, Sil85, Mur89]. Their success has been due not only to the graphical representation, useful in design phases, but mainly to the well-founded theory that allows to investigate a great number of logical properties of the behaviour of the system.

In the framework of Performance Evaluation, *queueing networks* (QN) are the most commonly used models for the analysis of computer systems [Kle76, LZGS84, Lav89]. Such models have the capability of naturally express *sharing of resources* and *queueing*, that are typical situations of traditional computer systems. Efficient solution algorithms, of polynomial complexity on the size of the model, have been developed for important classes of these models, contributing to their increasing success. Many proposals exist to extend the modelling power of queueing networks by adding various synchronization constraints to the basic model [SMK82, VZL87, CCS91b]. Unfortunately, the introduction of synchronization primitives usually destroys the *product form* solution, so that general parallel and distributed systems are not easily studied with this class of models.

More recently, many SPN models have been introduced as formalisms reflecting both the logical aspects, and capable of naturally represent synchronization and concurrency [TPN85, PNPM87, PNPM89, PNPM91]. One of the main problems in the actual use of SPN models for the quantitative evaluation of large systems is the explosion of the computational complexity of the analysis algorithms. In general, exact performance results are impossible to compute. Under important restrictions, enumerative techniques can be employed. For instance, assuming boundedness and exponentially distributed random variables for the transition firing times, performance indices can be computed through the numerical solution of a *continuous time Markov chain*, whose dimension is given by the size of the marking space of the model [Mol82, FN85]. Structural computation of exact performance measures has been only possible for some subclasses of nets, such as those with *state machine* topology. These nets, under certain assumptions on the stochastic interpretation are isomorphic to Gordon and Newell's networks [GN67], in queueing theory terminology. In the general case, efficient methods for the derivation of exact performance measures are still needed.

The final objective of analytical modelling is to obtain information about some performance measures of interest in the system, such as *productivity indices* (e.g., *throughput* of transitions), *responsiveness indices* (e.g., *response time* at places), and other derived *utilization* measures. Several possibilities can be explored depending upon accuracy of results and complexity of algorithms. In this paper we select *performance bounds computation* based on Petri net *structure techniques*, that usually lead to very *efficient solution algorithms*. We try to contribute to bridge the "historical" separation between qualitative and quantitative techniques in the analysis of SPN models: "It should however be stressed that the structural properties of SPN models are today used to either ease the model definition, or to compute very partial results" [AM90]. [AMBCC87] is an example for the first case. [Mol85] and [ZZ90] are examples of the second case: in [Mol85] throughput bounds are obtained under saturation conditions for the most basic SPN model [Mol82], while in [ZZ90] some transformation rules preserving throughput are suggested in a very informal way for a deterministic timing interpretation of net models. We centre our effort in the use of both structure theory of nets and transformation/reduction techniques for deriving efficient methods for performance evaluation. As in the case of qualitative analysis of "autonomous" Petri nets, more powerful results are expected for particular net subclasses (marked graphs, free choice nets...).

Several problems, preliminary to the *exact* analysis of stochastic Petri nets, can be considered that were trivially or easily solved for classical queueing networks in the past. The first of them is the meaning of an average behaviour of the model in the limit of time, i.e., the existence of a *steady-state* behaviour. The concept of *ergodicity*, classical in the framework of Markov processes, was introduced in the field of SPN by G. Florin and S. Natkin [FN85]. It allows to speak about the average behaviour estimated on the *long run* of the system, but it is valid only for very strong assumptions on the *probability distribution functions* (PDF) defining the timing of the model. For instance, deterministic duration of activities do not lead frequently to this kind of ergodic systems. *Weak ergodicity*, introduced in [CCCS89], allows the estimation of long run performances also in the case of deterministic models. Given that we will concentrate on performance bounds instead of exact performance indices, the discussion on ergodicity will not be addressed here. The reader is referred to [FN85, CCS91a, CCS91c].

The first step in the analysis of classical closed queueing networks is the computation of the *relative throughput of stations* or *visit ratios*. It is achieved by solving a system of *flow equations*, which, for each station, equates the rate of flow of customers into to the rate of flow out of the station [Kle75]. Only routing rates among stations are needed in order to do that, thus the visit ratios are the same for arbitrary service times of stations and distribution of customers in the network. The analogous problem is a bit more complicated for SPN. The computation of the relative throughput of transitions independently of the token distribution and of the service times of transitions is only possible for some net subclasses, like *FRT-nets*, a mild generalization of free choice nets [Cam90]. This computation needs the knowledge of the *routing of tokens* through the net system, based on the *deterministic routing* (fixed by the net structure) and the *conditional routing* (defining the resolution of free conflicts).

A complementary aspect to the definition of routing of tokens through the net system is the specification of the *semantics of enabling and firing* of transitions [AMBB+89]. Related concepts in the framework of queueing networks, such as the the *number of servers at each station*, can be redefined for stochastic Petri nets. If we consider *marking bounded* systems, it does not make sense to strictly speak about "infinite" number of servers at transitions. Therefore, a first goal must be to determine the real *maximum enabling degree* of transitions (or *enabling bound*), that will correspond to the number of servers used at them. The maximum number of servers available in steady state will be characterized by the *liveness bound*, a quantitative generalization of the concept of liveness of a transition.

The paper is organized as follows. In section 2, several aspects about the introduction of time in Petri net systems are considered. Relationships between stochastic Petri net systems and queueing networks with synchronizations are indicated. The computation of relative throughput of transitions is presented in section 3, while several concepts of degree of enabling of transitions are introduced and related in section 4. Sections 3 and 4 introduce concepts and results needed in the rest of the paper: the content of section 3 (section 4) is of primary importance for sections 5, 6, and 7.1 (6, 7.1, and 7.2). The computation of *insensitive* throughput bounds (bounds valid for any probability distribution of service times) is considered in sections 5 and 6. For the case of *Markovian Petri nets* some improvements of the insensitive bounds are achieved in section 7. The idea of deriving bounds for other performance indices from throughput bounds is briefly introduced in section 8. Finally, some conclusions are summarized in section 9.

2 Stochastic Petri nets and synchronized queueing networks

In the original definition, Petri nets did not include the notion of time, and tried to model only the logical behaviour of systems by describing the causal relations existing among events. Nevertheless the introduction of a timing specification is essential if we want to use this class of models for performance evaluation of distributed systems. In this section, some considerations are made about the different implications that the addition of a timing interpretation has in Petri net models. The close relations between queueing networks with synchronization primitives and stochastic Petri nets are remarked.

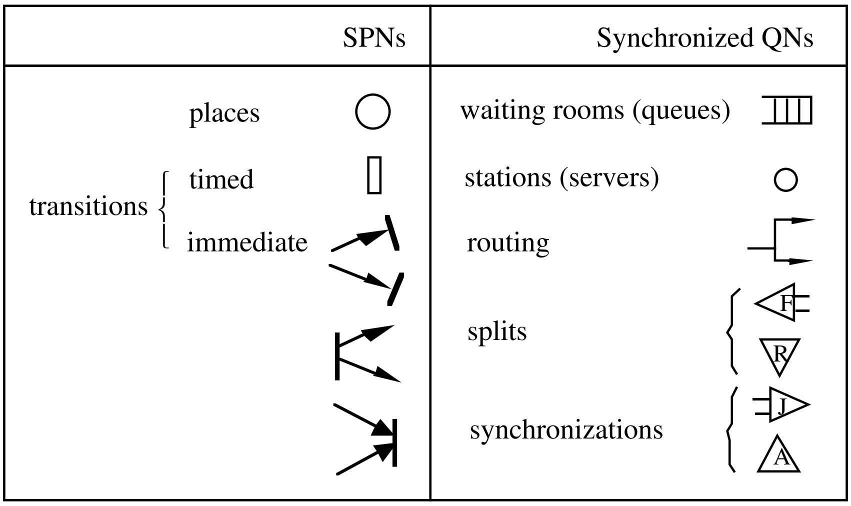
We assume the reader is familiar with the structure, firing rule, and basic properties of net models [Pet81, Sil85, Mur89] Let us just introduce some notations and terminology to be extensively used in the sequel. $\mathcal{N} = \langle P, T, Pre, Post \rangle$ is a net with $n = |P|$ places and $m = |T|$ transitions. We assume $\mathcal{N}$ to be strongly connected. 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). The support of a vector is the set of indices corresponding to non-null components, then the support $||Y||$ ($||X||$) of Y (X) is a subset of places (transitions). A semiflow is elementary if it has minimal support (in the sense of set inclusion) and the greatest common divisor of non-null elements is 1. M (M_0) is a marking (initial marking). $\langle N, M_0 \rangle$ is a net system (or marked net). The symbol σ represents a firable sequence, while $\vec{\sigma}$ is the firing count vector associated to σ: $\vec{\sigma}(t_i)$ is equal to the number of times t_i appears in 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 $\vec{\sigma} \geq 0$.

Among the net subclasses considered in the sequel, state machines, marked graphs, free choice nets, and simple nets are well-studied in the literature (see, for instance, [Mur89]) Finally, we call *mono-T-semiflow nets* [CCS91a] to those having a unique minimal T-semiflow.

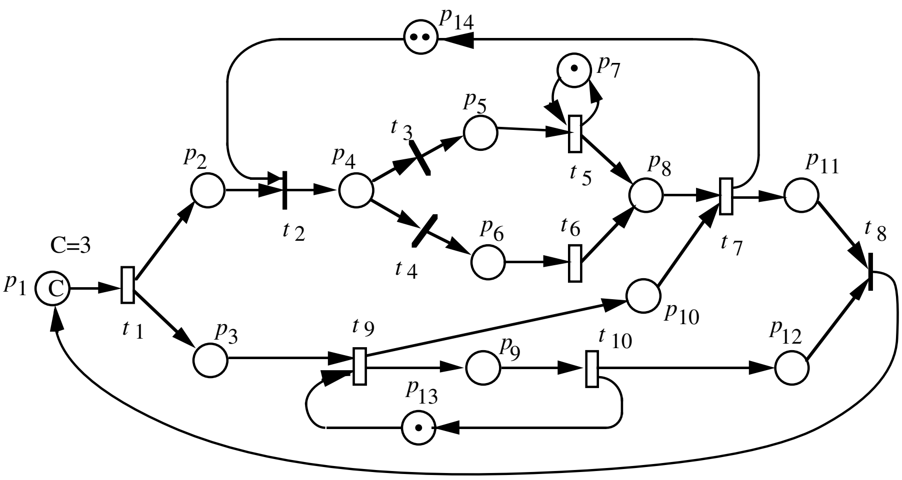
Stochastic Petri nets are defined through a stochastic interpretation of the net model, i.e., *"SPN = PN + stochastic interpretation"*. Looking at the topological and untimed behavioural analogy of strongly connected State Machines and the networks of queues, for certain stochastic interpretations, it can be informally stated that *"SPN = QN + synchronizations"*. The modelling paradigm of SPN in our context is fixed in section 2.1, while section 2.2 consider synchronized queueing networks (SQN) and SPN.

2.1 On Stochastic Petri Nets

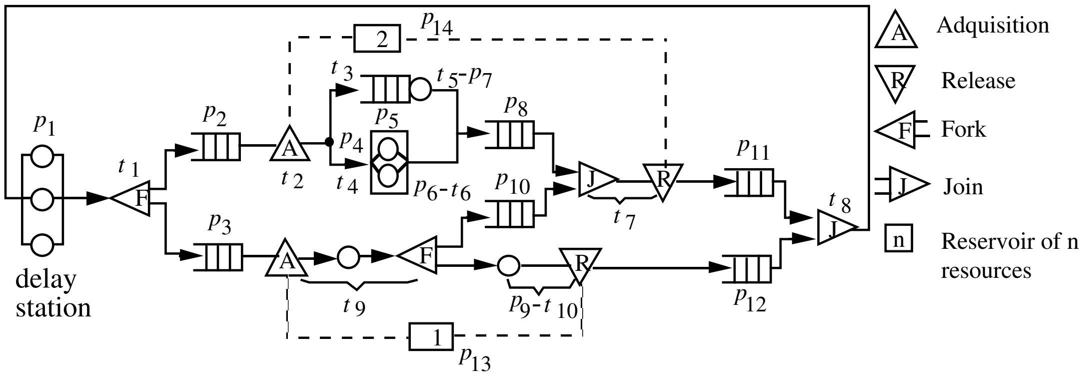
Time has been introduced in Petri net models in many different ways (see, for instance, references in surveys like [AM90, FFN91, Zub91]). Since Petri nets are bipartite graphs, historically there have been two ways of introducing the concept of time in them, namely, associating a time interpretation with either transitions [Ram74] or with places [Sif78]. Since transitions represent activities that change the state (marking) of the net system,



Informally: "SPNs = PNs + stochastic interpretation =
= QNs + general synchronization primitive"



(a) Stochastic Petri net system representation.



(b) Extended queueing network representation.

Figure 1: A stochastic net and the corresponding synchronized queueing network.

it seems natural to associate a duration with these activities (transitions). The latter has been our choice, i.e., we adopt a so called "t-timed interpretation". In the case of timed transition models, two different kinds of firing rules have been defined: *atomic* and *three-phases*. To be fully consistent with qualitative PN theory we adopt the first one, in which a "timed enabling" is eventually followed by an atomic firing (see, for instance, [AM90]). The three phases firing, "timed firing", changes the classical firing rule making some tokens disappear for a while (see, for instance, [Zub91]). Thus classical token conservation laws on places are weakened.

The stochastic timed interpretation of net models requires the specification of the *PDF of the random firing delays* and the *execution policy*. By execution policy it is undertood "the way in which the next transition to fire is chosen, and how the model keeps track of its past history" [AMBB+89].

Historically the first stochastic interpretations associated exponential PDF to the random firing delays, while the usual execution policy was very simple: *race* (that indicates that transitions compete for firing: the competition is won by the transition that samples the shortest delay). This basic model was completed in Generalized Stochastic Petri Nets (GSPN, see [AMBC84, AMBCC87]), adding *immediate* transitions (which fire in zero time with priority over timed transitions) and *inhibitor arcs*. Weights are associated with immediate transitions for the computation of firing probabilities in case of conflict. Immediate transitions allow to define conflict resolution policies independent of the timing specification.

A step further has been trying to increase the modelling power of GSPN and alternative models, allowing arbitrary PDF (deterministic in particular) for the random variables representing the firing delays. Under this circumstance the precise definition of the execution policy becomes crucial because the memoryless property of exponential PDF is lost. In [AMBB+89] the topic is considered in detail, defining some possible execution policies and their modelling and analysis consequences.

Our choice for the stochastic interpretation of net models is guided by the following principles:

1. Transitions to fire should be selected at the net level, independently of the firing delays. Therefore we are in the so called *preselection* execution policy paradigm [AMBB+89]: the activities with transitions that are enabled, but are not preselected, are not executed. The preselection policy is made explicit using *immediate transitions*. In other words, we are looking for something that is typical in QNs: the routing of customers is independent of service times.

   This choice will lead to easy to understand and manipulate models, allowing to state certain performance monotonicity results like in QN [SY86].

2. *Inhibitor arcs* are not allowed. This is a real constraint only for unbounded systems, where PN with inhibitor arcs have been shown to have a modelling power equivalent to Turing machines. But in this case many properties are undecidable! For bounded systems, an inhibitor arc is just a modelling convenience and can be removed, expressing the constraint with normal arcs and eventually new places (see, for instance, [Sil85]).

3. *Priorities* in the firing of transitions are forbidden, except for the two levels derived from the use of *immediate* and *timed* transitions.

4. Synchronizations may be immediate, and we discard the "pathological" situation consisting of a circuit in the net including only immediate transitions.

Under the above choices our stochastic interpretation grounds in a very general framework:

1. Firing delays are random variables with arbitrary PDF. They are assumed to be time and marking independent.

2. If a transition is enabled several times, let us say q times, then q firings progress in parallel at the same time. In other words, we assume the idea that the natural interpretation of parallelism in a PN model leads to an infinite-server semantics, in QN terminology. If this is not appropiated for the system because the number of servers should be constrained to k, then a place self-loop around the transition with k tokens will guarantee the k-server semantics.

3. Any policy for conflict resolution among immediate transitions is allowed (e.g., defined through a deterministic scheduler, through a probabilistic choice...).

In practice, for computing performance bounds, only the *mean firing (service) time* of transitions and *long run routing rates* will be needed by us. More precisely, s_i, the mean firing time for transition $t_i (i = 1, \ldots, m)$ is given, and 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 relative frecuency, $r_i/(\sum_{j=1}^{k} r_j)$. Note that the routing rates are assumed to be strictly positive, i.e., all possible outcomes of any conflict may fire.

In summary, we model services by means of 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.

The main price we pay for the adopted modelling paradigm is the unability to "directly" model situations like preempting scheduling disciplines, time-out mechanisms or unreliable processors which can "fail" during the processing stage. What we basically gain is the possibility of using many results from structure theory of Petri nets and queueing network theory.

A more restricted but easier to analyse stochastic interpretation, associating *time and marking independent exponential PDF* to the firing of transitions and *time and marking independent discrete probability distributions* to immediate transitions, will be called in the present framework *Markovian Petri net systems* (section 7).

2.2 Queueing networks with synchronizations

Many extensions have been proposed to introduce synchronization primitives into the queueing network formalism, in order to allow the modelling of distributed synchronous systems: *passive resources*, *fork and join*, *customer splitting*, etc. Some very restricted

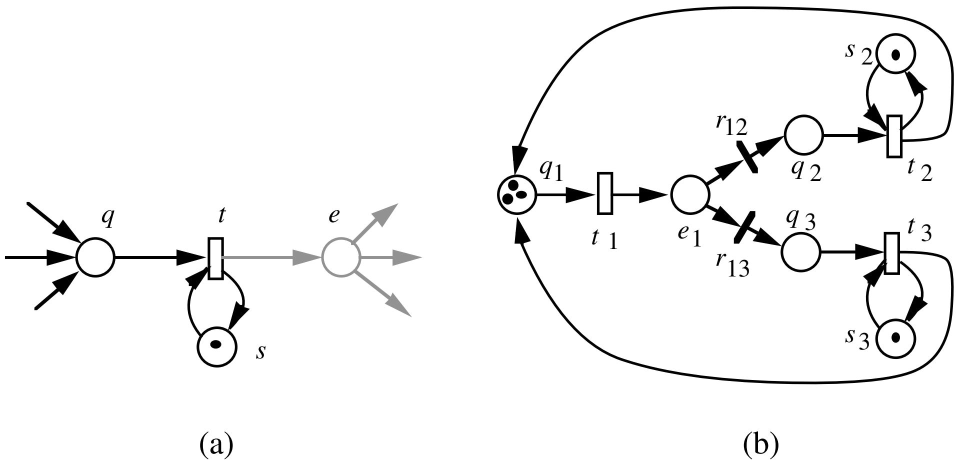


Figure 2: A Petri net representation of (a) a monoclass single server queue and (b) a monoclass queueing network.

forms of synchronization, such as some special use of passive resources [AMBCD86], preserve the *local balance property* [BCMP75] that allows the use of efficient algorithms for the computation of exact *product form solutions.* In general, however, these extensions destroy the local balance property, so the extended queueing models with synchronizations are used mainly as system descriptions for simulation experiments. Even the computation of bounds for these classes of models is not yet well developed.

In [VZL87], a comparison has been proposed between synchronized queueing networks and stochastic Petri nets, showing that *the two formalisms are roughly equivalent from a modelling point of view.* Here we show how the different queueing network models with synchronizations can be uniformly represented within a Petri net formalism.

A monoclass single server station [Kle75] can be modelled by a subnet of the type depicted in figure 2.a. An infinite server queue [Kle75] (i.e., a *pure delay node*) can be represented by a place to model the number of customers in the system and a timed transition connected with the place through an input arc to model departures. *Persistent* timed transitions represent service times of the nodes, while *conflicting* immediate transitions model the routing of customers moving from one node to the other. Queueing networks containing both *delay* and *finite server* nodes are thus naturally modelled by stochastic Petri nets of the type depicted in the example of figure 2.b (t_1 is a delay, while t_2 and t_3 are single server stations). Also in this more general context conflicting immediate transitions model the routing of customers among the stations, while persistent timed transitions model the service times.

On the other hand, stochastic nets can assume forms much more complex than the one illustrated in the example of figure 2.b. Figure 1.a, taken from [CCS91b], illustrates a more general stochastic Petri net that cannot be mapped onto a product form queueing network. In fact, this net can be mapped onto a queueing network extended with synchronization primitives (figure 1.b) [SMK82], in which such constructs as fork, join, and passive resources are used to map the effect of the pairs of transitions t_2–t_7 and t_9–t_{10}, respectively. These examples show how, using a Petri net formalism, extensions of product form queueing networks are represented with an analogous level of structural complexity of Jackson networks.

As a very particular case in which the interest of making a deep bridge between PN and QN theories appears, in [DLT90] it has be pointed out the isomorphism between Fork/Join Queueing Networks with Blocking (FJQN/B) and stochastic strongly connected Marked Graphs.

Finally, let us remark that stochastic Petri nets with weighted arcs (i.e., non-ordinary nets) can be used for the modelling of *bulk arrivals* and *bulk services* [Kle75], with *deterministic size of batches* (given by the weights of arcs).

3 Relative throughput of transitions: Visit ratios

One of the most usual indices of *productivity* in performance evaluation of computer systems is the *throughput* of the different components, i.e., the number of jobs or tasks processed by each component in the unit time. In classical queueing network models, components are represented by stations, and throughput of each component is the average number of service completions of the correspondent station per unit time. For stochastic Petri nets, since actions are represented with transitions, the throughput of a component is the number of firings per unit time of the corresponding transition.

3.1 Classical queueing networks: flow of customers

In a classical monoclass queueing network, the following system can be derived by equating the rate of *flow of customers* into each station to the rate of flow out of the station [Kle75]:

$$\overline{X}(j) = X_{0j} + \sum_{i=1}^{m} \overline{X}(i)\ r_{ij}, \qquad j = 1, \ldots, m \tag{1}$$

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

If the network is open (i.e., if there exists a station j with positive external arrival rate, $X_{0j} > 0$ and also customers can leave the system), then the above m equations are linearly independent, and the exact throughputs of stations can be derived (independently of the service times, s_i, $i = 1, \ldots, m$). This is not the case for closed networks. If $X_{0j} = 0$, $j = 1, \ldots, m$, then only $m - 1$ equations are linearly independent, and thus only ratios of throughputs can be determined. These *relative throughputs* which are often called *visit ratios*, denoted as v_i for each station i, summarize all the information given by the routing that we use for the computation of the throughput bounds. The visit ratios normalized, for instance, for station j are defined as:

$$v_i^{(j)} \stackrel{\text{def}}{=} \frac{\overline{X}(i)}{\overline{X}(j)}, \qquad i = 1, \ldots, m \tag{2}$$

For a restricted class of queueing networks, called *product form networks*, the exact steady-state solution can be shown to be a product of terms, one for each station, where the form of term i is derived from the visit ratio v_i and the service time s_i. The steady-state probability $\pi(\vec{n})$ of state $\vec{n} = (n_1, \ldots, n_m)^T$ (where n_i is the number of customers at center i, including those being served and those waiting) in a closed monoclass product form queueing network with m stations and N customers has the form:

$$\pi(n_1, \ldots, n_m) = \frac{1}{G(N)} \prod_{i=1}^{m} (D_i^{(j)})^{n_i} \tag{3}$$

where $D_i^{(j)}$ is the *average service demand* of customers from station i, defined as:

$$D_i^{(j)} \stackrel{\text{def}}{=} v_i^{(j)}\ s_i, \qquad i = 1, \ldots, m \tag{4}$$

and $G(N)$ is a *normalization constant* defined so that the $\pi(\vec{n})$ sum to 1.

We remark that the knowledge of average service demands is crucial for the computation of exact measures of product form queueing networks.

3.2 Stochastic Petri nets: flow of tokens

Concerning stochastic Petri nets, we assume also that the average service times of transitions are known. Then, in order to compute the average service demand of tokens from transitions, it is necessary to compute just the visit ratios or relative throughputs of transitions.

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

The counterpart of routing of customers in queueing networks consists both on the *net structure* $\mathcal{N}$ and the relative *routing rates at conflicts* (denoted $\mathcal{R}$) in stochastic Petri nets. Unfortunately, in the general Petri net case it is not possible to derive the visit ratios only from $\mathcal{N}$ and $\mathcal{R}$. Net systems can be constructed such that the visit ratios for transitions do depend on the net structure, on the routing rates at conflicts, but also on the initial marking (distribution of customers), and on the average service time of transitions:

$$\vec{v}^{(j)} = \vec{v}^{(j)}(\mathcal{N}, \mathcal{R}, M_0, \vec{s}) \tag{5}$$

where $\vec{v}^{(j)}$ and $\vec{s}$ denote the vectors with components $v_i^{(j)}$ and s_i, $i = 1, \ldots, m$, respectively.

As an example, let us consider the *simple net* depicted in figure 3. Transitions t_1 and t_3 are *immediate* (i.e., they fire in zero time). The constants $r_1, r_3 \in \mathbb{N}^+$ define the conflict resolution policy, i.e., when t_1 and t_3 are simultaneously enabled, t_1 fires with relative rate $r_1/(r_1 + r_3)$ and t_3 with $r_3/(r_1 + r_3)$. Let s_2 and s_4 be the average service times of t_2 and t_4, respectively. If $m_5 = 1$ (initial marking of p_5) then p_1 and p_3 are *implicit* [CS91], hence they can be deleted without affecting the behaviour! Thus a closed queueing network topology is derived. A product form queueing network can be obtained and the visit ratios, normalized for transition t_1 can be computed: $\vec{v}^{(1)} = (1, 1, r_3/r_1, r_3/r_1)^T$. If $m_5 = 2$ (different initial marking for p_5) then p_5 is now implicit, hence it can be deleted; two isolated closed tandem queueing networks are obtained and $\vec{v}^{(1)'} = (1, 1, s_2/s_4, s_2/s_4)^T$. Obviously $\vec{v}^{(1)} \neq \vec{v}^{(1)'}$, in general.

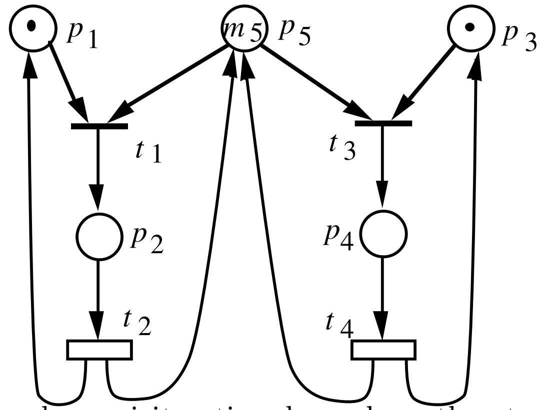


Figure 3: A simple net whose visit ratios depend on the structure, on the routing at conflicts, on the initial marking, and on the service times.

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

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). The computation of that vector is based on the two following facts:

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

$$C \cdot \vec{v}^{(j)} = 0 \tag{6}$$

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

$$\begin{array}{rcl} r_2\vec{v}^{(j)}(t_1) - r_1\vec{v}^{(j)}(t_2) & = & 0 \\ r_3\vec{v}^{(j)}(t_2) - r_2\vec{v}^{(j)}(t_3) & = & 0 \\ & \cdots & \\ r_k\vec{v}^{(j)}(t_{k-1}) - r_{k-1}\vec{v}^{(j)}(t_k) & = & 0 \end{array} \tag{7}$$

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

$$\mathcal{R} = \begin{pmatrix} \mathcal{R}_{T_1} \\ \vdots \\ \mathcal{R}_{T_r} \end{pmatrix} \tag{8}$$

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

$$\begin{pmatrix} C \\ \mathcal{R} \end{pmatrix} \tag{9}$$

is $m - 1$. Particularly interesting subclasses verifying this condition are structurally live and structurally bounded *mono-T-semiflow* nets [CCS91a] and structurally live and structurally bounded *free choice* nets [CCS91b]. We introduce now a more general class of structurally live and structurally bounded nets verifying the previous condition. In order to do that, we define an equivalence relation on the set of T-semiflows of the net. After that, the class of *FRT-nets* will be defined as nets having only one equivalence class for this relation.

Definition 3.1 *Let $\mathcal{N}$ be a Petri net and X_a, X_b two different T-semiflows of $\mathcal{N}$. X_a and X_b are said to be freely connected by places $P' \subset P$, denoted as $X_a \overset{P'}{\wedge} X_b$, iff $\exists t_a \in ||X_a||$, $t_b \in ||X_b||$ such that: $PRE[t_a] = PRE[t_b]$ and ${}^\bullet t_a = {}^\bullet t_b = P'$.*

Definition 3.2 *Let $\mathcal{N}$ be a Petri net and X_a, X_b two T-semiflows of $\mathcal{N}$. X_a and X_b are said to be freely related, denoted as $(X_a, X_b) \in FR$, iff one of the following conditions holds:*

1. *$X_a = X_b$,*
2. *$\exists P' \subset P$ such that $X_a \overset{P'}{\wedge} X_b$, or*
3. *$\exists X_1, \ldots, X_k$ T-semiflows of $\mathcal{N}$ and $P_1, \ldots, P_{k+1} \subset P$, $k \geq 1$, such that $X_a \overset{P_1}{\wedge} X_1 \overset{P_2}{\wedge} \ldots \overset{P_k}{\wedge} X_k \overset{P_{k+1}}{\wedge} X_b$.*

From the above definition the next property trivially follows:

Property 3.1 *FR is an equivalence relation on the set of T-semiflows of a net.*

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

Definition 3.3 [Cam90] *$\mathcal{N}$ is a net with freely related T-semiflows (FRT-net, for short) iff the introduction of the freely relation on the set of its T-semiflows induces only one equivalence class.*

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

Theorem 3.1 [Cam90] *Let $\langle \mathcal{N}, M_0 \rangle$ be a live and structurally bounded FRT-net system. Let C be the incidence matrix of $\mathcal{N}$ and $\mathcal{R}$ the previously introduced matrix. Then, the vector of visit ratios $\vec{v}^{(j)}$ normalized for transition t_j can be computed from C and $\mathcal{R}$ by solving the following linear system of equations:*

$$\begin{pmatrix} C \\ \mathcal{R} \end{pmatrix} \cdot \vec{v}^{(j)} = 0, \quad v_j^{(j)} = 1 \tag{10}$$

The reader can notice that a *rank condition* over the incidence matrix C exists underlying theorem 3.1: the system of equations (10) has a unique solution $\vec{v}^{(j)}$ if and only if $rank(C) = m - \delta - 1$, where δ is the rank of $\mathcal{R}$. For the particular case of free choice nets, a stronger result about the rank of the incidence matrix (that first appeared in [CCS91b]) can be formulated as:

Theorem 3.2 [ES91] *Let $\mathcal{N}$ be a free choice net. $\mathcal{N}$ is structurally live and structurally bounded iff it is conservative, consistent, and $rank(C) = m - 1 - (a - n)$, with $a = \sum_{p \in P, t \in T} PRE[p,t]$ (i.e., the number of input arcs to transitions).*

An important fact about this qualitative property suggested by the performance evaluation analysis is that many of Hack's classical results [Hac72] can be derived from it or the proof process (see [CCS91b] or [ES91]). On the other hand, theorem 3.2 gives a *polynomial* (on the net size) *time* method to decide if a given free choice net is structurally live and structurally bounded.

4 Number of servers at transitions: Enabling and liveness bounds

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.

In section 2.1 we explicitly adopted the convention that 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, M*:

$$E(t,M) \stackrel{\text{def}}{=} \max\{\ k \mid M \geq k\ PRE[t]\}$$

Therefore it can be said that at M, in transition t, $E(t,M)$ 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 $E(t,M) \leq q$.

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.

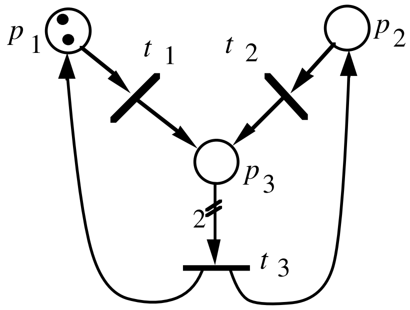


Figure 4: $E(t_1) = 2$, while $L(t_1) = 1$ (i.e., $L(t_1) < E(t_1)$).

Definition 4.1 [CCCS89] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The enabling bound of a given transition t of $\mathcal{N}$ is*

$$E(t) \stackrel{\text{def}}{=} \max \{ \ k \mid \exists \sigma, M_0[\sigma\rangle M : \ M \geq k \ PRE[t] \ \} \tag{11}$$

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.*

Definition 4.2 *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The marking bound of a given place p of $\mathcal{N}$ is*

$$B(p) \stackrel{\text{def}}{=} \max \{ \ M(p) \mid M_0[\sigma\rangle M \ \} \tag{12}$$

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 4.3 [CCS91a] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The liveness bound of a given transition t of $\mathcal{N}$ is*

$$L(t) \stackrel{\text{def}}{=} \max \{ \ k \mid \forall M' : M_0[\sigma\rangle M', \exists M : M'[\sigma'\rangle M \land M \geq k \ PRE[t] \ \} \tag{13}$$

The above definition generalizes the classical concept of liveness of a transition. In particular, a transition t is live if and only if $L(t) > 0$, 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 4.1 [CCS91a] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. For any transition t in $\mathcal{N}$, $E(t) \geq L(t)$ (see figure 4).*

Since for any *reversible* net system (i.e., such that M_0 can be recovered from any reachable marking: M_0 is a *home state*) the reachability graph is strongly connected, the following can be stated:

Property 4.2 [CCS91a] *Let $\langle \mathcal{N}, M_0 \rangle$ be a reversible net system. For any transition t in $\mathcal{N}$, $E(t) = L(t)$.*

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. Structural net theory has been developed from two complementary points of view: graph theory [Bes87] and mathematical programming (or more specifically linear programming and linear algebra) [SC88]. Let us recall our structural definition from the mathematical programming point of view; essentially in this case the reachability condition is substituted by the (in general) weaker (linear) constraint that markings satisfy the net state equation: $M = M_0 + C \cdot \vec{\sigma}$, with $M, \vec{\sigma} \geq 0$.

Definition 4.4 [CCCS89] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The structural enabling bound of a given transition t of $\mathcal{N}$ is*

$$SE(t) \stackrel{\text{def}}{=} \max \quad \{k \mid M = M_0 + C \cdot \vec{\sigma} \geq 0, \vec{\sigma} \geq 0 : M \geq k PRE[t] \ \} \qquad \text{(LPP1)}$$

Note that the definition of structural enabling bound reduces to the formulation of a *linear programming problem*, that can be solved in polynomial time [NRKT89].

Now let us remark the relation between behavioural and structural enabling bound concepts that follows from the implication "$M_0[\sigma\rangle M \ \Rightarrow \ M = M_0 + C \cdot \vec{\sigma} \ \wedge \ \vec{\sigma} \geq 0$".

Property 4.3 [CCS91a] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. For any transition t in $\mathcal{N}$, $SE(t) \geq E(t)$.*

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.

Definition 4.5 [SC88] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system. The structural marking bound of a given place p of $\mathcal{N}$ is*

$$SB(p) \stackrel{\text{def}}{=} \max \quad \{ \ M(p) \mid M = M_0 + C \cdot \vec{\sigma} \geq 0 \ , \ \vec{\sigma} \geq 0 \ \} \qquad \text{(LPP2)}$$

It is well-known that the structural marking bound of a place is, in general, greater than or equal to the marking bound of the same place (for instance, for the net in figure 4, the marking bound of p_2 is 1 while its structural marking bound is 2). For the particular case of live and bounded free choice systems, both the marking bound and the structural marking bound of a place are always the same [Esp90]. A similar result can be shown for the enabling bound, the structural enabling bound, and the liveness bound of transitions of such net subclass.

Theorem 4.1 [CCS91b] *Let $\langle \mathcal{N}, M_0 \rangle$ be a live and bounded free choice system. For any transition t in $\mathcal{N}$, $SE(t) = E(t) = L(t)$.*

Now, from the previous theorem and taking into account that for any transition t the computation of the structural enabling bound $SE(t)$ can be formulated in terms of the problem (LPP1), the following monotonicity property of the liveness bound of a transition with respect to the initial marking is obtained.

Corollary 4.1 [CCS91b] *If $\langle \mathcal{N}, M_0 \rangle$ is a live and bounded free choice system and $M'_0 \geq M_0$ then the liveness bound of t in $\langle \mathcal{N}, M'_0 \rangle$ is greater than or equal to the liveness bound of t in $\langle \mathcal{N}, M_0 \rangle$.*

The previous result appears to be a generalization (stated for the particular case of bounded nets) of the classical liveness monotonicity property for free choice systems (see, e.g., [Bes87]).

Once the computability of visit ratios and enabling/liveness bounds have been addressed using concepts and techniques from linear (algebra/programming) structure theory, bounds on throughput are presented in the following sections.

5 Insensitive upper bounds on throughput

In this section we present the computation of upper bounds for the throughput of transitions for stochastic Petri nets with general distribution of service times. The obtained bounds are called *insensitive* because they are valid for arbitrary forms of the probability distribution functions of service (including deterministic timing), since only mean values of random variables are used.

Let us just precise for stochastic Petri net systems the weak ergodicity notions for the marking and firing processes:

Definition 5.1 [CCS91b] *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:*

$$\lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau M_u \, du = \overline{M} < \vec{\infty}, \ \ a.s. \tag{14}$$

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:

$$\lim_{\tau \to \infty} \frac{\vec{\sigma}_\tau}{\tau} = \vec{\Sigma} < \vec{\infty}, \ \ a.s. \tag{15}$$

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

For bounded net systems, the existence of a home state (i.e., a marking reachable from any other) is a sufficient condition for weak marking ergodicity [Cam90].

5.1 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 [Lit61], which holds under very general (i.e., weak) conditions. This result can be applied to each place of a weakly ergodic net system. Denoting $\overline{M}(p_i)$ the limit average number of tokens at place p_i, $\vec{\Sigma}$ the limit vector of transition throughputs, and $\overline{R}(p_i)$ the average time spent by a token within the place p_i (average residence time at place p_i), the above mentioned relationship is stated as follows (see [FN85]):

$$\overline{M}(p_i) = (PRE[p_i] \cdot \vec{\Sigma})\ \overline{R}(p_i) \tag{16}$$

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.

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 the equation (16), $\vec{\Sigma}$ can be computed except for a scaling factor for important net system subclasses (see section 3):

$$\vec{\Sigma} = \frac{1}{\Gamma^{(j)}}\ \vec{v}^{(j)} \tag{17}$$

where $\vec{v}^{(j)}$ is the vector of visit ratios normalized for t_j and $\Gamma^{(j)}$ is the inverse of the limit throughput of transition t_j, that we call the *mean interfiring time* of that transition, i.e., the mean time between two consecutive firings of t_j.

The average residence time $\overline{R}(p_i)$ at places with more than one output transition is null because such transitions are considered immediate. For the places p_i with only one output transition, the average response time can be expressed as sum of the average waiting time due to a possible synchronization in the output transition and the mean service time associated with that transition. Thus the average residence times can be lowerly bounded from the knowledge of the mean service times of transitions, s_i, $i = 1, \ldots, m$, and the following system of inequalities can be derived from (16) and (17):

$$\Gamma^{(j)}\ \overline{M} \geq PRE \cdot \vec{D}^{(j)} \tag{18}$$

where $\vec{D}^{(j)}$ is the vector of average service demands, introduced in section 3.1: $D_k^{(j)} = v_k^{(j)} s_k$.

The limit average marking $\overline{M}$ is unknown. However, taking the product with a P-semiflow Y (i.e., $Y \geq 0$, $Y^T \cdot C = 0$, thus $Y^T \cdot M_0 = Y^T \cdot M = Y^T \cdot \overline{M}$ for all reachable marking M), the following inequality can be derived:

$$\Gamma^{(j)} \geq \max \left\{ \frac{Y^T \cdot PRE \cdot \vec{D}^{(j)}}{Y^T \cdot M_0} \mid Y^T \cdot C = 0\ ,\ Y \geq 0 \right\} \tag{19}$$

The previous lower bound for the mean interfiring time (or its inverse, an upper bound for the throughput) can be formulated in terms of a *fractional programming problem* [NRKT89] and later, after some considerations, transformed into a linear programming problem.

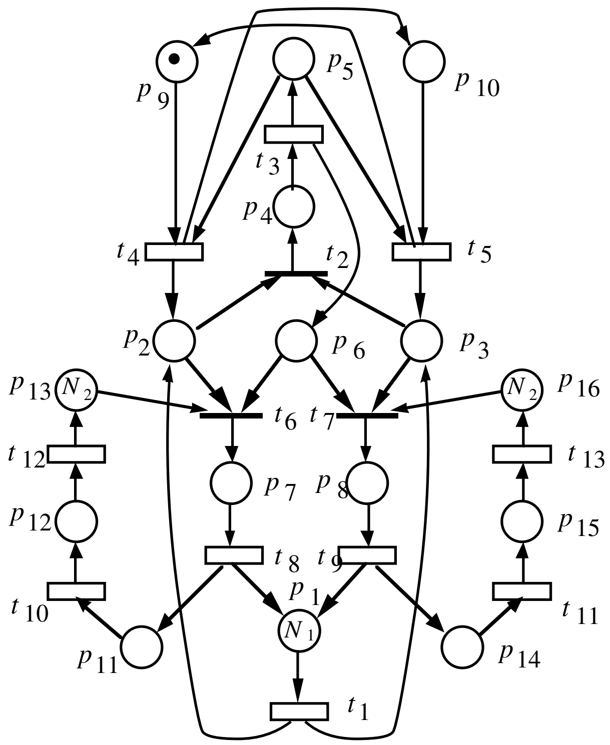


Figure 5: A live and bounded stochastic Petri net.

Theorem 5.1 [CCS91b] *For any net system, a lower bound for the mean interfiring time* $\Gamma^{(j)}$ *of transition* t_j *can be computed by solving the following linear programming problem:*

$$\Gamma^{(j)} \;\; \geq \max \;\; \{Y^T \cdot PRE \cdot \vec{D}^{(j)} | Y^T \cdot C = 0 \;,\; Y^T \cdot M_0 = 1 \;,\; Y \geq 0\} \qquad \text{(LPP3)}$$

The basic advantage of the previous theorem 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 [NRKT89]. Since for live and bounded free choice systems the computation of vector $\vec{v}^{(j)}$ (hence of $\vec{D}^{(j)}$) can be done by solving a linear system of equations (cf. theorem 3.1), the computation of a lower bound for the mean interfiring time (thus, of the upper bound on throughput) of a transition has worst case polynomial complexity on the net size.

In order to interpret theorem 5.1, let us consider the *mono-T-semiflow net* [CCS91a] depicted in figure 5. The unique minimal T-semiflow of the net is:

$$X = (2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)^T \qquad (20)$$

Therefore, according to (4), the vector of average service demands for transitions (if the visit ratios are normalized, for instance, for t_3) is

$$\vec{D}^{(3)} = (2s_1, 0, 2s_3, s_4, s_5, 0, 0, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13})^T \qquad (21)$$

because the vector of visit ratios is $\vec{v}^{(3)} = X$ (see section 3) and transitions t_2, t_6, and t_7 are assumed to be immediate ($s_2 = s_6 = s_7 = 0$).

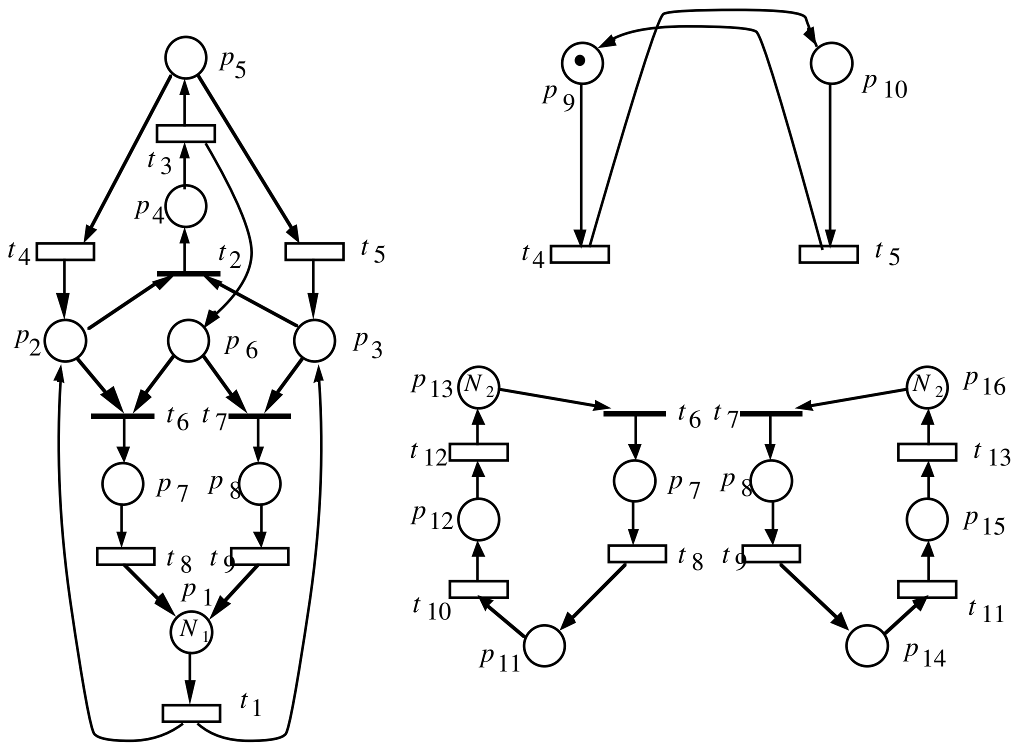


Figure 6: Subsystems of the net system in figure 5 generated by minimal P-semiflows.

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

$$\begin{aligned} Y_1 &= (2,1,1,2,1,1,2,2,0,0,0,0,0,0,0,0)^T \\ Y_2 &= (0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0)^T \\ Y_3 &= (0,0,0,0,0,0,1,0,0,0,1,1,1,0,0,0)^T \\ Y_4 &= (0,0,0,0,0,0,0,1,0,0,0,0,0,1,1,1)^T \end{aligned} \tag{22}$$

Then, the application of (LPP3) gives:

$$\begin{aligned} \Gamma^{(3)} \;\geq\; \max \{ \; & (4s_1 + 2s_3 + 2s_4 + s_5 + 2s_8 + 2s_9)/2N_1, \\ & s_4 + s_5, \\ & (s_8 + s_{10} + s_{12})/N_2, \\ & (s_9 + s_{11} + s_{13})/N_2 \; \} \end{aligned} \tag{23}$$

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 6.

The exact mean interfiring times of (all the transitions 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 interfiring time of t_3 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 $(4s_1+2s_3+2s_4+s_5+2s_8+2s_9)/2N_1$, 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.

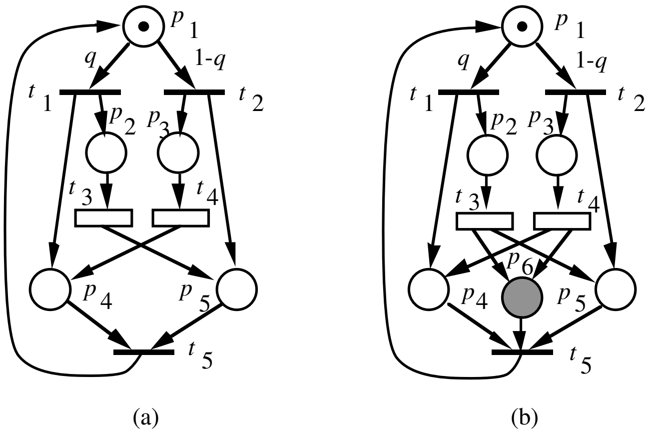


Figure 7: (a) A live and safe free choice system and (b) the addition of the implicit place p_6.

Therefore, the lower bound for the mean interfiring time of t_4 in the original net given by (23) is computed *looking at the "slowest subnet"* (net with minimum throughput for t_4) *generated by the elementary P-semiflows, considered in isolation.*

In the particular case of strongly connected marked graphs, the problem of finding an upper bound for the steady-state throughput (lower bound for the mean interfiring time) can be solved looking at the *mean interfiring time* associated with each elementary circuit (minimal P-semiflows for marked graphs) of the net, considered in isolation. These times can be computed making the summation of the mean service times of all the transitions involved in the P-semiflow (service time of the whole circuit), and dividing by the number of tokens present in it (customers in the circuit).

5.2 About the reachability of the bound

The above bound, that holds for any probability distribution function of service times of transitions, happens to be the same that has been obtained for strongly connected deterministically timed marked graphs by other authors (see for example [Ram74, Sif78, RH80]), but here it is considered in a practical linear programming form. For deterministically timed marked graphs, the reachability of this bound has been shown [Ram74, RH80]. Even more, it has been shown [CCCS89, CCCS90] that the previous bound cannot be improved, for the case of strongly connected marked graphs, only on the base of the knowledge of the coefficients of variation for the transition service times.

We remark that the importance of a tightness result for performance bounds lies in the fact that the bounds cannot be improved without increasing the information about the model (in particular, the moments of order greater than two of the associated random variables).

For the more general case of live and bounded free choice systems, the bound given by theorem 5.1 cannot be reached for some models, for any probability distribution function of service times. Let us consider, for instance, the live and safe free choice system in figure 7.a.

Let s_3 and s_4 be the mean 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 at place p_1. The vector of visit ratios (normalized for t_5) is

$$\vec{v}^{(5)} = (q, 1-q, q, 1-q, 1)^T \tag{24}$$

The elementary P-semiflows are

$$\begin{aligned} Y_1 &= (1,1,0,0,1)^T \\ Y_2 &= (1,0,1,1,0)^T \end{aligned} \tag{25}$$

Applying (LPP3) the following lower bound for the mean interfiring time of transition t_5 is obtained:

$$\Gamma^{(5)} \geq \max \{ \; qs_3, (1-q)s_4 \; \} \tag{26}$$

while the actual mean interfiring time for this transition is

$$\Gamma^{(5)} = qs_3 + (1-q)s_4 \tag{27}$$

independently of the higher moments of the probability distribution functions associated with transitions t_3 and t_4. Therefore the bound given by theorem 5.1 is non-reachable for the net system in figure 7.a.

Methods for the improvement of this bound have been presented in [CCS91c] and [CC91]. We just summarize here some ideas about them. The first one concerns the addition of *implicit places* to the net system. From a pure qualitative point of view, in [CS91] it is shown that the addition of "judicious" implicit places eliminates some of the *spurious* solutions of the linear relaxation of a net system (i.e., those integer solutions of $M = M_0 + C \cdot \vec{\sigma} \geq 0, \vec{\sigma} \geq 0$ that are non reachable). In [CCS91b], an analogous improvement is shown to hold at the performance level. Let us explain here this improvement by using again the example depicted in figure 7.a: consider the net in figure 7.b, where the implicit place p_6 has been added to the original net. The addition of implicit places can generate more elementary P-semiflows. In this case:

$$Y_3 = (1,1,1,0,0,1)^T \tag{28}$$

Then, the application of (LPP3) can eventually lead to an improvement of the previous bound. For this net system:

$$\Gamma^{(5)} \geq \max \{ \; qs_3, (1-q)s_4, qs_3 + (1-q)s_4 \; \} = qs_3 + (1-q)s_4 \tag{29}$$

which is exactly the actual mean interfiring time of t_5.

Details on this technique can be found in [CCS91c]. Moreover, it should be pointed out that the addition of implicit places does not guarantee the bound to be reachable. As an example, look at the net in figure 8.a. The exact mean interfiring time of t_7 for deterministic timing is:

$$\Gamma^{(7)} = \max \{ \; qs_3+s_6, (1-q)s_4+s_5, qs_3+(1-q)s_4+(1-q)s_5+qs_6, qs_5+(1-q)s_6 \; \}+s_7 \tag{30}$$

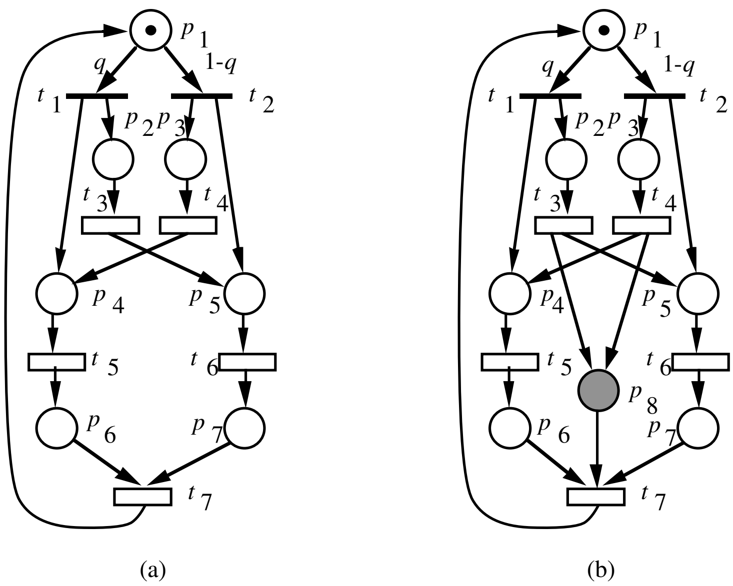


Figure 8: (a) A live and safe free choice system and (b) the addition of the implicit place p_8.

and its clearly greater than the bound obtained after the addition of the implicit place p_8 (figure 8.b):

$$\Gamma^{(7)} \geq \max \; \{ \; qs_3 + s_6, (1-q)s_4 + s_5, qs_3 + (1-q)s_4 \; \} + s_7 \tag{31}$$

The reader can check that addition of any set of implicit places to the system in figure 8.a does not allow to reach the value in (30).

Another approach for improving the throughput upper bound of theorem 5.1 is presented in [CC91], for the case of live and safe free choice systems. 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 5.1 for the case of marked graphs, because circuits (P-semiflows) of the marked graph are substituted now by multisets of circuits. The improvement is based on the application of a linear programming problem to a net obtained from the original one after a *transformation* of linear size increasing. The transformation, that is a modification of the *Lautenbach transformation for the computation of minimal traps in a net* [Lau87], will not be presented here (interested readers are referred to [CC91]). The application of this method to the net in figure 8.a gives exactly the mean interfiring time of t_7, given by (30), for deterministic service times of transitions.

5.3 Some derived results

Linear programming problems give an easy way to derive results and interpret them. Looking at (LPP3), the following *monotonicity property* can be obtained: the lower bound for the mean interfiring time of a transition does not increase if $\vec{s}$ (the mean service times vector) decreases or if M_0 increases.

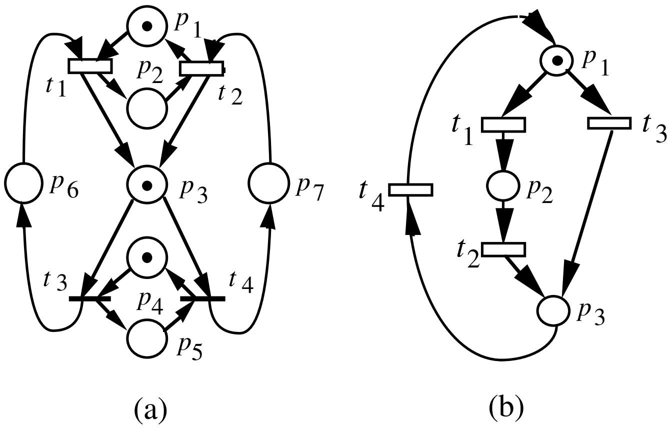


Figure 9: "Apparent improvements" lead to worse results: (a) The addition of a token to p_5 kills the net system (the sequence $\sigma = t_4$ leads to a deadlock); (b) Decreasing s_1 (mean service time of t_1), the mean interfiring time of t_4 increases if $s_2 \gg s_i$, $i = 1, 3$, and if exponential services and race policy are assumed.

Property 5.1 [CCCS89] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system and $\vec{s}$ the vector of mean service times of transitions.*

1. *For a fixed $\vec{s}$, if $M_0' \geq M_0$ (i.e., more resources) then the lower bound for the mean interfiring time of any transition of $\langle \mathcal{N}, M_0', \vec{s} \rangle$ computed through (LPP3) is less than or equal to the one of $\langle \mathcal{N}, M_0, \vec{s} \rangle$.*

2. *For a fixed M_0, if $\vec{s'} \leq \vec{s}$ (i.e., faster resources) then the lower bound for the mean interfiring time of any transition of $\langle \mathcal{N}, M_0, \vec{s'} \rangle$ computed through (LPP3) is less than or equal to the one of $\langle \mathcal{N}, M_0, \vec{s} \rangle$.*

For the case of live and bounded free choice systems, the above *monotonicity* properties for the bound hold also for the *exact* throughput. We recall that live and bounded free choice net systems can be decomposed into several strongly connected state machines (*P-components*) connected by means of synchronization transitions [Bes87]. Moreover, from the definition of free choice nets, if p_a and p_b are input places to a synchronization transition t, then t is the unique output transition of p_a and p_b. In other words, once a synchronization transition has been enabled in a free choice system, its firing is unavoidable. Then, because we assume (section 2.1) that all choices are among immediate transitions, if the service time of a transition decreases or the number of tokens at some place increases, the mean interfiring time of transitions can never increase: it is possible to increase for certain tokens the pure waiting time at some synchronizations (i.e., the time elapsed from the time instant in which the tokens arrive until the transition becomes enabled).

Property 5.2 *Let $\langle \mathcal{N}, M_0 \rangle$ be a live and bounded free choice system and $\vec{s}$ the vector of mean service times of transitions.*

1. *For a fixed $\vec{s}$, if $M_0' \geq M_0$ (i.e., more resources) then the mean interfiring time of any transition of $\langle \mathcal{N}, M_0', \vec{s} \rangle$ is less than or equal to the one of $\langle \mathcal{N}, M_0, \vec{s} \rangle$ (i.e., $\Gamma^{(j)'} \leq \Gamma^{(j)}$).*

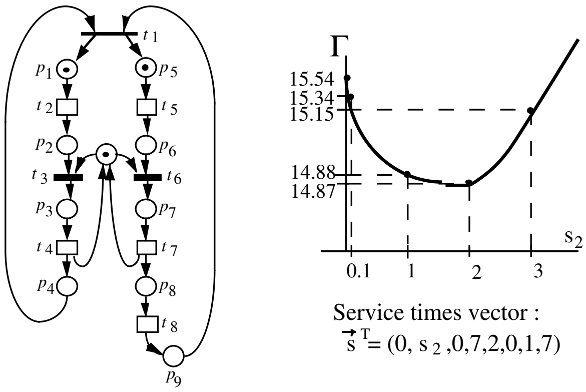


Figure 10: Increasing the mean service time s_2 of t_2 on the Markovian net system leads to better overall throughput (i.e., the mean interfiring time Γ of every transition decreases).

2. *For a fixed M_0, if $\vec{s'} \leq \vec{s}$ (i.e., faster resources) then the mean interfiring time of any transition of $\langle \mathcal{N}, M_0, \vec{s'} \rangle$ is less than or equal to the one of $\langle \mathcal{N}, M_0, \vec{s} \rangle$ (i.e., $\Gamma^{(j)'} \leq \Gamma^{(j)}$).*

The result stated in the above property does not hold for other net subclasses. For instance, increasing the number of initial resources (by adding one token to p_5) in the net system of figure 9.a, the obtained system reaches a *total deadlock*, therefore the throughput of the derived system is null.

On the other hand, the intuitive idea that decreasing the service time of a transition leads to a slower system is paradoxically wrong in general! Figure 10 shows a *Markovian Petri net system* (exponentially distributed service times of transitions) where increasing the mean service time s_2 of t_2, while $s_2 \in (0, 2)$, the throughput increases. Moreover, the statement 2 in the property 5.2 does not hold even for state machine topology using the basic stochastic interpretation in [Mol82] or [FN85]: all transitions are timed with exponential PDF and conflicts are solved with race policy. This "anomaly", illustrated in the extremely simple case of figure 9.b, appears because the routing and timing are *coupled* in this class of models.

Finally, an interesting interpretation of the problem (LPP3), that provides another example of possible interleaving between qualitative and quantitative analysis for stochastic net systems, is the following characterization of liveness for structurally live and structurally bounded free choice nets.

Corollary 5.1 *Assuming that $\vec{v}^{(j)} > 0$ and that there do not exist circuits containing only immediate transitions, liveness of structurally live and structurally bounded free choice nets can be decided in polynomial time, checking the boundedness of the problem (LPP3).*

This result is nothing more than deciding liveness by checking if all P-semiflows are marked (the linear programming problem is bounded if and only if for all $Y \geq 0$ such that $Y^T \cdot C = 0$, then $Y^T \cdot M_0 > 0$).

For more general net subclasses, if the solution of (LPP3) is unbounded (i.e., there exists an unmarked P-semiflow), since $\Gamma^{(j)}$ it is a lower bound for the mean interfiring time of transition t_j, the non-liveness can be assured (infinite interfiring time). Nevertheless, a net system can be non-live and the obtained lower bound for the mean interfiring time be finite (e.g., the mono-T-semiflow net in figure 9.a with the addition of a token in place p_5).

6 Insensitive lower bounds on throughput

In this section, lower bounds on throughput are presented, independent of the higher moments of the service time probability distribution functions, based on the computation of the vector of visit ratios for transitions as introduced in section 3 and on the transition liveness bounds, defined in section 4.

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. 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 upper bound, which in this case gives the same value).

This trivial lower bound has been improved in [CCS91b] for the case of live and bounded free choice systems based on the knowledge of the liveness bound $L(t)$ for all transitions t of the net system.

Theorem 6.1 [CCS91b] *For any live and bounded free choice system, an upper bound for the mean interfiring time* $\Gamma^{(j)}$ *of transition* t_j *can be computed as follows:*

$$\Gamma^{(j)} \leq \sum_{i=1}^{m} \frac{D_i^{(j)}}{L(t_i)} = \sum_{i=1}^{m} \frac{v_i^{(j)} s_i}{L(t_i)} \tag{32}$$

We recall (cf. theorem 4.1) that in the case of live and bounded free choice systems:

1. The liveness bound equals the structural enabling bound for each transition (Theorem 4.1) and this one can be computed by solving (LPP1).

2. The vector of visit ratios for transitions is obtained by solving the linear system of equations (10).

Therefore, the lower bound for the throughput of live and bounded free choice systems can be computed efficiently. Its worst case complexity is *polynomial time* on the net size.

The lower bound in performance given by the computation of theorem 6.1 can be shown [CCCS89] to be reachable for any marked graph topology and for some assignment of PDF to the service time of transitions. Therefore, if nothing but the average value is known about the PDF of the service time of transitions, the bound provided by Theorem 6.1 is *tight*.

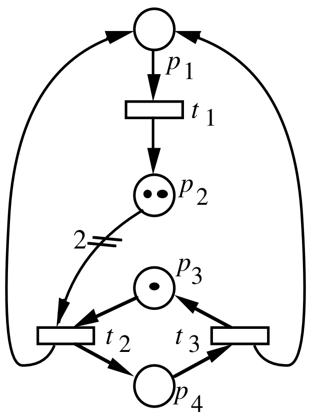


Figure 11: "Non-trivial" upper bound for the mean interfiring time cannot be applied.

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

An example showing that the bound presented in theorem 6.1 is not valid for non-free choice net systems is depicted in figure 11, where s_1, s_2, s_3 are the mean service times of transitions t_1, t_2, t_3, respectively. For this net, the vector of visit ratios normalized for transition t_2 is

$$\vec{v}^{(2)} = (2, 1, 1)^T \tag{33}$$

and the liveness bounds of transitions are given by $L(t_1) = 2$, $L(t_2) = 1$, and $L(t_3) = 1$. Thus, the theorem 6.1 would give the bound:

$$\Gamma^{(2)} \leq s_1 + s_2 + s_3 \tag{34}$$

If exponentially distributed random variables (with means s_1, s_2, s_3; $s_1 \neq s_3$) are associated with transitions, the steady-state mean interfiring time for transition t_2 is

$$\Gamma^{(2)} = s_1 + s_2 + s_3 + \frac{s_1^2}{2(s_1 + s_3)} \tag{35}$$

which is greater than the value obtained from the theorem 6.1, thus the "non-trivial" bound does not hold in general.

7 Throughput bounds for Markovian Petri net systems

In sections 5 and 6, insensitive bounds (valid for any probability distribution function of service times and for any conflict resolution policy) on throughput have been presented. The quality of the bounds is poor in some cases due to the fact that only mean values of the involved random variables have been used for the computation. In order to improve the bounds, it will be necessary to take into account more information from the form of the probability distribution functions.

Exponential distribution of service times 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, the maximum number of servers being defined by the liveness of the transitions. Marking independent discrete probability distributions are used for defining the solution of conflicts among immediate transitions. The general techniques in the literature for the analysis of this particular case of stochastic Petri net models 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 [AMBC84].

In summary, in this section we present better bounds for stochastic Petri net systems with exponentially distributed timed transitions, $L(t)$-server semantics, and marking independent discrete probability distributions for the resolution of conflicts, that we already call (section 2.1) *Markovian Petri net systems*, for short.

7.1 Embedded queueing networks

Insensitive lower bounds for the mean interfiring time of transitions were introduced in section 5 looking for the maximum of the mean interfiring time of transitions of isolated subnets generated by elementary P-semiflows. A more realistic computation of the mean 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 (section 4). 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 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 [CS92]) 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 [RL80, ZSEG82, ES83].

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: $PRE[t_1] = PRE[t_2]$. 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.

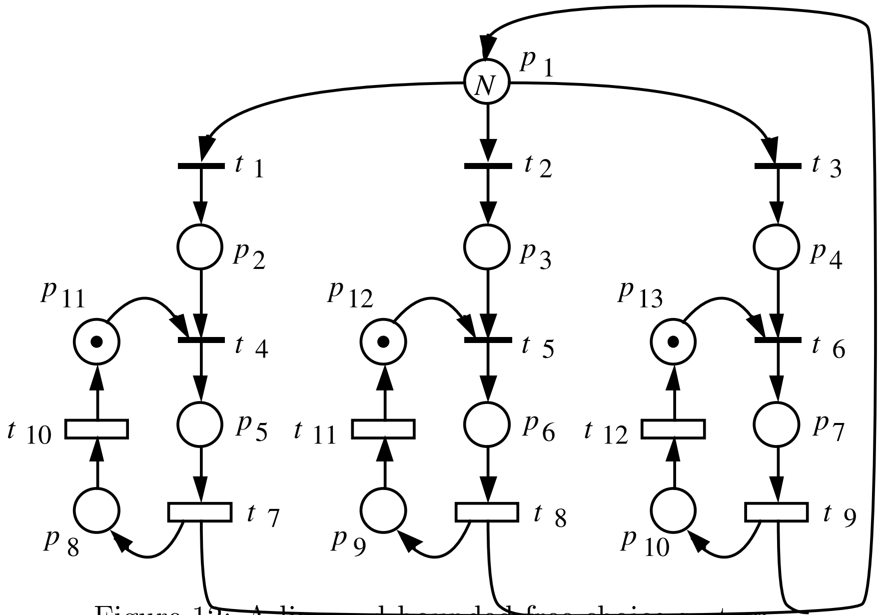


Figure 12: A live and bounded free choice system.

Definition 7.1 *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 $\mathcal{N}$: $PRE[t_j] = PRE[t_k]$.*

An improvement of the insensitive lower bound for the mean interfiring time of a transition t_j computed in theorem 5.1 can be eventually obtained computing the exact mean interfiring time of that transition in the RP-component generated by a minimal P-semiflow Y, with $L(t)$–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 Y can be selected among the optimal solutions of (LPP3) or it can be just a feasible *near-optimal* solution.

As an example, let us consider the net system depicted in figure 12. 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 elementary P-semiflows of the net are:

$$\begin{array}{l} Y_1 = (1,1,1,1,1,1,1,0,0,0,0,0,0)^T \\ Y_2 = (0,0,0,0,1,0,0,1,0,0,1,0,0)^T \\ Y_3 = (0,0,0,0,0,1,0,0,1,0,0,1,0)^T \\ Y_4 = (0,0,0,0,0,0,1,0,0,1,0,0,1)^T \end{array} \tag{36}$$

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 (LPP3) is

$$\Gamma^{(1)}_{(\mathrm{LPP3})} = \max\{30/N, 11, 11, 11\} \tag{37}$$

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 from Y_1 (mean

N	$\Gamma^{(1)}$	$\Gamma^{(1)}_{(Y_1)_L}$	$\Gamma^{(1)}_{(\mathrm{LPP3})}$
1	31.06	30	30
2	21.05	20	15
3	17.71	16.67	11
4	16.03	15	11
5	15.03	14	11
10	13.02	12	11
15	12.35	11.34	11

Table 2: Exact mean interfiring time of t_1, bounds obtained using (LPP3), and the improvements presented in this section, for different initial markings of p_1 in the net system of figure 12.

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 (LPP3) 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_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 [RL80]. This algorithm has $O(A^2B)$ 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). 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* [ZSEG82] or *throughput upper bounds hierarchies* [ES83]. 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 mean 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* [BCMP75].

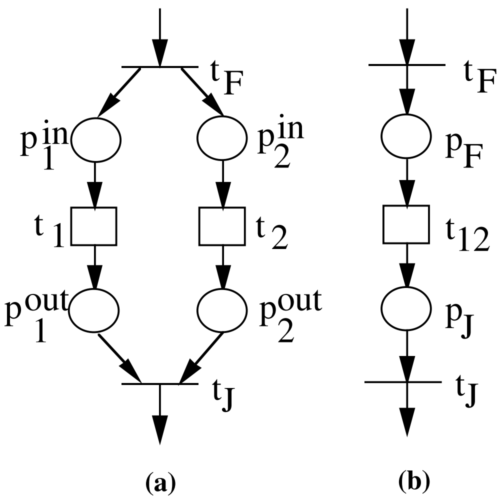


Figure 13: (a) Elementary fork-join and (b) its reduction.

7.2 Transformation techniques

The lower bounds for the throughput of transitions presented in section 6 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). This property is not always true as already mentionned (see, for instance, figure 10). Using the concept of *stochastic ordering* [Ros83], a pessimistic transformation is, for example, to substitute the PDF of a service (or token-subnet traversing) time by a *stochastically greater* PDF. Live and bounded free choice is a class of systems for which the above performance monotonicity property holds (property 5.2 is a particular case). Details about the techniques presented here can be found in [CSS91]. 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 section 6, exact analysis, or any other applicable technique.

In order to obtain better bounds (after these two steps) than the values computed in section 6, 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 13.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 13.b) with exponential service time and mean:

$$s_{12} = E[\max\{X_1, X_2\}] = s_1 + s_2 - \left(\frac{1}{s_1} + \frac{1}{s_2}\right)^{-1} \quad (38)$$

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 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 [CSS91] 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 14.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 14.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 14.c).

An example of application of all above transformation rules is depicted in figure 14 for a strongly connected marked graph with exponential timing. Let us assume that mean service times of transitions are: $s_i = 1$, $i = 1, 2, 3, 7, 8, 12, 13, 14$ and $s_i = 10$ otherwise.

In order to compute firstly the insensitive lower bounds on throughput introduced in section 6, it is necessary to derive the liveness bounds of transitions (section 4). In this case it is easy to see that $L(t_j) = 2$ for every transition t_j.

The vector of visit ratios of a marked graph is the unique minimal T-semiflow of the net: $\vec{v}^{(j)} = \mathbb{1}$, for all transitions t_j. Therefore, the insensitive upper bound (valid for any probability distribution function of service times) of the mean interfiring time of any transition of the net system is $\Gamma \leq 34$. This value can be reached for some distributions of service times (see comment on tightness on section 6). Nevertheless, if services are exponential the exact mean interfiring time of transitions is $\Gamma = 14.15$.

The quantitative results of the transformation process illustrated in figure 14 are shown in table 3. We remark that the bound has been improved in polynomial time from 34 to 19.2.

8 Bounds for other performance indices

Up to this point we just concentrated on throughput bounds. The purpose of this section is to bring the idea that given some throughput bounds, bounds for other performance

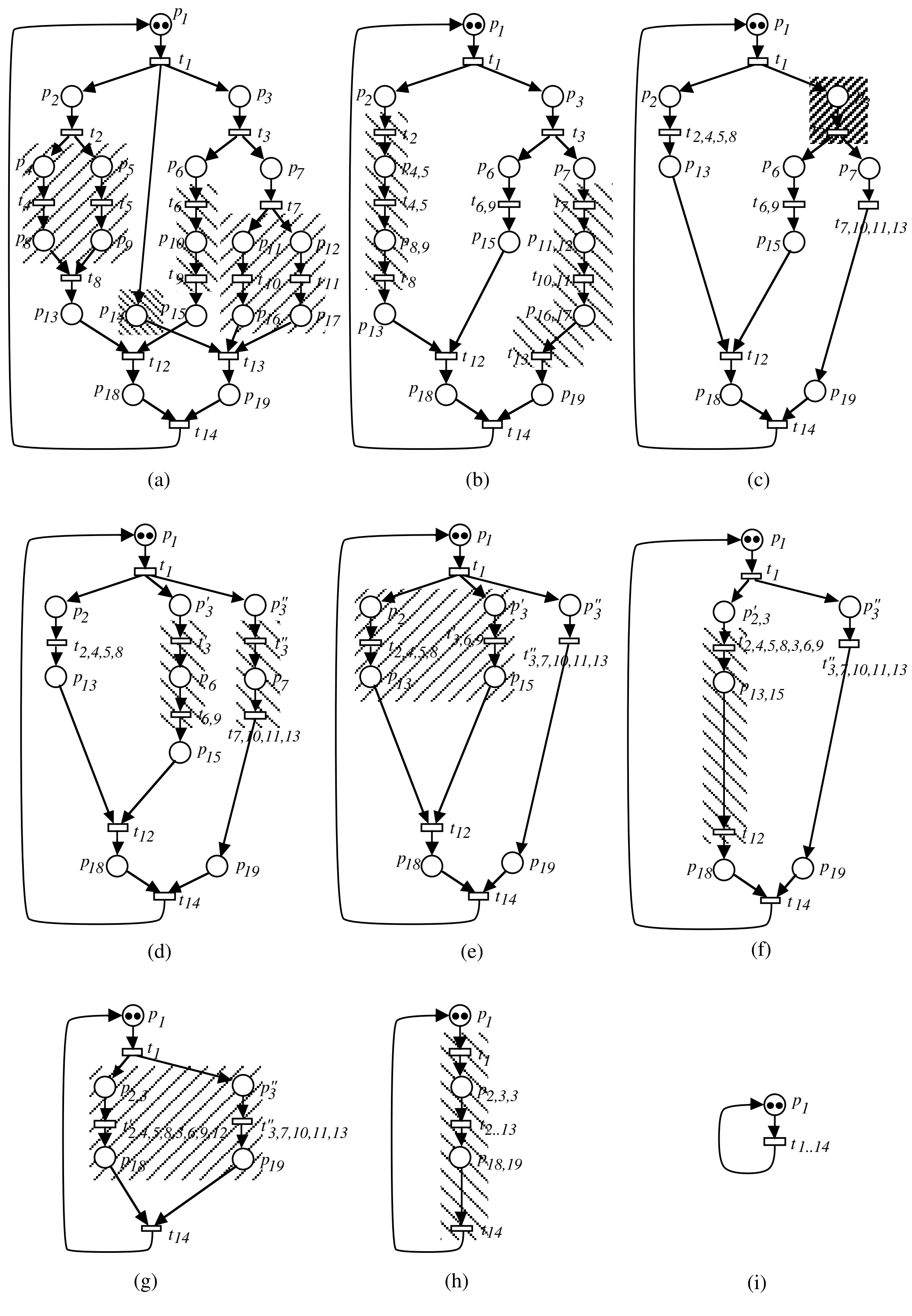


Figure 14: A complete reduction process. The relative error between insensitive bound and exact value diminishes from 140% to 35%.

System	Γ^{ub}	Relative error
Fig. 14.a	34	140 %
Fig. 14.b	29	105 %
Fig. 14.c	29	105 %
Fig. 14.d	29.5	108 %
Fig. 14.e	29.5	108 %
Fig. 14.f	24.8	75 %
Fig. 14.g	24.8	75 %
Fig. 14.h	19.2	35 %
Fig. 14.i	19.2	35 %

Table 3: Successive improvements of the upper bound for the mean interfiring time of transitions of the net in figure 14 and relative errors with respect to the exact value $\Gamma = 14.15$.

indices can be computed using classical formulas in QN theory such as Little's formula.

The number of tokens in a place defines the lenght of the represented queue (including the customers in service!). Thus it may be important to known bounds on average marking of places.

As an example, in [CCS91b] it has been shown that the following are lower and upper bounds for the average marking, $\overline{M}$:

$$\overline{M}^{lb} = PRE \cdot \mathcal{S} \cdot \vec{\Sigma}^{lb} \tag{39}$$

$$\overline{M}^{ub}(p) \quad = \quad \max \; \{ \; M(p) \mid B^T \cdot M = B^T \cdot M_0 \; , \; M \geq \overline{M}^{lb} \; \} \tag{LPP4}$$

where $\mathcal{S} = \text{diag } (s_i)$ and the rows of B^T are the basis of left annullers of C (the incidence matrix of the net).

As an interesting remark, the reader can check that a structural absolute bound for the marking of a place is given for conservative nets (i.e., $\exists Y > 0$, $Y^T \cdot C = 0$) by the following expression:

$$SB(p) \quad = \quad \max \; \{ \; M(p) \mid B^T \cdot M = B^T \cdot M_0 \; , \; M \geq 0 \; \} \tag{LPP5}$$

The constraint in (LPP5) being weaker than that in (LPP4) ($M \geq \overline{M}^{1b}$ is transformed into $M \geq 0$), it is obvious that $\overline{M}^{ub} \leq SB(p)$.

9 Conclusions

The main motivation to write this paper has been to show how the qualitative theory (and in particular the structural analysis techniques) of net systems can be useful in performance bounds computation for stochastic Petri net models. Several interesting questions that were easily answered in the past for classical queueing networks turn into non-trivial problems (ergodicity, visit ratios, number of servers at stations, performance bounds) in the case of stochastic Petri net systems. However, structural techniques can be used in order to solve those problems in an efficient way, at least for important subclasses of net systems. Moreover, the benefits of this approach have been not only for the quantitative understanding of the models but also for the qualitative point of view: some fundamental new results have appeared as by-products of the performance perspective. We remark the following points among those presented in this survey:

1. The (quoted) equation "SPNs = PNs + time = QNs + synchronization" must be always in mind in performance evaluation of stochastic Petri nets, since both qualitative theory of Petri nets and results from queueing theory are needed in order to solve the stated problems.

2. The computability of the vector of visit ratios from the different system parameters (structure, routing policy, initial marking, and service times) induces a new hierarchy of nets, being re-encountered well-known subclasses (e.g., marked graphs, free choice nets). Specially important appear those nets as FRT and subclasses (free choice, marked graphs…), whose vector of visit ratios does not depend neither on the initial marking nor in the service times (i.e., the vector of visit ratios can be computed using only structural and routing information).

3. The rank theorem [CCS91b, ES91], suggested by the performance approach, has important consequences in the qualitative theory of nets. Some extensions of that theorem appear in [CCS90b] for the characterization of structural liveness in general nets.

4. 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.

5. The liveness bound is also a quantitative generalization of the classical concept of liveness.

6. Performance bounds can be derived from structural components and properties: P-semiflows, T-semiflows, multisets of circuits, structural enabling bounds…

7. An step to extend the classical theory of qualitative transformation/reduction of nets has been achieved, including quantitative aspects, for deriving throughput bounds of Markovian Petri net system.

8. As in the case of (qualitative) structural theory of net systems, the derived performance-oriented results are specially powerful for some well-known subclasses (e.g., live and bounded free choice net systems).

Additional results to those presented here can be found in [CS90], related with exact performance analysis of a net subclass. In particular, for *totally open deterministic systems of Markovian sequential processes*, exact computation of limit throughput can be done in polynomial time, assuming *consistency* of the net and some *synchronic distance relations* among transitions.

In sections 5, 6, and 7, we focused on the computation of throughput bounds. From these results and using classical laws from queueing theory, bounds can be derived for other interesting performance indices (section 8), such as the *mean queue lengths* (or mean marking of places), or *mean response times* (or mean residence times of tokens at places).

In this paper, we have tried to clearly state concepts and results, illustrating them by means of examples and omitting the proofs (for more technical presentations the readers should consult the references). The net based examples have been choosed to illustrate the theory. Applications of some of the presented results in the manufacturing domain are considered in [CCS90a].

We are far from solving many of the problems related with performance analysis of Petri net systems. However, what is clear now is that such problems must be attacked by deeply bridging qualitative and quantitative aspects of the model and making use of several active fields like Petri net theory, graph theory, linear algebra, convex geometry, queueing networks, and applied stochastic processes.

Acknowledgements

This work was partially supported by the DEMON Esprit Basic Research Action 3148 and the Spanish Plan Nacional de Investigación, Grant PRONTIC-0358/89. The authors are indebted to G. Chiola of the Università di Torino and J. M. Colom of the Universidad de Zaragoza, co-authors of many of the results surveyed in this work. The comments and suggestions of three referees helped us for the improvement of the submitted draft.

References

[AM90] M. Ajmone Marsan. Stochastic Petri nets: An elementary introduction. In G. Rozenberg, editor, *Advances in Petri Nets 1989*, volume 424 of *LNCS*, pages 1–29. Springer-Verlag, Berlin, 1990.

[AMBB+89] 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.

[AMBC84] 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.

[AMBCC87] M. Ajmone Marsan, G. Balbo, G. Chiola, and G. Conte. Generalized stochastic Petri nets revisited: Random switches and priorities. In *Proceedings of the International Workshop on Petri Nets and Performance Models*, pages 44–53, Madison, WI, USA, August 1987. IEEE-Computer Society Press.

[AMBCD86] 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.

[BCMP75] 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.

[Bes87] E. Best. Structure theory of Petri nets: The free choice hiatus. In W. Brauer, W. Reisig, and G. Rozenberg, editors, *Advances in Petri Nets 1986 - Part I*, volume 254 of *LNCS*, pages 168–205. Springer-Verlag, Berlin, 1987.

[Cam90] 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.

[CC91] 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.

[CCCS89] 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.

[CCCS90] 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. To appear in *IEEE Transactions on Circuit and Systems*.

[CCS90a] J. Campos, J. M. Colom, and M. Silva. Performance evaluation of repetitive automated manufacturing systems. In *Proceedings of the Rensselaer's Second International Conference on Computer Integrated Manufacturing*, pages 74–81, Rensselaer Polytechnic Institute, Troy, NY, USA, May 1990. IEEE-Computer Society Press.

[CCS90b] J. M. Colom, J. Campos, and M. Silva. On liveness analysis through linear algebraic techniques. In *Proceedings of the Annual General Meeting of ESPRIT Basic Research Action 3148 Design Methods Based on Nets (DEMON)*, Paris, France, June 1990.

[CCS91a] 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.

[CCS91b] 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.

[CCS91c] 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*.

[CS90] 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.

[CS91] 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.

[CS92] J. Campos and M. Silva. Embedded queueing networks and the improvement of insensitive performance bounds for Markovian Petri net systems. Research Report GISI-RR-92-10, Departamento de Ingeniería Eléctrica e Informática, Universidad de Zaragoza, Spain, February 1992.

[CSS91] 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.

[DLT90] Y. Dallery, Z. Liu, and D. Towsley. Equivalence, reversibility and symmetry properties in fork/join queueing networks with blocking. Technical report, MASI 90-32, University Paris 6, 4 Place Jussieu, Paris, France, June 1990.

[ES83] D. L. Eager and K. C. Sevcik. Performance bound hierarchies for queueing networks. *ACM Transactions on Computer Systems*, 1(2):99–115, May 1983.

[ES91] 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.

[Esp90] J. Esparza. *Structure Theory of Free Choice Nets*. PhD thesis, Departamento de Ingeniería Eléctrica e Informática, Universidad de Zaragoza, Spain, June 1990. Research Report GISI-RR-90-03.

[FFN91] G. Florin, C. Fraize, and S. Natkin. Stochastic Petri nets: Properties, applications and tools. *Microelectronics and Reliability*, 31(4):669–698, 1991.

[FN85] G. Florin and S. Natkin. Les réseaux de Petri stochastiques. *Technique et Science Informatiques*, 4(1):143–160, February 1985. In French.

[GN67] W. J. Gordon and G. F. Newell. Closed queueing systems with exponential servers. *Operations Research*, 15:254–265, 1967.

[Hac72] M. H. T. Hack. Analysis of production schemata by Petri nets. M. S. Thesis, TR-94, M.I.T., Boston, USA, 1972.

[Kle75] L. Kleinrock. *Queueing Systems Volume I: Theory.* John Wiley & Sons, New York, NY, USA, 1975.

[Kle76] L. Kleinrock. *Queueing Systems Volume II: Computer Applications.* John Wiley & Sons, New York, NY, USA, 1976.

[Lau87] K. Lautenbach. Linear algebraic calculation of deadlocks and traps. In K. Voss, H. Genrich, and G. Rozenberg, editors, *Concurrency and Nets*, pages 315–336. Springer-Verlag, Berlin, 1987.

[Lav89] S. S. Lavenberg. A perspective on queueing models of computer performance. *Performance Evaluation*, 10:53–76, 1989.

[Lit61] J. D. C. Little. A proof of the queueing formula $L = \lambda W$. *Operations Research*, 9:383–387, 1961.

[LZGS84] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative System Performance.* Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, 1984.

[Mol82] M. K. Molloy. Performance analysis using stochastic Petri nets. *IEEE Transaction on Computers*, 31(9):913–917, September 1982.

[Mol85] 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.

[Mur89] T. Murata. Petri nets: Properties, analysis, and applications. *Proceedings of the IEEE*, 77(4):541–580, April 1989.

[NRKT89] 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.

[Pet81] J.L. Peterson. *Petri Net Theory and the Modeling of Systems.* Prentice-Hall, Englewood Cliffs, NJ, USA, 1981.

[PNPM87] *Proceedings of the International Workshop on Petri Nets and Performance Models*, Madison, WI, USA, August 1987. IEEE-Computer Society Press.

[PNPM89] *Proceedings of the 3^{rd} International Workshop on Petri Nets and Performance Models*, Kyoto, Japan, December 1989. IEEE-Computer Society Press.

[PNPM91] *Proceedings of the 4^{rd} International Workshop on Petri Nets and Performance Models*, Melbourne, Australia, December 1991. IEEE-Computer Society Press.

[Ram74] C. Ramchandani. *Analysis of Asynchronous Concurrent Systems by Petri Nets.* PhD thesis, MIT, Cambridge, MA, USA, February 1974.

[RH80] C. V. Ramamoorthy and G. S. Ho. Performance evaluation of asynchronous concurrent systems using Petri nets. *IEEE Transactions on Software Engineering*, 6(5):440–449, September 1980.

[RL80] M. Reiser and S. S. Lavenberg. Mean value analysis of closed multichain queueing networks. *Journal of the ACM*, 27(2):313–322, April 1980.

[Ros83] S. M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, NY, USA, 1983.

[SC88] M. Silva and J. M. Colom. On the computation of structural synchronic invariants in P/T nets. In G. Rozenberg, editor, *Advances in Petri Nets 1988*, volume 340 of *LNCS*, pages 386–417. Springer-Verlag, Berlin, 1988.

[Sif78] J. Sifakis. Use of Petri nets for performance evaluation. *Acta Cybernetica*, 4(2):185–202, 1978.

[Sil85] M. Silva. *Las Redes de Petri en la Automática y la Informática.* Editorial AC, Madrid, 1985. In Spanish.

[SMK82] C. H. Sauer, E. A. MacNair, and J. F. Kurose. The research queueing package: past, present, and future. In *Proceedings of the 1982 National Computer Conference.* AFIPS, 1982.

[SY86] J. G. Shanthikumar and D. D. Yao. The effect of increasing service rates in a closed queueing network. *Journal of Applied Probability*, 23:474–483, 1986.

[TPN85] *Proceedings of the International Workshop on Timed Petri Nets*, Torino, Italy, July 1985. IEEE-Computer Society Press.

[VZL87] M. Vernon, J. Zahorjan, and E. D. Lazowska. A comparison of performance Petri nets and queueing network models. In *Proceedings of the* 3^{rd} *International Workshop on Modelling Techniques and Performance Evaluation*, pages 181–192, Paris, France, March 1987. AFCET.

[ZSEG82] 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.

[Zub91] W. M. Zuberek. Timed Petri nets: Definitions, properties and applications. *Microelectronics and Reliability*, 31(4):627–644, 1991.

[ZZ90] W. M. Zuberek and M. S. Zuberek. Transformations of timed Petri nets and performance analysis. In *Proceedings of the Midwest Simposium on Circuits and Systems'90 (Special Session on Petri Net Models)*, pages 1–5, 1990.